Feature

Circuits and Systems for High-Throughput Biology

© DIGITAL VISION,GETTY IMAGES

C. Nardini, L. Benini, and G. De Micheli

Introduction

The beginning of this millennium has been marked by some remarkable scientific events, notably the completion of the first objective of the Human Genome Project [1], i.e., the decoding of 3 billion bases that compose the human genome. This success has been made possible by the advancement of bio-engineering, data processing and the collaboration of scientists from academic institutions and private companies in many countries. The availability of biological information through web-accessible open databases has stirred further research and enthusiasm. More interestingly, this has changed the way in which molecular biology is approached today since the newly available large amount of data require close interaction between information technology and life science in a way not appreciated

Christine Nardini and Luca Benini are with the Dipartimento di Elettronica Informatica e Sistemistica (DEIS), University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy, E-mail: cnardini@deis.unibo.it, E-mail: lbenini@deis.unibo.i., Giovanni de Micheli is with the Laboratoire des Systèmes Intégrés, EPFL, Bâtiment INF, Station 14, 1015 Lausanne, Switzerland, E-mail: giovanni.demicheli@epfl.ch.

 1531-636X/06/$20.00©2006 IEEE

before. Still much has to be accomplished to realize the potential impact of using the knowledge that we have acquired. Several great challenges are still open, such as diagnosing and treatment of a number of diseases, understanding details of the complex mechanisms that regulate life, predicting and controlling the evolution of several biological processes. Nevertheless, there is now unprecendented room to reach these objectives, because the underlying technologies that we master have been exploited only to a limited extent.

High-throughput biological data acquisition and processing technologies have shifted the focus of biological research from the realm of traditional experimental science (*wet biology*) to that of information science (*in silico biology*). Powerful computation and communication means can be applied to a very large amount of apparently incoherent data coming from biomedical research. The technical challenges that lie ahead include the interfacing between the information in biological samples and information, and its abstraction in terms of mathematical models and binary data that computer engineers are used to handle. For example, how can we automate costly, repetitive and time consuming processes for the analysis of data that must cover the information contained in a whole organism genome? How can we design a drug that triggers a specific answer? Anyone wearing the hat of a Circuit and System engineer would immediately realize that one important issue is the interfacing of the biological to the electrical world, which is often realized by microscopic probes, able to capture and manipulate bio-materials at the molecular level. A portion of costly and time consuming experiments and tests that we used to do *in vitro* and/or *in vivo*, can now be done *in silico*. The concept of *Laboratory (Lab) on Chip* (LoC) is the natural evolution of *System on Chip* (SoC) by using an array of heterogeneous technologies. Whether LoCs will be realized on a monolithic chip, or as a combination of modules, is just a technicality. The revolution brought by Labs on Chips is related to the rationalization of bioanalysis, the drastic reduction of sample quantities, and its portability to various environments.

We have witnessed the widespread distribution of complex electronic systems due to their low manufacturing costs. Also in this case, LoC costs will be key to their acceptance. But it is easy to foresee that LoCs may be mass produced, with post-silicon manufacturing technologies, where large production volumes correlate to competitive costs. At the same time, the reduction of size, weight and human intervention will limit operating costs and make LoCs competitive. Labs on Chips at medical *points of care* will fulfill the desire of fast and more accurate diagnosis. Moreover, diagnosis at home and/or at mass transit facilities (e.g., airports) can have a significant impact on the overall population health. LoCs for processing environmental data (e.g., pollution) may be coupled with wireless sensor networks to better monitor the planet.

The use of the information produced by the Human Genome Project (marking the beginning of the Genomic Era) and its further refinement and understanding (post-Genomic Era), as well as the consequences related to moral and legal implication for the betterment of society has just started.

In fact, the decoding of the Human Genome paved the way to a different approach to molecular biology, in that it is now possible to observe the interrelations among whole bodies of molecules such as genes, proteins, transcripts, metabolites in parallel (the so called *omic* data like genomes, proteomes, transcriptomes, metabolomes, etc.), rather than observe and characterize a single chain of a cascade of events (i.e. perform gen*omic* vs gen*etic* analyses). In other words, molecular biology underwent an important shift in the paradigm of research, from a reductionist to a more systemic approach (*systems biology*) for which models developed in engineering will be of primary importance.

Background

Before discussing in more details in the hardware and software tools made available for the monitoring of the molecular activity of living cells, a brief presentation of the modality by which information is processed in living systems is mandatory.

While the DNA molecule embeds almost all the information a cell can manage and produce, only a portion of it is actually used in any given cell under specific conditions, depending on the role and environment in which the cell lives (i.e., temperature, radiations, ongoing disease processes, etc.). This differentiation is managed in a complex process that results in the controlled *transcription* of targeted strands of DNA (genes) into mRNA (a particular type of RNA). DNA and RNA are nucleic acids made of basic units, *nucleotides*, of only 4 different types. mRNA is then *translated* into proteins, the real actuator of molecular processes, by means of a coding that allows to match every triplet of mRNA nucleotides (*codon*) with the corresponding *amminoacid*, the basic building unit of proteins. Translation, along with other processes, leads to the synthesis of a protein, unless other mechanisms such as the recently discovered RNA interference mechanism, interferes by blocking the mRNA translation. The whole process can be disrupted or strongly altered by mutations in the DNA original sequence due to the duplication or deletions of entire sequences, or to punctual mutations (*Single Nucleotide Polymorphisms*, SNPs).

The variety of information processed by living cells is nowadays mirrored in the plurality of devices designed to

monitor the quantities and qualities of molecules involved at any given stage of the information flow. It is in this diverse landscape that information technology is crucial to exploit the new bodies of data and to make it useful to society namely for i) the generation of efficient *omic* platform sensors, ii) the implementation of dedicated algorithms for mining the resulting data and iii) the definition of adapted approaches to integrate information from different platforms.

In the sequel, we will review first hardware supports and technologies that allow the studies of the whole bodies of molecules through the different steps of the biological information flow as depicted in Figure 1 (see also [2]), and then some of the algorithms and softwares for data processing and integration. We conclude with a perspective on where this field is projected toward the future.

Hardware

DNA sequencing is the process of finding the exact sequence of bases in a DNA sample. Sequencing has been the main technique to enable the achievement of the Human Genome Project. With sequencing it has been possible to determine the gene structure of *homo sapiens* and other species. Unfortunately, sequencing is a time consuming and complex operation. Up to now, there is no sequencer that can take a full DNA sample and analyze it. Hence, the DNA is typically first split into smaller pieces that are analyzed separately and then the information is combined together. Even though sequencing is today highly automated, it is still a time consuming and expensive task. Nevertheless, there is an open challenge to construct means for low-cost affordable sequencing [3]. Whereas the sequencing produced by the Human Genome Project aims at determining the exact sequence in a DNA sample of healthy individuals, it is often the case that it is important to determine the genomic alterations at different levels of the information transmission pipeline.

Most of the *omic* platforms are based on the properties of complementarity of molecular molecules. In fact, it is a distinctive feature of nucleic acids (DNA and RNA) that they are able to link (*hybridize*) to dual fragments. This ability is used in nature for self-replication and self-

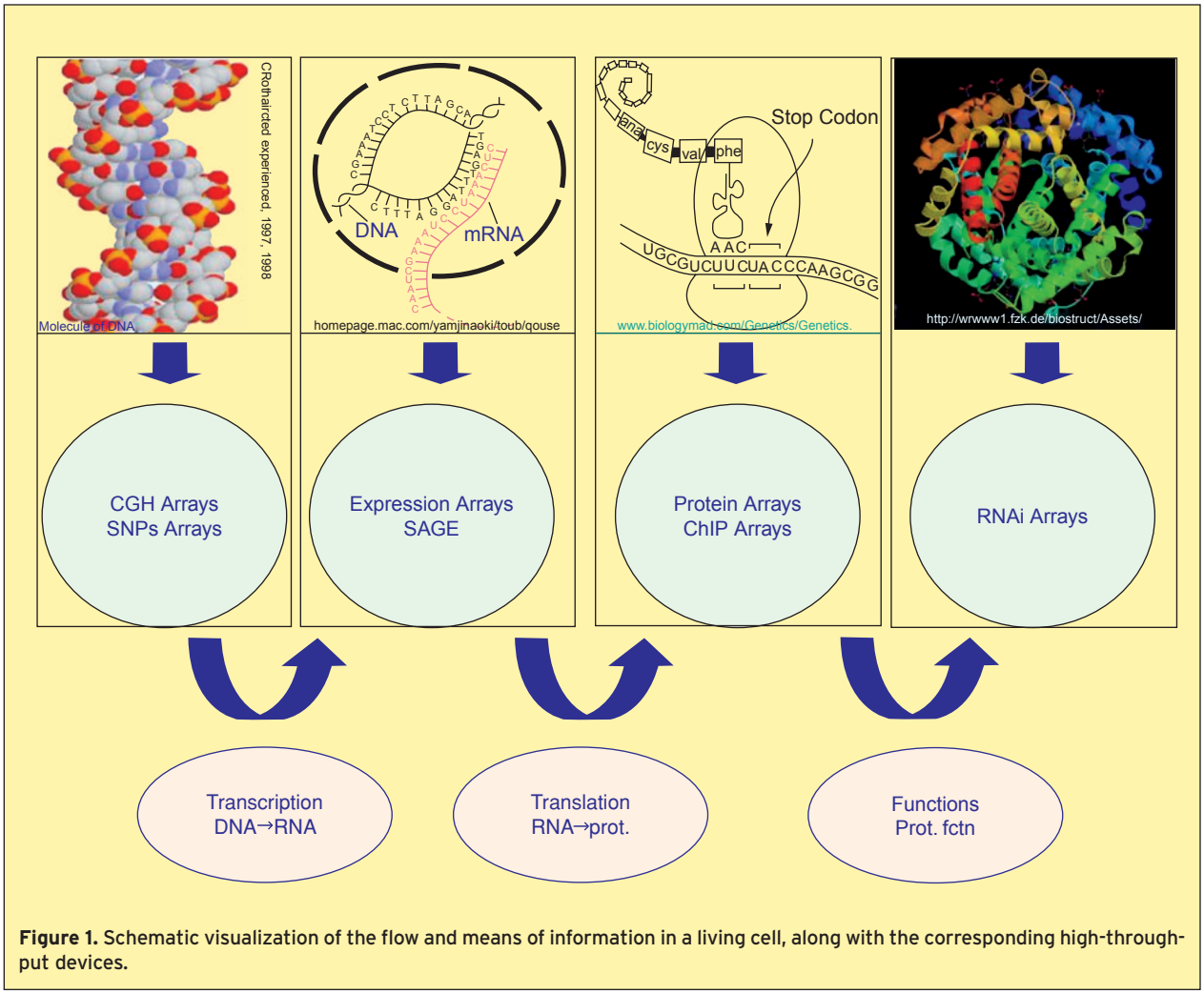


Figure 1. Schematic visualization of the flow and means of information in a living cell, along with the corresponding high-throughput devices.

repair of DNA molecules, and in sensing approaches it can be used for recognition. Similarly, the natural ability of antibodies to specifically link to selected proteins can be used for proteins' recognition. High-throughput devices take advantage of complementarity properties by generating arrays of reference molecules to which a mixture of appropriate but unknown molecules can attach themselves in proportion to their abundance, and thus they can be recognized and quantified.

Depending on the chemistry of the reference molecule, arrays can be used to capture DNA strands and quantify duplications or deletions (*Comparative Genomic Hybridization* –CGH- arrays) or evaluate punctual mutations (*Single Nucleotide Polymorphysm* –SNP- arrays).

Using mRNA information it is possible to measure the relative expression of the genes activated in the cells in a tissue, by either labeled or unlabeled techniques for sensing. Labeled techniques take their name from the fact that DNA samples are labeled with a fluorescent material, and the presence of a sample at a site is detected by optical scanning. Non-labeled techniques implement a direct reading of the binding to the probe, and are described later.

Among the labeled techniques, the first widespread technology is the cDNA microarray technology developed at Stanford by P. Brown and co-workers [4]. Probes are fabricated by depositing complementary DNA (cDNA) or synthetic oligonucleotides. These probes are processed to keep them unfolded. Conversely, on *in situ* built arrays [5], probes are synthesized directly on the substrate using, for example, light-directed chemical synthesis realized by photolithographic micro-fabrication techniques typical of the semiconductor industry as in the Affymetrix arrays (See Figure 2.1).

Recently, there have been several attempts to develop non-labeled techniques, with the objective of simplifying the readout process (by eliminating the laser scanning), reducing the noise, and possibly by combining intelligent processing on the same substrate, thus achieving a device that can carry out analysis on the filed (e.g., at a point of care). Some of these techniques are based on measuring the variation of capacitance between the probe base and the electrolyte [6], [7] (see Figure 2.2). This capacitance varies according to the fact that the probes are hybridized to the samples or not. Other approaches are based on ISFET technology [8], by fabricating transistors arrays whose gates host the probes. Thus, hybridization modulates the gain of the transistors, and an electrical readout circuit can provide a direct measure of gene expression levels. Finally, other techniques are based on ultraviolet light absorbance by the probes: the absorbance varies with hybridization and

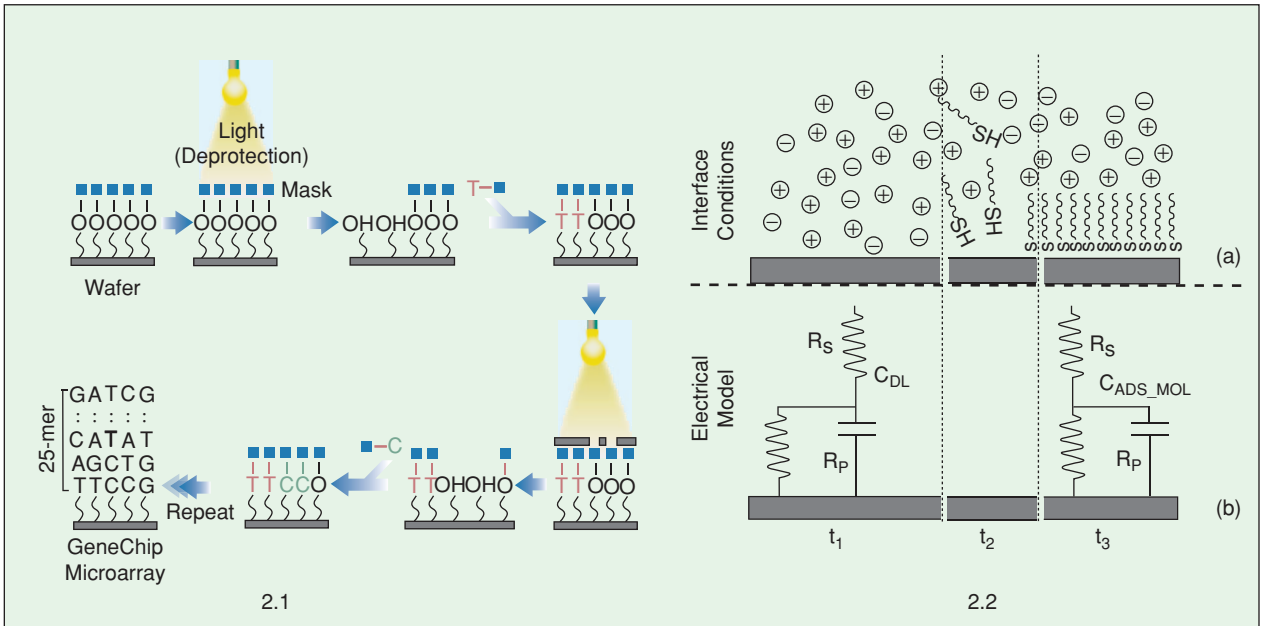


Figure 2. Shows examples of labeled and non-labeled techniques for gene expression analyses. Figure 2.1: In the Affymetrix technology, probes are synthesized in parallel by adding a nucleotide, (A, C, T or G), on top of the previously constructed chains. To define which oligonucleotide chain will receive a nucleotide, photolithographic masks with 20 square micron windows are placed over the coated wafer. When ultraviolet light is shone over the mask in a fabrication step, the exposed linkers become available for nucleotide coupling. These steps are iterated until the probes reach their full length, usually 25 nucleotides. Figure 2.2: schematic representation of the interface with funzionalization. (a) Transformation of the interface during oligonucleotides immobilization. (b) Corresponding electrical model: the capacitance and thus the overall impedance varies with hybridization.

that can be measured by the residual radiation impinging on sensors below the probes.

Independently from the technology adopted, microarrays for gene expression are limited to the evaluation of the expression of known genes that represent the target on the array. To overcome this limitation another high throughput technique has been devised: Serial Analysis of Gene Expression (SAGE, [9]). This approach does not require a reference slide, and in a nutshell, it collects from every molecule of mRNA extracted from a tissue a *tag*, made of 14 bases that only identify the corresponding mRNA. These tags are then glued together in *concatemers* that are multiplied through *Polymerase Chain Reaction* (PCR), a technique to amplify DNA subsequences), in order to obtain a large quantity of high-quality molecules that are then used to perform its sequencing. Finally, once the sequencing is performed, a computational analysis will evaluate the number of replicas of any given tag of which the concatemer is made, generating a frequency distribution. This technique is expensive, but does not require any previous knowledge of the mRNAs, and yields a precise quantitative measure of the mRNAs.

For the study of proteomes, several complex issues need to be addressed. In fact, the variety of proteins, (each one of the ~20,000 human genes can code for more than one protein), and of their techniques of interactions, is mirrored in the difficulty of reproducing systems that can mimic some complementarity strategy to capture the targets [10]. Proteins, in fact, can interact in natural processes among them to form bigger complexes, or with specific antibodies (protein produced by the immune system's cell), or with DNA fragments, moreover, the relationship among these molecules is the result of a complex balance between mechanical and chemical interactions, difficult to preserve in the arraying processes. The earliest device designed for the study of proteomes is based on the reaction antigene-antibody, a strategy commonly used by the immune system that is able to synthetize proteins, (antibody), able to capture and then lead to elimination of a given specific protein, (antigen), that has triggered the whole process. Several limitations arise, notably in the fact that the relation antibody-antigen is a very complex one and that antibodies need to be known for every specific protein. Moreover, despite this specificity there is evidence of the phenomenon of *cross-reactivity* that indicate the possibility of a protein to react with more than one molecule. For this reasons *sandwich* arrays have been designed to link each protein with two antibodies to increase specificity and sensitivity. To avoid the limitations due to the antibody design, efforts are now undertaken to the generate proteins directly from libraries of expressed sequences (nucleic acids sequences coding for proteins). Taking into account the importance of proteins interaction in such fields as drug design, this research area requires further improvements.

A more complex approach aims at taking care of the interaction between proteins and DNA, a complex relation in which a protein or a complex of interacting proteins enables the transcription of a gene that is, in turn, translated into a protein that can interact with the same complex, or with other proteins, or other DNA binding sites, in a very complex network interaction. To observe a part of these interaction with high parallelism, other types of devices have been designed [11]. Namely, the DNA regions where a protein (promoter) binds to activate (promote) a given gene transcription, can be identified by first capturing the protein with the corresponding antigene and then performing an experiment (chromatin immunoprecipitation, ChIP) that allows to isolate a complex of the promoter *and* the DNA binding site. The DNA binding site can be processed and then arrayed on a chip, (thus the name of Chip-chip arrays for this technique), and a comparison can be performed in the sequence of the binding site in healthy versus pathologic samples.

Finally, recent applications [12] have been devised to study processes ongoing in every single cell, rather than in the specific type of molecules extracted from a supposedly uniform tissue. This new approach is based on two biochemical mechanisms: *RNA interference* (RNAi) and *transfection*.

The first mechanism is a complex process discovered in 1998 in *c.elegans* that triggers the suppression of a gene expression, after the transcription process and before its translation into protein. This mechanism can be used to force the suppression of a gene expression after its transcription, by designing appropriate RNA molecules (siRNA) that through several steps lead to the *silencing* of the mRNA, impeding its encoding in amminoacid sequence. The second mechanism indicates a natural process that is the integration of foreign DNA or RNA into cells (in culture). Parallel transfection of hundreds of nucleic acids strands can be carried out in a microarray format, by preparing vectors containing appropriate siRNA. This array slide is enriched with molecules able to induce transfection, so that when the slide is covered with a lawn of adherent cells, these cells growing on top of the spots can be transfected, suppressing expression of specific proteins in spatially distinct groups of cells. Using both mechanisms, the study of the effect of the suppression of specific genes can be performed at cellular level with high parallelism [13].

Most of these approaches share in common the desire to perform—on a large scale—experiments that used to be performed with negligible or absent parallelism. From this perspective, engineering approaches are not only useful in the improvement of the single experiment

automatic setup, its reproducibility, but also in all the issues that descend from the optimization required when moving from custom to standardized, high-throughput processes.

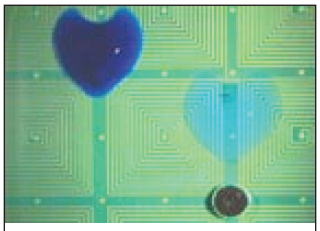


Figure 3. Two immobilized droplets after the passage of a droplet carrying magnetic particles through both of them.

Finally, in the post-genomic era, engineering principles are not only useful for high parallelization, but also for an efficient serialization. Laboratory on chips can now mix sensing methods with transport and other manipulations of DNA material. For example, *polymerase chain reaction* (PCR) is a well known technique to amplify DNA sub-sequences. PCR can be readily realized in hardware by creating micro-chambers with accurate temperature control. STMicroelectronics *In-Check* platform is an experimental product that combines DNA amplification with sensing.

Several transport techniques have been applied with micro-fluidic technologies. Such technologies aim at providing a versatile transport and separation means for small amounts of samples. Techniques ranging from electrophoresis [14] to magnetic droplet manipulation and transport have been successfully used [15], (Figure 3).

Software

Array-based bio-sensors are becoming an ubiquitous technology to be applied in the medical practice and in experimental biology, as well as in environment and food contamination control. For this reason, not only their construction and use, but also their interpretation are subject to continuous improvement. In particular we can identify three main levels of data mining approaches in which platforms for *omic* data are being used.

At the first level there are algorithms for data mining on the output of high-throughput platforms. This approach is used to characterize the interactions among sets of molecules of the same type (e.g., transcripts in micro-arrays for gene expression). A second level concerns the joint mining of the array outputs with other types of information, such as demographic or clinical ones. Finally, the last and more recent level of data mining aims at extracting information from multiple *omic* platforms and data sources. These approaches are used to obtain the most complete, integrated level of information, and often also to predict unknown features. In the following we describe these three levels of complexity in more detail (see Figure 4).

Arrays used for the evaluation of *omic* data produce the tables containing the measure of a given type of molecule abundance in different conditions. In general, the outcome of an experiment with these arrays is a matrix of real numbers, where rows correspond to molecules, columns to different arrays obtained under different conditions (e.g., different tissue samples or different patients), and entries to the abundance of the molecule, for the given condition (e.g., for gene expression arrays, its presence, possibly relative to some reference level). Then these matrices need to be preprocessed for normalization. Data analysis consists mainly in detecting common and consistent trends across conditions, and across molecules, or both. A typical approach in microarrays for gene expression (the more mature of these high throughput approaches) is to perform data clustering, that is determining the subset of genes and conditions expressing a common trend. Data clustering is a difficult problem and, in general, it is computationally intractable. Various heuristic algorithms have been proposed. The simplest category of clustering algorithm (hierarchical cluster, Self Organizing Maps SOM), solves the problem by grouping data based on a distance metric or variance reduction (Principal Component Analysis PCA), and produces a set of non-overlapping clusters which identify non-overlapping groups of genes [16]. Conversely, bi-clustering [17] is a set of techniques that cluster both genes and conditions, producing possibly overlapping clusters which are identified by proximity in a bi-dimensional metric (see Figure 5.1). In other words, a bi-cluster is a subset of the rows and the columns of the data matrix, such that for each row/column pair the distance $(a_{ij} - a_{kj}) - (a_{il} - a_{kl})$ is smaller than a given threshold. It has been argued that biclusters can reveal more biological information than other types of clusters. Indeed, they are compatible with our understanding of the cellular processes: we expect a subset of genes to be co-regulated and co-expressed under certain experimental conditions. In gene expression analysis, biclustering is more suitable for cases where genes have multiple functions.

Among the various approaches to bi-clustering, Yoon and others [18] developed an efficient algorithm that can compute all bi-clusters in a data set. Moreover, these clusters display *coherence* as compared to bi-clusters computed with other methods. Coherence is a statistical property that can be measured in terms of mean squared residues (see Figure 5.2).

Clustering can be applied to data in different forms, and in particular to genetic measurements of sequences of experiments done at different time points. In this case, the time-series of genetic expression value can shed light on the gene regulatory mechanisms.

As long as data mining algorithms are applied to the bare array values, techniques for genomic clustering are

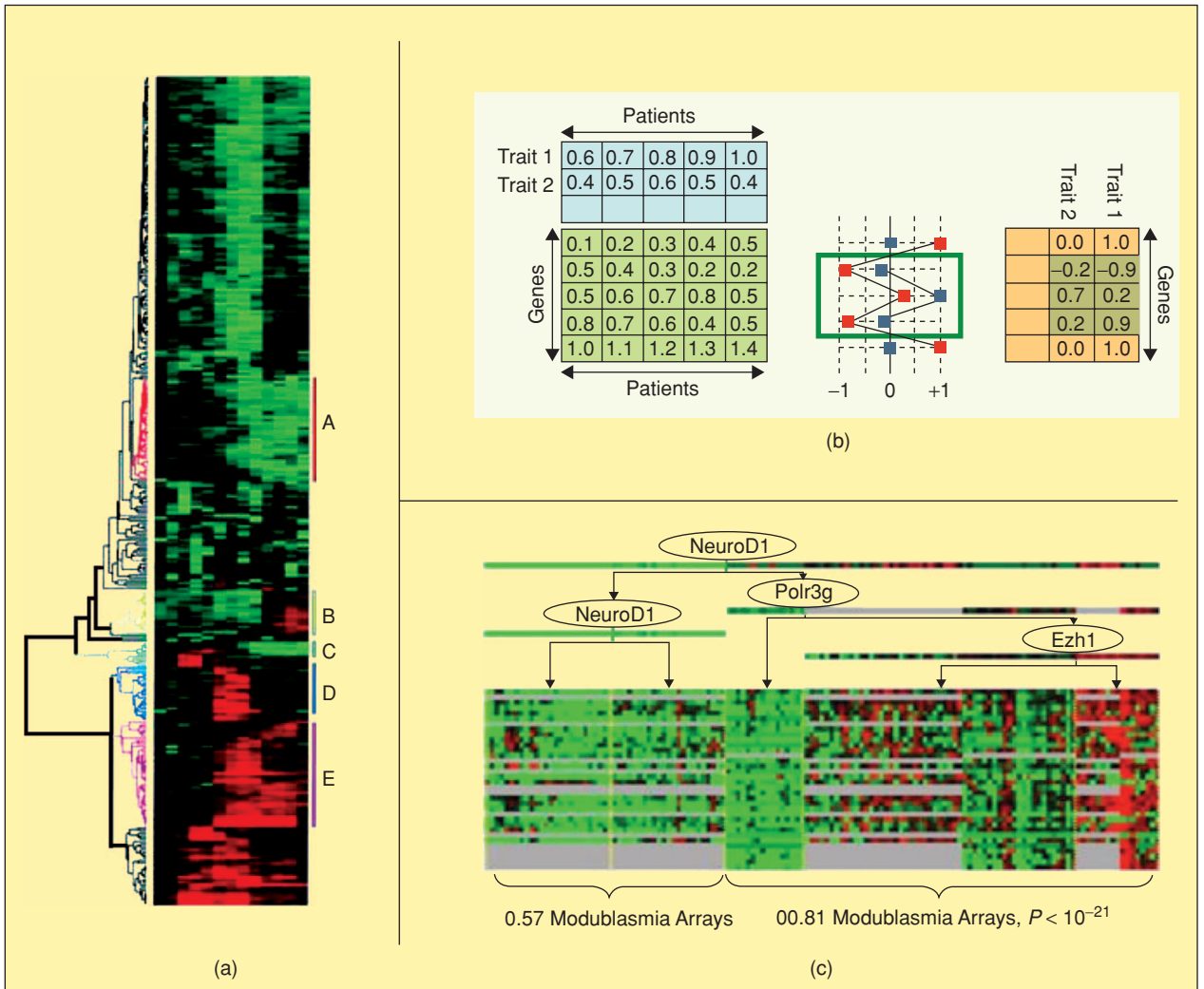
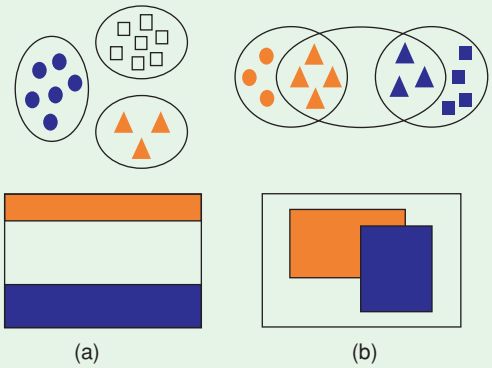


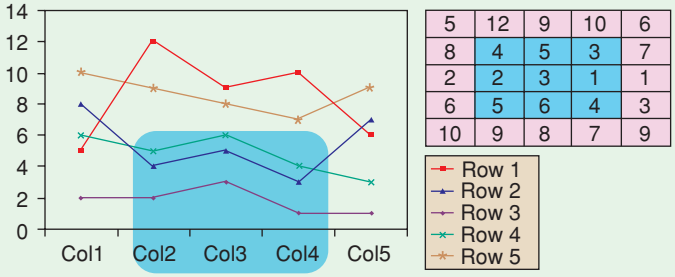
Figure 4. Analyses of increasing computational complexity for genomic data. Panel (a) depicts a heat-map, a typical representation of micro-array for gene expression, showing graphically the expression levels in a matrix, whose rows are genes and column conditions, after they have been reorganized using hierarchical clustering. Panel (b) shows an example of clinical-genomic analysis with a schematic representation of co-clustering. Input data is on the left, while the SAM correlation matrix is on the right. The diagram in the middle shows the trends and the green box shows the gene subset on which a coherent trend is detected. Panel (c) represents an example of analysis using module network, where information of regulators of transcription and expression data are used to detect causal relationships in the expression profile of a disease of interest.

called *unsupervised*. However, it is often useful to introduce a second layer of complexity, given by other types of information, such as clinical ones, to integrate and possibly complete the analysis (*supervised* analysis of microarray). The so called *Clinical-Genomics* is a new yet promising field of practical application of this approach and takes advantage of the information gathered by the long practical experience coded into clinical traits and the new, robust and detailed genomic ones. Early attempts were based on statistical correlation methods, such as using Spearman's coefficient. For instance, recent work has focused on establishing correlation between human genomic data and radiological traits [19]. In these studies, a set of clinical traits was manually extracted from medical images and then correlated to the data extracted from micro-arrays.

In this context various types of algorithms have been devised, such as *Significance Analysis of Microarray* (SAM, [20]) and *Gene Expression Enrichment Analysis* (GSEA, [21]). These algorithms extract a ranked set of genes candidate to be significantly related to a given external clinical trait. This ranking is based on the evaluation of various statistics (such as a modified t-test and Signal-to-Noise Ratio). Along this line of research an interesting approach called co-clustering has been defined by Yoon et al. [22]. This method is based on the following idea. First, a correlation matrix is constructed from the genetic and clinical data in matrix form using the statistic defined



5.1



5	12	9	10	6
8	4	5	3	7
2	2	3	1	1
6	5	6	4	3
10	9	8	7	9

5.2

Figure 5.1 Informal comparison of clustering and bi-clustering. **Figure 5.2:** Expression fluctuation of 5 genes in 5 experiments. The genes expressed by rows 2, 3 and 4 have coherent behavior in the experiments represented by columns 2, 3 and 4.

in SAM. Next, bi-clusters are searched for in this correlation matrix. This method was tested on data from *Acute Myelogenous Leukemia* (AML), with 6239 genes, 119 samples and 15 clinical parameters. The method found 43 clusters. Some of these have strong biological significance, for example by showing that the trait "survival" is correlated with genes that play a central role in the control of growth, differentiation and morphogenesis of normal and malignant cells.

Finally, a third layer of complexity is represented by the integration of diverse sources of *omic* data. These approaches take advantage of the diverse nature of the data available on publicly accessible databases and mostly of all possible interconnections that can be inferred for a given item of interest from several sources of knowledge. A very complex and challenging issue is given by the modeling of the numerous complex interactions that regulate genes behavior. In this context several efforts have been made to get more insight into the complex networks of interaction, a goal which is probably the most exciting and complex challenge of the post-genomic era. As an example, an interesting approach was devised by Segal et al. [23] who defined the *module network procedure*. This approach is based on probabilistic graphical models and uses information gathered from microarray for gene expression experiments, from databases collecting protein information, as well as from extended literature searches. These diverse sources of information are used to infer *regulatory modules*, that are a trinome made of a set of regulatory genes (*regulatory program*), that specify the behavior of coregulated genes (*regulated module*) under given conditions (*context*). Briefly, this method is based on the generation and iterative optimization of a regression tree, where nodes are represented by regulators (genes) whose over/under or unchanged expression decides the path to be followed in walking the tree further. The leaf of the tree are represented by a set of genes whose expression is coherent and influenced by all the regulators/nodes along the path. This approach allows to infer from expression experiments, that give a global but static image of the processes ongoing in the cell, information about reciprocal influences, offering a systemic and more dynamic picture of the ongoing events.

Another promising area of research for the efficient use of integrated information in *in silico* biology is research in multigenic hereditary diseases. The understanding of such diseases and possibly their control and treatment requires the individuation of all the genes that are disregulated or mutated. Traditionally the identification of genes candidate to be involved in such diseases necessitates tracing the genomes of several individuals from the same family, to highlight the area of the genome that is passed from generation to generation and thus is likely to encode the genes responsible for the disease. This process is extremely costly and time consuming, and inherently limiting the study of rare diseases since the number of individuals affected is very small. Conversely, public databases can be advantageously used to filter from the whole genome lists of candidate genes sharing features likely to cause multigenic diseases. Such characteristics are of very diverse nature. They range from the length of the protein encoded to the position of the gene on the genome and to the conditions and tissues in which the gene is expressed, and more (see [24] for a recent example of application tool).

As data mining gains momentum in extracting new information for the rapidly growing mass of *omic* data, the definition, construction and maintenance of controlled vocabularies for the systematization of knowledge, based on terms broadly accepted in the scientific

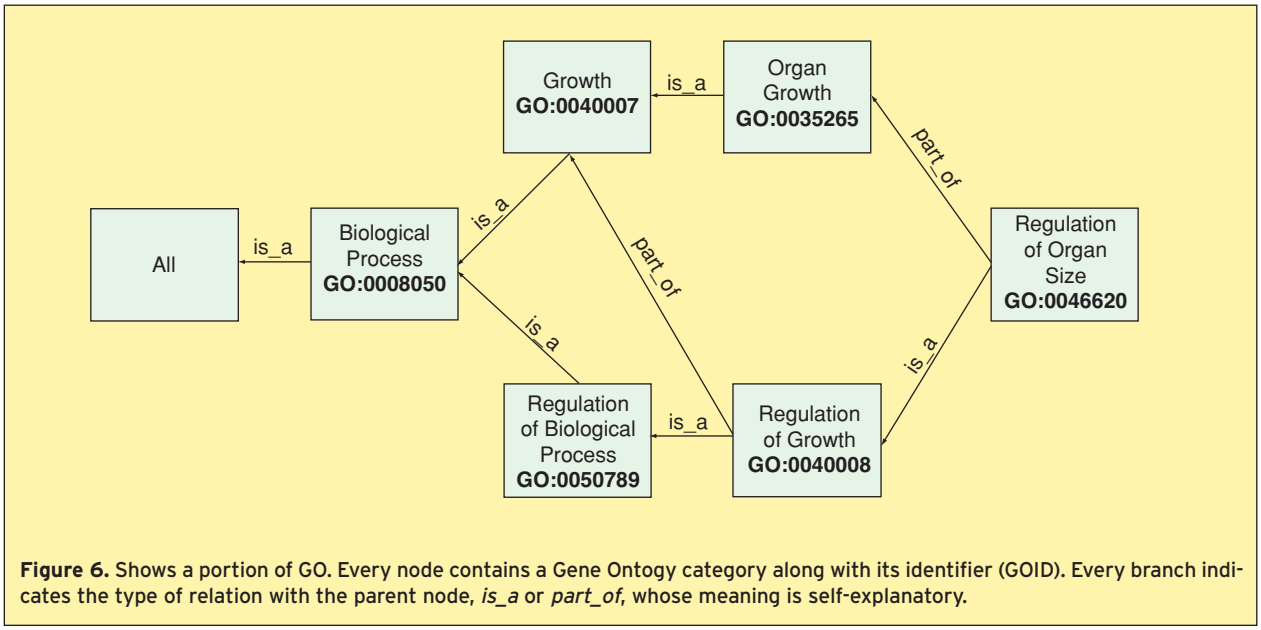


Figure 6. Shows a portion of GO. Every node contains a Gene Ontogy category along with its identifier (GOID). Every branch indicates the type of relation with the parent node, *is_a* or *part_of*, whose meaning is self-explanatory.

community is becoming an area of increasing interest. Several efforts are ongoing in this direction for the definition of ontologies, and controlled vocabularies based on a graphical structure that allows not only the classification of items of a given type, but also the quantitative evaluation of their semantic proximity. Two widely known examples of such vocabularies are *Gene Ontology* (GO, [25] and Figure 6) for the definition of genomic terms, and the *Medical Subject Heading* (MeSH, [26]) used for mining *PubMed* [27], the largest repository of medical and life science journals abstracts (MeSH is a part of a broader effort for a *Unified Medical Language System*, UML, [28]). These over-structures on the data allow not only a more efficient exchange of information among scientists, but also provide a remarkable contribution to the distillation of new knowledge. In fact, particularly for GO, a variety of tools have been designed to infer through statistical approaches the classification (*annotation*) of uncharacterized items in terms of known ones, based on some shared similarity. For example, genes that happen to be in a given set after a clustering process, are likely to act synergistically in a given cellular function. GO can then be used i) for validation purposes, to observe if the genes share in fact in large majority the same function, or ii) for discovery purposes, to extend the annotation of the most significant molecular role present in the cluster to the few un-annotated genes [29].

Perspectives

High-throughput biology is an experimental branch of biology that relies on specific circuits and systems for data acquisition and processing. Special hardware includes arrays that are functionalized to detect/capture different types of biological molecules as well as transport means on chip. These arrays exploit semiconductor and similar technologies, and benefit from integration because it is possible to combine probing with *in situ* computational engines, i.e., with processors that perform local data analysis. These processors can run programs implementing the data analysis algorithms. Micro-fluidic circuits play also an important role, as they provide us with the communication fabric to transport samples on chip. Moreover, several engineering approaches are required to improve the ability to capture different types of molecules, preserving their chemical and structural characteristics, while allowing high parallelism.

Overall Lab on Chips can be seen as the combination of sensing, transport and processing for performing biochemical analysis of small quantities. While LoCs are still expensive today, we believe that cost of operating them is much smaller than standard lab analysis and the unit cost will drop as mass production will develop. Moreover, portability is a major advantage of Lab on Chips. While this article has described only LoCs for biological analysis, LoCs can be used to synthesize small amounts of organic and/or inorganic samples by combining transport means for reagents to and from reaction microchambers.

While considering the entire bio-medical field, high-throughput analysis is ubiquitously present in research while its use in the day-by-day clinical practice is lagging. This is due to several factors such as the incompleteness of the *omic* information, the variety of techniques that evolution has implemented to preserve the continuity of the process of life, the intrinsic noise of

the high-throughput devices and the consequent variability of the results. For this precise reason, new circuit and probe architectures, as well as new devices, are needed to improve the certainty of the results and help to complete the characterization of the numerous processes that regulate cell life.

On the other hand new and improved algorithms, mathematical models and statistical approaches for the correct description of the body of data are required. Research in these fields may leverage the wealth of results in Circuit and System design and in data analysis algorithms. We are just at the beginning of a revolution that is bringing some aspects of biology and medicine closer to information science. The continuous and constant integration of medical and molecular knowledge with engineering approaches is a dynamic, incremental and crucial process leading to the efficient use, solution-oriented and possibly the improvement of tools and techniques used in engineering for both biological discoveries and medical treatment.

References

[1] Available: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

[2] A. Hsiao and M.D. Kuo, "High-througput biology in the post-genomic era," *Journal of Vascular Interventional Radiology*, to be published.

[3] J. Shendure, R. Mitra, C. Varma, and G. Church, "Advanced sequencing technologies: Methods and goals," *Nature*, vol. 5, pp. 335–344, May 2004.

[4] J. De Risi, L. Penland, P. Brown, M. Bittner, P. Meltler, M. Ray, Y. Chen, Y. Su, and M. Trent, "Use of a cDNA microarray to analyze gene expression patterns in human cancer," *Nature genetics*, vol. 14, no. 4, pp. 457–460, 1996.

[5] A. Pease et al, "Light-generated oligonucleotide arrays for rapid DNA sequencing analysis," *Proceedings of the National Academy of Sciences*, vol. 91, no. 11, pp. 5022–5026, 1994.

[6] C. Stagni, D. Esposti, C. Guiducci, C. Paulus, M. Schienle, Maugustyniak, G. Zuccheri, B. Samori, L. Benini, B. Ricco, and R. Thewes, "Fully electronic CMOS DNA detection array based on capacitance meausurement with on-chip analog to digital conversion," in *Proceedings ICSSC*, San Francisco, 2006.

[7] C. Guiducci, C. Stagni, G. Zuccheri, A. Bogliolo, L. Benini, A. Samorì, and B. Riccò, "DNA detection by integratable electronics," *Biosensors and Bioelectronics*, pp. 781–78719, (2004).

[8] G. Philippin and U. Bockelmann, "DNA detection on transistor array following mutation specific enzymatic apmplification," *Nanotech*, Montreux, 2005.

[9] S.L. Madden, C.J. Wang, and G. Landes, "Serial analysis of gene expression: from gene discovery to target identification," *Drug Discov Today*, vol. 5, no. 9, pp. 415–425, 2000.

[10] P. Bertone and M. Snyder, "Advances in functional protein microarray technology," *FEBS J.*, vol. 272, pp. 5400–5411, Nov. 2005.

[11] C.D. Herring, M. Raffaelle, T.E. Allen, E.I. Kanin, R. Landick, A.Z. Ansari, and B.Ø. Palsson, "Immobilization of escherichia coli RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays," *J. Bacteriology*, vol. 187, no. 17, pp. 6166–6174, 2005.

[12] D.B. Wheeler, A.E. Carpenter, and D.M. Sabatini, "Cell microarrays and RNA interference chip away at gene function," *Nat Genet*, vol. 37, pp. S25–S30, 2005.

[13] S. Mousses, N.J. Caplen, R. Cornelison, D. Weaver, M. Basik, S. Hautaniemi, A.G. Elkahloun, R.A. Lotufo, A. Choudary, E.R. Dougherty, E. Suh, and O. Kallioniemi, "RNAi microarray analysis in cultured mammalian cells," *Genome Res.*, vol. 13, p. 2341, 2003.

[14] U. Seger, S. Gawad, R. Johann, A. Bertsch, and P. Renaud, "Cell immersion and cell dipping in microfluidic devices," *LAB ON A CHIP*, vol. 4, no. 2, pp. 148–151, 2004.

[15] U. Lehmann, S. Hadjidj, V.K. Parashar, and M.A.M. Gijs, "Two dimensional magnetic manipulation of microdroplets on a chip," *Transducers'05*, 2005.

[16] J. Quackenbush, "Computationa Analysis of Micorarray Data," *Nat Rev Genet*, vol. 2, no. 6, pp. 418–427, June 2001.

[17] S. Madeira and A. Oliveira, "Biclustering algorithms for biological dada analysis: A survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[18] S. Yoon, C. Nardini, L. Benini, and G. De Micheli, "Discovering Coherent Biclusters from Gene Expression Data Using Zero-Suppressed Binary Decision Diagrams," *IEEE—Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 4, pp. 339–354, 2005.

[19] C. Nardini, S. Cha, D. Wang, M. Diehn, M. Mayo, L. Benini, G. De Micheli, and M. Kuo, "Approach for molecular characterization of glioblastoma multiforme: MRI correlation with cDNA expression profiles," in *AANS—Proceedings of the 6th Annual Meeting of the Neurological Surgeons*, San Francisco (www.ans.org), 2004.

[20] V. Tusher, R. Tibsharani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. National Academy of Science USA*, vol. 98, no. 9, pp. 5116–5121, Apr. 2001.

[21] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. National Academy of Science USA*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005.

[22] S. Yoon, L. Benini, and G. De Micheli, "Finding co-clusters of genes and clinical parameters," in *EMBC 2005*, no. 11.3.1.4, 2005.

[23] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data," *Nat Genet*, vol. 34, no. 2, pp 166–176, 2003.

[24] S. Rossi, D. Masotti, C. Nardini, E. Bonora, G. Romeo, E. Macii, L. Benini, and S. Volinia, "TOM: a web-based integrated approach for efficient identification of candidate disease genes," *Nucleic Acids Res.*, *Web Server Issue*, to be published.

[25] The Gene Ontology Consortium, "The Gene Ontology (GO) project in 2006.," *Nucleic Acids Res.*, vol. 34, pp. D322–D326, 2006.

[26] S.J. Nelson, M. Schopen, A.G. Savage, J.-L. Schulman, and N. Arluk, "The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation," in *Proceedings of the 11th World Congress on Medical Informatics*; 2004, CA. Amsterdam: IOS Press, pp. 67–69, 2004.

[27] Available: www.pubmed.gov

[28] Available: http://www.nlm.nih.gov/research/umls/umlsmain.html

[29] J.L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J.M. Mato, L.A. Martinez-Cruz, F.J. Corrales, and A. Rubio, "Correlation between gene expression and GO semantic similarity," *IEEE/ACM TCCB*, vol. 2 , no. 4, pp. 330–338, 2005.

Christine Nardini has the E.E. Master degree and focused Ph.D. research on bioinformatics, oriented to the analysis of data obtained from microarray for gene expression. She is currently post-doc at the University of Bologna. Research interests include adaptation and definition of statistically based algorithms for the extraction of associations between clinical and molecular data (clinical genomics) and for the efficient selection of genes candidate to be responsible of hereditary diseases. This research involves collaborations with the Computer Science Department at Stanford, The Center for Translational Medical Systems at UCSD (CA, USA), Unità di Genetica Medica S. Orsola, Dipartimento di Embriologia, DAMA (Data Mining and Analysis) Telethon Facility at Università di Ferrara, and École Polytechnique Fédérale de Lausanne (EPFL).

Luca Benini is an Associate Professor at the University of Bologna. He also holds a visiting faculty position at the École Polytecnique Fédérale de Lausanne (EPFL). He received the Ph.D. degree in electrical engineering from Stanford University in 1997. Dr. Benini's research interests are in the design of systems for ambient intelligence, from multi-processor systems-on-chip/networks on chip to energy-efficient smart sensors and sensor networks. He has published more than 250 papers in peer-reviewed international journals and conference proceedings, three books, several book chapters and two patents. He has been program chair and vice-chair of Design Automation and Test in Europe Conference. He has been a Member of the 2003 MEDEA+ EDA roadmap committee 2003. He is a member of the IST Embedded System Technology Platform Initiative (ARTEMIS): working group on Design Methodologies, a Member of the Strategic Management Board of the ARTIST2 Network of excellence on Embedded Systems and a Member of the Advisory group on Computing Systems of the IST Embedded Systems Unit.

He has been member of the technical program committee and organizing committee of several technical conferences, including the Design Automation Conference, International Symposium on Low Power Design, the Symposium on Hardware-Software Codesign. He is Associate Editor of the IEEE Transactions on Computer-Aided Design of Circuits and Systems and of the ACM Journal on Emerging Technologies in Computing Systems. He is a Senior Member of the IEEE.

Giovanni De Micheli is Professor and Director of the Integrated Systems Centre at EPF Lausanne, Switzerland, and President of the Scientific Committee of CSEM, Neuchatel, Switzerland. Previously, he was Professor of Electrical Engineering at Stanford University.

His research interests include several aspects of design technologies for integrated circuits and systems, with particular emphasis on synthesis, system-level design, hardware/software co-design and low-power design. He is author of: Synthesis and Optimization of Digital Circuits, McGraw-Hill, 1994, co-author and/or co-editor of five other books and of over 300 technical articles .

Dr. De Micheli is the recipient of the 2003 IEEE Emanuel Piore Award for contributions to computer-aided synthesis of digital systems. He is a Fellow of ACM and IEEE. He received the Golden Jubilee Medal for outstanding contributions to the IEEE CAS Society in 2000. He received the 1987 D. Pederson Award for the best paper on the IEEE Transactions on CAD/ICAS, two Best Paper Awards at the Design Automation Conference, in 1983 and in 1993, and a Best Paper Award at the DATE Conference in 2005.

He was President of the IEEE CAS Society in 2003, and he is currently President Elect of the IEEE Council on EDA and chairing the IEEE Product Package Committee . He is Program Chair of the Health and VLSI SOC conferences in 2006.

He was Editor in Chief of the IEEE Transactions on CAD/ICAS in 1987–2001. Dr. De Micheli was the Program Chair and General Chair of the Design Automation Conference (DAC) in 1996–1997 and 2000, respectively. He was the Program and General Chair of the International Conference on Computer Design (ICCD) in 1988 and 1989, respectively.