An introduction to DNA microarrays and some mathematical challenges behind them

Jean-Philippe Vert (Ecole des Mines de Paris)

（日本語要旨）

DNA マイクロアレイ技術および関連する数学的課題

原稿作成：阿久津達也（京大化学研究所）

DNA には多くの遺伝子が含まれているが、各遺伝子にはタンパク質に関する情報が記述され、必要に応じて遺伝子からタンパク質が合成される。生物の理解のためには、遺伝子がどのように制御されるか（すなわち、各タンパク質が、いつ、なぜ、どのように合成されるか）、各タンパク質はどのような機能を持つか、タンパク質どうし、もしくは、タンパク質が周囲の物質や刺激とどう相互作用するか、を明らかにすることが必要である。近年、開発された DNA マイクロアレイや DNA チップといった技術を用いることにより、様々な条件下において、各生物のほぼすべての遺伝子の発現（タンパク質の生成）量を間接的に同時に観測することができるようになりつつあるため、その情報解析手法の開発も重要な課題となっている。

1. DNA マイクロアレイ技術

DNA マイクロアレイは、あらかじめ各遺伝子配列より作成した一本鎖のプローブ DNA をガラス基板などの上に配置したものである。一方、細胞よりメッセンジャー RNA を通じて抽出した一本鎖の相補 DNA(cDNA) を蛍光物質などでラベルしておき、これらの cDNA とマイクロアレイ上のプローブ DNA をハイブリダイゼーションと呼ばれる技術を用いて結合させることにより、メッセンジャー RNA の量を間接的に測定する。

2. データの正規化

DNA マイクロアレイによる実験データは、画像データとして得られるため、まず、画像処理を行い各遺伝子に対応する領域の強度を求める。しかしながら、遺伝子により強度やその分布に大きな違いがあるため、データの正規化が必要となる。そのために、リファレンスとなる状態のデータと観測したデータを比較し、さらに統計解析などを行うことにより正規化するという手法が広く利用されている。

3. 教師なし学習

通常、数十から数百程度の異なる環境下で数千から数万個の遺伝子の発現量を観測する。これらの発現データの解析法としてクラスタリングと呼ばれる教師なし学習法（分類法）が広く利用されているが、その利用法は大きく二種類に分けられる。一つは遺伝子発現量の違いによる細胞の分類であり、もう一つは遺伝子発現量の変化の比較による遺伝子の分類である。前者は数万次元空間における数百点の分類に相当し、後者は数百次元空間における数万点の分類に相当する。クラスタリング手法は、階層的クラスタリング法と分割に基づくクラスタリング法に大きく分類されるが、どのようなクラスタリングを用いれば良いかはよくわかっていない。

4. 教師あり学習

遺伝子発現データは、がん細胞の分類にも利用できるが、その際には学習データの一部としてあらかじめ分類結果が与えられることが多い。このような場合の学習は教師あり学習と呼ばれる。教師あり学習の場合には、学習データに対するエラーが最小となる仮説を計算するといった戦略がよく用いられるが、発現データの場合には、数千から数万次元のデータを扱うため、過学習を起こし易い。高次元における学習は重要な数学的研究課題である。

5. システム生物学

遺伝子やタンパク質などの制御関係やパラメータを明らかにし、生物の数理モデルを作成し、より正確なシミュレーションを行うことは今後の重要な研究課題であり、システム生物学の名のもとに研究が進展しつつある。制御関係やパラメータの推定のためにベイジアンネットワークなどの確率モデルや微分方程式に基づく方法が研究されている。しかしながら、より多様なデータに対する数理モデルや推定方式など、研究すべき課題は多い。

An introduction to DNA microarrays and some mathematical challenges behind them

Jean-Philippe Vert
Ecole des Mines de Paris
35 rue Saint-Honore
77300 Fontainebleau
Jean-Philippe.Vert@mines.org

In each living organism, DNA contains a number of genes, ranging from several hundreds for simple bacteria to several tens of thousands for mammalians. Each gene encodes a protein, which can be synthesized when required from its blueprint on the DNA. In current biology, understanding living systems means to a large extent deciphering how genes are regulated (i.e., when, why and how proteins are synthesized), what are the functions of the proteins synthesized, and how they interact with each other and with their environment to form a living system.

A number of technological advances in the last two decades have contributed to provide answers to these questions. The genome sequencing technology enables to read the total DNA of any organism, including humans, and to detect genes. In order to characterize the functions of these genes and their regulations, a recent technology is playing a central role and is expected to be of increasing use in the coming years: DNA microarrays, which are the focus of this paper. DNA microarrays, or DNA chips, enable the monitoring of the quantity of messenger RNA[1] simultaneously for all the genes of a genome in a given condition. It has the potential to provide massive amounts of data about gene expression, and represents an invaluable analytic tool to help decipher the mechanisms behind life, at least at the gene expression level.

In this introductory paper, we first review the DNA microarray technology itself, and then present an overview of classical analysis performed on expression data, ranging from differential analysis of single genes to unsupervised clustering and supervised classification of genes or cells, and to the prediction of genome-wide regulatory systems. Rather than focusing on the relatively simple and classical statistical techniques used for these analysis, our goal is here to convince the reader that the DNA microarray technology is a breakthrough likely to completely modify our vision of living systems, and that new formalisms and mathematical tools need to be developed in order to be able to manipulate gene expression data and to model living systems.

1 The DNA microarray technology

The central dogma of molecular biology states that DNA, which carries the genetic information of living cells and organisms, is transmitted between generations, and that the information it contains it expressed when RNA molecules are synthesized and translated

[1]The intermediary molecule between DNA and a protein. When a proteins needs to be synthesized, the part of the DNA which contains its blueprint is copied into a RNA molecule (transcription), and the protein is synthesized by processing the RNA (translation).

into proteins. Proteins are ubiquitous molecules performing various tasks such as catalyzing chemical reactions, transmitting information or participating in structural components of the cell. The number of different proteins encoded in a genome varies from several hundreds for simple bacteria to several thousands for the budding yeast, to several tenth or hundreds of thousands for the human genome. While many genes, i.e., parts of the DNA which contain the information for the expression and the structure of a protein, have been identified thanks to the sequencing of a number of model organisms, the question of how these hundreds of thousands of proteins enable a human to live is still far from being understood. To help answering this vast question, being able to measure the quantity of all proteins in real time in a living cell would be useful. This remains impossible with the current technologies, but DNA microarrays provide a useful alternative. This technology is a tool to observe when and where genes are expressed, and represents the first analytical tool to monitor simply and on a large scale the RNA content of a cell, also called the transcriptome.

Array technologies encompass a wide class of technologies which monitor the combinatorial interaction of a set of molecules, such as DNA, RNA fragments or proteins, with a predetermined library of molecular probes. DNA microarrays, or DNA chips, are a particular class of arrays which measure the quantity of messenger RNA present in a living cell. Roughly speaking, a DNA chip is made of a small surface, usually made of glass or nylon membrane, on which a large number of known DNA molecules, called probes, are attached. The surface is usually divided as a grid into hundreds or thousands of cells, and each cell contains a large number of replications of a unique probe.

Historically, DNA chips are descendant of the Southern blot technique developed by Ed Southern more than 25 years ago [Sou75]. This first array was based on the observation that labeled nucleic acid fragments (e.g., single-stranded DNA fragments) could be used to detect complementary sequences attached on a solid support by hybridization. Hybridization refers to the fact that two complementary single-stranded nucleic acid molecules naturally form a double-helix maintained by hydrogen bonds between complementary bases. A Southern Blot experiments consists in fixing a large number of single-stranded DNA fragment extracted from a cell on a support, and trying to hybridize a radioactive genetic probe, i.e., a known single-stranded DNA sequence, to all DNA fragments fixed on the support. Following hybridization, an X-ray picture of the support highlights the fragments which hybridized with the probe, which indicate that they contain a sub-sequence complementary to the genetic probe.

DNA chips rely on the same principle of hybridization of nucleic acid molecules fixed on a support by labeled probed. The main differences are that the fixed nucleic acids are designed by the experimenter (e.g., to match known genes of an organism), and that their number can reach several tenth or hundreds of thousands. More precisely, starting form a living cell, all the messenger RNA are extracted, copied into complementary DNA strands and labeled with a small fluorescent chemical. The result is put in contact with a DNA chip, on top of which a large number of known DNA probes have been attached. After some time, when all molecules have enough time to visit all cells in the chip, those single-stranded solution cDNA which find a complementary DNA probe naturally hybridize to it to form a double-stranded DNA. After washing the solution, only the molecules which hybridized remain on the chip (see Figure 1). It is then possible to measure the quantity of hybridized material on each cell by detecting the quantity of fluorescent material in each cell. This gives an estimation of the quantity of messenger RNA in the initial living cell for all genes corresponding to the probes simultaneously. As an example, it is nowadays possible to buy chips with probes corresponding to all genes of the budding yeast, the fly

or humans, or cheaper chips more specialized with only those human genes known to play a role in a certain diseases like cancer.

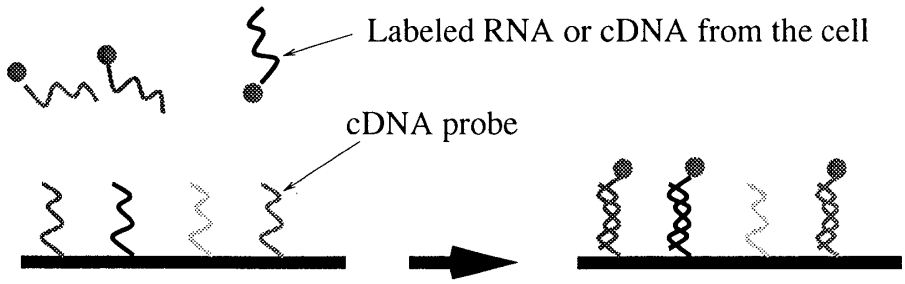

Figure 1: A DNA chip is made of a support where known probes are attached. When a solution containing RNA or cDNA extracted from a cell and labeled is in contact with the DNA chip, complementary strands hybridized and can be recognized by checking which probes form a labeled double-strand

Two mainstream strategies can be followed to manufacture DNA chips. One option is to individually synthesize each probe directly on the surface, i.e., to add the right nucleotides one by one incrementally in order to get the correct DNA sequences attached at the right position on the surface. This option is for instance used in the photolithographic method developed by Fodor [FRP+91] and commercialized by Affymetrix, Inc. It can be used to synthesize probes of up to 20-30 nucleotides, and is based on an efficient method for high density spatial synthesis of oligonucleotides. A second option is to first pre-synthesize the oligonucleotides of interest (e.g., using the PCR technology), and then to fix them on the chip. This technology enables the use of longer oligonucleotides (usually 100-5000 bases long), and has been popularized by the Patrick O. Brown laboratory at Stanford University who provides a methodology to manufacture affordable arrays [SSDB95]. This option has been very popular among academic research laboratories.

The result of an hybridization experiment by DNA chip is usually an image which reflects the quantity of hybridized material in each cell by the color or intensity of each spot. Various image analysis techniques enable to automatically isolate each cell on the image, and estimate the quantity of hybridized material. The result is therefore a series of numbers which show a global picture of the transcriptome of a cell at a given instant.

The transcriptome, i.e., the quantity of various messenger RNA in a cell, contains a lot of information about the state or the origin of a cell. While (almost) all cells of a human have the same DNA material, the transcriptomes of a skin cell and of a neural cell are likely to be very different, because each cell expresses the genes from DNA to RNA, and then to proteins, only for the proteins it needs. As a result, observing the transcriptome of a cell gives a lot of information about the origin and function of the cell. For a given cell type, the transcriptome is also likely to vary over time. For example, most cells sometimes grow and sometimes divide into two children cells. The proteins required during these different stages are obviously different, and the transcriptome reflects these differences at the expression level. An other interesting application of DNA chips is to observe diseases at the transcriptome level. For instance, many cancers seem to have very typical signatures at the transcriptome level, and can be observed or predicted with DNA chips. Finally, DNA chips are a revolutionary analytical tool to understand the behavior of living cells in terms of gene expression and regulation. By carefully designing a series of experiments where cells are submitted to various experimental conditions and performing DNA chip experiments at each stage, one can observe the variations of gene expression

between experiments and infer unknown regulation mechanisms or gene functions. This is certainly one of the most exciting applications of DNA chips in the coming years.

In terms of applications, DNA microarrays have tremendous potential in the drug discovery process, in particular to identify new drug targets such as genes over-expressed in a given disease, or to observe the reaction of an organism to a given therapy; in disease diagnosis, as many diseases are likely to be observable at a very early stage through gene profiling experiments, in which case they can often be better treated by traditional medicine than if they were discovered later; as a tool to help decide which drug is the more appropriate for a given patient, as many drugs are known to work only on a small population, which might be characterized at the transcriptome level; finally, in many areas of biology, in particular in systems biology which consists in considering a cell as a complex system of interacting elements.

2 Data normalization and single gene analysis

The result of a gene profiling experiment is an image, which can be translated into a series of numbers by various image processing methods. Typically, these numbers are the average intensity of the images on each cell, which is an increasing function of the quantity of hybridized material on the cell. In order to transform this numbers into estimates of the actual quantity of RNA, a calibration has to be carried out, as the relation between hybridization is not linear. For some chips, one can hybridize the RNA of two cells in different conditions simultaneously on the same chip, with two different colors as labels. In that case one gets a single spot for each probe, characterized by a color and an intensity. Here again, the relationships between quantities or RNA hybridized from each cell, the color and the intensity of the spot are not linear. The calibration is usually performed with statistical tools for regression and result in an hopefully unbiased estimate of the the quantity of hybridized RNA, or sometimes of the ratio of hybridized RNA between two cells.

The first direct use of these estimates is to check which genes have a very different expression level between two conditions, such as cells extracted from a metastatic versus non-metastatic derivatives or a tumor cell line. Typical applications consist in selecting a small set of genes observed to by particularly over or under-expressed in one of the conditions, in order to further analyze them using conventional methods. The revolution in biological research is that with the microarray technology, one get an objective and unbiased view of all genes in the same time, and is free to further analyze genes which never drew the attention of the medical community before.

3 Non-supervised clustering

While single gene analysis is by far the first use of microarrays in biomedical research nowadays, the availability of the expression levels for a large number of genes simultaneously suggests that a lot can be learned about the relationships between genes or between cells. Mathematically speaking, a gene profiling experiment characterizes a tissue sample or a set of cells by a point in a high-dimensional vector space (typically, with 1,000 - 100,000 dimensions). By observing the points for a number of gene profiling experiments, such as multiple time points from multiple cell lines treated independently with multiple growth factors, one can observe various correlations or similarities among genes and among samples.

A striking result of the first published gene profiling experiments was that very often, many genes seem to follow similar patterns of expression between various conditions, i.e., many subsets of genes form "clusters" when represented by vectors where each coordinate is the gene expression in one experiment. Biologically, one say that the genes are co-expressed, i.e., they are expressed and inhibited in the same time. This phenomenon is well-known for example in prokaryotes, where it is common to have several genes forming operons and being co-regulated. Even though operons don't exist in eukaryotes, such as humans, it turns out that apparent co-regulation is striking in many cases. Following this observation, a natural analysis to start understanding the relationships between genes is to cluster them in groups with similar expression profiles. Similarly, when a number of gene profiling experiment are performed, and when each experiment is seen as a high-dimensional vector of gene expressions, one can study the relative positions of these vectors and look for clusters which would correspond to samples with similar transcriptomes.

Clustering is useful as a visualization tool, and to quickly detect experimental artifacts, classes of cells (such as different types of cancer), or families of related genes (as co-regulated genes often participate to common biological processes). Clustering is performed almost systematically before any further analysis, because it can help getting a global vision of the data available.

Any introductory book on data mining describes various families of clustering algorithms, so we refer the interested reader to such references for more details about the algorithms. Roughly speaking, a clustering algorithm involves a distance measure for the objects to be clustered, which can be for instance the Euclidean distance between vectors in our case. Then a distance between sets of points must be defined; three classical variants include the single, average or complete linkage methods, where the distance between two sets of objects is respectively defined as the minimum, the average or the maximum distance between the points of each sets. From these basic ingredients, a number of techniques exist to obtain groups of similar points, some called hierarchical methods providing a hierarchy of clusters (from singletons to the whole set), other called partitioning methods providing only a set k groups, where k is pre-defined by the user. A classical hierarchical method is for instance the hierarchical clustering method, which starts from the set of all singletons, and then iteratively merge the two closest clusters together in order to get a hierarchy of clusters which can be represented as a dendogram. A classical example of partitioning method is the k-means clustering algorithm, which iteratively choses a predefined number of centroids in the space of objects, assigns each object to the closest centroid, adjust centroids as the centers of each obtained clusters, and iterates until convergence of the centroids.

Although clustering provides an appealing tool to quickly detect structures in the data, it has also pitfalls which are very often not known or misunderstood by practitioners. First, as the methods shortly described above suggest, a clustering algorithm always output a clustering of the points, whether or not a true underlying structure exists. Second, there is a lot of arbitrary in the choice of the method, of the distance between points, and results can differ a lot between different choices. Third, if several natural cluster structures exist among points (e.g., cells can come from male/female, sane/sick organisms, and be studied by different researchers), it is not clear what structure will be detected by a clustering algorithm. This is particularly problematic when few points are clustered in large dimension, which is often the case when samples are clustered from microarray data. Finally, while it is obvious mathematically speaking that gene clustering and tissue clustering belong to the same class of problems and can be tackled with the same techniques, there is however one difference which has almost never been pointed out: as the number of experiments is

usually orders of magnitude smaller than the number of genes (e.g., 100 experiments for 10,000 genes), clustering genes means clustering many points in a low-dimensional vector space, while clustering tissues means clustering few points in a high-dimensional space. It is then far from being obvious which clustering methods are relevant in which case.

4 Supervised classification

While clustering refers to the analysis of the positions of the points in a sometimes high-dimensional space, and to the discovery of possible hidden structures in the set of points (more precisely, a set of clusters or a hierarchy of clusters), different need arise when one wants to use microarrays to discriminate between two or more known classes, such as two types of cancers for tissue samples or functional classes for genes. In that case, a number of examples with a known class are given, and the goal of the analysis is to learn from these examples a rule or function to predict the class of any future examples. A straightforward example of classification problem is the development of a diagnosis tool for a type of cancer from microarray measurements: given the measurements for 50 patients with a given cancer, and 50 healthy patients, one need to find a rule which will be able to predict whether a new patient has a cancer or not from a simple gene profiling experiment.

This task is called (supervised) classification, as opposed to (unsupervised) clustering. In simple cases, clustering might be enough to discover that there are clearly two separate groups of points, which might correspond to two classes one would like to learn. However, in most situations, clustering algorithms are likely to discover clusters which do not correspond exactly to the separation into classes one is interested in, and supervised classification will work better (see a toy example in Figure 2).

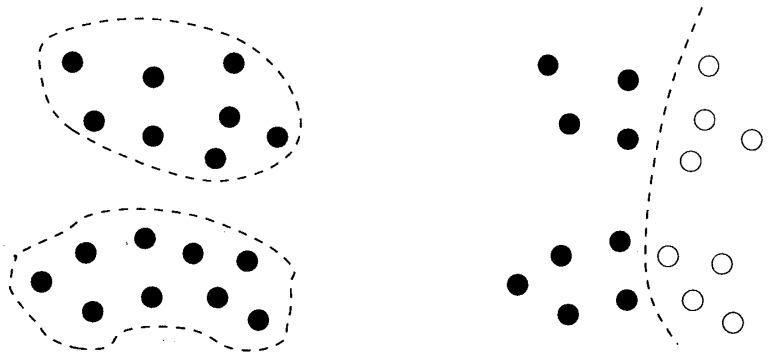

Figure 2: Difference between clustering and classification. On the left, only the positions of the points are given, and a natural separation of the points in two clusters can be found by clustering algorithms. On the right, each point has an associated label (black or white). The goal of classification is to detect a discrimination rule between each class of points.

Supervised classification has been an important research topic in the machine learning, artificial intelligence and statistical communities during the last decades, and a impressive list of methods have been developed. Rather than listing all methods, we limit ourselves in this contribution to a rapid overview of various issues is supervised classification, and invite the interested reader to consult more specific textbooks [Vap98, HTF01]. Roughly speaking, an algorithm for supervised classification observes a set of points together with their classes, and then picks a function which maps any possible object into a class. If we note $\mathcal{X}$ the space of objects, $\mathcal{A}$ the finite set of classes, then such an algorithm is defined by a set of functions $\mathcal{H} \subset \mathcal{X}^{\mathcal{A}}$ among which the algorithm can chose, and a mapping

$(\mathcal{X} \times \mathcal{A})^n \to \mathcal{H}$ (for any $n \geq 0$) which indicates which function in $\mathcal{H}$ is picked by the algorithm after seeing n objects and their classes. A convenient framework to study and design learning algorithms is to suppose that observations are independent realizations of a random variable with distribution P on $\mathcal{X} \times \mathcal{A}$, and that future data to be classified are also realizations of the same random variable. Under these hypotheses the performance of a any classifier $g \in \mathcal{H}$ can be quantified by its probability of mistake $R(g) = P(g(X) \neq Y)$, also called its risk, and the goal of a learning algorithm is to chose a function $\hat{g}$ with a risk as small as possible. However P is unknown a priori, and is only known through the observation of the n points sampled independently according to it. In particular, even though the risk $R(g)$ of a function $g \in \mathcal{H}$ is unknown, one can measure its empirical counterpart defined by $R_{emp}(g) = \frac{1}{n} \sum_{i=1}^{n} 1(Y_i \neq g(X_i))$. The main motivation behind the empirical risk is that, by the classical law of large numbers, for any $g \in \mathcal{H}$, the empirical risk $R_{emp}(g)$ converges almost surely to the risk $R(g)$. As a result, for a given set of n observations, it seems natural for a learning algorithm to chose one of the functions $g \in \mathcal{H}$ which minimizes the observable empirical risk. This very general approach is called *empirical risk minimization* (ERM), and is implemented under various forms in many learning algorithms.

The ERM principle, however, is not sufficient to ensure that one has a good learning algorithm. Suppose for example that the class of functions $\mathcal{H}$ is very large, perhaps equal to $\mathcal{X}^{\mathcal{A}}$. Then one can always find a function $g \in \mathcal{H}$ with very small empirical risk, which might not generalize well to unseen data (take for instance the function $g(x) = y$ if x has been observed with the class y, 0 otherwise). In this case, one talk about *overfitting*, which refers to the fact that the algorithm fits too much the observed data. The first main contribution to the theoretical analysis of this issue was the work of Vapnik and Chervonenkis in the 1970's, which observed that even though the law of large numbers ensures that $R_{emp}(g) \overset{a.s.}{\to} R(g)$ for each $g \in \mathcal{H}$ individually when the number of observations tends to infinity, it is not always true that $R_{emp}(\hat{g}) \overset{a.s.}{\to} R(\hat{g})$ where $\hat{g}$ is chosen in $\mathcal{H}$ by the ERM principle. The reason is that this is only true if one can ensure a sort of law of large numbers uniformly over the set $\mathcal{H}$, and Vapnik and Chervonenkis gave precise conditions for this law to hold. Intuitively, the conditions are expressed in terms of a measure of the size of the set $\mathcal{H}$ (called the VC dimension), and there is an equivalence between consistency of the ERM principle (i.e., $R_{emp}(\hat{g}) \overset{a.s.}{\to} R(\hat{g})$) and finiteness of the VC dimension of $\mathcal{H}$. Consequences of these results had huge influence in the design of learning algorithms in the last two decades: indeed, they show that a good algorithm must not only find functions with small empirical risk, but also control the complexity of the class of functions $\mathcal{H}$ it can pick. Finding a trade-off between these two constraints has been a major research topic in statistical learning theory recently.

The goal of this overview of the theory behind supervised classification was to convince the reader that it is not such an easy task, and that overfitting in particular is a dangerous phenomenon which often occurs when one tries to develop a complex learning algorithm to "mimic" nature, for instance. In particular, supervised classification for objects in a high-dimensional vector space, such as tissue samples characterized by tens of thousands of gene expressions, turns out to be a very difficult task in theory, particularly when a small number of samples are available. Theory would certainly consider making a good diagnosis tool from the observation of the gene expressions of 100 patients, for instance, an impossible task. However, this is typically a situation encountered in DNA microarray analysis, so tools and theory need to be developed in this context. A general question motivated by gene expression data, which is likely to require new mathematics and to motivate much research in machine learning and mathematical statistics in the coming

years, is therefore the following: how to learn and to perform statistical estimation when the number of points available is much smaller than the dimension of the space they live in?

5 Systems biology

At a higher level of abstraction, a major direction of biomedical research in the coming decades is likely to concern the modeling, understanding and simulation of biological systems involving a large number of elementary parts interacting together. The set of genes and of chemical compounds in a cell constitutes a natural basis to model life by modeling interactions among these elements, including gene regulation, catalysis of chemical reactions by enzymes, information transmission, physical interactions etc...

In order to develop a satisfactory and useful model of such biological systems, one needs a theoretical framework to represent mathematically the biological phenomena to be modeled, and experimental data to calibrate and confirm candidate models. With the development of the DNA microarray technology and other recent high-throughput technologies, experimental data seem to be preceding the development of satisfactory mathematical frameworks to incorporate them. The development of such a framework represents, to my opinion, one of the greatest challenges of biology in the post-genomics era, which can only be tackled with the participation of mathematicians coming from different disciplines, and which is likely to boost the research in new areas of mathematics.

The task is ill-posed and probably difficult. A number of biological evidences suggest various relationships among basic biological objects: genes have evolved from common ancestors during evolution; we know several examples of typical gene expression regulation mechanisms; the 3D structure of all molecules is known to play a crucial role in biological process, almost always based on physical interactions between molecules; global interaction or regulation networks are known to be very complex but seem to have typical topological structures; large biological systems seem to be very stable and resistant to variations in the environment (except during such events as death or development of a cancer), but individual molecules are sometimes very sensitive to tiny variations (e.g., the function of a protein can change completely when one out of several thousands amino acids is modified); etc. This list of biological evidences is far from being complete, but highlights the diversity of observations and evidences available today. A satisfactory mathematical framework for systems biology should be able to include these evidences, and many others.

With microarrays, massive data sets of gene expression levels can easily be generated. By submitting a cell to various experimental conditions, one can observe the evolution of the expression of all genes simultaneously, and observe correlations among genes or typical patterns of expression. To incorporate these data into a mathematical model, the simplest formalism is to consider the set of genes as a finite set $\mathcal{G}$, and a gene profiling experiment as a vector $v \in \mathbb{R}^{\mathcal{G}}$. DNA microarrays enable to study the evolution of v along different experiments, and to study the properties of the trajectories of v. Much research has been carried out in the recent years with this goal. On the one hand, several groups have proposed to model the evolution of the transcriptome as a dynamic system, satisfying an evolution equation of the form:

$$\frac{dv}{dt} = A(v(t)).$$

Various levels of complexities have been investigated for such models [BB01], ranging from boolean networks where v is a vector of binary numbers (each gene is considered expressed

or inhibited) and time is discretized, to continuous-time systems for real-valued vectors such as S-systems, defined by the following evolution equation:

$$\frac{dv_i}{dt} = \sum_k T_{i,k} \prod_j v_j^{g_{ijk}} - \sum_k U_{i,k} \prod_j v_j^{h_{ijk}} + I_i(t).$$

For each formalism, parameters of the evolution equation must be inferred from the observation of gene expression profiles on different experiments. Similar to problems which arise in supervised classification, the task is difficult in theory when not enough data are available. Choosing complex model is likely to enable a better approximation of a "true" dynamic system underlying gene expression evolution (for example, S-systems have universal approximation properties, while boolean models are obviously too restricted to represent a satisfactory model of gene expression evolution). However, learning parameters in S-systems is much more difficult than in a boolean model setting, and overfitting is more likely to occur. While much research has been devoted to these models, only limited success has been obtained, mainly for small models of the best studied regulatory switches in bacteria. As biological evidences suggest that the actual regulation of a single gene often involves a considerable number of other genes, such as transcription factors, as well as many other variables not observable, it seems that the "true" model itself is pretty complex, and one can be skeptical about the capacity of the dynamic system approach to uncover the "true" regulation mechanism in the short term.

An other school of thoughts worth mentioning is the probabilistic approach, which makes no dynamic system hypothesis but focuses on the characterization of the repartition of experiments in the high-dimensional vector space. In that case, the mathematical framework is still based on the discrete set of genes $\mathcal{G}$, but the regulation process is modeled by a probability measure on $\mathbb{R}^{\mathcal{A}}$. With the (dangerous) hypothesis that various gene profiling experiments are independent realization of a random variable in $\mathbb{R}^{\mathcal{A}}$, one can try to estimate this distribution. In particular, this is a way to detect correlations between several coordinates, i.e. , between the expression of several genes. This approach has been implemented recently with Bayesian graphical models [FLNP00], which enable to factorize a probability distribution for a high-dimensional variable through low-order correlations. Learning a graphical model from expression data results in a graph where genes are the nodes, and where cliques indicate the low-order correlations involved in the distribution learned. This has proved to be useful to detect regulatory relationships between genes, but also faces the formidable challenge of learning a distribution in high dimension for a very limited number of observations. Moreover, transforming correlations into causation is a difficult challenge faced by any probabilistic approach.

These two examples, the dynamic system and the probabilistic approaches, are just two illustrations of recent developments in mathematical modeling of biological systems. As data available in these cases come from microarray data, the formalism underlying this approaches is simply to consider the genes as a finite set, and the expression profiles as vectors. However, this approach is clearly limited to the analysis of the transcriptome, and is subject to many refinements in the future to incorporate more biological evidences as well as other types of data (such as metabolic pathway maps, structural or sequential information etc...). As interesting examples of different approaches to manipulate biological objects, one can cite for example the use of operators algebra [Kat01] or the development of inductive informatics using the ETS model for structural representation [GGK00]. Even though far from being mature, these attempts are very promising and suggest that important developments are going to result from the confrontation of post-genomics biology and mathematics in the coming year, for the benefits of both disciplines.

6 Conclusion

The DNA microarray technology, together with several other high-throughput technologies, is deeply affecting the outlook of biological research. It can provide a view of the transcriptome and its evolution, and represents an invaluable tool which is modifying our view of biological systems as well as the way biological problems are approached. While the first applications of this technology currently mainly focus on detection of over- or under-expressed genes in various conditions, and on further analysis of this genes using traditional tools, deeper understanding of the set of genes and their relationships are being obtained through unsupervised clustering, supervised classification or modeling of gene regulatory systems. These research directions, however, require new mathematical tools, likely to be more and more important in many scientific fields where high-throughput data generation technologies are emerging. Performing data mining or statistical inference in very large dimension remains theoretically difficult, but has to be performed. This issue is currently a major driving force in several fields related to learning theory, including machine learning and mathematical statistics.

As an analytical tool to observe the transcriptome, DNA chips have fostered the development of methods to decipher the gene regulation mechanisms. However these methods have still limited successes, and suggest that the transcriptome is only one projection of a much more complex object, a living organism. There are today no satisfactory formalism to describe, manipulate or simulated such living systems, and which could serve as natural formalisms to integrate not only gene expression data but also all sorts of data about genes, metabolisms, interactions, reactions to environment etc... The development of such formalisms is likely to be a *sine qua non* condition to achieve the promises of post-genomics, which is likely to become a discipline at the frontier of traditional biology, computer science and mathematics.

7 Acknowledgments

I am grateful to Kenji Ueno and Tsuyoshi Kato who offered me the opportunity to participate to the workshop "Mathematical aspects of molecular biology: toward new mathematics" that was held in Nara, Japan, on January 24-27, 2003. This text is based on the talk I gave there.

References

[BB01] James M. Bower and Hamid Bolouri. *Computational modeling of genetic and biochemical networks.* MIT Press, Cambridge, MA, 2001.

[FLNP00] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

[FRP+91] S.P. Fodor, J.L. Read, M.C. Pirrung, L. Stryer, A.T. Lu, and D. Solas. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251:767–773, 1991.

[GGK00] Lev Goldfarb, Oleg Golubitsky, and Dmitry Korkin. What is a structural representation? Technical report, University of New Brunswick, 2000. Technical report TR00-137.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer, 2001.

[Kat01] Tsuyoshi Kato. Operator dynamics in molecular biology. Technical report, I.H.E.S., 2001. Technical report IHES/M/01/41.

[Sou75] E.M. Southern. Detection of specific sequences among dna fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98:503–517, 1975.

[SSDB95] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complimentary DNA microarray. *Science*, 270:467–470, 1995.

[Vap98] Vladimir N. Vapnik. *Statistical Learning Theory.* Wiley, New-York, 1998.