SMGroup

Bioinformatics

Alberto Oliveira[1], Leandro Benevides[1], Diego Mariano[1], Edgar Aguiar[1], Thiago Sousa[1], Artur Silva[2] and Vasco Azevedo[1]*

[1]Universidade Federal de Minas Gerais/Instituto de Ciências Biológicas.

[2]Universidade Federal do Pará/ Instituto de Ciências Biológicas.

***Corresponding author:** Vasco Azevedo, Universidade Federal de Minas Gerais/Instituto de Ciências Biológicas, Minas Gerais, Brazil, Email: vasco@icb.ufmg.br

Published Date: February 15, 2015

How the study of nucleotide sequences, amino acids, the function of these biomolecules and the three-dimensional structures level, influence in biotechnological processes?

How may Bioinformatics help?

INTRODUCTION

In recent times we have seen a range of innovative studies that has revolutionized the biological sciences as we had never seen before. Studies of complete genome sequencing of pathogenic organisms has helped the understanding of the virulence mechanisms, the production of new drugs and the development of strategies to contain outbreaks [1-3]; studies of RNA transcription level and regulation of networks has helped to understand different types of cancer [4]; protein folding studies have been correlated with mutations in the DNA, and helped researchers to have new insights into the functioning of diseases previously poorly understood [5]; in vivo experiments with the aid of new technologies have helped to understand the damage to the spinal column and allowed to paralyzed limbs can be moved again [6].

These studies were made possible due to the evolution of information technology (IT), evidenced in the improving of software (algorithms and implementation techniques of parallel processes), and increasing processing power and storage information (improvement of hardware), which in recent times has allowed the storage and processing of huge amounts of data, the "big data" [7].

With the race to sequencing and mapping of the human genome, new and creative DNA sequencing techniques have contributed to the appearance of robust sequencing platforms, known as Next-Generation Sequencers (NGS),which allowed the large-scale sequencing, with both cost and time reduced. All these factors led to the rise of a new field of study, linking the biological sciences to computer science, called bioinformatics [8].

Bioinformatics is an area of research between biology, mainly molecular biology and computer science. It also covers other areas such as physics, chemistry, information science, engineering, mathematics and statistics.

One of the bioinformatics fundamentals is the modeling of a particular biological problem in order to treat it computationally. Thus, beyond the terms *in vivo* and *in vitro*, a new expression arises: *in silico*, representing experiments by computational simulation.

BIOINFORMATICS SOFTWARE

Computational experiments are performed using software, which are built with programming languages following a particular algorithm.

Software development is an important tool for bioinformatics. Knowing a programming language can be a differentiator for the professionals in this area. Recently, several groups of developers have provided libraries to assist the development of code for bioinformatics in various programming languages. Libraries such as BioPerl, BioJava, BioJS, BioRuby, BioPHP and BioPython have facilitated the development process of new biological software. However, knowledge of software programming is not essential the in this area, after all there are many biological software already developed for various platforms, especially for Linux operating systems [8]. Thus, the field of use of operating systems and specific software to your search area become of fundamental value for bioinformaticians.

Software: logical part of a computer; set of instructions to be interpreted and executed by a processor; the operating system and applications.
Algorithm: computational process that follows a set of actions to perform a specific task in a finite time interval; can be understood as the step-by-step to solving a problem computationally treated.
Hardware: physical part of computers; processors, boards and all electronic components in and out.
Operating System: range of programs responsible for linking the software and the hardware, ensuring operability of the system and allowing the user to get access to computer resources. Currently the most widely used operating systems are Windows, OS X, Linux, Android and IOS, the last two being exclusive both to smartphones and tablets.
Cloud computing: the use of the services by Internet through virtualized resources on remote servers [24].
Web browser: computer program that allows the user to access websites. Examples of web browsers: Firefox, Chrome and Internet Explorer.

However, with the recent strong adoption of cloud computing by the developer community and the improvement in the quality of internet connection, users could to work on collaborative projects easily through the World Wide Web. This phenomenon positively affected the manner in which software for bioinformatics are designed. The adoption of the Web has brought greater usability and interoperability for applications of bioinformatics, allowing software to be easily accessed on any operating system through a web browser.

The development of Web tools for bioinformatics has become popular due to low learning curve of Web programming languages (such as PHP, Python and JavaScript), the fast and easy way to access the software without the need to install through web browsers, the ease in performing cooperative work in parallel at distance, and especially the adoption of user-friendly interfaces. Thus, several research groups started to provide software implementation services for the World Wide Web, also allowing easy access to their databases [8]. Therefore, a large amount of information stored in databases could be accessed over the Web through electronic sites.

Note that the concept of database is very important for bioinformatics. But what is a database and what is its importance in representing biological problems?

WHAT IS DATABASE?

Databases are **data** sets organized in order to present comprehensible **information**. Thus, biological databases are data sets derived from biological experiments organized so that you can extract useful information through data mining techniques.

One of the first records of protein storage in databases occurred in the 60s, by Margaret Dayhoff et al. from the National Biomedical Research Foundation (NBRF). They have proposed the use of a substitution matrix, PAM (Point mutations accepted), to calculate the conservation of protein sequence and the mutation probability of a given amino acid. The methods proposed by Dayhoff et al. used a character (letter) to represent each amino acid, allowing a reduction in stored data to describe amino acid sequences [9]. However, biological databases are not just used for representation of proteins (proteomics field), but also in the fields of genomics (databases of genomic sequences) and transcriptomics (transcribed databases).

A biological database example is the PDB (Protein Data Bank). The PDB (http://www.wwpdb.org/) stores protein structural information, such as the spatial coordinates of atoms. There are databases that relate ontologies of information between genes and proteins such as the Gene Ontology (www.geneontology.org/). Another type of database is the KEGG (http://www.genome.jp/kegg/) that stores the biochemical interactions and metabolic pathways, in addition there is another database with information of the metabolic pathways that were cured (manually reviewed and corrected) the MetaCyc (http://www.metacyc.org/). It is possible to cite databases of networks RNA transcription regulation such as CMRegNet (http://lgcm.icb.ufmg.br/cmregnet).

A good computing practice is the management of databases through DBMS (Database Management Systems) that, in general, allow queries and data modification using SQL commands (Structured Query Language). However, many biological databases chose to use simple storage systems, such as text files, in order to facilitate the manipulation and analysis of these data using word processing tools written in programming languages. One example is the sequences databases.

Nucleic acid sequence databases are stored in different formats by three major institutes: NCBI (National Center for Biotechnology Information of the US National Library of Medicine), DNA Data Bank of Japan (National Institute of Genetics, Japan) and EMBL Data Library (European Bioinformatics Institute, UK). Another example is the UniProtKB, knowledgebase of proteins from the UniProt (Universal Protein Resource), that storing the known protein sequences. The UniProtKB is the union of two protein databases: Swiss-Prot and TrEMBL, and can be accessed through the Website UniProt (http://www.uniprot.org).

The patterns sequences storing were developed with the aim of speeding up information retrieval methods. Thus, storage patterns are directly related to the search algorithms of information, which will be presented below, and represent an important area of bioinformatics: sequence analyzes.

SEQUENCE ALIGNMENT

Sequence alignment is an important step toward structural and functional analysis of sequences. This is a method to compare two (pair-wise alignment) or more (multiple sequence alignment) sequences, searching for common character patterns and establishing residue–residue correspondence among related sequences.

Input Files (FASTA)

The main input file currently used in most sequence analysis is the fasta format [10]. This is a simple format broadly used when dealing nucleotides and amino acids sequences. The main characteristic which compose this file are the **header** and the **sequence**. Below is showed a example of this file.

>Header

S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S
E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U
E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C
E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E
Q U E N C E S E Q U E N C E S E Q U E N C E

>Streptococcus agalactiae - dnaA Chromosomal replication initiator protein DnaA forward

A T G G T A C A A T A T A A C A A T A A T T A T C C A C A A G A C A A T A A G G A A G A A G C T
A T G A C G G A A A A C G A A C A A C T A T T T T G G A A T A G A G T A C T A G A G C T A T C T C G
T T T C A A A T A G C A C C A G C A G C T T A T G A A T T T T T T G T T C T A G A G G C T A G A C T C
C T C A A A A T T G A A C A T C A A A C T G C A G T T A T T A C T T T A G A T A A C A T T G A A A T
G A A A A A G C T A T T C T G G G A A C A A A A T T T G G G G C C T G T T A T C C G T C T A G T C A
C G C G C T T T A G T T G G G G A C T C C C A G T A A A T A T C A C A C C A C C

This file can be represented by many others characters, as:

- **"-"** represent gaps, used when is made alignments in search of evolutionary patterns;
- **"."** can indicate other element in sequence, like heteroatoms in protein structures, for example water or ions;
- **"N"** Any nucleotide or amino acid;
- "\\" can indicates the end of the alignment, but this is not a common symbol;
- **"*"** indicate the translation stop;
- **"Upper and lower"** can be used to discriminate strong and weakly conserved residues respectively.

Nucleotides sequences can be represented by letter codes.

Table 1: Nucleotide letter code [10].

Nucleic Acid Symbol	Meaning
A	Adenine
T	**T**hymine
G	Guanine
C	Cytosine
U	Uracil
R	G or A (pu**R**ine)
Y	C, T or U (pYrimidine)
K	G, T or U (**K**etone)
M	A C (aMino groups)
S	G or C (**S**trong interaction)
W	A, T or U (Weak interaction)
B	not A (i.e. C, G, T or U) - **B** comes after A
D	not C (i.e. A, G, T or U) - D comes after C
H	not G (i.e., A, C, T or U) - **H** comes after G
V	H comes after G - V comes after U

The amino acid code can be divided into 24 characters, and 3 special codes.

Table 2: Amino acid letter code [10].

Amino Acid Symbol	Meaning
Z	Glutamic acid (E) or Glutamine (Q)
Y	Tyrosine
X	any
W	Tryptophan
V	Valine
U	Selenocysteine
T	Threonine
S	Serine
R	Arginine
Q	Glutamine
P	Proline
O	Pyrrolysine
N	Asparagine
M	Methionine
L	Leucine
K	Lysine
J	Leucine (L) or Isoleucine (I)
I	Isoleucine
H	Histidine
G	Glycine
F	Phenylalanine
E	Glutamic acid
D	Aspartic acid
C	Cysteine
B	Aspartic acid (D) or Asparagine (N)
A	Alanine
-	gap of indeterminate length
*	translation stop

There is no standard file extension for a text file containing sequences in FASTA format. The table below shows each extension and its respective meaning.

Extension	Meaning	Notes
fasta (.fas)	generic fasta	Any generic fasta file. Other extensions can be fa, seq, fsa
fna	fasta nucleic acid	Used to generically specify nucleic acids.
ffn	FASTA nucleotide coding regions	Contains coding regions for a genome.
faa	fasta amino acid	Contains amino acids. A multiple protein fasta file can have the more specific extension mpfa.
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA.

Local and Global Alignment

The sequence alignment is a tool used in bioinformatics with the objective to identify regions of similarity or identity a probable functional, structural or evolutionary relationships between the sequences. There are two ways to do this analysis, using local or global alignments.

The basic difference between local and global alignments is that in a local alignment, a query is checked to match to a substring (a portion of a sequence) of the subject (reference sequence), whereas in global alignment an end to end alignment with the subject is performed [11].

Local Alignment

5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'

||||||||||||||||||||||||

5' TACTCACGGATGAGGTACTTTAGAGGC 3'

Global Alignment

5' ACTACTAGACTACTTACGGGTCAGTTACTTTAGAGGCTTACAACCA 3'

||||||||||||||||||||||||||||||||||||||

5' ACTACTAGACT----ACGGGTC--TTACTTTAGAGGCTAACAACCA 3'

Suppose you are searching some information about your sequence in a database. This sequence has, approximately, 300 bp. The database in which you are doing this search has a large reference, around 2000 bp. When a global alignment is done, result demonstrates that your sequence is probably involved in alternative splice - specific event from eucariotic cells. However, when is done a local alignment, i.e. using just some fragments from your sequence, the result is more precision, showing that your sequence align in a specific region featuring one specific organism.

Multiple Sequence Alignment (MSA)

Multiple sequence alignment may be considered as an extension of pairwise alignment, which allow the alignment to have multiple related sequences in order to obtain the best matching of the sequences, and thus reveals more biological information. It allows you to

search and identify patterns in sequence in the whole sequence family. MSA presents several computational challenges. First, obtain an optimal alignment of several sequences that includes matches, mismatches, and gaps, and also that take into account the degree of variation in all of the sequences at the same time. Due to that, alternative methods had to be developed to speed up the calculations for multiple sequence alignment, including: (1) progressive alignment of the sequences, (2) iterative alignment, (3) alignments based on locally conserved patterns found in the same order in the sequences, and (4) use of statistical methods and probabilistic models of the sequences [12]. The automated alignment often contains misaligned regions that should be edit manually. It is need to check the alignment to correct obvious alignment errors and to remove sections of dubious quality. Biologically significant results can be improved by removing or inserting gaps. Corrections on an alignment can be made by manual editing, which requires practical experience [12]. Manual editing may be performed on the amino acid level for protein coding sequences. In this way, information about protein domain structure, secondary structure, or amino acid physicochemical properties can be taken into consideration [13].

ADVANCED BIOINFORMATICS

In this section we will address the theme "Advanced Bioinformatics" as a specific part, focused in evolutionary studies and structural analyses from biomolecules.

Phylogenetic and Molecular Evolution

The phylogeny arose from the human need to organize and classify the world around them in order to facilitate understanding and communication. Different systems were proposed for classifying organisms until XIX century, when Charles Darwin developed his theory of evolution, inspired in phenotypic variation of finches in the Galapagos Islands. Based on this theory, the classification of organisms has become not only natural, but also to present a prerequisite for common ancestry. According to this thinking, the organisms derived from each other, since the emergence of life on earth. Darwin represented this pattern through a branching scheme, where the branches represent the time between the ancestral and the new organism, and the nodes represents the organisms themselves. Later, this would be the first phylogenetic tree used to represent evolutionary processes [14]. Nowadays, the evolutionary analysis of ancestry is based on molecular data thanks to the advent of sequencing methods. Similarities and divergence among related biological sequences revealed by sequence alignment often have to be rationalized and visualized in the context of phylogenetic trees. Thus, molecular phylogenetics is a fundamental aspect of bioinformatics [9].

Phylogenetic trees

The evolutionary relationships among genes and organisms can be illustrated using a two-dimensional graph, called phylogeny or phylogenetic tree, which is comparable to a genealogy and shows which genes or organisms are most closely related. The phylogenetic trees receive terms

referring to the various parts of real trees (i.e. root, branch, nodes, and leaf). Terminal (external) nodes or leaves represents the existing taxa and are frequently called operational taxonomic units (OTUs). The trees can be drawn in cladogram or a phylogram. In a phylogram, the branch lengths represent the amount of evolutionary divergence. These kinds of trees are scaled and show the evolutionary relationships and information about the relative divergence time of the branches. In a cladogram, however, the terminal nodes lines up in a row. Their branch lengths are not proportional to the number of evolutionary changes and thus have no phylogenetic meaning. Furthermore, the phylogenetic trees can be rooted or unrooted. The main difference between them is that in the rooted phylogenetic trees there is one outgroup that represents the ancestor of all OTUs and is possible to infer the direction of evolutionary process. The unrooted phylogenetic trees only show the positions of the taxa and their relative relationships [15].

Methods for phylogenies inference

The main objective of molecular phylogenetics is to correctly reconstruct the evolutionary history based on the observed sequence divergence between organisms. The first step to constructing phylogenetic trees is alignment of sequences under study. After the alignment various methods that can be used to estimate the phylogeny of the sequences studied has been proposed [14]. There is three main methods to find the evolutionary trees: neighbor-joining (NJ), maximum parsimony (MP) and maximum likelihood (ML).

Neighbor-joining

Neighbor-joining is a distance method that employs the number of changes between each pair in a group of sequences to finding pairs of neighbors and thus produces a phylogenetic tree of this group. The NJ algorithm starts with a totally unresolved tree (star topology). The algorithm sequentially, based on a matrix of distances between all pairs of sequences, identifies the pair that presents the shortest distance, links this pair by a node (representing the common ancestor of the pair of sequences) and incorporates into the tree. The distances of each pair of sequence are recalculated for the new node, as well as the distances of all other sequences are recalculated for the new node. This process is then repeated replacing the pair of neighbors united by the new node and using the distances calculated in the previous step [12].

Maximum parsimony

The maximum parsimony is a qualitative method that assumes that within a range of phylogeny, that The maximum parsimony is a qualitative method that assumes that within a range of phylogeny, that phylogeny presents the lower number of evolutionary events (substitutions) should be the most probable to explain the alignment data. For each aligned position, the number of evolutionary changes to produce the observed sequence changes is calculated and the phylogenetic trees that require the smallest number are identified. This analysis continues for every position in the sequence alignment until the tree that requires the minimum number of changes is identified. Although fast, the parsimony methods fail to estimate the evolutionary relationship between a large number of sequences or sequences with a large amount of variation [14].

Maximum likelihood

Maximum likelihood (ML) is similar to the MP method and combines the alignment information to a statistical model able to better handle the probability of change of a nucleotide to another. Based on a certain evolutionary model, and assuming that each alignment of the site evolves so independent of the others, the likelihood for all possible nucleotide (or amino acid) states in the ancestral (internal) nodes is calculated. Because these likelihoods are very small numbers, their logarithms are usually added to give the logarithm likelihood of each tree. The most likely tree given the data is then identified. The main disadvantage of maximum likelihood methods is that they are computationally intense. However, with faster computers, the maximum likelihood method is seeing wider use and is being used for more complex models of evolution [14].

STRUCTURAL BIOINFORMATICS

The structural bioinformatics is summed up like the study of three-dimensional (3D) structures from biomolecules. In General, this area focuses in understanding the function, interactions and the molecular organization of proteins and some RNAs. The bioinformatics is able to analyze these molecules using a large arsenal of tool and thus understand the prediction of three-dimensional structures, i.e. proteins, that has been characterized with great impact in practical therapeutic applications and biotechnology.

Molecular Modeling

Understanding the physicochemical characteristics of the proteins that interact (for example in a network) and how an interaction is established at the molecular level can provide a biological understanding of the cells of organisms. The bioinformatics, also known by *In silico* methods, can then be used to investigate the interactions and predict how these molecules might interact at the atomic level. Some computational methods are wieldy to create these structures, referred to as template-based modeling, that includes both the **threading** techniques that return a full 3D description for the target and the **comparative** modeling [16].

Comparative

Currently, much of the three-dimensional structures are obtained by comparative modeling methods, which follow the Protocol held in the study of Sali (1997). This method is based on the use of homologous proteins for the prediction structures of sequences, thus the structure models are constructed from the residuals of the structure template that are aligned to the target sequence in the sequence comparison. Therefore, the quality of this alignment thus is critical for the accuracy achievable. The protocol (figure 1) pointed out four fundamental steps for the creation of the models: (i) identification of homologous structures that could be used as template for modeling; (ii) alignment of the target sequence with the template sequences; (iii) construction of the model for the target based on the information of the alignment; and (iv) model validation [17].

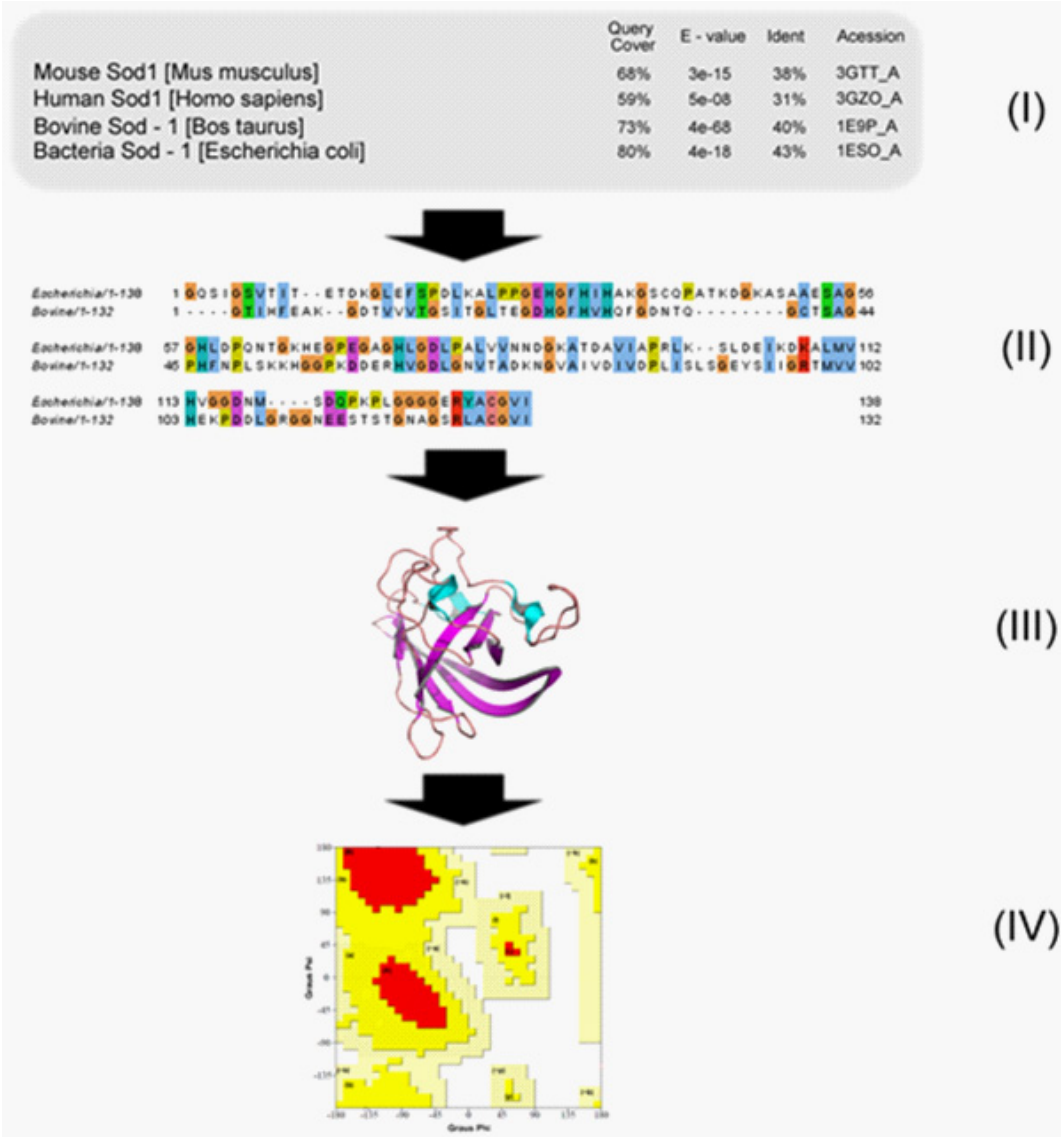

Figure 1: The comparative molecular modeling protocol scheme.

The important of the template-based model comes from template selection and sequence-template alignment, in addition to the difference between the structure and sequence template. It is essential the sequences have good values of identity, wherein it is expected that minimum values are around 30%. The step IV, about the validation, will be explained in the next subtopic.

Threading

When sequence identity is below 30%, it is complex to recognize the better template and generate accurate sequence-template alignments, so the resultant models have a wide range of accuracies. In general, proteins with low values of identity are said to be non-homologous with known proteins.

Considering that currently there are several proteins without experimental structures, even a rapid improvement in the accuracy of obtaining the model can have a significant impact on the prediction of large-scale structure and its applications. Given this difficulty, the protein threading method (Figure 2), also known as fold recognition, is a method of protein modeling that analyzes the folding of proteins with others thus is able to create models, i.e. this method uses the secondary structure (or little fragments of your target sequence) searches the same region in database of proteins for the prediction models [18].

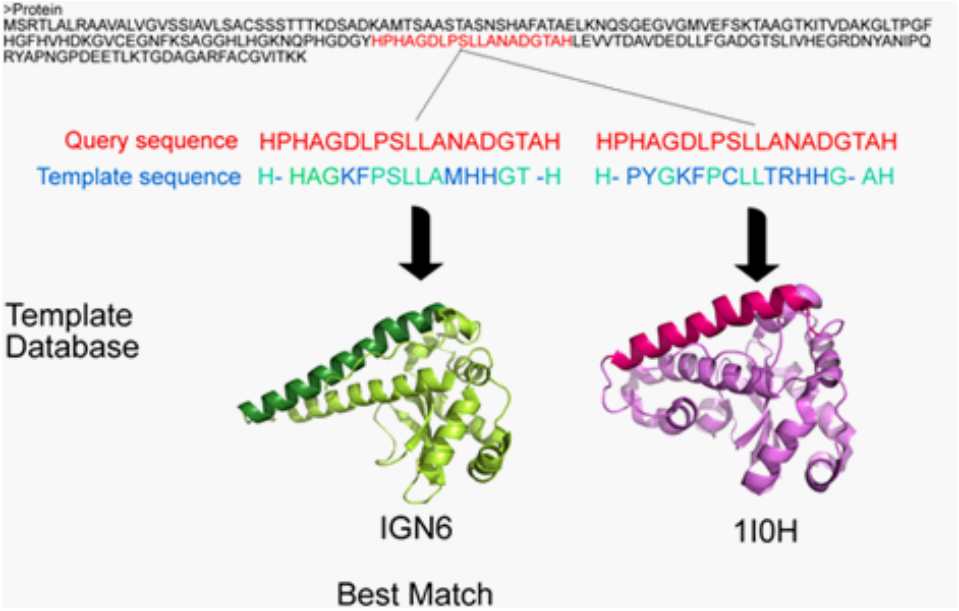

Figure 2: Modeling protocol scheme for threading

Ab Initio

Prediction of the spatial conformation of a protein from its primary structure is the one of the most important open problems in molecular biology, i.e. from its sequence of amino acids. Therefore, when no suitable structure templates can be found, Ab Initio methods can be used to predict the protein structure from the sequence information only. This is essential for a complete solution to the protein structure prediction problem; it can also help us understand the physicochemical principle of how proteins fold in nature. Currently, the accuracy of ab initio modeling (Figure 3) is low and the success is limited to small proteins (<100 residues) [19].

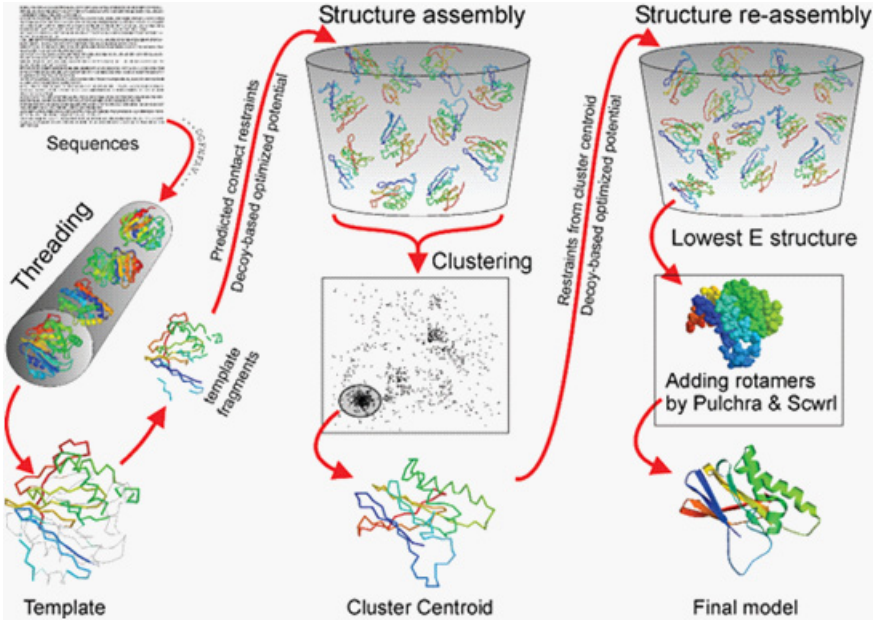

Figure 3: Here is a scheme from I-TASSER server. An internet service for protein structure and function predictions. It allows academic users to automatically generate high-quality predictions of 3D structure and biological function of protein molecules from their amino acid sequences [20].

Validations of Tree-Dimensional Models

Suppose we have two alternative conformations of some any protein, and we want analyze how much similar they are. For example, a model structure from molecular modeling, made by comparative modeling methods, need to have the conformation validated. This validation can be made using a distance matrix or known by "RMSD", in which the term mean Root-Mean-Square-Deviation [21]. One protein is typically represented by its virtual Cα atom chain of n residues or points. For a quantitative single-number measure of structural similarity between structures A and B:

$$D^2_{dis}(A,B) = (n(n-1)/2)^{-1} (d_{aij} - d_{bij})^2$$

Where *daij* and *dbij* are the correspondence distances between the ith and jth atoms or the "coordinate RMSD" after optimal rigid body superposition. Currently, many scientists made use of this method to align two or more protein structure to validate your spatial conformation. Thus, the root mean squared errors (deviations) function is defined as follows:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \delta_i^2}$$

Where δ is the distance between N pairs of equivalent atoms (usually $C\alpha$ and sometimes $C,N,O,C\beta$).

Other validation method is Ramachandran plot [22,23] or φ, ψ-plot that has remained nearly unchanged in the ensuing fifty years and continues to be an integral tool for protein structure research. His analysis is focused in the secondary structure of a protein that can be described by the torsion angles Φ (Phi) and Ψ (Psi) on which are repeated for each amino acid along the polypeptide chain. The allowed values of these angles, or acceptable to three-dimensional structures of proteins, are demonstrated at the Ramachandran plot (figure 4). This graph is based on analysis of 118 structures with a resolution of 2 angstroms (Å). To consider it a good model, it is expected that more than 90% of the amino acid residues are in the most favorable regions.

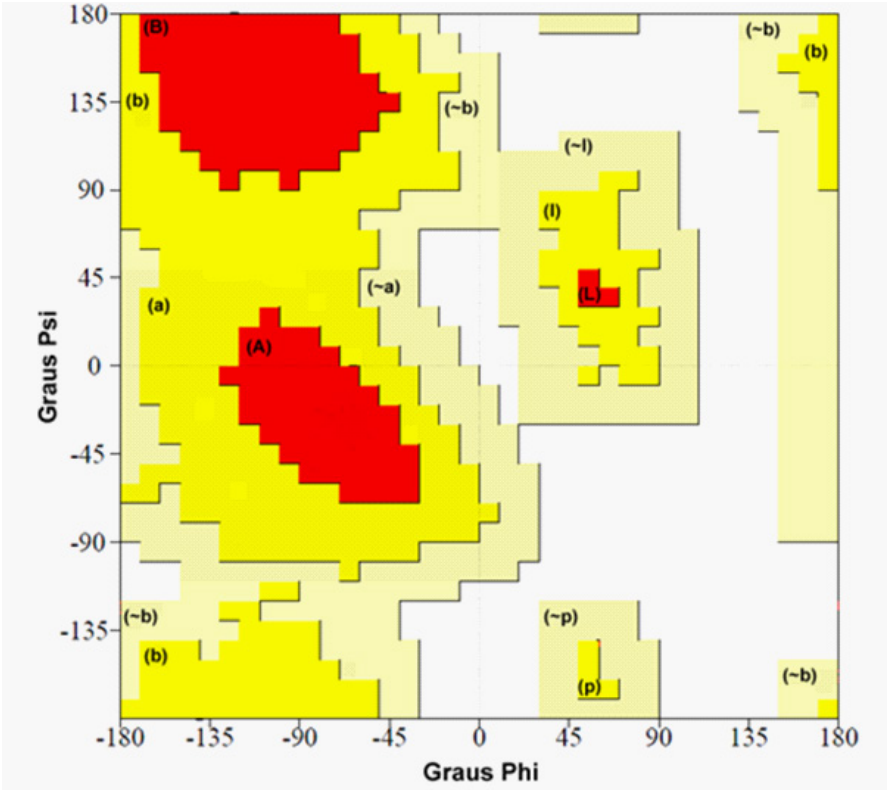

Figure 4: Ramachandran plot representation. The allowed regions are shown for amino acids: (A) more favorable regions of alpha *helix* (α-helix); (a) α-helix regions additionally permitted; (~a) α-helix regions generously allowed propellers; (B) β-sheet regions more favorable; (b) β-sheet regions further allowed; (~b) β-sheet generously allowed; (L) α-helix regions of left hand, more favorable; (I) α-helix regions of left hand propeller additionally allowed; (~i) α-helix regions of left hand propellers, generously allowed; (p) glycine regions additionally allowed; (~p) glycine regions generously allowed.

Box: Applications of bioinformatics in Biotechnology

The virtual screening (VS) techniques, through the use of computational methods, also known by *in silico* term, can help on search of organic compounds as promising therapeutic target ligands of interest, whether receptor agonists or antagonists, or enzyme inhibitors. The VS comprises two principal approach: The target structure-based virtual screening (I) and the screening ligand-based (II).

(I) This method considers the three-dimensional structure (3D) the therapeutic target, using the docking calculations as main strategy for selection of potential ligands with chemical characteristics, electronics and structural that promote interactions with the ligand site of the molecular target.

(II) Since the strategies ligand-based using organic molecules with biological activity known, working as molds for the screening in databases of new chemical entities with some level of similarity, sharing with these molds the same biological activity.

Thus, the main objective of a virtual screening is to identify the compounds of a library that are most likely that bind to the biological target of interest.

These targets may be parasites molecules or specific diseases. Since the ligands may be any small molecule found in nature, for example in plants. Thus, in environments with rich biodiversity vegetative can be focused to large production plants collections from biotechnological methods. Biotechnology is inserted in this context in order to study such plants with the aim to extract, purify and identify chemicals (small molecules) by physical-chemical analyzes with bioactive properties from them. Finally, this small molecules can be used in VS to search candidates therapeutic target against diseases or others biological problems.

BIOINFORMATICS LIMITS

Bioinformatics has been highlighting for some years as a powerful tool capable of providing a better and faster understanding of biological problems at the molecular level. However, it is limited by the frontiers of scientific computing, which is unable to show the entire biological setting complex.

To understand the advances in computational field and how they are directly related to the rise of bioinformatics, we can cite Moore's law. In 1965, Gordon Earl Moore, co-founder of Intel, stipulated that the processing power of a microchip would double every eighteen months, while in the same period, the price would drop by half. This prediction has had as a marketing strategy, but stimulated competition among technology companies, and in parallel, also served as a stimulus for the evolution of the computing power of the hardware, which in turn, allowed the emergence of bioinformatics. . However, it is important to understand the fine line between the data processing capacity and obtaining data by biological experiments, being one dependent on the other. Thus, to obtain good results *in silico*, it is necessary that biological experiments have produced good data.

An example of limitation found in bioinformatics is the prediction of protein structures. For these biomolecules, tend to lose their native three-dimensional structure easily to leave their ideal storage conditions. Thus, a structural modeling is intended to ensure the native conformation based on their primary amino acid sequence. First, think to try all possible conformations to get in the native structure of the protein, right? No, it is approach cannot be applied and this is exactly what Cyrus Levinhal affirmed in 1969, where states that if a protein was folding sequentially trying all possible conformations, it would take a period of time equivalent to the age of the universe, and this is at millisecond intervals under natural conditions. Based on this observation, the proteins shall directed by folding and orderly process. Currently, known to that the information necessary for the protein folding is contained in the amino acid sequence itself. Moreover, that

protein folding go through several alternative configurations with intermediate power stages, arriving in a lower level of state that would be the native state, like it was a map funnel-shaped. However, more research are still necessary to fully realize these *in silico* experiments. In short, it is still not possible to predict protein folding altogether, then we come in the next issue, we still lack computing power to simulate these conformations or are using the wrong approach?

The protein interactions are fundamental to almost all maintenance processes of integrity of cell. For example, when a structural analysis for protein is made using in silico methods the environment signalization and the biological environment are missing. Actually, the bioinformatics is not able to reproduce the same biological event. Due this is not possible get a real confirmation about the all structure analyses results.

Another big challenge of bioinformatics is the sequencing field and genome assembly. Sequencing platforms have gone through many evolutions in the last decades have enabled complete sequencing of genomes, reducing the time and expense thereof. However, the genome assembly process is directly impacted by sequencing, and this has computational cost determined by the complexity of the genome sequenced (Genome size and number of repetitive regions), the performance and quality of the sequencing platform used. However, the assembly of complex genomes is possible, but still requires a long time and has an extremely high cost.

The challenges of bioinformatics are still many and only time will tell if it will exceed its limits.

References

1. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011; 364: 730-739.
2. Lewis T, Loman NJ, Bingle L, Jumaa P, Weinstock GM. High-throughput whole-genome sequencing to dissect the epidemiology of Acinetobacter baumannii isolates from a hospital outbreak. J Hosp Infect. 2010; 75: 37-41.
3. Alexander Mellmann, Dag Harmsen, Craig A Cummings, Emily B Zentz, Shana R Leopold, et al. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS ONE. 2011.
4. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10: 57-63.
5. Roth DM, Hutt DM, Tong J, Bouchecareilh M, Wang N. Modulation of the maladaptive stress response to manage diseases of protein folding. PLoS Biol. 2014; 12: e1001998.
6. Shanechi MM, Hu RC, Williams ZM. A cortical-spinal prosthesis for targeted limb movement in paralysed primate avatars. Nat Commun. 2014; 5: 3237.
7. FRANKS Bill. Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics. New Jersey: John Wiley & Sons, Inc. 2012.
8. Arthur M Lesk. Introdução à Bioinformática (Tradução Ardala Elisa Breda Andrade, et al.). Oxford: Oxford University Press. 2008.
9. David W Mount. Bioinformatics: sequence and genome analysis. New York: Cold Spring Harbor Laboratory Press. 2001.
10. Tao Tao. Single Letter Codes for Nucleotides. NCBI Learning Center. National Center for Biotechnology Information. Retrieved 2012-03-15.
11. Polyanovsky VO, Roytberg MA, Tumanyan VG. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. Algorithms Mol Biol. 2011; 6: 25.
12. Jin Xiong. Essential Bioinformatics. New York: Cambridge University Press. 2006.
13. Hugo Verli. Bioinformática da Biologia à flexibilidade molecular. Porto Alegre: Bioinfo UFRGS. 2014.

14. Sergio Matioli, Flora Fernandes. Biologia Molecular e Evolução. 2nd edn. Brazil: Holos press. 2012.

15. Philippe Lemey, Marco Salemi, Anne-Mieke Vandamme. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. Cambrigde: Cambridge University Press. 2009.

16. Fiser A. Template-based protein structure modeling. Methods Mol Biol. 2010; 673: 73-94.

17. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D. Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci. 2007; Chapter 2: Unit 2.

18. Peng J, Xu J. Low-homology protein threading. Bioinformatics. 2010; 26: i294-300.

19. Jooyoung Lee, Sitao Wu, Yang Zhang. Ab Initio Protein Structure Prediction. In: Daniel John Rigden, editor. Houten: Springer Netherlands. 2009; 3-25.

20. Yang J, Yan R, Roy A, Xu D, Poisson J. The I-TASSER Suite: protein structure and function prediction. Nat Methods. 2014; 12: 7-8.

21. Maiorov VN, Crippen GM. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J Mol Biol. 1994; 235: 625-634.

22. ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963; 7: 95-99.

23. Hovmöller S, Zhou T, Ohlson T. Conformations of amino acids in proteins. Acta Crystallogr D Biol Crystallogr. 2002; 58: 768-776.

24. Borko Furht, Armando Escalante. Handbook of Cloud Computing. New York: Springer Science. 2010.