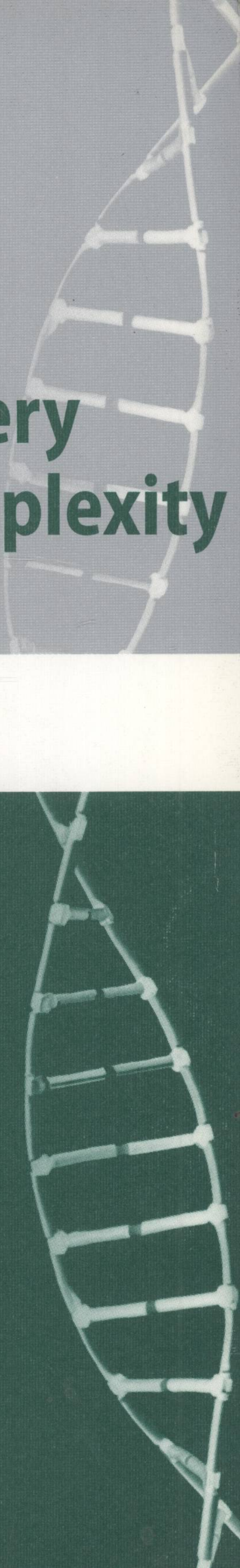
Karl Tuyls
Ronald Westra
Yvan Saeys
Ann Nowé (Eds.)

LNBI 4366

Knowledge Discovery and Emergent Complexity in Bioinformatics

First International Workshop, KDECB 2006
Ghent, Belgium, May 2006
Revised Selected Papers



Springer

Karl Tuyls Ronald Westra
Yvan Saeys Ann Nowé (Eds.)

Knowledge Discovery and Emergent Complexity in Bioinformatics

First International Workshop, KDECB 2006
Ghent, Belgium, May 10, 2006
Revised Selected Papers

Springer

Volume Editors

Karl Tuyls
Maastricht University, Faculty of Humanities and Science
Maastricht ICT Competence Center, 6200 MD Maastricht, The Netherlands
E-mail: k.tuyls@micc.unimaas.nl

Ronald Westra
Maastricht University, Department of Mathematics
6200 MD Maastricht, The Netherlands
E-mail: westra@math.unimaas.nl

Yvan Saeys
Ghent University
Technologiepark 927, 9052 Ghent, Belgium
E-mail: yvan.saeys@ugent.be

Ann Nowé
Vrije Universiteit Brussel, Faculty of Sciences (WE)
Department of Computer Science, Pleinlaan 2, 1050 Brussels, Belgium
E-mail: ann.nowe@vub.ac.be

Library of Congress Control Number: 2007920820

CR Subject Classification (1998): H.2.8, I.5, J.3, I.2, H.3, F.1-2

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-540-71036-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-71036-3 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12023929 06/3142 5 4 3 2 1 0

Preface

This book contains selected and revised papers of the International Symposium on Knowledge Discovery and Emergent Complexity in Bioinformatics (KDECB 2006), held at the University of Ghent, Belgium, May 10, 2006.

In February 1943, the Austrian physicist Erwin Schrödinger, one of the founding fathers of quantum mechanics, gave a series of lectures at Trinity College in Dublin titled "What Is Life? The Physical Aspect of the Living Cell and Mind." In these lectures Schrödinger stressed the fundamental differences encountered between observing animate and inanimate matter, and advanced some, at the time, audacious hypotheses about the nature and molecular structure of genes, some ten years before the discoveries of Watson and Crick. Indeed, the rules of living matter, from the molecular level to the level of supraorganic flocking behavior, seem to violate the simple basic interactions found between fundamental particles as electrons and protons. It is as if the organic molecules in the cell 'know' that they are alive. Despite all external stochastic fluctuations and chaos, process and additive noise, this machinery has been ticking for at least 3.8 billion years. Yet, we may safely assume that the laws that govern physics also steer these complex associations of synchronous and seemingly intentional dynamics in the cell. Contrary to the few simple laws that govern the interactions between the few really elementary particles, there are at least tens of thousands of different genes and proteins, with millions of possible interactions, and each of these interactions obeys its own peculiarities. There are different processes involved such as transcription, translation and subsequent folding. How can we ever understand the fundamentals of these complex interactions that emerge from the few empirical observations we are able to make?

The KDECB 2006 Workshop was a great success and provided a forum for the presentation of new ideas and results bearing on the conception of knowledge discovery and emergent complexity in bioinformatics. This event was organized in connection with the 15th Belgium-Netherlands Conference on Machine Learning, held in Ghent, Belgium. The goal of this workshop and this associated book is to increase awareness and interest in knowledge discovery and emergent complexity research in bioinformatics, and to encourage collaboration between machine learning experts, computational biology experts, mathematicians and physicists, so as to give a representative overview of the current state of affairs in this area. Next to a strong program with lectures by leading scientists in this multidisciplinary field, the book contains contributions on how knowledge can be extracted from sophisticated biological systems. Different disciplines, both 'wet' and 'dry,' have contributed to these developments and they will also benefit directly or indirectly from new, intelligent, computational techniques.

Hence, we welcomed scientists and practitioners from several European countries and different scientific areas in Ghent for the 1st Workshop on Knowledge Discovery and Emergent Complexity in Bioinformatics (KDECB 2006).

We hope that our readers will enjoy reading the efforts of the researchers.

Acknowledgements

Organizing a scientific event like KDECB, and editing an associated book, requires the help of many enthusiastic people. First of all, the organizers would like to thank the members of the Program Committee who guaranteed a scientifically strong and interesting LNBI volume. Secondly, we would like to express our appreciation to the invited speakers, Ricardo Grau, Reinhardt Guthke, William H. Majoros, Stan Maree, Grzegorz Rozenberg and Jean-Philippe Vert, for their distinguished contributions to the symposium program. Finally, we would also like to thank the authors of all contributions for submitting their scientific work to the KDECB symposium!

December 2006

Karl Tuyls
Ronald Westra
Yvan Saeys
Ann Nowé

Organization

Organizing Committee

Co-chairs:

Karl Tuyls
Ronald Westra
Yvan Saeys
Ann Nowé

Program Committee

Adelmo Cechin
Jeroen Donkers
Reinhard Gutkhe
Thomas Hamelryck
Nicolas Le Novere
Derek Linkens
Bernard Manderick
Ann Nowé
Yvan Saeys
Klaus Stiefel
Elena Tsiporkova
Bram Vanschoenwinkel
Ronald Westra

Lecture Notes in Bioinformatics

Vol. 4366: K. Tuyls, R. Westra, Y. Saeys, A. Nowé (Eds.), Knowledge Discovery and Emergent Complexity in Bioinformatics. IX, 183 pages. 2007.

Vol. 4360: W. Dubitzky, A. Schuster, P.M.A. Sloot, M. Schroeder, M. Romberg (Eds.), Distributed, High-Performance and Grid Computing in Computational Biology. X, 192 pages. 2007.

Vol. 4345: N. Maglaveras, I. Chouvarda, V. Koutkias, R. Brause (Eds.), Biological and Medical Data Analysis. XIII, 496 pages. 2006.

Vol. 4316: M.M. Dalkilic, S. Kim, J. Yang (Eds.), Data Mining and Bioinformatics. VIII, 197 pages. 2006.

Vol. 4230: C. Priami, A. Ingólfsdóttir, B. Mishra, H.R. Nielson (Eds.), Transactions on Computational Systems Biology VII. VII, 185 pages. 2006.

Vol. 4220: C. Priami, G. Plotkin (Eds.), Transactions on Computational Systems Biology VI. VII, 247 pages. 2006.

Vol. 4216: M.R. Berthold, R.C. Glen, I. Fischer (Eds.), Computational Life Sciences II. XIII, 269 pages. 2006.

Vol. 4210: C. Priami (Ed.), Computational Methods in Systems Biology. X, 323 pages. 2006.

Vol. 4205: G. Bourque, N. El-Mabrouk (Eds.), Comparative Genomics. X, 231 pages. 2006.

Vol. 4175: P. Bücher, B.M.E. Moret (Eds.), Algorithms in Bioinformatics. XII, 402 pages. 2006.

Vol. 4146: J.C. Rajapakse, L. Wong, R. Acharya (Eds.), Pattern Recognition in Bioinformatics. XIV, 186 pages. 2006.

Vol. 4115: D.-S. Huang, K. Li, G.W. Irwin (Eds.), Computational Intelligence and Bioinformatics, Part III. XXI, 803 pages. 2006.

Vol. 4075: U. Leser, F. Naumann, B. Eckman (Eds.), Data Integration in the Life Sciences. XI, 298 pages. 2006.

Vol. 4070: C. Priami, X. Hu, Y. Pan, T.Y. Lin (Eds.), Transactions on Computational Systems Biology V. IX, 129 pages. 2006.

Vol. 4023: E. Eskin, T. Ideker, B. Raphael, C. Workman (Eds.), Systems Biology and Regulatory Genomics. X, 259 pages. 2007.

Vol. 3939: C. Priami, L. Cardelli, S. Emmott (Eds.), Transactions on Computational Systems Biology IV. VII, 141 pages. 2006.

Vol. 3916: J. Li, Q. Yang, A.-H. Tan (Eds.), Data Mining for Biomedical Applications. VIII, 155 pages. 2006.

Vol. 3909: A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 612 pages. 2006.

Vol. 3886: E.G. Bremer, J. Hakenberg, E.-H.(S.) Han, D. Berrar, W. Dubitzky (Eds.), Knowledge Discovery in Life Science Literature. XIV, 147 pages. 2006.

Vol. 3745: J.L. Oliveira, V. Maojo, F. Martín-Sánchez, A.S. Pereira (Eds.), Biological and Medical Data Analysis. XII, 422 pages. 2005.

Vol. 3737: C. Priami, E. Merelli, P. Gonzalez, A. Omicini (Eds.), Transactions on Computational Systems Biology III. VII, 169 pages. 2005.

Vol. 3695: M.R. Berthold, R.C. Glen, K. Diederichs, O. Kohlbacher, I. Fischer (Eds.), Computational Life Sciences. XI, 277 pages. 2005.

Vol. 3692: R. Casadio, G. Myers (Eds.), Algorithms in Bioinformatics. X, 436 pages. 2005.

Vol. 3680: C. Priami, A. Zelikovsky (Eds.), Transactions on Computational Systems Biology II. IX, 153 pages. 2005.

Vol. 3678: A. McLysaght, D.H. Huson (Eds.), Comparative Genomics. VIII, 167 pages. 2005.

Vol. 3615: B. Ludäscher, L. Raschid (Eds.), Data Integration in the Life Sciences. XII, 344 pages. 2005.

Vol. 3594: J.C. Setubal, S. Verjovski-Almeida (Eds.), Advances in Bioinformatics and Computational Biology. XIV, 258 pages. 2005.

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005.

Vol. 3388: J. Lagergren (Ed.), Comparative Genomics. VII, 133 pages. 2005.

Vol. 3380: C. Priami (Ed.), Transactions on Computational Systems Biology I. IX, 111 pages. 2005.

Vol. 3370: A. Konagaya, K. Satou (Eds.), Grid Computing in Life Science. X, 188 pages. 2005.

Vol. 3318: E. Eskin, C. Workman (Eds.), Regulatory Genomics. VII, 115 pages. 2005.

Vol. 3240: I. Jonassen, J. Kim (Eds.), Algorithms in Bioinformatics. IX, 476 pages. 2004.

Vol. 3082: V. Danos, V. Schachter (Eds.), Computational Methods in Systems Biology. IX, 280 pages. 2005.

Vol. 2994: E. Rahm (Ed.), Data Integration in the Life Sciences. X, 221 pages. 2004.

Vol. 2983: S. Istrail, M.S. Waterman, A. Clark (Eds.), Computational Methods for SNPs and Haplotype Inference. IX, 153 pages. 2004.

Vol. 2812: G. Benson, R.D.M. Page (Eds.), Algorithms in Bioinformatics. X, 528 pages. 2003.

Vol. 2666: C. Guerra, S. Istrail (Eds.), Mathematical Methods for Protein Structure Analysis and Design. XI, 157 pages. 2003.

¥452.00元

Table of Contents

Knowledge Discovery and Emergent Complexity in Bioinformatics

Ronald Westra[1], Karl Tuyls[1], Yvan Saeys[2], and Ann Nowé[3]

[1] Department of Mathematics and Computer Science,
Maastricht University and Transnational University of Limburg,
Maastricht, The Netherlands
[2] Department of Plant Systems Biology, Ghent University,
Flanders Interuniversity Institute for Biotechnology (VIB),
Ghent, Belgium
[3] Computational Modeling Lab,
Vrije Universiteit Brussel,
Brussels, Belgium

1 Introduction

In February 1943, the Austrian physicist Erwin Schrödinger, one of the founding fathers of quantum mechanics, gave a series of lectures at the Trinity College in Dublin, entitled “What Is Life? The Physical Aspect of the Living Cell and Mind”. In these lectures Schrödinger stressed the fundamental differences encountered between observing animate and inanimate matter, and advanced some at the time audacious hypotheses about the nature and molecular structure of genes, some ten years before the discoveries of Watson and Crick.

Indeed, the rules of living matter, from the molecular level to the level of supraorganic flocking behavior, seem to violate the simple basic interactions found between fundamental particles as electrons and protons. It is as if the organic molecules in the cell ‘know’ that they are alive. Despite all external stochastic fluctuations and chaos, process and additive noise, this machinery is ticking for at least 3.8 billion years. Yet, we may safely assume that the laws that govern physics also steer these complex associations of synchronous and seemingly intentional dynamics in the cell. Contrary to the few simple laws that govern the interactions between the few really elementary particles, there are at least tens of thousands of different genes and proteins, with millions of possible interactions, and each of these interactions obeys its own peculiarities. There are different processes involved like transcription, translation and subsequent folding. How can we ever understand the fundamentals of these complex interactions that emerge from the few empirical observations we are able to make.

The KDECB 2006 Symposium, and this associated book, is intended to provide a forum for the presentation of new ideas and results bearing on the conception of knowledge discovery and emergent complexity in bioinformatics. The goal of this symposium is to increase awareness and interest in knowledge discovery and emergent complexity research in Bioinformatics, and encourage collaboration between Machine Learning experts, Computational Biology experts,

K. Tuyls et al. (Eds.): KDECB 2006, LNBI 4366, pp. 1–9, 2007.

Mathematicians and Physicists, and give a representative overview of the current state of affairs in this area. Next to a strong program with lectures of leading scientists in this multi-disciplinary field, we present contributions that cover on how knowledge can be extracted from, and complexity emerges in sophisticated biological systems. Different disciplines, both 'wet' and 'dry', have contributed to these developments and they will also benefit directly or indirectly from new, intelligent, computational techniques.

In the remainder of this document the three main themes of this book are introduced and discussed, namely, (i) Machine Learning for Bioinformatics, (ii) Mathematical modeling of gene-protein networks, and, (iii) Nature-inspired computation.

2 Machine Learning for Bioinformatics

During the past decades, advances in genomics have generated a wealth of biological data, increasing the discrepancy between what is observed and what is actually known about life's organisation at the molecular level. To gain a deeper understanding of the processes underlying the observed data, pattern recognition techniques play an essential role.

The notion of a *pattern* however, needs to be interpreted in a very broad sense. Essentially, we could define a pattern as everything that is the opposite of chaos. Thus the notion of *organisation* can be associated with a pattern. The goal of pattern recognition techniques then is to elucidate the organisation of the pattern, resulting in a wide range of subtasks such as recognition, description, classification, and grouping of patterns.

In bioinformatics, techniques to learn the theory *automatically* from the data (machine learning techniques) play a crucial role, as they are a first step towards interpreting the large amounts of data, and extracting useful biological knowledge from it. Machine learning techniques are generally applied for the following problems: classification, clustering, construction of probabilistic graphical models, and optimisation.

In classification (sometines also referred to as supervised learning) the goal is to divide objects into classes, based on the characteristics of the objects. The rule that is used to assign an object to a particular class is termed the classification function, classification model, or classifier. Many problems in bioinformatics can be cast into a classification problem, and well established methods can then be used to solve the task. Examples include the prediction of gene structures [4,26,37], which often is the first step towards a more detailed analysis of the organism, the classification of microarray data [17,21], and recently also classification problems related to text mining in biomedical literature [23]. The computational gene prediction problem is the problem of the automatic annotation of the location, structure, and functional class of protein-coding genes. A correct annotation forms the basis of many subsequent modeling steps, and thus should be done with great care. Driven by the explosion of genome data, computational approaches to identify genes have thus proliferated, thereby depending strongly on machine learning techniques.

A second class of problems in bioinformatics concerns the topic of clustering, also termed unsupervised learning, because no class information is known a priori. The goal of clustering is to find natural groups of objects (clusters) in the data that is being modeled, where objects in one cluster should be similar to each other, while being at the same time different from the objects in another cluster. The most common examples of clustering in bioinformatics concern the clustering of microarray expression data [10,19,39], and the grouping of sequences, e.g. to build phylogenetic trees [13].

Probabilistic graphical models [31] have proliferated as a useful set of techniques for a wide range of problems where dependencies between variables (objects) need to be modeled. Formally, they represent multivariate joint probability densities via a product of terms, each of which only involves a few variables. The structure of the problem is then modeled using a graph that represents the relations between the variables, which allows to reason about the properties entailed by the product. Common applications include the inference of genetic networks in systems biology [38] and Bayesian methods for constructing phylogenetic trees [34]. Other examples of applications of machine learning techniques in bioinformatics include the prediction of protein structure (which can be cast into an optimisation problem), motif identification in sequences, and the combination of different sources of evidence for analysis of global properties of bio(chemical) networks. In all of these domains, machine learning techniques have proven their value, and new methods are constantly being developed [25].

3 Modeling the Interactions Between Genes and Proteins

A prerequisite for the successful reconstruction of gene-protein networks is the way in which the dynamics of their interactions is modeled. The formal mathematical modeling of these interactions is an emerging field where an array of approaches are being attempted, all with their own problems and short-comings. The underlying physical and chemical processes involved are multifarious and hugely complex. This condition contrasts sharply with the modeling of inanimate Nature by physics. While in physics huge quantities of only a small amount of basic types of elementary particles interact in a uniform and deterministic way provided by the fundamental laws of Nature, the situation in gene-protein interactions deals with tens of thousands of genes and possibly some million proteins. The quantities thereby involved in the actual interactions are normally very small, as one single protein may be able to (in)activate a specific gene, and thereby change the global state of the system. For this reason, gene regulatory systems are much more prone to stochastic fluctuations than the interactions involved in normal inorganic reactions. Moreover, each of these interactions is different and involves its own peculiar geometrical and electrostatic details. There are different processes involved like transcription, translation and subsequent folding. Therefore, the emergent complexity resulting from gene regulatory networks is much more difficult to comprehend.

In the past few decades a number of different formalisms for modeling the interactions amongst genes and proteins have been presented. Some authors focus on specific detailed processes such as the circadian rhythms in *Drosophila* and *Neurospora* [16,18], or the cell cycle in *Schizosaccharomyces* (Fission yeast) [30]. Others try to provide a general platform for modeling the interactions between genes and proteins. For a thorough overview consult de Jong (2002) in [6], Bower (2001) in [3], and others [12,14,20].

Traditionally, much emphasis lay on static models, where the relations between genes and proteins are considered fixed in time. This was in line with the impressive developments in microarray technology that opened a window towards reconstructing static genetic and metabolic pathways, as for instance demonstrated in [36]. Successful static models are the Logical Boolean networks consult [2,3,5,1], and on Bayesian Networks consult [14,40,41]. In discrete event simulation models the detailed biochemical interactions are studied. Considering a large number of constituents, the approach aims to derive macroscopic quantities. More information on discrete event modeling can be found in[3].

In contrast to the static networks, the aim in modeling dynamic networks is to explain the macroscopic network complexity from the molecular dynamics and reaction kinetics. The approach to modeling the dynamical interactions amongst genes and proteins is by considering them as biochemical reactions, and thus representing them as traditional 'rate equations'. The concept of chemical rate equations, dating back to Van 't Hoff, consists of a set of differential equations, expressing the time derivative of the concentration of each constituent of the reaction as some rational function of the concentrations of all the constituents involved. In general, the syntax of the chemical reactions is mapped on the syntax of the rate equations, as e.g. in the Michaelis-Menten equation for enzyme kinetics. More on the physical basis of rate equations can be found in [48].

Though the truth of the underlying biochemical interactions between the constituents is generally accepted, the idea of representing them by rate equations involves a number of major problems. First of all, the rate equation is not a fundamental law of Nature like the great conservation laws of Energy and Momentum, but a statistical average over the entire ensemble of possible microscopic interactions. The applicability of the rate equation therefore relates to the law of large numbers. In normal inorganic reactions this requirement holds. However, in inhomogeneous mixtures or in fast reactions the actual dynamics will depart significantly from this average. Also in case of gene-, RNA-, and protein-interactions this condition will not hold as we will discuss later. Second, the Maxwell velocity distribution should apply, otherwise the collision frequency between the constituents would not be proportional to their concentrations, and details of the velocity distribution would enter. This condition is met easily in the presence of a solvent or an inert gas, but difficult to attain for macromolecules in a cytoplasm. The same holds for the distribution of the internal degrees of freedom of the constituents involved, such as rotational and vibrational energies. The distribution of their energies should have the same 'temperature' as in the Maxwell velocity distribution, otherwise this would affect the rate of the collisions that result in an actual chemical reaction. Also this condition is not easily met by gene-protein interactions. Finally, the temperature of the reaction should be constant in space and time - this condition may be accounted for in this context.

So, rate equations are statistical approximations that hold under above requirements. Under these conditions they predict the average number of reactive collisions. The actual observed number will fluctuate around this number, depending on the details of the microscopic processes involved. In case of biochemical interactions between genes and proteins at least some of the conditions will be violated and therefore the applicability of the concept of rate equations is valid only for genes with sufficient high transcription rates. This is confirmed by recent experimental findings by Swain and Elowitz [11], [35], [42], [43].

Dynamic gene-protein networks can lead to mathematical complexities in modeling and identification [27,28,8]. To overcome these problems, some authors have proposed to model them as piecewise linear models, as introduced by Glass and Kauffman [15]. Such models can be demonstrated to be memory-bounded Turing-machines [2]. de Jong *et al.* [6,7] have proposed qualitative piecewise linear models rather than a quantitative models, because the correct underlying multifarious mathematical expressions are not tractable. In spite of the intuitive attractiveness of this idea, there are a number of conceptual and practical problems in applying these techniques in practical situations. In biology piecewise linear behaviour is frequently observed, as for instance in embryonic growth where the organism develops by transitions through a number of well-defined 'check points'. Within each such checkpoint the system is in relative equilibrium. However, it should be mentioned that there is an ongoing debate on the modeling of gene-protein dynamics as *checkpoint mechanisms* versus *limit-cycle oscillators*, see [33,44].

Others have employed specific characteristics of the networks to construct a viable reconstruction algorithm, such as the sparsity and hierarchy in the network interactions [8,49,32].

4 Nature-Inspired Computing

In the sections above, we gave an overview of approaches and techniques from computer science and mathematics that are promising in order to model biological phenomena such as gene networks, protein structure, etc. We can however go one step further, and try to model the emergent collective intelligence, arising in nature from local, simple interactions between simple units, which can be biological cells, neurons as well as insects as ants and bees. Using insights from how this complexity and global intelligence emerges in nature, we can develop new computational algorithms to solve hard problems. Well known examples are Neural Networks and Genetic Algorithms. Whereas Neural Networks are inspired on the working of the brain, Genetic Algorithms are based on the model of natural evolution. Another natured inspired technique is reinforcement learning. Reinforcement learning [22,45] finds its roots in animal learning. It is well known that, by operand or instrumental conditioning, we can teach an animal to respond in some desired way. The learning is done by rewarding and punishing the learner appropriately, and as a result the likelihood of the desired behaviour is increased during the learning process, whereas undesired behaviour will become less likely.

The objective of a reinforcement learner is to discover a policy, meaning a mapping from situations to actions, so as to maximise the reinforcement it receives. The reinforcement is a scalar value which is usually negative to express a punishment, and positive to indicate a reward. Unlike supervised learning techniques, reinforcement learning methods do not assume the presence of a teacher who is able to judge the action taken in a particular situation. Instead the learner finds out what the best actions are by trying them out and by evaluating the consequences of the actions by itself. For many problems, such as planning problems, the consequences of the action are not immediately apparent after performing the action, but only after a number of other actions have been taken. In other words the selected action may not only affect the immediate reward/punishment the learner receives, but also the reinforcement it might get in subsequent situations, i.e. the delayed rewards or punishments. Reinforcement learning techniques such as Q-learning and Adaptive Critique techniques, can deal with this credit assignment problem and are guaranteed to converge to an optimal policy, as long as some conditions, such as the environment experienced by the learner should be Markovian and the learner should be allowed sufficient exploration, are met.

More recently other nature inspired techniques such as Ant Colony Optimisation (ACO) [9] received a lot of attention. ACO techniques are inspired by the behaviour of ants. It is well known that one single ant on its own cannot do anything useful, but a colony of ants is capable of performing complex behaviour. The complex behaviour emerges due to the fact that ants can communicate indirectly with each other, by laying a pheromone trail in the environment. This pheromone signal can be observed by other ants, and this will influence their own behaviour. The more pheromone is sensed by an ant in some direction, the more it will be attracted in that direction, and the more the pheromone will be reinforced. ACO algorithms have been successfully applied to complex graph problems such as large instances of the travelling salesman problem. ACO techniques are closely related to the Reinforcement Learning technique mentioned in the previous paragraph, however they do not come with straightforward convergence proofs. As is illustrated in [46] by Verbeeck *et al.* it is possible to provide a clean proof of convergence by expressing the mapping between the ACO pheromone updating mechanism and interconnecting learning automata [29]. The insight into the convergence issues of these algorithms is crucial in order to have a wider acceptance of these techniques.

Recent investigations [24,47] have also opened up the possibility of applying recruitment and navigational techniques from honeybees to computational problems as for instance foraging. Honeybees use a strategy named Path Integration. By employing this strategy, bees always know a direct path towards their destination and their home. Bees employ a direct recruitment strategy by dancing in the nest. Their dance communicates distance and direction towards a destination. Ants, on the other hand, employ an indirect recruitment strategy by accumulating pheromone trails. When a trail is strong enough, other ants are attracted to it and will follow this trail towards a destination. Both strategies provide the insects with an efficient way of foraging.

References

1. Arkin A., Ross J., McAdams H.H. (1994), Computational functions in biochemical reaction networks. *Biophys. Journal*, Vol. 67, pp. 560–578.
2. Ben-Hur A., Siegelmann H.T. (2004), Computation in Gene Networks. *Chaos*, Vol. 14(1) pp. 145–151.
3. Bower J.M., Bolouri H.(Editors) (2001), Computational Modeling of Genetic and Biochemical Networks. *MIT Press*, 2001.
4. Burge, C., Karlin, S. (1997), Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*, Vol. 268, pp. 78–94.
5. Davidson E.H. (1999), A View from the Genome: Spatial Control of Transcription in Sea Urchin Development, *Current Opinions in Genetics and Development*, Vol. 9, pp. 530–541.
6. de Jong H. (2002), Modeling and Simulation of Genetic RegulatorySystems: A Literature Review. *Journal of Computational Biology*, Vol. 9(1), pp. 67–103.
7. de Jong H., Gouze J.L., Hernandez C., Page M., Sari T., Geiselmann J. (2004), Qualitative simulation of genetic regulatory networks using piecewise-linear models. *B*ull Math Biol., Vol. 66(2), pp. 301–40.
8. D'haeseleer P., Liang S., Somogyi R. (2000), Genetic Network Inference: From Co-Expression Clustering to Reverse Engineering. *B*ioinformatics, Vol. 16(8), pp. 707–726.
9. Dorigo M. and Sttzle T. (2004), Ant Colony Optimization. MIT Press. (2004).
10. Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. (1998), Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, Vol. 95(25), pp 14863-14868
11. Elowitz M.B., Levine A.J., Siggia E.D., Swain P.S. (2002), Stochastic gene expression in a single cell. *Science*, Vol. 297, pp. 1183–1186.
12. Endy, D, Brent, R. (2001), Modeling Cellular Behavior. *Nature*, Vol. 409(6818), pp. 391–395.
13. Felsenstein, J. (2004), Inferring Phylogenies. Sinauer Associates, Sunderland, Mass.
14. Friedman, N., Linial, M., Nachman, I., Pe'er, D. (2000), Using Bayesian Networks to analyze expression data. *Journal of Computational Biology*, Vol. 7, pp. 601–620.
15. Glass L., Kauffman S.A. (1973), The Logical Analysis of Continuous Non-linear Biochemical Control Networks, *J.Theor.Biol.*, Vol. 39(1), pp. 103–129
16. Goldbeter A (2002), Computational approaches to cellular rhythms. *Nature*, Vol. 420, pp. 238–45.
17. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M. (1999), Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, Vol. 286, pp. 531–537.

18. Gonze D, Halloy J, and Goldbeter A (2004), Stochastic models for circadian oscillations : Emergence of a biological rhythm. *Int J Quantum Chem*, Vol. 98, pp. 228–238.
19. Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., Brown, P. (2000), Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, Vol. 1(2),research0003.10003.21.
20. Hasty J., McMillen D., Isaacs F., Collins J. J., (2001), Computational studies of gene regulatory networks: in numero molecular biology. *Nature Reviews Genetics*, Vol. 2(4), pp. 268– 279.
21. Inza, I., Larrañaga, P., Blanco, R., Cerrolaza, A.J. (2004), Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine*, special issue in "Data mining in genomics and proteomics", Vol.31(2), pp 91–103.
22. Kaelbling L.P., Littman L.M. and Moore A.W. (1996), Reinforcement learning: a survey, Journal of Artificial Intelligence Research, 4 (1996) 237-285.
23. Krallinger, M., Valencia, A. (2005), Text-mining and information-retrieval services for molecular biology. *Genome Biol.*, Vol. 6(7), 224
24. Lambrinos, D., Moller, R., Labhart, T., Pfeifer, R., and Wehner, R. (2000). A mobile robot employing insect strategies for navigation. Robotics and Autonomous Systems, Vol. 30, Nos. 12, pp. 3964.
25. Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armañanzas, R., Santafé, R., Pérez, A., Robles, V. (2006), Machine Learning in Bioinformatics. *Briefings in Bioinformatics*, Vol.7(1), pp. 86–112.
26. Mathé, C., Sagot, M.F., Schiex, T. and Rouzé, P. (2002), Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, Vol. 30(19), pp. 4103–17.
27. Mestl T., Plahte E., Omholt S.W. (1995*a*), A Mathematical Framework for describing and analysing and Analyzing Gene Regulatory Networks. *J. Theor. Biol.*, Vol. 176(2), pp. 291–300.
28. Mestl T., Plahte E., Omholt S.W. (1995*b*), Periodic Solutions in Systems of Piecewise-Linear Systems. *Synamic Stability of Systems*, Vol. 10(2), pp. 179–193.
29. Narendra K. and Thathachar M., Learning Automata: An Introduction, Prentice-Hall International, Inc, (1989).
30. Novak B, Tyson JJ (1997), Modeling the control of DNA replication in fission yeast. *Proc. Natl. Acad. Sci. USA*, Vol. 94, pp. 9147–9152.
31. Pearl, J. (1988), Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers, 1988.
32. Peeters R.L.M., Westra R.L. (2004), On the identification of sparse gene regulatory networks, *Proc. of the 16th Intern. Symp. on Mathematical Theory of Networks and Systems* (MTNS2004) Leuven, Belgium July 5-9, 2004
33. Rao, C.V., Wolf, D.M., Arkin, A.P. (2002), Control, exploitation and tolerance of intracellular noise. *Nature*, Vol. 420, pp. 231–237.
34. Ronquist, F., J. P. Huelsenbeck (2003), MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, Vol. 19, pp. 1572–1574.
35. Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S., Elowitz, M.B., Gene regulation at the single-cell level. *Science*, Vol. 307, pp. 1962.
36. Rustici, G., Mata, J., Kivinen, K., Lio, P., Penkett, C.J., Burns, G., Hayles, J., Brazma, A., Nurse, P., Bahler, J. (2004), Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics*, Vol. 36(8), pp. 809–17.