Editorial

Cancer systems biology: signal processing for cancer research

Olli Yli-Harja[1], Antti Ylipää[1], Matti Nykter[1], and Wei Zhang[2]

Abstract

In this editorial we introduce the research paradigms of signal processing in the era of systems biology. Signal processing is a field of science traditionally focused on modeling electronic and communications systems, but recently it has turned to biological applications with astounding results. The essence of signal processing is to describe the natural world by mathematical models and then, based on these models, develop efficient computational tools for solving engineering problems. Here, we underline, with examples, the endless possibilities which arise when the battle-hardened tools of engineering are applied to solve the problems that have tormented cancer researchers. Based on this approach, a new field has emerged, called cancer systems biology. Despite its short history, cancer systems biology has already produced several success stories tackling previously impracticable problems. Perhaps most importantly, it has been accepted as an integral part of the major endeavors of cancer research, such as analyzing the genomic and epigenomic data produced by The Cancer Genome Atlas (TCGA) project. Finally, we show that signal processing and cancer research, two fields that are seemingly distant from each other, have merged into a field that is indeed more than the sum of its parts.

Key words Systems biology, signal processing, gene regulation, methylation, glioblastoma

Cancer is recognized as a complex system with many genetic and molecular components that are tightly connected through mechanisms that cancer biologists are desperately trying to decipher in order to identify more effective approach to correct errors and cure the disease. However, this is a daunting task because the cancer genome can be altered in so many ways and the abnormalities exist at so many levels including genetic changes such as mutation, epigenetic changes such as DNA methylation and changes in microRNA expression. These changes are further connected through causal networks that are still mystery to cancer researchers for the most part. Research articles in these areas have been published in the early issues of the *Chinese Journal of Cancer*. Papers in this area will continue to be published in the current issue and future issues. This editorial does not intend to summarize these studies, rather it will provide a perspective of how the complex genomic and epigenomic data need to be viewed and processed to obtain insights into cancer biology.

Generic Signal Processing Methods Are Applicable in Diverse Fields of Biosciences

The two research paradigms in signal processing are formulating mathematical models of the natural world and developing algorithms to analyze it according to the models [1]. Traditionally, signal processing has enabled technologies with a wide scope of scientific and technical applications ranging from computer science and telecommunications to factory automation and robotics. Signal processing has played a crucial role in the development of such everyday technologies as television, radio, and personal portable communication devices. Only recently generic signal processing methods have also become an important catalyst in the future development of biology, paving the way to

Authors' Affiliations: [1]Department of Signal Processing, Tampere University of Technology, Tampere, FI-33101, Finland; [2]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.

Corresponding Author: Olli Yli-Harja, Department of Signal Processing, Tampere University of Technology, P.O. Box 553, Tampere, FI-33101, Finland. Email: olli.yli-harja@tut.fi.

important applications such as rational drug discovery and development, personalized medicine, and cancer care.

In unison with the report on New Biology by The National Academies, we argue that applying general data analysis methodology in biology has several benefits. Signal processing has been found to provide crucial links between theoretical and applied research, as well as between different disciplines in biosciences. One such clear benefit is that mathematical models of biological systems provide a common ground that facilitates more efficient communication between bioscientists. The sharing of similar computational tools paves way to sharing information and a new, more unified research paradigm, thus melting the barriers between disciplines. We can provide the biological research community in general with theoretical insight based on computational predictions, efficient ways to analyze data, and an effective means of integrating experimental results in a meaningful manner[2].

The Evolution of Mathematical Modeling: From Geometry to Deciphering Cancer Genomics

In the core of signal processing lie mathematical models. They are defined as being compact descriptions of natural systems using a mathematical language, but often have subtle differences in meaning depending on the background of the modeler. Recently, we have observed an interesting evolution in the role of mathematical modeling. The four traditional views, Pythagorean, Newtonian, Turing's, and Leonardo's views that have dominated the field, now give way to a fifth view: the biologists' view. The eldest of the four, Pythagorean view emphasizes the simplicity and the beauty of the models. This obsession stems from the need of finding closed form solutions to problems without computers. Secondly, Newtonian view emphasizes accurate prediction of the nature. With enough evidence gathered to support a model, it may ultimately be promoted to a generally accepted law of nature, such as Newton's law of universal gravitation. The dawn of computers paved the way for the third view, Alan Turing's view, dubbed by the famous British computer scientist. According to this view, it is required that the model can be described as a computer program, and that the program can be run in a feasible time. Fourth view, represented by the Italian artist and engineer Leonardo da Vinci, has always emphasized the applicability and usefulness of the models, thus combining aspects from the mathematicians', the physicists', and more recently,

the computer scientists' views.

Given the above, what is the view of a biologist to mathematical models? In biology, the word model itself is reserved for another purpose. A mouse model, for example, is by no means a simplified description of a human disease as could be misunderstood by a mathematician. Instead, it serves the purpose of creating a feasible experimental setup. A complete rethinking is required when we propose mathematical models to be used in the overwhelming complexity of biology. We propose that a useful biologists' view is to emphasize the use of the models in improving communication. A compact mathematical or graphical description of a biological system serves as a common ground for communication and provides a language that participants of a multidisciplinary research effort can depend on. Here simplicity is a virtue, just like in mathematics, because it increases the popularity of the model. This, in turn, increases the value of the model in communication. The model should also provide accurate prediction, but the high connectivity of biological systems to their environment poses serious problems in applying the physicist's view properly. It is conceivable, that in the future some models may achieve the status of a natural law in biology, for example, the long-standing hypothesis that life exists at the edge of chaos by Kauffman[3]. Finally, the utilitarian views of the computer scientist and the engineer are obviously useful in biology—a mathematical description of a biological system allows the development of efficient computational tools for e.g. prediction, feature extraction, filtering, classification, and system identification.

Systems Biology for Cancer Research Emerges from Efficient Use of Data

Using mathematical models to extract knowledge out of massive amounts of biological data is the essence of a field of science called systems biology. In cancer research, modern measurement data can provide information on individual genes and on the states cellular systems can adopt. On molecular level these states have characteristic patterns of gene expression due to genetic and epigenetic features, and on clinical level they can be seen as more aggressive or drug-resistant disease. The promise that systems biology holds for cancer research is to draw either weak associations or strong causal relationships from gene level to pathway level to phenotype. Given the abundance of mathematical modeling tools that were once used for modeling man-made electrical systems, and databases filled with high-throughput biological "omics" data, the

premises for a systems theoretical approach are thereby available.

Typical results of interdisciplinary efforts in cancer systems biology are models with various granularities describing the studied biological systems, usually tumor cells, cellular processes, or signaling pathways. The models help to generate new hypotheses to be tested through following experimental research, thus generating a research cycle consisting of both computational and experimental work. A close interplay between the experimental results, the predictions based on computational models, and the design of new experiments based on the model predictions is an essential part of the iterative systems biology research approach depicted in Figure 1. Application of signal processing therefore supports better targeting of research resources and a much more comprehensive knowledge building process in all life sciences, not just cancer research.

Glioblastoma Multiforme is The Proving Ground for Cancer Systems Biology

Diffuse gliomas are the most common type of primary brain tumors in adults and incidentally, also among the most extensively studied forms of cancer. A critical mass of research has now led to the application of cancer systems biology by pioneering glioma researchers, and with a great success. A shift in research focus for glioma is also justified with dire statistics: glioblastoma multiforme (GBM) is the most common, and unfortunately, also the most highly malignant glioma. GBM comprises 50% to 60% of all gliomas. The median overall survival for patients with GBM is less than one year, and the dismal prognosis has not significantly improved over the last five decades of modern cancer care [4-6]. In recent years, numerous chemotherapeutic regimens have been evaluated without significant improvement in patient survival. This disappointing failure underlines the urgent need to change the strategy for identifying novel molecular targets and the most appropriate chemical agents for intervention.

Identification of the driving molecular events, such as mutations in DNA or altered regulatory circuits, and understanding their impact in signaling pathways and biological processes are both highly critical areas of investigation in modern glioma research. During the last ten years, many studies have aimed at profiling and understanding the genomics and proteomics of glioma
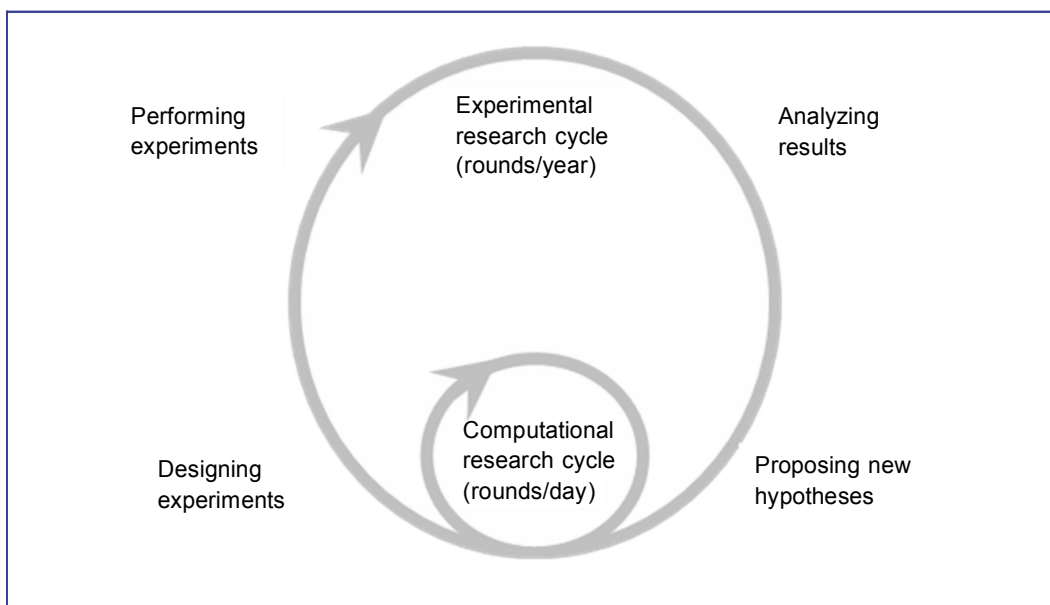

Figure 1. Research cycles of systems biology. The slowly rotating experimental research cycle consists of designing and performing experiments, analyzing their results, and finally proposing hypotheses based on the conclusions. Computational cycle is made up of the same constituents, but it uses mathematical models instead of, for example, mouse models. Simulating biological experiments on mathematical systems, and automatically analyzing the results, makes it feasible to propose new hypotheses in a fraction of the time it takes to complete an experimental cycle. Thus, spinning a rapid computational research cycle within an experimental cycle can accelerate research substantially.

using computational tools. The most recent effort was embodied by the Cancer Genome Atlas (TCGA) project[7] as well as the project at National Cancer Institute which formed the database Rembrandt [8]. Many exciting discoveries in glioma genomics have already been made, although a major challenge is the multiplicity of pathways activated in cancer and the difficulty of identifying the key targets in them. In recent years, numerous pathway approaches have been proposed to deal with such complexities. However, the sheer number of altered genes is still posing a serious problem for identifying proper markers for clinical translation, such as identifying potential genes for targeted therapy.

Not only true in glioma research, but even more generally, the emergence of vast amounts of biological high throughput data warrants computational tools that keep up with this positive, albeit rather problematic, development. We think that the tools should be developed based on the principles of data-driven signal processing and systems biology which is based on interpreting large amounts of genome-wide measurement data. Efficient use of these tools on multi-level biological data have resulted in numerous successes, including the findings of master regulators behind mesenchymal transformation of GBM cells[9], identification of four GBM subtypes [10], and a link between MGMT promoter methylation and a hypermutator phenotype[7]. Encouraged by these success stories, a consensus has emerged that we must move into understanding the genetics of the disease by integrated systems biological analyses. Signal processing methodologies can work as enabling technologies for this movement. To mention only a few examples, application of signal processing has led to development of more accurate tools for prediction of transcription factor binding to gene promoters, which is a necessity for identification of master regulators[11]. It has also contributed towards improved clustering and feature

selection methodologies that allow robust identification of cancer subtypes[12]. Furthermore, various machine learning and classification algorithms allow efficient reverse engineering of gene regulatory mechanisms[13]. We hope that integrating the data with signal processing and systems biology methods will help us build the genetic groundwork for gliomas and other malignancies alike[14].

The Signal Processing View on the Future of Cancer Research

New measurement platforms, such as microarrays and sequencing technologies, produce a huge mass of multi-modal and heterogeneous data for cancer researchers. High-throughput measurements are no doubt a necessity and have accordingly become commonplace in cancer research laboratories all over the globe. The volume of this data is increasing at a speed that effortlessly surpasses the rate of increase in computer efficiency known as Moore's law. Thus we are already losing the ability to cope with the incoming data. Development of new signal processing and systems biological methods is the answer to the analysis of all this data; it is also the glue that binds together biological experiments and mathematical models, as well as collaborative efforts between scientists. Among other leading journals, *Chinese Journal of Cancer* has also joined the frontlines of new systems biological cancer research effort. Hopefully such novel modeling approaches [15] will eventually result in new therapies to help patients who see little hope in the currently available treatment options. This is the challenge, opportunity, and the future of cancer research.

References

[1]   Moura JMF. What is signal processing? [J]. IEEE Signal Processing Mag, 2009,26(6):6.

[2]   The National Academies. A new biology for the 21st century [M]. Washington: The National Academies Press, 2009.

[3]   Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets [J]. J Theor Biol, 1969,22(3):437–467.

[4]   Behin A, Hoang-Xuan K, Carpentier AF, et al. Primary brain tumours in adults [J]. Lancet, 2003,361(9354):323–331.

[5]   Ohgaki H, Kleihues P. Genetic pathways to primary and secondary glioblastoma [J]. Am J Pathol, 2007,170(5):1445–1453.

[6]   Wykosky J, Fenton T, Fumari F, et al. Therapeutic targeting of epidermal growth factor receptor in human cancer: successes

and limitations [J]. Chin J Cancer, 2011,30(1):5–12.

[7]   Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways [J]. Nature, 2008,455(7216):1061–1068.

[8]   Madhavan S, Zenklusen JC, Kotliarov Y, et al. Rembrandt: helping personalized medicine become a reality through integrative translational research [J]. Mol Cancer Res, 2009,7(2):157–167.

[9]   Carro MS, Lim WK, Alvarez MJ, et al. The transcriptional network for mesenchymal transformation of brain tumours [J]. Nature, 2010,463(7279):318–325.

[10] Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and

NF1 [J]. Cancer Cell, 2010,17(1):98−110.

[11] Lähdesmäki H, Rust AG, Shmulevich I. Probabilistic inference of transcription factor binding from multiple data sources [J]. PLoS One, 2008,3(3):e1820.

[12] Monti S, Tamayo P, Mesirov J, et al. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data [J]. Mach Learn, 2003,52: 1−2;91−118.

[13] Shmulevich I, Dougherty ER, Kim S, et al. Probabilistic Boolean

Networks: a rule-based uncertainty model for gene regulatory networks [J]. Bioinformatics, 2002,18(2):261−274.

[14] Nykter M, Lähdesmäki H, Rust A, et al. A data integration framework for prediction of transcription factor targets: a BCL6 case study [J]. Ann N Y Acad Sci, 2009,1158:205−214.

[15] Ylipää A, Yli-Harja O, Zhang W, et al. A systems biological approach to identify key transcription factors and their genomic neighborhoods in human sarcomas [J]. Chin J Cancer, 2011,30(1):27−38.