Computational Discovery of Cis-Regulatory Sites in Eukaryotic Genomes by Collective Inference

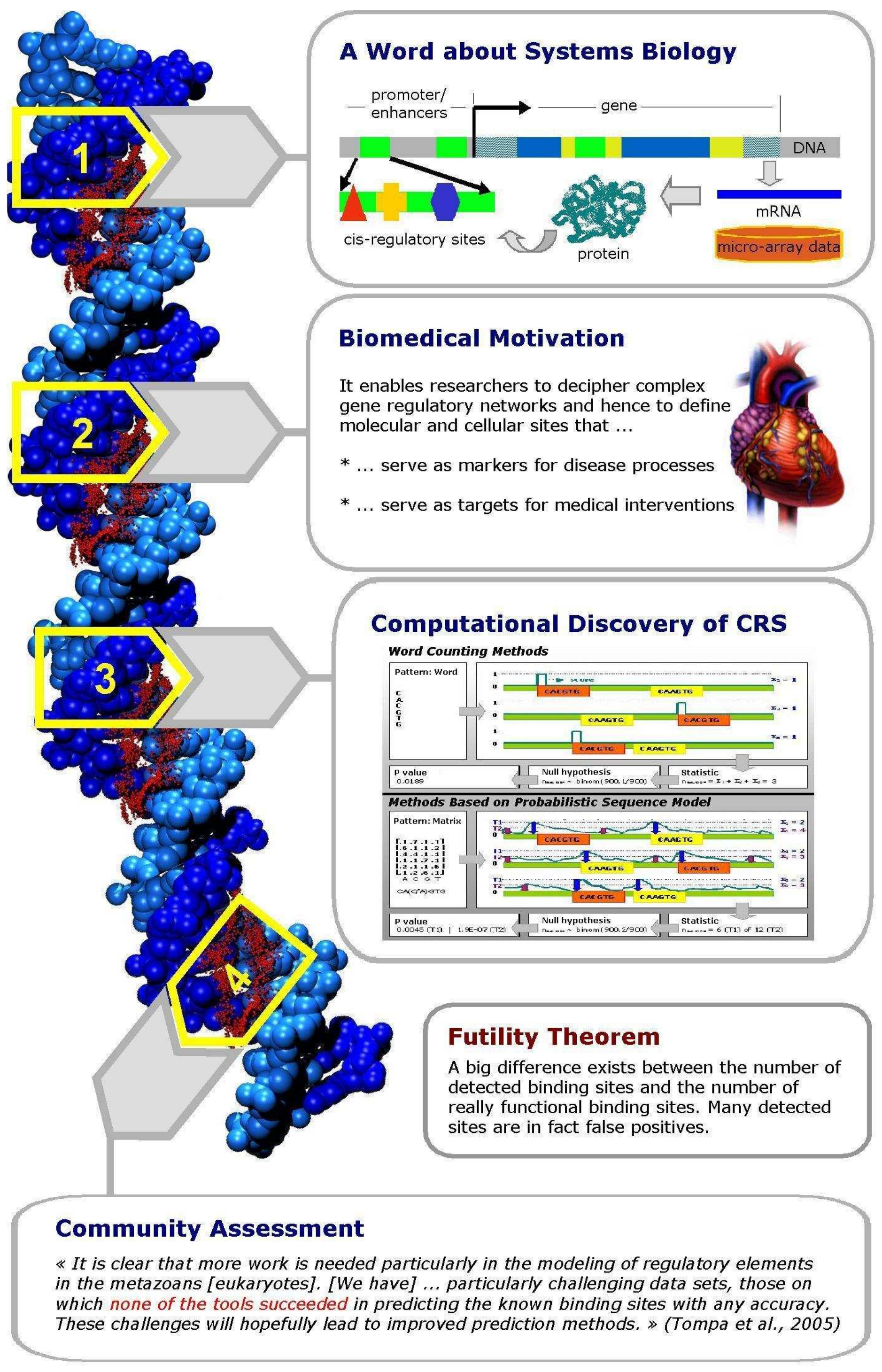
Wouter Van Delm Yves Moreau Bart De Moor

Dept. of Electrical Engineering (ESAT) - SCD - SISTA
Katholieke Universiteit Leuven, Belgium
`wvandelm@esat.kuleuven.be`

1 Problem Setting

Abstract: *Motif discovery in eukaryotic genomes is an abstraction process that describes regulatory DNA sequences as collections of cis-regulatory sites. The result enables researchers to decipher complex gene regulatory networks. In silico methods prioritize cis-regulatory sites and aim to increase the efficiency of arduous wet lab research. Computational discovery of cis-regulatory sites is however often an ill-posed problem, resulting in many false positive detections.*

1.1 Motif Discovery in Eukaryotes



A Word about Systems Biology

Biomedical Motivation

It enables researchers to decipher complex gene regulatory networks and hence to define molecular and cellular sites that ...

* ... serve as markers for disease processes

* ... serve as targets for medical interventions

Computational Discovery of CRS

Word Counting Methods

Methods Based on Probabilistic Sequence Model

Futility Theorem

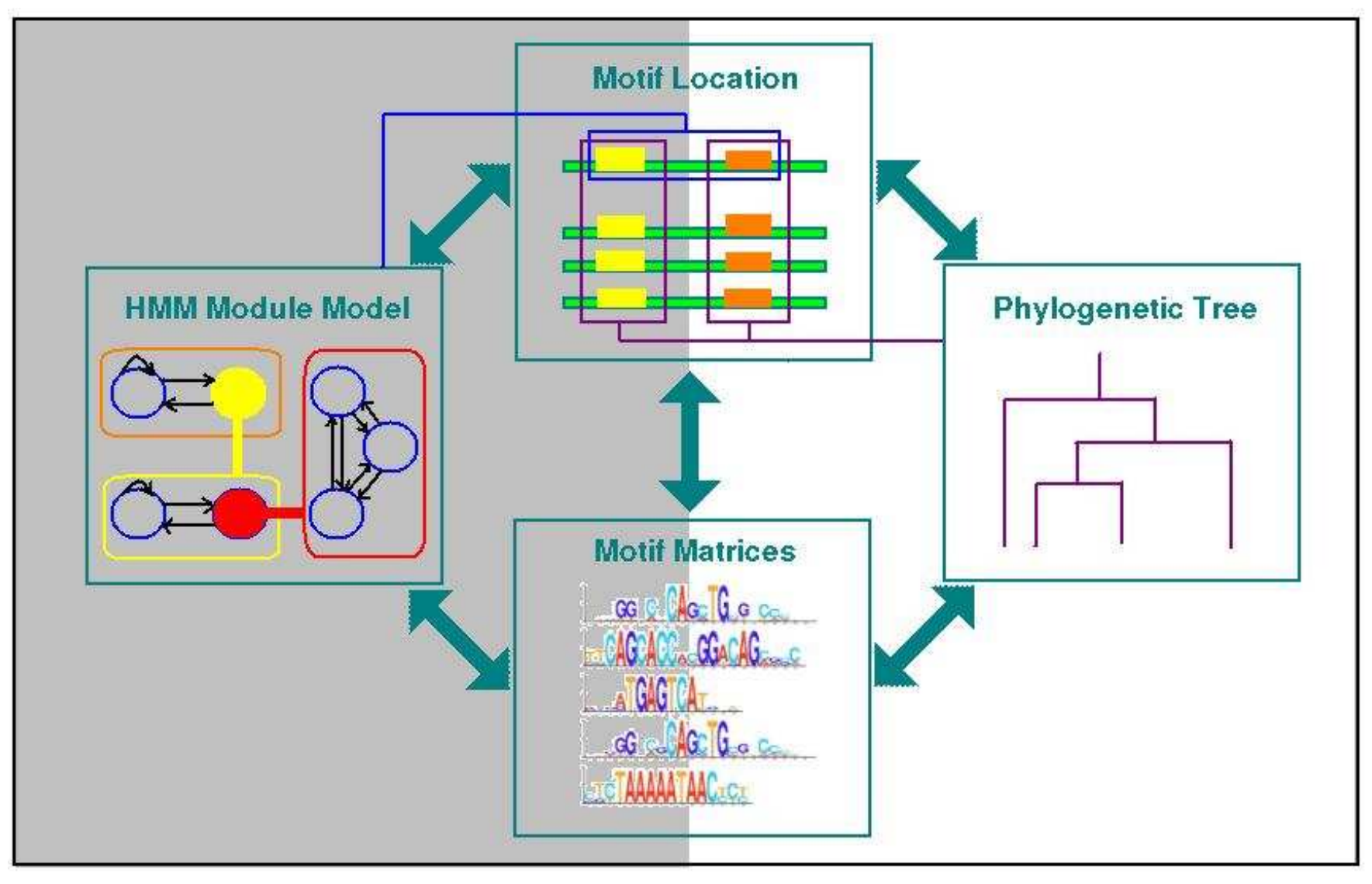
A big difference exists between the number of detected binding sites and the number of really functional binding sites. Many detected sites are in fact false positives.

Community Assessment

« It is clear that more work is needed particularly in the modeling of regulatory elements in the metazoans [eukaryotes]. [We have] ... particularly challenging data sets, those on which none of the tools succeeded in predicting the known binding sites with any accuracy. These challenges will hopefully lead to improved prediction methods. » (Tompa et al., 2005)

1.2 State of the Art

CRS Modules [1] vs Evolutionary Conservation [2]



Characteristics:

- Knowledge representation with graphical models
- Bayesian inference with Monte Carlo methods
- Modular computation on heterogeneous data

2 Approach

Abstract: *We formulate the problem as a constraint-based inference problem (CBIP) and infer a set of sites (cis-regulatory module) simultaneously by adding consistency constraints in an attempt to make the problem well-posed. A collective of computational agents can serve as a framework to solve the combined CBIP. Each agent infers part of biological reality given possibly different data sources. Synergy emerges as in bounded rational noncooperative games.*

2.1 Constraint-Based Inference in Games

A CBIP is a tuple $(\mathbb{X}, \mathbb{D}^*, R, \mathfrak{C})$ [3]. The general approach enhances knowledge integration and cross-fertilization of solution algorithms.

Example. A HMM in CBIP terms :

$$\mathbb{X}^t = \{O^t, H^t, \theta\}\ ,\ \mathbb{D}^* = \{\mathbb{D}_O, \mathbb{D}_H, \mathbb{D}_\theta\}$$
$$R = ([0, \inf), +, \times)\ ,\ \mathfrak{C}^t = \{\phi_O, \phi_H\}$$

with $\phi_O = p(O^t|H^t, \theta), \phi_H = p(H^t|H^{t-1}, \theta)$, $\mathfrak{C}^0 = \{p(\theta)\}$. ♦

The (Bayesian-like) inference task becomes:

$$\phi_{X_1}(x_1) \propto \bigoplus_{X_2} \bigotimes_{\phi\in\mathfrak{C}} \phi$$

We aim for a more tractable modular computation by considering the product constraint

$$\mathcal{Q}(x_1, x_2) = \mathcal{Q}_{X_1}(x_1) \otimes \mathcal{Q}_{X_2}(x_2) \approx \bigotimes_{\phi\in\mathfrak{C}} \phi$$

in analogy to a mean field approximation. A collective of agents $\mathcal{A}_{X_i}$ compute the $\mathcal{Q}_{X_i}$ by solving $\mathfrak{C}_i$ - the subsystems of constraints involving X_i - as generalized max-sat problems with Lagrangian

$$\mathcal{L}(\mathcal{Q}_{X_i}) = k_c(\mathcal{Q}_{X_i}) \oplus \bigoplus_{\phi_j\in\mathfrak{C}_i} (w_j \bigoplus_{X_{\mathfrak{C}_i}} k_{\phi_j}(\mathcal{Q}_{X_i}, \mathcal{Q}_{X_{\mathfrak{C}_i}\setminus X_i}))$$

Global dynamics are similar to that of bounded rational noncooperative games.

Example. Lagrangian (to minimize) and update rule of probability collective used in pilot study [4]:

$$\begin{aligned}\mathcal{L}(\mathcal{Q}_{X_i}) &= E_{\mathcal{Q}_{X_{\mathfrak{C}_i}}}[G] - TS(\mathcal{Q}_{X_i})\\ \mathcal{Q}_{X_i,t+1} &= \mathcal{Q}_{X_i,t} - \alpha\mathcal{Q}_{X_i,t}\{\frac{E[G|x_i] - E[G]}{T} + log(\mathcal{Q}_{X_i,t}) + S(\mathcal{Q}_{X_i,t})\}\end{aligned}$$

with $G = log(\prod_{p_j\in\mathfrak{C}_i} p_j)$ and $S(\mathcal{Q}_{X_i}) = -\int \mathcal{Q}_{X_i} log(\mathcal{Q}_{X_i})dX_i$. We use a sample based version by considering $\mathcal{Q}_{X_i,t}$ as proposal distribution for $\mathcal{Q}_{X_i,t+1}$. ♦

2.2 Parallel Computation in a MAS

Characteristics:

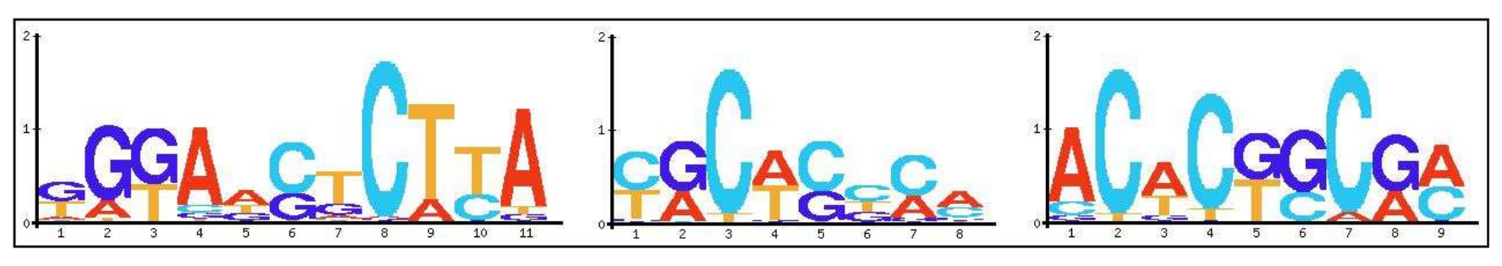
- Heterogeneous multi-agent system
- No central control, synchronous update
- Non-active / emergent cooperation
- Agents have alternative utility functions

3 Preliminary Results

Abstract: *In a pilot study, we implemented a 3-agent system (module structure, location, motif matrices). We observed the performance on a synthetic dataset and searched for heart-specific cis-regulatory modules in a small case study. Preliminary results suggest that the sketched approach can deal consistently with heterogeneous data and allows for efficient distributed processing. The observations encourage us to investigate further the design of the coordination scheme.*

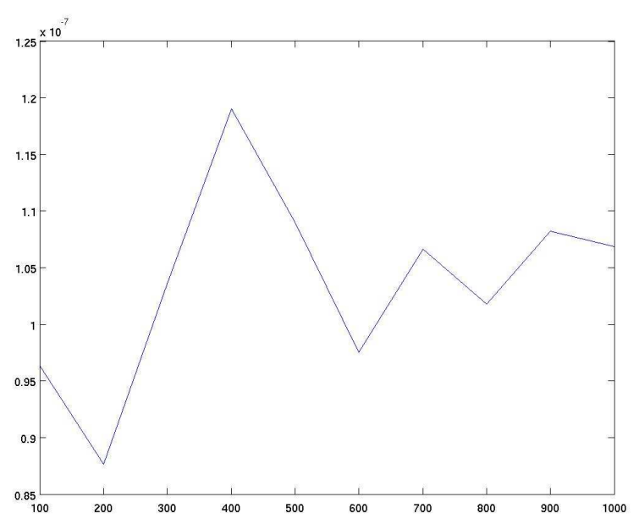
3.1 Performance Analysis

Synthetic data with realistic SNR.



The outcome of interest is the output of $\mathcal{A}_{matrix}$: a sample with combinations made out of 20 matrices.

Performance was measured on the average combination (PD over the 20 matrices). The effect of the sample size (variance) seemed to be important.



Sample size (X) versus performance (Y).

Since the game may have more than one Nash equilibrium, convergence to suboptimal solutions needs to be investigated more in the future.

3.2 Case Study

We searched for a heart specific cis-regulatory module. $\mathcal{A}_{matrix}$ used data from Transfac and MotifSampler. Validation of results with Endeavour was not convincing so far.

References

(1) Gupta M and Liu JS, 'De novo cis-regulatory module elicitation for eukaryotic genomes'. PNAS, 102.20, 7079-7084. (2005)
(2) Li X and Wong WH, 'Sampling motifs on phylogenetic trees'. PNAS, 102.27, 9481-9486. (2005)
(3) Chang L and Mackworth AK, 'Constraint-based inference: a bridge between constraint processing and probability inference'. Proceedings of Principle and Practice of Constraint Programming - CP 2005, Springer-Verlag. (2005)
(4) Wolpert DH, 'Information Theory The Bridge Connecting Bounded Rational Game Theory and Statistical Physics'. In Complex Engineering Systems, edited by Yaneer Bar-Yam and Dan Braha, Perseus books. (2005)

Acknowledgements

We thank dr. Torik Ayoubi of the Genome Center in Maastricht for providing the case study data. Research funded by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). Dr. Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Research supported by * Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, IDO (Genetic networks), several PhD/postdoc & fellow grants, * Flemish Government: o FWO: PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel) research communities (ICCoS, ANMMM, MLDM); o IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; * Belgian Federal Science Policy Office: IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling, 2002-2006); * EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain.