Poster Presentations

Expanding BioPAX format by integrating gene regulation

Irma Martinez-Flores[1,], Veronica Jimenez-Jacinto[2], Alejandra C. Lopez-Fuentes[1], Julio Collado-Vides[1]

[1]Program of Computational Genomics,
Center for Genomic Sciences, UNAM. Mexico,
[2]Biotechnology Institute, UNAM Mexico

GBS

Understanding the behavior of the cellular systems involves a complex series of biological relationships that make biological sense. Most of the biological pathways databases focus only on the metabolic pathways.

However, there are also other biological relationships that take part in cellular behavior. The different types of interactions in the cell include cell signaling pathways and genetic regulation.

RegulonDB is a database of the regulation of transcription initiation or regulatory network of the cell. It is also a model of the organization of genes in transcription units, operons and simple and complex regulons. In this sense, RegulonDB is a computational model of mechanisms of transcriptional regulation.

Currently, our laboratory is collaborating with BioPAX (Biological Pathway Exchange), which is a worldwide collaboration effort to create a format that makes the biological information exchange easier among different databases available on the Web (http://www.biopax.org). BioPAX format helps to improve the exchange of data between biological databases. We are collaborating on the expansion of the ontology that supports version 3 of BioPAX, which will include the data regarding transcriptional regulation of bacteria.

In this project, we explain the processes of mapping and exporting from RegulonDB data to BioPAX format. This allows to RegulonDB belong to the increasing group of the biological databases, including BioCyc, KEGG, Reactome, INOH and Cancer Cell Map with BioPAX format (http://www.pathguide.org). Having data available in BioPAX format makes easier for the databases to share information and provide a consistent format to facilitate integration of data from multiple sources. Also, it allows the addition of new types of data and permits the database to be compatible with other standards, such as SBML and CellML.

Translation of data from RegulonDB into BioPAX format has been a good test out of the schema. Having current RegulonDB data in BioPAX format helped us to find additional issues to improve the format; for instance, one of our more relevant contributions was to add new properties in the new class TemplateReaction of level 3, because there was not an appropriate attribute to represent direction of the transcription reaction. This improved the format, getting a better representation of genetic regulation. Also, it allowed us to create rules to validate errors or warnings that will be useful for the future in terms of validating gene regulation behavior.

Work Flow for the massive analysis of new transcription start sites in bacterial genomes

Veronica Jimenez-Jacinto[1], Enrique Morett[1], Alfredo Mendoza-Vargas[1], Ricardo Grande-Cano[1], Leticia Vega-Alvarado[2], Blanca Itzel Taboada[2], Julio Collado[3]

[1]Biotechnology Institute, UNAM Mexico,
[2]CCADET, UNAM, Mexico, [3]Center for Genomic Sciences. UNAM, Mexico

HTT

Background

This paper describes a workflow for the analysis of High throughput DNA sequence data, which allows us to identify new transcription start sites (TSS) in bacterial genomes, specifically in E. coli and Geobacter sulfurreducens.

Method

This workflow includes the following steps:

1. Filling out of the genomic information (genes, previously known promoter regions and transcription units, among others) in a relational database,
2. Alignment of the massively generated sequences against the genome of interest,
3. Classification of the information based on the characteristics of interest,
4. Incorporation to the database of the sequences obtained during the previous step, which have the possibility of being TSSs,
5. Application of mathematic tools to extract useful information from data, such as: Patterns of highly represented sequences (motifs) of codifying regions,

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

a. Analysis of distribution of sequences along the genome, c) New transcription start sites,
b. Validation of sequences with previously known TSSs,
c. Validation of computational predictions of TSSs lacking of solid evidence that justify their existence,
d. TSSs in antisense orientation,
e. Relationship of TSSs and other biological elements of major importance,
f. Identification of the gen or operon for each TSS,
g. Identification of signals that control gene expression.

6. Generation of a Web based interface for public use of their data.

Result

Development of a tool for visualizing the graphical mapping of sequences and compare it against a reference genome with applications of analysis.

Conclusions

The development of this integral system allows the management and analysis of large volumes of information in a systematic way. This development allows, with no doubt, to transform an enormous load of data, which would be otherwise manually unmanageable in highly relevant data.

BIREC: A New Web Based Bioinformatics Information Service

Nazim Rahman, Raheel Qamar, Shahid Nadeem Chohan

Department of Biosciences, COMSATS Institute of Information Technology, Islamabad, Pakistan

GBS

Background

Only two decades ago, the bottleneck in life science research was the scarcity of data where the challenge was to generate new data. Today, the challenge is to extract useful knowledge from the vast collection of data available through the Internet. Another problem which has emerged from this life science revolution is the bioinformatics user's ability to comprehend ever more complex tools, databases, new functionalities, new algorithms, updated tools, new database fields, new layouts, and other complexities. All this is changing so fast that the end-user bioinformatics books are often outdated at the time of publication. To address this issue, we have created a comprehensive Bioinformatics Information Resource and E-learning Center (BIREC) at http://www.birec.org.

Methods

In our opinion, none of the existing hierarchical resource classification methods are suitable for classifying bioinformatic resources as they show little if any hierarchy. In BIREC, we use tag-based classification. Each resource, such as a database, tool, tutorial or cheat sheet is considered a node. Each node can be assigned to multiple tags. Tags are keywords and they belong to several categories where each category describes a different aspect of the node. Since all relevant properties of a node are described by its tags, the node itself contains the instructions for its classification. Once we have nodes with tags, they can be clustered based on their tags. The clustering is dynamic and simply requires an SQL query with the relevant strings. The resulting output is displayed as a list using Views. BIREC uses Drupal Content Management System. The Content Creation Kit (CCK) module is used for custom data. Taxonomy module is used to classify with tags. Views and Taxonomy modules are used to generate lists of clustered nodes. Apart from classification, we have created fact sheets, case studies, strategies, eBooks, tutorials, cheat sheets, newsfeeds, and FAQ to facilitate understanding of the various aspects and utilization of bioinformatics resources.

Results

BIREC has been available online since January 1, 2007, and it is continuously being updated since then. Over 4000 visitors from 41 countries have viewed over 68,000 pages in the first two quarters.

Conclusions

To create a user-friendly and useful learning portal, it is essential to classify and present information in audience specific format. Tag-based classification strategy is capable of achieving this goal. However, the biggest hurdle in creating a meaningful classification is defining the tags. Experience has taught us that creating a useful tagging scheme is a step-wise process requiring planning and predefined objectives. Tags should be categorized. Each category should be created in response to an objective. When delegating tag allocation, the user should be provided with a list of keywords and their relevance in the classification.

Screening of novel inhibitors for MEK1 induced Breast Cancer-An Insilico Approach

Biplab Bhattacharjee[1,2], Jayadeepa R.M[1], Anantharamanan R[1], Samuel Sunil Pillay[2], Nirmala Kumari[2]. Sushil Kumar Middha[3]

[1]Institute Of Computational Biology Country India, [2]Brindavan College ,Bangalore, India, [3]College for Women Country India

GBS

Background

Breast cancer starts in the breast, usually in the inner lining of the milk ducts or lobules. MAPK1/3 and AKT1 belong to the serine and threonine kinase family and play important roles in estrogen induced cancer cell growth. Estrogen imbalance appears to hold the key to the understanding of breast cancer. Regulation of breast cancer growth by estrogen is mediated by estrogen receptors (ER) in nuclear and extranuclear compartments.PD98059 (MEK1Inhibitor) (http://www.arraybiopharma.com/ProductPipeline/Cancer/MEK.asp) served as a reference drug in our study. Target molecule was selected as MEK and its structure was retrieved from PDB (ID: 1s9j).

Methods

From literature, a survey of around 500 small natural molecules, which are responsible for inhibiting the biological processes important in causing cancer was taken as test drugs. Lipinski's Rule of Five was applied to evaluate drug likeness, pharmacological or biological activity on all the docked molecules. ADME Analysis was done for each molecule. Depending on Lipinski's rule, the molecules which were following the criteria for the same were subjected to receptor-ligand interaction study using ARGUSLAB docking tool.

Molecules showing a lower binding energy score, then, the reference drug was again docked in Quantum.Tox. Analysis was done comparing the screened molecules with the reference drug. The receptor ligand complex of the molecules was subjected to active site analysis in Swiss PDB Viewer (version 4.0.1) to find the amino acids contributing for the binding pocket. The hydrogen bond of all complexes was found by using this tool. Active site prediction of the receptor was done by Q Site Finder.

Results

Five natural compounds namely Caffeic Acid, Luteolin, Curcumin, Genistein, and Quercetin showed good docking scores in comparison with the reference drug(PD98059).). The docking score of the standard drug and the best performing natural molecule Curcumin was -20.91 and -27.12 in Quantum and -8.09 and -9.11794 in ArgusLab, respectively . Using ADME-TOX analysis it was found that Caffeic Acid, Luteolin, Curcumin, Genistein, and Quercetin was showing lower Ames test value than the reference molecule.

Conclusion

Curcumin, which is the principal curcuminoid of the popular Indian spice turmeric showed better ligand binding affinity towards the MEK1(Mitogen activated protein kinase kinase 1) and can play a role of potent inhibitor for the treatment of Breast Cancer. Curcumin also exhibited less toxicity effect than the study drug ((PD98059). Further animal studies need to be done to confirm the exact role and mechanism of Curcumin as a cancer chemo preventive agent for Breast cancer.

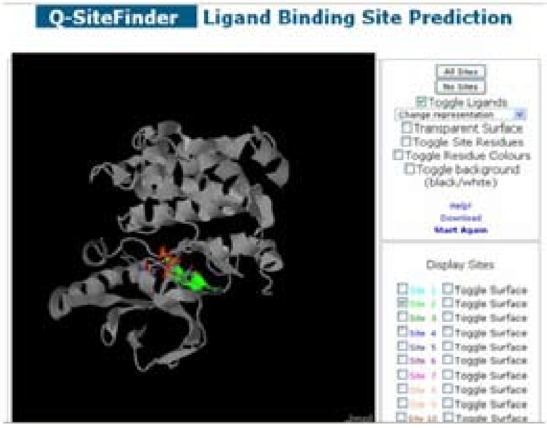
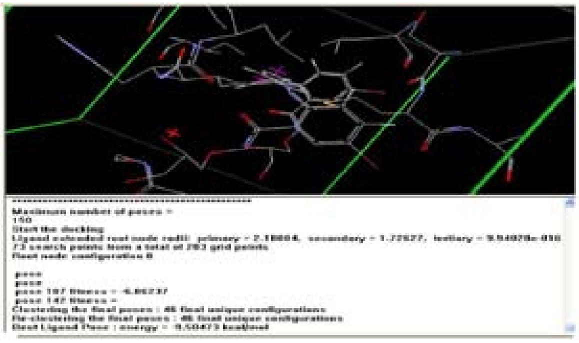

Q-SiteFinder Ligand Binding Site Prediction
All Sites
No Sites
☑ Toggle Ligands
Change representation
□ Transparent Surface
□ Toggle Site Residues
□ Toggle Residue Colours
□ Toggle background (black/white)
Help
Download
Start Again
Display Sites
□ Site 1 □ Toggle Surface
☑ Site 2 □ Toggle Surface
□ Site 3 □ Toggle Surface
□ Site 4 □ Toggle Surface
□ Site 5 □ Toggle Surface
□ Site 6 □ Toggle Surface
□ Site 7 □ Toggle Surface
□ Site 8 □ Toggle Surface
□ Site 9 □ Toggle Surface
□ Site 10 □ Toggle Surface

Fig 1. Q-SiteFinder showing the best active site



```
*****************************************************
Maximum number of poses =
150
Start the docking
Ligand extended root node radii: primary = 2.18604, secondary = 1.72627, tertiary = 9.54078e-016
73 search points from a total of 283 grid points
Root node configuration 8

 pose
 pose
 pose 107 fitness = -6.06737
 pose 142 fitness =
Clustering the final poses : 46 final unique configurations
Re-clustering the final poses : 46 final unique configurations
Best Ligand Pose : energy = -9.50473 kcal/mol
```

Fig 2. Docking of Curcumin with MEK1 in Argus Lab

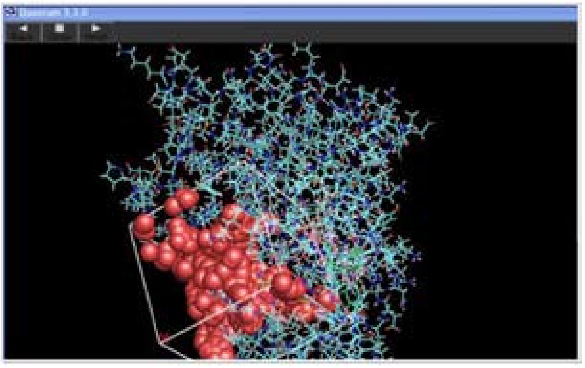
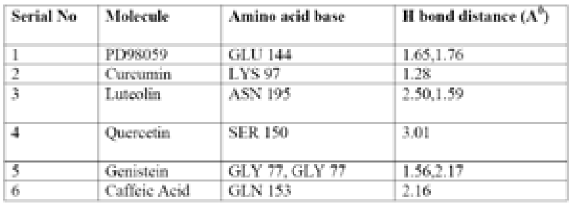
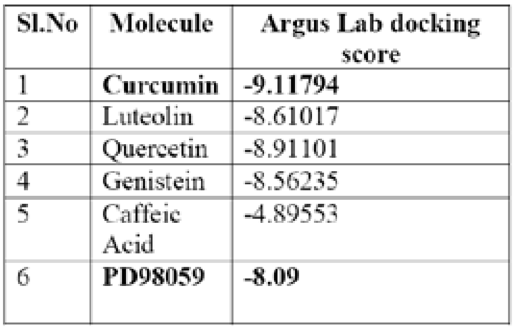


Fig 3. Docking of Curcumin with MEK1 using Quantum



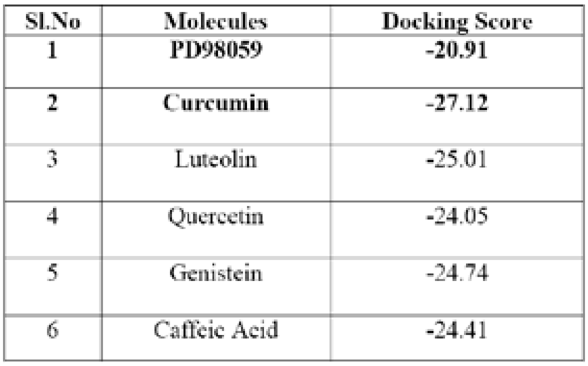
Serial No	Molecule	Amino acid base	H bond distance (A^0)
1	PD98059	GLU 144	1.65,1.76
2	Curcumin	LYS 97	1.28
3	Luteolin	ASN 195	2.50,1.59
4	Quercetin	SER 150	3.01
5	Genistein	GLY 77, GLY 77	1.56,2.17
6	Caffeic Acid	GLN 153	2.16

Table 1. Hydrogen Bond Distances of amino acid contributing to binding site with the molecules using SwissPDB Viewer.



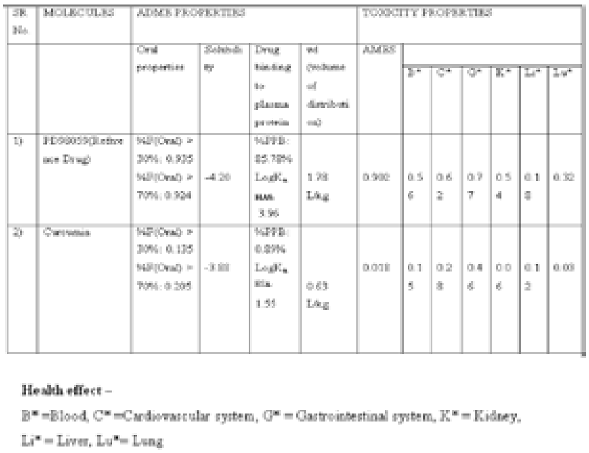
Sl.No	Molecule	Argus Lab docking score
1	**Curcumin**	**-9.11794**
2	Luteolin	-8.61017
3	Quercetin	-8.91101
4	Genistein	-8.56235
5	Caffeic Acid	-4.89553
6	**PD98059**	**-8.09**

Table 2. Docking Score of the natural molecules and the standard drug(PD98059) with MEK1 in Argus Lab



Sl.No	Molecules	Docking Score
1	**PD98059**	**-20.91**
2	**Curcumin**	**-27.12**
3	Luteolin	-25.01
4	Quercetin	-24.05
5	Genistein	-24.74
6	Caffeic Acid	-24.41

Table 3. Docking Score of the natural molecules and the standard drug(PD98059) with MEK1 in Quantum



SR No	MOLECULES	ADME PROPERTIES				TOXICITY PROPERTIES						
		Oral properties	Solubility	Drug binding to plasma protein	vd (volume of distribution)	AMES	B*	C*	G*	K*	Li*	Lu*
1)	PD98059(Reference Drug)	%F(Oral) > 30%: 0.935 %F(Oral) > 70%: 0.924	-4.20	%PPB: 85.78% LogKa HSA: 3.96	1.78 L/kg	0.902	0.56	0.62	0.77	0.54	0.18	0.32
2)	Curcumin	%F(Oral) > 30%: 0.135 %F(Oral) > 70%: 0.205	-3.88	%PPB: 0.89% LogKa HSA: 1.55	0.63 L/kg	0.018	0.15	0.28	0.46	0.06	0.12	0.03

Health effect –

B* =Blood, C* =Cardiovascular system, G* = Gastrointestinal system, K* = Kidney, Li* = Liver, Lu*= Lung

Table 4. ADMETOX Analysis of Curcumin and Standard Drug.

Efficient functional bioinformatics tools: towards understanding biological processes

Alberto Pascual-Montano, Jose Maria Carazo

National Center for Biotechnology-CSIC, Spain

OSB

Background

The post-genomics era has been characterized by the emergence of several high-throughput techniques (DNA microarrays, protein chips, chip-on-chip, etc) that have allowed researchers to study biological processes from a global perspective.

Nevertheless, the great potential of these technologies have resulted in a problem when researchers have had to deal with the vast amounts of data that are generated. In this context, the development of methodologies for the analysis and interpretation of such datasets has been a challenge for the bioinformatics community.

Methods

During the last few years our group, (http://bioinfo.dacya.ucm.es) have focused on the development of several methods and tools to assist researchers in the analysis and interpretation of genomics and proteomics data.

One of our main goals has been to provide to the research community with easy-to-access and easy-to-use tools implementing complex data analysis methodologies. In this way, we have generated a set of web-based applications that address three main fields; gene expression data analysis, functional interpretation of large lists of genes or proteins and automatic knowledge extraction from the biomedical literature.

In our effort to develop applications for a broad number of users able to deal with computing intensive problems, we have designed these tools incorporating efficient implementations of the algorithms and their variants, most of them using parallel and grid computing.

In addition, most of these tools can be accessed via web-services, which allow users to integrate them in their analysis workflows.

Results

We will review different methodologies and applications for functional bioinformatics in the context of automated data and text analysis in biology. In particular, we will focus on the following applications:

- GeneCodis (Gene Annotation Co-occurrence Discovery): http://genecodis.dacya.ucm.es [1]
- ChIPCodis (Mining Regulatory Transcription Factors): http://chipcodis.dacya.ucm.es [2]
- bioNMF (Non-negative Matrix Factorization in Biology): http://bionmf.dacya.ucm.es [3]
- SENT (Semantic Features in Text): http://sent.dacya.ucm.es [4]
- MOARA (Gene/Protein mention and normalization tool): http://moara.dacya.ucm.es

Conclusions

A new line of state of the art functional analysis applications will be presented, helping to set the scene for a review of open problems and current challenges in functional analysis.

References

1. Nogales-Cadenas R, Carmona-Saez P, Vazquez M, Vicente C, Yang X, Tirado F, Carazo JM, Pascual-Montano A: GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. Nucleic Acids Res 2009, 37(Web Server issue):W317-322.
2. Abascal F, Carmona-Saez P, Carazo JM, Pascual-Montano A: ChIPCodis: mining complex regulatory systems in yeast by concurrent enrichment analysis of chip-on-chip data. Bioinformatics 2008, 24(9):1208-1209.
3. Mejia-Roa E, Carmona-Saez P, Nogales R, Vicente C, Vazquez M, Yang XY, Garcia C, Tirado F, Pascual-Montano A: bioNMF: a web-based tool for nonnegative matrix factorization in biology. Nucleic Acids Res 2008, 36(Web Server issue):W523-528.
4. Vazquez M, Carmona-Saez P, Nogales-Cadenas R, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A: SENT: semantic features in text. Nucleic Acids Res 2009, 37(Web Server issue):W153-159.

Automatically Inferring the Regulom from Microarray Time-Series

Alexandru George Floares

Artificial Intelligence Department, Cancer Institute Cluj-Napoca Romania

IBSB

Background

The deluge of complex, high-throughput biological data is continuously increasing, and the necessity and benefits of systemic integrations become more and more evident. However, the progress is disappointingly slow. The greatest obstacle is the enormous difficulty of modeling, nonlinear, high dimensional, dynamical systems. To overcome it, efforts should be made to automate the modeling process. The final goal is to be able to understand, control and diagnose biological systems.

Methods

We outline a methodology for automatically inferring the concentration profiles of various molecular species, regulating transcription, and their mechanisms of action. It uses techniques from data mining and knowledge mining, guided by systems thinking, and it is illustrated with some of our results. Artificial intelligence (AI), as a data mining (DM) tool, is a key ingredient, being capable of automating complex model building. Knowledge mining, as a systematic use of existing knowledge bases and software instruments, greatly increases the AI power.

The systemic viewpoint dictates the kinds of models to build, how to build and analyze them, and what to expect from them. As the AI DM component of this methodology, we used RODES, a class of reverse engineering algorithms we developed for drug gene networks. RODES is based on genetic programming (GP) and neural networks (NN), and is capable of automatically building systems of ordinary differential equations (ODE), from microarray time-series data. The systemic viewpoint dictates the class of models, ODEs being considered adequate models of dynamical systems. By knowledge mining, we can find:

1. that ODEs are also adequate models for the underlying physical processes of the biochemical Networks.
2. which are the most accepted equations modeling transcription, and that regulation of transcription is crucial to the system,
3. the range or the values for the constants of the model, and
4. which are the important regulatory interactions in the networks, using software instruments like GeneGo, IPA, DAVID, CytoScape, etc.

Usually, due to various experimental constraints, essential information for networks ODE modeling is missing from data. One of the unique features of RODES is its ability to deal with the common but challenging situations of information (variables) missing from data. Due to the regulatory role of these nodes information is implicitly present in data.

Results

We proposed a methodology for automatically inferring the regulom from microarray time-series, using data mining and knowledge mining techniques, guided by systems thinking. The knowledge mining component helps finding the most probable inputs to each node of a biochemical network - tens instead of thousands, without using interactome knowledge. These drastically reduce the mathematical models searching space of the DM AI algorithms, and make them faster and scalable to high throughput data. RODES algorithms, as a DM AI component, applied to real pharmacogenomic microarray time-series data, discovered the transcription equation and its regulom for all investigated genes, with high accuracy (99.99%). The methodology considers transcription factors, drugs and drugs related compounds, and microRNAs as an equivalence class, the regulom. Thus, RODES automatically infers the regulom as hidden variables influencing microarray data, outperforming the accuracy and speed of other similar published algorithms, and scales better to high throughput data.

Conclusions

The main obstacle against systemic integration of high throughput data is the enormous difficulty of modeling high dimensional, non-linear, dynamical systems. Fortunately, modeling can be automated with the proposed methodology, combining data mining with artificial intelligence, and knowledge mining. Using RODES, the class of algorithms we developed, as the data mining artificial intelligence component, one can automatically build ordinary differential equations from time series microarray data, and even reconstruct the regulon - transcription factors, microRNAs, and drug related compounds - which is missing from data, as hidden variables.

Scaling relationship in the gene content of transcriptional machinery in bacteria

Ernesto Perez-Rueda[1], Sarath Chandra Janga[2], Agustino Martínez-Antonio[3]

[1]IBT-UNAM, [2]MRC Laboratory of Molecular Biology, [3]CINVESTAV-IPN

IBSB

Background

The metabolic, defensive, communicative and pathogenic capabilities of eubacteria depend on their repertoire of genes and ability to regulate the expression of them. Sigma and transcription factors have fundamental roles in controlling these processes. Here, we show that sigma, transcription factors (TFs) and the number of protein coding genes occurs in different magnitudes across 291 non-redundant eubacterial genomes.

Methods

In this work, we evaluated the gene content of the two main elements responsible for the regulation of transcription initiation, σ's and TFs, across 291 bacterial genomes. Families were identified based on hidden markov models of their DNA-binding domain.

Results

Our results indicate that most widely distributed families across eubacteria are small in size, while large families are relatively limited in their distribution across genomes. We also note that the diversity of extra-cytoplasmic sigma factors and TF families is constrained in larger genomes. Clustering of the distribution of transcription and sigma families across genomes suggests that functional constraints could force their co-evolution, as was observed in sigma54, IHF and EBP families. Our results also indicate that large families might be a consequence of lifestyle, as pathogens and free-living organisms were found to exhibit a major proportion of these expanded families.

Conclusions

We suggest that these differences can be explained based on the fact that the universe of TFs, in contrast to sigma factors, exhibits a greater flexibility for transcriptional regulation, due to their ability to sense diverse stimuli through a variety of ligand-binding domains by discriminating over longer regions on DNA, through their diverse DNA-binding domains, and by their combinatorial role with other sigmas and TFs. Our results suggest that understanding proteomes from an integrated perspective, as presented in this study, can be a general framework for uncovering the relationships between different classes of proteins.

Transcriptional machinery in bacteria is influenced by the domain organization of transcription factors

Nancy Rivera-Gomez[1], Lorenzo Segovia[2], Ernesto Pérez-Rueda[2]

[1]Center for Genemic Sciences, UNAM Mexico, [2]Institute of bitechnology UNAM Mexico

GBS

Escherichia coli, *Bacillus subtilis*, and *Corynebacterium glutamicum* represent three excellent bacterial model organisms for understanding diverse physiological and regulatory mechanisms in prokaryotic species. In these bacteria, the repertoire of transcription factors (TFs) has been elucidated, showing that they represent around 8% of their protein-coding genes. TFs comprise two-domain proteins, where the DNA-binding domain (DBD) has been well characterized. In counterpart, the ligand-binding and/or multimerization domain (LBD) has been loosely characterized. Here, we address the question of the degree to which TFs with winged Helix-Turn-Helix (wHTH) DBD are shared between these bacteria and how the domains beyond wHTH are influencing the regulatory response.

We searched for TFs with wHTH in the well-annotated regulatory databases RegulonDB, DBTBS and CoryneRegNet, and their domain organization was elucidated with the battery of Hidden Markov Models (HMM) deposited in Superfamily and PFAM databases. From this analysis, we found twenty-three different families of TFs, where the highest diversity of families identified resides in B. subtilis. Thirteen of these families are com-

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

mon to the three bacteria, such as LysR, MarR, and GntR regulating amino acid biosynthesis, antibiotic resistance and carbon source uptake related genes. Alternatively, seven families were identified exclusively in one of the three organisms, like PurR and Rrf2, exclusively identified in B. subtilis, suggesting likely lineage-specific events. Additionally, twenty-seven different ligand-binding domains (LBDs) were found, where 23 were identified in B. subtilis, 15 in E. coli, and 11 in C. glutamicum. From this analysis, we found specific associations between some families and their LBDs, such as the "Periplasmic binding protein II" domain and the wHTH in the LysR family. In counterpart, we identified diverse families associated with a high diversity of domains, such as GntR, IclR, and Lrp. Alternatively, and in order to identify universal and specific LBDs in a large scale, we analyzed 670 bacterial genomes, in which 85 LBDs associated to 39 different TFs families were identified. From their distribution, based on a clustering analysis using the Cluster and TreeView programs, we identified diverse groups of LBDs, those universally distributed, those LBDs exhibiting a lineage-specific distribution, and those with an erratic pattern. The results presented here describe a variety of designs of the TFs, where their LBDs allow versatility to perform multiple functions in the context of a regulatory network. From an evolutionary perspective, the TFs organization reflects the specialization of bacteria to survive in specific environments.

The necessity of clarifying concepts and terms related to transcriptional regulation in bacteria

Yalbi Itzel Balderas-Martínez, Alberto Santos-Zavaleta, Heladia Salgado, Julio Collado-Vides

Center for Genomic Sciences UNAM, Mexico

GBS

Since the general remarks of transcriptional regulation in bacteria proposed an original review (Jacob and Monod. J Mol Biol.1961; 3:318-56), new experimental knowledge is constantly being discovered and curated in databases. We have been adapting all the new data into the initial concepts, even if they do not match the original idea. This is an area in which some standardization, and codes of good practice, are needed.

A "mimicry phenomenon" can influence the scientific community in the sense that we frequently repeat what an author said without thinking if a given term is right or wrong, given that terms tend to change over time. This presented us the opportunity to update basic terms that have been used by scientists to explain similar or different concepts in their publications.

We thoroughly reviewed PubMed for classical publications related to transcriptional regulation in bacteria. We selected concepts used in Escherichia coli for it is a model that represents the major source of information of experimental data curated in databases, such as RegulonDB and EcoCyc. We compared classical concepts with the knowledge acquired through a number of analyses of the experimental information. Discrepancies were identified amongst several terms.

This could be due to a lack of monitored updates of the classical concepts with developing experimental information. There are: 1) terms to explain the same concept – the definition is very similar, 2) the same word to explain different concepts and 3) concepts that we need to update (See supplementary material). There is a necessity to standardize these terms in order to allow the formalization of theory that will be relevant to automated text mining, and a better representation of the bacterial physiology in databases. For this, we proposed definitions for Genomic Sciences and databases based on the knowledge acquired through experiments.

Mixing samples before or after expression analysis determines the final outcome

Elisabeth Tamayo[1], Antonio Muñoz[1], Rafael Fernandez-Muñoz[2], Antonio Granell-Richart[3], Oswaldo Trelles[1]

[1]Computer Architecture Department, University of Malaga, Spain, [2]Experimental Station La Mayora, Spain, [3]Fruit Genomics and Biotechnology Lab

GBS

Background

Messenger RNA samples are often pooled in microarray experiments to compensate for insufficient sample, reduce experimental costs or reduce overall variability. However, pooling

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

results in an irreversible loss of information which is in the core of recommendations for avoiding pooling at all. However, pooling can be beneficial when many subjects are pooled, provided that independent samples contribute to multiple pools. In this comparative study we determine at which extend pooling for a given condition could represent the general behaviour of a set of individual lines as a whole with the objective of identifying statistically significant changes in gene expression.

Material and methods

A complete set of experimental data obtained in the Framework of ESPSOL Spanish Project (ESPSOL Project [http://www.bitlab-es.com/espsol]) with Solanum lycopersicum has been used to obtain a set of differentially expressed genes using Prep+07 (Martin-Requena et al. BMC Bioinformatics.2009. 12; 10:16) including: a) empty spots removal, b) double-scan resolution, c) lowess adjust, d) intra/interslide replication, e) computing statistics for significance

Values (z-score and p-values)

The set was composed of 12 tomato TOM2 microarrays (http://www.operon.com/arrays/oligosets_Tomato.php) hybridized to samples representing two different experimental conditions, the first one with high levels and the second one with low levels in a character of fruit quality. A set of five biological replicates and one pool, made up of a mix of those five replicates, for each one are available. Three technical replicates for both individual replicates and pools were used.

P-values comparison

P-values coming from individual biological replicates and the pool were compared. Genes with a p-value lower than 0.05 in four of the five biological replicates were considered as significant genes and compared with genes selected at p-value <0.05 in pool set.

In a more relaxed exercise, the same experiment was performed selecting all the genes with at least one significant p-value of the five values. In addition, the experiment was made separately for the two classes (Condition 1 and 2).

Z-scores comparison

Using the genes z-scores, a two class t-test was performed to detect genes with a differential expression between both conditions. Genes with a lower t-value were compared with genes with a higher difference in log fold change (LFC) between pool from condition 1 and pool from condition 2. The analysis was performed using a difference of LFC of 2 and 1.

Results and conclusions

There is considerable disagreement about whether to pool individual samples (Kendziorski et al. Proc Natl Acad Sci. 2005; 102, 4252-4257). What implies a cost reduction or not to pool (Affymetrix (2004) Sample Pooling for Microarray Analysis (Affymetrix, San Diego), technical note).

The individual differences between pool and individual replicates are estimated using the p-value calculated for all the genes in each microarray. Results are shown at figure1.

For a global comprehension of the genes selected at any case using both strategies, z-scores were calculated and compared (figure 2).

Preliminary results show that there is an increased number of false positives, which can significantly alter the biological interpretation of the results.

A significant effect we can detect when pooling is the appearance of many empty values, due probably to the dilution of genes with very few mRNAs. To check this effect, we will make a new experiment in which genes will be separated in quartiles. Results will be available for the conference.

As it is shown in the results, pooling before the expression analysis decrease the variability in the samples, enhances the expression of genes with a higher signal (Figure 3) and reduces the expression of genes at low intensity rank (taking empty values for them) due to the mix at the same container.

Intronless Gene Database

Roddy Jorquera-Cifuentes[1,2], Rodrigo Ortiz[1,2], Carlos Wilson[3], Francisco J Ossandon[4,5], Juan Pablo Cárdenas[4], David S. Holmes[5]

[1]Center for Bioinformatics and Genome Biology and Facultad de Ciencias Biológicas, [2]Universidad Católica de Chile, [3]Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, [4]Center for Bioinformatics and Genome Biology, Fundación Ciencia para la Vida and Facultad de Ciencias Biológicas, [5]Universidad Andrés Bello, Santiago, Chile, [4]Center for Bioinformatics and Genome Biology, Fundación Ciencia para la Vida, Santiago,

Chile, [5]Center for Bioinformatics and Genome
Biology, Fundación Ciencia para la Vida,
MIFAB and Facultad de Ciencias Biológicas,
Universidad Andrés Bello, Santiago, Chile

GBS

Background

Eukaryotic genes are usually interrupted by introns. However, an increasing number of intronless genes (InGs) are being discovered. Although some of these are pseudogenes and are probably not functional, many have been demonstrated to be expressed, raising questions as to their origin, evolution and function. In order to address these questions on a massive comparative genome scale, it is necessary to construct a database of InGs. Such a database exists but it is not freely accessible to the public. Therefore, we decided to construct a new, publicly available, curated, searchable database which will be placed soon in a public domain and continually curated with quarterly updates of new genomes.

Methods

48 sequenced eukaryotic genomes were downloaded from the NCBI web page, including 13 vertebrates, 3 plants, 6 insects and 23 eukaryotic micro-organisms. Using Perl scripts and BioPerl Application Programming Interface (API), CDS were selected and those nucleotide sequences whose length matched that of the corresponding mature mRNA were selected for further study (InGs). InGs were classified into orthologs and paralogs using OrthoMCL and by potential function using KOG. The information was stored in a relational database built with My SQL Server 5.1.33.

Results

The database provides information on the occurrence, properties and genomic distribution of 148,127 InGs out of a total of 638,835 predicted genes from 48 eukaryotic species – or 23% intronless genes. The distribution of InGs ranges from only 2.6% in Caenorhabditis elegans to 94.6% in Saccharomyces cerevisiae. In the human genome, histones and G protein-coupled cell surface receptor genes (GPCRs) and other genes involved in signaling pathways. (The latter account for 803 out of a total of 4102 InGs = 19%) are particularly enriched in InGs. It has been suggested earlier that histone genes are predominantly intronless in order to expedite their rapid synthesis during DNA replication, but explanations for the frequent occurrence of InGs in the other functional categories now need to be sought. Of the 4102 InGs in human, 1308 (32%) are present in two or more copies per human genome and 1573 (38%) do not have orthologs in other species suggesting that they may have arisen from recent genetic events. However, annotation errors cannot be excluded in this analysis and the database will be a useful aid for the depuration of bad annotation.

Conclusions

The creation of a database of InGs provides an opportunity to pose questions relating to the origin, evolution and function of such genes. It will provide useful information for the “introns early” versus “introns late” debate. It could reveal special functional categories demanding a biological explanation and it could serve as a useful tool to improve genome annotation by comparative genome analysis.

Assessing regulon diversity in the acidithiobacillus genus by comparative genomics

Jorge Valdés[1], Julio Collado-Vides[2], David S. Holmes[1]

[1]Center for Bioinformatics and Genome Biology,
Fundación Ciencia para la Vida, and Facultad
de Ciencias Biológicas, Universidad Andrés
Bello, Santiago, Chile, [2]Center for Genomic
Sciences, UNAM, Mexico

IBSB

Background

Members of the acidithiobacillus genus are characterized by their ability to survive in extreme acidic environments (pH 1-3) and to derive energy from inorganic sources such as iron and reduced inorganic sulfur compounds.

These extreme features, in addition to enhanced heavy metal resistance, have enabled acidophilic microorganisms to be used for the recovery of metals (e.g. copper and gold) and for the bioremediation of polluted soils.

An analysis of the genome sequence of Acidithiobacillus ferrooxidans ATCC23270 and the recently sequenced genomes of A. thiooxidans and A. caldus have determined exclusive and shared functional modules involved in the biogeochemical cycling and nutrient assimilation of carbon, nitrogen and metals. To obtain a more detailed understanding of the regulatory architecture of these functional modules at the systems biology level, we conducted regulon comparative analysis in order to identify potential transcription factors and targets and their predicted response to environmental changes.

Methods

Pathways tools, RegulonDB and RSATools were used in order to manage the information, predict known transcription factors, regulated genes and transcription factor binding sites. Resulting information was manually curated and integrated with the experimental information available.

Results

A regulon network for each microorganism has been predicted and regulatory network architectures exclusive and shared among the three representatives of the acidithiobacillus genus have been identified. The main differences observed in the three genomes are in nitrogen metabolism, motility and chemotaxis, sulfur assimilation and hydrogen utilization.

Conclusions

A comparative assessment of regulon architecture and diversity in the acidithiobacillus genus provides a more comprehensive picture of metabolic variability and adaptation in extreme environments and helps to unravel their evolutionary history. The knowledge of regulon architecture forms the basis for the prediction of potential gene targets and binding sites for future experimental validation. It is also valuable in assessing responses to environmental changes and will pave the way for future systems biology approaches.

Plasmodium falciparum Chloroquine Resistance (Pfcrt) Mechanisms: an Intra-Erythrocytic Developmental Stage

Marion Olubunmi Adebiy[1]

[1]Covenant University, Niger

IBSB

Chloroquine (CQ) cheap and long history antimalaria has failed in the treatment of malaria. This work, therefore, is sought to expose the resistance mechanism(s) of Plasmodium falciparum (Pf) at the Intra-erythrocytic developmental stage.

By considering the activity involved at this stage and reviewing polymorphism within the food vacuolar membrane protein Pfcrt, chloroquine resistance polymorphism at that level will be determined.

The biochemical network of P.f and the gene expression data were downloaded from the genebank, NCBI, EMBL, plasmoDB and geneDB.

The data were performed as confirmed by the Blast and ClustalX programme using NCBI blast against the biochemical network of Pf, and mapped onto the enzymatic reaction nodes of the metabolic network.

The result shows that there was a variation in the targeted metabolic pathways of the erythrocytic cycle, likewise, the genes that codes for the enzymes of the metabolic pathways.

These methods give a better understanding of how resistance process occurs, as well as the important mechanisms that P.f deplores for resisting these anti-malaria drugs.

The knowledge therefore, facilitates the rationale to design new, effective and well tolerated antimalaria drugs.

In-silico Prediction of the Genetic Regulatory Interactions in Maurer's Cleft Pathway of Plasmodium falciparum

Itunuoluwa Marian Ewejobi[1], Svetlana Bulashevska[2], Benedikt Brors[2], Ezekiel Femi Adebiyi[1]

[1]Covenant University, Niger, [2]Dept. of Theoretical Bioinformatics. German Cancer Research Centre (DKFZ), Heidelberg, Germany

GBS

Background

For over a century, the significance of the discovery of Georg Maurer remained undiscovered but recent works show that Maurer's clefts are appreciated as a novel type of secretory organelle. Established by the malaria parasite within its host cell, Maurer's clefts play an essential role in directing proteins from the parasite to the erythrocyte surface [1]. Its intermediary role in the export of protein from the parasite across the cytoplasm of the host cell to the erythrocyte surface has recently caught scientific interest.

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

This is due to the fact that erythrocytes lack secretory organelles found in other eukaryotic cells as a result the parasite cannot rely on the host cell its protein needs. Therefore, it must establish de novo a secretory system in the host cell cytoplasm, in a compartment outside of its own confines. The generation of such a protein secretion system, which is extracellular from the parasite's perspective, is a remarkable accomplishment [1]. This work aims at providing more computational insight into the modalities of regulation of the genes found in the Maurer's cleft pathway of Plasmodium falciparum (P. falciparum).

Methods

Presently, about ninety three (93) genes are known that belong to the Maurer's cleft of P. falciparum from the dataset. This work takes a close look at the identified genes in this pathway and attempts to elucidate the genetic regulatory connections. We applied the Bayesian inference of the probabilistic model for reconstruction of the genetic regulatory interactions from microarray data.

This model produced great results when previously developed and tested on yeast S. Cerevisiae [2] and the glycolysis and apicoplast pathways of Plasmodium falciparum [3]. A Bayesian network model for a genetic network can be presented as a directed acyclic graph (DAG) with N nodes. The nodes may represent genes or proteins and the random variables Xi levels of activity [4].

Results

From a total of ninety three (93) genes, the regulations for fifty three (53) genes were found. Furthermore, quite a number of interesting groups of genes working together as well as interesting regulators can be seen, and we modeled a simultaneous gene activities map of regulatory interactions of these genes showing each gene with its corresponding activator and inhibitor. Using the query tool "Predicted Functional Interaction" from PlasmoDB, a couple of the predicted interactions were validated. We also hope to validate these predicted interactions biologically from available literature soon.

Conclusion

We have been able to predict functional interactions among genes in the Maurer's clefts of P. falciparum. This result will no doubt help biologists in the quest to understand the functionality of this important pathway

References

3. Friedrich Frischknecht and Michael Lanzer "The Plasmodium falciparum Maurer's clefts in 3D" , Molecular Microbiology (2008) 67(4), 687–691
4. Bulashevska, S. and Eils, R. 2005. "Inferring genetic regu regulatory logic from expression data". Bioinformatics, 21(11):2706–13.
5. Bulashevska, S., Adebiyi, E. F., Brors, B., and Eils, R. "New insights into the genetic regulation of Plasmodium Falciparum obtained by Bayesian modelling". Gene Regulation and System Biology, 1, 137-149, 2007
6. Jason T.L. Wang, Mohammed J. Zaki, Hannu T.T. Toivonen and Dennis Shasha "Data Mining in Bioinformatics" Springer-Verlag London Limited 2005

Portal of Practical Bioinformatics: Education of Bioinformatics in Slovakia

Matej Stano, Lubos Klucar

[1]Institute of Molecular Biology SAS, Slovakia

GBS

Background

Bioinformatics education is only partially established in Slovakia. Few independent groups at universities and Slovak Academy of Sciences are engaged in education and research in the field of bioinformatics. However, there is no university in Slovakia offering bachelor's or master's programme in bioinformatics.

Teaching of bioinformatics at Faculty of Natural Sciences, Comenius University (the largest university in Slovakia) is covered by two one-term courses: the first during the bachelor's degree study (since 1997) and the second in master's (since 2006). Courses are recommended above all for students of molecular biology, genetics and biochemistry.

The authors of this abstract lead lectures and exercises within these courses. In order to make educational process more interactive and efficient, we decided to make use of e-learning features and to develop a Web portal intended for teaching practical bioinformatics.

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

Methods

The educational Web portal is built up on well-proven LAMP platform (Linux, Apache, MySQL and PHP) and it is located on the server of Slovak EMBnet node.

Results

We created Portal of Practical Bioinformatics (PPB) - an educational Web portal dedicated to the teaching of bioinformatics in Slovakia (http://www.embnet.sk/edu/ppb/index.php?lang=en). PPB provides background information as well as practical exercises in one place. Its content is divided into five main sections: (i) theoretical lessons, (ii) practical exercises, (iii) problem tasks, (iv) test forms and (v) external links. Lessons guide students through the main topics and questions of bioinformatics.

Exercises make students familiar with basic bioinformatics methods and workflows. On the other hand, practical tasks represent advanced and more complex problems. Tests allow students to prove their knowledge in bioinformatics and links section cross-links the content to significant bioinformatics databases and tools. All sections of PPB are available in two languages, Slovak and English.

PPB is regularly used in practical lessons. User interface contains individual tasks that should be done as well as answer form where students should submit results of exercises. In the administrative interface, teachers may edit content of PPB, record presence of students during exercises and evaluate students' answers and results of practical tasks.

Conclusions

Slovakia is still lagging behind the other EU countries in organising bioinformatics communities. To be successful in our research activities, we need proper education in this branch of science. It is obvious that dynamic and ever evolving bioinformatics leads the battle also in introducing the new concepts to the educational process. In order to improve study of practical bioinformatics, we develop PPB and continuously make an effort to supplement, improve and actualize its content. The concept of PPB is one of the first attempts of this kind in the context of academic education in Slovakia.

Improving the Metabolic Pathway Alignment with Genetic Algorithm

Patricia G. Ortegon-Cano[1], Ernesto Perez-Rueda[2], Katya Rodríguez-Vázquez[1]

[1]IIMAS, UNAM, Mexico, [2]Biotechnology Institute, UNAM, México

GBS

Diverse computational methods have been developed for the sequence alignment problem. One of these methods is evolutionary computing, a subfield of artificial intelligence based on the idea of evolving solutions, implementing mechanisms that emulate nature. Genetic algorithms (GA) are one of the most popular techniques of evolutionary algorithms. They offer a clearly separation between the evaluation criteria (objective function) and the optimization process. Nowadays, the availability of wealth biological information allows the analysis of the interaction between different entities, and in particular metabolic pathways. The comparative analysis of different metabolic pathways aims to identify similarities among them and metabolic pathways of diverse organisms, which provide insights for identification of alternative pathways, and phylogenetic reconstruction, among others.

In this work, the pairwise alignment problem for metabolic pathways is considered from the enzymes perspective to find evolutionary relationships between different metabolic pathways and metabolisms. The proposed method is divided into two sections (Fig. 1); in the first one, the enzymes belonging to a particular pathway are transformed to sequences, whereas in the second section, these sequences are then aligned by a GA. The 64 maps of E. coli metabolism retrieved from KEGG pathway database were used as study case. These pathways were transformed to sequences by using Breadth First Search algorithm that infers the closer neighbor considering a common substrate. Because we are interested in the enzymes that are involved in these reactions, the compounds that are not product of any reaction were considered as initial nodes. The first three levels of Enzyme Commission (E.C.) numbers were considered to represent enzymes as a string or sequence. A database containing all these strings was constructed to be used in a posterior step to perform all against all comparisons. In order to generate and maximize the sequence alignments, we proposed a GA that uses a binary codification, where "0" represents enzymes and "1" gaps, where an individual represents a possible alignment which is evaluated using a score matrix for the possible matches and mismatches, highly penalized. The gap insertion is also highly penalized. The algorithm tries to find the best alignment with the maximum score.

Experiments were conduced with alignments of segments from a same pathway, identifying regions of metabolic pathways sharing a similar succession of E. C. steps, suggesting common catalysis, that are non-trivial to identify with traditional computational tools. Nowadays, we are working on the algorithm for multiple sequences alignment to identify diverse pathways with modular segments. The GA proposed has shown efficient results providing a good alignment in less than 100 iterations. These results allow us to make inferences about the evolutionary process in the metabolic pathways.

Stable transcriptional states in Escherichia coli: a sketch of its transcriptomic landscape.

Enrique Balleza Davila[1], Agustino Martínez-Antonio[2], Julio Collado-Vides[1]

[1]Center for Genomic Sciences, [2]CINVESTAV-IPN Mexico

IBSB

Transcriptional activity is modified in response to environmental/genetic changes within the restrictions of an underlying regulatory network. The full extent and variety of these modifications consistent with the network and environmental/genetic changes is currently unknown.

We find that, even existing a huge quantity of environmental/genetic variations, Escherichia coli adapts and responds with a much more restricted set of stable transcriptional states. We show this by analyzing E. coli's transcriptional activity in 242 different experimental conditions finding only 74 different stable transcriptional patterns that reproduce almost completely the total expression range of all genes.

Differences among conditions inducing the same transcriptional pattern are, mainly, genetic manipulations.

Complementarily, transcriptional patterns of cultures in different growth media or in different growth phases are, most probable, likely different. Transcriptional activity is even more constrained: there are large sets of genes with a constant expression across many different patterns. Degeneracy of transcriptional activity might be a source of organismal dynamic robustness.

Bilayer conformation changes induced by the antibiotic peptide MccJ25 binding: new insights in its mechanism of action

Torres Bugeau Clarisa, Avila César, Dupuy Fernando, Morero Roberto, Chehín Rosana

Instituto Superior de Investigaciones Biologicas, Chacabuco, 461- (4000) Tucuman, Argentina

GBS

Background

Microcin J25 (MccJ25) is a 21 aminoacid peptide active against Escherichia coli and Salmonella enteritidis strains. The structure of the peptide was elucidated based on mass spectrometry and nuclear magnetic resonance showing a distinctive lasso-structure. Convincing evidence that RNA polymerase is the main target for MccJ25 in E. coli was provided by our laboratory. In addition, the peptide activity on cell and model membranes has also been demonstrated. Characterization of the interaction of the peptide to bilayers is of central importance to understand its membrane activity. In order to elucidate the peptide binding to membranes, fluorescence experiments were performed with the MccJ25 I13W mutant, which shares both structural and microbiological characteristics with the wild-type peptide.

Methodology

All simulations were performed using the NAMD package and the CHARMM force field, specially adopted for lipids. A general protocol developed by Woolf and Roux was used to construct the initial configuration of the protein-membrane system. Simulations were carried out with a time step of 1 fs, with imposed 3D periodic boundary conditions, in the NPT ensemble with a semisotropic pressure of 1 bar. The bilayer models were then placed into rectangular boxes and solvated to a final size of 88x 80x75. Analysis of MD trajectories was performed using VMD. Infrared and flu-

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

orescence spectroscopy were used to validate the model obtained by MDS.

Results

Experimental studies were complemented with molecular dynamics (MD) simulations of MccJ25 I13W embedded on dipalmitoyl phosphatidylcholine (DPPC) bilayer. According to our results, the peptide was capable to penetrate into DPPC membranes via its ß-hairpin while the N-terminal eight-aminoacid ring, as well as threaded the C-terminal fragment stay adsorbed on the bilayer surface resulting in a stable complex. The maximal depth of insertion was observed for the residue G12 located at about 5,5 Å from the bilayer center. The membrane response to the peptide insertion was also evidenced from de MD trajectories since the simulations revealed bilayer changes as response to the peptide binding. The average thickness of the lipid bilayer was significantly enhanced in comparison with pure DPPC. Moreover, the order parameter of acyl chains of lipids was increased.

Conclusions

These results suggest that upon binding to membranes, MccJ25 could to induce the formation of ordered lipid domains. Considering that packing defects in boundaries of different fluidity domains have been proposed, this would explain the increment in membrane permeability observed using experimental techniques.

A computer simulation model of vector population replacement based on the Maternal-effect dominant embryonic arrest (MEDEA).

Mauricio Guevara[1], Edgar Vallejo[2],

[1]Computer Science Department, [2]ITESM, Mexico

HTT

Background

Creating effective mechanisms for controlling vector borne diseases like malaria and dengue is a major epidemiological concern worldwide, especially in developing countries. Genetic modification of organisms to confer disease refractoriness to them is now possible due to advances in modern experimental techniques. Natural selection alone is not likely to produce the rapid spread of a new gene in the wild, so complementary molecular mechanisms, such as MEDEA, have been recently proposed to expedite this process. Considering the life spawn of most disease vectors, studying their population genetics experimentally would be extremely difficult, so mathematical and computer simulations models are used. MEDEA is a biological mechanism that is used to favor the survival of offspring that possess a particular collection of genes. MEDEA consists of a toxin, an effector and an antidote. The offspring that inherits the toxin and the effector, but not the antidote will die. There is much hope that MEDEA would contribute to the spread of a disease resistant gene rapidly in wild vector populations.

Methods

The biological processes incorporated in our model are genetic mutation, migration and reproduction of a simulated vector population. In addition, we introduced the MEDEA mechanism during reproduction. Vectors are represented as haploid DNA sequences and they reproduced sexually using random mating. The computer model includes a turnover parameter to avoid unbounded growth of the population.

Results

The use of transposable elements (TEs) to confer immunity to a vector population was explored in early attempts to achieve population replacement. TEs often contribute to diminish the fitness of the carrier so this complicates the rapid spread and prevalence of modified genes in wild populations. We conducted a comprehensive collection of computer simulations and we found that with approximately 20% proportion of MEDEA vectors in the population, a 100% of gene fixation was consistently achieved so we deduced from these results that the MEDEA disease control strategy has more chances to succeed than its TEs counterpart.

Conclusions

Computer simulations would be useful as a preliminary approach for studying of studying disease control strategy based on population replacement and to identify a set of conditions to be fulfilled to make them effective. There is still a lot of studies and research to close the gap between the reality and the abstractions presented here. In spite of the later, we believe computer simulations are capable of modeling fundamental aspects of many biological mechanisms with a reasonable accuracy and will be increasingly useful for supporting the study of fundamental questions on the biology of organisms. In addition, with the advent of high-throughput sequencing technologies, combining computational and experimental approaches for studying population genetics is foreseeable, in our opinion.

Hydrophobicity and protein structure

Marta Bunster

[1]Universidad de Concepción, Chile

GBS

The concept of Hidrophobicity has been subject of many studies. Nevertheless, there is still controversy in their definition. In general it is considered as a relative value assigned to chemical groups or compounds associated to the possibility of interaction with water; in proteins it has been considered the entropic component of folding.

Physical chemists have been publishing hydrophobicity listings or the twenty natural amino acids and other compounds, based on solubility or atomic properties among others. Tanford published hydrophobicity indexes for some amino acids, completed later by Jones.

This index involved the water and ethanol partition coefficient for each amino acid, simulating the exterior and the interior of a protein. Later, Ponnuswamy proposed that, although the hydrophobicity for each residue was a precise value, its behavior in the protein context would reveal the true hydrophobic character. Based on that statement, he proposed the concept of environmental hydrophobicity or bulk hydrophobicity(<Hf>), which was the average of the hydrophobicity of one amino acid and the hydrophobicity of the amino acids contained on a sphere of 8A. 24 proteins were considered for this study. The result was a distribution of hydrophobicity for each amino acid. Later Cid, Bunster and collaborators repeated the calculation for 64 proteins clustered by structural class. Considering this small data base, different values of bulk hydrophobicities were obtained for each structural class.

The Protein Data Bank contains >60000 protein structures and a review of the relationship between <Hf> and structural classes was performed using a filtered database with a similarity <30% and resolutions better than 2,5A. This filtering procedure produced 7656 structures, clustered as 1207 αα 1209 ββ, 1576 α+β and 1838 α/β. The same analysis was performed according to domain folding, relating them with internal packing of proteins.

In all cases the <Hf> of the amino acids, considering the α/β structural class showed the higher values. In this study, a comparison among the values obtained for different structural classes is shown. Other structural clusterization such as folding and architecture of domains are also considered.

References

Cid, H., Bunster, M., Arriagada, E., and Campos, M. (1982) FEBS Letters 150, 247-254.

Cid. H., Bunster, M., Canales, M., and Gazitua, F.(1992) Prot. Eng. 5(5): 373-375.

Ponnuswamy, P. K., Phrabharakan, M., and Manavalan, P. (1980) B.B. Acta 623: 301-316.

Tanford, C. (1963) J. Am. Chem. Soc. 84: 4240-4247.

In-Silico Approach to Tracking and Controlling the Spread of Plasmodium falciparum within the Anopheles gambiae Mosquito mid-gut

Olugbenga Oluseun Oluwagbemi

Department of Computer and Information Sciences (Bioinformatics Unit) Covenant University Nigeria

GBS

Background

Malaria constitutes a major problem within Sub-Sahara Africa. Malaria parasites interactions in/at the mosquito mid-gut exhibit the characteristics of a complex adaptive system, which implies that some malaria parasites experience death, due to activity of some genes within the mosquito mid-gut, during this developmental cycle. Some other malaria parasites, however, survive and experience migration and proliferation. The aim of this research is to apply a Genetic algorithm approach to track and control malaria parasites spread in/at the mosquito mid-gut. This will act as an alternative strategy to malaria control.

Genetic algorithms are adaptive search algorithms, based upon the principles of evolution and natural selection. A key component of evolution is natural selection. Organisms less suited to their environment tend to die off. Organisms that are more suited to their current environment are most likely to survive. These surviving organisms produce offspring that will have many of the better qualities possessed by their parents. As a result, these children will tend to be "more suited" to their environment. These children will be more likely to survive to mate in future generations. This is analogous to Darwin's "survival of the fittest" theory, which is also analogous to the complex adaptive system characteristic exhibited by the interactions of the malaria parasites in/at the mosquito mid-gut.

Genetic algorithms possess the ability to search large and complex search spaces to efficiently determine near optimal solutions in reasonable time frames by simulating biological evolution, hence its choice for this work. Furthermore, Genetic algorithms have been successfully applied to determine the optimal path traveled by the traveling salesman in the 'Traveling Salesman Problem'.

Methods: Technical aspects of the methodology involved the implementation of the Genetic algorithm as an approach to tracking and controlling malaria parasites spread in/at the mosquito mid-gut using the Java Programming language. The Net bean IDE (Integrated Development Environment) 6.5 was used for this purpose. This methodology employed the use a computational programming approach to implement the genetic algorithm in Java Programming language. Modeling the activities of malaria parasites, using agent based models and software was also carried out. Work is still going on in this direction, as a complementary method to tracking their malaria parasites (Plasmodium falciparum) spread within the mosquito mid-gut.

Results

Preliminary results obtained showed (i) the possibility of tracking and obtaining the optimal path traveled by malaria parasites from one stage to the other, thus, providing opportunity to attack the parasites before they reach their final stage of maturity for onward transmission.(ii) the possibility of monitoring the velocity of propagation of the malaria parasites as they travel through the mosquito mid-gut, thus providing opportunity for slowing down their speed.

A Java programming approach of the dynamic interactions of malaria parasites (as agents) with their environment as a means of slowing down velocity of propagation of the parasites and ultimately achieving transmission blocking

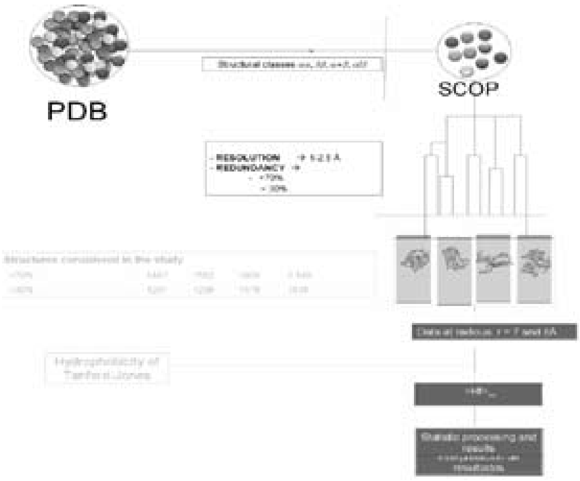

PDB
SCOP
Structural classes
- RESOLUTION
- REDUNDANCY
Structures considered in the study
Hydrophobicity of Tanford-Jones
Statistic processing and results

Figure 1. Flow diagram of the process

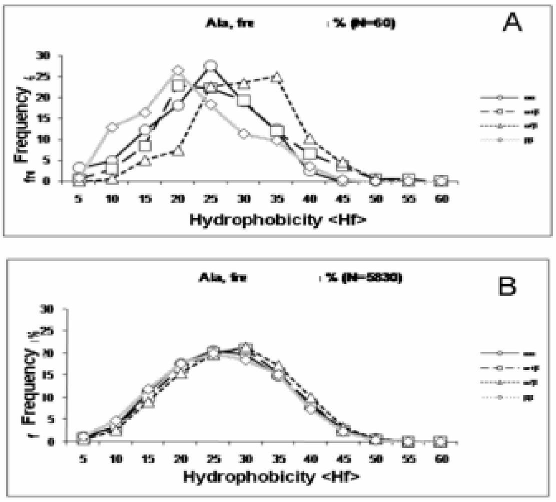


Figure 2. Representative result of the distribution of bulk hydrophobicity for Alanine using a data base of 64 proteins (A) and 7656 structures (B).

Conclusion

It is possible to gain insight towards achieving transmission blocking of malaria parasites within the mosquito mid-gut by applying the knowledge of high-level programming and computational concepts like Agent based model.

Reference

Osta , A.M, Christophides G. K., Vlachou D., and Kafatos F.C.,(2004), Innate immunity in the malaria vector Anopheles gambiae: comparative and functional genomics, The Journal of Experimental Biology,207,2552-2563

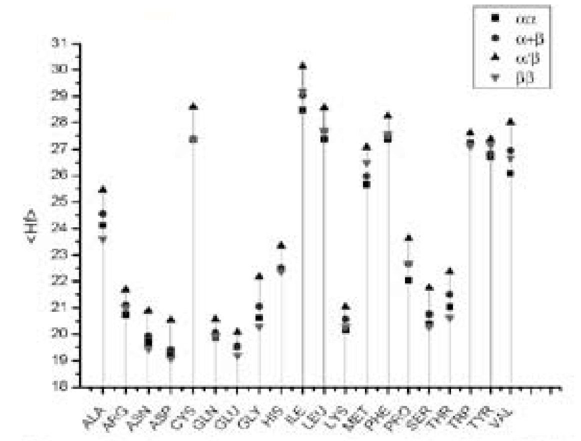
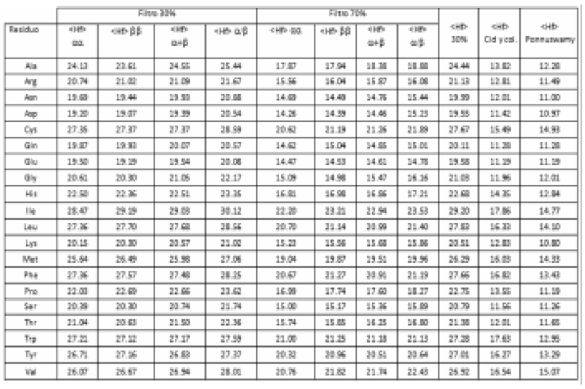


Figure 3. Schematic Representation of the Mean Bulk Hydrophobicity for the 20 amino acid residues.



Residuo	Fitro 30%				Fitro 70%				<Hf> 30%	<Hf> Cid y col.	<Hf> Ponnuswamy
	<Hf> αα	<Hf> ββ	<Hf> α+β	<Hf> α/β	<Hf> αα	<Hf> ββ	<Hf> α+β	<Hf> α/β			
Ala	24.13	23.61	24.55	25.44	17.87	17.94	18.38	18.88	24.44	13.82	12.28
Arg	20.74	21.02	21.09	21.67	15.56	16.04	15.87	16.08	21.13	12.81	11.49
Asn	19.68	19.44	19.93	20.88	14.68	14.49	14.76	15.44	19.99	12.01	11.00
Asp	19.20	19.07	19.39	20.54	14.26	14.39	14.46	15.23	19.55	11.42	10.97
Cys	27.35	27.37	27.37	28.59	20.62	21.19	21.26	21.89	27.67	15.49	14.93
Gln	19.87	19.93	20.07	20.57	14.62	15.04	14.85	15.01	20.11	11.28	11.28
Glu	19.50	19.19	19.54	20.08	14.47	14.53	14.61	14.78	19.58	11.19	11.19
Gly	20.61	20.30	21.05	22.17	15.09	14.98	15.47	16.16	21.08	11.96	12.01
His	22.50	22.36	22.51	23.35	16.81	16.98	16.86	17.21	22.68	14.35	12.84
Ile	28.47	29.19	29.03	30.12	22.28	23.21	22.94	23.53	29.20	17.86	14.77
Leu	27.36	27.70	27.68	28.56	20.70	21.14	20.99	21.40	27.83	16.33	14.10
Lys	20.15	20.30	20.57	21.02	15.23	15.56	15.68	15.86	20.51	12.83	10.80
Met	25.64	26.49	25.98	27.06	19.04	19.87	19.51	19.96	26.29	16.03	14.33
Phe	27.36	27.57	27.48	28.25	20.67	21.27	20.91	21.19	27.66	16.82	13.43
Pro	22.03	22.69	22.66	23.62	16.99	17.74	17.60	18.27	22.75	13.55	11.19
Ser	20.39	20.30	20.74	21.74	15.00	15.17	15.36	15.89	20.79	11.56	11.26
Thr	21.04	20.63	21.93	22.36	15.74	15.85	16.25	16.80	21.38	12.01	11.65
Trp	27.21	27.12	27.17	27.59	21.00	21.25	21.18	21.13	27.28	17.63	12.95
Tyr	26.71	27.16	26.83	27.37	20.32	20.96	20.51	20.64	27.01	16.37	13.29
Val	26.07	26.67	26.94	28.01	20.76	21.82	21.74	22.43	26.92	16.94	15.07

Table 1. <hf> coefficients for each amino acid residue considering their structural class and their redundancy.

Ezekiel Adebiyi, Gbenga Oluwagbemi and Seydou Doumbia, (2008), Modeling the Malaria parasite-Mosquito mid-gut cell interactions, Rocky Mountain Conference, U.S.A

Abraham, E.G., Islam, S., Srinivasan, P., Ghosh, A.K., Valenzuela, J.G., Ribeiro, J.M.C., Kafatos, F.C., Dimopoulos, G., and Jacobs-Lorena, M.,(2004), Analysis of the Plasmodium and Anopheles transcriptional repertoire during ookinete development and mid-gut invasion. Journal of Biol. Chem., 279(7) , 5573-5580.

Christophides, G. K., Zdobnov, E.,Barrillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P.T., Collins, F.H., Danielli, A., Dimopoulos, G., Hetru, C., Hoa, N.T., Hoffman, J.A., Kanzok, S.M., Letunic, I., Levashina, E.A., Loukeris, T.G., Lycett,G., Meister, S., Michael K., Moita L.F., Mueller H., Osta, M.A., Paskewitz, S.M., Reichhart, J., Rzhetsky, A., Troxler L., Vernick, K.D., Vlachou, D., Volz, J.,Von Mering, C., Xu, J., Zheng L., Bork P., Kafatos, F.C., (2002), Immunity-related genes and gene families in Anopheles gambiae; Science, 298, 159-

Computational Identification of functional related gene in Malaria

Oyelade Olanrewaju Jelili[1], Ezekiel Adebiyi[2], Benedict Brors[3], Roland Eils

[1]Covenant University, [2]Department of Computer and Information Sciences Covenant University, Niger

GBS

Plasmodium falciparum, the most severe form of malaria, causes 1.5-2.7 million deaths annually, mostly in Africa.

The most commonly used computational method for analyzing microarray gene expression data is clustering. This has been used by LeRoch et al.; 2003, and Bozdech et al.; 2003.

The results obtained have been used to classify genes into functional modules, namely, metabolisms and pathways.

The results obtained have left us with many putative functional genes.

Experimental results in the Hagai database (accessible also from www.plasmodb.org) provides limited information about this.

Recent work like Gangman Yi, Sing – Hoi Sze and Michael R. Thon; 2007 and Young et al.; 2008 introduce the use of Gene Ontology but the results are also still very limited in their application to Plasmodium falciparum (Oyelade et al.; 2008).

In this work, for the first time, with improved precision, we identify functional modules (i.e. groups of functional related genes and protein) using genomics-transcription factors and high throughput, large scale data, such as transcriptomic, proteomic and metabolic data.

Cleaning, assembling and annotating public sunflower ESTs sequences to create a curated unigene database to support gene expression studies

Paula Fernandez[1], Marcelo Soria[2], Dario Principi[3], Santiago Delfino[3], Ana Conesa[4], David Blesa[4], Joaquin Dopazo[4], Ruth Amelia Heinz[1], Norma Paniego[1]

[1]Instituto de Biotecnología INTA Castelar, Argentina, [2]Facultad de Agronomía, Universidad de Buenos Aires, Argentina, [3]Facultad de Ingeniería, Universidad de Buenos Aires,

Argentina, [4]Centro de Investigación Príncipe Felipe, Valencia, España

GBS

The presence of low-quality sequences in public ESTs databases affect the quality of unigene assemblies, which in turn negatively affect the design of expression-data based microarrays due to their sensitivity to probe specificity.

Methods

A set of 133,682 Helianthus annuus ESTs Genbank was downloaded. Then, they were screened for the presence of remnants of cloning or sequencing vectors using the UniVec database [1] and applying a BLASTN analysis optimized for short matches. Contaminating sequences located at either end of the EST were trimmed. ESTs containing contaminating vector sequence in the middle region were discarded. Ambiguity-rich (N's) regions on both ends of the sequences were trimmed using the trimseq program from the EMBOSS suite [2] and the poly-A or poly-T tails on the 3' and 5' ends were clipped using EMBOSS-trimmest. Finally, ESTs that were less than 100-bp long after cleaning were discarded. The CAP3 program [3] was used to assemble the cleaned ESTs. The orientation of the unigenes was estimated using BLASTX against the RefSeq protein database [4], the best hit was chosen and used to infer the correct orientation of each unigene. The sequences in the final assembly were analyzed with the Blast2GO program [5] (with the blastx option against the nr database) to infer electronic annotations and build a database of GO terms [6], Interpro domains [7] and KEGG links [8].

Results

Sequences were assembled with CAP3 into 28,089 singletons and 12,924 contigs. Using the Blast2GO program we could annotate 22,000 unigenes. A preliminary scan of the GO terms for biological processes and KEGG annotations showed that most of the main biochemical pathways are represented in our assembly.

Conclusions

We show the construction and validation of a unigene set for Helianthus annuus from public EST sequences. This unigene database will be used to design an oligonucleotide microarray to study the transcriptional profile of different sunflower accessions under biotic and abiotic stress conditions.

References

1. http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html
2. Rice P, Longden I, Bleasby A: EMBOSS: The European Molecular Biology Open Software Suite Trends in Genetics 2000, 16:276-277.
3. Xuang X, Madan A: CAP3: a DNA sequence assembly program. Genome Research 1999, 9:868-877.
4. http://www.ncbi.nlm.nih.gov/RefSeq/
5. Conesa A, Götz S, García Gómez J, Terol J, Talón M, Robles M: Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005, 21:3674-3676.
6. Ashburner M, Ball CA, Blake JA, Botstein D, Buttler H, Cherry JM, Davis AP, Dolinsky K, Dwight SS, Eppig JT: Gene Ontology tool for the unification of biology. Nature Genetics 2000, 25:25-29.
7. Hunter S, Apweiler R, Attwood T, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al: InterPro: the integrative protein signature database Nucleic Acids Research 2009, 37:D224-228.
8. KEGG: The Kyoto Encyclopedia of Genes and Genomes.

Ethnic Polymorphism Evaluation Tool (e-Pet)

Mariam Nersisyan, Deendayal Dinakarpandian

University of Missouri, Kansas United States

GBS

The analysis of sequences across different ethnic groups helps to identify features that are universally conserved in humans, as well as polymorphisms that might underlie racial differences in susceptibility to disease. Multiple sequence alignment methods can be used directly to detect regions of conservation within a set of sequences. In contrast, one needs to compare multiple sequence alignments to detect regions that are differentially conserved. The Ethnic Polymorphism Evaluation Tool (E-Pet) is an online tool that can be used to compare two sets of sequences. E-Pet is integrated with sequence data from the SNP500 Cancer Project and thus can be used to compare gene sequences of deidentified humans to highlight polymorphic differences between

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

African/African- American, Caucasian, Hispanic and PacificRim ethnic groups. Optionally, one may also explore the functional consequences of the conserved differences

Method

The input of the algorithms is a pair of ethnic groups chosen by the user, together with optional cutoff values for the degree of difference. Frequency matrices are constructed for the base composition at each SNP position for each gene for each ethnic group (percentage of the nucleotide at a particular position in an ethnic group). A differential metric is used to label each position. The simple metric chosen is the sum of the absolute values of the frequency differences of the nucleotides at each position. This is computed in two different ways. The first takes the frequency of each base separately, while the second considers the frequency of purines versus pyrimidines. Depending upon the cutoff values chosen, a given position within the gene may be considered significantly different, different or not at all different.

Results

We first carried a pairwise comparison across all possible pairs of ethnic groups. Then, we found all the distinct SNPs for each ethnic group by overlapping all pairwise comparisons involving that group (the overlap for Hispanic ethnic group was empty). After we ranked the SNPs in each group based on our metric and narrowed our analysis by picking only the top 5th percentile. For each gene we calculated the number of SNPs it has represented in each group and in the original dataset. This allowed us to measure how represented the gene was in each group. Caucasian group had only 2 genes C1S and RIPK2 that were represented by 20%, the rest were less then 10%. And since our cutoff was at least 50% this group was leftover. In African/African _ American group we found the following genes LOC389641 (100%), IKBKB and SLCO1B3 (80%), C4BPB, GSTA2 and PCTP (67%), IRF3 (62%), TP53BP1 (60%), POLB, CDKN1B, FCGRT, GSTT1, HEPH, MAOA, MTHFD2, RHOBTB2, RIS1, SLC22A8, SLC40A1, TMEM158, TRAF6 (50%) and in Pacific _ Rim FUT2 (100%), GSTM3 (67%), FES, TCTA, SDF4, MDN1, FMO5, FLJ33167, CCDC111 (50%). The project URL is http://134.193.129.29/ePet/ATP/index.php

Comparative large scale analysis of genes regulated by an homeoprotein in metazoarian genomes: feasibility & barriers.

Lucia Nikolaia López Bojórquez[1], Julian Esquivel Márquez[1], Julio Collado Vides[1], Carlos Valverde Rodriguez[2]

[1]Program of Computational Genomics, Center for Genomic Sciences, UNAM. Mexico [2]Instituto de Neurobiología UNAM, México

GBS

The genome wide identification of Transcription Factor Binding Sites (TFBS) in the regulatory regions of complex eukaryotes and the corresponding detection of genes targeted by particular transcription factors is helping to uncover the complex circuitry of transcriptional networks. Although significant advances have been made regarding the in-silico discovery of conserved TFBS, a comparative analysis aimed at describing changes in the transcriptional regulatory circuitry across species has barely been employed.

This project combines a genome wide detection of target genes for the homeotic protein Nkx2.1/TTF-1, essential in the development and function of the vertebrate thyroid axis, with a cross-species comparative analysis using representative metazoan genomes: human, Mus musculus, Gallus gallus, Danio rerio and Ciona intestinallis. The project is based in a position weight matrix (PWM) search, within a promoter collection. Most of the matrices required for this analysis are available in TRANSFAC and the promoter regions were collected from TRANSPRO (both accessible in BIOBASE). The analyses were performed using RSAtools, a suite of specialized programs for detecting regulatory elements.

Given the sequence and structural complexity of vertebrate genomes, our effort has been focused on the construction of high-quality matrices. For this purpose, the matrices were trained using different sets of vertebrate orthologs of well known thyroid genes (thyroglobulin, pendrin and thyroid peroxidase). After every training round, an analysis with matrix quality program was performed (see the work by A. Medina etal in this conference). At the moment, our matrices are capable of discriminating among different ortholog gene sets at the vertebrate class level.

The next step includes a search of genes co-regulated by transcription factors Pax-8 and TTF-2, in order to detect specific thyroid genes in the compared species. Additionally, the alignment or phylogenetic footprint from the thyroid genes 5´UTRs will yield clues about the evolution in the underlying regulatory networks in this gland, and their correlation with the developmental and phenotypic effects.

Integrative prediction of bacterial gene regulatory networks

Alejandra Eugenia Medina Rivera[1], Heladia Salgado Osorio[1], Julio Collado Vides[1]. Jacques van Helden

[1]Center for Genomics Sciences, UNAM, Mexico, [2]ULB, Belgium

GBS

Within the different regulatory systems co-existing in an organism, transcriptional regulation is one of great relevance. For this reason, modelling this system has become one major area in bioinformatics research. The transcriptional regulatory system modulates gene expression by means of several components present on the upstream-regulatory sequence of a gene: promoters, Transcription Factors (TF) binding sites, enhancers, etc.

The "Grammatical model of gene expression"1 integrates all the necessary elements for transcription initiation and its principal aim is to understand the complexity of regulation of gene expression in order to predict new elements involved in this process. Data to integrate the grammatical model of gene expression has been curated and stored at RegulonDB2. Nowadays, this database contains the most complete experimentally derived gene regulatory network of any living organism, that of Escherichia coli K-12.

TFs are proteins with a major part in transcriptional regulation, these proteins conform an important element in the grammatical model. Even with the amount of curated data, information for all TF Binding Sites (TFBS) is missing, still for a well curated organism as E. coli; therefore prediction and validation of TFBSs in gene promoters has become a major subject in bioinformatics, although, methods predicting new putative TFBSs have shown a high false positive rate.

In order to increase the predictive power of methods for TFBSs detection, we propose a new pipeline to predict and validate putative binding sites; this pipeline is focused on improving common approaches used to detect a TFBS. The proposed pipeline goes from analysing the quality of one of the most common detection methods of putative binding sites3, to validation based on phylogenetic conservation of TF-gene regulatory interaction.

References

1. Collado-Vides. Grammatical model of the regulation of gene expression. Proc Natl Acad Sci USA (1992) vol. 89 (20) pp. 9405-9
2. Gama-Castro et al. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Research (2008) vol. 36 (Database issue) pp. D120-4
3. Medina-Rivera, et al. Empirical and theoretical evaluation of transcription factor binding motifs. In prep.

Mapping oncogenes and tumor suppressors genes for cancer progression

Juan Emmanuel Martínez Ledesma[1], Victor Manuel Trevino Alvarado[1]

[1]Computer Science Department, Bioinformatics Group Tecnologico de Monterrey Campus Monterrey, Mexico

GBS

Background

The emergence of cancer is due mainly to alterations in 3 types of genes: oncogenes, tumor and metastasis suppressors, and stability genes (Vogelstein et al. Nat. Med. 2004; 10(8), 789-799). There is no known case where the alteration of one gene causes cancer (Vogelstein et al. 2004). Conversely, is widely accepted that only certain alterations in several genes simultaneously cause cancer and that these combinations are specific to every tissue (Wood et al. Science 2007; 318(5853), 1108-1113). To detect genes related to cancer stages, other groups including Oncomine have used adaptations of t-test. However, the differences in gene expression levels have to be drastic in order to be detected. For these reasons, we propose to map cancer-related genes to cancer stages using a hypergeometric test and a sliding-window spanning the continuous expression values.

Methods

We used cancer gene expression experiments from the GEO repository that includes cancer stage and at least 40 samples. To determine cancer-related genes, we used MSKCC. To determine the p-value associated to genes, we used the following pipeline: imputation of missing values using k-nearest-neighbors with k=15, log base 2 transformation whenever needed, quantile normalization, and filtering for cancer-related genes. For a given gene, values were sorted irrespective of its cancer stage. A moving-window was used to test whether the number of samples of each class was unexpectedly high. The window size was determined by the number of samples for each class. For every gene and stage, an h-statistic was designed taking the least hypergeometric p-value window. The p-value of the h-statistic was estimated by 10,000 sample stage permutations.

Results

In two prostate cancer datasets, we observed 244 and 289 cancer-related genes from 773 obtained from MSKCC. According to (Bigler et al. Prostate 2007; 67(14):1590-9) we correctly map the RAP2A to the primary cancer stage in prostate cancer, because they show that Rap2 is involved in androgen-mediated transcriptional and growth responses of human prostate cancer cells. We also map the androgen receptor gene to the metastasis stage in prostate cancer and agree with (Niu et al. PNAS 2008; 105(34):12182-7). They demonstrated that the prostate AR may function as both a suppressor and a proliferator to suppress or promote prostate cancer metastasis. Our results show that some of the p-values calculated with our algorithm differ from the obtained with a t-test and f-test.

Conclusions

We observtved better concordance of detected genes than using other methods. Therefore, the detection may be method-specific and requires considering different hypotheses to detect genes related to cancer stages. In addition, the inexistence of definitive cancer-related genes is a serious problem. We expect that our results may help to solve this issue.

Methionine Exploration of Trypanosomes Software Tool

Pilar Bulacio[1], Laura Angelone[1], Luis Esteban[2], Esteban Serra[3], Elizabeth Tapia[1]

[1]CIFASIS-Conicet Institute, Rosario Argentina, [2]Facultad de Ciencias Médicas, Santa Fe, Rosario Argentina, [3]Facultad de Ciencias Bioquímicas y Farmacéuticas, Suipacha 531, Rosario, Argentina

OSB

Background

The automatic identification of Translation Initiation Sites (TISs) remains a challenging problem for gene prediction. Briefly, standard tools such as Glimmer, MED-Star and GSFinder are based on looking for open reading frames with a statistically significant minimal length, which may work on prokaryotic sequences but not in eukaryotic ones (Gopal et al. Nucleic Acids Res. 2003, 31:5877–5885). A feasible justification is that scores derived from a trained statistical model considers only coding regions information. These scores make sense only for highly compact bacterial (prokaryotic) genomes, with high frequency of coding sequences. However, coding regions in protozoa frequently represent less than 10% of the genome (El-Sayed et al. Science 2005, 309:404–409), giving misleading training sets.

MET is a computational tool for TISs prediction in Trypanosomes. Its main goal is the simplicity and accuracy of its TIS prediction method. MET architecture, based on GSEA, consists of LOAD data, LEARN, PREDICT, ANALYSIS HISTORY, and REPORTS modules. The core process is done by PREDICT module which implements a heuristic that requires a knowledge model to classify sequences into CODing and NO CODing. Such a model can be inferred using LEARN module (AdaBoost DS). The final result is a ranking of potential TISs and corresponding p-values.

Methods

PREDICT module implements classification and exploration tasks. The first step classifies the input sequence S into COD, NO-COD subsegments (COD if $Pi > 0.5$ or NO COD if $Pi < 0.5$). If COD subsegments exist, the process starts with the parsing of S taking into account the first 10 ATGs: $Si \leftarrow S(ATGi,\dots,ATGn)$, $n \leq 10$. Inside this subsequence set, potential TISs are searched by a pruning process: potential TIS are those ATGs

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

preceding two coding subsegments, allowing one gap, i.e., COD-COD or COD-NO COD-COD. Once potential TISs have been identified, a MET score according to the probabilities of the initial classification is associated with each subsequence: Mi ← Prod p(Si,j), with j=1 to n. Finally, the statistical analysis (permutation tests) evaluates the most reliable TISs.

Results

The Trypanosoma Cruzi organism is analysed. T. Cruzi sequencing projects (http://tritrypdb.org) search regions of the DNA associated with the Chagas disease requiring the TIS discovery for gene identification.

TIS identification results with MET in T.Cruzi sequences are shown within and embedded browser consisting of a table with candidate TIS positions and p-values, and a graphical view of raw coding scores from the core AdaBoost classifier. The graphical MET output can be used for supplementary TIS inspection.

Conclusion

The availability of user friendly software is an important issue in current Bioinformatics research. TIS prediction with MET just requires a well-curated dataset of COD and NO COD sequences. As a result of its data-driven approach, MET may be well suited for TIS prediction in hard to analyze genomes like T.Cruzi.

Identifying Acidic Similarities in Retro-Transcribing Viral Proteomic Sequences using an Evolutionary Clustering Technique

Ramiro Garza-Dominguez, Antonio Quiroz-Gutierrez

Universidad Autónoma del Carmen, Mexico

GBS

According with Albert Szent-Györgyi hypothesis [1], the phenomena of double electronic mobility in biology has a special relevance in evolution and biomedicine. He proposed that such a pathology as cancer more than a disease itself is result of an inherited double electronic behavior in life molecules. The important process of electronic desaturation opened the way to development, differentiation and evolution. According to Szent-Györgyi investigations [1], the amino acid Lysine in proteins, plays a fundamental role in the process of electronic desaturation in life molecules. In [2], a Lysine-Arginine concentration analysis on a set of Retro-Transcribing viral proteomic sequences was presented. In this work, an Aspartic-Glutamic acid concentration analysis is described as a complementary study.

The Retro-Transcribing viruses are a group of viruses with an interesting and special feature: they contain a reverse transcription stage in their replication cycle. This reverse transcription process is highly error-prone, resulting in the introduction of many genetic mutations. This high mutation rate promotes genetic recombination and fast genetic variation. There are three taxonomical viral families that involve reverse transcription: Hepadnaviridae, Caulimoviridae and Retroviridae. The viral sequences for this experiment were selected through the Entrez Retrieval System.

Only two of the twenty amino acids in proteins have positive electric charge: Aspartic acid and Glutamic acid. These two are classified as acidic amino acids. An Aspartic-Glutamic acid concentration vector is calculated from the viral sequences and analyzed to identify correlations among species. The computational methodology is based on the descriptive data mining task of cluster analysis. The well known K-Means algorithm is used as the basic mechanism to partition the data into disjoint sets of points. A search strategy based on Evolutionary Programming is incorporated, in order to optimize the cluster structures generated by the K-Means algorithm, as described in [2]. In figure 1, the main steps of the computational strategy and the five-cluster structure are shown.

Experimental results show a number of interesting and unexpected similarities. In the context of the Retro-Transcribing viruses, it can be said that the viruses that infect vertebrates are characterized by a lower concentration of Glutamic acid compared to the viruses that infect plants. The members of the Orthohepadnavirus genus and the Deltaretroviruses show a very low concentration of Asp and Glu, specially the Human Hepatitis B virus. The viruses that infect birds, from different families and genera, show a close correlation to the Letiviruses in their Asp concentration. The Caulimoviruses are characterized by a high concentration of Asp. There are a number of unexpected similarities related to the Alpharetroviruses, as shown in figure 1. These

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

similarities could suggest bioelectronics relationships among viral proteomes.

References

1. Szent-Györgyi A., "The living state and cancer", National Foundation for Cancer Research, 1979.
2. Garza-Domínguez R. and E. Bautista-Thompson, "Finding Bioelectronics Correlations in Retro-Transcribing Viral Proteomic Sequences using an Evolutionary Clustering Technique", IEEE CPS Press, 2009

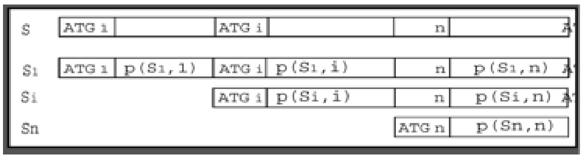


Figure 1.

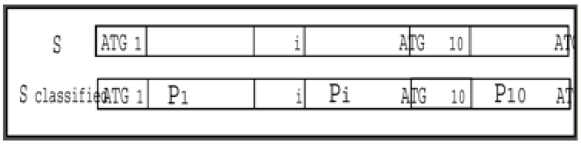
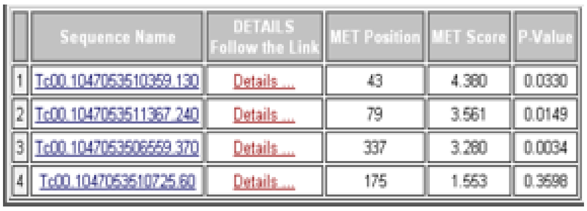


Figure 2.



	Sequence Name	DETAILS Follow the Link	MET Position	MET Score	P-Value
1	Tc00.1047053510359.130	Details ...	43	4.380	0.0330
2	Tc00.1047053511367.240	Details ...	79	3.561	0.0149
3	Tc00.1047053506559.370	Details ...	337	3.280	0.0034
4	Tc00.1047053510725.60	Details ...	175	1.553	0.3598

Figure 3.

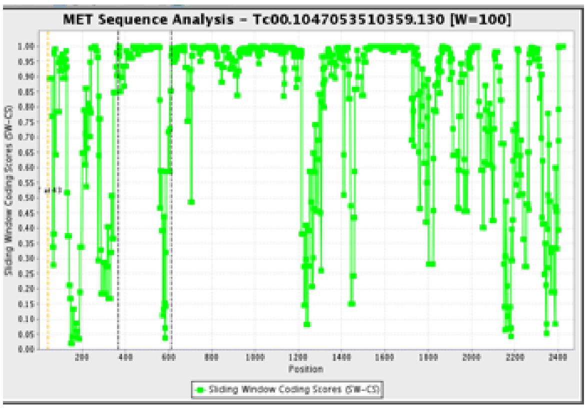


Figure 4.

Functional atlas of uncharacterized proteins in Escherichia coli

Sarath Chandra Janga

MRC Laboratory of Molecular Biology, University of Cambridge, United Kingdom

HTT

Background

One-third of the 4,225 protein-coding genes of Escherichia coli K-12 remain functionally unannotated. Many map to distant clades like Archaea, suggesting involvement in basic bacterial traits, while others appear restricted to E. coli, including pathogenic strains.

Results

To elucidate their biological roles, we performed an extensive proteomic survey using affinity-tagged E. coli strains and generated comprehensive genomic context inferences to derive a high-confidence compendium for virtually the entire proteome consisting of 5,993 putative physical interactions and 74,776 putative functional associations, most of which are novel. Clustering of the respective probabilistic networks revealed putative orphan membership in discrete multiprotein complexes and functional modules, while a machine-learning strategy based on network integration implicated the orphans in specific biological processes. We provide additional experimental evidence supporting orphan participation in protein synthesis, metabolism, cell adhesion and motility, and assembly of the bacterial cell envelope.

Conclusions

This resource provides a 'systems-wide' functional blueprint of a model microbe, with insights into the biological and evolutionary significance of previously uncharacterized proteins.

Comments

This study is a large scale analysis for the prediction of functions of uncharacterized genes in the bacterial genome, E. coli. It has been recently published in PLoS Biology. Please see the reference for further details about the work.

Reference

Hu P, Janga SC, Babu M, Díaz-Mejía JJ, Butland G, Yang W, Pogoutse O, Guo X, Phanse S, Wong P, Chandran S, Christopoulos C, Nazarians-Armavil A, Nasseri NK, Musso G, Ali M, Nazemof N, Eroukova V, Golshani A, Paccanaro A, Greenblatt JF, Moreno-

Hagelsieb G, Emili A.Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol. 2009 Apr 28;7(4):e96.

The HERACLES network: contributions of bioinformatics to the study of essential hypertension

Jana Selent, Ismael Zamora, Nuria Boada Centeno, Manuel Pastor

Research Unit on Biomedical Informatics (GRIB),

IMIM/UPF, Spain

GBS

HERACLES is a network of research groups collaborating to improve our understanding of the mechanisms of essential hypertension. The network involves groups from many different areas (physiopathology, cardiology, epidemiology, genomics, proteomics, molecular biology and bioinformatics) promoting collaborative research.

The work of HERACLES network is characterized as much by the transfer of knowledge "from bedside to bench" as by its converse: "from bench to bedside".

The current objectives of the HERACLES network are: a) to study the Ca+2-dependent K+ channels, the transient receptor potential (TRP) cation channels, and the Ca+2 dependent Cl- channels that are involved in vascular physiology; b) to study protein expression maps in plasma and cardiovascular tissue and their significance for drug treatment; c) to study the effect of flavonoids on ion transport and responses to oxidative stress, and d) to identify biomarkers of risk, prognosis, and treatment responses in extreme arterial hypertension phenotypes.

In the present work, we present an outline of the current collaborations that our group is carrying out with different experimental and clinic groups of the network.

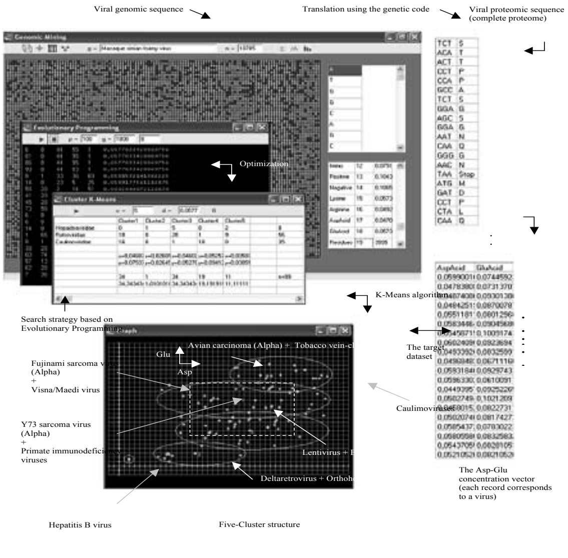

Viral genomic sequence
Translation using the genetic code
Viral proteomic sequence (complete proteome)
Genomic Mining
Evolutionary Programming
Optimization
Cluster K-Means
K-Means algorithm
Search strategy based on Evolutionary Programming
Graph
Glu
Asp
Avian carcinoma (Alpha) + Tobacco vein-cl
Fujinami sarcoma virus (Alpha) + Visna/Maedi virus
Y73 sarcoma virus (Alpha) + Primate immunodeficiency viruses
Lentivirus + E
Deltaretrovirus + Orthoh
Caulimoviruses
The target dataset
Hepatitis B virus
Five-Cluster structure
TCT S
ACA T
ACT T
CCT P
CCA P
GCC A
TCT S
GGA G
AGC S
GGA G
AAT N
CAA Q
GGG G
AAC N
TAA Stop
ATG M
GAT D
CCT P
CTA L
CAA Q
AspAcid GluAcid
The Asp-Glu concentration vector (each record corresponds to a virus)

Figure 1. The Evolutionary Clustering Technique

Figure 1.

RegulonDB: a new window to the genetic regulation of *Escherichia coli* k-12

Martin Peralta-Gil[1], Albero Santos-Zabaleta[1], Socorro Gama-Castro[1], Veronica Jímenez-Jacinto[1], Cesar Bonavides-Martines[1], Luis Muñiz-Rascado[1], Hilda Solano-Lira[1], Araceli Huerta[1], Alejandra Medina-Rivera[1], Heladia Salgado[1], Irma Martínez-Flores[1], Enrique Morett[2], Ingrid Keseler[3], Julio Collado-Vides[1]

[1]Program of Computational Genomics,
Center for Genomic Sciences, UNAM, Mexico,
[2]Biotechnology Institute, UNAM Mexico, [3]SRI International

Background

RegulonDB is a manually curated database that integrates biological knowledge of the mechanisms that regulate transcription initiation in Escherichia coli and the organization of genes and operons in the chromosome. RegulonDB contains detailed, accurate and up-to-date bibliographic information about operon organization, binding sites for transcription factors, promoters, terminators, and RNA regulatory elements. The user interface has a graphic representation and textual information about their sequences, location, evidence and references.

This database is being continuously complemented with computational analyses and predictions, which include weight matrices for some transcription factors, predictions of operons, transcription-factor binding sites, riboswitches, attenuators and computational predictions for promoters of five different sigma factors of the Sigma70 family and Sigma 54 family.

Methods

The first step of the curation process is searching for original published scientific articles using specific keywords related to transcriptional regulation in E. coli. Then, the articles are selected by means of different filters, and the information is annotated in the database. Every element annotated in the database is supported with strong and weak evidence given by the published papers.

On the other hand, we have initiated the annotation of some promoters and DNA binding sites with computational and experimental evidence, derivates from high-throughput experiments such as microarray, library of transcriptional reporters and ChIP-chip experiments.

Results and Conclusions

In RegulonDB, the complex regulatory network of E. coli is resumed in three navigable levels: genes, operons and regulons.

Our efforts, performed in the laboratory of the Dr. Enrique Morett through high-throughput experimental mapping of promoters, have contributed with a total of 317 TSSs for 259 TUs (from which 38 have more than one TSS) that have been mapped with the High Throughput Pyrosequencing Strategy (HTPS) with Roche's 454 GS instrument, out of which 263 are new (Mendoza-Vargas et al., PLOS ONE 2009, in press) and added to RegulonDB.

In relation to the transcription initiation information to date, according to RegulonDB (version 6.4), we have added 169 transcription factors (TFs), 232 conformations of the TFs, 1584 TFs binding sites, 2396 regulatory interactions, 75 effectors, 1771 promoters, 3356 transcription units, 221 terminators, 179 Shine Dalgarno sequences, and 4091 external references. On the other hand, we also have added information of other types of regulation: 751 attenuators, 20 riboswitches and 81 small RNAs.

Future expansions will include modeling of the regulatory pathways, starting from signals or environmental conditions that affect gene expression through several linked reactions that involve interactions among different molecules, such as proteins, RNA, DNA, and metabolites; the signal transduction pathway affecting the core of regulation, and the elicited physiological response.

We acknowledge the support from NIH-NIGMS grant number GM071962-06.

Genetic diversity of *Aedes aegypti* populations in Peru

Omar Alberto Caceres-Rey

Instituto Nacional de Salud, Peru

GBS

Background

Dengue is one of the most prevalent viral emergent infectious diseases worldwide. Since there is currently not any vaccine available, the prevention of the disease depends completely on control of the vector *Aedes aegypti* which carries the virus. Ae. aegypti is very efficient transmitting the dengue virus since it is highly anthropophilic. In 2001, it was reported that the mosquitoes that had been collected in 63% of the Peruvian territory had become a potential threat to public health. The geographic spreading of a species is frequently associated with its genetic divergence raising allopatric populations, this divergence is particularly important in this mosquito because it could affect its capacity of transmitting the dengue virus.

Methods

DNA was purified from 10 mosquitoes from each of the 9 endemic cities. A 314 pb fragment of ITS-2 gene from rDNA was amplified by PCR and then sequenced. The sequences obtained were aligned using ClustalW program. The neutral evolution of ITS-2 was determined with the Tajima Neutrality Test (D value). The genetic divergence was obtained with MEGA 4.1 software and the calculation of the phylogenetic distances was performed using the Tamura-Nei method. Finally, the phylogenetic tree was generated according to the Minimum Evolution model. To complement this approach, AFLP method was used to analyze the genetic variability of Aedes population using total genome, the bands obtained in the gels were scanned and analyzed using Gene Profiler 4.05 software. The matrix generated was used to build similitude dendrogram by UPGMA method using Treecon software. The variability degree between aedes populations (Fst) was calculated by RAPDFST software and the results was corroborated by AMOVA.

Results

The Tajima Neutrality Test showed four segregating sites. The nucleotide diversity estimated in the populations is slightly more than 0.5%. The

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

Tajima's statistic value was D = -0.037860, suggesting that the sequences are in (or near) neutral equilibrium and they can therefore be used to predict variation within populations. In the analysis of the populations, the overall mean genetic distance was 0.76% (SD = 0.004). The mean diversity for the entire population was 0.4% (SD = 0.0028). The phylogenetic tree showed the populations are divided in those which belong to the Coast and those which belong to the Sierra-forest. AFLP results showed that the Aedes populations have the same pattern, supporting our previous results. The total Fst value (genetic difference) was 0.113, for the Coastal cluster it was 0.106 and for the cluster Sierra - forest was 0.160. The AMOVA value was 14.99 % (p <0.001) endorsing the Fst value.

Conclusions

Ae. aegypti populations show two subpopulations, one circulating along the coast and another circulating between jungle and sierra cities. This last subpopulation shows a great differentiation in comparison with the other cluster. It is probable that the Andean mountain range is the responsible for the appearance of allopatric populations within *Ae. Aegypti*.

RegulonDB: challenges and strategies for modeling genetic regulation within a genomic perspective

Heladia Salgado, Verónica Jiménez-Jacinto, Luis J. Muñiz-Rascado, Hilda Solano, Irma Martinez-Flores, César Bonavides-Martínez, Shirley Alquicira-Hernández, Jair S. García-Sotelo, Liliana Porrón, Alejandra C. López-Fuetes, Víctor Del-Moral, Julio Collado-Vides

Program of Computational Genomics, Center for Genomic Sciences, UNAM, Mexico

GBS

Our laboratory has created RegulonDB database (DB), the most complete DB in transcriptional regulation of a nonpathogenic bacterium, Escherichia coli K-12 (Gama-Castro et al., 2008). We have acquired wide experience in the organization and curation of transcriptional regulation information; as well as in the development and implementation of visualization tools, allowing users to navigate in the genome (Genome browser). The user can identify co-regulators for a particular transcription factor, visualize neighbor genes in the regulatory network, and identify a set of genes predicted to be functionally related (Nebulon tool) (Janga SC et al., 2005). We have incorporated other useful graphical tools, which have been created by other groups such as Textpresso (Müller HM et al, 2004), a tool that allows to perform specific searches inside our DB, in the publishing repository created by the curators and in their notes and summaries written specifically for RegulonDB; the Web service, a software system for automated access by using SOAP protocols; and tools for the exchange of data, based in the BioPAX format. In addition, RegulonDB is complemented by computational analyses, and predictions of operons, promoters and binding sites of transcriptional factors. The computational elements behind make RegulonDB have been re-designed to allow a further expansion into a multigenomic DB. We are working hard in order to integrate the software engineering methodologies in our team. The information contained in RegulonDB is useful to researchers all over the world. In Collado-Vides, et al., 2009, it is possible to find a complete description of how researchers, both experimental and bioinformatitians, use RegulonDB; even if working with different organisms (Collado-Vides et al., 2009).

RegulonDB is considered to be "the golden standard" for the implementation of prediction methods, topologic analyses of networks and cellular models (Collado-Vides et al., 2009); therefore, our infrastructure shall allow to incorporating data about all kinds of genetic regulation in a multigenomic context. In the course of several years, the RegulonDB project has been enriched by the collaboration of the laboratory of Dr. Enrique Morett at IBt-UNAM, using high throughput technologies to experimentally map promoters and transcription units of the genome. This new and exciting avenue, together with the literature-based knowledge of the network, contributes to the world-wide effort to attain a comprehensive understanding at the molecular level of a single cell organism by encompassing experimental approaches, bioinformatics and systems modeling. The pioneer work performed in E. coli will enable the community to learn the limitations and strategies for a similar modeling of any other bacterial genome and will give lessons to the understanding of more complex organisms.

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

The Mexican National node of Bioinformatics, EMBNET: History and perspective

Romualdo Zayas-Lagunas, César Augusto Bonavides, Víctor Manuel del Moral, Alfredo José Hernández, Heladia Salgado, Santiago Sandoval, Jason Gunther Lomnitz, José M. Uriel Urquiza, María Guadalupe Loza, Julio Collado

Program of Computational Genomics, Centro de Ciencias Genómicas, UNAM, Mexico

GBS

During a meeting in Switzerland in 2000, EMBnet agreed to accept the membership application of the Mexican National node of Bioinformatics, EMBnet. The proposal included a perspective showing infrastructure capacity to have a site with periodically updated essential databases (GenBank, UNIPROT, and few others), the proposal of introductory courses of bioinformatics, to implement a website, and the personnel able to take care of all these responsibilities. The physical site and personnel are located within the Program of Computational Genomics, at the Center for Genomics at UNAM, in the city of Cuernavaca in the Morelos Campus of UNAM. This site had the aforementioned capacities thanks to the combination of the resources for "The Development of Genomic Sciences in Mexico: The Genome of Rhizobium etli" from the national council of Research (CONACYT) and the support for genomics in Cuernavaca within UNAMs plan of the former President of our university. This plan conceived both the development of a new undergraduate program in genomics, and the support of bioinformatics in Cuernavaca at the Institute of Biotechnology and the Center for Nitrogen Fixation –which became the current Center for Genomic Sciences. It this infrastructure that enabled the computational component associated to the annotation of the Rhizobium etli genome project, we generated the first full genome in our country. The major resources we offer to our community are: BLAST, RSATools, RegulonDB, RetliDB, wEMBOSS and a variety of computational tools for evolutionary studies. We also support running jobs on the cluster at CCG with 54 processors and grid initiatives inside the EELA (E-science grid facility for Europe and Latin America) project. In 2007, the General Direction of Computer Academic Services at UNAM (DGSCA) got interested in expanding its services to include bioinformatics. As a result, we have initiated the coordination of several institutes and centers for biomedical and biological research with the aim of integrating their resources in a single national node at UNAM´s level, and to enhance its portal services. Furthermore, within DGSCA, one of the larger clusters of Latin America, KanBalam (with more than 1300 processors) is now also running bioinformatic processes on an individual project-based strategy. The collection of these strategies is helping us to cope and stimulate bioinformatic research and education in our country. The URL of EMBNet Mexican Node is http://www.nnb.unam.mx.

Acknowledgments

We acknowledge the support of Dirección General de Servicios de Cómputo Académico, UNAM for the expansion of the national node, as well as to Jérôme Verleyen and Roberto Rodriguez from the Institute of Biotechnology, UNAM.

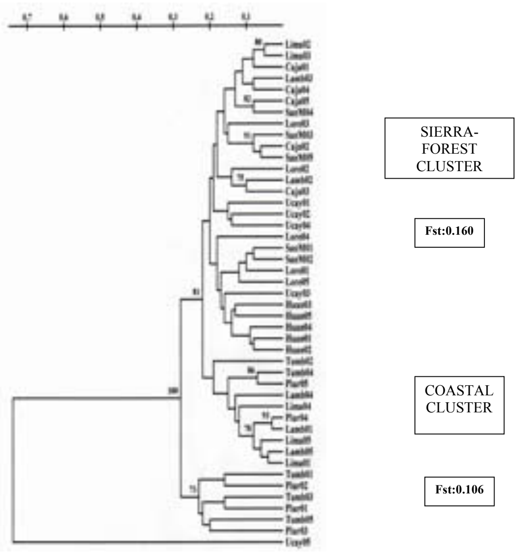


Fig 1. UPGMA Dendrogram of *Ae. aegypti* subpopulations (Boostrap >70) and its Fst values

High throughput, multigenome analysis using the AlterORF bioinformatics pipeline reveals both rampant mis-annotation and novel gene discovery

Francisco J. Ossandon[1,2,3], **Rene Sepulveda**[2], **David S. Holmes**[1,2,3]

[1]Center for Bioinformatics and Genome Biology, [2]Fundación Ciencia para la Vida and Depto. de Ciencias Biologicas, [3]Facultad de Ciencias de la Salud, Universidad Andres Bello, Santiago, Chile

HTT

Background

A particularly challenging problem in genome annotation is to attribute function to genes annotated as "hypothetical, no known function". These typically account for about 40% of all genes regardless of the genome. Some of these are "orphan" genes and are not found in any other genome; other "hypotheticals" are conserved across different species. Some orphan genes could encode species specific proteins and so are particularly interesting for evaluating novel metabolic potential and for understanding the evolution of genes and genomes. Several bioinformatics tools exist that help predict function of hypotheticals, but so far, none have been able to suggest function for more than a small percentage and the annotation of the others remains a formidable task.

Methods

We have developed a bioinformatic tool and database AlterORF (www.alterorf.cl), that is able to identify alternate open reading frames (ORFs) embedded within annotated genes. Using in-house Perl scripts, each gene and alternate ORF aminoacid sequence has been extracted from the genome annotations, and then compared through Blast to different domain databases (such as COG or Pfam); after which significant hits (Evalue ≤ 1e-5) are stored in a relational MSSQL database. Currently the database contains over 2 million genes and all their alternate ORFs of substantial length (potentially encoding 70 aminoacids or more) from over 700 completely sequenced prokaryotic genomes.

Results

Analysis of alternate ORFs in AlterORF reveals hundreds of examples where the alternate ORF has a significant hit with databases of motifs and domains (e.g. CDD, Pfam) and where the actual annotated gene is described as hypothetical and has no database match. This strongly suggests that the annotated gene has been incorrectly identified and that the alternate ORF is the real gene. We describe the evaluation of the following genomes of extremophile microorganisms using AlterORF: *Acidithiobacillus ferrooxidans* (2 strains), *Leptospirillum* type II, *Methylacidiphilum infernorum*, *Picrophilus torridus*, *Sulfolobus acidocaldarius*, *S. solfataricus*, *S. tokodaii*, *Thermodesulfovibrio yellowstonii*, *Thermoplasma acidophilum* and *T. volcanium*. Up to 60% of annotated hypotheticals proteins were discovered to have hits with motif and domain databases. Also, examples of novel genes and their suggested roles in metabolism will be described.

Conclusions

Deep bioinformatic analysis provided by AlterORF reveals the presence of both a substantial number of annotation errors and also potential new genes in the 11 extremophile microorganisms under review. A large number of the errors encountered result from inadequate or erroneous identification of genes, and many errors appear to escape the attention of expert human curators. It is anticipated that AlterORF will provide a service for improved genome annotation and new gene discovery.

Multigenome Analysis of Proteins from Extremely Acidic Environments

Francisco Arturo Duarte[1], **Rene Sepúlveda**[2], **David Holmes**[2,3]

[1]Center for Bioinformatics and Genome Biology, Chile, [2]Center for Bioinformatics and Genome Biology Fundacion Ciencia para la Vida MIFAB, [3]Depto. de Ciencias Biologicas, Facultad de Ciencias de la Salud

HTT

Background

Proteins outside the membrane of acidophilic bacteria must function at extremely acidic pHs (pH1-2). Proteins embedded in the membrane

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

will also have loops exposed to acid pH. A particularly challenging problem is how these proteins fold, make protein-protein contact and function at extremely acidic pHs. Another question is how membrane transporters, including those using proton motif force to drive uptake or discharge of ions, function when confronted by a DpH of 6 orders of magnitude across the periplasmic membrane (pH 6.5 inside to pH 1 outside). Whereas an enormous amount of information has been generated regarding the biochemical and biophysical bases of protein function at extreme temperatures, virtually nothing is known about the physicochemical determinants of acidic proteins prompting us to undertake this study.

Method

All proteins constituting the predicted proteomes of sixty-one completely sequenced extremophilic microorganisms, including 22 acidophilic, 21 neutrophilic, 9 alkaliphilic and 9 halophilic micro-organism, were subjected to an analysis of their subcellular destination using PsortB, and were binned according to cytoplasm, periplasmic membrane, periplasm and outer membrane. Binned proteins were then sorted into predicted orthologs according to life style using OrthoMCL. In-house PERL programs were scripted to calculate physicochemical features of the orthologs and the data was loaded into searchable tables using a MySQL-PHP interface.

Results

A searchable relational database was constructed of more than 6,000 proteins predicted to function in extreme environments, including very acidic conditions. Protein sequences were linked to information about organism habitat, cellular location, predicted function and multiple physicochemical parameters. Protein loops exposed to low pH tend to be shorter than their neutrophilic orthologs and to be richer in hydrophobic amino acids. In predicted membrane transport proteins such as aquaporins and potassium channels, amino acid changes were identified that help explain the selectivity of these channels for their substrate even in the presence of a huge extracellular concentration of protons.

Conclusions

The creation of DPAE opens up new opportunities for revealing fundamental properties of the structure and functions of proteins in acidophilic and other extreme conditions and might also contribute to the discovery of proteins with potential biotechnological applications.

High throughput prediction of small regulatory RNAs (srRNAs) in extremophilic bacteria.

Amir Shmaryahu[1,2] Claudia Lefimil[1,2], Eugenia Jedlicki, David S. Holmes[1,2].

Addres: [1]Center for Bioinformatics and Genome Biology, [2]Fundacion Ciencia para la Vida, Chile

GBS

Background

An increasing number of small regulatory RNAs (srRNAs), ranging in size from 70 to 500 nucleotides, have been shown to control critical pathways in microorganisms primarily by acting as regulators of either protein synthesis or protein activity. srRNAs are involved in the regulation of a large variety of processes such as plasmid replication, transposition and global genetic circuits that respond to environmental changes. srRNAs are proving to be multifunctional and have provided explanations for a number of previously mysterious regulatory effects. Despite the widespread occurrence and important function of srRNA genes, current automatic genome annotation programs have difficulty in predicting them. Their computational discovery is particularly challenging because they do not encode protein products, so bioinformatic tools such as open reading frame (ORF) identification, codon usage and Hidden Markov Model (HMM) searches based on conserved protein motifs cannot be used. In addition, srRNAs can be poorly conserved at the nucleotide sequence level and only some exhibit conserved secondary structure.

Methods

A novel bioinformatics pipeline was constructed for predicting srRNA genes. A database of intergenic DNA sequences (IGs) of selected extremophilic bacteria was created using in-house Bioperl scripts. The database was searched for intergenic regions that exhibited sequence similarity and conserved genetic context. Candidate sequences were then subjected to promoter and transcriptional termination predications. Intergenic regions that passed these screens were searched for sequence similarity to known srRNAs (Blast and Rfam) and for conserved secondary structure (Mfold).

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

Results

Thirty srRNA genes were predicted. Three of these (tmRNA, rnpB and 4.5S) have previously been identified in other organisms. Of the remaining 27 candidates, twelve were experimentally validated by Northern blotting and 5' RACE. One novel candidate, termed frr, was predicted to be involved in cellular iron uptake and homeostasis via the master iron regulator Fur.

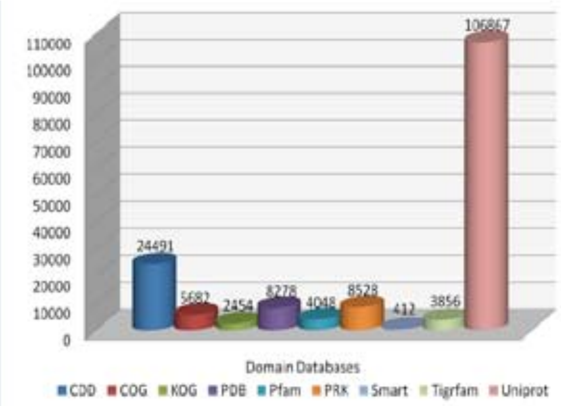
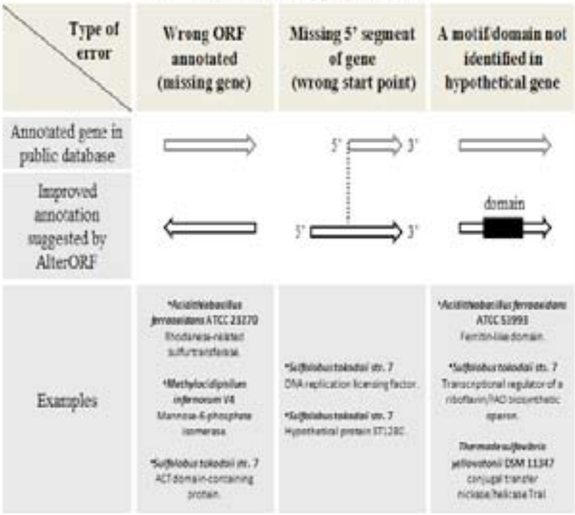
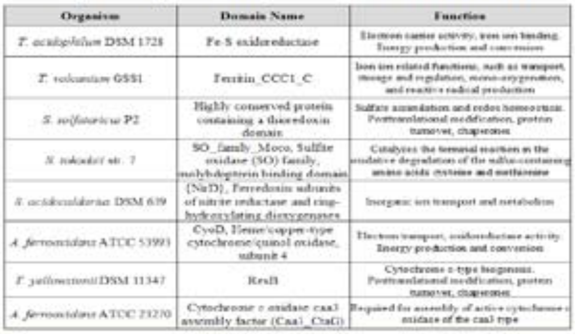


Figure 1. Alternate ORF's domain hits in database



Type of error	Wrong ORF annotated (missing gene)	Missing 5' segment of gene (wrong start point)	A motif/domain not identified in hypothetical gene
Annotated gene in public database		5' 3'	
Improved annotation suggested by AlterORF		5' 3'	domain
Examples	**Acidithiobacillus ferrooxidans* ATCC 23270 Rhodanese-related sulfurtransferase. **Methylacidiphilum infernorum* V4 Mannose-6-phosphate isomerase. **Sulfolobus tokodaii str. 7* ACT domain-containing protein.	**Sulfolobus tokodaii str. 7* DNA replication licensing factor. **Sulfolobus tokodaii str. 7* Hypothetical protein ST1280.	**Acidithiobacillus ferrooxidans* ATCC 53993 Ferritin-like domain. **Sulfolobus tokodaii str. 7* Transcriptional regulator of a riboflavin/FAD biosynthetic operon. *Thermodesulfovibrio yellowstonii* DSM 11347 conjugal transfer nickase/helicase TraI

Table 1. Annotation error types detected

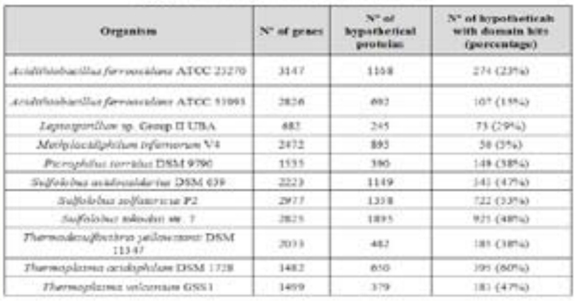


Organism	Domain Name	Function
T. acidophilum DSM 1728	Fe-S oxidoreductase	Electron carrier activity, iron ion binding. Energy production and conversion
T. volcanium GSS1	Ferritin_CCC1_C	Iron ion related functions, such as transport, storage and regulation, mono-oxygenation, and reactive radical production
S. solfataricus P2	Highly conserved protein containing a thioredoxin domain	Sulfate assimilation and redox homeostasis. Posttranslational modification, protein turnover, chaperones
S. tokodaii str. 7	SO_family_Moco, Sulfite oxidase (SO) family, molybdopterin binding domain	Catalyzes the terminal reaction in the oxidative degradation of the sulfur-containing amino acids cysteine and methionine
S. acidocaldarius DSM 639	(NirD), Ferredoxin subunits of nitrite reductase and ring-hydroxylating dioxygenases	Inorganic ion transport and metabolism
A. ferrooxidans ATCC 53993	CyoD, Heme/copper-type cytochrome/quinol oxidase, subunit 4	Electron transport, oxidoreductase activity. Energy production and conversion
T. yellowstonii DSM 11347	ResB	Cytochrome c-type biogenesis. Posttranslational modification, protein turnover, chaperones
A. ferrooxidans ATCC 23270	Cytochrome c oxidase caa3 assembly factor (Caa3_CtaG)	Required for assembly of active cytochrome c oxidase of the caa3 type

Table 2. Examples of "hypothetical proteins" with detected domains

Conclusion

A novel bioinformatics pipeline predicted twenty seven novel srRNA genes in a group of extremophilic bacteria. Some of the predictions were experimentally validated. One novel srRNA (frr) is thought to be involved in iron homeostasis, providing insight into a new regulatory mechanism for this important and universal metabolic capability.



Organism	N° of genes	N° of hypothetical proteins	N° of hypotheticals with domain hits (percentage)
Acidithiobacillus ferrooxidans ATCC 23270	3147	1168	274 (23%)
Acidithiobacillus ferrooxidans ATCC 53993	2826	692	107 (15%)
Leptospirillum sp. Group II UBA	482	245	73 (29%)
Methylacidiphilum infernorum V4	2472	893	50 (5%)
Picrophilus torridus DSM 9790	1535	390	148 (38%)
Sulfolobus acidocaldarius DSM 639	2223	1149	541 (47%)
Sulfolobus solfataricus P2	2977	1338	722 (53%)
Sulfolobus tokodaii str. 7	2825	1895	921 (48%)
Thermodesulfovibrio yellowstonii DSM 11347	2033	482	181 (38%)
Thermoplasma acidophilum DSM 1728	1482	650	395 (60%)
Thermoplasma volcanium GSS1	1499	379	181 (47%)

Table 3. Hypothetical proteins per organism

The spanish national bioinformatics institute

Allan Orozco

National Institute of Bioinformatics (INB), Spain

GBS

The INB (Spanish National Bioinformatics Institute, www.inab.org) is a technical platform of Genoma España, a Public Fundation for the development of Genomic an Proteomic research in Spain (www.gen- es.org/) The INB is organized in 9 specialized nodes including a partnership with the Barcelona Supercomputing Centre (www.bsc.org). During the five years it has been functioning, it has trained a new generation of bioinformaticians, 30 of which are now working in the institute. At the technical level, the INB has developed a large infrastructure of Web services, including 400 individual methods for the storage, access and exploration of the data in the context of large scale genome projects, a system for the archiving and direct execution of workflows in Taverna format (IWWEM, INB Web Workflow Enactor and Manager, http://moby- dev.inab.org/IWWEM/workflowmanager.html), a system for the creation of Web services under a defined on-

GBS - General Bioinformatics Subject
HTT - High Throughput Technologiest
OSB - Open Software for Bioinformatics
IBSB - The Interface of Bioinformatics & Systems Biology

tology and supervised documentation Efforts to implement integration and standardization have been developed some widgets that integrates web services coming from central repository, which retrieve and integrate biological data. Additionally the INB has developed in collaboration with the BioSapiens NoE (www.biosapiens.info) the CARGO [1] system for the representation of the results of bioinformatics analysis to end-users using a friendly technology of widget and desktops). Efforts to implement integration and standardization have been developed some widgets that integrates web services coming from central repository, which retrieve and integrate biological data. Also the INB has developed Gepas [2] and Babelomics [3]. GEPAS: is the web tool for microarray data analysis most extensively used in more than 40 modules (running as web servers with Biomoby technology). It can import data from different versions of different platforms (Affymetrix, Agilent, Codelink, Illumina and others). It provides facilities for normalization and preprocessing. Differential gene expression can be conducted with different tests for different experimental designs (case/controls, multiclass, continuous parameter, time-course, survival, etc.). Babelomics is a web-based suite of methods for the functional profiling of genome-scale experiments implemented in more than 20 modules (running as web servers with Biomoby technology). It implements different functional (GO, KEGG, Biocarta, interactions, etc.), regulatory (Transfac, CisRed, miRNAs) and other (tissues, etc.) databases containing biological knowledge in order to conduct different flavours of functional enrichment and gene-set enrichment tests. There are available tools denominated Geneid [4], consisting on an ab initio gene prediction. Geneid has been used to analyze many eukaryotic genomes. Geneid has also been used in the annotation of the 12 drosophila genomes, and it is also being used in the annotation of the cow and tomato genomes. Finally, the INB counts with the development of many tools, workflows, web services and else, that we are willing to show and explain to the genomic community.

References

1. http://cargo2.bioinfo.cnio.es
2. http://gepas.bioinfo.cipf.es/
3. http://babelomics2.bioinfo.cipf.es/
4. http://genome.imim.es/geneid.html