Partners

Supporters

open days 2014
BIOINFORMATICS

27 - 28 February
thursday friday

Amphitheater
B1 CP2
Universidade do Minho
Campus Gualtar
Braga

Organized by
Mestrado em Bioinformática
(UMinho)

Book of Abstracts

Organized by Mestrado em Bioinformática

Ana Alão Freitas
Brígida de Meireles
Diana Barros
Hugo Giesteira
Hugo Zenha
Luis Xandy Anjos
Nídia Ferreira
Pedro de Carvalho
Sara Carvalho
Vinício Oliveira
Vítor Silva

Scientific Committee
Miguel Rocha, Universidade do Minho
Isabel Rocha, Universidade do Minho

The Bioinformatics and Computational Biology fields integrate biological research efforts with data analysis, mathematical modeling, in silico simulation and optimization, aiming to gather large-scale databases and develop new computational technologies, to promote biological discovery and to help in tasks related with modeling and optimization, such as in the field of Systems Biology.

As the importance of Bioinformatics grows, the third edition of the Bioinformatics Open Days event takes place, organized by the students of the Masters in Bioinformatics course of the University of Minho and its Directive Committee. This event takes the form of a conference with several presentations of the latest work that is being developed in the Bioinformatics field. The Bioinformatics Open Days acts as a promoter and aims to boost the students' interest over the Bioinformatics area showing that it is an ever growing science and why it is so useful.

Contents

27 February
thursday

09h30 Registration / Coffee

10h00 Opening Session

10h30 Contributed Presentations

Tânia Mendes: "A flexible-docking approach for the design of novel cancer peptidomimetic drugs"

David Bowman: "Improving the performance of molecular dynamics simulations a noncomputacional approach"

Raquel Trindade: "Development of a chemical compounds database targeting amyloid proteins"

11h30 Invited lecture

Daniel Glez-Peña (U. Vigo): Bioinformatics Applications in the SING research Group. Case study: Mass-Up, a proteomics workbench.

12h30 Lunch

14h00 Contributed presentations

Gianluca Selvaggio: "Peroxiredoxin/Thioredoxin system: forecasting the peril"

Miguel Coimbra: "Computational metagenomic techniques applied to the prospection of genes linked with hydrocarbon biodegradation and surfactant production "

Natacha Leitão: "LLC - GNUMAP: Scalable, Precise and High-Coverage Genomics Mapping"

Diogo Pratas: "Insights into genomic evolution by sequence smashing"

Lara Silva: "Development of a semantic model for stroke patients"

Hugo J. Froufe: "Bioinformatic challenges in human genome and exome analysis"

Manuel Reis: "Computational prediction of microRNA targets in plant genomes"

16h30 Coffee-break

17h00 Session: Teaching in Bioinformatics

presentation of the Masters in Bioinformatics (U.Minho) and Bioinformatics and Computational Biology (U.Lisboa)

Round-table discussion

20h00 Dinner

28 February
friday

10h00 Contributed Presentations

José Carrilho: "Prediction relevant events in the life of a person with Alzheimer"

André Lamúrias: "Identifying interactions between chemical entities in Text"

Alessandro Bolli: "Organizing principles underlying microorganism's growth-robustness trade-off"

11h00 Coffee-break

11h30 Contributed Presentations

Mariana Xavier: "Image processing on animal cell cultures: a refined technique"

Franklin Nobrega: "Targeted therapy using phage technology: A computational and experimental breast cancer study"

12h00 Invited lecture

Nuno B. Morais (U. Lisboa): Analysis of Transcriptional Regulation and RNA Processing from RNA-Seq Data

13h00 Lunch

14h30 Session: Bioinformatics companies in Portugal

presentations and round table discussion

Participants:

BSim2 – Rui Brito

Bial – Marco Neves

CGC Genomics – Purificação Tavares

Biocant / Genoinseq – Conceição Egas

SilicoLife – Simão Soares (moderator)

17h00 Closing Session

Invited Speakers

Bioinformatics Applications in the SING research Group. Case study: Mass-Up, a proteomics workbench.

Daniel Glez-Peña
University of Vigo

Abstract

Since the last years, the SING research group has been developing many applications in several Bioinformatics areas, such as microarray analysis, functional genomics, phylogenetics and biomedical text mining. Recently, we have been involved in proteomics studies, where we have adapted and integrated many techniques needed to handle MALDI-TOF-MS data. In this area, we have developed Mass-Up, an all-in-one environment on top of our AIBench application framework, which enabling researchers to accomplish complex analyses in an easy way.

Analysis of Transcriptional Regulation and RNA Processing from RNA-Seq Data

Nuno Barbosa Morais
University of Lisbon

Abstract

The recent advent of next-generation sequencing (NGS) technologies is revolutionizing research in biological and medical sciences. Most molecular mechanisms associated with disease ultimately involve transcriptomic variation and RNA-Seq (the use of NGS to sequence cDNAs reversely transcribed from RNAs) can not only measure gene expression but also quantitatively reveal unknown transcripts and splicing isoforms. For instance, the investigation of cancer transcriptomes can now be expanded to alternative transcription, gene fusion, RNA editing, and non-coding RNAs.

A myriad of bioinformatics tools for RNA-Seq data analysis have allowed for ever more accurate mapping, transcriptome assembly, gene fusion detection, gene and isoform expression analysis, and alternative splicing detection.

However, few RNA-Seq surveys of cancer transcriptomes have taken the study of misregulation of transcription and RNA processing in cancer beyond the analysis of somatic mutations and differential gene expression.

In this seminar, methods for the analysis of some important levels of regulation of transcription and RNA processing from RNA-Seq data will be described. We will discuss, for example, the use of a junction-centric approach and the importance of mappability in the analysis of alternative splicing, techniques for the identification of alternative polyadenylation based on the analysis of 3′ UTR length, and the integration of other types of NGS data for the detection of long non-coding RNAs.

Selected Abstracts

Oral Presentations

A flexible-docking approach for the design of novel cancer peptidomimetic drugs

Tânia S. Mendes *, Vera L. Silva, Franklin L. Nobrega, Joana Azeredo, Leon D. Kluskens and Lígia R. Rodrigues*
taniamendes@ceb.uminho.pt/lrmr@deb.uminho.pt

Centre of Biological Engineering, Universidade do Minho
Campus de Gualtar, 4710-057 Braga, Portugal

Abstract

Cancer is the second leading cause of death worldwide and the lack of alternative therapies has kept patients dependent on classic chemotherapy. The most occurring form of cancer amongst women is breast cancer and the triple negative cell subtype (TNBC) is responsible for a high metastatic and mortality rate, as it presents molecular and genetic shifts, lacks specific targeting and responds poorly to existing therapies. Recent studies have provided convincing evidence that the therapeutic outcome of chemotherapy may be affected by the expression and activity of receptor tyrosine kinases and their phosphatase pathways (MAPK/ERK, PI3/AKT, among others). These proteins are implicated in mechanisms of cell survival, drug resistance and Epithelial-Mesenchymal Transition (EMT), and are therefore key targets for TNBC cancer cell subtype. However, many inhibitors developed for these targets have not succeeded at a clinical level and present low solubility.

As a result of the pronounced decline in productivity experienced by drug discovery efforts in the last years, novel approaches to the rational design of new drugs are now being pursued. A potential solution might be the use of natural or synthetic peptides and peptidomimetics targeting protein-protein interactions essential for signaling networks function. The combination of several bioinformatic approaches (docking, virtual screening, pharmacophore models, among others) allows the use of the vast amount of existing information on available compounds and protein-protein interactions in structural databases.

In this study we designed a procedure for small peptidomimetics structure-based rational drug design capable of blocking the active sites of SNAiL1, a protein that has been suggested as a potent repressor of E-cadherin expression and consequently, as an inducer of EMT transition in TNBCs. A random library was created using a composite approach for drug-like compound identification from the PubChem and Development Therapeutics NCI/NIH compound databases, which combined structure-based virtual screening (known motifs of peptide structures within proteins and small molecules) and Z-score comparison. Docking studies were performed to map the polypeptides activity and stability: (1) point alteration studies using non-natural aminoacids for helical stability over a wider range (since linear peptides adopt many confirmations in aqueous solution) using Ramachandran plot dihedral angles estimation; (2) quantitative structure-activity relationships (QSAR) using radical modification chemical studies and (3) umbrella sampling for dissociations studies. The peptidomimetic SNAiL1 model created suggested at least two radical modifications for a strong inhibition.

All authors contact details:

taniamendes@ceb.uminho.pt
vera.7.silva@gmail.com
franklin.nobrega@ceb.uminho.pt
jazeredo@deb.uminho.pt
kluskens@deb.uminho.pt
lrmr@deb.uminho.pt

Improving the performance of molecular dynamics simulations
A noncomputational approach

David Bowman 1,2,3, Armindo Salvador 3, Paulo Martel 1,2

1- PhD Program in Biochemistry and Bioinformatics
Univ. of the Algarve;
2- Institute for Biotechnology and Bioengineering (IBB)
Univ. of the Algarve;
3- Molecular Systems Biology Group, CNC
Univ. of Coimbra;

Abstract

Molecular dynamics simulations are important to the understanding of biomolecular systems and are used for research in the areas of: membrane dynamics, protein folding and unfolding, and small molecule behavior. These simulations can run for days or even weeks. Typically simulations are in the range of tens to hundreds of thousands of atoms. The primary performance problem is the calculation of the nonbonded force interactions (Coulomb and van der Waals) between the solvent (typically water) molecules. Interactions between water and solute and solute to solute represent a much smaller part of the computational cost of simulations.

Molecular dynamics simulation software use methods (e.g. lattice summation or spherical cutoffs) to reduce the number of interactions from O(N2) to O(NlogN) or O(N). GROMACS (generally considered to be the fastest) also exploits the latest in computer instructions, multi-core, and multiprocessor capabilities and tools such as MPI. Even with these techniques the computational requirement is extremely large and simulations of biomolecular processes needs to be performed at least in the range of nanoseconds to milliseconds. There also exist fundamental limitations on how finely a simulation can be divided and distributed across multiple processors or cores. As a simulation is split between an increasing number of cores/processors the communication cost between the cores/processors become a higher and higher percentage of the total time so that there are limits on the scalability. This limits the number of nanoseconds/hour that can be run for a given simulation. This also means that for simulations in the range of 10s of thousands of atoms will not perform better on the most powerful supercomputers than on an affordable 64 core server. We have recently developed methodologies and algorithms in both the C language and hand coded assembler routines method provide a 'non-computational' approach based on the definition of the simulation and the incremental assembly of pre-calculated results. A new mathematical model has also been developed providing a variable precision floating point calculation model based on the IEEE 754 standard to reduce the size of the pre-calculated results. This algorithm may be used in any application that is computationally intensive. There are no approximations other than those reflected in the reduced precision. The methodology also could benefit from new Intel computer instructions and these have been designed.

The results of this methodology show an improvement in the calculation of nonbonded force interactions in the inner loops of GROMACS for the Intel Core i7 processor for the water to water interaction simulation component of 14-15 times that of the GROMACS hand coded assembler routines. The methodology applied to mathematical functions in the C programming library (tan, sin, cos, log, pow) performs 30-125 times faster.

The methodology has been evaluated using proteins, small molecule and free energy simulations. Many microseconds of simulations have been run as part of the project for validation. The free energy studies showed that the results of the methodology were statistically equivalent to the standard GROMACS code.

Development of a Chemical Compounds Database Targeting Amyloid Proteins

Raquel Trindade a,#, Cândida G. Silva a,b, Carlos J. V. Simões b,c
Rui M. M. Brito a,b,*

a - Center for Neuroscience and Cell Biology
Universityof Coimbra, 3004-517 Coimbra, Portugal
b - Chemistry Department and Coimbra Chemistry Centre
Universityof Coimbra, 3004-535 Coimbra, Portugal
c - BSIM2 – Biomolecular Simulations
BiocantPark, 3060-197 Cantanhede, Portugal

Abstract

Amyloidoses encompass a broad spectrum of diseases in which normally innocuous proteins or their peptide fragments abnormally aggregate into cytotoxic species and eventually deposit in organs and tissues in the form of fibrils. Among some of the best known amyloid diseases are Alzheimer's, Parkinson's, type II diabetes mellitus and familial amyloid polyneuropathy (FAP). The search for strategies for early diagnosis and treatment of these disorders are among the most challenging areas in modern medicine. A growing number of active small molecules capable of interacting with amyloid aggregates and fibrils has been reported which opens the opportunity to gain a better understanding of the substructures encoded in these compounds and responsible for amyloid recognition. These substructures may then be a source of inspiration to develop new molecules with potential to be used as diagnostic tools or therapeutic agents.

Based on public data on compounds with activity information on amyloid fibrils available on ChEMBL database, we propose a workflow to retrieve these data, integrate complementary structural and chemical information and annotate it with hundreds of molecular descriptors, delivering the result in the form of an amyloid-focused relational database. This resource should act as a knowledge-driven infrastructure for the identification and characterization of features linked to the activity profiles that different small molecules exhibit on amyloid aggregates and fibrils, providing the solid grounds for knowledge extraction.

The workflow implemented for the population of the database involves several steps. First, all amyloid-related compounds were collected from ChEMBL, and the three-dimensional structure was obtained using a set of open-source and academic licensed tools such as Open-Babel and OpenEye software. Second, for all compounds, a wide range of molecular descriptors and fingerprints was calculated by powerful chemical computing tools such as ChemAxon software, CDK, Filter, Mold2 and OpenBabel. Third, importing mechanisms were created to store all the new information in the database.

A graphical user interface was also developed for users to access, visualize and edit data stored in the database. This interface allows users to perform different tasks: (i) add and remove compounds from the study, (ii) calculate and store molecular descriptors and fingerprints for each compound, and (iii) execute queries on specific compound characteristics. The results of the queries can then be employed in data mining, clustering and statistical treatment of the different compound features. Furthermore, to guarantee the consistency and security of the information in the database, different user roles have been created. The registration and management of users' profiles is also managed through the user interface.

Acknowledgments

This work is funded by ERDF - European Regional Development Fund through the COMPETE Programme (Operational Programme for Competitiveness) and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project PTDC/QUI-QUI/122900/2010.

Presenting author: raqueltrindade@qui.uc.pt
* Corresponding author: brito@ci.uc.pt

Peroxiredoxin/thioredoxin system: Forecasting the peril

Gianluca Selvaggio 1,2, Pedro Coelho 1, Armindo Salvador 1,3
1 - Centre for Neuroscience and Cell Biology
University of Coimbra
2 - Faculty of Science and Technology
New University of Lisbon
3 - Chemistry Department
Faculty of Science and Technology of the University of Coimbra

Corresponding Author: gianluca.selvaggio@gmail.com

Abstract

Organisms have a complex and highly differentiated cocktail of antioxidant defenses to cope with a wide army of harmful reactive species. The proteins involved in hydrogen peroxide (H2O2) elimination (e.g. peroxiredoxins) are among the most abundant ones in some cells, and likely cannot be substantially upregulated without outcompeting other important cellular functions for limited resources. However, in natural environments cells occasionally have to cope with high H2O2 concentrations for long periods, saturating the scavenging capability of the defense systems, thus irreversibly damaging the proteins. Due to their involvement in catalysis and folding maintenance, the irreversible oxidation of reactive thiols is very deleterious for organisms.
This outcome can be avoided if cells have mechanisms to “block” the thiols through reversible covalent modification, ”anticipating” the saturation or exhaustion of the defenses.

We term these mechanisms “anticipatory blocking”.

Here we hypothesize that the Peroxiredoxin / Thioredoxin / ThioredoxinReductase / Protein-Dithiol System (PTRTD System) drives anticipatory blocking of protein dithiols as disulfides. We examined the design requirements for such a system to operate effectively and we compared these requirements to the actual design in human erythrocytes. To that effect, we developed a minimal mathematical model of the PTRTD System and defined a set of quantitative performance criteria that embody the requirements for (a) efficient scavenging capacity, (b) low NADPH consumption, (c) effective signal propagation (e.g. disulfide switch, thiol redox signaling) and (d) effective anticipatory blocking control. We then sought the design principles (relationships among rate constants and species concentrations) that warrant satisfaction of all the criteria. These were as follows, for human erythrocytes: (i) the equilibrium constant for thiol-disulfide exchange between thioredoxin and the protein dithiol [T(SH)2 + PSS →TSS + P(SH)2] should be in the range $0.03 < K < 21$ to allow the protein to fully accumulate in the oxidized form as soon as the H2O2 concentration increases; (ii) the maximum flux of Thioredoxin reduction must be lower than the maximum flux of Peroxiredoxindisulfide reduction and formation. Additionally, we identified a trade-off between the robustness of signal transduction and the NADPH expenditure in the process. Human erythrocytes have a limited capacity for NADPH regeneration, and should thus sacrifice the former performance criteria to some extent in order to save NADPH.

A comparison of experimental data to the theoretical predictions above indicates that the design of the PTRTD system in human erythrocytes accomplishes effective integration between anticipatory blocking, antioxidant protection and redox signaling.

We acknowledge fellowship SFRH/BD/51576/2011 to GS and grant and grants PEst-C/SAU/LA0001/2013-2014, FCOMP-01-0124-FEDER-020978 financed by FEDER through the “ProgramaOperacionalFactores de Competitividade, COMPETE” and by national funds through “FCT, Fundação para a Ciência e a Tecnologia” (project PTDC/QUI-BIQ/119657/2010) PEST (LA).

Computational metagenomic techniques applied to the prospection of genes linked with hydrocarbon biodegradation and surfactant production

Miguel E. Coimbra 1(*), Jorge S. Oliveira 1, Luís M. S. Russo 1, Ana Tereza Vasconcelos 2, Lucymara F. A. Lima 3, AnaTeresa Freitas 1

(*)Correspondence: miguel.e.coimbra@ist.utl.pt

1 - INESC-ID/IST - Instituto de Engenharia e de Sistemas de Computadores/Instituto Superior Técnico

2 - Laboratório Nacional de Computação Científica Petrópolis, RJ, Brazil

3 - Universidade Federal do Rio Grande do Norte, Natal, RN

Abstract

This work is a coordinated multidisciplinary effort aimed at developing tools and procedures to identify:

(a) diversity patterns of taxonomy and microbial communities; (b) novel genes linked with hydrocarbon degradation and surfactant production, with a potential for bioremediation strategies and recuperation of mature petroliferous fields.

Metagenomics, a field that studies the DNA contained in an ecosystem, has developed rapidly in recent years. This was due to improvements and cost reductions of DNA sequencing technologies. Despite these positive developments, this field still faces serious computational challenges, manifested by the complexity and large size of the processed data. Several tools have been developed to face these difficulties, allowing for the identification of genes associated with the production of enzymes with diverse functions. Certain enzymes, in particular, are related to hydrocarbon degradation, a process during which there is production of adjuvant molecules, also called biosurfactants. These are amphipathic molecules, which play a role in several industries, such as petrochemistry, by facilitating the emulsion of liquids with different polarities.

The economic potential of applying surfactant-producing microorganisms to petroliferous reserves resides in the employment of the so called tertiary crude oil recovery techniques. By using compounds, namely surfactants, which improve the mobility of crude, up to 10% extra petrol may be recovered [1].

To respond to the need of optimized computational capabilities for identifying the specific type of genes herein described, we developed the MetaGen-FRAME framework [2], which incorporates a set of state-of-the-art software packages.

Analysis of five metagenomes from petrol drill residues was performed using this framework. Preliminary results allowed for the finding of some pathways relevant for biodegradation, notwithstanding the necessity of obtaining more sequencing data to enable the assembly of larger portions of the most abundant organisms in the analyzed environments.

ACKNOWLEDGEMENTS

This work was supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/-LA0021/2013.

REFERENCES

[1] R. Marchant, I. M. Banat, Microbial biosurfactants: challenges and opportunities for future exploitation, Trends in Biotechnology, Volume 30, Issue 11, November 2012, Pages 558-565, ISSN 0167-7799.

[2] Miguel C. V. E. Coimbra, MetaGen-FRAME: Metagenomics Data Analysis Framework Focused on Stressed Microbial Communities, MSc Thesis, Instituto Superior Técnico, Technical University of Lisbon, unpublished.

LLC-GNUMAP: Scalable, Precise, and High-Coverage Genomics Mapping

Natacha P. Leitão 1, João Leitão 2, and Francisco M. Couto 1

1 - LaSIGE, Department of Informatics, Faculty of Sciences
University of Lisbon

2 - CITI, Department of Informatics
Faculty of Sciences and Technology, NOVA University of Lisbon

Abstract

Next generation sequencing (NGS) technologies revolutionized biomedical research by bringing the promise of enabling the production of large amounts of sequence data at relatively low cost in a short time. NGS technologies are now being applied to a growing number of biological applications, most of which rely on accurate and fast read mapping mechanisms (i.e., aligning reads to a reference genome) in which a fundamental step is the ability to distinguish between technical sequencing errors and natural genetic variation. Many mapping tools have been developed, such as Bowtie[2], BWA[3], MAQ[4], and SHRiMP[5]. Each of these solutions rely on different approaches that have inherent limitations, either in terms of scalability (the time required to execute the mapping), coverage (the percentage of reads that are effectively mapped), and precision (mapping reads to the correct location in the reference genome). In particular, coverage is an extremely challenging aspect, as when a read occurs at multiple locations of the genome, finding the exact location from which it came is nearly impossible, and many existing solutions simply discard those reads. A key recent improvement in NGS technologies which might enable addressing the limited coverage of existing solutions is the quality information which is now included in every run; each position in a read correspond to one of the four bases, the quality score assigned to each position indicates the probability of the reported base in that position being incorrect. In fact, this information so far is being used by some solutions [2, 3, 4] through different approaches that focus in the probability of the called base being incorrect. A recent solution, named GNUMAP[1], has relied on this information to devise a rigorous probabilistic approach that uses the probabilities of all four bases being called. Therefore, this solution enables mapping of reads that occur in several regions, with a high coverage of approximately 70% albeit with no reported precision. In this manuscript, we propose a new mapping GNUMAP-base algorithm, dubbed LLC-GNUMAP, witch offers the potential not only to improve the coverage of GNUMAP, but also to improve its precision. At the core of LLC-GNUMAP lies a novel technique which increases the search space over the reference genome for each individual read, by using a hash-based approach that leverages quality information and biological constraints. As we increase the search space for each read, we designed our algorithm to be highly parallelized, allowing executions to span across an arbitrary number of machines - for instance in a cloud computing platform - with near linear scalability. Additionally, we propose a methodology to measure the precision of mapping algorithms based on artificial read sets, which are constructed to emulate common errors found in NGS datasets.

Acknowledgements

The authors would like to thank AlyssonBessani and Vinícius Cogo for the hardware provided, their availability and technical support. This work was financially supported by FCT through funding of LaSIGE Strategic Project, ref. PEst-OE/EEI/UI0408/2014.

References

[1] N. L. Clement, Q. Snell, M. J. Clement, P. C. Hollenhorst, J. Purwar, B. J. Graves, B. R. Cairns, and W. E. Johnson. The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. Bioinformatics, 26(1):38–45, 2010.

[2] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. GenomeBiol., 10:R25, 2009.

[3] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14):1754–1760, 2009.

[4] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. GenomeRes., 18:1851–1858, 2008.

[5] S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno. SHRiMP: Accurate Mapping of Short Color-space Reads. PLoSComput. Biol., 5(5):e1000386, 2009.

Corresponding Author

Natacha P. Leitao
Address: LaSIGE, Faculdade de Ciências da Universidade de Lisboa, Departamento de Informatica, Edifício C6 Piso 3, Campo Grande 1749 - 016 Lisboa.
E-mail: nleitao@lasige.di.fc.ul.pt.

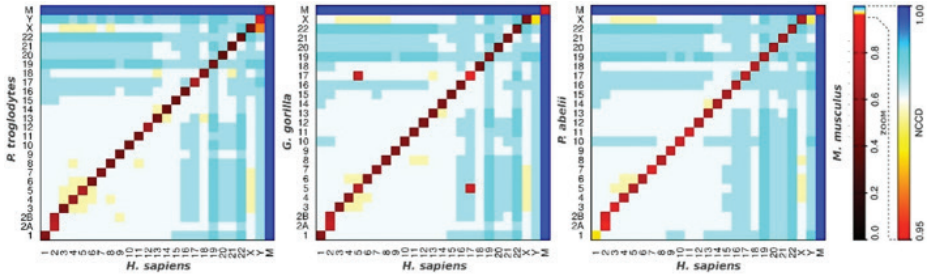
Insights into genomic evolution by sequence smashing

Diogo Pratas and Armando J. Pinho
IEETA / Dept of Electronics, Telecommunications and Informatics
Universityof Aveiro, 3810 –193 Aveiro, Portugal
pratas@ua.pt — ap@ua.pt

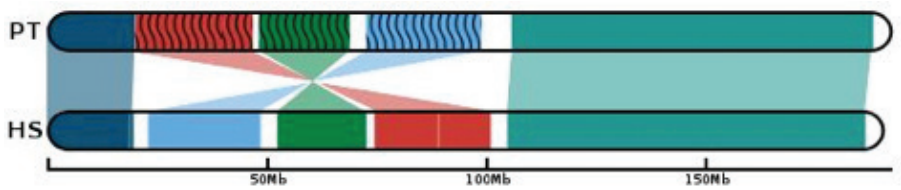
Abstract

Knowing how genomic sequences are organized and which types of re-organizations are implicated in speciation and macroevolutionary events along the time is funda-mental to understand the dynamics of chromosomal evolution. Common biological approaches to unveil these events (e.g., FISH) are very time-consuming and expensive, in contrast to computational solutions, such as recently proposed com-pressed-based metrics. The quantification of the common information between genomic sequences unveils how much they are similar and, when extended, where. Although these mea-sures are non-computa-ble, they can be approximated using a data compressor, and the bits required to compress a sequence, using exclusively the knowledge of another sequence, gives an upper bound on their mutual information.

We describe the foundations of these metrics and apply them in genomic anthro-pology, studying large-scale chromosomal rearrange-ments, such as inversions, translo-cations and fusions, between humans and the great apes. Accordingly, we measure distances between chromosomes of different species as a way to detect inter-species ho-mologies (using comparative normalized absolute measures). The figure below shows the information heatmaps between human and great apes.



Also, we show where the large-scale structures diverge between primate species. For that purpose, we use information maps, as can be seen in the figure below, between human and chimpanzee chromosomes 5.



Work supported in part by National Funds through FCT - Foundation for Science and Technology, in the context of the project PEst-OE/EEI/UI0127/2014, and by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 305444 "RD-Connect: An integrated platform connecting registries, biobanks and clinical bioinformatics for rare disease research".

Development of a semantic model for stroke patients

Lara Silva 1, Astrid Vicente 2, and Francisco Couto 1
1 - LaSIGE, Department of Informatics, Faculty of Sciences University of Lisbon
2 - DPSPDNT, Departamento de Promoção da Saúde e Prevenção de Doenças Não Transmissíveis
Instituto Nacional de Saúde Dr.Ricardo Jorge, Lisbon

Abstract

In current times, the amount of acquired data is quite superior to our human processing ability. However, in order to have computers aid researchers in this task, the information must be not only in a computer readable format, but standardized in case the data ranges different knowledge areas (like the biomedical field). These can be provided by the use of semantic web technologies. The merits of it's application in the biomedical field has already been observed by research groups [2, 3].

Besides being one of the most common death causes in Portugal, stroke is a leading cause of significant disability as well. 15 to 30% of stroke patients are permanently disabled and 20% require institutional care at 3 months after onset. Some clinical, demographic and genetic factors are known to be associated to stroke recovery processes [4], but there's still need for further research in this area.

The objective of this project is the creation of a semantic model for stroke that allows the integration of clinical and genetic data and the automated interconnection with outside resources (like SNOMED CT) in order to evaluate their individual and combined weight in the patient recovery score. This project uses graph databases, due to their flexibility, considering a possible future studies of stroke patient data without making the present data obsolete.

For this task we selected Protégé-OWL[1] for the semantic modelling, and Neo4J as the graph database engine. This database will be popu-

lated with genetic, clinical and demographic data from 416 stroke patients.

Acknowledgements

This work was financially supported by FCT through funding of LaSIGE Strategic Project, ref. PEst-OE/EEI/UI0408/2014.

References

[1] H. Knublauch, O. Dameron, and M. Musen. Weaving the Biomedical Semantic Web with the Protégé OWL Plugin, 2004.

[2] C. Machado, A. Freitas, and F. Couto. Enrichment analysis applied to disease prognosis. Ontologies in BiomedicineandLifeSciences, 4(21):1–26, 2012.

[3] C. Machado, D. Rebholz-Schuhmann, A. Freitas, and F. Couto. The semantic web in translational medicine: current applications and future directions. Briefings in Bioinformatics, pages1–15, 2013.

[4] S. C. Cramer, V. Procaccio and for the GAIN Americas and GAIN International Study Investigators. Correlation between genetic polymorphisms and stroke recovery. European Journal of Neurology, 2012.

Corresponding Author

Lara Silva
Address: LASIGE, Faculdade de Ciencias da Universidade de Lisboa, Departamento de informatica, Edifício C6 Piso 3, Campo Grande 1749 - 016 Lisboa.
E-mail: lara.ps.silva@gmail.com.

Bioinformatic challenges in human genome and exome analysis

Hugo J.C. Froufe, Xin Wang, Susana Carmona, Conceição Egas
Genoinseq, BiocantPark, 3060-197 Cantanhede, Portugal.

Abstract

Massively parallel sequencing has been proven revolutionary, shifting the paradigm of genomics to address biological questions at a genome-wide scale [1]. It is now possible to determine the individual genome or exome sequences for a large number of individuals, and the potential of this information for new discoveries in fields such as medicine is enormous [2]. As with all new, breaking technology, there are challenges to be overtaken. The amount of information produced by the new next generation sequencers is huge, and this is a problem to be addressed by bioinformaticians. In addition, the analysis of this information has to be automated and requires a large computational capacity since millions of sequences must be processed.

Many steps of the human genome and exome analysis and corresponding bioinformatics challenges will be addressed in this communication. One of the early steps is the mapping, were sequence reads are aligned against the reference human genome. However, this genome contains a number of DNA sequencing errors (0.01%) [3], which lead to incorrectly aligned regions. Furthermore, most mapping programs were implemented on heuristic algorithms to reduce processing time, but also decreasing accuracy and creating small imprecision in the alignment. This small imprecision will take a large effect in the next step: the variant calling.

In general, variants such as SNPs (single nucleotide polymorphisms) and small Indels (insertions and deletions) are relatively easy to identify. As alignment is not perfect, current methods have a false positive detection rate around 1% to 5% for SNPs and 10% to 30% for small Indels, depending in the software, sequencer platform and testing method. Therefore is vital to improve or create new mapping and variant calling methods to enhance the detection rate.

When variants are annotated, filtered and prioritized according to the study objectives, it is essential to use or create computational environments that can cope with massive datasets, with dozens or hundreds Gb of information, and assess the most likely impact of observed variants. These correlations need to be considered along with statistical association evidence in order to prioritize variants in terms of likelihood of influencing interest traits. Current platforms for variant annotation, filtering and prioritization, have several limitations, requiring additional development to respond to current needs.

Addressing these problems represents a challenge to bioinformatics working in –omics and solving them will promote new advances and discoveries in science.

[1] Koboldt D. C., Steinberg K. M., Larson D. E., Wilson R. K., Mardis E. R. The Next-Generation Sequencing Revolution and Its Impact on Genomics. Cell. 2010; 155(1): 27-38,

[2] Snyder M., Du J., Gerstein M.. Personal genome sequencing: current approaches and challenges. Genes Dev. 2010; 24(5): 423–431.

[3] Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Computational prediction of microRNA targets in plant genomes

Manuel Reis 1;2, Nuno D. Mendes 2;3, and Ana T. Freitas 1;2

1 - IST/UL, Av. Rovisco Pais, 1, 1049-001 Lisboa - Portugal
2 - KDBIO/INESC-ID, Rua Alves Redol, 9
1000-029 Lisboa – Portugal
3 - IBET, Av. Republica, Qta. do Marques
2780-157 Oeiras - Portugal

Abstract

MicroRNAs (miRNAs) are important posttranscriptional regulators and act by recognizing and binding to sites in their target messenger RNAs (mR-NAs). They are present in nearly all eukaryotes, in particular in plants, where they play important roles in developmental and stress response processes by targeting mRNAs for cleavage or translational repression. MiRNAs have been shown to have a crucial role in gene expression regulation, but so far only a few miRNA targets in plants have been experimentally validated. Based on the number of annotated genes, on the number of experimentally validated miRNAs and on the fact that one miRNA often regulates multiple genes, a long list of yet unidentied targets is to be expected. With this work, we evaluate two existing target prediction tools for plants - TAPIR and psRNATarget - and compare the results against validated targets by analysing their performance. We tested these tools using datasets from Arabidopsis thaliana and Oryzasativa. The transcriptome data was obtained from Phytozome (v9.0), and the miRNAs mature sequences from miRBase (release 20). Given the vast number of predictions that existing tools output, most of which are most likely false positives, we also introduce our tentative approach based on evolu-tionary properties and compare the results with both tools. This approach is based on the hypothesis that a transcript may show evidence of exhibiting a sequence bias towards either eliciting or avoiding target sites for a particular miRNA, if both are transcribed in an overlapping set of biological conditions. This bias is calculated as the log-odds score of

the probability of a Markov chain, which is trained over the sequence of a given transcript, generating the perfectly complementary sequence to a given miRNA, against the background probability of it occurring for a Markov chain modelling the entire set of transcripts. The evolutionary cri-terion could prove important when mitigating false positive predictions and thus reducing the large number of potential targets, facilitating downstream prediction by other methods. Additionally, we can also explore the concept of anti-targets as a complementary way to perform the functional annotation of miRNAs. Knowing that the transcript sequences that show evidence of clear target avoidance bias are involved in a certain biological process may indicate that co-expressed miRNAs may be participating in the same processes.

Keywords: plant microRNA target prediction, alignment-free sequence analysis Corresponding author: mreis@kdbio.inesc-id.pt.

Acknowledgments: This work was supported by national funds through FCT - Fundacão para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013 and CAMP: Computational Analysis of MicroRNA in Plants (PTDC/EIAEIA/122534/2010).

Predicting relevant events in the life of a person with Alzheimer

José Carilho 1, Oliver Schnell 1, JoãoMartins 1, Luís Carriço 1
Carlos Duarte 1, Francisco M Couto 1, and Tiago Guerreiro 1
1 - LaSIGE, Department of Informatics, Faculty of Sciences
University of Lisbon

Abstract

The number of people who are suffering from dementia and its most common form Alzheimer, increased significantly during the last years. One of the consequences of the Alzheimer's disease is the social isolation. The patients start to avoid their friends and family. With our project we aim to maintain the patients social active. Our approach is based on the automatic gathering of relevant contextual information in the patient's smartphone, such as, GPS indoor and outdoor location, pictures the patient explicitly takes and activity recognition through the use of the accelerometer, and enhance the creation capabilities of assistive technologies. The data collected from the smartphone will be used as an input for the inference motor. This motor will use machine learning techniques to build classifiers trained with previously curated data from the patient. The classifiers will aim at identifying life events with new data collected from the smartphone. As these events are detected by the inference motor, they are stored on a semantic network. The system will monitor the knowledge base and try to complete missing information by using free available online sources such as Google Maps or Wikipedia. Hereby, the user's location during the day can be described more detailed including information about the type of the place, the address or pictures of the visited haunts. However, since some of this very personal information can only be understood within a context and is in most cases only known by the patient's social contacts, a validation of the automatically retrieved data is required. Therefore, questions are posted on a social network site to ensure the retrieved data quality and moreover invite friends to enrich the won knowledge with additional media and more detailed information. This

will provide the system with the power, together with the inference motor, to detect routines in the patient life. A routine can be seen as a sequence of life events. Defining a routine as a sequence of events provides the system with the capacity to identify regularities on the patient routine, or if he goes outside his routine alert the caregivers.

Acknowledgements

This work was financially supported by FCT through funding of LaSIGE Strategic Project, ref. PEst-OE/EEI/UI0408/2014.

kluskens@deb.uminho.pt
lrmr@deb.uminho.pt

Identifying Interactions Between Chemical Entities in Text

Andre Lamurias and Francisco M. Couto
LaSIGE, Department of Informatics, Faculty of Sciences, University of Lisbon

Abstract

We developed Identifying Chemical Entities (ICE 1.0) which recognizes the chemical named entities mentioned in a given text and performs resolution of each entity to the Chemical Entities of Biological Interest [2](ChEBI) ontology [3].

Since interactions between chemical entities are also frequently reported in biomedical texts, we aim at developing a new method (ICE 2.0) capable of identifying interactions between entities in the same text, as proposed by the task 9.2 of SemEval 2013 [6]. For this task, an interaction is defined as when one chemical compound or drug influences the level or activity of another. We intend to adapt some of the approaches used for this competition, developing a new method that used semantic similarity for identifying and validating interactions.

The chemical entities recognized and mapped by ICE 1.0 would be the starting point for this new method. Then, the semantic similarity between each pair can be calculated to verify if both entities participate in a chemical process. We have implemented three semantic similarity measures in our system that can be calculated between every two terms of the ChEBI ontology. These semantic similarity measures we have implement are Resnik's similarity [5], simUI[1] and simGIC [4].

We propose a new method that uses semantic similarity, along with lexical features extracted from the original text, to train classifiers to predict if a pair of ChEBI terms represents an interaction. Our assumption is that two ChEBI terms that participate in an interaction have also a semantic relation defined in the ontology. We intend to make this method accessible from a web tool and web service.

Acknowledgements

This work was financially supported by FCT through the financial support of the SPNet project (PTDC/EBB-EBI/113824/2009) and through funding of LaSIGE Strategic Project, ref. PEst-OE/EEI/UI0408/2014.

References

[1] R. Gentleman. Visualizing and distances using GO. URL http://www. bioconductor. org/docs/vignettes. html, 2005.

[2] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleicacids research, 41(D1):D456–D463, 2013.

[3] A. Lamurias, T. Grego, and F. M. Couto. Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI. In BioCreativeChallengeEvaluation Workshop vol. 2, volume 489, page 75, 2013.

[4] C. Pesquita, D. Faria, H. Bastos, A. Falcao, and F. Couto. Evaluating GO-based semantic similarity measures. In Proc. 10th Annual-Bio-Ontologies Meeting, pages 37–40, 2007.

[5] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007, 1995.

[6] I. Segura-Bedmar, P. Martinez, and M. Herrero-Zazo. 2013 SemEval-2013 Task 9: Extraction of drug-drug interactions from biomed-ical texts. In In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval. Citeseer, 2013).

Corresponding Author

Andre Lamurias
Address: LASIGE, Faculdade de Ciências da Universidade de Lisboa, Departamento de informatica, Edifício C6 Piso 3, Campo Grande 1749 - 016 Lisboa
E-mail: alamurias@lasige.fc.ul.pt.

Organizing principles underlying microorganism's growth-robustness trade-off

Alessandro Bolli 1, Armindo Salvador 1,2
1 - Computational Systems Biology Group, Center for Neuroscience and Cell Biology, University of Coimbra;
2 - Chemistry Department, University of Coimbra, Portugal
Corresponding author: aless.bolli@gmail.com

Abstract

Microorganisms show some conserved relations between physiological state and environmental conditions. Growth Robustness Reciprocity (GRR) is an intriguing example: genetic and environmental factors that impair microorganism's vegetative performance (growth rate) enhance their ability to resist abiotic stresses, and vice-versa. Mechanistically, this relationship may be explained by regulatory interactions that determine higher expression of protection mechanisms in response to low growth rates. However, because the mechanisms themselves are not conserved between bacteria and eukaryotes, the observed GRR must result from convergent evolution. Why does natural selection favor such an outcome? We used mathematical models of optimal resource allocation in an idealized cellular self-replicating system to identify the key cellular functions and general evolutionary and physiological principles that may explain why GRR is widespread among microorganisms. These models account for the cell protein components involved in: (i) substrate uptake from environment, (ii) metabolic transformation of substrate to anabolic precursors, (iii) biosynthetic enzymes (e.g. ribosomes and lipid synthesizing enzymes), (iv) protein inactivation, representing stress, and (v) protein repair (REP). The relative fraction of each cell component is computed by adjusting the proportion of ribosomes engaged in the synthesis of each component. The protein pool proportions that maximize growth rate were computed through non-linear optimization. In the absence of the REP component, the cell model displays substrate concentra-

tion-dependent stress sensitivity: cell stasis (no growth) is reached at lower stress levels for proportionally lower substrate concentrations (i.e., lower initial growth rate). This is due to growth-related damage dilution: the higher the substrate availability, the highest the growth rate, the fastest the dilution of damaged proteins by newly synthesized proteins, the highest the stress that can be tolerated until the inactive pool retains all the cell resources necessary for growth. When the REP component is considered, resource optimization favors its higher expression at lower substrate availabilities. Consequently, maximal stress protection (i.e. higher REP-induced growth recovery) is obtained at low nutrient conditions where growth-related damage dilution is less effective. Overall, these results show that GRR can be explained by the interplay among three general principles. Namely, (a) natural selection for maximal growth rates, (b) inevitability of damage or errors, inactivating cellular components and thus decreasing growth, and (c) damage dilution by the growth-associated synthesis of new components.

We acknowledge fellowship SFRH/BPD/90065/2012 and grants PEst-C/SAU/LA0001/2013-2014,PEst-OE/QUI/UI0612/2013, FCOMP-01-0124-FEDER-020978 financed by FEDER through the "Programa Operacional Factores de Competitividade, COMPETE" and by national funds through "FCT, Fundação para a Ciência e a Tecnologia" (project PTDC/QUI-BIQ/119657/2010).

Image processing on animal cell cultures: a refined technique

Mariana Xavier, Daniela P. Mesquita, António L. Amaral, Eugénio C. Ferreira, Lígia R. Rodrigues and Leon D. Kluskens*

kluskens@deb.uminho.pt

Centre of Biological Engineering, Universidade do Minho
Campus de Gualtar, 4710-057 Braga, Portugal

Abstract

The process of microscopic animal cell counting can be a time-consuming process, resulting in a subjective analysis varying according to the researcher's perception. Regarding the ideal moment to divide the cells, the decision is performed in an empirical manner and is affected by the complexity of cell morphology and density. Searching for a way to overcome these problems, and considering the decreasing costs of computational data processing, a window was found for new methodologies to quickly characterize a given structure.

Advances in digital imaging allow the extraction of quantitative information, opposite to the qualitative and subjective evaluation of human analysis. Thus, microscopy image analysis techniques have gained, during the last years, an unquestionable role in several fields of research. The purpose of an image processing step resides in obtaining a final image holding significant information for a given application. These techniques should be automated as much as possible to avoid subjectivity. Thus, several segmentation techniques have been already proposed. For segmentation to take place, usually a threshold value(s) must be defined to allow the differentiation between the objects and background. Other methods, such as region growing, mathematical morphology and watershed are also used for this purpose. These are simple algorithms that when appropriately used can provide promising results and oftentimes with a low computation complexity. Nevertheless, the previous methods have some limitations, including non-uniform intensity variations, low-contrast images, irregular segmentation and over-segmentation. More sophisticated methods based on frame-

works of active contours (e.g. snakes, level-sets) or graph-cuts can also be applied to segment cells with positive results. Nonetheless, these algorithms present high computational complexity.

The main goal of this work was to develop an image processing tool using several algorithms in order to improve cell segmentation processing for different morphological cells and densities. For that purpose, different cells were used – MDA-MB-231 and -435, both cancer cell lines, and MCF-10-2A, a non-tumorigenic line. Cells were observed in a Leica DM IL inverted contrasting microscope, in phase-contrast at 100x total magnification, coupled with a Leica D-LUX 3 camera, ensuring the same acquisition conditions. Despite the variability in their morphology, preliminary results demonstrated that the segmentation process was fairly successfully. As a result, the previously described flaws were minimized, leading to more efficient animal cell culturing with less variability.

All authors contact details

mariana.nx91@gmail.com
{daniela, lpamaral, ecferreira, lrmr, kluskens}@deb.uminho.pt

Targeted therapy using phage technology: A computational and experimental breast cancer study

Franklin L. Nobrega*, Tânia S. Mendes, Mariana Xavier, Vera L. Silva, Joana Azeredo, Lígia R. Rodrigues and Leon D. Kluskens*

franklin.nobrega@ceb.uminho.pt / kluskens@deb.uminho.pt

Centre of Biological Engineering, Universidade do Minho
Campus de Gualtar, 4710-057 Braga, Portugal

Abstract

During the past two decades cancer biology knowledge has widely increased and shifted the paradigm of cancer treatment from nonspecific cytotoxic agents to selective, mechanism-based therapeutics. Initially, cancer drug design was focused on compounds that rapidly killed dividing cells. Though still used as the backbone of current treatments, these highly unspecific targeting drugs lead to significant toxicity for patients, narrowing the therapeutic index, and frequently lead to drug resistance. Therefore, cancer therapies are now based on cancer immunotherapy and targeted agents, whereas novel treatments are strategically combining both to improve clinical outcomes.

Despite the nanotechnology advances dictating the development of targeted therapies in diverse classes of nano-based carriers, virus-based vectors still remain highly used due to its biocompatibility and specificity for the target.

Particularly, bacteriophages are an interesting alternative 'nanomedicine' that can combine biological and chemical components into the same drug delivery system. The great potential of this novel platform for cancer therapy is the ability to genetically manipulate the virus-vector to display specific targeting moieties.

Phage display technology, a general technique used for detecting interfaces of various types of interacting proteins outside of the immunological context, allows the target agents to locate the target (with an increased selection process for the specific binding– termed biopan-

ning) and play their essential role inhibiting molecular pathways crucial for tumour growth and maintenance. Phage display specificity core is related with the binding of small peptides displayed at their coat or capsid proteins, enriched during biopanning. Bioinformatics plays an important role in testing and improving phage display libraries by effective epitope mapping, selecting from a large set of random peptides those with a high binding affinity to a target of interest.

In this work we demonstrate the screening of a manually constructed 7-mer peptide library of M13KE phage particles against MDA-MB-231 and -435 cancer cell lines. Two peptides – TLATVEV and PRLNVSP – with high affinity for the referred cells were identified, respectively. Based on computationally predicted epitopes based on the peptides extracted from this library the linear peptide sequence was docked onto known membrane proteins from the used cell lines and peptides-proteins interactions were mapped. Umbrella sampling studies were performed to predict the binding affinity and to improve future rational design of binding peptides to these cancer cells.

All authors contact details:

franklin.nobrega@ceb.uminho.pt
taniamendes@ceb.uminho.pt
mariana.nx91@gmail.com
vera.7.silva@gmail.com
jazeredo@deb.uminho.pt
lrmr@deb.uminho.pt
kluskens@deb.uminho.pt

open days