H24/3250

MONASH UNIVERSITY
THESIS ACCEPTED IN SATISFACTION OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
ON..... 4 October 2002 .....
.....[redacted].....
✓ Sec. Research Graduate School Committee

Under the copyright Act 1968, this thesis must be used only under the normal conditions of scholarly fair dealing for the purposes of research, criticism or review. In particular no results or conclusions should be extracted from it, nor should it be copied or closely paraphrased in whole or in part without the written consent of the author. Proper written acknowledgement should be made for any assistance obtained from this thesis.

EXAMINATION OF PHYSICOCHEMICAL PROPERTIES OF AMINO ACIDS WITHIN THE RESONANT RECOGNITION MODEL

A dissertation submitted for the award of the degree of

Doctor of Philosophy

by

Elena Pirogova

Bachelor in Chemical Engineering (Honours)

National Technical University of Ukraine, Kiev

Bioelectronics Group

Department of Electrical and Computer Systems Engineering

Monash University, Australia

November 2001

TABLE OF CONTENST

ABSTRACT

Proteins are one of the most complex and diverse classes of macromolecules found in the cell. They occur in all living cells, and their biological significance cannot be overemphasized. This fact was recognized by the German chemist G.T. Mulder in 1839, when he gave this class of compounds the name protein, which means "of prime importance".

Proteins are linear macromolecules built up of sequentially linked amino acids. All proteins are composed of the elements: carbon, nitrogen, oxygen and hydrogen. Most proteins also contain sulfur, and some have phosphorus and other elements, such as iron, zinc, or copper. Because of their diverse natures, proteins can be classified in several ways. Firstly proteins can be divided into two major classes: simple proteins, which yield only amino acids upon hydrolysis, and conjugated proteins, which yield amino acids and other organic and inorganic components upon hydrolysis. These other components are called prosthetic groups. A second classification is based on the physical characteristics of the protein molecule: globular proteins being soluble in water, and fibrous proteins insoluble in water. The biological importance of proteins results from their wide variety of functions. Proteins are the body's main dietary source of nitrogen and sulfur. In addition to their catalytic and structural functions, they make up the contractile system of muscles. As antibodies they are the defense system of the body, and as hormones they regulate the body's glandular activity. In the blood they maintain fluid balance, are part of the clotting mechanism, and transport oxygen and lipids. Proteins play crucial role in almost every biological process however they can only express their biological functions, interactions with their targets, when they achieve a certain active native conformation, so called three-dimensional structure (3-D). Both protein function and its active 3-D structure are determined by the sequence of amino acids in the protein molecule. There are many attempts to solve the problem of a protein's structure-function relationship.

There is evidence that biological processes can be induced or modulated by induction of light of particular haracteristic frequencies. The frequency selective effects of light on biological processes involving protein activation imply that protein activation involves energies of the same order and nature as electromagnetic irradiation of light. As there is evidence that certain charge could travel along the protein molecule, then charge moving through the protein backbone and passing different energy stages caused by different

amino acid side groups, can produce sufficient conditions for the specific electromagnetic radiation or absorption. These results lead to the conclusion that specificity of protein interactiionnare based on the resonant electromagnetic energy tansfer between interacting molecules, on a frequency specific for each observed function/interaction. These phenomena are analyzed in terms of the **Resonant Recognition Model (RRM)** [5,6]. It proposes that protein activities (interactions) are based on resonant electromagnetic energy transfer within the range of infrared and visible light. The RRM, which is central to this study, is a physicomathematical method for analysing protein and nucleic acid sequences [5,6]. Its aim is to find possible patterns of significance and thus assist in structure-function studies. The RRM is a model of protein-protein and protein-DNA interaction based on a significant correlation between spectra of the numerical presentation of amino acid or nucleotide sequences and their biological activity. The RRM analyses the correlation by converting the sequences of amino acids into numerical sequences and consequently transforming the numerical sequences into frequency domain using the Fourier transform.

The focus of this study is therefore directed at several things:

- To investigate the effect of various amino acid parameters on the determination of characteristic frequencies of the different analyzed protein groups

- To compare obtained results with the results of previous investigations where the Electron-Ion Interaction Potential (EIIP) was used to determine the characteristic profiles of biologically unrelated proteins in the RRM

- The investigation of the possible use of the experimentally measured parameter Ionization constant of amino acids (IC) instead of the EIIP within the RRM.

- The experimental measurement of dielectric and conduction properties of amino acids and analysis of their possible use for structure/function analysis of proteins within the RRM.

- To provide the physical basis for the EIIP.

This project is aimed at expanding an entirely new RRM approach to analysis of protein-protein interactions. Within the project different protein families were investigated using the Fourier transform. To convert protein sequences into relevant numbers of parameters related to the protein's biological function, the dielectric and conduction properties of 20 amino acid solutions were measured. As a result of this project an improved RRM model is proposed. This model is expected to become a valuable tool in analysis of proteins and DNA, in the modification and the design of new peptides with desired biological functions and thus it should be possible to produce new and more effective drugs and other biotechnological products.

ACKNOWLEDGMENT

The acknowledgment is paid to my supervisor Associate Professor Irena Cosic for her guidance, great support, encouragement and friendship throughout the years of my study and all stages of thesis preparation. I am also very grateful to Dr. Chafia T. de Trad for her support and assistance during the experimental part of my research project. I wish to thank Dr. George Simon from the Department of Materials Engineering for providing me with all necessary experimental equipment, software and his productive discussions and comments throughout the time of the experimental work. I would like to thank Mr. Brian Lithgow for reading and reviewing this thesis.

Beyond that I wish to express my gratitude to Monash University for providing an APA Scholarship that has made this study possible.

Thanks also go to all staff at the Department of Electrical and Computer Systems Engineering for providing an enjoyable working and research environment.

DECLARATION

It is hereby certified that except where the work of other authors has been specifically cited, this dissertation is entirely carried out by the candidate's own investigation.

It is also certified that this study has not already been accepted in substance for another degree, and is not being currently submitted in candidature for any other degree.

Candidate ... [redacted]

(Elena Pirogova)

INTRODUCTION

Proteins are crucial to virtually every biological process. As enzymes, they catalyze innumerable chemical reactions, which would otherwise occur slowly. They are active as carrier and storage molecules, in muscle contraction and in mechanical support [1]. As antibodies, they are responsible for immune protection and as receptors, in the nervous system, for the generation and transmission of nerve impulses [2].

Proteins are polymers built up from amino acids. Their great diversity and versatility is derived from the properties of the twenty different amino acid side chains that may be present. The protein has an unbranched structure and is formed by condensation reactions, which progressively add amino acid units to the polypeptide chain. Each monomer unit called amino acid is connected by a peptide bond and possesses the side chain of the amino acid from which it was formed [3].

As the polymerization process progresses, the polypeptide chain folds in a way, which is under the direction of the side-chain sequence, into the three-dimensional structure (3-D) that characterizes the active protein. A variety of intermolecular interactions influence the final protein conformation: electrostatic and torsional interactions, dispersion forces and solvent reaction [4]. If the biological function is considered as a selective interaction of the protein and its target, then a more fundamental problem of understanding the physical basis of this interaction and sequentially a selectivity of such process appears [5]. Once this understanding has been gained, it should be possible to design peptides and even proteins *de novo,* with the chosen biological function, and thus to produce new and more efficient drugs and other biotechnological products [5,6].

There have been many attempts to discover the rules governing the coding of the biological function into the sequence of amino acids within the protein. Typical approaches deal with either homology characterization of specific features of the primary and secondary structure of proteins or molecular modeling of protein tertiary structure [1,7]. Those approaches permit a significant insight into protein structure and active site location. Unfortunately they still do not provide sufficient knowledge about informational, structural and physicochemical parameters crucial to the selectivity of protein interactions, which can be used for the *de novo* design of peptide or protein analogous with desired biological

activity. The existing knowledge in the field of computer-aided molecular modeling and protein structure/function analysis can be classified in terms of the primary, secondary and tertiary structure analysis of proteins [8,9].

This report presents the author's investigation of the effect of various amino acid properties on the determination of characteristic patterns of different functional protein families within the RRM [5,6]. Also it includes the research of possible usage of experimentally measured amino acid parameters instead of the Electron-Ion Interaction Potential (EIIP) [54] in the RRM. Solving this problem will assist in the understanding of the functions of thousands of protein sequences discovered each year and generate great impact in the field of molecular biology and bioengineering.

CHAPTER 1

PROTEIN PROPERTIES

1.1. AMINO ACIDS

The particular function of a given protein is determined by the sequence of amino acids in the protein molecule. Amino acids are organic acids that have an amino group on the alpha carbon - the carbon next to the carboxyl group. The general structure of an amino acid is as follow [8]:

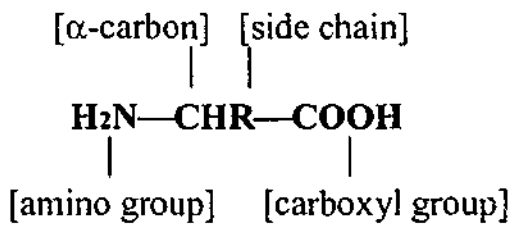


Figure 1.1. General structure of an amino acid.

The different R-group side chains on the amino acids distinguish one amino acid from another. Amino acids are so called because they contain in their molecular structure at least one primary amino group - NH (or cyclic imino group > NH) - and one carboxylic group. So they have both *basic* and *acidic* characteristics [1].

Amino acids have many uses in the body. They are used by cells to synthesize tissue protein for use in the formation of new cells or the repair of old cells, to synthesize enzymes, to form non-protein nitrogen-contained compounds such as nucleic acids or heme groups, or to form new amino acids. Amino acids that are not used in synthesis are converted in the liver to ammonia, carbon dioxide, water and energy.

Occurrence

Over eighty amino acids are now known to occur naturally. More may yet be discovered but their occurrence is rare. Twenty-six are constituents of proteins; twenty occurring frequently in protein structure.

Properties

All amino acids are white crystalline solids that melt indifferently at high temperatures (>200°C) and with some decomposition. They are soluble in water but they vary

considerably in their solubility. Furthermore the solubility of any acid increases as the solution is made basic or acidic. Most are insoluble in organic solvents. These properties are a direct consequence of the dual basic-acidic character of amino acids which gives rise to ionic behavior [7,8].

$$^{+}H_3N\underset{\underset{\displaystyle R}{|}}{C}HCH(NH\underset{\underset{\displaystyle R}{|}}{C}HCO)nNH\underset{\underset{\displaystyle R}{|}}{C}HCOO^{-}$$

Proteins may be generally classified into two categories: the *fibrous* proteins, which are insoluble in water and the water-soluble *globular* proteins. Molecules of fibrous proteins tend to be long and thread-like, and may be set side by side to form fibres and are often held together at many points by hydrogen bonds. Fibrous proteins often serve as structural materials. Globular proteins are typically folded to form spheroids with the hydrophobic portions of the proteins turned into the centre of the spheroid and the hydrophilic portions of the proteins turned outward to the surface. Due to their shape the contact area between molecules tends to be small and hence intermolecular forces also tend to be relatively weak and the proteins are relatively soluble. This solubility and resultant mobility are often exploited by biological systems to perform functions relating to the regulation and maintenance of biological processes [1,10].

It should also be noted that proteins along with nucleic acids exhibit the phenomenon of denaturation. Exposing a protein to heat or a strong reactive agent will cause it to lose its ability to perform physiological functions [8,11]. For example, albumin in eggs turns white when heated. In contrast, large polypeptides do not exhibit this phenomenon because, unlike proteins, they do not possess a stable secondary structure.

1.2 PROTEIN STRUCTURE

Proteins are complex molecules that can be classified according to specific primary, secondary, tertiary and quaternary structural characteristics.

Primary Structure: The Amino Acid Sequence and the Peptide Bond

The primary structure of a protein is given by the sequence of amino acids in the protein molecule. These amino acids are coupled together by covalent bonds called *peptide bonds.*

The peptide bond is an amide linkage, formed by joining the carboxyl group of one amino acid to the amino group of a second amino acid through elimination of water (a condensation reaction) (Fig.1.2) [10].

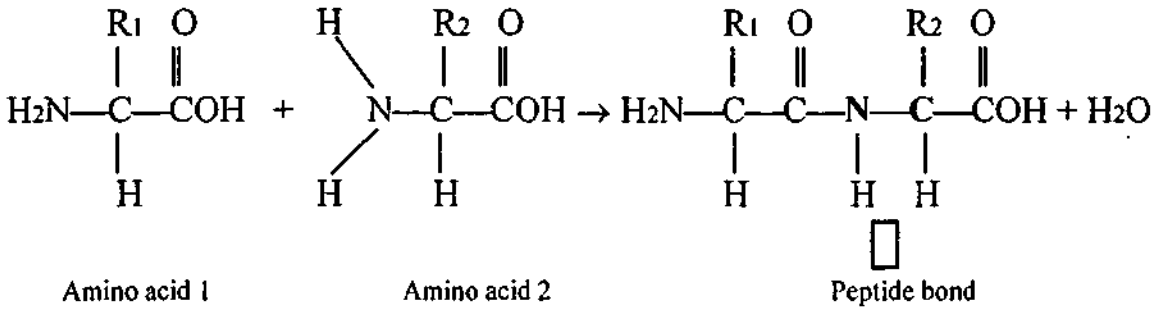


Figure 1.2 The peptide bond.

The peptide bond is stable in the face of changes in pH, in solvents, or in salt concentration and can be broken only by acid or base hydrolysis, or by specific enzymes. Two amino acids held together by a peptide bond are called dipeptide; three amino acids form a tripeptide and more than three form a polypeptide. There are some standard conventions for the naming of peptides and proteins. Since proteins contain very large numbers of amino acids, three-letter abbreviations of the amino acid names are used when writing the amino acid sequence. A dash or dot between the amino acid names represents the peptide bond. One end of the protein will consist of an amino acid having a free amino group so called the N-terminal end and it is listed first in the sequence of amino acids. At the other end of the protein there will be an amino acid having a free carboxyl group. This is the C-terminal amino acid, and it is the last amino acid listed in the sequence. The order of amino acids in a protein will determine its function, and is critical to its biological activity. A change of just one amino acid in the sequence can disrupt the entire protein molecule [12].

Secondary structure: Noncovalent Bonding

The secondary structure of a protein is a specific geometric arrangement of the amino acids. These configurations, resulting from hydrogen bonding, were established by Linus Pauling [1] using X-ray diffraction (Fig. 1.3a).

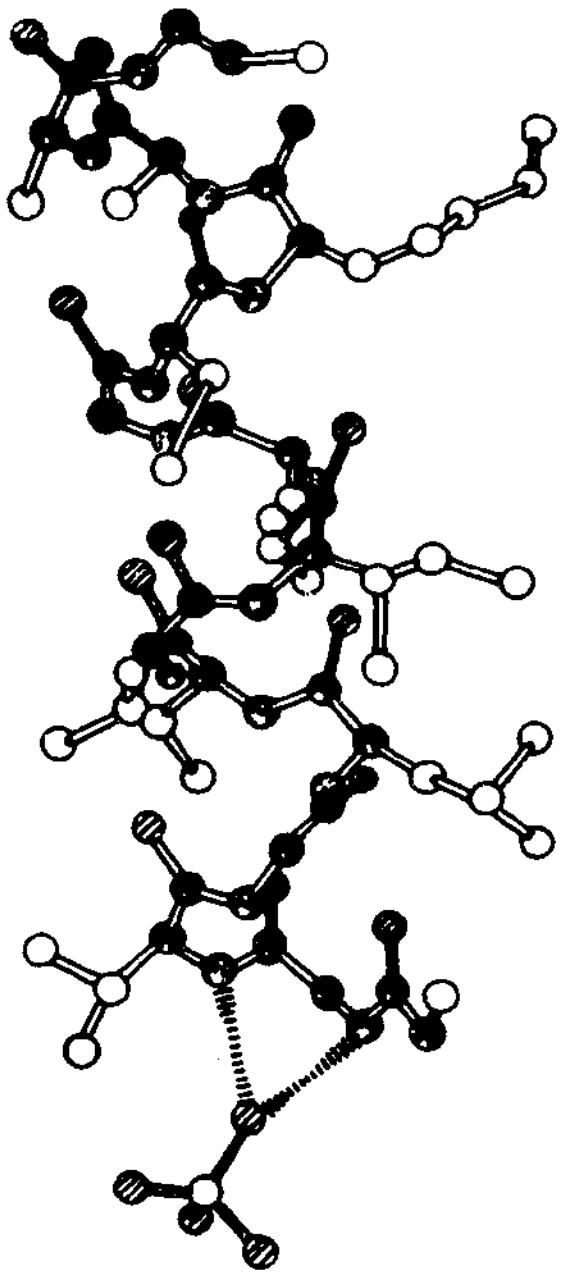


Figure 1.3a Alpha-helix.

α - Helix

Pauling determined that the polypeptide chains in the protein keratin were curled in an arrangement called α - helix. In this configuration the amino acids form loops in which the hydrogen on the nitrogen atom in the peptide bond is hydrogen bonded to the oxygen attached to the carbon atom of a peptide bond farther down the chain. There are 3.6 amino acids in each turn of the α-helix, and the R-groups on these amino acods extend outward from the helix. Hair and wool are made up of several coils of keratin wound around each other and held together by disulfide bridges.

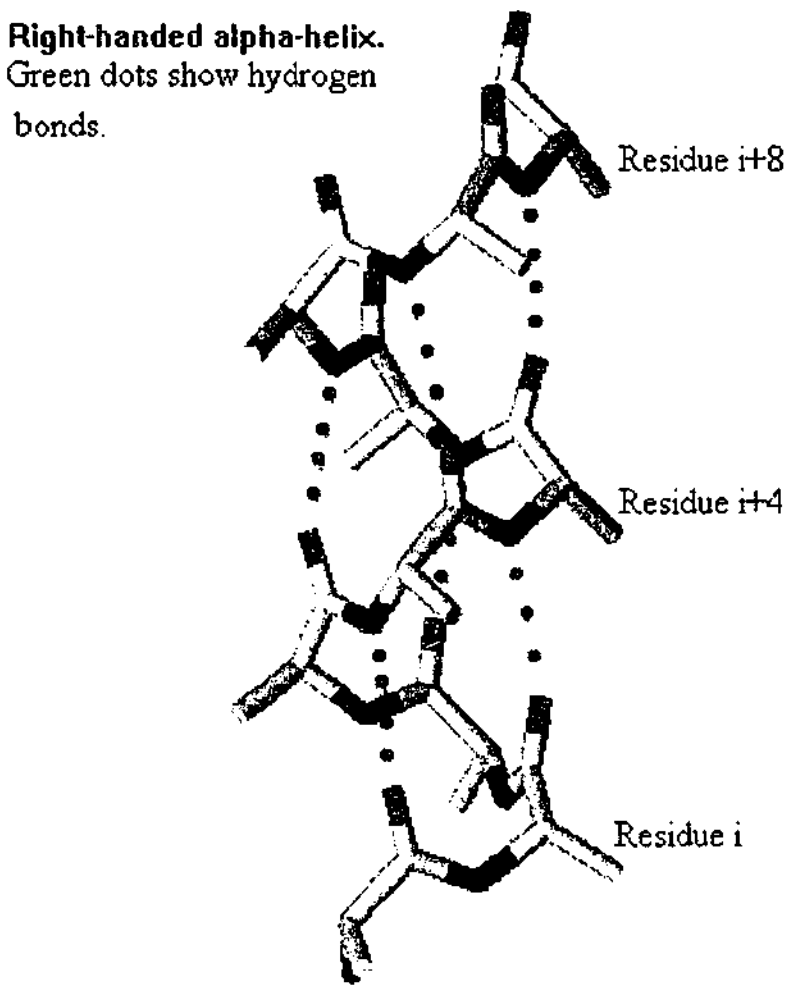


Figure 1.3b A right-handed α-helix. The α-helix has 3.6 residues per turn. The green dots show the hydrogen bond. [Courtesy, Dr. Berndt, Karolinska Institute, Sweden].

In this configuration, the amino acids form loops in which the hydrogen on the nitrogen atom in the peptide bond is hydrogen bonded to the oxygen attached to the carbon atom of a peptide bond farther down the chain (Fig.1.3a, 1.3b).

β - Configuration, or Pleated Sheet

The fibrous protein of silk has a different secondary structure. In silk, several polypeptide chains going in different directions are located next to each other; this gives the protein a zigzag appearance, hence the name pleated sheet (β-sheet) (Fig.1.4a, 1.4b). The chains are held together by hydrogen bonds, and the R-groups extend above and below the sheet. Each protein will have a specific secondary structure depending upon the amino acid

sequence. It should be noted that the hydrogen bonding is weak noncovalent bonding, and is easily disrupted by changes in pH, temperature, solvents or salt concentrations [9,10].

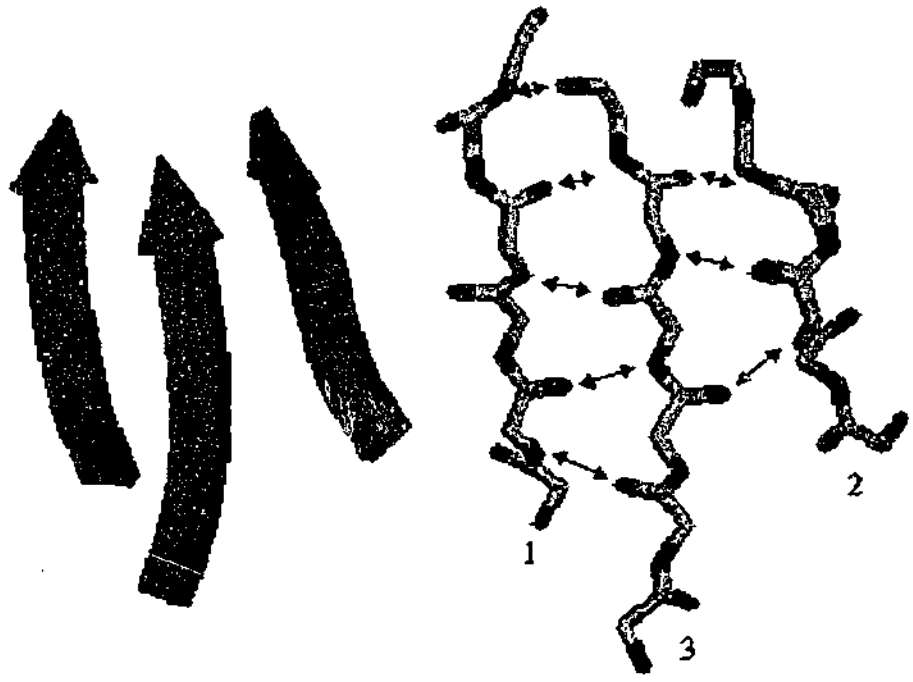


Figure 1.4a The three-stranded parallel beta sheet in thioredoxin (2TRX.PDB). The tree parallel strands are shown in both cartoon format (left) and in stick form containing backbone atoms N, Cα, C and O (right). Arrows identify hydrogen bonds. [Courtesy, Dr. Berndt, Karolinska Institute, Sweden].

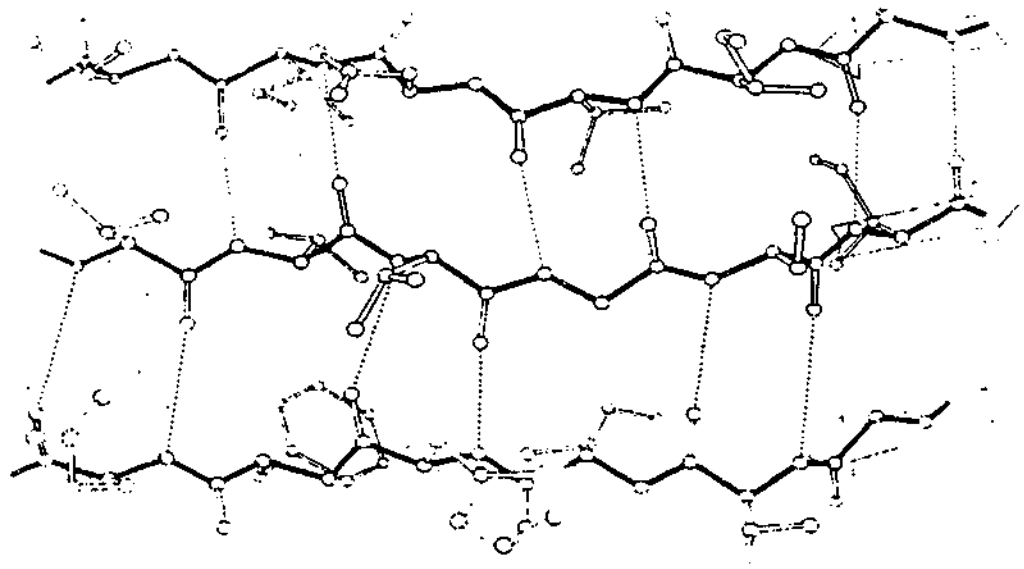


Figure 1.4b β-Sheet. . [Courtesy, Dr. Berndt, Karolinska Institute, Sweden].

Tertiary Structure: Globular Proteins

The tertiary structure refers to the total 3-D structure of the polypeptide units of the protein. It includes the conformation relationships in space of the side chain groups to the

polypeptide chain and the geometric relationship of distant regions of the polypeptide chain to each other. The native tertiary conformation is that of the lowest Gibbs free energy kinetically accessible to the polypeptide chains for the particular condition of ionic strength, pH and temperature of the solvent in which the folding process originates [4].

Tertiary structure is the three-dimensional structure (3-D) of globular proteins (Fig.1.5). In general, globular proteins are very tightly folded into a compact spherical form. This folding results from interactions between the R-group side chains of amino acids and may involve hydrogen bonding and the following interactions [3,4,10].

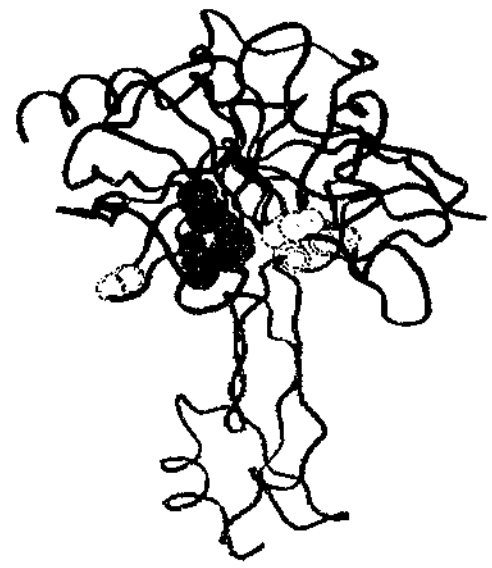


Figure 1.5 3-D structure of Trypsin Inhibitors [6].

These 3-D surface features are fundamental to protein function as they allow protein molecules to interact with other molecules in specific ways. For example, specific surface features are responsible for the binding of enzymes [13]. Proteins also form larger structures such as tubules or filaments. Therefore, knowledge of 3-D structure of protein, either by prediction or experimental measurements, will contribute to the understanding of protein functions that are so crucial in living systems.

Disulfide Bridges

Disulfide bridges are interactions that include a disulfide linkage between molecules of the amino acid cysteine. Sulfhydryl groups (—S—H) are easily oxidised to form a disulfide (—S—S—). These bridges are covalent linkages that can be ruptured by reduction, but that are stable in the face of changes in pH, solvents or salt concentrations [8,11].

Salt bridges

Salt bridges result from ionic interactions between charged carboxyl or amino side chains found on amino acids such as aspartic acid, glutamic acid, lysine and arginine. These linkages are particularly vulnerable to changes in pH [3,4].

Hydrophobic Interactions

Hydrophobic interactions occur between the nonpolar side chains of the amino acids in the protein molecule. Since the solvent water repulses them, these groups tent to be found on the inside of the molecule along with other hydrophobic groups. At normal pH and temperature, each protein will assume a shape that is energetically the most stable given specific sequence of amino acids and the various types of interactions we have mentioned. This shape is called the native state or native configuration of the protein [3,4,10,14,15].

Quaternary Structure: Two or More Polypeptide Chains

The quaternary structure refers to the structure and interactions of the non-covalent association of discrete polypeptide subunits into a multisubunit protein. Not all proteins have a quaternary structure. The quaternary structure of a protein is the manner in which separate polypeptide chains fit together in those proteins containing more than one chain. Hydrogen bonding, hydrophobic interactions and salt bridges may be involved on holding the chains in position [3].

1.3 DENATURATION AND FACTORS AFFECTING PROTEINS' STRUCTURE

Various changes in the environment of a protein can disrupt the complex secondary, tertiary or quaternary structure of the molecule. Disruption of the native state of the protein is called **denaturation,** and will cause the protein to lose its biological activity. Denaturation may or may not be permanent; in some cases the protein will return to its native state when the denaturing agent is removed. Denaturation may result in coagulation, with the protein being precipitated from solution. The process of denaturation involves the uncoiling of the protein molecule into a random state. Denaturing agents come in many forms, a few of which are as follows:

Heat

Heat causes an increase in thermal agitation of the molecule, disrupting hydrogen bonding and salt bridges. After gentle heating the protein can usually regain its native state, but violent heating will result in irreversible denaturation and coagulation of the protein. Similarly, heat used to sterilize equipment coagulates the protein of microorganisms. Heat can also be used to detect the presence of protein in urine; heated urine that turns cloudy indicates the presence of protein.

pH

Changes in pH have their greatest disruptive effect on hydrogen bonding and salt bridges. The polypeptide polylysine is composed entirely of the amino acid lysine, which has an amino group on its side chain. In acidic pH the side chains will all be positively charged and will repel each other, causing the molecule to uncoil. In basic pH, however, the side chains will be neutral. Therefore, they do not repel, and the molecule will coil into a α-helix.

Organic Solvents

Organic solvents such as alcohol or acetone are capable of forming hydrogen bonds themselves, which then complete with the hydrogen bonds naturally occurring in the protein. This causes denaturation and coagulation. A 70% alcohol solution is a good disinfectant since it will coagulate the protein in bacteria, thereby destroying these organisms.

Heavy Metal Ions

Heavy metal ions such as Pb^{2+}, Hg^{2+}, and Ag^{+} may disrupt the natural salt bridges of the protein by forming salt bridges of their own with the protein molecule. This usually causes coagulation of the protein. The heavy metal ions also bind with sulfhydryl groups, disrupting the disulfide linkages in the protein molecule and denaturing the protein. Such heavy metals are toxic to living organisms. As an antipode to heavy metal poisoning, substances high in protein, such as milk or egg whites, are given. These substances will bond with the metals while they are still in the stomach. The victim must then be made vomit before the metals are again released by the process of digestion.

Alkaloidal Reagents

Reagents such as tannic or picric acid affect salt bridges and hydrogen bonding, causing proteins to precipitate. Tannic acid is used to precipitate proteins in animal hides; this is the process of tanning used in the manufacture of leather.

Reducing Agents

Reducing agents disrupt disulfide bridges formed between cysteine molecules. For example, hair permanents work by reducing and disrupting the disulfide bridges in hair; when the hair is curled and oxidizing agents are applied, new disulfide bridges are formed.

Radiation

Radiation is similar to heat in its effect on proteins. For example, ultraviolet light will burn the skin, causing sunburn (structural protein: skin pigment melanin). It causes an increase in thermal agitation of the molecule, disrupting hydrogen bonding and salt bridges. If the influence is not too strong the structural protein can usually regain its native state.

1.4 INTERACTION FORCES

The forces between the atoms in a molecule and between molecules themselves determine the molecule's structure. A variety of forces are involved in the interactions of protein molecules including electron exchange forces, resonance forces, dipole forces, polarisation forces and van der Waals' forces [16]. We can classify those interactions into strong and weak interactions with a criterion to see whether or not the thermal motion will disrupt the interaction. The average thermal energy of one electron is in the order of kT, where k is Boltzmann's constant and T is the absolute temperature [16]. The strong interactions have a value much greater than kT, so they are unlikely to be disrupted by thermal motion. The primary structure of a protein is determined by such strong interactions through forming the covalent bonds. On the other hand, the weak interactions are of the same order as kT and contribute most to the forming of the higher-order structures. The weak interaction forms dipole-dipole interactions and hydrogen bonds. These bonds will be disrupted first when the protein molecule is heated, resulting in the loss of quaternary, tertiary and secondary structure in that order.

All these forces mentioned above are basically electrostatic forces, thus, they could be described by Coulomb's law. Although the electromagnetic forces between moving charges are too weak to account for the formation of protein structures, they may be crucially involved in protein's biorecognition, chemical binding and energy transfer [5]. Moreover, the fact that the most biological processes in any living organism are based on the selective interactions between particular biomolecules, makes this force unable to be ignored at all. The newly developed Resonant Recognition Model is mainly focused on investigating the behavior of the electromagnetic forces and fields within the protein structures [5,6].

CHAPTER 2

REVIEW OF PREVIOUS RESEARCH

2.1 PROTEIN'S STRUCTURE-FUNCTION ANALYSIS

A protein's biological function and its 3-D structure are in some way determined by the corresponding amino acid sequence [17]. Solving of the protein's folding problem and determination of the biological function from the amino acid sequence are two facets of one fundamental problem in today's molecular biology. This fundamental problem is to decipher the natural information encrypted within the amino acid sequence. Solving this problem will assist in the understanding of the functions of thousands of protein sequences discovered each year. Furthermore, it will facilitate the *de novo* designed of new protein sequences for the pharmaceutical industry [12].

There have been many attempts to discover the rules governing the coding of the biological function into sequence of amino acids within the protein. Typical approaches deal with either homology characterization of specific features of the primary and secondary structure of proteins or molecular modeling of protein tertiary structure. Although such approaches permit a significant insight into protein structure and active site location, they still do not provide sufficient knowledge about the informational, structural and physicochemical parameters crucial to the selectivity of protein interactions that can be used for *de novo* designed protein's analogies with the desired biological behavior [3,4,8].

Typical approaches fall into two distinct groups:

1) Model based analysis of amino acid sequences to simulate the folding process and search for information related to the biological function.

2) Empirical and statistical approaches inferred from known structures and functions.

2.1.1 Model based approaches

Energy minimization

The approaches based on energy minimization are mainly centered on attempting to solve the protein's folding problem. It is a thermodynamic assumption that the native structure of a globular protein corresponds to the minimum energy of that biological system [12]. This energy is postulated as Gibbs free energy [18]. Theoretically, the native structure of any molecule can be determined by searching the entire conformation space of all the degrees of freedom for the conformation with the lowest free energy. Once the 3-D structure of the protein is determined then it will be easy to determine the biological function either through experimental means or searching a database.

In practice, this is not possible due to the following reasons:

- The energy functions that drive the folding and stabilize the native structure are not well understood. Currently, the energy of a conformation is deduced from the combination of various factors such as hydrophobic interactions and hydrogen bonding [12]. Hence, in order to model the physical folding accurately, detailed full-atomic model of the protein and its surrounding solvent are required.

- The conformation space for a full-atomic model is too large for an exhaustive search to be conducted. A simplified model is needed so that the conformation space is small enough for the minimum free energy conformation to be located in a reasonable amount of time [19]. Examples of simplified models are the random walk model, reduced representation model, cylinder and sphere model, and lattice model [19-22].

The most direct approach in the protein structure prediction is a simulation of the folding process using a molecular dynamics simulation (MD) [23]. This procedure usually requires a full-atomic model of the analyzing protein molecule. A few shells of solvent molecules may be placed around the protein for counting of the solvating effects. The remaining solvent environment that cannot be explicitly included is mimicked either by adjusting the boundary conditions or replicating the system in three dimensions so that no water-vacuum interfaces remain. The classical equations of motion for the system are integrated numerically by solving Newton's equation of motion:

$$m_i (d^2r_i/dt^2) = -\nabla_i[U(r_1, r_2,...r_n)] \quad i=1,..N \tag{2.1}$$

where m_i is the mass of particle *i*, r_i is the position of particle *i*, U is the potential energy surface as a function of positions of all particles, N is the number of particles in the system.

The solutions to these equations determine the atomic positions and velocities as a function of time and allow the structure, folding pathways, diffusion and thermodynamics properties to be calculated. The potential energy surface U for proteins includes energy terms to represent chemical bonds, angles, torsions, rotations, van der Waals interactions, Coulombic interactions, etc. This is a complicated function of many parameters that must be differentiated with respect to the Cartesian positions of each atom for the numerical solution of Newton's equations of motion. Step size δt for the integration is usually chosen to be around 10^{-15} seconds (1fs). It is currently feasible to simulate for about 10^6-10^7 integration steps. This means that a simulation time is on the order of nanoseconds. However, the time scales of many processes in protein folding are determined by the rate of overcoming free-energy barriers, which make take milliseconds. Even for small peptides, the maximum simulation time achievable is a few nanoseconds, due to thee computational effort needed to evaluate every element in the full-atomic model. Thus for larger proteins it is impossible to obtain a statistically valid sample of the relevant conformations using a standard MD simulation [23].

Another powerful optimizing method is the Monte Carlo (MC) approach in which both the energy and temperature of the Boltzmann distribution guides the search for the global minimum. Instead of simulating the entire molecule in one step, a conformation change is introduced at one bond and the new energy is evaluated. If this new energy is lower, the new conformation is taken to be the new conformation for the molecule. If the new energy is higher, the new conformation is accepted with a Boltzmann probability that depends on the simulation temperature and the conformation energy [24]. Although the MC method can eventually locate the global minimum, the number of steps required to guarantee that is larger than the exponential of the number of steps to enumerate the whole search space [25].

To increase the simulation efficiency, simulated thermal annealing may be used with MC or MD simulation [26]. The protein molecule is initially simulated at a high temperature (eg.1000K). The temperature is gradually lowered until eventually a conformation freezes out. Finally, energy minimization is performed to locate the energy minimum in the vicinity where the protein has been frozen out.

Hydrophobicity profiles

The hydrophobicity of a residue in a protein is a measure of its relative comfort in a non-aqueous medium such as a non-polar solvent or a protein interior, compared with an aqueous medium [17]. Residues with hydrophobic side-chains tend to bury themselves within proteins, away from the aqueous surroundings. A protein can be transferred into a numerical series by replacing each amino acid using its hydrophobicity [17]. The graph of this numerical series is often referred as hydrophobicity profiles. The feature of the distribution of the hydrophobicity along the protein sequence will be revealed after applying a moving window. Rose [27] used a hydrophobicity profiles to predict turns while Kyte and Doolittle [28] predicted protein molecular segmentation into interior and exterior regions for both soluble and membrane-bound proteins. Eisenberg [29] applied hydrophobic moment distinguish transmemrbrane helices from amphiphilic helices.

The underlying assumptions of the hydrophobicity profiles are based on a simple two-solvent model, i.e., the protein and its aqueous surrounding can be viewed as two different solvent phases. The hydrophobicity profiles depend on the choice of hydrophobicity scale and the length of the moving window (the number of consecutive amino acid to be smoothed together). Although, the hydrophobicity profiles give some hints about a protein's structure, no published methods based on the hydrophobicity analysis is accurate enough for large protein samples. Overlap regions are always found [30].

Stochastic modeling and optimal filtering

White, Stultz and Smith [31,32] have proposed a stochastic signal model for recognized structural classification of single-domain proteins. This model treats protein as a "time series" of symbols containing signals that determined the protein's structural class. At first, a set of symbols is mathematically defined by using stochastic, nonlinear, state-space

model [32]. Then, an optimal filtering algorithm is adopted to compute the model likelihood, i.e., and the probability that a given primary sequence would have been generated by each of the candidate hypotheses. Finally, the protein folding structure is classified accordingly to that particular hypothesis which has sufficiently high probability. The Protein Sequence Analysis (PSA) algorithm is based on probabilistic Discrete State-space Models (DSMs) and optimal filtering and smoothing algorithms as described in the paper "Structural analysis based on state-space modeling" by C.M. Stultz, J.V. White, and T.F. Smith, Protein Science (1993), 2:305-314. The mathematical basis for the models and algorithms is presented in [31]. In the PSA, a single amino acid sequence is analyzed in one of three ways: using Type-1, Type-2, or WD-repeat DSMs. Type-1 models are for complete sequences from monomeric, single-domain, globular, water-soluble proteins in several recognized structural classes. Now, however, there is an exception, because a set of Type-1 DSMs have been included for transmembrane proteins with a beta-barrel fold like porin. In any case, the Type-1 analysis may be applied to sequences having lengths in the range from 40 to 350 residues. Type-1 DSMs are also appropriate for those subsequences of membrane-spanning proteins that are believed to extend beyond the membrane (based on a hydropathy profile). Type-2 models, in contrast, are for either partial or complete sequences from potentially large proteins that violate one or more of the modeling assumptions embodied in Type-1 models. For example, Type-2 models are appropriate for proteins that have one or more of the following properties: (1) they are multimeric; (2) they have more than one structural domain; or (3) they are not globular or soluble (e.g., membrane-spanning proteins). Type-2 models can be applied to sequences up to 1000 residues long. WD-repeat models are specially designed for the WD-repeat family of proteins. These models combine a particular Type-1 structural model with sequence-specific pattern information. They can be applied to sequences up to 1000 residues long.

Limitations of the PSA server and its libraries of DSMs are as follows:

1.When using Type-1 models to determine the probable tertiary structural class of the protein, the probability calculated for each structural class indicates the relative probability of that structural class compared with all of the other structural classes in their library. The probabilities are weights of support for the available DSMs as explanations of the analyzed sequence. The PSA server does not predict folds of proteins, for which they have no DSM. Moreover, the probabilities reported by this system are not the probabilities that the analysis is "correct" or "true." Rather, the probabilities indicate which of their DSMs are the best explanations of the analyzed sequence.

2. The current library of Type-1 models is primarily intended for monomeric, single-domain, soluble, globular proteins having sequences in the length range from 35 through 350 residues. The Type-2 models and WD-repeat models are based on fewer analysis assumptions and deal with sequences of any length. However, for practical reasons, sequences analyzed by the PSA System using Type-2 models and WD-repeat models are limited to 1000 residues. To analyze longer sequences, it is requested to split them into 1000-residue subsequences, with some overlap, and analyze each of the subsequences separately. This approach is based on a mathematical model. However, the biological and physical meaning of this model is a blur. Moreover, the treatment of the individual amino acid in a protein sequence as a symbol discards all physicochemical properties of amino acids that will play an important role in the formation of protein's natural conformations and biological functions.

2.1.2 Empirical and Statistical Methods

The existing knowledge in the field of computer-aided molecular modeling and protein structure/function analysis can be classified in terms of primary, secondary and tertiary structure analysis of proteins' [2].

Primary structure analysis

Primary structure analysis is concerned mainly with the search for homologies among amino acid sequences. The main concept behind this method is that proteins with the same biological function share amino acid sequence alignments and these homologous fragments carry the main information about the protein function [13,22]. However, similar sequences may appear occasionally in totally unrelated proteins through convergent evolution of sequences with similar properties [1,2,33].

On one hand, there are a number of cases of unexpected but significant resemblances between functionally dissimilar proteins, while on the other hand, there are cases of insignificant resemblance between functionally related peptides [1]. Thus, the available methods rarely reach a satisfactory level of predictive accuracy for the *de novo* prediction of the biological function from sequence similarities or for the rational peptide design.

Optimal alignment programs [34] are designed to distinguish between analogous and homologous sequences, but they are still based on sequences similarities. Problems remaining with optimal alignment programs include difficulties associated with the length of the sequence, string and insertion of gaps in order to increase the number of matching residues. Inserting the gaps too liberally and assigning the gap-weights arbitrarily can lead to biologically irrelevant alignments. Combination of primary, secondary and tertiary structure homologies to overcome some of these problems has recently been attempted [35].

Thus, the homology comparison and the alignment of protein sequences are the basis for the protein structure prediction and molecule design. A similarity search for comparison and alignment has two phases: first, the finding of such similarities, and second, the assessing of their significance [35].

Two-sequence comparison uses a match matrix to contain the results of all possible character comparisons between two sequences and visualize the regions of significant similarity [2]. The techniques developed for two-sequence comparison include simple dot-plot, threshold filtering, reduced alphabet and similarity scoring [2]. All these approaches are based on the expression of similarity by summing a score for each pairwise comparison of residues. The simple dot-plot is the simplest scheme that gives a score of unity to each match, and zero to every mismatch or deletion [36]. As an alternative to similarity scoring, one can score the distance between two subsequences in a particular alignment. The simplest distance-scoring scheme scores zero for each match, and unity for each mismatch or single residue deletion. The summed distance is the minimum number of individual changes required to transform one subsequence into the other, and can therefore be regarded as a measure of the evolutionary distance between them. Sometimes the background is so high that significant similarities may be cancelled. In order to increase the signal-to-noise ratio, threshold filtering is designed to systematically remove the dots derived from accidental pairing (noise) from the genuine similarity (signal). A flexible and efficient method of filtering is to put a dot only where the local concentration of matches exceeds a set threshold [2]. Alphabet reduction is to group some residues together to reduce the number of different symbols [2]. For example, in weakly homologous sub-sequences, lysine and arginine, or leucine and isoleucine are often regarded as functionally equivalent. Thus, before doing comparison, all lysine and leucine can be written as

arginine and isoleucine. This alphabet reduction can enhance signals. However, the number of random matches is also increased. Similarity scoring is used to recognize intermediate grades of similarity between match and mismatch when two residues are compared. Since Dayhoff and Eck proposed the first PAM (Percentage of Acceptable point Mutations) scoring matrix in 1968 [37], there have been a few types of similarity scoring matrixes. The first type scoring matrixes is based on the nucleic acids mutation of the corresponding amino acids. The second type is based on the physicochemical properties of amino acids (Rao matrix [38]). The third type is related to the substitutive relationship of amino acids within the group of proteins (PAM matrix and MD matrix [39]). Furthermore, protein's structural information is also used to construct the similarity-scoring matrix. This type of scoring matrixes is called structure-scoring matrix (RIS matrix and SCM matrix [40]). The use of similarity scoring scheme may produce a significant increase in the delectability of signals. However, the particular scoring scheme should be tailored to the similarity being sought. Another method is the multiple sequence comparison. The method is still based on the two-sequence comparison. Before doing multiple sequences comparison, the comparing order needs to be sorted. Generally, the higher sequence homology is the greater the reliability of the sequence comparison is. So, the comparing order is sorted by the homologous closeness of the sequences before doing multiple sequences comparison. There are two generally used multiple sequence comparison algorithms: the Feng and Doolitle method [41] and the Barton and Sternberg method [42]. Multiple sequence comparison is quite a complicated and time-consuming method as the combinatorial comparisons have to be taken into the account for a group of proteins.

Amino acid sequence comparison and alignment have achieved some success to extract the information pertinent to protein's structure and function. However, many different amino acid sequences give similar 3D-structures. Simple calculations have shown that there are a total of 10^{33} completely different side chain arrangements that can give similar polypeptide folds for an average-sized domain of 150 amino acid residues [43]. It is a huge number that cannot be solved by the current fastest computer. Another problem is that similar sequences may appear occasionally in totally functional unrelated proteins through convergent evolution of sequences with similar properties. In the mean time there are also cases of insignificant sequence resemblance between functionally related peptides. Actually, the comparing sequences should be identical lengths; otherwise gap deletion or

insertion will need to be adopted. Such "man made" matches can lead to biologically irrelevant alignments. Thus, only limited success has been achieved so far by this method and no general solution seems to be in sight [1,43].

Secondary structure analysis

The relationship between protein amino acid sequence and secondary structure is complex, reflecting the intricate thermodynamic and kinetic process of protein folding. The secondary structure analysis methods can be categorized into three classes: the empirical statistical methods, stereo-chemical methods and non-linear statistical methods [14]. In the absence of experimental knowledge about the protein secondary structure, various empirical, computer-aided algorithms are available for its prediction from the known primary structure [11], with results often presented in the form of preferred regions of regular secondary folding for a given peptide chain [10,44]. Most of these algorithms are based on the average probability that any particular amino acid residue will be found in a α -helical, β -sheet or "random coil" conformation [10,11,44]. While these approaches can be relatively successful in predicting secondary structure (with probabilities in the region of 60-70% with selected examples) [10], direct relationships to the protein function cannot generally be described with these methods.

Despite the fact that the degree of prediction is only up to 60% [45,46], existing methods for predicting protein secondary structures [47] are becoming more widely used as more amino acid sequences are determined from gene cloning and DNA sequencing techniques. Thus, there are growing needs for computational methods to obtain insight into structural and functional features associated with sequence information.

A multivariable analysis method for discrimination of the protein secondary structure segments [48,49] is based on a database analysis. There an amino acid sequence is represented by a sequence of numerical values corresponding to certain aspects of amino acid residues such as hydrophobicity and bulkiness. Given a collection of the amino acid indices [49] and given a library of calculation methods to extract specific patterns in the numerical profile representing the amino acid sequence, the method will automatically find the best sets of parameters that discriminate different groups of sequence data. In the present case, different secondary structure segments taken from the Protein Data Bank.

Once proper database are organized, the method can also be applied to other structural and functional features such as the prediction of various binding sites and modification sites in the amino acid sequences.

Tertiary structure analysis

The folding of the linear primary protein chain into the defined 3-D structure results in a spatial relationship between the various constituent amino acids which is found to be crucial for determination of the functional behavior of the protein. This widely accepted model of protein interactions proposes that selectivity of these interactions is based on the structural matching between active sites of interacting molecules. Experimentally, the tertiary structure and stability of proteins have been studied using such techniques as X-ray crystallography, circular dichroism, fluorescence spectroscopy and 2-D NMR [50]. The most well known tertiary structure information has been obtained through the use of X-ray crystallography [43]. The interaction of X-ray with electrons in protein molecules arranged in a crystal is used to obtain an electron-density map of the molecule, which can be then interpreted in terms of an atomic model. However, the X-ray diffraction method requires an enormous amount of work and is time consuming. Furthermore, a protein might not always be available in a crystalline form and the crystallization may cause protein structural distortion [43]. These limitations make the determination of protein's 3-D structures several orders that the determination of their primary structures, the amino acid sequences [15].

However, such methods as circular dichroism, fluorescence spectroscopy and 2-D NMR have been limited due to the need for relatively large amounts of protein and the inability of many techniques to detect low abundances of the conformation intermediates. These methods may also be limited by structural distortions caused by particular techniques, e.g. crystalization. The increasing data base of experimentally derived protein primary structures combined with the computer algorithms for performing molecular mechanics and dynamics has the potential to establish computational algorithms as a powerful tool to study protein tertiary structure and predict peptide/protein active conformations [9,51]. Recently, a new optimization method called Genetic Algorithm (GA) [52] has been recognized as a highly efficient searching algorithm and has begun to be applied in protein 3-D structure prediction [21]. The results obtained are still being anticipated. These

methods are still not able to predict protein 3-D structure solely from its sequence, and thus they usually use a number of constraints from experimental measurements, or they are based on the sequential and functional homology to the proteins with known 3-D structure.

Although all these techniques have significant success in prediction of protein structure and/or function, they are still not sufficient for understanding protein interactions, and consequently for the design of peptides or proteins with the desired bioactivity. The lack of knowledge about the informational parameters of protein sequence important to the protein biological function, as well as a lack of understanding of the physical processes, which are behind the biological activity of the proteins, has limited the success of these techniques.

2.2. ASSIGNING PHYSICOCHEMICAL PARAMETER

The information encoded in the amino acid sequence ultimately determines the 3-D structure and biological function of a protein under physiological conditions. In order to understand empirical relationships between the amino acid sequence, structural patterns and functional sites as well, statistical analyses can be made from collection of sequence data.

Each of the 20 amino acids has multifaceted properties that are responsible for the specificity and diversity of protein structure and function. A large body of experimental and theoretical research has been performed to characterize different kinds of properties of individual amino acids and to represent them in terms of the numerical index. In 1988 Nakai et al. collected 222 amino acid indices from research papers and investigated their interrelationships by the hierarchical cluster analysis [49]. The Cluster analysis of amino acid indexes method finds the best sets of parameters that discriminate different groups of sequence data [48,49]. One of the important factors in this analysis is how the amino acid sequence should be represented. The amino acid sequence may be considered as a sequence of numerical values reflecting various aspects of amino acid residues, such as: hydrophobicity and bulkiness. This analysis is intended as a practical guide for making use of suitable indices for structural and functional predictions. As reported in the paper [49] numerical indices representing various physicochemical and biochemical properties of the 20 amino acids have been collected from published literature. Because many of them

were highly correlated, hierarchical cluster analysis was performed using the correlation coefficient as a measure of similarity. They identified four major clusters: (i) helix and turn propensities, (ii) β-strand propensity, (iii) hydrophobicity and (iv) physicochemical properties. The database was then updated in 2000 [53] by adding of the newly published amino acid properties.

An Amino Acid Index Database [53] is a set of 20 numerical values representing any of the different physicochemical and biochemical properties of amino acids. As a follow-up to the previous study of Nakai, K., Kidera, A. & Kanehisa, M. [49], they have increased the size of database, which currently contains 402 published indices, and re-performed the single-linkage cluster analysis. The results basically confirmed the previous findings. The database may be accessed through the DBGET/LinkDB system at GenomeNet (http://www.genome.ad.jp/aaindex/).

With the rapid expansion of the protein databases, the identification of the biological function of newly sequenced proteins or the determination of their relationship with defined functional families, becomes a real problem. Therefore, the introduction of additional information concerning the relationship between amino acids within the protein sequence would be helpful. One of the most efficient approaches in this direction is Fourier analysis of parametric profiles representing the primary structure of the protein [54]. The main problem in this approach represents the proper choice of the amino acid parameter best suited to characterize the biological function of the analyzed proteins. Previous investigations in this direction [55] have shown that, among the 226 different analyzed parameters, **Electron-Ion interaction Potential (EIIP)** [54,55] has been chosen as the unique amino acid property and was recomended for the use for proteins' structure-function analysis within the RRM. However, the EIIP is only one of the many amino acid properties known. The question is as to whether any of the other known amino acid parameters from the newly updated database [53] can be used in the RRM analysis. When parameters from comprehensive database that displayed a significant correlation with EIIP values were used instead of the EIIP in the RRM the characteristic frequencies obtained were found to be greatly variable and to show no relation to the EIIP derived frequencies [56]. The results of the previous study [56] have shown the presence of periodicities across the range of parameters. The latter result is an indication of the difference in protein sequences. Different sequences will display varyingly demonstrable patterns of

periodicities [56]. This led to investigating the effect of non-significant parameters and the inclusion of the entire database of 402 published up to now different amino acid indicators in the RRM of several protein's families.

CHAPTER 3

THE RESONANT RECOGNITION MODEL

Proteins are "the work horses" of biological systems at the cellular level. They are responsible for various biological activities including enzymatic analysis, transport and storage, immune protection, intracellular communication and many other tasks. Proteins are linear macromolecules built up of sequentially linked amino acids. The information contained in the amino acid sequence determines the protein's chemical properties, chain conformation, protein's function and its specificity. Although both function and structure of a large number of proteins are becoming known throughout projects like Human Genome Project and are available through a number of data bases [4], the crucial problem of understanding how the biological function is "written" within the protein sequence still remains [5,6].

Once this understanding has been gained it should be possible to design peptides and even proteins *de novo* with the chosen biological function and thus to produce new and more efficient drugs and other biotechnological products [5,6].

With the rapid accumulation of databases of proteins' primary structures, there is an urgent need for theoretical approaches that are capable of analysing protein structure/function relationships. As described in the previous section, many attempts have been developed to predict the tertiary structure and biological function of protein from its sequence but only with limited success. These approaches have inherent limitations, i.e. they do not provide sufficient knowledge about the physical process for proteins folding and interacting. They also lack the informational, structural and physicochemical parameters crucial to the selectivity of protein interactions, which can be used for *de novo* design of peptide or protein analogous with the desired biological activity.

The new physicomathematical approach presented here is called the **Resonant Recognition Model** (RRM) [5,6,57-60]. The RRM, developed by I. Cosic [6], is based on the representation of the protein primary structure as a numerical series by assigning to

each amino acid a physical parameter value relevant to the protein's biological activity. A number of amino acid indices (402 have been published up to now [53]) have been found to correlate in some way with the biological activity of the whole protein. Previous investigations [55,61-63] have shown that the best correlation can be achieved with parameters, which are related to the energy of delocalised electrons of each amino acid. These findings can be explained by the fact that the electrons delocalised from the particular amino acid have the strongest impact on the electronic distribution of the whole protein. In previous studies [54,55], the energy of delocalised electrons (calculated as the electron-ion interaction pseudopotential, EIIP) of each amino acid residue was employed in the RRM approach. The resulting numerical series then represented the distribution of the free electrons' energies along the protein. This numerical series was then converted into a discrete Fourier spectrum, which carried the same information content about the arrangement of amino acids in the sequence as the original numerical sequence [36].

Approaches similar to the RRM, based on the Fourier transform and physical characteristics of amino acids, have been successfully applied by Mandell who has shown that the characteristic hydrophobic mass energy Fourier modes are signatures of isomorphism and immunological reactivates [64]. Viari *et al.* (1990) have used the RRM approach with scale independent coding to localize biologically relevant patterns in calcium-binding proteins [65].

Definition of Common Frequency Characteristics

The RRM is a physicomathematical model. It interprets the protein's sequence linear information using signal analysis methods. It comprises two stages. The first involves the transformation of the amino acid sequence into a numerical sequence. Each amino acid is represented by the value of the electron-ion interaction potential (EIIP) [54,55], which describes the average energy states of all valence electrons in particular amino acid. The EIIP values for each amino acid were calculated using the pseudopotentials general model [54,61]:

$$\langle \vec{k} + q \,|w|\, \vec{k} \rangle = 0.25\, Z \sin(\pi\, 1.04Z)/(2\pi) \qquad (3.1)$$

where q is a change of momentum $\vec{k}$ of the delocalised electron in the interaction with potential w, while

$$Z = (\sum Zi)/N \tag{3.2}$$

where Zi is the number of valence electrons of the i -th component of each amino acid and N is the total number of atoms in the amino acid. The EIIP values for 20 amino acids, as well as for five nucleotides (the whole procedure can be applied to the DNA and RNA), are shown in Table 3.1. A unique number can thus represent each amino acid or nucleotide, irrespective of its position in a sequence.

Numerical series obtained in this way are then analyzed by digital signal analysis methods in order to extract information pertinent to the biological function. The original numerical sequence is transformed to the frequency domain using the discrete Fourier transform (DFT). As the average distance between amino acid residues in a polypeptide chain is about 3.8 Å, it can be assumed that the points in the numerical sequence derived are equidistant [6]. For further numerical analysis, the distance between points in these numerical sequences is set at an arbitrary value d=1. Then, the maximum frequency in the spectrum is $F=1/2$ $d=0.5$. The total number of points in the sequence influences the resolution of the spectrum only. Thus, for N-point sequence the resolution in the spectrum is equal to $1/N$. The n-th point in the spectral function corresponds to the frequency $f = n/N$ [5,6].

In order to extract common spectral characteristics of sequences having the same or similar biological function, the following cross-spectral function was used:

$$Sn = XnYn^* \qquad n = 1,2,...,N/2 \tag{3.3}$$

where Xn are the DFT coefficients of the series $x(m)$ and Yn^* are complex conjugate DFT coefficients of the series $y(m)$. Peak frequencies in the amplitude cross-spectral function define common frequency components of the two sequences analyzed [5,6,66].

The whole procedure: **protein sequence → numerical series → amplitude spectra → cross-spectra** is shown in Fig. 3.1a. Fig 3.1b presents: (i) sequences of α and β human hemoglobins; (ii) graphical representation of the corresponding numerical sequences

obtained by replacing every amino acid with its EIIP value; (iii) spectra of both α and β human hemoglobins; (iv) cross-spectral function of the spectra presented in (iii). The prominent peaks denote common frequency components. The abscissa represents RRM frequencies, and the ordinate is the normalized intensity.

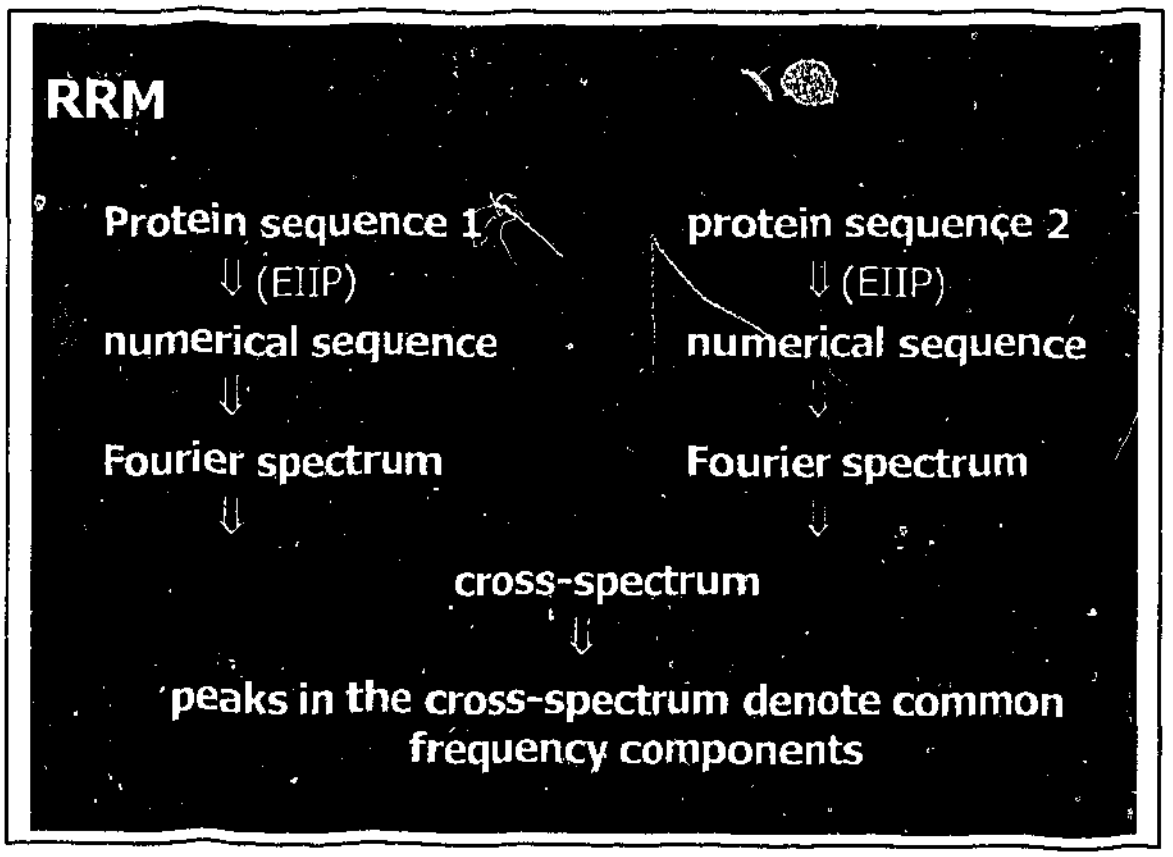


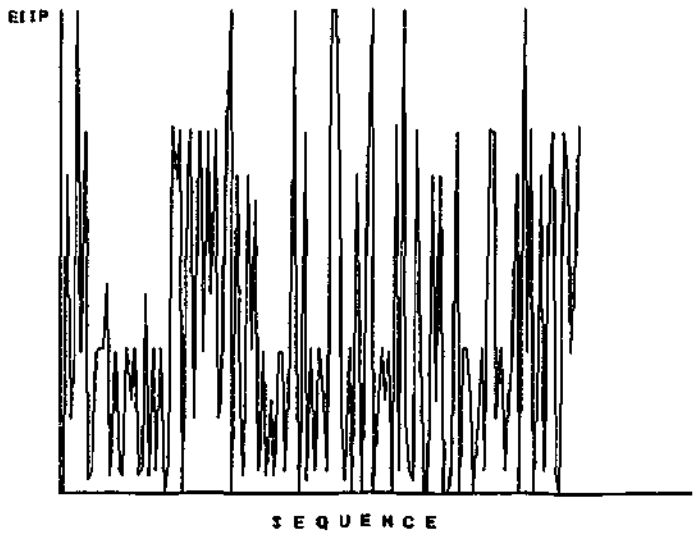
Figure 3.1a The Resonant Recognition Model (RRM) procedure

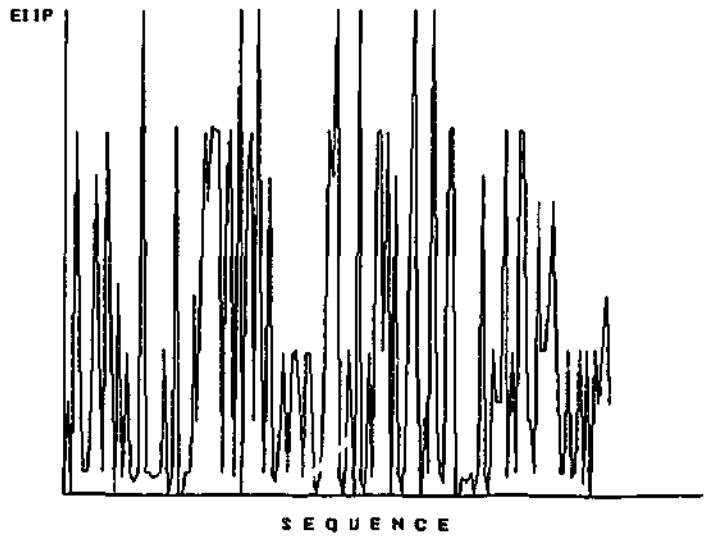
HUMAN α-HEMOGLOBIN

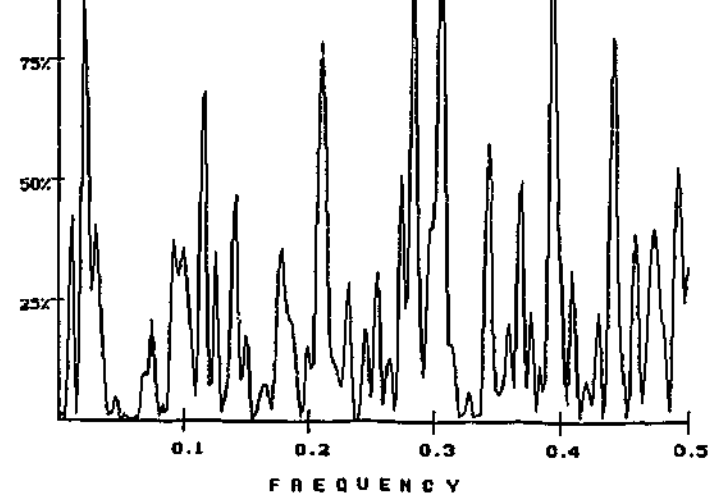
VLSPADKTNVKAAWGKVGAHAGEYGAEALER
MFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVAD
ALTNAVAHVDDMPNALSALSDLHAHKLRVDPV
NFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLA
SVSTVLTSKYR

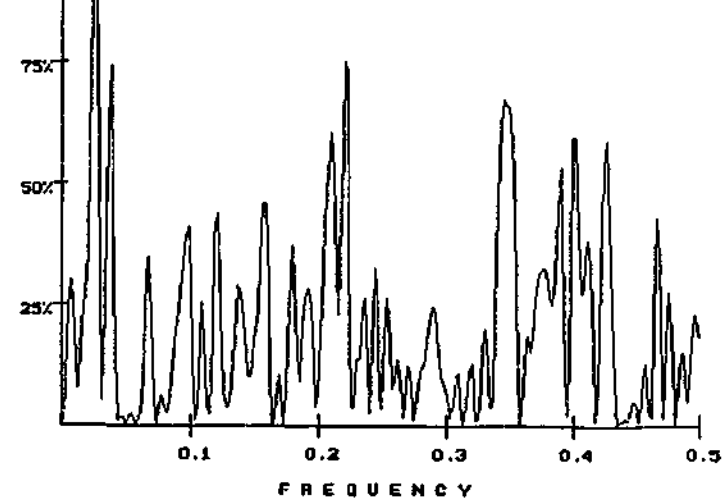
HUMAN β-HEMOGLOBIN

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVV
YPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVL
GAFSDGLAHLDNLKGTFATLSELHCDKLHVDPEN
FRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGV
ANALAHKYH









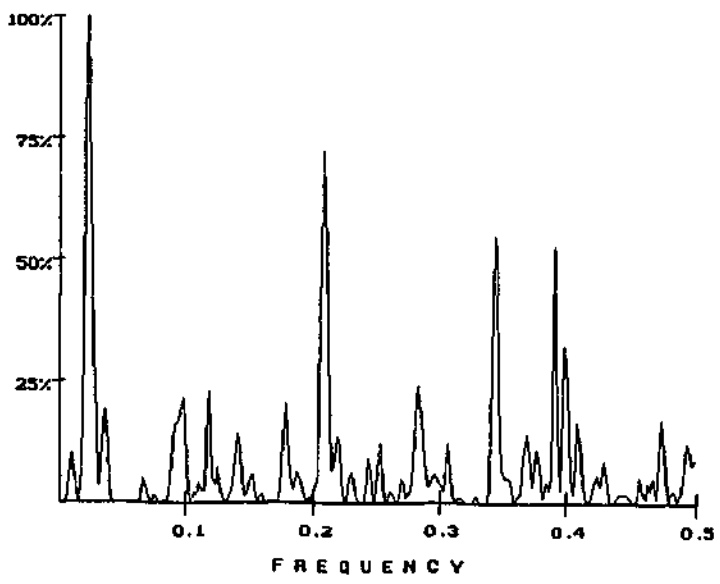


Figure 3.1b Graphical presentation of the RRM procedure [6].

Table 3.1. The Electron-Ion Interaction Potential (EIIP) Values for Nucleotides and Amino Acids

Nucleotide	Nucleotide	EIIP
A	A	0.1260
G	G	0.8606
T	T	0.1335
C	C	0.1340
Amino Acid	**Amino Acid**	**EIIP**
Leu	L	0.0000
Ile	I	0.0000
Asn	N	0.0036
Gly	G	0.0050
Val	V	0.0057
Glu	E	0.0058
Pro	P	0.0198
His	H	0.0242
Lys	K	0.0371
Ala	A	0.0373
Tyr	Y	0.0516
Trp	W	0.0548
Gln	Q	0.0761
Met	M	0.0823
Ser	S	0.0829
Cys	C	0.0829
Thr	T	0.0941
Phe	F	0.0946
Arg	R	0.0959
Asp	D	0.1263

To determine the common frequency components for a group of protein sequences, the absolute values of multiple cross-spectral function coefficients *M* have been calculated as follows:

$$/Mn/ = /X1n/ \cdot /X2n/ \ldots /XMn/ \qquad n = 1, 2, \ldots, N/2 \tag{3.4}$$

Peak frequencies in such a multiple cross-spectral function denote common frequency components for all sequences analyzed. Signal-to-noise ratio (S/N) for each peak is defined as a measure of similarity between sequences analyzed [5,6]. S/N is calculated as the ratio between signal intensity at the particular peak frequency and the mean value over the whole spectrum. The extensive experience gained from previous research suggests that an S/N ratio of at least 20 [57-59,67-70] can be considered significant. The multiple cross-spectral function for a large group of sequences with the same biological function has been named "consensus spectrum". The presence of a peak frequency with a significant S/N

ratio in a consensus spectrum implies that all of the analyzed sequences within the group have one frequency component in common [5,6]. This frequency is related to the biological function provided the following RRM criteria are met:

A: One peak only exists for a group of protein sequences sharing the same biological function

B: No significant peak exists for biologically unrelated protein sequences

C: Peak frequencies are different for different biological functions.

In previous studies, the above criteria have been tested with over 1000 proteins from 25 functional groups [5,6,57-59,67-70]. The regulatory DNA sequences were analyzed in the same way. The following fundamental conclusion was drawn from these studies: *each specific biological function of protein or regulatory DNA sequence is characterised by a single frequency.*

The RRM also provides a method to identify the individual amino acids that contribute mostly to a protein's specific biological function. Since the characteristic frequency correlates with the biological function, the positions of the amino acids that are most affected by the change of amplitude at the frequencies and consequently to the corresponding biological function. The inverse Fourier transform is used in this determining process. This "hot spots" identification method has been applied already to a number of examples. The previous studies with interleukin 2, SV40 enhancer, Ha-ras p21 oncogene product, glucagons, Cytochrome C, haemoglobins, myoglobins and lysozymes all have documented evidence that such predicted amino acids are found to be spatially clustered in the protein tertiary structure and to be positioned in and around the protein active site in some of those examples [5,6,58].

Signal-Noise Ratio Normalization

The frequency characteristic for an observed biological function is defined as the prominent peak frequency in the multiple cross-spectral function of the family of proteins having this function in common. This prominent peak denotes a common frequency

component for all the protein sequences analyzed. Any common frequency component can be considered as a characteristic of the observed function providing certain criteria are satisfied [6]. To quantify those criteria it is important initially to define the index, which will be a measure of the level of commonality for any frequency component in the cross-spectral function. However, this definition of S/N is sensitive to the protein signal length and the number of sequences that computed in the multiple cross-spectral function. Thus, it is not quite suitable for the development of the general criteria for a variety of protein families. There was a need to define S/N as a general comparable index that could be consistent and universally applied to the system under examination. This normalized S/N index has to be independent of the sequence length and the number of sequences analyzed. This index should measure the confidence of determining the characteristic frequency of a biological function. After an empirical and statistical testing, a new normalized S/N ratio (NSN) has been introduced to avoid the problem discussed above [71].

All these results have shown, the RRM represents a whole new view to protein activity, in particular protein-protein and protein-DNA interactions. The underlying hypothesis of this model is that certain periodicities in the distribution of energy of delocalised electrons along the protein molecule are critical parameters for protein biological function. This model allows extract the linear information contained in amino acid sequence and also provides a physical explanation of macromolecule's interaction processes. With a characteristic frequency identified by the RRM, it is possible then to design peptides of different length having desired periodicities in their distribution of energies of delocalised electrons along their sequence [5,6]. Thus, the RRM can identify signals that characterize protein biological functions. Applications of the RRM are mostly in the area of biotechnology, drug design and pharmacology.

CHAPTER 4

MATERIALS AND METHODS

4.1 THE RRM PROGRAM AND SEQUENCES USED

The RRM-EIIP program runs on IBM PC and takes as its input operator specified single sequences and performs the Fourier Transform using the EIIP parameter alone and displays the results. A modified version, here called the RRM program, allows the operator to also specify the amino acid parameter to be used for the Fourier Transform. Consensus Spectra are obtained by operator mediated selection of further sequences with which comparisons are made.

In this study, the RRM was used to determine the effect of various amino acid parameters on the characteristic profiles of different functional protein families. Twenty two different protein groups (glucagons, lysozymes, hemoglobins, cytochromes C, EGFs, FGFs, TNFs, TGFs, PDGs, TNFs, interleukins, myoglobins, viral oncogenes, proto-oncogenes, oncogenes, p53s, cysteine proteases, metallo-proteases, serine proteases, aspartic proteases, trypsins, protease inhibitors and AIDS related proteins) were used as test examples.

The protein sequences database is presented in Appendix D.

4.2 THE PARAMETER DATABASE

The amino acid parameters used and compared in this study were mostly obtained from the literature sources cited in Kawashima, S. and Kanehisa, M. et al. [53]. An amino acid index is a set of 20 numerical values representing any of the different physicochemical and biological properties of amino acids.

There are six main groups of parameters which are categorised by their first letter identifier. These groups are (**A**) alpha and turn propensities, (**B**) beta propensities, (**C**) composition, (**H**) hydrophobicity and (**P**) physicochemical properties and (**O**) other properties. The parameters falling in the **A, B, C, H, P** and **O** categories are arranged in

order of appearance in Kawashima *et al.*; in this reference an arrangement of these parameters based on statistical correlation, as determined by cluster analysis is presented. The amino acid parameters were numerically labeled according to the listing of Kawashima *et al.* using four number identifiers. 402 amino acid parameters were entered into the database. In cases where particular amino acid values were not available for a given parameter, they were replaced by zeros. This method was chosen rather than substituting the mean of the other values, as it was found that there is no significant difference between these two methods [49]. In those cases where no parameters were given in the references cited, no entries were made. In the amino acid parameter database, independent files identifiable by their alphanumeric indicator were created for each parameter, rather than bunching all of the parameters up as one large compendium. This was done to allow easy access to any single or combination of parameters. Additionally we have included into above database the Ionization constant of amino acid parameter (IC), which values were collected from other literature sources [72,73]. Furthermore, in this study we have analyzed the computational parameters presented as well as different mathematical combinations of these amino acid indices.

The parameter database is presented in Appendix C.

4.3 CORRELATION WITH THE EIIP

Correlation is a statistical technique that is used to measure and describe a relationship between two variables. Correlation requires two scores from each individual (one score from each of the two variables). Correlation can be classified into two basic categories: positive and negative. Positive correlation: two variables tend to move in the same direction. Negative correlation: two variables tend to move in the opposite direction. A perfect correlation always is identified by a correlation of 1.00 where as a correlation of 0 indicates no relationship at all. Thus when we include the positive or negative sign our correlation coefficient can range from -1 to 1.

The Pearson's Correlation.

r=(degree to which X and Y vary together)/(degree to which X and Y vary separately)=

=(covariability of X and Y)/(variability of X and Y separately)

$$SP = \sum (X - \bar{X})(Y - \bar{Y}) \qquad SP = \sum XY - \frac{\sum X \sum Y}{n} \qquad = \frac{SP}{\sqrt{SS_X SS_Y}}$$

where SS - the amount of variability for a single variable, SP - the amount of covariability between two variables, n - refers to the number of pairs of scores.

Once the database was established, the parameters were transferred to Microsoft EXCEL spreadsheets. Pearson product-moment correlation calculations were carried out sequentially between each amino acid parameter and the referential EIIP value in order to test for any linear relationship between them [74]. Parameters having correlation coefficients in excess of +/- 0.5 were deemed to be the most strongly correlated with EIIP and were thus isolated.

4.4 COMPARISON OF SPECTRA OBTAINED

The parameters, which significantly correlate with the EIIP, were reintroduced into the RRM program and characteristic frequencies determined using these new values. These frequencies were determined for different functional protein groups under examination and then compared to frequencies obtained using EIIP values.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 EXAMINATION OF DIFFERENT AMINO ACID PARAMETERS WITHIN THE RRM

5.1.1 Application of the Ionization Constant of Amino Acids for Protein Signal Analysis within the RRM

On going studies to predict the structure and function of proteins from the primary sequence have, as yet, proven unsuccessful although significant insight has been gained. The RRM is a model that interprets the protein primary sequence [6].

As it was mentioned previously, the assignment of a particular number to the physicochemical parameter for each amino acid is a critical initial step in all RRM calculations. This set of numbers should have a physical meaning related to the protein's biological function. Previously each amino acid in the protein sequence was represented by the value of the EIIP, the basic physical parameter in the RRM approach, as presented by I. Cosic [5,6,57-59,67-70]. The EIIP is deduced from the theory of metals. It originates in the physics of condensed matter [54,61]. V. Veljkovic suggested that this parameter might also play an important role in biochemical processes, as there is evidence that the electromagnetic field surrounding a molecule is able to take part in biomolecular processes [61]. The EIIP has been applied for the investigation of the biological activity of small organic molecules, particularly the carcinogens [61]. The initial approximation used for calculation of the EIIP values for each amino acid is that this potential is only a function of the average density of electrons in a molecule. Thus, the values of the EIIP for each amino acid were mathematically obtained from an approximate pseudopotential model [54]. This is, by nature, an extremaly complex physical quantity included the following approximations: (1) the quasi-local approximations, (2) the Slater's approximation, (3) the indispesable approximation and (4) the pseudoatomic approximation [61]. Particularly for this work the pseudoatomic approximation needs to be clarified specifically. In this approximation a complex system is treated as a homogeneous system composed of

pseudoatoms bearing the averaged characteristics of atomic types out of which the analyzed real system is composed. The EIIP was used in studying of the biological processes, since these processes are themselves the results of specific molecular interactions.

Although the EIIP was successfully used in the previous studies [5,6,55,57-59], in this investigation we have tried to improve the existing RRM approach by investigating the possible use of other physicochemical amino acid parameters. For this purpose we have investigated here the use of the Ionization constant of amino acid parameter (IC) instead of the EIIP to represent each amino acid in the protein sequence. In contrast with the EIIP [54,61], the IC parameter is a measurable physical property. If the protein conductivity was introduced, then charges moving through the protein backbone might produce electromagnetic irradiation or absorption with spectral characteritics corresponding to energy distribution along the protein. The RRM is capable of calculating these spectral characteristics [6]. In the RRM interactions between molecules are considered as the transfer of resonant energy between interacting molecules. On the other hand, every chemical reaction itself is considered as a transfer of free radicals (ions) between the reacting chemical species that involve the energy transform. To explain the selectivity of protein-protein interactions and taking into account that protein' structure and function are determined by the sequence of amino acids, it leads us to the investigation of the chemical properties of amino acids. In particular, the dual character of amino acids, which gives rise to ionic behavior that is chategorized by the ionization (dissociation) constant. The values of the IC parameter were collected [72,73] and tested for the use for structure/function analysis of different protein families within the RRM.

Acids and bases can be classified as proton donors and proton acceptors, respectively. This means that the conjugate base of a given acid will carry a net charge that is more negative than the corresponding acid. In biologically relevant compounds various weak acids and bases are encountered, e.g. the acidic and basic amino acids, nucleotides, phospholipids etc. Weak acids and bases in solution do not fully dissociate and, therefore, there is equilibrium between the acid and its conjugate base. This equilibrium can be calculated and is termed the equilibrium constant, Ka. This is also sometimes referred to as the dissociation constant as it pertains to the dissociation of protons from acids and bases.

The ionization reaction $(H—A)=(H^+)+(A^-)$ will be governed by an acid Ionization (Dissociation) Constant:

$$Ka = (H^+)\cdot(A^-)/(H—A), \qquad (5.1.1.1)$$

where (H^+) - concentration of positively charged ions in the solution, (A^-) - concentration of negatively charged ions in the solution, and (H—A) - concentration of a reactant.

Therefore, in obtaining the -log of both sides of the equation describing the dissociation of a weak acid we arrive at the following equation:

$$-\log Ka = -\log[H^+][A^-]/[H—A] \qquad (5.1.1.2)$$

Since as indicated above -logKa = pKa and taking into account the laws of logarithms:

$$pKa = -\log[H^+] - \log[A^-]/[H—A] \qquad (5.1.1.3)$$

$$pKa = pH - \log[A^-]/[H—A] \qquad (5.1.1.4)$$

From this equation it can be seen that the smaller the pKa value the stronger is the acid. This is due to the fact that the stronger an acid the more readily it will give up H^+ and, therefore, the value of [H—A] in the above equation will be relatively small. It is convenient to express Ka in a form analogous to pH:

$$pKa = -\log Ka \qquad (5.1.1.5)$$

When conditions are adjusted such that the acid is half ionized, $(H—A) = (A^-)$, and $Ka = (H^+)$ or pKa = pH. In other words, the term pKa is that pH at which an equivalent distribution of acid and conjugate base (or base and conjugate acid) exists in solution. It should be noted that around the pKa the pH of a solution does not change appreciably even when large amounts of acid or base are added. This phenomenon is known as buffering. In most biochemical studies it is important to perform experiments, that will consume H^+ or OH^- equivalents, in a solution of a buffering agent that has a pKa near the pH optimum for the experiment.

The experimental values in most cases are measured at 25°C. The widely used method for the determination of pKa is direct potentiometric titration while in some cases are optical methods [7]. Because pH is an important environmental variable in the study of proteins, it should be monitored within the direct potentiometric titration process. Titration measurements at high ionic strength predominantly reflect only electrostatic interactions between nearby charged groups. For intact proteins, a total potentiometric titration is a moderately simple experiment when sufficient quantities of proteins are available. However, the interpretation of the results is difficult because it is often not clear to which residue to assign a particular region of the titration curve. In these cases, other (more specific) ways of following the titration of individual residues must be selected, for instance optical methods. A typical spectroscopic experiment is simple. Electromagnetic radiation at a certain nominal wavelength is allowed to impinge on the sample. Then some properties of the radiation that emerges from the sample are measured. One of the simplest properties is the fraction of the incident radiation absorbed or dissipated by the sample. Not only the emergent intensity, but also the distribution of emergent frequencies, are sources of information.

The corresponding values of the IC parameter are taken from the references [72,73] and presented in Table 5.1.1.1 (column 3). The correlation coefficient between the EIIP values and IC values was calculated. This correlation coefficient is **-0.794** indicating a strong correlation [74] between these parameters (Table 5.1.1.1). Error was calculated as the ratio 1/N, where N is a number of the amino acid sequences in a particular protein group.

In this work each amino acid in the protein sequence is represented by the corresponding value of the IC parameter instead by the EIIP parameter as it was done previously. Numerical series obtained in this way are analyzed as previously by transforming them into the frequency domain using DFT in order to extract information pertinent to the biological function.

Sequences from different functional protein families (glucagons, lysozymes, hemoglobins, cytochrome C, EGFs, TGFs, interleukins, PDGs, TGFs and TNF) were analyzed, and a multiple cross-spectral analysis was performed for each protein group using the EIIP and IC parameter values. The S/N ratio value was calculated in the same way as previously.

Table 5.1.1.1 EIIP and IC Parameters and their Correlation Coefficient.

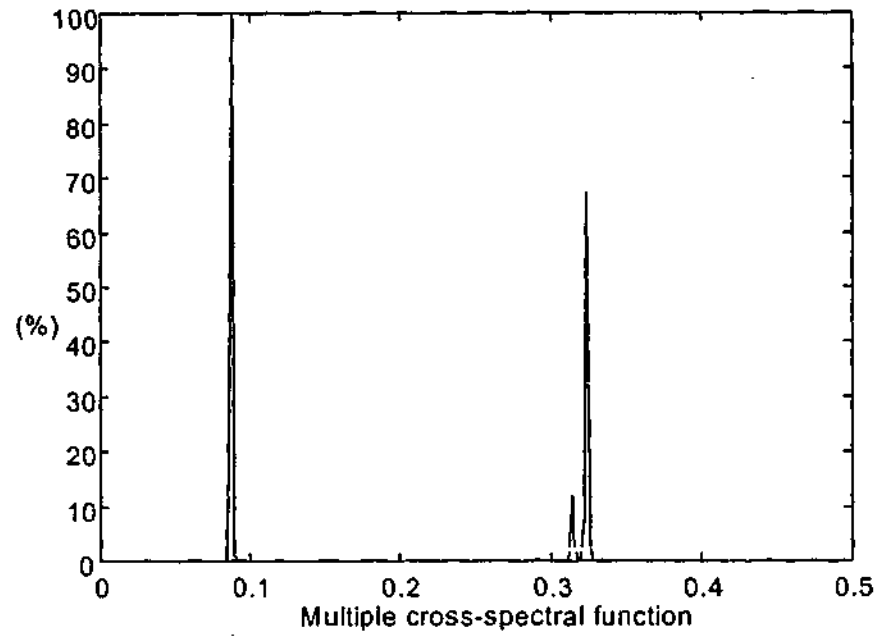
Amino Acid	EIIP	IC
L	0.0000	2.40
I	0.0000	2.40
N	0.0036	2.20
G	0.0050	2.46
V	0.0057	2.35
E	0.0058	2.30
P	0.0198	2.00
H	0.0242	2.30
K	0.0371	2.20
A	0.0373	2.30
Y	0.0516	2.20
W	0.0548	2.37
Q	0.0761	2.06
M	0.0823	2.17
S	0.0829	2.10
C	0.0829	1.96
T	0.0941	2.09
F	0.0946	1.98
R	0.0959	1.82
D	0.1263	1.88
Correl. Coefficient with EIIP.		-0.794

The peak frequencies, S/N and error values for each protein group are shown in Table 5.1.1.2.

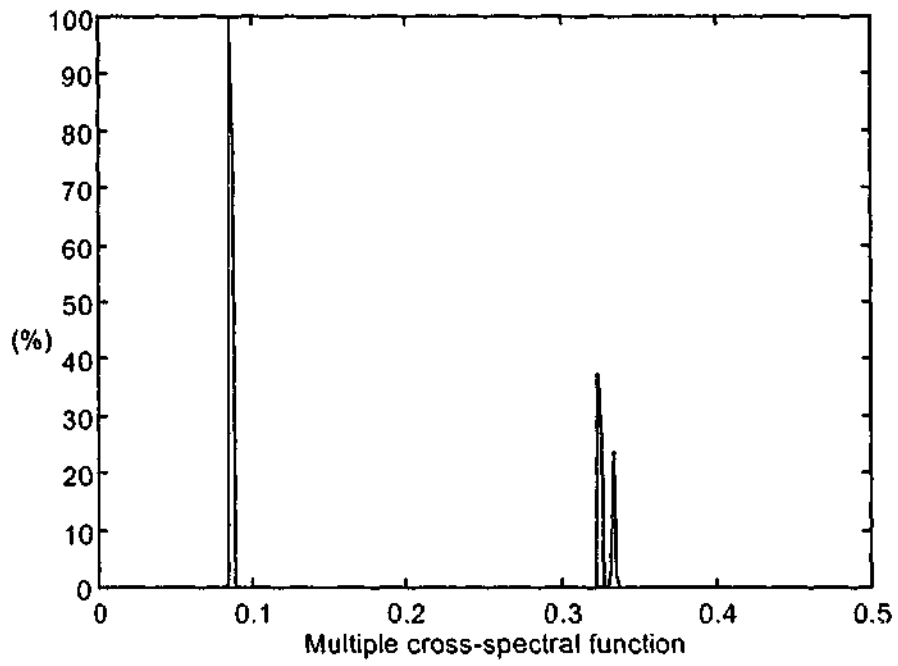
Table 5.1.1.2 Peak frequency and Signal-to-Noise Ratio Values for Protein Groups

Protein Group	EIIP		IC		No. seq.	Error
	Frequency	S/N	Frequency	S/N		
Glucagon	0.0879	122.7	0.0859	94.4	30	0.0333
Lysozyme	0.3281	238.3	0.3281	249.2	17	0.0588
Haemoglobin	0.0254	256.0	0.0254	256.0	85	0.0118
Cytochrome C	0.4746	247.4	0.4746	208.6	46	0.0217
EGF	0.0605	245.7	0.0605	255.8	73	0.0137
FGF	0.4551	255.9	0.4551	221.4	38	0.0263
TGF	0.0137	29.3	0.0117	38.1	3	0.33
Interleukin	0.0410	191.5	0.0957	192.2	11	0.0909
PDG	0.4199	75.1	0.1934	32.5	4	0.25
TNF	0.0527	119.0	0.0762	79.9	9	0.1111

As it was assumed, sequences with the same biological function share the same characteristic common frequency components. Each specific biological function is characterized by a single frequency.

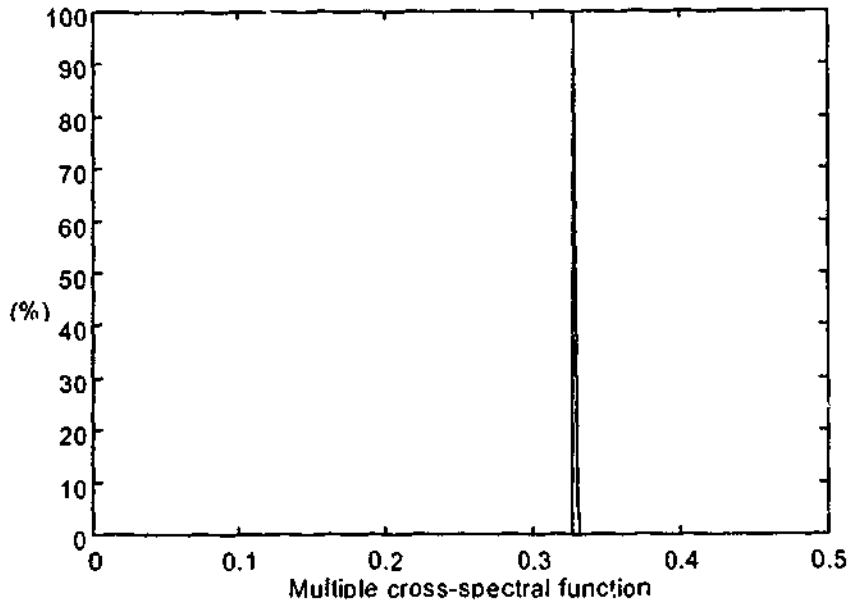


(a)

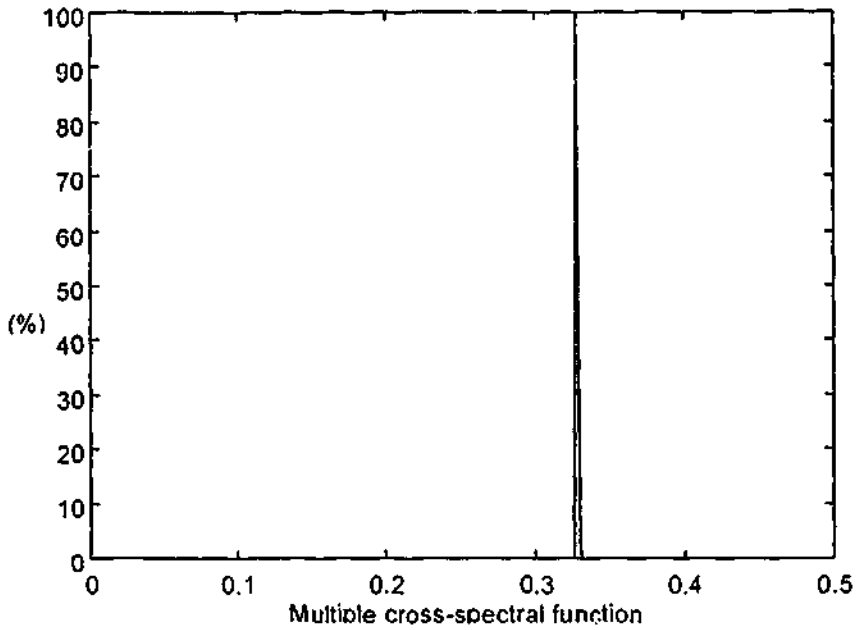


(b)

Figure 5.1.1.1 Multiple cross-spectral function of Glucagon proteins using (a) EIIP and (b) IC parameters.

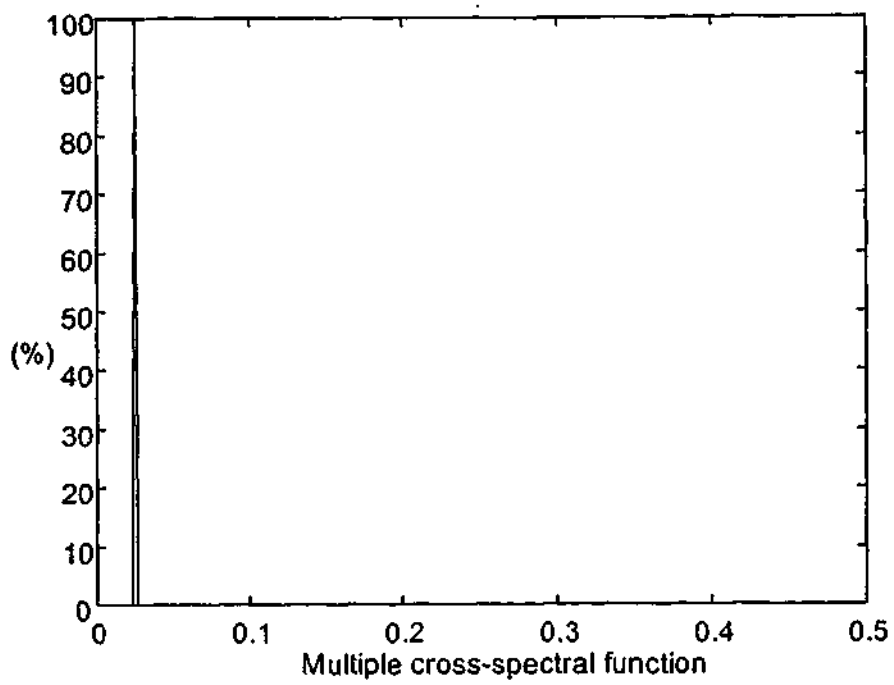


(a)

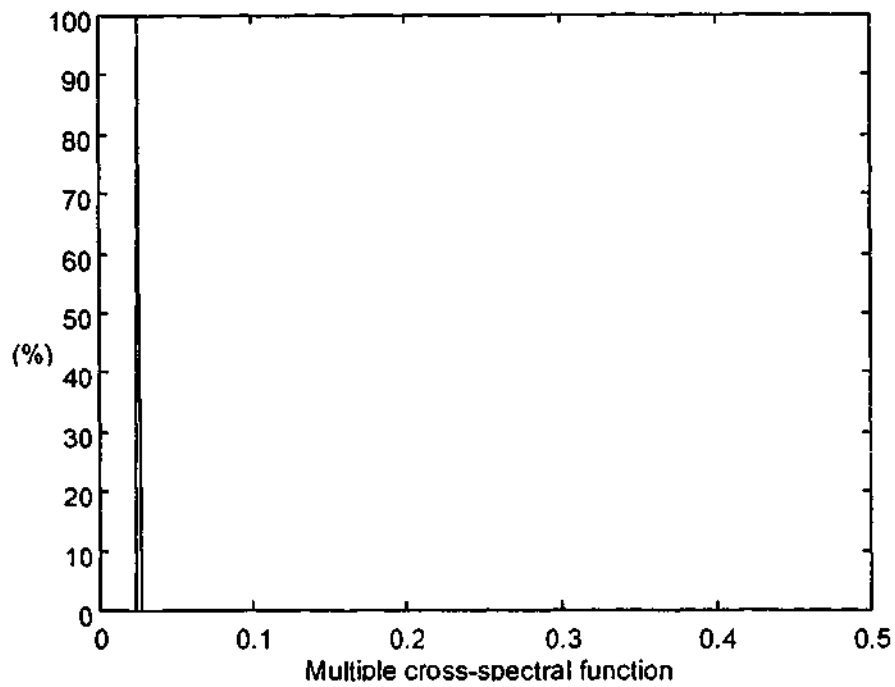


(b)

Figure 5.1.1.2 Multiple cross-spectral function of Lysozyme proteins using (a) EIIP and (b) IC parameters.

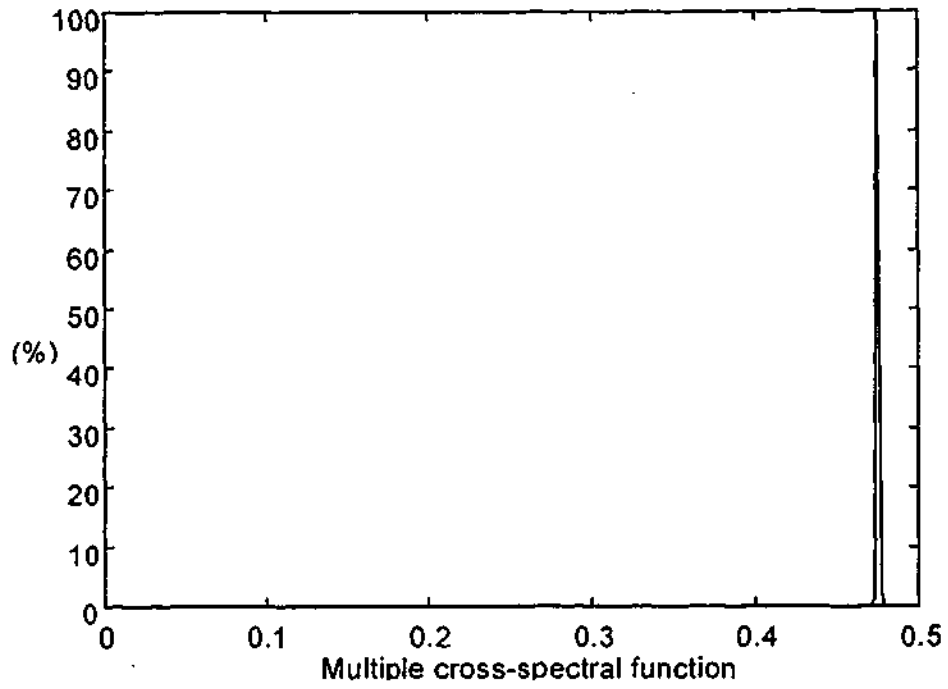


(a)

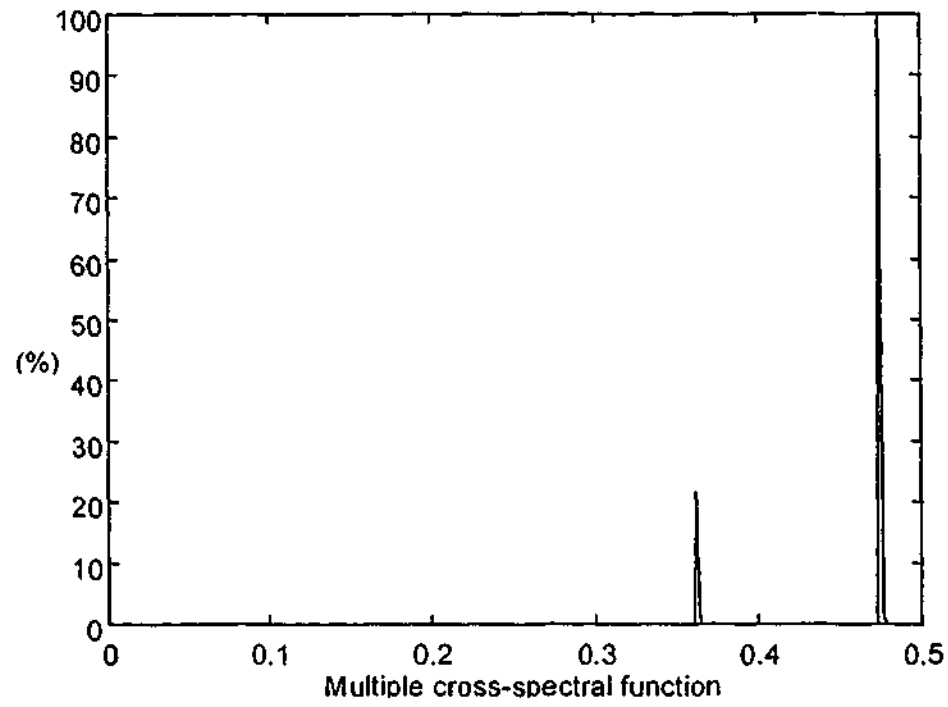


(b)

Figure 5.1.1.3. Multiple cross-spectral function of Hemoglobin proteins using (a) EIIP and (b) IC parameters.

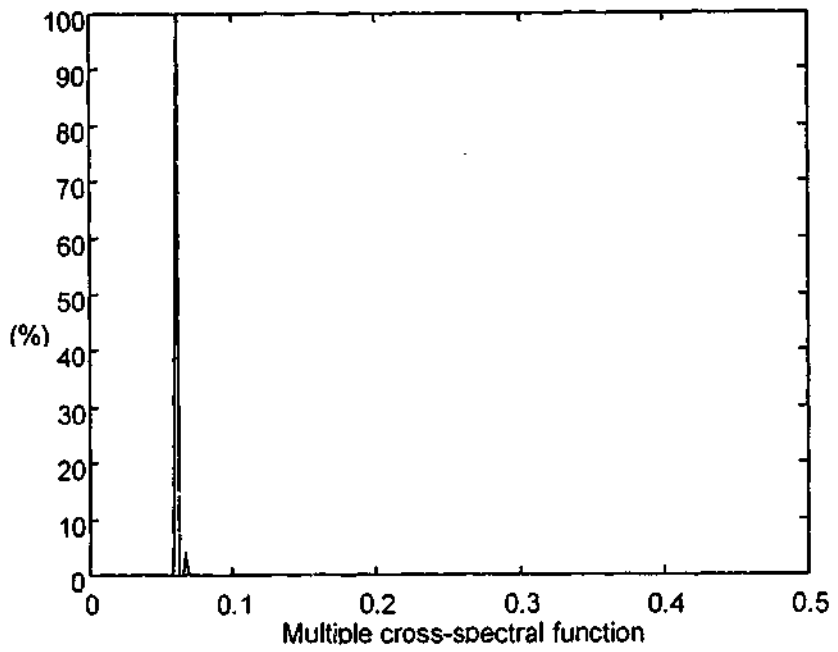


(a)

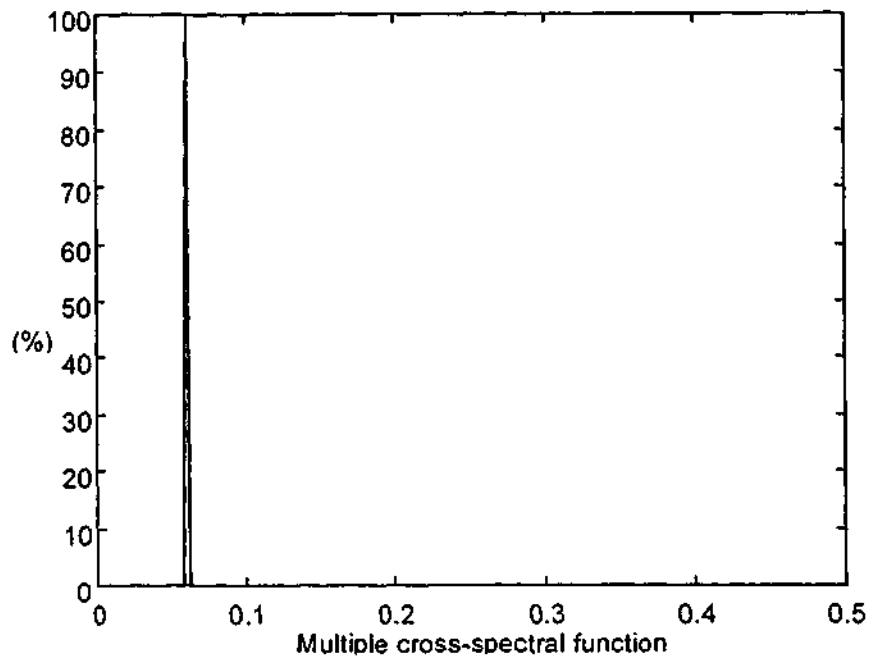


(b)

Figure 5.1.1.4. Multiple cross-spectral function of Cytochrome C proteins using (a) EIIP and (b) IC parameters.

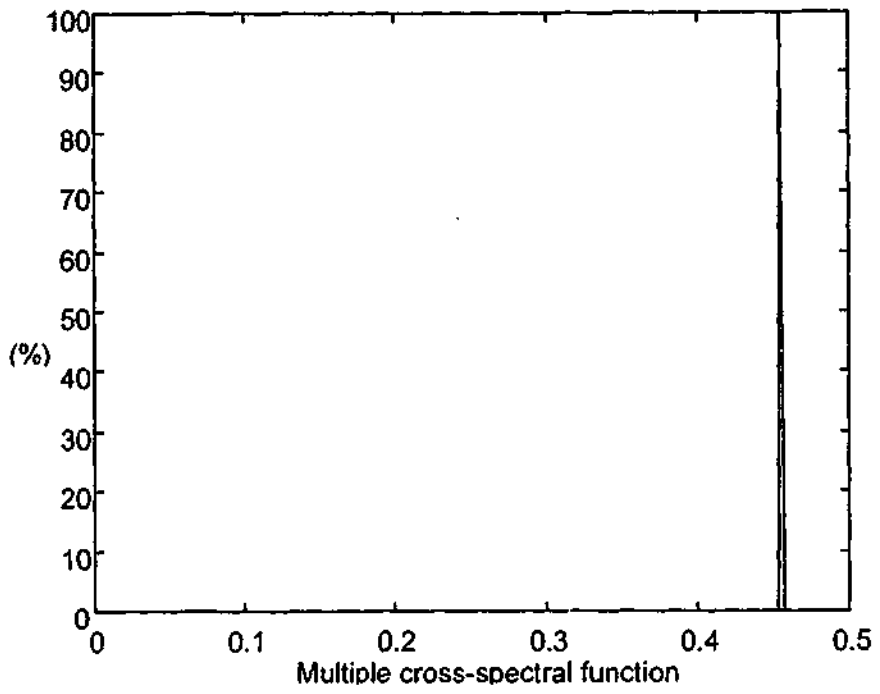


(a)



(b)

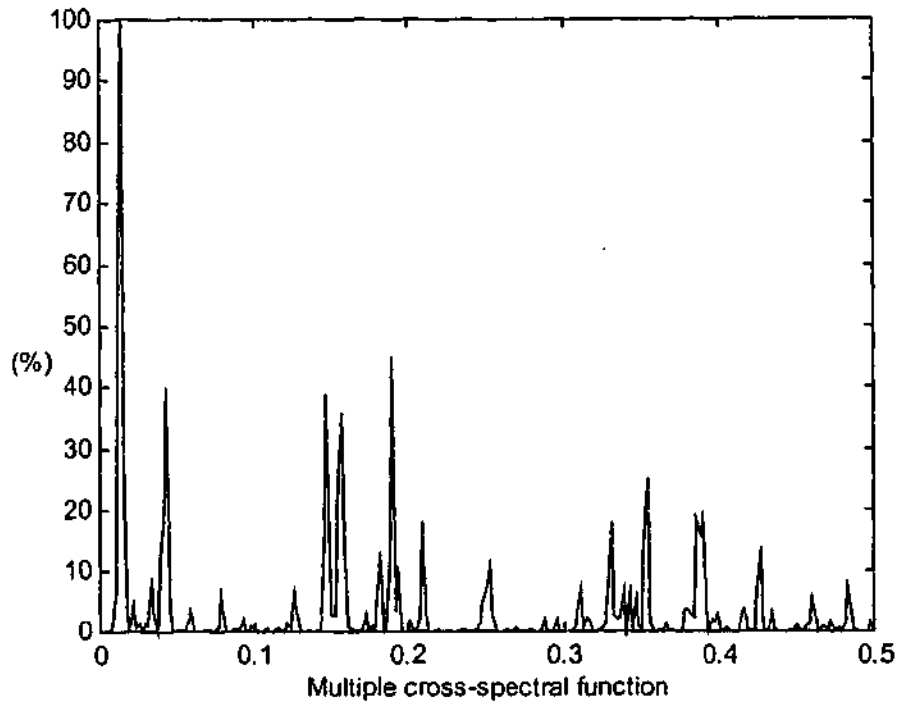
Figure 5.1.1.5 Multiple cross-spectral function of EGF proteins using (a) EIIP and (b) IC parameters.



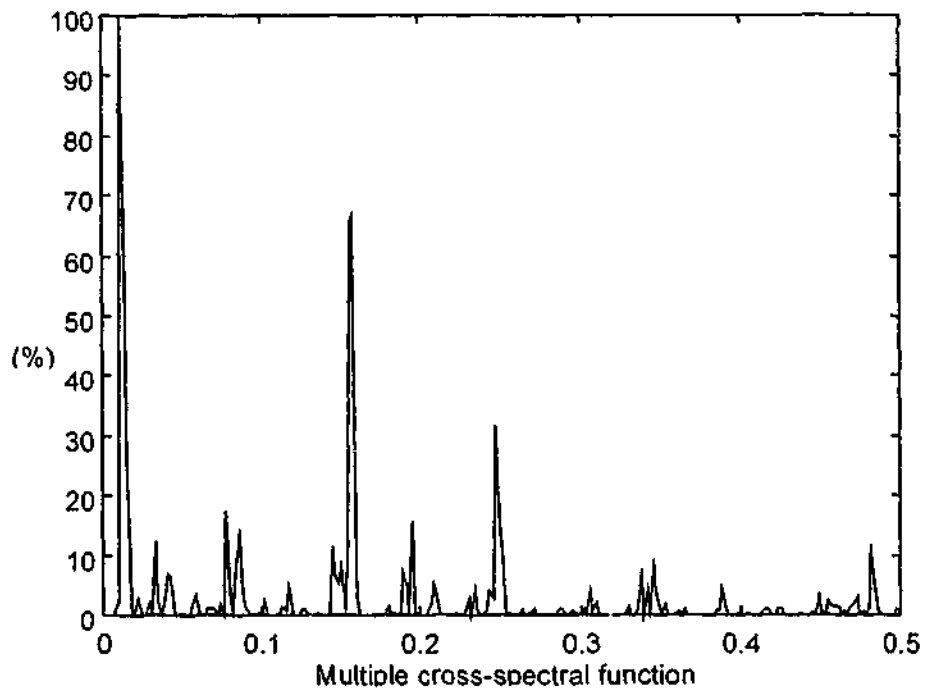
(a)

(b)

Figure 5.1.1.6 Multiple cross-spectral function of FGF proteins using (a) EIIP and (b) IC parameters.

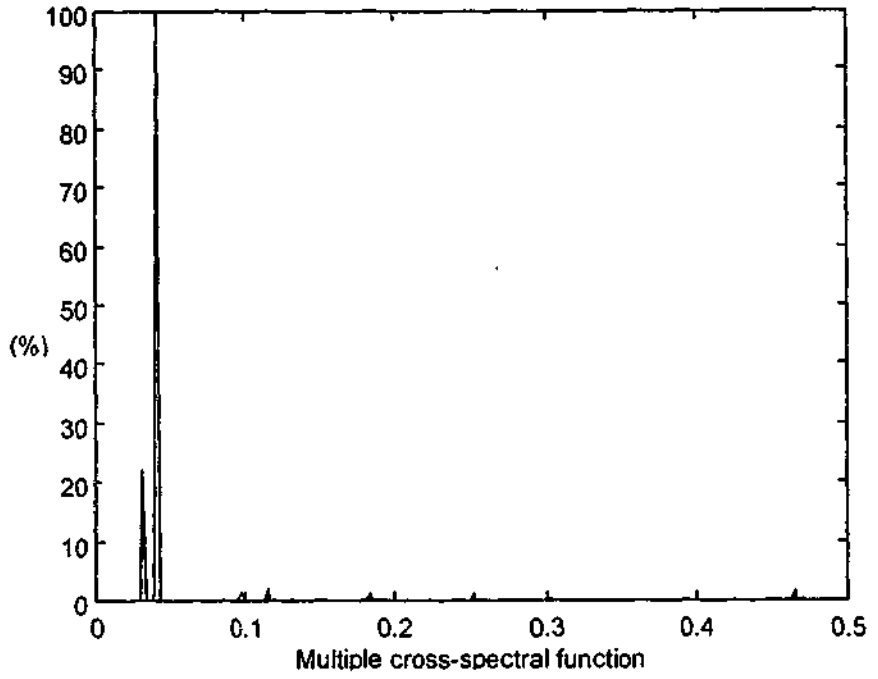


(a)

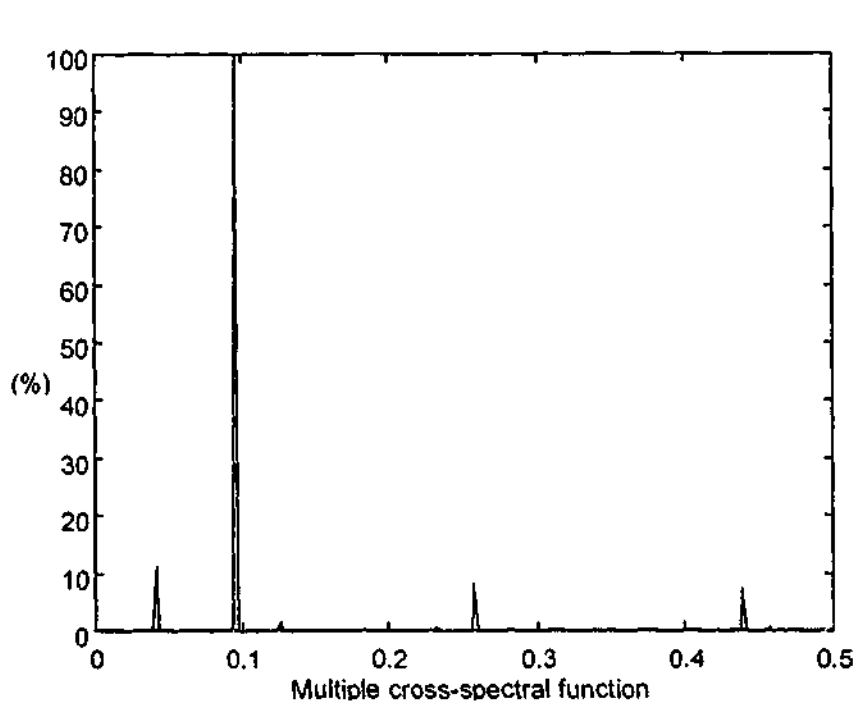


(b)

Figure 5.1.1.7 Multiple cross-spectral function of TGF proteins using (a) EIIP and (b) IC parameters.

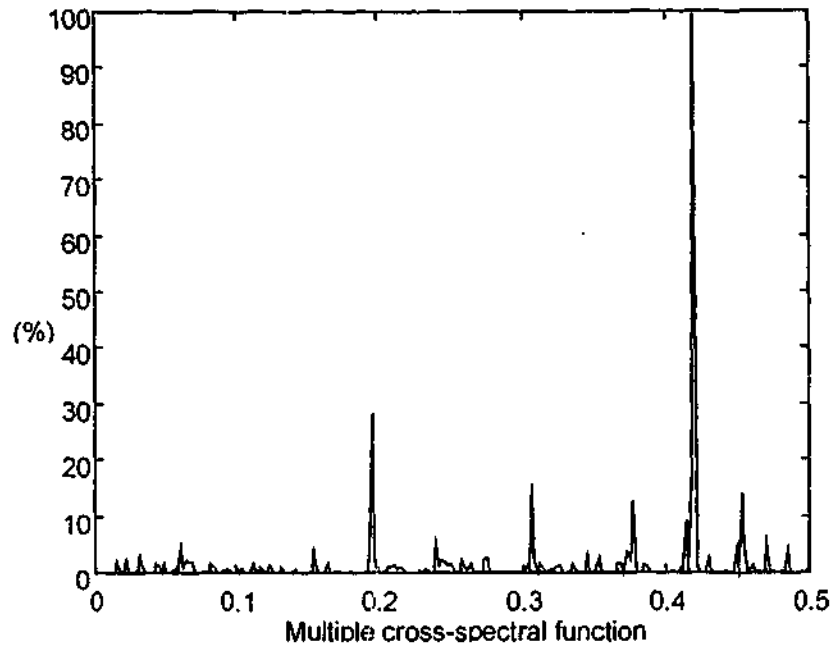


(a)



(b)

Figure 5.1.1.8 Multiple cross-spectral function of Interleukin proteins using (a) EIIP and (b) IC parameters.



(a)

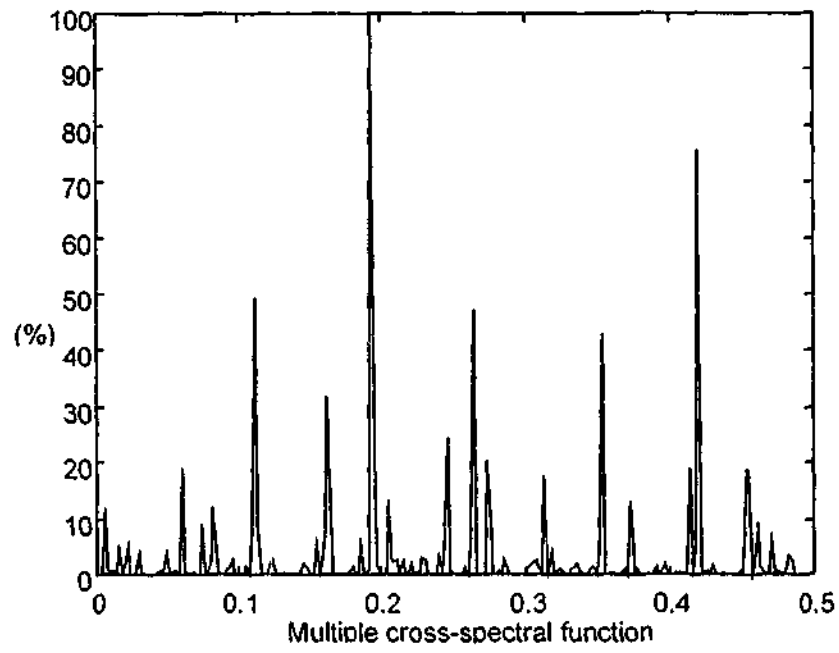
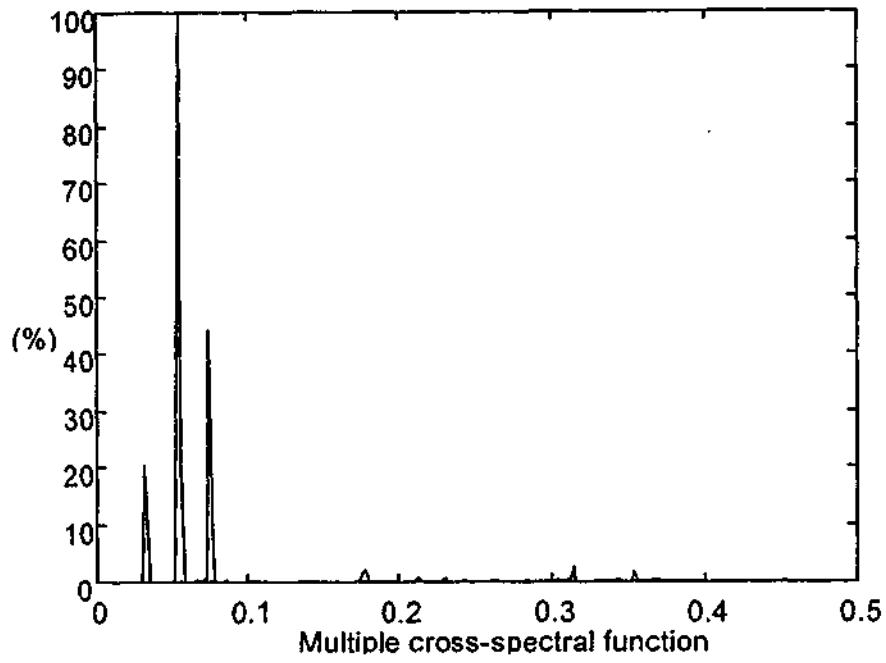
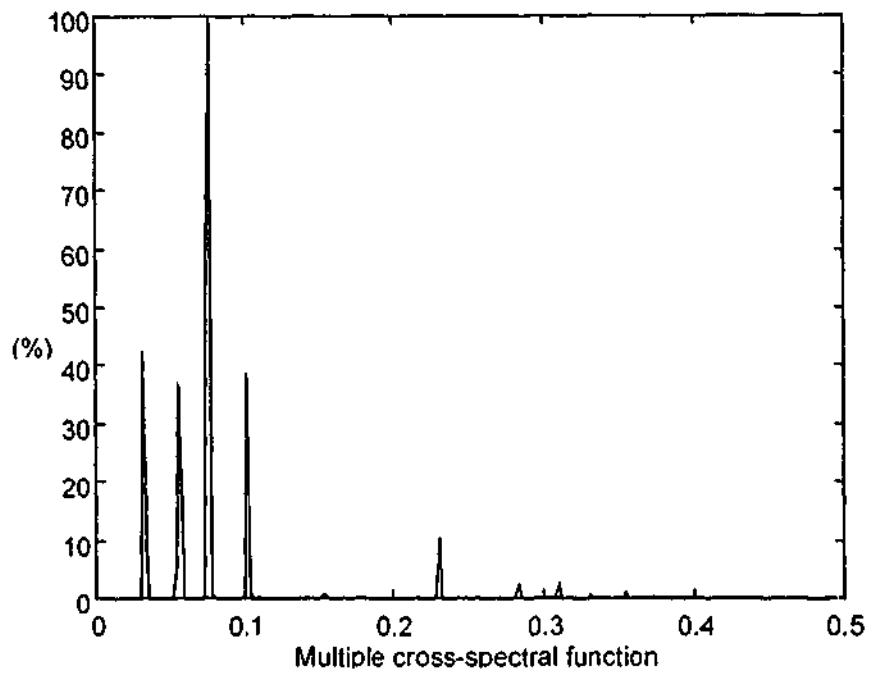


Figure 5.1.1.9 Multiple cross-spectral function of PDG proteins using (a) EIIP and (b) IC parameters.



(a)



(b)

Figure 5.1.1.10 Multiple cross-spectral function of TNF proteins using (a) EIIP and (b) IC parameters.

Results of this study have shown (Fig. 5.1.1.1-5.1.1.6) that the IC parameter generates in the consensus spectrum one dominant peak corresponding to a common biological activity of the selected proteins. The presence of a peak frequency with a significant S/N in a consensus spectrum implies that all of the analyzed amino acid sequences within the protein group have one frequency component in common. This common frequency component denotes the protein biological function of the analyzed protein family. This frequency is related to the biological function provided the following definition: (a) one peak only exists for a group of sequences sharing the same biological function, (b) no significant peak exists for biologically unrelated protein sequences. Results of this investigation reveal that the IC parameter is satisfied the RRM criteria A listed previously.

It could be observed (Table 5.1.1.2) that for Glucagon, Lysozyme, Haemoglobin, Cytochrome C, EGF and FGF proteins the same characteristic frequencies (within the calculation error) were obtained using both EIIP and IC parameter values. These similarities of the proteins characteristic frequencies are expected as comparable EIIP and IC parameters are strongly correlated (Table 5.1.1.1). This correlation between the parameters is reflected into the analogy of the consensus spectra of the pointed out proteins.

Figure 5.1.1.1 represents an example of the Glucagon consensus spectrum when they are obtained using EIIP and IC parameters respectively. Both spectra are very similar revealing the same common characteristic frequency peak at f=0.0879. The second significant frequency is at f=0.320 in both spectra but with different amplitude ratios. For glucagons particularly it was observed that these proteins have two characteristic frequencies f=0.0879 and f=0.320. It is difficult to distinguish which one among them is more important. In previous work [5,6] the frequency f=0.0879 was considered as the most common (characteristic) of Glucagon general activity. However, here we have discovered that f=0.320 is also, if not more, an inherited part of Glucagon characteristic.

It is interested to note that for other protein groups TGFs, Interleukins, PDGs and TNFs we observe (Table 5.1.1.2) different characteristic frequencies using EIIP and IC parameters. After careful examining the corresponding consensus spectra (Fig.5.1.1.7-5.1.1.10) it is clear that these observed spectra have more than one peak and those peaks are at the same frequencies for both EIIP and IC parameters but with different amplitude ratios. Thus, not

the same characteristic frequencies were recorded. This fact could be easily confirmed with the example of TNF proteins (Fig 5.1.1.10). Indeed all three main peaks identified by using the EIIP (Fig. 5.1.1.10a) parameter appear in the spectrum obtained using the IC (Fig. 5.1.1.10b) parameter but with different amplitude ratios.

Based on the results obtained and following the aim of the investigation of the possible use of the IC parameter in the RRM, we conclude that IC property is suitable parameter to use for structure/function analysis of unrelated proteins within this approach.

5.1.2 The use of the Ionization Constant of amino acids for protein signal analysis within the RRM - Application to Oncogenes

In this chapter we have compared results of the RRM investigation of oncogene products using both EIIP and IC parameters. In addition protein P53, as well known anti-oncogenic factor, have been analyzed.

The sequences from the following protein families: viral oncogenes, proto-oncogenes, oncogenes and p53 proteins have been investigated. A multiple cross-spectral analysis was performed for each analyzed protein group, using EIIP and IC values accordingly. The peak frequency and signal-to-noise values are shown in Table 5.1.2.1.

Table 5.1.2.1 Peak frequency and Signal-to-Noise Ratio Values for Protein Groups using selected analyzed EIIP and Ionization Constant parameters.

Parameter	**EIIP**		**Ionization constant**	
Protein Group	**Frequency**	**S/N**	**Frequency**	**S/N**
Viral Oncogene	0.1611	375.1	0.4844	368.4
Proto-oncogene	0.0576	219.4	0.4189	260.0
Oncogene	0.0332	504.6	0.4180	511.9
P53 protein	0.1943	155.9	0.1494	348.5

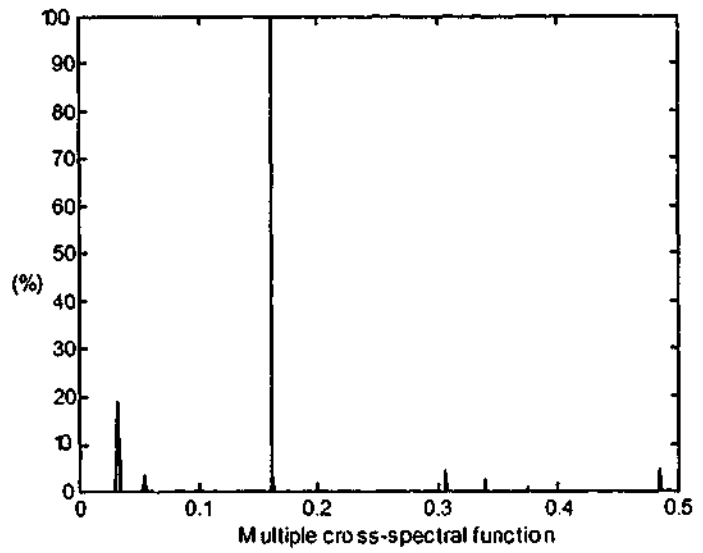
As it was assumed in the RRM theory, each specific biological function is characterized by a single frequency. Results obtained have shown that the proposed IC parameter generates in the consensus spectrum one dominant peak corresponding to a common biological activity of the selected proteins when performed according to criteria determined within the RRM.

It is interesting to note that for all protein groups selected for RRM analysis the characteristic frequencies, using both parameters EIIP and IC are different. After careful examining the corresponding consensus spectra (Fig.5.1.2.1-5.1.2.4) we observe that these spectra have more than one peak and those peaks are at the same frequencies for both parameters but with different amplitude ratios.

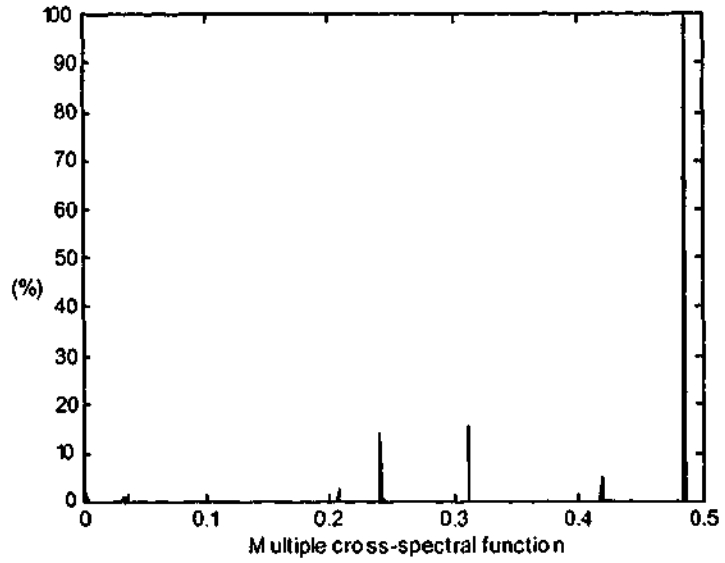
The results have shown that while EIIP can be used successfully to distinguish between oncogenes and proto-oncogenes (Fig.5.1.1.2-5.1.2.3), the IC can successfully identify

common characteristics for the whole oncogene protein family. In addition protein P53, as well known anti-oncogenic factor, have been analyzed. It has been shown that IC is more successful than EIIP in determining the p53 characteristic (Fig.5.1.2.4).

However, although in some cases the IC parameter presented is more suitable for the structure/function prediction, it is still difficult to conclude which one of these two comparable parameters solely would be the best parameter to use in the RRM.

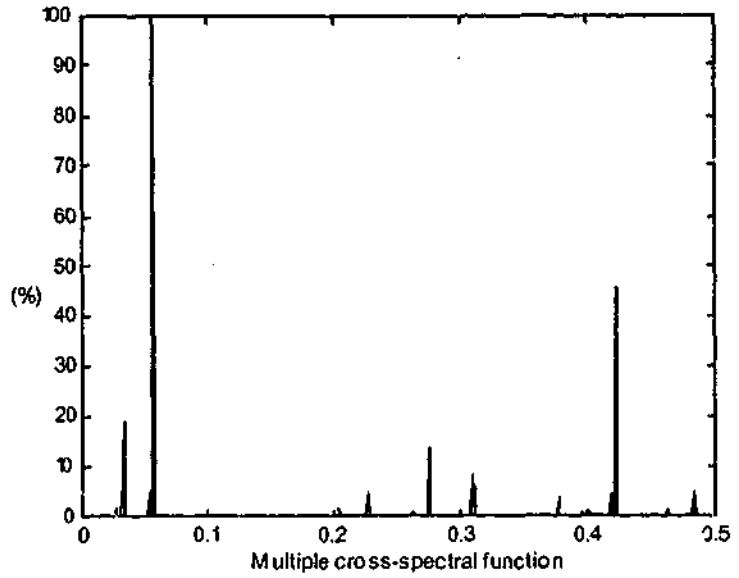


(a)

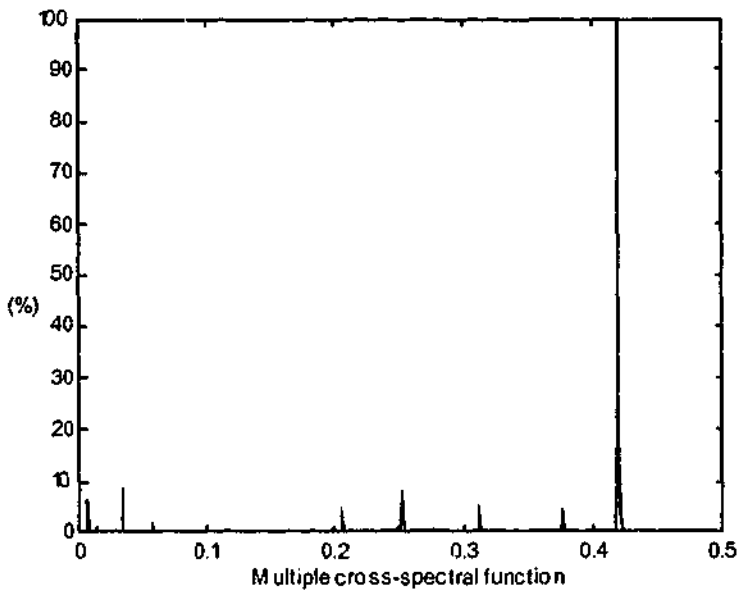


(b)

Figure 5.1.2.1 Multiple cross-spectral function of Viral Oncogene proteins using (a) EIIP and (b) IC parameters.

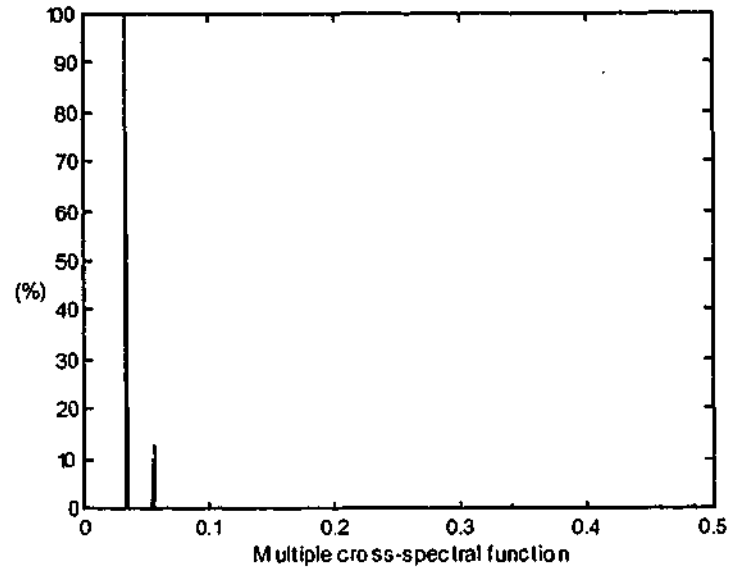


(a)

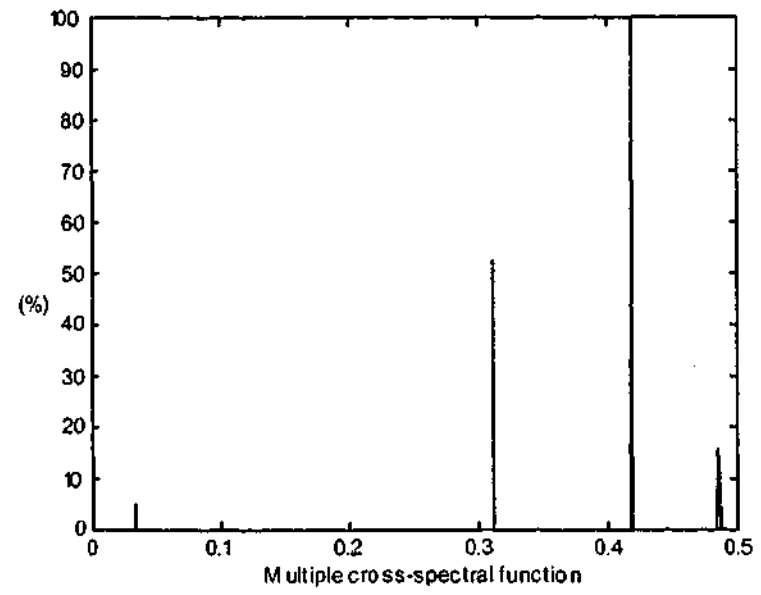


(b)

Figure 5.1.2.2 Multiple cross-spectral function of Proto-Oncogene proteins using (a) EIIP and (b) IC parameters.

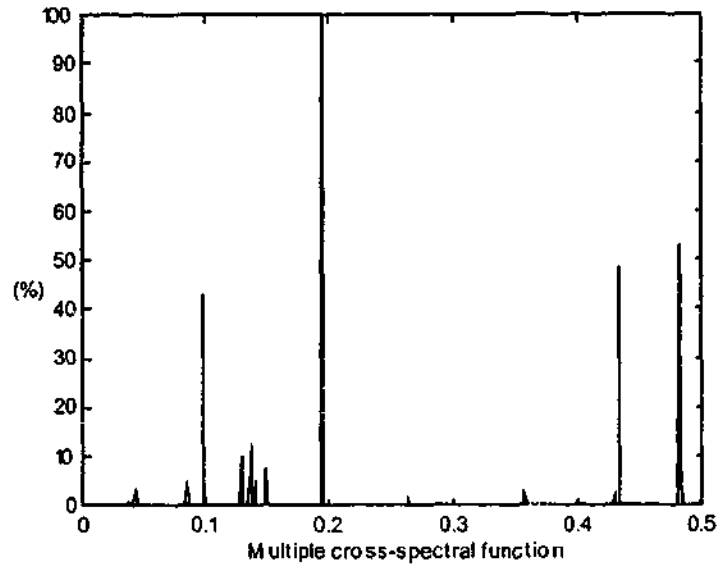


(a)

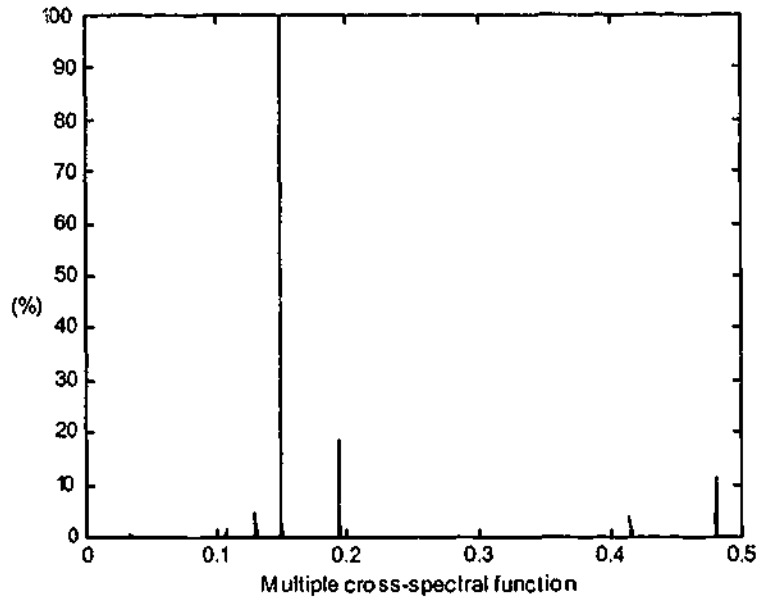


(b)

Figure 5.1.2.3 Multiple cross-spectral function of all Oncogene proteins using (a) EIIP and (b) IC parameters.



(a)



(b)

Figure 5.1.2.4 Multiple cross-spectral function of p53 proteins using (a) EIIP and (b) IC parameters.

5.1.3 The use of the Ionization Constant of amino acid for protein signal analysis within the RRM - Application to Proteases

The RRM has been employed here for the determination of biological profiles of protein groups under examination and for further investigation of the possible use of the IC parameter instead of the EIIP in the RRM approach.

The RRM analysis was performed for protease protein family using EIIP and IC parameters. As test examples the following protease subgroups:

- cysteine protease (23 sequences),
- metallo-protease (17 sequences),
- serine protease (38 sequences),
- aspartic protease (47 sequences),
- trypsin/chymotrypsin (15 sequences),
- AIDS related protease (25 sequences)
- protease inhibitors (26 sequences).

Sequences from all selected functional groups were investigated and a multiple cross-spectral analysis was performed for each protein group, using EIIP and IC parameter values consequently. As a result, protein characteristic frequencies for each protein functional group were obtained. The peak frequency and signal-to-noise values for each protease subgroup are shown in Table 5.1.3.1.

The assumption used here is that sequences with the same biological function share the same characteristic periodic component in the distribution of energy of delocalised electrons.

Table 5.1.3.1 Peak frequency and signal-to-noise ratio values for protein groups.

	EIIP		IC	
Protein Group	**Frequency**	**S/N**	**Frequency**	**S/N**
Cysteine	0.0596	197.9	0.4902	248.1
Metallo	0.2520	268.4	0.1699	165.9
Serine	[illegible]	[illegible]	[illegible]	[illegible]
Aspartic	0.1816	429.9	0.2178	167.4
Trypsin/Chymotrypsin	0.3457	178.3	0.0332	174.7
AIDS proteins	0.3447	481.0	0.0127	489.5
Inhibitors	[illegible]	[illegible]	[illegible]	[illegible]

It could be observed that in both cases (using EIIP and IC values) each specific biological function is characterized by a single frequency. Results obtained have shown that the analyzed IC parameter generates in consensus spectrum one dominant peak (Fig.5.1.3.1-5.1.3.7) corresponding to common biological activity of studied proteins and is satisfied all RRM criteria listed previously.

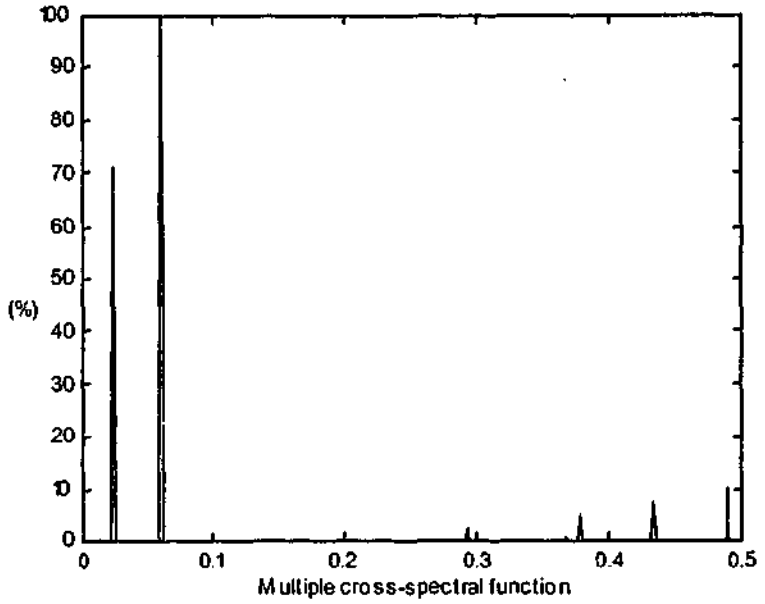
It could be observed from Table 5.1.3.1 that for serine proteases and protease inhibitiors the same characteristic frequencies (within the calculation error) were obtained using both EIIP and IC parameters. This similarity is to be expected, as there is a strong correlation - 0.794 between these two parameters. The analogy of the peak frequencies implies that in this particular frequency the analyzed protein group reveals the same specific biological function. Based on our previous investigations we conclude that the significance of correlation between selected parameters is then reflected into the analogy of the spectra of proteins studied. Furthermore, this analogy implies that we can detect then that particular characteristic frequency which is relevant to common biological behavior of the whole functional protein group. Therefore, the correlation coefficient value of analyzed parameter could be considered as one of the major factors influencing on the selection of the proposed parameter for its further use for structure/function analysis of unrelated proteins within the RRM.

Of interest is the fact that for the rest of the analyzed functional groups: cysteine, metallo, aspartic, trypsin/chymotrypsin proteases and AIDS related protcases, characteristic frequencies are different using the EIIP and IC parameters. Clearly, sequences with the same biological function share the same characteristic periodic component. Thus, the difference in characteristic frequencies recorded using the EIIP and IC parameters could be

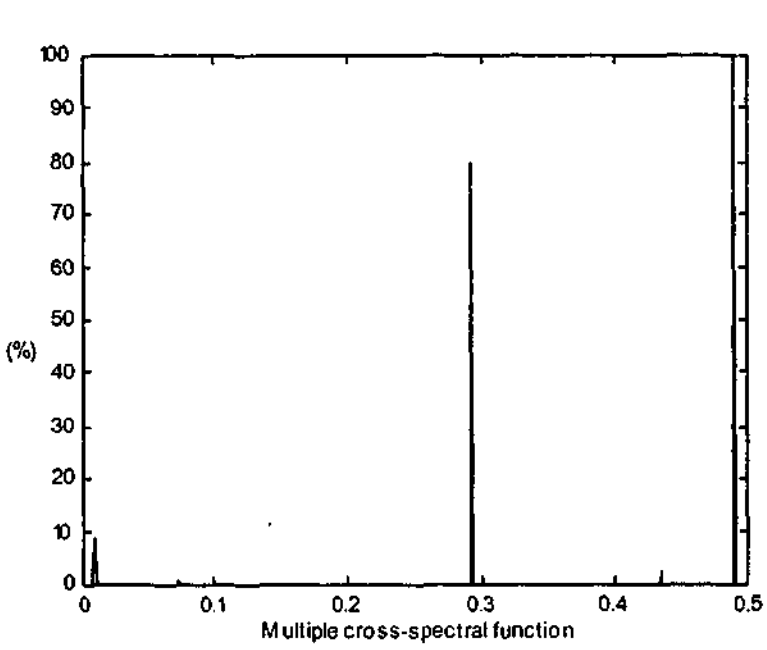
explained by the fact that these frequencies detected correspond to different specific biological function. Despite the diversity of characteristic frequencies obtained for mentioned above protein groups using EIIP and IC parameters, the IC parameter satisfy the RRM criteria C and generate one dominant peak in the consensus spectrum. Therefore we conclude that the introduced IC parameter is suitable for the use in the RRM analysis.

However, as the IC parameter is a measurable physical property, the use of the IC within the RRM might lead to a better understanding of the physical nature of the protein's biological function.

Finally, we have compared results of the RRM analysis of protease proteins using both EIIP and IC parameters. The results obtained have shown that IC parameter can successfully identify common characteristic frequency for all selected protease functional subgroups. However, although in some cases IC is more suitable for structure-function analysis of different protein families it is still difficult to conclude which one of two comparable parameters solely would be the best parameter to use for determination of biological profiles of proteins independent of their biological activity.

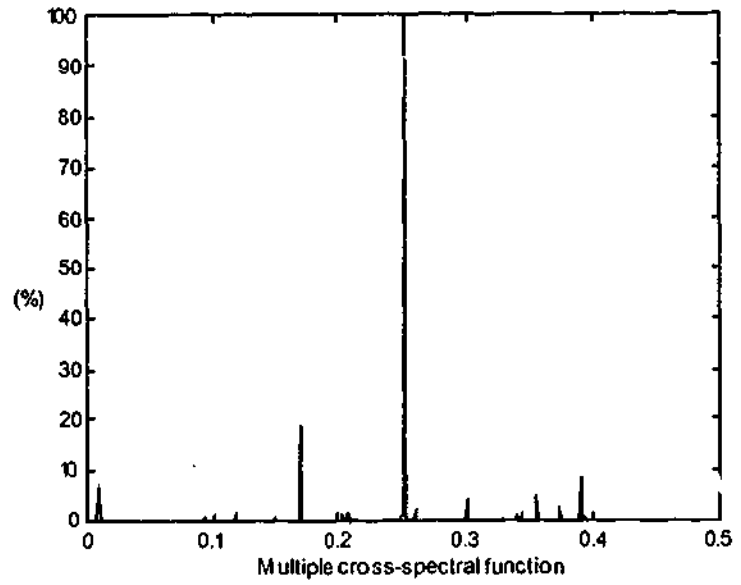


(a)

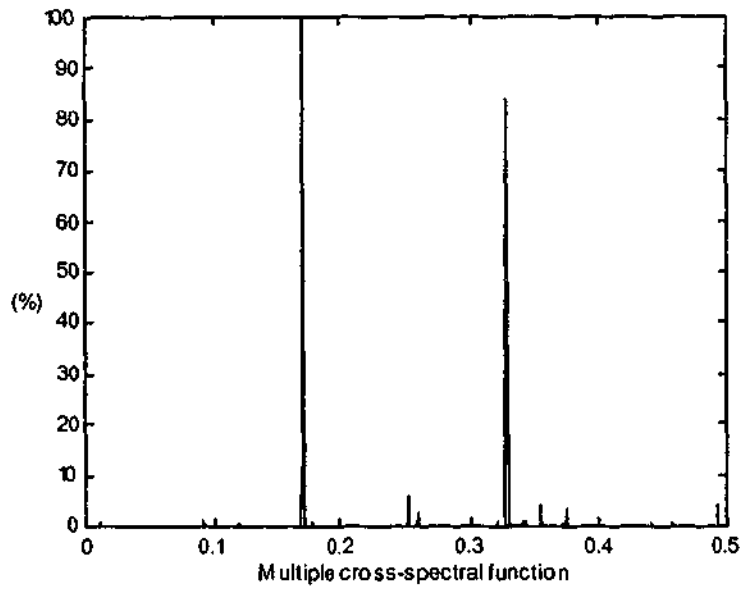


(b)

Figure 5.1.3.1 Multiple cross-spectral function of Cysteine proteases using (a) EIIP and (b) IC parameters.

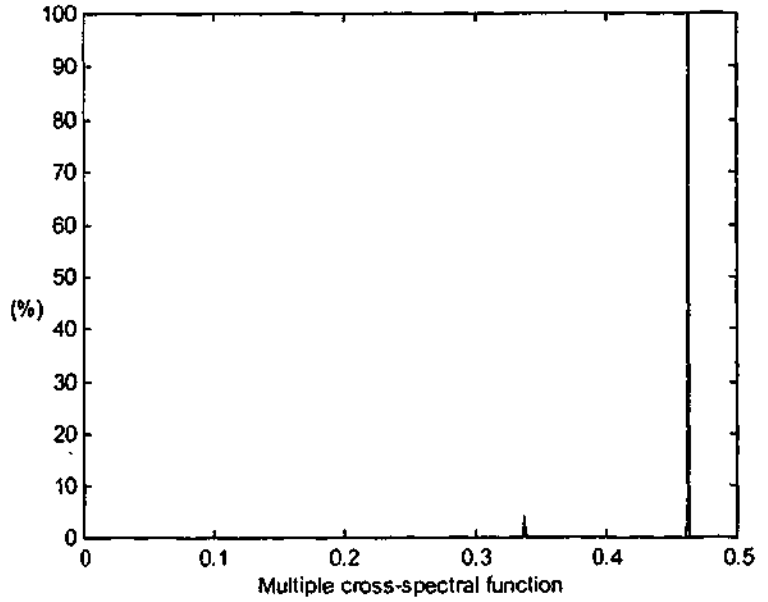


(a)

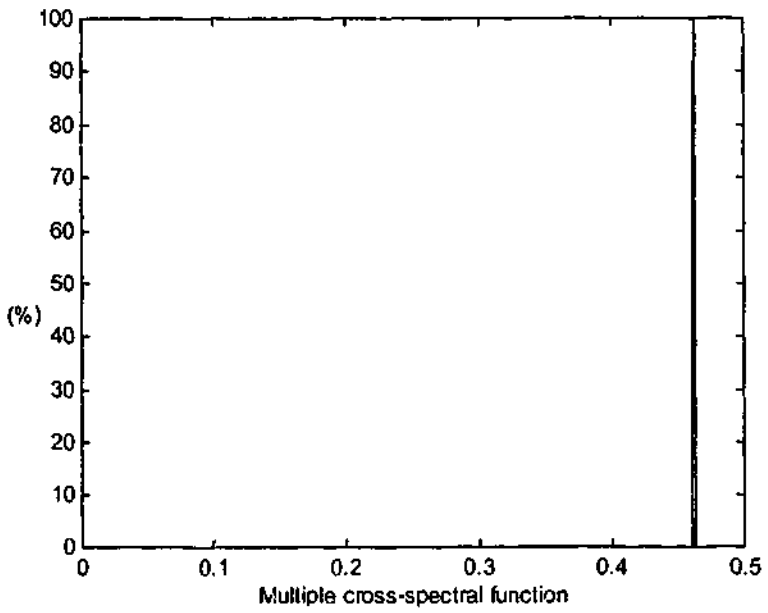


(b)

Figure 5.1.3.2 Multiple cross-spectral function of Metallo-proteases using (a) EIIP and (b) IC parameters.

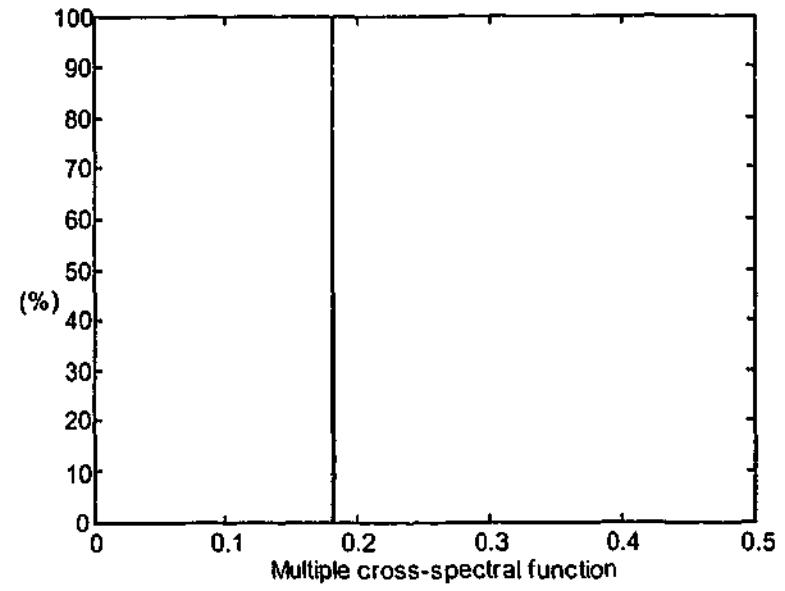


(a)

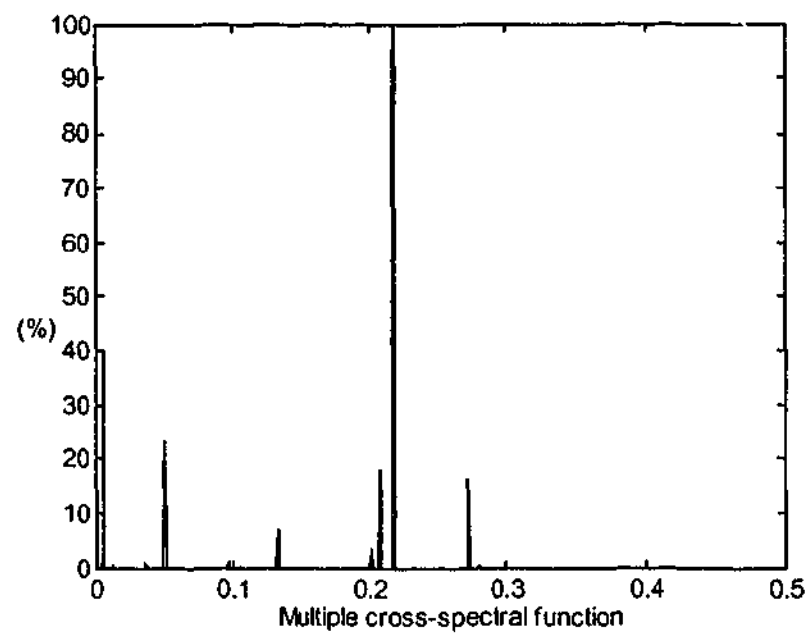


(b)

Figure 5.1.3.3 Multiple cross-spectral function of Serine proteases using (a) EIIP parameter and (b) IC parameters.

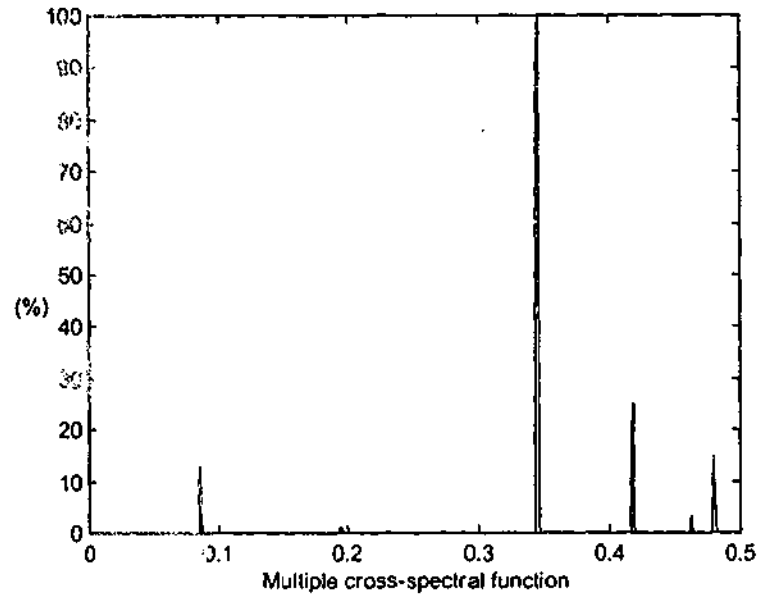


(a)

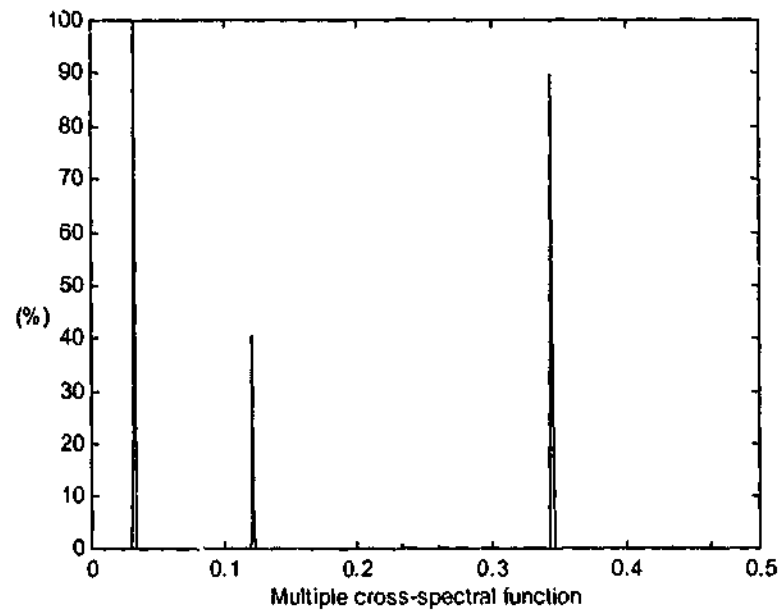


(b)

Figure 5.1.3.4 Multiple cross-spectral function of Aspartic proteases using (a) EIIP and (b) IC parameters.

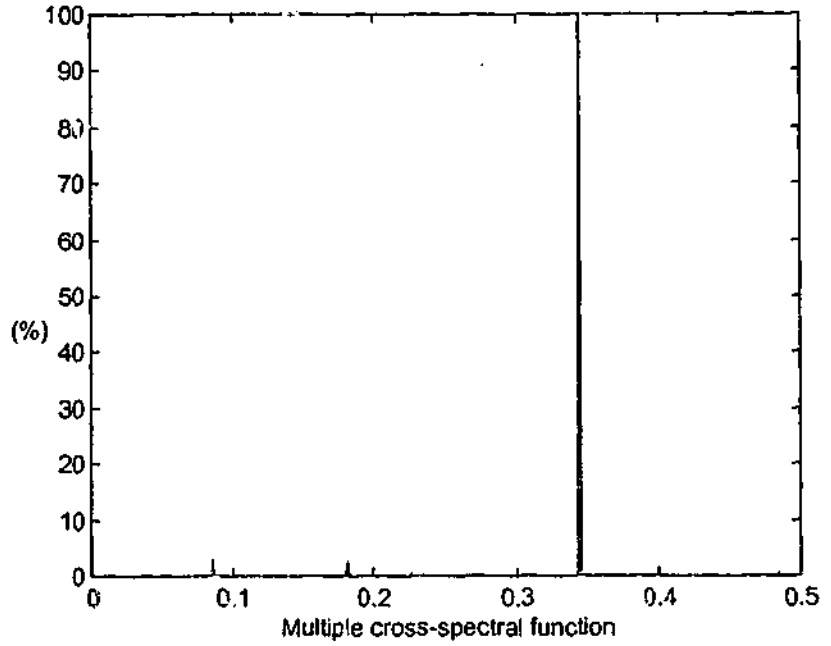


(a)

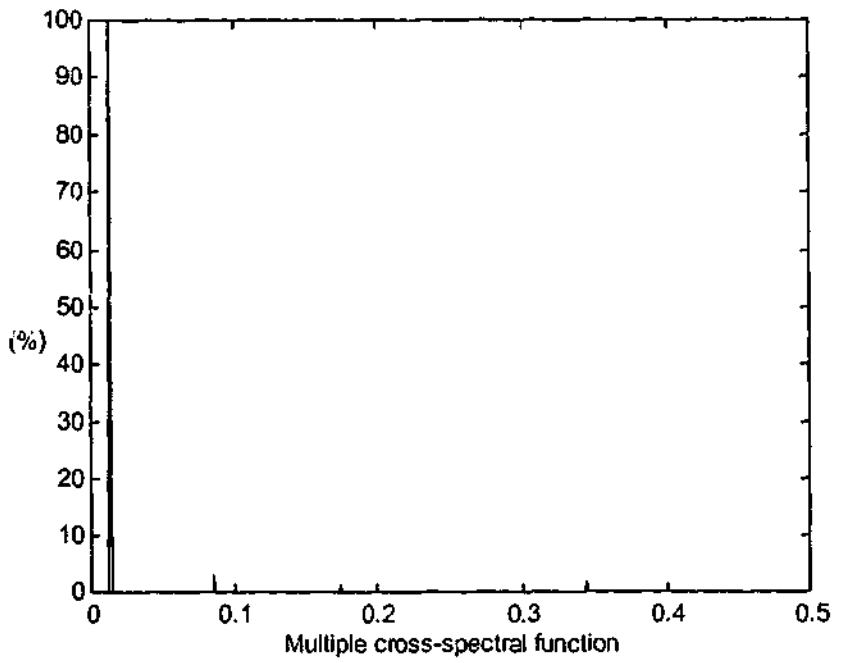


(b)

Figure 5.1.3.5 Multiple cross-spectral function of Trypsin/Chymotrypsin proteases using (a) EIIP and (b) IC parameter.

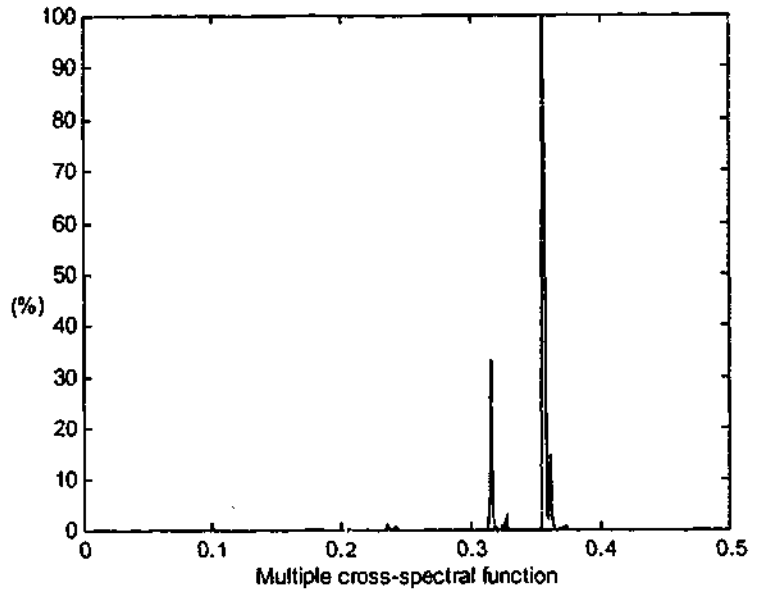


(a)

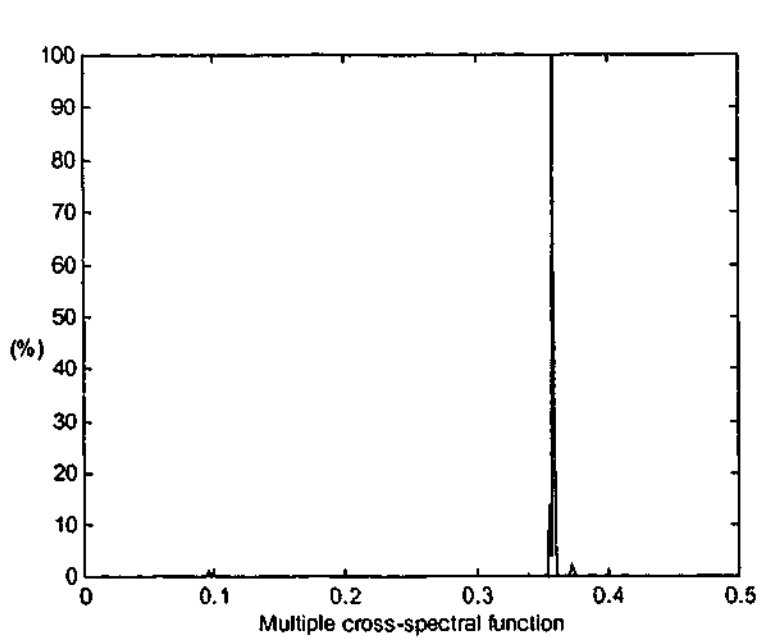


(b)

Figure 5.1.3.6 Multiple cross-spectral function of AIDS related proteins using (a) EIIP and (b) IC parameters.



(a)



(b)

Figure 5.1.3.7 Multiple cross-spectral function of protease Inhibitors using (a) EIIP parameter and (b) IC parameters.

5.1.4 Investigation of the amino acid properties presented in the Amino Acid Index Database for the use in the RRM approach

The RRM has been employed here to investigate the effect of various amino acid properties on the determination of functional patterns of the protein groups under examination.

In this study different amino acid properties have been compared. All of them were obtained from the **Amino Acid Index Database** [53]. The amino acid sequence is considered as a sequence of numerical values reflecting various aspects of amino acid residues, such as: hydrophobicity and bulkiness.

The complete listing of amino acid parameters database used in this study, along with the individual amino acid values and calculated correlation coefficients for each parameter when compared to the EIIP, can be found in Appendix C.

Table 5.1.4.1 Selected Parameter Values and their Correlation Coefficients with EIIP.

Amino Acid	Parameter			
	EIIP	**P001**	**H085**	**H371**
Ala	0.0373	4.35	-0.01	0.937
Arg	0.0959	4.38	0.04	1.725
Asn	0.0036	4.75	0.06	1.080
Asp	0.1263	4.76	0.15	1.640
Cys	0.0829	4.65	0.12	1.004
Gln	0.0761	4.37	0.05	1.078
Glu	0.0058	4.29	0.07	0.679
Gly	0.0050	3.97	0.00	0.901
His	0.0242	4.63	0.08	1.085
Ile	0.0000	3.95	-0.01	0.178
Leu	0.0000	4.17	-0.01	0.808
Lys	0.0371	4.36	0.00	1.254
Met	0.0823	4.52	0.04	0.886
Phe	0.0946	4.66	0.03	0.803
Pro	0.0198	4.44	0.00	0.748
Ser	0.0829	4.50	0.11	1.145
Thr	0.0941	4.35	0.04	1.487
Trp	0.0548	4.70	0.00	0.803
Tyr	0.0516	4.60	0.03	1.227
Val	0.0057	3.95	0.01	0.625
Correl. Coeffic. R		**0.5542**	**0.5643**	**0.6654**

There are six main groups of parameters which are categorized by their first letter identifier: (A) alpha and turn propensities, (B) beta propensity, (H) hydrophobicity, (C) composition, (P) physicochemical properties and (O) other properties.

Once the database was established, correlation calculations were carried out sequentially between each of the 402 parameter and the referential EIIP values in order to test for any linear relationship between them. Parameters having correlation coefficients in excess of +/- 0.5 were deemed to be the most strongly correlated with EIIP and were thus isolated [74].

These selected parameters are **P001-α-CH chemical shifts**, **H085-Localised electrical effect** and **H371- Normalized frequency of chain reversal D** [53]. The selected parameter values and their correlation coefficients with EIIP are shown in Table 5.1.4.1.

Sequences from six protein functional groups: glucagons, lysozymes, hemoglobins, myoglobins, and cytochromes C and EGFs were used as test examples. In this work each amino acid in the sequence was represented by corresponding value of selected parameter instead of the EIIP value as it was done previously [6,57-60,67-69]. Numerical series obtained in this way were analyzed and a multiple cross-spectral analysis was performed for each protein group. Peak frequency and S/N values obtained are presented in Table 5.1.4.2.

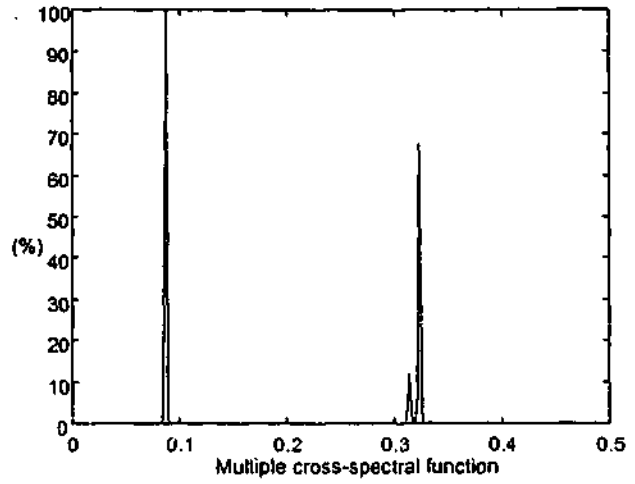
Table 5.1.4.2 Peak Frequency and Signal-to-Noise Values for Protein Group using Selected Parameters.

Parameter	**EIIP**		**P001**		**H085**		**H371**	
	Freq.	**S/N**	**Freq.**	**S/N**	**Freq.**	**S/N**	**Freq.**	**S/N**
Glucagon	0.0879	122.7	0.1364	166.3	0.1523	140.1	0.2520	95.7
Hemoglobin	0.0254	256.0	0.1367	256.0	0.4102	249.1	0.3926	256.0
Lysozyme			0.4707	120.5	0.0742	167.3		
Myoglobin			0.0254	255.2	0.1270	256.0		
Cytochrome C					0.4219	232.3	0.2012	150.3
EGF								

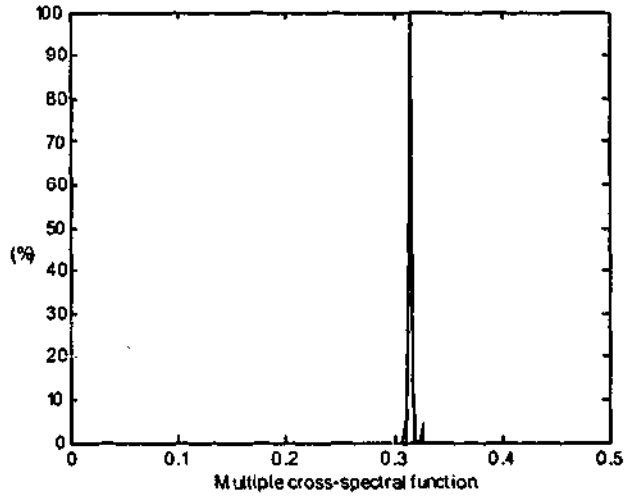
The assumption used here is that sequences with the same biological function share the same characteristic periodic component in the distribution of energy of delocalised electrons. Each specific biological function is characterized by a single frequency. A

multiple cross-spectral analysis was performed for each analyzed protein group, using new parameter values.

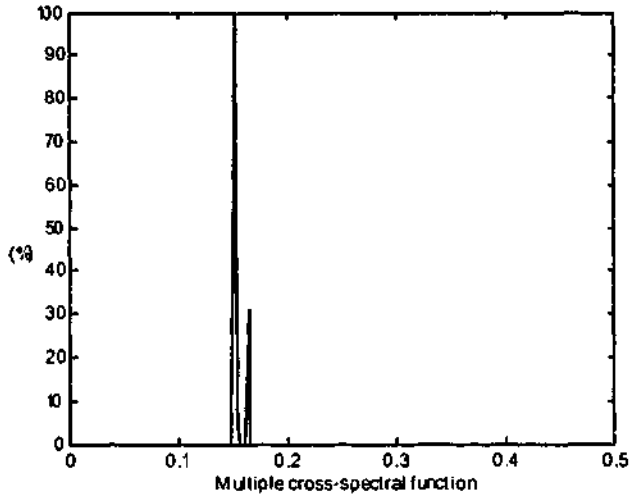
It could be observed from Table 5.1.4.2 that for Cytochrome C proteins the same characteristic frequencies were obtained using EIIP and P001 parameters, for which the values of the amino acid indexes are quite different (Table 5.1.4.1). The analogous results were found for Lysozyme and Myoglobin protein groups when we have obtained the same characteristic frequencies using EIIP and H085 parameters. For EGF proteins we have detected the same characteristic frequency (within the calculation error) or characteristic region using EIIP, P001, H085 and H371 parameters. The analogy of the peak frequencies in consensus spectrum implies that in these particular frequencies the analyzed protein groups reveal the same specific biological function. Furthermore, all these analyzed parameters are strongly correlated with EIIP that is indicated by the correlation coefficient values (Table 5.1.4.1). This correlation is reflected then into the similarity between spectra of proteins. In addition it is interest to note that for the rest analyzed functional groups characteristic frequencies are different. It is expected because, as it was mentioned above in the RRM criteria, the frequencies are different for different biological functions.



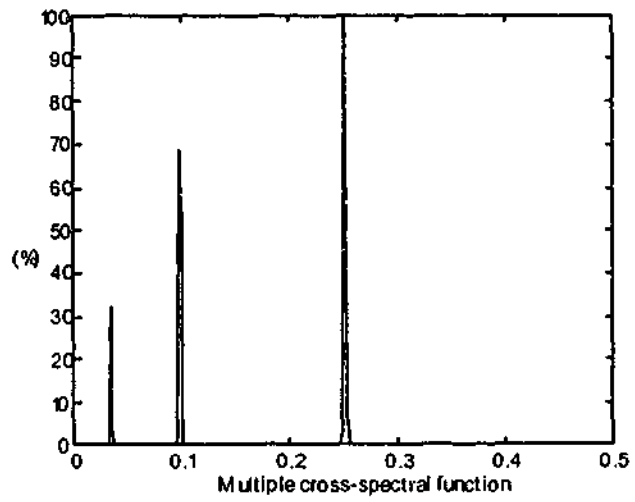
(a)



(b)

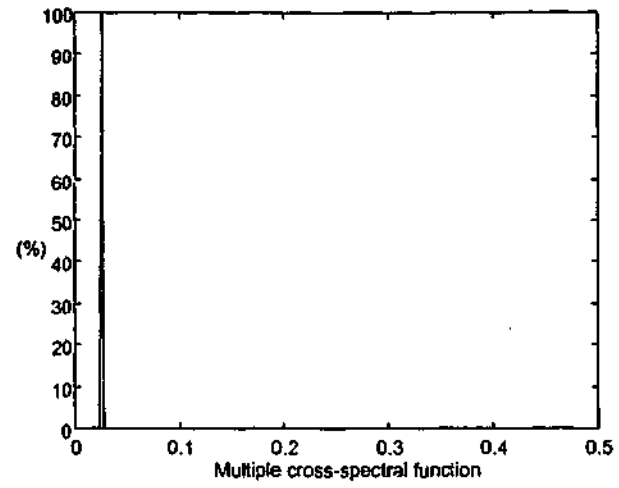


(c)

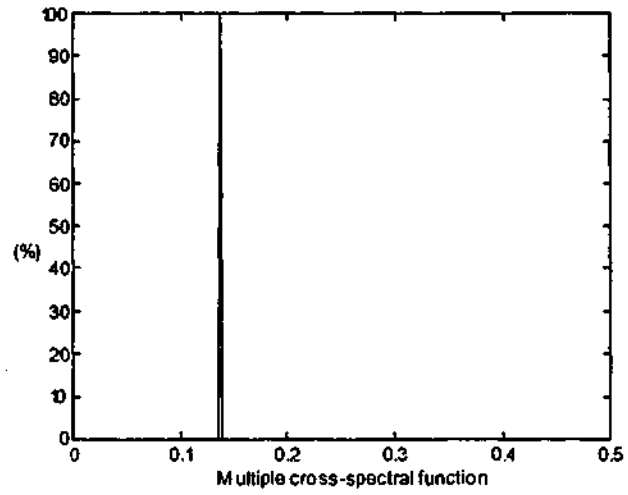


(d)

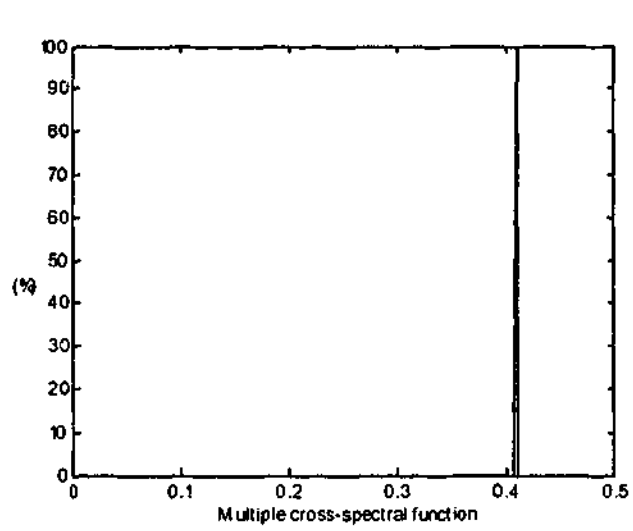
Figure 5.1.4.1 Multiple cross-spectral function of Glucagon proteins using parameters: (a) EIIP, (b) P001, (c) H085 and (d) H371.



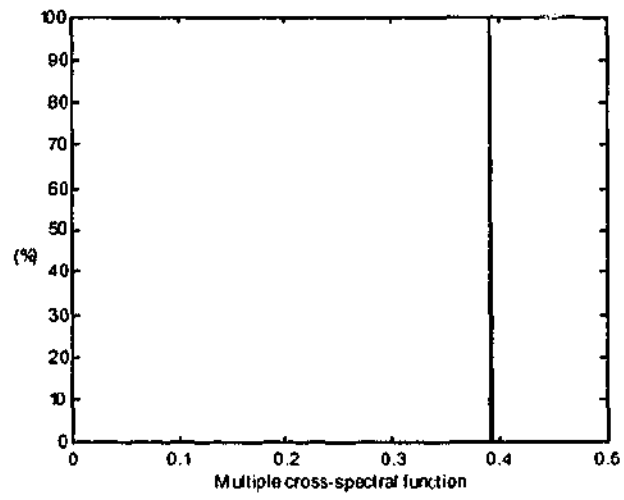
(a)



(b)

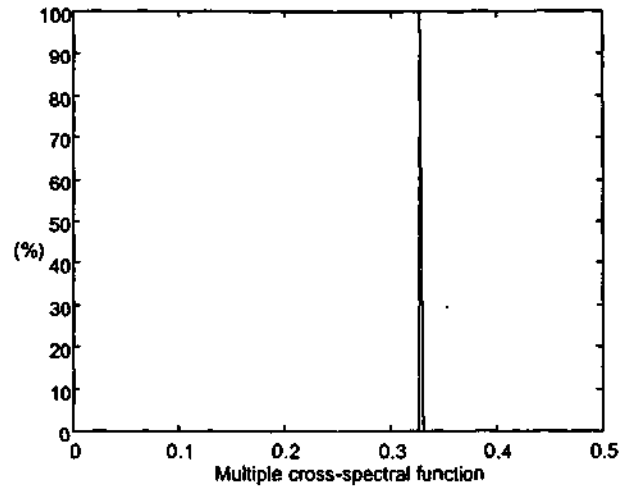


(c)

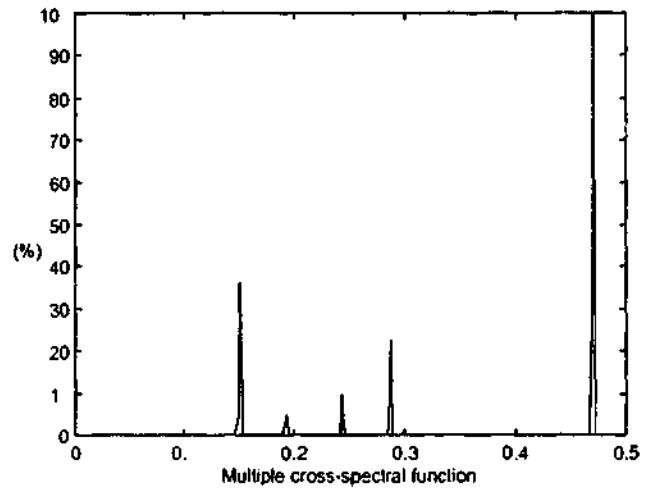


(d)

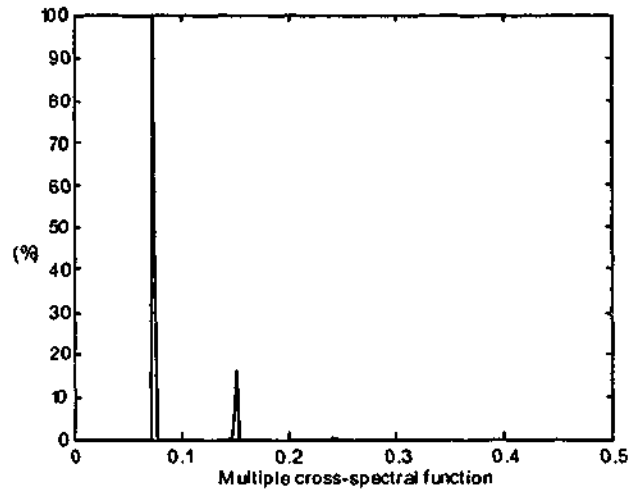
Figure 5.1.4.2 Multiple cross-spectral function of Hemoglobin using parameters: (a) EIIP, (b) P001, (c) H085 and (d) H371.



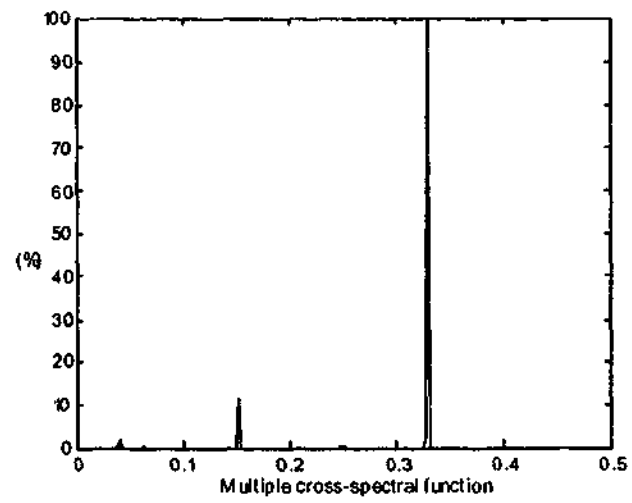
(a)



(b)

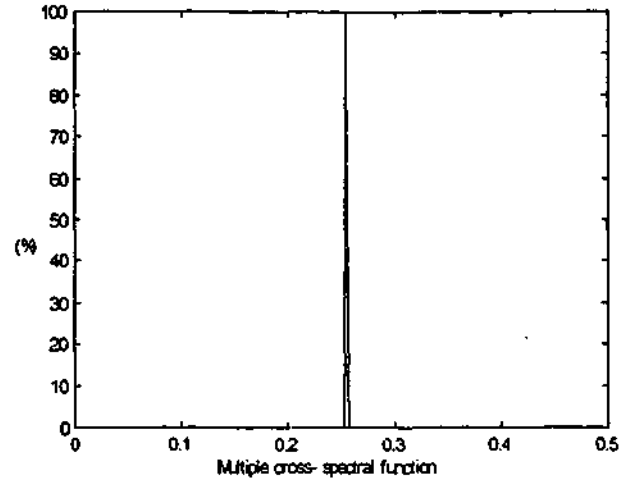


(c)

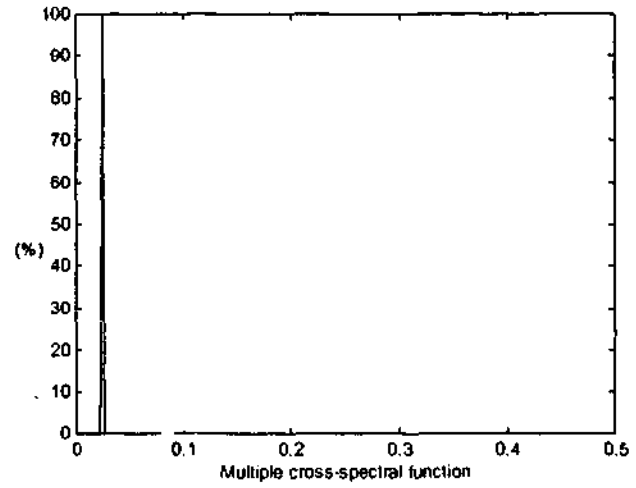


(d)

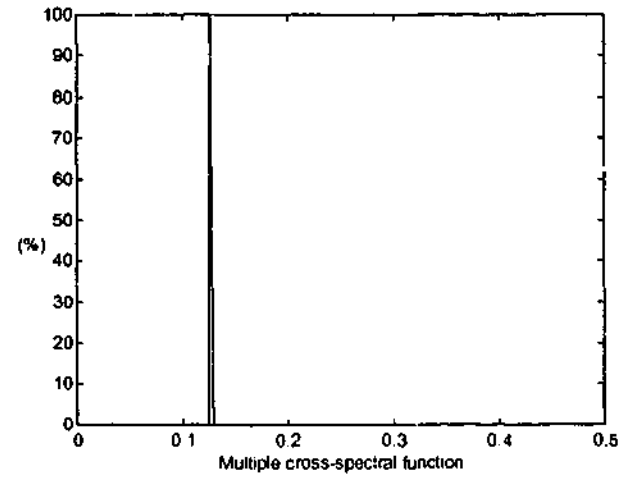
Figure 5.1.4.3 Multiple cross-spectral function of Lysozyme using parameters: (a) EIIP, (b) P001, (c) H085 and (d) H371.



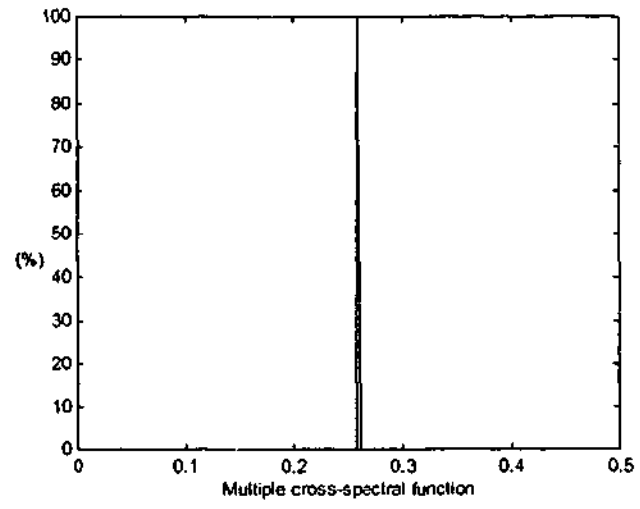
(a)



(b)

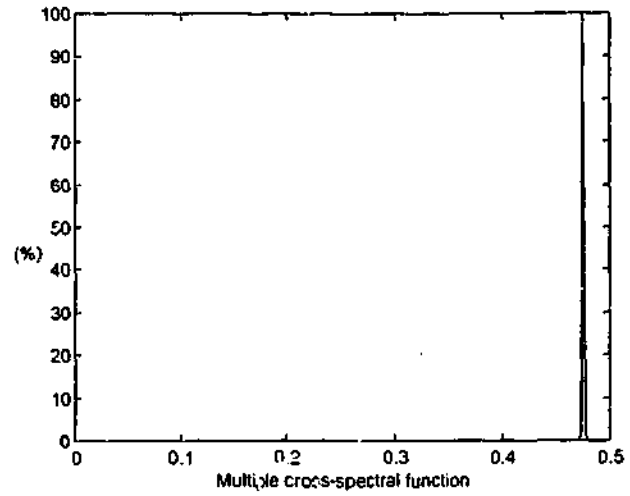


(c)

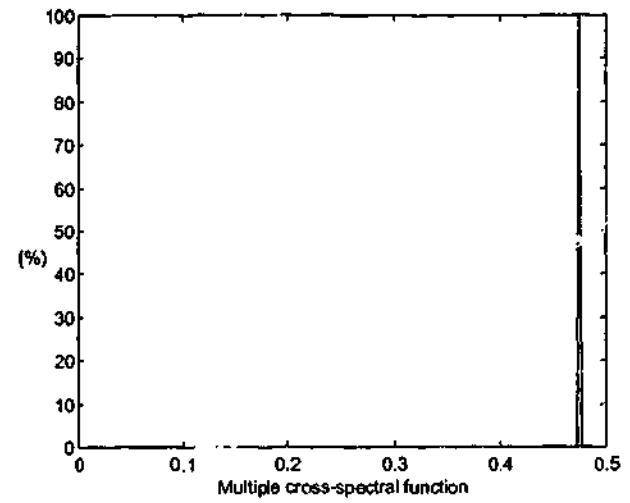


(d)

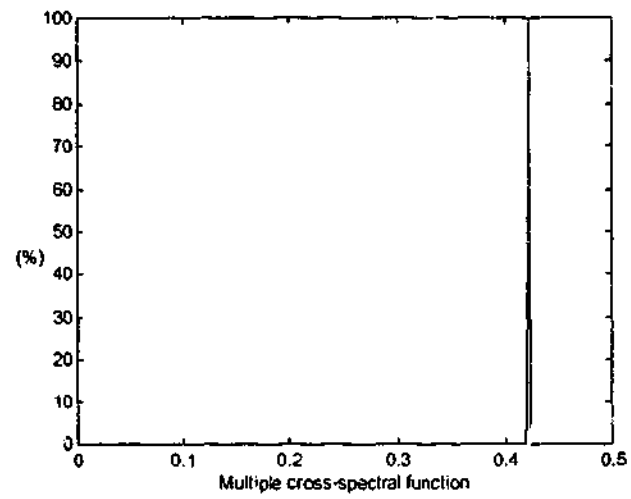
Figure 5.1.4.4 Multiple cross-spectral function of Myoglobin proteins using parameters: (a) EIIP, (b) P001, (c) H085 and (d) H371.



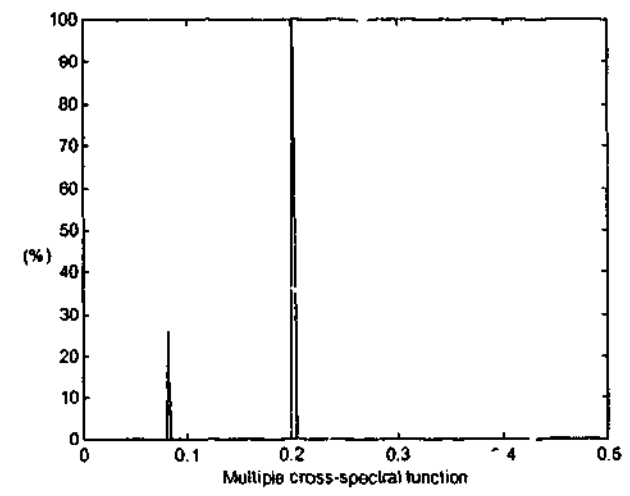
(a)



(b)

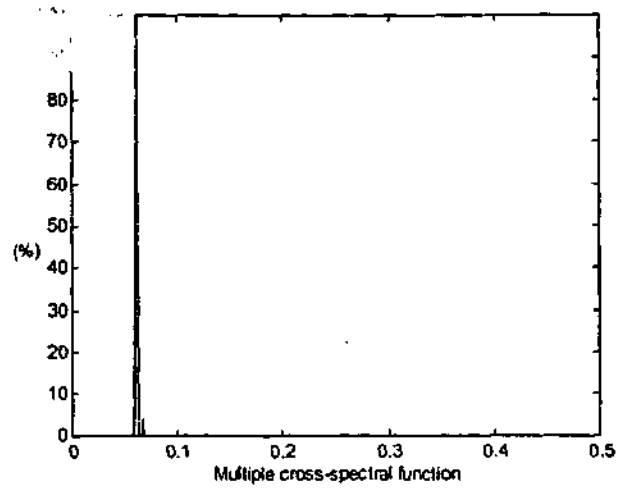


(c)

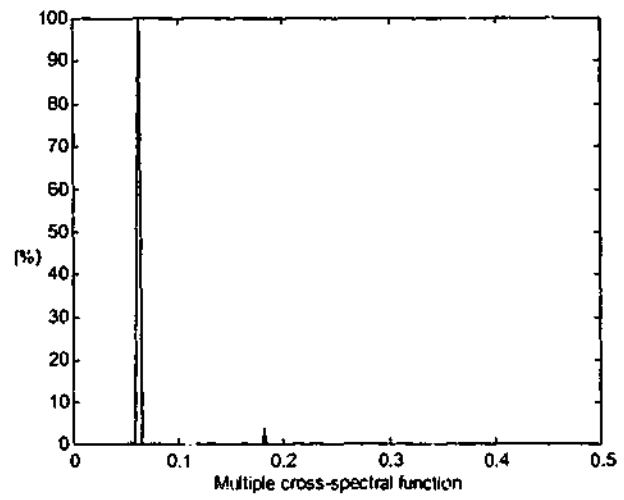


(d)

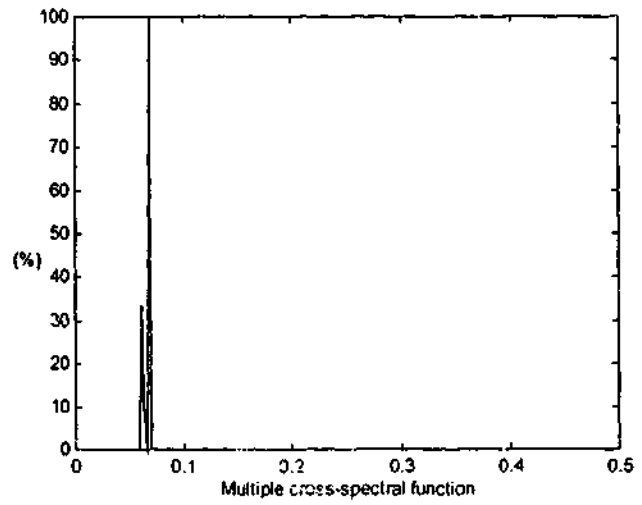
Figure 5.1.4.5 Multiple cross-spectral function of Cytochrome C proteins using parameters: (a) EllP, (b) P001, (c) H085 and (d) H371.



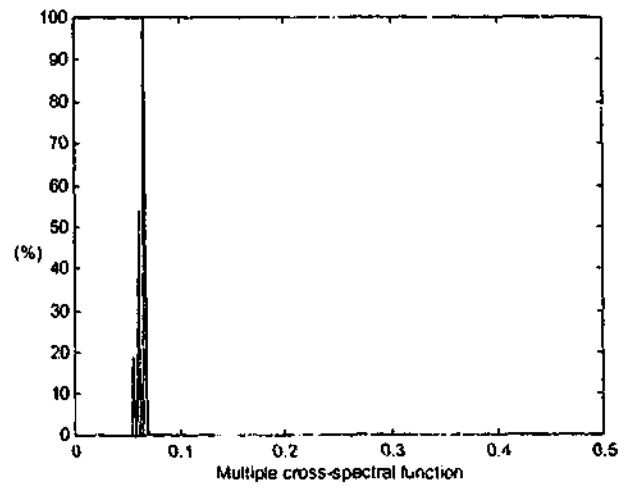
(a)



(b)



(c)



(d)

Figure 5.1.4.6 Multiple cross-spectral function of EGF using parameters: (a) EIIP, (b) P001, (c) H085 and (d) H371.

The aim of this study was to select those parameters among the presented database that are most suitable to represent the common biological behavior of the studied protein groups. Previously EIIP has been successfully used for the protein analysis within the RRM [57-60,67-69]. Results obtained have shown that all three parameters P001, H085 and H371 generate in consensus spectrum one dominant peak corresponding to common biological activity of selected protein groups and are performed according to the required RRM criteria (Fig.5.1.4.1-5.1.4.6). Thus, results reveal that P001, H085 and H371 parameters are suitable for the use within the RRM for structure-function analysis of proteins.

5.1.5 Development of new computational amino acid parameters for the use in protein structure-function analysis within the RRM

In this investigation the RRM was used to determine the effect of different amino acid parameters and their linear combinations on the periodicity (characteristic frequency) of certain protein groups. Nine functionally unrelated protein groups were used as test examples. Here, amino acid parameters are compared. Most of them were obtained, as previously, from the Amino Acid Index Database [53]. For this particular study we have chosen parameters strongly correlated with EIIP and recommended for the use based on the results of previous RRM analysis.

The comparable parameter database includes:

EIIP Electron-ion interaction potential [54,55]
A Helix coil equilibrium constant [49]
H Normalised frequency of chain reversal D [53]
IC Ionization constant [72,73]
and their different mathematical combinations.

Table 5.1.5.1 EIIP, Selected Parameter Values and their Correlation Coefficients

amino acid	EIIP	IC	A	H	IC+H	A+H	ICxH	AxH	AxHxIC
L	0.0000	2.40	2.2	0.808	3.208	3.008	1.939	1.778	4.267
I	0.0000	2.40	0.4	0.178	2.578	0.578	0.427	0.071	0.17
N	0.0036	2.20	2.6	1.080	3.28	3.68	2.376	2.808	6.178
G	0.0050	2.46	1.0	0.901	3.361	1.901	2.216	0.901	2.216
V	0.0057	2.35	0.8	0.625	2.975	1.425	1.469	0.500	1.175
E	0.0058	2.30	2.2	0.679	2.979	2.879	1.562	1.494	3.436
P	0.0198	2.00	0.0	0.748	2.748	0.748	1.496	0.000	0
H	0.0242	2.30	2.5	1.085	3.385	3.585	2.496	2.713	6.24
K	0.0371	2.20	2.2	1.254	3.454	3.454	2.759	2.759	6.07
A	0.0373	2.30	3.0	0.937	3.237	3.937	2.155	2.811	6.465
Y	0.0516	2.20	2.5	1.227	3.427	3.727	2.699	3.068	6.75
W	0.0548	2.37	2.5	0.803	3.173	3.303	1.903	2.008	4.759
Q	0.0761	2.06	2.7	1.078	3.138	3.778	2.221	2.911	5.997
M	0.0823	2.17	2.2	0.886	3.056	3.086	1.923	1.949	4.229
S	0.0829	2.10	3.7	1.145	3.245	4.845	2.405	4.237	8.898
C	0.0829	1.96	2.6	1.004	2.964	3.604	1.968	2.610	5.116
T	0.0941	2.09	2.7	1.487	3.577	4.187	3.108	4.015	8.391
F	0.0946	1.98	2.5	0.803	2.783	3.303	1.59	2.008	3.976
R	0.0959	1.82	2.7	1.725	3.545	4.425	3.14	4.658	8.478
D	0.1263	1.88	2.6	1.640	3.52	4.24	3.083	4.264	8.016
[redacted]					**0.328**				

The correlation coefficients between EIIP and selected parameters values have been calculated and presented in Table 5.1.5.1.

In this work each amino acid in the protein sequence is represented by the corresponding value of each selected parameter instead of the EIIP value as it was done previously [6]. Numerical series obtained in this way were analysed as previously by transforming them into frequency domain using DFT in order to extract information pertinent to the biological function. Peak frequencies in the multiple cross-spectral function denote common frequency components for the protein sequences analysed.

Sequences from different functional groups (glucagons, lysozymes, haemoglobins, cytochromes C, EGFs, FGFs, TNFs, interleukins and myoglobins) were investigated and a multiple cross-spectral analysis was performed for each group, using the EIIP and other investigated parameters values. The peak frequency and signal-to-noise values for each selected protein group are shown in Table 5.1.5.2.

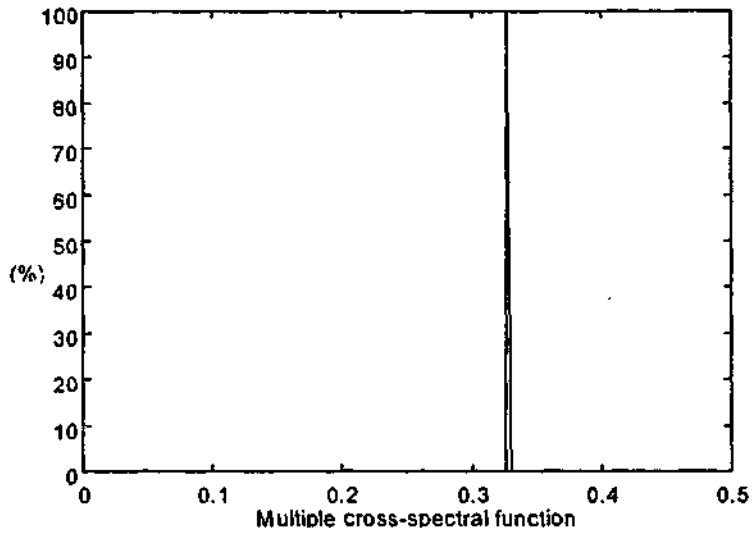
Table 5.1.5.2 Peak Frequency and Signal-to-Noise Values for Protein Groups.

Protein Group	EIIP	IC	A	H	IC+H	A+H	ICxH	AxH	AxHxIC
Cytochrome C			0.0254 104.1	0.2012 150.3	0.2012 245.6	0.2012 129.3	0.2012 202.1	0.2031 231.5	0.2031 226.7
Lysozyme			0.2422 97.2		0.4980 196.7	0.2422 106.1	0.0410 167.9		
Hemoglobin			0.4551 249.2	0.3926 256.0	0.3145 256.0	0.3926 231.2	0.4258 255.9	0.3926 255.9	0.3926 253.1
Glucagon			0.3730 106.1	0.2520 95.7	0.3730 82.9	0.1016 91.4	0.2520 224.0	0.0996 151.9	0.0996 119.0
TNF	0.0527 119.0		0.1836 92.0	0.0332 120.7	0.3086 28.1	0.2422 115.0	0.0332 84.9		
EGF									
FGF			0.4785 117.2		0.4570 118.9	0.4668 99.2			
Interleukin	0.0410 191.5	0.0957 192.2	0.0312 121.3	0.1133 169.4	0.3613 124.4	0.0566 86.2	0.3613 121.5	0.0312 76.0	0.4551 92.6
Myoglobin		0.0137 191.3	0.0820 239.1			0.0820 110.7			

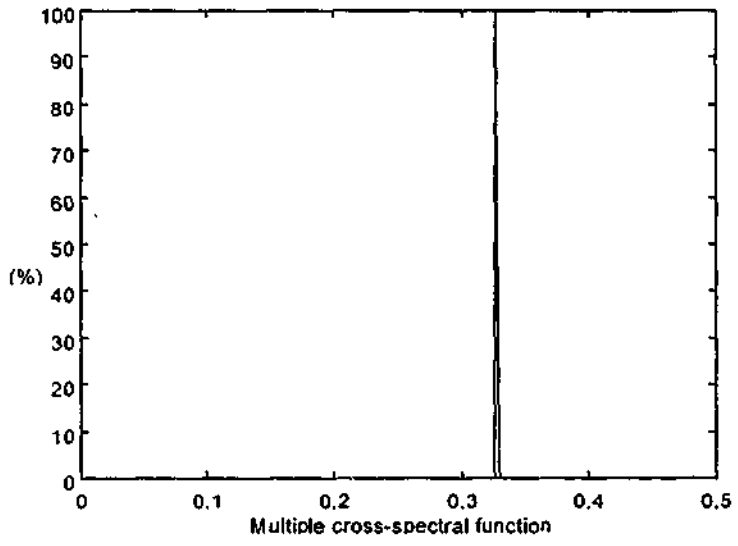
It could be observed from Table 5.1.5.2 that for Lysozyme the same characteristic frequencies were obtained using the EIIP, IC, H, AxH and AxHxIC parameters. For TNF protein group the same characteristic frequencies (within the calculation error) were obtained using the IC, AxH and AxHxIC parameters. For EGF proteins the same characteristic frequencies (within the calculation error) were obtained using all selected

parameters. For FGF the same characteristic frequencies obtained using the EIIP, IC, H, IcxH, AxH and AxHxIC. Analoqous results were obtained for Myoglobin using the EIIP, H, IC+H, IcxH, AxH and AxHxIC (Table 5.1.5.2). The similarities of characteristic frequencies obtained using mentioned above parameters are expected, as there is significant correlation [74] between these parameters (Table 5.1.5.1) and the EIIP. There is the strongest correlation (0.716) between the EIIP and the (AxH) parameters. The consensus spectrum of Lysozyme proteins using the EIIP, IC and (AxH) parameters are shown in Fig.5.1.5.1.

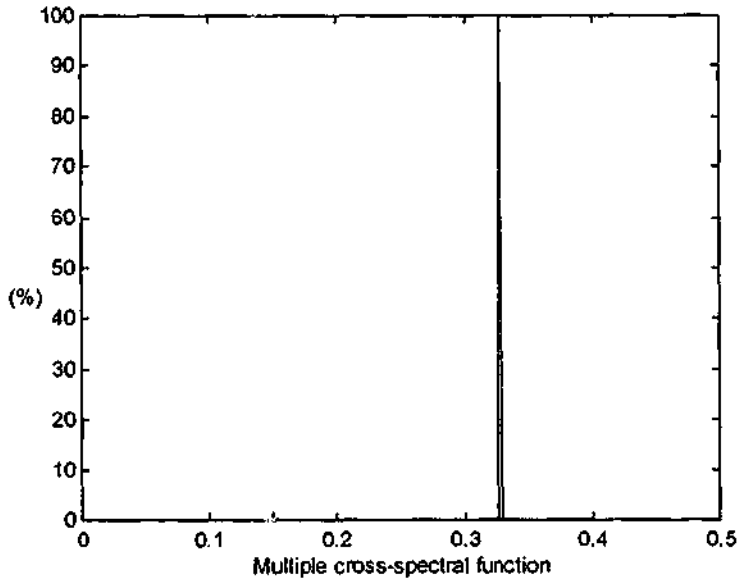
It is interested to note that for TNF protein group the main frequency is different when the EIIP is used rather than IC or (AxH) (Fig.5.1.5.2). From the figure 5.1.5.2 it could be observed that TNFs do not have only one but more significant common frequencies indicating their multifunctional activity. Thus, when using the EIIP one frequency f=0.0527 is more prominent than the other two while when using IC and (AxH) parameters the other frequency f=0.0762 is more prominent. Importantly, with all three pointed out parameters all three characteristic frequencies are significant, according to the S/N values [5,6].



(a)

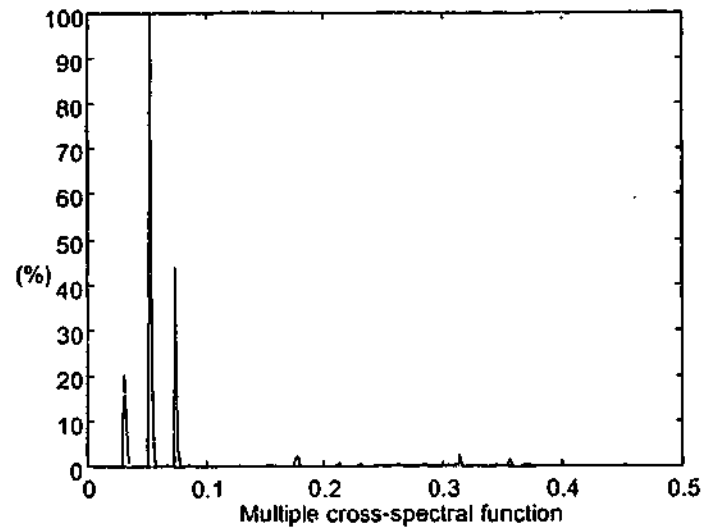


(b)

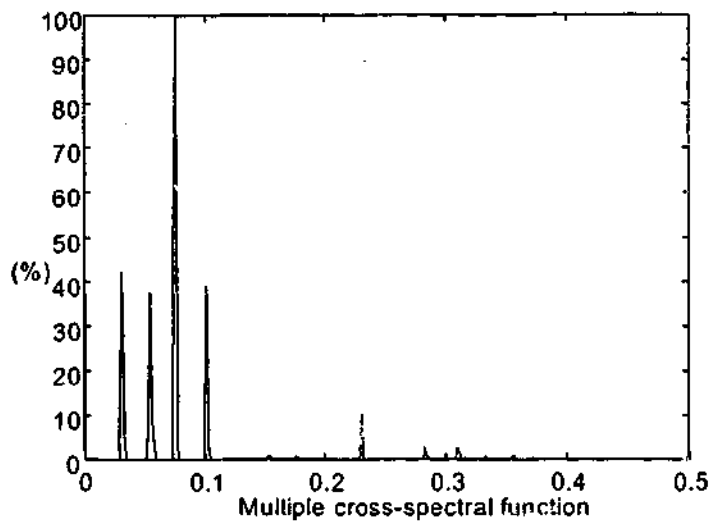


(c)

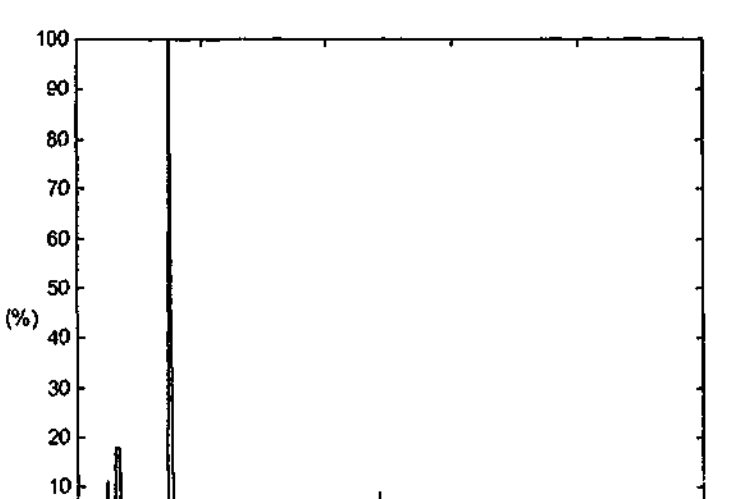
Figure 5.1.5.1 Multiple cross-spectral function of Lysozyme proteins using (a) EIIP; (b) IC and (c) AxH parameters.



(a)



(b)



(c)

Figure 5.1.5.2 Multiple cross-spectral function of TNF proteins using (a) EIIP, (b) IC and (c) AxH parameters.

The correlation coefficient between the EIIP and (AxH) parameters is 0.716 (Table 5.1.5.1). That indicates a strong correlation (> |0.5|) between these parameters [74]. The strong correlattion between the EIIP and (AxH) parameters is reflected into the analogy between spectra of the studied proteins. Results obtained here reveal that the proposed computational parameter (AxH) is also suitable for the determination of the characteristic frequencies of proteins in the RRM model. As long as it generates in the consensus spectrum one dominant peak corresponding to the common biological behavior of the studied proteins and satisfies the RRM criteria. However, after comparison of results obtained within the RRM analysis using the IC and (AxH) parameters, we conclude that the IC parameter is more appropriate for this approach. To decide finally which parameter is optimal for the use in the RRM further analysis is necessary.

Following the aim of searching for the best computational amino acid parameter for the use in the RRM, and based on results obtained from previous studied, we have continued to analyse the P001, H085 and H371 amino acid parameters extracted from the Amino Acid Index Database.

Table 5.1.5.3 Analyzed parameter values and their correlation coefficients with the EIIP.

	Parameter			
Amino Acid	**EIIP**	**IC**	**EE**	**ICEE**
L	0.0000	2.40	-0.01	2.41
I	0.0000	2.40	-0.01	2.41
N	0.0036	2.20	0.06	2.14
G	0.0050	2.46	0.00	2.46
V	0.0057	2.35	0.01	2.34
E	0.0058	2.30	0.05	2.23
P	0.0198	2.00	0.00	2.00
H	0.0242	2.30	0.08	2.22
K	0.0371	2.20	0.00	2.20
A	0.0373	2.30	-0.01	2.31
Y	0.0516	2.20	0.03	2.17
W	0.0548	2.37	0.00	2.37
Q	0.0761	2.06	0.07	2.01
M	0.0823	2.17	0.04	2.13
S	0.0829	2.10	0.11	1.99
C	0.0829	1.96	0.12	1.84
T	0.0941	2.09	0.04	2.05
F	0.0946	1.98	0.03	1.95
R	0.0959	1.82	0.04	1.78
D	0.1263	1.88	0.15	1.73
Cor. Coef.				

As we concluded in the previous study these parameters could be used for protein structure/function analysis within the RRM, as long as they satisfied all RRM criteria. However, these three selected parameters are not equally efficient accordingly to correlation coefficient and S/N values. Thus, we determined that P001, H085 and H371 parameters wouldn't be solely the best parameters to use in further protein modeling within the RRM.

Hence, for further improvement of parameters suitable for its application in the RRM approach, we have tried here different mathematical combinations of IC, P001, H085 and H371 parameters with the aim to find among them the most correlated with the EIIP parameter. This new parameter should be more significantly correlated with the EIIP than selected above parameters.

As a result of our calculations, we are proposing here a new computational parameter ICEE. This parameter is a simple mathematical combination of two previously selected parameters: IC (Ionization constant of amino acid) [72,73] and H085 (Localized electrical effect) [53]. We have chosen this parameter because of its strong correlation with the EIIP. The correlation coefficient is **-0.807** (Table 5.1.5.3).

Here, we have analyzed and compared the following amino acid parameters: EIIP, IC, EE (EE=H085 used in previous study, here renamed for further convenience) and computational parameter ICEE (ICEE=IC-EE). The selected parameter values are presented in Table 5.1.5.3.

The RRM was employed here for the proteins' analysis using proposed amino acid parameters. Sequences form different functional groups (glucagon, lysozyme, hemoglobin, cytochrome C, EGF, TGF, PDG, TNF, interleukin, myoglobin, viral oncogen, proto-oncogene, oncogene, and p53 protein) were investigated using EIIP, IC, EE and ICEE parameters. A multiple cross-spectral analysis was performed for each protein group using selected parameter values.

As a result protein characteristic frequencies for each protein functional group were obtained. The peak frequency and S/N values for each protein group are shown in Table 5.1.5.4.

Table 5.1.5.4 Peak frequency and signal-to-noise values for protein groups.

Protein group	Parameter							
	EIIP		IC		EE		ICEE	
					Freq.	S/N	Freq.	S/N
Glucagon					0.1523	140.1	0.3340	91.8
Lysozyme					0.0742	167.3		
Hemoglobin					0.4102	249.1		
Cytochrome C					0.4219	232.3		
EGF								
TGF								
PDG	0.4199	75.0			0.1797	21.5		
TNF	0.0527	119.0			0.1445	38.5		
Interleukin	0.0410	191.0			0.1016	236.5		
Myoglobin	0.2539	255.4	0.0137	191.3	0.1270	256.0	0.4824	186.5
Viral oncogene			0.4844	368.0			0.3115	236.5
Proto-oncogene	0.0576	219.0			0.1641	222.7		
Oncogene	0.0332	505.0			0.1641	347.6		
P53	0.1943	156.0			0.0381	289.1		

Results obtained have shown that all selected for analysis IC, EE and ICEE parameters generate in consensus spectrum one dominant peak corresponding to common biological activity of selected proteins and are performed accordingly to criteria determined within the RRM.

It could be observed from Table 5.1.5.4 for Lysozyme, Hemoglobin, Cytochrome C, EGF, TGF, PDG, TNF, Interleukin, Proto-oncogene, Oncogene and P53 proteins the same characteristic frequencies were obtained using parameters IC and ICEE. This similarity is expected, as both the IC and ICEE parameters are mutually strongly correlated (r=0.9837), that is reflected in the analogy of the consensus spectrums.

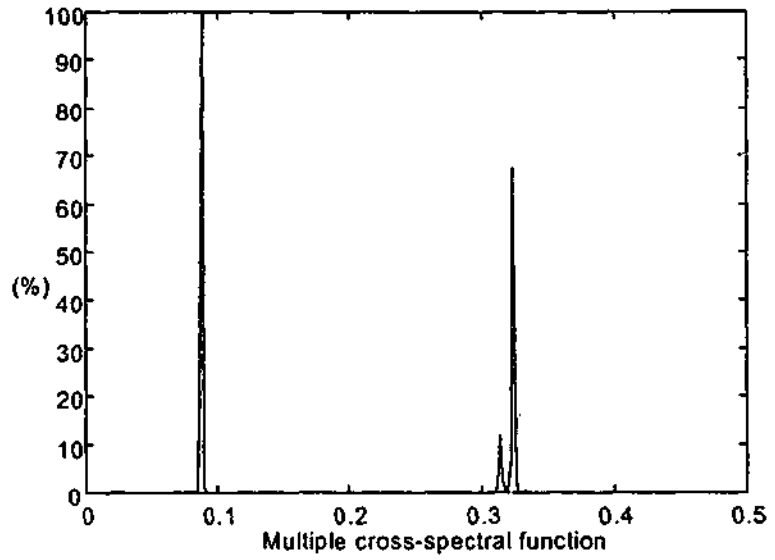
It should be noted that for Lysozyme, Hemoglobin, Cytochtome C, EGF and TGF the same characteristic frequencies were obtained using parameters EIIP, IC and ICEE (Fig.5.1.5.4-5.1.5.8). For EGF and TGF protein groups the same characteristic frequencies were obtained using all analyzed EIIP, IC, EE and ICEE parameters. The analogous result was found for Viral oncogene when we have obtained the similarity of the characteristic frequencies using the EIIP and EE parameters. These findings could be explained also by the significant correlation between the EIIP, IC and ICEE parameters (Table 5.1.5.3). This strong correlation is reflected into the analogy of the consensus spectrums.

It could be also observed from Table 5.1.5.4 that characteristic frequencies are different for the other protein groups. The following fundamental conclusion was drawn from the RRM

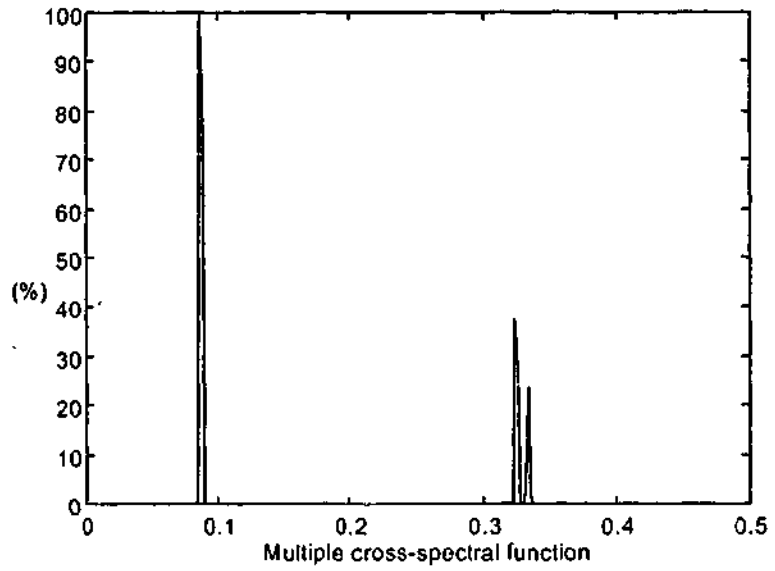
model: each specific biological function is characterised by a single frequency. Each RRM frequency characterises one biological function. The difference in the characteristic frequencies implies that we have detected different specific biological functions of the proteins studied using the analyzed parameters (satisfaction to RRM criteria (c), page 41). Thus, each of analyzed EIIP, IC, EE and ICEE parameters allows us to detect a single frequency relevant to the specific biological function of the studied protein sequences. That is a main criterion for the selection of the analyzed parameter for its further use within the RRM approach.

The aim of this study was to test the possible use of the proposed parameters IC, EE and ICEE instead of the EIIP for structure/function analysis of different proteins within the RRM. Results obtained reveal that among selected and comparable amino acid parameters only IC and ICEE are currently the most appropriate parameters for RRM analysis of different unrelated protein families. We can conclude this as far as these two parameters satisfy all RRM criteria, generate one prominent peak corresponding for biological activity of whole protein group and are highly correlated with EIIP.

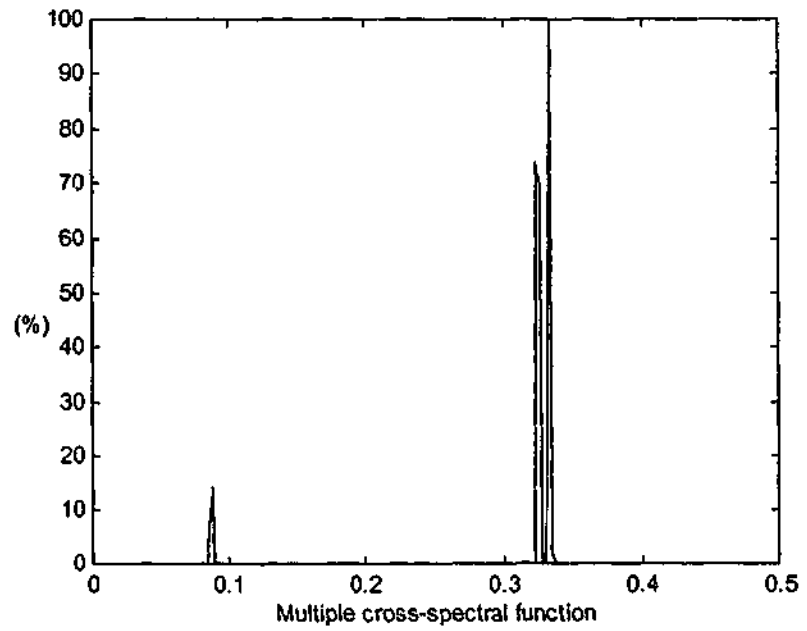
Despite the successful results it is still difficult finally to conclude which parameter would be solely the best parameter to use in the RRM. As we know the selectivity of protein interactions within the amino acid sequence could be identified if appropriate physical parameters are used. Thus, more research needs to be done in this direction. Therefore, we plan to measure for the first time the dielectric and conduction properties of twenty single amino acids and investigate the possible use of these new properties in the RRM. These newly measured parameters could then replace the EIIP values, which were mathematically calculated and include many approximations mentioned previously [54].



(a)

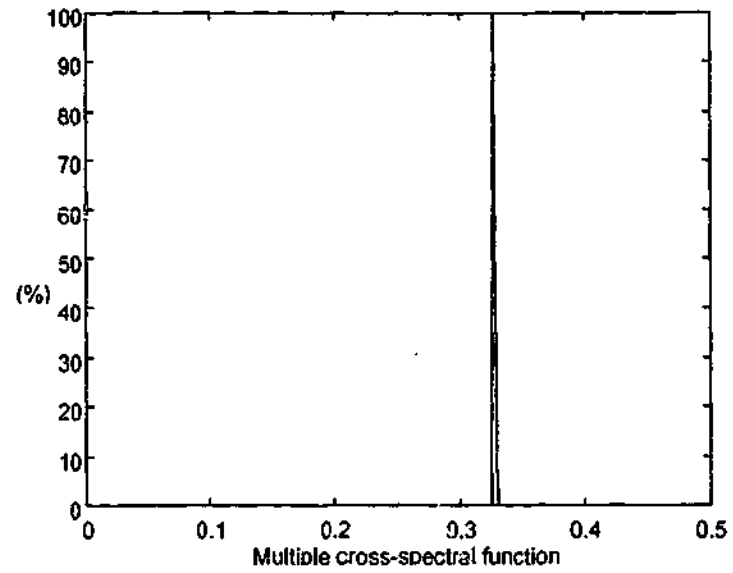


(b)

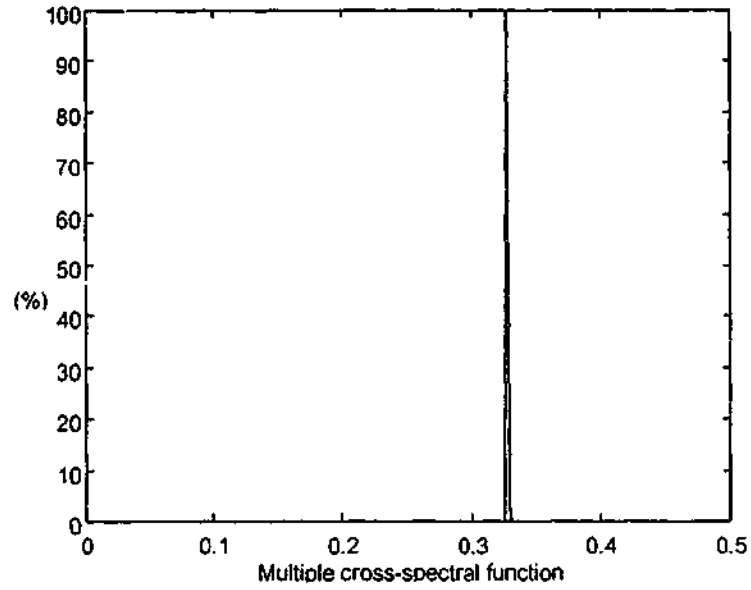


(c)

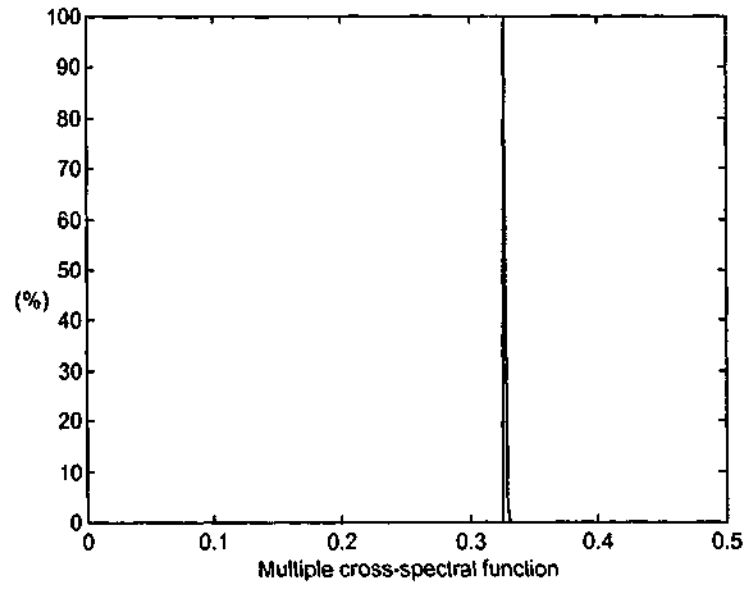
Figure 5.1.5.3 Multiple cross-spectral function of Glucagon proteins using (a) EIIP, (b) IC and (c) ICEE parameters.



(a)

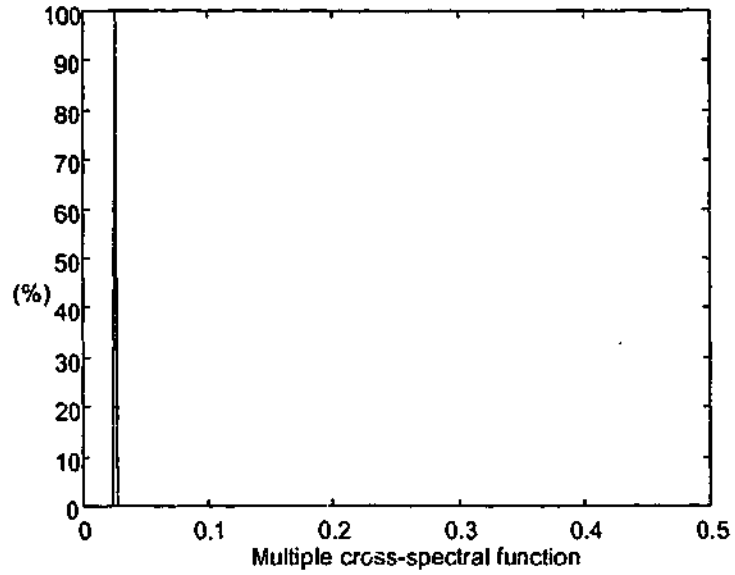


(b)

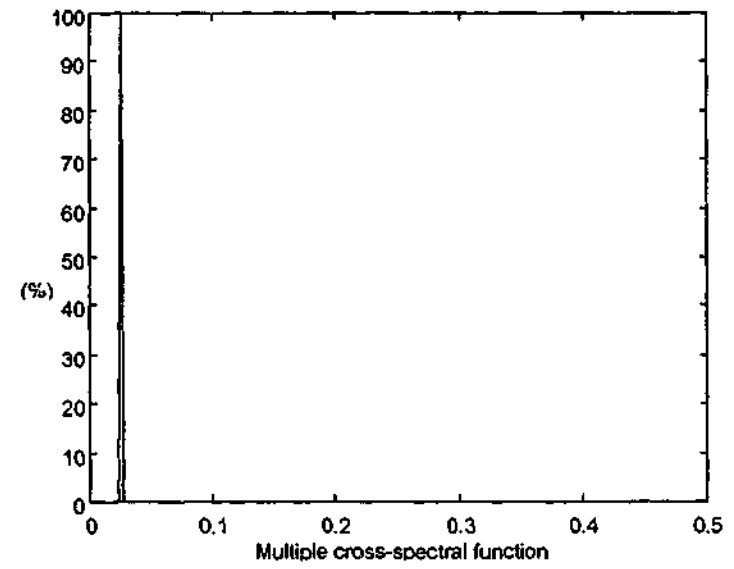


(c)

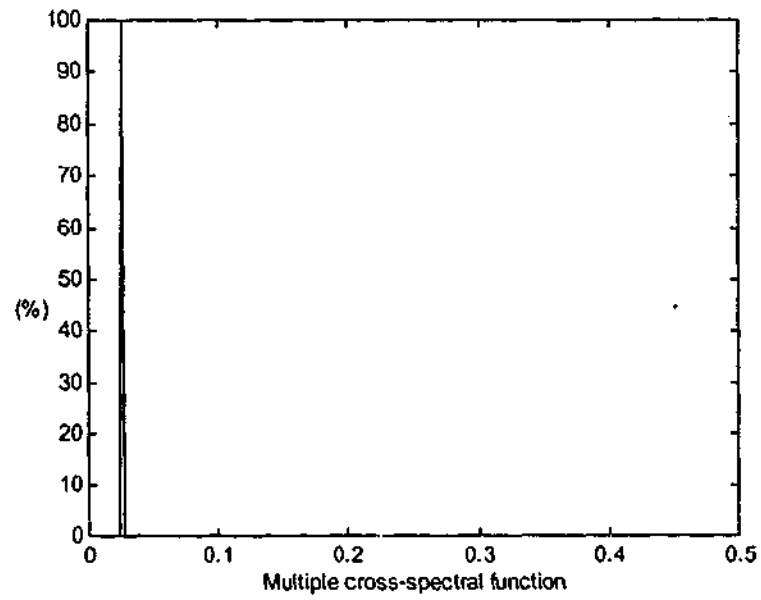
Figure 5.1.5.4 Multiple cross-spectral function of Lysozyme proteins using (a) EIIP, (b) IC and (c) ICEE parameters.



(a)

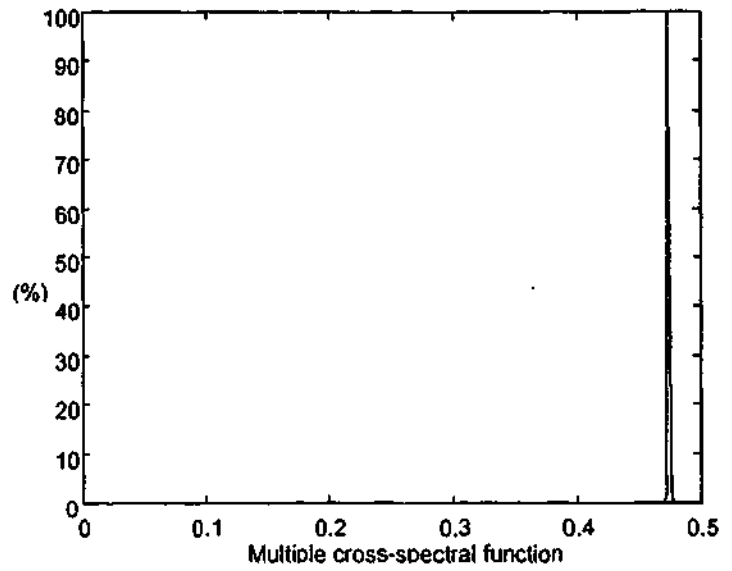


(b)

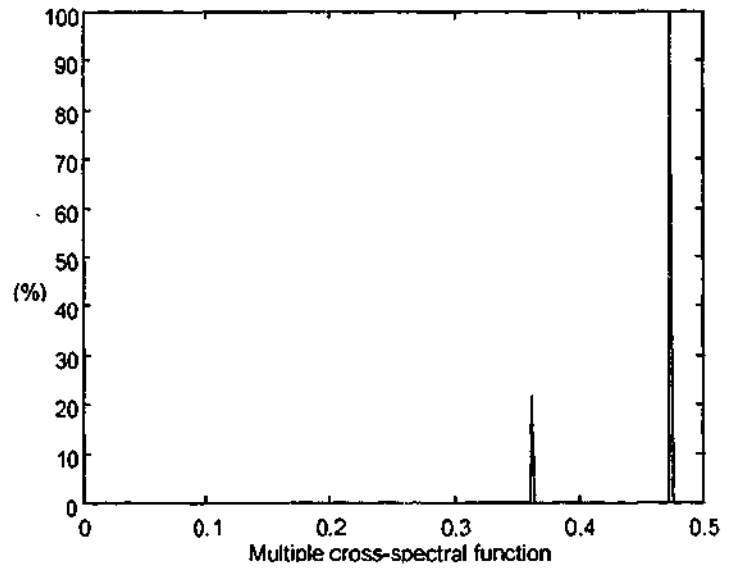


(c)

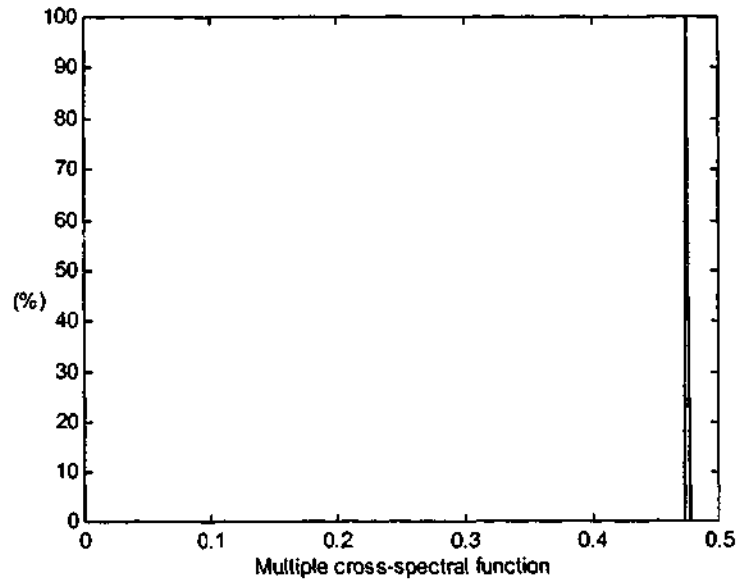
Figure 5.1.5.5 Multiple cross-spectral function of Hemoglobin proteins using (a) EIIP, (b) IC and (c) ICEE parameters.



(a)

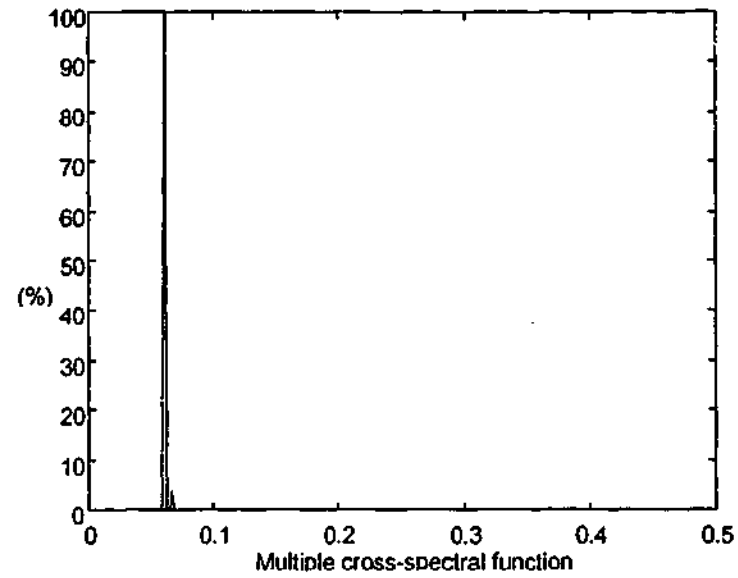


(b)

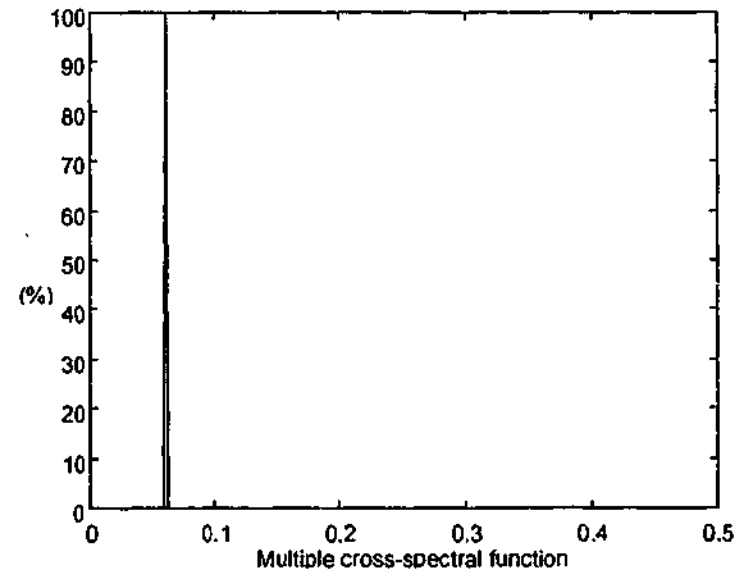


(c)

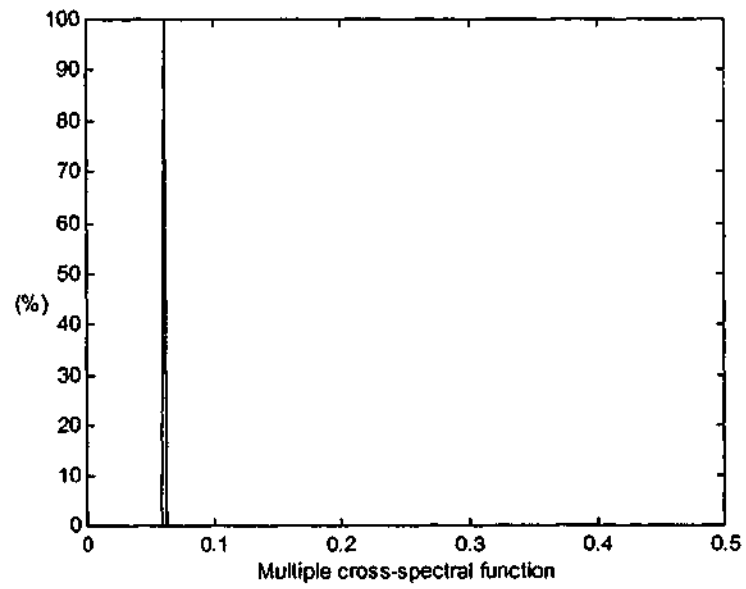
Figure 5.1.5.6 Multiple cross-spectral function of Cytochrome C proteins using (a) EIIP, (b) IC and (c) ICEE parameters.



(a)

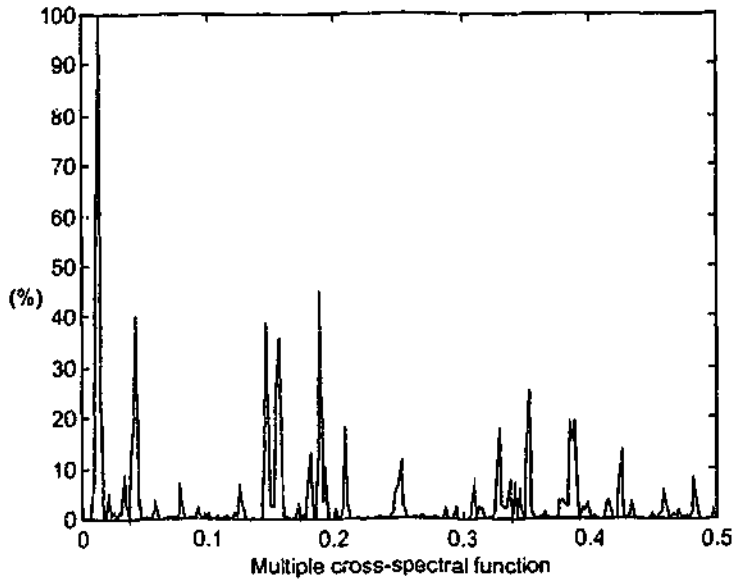


(b)

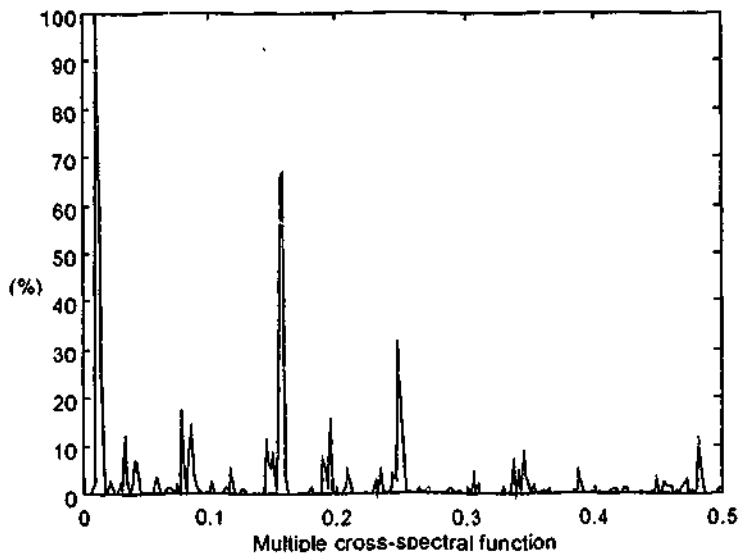


(c)

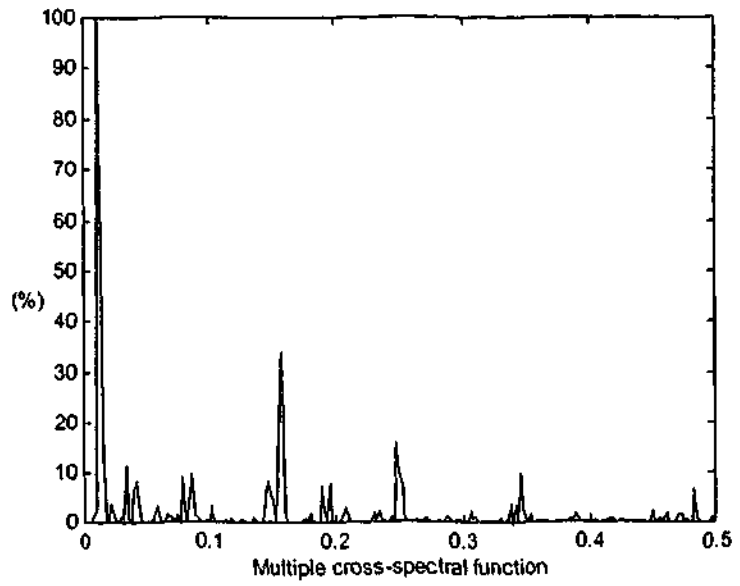
Figure 5.1.5.7 Multiple cross-spectral function of EGF proteins using (a) EIIP, (b) IC and (c) ICEE parameters.



(a)



(b)



(c)

Figure 5.1.5.8 Multiple cross-spectral function of TGF proteins using (a) EIIP, (b) IC and (c) ICEE parameters.

5.2 INVESTIGATION OF DIELECTRIC AND CONDUCTION PROPERTIES OF AMINO ACIDS

Szent-Gyorgyi [7ó] depicts a vision in which biology, once the study of whole living organism and its interations with the environment, becomes a science of the sub-molecular. The properties of living organisms become explicable in terms of the electrons which are responsible for the chemical reactions that "drive" the organism. The challenging aspects of biology from a modern electronic engineer's veiwpoint are threefold [75]:

(1) Biological systems store and process information using individual molecules, or sub-components of molecules as the storage medium. We may refer to this phenomenon as *biological information processing.*

(2) Biological systems are able to direct electronic (and ionic) currents along pathways that are also defined at the ultramicroscopic (molecular) level. This may refer to as *biological microelectronics.*

(3) Biological systems have developed a range of exquisitely sensitive transducting systems for receiving information about the state of thir environment. These are *biological microsensors.*

Electrochemical methods provide the means by which electronic devices and circuity can communicate directly (through the passage of d.c. currents) with electroactive biomolecules. If bioelectronic devices are to be consructed which exploit the properties of biological molecules, it is likely that in many cases electrochemical methods and concepts will be involved at the interface between the electronic and biological components. Dielectric spectroscopy has enabled investigations into the possible harmful effects of elctromagnetic field on biomolecular systems [75,76]. Amino acids ate the basic structural units from which proteins are formed. Depending on the pH of the local environment, amino acids can take a number of different ionized states: acid form, neutral form and base form. Each amino acid possesses a differetnt side chain and these can be small or relatively large. Each side chain also has a particular charge distribution. As it was mentioned previously the selectivity of protein interactions within the amino acid sequence could be identified if appropriate physical parameters are used [6]. Following the preliminary gains made through our previous studies, particularly the finding of the physicochemical parameters, which could be used in the RRM, there are several other steps that could be undertaken to facilitate and improve the current RRM approach. These steps are the dielectric measurements, which we have used to probe the physical changes

taking place in each particular amino acid [75-77]. Within this study we have measured the dielectric and conduction properties of twenty single amino acids for its use in the RRM. These newly measured parameter values will replace the electron-ion-interaction potential (EIIP) values, which were mathematically calculated from an approximate pseudopotential model [54]. This novel approach will provide us with more accurate amino acid parameters. It will allow us to improve active site prediction with the modified RRM model.

Techniques

Experimental dielectric bridge methods have been employed in some form for low molecular mass materials and polymers for more than forty years [77]. But never been used on amino acids.

These methods describe in more detail the use of frequency response analyzers and time domain reflectrometry, which are relatively new technique [75]. Although there are a number of different bridge configurations, the basic principle of operation is the same and involves the balancing of the impedance of the two opposite arm pairs of the bridge. One of these arms contains the sample, which may be modeled as a resistor R_s in parallel with a capacitor C_s.

The technique proposed [78] is designed to measure values of the sample capacitance (Cp, F) and conductance (G, S) over a frequency range 20-10MGz.

The main variables we are interested in measuring are:

C_p sample capacitance (μF)

G sample conductance (μS)

ε' sample dielectric constant (unitless)

tan δ sample loss tangent

These are explained here in more detail now.

During the experiment we have measured the values of the capacitance and conductance of 20 amino acid solutions. These new parameters values are given below. Based on results obtained we have calculated then the dielectric constant for each analyzed sample.

ε', dielectric constant:

In general terms, for a sample of thickness ***d*** and area ***A*** held between two parallel electrodes, the permittivity ε' (dielectric constant) can be calculated from the measured values of capacitance C_p using the equations:

$$\varepsilon' = C_p / C_o \qquad (5.2.1)$$

where C_p and C_o are the capacitance of the cell with a sample and an empty cell. The capacitance of an empty cell is given by

$$C_o = \varepsilon_o A / d \qquad (5.2.2)$$

where d and A are the area and thickness of the sample, respectively. ε_o is the permittivity of free space, a universal constant. C_p is the number that the machine experimentally determines. Therefore to work out ε' we need to accurately know the sample thickness and area. However, in our study we measure the capacitance and conductance of the prepared amino acid solutions. Thus, in our case (electrode is surrounded by the solution environment) we should measure experimentally the C_o of an empty cell as well. The experimental value of C_o is shown in Table 5.2.1.

tan δ, dielectric loss:

tan δ is independent of sample dimensions and equals

$$\tan\delta = \varepsilon''/\varepsilon' = (G/\omega)/C_p \qquad (5.2.3)$$

where $\omega = 2\pi$, C_p and G capacitance and conductance of the sample respectively.

A large source of error in the higher frequency region, which is particularly serious for conductive samples, is that of stray inductance. To minimize this phenomenon, a.c bridges are constructed symmetrically so that effects essentially cancel.

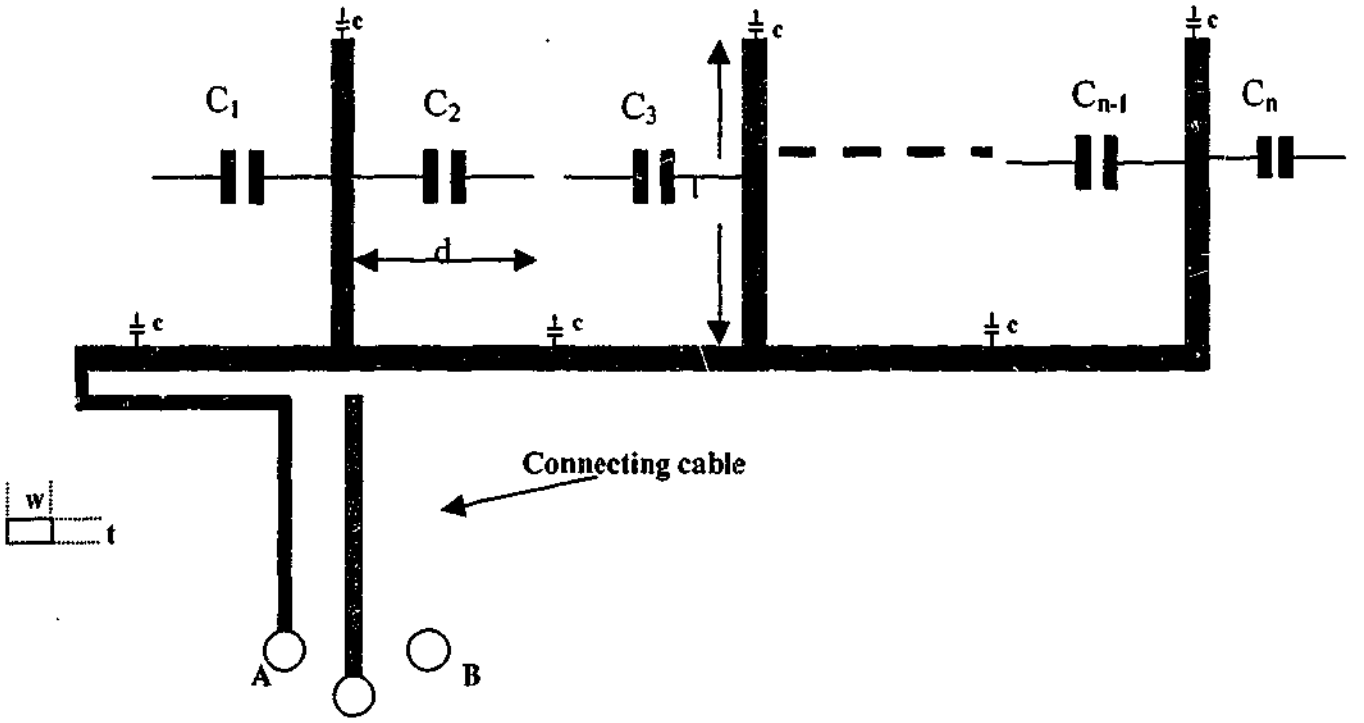
Sample preparation and experimental measurements

The preparation of amino acid samples for this study required additional laboratory work and was performed at the analytical laboratory at the Department of Chemical Engineering. As it was mentioned above all amino acids are white crystalline solids melted indifferently at high temperatures (>200°C) and with some decomposition. They are soluble in water but they vary considerably in their solubility. Furthermore the solubility of any acid increases as the solution is made basic or acidic. Most are insoluble in organic solvents. These properties are a direct consequence of the dual basic-acidic character of amino acids which gives rise to ionic behavior. Based on different physicochemical properties of analyzed amino acids we have used distilled water and 1M HCL as a solvent environment.

Alanine, Arginine, Glycine, Lysine, Proline, Serine, Threonine, Valine and Cysteine amino acids were dissolved in distilled water at room temperature using the magnetic stirrer. These amino acids were easily completely dissolved, obtained clear solutions. Asparagine and Histidine were dissolved in water using a hot plate magnetic stirrer, obtained clear light-yellow solution respectively. Methionine, Glutamine and Tryptophan were dissolved in water using the sonocation technique [81], obtained clear solutions. Leucine, Isoleucine, Aspartic Acid, Glutamic Acid, Phenylalanine and Tyrosine were dissolved completely in 1M HCL at room temperature using the magnetic stirrer, obtained clear solutions.

Hence, to run dielectric measurements we have prepared aqueous solutions of 0.5M concentration of the following amino acids: Alanine, Arginine, Glycine, Lysine, Proline, Serine, Threonine, Valine, Cysteine, Asparagine, Histidine, Methionine, Glutamine and Tryptophan. HCL solutions of Leucine, Isoleucine, Aspartic Acid, Glutamic Acid, Phenylalanine and Tyrosine have been prepared as well.

Then the experimental measurements of capacitance and conductance of our samples (amino acid solution, 60 ml) have been undertaken at the Department of Materials Engineering using Dr. George Simon's Software Win DETA V3.71 and HP4284 RLC Bridge (20-10MHz) [79,80]. The liquid cell used in these measurements consists of the glass container and the sensor attached. The description of the sensor is presented in Figure 5.2.1. The equipment and a glass container used in this experiment are pictured and shown in Figure 5.2.2. Results of this study are presented in Table 5.2.1 and 5.2.2.



Legenda

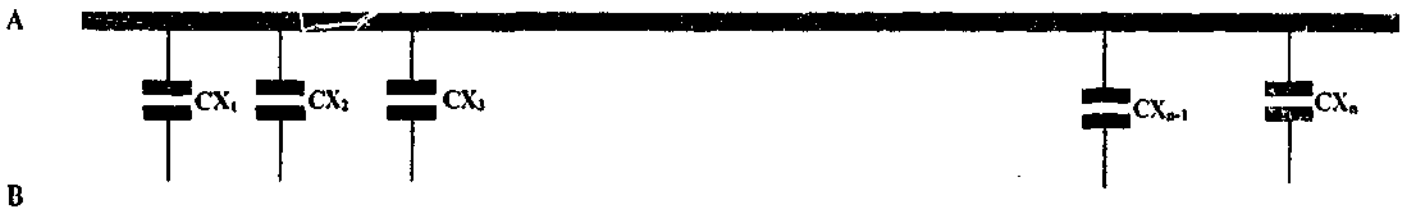
d = distance between elements
l = length of each element
w = width of each element
t = thickness of each element

Assumption:

$C_1 = C_2 = \ldots\ldots = C_n$ (1)
$CX_1 = C_1 + c = CX_2 = \ldots\ldots = CX_n$ (2)

CX_y (y = 1, ..., n) (3)
Length L = l + w

The capacitor plate area A is then: A = L*t



$$C_{tot} = \sum_{1}^{n} CX_y \quad \text{or (using (2))} \quad C_{tot} = n * CX_1$$

The dimensions of the electrodes on the IDEX sensor head are as follows:

Length = 0.950" Width = 0.004" Spacing = 0.004" Thickness = 0.001"

The electrodes cover an area of:

Area length = 0.950" Area width = 0.400"

Figure 5.2.1 Sensor

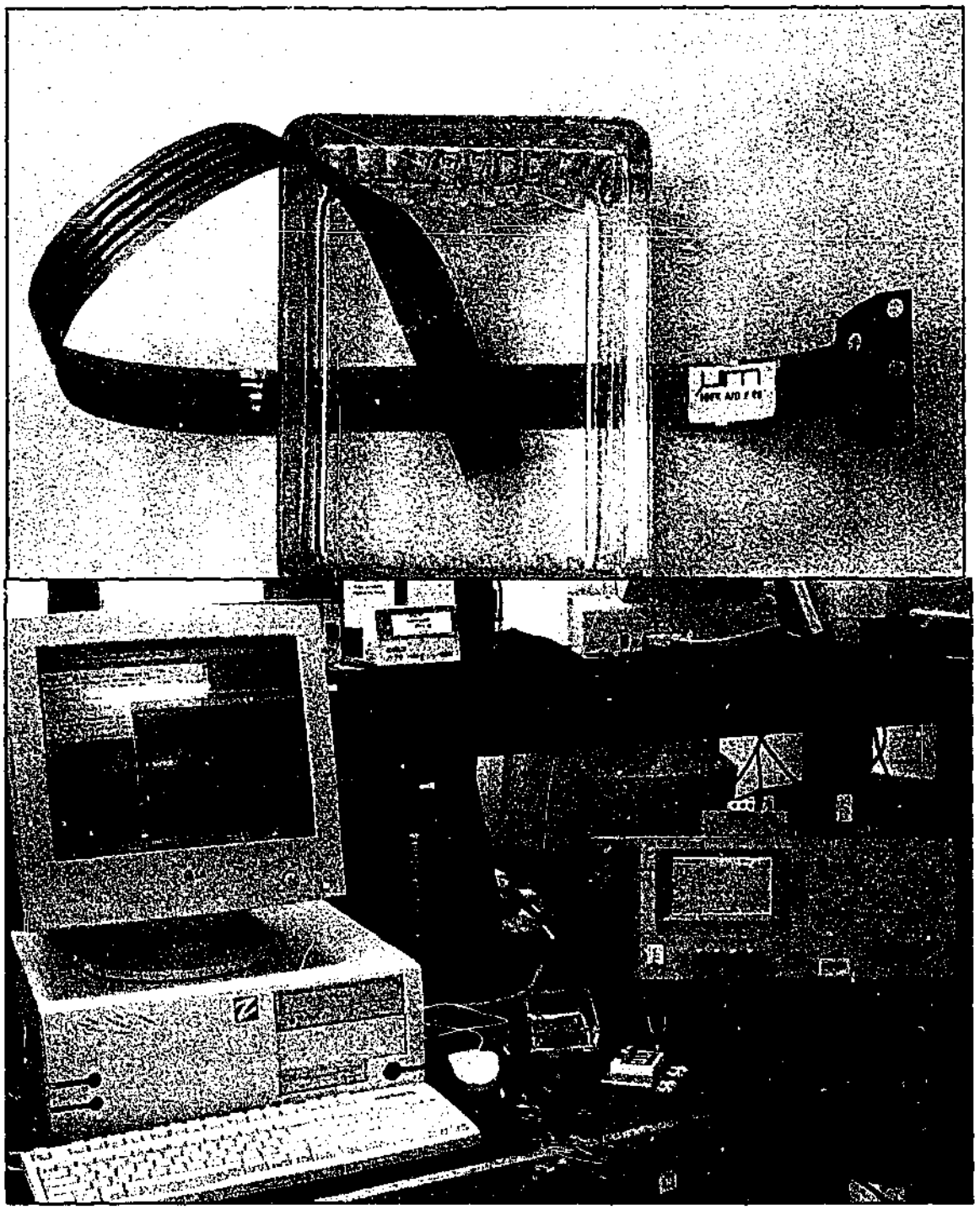


Figure 5.2.2 Liquid cell and equipment used.

Table 5.2.1 Capacitance (μF) and Conductance (μS) of Amino Acid Aqueous Solution.

Amino	Sample 1		Sample 2		Sample 3		Sample 4		Average Value		STD	
Acid	Cp	G	Cp	G	Cp	G	Cp	G	Cp	G	Cp	G
Alanine	2.277	303.11	1.878	272.24	1.863	274.47	1.993	280.51	2.003	282.58	0.1918	14.124
Arginine	11.82	433.50	12.27	447.61	12.43	456.91	12.58	460.78	12.28	449.70	0.3287	12.132
Asparagine	4.984	312.35	5.085	324.89	4.964	325.39	4.868	323.55	4.975	321.55	0.0890	6.179
Cysteine	3.970	399.48	3.956	410.53	3.917	415.70	3.876	418.61	3.930	411.08	0.0423	8.424
Glycine	3.362	287.47	3.418	297.07	3.512	301.55	3.503	305.59	3.449	297.92	0.0717	7.787
Histidine	6.204	279.51	6.530	315.68	6.357	324.16	6.015	323.27	6.277	310.66	0.2194	21.110
Lysine	8.986	304.32	9.339	320.33	9.401	325.35	9.470	327.87	9.299	319.47	0.2154	10.573
Proline	2.064	273.88	2.103	272.06	2.056	267.99	1.914	262.36	2.034	269.07	0.0828	5.108
Serine	2.082	223.83	2.160	239.55	2.193	243.90	2.214	247.53	2.162	238.70	0.0580	10.438
Threonine	2.165	267.26	2.278	288.95	2.279	290.78	2.292	292.75	2.254	284.94	0.0593	11.885
Valine	2.740	243.87	2.695	252.46	2.670	253.40	2.654	252.80	2.690	250.63	0.0375	4.525
Glutamine	6.467	317.79	5.994	301.06	5.858	293.23	5.738	276.65	6.014	297.18	0.3194	17.097
Methionine	4.756	406.81	4.829	415.21	4.380	395.31	3.306	349.60	4.318	391.73	0.7026	29.249
Tryptophan	2.179	219.40	2.143	215.97	2.148	213.78	2.118	211.24	2.147	215.10	0.0250	3.459

Table 5.2.2 Capacitance (μF) and Conductance (mS) of Amino Acids dissolved in HCL.

Amino	Sample 1		Sample 2		Sample 3		Sample 4		Average Value		STD	
Acid	Cp	G	Cp	G	Cp	G	Cp	G	Cp	G	Cp	G
Glutamic Acid	104.41	13.019	103.43	13.054	99.083	12.881	94.884	12.913	100.45	12.967	4.3746	0.0814
Tyrosine	81.298	14.453	82.702	15.246	86.057	14.683	88.878	14.076	84.734	14.615	3.4087	0.4897
Leucine	96.733	14.112	92.622	12.915	90.997	12.313	91.805	11.751	93.039	12.773	2.5503	1.0115
Isoleucine	84.423	14.753	81.422	12.734	84.525	11.764	86.713	11.202	84.271	12.568	2.1732	1.5606
Phenylalanine	97.267	9.282	95.934	9.164	95.687	9.109	96.519	9.094	96.352	9.162	0.7029	0.0853
Aspartic Acid	87.188	10.131	96.321	9.563	103.85	9.435	107.77	9.497	98.782	9.657	9.0728	0.3206

Distilled water: Cp=0.516 μF; G=124.097 μS; $\varepsilon'=0.0191 \times 10^6$ **HCL:** Cp=80.848 μF; G=15.701 mS; $\varepsilon'=2.9905 \times 10^6$

Empty cell: Co=27.035 pF; G=0.00034 μS

Table 5.2.1 Continued

Cp sample/ Cp H2O	G sample/ G H2O	ε'x10^6 sample	Δε'x10^6 amino acid	tan δ sample
3.882	2.277	0.0741	0.0550	22.465
23.798	3.624	0.4542	0.4351	5.8313
9.641	2.591	0.1841	0.1650	10.290
7.616	3.313	0.1457	0.1266	16.656
6.684	2.401	0.1276	0.1085	13.755
12.165	2.503	0.2322	0.2131	7.881
18.021	2.574	0.3440	0.3249	5.4706
3.942	2.168	0.0752	0.0561	21.065
4.190	1.923	0.0710	0.0519	17.581
3.602	2.296	0.0834	0.0643	20.130
5.213	2.020	0.0995	0.0804	14.836
11.655	2.395	0.2225	0.2034	7.869
8.368	3.157	0.1597	0.1406	14.446
11.161	1.733	0.0794	0.0603	15.953

Table 5.2.2 Continued

Cp sample/ Cp HCL	G sample/ G HCL	ε'x10^6 sample	Δε'x10^6 amino acid	tan δ sample
1.2425	0.8259	3.7156	0.7251	20.555
1.0481	0.9308	3.1342	0.1437	27.465
1.1508	0.8135	3.4114	0.4209	21.861
1.0423	0.8005	3.1171	0.1266	23.748
1.1918	0.5835	3.5640	0.5735	15.142
1.2218	0.6151	3.6539	0.6634	15.567

The changes of the $C_{p\ sample}$, ε'_{sample} and $\tan\delta_{sample}$ in a frequency range are shown in Fig.5.2.3-5.2.25.

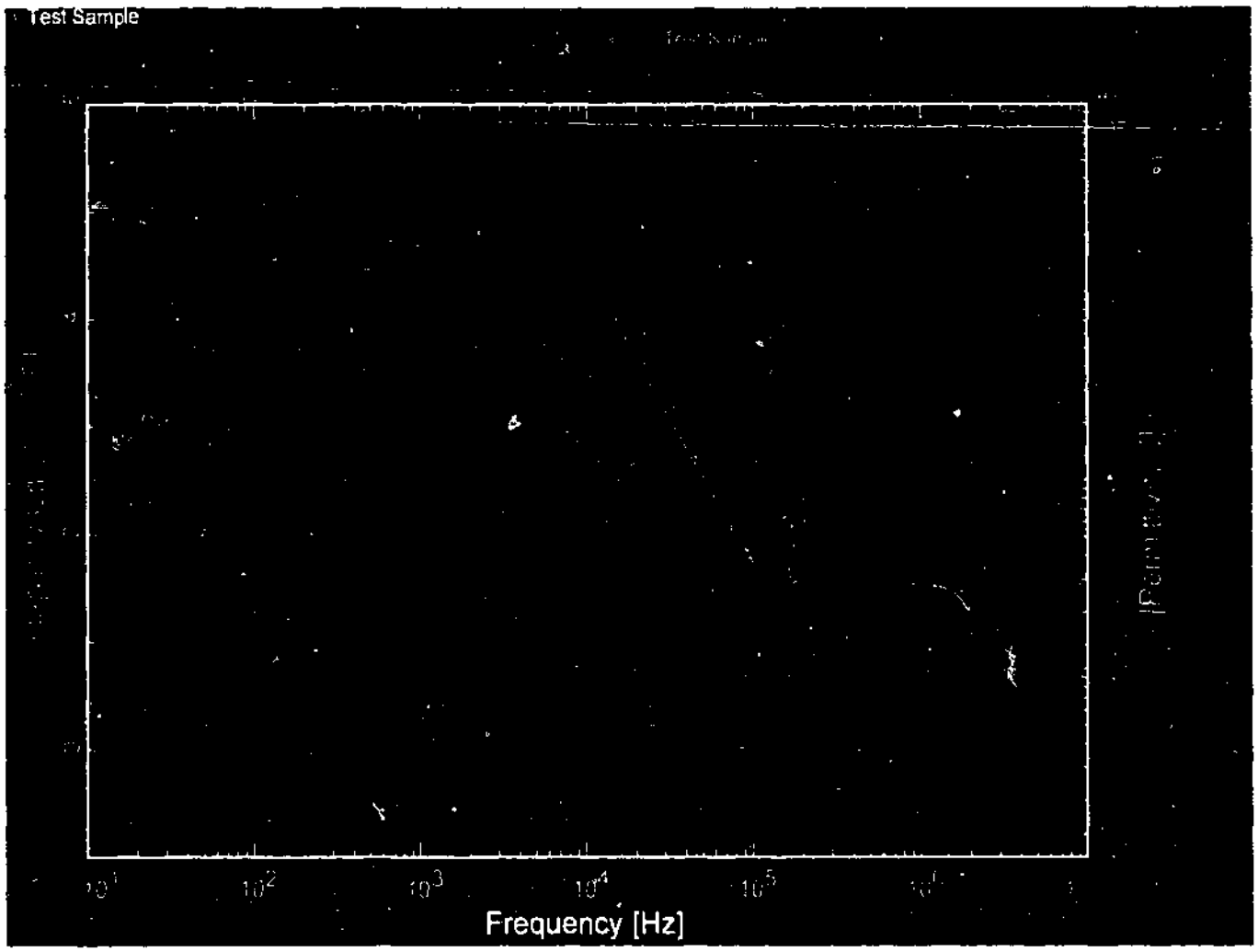


Figure 5.2.3 Example of Asparagine amino acid aqueous solution. The change of $C_{p\ sample}$ in a frequency range (red line), the change of ε'_{sample} in a frequency range (green line).

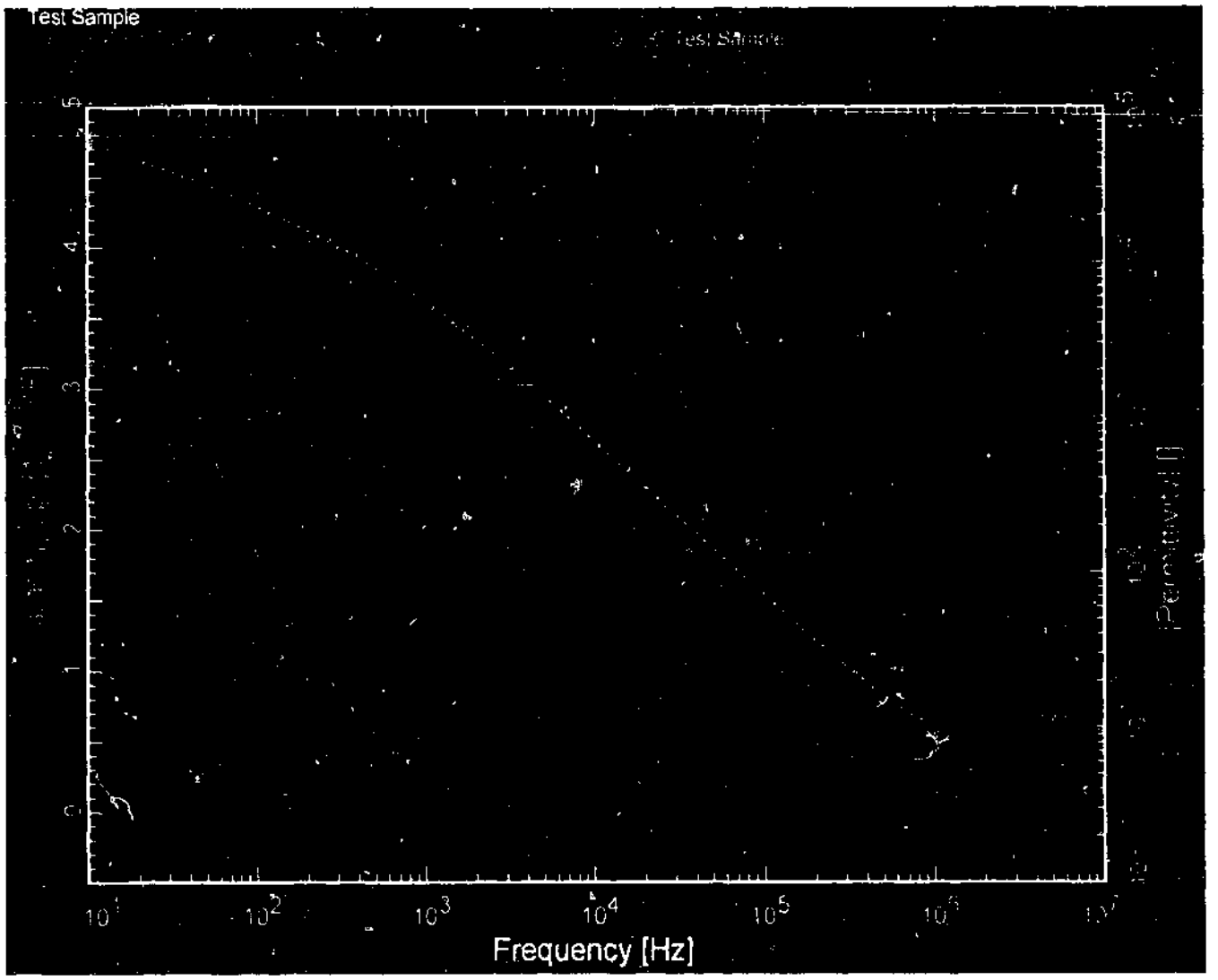


Figure 5.2.4 Example of Cysteine amino acid aqueous solution. The change of $C_{p\ sample}$ in a frequency range (red line), the change of ε'_{sample} in a frequency range (green line).

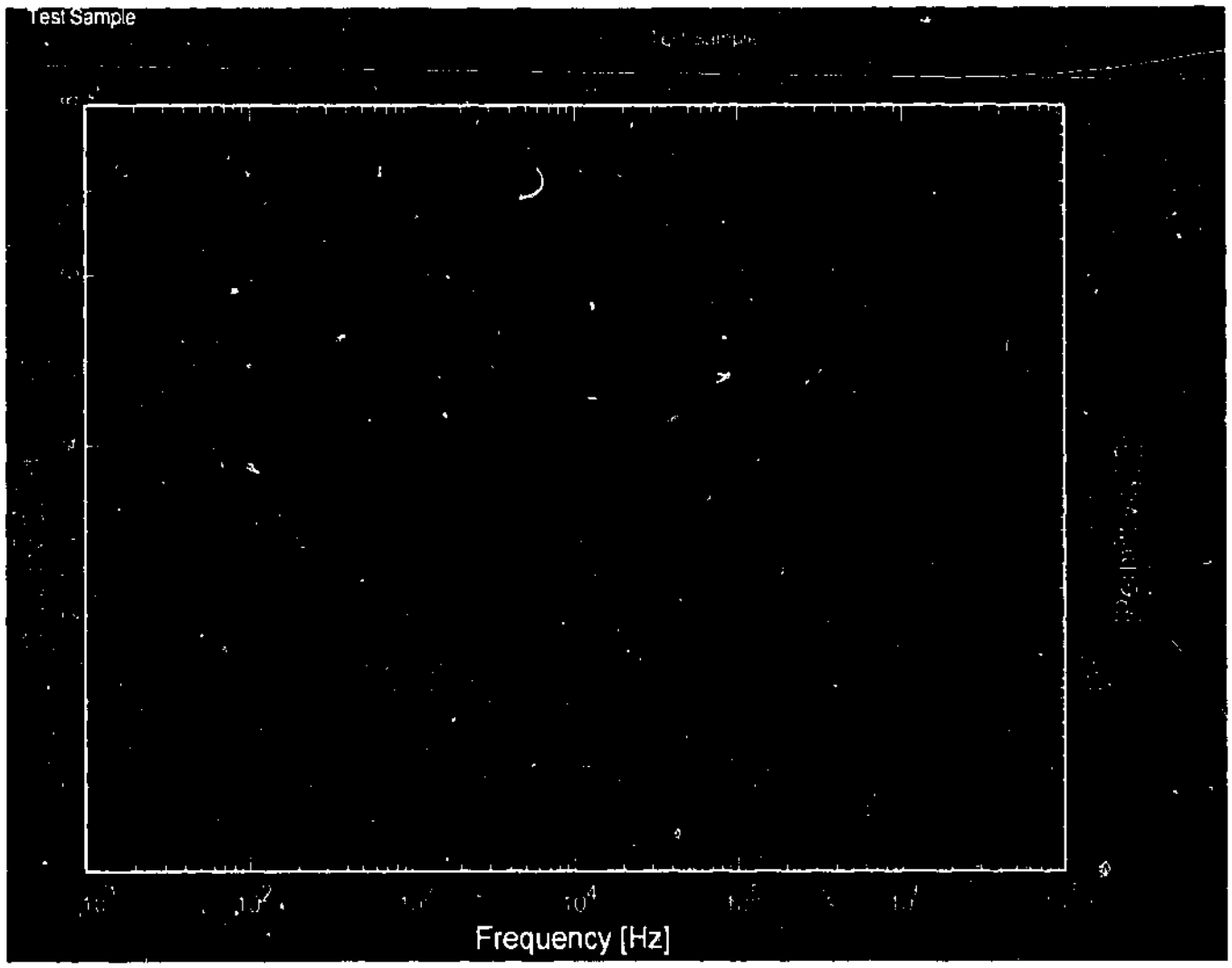


Figure 5.2.5 Example of Histidine amino acid aqueous solution. The change of $C_{p\,sample}$ in a frequency range (red line), the change of ε'_{sample} in a frequency range (green line).

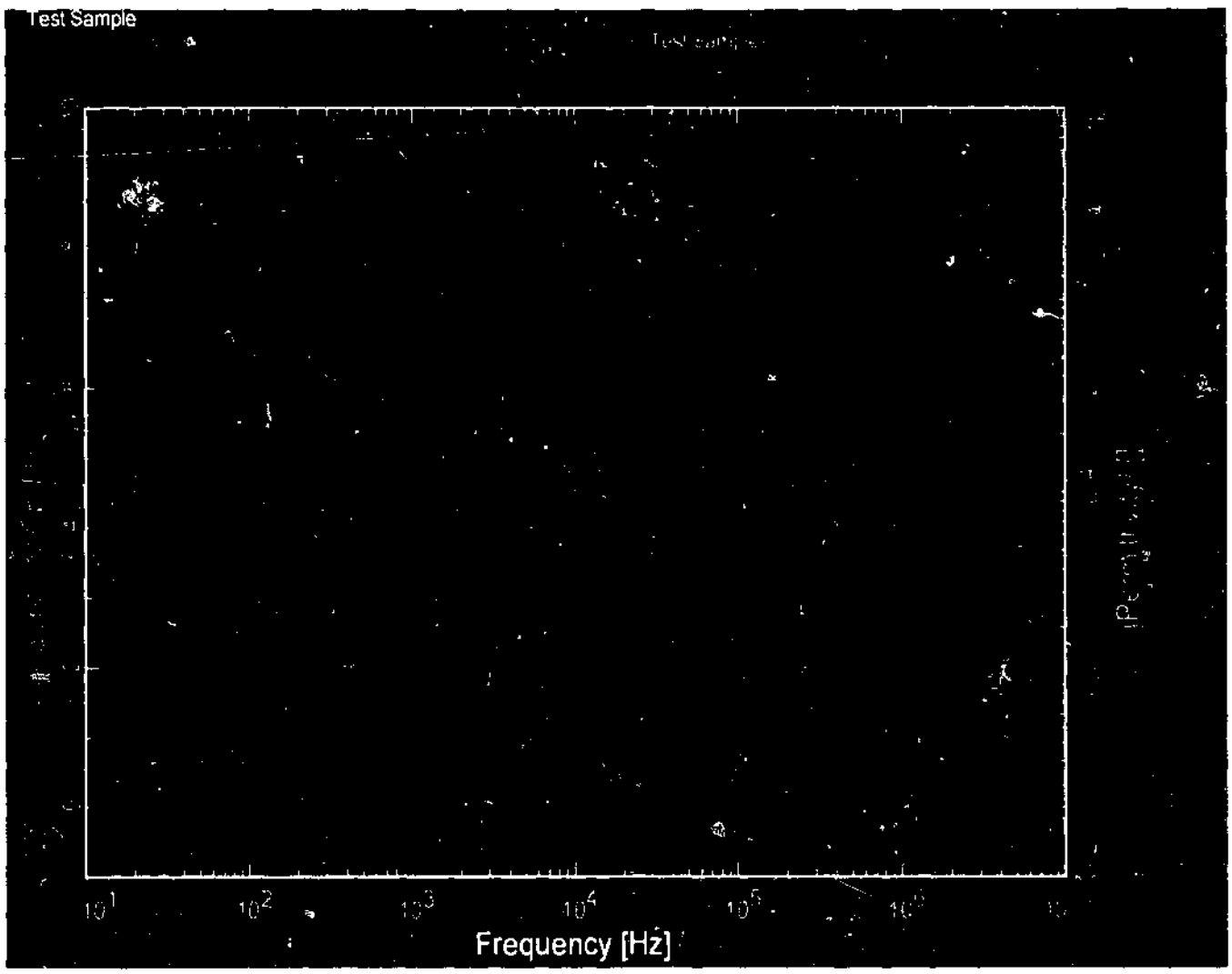


Figure 5.2.6 Example of Lysine amino acid aqueous solution. The change of $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

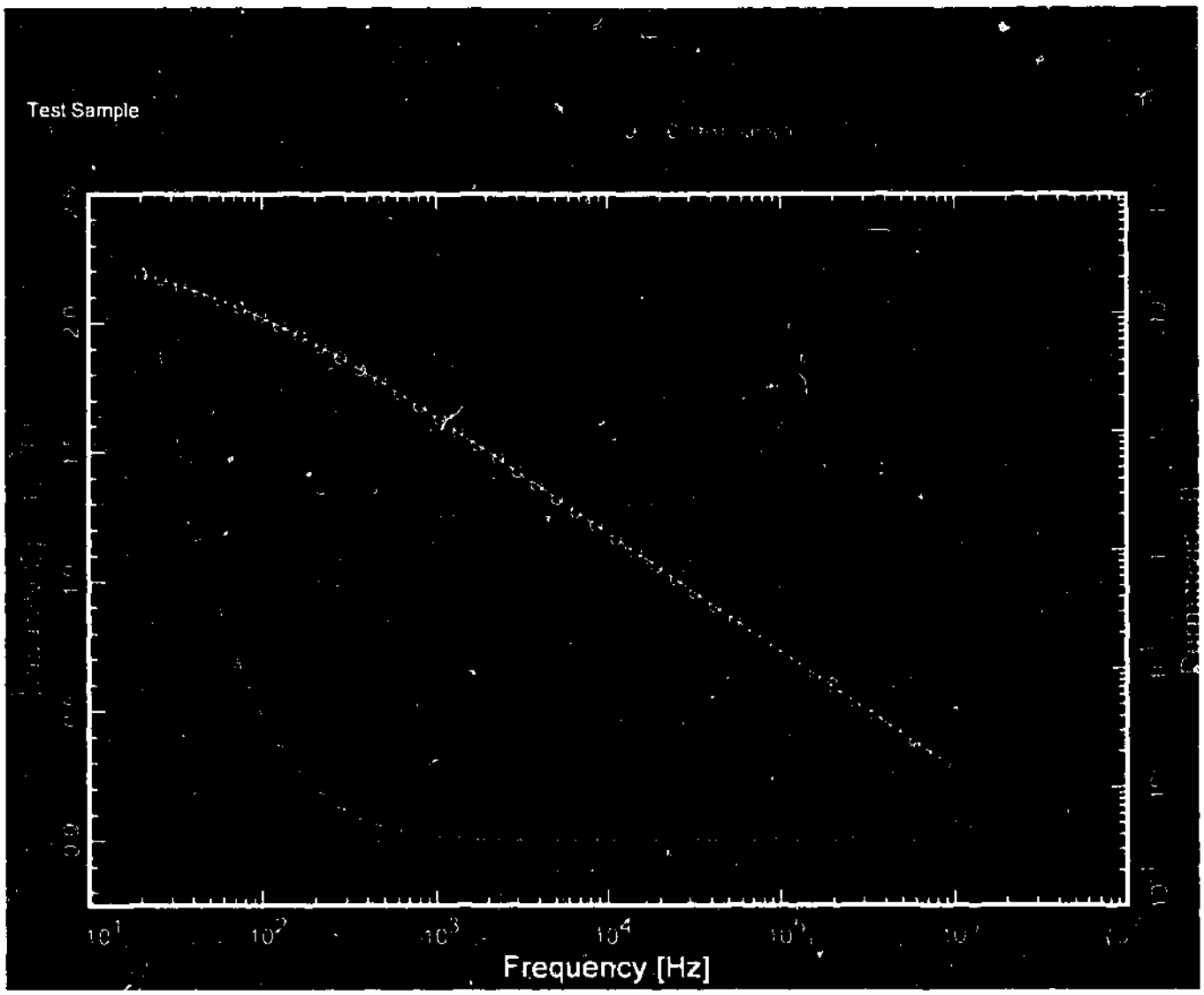


Figure 5.2.7 Example of Threinine amino acid aqueous solution. The change of $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

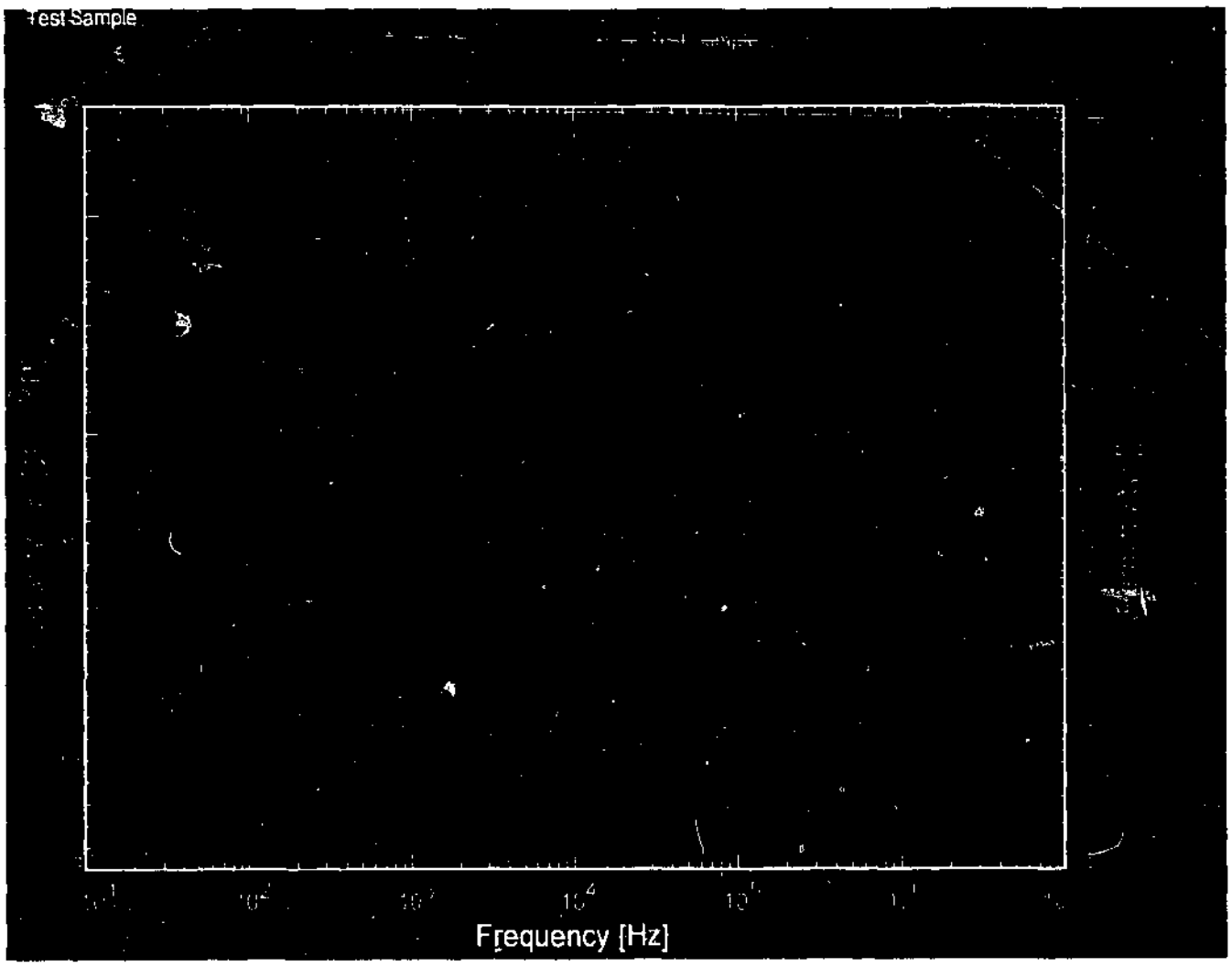


Figure 5.2.8 Example of Valine amino acid aqueous solution. The change of $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

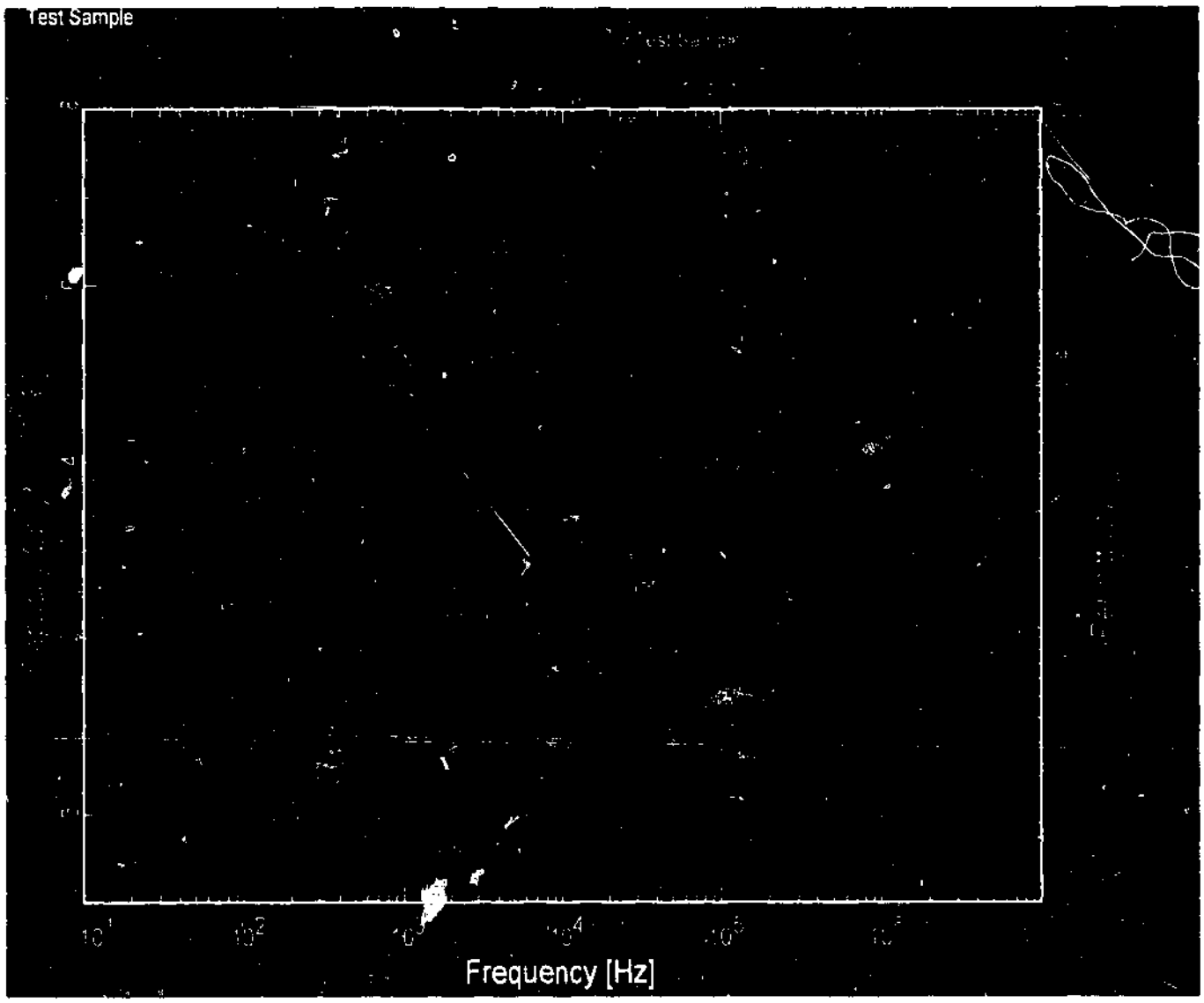


Figure 5.2.9 Example of Glutamine amino acid aqueous solution. The change of $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

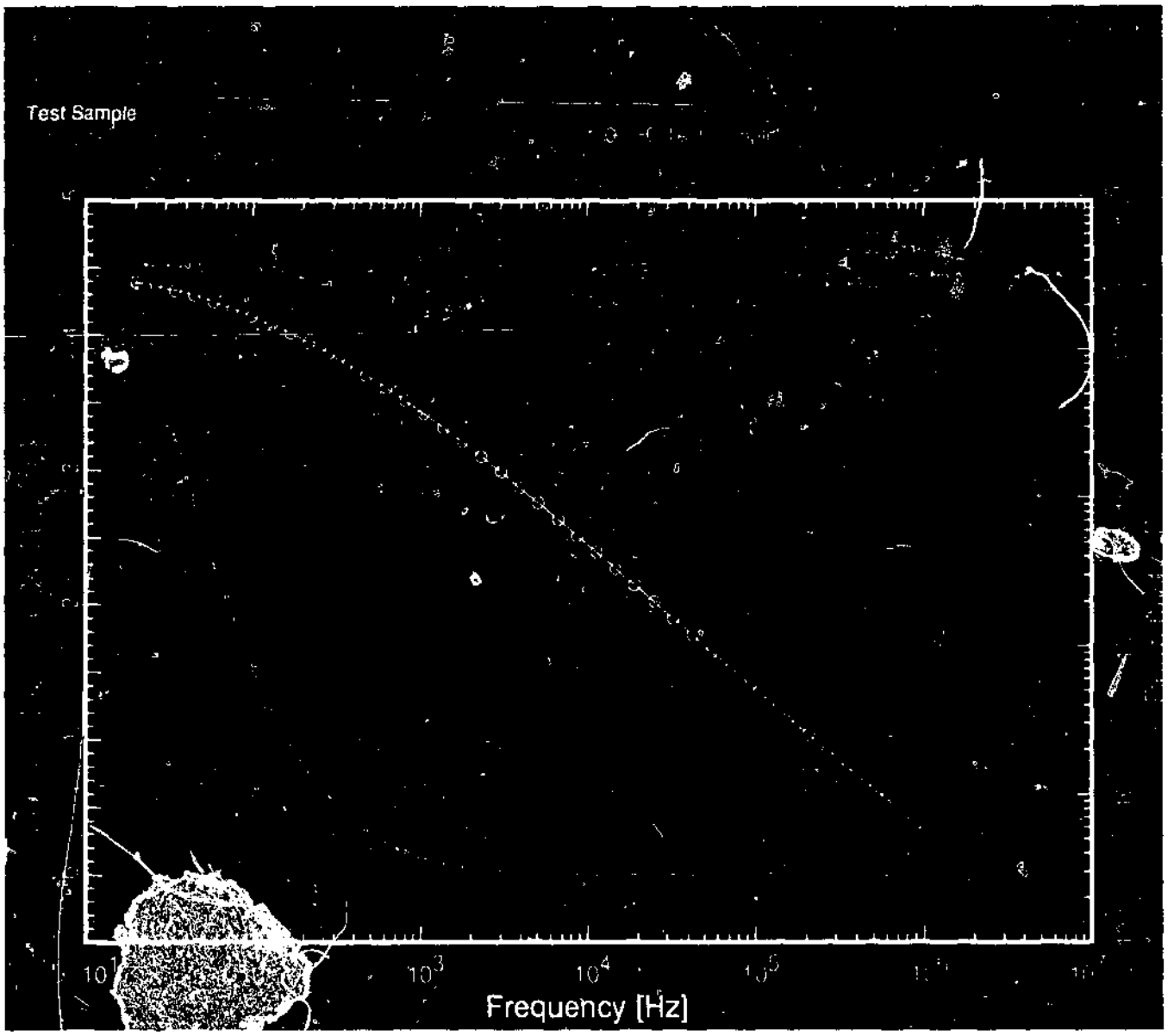


Figure 5.2.10 Example of Methionine amino acid aqueous solution. The change of $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

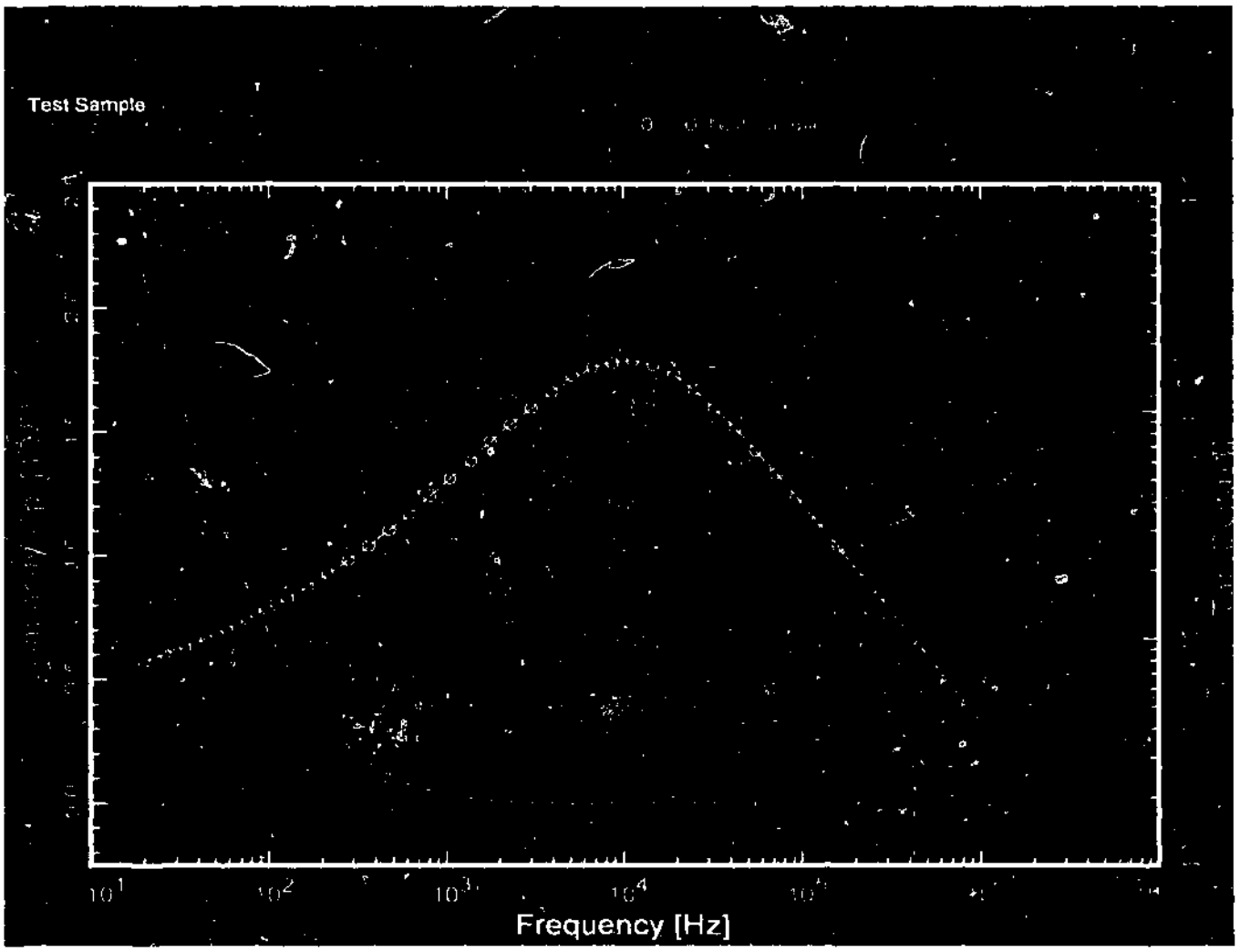


Figure 5.2.11 Example of Tryptophan amino acid aqueous solution. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of Tan (Delta) values in a frequency range (green line).

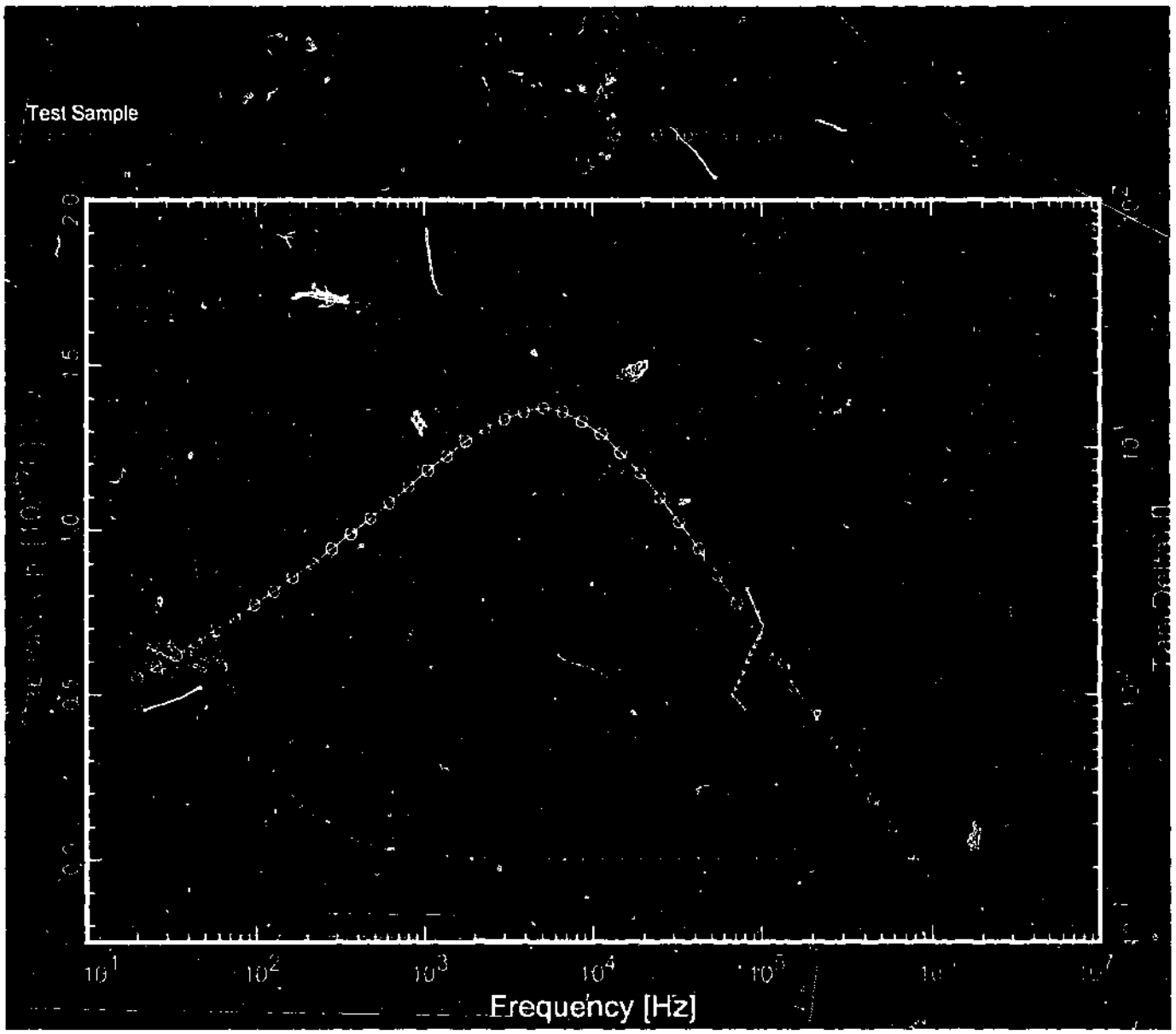


Figure 5.2.12 Example of Alanine amino acid aqueous solution. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of Tan (Delta) values in a frequency range (green line).

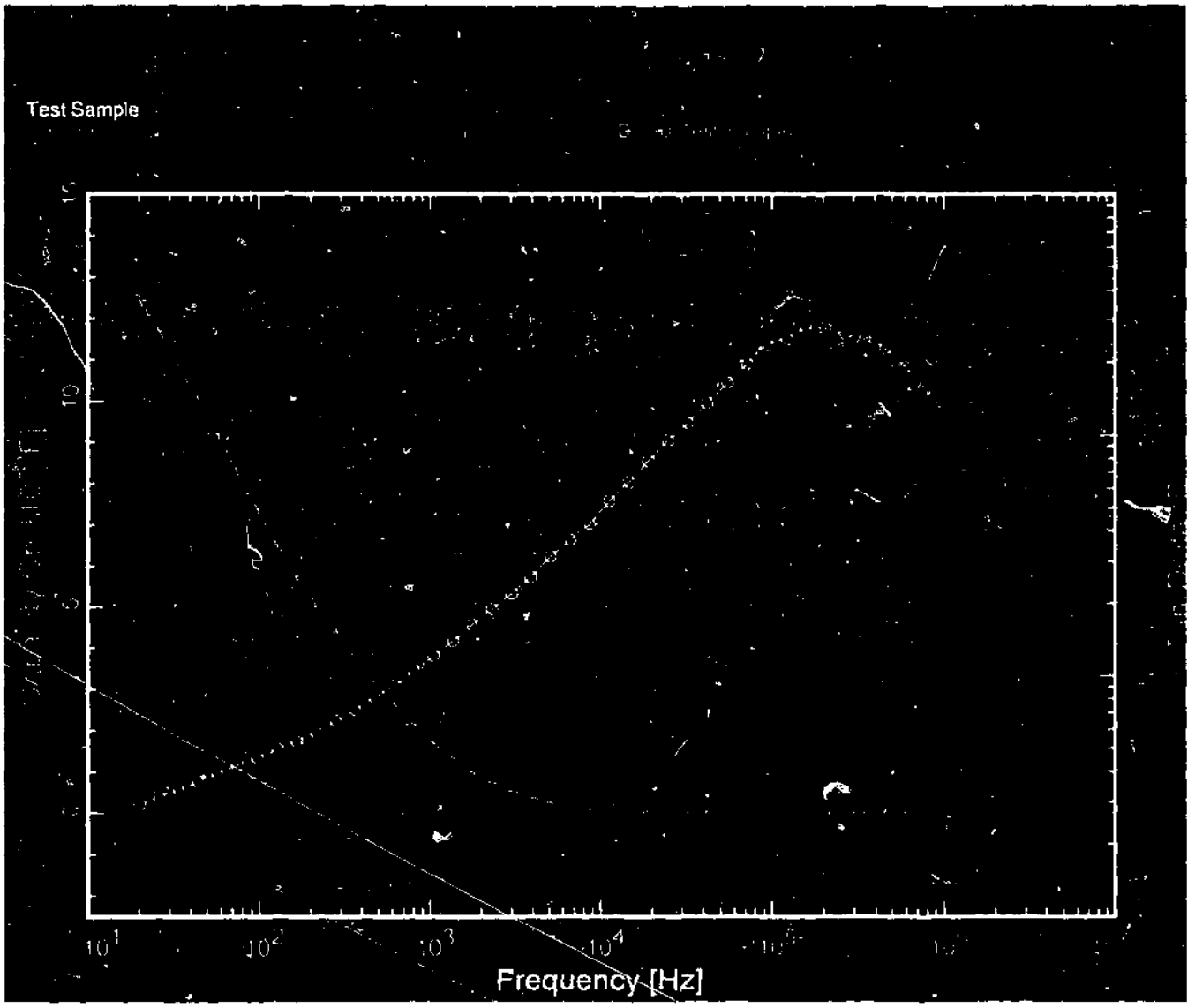


Figure 5.2.13 Example of Arginine amino acid aqueous solution. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of Tan (Delta) values in a frequency range (green line).

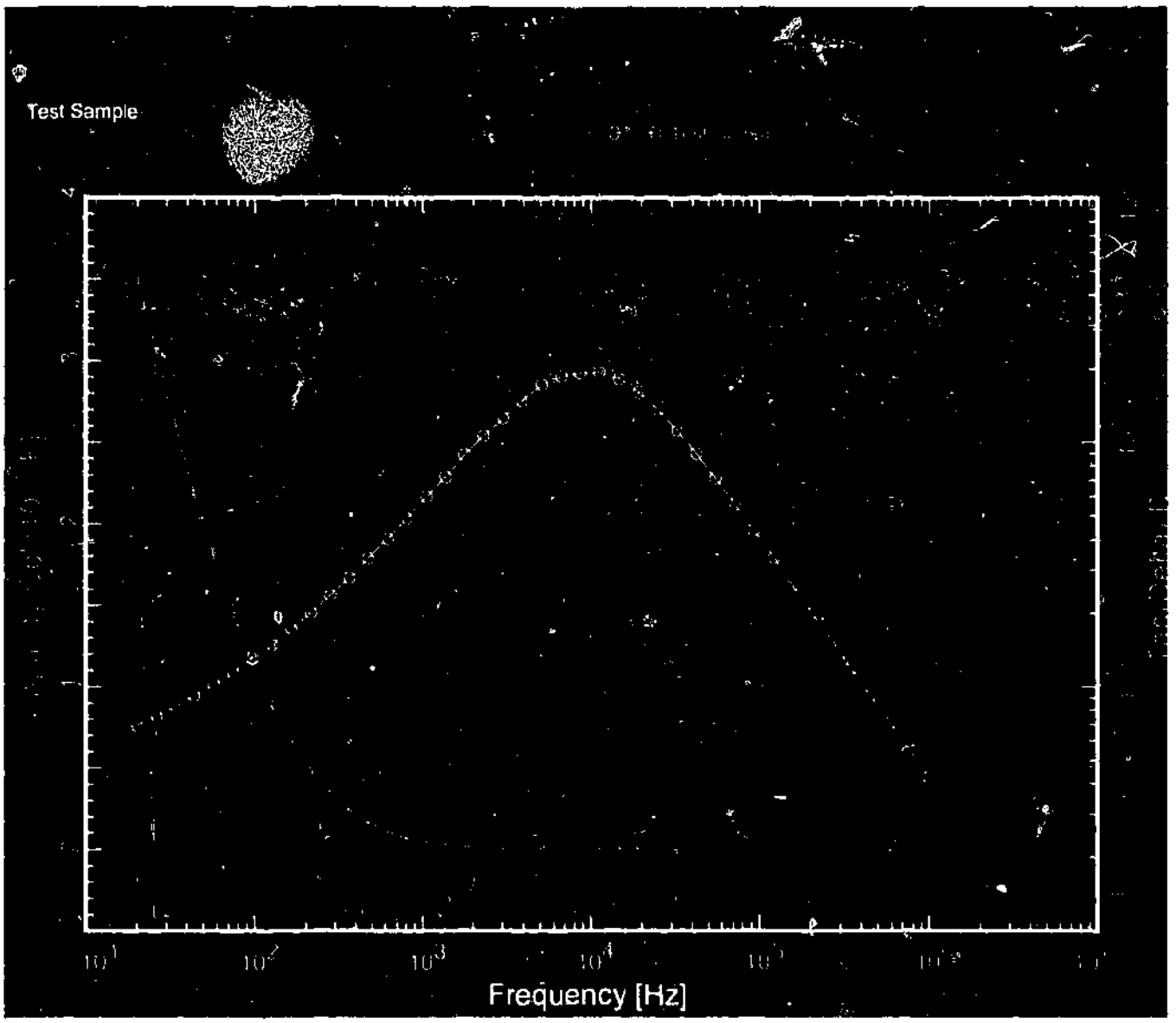


Figure 5.2.14 Example of Glycine amino acid solution. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of Tan (Delta) values in a frequency range (green line).

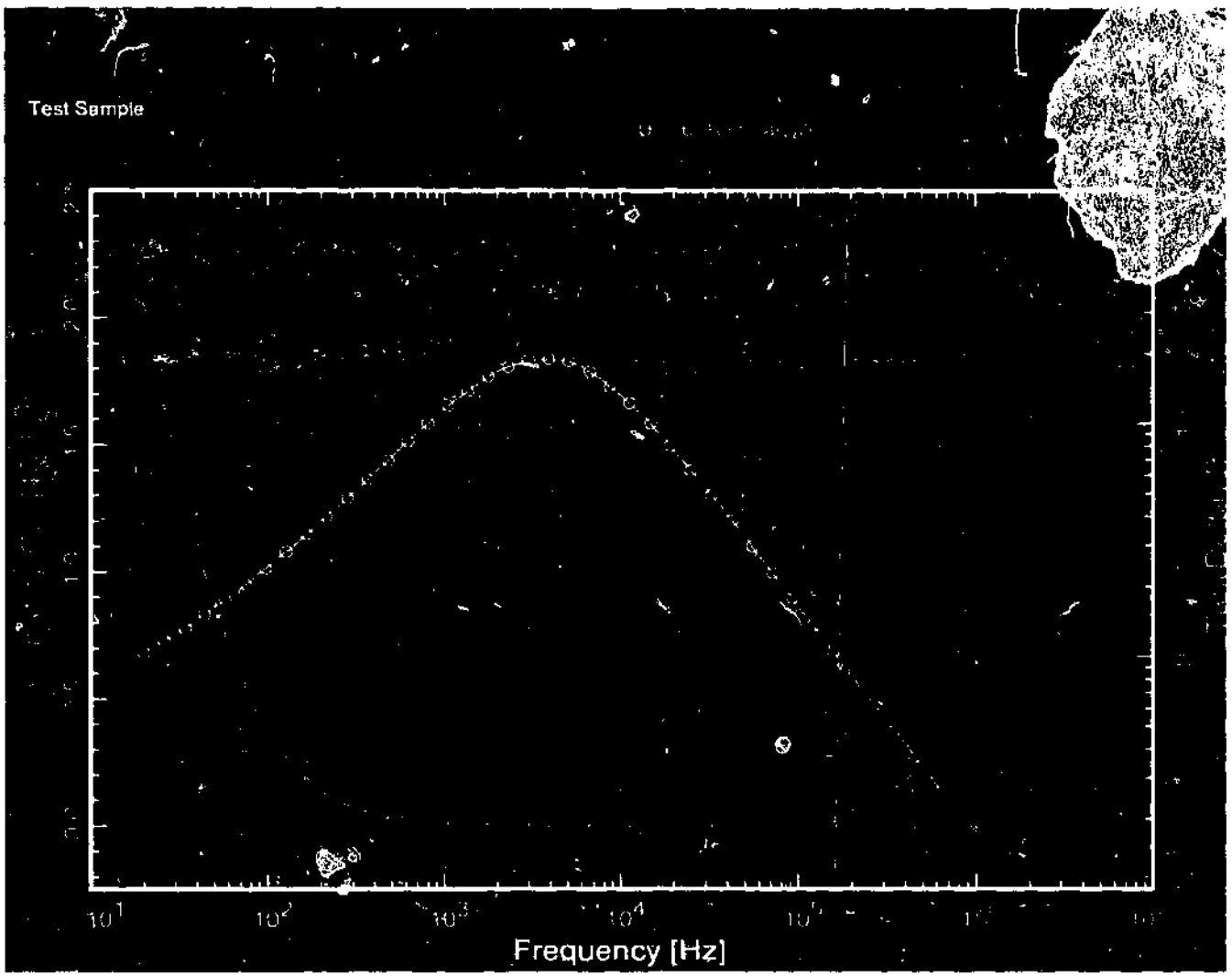


Figure 5.2.15 Example of Proline amino acid aqueous solution. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of Tan (Delta) values in a frequency range (green line).

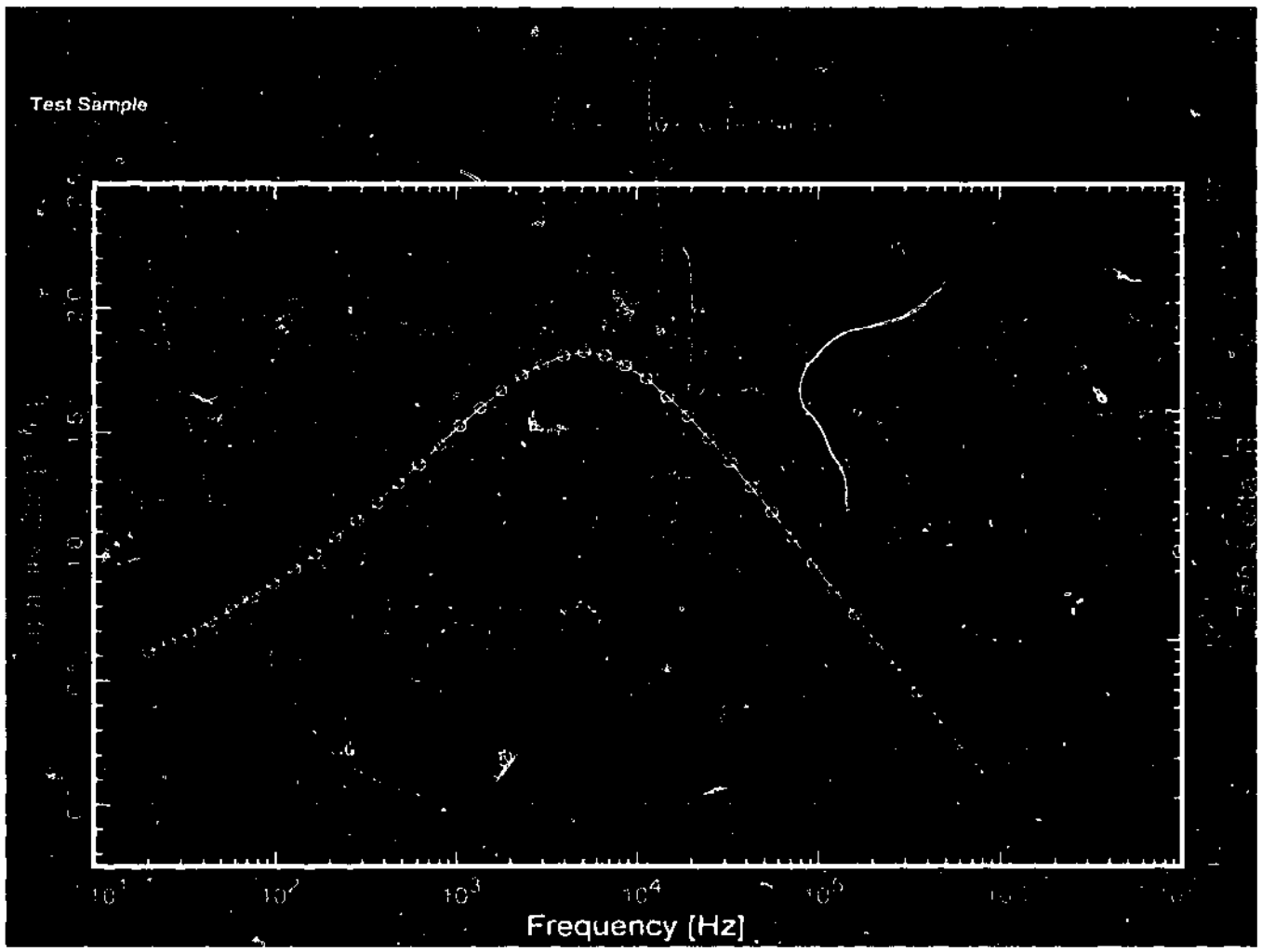


Figure 5.2.16 Example of Serine amino acid aqueous solution. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of Tan (Delta) values in a frequency range (green line).

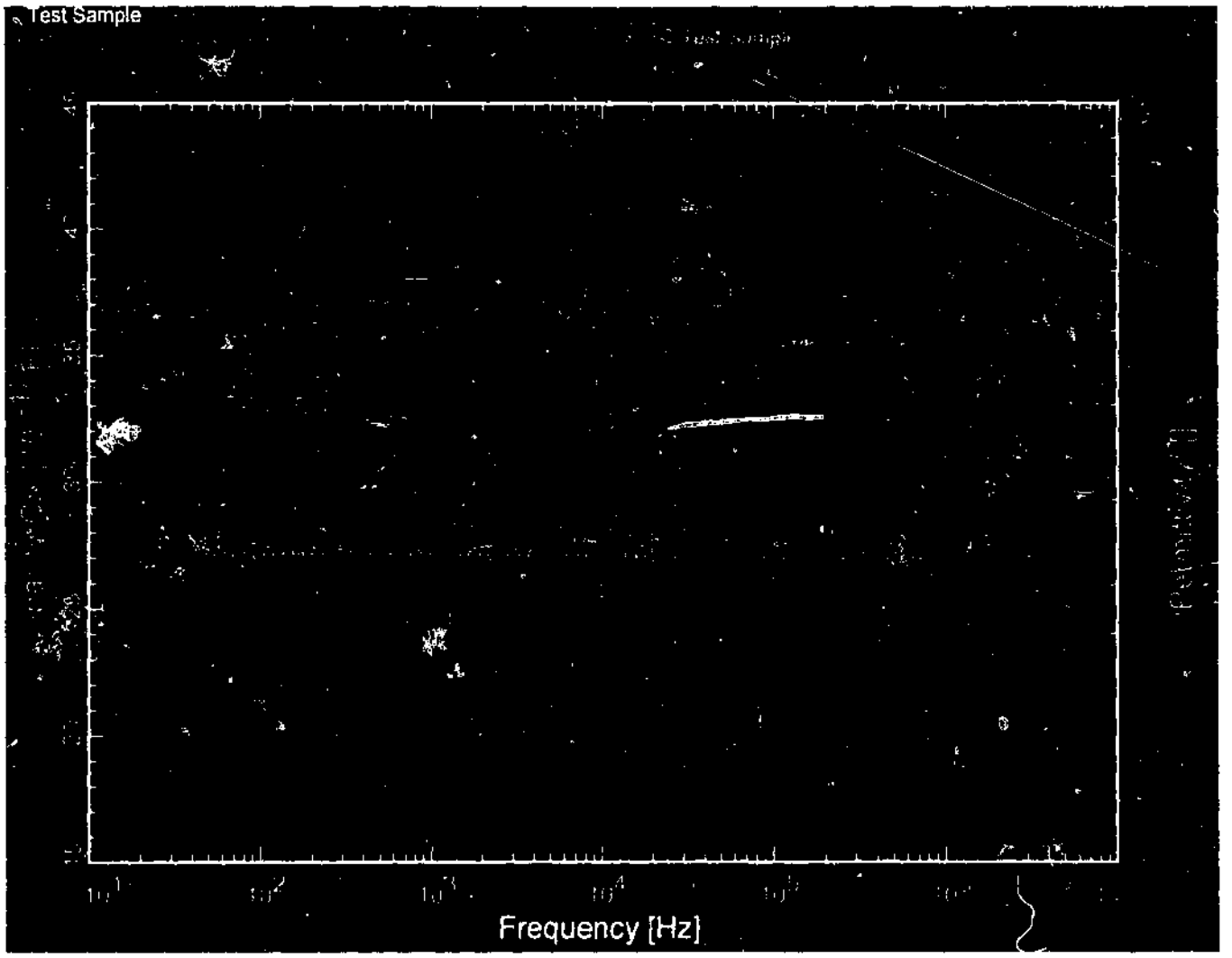


Figure 5.2.17 An empty cell. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

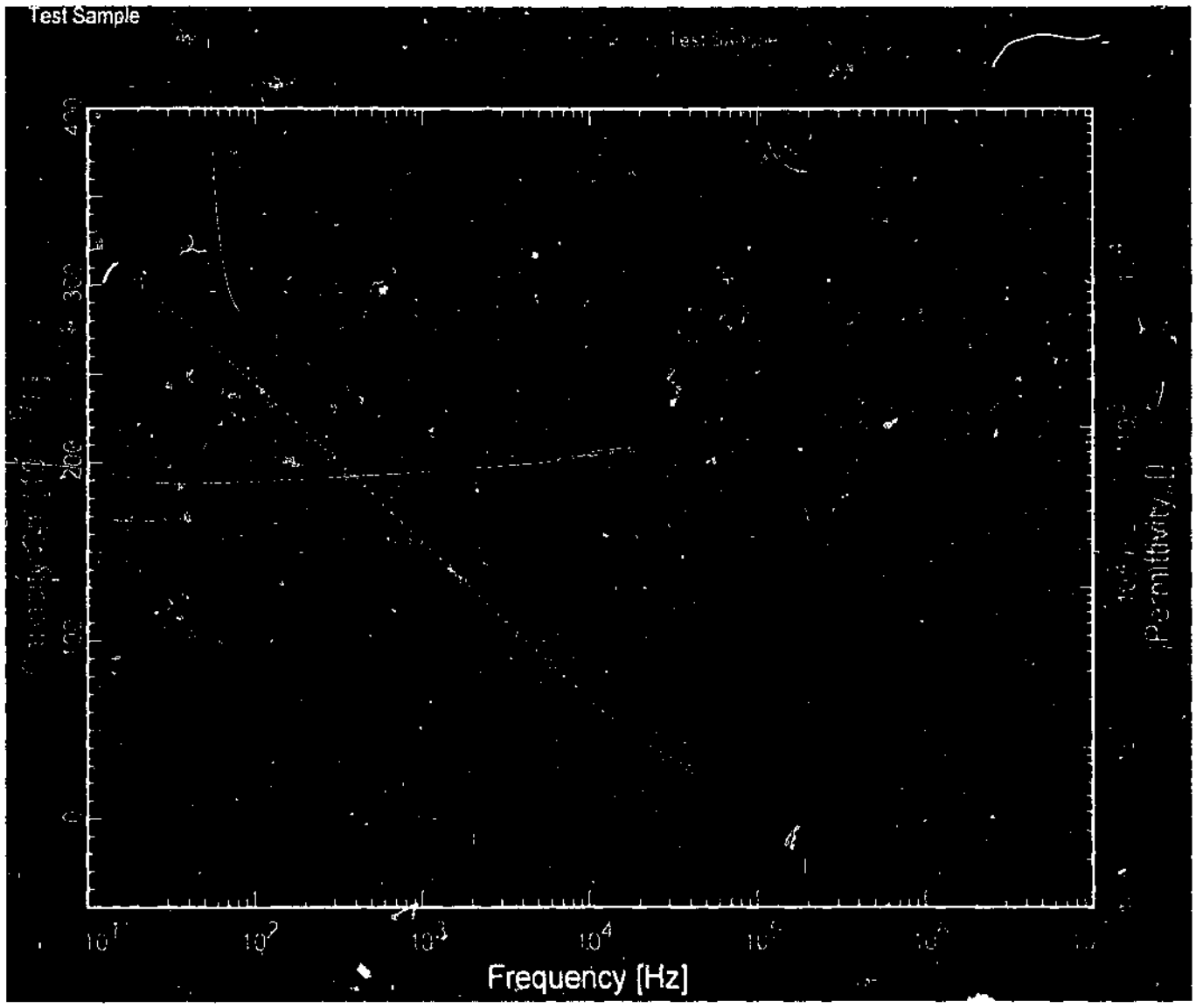


Figure 5.2.18 Example of distilled water. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

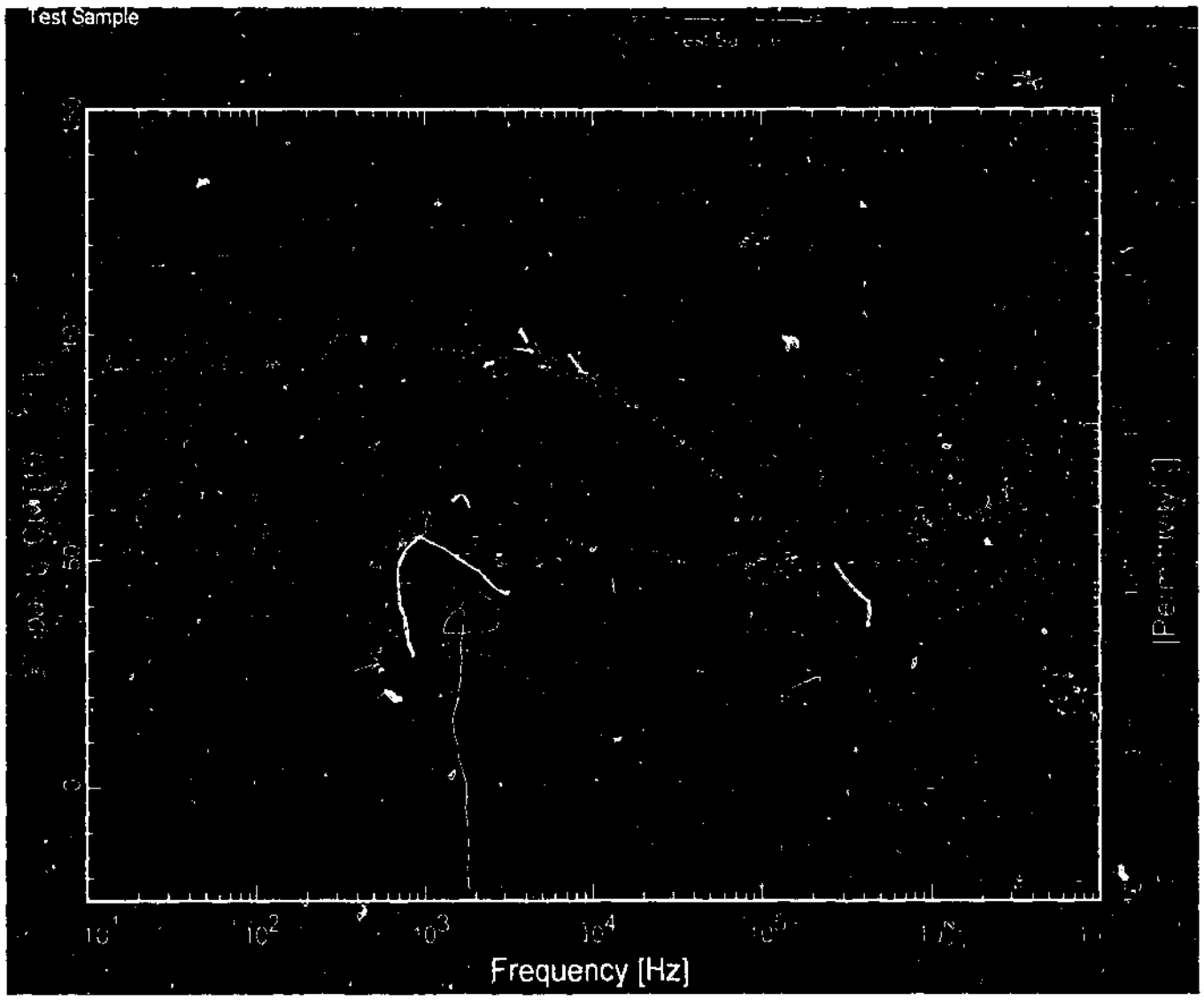


Figure 5.2.19 Example of Glutamic acid dissolved in 1M HCL. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

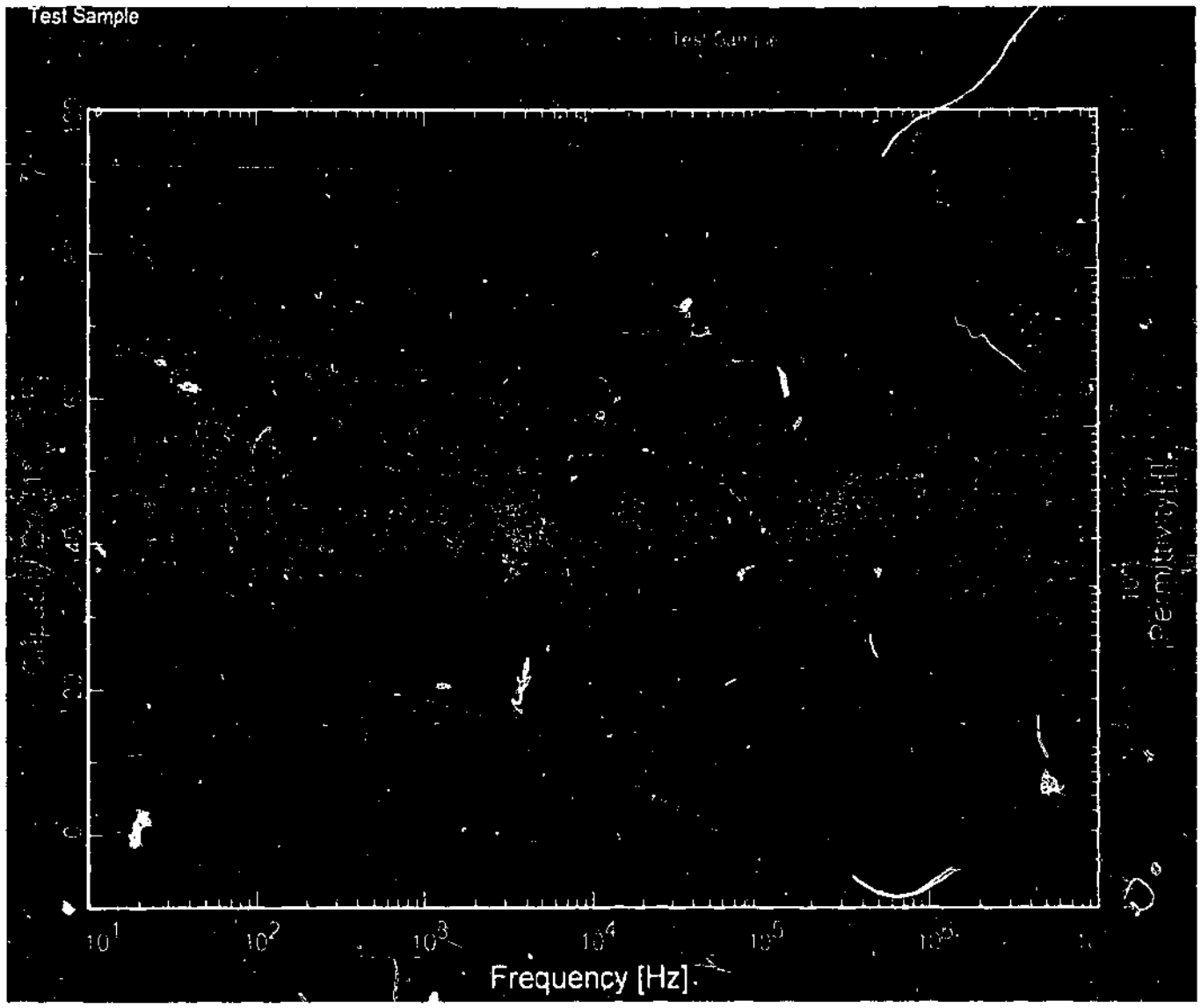


Figure 5.2.20 Example of Tyrosine amino acid dissolved in 1M HCL. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

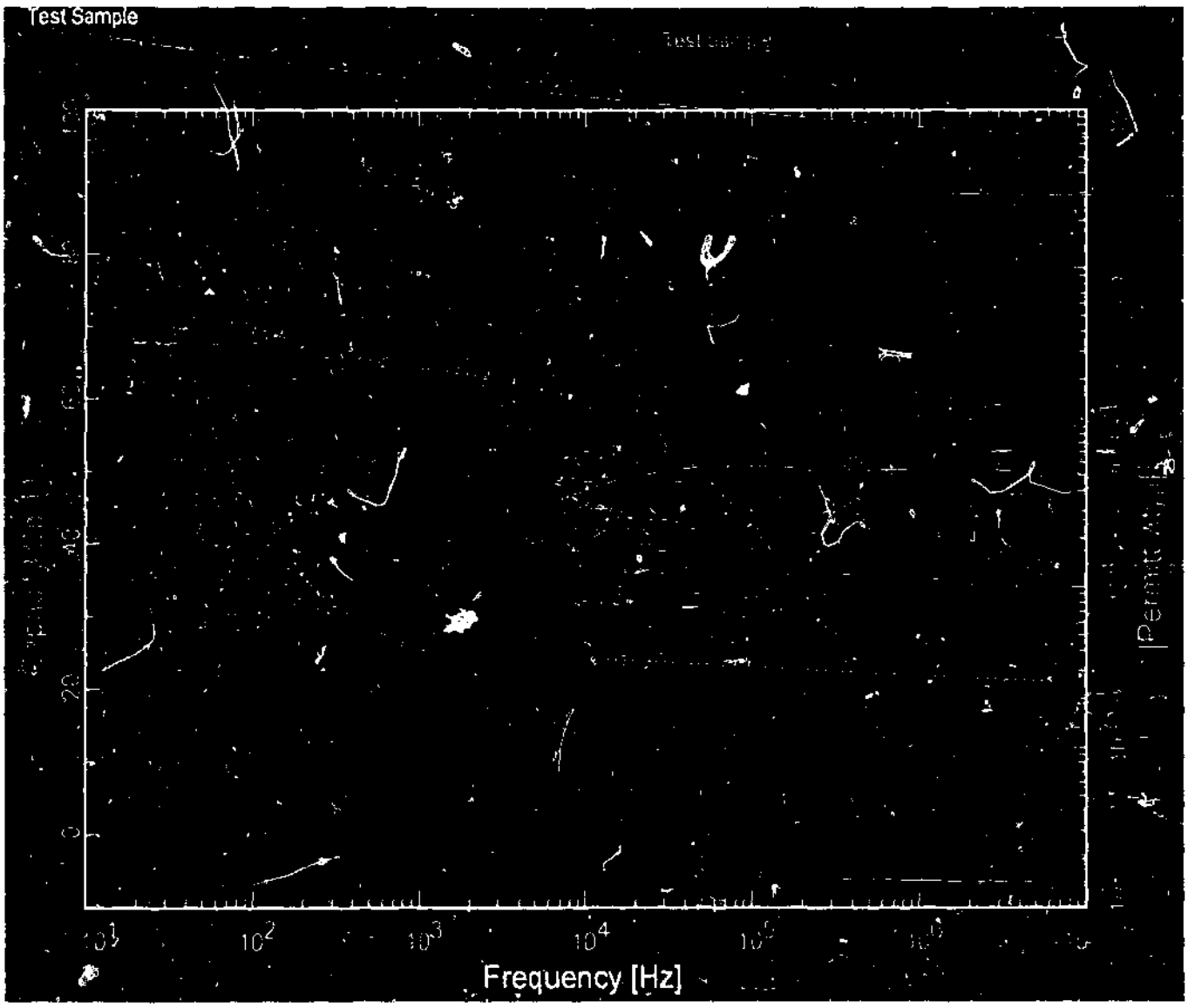


Figure 5.2.21 Example of Leucine amino acid dissolved in 1M HCL. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

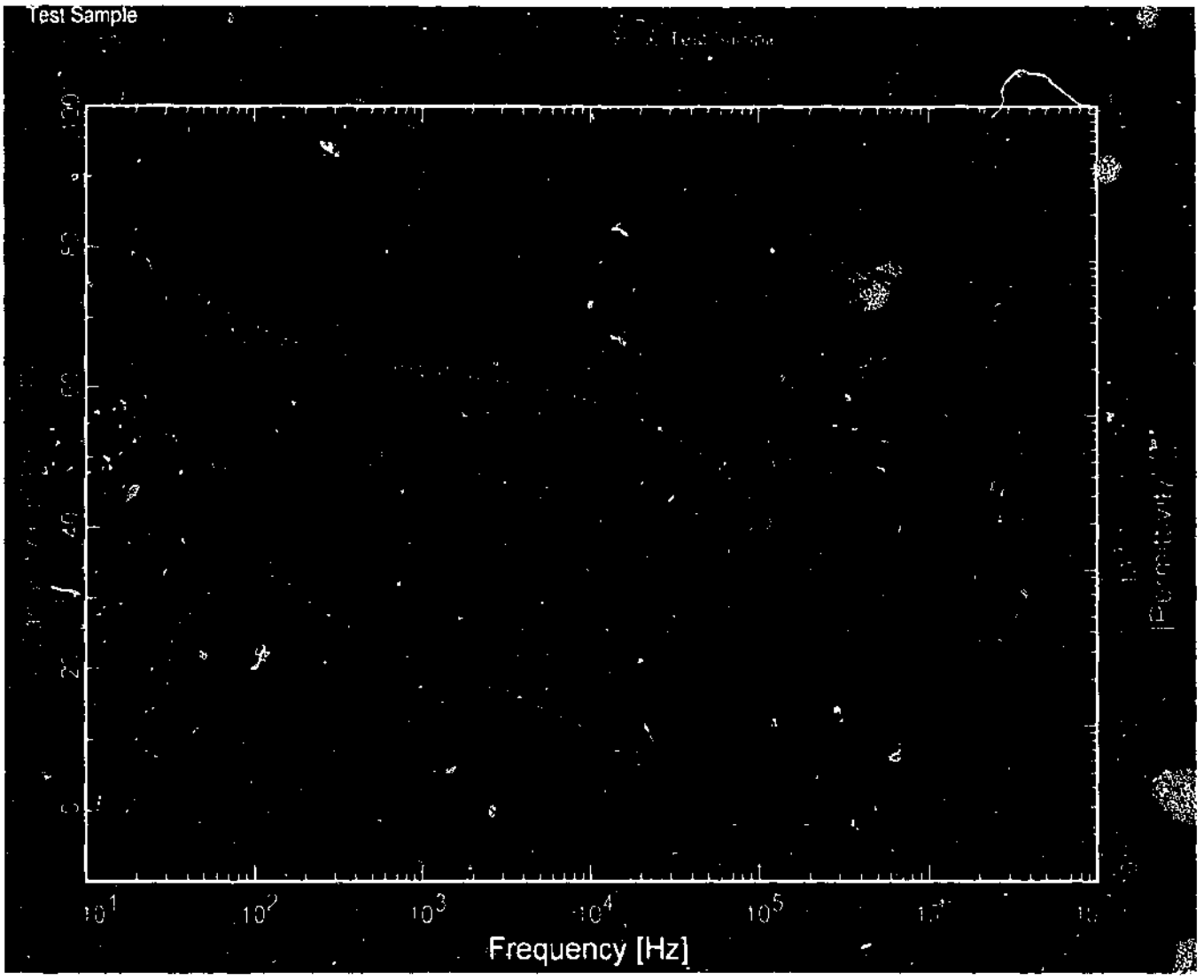


Figure 5.2.22 Example of Isoleucine amino acid dissolved in 1M HCL. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

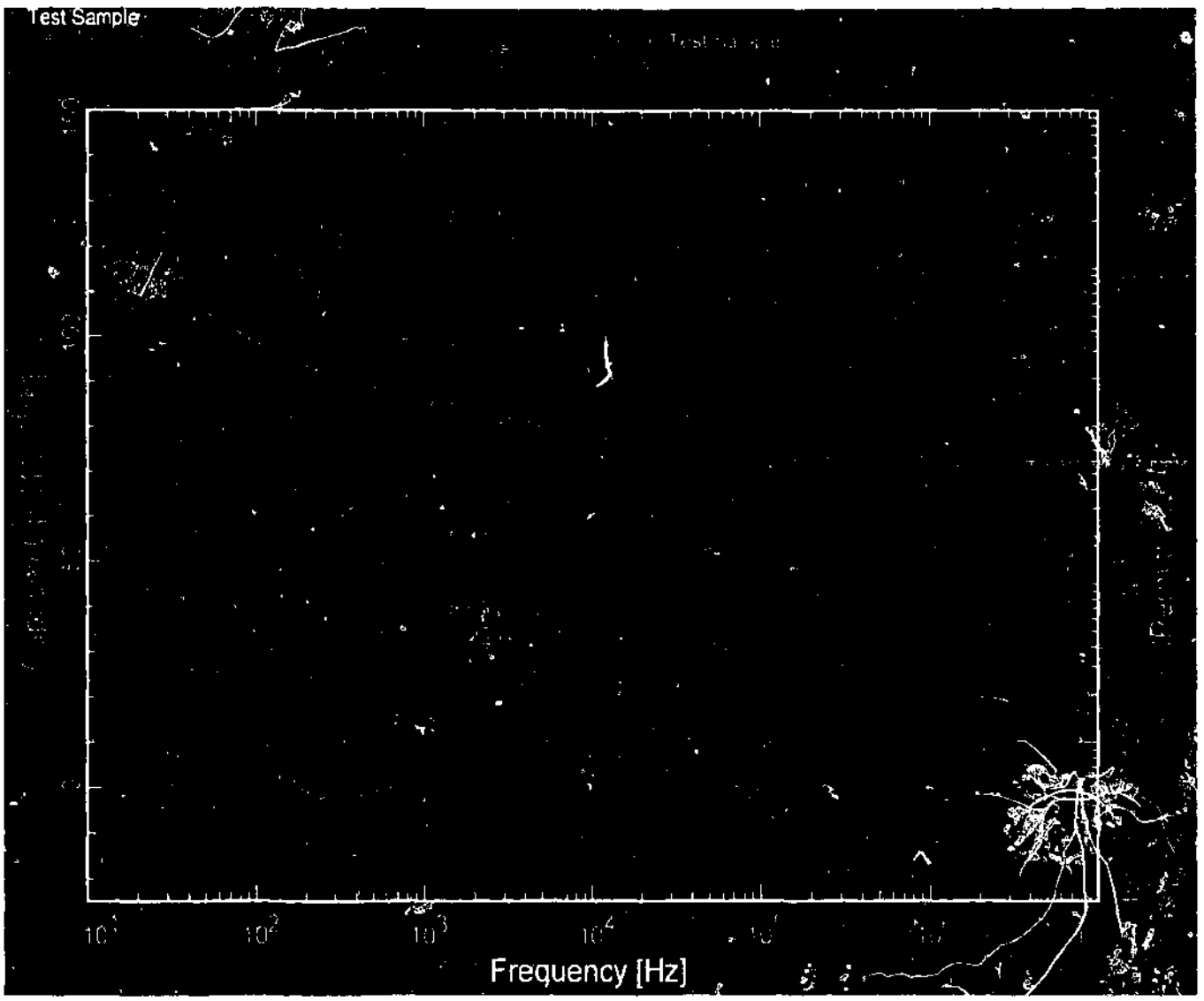


Figure 5.2.23 Example of Phenylalanine amino acid dissolved in 1M HCL. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

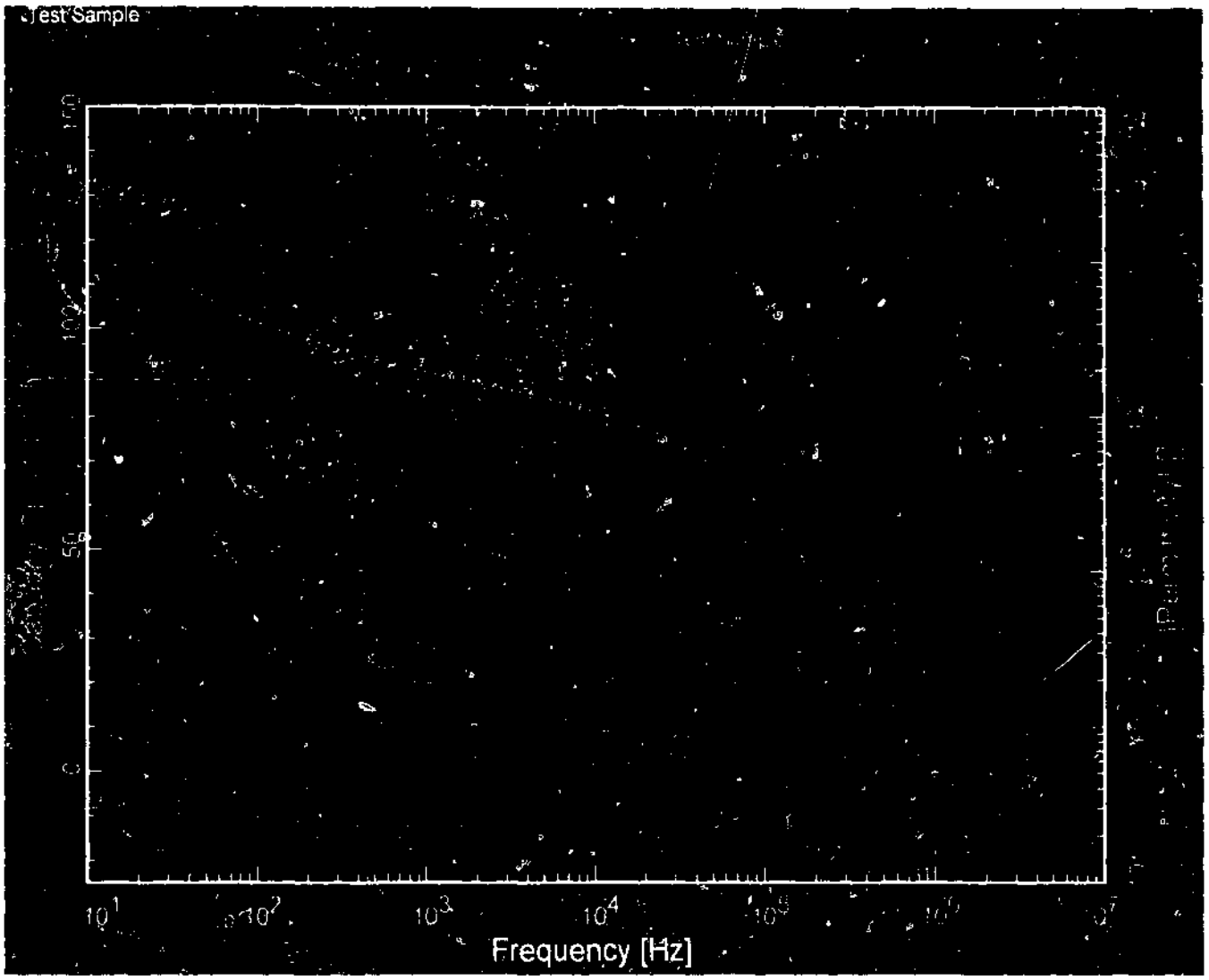


Figure 5.2.24 Example of Aspartic acid dissolved in 1M HCL. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

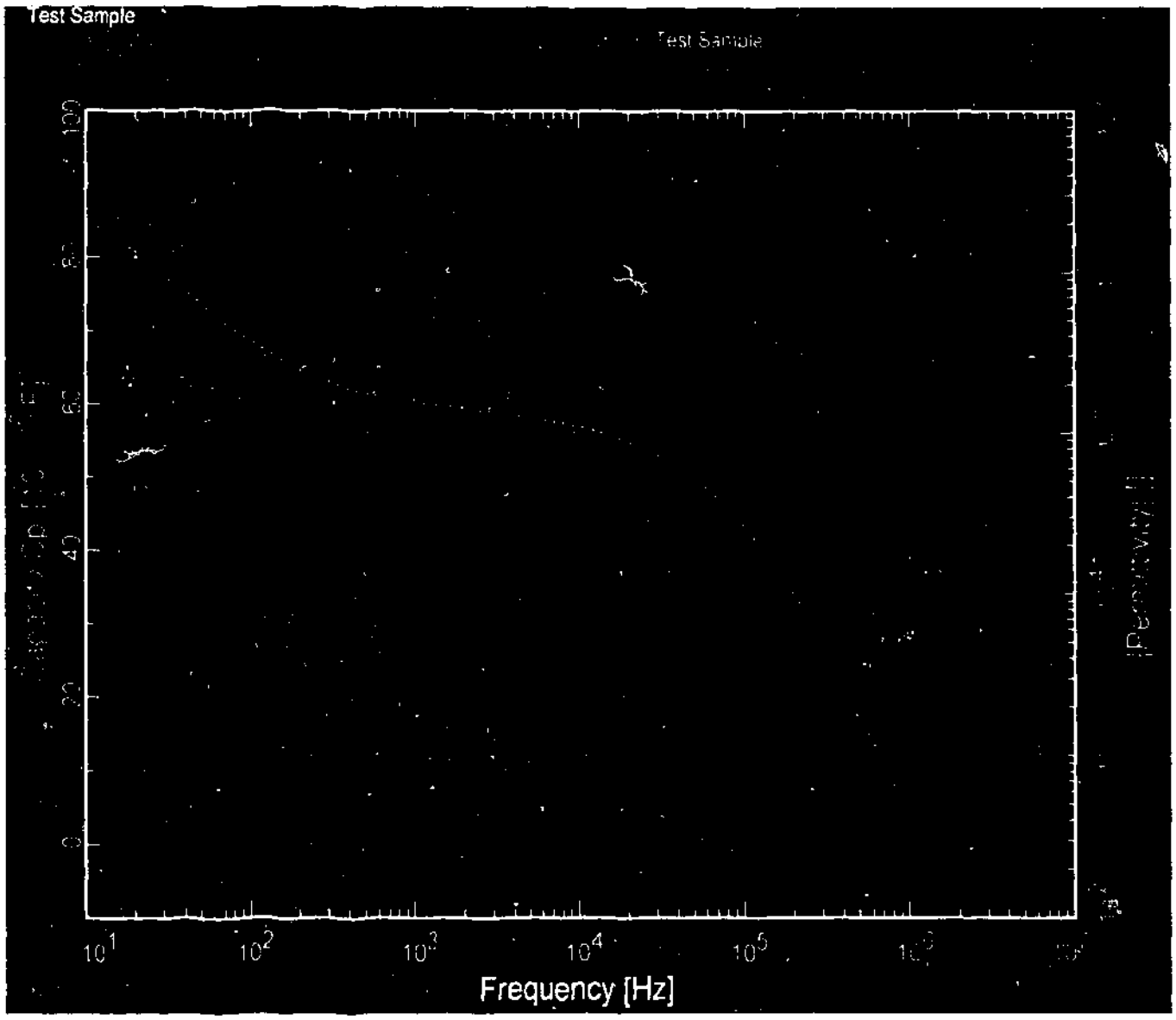


Figure 5.2.25 Example of 1M HCL. The change of the $C_{p\ sample}$ in a frequency range (red line), the change of the ε'_{sample} in a frequency range (green line).

Identification of the RRM characteristic frequencies using the new measured dielectric constant of amino acids parameter

As we mentioned above after completing the experimental measurements of capacitance and conductance of prepared amino acid solutions, we have calculated then the experimental values of new parameters: **dielectric constant (ε')** and **dielectric loss (tan δ)** for each particular amino acid sample as follows:

- **calculation of ε' of the sample**

$$\varepsilon'_{sample}=C_{psample}/C_o, \tag{5.2.4}$$

where C_o capacitance of the empty cell

- **calculation of ε' of distilled water and ε'of HCL respectively**

$$\varepsilon'_{water}=C_{pwater}/C_o \tag{5.2.5}$$

$$\varepsilon'_{HCL}=C_{pHCL}/C_o \tag{5.2.6}$$

- **calculation of Δε' amino acid for aqueous solutions**

$$\Delta\varepsilon'_{amino\ acid}=\varepsilon'_{sample}-\varepsilon'_{water} \tag{5.2.7}$$

and calculation of Δε' amino acid for HCL solutions

$$\Delta\varepsilon'_{amino\ acid}=\varepsilon'_{sample}-\varepsilon'_{HCL} \tag{5.2.8}$$

- **calculation of tan δ of the sample**

$$\tan\delta=(G_{sample}/\omega)/C_{psample} \tag{5.2.9}$$

where ω is an angular frequency, $\omega=2\pi$

The calculations of correlation coefficients were carried out as well to find out how significantly the EIIP and obtained new parameter values are correlated. The EIIP, dielectric constant and dielectric loss parameter values and their correlation coefficient are presented in Table 5.2.3.

The RRM was employed here for the structure/function analysis (determination of the characteristic profiles) of the proteins studied. Sequences from different functional groups: Glucagon, Lysozyme, FGF, Cytochrome C and EGF were investigated and results compared using the EIIP, $\Delta\varepsilon'_{\text{amino acid}}$ and tan δ parameters. To determine the characteristic frequencies of analyzed proteins a multiple cross-spectral analysis was performed for each protein group using the selected parameter values.

Table 5.2.3 EIIP and ε' Parameters and their Correlation Coefficient.

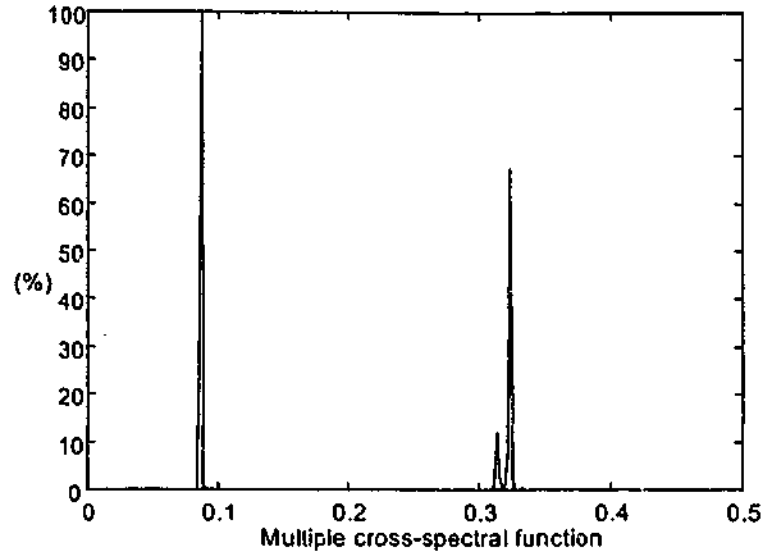
Amino Acid	EIIP	$\Delta\varepsilon'_{\text{amino acid}}$	$\tan\delta_{\text{sample}}$
L	0.0000	0.4209	21.861
I	0.0000	0.1266	23.748
N	0.0036	0.1650	10.290
G	0.0050	0.1085	13.755
V	0.0057	0.0804	14.836
E	0.0058	0.7251	20.555
P	0.0198	0.0561	21.065
H	0.0242	0.2131	7.881
K	0.0371	0.3249	5.471
A	0.0373	0.0550	22.465
Y	0.0516	0.1437	27.465
W	0.0548	0.0603	15.953
Q	0.0761	0.2034	7.869
M	0.0823	0.1406	14.446
S	0.0829	0.0519	17.581
C	0.0829	0.1266	16.656
T	0.0941	0.0643	20.130
F	0.0946	0.5735	15.142
R	0.0959	0.4351	5.831
D	0.1263	0.6634	15.567
Correl. Coefficient with EIIP.		0.1984	-0.1871

As a result of this analysis protein characteristic frequencies for each protein functional group were obtained. The peak frequency and S/N values for each protein group are shown in Table 5.2.4.

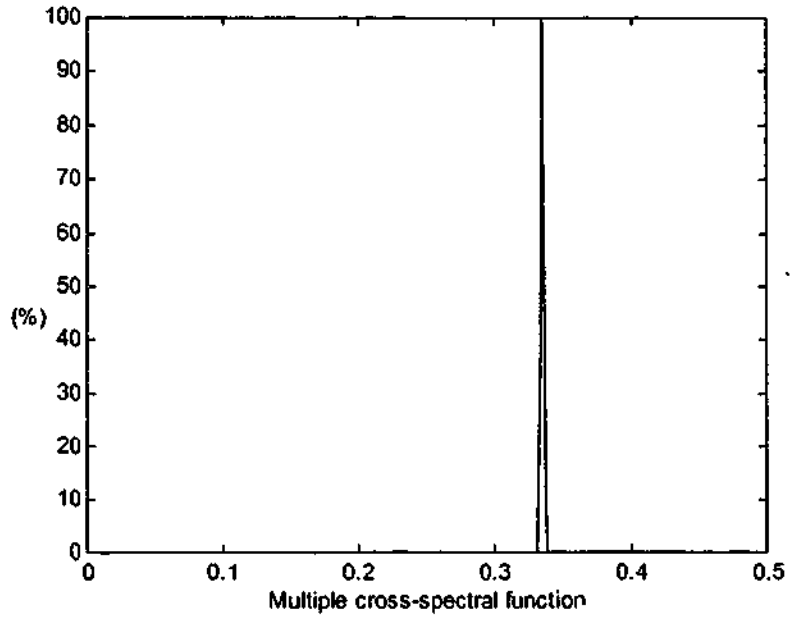
Table 5.2.4 Peak frequency and Signal-to-Noise Ratio Values for Protein Groups

Protein Group	EIIP		$\Delta\varepsilon'_{\text{amino acid}}$		$\tan\delta_{\text{sample}}$	
	Frequency	S/N	Frequency	S/N	Frequency	S/N
Glucagon	0.0879	122.7	0.3359	183.0	0.3789	205.8
Lysozyme	0.3281	238.3	0.0645	197.5	0.2520	250.6
FGF	0.4551	255.9	0.1055	147.4	0.0781	211.6
Cytochrome C	0.4746	247.4	0.2773	227.3	0.0645	195.5
EGF	0.0605	245.7	0.1230	102.4	0.4727	241.7
Myoglobin	0.2539	255.4	0.3066	253.0	0.0586	255.1

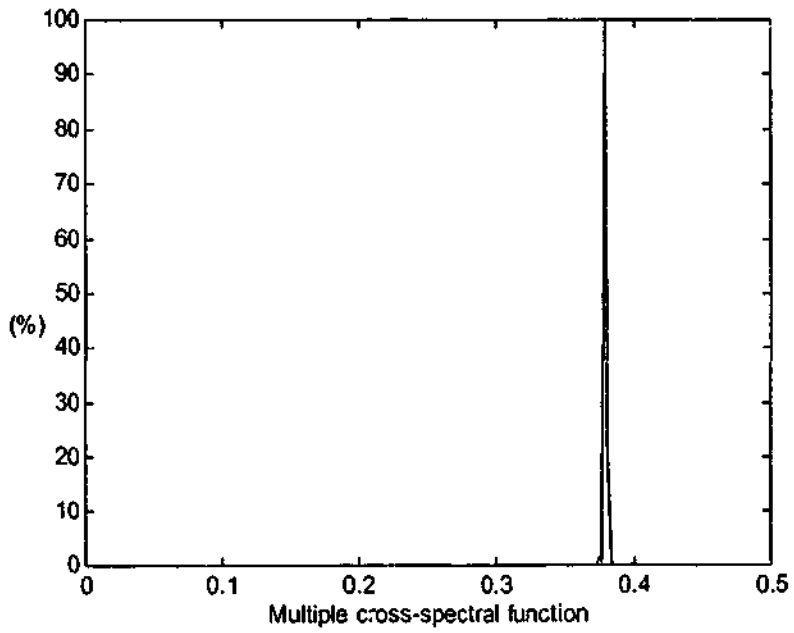
Consensus spectra for each selected protein group using the corresponding values of comparable EIIP, $\Delta\varepsilon'_{\text{amino acid}}$ and tan δ parameters are presented in Figure 5.2.3-5.2.8.



(a)

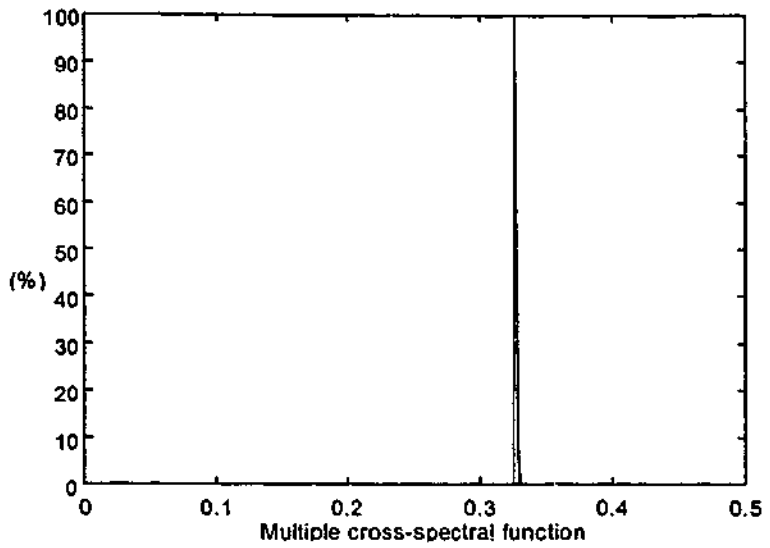


(b)

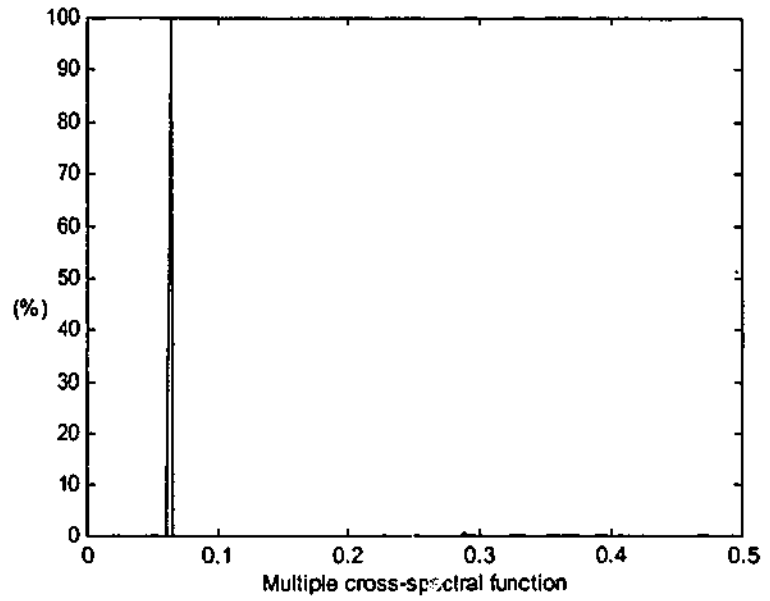


(c)

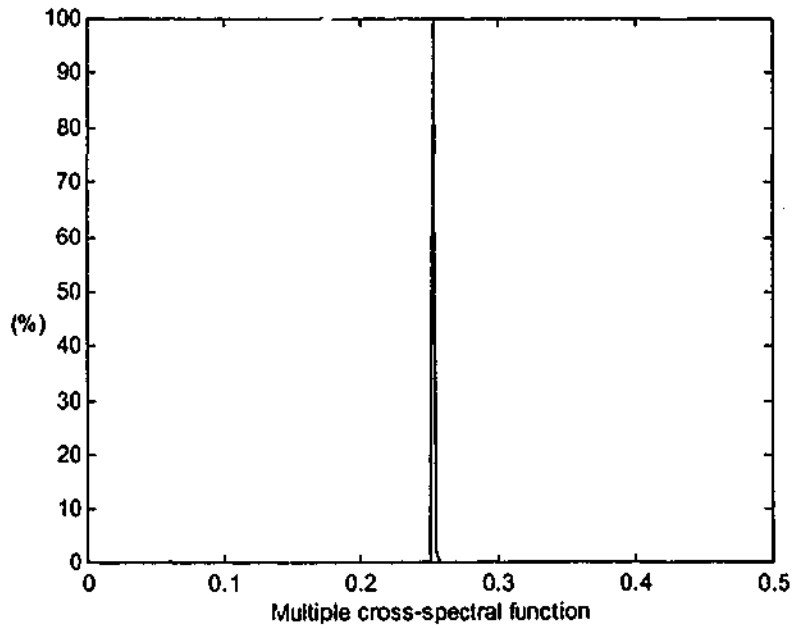
Figure 5.2.3 Multiple cross-spectral function of Glucagon proteins using (a) EIIP, (b) $\Delta\varepsilon'_{\text{amino acid}}$ and (c) tan δ parameters.



(a)

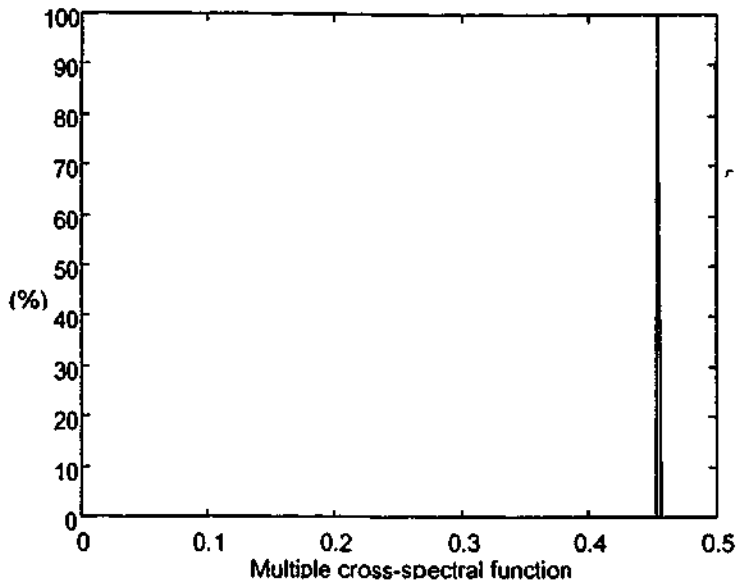


(b)

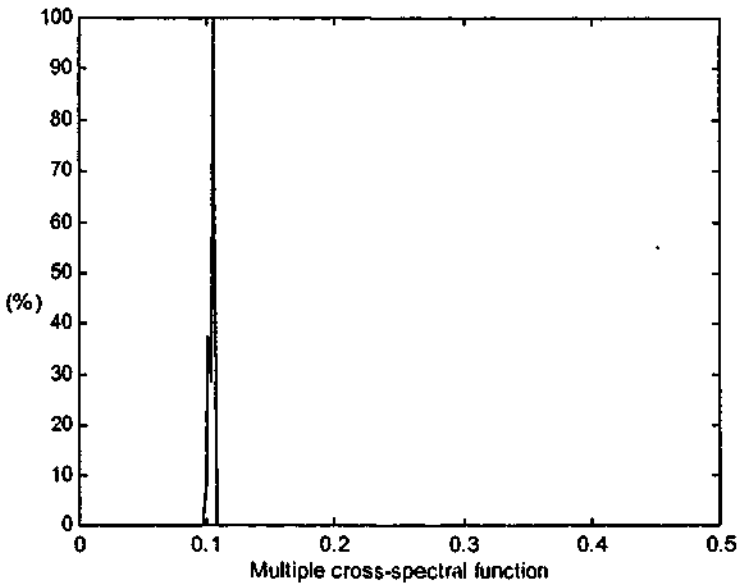


(c)

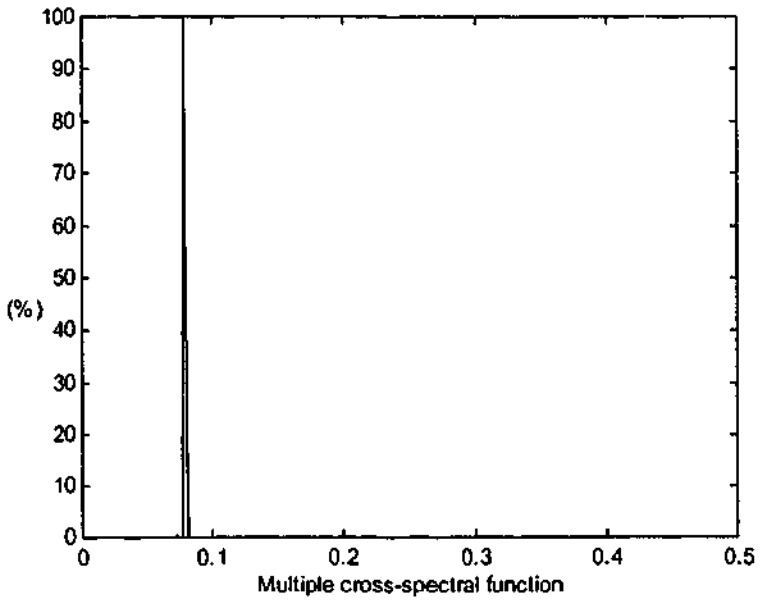
Figure 5.2.4 Multiple cross-spectral function of Lysozyme proteins using (a) EIIP, (b) $\Delta\varepsilon'_{\text{amino acid}}$ and (c) tan δ parameters.



(a)

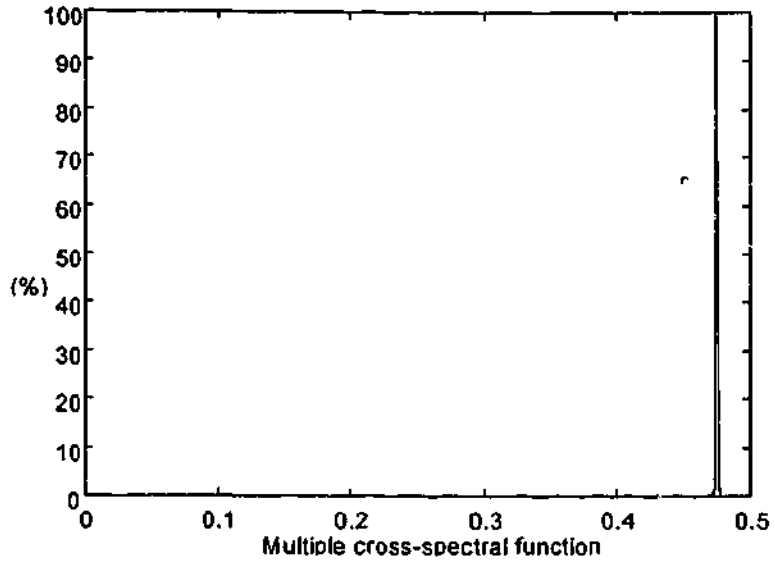


(b)

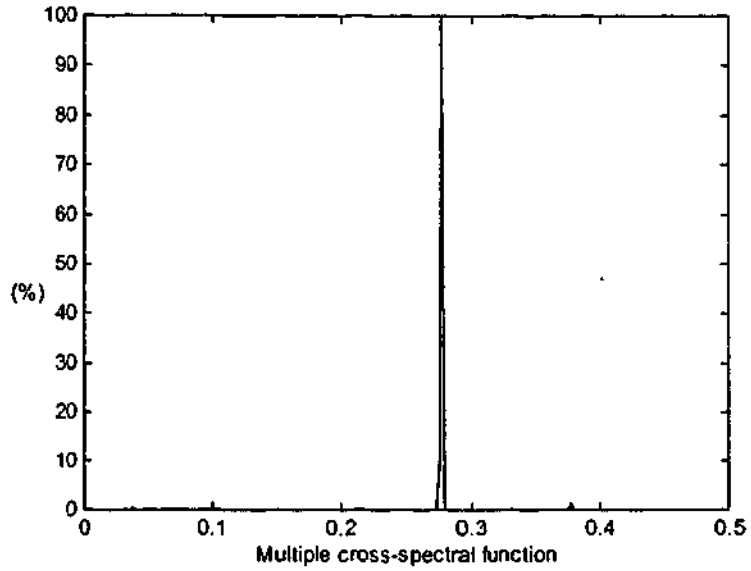


(c)

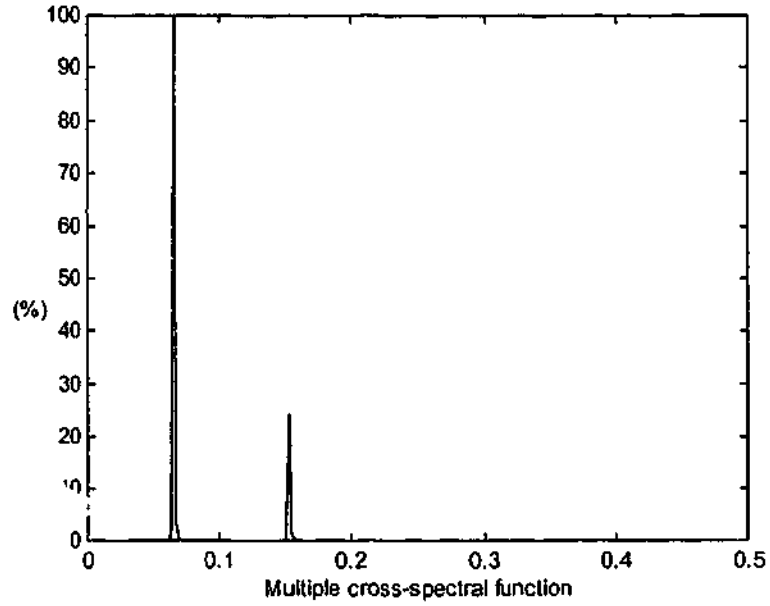
Figure 5.2.5 Multiple cross-spectral function of FGF proteins using (a) EIIP, (b) $\Delta\varepsilon'_{\text{amino acid}}$ and (c) tan δ parameters.



(a)

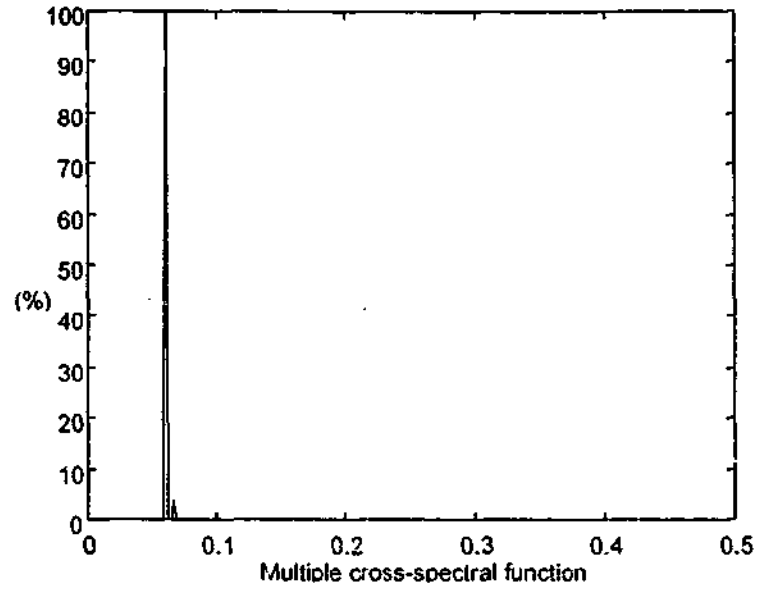


(b)

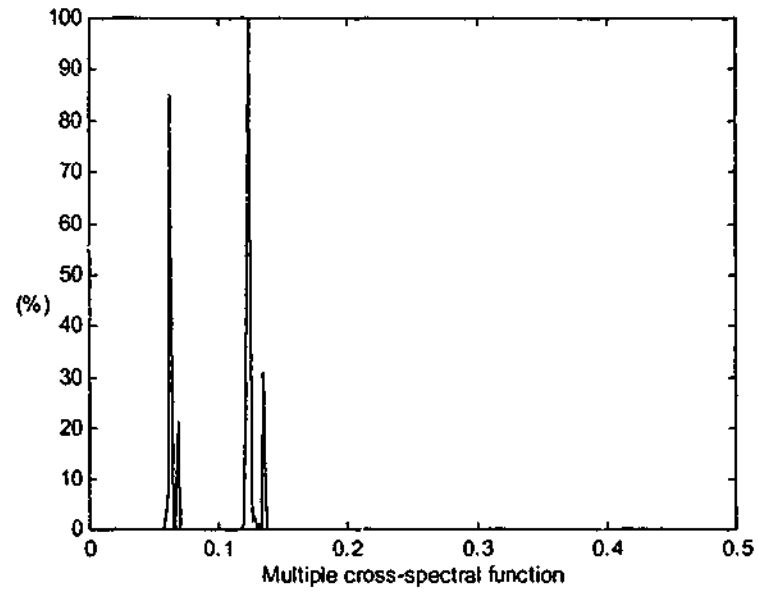


(c)

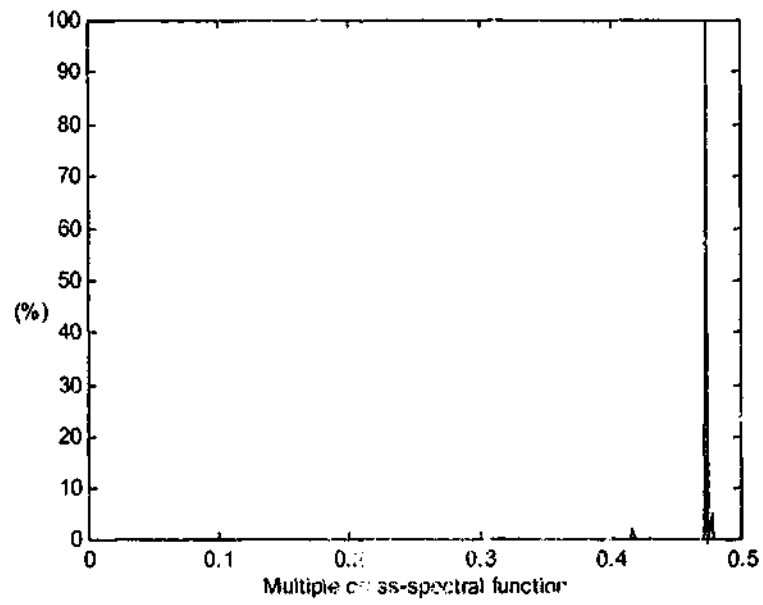
Figure 5.2.6 Multiple cross-spectral function of Cytochrome C proteins using (a) EIIP, (b) $\Delta\varepsilon'_{\text{amino acid}}$ and (c) tan δ parameters.



(a)

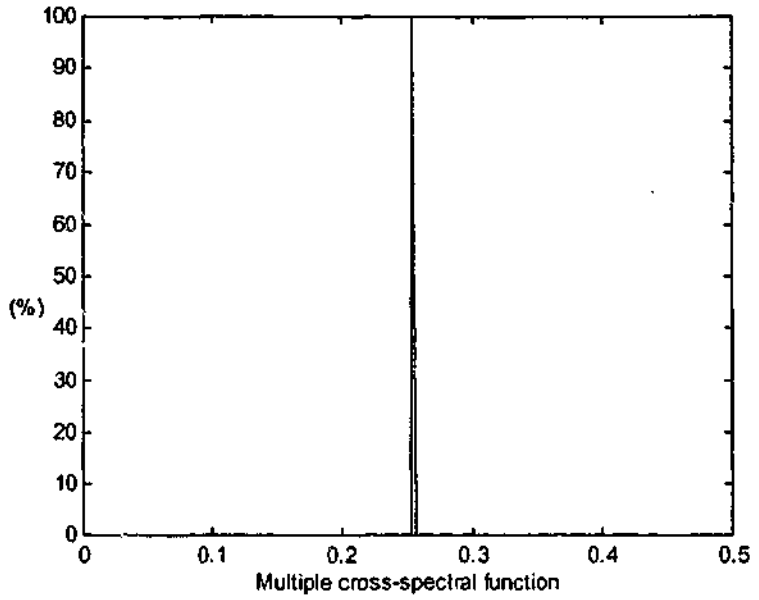


(b)

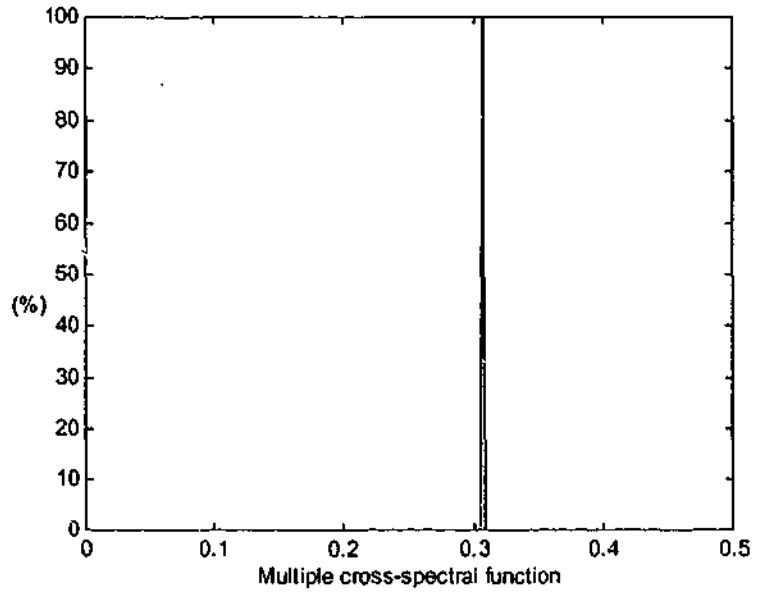


(c)

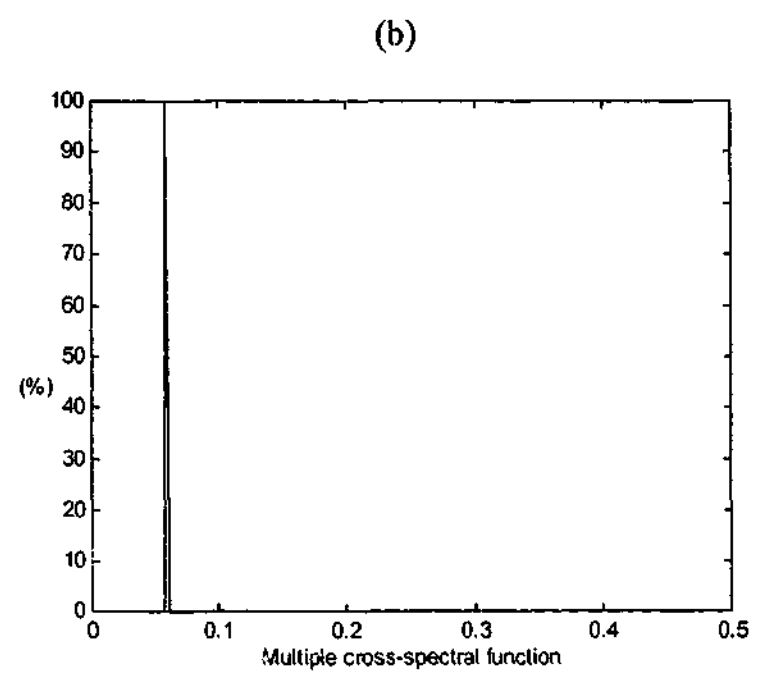
Figure 5.2.7 Multiple cross-spectral function of EGF proteins using (a) EIIP, (b) $\Delta\varepsilon'_{\text{amino acid}}$ and (c) tan δ parameters.



(a)



(b)



(c)

Figure 5.2.8 Multiple cross-spectral function of Myoglobin proteins using (a) EIIP, (b) $\Delta\varepsilon'_{\text{amino acid}}$ and (c) tan δ parameters.

Based on results of this study it should be noted that the RRM frequencies obtained are different for all analyzed protein groups using the EIIP, $\Delta\varepsilon'_{\text{amino acid}}$ and tan δ parameters (Table 5.2.4). This is expected, as comparable parameters are not significantly correlated. The correlation coefficient between the EIIP and the $\Delta\varepsilon'_{\text{amino acid}}$ is 0.1984 and the correlation coefficient between the EIIP and the tan δ parameters is -0.1871. However, it could be observed from the Fig.5.2.3-5.2.6, 5.2.8 that the $\Delta\varepsilon'_{\text{amino acid}}$ and tan δ parameters generate only one prominent peak in the consensus spectra of Glucagon, Lysozyme, Hemoglobin, Cytochrome C and Myoglobin proteins. This corresponds that all analyzed sequences within the protein group share the same biological function (satisfaction to criteria A, page 41). Thus, this frequency determined within the RRM denotes the specific biological function of studied proteins. It is interesting to note that for EGF protein group we observe three characteristics frequencies using the $\Delta\varepsilon'_{\text{amino acid}}$ parameter. Accordingly to criteria C (page 41) the peak frequencies are different for different biological functions. Thus, the $\Delta\varepsilon'_{\text{amino acid}}$ parameter allows us to detect not only one but three different biological functions for EGF proteins. Also it is impotant (Fig. 5.2.7) that among these three frequencies detected by the $\Delta\varepsilon'_{\text{amino acid}}$ parameter, one is the same as detected by the EIIP parameter. But these frequencies have different amplitude ratios. This defference in the amplitude ratio values explains why not the same frequencies were determined using the EIIP and $\Delta\varepsilon'_{\text{amino acid}}$ parameters.

Following the aim of the analysis of the possible use of the $\Delta\varepsilon'_{\text{amino acid}}$ and tan δ parameters for structure-function predictions within the RRM, we conclude that these parameters could be suitable for the determination of characteristic profiles of studied proteins. The parameters allow us to detect the particular RRM frequency characterized the specific biological function of the protein studied, whilst satisfying the required RRM criteria. However, the further analysis of other protein families as well as the "hot spots" analysis and the peptides'design within the RRM model would be needed to prove the suitability of the dielectric constant and loss parameters for this approach.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSIONS

Probably the most important problem in life science and bioengineering today is the understanding of the relationships between the primary sequence, the higher order structure (3-D structure) and the function of proteins. An application of the RRM approach for the analysis of different protein families with the aim of identifying the characteristic pattern implied within a protein sequence, which could be linked directly to the protein's structural and functional properties, has been investigated.

It has been previously shown that the EIIP, which is a basic parameter in the RRM model, presents a unique amino acid property that can be used for the analysis of different unrelated protein families. Although the EIIP was successfully used in our previous research there was many criticism against the artificial nature of this parameter as the EIIP was mathematically calculated using the following approximations:

Therefore, one the main aims of this study was the selection of the parameters representing different amino acid properties that are most suitable to represent the common biological behavior of the proteins studied and as well suitable for the use within the RRM approach instead of the EIIP or to understand and explain the physical meaning of the EIIP using related measurable amino acid properties.

The common procedure used for the selection of the amino acid parameters in the RRM involves two steps analysis:

1. The parameter generates in the consensus spectrum one dominant peak corresponding to a common biological activity of the studied proteins.

2. This obtained RRM characteristic frequency is related to the biological function provided the following criteria are met:

- **A:** One peak only exists for a group of protein sequences sharing the same biological function

- **B:** No significant peak exists for biologically unrelated protein sequences

- **C:** Peak frequencies are different for different biological functions

General conclusions are now drawn accordingly to the characteristics of the results achieved in previous chapters:

♦The IC parameter has been used in the RRM for the determination of the characteristic frequencies corresponding to the common biological activity of the following protein groups: glucagons, lysozymes, hemoglobins, cytochromes C, EGFs, FGFs, TNFs, TGFs, PDGs, TNFs, interleukins, myoglobins, viral oncogenes, proto-oncogenes, oncogenes, p53s, cysteine proteases, metallo-proteases, serine proteases, aspartic proteases, trypsins, protease inhibitors and AIDS related proteins. Comparable analysis of the characteristic frequencies obtained using the EIIP and IC parameter was performed as well. Results obtained reveal the suitability of the IC parameter to identify the biological profiles of the studied proteins.

♦Amino acid parameters: **P001**- α-CH chemical shifts, **H085**-Localized electrical effect, **H371**- Normalized frequency of chain reversal D and **A**- helix coil equilibrium constant have been chosen from the existing Amino Acid Index Database. These selected amino acid properties and their different linear combinations have been investigated within the RRM. As a result RRM characteristic frequencies were obtained for the following proteins: Cytochrome C, Lysozyme, Hemoglobin, Glucagon, TNF, EGF, FGF, Interleukin and Myoglobin. All mentioned above parameters are suitable for the use in structure-function analysis of proteins within the RRM as far as they generate one prominent peak in the consensus spectrum corresponding to the common biological behavior of sequences analyzed and satisfy the RRM criteria. Furthermore, the new computational parameter ICEE is introduced

and recommended for the use in further RRM analysis for the active sites predictions and the identification of the "hot spots".

♦The Dielectric constant (ε') and Dielectric loss (tan δ) parameters have been tested for its possible use in the RRM. These parameters are based on the values of capacitance and conductance obtained experimentally for 20 amino acids using the dielectric spectroscopy method. Results obtained satisfy the RRM criteria. This satisfaction allows us to conclude that the introduced parameters are suitable for the use in the RRM analysis.

♦Finally, the main conclusion is that the results of the whole study could be considered as a great support for the EIIP parameter. Although the EIIP is a mathematically calculated parameter, the physical nature of the EIIP, distribution of the energy of free electrons transfered along the protein, is very similar to the physical nature of the parameters selected from the Amino Acid Index Database, which values were experimentally measured by other authors. All investigated IC, P001, H085, H371 and ICEE parameters are strongly correlated with the EIIP. They describe the various amino acid properties related to the process of the charge moving through the protein backbone. Thus, the main concept of the RRM approach, based on the assumption that interactions between biomolecules are electromagnetic in nature, has been confirmed. As very similar results were obtained using the EIIP and ICEE (ICEE=IC-EE) parameteres, it leads us to the conclusion that the physical basis of the EIIP parameter is related to both the Ionization Constant (IC) and Localized electrical effect (EE=H085) parameters. This means that the EIIP's nature is relevant to the amino acid equilibrium and the localized electrical effect.

With the knowledge of the RRM characteristic frequencies for particular functions, it is possible to predict the protein's active site(s), functional mutations and even to design bioactive peptides with the desired biological functions.

6.2 LIMITATIONS OF THE STUDY

The comparable analysis of the use of various amino acids properties for structure/function predictions of biological profiles of different proteins within the RRM presented in this study is the subject to the following limitation:

- The measurements of experimental values of capacitance and conductance of amino acid solutions have been undertaken using two different solvents distilled water and HCL for the preparation of analyzed solutions. It is recommended for future work to design the measurements using the same solvent environment for all 20 analyzed amino acids, for instance we can use HCL or prepare less concentrated aqueous solutions of analyzed amino acids (<0.5M).

- Though extensive protein examples have been chosen to develop the protein functionality classification scheme, based on the assigning to each amino acid the proper amino acid physical property value, the examples are still far from exhaustive. The inclusion of all protein functional groups would rise the computation load dramatically and would be far beyond the author's capabiiity. However, partial examples would adversely affect the confidence rate and make the scheme difficult to become fully functioning.

As it was mentioned above the possibilities of the RRM are: (1) to define a particular function of a protein, based on the characteristics frequency obtained; (2) to predict functionally important amino acids within the protein sequence and thus to propose effective mutations; (3) to analyze and predict the possibility of macromolecular interactions; (4) to design sequences with desired spectral and, sequentially, functional characteristics [6].

- This study has investigated the application of new parameters suitable for the protein analysis within the first step of the whole RRM approach. This step (1) is the identification of the RRM characteristic frequency. Thus, the study is not covered the steps (2) - (4) of the RRM analysis, which are the predictions of the amino acid active sites, interactions and peptide design, based on the new parameters proposing for the use instead of the EIIP.

- No attempt has been taken to design the biophysical test to confirm predicted RRM characteristics of the studied proteins with frequency characteristics obtained experimentally for certain light-induced biomolecules.

6.3 FUTURE WORK

Following the preliminary gains made through this study, particularly the finding of the optimal physicochemical amino acid parameters, and based on the limitations of this research, there are several other steps that could be undertaken to facilitate and refine the work done. Particularly, after assigning the new parameter value to each amino acid in the protein sequence we should perform such research:

1) "Hot Spots" analysis of identifications of essential amino acids corresponding to the biological function of the protein sequence;

2) Analysis and predictions of the possibility of the interactions between proteins and its targets;

3) Design of the new peptides based on the predicted characteristic frequency;

4) Comparable analysis of the 3-D structures of the studied protein and designed peptide.

As a result of this project an improved RRM model is proposed. The concepts presented in this model lead to a completely new perspective on life processes at the molecular level. This is a step forwards to better understanding of the physical nature of biomolecular interactions and biological processes. The RRM is capable to bring a number of practical advantages to the fields of molecular biology, biotechnology and agriculture.

APPENDIX A

LIST OF PUBLICATIONS

1. **E.Pirogova,** Q.Fang, E.Lazoura, I.Cosic, (1998) "Analysis of Amino Acids Parameters in the Resonant Recognition Model", Proc. of the 2nd International Conference on Bioelectromagnetism, 71-72.

2. I.Cosic, **E.Pirogova,** (1998) "Protein Structure/Function Analysis Using the Resonant Recognition Model (RRM) - Application to Oncogenes", Proc. of the 4th Australian Molecular Modelling Conference and Workshop,

3. I.Cosic, **E.Pirogova,** (1998) "Applications of Ionization Constant of Amino Acids for Protein Signal Analysis Within the Resonant Recognition Model", Proc. of 20th Annual International Conference of the IEEE EMBS, Vol.20, No.2, 1072-1075.

4. **E.Pirogova,** I.Cosic, (1999) "Optimisation of physico-chemical properties of amino acids in structure/function analysis of proteins", Conference of the Victorian IEEE EMBS, Proc. Chapter of the IEEE EMBS, 203-206.

5. Cosic, I., Fang, Q., **Pirogova, E.,** (1999) "Modification of the RRM model using wavelet transform and Ionisation constant to predict protein active sites", IEEE EMBS, 21, Vol.2, 1215-1217.

6. Cosic, I., **Pirogova, E.** (2000), "Usage of Ionization constant of amino acids for protein signal analysis within the RRM - Application to Oncogene", IEEE-EMBS Asia-Pacific Conference on Biomedical Engineering, 413-414.

7. Cosic, I., **Pirogova, E.** (2000) "Usage of Ionization Constant of amino acids for protein signal analysis within the RRM - Application to Proteases", Proc. of the 6th Australian Molecular Modelling Conference, 94.

8. **Pirogova, E.,** Cosic, I., (2001) "Examination of amino acid indices within the Resonant Recognition Model", Proc. of the 2nd Conference of the Victorian Chapter of the IEEE EMBS, 124-127.

9. **Pirogova, E.**, Cosic, I., (2001) "Development of new computational amino acid parameter for structure/functional analysis of proteins within the RRM", Proc. of 23rd Int. Conf. of the IEEE.

10. **Pirogova, E.**, Cosic, I., (2001) Journal "Molecular Simulation", "Usage of IC for protein signal analysis within the RRM- Application to Proteases", (full paper accepted for publication).

11. **Pirogova, E.,** Cosic, I., Simon, G., (2001) "Experimental measurements of dielectric and conduction properties of amino acids", Proc. of the 7th Australian Molecular Modelling Conf., 97.

12. **Pirogova, E.** Cosic, I. (2001) "Introduction of new computational amino acid parameter for the use in the RRM", CABIOS (in preparation).

13. **Pirogova, E.,** Cosic, I., Simon, G. (2001) "Dielectric and conduction properties of amino acid solutions", Materials Forum (in preparation).

APPENDIX B

List of 20 Amino Acids with the 3-letter and 1-letter Abbreviations

AMINO ACID	3-letter	1-letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

APPENDIX C

AMINO ACID PARAMETERS AND THEIR CORRELATION COEFFICIENTS WITH EIIP

An amino acid index is a set of 20 numerical values representing any of the different physicochemical and biological properties of amino acids. The Amino Acid Index Database is a collection of published indices together with the result of cluster analysis using the correlation coefficient as the distance between two indices. This section currently contains 402 indices.

List of 402 Amino Acid Indices with the Result of Clustering

A. alpha and turn propensities

A005 BEGF750101 Conformational parameter of inner helix (Beghin-Dirkx, 1975)

A007 BEGF750103 Conformational parameter of beta-turn (Beghin-Dirkx, 1975)

A016 BUNA790101 alpha-NH chemical shifts (Bundi-Wuthrich, 1979)

A018 BUNA790103 Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich, 1979)

A019 BURA740101 Normalized frequency of alpha-helix (Burgess et al., 1974)

A022 CHAM830101 The Chou-Fasman parameter of the coil conformation (Charton-Charton, 1983)

A023 CHAM830102 A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet (Charton-Charton, 1983)

A037 CHOP780101 Normalized frequency of beta-turn (Chou-Fasman, 1978a)

A038 CHOP780201 Normalized frequency of alpha-helix (Chou-Fasman, 1978b)

A040 CHOP780203 Normalized frequency of beta-turn (Chou-Fasman, 1978b)

A042 CHOP780205 Normalized frequency of C-terminal helix (Chou-Fasman, 1978b)

A044 CHOP780207 Normalized frequency of C-terminal non helical region (Chou-Fasman, 1978b)

A047 CHOP780210 Normalized frequency of N-terminal non b region (Chou-Fasman, 1978b)

A048 CHOP780211 Normalized frequency of C-terminal non b region (Chou-Fasman, 1978b)

A049 CHOP780212 Frequency of the 1st residue in turn (Chou-Fasman, 1978b)

A050 CHOP780213 Frequency of the 2nd residue in turn (Chou-Fasman, 1978b)

A052 CHOP780215 Frequency of the 4th residue in turn (Chou-Fasman, 1978b)

A053 CHOP780216 Normalized frequency of the 2nd and 3rd residues in turn (Chou-Fasman, 1978b)

A060 CRAJ730101 Normalized frequency of middle helix (Crawford et al., 1973)

A062 CRAJ730103 Normalized frequency of turn (Crawford et al., 1973)

A074 FASG760103 Optical rotation (Fasman, 1976)

A075 FASG760104 pK-N (Fasman, 1976)

A090 FAUJ880113 pK-a(RCOOH) (Fauchere et al., 1988)

A091 FINA770101 Helix-coil equilibrium constant (Finkelstein-Ptitsyn, 1977)

A093 FINA910102 Helix initiation parameter at posision i,i+1,i+2 (Finkelstein et al., 1991)

A097 GEIM800101 a-helix indices (Geisow-Roberts, 1980)

A098 GEIM800102 a-helix indices for alpha-proteins (Geisow-Roberts, 1980)

A099 GEIM800103 a-helix indices for beta-proteins (Geisow-Roberts, 1980)

A100 GEIM800104 a-helix indices for alpha/beta-proteins (Geisow-Roberts, 1980)

A104 GEIM800108 Aperiodic indices (Geisow-Roberts, 1980)

A105 GEIM800109 Aperiodic indices for alpha-proteins (Geisow-Roberts, 1980)

A107 GEIM800111 Aperiodic indices for alpha/beta-proteins (Geisow-Roberts, 1980)

A119 ISOY800101 Normalized relative frequency of alpha-helix (Isogai et al., 1980)

A121 ISOY800103 Normalized relative frequency of bend (Isogai et al., 1980)

A122 ISOY800104 Normalized relative frequency of bend R (Isogai et al., 1980)

A124 ISOY800106 Normalized relative frequency of helix end (Isogai et al., 1980)

A138 KANM800101 Average relative probability of helix (Kanehisa-Tsong, 1980)

A140 KANM800103 Average relative probability of inner helix (Kanehisa-Tsong, 1980)

A160 LEVM780101 Normalized frequency of alpha-helix, with weights (Levitt, 1978)

A162 LEVM780103 Normalized frequency of reverse turn, with weights (Levitt, 1978)

A163 LEVM780104 Normalized frequency of alpha-helix, unweighted (Levitt, 1978)

A165 LEVM780106 Normalized frequency of reverse turn, unweighted (Levitt, 1978)

A166 LEWP710101 Frequency of occurrence in beta-bends (Lewis et al., 1971)

A171 MAXF760101 Normalized frequency of alpha-helix (Maxfield-Scheraga, 1976)

A176 MAXF760106 Normalized frequency of a region (Maxfield-Scheraga, 1976)

A186 NAGK730101 Normalized frequency of alpha-helix (Nagano, 1973)

A188 NAGK730103 Normalized frequency of coil (Nagano, 1973)

A223 PALJ810101 Normalized frequency of alpha-helix from LG (Palau et al., 1981)

A224 PALJ810102 Normalized frequency of alpha-helix from CF (Palau et al., 1981)

A227 PALJ810105 Normalized frequency of turn from LG (Palau et al., 1981)

A228 PALJ810106 Normalized frequency of turn from CF (Palau et al., 1981)

A229 PALJ810107 Normalized frequency of alpha-helix in all-alphaclass (Palau et al., 1981)

A230 PALJ810108 Normalized frequency of alpha-helix in alpha+beta class (Palau et al., 1981)

A231 PALJ810109 Normalized frequency of alpha-helix in alpha/beta class (Palau et al., 1981)

A235 PALJ810113 Normalized frequency of turn in all-alpha class (Palau et al., 1981)

A236 PALJ810114 Normalized frequency of turn in all-beta class (Palau et al., 1981)

A237 PALJ810115 Normalized frequency of turn in alpha+beta class (Palau et al., 1981)

A238 PALJ810116 Normalized frequency of turn in alpha/beta class (Palau et al., 1981)

A250 PRAM900102 Relative frequency in alpha-helix (Prabhakaran, 1990)

A252 PRAM900104 Relative frequency in reverse-turn (Prabhakaran, 1990)

A256 PTIO830101 Helix-coil equilibrium constant (Ptitsyn-Finkelstein, 1983)

A258 QIAN880101 Weights for alpha-helix at the window position of -6 (Qian-Sejnowski, 1988)

A259 QIAN880102 Weights for alpha-helix at the window position of -5 (Qian-Sejnowski, 1988)

A260 QIAN880103 Weights for alpha-helix at the window position of -4 (Qian-Sejnowski, 1988)

A261 QIAN880104 Weights for alpha-helix at the window position of -3 (Qian-Sejnowski, 1988)

A262 QIAN880105 Weights for alpha-helix at the window position of -2 (Qian-Sejnowski, 1988)

A263 QIAN880106 Weights for alpha-helix at the window position of -1 (Qian-Sejnowski, 1988)

A264 QIAN880107 Weights for alpha-helix at the window position of 0 (Qian-Sejnowski, 1988)

A265 QIAN880108 Weights for alpha-helix at the window position of 1 (Qian-Sejnowski, 1988)

A266 QIAN880109 Weights for alpha-helix at the window position of 2 (Qian-Sejnowski, 1988)

A267 QIAN880110 Weights for alpha-helix at the window position of 3 (Qian-Sejnowski, 1988)

A268 QIAN880111 Weights for alpha-helix at the window position of 4 (Qian-Sejnowski, 1988)

A269 QIAN880112 Weights for alpha-helix at the window position of 5 (Qian-Sejnowski, 1988)

A270 QIAN880113 Weights for alpha-helix at the window position of 6 (Qian-Sejnowski, 1988)

A274 QIAN880117 Weights for beta-sheet at the window position of -3 (Qian-Sejnowski, 1988)

A286 QIAN880129 Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)

A287 QIAN880130 Weights for coil at the window position of -3 (Qian-Sejnowski, 1988)

A288 QIAN880131 Weights for coil at the window position of -2 (Qian-Sejnowski, 1988)

A289 QIAN880132 Weights for coil at the window position of -1 (Qian-Sejnowski, 1988)
A290 QIAN880133 Weights for coil at the window position of 0 (Qian-Sejnowski, 1988)

A291 QIAN880134 Weights for coil at the window position of 1 (Qian-Sejnowski, 1988)

A292 QIAN880135 Weights for coil at the window position of 2 (Qian-Sejnowski, 1988)

A293 QIAN880136 Weights for coil at the window position of 3 (Qian-Sejnowski, 1988)

A294 QIAN880137 Weights for coil at the window position of 4 (Qian-Sejnowski, 1988)

A295 QIAN880138 Weights for coil at the window position of 5 (Qian-Sejnowski, 1988)

A296 QIAN880139 Weights for coil at the window position of 6 (Qian-Sejnowski, 1988)

A300 RACS820104 Average relative fractional occurrence in EL(i) (Rackovsky-Scheraga, 1982)

A304 RACS820108 Average relative fractional occurrence in AR(i-1) (Rackovsky-Scheraga, 1982)

A306 RACS820110 Average relative fractional occurrence in EL(i-1) (Rackovsky-Scheraga, 1982)

A308 RACS820112 Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)

A310 RACS820114 Value of theta(i-1) (Rackovsky-Scheraga, 1982)

A328 RICJ880107 Relative preference value at N4 (Richardson-Richardson, 1988)

A330 RICJ880109 Relative preference value at Mid (Richardson-Richardson, 1988)

A331 RICJ880110 Relative preference value at C5 (Richardson-Richardson, 1988)

A333 RICJ880112 Relative preference value at C3 (Richardson-Richardson, 1988)

A334 RICJ880113 Relative preference value at C2 (Richardson-Richardson, 1988)

A335 RICJ880114 Relative preference value at C1 (Richardson-Richardson, 1988)

A337 RICJ880116 Relative preference value at C' (Richardson-Richardson, 1988)

A338 RICJ880117 Relative preference value at C" (Richardson-Richardson, 1988)

A340 ROBB760101 Information measure for alpha-helix (Robson-Suzuki, 1976)

A342 ROBB760103 Information measure for middle helix (Robson-Suzuki, 1976)

A343 ROBB760104 Information measure for C-terminal helix (Robson-Suzuki, 1976)

A346 ROBB760107 Information measure for extended without H-bond (Robson-Suzuki, 1976)

A347 ROBB760108 Information measure for turn (Robson-Suzuki, 1976)

A348 ROBB760109 Information measure for N-terminal turn (Robson-Suzuki, 1976)

A349 ROBB760110 Information measure for middle turn (Robson-Suzuki, 1976)

A350 ROBB760111 Information measure for C-terminal turn (Robson-Suzuki, 1976)

A351 ROBB760112 Information measure for coil (Robson-Suzuki, 1976)

A352 ROBB760113 Information measure for loop (Robson-Suzuki, 1976)

A359 SNEP660101 Principal component I (Sneath, 1966)

A362 SNEP660104 Principal component IV (Sneath, 1966)

A363 SUEM840101 Zimm-Bragg parameter s at 20° C (Sueki et al., 1984)

A366 TANS770101 Normalized frequency of alpha-helix (Tanaka-Scheraga, 1977)

A367 TANS770102 Normalized frequency of isolated helix (Tanaka-Scheraga, 1977)

A369 TANS770104 Normalized frequency of chain reversal R (Tanaka-Scheraga, 1977)

A375 TANS770110 Normalized frequency of chain reversal (Tanaka-Scheraga, 1977)

A376 VASM830101 Relative population of conformational state A (Vasquez et al., 1983)

A391 WOLS870103 Principal property value z3 (Wold et al., 1987)

	Parameter										
Amino acid	**A005**	**A007**	**A016**	**A018**	**A019**	**A022**	**A023**	**A037**	**A038**	**A040**	**A042**
Ala	1.00	0.37	8.249	6.50	0.486	0.71	-0.118	0.66	1.42	0.74	1.20
Arg	0.52	0.84	8.274	6.90	0.262	1.06	0.124	0.95	0.98	1.01	1.25
Asn	0.35	0.97	8.747	7.50	0.193	1.37	0.289	1.56	0.67	1.46	0.59
Asp	0.44	0.97	8.410	7.00	0.288	1.21	0.048	1.46	1.01	1.52	0.61
Cys	0.06	0.84	8.312	7.70	0.200	1.19	0.083	1.19	0.70	0.96	1.11
Gln	0.44	0.64	8.411	6.00	0.418	0.87	-0.105	0.98	1.11	0.96	1.22
Glu	0.73	0.53	8.368	7.00	0.538	0.84	-0.245	0.74	1.51	0.95	1.24
Gly	0.35	0.97	8.391	5.60	0.120	1.52	0.104	1.56	0.57	1.56	0.42
His	0.60	0.75	8.415	8.00	0.400	1.07	0.138	0.95	1.00	0.95	1.77
Ile	0.73	0.37	8.195	7.00	0.370	0.66	0.230	0.47	1.08	0.47	0.98
Leu	1.00	0.53	8.423	6.50	0.420	0.69	-0.052	0.59	1.21	0.50	1.13
Lys	0.60	0.75	8.408	6.50	0.402	0.99	0.032	1.01	1.16	1.19	1.83
Met	1.00	0.64	8.418	0.00	0.417	0.59	-0.258	0.60	1.45	0.60	1.57
Phe	0.60	0.53	8.228	9.40	0.318	0.71	0.015	0.60	1.13	0.66	1.10
Pro	0.06	0.97	0.000	0.00	0.208	1.61	0.000	1.52	0.57	1.56	0.00
Ser	0.35	0.84	8.380	6.50	0.200	1.34	0.225	1.43	0.77	1.43	0.96
Thr	0.44	0.75	8.236	6.90	0.272	1.08	0.166	0.96	0.83	0.98	0.75
Trp	0.73	0.97	8.094	0.00	0.462	0.76	0.158	0.96	1.08	0.60	0.40
Tyr	0.44	0.84	8.183	6.80	0.161	1.07	0.094	1.14	0.69	1.14	0.73
Val	0.82	0.37	8.436	7.00	0.379	0.63	0.513	0.50	1.06	0.59	1.25
Correlat. coeffic. R with EIIP	**-0.2310**	**0.2819**	**0.1499**	**0.0296**	**-0.1714**	**0.0345**	**-0.1640**	**0.1399**	**-0.0067**	**0.1128**	**0.0380**

	Parameter										
Amino acid	**A044**	**A047**	**A048**	**A049**	**A050**	**A052**	**A053**	**A060**	**A062**	**A074**	**A075**
Ala	0.52	0.67	0.74	0.060	0.076	0.058	0.64	1.33	0.60	1.80	9.69
Arg	1.24	0.89	1.05	0.070	0.106	0.085	1.05	0.79	0.79	12.5	8.99
Asn	1.64	1.86	1.13	0.161	0.083	0.091	1.56	0.72	1.42	-5.60	8.80
Asp	1.06	1.39	1.32	0.147	0.110	0.081	1.61	0.97	1.24	5.05	9.60
Cys	0.94	1.34	0.53	0.149	0.053	0.128	0.92	0.93	1.29	-16.5	8.35
Gln	0.70	1.09	0.77	0.074	0.098	0.098	0.84	1.42	0.92	6.30	9.13
Glu	0.59	0.92	0.85	0.056	0.060	0.064	0.80	1.66	0.64	12.0	9.67
Gly	1.64	1.46	1.68	0.102	0.085	0.152	1.63	0.58	1.38	0.00	9.78
His	1.86	0.78	0.96	0.140	0.047	0.054	0.77	1.49	0.95	-38.5	9.17
Ile	0.87	0.59	0.53	0.043	0.034	0.056	0.29	0.99	0.67	12.4	9.68
Leu	0.84	0.46	0.59	0.061	0.025	0.070	0.36	1.29	0.70	-11.0	9.60
Lys	1.49	1.09	0.82	0.055	0.115	0.095	1.13	1.03	1.10	14.6	9.18
Met	0.52	0.52	0.85	0.068	0.082	0.055	0.51	1.40	0.67	-10.0	9.21
Phe	1.04	0.30	0.44	0.059	0.041	0.065	0.62	1.15	1.05	-34.5	9.18
Pro	1.58	1.58	1.69	0.102	0.301	0.068	2.04	0.49	1.47	-86.2	10.64
Ser	0.93	1.41	1.49	0.120	0.139	0.106	1.52	0.83	1.26	-7.50	9.21
Thr	0.86	1.09	1.16	0.086	0.108	0.079	0.98	0.94	1.05	-28.0	9.10
Trp	0.16	0.48	1.59	0.077	0.013	0.167	0.48	1.33	1.23	-33.7	9.44
Tyr	0.96	1.23	1.01	0.082	0.065	0.125	1.08	0.49	1.35	-10.0	9.11
Val	0.32	0.42	0.59	0.062	0.048	0.053	0.43	0.96	0.48	5.63	9.62
Correlat. coeffic. R with EIIP	**-0.1578**	**0.0404**	**0.0637**	**0.1973**	**0.1089**	**0.1294**	**0.1253**	**-0.0012**	**0.1887**	**-0.0192**	**-0.4190**

	Parameter										
Amino acid	**A090**	**A091**	**A093**	**A097**	**A098**	**A099**	**A100**	**A104**	**A105**	**A107**	**A119**
Ala	4.76	1.08	1.00	1.29	1.13	1.55	1.19	0.91	0.80	0.93	1.53
Arg	4.30	1.05	0.70	1.00	1.09	0.20	1.00	1.00	0.96	1.01	1.17
Asn	3.64	0.85	1.00	0.81	1.06	1.20	0.94	1.64	1.10	1.36	0.60
Asp	5.69	0.85	1.70	1.10	0.94	1.55	1.07	1.40	1.60	1.22	1.00
Cys	3.67	0.95	1.00	0.79	1.32	1.44	0.95	0.93	0.00	0.92	0.89
Gln	4.54	0.95	1.00	1.07	0.93	1.13	1.32	0.94	1.60	0.83	1.27
Glu	5.48	1.15	1.70	1.49	1.20	1.67	1.64	0.97	0.40	1.05	1.63
Gly	3.77	0.55	1.30	0.63	0.83	0.59	0.60	1.51	2.00	1.45	0.44
His	2.84	1.00	1.00	1.33	1.09	1.21	1.03	0.90	0.96	0.96	1.03
Ile	4.81	1.05	1.00	1.05	1.05	1.27	1.12	0.65	0.85	0.58	1.07
Leu	4.79	1.25	1.00	1.31	1.13	1.25	1.18	0.59	0.80	0.59	1.32
Lys	4.27	1.15	0.70	1.33	1.08	1.20	1.27	0.82	0.94	0.91	1.26
Met	4.25	1.15	1.00	1.54	1.23	1.37	1.49	0.58	0.39	0.60	1.66
Phe	4.31	1.10	1.00	1.13	1.01	0.40	1.02	0.72	1.20	0.71	1.22
Pro	0.00	0.71	13.0	0.63	0.82	0.21	0.68	1.66	2.10	1.67	0.25
Ser	3.83	0.75	1.00	0.78	1.01	1.01	0.81	1.23	1.30	1.25	0.65
Thr	3.87	0.75	1.00	0.77	1.17	0.55	0.85	1.04	0.60	1.08	0.86
Trp	4.75	1.10	1.00	1.18	1.32	1.86	1.18	0.67	0.00	0.68	1.05
Tyr	4.30	1.10	1.00	0.71	0.88	1.08	0.77	0.92	1.80	0.98	0.70
Val	4.86	0.95	1.00	0.81	1.13	0.64	0.74	0.60	0.80	0.62	0.93
Correlat. coeffic. R with EIIP	**0.1609**	**-0.060**	**-0.1708**	**-0.0069**	**0.0769**	**-0.0736**	**0.0453**	**-0.0061**	**-0.0457**	**-0.0175**	**0.1256**

	Parameter										
Amino acid	**A121**	**A122**	**A124**	**A138**	**A140**	**A160**	**A162**	**A163**	**A165**	**A166**	**A171**
Ala	0.78	1.09	1.09	1.36	1.45	1.29	0.77	1.32	0.79	0.22	1.43
Arg	1.06	0.97	1.07	1.00	1.15	0.96	0.88	0.98	0.90	0.28	1.18
Asn	1.56	1.14	0.88	0.89	0.64	0.90	1.28	0.95	1.25	0.42	0.64
Asp	1.50	0.77	1.24	1.04	0.91	1.04	1.41	1.03	1.47	0.73	0.92
Cys	0.60	0.50	1.04	0.82	0.70	1.11	0.81	0.92	0.79	0.20	0.94
Gln	0.78	0.83	1.09	1.14	1.14	1.27	0.98	1.10	0.92	0.26	1.22
Glu	0.97	0.93	1.14	1.48	1.29	1.44	0.99	1.44	1.02	0.08	1.67
Gly	1.73	1.25	0.27	0.63	0.53	0.56	1.64	0.61	1.67	0.58	0.46
His	0.83	0.67	1.07	1.11	1.13	1.22	0.68	1.31	0.81	0.14	0.98
Ile	0.40	0.66	0.97	1.08	1.23	0.97	0.51	0.93	0.50	0.22	1.04
Leu	0.57	0.44	1.30	1.21	1.56	1.30	0.58	1.31	0.57	0.19	1.36
Lys	1.01	1.25	1.20	1.22	1.27	1.23	0.96	1.25	0.99	0.27	1.27
Met	0.30	0.45	0.55	1.45	1.83	1.47	0.41	1.39	0.51	0.38	1.53
Phe	0.67	0.50	0.80	1.05	1.20	1.07	0.59	1.02	0.77	0.08	1.19
Pro	1.55	2.96	1.78	0.52	0.21	0.52	1.91	0.58	1.78	0.46	0.49
Ser	1.19	1.21	1.20	0.74	0.48	0.82	1.32	0.76	1.30	0.55	0.70
Thr	1.09	1.33	0.99	0.81	0.77	0.82	1.04	0.79	0.97	0.49	0.78
Trp	0.74	0.62	1.03	0.97	1.17	0.99	0.76	0.97	0.79	0.43	1.01
Tyr	1.14	0.94	0.69	0.79	0.74	0.72	1.05	0.73	0.93	0.46	0.69
Val	0.44	0.56	0.77	0.94	1.10	0.91	0.47	0.93	0.46	0.08	0.98
Correlat.											

coeffic. R with *EIIP*	**0.0064**	**-0.1564**	**0.0318**	**-0.0339**	**-0.0145**	**0.0691**	**0.0123**	**-0.0702**	**0.0520**	**0.3536**	**0.0437**

	Parameter										
Amino acid	**A176**	**A186**	**A188**	**A223**	**A224**	**A227**	**A228**	**A229**	**A230**	**A231**	**A235**
Ala	1.00	1.29	0.72	1.30	1.32	0.84	0.65	1.08	1.34	1.15	0.69
Arg	1.18	0.83	1.33	0.93	1.04	0.91	0.93	0.93	0.91	1.06	0.00
Asn	0.87	0.77	1.38	0.90	0.74	1.48	1.45	1.05	0.83	0.87	1.52
Asp	1.39	1.00	1.04	1.02	0.97	1.28	1.47	0.86	1.06	1.00	2.42
Cys	1.09	0.94	1.01	0.92	0.70	0.69	1.43	1.22	1.27	1.03	0.00
Gln	1.13	1.10	0.81	1.04	1.25	1.00	0.94	0.95	1.13	1.43	1.44
Glu	1.04	1.54	0.75	1.43	1.48	0.78	0.75	1.09	1.69	1.37	0.63
Gly	0.46	0.72	1.35	0.63	0.59	1.76	1.53	0.85	0.47	0.64	2.64
His	0.71	1.29	0.76	1.33	1.06	0.53	0.96	1.02	1.11	0.95	0.22
Ile	0.68	0.94	0.80	0.87	1.01	0.55	0.57	0.98	0.84	0.99	0.43
Leu	1.01	1.23	0.63	1.30	1.22	0.49	0.56	1.04	1.39	1.22	0.00
Lys	1.05	1.23	0.84	1.23	1.13	0.95	0.95	1.01	1.08	1.20	1.18
Met	0.36	1.23	0.62	1.32	1.47	0.52	0.71	1.11	0.90	1.45	0.88
Phe	0.65	1.23	0.58	1.09	1.10	0.88	0.72	0.96	1.02	0.92	2.20
Pro	1.95	0.70	1.43	0.63	0.57	1.47	1.51	0.91	0.48	0.72	1.34
Ser	1.56	0.78	1.34	0.78	0.77	1.29	1.46	0.95	1.05	0.84	1.43
Thr	1.23	0.87	1.03	0.80	0.86	1.05	0.96	1.15	0.74	0.97	0.28
Trp	1.10	1.06	0.87	1.03	1.02	0.88	0.90	1.17	0.64	1.11	0.00
Tyr	0.87	0.63	1.35	0.71	0.72	1.28	1.12	0.80	0.73	0.72	1.53

Val	0.58	0.97	0.83	0.95	1.05	0.51	0.55	1.03	1.18	0.82	0.14
Correlat. coeffic. R with EIIP	**0.2688**	**-0.0814**	**0.0317**	**-0.0710**	**0.0151**	**0.0725**	**0.2132**	**-0.0164**	**-0.0532**	**0.1615**	**0.1771**

	Parameter										
Amino acid	**A236**	**A237**	**A238**	**A250**	**A252**	**A256**	**A258**	**A259**	**A260**	**A261**	**A262**
Ala	0.87	0.91	0.92	1.29	0.78	1.10	0.12	0.26	0.64	0.29	0.68
Arg	1.30	0.77	0.90	0.96	0.88	0.95	0.04	-0.14	-0.10	-0.03	-0.22
Asn	1.36	1.32	1.57	0.90	1.28	0.80	-0.10	-0.03	0.09	-0.04	-0.09
Asp	1.24	0.90	1.22	1.04	1.41	0.65	0.01	0.15	0.33	0.11	-0.02
Cys	0.83	0.50	0.62	1.11	0.80	0.95	-0.25	-0.15	0.03	-0.05	-0.15
Gln	1.06	1.06	0.66	1.27	0.97	1.00	-0.03	-0.13	-0.23	0.26	-0.15
Glu	0.91	0.53	0.92	1.44	1.00	1.00	-0.02	0.21	0.51	0.28	0.44
Gly	1.69	1.61	1.61	0.56	1.64	0.60	-0.02	-0.37	-0.09	-0.67	-0.73
His	0.91	1.08	0.39	1.22	0.69	0.85	-0.06	0.10	-0.23	-0.26	-0.14
Ile	0.27	0.36	0.79	0.97	0.51	1.10	-0.07	-0.03	-0.22	0.00	-0.08
Leu	0.67	0.77	0.50	1.30	0.59	1.25	0.05	-0.02	0.41	0.47	0.61
Lys	0.66	1.27	0.86	1.23	0.96	1.00	0.26	0.12	-0.17	-0.19	0.03
Met	0.00	0.76	0.50	1.47	0.39	1.15	0.00	0.00	0.13	0.27	0.39
Phe	0.47	0.37	0.96	1.07	0.58	1.10	0.05	0.12	-0.03	0.24	0.06
Pro	1.54	1.62	1.30	0.52	1.91	0.10	-0.19	-0.08	-0.43	-0.34	-0.76
Ser	1.08	1.34	1.40	0.82	1.33	0.75	-0.19	0.01	-0.10	-0.17	-0.26
Thr	1.12	0.87	1.11	0.82	1.03	0.75	-0.04	-0.34	-0.07	-0.20	-0.10

Trp	1.24	1.10	0.57	0.99	0.75	1.10	-0.06	-0.01	-0.02	0.25	0.20
Tyr	0.54	1.24	1.78	0.72	11.05	1.10	-0.14	-0.29	-0.38	-0.30	-0.04
Val	0.69	0.52	0.50	0.91	0.47	0.95	-0.03	0.02	-0.01	-0.01	0.12
Correlat. coeffic. R with EIIP	**-0.0188**	**-0.1493**	**0.0189**	**0.0067**	**0.0019**	**0.0004**	**-0.0254**	**-0.0638**	**-0.0339**	**0.1408**	**-0.0506**

	Parameter										
Amino acid	**A263**	**A264**	**A265**	**A266**	**A267**	**A268**	**A269**	**A270**	**A274**	**A286**	**A287**
Ala	0.34	0.57	0.33	0.13	0.31	0.21	0.18	-0.08	-0.14	-0.43	-0.19
Arg	0.22	0.23	0.10	0.08	0.18	0.07	0.21	0.05	0.14	0.06	-0.07
Asn	-0.33	-0.36	-0.19	-0.07	-0.10	-0.04	-0.03	-0.08	-0.27	0.00	0.17
Asp	0.06	-0.46	-0.44	-0.71	-0.81	-0.58	-0.32	-0.24	-0.10	-0.31	-0.27
Cys	-0.18	-0.15	-0.03	-0.09	-0.26	-0.12	-0.29	-0.25	-0.64	0.19	0.42
Gln	0.01	0.15	0.19	0.12	0.41	0.13	-0.27	-0.28	-0.11	0.14	-0.29
Glu	0.20	0.26	0.21	0.13	-0.06	-0.23	-0.25	-0.19	-0.39	-0.41	-0.22
Gly	-0.88	-0.71	-0.46	-0.39	-0.42	-0.15	-0.40	-0.10	0.46	-0.21	0.17
His	-0.09	-0.05	0.27	0.32	0.51	0.37	0.28	0.29	-0.04	0.21	0.17
Ile	-0.03	0.00	-0.33	0.00	-0.15	0.31	-0.03	-0.01	0.16	0.29	-0.34
Leu	0.20	0.48	0.57	0.50	0.56	0.70	0.62	0.28	-0.57	-0.10	-0.22
Lys	-0.11	0.16	0.23	0.37	0.47	0.28	0.41	0.45	0.04	0.33	0.00
Met	0.43	0.41	0.79	0.63	0.58	0.61	0.21	0.11	0.24	-0.01	-0.53
Phe	0.15	0.03	0.48	0.15	0.10	-0.06	0.05	0.00	0.08	0.25	-0.31
Pro	-0.81	-1.12	-1.86	-1.4	-1.33	-1.03	-0.84	-0.42	0.02	0.28	0.14

Ser	-0.35	-0.47	-0.23	-0.28	-0.49	-0.28	-0.05	0.07	-0.12	-0.23	0.22
Thr	-0.37	-0.54	-0.33	-0.21	-0.44	-0.25	-0.16	-0.33	0.00	-0.26	0.10
Trp	0.07	-0.10	0.15	0.02	0.14	0.21	0.32	0.36	-0.10	0.15	-0.15
Tyr	-0.31	-0.35	-0.19	-0.10	-0.08	0.16	0.11	0.00	0.18	0.09	-0.02
Val	0.13	0.31	0.24	0.17	-0.01	0.00	0.06	-0.13	0.29	-0.10	-0.33
Correlat. coeffic. R with EIIP	**0.2097**	**-0.0833**	**0.0934**	**-0.0954**	**-0.1093**	**0.1986**	**-0.0816**	**-0.1646**	**-0.0020**	**-0.0466**	**-0.0457**

	Parameter										
Amino acid	**A288**	**A289**	**A290**	**A291**	**A292**	**A293**	**A294**	**A295**	**A296**	**A300**	**A304**
Ala	-0.25	-0.27	-0.42	-0.24	-0.14	0.01	-0.30	-0.23	0.08	0.78	1.48
Arg	0.12	-0.40	-0.23	-0.04	0.21	-0.13	-0.09	-0.20	-0.01	1.75	1.02
Asn	0.61	0.71	0.81	0.45	0.35	-0.11	-0.12	0.06	-0.06	1.32	0.99
Asp	0.60	0.54	0.95	0.65	0.65	0.78	0.44	0.34	0.04	1.25	1.19
Cys	0.18	0.00	-0.18	-0.38	-0.09	-0.31	0.03	0.19	0.37	3.14	0.86
Gln	0.09	-0.08	-0.01	0.01	0.11	-0.13	0.24	0.47	0.48	0.93	1.42
Glu	-0.12	-0.12	-0.09	0.07	0.06	0.09	0.18	0.28	0.36	0.94	1.43
Gly	0.09	1.14	1.24	0.85	0.36	0.14	-0.12	0.14	-0.02	1.13	0.46
His	0.42	0.18	0.05	-0.21	-0.31	-0.56	-0.20	-0.22	-0.45	1.03	1.27
Ile	-0.54	-0.74	-1.17	-0.65	-0.50	-0.09	-0.07	0.42	0.09	1.26	1.12
Leu	-0.55	-0.54	-0.69	-0.80	-0.80	-0.81	-0.18	-0.36	0.24	0.91	1.33
Lys	0.14	0.45	0.09	0.17	-0.14	-0.43	0.06	-0.15	-0.27	0.85	1.36
Met	-0.47	-0.76	-0.86	-0.71	-0.56	-0.49	-0.44	-0.19	0.16	0.41	1.41

Phe	-0.29	-0.47	-0.39	-0.61	-0.25	-0.20	0.11	-0.02	0.34	1.07	1.30
Pro	0.89	1.40	1.77	2.27	1.59	1.14	0.77	0.78	0.16	1.73	0.25
Ser	0.24	0.40	0.63	0.33	0.32	0.13	-0.09	-0.29	-0.35	1.31	0.89
Thr	0.16	-0.10	0.29	0.13	0.21	-0.02	-0.27	-0.30	-0.04	1.57	0.81
Trp	-0.44	-0.46	-0.37	-0.44	-0.17	-0.20	-0.09	-0.18	-0.06	0.98	1.27
Tyr	-0.19	-0.05	-0.41	-0.49	-0.35	0.10	-0.25	0.07	-0.20	1.31	0.91
Val	-0.45	-0.86	-1.32	-0.99	-0.70	-0.11	-0.06	0.29	0.18	1.11	0.93
Correlat. coeffic. R with EIIP	**0.1742**	**-0.0938**	**0.0878**	**-0.0173**	**0.1783**	**0.1419**	**0.0683**	**-0.1877**	**0.0387**	**0.2400**	**0.1209**

	Parameter										
Amino acid	**A306**	**A308**	**A310**	**A328**	**A330**	**A331**	**A333**	**A334**	**A335**	**A337**	**A338**
Ala	1.02	0.99	14.53	1.10	1.80	1.80	0.70	1.40	1.10	1.00	0.70
Arg	1.00	1.19	17.82	1.50	1.30	1.00	0.80	2.10	1.00	1.40	1.10
Asn	1.31	1.15	13.59	0.00	0.90	0.60	0.80	0.90	1.20	0.90	1.50
Asp	1.76	1.18	19.78	0.30	1.00	0.70	0.60	0.70	0.40	1.40	1.40
Cys	1.05	2.32	30.57	1.10	0.70	0.00	0.20	1.20	1.60	0.80	0.40
Gln	1.05	1.52	22.18	1.30	1.30	1.00	1.30	1.60	2.10	1.40	1.10
Glu	0.83	1.36	18.19	0.50	0.80	1.10	1.60	1.70	0.80	0.80	0.70
Gly	2.39	1.40	37.16	0.40	0.50	0.50	0.10	0.20	0.20	1.20	0.60
His	0.40	1.06	22.63	1.50	1.00	2.40	1.10	1.80	3.40	1.20	1.00
Ile	0.83	0.81	20.28	1.10	1.20	1.30	1.40	0.40	0.70	1.10	0.07
Leu	1.06	1.26	14.30	2.60	1.20	1.20	1.90	0.80	0.70	0.90	0.50

Lys	0.94	0.91	14.07	0.80	1.10	1.40	2.20	1.90	2.00	1.20	1.30
Met	1.33	1.00	20.61	1.70	1.50	2.70	1.00	1.30	1.00	0.80	0.00
Phe	0.41	1.25	19.61	1.90	1.30	1.90	1.80	0.30	0.70	0.10	1.20
Pro	2.73	0.00	52.63	0.10	0.30	0.30	0.00	0.20	0.00	1.90	1.50
Ser	1.18	1.50	18.56	0.40	0.60	0.50	0.60	1.60	1.70	0.70	0.90
Thr	0.77	1.18	21.09	0.50	1.00	1.50	0.70	0.90	1.00	0.80	2.10
Trp	1.22	1.33	19.78	3.10	1.50	1.10	0.40	0.40	0.00	0.40	2.70
Tyr	1.09	1.09	26.36	0.60	0.80	1.30	1.10	0.30	1.20	0.90	0.50
Val	0.88	1.01	21.87	1.50	1.20	0.40	1.30	0.70	0.70	0.60	1.00
Correlat. coeffic. R with EIIP	**-0.1138**	**0.3280**	**-0.1105**	**0.0272**	**0.1707**	**0.0211**	**-0.1965**	**0.1887**	**0.0691**	**-0.0946**	**0.1725**

	Parameter										
Amino acid	**A340**	**A342**	**A343**	**A346**	**A347**	**A348**	**A349**	**A350**	**A351**	**A352**	**A359**
Ala	6.5	6.7	2.3	0.0	-5.0	-3.3	-4.7	-3.7	-2.5	-5.1	0.239
Arg	-0.9	0.3	1.4	1.1	2.1	0.0	2.0	1.0	-1.2	2.6	0.211
Asn	-5.1	-6.1	-3.3	-2.0	4.2	5.4	3.9	-0.6	4.6	4.7	0.249
Asp	0.5	-3.1	-4.4	-2.6	3.1	3.9	1.9	-0.6	0.0	3.1	0.171
Cys	-1.3	-4.9	6.1	5.4	4.4	-0.3	6.2	4.0	-4.7	3.8	0.22
Gln	1.0	0.6	2.7	2.4	0.4	-0.4	-2.0	3.4	-0.5	0.2	0.26
Glu	77.8	2.2	2.5	3.1	-4.7	-1.8	-4.2	-4.3	-4.4	-5.2	0.187
Gly	-8.6	-6.8	-8.3	-3.4	5.7	-1.2	5.7	5.9	4.9	5.6	0.16
His	1.2	-1.0	5.9	0.8	-0.3	3.0	-2.6	-0.8	1.6	-0.9	0.205

Ile	0.6	3.2	-0.5	-0.1	-4.6	-0.5	-7.0	-0.5	-3.3	-4.5	0.273
Leu	3.2	5.5	0.1	-3.7	-5.6	-2.3	-6.2	-2.8	-2.0	-5.4	0.281
Lys	2.3	0.5	7.3	-3.1	1.0	-1.2	2.8	1.3	-0.8	1.0	0.228
Met	5.3	7.2	3.5	-2.1	-4.8	-4.3	-4.8	-1.6	-4.1	-5.3	0.253
Phe	1.6	2.8	1.6	0.7	-1.8	0.8	-3.7	1.6	-4.1	-2.4	0.234
Pro	-7.7	-22.8	24.4	7.4	2.6	6.5	3.6	-6.0	5.8	3.5	0.165
Ser	-3.9	-3.0	-1.9	1.3	2.6	1.8	2.1	.5	2.5	3.2	0.236
Thr	-2.6	-4.0	-3.7	0.0	0.3	-0.7	0.6	1.2	1.7	0.0	0.213
Trp	1.2	4.0	-0.9	-3.4	3.4	-0.8	3.3	6.5	1.2	2.9	0.183
Tyr	-4.5	-4.6	-0.6	4.8	2.9	3.1	3.8	1.3	-0.6	3.2	0.193
Val	1.4	2.5	2.3	2.7	-6.0	-3.5	-6.2	-4.6	-3.5	-6.3	0.255
Correlat. coeffic. R with EIIP	**0.0278**	**0.0540**	**0.1422**	**0.0306**	**0.3012**	**0.0757**	**0.2553**	**0.3610**	**-0.1534**	**0.2732**	**-0.1501**

	Parameter							
Amino acid	**A362**	**A363**	**A366**	**A367**	**A369**	**A375**	**A376**	**A391**
Ala	-0.062	1.071	1.42	0.946	1.194	0.842	0.135	0.09
Arg	-0.167	1.033	1.06	1.128	0.795	0.936	0.296	-3.44
Asn	0.166	0.784	0.71	0.432	0.659	1.352	0.196	0.84
Asp	-0.079	0.68	1.01	1.311	1.056	1.366	0.289	2.36
Cys	0.38	0.922	0.73	0.481	0.678	1.032	0.159	4.13
Gln	-0.025	0.977	1.02	1.615	1.29	0.998	0.236	-1.14
Glu	-0.184	0.97	1.63	0.698	0.928	0.758	0.184	-0.07
Gly	-0.017	0.591	0.5	0.36	1.015	1.349	0.051	0.3

His	0.056	0.85	1.2	2.168	0.611	1.079	0.223	1.11
Ile	-0.309	1.14	1.12	1.283	0.603	0.459	0.173	-1.03
Leu	-0.264	1.14	1.29	1.192	0.595	0.665	0.215	-0.98
Lys	-0.371	0.939	1.24	1.203	1.06	1.045	0.17	-3.14
Met	0.077	1.2	1.21	0.0	0.831	0.668	0.239	-0.41
Phe	0.074	1.086	1.16	0.963	0.377	0.881	0.087	0.45
Pro	-0.036	0.659	0.66	2.093	3.159	1.385	0.151	2.23
Ser	0.47	0.76	0.71	0.523	1.444	1.257	0.01	0.57
Thr	0.348	0.817	0.78	1.961	1.172	1.055	0.10	-1.4
Trp	0.05	1.107	1.05	1.925	0.452	0.881	0.166	0.85
Tyr	0.22	1.02	0.67	0.802	0.816	1.101	0.066	0.01
Val	-0.212	0.95	0.99	0.409	0.64	0.643	0.285	-1.29
Correlat. coeffic. R with EIIP	**0.4555**	**-0.0036**	**-0.1134**	**0.0679**	**-0.0150**	**0.2250**	**0.0315**	**0.1092**

B. beta propensity

B020 BURA740102 Normalized frequency of extended structure (Burgess et al., 1974)

B028 CHAM830107 A parameter of charge transfer capability (Charton-Charton, 1983)

B039 CHOP780202 Normalized frequency of beta-sheet (Chou-Fasman, 1978b)

B045 CHOP780208 Normalized frequency of N-terminal beta-sheet (Chou-Fasman, 1978b)

B046 CHOP780209 Normalized frequency of C-terminal beta-sheet (Chou-Fasman, 1978b)

B061 CRAJ730102 Normalized frequency of beta-sheet (Crawford et al., 1973)

B101 GEIM800105 beta-strand indices (Geisow-Roberts, 1980)

B102 GEIM800106 beta-strand indices for beta-proteins (Geisow-Roberts, 1980)

B103 GEIM800107 beta-strand indices for alpha/beta-proteins (Geisow-Roberts, 1980)

B106 GEIM800110 Aperiodic indices for beta-proteins (Geisow-Roberts, 1980)

B120 ISOY800102 Normalized relative frequency of extended structure (Isogai et al., 1980)

B139 KANM800102 Average relative probability of beta-sheet (Kanehisa-Tsong, 1980)

B141 KANM800104 Average relative probability of inner beta-sheet (Kanehisa-Tsong, 1980)

B161 LEVM780102 Normalized frequency of beta-sheet, with weights (Levitt, 1978)

B164 LEVM780105 Normalized frequency of beta-sheet, unweighted (Levitt, 1978)

B167 LIFS790101 Conformational preference for all beta-strands (Lifson-Sander, 1979)

B168 LIFS790102 Conformational preference for parallel beta-strands (Lifson-Sander, 1979)

B169 LIFS790103 Conformational preference for antiparallel beta-strands (Lifson-Sander, 1979)

B172 MAXF760102 Normalized frequency of extended structure (Maxfield-Scheraga, 1976)

B187 NAGK730102 Normalized frequency of beta-structure (Nagano, 1973)

B218 OOBM850101 Optimized beta-structure-coil equilibrium constant (Oobatake et al., 1985)

B221 OOBM850104 Optimized average non-bonded energy per atom (Oobatake et al., 1985)

B225 PALJ810103 Normalized frequency of beta-sheet from LG (Palau et al., 1981)

B226 PALJ810104 Normalized frequency of beta-sheet from CF (Palau et al., 1981)

B232 PALJ810110 Normalized frequency of beta-sheet in all-beta class (Palau et al., 1981)

B234 PALJ810112 Normalized frequency of beta-sheet in alpha/beta class (Palau et al., 1981)

B251 PRAM900103 Relative frequency in beta-sheet (Prabhakaran, 1990)

B257 PTIO830102 beta-coil equilibrium constant (Ptitsyn-Finkelstein, 1983)

B275 QIAN880118 Weights for beta-sheet at the window position of -2 (Qian-Sejnowski, 1988)

B276 QIAN880119 Weights for beta-sheet at the window position of -1 (Qian-Sejnowski, 1988)

B277 QIAN880120 Weights for beta-sheet at the window position of 0 (Qian-Sejnowski, 1988)

B278 QIAN880121 Weights for beta-sheet at the window position of 1 (Qian-Sejnowski, 1988)

B279 QIAN880122 Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)

B307 RACS820111 Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga, 1982)

B344 ROBB760105 Information measure for extended (Robson-Suzuki, 1976)

B345 ROBB760106 Information measure for pleated-sheet (Robson-Suzuki, 1976)

B368 TANS770103 Normalized frequency of extended structure (Tanaka-Scheraga, 1977)

	Parameter										
Amino acid	**B020**	**B028**	**B039**	**B045**	**B046**	**B061**	**B101**	**B102**	**B103**	**B106**	**B120**
Ala	0.288	0	0.83	[illegible]6	0.75	1.00	0.84	0.86	0.91	1.10	0.86
Arg	0.362	0	0.93	[illegible]	0.90	0.74	1.04	1.15	0.99	0.93	0.98
Asn	0.229	1	0.89	[illegible]	1.21	0.75	0.66	0.60	0.72	1.57	0.74
Asp	0.271	1	0.54	0.38	0.85	0.89	0.59	0.66	0.74	1.41	0.69
Cys	0.533	0	1.19	0.87	1.11	0.99	1.27	0.91	1.12	1.05	1.39
Gln	0.327	0	1.10	1.65	0.65	0.87	1.02	1.11	0.90	0.81	0.89
Glu	0.262	1	[illegible]	0.35	0.55	0.37	0.57	0.37	0.41	1.40	0.66
Gly	0.312	1	[illegible]	0.63	0.74	0.56	0.94	0.86	0.91	1.30	0.70
His	0.200	0	[illegible]	[illegible]	0.90	0.36	0.81	1.07	1.01	0.85	1.06
Ile	0.411	0	[illegible]	1.94	1.35	1.75	1.29	1.17	1.29	0.67	1.31
Leu	0.400	0	[illegible]	1.30	1.27	1.53	1.10	1.28	1.23	0.52	1.01
Lys	0.265	0	0.74	1.00	0.74	1.18	0.86	1.01	0.86	0.94	0.77
Met	0.375	0	1.05	1.43	0.95	1.40	0.88	1.15	0.96	0.69	1.06
Phe	0.318	0	1.38	1.50	1.50	1.26	1.15	1.34	1.26	0.60	1.16
Pro	0.340	0	0.55	0.66	0.40	0.36	0.80	0.61	0.65	1.77	1.16
Ser	0.354	0	0.75	0.63	0.79	0.65	1.05	0.91	0.93	1.13	1.09
Thr	0.388	0	1.19	1.17	0.75	1.15	1.20	0.14	1.05	0.88	1.24
Trp	0.231	0	1.37	1.49	1.19	0.84	1.15	1.13	1.15	0.62	1.17
Tyr	0.429	0	1.47	1.07	1.96	1.41	1.39	1.37	1.21	0.41	1.28
Val	0.495	0	1.70	1.69	1.79	1.61	1.56	1.31	1.58	0.58	1.40
Correlat. coeffic. R with EIIP	**0.0860**	**-0.1763**	**-0.0600**	**-0.0337**	**-0.1064**	**-0.0097**	**0.0134**	**0.1590**	**-0.0157**	**-0.0907**	**0.0832**

	Parameter										
Amino acid	**B139**	**B141**	**B161**	**B164**	**B167**	**B168**	**B169**	**B172**	**B187**	**B218**	**B221**
Ala	0.81	0.75	0.90	0.86	0.92	1.00	0.90	0.86	0.96	2.01	-2.49
Arg	0.85	0.79	0.99	0.97	0.93	0.68	1.02	0.94	0.67	0.84	2.55
Asn	0.62	0.33	0.76	0.73	0.60	0.54	0.62	0.74	0.72	0.03	2.27
Asp	0.71	0.31	0.72	0.69	0.48	0.50	0.47	0.72	0.90	-2.05	8.86
Cys	1.17	1.46	0.74	1.04	1.16	0.91	1.24	1.17	1.13	1.98	-3.13
Gln	0.98	0.75	0.80	1.00	0.95	0.28	1.18	0.89	1.18	1.02	1.79
Glu	0.53	0.46	0.75	0.66	0.61	0.59	0.62	0.62	0.33	0.93	4.04
Gly	0.88	0.83	0.92	0.89	0.61	0.79	0.56	0.97	0.90	0.12	-0.56
His	0.92	0.83	1.08	0.85	0.93	0.38	1.12	1.06	0.87	-0.14	4.22
Ile	1.48	1.87	1.45	1.47	1.81	2.60	1.54	1.24	1.54	3.70	-10.87
Leu	1.24	1.56	1.02	1.04	1.30	1.42	1.26	0.98	1.26	2.73	-7.16
Lys	0.77	0.66	0.77	0.77	0.70	0.59	0.74	0.79	0.81	2.55	-9.97
Met	1.05	0.86	0.97	0.93	1.19	1.49	1.09	1.08	1.29	1.75	-4.96
Phe	1.20	1.37	1.32	1.21	1.25	1.30	1.23	1.16	1.37	2.68	-6.64
Pro	0.61	0.52	0.64	0.68	0.40	0.35	0.42	1.22	0.75	0.41	5.19
Ser	0.92	0.82	0.95	1.02	0.82	0.70	0.87	1.04	0.77	1.47	-1.60
Thr	1.18	1.36	1.21	1.27	1.12	0.59	1.30	1.18	1.23	2.39	-4.75
Trp	1.18	0.79	1.14	1.26	1.54	0.89	1.75	1.07	1.13	2.49	17.84
Tyr	1.23	1.08	1.25	1.31	1.53	1.08	1.68	1.25	1.07	2.23	9.25
Val	1.66	2.0	1.49	1.43	1.81	2.63	1.53	1.33	1.41	3.50	-3.97
Correlat. coeffic. R	**-0.0403**	**-0.1626**	**-0.1053**	**0.0210**	**-0.0981**	**-0.3132**	**0.0353**	**-0.0032**	**0.0809**	**-0.2260**	**0.1537**

with EIIP											

	Parameter										
Amino acid	**B225**	**B226**	**B232**	**B234**	**B251**	**B257**	**B275**	**B276**	**B277**	**B278**	**B279**
Ala	0.81	0.90	0.89	0.98	0.90	1.00	-0.31	-0.10	-0.25	-0.26	0.05
Arg	1.03	0.75	1.06	1.03	0.99	0.70	0.25	0.19	-0.02	-0.09	-0.11
Asn	0.81	0.82	0.67	0.66	0.76	0.60	-0.53	-0.89	-0.77	-0.34	-0.40
Asp	0.71	0.75	0.71	0.74	0.72	0.50	-0.54	-0.89	-1.01	-0.55	-0.11
Cys	1.12	1.12	1.04	1.01	0.74	1.90	-0.06	0.13	0.13	0.47	0.36
Gln	1.03	0.95	1.06	0.63	0.80	1.00	0.07	-0.04	-0.12	-0.33	-0.67
Glu	0.59	0.44	0.72	0.59	0.75	0.70	-0.52	-0.34	-0.62	-0.75	-0.35
Gly	0.94	0.83	0.87	0.90	0.92	0.30	0.37	-0.45	-0.72	-0.56	0.14
His	0.85	0.86	1.04	1.17	1.08	0.80	-0.32	-0.34	-0.16	-0.04	0.02
Ile	1.47	1.59	1.14	1.38	1.45	4.00	0.57	0.95	1.10	0.94	0.47
Leu	1.03	1.24	1.02	1.05	1.02	2.00	0.09	0.32	0.23	0.25	0.32
Lys	0.77	0.75	1.00	0.83	0.77	0.70	-0.29	-0.46	-0.59	-0.55	-0.51
Met	0.96	0.94	1.41	0.82	0.97	1.90	0.29	0.43	0.32	-0.05	-0.10
Phe	1.13	1.41	1.32	1.23	1.32	3.10	0.24	0.36	0.48	0.20	0.20
Pro	0.75	0.46	0.69	0.73	0.64	0.20	-0.31	-0.91	-1.24	-1.28	-0.79
Ser	1.02	0.70	0.86	0.98	0.95	0.90	0.11	-0.12	-0.31	-0.28	0.03
Thr	1.19	1.20	1.15	1.20	1.21	1.70	0.03	0.49	0.17	0.08	-0.15
Trp	1.24	1.28	1.06	1.26	1.14	2.20	0.15	0.34	0.45	0.22	0.09
Tyr	1.35	1.45	1.35	1.23	1.25	2.80	0.29	0.42	0.77	0.53	0.34
Val	1.44	1.73	1.66	1.62	1.49	4.00	0.48	0.76	0.69	0.67	0.58

Correlat. coeffic. R with EIIP	**0.0076**	**-0.0698**	**0.1042**	**-0.0859**	**-0.1053**	**-0.0992**	**-0.0063**	**0.0403**	**0.0104**	**-0.0060**	**-0.0943**

	Parameter			
Amino acid	**B307**	**B344**	**B345**	**B368**
Ala	0.93	-2.30	-2.70	0.790
Arg	1.52	0.40	0.40	1.087
Asn	0.92	-4.10	-4.20	0.832
Asp	0.60	-4.40	-4.40	0.530
Cys	1.08	4.40	3.70	1.268
Gln	0.94	1.20	0.80	1.038
Glu	0.73	-5.00	-8.10	0.643
Gly	0.78	-4.20	-3.90	0.725
His	1.08	-2.50	-3.00	0.864
Ile	1.74	6.70	7.70	1.361
Leu	1.03	2.30	3.70	1.111
Lys	1.00	-3.30	-2.90	0.735
Met	1.31	2.30	3.70	1.092
Phe	1.51	2.60	3.00	1.052
Pro	1.37	-1.80	-6.60	1.249
Ser	0.97	-1.70	-2.40	1.093
Thr	1.38	1.30	1.70	1.214
Trp	1.12	-1.00	0.30	1.1144

Tyr	1.65	4.00	3.30	1.340
Val	1.70	6.80	7.10	1.428
Correlat. coeffic. R with EIIP	**0.1264**	**0.0333**	**0.0807**	**-0.0367**

C. Composition

C029 CHAM830108 A parameter of charge transfer donor capability (Charton-Charton, 1983)

C064 DAYM780101 Amino acid composition (Dayhoff et al., 1978b)

C116 HUTJ700101 Heat capacity (Hutchens, 1970)

C134 JOND920101 Relative frequency of occurrence (Jones et al., 1992)

C136 JUKT750101 Amino acid distribution (Jukes et al., 1975)

C137 JUNJ780101 Sequence frequency (Jungck, 1978)

C144 KARP850103 Flexibility parameter for two rigid neighbors (Karplus-Schulz, 1985)

C189 NAKH920101 AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa, 1992)

C190 NAKH920102 AA composition of CYT2 of single-spanning proteins (Nakashima-Nishikawa, 1992)

C191 NAKH920103 AA composition of EXT of single-spanning proteins (Nakashima-Nishikawa, 1992)

C192 NAKH920104 AA composition of EXT2 of single-spanning proteins (Nakashima-Nishikawa, 1992)

C193 NAKH920105 AA composition of MEM of single-spanning proteins (Nakashima-Nishikawa, 1992)

C194 NAKH920106 AA composition of CYT of multi-spanning proteins (Nakashima-Nishikawa, 1992)

C195 NAKH920107 AA composition of EXT of multi-spanning proteins (Nakashima-Nishikawa, 1992)

C196 NAKH920108 AA composition of MEM of multi-spanning proteins (Nakashima-Nishikawa, 1992)

C197 NAKH900101 AA composition of total proteins (Nakashima et al., 1990)

C198 NAKH900102 SD of AA composition of total proteins (Nakashima et al., 1990)

C199 NAKH900103 AA composition of mt-proteins (Nakashima et al., 1990)

C201 NAKH900105 AA composition of mt-proteins from animal (Nakashima et al., 1990)

C203 NAKH900107 AA composition of mt-proteins from fungi and plant (Nakashima et al., 1990)

C205 NAKH900109 AA composition of membrane proteins (Nakashima et al., 1990)

C207 NAKH900111 Transmembrane regions of non-mt-proteins (Nakashima et al., 1990)

C208 NAKH900112 Transmembrane regions of mt-proteins (Nakashima et al., 1990)

C301 RACS820105 Average relative fractional occurrence in E0(i) (Rackovsky-Scheraga, 1982)

	Parameter										
Amino acid	**C029**	**C064**	**C116**	**C134**	**C136**	**C137**	**C144**	**C189**	**C190**	**C191**	**C192**
Ala	0	8.6	29.22	0.077	5.300	685.0	0.892	[illegible]	0.88	5.15	5.04
Arg	1	4.9	26.37	0.051	2.60	382.0	0.901	[illegible]	[illegible]1	4.38	3.73
Asn	1	4.3	38.30	0.043	3.00	397.0	0.930	4.18	[illegible]5	4.81	5.94
Asp	0	5.5	37.09	0.052	3.60	400.0	0.932	[illegible]	[illegible]13	5.75	5.26
Cys	[illegible]	2.9	50.70	0.020	1.30	241.0	0.925	[illegible]	[illegible]69	3.24	2.20
Gln	1	3.9	44.02	0.041	2.40	313.0	0.885	4.76	4.68	4.45	4.50
Glu	0	6.0	41.84	0.062	3.30	427.0	0.933	7.82	9.34	7.05	6.07
Gly	0	8.4	23.71	0.074	4.80	707.0	0.923	6.80	7.72	6.38	7.09
His	1	2.0	59.64	0.023	1.40	155.0	0.894	2.70	2.15	2.69	2.99
Ile	0	4.5	45.0	0.053	3.10	394.0	0.872	3.48	1.80	4.40	4.32
Leu	0	7.4	48.03	0.091	4.70	581.0	0.921	8.44	8.03	8.11	9.88
Lys	1	6.6	57.10	0.059	4.10	575.0	1.057	6.25	6.11	5.25	6.31
Met	1	1.7	69.32	0.024	1.10	132.0	0.804	2.14	3.79	1.60	1.85
Phe	1	3.6	48.52	0.040	2.30	303.0	0.914	2.73	2.93	3.52	3.72
Pro	0	5.2	36.13	0.051	2.50	366.0	0.932	6.28	7.21	5.65	6.22
Ser	0	7.0	32.40	0.069	4.50	593.0	0.923	8.53	7.25	8.04	8.05
Thr	0	6.1	35.20	0.059	3.70	490.0	0.934	4.43	3.51	7.41	5.20
Trp	1	1.3	56.92	0.014	0.80	99.0	0.803	0.80	0.47	1.68	2.10
Tyr	1	3.4	51.73	0.032	2.30	292.0	0.837	2.54	1.01	3.42	3.32
Val	0	6.6	40.35	0.066	4.20	553.0	0.913	5.44	4.57	7.00	6.19
Correlat. coeffic. R with EIIP	**0.2884**	**-0.2853**	**0.0260**	**-0.3437**	**-0.2900**	**-0.3088**	**-0.1113**	**-0.2167**	**-0.2536**	**-0.2372**	**-0.4175**

	Parameter										
Amino acid	**C193**	**C194**	**C195**	**C196**	**C197**	**C198**	**C199**	**C201**	**C203**	**C205**	**C207**
Ala	9.90	6.69	5.08	9.36	7.99	3.73	5.74	5.88	5.39	9.25	10.17
Arg	0.09	6.65	4.75	0.27	5.86	3.34	1.92	1.54	2.81	3.96	1.21
Asn	0.94	4.49	5.75	2.31	4.33	2.33	5.25	4.38	7.31	3.71	1.36
Asp	0.35	4.97	5.96	0.94	5.14	2.23	2.11	1.70	3.07	3.89	1.18
Cys	2.55	1.70	2.95	2.56	1.81	2.30	1.03	1.11	0.86	1.07	1.48
Gln	0.87	5.39	4.24	1.14	3.98	2.36	2.30	2.30	2.31	3.17	1.57
Glu	0.08	7.76	6.04	0.94	6.10	3.00	2.63	2.60	2.70	4.80	1.15
Gly	8.14	6.32	8.20	6.17	6.91	3.36	5.66	5.29	6.52	8.51	8.87
His	0.20	2.11	2.10	0.47	2.17	1.55	2.30	2.33	2.23	1.88	1.07
Ile	15.25	4.51	4.95	13.73	5.48	2.52	9.12	8.78	9.94	6.47	10.91
Leu	22.28	8.23	8.03	16.64	9.16	3.40	15.36	16.52	12.64	10.94	16.22
Lys	0.16	8.36	4.93	0.58	6.01	3.36	3.20	2.58	4.67	3.50	1.04
Met	1.85	2.46	2.61	3.93	2.50	1.37	5.30	6.00	3.68	3.14	4.12
Phe	6.47	3.59	4.36	10.99	3.83	1.94	6.51	6.58	6.34	6.36	9.60
Pro	2.38	5.20	4.84	1.96	4.95	3.18	4.79	5.29	3.62	4.36	2.24
Ser	4.17	7.40	6.41	5.58	6.84	2.83	7.55	7.68	7.24	6.26	5.38
Thr	4.33	5.18	5.87	4.68	5.77	2.63	7.51	8.38	5.44	5.66	5.61
Trp	2.21	1.06	2.31	2.20	1.34	1.15	2.51	2.89	1.64	2.22	2.67
Tyr	3.42	2.75	4.55	3.13	3.15	1.76	4.08	3.51	5.42	3.28	2.68
Val	14.34	5.27	6.07	12.43	6.65	2.53	5.12	4.66	6.18	7.55	11.44
Correlat. coeffic. R with EIIP	**-0.4705**	**-0.2435**	**-0.3155**	**-0.3614**	**-0.3217**	**-0.2961**	**-0.3504**	**-0.3045**	**-0.4518**	**-0.3847**	**-0.3652**

	Parameter	
Amino acid	**C208**	**C211**
Ala	6.61	0.88
Arg	0.41	0.99
Asn	1.84	1.02
Asp	0.59	1.16
Cys	0.83	1.14
Gln	1.20	0.93
Glu	1.63	1.01
Gly	4.88	0.70
His	1.14	1.87
Ile	12.91	1.61
Leu	21.66	1.09
Lys	1.15	0.83
Met	7.17	1.71
Phe	7.76	1.52
Pro	3.51	0.87
Ser	6.84	1.14
Thr	8.89	0.96
Trp	2.11	1.96
Tyr	2.57	1.68
Val	6.30	1.56
Correlat. coeffic. R	**-0.2983**	**0.0389**

with EIIP		

H. Hydrophobicity

H002 ARGP820101 Hydrophobicity index (Argos et al., 1982)

H003 ARGP820102 Signal sequence helical potential (Argos et al., 1982)

H004 ARGP820103 Membrane-buried preference parameters (Argos et al., 1982)

H006 BEGF750102 Conformational parameter of beta-structure (Beghin-Dirkx, 1975)

H008 BHAR880101 Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)

H010 BIOV880101 Information value for accessibility; average fraction 35% (Biou et al., 1988)

H011 BIOV880102 Information value for accessibility; average fraction 23% (Biou et al., 1988)

H012 BROC820101 Retention coefficient in TFA (Browne et al., 1982)

H013 BROC820102 Retention coefficient in HFBA (Browne et al., 1982)

H014 BULH740101 Transfer free energy to surface (Bull-Breese, 1974)

H031 CHAM820102 Free energy of solution in water, kcal/mole (Charton-Charton, 1982)

H034 CHOC760102 Residue accessible surface area in folded protein (Chothia, 1976)

H035 CHOC760103 Proportion of residues 95% buried (Chothia, 1976)

H036 CHOC760104 Proportion of residues 100% buried (Chothia, 1976)

H041 CHOP780204 Normalized frequency of N-terminal helix (Chou-Fasman, 1978b)

H054 CIDH920101 Normalized hydrophobicity scales for alpha-proteins (Cid et al., 1992)

H055 CIDH920102 Normalized hydrophobicity scales for beta-proteins (Cid et al., 1992)

H056 CIDH920103 Normalized hydrophobicity scales for alpha+beta-proteins (Cid et al., 1992)

H057 CIDH920104 Normalized hydrophobicity scales for alpha/beta-proteins (Cid et al., 1992)

H058 CIDH920105 Normalized average hydrophobicity scales (Cid et al., 1992)

H066 DESM900101 Membrane preference for cytochrome b: MPH89 (Degli Esposti et al., 1990)

H067 DESM900102 Average membrane preference: AMP07 (Degli Esposti et al., 1990)

H068 EISD840101 Consensus normalized hydrophobicity scale (Eisenberg, 1984)

H069 EISD860101 Solvation free energy (Eisenberg-McLachlan, 1986)

H070 EISD860102 Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)

H071 EISD860103 Direction of hydrophobic moment (Eisenberg-McLachlan, 1986)

H073 FASG760102 Melting point (Fasman, 1976)

H076 FASG760105 pK-C (Fasman, 1976)

H077 FAUJ830101 Hydrophobic parameter p (Fauchere-Pliska, 1983)

H085 FAUJ880108 Localized electrical effect (Fauchere et al., 1988)

H086 FAUJ880109 Number of hydrogen bond donors (Fauchere et al., 1988)

H087 FAUJ880110 Number of full nonbonding orbitals (Fauchere et al., 1988)

H088 FAUJ880111 Positive charge (Fauchere et al., 1988)

H089 FAUJ880112 Negative charge (Fauchere et al., 1988)

H092 FINA910101 Helix initiation parameter at posision i-1 (Finkelstein et al., 1991)

H094 FINA910103 Helix termination parameter at posision j-2,j-1,j (Finkelstein et al., 1991)

H095 FINA910104 Helix termination parameter at posision j+1 (Finkelstein et al., 1991)

H108 GOLD730101 Hydrophobicity factor (Goldsack-Chalifoux, 1973)

H110 GRAR740101 Composition (Grantham, 1974)

H111 GRAR740102 Polarity (Grantham, 1974)

H113 GUYH850101 Partition energy (Guy, 1985)

H114 HOPA770101 Hydration number (Hopfinger, 1971), Cited by Charton-Charton (1982)

H115 HOPT810101 Hydrophilicity value (Hopp-Woods, 1981)

H125 ISOY800107 Normalized relative frequency of double bend (Isogai et al., 1980)

H129 JANJ780101 Average accessible surface area (Janin et al., 1978)

H130 JANJ780102 Percentage of buried residues (Janin et al., 1978)

H131 JANJ780103 Percentage of exposed residues (Janin et al., 1978)

H127 JANJ790101 Ratio of buried and accessible molar fractions (Janin, 1979)

H128 JANJ790102 Transfer free energy (Janin, 1979)

H132 JOND750101 Hydrophobicity (Jones, 1975)

H133 JOND750102 pK (-COOH) (Jones, 1975)

H142 KARP850101 Flexibility parameter for no rigid neighbors (Karplus-Schulz, 1985)

H143 KARP850102 Flexibility parameter for one rigid neighbor (Karplus-Schulz, 1985)

H145 KHAG800101 The Kerr-constant increments (Khanarian-Moore, 1980)

H146 KLEP840101 Net charge (Klein et al., 1984)

H147 KRIW790101 Side chain interaction parameter (Krigbaum-Komoriya, 1979)

H148 KRIW790102 Fraction of site occupied by water (Krigbaum-Komoriya, 1979)

H150 KRIW710101 Side chain interaction parameter (Krigbaum-Rubin, 1971)

H151 KYTJ820101 Hydropathy index (Kyte-Doolittle, 1982)

H152 LAWE840101 Transfer free energy, CHP/water (Lawson et al., 1984)

H153 LEVM760101 Hydrophobic parameter (Levitt, 1976)

H170 MANP780101 Average surrounding hydrophobicity (Manavalan-Ponnuswamy, 1978)

H178 MEEJ800101 Retention coefficient in HPLC, pH7.4 (Meek, 1980)

H179 MEEJ800102 Retention coefficient in HPLC, pH2.1 (Meek, 1980)

H180 MEEJ810101 Retention coefficient in NaClO4 (Meek-Rossetti, 1981)

H181 MEEJ810102 Retention coefficient in NaH2PO4 (Meek-Rossetti, 1981)

H182 MEIH800101 Average reduced distance for Ca (Meirovitch et al., 1980)

H183 MEIH800102 Average reduced distance for side chain (Meirovitch et al., 1980)

H184 MEIH800103 Average side chain orientation angle (Meirovitch et al., 1980)

H185 MIYS850101 Effective partition energy (Miyazawa-Jernigan, 1985)

H200 NAKH900104 Normalized composition of mt-proteins (Nakashima et al., 1990)

H202 NAKH900106 Normalized composition from animal (Nakashima et al., 1990)

H204 NAKH900108 Normalized composition from fungi and plant (Nakashima et al., 1990)

H206 NAKH900110 Normalized composition of membrane proteins (Nakashima et al., 1990)

H209 NAKH900113 Ratio of average and computed composition (Nakashima et al., 1990)

H210 NISK800101 8 * contact number (Nishikawa-Ooi, 1980)

H211 NISK860101 14 * contact number (Nishikawa-Ooi, 1986)

H212 NOZY710101 Transfer energy, organic solvent/water (Nozaki-Tanford, 1971)

H213 OOBM770101 Average non-bonded energy per atom (Oobatake-Ooi, 1977)

H215 OOBM770103 Long range non-bonded energy per atom (Oobatake-Ooi, 1977)

H220 OOBM850103 Optimized transfer energy parameter (Oobatake et al., 1985)

H222 OOBM850105 Optimized side chain interaction parameter (Oobatake et al., 1985)

H233 PALJ810111 Normalized frequency of beta-sheet in a+b class (Palau et al., 1981)

H239 PARJ860101 HPLC parameter (Parker et al., 1986)

H240 PLIV810101 Partition coefficient (Pliska et al., 1981)

H241 PONP800101 Surrounding hydrophobicity in folded form (Ponnuswamy et al., 1980)

H242 PONP800102 Average gain in surrounding hydrophobicity (Ponnuswamy et al., 1980)

H243 PONP800103 Average gain ratio in surrounding hydrophobicity (Ponnuswamy et al., 1980)

H244 PONP800104 Surrounding hydrophobicity in alpha-helix (Ponnuswamy et al., 1980)

H245 PONP800105 Surrounding hydrophobicity in beta-sheet (Ponnuswamy et al., 1980)

H246 PONP800106 Surrounding hydrophobicity in turn (Ponnuswamy et al., 1980)

H247 PONP800107 Accessibility reduction ratio (Ponnuswamy et al., 1980)

H248 PONP800108 Average number of surrounding residues (Ponnuswamy et al., 1980)

H253 PRAM820101 Intercept in regression analysis (Prabhakaran-Ponnuswamy, 1982)

H249 PRAM900101 Hydrophobicity (Prabhakaran, 1990)

H271 QIAN880114 Weights for beta-sheet at the window position of -6 (Qian-Sejnowski, 1988)

H272 QIAN880115 Weights for beta-sheet at the window position of -5 (Qian-Sejnowski, 1988)

H273 QIAN880116 Weights for beta-sheet at the window position of -4 (Qian-Sejnowski, 1988)

H283 QIAN880126 Weights for beta-sheet at the window position of 6 (Qian-Sejnowski, 1988)

H284 QIAN880127 Weights for coil at the window position of -6 (Qian-Sejnowski, 1988)

H285 QIAN880128 Weights for coil at the window position of -5 (Qian-Sejnowski, 1988)

H297 RACS820101 Average relative fractional occurrence in A0(i) (Rackovsky-Scheraga, 1982)

H299 RACS820103 Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga, 1982)

H311 RACS770101 Average reduced distance for Ca (Rackovsky-Scheraga, 1977)

H312 RACS770102 Average reduced distance for side chain (Rackovsky-Scheraga, 1977)

H313 RACS770103 Side chain orientational preference (Rackovsky-Scheraga, 1977)

H314 RADA880101 Transfer free energy from chx to wat (Radzicka-Wolfenden, 1988)

H315 RADA880102 Transfer free energy from oct to wat (Radzicka-Wolfenden, 1988)

H317 RADA880104 Transfer free energy from chx to oct (Radzicka-Wolfenden, 1988)

H318 RADA880105 Transfer free energy from vap to oct (Radzicka-Wolfenden, 1988)

H320 RADA880107 Energy transfer from out to in(95%buried) (Radzicka-Wolfenden, 1988)

H321 RADA880108 Mean polarity (Radzicka-Wolfenden, 1988)

H325 RICJ880104 Relative preference value at N1 (Richardson-Richardson, 1988)

H326 RICJ880105 Relative preference value at N2 (Richardson-Richardson, 1988)

H327 RICJ880106 Relative preference value at N3 (Richardson-Richardson, 1988)

H329 RICJ880108 Relative preference value at N5 (Richardson-Richardson, 1988)

H332 RICJ880111 Relative preference value at C4 (Richardson-Richardson, 1988)

H341 ROBB760102 Information measure for N-terminal helix (Robson-Suzuki, 1976)

H339 ROBB790101 Hydration free energy (Robson-Osguthorpe, 1979)

H354 ROSG850102 Mean fractional area loss (Rose et al., 1985)

H355 ROSM880101 Side chain hydropathy, uncorrected for solvation (Roseman, 1988)

H356 ROSM880102 Side chain hydropathy, corrected for solvation (Roseman, 1988)

H357 ROSM880103 Loss of Side chain hydropathy by helix formation (Roseman, 1988)

H358 SIMZ760101 Transfer free energy (Simon, 1976), Cited by Charton-Charton (1982)

H360 SNEP660102 Principal component II (Sneath, 1966)

H364 SUEM840102 Zimm-Bragg parameter s*1.0E4 (Sueki et al., 1984)

H365 SWER830101 Optimal matching hydrophobicity (Sweet-Eisenberg, 1983)

H371 TANS770106 Normalized frequency of chain reversal D (Tanaka-Scheraga, 1977)

H373 TANS770108 Normalized frequency of zR (Tanaka-Scheraga, 1977)

H377 VASM830102 Relative population of conformational state C (Vasquez et al., 1983)

H378 VASM830103 Relative population of conformational state E (Vasquez et al., 1983)

H379 VELV850101 Electron-ion interaction potential (Veljkovic et al., 1985)

H380 VENT840101 Bitterness (Venanzi, 1984)

H381 VHEG790101 Transfer free energy to lipophilic phase (von Heijne-Blomberg, 1979)

H382 WARP780101 Average interactions per side chain atom (Warme-Morgan, 1978)

H384 WERD780101 Propensity to be buried inside (Wertz-Scheraga, 1978)

H386 WERD780103 Free energy change of a(Ri) to a(Rh) (Wertz-Scheraga, 1978)

H387 WERD780104 Free energy change of e(i) to a(Rh) (Wertz-Scheraga, 1978)

H388 WOEC730101 Polar requirement (Woese, 1973)

H392 WOLR810101 Hydration potential (Wolfenden et al., 1981)

H389 WOLS870101 Principal property value z1 (Wold et al., 1987)

H393 YUTK870101 Unfolding Gibbs energy in water, pH7.0 (Yutani et al., 1987)

H394 YUTK870102 Unfolding Gibbs energy in water, pH9.0 (Yutani et al., 1987)

H395 YUTK870103 Activation Gibbs energy of unfolding, pH7.0 (Yutani et al., 1987)

H396 YUTK870104 Activation Gibbs energy of unfolding, pH9.0 (Yutani et al., 1987)

H398 ZIMJ680101 Hydrophobicity (Zimmerman et al., 1968)

H400 ZIMJ680103 Polarity (Zimmerman et al., 1968)

H401 ZIMJ680104 Isoelectric point (Zimmerman et al., 1968)

H402 ZIMJ680105 RF rank (Zimmerman et al., 1968)

	Parameter										
Amino acid	**H002**	**H003**	**H004**	**H006**	**H008**	**H010**	**H011**	**H012**	**H013**	**H014**	**H031**
Ala	0.61	1.18	1.56	0.77	0.357	16.0	44.0	7.30	3.9	-0.20	-0.368
Arg	0.60	0.20	0.45	0.72	0.529	-70.0	-68.0	-3.60	3.2	-0.12	-1.030
Asn	0.06	0.23	0.27	0.55	0.463	-74.0	-72.0	-5.70	-2.8	0.08	0.00
Asp	0.46	0.05	0.14	0.65	0.511	-78.0	-91.0	-2.90	-2.8	-0.20	2.06
Cys	1.07	1.89	1.23	0.65	0.346	168.0	90.0	-9.20	-14.3	-0.45	4.53
Gln	0.00	0.72	0.51	0.72	0.493	-73.0	-117.0	-0.30	1.8	0.16	0.731
Glu	0.47	0.11	0.23	0.55	0.497	-106.0	-139.0	-7.10	-7.5	-0.30	1.770
Gly	0.07	0.49	0.62	0.65	0.544	-13.0	-8.0	-1.20	-2.3	0.00	-0.525
His	0.61	0.31	0.29	0.83	0.323	50.0	47.0	-2.10	2.0	-0.12	0.00
Ile	2.22	1.45	1.67	0.98	0.462	151.0	100.0	6.60	11.0	-2.26	0.791
Leu	1.53	3.23	2.93	0.83	0.365	145.0	108.0	20.0	15.0	-2.46	1.070
Lys	1.15	0.06	0.15	0.55	0.466	-141.0	-188.0	-3.70	-2.5	-0.35	0.00
Met	1.18	2.67	2.96	0.98	0.295	124.0	121.0	5.60	4.1	-1.47	0.656
Phe	2.02	1.96	2.03	0.98	0.314	189.0	148.0	19.20	14.7	-2.33	1.060
Pro	1.95	0.76	0.76	0.55	0.509	-20.0	-36.0	5.10	5.6	-0.98	-2.240
Ser	0.05	0.97	0.81	0.55	0.507	-70.0	-60.0	-4.10	-3.5	-0.39	-0.524
Thr	0.05	0.84	0.91	0.83	0.444	-38.0	-54.0	0.80	1.1	-0.52	0.00
Trp	2.65	0.77	1.08	0..77	0.305	145.0	163.0	16.30	17.8	-2.01	1.600
Tyr	1.88	0.39	0.68	0.83	0.420	53.0	22.0	5.90	3.8	-2.24	4.910
Val	1.32	1.08	1.14	0.98	0.386	123.0	117.0	3.50	2.1	-1.56	0.401
Correlat. coeffic. R with EIIP	**-0.1652**	**-0.0087**	**-0.0288**	**0.0410**	**-0.0417**	**-0.0796**	**-0.0889**	**-0.0966**	**-0.1164**	**0.1473**	**0.2063**

	Parameter										
Amino acid	**H034**	**H035**	**H036**	**H041**	**H054**	**H055**	**H056**	**H057**	**H058**	**H066**	**H067**
Ala	25.0	0.38	0.20	1.29	-0.45	-0.08	0.36	0.17	0.02	1.56	1.26
Arg	90.0	0.01	0.00	0.44	-0.24	-0.09	-0.52	-0.70	-0.42	0.59	0.38
Asn	63.0	0.12	0.03	0.81	-0.20	-0.70	-0.90	-0.90	-0.77	0.51	0.59
Asp	50.0	0.15	0.04	2.02	-1.52	-0.71	-1.09	-1.05	-1.04	0.23	0.27
Cys	19.0	0.45	0.22	0.66	0.79	0.76	0.70	1.24	0.77	1.80	1.60
Gln	71.0	0.07	0.01	1.22	-0.99	-0.40	-1.05	-1.20	-1.10	0.39	0.39
Glu	49.0	0.18	0.03	2.44	-0.80	-1.31	-0.83	-1.19	-1.14	0.19	0.23
Gly	23.0	0.36	0.18	0.76	-1.00	-0.84	-0.82	-0.57	-0.80	1.03	1.08
His	43.0	0.17	0.02	0.73	1.07	0.43	0.16	-0.25	0.26	1.00	1.00
Ile	18.0	0.60	0.19	0.67	0.76	1.39	2.17	2.06	1.81	1.27	1.44
Leu	23.0	0.45	0.16	0.58	1.29	1.24	1.18	0.96	1.14	1.38	1.36
Lys	97.0	0.03	0.00	0.66	-0.36	-0.09	-0.56	-0.62	-0.41	0.15	0.33
Met	31.0	0.40	0.11	0.71	1.37	1.27	1.21	0.60	1.00	1.93	1.52
Phe	24.0	0.50	0.14	0.61	1.48	1.53	1.01	1.29	1.35	1.42	1.46
Pro	50.0	0.18	0.04	2.01	-0.12	-0.01	-0.06	-0.21	-0.09	0.27	0.54
Ser	44.0	0.22	0.08	0.74	-0.98	-0.93	-0.60	-0.83	-0.97	0.96	0.98
Thr	47.0	0.23	0.08	1.08	-0.70	-0.59	-1.20	-0.62	-0.77	1.11	1.01
Trp	32.0	0.27	0.04	1.47	1.38	2.25	1.31	1.51	1.71	0.91	1.06
Tyr	60.0	0.15	0.03	0.68	1.49	1.53	1.05	0.66	1.11	1.10	0.89
Val	18.0	0.54	0.18	0.61	1.26	1.09	1.21	1.21	1.13	1.58	1.33
Correlat. coeffic. R with EIIP	**0.2237**	**-0.2468**	**-0.2078**	**0.0110**	**-0.1906**	**-0.0413**	**-0.2483**	**-0.1743**	**-0.1755**	**0.0215**	**-0.0866**

	Parameter										
Amino acid	**H068**	**H069**	**H070**	**H071**	**H073**	**H076**	**H077**	**H085**	**H086**	**H087**	**H088**
Ala	0.25	0.67	0.0	0.00	297.0	2.34	0.31	-0.01	0.0	0.0	0.0
Arg	-1.76	-2.10	10.0	-0.96	238.0	1.82	-1.01	0.04	4.0	3.0	1.0
Asn	-0.64	-0.60	1.3	-0.86	236.0	2.02	-0.60	0.06	2.0	3.0	0.0
Asp	-0.72	-1.20	1.9	-0.98	270.0	1.88	-0.77	0.15	1.0	4.0	0.0
Cys	0.04	0.38	0.17	0.76	178.0	1.92	1.54	0.12	0.0	0.0	0.0
Gln	-0.69	-0.22	1.9	-1.00	185.0	2.17	-0.22	0.05	2.0	3.0	0.0
Glu	-0.62	-0.76	3.0	-0.89	249.0	2.10	-0.64	0.07	1.0	4.0	0.0
Gly	0.16	0.00	0.0	0.00	290.0	2.35	0.00	0.00	0.0	0.0	0.0
His	-0.40	0.64	0.99	-0.75	277.0	1.82	0.13	0.08	1.0	1.0	1.0
Ile	0.73	1.90	1.2	0.99	284.0	2.36	1.80	-0.01	0.0	0.0	0.0
Leu	0.53	1.90	1.0	0.89	337.0	2.36	1.70	-0.01	0.0	0.0	0.0
Lys	-1.10	-0.57	5.7	-0.99	224.0	2.16	-0.99	0.00	2.0	1.0	1.0
Met	0.26	2.40	1.9	0.94	283.0	2.28	1.23	0.04	0.0	0.0	0.0
Phe	0.61	2.30	1.1	0.92	284.0	2.16	1.79	0.03	0.0	0.0	0.0
Pro	-0.07	1.20	0.18	0.22	222.0	1.95	0.72	0.00	0.0	0.0	0.0
Ser	-0.26	0.01	0.73	-0.67	228.0	2.19	-0.04	0.11	1.0	2.0	0.0
Thr	-0.18	0.52	1.5	0.09	253.0	2.09	0.26	0.04	1.0	2.0	0.0
Trp	0.37	2.6600	1.6	0.67	282.0	2.43	2.25	0.0	1.0	0.0	0.0
Tyr	0.02	1.60	1.8	-0.93	344.0	2.20	0.96	0.03	1.0	2.0	0.0
Val	0.54	1.50	0.48	0.84	293.0	2.32	1.22	0.01	0.0	0.0	0.0
Correlat. coeffic. R with EIIP	**-0.3040**	**-0.2000**	**0.2711**	**-0.1507**	**-0.2916**	**-0.3778**	**-0.1241**	**0.5643**	**0.2816**	**0.3077**	**0.0364**

	Parameter										
Amino acid	**H089**	**H092**	**H094**	**H095**	**H108**	**H110**	**H111**	**H113**	**H114**	**H115**	**H125**
Ala	0.0	1.0	1.2	1.0	0.75	0.00	8.1	0.10	1.0	-0.5	1.34
Arg	0.0	0.7	1.7	1.7	0.75	0.65	10.5	1.91	2.3	3.0	2.78
Asn	0.0	1.7	1.2	1.0	0.69	1.33	11.6	0.48	2.2	0.2	0.92
Asp	1.0	3.2	0.7	0.7	0.00	1.38	13.0	0.78	6.5	3.0	1.77
Cys	0.0	1.0	1.0	1.0	1.00	2.75	5.5	-1.42	0.1	-1.0	1.44
Gln	0.0	1.0	1.0	1.0	0.59	0.89	10.5	0.95	2.1	0.2	0.79
Glu	1.0	1.7	0.7	0.7	0.00	0.92	12.3	0.83	6.2	3.0	2.54
Gly	0.0	1.0	0.8	1.5	0.00	0.74	9.0	0.33	1.1	0.0	0.95
His	0.0	1.0	1.2	1.0	0.00	0.58	10.4	-0.50	2.8	-0.5	0.00
Ile	0.0	0.6	0.8	1.0	2.95	0.00	5.2	-1.13	0.8	-1.8	0.52
Leu	0.0	1.0	1.0	1.0	2.40	0.00	4.9	-1.18	0.8	-1.8	1.05
Lys	0.0	0.7	1.7	1.7	1.50	0.33	11.3	1.40	5.3	3.0	0.79
Met	0.0	1.0	1.0	1.0	1.30	0.00	5.7	-1.59	0.7	-1.3	0.00
Phe	0.0	1.0	1.0	1.0	2.65	0.00	5.2	-2.12	1.4	-2.5	0.43
Pro	0.0	1.0	1.0	0.1	2.60	0.39	8.0	0.73	0.9	0.0	0.37
Ser	0.0	1.7	1.5	1.0	0.00	1.42	9.2	0.52	1.7	0.3	0.87
Thr	0.0	1.7	1.0	1.0	0.45	0.71	8.6	0.07	1.5	-0.4	1.14
Trp	0.0	1.0	1.0	1.0	3.00	0.13	5.4	-0.51	1.9	-3.4	1.79
Tyr	0.0	1.0	1.0	1.0	2.85	0.20	6.2	-0.21	2.1	-2.3	0.73
Val	0.0	0.6	0.8	1.0	1.70	0.00	5.9	-1.27	0.9	-1.5	0.00
Correlat. coeffic. R with EIIP	**0.1449**	**0.4195**	**0.1919**	**0.0711**	**-0.2056**	**0.3244**	**0.1057**	**0.0771**	**0.1411**	**0.1620**	**0.2453**

	Parameter										
Amino acid	**H129**	**H130**	**H131**	**H127**	**H128**	**H132**	**H133**	**H142**	**H143**	**H145**	**H146**
Ala	27.8	51.0	15.0	1.7	0.3	0.87	2.34	1.041	0.946	49.1	0.0
Arg	94.7	5.0	67.0	0.1	-1.4	0.85	1.18	1.038	1.028	133.0	1.0
Asn	60.1	22.0	49.0	0.4	-0.5	0.09	2.02	1.117	1.006	-3.6	0.0
Asp	60.6	19.0	50.0	0.4	-0.6	0.66	2.01	1.033	1.089	0.0	-1.0
Cys	15.5	74.0	5.0	4.6	0.9	1.52	1.65	0.960	0.878	0.0	0.0
Gln	68.7	16.0	56.0	0.3	-0.7	0.00	2.17	1.165	1.025	20.0	0.0
Glu	68.2	16.00	55.0	0.3	-0.7	0.67	2.19	1.094	1.036	0.0	-1.0
Gly	24.5	52.0	10.0	1.8	0.3	0.10	2.34	1.142	1.042	64.6	0.0
His	50.7	34.0	34.0	0.8	-0.1	0.87	1.82	0.982	0.952	75.7	0.0
Ile	22.8	66.0	13.0	3.1	0.7	3.15	2.36	1.002	0.892	18.9	0.0
Leu	27.6	60.0	16.0	2.4	0.5	2.17	2.36	0.967	0.961	15.6	0.0
Lys	103.0	3.0	85.0	0.05	-1.8	1.64	2.18	1.093	1.082	0.0	1.0
Met	33.5	52.0	20.0	1.9	0.4	1.67	2.28	0.947	0.862	6.8	0.0
Phe	25.5	58.0	10.0	2.2	0.5	2.87	1.83	0.930	0.912	54.7	0.0
Pro	51.5	25.0	45.0	0.6	-0.3	2.77	1.99	1.055	1.085	43.8	0.0
Ser	42.0	35.0	32.0	0.8	-0.1	0.07	2.21	1.169	1.048	44.4	0.0
Thr	45.0	30.0	32.0	0.7	-0.2	0.07	2.10	1.073	1.051	31.0	0.0
Trp	34.7	49.0	17.0	1.6	0.3	3.77	2.38	0.925	0.917	70.5	0.0
Tyr	55.2	24.0	41.0	0.5	-0.4	2.67	2.20	0.961	0.930	0.0	0.0
Val	23.7	64.0	14.0	2.9	0.6	1.87	2.32	0.982	0.927	29.5	0.0
Correlat. coeffic. R	**0.1648**	**-0.1972**	**0.1266**	**-0.1405**	**-0.1708**	**-0.1651**	**-0.4561**	**-0.0856**	**0.0789**	**0.1281**	**0.0030**

with EIIP											

	Parameter										
Amino acid	**H147**	**H148**	**H150**	**H151**	**H152**	**H153**	**H170**	**H178**	**H179**	**H180**	**H181**
Ala	0.28	4.32	4.6	1.8	-0.48	-0.5	12.97	0.5	-0.1	1.1	1.0
Arg	0.34	6.55	6.5	-4.5	-0.06	3.0	11.72	0.8	-4.5	-0.4	-2.0
Asn	0.31	6.24	5.9	-3.5	-0.87	0.2	11.42	0.8	-1.6	-4.2	-3.0
Asp	0.33	6.04	5.7	-3.5	-0.75	2.5	10.85	-8.2	-2.8	-1.6	-0.5
Cys	0.11	1.73	-1.0	2.5	-0.32	-1.0	14.63	-6.8	-2.2	7.1	4.6
Gln	0.39	6.13	6.1	-3.5	-0.32	0.2	11.76	-4.8	-2.5	-2.9	-2.0
Glu	0.37	6.17	5.6	-3.5	-0.71	2.5	11.89	-16.9	-7.5	0.7	1.1
Gly	0.28	6.09	7.6	-0.4	0.0	0.0	12.43	0.0	-0.5	-0.2	0.2
His	0.23	5.66	4.5	-3.2	-0.51	-0.5	12.16	-3.5	0.8	-0.7	-2.2
Ile	0.12	2.31	2.6	4.5	0.81	-1.8	15.67	13.9	11.8	8.5	7.0
Leu	0.16	3.93	3.25	3.8	1.02	-1.8	14.90	8.8	10.0	11.0	9.6
Lys	0.59	7.92	7.9	-3.9	-0.09	3.0	11.36	0.1	-3.2	-1.9	-3.0
Met	0.08	2.44	1.4	1.9	0.81	-1.3	14.39	4.8	7.1	5.4	4.0
Phe	0.10	2.59	3.2	2.8	1.03	-2.5	14.00	13.2	13.9	13.4	12.6
Pro	0.46	7.19	7.0	-1.6	2.03	-1.4	11.37	6.1	8.0	4.4	3.1
Ser	0.27	5.37	5.25	-0.8	0.05	0.3	11.23	1.2	-3.7	-3.2	-2.9
Thr	0.26	5.16	4.8	-0.7	-0.35	-0.4	11.69	2.7	1.5	-1.7	-0.6
Trp	0.15	2.78	4.0	-0.9	0.66	-3.4	13.93	14.9	18.1	17.1	15.1
Tyr	0.25	3.58	4.35	-1.3	1.24	-2.3	13.42	6.1	8.2	7.4	6.7
Val	0.22	3.31	3.4	4.2	0.56	-1.5	15.71	2.7	3.3	5.9	4.6
Correlat.											

coeffic. R *with EIIP*	**-0.0981**	**-0.0880**	**-0.1488**	**-0.2091**	**-0.1614**	**0.1847**	**-0.2769**	**-0.0949**	**-0.1191**	**-0.0868**	**-0.0706**

	Parameter										
Amino acid	**H182**	**H183**	**H184**	**H185**	**H200**	**H202**	**H204**	**H206**	**H209**	**H210**	**H211**
Ala	0.93	0.94	87.0	2.36	-0.6	-0.57	-0.70	0.34	1.61	0.23	-0.22
Arg	0.98	1.09	81.0	1.92	-1.18	-1.29	-0.91	-0.57	0.40	-0.26	-0.93
Asn	0.98	1.04	70.0	1.70	0.39	0.02	1.28	-0.27	0.73	-0.94	-2.65
Asp	1.01	1.08	71.0	1.67	-1.36	-1.54	-0.93	-0.56	0.75	-1.13	-4.12
Cys	0.88	0.84	104.0	3.36	-0.34	-0.3	-0.41	-0.32	0.37	1.78	4.66
Gln	1.02	1.11	66.0	1.75	-0.71	-0.71	-0.71	-0.34	0.61	-0.57	-2.76
Glu	1.02	1.12	72.0	1.74	-1.16	-1.17	-1.13	-0.43	1.50	-0.75	-3.64
Gly	1.01	1.01	90.0	2.06	-0.37	-0.48	-0.12	0.48	3.12	-0.07	-1.62
His	0.89	0.92	90.0	2.41	0.08	0.10	0.04	-0.19	0.46	0.11	1.28
Ile	0.79	0.76	105.0	4.17	1.44	1.31	1.77	0.39	1.61	1.19	5.58
Leu	0.85	0.82	104.0	3.93	1.82	2.16	1.02	0.52	1.37	1.03	5.01
Lys	1.05	1.23	65.0	1.23	-0.84	-1.02	-0.40	-0.75	0.62	-1.05	-4.18
Met	0.84	0.833	100.0	4.22	2.04	2.55	0.86	0.47	1.59	0.66	3.51
Phe	0.78	0.73	108.0	4.37	1.38	1.42	1.29	1.30	1.24	0.48	5.27
Pro	1.00	1.04	78.0	1.89	-0.05	0.11	-0.42	-0.19	0.67	-0.76	-3.03
Ser	1.02	1.04	83.0	1.81	0.25	0.30	0.14	-0.20	0.68	-0.67	-2.84
Thr	0.99	1.02	83.0	2.04	0.66	0.99	-0.13	-0.04	0.92	-0.36	-1.20
Trp	0.83	0.87	94.0	3.82	1.02	1.35	0.26	0.77	1.63	0.90	5.20
Tyr	0.93	1.03	83.0	2.91	0.53	0.20	1.29	0.07	0.67	0.59	2.15

Val	0.81	0.81	94.0	3.49	-0.60	-0.79	-0.19	0.36	1.30	1.24	4.45
Correlat. coeffic. R with EIIP	**0.1241**	**0.1117**	**-0.0663**	**-0.0569**	**-0.1066**	**-0.0540**	**-0.2483**	**-0.1364**	**-0.4067**	**-0.1477**	**-0.1044**

	Parameter										
Amino acid	**H212**	**H213**	**H215**	**H220**	**H222**	**H233**	**H239**	**H240**	**H241**	**H242**	**H243**
Ala	0.5	-1.895	-0.491	0.46	4.55	0.82	2.1	-2.89	12.28	7.62	2.63
Arg	0.0	-1.475	-0.554	-1.54	5.97	0.99	4.2	-3.30	11.49	6.81	2.45
Asn	0.0	-1.560	-0.382	1.31	5.56	1.27	7.0	-3.41	11.00	6.17	2.27
Asp	0.0	-1.518	-0.356	-0.33	2.85	0.98	10.0	-3.38	10.97	6.18	2.29
Cys	0.0	-2.035	-0.670	0.20	-0.78	0.71	1.4	-2.49	14.93	10.93	3.36
Gln	0.0	-1.521	-0.405	-1.12	4.15	1.01	6.0	-3.15	11.28	6.67	2.45
Glu	0.0	-1.535	-0.371	0.48	5.16	0.54	7.8	-2.94	11.19	6.38	2.31
Gly	0.0	-1.898	-0.534	0.64	9.14	0.94	5.7	-3.25	12.01	7.31	2.55
His	0.5	-1.755	-0.540	-1.31	4.48	1.26	2.1	-2.84	12.84	7.85	2.57
Ile	1.8	-1.951	-0.762	3.28	2.10	1.67	-8.0	-1.72	14.77	9.99	3.08
Leu	1.8	-1.966	-0.650	0.43	3.24	0.94	-9.2	-1.61	14.10	9.37	2.98
Lys	0.0	-1.374	-0.300	-1.71	10.68	0.73	5.7	-3.31	10.80	5.72	2.12
Met	1.30	-1.963	-0.659	0.15	2.18	1.30	-4.2	-1.84	14.33	9.83	3.18
Phe	2.5	-1.864	-0.729	0.52	4.37	1.56	-9.2	-1.63	13.43	8.99	3.02
Pro	0.0	-1.699	-0.463	-0.58	5.14	0.69	2.1	-2.50	11.19	6.64	2.46
Ser	0.0	-1.753	-0.455	-0.83	6.78	0.65	6.5	-3.30	11.26	6.93	2.60
Thr	0.4	-1.767	-0.515	-1.52	8.60	0.98	5.2	-2.91	11.65	7.08	2.55

Trp	3.4	-1.869	-0.839	1.25	1.97	1.25	-10.0	-1.75	12.95	8.41	2.85
Tyr	2.3	-1.686	-0.656	-2.21	2.40	1.26	-1.9	-2.42	13.29	8.53	2.79
Val	1.5	-1.981	-0.728	0.54	3.81	1.22	-3.7	-2.08	15.07	10.38	3.21
Correlat. coeffic. R with EIIP	**-0.0722**	**0.1883**	**0.0532**	**-0.4396**	**-0.1213**	**-0.0340**	**0.1868**	**-0.1726**	**-0.1692**	**-0.0920**	**0.0008**

	Parameter										
Amino acid	**H244**	**H245**	**H246**	**H247**	**H248**	**H253**	**H249**	**H271**	**H272**	**H273**	**H283**
Ala	13.65	14.60	10.67	3.70	6.05	0.305	-6.7	-0.18	-0.01	-0.19	-0.06
Arg	11.28	13.24	11.05	2.53	5.70	0.227	51.5	-0.13	0.02	0.03	0.02
Asn	12.24	11.79	10.85	2.12	5.04	0.322	20.1	0.28	0.41	0.02	0.10
Asp	10.98	13.78	10.21	2.60	4.95	0.335	38.5	0.05	-0.09	-0.06	0.24
Cys	14.49	15.50	14.15	3.03	7.86	0.339	-8.4	-0.26	-0.27	-0.29	-0.19
Gln	11.30	12.02	11.71	2.70	5.45	0.306	17.2	0.21	0.01	0.02	-0.04
Glu	12.55	13.59	11.71	3.30	5.10	0.282	34.3	-0.06	0.09	-0.10	-0.04
Gly	15.36	14.18	10.95	3.13	6.16	0.352	-4.2	0.23	0.13	0.19	0.17
His	11.59	15.35	12.07	3.57	5.8	0.215	12.6	0.24	0.22	-0.16	0.19
Ile	14.63	14.10	12.95	7.69	7.51	0.278	-13.0	-0.42	-0.27	-0.08	-0.20
Leu	14.01	16.49	13.07	5.88	7.37	0.262	-11.7	-0.23	-0.25	-0.42	-0.46
Lys	11.96	13.28	9.93	1.79	4.88	0.391	36.8	0.03	0.08	-0.09	-0.43
Met	13.40	16.23	15.00	5.21	6.39	0.280	-14.2	-0.42	-0.57	-0.38	-0.52
Phe	14.08	14.18	13.27	6.60	6.62	0.195	-15.5	-0.18	-0.12	-0.32	-0.33
Pro	11.51	14.10	10.62	2.12	5.65	0.346	0.8	-0.13	0.26	0.05	0.37
Ser	11.26	13.36	11.18	2.43	5.53	0.326	-2.5	0.41	0.44	0.25	0.43

Thr	13.00	14.50	10.53	2.60	5.81	0.251	-5.0	0.33	0.35	0.22	0.50
Trp	12.06	13.90	11.41	6.25	6.98	0.291	-7.9	-0.10	-0.15	-0.19	-0.32
Tyr	12.64	14.76	11.52	3.03	6.73	0.293	2.9	-0.10	0.15	0.05	0.35
Val	12.88	16.30	13.86	7.14	7.62	0.291	-10.9	-0.07	-0.09	-0.15	0.00
Correlat. coeffic. R with EIIP	**-0.3288**	**-0.1091**	**-0.0449**	**-0.2478**	**-0.1824**	**-0.1512**	**0.1803**	**0.0762**	**-0.1122**	**0.0657**	**0.1200**

	Parameter										
Amino acid	**H284**	**H285**	**H297**	**H299**	**H311**	**H312**	**H313**	**H314**	**H315**	**H317**	**H318**
Ala	-0.05	-0.19	0.85	0.82	0.934	0.941	1.16	1.81	0.52	1.29	1.42
Arg	0.06	0.17	2.02	2.60	0.962	1.112	1.72	-14.92	-1.32	-13.6	-18.6
Asn	0.00	-0.38	0.88	2.07	0.986	1.038	1.97	-6.64	-0.01	-6.63	-9.67
Asp	0.15	0.09	1.50	2.64	0.994	1.071	2.66	-8.72	0.00	0.00	0.00
Cys	0.30	0.41	0.90	0.00	0.900	0.866	0.50	1.28	0.00	0.00	0.00
Gln	-0.08	0.04	1.71	0.00	1.047	1.150	3.87	-5.54	-0.07	-5.47	-9.31
Glu	-0.02	-0.20	1.79	2.62	0.986	1.100	2.40	-6.81	-0.79	-6.02	-9.45
Gly	-0.14	0.28	1.54	1.63	1.015	1.055	1.63	0.94	0.00	0.94	2.39
His	-0.07	-0.19	1.59	0.00	0.882	0.911	0.86	-4.66	0.95	-5.61	-11.22
Ile	0.26	-0.06	0.67	2.32	0.766	0.742	0.57	4.92	2.04	2.88	0.11
Leu	0.04	0.34	1.03	0.00	0.825	0.798	0.51	4.92	1.76	3.16	0.52
Lys	-0.42	-0.20	0.88	2.86	1.040	1.232	3.90	-5.55	0.08	-5.63	-9.60
Met	0.25	0.45	1.17	0.00	0.804	0.781	0.40	2.35	1.32	1.03	-2.80

Phe	0.09	0.07	0.85	0.00	0.773	0.723	0.43	2.98	2.09	0.89	-2.85
Pro	0.31	0.04	1.47	0.00	1.047	1.093	2.04	0.00	0.00	0.00	0.00
Ser	-0.11	-0.23	1.50	1.23	1.056	1.082	1.61	-3.40	0.04	-3.44	-5.10
Thr	-0.06	-0.02	1.96	2.48	1.008	1.043	1.48	-2.57	0.27	-2.84	-5.15
Trp	0.19	0.16	0.83	0.00	0.848	0.867	0.75	2.33	2.51	-0.18	-8.39
Tyr	0.33	0.22	1.34	1.90	0.931	1.050	1.72	-0.14	1.63	-1.77	-7.74
Val	0.04	0.05	0.89	1.62	0.825	0.817	0.59	4.04	1.18	2.86	0.81
Correlat. coeffic. R with EIIP	**0.1342**	**0.2660**	**0.3196**	**-0.0249**	**0.1363**	**0.1208**	**0.1076**	**-0.3350**	**-0.1604**	**-0.2075**	**-0.1802**

	Parameter										
Amino acid	**H320**	**H321**	**H325**	**H326**	**H327**	**H329**	**H332**	**H341**	**H339**	**H354**	**H355**
Ala	-0.29	-0.06	1.2	1.6	1.0	1.4	1.3	2.3	-1.0	0.74	-0.67
Arg	-2.71	-0.84	0.7	0.9	0.4	1.2	0.8	-5.2	0.3	0.64	12.10
Asn	-1.18	-0.48	0.7	0.7	0.7	1.2	0.6	0.3	-0.7	0.63	7.23
Asp	-1.02	-0.80	0.8	2.6	2.2	0.6	0.5	7.4	-1.2	0.62	8.72
Cys	0.00	1.36	0.8	1.2	0.6	1.6	0.7	0.8	2.1	0.91	-0.34
Gln	-1.53	-0.73	0.7	0.8	1.5	1.4	0.2	-0.7	-0.1	0.62	6.39
Glu	-0.90	-0.77	2.2	2.0	3.3	0.9	0.7	10.3	-0.7	0.62	7.35
Gly	-0.34	-0.41	0.3	0.9	0.6	0.6	0.5	-5.2	0.3	0.72	0.00
His	-0.94	0.49	0.7	0.7	0.7	0.9	1.9	-2.8	1.1	0.78	3.82
Ile	0.24	1.31	0.9	0.7	0.4	0.9	1.6	-4.0	4.0	0.88	-3.02
Leu	-0.12	1.21	0.9	0.3	0.66	1.1	1.4	-2.1	2.0	0.85	-3.02

Lys	-2.05	-1.18	0.6	1.0	0.8	1.9	1.0	-4.1	-0.9	0.52	6.13
Met	-0.24	1.27	0.3	1.0	1.0	1.7	2.8	-3.5	1.8	0.85	-1.30
Phe	0.00	1.27	0.5	0.9	0.6	1.0	2.9	-1.1	2.8	0.88	-3.24
Pro	0.00	0.00	2.6	0.5	0.4	0.3	0.0	8.1	0.4	0.64	-1.75
Ser	-0.75	-0.50	0.7	0.8	0.4	1.1	0.5	-3.5	-1.2	0.66	4.35
Thr	-0.71	-0.27	0.8	0.7	1.0	0.6	0.6	2.3	-0.5	0.70	3.86
Trp	-0.59	0.88	2.1	1.7	1.4	1.4	2.1	-0.9	3.0	0.85	-2.86
Tyr	-1.02	0.33	1.8	0.4	1.2	0.2	0.8	-3.7	2.1	0.76	0.98
Val	0.09	1.09	1.1	0.6	1.1	0.8	1.4	-4.4	1.4	0.86	-2.18
Correlat. coeffic. R with EIIP	**-0.2936**	**-0.1172**	**-0.2847**	**0.3535**	**0.0606**	**0.1376**	**0.0450**	**0.0732**	**-0.1540**	**-0.0803**	**0.3242**

	Parameter										
Amino acid	**H356**	**H357**	**H358**	**H360**	**H364**	**H365**	**H371**	**H373**	**H377**	**H378**	**H379= EIIP**
Ala	-0.67	0.4	0.73	0.330	8.0	-0.40	0.937	0.328	0.507	0.159	0.0373
Arg	3.89	0.3	0.73	-0.176	0.1	-0.59	1.725	2.088	0.459	0.194	0.0959
Asn	2.27	0.9	-0.01	-0.233	0.1	-0.92	1.080	1.498	0.287	0.385	0.0036
Asp	1.57	0.8	0.54	-0.371	10.0	-1.31	1.640	3.379	0.223	0.283	0.1263
Cys	-2.00	0.5	0.70	0.074	26.0	0.17	1.004	0.000	0.592	0.187	0.0829
Gln	2.12	0.7	-0.10	-0.254	33.0	-0.91	1.078	0.000	0.383	0.236	0.0761
Glu	1.78	1.3	0.55	-0.409	6.00	-1.22	0.679	0.000	0.445	0.206	0.0058
Gly	0.00	0.0	0.00	0.370	0.1	-0.67	0.901	0.500	0.390	0.049	0.0050
His	1.09	1.0	1.10	-0.078	0.1	-0.64	1.085	1.204	0.310	0.233	0.0242

Ile	-3.02	0.4	2.97	0.149	55.0	1.25	0.178	2.078	0.111	0.581	0.0000
Leu	-3.02	0.6	2.49	0.129	33.0	1.22	0.869	0.414	0.619	0.083	0.0000
Lys	2.46	0.4	1.50	-0.075	1.0	-0.67	1.254	0.835	0.559	0.159	0.0371
Met	-1.67	0.3	1.30	-0.092	54.0	1.02	0.886	0.982	0.431	0.198	0.0823
Phe	-3.24	0.7	2.65	-0.011	18.0	1.92	0.803	1.336	0.077	0.682	0.0946
Pro	-1.75	0.9	2.60	0.370	42.0	-0.49	0.748	0.415	0.739	0.366	0.0198
Ser	0.10	0.4	0.04	0.022	0.1	-0.55	1.145	1.089	0.689	0.150	0.0829
Thr	-0.42	0.4	0.44	0.136	0.1	-0.28	1.487	1.732	0.785	0.074	0.0941
Trp	-2.86	0.6	3.00	-0.011	77.0	0.50	0.803	1.781	0.160	0.463	0.0548
Tyr	0.98	1.2	2.97	-0.138	66.0	1.67	1.227	0.000	0.060	0.737	0.0516
Val	-2.18	0.4	1.69	0.245	0.1	0.91	0.625	0.946	0.356	0.301	0.0057
Correlat. coeffic. R with EIIP	**0.1647**	**-0.1275**	**-0.2191**	**-0.3565**	**0.2163**	**-0.0576**	**0.6654**	**0.3970**	**0.0270**	**0.0007**	**1.0000**

	Parameter										
Amino acid	**H380**	**H381**	**H382**	**H384**	**H386**	**H387**	**H388**	**H392**	**H389**	**H393**	**H394**
Ala	0.0	-12.04	10.04	0.52	0.15	-0.07	7.0	1.94	0.07	8.5	6.8
Arg	0.0	39.23	6.18	0.49	-0.37	-0.40	9.1	-19.92	2.88	0.0	0.0
Asn	0.0	4.25	5.63	0.42	0.69	-0.57	10.0	-9.68	3.22	8.2	6.2
Asp	0.0	23.22	5.76	0.37	-0.22	-0.80	13.0	-10.95	3.64	8.5	7.0
Cys	0.0	3.95	8.89	0.83	-0.19	0.17	6.5	-1.24	0.71	11.0	8.3
Gln	0.0	2.16	5.41	0.35	-0.06	-0.26	8.6	-9.38	2.18	6.3	8.5
Glu	0.0	16.81	5.37	0.38	0.14	-0.63	12.5	-10.20	3.08	8.8	4.9

BioElectronics Group, Department of Electrical and Computer Systems Engineering

Gly	0.0	-7.85	7.99	0.41	0.36	0.27	7.9	2.39	2.23	7.1	6.4
His	0.0	6.28	7.49	0.70	-0.25	-0.49	8.4	-10.27	2.41	10.1	9.2
Ile	1.0	-18.32	8.72	0.79	0.02	0.06	4.9	2.15	-4.44	16.8	10.0
Leu	1.0	-17.79	8.79	0.77	0.06	-0.17	4.9	2.28	-4.19	15.0	12.2
Lys	0.0	9.71	4.40	0.31	-0.16	-0.45	10.1	-9.52	2.84	7.9	7.5
Met	0.0	-8.86	9.15	0.76	0.11	0.03	5.3	-1.48	-2.49	13.3	8.4
Phe	1.0	-21.98	7.98	0.87	1.18	0.40	5.0	-0.76	-4.92	11.2	8.3
Pro	0.0	5.82	7.79	0.35	0.11	-0.47	6.6	-3.68	-1.22	8.2	6.9
Ser	0.0	-1.54	7.08	0.49	0.13	-0.11	7.5	-5.06	1.96	7.4	8.0
Thr	0.0	-4.15	7.00	0.38	0.28	0.09	6.6	-4.88	0.92	8.8	7.0
Trp	1.0	-16.19	8.07	0.86	-0.12	-0.61	5.3	-5.88	-4.75	9.9	5.7
Tyr	1.0	-1.51	6.90	0.64	0.19	-0.61	5.7	-6.11	-1.39	8.8	6.8
Val	1.0	-16.22	8.88	0.72	-0.08	-0.11	5.6	1.99	-2.69	12.0	9.4
Correlat. coeffic. R with EIIP	**-0.2429**	**0.3216**	**-0.1650**	**-0.0257**	**-0.0695**	**0.0613**	**0.1183**	**-0.3674**	**0.1623**	**-0.3550**	**-0.2838**

	Parameter					
Amino acid	**H395**	**H396**	**H398**	**H400**	**H401**	**H402**
Ala	18.08	18.56	0.83	0.00	6.00	9.9
Arg	0.00	0.00	0.83	52.00	10.76	4.6
Asn	17.47	18.24	0.09	3.38	5.41	5.4
Asp	17.36	17.94	0.64	49.70	2.77	2.8
Cys	18.17	17.84	1.48	1.48	5.05	2.8
Gln	17.93	18.51	0.00	3.53	5.65	9.0
Glu	18.16	17.97	0.65	49.90	3.22	3.2

Gly	18.24	18.57	0.10	0.00	5.97	5.6
His	18.49	18.64	1.10	51.60	7.59	8.2
Ile	18.62	19.21	3.07	0.13	6.02	17.1
Leu	18.60	19.01	2.52	0.13	5.98	17.6
Lys	17.96	18.36	1.60	49.50	9.74	3.5
Met	18.11	18.49	1.40	1.43	5.74	14.9
Phe	17.30	17.95	2.75	0.35	5.48	18.8
Pro	18.16	18.77	2.70	1.58	6.30	14.8
Ser	17.57	18.06	0.14	1.67	5.68	6.9
Thr	17.54	17.71	0.54	1.66	5.66	9.5
Trp	17.19	16.87	0.31	2.10	5.89	17.1
Tyr	17.99	18.23	2.97	1.61	5.66	15.0
Val	18.30	18.98	1.79	0.13	5.96	14.3
Correlat. coeffic. R with EIIP	**-0.3374**	**-0.3417**	**-0.2178**	**0.1430**	**-0.0525**	**-0.1899**

P. Physicochemical properties

P001 ANDN920101 a-CH chemical shifts (Andersen et al., 1992)

P009 BIGC670101 Residue volume (Bigelow, 1967)

P015 BULH740102 Apparent partial specific volume (Bull-Breese, 1974)

P017 BUNA790102 a-CH chemical shifts (Bundi-Wuthrich, 1979)

P024 CHAM830103 The number of atoms in the side chain labelled 1+1 (Charton-Charton, 1983)

P026 CHAM830105 The number of atoms in the side chain labelled 3+1 (Charton-Charton, 1983)

P027 CHAM830106 The number of bonds in the longest chain (Charton-Charton, 1983)

P030 CHAM820101 Polarizability parameter (Charton-Charton, 1982)

P032 CHOC750101 Average volume of buried residue (Chothia, 1975)

P033 CHOC760101 Residue accessible surface area in tripeptide (Chothia, 1976)

P059 COHE430101 Partial specific volume (Cohn-Edsall, 1943)

P063 DAWD720101 Size (Dawson, 1972)

P065 DAYM780201 Relative mutability (Dayhoff et al., 1978a)

P072 FASG760101 Molecular weight (Fasman, 1976)

P078 FAUJ880101 Graph shape index (Fauchere et al., 1988)

P080 FAUJ880103 Normalized van der Waals volume (Fauchere et al., 1988)

P081 FAUJ880104 STERIMOL length of the side chain (Fauchere et al., 1988)

P082 FAUJ880105 STERIMOL minimum width of the side chain (Fauchere et al., 1988)

P083 FAUJ880106 STERIMOL maximum width of the side chain (Fauchere et al., 1988)

P096 GARJ730101 Partition coefficient (Garel et al., 1973)

P109 GOLD730102 Residue volume (Goldsack-Chalifoux, 1973)

P112 GRAR740103 Volume (Grantham, 1974)

P117 HUTJ700102 Absolute entropy (Hutchens, 1970)

P118 HUTJ700103 Entropy of formation (Hutchens, 1970)

P135 JOND920102 Relative mutability (Jones et al., 1992)

P149 KRIW790103 Side chain volume (Krigbaum-Komoriya, 1979)

P154 LEVM760102 Distance between Ca and centroid of side chain (Levitt, 1976)

P157 LEVM760105 Radius of gyration of side chain (Levitt, 1976)

P158 LEVM760106 van der Waals parameter R0 (Levitt, 1976)

P159 LEVM760107 van der Waals parameter e (Levitt, 1976)

P177 MCMT640101 Refractivity (McMeekin et al., 1964), Cited by Jones (1975)

P214 OOBM770102 Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)

P216 OOBM770104 Average non-bonded energy per residue (Oobatake-Ooi, 1977)

P217 OOBM770105 Short and medium range non-bonded energy per residue (Oobatake-Ooi, 1977)

P219 OOBM850102 Optimized propensity to form reverse turn (Oobatake et al., 1985)

P280 QIAN880123 Weights for beta-sheet at the window position of 3 (Qian-Sejnowski, 1988)

P281 QIAN880124 Weights for beta-sheet at the window position of 4 (Qian-Sejnowski, 1988)

P282 QIAN880125 Weights for beta-sheet at the window position of 5 (Qian-Sejnowski, 1988)

P316 RADA880103 Transfer free energy from vap to chx (Radzicka-Wolfenden, 1988)

P319 RADA880106 Accessible surface area (Radzicka-Wolfenden, 1988)

P353 ROSG850101 Mean area buried on transfer (Rose et al., 1985)

P361 SNEP660103 Principal component III (Sneath, 1966)

P383 WEBA780101 RF value in high salt chromatography (Weber-Lacey, 1978)

P390 WOLS870102 Principal property value z2 (Wold et al., 1987)

P397 ZASB820101 Dependence of partition coefficient on ionic strength (Zaslavsky et al., 1982)

P399 ZIMJ680102 Bulkiness (Zimmerman et al., 1968)

	Parameter										
Amino acid	**P001**	**P009**	**P015**	**P017**	**P024**	**P026**	**P027**	**P030**	**P032**	**P033**	**P059**
Ala	4.35	52.6	0.691	4.349	0.0	0.0	0.0	0.046	91.5	115.0	0.75
Arg	4.38	109.1	0.728	4.396	1.0	1.0	5.0	0.291	202.0	225.0	0.70
Asn	4.75	75.7	0.596	4.755	1.0	0.0	2.0	0.134	135.2	160.0	0.61
Asp	4.76	68.4	0.558	4.765	1.0	0.0	2.0	0.105	124.5	150.0	0.60
Cys	4.65	68.3	0.624	4.686	1.0	0.0	1.0	0.128	117.7	135.0	0.61
Gln	4.37	89.7	0.649	4.373	1.0	1.0	3.0	0.180	161.1	180.0	0.67
Glu	4.29	84.7	0.632	4.295	1.0	1.0	3.0	0.151	155.1	190.0	0.66
Gly	3.97	36.3	0.592	3.972	0.0	0.0	0.0	0.000	66.4	75.0	0.64
His	4.63	91.9	0.646	4.630	1.0	1.0	3.0	0.230	167.3	195.0	0.67
Ile	3.95	102.0	0.809	4.224	2.0	0.0	2.0	0.186	168.8	175.0	0.90
Leu	4.17	102.0	0.842	4.385	1.0	0.0	2.0	0.186	167.9	170.0	0.90
Lys	4.36	105.1	0.767	4.358	1.0	1.0	4.0	0.219	171.3	200.0	0.82
Met	4.52	97.7	0.709	4.513	1.0	1.0	3.0	0.221	170.8	185.0	0.75
Phe	4.66	113.9	0.756	4.663	1.0	1.0	4.0	0.290	203.4	210.0	0.77
Pro	4.44	73.6	0.730	4.471	0.0	0.0	0.0	0.131	129.3	145.0	0.76
Ser	4.50	54.9	0.594	4.498	1.0	0.0	1.0	0.062	99.1	115.0	0.68
Thr	4.35	71.2	0.655	4.346	2.0	0.0	1.0	0.108	122.1	140.0	0.70
Trp	4.70	135.4	0.743	4.702	1.0	1.5	5.0	0.409	237.6	255.0	0.74
Tyr	4.60	116.2	0.743	4.604	1.0	1.0	5.0	0.298	203.6	230.0	0.71
Val	3.95	85.1	0.777	4.184	2.0	0.0	1.0	0.140	141.7	155.0	0.86
Correlat. coeffic. R with EIIP	**0.5542**	**0.0582**	**-0.3184**	**0.4692**	**0.0878**	**0.2007**	**0.2563**	**0.1624**	**0.1159**	**0.1330**	**-0.4075**

	Parameter										
Amino acid	**P063**	**P065**	**P072**	**P078**	**P080**	**P081**	**P082**	**P083**	**P096**	**P109**	**P112**
Ala	2.5	100.0	89.09	1.28	1.00	2.87	1.52	2.04	0.28	88.3	31.0
Arg	7.5	65.0	174.20	2.34	6.13	7.82	1.52	6.24	0.10	181.2	124.0
Asn	5.0	134.0	132.12	1.60	2.95	4.58	1.52	4.37	0.25	125.1	56.0
Asp	2.5	106.0	133.10	1.60	2.78	4.74	1.52	3.78	0.21	110.8	54.0
Cys	3.0	20.0	121.15	1.77	2.43	4.47	1.52	3.41	0.28	112.4	55.0
Gln	6.0	93.0	146.15	1.56	3.95	6.11	1.52	3.53	0.35	148.7	85.0
Glu	5.0	102.0	147.13	1.56	3.78	5.97	1.52	3.31	0.33	140.5	83.0
Gly	0.5	49.0	75.07	0.00	0.00	2.06	1.00	1.00	0.17	60.0	3.0
His	6.0	66.0	155.16	2.99	4.66	5.23	1.52	5.66	0.21	152.6	96.0
Ile	5.5	96.0	131.17	4.19	4.00	4.92	1.90	3.49	0.82	168.5	111.0
Leu	5.5	40.0	131.17	2.59	4.00	4.92	1.52	4.45	1.00	168.5	111.0
Lys	7.0	56.0	146.19	1.89	4.77	6.89	1.52	4.87	0.09	175.6	119.0
Met	6.0	94.0	149.21	2.35	4.43	6.36	1.52	4.80	0.07	162.2	105.0
Phe	6.5	41.0	165.19	2.94	5.89	4.62	1.52	6.02	2.18	189.0	132.0
Pro	5.5	56.0	115.13	2.67	2.72	4.11	1.52	4.31	0.39	122.2	32.5
Ser	3.0	120.0	105.09	1.31	1.60	3.97	1.52	2.70	0.12	88.7	32.0
Thr	5.0	97.0	119.12	3.03	2.60	4.11	1.73	3.17	0.21	118.2	61.0
Trp	7.0	18.0	204.24	3.21	8.08	7.68	1.52	5.90	5.70	227.0	170.0
Tyr	7.0	41.0	181.19	2.94	6.47	4.73	1.52	6.72	1.26	193.0	136.0
Val	5.0	74.0	117.15	3.67	3.00	4.11	1.90	3.17	0.60	141.4	84.0
Correlat. coeff'n. R with EIIP	**0.0347**	**0.0187**	**0.2561**	**-0.0876**	**0.1758**	**0.2408**	**-0.0202**	**0.2273**	**0.0229**	**0.0502**	**0.0903**

	Parameter										
Amino acid	**P117**	**P118**	**P135**	**P149**	**P154**	**P157**	**P158**	**P159**	**P177**	**P214**	**P216**
Ala	30.88	154.33	100.0	27.5	0.77	0.77	5.2	0.025	4.34	-1.404	-9.475
Arg	68.43	341.01	83.0	105.0	3.72	2.38	6.0	0.200	26.66	-0.921	-16.225
Asn	41.70	207.90	104.0	58.7	1.98	1.45	5.0	0.100	13.28	-1.178	-12.480
Asp	40.66	194.91	86.0	40.0	1.99	1.43	5.0	0.100	12.00	-1.162	-12.144
Cys	53.83	219.79	44.0	44.6	1.38	1.22	6.1	0.100	35.77	-1.365	-12.210
Gln	46.62	235.51	84.0	80.7	2.58	1.75	6.0	0.100	17.56	-1.116	-13.689
Glu	44.98	223.16	77.0	62.0	2.63	1.77	6.0	0.100	17.26	-1.163	-13.815
Gly	24.74	127.90	50.0	0.00	0.00	0.58	4.2	0.025	0.00	-1.364	-7.592
His	65.99	242.54	91.0	79.0	2.76	1.78	6.0	0.100	21.81	-1.215	-17.550
Ile	49.71	233.21	103.0	93.5	1.83	1.56	7.0	0.190	19.06	-1.189	-15.608
Leu	50.62	232.20	54.0	93.5	2.08	1.54	7.0	0.190	18.78	-1.315	-15.728
Lys	63.21	300.46	72.0	100.0	2.94	2.08	6.0	0.200	21.29	-1.074	-12.366
Met	55.32	202.65	93.0	94.1	2.34	1.80	6.8	0.190	21.64	-1.303	-15.704
Phe	51.06	204.74	51.0	115.5	2.97	1.90	7.1	0.390	29.40	-1.135	-20.504
Pro	39.21	179.93	58.0	41.9	1.42	1.25	6.2	0.170	10.93	-1.236	-11.893
Ser	35.65	174.06	117.0	29.3	1.28	1.08	4.9	0.025	6.35	-1.297	-10.518
Thr	36.50	205.80	107.0	51.3	1.43	1.24	5.0	0.100	11.01	-1.252	-12.369
Trp	60.00	237.01	25.0	145.5	3.58	2.21	7.6	0.560	42.53	-1.030	-26.166
Tyr	51.15	229.15	50.0	117.3	3.36	2.13	7.1	0.390	31.53	-1.030	-20.232
Val	42.75	207.60	98.0	71.5	1.49	1.29	6.4	0.150	13.92	-1.254	-13.867
Correlat. coeffic. R	**0.1533**	**0.1403**	**0.0327**	**0.0818**	**0.2476**	**0.2176**	**-0.0780**	**0.1133**	**0.2455**	**0.2323**	**-0.1193**

with EIIP											

	Parameter										
Amino acid	**P217**	**P219**	**P280**	**P281**	**P282**	**P316**	**P319**	**P353**	**P361**	**P383**	**P390**
Ala	-7.020	1.34	-0.44	-0.31	-0.02	0.13	93.7	86.6	-0.110	0.89	-1.73
Arg	-10.131	0.95	-0.13	-0.10	0.04	-5.0	250.4	162.2	0.079	0.88	2.52
Asn	-9.424	2.49	0.05	0.06	0.03	-3.04	146.3	103.3	-0.136	0.89	1.45
Asp	-9.296	3.32	-0.20	0.13	0.11	-2.23	142.6	97.8	-0.285	0.87	1.13
Cys	-8.190	1.07	0.13	-0.11	-0.02	-2.52	135.2	132.3	-0.184	0.85	-0.97
Gln	-10.044	1.49	-0.58	-0.47	-0.17	-3.84	177.7	119.2	-0.067	0.82	0.53
Glu	-10.467	2.20	-0.28	-0.05	0.10	-3.43	182.9	113.9	-0.246	0.84	0.39
Gly	-5.456	2.07	0.08	0.45	0.38	1.45	52.6	62.9	-0.073	0.92	-5.36
His	-12.150	1.27	0.09	-0.06	-0.09	-5.61	188.1	155.8	0.320	0.83	1.74
Ile	-9.512	0.66	-0.04	-0.25	-0.48	-2.77	182.2	158.0	0.001	0.76	-1.68
Leu	-10.520	0.54	-0.12	-0.44	-0.26	-2.64	173.7	164.1	-0.008	0.73	-1.03
Lys	-9.666	0.61	-0.33	-0.44	-0.39	-3.97	215.2	115.5	0.049	0.97	1.41
Met	-10.424	0.70	-0.21	-0.28	-0.14	-3.83	197.6	172.9	-0.041	0.74	-0.27
Phe	-12.485	0.80	-0.13	-0.04	-0.03	-3.74	228.6	194.1	0.438	0.52	1.30
Pro	-8.652	2.12	-0.48	-0.29	-0.04	0.00	0.0	92.9	-0.016	0.82	0.88
Ser	-7.782	0.94	0.27	0.34	0.41	-1.66	109.5	85.6	-0.153	0.96	-1.63
Thr	-8.764	1.09	0.47	0.27	0.36	-2.31	142.1	106.5	-0.208	0.92	-2.09
Trp	-14.420	-4.65	-0.22	-0.08	-0.01	-8.21	271.6	224.6	0.493	0.20	3.65
Tyr	-12.360	-0.17	-0.11	0.06	-0.08	-5.97	239.9	177.7	0.381	0.49	2.32
Val	-8.778	1.32	0.06	0.11	-0.18	-2.05	157.2	141.0	-0.155	0.85	-2.53

Correlat. coeffic. R with EIIP	**0.0194**	**-0.0339**	**0.0786**	**0.1443**	**0.3349**	**-0.2116**	**0.2379**	**0.1165**	**0.0103**	**-0.0638**	**0.2731**

	Parameter	
Amino acid	**P397**	**P399**
Ala	-0.152	11.50
Arg	-0.089	14.28
Asn	-0.203	12.82
Asp	-0.355	11.68
Cys	0.000	13.46
Gln	-0.181	14.45
Glu	-0.411	13.57
Gly	-0.190	3.40
His	0.000	13.69
Ile	-0.086	21.40
Leu	-0.102	21.40
Lys	-0.062	15.71
Met	-0.107	16.25
Phe	0.001	19.80
Pro	-0.181	17.43
Ser	-0.203	9.47
Thr	-0.170	15.77
Trp	0.275	21.67
Tyr	0.000	18.03
Val	-0.125	21.57
Correlat.		

coeffic. R with EIIP	**0.0436**	**-0.1270**

. **Other properties**

O021 CHAM810101 Steric parameter (Charton, 1981)

O025 CHAM830104 The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)

O043 CHOP780206 Normalized frequency of N-terminal non helical region (Chou-Fasman, 1978b)

O051 CHOP780214 Frequency of the 3rd residue in turn (Chou-Fasman, 1978b)

O079 FAUJ880102 Smoothed upsilon steric parameter (Fauchere et al., 1988)

O084 FAUJ880107 N.m.r. chemical shift of alpha-carbon (Fauchere et al., 1988)

O123 ISOY800105 Normalized relative frequency of bend S (Isogai et al., 1980)

O126 ISOY800108 Normalized relative frequency of coil (Isogai et al., 1980)

O155 LEVM760103 Side chain angle theta(AAR) (Levitt, 1976)

O156 LEVM760104 Side chain torsion angle phi(AAAR) (Levitt, 1976)

O173 MAXF760103 Normalized frequency of zR (Maxfield-Scheraga, 1976)

O174 MAXF760104 Normalized frequency of left-handed alpha-helix (Maxfield-Scheraga, 1976)

O175 MAXF760105 Normalized frequency of zL (Maxfield-Scheraga, 1976)

O254 PRAM820102 Slope in regression analysis '1.0E1 (Prabhakaran-Ponnuswamy, 1982)

O255 PRAM820103 Correlation coefficient in regression analysis (Prabhakaran-Ponnuswamy, 1982)

O298 RACS820102 Average relative fractional occurrence in AR(i) (Rackovsky-Scheraga, 1982)

O302 RACS820106 Average relative fractional occurrence in ER(i) (Rackovsky-Scheraga, 1982)

O303 RACS820107 Average relative fractional occurrence in A0(i-1) (Rackovsky-Scheraga, 1982)

O305 RACS820109 Average relative fractional occurrence in AL(i-1) (Rackovsky-Scheraga, 1982)

O309 RACS820113 Value of theta(i) (Rackovsky-Scheraga, 1982)

O322 RICJ880101 Relative preference value at N" (Richardson- Richardson, 1988)

O323 RICJ880102 Relative preference value at N' (Richardson-Richardson, 1988)

O324 RICJ880103 Relative preference value at N-cap (Richardson-Richardson, 1988)

O336 RICJ880115 Relative preference value at C-cap (Richardson-Richardson, 1988)

O370 TANS770105 Normalized frequency of chain reversal S (Tanaka-Scheraga, 1977)

O372 TANS770107 Normalized frequency of left-handed helix (Tanaka-Scheraga, 1977)

O374 TANS770109 Normalized frequency of coil (Tanaka-Scheraga, 1977)

O385 WERD780102 Free energy change of e(i) to e(ex) (Wertz-Scheraga, 1978)

	Parameter										
Amino acid	**O021**	**O025**	**O043**	**O051**	**O079**	**O084**	**O123**	**O126**	**O155**	**O156**	**O173**
Ala	0.52	0.0	0.70	0.035	0.53	7.3	0.35	0,47	121.9	243.2	0.64
Arg	0.65	1.0	0.34	0.099	0.69	11.1	0.75	0.52	121.4	206.6	0.62
Asn	0.76	1.0	1.42	0.191	0.58	8.0	2.12	2.16	117.5	207.1	3.14
Asp	0.76	1.0	0.98	0.179	0.59	9.2	2.16	1.15	121.2	215.0	1.92
Cys	0.62	0.0	0.65	0.117	0.66	14.4	0.50	0.41	113.7	209.4	0.32
Gln	0.68	1.0	0.75	0.037	0.71	10.6	0.73	0.95	118.0	205.4	0.80
Glu	0.68	1.0	1.04	0.077	0.72	11.4	0.65	0.64	118.2	213.6	1.01
Gly	0.00	0.0	1.41	0.190	0.00	0.0	2.40	3.03	0.0	300.0	0.63
His	0.70	1.0	1.22	0.093	0.64	10.2	1.19	0.89	118.2	219.9	2.05
Ile	1.02	1.0	0.78	0.013	0.96	16.1	0.12	0.62	118.9	217.9	0.92
Leu	0.98	2.0	0.85	0.036	0.92	10.1	0.58	0.533	118.1	205.6	0.37
Lys	0.68	1.0	1.01	0.072	0.78	10.9	0.83	0.98	122.0	210.9	0.89
Met	0.78	1.0	0.83	0.014	0.77	10.4	0.22	0.68	113.1	204.0	1.07
Phe	0.70	1.0	0.93	0.065	0.71	13.9	0.89	0.61	118.2	203.7	0.86
Pro	0.36	0.0	1.10	0.034	0.00	17.8	0.43	0.63	81.9	237.4	0.50
Ser	0.53	0.0	1.55	0.125	0.55	13.1	1.24	1.03	117.9	232.0	1.01
Thr	0.50	0.0	1.09	0.065	0.63	16.7	0.85	0.39	117.1	226.7	0.92
Trp	0.70	1.0	0.62	0.064	0.84	13.2	0.62	0.63	118.4	203.7	1.00
Tyr	0.70	1.0	0.99	0.114	0.71	13.9	1.44	0.83	110.0	195.6	1.31
Val	0.76	0.0	0.75	0.028	0.89	17.2	0.43	0.76	121.7	220.3	0.87
Correlat. coeffic. R with EIIP	**-0.0236**	**-0.0772**	**-0.2386**	**0.1445**	**0.0624**	**0.1360**	**0.0714**	**-0.2977**	**0.2866**	**-0.2892**	**-0.0567**

	Parameter										
Amino acid	**O174**	**O175**	**O254**	**O255**	**O298**	**O302**	**O303**	**O305**	**O309**	**O322**	**O323**
Ala	0.17	1.13	0.175	0.687	1.58	0.30	0.40	0.00	17.05	0.7	0.7
Arg	0.76	0.48	0.083	0.590	1.14	0.90	1.20	0.00	21.25	0.4	0.4
Asn	2.62	1.11	0.090	0.489	0.77	2.73	1.24	4.14	34.81	1.2	1.2
Asp	1.08	1.18	0.140	0.632	0.98	1.26	1.59	2.15	19.27	1.4	1.4
Cys	0.95	0.38	0.074	0.263	1.04	0.72	2.98	0.00	28.84	0.6	0.6
Gln	0.91	0.41	0.093	0.527	1.24	0.97	0.50	0.00	15.42	1.0	1.0
Glu	0.28	1.02	0.135	0.669	1.49	1.33	1.26	0.00	20.12	1.0	1.0
Gly	5.02	3.84	0.201	0.670	0.66	3.09	1.89	6.49	38.14	1.6	1.6
His	0.57	0.30	0.125	0.594	0.99	1.33	2.71	0.00	23.07	1.2	1.2
Ile	0.26	0.40	0.100	0.564	1.09	0.45	1.31	0.00	16.66	0.9	0.9
Leu	0.21	0.65	0.104	0.541	1.21	0.96	0.57	0.00	10.89	0.9	0.9
Lys	1.17	1.13	0.058	0.407	1.27	0.71	0.87	0.00	16.46	1.0	1.0
Met	0.00	0.00	0.054	0.328	1.41	1.89	0.00	0.00	20.61	0.3	0.3
Phe	0.28	0.45	0.104	0.577	1.00	1.20	1.27	2.11	16.26	1.2	1.2
Pro	0.12	0.00	0.136	0.600	1.46	0.83	0.38	1.99	23.94	0.7	0.7
Ser	0.57	0.81	0.155	0.692	1.05	1.16	0.92	0.00	19.95	1.6	1.6
Thr	0.23	0.71	0.152	0.713	0.87	0.97	1.38	1.24	18.92	0.3	0.3
Trp	0.00	0.93	0.092	0.632	1.23	1.58	1.53	0.00	23.36	1.1	1.1
Tyr	0.97	0.38	0.081	0.495	0.68	0.86	1.79	1.90	26.49	1.9	1.9
Val	0.24	0.48	0.096	0.529	0.88	0.64	0.95	0.00	17.06	0.7	0.7
Correlat. coeffic. R with EIIP	**-0.2237**	**-0.2367**	**-0.1207**	**-0.0573**	**-0.0278**	**-0.1518**	**0.0673**	**-0.1568**	**-0.1665**	**-0.1161**	**-0.1161**

	Parameter					
Amino acid	**O324**	**O336**	**O370**	**O372**	**O374**	**O385**
Ala	0.5	0.8	0.497	0.289	0.945	0.16
Arg	0.4	0.9	0.677	1.380	0.364	-0.20
Asn	3.5	1.6	2.072	3.169	1.202	1.03
Asp	2.1	0.7	1.498	0.917	1.315	-0.24
Cys	0.6	0.4	1.348	1.767	0.932	-0.12
Gln	0.4	0.9	0.711	2.372	0.704	-0.55
Glu	0.4	0.3	0.651	0.285	1.014	-0.45
Gly	1.8	3.9	1.848	4.259	2.355	-0.16
His	1.1	1.3	1.474	1.061	0.525	-0.18
Ile	0.2	0.7	0.471	0.262	0.673	-0.19
Leu	0.2	0.7	0.656	0.000	0.758	-0.44
Lys	0.7	1.3	0.932	1.288	0.947	-0.12
Met	0.8	0.8	0.425	0.000	1.028	-0.79
Phe	0.2	0.5	1.348	0.393	0.622	-0.25
Pro	0.8	0.7	0.179	0.000	0.579	-0.59
Ser	2.3	0.8	1.151	0.160	1.140	-0.01
Thr	1.6	0.3	0.749	0.218	0.863	0.05
Trp	0.3	0.0	1.283	0.000	0.777	-0.33
Tyr	0.8	0.8	1.283	0.654	0.907	-0.42
Val	0.1	0.2	0.654	0.167	0.561	-0.46
Correlat. coeffic. R with EIIP	**0.2614**	**-0.3000**	**0.0669**	**-0.1275**	**-0.1009**	**-0.1193**

APPENDIX D

LIST OF AMINO ACID SEQUENCES USED IN THE STUDY

LYSOZYME (17 sequences)

1. Lzba
Lysozyme (EC 3.2.1.17) - Baboon
KIFERCELARTLKRLGLDGYRGISLANWVCLAKWESDYNTQATNYNPGDQSTDYGIFQIN
SHYWCNDGKTPGAVNACHISCNALLQDNITDAVACAKRVVSDPQGIRAWVAWRNHCQNRD
VSQYVQGCGV
2. Lzbo
Lysozyme c (EC 3.2.1.17) 2 - Bovine
KVFERCELARTLKKLGLDGYKGVSLANWLCLTKWESSYNTKATNYNPSSESTDYGIFQIN
SKWWCNDGKTPNAVDGCHVSCSELMENDIAKAVACAKKIVSEQGITAWVAWKSHCRDHDV
SSYVEGCTL
3. Lzbpt4
Lysozyme (EC 3.2.1.17) - Bacteriophages T4
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITK
DEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVRRCALINMVFQMGETGVAGFTNSLRM
LQQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDAYKNL
4. Lzch
Lysozyme c (EC 3.2.1.17) precursor - Chicken
MRSLLILVLCFLPLAALGKVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQA
TNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDG
NGMNAWVAWRNRCKGTDVQAWIRGCRL
5. Lzdk
Lysozyme c (EC 3.2.1.17) precursor - Duck
MKALLTLVFCLLPLAAQGKVYSRCELAAAMKRLGLDNYRGYSLGNWVCAANYESGFNTQA
TNRNTDGSTDYGILQINSRWWCDNGKTPRSKNACGIPCSVLLRSDITEAVRCAKRIVSDG
DGMNAWVAWRNRCRGTDVSKWIRGCRL
6. Lzdk3
Lysozyme c (EC 3.2.1.17) III - Duck
KVYERCELAAAMKRLGLDNYRGYSLGNWVCAANYESSFNTQATNRNTDGSTDYGILEINS
RWWCDNGKTPRAKNACGIPCSVLLRSDITEAVKCAKRIVSDGDGMNAWVAWRNRCKGTDV
SRWIRGCRL
7. Lzfer
Lysozyme c (EC 3.2.1.17) precursor - Ring-necked pheasant
MRSLLILVLCFLPLAAPGKVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFESNFNTGA
TNRNTDGSTDYGILQINSRWWCNDGRTPGSKNLCHIPCSALLSSDITASVNCAKKIVSDG
DGMNAWVAWRKHCKGTDVNVWIRGCRL
8. Lzgsg
Lysozyme g (EC 3.2.1.17) - Goose
RTDCYGNVNRIDTTGASCKTAKPEGLSYCGVSASKKIAERDLQAMDRYKTIIKKVGEKLC
VEPAVIAGIISRESHAGKVLKNGWGDRGNGFGLMQVDKRSHKPQGTWNGEVHITQGTTIL
INFIKTIQKKFPSWTKDQQLKGGISAYNAGAGNVRSYARMDIGTTHDDYANDVVARAQYY
KQHGY
9. Lzhu
Lysozyme (EC 3.2.1.17) - Human
KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNAGDRSTDYGIFQIN
SRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWRNRCQNRD
VRQYVQGCGV
10. Lzosg
Lysozyme g (EC 3.2.1.17) - Ostrich
RTGCYGDVNRVDTTGASCKSAKPEKLNYCGVAASRKIAERDLQSMDRYKALIKKVGQKLC
VDPAVIAGIISRESHAGKALRNGWGDNGNGFGLMQVDRRSHKPVGEWNGERHLMQGTEIL
ISMIKAIQKKFPRWTKEQQLKGGISAYNAGPGNVRSYERMDIGTTHDDYANDVVARAQYY
KQHGY

11. Lzove
Lysozyme c (EC 3.2.1.17) - Chachalaca
KIYKRCELAAAMKRYGLDNYRGYSLGNWVCAARYESNYNTQATNRNSNGSTDYGILQINS
RWWCNDGRTPGTKNLCHISCSALMGADIAPSVRCAKRIVSDGDGMNAWVAWRKHCKGTDV
STWIKDCKL
12. Lzpy
Lysozyme c (EC 3.2.1.17) - Pigeon
KDIPRCELVKILRRHGFEGFVGKTVANWVCLVKHESGYRTTAFNNNGPNSRDYGIFQINS
KYWCNDGKTRGSKNACNINCSKLRDDNIADDIQCAKKIAREARGLTPWVAWKKYCQGKDL
SSYVRGC
13. Lzqjeb
Lysozyme c (EC 3.2.1.17) - Bobwhite (tentative sequence)
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNSQATNRNTDGSTDYGVLQINS
RWWCNDGKTPGSRNLCNIPCSALLSSDITATVNCAKKIVSDGBGMNAWVAWRNRCKGTDV
QAWIRGCRL
14. Lzrt
Lysozyme (EC 3.2.1.17) - Rat
KTYERCEFARTLKRNGMSGYYGVSLADWVCLAQHESNYNTQARNYDPGDQSTDYGIFQIN
SRYWCNDGKTPRAKNACGIPCSALLQDDITQAIQCAKRVVRDPQGIRAWVAWQRHCKNRD
LSGYIRNCGV
15. Lzuh
Lysozyme c (EC 3.2.1.17) - Guineafowl (tentative sequence)
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNSQATNRNTDGSTDYGVLQINS
RWWCNDGRTPGSRNLCNIPCSALQSSDITATANCAKKIVSDGBGMNAWVAWRKHCKGTDV
RVWIKGCRL
16. Lzwk
Lysozyme c (EC 3.2.1.17) precursor - Cecropia moth (fragment)
CRSWQFALHCDAKRFTRCGLVQELRRRGFDETLMSNWVCLVENESGRFTDKIGKVNKNGS
RDYGLFQINDKYWCSKGSTPGKDCNVTCNQLLTDDISVAATCAKKIYKRIKFDAWYGWKN
HCQHGLPDISDC
17. Lzwsg
Lysozyme g (EC 3.2.1.17) - Black swan
RTDCYGNVNRIDTTGASCKTAKPEGLSYCGVPASKTIAERDLKAMDRYKTIIKKVGEKLC
VEPAVIAGIISRESHAGKVLKNGWGDRGNGFGLMQVDKRSHKPQGTWNGEVHITQGTTIL
TDFIKRIQKKFPSWTKDQQLKGGISAYNAGAGNVRSYARMDIGTTHDDYANDVVARAQYY
KQHGY

GLUCAGON (30 sequences)

1. Glu1l
GLUCAGON I PRECURSOR
MKRIHSLAGILLVLGLIQSSCRVLMQEADPSSSLEADSTLKDEPRELSNMKRHSEGTFSN
DYSKYLEDRKAQEFVRWLMNNKRSGVAEKRHADGTFTSDVSSYLKDQAIKDFVDRLKAGQ
VRRE
2. Glu2l
GLUCAGON II PRECURSOR
MTSLHSLAGLLLLMIIQSSWQMPDQDPDRNSMLLNENSMLTEPIEPLNMKRHSEGTFSND
YSKYLETRRAQDFVQWLKNSKRNGLFRRHADGTYTSDVSSYLQDQAAKDFVSWLKAGRGR
RE
3. Gluangl
H0708424A: glucagon
HSEGTFSNDYSKYLEDRKAQEFVRWLMNN
4. Glubov
H710325A: glucagon
HSQGTFTSDYSKYLDSRRAQDFVQWLMNT
5. Gluca
GLUCAGON PRECURSOR (FRAGMENT)
HSQGTFTNDYSKYMDTRRAQDFVQWLMSTXXXXXXXXXXXXXXXXYADAPYISDVYSYLQD
QVAKKWLKSGQDRRE

6. Glucatf
Z1105264B : glucagon like peptide
HADGTYTSDVSSYLQDQAAKDFITWLKSGQPKPE
7. Glucb
GLUCAGON PRECURSOR
MKSLYFVAGLFVMLVQGSWQRSLQNTEEKSSSFPAPQTDPLGDPDQINEDKRHSQGTFTS
DYSKYLDSRRAQDFVQWLMNTKRNKNNIAKRHDEFERHAEGTFTSDVSSYLEGQAAKEFI
AWLVKGRGRRDFPEEVNIVEELRRRHADGSFSDEMNTVLDSLATRDFINWLLQTKITDRK
8. Glucca
GLUCAGON PRECURSOR
MKSVYFVAGLFIMLAQGSWQRSLQDTEEKPRSVSASQTDMLDDPDQMNEDKRHSQGTFTS
DYSKYLDSRRAQQFLKWLLNVKRNRNNIAKRHDEFERHAEGTFTSDVSSYLEGQAAKEFI
AWLVKGRGRRDFPEEVAIVEELGRRHADGSFSDEMNTILDNLATRDFINWLIQTKITDRK
9. Glucch
GLUCAGON
HSQGTFTSDYSKHLDSRYAQEFVQWLMNT
10. Glucce
GLUCAGON PRECURSOR
MKMKSIYFIAGLLLMIVQGSWQNPLQDTEEKSRSFKASQSEPLDESRQLNEVKRHSQGTF
TSDYSKYLDSRRAQDFVQWLMSTKRNGQQGQEDKENDKFPDQLSSNAISKRHSEFERHAE
GTYTSDITSYLEGQAAKEFIAWLVNGRGRRE
11. Glucd
GLUCAGON
HSQGTFTSDYSKYLDSRRAQDFVQWLMST
12. Gluch
GLUCAGON
HTDGIFSSDYSKYLDNRRTKDFVQWLLSTKRNGANT
13. Gluchik
H753726A: glucagon
HSQGTFTSDYSKYLDSRRAQDFVQWLMST
14. Glucl
GLUCAGON PRECURSOR (FRAGMENT)
HSQGTFTNDYSKYLDTRRAQDFVQWLMSTKRSGGITXXXXXXXXHADGTYTSDVSSYLQD
QAAKKFVTWLKQGQDRRE
15. Glucm
GLUCAGON PRECURSOR
MKNIYIVAGFFCGAGQGSWQHSLQDTEEKSRSFPASQTDPLEDPDQINEDKRHSQGTFTS
DYSKYLDSRRAQDFVQWLMNTKRNRNNIAKRHDEFERHAEGTFTSDVSSYLEGQAAKEFI
AWLVKGRGRRDFPEEVTIVEELGRRHADGSFSDEMNTILDSLATRDFINWLIQTKITDKK
16. Gluco
GLUCAGON PRECURSOR
MKSIYFVAGLFVMLVQGSWQHPLQDTEEKPRSFSTSQTDLLDDPDQMNEDKRHSQGTFTS
DYSKFLDTRRAQDFLDWLKNTKRNRNEIAKRHDEFERHAEGTFTSDVSSYLEGQAAKEFI
AWLVKGRGRRDFPEEVTIVEELRRRHADGSFSDEMNTVLDHLATKDFINWLIQTKITDRK
17. Glucoho
Z1205373A: glucagon
HSEGTFSNDYSKYQEERMAQDFVQWLMNS
18. Glucp
GLUCAGON
HSEGTFSNDYSKYLETRRAQDFVQWLKNS
19. Glucr
GLUCAGON PRECURSOR
MKTVYIVAGLFVMLVQGSWQHAPQDTEENARSFPASQTEPLEDPDQINEDKRHSQGTFTS
DYSKYLDSRRAQDFVQWLMNTKRNRNNIAKRHDEFERHAEGTFTSDVSSYLEGQAAKEFI
AWLVKGRGRRDFPEEVAIAEELGRRHADGSFSDEMNTILDNLATRDFINWLIQTKITDKK
20. Glucrb
GLUCAGON
HSQGTFTSDYSKYLDSRRAQDFVQWLMNT
21. Glucs
GLUCAGON
HSEGTFTSDYSKYMDNRRAKDFVQWLMNT

22. Gluct
GLUCAGON
HSEGTFTSDYSKYLDNRRAKDFVQWLMNT
23. Gluduc
H721287A: glucagon
HSQGTFTSDYSKYLDTRRAQDFVQWLMST
24. Gluguipg
Z1112169A: glucagon
HSEGTFTSDYSKYLDNRRAKDFVQWLMNT
25. Gluhum
QQQQQMCSCCSMQQWWYYAAAAYWWQQMC
26. Glupig
H570104A: glucagon
HSQGTFTSDYSKYLDSRRAQDFVQWLMNT
27. Glurab
H720967A: glucagon
HSQGTFTSDYSKYLDSRRAQDFVQWLMNT
28. Glurat
H711649A: glucagon
HSQGTFTSDYSKYLDSRRAQDFVQWLMNT
29. Glutfish
Z1112169A: glucagon
HSEGTFTSDYSKYLDNRRAKDFVQWLMNT
30. Gluturk
H721913A: glucagon
HSQGTFTSDYSKYLDSRRAQDFVQWLMST

HEMOGLOBIN (91 sequences)

1. Habag
VLSPADKTNVKTAWGKVGGHGGEYGAEALERMFLSFPTTKTYFPHFDLSH
GSAQVKGHGKKVADALTLAAAHVDDMPSALSALSDLHAHKLRVDPVNFKL
LSHCLLVTLAAHHPAEFTPAVHASLDKFLASVSTVLTSKYR
2. Habaim
hemoglobin alpha-I chain - mandrill
VLSPADKKNVKAAWDKVGGHAGEYGAEALERMFLSFPTTKTYFPHFNLSHGSDQVKGHGK
KVADALTLAVGHVDDMPQALSKLSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA
VHASLDKFLASVSTVLTSKYR
3. Habay
hemoglobin alpha chain - yellow baboon
VLSPDDKKHVKAAWGKVGEHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSDQVNKHGK
KVADALTLAVGHVDDMPQALSKLSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA
VHASLDKFLASVSTVLTSKYR
4. Habd
hemoglobin alpha chain - Eurasian badger
VLSPADKANIKATWDKIGGHAGEYGGEALERTFASFPTTKTYFPHFDLSHGSAQVKGHGK
KVADALTNAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFLSSVSTVLTSKYR
5. Habdbm
hemoglobin alpha chain - beech marten
VLSPADKTNVKSTWDKIGGHAGEYGGEALERTFVSFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTLAVGHLDDLAGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFFSTVSTVLTSKYR
6. Habdr
hemoglobin alpha chain - ratel
VLSPSDKANVKATWDKIGGHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTNAVAHGDDLPMALSTLSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFFSTVSTVLTSKYR

7. Habrm
hemoglobin alpha chain - polar bear
VLSPADKSNVKATWDKIGSHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTTAAGHLDDLPGALSALSDLHAHKLRVDPVNFKFLSHCLLVTLASHHPAEFTPA
VHASLDKFFSAVSTVLTSKYR
8. Habrt
hemoglobin alpha chain - Asiatic black bear
VLSPADKSNVKATWDKIGSHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTTAAGHLDDLPGALSALSDLHAHKLRVDPVNFKFLSHCLLVTLASHHPAEFTPA
VHASLDKFFSAVSTVLTSKYR
9. Hacjb
hemoglobin alpha chain - black-tailed marmoset
VLSPADKSNVKAAWGKVGSHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
KVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPAEFTPA
VHASLDKFLASVSTVLTSKYR
10. Hadg
hemoglobin alpha-I chain - dog
VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTTAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPTEFTPA
VHASLDKFFAAVSTVLTSKYR
11. Hadg2
hemoglobin alpha-II chain - dog
VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTTAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPTEFTPA
VHASLDKFFTAVSTVLTSKYR
12. Hadgy
hemoglobin alpha chain - coyote (tentative sequence)
VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTTAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPTEFTPA
VHASLDKFFTAVSTVLTSKYR
13. Hafqg
hemoglobin alpha chain - giant panda
VLSPADKTNVKATWDKIGGHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTTAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPAEFTPA
VHASLDKFFSAVSTVLTSKYR
14. Hafql
hemoglobin alpha chain - lesser panda
VLSPADKTNVKSTWDKLGGHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTLAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFFSAVSTVLTSKYR
15. Hags
hemoglobin alpha-I and alpha-II chains - thick-tailed bush baby
VLSPTDKSNVKAAWEKVGGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALTNAVSHVDDMPSALSALSDLHAHKLRVDPVNFKLLSHCLLVTLACHHPAELTPA
VHASFDKFMASVSTVLTSKYR
16. Hago3
hemoglobin alpha-3 chain - lowland gorilla
VLSPADKTNVKAAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAZVKGHGK
KVAKALTBAVZHLDDMPNALSALSBLHAHKLRVBPVBFKLLNHCLLVTLAABFPSZFTPA
VHASVDKFLASVSTVLTSKYR
17. Hago3c
hemoglobin alpha-3 chain - chimpanzee
VLSPADKTNVKAAWGKVGAHAGZYGAEALERMFLSFPTTKTYFPHFDLSHGSAZVKGHGK
KVAKALSBAVZHLDDMPNALSALSBLHAHKLRVBPVBFKLLNHCLLVTLAABFPSZFTPA
VHASVDKFLASVSTVLTSKYR
18. Hahh
hemoglobin alpha chain - western European hedgehog
VLSATDKANVKTFWGKLGGHGGEYGGEALDRMFQAHPTTKTYFPHFDLNPGSAQVKGHGK
KVADALTTAVNNLDDVPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLALHHPADFTPA
VHASLDKFLATVATVLTSKYR

19. Haho
hemoglobin alpha chains - horse
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHGK
KVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPA
VHASLDKFLSSVSTVLTSKYR
20. Hahy
hemoglobin alpha chain - golden hamster
VLSAKDKTNISEAWGKIGGHAGEYGAEALERMFFVYPTTKTYFPHFDVSHGSAQVKGHGK
KVADALTNAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLANHHPADFTPA
VHASLDKFFASVSTVLTSKYR
21. Hahzs
hemoglobin alpha chain - spotted hyena
VLSSADKANIKATWDKIGGHGGEYGAEALERTFLCFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALALAAAHLDDLPSALSALSDLHAYKLRVDPVNFKLLSHCLLVTLAAHHPAEFTPA
VHSDLDKFLSSVSTVLTSKYR
22. Hajl
hemoglobin alpha chain - lion
VLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQAHGQ
KVADALTKAVVHINDLPNALSDLSDLHAYKLRVDPVNFKFLSHCLLVTLACHHPEEFTPA
VHASLDKFFSAVSTVLTSKYR
23. Hajua
hemoglobin alpha chain - jaguar
VLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQAHGQ
KVADALTKAVAHINDLPNALSDLSDLHAYKLRVDPVNFKFLSHCLLVTLACHHPEEFTPA
VHASLDKFFSAVSTVLTSKYR
24. Hale2r
hemoglobin alpha-II chain - ruffed lemur
VLSPADKNNVKSAWKAIGSHAGEHGAEALERMFLSFPPTKTYFPHFDLSHGSAQIKTHGK
KVADALTNAVNHIDDMPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPAEFTPA
VHASLDKFFAAVSTVLTSKYR
25. Haleir
hemoglobin alpha-I chain - ruffed lemur
VLSPADKNNVKSAWNAIGSHAGEHGAEALERMFLSFPPTKTYFPHFDLSHGSAQIKTHGK
KVADALTNAVNHIDDMPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPAEFTPA
VHASLDKFFAAVSTVLTSKYR
26. Halrn
hemoglobin alpha chain - slender loris
VLSPADKTNVKTAWEKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALTTAVSHVDDMPSALSALSDLHAHKLRVDPVNFKLLSHCLLVTLACHHPADFTPA
VHASLDKFLASVSTVLTSKYR
27. Halrs
hemoglobin alpha chain - slow loris
VLSPADKTNVKAAWEKVGSHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALTNAVSHVDDMPSALSALSDLHAHKLRVDPVNFKLLSHCLLVTLACHHPADFTPA
VHASLDKFLASVSTVLTSKYR
28. Hamkp
hemoglobin alpha chain - black-handed spider monkey
VLSPADKSNVKAAWGKVGGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
KVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPADFTPA
VHASLDKFLASVSTVLTSKYR
29. Hamks
hemoglobin alpha chain - sooty mangabey
VLSPDDKKHVKAAWGKVGEHAGEYGAEALERMFLSFPTTKTYFPHFNLSHGSDQVKGHGK
KVADALTLAVGHVDDMPHALSKLSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA
VHASLDKFLASVSTVLTSKYR
30. Hamnle
hemoglobin alpha-I chain - European mink
VLSPADKTNVKSTWDKIGGHAGEYGGEALERTFASFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALTNAVAHMDDLPGAMSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFFSAVSTVLTSKYR

31. Hamn2f
hemoglobin alpha-II chain - domestic ferret
VLSPADKTNVKSTWGKIGGHAGEYGGEALERTFASFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALTNAVAHVDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFFSAVSTVLTSKYR
32. Hamnif
hemoglobin alpha-I chain - domestic ferret
VLSPADKTNVKSTWDKIGGHAGEYGGEALERTFASFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALTNAVAHVDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFFSAVSTVLTSKYR
33. Hamqa
hemoglobin alpha chain - brown capuchin
VLSPADKTNVKTAWGKVGGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
KVADALSNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPADFTPA
VHASLDKFLASVSTVLTSKYR
34. Hamqb
hemoglobin alpha chain - red colobus (tentative sequence)
VLSPADKTNVKTAWGKVGGHGGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
KVADALTLAAAHVDDMPSALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPAEFTPA
VHASLDKFLASVSTVLTSKYR
35. Hams
hemoglobin alpha chains - mouse
VLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHGK
KVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTPA
VHASLDKFLASVSTVLTSKYR
36. Haoee
hemoglobin alpha chain - European mole
VLSGTDKSNIKAAWDKVGAHAGEYGAEALERTFTSFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALTNAVGHLDDLPGAMSALSDLHAHKLRVDPVNFKLLSHCLLVTLACHHPNDFTPA
VHASLDKFLATVSTVLTSKYR
37. Haol
hemoglobin alpha chain - Ehrenberg's mole-rat
VLSPEDKNHVRSTWDKIGGHGAEYGAEALERMFTSFPTTKTYFPHFDVSHGSAQVKAHGK
KVADALANAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLANHHPAEFTPG
VHASLDKFLASVSTVLTSKYR
38. Haote
hemoglobin alpha chain - Eurasian river otter
VLSPADKTNVKSTWDKIGGHAGEYGGEALERTFVSFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALTNAVAHMDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFFSAVSTVLTSKYR
39. Haotg
hemoglobin alpha chain - giant otter
VLSPADKTNVKATWDKIGGHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTNAVAHMDDLPAALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFFSTVSTVLTSKYR
40. Haoz
hemoglobin alpha chain - muskrat
VLSGEDKNNIKTAWGKIGGHAAEYGAEALERMFVVYPTTKTYFPHFDVSHGSGQVKAHGK
KVADALTTAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLANHIPADFTPA
VHASLDKFLASVSTVLTSKYR
41. Hapd
hemoglobin alpha chain - Amur leopard
VLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQAHGQ
KVADALTKAVAHINDLPNALSDLSDLHAYKLRVDPVNFKFLSHCLLVTLACHHPEEFTPA
VHASLDKFFSAVSTVLTSKYR
42. Hapdp
hemoglobin alpha chain - northern Persian leopard
VLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQTHGQ
KVADALTKAVAHINDLPNALSDLSDLHAYKLRVDPVNFKFLSHCLLVTLACHHPEEFTPA
VHASLDKFFSAVSTVLTSKYR

43. Harb
hemoglobin alpha chain - rabbit
VLSPADKTNIKTAWEKIGSHGGEYGAEAVERMFLGFPTTKTYFPHFDFTHGSEQIKAHGK
KVSEALTKAVGHLDDLPGALSTLSDLHAHKLRVDPVNFKLLSHCLLVTLANHHPSEFTPA
VHASLDKFLANVSTVLTSKYR
44. Harr
hemoglobin alpha chain - raccoon
VLSPADKANIKATWDKIGGHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTLAVGHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFFTSVSTVLTSKYR
45. Hartl
hemoglobin alpha-1 chain - rat
VLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFNHIDVSPGSAQVKAHGK
KVADALAKAGAHLBBVPZALSTLSDLHAHKLRVDPVNFKFLSHCLLVTLACHHPGDFTPA
MHASLDKFLASVSTVLTSKYR
46. Hartng
hemoglobin alpha chain - northern gundi
VLSAADKTNVKAAWDKIGGHGGEYGAEALERMFLSFPTTKTYFPHFDVSHGSAQVKAHGK
KVADALANAASHLDDLPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFLATVATVLTSKYR
47. Hatb
hemoglobin alpha chain - western tarsier
VLSPADKTNVKAAWDKVGGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSSQVKGHGK
KVADALTTAVGHIDNMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLACHHPADFTPA
VHASLDKFVASVSTVLTSKYR
48. Hats
hemoglobin alpha chain - common tree shrew
VLSPGDKSNIKAAWGKIGGQAPQYGAEALERMFLSFPTTKTYFPHFDMSHGSAQIQAHGK
KVADALSTAVGHLDDLPTALSALSDLHAHKLRVDPANFKLLSHCILVTLACHHPGDFTPE
IHASLDKFLANVSTVLTSKYR
49. Hatsm
hemoglobin alpha chain - house shrew
VLSANDKANVKAAWDKVGGQAANYGAEALERTFASFPTTKTYFPHYDLSPGSAQVKAHGK
KVADALTKAVGSMDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPADFTPA
VHASLDKFLASVSTVLTSKYR
50. Hatx
hemoglobin alpha chain - tiger
VLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQTHGQ
KVADALTKAVAHINNLPNALSDLSDLHAYKLRVDPVNFKFLSHCLLVTLACHHPEEFTPA
VHASLDKFFSAVSTVLTSKYR
51. Hauk1e
hemoglobin alpha-I chain - European polecat
VLSPADKTNVKSTWDKIGGHAGEYGGEALERTFASFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALTNAVAHMDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFFSAVSTVLTSKYR
52. Hauk2e
hemoglobin alpha-II chain - European polecat
VLSPADKTNVKSTWGKIGGHAGEYGGEALERTFASFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALTNAVAHMDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEFTPA
VHASLDKFFSAVSTVLTSKYR
53. Hbbag
hemoglobin beta chain - gelada baboon
VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPAAVMGNPKV
KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
54. Hbbam
hemoglobin beta chain - mandrill
VHLTPEEKTAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKV
KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH

55. Hbbay
hemoglobin beta chain - yellow baboon
VHLTPEEKNAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPAAVMGNPKV
KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
56. Hbcjb
hemoglobin beta chain - black-tailed marmoset
VHLTGEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMNNPKV
KAHGKKVLGAFSDGLTHLDNLKGTFAHLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPVVQAAYQKVVAGVANALAHKYH
57. Hbcz
hemoglobin beta chain - chimpanzee
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
58. Hbczp
hemoglobin beta chain - pygmy chimpanzee
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
59. Hbema
hemoglobin beta chain - Amazon manatee
VHLTPEEKALVIGLWAKVNVKEYGGEALGRLLVVYPWTQRFFEHFGDLSSASAIMNNPKV
KAHGEKVFTSFGDGLKHLEDLKGAFAELSELHCDKLHVDPENFRLLGNVLVCVLARHFGK
EFSPEAQAAYQKVVAGVANALAHKYH
60. Hbgi
hemoglobin beta chain - common gibbon (tentative sequence)
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
61. Hbgo
hemoglobin beta chain - gorilla
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
62. Hbho
hemoglobin beta chain - horse
VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPKV
KAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDPENFRLLGNVLVVVLARHFGK
DFTPELQASYQKVVAGVANALAHKYH
63. Hbhu
hemoglobin beta chain - human
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
64. Hbhua
hemoglobin beta chain - human
VHLTPVEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
65. Hbmkb
hemoglobin beta chain - red-crowned mangabey
VHLTPEEKVAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSNPDAVMGNPKV
KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
66. Hbmkg
hemoglobin beta chain - green monkey
VHLTPEEKTAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH

67. Hbmkh
hemoglobin beta chain - long-haired spider monkey
VHLTGEEKAAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMSNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQLQAAYQKVVAGVANALAHKYH
68. Hbmkk
hemoglobin beta chain - black spider monkey
VHLTGEEKAAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMSNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQLQAAYQKVVAGVANALAHKYH
69. Hbmkm
hemoglobin beta chain - moustached tamarin (tentative sequence)
VHLTGEEKSAVTTLWGKVNVEEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMNNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
70. Hbmkn
hemoglobin beta chain - douroucouli (tentative sequence)
VHLTGEEKAAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMNNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
71. Hbmkp
hemoglobin beta chain - black-handed spider monkey
VHLTGEEKAAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMSNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQLQAAYQKVVAGVANALAHKYH
72. Hbmks
hemoglobin beta chain - common squirrel monkey (tentative sequence)
VHLTGDEKAAVTALWGKVNVEDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMNNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
73. Hbmqa
hemoglobin beta chain - brown capuchin
VHLTAEEKSAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKV
KAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVATALAHKYH
74. Hbmqaa
hemoglobin beta chain - white-fronted capuchin
VHLTAEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKV
KAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVATALAHKYH
75. Hbmqb
hemoglobin beta chain - red colobus
VHLTPDEKNAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSTADAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
76. Hbmqc
hemoglobin beta chain - crab-eating macaque (tentative sequence)
VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKV
KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
77. Hbmqf
hemoglobin beta chain - brown-headed tamarin
VHLTGEEKSAVTTLWGKVNVEEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCNKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
78. Hbmqj
hemoglobin beta chain - Japanese macaque
VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKV
KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH

79. Hbmqn
hemoglobin beta chain - black-and-red tamarin (tentative sequence)
VHLTGEEKSAVTTLWGKVNVEEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMNNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
80. Hbmqp
hemoglobin beta chain - hanuman langur
VHLTPEEKAAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
81. Hbmqpm
hemoglobin beta chain - pig-tailed macaque
VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKV
KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
82. Hbmqr
hemoglobin beta chain - rhesus macaque
VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLLVYPWTQRFFESFGDLSSPDAVMGNPKV
KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK
EFTPQVQAAYQKVVAGVANALAHKYH
83. Hemahum
HUMANI ALFA HEMOGLOBIN
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH
GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKL
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
84. Hembhum
BETA-HUMANI HEMOGLOBIN
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLST
PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDP
ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
85. Hemghum
HUMAN GAMA GLOBIN
GHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSS
ASAIMGNPKVKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDP
ENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTGVASALSSRYH
86. Hemhag
HEMOGLOBIN HAGFISH (MYXINE GLUTINOSA) (N)
PITDHGQPPTLSEGDKKAIRESWPQIYKNFEQNSLAVLLEFLKKFPKAQD
SFPKFSAKKSHLEQDPAVKLQAEVIINAVNHTIGLMDKEAAMKKYLKDLS
TKHSTEFQVNPDMFKELSAVFVSTMGGKAAYEKLFSIIATLLRSTYDA
87. Hemlamp
HEMOGLOBIN RIVER LAMPREY (N)
PIVDSGSVAPLSAAEKTKIRSAWAPVYSNYETSGVDILVKFFTSTPAAQE
FFPKFKGMTSADQLKKSADVRWHAERIINAVNDAVASMDDTEKMSMKLRD
LSGKHAKSFQVDPQYFKVLAAVIADTVAAGDAGFEKLMSMICILLRSA
88. Hemphaa
ALFA A - CHAIN FROM HBA PHEASANT (IBID)
VLSAADKNNVKGIFTKIAGHAEEYGAEALERMFITYPSTKTYFPHFDLSH
GSAQIKGHGKKVVAALIEAVNHIDDITGTLSKLSDLHAHKLRVDPVNFKL
LGQCFLVVVAIHHPSALTPEVHASLDKFLCAVGTVLTAKYR
89. Hemphad
ALFA D - CHAIN FROM CHICKEN HBD(J.BIOCHEM..77(1975)1345-1347).
MLNAEDKKLIQQAWEKAASDQQEFGAEALVRMFTAYPQTKTYFPHFDLSP
GSDQIRGHGKKVLGALSNAVKNVDNLSQAMSELSNLHAYNLRVDPVNFKL
LSQCIEVVLAVHMGKDYTPEVHAAFDKFLSAVSAVLAEKYR
90. Hemphb
BETA CHAIN FROM HBA PHEASANT
VHWSAEEKQLITGLWGKVNVADCGAEALARLLIVYPWTQRFFASFGNLSS
PTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDP
ENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH

91. Horsehmo
behoglobin alpha&beta chains - horse
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHGK
KVADGLTLAVGHLDDLPGALSDLSNLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPA
VHASLDKFLSSVSTVLTSKYRVQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQ
RFFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDP
ENFRLLGNVLALVVARHFGKDFTPELQASYQKVVAGVANALAHKYH

CYTOCHROME C (47 sequences)

1. Cccarp
Cytochromes c - Carp (2 strains)
GDVEKGKKVFVQKCAQCHTVZBGGKHKVGPNLWGLFGRKTGQAPGFSYTBABKSKGIVWB
ZZTLMEYLZBPKKYIPGTKMIFAGIKKKGERADLIAYLKSATS
2. Ccch
Cytochrome c - Chicken and turkey
GDIEKGKKIFVQKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAEGFSYTDANKNKGITWG
EDTLMEYLENPKKYIPGTKMIFAGIKKKSERVDLIAYLKDATSK
3. Ccdesl
Cytochrome c - Desert locust
GVPQGDVEKGKKIFVQRCAQCHTVEAGGKHKTGPNLHGLFGRKTGQAPGFSYTDANKSKG
ITWDENTLFIYLENPKKYIPGTKMVFAGLKKPEERADLIAYLKESTK
4. Ccdfsh
Cytochrome c - Puget Sound dogfish
GDVEKGKKVFVQKCAQCHTVENGGKHKTGPNLSGLFGRKTGQAQGFSYTDANKSKGITWQ
QETLRIYLENPKKYIPGTKMIFAGLKKKSERQDLIAYLKKTAAS
5. Ccdog
Cytochrome c - Dog, southern elephant seal, and Schreibers's long-fingered bat
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFSYTDANKNKGITWG
EETLMEYLENPKKYIPGTKMIFAGIKKTGERADLIAYLKKATKE
6. Ccdon
Cytochrome c - Donkey and common zebra (tentative sequences)
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFSYTDANKNKGITWK
EETLMEYLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE
7. Ccefsh
Cytochrome c - Common European starfish
GQVEKGKKIFVQRCAQCHTVEKAGKHKTGPNLNGILGRKTGQAAGFSYTDANRNKGITWK
NETLFEYLENPKKYIPGTKMVFAGLKKQKERQDLIAYLEAATK
8. Ccemu
Cytochrome c - Emu
GDIEKGKKIFVQKCSQCHTVEKGGKHKTGPNLNGLFGRKTGQAEGFSYTDANKNKGITWG
EDTLMEYLENPKKYIPGTKMIFAGIKKKSERADLIAYLKDATSK
9. Ccff
Cytochrome c - Mediterranean fruit fly, fruit fly, horn fly, and greenbottle fly
GVPAGDVEKGKKLFVQRCAQCHTVEAGGKHKVGPNLHGLIGRKTGQAAGFAYTDANKAKG
ITWNEDTLFEYLENPKKYIPGTKMIFAGLKKPNERGDLIAYLKSATK
10. Ccfrog
Cytochrome c - Bullfrog
GDVEKGKKIFVQKCAQCHTCEKGGKHKVGPNLYGLIGRKTGQAAGFSYTDANKNKGITWG
EDTLMEYLENPKKYIPGTKMIFAGIKKKGERQDLIAYLKSACSK
11. Cchbee
Cytochrome c - Honeybee
GIPAGDPEKGKKIFVQKCAQCHTIESGGKHKVGPNLYGVYGRKTGQAPGYSYTDANKGKG
ITWNKETLFEYLENPKKYIPGTKMVFAGLKKPQERADLIAYIEQASK
12. Cchor

Cytochrome c - Horse
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWK
EETLMEYLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE

13. Cchp

Cytochrome c - Hippopotamus
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQSPGFSYTDANKNKGITWG
EETLMEYLENPKKYIPGTKMIFAGIKKKGERADLIAYLKQATNE

14. Cchum

Cytochrome c - Human
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG
EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

15. Cchworm

Cytochrome c - Cynthia moth and tobacco hornworm
GVPAGNAENGKKIFVQRCAQCHTVEAGGKHKVGPNLHGFYGRKTGQAPGFSYSNANKAKG
ITWGDDTLFEYLENPKKYIPGTKMVFAGLKKANERADLIAYLKESTK

16. Cckang

Cytochrome c - Eastern gray kangaroo
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLNGIFGRKTGQAPGFTYTDANKNKGIIWG
EDTLMEYLENPKKYIPGTKMIFAGIKKKGERADLIAYLKKATNE

17. Cclam

Cytochrome c - Pacific lamprey
GDVEKGKKVFVQKCSQCHTVEKAGKHKTGPNLSGLFGRKTGQAPGFSYTDANKSKGIVWN
QETLFVYLENPKKYIPGTKMIFAGIKKEGERKDLIAYLKKSTSE

18. Cclanug

Cytochrome c - Thermomyces lanuginosus
AKGGSFEPGDASKGANLFKTRCAQCHSVEQGGANKIGPNLHGLFGRKTGSVEGYSYTDAN
KQAGITWNEDTLFEYLENPKKFIPGTKMAFGGLKKNKDRNDLITYLKEATK

19. Ccms

Cytochrome c, testis-specific - Mouse
GDAEAGKKIFVQKCAQCHTVEKGGKHKTGPNLWGLFGRKTGQAPGFSYTDANKNKGIVWS
ZZTLMZYLZBPKKYIPGTKMIFAGIKKKSEREDLIKYLKQATSS

20. Ccnc

Cytochrome c - Neurospora crassa
GFSAGDSKKGANLFKTRCAQCHTLEEGGGNKIGPALHGLFGRKTGSVDGYAYTDANKQKG
ITWDENTLFEYLENPKKYIPGTKMAFGGLKKDKDRNDIITFMKEATA

21. Ccos

Cytochrome c - Ostrich
GDIEKGKKIFVQKCSQCHTVEKGGKHKTGPNLDGLFGRKTGQAEGFSYTDANKNKGITWG
EDTLMEYLENPKKYIPGTKMIFAGIKKKSERADLIAYLKDATSK

22. Ccpig

Cytochrome c - Pig, bovine, and sheep
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFSYTDANKNKGITWG
EETLMEYLENPKKYIPGTKMIFAGIKKKGEREDLIAYLKKATNE

23. Ccping

Cytochrome c - King penguin
GDIEKGKKIFVQKCSQCHTVEKGGKHKTGPNLHGIFGRKTGQAEGFSYTDANKNKGITWG
EDTLMEYLENPKKYIPGTKMIFAGIKKKSERADLIAYLKDATSK

24. Ccplpum

Cytochrome c - Pumpkin
ASFDEAPPGNSKAGEKIFKTKCAQCHTVDKGAGHKQGPNLNGLFGRQSGTTPGYSYSAAN
KNRAVIWEEKTLYDYLLNPKKYIPGTKMVFPGLKKPQDRADLIAYLKEATA

25. Ccprawn

Cytochrome c - Monsoon river-prawn
GDVEKGKKIFVQRCAQCHSAQANLKHKTGPNLNGLFGRQTGQASGYVYTDANKAKGITWQ
ADTLDVYLENPKKYIPGTKMVFAGLKKANERADLIAYLKQATNL

26. Ccrab

Cytochrome c - Rabbit, mouse, rat, California gray whale, Arabian camel, and gua
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAVGFSYTDANKNKGITWG
EDTLMEYLENPKKYIPGTKMIFAGIKKKDERADLIAYLKKATNE

27. Ccrape
Cytochrome c - Rape and cauliflower
ASFDEAPPGNSKAGEKIFKTKCAQCHTVDKGAGHKQGPNLNGLFGRQSGTTAGYSYSAAN
KNKAVEWEEKTLYDYLLNPKKYIPGTKMVFPGLKKPQDRADLIAYLKEATA
28. Ccsna
Cytochrome c - Brown garden snail
GZAZKGKKIFTQKCLQCHTVEAGGKHKTGPNLSGLFGRKQGQAPGFAYTDANKGKGITWK
NQTLFEYLENPKKYIPGTKMVFAGLKBZTERVHLIAYLZZATKK
29. Ccsnak
Cytochrome c - Eastern diamondback rattlesnake
GDVEKGKKIFSMKCGTCHTVEEGGKHKTGPNLHGLFGRKTGQAVGYSYTAANKNKGIIWG
DDTLMEYLENPKKYIPGTKMVFTGLKSKKERTDLIAYLKEATAK
30. Ccspmon
Cytochrome c - Spider monkey
GDVFKGKRIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQASGFTYTEANKNKGIIWG
EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
31. Cctun
Cytochrome c - Skipjack tuna
GDVAKGKKTFVQKCAQCHTVENGGKHKVGPNLWGLFGRKTGQAEGYSYTDANKSKGIVWN
ENTLMEYLENPKKYIPGTKMIFAGIKKKGERQDLVAYLKSATS
32. Cceturt
Cytochrome c - Snapping turtle
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLNGLIGRKTGQAEGFSYTEANKNKGITWG
EETLMEYLENPKKYIPGTKMIFAGIKKKAERADLIAYLKDATSK
33. Ccus
Cytochrome c - Ustilago sphaerogena
GFEDGDAKKGARIFKTRCAQCHTLGAGEPNKVGPNLHGLFGRKSGTVEGFSYTDANKKAG
QVWEEETFLEYLENPKKYIPGTKMAFGGLKKEKDRNDLVTYLREETK
34. Ccworm
Cytochrome c - Common brandling worm
GGIPAGDVEKGKTIFKQRCAQCHTVDKGGPHKTGPNLHGIFGRATGQAAGFAYTDANKSK
GITWTKDTLYEYLENPKKYIPGTKMVFAGLKNEKQRANLIAYLEQETK
35. Ccyea
Cytochrome c - Yeast (Issatchenkia orientalis)
PAPFEQGSAKKGATLFKTRC[illegible]HTI[illegible]AGGPHKVGPNLHGIFSRHSGQAEGYSYTDANKR
AGVEWAEPTMSDYLENPKKY[illegible]LKKAKDRNDLVTYMLEASK
36. Ccyeah
Cytochrome c - Yeast (Hansenula anomala)
PAPFKKGSEKKGATLFKTRCLQCH[illegible]EKGGPHKVGPN[illegible]EGYSYTDANIK
KAVEWSEQTMSDYLENPKKYIPGTKMAFGGLKKEKDR[illegible]
37. Ccyeas
Cytochrome c - Yeast (Schizosaccharomyces pombe)
PYAPGDEKKGASLFKTRCAQCHTVEKGGANKVGPNLHG[illegible]GQAEGFSYTEANKDRG
ITWBZZTLFAYLENPKKYIPGTKMAFAGFKKPADRNNVITYLKKATSE
38. Ccyeat
Cytochrome c - Yeast (Torulaspora hansenii)
PAPYEKGSEKKGANLFKTRCLQCHTVEEGGPHKVGPNLHGVVGRTSGQAQGFSYTDANKK
KGVEWTEQDLSDYLENPKKYIPGTKMAFGGLKKAKDRNDLITYLVKATK
39. Cytcbov
Cytochrome c1, heme protein - Bovine
SDLELHPPSYPWSHRGLLSSLDHTSIRRGFQVYKQVCSSCHSMDYVAYRHLVGVCYTEDE
AKALAEEVEVQDGPNEDGEMFMRPGKLSDYFPKPYPNPEAARAANNGALPPDLSYIVRAR
HGGEDYVFSLLTGYCEPPTGVSLREGLYFNPYFPGQAIGMAPPIYNEVLEFDDGTPATMS
QVAKDVCTFLRWAAEPEHDHRKRMGLKMLLMMGLLLPLVYAMKRHKWSVLKSRKLAYRPP
K
40. Cytcdm55
Cytochrome c553 - Desulfovibrio vulgaris (strain Miyazaki)
ADGAALYKSCVGCHGADGSKQAMGVGHAVKGQKADELFKKLKGYADGSYGGEKKAVMTNL
VKRYSDEEMKAMADYMSKL

41. Cytcdv3
Cytochrome c3 - Desulfovibrio vulgaris (strain Miyazaki)
APKAPADGLKMDKTKQPVVFNHSTHKAVKCGDCHHPVNGKENYQKCATAGCH
DNMDKKDKSAKGYYHAMHDKGTKFKSCVGCHLETAGADAAKKKELTGCKGSK
CHS
42. Cytchum
Cytochrome c - Human
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG
EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
43. Cytcpig
Cytochrome c - Pig, bovine, and sheep
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFSYTDANKNKGITWG
EETLMEYLENPKKYIPGTKMIFAGIKKKGEREDLIAYLKKATNE
44. Cytcrhm
Cytochrome c' - Rhodospirillum molischianum
QQSKPEDLLKLRQGLMQTLKSQWVPIAGFAAGKADLPADAAQRAENMAMVAKLAPIGWAK
GTEALPNGETKPEAFGSKSAEFLEGWKALATESTKLAAAAKAGPDALKAQAAATGKVCKA
CHEEFKQD
45. Cytcrhr
Cytochrome c' - Rhodospirillum rubrum
ADPAAYVEYRKSVLSATSNYMKAIGITLKEDLAVPNQTADHAKAIASIMETLPAAFPEGT
AGIAKTEAKAAIWKDFEAFKVASKKSQDAALELASAAETGDKAAIGAKLQALGGTCKACH
KEFKAD
46. Cytctunh
Cytochrome c - Skipjack tuna
GDVAKGKKTFVQKCAQCHTVENGGKHKVGPNLWGLFGRKTGQAEGYSYTDANKSKGIVWN
ENTLMEYLENPKKYIPGTKMIFAGIKKKGERQDLVAYLKSATS
47. Cytcyea
Cytochrome c1, heme protein precursor - Yeast (Saccharomyces cerevisiae)
MFSNLSKRWAQRTLSKSFYSTATGAASKSGKLTQKLVTAGVAAAGITASTLLYADSLTAE
AMTAAEHGLHAPAYAWSHNGPFETFDHASIRRGYQVYREVCAACHSLDRVAWRTLVGVSH
TNEEVRNMAEEFEYDDEPDEQGNPKKRPGKLSDYIPGPYPNEQAARAANQGALPPDLSLI
VKARHGGCDYIFSLLTGYPDEPPAGVALPPGSNYNPYFPGGSIAMARVLFDDMVEYEDGT
PATTSQMAKDVTTFLNWCAEPEHDERKRLGLKTVIILSSLYLLSIWVKKFKWAGIKTRKF
VFNPPKPRK

FGF (38 sequences)

1. Athr3hum
Antithrombin-III precursor - Human
MYSNVIGTVTSGKRKVYLLSLLLIGFWDCVTCHGSPVDICTAKPRDIPMNPMCIYRSPEK
KATEDEGSEQKIPEATNRRVWELSKANSRFATTFYQHLADSKNDNDNIFLSPLSISTAFA
MTKLGACNDTLQQLMEVFKFDTISEKTSDQIHFFFAKLNCRLYRKANKSSKLVSANRLFG
DKSLTFNETYQDISELVYGAKLQPLDFKENAEQSRAAINKWVSNKTEGRITDVIPSEAIN
ELTVLVLVNTIYFKGLWKSKFSPENTRKELFYKADGESCSASMMYQEGKFRYRRVAEGTQ
VLELPFKGDDITMVLILPKPEKSLAKVEKELTPEVLQEWLDELEEMMLVVHMPRFRIEDG
FSLKEQLQDMGLVDLFSPEKSKLPGIVAEGRDDLYVSDAFHKAFLEVNEEGSEAAASTAV
VIAGRSLNPNRVTFKANRPFLVFIREVPLNTIIFMGRVANPCVK
2. Bek
flg FGF receptor (EMBO,9,2685-2692)
MvSWgrficlvVvtmATLslARPSfslvedtTLePEepptkyqisqpEvyvaaPGesLevRCLlkD
aavIsWtkDGVhLgpnNRTvliGEylqikgatPrDSGLYACtaSrtvdSeTwYFmVNVtDAiSSgD
DeDDtdgaEdfvsensnnkrAPYWTntEKMEKrLHAVPAAnTVKFrCPagGnPmPTmRWLKNGKEF
KqeHRIGGYKVRnqhWSlIMeSVVPSDKGNYTCvVENEYGSINHTYhLDVVERSPHRPILQAGLPA
NastvvGgdVEFvCKVYSDaQPHIQWiKHvEkNGSKyGPDgLPYlkvLKaAGVNTTDKEiEVLyiR
NVtFEDAGEYTCLAGNSIGiSfHSAWLTVLpApgrekeitaSPdYLEIaIYCiGvFLIaCMVvtVI
LcrMKntTKKpDFsSQpAVHKLtKrIPLRRQVTVSAeSSsSMNSntpLVRittRLSStadTPMLAG
VSEYELPEDPkWEfPRDkLtLGKPLGEGCFGQVVmAEAvGiDKDKPkeavtVAVKMLKdDATEKDL
SDLvSEMEMMKMIGKHKNIINLLGACTQDGPLYVIVEYASKGNLREYLrARRPPGmEYsYdinrvP
EEQmtfKDLVSCtYQlARGMEYLASqKCIHRDLAARNVLVTEnNVMKIADFGLARDInnIDYYKKT

TNGRLPVKWMAPEALFDRvYTHQSDVWSFGVLmWEIFTLGGSPYPGiPVEELFKLLKEGHRMDKPaN
CTNELYMMMRDCWHAVPSQRPTFKQLVEDLDRIltLTtNeEYLDLSqPLeQYSPSyPDTRSSTCSSG
dDSVFSpdPmPyEPCLPqyphiNGsvKt

3. Fdgf1ahu

Z1205299A : platelet derived growth factor A
MRTLACLLLLGCGYLAHVLAEEAEIPREVIERLARSQIHSIRDLQRLLEIDSVGSEDSLD
TSLRAHGVHATKHVPEKRPLPIRRKRSIEEAVPAVCKTRTVIYEIPRSQVDPTSANFLIW
PPCVEVKRCTGCCNTSSVKCQPSRVHHRSVKVAKVEYVRKKPKLKEVQVRLEEHLECACA
TTSLNPDYREEDTGRPRESGKKRKRKRLKPT

4. Fgf

Fatal Germ Factor
PALPEDGGSGAFPPGHFKDPKRLYCKNGGFFLRIHPDGRVDGVREKS
DPHIKLQLQAEERGVVSIKGVCANRYLAMKEDGRLLASKCVTDECFF
FERLESNNYNTYRSRKYSSWYVALKRTGQYKLGPKTGPGQKAILFLP
MSAKS□

5. Fgf97

fgf97-120
RLESNNYNTYRSRKYSSWYVALKR□

6. Fgf97-12

FGFB$BOVIN:BASIC FIBROBLAST GROWTH FACTOR PRECURSOR
(BFGF)(PROSTATROPIN).155
RLESNNYNTYRSRKYSSWYVALKR□

7. Fgfa10op

fgf analog, opposite order, 10 a.a.
DERKQQKRED

8. Fgfa11op

fgf analog, opposite order, 10 a.a.
LDERKQQKRED

9. Fgfa12op

fgf analog, opposite order, 12 a.a.
LDERKQQKREDI

10. Fgfa13op

fgf analog, opposite order, 13 a.a.
DLDERKQQKREDI

11. Fgfa14op

fgf analog but opposite order of 14 a.a.
MWYRPDLDERKQQK

12. Fgfa16op

fgf analog but opposite order of 16 a.a.
MWYRPDLDERKQQKRE

13. Fgfa18op

fgf analog but opposite order of 18 a.a.
MWYRPDLDERKQQKREDI

14. Fgfa20op

fgf analog but opposite order of a.a.
MWYRPDLDERKQQKREDIDP

15. Fgfa23op

fgf analog but opposite order of a.a.
MWYRPDLDERKQQKREDIDPRYW

16. Fgfa9

fgf analog, opposite order, 10 a.a.
KQQKREDID

17. Fgfabov

FGFA$BOVIN: ACIDIC FIBROBLAST GROWTH FACTOR (AFGF) (PROSTATROPIN). 140
AA.
FNLPLGNYKKPKLLYCSNGGYFLRILPDGTVDGTKDRSDQHIQLQLCAESIGEVYIKSTE
TGQFLAMDTDGLLYGSQTPNEECLFLERLEENHYNTYISKKHAEKHWFVGLKKNGRSKLG
PRTHFGQKAILFLPLPVSSD

18. Fgfahum

FGFA$HUMAN: ACIDIC FIBROBLAST GROWTH FACTOR (AFGF) (HEPARIN-BINDING
GROWTH FACTO

FNLPPGNYKKPKLLYCSNGGHFLRILPDGTVDGTRDRSDQHIQLQLSAESVGEVYIKSTE
TGQYLAMDTDGLLYGSQTPNEECLFLERLEENHYNTYISKKHAEKNWFVGLKKNGSCKRG
PRTHYGQKAILFLPLPVSSD

19. Fgfamop

fgf analog but opposite order of 18 a.a.
YRPDLDERKQQKREDIDP

20. Fgfana

analog fgf 0.4521,1,-2.5
CWYRPDIDERKQQKREDLDPRYWM

21. Fgfana1

analog fgf 0.4521,1,-2.5
CWYRPDIDERKQ

22. Fgfana15

CWYRPDIDERKQQKR

23. Fgfana16

analog fgf 0.4521,1,-2.5 without C1
PDIDERKQQKREDLDP

24. Fgfana20

analog fgf 0.4521,1,-2.5 without C1
PDIDERKQQKREDLDPRYWM

25. Fgfana23

analog fgf 0.4521,1,-2.5 without C1
WYRPDIDERKQQKREDLDPRYWM

26. Fgfana24

analog fgf 0.4521,1,-2.5
CWYRPDIDERKQQKREDLDPRYWM

27. Fgfanaop

fgf analog but opposite order of a.a.
MWYRPDLDERKQQKREDIDPRYWC

28. Fgfanax

fgf analog f=0.4521,a=1,ph=-1.587
DERAQQHRGDLDPTYWTPDIDNDH

29. Fgfaretb

Z1502231A : acidic fibroblast growth factor
MAEGETTTFTALTEKFNLPLGNYKKPKLLYCSNGGYFLRILPDGTVDGTKDRSDQHIQLQ
LCAESIGEVYIKSTETGQFLAMDTDGLLYGSQTPNEECLFLERLEENHYNTYISKKHAEK
HWFVGLKKNGRSKLGPRTHFGQKAILFLPLPVSSD

30. Fgfbfrg

Z1501413A : basic fibroblast growth factor
MAAGSITTLPTESEDGGNTPFSPGSFKDPKRLYCKNGGFFLRINSDGRVDGSRDKSDSHI
KLQLQAVERGVVSIKGITANRYLAMKEDGRLTSLRCITDECFFFERLEANNYNTYRSRKY
SSWYVALKRTGQYKNGSSTGPGQKAILFLPMSAKA

31. Fgfbhum

Z1212265A : fibroblast growth factor,basic
MAAGSITTLPALPEDGGSGAFPPGHFKDPKRLYCKNGGFFLRIHPDGRVDGVREKSDPHI
KLQLQAEERGVVSIKGVCANRYLAMKEDGRLLASKCVTDECFFFERLESNNYNTYRSRKY
TSWYVALKRTGQYKLGSKTGPGQKAILFLPMSAKS

32. Fgfbrat

Z1409358A : basic fibroblast growth factor
MAAGSITSLPALPEDGGGAFPPGHFKDPKRLYCKNGGFFLRIHPDGRVDGVREKSDPHVK
LQLQAEERGVVSIKGVCANRYLAMKEDGRLLASKCVTEECFFFERLESNNYNTYRSRKYS
SWYVALKRTGQYKLGSKTGPGQKAILFLPMSAKS

33. Fgfdes2

FGF DESIGNED SEQ BR 2.
RLESggYgfwRSRKYSSWYVALKR

34. Fgfrec

FGF receptor (SCIENCE, 245, 57-60,7 july 1989)
MFTWRCLILWAVLVTATLSAARPAPTLPDQALPKANIEVESHSAHPGDLLQLRCRLRDDV
QSINWVRDGVQLPENNRTRITGEEVEVRDRVPEDSGLYACMTNSPSGSETTYFSVNVSDA
LPSAEDDDDEDDSSSEEKEADNTKPNQAVAPYWTYPEKMEKKLHAVPAAKTVKFKCPSGG
TPNPTLRWLKNGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENKYGSINHT

YQLDVVERSPHRPILQAGLPANKTVALGSNVEFVCKVYSDPQPHIQWLKHIEVNGSKIGP
DNLPYVQILKTAGVNTTDKEMEVLHLRNVSFEDAGEYTCLAGNSIGISHHSAWLTVLEAT
EQSPAMMTSPLYLEIIIYCTGAFLISCMVVTVIIYKMKSTTKKTDFNSQLAVHKLAKSIP
LRRQVTVSADSSSSMNSGVMLVRPSRLSSSGTPMLAGVSEYELPEDPRWELPRDRLILGK
PLGEGCFGQVVLAEAIGLDKDKPNRVTKVAVKMLKSDATEKDLSDLISEMEMMKMIGKHK
NIINLLGACTQDGPLYVIVEYASKGNLREYLQARRPPGMEYCYNPTRIPEEQLSFKDLVS
CAYQVARGMEYLASKKCIHRDLAARNVLVTEDNVMKIADFGLARDIHHIDYYKKTTNGRL
PVKWMAPEALFDRIYTHQSDVWSFGVLLWEIFTLGGSPYPGVPVEELFKLLKEGHRMDKP
SNCTNELYMMMRDCWHAVPSQRPTFKQLVEDLDRIVAMTSNQEYLDLSVPLDQYSPGFPA
TRSSTCSSGEDSVFSHDPLPDEPCLPRCPPHSHGALKRH

35. Fgfrecl

FGF receptor WITH leading sequense (SCIENCE, 245, 57-60,7 july 1989)
AQSLSSSRSSGMFTWRCLILWAVLVTATLSAARPAPTLPDQALPKANIEVESHSAHPGDL
LQLRCRLRDDVQSINWVRDGVQLPENNRTRITGEEVEVRDRVPEDSGLYACMTNSPSGSE
TTYFSVNVSDALPSAEDDDDEDDSSSEEKEADNTKPNQAVAPYWTYPEKMEKKLHAVPAA
KTVKFKCPSGGTPNPTLRWLKNGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCI
VENKYGSINHTYQLDVVERSPHRPILQAGLPANKTVALGSNVEFVCKVYSDPQPHIQWLK
HIEVNGSKIGPDNLPYVQILKTAGVNTTDKEMEVLHLRNVSFEDAGEYTCLAGNSIGISH
HSAWLTVLEATEQSPAMMTSPLYLEIIIYCTGAFLISCMVVTVIIYKMKSTTKKTDFNSQ
LAVHKLAKSIPLRRQVTVSADSSSSMNSGVMLVRPSRLSSSGTPMLAGVSEYELPEDPRW
ELPRDRLILGKPLGEGCFGQVVLAEAIGLDKDKPNRVTKVAVKMLKSDATEKDLSDLISE
MEMMKMIGKHKNIINLLGACTQDGPLYVIVEYASKGNLREYLQARRPPGMEYCYNPTRIP
EEQLSFKDLVSCAYQVARGMEYLASKKCIHRDLAARNVLVTEDNVMKIADFGLARDIHHI
DYYKKTTNGRLPVKWMAPEALFDRIYTHQSDVWSFGVLLWEIFTLGGSPYPGVPVEELFK
LLKEGHRMDKPSNCTNELYMMMRDCWHAVPSQRPTFKQLVEDLDRIVAMTSNQEYLDLSV
PLDQYSPGFPATRSSTCSSGEDSVFSHDPLPDEPCLPRCPPHSHGALKRH

36. Fgfrrm16

fgf analog but opposite order of 16 a.a.
MWYRPDLDERKQQKRE

37. Fgfrrm24

fgf analog but opposite order of a.a.
MWYRPDLDERKQQKREDIDPRYWC

38. Flg

flg FGF receptor (EMBO,9,2685-2692)
MWSWKCLLFWAVLVTATLCTARPSPTLPEQAQPWGAPVEVESFLVHPGDLLQLRCRLRDDVQSINWL
RDGVQLAESNRTRITGEEVEVQDSVPADSGLYACVTSSPSGSDTTYFSVNVSDALPSSEDDDDDDDS
SSEEKETDNTKPNRMPVAPYWTSPEKMEKKLHAVPAAKTVKFKCPSSGTPNPTLRWLKNGKEFKPDH
RIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILQAGLPANKTVA
LGSNVEFMCKVYSDPQPHIQWLKHIEVNGSKIGPDNLPYVQILKTAGVNTTDKEMEVLHLRNVSFED
AGEYTCLAGNSIGLSHHSAWLTVLEALEERPAVMTSPLYLEIIIYCTGAFLISCMVGSVIVYKMKSG
TKKSDFHSQMAVHKLAKSIPLRRQVTVSADSSASMNSGVLLVRPSRLSSSGTPMLAGVSEYELPEDP
RWELPRDRLVLGKPLGEGCFGQVVLAEAIGLDKDKPNRVTKVAVKMLKSDATEKDLSDLISEMEMMK
MIGKHKNIINLLGACTQDGPLYVIVEYASKGNLREYLQARRPPGLEYCYNPSHNPEEQLSSKDLVSC
AYQVARGMEYLASKKCIHRDLAARNVLVTEDNVMKIADFGLARDIHHIDYYKKTTNGRLPVKWMAPE
ALFDRIYTHQSDVWSFGVLLWEIFTLGGSPYPGVPVEELFKLLKEGHRMDKPSNCTNELYMMMRDCW
HAVPSQRPTFKQLVEDLDRIVALTSNQEYLDLSMPLDQYSPSFPDTRSSTCSSGEDSVFSHEPLPEE
PCLPRHPAQLANGGLKRR

EGF (73 sequences)

1. Agegf20

egf agonist
LNPKWQRDDDDRCWAPEINP

2. Agegf20s

egf agonist
LNPKYQRDDDDRsWAPEINP

3. Agegf20t

egf agonist
LNPKYQRDDDDRtWAPEINP

4. Agegf21

egf agonist
LNPKWQRDDDDRCWAPEINPH
5. Agegf21s
egf agonist
LNPKWQRDDDDRsWAPEINPH
6. Agegf21t
egf agonist
LNPKWQRDDDDRtWAPEINPH
7. Agegf22
egf agonist
LNPKWQRDDDDRCWAPEINPHY
8. Agegf23
egf agonist
LNPKWQRDDDDRCWAPEINPHYQ
9. Agegf23s
egf agonist
LNPKYQRDDDDRsWAPEINPHYQ
10. Agegf23t
egf agonist
LNPKYQRDDDDRtWAPEINPHYQ
11. Agegf24
egf agonist
LNPKWQRDDDDRCWAPEINPHYQR
12. Agegf25
egf agonist
LNPKWQRDDDDRCWAPEINPHYQRD
13. Agegf26
egf agonist
LNPKWQRDDDDRCWAPEINPHYQRDD
14. Agegf27
egf agonist
LNPKYQRDDDDRCWAPEINPHYQRDDD
15. Agegf28
egf agonist
LNPKYQRDDDDRCWAPEINPHYQRDDDD
16. Agegf28s
egf agonist
LNPKYQRDDDDRsWAPEINPHYQRDDDD
17. Agegf28t
egf agonist
LNPKYQRDDDDRtWAPEINPHYQRDDDD
18. Agegf29
egf agonist
LNPKYQRDDDDRCWAPEINPHYQRDDDDR
19. Agegf30-
egf agonist
KPNIGPAWCRDDDDRMWKPNLGPKWMRDDD
20. Agegf300
egf agonist
LNPKYQRDDDDRCWAPEINPHYQRDDDDRC
21. Agegf302
egf agonist
NIEPAWCRDDDDRQYHPNIEPAWCRDDDDR
22. Agegf303
egf agonist
YHEIIEPAWCRDDDDRQYHENLEPAWCRDD
23. Agegf305
egf agonist
EILEHYQTRDDDRTQYHEIIEHAQTRDDDD
24. Agegf308

egf agonist
PELIEHYQTDDDDRTQAHEILEHYQTRDDD
25. Agegf30s
egf agonist
LNPKYQRDDDDRsWAPEINPHYQRDDDDRC
26. Agegf30t
egf agonist
LNPKYQRDDDDRtWAPEINPHYQRDDDDRC
27. Anegf20
egf ant
CRDDDDRQYHPNIEPAWCRD
28. Anegf21
egf ant
CRDDDDRQYHPNIEPAWCRDD
29. Anegf22
egf ant
CRDDDDRQYHPNIEP[illegible]
30. Anegf23
egf ant
CRDDDDRQYHPNIE[illegible]PDC
31. Anegf24
egf ant
CRDDDDRQYHPNIEPAW[illegible]DR
32. Anegf25
egf ant
CRDDDDRQYHPNIEPAWCRDDDDRQ
33. Anegf26
egf antagonist
CRDDDDRQYHPNIEPAWCRDDDDRQW
34. Anegf27
egf antagonist
CRDDDDRQYHPNIEPAWCRDDDDRQWK
35. Anegf28
egf antagonist
CRDDDDRQYHPNIEPAWCRDDDDRQWKP
36. Anegf29
egf antagonist
CRDDDDRQYHPNIEPAWCRDDDDRQWKPN
37. Anegf300
egf antagonist
CRDDDDRQYKPNIEPAWCRDDDDRQWKPNL
38. Anegf301
egf antagonist
DDDRMWKPGLNPKWQRDDDDRCWAPVINPK
39. Anegf302
egf ant
DDRTQYHEILEHKQTRDDDRTQYHEIIEHK
40. Anegf305
egf ant
DDRQYHENIEPAWCRDDDDRQYHPNIEPAW
41. Antegf20
egf ant
DDRTQAHEIIEHYQTRDDDR
42. Antegf21
egf antagonist
DDRTQAHEIIEHYQTRDDDRT
43. Antegf22
egf ant
DDRTQAHEIIEHYQTRDDDRTQ
44. Antegf23

egf ant
DDRTQAHEIIEHYQTRDDDRTQA
45. Antegf24
egf ant
DDRTQAHEIIEHYQTRDDDRTQAH
46. Antegf25
egf ant
DDRTQAHEIIEHYQTRDDDRTQAHE
47. Antegf26
egf antagonist
DDRTQAHEIIEHYQTRDDDRTQAHEI
48. Antegf27
egf antagonist
DDRTQAHEIIEHYQTRDDDRTQAHEII
49. Antegf28
egf antagonist
DDRTQAHEIIEHYQTRDDDRTQAHEIIE
50. Antegf29
egf antagonist
DDRTQAHEIIEHYQTRDDDRTQAHEIIEH
51. Antegf30
egf antagonist
DDRTQAHEIIEHYQTRDDDRTQAHEIIEHY
52. Egfbnd
Epidermal growth factor-binding protein type B - Mouse
VVGGFNCEKNSQPWQVAVYYQKEHICGGVLLDRNWVLTAAHCYVDQYEVWLGKNKLFQEEPSAQHR
LVSKSFPHPGFNMSLLMLQTTPPGADFSNDLMLLRLSKPADITDVVKPIALPTKEPKPGSTCLASG
WGSITPTRWQKSDDLQCVFITLLPNENCAKVYLQKVTDVMLCAGEMGGGKDTCAGDSGGPLICDGI
LQGTTSNGPEPCGKPGVPAIYTNLIKFNSWIKDTMMKNA
53. Egfbndp
Epidermal growth factor-binding protein type B precursor - Mouse
MWFLILFPALSLGGIDAAPPLQSRVVGGFNCEKNSQPWQVAVYYQKEHIC
GGVLLDRNWVLTAAHCYVDQYEVWLGKNKLFQEEPSAQHRLVSKSFPHPG
FNMSLLMLQTTPPGADFSNDLMLLRLSKPADITDVVKPIALPTKEPKPGS
TCLASGWGSITPTRWQKSDDLQCVFITLLPNENCAKVYLQKVTDVMLCAG
EMGGGKDTCAGDSGGPLICDGILQGTTSNGPEPCGKPGVPAIYTNLIKFN
SWIKDTMMKNA
54. Egfchrec
Z1408242A : epidermal growth factor receptor
MGVRSPLSASGPRGAAVLVLLLLGVALCSAVEEKKVCQGTNNKLTQLGHVEDHFTSLQRM
YNNCEVVLSNLEITYVEHNRDLTFLKTIQEVAGYVLIALNMVDVIPLENLQIIRGNVLYD
NSFALAVLSNYHMNKTQGLRELPMKRLSEILNGGVKISNNPKLCNMDTVLWNDIIDTSRK
PLTVLDFASNLSSCPKCHPNCTEDHCWGAGEQNCQTLTKVICAQQCSGRCRGKVPSDCCH
NQCAAGCTGPRESDCLACRKFRDDATCKDTCPPLVLYNPTTYQMDVNPEGKYSFGATCVR
ECPHNYVVTDHGSCVRSCNTDTYEVEENGVRKCKKCDGLCSKVCNGIGIGELKGILSINA
TNIDSFKNCTKINGDVSILPVAFLGDAFTKTLPLDPKKLDVFRTVKEISGFLLIQAWPDN
ATDLYAFENLEIIRGRTKQHGQYSLAVVNLKIQSLGLRSLKEISDGDIAIMKNKNLCYAD
TMNWRSLFATQSQKTKIIQNRNKNDCTADRHVCDPLCSDVGCWGPGPFHCFSCRFFSRQK
ECVKQCNILQGEPREFERDSKCLPCHSECLVQNSTAYNTTCSGPGPDHCMKCAHFIDGPH
CVKACPAGVLGENDTLVWKYADANAVCQLCHPNCTRGCKGPGLEGCPNGSKTPSIAAGVV
GGLLCLVVVGLGIGLYLRRRHIVRKRTLRRLLQERELVEPLTP
55. Egfdrrec
Z1107191A : epidermal growth factor receptor
MNYDFDLFCAATARPPSRSRRVRRRLINVATFSLSLRITNYIVIGLDLIPCTLSYRLQII
RGRTLFSLSVEEEKYALFVTYSKMYTLEIPDLRDVLNGQVGFHNNYNLCHMRTIQWSEIV
SNGTDAYYNYDFTAPERECPKCHESCTHGCWGEGPKNCQKFSKLTCSPQCAGGRCYGPKP
RECCHLFCAGGCTGPTQKDCIACKNFFDEAVSKEECPPMRKYNPTTYVLETNPEGKYAYG
ATCVKECPGHLLRDNGACVRSCPQDKMDKGGECVPCNGPCPKTCPGVTVLHAGNIDSFRN
CTVIDGNIRILDQTFSGFQDVYANYTMGPRYIPLDPERREVFSTVKEITGYLNIEGTHPQ
FRNLSYFRNLETIHGRQLMESMFAALAIVKSSLYSLEMRNLKQISSGSVVIQHNRDLCYV
SNIRWPAIQKEPEQKVWVNENLRADLCGKF

56. Egfffrec

Epidermal growth factor receptor - Fruit fly

MNYDFDLFCAATARPPSRSRRVRRRLINVATFSLSLRITNYIVIGLDLIP
CTLSYRLQIIRGRTLFSLSVEEEKYALFVTYSKMYTLEIPDLRDVLNGQV
GFHNNYNLCHMRTIQWSEIVSNGTDAYYNYDFTAPERECPKCHESCTHGC
WGEGPKNCQKFSKLTCSPQCAGGRCYGPKPRECCHLFCAGGCTGPTQKDC
IACKNFFDEAVSKEECPPMRKYNPTTYVLETNPEGKYAYGATCVKECPGH
LLRDNGACVRSCPQDKMDKGGECVPCNGPCPKTCPGVTVLHAGNIDSFRN
CTVIDGNIRILDQTFSGFQDVYANYTMGPRYIPLDPERREVFSTVKEITG
YLNIEGTHPQFRNLSYFRNLETIHGRQLMESMFAALAIVKSSLYSLEMRN
LKQISSGSVVIQHNRDLCYVSNIRWPAIQKEPEQKVWVNENLRADLCGKF
LTILISVQHNIIMHIFAICREKWNHLLGSVQRGRLLGSWHGSVPYLQELQ
FQWHLHRRLWLYIQVSINSTQDKSNEHQLTDACYSPSVPTSLTIERARYA
IQSAGLAMELEQITARSASMRHSKTLPAEGRQVPRWVFLGVCASARAGIA
EPLAGRAVCRKCHPLCELCTNYGYHEQVCSKCTHYKRREQCETECPADHY
TDEEQRECFQRHPECNGCTGPGADDCKSCRNFKLFDANETGPYVNSTMFN
CTSKCPLEMRHVNYQYTAIGPYCAASPPRSSKITANLDVNMIFIITGAVL
VPTICILCVVTYICRQKQKAKKETVKMTMALSGREDSEPLRPSNIGANLC
KLRIVKDAELRKGGVLGMGAFGRVYKGVWVPEGENVKIPVAIKELLKSTG
AESSEEFLREAYIMASEEHVNLLKLLAVCMSSQMMLITQLMPLGCLLDYV
RNNRDKIGSKALLNWSTQIAKGMSYLEEKRLVHRDLAARNVLVRLLAGED
HDFGLAKLLSSDSNEYKAAGGKMPIKWLALECIRNRVFTSKSDVWAFGVT
IWELLTFGQRPHENIPAKDIPDLIEVGLKLEQPEICSLDIYCTLLSCWHL
DAAMRPTFKQLTTVFAEFARDPGRYLAILGDKFTRLPAYTSQDEKDLIRK
LAPTTDGSEAIAKPDDYLQPKAALGPSHRTDCTDEMPKLNRYCKDPSNKN
SSTGDDERDSSAREVGVGNLRLDLPVDEDDYLMPTCQPGPNNNNNMNNPN
QNNMAAVGVAAGYMDLIGVPVSVDNPEYLLNAQTLGVGESPIPTQTIGIP
VMGGPGTMEVKVPMPGSEPTSSDHEYYNDTQRSCSHASKPQHGDEGVGSS
RVGAIANEEGESCQVPLEAMRYAFAGCYLR

57. Egfgpig

EGF Ginuea Pig

QDAPGCPPSHDGYCLHGGVCMHIESLNTYACNCVIGYVGERCEHQDLDLWE

58. Egfh1rec

Z1006266A : epidermal growth factor receptor

PSGTAGAALLALLAALCPASRALEEKKVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEV
VLGNLEITYVQRNYDLSFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNMYYENSYALA
VLSNYDANKTGLKELPMRNLQEILHGAVRFSNNPALCNVESIQWRDIVSSDFLSNMSMDF
QNHLGSCQKCDPSCPNGSCWGAGEENCQKLTKIICAQQCSGRCRGKSPSDCCHNQCAAGC
TGPRESDCLVCRKFRDEATCKDTCPPLMLYNPTTYQMDVNPEGKYSFGATCVKKCPRNYV
VTDHGSCVRACGADSYEMEEDGVRKCKKCEGPCRKVCNGIGIGEFKDSLSINATNIKHFK
NCTSISGDLHILPVAFRGDSFTHTPPLDPQELDILKTVKEITGFLLIQAWPENRTDLHAF
ENLEIIRGRTKQHGQFSLAVVSLNITSLGLRSLKEISDGDVIISGNKNLCYANTINWKKL
FGTSGQKTKIISNRGENSCKATGQVCHALCSPEGCWGPEPRDCVSCRNVSRGRECVDKCK
LLEGEPREFVENSECIQCHPECLPQAMNITCTGRGPDNCIQCAHYIDGPHCVKTCPAGVM
GENNTLVWKYADAGHVCHLCHPNCTYGCTGPGLEGCPTNGPKIPSIATGMVGALLLLLVV
ALGIGLFMRRRHIVRKRTLRRLLQERELVEPLTPSGEAPNQALLRILKETEFKKIKVLGS
GAFGTVYKGLWIPEGEKVKIPVAIKELREATSPKANKEILDEAYVMASVDNPHVCRLLGI
CLTSTVQLITQLMPFGCLLDYVREHKDNIGSQYLLNWCVQIAKGMNYLEDRRLVHRDLAA
RNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWMALESILHRIYTHQSDVWSY
GVTVWELMTFGSKPYDGIPASEISSILEKGERLPQPPICTIDVYMIMVKCWMIDADSRPK
FRELIIEFSKMARDPQRYLVIQGDERMHLPSPTDSNFYRALMDEEDMDDVVDADEYLIPQ
QGFFSSPSTSRTPLLSSLSATSNNSTVACIDRNGLQSCPIKEDSFLQRYSSDPTGALTED
SIDDTFLPVPEYINQSVPKRPAGSVQNPVYHNQPLNPAPSRDPHYQDPHSTAVGNPEYLN
TVQPTCVNSTFDSPAHWAQKGSHQISLDNPDYQQDFFPKEAKPNGIFKGSTAENAEYLRV
APQSSEFIGA

59. Egfh431r

Z1011170A : epidermal growth factor receptor

STSRTPLLSSLSATSNNSTVACIDRNGLQSCPIKEDSFLQRYSSDPTGALTEDSIDDTFL
PVPEYINQSVPKRPAGSVQNPVYHNQPLNPAPSRDPHYQDPHSTAVGNPEYLNTVQPTCV
NSTFDSPAHWAQKGSHQISLDNPDYQQDFFPKEAKPNGIFKGSTAENAEYLRVAPQSSEF
IGA

60. Egfbprec

Z1301332A : epidermal growth factor precursor

MLLTLIILLPVVSKFSFVSLSAPQHWSCPEGTLAGNGNSTCVGPAPFLIFSHGNSIFRID
TEGTNYEQLVVDAGVSVIMDFHYNEKRIYWVDLERQLLQRVFLNGSRQERVCNIEKNVSG
MAINWINEEVIWSNQQEGIITVTDMKGNNSHILLSALKYPANVAVDPVERFIFWSSEVAG
SLYRADLDGVGVKALLETSEKITAVSLDVLDKRLFWIQYNREGSNSLICSCDYDGGSVHI
SKHPTQHNLFAMSLFGDRIFYSTWKMKTIWIANKHTGKDMVRINLHSSFVPLGELKVVHP
LAQPKAEDDTWEPEQKLCKLRKGNCSSTVCGQDLQSHLCMCAEGYALSRDRKYCEDVNEC
AFWNHGCTLGCKNTPGSYYCTCPVGFVLLPDGKRCHQLVSCPRNVSECSHDCVLTSEGPL
CFCPEGSVLERDGKTCSGCSSPDNGGCSQLCVPLSPVSWECDCFPGYDLQLDEKSCAASG
PQPFLLFANSQDIRHMHFDGTDYGTLLSQQMGMVYALDHDPVENKIYFAHTALKWIERAN
MDGSQRERLIEEGVDVPEGLAVDWIGRRFYWTDRGKSLIGRSDLNGKRSKIITKENISQP
RGIAVHPMAKRLFWTDTGINPRIESSSLQGLGRLVIASSDLIWPSGITIDFLTDKLYWCD
AKQSVIEMANLDGSKRRRLTQNDVGHPFAVAVFEDYVWFSDWAMPSVMRVNKRTGKDRVR
LQGSMLKPSSLVVVHPLAKPGADPCLYQNGGCEHICKKRLGTAWCSCREGFMKASDGKTC
LALDGHQLLAGGEVDLKNQVTPLDILSKTRVSEDNITESQHMLVAEIMVSDQDDCAPVGC
SMYARCISEGEDATCQCLKGFAGDGKLCSDIDECEMGVPVCPPASSKCINTEGGYVCRCS
EGYQGDGIHCLDIDECQLGVHSCGENASCTNTEGGYTCMCAGRLSEPGLICPDSTPPPHL
REDDHHYSVRNSDSECPLSHDGYCLHDGVCMYIEALDKYACNCVVGYIGERCQYRDLKWW
ELRHAGHGQQQKVIVVAVCVVVLVMLLLLSLWGAHYYRTQKLLSKNPKNPYEESSRDVRS
RRPADTEDGMSSCPQPWFVVIKEHQDLKNGGQPVAGEDGQAADGSMQPTSWRQEPQLCGM
GTEQGCWIPVSSDKGSCPQVMERSFHMPSYGTQTLEGGVEKPHSLLSANPLWQQRALDPP
HQMELTQ

61. Egfhrec

Z1006306A : epidermal growth factor receptor

KKIKVLGSGAFGTVYKGLWIPEGEKVKIPVAIKELREATSPKANKEILDEAYVMASVDNP
HVCRLLGICLTSTVQLITQLMPFGCLLDYVREHKDNIGSQYLLNWCVQIAKGMNYLEDRR
LVHRDLAARNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWMALESILHRIYT
HQSDVWSYGVTVWELMTFGSKPYDGIPASEISSILEKGERLPQPPICTIDVYMIMVKCWM
IDADSRPKFREL

62. Egfhrec2

Z1007208A : epidermal growth factor receptor

RCAWRRAFSNNPALCNVESIQWRDIVSSDFLSNMSMDFQNHLGSCKSVIQAVPMGACWGA
GEENCQKLTKIICAQQCSGRSRGKSPSDCCHHQCAAGCTGPRESDCLVCRKFRDEATCKD
TCPPLMLYNPTTYQMDVNPEGKYSFGATCVKKCPRNYVVTDHRACVRACGADSYEMEEDA
VRKCKKCEGPCRKVCNGIGIGEFKDSLSINATNIKHFKNCTSISGDLHILVAFRGDSFTH
TPPLDPQELDILKTVKEITGFLLIQAWPENRTDLHAFENLEIIRGRTKQHGQFSLAVVSL
NITSLGLRSLKEISDGDVIISGNKNLCYANTINWKKLFGTSGQKTKIISNRGENSCKATQ
GSAMPCAPRGCWGRARDCVSCRNVSRGRECVDKCNLLEGEPREFVENSECIQCHPECLPQ
AMNITCTGRGPDNCIQCAHYIDGPHCVKTCPAGVMGENNTLVWKYADAGHVCHLCHPNCT
YGCTGPGLEGCPTNGPKIPSIATGMVGALLLLLVVALGIGLFIGRRHIVRKATLRRLLQE
RELVEPLTPSGEAPNQALLRILKETEFKKIKVLGSGAFGTVYKGLWIPEGEKVKIPVAIK
ELREATSPKANKEILDEAYVMASVDNPHVCRLLGICLTSTVQLITQLMPSAGLLTDVREH
KDNIGSRYLLNWCVQIAKGMNYLEDRRLVHRDLAARNVLVKTPQHVKITDFGLAKLLGAE
EKEYHAEGGKVPIKWMALESILHPT HQSDVWSYGVTVWELMTFGSKPYDGIPASEISS
ILEKGERLPQPPICTID

63. Egfhrec3

Z1201275A : epidermal growth factor receptor

MRPSGTAGAALLALLAALCPASRALEEKKVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEV
VLGNLEITYVQRNYDLSFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNMYYENSYALA
VLSNYDANKTGLKELPMRNLQEILHGAVRFSNNPALCNVESIQWRDIVSSDFLSNMSMDF
QNHLGSCQKCDPSCPNGSCWGAGEENCQKLTKIICAQQCSGRCRGKSPSDCCHNQCAAGC
TGPRESDCLVCRKFRDEATCKDTCPPLMLYNPTTYQMDVNPEGKYSFGATCVKKCPRNYV
VTDHGSCVRACGADSYEMEEDGVRKCKKCEGPCRKVCNGIGIGEFKDSLSINATNIKHFK
NCTSISGDLHILPVAFRGDSFTHTPPLDPQELDILKTVKEITGFLLIQAWPENRTDLHAF
ENLEIIRGRTKQHGQFSLAVVSLNITSLGLRSLKEISDGDVIISGNKNLCYANTINWKKL
FGTSGQKTKIISNRGENSCKATGQVCHALCSPEGCWGPEPRDCVSCRNVSRGRECVDKCN
LLEGEPREFVENSECIQCHPECLPQAMNITCTGRGPDNCIQCAHYIDGPHCVKTCPAGVM
GENNTLVWKYADAGHVCHLCHPNCTYGCTGPGLEGCPTNGSYIVSHFPRSFYKMSVH

64. Egfhrfr

Z1109269A : epidermal growth factor receptor
MRPSGTAGAALLALLAALCPASRALEEKK

65. Egfhum

Epidermal growth factor - Human
NSDSECPLSHDGYCLHDGVCMYIEALDKYACNCVVGYIGERCQYRDLKWWELR

66. Ehgmous

EGF mouse
NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR

67. Egfmpr

Z0907234A : epidermal growth factor precursor
MPWGRRPTWLLLAFLLVFLKISILSVTAWQTGNCQPGPLERSERSGTCAGPAPFLVFSQG
KSISRIDPDGTNHQQLVVDAGISADMDIHYKKERLYWVDVERQVLLRVFLNGTGLEKVCN
VERKVSGLAIDWIDDEVLWVDQQNGVITVTDMTGKNSRVLLSSLKHPSNIAVDPIERLMF
WSSEVTGSLHRAHLKGVDVKTLLETGGISVLTLDVLDKRLFWVQDSGEGSHAYIHSCDYE
GGSVRLIRHQARHSLSSMAFFGDRIFYSVLKSKAIWIANKHTGKDTVRINLHPSFVTPGK
LMVVHPRAQPRTEDAAKDPDPELLKQRGRPCRFGLCERDPKSHSSACAEGYTLSRDRKYC
EDVNECATQNHGCTLGCENTPGSYHCTCPTGFVLLPDGKQCHELVSCPGNVSKCSHGCVL
TSDGPRCICPAGSVLGRDGKTCTGCSSPDNGGCSQICLPLRPGSWECDCFPGYDLQSDRK
SCAASGPQPLLLFANSQDIRHMHFDGTDYKVLLSRQMGMVFALDYDPVESKIYFAQTALK
WIERANMDGSQRERLITEGVDTLEGLALDWIGRRIYWTDSGKSVVGGSDLSGKHHRIIIQ
ERISRPRGIAVHPRARRLFWTDVGMSPRIESASLQGSDRVLIASSNLLEPSGITIDYLTD
TLYWCDTKRSVIEMANLDGSKRRRLIQNDVGHPFSLAVFEDHLWVSDWAIPSVIRVNKRT
GQNRVRLQGSMLKPSSLVVVHPLAKPGADPCLYRNGGCEHICQESLGTARCLCREGFVKA
WDGKMCLPQYYPILSGENADLSKEVTSLSNSTQAEVPDDDGTESSTLVAEIMVSGMNYED
DCGPGGCGSHARCVSDGETAECQCLKGFARDGNLCSDIDECVLARSDCPSTSSRCINTEG
GYVCRCSEGYEGDGISCFDIDECQRGAHNCAENAACTNTEGGYNCTCAGRPSSPGRSCPD
STAPSLLGEDGHHLDRNSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQT
RDLRWWELRHAGYGQKHDIMVVAVCMVSLVLLLLLGMWGTYYYRTRKQLSNPPKNPCDEP
SGSVSSSGPDSSSGAAVASCPQPWFVVLEKHQDPKNGSLPADGTNGAVVDAGLVSLPAAR
VSASDFMETEAPHRWNGNRAKLLDSTIK

68. Egfrat

EGF RAT epidermal growth factor
NSNTGCPPSYDGYCLNGGVCMYVESVDRYVCNCVIGYIGERCQHRDLR

69. Egfratpr

Z1414344A : epidermal growth factor
WKRWLHWVRNSNTGCPPSYDGYCLNGGVCMYVESVDRYVCNCVIGYIGERCQHRDLRWWN
WRHADYGQRHDITVVSVCVVALALLLLLGMWGTYYYRTRKQLSESSKKPSEESSSNVSSN
GPDSSGAGVSSGPQPWFVVLEEHQ

70. h1egfr

EGFS$HUMAN: EPIDERMAL GROWTH FACTOR RECEPTOR (VERSION 2) (GENE NAME:
EGFR) 797 A
RCAWRRAFSNNPALCNVESIQWRDIVSSDFLSNMSMDFQNHLGSCKSVIQAVPMGACWGA
GEENCQKLTKIICAQQCSGRSRGKSPSDCCHNQCAAGCTGPRESDCLVCRKFRDEATCKD
TCPPLMLYNPTTYQMDVNPEGKYSFGATCVKKCPRNYVVTDHRACVRACGADSYEMEEDA
VRKCKKCEGPCRKVCNGIGIGEFKDSLSINATNIKHFKNCTSISGDLHILVAFRGDSFTH
TPPLDPQELDILKTVKEITGFLLIQAWPENRTDLHAFENLEIIRGRTKQHGQFSLAVVSL
NITSLGLRSLKEISDGDVIISGNKNLCYANTINWKKLFGTSGQKTKIISNRGENSCKATQ
GSAMPCAPRGCWGRARDCVSCRNVSRGRECVDKCNLLEGEPREFVENSECIQCHPECLPQ
AMNITCTGRGPDNCIQCAHYIDGPHCVKTCPAGVMGENNTLVWKYADAGHVCHLCHPNCT
YGCTGPGLEGCPTNGPKIPSIATGMVGALLLLLVVALGIGLFIGRRHIVRKATLRRLLQE
RELVEPLTPSGEAPNQALLRILKETEFKKIKVLGSGAFGTVYKGLWIPEGEKVKIPVAIK
ELREATSPKANKEILDEAYVMASVDNPHVCRLLGICLTSTVQLITQLMPSAGLLTDVREH
KDNIGSRYLLNWCVQIAKGMNYLEDRRLVHRDLAARNVLVKTPQHVKITDFGLAKLLGAE
EKEYHAEGGKVPIKWMALESILHRIYTHQSDVWSYGVTVWELMTFGSKPYDGIPASEISS
ILEKGERLPQPPICTID

71. Megf1046

EGF mouse 10-46
YDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRD

72. Megf1746

H721003A : epidermal growth factor
GGVCMHIESLDSYTCNCVIGYSGDRCQTRD

73. Megf3446
H721003A : epidermal growth factor
VIGYSGDRCQTRD

MYOGLOBIN (65 sequences)

1. MYAQ
Myoglobin - American alligator
MELSDQEWKHVLDIWTKVESKLPEHGHEVIIRLLQEHPETQERFEKFKHMKTADEMKSSE
KMKQHGNTVFTALGNILKQKGNHAEVLKPLAKSHALEHKIPVKYLEFISEIIVKVIAEKY
PADFGADSQAAMRKALELFRNDMASKYKEFGYQG
2. MYBAO
Myoglobin - Olive baboon, red guenon, and hanuman langur (tentative sequences)
GLSDGEWQLVLNVWGKVEADIPSHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLELISESIIQVLQSKHP
GDFGADAQGAMNKALELFRNDMAAKYKELGFQG
3. MYBD
Myoglobin - Eurasian badger
GLSDGEWQLVLNVWGKVEADLAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKGSED
LKKHGNTVLTALGGILKKKGHQEAELKPLAQSHATKHKIPVKYLEFISDAIAQVLQSKHP
GNFAAEAQGAMKKALELFRNDIAAKYKELGFQG
4. MYBO
Myoglobin - Bovine
GLSDGEWQAVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED
LKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKVPIKYLEFISDAIIHVLHAKHP
SNFAADAQGAMSKALELFRNDAAEKYKVLGFHG
5. MYBTF
Myoglobin - Egyptian rousette
GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLFKFDKFKHLKSEDEMKASED
LKKHGATVLTALGGILKKKGQHEAQLKPLAQSHATKHKIPVKYLEFISEVIIQVLQSKHP
GDFGADAQGAMGKALELFRNDIAAKYKELGFQG
6. MYCA
Myoglobin - Carp
HDAELVLKCWGGVEADFEGTGGEVLTRLFKQHPETQKLFPKFVGIASNELAGNAAVKAHG
ATVLKKLGELLKARGDHAAILKPLATTHANTHKIALNNFRLITEVLVKVMAEKAGLDAGG
QSALRRVMDVVIGDIDTYYKEIGFAG
7. MYCH
Myoglobin (version 1) - Chicken
GLSDQEWQQVLTIWGKVEADIAGHGHEVLMRLFHDHPETLDRFDKFKGLKTEPDMKGSED
LKKHGQTVLTALGAQLKKKGHHEADLKPLAQTHATKHKIPVKYLEFISEVIIKVIAEKHA
ADFGADSQAAMKKALELFRDDMASKYKEFGFQG
8. MYCH2
Myoglobin (version 2) - Chicken (tentative sequence)
GLSDQEWQQVLTIWGKVEADIAGHGHEVLMRLFHDHPETLDRFDKFKGLKTPNEMKGSED
LKKHGATVLTQLGKILKQKGQHESDLKPLAQTHATKHKIPVKYLEFISEVIIKVIAEKHA
ADFGADSQAAMKKALELFRNDMASKYKEFGFQG
9. MYCJ
Myoglobin - Common marmoset (tentative sequence)
GLSDGEWQLVLNVWGKVEADIPSHGQEVLISLFKGHPETLEKFDKFKHLKSEDEMKASEE
LKKHGVTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPVKYLEFISDAIVHVLQKKHP
GDFGADAQGAMKKALELFRNDMAAKYKELGFQG
10. MYCZ
Myoglobin - Chimpanzee (tentative sequence)
GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLHSKHP
GDFGADAQGAMNKALELFRKDMASNYKELGFQG
11. MYDD
Myoglobin - Saddleback dolphin and bottle-nosed dolphin

GLSDGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKFKHLKTEADMKASED
LKKHGNTVLTALGAILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
AEFGADAQGAMNKALELFRKDIAAKYKELGFHG

12. MYDDAR

Myoglobin - Amazon dolphin

GLSDGEWQLVLNIWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKFKHLKTEAEMKASED
LKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
GDFGADAQAAMNKALELFRKDIAAKYKELGFHG

13. MYDDBS

Myoglobin - Saddleback dolphin

GLSDGEWQLVLNVWGKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEADMKASED
LKKHGDTVLTALGAILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
AQFGADAQGAMNKALELFRKDIAAKYKELGFHG

14. MYDE

Myoglobin - Red deer

GLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED
LKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKHP
SNFGADAQGAMSKALELFRNDMAAQYKVLGFQG

15. MYDG

Myoglobin - Dog, bat-eared fox, African hunting dog, and Cape fox

GLSDGEWQIVLNIWGKVETDLAGHGQEVLIRLFKNHPETLDKFDKFKHLKTEDEMKGSED
LKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPVKYLEFISDAIIQVLQSKHS
GDFHADTEAAMKKALELFRNDIAAKYKELGFQG

16. MYELA

Myoglobin - African elephant

GLSDGEWELVLKTWGKVEADIPGHGEFVLVRLFTGHPETLEKFDKFKHLKTEGEMKASED
LKKQGVTVLTALGGILKKKGHHEAEIQPLAQSHATKHKIPIKYLEFISDAIIHVLQSKHP
AEFGADAQAAMKKALELFRNDIAAKYKELGFQG

17. MYELI

Myoglobin - Indian elephant

GLSDGEWELVLKTWGKVEADIPGHGEFVLVRLFTGHPETLEKFDKFKHLKTEGEMKASED
LKKQGVTVLTALGGILKKKGHHEAEIQPLAQSHATKHKIPIKYLEFISDAIIHVLQSKHP
AEFGADAQGAMKKALELFRNDIAAKYKELGFQG

18. MYGC

Myoglobin - Grand galago

GLSDGEWQLVLKIWGKVEADLAGHGQDVLIRLFTAHPETLEKFDKFKNLKTADEMKASED
LKKHGVTVLTALGGILKKKGQHEAEIKPLAQSHATKHKIPVKYLEFISEAIIHVLQNKHS
GDFGTDVQGAMSKALELFRNDIAAKYKELGFQG

19. MYGI

Myoglobin - Agile gibbon (tentative sequence)

GLSDGEWQLVLNVWGKVEADIPSHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHP
GDFGADAQGAMNKALELFRKDMASNYKELGFQG

20. MYGO

Myoglobin - Mountain gorilla (tentative sequence)

GLSDGEWQLVLNVWGKVEADISGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHP
GDFGADAQGAMNKALELFRKDMASNYKELGFQG

21. MYHH

Myoglobin - Western European hedgehog

GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKDHPETLEKFDKFKHLKSEDEMKSSED
LKKHGTTVLTALGGILKKKGQHEAQLAPLAQSHANKHKIPVKYLEFISEAIIQVLKSKHA
GDFGADAQGAMSKALELFRNDIAAKYKELGFQG

22. MYHO

Myoglobin - Horse and common zebra

GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED
LKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKHP
GNFGADAQGAMTKALELFRNDIAAKYKELGFQG

23. MYHU

Myoglobin - Human

GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED

LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHP
GDFGADAQGAMNKALELFRKDMASNYKELGFQG

24. MYKGR

Myoglobin - Red kangaroo

GLSDGEWQLVLNIWGKVETDEGGHGKDVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
LKKHGITVLTALGNILKKKGHHEAELKPLAQSHATKHKIPVQFLEFISDAIIQVIQSKHA
GNFGADAQAAMKKALELFRHDMAAKYKEFGFQG

25. MYKS

Myoglobin - Casiragua

GLSDGEWQLVLNVWGKVEGDLSGHGQEVLIRLFKGHPETLEKFDKFKHLKAEDEMRASEE
LKKHGTTVLTALGGILKKKG AELAPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHP
GDFGADAQGAMSKALELFRNDIAAKYKELGFQG

26. MYLEM

Myoglobin - Weasel lemur

GLSDGEWQLVLNVWGKVEADVGGHGQEVLIRLFTGHPETLEKFDKFKHLKTADEMKASED
LKKHGTTVLTALGGILKKKGQHEAELKPLAQSHATKHKIPIKYLEFISDAIVHVLHSKHP
AEFGADAQAAMKKALELFRNDIAAKYKELGFQG

27. MYLP

Myoglobin - Potto (tentative sequence)

GLSDGEWQSVLNVWGKVEADLAGHGQEILIRLFTAHPETLEKFDKFKNLKTPDEMKASED
LKKHGVTVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISEAIIHVLQSKHP
GDFGADAQGAMNKALELFRNDIAAKYKELGFQG

28. MYLR

Myoglobin - Slow loris (tentative sequence)

GLSDGEWQSVLNVWGKVEADLAGHGQEILIRLFTAHPETLEKFDKFKNLKTPDEMKASED
LKKHGVTVLTALGGILKKKGQHEAEIKPLAQSHATKHKIPVKYLEFISGAIIHVLQSKHP
GDFGADAQGAMSKALELFRNDIAAKYKELGFQG

29. MYLZM

Myoglobin - Lace monitor

GLSDEEWKKVVDIWGKVEPDLPSHGQEVIIRMFQNHPETQDRFAKFKNLKTLDEMKNSED
LKKHGTTVLTALGRILKQKGHHEAEIAPLAQTHANTHKIPIKYLEFICEVIVGVIAEKHS
ADFGADSQEAMRKALELFRNDMASRYKELGFQG

30. MYMQC

Myoglobin - Brown capuchin (tentative sequence)

GLSDGEWQLVLNVWGKVEADIPSHGQEVLISLFKGHPETLEKFDKFKHLKSEDEMKASEE
LKKHGATVLTALGGILKKKGQHEAELKPLAQSHATKHKIPVKYLEFISDAIVHVLQKKHP
GDFGADAQGAMKKALELFRNDMAAKYKELGFQG

31. MYMQIM

Myoglobin - Crab-eating macaque (tentative sequence)

GLSDGEWQLVLNVWGKVEADIPSHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL
KKHGVTVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLELISESIIQVLQSKHPGD
FGADAQGAMNKALELFRNDMAAKYKELGFQG

32. MYMQN

Myoglobin - Night monkey

GLSDGEWQLVLNVWGKVEADVPSHGQEVLISLFKGHPETLEKFDKFKHLKSEDEMKASEE
LKKHGVTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPVKYLEFISDAIVHVLQKKHP
GDFGADAQGAMKKALELFRNDMAAKYKELGFQG

33. MYMQS

Myoglobin - Common squirrel monkey (tentative sequence)

GLSDGEWQLVLNIWGKVEADIPSHGQEVLISLFKGHPETLEKFDKFKHLKSEDEMKASEEL
KKHGTTVLTALGGILKKKGQHEAELKPLAQSHATKHKIPVKYLELISDAIVHVLQKKHPGD
FGADAQGAMKKALELFRNDMAAKYKELGFQG

34. MYMQW

Myoglobin - Common woolly monkey (tentative sequence)

GLSDGEWQLVLNIWGKVEADIPSHGQEVLISLFKGHPETLEKFDKFKHLKSEDEMKASEE
LKKHGVTVLTALGGILKKKGQHEAELKPLAQSHATKHKIPVKYLEFISDAIIHALQKKHP
GDFGADAQGAMKKALELFRNDMAAKYKELGFQG

35. MYMS

Myoglobin - Mouse

GLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSED
LKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIEVKKKRHS

GDFGADAQGAMSKALELFRNDIAAKYKELGFQG

36. MYOG

Myoglobin - Bornean orangutan (tentative sequence)

GLSDGEWQLVLNVWGKVEADIPSHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISESIIQVLQSKHP
GDFGADAQGAMNKALELFRKDMASNYKELGFQG

37. MYOI

Myoglobin - Southern American pika

GLSDGEWQLVLNVWGKVEADLAGHGQEVLIRLFKNHPETLEKFDKFKNLKSEDEMKGSDD
LKKHGNTVLSALGGILKKKGQHEAELKPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHP
GDFGADAQGAMSKALELFRNDMAAKYKELGFQG

38. MYOL

Myoglobin - Mole rat

GLSDGEWQLVLNVWGKVEGDLAGHGQEVLIKLFKNHPETLEKFDKFKHLKSEDEMKGSED
LKKHGNTVLTALGGILKKKGQHAAEIQPLAQSHATKHKIPIKYLEFISEAIIQVLQSKHP
GDFGADAQGAMSKALELFRNDIAAKYKELGFQG

39. MYOP

Myoglobin - North American opossum

GLSDGEWQLVLNAWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
LKKHGATVLTALGNILKKKGNHEAELKPLAQSHATKHKISVQFLEFISEAIIQVIQSKHP
GDFGGDAQAAMGKALELFRNDMAAKYKELGFQG

40. MYOR

Myoglobin - Duckbill platypus

GLSDGEWQLVLKVWGKVEGDLPGHGQEVLIRLFKTHPETLEKFDKFKGLKTEDEMKASAD
LKKHGGTVLTALGNILKKKGQHEAELKPLAQSHATKHKISIKFLEYISEAIIHVLQSKHS
ADFGADAQAAMGKALELFRNDMAAKYKEFGFQG

41. MYOSPW

sperm whale myoglobin

VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED
LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
GDFGADAQGAMNKALELFRKDIAAKYKELGYQG

42. MYOY

Myoglobin - Aardvark

GLSDAEWQLVLNVWGKVEADIPGHGQDVLIRLFKGHPETLEKFDRFKHLKTEDEMKASED
LKKHGTTVLTALGGILKKKGQHEAEIQPLAQSHATKHKIPVKYLEFISEAIIQVIQSKHS
GDFGADAQGAMSKALELFRNDIAAKYKELGFQG

43. MYPE

Myoglobin - Harbor porpoise and Dall's porpoise

GLSEGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKFKHLKTEAEMKASED
LKKHGNTVLTALGGILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
AEFGADAQGAMNKALELFRKDIATKYKELGFHG

44. MYPG

Myoglobin - Pig

GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
LKKHGNTVLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHP
GDFGADAQGAMSKALELFRNDMAAKYKELGFQG

45. MYPN

Myoglobin - Emperor penguin

GLNDQEWQQVLTMWGKVESDLAGHGHAVLMRLFKSHPETMDRFDKFRGLKTPDEMRGSED
MKKHGVTVLTLGQILKKKGHHEAELKPLSQTHATKHKVPVKYLEFISEAIMKVIAQKHAS
NFGADAQEAMKKALELFRNDMASKYKEFGFQG

46. MYRB

Myoglobin - Rabbit

GLSDAEWQLVLNVWGKVEADLAGHGQEVLIRLFHTHPETLEKFDKFKHLKSEDEMKASED
LKKHGNTVLTALGAILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISEAIIHVLHSKHP
GDFGADAQAAMSKALELFRNDIAAQYKELGFQG

47. MYRKJ

Myoglobin - Port Jackson shark

TEWEHVNKVWAVVEPDIPAVGLAILLRLFKEHKETKDLFPKFKEIPVQQLGNNEDLRKHG
VTVLRALGNILKQKGKHSTNVKELADTHINKHKIPPKNFVLITNIAVKVLTEMYPSDMTG
PMQESFSKVFTVICSDLETLYKEANFQG

48. MYSH
Myoglobin - Sheep
GLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED
LKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKHP
SNFGADAQGAMSKALELFRNDMAAEYKVLGFQG
49. MYSLG
Myoglobin - Gray seal and harbor seal
GLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKFKHLKSEDDMRRSED
LRKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSKHP
AEFGADAQAAMKKALELFRNDIAAKYKELGFHG
50. MYTG
Myoglobin - Australian echidna
GLSDGEWQLVLKVWGKVETDITGHGQDVLIRLFKTHPETLEKFDKFKHLKTEDEMKASAD
LKKHGGVVLTALGSILKKKGQHEAELKPLAQSHATKHKISIKFLEFISEAIIHVLQSKHS
ADFGADAQAAMGKALELFRNDMATKYKEFGFQG
51. MYTS
Myoglobin - Common tree shrew
GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKTEDEMKASED
LKKHGNTVLSALGGILKKKGQHEAEIKPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHP
GDFGADAQAAMSKALELFRNDIAAKYKELGFQG
52. MYTTG
Myoglobin - Green seaturtle
GLSDDEWNHVLGIWAKVEPDLTAHGQEVIIRLFQLHPETQERFAKFKNLTTIDALKSSEE
VKKHGTTVLTALGRILKQKNNHEQELKPLAESHATKHKIPVKYLEFICEIIVKVIAEKHP
SDFGADSQAAMKKALELFRNDMASKYKEFGFLG
53. MYTTM
Myoglobin - Map turtle
GLSDDEWHHVLGIWAKVEPDLSAHGQEVIIRLFQVHPETQERFAKFKNLKTIDELRSSEE
VKKHGTTVLTALGRILKLKNNHEPELKPLAESHATKHKIPVKYLEFICEIIVKVIAEKHP
SDFGADSQAAMRKALELFRNDMASKYKEFGFQG
54. MYTUY
Myoglobin - Yellowfin tuna
ADFDAVLKCWGPVEADYTTMGGLVLTRLFKEHPETQKLFPKFAGIAQADIAGNAAISAHG
ATVLKKLGELLKAKGSHAAILKPLANSHATKHKIPINNFKLISEVLVKVMHEKAGLDAGG
QTALRNVMGIIIADLEANYKELGFSG
55. MYVJ
Myoglobin - Viscacha
GLSDGEWQLVLNVWGKVEADLGGHGQEVLIRLFKGHPETLEKFDKFKHLKAEDEMRASED
LKKHGTTVLTALGGILKKRGQHAAELAPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHP
GDFGADAQAAMSKALELFRNDIAAKYKELGFQG
56. MYWHC
Myoglobin - California gray whale
VLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASED
LKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRHP
GDFGADAQAAMNKALELFRKDIAAKYKELGFQG
57. MYWHF
Myoglobin - Finback whale
VLTDAEWHLVLNIWAKVEADVAGHGQDILISLFKGHPETLEKFDKFKHLKTEAEMKASED
LKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRHP
ADFGADAQAAMNKALELFRKDIAAKYKELGFQG
58. MYWHH
Myoglobin - Humpback whale
VLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASED
LKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRHP
ADFGADAQAAMNKALELFRKDIAAKYKELGFQG
59. MYWHK
Myoglobin - Minke whale
VLSDAEWHLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASED
LKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRHP
AEFGADAQAAMNKALELFRKDIAAKYKELGFQG

60. MYWHL
Myoglobin - Killer whale
GLSDGEWQLVLNVWGKVEADLAGHGQDILIRLFKGHPETLEKFDKFKHLKTEADMKASED
LKKHGNTVLTALGAILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
AEFGADAQGAMNKALELFRKDIAAKYKELGFHG
61. MYWHP
Myoglobin - Sperm whale and dwarf sperm whale
VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED
LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
GDFGADAQGAMNKALELFRKDIAAKYKELGYQG
62. MYWHT
Myoglobin - Pilot whale
GLSDGEWQLVLNVWGKVEADLAGHGQDILIRLFKGHPETLEKFDKFKHLKTEADMKASED
LKKHGNTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
AEFGADAQGAMNKALELFRKDIAAKYKELGFHG
63. MYWHU
Myoglobin - Hubbs' whale
GLSEAEWQLVLHVWAKVEADLSGHGQEILIRLFKGHPETLEKFDKFKHLKSEAEMKASED
LKKHGHTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKHP
SDFGADAQGAMTKALELFRKDIAAKYKELGFHG
64. MYWHZ
Myoglobin - Goose-beaked whale
GLSEAEWQLVLHVWAKVEADLSGHGQEILIRLFKGHPETLEKFDKFKHLKSEAEMKASED
LKKHGHTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRHP
SDFGADAQAAMTKALELFRKDIAAKYKELGFHG
65. MYZC
Myoglobin - California sealion
GLSDGEWQLVLNIWGKVEADLVGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKRSED
LKKHGKTVLTALGGILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLQSKHP
GDFGADTHAAMKKALELFRNDIAAKYRELGFQG

TGF (3 sequences)

1. TGFAHUM
TGF alpha human
VVSHFNDCPDSHTQFCFHGTCRFLVQEDKPACVCHSGYVGARCEHADLLA

2. TGFARAT
TGF alpha RAT
VVSHFNKCPDSHTQYCFHGTCRFLVQEDKPACVCHSGYVGVRCEHADLLA

3. TGFBPIG
TGFB$PIG : TRANSFORMING GROWTH FACTOR BETA PRECURSOR (TGF-BETA) (TGF
TYPE 2). 3
MPPSGLRLLPLLLPLLWLLVLTPGRPAAGLSTCKTIDMELVKRKRIEAIRGQILSKLRLA
SPPSQGDVPPGPLPEAVLALYNSTRDRVAGESVEPEPEPEADYYAKEVTRVLMLESGNQI
YDKFKGTPHSLYMLFNTSELREAVPEPVLLSRAELRLLRLKLKVEQHVELYQKYSNDSWR
YLSNRLLAPSDSPEWLSFDVTGVVRQWLTRREAIEGFRLSAHCSCDSKDNTLHVEINGFN
SGRRGDLATIHGMNRPFLLLMATPLERAQHLHSSRHRRALDTNYCFSSTEKNCCVRQLYI
DFRKDLGWKWIHEPKGYHANFCLGPCPYIWSLDTQYSKVLALYNQHNPGASAAPCCVPQA
LEPLPIVYYVGRKPKVEQLSNMIVRSCKCS

TNF (9 sequences)

1. TNF
TUMOR NECROSIS FACTOR (IBID)
VRSSSRTPSDKPVAHVVANPQAEGQLQWLNRRANALLANGVELRDNQLVV

PSEGLYLIYSQVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETPEG
AEAKPWYEPIYLGGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGAFAL

2. TNFBBOV

TNF BETA BOVINE
LRGIGLTPSAAQPAHQQLPTPFTRGTLKPAAHLVGDPSTQDSLRWRANTDRAFLRHGF
SLSNNSLLVPTSGLYFVYSQVVFSGRGCFPRATPTPLYLAHEVQLFSPQYPFHVPLLS
AQKSVCPGPQGPWVRSVYQGAVFLLTRGDQLSTHTDGISHLLLSPSSVFFGAFAL

3. TNFBMOUS

TNF BETA MOUSE
LSGVRFSAARTAHPLPQKHLTHGILKPAAHLVGYPSKQNSLLWRASTDRAFLRHGFSL
SNNSLLIPTSGLYFVYSQVVFSGESCSPRAIPTPIYLAHEVQLFSSQYPFHVPLLSAQ
KSVYPGLQGPWVRSMYQGAVFLLSKGDQLSTHTDGISHLHFSPSSVFFGAFAL

4. TNFBOV

TUMOR NECROSIS FACTOR BOVINE (SA PAPIRA)
LRSSSQASSNKPVAHVVADINSPGQLRWWDSYANALMANGVKLEDNQLVVPAEGLYLI
YSQVLFRGQGCPPPPVLTHTISRIAVSYQTKVNILSAIKSPCHRETPEWAEAKPWYEP
IYQGGVFQLEKGDRLSAEINLPDYLDYAESGQVYFGIIAL

5. TNFBHUM

limpho toxin (Nature, 312,(1984/5) 720
LPGVGLTPSAAQTARQHPKMHLAHSTLKPAAHLIGDPSKQNSLLWRANTDRAF
LQDGF

CYSTEINE PROTEASES (23 sequences)

1. ACTNCHN1

Actinidin (EC 3.4.22.14) - Yangtao
LPSYVDWRSAGAVVDIKSQGECGGCWAFSAIATVEGINKITSGSLISLSEQELIDCGRTQ
NTRGCDGGYITDGFQFIINDGGINTZENYPYTAQDGDCDVALQDQKYVTIDTYENVPYNN
EWALQTAVTYQPVSVALDAAGDAFKQYASGIFTGPCGTAVDHAIVIVGYGTEGGVDYWIV
KNSWDTTWGEEGYMRILRNVGGAGTCGIATMPSYPVKYNN

2. ACTNKIW2

Actinidin -kiwi actin5pas
LPSYVDWRSAGAVVDIKSQGECGGCWAFSAIATVEGINKITSGSLISLSEQELIDCGRTQ
NTRGCDGGYITDGFQFIINDGGINTEENYPYTAQDGDCDVALQDQKYVTIDTYENVPYNN
EWALQTAVTYQPVSVALDAAGDAFKQYASGIFTGPCGTAVDHAIVIVGYGTEGGVDYWIV
KNSWDTTWGEEGYMRILRNVGGAGTCGIATMPSYPVKY

3. CATHCOW2

Z1409299A : cathepsin B
LPESFDAREQWPNCPTIKEIRDQGSCGSCWAFGAVEAISDRICIHSNGRVNVEVSAEDML
TCCGGECGDGCNGGEPSGAWNFWTKKGLVSGGLYNSHVGCRPYSIPPCEHHVNGSRPPCT
GEGDTPKCNKTCEPGYSPSYKEDKHFGCSSYSVANNEKEIMAEIYKNGPVEGAFSVYSDF
LLYKSGVYQHVSGEIMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQDHC
GIESEIVAGMPCT

4. CATHHEN1

Z1311264A : cathepsin L H
APRSVDWREKGYVTPVKDQGQCGSCWAFSTTGALEGQHFRKTGKLVSLSEQNLVDCSRPE
GNQGCNGGLMDQAFQYVQDNGGIDSEESYPYTAKDDEDCRYKAEYNAANDTGFVDIPQGH
ERALMKAVASVGPVSVAIDAGHSSFQFYQSGIYYEPDCSSEDLDHGVLVVGYGFEG

5. CATHHEN3

Z1312246A : cathepsin L
APRSVDWREKGYVTPVKDQGICGSCWAFNTTGALEGQHFFKTGKLVSLSEQNLVDCSRPE
GNQGCNGGLMDQAFQYVQDNGGIDSEESYPYTAKDDCRYKAEYNAAKDTGFVDIPQGHER
ALMKAVASVGPVSVAIDAGHSSFQFYQSGIYYEPDCSSEDLDHGVLVVGYGFEGKKYWIV
KNSWGEKWGDKGYQYMAKDRKNHCGIATAASYPLV

6. CATHHM13

Cathepsin B (EC 3.4.22.1) - Human
LPASFDAREQWPQCPTIKEIRDQGSCGSCWAFGAVEAISDRICIHTNVSVEVSAEDLLTC
CGSMCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCEHHVNGSRPPCTGE

GDTPKCSKICEPGYSPTYKQDKHYGYDSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLL
YKSGVYQHVTGEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQDHCGI
ESEVVAGIPRTD

7. CATHHM14

CATH$HUMAN: CATHEPSIN H PRECURSOR (EC 3.4.22.16). 335 AA.
YPPSVDWRKKGNFVSPVKNQGACGSCWTFSTTGALESAIAIATGKMLSLAEQQLVDCAQD
FNNYGCQGGLPSQAFEYILYNKGIMGEDTYPYQGKDGYCKFQPGKAIGFVKDVANITIYD
EEAMVEAVALYNPVSFAFEVTQDFMMYRTGIYSSTSCHKTPDKVNHAVLAVGYGEKNGIP
YWIVKNSWGPQWGMNGYFLIERGKNMCGLAACASYPIPLV

8. CATHHM2

Z1205366A : cathepsin B
VEVSAEDLLTCCGSMCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCEHH
VNGSRPPCTGEGDTPKCSKICEPGYSPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVE
GAFSVYSDFLLYKSGVYQHVTGEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFF
KILRGQDHCGIESEVVAGIPRTDQYWEKI

9. CATHHM4

Z1306235A : cathepsin G
IGGRESRPHSRPYMAYLQIQSPAGQSRCGGFLVREDFVLTAAHCWGSNINVTLGAHNIQR
RENTQQHITARRAIRHPQYNQRTIQNDIMLLQLSRRVRRNRNVNPVALPRAQEGLRPGTL
CTVAGWGRVSMRRGTDTLREVQLRVQRDRQCLRIFGSYDPRRQICVGDRRERKAAFKGDS
GGPLLCNNVAHGIVSYGKSSGVPPEVFTRVSSFLPWIRTTMRSFKLLDQMETPL

10. CATHHM5

Z1307261A : cathepsin D
GPIPEVLKNYMDAQYYGEIGIGTPPQCFTVVFDTGSSNLWVPSIHCKLLDIACWIHHKYN
SDKSSTYVKNGTSFDIHYGSGSLSGYLSQDTVSVPCQSASSASALGGVKVERQVFGEATK
QPGITFIAAKFDGILGMAYPRISVNNVLPVFDNLMQQKLVDQNIFSFYLSRDPDAQPGGE
LMLGGTDSKYYKGSLSYLNVTRKAYWQVHLDQVEVASGLTLCKEGCEAIVDTGTSLMVGP
VDEVRELQKAIGAVPLIQGEYMIPCEKVSTLPAITLKLGGKGYKLSPEDYTLKVSQAGKT
LCLSGFMGMDIPPPSGPLWILGDVFIGRYYTVFDRDNNRVGFAEAARL

11. CATHHM6

Z1404279A : Cathepsin H
CPYPPSVDWRKKGNFVSPVKNQGACGSCWTFSTTGALESAIAIATGKMLSLAEQQLVDCA
QDFNNHGCQGGLPSQAFEYILYNKGIMGEDTYPYQGKDGYCKFQPGKAIGFVKDVANITI
YDEEAMVEAVALYNPVSFAFEVTQDFMMYRTGIYSSTSCHKTPDKVNHAVLAVGYGEKNG
IPYWIVKNSWGPEWGMNGYFLIERGKNMCGLAACASYPIP

12. CATHHM7

Z1404279B : cathepsin L
APRSVDWREKGYVTPVKNQGQCGSCWAFSATGALPGQMFRKTGRLISLSEQNLVDCSGPQ
GNEGCNGGLMDYAFQYVQDNGGLDSEESYPYEATEESCKYNPKYSVAXDTGFVDIPKQEK
ALMKAVATVGPISVAIDAGHESFLFYKEGIYFEPNCSSEDMDHGVLVVGYGFESTNNKYW
LVKNSWGEEWGMGGYVKMAKDRRNHCGIASAASYPTV

13. CATHMSE1

Z1211378B : cathepsin B
LPETFDAREQWSNCPTIGQIRDQGSCGSCWAFGAVEAISDRTCIHTNGRVNVEVSAEDLL
TCCGIQCGDGCNGGYPSGAWNFWTKKGLVSGGVYDSHIGCLPYTIPPCEHHVNGSRPPCT
GEGDTPRCNKSCEAGYSPSYKEDKHFGYTSYSVSNSVKEIMAEIYKNGPVEGAFTVFSDF
LTYKSGVYKHEAGDMMGGHAIRILVWGVENGVPYWLAANSWNLDWGDNGFFKILRGENHC
GIESEIVAGIPRTDQYWGRF

14. CATHMSE2

Z1407219A : cathepsin L
MNLLLLLAVLCLGTALATPKFDQTFSAEWHQWKSTHRRLYGTNEEEWRRAIWEKNMRIIQ
LHNGEYSNGQHGFSMEMNAFGDMTNEEFRQVVNGYRHQKHKKGRLFQEPLMLKIPKSVDW
REKGCVTPVKNQGQCGSCWAFSASGCLEGQMFLKTGKLISLSEQNLVDCSHAQGNQGCNG
GLMDFAFQYIKENGGLDSEESYPYEAKDGSCKYRAEFAVANDTGFVDIPQQEKALMKAVA
TVGPISVAMDASHPSLQFYSSGIYYEPNCSSKNLDHGVLLVGYGYEGTDSNKNKYWLVKN
SWGSEWGMEGYIKIAKDRDNHCGLATAASYPVVN

15. CATHPIG1

Z1007243A : cathepsin D
GPIPEVLKNYMDAQYYGEIGIGTPPQCFTVVFDTGSSNLWVPSIHCKLLDIACWIHHKYN
SGKSSTYVKNGTTFAIHYGSGSLSGYLSQDTVSVPSNVGGIKVERQTFGEATKQPGLTFI
AAKFDGILGMAYPRISVNNVVPVFDNLMQQKLVDKDIFSFYLNRDPGAQPGGELMLGGID

SKYYKGSLDYHNVTRKAYWQIHMNQVAVGSSLTLCKGGCEAIVDTGTSLIVGQPEEVREL
GKAIGAVPLIQGEYMIPCEKVPSLPDVTVTLGGKKYKLSSENYTLKVSQAGQTICLSGFM
GMDIPPPGGPLWILGDVFIGRYYTVFDRDLNRVGLAEAA

16. CATHRAT1

Z1313245A : cathepsin L

MTPLLLLAVLCLGTALATPKFDQTFNAQWHQWKSTHRRLYGTNEEEWRRAVWEKNMRMIQ
LHNGEYSNGKHGFTMEMNAFGDMTNEEFRQIVNGYRHQKHKKGRLFQEPLMLQIPKTVDW
REKGCVTPVKNQGQCGSCWAFSASGCLEGQMFLKTGKLISLSEQNLVDCSHDQGNQGCNG
GLMDFAFQYIKENGGLDSEESYPYEAKDGSCKYRAEYAVANDTGFVDIPQQEKALMKPVA
TVGPISVAMDASHPSLQFYSSGIYYEPNCSSKDLDHGVLVVGYGYEGTDSNKDKYWLVKN
SWGKEWGMDGYIKIAKDRNNHCGLATAASYPIVN

17. CATHRAT2

Z1402240A : cathepsin H

YPSSMDWRKKGNVVSPVKNQGACGSCWTFSTTGALESAVAIASGKMMTLAEQQLVDCAQN
FNNHGCQGGLPSQAFEYILYNKGIMGEDSYPYIGKNGQCKFNPEKAVAFVKNVVNITLND
EAAMVEAVALYNPVSFAFEVTEDFMMYKSGVYSSNSCHKTPDKVNHAVLAVGYGEQNGLL
YWIVKNSWGSNWGNNGYFLIERGKNMCGLAACASYPIPQV

18. CATHRAT4

Cathepsin B (EC 3.4.22.1) - Rat

LPESFDAREQWSNCPTIAQIRDQGSCGSCWAFGAVEAMSDRICIHTNVNVEVSAEDLLTC
CGIQCGDGCNGGYPSGAGNFWTRKGLVSGGVYNSHIGCLPYTIPPCEHHVNGSRPPCTGE
GDTPKCNKMCEAGYSTSYKEDKHYGYTSYSVSDSEKEIMAEIYKNGPVEGAFTVFSDFLT
YKSGVYKHEAGDVMGGHAIRILGWGIENGVPYWLVANSWNVDWGDNGFFKILRGENHCGI
ESEIVAGIPRTQ

19. PAPAPAP1

Papain - papaya omega

LPENYDWRKKGAVTPVRHQGSSGSCWAFSAVATVEGINKIRTGKLVELSEQELVDCERRS
HGCKGGYPPYALEYVAKNGIHLRSKYPYKAKQGTCRAKQVGGPIVKTSGVGRVQPNNEGN
LLNAIAKQPVSVVVESKGRPFQLYKGGIFEGPCGTKVDHAVTAVGYGKSGGKGYILIKNS
WGTAWGEKGYIRIKRAPGNSPGVCGLYKSSYYPTKN

20. PAPAPAP2

Papain (EC 3.4.22.2) - Papaya

IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDRRS
YGCNGGYPWSALQLVAQYGIHYRNTPYYEGVQRYCRSREKGPYAAKTDGVRQVQPYNQGA
LLYSIANQPVSVVLQAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYNPGYILIKNSWGTG
WGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN

21. PAPAPAP3

Papain - chymopapain jbc 1989

YPQSIDWRAKGAVTPVKNQGACGSCWAFSTIATVEGINKIVTGNLLELSEQELVDCDKHS
YGCKGGYQTTSLQYVANNGVHTSKVYPYQAKQYKCRATDKPGPKVKITGYKRVPSNCETS
FLGALANQPLSVLVEAGGKPFQLYKSGVFDGPCGTKLDHAVTAVGYGTSDGKNYIIIKNS
WGPNWGEKGYMRLKRQSGNSQGTCGVYKSSYYPFKGFA

22. PAPAPAP4

Papain - ppiv febl 1989

LPESVDWRAKGAVTPVKHQGYCESCWAFSTVATVEGINKIKTGNLVELSEQELVDCDLQS
YGCNRGYQSTSLQYVAQNGIHLRAKYPYIAKQQTCRANQVGGPKVKTNGVGRVQSNNEGS
LLNAIAHQPVSVVVESAGRDFQNYKGGIFEGSCGTKVDHAVTAVGYGKSGGKGYILIKNS
WGPGWGENGYIRIRRASGNSPGVCGVYRSSYYPIKN

23. PAPAPAP5

Papain - Papaya papain5pas

IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDRRS
YGCNGGYPWSALQLVAQYGIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYNQGA
LLYSIANQPVSVVLQAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYGPNYILIKNSWGTG
WGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN

METALLO PROTEASES (17 sequences)

1. CARBBAC1

DACA$BACSU: PENICILLIN-BINDING PROTEIN 5 (D-ALANYL-D-ALANINE
CARBOXYPEPTIDASE) 3

ASSGKILYSKNADKRLPIASMTKMMTEYLLLEAIDQGKVKWDQTYTPDDYVYEISQDNSL
SNVPLRKDGKYTVKELYQATAIYSANAAAIAIAEIVAGSETKFVEKMNAKAKELGLTDYK
FVNATGLENKDLHGHQPEGTSVNEESEVSAKDMAVLADHLITDYPEILETSSIAKTKFRQ
GTDDEMDMPNWNFMLKGLVSEYKKATVDGLKTGSTDSAGSCFTGTAERNGMRVITVVLNA
KGNLHTGRFDETKKMFDYAFDNFSMKEIYAEGDQVKGHKTISVDKGKEKEVGIVTNKAFS
LPVKNGEEKNYKAKVTLNKDNLTAPVKKGTKVGKLTAEYTGDEKDYGFLNSDLAGVDLVT
KENVEKANWFVLTMRSIGGFFAGIWGSIVDTVTGWF

2. CARBBAR1

Serine carboxypeptidase (EC 3.4.16.1) - Barley

APQGAEVTGLPGFDGALPSKHYAGYVTVDEGHGRNLFYYVVESERDPGKDPVVLWLNGGP
GCSSFDGFVYEPGPFNFESGGSVKSLPKLHLNPYAWSKVSTMIYLDSPAGVGLSYSKNVS
DYETGDLKTATDSHTFLLKWFQLYPEFLSNPFYIAGESYAGVYVPTLSHEVVKGIQGGAK
PTINFKGYMVGNGVCDTIFDGNALVPFAHGMGLISDEIYQQASTSCHGNYWNATDGKCDT
AISKIESLISGLNIYDILEPCYHSRSGVPCMSDEVATAWLDNAAVRSAIHAQSVSAIGPW
LLCTDKLYFVHDAGSMIAYHKNLTSQGYRAIIFSGDHDMXVPFTGSEAWTKSLGYGVVDS
WRPWITNGQVSGYTEGYEHGLTFATIKGAGHTVPEYKPQEAFAFYSRWLAGSKL

3. CARBCOW1

Carboxypeptidase E - Bovine

RPQEDGISFEYHRYPELREALVSVWLQCAAVSRIYTVGRSFEGRELLVLELSDNPGVHEP
GEPEFKYIGNMHGNEAVGRELLIFLAQYLCNEYQKGNETIVQLIHNTRIHIMPSLNPDGF
EKAASQLGELKDWFVGRSNAQGIDLNRNFPDLDRIVYINEKEGGPNNHLLKNLKKIVDQN
TKLAPETKAVIHWIMDIPFVLSANLHGGDLVANYPYDETRSGSAHEYSSCPDDDIFQSLA
RAYSSFNPPMSDPDRPPCRKNDDDSSFVEGTTNGAAWYSVPGGMQDFNYLSSNCFEITVE
LSCEKFPPEETLKNYWEDNKNSLISYIQQIHRGVKGFVRDLQGNPIANATLSVEGIDHDV
TSAKDGDYWRLLVPGNYKLTASAPGYLAIAKKVAVPYSPAVRVDFELESFSERKEEEKEE
LMEWWKMMSETLNF

4. CARBCOW2

Carboxypeptidase A (EC 3.4.17.1) - Bovine

ARSTNTFNYATYHTLDEIYDFMDLLVAEHPQLVSKLQIGRSYEGRPIYVLKFSTGGSNRP
AIWIDLGIHSREWITQATGVWFAKKFTENYGQNPSFTAILDSMDIFLEIVTNPNGFAFTH
SENRLWRKTRSVTSSSLCVGVDANRNWDAGFGKAGASSSPCSETYHGKYANSEVEVKSIV
DFVKNHGNFKAFLSIHSYSQLLLYPYGYTYQSIPDKTELNQVAKSAVAALKSLYGTSYKY
GSIITTIYQASGGSIDWSYNQGIKYSFTFELRDTGRYGFLLPASQIIPTAQETWLGVLTI
MEHTVNN

5. CARBCOW3

Carboxypeptidase B (EC 3.4.17.2) - Bovine

TTGHSYEKYNNWETIEAWTEQVASENPDLISRSAIGTTFLGNTIYLLKVGKPGSNKPAVF
MDCGFHAREWISPAFCQWFVREAVRTYGREIHMTEFLDKLDFYVLPVVNIDGYIYTWTTN
RMWRKTRSTRAGSSCTGTDLNRNFDAGWCSIGASNNPCSETYCGSAAESEKESKAVADFI
RNHLSSIKAYLTIHSYSQMMLYPYSYDYKLPKNNVELNTLAKGAVKKLASLHGTTYSYGP
GATTIYPASGGSDDWAYDQGIKYSFTFELRDKGRYGFVLPESQIQPTCEETMLAIKYVTS
YVLEHL

6. CARBCOW4

Procarboxypeptidase A complex component III - Bovine

DGEDAVPYSWSWQVSLQYEKDGAFHHTCGGSLIAPDWVVTAGHCISTSRTYQVVLGEYDR
SVLEGSEQVIPINAGDLFVHPLWNSNCVACGNDIALVKLSRSAQLGDKVQLANLPPAGDI
LPNEAPCYISGWGRLYTGGPLPDKLQQALLPVVDYEHCSQWDWWGITVKKTMVCAGGDTR
SGCNGDSGGPLNCPAADGSWQVHGVTSFVSAFGCNTIKKPTVFTRVSAFIDWIDETIASN

7. CARBCRY1

Carboxypeptidase B (EC 3.4.17.2) - Crayfish

MDWTSYHDYDEINAWLDSLATDYPELASVEDVGLSYEGRTMKLLKLGKGGADKPIIFIDG
GIHAREWIAPSTVTYIVNEFVSNSATYDDILSNVNFYVMPTINPDGYAYTFTDDRLWRKT
RSETGSVLGCKGADPNRNWSFHWDEVGASDSPCSDIYAGPEPFSEVEMRNVRDQILEYAA
NIKVYLTFHSYSQLWMYPWGFTSDLPDDWQDLDTLATNAVDALTAVHGTRYEIGSSTNTI
YAAAGGSDDWAKGEGGVKYAYTIELRDTGNYGFLLPENQIIPTGEETFEGVKVVANFVKD
TYS

8. CARBPOT1

CARBOXYPEPTIDASE AALPHA (COX) (E.C.3.4.17.1) COMPLEX WITH POTATO
CARBOXYPEPTIDASARSTNTFNYATYHTLDEIYDFMDLLVAQHPELVSKLQIGRSYEGR
PIYVLKFSTGGSNRPAIWIDLGIHSREWITQATGVWFAKKFTENYGQNPSFTAILDSMDIFL
EIVTNPNGFAFTHSENRLWRKTRSVTSSSLCVGVDANRNWDAGFGKAGASSSPCSETYHGKY

ANSEVEVKSIVDFVKNHGNFKAFLSIHSYSQLLLYPYGYTTQSIPDKTELNQVAKSAVAA
LKSLYGTSYKYGSIITTIYQASGGSIDWSYNQGIKYSFTFELRDTGRYGFLLPASQIIPT
AQETWLGVLTIMEHTVNN

9. CARBPSD1

Folate hydrolase G2 (EC 3.4.-.-) precursor - Pseudomonas sp.
ALAQKRDNVLFQAATDEQPAVIKTLEKLVNIETGTGDAEGIAAAGNFLEAELKNLGFTVT
RSKSAGLVVGDNIVGKIKGRGGKNLLLMSHMDTVYLKGILAKAPFRVEGDKAYGPGIADD
KGGNAVILHTLKLLKEYGVRDYGTITVLFNTDEEKGSFGSRDLIQEEAKLADYVLSFEPT
SAGDEKLSLGTSGIAYVQVNITGKASHAGAAPELGVNALVEASDLVLRTMNIDDKAKNLR
FNWTIAKAGNVSNIIPASATLNADVRYARNEDFDAAMKTLEERAQQKKLPEADVKVIVTR
GRPAFNAGEGGKKLVDKAVAYYKEAGGTLGVEERTGGGTDAAYAALSGKPVIESLGLPGF
GYHSDKAEYVDISAIPRRLYMAARLIMDLGAGK

10. CARBRAT1

E1501328A : carboxypeptidase B
MLLLLALVSVALAHASEEHFDGNRVYRVSVHGEDHVNLIQELANTKEIDFWKPDSATQVK
PLTTVDFHVKAEDVADVENFLEENEVHYEVLISNVRNALESQFDSHTRASGHSYTKYNKW
ETIEAWIQQVATDNPDLVTQSVIGTTFEGRNMYVLKIGKTRPNKPAIFIDCGFHAREWIS
PAFCQWFVREAVRTYNQEIHMKQLLDELDFYVLPVVNIDGYVYTWTKDRMWRKTRSTMAG
SSCLGVRPNRNFNAGWCEVGASRSPCSETYCGPAPESEKETKALADFIRNNLSTIKAYLT
IHSYSQMMLYPYSYDYKLPENYEELNALVKGAAKELATLHGTKYTYGPGATTIYPAAGGS
DDWSYDQGIKYSFTFELRDTGFFGFLLPESQIRQTCEETMLAVKYIANYVREHLY

11. CARBRAT2

E1501328B : carboxypeptidase A1
ALSTDSFNYATYHTLDEIYEFMDLLVAEHPQLVSKIQIGNTFEGRPIHVLKFSTGGTNRP
AIWIDTGIHSREWVTQASGVWFAKKITKDYGQDPTFTAVLDNMDIFLEIVTNPDGFAYTH
KTNRMWRKTRSHTQGSLCVGVDPNRNWDAGFGMAGASSNPCSETYRGKFPNSEVEVKSIV
DFVTSHGNIKAFISIHSYSQLLLYPYGYTSEPAPDQAELDQLAKSAVTALTSLHGTKFKY
GSIIDTIYQASGSTIDWTYSQGIKYSFTFELRDTGLRGFLLPASQIIPTAEETWLALLTI
MDHTVKHPY

12. CARBSTP1

Muramoyl-pentapeptide carboxypeptidase (EC 3.4.17.8) - Streptomyces
griseus solv
NGCYTWSGTLSEGSSGEAVRQLQIRVAGYPGTGAQLAIDGQFGPATKAAVQRFQSAYGLA
ADGIAGPTFNKIYQLQDDDCTPVNFTYAELNRCNSDWSGGKVSAATARANALVTMWKLQA
MRHAMGDKPITVNGGFRSVTCNSNVGGASNSRHMYGHAADLGAGSQGFCALAQAARNHGF
TEILGPGYPGHNDHTHVAGGDGRFWSAPSCGI

13. CARBWHT1

Carboxypeptidase Y (EC 3.4.16.1) homolog - Wheat
MATTPRLASLLLLLALCAAAAGALRLPPDASFPGAQAERLIRALNLLPGRPRRGLGAGAE
DVAPGQLLERRVTLPGLPEGVGDLGHHAGYYRLPNTHDARMFYFFFESRGKKEDPVVIWL
TGGPGCSSELAVFYENGPFTIANNMSLVWNKFGWDKISNIIFVDPATGTGFSYSSDDRDT
RHDEAGVSNDLYDFLQVFFKKHPEFVKNDFFITGESYAGHYIPAFASRVHQGNKKNEGTH
INLKGFAIGNGLTDPAIQYKAYTDYALDMNLIQKADYDRINKFIPPCEFAIKLCGTDGKA
SCMAAYMVCNSIFNSIMKLVGTKNYYDVRKECEGKLCYDFSNLEKFFGDKAVRQAIGVGD
IEFVSCSTSVYQAMLTDWMRNLEVGIPALLEDGINVLIYAGEYDLICNWLGNSRWVHSME
WSGQKDFAKTAESSFLVDDAQAGVLKSHGALSFLKVHNAGHMVPMDQPKAALEMLRRFTQ
GKLKESVPEEEPATTFYAA

14. CARBYST3

E70003: CARBOXYPEPTIDASE Y EC 3.4.16.1
KIKDPKILGIDPNVTQYTGYLDVEDEDKHFFFWTFESRNDPAKDPVILWLNGGPGCSSLT
GLFFELGPSSIGPDLKPIGNPYSWNSNATVIFLDQPVNVGFSYSGSSGVSNTVAAGKDVY
NFLELFFDXFPEYVNKGQDFHIAGESYAHGYIPVFAXEILSHKDRNFNLTSVLIGNGLTD
PLTQYNYYEPMACGEGGEPSVLPSEECSAMEDSLERCLGLIESCYDSQSVWSCVPATIYC
NNAQLAPYQRTGRNVYDIRKDCEGGNLCYPTLQDIDDELNQDYVKEAVGAEVDHYESCNF
DINRNFLFAGDWMKPYHTAVTDLLNQDLPILVYAGDKDFINNTLGNKAWTDVLPWKYDEE
FASQKVRCWTASITDEVAGEVKSYKHFTYLRVFNGGHMVPFDVPENALSMVNEWIHGDFS
L

15. THERBAC1

Neutral proteinase (EC 3.4.24.4) - Bacillus cereus
VTGTNKVGTGKGVLGDTKSLNTTLSGSSYYLQDNTRGATIFTYDAKNRSTLPGTLWADAD
NVFNAAYDAAAVDAHYYAGKTYDYYKATFNRNSINDAGAPLKSTVHYGSNYNNAFWNGSQ

MVYGDGDGVTFTSLSGGIDVIGHELTHAVTENSSNLIYQNESGALNEAISDIFGTLVEFY
DNRNPDWEIGEDIYTPGKAGDALRSMSDPTKYGDPDHYSKRYTGSSDNGGVHTNSGIINK
QAYLLANGGTHYGVTVTGIGKDKLGAIYYRANTQYFTQSTTFSQARAGAVQAAADLYGAN
SAEVAAVKQSFSAVGVN

16. THERBAC2

Neutral proteinase (EC 3.4.24.4) - Bacillus sp.
ITGTSTVGVGRGVLGDQKNINTTYSTYYYLQDNTRGDGIFTYDAKYRTTLPGSLWADADN
QFFASYDAPAVDAHYYAGVTYDYYKNVHNRLSYDGNNAAIRSSVHYSQGYNNAFWNGSEM
VYGDGDGQTFIPLSGGIDVVAHELTHAVTDYTAGLIYQNESGAINEAISDIFGTLVEFYA
NKNPDWEIGEDVYTPGISGDSLRSMSDPAKYGDPDHYSKRYTGTQDNGGVHINSGIINKA
AYLISQGGTHYGVSVVGIGRDKLGKIFYRALTQYLTPTSNFSQLRAAAVQSATDLYGSTS
QEVASVKQAFDAVGVK

17. THERBAC3

Neutral proteinase (EC 3.4.24.4) precursor - Bacillus stearothermophilus
VAGASTVGVGRGVLGDQKYINTTYSSYYGYYYLQDNTRGSGIFTYDGRNRTVLPGSLWTD
GDNQFTASYDAAAVDAHYYAGVVYDYYKNVHGRLSYDGSNAAIRSTVHYGRGYNNAFWNG
SQMVYGDGDGQTFLPFSGGIDVVGHELTHAVTDYTAGLVYQNESGAINEAMSDIFGTLVE
FYANRNPDWEIGEDIYTPGVAGDALRSMSDPAKYGDPDHYSKRYTGTQDNGGVHTNSGII
NKAAYLLSQGGVHYGVSVNGIGRDKMGKIFYRALVYYLTPTSNFSQLRAACVQAAADLYG
STSQEVNSVKQAFNAVGVY

SERINE PROTEASES (38 sequences)

1. ELASHM1

E1304249A : elastase I
MLRLLVVASLVLYGHSTQDFPETNARVVGGTEAQRNSWPSQISLQYRSGSSWAHTCGGTL
IRQNWVMTAAHCVDRELTFRVVVGEHNLNQNDGTEQ VGVQKIVVHPYWNTDDVAAGYDI
ALLRLGQSVTLNSYVQLGVLPRAGTILANNSPCYITGWGLTRTNGQLAQTLQQAYLPTVD
YAILSSSSYWGSTVKNSMVCAGGDGVRSGCQGDSGGPLHCLVNGQYAVHGVTSFVSRLGC
NVTRKPTVFTRVSAYISWINNVIASN

2. ELASHM3

E1306262B : elastase IIB
MIRTLLLSTLVAGALSCGVSTYAPDMSRMLGGEEARPNSWPWQVSLQYSSNGQWYHTCGG
SLIANSWVLTAAHCISSSRIYRVMLGQHNLYVAESGSLAVSVSKIVVHKDWNSNQVSKGN
DIALLKLANPVSLTDKIQLACLPPAGTILPNNYPCYVTGWGRLQTNGALPDDLKQGRLLV
VDYATCSSSGWWGSTVKTNMICAGGDGVICTCNGDSGGPLNCQASDGRWEVHGIGSLTSV
LGCNYYYKPSIFTRVSNYNDWINSVIANN

3. ELASHM5

E1314212A : pancreatic elastase 2
MIRTLLLSTLVAGALSCGDPTYPPYVTRVVGGEEARPNSWPWQVSLQYSSNGKWYHTCGG
SLIANSWVLTAAHCISSSRTYRVGLGRHNLYVAESGSLAVSVSKIVVHKDWNSNQISKGN
DIALLKLANPVSLTDKIQLACLPPAGTILPNNYPCYVTGWGRLQTNGAVPDVLQQGRLLV
VDYATCSSSAWWGSSVKTSMICAGGDGVISSCNGDSGGPLNCQASDGRWQVHGIVSFGSR
LGCNYYHKPSVFTRVSNYIDWINSVIANN

4. ELASHM7

E1403370A : pancreatic elastase IIIA
MMLRLLSSLLLVAVASGYGPPSSHSSSRVVHGEDAVPYSWPWQVSLQYEKSGSFYHTCGG
SLIAPDWVVTAGHCISRDLTYQVVLGEYNLAVKEGPEQVIPINSEELFVHPLWNRSCVAC
GNDIALIKLSRSAQLGDAVQLASLPPAGDILPNKTPCYITGWGRLYTNGPLPDKLQQARL
PVVDYKHCSRWNWWGSTVKKTMVCAGGYIRSGCNGDSGGPLNCPTEDGGWQVHGVTSFVS
AFGCNFIWKPTVFTRVSAFIDWIEETIASH

5. ELASHM8

E1403370B : pancreatic elastase IIIB
MMLRLLSSLLLVAVASGYGPPSSRPSSRVVNGEDAVPYSWPWQVSLQYEKSGSFYHTCGG
SLIAPDWVVTAGHCISSSRTYQVVLGEYDRAVKEGPEQVIPINSGDLFVHPLWNRSCVAC
GNDIALIKLSRSAQLGDAVQLASLPPAGDILPNETPCYITGWGRLYTNGPLPDKLQEALL
PVVDYEHCSRWNWWGSSVKKTMVCAGGDIRSGCNGDSGGPLNCPTEDGGWQVHGVTSFVS
AFGCNTRRKPTVFTRVSAFIDWIEETIASH

6. ELASMSE1

E1301329A : elastase II
MIRTLLLSALVAGALSCGYPTYEVEDDVSRVVGGQEATPNTWPWQVSLQVLSSGRWRHNC
GGSLVANNWVLTAAHCLSNYQTYRVLLGAHSLSNPGAGSAAVQVSKLVVHQRWNSQNVGN
GYDIALIKLASPVTLSKNIQTACLPPAGTILPRNYVCYVTGWGLLQTNGNSPDTLRQGRL
LVVDYATCSSASWWGSSVKSSMVCAGGDGVTSSCNGDSGGPLNCRASNGQWQVHGIVSFG
SSLGCNYPRKPSVFTRVSNYIDWINSVMARN

7. ELASPIG2

E1306262C : elastase II
MIRALLLSTLVAGALSCGLPANLPQLPRVVGGEDARPNSWPWQVSLQYDSSGQWRHTCGG
TLVDQSWVLTAAHCISSSRTYRVVLGRHSLSTNEPGSLAVKVSKLVVHQDWNSNQLSNGN
DIALLKLASPVSLTDKIQLGCLPAAGTILPNNYVCYVTGWGRLQTNGASPDILQQGQLLV
VDYATCSKPGWWGSTVKTNMICAGGDGIISSCNGDSGGPLNCQGANGQWQVHGIVSFGSS
LGCNYYHKPSVFTRVSNYIDWINSVIANN

8. ELASRAT2

Proelastase precursor (EC 3.4.21.11) I - Rat
VVGGAEARRNSWPSQISLQYLSGGSWYHTCGGTLIRRNWVMTAAHCVSSQMTFRVVVGDH
NLSQNDGTEQYVSVQKIMVHPTWNSNNVAAGYDIALLRLAQSVTLNNYVQLAVLPQEGTI
LANNNPCYITGWGRTRTNGQLSQTLQQAYLPSVDYSICSSSSYWGSTVKTTMVCAGGDGV
RSGCQGDSGGPLHCLVNGQYSVHGVTSFVSSMGCNVSKKPTVFTRVSAYISWMNNVIAYT

9. ELASRAT3

Proelastase precursor (EC 3.4.21.11) II - Rat
VVGGQEASPNSWPWQVSLQYLSSGKWHHTCGGSLVANNWVLTAAHCISNSRTYRVLLGRH
SLSTSESGSLAVQVSKLVVHEKWNAQKLSNGNDIALVKLASPVALTSKIQTACLPPAGTI
LPNNYPCYVTGWGRLQTNGATPDVLQQGRLLVVDYATCSSASWWGSSVKTNMVCAGGDGV
TSSCNGDSGGPLNCQASNGQWQVHGIVSFGSTLGCNYPRKPSVFTRVSNYIDWINSVIAK
N

10. KALIGP1

Z1309183A : kallikrein
VIGGQECARDSHPWQAAVYYYSDIKCGGVLVDPQWVLTAAHCINDSNQVKLGRHNLFEDE
DTAQHFLVSQSVPHPDFYMSLLEPHNVLPNEDYSHDLMLLRLNQPAQITDSVQVMPLPTQ
EVQVGTTCRALGWGSIDPDPAHPVFPDELQCVGLEILPSKNCDDAHIAYVTSEMLCAGDL
AGGKDTCVGDSGGPLICDGVLQGLTSWGDSPCGVAHSPSLYTKVIEYREWIERTMADNP

11. KALIHM3

Z1404431A : urinary kallikrein
IVGGWECEQHSQPWQAALYHFSTFQCGGILVHRQWVLTAAHCISDNYQLWLGRHNLFDDE
NTAQFVHVSESFPHPGFNMSLLENHTRQADEDYSHDLMLLRLTEPADTITDAVKVVELPT
QEPEVGSTCLASGWGSIEPENFSFPDDLQCVDLKILPNDECEKAHVQKVTDFMLCVGHLE
GGKDTCVGDSGGPLMCDGVLQGVTSWGYVPCGTPNKPSVAVRVLSYVKWIEDTIAENS

12. KALIMSE3

Z1205271A : kallikrein,glandular
MRFLILFLALSLGGIDAAPPVQSRIVGGFNCEKNSQPWQVAVYRFTKYQCGGILLNANWV
LTAAHCHNDKYQVWLGKNNFLEDEPSAQHRLVSKAIPHPDFNMSLLNEHTPQPEDDYSND
LMLLRLKKPADITDVVKPIDLPTEEPKLGSTCLASGWGSITPVKYEYPDELQCVNIKLLP
NEDCAKAHIEKVTDDMLCAGDMDGGKDTCAGDSGGPLICDGVLQGITSWGPRPCGKPNVP
GIYTRVLNFNTWIRETMAEND

13. KALIMSE4

Z1208372A : kallikrein PMF2
MRFLILFLALSLGGIDAAPPLQSRVVGGFNCEKNSQPWQVAVYDNKEHICGGVLLEKNWV
LTAAHCYVDQYEVWLGKNKLFQEEPSAQHRLVSKSFPHPGFNMSLLTLKEIPPGADFSND
LMLLRLSKPADITDAVKPITLPTKESKLGSTCLASGWGSITPTKWQKPDDLQCVFLKLI
IKNCIENHNVKVTDVMLCAGEMSGGKNICKGDSGGPLICDSVLQGITSTGPIPCGKPGVP
AMYTNLIKFNSWIKDTMTKNS

14. KALIMSE5

Z1313201A : glandular kallikrein
MRFLILFLTLSLGGIDAAPPVQSRILGGFKCEKNSQPWQVAVYYLDEYLCGGVLLDRNWV
LTAAHCYEDKYNIWLGKNKLFQDEPSAQHRLVSKSFPHPDFNMSLLQSVPTGADLSNDLM
LLRLSKPADITDVVKPIDLPTTEPKLGSTCLASGWGSINQLIYQNPNDLQCVSIKLHPNE
VCVKAHILKVTDVMLCAGEMNGGKDTCKGGSGGPLICDGVLQGITSWGSTPCGEPNAPAI
YTKLIKFTSWIKDTMAKNP

15. KALIMSE6

Z1313201B : glandular kallikrein
MWFLILFLALSLGGIDAAPPLQSRVVGGFNCKKNSQPWQVAVYYQKEHICGGVLLDRNWV
LTAAHCYVDQYEVWLGKNKLFQEEPSAQHRLVSKSFPHPGFNMSLLMLQTIPPGADFSND
LMLLRLSKPADITDVVKPIALPTKEPKPGSKCLASGWGSITPTRWQKPDDLQCVFITLLP
NENCAKVYLQKVTDVMLCAGEMGGGKDTCRDDSGGPLICDGILQGTTSYGPVPCGKPGVP
AIYTNLIKFNSWIKDTMMKNA

16. KALIMSE7

Z1313201C : glandular kallikrein
MRFLILFLALSLGGIDAAPPVHSRIVGGFKCEKNSQPWHVAVYRYNEYICGGVLLDANWV
LTAAHCYYEENKVSLGKNNLYEEEPSAQHRLVSKSFLHPGYNRSLHRNHIRHPEYDYSND
LMLLRLSKPADITDVVKPIALPTEEPKLGSTCLASGWGSTTPFKFQNAKDLQCVNLKLLP
NEDCGKAHIEKVTDVMLCAGETDGGKDTCKGDSGGPLICDGVLQGITSWGFTPCGEPKKP
GVYTKLIKFTSWIKDTMAKNL

17. KALIMSE8

Z1402323A : glandular kallikrein
MWFLILFLALSLGGIDAAPPVQSRIFGGFNCEKNSQPWQVAVYRFTKYQCGGVLLNANWV
LTAAHCKNDKYQVWLGKNNFFEDEPSAQHRLVSKAIPHPDFNMSLLNEHTPQPEDDYSND
LMLLRLKKPADITDVVKPIDLPTEEPKLGSTCLASGWGSITPVIYEPADDLQCVNFKLLP
NEDCVKAHIEKVTDVMLCAGDMDGGKDTCMGDSGGPLICDGVLHGITSWGPSPCGKPNVP
GIYTKLIKFNSWIKDTIAKNA

18. KALIPIG1

Tissue kallikrein (EC 3.4.21.35), pancreatic - Pig
IIGGRECEKNSHPWQVAIYHYSSFQCGGVLVNPKWVLTAAHCKNDNYEVGWLRHNLFENE
NTAQFFGVTADFPHPGFNLSADGKDYSHDLMLLRLQSPAKITDAVKVLELPTQEPELGST
CEASGWGSIEPGPDDFEFPDEIQCVQLTLLQNTFCAHAHPBKVTESMLCAGYLPGGKDTC
MGDSGGPLICNGMWQGITSWGHTPCGSANKPSIYTKLIFYLDWIBBTITENP

19. KALIRAT4

Z1110144C : kallikrein S3
MWFLILFLALSLGQIDAAPPGQSRVVGGYNCETNSQPWQVAVIGTTFCGGVLIDPSWVIT
AAHCYSKNYRVLLGRNNLVKDEPFAQRRLVSQSFQHPDYIPVFMRNHTRQRAYDHNNDLM
LLHLSKPADITGGVKVIDLPTEEPKVGSICLASGWGMTNPSEMKLSHDLQCVNIHLLSNE
KCIETYKNIETDVTLCAGEMDGGKDTCTGDSGGPLICDGVLQGLTSGGATPCAKPKTPAI
YAKLIKFTSWIKKVMKENP

20. SUBTBAC1

E671056A : subtilisin
AQSVPYGVSQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDLKVAGGASMVPSETPNFQD
DNSHGTHVAGTVAALNNSIGVLGVAPSSALYAVKVLGDAGSGQYSWIINGIEWAIANNMD
VINMSLGGPSGSAALKAAVDKAVASGVVVVAAAGNEGSTGSSSTVGYPGKYPSVIAVGAV
DSSNQRASFSSVGPELDVMAPGVSIQSTLPGNKYGAYNGTSMASPHVAGAAALILSKHPN
WTNTQVRSSLQNTTTKLGDSFYYGKGLINVQAAAQ

21. SUBTBAC2

E721944A : subtilisin
AQSVPYGISQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDLNVRGGASFVPSETPNYQD
GSSHGTHVAGTIAALNNSIGVLGVAPSSALYAVKVLDSTGSGQYSWIINGIEWAISNNMD
VINMSLGGPSGSTALKTVVDKAVSSGIVVAAAAGNEGSSGSSSTVGYPAKYPSTIAVGAV
NSSNQRASFSSAGSELDVMAPGVSIQSTLPGGTYGAYNGTSMATPHVAGAAALILSKHPT
WTNAQVRDRLESTATYLGDSFYYGKGLINVQAAAQ

22. SUBTBAC3

Z0912207A : subtilisin DY
AQTVPYGIPLIKADKVQAQGYKGANVKVGIIDTGIAASHTDLKVVGGASFVSGESYNTDG
NGHGTHVAGTVAALDNTTGVLGVAPNVSLYAIKVLNSSGSGTYSAIVSGIEWATQNGLDV
INMSLGGPSGSTALKQAVDKAYASGIVVVAAAGNSGSSGSQNTIGYPAKYDSVIAVGAVD
SNKNRASFSSVGAELEVMAPGVSVYSTYPSNTYTSLNGTSMASPHVAGAAALILSKYPTL
SASQVRNRLSSTATNLGDSFYYGKGLINVEAAAQ

23. SUBTBAC6

Subtilisin Carlsberg (EC 3.4.21.14) - Bacillus subtilis
AQTVPYGIPLIKADKVQAQGFKGANVKVAVLDTGIQASHPDLNVVGGASFVAGEAYNTDG
NGHGTHVAGTVAALDNTTGVLGVAPSVSLYAVKVLNSSGSGSYSGIVSGIEWATTNGMDV
INMSLGGASGSTAMKQAVDNAYARGVVVVAAAGNSGNSGSTNTIGYPAKYDSVIAVGAVD
SNSNRASFSSVGAELEVMAPGAGVYSTYPTNTYATLNGTSMASPHVAGAAALILSKHPNL
SASQVRNRLSSTATYLGSSFYYGKGLINVEAAAQ

24. TRPCRAY1
Trypsin (EC 3.4.21.4) I - Crayfish
IVGGTDAVLGEFPYQLSFQETFLGFSFHFCGASIYNENYAITAGHCVYGDDYENPSGLQI
VAGELDMSVNEGSEQTITVSKIILHENFDYDLLDNDISLLKLSGSLTFNNNVAPIALPAQ
GHTATGNVIVTGWGTTSEGGNTPDVLQKVTVPLVSDAECRDDYGADEIFDSMICAGVPEG
GKDSCQGDSGGPLAASDTGSTYLAGIVSWGYGCARPGYPGVYTEVSYHVDWIKANAV
25. TRPDOG1
Trypsinogen (EC 3.4.21.4), cationic, precursor - Dog
IVGGYTCSRNSVPYQVSLNSGYHFCGGSLINSQWVVSAAHCYKSRIQVRLGEYNIAVSEG
GEQFINAAKIIRHPRYNANTIDNDIMLIKLSSPATLNSRVSAIALPKSCPAAGTQCLISG
WGNTQSIGQNYPDVLQCLKAPILSDSVCRNAYPGQISSNMMCLGYMEGGKDSCQGDSGGP
VVCNGELQGVVSWGAGCAQKGKPGVSPKVCKYVSWIQQTIAAN
26. TRPDOG2
Trypsinogen (EC 3.4.21.4), anionic, precursor - Dog
IVGGYTCEENSVPYQVSLNAGYHFCGGSLISDQWVVSAAHCYKSRIQVRLGEYNIDVLEG
NEQFI[illegible]AKVIRHPNYNSWILDNDIMLIKLSSPAVLNARVATISLPRACAAPGTQCLISG
WGNTL[illegible]STNYPELLQCLDAPILTQAQCEASYPGQITENMICAGFLEGGKDSCQGDSGGP
VVCNGELQGIVSWGYGCAQKNKPGVYTKVCNFVDWIQSTIAANS
27. TRPFISH1
Trypsinogen (EC 3.4.21.4) - Spiny dogfish
IVGGYECPKHAAPWTVSLNVGYHFCGGSLIAPGWVVSAAHCYQRRIQVRLGEHDISANEG
DETYIDSSMVIRHPNYSGYDLDNDIMLIKLSKPAALNRNVDLISLPTGCAYAGEMCLISG
WGNTMDGAVSGDQLQCLDAPVLSDAECKGAYPGMITNNMMCVGYMEGGKDSCQGDSGGPV
VCNGMLQGIVSWGYGCAERDHPGVYTRVCHYVSWIHETIASV
28. TRPFISH2
E750677A : trypsin
IVGGYECPKHAAPWTVSLNVGYHFCGGSLIAPGWVSAAHCYQRRIQVRLGEHDISANEGD
ETYIDSMVIRHPNYSGYDLDNDIMLIKLSKPAALNRNVDLISLPTGCAYAGEMCLISGWG
NTMDGAVSGDQLQCLDAPVLSDAECKGAYPGMITNNMMCVGYMEGGKDSCQGDSGGPVVC
NGMLQGIVSWGYGCAERDHPGVYTRVCHYVSWIHETIASV
29. TRPFLY1
Trypsinogen-like (EC 3.4.21.-) proenzyme precursor - Fruit fly
IVGGSATTISSFPWQISLQRSGSHSCGGSIYSANIIVTAAHCLQSVSASVLQVRAGSTYW
SSGGVVAKVSSFKNHEGYNANTMVNDIAVIRLSSSLSFSSSIKAISLATYNPANGASAAV
SGWGTQSSGSSSIPSQLQYVNVNIVSQSQCASSTYGYGSQIRNTMICAAASGKDACQGDS
GGPLVSGGVLVGVVSWGYGCAYSNYPGVYADVAVLRSWVVSTANSI
30. TRPHM1
E1403437A : cell specific trypsin like protease
MRNSYRFLASSLSVVVSLLLIPEDVCEKIIGGNEVTPHSRPYMVLLSLDRKTICAGALIA
KDWVLTAAHCNLNKRSQVILGAHSITREEPTKQIMLVKKEFPYPCYDPATREGDLKLLQL
TEKAKINKYVTILHLPKKGDDVKPGTMCQVAGWGRTHNSASWSDTLREVNITIIDRKVCN
DRNHYNFNPVIGMNMVCAGSLRGGRDSCNGDSGSPLLCEGVFRGVTSFGLENKCGDPRGP
GVYILLSKKHLNWIIMTIKGAV
31. TRPHOR1
Chymotrypsin II (EC 3.4.21.1) - Oriental hornet
IVGGTNAPRGKYPYQVSLRAPKHFCGGSISKRYVLTAAHCLVGKSEHQVTVGSVLLNKEE
AVYNAKELIVNKNYNSIRLINDIGLIRVSKDISFTQLVQPVKLPVSNTIKAGDPVVLTGW
GRIYVNGPIPNNLQQITLSIVNQQTCKSKHWGLTDSQICTFTKRGEGACHGDSGGPLVAN
GVQIGIVSYGHPCAIGSPNVFTRVYSFLDWIQKNQL
32. TRPHOR2
Chymotrypsin II (EC 3.4.21.1) - European hornet
IVGGTDAPRGKYPYQVSLRAPKHFCGGSISKRYVLTAAHCLVGKSKHQVTVHAGSVLLNK
EEAVYNAEELIVNKNYNSIRLINDIGLIRVSK[illegible]SYTQLVQPVKLPVSNTIKAGDPVVLT
GWGRIYVNGPIPNNLQQITLSIVNQQTCK[illegible]KHWGLTDSQICTFTKLGEGACDGDSGGPLV
ANGVQIGIVSYGHPCAVGSPNVFTRVYSFLDWIQKNQL
33. TRPMSE1
Z1301329B : trypsin
MSALLILALVGAAVAFPVDDDDKIVGGYTCRESSVPYQVSLNAGYHFCGGSLINDQWVVS
AAHCYKYRIQVRLGEHNINVLEGNEQFVDSAKIIRHPNYNSWTLDNDIMLIKLASPVTLN
ARVASVPLPSSCAPAGTQCLISGWGNTLSNGVNNPDLLQCVDAPVLPQADCEASYPGDIT
NNMICVGFLEGGKDSCQGDSGGPVVCNGELQGIVSWGYGCAQPDAPGVYTKVCNYVDWIQ

NTIADN

34. TRPPIG1

Trypsinogen (EC 3.4.21.4) - Pig

IVGGYTCAANSIPYQVSLNSGSHFCGGSLINSQWVVSAAHCYKSRIQVRLGEHNIDVLEG
NEQFINAAKIITHPNFNGNTLDNDIMLIKLSSPATLNSRVATVSLPRSCAAAGTECLISG
WGNTKSSGSSYPSLLQCLKAPVLSDSSCKSSYPGQITGNMICVGFLEGGKDSCQGDSGGP
VVCNGQLQGIVSWGYGCAQKNKPGVYTKVCNYVNWIQQTIAAN

35. TRPRAT1

Trypsinogen I (EC 3.4.21.4) precursor - Rat

IVGGYTCPEHSVPYQVSLNSGYHFCGGSLINDQWVVSAAHCYKSRIQVRLGEHNINVLEG
DEQFINAAKIIKHPNYSSWTLNNDIMLIKLSSPVKLNARVAPVALPSACAPAGTQCLISG
WGNTLSNGVNNPDLLQCVDAPVLSQADCEAAYPGEITSSMICVGFLEGGKDSCQGDSGGP
VVCNGQLQGIVSWGYGCALPDNPGVYTKVCNFVGWIQDTIAAN

36. TRPRAT2

Trypsinogen II (EC 3.4.21.4) precursor - Rat

IVGGYTCQENSVPYQVSLNSGYHFCGGSLINDQWVVSAAHCYKSRIQVRLGEHNINVLEG
DEQFINAAKIIKHPNFDRKTLNNDIMLIKLSSPVKLNARVATVALPSSCAPAGTQCLISG
WGNTLSSGVNEPDLLQCLDAPLLPQADCEASYPGKITDNMVCVGFLEGGKDSCQGDSGGP
VVCNGELQGIVSWGYGCALPDNPGVYTKVCNYVDWIQDTIAAN

37. TRPRAT3

Chymotrypsinogen B (EC 3.4.21.1) precursor - Rat

IVNGEDAIPGSWPWQVSLQDKTGFHFCGGSLISEDWVVTAAHCGVKTSDVVVAGEFDQGS
DEENIQVLKIAQVFKNPKFNMFTVRNDITLLKLATPAQFSETVSAVCLPNVDDDFPPGTV
CATTGWGKTKYNALKTPEKLQQAALPIVSEADCKKSWGSKITDVMTCAGASGVSSCMGDS
GGPLVCQKDGVWTLAGIVSWGSGVCSTS.PAVYSRVTALMPWVQQILEAN

38. TRPSTRP1

Trypsin (EC 3.4.21.4) - Streptomyces griseus

VVGGTRAAQGEFPFMVRLSMGCGGALYAQDIVLTAAHCVSGSGNNTSITATGGVVDLQSA
VKVRSTKVLQAPGYNGTGKDWALIKLAQPINQPTLKIATTTAYNQGTFTVAGWGANREGG
SQQRYLLKANVPFVSDAACRSAYGNELVANEEICAGYPDTGGVDTCQGDSGGPMFRKDNA
DEWIQVGIVSWGYGCARPGYPGVYTEVSTFASAIASAARTL

ASPARTIC PROTEASES (47 sequences)

1. AID1

Protease (EC 3.4.) - Simian AIDS retrovirus SRV-1

SRKPTTTPSGKRTEGPAPGPETSLWGGQLCSSQQKQPISKLTRATPGSAGLDLSSTSHTV
LTPEMGPQALSTGIYGPLPPNTFGLILGRSSITIKGLQVYPGVIDNDYTGEIKIMAKAVN
NIVTVPQGNRIAQLILLPLIETDNKVQQPYRGQGSFGSSDIYWVQPITCQKPSLTLWLDD
KMFTGLIDTGADVTIIKLEDWPPNWPITDTLTNLRGIGQSNNPKQSSKYLTWRDKENNSG
LIKPFVIPNLPVNLWGRDLLSQMKIMMCSPSDIVTAQMLAQGYSPGKGLGKNENGILHPI
PNQGQFDKKGFGNF

2. AID10

sor protein - Human immunodeficiency virus (HIV-2, isolate ROD)

MEEDKRWIVVPTWRVPGRMEKWHSLVKYLKYKTKDLEKVCYVPHHKVGWAWWTCSRVIFP
LKGNSHLEIQAYWNLTPEKGWLSSYSVRITWYTEKFWTDVTPDCADVLIHSTYFPCFTAG
EVRRAIRGEKLLSCCNYPRAHRAQVPSLQFLALVVVQQNDRPQRDSTTRKQRRRDYRRGL
RLAKQDSRSHKQRSSESPTPRTYFPGVAEVLEILA

3. AID11

nef protein - Human immunodeficiency virus I (HIV-1, isolate BR)

MGGKWSKMAGWSTVRERMRRAEPARERMRRAEPRAEPAADGVGAVSRDLEKHGAITSSNT
AATNADCAWLEAQEDEEVGFPVKPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIHSQQRQD
ILDLWVYHTQGYFPDWQNYTPGPGVRYPLTFGWCFKLVPVEPEKIEEANEGENNSLLHPM
SQHGMDDPEREVLQWRFDSRLAFHHMARELHPEYYKNC

4. AID12

nef protein - Human immunodeficiency virus (isolate HIV-Zr6)

MGGRWSKSSIVGWPAIRERIRRTDPRRTDPAADGVGAASRDLEKHGAITSSNTRDTNADC
AWLEAQEESEEVGFPVRPQVPLRPMTYKLAVDLSHFLKEKGGLEGLIWSKKRQEILDLWV
YNTQGIFPDWQNYTPGPGIRYPLTFGWCFELVPVDPREVEEATEGETNCLLHPVCQHGME

DTEREVLKWRFNSRLAFEHKAREMHPEFYKDC

5. AID13

nef protein - Human immunodeficiency virus (HIV-2, isolate ROD)

MGASGSKKHSRPPRGLQERLLRARAGACGGYWNESGGEYSRFQEGSDREQKSPSCEGRQY
QQGDFMNTPWKDPAAEREKNLYRQQNMDDVDSDDDDQVRVSVTPKVPLRPMTHRLAIDMS
HLIKTRGGLEGMFYSERRHKILNIYLEKEEGIIADWQNYTHGPGVRYPMFFGWLWKLVPV
DVPQEGEDTETHCLVHPAQTSKFDDPHGETLVWEFDPLLAYSYEAFIRYPEEFGHKSGLP
EEEWKARLKARGIPFS

6. AID2

gag polyprotein - Human immunodeficiency virus (HIV-1, isolate CDC-451)

MGARASVLSGGELDRWEKIRLRPGGKKQYRLKHIVWASRKLERFAVNPGLLETSKGCRQI
LGQLQPSLQTGSEELRSLYNTVATLYCVHQRIEVRDTKEALDKIEEEQNKSKKKAQQAAA
DTGNSSQVSQNYPIVQNLQGQMVHQAISPRTLNAWVKVIEEKAFSPEVIPMFAALSEGAT
PQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRLHPVHAGPIAPGQMREPRGSDIAGTT
STLQEQIGWMTNNPPTPVGEIYKRWIILGLNKIVRMYSPISILDIRQGPKEPFRDYVDRF
YKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGVGGPGHKA
RVLAEAMSQVTNSATIMMQRGNFRRQGKTVKCFNCGKEGHIARNCKAPRKKGCWKCGREG
HQMKDCTERQANFLGKIWPSHKGRPGNFLQSRPEPTAPPEESFRFGDETTTPSQKQEPRD
KELYPLASLRSLFGNDPSSQ

7. AID3

gag polyprotein - Human immunodeficiency virus (HIV-2, isolate ROD)

MGARNSVLRGKKADELERIRLRPGGKKKYRLKHIVWAANKLDRFGLAESLLESKEGCQKI
LTVLDPMVPTGSENLKSLFNTVCVIWCIHAEEKVKDTEGAKQIVRRHLVAETGTAEKMPS
TSRPTAPSSEKGGNYPVQHVGGNYTHIPLSPRTLNAWVKLVEEKKFGAEVVPGFQALSEG
CTPYDINQMLNCVGDHQAAMQIIREIINEEAAEWDVQHPIPGPLPAGQLREPRGSDIAGT
TSTVEEQIQWMFRPQNPVPVGNIYRRWIQIGLQKCVRMYNPTNILDIKQGPKEPFQSYVD
RFYKSLRAEQTDPAVKNWMTQTLLVQNANPDCKLVLKGLGMNPTLEEMLTACQGVGGPGQ
KARLMAEALKEVIGPAPIPFAAAQQRKAFKCWNCGKEGHSARQCRAPRRQGCWKCGKPGH
IMTNCPDRQAGFLGLGPWGKKPRNFPVAQVPQGLTPTAPPVDPAVDLLEKYMQQGKRQRE
QRERPYKEVTEDLLHLEQGETPYREPPTEDLLHLNSLFGKDQ

8. AID4

pol polyprotein - Human immunodeficiency virus (HIV-2, isolate ROD)

TGRFFRTGPLGKEAPQLPKGPSSAGAQTNSTPSGSSSGSTGEIYAAREKTERAERETIQG
SDRGLTAPRAGGDTIQGATNRGLAAPQFSLWKRPVVTAYIEGQPVEVLLDTGADDSIVAG
IELGNNYSPKIVGGIGGFINTKEYKNVEIEVLNKKVRATIMTGDTPINIFGRNILTALGM
SLNLPVAKVEPIKIMLKPGKDGPKLRQWPLTKEKIEALKEICEKMEKEGQLEEAPPTNPY
NTPTFAIKKKDKNKWRMLIDFRELNKVTQDFTEIQLGIPHPAGLAKKRRITVLDVGDAYF
SIPLHEDFRPYTAFTLPSVNNAEPGKRYIYKVLPQGWKGSPAIFQHTMRQVLEPFRKANK
DVIIIQYMDDILIASDRTDLEHDRVVLQLKELLNGLGFSTPDEKFQKDPPYHWMGYELWP
TKWKLQKIQLPQKEIWTVNDIQKLVGVLNWAAQLYPGIKTKHLCRLIRGKMTLTEEVQWT
ELAEAELEENRIILSQEQEGHYYQEEKELEATVQKDQENQWTYKIHQEEKILKVGKYAKV
KNTHTNGIRLLAQVVQKIGKEALVIWGRIPKFHLPVEREIWEQWWDNYWQVTWIPDWDFV
STPPLVRLAFNLVGDPIPGAETFYTDGSCNRQSKEGKAGYVTDRGKDKVKKLEQTTNQQA
ELEAFAMALTDSGPKVNIIVDSQYVMGISASQPTESESKIVNQIIEEMIKKEAIYVAWVP
AHKGIGGNQEVDHLVSQGIRQVLFLEKIEPAQEEHEKYHSNVKELSHKFGIPNLVARQIV
NSCAQCQQKGEAIHGQVNAELGTWQMDCTHLEGKIIIVAVHVASGFIEAEVIPQESGRQT
ALFLLKLASRWPITHLHTDNGANFTSQEVKMVAWWIGIEQSFGVPYNPQSQGVVEAMNHH
LKNQISRIREQANTIETIVLMAIHCMNFKRRGGIGDMTPSERLINMITTEQEIQFLQAKN
SKLKDFRVYFREGRDQLWKGPGELLWKGEGAVLVKVGTDIKIIPRRKAKIIRDYGGRQEM
DSGSHLEGAREDGEMA

9. AID5

env polyprotein - Human immunodeficiency virus I (HIV-1, isolate BR)

MRVKGIKKNYQHLWRWGGMMLLGILMICSATDKLWVTVYYGVPVWKEANTTLFCASDAKA
YDTEIHNVWATHACVPTDPNPQELVMGNVTENFNMWKNDMVEQMHEDIISLWDQSLKPCV
KLTPLCLTLNCHDFNATNATSNSGKMMEGGEMKNCSFNITTSIRDKMQKEYALFYKLDIV
PIDNDKTNTRYRLISCNTSVITQACPKVTFEPIPIHYCAPAGFAILKCNNKKFNGTGPCT
NVSTVQCTHGIRPVVSTQLLLNGSLAEEEVVIRSENFTNNVKTIIVQLNESVEINCTRPN
NNTRKRITMGPGRVYYTTGQIIGDIRRAHCNLSRSKWENTLKQIVTKLRVQFKNKTIVFN
RSSGGDPEIVMHSFNCGGEFFYCNTTQLFNSTWYRNTTGNITEGNSPITLPCRIKQIINM
WQEVGKAMYAPPIRGQIKCSSNITGLLLTRDGGNNNETTDTEIFRPGGGNMRDNWRSELY
KYKVVKIEPLGVAPTKAKRRVVQREKRAVGLGALFLGFLGAAGSTMGAASLTLTVQARLL

LSGIVQQQNNLLMAIEAQQHMLELTVWGIKQLQARVLAVERYLKDQQLLGIWGCSGKLIC
TTAVPWNASWSNKSLSDIWDNMTWMEWEREIDNYTNLIYSLIEDSQIQQEKNEKELLELD
KWASLWNWFNITNWLWYIKIFIMIVGGLIGLRIVFAVLSIVNRVRQGYSPLSFQTRLPGR
RGPDRPEGIEEEGGERDRDRSSPLVDGFLALFWVDLRSLFLFSYHRLRDLLLIVTRIVEL
LGRRGWEVLKYWWNLLQYWSQELKNSAVSLLNATAIAVGERTDRAIEVVQRAFRAILHIP
RRIRQGLERALQ

10. AID6

env polyprotein - Human immunodeficiency virus (HIV-1, isolate CDC-451)
MAMRAKGIRKNCQHLWRWGTMLLGMLMICSAAANLWVTVYYGVPVWKEATTTLFCASDAK
AYDTEAHNVWATHACVPTNPNPQEVVLENVTENFNMWKNNMVEQMHEDIISLWDQSLKPC
VKLTPLCVTLNCTDLNTNNTTNTTELSIIVVWEQRGKGEMRNCSFNITTSIRDKVQREYA
LFYKLDVEPIDDNKNTTNNTKYRLINCNTSVITQACPKVSFEPIPIHYCTPTGFALLKCN
DKKFNGTGPCTNVSTVQCTHGIRPVVSTQLLLNGSLAEEEVVIRSENFTNNAKTIIVQLN
VSVEINCTRPNNHTRKRVTLGPGRVWYTTGEILGNIRQAHCNISRAQWNNTLQQIATTLR
EQFGNKTIAFNQSSGGDPEIVMHSFNCGGEFFYCNSTQLFNSAWNVTSNGTWSVTRKQKD
TGDIITLPCRIKQIINRWQVVGKAMYALPIKGLIRCSSNITGLLLTRDGGGENQTTEIFR
PGGGDMRDNWRSELYKYKVVKIEPLGVAPTKAKRRVVQREKRAVGMLGAMFLGFLGAAGS
TMGATSMALTVQARQLLSGIVQQQNNLLRAIKAQQHLLQLTVWGIKQLQARILAVERYLK
DQQLLGFWGCSGKLICTTAVPWNASWSNKTLDQIWNNMTWMEWDREIDNYTHLIYTLIEE
SQNQQEKNQQELLQLDKWASLWTWSDITKWLWYIKIFIMIVGGLIGLRIVFAVLSIVNRV
RQGYSPLSFQTLLPNPRGPDRPEGTEEGGGERGRDGSTRLVHGFLALVWDDLRSLCLFSY
HRLRDLLLIVARIVELLGRRGWEVLKYWWNLLQYWSQELKNSAVSLVNVTAIAVAEGTDR
VIEVVQRIYRAFLHIPRRIRQGFERALL

11. AID7

env polyprotein precursor - Human immunodeficiency virus (isolate HIV-Zr6)
LLGILMICSAADNLWVTVYYGVPVWKEATTTLFCASDAKSYKTEAHNIWATHACVPTDPN
PQEIELENVTENFNMWRNNMVEQIHEDIISLWDQSLKPCVKLTPLCVTLNCTDESDEWMG
NVTGKNVTEDIRMKNCSFNITTVVRDKTKQVHALFYRLDIVPIDNDNSTNSTNYRLINCN
TSAITQACPKVSFEPIPIHYCAPAGFAILKCRDKRFNGTGPCTNVSTVQCTHGIRPVVST
QLLLNGSLAEEEIIIRSENLTNNAKIIIVQLNESVAINCTRPYKNTRQSTPIGLGQALYT
TRGRTKIIGQAHCNISKEDWNKTLQRVAIKLGNLLNKTTIIFKPSSGGDAEITTHSFNCG
GEFFYCNTSGLFNSTWNINNSEGANSTESDNKLITLQCRIKQIINMWQGVGKAMYAPPIE
GQINCSSNITGLLLTRDGGTNNSSNETFRPGGGDMRDNWRSELYKYKVVKIEPLGVAPTK
AKRRVVEREKRAIGLGAMFLGFLGAAGSTMGAASVTLTVQARQLMSGIVQQQNNLLRAIE
AQQHLLQLTVWGIKQLQARILAVERYLKDQQLLGIWGCSGKLICTTTVPWNSSWSNRSLN
DIWQNMTWMEWEREIDNYTGLIYRLIEESQTQQEKNEQELLELDKWASLWNWFNITQWLW
YIKIFIMIVGGLIGLRIVFAVLSLVNRVRQGYSPLSFQTLLPAPREPDRPEGIEEEGGER
GRDRSIRLVNGFSALIWDDLRNLCLFSYHRLRDLILIAARIVELLGRRGWEALKYLWNLL
QYWSRELRNSASSLLDTIAIAVAEGTDRVIEIVRRTYRAVLNVPTRIRQGLERLLL

12. AID8

env polyprotein precursor - Human immunodeficiency virus (HIV-2, isolate ROD)
YCTQYVTVFYGVPTWKNATIPLFCATRNRDTWGTIQCLPDNDDYQEITLNVTEAFDAWNN
TVTEQAIEDVWHLFETSIKPCVKLTPLCVAMKCSSTESSTGNNTTSKSTSTTTTTPTDQE
QEISEDTPCARADNCSGLGEEETINCQFNMTGLERDKKKQYNETWYSKDVVCETNNSTNQ
TQCYMNHCNTSVITESCDKHYWDAIRFRYCAPPGYALLRCNDTNYSGFAPNCSKVVASTC
TRMMETQTSTWFGFNGTRAENRTYIYWHGRDNRTIISLNKYYNLSLHCKRPGNKIVKQIM
LMSGHVFHSHYQPINKRPRQAWCWFKGKWKDAMQEVKETLAKHPRYRGTNDTRNISFAAP
GKGSDPEVAYMWTNCRGEFLYCNMTWFLNWIENKTHRNYAPCHIKQIINTWHKVGRNVYL
PPREGELSCNSTVTSIIANIDWQNNNQTNITFSAEVAELYRLELGDYKLVEITPIGFAPT
KEKRYSSAHGRHTRGVFVLGFLGFLATAGSAMGAASLTVSAQSRTLLAGIVQQQQQLLDV
VKRQQELLRLTVWGTKNLQARVTAIEKYLQDQARLNSWGCAFRQVCHTTVPWVNDSLAPD
WDNMTWQEWEKQVRYLEANISKSLEQAQIQQEKNMYELQKLNSWDIFGNWFDLTSWVKYI
QYGVLIIVAVIALRIVIYVVQMLSRLRKGYRPVFSSPPGYIQQIHIHKDRGQPANEETEE
DGGSNGGDRYWPWPIAYIHFLIRQLIRLLTRLYSICRDLLSRSFLTLQLIYQNLRDWLRL
RTAFLQYGCEWIQEAFQAAARATRETLAGACRGLWRVLERIGRGILAVPRRIRQGAEIAL
L

13. AID9

sor protein - Human immunodeficiency virus (isolate HIV-Zr6)
MENRWQVMIVWQVDRMRIRTWKSLVKHHMYVSKKASRWFYRHHYDSPHPKISSEVHIPLG

EARLVVKTYWGLHTGERDWHLGQGVSIEWRKRRYSTQVDPGLADQLIHMYYFDCFSEAAI
RKAILGHIVSHRCEYQAGHSKVGSLQYLALTALIAPKKIKPPLPSVRKLTEDRWNKPQKT
KGHKGAIQ

14. CHYMCOW1

E0510208A : chymosin B

GEVASVPLTNYLDSQYFGKIYLGTPPQEFTVLFDTGSSDFWVPSIYCKSNACKNHQRFDP
RKSSTFQNLGKPLSIHYGTGSMQGILGYDTVTVSNIVDIQQTVGLSTQEPGDVFTYAEFD
GILGMAYPSLASEYSIPVFDNMMNRHLVAQDLFSVYMDRDGQESMLTLGAIDPSYYTGSL
HWVPVTVQQYWQFTVDSVTISGVVVACEGGCQAILDTGTSKLVGPSSDILNIQQAIGATQ
NQYGEFDIDCDNLSYMPTVVFEINGKMYPLTPSAYTSQDQGFCTSGFQSENHSQKWILGD
VFIREYYSVFDRANNLVGLAKAI

15. CHYMLAM1

lamb chymosin nar 1990

GEVASVPLTNYLDSQYFGKIYLGTPPQEFTVLFDTGSSDFWVPSIYCKSNACKNHQRFDP
RKSSTFQNLGKPLSIRYGTGSMQGILGYDTVTVSNIVDIQQTVGLSTQEPGDVFTYAEFD
GILGMAYPSLASEYSVPVFDNMMDRRLVAQDLFSVYMDRSGQGSMLTLGAIDPSYYTGSL
HWVPVTLQKYWQFTVDSVTISGAVVACEGGCQAILDTGTSKLVGPSSDILNIQQAIGATQ
NQYGEFDIDCDSLSSMPTVVFEINGKMYPLTPYAYTSQEEGFCTSGFQGENHSHQWILGD
VFIREYYSVFDRANNLVGLAKAI

16. HTLV1

Protease (EC 3.4.21.) - T-cell leukemia virus (HTLV-I)

HSTPKKLHRGGGLTSPPTLQQVFLNQDPASILPVIPLDPARRPVIKAQVDTQTSHPKTIE
ALLDTGADMTVLPIALFSSNTPLKNTSVLGAGGQTQDHFKLTSLPVLIRLPFRTTPIVLT
SCLVDTKNNWAIIGRDALQQCQGVLYLPEAKGPPVILPIQAPAVLGLEHLPRPPQISQFP
LNQNASRPCNTWSGRPWRQAISNPTPGQEITQYSQLKRPMEPGDSSTTCGPLTL

17. HTLV10

env polyprotein - T-cell leukemia virus (HTLV-II)

MGNVFFLLLFSLTHFPLAQQSRCTLTIGISSYHSSPCSPTQPVCTWNLDLNSLTTDQRLH
PPCPNLITYSGFHKTYSLYLFPHWIKKPNRQGLGYYSPSYNDPCSLQCPYLGCQAWTSAY
TGPVSSPSWKFHSDVNFTQEVSQVSLRLHFSKCGSSMTLLVDAPGYDPLWFITSEPTQPP
PTSPPLVHDSDLEHVLTPSTSWTTKILKFIQLTLQSTNYSCMVCVDRSSLSSWHVLYTPN
ISIPQQTSSRTILFPSLALPAPPSQPFPWTHCYQPRLQAITTDNCNNSIILPPFSLAPVP
PPATRRRRAVPIAVWLVSALAAGTGIAGGVTGSLSLASSKSLLLEVDKDISHLTQAIVKN
HQNILRVAQYAAQNRRGLDLLFWEQGGLCKAIQEQCCFLNISNTHVSVLQERPPLEKRVI
TGWGLNWDLGLSQWAREALQTGITILALLLLVILFGPCILRQIQALPQRLQNRHNQYSLI
NPETML

18. HTLV11

nef protein - AIDS virus HTLV-III (T-cell leukemia virus BH10)

MGGKWSKSSVVGWPAVRERMRRAEPAADGVGAASRDLEKHGAITSSNTAANNADCAWLEA
QEEEEVGFPVTPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIHSQRRQDILDLWIYHTQGY
FPDQNYTPGPGIRYPLTFGWCYKLVPVEPEKLEEANKGENTSLLHPVSLHGMDDPEREVL
EWRFDSRLAFHHMARELHPEYFKNC

19. HTLV12

nef protein - AIDS virus HTLV-III (T-cell leukemia virus 12)

WKGFCYKMGGKWSKSSVVGWPAVRERMRRAEPAADGVGAASRDLEKHGAITSSNTAANNA
ACAWLEAQEEEKVGFPVTPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIHSQRRQDILDLW
IYHTQGYFPDWQNYTPGPGIRYPLTFGWCYKLVPVEPEKLEEANKGENTSLLHPVSLHGM
DDPEREVLEWRFDSRLAFHHVARELHPEYFKNC

20. HTLV2

Protease (EC 3.4.21.) - T-cell leukemia virus I (HTLV-I, Caribbean isolate)

KTPPKSAVLPVRESRPLESGLHSASSSPWAMPPMSRSNSLEARLPPPKAHYPRTRARGGC
PPIRSPRRHPTPKKLHRGGGLTSPPTLQQVLPNQDPTSILPVIPLDPARRPVIKAQIDTQ
TSHPKTIEALLDTGADMTVLPIALFSSNTPLKNTSVLGAGGQTQDHFKLTSLPVLIRLPF
RTTPIVLTSCLVDTKNNWAIIGRDALQQCQGVLYLPEAKRPPVILPIQAPAVLGLEHLPR
PPEISQFPLNQNASRPCNTWSGRPWRQAISNPTPGQEITQYSQLKKPMEPGDSSTTCGPL
TL

21. HTLV3

gag polyprotein - T-cell leukemia virus I (HTLV-I, Caribbean isolate)

MGQIFSRSASPIPRPPRGLAAHHWLNFLQAAYRLEPGPSSYDFHQLKKFLKIALETPVWI
CPINYSLLASLLPKGYPGRVNEILHILIQTQAQIPSRPAPPPPSSSTHDPPDSDPQIPPP

YVEPTAPQVLPVMHPHGAPPNHRPWQMKDLQAIKQEVSQAAPGSPQFMQTIRLAVQQFDP
TAKDLQDLLQYLCSSLVASLHHQQLDSLISEAETRGITGYNPLAGPLRVQANNPQQQGLR
REYQQLWLAAFAALPGSAKDPSWASILQGLEEPYHAFVERLNIALDNGLPEGTPKDPILR
SLAYSNANKECQKLLQARGHTNSPLGDMLRACQAWTPKDKTKVLVVQPKKPPPNQPCFRC
GKAGHWSRDCTQPRPPPGPCPLCQDPTHWKRDCPRLKPTIPEPEPEEDALLLDLPADIPH
PKNSIGGEV

22. HTLV4

gag polyprotein - T-cell leukemia virus (HTLV-II)

MGQIHGLSPTPIPKAPRGLSTHHWLNFLQAAYRLQPRPSDFDFQQLRRFLKLALKTPIWL
NPIDYSLLASLIPKGYPGRVVEIINILVKNQVSPSAPAAPVPTPICPTTTPPPPPPPSPE
AHVPPPYVEPTTTQCFPILHPPGAPSAHRPWQMKDLQAIKQEVSSSALGSPQFMQTLRLA
VQQFDPTAKDLQDLLQYLCSSLVVSLHHQQLNTLITEAETRGMTGYNPMAGPLRMQANNP
AQQGLRREYQNLWLAAFSTLPGNTRDPSWAAILQGLEEPYCAFVERLNVALDNGLPEGTP
KEPILRSLAYSNANKECQKILQARGHTNSPLGEMLRTCQAWTPKDKTKVLVVQPRRPPPT
QPCFRCGKVGHWSRDCTQPRPPPGPCPLCQDPSHWKRDCPQLKPPQEEGEPLLLDLPSTS
GTTEEKNSLRGEI

23. HTLV5

gag polyprotein - AIDS virus HTLV-III (T-cell leukemia virus, BH10)

MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERFAVNPGLLETSEGCRQI
LGQLQPSLQTGSEELRSLYNTVATLYCVHQRIEIKDTKEALDKIEEEQNKSKKKAQQAAA
DTGHSSQVSQNYPIVQNIQGQMVHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGAT
PQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPVHAGPIAPGQMREPRGSDIAGTT
STLQEQIGWMTNNPPIPVGEIYKRWIILGLNKIVRMYSPTSILDIRQGPKEPFRDYVDRF
YKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGVGGPGHKA
RVLAEAMSQVTNTATIMMQRGNFRNQRKMVKCFNCGKEGHTARNCRAPRKKGCWKCGKEG
HQMKDCTERQANFLGKIWPSYKGRPGNFLQSRPEPTAPPFLQSRPEPTAPPEESFRSGVE
TTTPPQKQEPIDKELYPLTSLRSLFGNDPSSQ

24. HTLV6

pol polyprotein - T-cell leukemia virus I (HTLV-I, Caribbean isolate)

GKKAACNLANTGASCPWARTPPKAPRNQPVPFKPERLQALQHLVRKALEAGHIEPYTGPG
NNPVFPVKKANGTWRFIHDLRATNSLTIDLSSSSPGPPDLSSLPTTLAHLQTIDLKDAFF
QIPLPKQFQPYFAFTVPQQCNYGPGTRYAWRVLPQGFKNSPTLFEMQLAHILQPIRQAFP
QCTILQYMDDILLASPSHADLQLLSEATMASLISHGLPVSENKTQQTPGTIKFLGQIISP
NHLTYDAVPKVPIRSRWALPELQALLGEIQWVSKGTPTLRQPLHSLYCALQRHTDPRDQI
YLNPSQVQSLVQLRQALSQNCRSRLVQTLPLLGAIMLTLTGTTTVVFQSKQQWPLVWLHA
PLPHTSQCPWGQLLASAVLLLDKYTLQSYGLLCQTIHHNISTQTFNQFIQTSDHPSVPIL
LHHSHRFKNLGAQTGELWNTFLKTTAPLAPVKALMPVFTLSPVIINTAPCLFSDGSTSQA
AYILWDKHILSQRSFPLPPPHKSAQRAELLGLLHGLSSARSWRCLNIFLDSKYLYHYLRT
LALGTFQGRSSQAPFQALLPRLLSRKVVYLHHVRSHTNLPDPISRLNALTDALLITPVLQ
LSPADLHSFTHCGQTALTLQGATTTEASNILRSCHACRKNNPQHQMPQGHIRRGLLPNHI
WQGDITHFKYKNTLYRLHVWVDTFSGAISATQKRKETSSEAISSLLQAIAYLGKPSYINT
DNGPAYISQDFLNMCTSLAIRHTTHVPYNPTSSGLVERSNGILKTLLYKYFTDKPDLPMD
NALSIALWTINHLNVLTNCHKTRWQLHHSPRLQPIPETHSLSNKQTHWYYFKLPGLNSRQ
WKGPQEALQEAAGAALIPVSASSAQWIPWRLLKRAACPRPVGGPADPKEKDHQHHG

25. HTLV7

pol polyprotein - T-cell leukemia virus (HTLV-II)

HRSRPYGYTPDTRARAGKAPRHPDPRRQWANQHPVQTPPNPPTHILALPKVPRYPFLLPL
RHPQQMDHHWKGRPTTMPGASIPPRRPQPPPIAANSHSKHHRPRTPSPTSPSGPISFKPE
RLQALNDLVSKALEAGHIEPYSGPGNNPVFPVKKPNGKWRFIHDLRATNAITTTLTSPSP
GPPDLTSLPTALPHLQTIDLTDAFFQIPLPKQYQPYFAFTIPQPCNYGPGTRYAWTVLPQ
GFKNSPTLFEQQLAAVLNPMRKMFPTSTIVQYMDDILLASPTNEELQQLSQLTLQALTTH
GLPISQEKTQQTPGQIRFLGQVISPNHITYESTPTIPIKSQWTLTELQVILGEIQWVSKG
TPILRKHLQSLYSALHGYRDPRACITLTPQQLHALHAIQQALQHNCRGRLNPALPLLGLI
SLSTSGTTSVIFQPKQNWPLAWLHTPHPPTSLCPWGHLLACTILTLDKYTLQHYGQLCQS
FHHNMSKQALCDFLRNSPHPSVGILIHHMGRFHNLGSQPSGPWKTLLHLPTLLQEPRLLR
PIFTLSPVVLDTAPCLFSDGSPQKAAYVLWDQTILQQDITPLPSHETHSAQKGELLALIC
GLRAAKPWPSLNIFLDSKYLIKYLHSLAIGAFLGTSSHQTLQAALPPLLQGKTIYLHHVR
SHTNLPDPISTFNEYTDSLILAPLVPLTPQGLHGLTHCNQRALVSFGATPREAKSLVQTC
HTCQTINSQHHMPRGYIRRGLLPNHIWQGDVTHYKYKKYKYCLHVWVDTFSGAVSVSCKK
KETSCETISAVLQAISLLGKPLHINTDNGPAFLSQEFQEFCTSYRIKHSTHIPYNPTSSG
LVERTNGVIKNLLNKYLLDCPNLPLDNAIHKALWTLNQLNVMNPSGKTRWQIHHSPPLPP

IPEASTPPKPPPKWFYYKLPGLTNQRWKGPLQSLQEAAGAALLSIDGSPRWIPWRFLKKA
ACPRPDASELAEHAATDHQHHG

26. HTLV8

env polyprotein precursor - AIDS virus HTLV-III (T-cell leukemia virus, BH10)

TEKLWVTVYYGVPVWKEATTTLFCASDAKAYDTEVHNVWATHACVPTDPNPQEVVLVNVT
ENFNMWKNDMVEQMHEDIISLWDQSLKPCVKLTPLCVSLKCTDLKNDTNTNSSSGRMIME
KGEIKNCSFNISTSIRGKVQKEYAFFYKLDIIPIDNDTTSYTLTSCNTSVITQACPKVSF
EPIPIHYCAPAGFAILKCNNKTFNGTGPCTNVSTVQCTHGIRPVVSTQLLLNGSLAEEEV
VIRSANFTDNAKTIIVQLNQSVEINCTRPNNNTRKSIRIQRGPGRAFVTIGKIGNMRQAH
CNISRAKWNNTLKQIDSKLREQFGNNKTIIFKQSSGGDPEIVTHSFNCGGEFFYCNSTQL
FNSTWFNSTWSTKGSNNTEGSDTITLPCRIKQIINMWQEVGKAMYAPPISGQIRCSSNIT
GLLLTRDGGNSNNESEIFRPGGGDMRDNWRSELYKYKVVKIEPLGVAPTKAKRRVVQREK
RAVGIGALFLGFLGAAGSTMGAASMTLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTV
WGIKQLQARILAVERYLKDQQLLGIWGCSGKLICTTAVPWNASWSNKSLEQIWNNMTWME
WDREINNYTSLIHSLIEESQNQQEKNEQELLELDKWASLWNWFNITNWLWYIKLFIMIVG
GLVGLRIVFAVLSVVNRVRQGYSPLSFQTHLPIPRGPDRPEGIEEEGGERDRDRSIRLVN
GSLALIWDDLRSLCLFSYHRLRDLLLIVTRIVELLGRRGWEALKYWWNLLQYWSQELKNS
AVSLLNATAIAVAEGTDRVIEVVQGAYRAIRHIPRRIRQGLERILL

27. HTLV9

env polyprotein - T-cell leukemia virus I (HTLV-I, Caribbean isolate)

MGKFLATLILFFQFCPLILGDYSPSCCTLTVGVSSYHSKPCNPAQPVCSWTLDLLALSAD
QALQPPCPNLVSYSSYHATYSLYLFPHWIKKPNRNGGGYYSASYSDPCSLKCPYLGCQSW
TCPYTGAVSSPYWKFQQDVNFTQEVSHLNINLHFSKCGFSFSLLVDAPGYDPIWFLNTEP
SQLPPTAPPLLSHSNLDHILEPSIPWKSKLLTLVQLTLQSTNYTCIVCIDRASLSTWHVL
YSPNVSVPSPSSTPLLYPSLALPAPHLTLPFNWTHCFDPQIQAIVSSPCHNSLILPPFSL
SPVPTLGSRSRRAVPVAVWLVSALAMGAGVAGRITGSMSLASGKSLLHEVDKDISQLTQA
IVKNHKNLLKIAQYAAQNRRGLDLLFWEQGGLCKALQEQCCFLNITNSHVSILQERPPLE
NRVLTGWGLNWDLGLSQWAREALQTGITLVALLLLVILAGPCILRQLRHLPSRVRYPHYS
LINPESSL

28. PENJAN1

Penicillopepsin (EC 3.4.23.6) - Penicillium janthinellum

AASGVATNTPTANDEEYITPVTIGGTTLNLNFDTGSADLWVFSTELPASQQSGHSVYNPS
ATGKELSGYTWSISYGDGSSASGNVFTDSVTVGGVTAHGQAVQAAQQISAQFQQDTNNDG
LLGLAFSSINTVQPQSQTTFFDTVKSSLAQPLFAVALKHQQPGVYDFGFIDSSKYTGSLT
YTGVDNSQGFWSFNVDSYTAGSQSGDGFSGIADTGTTLLLLBDSVVSQYYSQVSGAQQDS
NAGGYVFDCSTNLPDFSVSISGYTATVPGSLINYGPSGDGSTCLGGIQSNSGIGFSIFGD
IFLKSQYVVFDSDGPQLGFAPQA

29. PEPASP1

Z1211227A: aspergillopepsin A

SKGSAVTTPQNNDEEYLTPVTVGKSTLHLDFDTGSADLWVFSDELPSSERTGHNVYTPSS
SATKLSGYTWNISYGNGSSASGDVYRDTVTVGGVTNTKEAVQAASKISSEFZZVBGGZBS
GAZAYSSINTVQPKAQTTFFDTVKSQLNSPLFAVQLKHDAPGVYDFGYIBBSKYTGSITY
TDADSSEGYWGFNPNGYSIGDSSSSGFSAIADTGTTLILLDDEIVLNGSZVSGQANQEAD
GGYVFBCSTTPPDFTGXIGDYKAVGPKYINYAPSBTPSTCFGGIQSNSGLGLSILGDVFL
KSQYVVFDSQGPKLGFAAQA

30. PEPEND1

Z1311269A: endothiapepsin

STGSATTTPIDSLDDAYITPVQIGTPAQTLNLDFDTGSSDLWVFSSETTASEVDGQTIYT
PSKSTTAKLLSGATWSISYGDGSSSSGDVYTDTVSVGGLTVTGQAVESAKKVSSSFTEDS
TIDGLLGLAFSTLNTVSPTQQKTFFDNAKASLDSPVFTADLGYHAPGTYNFGFIDTTAYT
GSITYTAVSTKQGFWEWTSTGYAVGSGTFKSTSIDGIADTGTTLLYLPATVVSAYWAQVS
GAKSSSSVGGYVFPCSATLPSFTFGVGSARIVIPGDYIDFGPISTGSSSCFGGIQSSAGI
GINIFGDVALKAAFVVFNGATTPTLGFASK

31. PEPEND2

E70087:ACID PROTEINASE (ENDOTHIAPEPSIN)

AGTTTTTTIDSLDDAYIVVIQLGTPAATLNLNFDTGSADLWVFAGGSSNNSGHNVYNSAA
SSASATLSTGKWSIAYAAGASASGLNRDGTVTVGGVTAHGQAVQAAAAAFAAQATNDGLL
GLAFSSINTVQGSQNLFFVQGKSLAQTPLFSVYLAHQQGVYDFGLIDSSAYTGSLNYTGV
ASQGFWSFNVDSYTAGSGSTIGDVSSGIADTGTTVLYLPTSVVSAYYSQVSGATDSAAGG
LDITCDGLTAMPAGVGSTKASVPLSTAYGGQGNGSSGCLGGIQSAAGIGLLIFGDIFIKA

SYVVFDSNNPNLGFAPAA

32. PEPHEN1

Pepsinogen A (EC 3.4.23.1) - Chicken

TATESYEPMTNYMDASYYGTISIGTPQQDFSVIFDTGSSNLWVPSIYCKSSACSNHKRFD
PSKSSTYVSTNETVYIAYGTGSMSGILGYDTVAVSSIDVQNQIFGLSETEPGSFFYYCNF
DGILGLAFPSISSSGATPVFDNMMSQHLVAQDLFSVYLSKDGETGSFVLFGGIDPNYTTK
GIYWVPLSAETYWQITMDRVTVGNKYVACFFTCQAIVDTGTSLLVMPQGAYNRIIKDLGV
SSDGEISCDDISKLPDVTFHINGHAFTLPASAYVLNEDGSCMLGFENMGTPTELGEQWIL
GDVFIREYYVIFDRANNKVGLSPLS

33. PEPHIZ1

Z1402278A: hizopuspepsin I

AGVGTVPMTDYGNDIEYYGQVTIGTPGKKFNLDFDTGSSDLWIASTLCTNCGSRQTKYDP
NQSSTYQADGRTWSISYGDGSSASGILAKDNVNLGGLLIKGQTIELAKREAASFASGPND
GLLGLGFDTITTVRGVKTPMDNLISQGLISRPIFGVYLGKAKNGGGGEYIFGGYDSTKFK
GSLTTVPIDNSRGWWGITVDRATVGTSTVASSFDGILDTGTTLLILPNNIAASVARAYGA
YDNGDGTYTISCDTSRFKPLVFSINGASFQVSPDSLVFEEYQGQCIAGFGYGDFDFAIIG
DTFLKNNYVVFNQGVPEVQIAPVAE

34. PEPHM1

Pepsinogen A (EC 3.4.23.1) precursor - Human

VDEQPLENYLDMEYFGTIGIGTPAQDFTVVFDTGSSNLWVPSVYCSSLACTNHNRFNPED
SSTYQSTSETVSITYGTGSMTGILGYDTVQVGGISDTNQIFGLSETEPGSFLYYAPFDGI
LGLAYPSISSSGATPVFDNIWNQGLVSQDLFSVYLSADDQSGSVVIFGGIDSSYYTGSLN
WVPVTVEGYWQITVDSITMNGEAIACAEGCQAIVDTGTSLLTGPTSPIANIQSDIGASEN
SDGDMVVSCSAISSLPDIVFTINGVQYPVPPSAYILQSEGSCISGFQGMNLPTESGELWI
LGDVFIRQYFTVFDRANNQVGLAPVA

35. PEPMON1

Pepsinogen A (EC 3.4.23.1) precursor - Rhesus macaque

IDEQPLENYLDVEYFGTIGIGTPAQDFTVIFDTGSSNLWVPSVYCSSLACTNHNLFNPQD
SSTYQSTSGTLSITYGTGSMTGILGYDTVQVGGISDTNQIFGLSETEPGSFLYYAPFDGI
LGLAYPSISSSGATPVFDNIWDQGLVSQDLFSVYLSADDQSGSVVIFGGIDSSYYTGSLN
WVPVSVEGYWQISVDSITMNGEAIACAEGCQAIVDTGTSLLTGPTSPIANIQSDIGASEN
SDGEMVVSCSAISSLPDIVFTINGVQYPLPPSAYILQSQGSCTSGFQGMDVPTESGELWI
LGDVFIRQYFTVFDRANNQVGLAPVA

36. PEPMON3

Pepsinogen C (EC 3.4.23.3) - Japanese macaque

SVSYEPMAYMDAAYFGEISIGTPPQNFLVLFDTGSSNLWVPSVYCQSQACTSHSRFNPSE
SSTYSTNGQTFSLQYGSGSLTGFFGYDTLTVQSIQVPNQEFGLSENEPGTNFVYAQFDGI
MGLAYPTLSVDGATTAMQGMVQEGALTSPIFSVYLSDQQGSSGGAVVFGGVDSSLYTGQI
YWAPVTQELYWQIGIEEFLIGGQASGWCSEGCQAIVDTGTSLLTVPQQYMSALLQATGAQ
EDEYGQFLVNCNSIQNLPTLTFIINGVEFPLPPSSYILNNNGVCTVGVEPTYLSAQNSQP
LVYWILGDVFLRSYYSVYDLSNNRVGFATAA

37. PEPPIG2

E740190A: pepsin

IGDEPLENYLDTEYFGTIGIGTPAQDFTVIFDTGSSNLWVPSVYCSSLACSDHNQFNPDD
SSTFEATSQELSITYGTGSMTGILGYDTVQVGGISDTNQIFGLSETEPGSFLYYAPFDGI
LGLAYPSISASGATPVFDNLWDQGLVSQDLFSVYLSSNDDSGSVVLLGGIDSSYYTGSLN
WVPVSVEGYWQITLDSITMDGETIACSGGCQAIVDTGTSLLTGPTSAIANIQSDIGASEN
SDGEMVISCSSIDSLPDIVFTINGVQYPLSPSAYILQDDDSCTSGFEGMDVPTSSGELWI
LGDVFIRQYYTVFDRANNKVGLAPVA

38. PEPPROA

yeast proteinase a jbc 1987

GGHDVPLTNYLNAQYYTDITLGTPPQNFKVILDTGSSNLWVPSNECGSLACFLHSKYDHE
ASSSYKANGTEFAIQYGTGSLEGYISQDTLSIGDLTIPKQDFAEATSEPGLTFAFGKFDG
ILGLGYDTISVDKVVPPFYNAIQQDLLDEKRFAFYLGDTSKTENGGEATFGGIDESKFKG
DITWLPVRRKAYWEVKFEGIGLGDEYAELESHGAAIDTGTSLITLPSGLAEMINAEIGAK
KGWTGQYTLDCNTRDNLPDLIFNFNGYNFTIGPYDYTLEVSGSCISAITPMDFPEPVGPL
AIVGDAFLRKYYSIYDLGNNAVGLAKAILKNNYVVFNQGVPEVQIAPVAQ

39. PEPRAT3

PEPC$RAT: GASTRICSIN PRECURSOR (EC 3.4.23.3) (PEPSINOGEN C). 392 AA.

SVLYEPMAYMDASYFGEISIGTPPQNFLVLFDTGSSNLWVSSVYCQSEACTTHARFNPSK
SSTYYTEGQTFSLQYGTGSLTGFFGYDTLTVQSIQVPNQEFGLSENEPGTNFVYAQFDGI

MGLAYPGLSSGGATTALQGMLGEGALSQPLFGVYLGSQQGSNGGQIVFGGVDKNLYTGEI
TWVPVTQELYWQITIDDFLIGDQASGWCSSQGCQGIVDTGTSLLVMPAQYLSELLQTIGA
QEGEYGEYFVSCDSVSSLPTLSFVLNGVQFPLSPSSYIIQEDNFCMVGLESISLTSESGQ
PLWILGDVFLRSYYAIFDMGNNKVGLATSV

40. PEPRHZ2

E70086: ACID PROTEINASE EC 3.4.23.9
GVGTVPMTDYGNDVEYYGQVTIGTPGKSFNLNFDTGSSNLWVGSVQCQASGCKGGRDKFN
PSDGSTFKATGYDASIGYGDGSASGVLGYDTVQVGGIDVTGGPQIQLAQRLGGGGFPGDN
DGLLGLGFDTLSITPQSSTNAFDQVSAQGKVIQPVFVVYLAASNISDGDFTMPGWIDNKY
GGTLLNTNIDAGEGYWALNVTGATADSTYLGAIFQAILDTGTSLLILPDEAAVGNLVGFA
GAQDAALGGFVIACTSAGFKSIPWSIYSAIFEIITALGNAEDDSGCTSGIGASSLGEAIL
GDQFLKQQYVVFDRDNGIRLAPVA

41. RENHM1

RENI$HUMAN: RENIN PRECURSOR, RENAL (EC 3.4.23.15) (GENE NAME: REN). 406 AA.
LTLGNTTSSVILTNYMDTQYYGEIGIGTPPQTFKVVFDTGSSNVWVPSSKCSRLYTACVY
HKLFDASDSSSYKHNGTELTLRYSTGTVSGFLSQDIITVGGITVTQMFGEVTEMPALPFM
LAEFDGVVGMGFIEQAIGRVTPIFDNIISQGVLKEDVFSFYYNRDSENSQSLGGQIVLGG
SDPQHYEGNFHYINLIKTGVWQIQMKGVSVGSSTLLCEDGCLALVDTGASYISGSTSSIE
KLMEALGAKKRLFDYVVKCNEGPTLPDISFHLGGKEYTLTSADYVFQESYSSKKLCTLAI
HAMDIPPPTGPTWALGATFIRKFYTEFDRRNNRIGFALAR

42. RENMSE1

Z0809326A: renin
SSLTDLISPVVLTNYLNSQYYGEIGIGTPPQTFKVIFDTGSANLWVPSTKCSRLYLACGI
HSLYESSDSSSYMENGDDFTIHYGSGRVKGFLSQDSVTVGGITVTQTFGEVTELPLIPFM
LAQFDGVLGMGFPAQAVGGVTPVFDHILSQGVLKEKVFSVYYNRGPHLLGGEVVLGGSDP
EHYQGDFHYVSLSKTDSWQITMKGVSVGSSTLLCEEGCEVVVDTGSSFISAPTSSLKLIM
QALGAKEKRLHEYVVSCSQVPTLPDISFNLGGRAYTLSSTDYVLQYPNDKLCTVALHAMD
IPPPTGPVWVLGATFIRKFYTEFDRHNNRIGFALAR

43. RENMSE2

Gamma-renin (EC 3.4.23.15) precursor - Mouse
IVGGFKCEKNSQPWQVAVYYHKEHICGGVLLDRNWVLTAAHCYVDECEVWLGKNQLFQEE
PSAQNRLVSKSFPHPGFNMTLLTFEKLPPGADFSNDLMLLRLSKPADITDVVKPIDLPTK
EPKLDSTCLVSGWGSITPTKWQKPDDLQCMFTKLLPNENCAKAYLLKVTDVMLCTIEMGE
DKGPCVGDSGGPLICDGVLQGTVSIGPDPCGIPGVSAIYTNLVKFNSWIKDTMMKNA

44. RENMSE3

RENI$MOUSE: RENIN PRECURSOR, RENAL (EC 3.4.23.15) (GENE NAME: REN1). 402 AA.
PSLTNLTSPVVLTNYLNTQYYGEIGIGTPPQTFKVIFDTGSANLWVPSTKCSRLYLACGI
HSLYESSDSSSYMENGSDFTIHYGSGRVKGFLSQDSVTVGGITVTQTFGEVTELPLIPFM
LAKFDGVLGMGFPAQAVGGVTPVFDHILSQGVLKEEVFSVYYNRGSHLLGGEVVLGGSDP
QHYQGNFHYVSISKTDSWQITMKGVSVGSSTLLCEEGCAVVVDTGSSFISAPTSSLKLIM
QALGAKEKRIEEYVVNCSQVPTLPDISFDLGGRAYTLSSTDYVLQYPNRRDKLCTLALHA
MDIPPPTGPVWVLGATFIRKFYTEFDRHNNRIGFALAR

45. RENRAT1

Z1310378A: renin
KSSFTNVTSPVVLTNYLDTQYYGEIGIGTPSQTFKVIFDTGSANLWVPSTKCGPLYTACE
IHNLYDSSESSSYMENGTEFTIHYGSGKVKGFLSQDVVTVGGIIVTQTFGEVTELPLIPF
MLAKFDGVLGMGFPAQAVDGVIPVFDHILSHEVLKEEVFSVYYSRESHLLGGEVVLGGSD
PQHYQGNFHYVSISKAGSWQITMKGVSVGPATLLCEEGCMAVVDTGTSYISGPTSSLQLI
MQALGVKEKRANNYVVNCSQVPTLPDISFYLGGRTYTLSNMDYVQKNPFRNDDLCILALQ
GLDIPPPTGPVWVLGATFIRKFYTEFDRHNNRIGFALAR

46. RENRHZ1

Aspartic proteinase (EC 3.4.23.6) - Rhizomucor pusillus
AEGDGSVDTPGLYDFDLEEYAIPVSIGTPGQDFYLLFDTGSSDTWVPHKGCDNSEGCVGK
RFFDPSSSSTFKETDYNLNITYGTGGANGIYFRDSITVGGATVKQQTLAYVDNVSGPTAE
QSPDSELFLDGIFGAAYPDNTAMEAEYGDTYNTVHVNLYKQGLISSPVFSVYMNTNDGGG
QVVFGGVNNTLLGGDIQYTDVLKSRGGYFFWDAPVTGVKIDGSDAVSFDGAQAFTIDTGT
NFFIAPSSFAEKVVKAALPDATESQQGYTVPCSKYQDSKTTFSLVLQKSGSSSDTIDVSV
PISKMLLPVDKSGETCMFIVLPDGGNQFIVGNLFLRFFVNVYDFGKNRIGFAPLASGYEN
D

47. RENRHZ2
CARP$RHIMI: ASPARTATE PROTEASE PRECURSOR (EC 3.4.23.6) (MUCOR RENNIN).
430 AA.
AAADGSVDTPGYYDFDLEEYAIPVSIGTPGQDFLLLFDTGSSDTWVPHKGCTKSEGCVGS
RFFDPSASSTFKATNYNLNITYGTGGANGLYFEDSIAIGDITVTKQILAYVDNVRGPTAE
QSPNADIFLDGLFGAAYPDNTAMEAEYGSTYNTVHVNLYKQGLISSPLFSVYMNTNSGTG
EVVFGGVNNTLLGGDIAYTDVMSRYGGYYFWDAPVTGITVDGSAAVRFSRPQAFTIDTGT
NFFIMPSSAASKIVKAALPDATETQQGWVVPCASYQNSKSTISIVMQKSGSSSDTIEISV
PVSKMLLPVDQSNETCMFIILPDGGNQYIVGNLFLRFFVNVYDFGNNRIGFAPLASAYEN
E

TRYPSIN/CHYMOTRYPSIN (15 sequences)
Full description is listed previously

1. TRPCRAY1
2. TRPDOG1
3. TRPDOG2
4. TRPFISH1
5. TRPFISH2
6. TRPFLY1
7. TRPHM1
8. TRPHOR1
9. TRPHOR2
10. TRPMSE1
11. TRPPIG1
12. TRPRAT1
13. TRPRAT2
14. TRPRAT3
15. TRPSTRP1

AIDS RELATED PROTEINS (25 sequences)
Full description is listed previously

1. AID1
2. AID10
3. AID11
4. AID12
5. AID13
6. AID2
7. AID3
8. AID4
9. AID5
10. AID6
11. AID7
12. AID8
13. AID9
14. HTLV1
15. HTLV10
16. HTLV11
17. HTLV12
18. HTLV2
19. HTLV3
20. HTLV4
21. HTLV5
22. HTLV6
23. HTLV7

24. HTLV8
25. HTLV9

PROTEASE INHIBITORS (26 sequences)

1. ANTIASP1
Proteinase inhibitor - Eggplant
QICTNNCAGRKGCSYFSEDGTFICKGESNPENPKACPRNCDGRIAYGICPLS
2. ANTIASP2
Proteinase inhibitor PTI - Potato
RICTNCCAGYKGCNYYSANGAFICEGESDPKNPNVCPRNCDTNIAYSKCLR
3. ANTIASP3
Proteinase inhibitor PCI-I - Potato
PICTNCCAGYKGCNYYSANGAFICEGQSDPKKPKACPLNCDPHIAYSKCPRS
4. ANTIASP4
Proteinase A inhibitor 3 - Yeast (Saccharomyces cerevisiae)
MNTDQQKVSEIFQSSKEKLQGDAKVVSDAFKKMASQDKDGKTTDADESEKHNYQEQYNKL
KGAGHKKE
5. ANTIMET1
Metalloproteinase inhibitor precursor - Human
CTCVPPHPQTAFCNSDLVIRAKFVGTPEVNQTTLYQRYEIKMTKMYKGFQALGDAADIRF
VYTPAMESVCGYFHRSHNRSEEFLIAGKLQDGLLHITTCSFVAPWNSLSLAQRRGFTKTY
TVGCEECTVFPCLSIPCKLQSGTHCLWTDQLLQGSEKGFQSRHLACLPREPGLCTWQSLR
SQIA
6. ANTIMET2
Metalloproteinase inhibitor - Streptomyces nigrescens
APSCPAGSLCTYSGTGLSGARTVIPASDMEKAGTDGVKLPASARSFANGTHFTLRYGPAR
KVTCVRFPCYQYATVGKVAPGAQLRSLPSPGATVTVGQDLGD
7. ANTIP10
Venom basic protease inhibitor I - Western sand viper
QDHPKFCYLPADPGRCKAHIPRFYYDSASNKCNKFIYGGCPGNANNFKTWDECRQTCGAS
A
8. ANTIP11
Venom basic protease inhibitor III - Sand viper
RDRPKFCYLPADPGRCLAYMPRFYYNPASNKCEKFIYGGCRGNANNFKTWDECRHTCVAS
GIQPR
9. ANTIP12
Basic protease inhibitor - Red sea turtle
QGDKRDICRLPPEQGPCKGRIPRYFYNPASRMCESFIYGGCKGNKNNFKTKAECVRACRP
PERPGVCPKTSGPGICLHGCDSDSDCKEGQKCCFDGCGYICLTVAPSGSP
10. ANTIP13
Protease inhibitor (PSTI type), submandibular - Dog
GPPPAIGRGVBCSBYKGKGSEIACPRLHZPICGTDHKTYSNZCMLCAFTLDKKFZVRKLQ
DTACDIECTEYSDMCTMDYRPLCGSDGKNYSNKCSFCNAVKKSRGTIFLAKHGEC
11. ANTIP14
Protease B inhibitors 2 and 1 - Yeast (Saccharomyces cerevisiae)
TKNFIVTLKKNTPDVEAKKFLDSVHHAGGSILHEFDIIKGYTIKVPDVLHLNKLKEKHND
VIENVEEDKEVHTN
12. ANTIPRO1
Serum basic protease inhibitor - Bovine
TERPDFCLEPPYTGPCKAAMIRYFYNAKAGFCETFVYGGCRAKSNNFKSAEDCMRTCGGA
13. ANTIPRO2
Venom basic protease inhibitor I - Black mamba
QPLRKLCILHRNPGRCYQKIPAFYYNQKKKQCZGFTWSGCGGNSNRFKTIEECRRTCIRK
14. ANTIPRO3
Venom basic protease inhibitor I homolog - Eastern green mamba
QPRRKLCILHRNPGRCYDKIPAFYYNQKKKQCERFDWSGCGGNSNRFKTIEECRRTCIG

15. ANTIPRO4
Venom basic protease inhibitor K - Black mamba and eastern green mamba
AAKYCKLPLRIGPCKRKIPSFYYKWKAKQCLPFDYSGCGGNANRFKTIEECRRTCVG
16. ANTIPRO5
Venom basic protease inhibitor B - Black mamba
RPYACELIVAAGPCMFFISAFYYSKGANKCYPFTYSGCRGNANRFKTIEECRRTCVV
17. ANTIPRO6
Venom basic protease inhibitor E - Black mamba
LQHRTFCKLPAEPGPCKASIPAFYYNWAAKKCQLFHYGGCKGNANRFSTIEKCRHACVG
18. ANTIPRO7
Venom basic protease inhibitor II - Ringhals
RPDFCELPAETGLCKAYIRSFHYNLAAQQCLQFIYGGCGGNANRFKTIDECRRTCVG
19. ANTIPRO8
Venom basic protease inhibitor II - Cape cobra
RPRFCELPAETGLCKARIRSFHYNRAAQQCLEFIYGGCGGNANRFKTIDECHRTCVG
20. ANTIPRO9
Venom basic protease inhibitor II - Russell's viper
HDRPTFCNLAPESGRCRGHLRRIYYNLESNKCKVFFYGGCGGNANNFETRDECRETCGGK
21. ANTISER1
Antileukoproteinase 1 - Human
SGKSFKAGVCPPKKSAQCLRYKKPECQSDWQCPGKKRCCPDTCGIKCLDPVDTPNPTRRK
PGKCPVTYGQCLMLNPPNFCEMDGQCKRDLKCCMGMCGKSCVSPVKA
22. ANTISER2
Acrosin inhibitor I (PSTI type) - Bovine
EIYFEPDFGFPPDCKVYTEACTREYNPICDSAAKTYSNECTFCNEKMNNDADIHFNHFGE
CEY
23. ANTISER3
Serine proteinase inhibitor 1 - Vaccinia virus
MDIFKELILKHTDENVLISPVSILSTLSILNHGAAGSTAEQLSKYIENMNENTPDDNNDM
DVDIPYCATLATANKIYGSDSIEFHASFLQKIKDDFQTVNFNNANQTKELINEWVKTMTN
GKINSLLTSPLSINTRMTVVSAVHFKAMWKYPFSKHLTYTDKFYISKNIVTSVDMMVSTE
NNLQYVHINELFGGFSIIDIPYEGNSSMVIILPDDIEGIYNIEKNITDEKFKKWCGMLST
KSIDLYMPKFKVEMTEPYNLVPILENLGLTNIFGYYADFSKMCNETITVEKFLHTTFIDV
NEEYTEASAVTGVFMTNFSMVYRTKVYINHPFMYMIKDNTGRILFIGKYCYPQ
24. ANTISER4
Ovomucoid (PSTI-type protease inhibitor) 1 - Japanese quail
VEVDCSRFPNTTNEEGKDEVVCPDELRLICGTDGVTYNHECMLCFYNKEYGTNISKEQDG
ECGETVPMDCSRYPNTTSEDGKVTILCTKDFSFVCGTDGVTYDNECMLCAHNVVQGTSVG
KKHDGECRKELAAVSVDCSEYPKPACPKDYRPVCGSDNKTYSNKCNFCNAVVESNGTLTL
NHFGKC
25. ANTISER7
Ovomucoid (PSTI-type protease inhibitor) precursor - Chicken
MAMAGVFVLFSFVLCGFLPDAAFGAEVDCSRFPNATDKEGKDVLVCNKDLRPICGTDGVT
YTNDCLLCAYSIEFGTNISKEHDGECKETVPMNCSSYANTTSEDGKVMVLCNRAFNPVCG
TDGVTYDNECLLCAHKVEQGASVDKRHDGGCRKELAAVSVDCSEYPKPDCTAEDRPLCGS
DNKTYGNKCNFCNAVVESNGTLTLSHFGKC
26. ANTISER8
Plasminostreptin (PSTI-type protease inhibitor) - Streptomyces sp.
GLYAPSALVLTMGHGNSAATVNPERAVTLNCAPTASGTHPAALQACAELRGAGGDFDALT
VRGDVACTKQFDPVVVTVDGVWQGKRVSYERTFANECVKNSYGMTVFTFLLNNSLLVPTS
GIYFVYSQVVFSGKAYSPKATSSPLYLAHEVQLFSSQYPFHVPLLSSQKMVYPGLQEPWL
HSMYHGAAFQLTQGDQLSTHTDGIPHLVLSPSTVFFGAFAL

6. TNFHUM
tumor necrosis factor (Nature, 312 (1984/5) 720
VRSSSRTPSDKPVAHVVANPQAEGQLQWLNRRANALLANGVELRDNQLVV
PSEGLYLIYSQVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSP
CQRETPEGAEAKPWYEPIYLGGVFQLEKGDRLSAEINRPDYLDFAESGQV
YFGAFAL□

7. **TNFM1**
TNF misa r-152 (Nature, 320, 584, (1986))
LRSSSQNSSDKPVAHVVANHQVEEQLEWLSQRANALLANGMDLKDNQLVV
PADGLYLVYSQVLFKGQGCPDYVLLTHTVSRFAISYQEKVNLLSAVKSPC
PKDTPEGAELKPWYEPYYLGGVFQLEKGDQLSAEVNLPKYLDFAESGQVY
FgVIAL
8. **TNFMIS**
TNF misa r-152 (Nature, 320, 584, (1986))
LRSSSQNSSDKPVAHVVANHQVEEQLEWLSQRANALLANGMDLKDNQLVV
PADGLYLVYSQVLFKGQGCPDYVLLTHTVSRFAISYQEKVNLLSAVKSPC
PKDTPEGAELKPWYEPYYLGGVFQLEKGDQLSAEVNLPKYLDFAESGQVY
FRVIAL
9. **TNFRAB**
TNF RABBIT
LRSASRALSDKPLAHVVANPQVEGQLQWLSQRANALLANGM
KLTDNQLVVPADGLYLIYSQVLFSGQGCRSYVLLTHTVSRFAVSYPNKVNLLSAIKSPC
HRETREEAEPMAWYEPIYLGGVFQLEKGDRLSTEVNQPEYLDLAESGQVYFGIIAL

PDG (4 sequences)

1. PDGFAHUM
Z1008276A : platelet derived growth factor A
MNRCWALFLSLCCYLRLVSAEGDPIPEELYEMLSDHSIRSFDDLQRLLHGDPGEEDGAEL
DLNMTRSHSGGELESLARGRRSLGSLTIAEPAMIAECKTRTEVFEISRRLIDRTNANFLV
WPPCVEVQRCSGCCNNRNVQCRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKC
ETVAAARPVTRSPGGSQEQRAKTPQTRVTIRTVRVRRPPKGKHRKFKHTHDKTALKETLG
A
2. PDGFAXEN
Z1411343A : platelet derived growth factor A
MRIWAWILLLSVQCSYLSPSLGEEAEIPQELIERLAHSEIRSISDLQRLLDIDSVGGGED
ASAANIRSQKHDFHHNRLVPEKRSVPSRRKRSVEEAVPAICKTRTVIYEIPRSQIDPTSA
NFLIWPPCVEVKRCTGCCNTSSVKCQPSRIHHRSVKVAKVEYVRKKPKLKEVLVRLEEHL
ECTCTANSNSDYREEETGFFTSPALVLTGRTRETGKKQKRKKLKPT
3. PDGFHUM
Z1007147A : platelet derived growth factor
LVSARQGDPIPEELYEMLSDHSIRSFDDLQRLLHGDPGEEDGAELDLNMTRSHSGGELES
LARGRRSLGSLTIAEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCN
NRNVQCRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKCETVAAARPVTRSPGG
SQEQRAKTPQTRVTIRTVRVRRPPKGKHRKFKHTHDKTALKETLGA
4. PDGFRECM
Z1210315A : platelet derived growth factor receptor
MGLPGVIPALVLRGQLLLSVLWLLGPQTSRGLVITPPGPEFVLNISSTFVLTCSGSAPVM
WEQMSQVPWQEAAMNQDGTFSSVLTLTNVTGGDTGEYFCVYNNSLGPELSERKRIYIFVP
DPTMGFLPMDSEDLFIFVTDVTETTIPCRVTDPQLEVTLHEKKVDIPLHVPYDHQRGFTG
TFEDKTYICKTTIGDREVDSDTYYVYSLQVSSINVSVNAVQTVVRQGESITIRCIVMGND
VVNFQWTYPRMKSGRLVEPVTDYLFGVPSRIGSILHIPTAELSDSGTYTCNVSVSVNDHG
DEKAINISVIENGYVRLLETLGDVEIAELHRSRTLRVVFEAYPMPSVLWLKDNRTLGDSG
AGELVLSTRNMSETRYVSELILVRVKVSEAGYYTMRAFHEDDEVQLSFKLQVNVPVRVLE
LSESHPANGEQTIRCRGRGMPQPNVTWSTCRDLKRCPRKLSPTPLGNSSKEESQLETNVT
FWEEDQEYEVVSTLRLRHVDQPLSVRCMLQNSMGGDSQEVTVVPHSLPFKVVVISAILAL
VVLTVISLIILIMLWQKKPRYEIRWKVIESVSSDGHEYIYVDPVQLPYDSTWELPRDQLV
LGRTLGSGAFGQVVEATAHGLSHSQATMKVAVKMLKSTARSSEKQALMSELKIMSHLGPH
LNVVNLLGACTKGGPIYIITEYCRYGDLVDYLHRNKHTFLQRHSNKHCPPSAELYSNALP
VGFSLPSHLNLTGESDGGYMDMSKDESIDYVPMLDMKGDIKYADIESPSYMAPYDNYVPS
APERTYRATLINDSPVLSYTDLVGFSYQVANGMDFLASKNCVHRDLAARNVLICEGKLVK
ICDFGLARDIMRDSNYISKGSTYLPLKWMAPESIFNSLYTTLSDVWSFGILLWEIFTLGG
TPYPELPMNDQFYNAIKRGYRMAQPAHASDEIYEIMQKCWEEKFETRPPFSQLVLLLERL
LGEGYKKKYQQVDEEFLRSDHPAILRSQARFPGIHSLRSPLDTSSVLYTAVQPNESDNDY
IIPLPDPKPDVADEGLPEGSPSLASSTLNEVNTSSTISCDSPLELQEEPQQAEPEAQLEQ
PQDSGCPGPLAEAEDSFL

INTERLEUKIN (11 sequences)

1. IL1ABOV
Z1413325A : interleukin 1alpha
MAKVPDLFEDLKNCYSENEDYSSEIDHLSLNQKSFYDASYEPLREDQMNKFMSLDTSETS
KSKLSFKENVVMVAASGKILKKRRLSLNQFITDDDLEAIANNTEEEIIKPRSAHYSFQSN
VKYNFMRVIHQECILNDALNQSIIRDMSGPYLTATTLNNLEEAVKFDMVAYVSEEDSQLP
VTLRISKTQLFVSAQNEDEPVLLKEMPETPKIIKDETNLLFFWEKHGSMDYFKSVAHPKL
FIATKQEKLVHMASGPPSITDFQILEK
2. IL1ABOV2
Z1413343A : interleukin 1alpha
MAKVPDLFEDLKNCYSENEDYSSEIDHLSLNQKSFYDASYEPLREDQMNKFMSLDTSETS
KTSKLSFKENVVMVAASGKILKKRRLSLNQFITDDDLEAIANNTEEEIIKPRSAHYSFQS
NVKYNFMRVIHQECILNDALNQSIIRDMSGPYLTATTLNNLEEAVKFDMVAYVSEEDSQL
PVTLRISKTQLFVSAQNEDEPVLLKEMPETPKIIKDETNLLFFWEKHGSMDYFKSVAHPK
LFIATKQEKLVHMASGPPSITDFQILEK
3. IL1AHUM
Z1107273A : interleukin 1alpha
MAKVPDMFEDLKNCYSENEEDSSSIDHLSLNQKSFYHVSYGPLHEGCMDQSVSLSISETS
KTSKLTFKESMVVVATNGKVLKKRRLSLSQSITDDDLEAIANDSEEEIIKPRSAPFSFLS
NVKYNFMRIIKYEFILNDALNQSIIRANDQYLTAAALHNLDEAVKFDMGAYKSSKDDAKI
TVILRISKTQLYVTAQDEDQPVLLKEMPEIPKTITGSETNLLFFWETHGTKNYFTSVAHP
NLFIATKQDYWVCLAGGPPSITDFQILENQA
4. IL1BBOV
Z1413325B : interleukin 1beta
MATVPEPINEMMAYYSDENELLFEADDPKQMKSCIQHLDLGSMGDGNIQLQISHQFYNKS
FRQVVSVIVAMEKLRNSAYAHVFHDDDLRSILSFIFEEEPVIFETSSDEFLCDAPVQSIK
CKLQDREQKSLVLASPCVLKALHLLSQEMNREVVFCMSFVQGEERDNKIPVALGIKDKNL
YLSCVKKGDTPTLQLEEVDPKVYPKRNMEKRFVFYKTEIKNTVEFESVLYPNWYISTSQI
EERPVFLGHFRGGQDITDFRMETLSP
5. IL1BHUM
Z1107273B : interleukin 1beta
MAEVPELASEMMAYYSGNEDDLFFEADGPKQMKCSFQDLDLCPLDGGIQLRISDHHYSKG
FRQAASVVVAMDKLRKMLVPCPQTFQENDLSTFFPFIFEEEPIFFDTWDNEAYVHDAPVR
SLNCTLRDSQQKSLVMSGPYELKALHLQGQDMEQQVVFSMSFVQGEESNDKIPVALGLKE
KNLYLSCVLKDDKPTLQLESVDPKNYPKKKMEKRFVFNKIEINNKLEFESAQFPNWYIST
SQAENMPVFLGGTKGGQDITDFTMQFVSS
6. IL1BHUML
Z1304174B : interleukin 1 beta
MAKVPDMFEDLKNCYSENEEDSSSIDHLSLNQKSFYHVSYGPLHEGCMDQSVSLSISETS
KTSKLTFKESMVVVATNGKVLKKRRLSLSQSITDDDLEAIANDSEEEIIKPRSAPFSFLS
NVKYNFMRIIKYEFILNDALNQSIIRANDQYLTAAALHNLDEAVKFDMGAYKSSKDDAKI
TVILRISKTQLYVTAQDEDQPVLLKEMPEIPKTITGSETNLLFFWETHGTKNYFTSVAHP
NLFIATKQDYWVCLAGGPPSITDFQILENQA
7. IL1BMUR
Z1302275A : interleukin 1beta
MATVPELNCEMPPFDSDENDLFFEVDGPQKMKGCFQTFDLGCPDESIQLQISQQHINKSF
RQAVSLIVAVEKLWQLPVSFPWTFQDEDMSTFFSFIFEEEPILCDSWDDDDNLLVCDVPI
RQLHYRLRDEQQKSLVLSDPYELKALHLNGQNINQQVIFSMSFVQGEPSNDKIPVALGLK
GKNLYLSCVMKDGTPTLQLESVDPKQYPKKKMEKRFVFNKIEVKSKVEFESAEFPNWYIS
TSQAEHKPVFLGNNSGQDIIDFTMESVSS
8. IL1HUMPR
Z1102279A : interleukin 1 precursor
MAEVPKLASEMMAYYSGNEDDLFFEADGPKQMKCSFQDLDLCPLDGGIQLRISDHHYSKG
FRQAASVVVAMDKLRKMLVPCPQTFQENDLSTFFPFIFEEEPIFFDTWDNEAYVHDAPVR
SLNCTLRDSQQKSLVMSGPYELKALHLQGQDMEQQVVFSMSFVQGEESNDKIPVALGLKE
KNLYLSCVLKDDKPTLQLESVDPKNYPKKKMEKRFVFNKIEINNKLEFESAQFPNWYIST
SQAENMPVFLGGTKGGQDITDFTMQFVSS
9. IL1MSREC
Z1410355A : interleukin 1 receptor
MENMKVLLGLICLMVPLLSLEIDVCTEYPNQIVLFLSVNEIDIRKCPLTPNKMHGDTIIW

YKNDSKTPISADRDSRIHQQNEHLWFVPAKVEDSGYYYCIVRNSTYCLKTKVTVTVLEND
PGLCYSTQATFPQRLHIAGDGSLVCPYVSYFKDENNELPEVQWYKNCKPLLLDNVSFFGV
KDKLLVRNVAEEHRGDYICRMSYTFRGKQYPVTRVIQFITIDENKRDRPVILSPRNETIE
ADPGSMIQLICNVTGQFSDLVYWKWNGSEIEWNDPFLAEDYQFVEHPSTKRKYTLITTLN
ISEVKSQFYRYPFICVVKNTNIFESAHVQLIYPVPDFKNYLIGGFIILTATIVCCVCIYK
VFKVDIVLWYRDSCSGFLPSKASDGKTYDAYILYPKTLGEGSFSDLDTFVFKLLPEVLEG
QFGYKLFIYGRDDYVGEDTIEVTNENVKKSRRLIIILVRDMGGFSWLGQSSEEQIAIYNA
LIQEGIKIVLLELEKIQDYEKMPDSIQFIKQKHGVICWSGDFQERPQSAKTRFWKNLRYQ
MPAQRRSPLSKHRLLTLDPVRDTKEKLPAATHLPLG

10. IL1MURPR

Z1101396A : interleukin 1 precursor

MAKVPDLFEDLKNCYSENEDYSSAIDHLSLNQKSFYDASYGSLHETCTDQFVSLRTSETS
KMSNFTFKESRVTVSATSSNGKILKKRRLSFSETFTEDDLQSITHDLEETIQPRSAPYTY
QSDLRYKLMKLVRQKFVMNDSLNQTIYQDVDKHYLSTTWLNDLQQEVKFDMYAYSSGGDD
SKYPVTLKISDSQLFVSAQGEDQPVLLKELPETPKLITGSETDLIFFWKSINSKNYFTSA
AYPELFIATKEQSRVHLARGLPSMTDFQIS

11. IL1RABPR

Z1110252B : interleukin 1 precursor

MAKVPDLFEDLKNCFSENEEYSSAIDHLSLNQKSFYDASYEPLHEDCMNKVVSLSTSETS
VSPNLTFQENVVAVTASGKILKKRRLSLNQPITDVDLETNVSDPEEGIIKPRSVPYTFQR
NMRYKYLRIIKQEFTLNDALNQSLVRDTSDQYLRAAPLQNLGDAVKFDMGVYMTSEDSIL
PVTLRISQTPLFVSAQNEDEPVLLKEMPETPRIITDSESDILFFWETQGNKNYFKSAANP
QLFIATKPEHLVHMARGLPSMTDFQIS

VIRAL ONCOGENE (30 sequences)

1. ABLMUR

Kinase-related transforming protein (abl) (EC 2.7.1.-) - Abelson murine leukemia

YITPVNSLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRSISLRYEGRVYHY
RINTASDGKLYVSSESRFNTLAELVHHHSTVADGLITTLHYPAPKRNKPTIYGVSPNYDK
WEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIK
HPNLVQLLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVSAVVLLYMATQISSAMEYL
EKKNFIHRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYN
KFSIKSDVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCPEKVYELMRA
CWQWNPSDRPSFAEIHQAFETMFQESSISDEVEKELGKRGTRGGAGSMLQAPELPTKTRT
CRRAAEQKASPPSLTPKLLRRQVTASPSSGLSHKKEATKGSASGMGTPATAEPAPPSNKV
GLSKASSEEMRVRRHKHSSESPGRDKGRLAKLKPAPPPPPACTGKAGKPAQSPSQEAGEA
GGPTKTKCTSLAMDAVNTDPTKAGPPGEGLRKPVPPSVPKPQSTAKPPGTPTSPVSTPST
APAPSPLAGDQQPSSAAFIPLISTRVSLRKTRQPPERIASGTITKGVVLDSTEALCLAIS
RNSEQMASHSAVLEAGKNLYTFCVSYVDSIQQMRNKFAFREAINKLESNLRELQICPATA
SSGPAATQDFSKLLSSVKEISDIVRR

2. SRCRAV

Kinase-related transforming protein (src) (EC 2.7.1.-) - Rous sarcoma virus

MGSSKSKPKDPSQRRRSLEPPDSTHHGGFPASQTPNKTAAPDTHRTPSRSFGTVATEPKL
FGGFNTSDTVTSPQRAGALAGGVTTFVALYDYESWIETDLSFKKGERLQIVNNTEGNWWL
AHSLTTGQTGYIPSNYVAPSDSIQAEEWYFGKITRRESERLLLNPENPRGTFLVRESETT
KGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFSSLQQLVAYYSKHADGLCHR
LTNVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTLKPG
TMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVIEYMSKGSLLDFLKGEMGKYLRL
PQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARLIEDNEYTARQ
GAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMGNGEVLDRVERGYR
MPCPPECPESLHDLMCQCWRRDPEERPTFEYLQAQLLPACVLEVAE

3. YESAV

Kinase-related transforming protein (yes) (EC 2.7.1.-) - Avian sarcoma virus Y73

DKGPAMKYRTDNTPEPISSHVSHYGSDSSQATQSPAIKGSAVNFNSHSMTPFGGPSGMTP
FGGASSSFSAVPSPYPSTLTGGGTVFVALYDYEARTTDDLSFKKGERFQIINNTEGDWWE

ARSIATGKTGYIPSNYVAPADSIQAEEWYFGKMGRKDAERLLLNPGNQRGIFLVRESETT
KGAYSLSIRDWDEVRGDNVKHYKIRKLDNGGYYITTRAQFESLQKLVKHYREHADGLCHK
LTTVCPTVKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTKVAIKTLKLG
TMMPEAFLQEAQIMKKLRHDKLVPLYAVVSEEPIYIVTEFMTKGSLLDFLKEGEGKFLKL
PQLVDMAAQIADGMAYIERMNYIHRDLRAANILVGDNLVCKIADFGLARLIEDNEYTARQ
GAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELVTKGRVPYPGMVNREVLEQVERGYR
MPCPQGCPESLHELMKLCWKKDPDERPTFEYIQSFLEDYFTAAEPSGY

4. ROSASRC

Kinase-related transforming protein (ros) (EC 2.7.1.-) - Avian sarcoma virus UR2

DTVTSPDITAIVAVIGAVVLGLTSLTIIILFGFVWHQRWKSRKPASTGQIVLVKEDKELA
QLRGMAETVGLANACYAVSTLPSQAEIESLPAFPRDKLNLHKLLGSGAFGEVYEGTALDI
LADGSGESRVAVKTLKRGATDQEKSEFLKEAHLMSKFDHPHILKLLGVCLLNEPQYLILE
LMEGGDLLSYLRGARKQKFQSPLLTLTDLLDICLDICKGCVYLEKMRFIHRDLAARNCLV
SEKQYGSCSRVVKIGDFGLARDIYKNDYYRKRGEGLLPVRWMAPESLIDGVFTNHSDVWA
FGVLVWETLTLGQQPYPGLSNIEVLHHVRSGGRLESPNNCPDDIRDLMTRCWAQDPHNRP
TFFYIQHKLQEIRHSPLCFSYFLGDKESVAPLRIQTAFFQPL

5. FPS

Kinase-related transforming protein (fps) (EC 2.7.1.-) - Fujinami sarcoma virus

ASGQLHRPQPQEHTSTSAAAGTWRLTQASESRHRLPHCSAAPSHQDHSAMGFGPELWCPK
GHTELLRLQDSELRLLELMKKWMSQRAKSDREYAGMLHHMFSQLEKQEGLGHLRATDHSS
QIGESWWVLASQTETLSQTLRRHAEELAAGPLAKLSILIRDKQQLRKVFSEQWQQLSQEY
AWTTQQEVEKLKAQYRSLVRDSTQAKRKYQEASKDKEREKAKEKYVRSLSKLYALHNQYV
LAVQAAALHHHHHYQRALPTLHESLYSLQQEMVLVLKEILGEYCSITSLVQEDVLAIHQK
VAHAVEMIDPATEYSSFVQCHRYDSEVPPAVTFDESLLEEAENLEPGELQLNELTIESVQ
HSLTSIEEELLASRKAVSSKEQRVWELQVELRGEELALSPGERVHLLGKRQGLREAQQQL
QGLVCAQAKLQAQRDMLANKLAELGSEEPPPALPLQEDRQSARSTDQERSGVTALKTIKN
HISGIFSPRFSLPPPVPLIPEVQKPLCQQAWYHGAIPRSEVQELLKYSGDFLVRESQGKQ
EYVLSVLWDGQPRHFIIQAADNLYRLEDDGLPTIPLLIDHLLQSQRPITRKSGIVLTRAV
LKDKWVLNHEDVLLGERIGRGNFGEVFSGRLRADNTPVAVKSCRETLPPELKAKFLQEAR
ILKQCNHPNIVRLIGVCTQKQPIYIVMELVQGGDFLSFLRSKGPRLKMKKLIKMMENAAA
GMEYLESKHCIHRDLAARNCLVTEKNTLKISDFGMSRQEEDGVYASTGGMKQIPVKWTAP
EALNYGWYSSESDVWSFGILLWEAFSLGAVPYANLSNQQTREAIEQGVRLEPPEQCPEDV
YRLMQRCWEYDPHRRPSFGAVHQDLIAIRKRHR

6. MILAVI

Kinase-related transforming protein (mht or mil) (EC 2.7.1.-) - Avian retrovirus

PTMPVDSRIIEDAIRNHSESASPSASSGSPNNMSPTGWSQPKTPVPAQRERAPGTNTQEK
NKIRPRGQRDSSYYWEIEASEVLLSTRIGSGSFGTVYKGKWHGDVAVKILKVVDPTPEQF
QAFRNEVAVLRKTRHVNILLFMGYMTKDNLAIVTQWCEGSSLYKHLHVQETKFQMFQLID
IARQTAQGMDYLHAKNIIHRDMKSNNIFLHGGLTVKIGDFGLATVKSRWSGSQQVEQPTG
SILWMAPEVIRMQDSNPFSFQSDVYSYGIVLYELMTGELPYSHINNRDQIIFMVGRGYAS
PDLSKLYKNCPKAMKRLVADCLKKVREERPLFPQILSSIALLQHSLPKINRSASEPSLHR
ASHTEDINSCTLTSTRLPVF

7. ERBBAV

Kinase-related transforming protein (erbB) - Chicken and avian erythroblastosis

MKCAHFIDGPHCVKACPAGVLGENDTLVWKYADANAVCQLCHPNCTRGCKGPGLEGCPNG
SKTPSIAAGVVGGLLCLVVVGLGIGLYLRRRHIVRKRTLRRLLQERELVEPLTPSGEAPN
QAHLRILKETEFKKVKVLGFGAFGTVYKGLWIPEGEKVTIPVAIKELREATSPKANKEIL
DEAYVMASVDNPHVCRLLGICLTSTVQLITQLMPYGCLLDYIREHKDNIGSQYLLNWCVQ
IAKGMNYLEERHMVHRDLAARNVLVKTPQHVKITDFGLAKQLGADEKEYHAEGGKVPIKW
MALESILHRIYTHQSDVWSYGVTVWELMTFGSKPYDGIPASEISSVLEKGERLPQPPICT
IDVYMIMVKCWMSDADSRPKFRELIAEFSKMARDPPRYLVIQGDERMHLPSPTDSKFYRT
LMEEEDMEDIVDADEYLVPHQGFFNSPSTSRTPLLSSLSATSNNSATNCIDRNGGHPVRE
DGFLPAPEYVNQLMPKKPSTAMVQNQIYNYISLTAISKLPIDSRYQNSHSTAVDNPEYLE

8. MOSMUR

Kinase-related transforming protein (mos) - Moloney murine sarcoma virus

MAHSTPCSQTSLAVPNHFSLVSHVTVPSEGVMPSPLSLCRYLPRELSPSVDSRSCSIPLV
APRKAGKLFLGTTPPRAPGLPRRLAWFSIDWEQVCLMHRLGSGGFGSVYKATYHGVPVAI

KQVNKCTEDLRASQRSFWAELNIAGLRHDNIVRVVAASTRTPEDSNSLGTIIMEFGGNVT
LHQVIYDATRSPEPLSCRKQLSLGKCLKYSLDVVNGLLFLHSQSILHLDLKPANILISEQ
DVCKISDFGCSQKLQDLRGRQASPPHIGGTYTHQAPEILKGEIATPKADIYSFGITLWQM
TTREVPYSGEPQYVQYAVVAYNLRPSLAGAVFTASLTGKALQNIIQSCWEARGLQRPSAE
LLQRDLKAFRGTLG

9. FES

Kinase-related transforming protein (fes) (EC 2.7.1.-) - Feline sarcoma virus

AARADGTMGFSSELCSPQGHGAEQQMQEAELRLLEGMRKWMAQRVKSDREYAGLLHHMSL
QDGGGRGTGPYSPISQSWAEITSQTEGLSRLLRQHAEDLNSGPLSKLGLLIRERQQLRKT
YSEQWQQLQQELTKTHNQDIEKLKSQYRALARDSAQARRKYQEASKDKDRDKAKLEQLGP
GEPPPVLLLQDDRHSTSSSEQEREGGRTPTLEILKSHISGIFRPKFSLPPPLQLVPEVQK
PLHEQLWYHGALPRAEVAELLTHSGDFLVRESQGKQEYVLSVLWDGQPRHFIIQSADNLY
RPEGDGFASIPLLVDHLLRSQQPLTKKSGIVLNRAVPKDKWVLNHEDLVLGEQIGRGNFG
EVFSGRLRADNTLVAVKSCRETLPPDIKAKFLQEAKILKQYSHPNIVRLIGVCTQKQPIY
IVMELVQGGDFLTFLRTEGARLRMKTLLQMVGDAAAGMEYLESKCCIHRDLAARNCLVTE
KNVLKISDFGMSREAADGIYAASGGLRQVPVKWTAPEALNYGRYSSESDVWSFGILLWET
FSLGASPYPNLSNQQTREFVEKGGRLPCPELCPDAVFRLMEQCWAYEPGQRPSFSAIYQE
LQSIRKRHR

10. FPSAV

Kinase-related transforming protein (fps) (EC 2.7.1.-) - Avian sarcoma virus PRC

ASGQLHRPQPQEHTSTSAAAGTWRHTQASESRHRLPHCSAAPSHQDHSAMGFGPELWCPK
GHSELLRLQDSELRLLELMKKWMSERAKSDREYAGMLHHMFSQLGSEEPPPALPLQEDRQ
SVCSTDQERSGVTALETIKNHISGIFSPRFSLPPPVPLIPEVQKPLCQQAWYHGAIPRSE
VQELLKCSGDFLVRESQGKQEYVLSVLWDGQPRHFIIQAADNLYRLEGDGFPTIPLLIDH
LLQSQQPITRKSGIVLTRAVLKDKWVLNHEDVLLGERIGRGNFGEVFSGRLRADNTPVAV
KSCRETLPPELKAKFLQEARILKQYNHPNIVRLIGVCTQKQPIYIVMELVQGGDFLSFLR
SKGPHLKMKELIKMMENAAAGMEYLESKHCIHRDLAARNCLVTEKNTLKISDFGMSRQEE
DGVYASTGGMKQIPVKWTAPEALNYGRYSSESDVWSFGILLWEAFSLGAVPYANLSNQQT
REAIEQGVRLEPPEQCPEDVYRLMQRCWEYDPRRRPSFGAVHQDLIAIRKRHR

11. FGR

Kinase-related transforming protein (fgr) (EC 2.7.1.-) - Feline sarcoma virus (s

ARALCRPAVCRPRPLPPLPPTAMEEEVAALVIDNGSGMCKAGFAGDDAPRAVFPSIVGRP
RHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAP
EEHPVLLTEAPLNPKANREKMTQIMFETFNIPSNYVAPVDSIQAEEWYFGKIGRKDAERQ
LLSPGNARGAFLVRESETTKGAYSLSIRDWDEARGDHVKHYKIRKLDTGGYYITTRAQFN
SVQELVQHYVEVNDGLCHLLTAACTTMKPQTMGLAKDAWEISRSSITLQRRLGTGCFGDV
WLGMWNGSTKVAVKTLKPGTMSPKASLEEAQIMKLLRHDKLVQLYAVVPEEPIYIVTEFM
CHGSLLEFLKDQEGQDLTLPQLVDMAAQVAEGMAYMERMDYIHRDLRAANILVGERLVCK
IADFGLARLIEDNEYNPRQGAKFPIKWTAPEAALFGRFTIKSDVWSFGILLTELISKGRV
PYPGMNNREVLEQVEHGYHMPCPPGCPASLYEAMEQTWRLDPEERPTFEYLQSFLEDYFN
GPQQN

12. FMS

Kinase-related transforming protein (fms) (EC 2.7.1.-) - Feline sarcoma virus (s

RMPSGPGHYGASAETPGPRPPLCPASSCCLPTEAMGPRALLVLLMATAWHAQGVPVIQPS
GPELVVEPGTTVTLRCVGNGSVEWDGPISPHWNLDLDPPSSILTTNNATFQNTGTYHCTE
PGNPRGGNATIHLYVKDPARPWKVLAQEVTVLEGQDALLPCLLTDPALEAGVSLVRVRGR
PVLRQTNYSFSPWHGFTIHKAKFIENHVYQCSARVDGRTVTSMGIWLKVQKDISGPATLT
LEPAELVRIQGEAAQIVCSASNIDVNFDVSLRHGDTKLTISQQSDFHDNRYQKVLTLNLD
HVSFQDAGNYSCTATNAWGNHSASMVFRVVESAYSNLTSEQSLLQEVTVGEKVDLQVKVE
AYPGLESFNWTYLGPFSDYQDKLDFVTIKDTYRYTSTLSLPRLKRSESGRYSFLARNAGG
QNALTFELTLRYPPEVRVTMTLINGSDTLLCEASGYPQPSVTWVQCRSHTDRCDESAGLV
LEDSHSEVLSQVPFYEVIVHSLLAIGTLEHNRTYECRAFNSVGNSSQTFWPISIGAHTPL
PDELLFTPVLLTCMSIMALLLLLLLLLLYKYKQKPKYQVRWKIIESYEGNSYTFIDPTQL
PYNEKWEFPRNNLQFGKTLGTGAFGKVVEATAFGLGKEDAVLKVAVKMLKSTAHADEKEA
LMSELKIMSHLGQHENIVNLLGACTHGGPVLVITEYCCYGDLLNFLRRQAEAMPGPSLSV
GQDPEAGAGYKNIHLEKKYVRRDSGFSSQGVDTYVEMRPVSTSSSNDSFSEEDLGKEDGR
PLELRDLLHFSSQVAQGMAFLASKNCIHRDVAARNVLLTSGRVAKIGDFGLARDIMNDSN

YIVKGNARLPVKWMAPESIFDCVYTVQSDVWSYGILLWEIFSLGLNPYPGILVNSKFYKL
VKDGYQMAQPAFAPKNIYSIMQACWALEPTRRPTFQQICSLLQKQAQEDRRVPNYTNLPS
SSSSRLLRPWRGPPL

13. KIT

Kinase-related transforming protein (kit) - Feline sarcoma virus (strain Hardy-Z

EEINGNNYIDPTQLPYDHKWEFPRNRLSFGKTLGAGAFGKVVEATAYGLIKSDAAMTVAV
KMLKPSAHLTEREALMSELKVLSYLGNHMNIVNLLGACTVGGPTLVITEYCCYGDLLNFL
RRKRDSFICSKQEDHAEVALYKNLLQSKESSCNDSTNEYMDMKPGVSYVVPTKADKRRSA
RIGSYIERDVTPAIMEDGELALDLEDLLSFSYQVAKGMAFLASKNCIHRDLAARNILLTH
GRITKICDFGLARDIKNDSNYVVKGNARLPVKWMAPESIFNCVYTFESDVWSYGIFLWEL
FSLGSSPYPGMPVDSKFYKMIKEGFRMLSPEHAPAEMYDIMKTCWDADPLKRPTFKQIVQ
LIEKQKNKNE

14. FOS

Transforming protein (fos) - FBJ murine osteosarcoma virus

MMFSGFNADYEASSFRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF
IPTVTATSTSPDLQWLVQPTLVSSVAPSQTRAPHPYGLPTQSAGAYARAEMVKTVSGGRA
QSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDKKSALQ
TEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEMSVASLDLTGGLPEASTPESEEAFT
LPLLNDPEPKPSLEPVKSISNVELKAEPFDDFLFPASSRPSGSETSRSVPNVDLSGSFYA
ADWEPLHSNSLGMGPMVTELEPLCTPVVTCTPLLRLPELTHAAGPVSSQRRQGSRHPDVP
LPELVHYREEKHVFPQRFPST

15. MYBAV

Transforming protein (myb) - Avian myeloblastosis virus

MRRKVEQEGYPQESSKAGPPSATTGFQKSSHLMAFAHNPPAGPLPGAGQAPLGSDYPYYH
IAEPQNVPGQIPYPVALHINIINVPQPAAAAIQRHYTDEDPEKEKRIKELELLLMSTENE
LKGQQALPTQNHTANYPGWHSTTVADNTRTSGDNAPVSCLGEHHHCTPSPPVDHGCLPEE
SASPARCMIVHQSNILDNVKNLLEFAETLQLIDSFLNTSSNHENLNLDNPALTSTPVCGH
KMSVTTPFHKDQTFTEYRKMHGGAV

16. MYCAV

Transforming protein (myc) - Avian myelocytomatosis virus MC29

QAAAAAMPLSASLPSKNYDYDYDSVQPYFYFEEEEENFYLAAQQRGSELQPPAPSEDIWK
KFELLPMPPLSPSRRSSLAAASCFPSTADQLEMVTELLGGDMVNQSFICDPDDETFVKSI
IIQDCMWSGFSAAAKLEKVVSEKLATYQASRQEGGPAAASRPGPPPSGPPPPPAGPAASA
GLYLHDLGAAAADCIDPSVVFPYPLSERAPRAAPPGANPAALLGVDTPPTTSSDSEEEQE
EDEEIDVVTLAEANESESSTESSTEASEEHCKPHHSPLVLKRCHVNIHQHNYAAPPSTKV
EYPAAKRLKLDSGRVLKQISNNRKCSSPRTLDSEENDKRRTHNVLERQRRNELKLRFFAL
RDQIPEVANNEKAPKVVILKKATEYVLSLQSDEHKLIAEKEQLRRRREQLKHNLEQLRNS
RA

17. RASHM

Transforming protein (H-ras) - BALB murine sarcoma virus

MTEYKLVVVGAKGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPMVLVGNKCDL
AARTVESRQAQDLARSYGIPYIKTSAKTRQGVEDAFYTLVREIRQHKLRKLNPPDESGPG
CMSCKCVLS

18. RASK1M

Transforming protein 1 (K-ras) - Kirsten murine sarcoma virus

MTEYKLVVVGASGVGKSALTIQLIQNHFVDEYDPTIQDSYRKQVVIDGETCLLDILDTTG
QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQLKRVKDSEDVPMVLVGNKCDL
PSRTVDTKQAQELARSYGIPFIETSAKTRQRVEDAFYTLVREIRQYRLKKISKEEKTPGC
VKIKKCVIM

19. MYCPRO

Transforming protein (myc) - Feline leukemia provirus

AGREHLRLYRQLLLAGLRGAARRPTNLAQVKQVVQGIVENPQAAAMPLNVSFANRNYDLD
YDSVQPYFYCDEEENFYQQQQQSELQPPAPSEDIWKKFELLPTPPLSPSRRSGLCSPSYV
AFASFSPRGDDDGGGGSFSTADQLEMVTELLGGDMVNQSFICDPDDETFIKNIIIQDCMW
SGFSAAAKLVSEKLASYQAARKDSGSPSPARGPGGCPTSSLYLQDLTAAASECIDPSVVF
PYPLNDSSSPKPCASPDSAAFSPSSDSLLSSAESSPRASPEPLALHEETPPTTSSDSEEE
QEEEEEIDVVSVEKRQPPAKRSESGSPSAGGHSKPPHSPLVLKRCHVPTHQHNYAAPPST
RKDYPAAKRAKLDSGRVLKQISNNRKCISPRSSDTEENDKRRTHNVLERQRRNELKRSFF
ALRDQIPELENNEKAPKVVILKKATAYILSVQAGEQKLISEKDLLRKRREQLKHKLEQLR

NSCA

20. RAFMSRC

Kinase-related transforming protein (raf) (EC 2.7.1.-) - Murine sarcoma virus 36

EKNKIRPRGQRDSSYYWKMEASEVMLSTRIGSGSFGTVYKGKWHGDVAVKILKVVDPTPE
QLQAFRNEVAVLRKTRHVNILLFMGYMTKDNLAIVTQWCEGSSLYKHLHVQETKFQMFQL
IDIARQTAQGMDYLHAKNIIHRDMKSNNIFLHEGLTVKIGDFGLATVKSRWSGSQQVEQP
TGSVLWMAPEVIRMQDDNPFSFQSDVYSYGIVLYELMAGELPYAHINNRDQIIFMVGRGY
ASPDLSRLYKNCPKAIKRLVADCVKKVKEERPLFPQILSSIELLQHSLPKINRSAPEPSL
HRAAHTEDINACTLTTSPRLPVF

21. SIS

PDGF-related transforming protein (sis) - Simian sarcoma virus

MTLTWQGDPIPEELYKMLSGHSIRSFDDLQRLLQGDSGKEDGAELDLNMTRSHSGGELES
LARGKRSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCN
NRNVQCRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKCEIVAAARAVTRSPGT
SQEQRAKTTQSRVTIRTVRVRRPPKGKHRKCKHTHDKTALKETLGA

22. ERBA

Transforming protein (erbA) - Avian erythroblastosis virus

EDTRWLDGKHKRKSSQCLVKSSMSGYIPSCLDKDEQCVVCGDKATGYHYRCITCEGCKSF
FRRTIQKNLHPTYSCTYDGCCVIDKITRNQCQLCRFKKCISVGMAMDLVLDDSKRVAKRK
LIEENRERRRKEEMIKSLQHRPSPSAEEWELIHVVTEAHRSTNAQGSHWKQRRKFLLEDI
GQSPMASMLDGDKVDLEAFSEFTKIITPAITRVVDFAKNLPMFSELPCEDQIILLKGCCM
EIMSLRAAVRYDPESETLTLSGEMAVKREQLKNGGLGVVSDAIFDLGKSLSAFNLDDTEV
ALLQAVLLMSSDRTGLICVDKIEKCQESYLLAFEHYINYRKHNIPHFWSKLLMKVADLRM
IGAYHASRFLHMKVECPTELPPRRCRALQILGSILPFV

23. REL

Transforming protein (rel) - Avian reticuloendotheliosis virus

MDFLTNLRFTEGISEPYIEIFEQPRQRGTRFRYKCEGRSAGSIPGEHSTDNNKTFPSIQI
LNYFGKVKIRTTLVTKNEPYKPHPHDLVGKGCRDGYYEAEFGPERQVLSFQNLGIQCVKK
KDLKESISLRISKKINPFNVPEEQLHNIDEYDLNVVRLCFQAFLPDEHGNYTLALPPLIS
NPIYDNRAPNTAELRICRVNKNCGSVKGGDEIFLLCDKVQKDDIEVRFVLGNWEAKGSFS
QADVHRQVAIVFRTPPFLGDITEPITVKMQLRRPSDQAVSEPVDFRYLPDEEDPSGNKAK
RQRSTLAWQKPIQDCGSAVTERPKAAPIPTVNPEGKLKKEPNMFSPTLMLPGLGTLSSSQ
MYPACSQMPTQPAQLGPGKQDTLHSCWQQLYSPSPSASSLLSLHSHSSFTAEVPQPGAQG
SSSLPAYNPLNWPDEKNSSFYRNFGNTHGMGAALVSAAGMQSVSSSSIVQGTHQASATTA
SIMTMPRTPGEVPFLRQQVGYRS

24. NEU

NEU$RAT : NEU ONCOGENE (P185) (EPIDERMAL GROWTH FACTOR RECEPTOR-RELATED 1260 A

MIIMELAAWCRWGFLLALLPPGIAGTQVCTGTDMKLRLPASPETHLDMLRHLYQGCQVVQ
GNLELTYVPANASLSFLQDIQEVQGYMLIAHNQVKRVPLQRLRIVRGTQLFEDKYALAVL
DNRDPQDNVAASTPGRTPEGLRELQLRSLTEILKGGVLIRGNPQLCYQDMVLWKDVFRKN
NQLAPVDIDTNRSRACPPCAPACKDNHCWGESPEDCQILTGTICTSGCARCKGRLPTDCC
HEQCAAGCTGPKHSDCLACLHFNHSGICELHCPALVTYNTDTFESMHNPEGRYTFGASCV
TTCPYNYLSTEVGSCTLVCPPNNQEVTAEDGTQRCEKCSKPCARVCYGLGMEHLRGARAI
TSDNVQEFDGCKKIFGSLAFLPESFDGDPSSGIAPLRPEQLQVFETLEEITGYLYISAWP
DSLRDLSVFQNLRIIRGRILHDGAYSLTLQGLGIHSLGLRSLRELGSGLALIHRNAHLCF
VHTVPWDQLFRNPHQALLHSGNRPEEDLCVSSGLVCNSLCAHGHCWGPGPTQCVNCSHFL
RGQECVEECRVWKGLPREYVSDKRCLPCHPECQPQNSSETCFGSEADQCAACAHYKDSSS
CVARCPSGVKPDLSYMPIWKYPDEEGICQPCPINCTHSCVDLDERGCPAEQRASPVTFII
ATVEGVLLFLILVVVVGILIKRRRQKIRKYTMRRLLQETELVEPLTPSGAMPNQAQMRIL
KETELRKVKVLGSGAFGTVYKGIWIPDGENVKIPVAIKVLRENTSPKANKEILDEAYVMA
GVGSPYVSRLLGICLTSTVQLVTQLMPYGCLLDHVREHRGRLGSQDLLNWCVQIAKGMSY
LEDVRLVHRDLAARNVLVKSPNHVKITDFGLARLLDIDETEYHADGGKVPIKWMALESIL
RRRFTHQSDVWSYGVTVWELMTFGAKPYDGIPAREIPDLLEKGERLPQPPICTIDVYMIM
VKCWMIDSECRPRFRELVSEFSRMARDPQRFVVIQNEDLGPSSPMDSTFYRSLLEDDDMG
DLVDAEEYLVPQQGFFSPDPTPGTGSTAHRRHRSSSTRSGGGELTLGLEPSEEGPPRSPL
APSEGAGSDVFDGDLAMGVTKGLQSLSPHDLSPLQRYSEDPTLPLPPETDGYVAPLACSP
QPEYVNQSEVQPQPPLTPEGPLPPVRPAGATLERPKTLSPGKNGVVKDVFAFGGAVENPE
YLVPREGTASPPHPSPAFSPAFDNLYYWDQNSSEQGPPPSNFEGTPTAENPEYLGLDVPV

25. RASHAM
Transforming protein (H-ras) - Harvey murine sarcoma virus
MPAARAAPAADEPMRDPVAPVRAPALPRPAPGAVAPASGGARAPGLAAPVEAMTEYKLVV
VGARGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTTGQEEYSAMR
DQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPMVLVGNKCDLAGRTVESR
QAQDLARSYGIPYIETSAKTRQGVEDAFYTLVREIRQHKLRKLNPPDESGPGCMSCKCVL
S
26. MYCHBI
Transforming protein (myc) - Avian myelocytomatosis virus HBI
QAAAAAMPLSASLPSKNYDYDYDSVQPYFYFEEEEENFYLAAQQRGSELQPPAPSEDIWK
KFELLPMPPLSPSRRSSLAAASCFPSTADQLEMVTELLGGDMVNQSFICDPDDESFVKSI
IIQDCMWSGFSAAAKLEKVVSEKLATYQASRREGGPAAASRPGPPPSGPPPPPAGPAASA
GLYLHDLGAAAADCIDPSVVFPYPLSERAPRAAPPGANPAALLGVDTPPTTSSDSEEEQE
EDEEIDVVTLAEANESESSTESSTEASEEHCKPHHSPLVLKRCQVNIHQHNYAAPPSTKV
EYPAAKRLKLDSGRVLKQISNNRKCSSPRTLDSEENDKRRTHNVLERQRRNELKLRFFAL
RDQIPEVANNEKAPKVGILKKATEYVLSIQSDEHRLIAEKEQLRRRREQLKHNLEQLKNS
RA
27. MOSHT
Kinase-related transforming protein (mos) (EC 2.7.1.-) - Moloney murine sarcoma
MARSTPCSQTSLAVPTHFSLVSHVTVPSEGVMPSPLSLCRYLPRELSPSVDSRSCSIPLV
APRKAGKLFLGTTPPRAPGLPRRLAWFSIDWEQVCLMHRLGSGGFGSVYKATYHGVPVAI
KQVNKCTKDLRASQRSFWAELNIARLRHDNIVRVVAASTRTPEDSNSLGTIIMEFGGNVT
LHQVIYGATRSPEPLSCREQLSLGKCLKYSLDVVNGLLFLHSQSILHLDLKPANILISEQ
DVCKISDFGCSQKLQDLRCRQASPHHIGGTYTHQAPEILKGEIATPKADIYSFGITLWQM
TTREVPYSGEPQYVQYAVVAYNLRPSLAGAVFTASLTGKTLQNIIQSCWEARALQRPGAE
LLQRDLKAFRGALG
28. ERBBA
Kinase-related transforming protein (erbB) - Avian erythroblastosis virus
MKCAHFIDGPHCVKACPAGVLGENDTLVRKYADANAVCQLCHPNCTRGCKGPGLEGCPNG
SKTPSIAAGVVGGLLCLVVVGLGIGLYLRRRHIVRKRTLRRLLQERELVEPLTPSGEAPN
QAHLRILKETEFKKVKVLGSGAFGTIYKGLWIPEGEKVKIPVAIKELREATSPKANKEIL
DEAYVMASVDNPHVCRLLGICLTSTVQLITQLMPYGCLLDYIREHKDNIGSQYLLNWCVQ
IAKGMNYLEERRLVHRDLAARNVLVKTPQHVKITDFGLAKLLGADEKEYHAEGGKVPIKW
MALESILHRIYTHQSDVWSYGVTVWELMTFGSKPYDGIPASEISSVLEKGERLPQPPICT
IDVYMIMVKCWMIDADSRPKFRELIAEFSKMARDPPRYLVIQGDERMHLPSPTDSKFYRT
LMEEEDMEDIVDADEYLVPHQGFFNSPSTSRTPLLSSLSATSNNSATNCIDRNGQGHPVR
EDSFVQRYSSDPTGNFLEESIDDGFLPAPEYVNQLMPKKPSTAMVQNQIYNFISLTAISK
LPMDSRYQNSHSTAVDNPEYLNTNQSPLAKTVFESSPYWIQSGNHQINLDNPDYQQDFLP
TSCS
29. SRCROS
Kinase-related transforming protein (src) (EC 2.7.1.-) - Rous sarcoma virus
MGSSKSKPKDPSQRRHSLEPPDSTHHGGFPASQTPDETAAPDAHRNPSRSFGTVATEPKL
FWGFNTSDTVTSPQRAGALAGGVTTFVALYDYESWTETDLSFKKGERLQIVNNTEGDWWL
AHSLTTGQTGYIPSNYVAPSDSIQAEEWYFGKITRRESERLLLNPENPRGTFLVRKSETA
KGAYCLSVSDFDNAKGPNVKHYKIYKLYSGGFYITSRTQFGSLQQLVAYYSKHADGLCHR
LANVCPTSKPQTQGLAKDAWEIPRESLRLEAKLGQGCFGEVWMGTWNDTTRVAIKTLKPG
TMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVIEYMSKGSLLDFLKGEMGKYLRL
PQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARLIEDNEYTARQ
GAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLDQVERGYR
MPCPPECPESLHDLMCQCWRKDPEERPTFKYLQAQLLPACVLEVAE
30. FES2
Kinase-related transforming protein (fes) (EC 2.7.1.-) - Feline sarcoma virus
HPREQVQLLAKKQVLQEALQALQVALCSQAKLQAQRELLQAKLEQLGPGEPPPVLLLQDD
RHSTSSSEQEREGGRTPTLEILKSHISGIFRPKFSLPPPLQLVPEVQKPLHEQLWYHGAL
PRAEVAELLTHSGDFLVRESQGKQEYVLSVLWDGQPRHFIIQSADNLYRPEGDGFASIPL
LVDHLLRSQQPLTKKSGIVLNRAVPKDKWVLNHEDLVLGEQIGRGNFGEVFSGRLRADNT
LVAVKSCRETLPPDIKAKFLQEAKILKQYSHPNIVRLIGVCTQKQPIYIVMELVQGGDFL
TFLRTEGARLRMKTLLQMVGDAAAGMEYLESKCCIHRDLAARNCLVTEKNVLKISDFGMS

REEADGVYAASGGLRLVPVKWTAPEALNYGRYSSESDVWSFGILLWETFSLGASPYPNLS
NQQTREFVEKGGRLPCPELCPDAVFRLMEQCWAYEPGQRPSFSAFYQELQSIRKRHR

PROTO-ONCOGENE (15 sequences)

1. RASKMOUS
Transforming protein (K-ras) - Mouse
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDL
PSRTVDTKQAQELARSYGIPFIETSAKTRQRVEDAFYTLVREIRQYRLKKISKEEKTPGC
VKIKKCVIMGVDDAFYTLVREIRKHKEKMSKDGKKKKKKSRTRCTVM
2. RASK2HUM
Transforming protein 2 (K-ras) - Human
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDL
PSRTVDTKQAQDLARSYGIPFIETSAKTRQGVDDAFYTLVREIRKHKEKMSKDGKKKKKK
SKTKCVIM
3. RASKHUM
Transforming protein 1 (K-ras) - Human
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDL
PSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYTLVREIRQYRLKKISKEEKTPGC
VKIKKCIIM
4. SRCCH
Kinase-related transforming protein (src) (EC 2.7.1.-) - Chicken
MGSSKSKPKDPSQRRRSLEPPDSTHHGGFPASQTPNKTAAPDTHRTPSRSFGTVATEPKL
FGGFNTSDTVTSPQRAGALAGGVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWL
AHSLTTGQTGYIPSNYVAPSDSIQAEEWYFGKITRRESERLLLNPENPRGTFLVRESETT
KGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFSSLQQLVAYYSKHADGLCHR
LTNVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTLKPG
NMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLDFLKGEMGKYLRL
PQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARLIEDNEYTARQ
GAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLDQVERGYR
MPCPPECPESLHDLMCQCWRRDPEERPTFEYLQAFLEDYFTSTEPQYQPGENL
5. ERBBCH
Virus-induced, kinase-related transforming protein (gag-env-erbB) -
Chicken
MEAVIKAFLTGYPGETSKKDSKKKPPATSKKDPEKTPLLPTRVNYILIIGVLVLCEVTGG
PDHCMKCAHFIDGPHCVKACPAGVLGENDTLVWKYADANAVCQLCHPNCTRGCKGPGLEG
CPNGSKTPSIAAGVVGGLLCLVVVGLGIGLYLRRRHIVRKRTLRRLLQERELVEPLTPSG
EAPNQAHLRILKETEFKKVKVLGSGAFGTVYKGLWIPEGEKVKIPVAIKELREATSPKAN
KEILDEAYVMASVDNPHVCRLLGICLTSTVQLITQLMPYGCLLDYIREHKDNIGSQYLLN
WCVQIAKGMNYLEERRLVHRDLAARNVLVKTPQHVKITDFGLAKLLGADEKEYHAEGGKV
PIKWMALESILHRIYTHQSDVWSYGVTVWELMTFGSKPYDGIPASEISSVLEKGERLPQP
PICTIDVYMIMVKCWMIDADSRPKFRELIAEFSKMARDPPRYLVIQGDERMHLPSPTDSK
FYRTLMEEEDMEDIVDADEYLVPHQGFFNSPSTSRTPLLSSLSATSNNSATNCIDRNGQG
HPVREDSFVQRYSSDPTGNFLEESIDDGFLPAPEYVNQLMPKKPSTAMVQNQIYNNISLT
AISKLPMDSRYQNSHSTAVDNPEYLNTNQSPLAKTVFESSPYWIQSGNHQINLDNPDYQQ
DFLPNETKPNGLLKVPAAENPEYLRVAAPKSEYIEASA
6. MOSMS
Kinase-related transforming protein (mos) - Mouse
MWLVLRIKEEGKGTGIEGSNLPCSQTSLAVPTHFSLVSHVTVPSEGVMPSPLSLCRYLPR
ELSPSVDSRSCSIPLVAPRKAGKLFLGTTPPRAPGLPRRLAWFSIDWEQVCLMHRLGSGG
FGSVYKATYHGVPVAIKQVNKCTKDLRASQRSFWAELNIARLRHDNIVRVVAASTRTPED
SNSLGTIIMEFGGNVTLHQVIYGATRSPEPLSCREQLSLGKCLKYSLDVVNGLLFLHSQS
ILHLDLKPANILISEQDVCKISDFGCSQKLQVLRCRQASPHHIGGTYTHQAPEILKGEIA
TPKADIYSFGITLWQMTTREVPYSGEPQYVQYAVVAYNLRPSLAGAVFTASLTGKTLQNI
.QSCWEARALQRPCAELLQRDLKAFRGALG

7. MOSHUM
Kinase-related transforming protein (mos) - Human
MPSPLALRPYLRSEFSPSVDARPCSSPSELPAKLLLGATLPRAPRLPRRLAWCSIDWEQV
CLLQRLGAGGFGSVYKATYRGVPVAIKQVNKCTKNRLASRRSFWAELNVARLRHDNIVRV
VAASTRTPAGSNSLGTIIMEFGGNVTLHQVIYGAAGHPEGDAGEPHCRTGGQLSLGKCLK
YSLDVVNGLLFLHSQSIVHLDLKPANILISEQDVCKISDFGCSEKLEDLLCFQTPSYPLG
GTYTHRAPELLKGEGVTPKADIYSFAITLWQMTTKQAPYSGERQHILYAVVAYDLRPSLS
AAVFEDSLPGQRLGDVIQRCWRPSAAQRPSARLLLVDLTSLKAELG
8. FMSHUM
Kinase-related transforming protein precursor (fms) (EC 2.7.1.-)-Human
MGPGVLLLLLVATAWHGQGIPVIEPSVPELVVKPGATVTLRCVGNGSVEWDGPASPHWTL
YSDGSSSILSTNNATFQNTGTYRCTEPGDPLGGSAAIHLYVKDPARPWNVLAQEVVVFED
QDALLPCLLTDPVLEAGVSLVRVRGRPLMRHTNYSFSPWHGFTIHRAKFIQSQDYQCSAL
MGGRKVMSISIRLKVQKVIPGPPALTLVPAELVRIRGEAAQIVCSASSVDVNFDVFLQHN
NTKLAIPQQSDFHNNRYQKVLTLNLDQVDFQHAGNYSCVASNVQGKHSTSMFFRVVESAY
LNLSSEQNLIQEVTVGEGLNLKVMVEAYPGLQGFNWTYLGPFSDHQPEPKLANATTKDTY
RHTFTLSLPRLKPSEAGRYSFLARNPGGWRALTFELTLRYPPEVSVIWTFINGSGTLLCA
ASGYPQPNVTWLQCSGHTDRCDEAQVLQVWDDPYPEVLSQEPFHKVTVQSLLTVETLEHN
QTYECRAHNSVGSGSWAFIPISAGAHTHPPDEFLFTPVVVACMSIMALLLLLLLLLLYKY
KQKPKYQVRWKIIESYEGNSYTFIDPTQLPYNEKWEFPRNNLQFGKTLGAGAFGKVVEAT
AFGLGKEDAVLKVAVKMLKSTAHADEKEALMSELKIMSHLGQHENIVNLLGACTHGGPVL
VITEYCCYGDLLNFLRRKAEAMLGPSLSPGQDPEGGVDYKNIHLEKKYVRRDSGFSSQGV
DTYVEMRPVSTSSNDSFSEQDLDKEDGRPLELRDLLHFSSQVAQGMAFLASKNCIHRDVA
ARNVLLTNGHVAKIGDFGLARDIMNDSNYIVKGNARLPVKWMAPESIFDCVYTVQSDVWS
YGILLWEIFSLGLNPYPGILVNSKFYKLVKDGYQMAQPAFAPKNIYSIMQACWALEPTHR
PTFQQICSFLQEQAQEDRRERDYTNLPSSSRSGGSGSSSSELEEESSSEHLTCCEQGDIA
QPLLQPNNYQFC
9. FOSHUM
Transforming protein (fos)-Human
MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCTDLAVSSANF
IPTVTAISTSPDLQWLVQPALVSSVAPSQTRAPHPFGVPAPSAGAYSRAGVVKTMTGGRA
QSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQ
TEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEMSVASLDLTGGLPEVATPESEEAFT
LPLLNDPEPKPSVEPVKSISSMELKTEPFDDFLFPASSRPSGSETARSVPDMDLSGSFYA
ADWEPLHSGSLGMGPMATELEPLCTPVVTCTPSCTAYTSSFVFTYPEADSFPSCAAAHRK
GSSSNEPSSDSLSSPTLLAL
10. MYCBUR
Transforming protein (myc), Burkitt lymphoma - Human
MPLNVTITNKNYDLDYDSVQPYFYCDEEENFYQQQQQSDLQPPAPSEDIWKKFELLLPNP
PLSPSRRSGLCSPSYVAVTPFSLRGDNDDGGGNFSTADQLEMVTELLGGDMVNQNFICDP
GDETFIKNIIIQDCMWSGFSAAAKLVSEKVASYQAARKDSGSPNPARGHSVSSTSSLYLQ
DLSAAASECIDPSVVFPYPLNDSRSPKSCASQDSSAFSPSSDSLLSSTESAPQGSPEPLV
FHEETSPTTSSDSEEEQEDEEEIDVVSVEKRQAPGKRSESGSPSAGGHSKPPHSPLVLKR
CHVSTHQHNYAAPPSTRKDYPAAKRVKLDSVRVLRQISNNRKCTSPRSSDTEENVKRRTH
NVLERQRRNELKRSFFALRDQIPELENNEKAPKVVILKKATAYILSVQAEEQKLISEEDL
LRKRREQLKHKLEQLRNSCA
11. MYCCH
Transforming protein (myc) - Chicken
MPLSASLPSKNYDYDYDSVQPYFYFEEEEENFYLAAQQRGSELQPPAPSEDIWKKFELLP
TPPLSPSRRSSLAAASCFPSTADQLEMVTELLGGDMVNQSFICDPDDESFVKSIIIQDCM
WSGFSAAAKLEKVVSEKLATYQASRREGGPAAASRPGPPPSGPPPPPAGPAASAGLYLHD
LGAAAADCIDPSVVFPYPLSERAPRAAPPGANPAALLGVDTPPTTSSDSEEEQEEDEEID
VVTLAEANESESSTESSTEASEEHCKPHHSPLVLKRCHVNIHQHNYAAPPSTKVEYPAAK
RLKLDSGRVLKQISNNRKCSSPRTSDSEENDKRRTHNVLERQRRNELKLSFFALRDQIPE
VANNEKAPKVVILKKATEYVLSIQSDEHRLIAEKEQLRRRREQLKHKLEQLRNSRA
12. INT1HU
Transforming protein (int-1) - Human
MGLWALLPGWVSATLLLALAALPAALAANSSGRWWGIVNVASSTNLLTDSKSLQLVLEPS
LQLLSRKQRRLIRQNPGILHSVSGGLQSAVRECKWQFRNRRWNCPTAPGPHLFGKIVNRG
CRETAFIFAITSAGVTHSVARSCSEGSIESCTCDYRRRGPGGPDWHWGGCSDNIDFGRLF
GREFVDSGEKGRDLRFLMNLHNNEAGRTTVFSEMRQECKCHGMSGSCTVRTCWMRLPTLR

AVGDVLRDRFDGASRVLYGNRGSNRASRAELLRLEPEDPAHKPPSPHDLVYFEKSPNFCT
YSGRLGTAGTAGRACNSSSPALDGCELLCCGRGHRTRTQRVTERCNCTFHWCCHVSCRNC
THTRVLHECL

13. INT1MS

Transforming protein (int-1) - Mouse mammary tumor virus

MGLWALLPSWVSTTLLLALTALPAALAANSSGRWWGIVNIASSTNLLTDSKSLQLVLEPS
LQLLSRKQRRLIRQNPGILHSVSGGLQSAVRECKWQFRNRRWNCPTAPGPHLFGKIVNRG
CRETAFIFAITSAGVTHSVARSCSEGSIESCTCDYRRRGPGGPDWHWGGCSDNIDFGRLF
GREFVDSGEKGRDLRFLMNLHNNEAGRTTVFSEMRQECKCHGMSGSCTVRTCWMRLPTLR
AVGDVLRDRFDGASRVLYGNRGSNRASRAELLRLEPEDPAHKPPSPHDLVYFEKSPNFCT
YSGRLGTAGTAGRACNSSSPALDGCELLCCGRGHRTRTQRVTERCNCTFHWCCHVSCRNC
THTRVLHECL

14. RASH1H

Transforming protein p21 (H-ras-1) - Human

MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPMVLVGNKCDL
AARTVESRQAQDLARSYGIPYIETSAKTRQGVEDAFYTLVREIRQHKLRKLNPPDESGPG
CMSCKCVLS

15. HRAS

Harvey murine sarcoma virus oncogene

MTGYKLVVVGAGGVGKSAVTIQLIQNHFVDEYDPTIEDSYRKQVVIDGET
SLLDILDTTGQTTYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQI
KRVYDSDDVPMVLVGNKCDVAGRTVESRQAQDLARSYGIPYIETSAKTRQ
GVEDAFYTLVREIRQH

ONCOGENE

(45 sequences including viral and non-viral oncogenes)

Description of all following sequences is listed above

1. ABLMUR
2. SRCRAV
3. YESAV
4. ROSASRC
5. FPS
6. MILAVI
7. ERBBAV
8. MOSMUR
9. FES
10. FPSAV
11. FGR
12. FMS
13. KIT
14. FOS
15. MYBAV
16. MYCAV
17. RASHM
18. RASK1M
19. MYCPRO
20. RAFMSRC
21. SIS
22. ERBA
23. REL
24. NEU
25. RASHAM
26. MYCHBI
27. MOSHT
28. ERBBA
29. SRCROS
30. FES2

31. RASKMOUS
32. RASK2HUM
33. RASKHUM
34. SRCCH
35. ERBBCH
36. MOSMS
37. MOSHUM
38. FMSHUM
39. FOSHUM
40. MYCBUR
41. MYCCH
42. INT1HU
43. INT1MS
44. RASH1H
45. HRAS

P53 (13 sequences)

1. P53CDVIR

Z1209275A : antigen p53 pCD tumor

MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDDLLLPQDVEEFFEGPSEA
LRVSGAPAAQDPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVM
CTYSPPLNKLFCQLAKTCPVQLWVSATPPAGSRVRAMAIYKKSQHMTEVVRRCPHHERCS
DGDGLAPPQHLIRVEGNLYPEYLEDRQTFRHSVVVPYEPPEAGSEYTTIHYKYMCNSSCM
GGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDRRTEEENFRKKEVLCPELPPGS
AKRALPTCTSASPPQKKKPLDGEYFTLKIRGRKRFEMFRELNEALELKDAHATEESGDSR
AHSSYLKTKKGQSTSRHKKTMVKKVGPDSD

2. P53CH

Z1503247A : p53 nuclear oncoprotein

MAEEMEPLLEPTEVFMDLWSMLPYSMQQLPLPEDHSNWQELSPLEPSDPPPPPPPPPLPL
AAAAPPPLNPPTPPRAAPSPVVPSTEDYGGDFDFRVGFVEAGTAKSVTCTYSPVLNKVYC
RLAKPCPVQVRVGVAPPPGSSLRAVAVYKKSEHVAEVVRRCPHHERCGGGTDGLAPAQHL
IRVEGNPQARYHDDETTKRHSVVVPYEPPEVGSDCTTVLYNFMCNSSCMGGMNRRPILTI
LTLEGPGGQLLGRRCFEVRVCACPGRDRKIEEENFRKRGGAGGVAKRAMSPPTEAPEPPK
KRVLNPDNEIFYLQVRGRRRYEMLKEINEALQLAEGGSAPRPSKGRRVKVEGPQPSCGKK
LLQKGSD

3. P53DEOHU

Z1103198A : protein p53

KTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAM
AIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPY
EPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGR
DRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEM
FRELNEALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

4. P53H1HUM

Z1212347A : antigen p53-H-1,tumor

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPRVAPAPATPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

5. P53HUM

CT53$HUMAN: CELLULAR TUMOR ANTIGEN P53 (PHOSPHOPROTEIN P53). 393 AA.

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVHVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG

GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

6. P53HUM43

Z1201270A : antigen p53,tumor

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVHVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

7. P53M8VIR

Z1209275B : antigen p53 M8 tumor

MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDDLLLPQDVEEFFEGPSEA
LRVSGAPAAQDPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVM
CTYSPPLNKLFFQLAKTCPVQLWVSATPPAGSRVRAMAIYKKSQHMTEVVRRCPHHERCS
DGDGLAPPQHLIRVEGNLYPEYLEDRQTFRHSVVVPYEPPEAGSEYTTIHYKYMCNSSCM
GGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDRRTEEENFRKKEVLCPELPPGS
AKRALPTCTSASPPQKKKPLDGEYFTLKIRGRKRFEMFRELNEALELKDAHATEESGDSR
AHSSLQPRAFQALIKEESPNC

8. P53MOUSE

CT53$MOUSE: CELLULAR TUMOR ANTIGEN P53. 390 AA.

MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDDLLLPQDVEEFFEGPSEA
LRVSGAPAAQDPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVM
CTYSPPLNKLFCQLVKTCPVQLWVSATPPAGSRVRAMAIYKKSQHMTEVVRRCPHHERCS
DGDGLAPPQHLIRVEGNLYPEYLEDRQTFRHSVVVPYEPPEAGSEYTTIHYKYMCNSSCM
GGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDRRTEEENFRKKEVLCPELPPGS
AKRALPTCTSASPPQKKKPLDGEYFTLKIRGRKRFEMFRELNEALELKDAHATEESGDSR
AHSSYLKTKKGQSTSRHKKTMVKKVGPDSD

9. P53MSLIV

Z1011200A : antigen p53,cellular tumor

MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDDLLLPQDVEEFFEGPSEA
LRVSGAPAAQDPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVM
CTYSPPLNKLFCQLVKTCPVQLWVSATPPAGSRVRAMAIYKKSQHMTEVVRRCPHHERCS
DGDGLAPPQHLIRVEGNLYPEYLEDRQTFRHSVVVPYEPPEAGSEYTTIHYKYMCNSSCM
GGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDRRTEEENFRKKEVLCPELPPGS
AKRALPTCTSASPPQKKKPLDGEYFTLKIRGRKRFEMFRELNEALELKDAHATEESGDSR
AHSSYLKTKKGQSTSRHKKTMVKKVGPDSD

10. P53MUR

Z1001197A : antigen p53,tumor

MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDDLLLPQDVEEFFEGPSEA
LRVSGAPAAQDPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVM
CTYSPPLNKLFCQLAKTCPVQLWVSATPPAGSRVRAMAIHKKSQHMTGVVRRCPHHERCS
DGDGLAPPQHLIRVEGNLYPEYLEDRQTFRHSVVVPYEPPEAGSEYTTIHYKYICNSSCM
GGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDRRTEEENFRKKEVLCPELPPGS
AKRALPTCTSASPPQKKKPLDGEYFTLKIRGRKRFEMFRELNEALELKDAHATEESGDSR
AHSSYLKTKKGQSTSRHKKTMVKKVGPDSD

11. P53MURCA

Z1008277A : protein p53

MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDDLLLPQDVEEFFEGPSEA
LRVSGAPAAQDPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVM
CTYSPPLNKLFCQLAKTCPVQLWVSATPPAGSRVRAMAIYKKSQHMTEVVRRCPHHERCS
DGDGLAPPQHLIRVEGNLYPEYLEDRQTFRHSVVVPYEPPEAGSEYTTIHYKYMCNSSCM
GGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDRRTEEENFRKKEVLCPELPPGS
AKRALPTCTSASPPQKKKPLDGEYFTLKIRGRKRFEMFRELNEALELKDAHATEESGDSR
AHSSYLKTKKGQSTSRHKKTMVKKVGPDSD

12. P53RAT

Z1503248A : p53 nuclear oncoprotein

MEDSQSDMSIELPLSQETFSCLWKLLPPDDILPTTATGSPNSMEDLFLPQDVAELLEGPE
EALQVSAPAAQEPGTEAPAPVAPASATPWPLSSSVPSQKTYQGNYGFHLGFLQSGTAKSV
MCTYSISLNKLFCQLAKTCPVQLWVTSTPPPGTRVRAMAIYKKSQHMTEVVRRCPHHERC
SDGDGLAPPQHLIRVEGNPYAEYLDDRQTFRHSVVVPYEPPEVGSDYTTIHYKYMCNSSC

MGGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDRRTEEENFRKKEEHCPELPPG
SAKRALPTSTSSSPQQKKKPLDGEYFTLKIRGRERFEMFRELNEALELKDARAAEESGDS
RAHSSYPKTKKGQSTSRHKKPMIKKVGPDSD

13. P53XL

Z1404409A : p53 related gene

MEPSSETGMDPPLSQETFELLWSLLPDPLQTVTCRLDNLSEFPDYPLAADMTVLQEGLMG
NAVPTVTSCAVPSTDDYAGKYGLQLDFQQNGTAKSVTCTYSPELNKLFCQLAKTCPLLVR
VESPPPRGSILRATAVYKKSEHVAEVVKRCPHHERSVEPGEDAAPPSHLMRVEGNLQAYY
MEDVNSGRHSVCVPYEGPQVGTECTTVLYNYMCNSSCMGGMNRRPILTIITLETPQGLLL
GRRCFEVRVCACPGRDRRTEEDNYTKKRGLKPSGKRELAHPPSSEPPLPKKRLVVVDDDE
EIFTLRIKGRSRYEMIKKLNDALELQESLDQQKVTIKCRKCRDEIKPKKGKKLLVKDEQP
DSE

CYSTEINE PROTEASES (23 sequences)

1. ACTNCHN1

Actinidin (EC 3.4.22.14) - Yangtao

LPSYVDWRSAGAVVDIKSQGECGGCWAFSAIATVEGINKITSGSLISLSEQELIDCGRTQ
NTRGCDGGYITDGFQFIINDGGINTZENYPYTAQDGDCDVALQDQKYVTIDTYENVPYNN
EWALQTAVTYQPVSVALDAAGDAFKQYASGIFTGPCGTAVDHAIVIVGYGTEGGVDYWIV
KNSWDTTWGEEGYMRILRNVGGAGTCGIATMPSYPVKYNN

2. ACTNKIW2

Actinidin -kiwi actin5pas

LPSYVDWRSAGAVVDIKSQGECGGCWAFSAIATVEGINKITSGSLISLSEQELIDCGRTQ
NTRGCDGGYITDGFQFIINDGGINTEENYPYTAQDGDCDVALQDQKYVTIDTYENVPYNN
EWALQTAVTYQPVSVALDAAGDAFKQYASGIFTGPCGTAVDHAIVIVGYGTEGGVDYWIV
KNSWDTTWGEEGYMRILRNVGGAGTCGIATMPSYPVKY

3. CATHCOW2

Z1409299A : cathepsin B

LPESFDAREQWPNCPTIKEIRDQGSCGSCWAFGAVEAISDRICIHSNGRVNVEVSAEDML
TCCGGECGDGCNGGEPSGAWNFWTKKGLVSGGLYNSHVGCRPYSIPPCEHHVNGSRPPCT
GEGDTPKCNKTCEPGYSPSYKEDKHFGCSSYSVANNEKEIMAEIYKNGPVEGAFSVYSDF
LLYKSGVYQHVSGEIMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQDHC
GIESEIVAGMPCT

4. CATHHEN1

Z1311264A : cathepsin L H

APRSVDWREKGYVTPVKDQGQCGSCWAFSTTGALEGQHFRKTGKLVSLSEQNLVDCSRPE
GNQGCNGGLMDQAFQYVQDNGGIDSEESYPYTAKDDEDCRYKAEYNAANDTGFVDIPQGH
ERALMKAVASVGPVSVAIDAGHSSFQFYQSGIYYEPDCSSEDLDHGVLVVGYGFEG

5. CATHHEN3

Z1312246A : cathepsin L

APRSVDWREKGYVTPVKDQGICGSCWAFNTTGALEGQHFFKTGKLVSLSEQNLVDCSRPE
GNQGCNGGLMDQAFQYVQDNGGIDSEESYPYTAKDDCRYKAEYNAAKDTGFVDIPQGHER
ALMKAVASVGPVSVAIDAGHSSFQFYQSGIYYEPDCSSEDLDHGVLVVGYGFEGKKYWIV
KNSWGEKWGDKGYQYMAKDRKNHCGIATAASYPLV

6. CATHHM13

Cathepsin B (EC 3.4.22.1) - Human

LPASFDAREQWPQCPTIKEIRDQGSCGSCWAFGAVEAISDRICIHTNVSVEVSAEDLLTC
CGSMCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCEHHVNGSRPPCTGE
GDTPKCSKICEPGYSPTYKQDKHYGYDSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLL
YKSGVYQHVTGEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQDHCGI
ESEVVAGIPRTD

7. CATHHM14

CATH$HUMAN: CATHEPSIN H PRECURSOR (EC 3.4.22.16). 335 AA.

YPPSVDWRKKGNFVSPVKNQGACGSCWTFSTTGALESAIAIATGKMLSLAEQQLVDCAQD
FNNYGCQGGLPSQAFEYILYNKGIMGEDTYPYQGKDGYCKFQPGKAIGFVKDVANITIYD
EEAMVEAVALYNPVSFAFEVTQDFMMYRTGIYSSTSCHKTPDKVNHAVLAVGYGEKNGIP
YWIVKNSWGPQWGMNGYFLIERGKNMCGLAACASYPIPLV

8. CATHHM2
Z1205366A : cathepsin B
VEVSAEDLLTCCGSMCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCEHH
VNGSRPPCTGEGDTPKCSKICEPGYSPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVE
GAFSVYSDFLLYKSGVYQHVTGEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFF
KILRGQDHCGIESEVVAGIPRTDQYWEKI
9. CATHHM4
Z1306235A : cathepsin G
IGGRESRPHSRPYMAYLQIQSPAGQSRCGGFLVREDFVLTAAHCWGSNINVTLGAHNIQR
RENTQQHITARRAIRHPQYNQRTIQNDIMLLQLSRRVRRNRNVNPVALPRAQEGLRPGTL
CTVAGWGRVSMRRGTDTLREVQLRVQRDRQCLRIFGSYDPRRQICVGDRRERKAAFKGDS
GGPLLCNNVAHGIVSYGKSSGVPPEVFTRVSSFLPWIRTTMRSFKLLDQMETPL
10. CATHHM5
Z1307261A : cathepsin D
GPIPEVLKNYMDAQYYGEIGIGTPPQCFTVVFDTGSSNLWVPSIHCKLLDIACWIHHKYN
SDKSSTYVKNGTSFDIHYGSGSLSGYLSQDTVSVPCQSASSASALGGVKVERQVFGEATK
QPGITFIAAKFDGILGMAYPRISVNNVLPVFDNLMQQKLVDQNIFSFYLSRDPDAQPGGE
LMLGGTDSKYYKGSLSYLNVTRKAYWQVHLDQVEVASGLTLCKEGCEAIVDTGTSLMVGP
VDEVRELQKAIGAVPLIQGEYMIPCEKVSTLPAITLKLGGKGYKLSPEDYTLKVSQAGKT
LCLSGFMGMDIPPPSGPLWILGDVFIGRYYTVFDRDNNRVGFAEAARL
11. CATHHM6
Z1404279A : Cathepsin H
CPYPPSVDWRKKGNFVSPVKNQGACGSCWTFSTTGALESAIAIATGKMLSLAEQQLVDCA
QDFNNHGCQGGLPSQAFEYILYNKGIMGEDTYPYQGKDGYCKFQPGKAIGFVKDVANITI
YDEEAMVEAVALYNPVSFAFEVTQDFMMYRTGIYSSTSCHKTPDKVNHAVLAVGYGEKNG
IPYWIVKNSWGPEWGMNGYFLIERGKNMCGLAACASYPIP
12. CATHHM7
Z1404279B : cathepsin L
APRSVDWREKGYVTPVKNQGQCGSCWAFSATGALPGQMFRKTGRLISLSEQNLVDCSGPQ
GNEGCNGGLMDYAFQYVQDNGGLDSEESYPYEATEESCKYNPKYSVAXDTGFVDIPKQEK
ALMKAVATVGPISVAIDAGHESFLFYKEGIYFEPNCSSEDMDHGVLVVGYGFESTNNKYW
LVKNSWGEEWGMGGYVKMAKDRRNHCGIASAASYPTV
13. CATHMSE1
Z1211378B : cathepsin B
LPETFDAREQWSNCPTIGQIRDQGSCGSCWAFGAVEAISDRTCIHTNGRVNVEVSAEDLL
TCCGIQCGDGCNGGYPSGAWNFWTKKGLVSGGVYDSHIGCLPYTIPPCEHHVNGSRPPCT
GEGDTPRCNKSCEAGYSPSYKEDKHFGYTSYSVSNSVKEIMAEIYKNGPVEGAFTVFSDF
LTYKSGVYKHEAGDMMGGHAIRILVWGVENGVPYWLAANSWNLDWGDNGFFKILRGENHC
GIESEIVAGIPRTDQYWGRF
14. CATHMSE2
Z1407219A : cathepsin L
MNLLLLLAVLCLGTALATPKFDQTFSAEWHQWKSTHRRLYGTNEEEWRRAIWEKNMRIIQ
LHNGEYSNGQHGFSMEMNAFGDMTNEEFRQVVNGYRHQKHKKGRLFQEPLMLKIPKSVDW
REKGCVTPVKNQGQCGSCWAFSASGCLEGQMFLKTGKLISLSEQNLVDCSHAQGNQGCNG
GLMDFAFQYIKENGGLDSEESYPYEAKDGSCKYRAEFAVANDTGFVDIPQQEKALMKAVA
TVGPISVAMDASHPSLQFYSSGIYYEPNCSSKNLDHGVLLVGYGYEGTDSNKNKYWLVKN
SWGSEWGMEGYIKIAKDRDNHCGLATAASYPVVN
15. CATHPIG1
Z1007243A : cathepsin D
GPIPEVLKNYMDAQYYGEIGIGTPPQCFTVVFDTGSSNLWVPSIHCKLLDIACWIHHKYN
SGKSSTYVKNGTTFAIHYGSGSLSGYLSQDTVSVPSNVGGIKVERQTFGEATKQPGLTFI
AAKFDGILGMAYPRISVNNVVPVFDNLMQQKLVDKDIFSFYLNRDPGAQPGGELMLGGID
SKYYKGSLDYHNVTRKAYWQIHMNQVAVGSSLTLCKGGCEAIVDTGTSLIVGQPEEVREL
GKAIGAVPLIQGEYMIPCEKVPSLPDVTVTLGGKKYKLSSENYTLKVSQAGQTICLSGFM
GMDIPPPGGPLWILGDVFIGRYYTVFDRDLNRVGLAEAA
16. CATHRAT1
Z1313245A : cathepsin L
MTPLLLLAVLCLGTALATPKFDQTFNAQWHQWKSTHRRLYGTNEEEWRRAVWEKNMRMIQ
LHNGEYSNGKHGFTMEMNAFGDMTNEEFRQIVNGYRHQKHKKGRLFQEPLMLQIPKTVDW
REKGCVTPVKNQGQCGSCWAFSASGCLEGQMFLKTGKLISLSEQNLVDCSHDQGNQGCNG
GLMDFAFQYIKENGGLDSEESYPYEAKDGSCKYRAEYAVANDTGFVDIPQQEKALMKPVA

TVGPISVAMDASHPSLQFYSSGIYYEPNCSSKDLDHGVLVVGYGYEGTDSNKDKYWLVKN
SWGKEWGMDGYIKIAKDRNNHCGLATAASYPIVN

17. CATHRAT2

Z1402240A : cathepsin H
YPSSMDWRKKGNVVSPVKNQGACGSCWTFSTTGALESAVAIASGKMMTLAEQQLVDCAQN
FNNHGCQGGLPSQAFEYILYNKGIMGEDSYPYIGKNGQCKFNPEKAVAFVKNVVNITLND
EAAMVEAVALYNPVSFAFEVTEDFMMYKSGVYSSNSCHKTPDKVNHAVLAVGYGEQNGLL
YWIVKNSWGSNWGNNGYFLIERGKNMCGLAACASYPIPQV

18. CATHRAT4

Cathepsin B (EC 3.4.22.1) - Rat
LPESFDAREQWSNCPTIAQIRDQGSCGSCWAFGAVEAMSDRICIHTNVNVEVSAEDLLTC
CGIQCGDGCNGGYPSGAGNFWTRKGLVSGGVYNSHIGCLPYTIPPCEHHVNGSRPPCTGE
GDTPKCNKMCEAGYSTSYKEDKHYGYTSYSVSDSEKEIMAEIYKNGPVEGAFTVFSDFLT
YKSGVYKHEAGDVMGGHAIRILGWGIENGVPYWLVANSWNVDWGDNGFFKILRGENHCGI
ESEIVAGIPRTQ

19. PAPAPAP1

Papain - papaya omega
LPENYDWRKKGAVTPVRHQGSSGSCWAFSAVATVEGINKIRTGKLVELSEQELVDCERRS
HGCKGGYPPYALEYVAKNGIHLRSKYPYKAKQGTCRAKQVGGPIVKTSGVGRVQPNNEGN
LLNAIAKQPVSVVVESKGRPFQLYKGGIFEGPCGTKVDHAVTAVGYGKSGGKGYILIKNS
WGTAWGEKGYIRIKRAPGNSPGVCGLYKSSYYPTKN

20. PAPAPAP2

Papain (EC 3.4.22.2) - Papaya
IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDRRS
YGCNGGYPWSALQLVAQYGIHYRNTPYYEGVQRYCRSREKGPYAAKTDGVRQVQPYNQGA
LLYSIANQPVSVVLQAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYNPGYILIKNSWGTG
WGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN

21. PAPAPAP3

Papain - chymopapain jbc 1989
YPQSIDWRAKGAVTPVKNQGACGSCWAFSTIATVEGINKIVTGNLLELSEQELVDCDKHS
YGCKGGYQTTSLQYVANNGVHTSKVYPYQAKQYKCRATDKPGPKVKITGYKRVPSNCETS
FLGALANQPLSVLVEAGGKPFQLYKSGVFDGPCGTKLDHAVTAVGYGTSDGKNYIIIKNS
WGPNWGEKGYMRLKRQSGNSQGTCGVYKSSYYPFKGFA

22. PAPAPAP4

Papain - ppiv febl 1989
LPESVDWRAKGAVTPVKHQGYCESCWAFSTVATVEGINKIKTGNLVELSEQELVDCDLQS
YGCNRGYQSTSLQYVAQNGIHLRAKYPYIAKQQTCRANQVGGPKVKTNGVGRVQSNNEGS
LLNAIAHQPVSVVVESAGRDFQNYKGGIFEGSCGTKVDHAVTAVGYGKSGGKGYILIKNS
WGPGWGENGYIRIRRASGNSPGVCGVYRSSYYPIKN

23. PAPAPAP5

Papain - Papaya papain5pas
IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDRRS
YGCNGGYPWSALQLVAQYGIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYNQGA
LLYSIANQPVSVVLQAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYGPNYILIKNSWGTG
WGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN

METALLO PROTEASES (17 sequences)

1. CARBBAC1

DACA$BACSU: PENICILLIN-BINDING PROTEIN 5 (D-ALANYL-D-ALANINE
CARBOXYPEPTIDASE) 3
ASSGKILYSKNADKRLPIASMTKMMTEYLLLEAIDQGKVKWDQTYTPDDYVYEISQDNSL
SNVPLRKDGKYTVKELYQATAIYSANAAAIAIAEIVAGSETKFVEKMNAKAKELGLTDYK
FVNATGLENKDLHGHQPEGTSVNEESEVSAKDMAVLADHLITDYPEILETSSIAKTKFRQ
GTDDEMDMPNWNFMLKGLVSEYKKATVDGLKTGSTDSAGSCFTGTAERNGMRVITVVLNA
KGNLHTGRFDETKKMFDYAFDNFSMKEIYAEGDQVKGHKTISVDKGKEKEVGIVTNKAFS
LPVKNGEEKNYKAKVTLNKDNLTAPVKKGTKVGKLTAEYTGDEKDYGFLNSDLAGVDLVT
KENVEKANWFVLTMRSIGGFFAGIWGSIVDTVTGWF

2. CARBBAR1
Serine carboxypeptidase (EC 3.4.16.1) - Barley
APQGAEVTGLPGFDGALPSKHYAGYVTVDEGHGRNLFYYVVESERDPGKDPVVLWLNGGP
GCSSFDGFVYEPGPFNFESGGSVKSLPKLHLNPYAWSKVSTMIYLDSPAGVGLSYSKNVS
DYETGDLKTATDSHTFLLKWFQLYPEFLSNPFYIAGESYAGVYVPTLSHEVVKGIQGGAK
PTINFKGYMVGNGVCDTIFDGNALVPFAHGMGLISDEIYQQASTSCHGNYWNATDGKCDT
AISKIESLISGLNIYDILEPCYHSRSGVPCMSDEVATAWLDNAAVRSAIHAQSVSAIGPW
LLCTDKLYFVHDAGSMIAYHKNLTSQGYRAIIFSGDHDMXVPFTGSEAWTKSLGYGVVDS
WRPWITNGQVSGYTEGYEHGLTFATIKGAGHTVPEYKPQEAFAFYSRWLAGSKL
3. CARBCOW1
Carboxypeptidase E - Bovine
RPQEDGISFEYHRYPELREALVSVWLQCAAVSRIYTVGRSFEGRELLVLELSDNPGVHEP
GEPEFKYIGNMHGNEAVGRELLIFLAQYLCNEYQKGNETIVQLIHNTRIHIMPSLNPDGF
EKAASQLGELKDWFVGRSNAQGIDLNRNFPDLDRIVYINEKEGGPNNHLLKNLKKIVDQN
TKLAPETKAVIHWIMDIPFVLSANLHGGDLVANYPYDETRSGSAHEYSSCPDDDIFQSLA
RAYSSFNPPMSDPDRPPCRKNDDDSSFVEGTTNGAAWYSVPGGMQDFNYLSSNCFEITVE
LSCEKFPPEETLKNYWEDNKNSLISYIQQIHRGVKGFVRDLQGNPIANATLSVEGIDHDV
TSAKDGDYWRLLVPGNYKLTASAPGYLAIAKKVAVPYSPAVRVDFELESFSERKEEEKEE
LMEWWKMMSETLNF
4. CARBCOW2
Carboxypeptidase A (EC 3.4.17.1) - Bovine
ARSTNTFNYATYHTLDEIYDFMDLLVAEHPQLVSKLQIGRSYEGRPIYVLKFSTGGSNRP
AIWIDLGIHSREWITQATGVWFAKKFTENYGQNPSFTAILDSMDIFLEIVTNPNGFAFTH
SENRLWRKTRSVTSSSLCVGVDANRNWDAGFGKAGASSSPCSETYHGKYANSEVEVKSIV
DFVKNHGNFKAFLSIHSYSQLLLYPYGYTTQSIPDKTELNQVAKSAVAALKSLYGTSYKY
GSIITTIYQASGGSIDWSYNQGIKYSFTFELRDTGRYGFLLPASQIIPTAQETWLGVLTI
MEHTVNN
5. CARBCOW3
Carboxypeptidase B (EC 3.4.17.2) - Bovine
TTGHSYEKYNNWETIEAWTEQVASENPDLISRSAIGTTFLGNTIYLLKVGKPGSNKPAVF
MDCGFHAREWISPAFCQWFVREAVRTYGREIHMTEFLDKLDFYVLPVVNIDGYIYTWTTN
RMWRKTRSTRAGSSCTGTDLNRNFDAGWCSIGASNNPCSETYCGSAAESEKESKAVADFI
RNHLSSIKAYLTIHSYSQMMLYPYSYDYKLPKNNVELNTLAKGAVKKLASLHGTTYSYGP
GATTIYPASGGSDDWAYDQGIKYSFTFELRDKGRYGFVLPESQIQPTCEETMLAIKYVTS
YVLEHL
6. CARBCOW4
Procarboxypeptidase A complex component III - Bovine
DGEDAVPYSWSWQVSLQYEKDGAFHHTCGGSLIAPDWVVTAGHCISTSRTYQVVLGEYDR
SVLEGSEQVIPINAGDLFVHPLWNSNCVACGNDIALVKLSRSAQLGDKVQLANLPPAGDI
LPNEAPCYISGWGRLYTGGPLPDKLQCALLPVVDYEHCSQWDWWGITVKKTMVCAGGDTR
SGCNGDSGGPLNCPAADGSWQVHGVTSFVSAFGCNTIKKPTVFTRVSAFIDWIDETIASN
7. CARBCRY1
Carboxypeptidase B (EC 3.4.17.2) - Crayfish
MDWTSYHDYDEINAWLDSLATDYPELASVEDVGLSYEGRTMKLLKLGKGGADKPIIFIDG
GIHAREWIAPSTVTYIVNEFVSNSATYDDILSNVNFYVMPTINPDGYAYTFTDDRLWRKT
RSETGSVLGCKGADPNRNWSFHWDEVGASDSPCSDIYAGPEPFSEVEMRNVRDQILEYAA
NIKVYLTFHSYSQLWMYPWGFTSDLPDDWQDLDTLATNAVDALTAVHGTRYEIGSSTNTI
YAAAGGSDDWAKGEGGVKYAYTIELRDTGNYGFLLPENQIIPTGEETFEGVKVVANFVKD
TYS
8. CARBPOT1
CARBOXYPEPTIDASE AALPHA (COX) (E.C.3.4.17.1) COMPLEX WITH POTATO
CARBOXYPEPTIDASE
ARSTNTFNYATYHTLDEIYDFMDLLVAQHPELVSKLQIGRSYEGRPIYVLKFSTGGSNRP
AIWIDLGIHSREWITQATGVWFAKKFTENYGQNPSFTAILDSMDIFLEIVTNPNGFAFTH
SENRLWRKTRSVTSSSLCVGVDANRNWDAGFGKAGASSSPCSETYHGKYANSEVEVKSIV
DFVKNHGNFKAFLSIHSYSQLLLYPYGYTTQSIPDKTELNQVAKSAVAALKSLYGTSYKY
GSIITTIYQASGGSIDWSYNQGIKYSFTFELRDTGRYGFLLPASQIIPTAQETWLGVLTI
MEHTVNN
9. CARBPSD1 Folate hydrolase G2 (EC 3.4.-.-) precursor - Pseudomonas sp.
ALAQKRDNVLFQAATDEQPAVIKTLEKLVNIETGTGDAEGIAAAGNFLEAELKNL
GFTVTRSKSAGLVVGDNIVGKIKGRGGKNLLLMSHMDTVYLKGILAKAPFRVEG

DKAYGPGIADDKGGNAVILHTLKLLKEYGVRDYGTITVLFNTDEEKGSFGSRDLIQ
EEAKLADYVLSFEPTSAGDEKLSLGTSGIAYVQVNITGKASHAGAAPELGVNALVE
ASDLVLRTMNIDDKAKNLRFNWTIAKAGNVSNIIPASATLNADVRYARNEDFDAA
MKTLEERAQQKKLPEADVKVIVTRGRPAFNAGEGGKKLVDKAVAYYKEAGGTLG
VEERTGGGTDAAYAALSGKPVIESLGLPGFGYHSDKAEYVDISAIPRRLYMAARLIMDLGA
GK

10. CARBRAT1

E1501328A : carboxypeptidase B
MLLLLALVSVALAHASEEHFDGNRVYRVSVHGEDHVNLIQELANTKEIDFWKPDSATQVK
PLTTVDFHVKAEDVADVENFLEENEVHYEVLISNVRNALESQFDSHTRASGHSYTKYNKW
ETIEAWIQQVATDNPDLVTQSVIGTTFEGRNMYVLKIGKTRPNKPAIFIDCGFHAREWIS
PAFCQWFVREAVRTYNQEIHMKQLLDELDFYVLPVVNIDGYVYTWTKDRMWRKTRSTMAG
SSCLGVRPNRNFNAGWCEVGASRSPCSETYCGPAPESEKETKALADFIRNNLSTIKAYLT
IHSYSQMMLYPYSYDYKLPENYEELNALVKGAAKELATLHGTKYTYGPGATTIYPAAGGS
DDWSYDQGIKYSFTFELRDTGFFGFLLPESQIRQTCEETMLAVKYIANYVREHLY

11. CARBRAT2

E1501328B : carboxypeptidase A1
ALSTDSFNYATYHTLDEIYEFMDLLVAEHPQLVSKIQIGNTFEGRPIHVLKFSTGGTNRP
AIWIDTGIHSREWVTQASGVWFAKKITKDYGQDPTFTAVLDNMDIFLEIVTNPDGFAYTH
KTNRMWRKTRSHTQGSLCVGVDPNRNWDAGFGMAGASSNPCSETYRGKFPNSEVEVKSIV
DFVTSHGNIKAFISIHSYSQLLLYPYGYTSEPAPDQAELDQLAKSAVTALTSLHGTKFKY
GSIIDTIYQASGSTIDWTYSQGIKYSFTFELRDTGLRGFLLPASQIIPTAEETWLALLTI
MDHTVKHPY

12. CARBSTP1

Muramoyl-pentapeptide carboxypeptidase (EC 3.4.17.8) - Streptomyces
griseus soiv
NGCYTWSGTLSEGSSGEAVRQLQIRVAGYPGTGAQLAIDGQFGPATKAAVQRFQSAYGLA
ADGIAGPTFNKIYQLQDDDCTPVNFTYAELNRCNSDWSGGKVSAATARANALVTMWKLQA
MRHAMGDKPITVNGGFRSVTCNSNVGGASNSRHMYGHAADLGAGSQGFCALAQAARNHGF
TEILGPGYPGHNDHTHVAGGDGRFWSAPSCGI

13. CARBWHT1

Carboxypeptidase Y (EC 3.4.16.1) homolog - Wheat
MATTPRLASLLLLLALCAAAAGALRLPPDASFPGAQAERLIRALNLLPGRPRRGLGAGAE
DVAPGQLLERRVTLPGLPEGVGDLGHHAGYYRLPNTHDARMFYFFFESRGKKEDPVVIWL
TGGPGCSSELAVFYENGPFTIANNMSLVWNKFGWDKISNIIFVDPATGTGFSYSSDDRDT
RHDEAGVSNDLYDFLQVFFKKHPEFVKNDFFITGESYAGHYIPAFASRVHQGNKKNEGTH
INLKGFAIGNGLTDPAIQYKAYTDYALDMNLIQKADYDRINKFIPPCEFAIKLCGTDGKA
SCMAAYMVCNSIFNSIMKLVGTKNYYDVRKECEGKLCYDFSN.EKFFGDKAVRQAIGVGD
IEFVSCSTSVYQAMLTDWMRNLEVGIPALLEDGINVLIYAGEYDLICNWLGNSRWVHSME
WSGQKDFAKTAESSFLVDDAQAGVLKSHGALSFLKVHNAGHMVPMDQPKAALEMLRRFTQ
GKLKESVPEEEPATTFYAA

14. CARBYST3

E70003: CARBOXYPEPTIDASE Y EC 3.4.16.1
KIKDPKILGIDPNVTQYTGYLDVEDEDKHFFFWTFESRNDPAKDPVILWLNGGPGCSSLT
GLFFELGPSSIGPDLKPIGNPYSWNSNATVIFLDQPVNVGFSYSGSSGVSNTVAAGKDVY
NFLELFFDXFPEYVNKGQDFHIAGESYAHGYIPVFAXEILSHKDRNFNLTSVLIGNGLTD
PLTQYNYYEPMACGEGGEPSVLPSEECSAMEDSLERCLGLIESCYDSQSVWSCVPATIYC
NNAQLAPYQRTGRNVYDIRKDCEGGNLCYPTLQDIDDELNQDYVKEAVGAEVDHYESCNF
DINRNFLFAGDWMKPYHTAVTDLLNQDLPILVYAGDKDFINNTLGNKAWTDVLPWKYDEE
FASQKVRCWTASITDEVAGEVKSYKHFTYLRVFNGGHMVPFDVPENALSMVNEWIHGDFS
L

15. THERBAC1

Neutral proteinase (EC 3.4.24.4) - Bacillus cereus
VTGTNKVGTGKGVLGDTKSLNTTLSGSSYYLQDNTRGATIFTYDAKNRSTLPGTLWADAD
NVFNAAYDAAAVDAHYYAGKTYDYYKATFNRNSINDAGAPLKSTVHYGSNYNNAFWNGSQ
MVYGDGDGVTFTSLSGGIDVIGHELTHAVTENSSNLIYQNESGALNEAISDIFGTLVEFY
DNRNPDWEIGEDIYTPGKAGDALRSMSDPTKYGDPDHYSKRYTGSSDNGGVHTNSGIINK
QAYLLANGGTHYGVTVTGIGKDKLGAIYYRANTQYFTQSTTFSQARAGAVQAAADLYGAN
SAEVAAVKQSFSAVGVN

16. THERBAC2
Neutral proteinase (EC 3.4.24.4) - Bacillus sp.
ITGTSTVGVGRGVLGDQKNINTTYSTYYYLQDNTRGDGIFTYDAKYRTTLPGSLWADADN
QFFASYDAPAVDAHYYAGVTYDYYKNVHNRLSYDGNNAAIRSSVHYSQGYNNAFWNGSEM
VYGDGDGQTFIPLSGGIDVVAHELTHAVTDYTAGLIYQNESGAINEAISDIFGTLVEFYA
NKNPDWEIGEDVYTPGISGDSLRSMSDPAKYGDPDHYSKRYTGTQDNGGVHINSGIINKA
AYLISQGGTHYGVSVVGIGRDKLGKIFYRALTQYLTPTSNFSQLRAAAVQSATDLYGSTS
QEVASVKQAFDAVGVK
17. THERBAC3
Neutral proteinase (EC 3.4.24.4) precursor - Bacillus stearothermophilus
VAGASTVGVGRGVLGDQKYINTTYSSYYGYYYLQDNTRGSGIFTYDGRNRTVLPGSLWTD
GDNQFTASYDAAAVDAHYYAGVVYDYYKNVHGRLSYDGSNAAIRSTVHYGRGYNNAFWNG
SQMVYGDGDGQTFLPFSGGIDVVGHELTHAVTDYTAGLVYQNESGAINEAMSDIFGTLVE
FYANRNPDWEIGEDIYTPGVAGDALRSMSDPAKYGDPDHYSKRYTGTQDNGGVHTNSGII
NKAAYLLSQGGVHYGVSVNGIGRDKMGKIFYRALVYYLTPTSNFSQLRAACVQAAADLYG
STSQEVNSVKQAFNAVGVY

SERINE PROTEASES (38 sequences)

1. ELASHM1
E1304249A : elastase I
MLRLLVVASLVLYGHSTQDFPETNARVVGGTEAQRNSWPSQISLQYRSGSSWAHTCGGTL
IRQNWVMTAAHCVDRELTFRVVVGEHNLNQNDGTEQYVGVQKIVVHPYWNTDDVAAGYDI
ALLRLGQSVTLNSYVQLGVLPRAGTILANNSPCYITGWGLTRTNGQLAQTLQQAYLPTVD
YAILSSSSYWGSTVKNSMVCAGGDGVRSGCQGDSGGPLHCLVNGQYAVHGVTSFVSRLGC
NVTRKPTVFTRVSAYISWINNVIASN
2. ELASHM3
E1306262B : elastase IIB
MIRTLLLSTLVAGALSCGVSTYAPDMSRMLGGEEARPNSWPWQVSLQYSSNGQWYHTCGG
SLIANSWVLTAAHCISSSRIYRVMLGQHNLYVAESGSLAVSVSKIVVHKDWNSNQVSKGN
DIALLKLANPVSLTDKIQLACLPPAGTILPNNYPCYVTGWGRLQTNGALPDDLKQGRLLV
VDYATCSSSGWWGSTVKTNMICAGGDGVICTCNGDSGGPLNCQASDGRWEVHGIGSLTSV
LGCNYYYKPSIFTRVSNYNDWINSVIANN
3. ELAS[illegible]5
E131421[illegible] : [illegible]reatic elastase 2
MIRTLLLS [illegible]GGDPTYPPYV[illegible]GEEARPNSWPWQVSLQYSSNGKWYHTCGG
SLIANSWVLT[illegible]SSSRTYRVGI[illegible]RSGSLAVSVSKIVVHKDWNSNQISKGN
DIALLKLANP[illegible]TDKIQLACLPE[illegible]YVTGWGRLQTNGAVPDVLQQGRLLV
VDYATCSSSAWWGSSVKTSMICAG[illegible]GGPLNCQASDGRWQVHGIVSFGSR
LGCNYYHKPSVFTRVSNYIDWINS[illegible]
4. ELASHM7
E1403370A : pancreatic el[illegible]e IIIA
MMLRLLSSLLLVAVASGYGPPSSHSS[illegible]VVHGEDAVPYSWPWQVSLQYEKSGSFYHTCGG
SLIAPDWVVTAGHCISR[illegible]LTYQVVLGEYNLAVKEGPEQVIPINSEELFVHPLWNRSCVAC
GNDIALIKLSRSAQLGDAVQLASLPPAGDILPNKTPCYITGWGRLYTNGPLPDKLQQARL
PVVDYKHCSRWNWWGSTVKKTMVCAGGYIRSGCNGDSGGPLNCPTEDGGWQVHGVTSFVS
AFGCNFIWKPTVFTRVSAFIDWIEETIASH
5. ELASHM8
E1403370B : pancreatic elastase IIIB
MMLRLLSSLLLVAVASGYGPPSSRPSSRVVNGEDAVPYSWPWQVSLQYEKSGSFYHTCGG
SLIAPDWVVTAGHCISSSRTYQVVLGEYDRAVKEGPEQVIPINSGDLFVHPLWNRSCVAC
GNDIALIKLSRSAQLGDAVQLASLPPAGDILPNETPCYITGWGRLYTNGPLPDKLQEALL
PVVDYEHCSRWNWWGSSVKKTMVCAGGDIRSGCNGDSGGPLNCPTEDGGWQVHGVTSFVS
AFGCNTRRKPTVFTRVSAFIDWIEETIASH
6. ELASMSE1
E1301329A : elastase II
MIRTLLLSALVAGALSCGYPTYEVEDDVSRVVGGQEATPNTWPWQVSLQVLSSGRWRHNC
GGSLVANNWVLTAAHCLSNYQTYRVLLGAHSLSNPGAGSAAVQVSKLVVHQRWNSQNVGN
GYDIALIKLASPVTLSKNIQTACLPPAGTILPRNYVCYVTGWGLLQTNGNSPDTLRQGRL

LVVDYATCSSASWWGSSVKSSMVCAGGDGVTSSCNGDSGGPLNCRASNGQWQVHGIVSFG
SSLGCNYPRKPSVFTRVSNYIDWINSVMARN

7. ELASPIG2

E1306262C : elastase II

MIRALLLSTLVAGALSCGLPANLPQLPRVVGGEDARPNSWPWQVSLQYDSSGQWRHTCGG
TLVDQSWVLTAAHCISSSRTYRVVLGRHSLSTNEPGSLAVKVSKLVVHQDWNSNQLSNGN
DIALLKLASPVSLTDKIQLGCLPAAGTILPNNYVCYVTGWGRLQTNGASPDILQQGQLLV
VDYATCSKPGWWGSTVKTNMICAGGDGIISSCNGDSGGPLNCQGANGQWQVHGIVSFGSS
LGCNYYHKPSVFTRVSNYIDWINSVIANN

8. ELASRAT2

Proelastase precursor (EC 3.4.21.11) I - Rat

VVGGAEARRNSWPSQISLQYLSGGSWYHTCGGTLIRRNWVMTAAHCVSSQMTFRVVVGDH
NLSQNDGTEQYVSVQKIMVHPTWNSNNVAAGYDIALLRLAQSVTLNNYVQLAVLPQEGTI
LANNNPCYITGWGRTRTNGQLSQTLQQAYLPSVDYSICSSSSYWGSTVKTTMVCAGGDGV
RSGCQGDSGGPLHCLVNGQYSVHGVTSFVSSMGCNVSKKPTVFTRVSAYISWMNNVIAYT

9. ELASRAT3

Proelastase precursor (EC 3.4.21.11) II - Rat

VVGGQEASPNSWPWQVSLQYLSSGKWHHTCGGSLVANNWVLTAAHCISNSRTYRVLLGRH
SLSTSESGSLAVQVSKLVVHEKWNAQKLSNGNDIALVKLASPVALTSKIQTACLPPAGTI
LPNNYPCYVTGWGRLQTNGATPDVLQQGRLLVVDYATCSSASWWGSSVKTNMVCAGGDGV
TSSCNGDSGGPLNCQASNGQWQVHGIVSFGSTLGCNYPRKPSVFTRVSNYIDWINSVIAK
N

10. KALIGP1

Z1309183A : kallikrein

VIGGQECARDSHPWQAAVYYYSDIKCGGVLVDPQWVLTAAHCINDSNQVKLGRHNLFEDE
DTAQHFLVSQSVPHPDFYMSLLEPHNVLPNEDYSHDLMLLRLNQPAQITDSVQVMPLPTQ
EVQVGTTCRALGWGSIDPDPAHPVFPDELQCVGLEILPSKNCDDAHIAYVTGTMLCAGDL
AGGKDTCVGDSGGPLICDGVLQGLTSWGDSPCGVAHSPSLYTKVIEYREWIERTMADNP

11. KALIHM3

Z1404431A : urinary kallikrein

IVGGWECEQHSQPWQAALYHFSTFQCGGILVHRQWVLTAAHCISDNYQLWLGRHNLFDDE
NTAQFVHVSESFPHPGFNMSLLENHTRQADEDYSHDLMLLRLTEPADTITDAVKVVELPT
QEPEVGSTCLASGWGSIEPENFSFPDDLQCVDLKILPNDECEKAHVQKVTDFMLCVGHLE
GGKDTCVGDSGGPLMCDGVLQGVTSWGYVPCGTPNKPSVAVRVLSYVKWIEDTIAENS

12. KALIMSE3

Z1205271A : kallikrein,glandular

MRFLILFLALSLGGIDAAPPVQSRIVGGFNCEKNSQPWQVAVYRFTKYQCGGILLNANWV
LTAAHCHNDKYQVWLGKNNFLEDEPSAQHRLVSKAIPHPDFNMSLLNEHTPQPEDDYSND
LMLLRLKKPADITDVVKPIDLPTEEPKLGSTCLASGWGSITPVKYEYPDELQCVNLKLLP
NEDCAKAHIEKVTDDMLCAGDMDGGKDTCAGDSGGPLICDGVLQGITSWGPRPCGKPNVP
GIYTRVLNFNTWIRETMAEND

13. KALIMSE4

Z1208372A : kallikrein PMF2

MRFLILFLALSLGGIDAAPPLQSRVVGGFNCEKNSQPWQVAVYDNKEHICGGVLLERNWV
LTAAHCYVDQYEVWLGKNKLFQEEPSAQHRLVSKSFPHPGFNMSLLTLKEIPPGADFSND
LMLLRLSKPADITDAVKPITLPTKESKLGSTCLASGWGSITPTKWQKPDDLQCVFLKLLP
IKNCIENHNVKVTDVMLCAGEMSGGKNICKGDSGGPLICDSVLQGITSTGPIPCGKPGVP
AMYTNLIKFNSWIKDTMTKNS

14. KALIMSE5

Z1313201A : glandular kallikrein

MRFLILFLTLSLGGIDAAPPVQSRILGGFKCEKNSQPWQVAVYYLDEYLCGGVLLDRNWV
LTAAHCYEDKYNIWLGKNKLFQDEPSAQHRLVSKSFPHPDFNMSLLQSVPTGADLSNDLM
LLRLSKPADITDVVKPIDLPTTEPKLGSTCLASGWGSINQLIYQNPNDLQCVSIKLHPNE
VCVKAHILKVTDVMLCAGEMNGGKDTCKGGSGGPLICDGVLQGITSWGSTPCGEPNAPAI
YTKLIKFTSWIKDTMAKNP

15. KALIMSE6

Z1313201B : glandular kallikrein

MWFLILFLALSLGGIDAAPPLQSRVVGGFNCKKNSQPWQVAVYYQKEHICGGVLLDRNWV
LTAAHCYVDQYEVWLGKNKLFQEEPSAQHRLVSKSFPHPGFNMSLLMLQTIPPGADFSND
LMLLRLSKPADITDVVKPIALPTKEPKPGSKCLASGWGSITPTRWQKPDDLQCVFITLLP
NENCAKVYLQKVTDVMLCAGEMGGGKDTCRDDSGGPLICDGILQGTTSYGPVPCGKPGVP

AIYTNLIKFNSWIKDTMMKNA

16. KALIMSE7

Z1313201C : glandular kallikrein

MRFLILFLALSLGGIDAAPPVHSRIVGGFKCEKNSQPWHVAVYRYNEYICGGVLLDANWV
LTAAHCYYEENKVSLGKNNLYEEEPSAQHRLVSKSFLHPGYNRSLHRNHIRHPEYDYSND
LMLLRLSKPADITDVVKPIALPTEEPKLGSTCLASGWGSTTPFKFQNAKDLQCVNLKLLP
NEDCGKAHIEKVTDVMLCAGETDGGKDTCKGDSGGPLICDGVLQGITSWGFTPCGEPKKP
GVYTKLIKFTSWIKDTMAKNL

17. KALIMSE8

Z1402323A : glandular kallikrein

MWFLILFLALSLGGIDAAPPVQSRIFGGFNCEKNSQPWQVAVYRFTKYQCGGVLLNANWV
LTAAHCHNDKYQVWLGKNNFFEDEPSAQHRLVSKAIPHPDFNMSLLNEHTPQPEDDYSND
LMLLRLKKPADITDVVKPIDLPTEEPKLGSTCLASGWGSITPVIYEPADDLQCVNFKLLP
NEDCVKAHIEKVTDVMLCAGDMDGGKDTCMGDSGGPLICDGVLHGITSWGPSPCGKPNVP
GIYTKLIKFNSWIKDTIAKNA

18. KALIPIG1

Tissue kallikrein (EC 3.4.21.35), pancreatic - Pig

IIGGRECEKNSHPWQVAIYHYSSFQCGGVLVNPKWVLTAAHCKNDNYEVGWLRHNLFENE
NTAQFFGVTADFPHPGFNLSADGKDYSHDLMLLRLQSPAKITDAVKVLELPTQEPELGST
CEASGWGSIEPGPDDFEFPDEIQCVQLTLLQNTFCAHABPBKVTESMLCAGYLPGGKDTC
MGDSGGPLICNGMWQGITSWGHTPCGSANKPSIYTKLIFYLDWIBBTITENP

19. KALIRAT4

Z1110144C : kallikrein S3

MWFLILFLALSLGQIDAAPPGQSRVVGGYNCETNSQPWQVAVIGTTFCGGVLIDPSWVIT
AAHCYSKNYRVLLGRNNLVKDEPFAQRRLVSQSFQHPDYIPVFMRNHTRQRAYDHNNDLM
LLHLSKPADITGGVKVIDLPTEEPKVGSICLASGWGMTNPSEMKLSHDLQCVNIHLLSNE
KCIETYKNIETDVTLCAGEMDGGKDTCTGDSGGPLICDGVLQGLTSGGATPCAKPKTPAI
YAKLIKFTSWIKKVMKENP

20. SUBTBAC1

E671056A : subtilisin

AQSVPYGVSQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDLKVAGGASMVPSETPN QD
DNSHGTHVAGTVAALNNSIGVLGVAPSSALYAVKVLGDAGSGQYSWIINGIEWAIANNMD
VINMSLGGPSGSAALKAAVDKAVASGVVVVAAAGNEGSTGSSSTVGYPGKYPSVIAVGAV
DSSNQRASFSSVGPELDVMAPGVSIQSTLPGNKYGAYNGTSMASPHVAGAAALILSKHPN
WTNTQVRSSLQNTTTKLGDSFYYGKGLINVQAAAQ

21. SUBTBAC2

E721944A : subtilisin

AQSVPYGISQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDLNVRGGASFVPSETPNYQD
GSSHGTHVAGTIAALNNSIGVLGVAPSSALYAVKVLDSTGSGQYSWIINGIEWAISNNMD
VINMSLGGPSGSTALKTVVDKAVSSGIVVAAAAGNEGSSGSSSTVGYPAKYPSTIAVGAV
NSSNQRASFSSAGSELDVMAPGVSIQSTLPGGTYGAYNGTSMATPHVAGAAALILSKHPT
WTNAQVRDRLESTATYLGDSFYYGKGLINVQAAAQ

22. SUBTBAC3

Z0912207A : subtilisin DY

AQTVPYGIPLIKADKVQAQGYKGANVKVGIIDTGIAASHTDLKVVGGASFVSGESYNTDG
NGHGTHVAGTVAALDNTTGVLGVAPNVSLYAIKVLNSSGSGTYSAIVSGIEWATQNGLDV
INMSLGGPSGSTALKQAVDKAYASGIVVVAAAGNSGSSGSQNTIGYPAKYDSVIAVGAVD
SNKNRASFSSVGAELEVMAPGVSVYSTYPSNTYTSLNGTSMASPHVAGAAALILSKYPTL
SASQVRNRLSSTATNLGDSFYYGKGLINVEAAAQ

23. SUBTBAC6

Subtilisin Carlsberg (EC 3.4.21.14) - Bacillus subtilis

AQTVPYGIPLIKADKVQAQGFKGANVKVAVLDTGIQASHPDLNVVGGASFVAGEAYNTDG
NGHGTHVAGTVAALDNTTGVLGVAPSVSLYAVKVLNSSGSGSYSGIVSGIEWATTNGMDV
INMSLGGASGSTAMKQAVDNAYARGVVVVAAAGNSGNSGSTNTIGYPAKYDSVIAVGAVD
SNSNRASFSSVGAELEVMAPGAGVYSTYPTNTYATLNGTSMASPHVAGAAALILSKHPNL
SASQVRNRLSSTATYLGSSFYYGKGLINVEAAAQ

24. TRPCRAY1

Trypsin (EC 3.4.21.4) I - Crayfish

IVGGTDAVLGEFPYQLSFQETFLGFSFHFCGASIYNENYAITAGHCVYGDDYENPSGLQI
VAGELDMSVNEGSEQTITVSKIILHENFDYDLLDNDISLLKLSGSLTFNNNVAPIALPAQ
GHTATGNVIVTGWGTTSEGGNTPDVLQKVTVPLVSDAECRDDYGADEIFDSMICAGVPEG

GKDSCQGDSGGPLAASDTGSTYLAGIVSWGYGCARPGYPGVYTEVSYHVDWIKANAV

25. TRPDOG1

Trypsinogen (EC 3.4.21.4), cationic, precursor - Dog

IVGGYTCSRNSVPYQVSLNSGYHFCGGSLINSQWVVSAAHCYKSRIQVRLGEYNIAVSEG
GEQFINAAKIIRHPRYNANTIDNDIMLIKLSSPATLNSRVSAIALPKSCPAAGTQCLISG
WGNTQSIGQNYPDVLQCLKAPILSDSVCRNAYPGQISSNMMCLGYMEGGKDSCQGDSGGP
VVCNGELQGVVSWGAGCAQKGKPGVSPKVCKYVSWIQQTIAAN

26. TRPDOG2

Trypsinogen (EC 3.4.21.4), anionic, precursor - Dog

IVGGYTCEENSVPYQVSLNAGYHFCGGSLISDQWVVSAAHCYKSRIQVRLGEYNIDVLEG
NEQFINSAKVIRHPNYNSWILDNDIMLIKLSSPAVLNARVATISLPRACAAPGTQCLISG
WGNTLSSGTNYPELLQCLDAPILTQAQCEASYPGQITENMICAGFLEGGKDSCQGDSGGP
VVCNGELQGIVSWGYGCAQKNKPGVYTKVCNFVDWIQSTIAANS

27. TRPFISH1

Trypsinogen (EC 3.4.21.4) - Spiny dogfish

IVGGYECPKHAAPWTVSLNVGYHFCGGSLIAPGWVVSAAHCYQRRIQVRLGEHDISANEG
DETYIDSSMVIRHPNYSGYDLDNDIMLIKLSKPAALNRNVDLISLPTGCAYAGEMCLISG
WGNTMDGAVSGDQLQCLDAPVLSDAECKGAYPGMITNNMMCVGYMEGGKDSCQGDSGGPV
VCNGMLQGIVSWGYGCAERDHPGVYTRVCHYVSWIHETIASV

28. TRPFISH2

E750677A : trypsin

IVGGYECPKHAAPWTVSLNVGYHFCGGSLIAPGWVSAAHCYQRRIQVRLGEHDISANEGD
ETYIDSMVIRHPNYSGYDLDNDIMLIKLSKPAALNRNVDLISLPTGCAYAGEMCLISGWG
NTMDGAVSGDQLQCLDAPVLSDAECKGAYPGMITNNMMCVGYMEGGKDSCQGDSGGPVVC
NGMLQGIVSWGYGCAERDHPGVYTRVCHYVSWIHETIASV

29. TRPFLY1

Trypsinogen-like (EC 3.4.21.-) proenzyme precursor - Fruit fly

IVGGSATTISSFPWQISLQRSGSHSCGGSIYSANIIVTAAHCLQSVSASVLQVRAGSTYW
SSGGVVAKVSSFKNHEGYNANTMVNDIAVIRLSSSLSFSSSIKAISLATYNPANGASAAV
SGWGTQSSGSSSIPSQLQYVNVNIVSQSQCASSTYGYGSQIRNTMICAAASGKDACQGDS
GGPLVSGGVLVGVVSWGYGCAYSNYPGVYADVAVLRSWVVSTANSI

30. TRPHM1

E1403437A : cell specific trypsin like protease

MRNSYRFLASSLSVVVSLLLIPEDVCEKIIGGNEVTPHSRPYMVLLSLDRKTICAGALIA
KDWVLTAAHCNLNKRSQVILGAHSITREEPTKQIMLVKKEFPYPCYDPATREGDLKLLQL
TEKAKINKYVTILHLPKKGDDVKPGTMCQVAGWGRTHNSASWSDTLREVNITIIDRKVCN
DRNHYNFNPVIGMNMVCAGSLRGGRDSCNGDSGSPLLCEGVFRGVTSFGLENKCGDPRGP
GVYILLSKKHLNWIIMTIKGAV

31. TRPHOR1

Chymotrypsin II (EC 3.4.21.1) - Oriental hornet

IVGGTNAPRGKYPYQVSLRAPKHFCGGSISKRYVLTAAHCLVGKSEHQVTVGSVLLNKEE
AVYNAKELIVNKNYNSIRLINDIGLIRVSKDISFTQLVQPVKLPVSNTIKAGDPVVLTGW
GRIYVNGPIPNNLQQITLSIVNQQTCKSKHWGLTDSQICTFTKRGEGACHGDSGGPLVAN
GVQIGIVSYGHPCAIGSPNVFTRVYSFLDWIQKNQL

32. TRPHOR2

Chymotrypsin II (EC 3.4.21.1) - European hornet

IVGGTDAPRGKYPYQVSLRAPKHFCGGSISKRYVLTAAHCLVGKSKHQVTVHAGSVLLNK
EEAVYNAEELIVNKNYNSIRLINDIGLIRVSKDISYTQLVQPVKLPVSNTIKAGDPVVLT
GWGRIYVNGPIPNNLQQITLSIVNQQTCKFKHWGLTDSQICTFTKLGEGACDGDSGGPLV
ANGVQIGIVSYGHPCAVGSPNVFTRVYSFLDWIQKNQL

33. TRPMSE1

Z1301329B : trypsin

MSALLILALVGAAVAFPVDDDDKIVGGYTCRESSVPYQVSLNAGYHFCGGSLINDQWVVS
AAHCYKYRIQVRLGEHNINVLEGNEQFVDSAKIIRHPNYNSWTLDNDIMLIKLASPVTLN
ARVASVPLPSSCAPAGTQCLISGWGNTLSNGVNNPDLLQCVDAPVLPQADCEASYPGDIT
NNMICVGFLEGGKDSCQGDSGGPVVCNGELQGIVSWGYGCAQPDAPGVYTKVCNYVDWIQ
NTIADN

34. TRPPIG1

Trypsinogen (EC 3.4.21.4) - Pig

IVGGYTCAANSIPYQVSLNSGSHFCGGSLINSQWVVSAAHCYKSRIQVRLGEHNIDVLEG
NEQFINAAKIITHPNFNGNTLDNDIMLIKLSSPATLNSRVATVSLPRSCAAAGTECLISG

WGNTKSSGSSYPSLLQCLKAPVLSDSSCKSSYPGQITGNMICVGFLEGGKDSCQGDSGGP
VVCNGQLQGIVSWGYGCAQKNKPGVYTKVCNYVNWIQQTIAAN

35. TRPRAT1

Trypsinogen I (EC 3.4.21.4) precursor - Rat

IVGGYTCPEHSVPYQVSLNSGYHFCGGSLINDQWVVSAAHCYKSRIQVRLGEHNINVLEG
DEQFINAAKIIKHPNYSSWTLNNDIMLIKLSSPVKLNARVAPVALPSACAPAGTQCLISG
WGNTLSNGVNNPDLLQCVDAPVLSQADCEAAYPGEITSSMICVGFLEGGKDSCQGDSGGP
VVCNGQLQGIVSWGYGCALPDNPGVYTKVCNFVGWIQDTIAAN

36. TRPRAT2

Trypsinogen II (EC 3.4.21.4) precursor - Rat

IVGGYTCQENSVPYQVSLNSGYHFCGGSLINDQWVVSAAHCYKSRIQVRLGEHNINVLEG
DEQFINAAKIIKHPNFDRKTLNNDIMLIKLSSPVKLNARVATVALPSSCAPAGTQCLISG
WGNTLSSGVNEPDLLQCLDAPLLPQADCEASYPGKITDNMVCVGFLEGGKDSCQGDSGGP
VVCNGELQGIVSWGYGCALPDNPGVYTKVCNYVDWIQDTIAAN

37. TRPRAT3

Chymotrypsinogen B (EC 3.4.21.1) precursor - Rat

IVNGEDAIPGSWPWQVSLQDKTGFHFCGGSLISEDWVVTAAHCGVKTSDVVVAGEFDQGS
DEENIQVLKIAQVFKNPKFNMFTVRNDITLLKLATPAQFSETVSAVCLPNVDDDFPPGTV
CATTGWGKTKYNALKTPEKLQQAALPIVSEADCKKSWGSKITDVMTCAGASGVSSCMGDS
GGPLVCQKDGVWTLAGIVSWGSGVCSTSTPAVYSRVTALMPWVQQILEAN

38. TRPSTRP1

Trypsin (EC 3.4.21.4) - Streptomyces griseus

VVGGTRAAQGEFPFMVRLSMGCGGALYAQDIVLTAAHCVSGSGNNTSITATGGVVDLQSA
VKVRSTKVLQAPGYNGTGKDWALIKLAQPINQPTLKIATTTAYNQGTFTVAGWGANREGG
SQQRYLLKANVPFVSDAACRSAYGNELVANEEICAGYPDTGGVDTCQGDSGGPMFRKDNA
DEWIQVGIVSWGYGCARPGYPGVYTEVSTFASAIASAARTL

ASPARTIC PROTEASES (47 sequences)

1. AID1

Protease (EC 3.4.) - Simian AIDS retrovirus SRV-1

SRKPTTTPSGKRTEGPAPGPETSLWGGQLCSSQQKQPISKLTRATPGSAGLDLSSTSHTV
LTPEMGPQALSTGIYGPLPPNTFGLILGRSSITIKGLQVYPGVIDNDYTGEIKIMAKAVN
NIVTVPQGNRIAQLILLPLIETDNKVQQPYRGQGSFGSSDIYWVQPITCQKPSLTLWLDD
KMETGLIDTGADVTIIKLEDWPPNWPITDTLTNLRGIGQSNNPKQSSKYLTWRDKENNSG
LIKPFVIPNLPVNLWGRDLLSQMKIMMCSPSDIVTAQMLAQGYSPGKGLGKNENGILHPI
PNQGQFDKKGFGNF

2. AID10

sor protein - Human immunodeficiency virus (HIV-2, isolate ROD)

MEEDKRWIVVPTWRVPGRMEKWHSLVKYLKYKTKDLEKVCYVPHHKVGWAWWTCSRVIFP
LKGNSHLEIQAYWNLTPEKGWLSSYSVRITWYTEKFWTDVTPDCADVLIHSTYFPCFTAG
EVRRAIRGEKLLSCCNYPRAHRAQVPSLQFLALVVVQQNDRPQRDSTTRKQRRRDYRRGL
RLAKQDSRSHKQRSSESPTPRTYFPGVAEVLEILA

3. AID11

nef protein - Human immunodeficiency virus I (HIV-1, isolate BR)

MGGKWSKMAGWSTVRERMRRAEPARERMRRAEPRAEPAADGVGAVSRDLEKHGAITSSNT
AATNADCAWLEAQEDEEVGFPVKPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIHSQQRQD
ILDLWVYHTQGYFPDWQNYTPGPGVRYPLTFGWCFKLVPVEPEKIEEANEGENNSLLHPM
SQHGMDDPEREVLQWRFDSRLAFHHMARELHPEYYKNC

4. AID12

nef protein - Human immunodeficiency virus (isolate HIV-Zr6)

MGGRWSKSSIVGWPAIRERIRRTDPRRTDPAADGVGAASRDLEKHGAITSSNTRDTNADC
AWLEAQEESEEVGFPVRPQVPLRPMTYKLAVDLSHFLKEKGGLEGLIWSKKRQEILDLWV
YNTQGIFPDWQNYTPGPGIRYPLTFGWCFELVPVDPREVEEATEGETNCLLHPVCQHGME
DTEREVLKWRFNSRLAFEHKAREMHPEFYKDC

5. AID13

nef protein - Human immunodeficiency virus (HIV-2, isolate ROD)

MGASGSKKHSRPPRGLQERLLRARAGACGGYWNESGGEYSRFQEGSDREQKSPSCEGRQY
QQGDFMNTPWKDPAAEREKNLYRQQNMDDVDSDDDDQVRVSVTPKVPLRPMTHRLAIDMS
HLIKTRGGLEGMFYSERRHKILNIYLEKEEGIIADWQNYTHGPGVRYPMFFGWLWKLVPV

DVPQEGEDTETHCLVHPAQTSKFDDPHGETLVWEFDPLLAYSYEAFIRYPEEFGHKSGLP
EEEWKARLKARGIPFS

6. AID2

gag polyprotein - Human immunodeficiency virus (HIV-1, isolate CDC-451)

MGARASVLSGGELDRWEKIRLRPGGKKQYRLKHIVWASRKLERFAVNPGLLETSKGCRQI
LGQLQPSLQTGSEELRSLYNTVATLYCVHQRIEVRDTKEALDKIEEEQNKSKKKAQQAAA
DTGNSSQVSQNYPIVQNLQGQMVHQAISPRTLNAWVKVIEEKAFSPEVIPMFAALSEGAT
PQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRLHPVHAGPIAPGQMREPRGSDIAGTT
STLQEQIGWMTNNPPTPVGEIYKRWIILGLNKIVRMYSPISILDIRQGPKEPFRDYVDRF
YKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGVGGPGHKA
RVLAEAMSQVTNSATIMMQRGNFRRQGKTVKCFNCGKEGHIARNCKAPRKKGCWKCGREG
HQMKDCTERQANFLGKIWPSHKGRPGNFLQSRPEPTAPPEESFRFGDETTTPSQKQEPRD
KELYPLASLRSLFGNDPSSQ

7. AID3

gag polyprotein - Human immunodeficiency virus (HIV-2, isolate ROD)

MGARNSVLRGKKADELERIRLRPGGKKKYRLKHIVWAANKLDRFGLAESLLESKEGCQKI
LTVLDPMVPTGSENLKSLFNTVCVIWCIHAEEKVKDTEGAKQIVRRHLVAETGTAEKMPS
TSRPTAPSSEKGGNYPVQHVGGNYTHIPLSPRTLNAWVKLVEEKKFGAEVVPGFQALSEG
CTPYDINQMLNCVGDHQAAMQIIREIINEEAAEWDVQHPIPGPLPAGQLREPRGSDIAGT
TSTVEEQIQWMFRPQNPVPVGNIYRRWIQIGLQKCVRMYNPTNILDIKQGPKEPFQSYVD
RFYKSLRAEQTDPAVKNWMTQTLLVQNANPDCKLVLKGLGMNPTLEEMLTACQGVGGPGQ
KARLMAEALKEVIGPAPIPFAAAQQRKAFKCWNCGKEGHSARQCRAPRRQGCWKCGKPGH
IMTNCPDRQAGFLGLGPWGKKPRNFPVAQVPQGLTPTAPPVDPAVDLLEKYMQQGKRQRE
QRERPYKEVTEDLLHLEQGETPYREPPTEDLLHLNSLFGKDQ

8. AID4

pol polyprotein - Human immunodeficiency virus (HIV-2, isolate ROD)

TGRFFRTGPLGKEAPQLPRGPSSAGADTNSTPSGSSSGSTGEIYAAREKTERAERETIQG
SDRGLTAPRAGGDTIQGATNRGLAAPQFSLWKRPVVTAYIEGQPVEVLLDTGADDSIVAG
IELGNNYSPKIVGGIGGFINTKEYKNVEIEVLNKKVRATIMTGDTPINIFGRNILTALGM
SLNLPVAKVEPIKIMLKPGKDGPKLRQWPLTKEKIEALKEICEKMEKEGQLEEAPPTNPY
NTPTFAIKKKDKNKWRMLIDFRELNKVTQDFTEIQLGIPHPAGLAKKRRITVLDVGDAYF
SIPLHEDFRPYTAFTLPSVNNAEPGKRYIYKVLPQGWKGSPAIFQHTMRQVLEPFRKANK
DVIIIQYMDDILIASDRTDLEHDRVVLQLKELLNGLGFSTPDEKFQKDPPYHWMGYELWP
TKWKLQKIQLPQKEIWTVNDIQKLVGVLNWAAQLYPGIKTKHLCRLIRGKMTLTEEVQWT
ELAEAELEENRIILSQEQEGHYYQEEKELEATVQKDQENQWTYKIHQEEKILKVGKYAKV
KNTHTNGIRLLAQVVQKIGKEALVIWGRIPKFHLPVEREIWEQWWDNYWQVTWIPDWDFV
STPPLVRLAFNLVGDPIPGAETFYTDGSCNRQSKEGKAGYVTDRGKDKVKKLEQTTNQQA
ELEAFAMALTDSGPKVNIIVDSQYVMGISASQPTESESKIVNQIIEEMIKKEAIYVAWVP
AHKGIGGNQEVDHLVSQGIRQVLFLEKIEPAQEEHEKYHSNVKELSHKFGIPNLVARQIV
NSCAQCQQKGEAIHGQVNAELGTWQMDCTHLEGKIIIVAVHVASGFIEAEVIPQESGRQT
ALFLLKLASRWPITHLHTDNGANFTSQEVKMVAWWIGIEQSFGVPYNPQSQGVVEAMNHH
LKNQISRIREQANTIETIVLMAIHCMNFKRRGGIGDMTPSERLINMITTEQEIQFLQAKN
SKLKDFRVYFREGRDQLWKGPGELLWKGEGAVLVKVGTDIKIIPRRKAKIIRDYGGRQEM
DSGSHLEGAREDGEMA

9. AID5

env polyprotein - Human immunodeficiency virus I (HIV-1, isolate BR)

MRVKGIKKNYQHLWRWGGMMLLGILMICSATDKLWVTVYYGVPVWKEANTTLFCASDAKA
YDTEIHNVWATHACVPTDPNPQELVMGNVTENFNMWKNDMVEQMHEDIISLWDQSLKPCV
KLTPLCLTLNCHDFNATNATSNSGKMMEGGEMKNCSFNITTSIRDKMQKEYALFYKLDIV
PIDNDKTNTRYRLISCNTSVITQACPKVTFEPIPIHYCAPAGFAILKCNNKKFNGTGPCT
NVSTVQCTHGIRPVVSTQLLLNGSLAEEEVVIRSENFTNNVKTIIVQLNESVEINCTRPN
NNTRKRITMGPGRVYYTTGQIIGDIRRAHCNLSRSKWENTLKQIVTKLRVQFKNKTIVFN
RSSGGDPEIVMHSFNCGGEFFFCNTTQLFNSTWYRNTTGNITEGNSPITLPCRIKQIINM
WQEVGKAMYAPPIRGQIKCSSNITGLLLTRDGGNNNETTDTEIFRPGGGNMRDNWRSELY
KYKVVKIEPLGVAPTKAKRRVVQREKRAVGLGALFLGFLGAAGSTMGAASLTLTVQARLL
LSGIVQQQNNLLMAIEAQQHMLELTVWGIKQLQARVLAVERYLKDQQLLGIWGCSGKLIC
TTAVPWNASWSNKSLSDIWDNMTWMEWEREIDNYTNLIYSLIEDSQIQQEKNEKELLELD
KWASLWNWFNITNWLWYIKIFIMIVGGLIGLRIVFAVLSIVNRVRQGYSPLSFQTRLPGR
RGPDRPEGIEEEGGERDRDRSSPLVDGFLALFWVDLRSLFLFSYHRLRDLLLIVTRIVEL
LGRRGWEVLKYWWNLLQYWSQELKNSAVSLLNATAIAVGERTDRAIEVVQRAFRAILHIP
RRIRQGLERALQ

10. AID6

env polyprotein - Human immunodeficiency virus (HIV-1, isolate CDC-451)

MAMRAKGIRKNCQHLWRWGTMLLGMLMICSAAANLWVTVYYGVPVWKEATTTLFCASDAK
AYDTEAHNVWATHACVPTNPNPQEVVLENVTENFNMWKNNMVEQMHEDIISLWDQSLKPC
VKLTPLCVTLNCTDLNTNNTTNTTELSIIVVWEQRGKGEMRNCSFNITTSIRDKVQREYA
LFYKLDVEPIDDNKNTTNNTKYRLINCNTSVITQACPKVSFEPIPIHYCTPTGFALLKCN
DKKFNGTGPCTNVSTVQCTHGIRPVVSTQLLLNGSLAEEEVVIRSENFTNNAKTIIVQLN
VSVEINCTRPNNHTRKRVTLGPGRVWYTTGEILGNIRQAHCNISRAQWNNTLQQIATTLR
EQFGNKTIAFNQSSGGDPEIVMHSFNCGGEFFYCNSTQLFNSAWNVTSNGTWSVTRKQKD
TGDIITLPCRIKQIINRWQVVGKAMYALPIKGLIRCSSNITGLLLTRDGGGENQTTEIFR
PGGGDMRDNWRSELYKYKVVKIEPLGVAPTKAKRRVVQREKRAVGMLGAMFLGFLGAAGS
TMGATSMALTVQARQLLSGIVQQQNNLLRAIKAQQHLLQLTVWGIKQLQARILAVERYLK
DQQLLGFWGCSGKLICTTAVPWNASWSNKTLDQIWNNMTWMEWDREIDNYTHLIYTLIEE
SQNQQEKNQQELLQLDKWASLWTWSDITKWLWYIKIFIMIVGGLIGLRIVFAVLSIVNRV
RQGYSPLSFQTLLPNPRGPDRPEGTEEGGGERGRDGSTRLVHGFLALVWDDLRSLCLFSY
HRLRDLLLIVARIVELLGRRGWEVLK[illegible]SLKNSAVSLVNVTAIAVAEGTDR
VIEVVQRIYRAFLHIPRRIRQGFEP

11. AID7

env polyprotein precurs[illegible]odeficiency virus (isolate HIV-
Zr6)

LLGILMICSAADNLWVTVYYGVP[illegible]GDAKSYKTEAHNIWATHACVPTDPN
PQEIELENVTENFNMWRNNMVEQI[illegible]SLKPCVKLTPLCVTLNCTDESDEWMG
NVTGKNVTEDIRMKNCSFNITTVVRL[illegible]LFYRLDIVPIDNDNSTNSTNYRLINCN
TSAITQACPKVSFEPIPIHYCAPAGFA[illegible]RDKRFNGTGPCTNVSTVQCTHGIRPVVST
QLLLNGSLAEEEIIIRSENLTNNAKIIIVQLNESVAINCTRPYKNTRQSTPIGLGQALYT
TRGRTKIIGQAHCNISKEDWNKTLQRVAIKLGNLLNKTTIIFKPSSGGDAEITTHSFNCG
GEFFYCNTSGLFNSTWNINNSEGANSTESDNKLITLQCRIKQIINMWQGVGKAMYAPPIE
GQINCSSNITGLLLTRDGGTNNSSNETFRPGGGDMRDNWRSELYKYKVVKIEPLGVAPTK
AKRRVVEREKRAIGLGAMFLGFLGAAGSTMGAASVTLTVQARQLMSGIVQQQNNLLRAIE
AQQHLLQLTVWGIKQLQARILAVERYLKDQQLLGIWGCSGKLICTTTVPWNSSWSNRSLN
DIWQNMTWMEWEREIDNYTGLIYRLIEESQTQQEKNEQELLELDKWASLWNWFNITQWLW
YIKIFIMIVGGLIGLRIVFAVLSLVNRVRQGYSPLSFQTLLPAPREPDRPEGIEEEGGER
GRDRSIRLVNGFSALIWDDLRNLCLFSYHRLRDLILIAARIVELLGRRGWEALKYLWNLL
QYWSRELRNSASSLLDTIAIAVAEGTDRVIEIVRRTYRAVLNVPTRIRQGLERLLL

12. AID8

env polyprotein precursor - Human immunodeficiency virus (HIV-2, isolate
ROD)

YCTQYVTVFYGVPTWKNATIPLFCATRNRDTWGTIQCLPDNDDYQEITLNVTEAFDAWNN
TVTEQAIEDVWHLFETSIKPCVKLTPLCVAMKCSSTESSTGNNTTSKSTSTTTTTPTDQE
QEISEDTPCARADNCSGLGEEETINCQFNMTGLERDKKKQYNETWYSKDVVCETNNSTNQ
TQCYMNHCNTSVITESCDKHYWDAIRFRYCAPPGYALLRCNDTNYSGFAPNCSKVVASTC
TRMMETQTSTWFGFNGTRAENRTYIYWHGRDNRTIISLNKYYNLSLHCKRPGNKIVKQIM
LMSGHVFHSHYQPINKRPRQAWCWFKGKWKDAMQEVKETLAKHPRYRGTNDTRNISFAAP
GKGSDPEVAYMWTNCRGEFLYCNMTWFLNWIENKTHRNYAPCHIKQIINTWHKVGRNVYL
PPREGELSCNSTVTSIIANIDWQNNNQTNITFSAEVAELYRLELGDYKLVEITPIGFAPT
KEKRYSSAHGRHTRGVFVLGFLGFLATAGSAMGAASLTVSAQSRTLLAGIVQQQQQLLDV
VKRQQELLRLTVWGTKNLQARVTAIEKYLQDQARLNSWGCAFRQVCHTTVPWVNDSLAPD
WDNMTWQEWEKQVRYLEANISKSLEQAQIQQEKNMYELQKLNSWDIFGNWFDLTSWVKYI
QYGVLIIVAVIALRIVIYVVQMLSRLRKGYRPVFSSPPGYIQQIHIHKDRGQPANEETEE
DGGSNGGDRYWPWPIAYIHFLIRQLIRLLTRLYSICRDLLSRSFLTLQLIYQNLRDWLRL
RTAFLQYGCEWIQEAFQAAARATRETLAGACRGLWRVLERIGRGILAVPRRIRQGAEIAL
L

13. AID9

sor protein - Human immunodeficiency virus (isolate HIV-Zr6)

MENRWQVMIVWQVDRMRIRTWKSLVKHHMYVSKKASRWFYRHHYDSPHPKISSEVHIPLG
EARLVVKTYWGLHTGERDWHLGQGVSIEWRKRRYSTQVDPGLADQLIHMYYFDCFSEAAI
RKAILGHIVSHRCEYQAGHSKVGSLQYLALTALIAPKKIKPPLPSVRKLTEDRWNKPQKT
KGHKGAIQ

14. CHYMCOW1

E0510208A : chymosin B

GEVASVPLTNYLDSQYFGKIYLGTPPQEFTVLFDTGSSDFWVPSIYCKSNACKNHQRFDP

RKSSTFQNLGKPLSIHYGTGSMQGILGYDTVTVSNIVDIQQTVGLSTQEPGDVFTYAEFD
GILGMAYPSLASEYSIPVFDNMMNRHLVAQDLFSVYMDRDGQESMLTLGAIDPSYYTGSL
HWVPVTVQQYWQFTVDSVTISGVVVACEGGCQAILDTGTSKLVGPSSDILNIQQAIGATQ
NQYGEFDIDCDNLSYMPTVVFEINGKMYPLTPSAYTSQDQGFCTSGFQSENHSQKWILGD
VFIREYYSVFDRANNLVGLAKAI

15. CHYMLAM1

lamb chymosin nar 1990

GEVASVPLTNYLDSQYFGKIYLGTPPQEFTVLFDTGSSDFWVPSIYCKSNACKNHQRFDP
RKSSTFQNLGKPLSIRYGTGSMQGILGYDTVTVSNIVDIQQTVGLSTQEPGDVFTYAEFD
GILGMAYPSLASEYSVPVFDNMMDRRLVAQDLFSVYMDRSGQGSMLTLGAIDPSYYTGSL
HWVPVTLQKYWQFTVDSVTISGAVVACEGGCQAILDTGTSKLVGPSSDILNIQQAIGATQ
NQYGEFDIDCDSLSSMPTVVFEINGKMYPLTPYAYTSQEEGFCTSGFQGENHSHQWILGD
VFIREYYSVFDRANNLVGLAKAI

16. HTLV1

Protease (EC 3.4.21.) - T-cell leukemia virus (HTLV-I)

HSTPKKLHRGGGLTSPPTLQQVFLNQDPASILPVIPLDPARRPVIKAQVDTQTSHPKTIE
ALLDTGADMTVLPIALFSSNTPLKNTSVLGAGGQTQDHFKLTSLPVLIRLPFRTTPIVLT
SCLVDTKNNWAIIGRDALQQCQGVLYLPEAKGPPVILPIQAPAVLGLEHLPRPPQISQFP
LNQNASRPCNTWSGRPWRQAISNPTPGQEITQYSQLKRPMEPGDSSTTCGPLTL

17. HTLV10

env polyprotein - T-cell leukemia virus (HTLV-II)

MGNVFFLLLFSLTHFPLAQQSRCTLTIGISSYHSSPCSPTQPVCTWNLDLNSLTTDQRLH
PPCPNLITYSGFHKTYSLYLFPHWIKKPNRQGLGYYSPSYNDPCSLQCPYLGCQAWTSAY
TGPVSSPSWKFHSDVNFTQEVSQVSLRLHFSKCGSSMTLLVDAPGYDPLWFITSEPTQPP
PTSPPLVHDSDLEHVLTPSTSWTTKILKFIQLTLQSTNYSCMVCVDRSSLSSWHVLYTPN
ISIPQQTSSRTILFPSLALPAPPSQPFPWTHCYQPRLQAITTDNCNNSIILPPFSLAPVP
PPATRRRRAVPIAVWLVSALAAGTGIAGGVTGSLSLASSKSLLLEVDKDISHLTQAIVKN
HQNILRVAQYAAQNRRGLDLLFWEQGGLCKAIQEQCCFLNISNTHVSVLQERPPLEKRVI
TGWGLNWDLGLSQWAREALQTGITILALLLLVILFGPCILRQIQALPQRLQNRHNQYSLI
NPETML

18. HTLV11

nef protein - AIDS virus HTLV-III (T-cell leukemia virus BH10)

MGGKWSKSSVVGWPAVRERMRRAEPAADGVGAASRDLEKHGAITSSNTAANNADCAWLEA
QEEEEVGFPVTPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIHSQRRQDILDLWIYHTQGY
FPDQNYTPGPGIRYPLTFGWCYKLVPVEPEKLEEANKGENTSLLHPVSLHGMDDPEREVL
EWRFDSRLAFHHMARELHPEYFKNC

19. HTLV12

nef protein - AIDS virus HTLV-III (T-cell leukemia virus 12)

WKGFCYKMGGKWSKSSVVGWPAVRERMRRAEPAADGVGAASRDLEKHGAITSSNTAANNA
ACAWLEAQEEEKVGFPVTPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIHSQRRQDILDLW
IYHTQGYFPDWQNYTPGPGIRYPLTFGWCYKLVPVEPEKLEEANKGENTSLLHPVSLHGM
DDPEREVLEWRFDSRLAFHHVARELHPEYFKNC

20. HTLV2

Protease (EC 3.4.21.) - T-cell leukemia virus I (HTLV-I, Caribbean isolate)

KTPPKSAVLPVRESRPLESGLHSASSSPWAMPPMSRSNSLEARLPPPKAHYPRTRARGGC
PPIRSPRRHPTPKKLHRGGGLTSPPTLQQVLPNQDPTSILPVIPLDPARRPVIKAQIDTQ
TSHPKTIEALLDTGADMTVLPIALFSSNTPLKNTSVLGAGGQTQDHFKLTSLPVLIRLPF
RTTPIVLTSCLVDTKNNWAIIGRDALQQCQGVLYLPEAKRPPVILPIQAPAVLGLEHLPR
PPEISQFPLNQNASRPCNTWSGRPWRQAISNPTPGQEITQYSQLKKPMEPGDSSTTCGPL
TL

21. HTLV3

gag polyprotein - T-cell leukemia virus I (HTLV-I, Caribbean isolate)

MGQIFSRSASPIPRPPRGLAAHHWLNFLQAAYRLEPGPSSYDFHQLKKFLKIALETPVWI
CPINYSLLASLLPKGYPGRVNEILHILIQTQAQIPSRPAPPPPSSSTHDPPDSDPQIPPP
YVEPTAPQVLPVMHPHGAPPNHRPWQMKDLQAIKQEVSQAAPGSPQFMQTIRLAVQQFDP
TAKDLQDLLQYLCSSLVASLHHQQLDSLISEAETRGITGYNPLAGPLRVQANNPQQQGLR
REYQQLWLAAFAALPGSAKDPSWASILQGLEEPYHAFVERLNIALDNGLPEGTPKDPILR
SLAYSNANKECQKLLQARGHTNSPLGDMLRACQAWTPKDKTKVLVVQPKKPPPNQPCFRC
GKAGHWSRDCTQPRPPPGPCPLCQDPTHWKRDCPRLKPTIPEPEPEEDALLLDLPADIPH
PKNSIGGEV

22. HTLV4

gag polyprotein - T-cell leukemia virus (HTLV-II)

MGQIHGLSPTPIPKAPRGLSTHHWLNFLQAAYRLQPRPSDFDFQQLRRFLKLALKTPIWL
NPIDYSLLASLIPKGYPGRVVEIINILVKNQVSPSAPAAPVPTPICPTTTPPPPPPPSPE
AHVPPPYVEPTTTQCFPILHPPGAPSAHRPWQMKDLQAIKQEVSSSALGSPQFMQTLRLA
VQQFDPTAKDLQDLLQYLCSSLVVSLHHQQLNTLITEAETRGMTGYNPMAGPLRMQANNP
AQQGLRREYQNLWLAAFSTLPGNTRDPSWAAILQGLEEPYCAFVERLNVALDNGLPEGTP
KEPILRSLAYSNANKECQKILQARGHTNSPLGEMLRTCQAWTPKDKTKVLVVQPRRPPPT
QPCFRCGKVGHWSRDCTQPRPPPGPCPLCQDPSHWKRDCPQLKPPQEEGEPLLLDLPSTS
GTTEEKNSLRGEI

23. HTLV5

gag polyprotein - AIDS virus HTLV-III (T-cell leukemia virus, BH10)

MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERFAVNPGLLETSEGCRQI
LGQLQPSLQTGSEELRSLYNTVATLYCVHQRIEIKDTKEALDKIEEEQNKSKKKAQQAAA
DTGHSSQVSQNYPIVQNIQGQMVHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGAT
PQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPVHAGPIAPGQMREPRGSDIAGTT
STLQEQIGWMTNNPPIPVGEIYKRWIILGLNKIVRMYSPTSILDIRQGPKEPFRDYVDRF
YKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGVGGPGHKA
RVLAEAMSQVTNTATIMMQRGNFRNQRKMVKCFNCGKEGHTARNCRAPRKKGCWKCGKEG
HQMKDCTERQANFLGKIWPSYKGRPGNFLQSRPEPTAPPFLQSRPEPTAPPEESFRSGVE
TTTPPQKQEPIDKELYPLTSLRSLFGNDPSSQ

24. HTLV6

pol polyprotein - T-cell leukemia virus I (HTLV-I, Caribbean isolate)

GKKAACNLANTGASCPWARTPPKAPRNQPVPFKPERLQALQHLVRKALEAGHIEPYTGPG
NNPVFPVKKANGTWRFIHDLRATNSLTIDLSSSSPGPPDLSSLPTTLAHLQTIDLKDAFF
QIPLPKQFQPYFAFTVPQQCNYGPGTRYAWRVLPQGFKNSPTLFEMQLAHILQPIRQAFP
QCTILQYMDDILLASPSHADLQLLSEATMASLISHGLPVSENKTQQTPGTIKFLGQIISP
NHLTYDAVPKVPIRSRWALPELQALLGEIQWVSKGTPTLRQPLHSLYCALQRHTDPRDQI
YLNPSQVQSLVQLRQALSQNCRSRLVQTLPLLGAIMLTLTGTTTVVFQSKQQWPLVWLHA
PLPHTSQCPWGQLLASAVLLLDKYTLQSYGLLCQTIHHNISTQTFNQFIQTSDHPSVPIL
LHHSHRFKNLGAQTGELWNTFLKTTAPLAPVKALMPVFTLSPVIINTAPCLFSDGSTSQA
AYILWDKHILSQRSFPLPPPHKSAQRAELLGLLHGLSSARSWRCLNIFLDSKYLYHYLRT
LALGTFQGRSSQAPFQALLPRLLSRKVVYLHHVRSHTNLPDPISRLNALTDALLITPVLQ
LSPADLHSFTHCGQTALTLQGATTTEASNILRSCHACRKNNPQHQMPQGHIRRGLLPNHI
WQGDITHFKYKNTLYRLHVWVDTFSGAISATQKRKETSSEAISSLLQAIAYLGKPSYINT
DNGPAYISQDFLNMCTSLAIRHTTHVPYNPTSSGLVERSNGILKTLLYKYFTDKPDLPMD
NALSIALWTINHLNVLTNCHKTRWQLHHSPRLQPIPETHSLSNKQTHWYYFKLPGLNSRQ
WKGPQEALQEAAGAALIPVSASSAQWIPWRLLKRAACPRPVGGPADPKEKDHQHHG

25. HTLV7

pol polyprotein - T-cell leukemia virus (HTLV-II)

HRSRPYGYTPDTRARAGKAPRHPDPRRQWANQHPVQTPPNPPTHILALPKVPRYPFLLPL
RHPQQMDHHWKGRPTTMPGASIPPRRPQPPPIAANSHSKHHRPRTPSPTSPSGPISFKPE
RLQALNDLVSKALEAGHIEPYSGPGNNPVFPVKKPNGKWRFIHDLRATNAITTTLTSPSP
GPPDLTSLPTALPHLQTIDLTDAFFQIPLPKQYQPYFAFTIPQPCNYGPGTRYAWTVLPQ
GFKNSPTLFEQQLAAVLNPMRKMFPTSTIVQYMDDILLASPTNEELQQLSQLTLQALTTH
GLPISQEKTQQTPGQIRFLGQVISPNHITYESTPTIPIKSQWTLTELQVILGEIQWVSKG
TPILRKHLQSLYSALHGYRDPRACITLTPQQLHALHAIQQALQHNCRGRLNPALPLLGLI
SLSTSGTTSVIFQPKQNWPLAWLHTPHPPTSLCPWGHLLACTILTLDKYTLQHYGQLCQS
FHHNMSKQALCDFLRNSPHPSVGILIHHMGRFHNLGSQPSGPWKTLLHLPTLLQEPRLLR
PIFTLSPVVLDTAPCLFSDGSPQKAAYVLWDQTILQQDITPLPSHETHSAQKGELLALIC
GLRAAKPWPSLNIFLDSKYLIKYLHSLAIGAFLGTSSHQTLQAALPPLLQGKTIYLHHVR
SHTNLPDPISTFNEYTDSLILAPLVPLTPQGLHGLTHCNQRALVSFGATPREAKSLVQTC
HTCQTINSQHHMPRGYIRRGLLPNHIWQGDVTHYKYKKYKYCLHVWVDTFSGAVSVSCKK
KETSCETISAVLQAISLLGKPLHINTDNGPAFLSQEFQEFCTSYRIKHSTHIPYNPTSSG
LVERTNGVIKNLLNKYLLDCPNLPLDNAIHKALWTLNQLNVMNPSGKTRWQIHHSPPLPP
IPEASTPPKPPPKWFYYKLPGLTNQRWKGPLQSLQEAAGAALLSIDGSPRWIPWRFLKKA
ACPRPDASELAEHAATDHQHHG

26. HTLV8

env polyprotein precursor - AIDS virus HTLV-III (T-cell leukemia virus, BH10)

TEKLWVTVYYGVPVWKEATTTLFCASDAKAYDTEVHNVWATHACVPTDPNPQEVVLVNVT
ENFNMWKNDMVEQMHEDIISLWDQSLKPCVKLTPLCVSLKCTDLKNDTNTNSSSGRMIME
KGEIKNCSFNISTSIRGKVQKEYAFFYKLDIIPIDNDTTSYTLTSCNTSVITQACPKVSF
EPIPIHYCAPAGFAILKCNNKTFNGTGPCTNVSTVQCTHGIRPVVSTQLLLNGSLAEEEV
VIRSANFTDNAKTIIVQLNQSVEINCTRPNNNTRKSIRIQRGPGRAFVTIGKIGNMRQAH
CNISRAKWNNTLKQIDSKLREQFGNNKTIIFKQSSGGDPEIVTHSFNCGGEFFYCNSTQL
FNSTWFNSTWSTKGSNNTEGSDTITLPCRIKQIINMWQEVGKAMYAPPISGQIRCSSNIT
GLLLTRDGGNSNNESEIFRPGGGDMRDNWRSELYKYKVVKIEPLGVAPTKAKRRVVQREK
RAVGIGALFLGFLGAAGSTMGAASMTLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTV
WGIKQLQARILAVERYLKDQQLLGIWGCSGKLICTTAVPWNASWSNKSLEQIWNNMTWME
WDREINNYTSLIHSLIEESQNQQEKNEQELLELDKWASLWNWFNITNWLWYIKLFIMIVG
GLVGLRIVFAVLSVVNRVRQGYSPLSFQTHLPIPRGPDRPEGIEEEGGERDRDRSIRLVN
GSLALIWDDLRSLCLFSYHRLRDLLLIVTRIVELLGRRGWEALKYWWNLLQYWSQELKNS
AVSLLNATAIAVAEGTDRVIEVVQGAYRAIRHIPRRIRQGLERILL

27. HTLV9

env polyprotein - T-cell leukemia virus I (HTLV-I, Caribbean isolate)
MGKFLATLILFFQFCPLILGDYSPSCCTLTVGVSSYHSKPCNPAQPVCSWTLDLLALSAD
QALQPPCPNLVSYSSYHATYSLYLFPHWIKKPNRNGGGYYSASYSDPCSLKCPYLGCQSW
TCPYTGAVSSPYWKFQQDVNFTQEVSHLNINLHFSKCGFSFSLLVDAPGYDPIWFLNTEP
SQLPPTAPPLLSHSNLDHILEPSIPWKSKLLTLVQLTLQSTNYTCIVCIDRASLSTWHVL
YSPNVSVPSPSSTPLLYPSLALPAPHLTLPFNWTHCFDPQIQAIVSSPCHNSLILPPFSL
SPVPTLGSRSRRAVPVAVWLVSALAMGAGVAGRITGSMSLASGKSLLHEVDKDISQLTQA
IVKNHKNLLKIAQYAAQNRRGLDLLFWEQGGLCKALQEQCCFLNITNSHVSILQERPPLE
NRVLTGWGLNWDLGLSQWAREALQTGITLVALLLLVILAGPCILRQLRHLPSRVRYPHYS
LINPESSL

28. PENJAN1

Penicillopepsin (EC 3.4.23.6) - Penicillium janthinellum
AASGVATNTPTANDEEYITPVTIGGTTLNLNFDTGSADLWVFSTELPASQQSGHSVYNPS
ATGKELSGYTWSISYGDGSSASGNVFTDSVTVGGVTAHGQAVQAAQQISAQFQQDTNNDG
LLGLAFSSINTVQPQSQTTFFDTVKSSLAQPLFAVALKHQQPGVYDFGFIDSSKYTGSLT
YTGVDNSQGFWSFNVDSYTAGSQSGDGFSGIADTGTTLLLLBDSVVSQYYSQVSGAQQDS
NAGGYVFDCSTNLPDFSVSISGYTATVPGSLINYGPSGDGSTCLGGIQSNSGIGFSIFGD
IFLKSQYVVFDSDGPQLGFAPQA

29. PEPASP1

Z1211227A: aspergillopepsin A
SKGSAVTTPQNNDEEYLTPVTVGKSTLHLDFDTGSADLWVFSDELPSSERTGHNVYTPSS
SATKLSGYTWNISYGNGSSASGDVYRDTVTVGGVTNTKEAVQAASKISSEFZZVBGGZBS
GAZAYSSINTVQPKAQTTFFDTVKSQLNSPLFAVQLKHDAPGVYDFGYIBBSKYTGSITY
TDADSSEGYWGFNPNGYSIGDSSSSGFSAIADTGTTLILLDDEIVLNGSZVSGQANQEAD
GGYVFBCSTTPPDFTGXIGDYKAVGPKYINYAPSBTPSTCFGGIQSNSGLGLSILGDVFL
KSQYVVFDSQGPKLGFAAQA

30. PEPEND1

Z1311269A: endothiapepsin
STGSATTTPIDSLDDAYITPVQIGTPAQTLNLDFDTGSSDLWVFSSETTASEVDGQTIYT
PSKSTTAKLLSGATWSISYGDGSSSSGDVYTDTVSVGGLTVTGQAVESAKKVSSSFTEDS
TIDGLLGLAFSTLNTVSPTQQKTFFDNAKASLDSPVFTADLGYHAPGTYNFGFIDTTAYT
GSITYTAVSTKQGFWEWTSTGYAVGSGTFKSTSIDGIADTGTTLLYLPATVVSAYWAQVS
GAKSSSSVGGYVFPCSATLPSFTFGVGSARIV DYIDFGPISTGSSSCFGGIQSSAGI
GINIFGDVALKAAFVVFNGATTPTLGFASK

31. PEPEND2

E70087:ACID PROTEINASE (ENDOTHIAPEPSIN)
AGTTTTTTIDSLDDAYIVVIQLGTPAATLNLNFDTGSADLWVFAGGSSNNSGHNVYNSAA
SSASATLSTGKWSIAYAAGASASGLNRDGTVTVGGVTAHGQAVQAAAAAFAAQATNDGLL
GLAFSSINTVQGSQNLFFVQGKSLAQTPLFSVYLAHQQGVYDFGLIDSSAYTGSLNYTGV
ASQGFWSFNVDSYTAGSGSTIGDVSSGIADTGTTVLYLPTSVVSAYYSQVSGATDSAAGG
LDITCDGLTAMPAGVGSTKASVPLSTAYGGQGNGSSGCLGGIQSAAGIGLLIFGDIFIKA
SYVVFDSNNPNLGFAPAA

32. PEPHEN1

Pepsinogen A (EC 3.4.23.1) - Chicken
TATESYEPMTNYMDASYYGTISIGTPQQDFSVIFDTGSSNLWVPSIYCKSSACSNHKRFD
PSKSSTYVSTNETVYIAYGTGSMSGILGYDTVAVSSIDVQNQIFGLSETEPGSFFYYCNF

DGILGLAFPSISSSGATPVFDNMMSQHLVAQDLFSVYLSKDGETGSFVLFGGIDPNYTTK
GIYWVPLSAETYWQITMDRVTVGNKYVACFFTCQAIVDTGTSLLVMPQGAYNRIIKDLGV
SSDGEISCDDISKLPDVTFHINGHAFTLPASAYVLNEDGSCMLGFENMGTPTELGEQWIL
GDVFIREYYVIFDRANNKVGLSPLS

33. PEPHIZ1

Z1402278A: hizopuspepsin I
AGVGTVPMTDYGNDIEYYGQVTIGTPGKKFNLDFDTGSSDLWIASTLCTNCGSRQTKYDP
NQSSTYQADGRTWSISYGDGSSASGILAKDNVNLGGLLIKGQTIELAKREAASFASGPND
GLLGLGFDTITTVRGVKTPMDNLISQGLISRPIFGVYLGKAKNGGGGEYIFGGYDSTKFK
GSLTTVPIDNSRGWWGITVDRATVGTSTVASSFDGILDTGTTLLILPNNIAASVARAYGA
YDNGDGTYTISCDTSRFKPLVFSINGASFQVSPDSLVFEEYQGQCIAGFGYGDFDFAIIG
DTFLKNNYVVFNQGVPEVQIAPVAE

34. PEPHM1

Pepsinogen A (EC 3.4.23.1) precursor - Human
VDEQPLENYLDMEYFGTIGIGTPAQDFTVVFDTGSSNLWVPSVYCSSLACTNHNRFNPED
SSTYQSTSETVSITYGTGSMTGILGYDTVQVGGISDTNQIFGLSETEPGSFLYYAPFDGI
LGLAYPSISSSGATPVFDNIWNQGLVSQDLFSVYLSADDQSGSVVIFGGIDSSYYTGSLN
WVPVTVEGYWQITVDSITMNGEAIACAEGCQAIVDTGTSLLTGPTSPIANIQSDIGASEN
SDGDMVVSCSAISSLPDIVFTINGVQYPVPPSAYILQSEGSCISGFQGMNLPTESGELWI
LGDVFIRQYFTVFDRANNQVGLAPVA

35. PEPMON1

Pepsinogen A (EC 3.4.23.1) precursor - Rhesus macaque
IDEQPLENYLDVEYFGTIGIGTPAQDFTVIFDTGSSNLWVPSVYCSSLACTNHNLFNPQD
SSTYQSTSGTLSITYGTGSMTGILGYDTVQVGGISDTNQIFGLSETEPGSFLYYAPFDGI
LGLAYPSISSSGATPVFDNIWDQGLVSQDLFSVYLSADDQSGSVVIFGGIDSSYYTGSLN
WVPVSVEGYWQISVDSITMNGEAIACAEGCQAIVDTGTSLLTGPTSPIANIQSDIGASEN
SDGEMVVSCSAISSLPDIVFTINGVQYPLPPSAYILQSQGSCTSGFQGMDVPTESGELWI
LGDVFIRQYFTVFDRANNQVGLAPVA

36. PEPMON3

Pepsinogen C (EC 3.4.23.3) - Japanese macaque
SVSYEPMAYMDAAYFGEISIGTPPQNFLVLFDTGSSNLWVPSVYCQSQACTSHSRFNPSE
SSTYSTNGQTFSLQYGSGSLTGFFGYDTLTVQSIQVPNQEFGLSENEPGTNFVYAQFDGI
MGLAYPTLSVDGATTAMQGMVQEGALTSPIFSVYLSDQQGSSGGAVVFGGVDSSLYTGQI
YWAPVTQELYWQIGIEEFLIGGQASGWCSEGCQAIVDTGTSLLTVPQQYMSALLQATGAQ
EDEYGQFLVNCNSIQNLPTLTFIINGVEFPLPPSSYILNNNGVCTVGVEPTYLSAQNSQP
LVYWILGDVFLRSYYSVYDLSNNRVGFATAA

37. PEPPIG2

E740190A: pepsin
IGDEPLENYLDTEYFGTIGIGTPAQDFTVIFDTGSSNLWVPSVYCSSLACSDHNQFNPDD
SSTFEATSQELSITYGTGSMTGILGYDTVQVGGISDTNQIFGLSETEPGSFLYYAPFDGI
LGLAYPSISASGATPVFDNLWDQGLVSQDLFSVYLSSNDDSGSVVLLGGIDSSYYTGSLN
WVPVSVEGYWQITLDSITMDGETIACSGGCQAIVDTGTSLLTCPTSAIANIQSDIGASEN
SDGEMVISCSSIDSLPDIVFTINGVQYPLSPSAYILQDDDSCTSGFEGMDVPTSSGELWI
LGDVFIRQYYTVFDRANNKVGLAPVA

38. PEPPROA

yeast proteinase a jbc 1987
GGHDVPLTNYLNAQYYTDITLGTPPQNFKVILDTGSSNLWVPSNECGSLACFLHSKYDHE
ASSSYKANGTEFAIQYGTGSLEGYISQDTLSIGDLTIPKQDFAEATSEPGLTFAFGKFDG
ILGLGYDTISVDKVVPPFYNAIQQDLLDEKRFAFYLGDTSKTENGGEATFGGIDESKFKG
DITWLPVRRKAYWEVKFEGIGLGDEYAELESHGAAIDTGTSLITLPSGLAEMINAEIGAK
KGWTGQYTLDCNTRDNLPDLIFNFNGYNFTIGPYDYTLEVSGSCISAITPMDFPEPVGPL
AIVGDAFLRKYYSIYDLGNNAVGLAKAILKNNYVVFNQGVPEVQIAPVAQ

39. PEPRAT3

PEPC$RAT: GASTRICSIN PRECURSOR (EC 3.4.23.3) (PEPSINOGEN C). 392 AA.
SVLYEPMAYMDASYFGEISIGTPPQNFLVLFDTGSSNLWVSSVYCQSEACTTHARFNPSK
SSTYYTEGQTFSLQYGTGSLTGFFGYDTLTVQSIQVPNQEFGLSENEPGTNFVYAQFDGI
MGLAYPGLSSGGATTALQGMLGEGALSQPLFGVYLGSQQGSNGGQIVFGGVDKNLYTGEI
TWVPVTQELYWQITIDDFLIGDQASGWCSSQGCQGIVDTGTSLLVMPAQYLSELLQTIGA
QEGEYGEYFVSCDSVSSLPTLSFVLNGVQFPLSPSSYIIQEDNFCMVGLESISLTSESGQ
PLWILGDVFLRSYYAIFDMGNNKVGLATSV

40. PEPRHZ2
E70086: ACID PROTEINASE EC 3.4.23.9
GVGTVPMTDYGNDVEYYGQVTIGTPGKSFNLNFDTGSSNLWVGSVQCQASGCKGGRDKFN
PSDGSTFKATGYDASIGYGDGSASGVLGYDTVQVGGIDVTGGPQIQLAQRLGGGGFPGDN
DGLLGLGFDTLSITPQSSTNAFDQVSAQGKVIQPVFVVYLAASNISDGDFTMPGWIDNKY
GGTLLNTNIDAGEGYWALNVTGATADSTYLGAIFQAILDTGTSLLILPDEAAVGNLVGFA
GAQDAALGGFVIACTSAGFKSIPWSIYSAIFEIITALGNAEDDSGCTSGIGASSLGEAIL
GDQFLKQQYVVFDRDNGIRLAPVA
41. RENHM1
RENI$HUMAN: RENIN PRECURSOR, RENAL (EC 3.4.23.15) (GENE NAME: REN). 406
AA.
LTLGNTTSSVILTNYMDTQYYGEIGIGTPPQTFKVVFDTGSSNVWVPSSKCSRLYTACVY
HKLFDASDSSSYKHNGTELTLRYSTGTVSGFLSQDIITVGGITVTQMFGEVTEMPALPFM
LAEFDGVVGMGFIEQAIGRVTPIFDNIISQGVLKEDVFSFYYNRDSENSQSLGGQIVLGG
SDPQHYEGNFHYINLIKTGVWQIQMKGVSVGSSTLLCEDGCLALVDTGASYISGSTSSIE
KLMEALGAKKRLFDYVVKCNEGPTLPDISFHLGGKEYTLTSADYVFQESYSSKKLCTLAI
HAMDIPPPTGPTWALGATFIRKFYTEFDRRNNRIGFALAR
42. RENMSE1
Z0809326A: renin
SSLTDLISPVVLTNYLNSQYYGEIGIGTPPQTFKVIFDTGSANLWVPSTKCSRLYLACGI
HSLYESSDSSSYMENGDDFTIHYGSGRVKGFLSQDSVTVGGITVTQTFGEVTELPLIPFM
LAQFDGVLGMGFPAQAVGGVTPVFDHILSQGVLKEKVFSVYYNRGPHLLGGEVVLGGSDP
EHYQGDFHYVSLSKTDSWQITMKGVSVGSSTLLCEEGCEVVVDTGSSFISAPTSSLKLIM
QALGAKEKRLHEYVVSCSQVPTLPDISFNLGGRAYTLSSTDYVLQYPNDKLCTVALHAMD
IPPPTGPVWVLGATFIRKFYTEFDRHNNRIGFALAR
43. RENMSE2
Gamma-renin (EC 3.4.23.15) precursor - Mouse
IVGGFKCEKNSQPWQVAVYYHKEHICGGVLLDRNWVLTAAHCYVDECEVWLGKNQLFQEE
PSAQNRLVSKSFPHPGFNMTLLTFEKLPPGADFSNDLMLLRLSKPADITDVVKPIDLPTK
EPKLDSTCLVSGWGSITPTKWQKPDDLQCMFTKLLPNENCAKAYLLKVTDVMLCTIEMGE
DKGPCVGDSGGPLICDGVLQGTVSIGPDPCGIPGVSAIYTNLVKFNSWIKDTMMKNA
44. RENMSE3
RENI$MOUSE: RENIN PRECURSOR, RENAL (EC 3.4.23.15) (GENE NAME: REN1). 402
AA.
PSLTNLTSPVVLTNYLNTQYYGEIGIGTPPQTFKVIFDTGSANLWVPSTKCSRLYLACGI
HSLYESSDSSSYMENGSDFTIHYGSGRVKGFLSQDSVTVGGITVTQTFGEVTELPLIPFM
LAKFDGVLGMGFPAQAVGGVTPVFDHILSQGVLKEEVFSVYYNRGSHLLGGEVVLGGSDP
QHYQGNFHYVSISKTDSWQITMKGVSVGSSTLLCEEGCAVVVDTGSSFISAPTSSLKLIM
QALGAKEKRIEEYVVNCSQVPTLPDISFDLGGRAYTLSSTDYVLQYPNRRDKLCTLALHA
MDIPPPTGPVWVLGATFIRKFYTEFDRHNNRIGFALAR
45. RENRAT1
Z1310378A: renin
KSSFTNVTSPVVLTNYLDTQYYGEIGIGTPSQTFKVIFDTGSANLWVPSTKCGPLYTACE
IHNLYDSSESSSYMENGTEFTIHYGSGKVKGFLSQDVVTVGGIIVTQTFGEVTELPLIPF
MLAKFDGVLGMGFPAQAVDGVIPVFDHILSHEVLKEEVFSVYYSRESHLLGGEVVLGGSD
PQHYQGNFHYVSISKAGSWQITMKGVSVGPATLLCEEGCMAVVDTGTSYISGPTSSLQLI
MQALGVKEKRANNYVVNCSQVPTLPDISFYLGGRTYTLSNMDYVQKNPFRNDDLCILALQ
GLDIPPPTGPVWVLGATFIRKFYTEFDRHNNRIGFALAR
46. RENRHZ1
Aspartic proteinase (EC 3.4.23.6) - Rhizomucor pusillus
AEGDGSVDTPGLYDFDLEEYAIPVSIGTPGQDFYLLFDTGSSDTWVPHKGCDNSEGCVGK
RFFDPSSSSTFKETDYNLNITYGTGGANGIYFRDSITVGGATVKQQTLAYVDNVSGPTAE
QSPDSELFLDGIFGAAYPDNTAMEAEYGDTYNTVHVNLYKQGLISSPVFSVYMNTNDGGG
QVVFGGVNNTLLGGDIQYTDVLKSRGGYFFWDAPVTGVKIDGSDAVSFDGAQAFTIDTGT
NFFIAPSSFAEKVVKAALPDATESQQGYTVPCSKYQDSKTTFSLVLQKSGSSSDTIDVSV
PISKMLLPVDKSGETCMFIVLPDGGNQFIVGNLFLRFFVNVYDFGKNRIGFAPLASGYEN
D
47. RENRHZ2
CARP$RHIMI: ASPARTATE PROTEASE PRECURSOR (EC 3.4.23.6) (MUCOR RENNIN).
430 AA.
AAADGSVDTPGYYDFDLEEYAIPVSIGTPGQDFLLLFDTGSSDTWVPHKGCTKSEGCVGS

RFFDPSASSTFKATNYNLNITYGTGGANGLYFEDSIAIGDITVTKQILAYVDNVRGPTAE
QSPNADIFLDGLFGAAYPDNTAMEAEYGSTYNTVHVNLYKQGLISSPLFSVYMNTNSGTG
EVVFGGVNNTLLGGDIAYTDVMSRYGGYYFWDAPVTGITVDGSAAVRFSRPQAFTIDTGT
NFFIMPSSAASKIVKAALPDATETQQGWVVPCASYQNSKSTISIVMQKSGSSSDTIEISV
PVSKMLLPVDQSNETCMFIILPDGGNQYIVGNLFLRFFVNVYDFGNNRIGFAPLASAYEN
E

TRYPSIN/CHYMOTRYPSIN (15 sequences)
Full description is listed previously

1. TRPCRAY1
2. TRPDOG1
3. TRPDOG2
4. TRPFISH1
5. TRPFISH2
6. TRPFLY1
7. TRPHM1
8. TRPHOR1
9. TRPHOR2
10. TRPMSE1
11. TRPPIG1
12. TRPRAT1
13. TRPRAT2
14. TRPRAT3
15. TRPSTRP1

AIDS RELATED PROTEINS (25 sequences)
Full description is listed previously

1. AID1
2. AID10
3. AID11
4. AID12
5. AID13
6. AID2
7. AID3
8. AID4
9. AID5
10. AID6
11. AID7
12. AID8
13. AID9
14. HTLV1
15. HTLV10
16. HTLV11
17. HTLV12
18. HTLV2
19. HTLV3
20. HTLV4
21. HTLV5
22. HTLV6
23. HTLV7
24. HTLV8
25. HTLV9

PROTEASE INHIBITORS (26 sequences)

1. ANTIASP1
Proteinase inhibitor - Eggplant
QICTNNCAGRKGCSYFSEDGTFICKGESNPENPKACPRNCDGRIAYGICPLS
2. ANTIASP2
Proteinase inhibitor PTI - Potato
RICTNCCAGYKGCNYYSANGAFICEGESDPKNPNVCPRNCDGRIAYSKCLR
3. ANTIASP3
Proteinase inhibitor PCI-I - Potato
PICTNCCAGYKGCNYYSANGAFICEGQSDPKKPKACPLNCDPRIAYSKCLRS
4. ANTIASP4
Proteinase A inhibitor 3 - Yeast (Saccharomyces cerevisiae)
MNTDQQKVSEIFQSSKEKLQGDAKVVSDAFKKMASQDKDGKTTDADLSEKHNYQEQYNKL
KGAGHKKE
5. ANTIMET1
Metalloproteinase inhibitor precursor - Human
CTCVPPHPQTAFCNSDLVIRAKFVGTPEVNQTTLYQRYEIKMTKMYKGFQALGDAADIRF
VYTPAMESVCGYFHRSHNRSEEFLIAGKLQDGLLHITTCSFVAPWNSLSLAQRRGFTKTY
TVGCEECTVFPCLSIPCKLQSGTHCLWTDQLLQGSEKGFQSRHLACLPREPGLCTWQSLR
SQIA
6. ANTIMET2
Metalloproteinase inhibitor - Streptomyces nigrescens
APSCPAGSLCTYSGTGLSGARTVIPASDMEKAGTDGVKLPASARSFANGTHFTLRYGPAR
KVTCVRFPCYQYATVGKVAPGAQLRSLPSPGATVTVGQDLGD
7. ANTIP10
Venom basic protease inhibitor I - Western sand viper
QDHPKFCYLPADPGRCKAHIPRFYYDSASNKCNKFIYGGCPGNANNFKTWDECRQTCGAS
A
8. ANTIP11
Venom basic protease inhibitor III - Sand viper
RDRPKFCYLPADPGRCLAYMPRFYYNPASNKCEKFIYGGCRGNANNFKTWDECRHTCVAS
GIQPR
9. ANTIP12
Basic protease inhibitor - Red sea turtle
QGDKRDICRLPPEQGPCKGRIPRYFYNPASRMCESFIYGGCKGNKNNFKTKAECVRACRP
PERPGVCPKTSGPGICLHGCDSDSDCKEGQKCCFDGCGYICLTVAPSGSP
10. ANTIP13
Protease inhibitor (PSTI type), submandibular - Dog
GPPPAIGRGVBCSBYKGKGSEIACPRLHZPICGTDHKTYSNZCMLCAFTLDKKFZVRKLQ
DTACDIECTEYSDMCTMDYRPLCGSDGKNYSNKCSFCNAVKKSRGTIFLAKHGEC
11. ANTIP14
Protease B inhibitors 2 and 1 - Yeast (Saccharomyces cerevisiae)
TKNFIVTLKKNTPDVEAKKFLDSVHHAGGSILHEFDIIKGYTIKVPDVLHLNKLKEKHND
VIENVEEDKEVHTN
12. ANTIPRO1
Serum basic protease inhibitor - Bovine
TERPDFCLEPPYTGPCKAAMIRYFYNAKAGFCETFVYGGCRAKSNNFKSAEDCMRTCGGA
13. ANTIPRO2
Venom basic protease inhibitor I - Black mamba
QPLRKLCILHRNPGRCYQKIPAFYYNQKKKQCZGFTWSGCGGNSNRFKTIEECRRTCIRK
14. ANTIPRO3
Venom basic protease inhibitor I homolog - Eastern green mamba
QPRRKLCILHRNPGRCYDKIPAFYYNQKKKQCERFDWSGCGGNSNRFKTIEECRRTCIG
15. ANTIPRO4
Venom basic protease inhibitor K - Black mamba and eastern green mamba
AAKYCKLPLRIGPCKRKIPSFYYKWKAKQCLPFDYSGCGGNANRFKTIEECRRTCVG
16. ANTIPRO5
Venom basic protease inhibitor B - Black mamba
RPYACELIVAAGPCMFFISAFYYSKGANKCYPFTYSGCRGNANRFKTIEECRRTCVV

17. ANTIPRO6
Venom basic protease inhibitor E - Black mamba
LQHRTFCKLPAEPGPCKASIPAFYYNWAAKKCQLFHYGGCKGNANRFSTIEKCRHACVG
18. ANTIPRO7
Venom basic protease inhibitor II - Ringhals
RPDFCELPAETGLCKAYIRSFHYNLAAQQCLQFIYGGCGGNANRFKTIDECRRTCVG
19. ANTIPRO8
Venom basic protease inhibitor II - Cape cobra
RPRFCELPAETGLCKARIRSFHYNRAAQQCLEFIYGGCGGNANRFKTIDECHRTCVG
20. ANTIPRO9
Venom basic protease inhibitor II - Russell's viper
HDRPTFCNLAPESGRCRGHLRRIYYNLESNKCKVFFYGGCGGNANNFETRDECRETCGGK
21. ANTISER1
Antileukoproteinase 1 - Human
SGKSFKAGVCPPKKSAQCLRYKKPECQSDWQCPGKKRCCPDTCGIKCLDPVDTPNPTRRK
PGKCPVTYGQCLMLNPPNFCEMDGQCKRDLKCCMGMCGKSCVSPVKA
22. ANTISER2
Acrosin inhibitor I (PSTI type) - Bovine
EIYFEPDFGFPPDCKVYTEACTREYNPICDSAAKTYSNECTFCNEKMNNDADIEFNHFGE
CEY
23. ANTISER3
Serine proteinase inhibitor 1 - Vaccinia virus
MDIFKELILKHTDENVLISPVSILSTLSILNHGAAGSTAEQLSKYIENMNENTPDDNNDM
DVDIPYCATLATANKIYGSDSIEFHASFLQKIKDDFQTVNFNNANQTKELINEWVKTMTN
GKINSLLTSPLSINTRMTVVSAVHFKAMWKYPFSKHLTYTDKFYISKNIVTSVDMMVSTE
NNLQYVHINELFGGFSIIDIPYEGNSSMVIILPDDIEGIYNIEKNITDEKFKKWCGMLST
KSIDLYMPKFKVEMTEPYNLVPILENLGITNIFGYYADFSKMCNETIYVEKFLHTTFIDV
NEEYTEASAVTGVFMTNFSMVYRTKVYINHPFMYMIKDNTGRILFIGKYCYPQ
24. ANTISER4
Ovomucoid (PSTI-type protease inhibitor) 1 - Japanese quail
VEVDCSRFPNTTNEEGKDEVVCPDELRLICGTDGVTYNHECMLCFYNKEYGTNISKEQDG
ECGETVPMDCSRYPNTTSEDGKVTILCTKDFSFVCGTDGVTYDNECMLCAHNVVQGTSVG
KKHDGECRKELAAVSVDCSEYPKPACPKDYRPVCGSDNKTYSNKCNFCNAVVESNGTLTL
NHFGKC
25. ANTISER7
Ovomucoid (PSTI-type protease inhibitor) precursor - Chicken
MAMAGVFVLFSFVLCGFLPDTAFGAEVDCSRFPNATDKEGKDVLVCNKDLRPICGTDGVT
YTNDCLLCAYSIEFGTNISKEHDGECKETVPMNCSSYANTTSEDGKVMVLCNRAFNPVCG
TDGVTYDNECLLCAHKVEQGASVDKRHDGGCRKELAAVSVDCSEYPKPDCTAEDRPLCGS
DNKTYGNKCNFCNAVVESNGTLTLSHFGKC
26. ANTISER8
Plasminostreptin (PSTI-type protease inhibitor) - Streptomyces sp.
GLYAPSALVLTMGHGNSAATVNPERAVTLNCAPTASGTHPAALQACAELRGAGGDFDALT
VRGDVACTKQFDPVVVTVDGVWQGKRVSYERTFANECVKNSYGMTVFTF

REFERENCES

1. Dockerson, R.E. & Geis, I., (1969) The structure and action of proteins, Harper & Row Publishers, New York.

2. Bishop, M. & Rawlings, C., (1987) Nucleic acid and protein sequence analysis - a practical approach, IRL Press, ed.

3. Darby, N. & Creighton, T., (1993) Protein structure, IRL Press.

4. Nall, B. & Dill, K., (1991) Conformations and forces in protein folding, American Association for the Advancement of Science.

5. Cosic, I., (1994) Macromolecular bioactivity: is it resonant interaction between macromolecules? - theory and application, IEEE Trans. on Biomedical Engineering, 41, 1101-1114.

6. Cosic, I., (1997) The Resonant Recognition Model of macromolecular bioactivity, Birkhouser.

7. Kane, K. (ed.) (1993) Biochemistry, Wm. C. Brown Publishers, IA.

8. Schulz, G.E., (1979) Principles of protein structure, New York, Springer-Verlag.

9. Mackay, D.H.J., Cross, A.J. & Hagler, A.T. (1989) Prediction of Protein Structure and Principles of Protein Conformation, G.D. Fasman, Ed.New York, 317-358.

10. Fasman, G.D., (1989) Prediction of protein structure and the principles of protein conformation, Plenum Press.

11. King, J. (1989) Deciphering the rules of protein folding, Chem. and Eng. News, Vol.32.

12. Chan, H.S. & Dill, K.A., (1993) The protein folding problem, Physics Today, 46, 24-32.

13. Ebbing, D.D. (1987) General Chemistry. Houghton Miffin Company, Boston.

14. King, R.D. & Sternberg, M.J., (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction, Potein Science, 5, 2296-2310.

15. Bowie, J.U., Luthy, R. & Eisenberg, D., (1991) A method to identify protein sequences that fold into a known three-dimensional structure, Science, 235, 164-170.

16. Ehrensten, G., Lecar, H. (1982) Biophysics, Academic Press.

17. Lehninger, A.L., Nelson, D.L., Cox, M.M. (1993) Principles of Biochemistry. Worth Publisher, New York.

18. Anfinsen, C.B., (1973) Principles that govern the folding of protein chains, Science, 181, 223-230.

19. Sali, A., Shakhnovich, E. & Karplus, M., (1994) How does a protein fold?, Nature, 369, 248-251.

20. Gregoret, L.M. & Cohen, F.E., (1991) Protein folding:effect of packing density on chain conformation, Journal of Molecular Biology, 219, 109-122.

21. Sun, S., (1993) Reduced representation of protein structure prediction: Statistical potential and genetic algirithm, Protein Science, 2, 762-785.

22. Gunn, J.R., Monge, A., Friesner, R.A., & marshall, C.H., (1994) Hierarchical algorithm for computer modelling of protein tertiary structure: Folding of Myoglobin to 6.2 A Resolution, Journal of Physical Chemistry, 98, 702-711.

23. Brooks, C.L.III & Case, D.A., (1993) Simulatin of peptide conformational dynamics and thermodynamics, Chemical Reviews, 93, 2487-2502.

24. Wilson, S.R. & Cui, W., (1990) Applications of simulated annealing to peptides, Biopolymers, 29, 225-235.

25. Unger, R. & Moult, J., (1993) Genetic algorithm for protein folding simulations, Journal of Molecular Biology, 231, 75-81.

26. Laughton, C.A., (1994) A study of simulated anneling protocols for use with molecular dynamics in protein structure prediction, Protein Engineering, 7, 235-241.

27. Rose, G.D. (1978) Prediction of chain turns in globular proteins on a hydrophobic basis. Nature, 272: 586-590.

28. Kyte, J. & Doolitle, R.F., (1982) A simple method for displaing the hyropathic character of a protein, Journal of Molecular Biology, 157(6):105-142.

29. Eisenberg, D. (1984) Three-dimensional structure of membrane and surface proteins. [Review]. Annual Review of Biochemistry, 53:595-623.

30. Lesk, A.M. ed. (1988) Computational molecular biology. Oxford University Press

31. White, J.V., Stultz, C.M., Smith, T.F. (1994) Protein Classification by Stochastic Modeling and Optimal Filtering of Amino Acid Sequence. Mathematical Biosciences, 119:35-75.

32. White J.V. (1988) Modeling and filtering for discretely valued time series, in Spall J.C. ed., Bayesian analysis of time series and dynamics models, Dekker, New York, Chapter 10.

33. Doolitle, R. (1981) Similar amino acid sequences: Chance or common ancestry?, Science, Vol.214, 149-159.

34. Devereux, J, Haeberly, P. & Smithies, O. (1984) A comprehencive set of sequence analysis programs for the VAX, Nucl.Acids Res., Vol. 12, 387-395.

35. Breg, J.N., van Opheusden, J.H.J., Burgering, M.J.M. & Captein, R. (1990) Structure of arc repressor in solution: Evidence for a family of β-sheet DNA-binding proteins, Nature, Vol.346, 586-589.

36. Needleman, S. & Wunssch, C.D., (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol., 48:443.

37. Dayhoff, M.O. & Eck, R.V., (1970) MASSPEC: a computer program for complete sequence analysis of large proteins from mass spectrometry data of a single sample, Computers in Biology & Medicine, 1(1):5-28.

38. Mohana Rao, J.K., (1987) New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters, International Journal of Peptide & Protein Research, 29(2): 276-281.

39. Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978) A model for evolutionary change, Atlas of protein sequence and structure, Vol. 5, Suppl. 3, National Biomedical Research Foundation, Washington DC, 345-358.

40. Taylor, W.R. & Orengo, C.A., (1989) A holistic approach to protein structure alignment, Protein Engineering, 2(7):505:19.

41. Feng, D.F. & Doolitle, R.F., (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic treec, Journal of Molecular Evolution, 25(4):

42. Barton, G.J. & Sternberg, M.J., (1987) A strategy for the rapid multiple alignment of protein sequences: Confidence levels from tertiary structure comparisons, Journal of Molecular Biology, 198(2):327-37.

43. Branden, C. & Tooze, J. (1992) Introduction to proteinstructure.

44. Chou, P.Y. & Fasman, G.D. (1978) Annu. Review in Biochemistry, 47, 251-275.

45. Nischikawa, K. (1983) Biochim. Biophys. Acta, 748, 285-329.

46. Kabsch, W. & Sander, C. (1983) FEBS Lett., 155, 179-182.

47. Schultz, G.E. (1988) Annu.Rev.Biopfys.

48. Kanehisa, M. (1988) A multivariate analysis method for discriminating protein secondary structure segments, J. Prot.Eng, Vol.2, 87-92.

49. Nakai, K., Kidera, A. & Kanehisa, M. (1988) Cluster analysis of amino acid indices for prediction of protein structure and function, J. Protein Engineering, 2, 93-100.

50. Schmid, E.D., Schneider, F.W. & Siebert, F. (1988) Spectroscopy of biological molecules New advances, New York: Wiley.

51. Karplus, P.A. & Schulz, G.E. (1985) Naturwissenschaten, 72, 212-213.

52. Goldberg, D.E., (1989) Genetic algorithms in search, optimisation and machine learning, Addison Wesley, New York.

53. Kawashima, S., Kanehisa, M. (2000) Aaindex: amino acid index database. Nucleic Acids Res. 28,374.

54. Veljkovic, V. & Slavic, I. (1972) General model of pseudopotentials, Physica Review, 29, 105-108.

55. Lasovic, J. (1996) CABIOS, Vol.12, No 6, 553-562.

56. Birch, S., West, R. & Cosic, I. (1995) Preliminary expansion of the RRM to incorporate Multi Variable Analysis, J. Austr. Physic. and Eng. Sciences in Medicine.

57. Cosic, I., Vojisavljevic, V. & Pavlovic, M., (1989) The relationship of the resonant recognition model to effects of low-intensity light on cell growth, Int. J. Radiat. Biol., 56, 179-191.

58. Cosic, I., Hodder, A.N., Aguilar, M.I. & Hearn, M.T.W., (1991) Resonant Recognition model and protein topography: Model studies with myoglobin, hemoglobin and lysozyme, Eur. J. Biochem., 198, 113-119.

59. Cosic, I., Nesic, D., Pavlovic, M. & Williams, R., (1986) Enhancer binding proteins predicted by informational spectrum method, Biophys. Biochem. Comm., 141, 831-839.

60. Cosic, I. & Fang, Q., (1995) Problems in using FFT versus DFT in the resonant recognition model, IEEE EMBS, 17.

61. Veljkovic, I. (1980) A theoretical approach to preselection of cancerogens and chemical carcinogenesis, Gordon and Breach, New York.

62. Charton, B.I. (1981) Electr. Effect Sustituent Constants for Correlation Analysis, Prog. Phys. Organic Chemis., 13, 119-251.

63. Charton, B.I. (1982) The structural dependence of amino acid hydrophobicity parameters, J. Theoret. Biology, 99, 629-644.

64. Mandell, A. (1988) Chaos and Order in the Natural Science.

65. Viari, A., Soldano, H. & Olliveier, E. (1990) A scale-independent signal processing method for sequence analysis, Comput.Appl.Bioscien., Vol.6, 71-80.

66. Tanaka, S. & Scheraga, H.A. (1974) Macromolecules, 10, 9-20.

67. Cosic, I., Pavlovoc, M. & Vojisavljevic, V., (1989) Prediction of "hot spots" in interleukin-2 based on informational spectrum characteristics of growth regulating factors, Biochimie, 71, 333-342.

68. Cosic, I. & Nesic, D., (1988) Prediction of "hot spots" in SV40 enhancer and relation with experimental data, Eur. Journal Biochemistry, 170, 247-252.

69. Cosic, I. & Hearn, M.T.W., (1991) "Hot spot" amino acid distribution in Ha-ras Oncogene Product p21: Relationship to Guanine Binding Site, J. Molecular Recognition, 4, 57-62.

70. Veljkovic, V., Cosic, I., Dimitrijevic, B. & Pavlovic, D. (1985) IEEE Trans. Biomed. Eng., 32, 337-341.

71. Fang, Q. (1999) Protein sequence analysis: Anapproach based on digital signal processing and the RRM. PhD Thesis.

72. IUPAC Dissociation constants of organic acids in aqueous solutions.

73. IUPAC Ionisation constants of organic acids in aqueous solutions.

74. Oyster, C.K., Hanten, W.P. & Llorens, L.A. (1987) Introduction to research : A guide for the health science professional, Lippincott, Philadelphia PA, 170-176.

75. Bone, S. & Zaba, B. (1992) Bioelectronics, John Wiley & Sons, New York.

76. Szent-Gyorgyi, A. (1968) Bioelectronics. A study in cellular regulations, defense and cancer.

77. Smyth, C.P. (1955) Dielectric behavior and structure. Dielectric constant and loss, dipole moment and molecular structure.

78. Heine, V., Koen, M. & Uir, D., (1973) Theory of pseudopotentials, MIR, Moscow.

79. Simon, G.P. (1994) Dielectric relaxation Spectroscopy of thermoplastic polymers and blends. Materials Forum, 18, 235-264.

80. Simon, G. P., Beatty, C. L., Andrews, S.R., Shinton, S., Williams, G. (1998) Dielectric relaxation spectroscopy of a styrene/n-butyl methacrylate copolymer series. PolymerVol. 39, No.24, 5929-5937.

81. Mason, T. G. (1990) Sonochemistry: the uses of ultrasound in chemistry. Cambridge: Royal Society of Chemistry.