

GENetic NETWORKs: Emergence and Complexity

François Képès

Genopole, CNRS UPS3201, PRES UniverSud Paris, University of Evry

Genopole Campus 1 - Genavenir 6

5 rue Henri Desbruères - F-91030 EVRY cedex

Telephone : 331 69 47 44 31 - Mobile : 336 70 87 11 46

Assistant : 331 69 47 44 30 - Fax : 331 69 47 44 37

Francois.Kepes@epigenomique.genopole.fr

We dwell in and are inhabited by complex systems. Think for instance of our human societies or of the web, and of our cells or of our organs. Complex systems share a number of hallmark features, often including emergence, multiple scales and satisfaction of multiple goals. Emergence is sometimes defined in loose terms by stating that the whole is more than the sum of its parts. Comprehending a complex object often requires simultaneous descriptions at several — time or space — scales. A complex object sometimes can successfully tackle multiple challenges, *e.g.* we thrive in cold and hot weather.

This chapter describes the research results achieved by the partners of the GENNETEC consortium (Fig. 1), who took their inspiration from observations on networks of genetic interactions (Fig. 2). From this bio-inspiration, the consortium provides a strict definition of emergence. It goes further in proposing a framework to simultaneously describe a given complex object at multiple scales. It also addresses the optimization and potentially the design of a complex system which would respond successfully to different challenges. Through this theoretical core, the GENNETEC project¹ is driven towards two applications. The first application is software engineering, a branch of computer science. The second application is the discovery or completion of genetic networks from incomplete data, a branch of bioinformatics. The latter application is subject to development into a commercial product by a company.

¹ The GENETEC project started on the 1st September, 2006, and its duration was 39 months. See the project website at <http://gennetec.csregistry.org/> for more details.


<div>  <div>TEAMS & PARTICIPANTS</div> </div>			
Participant no.	Participant organisation name	Participant org. short name Main researchers	Country
CO01	Centre National de la Recherche Scientifique, Evry	CNRS M. Aiguier, P. Le Gall, F. Képès, M. Manceny, M. Mabrouki, M. Choukroune, J. Hérisson, I. Junier, M. Elati	France
P02	Institute for Scientific Interchange, Turin	ISI M. Weigt, M. Leone, A. Pagnani, A. Braunstein, Sumedha, A. Procaccini, A. Lage-Castellanos, B. Lunt	Italy
P03	Instituto Superior Técnico, Lisbon	IST R. Dilao, D. Muraro	Portugal
P04	INRIA - Université de Paris-Sud, Orsay	INRIA M. Schoenauer, O. Martin, M. Nicolau, T. Jörg, P. Sulc, D. Fichera	France
P05	International Centre for Theoretical Physics, Trieste	ICTP R. Zecchina, M. Marsili, G. Bianconi, T. Galla, Chatterjee	Italy
P06	NORAY BIOINFORMATICS, Bilbao	NB J. Font, D. Fernandez, E. Gonzalez, I. Bilbao	Spain

Figure 1.

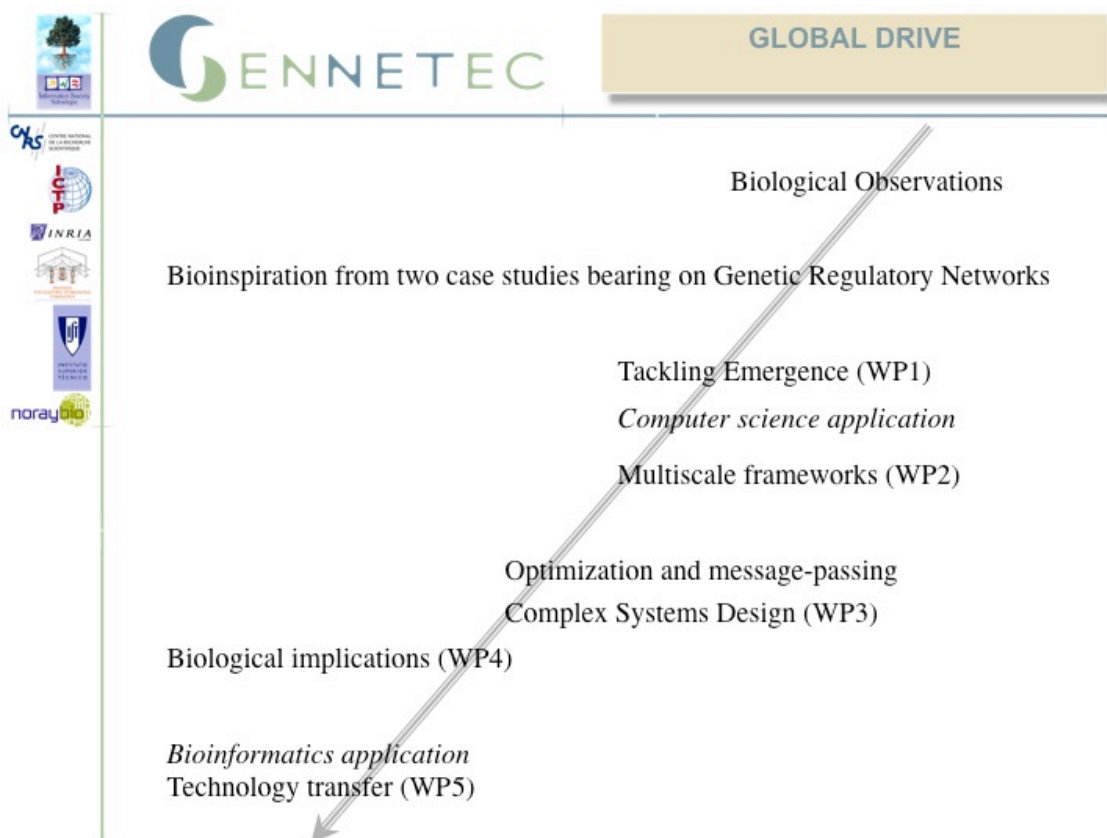


Figure 2.

In the following, a few significant achievements of the GENNETEC project are summarized. For the sake of simplicity, only the major senior author is mentioned for a given workpiece. However, it must be stressed that most GENNETEC achievements were a consequence of tight cross-disciplinary work involving several teams, which would not have been possible without support at the European level. Each team comprised junior scientists who are listed in Fig. 1.

1. Properties of networks of regulatory interactions

Networks of regulatory interactions exhibit various types of properties. The network topology is of interest, but we also would like to access to the behaviour of these networks in time and space. While unfolding networks in time to provide a dynamics is quite common, it is an originality of GENNETEC to also unfold them in the cellular space.

1.1. Network topology

When assessing some property of a network, it would be interesting to also measure how relevant is this property in defining the specificity of the natural network with respect to a set of randomized

networks. Matteo Marsili et al. devised a universal indicator to estimate exactly that relevance. It was applied to demonstrate feature significance where other methods failed. For instance in Fig. 3 it is visually apparent that a pattern exists that links the abundances of proteins 1 and 2 (plotted on both axis) and the probability that they interact (color density). More generally, the concept of randomized network ensembles has been proposed by Ginestra Bianconi as a tool to characterize empirical networks and their constituents with respect to various levels of constraints on the generation of these networks.

As mentioned above, a hallmark of complex systems is its multi-scale nature. Traditionally, the multiple scales have hampered proper clustering of network components that would belong together given some criterion. Martin Weigt et al. have applied advanced methods from statistical physics to devise an algorithm able to recognize multi-scale data organization. It suffices to teach the algorithm where a few data points belong, to obtain a very efficient clustering of data sets, even though they may have involved or contorted shapes (Fig. 4).

Olivier Martin et al. addressed the difficult problem of sampling and counting systems with multiple scales and devised a very efficient algorithm. They applied their algorithm to counting the huge number of RNA sequences that would fold to a given conformation.

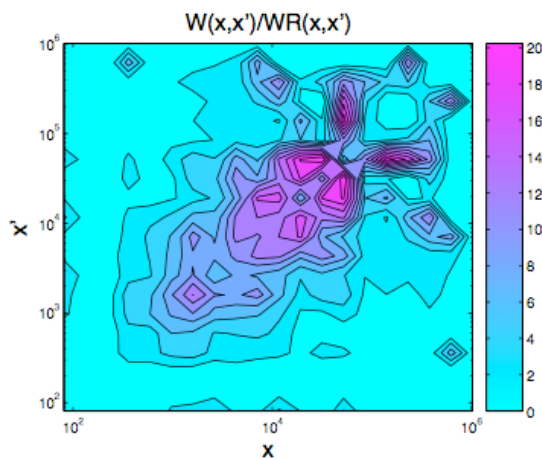


Figure 3.

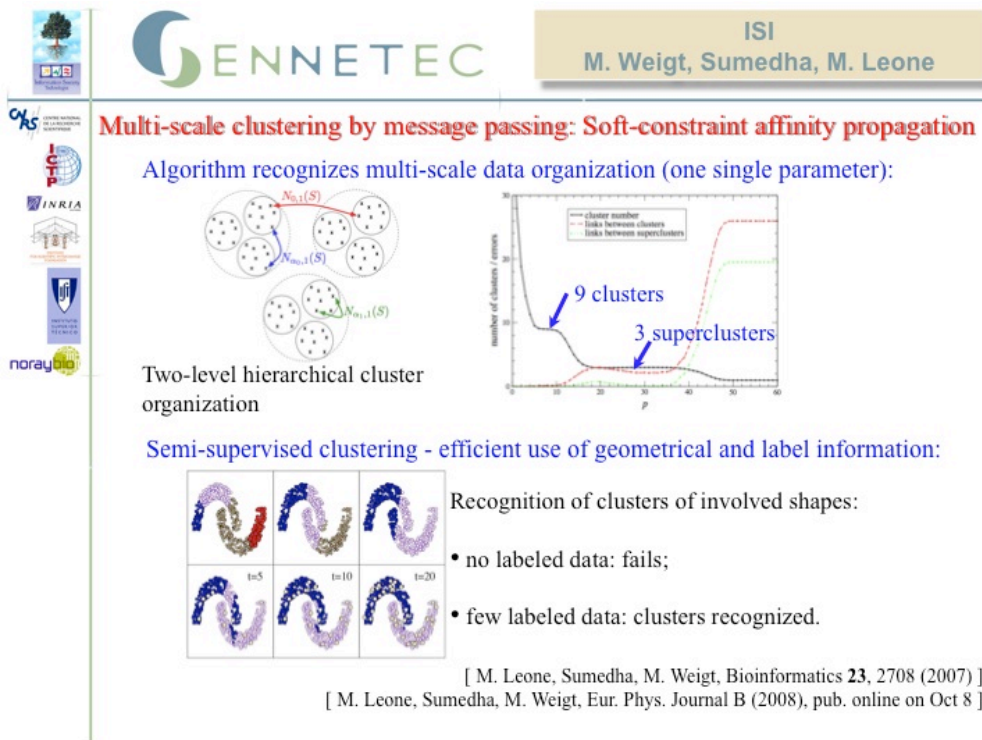


Figure 4.

Several partners focussed on an artificial network of regulatory interactions proposed by Wolfgang Banzhaf in previous work. Marc Schoenauer et al. used it to generate networks with scale-free topology and exceptional evolvability. They successfully evolved these networks to tackle the pole balancing problem, a benchmark for reinforcement learning (Fig. 5).

Olivier Martin showed that the Banzhaf regulatory networks could spontaneously display a small number (typically one) of essential interactions between pairs of regulators; given that this parsimony feature had not been introduced explicitly in the model, it could be considered as an emergent property of the system.

1.2. Network unfolding in time and space

Logical analysis allows to enumerate all the possible dynamics that are consistent with a given network topology. However, for even a small network, the number of possible dynamics can be enormous. Prior knowledge may be used to decrease the number of possible dynamics. This prior knowledge may consist of temporal properties (A always occurs after B) or, more novel, of spatial properties (A and B are close neighbors in space and thus their network interaction is physically stronger than other interactions). Marc Aiguier, Pascale Le Gall et al. have exploited this notion of privileged interaction based on spatial information to reduce the number of dynamics to be considered for small biological networks and for Banzhaf's artificial network (Fig. 6).

A complex system is characterized by an important set of components that mutually interact. This complexity implies the emergence of properties which are true properties at the local level (subnetwork) but questioned at the global level (whole network). The challenge is then to give conditions that the system must fulfill to not possess such properties, *i.e.* to ensure that the whole equals the sum of its parts. This has applications in software engineering, where precisely emergent properties are undesirable because they oppose the engineering principle of construction. Marc Aiguier, Pascale Le Gall et al. were able to demonstrate conditions that preserve properties through embedding of a well-characterized subnetwork into a larger network (Fig. 7; the issue is whether an additional regulator with two arrows pointing towards the subnetwork can be added while preserving the dynamical subnetwork properties symbolized below).

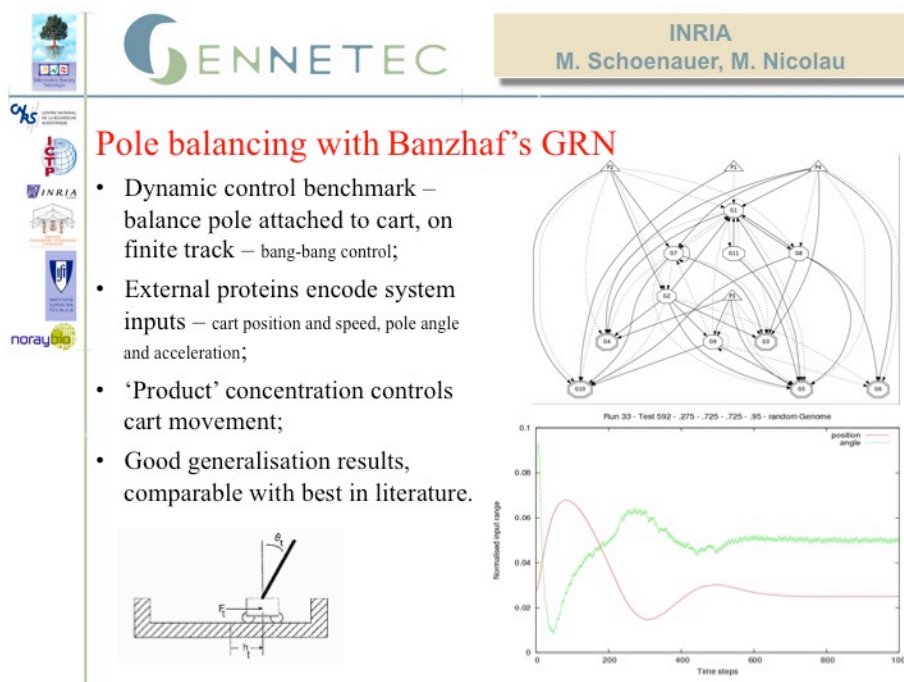


Figure 5.

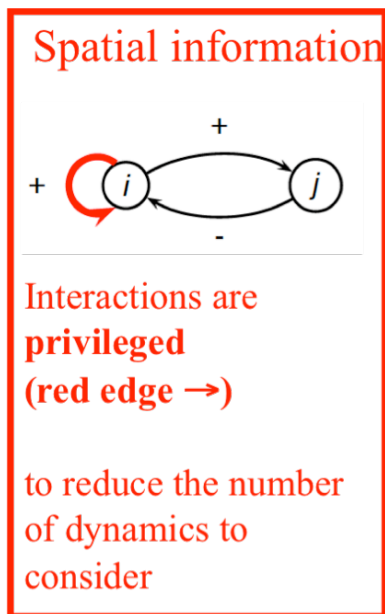


Figure 6.

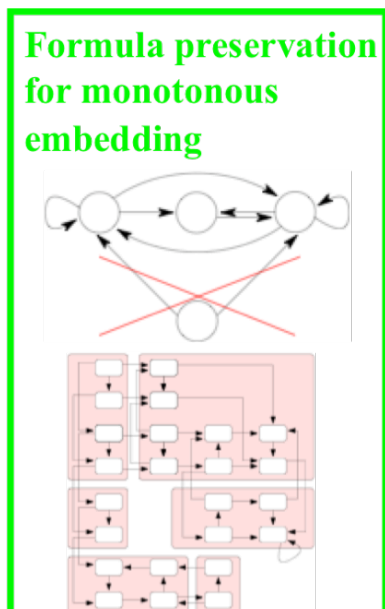


Figure 7.

2. Optimization of a system with multiple objectives

As an example system with multiple objectives and an underlying network of regulatory interactions, Rui Dilão et al. used pattern formation in fly early development. The novelty was to introduce an evolutionary algorithm strategy to optimize the regulatory network for two simultaneous objectives. It was applied to predict the distribution of regulators along the fly embryo (Fig. 8). Among them, the HB and KNI regulators were predicted correctly as compared to

experimental data. For HKB however, no experimental data are currently available and this prediction may be used to guide experiments.

Another success of this partner is their prediction in 2008 that the gradient-forming molecule responsible for the very first steps of fly development was the mRNA encoding the BCD protein (bicoid) rather than the BCD protein itself, as was believed so far. An article from an external group appeared in 2009 that demonstrated that the mRNA gradient was indeed key to morphogenesis (Fig. 9).

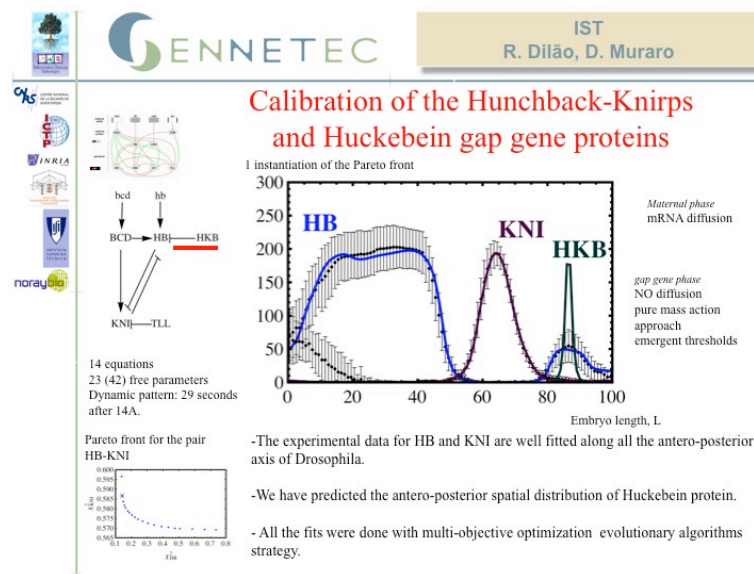


Figure 8.

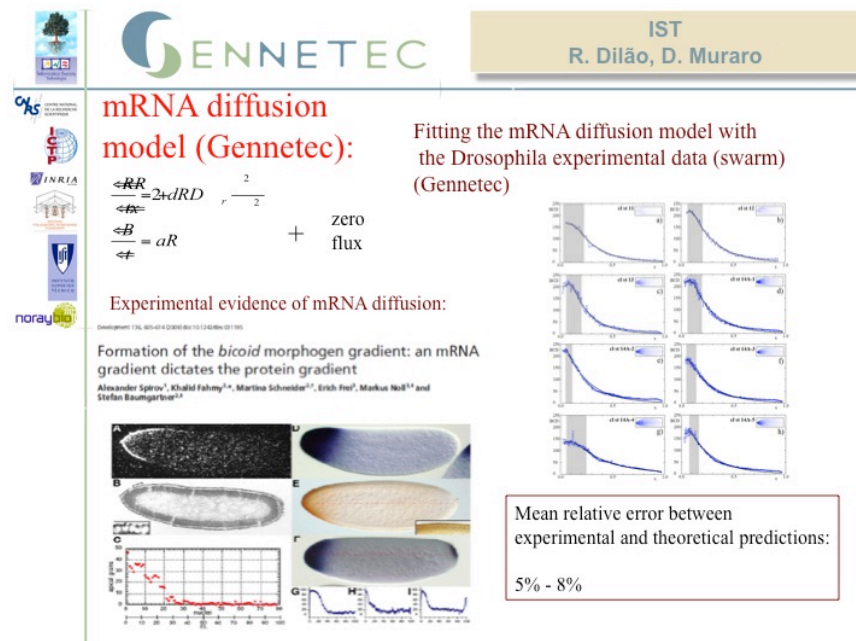


Figure 9.

3. Application to chromosome folding and organization and to discovery of regulatory networks

3.1. Chromosome folding

It was known for some time that in active cells, the first step of gene expression, transcription, occurs at discrete sites rather than in a diffuse manner. Such discrete sites are called transcription factories. Martin Weigt et al. proposed a coarse-grained model of chromosome folding under entropic forces. François Képès et al. subsequently refined the physical model of chromosome folding, by introducing pairwise interactions at defined sites along the polymer to mimic their specific binding by bivalent proteins like transcription regulators. It allowed them with Olivier Martin to demonstrate a transition between a very ordered and a less-ordered state of condensed polymer folding, and to account for the observed transcription factories through a thermodynamically driven self-organizing process (Fig. 10). This physical model was used to show the importance of the arrangement of interacting genes along the chromosome to obtain transcription factories. In particular, periodical gene positioning favors a coil (solenoidal) arrangement.

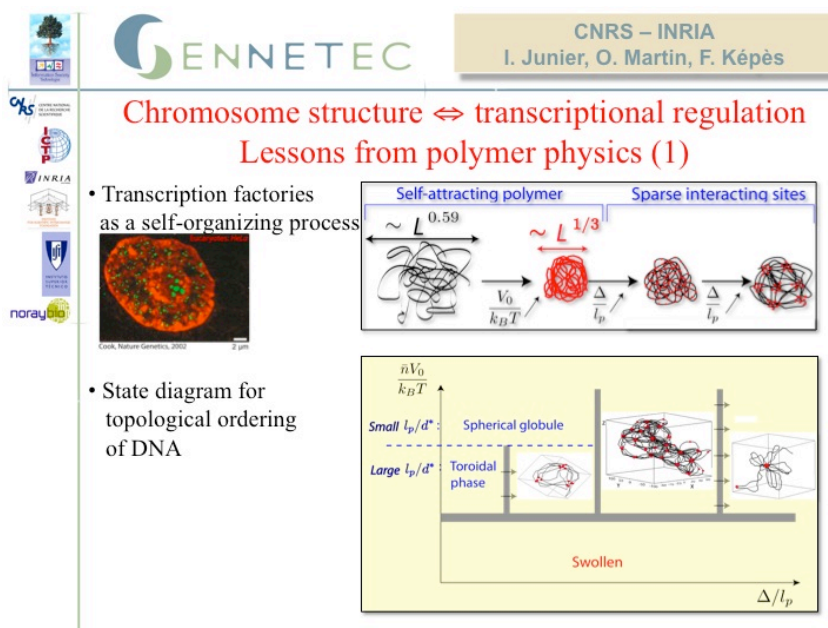


Figure 10.

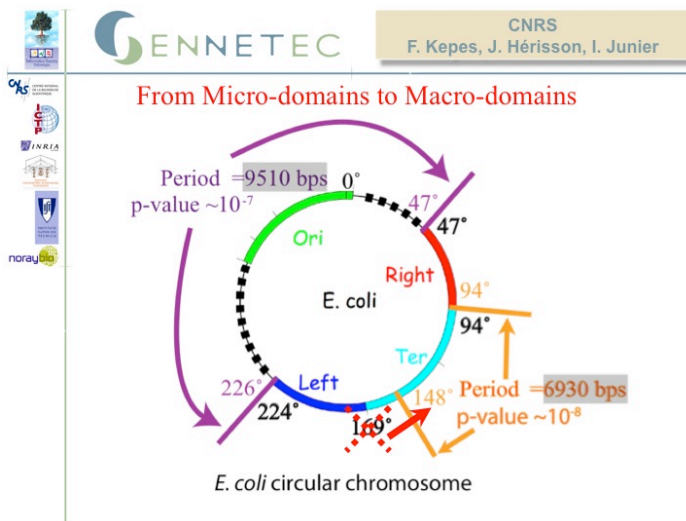



Figure 11.


3.2. Chromosome organization

François Képès et al. devised a new algorithm to detect the periodical positioning of interacting genes along chromosomes, starting from sparse and noisy datasets. This algorithm allowed the discovery of a number of new features in chromosomes from various species. As an illustration, Fig. 11 shows that the algorithm is able to automatically detect the frontiers of the so-called macro-domains, *i.e.* regions of bacterial chromosomes where genes preferentially interact with others in the same region rather than outside this region; it had taken heavy bench experimentation to detect such frontiers. One common observation made with this tool is that interacting genes are in part periodically spaced along chromosomes, which favors solenoidal folding as mentioned above in section 3.1.

3.3. Discovery of regulatory networks and technology transfer

Discovery of regulatory networks through bioinformatics has so far been exclusively based on searching short consensus sequences in the genome sequence. This process generates a high proportion of false hits. The idea is to use positional information as explained above to cut down on false hits. François Képès et al. brought under GENNETEC the proof of principle that positional information decreased the proportion of false hits for two bacteria, yeast and human. Julio Font et al. transferred this observation into a scalable, user-friendly software that optimally combines both types of information to complete genetic networks in various organisms (Fig. 12).





NorayBio Research results

A new machine learning framework for the inference of interaction networks

Two methods for the prediction of the transcription factor binding sites based on the combination of sequence and position have been developed.

1. The first one follows the stacked generalization schema to combine the base level sequential and positional classifiers using naive Bayes and lazy K-star as meta-classifiers.
2. The second method is based on Adaboost. We have varied the algorithm to choose at each iteration among several types of weak classifiers.

- Tests were performed for *E. coli*, *B. subtilis*, *C. glutamicum* and *S. cerevisiae*. The results show how the positional information contributes to the filtering of many false positives, obtaining this way a more accurate prediction of the binding positions for a certain transcription factor.

This work is summarized in the paper entitled: "PreCislon: PREdiction of CIS-regulatory elements based on positIOn", in preparation, that will be submitted to Bioinformatics.




This work has been carried out in close collaboration between: 

Figure 12.

This software will be made available free of charge to the academia, while an advanced version may be commercialized by NorayBio at a later stage (Fig. 13).







NorayBio Technology Transfer results



Development of two software packages that integrate the algorithms developed in the project

The task of technology transfer has produced two software packages:

- WP4/WP5 → application for prediction of TFBS
PreCislon : PREdiction of CIS-regulatory elements based on positIOn.

- WP3/WP5 → Integration of the algorithms developed by WP3 into a software package

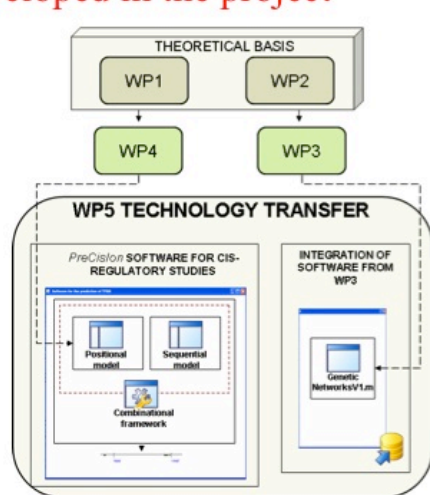


Figure 13.

4. Conclusion

The main achievements of the GENNETEC project were to propose some new concepts and tools based on observations on networks of genetic interactions. These new concepts and tools include the use of randomized network ensembles as a tool to characterize empirical networks and their constituents; algorithms for sampling and counting systems with multiple scales, such as RNA sequences; considering interactions based on spatial information to characterize network dynamics; and identifying conditions that preserve fundamental network properties. Another domain was modelling systems with multiple objectives and an underlying network of regulatory interactions - an evolutionary algorithm strategy was proposed to optimize the regulatory network for two simultaneous objectives. In general terms, these proposals were coherently organized in a framework allowing to simultaneously describe a given complex object at multiple scales and potentially design a system composed by several of these complex objects. Several applications were developed, namely in the domain of discovery or completion of genetic networks from incomplete data.

Bibliography

M'Barka Mabrouki, Marc Aiguier, Jean-Paul Comet, Pascale Le Gall, Adrien Richard, “Embedding of Biological Regulatory Networks and Property Preservation”. Internal Report - Extended version of the conference paper AB'08. In progress.

Matthieu Manceny, Marc Aiguier, Pascale Le Gall, Joan Herisson, Ivan Junier, Francois Kepes, “Spatial Information to Restrict the Dynamics of Genetic Regulatory Networks”. In BICoB 2009 - International Conference on BioInformatics and Computational Biology., LNBI 5462, pp. 270-281, 2009.

Marc Aiguier, Pascale Le Gall, M'Barka Mabrouki, “Complex software systems : Formalization and Applications”. (IARIA) International Journal on Advances in Software , 2:47-62, 2009. Invited paper - Extended version of the conference paper ICSEA'08.

M'Barka Mabrouki, Marc Aiguier, Jean-Paul Comet, Pascale Le Gall, "Property preservation along embedding of biological regulatory networks". In AB 2008 - 3rd International Conference on Algebraic Biology., LNCS 5147, pp. 125-138, 2008.

Marc Aiguier, Pascale Le Gall, M'Barka Mabrouki, "A formal definition of complex software". In ICSEA 2008 - 3rd International Conference on Software Engineering Advances, IEEE Computer Society Press, 2008.

Marc Aiguier, Pascale Le Gall, M'Barka Mabrouki, "Emergent properties in reactive systems". In APSEC 2008 - 15th Asia-Pacific Software Engineering Conference., IEEE Computer Society Press, 2008.

Ginestra Bianconi, "The entropy of randomized network ensembles". Europhys.Lett. 81, 28005 (2008).

Ginestra Bianconi, "The entropy of network ensembles". Phys.Rev. E 79, 016114 (2009).

G.Bianconi N. Gulbahce A. E. Motter, "Local structure of directed networks". Phys. Rev. Lett. 100,118701 (2008).

G. Bianconi and N. Gulbahce, "Algorithm for counting large directed loops". Journal of Physics A 41, 224003 (2008).

T. Galla, "Two-population replicator dynamics and number of Nash equilibria in matrix games",. Exploring the Frontiers of Physics EPL, 78(2007).

G. Bianconi, Anthony C.C. Coolen, Conrad J. Perez Vicente, "Entropies of complex networks with hierarchically constrained topologies". PHYSICAL REVIEW E 78, 016114 2008.

Y. Yoshino, T. Galla, K. Tokita, "Statistical mechanics and stability of a model eco-system". Journal of Statistical Mechanics: Theory and Experiment, J. Stat. Mech. (2007) P09003.

S. Bradde, G. Bianconi. “Percolation transition and distribution of connected components in generalized random network ensembles”. *Journal of Physics A*, 42, 195007 (2009).

S. Bradde, G. Bianconi. “Percolation transition in correlated hypergraphs”. *J. Stat. Mech.* (2009) P07028.

G. Bianconi, P. Pin and M. Marsili. “Assessing the relevance of node features for network structure”. *PNAS* 106 11433 (2009).

G. Bianconi, L. Ferretti and S. Franz. “Non-neutral theory of biodiversity”. *EPL*, 87, 28001 (2009).

Michele Leone, Sumedha, Martin Weigt. “Clustering by soft-constraint affinity propagation: Applications to gene-expression data”. *Bioinformatics* 23, 2708 (2007); DOI: 10.1093/bioinformatics/btm414

Michele Leone, Sumedha, Martin Weigt. “Unsupervised and semi-supervised clustering by message passing: Soft-constraint affinity propagation”. *European Physical Journal B* (2008); DOI: 10.1140/epjb/e2008-00381-8

G. Bianconi and R. Zecchina. “Viable fluxes in the metabolic network”. *Networks and Heterogeneous Media* 3, 361 (2008).

Thomas Jorg, Olivier C Martin, Andreas Wagner. “Neutral network sizes of biological RNA molecules can be computed and are not atypically small”. *BMC Bioinformatics* 2008, 9:464 doi: 10.1186/1471-2105-9-464

Rui Dilão, “Emergence of a Collective Steady State and Symmetry Breaking in Systems of two Identical Cells”. In *BIOMAT 2006, International Symposium on Mathematical and Computational Biology*, R. P. Mondaini et R. Dilão (ed.), pp. 25-36, World Scientific, 2007.

G. Bianconi, “Flux distribution of the metabolic network close to optimal biomass production”. *Phys. Rev. E* 78 035101 (2008).

Rui Dilão, Daniele Muraro, Miguel Nicolau, Marc Schoenauer. “Validation of a morphogenesis model of *Drosophila* early development by a multi-objective evolutionary optimization algorithm”. In *EvoBIO'09, Proc. 7th European Conference on Evolutionary Computation, ML and Data Mining in BioInformatics*, April 2009. Best Paper Award.

Miguel Nicolau, Marc Schoenauer. “Evolving Specific Network Statistical Properties using a Gene Regulatory Network Model”. In *GECCO 2009, Genetic and Evolutionary Computation Conference*, ACM Press, 2009.

Miguel Nicolau, Marc Schoenauer, “Evolving Scale-Free Topologies using a Gene Regulatory Network Model”. In *CEC 2008, IEEE Congress on Evolutionary Computation*, pp. 3748-3755, IEEE Press, 2008.

Olivier C Martin, Andreas Wagner. “Effects of Recombination on Complex Regulatory Circuits”. *Genetics* 2009, 183: 673-684.

A. Lage-Castellanos, A. Pagnani, M. Weigt. “Statistical mechanics of sparse generalization and model selection”. *J. Stat. Mech.* P10009 (2009).

B. Lunt, H. Szurmant, A. Procaccini, J.A. Hoch, T. Hwa, M. Weigt. “Inference of Direct Residue Contacts in Two-Component Signaling”. *Methods in Enzymology* (2009).

A. Lage-Castellanos, A. Pagnani, M. Weigt. “Sparse generalization and graphical model selection”. *IEEE Proc. 47th Ann. Allerton Conf. on Communication, Control, and Computing* (2009).

C. Martelli, A. De Martino, E. Marinari, M. Marsili, I. Perez Castillo. “Identifying essential genes in *E. coli* from a metabolic optimization principle”. *PNAS* 106 2607 (2009).

Zhang, X., Furtlehner, C., Sebag, M.. “Data Streaming with Affinity Propagation”. in Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2008.

Zhang, X., Furtlehner, C., Sebag, M.. “Distributed and Incremental Clustering Based on Weighted Affinity Propagation”. In Fourth European Starting AI Researcher Symposium (STAIRS) (2008).

Zhang, X. , Furtlehner, C. , Perez, J. , Germain, C. and Sebag, M.. “Toward Autonomic Grids: Analyzing the Job Flow with Affinity Streaming”. In: 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). Paris France. 2009.

Furtlehner, C., Sebag, M., Zhang, X. . “Scaling Analysis of the Affinity Propagation Algorithm”. Technical Report INRIA-RR-7046 (2009).