SIGNAL PROCESSING CHALLENGES IN THE POST-GENOMIC ERA

Amir A. Handzel

BG Medicine, Inc.,
40 Bear Hill Road,
Waltham MA 02451, USA
AHandzel@BeyondGenomics.com

ABSTRACT

Molecular biology has been undergoing a revolution whose most visible milestone is the complete sequencing of the human genome. This revolution has been propelled by explosive advances in technology accompanied by changes in fundamental concepts. Progress in this newly shaped field requires innovative approaches and cross-disciplinary fertilization. I highlight here several areas of research which pose interesting challenges to the signal processing community. One promising theme for success is the synergy of established signal processing techniques with specific characteristics of molecular biology. These questions also give the opportunity to develop new signal processing methods inspired by biology.

1. INTRODUCTION

The substance of life are natural polymers, a modular molecular structure whose combinatorial flexibility enables the coding of information as strings of symbols, as well as allowing the creation of an enormously diverse universe of molecules. DNA molecules made of nucleotides encode and store the information of an organism; the similar but more volatile RNA molecules are central to the production of proteins from DNA and to the regulation of this process; proteins made of amino acids constitute most of the structural and functional material in the living world; and additional molecules, including sugars and lipids, that have vital roles in cell structure, metabolism and other functions.

Since its early days, the central dogma of molecular biology was a linear flow process, whereby genes are transcribed to messenger RNA (mRNA) which then travels to the cell's factory to produce proteins which are later chemically degraded. This view has changed dramatically. We know now that there is an intricate web of feedback between various components of the above chain. Gene expression, i.e. the transcription of genes in order to produce a protein, is regulated in a highly complex and dynamic manner, ever changing as function of intra- and extra-cellular signals, and regulated by various layers of controling molecules with multiple feedback loops.

Driving this new understanding are leap advances in technology and the experimental information that they enable to collect. A motivating force for technological progress in the 90's was the goal of the complete sequencing of the human genome, and its formal attainment in 2001 is an important milestone in biology. Most of the new technologies fall into four categories: 1) DNA sequencers, 2) arrays that measure mRNA levels, i.e. gene expression, 3) analytical chemistry tools, namely mass spectrometers coupled with chromatography for separation, and 4) molecular imaging. All these are characterised by high through-put capability relative to older methods. Effort is being invested now in several directions: to increase speed and volume, to decrease the price of experiments and to develop methods that will allow for frequent measurements in time, which could open new vistas onto the dynamics of molecular systems.

Biological signal processing on the molecular and cellular levels differs from what engineers and signal processing practitioners are used to. The reasons for this are twofold. First, biological signals are not merely symbols that are manipulated in a computational way, rather they have a physical substrate that not only carries the signal, but is an inherent part of the computation itself. Molecular signaling is achieved through the matching of three-dimensional molecular structure and electric charge distribution. Second, biological systems were not designed from start to finish in a comprehensive manner, but are the result of the complex process of evolution whereby a series of changes and additions to the system brought it from an initial state to the current one through a convoluted path under various forms of selection pressure. In contrast to the usual engineering mode of operation, the task of this field is to reverse engineer and uncover a system whose working principles are unfamiliar.

In the following sections I describe several important areas of current research in which signal processing (SP) plays a role. For related aspects in terms of systems and control theory see [1].

2. SEQUENCE ANALYSIS

The functionality of proteins is largely determined by their three dimensional structure and electric charge distribution. These, in turn, are dictated by the sequence of amino acids, the "letters" that constitute a protein. Each amino acid is encoded by a triplet of DNA bases — a *codon* — in the genomic sequence. Indeed, SP has often been used to study properties of genome sequences as discrete statistical signals, using methods such as Hidden Markov Models [2]. Yet only a small fraction of the total DNA constitutes genes that encode for proteins. Various classical SP tools have been used to analyse DNA sequences and to predict the regions of protein coding genes: Fourier analysis, joint space-frequency (spectrogram) analysis [3], digital filters [4] and more. A salient signature of coding regions in the genome is a spike in the Fourier domain due to the period-three of codons and bias in nucleotide distribution in the codon map. Protein coding regions also exhibit long-range correlations, revealed as *1/f* power spectrum, which are assumed to be a signature of the mechanisms by which genomes evolved.

A different approach to sequence analysis is based on the comparison of the genomes of several species. The goal is to find segments which are evolutionarily conserved amongst different species to a degree that exceeds the average evolutionary sequence distance. Such segments may then be presumed to have been retained to perform some function, for which more direct confirmation is ultimately required. Surprisingly, numerous conserved segments have been recently discovered in mammalian genomes whose function, if any, remains a mystery [5]. Some of these may be genes that code for RNA molecules as end products, namely RNA genes. Indeed, several long-known RNAs are part of the protein production machinery (tRNA and rRNA). A parallel surprising finding has been that significant portions of the genome that are transcribed into RNA are never translated to protein products and are not known RNA genes. Called non-coding RNA (ncRNA), these molecules probably serve a diverse spectrum of vital functions, including the complex regulation and modulation of the protein production chain itself [6]. The abundance of ncRNA correlates well with an organism's complexity — very high in humans, for example. This correlation has been suggested to be a causal relation, ncRNA being responsible, in part, for an organism's complexity through regulatory control, thus allowing orders of magnitude greater flexibility in the use of an existing set of protein genes.

Current research efforts focus on charting the ncRNA world, its interactions and functions. But ncRNA segments are harder to tackle with standard SP methods, and their characteristics are more elusive than for protein coding regions. Absent are typical gene sequence statistics and the protein coding signature of codon periodicity [7]. Some functionally important ncRNA, called micro RNA (miRNA), are very short segments only 20-25 bases long. Their brevity defies analysis by means such as the Fourier transform. Complicating the task, RNA molecules often fold upon themselves creating a secondary structure, which can be critical for their function. This long range connection is reflected in their sequence (the primary structure). New, more sophisticated, SP tools are therefore required. One such candidate is a framework that generalizes HMM, called Stochastic Context-Free Grammars (SCFG) [7, 8], but much more SP development work is needed.

Substantial challenges remain also in relating amino acid sequences to protein structure and function. In this area, a recent attempt was made to identify repeat structural motifs in proteins using wavelet analysis [9].

3. BIOCHEMICAL NETWORKS: SIGNALING AND CONTROL

Given the substrate of intermolecular communication, what is the nature of the signals at the circuit and network level and how are they processed? The basic unit of communication is a discrete single molecular interaction, but often signals comprise numerous such events, amounting to a *de facto* continuous signal. The levels of various proteins and metabolites are regulated and maintained at precise quantities inside cells and up to the whole organism, e.g. blood glucose concentration. It was recently discovered that analog signals can produce digital pulses in critical cellular pathways [10]. Identifiying such new modes of cellular signal processing and deciphering the mechanisms by which they are realised in biochemical circuitry constitute fertile area for research. A novel approach in this direction is to construct cellular communication systems by design and to study their behaviour under controlled conditions [11]. This enables the estimation of system parameters.

Molecular signals are not perfect — stochasticity in molecular processes and variability in cellular conditions confound signals with noise, for example in gene expression and protein production [12]. Stochasticity is accentuated in such processes by the fact that only few molecules may be involved. Despite the ubiquity of molecular noise, organisms are not passively subject to its mercy, but rather evolved mechanisms to control and attenuate it. Correction of errors in DNA replication down to the average rate of 10^{-9} nucleotide is well known. Various mechanisms have been proposed and studied for attenuation of noise in biochemical circuits such as cascades [13] and gene expression in genetic oscillators [14]. An intriguing possibility has been raised that the effective amount of noise in gene expression and protein production are selected for evolutionarily depending on how vital the protein is to the organism's survival [15]. It is also thought that stochasticity may

confer advantage in some cases at the population level [16]. An example of noise analysis in simple circuits using the frequency domain of the system is given in [17].

Above the small circuit level we are interested in identifying complete operational modules of regulatory networks of genes and their expression, and of biochemical networks. These are often extracted from broad profile measurements of gene expression, proteins and metabolites. Experimental data are usually organized in a matrix of varibles (e.g. genes) vs. samples, where samples may be representatives of a population or taken in different conditions. A promising family of methods, called biclustering, finds meaningful submatrices in the data matrix [18]. These subclusters correspond to genes that are expressed similarly in a subgroup of samples or conditions. Some biclustering algorithms also account for issues of statistical significance [19]. Another approach is to use prior information about a network's structure [20].

The final goal is the reconstruction of the complete global network of interaction between genes, proteins and metabolites. An extensive literature exists on the subject, especially in the context of gene regulatory networks. Two examples are Relevance Networks in gene expression [21] and Correlation Networks for integrative systems biology [22].

4. STATISTICS AND DATA ANALYSIS

Analysing experimental data about systems relies heavily on statistics. Numerous measurements are collated and passed through statistical tests, such as *t*-test and ANOVA. The fundamental difficulty of the field is that in many experiments the number of measured variables is several thousands to tens of thousands, while the number of samples is only several dozen at best. The problem exhibits itself in various ways. Univariate hypothesis testing is guaranteed to result in numerous false positives because of the high number of multiple tests. Although these can be controlled by severely restricting the acceptance level in multiple tests, the challenge is to do so without discarding most of the real information in the data. False Discovery Rate (FDR) correction is an attempt to balance these two requirements. Classification and prediction based on such severe undersampling faces extreme overfitting to the sample set at hand [23, 24]. Similar problems afflict feature selection, i.e. the identification of subsets of variables which are thought to be meaningful in a process. Various attempts are being made to overcome these problems, for example by finding the relation between sample size and optimal number of features for various classifiers and class distributions [25]. Applying methods for dealing with multiple tests, such as FDR, to correlation structure of gene sets (called gene co-expression) [26] enables the reconstruction of networks, as in [21, 22].

As basic principles of molecular systems are gradually elucidated, temporal aspects are becoming more important to fully understand their dynamic operation [27]. This requires the development of suitable analysis methods [28].

5. TECHNOLOGICAL MEANS

The previous sections dealt with primary scientific questions in the field. In order to obtain useful data to study them, various technological means are used, many of which are still in the process of development. SP is essential in transforming raw instrument read-out into useful data. Arrays that measure gene expression levels (mRNA) are typically photographed and initially analysed with image processing steps. Statistical SP procedures are then applied to data before it can be used. These include the identification and removal of noise and out-lier data, scaling and more.

Deriving information by analytical chemistry tools for small and large molecules (metabolites and proteins) with mass spectrometers requires a host of steps: separating spectrometric peaks of individual molecules, alignment of peaks of the same molecule in different spectra (coming from different biological samples), pattern matching of measured spectra to spectra in data bases, and data normalization.

Signal processing will also play a critical role in completely new technologies. Following the revolution of medical imaging by CT, MRI and PET, new ways are being developed now to image tissues and organs on the molecular level. Notable among the Molecular Imaging technologies is the innovative use of mass spectrometers directly onto tissue slices, instead of on chemically processed samples, thus preserving tissue integrity. Unprecedented spatiotemporal mapping of protein abundance is achieved on full organ scale and at $50\mu m$ resolution [29]. This method, which is still in its infancy, enables detailed study of the dynamics of proteins in tissues and organs.

6. CLOSING REMARK

A key lesson from the application of signal processing to molecular biology has been that progress is achieved by astute adaptation of established techniques to the specific characteristics and pecularities of biological phenomena. The field being young, however, there is ample room for developing novel signal processing methodologies and, in return, having signal processing practice in other engineering domains be inspired by that which is learned from the natural world.

7. ACKNOWLEDGEMENT

I thank Noam Shoresh for fruitful discussions on the subject.

8. REFERENCES

[1] Eduardo D. Sontag, "Some new directions in control theory inspired by systems biology," *Systems Biology*, vol. 1, no. 1, pp. 9–18, 2004.

[2] Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.

[3] David Sussillo, Anshul Kundaje, and Dimitris Anastassiou, "Spectrogram analysis of genomes," *EURASIP J. on Applied Signal Processing*, vol. 2004, no. 1, pp. 29–42, 2004.

[4] P. P. Vaidyanathan and Byung-Jun Yoon, "The role of signal-processing concepts in genomics and proteomics," *J. of the Franklin Inst. Eng. Appl. Math.*, vol. 341, no. 1-2 (Special Issue on Genomics), pp. 111–135, 2004.

[5] Daniel J. Gaffney and Peter D. Keightley, "Unexpected conserved non-coding DNA blocks in mammals," *Trends Genet.*, vol. 20, no. 8, pp. 332–7, Aug. 2004.

[6] John S. Mattick, "Non-coding RNAs: the architects of eukaryotic complexity," *EMBO Rep.*, vol. 2, no. 11, pp. 986–991, 2001.

[7] Sean R. Eddy, "Computational genomics of noncoding RNA genes," *Cell*, vol. 109, no. 2, pp. 137–140, Apr. 2002.

[8] Robin D. Dowell and Sean R. Eddy, "Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction," *BMC Bioinformatics*, vol. 5, no. 1, 2004.

[9] Kevin B. Murray, Denise Gorse, and Janet M. Thornton, "Wavelet transforms for the characterization and detection of repeating motifs," *J. Mol. Biol.*, vol. 316, no. 2, pp. 314–363, 2002.

[10] Galit Lahav, Nitzan Rosenfeld, Alex Sigal, Naama Geva-Zatorsky, Arnold J. Levine, Michael B. Elowitz, and Uri Alon, "Dynamics of the p53-mdm2 feedback loop in individual cells," *Nat. Gen.*, vol. 36, no. 2, pp. 147–150, 2004.

[11] Subhayu Basu, Rishabh Mehreja, Stephan Thiberge, Ming-Tang Chen, and Ron Weiss, "Spatiotemporal control of gene expression with pulse-generating networks," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 17, pp. 6255–6360, Dec. 2004.

[12] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia, and Peter S. Swain, "Stochastic gene expression in a single cell," *Science*, vol. 297, pp. 1183–6, 16 August 2002.

[13] Mukund Thattai and Alexander van Oudenaarden, "Attenuation of noise in ultrasensitive signaling cascades," *Biophys. J.*, vol. 82, no. 6, pp. 2943–2950, June 2002.

[14] José M. G. Vilar, Hao Yuan Kueh, Naama Barkai, and Stanislas Leibler, "Mechanisms of noise-resistance in genetic oscillators," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 9, pp. 5988–5992, Apr. 2002.

[15] Hunter B. Fraser, Aaron E. Hirsh, Guri Giaever, Jochen Kumm, and Michael B. Eisen, "Noise minimization in eukaryotic gene expression," *PLoS Biol.*, vol. 2, no. 6, pp. 834–8, June 2004.

[16] Adam Arkin, John Ross, and Harley H. McAdams, "Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells," *Genetics*, vol. 149, pp. 1633–1648, 1998.

[17] Michael L. Simpson, Chris D. Cox, and Gary S. Sayler, "Frequency domain analysis of noise in autoregulated gene circuits," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 8, pp. 4551–6, 2003.

[18] Sara C. Madeira and Arlindo L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE Trans. Comp. Biol. Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[19] Amos Tanay, Roded Sharan, and Ron Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics*, vol. 18, no. Suppl. 1, pp. S136–S144, 2002.

[20] James C. Liao, Riccardo Boscolo, Young-Lyeol Yang, Linh My Tran, Chiara Sabatti, and Vwani P. Roychowdhury, "Network component analysis: Reconstruction of regulatory signals in biological systems," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 26, pp. 15522–7, Dec. 2003.

[21] Atul J. Butte, Pablo Tamayo, Donna Slonim, Todd R. Golub, and Isaac S. Kohane, "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks," vol. 97, no. 22, pp. 12182–6.

[22] Clary Clish et al., "Integrative biological analysis of the APOE*3-leiden transgenic mouse," *OMICS*, vol. 8, no. 1, pp. 3–13, 2004.

[23] Edward R. Dougherty, "Small sample issues for microarray-based classification," *Comparative and Functional Genomics*, vol. 2, pp. 28–34, 2001.

[24] Amir A. Handzel, "Avoiding erroneous prediction with cross validation in transcriptomics, proteomics and metabolomics," 2004, Genomic Signal Processing and Statistics (GENSIPS) 2004.

[25] Jianping Hua, Zixiang Xiong, James Lowey, Edward Suh, and Edward R. Dougherty, "Optimal number of features as a function of sample size for various classification rules," *Bioinformatics*, 2004, to appear.

[26] Homin K. Lee, Amy K. Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis, "Coexpression analysis of human genes across many microarray data sets," *Genome Res.*, vol. 14, pp. 1085–1094, 2004.

[27] Alon Zaslaver, Avi E. Mayo, Revital Rosenberg, Pnina Bashkin, Hila Sberro, Miri Tsalyuk, Michael G. Surette, and Uri Alon, "Just-in-time transcription program in metabolic pathways," *Nat. Gen.*, vol. 36, no. 5, pp. 486–491, 2004.

[28] Xin Lu, Wen Zhang, Zhaohiu S. Qin, Kurt E. Kwast, and Jnu S. Liu, "Statistical resynchronization and bayesian detection of periodically expressed genes," *Nucleic Acids Res.*, vol. 32, no. 2, pp. 447–455, 2004.

[29] Pierre Chaurand, Sarah A. Schwartz, and Richard M. Caprioli, "Imaging mass spectrometry: a new tool to investigate the spatial organization of peptides and proteins in mammalian tissue sections," *Cur. Opinion Chem. Biol.*, vol. 6, no. 5, pp. 676–681, Oct. 2002.