High Parallelism, Portability, and Broad Accessibility: Technologies for Genomics

CARLOTTA GUIDUCCI and CHRISTINE NARDINI
DEIS University of Bologna

Biotechnology is an area of great innovations that promises to have deep impact on everyday life thanks to profound changes in biology, medicine, and health care. This article will span from the description of the biochemical principles of molecular biology to the definition of the physics that supports the technology and to the devices and algorithms necessary to observe molecular events in a controlled, portable, and highly parallel manner. Throughout this discussion, constant attention will be given to the ultimate goals and applications of these innovations as well as to the related issues.

Categories and Subject Descriptors: H.4.0 [**Information Systems Applications**]: General

General Terms: Algorithms, Measurement, Reliability

Additional Key Words and Phrases: Genomics, biosensors, microarrays, point-of-care diagnostics

ACM Reference Format:

Guiducci, C. and Nardini, C. 2008. High Parallelism, Portability, and Broad Accessibility: Technologies for Genomics. ACM J. Emerg. Technol. Comput. Syst. 4, 1, Article 3 (March 2008), 39 pages. DOI = 10.1145/1330521.1330524 http://doi.acm.org/10.1145/1330521.1330524

1. INTRODUCTION

Technology is often perceived as a mean to improve life quality which, in turn, recalls the fundamental concept of health, its definition, preservation, and recovery. Although the last century has offered unprecedent advances in this direction, witnessed by extended life span, medicine is suffering from being still strongly based on empirical approaches. The improvements to come in life quality crucially depend upon the advances in medicine as an evidence-based science. The area dedicated to the quantification and comprehension of the relevant causes behind the effects observed in medicine is molecular biology.

Molecular biology has been the protagonist of life science in the last half century, and it consists of the study of biological events and dynamics at the

Authors' address: University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy; email: christine.nardini@unibo.it.

molecular level [Alberts et al. 1989]. In particular, it is concerned with the description of mechanisms regulating the chemical and functional interactions among the molecular actors playing in the cell. Due to the complex interplay and the nanomicroscopic dimensions of the objects under study, advanced technologies have had an important role in the development of molecular biology. Moreover, new achievements in molecular biology have in turn called for more advanced technological solutions in a cycle of design of novel tools and biological discoveries. It follows that this research area attracts the interests of investors from public and private institutions and appears to be a fast evolving and challenging area of research. In particular, enormous potential is perceived in the integration of information technologies and life sciences. This requires interdisciplinary competences for the design of innovative tools such as microfabricated devices for molecular analysis, screening and diagnostic, molecular manipulation, and data mining. In order to capture the principles that underlie these emerging fields, it is first necessary to understand the basic rules that govern the biology of the cell, namely, the network of events that allows the transmission of information in living cells.

Information related to this pipeline has been cumulated over the last 50 years by molecular biologists starting from the identification of the DNA structure in 1953 [Watson and Crick 1953a, 1953b]. From there forward, the identification of the fundamental mechanisms of communication (i.e., the mechanisms of transcription and translation that will be detailed in Section 2) have contributed to a deeper understanding of the basic principles that govern the cell. Until the end of the last century, life scientists have occumulated knowledge of specific pathways and modes of molecular activation. Several fundamental mechanisms were unveiled in these years, mainly by deepening the understanding of crucial cellular functions. However, toward the end of the last century, several types of concurrent interests and pressures revealed the growing need for a different approach: this impulse was oriented toward a more systemic understanding of the interactions among pathways rather than limited to the local investigation of a specific pathway. Some of the strongest motivations were the envisioning of potential applications of this still theoretical and qualitative information to areas of current practice such as clinical and pharmacological applications, environmental monitoring, as well as the study of the evolution of populations. All of this was boosted by concurrent technological advancements that enabled automated and high-throughput molecular analyses, namely, the development of high parallel platforms based on molecular micropatterning [Brown and Botstein 1999], the conceiving of the Polymerase Chain Reaction (PCR) [Mullis 2003], and the first automated sequencing techniques [Maxam and Gilbert 1977; Sanger and Coulson 1975].

The creation of the Human Genome Project Consortium in 1995 [Collins et al. 1998] that marks the beginning of the so called Genomic Era can be defined as the formal event that led to the convergence of all of the aforementioned pressures and motivations. Among the objectives of this Consortium were the identification of the human genes, the complete sequencing of the human genome, the storage of these data in publicly accessible databases, the improvement of tools for their analysis, the transfer of technologies to the private sector, and the attention to ethical, legal, and social issues (ELSI) arising from the project. It is worth noting that such an ambitious program could not be undertaken nor even imagined without the synergistic contribution of the competences from different scientific areas: biology, biophysics, physics, material science, computer science, statistics, medicine, and genetics.

The project is far from being completed, however, the first outstanding achievement, reached in 2003 (beginning of the post Genomic Era, [Collins et al. 2003]) with the first draft of the complete decoding of the human genome, carried much more than the mere yet enormous amount of information decoded. After this milestone, life sciences have undergone a major change often defined as a shift in perspective from the reductionist to the systemic approach [Hocquette 2005]. Actually, despite the fact that systems theory is a well established field, its application in life science had not been considered as a viable option to explain biological events beforehand. The genetic approach (single gene study) is currently being integrated into a more comprehensive genomic approach. This novel perception of biological process as a network of events more than the sum of independent pathways is leading to a more holistic view of diseases and is commonly referred to as the study of *omic* data [Quackenbush 2007]. The approach allows for the definition of more comprehensive molecular portraits of diseases, and, importantly, permits the perception of both commonalities and differences among molecular profiles of individuals. This will lead, through the translation of the newly acquired knowledge into medicine, to personalized approaches in clinical diagnostics and therapy by providing individual monitoring at the molecular level. In the coming decades, the consequences of this will bring fundamental changes in high impact areas of human activity such as medical research and practice. Given the high potential of these achievements, huge efforts are underway to continue to unveil the details of the molecular mechanisms and to develop robust, performing tools of molecular analysis.

Current tools and technology allow screening molecular activity at a genome-wide level for the different molecules involved in cellular information processing (omic data processing). DNA, RNA, and proteins are the fundamental actors of such a system. Through complex and nonlinear mechanisms they are, in fact, used by the cell to exchange information and accomplish functions that several endogenous or exogenous factors can impair. Current tools allow the highly parallel monitoring of a number of such incorrect functions involving different types of defects such as DNA mutation, RNA over/underexpression, disequilibria in protein patterns, and more. Among the many challenges presented, two fields of development are capturing remarkable interests and stirring relevant efforts: (i) omic data analysis (Section 4) and (ii) translation to medical practice via the implementation of point-of-care devices (Section 5). Briefly, omic data analyses imply analytical approaches to the data to infer knowledge from genome-wide molecular screenings. In contrast, the implementation of point-of-care devices requires the synthesis of established knowledge (obtained from the reductionist approach) with systemic-inferred information (from genome-wide screens). This synthesis aims at selecting a restricted number of highly specific molecules to perform efficient and complete molecular monitoring.

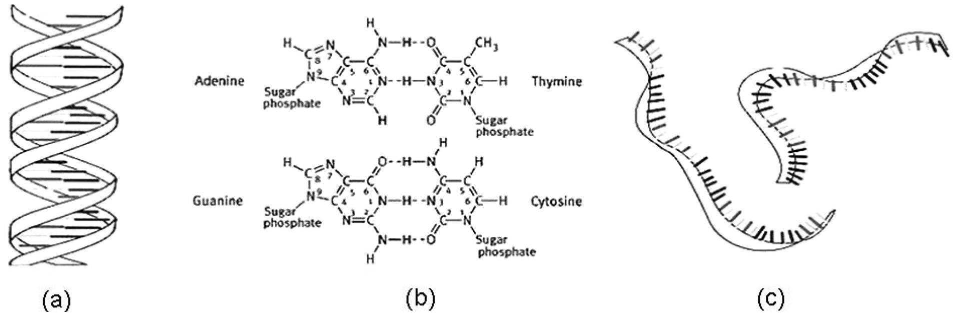

Fig. 1. On the left-hand side (a), the double helix form of the DNA filament and, on the right-hand side (c), two separate fragments are shown. In the middle (b), the hydrogen chemical bonds that gave rise to the DNA skeleton and to the stability of the double helix are depicted. Guanine and Cytosine are bound by three hydrogen bonds, while Thymine and Adenine only by two. This confers specificity to the base-pair recognition process.

In the following, we will discuss these two related areas of development, however, the reader should be aware of the fact that the potential of this revolution extends beyond these areas to a number of emerging fields such as synthetic biology, environmental monitoring, and in vivo drug delivery.

This work is organized as follows. Section 2 will provide the reader with the fundamentals of molecular biology, Section 3 will describe in more detail the specific platforms for omic data analysis and their application, focusing on the underlying technology. Section 4 will give an overview of the techniques adopted to process omic data, their evolution, and the type of applications at which they are aimed. Section 5 will present the state-of-the-art of the techniques, devices, and chips developed for low-cost, portable clinical tests.

2. FUNDAMENTAL CONCEPTS IN MOLECULAR BIOLOGY

The fundamental dogma of molecular biology represents a simplistic yet meaningful way to explain the basic flux of information in the cell (from DNA encoded information to cellular functions). This dogma involves the description of 3 types of molecules DNA, RNA, and proteins and their complex interactions. DNA is a long molecule of nucleic acids made of 4 basic building blocks: Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). A DNA filament is a strand of virtually any combination of variable length of such bases. The DNA molecule consists of two such filaments coiled together in a double helix shape. The double filament structure is made possible by the fundamental property of *complementarity* of the 4 bases to form only two possible couplings: A-T and C-G. Thus, each filament represent the complementary blueprint of the other (see Figure 1).

The DNA of a cell is said to *encode* its genome, that is, the human genome consists of about 3 billion base pairs (*bp*) that are grouped and compacted in 23 coiled structures called *chromosomes*. Other organisms have different length and different numbers of chromosomes[1]. Every cell uses different portions of the information encoded in its DNA according to its role, its environment, its life cycle phase, and its need for communication with other cells in the environment (signaling pathways). The information extraction, decoded on demand from DNA, involves several intermediate molecules. The second molecule (after DNA) of interest in this pipeline is RNA, another nucleic acid polymer whose constitutive bases are A,C,G, and Uracil (U) instead of Thymine (T) (both U and T are complementary to A). RNA is synthesized from DNA by a copy process. Briefly, the 2 DNA filaments are separated (*denatured*), and one of the two filaments is used as the blueprint to build, one base at a time, the complementary messenger RNA (mRNA) single filament (transcription process). Interestingly, not all the DNA segments are transcribed into RNA, only the so-called *exons* that are usually separated by untranscribed sequences called *introns*. The transcription process is mediated by a specific family of enzymes: the RNA polymerases. mRNA is the vehicle in the synthesis of proteins that represents the principal actors of the molecular activity. Thanks to other types of RNA (rRNA, tRNA), the mRNA strand is converted into a sequence of aminoacids, the proteins' building blocks. This conversion (translation process) is mediated by the *ribosome* (an organelle in the cell) which has a double structure that binds on one side to a specific triplet of bases (coded in sequence on the mRNA strand) and that is associated on the other side to one of the 20 existing aminoacids. Since there are 4^3 possible triplets of the 4 existing bases, each amino acid can be obtained by more than one triplet of base combinations this redundancy is a reason for the robustness of the translation process. A sequence of aminoacids is not yet a functional protein, in fact a protein becomes fully functional when its 3-dimensional structure is achieved. This represents another sequence of steps: the folding step (secondary and tertiary structure) and sometimes the assembly with other proteins to form a complex (quaternary structure).

The generation of a fully functional protein can be interrupted in any of the described stages; in fact, specific molecules control the perfect functionality of each step. Chaperone proteins are devoted to check the correct folding and assembly, while short RNA (shRNA) can bind to the mRNA and lead to its destruction before it reaches the ribosome [Pederson 2004]. The production of all the interacting molecules can be inhibited by exogenous (such as nutrients, pollutants and so on) and also endogenous factors (small molecules or proteins produced by the cell). This results in an extremely sophisticated process of (auto) regulation of the cell (Figure 2). For example, the expression of a gene (i.e. the transcription of its DNA in RNA) is regulated by a promoter which is a protein produced by another gene in the same cell. For more details the reader can refer to Alberts et al. [1989].

[1] Overall, cells, of more recent evolution have their chromosomes contained into a specific structure of the cell the nucleus. These cells are called *eukaryotes*. Simpler organisms are not provided with a nucleus and have the genetic material directly inside the cell (cytoplasm). These are called prokaryotes.

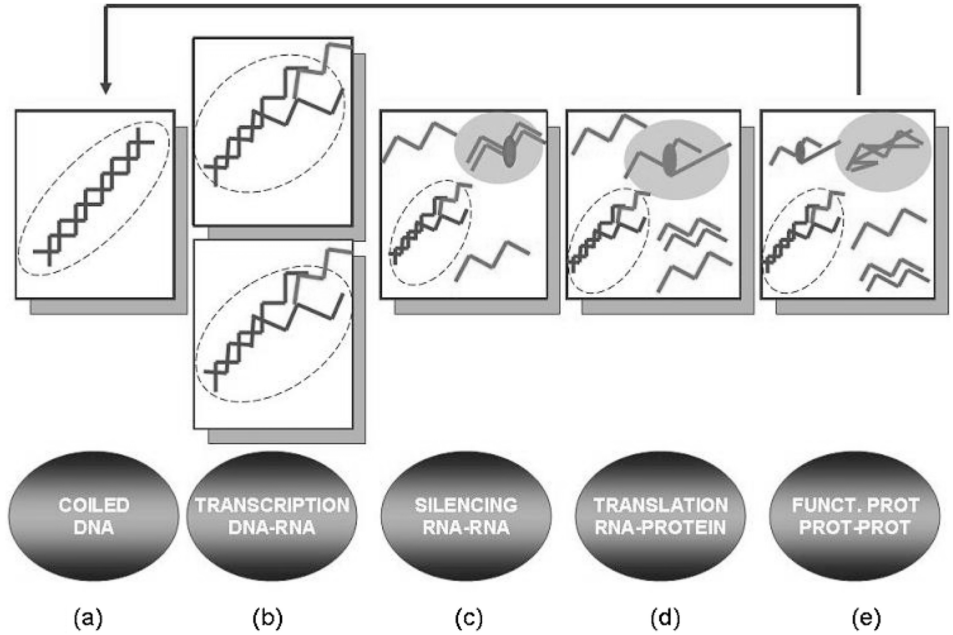

Fig. 2. Expanded graphical version of the central dogma of molecular biology with the representation of some of the regulatory processes that occur in the cell. From left-to-right-coiled DNA (coiled, double-stranded molecule) is unfolded and transcribed into RNA (single-stranded molecule). Depending on the exact sequence of the encoding strands of DNA (exons) that are transcribed, the same gene can give rise to different proteins (variants of the gene). This fact is depicted with the two boxes in (b). Once the mRNA molecule exits the nucleus (nucleus membrane represented by a dotted line), it can be destroyed (degraded) by a complex mechanism involving several molecules, highlighted in gray in (c) (here oversimplified with a complex molecule of oval shape, representing the RISC complex described in more details in Section 3.1.3). If the RNA reaches the ribosome (highlighted in gray oval-shaped molecule), it can be translated into the corresponding sequence of aminoacids, (d). Form there, with the help of the chaperone-proteins, if the aminoacids sequence is perfect, it is then folded into the appropriate 3-dimensional structure of the protein, highlighted in gray in Panel (e). The transcription process of a gene continues until the appropriate signal is not sent back to the DNA through some specific inhibitory mechanism.

Currently, several of the molecular stages of this complex pipeline can be screened at the genome-wide level with suitable tools. The next section will focus on the technology and application of such techniques.

3. MOLECULAR RECOGNITION BASED ON AFFINITY REACTIONS

This section will give an overview on the most widespread high-throughput platforms existing for the molecular screening at genome-wide level (*omic screens*). Platforms presentations are organized into 3 sections following the type of interacting molecules. Interactions among nucleic acids (Section 3.1) include DNA-DNA, DNA-RNA, and RNA-RNA. These 3 types of interactions are grouped together because they exploit the same mechanism for recognition: *hybridization* that consists of the binding of complementary nucleotides from two different molecular sequences. However, as will be detailed later, the exploitation of hybridization leads to very different types of information, depending on the specific molecules under study. In particular, DNA-DNA interactions are used to study mutations, DNA-RNA to oversee gene expression, and RNA-RNA for the observation of a particular form of gene expression inhibition. Interactions among proteins (Section 3.2) use a different property of these molecules, namely, the ability for a given protein (antibody) to recognize another specific protein (antigen). Finally (Section 3.3), the interactions among DNA and proteins relie on a combination of the two described techniques. These studies have only recently been made possible with the use of high parallelism, which assists in the understanding of how particular proteins (promoters) interact with DNA to allow the expression of genes to start, reinforce, or interrupt some molecular functions. For each of the 3 interactions we define (Sections 3.1, 3.2, 3.3), we will also give a description of the chemical-mechanical interaction (Section *Interaction*, deepening the information given in Section 2), of the technology available (Section *Technology*, developed in Section 5), and of the platform used (Section *Use*, further described in Section 4). Section 3.1 describes the chips for which most experience has been accumulated since they represent the first attempt to design genome-wide screens. For this reason, this section is more detailed. In fact, since the first design and use of these chips, several problems have been approached and solved (for a survey, see Yoon [2006]).

As a final introductory note to this section, it is worth noting that high throughput sequential (versus parallel) technology exists that deals with DNA synthesis. Although this is not the focus of our article, these technologies are the base that allowed for the massive sequencing of whole genomes *that permitted* for the Human Genome Project. In fact, before the advent of innovative approaches (i.e., Pyrosequencing [Ronaghi 2001]) that can sequence roughly 100 megabases in a 7-hour run with a single machine, sequencing was an extremely expensive and time-consuming process [Sanger and Coulson 1975]. Basically, this method allows sequencing of a single strand of DNA, thanks to the synthesis of the complementary strand along it. The method exploits the fact that each time a nucleotide, A, C, G, or T is incorporated into the growing complementary chain, a cascade of events is triggered which ultimately produces a light signal detected by a charge coupled device (CCD) camera. Since only one type of nucleotides are put into the sequencer at a time and since each light signal is proportional to the number of nucleotides incorporated, a software approach can be used to recognize the number of incorporated nucleotide, complementary to the sequence read.

3.1 Nucleic Acids Interactions

3.1.1 *Interaction.* As mentioned in Section 2, two nucleic acids strands are able to reversibly bind or separate thanks to the highly-specific affinity between single complementary bases. This same principle holds for different processes, namely, the denaturation (separation) of two single strands of a DNA molecule, the folding between parts of the same RNA molecule, and the transcription of RNA from DNA by means of the RNA polymerase (the enzyme devoted to enabling of the transcription process). The binding or separation of

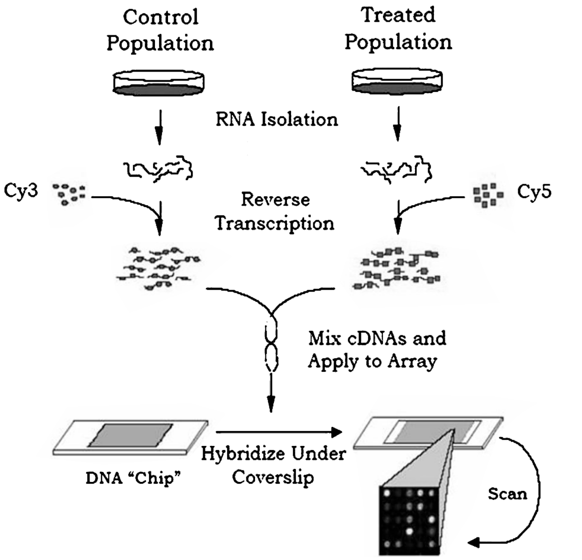

Fig. 3. A schematic representation of a comparative genetic analysis performed with double channel microarrays. Two different cell populations are taken into consideration, one for analysis (e.g., after treatment) and one for control. Form both cell populations, RNA is isolated, reversely transcribed, and labeled, adding two different dyes (in this case Cy3, Cy5). The mixture of both labeled RNAs is then spread on the array, where competitive hybridization can take place. A scan of the slide is then performed to read out fluorescence corresponding to different expression activity. Adapted from www.uwec.edu.

two sequences is regulated both by the sequence characteristics and by physicochemical parameters. Coupled forms of complementary molecules are always favored since their Gibbs free energy is higher in the bound form, and thus additional energy is required to go back to a denatured state. Moreover, molecules characteristics can enhance the ease of binding, in fact, longer complementary sequences as well as abundance of C-G couples lead to stronger bonds since this also increases the Gibbs free energy. However, it is always possible to open these bonds by increasing the temperature of the solution up to a point typical of each molecule. In fact, given the constancy of the solution characteristics, it is possible to identify the so called *melting temperature*, that is, the temperature at which half of the molecules in the solution have been separated. The melting temperature increases also with the concentration of salts of the solution, positive ions help to stabilize the bonds between nucleic acid bases which tend to repel each other because of their negative charge.

3.1.2 *Technology.* As previously mentioned, two strands of nucleic acids of different length coming from different species that bind together undergo hybridization. In this process, the complementary segment can be shorter than the hybridizing molecules. This phenomenon is exploited in arrayed analysis tools to test the level of complementarity between species immobilized on the array sites and species present in solution. The fragments diluted in the sample solution (bulk molecules) are preprocessed and coupled with a permanent marker molecule which, once the fragment has been captured by a complementary strand on the surface (here called surface molecules), is able to generate a localized detectable signal (see Figure 3 for a representation of the process).

The genetic material employed in microarray devices both as surface and bulk molecules is extracted from cells, suitably purified and fragmented. This represents genomic DNA, RNA, or cDNA (the molecule obtained using RNA as a blueprint, known as *retrotranscription*). In general, cDNA is preferred to RNA because it is more resistant to degradation and manipulation.

Because of the many variables and technological possibilities that influence and can be used to design high-throughput hybridization experiment arrays (commonly called microarrays), their classification can be performed in a number of ways depending, for example, on the type of molecules used, the number of sites, or the transducer required to scan the array as it described in the following.

When focusing on the type of molecules adopted, there are two fundamental types of arrays. In one type, both surface molecules and bulk molecules are fragments of cellular extracts and their length extends from 20 to hundreds of base pairs. These systems aim at comparing the expression of two cellular entities, one taken as reference and the other is the sample under test. For this reason, they are also called *double-channel microarrays*. A different microarray fabrication technique uses synthesized short probes (15–25bp) of known sequence as surface molecules which are able to capture specific fragments in solution; these arrays are called *single-channel*. Probes that differ in one or few bases are also synthesized and are used to test the presence of mismatches or single polymorphisms (see Section 3.1.3).

Surface molecules are organized in localized region called sites. Depending on the immobilization chemistry (covalent or by adsorption), the molecules length, and the sites area, the quantity of molecules per site varies by orders of magnitude. Thus, microarrays are also usually classified according to the number of sites on a single support and the type of probes spotted by the user (surface molecules) or synthesized by the company selling the microarray technology. According to the first option, users must have access to a spotting system that delivers drops of 5nl containing the surface molecules to be immobilized. Spotters can be used which work both with cDNA strands from 500bp to 5,000bp as well as with presynthesized probes. The number of sites ranges from 5,000 to 40,000 arranged on a microscope slide, and the diameter of the sites is usually between 150μm and 200μm. Alternatively, some companies offer microarrays with already immobilized probes. These solutions allow higher quality and density of the arrays. Affymetrix produces devices with a site density up to $6 \cdot 10^6$ sites on a surface of $1.28 \cdot 1.28\text{cm}^2$. Probes of 25bp are synthesized one base at a time with a photolithographic technique derived from the microelectronic process. As probes are particularly short compared to the total length of a gene, around 20 sites with different sequences are needed to specifically identify one gene. Consequently, one chip is almost able to perform whole genome tests (25,000 is the current upper limit of the estimated number of genes of the human genome). By inkjet printing of presynthesized molecules, Agilent is able to obtain almost the same results in terms of numbers of genes on a single support. In fact, the density is smaller (44,000) but the length of the probes is longer (60bp).

Another common feature used to classify microarray types is the transducer needed to scan the array as determined by the nature of the marker molecule. Fluorophores are generally used for this (labeled technique), nevertheless the disadvantages of this technology and new applications envisaged for gene-based tests have boosted the research in the direction of finding alternative (non-labeled) techniques. The drawbacks of fluorescence-based measurements are due to the fact that errors are introduced both in the fluorescence signal-reading process [Romualdi et al. 2003] and the labeling of the sample [Churchill 1953]. Moreover, the current setup needed for scanning fluorescent arrays is not compatible with the development of portable systems. For this reason, the integration of optoelectronic detection on existing fluorescence microarrays has been attempted but is still far from acheivable [Thrush et al. 2006a]. On the one hand, many efforts have been directed toward the development of label-free techniques. On the other hand, new types of labeling have been employed to implement more reliable and portable systems, namely electroactive molecules, chemiluminescent labels, metal, and semiconductor nanoparticles. The latter can be employed in systems for mass, electrochemical, and optical sensing. More details on portable systems will be presented in Section 5.

The ability of an array to capture only complementary molecules is called *specificity*. Sometimes nonspecificity is caused by the affinity between bulk molecules and the array surface. This can be reduced by a suitable passivation chemistry which, on the one hand, should not repel molecules from the probes. On the other hand, the bonds between noncomplementary molecules can be limited by applying stringent condition to the system at the moment of the hybridization (i.e., increasing temperature, lowering ionic strength, and so on).

Sensitivity is practically defined as the minimum amount of bulk molecules that can be reproducibly detected. This parameter can be expressed as the minimum detectable number of molecules on one array site or as the concentration of the bulk molecules in the sample solution for the hybridization. These two ways of expressing sensitivity are inherently different, in fact, given the same sample concentration, the final number of molecules hybridized on the sites depends on the hybridization time, the sample volume, and the site area (this last can differ also by orders of magnitude). In most systems, sensitivity is enhanced by previously amplifying the fragments of cellular extract under test. Before 1999, when amplification was not yet an established technique, 200μg of genetic material were needed to perform a microarray analysis. Now, thanks to amplification, 0.25μg (i.e. 10^4 cells) are sufficient.

From these considerations, it follows that different experimental conditions have a large impact on the performance of the tool, and this is one of the reason why a comparison between microarrays results is not trivial. To face this problem, some years ago a study on the discordances between commercial platforms for gene expression employed with the same RNA samples [Tan et al. 2003] gave birth to the MicroArray Quality Control Project (MAQC), led by the U.S. Food and Drug Administration (FDA) which was intended to enhance homogeneity and to assess reproducibility of microarray results. From this study, it emerged that the identification (and quantification) of a gene in a sample is not trivial for several reasons. First, the mRNA transcribed from a gene can be composed by different combinations of its exons. Thus, for the same gene (portion of DNA), several mRNAs are produced (variants). Microarray providers define different sets of probes that can capture a partial number of these variants since a complete list is not yet available. For this reason, different platforms are able to detect different variants, all named after the same gene. Moreover, since probes are short sequences that represent only portions of a gene, they have nonnegligible probability to match parts of unrelated genes. Finally, probes used in commercial platform sometimes differ slightly from the sequences reported in the RefSeq public database (that provides a stable reference for genome annotation [Pruitt et al. 2007]), considered the gold standard for gene sequences. Companies are focusing many efforts on smoothing discrepancies because the reproducibility of gene expression results is a crucial issue in assesing the applicability of microarray technology in clinical diagnostics. Nevertheless, as it remains difficult to suppress variations in the quantity of gene products captured when working with different probes, a different approach has to be accepted for result comparison. In fact, the homogeneity of results should not be considered at the level of a single genes activities but rather at a higher level of cellular activity in which the genes resulting from different analyses are involved. For examples, if two platforms give rise to two different sets of genes, only partially overlapping but all involved in the same cell signaling process or other pathway, then the results should and will be considered coherent.

3.1.3 *Use.* Nucleic acid interactions are used to investigate a large number of biological questions related to the molecule's function. In the following, we will describe mRNA, DNA, and *d*ouble *s*trand RNA (dsRNA) genome-wide studies. The latter represents a recently discovered type of molecule, whose activity can impede gene translation into protein after transcription.

The most widespread analyses concerns the genome-wide screening of the transcriptome of the cell, that is, the total amount of mRNA transcribed in a cell [Brown and Botstein 1999]. This study is performed with RNA-cDNA arrays also called *microarrays for gene expression* and gives a snapshot of the molecular transcription activity in the cell at the time of the experiment. In fact, mRNA is produced according to the needs of the cell and, because of this characteristic, microarrays for gene expression have been widely used to screen cancers and assess if and how it is possible to recognize some taxonomy at the molecular level not visible at the macroscopic level. This is useful for better targeted therapy delivery in diseases characterized by strong heterogeneity.

DNA screens (instead of RNA-cDNA screens) can also be performed. These tests have a different goal, namely, they investigate DNA mutations, implying deletion or multiplication of DNA strands. DNA can undergo mutations of various types that can be classified based on the mutation's size. The study of an entire chromosomal duplication is called *cytogenetics*, and it is performed on chromosomes during *metaphase*, a step in the cell cycle where chromosomes are easier to observe. The genome-wide screening of smaller deletions can be studied by means of *comparative genomic hybridization* (CGH) arrays, [Shaffer and Bejjani 2006]. The same technique used in cDNA arrays is used but both the surface and bulk molecules correspond to DNA strands instead of mRNA (or cDNA) ones. Punctual mutations in the DNA are called Single Nucleotide Polymorphisms (SNP). These represent variations of one single nucleotide in the known sequence of DNA of a given species. The mapping of all the human SNPs is the objective of the Hap Map Project [Thorisson et al. 2005] and represents a novel frontier in the study of individual genomes (*genotyping*) with relevant interest in the study of hereditary diseases. In fact, groups of SNPs that are present in tight association can be used as markers to recognize specific (possibly mutated) portions of DNA. Before the advent of SNP studies, the markers adopted were less specific. Therefore, the use of SNP as markers for genotyping can make the process cheaper and more efficient.

Finally, there is another type of RNA interaction, whose discovery is very recent [Fire et al. 1998] called RNA interference (RNAi). This mechanism relies on a double-stranded RNA (dsRNA), which after several steps of processing, can hybridize to the mRNA and drive it to destruction. This mechanism is thought to originate from the necessity to protect the cell against viruses whose genetic material can be blocked and destroyed before it infects and eventually kills the hosting cell. During evolution, this ability developed further and is now capable of blocking translation of endogenous RNA. The process requires the presence of the dicer enzyme, which is able to cut the dsRNA into shorter fragments (20–25bp), called *s*mall *i*nterfering RNA (siRNA) and microRNA (miRNA) when exogenous and endogenous RNAs, respectively, are involved. One of the two strands (called the guide strand) hybridizes to another complex called the RNA-Induced Silencing Complex (RISC). The silencing (inhibition) of a transcribed gene occurs when its corresponding mRNA strand hybridizes to the guide strand. The mRNA is finally degraded by the RISC complex. This mechanism is used in research to silence genes in the posttranslational phase and allows us to observe the effects of a gene's absence on the cell activity. Silencing by RNA interaction is preferred to previous techniques since it is simpler than the *knock-out* of a gene. Knock-out requires the genetic engineering of the organism that carries the gene of interest so that it is made inoperative. In research, siRNAs have to be designed, and usually they are made of a sequence complementary to the gene of interest. These siRNA are then introduced into the cell, exploiting *transfection*, a process that enables the membrane of eukaryotic cells to uptake molecules of various sizes. Thus, several siRNA are synthesized and spotted on an array; cells ready for transfection are laid on the spots in conditions that keep them alive and observed to assess if the mRNA of interest could finally be translated into a protein or was destroyed before translation. Since this process eliminates a large part, but not all of the transcript of a given gene, it is called *knock-down* (to distinguish it from knock-out). A whole new area of research is interested in the identification of genes that code for such RNA. Since the mechanisms of action are not completely known, prediction and inference by mean of computational techniques are important for predicting such genes. For a review see Yoon and de Micheli [2006].

3.2 Protein-Protein Interactions

3.2.1 *Interaction.* Proteomics (the study of proteins at genome-wide scale) represents a very complex area of research due to the large existing number of proteins (the estimate for the human proteome is 500,000 proteins), the complexity of the molecules' structure, and the variety of interactions they can have. Proteins, in fact, can be described in terms of their aminoacids sequence (primary structure), their folding into basic 3-dimensional structures, such as alpha helices and beta sheets (secondary structure), their complex folding built on the secondary structure (tertiary structure), and their association with other tertiary structures (quaternary structure). Proteins can build complexes by combining several units of the same protein (such as hemoglobin, made of 4 identical subunits) or by combining different proteins.

A very specific and relevant case of protein-protein interaction is given by the complex antibody-antigen. *Antibodies* are proteins used by the immune system to recognize specific molecules potentially dangerous for the organism, called *antigens*. Given the specificity of the coupling mechanisms, antibodies are often used as a reference to recognize the corresponding antigen protein. Because these interactions are based both on chemical and mechanical affinity, high throughput screens are complex to obtain since they should preserve the 3-dimensional structure of the molecule [Zhu and M.Snyder 2003].

3.2.2 *Technology*. The design of protein arrays is characterized by much more stringent requirements than DNA or RNA microarrays. This is due to several reasons connected to the importance of their 3-dimensional structure. First, proteins interact through a specific active site which has to be accessible when the protein is immobilized on a surface. This can be achieved by placing a suitable linker termination on an inactive side of the protein. The linker must be chosen in such a way that it provides a stable immobilization. As can be the case for nucleic acids fixing, the linker can provide a covalent bond with the surface or be a part of an affinity couple (e.g., biotin-streptavidin). Second, the sensing properties of proteins strongly depend on environmental conditions. Thus, to be able to characterize reliably the interaction between two proteins on a surface, it is necessary to reproduce specific bulk chemistry and physics. This implies the use of specific reagents and solutions that should not interfere with the immobilization chemistry or with the detection technique. Finally, the most stringent requirements for protein activity are proper moisturizing and sufficient steric mobility. Surface adsorption on hydrophobic plastic-like polystyrene—employed in more traditional protein analysis tools—allow for good protein activity [Silzel et al. 1998]. However, this impedes high-density patterning and is responsible for the important fluorescence background signal.

To better capture all the properties of proteins, other approaches that try to mimic bulk conditions have been investigated. Among these, classical agarose gel depositions [Afanassiev 2000] and nanowells [Zhu 2000] have been tested with encouraging results since they help to preserve the 3-dimensional structure of the molecules. However, both solutions have some drawbacks: agarose does not allow for the easy exchange of buffer solutions needed in multistep analysis, and nanowell technology is not compatible with commercial arrayers. Besides the study of chemomechanical interactions events several efforts have been made to shed light on the dynamics of these events studied with high parallelism. Some approaches, namely fluorescence and chemiluminescence, have been simply derived from standard microarrays for nucleic acid interactions (Section 3.1.2).

Among the more advanced detection techniques, surface plasmon resonance has proven to be particularly suitable in the real-time study of binding and separation kinetics of protein-protein interactions [Georgiadis et al. 2000]. Surface plasmon resonance is a phenomenon that consists of the propagation of electromagnetic energy along a metal surface that is due to the absorption of light. This propagation occurs at specific light energy, which depends on the refractive index at the metal solution interface. The interface changes its refractive index when biomolecules that are in the solution react on the metal. This technique, which requires immobilization chemistry specific to gold, has the advantage that it is label-free, but still lacks high-parallelism implementations. Single-molecule techniques have also been tested, and approaches based on Atomic Force Microscopy (AFM) are able to detect relevant morphological details of the interaction [Jones et al. 1998; Hinterdorfer et al. 1996].

3.2.3 *Use.* Depending on the application, protein microarrays can be classified into two categories: (i) functional microarrays to detect biochemical activity or understand drugs mechanisms of action and (ii) analytical microarrays used to detect the presence of specific proteins in unlabeled mixtures, exploiting their antibody-antigen properties. Because proteins are the final actors of cellular functions, understanding their distribution and abundance is relevant to shed light on all the molecular activities ongoing in the cell. These screening tools are the most advanced candidates in the point-of-care applications, described in Section 5. In fact, since several classes of diseases are known to produce abundant quantities of specific proteins (often called *markers* in the affected organisms), the detection of such markers can be extremely useful to perform fast screening of a large set of diseases.

3.3 DNA-Protein Interactions

3.3.1 *Interaction.* Interplays between DNA and proteins represent a crucial step in the regulation of cell activity. The structure of a gene includes not only the set of sequences that are actually transcribed (exons) and other noncoding regions (introns) but also the promoter region. This region is recognized by proteins known as transcription factors (TF) that bind to the promoter sequences and recruit the RNA polymerase, the enzyme that allows the actual synthesis of mRNA. Promoter regions can also be assisted in the transcription process by other DNA sequences called enhancers or silencers. These are regions where proteins that enable or disable the actual transcription of the DNA can bind, facilitating or impeding the transcription process. These proteins can be exogenous or endogenous, and subtle controls of their production are the key of the regulation of the cell mechanisms. The study of promoters and transcription factors is in close interaction with *histones*, the most common proteins that envelope the chromosomes and constitute the *chromatin*. Histones modifications are necessary for promoters, transcription factors and RNA polymerase to access the coiled DNA and start transcription. Some of these modifications can also be inherited, somehow invalidating the idea that all hereditary information is contained in the DNA. *Epigenomics* is the emerging field studies these nongenomic data (for more information, see the Human Epigenome Project [Esteller 2006]).

3.3.2 *Technology.* A recently developed technology has been proposed to study the DNA-protein interaction. Since the relation promoter gene can be multiple, this technique aims at associating promoters to all the target genes of a given promoter (protein-DNA interaction) by associating a precipitation technique with microarray technology (see Figure 4 for a depicted description of the process). First, the precipitation technique, Chromatin ImmunoPrecipitation (ChIP), is applied. Specifically, DNA and promoter proteins are allowed to crosslink in vivo under suitable conditions to form a promoter-DNA complex. Then, antigens corresponding to the promoter proteins and associated to agarose beads are added to the mixture to capture the appropriate promoters-DNA complexes. Beads are precipitated by centrifugation along with the complexes bead-promoter-DNA that are then separated from all other molecules. DNA is finally extracted and fragmented by sonication (use of high frequency sound waves to disrupt DNA molecules) in strands of 0.2–2kb and fluorescently labeled. The technique is then coupled with a microarray approach (Section 3.1): the device is exposed to a mixture of the input DNA labeled with a different fluorophores and of the DNA fragments obtained by immunoprecipitation. This technique reveals the binding position of specific promoters on the genome. Globally this screening tool is often referred to as ChIP-chip. Due to the quantity of material produced with each ChIP (10-100 nanograms), 50 experiments are needed to perform a microarray analysis.

3.3.3 *Use.* These platforms are used to infer the relationship between DNA and proteins, which represents a fundamental step of genomic regulation. For these reasons, ChIP-chips have gained interest in recent years given the general enhanced attention toward the complex network of genomic interaction. These tools are used to screen and identify binding targets or protein factors without prior knowledge since, for many genes, the transcription factor is unknown. This information is relevant to possibly controlling the transcription or the silencing of a gene of interest, taking into account all the complex network of interactions that influence gene activity. These tools are also used to map protein binding locations for a genome-wide binding profile and to identify the promoters distribution and histones modifications. Notably, these tools are used to provide (intermediate) validation of algorithmic modeling of networking interactions, described in Section 4.2.

4. PROCESSING OMIC DATA

Given the portrait just depicted, the introduction of informatics in the realm of biology appears to be an obvious step. In fact, the automation of analyses has become a crucial requirement given the large number of data (points) researchers have come to deal with as a result of genome-wide screens. Notably, bioinformatics has become relevant to two main biological interests: the study of protein structures and the analysis of genome-wide expression data. Both

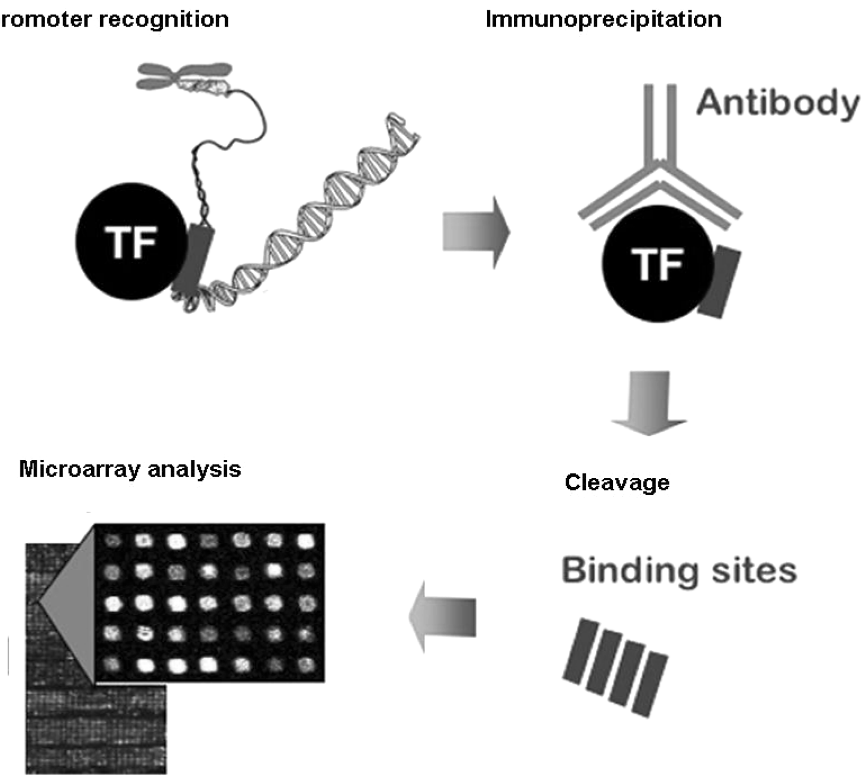

Fig. 4. Schematic view of the ChIP-chip technology. From top left and clockwise: crosslink between the DNA and the transcription factor (TF) to form a complex TF-DNA; addition of the antibody corresponding to the transcription factor to form the complex Antibody-TF-DNA; isolation of the DNA strands of interests (promoter regions) fragmentation by cleavage from the complex Antibody-TF-DNA; spreading of the selected fragments on a microarray slide to identify them. Adapted from www.chiponchip.org.

areas require handling an enormous number of variables which cannot be done without automated and powerful computing systems.

It is worth noting that the power and potential that lies behind the introduction of informatics into the field of biology is not the mere ability to treat a larger amount of data than what could be managed manually. It is more the possibility to model, then simulate, and thus predict molecular biological events. There has been a lot of controversy on this theme among computer and life scientists concerning the feasibility and robustness of such predictive ability. However, it now appears clearly that computer and life sciences cannot progress at this time and in this area without collaborating synergistically. Biological information is not complete and standardized enough to be treated as any datapoint, and molecular data have a limited potential to deliver innovative information unless they are studied and approached with automated and powerful tools.

This debate coagulated into the definition of in silico biology. As opposed to *wet* biology performed in a traditional molecular biology laboratory, in silico biology offers predictions of molecular behaviors simulated on computers. Once the model is understood, this can save large amounts of time and money, reducing the number of experiments that need to be performed for final validation. As highlighted previously, the construction of the model requires experimental data, and it is not (yet) possible to predict or understand a complete molecular behavior merely using simulation. However, in silico results can be extremely helpful in the advancement of molecular biology-related sciences as will be discussed in Section 4.2 on the potential of systems biology (Section 4.2) or in specific areas of research, oriented, for example, to the identification of genes involved in the development of complex hereditary diseases.

The study of complex hereditary diseases is a fundamental area of research that aims at the identification of mutated genes responsible for diseases transmitted from parents to offsprings. Traditionally, this required the identification and genotyping of large numbers of affected patients' genomes. This process is costly, and when studying rare syndromes, it can be very difficult to collect a statistically meaningful number of samples. Approaches typical of in silico biology take advantage of all the information already publicly available to infer association among genes with no additional experimental expenses. We developed one such tool (TOM, [Rossi et al. 2006; Masotti et al. 2007]), based on the information extracted from gene expression arrays to infer, after several statistical filtering steps and the addition of information from external databases, the genes that are more likely to be associated and localized in a genomic area known to be relevant for the malady. Several such approaches exist (as a sample, Tiffin et al. [2005], Turner et al. [2003], Franke et al. [2006]), based on different principles of association (literature search, coexpression, common evolutionary origin etc.) and aiming at the identification of the candidate genes list. This list represents an extremely valuable input for geneticists to further explore the molecular origin of a disease.

In the following, we will discuss the effects of the revolution introduced by the advent of bioinformatics in the Genomic (1995–2003) and post-Genomic Era (2003 on) in the analysis of omic data, namely, what types of analysis are and were adopted and with which aim, how they evolved, and how they impacted the scientific areas of life and exact sciences. This analysis will try to follow a time-based evolution. It will move from the first high throughput analyses that microarrays for gene expression made available (Section 4.1). All the steps of this evolution involved the solution of issues related to different branches of science, from the algorithmic treatment of the data, to statistics, to biology and medicine. We will then move to the integrated and interdisciplinary view typical of systems biology (Section 4.2).

Globally, this overview is meant to explicitly show how and why the results of the analyses based on the platforms depicted in (Section 3) are attracting so much effort and interest, and what expectations are during these efforts.

4.1 Bioinformatics

All the major players in the Genomic and early post-Genomic Era were committed to face new problems related to the large amount of data involved. This implied the development of relevant efforts in several areas of science, all leading to solutions tailored to answer unaddressed questions arising from the availability of a large number of molecular biology experiments obtained in parallel. Namely, problems and solutions were represented by (i) algorithmic efforts mainly oriented to clustering (although other approaches toward molecular similarity measurement for recognition and phylogenetic approaches have been developed [Lipman and Pearson 1985; Altschul et al. 1997; Cameron et al. 2004]) from the computer science point of view (Section 4.1.1); (ii) biological characterization efforts for molecular signatures extraction from the life science point of view (Section 4.1.2); (iii) statistical efforts for searching solutions to the curse of dimensionality problem (Section 4.1.3); and (iv) a rethinking of the medical perspective, leading to the identification of new paradigms in medicine (personalized and evidence-based medicine) (Section 4.1.4).

With a different focus, these topics all represent the necessity to face the manipulation of an abundant mass of data with poor (there are still many unknown genes) or fragmented characterization (some molecular pathways are extremely well known). Interestingly, these four approaches have not moved in close synergy but have rather built parallel paths of solutions. This represents an emblematic example of the novel types of problems involving interdisciplinary approaches, an issue that has been typical of the early post-Genomic Era. In fact, biology requires as a crucial precondition the tight collaboration among researchers from different backgrounds. However, until the advent of systems biology, this has remained an important, mostly unaddressed issue. Notwithstanding this limitation, this early post-Genomic phase has shed partial light on difficult molecular problems, notably offering the possibility to provide some taxonomy in cancers whose prominent characteristic and lethal strength is heterogeneity.

4.1.1 *The Algorithmic Effort.* The algorithmic field has been massively involved in the study of molecular data thanks to the introduction of high-throughput devices and has invested large (although nonexclusive) resources efforts in the optimization of clustering approaches. Clustering, in fact, defines the classification of items into different groups (clusters), so that the data in each group ideally share some common features that differentiate them from the other clusters. Often these features are defined by proximity according to some defined distance measure. This aspect is extremely important to provide some organization and structure in the large amount of data provided by high-throughput biology. In this field, microarray results are viewed as textual tables with rows representing genes and columns representing patients (or samples); rows are called gene expression profiles and columns are called sample molecular profiles. The hypothesis underlying the development of this area is the assumption that genes sharing similar profiles of expression across a large number of samples are likely to be part of the same molecular function (also called module). Similarly, samples' molecular profiles that share consistent similarity are likely to present similar aspects (phenotypes). For this reason, algorithms have been imported from the data mining area or developed ad hoc to cluster together either a gene's vectors (expression profiles or rows of the table) or sample's vectors (samples molecular profiles, or columns of the

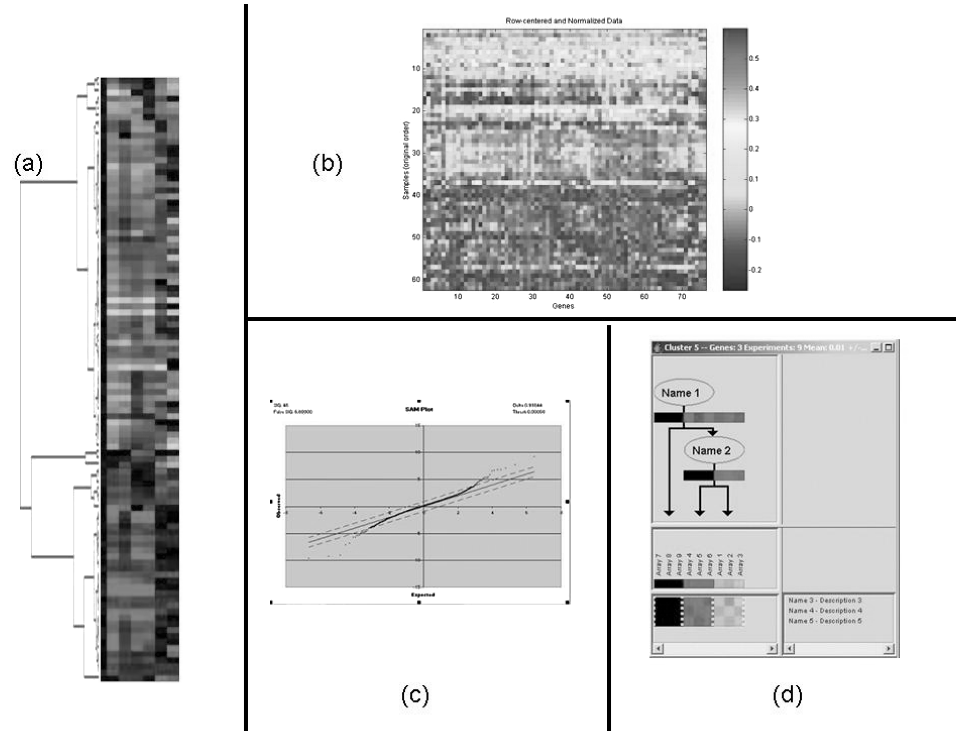

Fig. 5. Depicts examples of graphical outputs of clustering algorithms. Figure 5(a) shows an example of hierarchical clustering applied to microarray gene expression profiles. On the left-hand side, the hierarchical tree indicates the proximity of various genes (generated with cluster form sample given in, rana.lbl.gov/EisenSoftware.htm). Figure 5(b) depicts the graphical output of the CTWC algorithm genes and samples are grouped together simultaneously (with permission from ctwc.weizmann.ac.il). Figure 5(c) shows the graphical output of SAM (generated with SAM from the sample two class unpaired data given as example in the software source http://www-stat.stanford.edu/t̃ibs/SAM/). Figure 5(d) shows the graphical output of the network module approach; nodes represent the value of some molecules of interest (called NameX), and the branches of the tree lead to the coherent over/under expression of a number of genes. Depending on the presence or absence of the NameX molecule, the cascade of decisions that lead to the leaves allow the definitions of sets of genes and samples (here called arrays) that share some coherence (with permission from genomica.weizmann.ac.il).

table). The most widespread of these approaches are hierarchical clustering, k-mean clustering, Principal Component Decomposition Analysis (PCA) and Self-Organizing Maps (SOM) [Quackenbush 2001]. Mostly, these approaches relay on the evaluation of some distance among genes (or samples) profiles or distance to some reference profile. The distance is considered meaningful when a threshold below a given value defines the cluster. These approaches obtained great success among medical and biological researchers, mainly due to the fact they are very intuitive to visualize (notably hierarchical clustering [Eisen et al. 1998], Figure 5(a)). All early biological discovery relies on these methods (as a sample see Alizadeh et al. [2000]; Lapointe et al. [2004]; Ramaswamy et al. [2003]) that carry, however, an heavy burden of limitation. In fact, biological reality is more complex than what the hierarchical clustering assumption states since the algorithms described previously can only assign a gene to a single cluster, while most genes are known to participate in several functions.

For these reasons, another class of algorithms was developed, the so-called bi-clustering algorithms [Madeira and Oliveira 2004]. Based on different techniques, they are able to cluster contemporary genes and samples and can create overlapping clusters. In contrast to the classical clustering algorithms, the metric of a bi-cluster is defined in two dimensions. A *bicluster* is a subset of the rows and the columns of the data matrix such that the distance between each element and the average values across the element considered in the cluster are contained within a given threshold value. We participated in the development of one such algorithm [Yoon et al. 2005] for an approach based on the *pCluster* algorithm [Wang et al. 2002], a biclustering method with the ability to recognize similarity among genes even if expressed only in a subset of experiments. Our algorithm represents an innovation with respect to other biclustering algorithms like the work of Cheng and Church [2000] whose δ-biclusters definition could not hold for subsets of the δ-biclusters. This fact is counterintuitive and can become a serious drawback especially since these algorithms are designed to be used to support further biological validation required to assess any type of results related to functional biology. This algorithm is also innovative with respect to the Coupled Two-Way Clustering (CTWC) by Getz et al. ([2000], see Figure 5(b)), another popular biclustering algorithm, that requires the tuning of sometimes nonintuitive parameters. Since clustering is only one of several complex steps in the complete analysis of a microarray dataset, a complex tuning can represent an obstacle in the choice of the clustering algorithm.

The enforced ability to handle a large amount of data suggested the introduction in the analysis of nongenomic data. In fact, information related to other biological markers can be highly informative. Because this approach adds a third dimension in the clustering effort (genes, samples, nongenomic data), it is sometimes referred to as triclustering. Several approaches have been devised for this, based either on statistical information or built on the biclustering method. The most widespread methods rely on statistical information and are known as significance analysis of microarrays (SAM [Tusher et al. 2001], see Figure 5(c)) and gene expression enrichment analysis (GSEA [Mootha et al. 2003]). Both extract a list of genes that are more likely to be involved in the process described in the third dimension. In contrast, the coclustering method [Yoon et al. 2005] is an extension of a biclustering approach, and first evaluates an association matrix constructed from the genomic and nongenomic data using the statistics defined in SAM for measuring the association. Finally, another type of approach was devised by Segal et al. [2003] who defined the module network procedure (see Figure 5(d)). This approach represents a good balance between the rigor of the algorithmic method design and the quantity and quality of experimental information adopted, and probably represents one of the first examples of interdisciplinary efforts in this direction. In fact, the approach is based on probabilistic graphical models and uses information from a variety of omic data from microarrays for gene expression experiments to databases collecting protein information and from extended literature searches. These diverse sources of information are used to infer *regulatory modules*, defined by three entities: a set of regulatory genes (regulatory program) that specify the behavior of coregulated genes (regulated module) under given conditions (context). The 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 further walking the tree. The leaf of the tree is represented by a set of genes whose expression is coherent and influenced by all the regulators/nodes along the path. This approach was recently used [Segal et al. 2007] with radiological data as regulators, leading to the extraction in the final nodes of the tree of sets of genes responsible or strongly related to specific radiological aspects. Such an approach applied to liver cancer represents an important proof of concept of the possibility to use noninvasive common diagnostic tools (Computed Tomography scans, CT) to infer important hidden molecular information for better diagnosis and therapy.

4.1.2 *The Biological Effort.* While the cluster represents the atomic unit of interest in the algorithmic area, the biological approach is interested in the so-called *molecular signatures*. These are sets of genes, obtained with some clustering approach, but that also have another important property: it is possible to characterize the genes' set role and function and associate it to some specific clinical outcome. The most common signatures are related to survival expectancy and represent sets of genes whose consistent activation or silencing is an important marker of the different survival times of individuals affected by a given disease (e.g., Lapointe et al. [2004], Bullinger et al. [2007]).

The necessity to simplify the process of understanding the causes that make such sets of genes so crucial for survival was fostered and, in turn, boosted, by the development of a very interesting tool, the gene ontologies. Briefly, *ontologies* represent an organized set of definitions applied to a homogeneous set of items. In this area, the Gene Ontology Consortium represents a success story in the standardization of genes identification. In particular, gene ontology (GO [Consortium 2001]) is a controlled vocabulary based on a graph structure that allows not only the classification of items of a given type but also, and importantly, the quantitative evaluation of their semantic proximity. It is made of three subontologies, each one defining a different aspect that characterizes a gene: its molecular function, the cellular components active in it, and the biological process in which it is involved. Each known gene from several organisms is annotated in each one of these ontologies with a unique identifier. The level of reliability of the annotation is stored as information in the ontology, and several tools have been provided both by the GO Consortium as well as by other independent groups to mine gene sets, making use of GO ([Al-Shahrour et al. 2004; J.L.Sevilla et al. 2005; Lee et al. 2005]). The most common use of GO is the characterization of a gene set through the *enrichment analysis*, a statistical approach designed to infer the classification (annotation) of uncharacterized items in terms of known ones based on the number of occurrences of a given subtype of identifiers in the observed gene set. 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 for validation purposes to observe if the genes share in large part the same function or, for discovery purposes, to extend the annotation of the most significant molecular role present in the gene set to the unannotated genes [J.L.Sevilla et al. 2005].

4.1.3 *The Statistical Effort.* When dealing with omic data, the statistical approach also challenges researchers with novel problems. In fact, the very design of a microarray study is different and opposite to previous classical studies. Clinical studies, oriented to the identification of recurrences and commonalities among patients to infer and influence specific mechanisms, are normally built on a large number of patients and a reduced number of monitored variables. Microarray studies are designed in the opposite manner since, given their cost, the number of samples is often of a few tens, while the number of tests performed (gene expression) is of several thousands. For this reason, all the statistical procedures normally adopted require modification and correction for multiple hypothesis testing. In fact, it is also intuitively clear that when a hypothesis is tested against a large number of tests, the probability of obtaining a statistically significant answer based only on random chance increases. Statistical testing requires the formulation of the hypothesis to test and its alternative (known as null H_0 and alternative H_1 hypothesis respectively), and it assesses the likelihood (α-level or p-value) of a given test to be the real expression of some similarity to a known distribution of values or to occur by chance. When dealing with numerous tests at one time, p-values need to be readjusted in a more conservative way, accounting for the so called *multiple hypothesis testing* issue. The most simplistic solution to this issue is the Bonferroni correction [Sokal and Rohlf 2003] that simply multiplies the actual p-value of every single test by the total number of tests observed. However, this approach is not considered feasible in omic studies as it often produces results that are too conservative, thus this approach can fail to reject the null hypothesis (H_0) when it should, leading to erroneous conclusions (*type I error* or *false positive*). An alternative and less conservative approach to this problem is the generation of a random distribution based on random resampling or on the generation of scores obtained from the randomization of the data. Such approaches allow for building a distribution that represents the population's behavior and can thus be used to test the hypothesis of interest [Sokal and Rohlf 2003]. When operating with microarray data, a novel statistic, the false discovery rate (FDR), has been introduced [Benjamini and Hochberg 1995; Storey and Tibshirani 2003; Tusher et al. 2001]. Like the p-value, the FDR measures the false positives. However while the p-value controls the number of false positive over the number of truly null tests, the FDR controls the number of false positive over the fraction of significant tests.

4.1.4 *The Medical Effort.* The change in the medical area is more subtle, probably less visible, but is also the one with the stronger longterm impact. Briefly, the possibility to observe genome-wide variations has triggered a change in the perception of health and cure, leading to the emergence of novel types of medicine and paving the way for an actual personalized medicine. This change is rooted in the possibility offered to researchers to observe a comprehensive amount of data that defines the molecular portrait of an individual. This is the fundamental information on which personalized medicine can operate to quantify information that can explain the complexity of a disease [West et al. 2006]. Historically, microarrays, the first *omic* platform, have opened the door to a tangible way of analyzing biological data in a completely different way, allowing scientists from a number of academic institutions to observe data from a systemic point of view, processing data with a parallelism that was not even thinkable before the advent of such technology. Thus, not only can signatures be a useful tool for diagnostic and prognostic assessment, but the integration of several such signatures represent the next challenge that must be undertaken in order to increment the risk assessment prediction power. Several discipline areas are now under development for the effective use of these novel concepts. Pharmacogenomics and chemogenomics are some of these novel research fields that can help to define medical treatment based on principles adapted to each patient, making use of synergistic contributions of several research areas [di Bernardo1 et al. 2005]. The importance of pharamcogenomics lies in its potential impact both on economic and medical areas. In fact, this branch of pharmacology deals with the influence of genetic variation on drug response in patients that can be inferred thanks to the relationship evaluated by correlating gene expression or DNA mutations (mostly single-nucleotide polymorphisms) with a drug's efficacy. Such a personalized approach can avoid the number of side effects and can improve and optimize drug therapy. Institutions such as the Institute for Systems Biology in Seattle led by Lee Hood are devoted to the effective realization of the so-called 3P medicine (predictive, preventive and personalized) integrating the concepts and techniques described. All these innovations can be broadly grouped under the development of systems biology described in Section 4.2.

4.2 Systems Biology

In contrast with Section 4.1, this section cannot be built with the separate contribution of different research areas since systems biology is the result of the effort of building an interdisciplinary approach to molecular biology. For this same reason, systems biology is among the most promising areas of research emerging in the post-Genomic era. It carries promises to that answer some of the more dramatic questions related to the complex and systemic functioning of healthy and diseased mechanisms.

Globally, the approaches described in Section 4.1 proved to be useful in uncovering information relevant to cancers, one of the most complex and widespread diseases. Cancers, in fact, are often characterized by high heterogeneity that impedes correct taxonomy of the malady and, consequently appropriate, personalized therapy. To improve upon this, more and more often scientists from all backgrounds are interested in adding nongenomic data to microarray analyses in order to guide the information extraction with other types of meaningful data such as clinical traits, survival information, and so on. However, static relationships among transcripts have proven to be insufficient to explain the interplay among genes. In fact, while signatures can be used as markers for diseases, they are too simple in their structure to account for the molecular mechanisms that lead to the actual coexpression. To face this new challenge, that aims at understanding more complex and dynamic relationships among gene activity, network approaches and graph theories have become the main tool. They promise to permit the understanding of cause-effect relationships among genetic synergies. The wealth of approaches produced so far is comparable to the work produced in the definition of data-mining algorithms for microarrays in earlier times. There are two main sets of approaches: the ones based on *physical interactions* that search for actual relationships among molecules (such as protein-protein interactions (see Section 3.2) or transcription factor and target gene (Section 3.3) or general metabolic networks that break down metabolism pathways into their respective reactions and enzymes); and the ones called *influence interactions* that represent relationships that can be inferred (such as indirect interactions that lead to coexpression (Section 3.1), possibly mediated through unknown intermediate molecules [Bansal et al. 2007] also known as Gene regulatory Networks). Genome-wide studies now offer the possibility of investigating both types of interactions. However, literature and available data such as expression data are richer in the influence interaction, area, and the wealth of data and the experience accumulated in this area is also greater. Several efforts are under way to elucidate the relationships occurring among genes, to uncover the potential pathways, cause-effect relationships that are missing or disrupted in diseased cells. Interestingly, the tools designed in this novel era of molecular biology always appear to be published accompanied by validation on real data, a sign of the growing interaction among scientists from different disciplines. Along this path, the typical process of reverse engineering in systems biology requires the identification of a model based on simulated data and its iterative adjustment and correction by validation on real data, experimentally obtained. Experimental techniques might vary with the network under observation and can include the whole range of experiment presented in Section 3.

From the point of view of the model, there are three main techniques used to perform reverse engineering: (i) ordinary differential equation (ODE) models; (ii) Boolean equation models ; (iii) Bayesian models (for a simple example, see Figure 6). In general, models are graphs $G\{X, E\}$, where X is a set of N vertices $x_i, i = 1..N$, and E a set of edges $e_i, i = 1..N$. If the edges indicate that the relationship between the nodes moves from one node to the other, the graph is said to be *directed*, and it is possible to infer the flow of activation from one gene to another. If the graph is *undirected* this inference is not possible.

4.2.1 *Ordinary Differential Equation Models.*

Ordinary differential equation models are deterministic approaches. Although this tutorial focuses on gene networks, it is worth mentioning that the flexibility of ODE enables researches to approach the modeling of a variety of systems and behaviors, with this method ranging from the inference of the physiological properties of a cell from the network of proteins interactions [Tyson et al. 2001] to the details of the dynamic activity of an operon (set of genes working on the same function) in *Escherichia coli* [M.Santillán et al. 2007]. For a review see also You [2004].

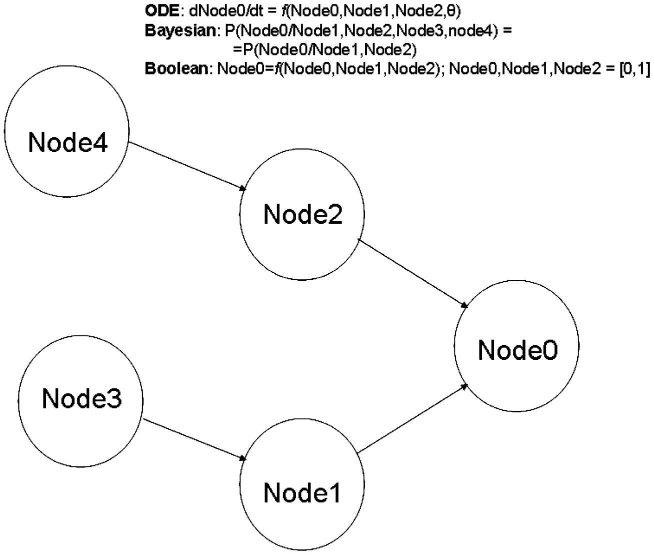
ODE: dNode0/dt = *f*(Node0,Node1,Node2,θ)
Bayesian: P(Node0/Node1,Node2,Node3,node4) =
=P(Node0/Node1,Node2)
Boolean: Node0=*f*(Node0,Node1,Node2); Node0,Node1,Node2 = [0,1]
Node4
Node2
Node0
Node3
Node1

Fig. 6. This simple network shows a directed graph and the expression of the gene represented by Node 0, depending on the model adopted. Notice that the Bayesian example assumes that Node 0 is conditionally independent of Node 3 and Node 4, given Nde 1 and Node 2. Notice also that the ODE and Boolean model have output also dependent on the value of the node itself.

As far as gene networks are considered, this approach is able to relate changes in an abundance of gene transcripts. These variations are due to changes in the concentration of some transcripts (called perturbations) that, in turn, influence other transcription rates in the cell. Formally, this is represented as $x_i(t) = f_i(x_1, \ldots, x_N, u, \theta_i)$, where θ_i represents the edges of the network, that is, interactions among the N genes, $x_i(t), i = 1..N$ represents the abundance of the transcript i, and u is the perturbation of the system. The modeling involves the identification of the parameters θ_i (corresponding to the edges e_i). Once the model is validated, perturbations variations can be performed and the consequent evolution of the genomic transcripts can be predicted [Gardner et al. 2003; di Bernardo1 et al. 2005; Bonneau et al. 2006].

4.2.2 *Boolean Models.* Boolean equations models are also deterministic models. In particular, their use in the description of gene network is defined as a a system of N binary-state nodes (genes) with K inputs to each node representing regulatory mechanisms. In Boolean approaches, only two values (0/1) represent the status of a gene expressed or inactive. The activity of a gene is thus a binary value function of the K binary states of the gene of interest's parents $x_i(t) = f_i(x_1, \ldots, x_K)$, $x_i = [0, 1]$. These approaches are broadly studied since several algorithms and techniques exist that can handle Boolean functions, however, binary values are often not enough to explain genetic relationships. A novel approach, called a probabilistic Boolean network has been designed by Shmulevich et al. [2000], that includes the uncertainty of the data used to generate the model to design the prediction. These networks are very appealing since they are based on the principles that regulate Boolean functions that are well understood and standardized in a number of libraries. However, their complexity grows very quickly and becomes untreatable for medium-size networks.

4.2.3 *Bayesian Models.* Bayesian Models are probabilistic approaches to the gene network problem. In this case, the expression value X_i is a random variable, and the relationship among the expression values is described by a joint probability distribution $P(X_1, .., X_N) = \prod_{i=1}^{N} P(X_i = x_i|X_j = x_j, X_{j+p} = x_{j+p})$, where the $p + 1$ genes on the right-hand side of the formula represent the genes on which the probability is conditioned, and are called the parents of gene i, for which they represent the regulators. In contrast to the two previous models, these edges can only assume positive values (they represent a probability $p \in [0, 1]$).

4.2.4 *Further Developments.* It is important to advise the reader that the expansion of the types and amount of information included in these analyses is in continuous development. In fact, novel areas of research, such as *biodiversity informatics*, aim at the creation, integration, analysis, and understanding of information regarding biological diversity. Not only, can different layers of molecular information can be usefully included to enrich each other, but the development of methods to organize knowledge at the organism level can be of great help in understanding the evolution and ultimately the role of various mechanisms occurring in organisms across the entire tree of life [Sarkar 2007].

Moreover, the methods developed to reconstruct the complex interactions among genes, both with approaches typical of bioinformatics and with reverse engineering algorithms, allow for the simulation of the evolution of such interactions. Importantly, the implicit models adopted in reverse engineering approaches that allow for the definition of efficient frameworks to simulate knock-out experiments are difficult and expensive to be performed in vivo.

Finally, the integrative approach that characterizes systems biology, along with the possibility of obtaining the synthesis of DNA sequences with case and at affordable cost are now being adopted in another extremely promising field, synthetic biology [Endy 2005]. While systems biology aims at uncovering the relationships ongoing in real molecular systems (analysis), synthetic biology aims at building and engineering (small) molecular systems (synthesis). This area was able to develop and take advantage of another important product of the Human Genome Project, DNA synthesis. In fact, to achieve the decoding of full genomes, the development of DNA decoders and synthesizer has received a strong boost. Given the larger availability of tools for genetic manipulation, it is now possible to mount genetic circuits using basic standardized units (biobricks, www.biobricks.org) as is done in the International Genetically Engineered Machine Competition (iGEM, www.igem2007.com), the undergraduate competition in synthetic biology that takes place every year at MIT in Cambridge, MA. Although the synthesis of (complex) genetic circuits is now a concrete possibility, several limitations or caveats exist. The challenges ahead are still big and involve areas such as extended standardization and ethical issues [Arkin and Fletcher 2004].

5. POINT-OF-CARE SYSTEMS

While Section 4 deals with promising principles to be used in mostly still to come widespread devices, this section will give an overview on methods and approaches that are much closer to actual application in several fields from health care to environmental monitoring.

In particular, point-of-care devices for the implementation of outpatient clinics—to enable high-level quality-of-care outside laboratories and hospitals—are extremely desirable for reasons related to cost, comfort, and efficiency. These systems can address multiple types of diagnoses such as RNA expression analysis for the detection of viral infections, of biohazards, or the fast monitoring of a specific set of biomarkers. The aim of a diagnostic or, in general, of distributed analysis system is basically to implement a low-cost, portable, simplified, yet reliable version of existing genetic and genomic analysis systems. Assays on genes and proteins are performed by laboratory setups on a variety of platforms (arrays, gels, filters) and involve different techniques for separation (mass spectrometry, capillary electrophoresis blotting), amplification (rolling-circle amplification, Polymerase Chain Reaction, Reverse Transcriptase PCR), labeling or signal amplification (radioactivity, fluorescence, chemiluminescence, electrochemical). Still, the most suitable and convenient method for diagnostics has not been identified yet [Fortina 2002]. In order to create a point-of-care gene-based detection system, expertise in molecular biology and engineering must converge. The aim is to obtain miniaturized devices for analysis that can then be mass produced and are superior in performance with respect to existing systems.

Issues related to these goals are multiple. For instance, complexity reduction is a crucial objective. Such a goal refers to minimizing the number of sites needed to perform a specific and reliable analysis. Interestingly, this can be achieved, for example, downstream of omic analyses (see Section 4) where the identification of signatures and pathways is performed and validated. Furthermore, the reduction of the number of sites diminishes the global cost of the fabrication and this impacts, for instance, the amount of reagents required. This and other issues, oriented to cost reduction and portability, define a complete pipeline of steps crucial to the development of an effective point-of-care device implementation.

This section will approach the device implementation problem from a wide perspective, with particular emphasis on portability (Section 5.1), system integration (Section 5.2), related biocompatibility issues (Section 5.3), and finally on a means for easing the use of such devices (Section 5.4). In order to concretize challenges, issues, and solutions related to these steps, we will support the description with examples related to the point-of-care systems development that have been faced during our experimental activity.

5.1 Detection Techniques for Portability

The need for portability of genetic tests fuels innovation in many fields of biomolecular sensing. Portability can be achieved by a reduction in size until complete miniaturization of existing biosensing systems (biosensors) has been acheived with great care to avoid loss of functionality. The phases toward the development of a portable device from existing techniques can be described as follows. Phase 1 concerns the reduction in size of the measurement unit, the supplies, and the source of excitation. Then, Phase 2 impacts the transducing systems by developing ad hoc miniaturized and tunable devices to measure physical and chemical parameters modified by molecular binding events. Finally, the measurement systems and sites have to converge to a unique support (Phase 3).

Usign the three-phase path described, we will present two examples. Let us consider first impedance measurements on electrode interfaces. In this frame, Phase 1 consists of the development of the impedance instrumentation in a compact and automated version [Riepl et al. 1999]. Phase 2, has been approached with strong innovation by measuring the binding event by sensing capacitance changes with a suitable charge-based technique [Stagni et al. 2005]. In Phase 3, two different capacitance measurement techniques implemented on a silicon chip have been successfully applied to the detection of the hybridization between nucleotide strands [Stagni et al. 2006a, 2006b].

A second example concerns the piezoelectric technique [Willner et al. 2002]. The systems developed based on this mass-sensing technique by means of changes in the resonance frequency are well known for their ability to resolve nanograms (i.e., layers of biomolecules) on their surface. Among piezoelectric devices, quartz-crystal microbalances and surface acoustic wave devices have both given outstanding results. Moreover, they have the desirable advantage of transducing the mass changes directly into an electrical signal that can be easily conditioned and processed. Bulky and compact implementations with different degrees of automation already exist (Phase 1 and Phase 2), but recently advancements toward integration, Phase 3, have been achieved by Brederlow et al. [2003] who developed film acoustic wave devices integrated on a CMOS chip and were able to detect the presence of a Bovine Serum Albumine surface layer.

Other techniques that allow high-precision biosensing have not yet undergone all of the three phases. In some cases, the portability path is limited by the characteristics of the detection technologies. In particular, the surface plasmon resonance technique, based on the measurement of the refractive index changes at a gold solution interface, is well known, but it is still available only as benchtop instrumentation. This is due to the fact that the technique must be equipped with a quite complex optical part which consists of a monochromatic laser beam and an apparatus for precisely directing the laser on the metal back to detect the reflected intensity [McDonnell 2001].

5.2 Integration

Existing devices integrate only a few of the basic steps involved in molecular detection, which is a challenging and complex analysis demand (see Figure 7). The steps can be classified according to their functions: (i) presensing steps: extraction, separation, amplification; (ii) sensing steps: sensing, transduction for the generation of electrical signals; and (iii) readout steps: signal conditioning and data processing.

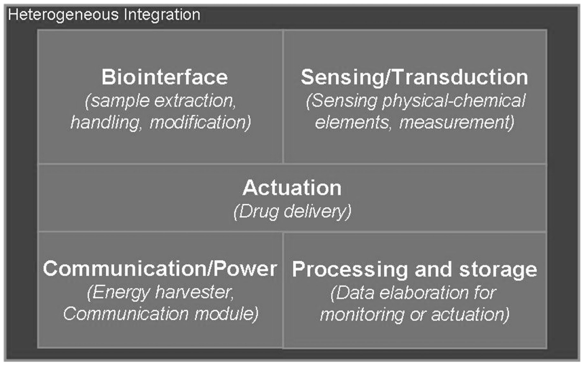
Heterogeneous Integration
Biointerface
(sample extraction, handling, modification)
Sensing/Transduction
(Sensing physical-chemical elements, measurement)
Actuation
(Drug delivery)
Communication/Power
(Energy harvester, Communication module)
Processing and storage
(Data elaboration for monitoring or actuation)

Fig. 7. Lab-on-a-chip components. A stand-alone, miniaturized device for biomolecular analysis and gene-based tests integrates on a single device all the functions indicated: a biointerface to extract, dilute, and manage the molecules to be analyzed; transducers, sensing, and conditioning circuits; circuits for data elaboration; communication, and power generation capabilities; an actuation device controlled by the sensing and elaboration functions.

Integration of analysis systems is based on microtechnology to create structures using both standard microelectronics materials and nonconventional materials. Silicon, glass, and metals are employed to integrate data and signal processing, semiconductor sensors, microchannels, electromechanical microcomponents and thermal actuators. New concept materials of a polymeric nature (e.g., polydimethylsiloxane) are preferred for their biocompatibility and their suitability to be patterned to create channels for handling and processing biological samples [LaVan et al. 2001]. The ability to integrate the presensing steps is crucial since it is the most expensive step involved at present in gene-based tests.

Importantly, integration should be coupled with the implementation of configurability all through the fabrication process. This concept aims at defining of systems based on a common μ-fab platform to be employed in various applications, the costs of which will be drastically reduced by batch production (see Figure 8). The platform would host the semiconductor sensors and the proper circuitry for signal detection, conditioning, and processing, which could be selected and configured for a specific application. Heating-control elements for the pretreatment of the samples could also be integrated at this stage. Further offline technological steps could add specific features to the integrated sensors by depositing layers whose processing might be undesirable in standard μ-fab. Other steps aimed at targeting the systems for the desired microfluidic functions and patterning of the molecular probes corresponding to the analytes should follow. Improved speed and reliability could be achieved by allowing different analytes to be tested simultaneously with different probes patterned on microsites. Of course, as a further feature, the integration of batteries and communication modules would lead to stand-alone, self-contained controllable-programmable onfield systems.

We will now describe three approaches to functions integration specifically addressing DNA sensing.

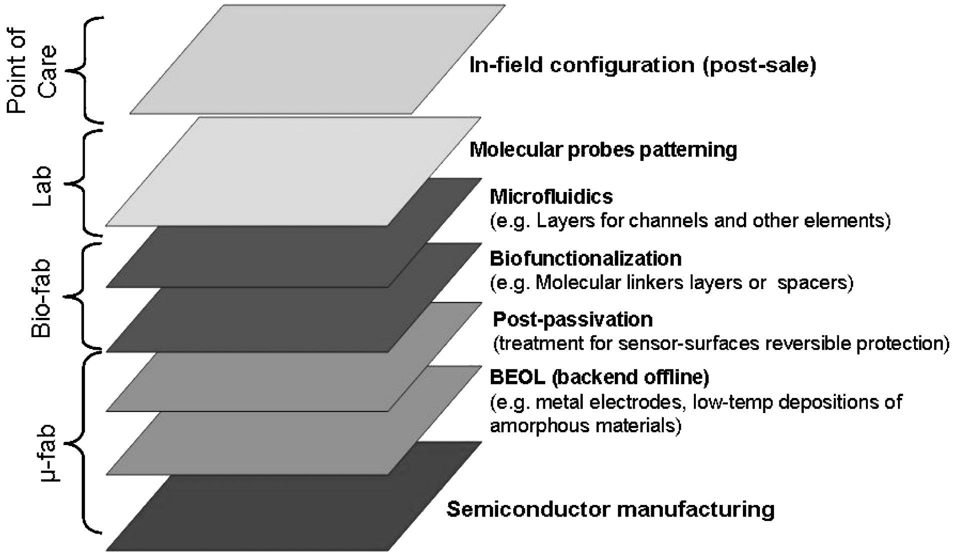
Point of Care
Lab
Bio-fab
µ-fab
In-field configuration (post-sale)
Molecular probes patterning
Microfluidics
(e.g. Layers for channels and other elements)
Biofunctionalization
(e.g. Molecular linkers layers or spacers)
Post-passivation
(treatment for sensor-surfaces reversible protection)
BEOL (backend offline)
(e.g. metal electrodes, low-temp depositions of amorphous materials)
Semiconductor manufacturing

Fig. 8. Configurability in integrated analysis systems. The items on the right correspond to the different implementation layers. On the left, the facilities in charge of the development of the different implementation and configuration steps are indicated. The device implemented by the μ-fab facility will be a configurable platform. The other steps could target the final systems (e.g., biofunctionalization, molecular probes patterning), or add other degrees of configurability (e.g., microfluidics).

The first approach focuses on integration of microfluidics. This is a crucial aspect for sensing devices as it enhances reliability by avoiding sample pollution due to manipulation and allows for the management of small sample volumes. Devices developed with microfluidic parts that perform active functions on the sample are generally named *Lab-on-a-chip*. Bulk and surface micromachining performed with sophisticated etching, patterning, and deposition techniques are at the basis of channel implementation. One of the most relevant microfabricated implementations on chip is the Polymerase Chain Reaction (PCR) molecular amplification. This approach has been widely investigated exploiting the good properties of thermal conductivity of silicon and its ability to easily integrate thermal resistances [Liu et al. 2004; Lagally et al. 2001, 2004]. A large number of basic fluidic components have been assembled in different ways to perform other common chemical processes [Kelly and Wolley 2003; Woolley et al. 1997; Shi et al. 1999]. Besides fluid-mechanics based techniques, an emerging technology based on droplets handling by means of arrays of electrodes, is gaining interest. Few groups demonstrated the potential for controling the movement of several drops at a time on a single chip and the potential for successfully performing mixing of reagents for chemical and biochemical reactions. These arrays, known as digital microfluidic chips, can be easily reconfigured by applying different algorithms to the droplet movement [Gong and Kim 2005; Xu et al. 2007].

The second approach concerns the integration of the existing fluorescence detection techniques applied on microarrays. The aim is to develop arrays of solid-state optical devices able to detect onsite the emitted light of the fluorescent labels. These photodiode arrays are meant to be both coupled with glass structures [Thrush et al. 2005; Kamei et al. 2003; Webster et al. 2001; Nakanishi et al. 2001], and to become active sensing substrates for the molecular spotted arrays [Xu et al. 2005; Misiakos et al. 2004; Fixe et al. 2004]. In the latter approach, probes may be spotted directly on the top of the chip by functionalizing the silicon dioxide passivation. Nevertheless, the integration of optoelectronic detection on existing fluorescence microarrays is still far achieved [Thrush et al. 2006b] because of the nontrivial design issues that exist due to the need for integrated filters in order to screen from being the excitation light.

Finally, the third approach uses semiconductor sensors for biomolecular detection. One possibility is to employ silicon nitride cantilevers developed by micromachining techniques and commonly used in MEMS or in atomic force microscopy. Recent papers reported the sensing of DNA and protein-binding events [Wu et al. 2001; Fritz et al. 2001] where authors were able to detect DNA hybridization, including accurate positive/negative detection of one base pair of mismatches. The detection principle is based on mass, or on the surface stress at the cantilever by measuring the change of the resonant frequency, or of bending, respectively. Another technique investigates electrolyte/insulator/semiconductor devices for biosensing purposes. These approaches include (i) a 1-dimensional vertical capacitor similar to MOS structures where metal has been substituted with an electrolyte conductive solution and (ii) field-effect transistors where the gate has been substituted as well. Such devices have been widely employed in the last twenty years in electrochemistry and analytical chemistry for ion and pH sensing. More recently DNA hybridization occurring at the interface between the electrolyte and the insulator has been detected. The biomolecular binding on the insulator is detected by the field effect of the DNA negative charge [Uslu et al. 2004; Fritz et al. 2000; Pouthas et al. 2004; Cloarec et al. 2002].

5.3 Compatibility

5.3.1 *Prevention of Chip Damages.* Some of the procedures involved in the array preparation (see Figure 9) have been demonstrated to permanently damage silicon chips developed for molecular sensing. In our experience, we tested these issues using a chip fabricated with standard CMOS technology plus a BackEnd OffLine (BEOL) process for electrodes implementation. We could show that the use of strong acid surface cleaning, such as piranha surface cleaning normally used in substrates preparation ($1 : 3\, H_2O_2/H_2SO_4$) [Riepl et al. 1999], can cause extensive damage to passivation and interconnects in as little as 30s. Polishing, a common method for electrode preparation [Janek et al. 1997], can be damaging for chips. Thus, the cleaning procedures should be restricted to ethanol bathes, ultrasonic bathes and oxygen plasma which are well known procedures for silicon chips and, eventually, electrochemical stripping techniques [Peterlinz and Georgiadis 1996].

Immediately after cleaning, the surface must be biofunctionalized with probemolecules in order to form a dense, compact, and homogeneous

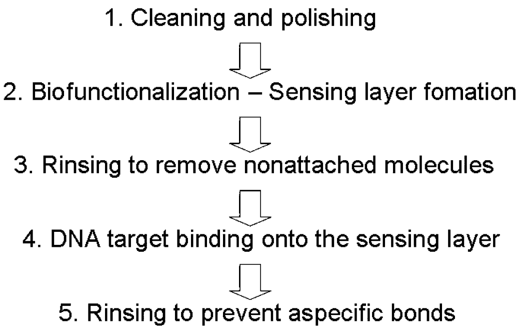
1. Cleaning and polishing
2. Biofunctionalization – Sensing layer fomation
3. Rinsing to remove nonattached molecules
4. DNA target binding onto the sensing layer
5. Rinsing to prevent aspecific bonds

Fig. 9. This flowchart describes the sequential steps involved in DNA recognition by means of electrodes.

biomolecular layer. Biofunctionalization can be very delicate with respect to chip integrity. Molecules can be deposited on the electrodes by different techniques, namely, immersion, microspotting, or inkjet printing and should be dissolved in a solution of salts to avoid electrostatic repulsion between charged molecules during layer formation. Then, the surface should be rinsed with ultrapure water or buffers to remove noncovalently attached probe molecules and to obtain a layer of ordered and well-oriented receptors. These procedures can potentially be dangerous for the chip surface, thus low-impact biofunctionalization procedures have to be selected to limit the damage. An incubation time of 5hrs in a 1M NaCl ensures 90% layer formation [Georgiadis et al. 2000]. Furthermore, the ideal immobilization process should be driven by microspotting techniques or inkjet printing techniques to limit the wet chip surface. Finally, molecular recognition is performed by spreading a buffer solution on the electrodes under suitable physical and chemical parameters, namely, temperature, ionic strength, pH, presence of surfactant. These conditions are extremely important to ensure efficient and specific sensing but, again, they should respect the device constraints. In particular, temperature should never exceed 70°C for noncritical processes, even if specificity could be enhanced by a high hybridization temperature. Being aware of such compatibility issues helps to improve the BEOL process in terms of the choice of the upper passivation and the employment of metals for ensuring a more robust adhesion to silicon dioxide in order to protect the paths toward the inside of the chip. In particular, the process defined by Infineon Technologies that takes into account all the aforementioned precautions, was robust in different biochemical operations [Stagni et al. 2006; Schienle et al. 2004].

5.3.2 *Interference of the System on Molecular Functions.* As mentioned the measurement system of an array can interfere with the effectiveness and dynamics of the binding. The primary reason concerns the fact that the binding must take place on a surface instead of in bulk. This can cause steric constraints and/or noncompatibility with the surface components (see Section 3.2). Moreover, a crucial issue in the choice of the immobilization technique concerns the fact that the probe layer or a prefunctionalization layer would probably be stored for a certain time before onfield use. Several immobilization techniques have been considered according to these requirements [Pirrung 2002] but a suitable procedure should be customized in every single case of molecular couple and chip type. In general, the employment of inert and biocompatible polymeric materials such as PDMS polydimethylsiloxane to develop the microfluidic, substrate and packaging parts is a common feature in lab-on-a-chip devices [Fujii 2002].

A different issue may concern the solicitation or excitation needed to perform the detection, that is, the measure the state of the molecular compound on the site. For example, in case of measurements by light absorbance, the energy absorbed by DNA molecules can cause damages to their structure [Sauer et al. 2001]. In particular, nucleic acid bases that absorb light energy in the UV range with a peak around 260nm and 10 J/m^2 is usually considered a safe intensity. Electrochemical and electrical techniques are the most critical; some of the bases that form the DNA molecules have electroactive behavior on electrodes if polarized to critical potential [Thorp 1998]. In particular, the irreversible oxidation of Guanine and Adenine on electrodes polarized at around 1V vs. Ag/AgCl reference electrode has been described in detail [Oliveira-Brett et al. 2002]. Finally, as discussed in Section 3.2, mechanical solicitation can be critical for more fragile molecules as proteins.

5.4 Simplification of Procedures for the End User

A suitable choice of the detection technique that minimizes the sample preparation steps which leads to improved reliability of the assay while suppressing possible sources of pollution, and errors and simplifying the onfield procedure can be done by the end user. Ideally, biochemical and mechanical processes like amplification and separation should be avoided unless they are implemented on the system in self-controlled, stand-alone parts. In particular, much efforts have been expended in order to develop techniques that perform direct detection of the binding event, that is, that avoid the use of labels to be associated and attached to the target molecules [Guiducci et al. 2006; Georgiadis et al. 2000; Willner et al. 2002]. Among the cited techniques, impedance, quartz-crystal microbalances, and surface plasmon resonance are label-free methods.

6. CONCLUSION AND PERSPECTIVES

The fast and frequent discovery and development of enabling technologies is a distinctive characteristic of our era. Roughly speaking, novel technologies arise thanks to the mastering of naturally occurring phenomena as well as to the innovative association of known techniques from assessed fields. Both these circumstances have contributed to the diffusion of biotechnologies, a fast growing area accompanied by a plethora of novel approaches, theories, and ultimately innovative paradigms impacting science and technology.

The most high-technology platforms that are involved in this area move along two main paths, one more focused on the portability and broad accessibility of genomic analyses, and one more oriented to the systemic approach in the discovery of molecular mechanisms through dense parallelization. The translation of both these innovations into practice requires the integration of these two paths. This implies the possibility of performing genome-wide screens at affordable prices, in real time and with reliable outputs. To meet these goals, several requirements need to be fulfilled. At a first stage, all the different platforms applied to the same type of omic data must be able to provide comparable results, and this comparability must become a routine, broadly acknowledged process, thus expanding the results of initiatives like the MAQC and other studies to come. At a higher lever, different types of omic data must be interpreted coherently, that is, knowledge at the genomic level has to be meaningful for the proteomic level and so on. Finally, not only must these results be coherent but they must also be able to allow for specific and sensitive testing in assessed relationship with diseases and healthy mechanisms for improved prevention, diagnosis, and follow-up.

Only when these goals are met with the actuation of personal medicine will become feasible. Personalized medicine represents the final outcome of a systemic approach to the disease, while taking into account the uniqueness of each patient. All developed societies heavily encourage and promote such a vision. This, in fact, promises to greatly improve the efficient management of all sorts of resources in health care from reduced side effects due to inappropriate drug dosage, to improved patient engagement, enhancing proactive actions and attitudes toward health recovery or disease awareness.

In conclusion, this area is extremely lively and innovations occur at a very fast pace. Both industry and academic institutions provide almost weekly noticeable advancements in one or more of the areas affecting the evolution of the broad field of biotechnology. This is actually one of the reasons for the dynamism of this area since every innovation fuels enhancements in other related fields or triggers new exploratory studies, technological designs, and projects.

To capture as much as possible of this dynamism, we first introduced the background on which this area fired its development, moving from the Human Genome Project, the easiest and clearer time-point to define. We then briefed the reader on the basics of molecular biology, a necessary step to understand the complexity, challenges, and difficulty arising in the management of such data. With this background, we moved to an overview on the platforms currently available to the highly parallel study of molecular interactions. Further, we described the approaches defined to mine knowledge from the readout of such platforms, and, finally, we delineated the technological innovations now available with special attention to the characteristics of portability. This article attempts to describe a large, but possibly incomplete, portion of the wide range of innovations occurring in the biotechnology area. Our article is oriented to both highly parallel and low-cost portable devices and their use, with the intention of offering an overview on the evolution that accompanied the most recent breakthroughs, paying attention to the goals, issues, and hopes that are still driving research and innovation.

ACKNOWLEDGMENTS

The authors would like to thank Professor Luca Benini for all the constructive conversation that permitted this work and for his invaluable help.

REFERENCES

Afanassiev V., Hanemann V. 2000. Preparation of dna and protein microarrays on glass slides coated with an agarose film. *Nucleic Acids Res. 28*, 126, e66–e66.

Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. 2004. FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics 20*, 4, 578–580.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. 1989. *Molecular Biology of the Cell*. Garland Publishing, New York, NY.

Alizadeh, A., Eisen, M., Davis, R., C. Ma, I. L., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Jr, J. H., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, J., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P., , and Staudt, L. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature 403*, 6769, 503–511.

Altschul, S., Madden, T., Schäaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res. 25*, 3389–3402.

Arkin, A. P. and Fletcher, D. A. 2004. Fast, cheap and somewhat in control. *Genome Biol. 7*, 11.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. 2007. How to infer gene networks from expression profiles. *Molec. Syst. Biol. 3*, 78.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J.R. Stat. Soc. B 57*, 289–300.

Bonneau, R., Reiss, D., Shannon, P., Facciotti, M., Hood, L., Baliga, N., and Thorsson, V. 2006. The inferelator: An algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol. 7*, 5, R36.

Brederlow, R., Zauner, S., Scholtz, A., Aufinger, K., Simburger, W., Paulus, C., Martin, A., Timme, M. F. H.-J., , Heiss, H., Marksteiner, S., Elbrecht, L., Aigner, R., and Thewes, R. 2003. Biochemical sensors based on bulk acoustic wave resonators. In *IEEE International Electron Devices Meeting (IEDM'03)*.

Brown, P. and Botstein, D. 1999. Exploring the new world of the genome with DNA microarrays. *Nat. Genet. 21*, 1, 33–37.

Bullinger, L., Rücker, F. G., Kurz, S., Du, J., Scholl, C., Sander, S., Corbacioglu, A., Lottaz, C., Krauter, J., Fröhling, S., Ganser, A., Schlenk, R. F., Döhner, K., Pollack, J. R., and Döhner, H. 2007. Gene-expression profiling identifies distinct subclasses of core binding factor acute myeloid leukemia. *Blood 110*, 4, 1291–1300.

Cameron, M., Williams, H., and Cannane, A. 2004. Improved gapped alignment in BLAST. *IEEE/ACM Trans. Computat. Biol. Bioinform. 1*, 3, 116–129.

Cheng, Y. and Church, G. M. 2000. Biclustering of expression data. In *Proceedings of the International Conference on Intelligent Systems in Molecular Biology. 8*, 93–103.

Churchill, G. A. 1953. Fundamentals of experimental design for cDNA microarrays. *Nature (Genetic Supplement) 32*, 490–495.

Cloarec, J. P., Deligianis, N., Martin, J. R., Lawrence, I., Souteyrand, E., Polychronakos, C., and Lawrence, M. F. 2002. Immobilization of homooligonucleotide probe layers onto Si/SiO2 substrates: Characterization by electrochemical impedance measurements and radiolabelling. *Biosens. Bioelectron. 17*, 405–412.

Collins, F. S., Morgan, M., and Patrinos, A. 2003. The human genome project: Lessons from large-scale biology. *Science 300*, 5617, 286–290.

Collins, F. S., Patrinos, A., Chakravarti, E. J. A., Gesteland, R., Walters, L., and the members of the DOE and NIH planning groups. 1998. New goals for the U.S. human genome project: 1998-2003. *Science 282*, 5389, 682–689.

Consortium, T. G. O. 2001. Creating the gene ontology resource: Design and implementation. *Genome Res. 11*, 8, 1425–1433.

di Bernardo1, D., Thompson, M. J., Gardner, T. S., Chobot, S. E., Eastwood, E. L., Wojtovich, A. P., Elliott, S. J., Schaus, S. E., and Collins, J. J. 2005. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol. 23*, 377–383.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. 95*, 25, 14863–14868.

Endy, D. 2005. Fundations for engineering biology. *Nature 438*, 24, 449–453.

Esteller, M. 2006. The necessity of a human epigenome project. *Carcinogenesis 27*, 1121–1125.

Fire, A., Xu, S., Montgomery, M., Kostas, S., and anc C. Mello, S. D. 1998. Potent and specific genetic interference by double-stranded RNA in *caenorhabditis elegans*. *Nature 391*, 6669, 806–811.

Fixe, F., Chu, V., Prazeres, D. M. F., and Conde, J. P. 2004. Thin film micro arrays with immobilized DNA for hybridization analysis. *Anal. Chem. 32*, 9, e70.

Fortina. 2002. Molecular diagnostics: hurdles for clinical implementation. *Trend Molec. Medicine 8*, 6, 264–266.

Franke, L., van Bakel, H., Fokkens, L., de Jong, E. D., Petersen, M.-E., and Wijmenga, C. 2006. Reconstruction of functional human gene network, with an application for prioritizing positional candidate genes. *Amer. J. Hum. Genetics 78*, 6, 1011–1025.

Fritz, J., Cooper, E. B., Gaudet, S., Sorger, P. K., and Manalis, S. R. 2000. Electronic detection of DNA by its intrinsic molecular charge. *Proc. Natl. Acad. Sci. 99*, 8, 14142–14146.

Fritz, J., Lang, M. K. B. H. P., Rothuizen, H., Vettiger, P., Meyer, E., Guntherodt, H. J., and Gimzewski, C. G. J. K. 2001. Translating biomolecular recognition into nanomechanics. *Sci. 288*, 316–318.

Fujii, T. 2002. PDMS-based microfluidic devices for biomedical applications. *Microelectron. Engin. 61–62*, 907–914.

Gardner, T. S., di Bernardo, D., Lorenz, D., and Collins, J. J. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science 301*, 5629, 102–105.

Georgiadis, R., Peterlinz, K. P., and Peterson, A. W. 2000. Quantitative measurements and modeling of kinetics in nucleic acid monolayer films using spr spectroscopy. *J. Amer. Chem. Soc. 122*, 3166–3173.

Getz, G., Levine, E., and Domany, E. 2000. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. 97*, 22, 12079–12084.

Guiducci, C., Stagni, C., Brocchi, M., Lanzoni, M., Riccó, B., Nascetti, A., Caputo, D., and Cesare, A. D. 2006. *Innovative Optoelectronic Approaches to Biomolecular Analysis with Arrays of Silicon Devices*. Springer.

Gong M. and Kim, C.-J. 2005. Two-dimensional digital microfluidic system by multilayer printed circuit board. In *Proceedings of the 18th IEEE International Conference on MEMS*.

Hinterdorfer, P., Baumgartner, W., Gruber, H. J., Schilcher, K., and Schindler, H. 1996. Detection and localization of individual antibody-antigen recognition events by atomic force microscopy. *Proc. Natl. Acad. Sci. 93*, 8, 3477–3481.

Hocquette, J. F. 2005. Where are we in genomics? *J. Physiol. Pharmacol. 56*, 3, 37–70.

Janek, R. P., Fawcett, W. R., and Ulman, A. 1997. Impedance spectroscopy of self-assembled monolayers on Au(111): Evidence for complex doublelayer structure in aqueous NaClO4 at the potential of zero charge. *J. Phys. Chem. B 101*, 8550–8558.

Jones, V. W., Kenseth, J. R., Porter, M. D., Mosher, C. L., and Henderson, E. 1998. Microminiaturized immunoassays using atomic force microscopy and compisitionally patterned antigen arrays. *Anal Chem 70*, 1233–1241.

Kamei, T., Paegel, B. M., Scherer, J. R., , Street, A. M. S. R. A., and Mathies, R. A. 2003. Integrated hydrogenated amorphous si photodiode detector for microfluidic bioanalytical devices. *Anal. Chem. 75*, 20, 5300–5305.

Kelly, T. R. and Wolley, A. T. 2003. Thermal bonding of polymeric capillary electrophoresis microdevices in water. *Anal. Chem. 75*, 8, 1941–1945.

Lagally, E. T., Medintz, I., and Mathies, R. A. 2001. Single-molecule DNA amplification and analysis in an integrated microfluidic device. *Anal. Chem. 73*, 565–570.

Lagally, E. T., Scherer, J. R., Blazej, R. G., Toriello, N. M., Diep, B. A., Ramchandani, M., Sensabaugh, G. F., Riley, L. W., and Mathies, R. A. 2004. Integrated portable genetic analysis microsystem for pathogen/infectious disease detection. *Anal. Chem. 76*, 3162–3170.

Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A. M., Tibshirani, R., Botstein, D., Brown, P. O., Brooks, J. D., and Pollack, J. R. 2004. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. 101*, 3, 811–816.

LaVan, D., Lynn, D., and Langer, R. 2001. Moving smaller in drug discovery and delivery. *Nat. Rev. 1*, 77.

Lee, H. K., Braynen, W., Keshav, K., and Pavlidis, P. 2005. ErmineJ: Tool for functional analysis of gene expression data sets. *BMC Bioinform. 6*, 269, doi:10.1186/1471–2105–6–269.

Lipman, D. and Pearson, W. 1985. Rapid and sensitive protein similarity searches. *Science 227*, 4693, 1435–1441.

Liu, R. H., Yang, J., Lenigk, R., Bonanno, J., and Grodzinski, P. 2004. Self-contained, fully integrated biochip for sample preparation, polymerase chain reaction amplification, and DNA microarray detection. *Anal. Chem. 76*, 1824–1831.

Madeira, S. C. and Oliveira, A. L. 2004. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Computat. Biol. Bioinform. 1*, 1, 24–45.

Masotti, D., Nardini, C., Rossi, S., Bonora, E., Romeo, G., Volinia, S., and Benini, L. 2007. TOM: enhancement and extension of a tool suite for in silico approaches to multigenic complex disorders. *Bioinform.* To appear.

Maxam, A. M. and Gilbert, W. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. 74*, 2, 560–564.

McDonnell, J. M. 2001. Surface plasmon resonance: towards an understanding of the mechanisms of biological molecular recognition monolayer films using SPR spectroscopy. *Curr. Opin. Chem. Biol. 5*, 5, 572–577.

Misiakos, K., Kakabakos, S. E., Petrou, P. S., and Ruf, H. 2004. A monolithic silicon optoelectronic transducer as a real-time affinity biosensor. *Sens. and Act. B 76*, 5, 1366–1373.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. 2003. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet. 34*, 3, 267–273.

Mullis, K. B. 2003. The first polymerase chain reaction. *Scientist 17*, 4, 11.

Nakanishi, H., Nishimoto, T., Arai, A., Abe, H., Kanai, M., Fujiyama, Y., and Yoshida, T. 2001. Fabrication of quartz microchips with optical slit and development of a linear imaging uv detector for microchip eletrophoresis systems. *Anal. Chem. 22*, 2, 230–234.

Oliveira-Brett, A. M., Diculescu, V., and Piedade, J. 2002. Electrochemical oxidation mechanism of guanine and adenine using a glassy carbon microelectrode. *Bioelectrochem. 55*, 61–62.

Pederson, T. 2004. RNA interference and mRNA silencing, 2004: How far will they reach? *Molec. Biol. Cell 15*, 2, 407–410.

Peterlinz, K. and Georgiadis, R. 1996. In situ kinetics of self-assembly by surface plasmon resonance spectroscopy. *Langmuir 12*, 4731–4740.

Pirrung, M. C. 2002. How to make a DNA. *Angew. Chem. Int. Ed. 41*, 1276–1289.

Pouthas, F., Gentil, C., Cote, D., and Bockelmann, U. 2004. Dna detection on transistor arrays following mutation specific enzymatic amplification. *Appl. Phys. Lett. 84*, 1594–1596.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. 2007. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res. 35*, D61–D65.

Quackenbush, J. 2001. Computationa analysis of micorarray data. *Nat Rev Genet 2*, 6, 418–427.

Quackenbush, J. 2007. Extracting biology from high-dimensional biological data. *J. Exp. Biol. 210*, 1507–1517.

Ramaswamy, S., Ross, K. N., Lander, E. S., and Golub, T. R. 2003. A molecular signature of metastasis in primary solid tumors. *Nat. Genet. 33*, 1, 49–54.

Riepl, M., Mirsky, V. M., Novotny, I., Tvarozek, V., Rehacek, V., and Wolfbeis, O. S. 1999. Optimization of capacitive affinity sensors: drift suppression and signal amplification. *Analytica Chimica Acta 392*, 1, 77–84.

Romualdi, C., Trevisan, S., Celegato, B., Costa, G., and Lanfranchi, G. 2003. Improved detection of differentially expressed genes in microarray experiments through multiple scanning and image integration. *Nucleic Acid Res. 31*, 23, e149–1–e149–8.

Ronaghi, M. 2001. Pyrosequencing sheds light on DNA sequencing. *Genome Res. 11*, 1, 3–11.

Rossi, S., Masotti, D., Nardini, C., Bonora, E., Romeo, G., Macii, E., Benini, L., and Volinia, S. 2006. TOM: a web-based integrated approach for efficient identification of candidate disease genes. *Nucleic Acids Res. 34*, doi:10.1093/nar/gkl340, W285–W292.

Sokal, R. R. and Rohlf, F. J. 2003. *Biometry*. Freeman, New York, NY.

Sanger, F. and Coulson, A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Molec. Biol. 94*, 3, 441–448.

Santillán, M., Mackey, M. C., and Zeron, E. S. 2007. Origin of bistability in the *lac* operon. *Biophys. J. 92*, 3830–3842.

Sarkar, I. N . 2007. Biodiversity informatics: Organizing and linking information across the spectrum of life. *Brief Bioinform.*, To appear.

Sauer, I. T. I, Wang, K., and Puglisi, J. D. 2001. *Chemistry:Principles and Applications in Biological Sciences*. Prentice Hall.

Schienle, M., Frey, A., Hofmann, F., Holzapfl, B., Paulus, C., Schindler-Bauer, P., and Thewes, R. 2004. A fully electronic DNA sensor with 128 positions and in-pixel A/D conversion. *Digest of Technical Papers. IEEE International 1*, 220–524.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. 2003. Module networks. identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet. 34*, 2, 166–176.

Segal, E., Sirlin, C. B., Ooi, C., Adler, A. S., Gollub, J., Chen, X., Chan, B. K., Matcuk, G. R., Barry, C. T., Chang, H. Y., and Kuo, M. D. 2007. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nature Biotechn. 25*, 6, 675–680.

Sevilla, J. L. Segura, V., Podhorski, A., Guruceaga, E., Mato, J., Martinez-Cruz, L. A. Corrales, F. J. and Rubio, A. 2005. Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans. Computat. Biol. Bioinform. 2*, 4, 330–338.

Shaffer, L. G. and Bejjani, B. A. 2006. A cytogeneticists perspective on genomic microarrays. *Hum. Reprod. 15*, 1, R57–R66.

Shi, Y., Simpson, P. C., Scherer, J. R., Wexler, D., Skibola, C., Smith, M. T., and Mathies, R. A. 1999. Microfabricated capillary electrophoresis amino acid chirality analyzer for extraterrestrial exploration. *Anal. Chem. 71*, 5354–5361.

Shmulevich, I., Dougherty, E. R., and Zhang, W. 2002. Gene perturbation and intervention in probabilistic boolean networks. *Bioinform. 18*, 10, R57–R66.

Silzel, J. W., Cercek, B., Dodson, C., Tsay, T., and Obremski, R. J. 1998. Mass-sensing, multianalyte microarray immunoassay with imaging detection. *Clinical Chem 44*, 2036–2043.

Stagni, C., Guiducci, C., Benini, L., Riccó, B., Carrara, S., Samor, B., Paulus, C., Schienle, M., Augustyniak, M., and Thewes, R. 2006. CMOS sensor array with integrated A/D conversion based on label-free capacitance measurement. *J. Solid-State Circ. 41*, 12, 2956–2964.

Stagni, C., Guiducci, C., Lanzoni, M., Benini, L., and Riccó, B. 2005. Hardware-software design of a smart sensor for fully-electronic DNA hybridization detection. In *Proceedings of the (IEEE) Design, Automation and Test Conference, 3*, 198–203.

Stagni. C., Guiducci, C., Riccó, B., Carrara, S., amd M. Schienle, C. P., and Thewes, R. 2006. Fully-electronic label-free DNA sensor chip. *IEEE Trans. Circ. Syst. I*.

Storey, J. D. and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc. Not. Acod, Sci.* (PNAS) *10*, 16, 9440–9445.

Tan, P., Donwney, T. J., Spitznagel, E. L., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., and Cam, M. C. 2003. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res. 31*, 5676–5684.

Thorisson, G. A., Smith, A. V., Krishnan, L., and Stein, L. D. 2005. The International HapMap Project web site. *Genome Res. 15*, 1592–1593.

Thorp, H. H. 1998. Cutting out the middleman: DNA biosensors based on electrochemical oxidation. *Trends Biotechn. 16*, 117–121.

Thrush, E., Levi, O., Ha, W., Wang, K., Smith, S. J., and Harris, J. S. 2006a. Integrated biofluorescence sensor. *J. Chromatography A 1013*, 103–111.

Thrush, E., Levi, O., Ha, W., Wang, K., Smith, S. J., and Harris, J. S. 2006b. Integrated biofluorescence sensor. *Journal of Chromatography A 1013*, 103–111.

Thrush, E., Levi, O., Smith, L. J. C. J. D. A. K. S. J., Moerner, W. E., and Harris, J. 2005. Monolithically integrated semiconductor fluorescence sensor for microfluidic applications. *Sens. and Act. B 105*, 2, 393–399.

Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B., and Hide, W. A. 2005. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res. 33*, 5, 1544–1552.

Turner, F., Clutterbuck, D., and Semple, C. 2003. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol 4*, 11, R75.

Tusher, V. G., Tibshirani, R., and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. 98*, 9, 5116–5121.

Tyson, J., Chen, K., and Novak, B. 2001. Network dynamics and cell physiology. *Nat. Rev. Molec. Cell. Biol. 2*, 908–916.

Uslu, F., Ingebrandt, S., Mayer, D., Bcker-Meffert, S., Odenthal, M., and Offenhusser, A. 2004. Label-free fully electronic nucleic acid detection system based on a field-effect transistor device. *Biosen. Bioelectron. 19*, 12, 1723–1731.

Wang, H., Wang, W., and ans P.S.Yu, J. Y. 2002. Clustering by pattern similarity in large data sets. In *Proceedings of ACM SIGMOD*.

Watson, J. D. and Crick, F. H. C. 1953a. Genetical implications of the structure of deoxyribonucleic acid. *Nature 171*, 964–967.

Watson, J. D. and Crick, F. H. C. 1953b. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature 17*, 737–738.

Webster, J. R., Burke, M. A. B. D. T., and Mastrangelo, C. H. 2001. Monolithic capillary electrophoresis device with integrated fluorescence detector. *Anal. Chem. 73*, 7, 1622–1626.

West, M., Ginsburg, G. S., Huang, A. T., and Nevins, J. R. 2006. Embracing the complexity of genomic data for personalized medicine. *Genome Res. 16*, 559–566.

Willner, I., Patolsky, F., Weizmann, Y., and Willner, B. 2002. Amplified detection of single-base mismatches in DNA using microgravimetric quartz-crystal-microbalance transduction. *Talanta 56*, 847–856.

Woolley, A. T., Sensabaugh, G. F., and Mathies, R. A. 1997. High-speed dna genotyping using microfabricated capillary array electrophoresis chips. *Anal. Chem. 69*, 2181–2186.

Wu, G., Ram, D., Hansen, K., Thundat, T., and Majumdar, R. J. C. A. 2001. Bioassay of prostate-specific antigen (PSA) using microcantilevers. *Nature Biotechn. 19*, 856–860.

Xu, C., Li, J., Wang, Y., Cheng, L., Lu, Z., and Chan, M. 2005. A CMOS-compatible DNA microarray using optical detection together with a highly sensitive nanometallic particle protocol. *IEEE Electron. Device Lett. 26*, 4, 240–242.

Xu, T., Hwang, W., Su, F., and Chakrabarty, K. 2007. Automated design of pin-constrained digital microfluidic biochips under droplet-interference constraints. *ACM J. Emerg. Techn. Comput. Syst. 3*, Article 14.

Yoon, S. 2006. Technologies and analysis methods for detecting gene expression by DNA microarrays. *IEEE Techno. Surv.*

Yoon, S., Benini, L., and Micheli, G. D. 2005. Finding co-clusters of genes and clinical parameters. In *(IEEE-EMBS). 27th Annual International Conference of the Engineering in Medicine and Biology Society*. 906–912.

Yoon, S. and de Micheli, G. 2006. Computational identification of micrornas and their targets. *Birth Defect Res. (Part C) 78*, 2, 118–128.

Yoon, S., Nardini, C., Benini, L., and de Micheli, G. 2005. Discovering coherent biclusters from gene expression data using zero-suppressed binary decision diagrams. *IEEE/ACM Transa. Computati. Biolo. Bioinform. 2*, 3.

You, L. 2004. Toward computational systems biology. *Cell Biochem. Biophy. 40*, 2, 167–184.

Zhu, H. and M.Snyder. 2003. Protein chip technology. *Curr. Opini. Chem. Biol. 7*, 55–63.

Zhu, H., Klemic, J. F. 2000. Analysis of yeast protein kinases using protein chips. *Nature Genetics 26*, 283–289.

Received October 2007; revised December 2007; accepted December 2007 by Krishnendu Chakrabarty