Symposium on Scientific Research at the Technical University of Lisbon
Lisbon, Portugal, 2 and 3 of February

Bioinformatics: A New Approach for the Challenges of Molecular Biology

Arlindo L. Oliveira[1], Ana T. Freitas[1], and Isabel Sá-Correia[2]

[1]IST/INESC-ID
R. Alves Redol, 1000 Lisboa
{aml,atf}@inesc-id.pt

[2] IST/CEBQ
Av. Rovisco Pais, 1000 Lisboa
isacorreia@ist.utl.pt

Keywords: Bioinformatics, Molecular Biology, Biotechnology, Regulatory Networks.

Abstract. *We describe the research being undertaken by the ALGOS/KDBIO and Biological Sciences groups of Instituto Superior Técnico on the field of bioinformatics and computational biology, with emphasis on the efforts under way to develop new approaches, methods and algorithms for the determination of gene regulatory networks. We put the field in perspective by first looking at recent developments in the field of bioinformatics, and how these developments contributed to the advance of science. We then describe the approach that is being followed, based on the development of algorithms and information systems for the problems of motif detection, gene expression analysis and inference of gene regulatory networks. We conclude by pointing out possible directions for future research in the fields of systems biology and synthetic biology, two critical areas for the development of science in the coming years.*

1 INTRODUCTION

The analysis of biological systems using computational tools has emerged as an autonomous area that is central to the advancement of science. The ability to sequence efficiently the genomes of many organisms, including *Homo sapiens*, can be seen as the first significant impact of bioinformatics in the study of complex biological systems. However, this ability represents only the first step in a long journey that will, in the end, contribute to the development of algorithms, methods and systems that will be able to simulate, *in-silico*, complex biological systems.

The techniques required to achieve this ambitious goal will come from the confluence of a number of fields of knowledge. Bioinformatics has emerged as a field that integrates contributions from many different fields to analyze efficiently the large volumes of data made available by modern technologies for sequencing and global expression analysis.

This article starts, in section 2, by putting in perspective the recent history of Bioinformatics, giving a necessarily brief overview of the techniques developed in the past decades, and the most important results obtained.

It then goes on to describe, in section 3, the approaches that are currently being taken by two research groups at IST in a number of important areas, which include the development of biological information systems, methods for the analysis of promoter regions of genes and algorithms for the treatment of results obtained from global expression analysis. These approaches have one major goal in common, that of modeling in a precise and effective way the dynamics of gene and metabolic networks that govern the life cycles of biological systems.

Finally, section 4 lists a number of promising areas for future research and provides a number of educated guesses on the possible impacts of this areas of research in the future of biotechnology and the natural sciences.

2 COMPUTERS, SEQUENCES AND GENOMES

When, in 1953, Watson and Crick [1] identified the double helix of the DNA as the repository for genetic information, stored in digital format in a molecule that, until then, was relatively uninteresting, it did not become immediately obvious that computers would play such a fundamental role in understanding biological systems. At the time, computers were so rare and unknown that the parallel between the encoding of genetic information in DNA and the stored programs of digital computers did not become immediately obvious.

Roughly 25 years after the discovery of the DNA, advances in technology lead to what can be considered the first convergence between biology and computation. Soon after the first sequencing of a fragment of DNA, the first algorithms that aimed at reconstructing DNA sequences from fragments were developed [2].

Far from being consensual, the projects of sequencing whole genomes have met a number of objections [3]. These objections were centered not only on the technical difficulties perceived in such an undertaking, but also on the perception that this information would not be specially useful, given the challenges that would face a biologist interested in exploring it in an effective way.

The first realistic proposal for the sequencing of the human genome appeared in 1985. However, only in 1988 did the Human Genome Project [4] make clear that such an ambitious objective was within reach. The selection of model organisms with increasing degrees of complexity enabled not only the progressive development of the technologies necessary for the sequencing of the human genome, but also the progressive development of research in genomics of the communities involved in the study of those model organisms.

A number of model organisms was selected to be sequenced, including, *Haemophilus influenza* (1.8Mb), *Escherichia coli* (5Mb), *Saccharomyces cerevisiae* (12Mb), *Caenorhabditis elegans* (100Mb), *Arabidopsis thaliana* (125Mb), *Drosophila melanogaster* (200Mb).

With the development of better sequencing techniques, the genome sequences of these organisms were successively obtained; *H. influenza* in 1995 [5], *S. cerevisiae* in 1996 [6], *E. coli* in 1997 [7], *C. elegans* in 1998 [8] and *A. thaliana* and *D. melanogaster* in 2000 [9] [10].

The final steps of the process that lead to the simultaneous publication of the human genome papers in Science [11] and Nature [12], using two different (but not totally independent) approaches are well known. The sequence of the human genome brought the additional somewhat surprising information that our genome is not larger (3000Mb) nor more complex than that of the mouse. The approach being taken for human genomic sequencing was originally the same as the one used for the *S. cerevisiae* and *C. elegans* genomes, based on the serial sequencing of overlapping arrays of large DNA inserts in *E. coli* clones. The alternative approach, taken by Celera, was to apply shotgun sequencing to the whole human genome, an approach made possible only by advances in algorithms and bioinformatics [13].

At the end of 2005, there are 321 completely sequenced genomes, of which 281 are bacterial and 40 are eukaryotes. At the time of this writing, more than 1350 new sequencing projects are currently under way to obtain the genomes of more than 800 prokaryotes and 550 eukaryotes.

Having the sequences available, however, is only the beginning of the long path that will lead to the full understanding of biological systems. Researchers in bioinformatics have been concerned with the development of algorithms for the analysis, manipulation and annotation of biological sequences since the connection between biology and computation became obvious. In the last 20 years, a number of important algorithms for sequence analysis were developed, in order to cope with the increasingly large amounts of data available. Amongst these, alignment methods, both exact [14][15] and approximate [16][17] have become an indispensable tool for scientists, and are now part of the toolbox of every molecular biologist.

Gene finding in sequences, using either automatic or semi-automatic procedures, is another field where algorithms have come to the rescue of biologists. Accurate identification of genes is dependent on the accurate identification of signals, namely on the accurate identification of splice sites and of start and stop codons. A large number of methods has been proposed for this problems, and they can be classified into two broad classes: homology based methods and ab initio methods.

Homology based methods use the fact that conservation of DNA sequences leads to strong similarities in the coding sequences of related genes, a fact that can be used to predict coding regions in unlabeled DNA sequences. They use sequence alignment methods like BLAST [16] or FASTA [17] to identify highly conserved regions in the DNA that provide evidence of the existence of signals and/or coding regions. These methods are limited in their ability to discover new genes encoding proteins that are not sufficiently similar to those encoded by already annotated DNA sequences.

Ab initio based methods use pattern recognition algorithms to learn, from labeled sequences, the rules that can be used to recognize the relevant signals. Among the pattern recognition algorithms that have been used, neural networks, decision trees and hidden Markov models have been extensively applied.

In practice, effective gene finding methods combine both approaches, and are usually difficult to classify squarely as either homology based or ab initio. A number of well known gene finding programs have been made available, amongst which GRAIL [18], MZEF [19], GENSCAN [20] and GENEID [21] deserve to be mentioned. Performance of the different methods varies with a number of factors [22] but, when coupled with human based annotation,

these methods are adequate to perform the identification of the coding regions in the genomes, and do not represent the limiting factor for the advancement of biological knowledge.

The first years of the XXI century have, therefore, led to the sequencing, mapping and annotation of the genes that control the development of all life forms. Using all this information to understand biological phenomena appears now as the major challenge of this coming century.

3 UNDERSTANDING REGULATORY NETWORKS

As the knowledge about biology advanced, pushed by advances in sequencing and annotation, it became clear that the next steps in the understanding of biological systems depend critically on the identification of the complex regulatory networks that are present in every living organism. In fact, knowing where the genes are, and the proteins they code for, is of little help if one does not know the relations between genes, proteins and the other chemical compounds that are present in every cell. Understanding biology from a systemic standpoint came to be known as Systems Biology, and is the main challenge of the next decades in the natural sciences.

Information needed for this task comes from two main sources: genomic sequence data and whole-genome measures of gene expression obtained using microarrays and quantitative proteomics. Using sequence and gene expression data together with phylogenetic information to infer network structures has emerged as the only realistic avenue that can be pursued in order to address the challenge of identifying, mapping and documenting the complex architectures of gene regulatory networks of a living organism, the central goal of systems biology. To achieve this goal, we need to be successful in, at least, four major areas.

In the first place, we need to successfully integrate, curate and make widely available existing knowledge about regulatory mechanisms in different organisms. Model organisms, such as *S. cerevisiae*, represent an important intermediate step, because they can more easily be used in experiments than more complex organisms. To support the experimental procedures needed to validate mathematical models, we need to develop information systems that make available, in an integrated and uniform way, the vast amounts of knowledge that already exist about specific gene regulatory mechanisms.

In the second place, we need to develop algorithms, methods, models and systems that can guide researchers in their search for still unknown regulatory mechanisms. This will help us completing our understanding of the rules that govern gene regulation and the interaction between the many other components of living organisms.

In the third place, we need to develop effective methods for the analysis of gene expression data, in order to help researchers in their quest for understanding the massive amounts of data generated by modern technologies such as microarrays.

Finally, we will need to integrate all this information into a coherent model of biological networks that can be used to model, simulate and predict responses of organisms to specific conditions. Although the complete achievement of such a goal is still far off in the future, the development of these enabling tools stands in the critical path, and has emerged as a central goal in systems biology.

The next sections describe some of the efforts currently being undertaken at IST in each of these areas, while also providing some broader perspective on the important issues at stake.

3.1 Databases for gene regulatory information

The availability of the complete sequence of a number of organisms implied a significant change in biology and biotechnology in the last decade. Making these sequences available to the scientific community has emerged as an important objective in itself. Creating and making widely available biological databases with sequence and annotation information represents now a very significant part of the activity of the scientific community. More than 850 biological databases are listed in one single source of information [23], and many more are available and listed in other places.

No single group can contribute with a significant fraction of the total information that is made available and every research community focuses in specific sub-fields. The IST groups have focuses on documenting, organizing and making publicly available information about gene regulation mechanisms in Yeast, complementing the many available databases available for this organism.

Since the release of the complete genome sequence of *S. cerevisiae*, a number of computational methods and tools have become available to support research related with this organism. Most significant for the Yeast community, the *Saccharomyces cerevisiae* database (SGD) [24], and other databases specialized in Yeast, like CYGD, the Comprehensive Yeast Genome Database [25] or YRC, the Yeast Resource Center [26], make available extensive information on molecular biology and genetics of *S. cerevisiae*.

The precise coordinated control of gene expression is accomplished by the interplay of multiple regulatory mechanisms. The transcriptional machinery is recruited to the promoter leading to the transcription of the downstream gene through the binding of transcription regulatory proteins to short nucleotide sequences occurring in gene promoter regions. To support the analysis of the promoter sequences in the yeast genome, a set of software tools is provided by RSAT (Regulatory Sequences Analysis Tools [27]). RSAT makes available pattern matching methods, supporting the search for given nucleotide sequences (e. g. transcription factor binding sites) within the promoter region of chosen genes, thus leading to the identification of putative target genes for specific transcription factors. However, the transcription factor binding sites, used for pattern matching in RSAT, have to be provided by the user, since RSAT does not hold a database of transcription factor binding sites. Existing databases do not fill this gap. The IST groups involved in this research have therefore developed a database of known regulatory associations between genes in this organism, YEASTRACT [28].

YEASTRACT (YEAst Search for Transcriptional Regulators And Consensus Tracking; www.yeastract.com) database is a repository of more than 12500 regulatory associations between genes and transcription factors in *S. cerevisiae* and includes the description of 269 specific DNA binding sites for 106 characterized transcription factors. This publicly available database was developed to provide assistance in three major issues: identification of documented and potential regulatory associations for an ORF/Gene; microarray data clustering based on regulatory associations; search for a DNA motif within known transcription factor binding sites and promoter regions. In the first three months of 2006, more than 150 different groups from 36 different countries have performed over 80000 queries using YEASTRACT.

Currently, new algorithms for the analysis of biologically significant over-represented motifs in promoter regions are under development. When available, these tools will be integrated with YEASTRACT and will simplify the analysis of the complex regulatory mechanisms underlying transcriptional response in Yeast.

3.2 Detection of cis-regulatory modules in promoter regions

Algorithms for identifying relevant motifs in promoter regions have been known for a number of years. Two main types of models have appeared in the literature: pattern and weight matrix. Each of these models was developed together with corresponding methods for inferring regulatory signals in DNA sequences. More recently, the sequencing of different genomes from closely related species has enabled the use of evolutionary information through genome comparative approaches to improve the identification process.

The first methods for detecting promoter regions in DNA sequences [29][30] looked for a binding site composed of adjacent nucleotides. In the search for more complex cis-regulatory models methods have appeared that extract DNA sites composed by non-contiguous nucleotides.

Existing methods obtain comparable results. However, although these results have appeared promising for a long time, the algorithms have so far fallen short of becoming a truly systematic and automatic method for identifying signals, particularly at the scale of whole genomes.

There are many difficult issues related with this problem, including the ones around the computational hardness of most variants of the problem. One major difficulty however is that it has not been able to arrive at any accurate method for estimating, either probabilistically or combinatorially, the parameters that should be passed to the algorithms, that is, the characteristics, even roughly defined, of the motifs to be identified. This problem is often alleviated by using available biological knowledge to do a pre-processing of the data, thus reducing it to sizes and levels of noise that are more manageable by current methods. Even in such cases, important information may be missed leading to the repeated identification of what is already reasonably well known. More importantly, biological knowledge, even partial, is not always available, in particular at the scale of the full regulatory system of an organism.

Research developed at IST has attacked this problem from a number of different perspectives.

One approach is based on the development of more efficient algorithms for the discovery of complex motifs [31][32]. These methods use sophisticated string processing techniques to process large volumes of sequence data, and efficiently identify over-represented motifs in promoter regions. Current research is focused on applying these techniques to the derivation of more general models, and on using additional biological information related with the structure of the DNA to improve the quality of the results obtained.

We have explored the use of massively parallel computation to achieve significant speedups in this computationally difficult problem [33]. The application of the Grid based computing paradigm enlarges the range of problems that can be tackled by the algorithms, using available and unused CPU time.

Current research is focused on the development of methods that can be used to accurately identify the parameters that should be passed to motif finding algorithms, relieving biologists from this difficult task. Preliminary results obtained using this approach, coupled with a more accurate assessment of the statistical significance of the motifs, have already led to the discovery of previously unknown biological knowledge.

3.3 Analysis of gene expression data

Additional information about regulatory networks comes from gene expression data. The experimental analysis of whole, complex genetic networks was revolutionized about ten years ago by the development of DNA microarrays [34]. This technology enables genome-wide analysis of gene expression and is now widely used to obtain data about the interactions between genes.

Since then, microarrays have given thousands of snapshots of gene expression in many organisms, including *S. cerevisiae*, using a broad panel of experimental conditions [35]. The coupling of genetic engineering and microarrays has also been extensively used to identify the target genes and physiological impact of many yeast transcription factors. More than 100 million of individual expression data are now available just for *S. cerevisiae*.

These large amounts of data created the need for novel computational tools that perform gene expression data analysis. Since microarrays can provide a snapshot of the expression level of all the genes in a cell at a given time, and since it has been demonstrated that gene expression is a fundamental link between genotype and phenotype, the analysis of gene expression data is bound to play a major role in our understanding of biological processes and systems including gene regulation, development, evolution and disease mechanisms. Other sources of gene expression data like quantitative proteomics [36] also provide important information that should be processed using adequate computational tools.

Clustering of genes and/or conditions has been used, with some success, to pursue the objectives of understanding regulatory mechanisms, starting from gene expression data. However, one should expect subsets of genes to be co-regulated and co-expressed under certain experimental conditions, but to behave almost independently under other conditions. Discovering such local expression patterns may be the key to uncovering many genetic regulatory pathways that are not apparent otherwise. It is therefore important to move beyond the standard clustering paradigm, and to develop approaches capable of discovering local patterns in microarray data and identifying regulatory mechanisms that adequately model the observed patterns.

In this context, biclustering algorithms [37][38] represent a powerful mechanism for gene expression data analysis and, therefore, for the identification of co-regulated genes and potential gene regulatory networks. Unlike clustering algorithms, biclustering algorithms identify groups of genes that show similar activity patterns under a specific subset of the experimental conditions. Biclustering is thus particularly relevant when only a subset of the genes participates in a cellular process of interest, when an interesting cellular process is active only in a subset of the conditions or when a single gene may participate in multiple pathways that may or not be co-active under all conditions.

Data from time-series is specially interesting for researchers interested in identifying regulatory networks, since it gives information not only about the connections between genes, but also about the dynamics of such connections. Algorithms developed specifically for the analysis of time-series gene expression data have been developed by our research group [39], and will be used in an integrated system for the analysis of sequence and expression information that will aim at identifying the complex gene regulatory mechanisms present in higher organisms.

3.4 Modeling gene regulatory networks

Once high quality sequence and expression data is available, accurate models for the transcription mechanisms and the promoter regions have been identified, and appropriate techniques for the analysis of expression data are in place, the next challenge faced by researchers is the assembly of this information into a coherent representation of a biological interaction network.

Biological networks are abstract representations of functional interactions, and are usually modeled using a graph-based formalism. In the case of genetic networks, the approaches more commonly proposed use nodes to model genes, and links to model regulation between genes. Each node represents a single variable which can have multiple values, from a discrete or continuous range.

Many methods have been proposed for the identification of genetic networks, using a number of heuristics and biases to guide and limit the search, ranging from connectivity considerations [40] to differential analysis using mutant strands [41]. A number of such partial networks have already been identified for Yeast, but, to date, this analysis is partial and has been done mostly by hand since we are still lacking the tools to perform an automatic, high-confidence identification of such mechanisms. A complete analysis of detailed interactions between transcription factors and genes in Yeast has been performed [42] and has been instrumental in the understanding of the regulatory mechanisms in this organism. However, it has not contributed significantly to the advance of our models for these regulatory mechanisms, since the methods used do not give additional information about the reasons why a given transcription factor regulates a particular gene.

Most current methods proposed to date (e.g., [43][44]) work with single sources of information, e.g., with transcriptional control deduced from large-scale microarray experiments. However, integrating evidence obtained from as many different sources as available, coming from sequence and expression analysis, as well as from other data such as proteomics and general metabolism, in such a way as to identify the characteristics, structure and evolution of genetic regulation networks is the key idea that will support the next quantum step in the understanding of living organisms. Such integration may also enable to considerably reduce the amount of noise resulting from erroneous or incomplete data.

Approaches proposed to date fall short in their ability to seamlessly, efficiently and accurately integrate evidence coming from many different and heterogeneous sources. Indeed, the integration of information coming from just gene expression and metabolism for the purpose of identifying genetic networks is currently been addressed by only a few research groups in the world. Even at this limited level, integration poses formidable problems since determining if and when gene expression correlates with metabolic flux, via translation, enzyme regulation etc., is already not a trivial question.

A high-confidence, automatic identification of whole networks is inaccessible with current methods due to the computational complexity of the problem. One possible route to overcome this barrier is to arrive at a better understanding of the structure of biological networks, in particular its potentially modular characteristic.

Most existing methods that perform some integration, for instance of sequence with expression data, or of expression data with metabolic topology information, do not consider all available information together, or assume analysis of part of the data has already been performed thereby splitting the inference process into different, independent steps. We are currently aiming at developing methods that will integrate seamlessly information from several different sources.

A number of important gene regulatory networks in *S. cerevisiae* are being used as test platforms by the teams involved in this project. For its theoretical and practical importance, the regulatory networks related with the activation of gene *FLR1* in the presence of benomyl and mancozeb fungicides [45] have been chosen as one experimental platform and are being used to validate the hypotheses obtained using computational methods.

4 FUTURE WORK

The potential impact of this line of research in science and society is so large that it is hard to estimate accurately at this point. Even partial success, leading to methods that can infer, with high confidence, regulatory networks, would greatly improve our knowledge of biological systems and open the door to advances on such diverse fields as cancer treatment, drug design, disease control and food technology, to name only a few application areas.

It is now commonly accepted that the complexity of living beings comes, to a large extent, from the dynamic of biological networks, in general, and of genetic networks, in particular. A clear understanding of the way the building blocks of genetic networks are reused by nature has the potential to uncover many complex issues that are presently stopping us from fully understanding biological systems. The potential advantages obtained by a significant advance in this area far outweigh the risks inherent to the development of radically new techniques and models.

Advances in our abilities to model, analyze and simulate genetic networks will also lead to breakthroughs in an emerging field that will become extremely important in the next decade, that of synthetic biology.

Using computer models of cells to design and fabricate biological components and systems that do not already exist in the natural world is the aim of this new discipline. Synthetic biology takes off where systems biology ends, the study of complex biological systems as integrated wholes, using modeling and simulation tools. The aim is to build artificial biological systems using many tools and experimental techniques developed for systems biology. The focus will shift to finding new ways of taking parts of natural biological systems, characterizing and simplifying them, and using them as components of a newly engineered biological system.

The research now under way at Lisbon Technical University, together with that being developed in many other research centers worldwide will contribute to an old dream of humanity, that of understanding the grand design of life on earth.

ACKOWLEDGEMENTS

The authors would like to acknowledge the contributions of past and present members of the ALGOS/KDBIO group of INESC-ID and the Biological Sciences Group of CEBQ. In particular, we acknowledge the contributions of Miguel Teixeira, Pedro Monteiro, Pooja Jain, Sandra Tenreiro, Alexandra Fernandes, Nuno Mira, Marta Alenquer, Alexandra Carvalho, Sara Madeira, Nuno Mendes, Ana Ramalho, Luís Coelho, Ana Casimiro, José Caldas, Miguel Bugalho, Alexandre Francisco, Artur Lourenço, Luís Russo, Christian Nogueira, Carlos Oliveira, Rodrigo Moisés, Orlando Anunciação, Óscar Lopes and Susana Vinga. They also thank external collaborators that have helped in a number of initiatives, including Marie France Sagot and André Goffeau.

This research was supported by FEDER, FCT and the POSI, POCTI and PDCT programs under projects POSI/EIA/57398/2004, POSI/SRI/47778/2002, POCTI/AGG/38110/2001, POCTI/AGR/45347/2002, POCTI/BME/46526/2002, POSI/EIA/57398/2004 and PDCT/BIO/56838/2004.

REFERENCES

[1] J. Watson and F. Crick*, A structure for Deoxyribose Nucleic Acid,* Nature, 171, pp. 737:738, 1953.

[2] R. Staden, *A new computer method for the storage and manipulation of DNA gel reading data* Nucleic Acids Research, 25;8(16), pp. 3673:94, 1980.

[3] R. A. Gibbs, R.A. *Pressing ahead with human genome sequencing*. Nature Genetics,. 11**,** pp. 121:125, 1995.

[4] Commission on Life Sciences, National Research Council. *Mapping and Sequencing the Human Genome*, National Academy Press: Washington, D.C., 1988.

[5] R. D. Fleischmann, A. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick et al. *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd*, Science 269(5223), pp. 496:512, 1995.

[6] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S. G. Oliver. *Life with 6000 genes*. Science. 274(546), pp. 563:567, 1996.

[7] Blattner et al.. *The complete genome sequence of Escherichia coli K-12*, Science, 277(5331), pp. 1453:1474, 1997.

[8] The *C. elegans* Sequencing Consortium, *Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology,* Science 282(5396), pp. 2012:2018, 1998.

[9] The *Arabidopsis* Initiative, *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana*, Nature 408, pp. 796:815 2000.

[10] M. Adams et al. Science, *The Genome Sequence of Drosophila melanogaster*, Science 287(5461), pp. 2185:2195, 2000.

[11] J. C. Venter et al, *The Sequence of the Human Genome*, Science, 291(5507), pp. 1304:1351, 2001.

[12] International Human Genome Sequencing Consortium, *Initial sequencing and analysis of the human genome*, Nature 409, pp. 860:921, 2001.

[13] J. L. Weber and E. W. Myers, *Human Whole-Genome Shotgun Sequencing*, Genome Research, 7(5), pp. 401:409, 1997.

[14] S. B. Needleman and C. D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. Journal of Molecular Biology 48, pp. 443:453, 1970.

[15] T. F. Smith and M. S. Waterman, *Identification of Common Molecular Subsequences*. Journal of Molecular Biology 147, pp. 195:197, 1981.

[16] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *Basic local alignment search tool* Journal of Molecular Biology, 215(3):403:10, 1990.

[17] W. R. Pearson and D.J. Lipman. *Improved tools for biological sequence comparison*. Proceedings of the National Academy of Sciences, 85, pp. 2444:2448, 1988.

[18] E. Uberbacher and R. Mural. *Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach*. Genetics, 88, pp. 11261:11265, 1991.

[19] M. Zhang. *Identification of protein coding regions in human genome by quadratic discriminant analysis*, Genetics, 94, pp. 565:568, 1997.

[20] C. Burge and S. Karlin. *Prediction of complete gene structures in human genomic DNA*. Journal of Molecular Biology, 268, pp. 78:94, 1997.

[21] R. Guigó, S. Knudsen, N. Drake and T. F. Smith. *Prediction of gene structure*. Journal of Molecular Biology, 226, pp. 141:157, 1992.

[22] P. Monteiro, A. Ramalho, A. T. Freitas and A. L. Oliveira, *DECIDE A Gene Finding Evaluation Tool*, Proceedings of BKDB2005 - Bioinformatics: Knowledge Discovery in Biology, pp. 68:72, 2005.

[23] M. Y. Galperin, The *Molecular Biology Database Collection: 2006 update* Nucleic Acids Research, 34: pp. D3:D5, 2006.

[24] J. M. Cherry, C. Adler, C. A. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng and D. Botstein, *SGD: Saccharomyces Genome Database*. Nucleic Acids Research., 26, pp. 73:79, 1998.

[25] U. Güldener, M. Münsterkötter, G. N. Kastenmüller, J. Strack, C. Lemer, J. Richelles, S. J. Wodak, J. García-Martínez, J. E. Pérez-Ortín, H. Michael, A. Kaps, E. Talla, B. Dujon, B. André, J. L. Souciet, J. D. Montigny, E. Bon, C. Gaillardin and H. W. Mewes CYGD: the comprehensive yeast genome database, *Nucleic Acids Research*, 33, pp. D364:D368, 2005.

[26] M. Riffle, L. Malmström and T. N. Davis, *The yeast resource center public data repository*, Nucleic Acids Research, 33, D378-D382, 2005.

[27] J. van Helden, *Regulatory sequence analysis tools*. Nucleic Acids Research, 31, pp. 3593:3596, 2003.

[28] M. C. Teixeira, P. Monteiro, P. Jain, S. Tenreiro, A. R. Fernandes, N. P. Mira, M. Alenquer, A. T. Freitas, A. L. Oliveira and I. Sá-Correia, *The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae*, Nucleic Acids Research, 34, pp. D446:D451, 2006.

[29] M. F. Sagot. *Spelling approximate repeated or common motifs using a suffix tree*, Proceedings of Latin'98, LNCS 1380, pp. 111:127, 1998.

[30] J. van Helden, A. F. Rios and J. Collado-Vides. *Discovering regulatory elements in non-coding sequences by analysis of spaced dyads*. Nucleic Acids Research, 28, pp. 1808:1818, 2000.

[31] A. M. Carvalho, A. T. Freitas, A. L. Oliveira and M. F. Sagot, *An efficient algorithm for the identification of structured motifs in DNA promoter sequences*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 3(2), 2006.

[32] A. M. Carvalho, A. T. Freitas, A. L. Oliveira and M. F. Sagot, *A highly scalable algorithm for the extraction of cis-regulatory regions,* Proceedings of the 3rd Asia Pacific Bioinformatics Conference, pp. 273:282, 2005.

[33] A. M. Carvalho, A. T. Freitas, A. L. Oliveira and M. F. Sagot, *A parallel algorithm for the extraction of structured motifs,* Proceedings of the 19th ACM Symposium on Applied Computing, pp. 147:153, 2004.

[34] J. L. DeRisi, V. R. Iyer, P. O. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale*. Science. 278(5338). pp. 680-686, 1997.

[35] S. le Crom, F. Devaux, C. Jacq, P. Marc, *yMGV: helping biologists with yeast microarray data mining*, Nucleic Acids Research, 30, pp. 76:9, 2002.

[36] J. E. Celis, M. Kruhøffer, I. Gromova, C. Frederiksen, M. Østergaard, T. Thykjaer, P. Gromov, J. Yu, H. Pálsdóttir, N. Magnusson and T. F. Ørntoft, *Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics.* FEBS Letters 480, pp. 2:16, 2000.

[37] Y. Cheng and G. M. Church, *Biclustering of Expression Data,* Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pp. 93:103, 2000.

[38] S. C. Madeira and A. L. Oliveira, *Biclustering algorithms for biological data analysis: A survey*. IEEE/ACM Transactions on Computational Biology and Bioinformics 1(1), pp. 24:45, 2004.

[39] S. C. Madeira and A. L. Oliveira, *A Linear Time Biclustering Algorithm for Time Series Gene Expression Data*, Proceedings of the 5th Workshop on Algorithms in Bioinformatics, LNCS 3692, pp. 39:52 , 2005.

[40] X. Zhou, X. Wang, R. Pal, I. Ivanov, M. Bittner and E. Dougherty *A Bayesian connectivity based approach to constructing probabilistic gene regulatory networks*. Bioinformatics, 20 pp. 2918:2927, 2004.

[41] K. Kyoda, K. Baba, S. Onami S and H. Kitano, *DBRF–MEGN method: an algorithm for deducing minimum equivalent gene networks from large-scale gene expression profiles of gene deletion mutants*, Bioinformatics. 20, pp. 2662:2675, 2004.

[42] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford and R. A. Young, *Transcriptional regulatory networks in Saccharomyces cerevisiae*. Science 298, pp. 799:804, 2002.

[43] J. Ihmels, S. Bergmann and N. Barkai, *Defining transcription modules using large-scale gene expression data*, Bioinformatics, 20, pp. 1993:2003, 2004.

[44] M. Zou and S. D. Conzen, *A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data*. Bioinformatics, 21(1), pp. 71:79, 2005.

[45] S. Tenreiro, A. R. Fernandes and I. Sá-Correia *Transcriptional activation of FLR1 gene during Saccharomyces cerevisiae adaptation to growth with benomyl: role of Yap1p and Pdr3p*. Biochemical and Biophysical Research Communications, 280, pp. 216:222, 2001.