Escola Gaúcha de Bioinformática

27 a 31 de Julho de 2015
Centro de Eventos do Instituto de Informática da UFRGS

A Bioinformática é um dos campos de pesquisas com maior crescimento nos últimos anos e cujo papel nas diferentes áreas do conhecimento biológico tem aumentado de forma exponencial. Este crescimento se deve a inúmeros fatores, dos quais se destacam: (i) geração de dados biológicos em grandes volumes, provenientes do desenvolvimento das chamadas "técnicas de larga-escala", das quais incluem-se a genômica, transcriptômica e proteômica; (ii) acesso facilitado a bancos de dados desenvolvidos especialmente para a estocagem dos dados biológicos; (iii) geração de algoritmos e sistemas informatizados capazes de processar os diferentes dados biológicos e (iv) diminuição dos custos associados com processadores, memórias e sistemas de armazenamento, facilitando o acesso de diferentes grupos de pesquisas à computadores com alto poder e/ou desempenho para o processamento de dados. Conectado a estes pontos está o fato de que a Bioinformática, como campo de pesquisas, alia de forma transdisciplinar as áreas das Ciências Biológicas e das Exatas (Ciência da Computação, Física, Matemática, entre outras), proporcionando uma troca bastante dinâmica de conhecimentos entre os pesquisadores.

A Escola Gaúcha de Bioinformática, EGB, realizada de 27 a 31 de julho de 2015, foi organizada de forma integrada pelo Instituto de Informática e pelo Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul, buscando construir um espaço de formação, integração, qualificação e desenvolvimento das atividades de pesquisa envolvendo o emprego de métodos computacionais no estudo de sistemas biológicos. Sob uma perspectiva interdisciplinar de problemas biológicos e terapêuticos, o evento agregou tanto abordagens através de sequências (de nucleotídeos ou aminoácidos), de biologia de sistemas e de larga escala, quanto de estruturas 3D (e suas conformações) ao suporte e desenvolvimento possibilitados pela Ciência da Computação. Durante a realização da EGB foram ofertadas aulas, mini-cursos e atividades envolvendo assuntos do estado da arte da área de Bioinformática, os quais não são frequentemente cobertos nos cursos graduação e ou pós-graduação no estado do RS. Além disto, o evento induziu futuras colaborações entre alunos e pesquisadores de diferentes instituições de ensino e pesquisa do Rio Grande do Sul e Brasil. A Escola também contribuiu com a qualificação de profissionais que atuam na área de Bioinformática no estado do Rio Grande do Sul e também com o desenvolvimento e expansão desta área de pesquisa. A EGB 2015 contou com a participação efetiva de 159 pessoas. Durante a semana de realização da Escola estiveram presentes estudantes de graduação e pós-graduação, bem como de profissionais das áreas de Ciências Biológicas, Matemática, Física, Química, Engenharias e Ciências da Computação e as suas respectivas áreas correlatas.

Durante a EGB foram realizados 7 mini-cursos teóricos, práticos e teórico-práticos em nível básico e avançado, cada um com carga horária de 15hs. Um público de 133 pessoas participaram dos mini-cursos no período 27 a 31 de Julho de 2015. Os mini-cursos realizados e o número de participantes em cada mini-curso foram: Mini-curso 01: Análises transcriptômicas e biologia de sistemas - avançado: 31 participantes; Mini-curso 02: Programação Java para Bioinformática - básico: 11 participantes; Mini-curso 03: Introdução à Biologia de Sistemas - básico: 36 participantes; Mini-curso 04: Análise e interpretação de dados estruturais e conformacionais - avançado: 9 participantes; Mini-curso 05: Programação Python para Bioinformática - básico: 30 participantes; Mini-curso 06: Uso de pacote R para a Biologia de Sistemas - básico: 8 participantes; Mini-curso 07: Biologia Estrutural para iniciantes - básico: 8 participantes. Durante a EGB 2015 foram aceitos e apresentados, na forma de pôsteres e sessões colaborativas, 42 trabalhos científicos. Os trabalhos foram submetidos na forma de resumos, e encontram-se nas próximas páginas.

A Comissão Organizadora agradece: a presença de estudantes, professores, pesquisadores e palestrantes; ao apoio financeiro recebido pela FAPERGS, CNPq, CAPES e PROPESQ; ao Instituto de Informática da UFRGS e sua equipe técnica pelo suporte recebido durante o evento; e aos estudantes do Instituto de Informática e do Centro de Biotecnolgia da UFRGS que estiveram envolvidos com a organização da EGB 2015.

Porto Alegre, 30 de Maio de 2016.

Márcio Dorn – Inf – UFRGS
Diego Bonatto – CBiot – UFRGS
Hugo Verli – CBiot - UFRGS

Identification of Transcription Factors that Act as Master Regulators in Alzheimer's Disease and Parkinson's Disease

Marco Antônio De Bastiani[1]
Carolina Chatain[1]
Mauro Antônio Castro[2]
Fábio Klamt[1]

Alzheimer's Disease (AD) and Parkinson's Disease (PD) are the two most common neurodegenerative disorders. It is estimated that more than 45 million people worldwide suffer from one of these pathologies. Despite the large investment in the neuroscience field, etiology and molecular mechanisms underlying neuronal death remain unclear. Recently, a small proportion of cases of AD and PD have been attributed to mutations in specific genes. However, the many pathways in which their gene products are involved and the interaction with other factors that might lead to neuropathological changes are still poorly understood. In the current work, microarray data acquired from NCBI public repository Gene Expression Omnibus (GEO) (GSE60862) was used to determine normal tissue-specific transcriptional networks for hippocampus and *substantia nigra*, structures specifically damaged in AD and PD, respectively. Case-control studies (GSE5281 and GSE8397), also obtained from GEO, were used to establish gene expression signatures for the two neurodegenerative diseases and identify transcription factors (TFs) that are pivotal to modulate phenotypic changes from normal to pathological contexts, called master regulators (MRs), applying MRA and GSEA algorithms. As results, we identified 117 important TFs regulating gene expression in hippocampus and 123 in *substantia nigra*. We proposed 17 MRs involved with AD and 28 with PD, some of which have already been described in the process of neurodegeneration (such as YY1, HMG20A, RREB-1and SLC3A9) and others not related to the pathologies so far. We believe that these results might help in the understanding of AD and PD and lead to the discovery of targets for potential therapeutic intervention.

[1] Universidade Federal do Rio Grande do Sul
{marco.bastiani@ufrgs.br, carolpchatain@gmail.com, 00025267@ufrgs.br}
[2] Universidade Federal do Paraná {mauro.a.castro@gmail.com}

GROMOS 53A6 Force Field Parameters for Chalcones

Elisa Beatriz de Oliveira John[1]
Pablo Ricardo Arantes[1]
Hugo Verli[1]

Chalcones are compounds extensively distributed in plants and considered as the precursors for flavonoid synthesis. Their structure consists of an open chain in which two aromatic rings are joined by a three-carbon α, β-unsaturated carbonyl system. Chalcones are classified according to the A- and B- ring substitution by OH and OCH_3 groups, and different substitution patterns can affect their biological activities. Since they present many pharmacological activities, these biomolecules are being intensively studied and modified. Computational methods such as molecular dynamics (MD) simulations can provide microscopic information that may be difficult to obtain from other experiments. Accurate force fields are essential for describing biological systems in a MD simulation, thus a parameter set associated to a certain compound need to be carefully calibrated to ensure reliable results. There are no validated parameters for the α, β-unsaturated carbonyl group in the GROMOS force field, therefore the present work intends to provide a new parameter set for the simulation of chalcones. We employed a protocol for parameter optimization combining ab initio calculations, which provide the quantum mechanical (QM) potential energy profile (performed by GAMESS), and MD simulations that provide the molecular-mechanical (MM) potential energy profile, using the GROMACS simulation suite and GROMOS53A6 force field. A fitting of MM to QM torsional profiles was performed for each of three dihedrals of interest within the basic structure of chalcone. For the dihedrals including the rings, the fitting has already generated parameters that reproduce well the QM potential energy in the MM calculations. Additionally, the density and enthalpy of vaporization values are in good agreement with experimental data. When completed, we expect that such parameters will be able to properly describe the conformational distribution of chalcones, a starting point to further studies on the biological role of such molecules, at the atomic level.

[1] Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Caixa Postal 15005
{elisajohn@live.com; hverli@cbiot.ufrgs.br}

Effects of D-myo-inositol 3,4,5,6-tetrakisphosphate (TMI) binding on antithrombin

Pablo Ricardo Arantes[1]
Horacio Pérez-Sánchez[2]
Hugo Verli[1]

Antithrombin (AT), a serine protease inhibitor, circulates in blood in two major isoforms, α and β, which differ in their amount of glycosylation and affinity for heparin. After binding to this glycosaminoglycan, the native AT conformation, relatively inactive as a protease inhibitor, is converted to an activated form. Recently a new compound, named TMI, was discovered with nanomolar affinity to antithrombin and shown to be able to induce a partial activation of antithrombin, which in turn facilitates the interaction with heparin. In this context, the present work intends to characterize the effects of TMI binding on AT structure and dynamics through a series of MD simulations. The GROMACS package was employed with GROMOS 43a1 force field, the native and activate α-AT unbounded and bounded to heparin and TMI were selected for simulations under SPC water models, the PME method was applied in the calculation of electrostatic interactions, and simulations were performed in triplicate at 310K for 0.2µs. Molecular mechanics Poisson - Boltzmann surface area (MMPBSA) was used to estimate binding free energy of heparin and TMI, showing a correlation to the experimentally determined affinities. The reactive center loop (RCL), between residues Glu377 to Leu400, on the TMI-bounded AT simulations shown a flexibility behavior similar to the observed for heparin-bounded AT. As well, TMI also exposes the P1 residue of antithrombin (Arg393), as observed during the heparin-bounded AT. Combined, these data provides atomic details for TMI induced partial activation of AT, and may constitute a basis for future studies aiming TMI structural optimization.

Supported by: CNPq, CAPES and FAPERGS

[1]Centro de Biotecnologia, UFRGS
[2] Bioinformatics and High Performance Computing Research Group, Universidad Católica San Antonio de Murcia (UCAM), Spain
{pabloarantes@cbiot.ufrgs.br; hperez@ucam.edu; hverli@cbiot.ufrgs.br}

The complete chloroplast genome sequence of neotropical Myrtaceae Eugenia uniflora: organization and phylogenetic relationships

Eguiluz M. Maria[1]
Rodrigues F. Nureyev[1]
Guzman Frank[1]
Yuyama Priscila[2]
Margis Rogerio[123]

Eugenia uniflora is plant native to tropical America with pharmacological and ecological importance. The complete chloroplast (cp) genome sequence of Eugenia uniflora, the first sequenced member of the neotropical myrtaceae family, is reported here. The genome is 158,445 bp in length and exhibits a typical quadripartite structure of the large (LSC, 87 459 bp) and small (SSC, 18 318 bp) single-copy regions, separated by a pair of inverted repeats (IRs, 26 3334 bp). It contains 111 unique genes, including 77 protein-coding genes, 30 tRNAs and four rRNAs. The genome structure, gene order, GC content and codon usage are similar to the typical angiosperm cp genomes. Entire cp genome comparison of E.uniflora and three others Myrtaceas revealed an expansion in the intergenic spacer located between IRA/large single copy (LSC) border and the first gene of LSC region, driven by sequence of 43 bp. Simple sequence repeat (SSR) analysis revealed that most SSRs are AT-rich, which contribute to the overall AT richness of the cp genome. Additionally, fewer SSRs are distributed in the protein-coding sequences compared to the noncoding regions. Phylogenetic analysis demonstrated a close relationship between Eugenia uniflora and Syzygium cuminis in Myrtaceae. The complete cp genome sequence of E.uniflora reported here will facilitate population, phylogenetic and evolutionary studies

[1] PPGGBM, Departamento de Genética, Universidade Federal do Rio Grande do Sul - UFRGS,Porto Alegre, Rio Grande do Sul, Brazil
[2] Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil
[3] Departamento de Biofisica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil

Insights on the Dynamics and Thermostability of an Archaeal Oligosaccharyltransferase

Conrado Pedebos[1]
Hugo Verli[1]

N-glycosylation is one of the most prevalent post-translational modifications. The addition of newly formed glycan chains to a nascent polypeptide is fulfilled by the oligosaccharyltransferases PglB (Bacteria), AglB (Archaea), and Stt3 (catalytic subunit in Eukarya). Previous crystallographic data for these enzymes identified structural units with distinct functions, such as catalysis (CC) and structural stability (IS, P1, and P2). Studies regarding these enzymes and its units may support the development of vaccines and glycoprotein engineering. Therefore, here, we examine the predicted role of P1 unit from AglB, aiming to contribute with insights for the comprehension of AglB thermostability. We performed molecular dynamics simulations of the following systems: i) wild type AglB, ii) AglB lacking the P1 unit (AglBΔP1), and iii) P1 unit, as a control. The force field employed was GROMOS54a7, at a temperature of 356 K, in the presence of explicit aqueous solvent, catalytic ion, and a membrane bilayer composed of palmitoyl-oleoyl-phosphatidylethanolamine lipids. AglB maintained a stable behavior during simulations, while the AglBΔP1 structure became highly unstable, losing contacts and secondary structure elements, although not influencing the large transmembrane section. Nevertheless, AglBΔP1 maintained some organization at the catalytic site, such as the coordination of the metal ion with the catalytic residues, which indicates the many roles played by Zn2+, acting both as a catalytic and as a structural ion. Additionally,we detected regions with conserved residues that preserves the strong interaction between P1 unit and AglB, and could be the target of mutational studies. This data may provide the basis for the engineering of thermostable oligosaccharyltransferases from different species. The insertion of a P1 unit at the C-terminal end of the bacterial PglB would be the first step on the generation of chimeric OSTs with biotechnological applications.

Supported by: CNPq, CAPES and FAPERGS

[1] Centro de Biotecnologia, UFRGS, Caixa Postal 15005
{conrado.pedebos@ufrgs.br; hverli@cbiot.ufrgs.br}

A comparison of molecular dynamics simulations using GROMACS with GPU and CPU

Alex Dias Camargo[1],
Adriano V. Werhli[1],
Karina dos Santos Machado[1]

1 Introduction

According to Zhou et al. [13] mathematical modelling is central to systems and synthetic biology . The use of simulations is a common means for analysing these models and it is, usually, computationally intensive. High-performance computing holds the key to making significant biomolecular calculations [4] and the complexity of computational calculations has historically been extremely high [3]. The Molecular Dynamics (MD) simulations share a need for fast and efficient software that can be deployed on massive scale in distributed computing [7]. MD is based on Newton's perception, that from the starting position, it is possible to calculate the next position and velocities of the atoms in a small time interval and the forces in the new position [2]. MD is used in physics, biology and chemistry where systems of several million of atoms are simulated for days or weeks prior to completion [12]. Hardware advances brings supercomputing power to the desktop computer, thus facilitating the widespread use of parallel algorithms by bioinformaticians [11]. GPUs (Graphics Processing Units) are different from CPUs (Central Processing Units) in several fundamental ways that impact how they can be executed [10]. Various MD implementations that utilize GPUs to gain notable performance over CPUs have been described [1, 8, 9]. In this paper we show the performance of MD simulations using versions of GROMACS (Groningen Machine for Chemical Simulation) on GPU and CPU. The GPU performance was evaluated and it is shown that it can be more than 50% faster than a conventional method running on a multiple CPU core. This paper is organized as follows: Important features of the GPU architecture and CUDA (Compute Unified Device Architecture) programming model are described in Section 2. Section 3 introduces the MD and the software suite GROMACS. The obtained results are described in Section 4. Finally, Section 5 presents the discussion and conclusion.

2 GPU architecture and CUDA

A GPU is a massively parallel computer designed to accelerate computationally intensive applications which operate in a single-instruction multiple-thread (SIMT) mode [11]. With the enhanced programmability of GPUs, these chips are now capable of performing more than the specific graphics computations they were originally designed [4]. CUDA Toolkit provides a comprehensive development environment for developers building GPU-accelerated applications. It includes a compiler for NVIDIA GPUs, math libraries and tools for optimizing the performance of applications [6]. We have implemented the proposed application using CUDA Toolkit 7.

[1] Programa de Pós-graduação em Computação, FURG, Caixa Postal 96201-900
{alexcamargo, werhli, karina.machado}@furg.br

3 Molecular Dynamics Simulation and GROMACS

The MD emerged as one of the first simulation methods from the pioneering applications to the dynamics of liquids by Alder and Wainwright and by Rahman in the late 1950s and early 1960s [5]. Modeling and visualization of atomic level details provides insight into the function and dynamics of biomolecular structures [10]. The core of process is the computation of distances between pairs of atoms, and subsequently the forces and/or energies that they exert on each other [12]. In accordance with [3], a good MD simulation in GROMACS depends on three factors: the speed of the computational part in isolation, the efficiency of the parallel and communication algorithms, and the efficiency of the communication itself. In this paper we use GROMACS[1] version 5, since this software has native support for GPUs, in comparison by version 4 (OpenMP trheads native). Therefore, it is possible to perform high-throughput MD simulations. Algorithm 1 [12] illustrates the pseudocode for a MD simulation on GPU.

Algorithm 1. Pseudocode for GPU MD simulations [12]

```
Initialize the position, velocity, and force of atoms
Allocate memory for the atoms on the GPU
Transfer force, velocity, and position data to GPU
for all steps do
  Launch position GPU kernel
        Update head atom positions
        Half-update head atom velocity basedon previous force
        Zero head atom forces
        Launch force GPU kernel
  for all atoms do
    Compute force exerted by atom on head atom if within cutoff distance
  end for
Complete update of head atom velocity based on current force
end for
Transfer data from GPU to CPU
```

4 Performance Results

This paper compares the use of the software suite GROMACS running entirely on a GPU and CPU. CPUs typically provide a small number of very fast processing units, whereas GPUs have a large number of slower processing units [3]. We considered as a case study the main steps of a MD simulation of a system containing the protein Lysozyme (PDB code 1AKI) in a box of water having 38,790 atoms including solvent and ions for general purposes.

The sequential MD simulations tests were executed on identical hardware on a PC with an Intel Xeon CPU X5675 – 3,6GHz (12 processor cores), RAM 12GB, HD 1TB, running Ubuntu 14.04 64 bit Desktop Edition. A single GPU NVIDIA Quadro 600 (96 CUDA cores) was used. Table 1 shows the performance comparison between the two run methods, GROMACS 4 on CPU and GROMACS 5 on GPU, for main tasks of MD.

[1] http://www.gromacs.org/downloads

Table 1. Comparison of runtimes

TASK	Gromacs 4 (s)	Gromacs 5 (s)
Energy minimization	31,77	16,27
Equilibration (phase 1)	6207,78	2188,51
Equilibration (phase 2)	5960,86	2238,95
Production MD	22766,65	18267,44

In energy minimization, the structure is relaxed avoiding steric clashes or wrong geometry. Maximum number of steps (nsteps) to perform was 5000. This task demands less throughput data. The GROMACS 5 gain over GROMACS 4 total time reached 49%. In equilibration (phase 1 and phase 2), the structure is to brought to simulation temperature and establish the ideal orientation about the solute, also it is needed to stabilize the pressure of the system. In lines 3 and 4 of Table 1 it can be seen that the GROMACS 5 performance over GROMACS 4 total time was respectively 65% and 62% faster respectively, with nsteps = 5000. Upon completion of the two equilibration phases, we are ready to run production MD for data collection. With nsteps = 500000, GROMACS 5 perfomance also exceeded by 20% of the GROMACS 4 total time. These results show the advantages of GPU use whenever we have a high throughput data.

5 Discussion and conclusion

MD simulations of macromolecules are extremely computationally demanding, which makes them a natural candidate for implementation on GPUs [10]. The time required for such a step depends normally on the system. Advances in performance make it possible to calculate complex biomolecular interactions and function using a single desktop computer.

In this paper we disucss how MD simulations can benefit from the computing power of GPUs. The results show that the performance of MD simulations of proteins on GPU using GROMACS is attractive. As future work we intend to run MD simulations with different proteins and MD parameters, also considering other GPUs.

Acknowledgment

This work was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

References

[1] Anderson, Joshua A., Chris D. Lorenz, and Alex Travesset. "General purpose molecular dynamics simulations fully implemented on graphics processing units." Journal of Computational Physics 227.10 (2008): 5342-5359.

[2] Astuti, A. D., and A. B. Mutiara. "Performance Analysis on Molecular Dynamics Simulation of Protein Using GROMACS." arXiv preprint arXiv:0912.0893 (2009).

[3] Hess, Berk, et al. "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation." Journal of chemical theory and computation 4.3 (2008): 435-447.

[4] Liu, Weiguo, et al. "Molecular dynamics simulations on commodity GPUs with CUDA." High Performance Computing–HiPC 2007. Springer Berlin Heidelberg, 2007. 185-196.

[5] Meller, Jaro. "Molecular dynamics." ELS (2001).

[6] NVIDIA, CUDA. "CUDA Toolkit Documentation - v7.0 ". Available: https://docs.nvidia.com/cuda/. (2015).

[7] Pronk, Sander, et al. "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit." Bioinformatics (2013): btt055.

[8] Schmid, Nathan, Mathias Bötschi, and Wilfred F. Van Gunsteren. "A GPU solvent–solvent interaction calculation accelerator for biomolecular simulations using the GROMOS software." Journal of computational chemistry 31.8 (2010): 1636-1643.

[9] Stone, John E., et al. "Accelerating molecular modeling applications with graphics processors." Journal of computational chemistry 28.16 (2007): 2618-2640.

[10] Rodrigues, Christopher I., et al. "GPU acceleration of cutoff pair potentials for molecular modeling applications." Proceedings of the 5th conference on Computing frontiers. ACM, (2008).

[11] Vouzis, Panagiotis D., and Nikolaos V. Sahinidis. "GPU-BLAST: using graphics processors to accelerate protein sequence alignment." Bioinformatics 27.2 (2011): 182-188.

[12] Walters, John Paul, et al. "Accelerating Molecular Dynamics Simulations with GPUs." ISCA PDCCS. (2008).

[13] Zhou, Yanxiang, et al. "GPU accelerated biochemical network simulation." Bioinformatics 27.6 (2011): 874-876.

Dynamics of the Complete, Membrane Soaked hTLR4

Carla Carvalho de Aguiar[1]
Hugo Verli[1]

Human Toll-Like Receptor 4 (hTLR4) is a transmembrane protein of the immune system. Together with MD-2, a co-receptor, hTLR4 recognizes bacterial lipopolysaccharide. hTLR4 posses three domains: an extracellular domain (ECD; for ligand recognition); an helical transmembrane domain (TM); and an intracellular domain (TIR; for signaling). In a previous work, our group described the influence of the co-receptor over the hTLR4 ECD dynamics. As this model included only the receptor ECD, the present work aims to study the relation of hTLR4 and MD-2 within membranes, searching for clues to the receptor conformational activation. For this, comparative modeling (Modeller v.9) was employed to produce a model of the complete hTLR4, inserted in a POPC membrane as a single (i, hTLR4) and heterodimeric (ii, hTLR4-MD-2) systems. Both i and ii were submitted to molecular dynamics simulations using GROMACS and GROMOS 53a6 force field. We observed that MD-2 was able to interfere in hTLR4 conformational space, whereas it does not seems to be related to the principal movement of ECD. In both systems, the TIR and ECD became closer to the membrane, and Leu661 of TM seems to be involved in the conformational connection between TM and TIR. Moreover, dynamical network analysis shows that MD-2 induced the formation of a more adherent community of conformational connected residues in TIR. Once we had inferred, previously, a main movement of ECD in solution, now we could identified this behavior without the bias of the absence of other domains or membrane. Furthermore, TM mobility could be influencing TIR dynamics in relation to membrane through specific residues.

Supported by: CNPq, CAPES and FAPERGS

[1]Centro de Biotecnologia, UFRGS, Caixa Postal 15005
{carlaaguiar@cbiot.ufrgs.br; hverli@cbiot.ufrgs.br}

A proposal to the manipulatioon of a set of protein structures from PDB

Vinicius Rosa Seus [1]
Karina dos Santos Machado[2]
Adriano Velasque Werhli[3]

Protein Data Bank (PDB) is a public web database with more than 100,000 biological macromolecular structures. With this large amount of protein structures available on PDB the use of tools for acquisition and analysis of specific sets of biological macromolecules is a necessity. Hence, in this work we propose the development of a tool for acquiring, storing and analyzing specific sets of proteins from PDB. The proposed tool runs on desktop environment allowing the user to acquire the structures from the RESTful web-service provided by PDB server. After the acquisition of a set of interesting PDBs the user can manipulate these data in an off-line environment through a local database that stores the information about the characteristics of the structures, for example, ligands, mutations, residues, sequences and docking results. The protein files are locally stored on users' computer and can be used, for instance, for molecular docking simulations and alignment of sequences and structures. Having a set of proteins of interest available locally and using our proposed tool the user can perform analysis related to alignments and visualize important proteins characteristics improving the knowledge about specific target. Besides, the user can select PDB files to be visualized on a graphical environment that is integrated in our tool. Other features are related to the exporting of sequence alignments results in csv format or exporting sequences that have a similar identity in a format that can be easily loaded on graph tools. These alignments allow user to visualize which proteins are similar and discard those that are not. As future work we propose to conduct an study using our tool for acquiring and analyzing a set of betaglucosidase molecules, which is the enzyme responsible for the extraction of fermentable sugars from the sugar cane bagasse used in the industry for bio-ethanol production.

[1] C3 - Centro de Ciências Computacionais, FURG, Av. Itália km 8 Bairro Carreiros, Brazil, Rio Grande do Sul, Rio Grande {viniciusseus@furg.br}
[2] C3 - Centro de Ciências Computacionais, FURG, Av. Itália km 8 Bairro Carreiros, Brazil, Rio Grande do Sul, Rio Grande {karina.machado@furg.br}
[3] C3 - Centro de Ciências Computacionais, FURG, Av. Itália km 8 Bairro Carreiros, Brazil, Rio Grande do Sul, Rio Grande {werhli@furg.br}

Mother´s Vaginal Microbiota Shares Few Phylotypes With Preterm Infants

Priscila Caroline Thiago Dobbler[1]

Miriane Acosta Saraiva[2]

Andréa Lúcia Corso[2]

Rita de Cassia Silveira[2]

Renato Soibelmann Procianoy[2]

Luiz Fernando Wurdig Roesch[1]

Evidences generated with new cultivation free molecular tools, challenges the established dogma that the fetus remains sterile until delivery. Microbes derived from the mother's vagina, gut or placenta might colonize the fetus's gastrointestinal tract, however, the individual contribution of each mother's body site to such early colonization is still unknown. Here we determined the percentage of microbes shared between the mother's vagina and the newborn. For this analysis, we performed microbial DNA extraction of a single vaginal swab from 11 pregnant women and second evacuation of their respective preterm infants. The gestational period ranged from 27 to 32 weeks. Nine mothers gave birth through C-section and two had normal (vaginal) delivery. One mother gave birth to twins. Following DNA extraction, the V3 region of the 16S rRNA gene was amplified from all samples, using the 515F and 806R primers and the amplicons were sequenced using the Ion Torrent-PGM technology. Quality filtering and assembling of sequences into Operational Taxonomic Units were performed according to the Brazilian Microbiome Project pipeline, available at: http://www.brmicrobiome.org. Microbial phylotypes shared between mother´s vagina and preterm infants varied among samples. Mothers shared 0 to 42.59% (Average of 13.68% and SD = 13%) of their vaginal microbiota with their infants. Moreover, the newborn's microbiome comprised of around 0 to 34% (Average of 15% and SD = 10%) of phylotypes that were similar to the mother's vagina. In summary, our results indicated that the microbiota from mother's vagina is a source of colonization of the fetus's gastrointestinal tract but the number of microbes shared between mother and the fetus is variable and in some cases, no microbes derived from the mother's vagina. These results indicated that the fetus might also receive microbes from another source, such as the placenta or the mother´s gastrointestinal tract.

[1] Universidade Federal do Pampa – São Gabriel, UNIPAMPA, Endereço: Av. Antônio Trilha, 1847 - São Gabriel - RS - CEP: 97300-000
{Priscila Caroline Thiago Dobbler, prisciladobbler@gmail.com} {Miriane Acosta Saraiva, mirisdabio@gmail.com} {Luiz Fernando Wurdig Roesch, luizroesch@unipampa.edu.br}
[2] Universidade Federal do Rio Grande do Sul, UFRGS, Av. Paulo Gama, 110, Porto Alegre
{Andréa Lúcia Corso, andrea.lucia@terra.com.br} {Rita de Cassia Silveira, drarita.c.s@gmail.com}

CrossTope: how this database can impact new vaccine strategies development

FREITAS, Martiela Vaz[1]
BRAGATTE, Marcelo Alves[1]
ANTUNES, Dinler Amaral[1,2]
RIGO, Maurício Menegatti[1]
MENDES, Marcus de Almeida[1]
SINIGAGLIA, Marialva[1]
VIEIRA, Gustavo Fioravanti[1]

One of the greatest challenges in immunology is the discovery of appropriate targets to vaccine development. The choice relies on the identification of elements responsible for the stimulation of immune responses. These features could define the cross-reactivity among different pMHC-I complexes (one T Cell Receptor-TCR could recognize more than one pMHC-I). Focusing on cross-reactivity, our group developed the CrossTope Data Bank, a curated repository of 3D structures of pMHC-I complexes.]One of the greatest challenges in immunology is the discovery of appropriate targets to vaccine development. The choice relies on the identification of elements responsible for the stimulation of immune responses. These features could define the cross-reactivity among different pMHC-I complexes (one T Cell Receptor-TCR could recognize more than one pMHC-I). Focusing on cross-reactivity, our group developed the CrossTope Data Bank, a curated repository of 3D structures of pMHC-I complexes. One of the greatest challenges in immunology is the discovery of appropriate targets to vaccine development. The choice relies on the identification of elements responsible for the stimulation of immune responses. These features could define the cross-reactivity among different pMHC-I complexes (one T Cell Receptor-TCR could recognize more than one pMHC-I). Focusing on cross-reactivity, our group developed the CrossTope Data Bank,a curated repository of 3D structures of pMHC-I complexes. One of the greatest challenges in immunology is the discovery of appropriate targets to vaccine development. The choice relies on the identification of elements responsible for the stimulation of immune responses. These features could define the cross-reactivity among different pMHC-I complexes (one T Cell Receptor-TCR could recognize more than one pMHC-I). Focusing on cross-reactivity, our group developed the CrossTope Data Bank, a curated repository of 3D structures of pMHC-I complexes.

[1] Núcleo de Bioinformática do Laboratório de Imunogenética, Departamento de Genética,
Universidade Federal do Rio Grande do Sul, Postcode 91501-970, Brazil; {imunoinfo@gmail.com}

[2] Department of Computer Science, Rice University, Houston, Texas, 77005, USA.
{dinler@gmail.com}

The Proteins are Linearized before the Proteasome Degradation: Are you sure?

TARABINI, Renata Fioravanti[1]
GUTIERRES, Matheus de Bastos Balbé[1]
FREITAS, Martiela Vaz[1]
SINIGAGLIA, Marialva[1]
VIEIRA, Gustavo Fioravanti[1]

The proteasome is a protein complex composed by a central pore with proteolytic activity combined with two regulatory components that recognize polyubiquitinated proteins and direct them to the catalytic core for degradation. This pathway is the main mechanism of protein digestion in eukaryotes and presents an important role in processing and presentation of epitopes to T cells in cell-mediated immunity. It is well known that the substrate enters the proteasome central cavity lacking its native 3D-structure, and then passes to the substrate binding channel where the peptide bond hydrolysis takes place. Research has focused efforts manly on the substrate primary sequence influence to the substrate processing inside the channel and still remains dubious the existence of some level of secondary structure. This fact suggest the hypothesis that spatial characteristics among the spectrum of secondary structure (beta sheets and alpha helices), could drive preferential cleavage sites. Our group has been working on a way to integrate and automate the interpretation of the outputs from Netchop 3.1 and Stride programs. Given a pdb file as an input for our workflow, Stride program assigns secondary structure for each amino acid (AA) and generate files in fasta format, which are used as input to the Netchop 3.1 program. In this program, the cleavage sites are predicted and recovered when an associated probability (higher than 70%) is found. The script was developed using python, and the algorithm consists in the correlation of proteolytic sites with their related secondary structure. We evaluated all nonredundant structures from human and viral proteins hosted in Protein Data Bank. Our initial results point to a tendency for the occurrence of cleavage in beta strands conformation, what would represent the presence of some structuration degree of the substrate before the degradation by the proteasome.

Finantial supported: Bill & Mellinda Gates Foundation, Fapergs, Capes.

[1] Núcleo de Bioinformática do Laboratório de Imunogenética, Departamento de Genética, Universidade Federal do Rio Grande do Sul, Postcode 91501-970, Brazil
{imunoinfo@gmail.com}

Differentially Expressed Genes During Salt Stress in Rice (*Oryza sativa* L.)

Daniel da Rosa Farias[1], Luis Willian Pacheco Arge[1], Marcelo Nogueira do Amaral[2], Rodrigo Danielowski[1], Letícia Carvalho Benitez[2], Eugenia Jacira Bolacel Braga[2], Antônio Costa de Oliveira[1] and Luciano Carlos da Maia[1]

Rice is the staple food of more than two thirds of the world´s population, it is the second most widely grown cereal in the world, and Brazil is the largest producer outside Asia. This cereal is sensitive to salinity, which affects one fifth of irrigated lands worldwide. Developments in molecular marker and bioinformatics analyses became important tools to understand the gene responses related to abiotic and biotic stresses. Thus, this study aims to identify differentially expressed genes (DEGs) present in QTLs associated to salt tolerance in rice (cv. "BRS Querencia"). RNA-seq technique was used to analyze the transcriptional profile and elucidate the DEGs of rice under salt stress, total RNA was extracted from leaves in stage V3, exposed to stress for 24 hours. The paired-end sequencing of the cDNA libraries was performed on a HiSeq 2500 R Illumina platform. For the analysis and visualization of the read qualities, the software FastQC was used. Thereafter, the program Trimmomatic was used to remove low-quality bases and adapters from each library. The reads were mapped against the rice reference genome from The Rice Annotation Project Database, using the software TopHat, which uses software Bowtie to mapping. Software HTseq was used to count reads, and identification of DEGs was performed using software R with edgeR packaged (Bioconductor repository). Expression levels were normalized by the method RPKM considering differentially expressed genes with $P<0.01$ by Fisher's exact test. After a literature search, nine QTLs associated with salt tolerance in rice were used in this study. To align the DEGs in QTLs regions, the software BLAST was used. A total of 40 DEGs identified by analysis of RNAseq associated with QTLs were found. Thereby, these genes may be important for marked-assisted selection in breeding programs, improving the precision and efficiency in the selection for salt tolerance in rice.

Research supported by the following funding agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Rio Grande do Sul (FAPERGS).

[1] Plant Genomics and Breeding Center, Universidade Federal de Pelotas, Caixa Postal, 354
fariasdr@gmail.com
[2] Department of Plant Physiology, Universidade Federal de Pelotas, UFPel.

Characterization of Transposable Elements in Leptopilina boulardi Genome

Filipe Zimmer Dezordi[1]
Alexandre Freitas da Silva[1]
Elgion Lucio da Silva Loreto[2]
Paulo Marcos Pinto[1]
Gabriel da Luz Wallau[3]

Transposable elements (TEs) are mobile genetic elements present in the genome of most organisms studied so far [1]. TEs are grouped into two types, Type I transposons which uses a DNA intermediate in their mobilization process, and Type II that uses an RNA intermediate [2]. The characterization of these elements in new sequenced genomes becomes important to better understand the genome evolution as the ratio between mobile and non-mobile genome components. This study aimed to characterize the mobilome present in the wasp Leptopilina boulardi genome. After obtaining the wasp genome sequences, bioinformatics analyzes were conducted with RepeatExplorer tool, which uses a new approach based on clusterization of highly repetitive reads from next generation sequencing platforms. Our preliminary data shows several TEs inhabit L. boulardi genome, both Type 1 and Type 2 TEs. After an initial characterization we analyzed the TEs clusters with CENSOR tools to determine which homologues are present in RepBase and we are going to use ORF Finder to search for open reading frames, Basic Local Alignment Search Tool to search similarity against the NCBI. We obtained a total of 53 clusters representing the TE's. Where the elements of the Type I represented a total of 2.823% of the genome while the Type II represented a total of 2.37%. The elements that were present in greater amounts were LTR-Gypsy, representing 1,921%. The abundance of each TEs family is highly variable in insect genomes but, in general, Type 2 elements are more abundant than Type 1 although large variations in the contribution of each TE order are observed [3], this parameter is not found in L. boulardi genome, where type 1 elements are found in greater number, suggesting that the abundance of TEs in this genome can be result of a different host genome evolutionary history than other insects.

We thank CAPES, CNPq and Fapergs.

[1]Laboratório de Proteômica Aplicada, UNIPAMPA
{Filipe Zimmer Dezordi, zimmer.filipe@gmail.com; Alexandre Freitas da Silva, alexfreitasbiotec@gmail.com; Paulo Marcos Pinto, paulomarcospinto@gmail.com}
[2]Departamento de Bioquimica e Biologia Molecular, UFSM
{Elgion Lucio da Silva Loreto, elgionl@gmail.com}
3 Departamento de Ecologia, Zoologia e Genética, UFPEL
{Gabriel da Luz Wallau, gabriel.wallau@gmail.com)

Potential use of GROMOS force field parameters for organic molecules

Marcelo Depólo Polêto[1]
Hugo Verli[1]

Techniques as docking and molecular dynamics have been applied around the world to rationally engineer molecules aiming to tackle different purposes. So far, force fields for organic compounds, like GAFF, have been designed to reproduce quantumly calculated parameters and, thus, do not take in account condensed phase terms. One recent exception was OPLS/AA applied to organic molecules (Caleman et al., 2011 - dx.doi.org/10.1021/ct200731v). With this in mind, the aim of this work is to create a small molecule force field based on GROMOS. Thus, charges, bonded and non-bonded parameters of GROMOS54A7 were used to create organic compounds topologies. After, we have simulated their properties using GROMACS 5.0.4, applying the same protocol as Caleman et al.. Therefore, density and enthalpy of vaporization were calculated for each compound, along with absolute error in relation to experimental values. Absolute error values were used to determine whether the quality of each topology was acceptable. When required, manual adjusts on topology were done to guarantee its confidence. So far benzen, 1H-pyrrole, fluorobenzen, 1,2-difluorobenzene, 1,3-difluorobenzen, 1,2,3,5-tetrafluorobenzen, pyridine and pyrimidine were succesfully parameterized. Regarding density, all compounds had absolute error below 5%, except 1H-pyrrole(5,05%) and pyrimidine (8.46%). Regarding enthalpy of vaporization, all compounds also had absolute error values below 5%, except 1H-pyrrole (9.30%), 1,2-difluorobenzen (5.76%) and 1,2,3,5-tetrafluorobenzen (10,38%). Moreover, other molecules are being built in order to expand our database and other properties will be included in our analyzes to guarantee a good description of more termodynamical properties. These preliminary results suggest the potential use of GROMOS43A7 parameters to create organic molecules with good agreement with their physical-chemical properties, and can lead to a rich database for drug design purposes.

Supported by: CNPq, CAPES and FAPERGS

[1] Centro de Biotecnologia, UFRGS, Caixa Postal 15005
{marcelodepolo@gmail.com; hverli@cbiot.ufrgs.br}

Genetic Mapping of Diseases with NGS using Big Data Analysis

Julio C. S. dos Anjos[1]
Junior F. Barros[1]
Raffael B. Schemmer[1]
Claudio F. R. Geyer[1]
Ursula Matte[2]

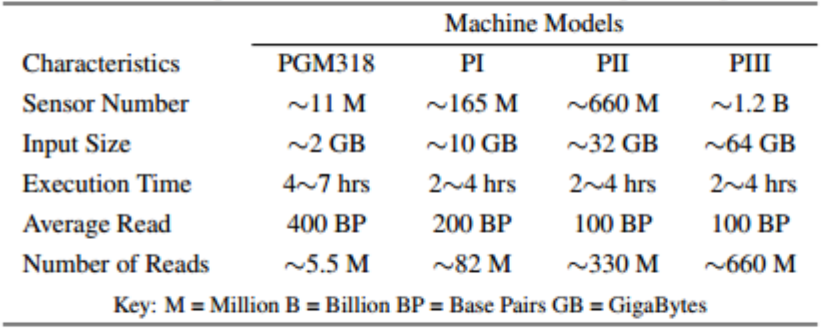
Abstract: The development of advanced genetic sequencing techniques has allowed creating what is called Next-Generation Sequencing (NGS) to provide new advances into genetics field. However, new problems related to processing of significant amounts of data were discovered and begin to turn out complexity to solve genetic sequencing problems. The necessity of new methods of processing grows the towards the NGS technology. Also a fast and accurate analysis is relevant in the context of diseases due to time for treatment. This study proposes the development of genetic notation systems through the MapReduce framework to allow disease analyses in an automated way to provide tools for researchers and doctors in a clinical environment.

Introduction

The information provided by DNA is crucial to improve the development of several areas of biological research. The studies about new pharmaceutical substances, foodstuffs, pesticides and agricultural products are clearly benefited from biotechnology and their insights [1]. One the most focused fields in terms of efforts is the clinical analysis, which is strongly and constantly improved by research, in case specially the detection of pathologies will be centered in this work. When looking a few years ago, DNA sequencing was considered an expensive technique, but it has changed since the Next-Generation Sequencing appears and a migration happened from Sanger. Next-Generation Sequencing or then called as NGS could provide a great solution looking the cost-effective relation in comparison to Sanger, although, NGS is a technique that produces a big amount of data after a sequencing process. As example shown in Table 1, NGS machines particularly those that use ion semiconductors (Ion Torrent PGM or Proton families) are able to sequence million of base pairs in few hours, computationally it means high throughput of data very quickly.

[1]Institute of Informatics, UFRGS, PoBox 15064
{jcsanjos@inf.ufrgs.br, jfbarros@inf.ufrgs.br, raffael.schemmer@inf.ufrgs.br, geyer@inf.ufrgs.br}
[2]Genetic Therapy Laboratory, Clinical Hospital of Porto Alegre, UFRGS
{umatte@hcpa.edu.br}

Table 1. Technical specifications of Ion Torrent sequencing machines.



	Machine Models			
Characteristics	PGM318	PI	PII	PIII
Sensor Number	~11 M	~165 M	~660 M	~1.2 B
Input Size	~2 GB	~10 GB	~32 GB	~64 GB
Execution Time	4~7 hrs	2~4 hrs	2~4 hrs	2~4 hrs
Average Read	400 BP	200 BP	100 BP	100 BP
Number of Reads	~5.5 M	~82 M	~330 M	~660 M

Key: M = Million B = Billion BP = Base Pairs GB = GigaBytes

Therefore, the techniques used nowadays for processing the unprecedent amount of data generated by NGS are limited and falls when lack the scalability or response to ease the diagnosis process. Several variations of human genome can be determined by massive parallel sequencing. However, many of these are not clinically relevant, the need of methods to discriminate between disease-causing mutation and normal genetic variability is a short run-time [4]. The purpose of this work is to present a model of system that can assist doctors in clinical diagnosis of patients by conducting an analysis of the genetic mutations contained in their DNA. The scientific goal is to provide an efficient and robust method for the genetic mapping of diseases through NGS Big Data.

Mutations, Polymorphism and Clinical Genetics

A mutation is defined as a change in the nucleotide sequence of an organism, it can be caused by an irreparable damage suffered by the genome, errors in the replication process or insertion/deletion of DNA fragments by mobile genetic elements. Several studies such as [5] suggests that mutation changes the proteins that a gene produces and it causes dangerous consequences for the organism. Polymorphism is a kind of mutation that takes place with a frequency greater than 1% in a population and can be divided into distinct, well defined classes. According [6] there are three classes of mutations, those alter the number of chromosomes in a cell (genomic mutations), those alter structures of specific chromosomes (chromosomal mutations) and mutations that change individual genes (gene mutation).

MapReduce

MapReduce is a programming framework that abstracts the complexity of parallel applications by partitioning and scattering data sets across hundreds or thousands of machines, and by bringing computation and data closer [2]. In case, the MapReduce is an attractive solution to imply as it offers a parallel distributed solution for processing in high volumes of data such as those produced by sequencers as it free certain difficulties from programmers when developing distributed systems. Beyond that, using Apache Hadoop implementation, it has its own solution for distributed file system called HDFS. Using MapReduce to develop an application applied to bioinformatics could provide a reduction of complexity [3] focusing efforts to solve problems.

Description of the Model

This study implements a genetic mapping by using MapReduce to analyze and process the amount of data generated by the sequencing process. Three stages compose this work through the Hadoop implementation. A filtering step executes a script to preprocessing that saves in memory the gene from the patient to form a single input. This filtering occurs before to begin the first phase of MapReduce processing. In the first stage, the patient's gene is sent to a distributed file system to be processed from nodes. In the second stage, each gene with its respective identifier is evaluated against to the reference genome to find a gene position that has some anomaly using CCDS database. In the second stage, if some anomaly was found, a key/value is emitted in memory for each node. Where the key is the patient identifier and the value, the gene position changed. This model assumes that the input file is formed by different kinds of genes and patients. The third stage is to compose the annotation by using the MapReduce processing. Figure 1 illustrates this process.

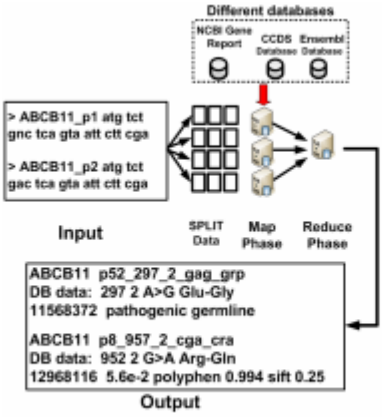

Different databases
NCBI Gene Report
CCDS Database
Ensembl Database
> ABCB11_p1 atg tct gnc tca gta att ctt cga
> ABCB11_p2 atg tct gac tca gta att ctt cga
Input
SPLIT Data
Map Phase
Reduce Phase
ABCB11 p52_297_2_gag_grp
DB data: 297 2 A>G Glu-Gly
11568372 pathogenic germline
ABCB11 p8_957_2_cga_cra
DB data: 952 2 G>A Arg-Gln
12968116 5.6e-2 polyphen 0.994 sift 0.25
Output

Figure 1. Flowchart of the proposal.

In the Map phase, a Combine execution compares the gene anomaly found against both Ensembl and Gene Report databases. The search process uses the position and the name of the gene to query the databases, if a pathology is reported, the Map emits a key/value pair intermediate. The key is formed by gene's position and the value within a tuple of a reference gene and patient's pathology. After, the reduce function emits a new key/value pair with the information associated about all pathologies found for each patient and saves on HDFS. Only mutations found in databases are written in the output, followed by messages from the associated pathological informations.

Conclusion

This study proposes a model for genetic mapping of diseases through Big Data analysis from different patients simultaneously in a scalable fashion, to support researchers and doctors in an automated way. MapReduce framework establishes a coherent model to the development of genetic notation solutions and provides a high level of parallelism in cluster or cloud environments.

References

[1] William J, T. and Palladino, M. A. (2012). Introduction to Biotechnology, volume 1. Pearson, 3rd edition.

[2] Dean, J. and Ghemawat, S. (2010). MapReduce - A Flexible Data Processing Tool. Communications of the ACM, 53(1):72–77.

[3] Zou, Q., Li, X.-B., Jiang, W.-R., Lin, Z.-Y., Li, G.-L., and Chen, K. (2014). Survey of MapReduce frame operation in bioinformatics. Briefings in Bioinformatics, 15(4):637–647.

[4] Frebourg, T. (2014). The challenge for the next generation of medical geneticists. Hum Mutat, 35(8):909–11.

[5] Johnsen, J. M., Nickerson, D. A., and Reiner, A. P. (2013). Massively parallel sequencing: the new frontier of hematologic genomics. Blood, 122(19):3268–3275. [6] Nussbaum, R., McInnes, R., and Willard, H. (2013). Thompson Genetics in Medicine. Elsevier Science Publishers B. V., 7th edition.

Analyzing Microarray Data Using Gene Regulatory Networks Cycles

Fabiane Cristine Dillenburg[1], Alfeu Zanotto-Filho[2], José Cláudio Fonseca Moreira[2],
Leila Ribeiro[1], Luigi Carro[1]

Introduction

Gene expression provides information for building models of biological systems. Gene expression analysis comparing normal and neoplastic tissues have been used to identify genes associated with tumor genesis and potential therapeutic targets [1]. Genomic high-throughput technologies, such as microarrays, may considerably facilitate the molecular profiling of human tumors. Thousands of genes can now be analyzed using a single microarray hybridization chip [2]. The expression profile from a single tumor reflects the state of events of an individual malignancy at a certain time point. To generalize the findings and provide conclusive evidence for the involvement of a molecular alteration, it is often necessary to analyze several hundred tumors. Using traditional molecular pathology, such verification could take several months, or even years, to reach completion. To facilitate translational research in a large-scale manner, new techniques are needed. One of the main research areas in systems biology concerns the discovery of biological regulatory pathways or networks from microarray datasets [3]. A gene regulatory network (GRN) consists of a great number of genes whose expression levels affect each other in various ways. Computational models of GRNs can take a variety of forms, include models comprised of directed and undirected graphs. The process of constructing a regulatory network that explains some behavior of the cell using microarrays can be done in two steps: first, the set of genes that is thought to be relevant for this biological process is selected and measured in the microarray (for all the samples); then, the gene expression data is analyzed to generate graphs that represent the desired regulatory network. In this work, we present a new way of analyzing microarray datasets, based on the different kind of cycles found among genes of the GRN constructed using quantized data obtained from the microarrays. A cycle is a closed walk with all vertices (and hence all edges) distinct (except the first and last vertices). Thanks to the new way of finding relations among genes, a more robust interpretation of gene correlations is possible. Furthermore, the cycles help differentiate, measure and explain the phenomena identified in healthy tissue and diseased tissue. We use the proposed methodology to analyze the genes of three networks closely related with cancer - apoptosis, glucolysis and cell cycle - in tissues of the most aggressive type of brain tumor (Gliobastomamultiforme – GBM) and in healthy tissues. Because most patients with GBMs die in less than a year, and essentially no patient has long-term survival, these tumors have drawn significant attention [4].

Methodology

Firstly, we selected the main genes involved in the pathway of interest. The gene expression analysis comparing normal and GBM tissues was performed from previously published and characterized database comprising 276 GBM samples of all histology

[1] Instituto de Informática, UFRGS, Porto Alegre, Rio Grande do Sul, Brazil.
{fabiane.dillenburg, leila, carro@inf.ufrgs.br}

[2] Departamento de Bioquímica - ICBS, UFRGS, Porto Alegre, Rio Grande do Sul, Brazil.
{alfeuzanotto@hotmail.com, 00006866@ufrgs.br}

compared to 8 brain samples of non-neoplastic white matter tissue. The raw data for this study is available as experiment number GSE16011 in the Gene Expression Omnibus[2]. Experimental data used in the analysis are available in AffymetrixGeneChip Human Genome U133 Plus 2.0 Array format. The analyses of Affymetrix microarray data were performed using R[3] and Bioconductor[4]. Our analysis method consists of the following steps:

1) **Preprocessing of Affymetrix microarray data.** This step consists of importing the raw data in and summarizing the expression values per each probe set. The process of summarizing expression values is constituted of (a) background correction; (b) normalization; (c) summarization. All these operations are supported in the Bioconductor package *affy*.

2) **Annotation data.** The purpose of the annotation is to provide detailed information about the data. These operations are supported in the Bioconductor package *annotate* and *hgu133plus2.db*. We created a dataframe with the features names, the genes symbols and the summarized data samples. We then selected the data from the genes of interest through its symbol. These data were divided in GBM samples and control samples.

3) **Sigmoidal normalization.** This step reduces the influence of extreme values or outliers in the data without removing them from the dataset. Data are nonlinearly transformed by using a sigmoidal function [5] and the normalized values range from 0 to 1.

4) **Spearman's Correlation.** Correlation is used to discover sets of genes with similar expression profiles. Spearman's rank correlation coefficient is non-parametric and allows one to identify whether two variables (genes) relate in a monotonic function.

5) **Graphs.** The undirected graphs (representing the GRN) are constructed by computing a correlation coefficient for each pair of genes. If the coefficient is above a certain threshold and is statistically significant ($p<0.05$), the gene pair gets connected in the graph, if not, it remains unconnected. We used the R package *igraph* for manipulating undirected graphs. The adjacency matrix used is a matrix with correlation coefficients greater than the defined threshold.

6) **Cycles.** In order to seek the biological explanation of the observed gene association, we decided to look for cycles in the gene network. We used a C++ implementation of Johnson's algorithm [6] to find the cycles in the graphs. Feedback mechanisms are very common in biological networks. Our hypothesis is that negative feedback could allow one to find relations among genes that could help explain the stability of the regulatory process within the cell. Positive feedback cycles, on the other hand, can show the amount of imbalance a certain cell is suffering when in that state. The genes of interest are of two types: activators and inhibitors. We assume that a cycle of a graph is positive when the number of inhibitors in the cycle is zero. Similarly, it is said to be negative when the number of inhibitors in the cycle is greater than or equal to one.

Results

We use the proposed methodology to analyze the genes of three networks closely related with cancer - apoptosis, glucolysis and cell cycle. The results evidence differences between the GRNs of the three networks among the control samples and GBM samples. The cycles of the control graphs use about all the genes of each network; while the cycles of the GBM graphs using a small group of genes of each network. In apoptosis, only a few cycles

[2] http://www.ncbi.nlm.nih.gov/geo/

[3] http://www.r-project.org/

[4] http://www.bioconductor.org/

were found in the GBM graph, it would indicate that the cell cannot die [7]. In the cell cycle, the greater number of cycles has been found in GBM, which might indicate that the tumor has more active cell cycle mechanisms, since it is more proliferate [7]. Analyzing the most common genes found in the cycles, we observed that the t-test with 0.05 significance level indicated that there is no significant difference between the average of the gene expression level of the control samples and the GBM samples of some genes of the three pathways. Thus, there is a gain of information with the analysis using cycle to analysis of the gene expression level, whereas cycles highlight the difference between the control samples and the GBM samples.

Discussion

In literature, some articles focus on how to detect biologically meaningful modules [8] and recurring patterns called motifs [9] in networks. Our analysis is focused on the genes from a pathway so the goal was not to identify modules, pathways or motifs, but rather to better understand the relationships of the genes of the pathway of interest and its variations between samples of GBM and control to get insights on how alterations in the levels of inhibitors may affect the activation of the pathway based on target genes evaluation. The proposed approach innovates by using the existing cycles in the network for analysis, instead of using the connectivity of the whole network or the intramodular connectivity as the measure of node importance as other approaches do [10], thus providing a different and potentially fruitful strategy to analyze complex interactions in pathways.

References

[1] PARMIGIANI, G. et al. The Analysis Of Gene Expression Data: an overview of methods and software. In: PARMIGIANI, G. et al. (Ed.). The Analysis Of Gene Expression Data: methods and software. New York: Spring Verlag, 2003.

[2] STEKEL, D. Microarray Bioinformatics. Cambridge University Press, 2003.

[3] ALTAY, G.; EMMERT-STREIB, F. Inferring the conservative causal core of gene regulatory networks. BMC Systems Biology, [S.l.], v.4, p.132, 2010.

[4] MRUGALA, M. M. Advances and challenges in the treatment of GBM: a clinician's perspective. Discov. Med., [S.l.], v.15, n.83, p.221–230, 2013.

[5] PRIDDY, K. L.; KELLER, P. E. Artificial Neural Networks: an introduction. Bellingham: SPIE (The International Society for Optical Engineering) Press, 2005.

[6] JOHNSON, D. B. Finding All the Elementary Circuits of a Directed Graph. SIAM Journal of Computing, [S.l.], v.4, n.1, p.77–84, 1975.

[7] HANAHAN, D. WEINBERG, RA. Hallmarks of cancer: the next generation. Cell. v.144, n.5, p.646-74, 2011.

[8] LANGFELDER, P.; HORVATH, S. Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology, v.1, n.1, p.54, 2007.

[9] ALON, U. Network motifs: theory and experimental approaches. Nat Rev Genet, v.8, n.6, p.450-61, 2007.

[10] MA, S. et al. Incorporating gene co-expression network in identification of cancer prognosis markers. BMC Bioinformatics, v.11, n.271, 2010.

Molecular Dynamics of Stapled Peptides

Bianca Villavicencio[1]
Rodrigo Ligabue-Braun[1]
Hugo Verli[1]

Peptides are promising for drug development research due to their reduced size, their ability to traverse membranes, and a reasonable contact surface for interacting with target proteins. However, without its complete proteic framework, a peptideâs secondary structure can be easily destabilized, allowing the loss of its biologically relevant conformation. In α-helices, a strategy to address this issue is the addition of a molecular staple linking one or more turns of the helix through an all-hydrocarbon bridge. Studies of this type of structure yielded good results both in vitro and in vivo. Simultaneously, predictive models are still scarce, but would reduce research costs and allow a better understanding of stapled peptides and their potential use. One of such models is achieved by simulations by molecular dynamics, which renders information about conformational changes in the peptide. In order to allow their application in protein engineering, we parameterized and simulated both stapled and unstapled peptides, starting as either helices or coiled molecules with the potential to acquire defined structures. We evaluated staples of 6 or 8 carbons with different configurations (R-S). All simulations were carried out with the GROMOS54a7 force field in the Gromacs simulation suite. Results indicate that, while unstapled peptides tend to lose their helical form, stapled peptides tend to remain helical. In the case of initially coiled stapled peptides, the trend is to become helical. The type of staple used seems to influence the degree of helical content acquired. These alterations in secondary structure content seem to agree with experimental circular dichroism results. These results might allow the implementation of simulation protocols for stapled peptides, and also seem to describe with confidence the conformational behaviour of such peptides at the atomic scale. Further simulations with different force fields are being performed in order to obtain more data for comparison.

Supported by: CNPq, CAPES and FAPERGS

[1]Centro de Biotecnologia, UFRGS, Caixa Postal 15005
{bia.villavicencio@gmail.com; hverli@cbiot.ufrgs.br}

Entering Inflammaging pathways into Ontocancro 3.0

Laís Falcade[1]

Giovani R. Librelotto[1]

Rômulo Stringhini[1]

In the natural process of life, people are born, grow, age and die. The degeneration of the cells that occurs in this process is unconditionally, in fact, the inability to renew cell, and this failure is that live various diseases. Knowing that every human being older and, by consequence, are prone to more degrading inflammatory processes than in youth, the search for longevity is a fascination for many scholars. A topic that has been researched in recent years in this area is Inflammaging, which deals with the chronic inflammatory state that comes with aging. This work has promoted the development of an ontology oriented to the study of inflammaging being unified to Ontocancro. In the context of this work, ontology has the purpose to map the existing knowledge in the genetic framework of chronic inflammatory state elapsed by aging, such as Alzheimer's disease, Type 2 diabetes, COPD or pulmonary disorder Chronic Obstructive the Thyroid Cancer of Pancreas of Colon and Rectal Adrenocortical glands, as well as samples of each disease, metabolic pathways and genes present in this process. All data were defined based on current research literature. The dissertation presents results in a comparative analysis between the tracks of the natural process of cell development with inflammation pathways inserted after this study, and you can see that both are intertwined from the genetic interaction. To visually verify these connections used the String, a genetic link processor that provides graphs relationship between genes, showing the intensity ratio within a bonding means or more lanes. Thus, there was a comparative study of two routes; detected genes of intersection between them have intensive connections to a degree of 90% reliability.

[1] Programa de Pós-Graduação em Informática – Universidade Federal de Santa Maria

Alternative Splicing Analysis in Rice Under Iron Overload

Artur Teixeira de Araujo Junior[1]; Marcelo Nogueira do Amaral[2]; Luis Willian Pacheco Arge[2]; Daniel da Rosa Farias[2]; Railson Schreinert dos Santos[2]; Danyela de Cássia Oliveira[2]; Solange Ferreira da Silveira Silveira[2]; Eugenia Jacira Bolacel Braga[2]; Luciano Carlos da Maia[2]; Antonio Costa de Oliveira[3]

Introduction

Rice (Oryza sativa L.) is an important crop, being the second most produced cereal worldwide. In Brazil, rice is mainly produced in ecosystems called "terras baixas" (lowland). These conditions of waterlogging cause a large problem: iron overload. This problem culminates with large losses in crop production [1], possibly leading to a reduction of up to 100% in yield, depending on the concentration of reduced iron in soil solution and on the tolerance of rice cultivars [2]. Recent studies report that environmental stresses can affect the mechanism of alternative splicing, causing changes in the expression of different genes and affecting especially those involved in post-transcriptional modifications. Due the little information on this subject, this study aimed to evaluate and quantify the different junction sites and alternative splicing events in the transcriptome of rice cultivar BRS Querência under normal and high iron concentrations.

Materials and Methods

Seeds of rice (cv. BRS Querência) were germinated in a growth chamber for 7 days with a photoperiod of 16 hours and a temperature of 25 ± 2 °C. After this period, seedlings were transferred to plastic trays containing pre-washed sand and kept in a greenhouse with alternate irrigation (every 2 days) with nutrient solution [3] and water. Plants were subjected to iron overload at the seedling stage, adding 300 mg L^{-1} Fe^{+2} to the nutrient solution, for 24 hours. Untreated plants remained in normal nutrient solution for the same time, to be used as control. The total RNA was extracted from 100 mg of tissue (leaves) using of Plant RNA Reagent Purelink®. For the preparation of libraries was used TruSeq RNA Sample Preparation v2 (Illumina®) kit, following manufacturer's recommendations. RNA-Seq analysis was performed using HiSeq 2500 platform (Illumina®) with paired-end 2 x 100 reads (two reads of 100 bp). In order to evaluate read quality, the software FastQC Ver. 0.11.2 was used. After that, the software Trimmomatic Ver. 0.32 [4] was used toremove the adapters and low quality base reads of each libray. Then the reads were mapped using O. sativa Nipponbare (IRGSP build 4.0) as reference in TopHat Ver. 2.0.11 [5]. After alignment, the software MapSplice [6] and SpliceGrapher [7] were used to analyze of the junction sites and alternative splicing events, respectively. Predicted Rice Interactome Network - PRIN [8] were used to demonstrate probable protein-protein interactions.

[1] Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900 (arturtaj@hotmail.com)

[2] Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900

[3] Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900 (acostol@cgfufpel.org)

Results and Discussion

A total of 127,781 alternative splicing junctions sites were identified in plants under normal conditions (control), while 123,682 were found in stressed plants. When these results are compared with the literature, a significant difference in the proportion of these junction sites into the three main categories described is found (Figure 1A). These results indicate that there is no big change in splicing processes due to iron overload stress. As expected most common sites are shown to be the canonical, involving 98,85% and 98,91% in control and stress condition respectively, consistently with other results in the literature [9, 10]. Also that we have the same-canonical (0,73% in control and 0,70% in stressed) and non-canonical junctions sites (0,42% in control and 0,40% in stressed). A deeper alternative splicing analysis showed that the control plants presented a higher number of events (23,307) when compared to the treated ones (8,244). Furthermore it is interesting to note that intron retention was the most frequent event in both situations, with 10,281 (44.1%) for control and 3,909 (47.4%) for stressed (Figure 1B and C). Similar results were also detected in others studies, not just for rice but also for other plants species [10, 11]. The second major event is the alternative 3' splice site (5,261/22.6% and 1,802/21.9%), followed by exon skipping (4,413/18.9% and 1,429/17.3%) and the alternative 5' splice site (3,352/14.4% and 1,104/13.4%). Some studies demonstrate that the intron retention can even form truncated proteins and these proteins can make important functions in the adaptation of plants to stress conditions [12].

The alternative splicing influences are present in almost all aspects of protein functions, making it a central element of gene expression regulation. In this case, changes promoted by alternative splicing in the expression of proteins related to post-translational modifications were analyzed. These proteins are responsible for diversifying the functions of others proteins and for the dynamical coordination of signaling networks [13]. Here we highlight that, 25 differentially expressed proteins were found related to post-translational modifications, and from these, 20 were downregulated, while 5 were upregulated. When these proteins were analyzed, kinases and phosphatase were found to be the most common classes, where 14 assignments were found. These proteins are responsible for two important biological processes: phosphorylation with 14.29% assignments (3 downregulated) and dephosphorylation with 85.71% assignments (14 downregulated and 4 upregulated), and the most common molecular functions that are transferase activity and ATP binding both with 16 downregulated and 4 upregulated assignments. PRIN presented data about interactions with other differentially expressed proteins for just one locus (LOC_Os02g34600) in this experiment. Results demonstrated that alterations in the expression of this protein correlates with other 6 (Figure 1D), having 5 negative correlations with co-expression between -0.1879 and -0.021, and 1 positive correlation (0.1756). As demonstrated in Figure 1E the metabolomic correlation network, indicates protein-folding metabolism which presented negative co-expression with the post-translational modifications, and post-translational modifications metabolism, presenting positive co-expression with the RNA processing and negative co-expression with ribosomal protein synthesis and another non-assigned metabolism.

From these results one can conclude that BRS Querência demonstrated change in alternative splicing and the post-translational modification proteins promoted by the iron stress. This initial data is important for the understanding on how plant performs their responses to stress and how this response may be related to the t olerance of this cultivar. The identification of transcripts can aid its use in genetic engineering and marker development.

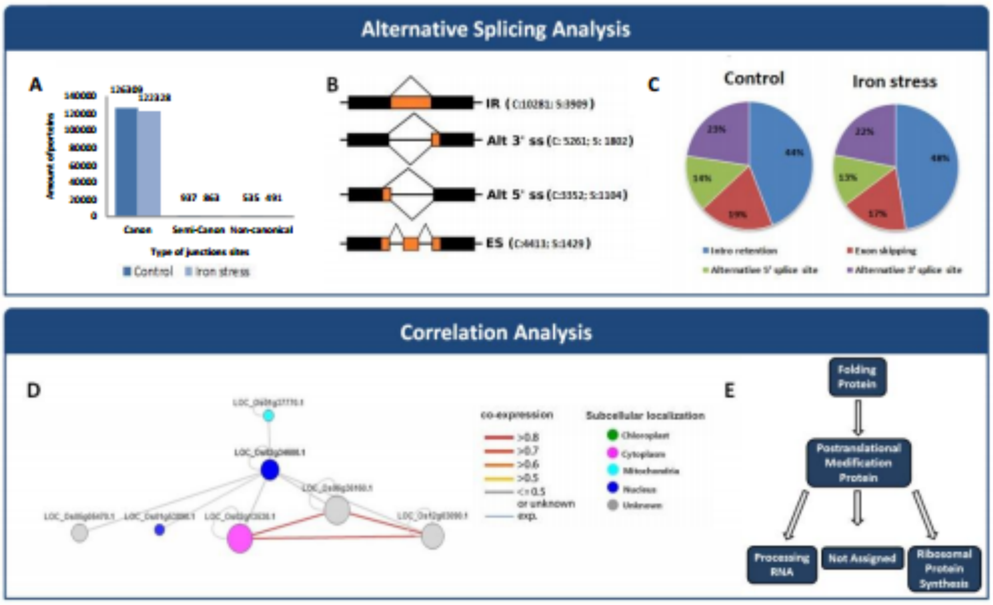


Figure 1 Alternative splicing analysis. (A) Types of alternative splicing junctions sites.(B) Schematic representation of the most frequent splicing event identified in the v2 prediction: intron retention (IR), alternative 3' splicing site (Alt 3' ss), alternative 5' splicing site (Alt 5'), exon skipping (ES). The number of events is reported between brackets in control (C) and stress condition (S). (C) Pie chart showing the percentage distribution of alternative splicing events. (D) Protein-protein interactions network. (E) Correlation metabolomics network.

References

[1] VAHL, L.C., Iron toxicity in rice genotypes irrigated by flooding. Ph.D. Thesis. Federal University of Rio grande do Sul, Porto Alegre,1991.

[2] SAHRAWAT, K. L. Iron toxicity in wetland rice and the role of other nutrients. Journal of Plant Nutrition, v. 27, p. 1471-1504, 2004.

[3] YOSHIDA, S.; FORNO, D.A.; COCK, J.H.; GOMEZ, K.A. Laboratory manual for physiological studies of rice. The Philippines: International Rice Research Institute(IRRI), 1972. 3v.

[4] BOLGER, A.M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, Reino Unido, v.30, n.15, p.2114 -2120, 2014.

[5] TRAPNELL, C.; PACHTER, L.; SALZBERG, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, Reino Unido, v. 25, n. 9, p. 1105-1111, 2009.

[6] WANG, K.; SINGH, D.; ZENG, Z.; COLEMAN, S.J.; HUANG, Y.; SAVICH, G.L.; HE,X.; MIECZKOWSKI, P.; GRIMM, S.A.; PEROU, C.M; MACLEOD, J.N.; CHIANG, D.Y.;

PRINS, J.F.; LIU, J. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research, Reino Unido, v.38, n.18, p.1-14, 2010.

[7] ROGERS, M.F.; THOMAS, J.; REDDY, A.S.N.; BEN-HUR, A. SpliceGrapher: Detecting patterns of alternative splicing from RNA-seq data in the context of gene models and EST data. Genome Biology, Reino Unido, v. 13, n.4, p.1-17, 2012.

[8] GU H.B., ZHU P.C.; CHEN M. PRIN: a pedicted rice interactome network. BMC Bioinformatics, 2011.

[9] AMARAL, M.N.; ARGE, L.W.P; BENITEZ, L.C.; MORAES, G.P.; MAIA, L.C.; BRAGA, E.J.B. Eventos de splicing alternativo associados ao estresse salino em plantas de arroz. ENPOS, 2014.

[10] VITULO, N.; FORCATO, C.; CARPINELLI, E.C.; TELATIN, A.; CAMPAGNA, D; D'ANGELO, M.; ZIMBELLO, R.; CORSO, M.; VANNOZZI, A.; BONGHI, C.; LUCCHIN, M; VALLE, G. A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biology, Reino Unido, v.14, n.99, p.1-16, 2014.

[11] LU, T., LU, G., FAN, D., ZHU, C., LI, W., ZHAO, Q., FENG, Q., ZHAO, Y., GUO, Y., LI, W. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res. 20, 1238–1249, 2010.

[12] MATSUKURA, S.; MIZOI, J.; YOSHIDA, T.; TODAKA, D.; ITO, Y.; MARUYAMA, K.; SHINOZAKI K.; YAMAGUCHI-SHINOZAKI K. Comprehensive analysis of rice DREB2 type genes that encode transcription factors involved in the expression of abiotic stress-responsive genes. Molecular Genetics and Genomics, Alemanha, v.283, n.2, p.185-196, 2010.

[13] WANG, Y.C.; PETERSON, S.E.; LORING. J.F. Protein post-translational modifications and regulation of pluripotency in human stem cells. Cell Research, 2014, 24:143-160.

Workflow for Illumina Amplicon Analysis of Soil Communities under Pollutant Stress

Daiana Lima-Morales[1,2], Ruy Jáuregui[1,3], Amélia Camarinha-Silva[1,4], Ramiro Vilchez-Vargas[1,5], Dietmar H. Pieper[1]

There is still a general lack of knowledge how microbial communities react under pollutant stress, further more it is known that microorganisms enriched in the laboratory often don't play an important role in in-situ biodegradation. Disregarding the need of cultivation, molecular techniques allow studying the microbial diversity in-situ. To analyze which microorganisms are important in pollutant soils, we upgraded a high throughput method using the Illumina platform, to amplify the V5/V6 region of the 16S rRNA gene. In total 2.003.786 reads were obtained. Only reads of a minimum of 115 nt in length (29 nt of primer/barcode and 86 nt of 16S rRNA gene sequence) were considered. Truncated reads that had an N character, mismatches within primers and barcodes or more than 8 homopolymer stretches were discarded. All sequences, totaling 1.671.472 reads, were split into 63 files according to their unique barcode and collapsed into unique representative reads. Using Mothur, these reads were pre-clustered into 24,000 unique representative reads and afterwards filtered. A representative reads was kept if: a) it was present in at least one sample in a relative abundance >0,1% of the total sequences of that sample or b) it was present in at least 3 samples or c) it was present in a copy number of at least 10 in at least one sample. Phylotypes were then generated by clustering at 98% similarity using USEARCH and assigned to a taxonomic affiliation based on RDP classifier . Applying this workflow we could reduce the number of representative reads to 501 phylotypes, a computational manageable number, without compromising the fine scale soil community composition. Out of 501 phylotypes, 470 could be identified to phylum level, 432 to class level, 361 to order level and 312 to family level.

Financial Support: CNPq-Conselho Nacional de Desenvolvimento Científico e Tecnológico-Brazil and BACSIN from the European Commission.

[1] Microbial Interactions and Processes Research Group (MINP), Department of Medical Microbiology, Helmholtz Centre for Infection Research, Braunschweig, Germany

Current addresses: Laboratório de Pesquisa em Resistência Bacteriana (LABRESIS), Centro de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, Brazil; AgResearch, Palmerston North, New Zealand; Institute of Animal Nutrition, University of Hohenheim, Stuttgart, Germany; Laboratory of Microbial Ecology and Technology (LabMET), University of Ghent, Ghent, Belgium

ExOmin: Boosting Variant Priorization Identified by Whole-Exome Sequencing

Pedro B. Marin[1]'Delva Leão[2], Junior F. Barros[1], Julio C. S. Anjos[1],Filippo P. Vairo[2], Jonas A. M. Saute[2], Claudio R. Geyer[1],Ursula Matte[2]

The Whole-Exome Sequencing (WES) is one of the main applications of Next-Generation Sequencing increasingly used into medicine to elucidate the molecular basis of mendelian diseases without defined etiology by tradicional methodologies. The strategy to prioritize the causal genetic variant among thousands found on a WES analysis as well as the connection between phenotype and genotype specifically responsible by the condition investigated are crucial and laborious tasks. Such complexity has stimulated the developing of methodologies to assist this process. The majority of available prioritization tools are paid, overly complex or do not meet all aspects needed to perform an integrated analysis. Seeking to attend a real demand of medical geneticists of Clinical Hospital of Porto Alegre (HCPA), that face in prioritization step a relevant obstacle to define the diagnosis of patients attended at Genetic Medical Service (SGM), we propose the developing of ExOmin tool. This tool will comprise different modules and will use a strategy designed to combine variant annotation, variant prioritization, phenotype/genotype relation and selection of causal variant among probable variants. The main modules will be: 1)annotation of .VCF file with the open-source software ANNOVAR; 2) prioritization of variants using heuristic ilters; 3) integration of phenotype/genotype through OMIM database; 4) sorting of probable variants using conservation and pathogenicity predictors. ExOmim will be implemented as a workflow, where a .VCF file processed in different modules will result in fewer candidates variants to be defined as responsible for observed clinic features. It's expected that these modules can be used independently as required by the investigator. Once implemented, we hope that ExOmin boost the process of WES analysis at SGM/HCPA and that patients without defined diagnosis attended by this facility have their diagnoses with greater accuracy and in less time.

[1] Instituto de Informática, UFRGS, Caixa Postal 15064
{Pedro B. Marin, pedrobmarin@gmail.com; Junior F. Barros, jfbarros@inf.ufrgs.br;
Julio C. S. Anjos, julio.c.s.anjos@gmail.com; Claudio R. Geyer, geyer@inf.ufrgs.br}
[2] Serviço de Genética Médica, Hospital de Clínicas de Porto Alegre, Caixa Postal 15039
{Delva Leão, dleao@hcpa.edu.br; Fillipo P. Vairo, fvairo@hcpa.edu.br; Jonas A. M. Saute, jsaute@hcpa.edu.br; Ursula Matte, umatte@hcpa.edu.br}

Bioinformatics applied to QTLs related to absorption and accumulation of arsenic in rice

Railson Schreinert dos Santos[1]
Eduardo Venske[1]
Artur Teixeira de Araújo Júnior[1]
Daniel da Rosa Farias[1]
Antônio Costa de Oliveira[1]

Introduction

Rice (Oryza sativa L.) is a staple food crop for half the world's population. This cereal is an important source of carbohydrates and other nutrients to the population and ha s a high socioeconomic importance to the producing countries. Due to its high consumption, the presence of toxic elements such as arsenic in its grains can be of concern. Studies have shown that rice is a major source of arsenic to humans through diet [1, 2], with a range of genotoxic effects [3, 4]. Countries such as Europe, United States [5, 6] and even Brazil [7], can have dangerous arsenic levels in rice grains. Arsenic is considered the most toxic xenobiotic and a carcinogen of the highest class of non-metals. Therefore, there is no minimum content that may be considered safe or harmless [8]. Breeding efforts to lower arsenic content in rice grains seem to be promising, since different studies show significant genetic variability for both arsenic accumulation and speciation. QTL (Quantitative Trait Locus), physiological processes, and genes responsible for these answers have also been studied [3, 9, 10]. Here we highlight one important QTL related to the change in the concentration of arsenic in shoots of seedlings. This QTL, generically called AsS, is flanked by RM318 and RM450 markers on rice chromosome 2, and accounts for 24.4% of the expected phenotypic variation [9]. The present investigation aimed to evaluate the use of some bioinformatics tools to analyze AsS, with the purpose of identifying genes related to the change in the concentration of arsenic in rice seedlings.

Materials and Methods

In order to identify the genes within the AsS locus (chr2:28634200..29637600) the genomic sequences of the MSU Rice Genome Annotation Project [11] has been used. The 145 identified genes were subjected to a differential analysis in Genevestigator [12] using the database obtained from a study of transcriptional expression in roots of cultivars with contrasting phenotypes for tolerance to arsenate [13].

[1]Plant Genomics and Breeding Center, UFPEL, PO Box 354

{railsons.faem@ufpel.edu.br; eduardo.venske@yahoo.com.br; arturtaj@hotmail.com; fariasdr@gmail.com; acostol@cgfufpel.org}

The changes of expression were evaluated using a p value ≤ 0.05 to select genes differentially expressed when under stress. The genes that showed to change its expression under stress had their similarity of expression evaluated using hierarchical clustering in the same experiment [13]. These genes were then evaluated for transcriptional expression in different organs, using the databases of Kudo et al., 2013 [14] and Norton et al., 2008 [15].

Results and Discussion

Analysis of the chromosomal region of AsS demonstrated the presence of 145 genes in this locus (data not shown). Genes that had transcriptional expression data available in the experiment of Norton et al., 2008 [13] were evaluated, and 3 loci were considered differentially expressed when plants were under 1 ppm of arsenate. The results of differential expression analysis are shown in Figure 1.

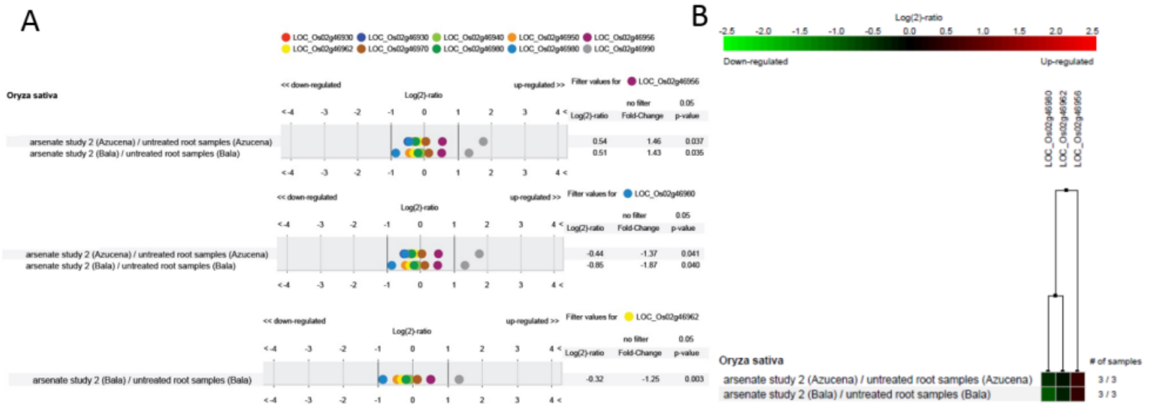


Figure 1. Differential expression of genes at the locus AsS. (A) Differential expression in plants under arsenate stress (1 ppm); (B) Hierarchical clustering analysis.

An increase in expression of LOC_Os02g46956 (chr2:28669044..28672392) and repression of LOC_Os02g46980 (chr2:28683849..28681681) and LOC_Os02g46962 (chr2:28673002..28676872) were observed. It is important to highlight that the last locus was only repressed in Bala, a tolerant genotype. To get an idea of the expression of these genes in different organs another analysis has been performed, generating the graph of Figure 2.

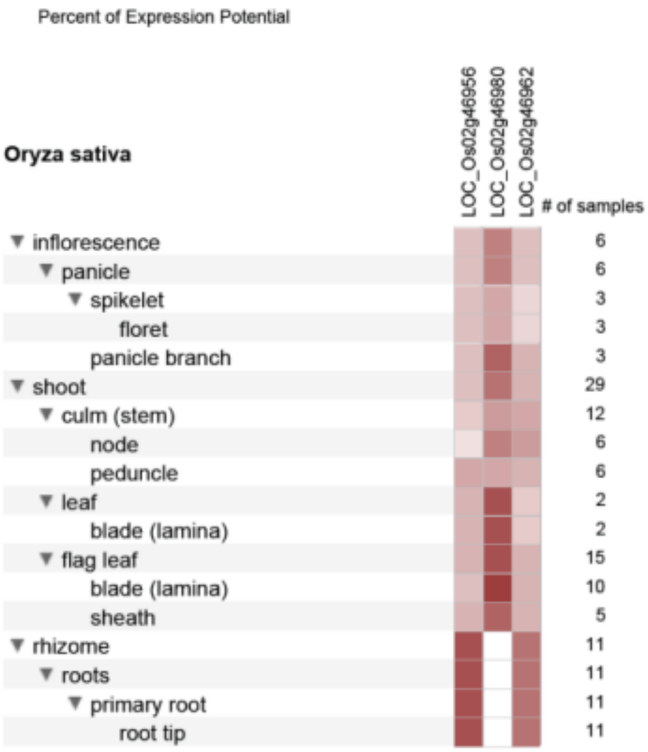


Figure 2. Expression of the three differentially expressed genes along rice anatomy.

The analysis of Figure 2 shows high expression potential of LOC_Os02g46956 and LOC_Os02g46962, another indicative of the probable importance in response to arsenate stress. Further studies involving characterization of these genes through structure prediction and modifications of these with molecular biology tools are yet to be conducted. The search for mutants and characterization these, are other important steps in elucidating the role of these genes and aid plant breeding. The results presented here show great utility of bioinformatics in the analysis of QTL related to absorption and accumulation of arsenic in rice plants. Results obtained from the analysis this and of other QTLs, also carried out by our laboratory (unpublished data), should help the understanding of the dynamics of arsenic in rice plants and their breeding to obtain safer crops and reducing the risk of cancer due to the intake of arsenic in food.

References

[1] Heinkens A. Arsenic Contamination of Irrigation Water, Soil and Crops in Bangladesh: Risk Implications for Sustainable Agriculture and Food Safety in Asia. Bangkok: FAO, 2006.

[2] Meharg AA, Williams PN, Adomako E, Lawgali YY, Deacon C, Villada A, et al. Geographical variation in total and inorganic arsenic content of polished (white) rice. Environmental science & technology, vol. 43, no. 5, pp. 1612-7. 2009.

[3] Meharg AA and Zhao FJ. Arsenic and rice. Netherlands: Springer Netherlands; vol. 172, pp. 2012.

[4] Banerjee M, Banerjee N, Bhattacharjee P, Mondal D, Lythgoe PR, Martínez M, et al. High arsenic in rice is associated with elevated genotoxic effects in humans. Sci Rep, vol. 3, 2013.

[5] Williams PN, Price AH, Raab A, Hossain SA, Feldmann J and Meharg AA. Variation in arsenic speciation and concentration in paddy rice related to dietary exposure. Environmental science & technology, vol. 39, no. 15, pp. 5531-40, 2005.

[6] Zavala YJ, Duxbury JM. Arsenic in rice: I. Estimating normal levels of total arsenic in rice grain. Environmental science & technology, vol. 42, no. 10, pp. 3856-60, 2008.

[7] Batista BL, Souza JM, De Souza SS and Barbosa F, Jr. Speciation of arsenic in rice and estimation of daily intake of different arsenic species by Brazilians through rice consumption. Journal of hazardous materials, vol. 191, n. 1-3, pp. 342-8. 2011.

[8] Smith AH, Lopipero PA, Bates MN and Steinmaus CM. Public health. Arsenic epidemiology and drinking water standards. Science, 296, no. 5576, pp. 2145-6. 2002.

[9] Zhang J, Zhu YG, Zeng DL, Cheng WD, Qian Q and Duan GL. Mapping quantitative trait loci associated with arsenic accumulation in rice (Oryza sativa). The New phytologist, vol. 177, no. 2, pp. 350-5. 2008.

[10] Norton GJ, Pinson SR, Alexander J, McKay S, Hansen H, Duan GL, et al. Variation in grain arsenic assessed in a diverse panel of rice (Oryza sativa) grown in multiple sites. The New phytologist, vol. 193, no. 3, pp. 650-64. 2012.

[11] Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Research,vol. 35, pp. D883-D887, 2007.

[12] Zimmermann P, Hennig L and Gruissem W. Gene-expression analysis and network discovery using Genevestigator. Trends in plant science. vol. 10, n. 9, p. 407-9. 2005.

[13] Norton GJ, Nigar M, Williams PN, Dasgupta T, Meharg AA, Price AH. Rice–arsenate interactions in hydroponics: a three-gene model for tolerance. Journal of Experimental Botany, vol. 59, no. 8, pp. 2277-2284. 2008.

[14] Kudo T, Akiyama K, Kojima M, Makita N, Sakurai T and Sakakibara H. UniVIO: a multiple omics database with hormonome and transcriptome data from rice. Plant & cell physiology, vol. 54, no. 2, pp. e9. 2013.

[15] Norton GJ, Aitkenhead MJ, Khowaja FS, Whalley WR and Price AH. A bioinformatic and transcriptomic approach to identifying positional candidate genes without finemapping: an example using rice root-growth QTLs. Genomics, vol. 92, no. 5, pp. 344-52.

2008.

Functional analysis of microRNA and transcription factor co-regulatory network in pathological cardiac hypertrophy

Mariana R. Mendoza[1], Adriano V. Werhli[2], Graziela H. Pinto[1,3],Daiane Silvello[1,3], Carolina R. Cohen[1,4], Nadine O. Clausell[1,3], Luis E. P. Rohde[1,3], Andréia Biolo[1,3]

Introduction

Pathological cardiac hypertrophy is the increase in heart mass that occurs as response to pressure or volume overload in disease settings, or to myocardial infarction and cardiomyopathies. Although compensatory at first stage, sustained cardiac hypertrophy is detrimental and increases the risk of heart failure. Several well characterized structural, functional and molecular changes underlie this process, including cell death, fibrosis, cardiac dysfunction and altered gene expression. Recently, microRNAs (miRNAs) emerged as important post-transcriptional regulators of gene expression and have been implicated in cardiac development and several cardiovascular disorders [3]. Although their involvement with cardiac hypertrophy has been discussed in literature [2], their specific regulatory mechanisms, as well as the patterns of their cooperation in the synergistic regulatory network with transcription factors (TFs) have rarely been studied in this scenario. This study aims at constructing and analyzing the miRNA-TF co-regulatory network involved in cardiac hypertrophy using a bioinformatics approach, and help decipher major regulators contributing to this phenotype and their respective regulatory mechanisms and biological functions.

mRNA and miRNA expression profiles

Large-scale expression profiling of mRNAs and miRNAs related to an in vitro model of cardiac hypertrophy were downloaded from Gene Expression Omnibus (GEO) database (accession numbers GSE60291 and GSE60292). The study was carried out to identify RNAs in human cardiomyocytes that show differential expression upon induction of hypertrophy, assessing expression levels of miRNAs and mRNAs in triplicates for controls and endothelin 1 (ET-1) stimulated human cardiomyocytes [1].

[1]Laboratório de Pesquisa Cardiovascular, Centro de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre, Porto Alegre, RS, Brazil. E-mail address of corresponding author (MRM): {mrmendoza@inf.ufrgs.br}

[2]Centro de Ciências Computacionais, Universidade Federal do Rio Grande (FURG), Rio Grande, RS, Brazil

[3]PPG em Ciências da Saúde: Cardiologia e Ciências Cardiovasculares, Universidade Federal do Rio Grande do Sul,Porto Alegre, RS, Brazil

[4]PPG em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

Differential expression and functional enrichment analysis

To collect genes and miRNAs involved in the physiopathology of cardiac hypertrophy, we performed differential expression analysis based on the large-scale expression profiles from GEO. Original mRNA expression profile (GSE60291) was analyzed with affy and limma R packages, adopting background correction and quantile normalization by Robust Multi-array Average algorithm. Genes were considered differentially expressed between both groups if presenting absolute log2 fold changes greater than 1.5 and a false discovery rate (FDR) < 0.05. Differential expression analysis of miRNAs (GSE60292) was carried out with DEQSeq R package. MicroRNAs with minimum up- or down-regulation ratio of 1.5 (0.58 in a log scale) and FDR < 0.05 were considered significantly different between both groups. Following this analysis, we performed functional enrichment of differentially expressed genes (DEGs) using Gene Set Enrichment Analysis (GSEA) to extract biological insights such as the main molecular mechanisms deregulated upon induction of hypertrophy.

Construction and analysis of miRNA-TF co-regulatory network

A miRNA-TF co-regulatory network was constructed using a compilation of experimentally verified and confidentially predicted miRNA-gene (including TF genes), TF-miRNA and TF-gene interactions. Interactome data were retrieved from the following sources: i) ChipBase, HTRIdb and ORegAnno for validated TF-gene interactions; ii) miRTarBase and starBase for validated miRNA-gene interactions; iii) TargetScan for predicted miRNA-gene interactions; iv) ChipBase for validated TF-miRNA interactions. Only interactions in which both regulator and target were found to be differentially expressed in cardiac hypertrophy were considered for network construction. Once the network was generated, its structural properties, such as node degree and betweenness, were analyzed to investigate central genes contributing to network stability and communication. We defined hub nodes as the top 5% highest-degree nodes and bottleneck nodes as nodes with betweenness value standing one standard deviation above the mean. Finally, we analyzed the patterns of the interplay between miRNAs and TFs in the synergistic regulatory network by identifying significant network motifs using the FANMOD tool [4].

Differentially expressed genes and miRNAs

Based on GEO expression profiling datasets, we identified 496 differentially expressed genes (DEGs), including 24 TFs, and 273 differentially expressed miRNAs (DEMs) between both groups. A total of 201 genes, 16 TFs and 143 miRNAs were up-regulated upon induction of cardiac hypertrophy, while a total of 273 genes, 6 TFs and 130 miRNAs were down-regulated. Among the most deregulated genes and miRNAS, we identified significant overexpression of NPPB, ACTA1, CCL4, SERPINE1, SOX9, miR-1322, miR-21, miR-221 and miR-208a. According to GSEA, up-regulated genes are significantly enriched in regulation of MAPK activity, inflammatory and immune response, cytokine production and dephosphorylation, while down-regulated genes are mainly involved in distinct cell cycle phases, including DNA replication, mitosis and sister chromatid segregation.

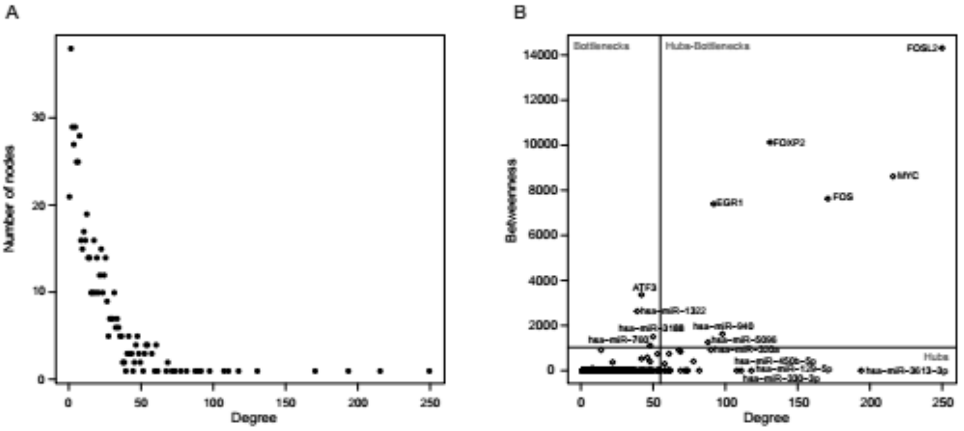


Figure 1. A) Degree distribution of the constructed miRNA-TF co-regulatory network. B) Degree vs. betweenness plot, defining hubs, bottlenecks and hub-bottlenecks nodes.

The cardiac hypertrophy miRNA-TF co-regulatory network

The miRNA-TF co-regulatory network included 6353 interactions among 617 nodes, recruiting 415 genes, 24 TFs and 178 miRNAs from those differentially expressed. Results of the node degree distribution reflected a scale-free topology, in which most nodes had low degrees and only a few nodes had high degrees (Fig. 1-A). Therefore, hub nodes might play major roles in network stability. Analysis of centrality measures identified several high degree nodes that may represent important network regulators, such as TFs MYC, FOSL2, FOS and EGR1, and miRNAs miR-3613-3p, miR-129-5p, miR-320a and miR-330-3p. Among the central nodes, we found that several hubs are also network bottlenecks, thus playing essential role in the control of information flow (Fig. 1-B). Network motif analysis identified significantly overrepresented patterns of interactions between TFs and miRNAs in the co-regulatory network (Fig. 2). A total of 18 motif types with size 3 were found englobing miRNA-TF synergistic regulation, among which we identified miRNA-mediated feedforward loop (FFL) regulation (Fig. 2-A), miRNA-mediated cascade regulation (B), TF-mediated FFL regulation (C), co-regulation (D) and miRNA simultaneous regulation (E). P-values were computed based on 1000 random simulations, comparing the frequency of motifs between randomly generated networks and the constructed network. Further analysis of these motifs and their constituents are needed to investigate common associations between miRNA and TF regulation.

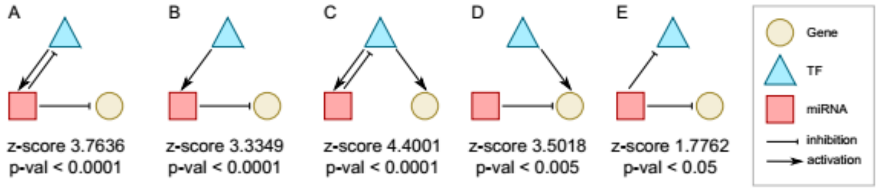


Figure 2. Over-represented motifs in the constructed miRNA-TF co-regulatory network.

Discussion

In this work we constructed a miRNA-TF co-regulatory network involved in the development of pathological cardiac hypertrophy, which allows us to analyse major regulators and regulatory patterns contributing to this mechanism and better understand its physiopathology. We identified important network regulators, such as TFs MYC, FOSL2, EGR1 and miRNAs miR-3613-3p, miR-129-5p, miR-330-3p, as well as significant miRNA-TF synergistic regulatory motifs that may be relevant for a better understanding of the molecular changes underlying cardiac hypertrophy. Further analysis of these data and a deeper investigation of network motifs and structural properties may help reveal common and essential patterns of interplay between miRNAs and TFs, as well as elucidate the regulatory role of miRNAs in pathological cardiac hypertrophy.

References

[1] P. Aggarwal, A. Turner, A. Matter, and et al. RNA expression profiling of human iPSCderived cardiomyocytes in a cardiac hypertrophy model. PLoS ONE, 9(9):e108051, 09 2014.

[2] P. A. Da Costa Martins and L. J. De Windt. MicroRNAs in control of cardiac hypertrophy. Cardiovascular Res., 93(4):563–572, 2012.

[3] V. Divakaran and D. L. Mann. The emerging role of microRNAs in cardiac remodeling and heart failure. Circ Res., 103(10):1072–1083, 2008.

[4] S. Wernicke and F. Rasche. Fanmod: a tool for fast network motif detection. Bioinformatics, 22(9):1152–1153, 2006.

In Silico Analysis of Gene Expression Profiling in Genes Linked to Autophagy and Pluripotency in Microarray Databases of IPSC, ESCS and Somatic Cells

PEREIRA, MB[1]
DALBERTO, TP[1]
LENZ, G[1]

Induced pluripotent stem cells (iPSC) have been widely explored as therapeutic tools in regenerative medicine and drug screening. The efficiency of the cellular reprogramming technique for the iPSC generation and of the cell differentiation process seem to be related to changes in energetic metabolism of these cells. Autophagy, in turn, appears to be important to the success of reprogramming. The identification of specific metabolic pathways and factors that regulate the destination of stem cells is important to the efficiency of reprogramming and to control the differentiation and destination of iPSCs. In this study, the in silico analysis of the correlation between groups of pluripotency genes and autophagy genes was performed using 12 microarray databases including ESCs, iPSCs and somatic cells. Samples from each database were grouped into clusters and the analysis of the correlation between groups of pluripotency genes and autophagy genes was done for the different cell types. Our findings indicate that pluripotency genes, such as Sox2, c-Myc and Lin28, are correlated with important genes involved in the autophagy induction in ESCs and iPSCs, suggesting the interaction between these two pathways and contributing to the better understanding of the relationship between them and their respective signaling pathways.

[1]Laboratório de Sinalização e Plasticidade Celular, Instituto de Biociências, UFRGS, Caixa Postal 9999 {Mariana Brutschin Pereira, mbrutschin@gmail.com} {Tiago Pires Dalberto, tiagodalberto@hotmail.com} {Guido Lenz, gulenz@gmail.com}

Multi-objective optimization in Bioinformatics

M. Villalobos Cid[1]

M. Inostroza-Ponta[1]

Many problems in bioinformatics and computational biology can be formulated as optimization problems: motifs discovery, expression data analysis, sequence alignment problems, gene regulation networks, protein structure prediction and recognition, single nucleotide polymorphism problem (SNP), drugs design, directed evolution analysis and optimization of biochemical processes (Biochemical informatics), among others. In general, single objective optimization problems are solved using algorithms based on techniques like linear and nonlinear programming, mixed integer programing, bilevel and dynamic optimization. In bioinformatics, most of the optimization problems belong to the NP-hard class of problems, i.e., computationally intractable (it means that there is no polynomial algorithm to find a solution in a reasonable time). An approach to deal with these problems is the use of metaheuristics algorithms. Researchers have used single objective optimization based models to deal with biological problems applying a wide number of metaheuristic techniques: simulated annealing, variable neighborhood search, particle swarm optimization, memetic algorithms, tabu search, evolutionary algorithms and other bioinspired techniques. However, the globally optimal solution is not guaranteed due to local minimum stagnation, lack of gradient and data noise influence. Biological optimization problems generally involve numerous objectives, which may consider different aspects of the solutions, which can be incommensurable and often, partially or wholly in conflict. Using multiobjective optimization formulation is possible to model these problems, decompose one objective function into more functions to reduce the probability of local minimum stagnation, add goals to solve problems with gradient and reduce noise using secondary bias function. However, finding a set of Pareto-optimal solutions and chosing only one solution using a decision maker increases the computational cost. We show a comparison between single and multi objective optimization based models in sequence alignment and protein structure prediction. The goal is to highlight the advantages on using a multiobjective approach despite the extra computational effort needed.

[1]Departamento de Ingeniería Informática, Universidad de Santiago de Chile, USACH

{manuel.villalobos, mario.inostroza@usach.cl}

A medical bioinformatics approach for metabolic inherited disorders associated with cancer: A case study on D2/L2-Hydroxyglutaric Aciduria related to Glioma with IDH1/2 deficiency

Vallejo-Ardila Dora L.[1,2], Ida Vanessa D. Schwartz[1,2], Fernanda SperbLudwig[1,2]

After Otto Warburg's work "On the Origin of Cancer Cells" was published in 1956, other studies of cancer metabolism have been reinforcing the concept that aberrant energetic metabolism occurs in cancer cells prior to the sequence of events that results in carcinogenesis. The availability of whole genome analyses data has facilitated the discovery of clinically relevant genetic alterations in cancer and inborn errors of metabolism, but still remains unclear what would be the potential role for those metabolic enzymes in cancer development. This fact provides us not only with the challenge to assess functional associations between an experimentally derived gene set of interest and a database of known gene sets, but also to integrate biological knowledgebase and analytic tools aiming at systematically exploiting biological meaning from large gene list. A case study on L2-hydroxyglutaric Aciduria related to Glioma with IDH1/2 deficiency is demonstrated by using network-based set enrichment analysis (EnrichNet) and modular enrichmen analysis (MEA). We applied a medical bioinformatics approach for comparing two types of methods of enrichment analysis to extract the biological insight into the pathogenesis of metabolic inherited disorders associated with cancer. Availability: EnrichNet: network-based gene set enrichment analysis is available at: http://www.enrichnet.org and The Database for Annotation, Visualization and Integrated Discovery (DAVID) is available at: http://david.abcc.ncifcrf.gov.

[1]Universidade Federal do Rio Grande do Sul (UFRGS)

[2]Hospital de Clínicas de Porto Alegre (HCPA)

A Knowledge-based Particle Swarm Optimization for the Protein Structure Prediction Problem

Mariel Barbachan e Siva[1]
Bruno Borguesan[2]
Márcio Dorn[2]

Proteins perform a wide range of functions in living organisms and consist of one or more chains of amino acids linked together by peptide bonds. This linear sequence is called the primary structure of the protein and tends to organize itself in regular conformations in space, these organization patterns are the secondary structure, the protein folding results in the tertiary structure which is closely related to the protein function. The three-dimensional protein structure prediction (PSP) is one of the most important problems of structural bioinformatics. Several computational strategies have been proposed for this problem, either using experimental information or not, the latter have a high computational cost associated. Metaheuristics are often used in attempts to solve the PSP problem due to their ability to find satisfactory solutions with lower computational effort than exact methods. Particle Swarm Optimization (PSO) is a metaheuristic of stochastic optimization inspired by the social behavior of animals. The algorithm works with a swarm of candidate solutions (particles) moving around the search space according to a fitness function. The movement of the particles is guided by their own best position in the search space and the best position of the swarm. This work is based on the fact that the native conformation of a protein corresponds to the one with the lowest potential energy. We developed a knowledge-based PSO for PSP problem. In our method, each particle is a candidate protein structure represented by a set of main and side chain torsion angles. Our knowledge-based swarm is generated by two different datasets. The first one have its structural information extracted from experimental protein templates by a clustering strategy combined with artificial neural networks and the second one from a database constructed by homology with protein fragments containing torsion angles information based on amino acid sequences and corresponding secondary structure. Preliminary results indicate that the approach using these datasets lead to better potential energy and RMSD results compared to swarms generated without this information.

Acknowledgements

This work was partially supported by grants from FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9).

[1] Instituto de Biociências, UFRGS
{mariel.barbachan@ufrgs.br}
[2] Instituto de Informática, UFRGS
{mdorn, bborguesan@inf.ufrgs.br}

Analysis of Protein Interaction in the Cancer Development

Luiz Henrique Rauber[1]
Cristhian AugustoBugs[2]
Éder Maiquel Simão[3]

The activation of Genome maintenance mechanisms (GMM) pathways such as: cell cycle (CC), DNA damage response (DDR) and apoptosis (APO) significantly contribute to tumor development. In previous studies, it was found in pre -cancerous activation process there is an anti-cancer barrier, which is responsible for prevention of tumor progression. The identification of the genes expressed during activation of the anti -cancer barrier, associated interactions in GMM pathways, becomes a complementarity to the study of the evolution of cancer. In this work, the objective was to investigate the activation of anti -cancer barrier in pre-cancer and cancer present in the adrenal gland tissue, colon, pancreas and thyroid follicles, using networks of interaction between proteins. To describe this barrier was proposed the modeling interaction networks between the proteins of MMG pathways using the Cytoscape software. The results obtained with the most prominent genes in expression and quantity of interactions were compared with the results of previous publications and reconfirmed the relevance of CDKN1A, CHEK1, ATR, TP53, MRE11A and XRCC4 genes. This analysis allowed the identification of other genes complementary to previous studies as SKP2, CCNO, FADD, RAD50, NBN, BIRC3, CDK2 and XRCC6 genes. These genes are associated and complement the studies on the activation of anti-cancer barrier. These considerations highlight that it is important to note the entire biological systemic and immersive context.

[1]Curso de Ciência da Computação, URI, CEP 97700-000
{Luiz Henrique Rauber, luiz.rauber@gmail.com}
[2]UNIPAMPA, CEP 97300-000
{Cristhian Augusto Bugs, cristhianbugs@gmail.com}
[3]Programa de Pós Graduação em Nanociências, UNIFRA, CEP 97010-491
{Éder Maiquel Simão , edersimao@gmail.com}

Analysis of Histones Deacetylases Expression and Post-transcriptional Regulation in Pancreatic Ductal Adenocarcinoma

Cleandra Gregório[1,2], Bárbara Alemar[1,2], Mariana R. Mendoza[3], Alessandro Osvaldt[4,5], Rudinei Correia[2], Raquel C. Rivero[5,6], Simone M. S. Machado[5], Patrícia Ashton-Prolla[1,2]

Pancreatic ductal adenocarcinoma (PDAC) is a highly lethal and aggressive disease. The disruption of histone acetylation through histones deacetylases (HDACs) and expression regulation by miRNAs can lead to tumor development. In this study we assessed HDAC1, HDAC2, HDAC3 and HDAC7 expression in PDAC and non-tumoral tissue samples using experimental and bioinformatics analysis, correlated their expression levels with clinicopathological features in patients and performed in silico investigation of regulation by miRNAs. Expression levels of HDAC1, 2, 3 and 7 were measured by qRT-PCR from 25 PDAC and 23 non-tumoral adjacent tissues and their association with clinicopathological parameters was analyzed. Differential expression (DE) and correlation analyses of HDACs and miRNAs in PDAC was performed using five GEO microarray datasets. Potential miRNA-HDACs relationships were collected from miRNA interaction databases. $P<0.05$ was considered statistically significant. We found reduced expression in PDAC compared with non-tumoral tissues for all HDACs analyzed, with $P<0.05$ for HDAC1, 2 and 3. However, fold-changes were very small and not biologically relevant. Strong positive correlation was observed between HDAC1 and HDAC3 ($P=0.003$), and moderate negative correlation between HDAC1 and HDAC7 ($P=0.017$) and HDAC3 and HDAC7 ($P=0.032$). None of HDACs expression was correlated with clinicopathological features. DE analysis suggested significant up-regulation of HDAC1, 2 and 7, and down-regulation of HDAC3, albeit all of them associated with small fold changes. 728 miRNAs (44 DE) were retrieved by bioinformatics analysis as HDACs regulators, and 125 predicted miRNA-HDAC pairs had negative expression correlation. Twenty miRNAs were DE and negatively correlated with their targets, thus representing promising regulatory mechanisms to be further investigated. Our results indicate that there may be a role of HDACs in the pathogenesis of this tumor, but differential expression between groups was subtle. Biologically relevant role for HDAC1, 2, 3 and 7 in pancreatic carcinogenesis is questionable.

Financial support: (Grant #559814/2009-7), CAPES and Fundo de Incentivo à Pesquisa-FIPE, Hospital de Clínicas de Porto Alegre (Grant 10-0162 and 11-0510)

[1] Programa de Pós-graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil. Corresponding author: {cleandra.gregorio@gmail.com}

[2] Laboratório de Medicina Genômica, Centro de Pesquisa Experimental - Hospital de Clínicas de Porto Alegre.

[3] Laboratório de Pesquisa Cardiovascular, Centro de Pesquisa Experimental - Hospital de Clínicas de Porto Alegre

[4] Grupo de Vias biliares e Pâncreas - Hospital de Clínicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil.

[5] Hospital de Clínicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil.

[6] Departamento de Patologia, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil.

Relational Database for Genetic Analysis

Timóteo Matthies Rico[1]
Andrea von Groll[1]
Karina dos Santos Machado[2]
Pedro Eduardo Almeida da Silva[1]

Biological data that are stored in public database are generally available in flat files. Among several types of this format, one of the most popular is the FASTA format. However, providing those data in flat files is often inadequate for complex queries. The softwares that analyse biological data in flat files format are limited when complex filters and crossinformation are needed. Therefore, this study aims the implementation of a relational database for effective organization and analysis of data related to genre, species, strains, genes, and nucleotide of any type of organism. Furthermore, we developed a software that insert biological data into the proposed database, based on multiple alignment FASTA files. The database was developed in the Database Management System PostgreSQL version 9.3 and the software was developed in Java platform. Sequences from Mycobacterium tuberculosis were downloaded in FASTA format and inserted into the database to test the database schema and the software's feature of data import. Results show the use of the developed database allows the extraction of different information as: gene relation per species/strains; distinct sequence per gene; which strains have the same sequence in determined gene. Besides, important queries as identifying conserved regions and the molecular markers were also developed. The database ensures data integrity, avoids redundancy, and ensures high performance, even in complex queries, using appropriately its features such as primary and foreign keys, unicity, and indexes in specific fields. As future work we propose the development of an interface to generate appropriate data mining input files using the stored data on the proposed database.

[1]Programa de Pós graduação em Ciências da Saúde, Universidade Federal do Rio Grande (FURG)
{timoteomr@gmail.com, avongrol@hotmail.com, pedrefurg@gmail.com}
[2]Centro de Ciências Computacionais, Universidade Federal do Rio Grande (FURG)
{karinaecomp@gmail.com}

Identification and Analysis of Transcription Factors in Rice Under Iron Overload

Artur Teixeira de Araujo Junior[1]; Daniel da Rosa Farias[2]; Railson Schreinert dos Santos[2]; Danyela de Cássia Oliveira da Silva[2]; Solange Ferreira da Silveira Silveira[2]; Camila Fernanda de Oliveira Junkes[3]; Luciano Carlos da Maia[2]; Antonio Costa de Oliveira[4]

1 Introduction

Rice (*Oryza sativa* L.) is the staple food for more than two thirds of the world´s population, and the second most widely grown cereal in the world. Irrigated rice is a crop with large importance in the state of Rio Grande do Sul (Brazil), accounting for two thirds of total production in the country. When rice is under conditions of iron excess, symptoms of toxicity are observed and may cause a decrease in productivity [1].

To respond and adapt to diverse abiotic stresses, as iron overload, many plant genes are induced or repressed, changing the levels of a variety of proteins. In this context transcription factors play an important role, since they do not only regulate growth and development, but also response to biotic and abiotic stresses [2]. These proteins have the capacity to recognize and bind to specific DNA sequences, regulating negatively or positively the transcription of a target gene [3]. Recently, several studies aiming to identify genes responsive to stress by iron toxicity have been described, however, in the literature there is a predominance of studies on iron deficiency. Due to little information about this topic, this work aimed to perform an analysis of these transcription factors. Here we managed to discover the transcription factor family which had its transcriptional activity altered more frequently in response to iron excess, predicted the protein-protein interaction of these genes and performed an analysis to detect conserved motifs in the promoter region of these genes.

2 Materials and Methods

According to the study of [4], differentially expressed genes under iron stress were selected. This genes where then characterized as transcription factors through Plant Transcription Factor Database (PlantTFDB) [5]. Their domains were identified in the Protein Families Database (Pfam) [6], based on the sequences predicted by Expert Protein Analysis System (ExPASy) [7]. Promoter phylogenetic analysis was performed using the sequences 1 kb upstream of each gene, using Molecular Evolutionary Genetics Analysis 6 - MEGA 6 [8] with ClustalW [9] and NeighborJoining [10] with 10,000 replicates of bootstrap. The identification of conserved motifs was performed using the same sequences in the Multiple Motif In Elicitation - MEME 4.4.0 [11], using a width ranging from 6 to 65 nucleotides using both strands, when analyzing promoter sequences, and 6 to 65 amino acids, when analyzing protein sequences, and a maximum of 10 motifs. Subsequently, these proteins were analyzed with the Predicted Rice Interactome Network – PRIN [12] and GENEVESTIGATOR [13] to verify protein-protein interactions and gene-expression analysis, respectively.

[1] Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900 (arturtaj@hotmail.com)
[2] Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900
[3] Universidade Federal do Rio Grande do Sul, UFRGS, Caixa Postal 110 CEP 90040-060
[4] Universidade Federal de Pelotas, UFPel, Caixa Postal 354 CEP 96010-900 (acostol@cgfufpel.org)

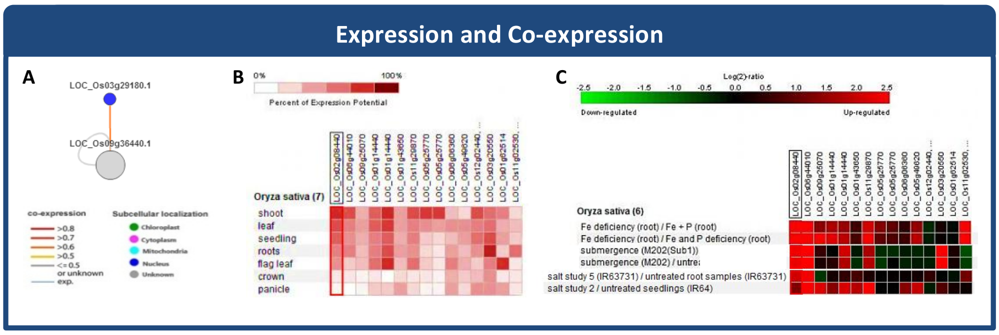
3 Results and Discussion

We identified the presence of 50 sequences characterized as transcription factors. Of these only one was down-regulated (Os02g0677300) while the others had an up-regulation. We analyzed the correlation of these proteins through PRIN, and found four recognized proteins, but only LOC_Os09g36440.1 presented correlation with another differentially expressed protein (LOC_Os03g29180.1). This co-expression value was of 0.6344. In this sense, the lack of information about protein-protein interaction on the sequences found in this work suggests a wide diversity, which is may not be well represented in the database.

We found 21 different domains characterized as transcription factors, wherein the most frequent was WRKY DNA-binding domain, with 15 occurrences, then the AP2 domain (6 occurrences) and Myb-like DNA-binding domain (5 occurrences). According to [14] the WRKY proteins are newly identified transcription factors involved in many plant processes, including plant responses to biotic and abiotic stresses. We analyzed the expression of these 15 WRKY genes in GENEVESTIGATOR, demonstrating that generally, when we observe the plant tissue we have many of these genes not expressed or have a low potential for expression in leaf, justifying this increase obtained in expression may have been promoted by stress conditions. For other stresses, we have iron deficiency stress with usually a high expression of these genes, the submergence stress has usually a down expression or no change in the expression of these genes and excess salt stress showed mostly a non-altered expression, due this we can see that the stress response mechanism which iron unleashed the expression of these genes following a different route response to other stresses.

The phylogenetic analysis of these 15 genes, formed three different groups when their amino acid sequence was analyzed, and other three groups when analyzing the nucleotide sequence of their promoter regions. The data obtained in classification of the amino acid sequences are confirmed by [14,15].

According to these results, we can conclude that changes in transcript accumulation of transcription factors are an important form of adaption to iron stress in *Oryza sativa* L. ssp. *japonica* cv. Nipponbare. There is still little information about iron stress in rice, and this work can enrich the databases aimed for breeding purposes. The identification of regulators of transcription can aid its use in genetic engineering and marker development.


Expression and Co-expression

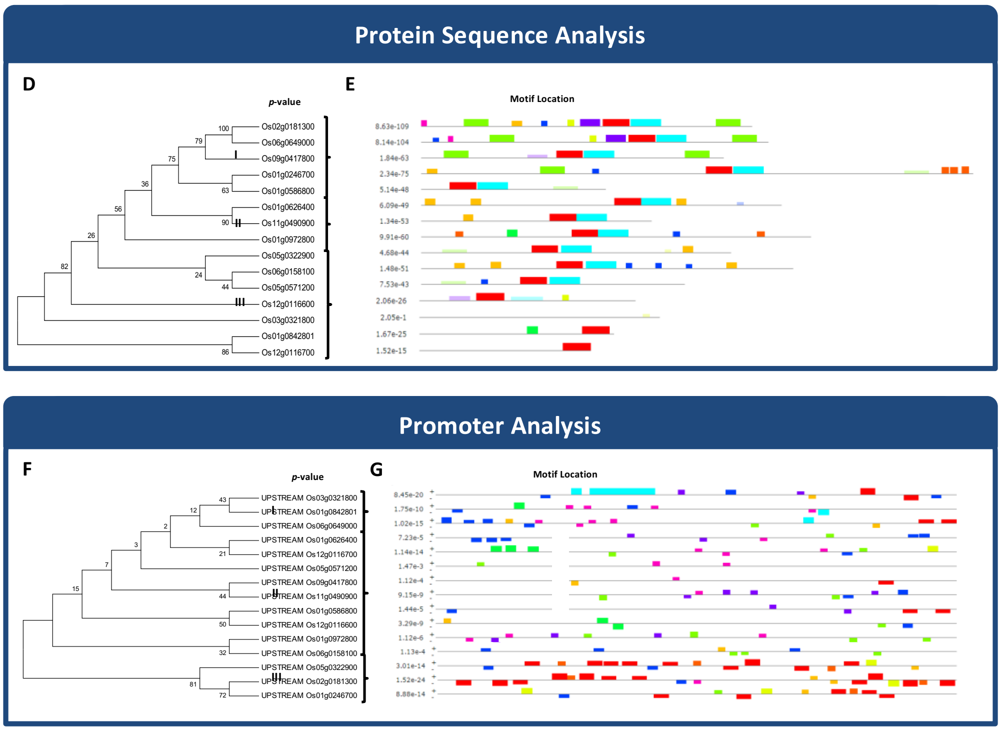


Figure 1 (A) Protein-protein interactions network. Expression analysis: (B) Localization and (C) Stress condition. (D) Phylogenetic tree and (E) Motifs found on WRKY proteins. (F) Phylogenetic tree and (G) Motifs found in the promoter region (1 kb upstream) of the analyzed WRKY genes.

References

[1] Pierre JL, Fontecave M. Iron and activated oxygen species in biology: the basic chemistry. Biometals. Sep;12(3):195-9. 1999.

[2] Zhu JK. Salt and drought stress signal transduction in plants. Annu Rev Plant Biol. 53:247–73. 2002.

[3] Krol R, Blom J, Winnebald J, Berhörster A, Barnett MJ, Goesmann A, Baumbach J and Becker A. RhizoRegNet—A database of rhizobial transcription factorsand regulatory networks. Journal of Biotechnology 155, 127– 134. 2011.

[4] Finatto T, Oliveira AC, Chaparro C, Maia LC, Farias DR, Woyann LG, Mistura CC, Soares-Bresolin AP, Llauro C, Panaud O and Picault N. Abiotic stress and genome dynamics: specific genes and transposable elements response to iron excess in rice. Rice. 8:13. 2015.

[5] Jin JP, Zhang H, Kong L, Gao G and Luo JC. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Research, 42(D1):D1182-D1187. 2014.

[6] BATEMAN A. et al. The Pfam protein families database. Nucleic Acids Research, Oxford, v. 30, p. 276-280, 2002.

[7] GASTEIGER E. et al. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research, Oxford, v. 31, p. 3784-3788, 2003.

[8] TAMURA, K. et al. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution, Oxford, v. 30, p. 2725–2729, 2013.

[9] 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 Research, Oxford, v. 22, p. 4673-4680, 1994.

[10] SAITOU N.; NEI M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, Oxford, v. 4, p. 406-425, 1987.

[11] BAILEY T.L. et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research, Oxford, p. W202-W208, 2009.

[12] GU H.B., ZHU P.C.; CHEN M. PRIN: a pedicted rice interactome network. BMC Bioinformatics, 2011.

[13] Zimmermann P, Hirsch-Hoffmann M, Hennig L, and W Gruissem GENEVESTIGATOR: Arabidopsis Microarray Database and Analysis Toolbox Plant Physiology 136 1 , 2621-2632. 2004.

[14] Zhang Y, Wang L. The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. BMC Evol Biol. Jan 3;5:1. 2005.

[15] Eulgem T, Rushton PJ, Robatzek S and Somssich IE. The WRKY superfamily of plant transcription factors. Trends Plant Sci. May;5(5):199-206. 2000.

A Distributed Knowledge-based Genetic Algorithm for Protein Structure Prediction

Jonas da S. Bohrer[1]
Bruno Borguesan[1]
Márcio Dorn[1]

The study of proteins and the prediction of their three-dimensional (3-D) structure is one of the most challenging problems in Structural Bioinformatics. Over the last years, several computational strategies have been proposed as a solution to this problem. As revealed by recent CASP experiments, the best results have been achieved by knowledge-based methods, but further research remains to be done. In this work, we propose a distributed knowledge-based Genetic Algorithm to predict the 3-D structure of proteins. A Genetic Algorithm (GA) is an adaptive heuristic search algorithm based on evolutionary ideas. GAs are modeled through the use of a population of individuals that undergo selection in the presence of variation-inducing operators such as mutation and recombination (crossover). A fitness function is used to evaluate individuals, and reproductive success varies with fitness. The success of the GA depends mainly on the balanced exploration of the solution space. When this balance is disproportionate, a premature convergence problem can occur, and the GA will lose efficacy. In protein structure prediction, the roughness of the protein energy surface poses a significant challenge to optimization techniques such as GAs. One approach to deal with this problem considers the use of a Distributed Genetic Algorithm. The basic idea of a distributed GA is to keep, in parallel, independent populations. We proposed a Distributed Genetic Algorithm based on the conformational preference of amino acid residues in experimentally-determined proteins. Each population incorporates an Angle Probability List (APL) derived from experimental data to generate the initial populations and new individuals to increase the diversity of the model. The proposed method has been tested with eight protein sequences. As corroborated by experiments, the developed method can produce accurate predictions, where the 3-D protein structures are comparable to their corresponding experimental ones. When compared with other first principle prediction methods that use database information, our approach presents advantages in terms of demanded time to produced native-like 3-D structures of proteins.

[1]Instituto de Informática, UFRGS, Porto Alegre, RS, Brazil.

{jonas.silveirabohrer,bborguesan,mdorn@inf.ufrgs.br}

OncoProSim: A tool for in silico tumor evolution analysis

Darlan Conterno Minussi[1,2]
Bernardo Henz[4]
Eduardo Cremonese Filippi-Chiela[3]
Manuel Menezes de Oliveira[4]
Guido Lenz[1,2]

The current knowledge regarding tumor evolution, until now, has been obtained from what we can gather through biopsy samples. In spite of the importance of these samples to diagnosis and treatment, they represent only a glimpse of the whole tumor evolutionary path, whereas the majority of the tumor development remains hidden. In this work, we attempt to simulate the clonal evolution of different types of tumors, focusing on the molecular mechanisms that lead to tumor progression. In order to do that, we generate cells with two arrays that represent the diploid characteristic of the human genome and, in case of cell division, mutations can be inserted in the genome that may alter the default probabilities of proliferation and death. With the help of a pseudorandom number generator, we can use our model to simulate different patterns of tumor evolution and investigate the effects of different mutational spectrum in the incidence of distinct tumors. Moreover, we can use our model to reproduce essential characteristics of the tumor genome such as: the synergy between oncogenes and tumor suppressor genes, changes in mutation frequency with distinct fitness values and even simulate tumor treatment. Therefore, with the increasing knowledge in tumor biology, we believe that our model can offer a unique perspective of tumor evolution, allied to the speed and reproducibility that only in silico models are capable of offering.

[1]Programa de Pós-Graduação em Biologia Celular e Molecular, UFRGS (darlanminussi@gmail.com)
[2]Instituto de Biociências, Departamento de Biofísica, UFRGS
[3]Faculdade de Medicina, UFRGS
[4]Instituto de Informática, UFRGS

Side-Chain conformational analysis of the multi-dependent rotamer preferences of proteins

Bruno Borguesan[1]
Mariel Barbachan e Silva[2]
Márcio Dorn[1]

Rotamer libraries are commonly used to correctly assign the dihedral side-chain angles for amino acid residues in protein structures. Rotamer libraries are widely used to assist the problems of Structural Bioinformatics like the protein structure prediction, protein design, structure refinement, homology modeling, and X-ray and NMR structure validation. These libraries are mostly classified as backbone-dependent, backbone-independent and secondarystructure-dependent. The first group are libraries that consists of rotamer frequencies, mean dihedral angles, and variances as a function of the backbone dihedral angles. The second group makes no reference to backbone conformation, but use side-chain information from all experimentally determined protein structures available. The third group present rotamer frequencies and dihedral angles for each secondary structures, such as α-helix, β-sheet and coil. However, even when these romater libraries are applied an enormous possibility of sidechain angles can be allowed. To reduce the complexity of the side-chain search within the Tertiary Protein Structure Prediction problem, we propose a novel approach that combine backbone-dependent, amino-acid-dependent and secondary-structure-dependent. To develop this method, we performed a study with 6,650 protein structures to analyze the conformational preferences for the side-chain angles of each amino acid residue, computed with PROMOTIF, in the secondary structure assigned by STRIDE. From that point forward, we combined the side-chain angles of each amino acid residue in the secondary structure with the frequencies of backbone dihedral angles. Based on this analysis, we developed a rotamer library that combine all these conformational preferences to assign the dihedral side-chain angles for amino acid residues in protein structures. This multi-dependent rotamer library was already used to assist implementations with PyRosetta in tertiary structure prediction.

This work was partially supported by grants from FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9).

[1]Institute of Informatics, UFRGS, Av. Bento Gonçalves 9500, 91501-970 Porto Alegre, RS, Brazil{bborguesan,mdorn@inf.ufrgs.br}

[2]Institute of Biosciences, UFRGS, Av. Bento Gonçalves 9500, 91501-970 Porto Alegre, RS, Brazil {mariel.barbachan@ufrgs.br}

A Multi-Agent Approach for the 3-D Protein Structure Prediction Problem

Leonardo de L. Corrêa[1]
Márcio Dorn[1]

The prediction of the three-dimensional structure of proteins is an important research area in Structural Bioinformatics. Proteins are present in all living systems, performing different functions. The function performed by a protein is strictly related to its adopted conformation. This is the main motivation for researchers in the field, besides the large gap between the number of known protein sequences and known 3-D protein structures. The protein structure prediction (PSP) problem is classified in computational theory as a NP-complete problem due the high dimensionality and complexity that the search space can assume, even for a small protein. Therefore, there is not any method capable to achieve the optimal solution. Currently, to predict the 3-D structure of a protein only from the amino acids sequence (ab initio methods), a wide range of optimization algorithms are being applied to find approximated solutions, as well as previous knowledge of known protein structures stored in a protein data bank in order to improve these methods and reduce the search space. In this work, we propose a multi-agent system to deal with the PSP problem. Multiagent systems are being used to face complex problems, devising tasks among the agents of the system and exploring a more distributed approach. An agent is a computer process that runs in an independent way, where under some given circumstances, interacts and cooperates with the other agents to solve a major problem. In our approach, we employed a population of thirteen agents that were organized in a structured ternary tree, whereas each agent performs specific roles and interacts in an attempt to obtain good solutions to the problem. We also incorporate, as behaviors of the agents, concepts of evolutionary-based algorithms, such as crossover and swap operators, and a Simulated Annealing implementation as a local search method. The algorithm also uses the information stored in the Protein Data Bank through an Angle Probability List (APL) to reduce the search space. The APL was partitioned into sub-groups, according to the combination of amino acids residues and their respective secondary structures, resulting in different patterns. The population of agents was divided into subpopulations, to increase the diversity in the space of solutions, each agent could see only one pattern of the APL. We used the Rosetta energy function as scoring function. The system was tested against three protein sequences, where for each one the algorithm was run four times for eight hours. Preliminary results show that the proposed approach has a promising ability to predict good three-dimensional conformations in terms of potential energy and RMSD measurement when compared with the experimental protein structures.

This work was partially supported by grants from FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9)

[1]Institute of Informatics, Federal University of Rio Grande do Sul, Post Code 24105
{llcorrea, mdorn@inf.ufrgs.br}

Effects of Zn^{+2}, N-Glycosylation and Dimerization on Auxin Binding Protein 1 (ABP1) Conformation and its Interaction with Auxin

Cibele Tesser da Costa[1], Conrado Pedebos[1], Hugo Verli[1], Arthur Germano Fett-Neto[1]

Auxin is a critical phytohormone for plant growth and development, influencing aspects of cell division, elongation and differentiation. One natural form of auxin is the indole-3-acetic acid (IAA), whereas 1-naphtalene acetic acid (NAA) is a synthetic form. Auxin Binding Protein 1 (ABP1) is an auxin receptor that plays key roles in fast nontranscriptional responses. Its structure was previously established from maize at 1.9 Å resolution and each subunit of ABP1 is glycosylated through a high mannose-type glycan structure. In dicot species it has an additional glycosylation site. We aimed at expanding the knowledge of ABP1 structural biology employing molecular dynamics simulations of the complete models of the oligomeric glycosylated proteins for maize and Arabidopsis thaliana with or without auxins. In maize, both coordination to Zn^{+2} and glycosylation were able to promote increased conformational stability whereas in Arabidopsis we observed only a minor effect of both factors. Most of such stabilization effect was located on N- and Cterminal regions. C-terminal regions of ABP1 from both species contain an α-helix and in the performed simulations this helix unfolded, assuming a more extended structure in all replica simulations of maize ABP1. In Arabidopsis the helix seems to be more stable, beingpreserved in most of the monomeric simulations and unfolding when the protein was in the dimeric form. When Arabidopsis ABP1 was bound to IAA or NAA, the glycosylation structures arranged around the proteins, covering the putative site of entrance or egress of the ligand. Along the simulation, when NAA was bound to the proteins, the fold was more similar to the crystal structure and the complex showed higher stability compared to IAA binding. This work contributed to a better understanding of the effects of Zn^{+2}, glycosylation and dimerization in the structure of the protein ABP1 and its interaction with auxins.

Support: Capes, CNPq and Fapergs

[1]Center for Biotechnology, Federal University of Rio Grande do Sul (UFRGS), CP 15005, Porto Alegre, RS, 91501-970, Brazil

Memetic algorithm for Docking's problem with a rigid protein and a flexible ligand

Felipe Gonzalez Foncea[1]
Mario Inostroza-Ponta[1]
Márcio Dorn[2]

Docking is a problem of bioinformatics, which has achieved acceptance over the past 20 years, due to advances in the field of molecular biology. The search for new structures has defined the Docking as a viable alternative for finding an optimal bonding between two molecules, mainly a protein and a ligand. Protein-ligand structure can generate alterations in signal transduction process, which is a key factor in the chemical processes of a biological body. This knowledge will validate the creation or discovery of new types of drugs. As a research proposal, It has been worked to generate a new alternative for Docking problema, through a memetic algorithm and the Rosetta software suite as a scoring function, which evaluates the level of protein-ligand affinity through free energy binding. The model is restricted to use proteins without any change on the structure during all the evaluation process. However ligand will be able to change their structure to generate more alternatives of solution. The flexibility of the protein structure will be produced by the algorithm, that changes the dihedral angles of the ligand. The results will be evaluated by the scoring function. The choice of a rigid protein and a flexible ligand is focused on getting a good result without high computing power. In addition it would be possible to use this model as a local search for more complex models both protein and ligand flexible. The objective of this algorithm is solve most of the problems associated to Docking. The results will be analyzed by data quality through RMSD (Root-Mean-Square Deviation) and it will be compared with other Docking programs like AutoDock and Dockthor.

[1] Departamento de Ingeniería Informática, Universidad de Santiago de Chile
{felipe.gonzalez.f@usach.cl, mario.inostroza@usach.cl}
[2] Institute of Informatics, Federal University of Rio Grande do Sul
{mdorn@inf.ufrgs.br}

Análises de elementos transponíveis em *Angiostrongylus cantonensis* (*Nemathelminthes: Angiostrongylidae*)

Alice Giovana Buzetto[1]
Leandro de Mattos Pereira[1]
Gabriel da Luz Wallau[1]
Amaranta Ramos Rangel[1]
Alessandra Loureiro Morassutti[1]
Carlos Graeff-Teixeira[1]

Tansposons são elementos móveis do genoma que se fazem presentes em todos os organismos, correspondendo até 50% do genoma em algumas espécies. Pela sua natureza móvel, possuem grande importância genômica, sendo fontes de mutações espontâneas, rearranjos cromossômicos, produzem efeitos na evolução das espécies e podem ser utilizados como marcadores moleculares. O estudo dos transposons em uma espécie de nematódeo com importância em saúde pública, *Angiostrongylus cantonensis*, tem como objetivo identificar a presença destes elementos móveis a fim de compreender seu papel na evolução da espécie e a nível molecular, bem como sua possível relação com os mecanismos de patogenicidade. Foram utilizados dois métodos de análises para a identificação dos transposons: análise por homologia (TBLASTN) e *ab initio* e homologia (RepeatExplorer). A análise baseada em homologia foi aplicada utilizando o algoritmo TBLASTN, a qual consistiu em comparar as sequências de transposons de *Caenorhabditis elegans* com sequências genômicas de *A. cantonensis*. Enquanto que a análise híbrida (RepeatExplorer), executa uma análise global no genoma de interesse. Os dados preliminares apresentam 1.149 sequências de elementos repetitivos, obtidas na análise por TBLASTN, curadas no banco de dados RepBase, utilizando o algoritmo CENSOR. Destas sequências, observam-se retrotransposons e transposons, classificados em classe e família, tendo destaque a família de transposon Mariner/Tc1, por compreender maior número de repetições. As análises com o programa RepeatExplorer revelam 124 clusters, compreendendo sequências repetitivas de retrotransposon e transposon. Os resultados obtidos até o momento, associados a dados já publicados com outros organismos, sugerem a participação destes elementos na evolução das espécies, uma vez que identificam-se alguns transposons de mesma classe e família em diversos organismos, previamente descritos e depositados em banco de dados.

[1] Laboratório de Biologia Parasitária - Pontifícia Universidade Católica do Rio Grande do Sul

A Python-Based API of the 3D-Tree Algorithm

Aline Kronbauer[1] ;Bruna L. Balbinot[1], Leonardo A. Schmidt[1], Leonardo N. Machado[1],
Raphael G. Nascimento e Silva[1], Karina S. Machado[2], Ana T. Winck[1]

Rational Drug Design is an important field of bioinformatics, treating mainly about interaction between macromolecules, called as receptors, and small molecules, called as ligands. By means of molecular docking and the estimated Free Energy of Binding (FEB) we can identify the best bind between the molecules. Many studies consider the receptor as a rigid structure, ignoring its flexibility. Aiming at reaching better results we make use of several molecular docking experiments, being that each of them run on a given receptor conformation generated by molecular dynamic (MD) simulations. However, we know that such strategy generates a lot of data, turning it a timing-consuming process. Different methods have been developed to treat this issue, and some of them try to select the most promising conformations from the whole MD. In this work we focus on the 3D-Tree algorithm, which identify the best pose of relevant receptorâs atoms that leads to a good estimated FEB value. We are implementing a Python-based version of the 3D-Tree algorithm, aiming to turn available an API to provide the whole process of this algorithm. In summary, the 3D-Tree algorithm treats spatial coordinates in a x, y, z format, and induce a decision-tree that predicts FEB. The algorithm uses the spatial coordinates to binary split a node, where the edges evaluate whether the atom is part of a given block. Hence, we develop a full preprocessing step, where a set of PDB files can be uploaded to generate a dataset fitting the 3D-Tree file format. Having these dataset, the API is able to submit them to the next two main steps of the algorithm, which regards the block generation and the induction decision-tree step. Currently, we are testing our implementation on the AcrB protein, which contains 7.639 atoms and 1.001 PDB files obtained by MD simulations.

Acknowledgments: This work was supported by grants from MCT/CNPq 14/2013 to ATW and KSM. AK is supported by a FAPERGS scholarship. The 3D-Tree algorithm is a PhD theses developed at PUCRS from 2008 to 2012.

[1]Grupo de Pesquisa em Sistemas Inteligentes - UFSM
alyne.k@gmail.com,{bbalbinot,lschmidt,lmachado,rnascimento,ana}@inf.ufsm.br
[2] Grupo de Pesquisa em Biologia Computacional - FURG
karina.machado@furg.br

Data Mining Applied to Molecular Docking Experiments

Luisa Rodrigues Cornetet[1]
Michael González-Durruthy[2]
José M. Monserrat[2]
Adriano V. Werhli[3]
Karina dos Santos Machado[3]

Introduction

Since Bioinformatics experiments as molecular docking and virtual screening usually produce a lot of data, it is really important to apply data mining techniques to discover useful information about this generated data. In this paper, we propose to apply data mining classification algorithms into the results of molecular docking experiments with carbon nanotubes and the protein ANT [ant].

Molecular docking is a computer technique that helps to understand the interaction between biological macromolecules, called receptors, and some small molecule that are inhibitor candidates, called ligand. The analysis of molecular docking allows to understand if a ligand acts as an inhibitor or antagonist during a physiological process [bioinformatica]. For the molecular docking, we perform the experiments using Autodock Vina [vina]. Autodock Vina is a open-source software developed by Trott et al [vina]. As the results of the molecular docking simulation we obtain the conformations of the ligand and the Free Energy of Binding (FEB) of the receptor-ligand complex. Also, we used a framework [luisa] to make the execution easier and calculate the minimum distances between atoms of the aminoacid and the ligand on the results.

Metodology

In this section we will present the receptor and the ligands used in the docking experiments, as well as the pre-processing of the data for the data mining process.

As receptor in the molecular docking experiments we consider the protein ANT [ant], or Adenine Nucleotide Translocator. This protein is located on the inner wall of mitochondria and is an essential part on the process of energy production for a eukaryotic cell. ANT acts as a carrier, and performs the exchange between ADP and ATP. This exchange works with importing ADP and exporting ATP across the inner membrane of mitochondria. The structure of this protein was obtained from Protein Data Bank (PDB) [pdb] with the PDB ID 1OKC.

[1] Centro de Ciências Computacionais, FURG.
{cornetet.luisa@gmail.com}
[2] Instituto de Ciências Biológicas, FURG.
[3] Centro de Ciências Computacionais, FURG.

As ligand we have considered several types of carbon nanotubes. The carbon nanotubes are a material made of a graphene sheet in a tube form, and the diameter in nanometric scale [SWCNT]. In this work we have varied parameters such as the Hamada Index and the oxidation of the carbon nanotube. All the experiments were performed with single wall carbon nanotubes (SWCNT). The Hamada Index [hamada] is expressed by two parameters,n and m, that classify the carbon nanotubes for its diameter and helicity. As de n and m parameter vary, we have three classifications for this nanotubes: armchair (when m = n), chiral (when m 6 = n and m = 0) and zigzag (when m = 0). We have used 7 armchair nanotubes, 31 chiral nanotubes and 7 zigzag nanotubes. About the oxidation, we used pristine nanotubes, and also nanotubes with the oxidation groups carboxyl (COOH) and hydroxyl (OH), in all the previous described nanotubes.

As ligand we have considered several types of carbon nanotubes. The carbon nanotubes are a material made of a graphene sheet in a tube form, and the diameter in nanometric scale [SWCNT]. In this work we have varied parameters such as the Hamada Index and the oxidation of the carbon nanotube. All the experiments were performed with single wall carbon nanotubes (SWCNT). The Hamada Index [hamada] is expressed by two parameters, n and m, that classify the carbon nanotubes for its diameter and helicity. As de n and m parameter vary, we have three classifications for this nanotubes: armchair (when m = n), chiral (when m 6 = n and m = 0) and zigzag (when m = 0). We have used 7 armchair nanotubes, 31 chiral nanotubes and 7 zigzag nanotubes. About the oxidation, we used pristine nanotubes, and also nanotubes with the oxidation groups carboxyl (COOH) and hydroxyl (OH), in all the previous described nanotubes.

Results and discussion

In this section we describe and analyze the data mining techniques applied to the results of the molecular docking experiments. For the data mining algorithms we used Weka [wekaUpdate], that provides an environment for some techniques as automatic classification, clustering and regression [weka].

We applied J48 [j48] and REPTree [reptree] classification algorithms. In both cases we validate the obtained models with a 10-fold cross validation. Figure 1 shows the decision tree generated by J48 algorithm. This model had accuracy of 73.88%. The root node of the tree had divided the instances according to the Oxidation. According to Figure 1 we can observe that distances between the aminoacid Arg 234 and the nanotubes greater than 13.71Å produces better results, we believe it occurs because this aminoacid is a little lower in the protein, and the nanotubes were in a position above.

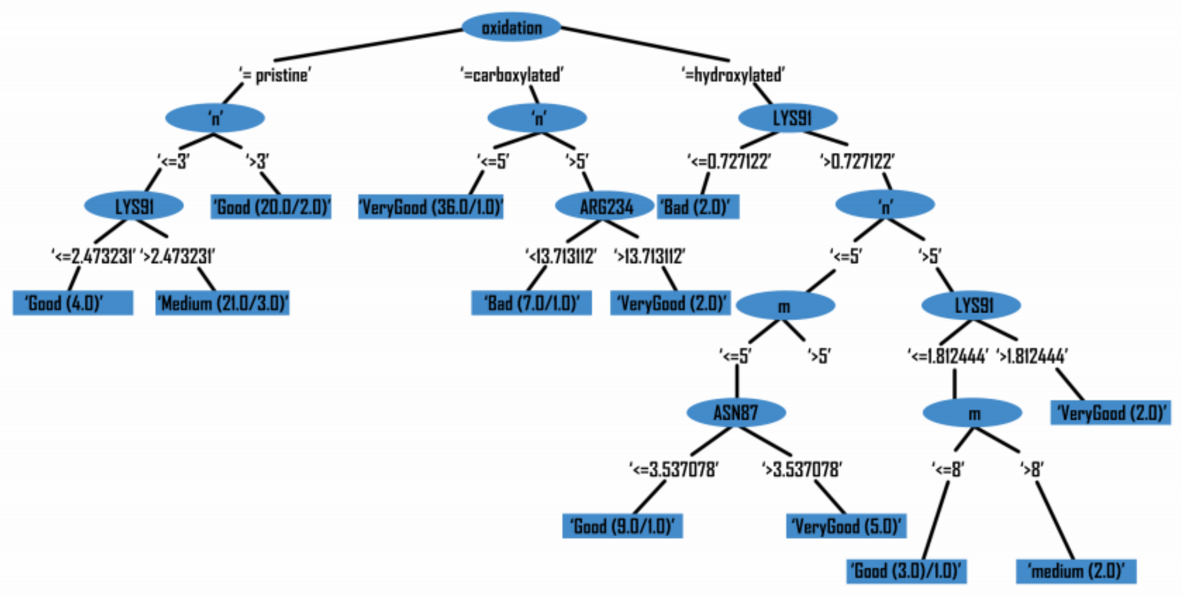


Figure 1. J48 results

Figure 2 shows the decision tree created by REPTree algorithm. It correctly classified 72.38% of the instances. As well as J48 decision tree, the root node of the tree is Oxidation. First of all, it says that every carboxylated nanotube has a Very Good FEB result, since the best experiment results are related to nanotubes that has the carboxyl group. Besides Good or Bad FEB results for pristine nanotubes is related to its size: the biggest ones ($n \geq 3.5$) has Good FEB values, while the small ones ($n < 3.5$) were classified as Medium. Finally the hydroxylated nanotubes presented Medium results for the ones with $n \geq 6.5$, and Good or Very Good for the others. It is also observed that very small distances between Lys 91 (less than 2.43) and the nanotubes obtained docking results with Bad FEB values and bigger values of distance between this aminoacid produces a Good result. We can explain this because very small distances between Lys 91 and the nanotubes might produce some collision of atoms.

Conclusion

Based on the presented results, we conclude that applying data mining techniques is very effective to discover important information about the ANT-Nanotubes molecular docking experiments. The results gave us a different point of view about the interactions between this target protein and different carbon nanotubes, allowing us to do further research about the results of the experiments.

Analysis of the metabolic potential of Angiostrongylus cantonensis

Amaranta Ramos Rangel[1], Leandro de Mattos Pereira[1], Alice Giovana Buzetto[1], Alessandra Loureiro Morassuti[1], Carlos Graeff-Teixeira[1]

Angiostrongylus cantonensis, the rat lungworm, is an etiological agent of eosinophilic meningitis in humans, has as hosts: Veronicellidae (intermediate), Rattus norvegicus (definitive) and Homo sapiens (accidental). This species has been founded in Southeast Asia, Pacific Basin and South of the Brazil (Porto Alegre). Currently, little is known about the metabolic pathways and biologic processes present in the genome and transcriptome of Angiostrongylus. Genome-wide analyses using computational biology are of fundamental importance for elucidating molecular mechanisms, metabolic pathways and molecular marker to diagnostics. From a holistic approach, the aim of this study was to evaluate the metabolic potential of A. cantonensis through of annotation of all metabolic pathways present in the genome and transcriptome of this parasite. We obtained the predicted proteome in the database WormDB and the transcriptome was sequenced by Macrogen with the Illumina platform. The Trinity software was used to de novo assembled of transcriptome and Trinotate, Blast2GO for functional annotation. Through of the AnEnPi we mapping all metabolic pathways present in the genome and transcriptome of A. cantonensis, including possible alternative routes. Our preliminary results shows that the predicted proteome (11,998 sequences) consists of 634 enzymatic activities, with 136 Oxidoreductases (Ecs1), 226 Transferases (ECS2), 193 hydrolases (Ecs3), 34 Lyases (Ecs4), 28 Isomerases (Ecs5) ligases (Ecs5), totaling 33% of the proteome. Among the metabolic pathways, we found metabolic pathways important for the survival of the parasite, such as: drug metabolism, carbohydrates metabolism and synthesis of nitrogenous bases.

Supported by Capes, Brazil. Acknowledge: Laboratório de Biologia Parasitária, PUCRS; Programa de Pósgraduação em Biologia Celular e Molecular, PUCRS.

[1]Laboratório de Biologia Parasitária- PUCRS
Pontifícia Universidade Católica do Rio Grande do Sul.

Development of a Genetic Algorithm for Protein-Ligand Redocking

Eduardo Spieler de Oliveira[1]
Márcio Dorn[1]

Molecular docking problems refer to the prediction of the conformation between a small molecule (ligand) and a receptor molecule with minimum binding energy. The quality of this binding depends on the score function and on the computation methods applied. The redocking method is often performed to verify if the docking parameters are able to recover a known complex structure and interactions. The development of docking methods include the use of metaheuristics to improve prediction capability. In this work, a Genetic Algorithm (GA) was developed in order to search efficiently the conformational space of the Protein-Ligand and find the minimum binding energy redocked structure. As a first study both structures were considered rigid and Rosetta score function was used to calculate the energy of the docking structures. Random perturbations of positions, such as translations and rotations, called individuals, were made in the ligand to search the ideal biding location. A population of one hundred individuals was created. The algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm randomly selects individuals from the current population and uses them as parents to produce the offspring for the next generation. The structure 1ENY from the Protein Data Bank was used as a reference. The algorithm ran fives times, each one for eight hours, over three hundred generations were tested toward an optimal solution. The final structure was compared with the original crystallographic structure in therms of root-mean-square deviation. Preliminary results show that the proposed algorithm can find good solutions in terms of RMSD when compared with the experimental crystallographic 3D structure.

Acknowledgements: This work was partially supported by grants from FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9).

[1] Instituto de Informática, UFRGS, Porto Alegre, RS, Brazil. {esolivera,mdorn@inf.ufrgs.br}