

# Bioinformatics and the Systems Biology Hierarchy: *An Overview*

Dr. Eric Werner  
 President, Cellnomica, Inc.  
 Fort Myers, Florida and  
 Munich, Germany  
 eric.werner@cellnomica.com  
[www.cellnomica.com](http://www.cellnomica.com)

**Abstract**  
 A short overview and perspective on present and future trends in bioinformatics and systems biology aimed at CEO's and decision makers involved in strategic development of biotechnology and pharmaceutical companies is presented. The emphasis is not on details; some reviews are cited for that, but rather on the vision of where we are heading and to assist the decision maker in strategic planning of research and development. Research budgets can be in the hundreds of millions and billions of dollars. No CEO or decision maker

can afford to be ignorant of the broad directions of the software tools that may help the company to position itself in rapidly evolving fields such as biotechnology and pharmacogenomics. Indeed, brand new markets are opening up that offer unprecedented opportunities as well as giant challenges to traditional pharmaceutical and biotech companies. Much like the beginnings of the PC market, large corporations will lose market share and even be replaced entirely if they do not face up and adapt to these challenges.

## Bioinformatics and Systems Biology Hierarchy



**Figure 1** The bioinformatics and systems biology organizational hierarchy contains six distinct levels: **Level 1.** The genomic and proteomic databases, **Level 2.** Bioinformatic analysis tools, **Level 3.** Mathematics and formal theories [9], **Level 4.** Regulatory networks and pathways [8], **Level 5.** Single cell modeling [5, 6, 10, 7, 10], **Level 6.** Multicellular systems modeling.

## Introduction

The aim of this essay is to give CEO's and other decision makers in charge of reviewing and investing in new technology a brief overview of the important future trends in the area of bioinformatics and systems biology. We start with an overview of the phenomena of data drowning and the software challenges that accompany it. Traditional approaches to research in the pharmaceutical industry have followed a bottom-up strategy. Under it, a systematic investigation of all possible combinations will eventually lead to a solution to the problem in question. In the genomic and proteomic field this is a recipe for failure. The combinatorics are so large that no systematic bottom-up strategy alone will suffice to understand genomic and proteomic data. Instead a top-down approach needs to complement the bottom-up strategy. On the top-down approach, system level models help the researcher understand the data. Much like a manager or engineer uses system level analysis, the researcher is helped to see the forest from the trees and begin to develop an understanding or semantics of genomic and proteomic data. The system level approach means we must model not just the function of the genome within the cell but within a multicellular dynamic context, which makes up tissue, organs and organisms.

## Data Drowning and Missing Meaning

Pharmaceutical companies are facing a serious research problem. The problem is not one of scarcity but one of plenitude. We have a surfeit of information from genome and proteome projects. Genome projects are generating several types of data. First, we have the basic DNA sequences that make up the genome. Some of the subsequences represent genes. We cannot ignore the regions that do not code for genes, the so called noncoding regions, because they may contain control elements such as promoters that tell when genes are switched on and off [1]. Genes can be in active or inactive states. Genes are

transcribed into RNA and RNA is translated into protein. When a gene is active it results in the production of RNA and then protein. The gene is expressed either as RNA or as protein. This generates the next type of data, namely, RNA expression profiles and protein expression profiles. Thousands of genes may be expressed at any given instant. Hence, each expression profile involves a mass of data that indicates which measured genes are expressed and to what degree. Furthermore, unlike the cell's DNA, the expression profiles are not static but time dependent. So as the cell develops its RNA and protein expression profiles will change. Hence, we get a massive amount of data at each time step we choose to measure. Such data is needed to understand the networks of interactions between genes and between cell signaling pathways that relate the genome to the outside world. Beyond the crisis in constructing coherent and useful databases to handle the sheer volume of data [2], there is a more fundamental and more important problem: The central problem is not the amount of data, it is rather that the meaning of the data is missing.

## Software and the Search for Meaning

Anyone who has ever looked at DNA sequences, or RNA or protein expression data, knows it is hopeless without the aid of software to analyze or in any way make sense of the data. However, there lies the difficulty. Any software system is ultimately based on hypotheses, theories and models of what things are and how things are supposed to work. At present we have no complete theory or model as to what happens in the cell or what happens in multicellular systems. Instead we have an eclectic assortment of different software packages each with its strengths and weaknesses. Each has its own view of the biological world. It is a situation much like the blind men trying to understand the elephant. Whenever someone has a new approach a company springs up that attempts to market that view. The

buyer of technology is faced with the problem that he probably needs many of the new approaches, but once he has them integrating the different packages can be difficult if not impossible. We need an integrated system level view of what are highly dynamic and complex living organisms.

Note, sequence analysis and comparison software is the dominant method of assigning meaning to DNA and protein sequences. One looks for similar, homologous sequences and simply carries over the meaning that has been worked out for the old sequence to the new sequence. This process however can only be as good as the quality of the analysis of meaning of the old sequence. Even if the analysis of the original sequence is good, small differences in sequence can result in radically different functions of the DNA or protein sequence. Even more devastating is the problem that often the annotation of original, reference sequences is a rather informal process subject to the observational acumen and caprice of individual researchers. Finally, there is the problem of combinatorial complexity of a bottom up approach to understanding the genome. The idea of following the research methodology inspired by the Central Dogma [3], [4] will not give an adequate understanding of genes and their function. In recent years and even more so now it has become evident that a system level understanding of the cell and multicellular phenomena is indispensable.

## Systems Biology and the Semantics of Genomes

In view of the massive and for all practical purposes, incoherent data stream coming out of commercial and university laboratories, a recent trend is to use a system level, in silico approach to analyze and model the phenomena observed in living systems, in vivo and in vitro. Under this approach, understanding the meaning and function of states and processes in a living system is the key

to understanding the experimental data that is measured. For example, to make sense of gene expression data, the data is analyzed in terms of our understanding as realized in models of the dynamics of the cell. This process of establishing a correlation between meaning or function and data is part of the methodology that we refer to as giving a semantics to the genome. The end result of this is a formal working model of how systems of cells function dynamically in the growth, development and maintenance of the organism in the context of its environment. Ideally, this model can then be used to predict and explain the data that is observed. Practically, the process of building such a model involves a feedback loop, where predictions made by the model are compared with experimental data. If the prediction is correct, we have a confirmation of the model. Otherwise, the model is modified and a new round of predictions is made. This continues until we have a model that is consistent in making verifiable and veridical predictions. In the process of building and adapting the model, our understanding of the semantics of genomes and the cell increases. It points the way to new experiments and new possibilities in the treatment of diseases.

### **Advantages of Systems Biology Approach**

One of the key advantages of system level models and theories is that they are able to integrate data from system level experiments in developmental biology, tissue regeneration, wound healing, cancer studies and organ generation with lower level data such as DNA sequencing data, proteomic data, and gene expression data. This is in sharp contrast to bioinformatic databases that only associate low level data loosely with imprecise descriptions of higher level systemic properties. For example, gene mutations are annotated with information about the effect of the mutation on the development of the organism. But, this is left at the level

of description and does not explain or integrate the phenomena in question.

Thus, databases are not enough. Systemic models lead to a deeper understanding and semantics of genomes. For drug development and genomic engineering we need to be able to understand the precise interrelationships and functional organization of cells and genomes. A database does not do that.

Why is this important? For example in wound healing, tissue regeneration, cancer, differentiation, bilateral symmetry, cell signaling, tissue engineering, pattern formation, and homeotic mutations, a database will only show associations of what already has been discovered. In each case the database may associate gene expression profiles with the phenomena in question, but it does not give us control of the phenomena. A good model can do that. The model determines the semantics of the genome and the cell state. It can be used to make predictions and influence decisions in a much more direct way than a database can.

Another advantage of system level models is made clear by a simple example. Consider a gear in a mechanical clock. We may know it must be related to another gear. But we will not fully understand the function of the gear in the clock unless we understand its function in the entire system of interacting parts. In one clock the gear may function to move the second hand, in another clock the very same gear may move the hour hand. The point is that there is no inherent property of the gear that makes it an hour hand as opposed to a second or minute hand gear or whatever. Only in the context of the entire system can we give meaning to the gear. Similarly, genes are parts that generate other parts of a complex system. To understand the meaning or semantics of genes we need to have a system level model of the genome, the cell and the multicellular context.

### **Single Cell Modeling**

There are several important attempts at single cell simulations (\*see Chart). One is an initiative in Japan with the goal to simulate every aspect of cell functioning of a minimal cell. A minimal cell is a cell that is able to take in nutrition and do the minimal processing to survive. The system is called e-cell [5]. It uses differential equations to model the chemical pathways in the cell. Another effort is being made in America with a system called vcell [6]. It is a similar approach to the Japanese using differential equations to model the basic chemical pathways in the cell. The idea is to come up with predictive models that can be manipulated to study what would happen if a pathway were to be interfered with or changed in some way [7][8]. The models predictions are compared with the gene expression levels in real cells. Through these methods important insights are gained into the workings of the complex cellular system. The work has great potential applicability to finding drug targets more quickly and thereby reducing the cost of drug development enormously. Cell models can also assist the researcher in seeing potential problems such as adverse drug interactions. That can be invaluable as a proactive and preventative measure.

### **In Silico Multicellular Systems**

Because modeling a single cell gives little insight into multicellular processes and in particular multicellular diseases, it is inevitable that single cell simulations must be complemented by in silico multicellular systems. (\*see Fig. 1) Cellnomica, Inc. is a company that has software that models many interesting and useful multicellular processes including cell division, cell differentiation, cell signaling, chemical gradients, bilateral symmetry, tissue development, tissue interactions, mutations including homeotic mutations, and multicellular diseases such as some types of cancer. At the level of the cell the software models the cell at minimal but sufficient level of detail to include, for example, cell

signaling networks. At the level of the genome, genome networks are not only modeled but also integrated with the rest of the cellular and multicellular model so that changes in the network can be observed in the developing, dynamic 4-dimensional structure. What is significant is that a balance is reached between sufficient detail to model such phenomena, but not so much detail as to make the system incomprehensible. The software is extensible and has the potential to be integrated with software packages that have more detailed models of single cells.

Perhaps the greatest strength of such a system level view of the function of the genome in the context of the cell and its multicellular environment is the insight and global understanding it gives the scientist attempting to unlock the meaning of the genome and the proteome. The ability to mutate the genome and see the effects immediately in the developing in silico multicellular system gives the researcher an unprecedented tool to test his hypotheses and models. Given that a normal in vivo or in vitro mutation can take months to perform in the lab, reducing the time to two or three seconds can be a significant benefit to the researcher attempting to understand and design his experiment.

### **Reverse and Forward Engineering Living Systems**

Another important aspect of the systems biology approach is that in silico models of cells and pathways can be used not just to reverse engineer the workings of cellular systems but also to forward engineer or design biological systems with desired properties. It opens up a new world of possibilities that go beyond the traditional reactive approaches to living systems. Specifically, because we now have the ability to manipulate genomes to express genes at will, genomes are becoming the objects of new designs for new functions. The manipulation of the genomes of model organisms such as pigs in order enable the production of organs that are not

rejected by humans, is a purely biological example. It is becoming commonplace to manipulate regulatory networks in genomes [7]. The danger is that without a system wide understanding of the genome and the effects of transforming that genome, we can easily have unwanted and costly side effects.

If one combines this with the possibility of designing and predicting in silico the effects of much more complex genome networks, then brand new markets open up that traditional pharmaceutical and biotech companies should not ignore. Since there are only a limited number of drug targets, the life span of a traditional pharmaceutical company is limited. Therefore, pharmaceutical companies must open themselves to new markets if they are to survive. The traditional division between living, in vivo or in vitro, and nonliving, in silico systems is disappearing.

Nanotechnology the science of engineering at the molecular level is a field where designed systems are often inspired by living systems. And, nanomachines may transform living systems. We expect systems biology and nanotechnology to interrelate more and more as our understanding of both these fields increases.

### **Drug Delivery and System Modeling**

Clearly the problem of drug delivery can benefit from software systems that model not just a single cell, but also a system of cells that interact with the drug and with each other. Since drugs can have different effects at different times, it is important to model not just a static group of cell-like objects or some material that has tissue like properties. It is far better to model a dynamic multicellular system where the cells, that make it up, change in time in response to extra cellular, intracellular, and genomic signals, as well as, externally induced chemical signals. The advantages of an in silico model of such complex dynamic processes is that we can view system in action in a way that is impossible in

the lab. For example, we can cut the organism in half and look at the interactions between the cells without disturbing the organism. Furthermore, because the processes are so complex, the aid of a simulation that can be stopped at any time and nondestructively dissected in time and space provides insights that cannot be predicted by reasoning alone. While we are not there yet, software systems like Cellnomica's together with e-cell or vcell show that we may be closer than many would have expected even just a year ago.

### **Perspectives**

We have seen that data is growing exponentially but our understanding of the data is growing at a snail's pace by comparison. We have seen there are great challenges for both the biotech and pharmaceutical industry. We can no longer do business as usual. The data presents a great opportunity to find new solutions to the problem of disease. It also offers opportunities that previously did not even exist. These include new approaches to tissue engineering and design, in silico drug delivery testing, and reduction of the need for animal testing. It may give us new insight to the problems of multicellular diseases such as cancer. While there are many individual software packages, it is essential to have software that allows the researchers to integrate the massive data confronting them. Systems biology is now becoming a strong new trend that cannot be ignored by more traditional drug companies. To adequately understand the data one must include a system level approach that complements the traditional bottom up chemical analysis. The system level, top-down approach can save a team years of research time, by eliminating unpromising experimental scenarios and pointing out promising avenues of research and development.

Managers are familiar with system level approaches. They use them every day in their strategic planning. One must have an overview of the components and what they are doing

to be able to plan and take proactive and effective action. Research must, of course, also be managed to be effective. Today, the utilization of genomic, proteomic and other biological data requires an integrated system level understanding of how things work in an organism. This in turn helps to determine what data is important. And this helps in the setting of research and development goals as well as helping to formulating the overall research agenda of a company.

[1] T. Werner, Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data, *Pharmacogenomics* 2001 Feb;2(1):25-36 (review of promoters)

[2] D. Frishman, K. Heumann, A. Lesk, HW. Mewes, *Comprehensive, comprehensible, distributed and intelligent databases: current status*, *Bioinformatics* 1998;14(7):551-61. (review of bioinformatic databases)

[3] F. Crick, Central Dogma of Molecular Biology, *Nature*, 227, 561-563 (1970)

[4] E. Werner, *Genome Semantics, Systems Biology and the Central Dogma*, forthcoming, 2002.

[5] M. Tomita, Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol.* 2001 Jun;19(6):205-10. (e-cell review)

[6] J. Schaff, L.M. Loew, (1999) "The Virtual Cell", *Pacific Symposium on Biocomputing*, 4:228-239 (overview of vcell)

[7] T. Ideker, et. al., A New Approach to Decoding Life: Systems Biology, *Annu. Rev. Genomics Hum. Genet.*, 2:343-72, 2001.(review)

[8] E. Davidson, *Genomic Regulatory Systems: Development and Evolution*, San Diego, CA: Academic Press, 2001. (Regulatory networks)

[9] E. Werner, "Logical Foundations of Distributed Artificial Intelligence", *Foundations of Distributed AI*, G. O'Hare and N. Jennings (eds.), Wiley Publishers, 1996. (Introduction of multiagent systems)

[10] D. Endy, L. You, J. Yin, IJ Molineux, Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. *Proc Natl Acad Sci U S A* 2000 May 9;97(10):5375-80 (T7 simulation)

[11] [www.cellnomica.com](http://www.cellnomica.com) (multi-cellular systems biology software)