BRAZILIAN SYMPOSIUM ON BIOINFORMATICS

Ferradura Resort Hotel
Búzios, Rio de Janeiro, Brazil
August 31th to September 3rd, 2010

BSB 2010 Poster Proceedings

Conference Chair
Sérgio Lifschitz
Department of Informatics - Pontificial Catholic University/Rio de Janeiro, Brazil

Volume Editors
André Ponce de Leon F. de Carvalho – USP, São Carlos – Brazil
Maria Emília M. T. Walter – University of Brasília – Brazil

Promotion
Brazilian Computer Society – SBC

Brazilian Symposium on Bioinformatics (2010 : Búzios, RJ)
BSB 2010 poster proceedings, August, 31 to September, 3, 2010, Búzios, Rio de Janeiro, Brasil / conference chair Sérgio Lifschtiz ; volume editors André Ponce de Leon F. de Carvalho, Maria Emília M. T. Walter. -- [Búzios] : Brazilian Computer Society, [2010].

1 CD-ROM.

ISSN 2178-5120

1. Bioinformática – Congresso. I. Lifschitz , Sérgio. II. Carvalho, André Ponce de Leon F. de. III. Walter, Maria Emília M. T. IV. Título.

Preface

The Brazilian Symposium on Bioinformatics (BSB 2010) took place in Búzios(Rio de Janeiro), Brazil, August 31st to September 3rd, 2010, at Ferradura Resort Hotel.

BSB is a traditional event that has been promoted by the Brazilian Computer Society's (SBC) through its special committee for computational biology (CEBioComp). BSB 2010 was the 5th symposium of the series, though BSB is a new name for its predecessor called Brazilian Workshop on Bioinformatics (WOB). This previous event had three consecutive editions in 2002 (Gramado, Rio Grande do Sul), 2003 (Macaé, Rio de Janeiro), and 2004 (Brasilia, Distrito Federal). Therefore BSB 2010 is already the 8^{th} event of the brazilian bioinformatics community.

BSB 2010 was held co-located with the International Workshop on Genomic Databases (as in 2007) and also with EBB - the brazilian school on bioinformatics. BSB full papers and extended abstracts are published in a special volume of Lecture Notes on Bioinformatics (LNBI) (Springer-Verlag, Germany). However, in order to increase the participation in our event we have decided to create another volume called "BSB poster proceedings" that brings extended abstract of ongoing work in progress.

The BSB organization would like to thank the SBC CEBioComp steering committee members for their support with the whole event, particularly making this poster proceedings feasible. We expect that the readers will get a brief overview of recent initiatives in many different research groups in Brazil, what may contribute to motivate more people to work with bioinformatics and related issues.

Sérgio Lifschitz
BSB 2010 General Chair
August 30th, 2010

Organization

BSB 2010 was organized by the Department of Informatics - Pontifical Catholic University of Rio de Janeiro/Brazil.

Executive Committee

Conference Chair:	Sérgio Lifschitz Pontifical Catholic University of Rio de Janeiro Brazil
Local Arrangements:	Ana Carolina Almeida Carlos Juliano Moura Viana Cristian Tristão Márcia Mártyres Bezerra Paulo Roberto Gomes Renato Marroquín Mogrovejo Pontifical Catholic University of Rio de Janeiro Brazil

Steering Committee

André C.P.L. de Carvalho	(USP, Brazil)
Carlos Eduardo Ferreira	(USP, Brazil)
Kátia Guimarães	(UFPE, Brazil)
Francisco M. Salzano	(UFRGS, Brazil)
João Carlos Setubal	(Virginia Tech, USA)
Osmar Norberto de Sousa	(PUC-RS, Brazil)
Sérgio Lifschitz	(PUC-Rio, Brazil)

Table of Contents

Selected works

A Agent-Based Simulation Tool of Biological Immune System: a Case Study of Autoimmune Diseases

Maurilio de Araujo Possi[1], Alcione de Paiva Oliveira[1], Vladimir Oliveira Di Iório[1], and Cristina Maria Ganns Chaves Dias[2]

[1] Departamento de Informática - Universidade Federal de Viçosa
[2] Departamento de Medicina e Enfermagem - Universidade Federal de Viçosa

Abstract. The immune system defines an area of knowledge of great importance to science, and its study can make important contributions both to biology and to the computer science. In biology, the benefits are obvious, and range from the most effective treatment of infections until a cure for cancer. On the other hand, the increasing needs to deal with complex and dynamic problems, demand new approaches to creating more sophisticated solutions in computer science. One such approach is precisely to seek inspiration in nature, especially in biological systems, in building these solutions. Among these systems, the immune system is a major concern due to its unique features. However, the immune system is still poorly known, and its mechanisms still can not be fully applied, neither by biology nor by the computer science. One way to try to understand it better is building research tools "in-silico" of it. The aim of this paper is to propose a tool for the simulation of the immune system, highlighting its usefulness through the study of autoimmunity.

1 Introduction

The importance of the immune system in humans is clearly evidenced by clinical observation of individuals who, because of some immune deficiency, have become susceptible to severe infections, possibly fatal (Abbas, 2003). Thus, the study and better understanding of it turns out to be extremely relevant to science.

Moreover, for biology (and medicine), the better understanding of this system could provide many benefits, such as the development of new vaccines, the success in transplantation of tissues, the cure for cancer, AIDS, among others.

As for the computer science, the study of complex biological systems has become critical to the advancement of technology. This is because, with the evolution of society, also evolved the problems that the computer science needs to deal, becoming increasingly complex and dynamic. Hence, it is necessary to create increasingly sophisticated solutions to cope with the demand for computing systems increasingly complex.

To try to meet this demand a new approach in developing computing solutions has emerged. Seeking inspiration in biological systems to solve these new

problems, with the same simplicity and elegance that nature does. These systems became known as bio-inspired systems.

Among the various systems that have interesting mechanisms, one stands out: the vertebrate immune system. That's because it has several features such as high parallelism, distribution, organization and adaptation, which makes it a great source of inspiration for computational solutions. Such solutions could use metaphors of the immune system to implement features such as machine learning, anomaly detection, data mining, computer security, adaptive behavior, fault tolerance and pattern recognition, among others.

However, this is not an easy task. First, because the immune system is very complex, being cited as the most challenging topic of biology (Rapin, 2010). Second, because there is still much to discover about the system. Although much has been discovered, many of its mechanisms and their cellular interactions remain incognito for scientists and can not be used neither by biology nor by computer system.

According to Li (2009), it is imperative that we make models of the immune system to learn the functions of each component and the internal mechanisms of that system, so that we can use all the potential of the immune system in biology, and create bio-inspired computer applications in its many useful properties. For biology, he said, models of the immune system are able to simulate with some degree of accuracy mechanisms responsible for various diseases. There are many hypotheses about how the immune system responds to viral infections, but the question remains about the effectiveness of these hypotheses to describe the observed phenomena. Computer models of the immune system may help researchers to understand their mechanisms and also verify their hypotheses. Moreover, they can use this new understanding and inspiration to develop new drugs and use the same models to test their effectiveness. Moreover, Li (2009) also states that use computer models of the immune system is not only cheaper than in-vivo studies, but is also faster.

As for computing, Li (2009) highlights some advantages of modeling the immune system. Among them better understand the bio-inspired algorithms and use these algorithms in the improvement of intelligent and adaptive systems: modeling the immune system is a way to break the bottleneck that exists in the application of metaphors of the immune system in engineering and computing. Through modeling of the immune system functions is possible to know the system better and how the microscopic nonlinear interactions works in the cell level, which can provide guidelines for building computing methods that are distributed and parallel.

It is in this context that we present this work, an approach "in-silico" study, that suggests the use of a multi-agent system to simulate the behavior of the immune system to try to understand it better, and so contribute both to the biology as for to computing. Obviously, it is not feasible, at least for now, to simulate the immune system as a whole. Instead, what we suggest is to simulate a relevant part of the immune system for the study of autoimmunity.

2 Modeling the Immune System

As previously mentioned, modeling the immune system, or part thereof, is not an easy task. It has a lot of non-linear iterations between cells and has the ability to self-regulation in a dynamic environment. All this makes the modeling task very complex and therefore only a few mechanisms, cells and tissues, which are relevant to what one want to study, are included in the model. In this context, the techniques of analysis, modeling and simulation are no longer viable, as it is too complex to model the interactions between each entity of the system. One new way to do it is using multi-agent systems (MAS).

In the MAS models we abstract the interactions and focus only on the individual rules of the entities, leaving the behavior of the system emerges by it self. Moreover, a major advantage of using MAS in the simulation of complex systems, like the immune system, is that it is more natural to create the model, using metaphors that allow direct analogy with the real world. Added to this, we have the evolution of computing power, which meant that systems with large populations of agents started to be viable to be implemented and simulated.

This paper proposes the use of multi-agent systems as a platform for the implementation of the model of the immune system. This is because the use of MAS has been used for years to simulate, understand and predict complex behaviors arising from various dynamical systems. The agent-based simulation is closer to reality than models using other approaches (Li, 2009). In the MAS models, agents can be heterogeneous, with different rules in different groups. These models can explore the emergence of complex functions at the macroscopic level from stochastic microscopic interactions. Using this technique it is possible to verify hypotheses about how cells interact with each other. Li (2009) states that, without doubt, the use of MAS is the best method for modeling complex systems. Also according to Folcik (2007), the immune system is a complex system, one of the most complex biology. MAS-based modeling technique is the most recommended to study these systems, that has a complex and nonlinear behavior. Given the arguments in the literature, the technique to be used for modeling the immune system is the use of multi-agent systems. But, how to create a agent-based model of the immune system?

According to Macal (2009), identify agents, specify their behavior correctly and properly represent their interactions are the key to developing agent-based models. In contrast, we can not underestimate the immune system: Try to model it in its entirety is not feasible. Instead, this project will include in the model only the cells, molecules and mechanisms relevant to understanding the autoimmune pathology chosen. This abstraction is a necessary step in the translation of real-world systems to mathematical models and simulation, and its goal is to achieve the highest possible level of granularity but that still can play effectively the system's behavior in a pre-specified level of interest (Folcik, 2007).

The level of granularity of the model will be the cell-as-agent, i.e., each cell is represented by an agent. Each cell represents a level of abstraction defined and there are many published data on their behavior in response to external

stimuli (Folcik, 2007). This behavior can be abstracted as state machines and encapsulated into classes, generating agents.

The agents are reactive, i.e., their behavior is a response to environmental stimuli, which contain other agents and detectable signals, representing the chemical signals. The internal rules of the agents will be drawn from literature, coded and tested, so that one observes its consequences in system behavior, as described in the literature. Some agents, depending on which cell are representing, will move randomly, while others will follow signs through their gradients, simulating the chemical attraction involved in the immune system.

Substances such as cytokines and chemokines will be represented as signs that will spread, with some degradation factor, forming gradients. This diffusion of substances will be implemented using a feature of the framework, which will simulate the dispersion of the substances based on an evaporation rate, based on literature. Substances that have similar roles, according to literature, will be grouped into a single signal. The system behavior emerges from interactions between society (or population) of agents.

The representation of space will be made through a two-dimensional mesh, where the spaces will be occupied by a single agent. The locomotion of the agents will be done by changing the space that it occupies to one of eight adjacent spaces available, i.e., each agent can move to spaces included in the neighborhood of Moore.

The progression of time will be simulated using features of the framework to be chosen. Typically, these biological simulation's frameworks define the time stochastically, through discrete intervals of time, known as ticks. At each tick, the agents will analyze the environment and act according to their internal rules.

The feasibility of the model will be verified through case studies. First, we will try to observe in the computational model the same behaviors described for the healthy biological immune systems. Later, we will introduce weaknesses in this model, according to an experimental autoimmune disease to be chosen, so that we can verify if the expected behavior will emerge from the computational model as described in experimental model of disease.

References

1. Abbas, A.K. and Lichtman, A.H., Imunologia Básica: Funções e Distúrbios do Sistema Imune, (2003).
2. Folcik, V.A., An, G.C. and Orosz, C.G., "The Basic Immune Simulator: an agent-based model to study the interactions between innate and adaptive immunity.," Theoretical biology & medical modelling, vol. 4, (2007), p. 39.
3. Li, X. Wang, Z., Lu, T. and Che, X., "Modelling Immune System: Principles, Models, Analysis and Perspectives," Journal of Bionic Engineering, vol. 6, (2009), pp. 77-85.
4. Macal, C.M. and North, M.J., "AGENT-BASED MODELING AND SIMULATION," Simulation, (2009), pp. 86-98.
5. Rapin, N., Lund, O., Bernaschi, M. and Castiglione, F., "Computational immunology meets bioinformatics: the use of prediction tools for molecular binding in the simulation of the immune system.," PloS one, vol. 5, (2010), p. e9862.

A Cloud-based Method For Comparing Three Genomes

Renato M. Mogrovejo[1], Carlos Juliano M. Viana[1], Cristian Tristão[1], Márcia Mártyres Bezerra[2], and Sérgio Lifschitz[1]

[1] Pontifical Catholic University of Rio de Janeiro - PUC-Rio
{ctristao,cviana,rmogrovejo,sergio}@inf.puc-rio.br
[2] Oswaldo Cruz Institute - FIOCRUZ - RJ, Brazil
marciamb@ioc.fiocruz.br

Abstract. Genome Comparison is a very important and common task in Bioinformatics, since their results represent the beginning of some other, maybe more complex, tasks. The present work aims to explore new possibilities for finding genome's similarities by extending the previous 3GC method [10], in order to obtain a broader set of similar genes among genomes. This can lead to interesting biological clues about proteins similarities related to particular issues. The new approach tries to find out all related sequences that could have been early dismissed by the traditional 3GC approach. We will briefly describe the possibilities and challenges while taking this bioinformatic's problem into the Cloud.

1 Introduction

Groups of researchers need to host, process, analyze, and share large volumes of unprecedented multidisciplinary data. These needs can be alleviated by cloud computing. An advantage of cloud computing is the parallelism, which can be achieved by allowing access to millions of concurrent users, or by exploiting some possible parallel characteristics of specific data analysis tasks [5]. This computational solution offers many possibilities for scientific researches.

Nevertheless, as explain by [4, 3, 2], the importance of computational science has stopped being a need and become an urgency for scientific discovery. This is why academia is also interested on science's data challenges.

The present work aims to explore new possibilities for finding genome's similarities by extending previous research work done by [10]. Common sequences in different genomes can give some clues about metabolic pathways and proteins related with some particular issue. The three genome comparison (3GC) method looks for pairs or triples of common sequences in a greedy manner. Because of that, some of common sequences are dismissed. Therefore, we are interested in getting all pairs or triples of common sequences that could have been early dismissed during this computational process.

The project's main objective is to compare genome sequences of pathogenic and non-pathogenic organisms of the Protein World Database - PWD [8], trying to identify other genes involved on human pathogenicity.

Our current work is based on using cloud computing technology to improve the 3GC method, not only in terms of time performance, but also in the amount of common sequences obtained. This is because we will be able to process more data in less time and not dismissing previously dismissed ones. This work is only a part of our main goal.

2 Conceptual Design

The three genome comparison approach [10] was developed in order to provide a comparison among three genomes using a Venn-Euler diagram to represent them. The sequence similarity was used to select the sequences that have to be assigned to each region of the diagram. The 3GC method tries to avoid dealing with some complicated cases as illustrated in [10].

The method begins finding as many as possible similar sequences among the three genomes, taking in consideration a given score for each three sequence group. The sequence triplet considered similar are called **triangles**. Each triangle has a weight value which is any metric evaluating each pair of similar sequences. After the triangle has been processed, the method tries to find out as many as possible common sequences between pairs of genomes. Like triangles, 3GC takes into consideration a score value for each pair of sequences called **edges**. Finally the remaining not similar sequences are considered specific of each genome, and then they added to a specific genome region in the diagram.

Specifically, the first stage of 3GC is to assess the weight of every pair of sequences from distinct genomes. Then, the triangles computation is performed in a non-increasing order of their weights. After that, each triangle is assigned to a common region of the diagram, until there are no more triangles with scores greater than a specific threshold T_t. After a triangle Δ_i is included into a specific diagram region, we do not process more triangles or edges sharing the sequences from the Δ_i triangle. Similarly to the second stage, the following one starts with edges been processed in a non-increasing order of their weights. One by one, the edges are assigned to a specific region of the diagram, until there are no more edges with greater scores than a specific threshold T_e. After that, the remaining sequences, which are not member of triangles or edges, are assigned to a specific region of the diagram. These sequences are considered specific of each genome. The general description of the method, as the running time and an experiments performed are fully explained in [10].

3 Implementation Design

The `MapReduce` [6] computing paradigm has been taken by the industry almost as a `de facto` standard for massive dataset computations. MapReduce applications are usually deployed in huge cloud computing infrastructures such as Amazon EC2, Google's, among others.This two-phase model consists in partitioning the data across large number of processors, each of them can analyze a subset of data. However, all solutions have to be designed to fit this model, which most of the time comes unnatural for unexperienced users.

Even though some believe that MapReduce-based systems are best suited for analytical operations, extra complexity is added by using its computation model. Besides, `MapReduce` does not have key features used for analysis of structured data, and it does not have the benefits of an ETL process before analyzing the data. These limitations emphasize the need for systems which integrate declarative query languages from the database world into MapReduce-like software. In this manner, hybrid systems such as HBase, Pig, Scope, Hive, among others, have been proposed.

The Pig system [7] compiles Pig Latin expressions into a sequence of MapReduce jobs, and orchestrates the execution of these jobs on Hadoop [1], a MapReduce open-source implementation. We decided to use Pig because it is a high-level programming language which is simple, and easy to learn. This results in better programming productivity.

Using Pig, users express data analysis tasks as queries similar to SQL. This represents an important possibility in expressing the method to compare three genomes as a group of *simple* queries. But as stated before, our work aims to infer the relationship between three genomes. This would mean trying to express transitive clousure in queries. This is not possible in any languages similar to SQL because SQL-like language does not have enough expressiveness to represent such a thing (first order logic expressive power).

We implemented our version of the three genomes comparison by performing a series of CO-GROUP and FLATTEN operations in order to generate a three way relationship based on our BLAST comparison files. Next, we created a User Defined Function (UDF) to sort the contents of the generated bags using the NCBI GI from the BLAST comparison files. This results on sorted relationships. The repeated ones are then dismissed because they represent the same genomes relationships.

We are planning to extend the experimental work presented by [10], by incorporating new human pathogens genomes from the PWD database [8]. This might result in new biological inferences on genes related to human pathogenesis. Incorporating new similarity data represents a higher computational cost that we will be able to manage using these new computing paradigms.

At this point we are studying data visualization techniques in order to share our current and future results with the research community.

4 Conclusions and Future Works

Nowadays, science is running into big data problems, but it does not have the tools the commercial world has to infer meaning from data, and to exploit this meaning. This is mostly due to the fact that science's problems cannot be completely modeled by just a few features, and also because sometimes scientific problems do not have a high enough economic value.

Nevertheless, cloud computing has become an appealing opportunity to access big computational resources, and to use an utility cost model, just paying for what is used. There are many cloud computing solutions being offered, many of them provide new interesting possibilities to academia, but, as pointed by [9], it

also presents limitations worth studying before hand. These limitations are such as moving big data in and out of our provider's infrastructure(higher costs), data privacy (regional regulatory laws), data management in general, among others limitations.

In addition, data mining applications for scientific discovery mainly requires managing and organizing big data sets. These tasks are suitable for cloud computing deployment because they imply moving data once into the cloud, and then performing computations on it to construct models, or to infer relevant information. Cloud computing environments can make these computationally demanding tasks available for the research community[9]. Finally, as all the process are made on the web, these result can easily be made available for all scientific community.

Our work tries to leverage the bioinformatic's data deluge problem by first deploying our solution on our private cloud, and then use other scientific data sources to obtain more biologically relevant information. As pointed out by [5], data-intensive science applications consist of three basic activities: capture, curation, and analysis. Our ongoing work is part of a series of activities performed in our research laboratory. In that way, our next step is to use the similarity data from the Protein World DB in order to obtain higher-order biological relationships. We have decided to use this PWD database due to the fact that it is also an ongoing research project on our laboratory.

References

1. Hadoop (2010), available at URL: http://hadoop.apache.org/. Accessed on August 1th of 2010.
2. Acebrón, J.A., Spigler, R.: Supercomputing applications to the numerical modeling of industrial and applied mathematics problems. J. Supercomput. 40(1), 67–80 (2007)
3. Dongarra, J., et al., T.S.: High-Performance Computing: Clusters, Constellations, MPPs, and Future Directions. Computing in Science and Engg. 7(2), 51–59 (2009)
4. G., S.C., et al., P.B.: Challenges and Opportunities in Preparing Students for Petascale Computational Science and Engineering. Computing in Science and Engineering 11, 22–27 (2009)
5. Hey, T., et al., S.T. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research (2009)
6. J., D., S., G.: MapReduce: simplified data processing on large clusters. Commun. ACM 51(1), 107–113 (2008)
7. Olston, C., et al., B.R.: Pig latin: a not-so-foreign language for data processing. In: SIGMOD '08: Proceedings of the 2008 ACM SIGMOD. pp. 1099–1110. ACM (2008)
8. Otto, T.D., et al., S.L.: ProteinWorldDB: querying radical pairwise alignments among protein sets from complete genomes. Bioinformatics 26(5), 705–707 (2010)
9. Schadt, E.E., et al., M.D.L.: Computational solutions to large-scale data management and analysis. Nature Review Genetics 11(9), 647–657 (2010)
10. Telles, G.P., Almeida, N., Viana, C., et al., M.E.M.T.W.: A Method for Comparing Three Genomes. Brazilian Symposium on Bioinformatics - BSB 3594, 160–169 (2005)

A possible three-dimensional model for the enzyme chorismate synthase from *Plasmodium falciparum*

Carla Carvalho de Aguiar[1], Danieli F. Figueiredo[1,3], Osmar Norberto de Souza[1,2]

[1]Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas – LABIO, [2]Faculdades de Informática e [3]Biociências, Pontifícia Universidade Católica do Rio Grande do Sul – PUCRS, Av. Ipiranga, 6681, Prédio 32 – Sala 602, 90619-900, Porto Alegre, RS, Brasil
{carla.aguiar, danieli.figueiredo, osmar.norberto}@pucrs.br

Abstract. Presently, chorismate synthase is the only of the seven enzymes from the shikimate pathway identified and characterized in *Plasmodium falciparum*, the organism that causes cerebral malaria. Chorismate synthase is considered a promising drug target against this deadly disease spread worldwide. Here we propose a three-dimensional model for this enzyme using comparative homology modeling. We hope that this model can, in the future, aid in the discovery of novel inhibitor-like small molecules capable of becoming anti-malaria drugs.

Keywords: Chorismate synthase, *Plasmodium falciparum*, Shikimate pathway, Malaria.

1 Introduction

The shikimate pathway plays an important role in the biosynthesis of essential compounds for growth and survival of bacteria, plants, algae, fungi and apicomplexan parasites [1]. The pathway represents a potential target for new drugs since it is absent in mammals and present in organisms responsible for human diseases, such as toxoplasmosis, malaria and tuberculosis [2]. The pathway has seven enzymatic steps, along which phosphoenol pyruvate and erythrose 4-phosphate are converted to chorismate [3].

Of the seven enzymes in the shikimate pathway, the last, chorismate synthase (CS), at present, is the only enzyme well characterized and identified in the agent of cerebral malaria: the apicomplexan parasite *Plasmodium falciparum* (Pf) [1,2]. CS catalyzes the conversion of the substrate 5-enolpyruvylshikimate 3-phosphate (EPSP) to chorismate, in a reaction described as unique in nature [4]. In this process, there is a strict requirement for the presence of reduced flavin mononucleotide (FMN) cofactor, which is not consumed during the process and that, according to its regeneration, classifies CS enzymes as mono- or bifunctional [4,5]. Investigations are underway aiming at elucidating the mechanisms involved in this process [4,6].

The sequences of CS enzymes range from 360-400 amino acids. However, those from Apicomplexa do not follow this pattern, being much longer [7]. Pf CS (CS_Pf) has 527 amino acids, of which 147 are arranged in an unusual way, presenting

themselves as long insertions in the alignments of this enzyme with those from other organisms (Figure 1) [8]. Pf is the most widespread and virulent of the four malaria species and has shown resistance to standard drug treatments [2,3]. The identification of the shikimate pathway in Pf and the detection of a gene encoding the enzyme CS appear to represent an incentive to target this pathway for the development of new drugs [1]. A 3-D model for the CS_Pf enzyme, proposed here, represents a contribution to advance in development of new anti-malarial drugs.

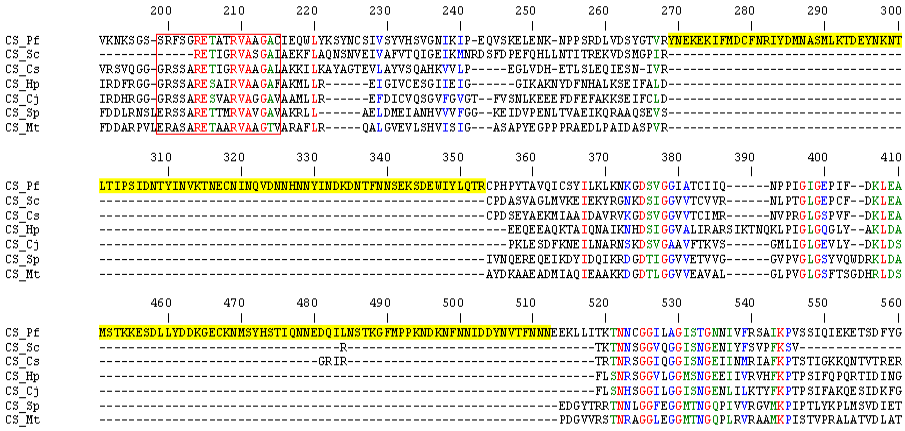


Fig. 1. ClustalW multiple sequence alignment of the CS_Pf sequence. Only part of the multiple alignment is shown with examples of the long insertions, in yellow boxes, in the CS_Pf sequence as compared to CS sequences from other organisms, namely: CS_Sc: *Saccharomyces cerevisiae*; Cs_Cs: *Corydalis sempervirens*; CS_Hp: *Helicobacter pylori*; Cs_Cj: *Campilobacter jejuni*; CS_Sp: *Streptococcus pneumoniae*; CS_Mt: *Mycobacterium tuberculosis*. In the red box is one of the three CS signatures.

2 Materials and methods

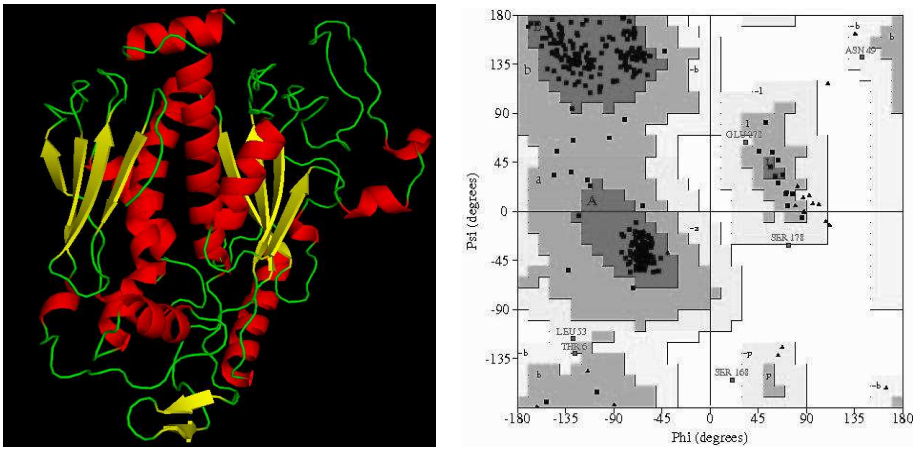
The first step in comparative homology modeling is the identifications of the target sequence, CS_Pf (GenBank Access Number: XP_966212.1), and the sequences and structures of its homologous enzymes that can be used as templates for modeling [9]. Identification of homologous templates was accomplished with the BLASTp [10] program by “blasting” the CS_Pf sequence against the Protein Data Bank (PDB) [11]. As a result, we found the structures of CS from *Saccharomyces cerevisiae* (CS_Sc, PDB ID: 1R53) and *Streptococcus pneumoniae* (CS_Sp, PDB ID: 1QXO). They were both resolved by x-ray diffraction at 2.2 Å and 2.0 Å, respectively [6,12]. They will be used as templates to generate a 3-D model for CS_Pf.

Multiple sequence alignments of the CS_Pf and the templates were performed with the program ClustalW [13] and the best alignment was used in the program MODELLER9v6 [9], which models protein structure by satisfaction of spatial restraints. With the alignment data and the pdb file of the templates, MODELLER can generate as many models as needed. The models stereochemistry was evaluated with

the program PROCHECK [14], and VERIFY 3D [15] was used to assess how well the CS_Pf sequence fitted the template structure. These analyses conclude the validation of the generated models.

3 Results and Discussion

In the search for templates for CS_Pf, seven results were obtained. The sequence of CS_Sc was chosen due to its highest degree of similarity to the target sequence, with 31,0 % identity. However, its 3-D structure has many missing amino-acids residues. This, in turn, motivated the construction of a chimera for the 3-D model of CS_Pf. We looked for a complementary structure for the alignment of CS_Pf and CS_Sc. The structure of CS_Sp complements the proposed alignment, and has the interesting additional feature which is the presence of its substrate and cofactor. The alignment between the templates and target sequences preserved the three regions of signature patterns of CS (Figure 1), which are rich in basic residues. The long insertions (Figure 1) in the target sequence were manually removed. This action is not expected to affect the core structure of the 3-D model of the enzyme. Ten models were generated for the enzyme CS_Pf and the best one was selected based on the analyses with the program PROCHECK [14]. The best model (Figure 2a) showed 326 (90.8 %) of non-glycine and non-proline residues in the most favorable regions of the Ramachandran plot (Figure 2b).



(a) (b)

Fig. 2. (a) Ribbon representation of the backbone 3-D model of CS_Pf generated with PyMol. The β-sheets are colored yellow, α-helices are in red, and turns and loops are green. Its folds into an α/β 4-layer sandwich. (b) The Ramachandran plot for the best 3-D model of CS_Pf showing the majority of the amino-acid residues in the most favorable regions (A and B).

The analyses of the proposed model are in the initial stage but have shown promising qualitative results. Further assessments regarding the implications of the insertions' removals on the structure and function of CS_Pf enzyme will be explored with molecular docking and dynamics simulations.

Acknowledgments. This work was supported by MCT/CNPq grants 410505/2006-4 and 312027/2006-0 to ONS. ONS is a CNPq Research Fellow. CCA was supported by a FAPERGS PIBIC scholarship. DFF is supported by a PUCRS-BPA scholarship program.

References

1. McRobert, L., Jiang, S., Stead, A., McConkey, G.A.: *Plasmodium falciparum*: Interaction of shikimate analogues with antimalarial drugs. Exp. Parasitol. 111, 178-181 (2005)
2. McConkey, G.A.: Targeting the shikimate pathway in the malaria parasite *Plasmodium falciparum*. Antimicrob. Agents Chemother. 43, 175-177 (1999)
3. Roberts, C.W., Roberts, F., Lyons, R.E., Kirisits, M.J., Mui, E.J., Finnerty, J., Johnson, J.J., Ferguson, D.J.P., Coggins, J.R., Krell, T., Coombs, G.H., Milhous, W.K., Kyle, D.E., Tzipori, S., Barnwell, J., Dame, J.B., Carlton, J., McLeod, R.: The shikimate pathway and its branches in Apicomplexa parasites. J. Infect. Dis. 185(Suppl 1), 25-36 (2002)
4. Herrmann, K.M., Weaver, L.M.: The shikimate pathway. Annu. Rev. Plant Physiol. Plant Mol. Biol. 50, 473-503 (1999)
5. Ehammer, H., Rauch, G., Prem, A., Kappes, B., Macheroux, P.: Conservation of NADPH utilization by chorismate synthase and its implications for the evolution of the shikimate pathway. Mol. Microbiol. 65, 1249-1257 (2007)
6. Quevillon-Cheruel, S., Leulliot, N., Meyer, P., Graille, M., Bremang, M., Blondeau, K., Sorel, I., Poupon, A., Janin, J., van Tilbeurgh, H.: Crystal structure of the bifunctional chorismate synthase from *Saccharomyces cerevisiae*. J. Biol. Chem. 279, 619-625 (2004)
7. Fernandes, C.L., Breda, A., Santos, D.S., Basso, L.A., Norberto de Souza, O.: A structural model for Chorismate synthase from *Mycobacterium tuberculosis* in complex with coenzyme and substrate. Comp. Biol. Med. 37, 149-158 (2007)
8. Roberts, F., Roberts, C.W., Johnson, J.J., Kyle, D.E., Krell, T., Coggins, J.R., Coombs, G.H., Milhous, W.K., Tzipori, S., Ferguson, D.J., Chakrabarti, D., McLeod, R.: Evidence for the shikimate pathway in apicomplexan parasites. Nature. 393, 801-805 (1998)
9. Sali, A., Blundell, T.L.: Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815 (1993)
10. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., Madden, T.L.: NCBI BLAST: a better web interface. Nucleic Acids Res. 36(Suppl 2), W5-W9 (2008)
11. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne P.E.: The Protein Data Bank. Nucleic Acids Res. 28, 235-242 (2000)
12. Maclean, J., Ali, S.: The structure of chorismate synthase reveals a novel flavin binding site fundamental to a unique chemical reaction. Structure. 11, 1499-1511 (2003)
13. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-4680 (1994)
14. Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M.: PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26, 283–291 (1993)
15. Lüthy, R., Bowie, J.U., Eisenberg, D.: Assessment of protein models with three-dimensional profiles. Nature. 356, 83-85 (1992)

A Provenance Model for Bioinformatics Workflows

Luciana Almendra Gomes[1], Sérgio Lifschitz[1] , Philippe Picouet[2] ,
Priscila V. S. Z. Capriles[3], Laurent E. Dardenne[3]

[1] Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio,
[2] Ecole Nationale Supérieure des Télécommunications de Bretagne - ENST Bretagne,
[3] LNCC/MCT
lgomes@inf.puc-rio.br, sergio@inf.puc-rio.br, Philippe.Picouet@telecom-bretagne.eu,
capriles@lncc.br, dardenne@lncc.br

Abstract. Some SWfMS (Scientific Workflow Management Systems) are being used to construct and execute bioinformatics experiments. An important feature of these systems is the ability to automatically capture data provenance which consists in metadata about the prototype and execution of the experiment. In this work we aim to produce a bioinformatics provenance model to the BioSide system using the MHOLline workflow as a case study.

Keywords: provenance, scientific workflows

1 Introduction

Scientific workflow management systems (SWfMS) are being used to enact bioinformatics workflows, which before these systems could only be automated by the use of scripts. These systems help the scientists to construct and manage workflows, offering in general a drag-and-drop interface where they can add and link processes in a visual way. One important contribution of SWfMS is the automatic capture of data provenance, which consists in metadata about the execution (retrospective provenance) and specification (prospective provenance) of the experiment [1].

The existence of different ways that the SWfMS and other systems capture, store and query the data provenance was one of the motivations to the Provenance Challenges [2]. The First Provenance Challenge was proposed to produce a better understanding of the characteristics and capabilities of these different ways to manage data provenance. The Second Provenance Challenge was about interoperability between systems, which results motivated the specification of an abstract model to provenance, the Open Provenance Model. OPM was evaluated in the Third Provenance Challenge, which led to the last revision of the model [3].

One of the goals of OPM is to allow developers to build tools that operate on the model, making easier the comprehension of the data provenance that the system produces, and facilitating the exchange of these data with other systems.

This work intends to produce a model of provenance based on OPM to BioSide [4] system, which is a SWfMS primarily designed for biological experiments. We will

use as case study the MHOLline workflow [5] which will help us identify some important features of provenance to Bioinformatics experiments. The next sessions will give a short description of BioSide, MHOLline and a conclusion with the expected contributions of this work.

2 The BioSide System

BioSide [4] is a SWfMS which allows the construction of data-flows. The system was developed to a specific project [6] to help scientists in the construction, parameterization and reuse of biological workflows. BioSide is attractive because it has a simple interface, with very easy interaction. But the system lacks important features of provenance. There is no provenance model, and no way to query provenance information. Each execution of a workflow is registered in a specific folder with the description of the workflow, the intermediary data and results generated and logs of execution. These past executions are displayed in the interface, but cannot be queried, and they are not described by a specific model.

We aim to develop a data provenance model for BioSide, specifying which data will be registered and how it will be stored and queried. We will compare our solution with other provenance models, like those used by Taverna [7] and Vistrails [8]. This comparison will focus on whether or not these models supply the provenance needs of the MHOLline workflow, a biological workflow that we will use as a case study.

3 The MHOLline Workflow

MHOLline [5] is a biological workflow used for protein structure comparative modeling. Scientists over the world can access MHOLline page [5] and submit a file of protein sequences (in the FASTA format) containing one or more sequences. A link to the final results are sent via email to the user, or downloaded in the page. Nowadays the workflow is running as a script in Perl language, called by a PHP application which receives the requests from the users.

We aim to implement the MHOLline workflow in the BioSide system, and use it as a model to observe what biological workflows, specially sequence processing workflows, need in terms of provenance.

MHOLline already does automatically some registries which are “classic” retrospective provenance data, for example, the duration of each process and whether the execution terminated successfully. Naturally we will try to *cover those classic provenance data required by the MHOLline with our model*, but we are interested in raise all another provenance requisites this workflow has. We have identified one unusual requisite that maybe can be generalized for any workflow, which will be explained as follows.

Provenance of n-composed workflow experiment. MHOLline starts with an input file containing one or more sequences in FASTA format. Each of those sequences is aligned with all sequences of PDB (Protein Data Bank) [9,10] using the BLAST

program [11]. Then the BATS (Blast Automatic Targeting for Structures) program is executed to give a MHOLline score for each BLAST alignment and it choose for each input sequence the best PDB sequence alignment, called “champ sequence”. This champ sequence is the only one that will be used by the MODELLER program [12] to finally construct the three-dimensional (3D) model for the respective input protein sequence.

However there is an improvement we want to provide which is a kind of refinement of the experiment: at the end of the usual execution, the user will be allowed to choose from the BATS report any other sequences of PDB in order to construct other 3D models. A similar workflow will run, starting in the process that runs right after the BATS program (FILTERS program), which will use as input data that come from the first execution. Then the user can compare that one generated with the BATS choice and this new one, generated with his own choices.

In terms of process modeling, there are several ways to define this improvement in the MHOLline workflow, which will not be treated in this work. We have chose to model this required process as two different workflow definitions, the original one and other that starts at a new process which will capture the user choice of sequences. After this new interaction process the workflow will continue from the FILTERS program.

Our main concern here is with the *provenance registry of an experiment which is composed by two or more different workflow executions of different or equal workflow definitions*. The workflow’s final and intermediary outputs may be used as input to other executions, which leads to a problem of how can we reuse intermediary output data from executions which already happened.

4 Conclusion

Is this work we aim to produce a model of Provenance which could be used by BioSide or another interested SWfMS . This model consists in which data will be captured and how it will be stored and queried. It should have an easy way to export to OPM in order to promote better understanding and to allow exchange of provenance data between BioSide and other systems.

We are using MHOLline workflow as motivation to raise some provenance registry requisites for bioinformatics workflows. Our goal is to satisfy both classic requirements which are in general present in all provenance systems, and unusual requirements, like experiments composed by more than one workflow.

Acknowledgments

We thank Sébastien Bigaret (Télécom Bretagne - France) for his complete support and help with BioSide system. The work described here is partially funded by CNPq.

References

1. S. M. S. Cruz, M. L. M. Campos, M. Mattoso, "Towards a Taxonomy of Provenance in Scientific Workflow Management Systems ", SERVICES I, pp. 259-266, 2009.
2. Provenance Challenge Wiki. Available at URL: http://twiki.ipaw.info/bin/view/Challenge/
3. L. Moreau et al., "The Open Provenance Model core specification (v1.1)", Future Generation Computer Systems. Article in Press, Corrected Proof. 2010. doi:10.1016/j.future.2010.07.005
4. S. Bigaret, P. Meyer, "BioSide : from bioinformatics needs to a generic workflow engine", 2nd Decision Deck Developers Days, 03-04 december 2008, Paris, France, 2008.
5. MHOLline Workflow. Available at URL: www.mholline.lncc.br.
6. The CapExBio Project. Available at URL: http://en.academic.ru/dic.nsf/enwiki/10106361.
7. D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, and T. Oinn, "Taverna: a tool for building and running workflows of services", Nucleic Acids Research, vol. 34, iss. Web Server issue, pp. 729-732, 2006.
8. S. Callahan, J. Freire, E. Santos, C. Scheidegger, C. Silva, H. Vo, "VisTrails: Visualization meets Data Management". In: Proc SIGMOD 2006, pp. 745-747.
9. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) The Protein Data Bank, Nucleic Acids Research, 28: 235-242.
10. RCSB PDB. The Protein Data Bank. Available at URL: www.pdb.org
11. S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman, "Basic Local Alignment Search Tool", J. Mol. Biol., 215 (1990), pp. 403-410.
12. N. Eswar, M. A. Marti-Renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali, "Comparative Protein Structure Modeling With MODELLER", Current Protocols in Bioinformatics, John Wiley & Sons, Inc., Supplement 15, 5.6.1-5.6.30, 2006.

Ab initio Protein Structure Prediction via Genetic Algorithms using a Coarse-grained Model for Side Chains

Priscila V. S. Z. Capriles, Fábio L. Custódio, and Laurent E. Dardenne

Laboratório Nacional de Computação Científica - LNCC/MCT
Av. Getúlio Vargas, 333 - Quitandinha - RJ - Brazil
Tel.: +55-24-22336121 - {capriles,flc,dardenne}@lncc.br
http://www.gmmsb.lncc.br

1 Introduction

The knowledge of the three-dimensional (3D) conformation of the native structure of proteins presents important applications in biotechnological researches aiding in the *de novo* protein design, structure based drug design and refinement of theoretical models obtained by comparative modelling (CM) [1] [2]. However, experimental techniques to solve 3D conformations has not followed the increasing number of sequenced genomes. This also limits the use of CM techniques that use solved homologous structures to predict the 3D structure of proteins.

As an alternative, *ab initio* protein structure prediction (PSP) method consists in determining the 3D native structure of proteins from their amino acid sequences, using only physical principles. Usually, PSP methods assume that the conformation the protein adopts under physiological conditions is the conformation with the lowest Gibbs free energy (thermodynamic hypothesis) [3]. Then, dealing with it as an optimization problem, PSP can be decomposed in two sub-problems: (i) to define an appropriate energy function that places the native structure on the global minimum and is able to discriminate correct from incorrect folds, and (ii) to develop an efficient and robust search strategy [9].

These involve the optimization of a computationally expensive energy function with thousands of degrees of freedom associated with complex energy landscapes, that is, highly degenerated (including multiple minima), with massive multi-modality (roughness), and large regions of unfeasible conformations [9]. Generally, the PSP optimization problem is carried out by metaheuristic, amongst them the Genetic Algorithms (GA). GA are inspired by the Darwinian priciples of evolution [7, 8] and due to the fact that they works with a population of candidate solutions, GA can mimics the diversity nature of the PSP problem. However, to reach sub- or optimal solutions, GA requires a large number of function evaluations limiting their usage in expensive problems such as PSP [10].

One possible solution to this problem is the use of simplified representation of the system. In the coarse-grained (CG) model, each amino acid representation is reduced in a set of few interactions site [12]. Therefore, the high number of degrees of freedom are averaged in a small one, decreasing the computational time required in the evaluation of each individual in the population of GA.

In previous work, we presented a crowding-based steady-state GA (CSSGA) developed for the all-atom *ab initio* PSP problem [9], using a similarity-based surrogate model to reduce computational time cost. In this work, we replaced the surrogate model for a CG model to represent side chains of amino acids. The coarse-grained CSSGA was applied to a test set of proteins, and this adaptation resulted in improvements in the performance of the algorithm and in reducing computational time.

2 Representation of Protein Structure

In real proteins, the atoms are connected following specific geometries (according to the types of atoms participating in the link). Since the atomic composition of the protein does not change during the PSP, these geometries can be automatically fulfilled by using an internal coordinates representation [4] and performing movements in the structure by changing the torsion angles [9]. In reality, the distances and angles of covalent bonds are not fixed, but the changes that occur around the equilibrium values are usually small. Thus, an approach commonly used by PSP methods is to keep the geometry of chemical bonds fixed during the search.

In this work, the conformation of a peptide chain can be defined by a series of torsion angles (backbone dihedral angles) ϕ and ψ based on [5]. The positioning and representation of coarse grained side chains were adapted from [11], *i.e.*, each side chain is then represented by a single "atom" placed at the geometrical center of the true side chain. The position of this atom remains fixed in relation to the C_α during conformation changes, thus movements of the side chains (by rotamer libraries or not) are no longer modeled.

3 Energy Function

The energy function from the interaction between the atoms of the backbone is calculated using the classical molecular force field GROMOS96 [5]. Interactions involving side chains is calculated using parameters from the coarse-grained force field OPEPv3 [11]. The force field terms modeling bond geometries are not used as they are invariant. The total energy function has the following form:

$$E_{total}(r_i) = E_{torc} + E_{coul} + E_{LJ}. \tag{1}$$

The energy barriers from the rotation of bonds (proper dihedral potential) and the electrostatic interactions (Coulomb potential) are respectively given by

$$E_{torc} = \sum_{n}^{N_\phi} K_{\phi n}[1 + \cos(n_n \phi_n - \delta_n)], \tag{2}$$

$$E_{coul} = \sum_{i \leq j}^{N_{atoms}} \frac{q_i q_j}{4\pi\epsilon_0 \epsilon_r(r_{ij}) r_{ij}}, \tag{3}$$

where $K_{\phi n}$ is the energy constant associated with the torsion of a bond, ϕ_n is the torsion angle, n_n is the period, δ_n is the phase angle, r_{ij} is the distance between atoms i and j, q_i and q_j are the atomic charges of atoms i and j, and ϵ_r is a

distance dependent dielectric sigmoid function, which models the attenuation of the attraction between distant charges, caused by water as a solvent, and preserves the strong attraction at shorter distances [6].

The Lennard-Jones potential (E_{LJ}) models represents the atomic repulsion at very small distances and the attraction of van der Waals (vdW). In this work, we use a special formulation for interactions involving the coarse-grained side chains, as follows:

$$E_{LJ} = \sum_{i \leq j}^{N_{atoms}} \begin{cases} -\sigma_{ij} \left(\frac{A_{ij}}{r_{ij}}\right)^6 & \text{if } \sigma_{ij} < 0, \\ \sigma_{ij} \left[\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{B_{ij}}{r_{ij}}\right)^6\right] & \text{else,} \end{cases} \tag{4}$$

where A_{ij} and B_{ij} are Lennard-Jones parameters dependent of the atomic type, and σ_{ij} is the energy well depth at the minimum, based on OPEPv3 [11]. For backbone-backbone interactions, only the 12-6 potential is calculated and σ_{ij} is set to 1.

4 Results and Discussion

We tested a set of eight small proteins and compared the obtained results with those from the all-atom CSSGA (Table 1). The control test using poly-alanine sequences (18ALA and 23ALA) showed structural and time results similar to those obtained with the all-atom CSSGA, that is, the lowest energy conformation, and the native conformation were found. This guarantee the maintenance of the CSSGA efficiency. Except for PDB1AMB, the coarse-grained CSSGA modeled structures with comparable quality (similarity to the known native structure) to those modeled under the all-atom model. Furthermore, by using a simplified sidechain representation the computational time required for performing the same number of function evaluations decreased about 50%. Some sample structural results are presented in Figure 1 (see Appendix).

Acknowledgments. The Brazilian National Council of Research (CNPq) and the FAPERJ Foundation have supported this work.

References

1. Röthlisberger, D., *et al*: Kemp elimination catalysts by computational enzyme design. Nature 453(7192), 190–195 (2008).
2. Davis, I.W., Baker, D.: RosettaLigand docking with full ligand and receptor flexibility. J. Mol. Biol. 385(2), 381–392 (2009).
3. Anfisen, C.B.: Principles that govern the folding of proteins. Science 181, 187 (1973).
4. Jensen, F.: Introduction to computational chemistry. Wiley New York (1999).
5. Schuler, L.D., *et al*: An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. J. Comput. Chem. 22(11), 1205–1218 (2001).
6. Arora, N., Jayaram, B.: Strength of hydrogen bonds in alpha-helices. J. Comput. Chem. 18(9), 1245–1252 (1997).

Table 1. Detailing of test set and results of all-atom and coarse-grained CSSGA.

ID (Class)	Length	Atoms[a]	Atoms[b]	RMSD[c]	RMSD[d]	Time[e]	Time[f]
18ALA (α)	18	111	111	-	0.281[g]	40.28	44.46
23ALA (α)	23	141	141	-	0.289[g]	48.31	50.21
BETA3S(β)	20	233	121	-	5.567[g]	87.31	39.55
PDB1L2Y (α)	20	198	120	3.343	3.571	69.11	46.50
PDB1AMB (α)	28	305	169	1.395	3.256	127.35	64.15
PDB1VII (α)	36	389	217	3.389	4.119	173.58	82.24
PDB1E0L (β)	37	410	223	6.764	6.060	217.49	88.27
PDB1E0G ($\alpha + \beta$)	48	500	288	5.736	5.731	306.26	128.44

Averaged results from 10 runs of 2×10^6 function evaluations (processor Intel Xeon E5520). Total number of atoms with [a]all-atom representation and [b]CG representation of side chains. RMSD (in $\mathring{A}$) calculated between the [c]backbone of the known structure and the structure with all-atom representation, and [d]the backbone of the known structure and the structure with CG representation. Elapsed time (in minutes) spent in [e]all-atom CSSGA and [f]coarse-grained CSSGA. [g]RMSD (in $\mathring{A}$) calculated between structure with lowest energy in all-atom and CG representations.

7. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI, USA (1975).
8. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Publishing Co., Reading, Mass., USA (1989)
9. Custódio, F.L., *et al*: Full-Atom *Ab Initio* Protein Structure Prediction with a Genetic Algorithm using a Similarity-based Surrogate Model. In: WCCI 2010 IEEE World Congress on Computational Intelligence, pp. 2554–2561. IEEE Press (2010).
10. Fonseca, L.G., *et al*: A similarity-based surrogate model for enhanced performance in genetic algorithms. OPSEARCH 46(1), 89–107 (2010).
11. Maupetit, J., *et al*: A coarse-grained protein force field for folding and structure prediction. PROTEINS 69, 394–408 (2007).
12. Kozlowska, U., *et al*: Determination of side-chain-rotamer and side-chain and backbone virtual-bond-stretching potentials of mean force from AM1 energy surfaces of terminally-blocked amino-acid residues, for coarse-grained simulations of protein structure and folding. I. The method. J. Comput. Chem. 31, 1143–1153 (2010).
13. Westbrook, J., *et al*: The Protein Data Bank: unifying the archive. Nucleic Acids Res. 30(10), 245–248 (2002).
14. Humphrey, W., *et al*: VMD – Visual Molecular Dynamics. J. Mol. Graph. 14, 33–38 (1996).

Appendix: Cartoon representation of 3D structures predicted via coarse-grained CSSGA

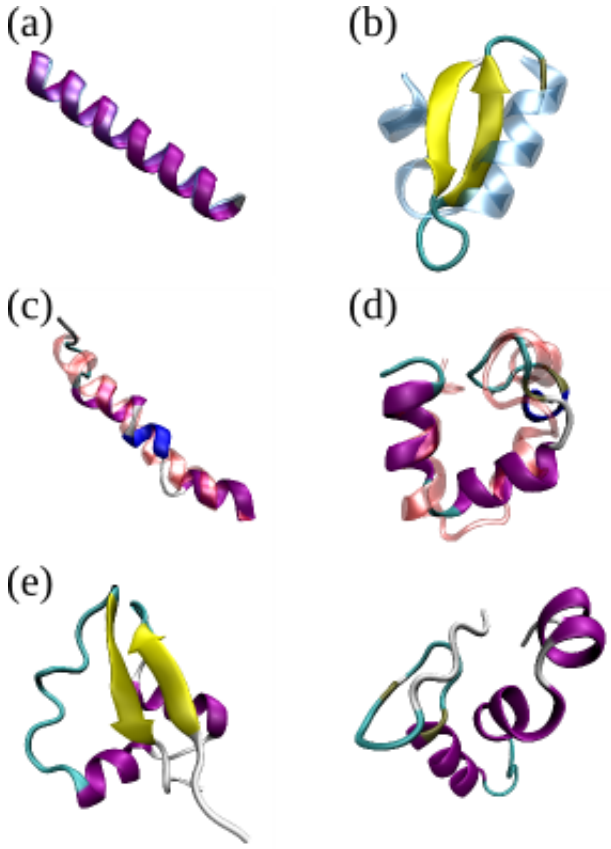


Fig. 1. Cartoon representation of 3D structures predicted via coarse-grained CSSGA.

In blue are represented the best results with all-atom CSSGA. In red are represented the structures from PDB (Protein Data Bank) [13]. The structural alignment was performed using VMD program [14]: (a) 23ALA; (b) BETA3S; (c) PDB1AMB; (d) PDB1VII. In (e) is presented the PDB1E0G (left) and the structure predicted in this work using CG representation (right).

An algorithm to search and repair errors and nonconformities in a biological database

Flávia G. Silva, Kátia P. Lopes, Sandro R. Dias

Faculdade Anhanguera, Departamento de Sistemas de Informação

Av. dos Andradas, 436, Centro, Belo Horizonte - MG, Brasil

flaviagomes.silva@yahoo.com.br, katiaplopes@gmail.com,

sandrord@gmail.com

Abstract. With the technology and scientific research advances in molecular biology, the volume of the generated information has increased exponentially, at this moment were created the biological databases. But, despite periodic reviews, several errors and no conformities were found in the handling of PDB (Protein Data Bank) proteins [1]. So, the goal of this article is to implement and present an algorithm in the PERL language, to scan the PDB bank for errors and nonconformities, to identify and suggest corrections for the errors found in order to contribute positively to subsequent research based on this database.

Keywords: PERL, database PDB, bioinformatic.

1 Introduction

A great event in modern molecular biology was not only the DNA structure discovery, but also the conclusion that DNA was the substance that carried genetic information of nucleic acids to proteins [5]. Since then, with the technology and scientific research advances, well as methods used in the sequencing of biological information, the volume of the generated information has increased exponentially, making a challenge to find the best way of storing and handling such information, because several DBMS (Database Management System) are developed to deal just organizational data [7]. So, the biological databases were created to deal these data. The figure 1 shows the steps taken to storage segments (sequences of genes) of DNA in a database.

Currently there are a lot of biological information in public databases because scientists around the world use this data to their research. In these databases, were discovered evolutionary information of genes, chromosome location, structural information of proteins, functional information about molecules, enzymes and disease [6].

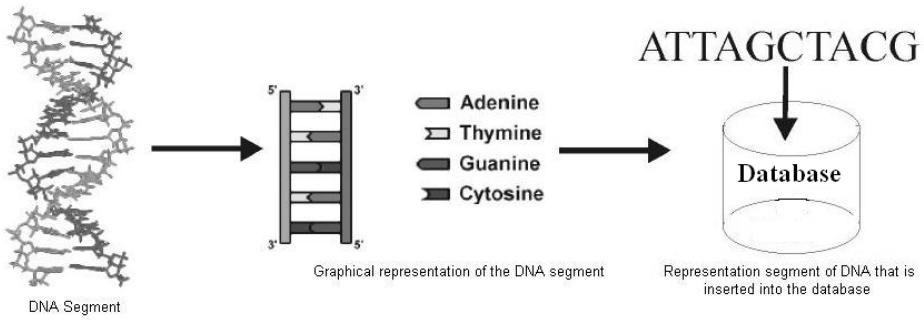


Fig. 1 - Steps taken to hold a series of genes in a DNA database.

Adapted: WIECZOREK, LEAL, 2002 [7].

Specifically for this work was used biological database Protein Data Bank (PDB), available at: http://www.pdb.org [1]. It has three dimensional informations about protein structures, nucleic acids and complex sets, primary and secondary structure of proteins, as well as angles and distances between atoms. These data are stored in a set of files containing a standardized nomenclature: flat files. These files are submitted to periodic review, order to maintain the trust of the stored data. They are organized, so that the researcher can draw the necessary information. They contain a header with information about the stored object, and then all the related data. But, not all files following the pattern proposed or errors go unnoticed by the researcher, which can avoid the process of manipulation and analysis of data for other researchers. Then, this study aims to present an algorithm implemented using PERL (Practial Extract and Report Language) to scan the full bank of PDB (Protein Data Bank) for errors and nonconformities, based on document: “Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description” available at: http://mmcif.pdb.org/dictionaries/mmcif_pdbx.dic/Index/.[4]

2 Objectives

Implement and present a script in PERL that scans the full bank of PDB for errors and nonconformities to identify and suggest corrections for the errors found in order to contribute positively to subsequent research based in format presented in the documentation from the PDB. So, expected to contribute positively to subsequent research based on this database.

3 Methodology

The used methodologies are:

- Programming language: PERL. This language was chosen because it is very adopted for professionals of bioinformatics area, simple and very syntactically rich. Besides being recognized as a sophisticated language, with characteristics of high level and possess strong point as the manipulation of text and a good connectivity with databases using specific libraries 1 [5].

- Database Management System: MySQL. It was chosen because widely known and used, provides robustness to work with identified [3] some of the problems listed below (not an exaustive list):

    1. The atoms are not numbered consecutively indicating that not all the atoms were resolved, whatever, not all the atoms that make up the protein are present in the file.
    2. The number of residue is not consecutive indicating the lack of one.
    3. The number of residue contains, besides the number, a letter.
    4. Many residue can have the same number.

Then, after spending time and effort to solve these errors, we observed the need of a treatment for them, and it would be interesting to suggest corrections for a greater contribution to other researchers.

This work is under development, but the PDB files were collected and stored on a server, it is being updated weekly, since the PDB is very dynamic. So, the script developed in PERL language will run that database through the following steps: 1) Identify errors and nonconformities as standard documentation PDB. 2) Generate log with identification file and their problems. 3) Correct the problems. 4) During this step, data will be processed and the data result will be stored in a database for use by the MySQL DBMS for better data analysis and reporting.

4 Future work

In this moment, from the work of Dias & Nagem (2009) and based in the document: “Protein Data Bank Contents Guide - Atomic Coordinate Entry Format Description”, available at the site of PDB, we could note errors and nonconformities in the PDB database, and export the incorrect base for a local server to run the scripts. As a result, we expect to suggest corrections of errors and to generate reports and statistics from the data handled since part of this protein database provides a wealth of information.

Acknowledgementes: Faculdade Anhanguera.

[1] http://www.cpan.org/

5 References

1. Berman, H.M. et al. The Protein Data Bank. Oxford Journals, 2000, Vol. 28, No. 1 235-242, Disponível em: < http://nar.oxfordjournals.org/cgi/content/full/28/1/235>. Acesso em: 8 nov 2009.
2. Dias, S. R. ; Nagem, R. A. P. Residue-residue interaction database: use in the modification of proteins. In: International Network of Protein Engineering Centers, 2009, Ubatuba/SP. International Network of Protein Engineering Centers Meeting Abstract Book, 2009.
3. Lopes, K. P. ; Dias, S. R. Comparação entre diferentes tecnologias em banco de dados para manipulação rápida e ubíqua de dados biológicos. In: III Encontro de Modelagem Computacional – Laboratório Nacional de Computação Científica - 2010, Petrópolis /RJ.
4. Protein Data Bank contents guide: Atomic coordinate entry format description. Version 3.20. Disponível em: <http://mmcif.pdb.org/dictionaries/mmcif_pdbx.dic/Index/> Acesso em: 17 jun 2010.
5. Prosdocimi, F. et al. Bioinformática: Manual do usuário. Um guia básico e amplo sobre os diversos aspectos dessa nova ciência. Biotecnologia, Ciência e Desenvolvimento. N.29 P.1-14, 2003.
6. Souto, M.C.P. Banco de Dado Biológicos. Rio Grande do Norte: Universidade Federal do Rio do Grande (UFRN), 2004. 53 slides: color. Acompanha texto.
7. Wieczorek, E.M.; Leal, E. Caminhos e Tendências do uso de banco de dados em bioinformática. IV Encontro de Estudantes de Informática do Estado do Tocantins. P.1-9, 2002.

APPROACHING PROTEIN FOLDING THROUGH NEURAL NETWORKS

BELLINI, R. G.[1] – RIBEIRO, T. S.[1] – FIGUEIREDO, K.[1,2] – PACHECO, M. A.[1]

[1]*Applied Computational Intelligence Laboratory – ICA*
Pontificial Catholic University/Rio de Janeiro-Brazil

[2]*UEZO /Rio de Janeiro-Brazil*

reinaldo.bellini@gmail.com – thiagoribeiro1@gmail.com – karla.figueiredo@gmail.com – marco@ele.puc-rio.br

Abstract: The understanding of protein folding is supposed to bring answers for a number of diseases such as Alzheimer`s disease. The sequence of amino acids determines 3D geometry of a protein. This paper presents a preliminary study on the protein folding problem through the geometric properties between α-carbons, using standard physicochemical properties of amino acids which were modeled using an artificial neural network. We investigate the angle between α-carbons without considering the neighborhood interactions which influence the angle formation. The main idea is to evaluate the ability of an artificial neural network to find the angle between a triad of α-carbons by just using their respective geometric position. The neural network presented good results on finding those angles, with an error rate associated to the degeneration of the aminoacids`s angle.

Keywords: Protein-folding, neural network, α-carbons, protein backbone

1. Introduction

Proteins are the most important structures in the human body because their association with several processes such as hair protection and inhibition of a specific disease. The incorrect folding of proteins are associate with serious diseases such as Alzheimer`s disease. These structures are built through amino acids, little structures with many physical chemistry proprieties. The sequence of amino acids determines 3D geometry of a protein [2]. The investigation of how proteins fold will allow the understanding of several diseases since diseases are related with an incorrect fold of the chain. This paper aims to conduct a preliminary study on the protein folding problem through the geometric properties between α-carbons, using standard physicochemical properties of amino acids which are modeled using an artificial neural network [3]. Neural Networks are nonlinear computational models inspired by the structure and operation of the human brain that are capable

of performing the following tasks: learning, association, generalization and abstraction. A Neural Network is composed of several highly interconnected processing elements (artificial neurons) that perform simple operations and transmit their results to the neighboring processors figure1.RNs are trained so that when an input set is applied, a desired output set is produced. During training, the network's weights gradually converge to certain values in order that the application of the input vectors may yield the necessary outputs.

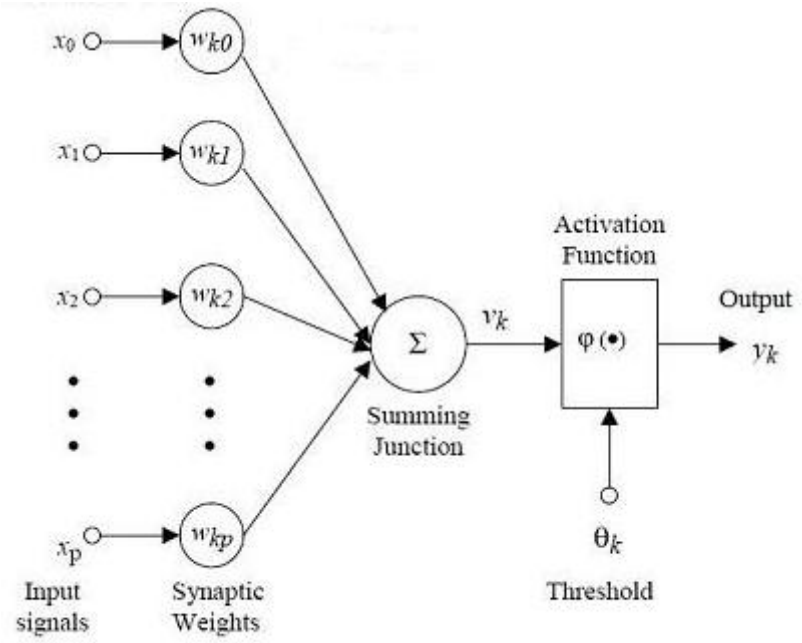


Figure 1: Neural Network

2. Methodology

The structures of proteins used in this experiment were obtained from crystallographic data deposited in a protein database (*Protein Data Bank*) [1]. Once collected, the data were first pre-processed. This procedure consisted of cleaning the base by extracting the information from the α-carbon and calculating the angle between them. We used 49,932 triads of amino acids, divided into 3 parts: 30,549 for training the network, 10,183 to perform the early-stopping and 9,200 for testing. The following attributes were used as inputs to the network: Side Chain Polarity, Side Chain Charge, Hydrophobicity, Average Residue Mass, Percent Buried Residues, van der Walls Volume.

3. Results

Real	Predicted	Error
96.3180	96.3308	0.0130%
99.8364	99.8276	0.0090%
107.1139	107.1179	0.0040%

Table 1: Difference between the actual and predicted angle, considering the smallest percentage error found.

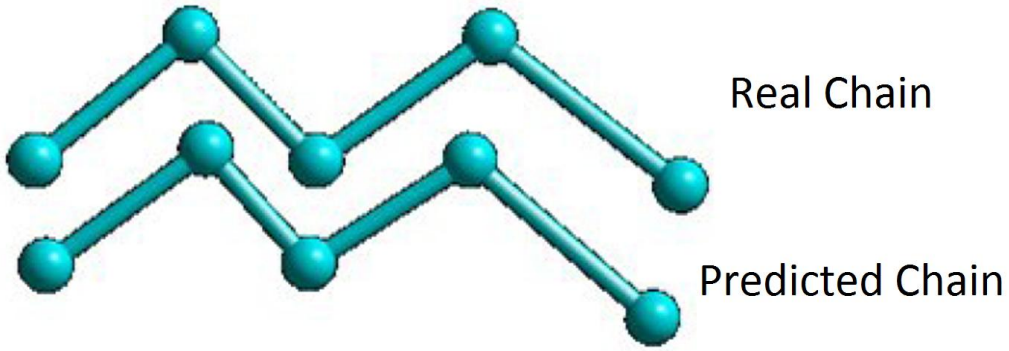


Image 1: Conformational difference between the real and predicted chain, considering the smallest percentage error found.

4. Conclusions

Even not considering a neighborhood that interferes in the folding process, the neural network was able to find the angle with good accuracy, making the real and predicted chain with similar geometry. Thus, the actual result shows that neural network is a promising technique to address the problem of protein folding as it can provide advances to a better understanding of this problem.

5. References

[1] Protein Data Bank (PDB)http://www.pdb.org/pdb/home/home.do

[2] Lodish, Berk, Baltimore, et al. Molecular Cell Biology.4 edn.Revinter,2000.

[3] Simon Haykin, Neural Networks: A Comprehensive Foundation.2 edn. Bookman,2007.

[4] Chistina R. Crecca, Adrian E. Roitberg. Using Distances Between α-Carbons to Predict Protein Structure. Wiley InterScience.2008.

Computational analysis of small RNAs libraries of sugarcane cultivars submitted to drought stress

Flávia Thiebaut[1*], Clícia Grativol[1], Cristian A. Rojas[1], Renato Vicentini[2], Adriana S. Hemerly[1], Paulo C. G. Ferreira[1]

[1] Laboratorio de Biologia Molecular de Plantas, Instituto de Bioquímica Médica Universidade Federal do Rio de Janeiro, Avenida Brigadeiro Trompowski Ed. CCS, 21941-902, Rio de Janeiro, RJ, Brazil. [2] Laboratório de Bioinformatica e Biologia de Sistemas, Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Campinas, SP, Brazil.
*thiebaut@bioqmed.ufrj.br

Abstract. Small non-coding RNAs in plants have been investigated for their important function as post-transcriptional regulators. Computational characterization of sRNAs from plant species, whose genomes are not yet sequenced has been increased in recent years. In the present study, we have used deep sequencing data to identify miRNAs and siRNAs expressed in sugarcane roots submitted to drought stress. We found 6,971,203 reads in sRNA libraries and these were classified into known mature, new miRNAs and siRNAs. These findings broaden the scope of understanding the gene regulation through sRNA in sugarcane subjected to drought stress.

Key words: Sugarcane, sRNAs, deep sequencing

1 Introduction

A novel system of gene regulation has emerged recently, the mechanism of RNA interference [1]. Key components of this regulatory process are denominated small non-coding RNAs (sRNAs) [2]. These sRNAs are classified into microRNAs (miRNAs) and small interfering RNAs (siRNAs), according to the biogenesis pathway and manner of action [3]. They play a role in regulating the expression of messenger RNA-target by cleavage or repression of translation, and methylation of DNA-target, respectively [4]. Both are small endogenous RNAs 20-25 nucleotides in length [5].

The siRNAs are not phylogenetically conserved, in contrast with the majority of miRNAs. Based on sequence similarity, the miRNAs were clustered in families [6]. MiRNA families are composed of secondary structures that can be result in identical or very similar mature miRNAs, differing by the number of family members [7]. MiRNA-guided gene regulation is essential for normal growth and development of the plant [8], as well as for adaptation to stress conditions [9]. Several studies have reported the involvement of plants miRNAs in response to stress conditions, and numerous microRNA targets are genes related to stress response [10; 2]. The number

tends to grow with advanced sequencing techniques that have been used to plant small RNAs library construction.

Sugarcane has one of the most complex plant genomes with a variable ploidy number [11]. The sRNAs analysis in sugarcane subject to drought stress may help to better understand the molecular mechanisms of response of this crop and related species, and contribute to breeding programs. Our goal was to study the involvement of sRNAs, specially the miRNAs, in the drought tolerance in sugarcane root.

2 Material and Methods

2.1 Plant materials, sRNA library construction and deep sequencing

Stalks of sugarcane cultivars, with different drought sensitivities, were germinated and grown in a greenhouse. After three months, the plants were exposed to drought stress. Treated and control roots were collected after 0 and 24 hours of treatment, respectively. Total RNA from each sample was extracted with Trizol reagent (Invitrogen®, USA) as described by the manufacturer.

A total number of four sRNA libraries for deep sequencing were prepared from RNA pool of sensitive and tolerant sugarcane cultivars submitted to drought stress and control plants. The sRNA libraries were sequenced by Illumina/Solexa. in the Cold Spring Harbor Laboratory, USA.

2.2 Processing of deep sequencing data

The four libraries of sequencing reads were parsed to remove the 3'-adaptors. The libraries were subjected to removal of other RNAs (rRNA, tRNA, Rfam, degradation, etc). The unique sequence obtained were mapped to the *Sorghum bicolor* genome sequences (http://www.plantgdb.org), private *Saccharum officinarum* EST database, to the plant repeat databases (http://plantrepeats.plantbiology.msu.edu/) and were aligned to the miRNAs in miRBase (release 13.0, http://microrna.sanger.ac.uk), with BLASTn. The reads that match to these sequences with 0-4 mismatch were retained for further analysis. The sequences that closely matched the previously known plant mature miRNAs were included in the set of miRNA candidates and those were classified in miRNA families.

3 Results and Discussion

A total of 6,971,203 reads were obtained from four sRNAs libraries. Raw sequence reads were parsed to remove the adaptors and grouped according to number the unique sequences. After that, we obtained 1,031,824 unique sequences that were treated to removal others RNAs, resulting in 993,406 unique sequences. Most mature miRNAs are evolutionarily conserved among species within the plant kingdom. This information enables us to computationally predict new miRNA homologs or orthologs in different plant species. Therefore, we used all previously known plant mature miRNAs from miR registry to search for homologs of miRNAs in the publicly available miRBase, *Sorghum bicolor* genome and *Saccharum officinarum* EST database. With unique sequences mapping against publicly available database, we obtained 792,907 no hits sequence, 189,554 new miRNAs candidates and 1,399

known miRNAs grouped in 30 families. Moreover, the mappings against Plant Repeat Databases show 9,546 siRNAs candidates. Figure 1 shows filtering procedure and mapping results.

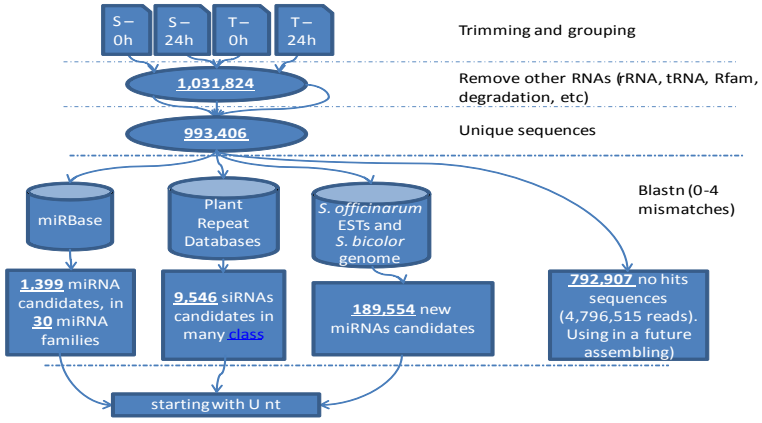


Fig. 1. Summary of results obtained after filtering procedure and mapping. S= sensitive sugarcane cultivar and T= tolerant sugarcane cultivar

After the classification of sequences in known miRNA, new miRNAs candidates, siRNAs and specific siRNAs (siRNAs candidates that show no matches with *S. bicolor* were classified as *S. officinarum* specific), we verified the first nucleotide preference of each group. The results show that new miRNAs, siRNAs and specific siRNAs have Adenine (A) as first nucleotide in main sequences while conserved miRNAs have Uracile (U) (Fig. 2). Recent studies reported that most known miRNAs have U in the first nucleotide of the sequence what was corroborated with the result showed here. This characteristic is mainly due the action pathway of miRNAs, where the AGORNAUTE 1 (AGO1) recognizes specifically the first nucleotide in the sequence [12].

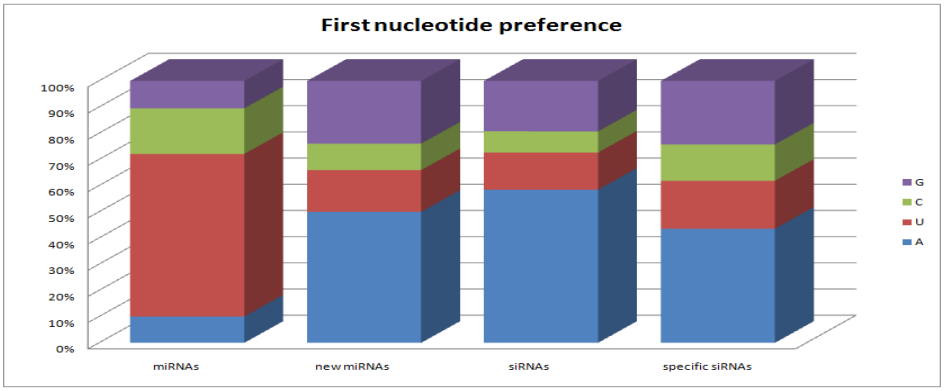


Fig. 2. The percentage of first nucleotide preference of known miRNAs, new miRNAs, siRNAs and specific siRNAs (A= adenine, U= uracile, G= guanine and C= cytosine).

4 **Flávia Thiebaut1*, Clícia Grativol1, Cristian A. Rojas1, Renato Vicentini2, Adriana S. Hemerly1, Paulo C. G. Ferreira1**

Exhaustive small RNA sequencing has been performed in plants using next generation sequencing. These analyses show that each species of sRNAs presents a sequence accumulation of determinate size. miRNA size ranges from 20 to 24 nt, and siRNA size ranges from 21 to 24 nt, being the 21nt miRNAs and 24nt siRNAs sequences most abundant [13]. As shown in figure 3, the sRNAs sequence size distribution before the trimming and filtering procedure presented 21nt and 24nt species in a greater amount. When we analyzed the unique sequences size distribution was seen that the 24nt species was most abundant. These results suggest that the miRNAs are most express presenting many repetitive reads. After the trimming and filtering of ambiguous sequences, this 21nt species was decreased.

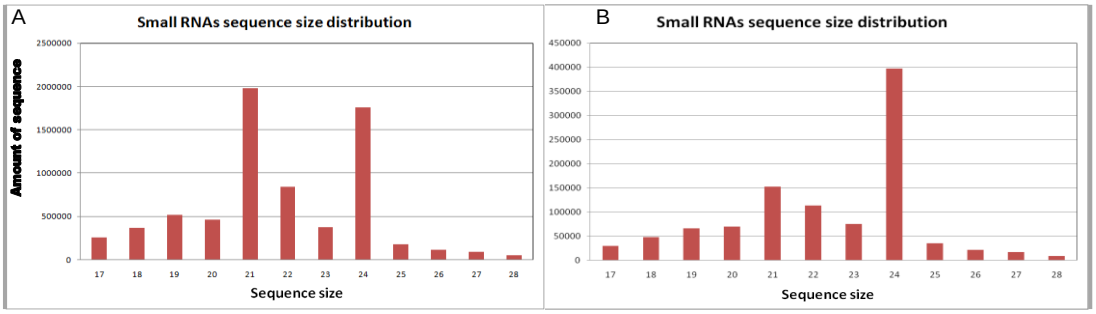


Fig. 3. Small RNAs sequence size distribution. A total of sRNAs generated on the libraries (A) and non-redundants sRNAs (B).

Another feature of sRNAs is the classification the sequences in families. This feature is unique to miRNAs for being phylogenetically conserved. The family classification is based on the miRNA precursor sequence which can lead to identical or similar matures miRNAs [7]. Figure 4 shows the classification of miRNAs in families' the miRNAs identified in the libraries. Known miRNAs were classified in 30 families (miR408, miR435, miR528, miR535, miR1432, miR160, miR415, miR827, miR1436, miR415, miR530, miR414, miR395, miR172, miR399, miR437, miR529, miR397, miR398, miR390, miR393, miR396, miR444, miR167, miR164, miR169, miR319, miR171, miR166, miR156, miR159 and miR168). The most abundant families in sugarcane roots were miR168 (with 350 members) and miR159 (with 305 members).

In conclusion the results are based on the computational approach for sRNAs identification from plant species whose genome is not yet sequenced. We have identified known mature and new miRNAs and siRNAs expressed in sugarcane roots submitted to drought stress. This is a first step towards the identification of new miRNAs in sugarcane and further precursor studies may verify these analysis. Thus, the identification of sRNAs differential expression can serve as an initial point for the characterization of gene regulation pathways involved with drought stress in sugarcane, an important economic crop.

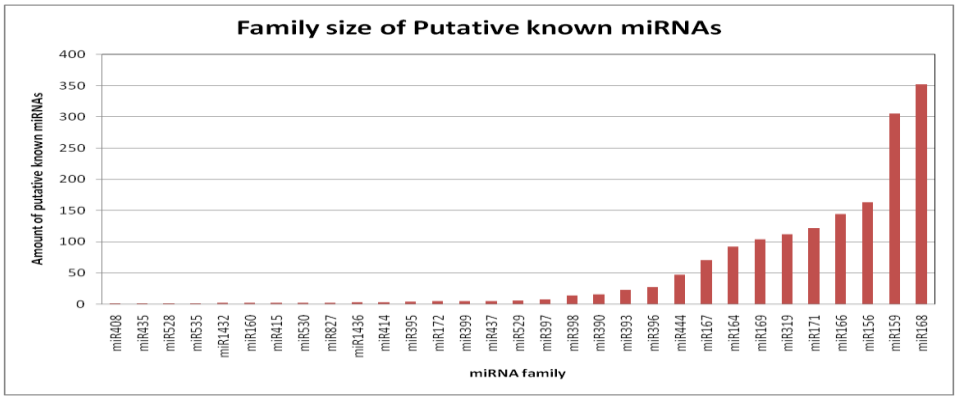


Fig. 4. Classification and amount of family members of putative known miRNAs.

4 References

1. Tijsterman M., ketting R.F. & Plasterk R.H. (2002) The genetics of RNA silencing. *Annual Review Genetics* 36, 489-519.
2. Phillips J.R., Dalmay T., Bartel D. (2007) The role of small RNAs in abiotic stress. *FEBS Letters* 581, 3592-3597.
3. Ghildiyal, M., Zamore, P.D. (2009) Small silencing RNAs: an expanding universe. *Nature Reviews Genetics* 10, 94-108.
4. Ramachandran, V., Chen X. (2008) Small RNA metabolism in Arabidopsis. *Trends Plant Science* 13, 368-374.
5. Zhang B., Pan X., Cobb G.P. & Anderson T.A. (2006) Plant microRNA: A small regulatory molecule with big impact. *Development Biology* 289, 3-16.
6. Sunkar R. & Zhu J.K. (2004) Novel and stress-regulated microRNAs an other small RNAs from *Arabidopsis. Plant Cell* 16, 2001-2019.
7. Jones-Rhoades M.W., Bartel D.P., Bartel, B. (2006) MicroRNAs and their regulatory roles in plants. *Annual Review of Plant Biology* 57, 19-53.
8. Rubio-Somoza I., Cuperus J.T., Weigel D. & Carrington J.C. (2009) Regulation and functional specialization of small RNA-target nodes during plant development. *Current Opinion in Plant Biology* 12, 622-627.
9. Zhou X., Wang G., Sutoh K., Zhu J.K. & Zhang W. (2008) Identification of cold-inducible microRNAs in plants by transcriptome analysis. *Biochimica et Biophysica Acta* 1779, 780-788.
10. Jones-Rhoades M.W. & Bartel D.P. (2004) Computational identification of plant microRNAs and their targets including a stress-induced miRNA. *Molecular Cell* 14, 787-799.
11. Ingelbrecht I.L., Irvine J.E. & Mirkov E. (1999) Posttranscriptional gene silencing in transgenic sugarcane. Dissection of homology-dependent virus resistance in a monocot that has a complex polyploidy genome. *Plant Physiology* 119, 1187-1197.
12. Baumberger N.; Baulcombe D.C. (2005) Arabidopsis ARGONAUTE1 is an RNA slicer that selectively recruits microRNAs and short interfering RNAs. *Proceedings of the National Academy of Sciences* 102, 11928-11933.
13. Vaucheret H. (2006) Post-transcriptional small RNA pathways in plants: mechanisms and regulations. *Genes & Development* 20, 759-771.

Development of a fully-flexible receptor-based ligand filter to accelerate virtual screening

Christian V. Quevedo[1,2,4], Ivani Pauli[1,3,5], Osmar Norberto de Souza[1,3-5] and Duncan D. Ruiz[2,4]

[1]Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas – LABIO,
[2]Grupo de Pesquisa e Inteligência de Negócios – GPIN,
[3]Instituto Nacional de Ciências e Tecnologia em Tuberculose – INCT-TB,
[4]Faculdade de Informática e [5]Biociências, Pontificia Universidade Católica do Rio Grande do Sul – PUCRS, Av. Ipiranga, 6681, Sala 32/628, 90619-900, Porto Alegre, RS, Brasil

{christian.quevedo, ivani.pauli}@acad.pucrs.br, {osmar.norberto, duncan.ruiz}@pucrs.br

Abstract. Public databases provide over 20 million ligands to users. *In silico* testing with this high volume of data is computationally very expensive. Researchers have been seeking for solutions aimed at reducing the number of ligands to be tested on their target receptors. However, there still is no method to effectively reduce this large number into a manageable one. This is a major challenge of rational drug design. This work presents the current state of a heuristic function we are developing to perform virtual screening (VS) with available ligands, and to filter the most promising candidates. This function is based on the receptor's substrate binding cavity geometry. The heuristic should filter only the ligands compatible with the cavity, derived from a fully-flexible model of the receptor. This filter is expected to greatly improve VS with molecular docking by avoiding ligands that do not fit in the receptor's substrate binding cavity.

Keywords: Rational drug design, databases of ligands, ligand filtering.

1 Introduction

Current virtual screening (VS) protocols of ligands against a target receptor can be a very time consuming task with low chances of finding good receptor inhibitor candidates. To reduce costs and optimize drug development, researchers worldwide are working with the goal of reducing the time invested in selecting ligands, thus reducing the amount of ligands to be tested in receptor-ligand molecular docking experiments. For instance, C. Lipinski [1], introduced a heuristic method to classify the possible success or failure of a ligand based on the presence of features in molecules already approved in the early stages of clinical testing *in vivo*, reducing the number of ligands candidates to be an inhibitor. Although this method is widely used, it can generate many inaccurate results [2]. Moreover, most existing heuristics consider only the properties of the ligands to select candidates without considering the characteristics of the target receptor. Therefore, development of new methods to filter

large amount of binders and, simultaneously, provide a high rate of success, considering the binding site properties of the flexible target receptor, is still a problem to be solved.

2 Materials and Methods

To develop of a filter capable of determining the chances of a ligand to fit into a receptor binding cavity, based on its geometric features, first it is necessary to identify the atoms that define the structure of the binding cavity. Currently, there are several programs that can identify cavities in macromolecules using geometry-based criteria. CASTp [3] is one of these software. It processes a PDB file of a target receptor and calculates the areas and volumes of all the cavities present in the receptor. Additionally, it provides a list of the receptor's residues that make up each identified cavity. However, we treat the receptor not as a single, rigid crystal structure, but rather as an ensemble of conformations obtained from a molecular dynamics (MD) simulation of the receptor [4]. We call this representation a fully-flexible-receptor (FFR) model. Hence, we developed a computer program to submit the FFR model to and retrieve its results from web server. This program, divided in two parts, was developed in Ruby programming language to submit PDB files of the FFR model, and a C program to process the CASTp results.

To validate our experiments we consider the InhA enzyme from *Mycobacterium tuberculosis*. It was chosen because it represents an attractive target for developing new anti-tuberculosis drugs. InhA is a highly flexibility protein. In the present work, the InhA FFR model was based on 3,100 conformations or snapshots obtained from its MD simulation [4]. Herein it will be called the InhA_FFR model.

Empirical studies have shown that binding sites correspond to the largest cavities found in proteins [5-7]. To evaluate its major binding cavities geometries we performed two analyses at CASTp: one that calculated the major binding site cavity and another that calculated only the substrate binding cavity. The major binding cavity of InhA comprises the coenzyme and substrates' binding cavities, respectively. Fig. 1 shows the volume variations of the major binding cavities of the InhA_FFR model.

It is important to note that some of snapshots of the InhA_FFR model have volumes that can be considered *outliers*. In spite of this, we used all conformations generated because, currently, there is no heuristics to discard conformations. Here we concentrate on the substrate binding cavity since we expect to find ligands capable of binding in this cavity. By processing the CASTp results for the InhA_FFR model substrate binding cavity we found an average of 72 different atoms (from different residues) that determines this cavity. This value is too high to apply a brute force algorithm to test out each possible binding position of a ligand. As a consequence, it is essential to develop heuristics to reduce the number of atoms used to determine the binding site cavity.

To identify which are the atoms/residues most relevant in the substrate cavity, we performed a detailed study, analyzing the 33 structures of InhA currently available in the PDB. Using the software LIGPLOT [8], interactions between residues of the

receptor and different ligands conveniently positioned in the substrate cavity were calculated.

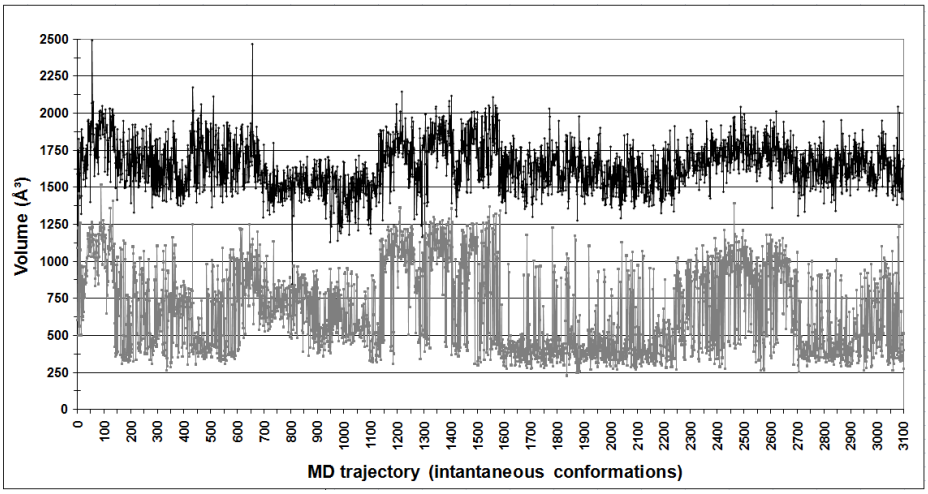


Fig. 1. InhA_FFR model binding cavities' analyses with the CASTp [3]. The binding cavities' volumes are shown as a function of the instantaneous conformations along the MD simulation trajectory [4]. The major binding cavity's volume (in black) has an average of 1,647 Å³. A similar measure for the crystal structure (PDB ID: 1ENY) equals to 1,657 Å³. The substrate binding cavity's volume (dark grey) averages 674.56 Å³ in the InhA_FFR model while it equals to 598.20 Å³ in the crystal structure.

Table I shows the residues that had at least 5% of all possible interactions established within the substrate binding cavity. Nine amino-acids residues were selected, amounting to 76 heavy atoms. The intersection of atoms from this analysis and those obtained with CASTp reduced the number of atoms defining the substrate binding cavity from 76 to 60 and then to 23.

Table 1. Analysis of the 33 InhA PDB structures. Frequency of interactions between residues of the InhA-NADH complex with ligands located in the substrate binding cavity.

Residue	Interactions established (separated structure)	Interactions established with drug (total)	Interactions established with drug (%)
TYR158	13(O) + 4(D) + 1(S)	18	10,53
GLY96	12(O) + 4(D)	16	9,36
MET103	12(O) + 3(D) + 1(S)	16	9,36
PHE149	10(O) + 4(D) + 1(S)	15	8,77
MET199	9(O) + 4(D) + 1(S)	14	8,19
ILE215	9(O) + 3(D) + 1(S)	13	7,60
LYS165	7(O) + 4(D)	11	6,43
PHE97	7(O) + 4(D)	11	6,43
ALA157	8(O) + 1(S)	9	5,26

The second column is separated by the number of interactions with the type of structure:
O: Other ligands D: Derivatives of triclosan S: Substrate

3 Conclusion and future work

The heuristic function to filter the ligands is still under development. So far, in this work we identified the geometric properties of the InhA_FFR model binding cavities. From these identifications, the next step is to develop a methodology capable of comparing the geometric features of the InhA_FFR model substrate binding cavity with ligands 3-D coordinates, in order to test out the geometric fitting between them.

Acknowledgments. CVQ and IP are supported by CAPES M. Sc. fellowships. ONS is a CNPq Research Fellow.

References

1. Lipinski, C.A.: Drug-like properties and the causes of poor solubility and poor permeability. J. Pharmacol. Toxicol. Methods, 44, 235-249 (2000)
2. Kadam, R.U., Roy, N.: Recent Trends in Drug-Likeness Prediction: A Comprehensive Review of In Silico Methods. Indian J Pharm Sci, 69, 609-615 (2007)
3. Liang, J., Edelsbrunner, H., Woodward, C.: Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci, 7, 1884-1897 (1998)
4. Schroeder, E.K, Basso, L.A., Santos, D.S., Norberto de Souza, O.: Molecular dynamics simulation studies of the Wild-Type, I21V, and I16T Mutants of Isoniazida-Resistant Mycobacterium tuberculosis Enoyl Reductase (InhA). In Complex with NADH: Toward the Understanding of NADH-InhA Different Affinities, Biophys J, 89, 876-884 (2005)
5. Laskowski, R.A., Luscombe, N.M., Swindells, M.B., Thornton, J.M.: Protein clefts in molecular recognition and function. Protein Sci, 5, 2438-2452 (1996)
6. Sotriffer, C., Klebe, G.: Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. Farmaco, 57, 243-251 (2002)
7. Campbell, S. J., Gold, N.D., Jackson, R.M., Westhead, D.R.: Ligand binding: functional site location, similarity and docking. *Curr* Opin Struct Biol, 13, 389-395 (2003)
8. Wallace, A.C., Laskowski, R.A., Thornton, J.M.: LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng, 8, 127-134 (1995)

Human Disease: domain ontology to simulate biological models

Daniele Palazzi[1,2], Ely Edison Matos[1,2], Fernanda Campos[1,2], Regina Braga[1,2], Elaine Coimbra[3]

[1]Software Quality Research Group
[2] Master Program in Computational Modeling
[3] Department of Parasitology and Microbiology
Federal University of Juiz de Fora
36036-900 – Juiz de Fora – MG – Brazil
{daniele.palazzi,ely.matos, fernanda.campos, regina.braga, elaine.coimbra}@ufjf.edu.br

Abstract. This paper describes CelO - Human Disease ontology and its development process using QDAontology - Quality Driven Approach for e-Science Ontologies. This approach is composed of stages, activities, participants, artifacts and quality criteria. Ontology artifacts illustrate the stage documentation. CelO ontology captures both structure of a cell model and properties of functional components. We use this ontology in a Web project (CelOWS) to describe, query and compose CellML models, using semantic web services.

Keywords: Ontology, Human disease, Biological Models.

1 Introduction

Distributed data, computation, models and instruments at unprecedented scales are changing researchers way of working. The analysis of large amounts of widely distributed data is becoming commonplace. Experimental apparatus and simulation systems are collaboration activities and data is shared among researchers. Model simulation tools should orchestrate the steps of scientific discovery. Disciplines like astronomy, biology, chemistry, environmental science, engineering, geosciences, medicine, physics, and social sciences are benefiting from the use and development of eScience frameworks to enhance simulation.

Our research aims at using and discussing ontologies and semantic rules for creation, validation, storage and sharing of biological models, through the use of a framework named CelOWS [3]. It is based on Service Oriented Architecture (SOA) together with semantic annotations provided by an ontology named CelO (Cell Component Ontology). It allows the storage, search, reuse, composition and execution of described models using the CelO ontology. The main goal of the CelOWS is composition of a new model from reuse of different models from a specific repository.

As ontology development needs a well defined process, and considering our work in eScience projects we defined QDAontology - Quality Driven Approach for e-Science Ontologies [2]. It is composed of six stages. For each stage the model describes (i) participants, (II) activities, (III) artifacts (IV) quality criteria. In this paper we present CelO – Human Desease ontology development and describe how it fits a tool to simulate biological models. The rest of the paper is organized as follows: Section 2 presents CelO-Human Disease ontology and its development process. In Section 3, we highlight some conclusions.

2 Cell Component Ontology

CelO ontology [20] [5] was developed in two cycles: the first domain is related to representation of biological models, mainly the electrophysiology cell models. For second version an expansion was made adding the sub-domain of human disease cells. QDAontology approach considers an ontology engineering process with all components. For each stage the model describes (i) participants, (II) activities, (III) artifacts and (IV) quality criteria. There are six stages: Specification, Conceptualization, Formalization, Implementation, Integration and Evolution. Each cycle (all round up stages) ends up with a prototype version. Stages are composed of specific activities, which generate specific artifacts, as milestones of the development process. Activities can integrate different stages of the same cycle. Artifacts are responsible for documentation and they improve quality of development process. Moreover, a description of involved participants and their roles in different stages and activities are described. Alignment mechanism for interoperability between ontologies was selected to attend the integration step. A tool was developed for OBO and OWL conversion that follows three steps: terms comparison, in the lexical level, and then application of a similarity measure. CelO ontology documentation is available at http://celo.mmc.ufjf.br.

Stage 1- Specification

This stage aims to generate a document that identifies the objectives, specifies users, set of terms to be represented, characteristics and the granularities of the ontology to be developed. It is also time to identify existing similar ontologies and/or complementary domains and decide about the implementation. In this stage a set of domain questions must also be specified to limit the target area.

Stage 2 - Conceptualization

Conceptualization organizes and structures the domain knowledge as significant models using external representations. As participants we have the Ontological Engineer and the Stakeholder. This stage presents the activities Knowledge Acquisition, Evaluation, Documentation, Configuration Management, Planning, Quality and Environment. Glossary and dictionary of terms are the most important artifacts generated in this stage, including domain concepts, class instances and attributes, synonymous and acronymous.

Stage 3 - Formalization

Formalization changes conceptual model into a formal or half-computable model, what makes possible the requirement transformation and the definition of terms in

ontology project, suggesting an architectural model and adapting the project for implementation. In this stage activities are supported by developer that builds the ontology in a formal language, implementing what was specified. Moreover, ontological engineer and stakeholder also participate. The activities that integrate the stage are: Knowledge Acquisition, Evaluation, Documentation, Configuration Management, Quality and Environment. Artifacts generated in this stage are presented in a classification tree of the concepts and in a diagram of partial relationships,that allow an intermediate representation of the domain. In these diagrams the relations are shown and are defined as "is one", "is a type of", "it is part of" or others that better represent the domain.

Stage 4 - Implementation

Implementation builds computable models using appropriate tools for the development of ontologies. It defines formal organization of terms, concepts and relations to allow reasoning, through the definition of rules and constrains. The result is a codified ontology in a formal language, ready to be used in different applications. The following activities are part of this stage: Evaluation, Documentation, Configuration Management, Quality and Environment. In this stage, the artifact is the proper ontology, in a formal language. Protégé tool was used.

Stage 5 - Integration

Integration considers reuse and sharing of the ontology with other ones from the same domain. The interoperability between ontologies is an essential factor and this stage guarantees sharing and exchanging of information among applications. Techniques of ontological mapping must be used, as mapping, alignment, combination and integration. This stage is concerned with the reuse of ontologies and mapping between them. The activities are: Evaluation, Documentation, Configuration Management, Quality and Environment. As described before we developed a special tool for that, not considered in this paper.

Stage 6 - Evolution

Evolution has as main goal to support the knowledge enrichment of ontology, adding new concepts or expanding existing ones with new acquired knowledge. It allows not only the ontology enrichment but, also, its expansion and maintenance, when necessary. It is supported by the activities: Evaluation, Documentation, Configuration Management, Planning, Quality and Environment. To represent the ontology evolution, graphical representations can be used. At this step a final version with the evolution is ready (figure 1).

3 Conclusions

Cells are basic units to human body and functional elements are understood as components of the whole cell. The increasing volume and distribution of data and processes in Bioinformatics allowed the discovery of new biological information an easier process. Build ontology in this area is part of an international effort to understand proposed components models in different complexity levels. Considering this aspect, we developed Cell Component Ontology (CelO) to describe semantics of biological models. We use CelO in CelOWS tool to help scientists to compose cell models trough CellML components. Semantic level and ontology documentation are

important to exchange information among groups involved in similar research. Our research considers the context of infrastructures for e-Science using innovative technologies, based on well established standards. Future works will focus on integration of biological domain ontologies in scientific workflows, as a semantic solution.

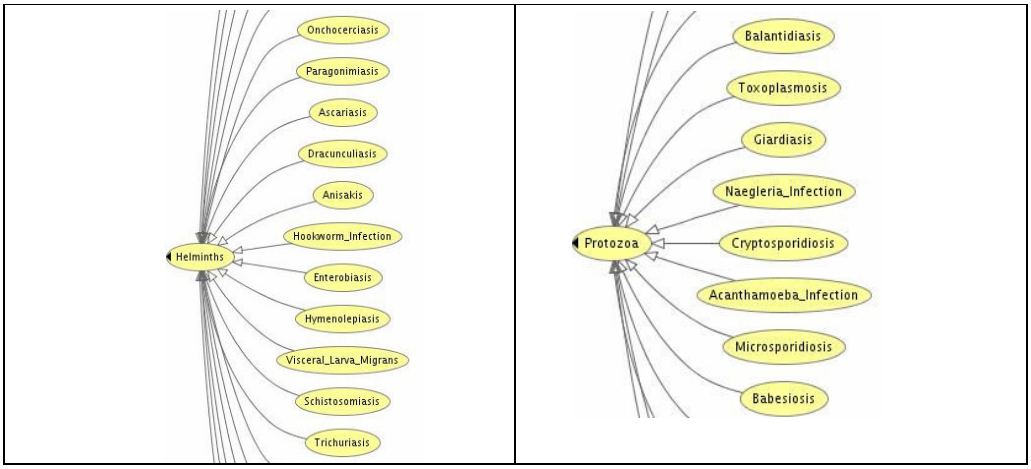


Fig. 1. Part of the CelO- Human Disease ontology.

Acknowledgement: This work has been partially supported by FAPEMIG.

References

1. CellML. 2008. "CellML – Cell Markup Language". http://www.cellml.org. Retrieved 2009/03/30.
2. Palazzi, Daniele ; Matos, E. E. S. ; Campos, F. ; Braga, R . Development Approach for Modeling Biological Ontologies. In: Joint 5th International Workshop on Vocabularies, Ontologies and Rules for The Enterprise (VORTE) - International Workshop on Metamodels, Ontologies and Semantic Technologies (MOST), 2010, Vitoria.
3. Matos, E. E. S., Campos, F., Braga, R., Palazzi, D. CelOWS: a ontology based framework for the provision of semantic web service related to biological models. Journal of Biomedical Informatics, v. 43, p. 125-136, 2009.Matos, E. E. S., Campos, F., Braga, R., Palazzi, D. CelOWS: a ontology based framework for the provision of semantic web service related to biological models. Journal of Biomedical Informatics, v. 43, p. 125-136, 2009.4 BibTeX Entries

Identification and Classification of ncRNAs in Trypanosoma cruzi: A Multistep Approach

Priscila Grynberg[1], Mainá Bitar[2], Alexandre Paschoal[3], Alan M. Durham[3], Glória R. Franco[1]

[1]Department of Biochemistry and Immunology,
Federal University of Minas Gerais, Belo Horizonte, MG, Brazil.
[2]Biophysics Institute, Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil.
3Mathematics and Statistics Institute, University of São Paulo, São Paulo, SP, Brazil.

maina@biof.ufrj.br

Abstract: Non-coding RNAs (ncRNAs) prediction has become a vast field of research and several classes of ncRNAs with different regulatory, catalytic and structural functions have been discovered. We intend to predict and classify non-coding rnas (nc-rnas) from Trypanosoma cruzi, the causative agent of Chaga's Disease. Besides the application of known methods for in silico nc-rnas prediction and classification, we also intend to propose a new method based on energy parameters to predict and classify putative novel nc-rnas with no evidence in searches against nc-rnas databases.

Keynotes: Trypanosoma cruzi, non-coding RNA, in silico prediction

1 Introduction

Non-coding RNAs (ncRNAs) prediction has become a vast field of research and several classes of ncRNAs with different regulatory, catalytic and structural functions have been discovered. To enable these new molecular characterizations, computer-based techniques were developed and coupled with experimental designs [1]. Trypanosomatids as Trypanosoma cruzi, Trypanosoma brucei and Leishmania sp. are the etiologic agents of high-incidence tropical diseases. Considering that genes in trypanosomatids are transcribed as polycistronic units, ncRNAs are likely to have fundamental roles in gene expression mechanisms.

Few years ago, three kinetoplastid genomes have been finalized, and a recent study to predict ncRNAs in Leishmania braziliensis and T. brucei has been published [2]. Similarly, we propose to predict and classify ncRNAs for the complete genome of T. cruzi. For this purpose, we used eQRNA [3], an algorithm for comparative analysis of biological sequences that performs probabilistic inference on genomic alignments. eQRNA identifies conserved sequences that show mutational patterns consistent with a preserved RNA secondary structure, as opposed to conserved coding frames and/or other genomic features.

2 Results and Discussion

The entire genomes of T. brucei and T. cruzi were used to generate the initial alignments submitted to eQRNA, and 4195 ncRNA candidate sequences equal to or longer than 30 nucleotides were found. The candidate sequences were used for blastx search (e-value = 10e-05) against T. cruzi annotated proteins. 2813 candidates matched protein-coding sequences and the remaining 1382 candidates were submitted to a pipeline that included search against 25 different ncRNA databases, ab initio RNA tools and structural analysis. 1301 candidates had no evidence to be classified as ncRNAs and 49 candidates were tRNAs or rRNAs. Twenty-nine candidates presented similarity with ncRNAs from several databases. Three were considered false-positives. This methodology is schematized (Figure1).

For cases where no functional prediction was accomplished, an energy-based approach to characterize native-like structures of ncRNAs was applied. This new methodological tool takes into account that native-like structures of macromolecules are likely to present a lower energy than structures generated at random with no biological function. To test this statement in the context of ncRNAs, we developed a software tool coupled with the Vienna RNA package [4] for energetic assessment of RNA molecules. The new approach generates a set of random RNA sequences, carrying the same base composition as the original RNA sequence under analysis, but differently organized. Two-dimensional structures are produced for each of those random sequences and their energy is evaluated with RNA-fold from Vienna package [4]. Whenever the mean value of energies associated with the set of random structures is higher than the value obtained for the ncRNA under study, we consider this a native-like structure, with possible biological function. An additional strategy aims to classify ncRNAs according to their RNA family. This computer-based tool consists of an energetic map that takes into account the length of the RNA sequences, and their RNA family. This map is generated from a 2D graph of length vs. energy, with points colored by RNA family type. All sequences used in this methodology were retrieved from RNAstrand [5]. It is clear that RNA families form clusters in this graph, generating an RNA family classification map, were a recently identified RNA sequence can be classified according to its type. Both methodological approaches for ncRNA classification are under analysis concerning their efficiency.

3 Conclusions and perspectives

Preliminary tests point to an appropriate performance of these new computer-based methodologies. Other in silico approaches which make use of energy parameters will be employed to test the validity of the identified ncRNAs and to group these sequences according to family.

Our next goal is to identify putative regulatory ncRNAs that may be directed to UTR elements by matching them to a catalog of 5' and 3' UTR sequences of T. cruzi transcripts retrieved from dbEST.

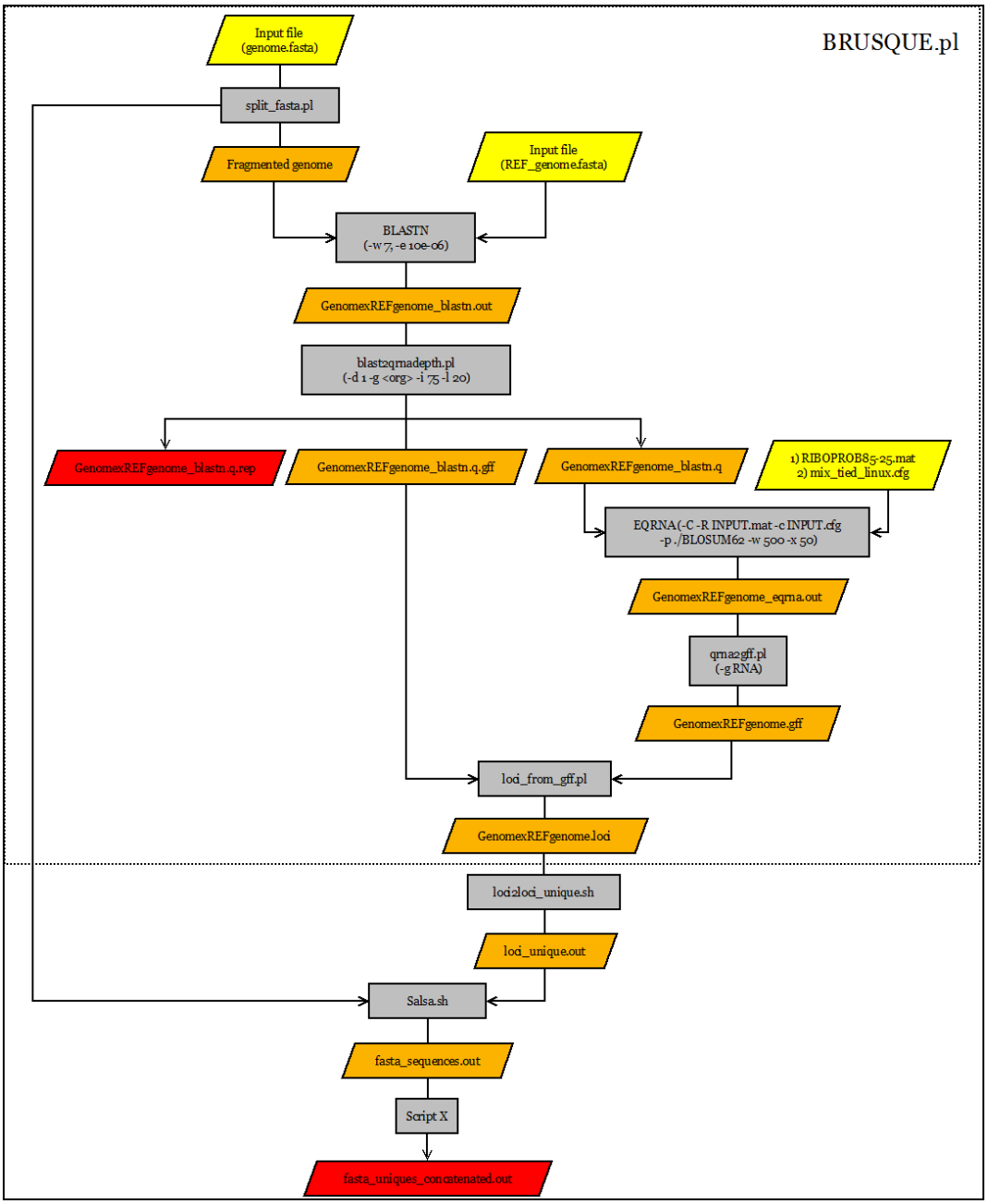


Figure1. Annotation pipeline

References

1. McCutcheon JP, Eddy SR: Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nucleic Acid Research, 2003. 31(14): 4119-4128
2. Mao Y, Najafabadi SH, Salavati R: Genome-wide computational identification of functional RNA elements in Trypanosoma brucei. BMC Genomics, 2009. 10:355-366
3. Yu N, Cho KH, Cheng Q, Tesorero RA: A hybrid computational approach for the prediction of small non-coding RNAs from genome sequences. International Conference on Conputational Science Engineering (IEEE), 2009. 2:1071-1076.
4. Zuker M, Stiegler P: Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. Nucleic Acid Research, 1981. 9: 133-148
5. Andronescu M, Bereg V, Hoos HH, and Condon A: RNA STRAND: The RNA Secondary Structure and Statistical Analysis Database. BMC Bioinformatics, 2008. 9(1):340.

Improving Biomarker Identification through Ensemble Feature Rank Aggregation*

Ronaldo C. Prati

Centro de Matemática, Computação e Cognição
Universidade Federal do ABC - UFABC
ronaldo.prati@ufabc.edu.br

Abstract. Feature selection techniques have been used for biomarker identification for a long time. However, considering the heuristic nature of these techniques and sampling variations due to low sample size common in bioinformatics studies, this task is known to be unstable. In this paper we investigate the use of different rank aggregation methods to construct ensembles of feature selection techniques, aiming to overcome the stability of the techniques. Preliminary results on miRNA sequences show the effectiveness of the approach.

1 Introduction

Recent advances in biological studies allow the discovery of biomarkers for diagnosis of complex diseases at the molecular level. A biomarker may be defined as "a characteristics that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [1]. In bioinformatics, the problem of biomarker identification is often framed as a feature selection task, where the objective is to select the most discriminating features for classification.

Although numerous different feature selection algorithms are available, they do not necessarily produce the same subset of candidate features if we apply them to the same dataset. Moreover, even the same feature selection algorithm applied to slight different random samples of a dataset do not produce the same subsets. In bioinformatics, this problem is worsened due to the low sample size common in most domains. Ensemble techniques were often used in machine learning to improve classification accuracy. Recently, ensembles of feature selection algorithms have been proposed to overcome the instability of feature selection methods in the context of biomarker detection [2, 3].

This paper compares three approaches for ensemble feature selection based on ranking aggregation. Two of them have been previously used in this context [4, 5], but we are not aware of a comparison of them. Furthermore, we are not aware of the use of the third method in this context. Experiments were conducted using a dataset of microRNA expression. The evaluation used three different learning algorithms and three different performance metrics. Results show the suitability of using rank aggregation for improving biomarker identification.

* This work is supported by CNPq

2 Rank aggregation for ensemble feature selection

The rank aggregation problem is to combine many different rankings on the same set of alternatives in order to obtain a combined "consensus" rank with (hopefully) a better ordering. More formally, a ranking aggregation method takes as input a set of k ranked lists $R = \{R_1, R_2, \ldots, R_k\}$ and produces as output another ranking R_A. $R_i(1)$ is the highest ranked (Rank 1) object in the list i and in general $R_i(m) = o_j$ is to say that object o_j has rank m in the list R_i.

Ensemble feature selection methods based on rank aggregation take as input different base-rankings of features and combine them in order to obtain a better final ranking of features. Many feature selection algorithms use feature ranking as a principal or auxiliary step because of its simplicity, scalability, and good empirical success [6]. Therefore, we have a large set of options in order to construct the base-rankings.

Numerous ranking aggregation methods have been proposed. In this paper, we evaluate three of them in order to compute the aggregate feature ranking, namely Average Rank (AVG), Markov Chain (MC), and Schwartz Sequential Dropping (SSD). AVG computes the average rank through all base-rankings. The final ranking is then constructed based on the average ranking. It has been used in [4] to construct ensembles for biomarker identification. For MC, the consensus ranking is found as the stationary distribution of an appropriately defined Markov chain over the set of features. The states of the chain correspond to the features and the transitions among the states are proportional to the sum of the distance in the ranked list among all base-rankings. A procedure similar to PageRank is used to find the stationary distribution. It has been used in [5] to construct ensembles for biomarker identification. SSD computes the aggregate ranking by sequentially computing the Schwartz set. A feature A beats feature B if A appears above B in all base-rankings. The Schwartz set is the innermost unbeaten set, or the smallest set of features such that any feature outside the set beats none inside. If no defeats exist among the Schwartz set, then its members are the winners (plural only in the case of a tie, which must be resolved by another method). Otherwise, drop the weakest defeat among the Schwartz set, determine the new Schwartz set, and repeat the procedure. In the aggregate ranking, features are ranked according to the order they appear as winner in the Schwartz sets. To our knowledge, SSD has not been used for ensemble feature construction for biomarker discovery.

3 Results

Experiments were conducted using miRNAcancer dataset. This dataset refers to expression of microRNAs (miRNAs), a class of small noncoding RNA species with critical functions across various biological processes. We used the miGCM (miRNA Global Cancer Map) collection [7]. It contains 217 numerical attributes and 218 examples, distributed into 3 classes. Class 1 contains 46 (21.10%), representing normal human tissues, class 2 contains 140 (64.22%), representing human cancer specimens and class 3 contains 32 (14.67%) representing cell lines.

The experiments were conducted using the Weka[1] toolkit. As base-rankings, we used all the feature ranking algorithms implemented in Weka: χ^2, Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty (SU), ReliefF (RF), One Rule (1R) and SVM-REF (SR). Among them, SVM-REF is normally acknowledge as the state-of-art single feature selection method for bioinformatics [6]. AVG, MC and SSD were implemented in Java.

As the "true" best subset of biomarkers is unknown, the evaluation of feature subsection algorithms is often carried out indirectly by assessing the performance of a predictive model induced using the subset of features as an indirect measure. To avoid a bias due to a particular choice of measure/learning algorithm, we used three different learning algorithms and performance measures. The learning algorithms are SVM, J48 and Naive Bayes (NB) and the performance measures are misclassification error rate, F1 and area under the ROC curve (AUC).

To evaluate the methods, we used the same experimental design as in [3]. This approach follows a stepwise procedure, where the best ranked feature is used to induce a model using a learning algorithm, and the performance of the model is measured into a separated test set. A new feature is included according to the ranking provided by the feature ranking method (either single or aggregated ranking) in each step, and the process is repeated until all features were included. This process leads to a "performance curve", with the percentage of features into the x-axis and the measure into the y-axis. We use 10-fold cross validation and each cross validation run was repeated 10 times (10x10-fold cross validation).

Table 1. Average area under(bellow) the performance curve

Alg.	Measure	χ^2	GR	IG	1R	RL	SR	SU	AVG	MC	SSD
J48	AUC	0.792	0.788	0.780	0.837	0.768	0.808	0.779	0.906	0.810	0.869
	Error	0.781	0.783	0.772	0.826	0.774	0.800	0.770	0.902	0.799	0.861
	F1	0.835	0.837	0.826	0.865	0.829	0.845	0.826	0.924	0.846	0.891
NB	AUC	0.788	0.783	0.790	0.851	0.811	0.829	0.788	0.878	0.901	0.892
	Error	0.669	0.661	0.669	0.745	0.696	0.700	0.665	0.769	0.810	0.773
	F1	0.704	0.700	0.703	0.769	0.716	0.719	0.700	0.795	0.839	0.795
SVM	AUC	0.860	0.863	0.858	0.928	0.859	0.884	0.859	0.938	0.849	0.935
	Error	0.876	0.876	0.875	0.928	0.876	0.880	0.875	0.936	0.834	0.928
	F1	0.916	0.915	0.916	0.948	0.915	0.912	0.915	0.953	0.873	0.945

Table 1 shows the average area bellow the performance curve for F1 and AUC and the average area above the performance curve for Error. Overall, we can observe an inprovement in performance of the aggregate methods when comparated to the base methdos. This improvement is verified for all three learning algorithms and three different performance measures. We ran the Friedman test, at 95% of confidence level. The Friedman test is the nonparametric equivalent of the repeated-measures ANOVA, and the pos-hoc Nemenyi test, which is sim-

[1] `http://www.cs.waikato.ac.nz/ml/weka/`

ilar to the Tukey test for ANOVA and is used when all methods are compared to each other. The results of this test is depicted in Figure 1. The test shows that, for this dataset, apart from 1R and SVM-REF, the aggregate methods are statistically better than the base-rankings. MC performed best overall.

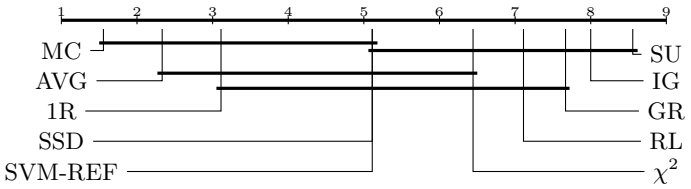


Fig. 1. Results of the Nemenyi test. Methods are ordered by performance from left to wright. A thick line joining two or more methods indicates that there is no significantly difference among methods.

4 Concluding remarks

This paper compares three different rank aggregation methods to construct ensembles of feature selection algorithms based on ranking. Experimental results using miRNA Global Cancer Map collection show that the aggregate rankings outperform the base-ranking methods. Future research direction includes the experimentation with a large set of datasets, specially gene expression datasets.

References

1. Biomarkers Definitions Working Group: Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clin Pharmacol Ther **69**(3) (2001) 89–95
2. He, Z., Yu, W.: Stable feature selection for biomarker discovery. `http://arxiv.org/abs/1001.0887` (2010)
3. Slavkov, I., Ženko, B., Džeroski, S.: Evaluation method for feature rankings and their aggregations for biomarker discovery. In: 3th Intern Workshop on Machine Learning in Systems Biology, Helsinki University Printing House (2009) 115–125
4. Abeel, T., Helleputte, T., de Peer, Y.V., Dupont, P., Saeys, Y.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics **26**(3) (2010) 392–398
5. Dutkowski, J., Gambin, A.: On consensus biomarker selection. BMC Bioinformatics **8**(S-5) (2007)
6. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46**(1-3) (2002) 389–422
7. Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E.A., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., Downing, J.R., Jacks, T., Horvitz, H.R., Golub, T.R.: Microrna expression profiles classify human cancers. Nature **435** (2005) 834–838

In silico characterization of *Rhodnius prolixus* lipophorin receptor

Vinicius Vieira de Lima[1], David Majerowicz [1], Glória R.C. Braz[2], Rafael Dias Mesquita[3], Katia C. Gondim[1]

[1] - Instituto de Bioquímica Médica, Universidade Federal do Rio de Janeiro, Brazil
[2] - Departamento de Bioquímica, Instituto de Química, Universidade Federal do Rio de Janeiro, Rio de *Janeiro*, Brazil
[3] - Departamento de Biotecnologia, Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro, Rio de Janeiro, Brazil.
Tel: +55 21 2562 6785
Fax: +55 21 2270 8647
katia@bioqmed.ufrj.br

Supported by CNPq, FAPERJ and INCT-EM

Abstract. Lipids have different functions in metabolism and cellular physiology and, as they are non-polar substances, they need carrier molecules to be transported in the organisms. In insect hemolymph they are transported by a major lipoprotein, lipophorin. In Rhodnius prolixus, a vector of Chagas disease, it has been shown that lipophorin specifically binds to receptors belonging to the superfamily of LDL receptors, present on the surface of midgut and fat body cells. In this study, it was found, with the use of bioinformatics tools, a partial sequence of lipophorin receptor gene from Rhodnius prolixus (RpLpR) and its amino acid sequence. Using these sequences, the RpLpR amino acid sequence was compared with those of other insect LpRs and their high homology was confirmed. Possible phosphorylation sites and the transmembrane helices of RpLpR were also predicted. All these results provide data for the characterization of lipophorin receptor in Rhodnius prolixus.

Keywords. Rhodnius prolixus, Lipid metabolism, Low density lipoprotein receptor, Lipophorin Receptor.

1 Results

In a blood meal, R. prolixus ingests a great amount of lipids. These lipids are digested and absorbed by midgut cells and then transported to lipophorin, a major lipoprotein in insect hemolymph [1, 2]. Lipophorin transfers these lipids to other tissues, interacting to specific binding sites on the cell surface [3, 4], probably the lipophorin receptor (LpR). Lipophorin receptor structure and function are poorly understood, although they seem to play a critical role on insect lipid metabolism since it can possibly regulate the lipid uptake by the cell. To investigate whether Rhodnius prolixus adults express the LpR gene, a RT-PCR reaction was performed using specific primers (data not shown) and it was observed that all analyzed organs express the RpLpR . The next step was an in silico characterization. Through the NCBI database (National Center for Biotechnology Information), the Acyrthosiphon pisum (access number GenBank: XP_001946703.1) LpR protein sequence was found. We used it as a query for the tBLASTn algorithm to search the contig that might contain RpLpR on R. prolixus genome. Two contigs that matched the query were found, the contigs 977.6 (access number GenBank: ACPB01046953.1) and 977.8 (access number GenBank: ACPB01046955.1), that were united and converted to a positive open reading frame for posterior analysis. Then, with the use of GeneWise software [5], [6] a comparison of the two R. prolixus contigs with the A. pisum LpR was performed and as a result the LpR partial protein sequence was obtained (Fig. 1). All expected conserved domains were observed, like the cysteine-rich ligand binding domain (LBD), an important region of interaction with lipoproteins [7], the EGFD with the YWXD motif and the intracellular domain with the characteristic FDNPVY motif of LDLRs [8]. To analyze the similarity relationship of other insect LpRs with RpLpR, a dendrogram was created using the Neighbor-Joining method [9] with the MEGA 4.0 software (Fig. 2). It was shown that LpR is highly conserved among several insect species such as R. maderae (Access number GenBank: BAE00010.1), L. migratoria (Access number GenBank: CAA03855.1), P. humanus (Access number GenBank: XP_002428384.1), B. mori (Access number GenBank: BAE71409.1), G. mellonella (Access number GenBank:

ABF20542.1), D. melanosgaster (Access number GenBank: NP_001163734.1), A. aegypti (Access number GenBank: AAQ16410.1), and A. pisum (Access number GenBank: XP_001946703.1). Next, a search for phosphorylation sites on RpLpR was performed using NetPhos 2.0 Server [10, 11] and as a result many serine, threonine and tyrosine residues with high probability of phosphorylation were found (Fig. 3). Then, it was analyzed which part of the receptor is attached to the plasma membrane, and for so the TMHMM server v. 2.0 [12] was used. Only one region at the C-terminus of the RpLpR sequence was predicted to have transmembrane helices (Fig.4). The conclusion of this work is that RpLpR is a real member of the LDLR superfamily and has high homology with other insect LpRs. This receptor is possibly phosphorylated and only one region containing transmembrane helices was found. This set of results contributes to the understanding of the biological function of this protein in insects.

>Partial RpLpR amino acid sequence

CKPEEFRCHSGRCIPQYWQCDKEPDCPDESDEDPNTCNFGKCNADELQCSAHECVPLAWVCDGVRDCKNNVDEKNCKEKLCQ
AEEFTCRAAPGECVPLTWMCDDATDCTDGSDERACNETCRSDEFTCGNGKCIQQIWACDMDEDCEDGSDEKNCSERVCASNE
FQCEDHSCILAKWKCDGDYDCHDNSDEKGCTNAPEPVSSCLSKEFKCPDHLTCIHQTWVCDGDPDCPDGADESPQLCTNLTC
RPDQFQCSSRQCIPGHFHCNGQSDCVDGSDEQDCGNLLYCVIYLDYSDDTVRNTDLEIKKVCDIILENSLGPYVIIIRHSGK
YEVCQYVIELHFRCTPIPRFTRKIGSRENHINECDNPGACSQSCINEKGTFKCECEEGYLRDPHDRTRCKATEGHASLLFAR
RHDIRKISLDHHEMTSIVNDTKSATALDFVFRTGMIFWSDSSEQKIYKAPIDEGSERTIVISNEVTMCDGLAVDWLYSHIYW
TDTGKNTIELANYEGSMHKTLIRDSLDEPRAIALNPLDGWMYWSDWGKEGRIERAGLDGSHRQVIVSYDIRWPNGLTLDLVR
KRLYWVDAKLNLISSVNYDGSGRRVVLRSTETLHHPFSISVFEDFVYWTDWDKQAIFKANKFDGSNVTAITAIRMLQNPMVV
HVYHPYRQPDGANHCAAVNGHCSHLCLPAPQINPRSPKISCACPDGLVLMKDGLMCTDQGQVHRKIVLHNHKEDRPHDMFDE
ADSGVIASVVIAGISVFLAFASMIVFIIYRHYLRRNVTSMNFDNPVYRKTTEDQFSLAKSHFQTQRIYPATVAEE

Figure 1. The partial RpLpR amino acid sequence was predicted using the A. pisum LpR as a query in Rhodnius prolixus genome. Conserved motifs are showed in red characters.

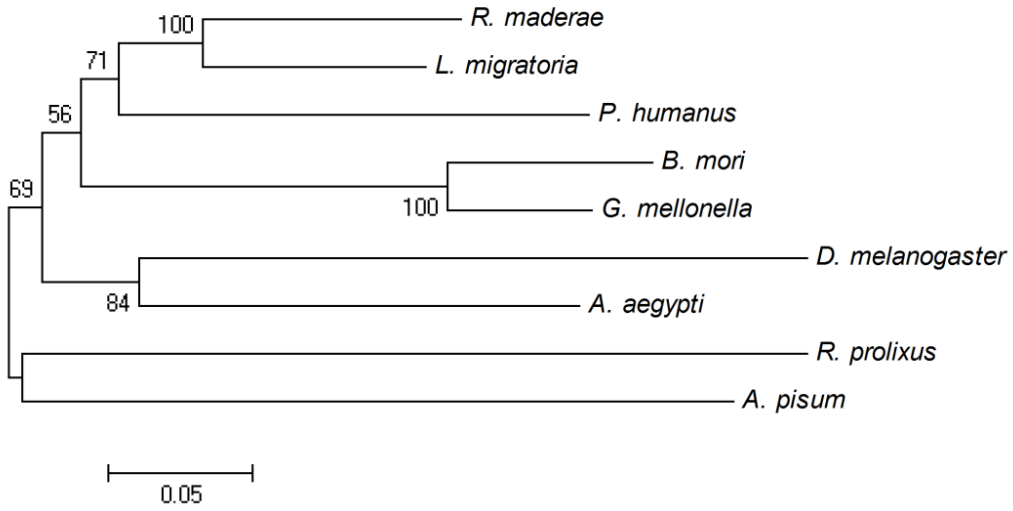


Figure 2. Dendrogram done on MEGA 4.0 showing high conservation among LpR of several insect species.

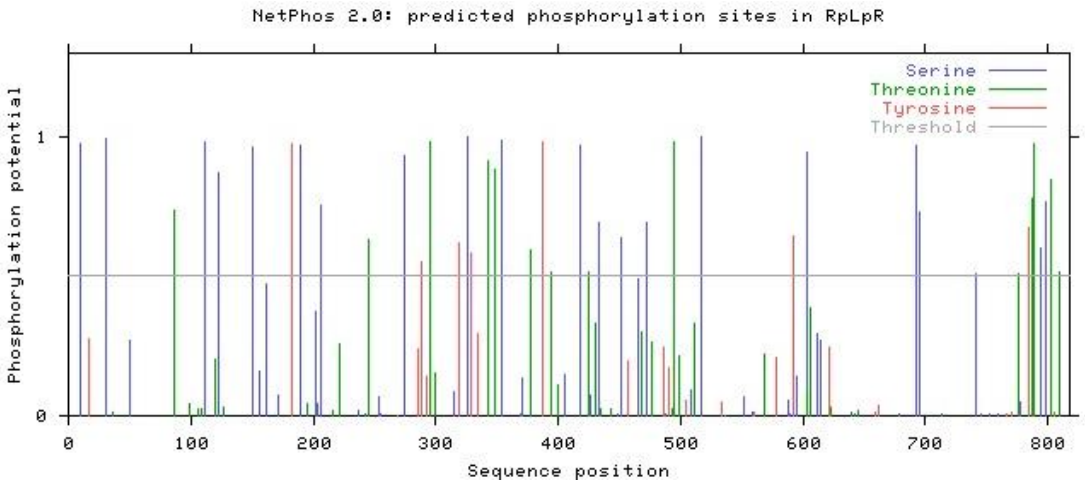


Figure 3. NetPhos 2.0 graph showing probability of serine, threonine and tyrosine residues to be phosphorylated in RpLpR protein sequence.

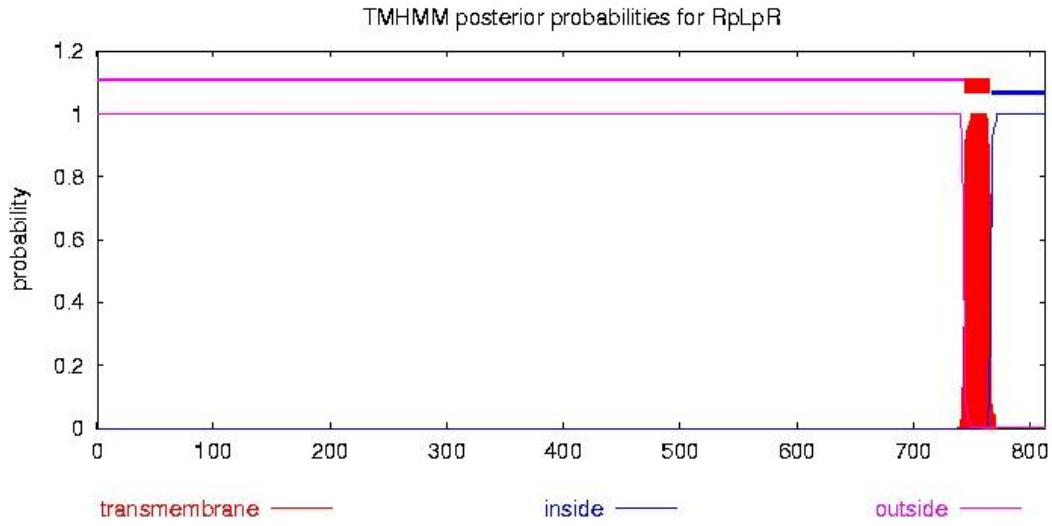


Figure 4. TMHMM analysis of RpLpR protein sequence, showing only one region with high probability of containing transmembrane helices.

References

1 CHINO, H.; DOWNER, R.G.H., WYATT, G.R., GILBERT, L.I. Lipophorins, a major class of lipoproteins of insect haemolymph. Insect Biochem. 11: 491 (1981).

2 SHAPIRO, J.P., LAW, J.H., WELLS, M.A. Lipid transport in insects. Annu Rev Entomol. 33: 297-318 (1988).

3 GONDIM, K.C., WEELS, M.A. Characterization of lipophorin binding to the midgut of larval Manduca sexta. Insect Biochem. Mol. Biol. 30:405-413 (2000).

4 GRILLO, L.A.M., PONTES, E.G., GONDIM, K.C. Lipophorin interaction with the midgut of Rhodnius prolixus: characterization and changes in binding capacity. Insect Biochem. Mol. Biol. 33:429-438 (2003).

5 Wise2 - Intelligent algorithms for DNA searches, http://www.ebi.ac.uk/Tools/Wise2/index.html.

6 BIRNEY, E., CLAMP, M., DURBIN, R. GeneWise and Genomewise. Genome Res. 14(5):988-95 (2004).

7 GOLDSTEIN, J.L., BROWN, M.S. Binding and degradation of low density lipoproteins by cultured human fibroblasts. Comparison of cells from a normal subject and from a patient with homozygous familial hypercholesterolemia. J Biol Chem. 249(16):5153-62 (1974).

8 DAVIS, C.G., VAN DRIEL, I.R., RUSSELL, D.W., BROWN, M.S., GOLDSTEIN, J.L. The low density lipoprotein receptor. Identification of amino acids in cytoplasmic domain required for rapid endocytosis. J Biol Chem. 262(9):4075-82 (1987).

9 SAITOU, N., NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4(4):406-25 (1987).

10 NetPhos 2.0 Server, http://www.cbs.dtu.dk/services/NetPhos/.

11 BLOM, N., GAMMELTOFT, S., BRUNAK, S. Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. Journal of Molecular Biology: 294(5): 1351-1362 (1999).

12 TMHMM server v. 2.0, http://www.cbs

PWD: Logical schema and ETL Process

Cristian Tristão[1], Carlos Juliano M. Viana[1], Márcia Mártyres Bezerra[2], Renato Marroquin Mogrovejo[1], Sérgio Lifschitz[1], and Antônio Basílio de Miranda[2]

[1] Pontifical Catholic University of Rio de Janeiro - PUC-Rio
{ctristao,cviana,rmogrovejo,sergio}@inf.puc-rio.br
[2] Oswaldo Cruz Institute - FIOCRUZ - RJ, Brazil
marciamb@ioc.fiocruz.br, antonio@fiocruz.br

Abstract. The development of DNA automated sequencing methods on large scale, coupled with the development of high performance computing technologies and more efficient algorithms have resulted in the generation of large amounts of data. This fact has enabled the scientific community to study the structure, organization and evolution of genomes. Today the main challenge is organizing, storing and making available all this biological data without compromising the interpretation and understanding of biological systems, and their interactions. This manuscript describes the logical schema used to store the data from the Genome Comparison Project (GCP) together with data from various public sources, e.g. RefSeq, Swissprot, NCBI Taxonomy, Pfam, KEGG and GO, as well as discusses the drawbacks of using a data Extraction, Transformation, and Loading process (ETL). The proposed schema allows obtaining relevant information about the proteins similarity, basic process in the functional annotation and in the discovery of new proteins.

1 Introduction

The availability of numerous organisms complete genome sequences, associated with the computational progress occurred in the last few decades have provided an opportunity to use holistic approaches for meticulous genome structure studies, as well as for gene prediction and functional classification. However, the large amounts of generated data become a factor that could derail the storage, manipulation and availability of data. Today, the ability to manage data is as important as the processing capacity.

The Protein World Database (PWD) [7, 4] is a result of the Genome Comparison Project (GCP) [3], an ongoing research project that compared more than 3.8 million of proteins sequences in a pairwise manner using the SSEARCH program, an Smith Waterman algorithm implementation. The GCP is an initiative among the Functional Genomics and Bioinformatics Laboratory - Fiocruz [2], the World Community Grid^{TM}(WCG) [5], and the PUC-Rio Bioinformatics Laboratory [1]. This manuscript presents a data logical model and describes the ETL process of the PWD.

2 Logical Project

The logical schema generation **Fig.1** occurred after the conceptual schema definition of the PWD, as described in [6]. Some basic mapping rules were used and, when necessary, some modeling decisions were taken. According to one of these rules, all entities in the conceptual schema became tables in the logical schema. Thus, the PROTEIN, ORF_T, GENOMIC SEQUENCE, CDS, GENE, TAXONOMY, RANK, DOMAIN, GENE_ONTOLOGY and ENZYME entities were mapped respectively to protein, orf_t, genomicsequence, cds, gene, taxonomy, tax_rank, domain, gene_ontology and enzyme tables. Regarding the entity relationships, we applied the mapping rules and made the following decisions:

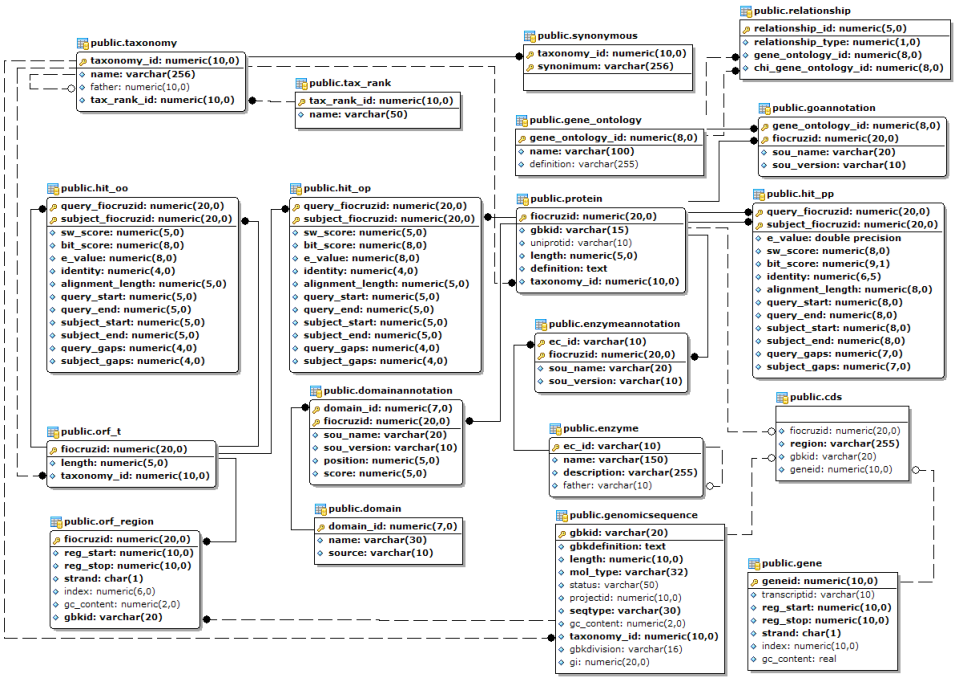


Fig. 1. Protein World Database: logical schema.

– Hits - represents the similarity relationship between proteins and tORFs (translated "Open Reading Frame - ORF") compared in the GCP. It is a "many-to-many" relationship, so it was necessary to create a table for each relationship: hit_oo, relationship between tORFs; hit_pp, relationship between proteins; and hit_op, relationship between tORFs and proteins.
– Taxonomy - has a "one-to-many" self-relationship, then the taxonomy table gets a new attribute called "father" (foreign key for taxonomy). Because of the "one-to-many" relationship between rank and taxonomy the taxonomy

table gets the reference to the attribute related to rank, called tax_rank_id. The taxonomy entity has also a "one-to-many" relationship with the orf_t and protein tables, so both receive the reference attribute taxonomy_id. Synonymous is a taxonomy compound attribute, therefore we must convert this attribute into a new table called synonymous, and add to it the synonimum and taxonomy_id (foreign key for taxonomy) attributes.

- Central Dogma of Molecular Biology - the relationship between orfs and genomic sequences is a "one-to-many" relationship. The mapping rule instructs to insert into the orf_t table a reference to the genomicsequence table. However, due to organizational and conceptual reasons, the orf_region table was created to represent this relationship. CDS references gene and genomicsequence because the "one-to-many" relationship, and references protein for managerial reasons.
- Annotations - are "many-to-many" relationships of proteins with their annotated properties in public databases, such as enzymatic activities, functional domains, and gene ontology (GO) terms. Therefore, tables were created to represent them, called respectively enzymeannotation, domainannotation and goannotation. Moreover, the relationship table was created to represent the GO "many-to-many" self-relationship.

3 Extraction, Transformation and Loading

After generating the logical schema, the next step was the implementation of the Protein World DB, and the definition of the Extraction, Transformation and Loading process (ETL) as described in **Fig.2**.

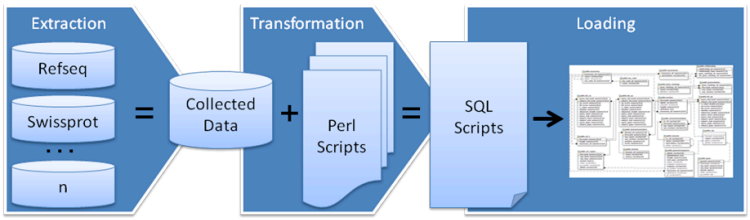


Fig. 2. ETL process.

The logical schema was implemented in the Relational Database Management System (RDBMS) PostgreSQL version 8.4. The first stage of the ETL involves extracting data from different sources, which usually use different formats and data organization, such as: (a) RefSeq, (b) Swissprot, (c) Taxonomy NCBI, which constitutes the core of the information, and (d) Pfam - domains, (e) KEGG - metabolic pathways, and (f) Gene Ontology. The transformation stage involves applying a series of rules and/or functions to the extracted files in order to transform them into the appropriate format for the load task. For this phase,

scripts based on Perl programming language were used to generate SQL scripts. The last task of the ETL process was the execution of these SQL scripts to insert the formatted data.

4 Conclusions and Future Works

The Genome Comparison Project has a great multidisciplinary potential, making possible the exchange of experiences between different research areas. In addition, the Protein World DB becomes an important data source to answer several questions about the organization, structure and annotation of genes, proteins and genomes. The generated database will be available to the scientific community, becoming a repository of great value to all involved researchers.

Queries of scientific interest are currently being developed and they will be available to the scientific community. Regarding the Very Large Database (VLDB), effective solutions of data access should also be proposed to avoid some possible data processing bottlenecks. Another interesting possibility would be to investigate the use of distributed databases. Furthermore, new approaches based on the MapReduce [9, 8] programming model could be used as a distributed scalable database system. This could allow researchers to develop higher-throughput analysis pipelines in order to process data of some ever-growing biological databases and genome sequences.

References

1. Data Management and Bioinformatics Laboratory, 2010. Available at URL: http://www.inf.puc-rio.br/~blast/.
2. Functional Genomics and Bioinformatics Laboratory - Fiocruz, 2010. Available at URL: http://www.dbbm.fiocruz.br/labwim/bioinfoteam/index.pl?action=home.
3. Genome Comparison Project, 2010. Available at URL: http://www.dbbm.fiocruz.br/GenomeComparison/.
4. Protein World Database, 2010. Available at URL: http://proteinworlddb.org/.
5. World Community Grid. Available at URL: http://www.worldcommunitygrid.org/., 2010.
6. Tristão C., Miranda A.B., and Lifschitz S. A conceptual data model involving protein sets from complete genomes: a biological point of view. Technical Report MCC 27/09, PUC-Rio, 2009.
7. Otto Thomas Dan, Marcos Catanho, Cristian Tristão, Márcia Bezerra, Renan Mathias Fernandes, Guilherme Steinberger Elias, Alexandre Capeletto Scaglia, Bill Bovermann, Viktors Berstisand Sérgio Lifschitz, Antônio Basílio de Miranda, and Wim Degrave. ProteinWorldDB: querying radical pairwise alignments among protein sets from complete genomes. *Bioinformatics*, 26(5):705–707, 2010.
8. Christopher Olston., Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. Pig latin: a not-so-foreign language for data processing. *SIGMOD*, pages 1099–1110, 2008.
9. Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Anthony, Hao Liu, and Raghotham Murthy. Hive - a petabyte scale data warehouse using hadoop. *ICDE*, pages 996–1005, 2010.

Search and Rational Design of Inactive Anionic Sequences to Enable Antimicrobial Activity

William F. Porto[1], Ludovico Migliolo[1], Osmar N. Silva[1,2] and Octávio L. Franco[1,2]

[1]Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, DF, Brazil, 70790-160
[2]Pós-Graduação em Genética e Imunologia, Universidade Federal de Juiz de Fora, MG, Brazil
{williamfp7, kingvico, osrmarns, ocfranco} @gmail.com

Abstract. Conventional antibiotics' activities against pathogenic bacteria have decreased. Antimicrobial peptides appear as an alternative to control these pathogens. Moreover, these peptides have been commonly redesigned in order to improve their biological activity. This work was carried out to develop a novel method to enable antimicrobial activities from inactive sequences, through amino acids substitutions. To this end, anionic residues were replaced by lysine residues, the most common amino acid in antimicrobial peptides. Antimicrobial activity from original and redesigned sequences was predicted through Collection of Antimicrobial Peptides Algorithms; 99.1% of original sequences had been predicted as non-antimicrobial based on a mixed approach; after the redesign, 65% were predicted as antimicrobial. Through this method, novel antimicrobial peptides can be generated and these sequences can be used in novel strategies to control pathogenic bacteria.

Keywords: Antimicrobial Peptides, Rational Design, Antimicrobial Activity Prediction.

1 Introduction

In recent decades conventional antibiotics' activity against pathogenic bacteria has decreased, due to the development of bacterial resistance [1]. Antimicrobial peptides (AMP) appear as an alternative to control those pathogens [2]. Modifications in peptides' primary structure may lead to modifications in their lethal activity. Rational design is a common practice to increase or enable antimicrobial activity [3].

AMPs may share some properties that could modulate their activity, especially positive net charge, since this is essential to interact with negative membranes; hydrophobicity, which provides a major affinity to lipids; and amphipathic structure, joining the two first properties [4, 5].

Mining free protein databases is an important tool to search of novel AMPs [6]. Although AMPs vary between 7 and 59 amino acid residues [1], approaches targeting the smaller peptides must be made, as these have a greater potential than larger peptides, since they are easier to synthesize and may have similar activity [7].

This describes a method to mine unusual candidates in databases, based on amphipathic characteristics, without evidences of the tertiary structure and activity to redesign to enable the antimicrobial activity.

2 Material and Methods

Starting from the NCBI's Non Redundant Protein Database (NR - ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz), all anionic peptides with 18 amino acid residues and a ratio of hydrophobic to charge residues of 1:1 or 2:1 were extracted. Subsequently the redundant sequences in 80% or more were removed through Jalview software [8]. Based on the original set, the redesigned set was generated with a single amino acid substitution: the aspartic acids and the glutamic acids were replaced by lysine residues.

The antimicrobial activities from the original and redesigned sets were predicted through Collection of Anti-Microbial Peptides (CAMP) algorithms [9], SVM, Random Forest and Discriminant Analysis. Figure 1 shows an overview of proposed method.

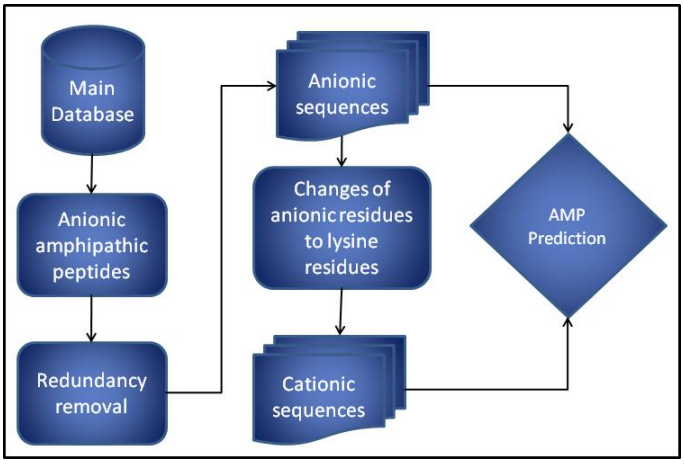


Fig. 1. Flowchart of proposed data mining method. NR was the main database.

3 Results and Discussion

Database search is a strategy to find unusual AMPs [6]; NR has been chosen in order to increase the range of proteins without use random approaches to generate the sequences to redesign. Firstly, the anionic anphipathic peptides were selected, since they have a little chance of displaying antimicrobial activity, even with an anphipathic

structure [1, 4]. In this way, changing the anionic residues to lysine residues, the sequences become candidates for AMP, increasing their affinity to anionic lipids [9].

From NR, 388 anionic peptides with 18 residues were extracted. In order to avoid very similar results in the predictions, sequences with redundancy over 80% were discarded, leaving 329 sequences in the original set. The redesigned set had the same number of sequences from original set. Figure 2 shows a frequency logo from the 388 anionic sequences, the most frequent amino acids residues are leucine, alanine, glicine, serine, glutamic acid and aspartic acid, which may provide an amphipathic and flexible structure.

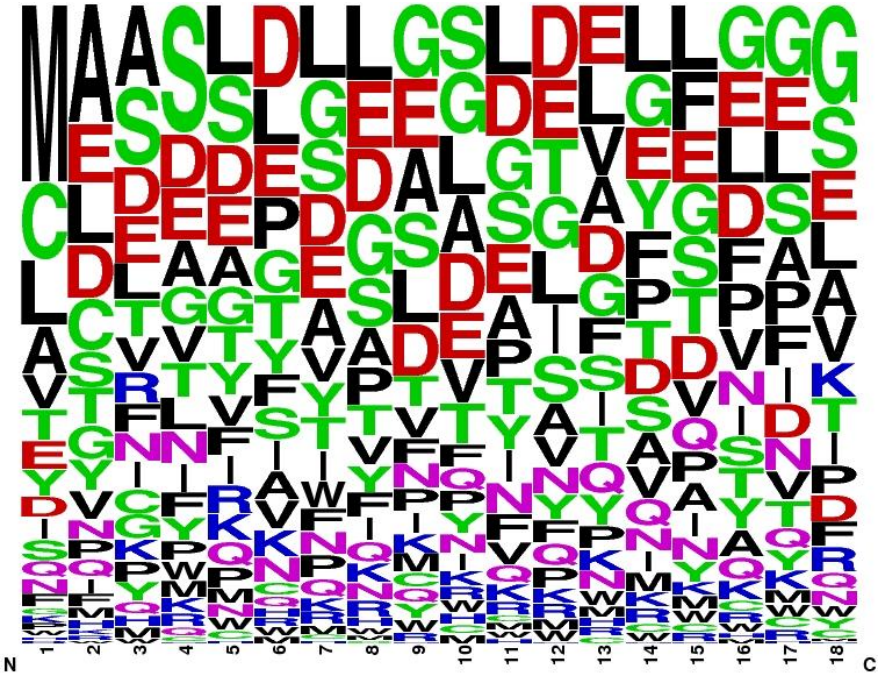


Fig. 2. Frequency logo from original set, generated by WebLogo [10].

Table 1 shows the results of antimicrobial activity prediction. The sequence had been predicted as positive in Mixed Approach when two or three algorithms predicted it as positive. Based on this approach, the redesign method may allow activity from 65% of original sequences, which in their original state shows only 0.9% of predicted antimicrobial activity. This shows that single amino acids substitutions may allow the antimicrobial activity.

Table 1. Positive prediction of antimicrobial activity through CAMP algorithms.

Set of Sequences	SVM	Random Forest	Discriminant Analysis	Mixed Approach
Original	9	2	4	3
Redesigned	197	208	235	214

4 Conclusion

Due to the easy implementation of this method, lots of novel candidates to AMP can be generated; and many of these candidates can be used as a novel strategy against pathogenic bacteria. Despite the efficacy of redesign method, the great challenge is how to choose the best candidates for the *in vitro* test, because the prediction algorithms tell us if the peptide has the potential to be an AMP. We therefore have 214 candidates for AMP. In future studies, a classifier method will be implemented, in order to select the best candidates for *in vitro* tests and structural predictions, generating more detailed information about the changes caused by the redesign method.

Acknowledgments. This work was supported by FAPDF, CNPq, CAPES and UCB.

References

1. Brogden, K. A.: Antimicrobial Peptides: Pore Formers or Metabolic Inhibitors in Bacteria? Nature Microbiology. 3, 238--250 (2005)
2. Mandal, S.M., Dey, S., Mandal, M., Sarkar, S., Maria-Neto, S., Franco, O.L.: Identification and Structural Insights of Three Novel Antimicrobial Peptides Isolated from Green Coconut Water. Peptides. 30, 633--637 (2008)
3. Pacor, S., Giangaspero, A., Bacac, M., Sava, G., Tossi, A.: Analysis of the Cytotoxicity of Synthetic Antimicrobial Peptides on Mouse Leucocytes: Implications form Systemic Use. J. Antimicrob. Chemother. 50, 339--348 (2002)
4. Drin, G., Antonny, B.: Amphipathic Helices and Membrane Curvature. FEBS Letters. 584, 1840--1847 (2010)
5. Dathe, M., Wieprecht, T., Nikolenko, H., Handel, L., Maloy, W.L., MacDonald, D.L., Beyermann, M., Bienert, M.: Hydrophobicity, Hydrophobic Moment and Angle Subtended by Charged Residues Modulate Antibacterial and Hemolytic Activity of Amphipathic Helical Peptides. FEBS Letters. 403, 208--212 (1997)
6. Fernandes, F. C., Porto, W. F., Franco, O. L.: A Wide Antimicrobial Peptides Search Method Using Fuzzy Modeling. LNBI. 5676, 147--150 (2009)
7. Ahn, H. S., Cho, W., Kang, S. H., Ko, S. S. Park, M. S., Cho, H., Lee, K. H.: Design and Synthesis of Novel Antimicrobial Peptides on the Basis of Alpha Helical Domain of Tenecin 1, an Insect Defensin Protein, and Structure-Activity Relationship Study. Peptides. 27, 640--648 (2006)
8. Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M., Barton, G. J.: Jalview Version 2 - A Multiple Sequence Alignment Editor and Analysis Workbench. Bioinformatics. 25 (9), 1189--1191 (2009)
9. Thomas, S., Karnik, S., Barai, R. S., Jayaraman, V. K., Idicula-Thomas, S.: CAMP: A Useful Resource for Research on Antimicrobial Peptides. Nucl. Acids Res. 38, D774--D780 (2010)
10. Crooks G.E., Hon, G., Chandonia, J.M, Brenner, S.E.: WebLogo: A Sequence Logo Generator. Genome Res. 14, 1188--1190 (2004)

References

1. Maurilio de Araujo Possi, Alcione de Paiva Oliveira, Vladimir Oliveira Di Iório, Cristina Maria Ganns Chaves Dias. A agent-based simulation tool of biological immune system: a case study of autoimmune diseases (work-in-progress). In *BSB 2010 Poster Proceedings*, pages 7-10. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
2. Renato M. Mogrovejo, Carlos Juliano M. Viana, Cristian Tristão, Márcia Mártyres Bezerra and Sérgio Lifschitz. A Cloud-based Method For Comparing Three Genomes. In *BSB 2010 Poster Proceedings*, pages 11-14. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
3. Carla Carvalho de Aguiar, Danieli F. Figueiredo, Osmar Norberto de Souza. A possible three-dimensional model for the enzyme chorismate synthase from Plasmodium falciparum. In *BSB 2010 Poster Proceedings*, pages 15-20. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
4. Luciana da Silva Almendra Gomes, Sérgio Lifschitz, Priscila V. S. Z. Capriles, Laurent E. Dardenne. A Provenance Model for Bioinformatics Workflows. In *BSB 2010 Poster Proceedings*, pages 19-22. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
5. Capriles, P. V. S. Z.; Custódio, F. L.; Dardenne, L. E. Ab initio Protein Structure Prediction via Genetic Algorithms using a Coarse-grained Model for Side Chains. In *BSB 2010 Poster Proceedings*, pages 23-27. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
6. Flávia G. Silva, Kátia P. Lopes, Sandro R. Dias . An algorithm to search and repair errors and non conformities in a biological database. In *BSB 2010 Poster Proceedings*, pages 28-31. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
7. Bellini R. G., Ribeiro T. S., Figueiredo K., Pacheco M. A. C. Approaching Protein Folding Through Neural Networks. In *BSB 2010 Poster Proceedings*, pages 32-34. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
8. Flávia Thiebaut, Clícia Grativol, Cristian A. Rojas, Renato Vicentini, Adriana S. Hemerly, Paulo C. G. Ferreira. Computational analysis of small RNAs libraries of sugarcane cultivars submitted to drought stress. In *BSB 2010 Poster Proceedings*, pages 35-39. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
9. Christian V. Quevedo, Ivani Pauli, Osmar Norberto de Souza and Duncan D. Ruiz. Development of a filter of molecular descriptors aiming to select the most promising ligands to a flexible receptor. In *BSB 2010 Poster Proceedings*, pages 40-43. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
10. Daniele Palazzi, Ely Edison Matos, Fernanda Campos, Regina Braga, Elaine Coimbra. Human Disease: domain ontology to simulate biological models. In *BSB 2010 Poster Proceedings*, pages 44-47. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
11. Priscila Grynberg, Mainá Bitar, Alexandre Paschoal, Alan M. Durham, Glória R. Franco. Identification and Classification of ncRNAs in Trypanosoma cruzi: A Multistep Approach. In *BSB 2010 Poster Proceedings*, pages 48-51. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).

12. Ronaldo C. Prati. Improving Biomarker Identication through Ensemble Feature Rank Aggregation. In *BSB 2010 Poster Proceedings*, pages 52-55. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
13. Vinicius Vieira de Lima, David Majerowicz, Glória R.C. Braz, Rafael Dias Mesquita, Katia C. Gondim. In silico characterization of Rhodnius prolixus lipophorin receptor. In *BSB 2010 Poster Proceedings*, pages 56-58. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
14. Cristian Tristão, Carlos Juliano M. Viana, Márcia Mártyres Bezerra, Renato Marroquin Mogrovejo, Sérgio Lifschitz and Antônio Basílio de Miranda. PWD: Logical schema and ETL Process. In *BSB 2010 Poster Proceedings*, pages 59-62. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).
15. William F. Porto, Ludovico Migliolo, Osmar N. Silva and Octávio L. Franco. Search and Rational Design of Inactive Anionic Sequences to Enable Antimicrobial Activity. In *BSB 2010 Poster Proceedings*, pages 63-66. Búzios, Rio de Janeiro, Brazil, September 2010. Held in conjunction with the International Workshop on Genomic Databases (IWGD 2010).