BIOINFORMATICS
OPEN DAYS

Detailed Program

Braga

1st and 2nd March, 2012

Universidade do Minho

Contents

Scientific Committee

- Miguel Rocha - CCTC, University of Minho
- Isabel C. Rocha - IBB/CEB, University of Minho

Organizing Committee

- Alberto Noronha
- André Santos
- Carina Lourenço
- Edilana Gomes
- Emanuel Gonçalves
- Inês Martins
- João Cardoso
- João Saraiva
- José Fernandes
- Rita Vilaça
- Renato Ribeiro
- Sara Fonseca
- Sofia Fernandes

Short Program

March 1st

09:30-10:00	Registration
10:00-10:30	Opening Session: Isabel Rocha; Manuel Mota (CEB-UM/IBB); João Luis Sobral (CCTC-UM)
10:30-11:00	*New tools for in silico metabolic engineering*, Isabel Rocha
11:00-11:30	Coffee Break / Posters
11:30-12:00	*In silico screening of transthyretin amyloid inhibitors*, Rui Brito
12:00-12:30	*Computational Phylogenomics*, David Posada
12:30-13:00	*Structural bioinformatics towards the understanding of molecules and life*, Cláudio Soares
13:00-14:30	Lunch
14:30-15:00	*Exploring the Semantic of Biomedical Ontologies*, Francisco Couto
15:00-15:30	*Multi-Relational Learning for the Life Sciences*, Rui Camacho
15:30-16:00	*Challenges and solutions on genotype-to-phenotype integration,* José Luís Oliveira
16:00-16:30	Coffee Break
16:30-17:00	Forum*: Experiences and dialogues in Portuguese Bioinformatics education*

March 2nd

10:00-11:00 Poster Session / Breakfast

1. **Unraveling the molecular mechanisms of ABC transporters: Insights from molecular dynamics simulations**. A. Sofia F. Oliveira, António M. Baptista, and Cláudio M. Soares
2. **HOXA9 transcriptome analysis in Glioblastoma: The bioinformatics beyond the bench.** Ana Xavier-Magalhães, Céline S. Gonçalves, Miguel Rocha, Bruno M. Costa
3. **Data Mining Strategies for the Analysis of Protein Folding and Unfolding Simulation Data.** Cândida G. Silva, Pedro Gabriel Ferreira, Paulo J. Azevedo, Rui M. M. Brito
4. **Reconstruction of a genome-scale metabolic model for the filamentous fungus Ashbya gossypii.** Daniel Gomes, Oscar Dias, Eugénio Ferreira, Lucília Domingues, Isabel Rocha
5. **Comparison of H. pylori in silico metabolic model predictions with experimental data.** D.M. Correia, M.L.R. Cunha, N.F. Azevedo, M.J. Vieira, I. Rocha
6. **A Computational Modeling Study on African Swine Fever Virus encoded Proteins with Predicted Structural Repeats**. David Henriques
7. **Efficient high-throughput read mapping for re-sequencing applications**. Francisco Fernandes, Paulo G.S. da Fonseca, Luis M.S. Russo, Arlindo L. Oliveira and Ana T. Freitas
8. **Bacteriophage phylogeny revisited**. Franklin L. Nóbrega, Graça Pinto, Joana Azeredo and Leon D. Kluskens
9. **Insights into phage endolysins.** Franklin L. Nóbrega, Hugo Oliveira, Luís D. R. Melo, Sílvio B. Santos, Nuno Cerca
10. **Extraction and Characterization of Biologically Relevant Relations in Biomedical Literature.** Hugo Costa, Isabel Rocha e Anália Lourenço
11. **Semantic Similarity in the Biomedical Domain**. João D. Ferreira and Francisco M. Couto
12. **Highlighting Metabolic Strategies using Network Analysis over Strain Optimization Results**. José Pedro Pinto, Isabel Rocha, and Miguel Rocha
13. **Dynamic Programming Algorithms for Biobjective Sequence Alignment.** Maryam Abbasi, Luís Paquete, Miguel Pinheiro
14. **Detection of change points in mitochondrial DNA**. Nora M. Villanueva, Miguel M. Fonseca, Marta Sestelo and Javier Roca-Pardiñas

15. **Genome-Wide Analysis of Single Nucleotide Substitutions in Mycobacterium tuberculosis.** Osório N.S., Saraiva M.S., Pedrosa J., Castro A.G., Rodrigues F.
16. **Assembling genome-wide transporter system annotations.** Oscar Dias, Miguel Rocha, Eugénio Ferreira, Isabel Rocha
17. **OPTFERM – A Computational Platform for the Optimization of Fermentation Processes.** Orlando Rocha, Paulo Maia, Isabel Rocha, Miguel Rocha
18. **Automating, Gathering and Manipulating Genetic Information at the Sequence Level.** Paulo Gaspar, José Luís Oliveira
19. **A Generic Multi-Criterion Approach for Mutant Strain Optimization.** Paulo Maia, Isabel Rocha and Miguel Rocha
20. **In silico metabolic Engineering tool for simulation and optimization of microbial strains accounting integrated metabolic/regulatory information.** Paulo Vilaça, Miguel Rocha and Isabel Rocha
21. **COEUS: Semantic Web Application Framework**. Pedro Lopes and José Luís Oliveira
22. **C-based Metabolic Flux Analysis.** Carreira R., Carneiro S., Villas-boas S., Rocha I. and Rocha M.
23. **Search for coherent gene clusters that predict invasiveness of Streptococcus pneumoniae strains**. Rui Catarino, Sandra Aguiar, Mário Ramirez, Francisco Pinto
24. **In silico optimization of the production of amino-acids in Escherichia coli**. Rui Pereira, Paulo Vilaça, Miguel Rocha, Isabel Rocha
25. **Computational tools for strain optimization by adding reactions**. S. Correia, M. Rocha
26. **Application of Machine Learning techniques to the discovery and annotation of Transposons in genomes**. Tiago Loureiro, Nuno A. Fonseca, Rui Camacho
27. **Parallelizing SuperFine**. Diogo Neves
28. **Prognostic prediction using clinical expression time series: towards a supervised learning approach based on meta-biclusters**. André V. Carreiro, Artur J. Ferreira, Mário A. T. Figueiredo and Sara C. Madeira
29. **In silico screening of transthyretin amyloid inhibitors**. Carlos J. V. Simões, Catarina S. H. Jesus, Rui M. M. Brito

11:00-11:30 *Relating design to function in cellular protection against hydrogen peroxide,* Armindo Salvador

11:30-12:00 *Translating bioinformatics to human health,* José Leal

12:00-13:00 Selected abstracts

30. **A Computational Modeling Study on African Swine Fever Virus encoded Proteins with Predicted Structural Repeats**. Elsa S. Henriques and Rui M. M. Brito (oral presentation)
31. **Recognition and resolution of chemical entities to ChEBI**. Tiago Grego, Francisco M. Couto (oral presentation)
32. **Prognostic prediction using clinical expression time series: towards a supervised learning approach based on meta-biclusters**. André V. Carreiro, Artur J. Ferreira, Mário A. T. Figueiredo and Sara C. Madeira (oral presentation)

13:00-14:30 Lunch

14:30-15:00 *SilicoLife,* Simão Soares

15:00-15:30 *TAPYR: An efficient high-throughput sequence aligner for re-sequencing applications,* Ana Teresa Freitas

15:30-16:00 *A look on the sweet side of life – through genome-scale metabolic models,* Kiran Patil

Invited Speakers

A look on the sweet side of life – through genome-scale metabolic models

Kiran Raosaheb Patil

Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidleberg.

Abstract

Availability of plant-based carbohydrates (esp. sugars) presents a challenging task for biochemical engineering – developing microbial cell factories that can economically convert these renewable resources into molecules of socio-economic relevance. System-level analysis and modeling of microbial metabolic networks can significantly aid towards this goal and thereby contribute for creating a sustainable chemical industry. To this end, system-wide metabolic modeling tools based on evolutionary programming will be presented along with few examples of successful application to *in vivo* systems. The examples include production of Vanillin in baker's yeast with glucose as a carbon source. Understanding and modulating sugar metabolism is of paramount importance not only in industrial biotechnology, but also for fighting metabolic diseases. Indeed, excessive sugar intake (among others) has been identified as a major risk factor for Type 2 diabetes mellitus (T2DM) – one of the main threats to human health in the 21st century. At the level of gene expression, multiple metabolic pathways are dysregulated in diabetes and in individuals at risk for diabetes; which of these pathways are central to disease pathogenesis, remains a key question. Topological and regulatory complexity of cellular metabolic network presents a considerable challenge in pinpointing key molecular mechanisms and biomarkers associated with insulin resistance and type 2 diabetes. This problem can be addressed by using a novel methodology that integrates gene expression data with human cellular metabolic network. The methodology will help in the development of new therapeutic agents and future clinical diagnostics for type 2 diabetes.

References

1. Patil K. R., Rocha I., Forster J. & Nielsen J. Evolutionary programming as a platform for in silico metabolic engineering. ***BMC Bioinformatics*****. 6**, 308 (2005).
2. Zelezniak A., Pers T.H., Soares S., Patti M.E. & Patil K.R. Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes. ***PLoS Computational Biology,*** **6**:4 (2010).

Structural bioinformatics towards the understanding of molecules and life.

Claudio M. Soares

Protein Modelling Laboratory, Instituto de Tecnologia Química e Biológica - Universidade Nova de Lisboa, Oeiras, Portugal

Abstract

An overview of the work developed in the Protein Modelling Laboratory will be presented. This will include a very short description of methodologies used, prior to examples of research pursued in this laboratory. These examples are quite varied and will focus on the more recent research, which will range from electron and proton transfer in biological redox chains, molecular mechanisms in ABC transporters, membrane fusion mechanisms driven by Haemagglutinin from the Influenza Virus and simulation of proteins involved in biotechnologically relevant processes.

Relating design to function in cellular protection against hydrogen peroxide

Armindo Salvador

Abstract

Erythrocytes, like most other cells, carry two types of enzymatic defenses against hydrogen peroxide. Namely, catalases, which dismutate hydrogen peroxide, and peroxidases, which reduce it. The latter process requires a metabolic supply of reducing equivalents, making its operation more expensive than that of the former. Why then don't cells rely solely on catalase for protection? In this talk I will examine this and other aspects of the design of antioxidant protection in human erythocytes.

TAPYR: An efficient high-throughput sequence aligner for re-sequencing applications

Francisco Fernandes[1,2], Paulo G.S. da Fonseca[1], Luis M.S. Russo[1,2], Arlindo L. Oliveira[1,2], Ana T. Freitas[1,2]

[1]Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento (INESC-ID), R. Alves Redol 9, 1000-029 Lisboa, Portugal

[2]Instituto Superior Técnico/Universidade Técnica de Lisboa (IST/UTL), Av. Rovisco Pais, 1049-001 Lisboa, Portugal

Email: fjdf@kdbio.inesc-id.pt; pgsf@kdbio.inesc-id.pt; lsr@kdbio.inesc-id.pt; aml@kdbio.inesc-id.pt; atf@kdbio.inesc-id.pt;

Abstract

During the last two decades most laboratories used Sanger's "shotgun" method in many significant large-scale sequencing projects, being this method considered the 'gold standard' in terms of both read length and sequencing accuracy. Recently, several next generation sequencing (NGS) technologies have emerged, including the GS FLX (454) Genome Analyzer, the Illumina's Solexa 1G Sequencer, the SOLiDTM and the Ion Torrent Systems, which are able to generate three to four orders of magnitude more sequences and are considerably less expensive than the Sanger method. However, the read lengths of NGS technologies create important algorithmic challenges. While the 454 platform is able to obtain reads in the 400-600 base pairs (bp), the Illumina's Solexa 1G Sequencer and the Ion Torrent Systems present reads with an average length of 100 bp and the SOLiD platform is currently limited to 25-50 bp.

Several assembly tools have recently been developed for generating assemblies from short, unpaired sequencing reads. However, the sheer volume of data generated by these technologies and the need to align reads to increasing large reference genomes limit the applicability of standard methods.

One way to speed up the read alignment task is to resort to software based on approximate indexing technologies. Indexed alignment algorithms, which preprocess the reference genome into an index data structure that can then be searched, correspond to more efficient approaches. On one hand it can discard irrelevant portions of the reference genome much more efficiently. On the other hand the computation on relevant regions can be factored out. However, building indexes is time and space consuming. State of the art algorithms are using techniques from a new class of indexes, compressed indexes, which have smaller space requirements by using data compression techniques to eliminate regularities in the indexes.

In this work we present TAPyR (http://www.tapyr.net) [1] a new method for the alignment of NGS reads that uses compressed indexing build an index of the reference genome sequence to accelerate the alignment. Being firstly proposed to handle the 454 GS FLX data, it can also be used with Illumina and Ion Torrent data. Like other algorithms, TAPyR uses in a second stage a multiple seed heuristic to anchor the best candidate alignments. This heuristics has the advantage that it dispenses the need of determining the number and length of the seeds beforehand, relying on the assumption that the optimal alignments are mostly composed of relatively large chunks of exact matches interspersed by small, possibly gapped, divergent regions. At the ultimate stage banded dynamic programming is used to finish up the candidate multiple seed alignments considering user-specified error constraints.

TAPyR was evaluated against other mainstream mapping tools namely BWA-SW [2], SSAHA2 [3], Segemehl [4], GASSST [5], and Newbler [6]. The analyses were performed with real and simulated data sets. As the results show the new method manages to achieve convincing performance in terms of speed and in terms of the number and precision of aligned reads. In fact, TAPyR has displayed class-leading CPU-time performance and excellent use of input reads in comparison to other mainstream tools.

References

1.
2. Li H, Durbin R: Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010, 26(5):589-95.
3. Ning Z, Cox AJ, Mullikin JC: SSAHA: a fast search method for large DNA databases. Genome Res 2001, 11(10):1725-9.

4. Ho_mann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackerm□uller J: Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 2009, 5(9):e1000502.
5. Rizk G, Lavenier D: GASSST: global alignment short sequence search tool. Bioinformatics 2010, 26(20):2534-40.
6. Droege M, Hill B: The Genome Sequencer FLX System - longer reads, more applications, straight forward bioinformatics and more complete data sets. J Biotechnol 2008, 136(1-2):3-10.

Computational Phylogenomics

David Posada

Universidade de Vigo

Abstract

The massive amount of data resulting from next-generation sequencing techniques has opened a new era in phylogenomics, or phylogenetics at large. In this talk I will offer an overview of some the most important computational challenges evolutionary biologists will be confronted with to understand the evolutionary relationships among genomes. This will include dealing with unprecedented amounts of information and with obvious conflicting signals from different genomic regions. We should anticipate an exponentially large computational burden that could restrict the implementation of appropriately complex statistical models and/or the completion of thorough large-scale analyses. In this regard, the implementation of High Performance Computing (HPC) techniques will become essential to overcome these limitations. Parallel computing is increasingly common in bioinformatics, and phylogenomics will not be an exception.

Exploring the Semantic of Biomedical Ontologies

Francisco Couto

Abstract

The analysis of complex biomedical entities and events, such as disease and epidemiological models, is challenging due to their multiple domain features, and thus to accurately describe them we need to use concepts from multiple biomedical ontologies, such as gene mutations, protein functions, anatomical parts and phenotypes. The usefulness of ontological annotations to interlink and interpret biomedical information is widely recognized, particularly for retrieving related information. This relatedness can be captured by semantic similarity measures that return a numerical value reflecting the closeness in meaning between two ontology concepts or two annotated entities. These measures have been successfully applied to biomedical ontologies, particularly to the Gene Ontology, for comparing proteins based on the similarity of their functions.

This talk will present ongoing efforts that apply semantic similarity measures to perform:

- Chemical entity recognition and mapping
- Ontology matching and extension
- Enzyme family coherency assessment
- Epidemiological data retrieval

Challenges and solutions on genotype-to-phenotype integration

José Luís Oliveira

University of Aveiro, DETI / IEETA, 3810-193 Aveiro, Portugal; Contact: jlo@ua.pt

Abstract

The decoding and annotation of the human genome have been leading to a better understanding of the evolution process of the species and the relation between diseases and genes. The results of human variome projects may open new frontiers in our understanding and treatment of diseases. The discovery of novel relationships between simple sequence changes and the diseases that result from them is essential to underpin the future prospect of custom drug design and personalized patient care. However, management and processing all generated data as well the scientific information that can be obtained from those studies are still an open issue.

Locus specific databases are typically closed systems, filled with heterogeneous tools and non-standardized legacy systems. This is a major drawback for data exchanges, aggregation in external systems or integration of resources. Researchers need access to miscellaneous resources and features whilst browsing gene variants: gene loci and variant information should be complemented with related proteins, pathways or published literature, among others. Nowadays, this is not possible without a complex data analysis workflow involving interactions with various distinct applications.

These latter requirements leverage the requirement for the development of new software tools that aggregate both high-quality LSDB genomic variation datasets and connections to external resources.

The GEN2PHEN project aims to unify human genetic variation databases and to put in place the main building blocks needed to move substantially from G2P disaggregated databases situation towards the ultimate future of a complete biomedical knowledge environment. This will consist of a European-centred but globally networked hierarchy of bioinformatics GRID-linked databases, tools and standards, all tied into the Ensembl genome browser.

Multi-Relational Learning for the Life Sciences

Rui Camacho

Abstract

Inductive Logic Programming (ILP) is a major field in Machine Learning with important applications in (Relational) Data Mining. The fundamental goal of an ILP system is to construct models for data. The main ingredients for ILP are: i) background knowledge; and ii) observations (usually called examples in the ILP literature). A characteristic of ILP is that both data and models are expressed in a subset of First Order Logic (FOL) providing and expressive language to encode both data and models. Background knowledge is a set of predicates encoding all the information that the domain expert finds relevant for constructing the models. ILP have been applied not only to classification problems but also clustering, regression problems and Association Rule Discovery.

Major advantages of ILP that make it adequate for the Life Science applications include its facility to encode data with structure, the facility of using data from diverse sources and encoded in different formats and the induction comprehensible models. It is also easy for an ILP system to combine numerical reasoning with symbolic relations within the same model.

Together with colleagues from University of Porto we have applied ILP to proteomics and Rational Drug Design applications. We have developed LogCHEM, a tool that allows a chemist to graphically control the search for interesting patterns in chemical fragments. LogCHEM couples an ILP system with a molecular visualisation software, thus leveraging the flexibility of ILP while addressing the SAR task. LogCHEM can input data from chemical representations, such as MDL's SDF file format, and display molecules and matching patterns using visualisation tools such as VMD. It has been demonstrated that LogCHEM can be used to mine effectively large chemoinformatics data sets.

In silico screening of transthyretin amyloid inhibitors

Carlos J. V. Simões, Catarina S. H. Jesus, Rui M. M. Brito

Abstract

Amyloid diseases embody a wide spectrum of acquired, inherited, or infectious pathologies that includes diseases, such as Alzheimer's disease, type 2 diabetes, familial amyloid polyneuropathy (FAP), or cardiomiopathy (FAC), and several others with significant socio-economic impact. Although caused by different proteins, which do not share sequence or structural homology, amyloid diseases are triggered by the formation of highly ordered protein aggregates which are cytotoxic, and thus these diseases share common molecular mechanisms.

As a model to study amyloid formation, we have been using the protein transthyretin (TTR), a homo-tetrameric protein present in the blood plasma and cerebral spinal fluid, and implicated in the deposition of fibrils in peripheral nerves and heart tissue in FAP, FAC and Senile Systemic Amyloidosis (SSA). Amyloid formation by TTR involves the dissociation of the native tetrameric form of the protein, partial unfolding of the subunits and aggregation into soluble oligomers which grow into insoluble oligomers and eventually mature amyloid fibrils.

With the goal of finding efficient therapeutic approaches against TTR amyloid, we have started a virtual screening program to find compounds with potential to stabilize the native tetrameric form of TTR and consequently inhibit amyloid formation. We have used several receptor-based and ligand-based virtual screening methods, ranging from 2D and 3D similarity searches to molecular docking. We have used an initial virtual library of more than 11 million compounds, and large scale molecular docking has been performed on the Ibercivis volunteer computing platform (www.ibercivis.net) through the AMILOIDE project. This search culminated in the identification of several compounds with inhibitory activity for TTR amyloid formation *in vitro*.

Selected Abstracts

A Computational Modeling Study on African Swine Fever Virus encoded Proteins with Predicted Structural Repeats

Elsa S. Henriques and Rui M. M. Brito

ehenriques@qui.uc.pt

Center for Neuroscience and Cell Biology

Abstract

All metazoan cells have the intrinsic capacity to sense virus infection and restrict viral replication and spread. Viral pathogens hijack host cellular machinery for their own replication and evade the surveillance of the host immune system in a highly antagonistic manner. Viral proteins tend to bind to and mimic existing within-host protein-protein interfaces otherwise occupied by multiple, transiently bound regulators **[1]**. 3D structural information provides a mechanistic and high-resolution view of the interactions involved, but for many virus this information is often scarce, this being the case of the African Swine Fever Virus (ASFV). Of the over 150 genes encoded by its complex DNA genome, only for two proteins has the 3D structure been experimentally resolved so far, none of them involved in host cell function modulation. This is meager, specially considering that ASFV is the aetiological agent of a highly contagious lethal pig disease against which no vaccine has ever been obtained. Like many other viral proteins known to be involved in the so called interface mimicry, a number of such ASFV components share only remote sequence relationships with their putative host targets. This does not imply that they can not feature extensive structural overlap with those targets **[1]**, as it is most likely the case of those ASFV genes that target and are predicted to fold into repeat domains like Leu-rich repeats (LRR) or ankyrin-repeats. Such domains are built from tandems of several repeats, with a canonical aminoacid pattern but otherwise low sequence similarity. Yet, they share a common structural framework and a basic shape that is particularly suitable for protein-protein interactions. This scenario prompts for a challenging computational modeling exercise of the ASFV proteins in question. To illustrate it, here we propose a structural, full atomistic computational model of the ASFV protein pI329L, which was recently found to inhibit the Toll-like receptor 3 (TLR3) signaling pathway and is predicted to be a trans-membrane protein containing extracellular LRR, much like the TLR3 itself **[2,3]**. The model was built with a combined strategy involving PHYREv0.1 automated folding prediction, ITASSER free modeling via structural fragment-excised reassembly, a customized template-based LRRassembly approach, restrained homology modeling within Modeller.9v4 and experimental guidance. Following the best estimate fold predictions for the putative ecto- (extracellular) and cyto- (intracellular) domains of pI329L, initial crude models for these two domains were separately generated using free modeling. For the cyto-domain, a curated PDB-search by secondary structure content retrieved an apposite protein structural domain that was used for an ensuing template-based refinement. For the ecto-domain, the

repeat organization of the crude LRR-like solenoid model required reassignment. This was achieved by pattern-matching it to five LRR-representative proteins (including TLR3), sequence-aligned by LRRconsensus secondary structure. The resulting multiple-alignment was then used for homology refinement of the structure, to which the trans-membrane helix connecting the two domains was lastly grafted using restrained modeling. The pI329L final model exhibits a plausible fold and good structural quality. The way the cyto-domain can superpose the TLR3 cytoplasmic TIR-domain corroborates the experimental evidence that pI329L downstream inhibition occurs at the TLR3/TRIF level **[3]**. For this superposition to take place, a decoy-like pI329L:TLR3 heterodimer at extracellular level must first occur, and the modeled viral LRR-solenoid features the apposite interface mimicry for it, while avoiding a TLR3 key region that would otherwise promote the transcriptional activation of important immuno-modulatory genes **[4]**.

References

1. Franzosa, E.A., Xia, Y. 2011. PNAS, June 16, epub ahead of print.
2. de Oliveira, V.L., Almeida, S.C.P., Soares, H.R., Crespo, A., Marshall-Clarke, S., Parkhouse, R.M.E. 2011. Arch Virol. 156: 597–609.
3. Henriques, E.S., Brito, R.M., Soares, H., Ventura, S., de Oliveira, V.L., Parkhouse R.M. 2011. Protein Sci. 20: 247-255.

4. Henriques et al., submitted.

Recognition and resolution of chemical entities to ChEBI.

Tiago Grego[1], Francisco M. Couto[2]

[1,2] Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa

[1] tgrego@fc.ul.pt (corresponding author)

[2] fcouto@di.fc.ul.pt

Abstract

Virtually all biological phenomena have their roots in chemical processes, and it is impossible to understand the complex signaling and metabolic networks of living systems without taking into account the chemical entities that take part in it.

Since the late 1960's, Bioinformatics emerged and was applied in the creation and maintenance of public biomedical databases, but only recently public chemical databases such as ChEBI (Chemical Entities of Biological Interest) were released providing high quality manually curated chemical data organized in an ontology.

However, the largest repository of scientific data is the scientific literature, containing millions of documents such as scientific publications and patents in unstructured natural language text.

Within the literature exists extensive information that is not covered in other knowledge resources, thus database curators manually analyze and annotate the literature to grow and validate the data in the databases, but this is a tedious, time consuming and costly process.

Fortunately, this process have been addressed by automatic text mining systems that already shown to be helpful in speeding up some steps of this process, namely performing named entity recognition and entity resolution.

Two main approaches are used by such systems:

Dictionary based approaches require domain terminologies to find and map matching entities in the text and depend on the availability and completeness of these terminologies.

Whatizit is a popular text processing system that uses a dictionary based approach for identifying a wide variety of biomedical terms by using several pipelines based on specific terminologies, including ChEBI.

Machine learning based approaches require an annotated corpus which is used to learn a model that can be applied in the named entity recognition of new text. An entity resolution module is required to perform the mapping of recognized entities.

Recently, a joint team of curators from ChEBI and the European Patent Office released a gold standard corpus composed by 40 patent documents manually annotated with 18061 chemical entities, from which 9696 could be mapped to ChEBI. With the availability of this this corpus we developed a Conditional Random Fields based machine learning method for chemical entity recognition and a lexical similarity based method for chemical entity resolution to the ChEBI database, and compared our methods with Whatizit [1].

We found that dictionary based method can already provide competitive results in recognizing chemical named entities obtaining F-measures of up to 70% for partial matching and 32% for exact matching assessment in the full gold standard. However the developed machine learning method consistently outperformed it obtaining F-measures of 77% for partial matching and 57% for exact matching.

A known drawback of dictionary based methods is the inability to recognize entities not present in the dictionary used, so an evaluation was performed using only the mapped entities in the gold standard. Still, the machine learning method outperformed the dictionary based method obtaining 66% and 57% F-measure for respectively partial and exact matching against the 54% and 39% obtained using Whatizit.

Taking into consideration not only entity recognition but also resolution, our methods were able to obtain an F-measure of 51% for partial matching and 47% for exact matching, while the dictionary method could only obtain 37% and 32% respectively.

Overall, we demonstrated that a completely dictionary independent machine learning entity recognition method and a lexical similarity resolution method can surpass dictionary based methods in recognizing chemical compounds and mapping them to the ChEBI database.

Tools like the one presented here [2] can be helpful to gather more knowledge about biomedical entities involved in complex biological processes, such as metabolic networks. Thus, we intend to apply our tool to the project SPNet, which aims to uncover network motifs associated with virulence in pneumococcus. After the identification of genes with an impact in virulence, the analysis of the transcriptional and metabolic network for neighbor genes can locate other direct or indirect target genes involved in the virulence.

The metabolic networks comprise the chemical reactions within the organism, thus to understand them a characterization of their chemical entities is required. This will be the aim for the application of our tool to SPNet, in order to identify the chemical entities in literature, which may provide more information for the integration with the genomic and proteomic data that may uncover the network motifs associated with the virulence of the bacteria.

References

1. Rebholz-Schuhmann et al, Text processing through Web services: calling Whatizit. Bioinformatics, 24(2) 296-298, 2008
2. Grego T et al, Identification of chemical entities in patent documents. Lecture Notes in Computer Science, vol.5518, 942–949, 2009

Prognostic prediction using clinical expression time series: towards a supervised learning approach based on meta-biclusters

André V. Carreiro[1], Artur J. Ferreira[2], Mário A. T. Figueiredo[3] and Sara C. Madeira[1]

[1]KDBIO group, INESC-ID, Lisbon, and Instituto Superior Técnico, Technical University of Lisbon, Portugal, e-mail: acarreiro@kdbio.inesc-id.pt (corresponding author), sara.madeira@ist.utl.pt

[2]Instituto de Telecomunicações, Lisbon, and Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal,e-mail: arturj@isel.pt

[3]Instituto de Telecomunicações, Lisbon, and Instituto Superior Técnico, Technical University of Lisbon, Portugal,e-mail: mtf@lx.it.pt

Abstract

Biclustering has been recognized as a remarkably effective method for discovering local temporal expression patterns and unraveling potential regulatory mechanisms, critical to understand complex biomedical processes, such as disease progression and drug response. In this work, we propose a classification approach based on meta-biclusters (a set of similar biclusters) applied to prognostic prediction. These biclusters thus represent temporal expression profiles potentially involved in the transcriptomic response of a set of patients to a given disease or treatment. We use real clinical expression time series to predict the response of patients with multiple sclerosis to treatment with Interferon-β. The main advantages of this strategy are the interpretability of the results and the reduction of data dimensionality, due to biclustering. Preliminary results anticipate the possibility of recognizing the most promising genes and time points explaining different types of response profiles, according to clinical knowledge. The impact on the classification accuracy of different techniques for unsupervised discretization of the data is also studied.

Accepted submissions

Unraveling the molecular mechanisms of ABC transporters: Insights from molecular dynamics simulations

A. Sofia F. Oliveira[1], António M. Baptista[1], and Cláudio M. Soares[1]

[1]Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal (asfo@itqb.unl.pt)

Abstract

ATP-Binding Cassette (ABC) transporters are ubiquitous membrane proteins that use the energy from ATP-binding or/and hydrolysis to actively transport substrates (named allocrites) across membranes against the concentration gradient [1]. The allocrites are chemically very diverse, ranging from small ions to polypeptides. These transporters are found virtually in all living organisms and mutations in several members of this family have been associated with genetic diseases in humans (such as cystic fibrosis [2] and Tangier disease [3]) and to multidrug resistance in bacteria, fungi, yeasts, parasites and mammals.

ABC transporters can be divided in two groups: importers, which translocate allocrites to the cellular interior, and exporters, which do the opposite. Independently of the transport directionality, ABC transporters are usually composed by a minimum ‘‘functional core’’ formed by four modules [1]: two transmembrane domains (TMDs) and two catalytic domains (NBDs).

However, and despite the large amount of experimental and theoretical data available for several family members many fundamental questions about the ABC-family remain unanswered until this moment. In particular, it is still not clear which are the conformational changes induced by ATP-hydrolysis in the NBDs, nor how these rearrangements are ‘‘transmitted’’ to the TMDS, in order to drive allocrite translocation.

The main objective of this work is to identify and map the structural changes occurring during an ATP-hydrolytic cycle in three distinct systems: an isolated NBD dimer from *Methanococcus jannaschii* (MJ0796) [4], a full length ABC exporter from *Staphylococcus Aureaus* (Sav1866) [5] and a complete ABC import system from *Escherichia coli* ($MalFGK_2E$) [6]. In this work, we present the results of three computational studies [4-6] using extensive Molecular Dynamics (MD) simulations of several intermediates states of the ATP-cycle. Our MD simulations allowed us to identify that the conformational rearrangements occurring during the ATP-cycle are not restricted to the NBDs, but extend to the TMDs external regions. Additionally, in the context of the complete transporters, we were also able to identify the atomic details associated with the NBD dimer opening upon hydrolysis. Lastly, we suggest a general mechanism for coupling hydrolysis and energy transduction to allocrite translocation (independently of the transporter directionality), in which the NBDs “helical sub-domain region” and the TMDs ‘‘coupling helices’’ are the keystones.

References

1. C.F. Higgins *Annu Rev Cell Biol* **8** (1992) 67.
2. J.R. Riordan et al. *Sciencel* **245** (1989) 1066.
3. M. Dean et al. *Genome Res* **11** (2001) 1156.
4. A. S. F. Oliveira et al. *J Phys Chem B* **114** (2010) 5486.
5. A. S. F. Oliveira et al. *Proteins* **79** (2011) 1977.
6. A. S. F. Oliveira et al. *PLoS Comput Biol* **7** (2011) e1002128.

HOXA9 transcriptome analysis in Glioblastoma: The bioinformatics beyond the bench

Ana Xavier-Magalhães[1,2], Céline S. Gonçalves[1,2], Miguel Rocha[3], Bruno M. Costa[1,2]

[1]Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal; [2]ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal; [3]Department of Informatics, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal.

Abstract

Introduction: Gliomas, originating from the predominant glial tissue, are the most common primary tumors of the central nervous system (CNS) in adults. The most common form is glioblastoma (GBM, WHO grade IV), and the median survival of the affected patients is approximately 1 year after diagnosis. The identification of molecular features specific of GBM may aid in defining prognostic subgroups. Some putative prognostic biomarkers in GBM patients include *MGMT* methylation and *HOXA9* expression, but the establishment of additional clinically-relevant biomarkers is imperative. Overexpression of HOXA9 is associated with pro-proliferative and pro-invasion properties and with a poor prognosis. Since HOXA9 is a transcription factor, we hypothesized that it transcriptionally regulates a specific set of genes that may be the true biological effectors. In this way, we aimed to analyze its transcriptome to identify novel prognostic biomarkers and therapeutic targets.

Material and Methods: We performed expression microarrays in *HOXA9*-negative and *HOXA9*-overexpressing U87MG cells, and the resulting transcriptome was studied using bioinformatics tools (e.g. DAVID, Bioconductor). In order to validate the microarray results, a subset of the over- and under-expressed genes were selected and tested by RT-PCR. To verify which of the differentially expressed genes are direct targets of HOXA9, an *in silico* analysis was performed using Biobase software. By RT-PCR we inquired a few of these putative direct-targets in GBM cell-lines A172 (endogenously express *HOXA9*) and U87-MG (retroviraly infected to overexpress *HOXA9*, and the respective empty control-MSCV). To confirm if the putative direct-targets are transcriptional targets, we performed chromatin immunoprecipitation (ChIP) in U87-HOXA9, U87-MSCV and A172 cell-lines. As *HOXA9* expression has prognostic value in GBM patients, we analyzed if the HOXA9 targets are associated with patients' survival in two different cohorts of patients (TCGA and Rembrandt databases).

Results: By analyzing gene expression data with Bioconductor, we found 1537 genes upregulated and 1917 downregulated in U87MG-HOXA9 cell line relatively to U87MG-MSCV. DAVID analysis showed several cancer-related pathways upregulated in HOXA9^{+} cells, as "Pathways in Cancer" (e.g., Wnt signaling) and "Cell Cycle", and some pathways downregulated, as "Focal Adhesion" and "Cell Adhesion Molecules". Taking this into account, by RT-PCR we found that the expression level of *WNT6,* a gene of the WNT signaling pathway, is higher in GBM cell lines that overexpress *HOXA9*. Oncomine analysis showed that *WNT6* is frequently co-expressed with *HOXA9* in primary GBM. ChIP experiments showed that *HOXA9* transcriptionally activates *WNT6* by directly binding to its promoter region. Finally, as *HOXA9* may be considered as a prognostic factor, we found that *WNT6* overexpression is significantly associated with worse survival in GBM patients' in two independent sets of patients (TCGA and Rembrandt databases).

Conclusion: Bioinformatics is playing a vital role in biological sciences, as it allows the systematical analysis of biologically relevant information. Its relevance arises especially among the enormous amounts of raw data generated within genome-wide projects, providing a reduction in high quantities of information to small subsets, easily inquired on the bench. Based on bioinformatics clues, here we were able to identify *WNT6* as a direct target of HOXA9, and how its expression associates with GBM patients' survival.

Data Mining Strategies for the Analysis of Protein Folding and Unfolding Simulation Data

Cândida G. Silva[1], Pedro Gabriel Ferreira[2], Paulo J. Azevedo[3], Rui M. M. Brito[1,4]

[1]Grupo de Biologia Estrutural e Computacional, Centro de Neurociências, Universidade de Coimbra, Coimbra, Portugal

[2]Bioinformatics and Genomics Research Group, Centre de Regulacion Genomica, Barcelona, Espanha

[3]Departamento de Informática, Universidade do Minho, Braga, Portugal

[4]Departamento de Química, Universidade de Coimbra, Coimbra, Portugal

Abstract

In recent years, a number of promising comparative-based approaches to analyze results from multiple molecular dynamics (MD) simulations have been reported in the literature, highlighting the fundamental role these tools play in analyzing the enormous amount of new data that it is currently being generated and in providing the basis for a new understanding of such complex processes as protein folding and unfolding.

Using multiple MD unfolding simulations of the amyloidogenic protein transthyretin (TTR), we explored the usefulness of data mining to help inferring rules for the folding and unfolding behavior in amyloidogenic proteins, and eventually to help predict unfolding behavior or even structural determinants of native protein folding and stability. Human TTR is a homotetrameric protein involved in amyloid pathologies, such as Familial Amyloid Polyneuropathy (FAP). High temperature molecular dynamics (MD) simulations had previously been performed to explore the unfolding routes of monomeric species of wild-type TTR and its highly amyloidogenic variant L55P, in order to study the unfolding mechanisms at the source of its pathological behavior [1].

We started by applying cluster analysis to group conformational states sharing similar structural properties across multiple simulations of TTR monomeric species. The structural similarity between any two conformations was measured using the root mean square (RMS) distance between the alpha carbon (C-alpha) atoms. The selection of cluster representatives was done based on four distinct molecular properties of the conformations: number of hydrogen bonds, number of native contacts, value of non-polar solvent accessible surface area, and number of amino-acid residues in regular secondary structure. In the case of L55P-TTR, two interesting cluster representatives of partially unfolded conformations were found [1].

Instead of looking directly into the conformations sampled from the MD simulations, an alternative analysis of the unfolding conformational landscape of a protein can be attained by monitoring changes in several molecular properties of the protein along multiple unfolding simulations. The solvent accessible surface area (SASA) is one of the molecular properties that might be calculated for each MD trajectory. The study of the SASA variation of each individual amino-acid residue provides a greater level of detail on the individual contributions for the folding or unfolding processes, shedding light on potential coordinated behavior of residues far apart in the protein primary structure, or even far apart in the three-dimensional structure. We proposed the application of two different methods - cluster analysis and association rules - to study the SASA profiles of individual amino-acid residues across multiple MD unfolding simulations of WT- and L55P-TTR searching for groups of amino-acid residues playing a coordinated role in protein unfolding.

First, we propose a hierarchical clustering-based method to search for prevalent correlations among groups of amino-acid residues across multiple WT- and L55P-TTR unfolding trajectories. Our results show that for both proteins the most prevalent correlations are found in the most stable core of TTR comprised by beta-strands A, B, E and G. However, for L55P-TTR these correlations are weaker, reflecting the greater fluctuations in the SASA profiles of the amino-acid residues across different simulations. These results

show that the protein core is more stable in the wild-type than in the highly amyloidogenic variant L55P-TTR, suggesting that the stability of the TTR monomer is affected by the mutation [2].

Additionally, it is known that the interactions between hydrophobic residues are important in restricting the conformational space available to a protein as it folds. An association rule mining algorithm was applied to the SASA variation profiles of the hydrophobic amino-acid residues of WT- and L55P-TTR along the MD unfolding trajectories to elucidate potential relationships among these residues in TTR unfolding. For both proteins, we identified a group of hydrophobic residues distributed in beta-strands A, B, D, E and G behaving in a coordinated fashion across multiple simulations of WT- and L55P-TTR. Furthermore, we observed that in L55P-TTR the hydrophobic residues become more exposed to the solvent and earlier in the simulation. Moreover, the beta-sheet DAGH is more destabilized in L55P- than in WT-TTR [3].

References

1. Rodrigues, J.R., Simões, C.J.V., Silva, C.G., Brito, R.M.M. (2010) Protein Sci. 19(2), pp. 202.
2. Ferreira, P.G., Silva, C.G., Brito, R.M.M., Azevedo, P.J. (2007) Proc. IEEE CIBCB'07, pp. 461.
3. Azevedo, P.J., Silva, C.G, Rodrigues, J.R., Loureiro-Ferreira, N., Brito, R.M.M. (2005) LNCS, vol. 3725, pp. 329.

Reconstruction of a genome-scale metabolic model for the filamentous fungus *Ashbya gossypii*

Daniel Gomes, Oscar Dias, Eugénio Ferreira, Lucília Domingues, Isabel Rocha

IBB – Institute for Biotechnology and Bioengineering, Centre of Biological Engineering

Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal

E-mail {danielg_gomes, odias, ecferreira, luciliad, irocha}@deb.uminho.pt

Abstract

Systems biology has recently arisen as a promising and powerful tool for process development and optimization. Metabolic models are one of its different methodologies with high interest and applicability since it allows the simulation of cells behavior under different environments and/or specific genetic variations. The fast-growing number of sequenced genomes may have contributed to this phenomenon, as the sequenced genome is the starting point from where it is possible to associate by homology a specific function to the genes of a microorganism. Afterwards, from this entire set of enzymatic functions we can possibly identify the main metabolic pathways of a specific microorganism allowing this way the construction of its metabolic network.

One of the genomes already sequenced is from the fungus *Ashbya gossypii*, an industrially relevant microorganism intensively used for riboflavin production. Despite the high similarity with *Saccharomyces cerevisiae* genome *A. gossypii* presents a lower level of complexity containing only 4726 protein-coding genes distributed over seven chromosomes. The aim of this work is to construct a metabolic model for *A. gossypii* based on its genome and from which we can retrieve valuable information concerning specific metabolic pathways and the optimum conditions for the production of interesting compounds such as riboflavin.

The initial stage of this process consisted in the collection of all metabolic-relevant genes through a manual re-annotation of *A. gossypii* genome. Despite being a manual procedure, this step was performed using the user-friendly software – *merlin* – that provided an automatic annotation for each gene, speeding up the entire process. The function automatically assigned by this application was manually analyzed, being accepted or replaced by another one. Each metabolic gene was assigned to an Enzyme Commission(EC) number that corresponds to a specific enzyme. For such procedure several databases were used such as UniProt, SGD, AGD, ExPASy and BRENDA.

At the end of this phase, a total of 1429 genes were assigned among the different enzymatic families. Such distribution was considerably heterogeneous: 35,4 % hydrolases; 35,8 % transferases; 28,8 % other enzymatic families. Of the 1429 genes 59 were assigned to different EC numbers and among these 36 % have EC numbers from different enzymatic families.

The next stage of the reconstruction process, being performed at present, involves the elaboration of the set of metabolic reactions and their curation regarding stoichiometry, balance of charges and localization inside the cell. For this purpose, information from genome re-annotation is crossed with curated models from closely related microorganisms such as the iMM904 (Mo et al., 2009) and the iIN800 (Nookaew et al., 2008) from *Saccharomyces cerevisiae*. To complete this process regarding information that was not found in the curated models, reactions databases like BRENDA or KEGG will be used to retrieve such data.

At the end of this phase, once we have the complete set of curated metabolic reactions, we will be able to construct a system of m equations (metabolites) and n variables (reactions) that is the base for the optimization studies. The next step will be the determination of biomass composition that encounters another key element for the optimization studies.

Acknowledgements: The authors thank the financial support of Fundaçãopara a Ciência e aTecnologia (FCT), Portugal: project AshByofactoryPTDC/EBB-EBI/101985/2008.

Comparison of *H. pylori in silico* metabolic model predictions with experimental data

D.M. Correia[1], M.L.R. Cunha[1], N.F. Azevedo[1], M.J. Vieira[1], I. Rocha[1]

[1]IBB - Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, Braga, Portugal

Abstract

The Systems Biology approach has been replacing the reductionist view that dominated biology research in the last decades. Present biochemical knowledge and genomic databases allowed the development of metabolic models for several organisms, which, however, are still incomplete. The availability of the genome sequence of *H. pylori* has allowed the construction of a genome-scale metabolic model for this organism. The purposes of this work were to study the growth of *H. pylori* in a chemically defined medium, to compare the experimental data obtained with the simulated data supplied by the model and analyse the composition of the *in silico* media used.

Cultures were grown at 37ºC under microaerophilic conditions in Ham´s F-12 medium supplemented with fetal bovine serum. Optical density and the counting of CFU/mL were performed for assessing the growth. OptFlux, a software platform for metabolic engineering, which includes several tools such as flux balance analysis (FBA) was employed for simulate the behavior of wild type *H. pylori* under the conditions used *in vivo*.

The simultaneous use of both approaches allows to correct the *in silico* model, and on the other hand, to rationally adjust the medium components present in F-12. For instance pimelate, that has been considered to be essential in the latest metabolic model, is lacking in F-12 and is likely to be redundant in the model.

Our future work is not only to improve the genome-scale metabolic model, but also, identify potential targets for design more effective drugs for the inactivation of *H. pylori.*

Calibration of Logic-Based ODE Models

David Henriques

Abstract

Here we approach the problem of calibrating logic-based ODE models in the context of intracellular protein signaling networks. To achieve this we extended the functionalities from the tool CellNetOptimizer, used for the analysis of discrete logic models, combining it with efficient software for model simulation and analysis as well as parameter estimation. Additionally we apply the methods to three case studies.

Logic-based ODE models are a dynamic variant of logic models where networks can be modeled without biochemical information, in a simpler manner compared with mechanistic biochemical models.

In the first case study we use a logic-based ODE model to reproduce the behavior of a mechanistic biochemical one. While using fewer parameters and dynamic states logic-based ODE model was able to reproduce the mechanistic one quite well.

In the second case study we tested our ability to reverse engineer a set of parameters from a logic-based ODE model while using different *in silico* generated data-sets with different quantity and quality of information. Here we were generally successful.

Finally in the last case study we performed parameter estimation in a realistically sized model against real experimental data in the presence of multiple cytokines (used as stimuli) and small molecule inhibitors. Here the results were limited, nevertheless, possible causes for this are discussed.

The preliminary results obtained are promising, however, further analysis and comparative studies with other modeling approaches are needed in order to validate logic based ODE models as a tool of great general applicability for modeling intracellular protein signaling. Here we present the methods used in this study and provide a computational framework to achieve this goal.

Efficient high-throughput read mapping for re-sequencing applications

Francisco Fernandes[1,2] (fjdf@kdbio.inesc-id.pt), Paulo G.S. da Fonseca[1] (pgsf@kdbio.inesc-id.pt), Luis M.S. Russo[1,2] (lsr@kdbio.inesc-id.pt), Arlindo L. Oliveira[1,2] (aml@inesc-id.pt) and Ana T. Freitas[1,2*] (atf@inesc-id.pt)

[1] Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento (INESC-ID), R. Alves Redol 9, 1000-029 Lisboa, Portugal.

[2] Instituto Superior Técnico – Universidade Técnica de Lisboa (IST/UTL), Av. Rovisco Pais, 1049-001, Lisboa, Portugal.

Abstract

Over the past few years, new massively parallel DNA sequencing (MPDS) technologies based on the sequencing-by-synthesis approach have emerged. These platforms generate massive amounts of data per run, greatly reducing the cost of DNA sequencing compared to the traditional capillary electrophoresis method. However, these techniques also raise important computational difficulties due to the huge volume of data produced and their specific error characteristics. One of the main applications of these technologies consists in sequencing a new genome or region of interest using a homologous sequence of another individual of the same species or a taxonomically related species as a reference. A crucial step of this task corresponds to mapping the sequenced fragments, or reads, onto the reference sequence. The sheer volume of data generated by MPDS technologies (to the order of hundreds of gigabases per run), and the need to align reads to large reference genomes limit the applicability of standard techniques.

We have recently proposed a new method for the alignment of MPDS reads *[1]*. Our initial efforts focused on high-throughput pyrosequencing data, like those produced by the Roche 454 GS FLX platform. However, the heuristic employed by our method has proven to be sufficiently generic to handle reads with different lengths and error characteristics equally as well. Our proposed tool, called TAPyR, builds a Burrows-Wheeler Transform (BWT) based index of the reference sequence to accelerate the alignment. It then employs a multiple seed heuristic to anchor the best candidate alignments. Contrary to other seed-based alignment tools, our strategy adds more flexibility by dispensing with the need of determining the number and length of the seeds beforehand. A banded dynamic programming is used to finish up the candidate multiple seed alignments considering user-specified error constraints.

We evaluated TAPyR against a host of other mainstream mapping tools using a comprehensive collection of real and simulated data sets with the objective of assessing their efficiency and accuracy in the context of re-sequencing projects. The results show that TAPyR compares favorably to the evaluated tools. Our proposed solution displayed class-leading CPU-time performance and excellent use of input reads, managing to map more reads than its competitors at comparable settings. Our tests with simulated data also showed that TAPyR was more accurate than the other tools, consistently mapping a high percentage of simulated reads back to their original positions, despite a considerable level of introduced noise. Memory requirements are also on par with the best in this category of tools, being linearly proportional to the size of the input reference sequence by a small factor. Based on these results, we propose that TAPyR constitutes an advantageous alternative for re-sequencing projects. The tool, currently available from www.tapyr.net, complies with standard IO file formats and is straightforward to use, requiring almost no external parameterization.

References

1. Fernandes et al. Efficient alignment of pyrosequencing reads for re-sequencing applications. BMC Bioinformatics (2011) vol. 12 pp. 163.

Bacteriophage phylogeny revisited

Franklin L. Nóbrega[1], Graça Pinto[2], Joana Azeredo[3] and Leon D. Kluskens[4*]

IBB – Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, Universidade do Minho, Campus de Gualtar, Braga, Portugal

*Corresponding author, [1]franklin.nobrega@ceb.uminho.pt, [2]gracap88@gmail.com, [3]jazeredo@deb.uminho.pt, [4]kluskens@deb.uminho.pt

Abstract

Bacteriophages or phages are viruses that only infect bacteria. The International Committee on Taxonomy of Viruses classified these viruses in accordance with the morphology of their free virion particles and type and size of their genome. This system fails on the classification of several phages, which have their genome already sequenced. It also requires a morphological analysis by transmission electron microscopy, which is very expensive and time consuming [1]. In 2002 Rohwer and Edward proposed the only sequence-based system existing up to this moment. Thus, it is of utmost importance to develop new systems for bacteriophage classification that take into consideration the genomic and proteomic information already available [2].

The purpose of this study is to establish a new method for the classification of phage based on the genetic information available. The principal objective is to cluster the bacteriophages in different family and types. To create a new *phylogenetic tree* we analysed all 670 available genome sequences deposited in the *GenBank* database. Sequences were aligned using the *T-coffee* program [3]. A genetic marker for the construction of the phylogenetic tree was designed by creating a *concatenate* of different gene products that presented the highest similarity. In other words, the most conserved gene products were used to form a broader genetic marker. The method allows the use of a single, created genetic marker to classify unknown phages with existing phage types and families. A comparison to existing methods is discussed.

References

1. Calendar, R., 2006. *The Bacteriophages* 2nd ed. R. Calendar, ed., New York: Oxford University;
2. Rohwer, F. & Edwards, R., 2002. *The Phage Proteomic Tree : a Genome-Based Taxonomy for Phage*, Society, 184(16), pp. 4529-4535;
3. Notredame C, Higgins DG, Heringa J., 2000, T-Coffee: *A novel method for fast and accurate multiple sequence alignment*, Journal of Molecular Biology, 302(1), pp. 205-17.

Insights into phage endolysins

Franklin L. Nóbrega [1], Hugo Oliveira [2], Luís D. R. Melo [3], Sílvio B. Santos [4], Nuno Cerca [5], Eugénio C. Ferreira [6], Joana Azeredo [7] and Leon. D. Kluskens [8*]

*Corresponding author, [1]franklin.nobrega@ceb.uminho.pt, [2]hugooliveira@deb.uminho.pt, [3]lmelo@deb.uminho.pt, [4]silviosantos@ceb.uminho.pt, [5]nunocerca@ceb.uminho.pt, [6]ecferreira@deb.uminho.pt, [7]jazeredo@deb.uminho.pt, [8]kluskens@deb.uminho.pt

Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4700-057, Braga, Portugal

Abstract

(Bacterio)phages are viruses that specifically infect bacteria, and thus are harmless to humans, animals, and plants. They are the most abundant microorganisms on the planet (estimated to be 1031 on Earth) in a ratio of 10 times more than bacteria (1). Consequently, even rare phage-induced events are frequent at the global level. Therefore, they have a staggering ecological impact on the bacterial population and in the evolution of bacterial genomic structure upon virus-host interactions, acting as agents in the recycling of organic matter and presenting a valuable tool in molecular biology and epidemiology. Regarding the diversity of phages, they can have different types of replication mechanisms, morphologies, nucleic acid composition and genome sizes. Over the last decade improvements on phages genome sequencing and progresses in genomic research have revealed information on open reading frames of proteins of interest (2).

Increasing interest has been given to phage (endo)lysins in molecular biology, biotechnology and medicine. Lysins are phage lytic enzymes that break down the peptidoglycan of the bacterial cell wall at the terminal stage of the phage reproduction cycle, in order to release the phage progeny with the consequent death of the bacterial cells (3).

The number of phage genomes deposited in GenBank has been increasing exponentially in the last years. However, no effort has been made so far to understand the relation between lysins and their phage family and host species, presenting challenges in their annotation, comparative analysis, and representation.

The almost 700 complete phage genomes deposited in the NCBI database were searched for the presence of lysins by making use of the Pfam (4) identified domains and BLAST comparison of putative or unidentified complete genome against known lysins. In approximately 5% of the phage genomes it was not possible to identify any lysin. The identified enzymes were used to construct a phylogenetic tree with Phylip (5), using Neighbor-Joining, Maximum Likelihood and Parsimony algorithms (6). From the resulting tree, we were able to present a phage-lysin characterization network analysis taking into account the lysin aminoacid sequence and the different phage classes (Family/Genus) and host species to study their evolutionary stories. Regarding the phage families, muramidases, amidases and peptidases are the largest type of lysins in *Myoviridae*, *Podoviridae* and *Siphoviridae* phages respectively. Grouped data will also be used to identify conserved domains among lysins of different phages which will play an important role in the annotation of the still unidentified lytic cassette of phages with sequenced genomes.

References

1. Breitbart M, Rohwer F: Here a virus, there a virus, everywhere the same virus? Trends in Microbiology 2005, 13:278-284.
2. Calendar R, The Bacteriophages, Oxford University Press, 2006
3. Fischetti,V.A. 2010. Bacteriophage endolysins: a novel anti-infective to control Gram-positive pathogens. Int. J. Med. Microbiol. 300(6):357-362
4. M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn. 2012. The Pfam protein families database. Nucleic Acids Research. 40:D290-D301

5. Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.
6. Felsenstein, J., 2004. Inferring Phylogenies. Sinauer, Sunderland.

Extraction and Characterization of Biologically Relevant Relations in Biomedical Literature

Hugo Costa[1,2], Isabel Rocha[2] e Anália Lourenço[2]

[1] Department of Informatics / CCTC - University of Minho

[2] IBB - Institute for Biotechnology and Bioengineering

Centre of Biological Engineering - University of Minho

Campus de Gualtar, 4710-057 Braga - PORTUGAL

Abstract

The development of biomedical research in recent years has produced a large amount of information mostly spread throughout scientific literature that is not directly available for computational processing. Due to this wide range of scientific publications, it is hard for the community to perform tasks such as collecting, processing and analyzing relevant information in literature. Therefore, there is an important need for methods that contribute to the automation of data collecting from the literature, transforming dispersed natural language into useful knowledge, making it a valuable resource to research.

In this context, the role of Biomedical Literature Mining techniques is crucial to aid in filling the gaps between the information in literature and other data sources. Automated processes have gradually replaced manual curation, and tools for named entity recognition in text have already a considerable degree of maturity. Their application for relation extraction from the biomedical literature is the next step.

In this work, based on natural language processing systems, a conceptual model has been created containing the workflow for extraction and characterization of relations from biomedical texts. This model makes use of annotation schemes working with previously annotated texts where biological entities have been identified, combining syntactic and semantic layers.

Considering this conceptual model we developed the Rel@tioN platform that combines natural language processes from the GATE platform and AIBench, a lightweight and non-intrusive workbench for the development of applications.

Under this platform, two case studies have been considered for evaluation: first, we used a corpus described in the literature, the GENIA corpus, related to human transcription factors; the second case study comes from our research group being related to the stringent response in Escherichia coli bacteria.

Semantic Similarity in the Biomedical Domain

João D. Ferreira & Francisco M. Couto

joao.ferreira@lasige.di.fc.ul.pt (corresponding author),
fcouto@di.fc.ul.pt

Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa

Abstract

One of the most important aspects in biomedical informatics is the ability to determine whether two entities are related to each other. For instance, in genetics, similarity between genetic products is often associated with one or more functions being shared among them; in chemistry, similarity in molecular structure correlates to a similar biological role; in medicine, similarity in two clinical cases is a strong argument towards a similar diagnosis.

But finding a way to compare these entities is not trivial and strongly depends on the entities being compared. For gene products, similarity can be calculated by comparing theirs sequences; for chemical compounds, by comparing the graphs that represent their structure. While these methods have been shown to be useful, they may fail in some cases (e.g., L-serine and D-serine, although extremely similar in structure, have very different biological roles). Moreover, in some other cases, like the medical example given above, there is not an easy way to extract a similarity measure: how to generate a numeric value to the degree of similarity between two clinical cases?

This gap can be filled in by the notion of ontologies. These are machine-readable representations of knowledge, a way to make computers aware of given facts, which are, themselves, simple statements expressing a relation between concepts: for example, “an Arm is a Limb”, “a Hand is adjacent to an Arm”, or “Fever is a symptom of the Flu”. Thus, ontologies allow computers to understand the meaning behind the concept. By explicitly providing the relations between these concepts to a computer, we open up the possibility to employ computational power to automatically explore them.

One of the technologies enabled by the use of ontologies is indeed the calculation of similarity between the concepts they represent, a technology also known as semantic similarity [1]. Since an Arm is a Limb and a Leg is also a Limb, they are more similar than, e.g., an Arm and a Torso. Or, because Fever is a symptom of the Flu, the two concepts are more related to each other than Fever and Leg. By exploiting ontologies and the relations they contain, we can therefore create measures that can compare these concepts and the entities they represent. In the biomedical field, several fields have already benefited from this notion of semantic similarity, including (i) protein functions; (ii) chemical compounds; (iii) symptoms; and (iv) anatomical concepts.

Despite these previous achievements, the usefulness of this technology lies well beyond the mere capability of comparing ontological concepts. Consider the Gene Ontology, which contains a few thousand gene functions. This ontology can be used to “annotate” proteins, i.e., it can be used to explicitly state that “protein A has function X”. By doing so, proteins are brought up to the world of machine-readable semantics. Computers can leverage on these annotations to compare proteins not only by their sequence (using classical methods such as BLAST) but by their functions as well. Imagine that a clinical case is annotated with some anatomical concepts, several symptoms, and a set of altered chemical compounds in blood screenings. By using semantic similarity over all these concepts, we can achieve the goal of objectively measuring the degree of relatedness between two clinical cases, eventually leading to diagnostic and treatment of the patient.

In the course of my PhD program, I have created and applied two measures of semantic similarity in the biomedical domain: one for chemical compounds, using the ontology for Chemical Entities of Biological Interest (ChEBI) and another for anatomical concepts, using the Foundational Model of Anatomy (FMA) ontology. To evaluate the similarity measure in ChEBI, I applied it a classification problem. The results show that semantic similarity can be used to enhance the performance of predicting whether small molecules are able to cross the blood-brain barrier, thus establishing that the measure does indeed reflect biological similarity.

The work done on semantic similarity for the FMA [2] consists of a generic measure of semantic relatedness that can potentially be applied to any ontology. More interestingly, it can be applied to multiple ontologies, using for that effect links between two ontologies. For example, the concepts of the Human Phenotype Ontology (HPO), an ontology containing human symptoms, make some cross-references to anatomical concepts on FMA. By treating these as bridges between FMA and HPO, it is possible to extract relations between anatomical concepts and symptoms, thereby bringing the notion of semantic similarity to the multi-domain level.

References

1. Pesquita C, et al. (2009). Semantic Similarity in Biomedical Ontologies. PLoS Comput Biol, 5(7): e1000443.
2. Ferreira JD & Couto FM (2010). Generic semantic relatedness measure. Proceedings of the ICBO 2012.

Highlighting Metabolic Strategies using Network Analysis over Strain Optimization Results

José Pedro Pinto1;2, Isabel Rocha2, and Miguel Rocha1

[1] Department of Informatics / CCTC - University of Minho

[2] IBB - Institute for Biotechnology and Bioengineering

Centre of Biological Engineering - University of Minho

Campus de Gualtar, 4710-057 Braga - PORTUGAL

Abstract

The field of Metabolic Engineering has been growing, supported by the increase in the number of annotated genomes and genome-scale metabolic models. In silico strain optimization methods allow to create mutant strains able to overproduce certain metabolites of interest in Biotechnology. Thus, it is possible to reach (near-) optimal solutions, i.e. strains that provide the desired phenotype in computational phenotype simulations. However, the validation of the results involves understanding the strategies followed by these mutant strains to achieve the desired phenotype, studying the different use of reactions/pathways by the mutants. This is quite complex given the size of the networks and the interactions between (sometimes distant) components. The manual verification and comparison of phenotypes is typically impossible.

Here a methodology to validate in silico results though the use network topology analysis is proposed, our method is based in two algorithms the first, called simulation filtering, uses a metabolic and the results of a in silico to create a smaller network which is a "snapshot" of the metabolism in the simulated conditions, the second, called multiple topological network comparison, compares one metabolic network with a set of similar networks in order to identify the more common differences.

Our method identifies the more commune alterations that occur from the wildtype when an organism is manipulated thus highly contributing to elucidate the strategies that lead to successful mutants.

Dynamic Programming Algorithms for Biobjective Sequence Alignment

Maryam Abbasi[1], Luís Paquete[2], Miguel Pinheiro[3]

[1] CISUC, Department of Informatics Engineering, University of Coimbra, Portugal, maryam@dei.uc.pt;

[2] CISUC, Department of Informatics Engineering, University of Coimbra, Portugal, paquete@dei.uc.pt;

[3] Biocant, Biotechnology Innovation Center, Cantanhede, Portugal, Monsanto@biocant.pt;

Abstract

In this work, the multiobjective formulation of the pairwise sequence alignment problem is considered, where a score vector function takes into account simultaneously two components, similarity score and indels or gaps. Similarity score for amino acids and nucleotide sequences is obtained by a substitution matrix, such as BLOSUM or PAM, and by the subtraction of the number of matches and mismatches, respectively. This formulation is of interest to the practitioner since it allows to get rid of parameters in the usual single weighted-sum objective function and to explore a tractable set of alignments that are not reachable by any other methods [1]. In this work, explicit recurrence equations for the dynamic programming algorithms are given as well as a new pruning technique that improves the run-time. Furthermore, the algorithm is applied to 7 PAP genes of Candida clade [2] and the results are compared with those of a distance tree, found by simulating the divergence of these sequences from a common ancestor. Similar conclusions were achieved by the two methods. Extensions for multiple sequence alignment are also discussed.

References:

1. M. A. Roytberg, M. N. Semionenkov, and O. Yu. Tabolina. Pareto-optimal alignment of biological sequences. Biophysics, 44(4):565-577, (1999).
2. G. Butler et al. Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature 459, 657–662 (2009).

Detection of change points in mitochondrial DNA

Nora M. Villanueva[1], Miguel M. Fonseca[2], Marta Sestelo[1] and Javier Roca-Pardiñas[1]

nmvillanueva@uvigo.es[*], mig.m.fonseca@gmail.com,
sestelo@uvigo.es, roca@uvigo.es

[1]Department of Statistics and O. R.

[2]Department of Biochemistry, Genetics and Immunology

[*]Corresponding author

University of Vigo (Spain)

Abstract

The discovery that mutations in mitochondrial DNA (mtDNA) can cause human diseases has also increased the interest of the scientific community in understanding mtDNA evolution or its maintenance. Replication mechanism, in which the strands are exposed to an elevated mutational damage, has been described as one of the main sources of compositional bias in mitogenomes [3, 6]. In this work we present an R package [5], seq2R. With this package we could detect singularities in the nucleotide composition. Kernel smoothing techniques [7] have been implemented in order to estimate the skew profile and its first derivative. Bootstrap methods [1, 2] have been used to construct confidence intervals for these estimates, and binning techniques [4] were applied to speed up computation. In addition, this package allows plotting the estimates obtained and makes inferences about the change points (or singularities).

Keywords: mtDNA, R package, bootstrap, kernel, binning, change points

References:

1. Efron, B., 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1–26.
2. Efron, E., Tibshirani, R. J., 1993. An introduction to the Bootstrap. Chapman and Hall, London.
3. Faith J. J., Pollock D. D. 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics*, 165: 735–745.
4. Fan, J., Marron, J., 1994. Fast implementation of nonparametric curve estimators. Journal of Computational and Graphical Statistics 3, 35–56.
5. R Development Core Team. 2009. R: A language and enviroment for statistical computing. R Foundation for Statistical Computing. Viena, Austria.
6. Reyes A, Gissi C, Pesole G, Saccone C. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol*, 15: 957–966.
7. Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman & Hall, London.

Genome-Wide Analysis of Single Nucleotide Substitutions in Mycobacterium tuberculosis

Osório N.S.[1,2] (nosorio@ecsaude.uminho.pt);
Saraiva M.S.[1,2] (msaraiva@ecsaude.uminho.pt);
Pedrosa J.[1,2] (jpedrosa@ecsaude.uminho.pt);
Castro A.G.[1,2] (acastro@ecsaude.uminho.pt);
Rodrigues F.[1,2] (frodrigues@ecsaude.uminho.pt).

[1]Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal.

[2]ICVS/3B's-PT Government Associate Laboratory, Braga/Guimarães, Portugal

Abstract

The strong decrease in Tuberculosis (TB) incidence in the second half of the 20th century in many developed countries led to the erroneous assumption that TB would be rapidly eradicated. During this period the amount of research in this field declined impairing the development of better therapeutic and diagnostic tools [1]. This fact associated with the increase in the number of individuals with HIV-AIDS, diabetes or/and poor compliance to antibiotic treatment regimens contributed to a reemergence of TB epidemics. Altogether these factors have also contributed to the spread of drug resistant Mycobacterium tuberculosis strains [2]. The complete genomic sequencing of different M. tuberculosis clinical isolates [3-5] results in a rapidly increasing set of sequencing data that is now available to be analyzed in different ways. The understanding of how M. tuberculosis genetic variability contributes to the differential success of each isolate has the potential to unravel new diagnostic tools and improve TB treatment.

In this work we performed whole genome multiple sequence alignment from 62 M. tuberculosis strains from different phylogeographic lineages and with varying antibiotic susceptibility status. We compared these genomes with that of H37Rv and identified in high coverage areas a total of 3972 single nucleotide substitutions (SNS) with frequencies in the population between 5 and 80%. The vast majority of the SNS were in coding sequences. Results show that M. tuberculosis genes have in average 2.03 SNS and only 6% of the genes have 5 or more SNS. Functional annotation of the cluster with the higher number of SNS revealed the presence of a high frequency of genes involved in cell wall/membrane/envelope biogenesis. This work further explores the genetic diversity and evolutionary patterns in the population of M. tuberculosis. NSO is supported by SFRH / BPD / 33959 / 2009.

References

1. Koul A, Arnoult E, Lounis N, Guillemont J, Andries K: The challenge of new drug discovery for tuberculosis. Nature 2011, 469:483–490.
2. Hegreness M, Shoresh N, Damian D, Hartl D, Kishony R: Accelerated evolution of resistance in multidrug environments. Proceedings of the National Academy of Sciences 2008,105:13977.
3. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG: Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 1998, 393:537–544.
4. Ioerger TR, Koo S, No E-G, Chen X, Larsen MH, Jacobs WR, Pillay M, Sturm AW, Sacchettini JC: Genome Analysis of Multi- and Extensively-Drug-Resistant Tuberculosis from KwaZulu-Natal, South Africa. PLoS ONE 2009, 4:e7778.
5. Mostowy S, Onipede A, Gagneux S, Niemann S, Kremer K, Desmond EP, Kato-Maeda M, Behr M: Genomic analysis distinguishes Mycobacterium africanum. Journal of clinical microbiology2004, 42:3594.

Assembling genome-wide transporter system annotations

Oscar Dias[1], Miguel Rocha[2], Eugénio Ferreira[1], Isabel Rocha[1]

[1]IBB-Institute for Biotechnology and Bioengineering / Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

[2] Department of Informatics / CCTC, University of Minho, Campus de Gualtar, 4710-057 Braga – PORTUGAL

Abstract

Over the years, several genome-scale metabolic models have been released with compartmentation information. Models such as iMH805/775 [1] (15 compartments) and iMM904 [2] (8 compartments) for Saccharomyces cerevisiae or iRS1563 [3] for Zea mays (6 compartments) include reactions performed in specific cellular organelles, such as mitochondria, chloroplasts (in photosynthetic organisms), lysosomes, cell nucleus or the Golgi apparatus, etc. Thus, cells have specific structures, the transport systems, to assist on the metabolites relocation.

Cellular transport systems are described in databases such as TCDB (http://www.tcdb.org/) maintained by the Saier Lab Bioinformatics Group, the TransportDB (http://www.membranetransport.org/) or the YTPdb (http://homes.esat.kuleuven.be/~sbrohee/ytpdb/index.php/Main_Page) yeast transport protein database. TCDB proposed a classification system, analogous to the Enzyme Classification System (EC Number) [4] though including phylogenetic information, for transport proteins. The phylogenetic information included in the proposed nomenclature is very restrictive as it assigns a distinct TC Number to each transport protein identified in this database. Hence, in this work we propose a system to detect and classify potential transport proteins for a given genome.

A TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) [5] search was performed to identify potential transmembrane domains in the genome. Next, Smith-Waterman (SW) [6] alignments were performed on the transport candidate genes, against the TCDB database [7], to identify sequences encoding proteins with sequences similarities to the known transport systems.

Instead of assigning a similar TC number to the transport system, the classification of the metabolites is proposed, according to the frequency of the metabolite in the SW search and the taxonomy resemblance of each hit organism to the case study organism, and not the transporter systems. The membrane where the transporter is located is predicted by the WoLF PSORT (http://wolfpsort.org/) [8] software.

References

1. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novère N, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasić I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kirdar B, Penttilä M, Klipp E, Palsson BØ, Sauer U, Oliver SG, Mendes P, Nielsen J, Kell DB: A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nature Biotechnology 2008, 26:1155-1160.
2. Mo ML, Palsson BØ, Herrgård MJ: Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC systems biology 2009, 3:37.
3. Saha R, Suthers PF, Maranas CD: Zea mays iRS1563: A Comprehensive Genome-Scale Metabolic Reconstruction of Maize Metabolism. PLoS ONE 2011, 6:e21784.
4. Barrett AJ, Canter CR, Liebecq C, Moss GP, Saenger W, Sharon N, Tipton KF, Vnetianer P, Vliegenthart VFG: Enzyme Nomenclature. San Diego: Academic Press; 1992:862.
5. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology 2001, 305:567-80.
6. Smith TF, Waterman MS: Identification of common molecular subsequences. Journal of molecular biology 1981, 147:195-7.
7. Saier MH: A functional-phylogenetic classification system for transmembrane solute transporters. Microbiology and molecular biology reviews : MMBR 2000, 64:354-411.
8. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: WoLF PSORT: protein localization predictor. NUCLEIC ACIDS RESEARCH 2007, 35:W585-W587.

OPTFERM – A Computational Platform for the Optimization of Fermentation Processes

Orlando Rocha[1,2], Paulo Maia[1,2], Isabel Rocha[1], Miguel Rocha[2]

[1]IBB – Institute for Biotechnology and Bioengineering / Centre for Biological Engineering

[2]CCTC – Computer Science and Technology Center / Dep. Informatics

Universidade do Minho

Campus de Gualtar, 4710-057 Braga, Portugal

E-mails: {orocha,pmaia,irocha}@deb.uminho.pt, mrocha@di.uminho.pt

Abstract

Numerous products such as antibiotics, proteins, amino-acids and other chemicals are produced using fermentation processes. These systems are affected by biochemical and chemical phenomena as well as environmental conditions. Consequently, several computational tools have been designed and implemented for modeling, simulation and optimization, sharing a common purpose: increase the production yield of the final product.

We present OptFerm, a computational platform for the simulation and optimization of fermentation processes. The aim of this project is to offer a platform-independent, user-friendly, open-source and extensible environment for the improvement of Bioengineering processes. This tool is focused in optimizing a feeding trajectory to be fed into a fed-batch bioreactor and to calculate the best concentration of nutrients to initiate the fermentation. Furthermore, a module for the estimation of kinetic and yield parameters has been developed, allowing the use of experimental data obtained from batch or fed-batch fermentations to reach the best possible model setup. The features present in this tool allow the users to analyze the robustness of a fed-batch model, compare simulated with experimental data, determine unknown parameters and optimize feeding profiles.

The software was built using a component-based modular development methodology, using Java as the programming language. AIBench, a Model-View-Control based Java application framework was used as the basis to implement the different data objects and operations, as well as their graphical user interfaces. Moreover, this allows the tool to be easily extended with new modules, which are currently being developed.

Keywords: Fermentation processes, open-source software tools, process simulation and optimization.

Automating, Gathering and Manipulating Genetic Information at the Sequence Level

Paulo Gaspar *, José Luís Oliveira

{paulogaspar, jlo}@ua.pt

IEETA, Universidade de Aveiro

Abstract

Background*:* Numerous software applications to deal with synthetic gene redesign already exist, granting the field of heterologous expression a significant support [1, 2]. However, their dispersion forces the access to different tools and online services in order to complete one single project. Analyzing codon usage, calculating CAI, aligning orthologs and optimizing genes are just a few examples. The extent of available online services further suggests the need for an integration of tools. Furthermore, in heterologous expression, analysing all known factors that influence protein synthesis is essential to increase the chance of success.

Results*:* A software tool, EuGene, was developed to gather several main tools and web services frequently involved in gene synthetic design and analysis. In a seamless automatic form, and using only the given genes and genomes, EuGene calculates or retrieves genome data on codon usage (RSCU and CAI), codon context (CPS and CPB) [3], GC content, protein secondary and tertiary structures, gene orthologs, species housekeeping genes, performs gene alignments, and identifies gene and genome names. Moreover, the main function of EuGene is analyzing and redesigning gene sequences integrating information about hidden stop-codons, codon and nucleotide repeats, deleterious sites, GC content, codon context and usage. This is achieved using the best and fastest algorithms to maximize the resulting sequence.

Conclusions*:* EuGene is implemented with a user-friendly interface for an intuitive manipulation of gene sequences by biologists. Through the access to various online resources and offline integrated tools, numerous important aspects of a gene can be easily retrieved and/or calculated. Also, using advanced machine learning and mathematical concepts, EuGene is able to return several equivalent solutions to a gene optimization project, amenable to in vitro or in vivo testing. The auto-update and plug-in systems allows a continuous extension of EuGene functionality to automatically keep up with research requirements.

Acronyms

CAI – Codon Adaptation Index

RSCU – Relative Synonymous Codon Usage

CPB – Codon-Pair Bias

CPS – Codon-Pair Score

References

1. Welch, M., et al., Design parameters to control synthetic gene expression in Escherichia coli. PLoS One, 2009. 4(9): p. e7002.
2. Raymond, A., et al., Combined protein construct and synthetic gene engineering for heterologous protein expression and crystallization using Gene Composer. Bmc Biotechnology, 2009. 9.
3. Moura, G., et al., Large Scale Comparative Codon-Pair Context Analysis Unveils General Rules that Fine-Tune Evolution of mRNA Primary Structure. PLoS One, 2007. 2(9).

* Corresponding Author

A Generic Multi-Criterion Approach for Mutant Strain Optimization

Paulo Maia[1,2,*], Isabel Rocha[1] and Miguel Rocha[2]

[1] IBB-Institute for Biotechnology and Bioengineering / Centre of Biological Engineering, Universidade do Minho, 4710-057 Campus de Gualtar, Braga, Portugal

[2] Department of Informatics / CCTC, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

* contact: paulo.maia@deb.uminho.pt

Abstract

Motivation: The identification of genetic modifications that can lead to mutant strains that overproduce compounds of industrial interest is a challenging task in Metabolic Engineering. Evolutionary Algorithms and other metaheuristics have provided successful methods for solving the underlying *in silico* bi-level optimization problems (e.g. to find the best set of gene knockouts) [1]. Although these algorithms perform well in some criteria, they lose sense of the inner multi-objective nature of these problems.

Results: In this work, these tasks are viewed as multi-objective optimization problems and algorithms based on multi-objective EAs are proposed. The objectives include maximizing the production of the compound of interest, maximizing biomass and minimizing the number of knockouts. Furthermore, a generalization to integrate multiple-criterion capabilities into single-objective algorithms is proposed and implemented as an ensemble method. This new approach allows taking advantage of the solution space sampling capabilities of some algorithms (e.g. Simulated Annealing), while generating the set of solutions (Pareto-front) according to the multiobjective premises. The algorithms are validated with two case studies, where E. coli is used to produce succinate and lactate. Results show that this option provides an efficient alternative to the previous approaches, returning not a single solution, but rather sets of solutions that are trade-offs among the distinct objective functions.

Availability: Algorithms are implemented as a plug-in for the open-source OptFlux [2] platform available in the site *http://www.optflux.org*.

References

1. Rocha,M., Maia,P., Mendes,R., Pinto,J.P., Ferreira,E.C., Nielsen,J., Patil,K.R. and Rocha,I. Natural computation meta-heuristics for the in silico optimization of microbial strains. *BMC Bioinformatics*, **9,** 499, 2008.
2. Isabel Rocha , Paulo Maia , Pedro Evangelista , Paulo Vilaça , Simão Soares , José P Pinto , Jens Nielsen , Kiran R Patil , Eugénio C Ferreira and Miguel Rocha. OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Systems Biology* 2010, **4**:45, 2010

In silico metabolic Engineering tool for simulation and optimization of microbial strains accounting integrated metabolic/regulatory information

Paulo Vilaça[1,2,3], Miguel Rocha[3] and Isabel Rocha[2]

[1]SilicoLife Lda, Spinpark, Avepark Apart. 4152 4806-909 Guimarães PORTUGAL

[2]IBB-Institute for Biotechnology and Bioengineering, Centre of Biological Engineering,

[3]CCTC - Computer Science and Technology Center,

Universidade do Minho, 4710-057 Campus de Gualtar, Braga, Portugal.

[1]pvilaca@silicolife.com, [2]irocha@deb.uminho.pt,
[3]mrocha@di.uminho.pt

Abstract

Recently, a number of methods and tools have been proposed to allow the use of genome-scale metabolic models for the phenotype simulation and optimization of microbial strains, within the field of Metabolic Engineering (ME). One of the limitations of most of these algorithms and tools is the fact that only metabolic information is taken into account, disregarding knowledge on regulatory events. OptFlux (http://www.optflux.org) is an open-source platform that includes several tools to support in silico Metabolic Engineering (ME), including functionalities to load genome-scale metabolic models in several formats, to simulate the phenotype of both wild type and mutant strains using steady-state approaches (e.g. Flux Balance Analysis-FBA[1]) and also to perform strain optimization tasks (e.g. finding the best sets of knockouts for the production of a given metabolite).

In this work, a novel plug-in for this platform is presented, allowing the use of gene regulatory models, represented as Boolean networks. This plug-in links the regulatory model to its corresponding metabolic model, creating an integrated model and allowing its use for the phenotype simulation and strain optimization tasks. This is the first software application that allows the simulation of integrated regulatory/metabolic models.

The phenotype simulation using integrated models is conducted by firstly simulating the Boolean network and, afterwards, identifying the disabled genes in the final state and considering those as knockouts for the metabolic simulation that is conducted using FBA or alternative methods (such as MOMA or ROOM). The user can define the initial state of the Boolean network, use different environmental conditions and also simulate mutants by imposing a set of gene deletions.

The strain optimization operation provides interfaces to identify sets of genes deletions that are able to maximize a given objective function related to the production of a given metabolite with industrial interest. Two meta-heuristic optimization methods (Evolutionary Algorithms and Simulated Annealing) are used in this task[2].

References

1. K.J. Kauffman, P. Prakash, and J.S. Edwards. Advances in flux balance analysis. Curr Opin Biotechnol, 14:491-496, 2003.
2. M. Rocha, P. Maia, R. Mendes, J.P. Pinto, E.C. Ferreira, J. Nielsen, K.R. Patil,and I. Rocha. Natural computation meta-heuristics for the in silico optimization of microbial strains. BMC Bioinformatics, 9, 2008.

COEUS: Semantic Web Application Framework

Pedro Lopes and José Luís Oliveira
{pedrolopes, jlo}@ua.pt
DETI/IEETA, Universidade de Aveiro
Campus Universitário de Santiago
3810 – 193 Aveiro
Portugal

Abstract

As the “omics” revolution unfolds, the growth in data quantity and diversity is pushing forward the need for pioneering bioinformatics software, capable of significantly improving the research workflow. To cope with these computer science demands, biomedical software engineers are adopting emerging Semantic Web technologies that better suit the life sciences domain. The latter complex innate relationships are easily mapped into Semantic Web graphs, enabling a superior understanding of collected knowledge. Despite the increased awareness regarding Semantic Web technologies in bioinformatics, its usage is still diminished. With COEUS, we introduce a new Semantic Web framework targeting biomedical software developers, and comprising the skeleton for integrating, triplifying and exploring data, enabling the creation of edge-of-breed Semantic Web information systems. This framework aims at a streamlined application development cycle, adopting a “Semantic Web in a box” approach, with a package including advanced data integration and triplification tools, base ontologies, a web-oriented engine and a flexible exploration API [1]. The platform, targeted at life sciences developers [2], provides a complete application skeleton ready for rapid application deployment, and is available free as open source at http://bioinformatics.ua.pt/coeus/.

1. P. Lopes and J. L. Oliveira, "Towards knowledge federation in biomedical applications", presented at the Proceedings of the 7th International Conference on Semantic Systems, Graz, Austria, 2011.
2. P. Lopes, José, and L. Oliveira, "A semantic web application framework for health systems interoperability," presented at the Proceedings of the first international workshop on Managing interoperability and complexity in health systems, Glasgow, Scotland, UK, 2011.

^{13}C-based Metabolic Flux Analysis

Carreira R.[123], Carneiro S.[2], Villas-boas S.[3] ,Rocha I.[2] and Rocha M.[1]

[1]CCTC-Computer Science and Technology Center, Informatics Department, University of Minho, Braga

[2]IBB-Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, University of Minho, Braga

[3]School of Biological Sciences, University of Auckland, New Zealand

email: rafaelcc@di.uminho.pt

Abstract

The determination of fluxes is an important parameter to define the extent to which enzymes participate in metabolic networks and to simulate organism behaviour to various types of genetic and environmental perturbations. Metabolic flux analysis is the ultimate measurement of metabolic pathway activity during steady-state conditions, operating as a valuable tool in the detection of physiological alterations and to describe cell phenotypes. Since intracellular metabolic fluxes cannot be directly measured, available methodologies often estimate fluxes by applying mass balances around intracellular metabolites and from experimentally determined nutrient uptake and product secretion rates. ^{13}C-isotopic tracing is a technique that can also be used to measure fluxes. When cells are grown on a ^{13}C-labeled carbon substrate, the ^{13}C-labelling pattern in their proteinogenic amino acids can be determined through nuclear magnetic resonance or mass spectrometry (e.g.; GC-MS).

In this work we adapted a sensitive and high-throughput GC-MS method for measurement of metabolic flux distribution based on methylchlorofomate (MCF) derivatization to convert the amino acids into volatile compounds. Using the ^{13}C-labelling distribution of these compounds we determine the metabolic flux ratios in the central carbon metabolism. Great part of this work was the development of a flexible software to calculate flux ratios and estimate flux distribution in different metabolic models. Although our case studies are based on GC-MS coupled to MCF derivatization, this software is generic for different mass spectrometric methods.

Search for coherent gene clusters that predict invasiveness of *Streptococcus pneumoniae* strains.

Rui Catarino[1§], Sandra Aguiar[2], Mário Ramirez[2], Francisco Pinto [1§]

[1] Enzimology group, Chemistry and Biochemistry Center, Faculty of Sciences of the University of Lisbon, Edifício C8, Campo Grande, 1749-016 Lisboa, Portugal

[2] Laboratory of Microbiology, Molecular Medicine Institute, Faculty of Medicine of Lisbon, Av. Professor Egas Moniz, 1649-028 Lisboa, Portugal

[§]Corresponding author

Email addresses:

RC: rrcatarino@gmail.com

SA: siaguiar@fm.ul.pt

MR: ramirez@fm.ul.pt

FP: frpinto@fc.ul.pt

Abstract

Streptococcus pneumonia is a pathogenic bacteria responsible for several human diseases, such as pneumonia, meningitis and sepsis. It has been used as a model organism in molecular biology for more than 50 years, being historically associated with the discovery of DNA as the molecular carrier of genetic information. Nevertheless, the treatment of the pneumococcal infections is facing new challenges. Existing vaccines target a subset of pneumococcal serotypes, and serotype replacement events have been reported as a consequence of vaccine selective pressure. On the other hand, being naturally transformable, pneumococcus can rapidly acquire resistance to antibiotics. While the research for new vaccines and antibiotics is of extreme importance, the full understanding of its life cycle and invasion mechanisms may unveil new therapeutic targets.

Any pneumococcal disease is preceded by an asymptomatic colonization stage in the human nasopharynx. The transition from colonization to invasion is known to depend on both human and pathogen factors. In this work we aim to computationally identify pneumococcal genetic factors that influence the likelihood of those invasion events.

For this purpose, we analyze microarray based comparative genomic hybridization data of 72 strains of pneumococcus. The microarrays contain genes from three sequenced pneumococcal strains (TIGR4, R6 and G54) and were analyzed using a previously optimized method [1,2].

We propose to select genes that, individually or in a coordinated way, affect the frequency of invasion transitions among all colonization events, which we denominate as invasiveness. Each strain was classified as Invasive, Neutral or Colonizer according to a previous study [3] that compared the frequencies with which strains were recovered from an asymptomatic carrier or from invasive disease episodes.

To detect coordinated sets of genes, we built a network based on a distance score, calculated with two parameters: gene co-occurrence and dissimilarity of association with invasiveness. The co-occurrence was measured as a jaccard distance. A normalized association with invasiveness was accessed through the quantile of the Fisher's exact test distribution setting the dissimilarity in each pair of genes as the difference between their Fisher's normalized quantile. The distance score was then determined as the maximum of the two individual distances, assuring the coherence of close genes in both properties.

The network was then explored to find gene clusters that could predict invasiveness. Each cluster was founded with a single gene and then grown with its closest neighbors. The growth was stopped when it no longer added any predictive power to the cluster.

Resulting clusters with significant predictive power were evaluated for enrichment in gene functional annotation categories.

1. Pinto FR, Aguiar SI, Melo-Cristino J, Ramirez M: Optimal control and analysis of two-color genomotyping experiments using bacterial multistrain arrays. *BMC Genomics* 2008, 9:230.
2. Cardoso L: Classificação de genes em hibridação genómica comparativa de estirpes de Streptococcus pneumoniae. *Tese de Mestrado em Bioestatística, FCUL*, 2009.
3. Sá-Leão R, Pinto F, Aguiar S, Nunes S, Carriço JA, Frazão N, Gonçalves-Sousa N, Melo-Cristino J, de Lencastre H, Ramirez M: Analysis of invasiveness of pneumococcal serotypes and clones circulating in Portugal before widespread use of conjugate vaccines reveals heterogeneous behavior of clones expressing the same serotype. *J Clin Microbiol.* 2011, 49:1369.

In silico optimization of the production of amino-acids in Escherichia coli

Rui Pereira[1], Paulo Vilaça[1,2], Miguel Rocha[2], Isabel Rocha[1]

[1]IBB-Institute for Biotechnology and Bioengineering / Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

[2] Department of Informatics / CCTC, University of Minho, Campus de Gualtar, 4710-057 Braga – PORTUGAL

Abstract

The increasing need to replace chemical synthesis of compounds of interest by more environmentally friendly biological processes is driving the research for microbial cell factories. The industrial production of amino and organic acids includes several examples of success stories using microorganisms to convert inexpensive substrates into added value products. Traditionally, the design of such microbes relied on cycles of random mutagenesis followed by phenotypic selection [1], but a deeper knowledge of the microbial physiology allowed a more rational approach to this optimization problem [2,3]. However, this task is not straightforward, since the cell metabolism has proved to be highly complex and hard to predict.

One of the approaches to tackle this problem is to use Systems Biology simulation tools to predict the microorganism behavior when subjected to genetic modifications. Using genome scale stoichiometric models, such as the latest iAF1260 for *Escherichia coli* [4] one can simulate a great diversity of possible metabolic phenotypes under steady state conditions by imposing flux-balance constrains. The use of flux balanced analysis (FBA) allows the determination of flux values through all the reactions in the network under a set of environmental conditions and genetic manipulations, by using an objective function, such as the maximization of growth [5]. In this work, we used genetic algorithms, such as OptGene [6] to search for sets of gene knockouts that result in the overproduction *in silico* of amino-acids in *Escherichia coli.*

From all the proteinogenic amino-acids, glycine yielded the best results in the optimizations. A careful analysis of the *in silico* flux distribution in some of the mutants revealed an interesting and non-intuitive mechanism behind glycine accumulation. Furthermore, in these mutants the growth is coupled to the production of glycine, which makes them excellent candidates for *in vivo* implementation.

We are reaching a point where bioinformatics tools are advanced enough to aid in complex tasks, such as the optimization of microbial cell factories. Here we described an effort to optimize *in silico* the production of amino-acids in *Escherichia coli*, which resulted in the discovery of a potential set of knock-outs that leads to glycine overproduction. This serves to show the increasing importance of *in silico* optimizations to aid in the metabolic engineering projects, especially to search for non-intuitive beneficial genome modifications.

References

1. K. Okamoto et al. *Bioscience, Biotechnology, and Biochemistry* **61** (1997) 1877-1882.
2. A. Ozaki et al. *Agricultural and Biological Chemistry* **49** (1985) 2925–2930.
3. M. Ikeda *Applied Microbiology and Biotechnology* **69** (2005) 615-626.
4. A.M. Feist et al. *Metabolic Engineering* **12** (2010) 173-186.
5. K.J. Kauffman et al. *Current Opinion in Biotechnology* **14** (2003) 491-496.
6. K. Patil et al. *BMC Bioinformatics* **6** (2005) 308.

Computational tools for strain optimization by adding reactions

S. Correia, M. Rocha

CCTC, University of Minho, Campus de Gualtar, Braga, Portugal

e-mail: scorreia@di.uminho.pt, mrocha@di.uminho.pt

Abstract

One of the greatest challenges in Metabolic Engineering consists in the identification of sets of genes to be manipulated in order to be able to produce compounds of industrial interest in a specific organism.

This process is based on solving optimization problems that consists in finding sets of reactions to be added to a wild-type strain, as well as complementary sets of reactions to be removed, so that the strain becomes able to maximize the production of a given compound, while seeking to preserve its viability. The underlying optimization problems have found in Evolutionary Algorithms (EAs) a successful approach.

In this work, methods for the simulation and optimization of strains are proposed, allowing the addition of metabolic heterologous reactions that are external to the metabolic model making the organism able to produce the desired compound. Furthermore, these methods can be useful in the reconstruction of models, by filling gaps in the metabolic network.

These simulation and optimization methods here presented have been tested in several case studies, in which the metabolic model of E. coli is manipulated to produce vanillin, a compound with high added value and industrial interest. The referred simulation and optimization methods were implemented in the OptFlux platform, which provides several features to support Metabolic Engineering tasks.

Application of Machine Learning techniques to the discovery and annotation of Transposons in genomes

Tiago Loureiro[1], Nuno A. Fonseca[2], Rui Camacho[3]

[1,3]FEUP

[2]EMBL-EBI

[1]tiago.loureiro@fe.up.pt, [2]nunofonseca@acm.org,
[3]rcamacho@fe.up.pt

Abstract

Transposons or transposable elements (TEs) are a large class of repetitive DNA sequences that have the ability to move within a given genome. Their contribution to genome structure and evolution has generated increasing interest in developing new methods for their computational analysis.

Several methods have been developed to discover and annotate transposable elements [1] [2]. These methods can be classified in four main categories: De novo, Homology-based, Structure-based, and Comparative genomic.

De Novo methods compare several sub-sequences of a given genome and if they are repeated several times within that genome then they can potentially be transposons. The key step on this approach is to distinguish transposons from all other repeat classes. The Homology based methods sets identify transposable elements in DNA sequences by finding subsequences similar to known transposable elements. Structure based methods use prior knowledge about the common TEs structural features, such as long terminal repeats to identify potential TEs on a given genome. Finally comparative genomics methods use multiple alignments of closely related genomes to detect large changes between genomes.

The different TE detection methodologies are implemented in several software tools [3] [4] [5] [6] [7] [8] [9] [10] [11] [12]. Each tool has its own strengths and weaknesses, hence some achieve better detection rates on specific TE categories. TEs detection agreement between different tools is yet another problem as different tools can identify or not a given TE and even if identified they can disagree on its length and start and end positions within the genome.

We believe that integration of multiple approaches will further advance the computational detection of transposable elements and, based on this idea, our aim is to integrate different tools combining their strengths to improve the overall TE detection accuracy.

To achieve this we will propose the following approach:

1. Create several artificial data set of DNA sequences and their respective annotations to be processed by transposon detection tools;
2. Evaluate the transposon detection tools on the artificial datasets by gathering the transposable elements predicted and verifying the correctness by using the annotations;
3. Use the errors and correct predictions made by the annotations tools together with a characterization of the TEs and their context to assemble a Machine Learning data set;
4. Build a machine learning classifier to predict, using the features mentioned above, if a transposable element prediction is correct;
5. Use this classifier to combine the annotation tools.

We propose to combine the predictions of different tools with the expectation that accuracy can be improved. To this end we will create a model, using machine learning techniques available in

Weka [13], that can later be used to make the predictions, and, at same time, provide the user with a probability of the prediction being correct.

The features used by the model will consist of the predictions made by the TE annotation tools combined with other features that describe the DNA subsequence annotated and their context in each DNA sequence, together with the information if the annotation is correct or incorrect (based on the simulated data set annotation). The models produced by the different machine learning techniques and parameters will be compared using 10 fold cross-validation. The best model will be then used to be integrated into a tool to annotate transposable elements.

References

1. Surya Saha, Susan Bridges, Zenaida Magbanua, and Daniel Peterson. Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences. Tropical Plant Biology, 1(1):85–96, March 2008.
2. Casey M Bergman and Hadi Quesneville. Discovering and detecting transposable elements in genome sequences. Briefings in Bioinformatics, 8(6):382–392, November 2007.
3. Jerzy Jurka, Paul Klonowski, Vadim Dagman, and Paul Pelton. Censora program for identification and elimination of repetitive elements from DNA sequences. Computers & Chemistry, 20(1):119–121, 1996.
4. Zhao Xu and Hao Wang. LTR FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research, 35(suppl 2):W265–W268, 2007.
5. Mina Rho and Haixu Tang. MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. Nucleic Acids Research, 37(21):e143, 2009.
6. Yujun Han and Susan R Wessler. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Research, 2010.
7. Robert C Edgar and Eugene W Myers. PILER: identification and classification of genomic repeats. Bioinformatics, 21(suppl 1):i152–i158, 2005.
8. Zhirong Bao and Sean R Eddy. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Research, 12(8):1269–1276, 2002.
9. N Volfovsky, B J Haas, and S L Salzberg. A clustering method for repeat analysis in DNA sequences. Genome Biol, 2(8):RESEARCH0027+, 2001.
10. Green P Smit AFA, Hubley R. RepeatMasker Open-3.0.
11. Ryan Kennedy, Maria Unger, Scott Christley, Frank Collins, and Gregory Madey. An automated homology-based approach for identifying transposable elements. BMC Bioinformatics, 12(1):130, 2011.
12. F Hormozdiari, I Hajirasouliha, P Dao, F Hach, D Yorukoglu, C Alkan, E E Eichler, and S C Sahinalp. Next-generation VariationHunter: Combinatorial algorithms for transposon insertion discovery. Bioinformatics, 26(12):i350–i357, 2010.
13. University of Waikato. Weka. http://www.cs.waikato.ac.nz/ml/weka/.

Parallelizing SuperFine*.*

Diogo Neves

Abstract

The estimation of the Tree of Life, a rooted binary tree representing how all extant species evolved from a common ancestor, is one of the grand challenges of modern biology. Research groups around the world are attempting to estimate evolutionary trees on particular sets of species (typically clades, or rooted subtrees), in the hope that a final "supertree" can be produced from these smaller estimated trees through the addition of a "scaffold" tree of randomly sampled taxa from the tree of life. However, supertree estimation is itself a computationally challenging problem, because the most accurate trees are produced by running heuristics for NP-hard problems. In this paper we report on a study in which we parallelize SuperFine, the currently most accurate and efficient supertree estimation method. We explore performance of these parallel implementations on simulated data-sets with 1000 taxa and biological data-sets with up to 2,228 taxa. Our study reveals aspects of SuperFine that limit the speed-ups that are possible through the type of outer-loop parallelism we exploit.

References

1. M. Bansal, J. G. Burleigh, O. Eulenstein, and D. Fernández-Baca. Robinson-Foulds supertrees. Algorithms for Molecular Biology, 5:18, 2009.
2. B. R. Baum. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon, 41:3-10, 1992.
3. B. R. Baum and M. A. Ragan. The MRP method. In O. R. P. Bininda-Emonds, editor, Phylogenetic Supertrees: combining information to reveal The Tree Of Life, pages 17-34. Kluwer Academic, Dordrecht, the Netherlands, 2004.
4. K. Liu, T. J. Warnow, M. T. Holder, S. Nelesen, J. Yu, A. Stamatakis, and C. R. Linder. SATé-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Syst Biol, 2011. In press.
5. K. C. Nixon. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics, 15(4):407-414, 1999.
6. M. A. Ragan. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol., 1:53-58, 1992.
7. U. Roshan, B. M. E. Moret, T. L. Williams, and T. Warnow. Performance of supertree methods on various data-set decompositions. In O. R. P. Bininda-Emonds, editor, Phylogenetic Supertrees: combining information to reveal The Tree Of Life, pages 301-328. Kluwer Academic, Dordrecht, the Netherlands, 2004.
8. M. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. J Classification, 9(1):91-116, 1992.
9. M. S. Swenson, F. Barbançon, C. Linder, and T. Warnow. A simulation study comparing supertree and combined analysis methods using SMIDGen. In Proceedings of the 2009 Workshop on Algorithms in Bioinformatics (WABI), pages 333-344, 2009.
10. M. S. Swenson, R. Suri, C. Linder, and T. Warnow. An experimental study of Quartets MaxCut and other supertree methods. In Proceedings of the 2010 Workshop on Algorithms in Bioinformatics (WABI), 2010.
11. M. S. Swenson, R. Suri, C. R. Linder, and T. Warnow. SuperFine: Fast and accurate supertree estimation. Syst Biol, 2011. In press.

Prognostic prediction using clinical expression time series: towards a supervised learning approach based on meta-biclusters

André V. Carreiro[1], Artur J. Ferreira[2], Mário A. T. Figueiredo[3] and Sara C. Madeira[1]

[1]KDBIO group, INESC-ID, Lisbon, and Instituto Superior Técnico, Technical University of Lisbon, Portugal, e-mail: acarreiro@kdbio.inesc-id.pt (corresponding author), sara.madeira@ist.utl.pt

[2]Instituto de Telecomunicações, Lisbon, and Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal,e-mail:arturj@isel.pt

[3]Instituto de Telecomunicações, Lisbon, and Instituto Superior Técnico, Technical University of Lisbon,Portugal,e-mail: mtf@lx.it.pt

Abstract

Biclustering has been recognized as a remarkably effective method for discovering local temporal expression patterns and unraveling potential regulatory mechanisms, critical to understand complex biomedical processes, such as disease progression and drug response. In this work, we propose a classification approach based on meta-biclusters (a set of similar biclusters) applied to prognostic prediction. These biclusters thus represent temporal expression profiles potentially involved in the transcriptomic response of a set of patients to a given disease or treatment.

We use real clinical expression time series to predict the response of patients with multiple sclerosis to treatment with Interferon-β. The main advantages of this strategy are the interpretability of the results and the reduction of data dimensionality, due to biclustering. Preliminary results anticipate the possibility of recognizing the most promising genes and time points explaining different types of response profiles, according to clinical knowledge. The impact on the classification accuracy of different techniques for unsupervised discretization of the data is also studied.

Design of a biosynthetic pathway for curcumin production in Escherichia coli

Daniel Machado, Lígia Rodrigues and Isabel Rocha

IBB-Institute for Biotechnology and Bioengineering/Centre of Biological Engineering

University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

{dmachado,lrmr,irocha}@deb.uminho.pt

Abstract

Curcumin is the yellow pigment from turmeric, a well known culinary spice produced from the herb Curcuma longa. Research over the last years has shown that curcumin presents a wide range of pharmacological effects, including anti-inflammatory, anti-oxidant and anticarcinogenic activity. Given its potential application in cancer treatment, there is an interest for industrial production of this natural compound. This work consists on a synthetic biology approach for the design of a heterologous pathway for curcumin synthesis in Escherichia coli, a widely used microbe in industrial biotechnology. Using pathway databases and literature research we have selected the best gene candidates for heterologous expression of a curcumin synthesis pathway in E. coli. The DNA sequences for these genes were retrieved from public databases and can be readily synthesized for insertion into the host using molecular biology techniques. The inclusion of this pathway in a recent genome-scale reconstruction of the metabolism of E. coli has enabled the in silico analysis of the production capabilities for this host. We have analysed the theoretical production yields and biomass growth under different experimental conditions. Using this model we have also searched for potential gene knockouts that partially redirect the metabolic flux to the heterologous pathway without compromising cellular growth. In overall, the methods used in this work allow the selection of the most suitable combination of experimental conditions and genetic manipulations for the design of an efficient biosynthetic pathway for curcumin production in E.coli.