eISBN: 978-1-68108-269-1
ISBN: 978-1-68108-270-7

POLARITY INDEX IN PROTEINS-A BIOINFORMATICS TOOL

Carlos Polanco

Bentham e Books

Polarity Index in Proteins– A Bioinformatics Tool

Authored By:

Carlos Polanco

Faculty of Sciences
Universidad Nacional Autónoma de México
México

Polarity Index in Proteins - A Bioinformatics Tool

Author: Carlos Polanco

eISBN (Online): 978-1-68108-269-1

ISBN (Print): 978-1-68108-270-7

First Published in 2016

BENTHAM SCIENCE PUBLISHERS LTD.

End User License Agreement (for non-institutional, personal use)

This is an agreement between you and Bentham Science Publishers Ltd. Please read this License Agreement carefully before using the ebook/echapter/ejournal (**"Work"**). Your use of the Work constitutes your agreement to the terms and conditions set forth in this License Agreement. If you do not agree to these terms and conditions then you should not use the Work.

Bentham Science Publishers agrees to grant you a non-exclusive, non-transferable limited license to use the Work subject to and in accordance with the following terms and conditions. This License Agreement is for non-library, personal use only. For a library / institutional / multi user license in respect of the Work, please contact: permission@benthamscience.org.

Usage Rules:

1. All rights reserved: The Work is the subject of copyright and Bentham Science Publishers either owns the Work (and the copyright in it) or is licensed to distribute the Work. You shall not copy, reproduce, modify, remove, delete, augment, add to, publish, transmit, sell, resell, create derivative works from, or in any way exploit the Work or make the Work available for others to do any of the same, in any form or by any means, in whole or in part, in each case without the prior written permission of Bentham Science Publishers, unless stated otherwise in this License Agreement.
2. You may download a copy of the Work on one occasion to one personal computer (including tablet, laptop, desktop, or other such devices). You may make one back-up copy of the Work to avoid losing it. The following DRM (Digital Rights Management) policy may also be applicable to the Work at Bentham Science Publishers' election, acting in its sole discretion:

- 25 'copy' commands can be executed every 7 days in respect of the Work. The text selected for copying cannot extend to more than a single page. Each time a text 'copy' command is executed, irrespective of whether the text selection is made from within one page or from separate pages, it will be considered as a separate / individual 'copy' command.
- 25 pages only from the Work can be printed every 7 days.

3. The unauthorised use or distribution of copyrighted or other proprietary content is illegal and could subject you to liability for substantial money damages. You will be liable for any damage resulting from your misuse of the Work or any violation of this License Agreement, including any infringement by you of copyrights or proprietary rights.

Disclaimer:

Bentham Science Publishers does not guarantee that the information in the Work is error-free, or warrant that it will meet your requirements or that access to the Work will be uninterrupted or error-free. The Work is provided "as is" without warranty of any kind, either express or implied or statutory, including, without limitation, implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the results and performance of the Work is assumed by you. No responsibility is assumed by Bentham Science Publishers, its staff, editors and/or authors for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products instruction,

Limitation of Liability:

General:

Bentham Science Publishers Ltd.
Executive Suite Y - 2
PO Box 7917, Saif Zone
Sharjah, U.A.E.
Email: subscriptions@benthamscience.org

BENTHAM
SCIENCE

CONTENTS

Knowing is not enough; we must apply.
Willing is not enough; we must do.

-Goethe

FOREWORD

Nowadays technical dimensions in Bioinformatics are of ever increasing importance in the solution of environmental and biological problems, as they provide unprecedented tools for medical doctors and scientist that will help them in the advancement of disease diagnosis and drug development. In this book, Dr. Carlos Polanco overcomes the usual gap between algorithm developers and application designers and application users, describing and illustrating with examples a practical mathematical-computational algorithm named Polarity Index Method whose metric evaluates thoroughly the polar profile of a peptide or protein and predicts the main function associated to them with a high level of efficiency. This book can be used as an introduction in Computational Proteomics for those interested in biological algorithms, as well as a practical Bionformatics tool for the seasoned Researcher.

Jorge Alberto Castañón González
Department of Critical Care Medicine
Department of Biomedical Research
Hospital Juárez de México
México

PREFACE

Polarity is a physico-chemical property that characterizes the electromagnetic stability of a protein and can predict its plausible pathogenic action. For this reason, it is not surprising to find polarity as a major actor in most mathematical-computational algorithms that seek to characterize peptides and proteins. In this work I summarize the seven-year research oriented towards the study of this electromagnetic property. The reader will find in first instance, a classification of the algorithms known for this purpose, as inclusive as possible, and a description of an algorithm designed by us called polarity index method, expressing as a metric, the peptide polarity from all possible polar interactions that can occur when reading its linear sequence. You will also find the method makes it possible to reproduce the main classification of peptide proteins found in different databases, with a high degree of discriminative efficiency. Addition- ally I present the improvements the method has undergone in the recent years, and the knowledge, acquired in the process which allowed us to expand the its discriminating ability, and at the same time ameliorate its computational design. As a result of these improvements, the reader will find that this method is oriented to the identification of the possible selectivity some peptides have towards specific membranes. This group of peptides is now considered basic for the design of new pharmaceutical drugs. We have studied two peptide groups identified by Polarity index method: cell penetrating peptides, and natively unfolded proteins. The first group is closely related to the toxicity of a peptide, from the structural point of view, and it correlates with its ability to permeate the pathogenic membrane. This structural feature is also identified by the method that selects it from different protein groups, finding unknown features in the groups studied by non-experimental methods. The last group, natively unfolded proteins keeps a close relationship with a group of neurodegenerative diseases that are classified under the term Amyloidosis. The reader will find that just as the method identifies each of these groups it also differentiates their counterpart, the natively folded protein group, which includes neurons. We believe the results achieved with this method, that only measures the peptide polarity, can help the reader to improve predictive algorithms and to observe, from another perspective, how the electromagnetic balance of the protein provides enough information about the function of the protein itself. There is also a section oriented to the computational and mathematical aspect of the method, particularly for its computational implementation in personal computers and supercomputers. We consider this section very important because the method will be used for the manufacture of peptides or proteins, therefore the user will find it very useful. The mathematical aspect of the method was carefully developed in order to show the reader the importance of identifying certain regularities in the peptide polarity profile called catastrophic bifurcations points. We conclude with the results of our research about the possible proteins that should have been presented 4 billion years ago. The reader will find that when I computationally presented the experiments of Stanley Miller & Harold Clayton Urey, Sidney Walter Fox & Kaoru Harada, and Bernd Michael Rode, they were oriented to produce a considerable amount of prebiotic proteins, from the assumption of each experiment, I evaluated each set of proteins produced by polarity index method finding that there was a similar pattern in the four models, which in addition is coincident with the profile of the proteins known today, and when assessing the restrictions of each model, I came across that the abundance was a decisive factor in the profile of the proteins known today. The author hope that the reader interested in Proteomics and Bioinformatics will find to the material presented here useful, and those who start studying this field, will find this information motivating. It is a pleasure to thank Concepcio´n Celis Jua´rez whose suggestions and proof-reading have greatly improved the original manuscript, and also I acknowledge the Computer Science department at Institute for Nuclear Sciences at the Universidad Nacional Autonoma de México for support.

CONFLICT OF INTEREST

The author declared no conflict of interest regarding the contents of each of the chapters of this book.

Carlos Polanco
Faculty of Sciences
Universidad Nacional Autónoma de México
México

ACKNOWLEDGEMENTS

Special acknowledgment to Concepción Celis Juárez whose suggestions and proof- reading have greatly improved the original manuscript, Helmuth Eckerle by the design of the figures, and also I acknowledge the Computer Science department at Institute for Nuclear Sciences at the Universidad Nacional Autónoma de México for support.

Acronyms

3–D space	=	Three dimensional real space.
A°	=	1 angstrom = 1.0 × 10^{-10} meters.
cal/mol	=	calories per mole.
mM	=	1 millimolar = 1.0 mol × m^3.

Part I-PRELIMINARIES

This unit introduces the reader to the basic terminology used in the macromolecular structures throughout the book. For this reason, it has been ensured that the section is self-contained, so that, the unfamiliar reader can follow up the subject. This unit is divided into two sections: Macromolecules and Electromagnetic stability. The first section aims to characterize macromolecules morphologically, and the second section describes the various electromagnetic forces acting on a subgroup of macromolecular polymers called proteins.

CHAPTER 1

Macromolecules

Abstract: This chapter describes the structure of a group of polymers responsible for regulating the basic functions and inheritance of living organisms: the proteins. The characterization of these organic units involves a very complex problem due to its multi-factorial nature. To develop this topic I showed the difficulty when trying to differentiate living matter from non-living matter and then I proceeded to find distinctions at cellular level. At this stage I identified two main groups with three types of major macromolecular structures, where proteins are included, and four levels of structural complexity known as primary, secondary, tertiary and quaternary structures, from which proteins get their morphologic features.

Keywords: α-amino acids, Amide group, Eukaryotes, Carboxyl group, Covalent bond, Gene, Ionic bond, Macromolecules, Monomers, Non covalent bond, Nucleic acids, Peptide bond, Polymers, Polymerization, Polysaccharides, Prokaryotes, Proteins, Van der Waals' forces.

1.1 . THE LIVING AND NON-LIVING ENTITIES

In this chapter, I focus on the Structural characterization of the proteins as the main group of macromolecules regulating the functions and inheritance in living organisms. Although it is not conclusive that these functional units are exclusive of living organisms, the fact to this day is that it has not been found an effective discriminant showing a clear division between the living and non-living organisms. Let us consider as an approach to this reasoning two species that at first glance are totally incompatible: a human being and a crystal. To the naked eye there are clear differences between both groups and I could immediately suppose their chemical composition is completely different, however when comparing the constituents of both organisms I find they are composed of the

Carlos Polanco

same elements. 97.9% of all living organisms are made of only six chemical elements: carbon (C), hydrogen (H), oxygen (O), nitrogen (N), phosphorus (P) and sulfur (S); 2% are made of calcium (Ca), sodium (Na), potassium (K), magnesium (Mg) and chlorine (Cl); and 0.1% are made of small quantities of manganese (Mn), iron (Fe), cobalt (Co), copper (Cu), zinc (Zn), boron (B), aluminum (Al), vanadium (V), molybdenum (Mo), iodine (I) and silicon (Si), amongst others. Living matter also shares the same elements, so in our quest to find the difference between living and non-living organisms I could consider very specific functions, such as the assimilation of nutrients and reproduction, as proper of living organisms, however this is not so as these functions are also shared by the non-living organisms. I can mention three examples [1] of non-living- living world: salt crystals increase their mass when they are dissolved in a substance of the same nature; the chemical reaction of dyes in the same solution that penetrate and adhere to crystals; and the reconstruction of crystals broken facets and angles when placed in a saturated solution of the same substance. Therefore the functions, I think are exclusive of living organisms are not. Although these cases show the inherent complexity between the living and non-living organisms, at a cellular level it is possible to characterize two major groups: prokaryotes & eukaryotes, the difference lies in their structure. Prokaryotic organisms (from 3,500 million years) only have an external membrane while eukaryotic organisms (from 3,500 million years) apart from having a cell with an external membrane they have internal biological subunits, all of them protected by membranes showing a more complex structure. Both groups share basically 33 organic molecules forming three major macromolecular groups in the biological process: nucleic acids, polysaccharides and proteins. These macromolecular groups are responsible for featuring, regulating functions and determining inheritance in the organisms. Essentially they are polymers that are formed by a chemical process called Polymerization and its constituent elements are known as monomers. Each group has different variants in terms of the number and type of their monomers. In nucleic acids there are four monomers called nucleotides; the polysaccharides consist of six monomers known as hexoses, and proteins have twenty α–amino acids. The polymerization process that generates a polymer is influenced by the electromagnetic forces (covalent bond, and non covalent bonds), between monomers until the monomer reaches its electromagnetic balance:

Covalent bond This is a strong bond forming the structure of the polymer.

Non covalent bond Here are mainly three different types of intermolecular forces.

Ionic bond It is the result of the attraction and repulsion between distinctly separate electrical charges

Hydrogen bond It is the result of the bond between atoms of hydrogen and nitrogen.

Van der Walls These forces are the result of the attractive or repulsive forces between close molecules, these are weak bonds responsible for the final polymer spatial conformation. See Chapter 2 for more details in the description of covalent and non covalent bonds.

1.2. PROTEINS

In Macromolecules, the group of proteins is by itself of vital importance for living organisms, as proteins participate in all cell functions having each a specific task. The chemical process called Polymerization only considers 20 α–amino acids (see Table **1.1**) from all the amino acids known, this process makes the protein adopt a different spatial conformation or arrangement denominated primary, secondary, tertiary and quaternary structure. Each Polymer takes a unique form, and in that sense I can say there is a direct relation between the primary and tertiary or quaternary structures.

Table 1.1. Amino acids summary table.

α –amino acid	3–Letter	1–Letter Side	Chain	Polarity	%	Essential
Alanine	Ala	A	–CH_3	nonpolar	7.8	No
Arginine	Arg	R	–$(CH_2)_3$NH–C(NH)NH_2	basic polar	5.1	Cond
Asparagine	Asn	N	–CH_2CONH_2	polar	4.3	No
Aspartic acid	Asp	D	–CH_2COOH	acidic polar	5.3	No
Cysteine	Cys	C	–CH_2SH	nonpolar	1.9	Cond
Glutamic acid	Gly	E	–CH_2CH_2COOH	acidic polar	6.3	Cond
Glutamine	Gln	Q	–$CH_2CH_2CONH_2$	polar	4.2	No
Glycine	Gly	G	–H	nonpolar	7.2	Cond
Histidine	His	H	–CH_2–$C_3H_3N_2$	basic polar	2.3	Yes
Isoleucine	Ile	I	–$CH(CH_3)CH_2CH_3$	nonpolar	5.3	Yes

CHAPTER 2

Electromagnetic Stability

Abstract: This chapter introduces the covalent and non-covalent intermolecular bonds as major actors in the basic macromolecule structure. It states the importance of the ionization energy in bonds and it determines the electronegativity or polarity quantification as an effective discriminator for the type of bond. With this quantification, amino acids are classified into four groups: polar basic, polar acidic, polar neutral and non-polar.

Keywords: Bonds, Chemical bonds, Electronegativity, Hydrogen bonds, Hydrophobic forces, Ionization energy, Intermolecular bonds, Macromolecules, Polarity, Van der Waals forces, Valence electrons.

2.1. IONIZATION ENERGY

The ionization energy [1] is the minimum amount of energy required to separate an electron from an atom at a distance where no electrostatic interaction between the ion and electron exists [2], its measurement unit is kcal/mol (cal/mol, see Glossary). This energy is used to form chemical bonds between atoms, particularly with the most distant electrons from the atomic nucleus as they require less ionization energy to be attracted.

2.2. CHEMICAL BONDS

The final structure of all molecules in 3–D space is the result of the electromagnetic stability reached by the chemical bonds in atoms when they share or give valence electrons [3], these electrons are always found in the outermost shell of the atomic nucleus. It is important to note that not all electrons form bonds, the inner shell electrons are excluded, as the ionization energy required here would be larger than the energy required for the valence electrons in the

Carlos Polanco

outermost shell. The quantification of the ionization energy makes possible the classification of bonds into two groups: covalent bonds if they share valence electrons or non-covalent bonds if they gain or lose electrons.

2.2.1. Covalent Bond

If the atom bond is sharing instead of giving valence electrons, the chemical bond will be covalent [4], there can be double or triple bonds when four or six valence electrons are shared (two or three by each atom), these bonds are stronger but at the same time less stable than non-covalent bonds (see Fig. **2.1**). When two atoms share a valence electron the form bonds can get in 3–D space is restricted. Covalent bonds actively participate in the structure adopted by the molecule although not in a definitive way as non-covalent bonds also have influence.

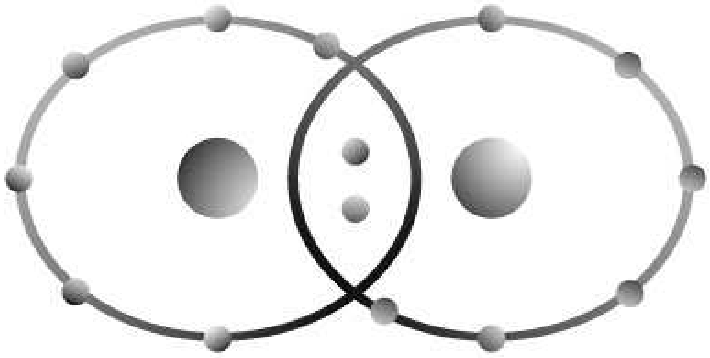

Fig. (2.1). Covalent bond is when two or more atoms share one or more electrons. This force is 5 times stronger here than in non-covalent bonds.

Covalent bonds are featured by sharing valence electrons between atoms; however it is possible to have one of the atoms with more attracting force on the electrons valence than the other, due to the positive charge in its atomic nucleus, which makes the distance between the shared valence electrons that atom nucleus shorter (see Fig. **2.2**). The attraction force between the shared valence electrons is called electronegativity, this property acts at an atomic level but it can also be extended to the macromolecule structure, giving rise to an important sub-classification for covalent bonds: polar bonds and non-polar bonds.

Polar bond Is when the electronegativity difference is greater than 0.5 but less than 1.7.

Non-Polar bond Is when the electronegativity difference is less than 0.5.

The chemical polarity or bond polarity occurs when there is an asymmetric distribution of electrons in the bond shared by two atoms. This takes place when the atoms of a molecule have different electronegativity, for instance, in a molecule formed by two identical atoms $A_i - A_i$, the electronegativity will be the same and the bond will be non-polar, therefore the electron distribution around the two atoms will be symmetric; whereas in a molecule formed by different atoms *e.g.* $A_i - A_j$, where atom A_j is more electronegative than atom A_i, it will attract more the electrons from the chemical bond forming an asymmetrical distribution. I can say that bond $A_i - A_j$ is polar and although the molecule, in general terms is neutral, due to this asymmetric distribution A_j will have a negative density charge which will be represented as δ −, and A_i will have a positive density charge represented as δ +. This statement gives rise to another classification for polar bonds into four types: polar basic (δ +), polar acidic (δ −), polar neutral or polar (δ +/−), and non-polar. These four types of polar bonds can be observed particularly in living organism (see Table **1.1**).

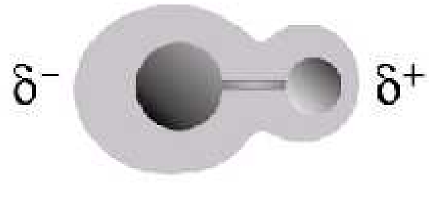

Fig. (2.2). Dipole. Molecular orientation where the positive end of a dipole (δ +) s near the negative end of another (δ −), exerting attraction.

The final characterization of a molecule is the result of the bond electromagnetic

Part II-FOUNDATIONS

This part constitutes the core of the book. It explains the supervised method called the polarity index method, and it shows the results obtained when the method is applied to several groups of proteins and peptides. It basically explains how the polar profile of a protein is capable of predicting the main action associated with it. This unit is divided into four sections: the first section aims to characterize the polarity index method; the second describes the mathematical calculation; the third is about the computing platform and architecture required to execute the programs; and the last chapter shows the first results obtained by the polarity index method on the group of antimicrobial peptides.

CHAPTER 3

Polarity Index Method

Abstract: This chapter will describe one supervised algorithm that is frequently used to predict the function of proteins, the quantitative structure-activity relationship model (QSAR). I will discuss its advantages and disadvantages, and I will present a model called the polarity index method. This model shows a high degree of efficiency in predicting the main function of peptides and proteins by inspecting the linear sequence of macromolecules to evaluate a single physico-chemical property, polarity.

Keywords: Bioinformatics algorithms, Deterministic algorithms, Macromolecules, Physico–chemical properties, Polarity index method, Proteomics, Quantitative structure activity relationship models, Supervised algorithms, Stochastic algorithms, *Training data*.

3.1. SUPERVISED ALGORITHMS

A supervised algorithm [1] is a mathematical algorithm that searches for a particular feature within a group of elements provided for a study. It differs from an unsupervised algorithm, which explores the group studied and infers from it the distinctive characteristic that identifies the group. Supervised algorithms do not perform this search.

3.2. QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP MODELS

Most bioinformatics algorithms oriented to proteomics use this type of model; its use has been constantly growing during the last five decades. Quantitative structure-activity relationship (QSAR) models always require the input of a search pattern, provided either explicitly or implicitly, in the *training data*. QSAR models [2] are particularly suited to the identification of a specific characteristic

Carlos Polanco

of the organism being studied from another of its properties, usually a physico-chemical [3] or structural property. In this type of model, there are deterministic [4] and stochastic [3, 5] implementations; the latter being the easiest to computationally implement, as stochastic models do not require one to retain much information for their implementation, and they can handle several properties simultaneously

3.2.1. Advantages

It is not necessary to have a complete understanding of the phenomenon being studied to use QSAR models; it is sufficient to feed the models with representative training data of the profile being searched. These models usually require a subsequent refinement to improve their efficiency and minimize the false-positives they always generate. They are used as a first filter to reduce the number of candidates evaluated; their design is not complex, and their efficiency depends on the profile being searched in the training data.

3.2.2. Disadvantages

QSAR models do not always produce a small number of candidates; therefore, their efficiency is frequently not documented in detail, to the detriment of their attributes, as the quality of their results does not depend on the model, but on the profile being searched. Because these are approximation models, their main disadvantage is that always produce candidates; therefore it is advisable to index the candidates with a percentage of relative effectiveness.

3.2.3. Minimizing Risk

Consider minimizing false-positives continually, as it is a maximizing/minimizing method. The strategy to minimize false-positives implies the implementation of two sets of information from the beginning: the training data, which is representative of the property evaluated, and the test set of false positives that will allow the efficient calibration of the method. Both of them will be necessary for the double-blind statistical test [6] applied to the QSAR model.

3.3. POLARITY INDEX METHOD

The polarity index method [8 - 10] is a QSAR method that measures the polarity of a protein by reading its primary structure and identifying the main action of the peptide or protein. This chapter will describe in detail how the method works. The following chapters will show the results obtained with this method.

3.3.1. Matrices $\sum^{n}_{i=1} Q_i$ and Q_i

The method starts by converting each sequence from the training data to its polar format, *e.g.*, to convert the sequence WFQNRRMKWKK to its numerical equivalent, I substitute each amino acid in the sequence with its polar equivalent according to (Table **3.1**). The result will be the following numerical sequence 44331141411.

Table 3.1. Polar classification.

Symbol	Category	1–letter code	Numerical equivalence
P+	Basic polar	H, K, R	1
P–	Acidic polar	D, E	2
N	Polar	C, G, N, Q, S, T, Y	3
NP	Nonpolar	A, F, I, L, M, P, V, W	4

Data from [7] 20 amino acid classification differentiated by their side-chain according to their polarity characteristics

From this numerical sequence I build matrix Q_i, taking each number in the sequence from left to right by pairs, one at a the time, and adding the incidents in $Q(i, j)$ where i ={P+, P–, N, NP} and j = {P+, P–, N, NP}. For example, taking the numerical sequence 44331141411 by pairs according to the instructions, the first element is $(i, j) = (4, 4)$, the second $(i, j) = (4, 3)$, the third $(i, j) = (3, 3)$, and so forth until the last element $(i, j) = (1, 1)$. The incidents are recorded in matrix Q_i (see Eq. 3.1). The same procedure is carried out for the n sequences forming the *training data*, generating an equal number of matrices Q_i.

$$Q_1 = \begin{pmatrix} 2 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 2 & 0 & 1 & 1 \end{pmatrix} \qquad \textbf{(3.1)}$$

CHAPTER 4

Mathematical Foundation

Abstract: This chapter explains the mathematical background used for polarity index method described in Chapter 3 and its relation with the catastrophic bifurcation points located in the geometric representation of the relative frequencies of a protein group. I discuss the singularities and regularities of this geometric representation and how this metric identifies the main action of a protein with a high level of accuracy. I introduce the concepts of maximum points, minimum points and saddle points, I also calculate the smooth curve from matrix $Q_i + \sum_{i=1}^{n} Q_i$ (Sect. 3.3.2), and I justify the exhaustiveness of the metric.

Keywords: Catastrophe theory, Catastrophic bifurcation points, Critical points, Differential calculus, Distributed computing, Function, Polarity index method, Regularity, Serial computing, Singularity, Smooth curves.

4.1. SMOOTH CURVES

The graph of a function [1] is only the geometric representation of a precise rule (or function) showing a pattern. Fig. (**4.1**) shows the graph of the function sin(x), evaluated in each of the points in the interval [-3.14, 3.14] (see x–axis), note this graph is contained in the interval [-1, 1] (y–axis). Not all these representations are smooth curve graphs, you can see the graph of the function abs(x) (Fig. **4.2**).

4.1.1. Critical Points

Critical points [1, 2] are those points on the x–axis where the graph of the function reaches the maximum or minimum values. In the function sin(x) (Fig. **4.1**), these two points are reached in the values $y = 1$, and $y = 1$, on the y–axis, and $x = 1.57$ and $x = 1.57$, on the x–axis respectively. There are other points equally important that are also critical points, they are called saddle points or

Carlos Polanco

inflection points and are located on the x–axis where the graph concavity changes. To illustrate this, see point $x = 0$, located on the x–axis, before this point the concavity of the function graph points up, after this point the concavity points down. In summary the function sin(x) has in the interval [-3.14, 3.14] three critical points: a minimum point $x = 1.57$, a maximum point $x = 1.57$, and a saddle point $x = 0$, all of them located on the x–axis.

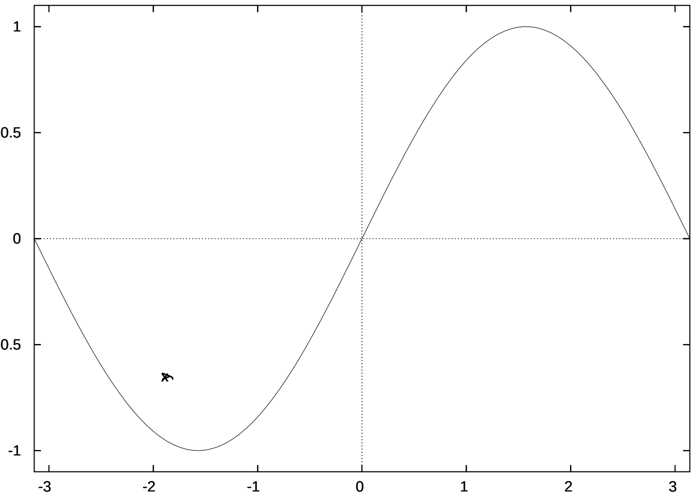

Fig. (4.1). Geometric representation of the function sin *(x)*.

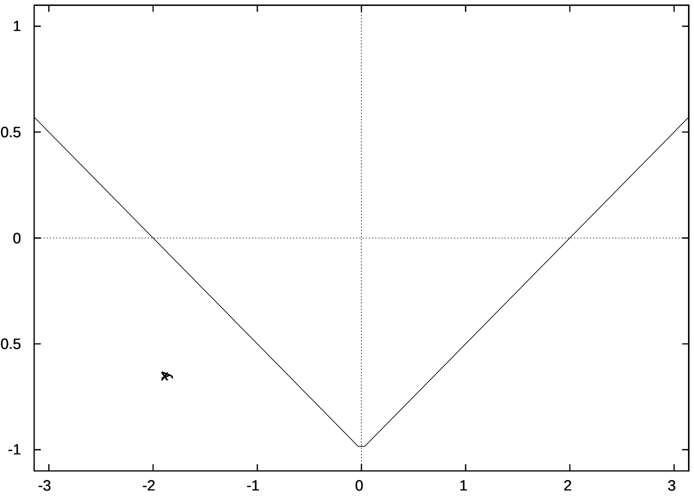

Fig. (4.2). Geometric representation of the function abs *(x)*.

4.2. SINGULARITIES AND REGULARITIES

Usually a singularity occurs just before and after a point where there is a regularity in the graph of the function. It is not always in the saddle points where a singularity occurs. Let us observe point x = 3 in the graph of the function (2x-8)/(4x-12) (Fig. **4.3**), where I can see a regularity between the before and after on the x–axis, however x = 3 is not a saddle point. 4.4 Matrix $Q_i + \Sigma_{i=1}^{n} Q_i$.

The same can be seen in point x = 0, (Fig. **4.1**). The importance of singularities is that regularities are observed when singularities are present [3].

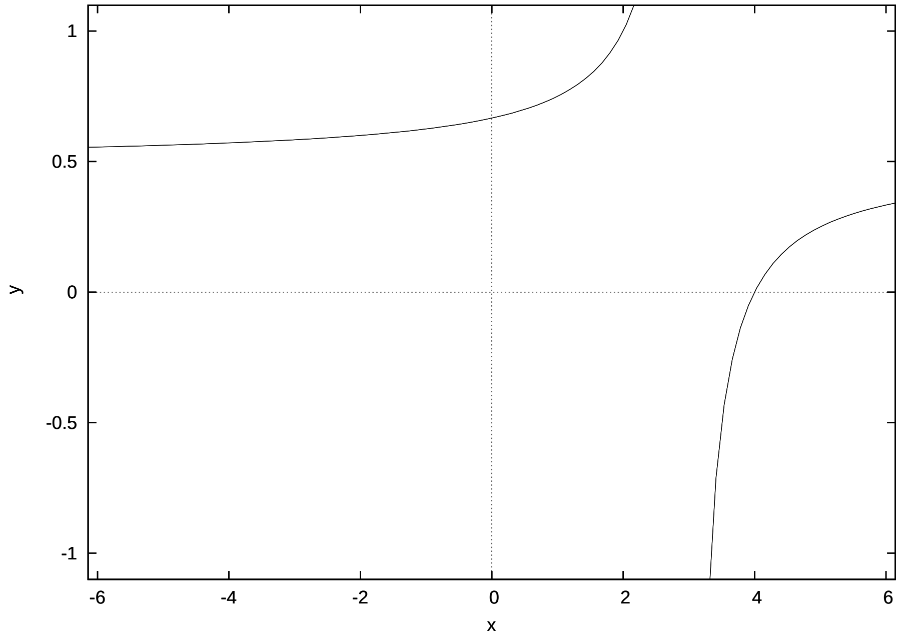

Fig. (4.3). Geometric representation of the function (2x-8)/(4x-12).

4.3. CATASTROPHIC BIFURCATION POINTS

The points where I find singularities [4] are called catastrophic bifurcation points [5, 6], they are particularly useful to study dynamic systems representing biological phenomena that can not always be accurately described by differential calculus. These points have a specific application in the analysis of the polymerization process of proteins, and in the Proteomics [7] field in general. They are a fundamental part of the Catastrophe theory which represents the tendency of structural stable systems to manifest discontinuities, that can produce

CHAPTER 5

Computational Implementation

Abstract: This chapter discusses the computational profile required for the implementation of the polarity index method (Sect. 4) from two platforms: distributed computing and serial computing. It also shows the advantages and disadvantages of each one of them and the processing time used by the method. Finally, it describes the degree of automation reached by the method.

Keywords: Antimicrobial peptides, Distributed computing, Microprocessing, Polarity index method, Proteomics, Serial computing, Training data.

5.1. TYPES OF PROCESSING

Here I will discuss two computational platforms: distributed computing; and serial computing [1]. The former allocates the tasks in two or more processors; while the Polarity index method (Sect. 4) has two stages. The first stage determines the profile searched (see matrix $\frac{1}{n_t}$ where n_t is the maximum processing time. This will occur provided the tasks or programs are independent, but if their nature implies a relation between the processes, this platform will probably increase the processing time by a factor of $1.2n_t$. For instance, if a task has four routines and all of them are independent from each other it will be possible to make the most of a platform of multiple nodes, assigning a task to each processor, this way the total processing time will be $\frac{1}{n_t}$ where n_t will be the average processing time for the tasks. However, if the tasks to be performed depend on each other, the transit of information between the nodes will increase substantially and instead of reducing the processing time it will be increased, as the platform is not fully maximized, having only an excessive traffic of information. A serial computing platform corresponds to computers that only have a CPU processor, in this platform the tasks are concurrent, that is one after the other. It is essential to always bear in mind the nature of the process to be run on a platform, so one can choose the computational platform that best suits the needs of the process.

5.2. MICROPROCESSING

During the last two decades the processors that were originally for video graphics video have been designed for different scientific applications such as Field-Programmable Gate Arrays (FPGA) [2], and more recently, the Graphics Processing Unit (GPU) [3]. They even have their own ROM–memory and RAM–memory, in such a way that they can replicate hundreds of times the work of only one processor. The most versatile first architecture language is Handel–C [4], its programming syntax corresponds to a low level language but its numerical libraries demand excessive processing memory. The language used for second architecture is called CUDA [5], it has the syntax of a high level language and has numerical libraries integrated, which are transparent to the user, making the codification of tasks very friendly. A microprocesing processor can reside in a laptop, it allows replicating hundreds of processors without any additional requirements of space, and the interface of the CPU is also transparent to the user.

5.3. AIMS & FOCUS

while the second stage evaluates the different groups of proteins, both tasks are profile searched (see matrix $Q_i + \Sigma_{ni=1} Q_i$ Sect. 3.3.2) using the training data, while the second stage evaluates the different groups of proteins, both tasks are performed simultaneously. This is the reason the structure of the polarity index method runs optimally in a distributed computing platform, because the tasks are independent. The program is capable to analyze each protein group in every processing node, or to divide one group in *n* available nodes. This is particularly useful if it is required to evaluate 20^n peptides of *n* length, this way the possible combinatorial of the group can be divided between all the available nodes. Polarity index methodindexpolarity index method can be very useful in Proteomics to explore the change of the polar profile of a peptide group of fixed *n* length (Sect. 11).

5.4. CONSIDERATIONS

In this chapter, I discussed the advantages and disadvantages of the two platforms where polarity index can be run: distributed computing and serial computing. It was emphasized the processing time required, as in a future application it will be used to evaluate 20n peptides of *n* length (Sect. 11). In the next chapter, I will show the main results obtained when the method is applied to antimicrobial peptides. This group is of great interest because of the incidence, and prevalence of associated diseases, and where this method has shown a high discriminative capacity.

REFERENCES

[1] Loewe L. Global computing for bioinformatics. Brief Bioinform 2002; 3(4): 377-88. [http://dx.doi.org/10.1093/bib/3.4.377] [PMID: 12511066]

[2] Herbordt MC, Vancourt T, Gu Y, *et al.* Achieving high performance with FPGA-based computing. Computer (Long Beach Calif) 2007; 40(3): 50-7. [http://dx.doi.org/10.1109/MC.2007.79] [PMID: 21603088]

[3] Pratx G, Xing L. GPU computing in medical physics: A review. Med Phys 2011; 38(5): 2685-97. [http://dx.doi.org/10.1118/1.3578605] [PMID: 21776805]

[4] Polanco González C, Nuño Maganda MA, Arias-Estrada M, del Rio G. An FPGA implementation to detect selective cationic antibacterial peptides. PLoS One 2011; 6(6): e21399. [http://dx.doi.org/10.1371/journal.pone.0021399] [PMID: 21738652]

[5] Park S, Shin SY, Hwang KB. CFMDS: CUDA-based fast multidimensional scaling for genome-scale data. BMC Bioinformatics 2012; 13 (Suppl. 17): S23. [http://dx.doi.org/10.1186/1471-2105-13-S17-S23] [PMID: 23282007]

CHAPTER 6

Pathogenic Bacteria

Abstract: This chapter includes the identification of a group of peptides known as antimicrobial peptides. They have an important role in the immune system of all living organisms. The identification of the main function associated to each peptide was made by calculating the polar profile of the peptide with the QSAR method called polarity index, already described in Sect. 3.3. The peptides computationally tested were taken from two different databases: Antimicrobial Peptides Database (APD2), and Uniprot Database, which included the set of peptides called selective cationic amphipatic antibacterial peptides.

Keywords: Antibacterial peptides, Antimicrobial peptides, Antrax, Immune system, Gram staining, Serial computing, Listeriosis, Living organisms, Matrix $\sum_{i=1}^{n} Q_i$, Selective cationic amphipatic antibacterial peptides.

6.1. FUNCTIONALITY

In this work the main function of a protein will be interpreted as the main action a protein has towards a microorganism that has been previously identified through experimental tests in labs and has been reported in different scientific journals and registered in several databases. A brief inspection to any of those databases will show that a peptide hardly has only one function associated to it, there are usually several actions associated to a peptide. This factor should be considered when using this information. Nowadays I have a significant number of databases; some of them focus on a particular group of peptides or proteins, while others are more general. It is important to point out that with an ever increasing number of databases, it is essential to make sure they receive adequate maintenance and are regularly updated. In this chapter, I have chosen two databases: Antimicrobial Peptides Database (APD2) [1], and Uniprot Database [2], as well as the set of selective cationic amphipatic antibacterial peptides, described in Sect. 6.3.3 from the work of Del Rio *et al*. [3]. I took several groups of peptides and proteins from them and identified the main action of these groups using the QSAR polarity index method described in Sect. 3.3 the main action of those groups.

6.2. CURRENT TAXONOMY

I took from Antimicrobial Peptides Database (APD2) [1] the groups: bacteria (sub classified in Gram + and Gram –), fungi, and cancer cells. From Uniprot Database [2] I took the group of proteins associated to influenza type A-H1N1 and Microbacterium tuberculosis (both located in homo sapiens), and I also took 30 peptides called selective cationic amphipatic antibacterial peptides (SCAAP) [3]. The following classification uses as *training data* peptides with unique action, as already described in Sect. 3.2. In order to find them, it was necessary to search all groups and make sure a peptide or protein was only in one of them, therefore the training data was formed only by those peptides with a unique action.

6.3. BACTERIA

Antibacterial peptides [4] are those that inhibit or nullify the growth of bacteria. A protein is more toxic to bacteria when there is a small amount of protein diluted in the experimental tests. One technique used to differentiate and identify two groups of bacteria by color is the Gram staining technique [5, 6]. Gram + are the bacteria that get dark blue or violet, while Gram– are the bacteria pink or red (gram staining, see Glossary). The bacteria that cause diseases are usually different from those living in our body. Bacterial diseases normally appear after surgical interventions, accidents, or any other cause that depletes the immune system.

6.3.1. Gram + Bacteria

Some of the infections caused by this bacteria group in human beings are:

Erisipelotricosis This is a slow progression cutaneous infection caused by the *bacterium erysipelothrix rhusiopathiae* that grows mainly in a medium with dead or decaying matter.

Listeriosis This illness is caused by the *bacterium listeria monocyitogenes*; it has a wide variety of symptoms. It is not frequent in humans, however when it occurs it is extremely serious. It is featured by causing low morbidity but high mortality (30%) and in the case of pregnant women, infants and elderly over the age of 70, it rises up to 70%.

Antrax *The bacterium bacillus anthracis* is the cause of this illness, it usually infects the skin, lungs and the gastrointestinal tract, it is highly contagious and life threatening; it is generally transmitted by animals. In its inactive state (spores) they can live in animal fur or on the earth for decades. It can also be transmitted by eating contaminated meat or by inhaling the spores.

Bioinformatics Test

When polarity index method (Sect. 3.3) evaluated the set of peptides from the Gram + bacteria group, matrix $\sum_{i=1}^{n} Q_i$, described by Eq. 6.1, was generated. When the graph of this matrix is plotted, as instructed in Sect. 4.1, Fig. (**6.1**) is displayed.

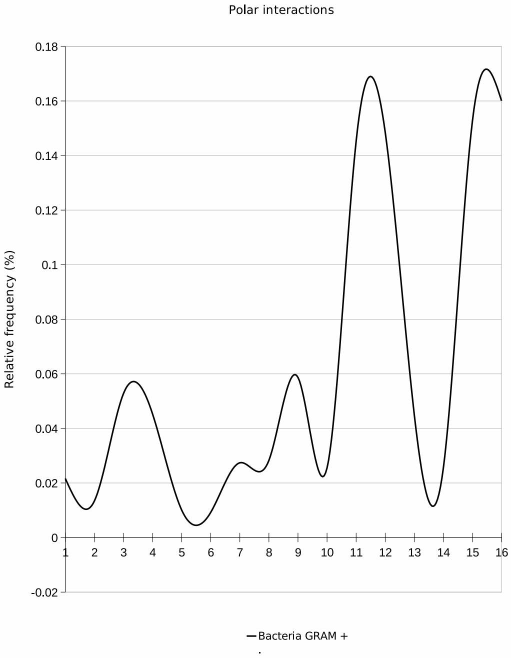

Fig. (6.1). Relative frequency distribution of Gram + bacteria group [1, 7 - 9].

Part III-STRUCTURAL PROPERTIES

This part of the book describes the application of the supervised polarity index method for the identification of the structural properties associated to proteins. It includes an analysis of three structural groups of proteins: cell penetrating peptides, amyloid proteins and globular and fibrous proteins. It is known that the first group of proteins penetrates the membrane subject, the second concentrates in proteins related to Amyloidosis and the third group characterizes a major group of proteins for their globular and fibrous morphology.

CHAPTER 7

Cell Penetrating Peptides

Abstract: This chapter describes the identification of a group of peptides called Cell penetrating peptides (CPP) that are featured by their ability to penetrate the membrane of different microorganisms. They were identified using the polarity index method described in Sect. 3.3. For this purpose two classifications were taken into account: non endocytic and endocytic pathway uptake mechanisms. The peptides studied were taken from CPPsite Database (CPPsite), and the set of selective cationic amphipatic antibacterial peptides (SCAAP) described in Sect. 6.3.3. The comparative study of these two groups made possible the identification of a particular reason for the toxicity of the selective cationic amphipatic antibacterial peptides.

Keywords: Antibacterial peptides, Carpet-like model, Cell penetrating peptides, Endocytic pathway, Inverted micelle formation, Matrix $\sum_{i=1}^{n} Q_i$, Membrane thinning model, Non endocytic pathway, Polarity index method, Pore formation, Toroidal model.

7.1. DESCRIPTION

It is known that peptide toxicity is partially related to the mechanism they use to penetrate the cell membrane of a microorganism, this process is called transduction [1], and the peptides with this characteristic are called Cell penetrating peptides (CPP) [2]. One of the first uses CPPs had was as cargo or Trojan peptides [3] to transport other peptides with a distinctive toxic action towards a target pathogen agent. CPPsite [4] is a database specialized in this type of peptides with a significant number of CPPs grouped under different classifications, being one of them the transduction mechanism of a peptide. This classification considers two groups: *endocytic pathway*, where the transduction process depends on the lipid - aqueous medium and the membrane; and *non*

Carlos Polanco

endocytic pathway, where the transduction is independent of the medium.

7.1.1. Non Endocytic Pathway

The non endocytic pathway can include different mechanisms such as: inverted micelle formation [5], pore formation [6], the carpet–like model [7] and the membrane thinning model [8]. The initial stage of all these mechanisms is related to the different electro-magnetic potential in peptides and the membrane subject, at a later stage the concentration of the peptide, the amino acids forming the CPPs and the lipid composition in each model membrane studied play an important role. In general terms it is more common to find a high number of CPPs and primary amphipatic CPPs with non endocytic pathway uptake mechanism.

Inverted micelle formation This is a model where the internalization occurs with the positive CPPs and the negative lipid membrane, although the hydrophobic amino acids from the CPPs and the membrane subject can also participate.

Pore formation In this classification there are two models, barrel and toroidal. In the former CPPs form a barrel by which hydrophobic amino acids are close to the lipid chains, and hydrophilic amino acids form the central pore. In the latter model, lipids bend in a way that CPPs are always close to the headgroup

Carpet–like model The interaction between negatively charged membrane, and positively CPPs result in a carpeting form, and the thinning of the membrane form.

Bioinformatics Test

When polarity index method (Sect. 3.3) evaluated the set of cell penetrating peptides non endocytic pathway, generated matrix $\sum_{i=1}^{n} Q_i$ described by Eq. 7.1 . When this matrix was plotted as mentioned in Sect. 4.1 as shown in Fig. (**7.1**) .

$$\sum_{i=1}^{n} Q_i = \begin{pmatrix} 0.13 & 0.00 & 0.05 & 0.14 \\ 0.01 & 0.00 & 0.00 & 0.01 \\ 0.05 & 0.01 & 0.04 & 0.06 \\ 0.13 & 0.01 & 0.07 & 0.21 \end{pmatrix} \qquad \textbf{(7.1)}$$

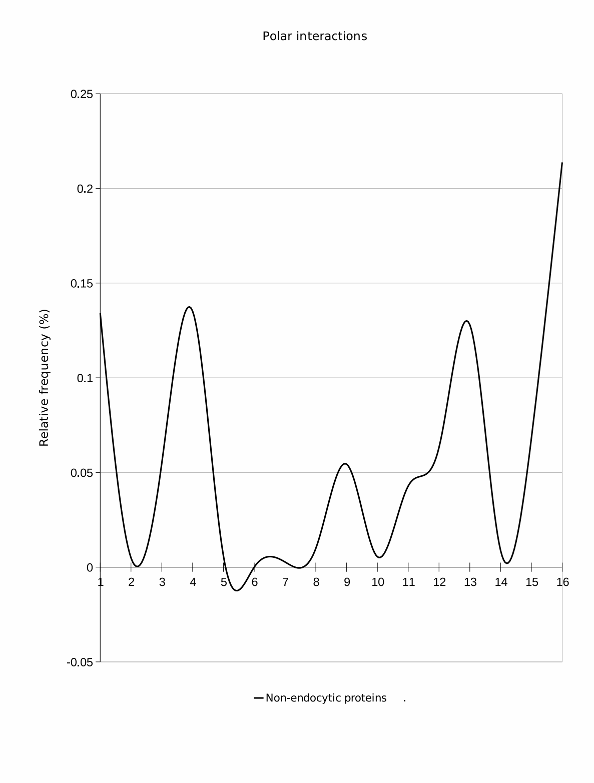

Fig. (7.1). Relative frequency distribution of CPP non endocytic pathway set [4, 9 - 16]

7.1.2. Endocytic Pathway

The endocytic pathway uptake mechanism considers several types of transduction, in-cluing the phagocytosis of large particles. This pathway is associated with the inwards folding of the outer surface of the plasma membrane, which results in the formation of vesicles.

Bioinformatics Test

When the set of CPPs endocytic pathway was evaluated by polarity index method, according to description in Sect. 3.3, it generated matrix $\sum_{i=1}^{n} Q_{i}$, described by Eq. 7.2. When I plotted this matrix as mentioned in Sect. 4.1 as shown in Fig. **(7.2)**

CHAPTER 8

Amyloid Proteins

Abstract: This chapter describes the main results obtained when the polar profile of peptides associated to Amyloidosis is determined by polarity index method (Sect. 3.3), and the relation these peptides have with a group of proteins identified by their structure-function called natively folded proteins, partially folded proteins, and natively unfolded proteins. Amyloidosis is a term that groups a set of mental diseases with neurodegenerative characteristics, caused by the agglomeration of natively unfolded proteins on the neurons and lipoproteins. This chapter shows that the proteins that express neurons are natively folded proteins.

Keywords: Amyloid fibrils, Amyloidosis, Bacteria, Chylomicrons carry triglycerides, Fungi, High density lipoproteins, Intermediate density lipoproteins, Lipids, Lipoproteins, Low density lipoproteins, Natively folded proteins, Natively unfolded proteins, Neurodegenerative diseases, Neurons, Partially folded proteins, Polar profile, Polarity index method, Very low density lipoproteins, Virus.

8.1. PROTEIN FOLDING

The main function of a protein is strongly associated to the form or structure it adopts in a lipid–aqueous medium. This form is closely related to the electromagnetic balance [1] of its components, the amino acids, but it is also related to the electromagnetic balance of the receptor, in such a way that together the protein and the receptor form a new electromagnetic balance.

8.2. POLAR PROFILE

Proteins that adopt the typical structure of a generic group called natively folded proteins [2, 3] have a function associated to a protein or peptide group. On the other hand proteins that do not adopt a structure, *i.e.* natively unfolded proteins

Carlos Polanco

[2, 3], were thought to have an unknown function. Today I know that natively unfolded proteins are the cause of mental chronic degenerative diseases grouped under the term Amyloidosis. They tend to agglomerate on the neurons [4], affecting their electromagnetic balance. In the folded/unfolded process there are also intermediate states that originate a group known as partially folded proteins [2, 3]. This group such as the natively unfolded proteins is strongly associated to Amyloidosis. There are also proteins that transport lipids known as lipoproteins [5], but as I will see later on, their polar profile differs from the natively unfolded protein group.

8.3. LIPOPROTEINS

Lipoproteins have as main function the transport of lipids, a high percentage of them is associated to Amyloidosis, in their classification most of the proteins fall in one of these sub-groups: Chylomicrons carry triglycerides, and High density lipoproteins [6, 7].

Chylomicrons carry triglycerides "From the intestines to the liver, skeletal muscle,and to adipose tissue" [6, 7].
Very low density lipoproteins the liver to adipose tissue" [6, 7].
Intermediate density lipoproteins Intermediate density lipoproteins "Are intermediate between Very low density lipoproteins and Low density lipoproteins. They are not usually detectable in the blood" [6, 7].
Low density lipoproteins "Carry cholesterol from the liver to cells of the body. Low density lipoproteins are sometimes referred to as the *bad cholesterol* lipoprotein" [6, 7].
High density lipoproteins "Collect cholesterol from the body's tissues, and bring it back to the liver. High density lipoproteins are sometimes referred to as the *good cholesterol* lipoprotein" [6, 7].

Bioinformatics Test

When the set of lipoproteins was evaluated by polarity index method [8–14] (Sect. 3.3) it generated matrix $\Sigma_{i=1}^{n} Q_i$ described by Eq. 8.1. When this matrix was plotted as instructed in Sect. 4.1, as shown in Fig. (**8.1**).

$$\sum_{i=1}^{n} Q_i = \begin{pmatrix} 0.02 & 0.01 & 0.05 & 0.05 \\ 0.01 & 0.02 & 0.04 & 0.05 \\ 0.05 & 0.04 & 0.13 & 0.14 \\ 0.05 & 0.05 & 0.14 & 0.15 \end{pmatrix} \qquad \textbf{(8.1)}$$

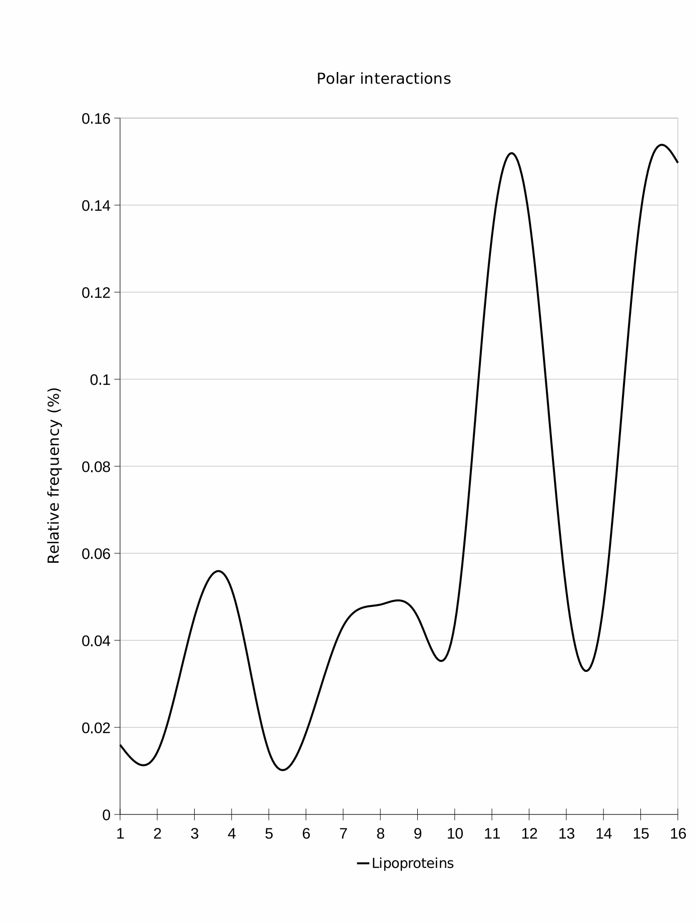

Fig. (8.1). Relative frequency distribution of CPP non endocytic pathway set [4, 8, 9, 11, 13, 14].

8.4. NATIVELY UNFOLDED PROTEINS

There is a group of natural proteins resistant to proteolysis and not soluble in water that do not adopt a typical structure. They are toxic when they settle in an organ and are known as natively unfolded proteins [2, 3]. In some cases these

CHAPTER 9

Globular & Fibrous Proteins

Abstract: In this chapter, I discuss the main results obtained when calculating the polar profile of two important groups of proteins: globular and fibrous proteins, by polarity index method (Sect. 3.3). This classification, unlike other proteins mentioned in previous chapters of the book, is not a sub-division but a major classification from which all protein classifications come from.

Keywords: Fibrous proteins, Globular proteins, Polar profile, Polarity index method.

9.1. PRELIMINARIES

Proteins are morphologically divided into two major groups: fibrous and globular proteins. Fibrous proteins have a tertiary structure similar to a fiber and are not soluble in water, diluted saline solution, organic solvents, diluted acids and alkalis. Globular proteins, on the other hand, have spherical shape and are mainly soluble in one of those solutions; their morphological identification is fully documented, and they are easily recognized with lab techniques.

9.2. GLOBULAR PROTEINS

These proteins require the solubility of the blood or any other aqueous medium in cells and tissues. Their form is well defined: hydrophobic amino acids are in the interior of the protein, while amphiphilic amino acids are in the exterior interacting with water *e.g.* hemoglobin and enzymes.

Bioinformatics Test

When the set of globular proteins was evaluated by polarity index method [1 - 7]

Carlos Polanco

(Sect. 3.3), it generated matrix $\Sigma_{i=1}^{n} Q_i$ described by Eq. 9.1. When this matrix was plotted, as instructed in Sect 4.1, as shown in Fig. (**9.1**) .

$$\sum_{i=1}^{n} Q_i = \begin{pmatrix} 0.02 & 0.01 & 0.05 & 0.05 \\ 0.01 & 0.02 & 0.04 & 0.05 \\ 0.05 & 0.04 & 0.13 & 0.14 \\ 0.05 & 0.05 & 0.14 & 0.15 \end{pmatrix} \qquad \textbf{(9.1)}$$

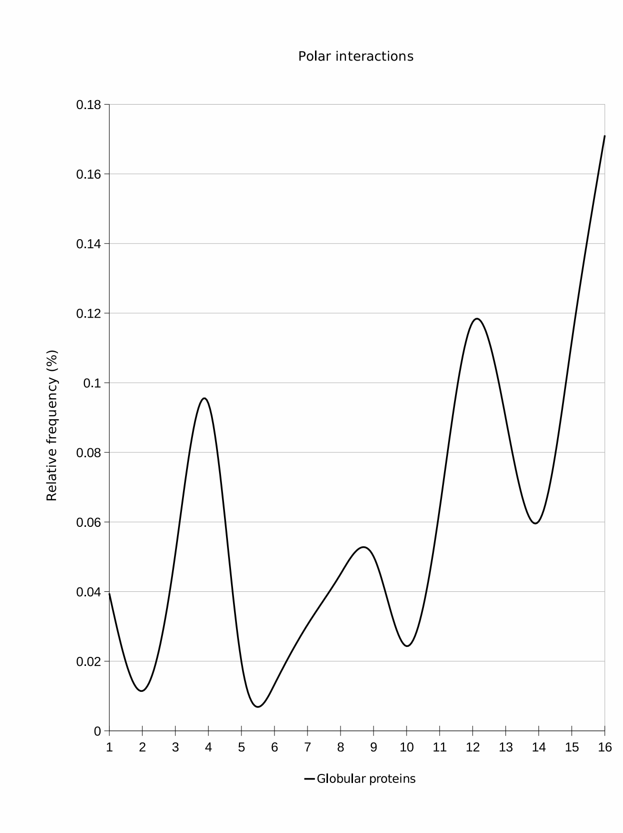

Fig. (9.1). Relative frequency distribution of the globular protein group [9].

9.3. FIBROUS PROTEINS

The main function of fibrous proteins is to provide mechanical support to organisms; they are water insoluble and form filaments, *e.g.* α–keratin from hair and nails and collagen from skin, teeth and bones.

Bioinformatics Test

When polarity index method [1 - 7] evaluated the set of fibrous proteins [1 - 7] (Sect. 3.3), matrix $\Sigma_{i=1}^{n} Q_i$, described by Eq. 9.2, was generated. When I plotted the graph as mentioned in Sect. 4.1, as shown in Fig. (**9.2**).

$$\sum_{i=1}^{n} Q_i = \begin{pmatrix} 0.02 & 0.02 & 0.05 & 0.05 \\ 0.02 & 0.02 & 0.03 & 0.05 \\ 0.04 & 0.04 & 0.11 & 0.12 \\ 0.06 & 0.05 & 0.13 & 0.16 \end{pmatrix} \qquad \textbf{(9.2)}$$

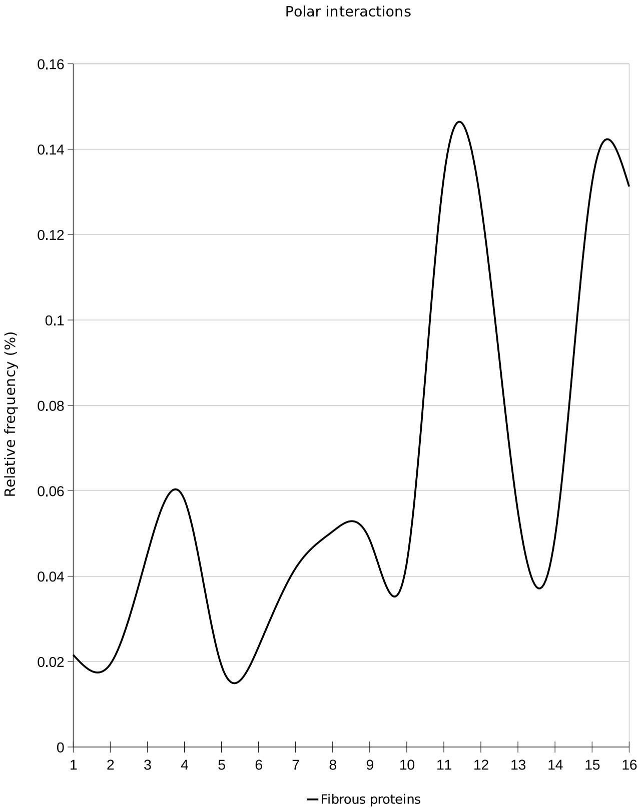

Fig. (9.2). Relative frequency distribution of fibrous proteins group [10].

Part IV-BIOGENESIS & TRENDS

In this part, I describe the results of the biogenetic experiments of Stanley Miller & Harold Clayton Urey, Sidney Walter Fox & Kaoru Harada, and Bernd Michael Rode obtained by computational modeling. I also analyze the proteins produced by each experiment, and in the last section, I present what will be, in the opinion of the authors, the future trend of computational proteomics.

CHAPTER 10

Biogenetic Experiments

Abstract: This chapter describes the experiments related to the origin of life developed by Miller & Urey, Fox & Harada, and Rode oriented to the polymerization of the prebiotic proteins. It is explained how each experiment with a different number of amino acids, proportion and methodology, made it possible to find a polar profile for their dipeptides and proteins, and how this information was used to study the trend of each experiment. The results show a common polar profile with the most preserved genes from three microorganisms.

Keywords: Biogenesis, Dipeptides, Fox & Harada experiment, Genes, Miller & Urey experiment, Origin of life, Prebiotic proteins, Rode experiment.

10.1. BIOGENESIS

The theory of biogenesis received acceptance with Louis Pasteur. This theory presents the following two issues: (i) life forms produce other life forms; and (ii) if life arose from another life form, from where did the first life form originate? The latter is known as the autotrophs theory, which has two statements: (i) a complex organism originated in a simple environment; and (ii) a simple organism originated in a complex environment. There is also the theory that life originated from heterotroph forms (*i.e.*, life forms that cannot produce their own food and although they can produce some compounds, they must depend on an external source to feed themselves). Humans and animals are heterotrophs. This hypothesis states that simple organisms slowly evolved from non-living matter under specific environmental conditions. Charles Darwin's theory uses this approach. Darwin's sustained the idea that in the warmth of a small lagoon with phosphoric and ammonium salts, light, heat, and electricity life could have been formed through

Carlos Polanco

chemical processes creating a protein compound where complex changes could be possible.

10.2. MILLER & UREY: PRIMORDIAL SOUP

The Miller & Urey experiment was significant, as it demonstrated that spontaneous generation of life was far more likely than expected. From the experimental point of view, there is no record of the controlled conditions in which the experiment was conducted; however, the experiment confirms that a certain number of biologically important amino acids can be synthesized with the application of electrical sparks to a gaseous mixture of ammonia, hydrogen, water vapor, and a simple organic substance, methane. The necessary chemical reactions to produce complex organic substances were completed under certain conditions, such as high temperatures, pressure, ultraviolet radiation, and electrical sparks. The elements were placed in a device and heated to induce evaporation, and then electrical sparks were applied periodically to simulate lightning. Scientists were surprised with the results; after a week there were several amino acids. After disclosure, several experiments were designed to verify the results; in all cases they showed the formation of amino acids.

Bioinformatics Test

When the set of proteins from the modeled Miller & Urey [1] experiment was evaluated by polarity index method [2 - 7] (Sect. 3.3), it generated matrix $\sum_{i=1}^{n} Q_i$, , described by Eq. 10.1. When I plotted this matrix as mentioned in Sect. 4.1 as shown in Fig. (**10.1**). The computational-mathematical model builds peptides in a linear format. It considers a group of 21 amino acids (Table **10.1**), where only 11 of them G, A, V, L, I, P,D, E, S, T, K are nowadays identified as basic amino acids, while the others {G, A, V, L, I, P, D, E, S, T, K} are classified as prebiotic amino acids (Table **10.1**).

$$\sum_{i=1}^{n} Q_i = \begin{pmatrix} 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.01 & 0.02 \\ 0.00 & 0.01 & 0.10 & 0.20 \\ 0.00 & 0.02 & 0.21 & 0.44 \end{pmatrix} \qquad \textbf{(10.1)}$$

10.3. FOX & HARADA: PROTEINOIDS

Here, I quote the Fox & Harada [9] experimental procedure. “Ten grams of L-glutamic acid was heated at 175-180° until molten (about 30 min.) after which period it had been largely converted to the lactam. At this time, 10g of DLaspartic acid and 5g, of the mixture of the sixteen basic and neutral (BN) amino acids were added. The solution was then maintained at 170°, 2° under an atmosphere of nitrogen for varying periods of time. Within a period of a few hours considerable gas was evolved, and the colour of the liquid changed to amber. When the inside of the tubes were observed and chromatograms were taken, it showed the presence of structures or proteinoids.

Table 10.1. Amino acid composition of the Miller experiment.

α–amino acid	1– Letter	PRO	SPK	FTT	MET	Abundance	Polarity
Glycine	G	X	++++	X	++++	440.0	nonpolar
Alanine	A	X	++++	X	+++	790.0	nonpolar
Valine	V	X	++	?	+++	19.5	nonpolar
Leucine	L	X	++			11.3	nonpolar
Isoleucine	I	X	++	?		4.8	nonpolar
Proline	P	X	+	?	+++	1.5	nonpolar
Aspart ic Acid	D	X	++	X	+++	34.0	acidic polar
Glutamic acid	E	X	++	X	+++	7.7	acidic polar
Serine	S	X	++			5.0	polar
Threonine	T	X	+			0.8	polar
Lysine	K	X				1.2	basic polar
Sarcocine	1					55.0	nonpolar
N–Ethylglycine	2					33.0	nonpolar
N–Methylalanine	3					15.0	nonpolar
β –Alanine	4		++	X	+++	18.8	nonpolar
β –Amino–η–butyric acid	5		+		++	0.3	nonpolar
β –Aminoisobutyric acid	6		+	?	++	0.3	nonpolar
γ –Aminobutyric acid	7					2.4	nonpolar
Norvaline	8		+++	X	+++	61.0	nonpolar
α–Amino–η–Butyric acid	9		++	?	+++	270.0	nonpolar

CHAPTER 11

Future Directions

Abstract: This final chapter describes what will be, in the opinion of the authors, the future trend for the construction of synthetic proteins in the bioinformatics field, their potential use in the pharmaceutical industry, and the impact that distributed computing will have on the algorithms designed to predict the main action of proteins.

Keywords: Distributed computing, Massive databases, Mathematical algorithms, Prebiotic proteins, Proteomics, Synthetic proteins.

11.1. PROTEOMICS TRENDS

The current trend in medicine is to focus on knowing the genes and their expression in proteins to find out the changes that affect health. This knowledge can be used to design pharmaceutical drugs or curative strategies that use the intrinsic ability that the human body has to build to maintain and repair itself. In order to determine the future trend [1, 2] of Proteomics, I have to point out that there is a wide range of mathematical algorithms, some demanding more computational resources than others. Their metrics usually consider two or more physico–chemical properties; however, their efficiency seems to be very limited as they are not 100% accurate. Total accuracy would mean a complete understanding of the principles behind the mechanics of the property This work examines the efficiency of a method based on only one property, polarity; and although this property has been extensively used and reported in several publications, the way it is used here has been completely different since an array with every possible polarity interaction was built. The method called polarity index requires to be trained with the *training data* of the group studied, once this is done it becomes a very agile algorithm. It is essential to deepen the understanding of the physico–chemical properties that feature a protein; these

Carlos Polanco

properties were determined more than five decades ago and now it is necessary to examine them under a different perspective. Polarity is included in 60% of all bioinformatics algorithms, however all metrics consider it as a value. This work shows that this property must not be studied in such a limited way, as its exhaustive analysis provides more understanding to the Proteomics field. Under Prebiotic proteomics, I have explored three important experiments Miller & Urey, Fox & Harada, and Rode focused on the prebiotic protein polymerization (Chapter . 10); the polarity profile resulting from these experiments was compared finding not only a coincidental trend but a way to verify the results of the experiments (Sect. 10.1). The purpose of this broad study is to understand the remote past of proteins and from that knowledge comprehend the functional divergences of the current protein groups. Bearing that in mind I think that the future trend of this field should consider a different approach and evaluate the metrics from another perspective, interpreting in other ways what I know about prebiotic proteins.

11.2. DISTRIBUTED COMPUTING

Distributed computing dates from the 50s decade, it is a computer network sharing their computational resources with other processors in the system. Processing power, memory, and data storage are available to users to perform tasks. A distributed network can be as simple as having a group of similar equipments working with the same operative system, or as complex as an interconnected net of systems working with a specific computational platform; with this technology different computers in a network share one or more of their resources. In this type of systems where all the resources are shared, the processor network can be turned into a supercomputer, and with a suitable interface the access to these systems will not differ greatly from accessing the resources of a local machine, enabling all authorized computers to share high processing power and storage capacity. Distributed computing is a resource with high impact on the Proteomics field, especially for those algorithms with highly independent tasks. The optimization of these computational tools will make possible the analysis of specific protein regions in databases where protein sequences are located, since massive databases have tens of thousands of protein sequences identified and though short proteins tend to decrease in organisms, it is not the case of large proteins.

11.3. SYNTHETIC PROTEINS

Synthetic proteins are manmade molecules using different techniques particularly computational algorithms. The first step to make this type of proteins is to modify certain amino acids in the linear sequence of the protein; they are called hybrid proteins. Both natural and synthetic proteins exhibit a wide range of applications in Biomedicine and drug therapy however, it is necessary to complement the existing databases to include reliable indicators, such as toxicity, to allow the construction of mathematical algorithms oriented to specific fields, as it is the particular case of the proteins with antibacterial action.

11.4. BIOINFORMATICS

The future development of Biomedicine is a multidisciplinary work where Biotechnology, Health Bioinformatics and Telemedicine participate, not only to have a curative and palliative health system but also to prevent diseases. In the recent years Bioinformatics has directed efforts to store biological information in several databases, for that purpose different data mining techniques have been applied to integrate the scattered information and select it automatically. This has evidenced the need of new tools to deal with the information as new techniques to find and extract data are being used. The trend of this field points to new programs totally parallelized run in mathematical co-processors called General–purpose computing on graphics processing units. The power of these systems makes possible the collection of large amounts of data, in a very short period of time, to be analyzed with informatic techniques obtaining useful information that can be applied to biomedical research.

11.5. CONSIDERATIONS

In this chapter, I have seen that the design of better algorithms has very little to do with computational development but with the granularity and sensitivity of the metrics involved and the adequate exploration of protein regions in the databases. In the next decades pharmaceutical drug design will be faced with a considerable reduction in experimental testing on animals; therefore, new algorithms under the guidelines mentioned above must be designed, but it will also be necessary to design computational biological scenarios to minimize the number of synthetic

Appendix A-Computational Tools

Abstract: This section discusses the Linux scripts and the FORTRAN–77 source codes used for the automatic execution of the polarity index method, in a LINUX Fedora–14 platform.

A.1. PRELIMINARIES

Polarity index method program was written in FORTRAN–77 [1] to be run in a LINUX Fedora–14 platform [2], LINUX scripts must be considered in case of making any changes before execution. The program is completely automated, however it is neces sary a preparatory procedure (Sect. A.2) to adequate the format of the files that will be analyzed and create some files that will be required. The processing time will depend on the number of proteins analyzed, therefore it is suggested to have a nohup command.

A.2. PREPARATION FILES

Polarity index method program requires the input file in POLAR–format or numeric code {1,2,3, y 4 } (Sect. 3.3.1). To achieve this it is necessary to copy the group of proteins in FASTA–format in the file namegroup.single [multiple].dat0, after that the file preparacionarchivos script (Tables **A.1**, **A.2**, **A.11**) ./preparacionarchivos namegroup unico must be executed. This script generates namegroup. single [multiple].dat2. The procedure must be repeated for each file that will be evaluated. Then the grupos.dat file must be created, by placing in the first line the namegroup to be evaluated and below the name of the other groups (Table **A.3**). Now the preparatorio3 script can be executed (./preparatorio3 namegroup unico).

Carlos Polanco

Table A.1. Previous process (part 1) to *polarity index method* program.

```
#! /bin/sh
#
#       Author     Carlos Polanco
#       Date       January, 2013.
#       Script     preparacionarchivos
#       email:     polanco@unam.mx
#
#   INPUT FORMAT: name unico
#   THERE SHOULD BE FASTA—format globular.unico.dat0,
#   AND fase2.f WHICH IS THE COMPLEMENTARY PART OF THE
#   PROGRAM'S FORMATTING.

clear
#
#   CONVERT FASTA—format TO POLAR—format.
#
./depurador $1 $2

largom=‘head −1 cifra.txt | cut −d” ” −f1‘
echo ”      implicit none” > fase1.f
echo ”      character ∗ 1 arreglo($largom)
echo ”      character ∗ 1arreglo2(40000)” >> fase1.f
echo ”      character ∗ 1 convert” >> fase1.f
echo ”      integer n, m, i” >> fase1.f
echo ”      open (1,file= ’entrada.txt’)” >> fase1.f
echo ”      open (2,file= ’salida.txt’)” >> fase1.f
echo ” 200  format ($largom(A1))” >> fase1.f
echo ” 300  format (40000(A1))” >> fase1.f
echo ”      n = $largom” >> fase1.f
echo ”      m = 40000” >> fase1.f
cat fase2.f >> fase1.f
gfortran fase1.f −o fase1
# PROGRAM TO CONVERT FASTA—format TO POLAR—format
#

./fase1
```

preparacionarchivos script that converts the FASTAformat POLAR–format for each protein groups.

Table A.2. Previous process (part 2) to *polarity index method* program.

```
#! /bin/sh
#
#       Author     Carlos Polanco
#       Date       January , 2013.
#       Name       preparacionarchivos
#       Script     depurador
#       email:     polanco@unam.mx
#
#  THIS SCRIPT RECEIVES THE FILE IN FASTA—format
#  GENERATES entrada.txt AND GIVES ITS MAXIMUM LENGTH
#  OF RECORD TO BE EXECUTED IN pasodefinitivo.f
clear
rm longitudes.txt
rm pad.awk
#echo ”STEP 1”
awk ’BEGIN{RS=”>sp”}NR>1{sub(”\n”,”\t”); gsub(”\n”,””);
print RS$0}’ $1.$2.dat0 | cut −f2 > $1.$2.dat1
IFS=” ”
while read A
do
   echo ”$A” | wc −c |cut −d” ” −f1 >> longitudes.txt
done < $1.$2.dat1
IFS=$SAVEDIFS
sort −n longitudes.txt | tail −1 > longitudes.txt1
long=‘echo | awk −F” ” ’{ print ($1)}’ longitudes.txt1 ‘
viene=‘echo s’\’n’”’‘
otro=‘echo ’,$0’‘
par=‘echo printf ’”’‘
echo ”$long” > cifra.txt
#echo ”STEP 2”
echo ”{” > pad.awk
echo ”$par%−$long$viene$otro” >> pad.awk
echo ”}” >> pad.awk
awk −f pad.awk $1.$2.dat1 > $1.$2.dat3
mv $1.$2.dat3  entrada.txt
```

depurador script that converts the FASTA–format POLAR–format for each protein.

Table A.3. Profile of *grupos.dat* file.

```
{\bf namegroup}
bacteriasplus
bacteriasminus
scaaps
tuberculosis
h1n1
fungi
cancer
```

grupos.dat file used for *preparacion3* script.

A.3. POLARITY INDEX METHOD PROGRAM

Polarity index method program is executed by huella3 script (Table **A.6**), once the files have been converted from FASTA–format to POLAR–format with the preparatorio3 script (Sect. A.2). Before executing huella3 script it is necessary to verify that the grupos.dat and analisis.dat files (Tables **A.3-A.5**) keep the same order (the latter file has to be generated), making sure that the first group of each file in namegroup is the one to be analyzed. It is recommended that the execution of huella3 (Table **A.6**) script uses a nohup command in this way (nohup ./huella3 namegroup unico), since this procedure can take hours of processing time in the equipment. The above mentioned method is divided into three parts (Tables **A.7** – **A.9**). Section (A.7) is used by huella3 script (Table **A.6**) to define the variables, section (A.8) defines the polarity profile of the namegroup and section (A.9) identifies the incidents. None of these sections is implemented by the user of the program.

Table A.4. Process (part 1) *polarity index method* program.

```
#! /bin/sh
#
#      Author     Carlos Polanco
#      Date       January , 2013.
#      Script     huella3
#      email:     polanco@unam.mx
#
clear
total=0
limite=100
while [ $limite −ge 75 ]
do
   rm totales.txt
   rm cifras.txt
   rm imagen.txt
#  INDICATE THE TARGET GROUP!
   ./dispara $1 $2

   IFS=” ”
   while read A
      total =0
      do
         ./huella $1 $A $2 >> totales.txt
         let ”total++”
         echo ”$total” >> cifras.txt
      done < grupos.dat
      IFS=$SAVEDIFS
      paste cifras.txt totales.txt >> imagen.txt
      echo ”COMPLETED PROCESS”
      limite=‘head −1 totales.txt | cut −d” ” −f1‘
      if test $limite −ge 75
      then
        echo ”$limite”
        cat imagen.txt > anteriortotales.txt
        cat analisis.dat > anterioranalisis.dat
        gfortran maximos.f
        ./a.out
      else
        echo ”IT REACHED THE LIMIT:  75%”
      fi
done
```

huella3 script iterates until the percentage of efficiency is reached.

Table A.5. Previous process to *polarity index method* program.

```
{\bf namegroup} 0 0
bacteriasplus 0 0
bacteriasminus 0 0
scaaps 0 0
tuberculosis 0 0
h1n1 0 0
fungi 0 0
cancer 0 0
```

analisis.dat file used for *huella3* script.

Table A.6. Internal process *huella3* script.

```
#! /bin/sh
#
#       Author     Carlos Polanco
#       Date       January, 2013.
#       Script     dispara
#       email:     polanco@unam.mx
#
clear
head −1 analisis.dat > analisis1.dat
IFS=” ”
while read A B C
do
   cat valores$1.$A.$2.net > objetivo.net
   gfortran coincidenciasotros.f
   ./a.out
   gfortran discriminante.f
   ./a.out
   cat temporal.net > sesgos$1.$A.$2.net
done < analisis.dat
IFS=$SAVEDIFS
IFS=” ”
while read A B C
do
   cat valores$1.$A.$2.net > objetivo.net
   cat $A.unico.dat2 > archivoperfil.net
   gfortran producetabla.f
   ./a.out > tablasdelperfil.f
done < analisis1.dat
IFS=$SAVEDIFS
rm losotrossesgos.f
cat analisis1.dat analisis.dat > tanalisis.dat
uniq −u tanalisis.dat > analisis2.dat
rm tanalisis.dat
IFS=” ”
while read A B C
do
  cat sesgos$1.$A.$2.net > temporal.net
  echo ”$B $C” > decision.txt
  gfortran generador.f
  ./a.out >> losotrossesgos.f
done < analisis2.dat
IFS=$SAVEDIFS
rm perfilesmaestro.net
gfortran coincidenciasmaestro.f
./a.out > nousado.txt
cat vieneP1.f > todo.f
```

```
titulo=‘head −1 analisis1.dat | cut −d” ” −f1‘
echo ” ” >> todo.f
echo ”c      $titulo” >> todo.f
echo ” ” >> todo.f
cat tablasdelperfil.f >> todo.f
cat vieneP2.f >> todo.f
echo ”c      Inicio ” >> todo.f
cat losotrossesgos.f >> todo.f
echo ”c      Final ” >> todo.f
echo ” ” >> todo.f
cat perfilesmaestro.net >> todo.f
echo ”       return” >> todo.f
echo ”       end” >> todo.f
cp viene3.f anteriorviene3.f
mv todo.f viene3.f
rm analisis1.dat
rm analisis2.dat
rm tablasdelperfil.f
```

dispara script evaluates each protein group.

Table A.7. Section (1) *polarity index method* program.

```
c       Author     Carlos Polanco
c       Date       January, 2013.
c       Program    vieneP1.f
c       email:     polanco@unam.mx
c
        implicit none
        integer arreglo(40000), suma, espe
        integer n,j,i,k, total, linealA(16)
        integer maestro(16),matriz(4,4)
        integer valores(4,2),cocientes1
        integer cocientes2
        real peso(4,4), lineal(16), comodin
        open (1,file=”candidato.dat”)
        open (2,file=”valor.dat”)
 200    format (40000(I1))
 34     format (16(f5.2,1x))
 35     format (f6.2,I2)
 53     format (A3,16(1x,I2),2x,I2,3x,I2)
c
c       RELATIVE FREQUENCY POSITION OF AMINO ACID IN
c       THIS GROUP.
c
```

Definition of variables used by *polarity index method* program.

Table A.8. Section (2) *polarity index method* program.

```
c        Author     Carlos Polanco
c        Date       January, 2013.
c        Program    Polarity profile
c        email:     polanco@unam.mx
c
c
c        RELATIVE FRECUENCY POSITION OF AMINO ACID IN
c        THIS GROUP
c
        peso ( 1, 1)    =    0.0216293484/  0.1600571871
        peso ( 1, 2)    =    0.0134101966/  0.1600571871
        peso ( 1, 3)    =    0.0527756102/  0.1600571871
        peso ( 1, 4)    =    0.0452774391/  0.1600571871
        peso ( 2, 1)    =    0.0100936964/  0.1600571871
        peso ( 2, 2)    =    0.0092285220/  0.1600571871
        peso ( 2, 3)    =    0.0273971763/  0.1600571871
        peso ( 2, 4)    =    0.0282623488/  0.1600571871
        peso ( 3, 1)    =    0.0586876348/  0.1600571871
        peso ( 3, 2)    =    0.0255226325/  0.1600571871
        peso ( 3, 3)    =    0.1452050358/  0.1600571871
        peso ( 3, 4)    =    0.1485215276/  0.1600571871
        peso ( 4, 1)    =    0.0449890457/  0.1600571871
        peso ( 4, 2)    =    0.0255226325/  0.1600571871
        peso ( 4, 3)    =    0.1527032107/  0.1600571871
        peso ( 4, 4)    =    0.1600571871/  0.1600571871

        maestro ( 1)    =    16
        maestro ( 2)    =    15
        maestro ( 3)    =    12
        maestro ( 4)    =    11
        maestro ( 5)    =     9
        maestro ( 6)    =     3
        maestro ( 7)    =     4
        maestro ( 8)    =    13
        maestro ( 9)    =     8
        maestro (10)    =     7
        maestro (11)    =    14
        maestro (12)    =    10
        maestro (13)    =     1
        maestro (14)    =     2
        maestro (15)    =     5
        maestro (16)    =     6
```

Polarity profile of the target group generated by *dispara* script.

Table A.9. Section (3) *polarity index method* program.

```
c       Author      Carlos Polanco
c       Date        January, 2013.
c       Program     vieneP2.f
c       email:      polanco@unam.mx
c
        do i= 1,4
           do j = 1,4
           matriz(i,j) = 0
           enddo
        enddo
        k      = 0
        n      = 40000
        total  = 0
        suma   = 0
        espe   = 0
c
c       ROUTINE TO GET THE SEQUENCE, AND DETERMINES THE
c       ABSOLUTE FREQUENCY DISTRIBUTION OF AMINO ACID IN
c       THE SEQUENCE
c
        read (1,200) (arreglo(i),i=1,n)
        do i= 1,(n−1)
           if (arreglo(i) .eq. 0) goto 100
           total = total + 1
           matriz((arreglo(i)),(arreglo(i+1))) =
     &     matriz((arreglo(i)),(arreglo(i+1))) + 1
        enddo
100     do i= 1,4
           do j = 1,4
              k = k + 1
              lineal(k) = matriz(i,j)+peso(i,j)/(total∗1.)
              linealA(k) = 0
           enddo
        enddo
        do i = 1, 16
           write (2,35) lineal(i), i
        enddo
        close (1)
        close (2)
        call system (”sort −nr valor.dat
        > valorordenado.dat”)
        call system (”rm valor.dat”)
        call system (”mv valorordenado.dat valor.dat”)
c
c       ROUTINE TO EVALUTE IF THE PROTEIN IS OR NOT
c       CANDIDATE
```

```
c
      open (3,file="valor.dat")
      do i= 1,16
         read (3,35,END=101) comodin, linealA(i)
      enddo
      call coincidencias(linealA ,maestro ,suma,espe)
      if (espe .ge. 1) then
        write (6,53)"Yes",(linealA(i),i=1,16),espe ,suma
      else
        write (6,53)"No ",(linealA(i),i=1,16),espe ,suma
      endif
      call system ("rm valor.dat")
      close(3)

 101  stop
      end
c
c     SUBROUTINES AND FUNCTIONS
c
      subroutine coincidencias(linealA ,maestro ,suma,espe)
      integer i,j, cocientes1 , cocientes2
      integer linealA(16),maestro(16)
      integer suma, espe
      espe = 1
c
c     ROUTINES FOR IDENTIFY THE COINCIDENCES
c
```

Routines to identify the polar profile of each protein by *polarity index method* program.

A.4. OUTPUT DATASETS

Polarity index method program reports the results in two files: (i) anterioranalisis.dat file (Table **A.10**), that shows (00) when that group has a low detection percentage compared to the protein group namegroup, or (1, x) where x states the number of iterations effected to reduce the percentage of that particular group *i.e.* an x different than the value zero means that the group is very similar to the namegroup, and this similarity is reduced; and (ii) the anteriortotales.txt file (Tables **A.11-A.13**) gives the percentage of similarity compared to namegroup, e.g. the method grades as 99% efficiency the namegroup, and 10% the scaaps group. If the similarity between groups needs to be reduced, it will be necessary to reduce the efficiency of the method (Sect. A.3).

Table A.10. Final status of *anterioranalisis.dat* file.

```
namegroup 0 0
bacteriasplus 0 0
bacteriasminus 0 0
scaaps 0 0
tuberculosis 1 5
h1n1 0 0
fungi 1 2
cancer 0 0
```

anterioranalisis.dat file produced by *huella3* script.

Table A.11. Final status of *anteriortotales.txt* file.

```
1 99
2 35
3 22
4 10
5 27
6 19
7 22
8  3
```

anteriortotales.txt file produced by *huella3* script.

Table A.12. Internal process of *dispara* script.

```
c       Author     Carlos Polanco
c       Date       January, 2013.
c       Program    generador.f
c       email:     polanco@unam.mx
c
        implicit none
        integer j,i, linealmaestro(16),linealotros(16)
        integer tope
        integer totalotros, totalmaestro, k,unavez
        integer proceso,inicio
        real    datomaes, datootro
 20     format (6x,”if (linealA(”,I2,”) .eq. ”,I2,”)
        espe = 0”)
        inicio = 0
        tope   = 0
        proceso = 0
        open (1,file=”temporal.net”)
        open (2,file=”decision.txt”)
        read (2,*, END=200) proceso, tope
        if (proceso .eq. 1) then
80         if (inicio .lt. tope) then
              read (1,*,END=200) datomaes, datootro, i,j
              write(6,20) i,j
              inicio = inicio + 1
              goto 80
           endif
        endif
200     close(1)
        close(2)
        stop
        end
```

generador.f program used by *dispara* script.

Table A.13. Internal process of *dispara* script.

```
c       Author     Carlos Polanco
c       Date       January, 2013.
c       Program    producetabla.f
c       email:     polanco@unam.mx
c
        implicit none
        integer arreglo(40000)
        integer matriz(4,4)
        real total, cifra
        integer n, mayor, i,j,k
        open (1,file=”archivoperfil.net”)
        open (2,file=”muestra.txt”)
        open (3,file=”muestracompleta.txt”)

 34     format (f14.10,1x,I2)
 40     format (f14.10)
 200    format (40000(I1))
 300    format (”        peso (”,I2,”,”,I2,”)      =    ”,
        f14.10,”/”,f14.10)
 400    format (”        maestro (”,I2,”)      =      ”,I2)
 500    format (” ”)

        n = 40000
        do i= 1,4
           do j = 1,4
           matriz(i,j) = 0
           enddo
        enddo
        total = 0
        k     = 0
        mayor = 0
50      read (1,200,END=100) (arreglo(i),i=1,n)
        do i= 1,(n−1)
           if (arreglo(i) .eq. 0) goto 50
           if (arreglo(i) .ne. 0) total = total +1
           matriz((arreglo(i)),(arreglo(i+1))) =
     &     matriz((arreglo(i)),(arreglo(i+1))) + 1
        enddo
        goto 50

100     do i= 1,4
           do j = 1,4
              k = k + 1
             write (2,40) matriz(i,j)/total
             write (3,34) matriz(i,j)/total ,k
           enddo
```

```
      enddo

      close (1)
      close (2)
      close (3)
      call system (”sort −r muestra.txt >
      muestraordenada.txt”)
      call system (”head −1 muestraordenada.txt >
      muestraordenada2.txt”)
      open (1,file=”muestra.txt”)
      open (2,file=”muestraordenada2.txt”)
      read (2,40) total
      write (6,500)
      do i = 1, 4
         do j = 1, 4
            read (1,40) cifra
            write (6,300) i, j, cifra, total
         enddo
      enddo
      write (6,500)
      write (6,500)
      call system (”sort −nr muestracompleta.txt>
      muestracompleta2.txt”)
      open (3,file=”muestracompleta2.txt”)
      do i = 1, 16
         read (3,34) cifra, k
         write (6,400) i, k
      enddo
      stop
      end
```

producetabla.f program used by *dispara* script.

Appendix B-Protein Databases

B.1. APD2 DATABASE

The Antimicrobial Peptides database (APD2) has more than 2400 peptides; it has a very simple and comprehensible search procedure. “APD2 database consists of a pipeline of search functions. You can search for peptide information using APD ID, peptide name, amino acid sequence, peptide motif, chemical modification, length, charge, hydrophobic content, PDB ID, 3D structure, methods for structural determination, peptide source organism, peptide family name, life kingdoms/domains (bacteria, archaea, protists, fungi, plants, animals), antimicrobial activities, synergistic effects, target microbes, molecular targets, mechanism of action, contributing authors, and year of publication [1].

B.2. AMYPDB DATABASE

AMYPdb is a database for amyloid precursor proteins, and to the large scale signature analysis of those proteins. The last update of the structure and the interface of AMYPdb was in 2008. The protein files are from UniProt [2].

B.3. UNIPROT DATABASE

The Universal Protein Resource (UniProt) is a computational resource for protein se-quence and annotation data. It includes all types of proteins and peptides, its search system requires the use of training manuals, and it is constantly updated. “UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR). Across the three institutes more than 100 people are involved through different tasks such as database curation, software development and support [3].

B.4. CPPSITE DATABASE

“CPPsite is a database of experimentally validated Cell Penetrating Peptides (10–30 amino acids). Importance of CPPsite: CPPs have tremendous therapeutic applications. These are widely used to promote intracellular uptake of conjugated cargos (nucleic acids, peptide nucleic acids, proteins, drugs, liposomes etc.) and thus play role to overcome the problem of poor delivery and low bioavailability of therapeutic molecules. CPP conjugated drugs when delivered *in vivo* have shown promising results with high efficacy. Many CPP–conjugated compounds are under clinical trials. CPPsite database provides comprehensive information on CPPs, which may be helpful to scientific community working in the area of peptide based drug discovery. What type of information it has: CPPsite database’s current version contains

Carlos Polanco

comprehensive information of 843 CPPs with multiple entries in terms of peptide sequence, source origin, localization, uptake efficiency, uptake mechanism, hydrophobicity, charge" [4].

B.5. SCAAP DATASET

This dataset is formed by 30 peptides verified as selective cationic amphipathic antibacterial peptides, and it includes their toxicity [5].

B.6. AMYLOIDOSIS DATASETS

This dataset includes Supplementary files divided in three groups containing fragments of proteins, natively unfolded proteins, natively folded proteins, and partially folded proteins [6].

B.7. CONSIDERATIONS

All programs and scripts necessary to run Polarity index method program can be requested at polanco@unam.mx.

REFERENCES

[1] Wang G, Li X, Wang Z. APD2: the updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res 2009; 37(Database issue): D933-7. [http://dx.doi.org/10.1093/nar/gkn823] [PMID: 18957441]

[2] Apweiler R, Bairoch A, Wu CH, *et al.* UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 2004; 32(Database issue): D115-9. [http://dx.doi.org/10.1093/nar/gkh131] [PMID: 14681372]

[3] Gautam A, Singh H, Tyagi A, *et al.* CPPsite: a curated database of cell penetrating peptides Database (Oxford) 2012. [http://dx.doi.org/10.1093/database/bas015]

[4] del Rio G, Castro-Obregon S, Rao R, Ellerby HM, Bredesen DE. APAP, a sequence-pattern recognition approach identifies substance P as a potential apoptotic peptide. FEBS Lett 2001; 494(3): 213-9. [http://dx.doi.org/10.1016/S0014-5793(01)02348-1] [PMID: 11311243]

[5] Pawlicki S, Le Béchec A, Delamarche C. AMYPdb: a database dedicated to amyloid precursor proteins. BMC Bioinformatics 2008; 9: 273. [http://dx.doi.org/10.1186/1471-2105-9-273] [PMID: 18544157]

[6] Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK. Comparing and combining predictors of mostly disordered proteins. Biochemistry 2005; 44(6): 1989-2000. [http://dx.doi.org/10.1021/bi047993o] [PMID: 15697224]

GLOSSARY

Å The angstrom (symbol: Å) is a unit of length. 1 angstrom = 1.0 × 10^{-10} meters, it is often used to express the sizes of molecules, atoms, and the lengths of chemical bonds [1].

3–D space: It is the set of points can thus be illustrated with a physical body, is called 3-dimensional Euclidean space. It is represented by the symbol R^3 [2].

Aleksandr Mikhailovich Lyapunov: (June 6, 1857 – November 3, 1918) was a Russian mathematician, mechanician and physicist. His surname is sometimes Romanized as Ljapunov, Liapunov, Liapounoff or Ljapunow. Lyapunov is known for his development of the stability theory of a dynamical system, as well as for his many contributions to mathematical physics and probability theory [3].

Algorithm: Is an effective method expressed as a finite list of well-defined instructions for calculating a function. Starting from an initial state and initial input (perhaps empty), the instructions describe a computation that, when executed, proceeds through a finite number of well-defined successive states, eventually producing "output" and terminating at a final ending state [4].

Andrey Andreyevich Markov: (June 14, 1856 – July 20, 1922) was a Russian mathematician. He is best known for his work on stochastic processes. A primary subject of his research later became known as Markov chains and Markov processes [5].

Bioinformatics: Is the application of computer technology to the management of biological information. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development [6].

Boiling point: The boiling point of a substance is the temperature at which the vapor pressure of the liquid equals the pressure surrounding the liquid and the liquid changes into a vapor [7].

Cal/mol: The calories per mole (symbol: cal/mol) is an unit of energy per amount of material. Energy is measured in calories, and the amount of material is measured in moles [8].

Catastrophe theory: In mathematics, catastrophe theory is a branch of bifurcation theory in the study of dynamical systems; it is also a particular special case of more general singularity theory in geometry. Bifurcation theory studies and classifies phenomena characterized by sudden shifts in behavior arising from small changes in circumstances, analysing how the qualitative nature of equation solutions depends on the parameters that appear in the equation.

Carlos Polanco

This may lead to sudden and dramatic changes, for example the unpredictable timing and magnitude of a landslide [9].

Collagen: Is the main structural protein of the various connective tissues in animals [10].

Electronegativity: Is a chemical property that describes the tendency of an atom or a functional group to attract electrons (or electron density) towards itself. An atom's electronegativity is affected by both its atomic number and the distance at which its valence electrons reside from the charged nucleus. The higher the associated electronegativity number, the more an element or compound attracts electrons towards it. Electronegativity cannot be directly measured and must be calculated from other atomic or molecular properties. Several methods of calculation have been proposed, and although there may be small differences in the numerical values of the electronegativity, all methods show the same periodic trends between elements [11].

Gram staining: Is an auxiliary technique used in microscopic techniques used to enhance the clarity of the microscopic image. Stains and dyes are widely used in the scientific field to highlight the structure of the biological specimens, cells, tissues, etc. This technique differentiates bacteria into two subgroups: Gram + (positive), and Gram– (negative) [12].

Heat capacity: Heat capacity, or thermal capacity, is the measurable physical quantity of heat energy required to change the temperature of an object or body by a given amount [13].

Heat of fusion: Is the total energy change, resulting from heating a given quantity of a substance to change its state from a solid to a liquid. The temperature at which this occurs is the melting point [14].

Heat of vaporization: Is the total energy change required to transform a given quantity of a substance from a liquid into a gas at a given pressure [15].

Linear differential equations: Are differential equations having differential equation solutions which can be added together to form other solutions. They can be ordinary or partial. The solutions to linear equations form a vector space (unlike non-linear differential equations) [16].

Lyapunov function: Are scalar functions that may be used to prove the stability of an equilibrium of an ODE. Named after the Russian mathematician Aleksandr Mikhailovich Lyapunov, Lyapunov functions are important to stability theory and control theory. A similar concept appears in the theory of general state space Markov Chains, usually under the name Foster-Lyapunov functions. For many classes of ODEs, the existence of Lyapunov functions is a necessary and sufficient condition for stability. Whereas there is no general technique for constructing Lyapunov functions for ODEs, in many specific cases, the construction of Lyapunov functions is known. For instance, quadratic functions suffice for systems with one state; the solution of a particular linear matrix inequality provides Lyapunov functions for linear systems; and conservation laws can often be used to construct Lyapunov functions for physical systems [17].

Markov chains: A Markov chain, named after Andrey Markov, is a mathematical system that undergoes transitions from one state to another on a state space. It is a random process usually characterized as memoryless: the next state depends only on the current state and not on the sequence of events that preceded it. This specific kind of ”memorylessness” is called the Markov property. Markov chains have many applications as statistical models of real-world processes [18].

Markov process: or Markoff process, named after the Russian mathematician Andrey Markov, is a stochastic process that satisfies the Markov property. A Markov process can be thought of as ”memoryless”: generally speaking, a process that satisfies the Markov property if one can make predictions for the future of the process based solely on its present state just as well as one could know the process’s full history *i.e.* conditional on the present state of the system, with its future and past being independent [19].

Melting point: Is the temperature at which it changes state from solid to liquid at atmospheric pressure. At the melting point the solid and liquid phase exist in equilibrium [20].

Ordinary differential equation: ODE is an equation containing a function of one independent variable and its derivatives. The term ”ordinary” is used in contrast with the term partial differential equation which may be with respect to more than one independent variable [21].

Proteomics: Is the large-scale study of proteins, particularly their structures and functions. Proteins are vital parts of living organisms, as they are the main components of the physiological metabolic pathways of cells [22].

Relative permittivity: It reflects the extent to which the electrostatic lines of flux]23] are concentrated in a material under given conditions.

Ren´e Frederic Thom: (September 2, 1923 – October 25, 2002) was a French mathematician. He made his reputation as a topologist, moving on to aspects of what would be called singularity theory; he became world-famous among the wider academic community and the educated general public for one aspect of this latter interest, his work as founder of catastrophe theory (later developed by Erik Christopher Zeeman) [9].

Stability theory: In mathematics, stability theory addresses the stability of solutions of differential equations and of trajectories of dynamical systems under small perturbations of initial conditions. The heat equation, for example, is a stable partial differential equation because small perturbations of initial data lead to small variations in temperature at a later time as a result of the maximum principle. One must specify the metric used to measure the perturbations when claiming a theorem is stable. In partial differ-ential equations one may measure the distances between functions using Lp norms or the sup norm, while in differential geometry one may measure the distance between spaces using the Gromov–Hausdorff distance [24].

Solubility: Is the property of a solid, liquid, or gaseous chemical substance called so-lute to dissolve in a solid, liquid, or gaseous solvent to form a homogeneous solution of the solute in the solvent [25].

Stochastic process: In probability theory, a stochastic process, or sometimes random process, is the counterpart to a deterministic process (or deterministic system). Instead of dealing with only one possible reality of how the process might evolve under time (as is the case, for example, for solutions of an ordinary differential equation), in a stochastic or random process there is some indeterminacy in its future evolution de-scribed by probability distributions. This means that even if the initial condition (or starting point) is known, there are many possibilities the process might go to, but some paths may be more probable and others less so [26].

Surface tension: Is a contractive tendency of the surface of a liquid that allows it to resist an external force [27].

REFERENCES

[1] Angstrom In Wikipedia, The Free Encyclopedia , [cited 2014 Oct 7]; Available from: http://en.wikipedia.org/w/index.php?title=Angstrom&oldid=627571610.

[2] 3–D space. In Wikipedia, The Free Encyclopedia. Available from: http://en.wikipedia.org/w/index.php?title=3D space&oldid=366953676

[3] Lyapunov AM. In Wikipedia, The Free Encyclopedia , [cited 2014 Oct 7]; Available from: http://en.wikipedia.org/w/index.php?title=AleksandrMikhailovichLyapunov&oldid=528671617.

[4] Algorithm In Wikipedia, The Free Encyclopedia , [cited 2014 Oct 7]; Available from: http://en.wikipedia.org/w/index.php?title=Algorithm&oldid=627455989.

[5] Markov AA. In Wikipedia, The Free Encyclopedia , [cited 2014 Oct 7]; Available from: http://en.wikipedia.org/w/index.php?title=AndreyAndreyevichMarkov&oldid=17324162.

[6] Bioinformatics In Wikipedia, The Free Encyclopedia , [cited 2014 Oct 6]; Available from: http://en.wikipedia.org/w/index.php?title=Bioinformatics&oldid=628420745.

[7] Boiling point In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 23]; Available from: http://en.wikipedia.org/w/index.php?title=Boilingpoint&oldid=626714530.

[8] Calorie In Wikipedia, The Free Encyclopedia , [cited 2014 Oct 3]; Available from: http://en.wikipedia.org/w/index.php?title=Calorie&oldid=628119137.

[9] Thom RF. Rene Frederic Thom In Wikipedia, The Free Encyclopedia , [cited 2014 Aug 10]; Available from: http://en.wikipedia.org/w/index.php?title=ReneFredericThom&oldid=17638884.

[10] Collagen In Wikipedia, The Free Encyclopedia , [cited 2014 Oct 7]; Available from: http://en.wikipedia.org/w/index.php?title=Collagen&oldid=628490937.

[11] Electronegativity In Wikipedia, The Free Encyclopedia , [cited 2014 Oct 7]; Available from: http://en.wikipedia.org/w/index.php?title=Electronegativity&oldid=628641682.

[12] Gram InWikipedia, The Free Encyclopedia , [cited 2014 Sep 9]; Available from: http://en.wikipedia.org/w/index.php?title=Gram&oldid=624803603.

[13] Heat capacity In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 28]; Available from: http://en.wikipedia.org/w/index.php?title=Heat capacity&oldid=627465184.

[14] Heat of fusion In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 9]; Available from: http://en.wikipedia.org/w/index.php?title=Heatoffusion&oldid=106179800.

[15] Enthalpy of vaporization In Wikipedia, The Free Encyclopedia , [cited 2014 Aug 27]; Available from: http://en.wikipedia.org/w/index.php?title=Enthalpyofvaporization&oldid=623080840.

[16] Linear differential equations In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 9]; Available from: http://en.wikipedia.org/w/index.php?title=Lineardifferentialequations&oldid=43055150.

[17] Lyapunov function In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 9]; Available from: http://en.wikipedia.org/w/index.php?title=Lyapunovfunction&oldid=616841899.

[18] Markov chains In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 9]; Available from: http://en.wikipedia.org/w/index.php?title=Markov chains&oldid=16508901.

[19] Markov process In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 9]; Available from: http://en.wikipedia.org/w/index.php?title=Markovprocess&oldid=627673571.

[20] Melting point In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 9]; Available from: http://en.wikipedia.org/w/index.php?title=Meltingpoint&oldid=628071783.

[21] Ordinary differential equation In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 11]; Available from: http://en.wikipedia.org/w/index.php?title=Ordinarydifferentialequation&oldid=625070330.

[22] Proteomics In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 11]; Available from: http://en.wikipedia.org/w/index.php?title=Proteomics&oldid=628508741.

[23] Relative permittivity In Wikipedia, The Free Encyclopedia , [cited 2014 Aug 10]; Available from: http://en.wikipedia.org/w/index.php?title=Relativepermittivity&oldid=622047565.

[24] Stability theory In Wikipedia, The Free Encyclopedia , [cited 2014 Aug 1]; Available from: http://en.wikipedia.org/w/index.php?title=Stabilitytheory&oldid=619406379.

[25] Solubility In Wikipedia, The Free Encyclopedia , [cited 2014 Oct 7]; Available from: http://en.wikipedia.org/w/index.php?title=Solubility&oldid=627348092.

[26] Stochastic process In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 2]; Available from: http://en.wikipedia.org/w/index.php?title=Stochasticprocess&oldid=623885621.

[27] Surface tension In Wikipedia, The Free Encyclopedia , [cited 2014 Sep 25]; Available from: http://en.wikipedia.org/w/index.php?title=Surfacetension&oldid=627059380.

SUBJECT INDEX