TIN2006: Identifying complex information in biology: From data collections to network-organized knowledge

Principal Investigator:
Igor Zwir

Members:
Rocio Romero Zaliz
Oscar Harari
Patrizia Anders

Coral del Val
Cristina Rubio Escudero
Agnes Hotz-Wagenblatt

Abstract

The multiple levels at which selection acts in biology always conceals the answer to the question of what does a biological system do? Because of this difficulty, it is unlikely that even a mature theory of program machines will be adequate to explain biological systems. However, even in the absence of grand theory, one can work on intermediate steps. Here we describe several avenues worth exploring. We are devoted to provide a computational framework and methodology that can generate broad sets of optimal hypotheses to answer the above question, describing biological systems from distinct perspectives. In other words, we would like to identify, simulate and predict genotypic-phenotypic relations and, more pretentiously, to shed light in their subjacent mechanisms.

Keywords: Computational biology, Bioinformatics, Machine Learning, Conceptual Clustering, Evolutionary Computation, Gene Expression, Microarrays, Genomic Sequence Analysis, Prokaryotes, Eukaryotes
1 Project objectives

Although a critical challenge of the postgenomic era is to understand how genes are differentially regulated, development and implementation of biological data repositories containing representations of complex regulatory networks has not been matched by an increased availability of tools and structures permitting the search of those databases in terms that are close to the language of the intended users.

To address the problem of identifying information networks in biology, we propose a generalized conceptual clustering methodology that retrieves system descriptions, combines information of different sources (e.g., genomic, proteomic, transcriptomic data), performs annotation and predicts new cases based on that knowledge. This methodology is based on applying conceptual clustering and multivariate, multiobjective, multimodal and mutidecision optimization techniques, and combines metaheuristics based on evolutionary computation, fuzzy logic and control techniques. The computational approach is combined with molecular biology techniques, including microarray, Chromatin immunoprecipitation, and real time PCR, to validate and update the computational discovered information networks.

Summarizing, we propose a complete methodology devoted to 1. Discover regulatory elements by identifying their features. 2. Relate regulatory elements to identify regulatory profiles (i.e., set of genes sharing common sets of features). 3. Annotate profiles. 4. Predict profiles. 5. Evolve profiles. 6. Identify and simulate the profile dynamic. 7. Perform molecular engineering of profiles. 8. Extend prokaryotic profiles to eukaryotic organisms.

We approached these objectives having a distinguished and experienced team and collaborators who facilitated the developed research program. Dr. Zwir and Dr. del Val have conducted this team. They both designed and scheduled the program that allowed us to satisfy the objectives of the project. They complemented each other having machine learning and molecular biology backgrounds, and carried out the difficult task of developing a new interdisciplinary research group at the DECSA1, University of Granada, Spain. Both of the researchers closely collaborated with international institutions including SRI International (USA), Howard Hughes Medical Institute (HMMI) at Washington University School of Medicine, St. Louis, MO, and at Janelia Farm, WA, USA, and the German Cancer Research Center, Heidelberg, Germany. These collaborations constituted the fundamental source of biological problems and experimental validation of the research carried out by the group. Indeed, the other members had performed an incredible contribution to the achievements of this project. Dr. Romero Zaliz, Dr. Cristina Rubio Escudero, and Dr. Harari designed, implemented and applied most of the computational tools used to solve the former problems as it is demonstrated by the publications of the group. The PhD. Student Patrizia Anders is now carrying some of this work out. In addition, discussions with Dr. Hotz-Wagenblatt played a very useful role.

Initially we developed a conceptual clustering methodology to identify, describe, annotate and predict genetic networks and applied it to *cis*-regulatory networks in prokaryotic organisms. Then
we studied the evolution of these regulatory systems in different genomes and analyzed their dynamic properties including temporal binding and transcription. We joined the cis-and the dynamic descriptions of the system and developed a strategy to predict gene expression from cis-regulatory elements. We also approached particular regulatory elements such as RNA, which resulted in fundamental gene-control processes. Numerous mRNAs in prokaryotes carry complex folded domains, which are known as riboswitches, within the non-coding portions of their polynucleotide chains. Finally, we applied molecular engineering to validate the obtained results. We extended our approach to eukaryotes and found phenotypic and genotypic profiles of patients suffering diseases including schizophrenia and inflammatory problems. All of this work has been done in stretch collaboration with:

- Dr. Suhais Lab, Department of Molecular Biophysics, German Cancer Research Center, Heidelberg, Germany (http://genome.dkfz-heidelberg.de/)
- Dr. Wiemman Lab, Genome Analysis, German Cancer Research Center, Heidelberg, Germany (http://www.dkfz.de/mga/)
- Dr. Jiménez Zurdo EEZ (CSIC), Granada, Spain.
- Dr. Eddy and Dr. Rivas HHMI Janelia Farm, WA, USA
- Groisman Lab, Department of Molecular Microbiology, Howard Hughes Medical Institute, Washington University School of Medicine, St. Louis, MO, USA. (http://www.hhmi.org/)
- SRI Internacional, Artificial Intelligence Center, Menlo Park, CA, USA (http://www.ai.sri.com/).
- Grupo SCIS, Soft Computing and Intelligent Information systems University of Granada, Granada, Spain (http://sci2s.ugr.es).
- Dr. de Erausquin Lab., Department of Psychiatry, School of Medicine, Washington University, St. Louis, MO, USA.
- Dr. Cobb Lab, Center for Critical Illness and Health Engineering, Washington University, St. Louis, MO, USA.

2 Project achievements

We have been working in different areas including creating a research program, developing a computational biology group or lab, generation of resources to support the research, and finally, development of a teaching program to communicate the acquired knowledge.

The research program: identifying, describing and predicting biological networks

We have focused our research in identifying genetic networks in both prokaryotic and eukaryotic organisms covering the following topics.
**How a transcriptional activator achieves differential expression of its target genes.** The primary mechanism governing inducible gene expression operates at the level of transcription initiation whereby DNA binding proteins recognize specific sequences in promoters to activate or repress transcription by RNA polymerase (RNAP) [1] Whereas certain regulators control transcription of a single gene or operon, others – global regulators – can govern expression of dozens or even hundreds of genes. Because the encoded gene products are typically required in different amounts, a global regulator must act distinctively at its target promoters to produce disparate outputs for a given input. We study the critical elements [2] that enable a global regulator like PhoP to differentially control its target genes [3] We described the binding and transcription dynamic behavior of promoters controlled by the PhoP protein in the bacterium *Salmonella enterica* serovar Typhimurium and *E. coli* [4]. We could deduce the critical promoter features that determine differential gene expression. Taken together with *in vitro* binding assays, the use of engineered promoters and *in vitro* transcription experiments, we demonstrate that PhoP utilizes distinct mechanisms to promote transcription of its target genes.

A resulting mathematical model can predict gene expression behavior from *cis*-acting promoter features. Indeed, we traced our initial findings in *Salmonella* through the evolution of the enterobacteria. The original question of what features are responsible for certain kinetics, will be now reformulated as what happened in the *Yersinia* genome with those promoters having an early and high dynamic behavior in *Salmonella*. Preliminary studies suggest that genes and *cis*-features can be gained and lost along the phylogenetic tree of gamma-enterobacteria [5] following a modular model of evolution (O.H. et al, submitted manuscript). More recently, we have studied the kinetic activation overshoots resulting from the interconversion of bi-functional histidine sensor kinases between kinase and phosphatase status in the same protein, which constitutes a two component system (I.Z. et al., manuscript in preparation).

**RNA mediates gene control in regulatory networks.** Accurate identification of novel, functional noncoding (nc) RNA features in genome sequence has proven more difficult than for gene coding proteins. Current algorithms identify and score potential RNA secondary structures on the basis of thermodynamic stability, conservation, and/or covariance in sequence alignments. The principal drawback of all existing methods is the vast number of false positives that make impossible the validation of results in the lab, and thus the extraction of new regulation knowledge [6] We had developed in collaboration with Dr. E Rivas (HHMI Janelia Farm) a strategy that combines the use of Hidden Markov Models and the optimization structure thermodynamic stability through the evolution [7], taking into consideration the problem of false-positive predictions by conducting a careful characterization of the theoretical reduction in the number of false positives that the combination of two methods with different algorithms would bring. Our strategy presents sensitivity values as good as for each of the individual methods used (81%), and specificity was substantially increased, indeed we had no false positives for our tests. Using this methodology, we performed a whole-genome screen to identify novel sRNAs in the legume endosymbiont *S. meliloti*

We identified and mapped experimentally seven novel transcripts that accumulated differentially in free-living and symbiotic conditions. These findings suggest novel regulatory
functions for sRNAs related to the interactions of a-proteobacteria with their eukaryotic hosts [8]. The resulting predictions were validated in the laboratory of Dr. Jimenez-Zurdo (Estacion experimental del Zaidin, CSIC, Granada).

A map function of schizophrenia risk independent of psychosis. Several metabolomics studies have recently been conducted in an attempt to better define pathways modified in schizophrenia and its treatment [9] We designed a study to test the hypothesis that dopaminergic deficits are sufficient to distinguish subjects at risk of, or affected with, schizophrenia from appropriate matched controls. To test this possibility we looked for evidence of dopaminergic deficits in untreated patients with schizophrenia and their unaffected siblings. Therefore, assessed measures of brain dopaminergic function were expected to vary in opposite directions at the extremes of the variable domains [10] We developed a machine learning method specifically designed to recover phenotype/genotype relationships without assumptions regarding prevalence or disease model (specifically, our method is neutral to “multiple common variations with small effects” vs. “multiple rare variations with large effects”) (G.E. et al., manuscript in preparation).

Differential expression profiles over time, treatment, patient, or other experimental conditions dissect different inflammatory phenotypes. We have created a methodology that successfully extracts all reliable targets from microarray experiments providing statistically meaningful results by means of combining the advantages of several microarray analysis methods [11, 12] The gene information is extracted based on the differential profiles genes exhibit over time, treatment, patient, or other experimental conditions. This environment also fuses genetic information from different sources including experimental knowledge and biological databases. We have achieved the functional annotation of the gene expression information, grouped in differential profiles, obtained from microarray experiments, with data from biological databases, besides the comparison of gene network creation methods and fusion of the information given by these genetic networks with the already known problem related.

Particularly, the Gene Ontology (GO) vocabulary has been extensively explored to analyze the functions of coexpressed genes. However, despite its extended use in Biology and Medical Sciences, there are still high levels of uncertainty about which ontology (i.e. Molecular Process, Cellular Component or Molecular Function) should be used, and at which level of specificity. We proposed a web tool [11], Onto-CC, as an automatic method specially suited for independent explanation/validation of gene grouping hypotheses (e.g. coexpressed genes) based on GO clusters (i.e. expression versus GO). Onto-CC [13] approach reduces the uncertainty of the queries by identifying optimal conceptual clusters that combine terms from different ontologies simultaneously, as well as terms defined at different levels of specificity in the GO hierarchy. Therefore, we can generate alternative and, still, optimal explanations of queries that can provide new insights for a given problem.

Summary of the computational framework and methodology designed to identify and predict biological networks. Here we synthesize our computational approach developer as a result of the TIN2006 project, which has been used to generate solutions in the above described biological
problems. The proposed methodology is composed of several steps that can be iteratively and dynamically applied \([4, 14]\): 1) we encode the data into databases based on a degree of matching with a feature he instead of discrete Boolean data using fuzzy sets \([15]\) (or non parametric distributions). 2) We identify patterns, where one observation can support more than one pattern using fuzzy clustering. 3) We identify patterns defined at distinct granularity levels (i.e., few features with plenty of observations vs. several features supporting few observations) and distinct sets of features (i.e., structural vs. static data \([11]\)) using conceptual clustering. We avoid consensus methods that tend to homogenize features among observations, often hampering the discovery of novel patterns. 4) We extract optimal and diverse partitions of the data, describing a system from different points of view by evaluating clusters using multiobjective and multimodal optimization techniques. 5) We independently search for patterns within each domain of knowledge (e.g., latent classes) and a posteriori integrate the different sources into relations to avoid conditioning patterns to preexisting ones (i.e., unsupervised learning). This fusion of knowledge is based on coincidence analysis (i.e., probability of intersection and multi-valued logic techniques). 6) The identified relations constitute the basis of a labeling process, which can later derive in the construction of classifiers (e.g., supervised learning). Particularly, we use voting multi-classifiers \([16]\) for the inference/prediction process. 7) We trace clusters through evolution using population clustering, and search for policies of evolution using game theory \([17]\) and reinforcement learning techniques \([18]\). 8) We describe the system and simulate its dynamics using forward and reverse engineering to identify kinetic reactions \([19]\). We also use sampling and probabilistic techniques to identify the architecture of the system \([20]\) (e.g., genetic network).

We demonstrated in our publications that the followed methodology is crucial for discovering novel and useful knowledge in biology. The power of the proposed methodology is that it uses the advantages of several methods and combines them in a hybrid research project. In addition to the biological findings, we demonstrated the bioinformatics usefulness of our approach by strict methodological comparisons with other methods \([13, 14]\).

**Creation of a computational biology group: The M4M Lab**

We have developed a new interdisciplinary group composed of computer scientist and molecular biologists and constituted the M4M lab (m4m.es), which is a laboratory of the Research Group Soft Computing and Intelligent Information Systems. Dr. Igor Zwir and Dr. Del Val combining their experience in machine learning, computational biology, bioinformatics and molecular biology have conducted this lab. The lab is also composed of Dr. Rocio Romero Zaliz, an assistant professor at the University of Granada who was a PhD student advised by Dr. Zwir and defended her thesis in computational biology. Dr. Cristina Rubio, currently assistant professor at the Seville University, defended her European thesis in analysis of gene expression in patients suffering inflammatory diseases. The PhD student Oscar Harari is currently in the process of defending his European thesis in prokaryotic and eukaryotic regulatory networks. More recently Patrizia Anders, who has a scholarship derived from our TIN2006 project, started to work in bacterial networks. We also advised other students from Universidad Metropolitana de Chile, Universidad Autónoma de Madrid and Karls-Ruprecht University.
Generation of research projects
We obtained the project termed “Caracterización y predicción de redes de expresión génica mediante el análisis y modelado de elementos regulatorios cis y post-transcripcionales: de bacterias a organismos superiores” from the Junta de Andalucía, whose principal investigator is Dr. del Val. In addition, we have been supported by the Groisman lab, Howard Hughes Medical Institute, and by other collaborators for the experimental developments and data. In addition, we have applied to several national and international projects (see below). The most relevant consist of the development of the “Consorcio andaluz de ómica: nuevas tecnologías al servicio de la calidad de vida en Andalucía”, which pretend to join several companies from the Campus de la Salud, Spanish Hospitals and the University of Granada.

Developing a teaching program
We have designed a new graduate and post-graduate course termed “Introduction to computational biology” at the DECSAI, University of Granada, which is the link of our lab to the students. This is the third year that we are teaching and we have had an average of 30 students, suggesting that they are interested in the proposed subjects. Indeed we conducted the course termed “Bioinformatics” at the PhD program carried out at the DECSAI, University of Granada and also we are teaching in courses from other European and American Universities.

Analysis of results
The support received from the TIN2006 played a critical role in the development of our lab. None of our recent activities could be done without this help. However, we have some suggestions, which results from our recent experience. Since ours is an interdisciplinary lab, sometimes we need to spend recourses in items that are not explicitly detailed in the program (e.g., temporal hiring technicians). Moreover, we specifically need postdoctoral positions, which at this stage are fundamental for improving the performance of the project. Finally, we think that starting a new lab in a novel field needs at least 5 years funding to begin with.

3 Result indicators
In the following we explicitly specify the publications resulting from the project (Conferences and Lecture Notes were omitted for space constraints):

Publications in peer reviewed journals
O. Harari, C. del Val, D. Shin, H. Huang, and E. A. Groisman, I. Zvir,. EXTRACTING PROMOTER FEATURES FROM GENOMIC DATASETS: TOWARDS AN ANNOTATION OF GENOME REGULATORY REGIONS. BMC Bioinformatics, 2009. (To be published in the special issue corresponding to the IEEE international Conference on Bioinformatics and Biomedicine (BIBM08)).


Submitted and in preparation manuscripts

J. C. Perez, D. Shin, I. Zwir, T. Latifi and E. A. Groisman. EVOLUTION OF A BACTERIAL REGULON CONTROLLING VIRULENCE AND MG2+ HOMEOSTASIS.


I. Zwir, H. Huang, D. Shin, W. Sik Yeo, A. Mitrophanov, A. Kato, and E. A. Groisman. BI-FUNCTIONAL HISTIDINE SENSOR KINASES INTERCONVERT BETWEEN KINASE AND PHOSPHATASE STATES TO GENERATE AN ACTIVATION SURGE.


Book chapters


Additional research support
Project title: Characterization and prediction of gene expression networks by modeling and analyzing cis and post-transcriptional regulatory elements: from bacterial to superior organisms (TIC-02788)

Finance provider: Proyecto de excelencia de la Junta de Andalucía (Spain).

Participating entities: M4M LAB (www.m4m.es)

Duration: from 2007 to 2010

Head of research: Dra. Coral del Val

Number of researchers taking part: 6

PhD. thesis

- **O. Harari** (2009). PREDICTING PROKARYOTIC AND EUKARYOTIC GENE NETWORKS BY FUSING DOMAIN KNOWLEDGE WITH CONCEPTUAL CLUSTERING ALGORITHMS, DECSAI, UGR, European PhD. Program. Currently Post doctoral studies at Psychiatry Dept., Harvard University, USA.


- **C. Rubio Escudero** (2007). IDENTIFYING GENE PROFILES BY REVERSE PROBLEM SOLVING: FROM GROUPING GENE EXPRESSIONS TO COMBINING MICROARRAY ANALYSIS METHODS, DECSAI, UGR, European PhD. Program. Currently Assistant Professor at University of Sevilla, Spain.

- **P. Anders** (present). PHYLOGENETIC EVOLUTION IN PROKARYOTIC CIS REGULATORY NETWORKS, DECSAI, UGR, European PhD. Program.

- **L. Herrera** (present). MULTIOBJECTIVE WEB MINING: TOWARDS IDENTIFYING BIOLOGICALLY MEANINGFUL INFORMATION. Universidad Metropolitana de Chile, Chile – UGR, Spain.

- **P. Yankelevich** (present). INTEGRATING GENOMIC, PROTEOMIC AND TRANSCRIPTOMIC INFORMATION BY USING CONCEPTUAL CLUSTERING TECHNIQUES. Universidad Autóinoma de Madrid, Spain.

Fellowships

Most of the research groups spend time at the following institutions:

- Dr. Groisman Lab, Howard Hughes Medical Institute (HMMI), Washington University School of Medicine, St Louis, USA. Dpt. of Molecular Microbiology,

- Dr. Eddy and E. Rivas Lab., HHMI, JANELIA FARM, VA, USA.
- HUSAR Bioinformatics Group y Div. of Genomic Analysis, Dept. of Molecular Biophysics, German Cancer Research Center (DKFZ), Germany.
- Dr. Gabriel de Erausquin Lab., Dept. of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA.
- Dr. Cobb, Cellular Injury and Adaptation Laboratory, Washington University School of Medicine, St. Louis, MO, USA.
- Dr. Kathleen Marchal, CMFG (Centre of Microbial and Plant Genetics) K.U. Leuven, Kasteelpark Arenberg, Leuven-Heverlee Belgium

Projects in development

- **Coupling quorum sensing and nutritional signals to modulate virulence gene expression, motility and biofilm formation in V. cholerae**, NIH project, in collaboration with Alberto Pascual Montano, Computer Architecture Department, Facultad de Ciencias Físicas, Universidad Complutense de Madrid, Integromics, and Jorge A. Benitez, Morehouse School of Medicine Department of Microbiology, Biochemistry and Immunology, SW Atlanta, Georgia, USA.
- **Refinement of bacterial annotation and regulatory networks for white biotechnology applications** in collaboration with HUSAR Bioinformatics Laben el German Cancer Research Center (DKFZ), Ministerio de Educación y Ciencia de España
- **Andalusian group of Omics** in collaboration with Consejería de, Innovación Ciencia y Empresa, and several companies from el Campus de la Salud de Granada, Spain.

Teaching

Design and development of the teaching materials of the courses:

- Computational Biology, DECSAI, University of Granada, Spain.
- Applications in Bioinformatics, Master in softcomputing and intelligent systems, DECSAI, University of Granada, Spain.
- Developments of tools for teaching bioinformatics, teaching innovation project (2007), University of Granada, Spain.
- Fundaments and Applications to Bioinformatics, PhD. Program, Universidad Tecnológica Metropolitana de Chile, Chile – UGR, Spain
- Interaction Networks and Array Analysis, Master Program in Computational Biotechnology, University of San Pablo, Spain.
4 References