Minireview

Genome semantics, in silico multicellular systems and the Central Dogma

Eric Werner

Cellnomica, Inc., P.O. Box 1422, Fort Myers, FL 33928-1422, USA

Accepted 1 February 2005

Available online 14 February 2005

Edited by Robert Russell and Giulio Superti-Furga

Abstract **Genomes with their complexity and size present what appears to be an impossible challenge. Scientists speak in terms of decades or even centuries before we will understand how genomes and their hosts the cell and the city of cells that make up the multicellular context function. We believe that there will be surprisingly quick progress made in our understanding of genomes. The key is to stop taking the Central Dogma as the only direction in which genome research can scale the semantics of genomes. Instead a top-down approach coupled with a bottom-up approach may snare the unwieldy beast and make sense of genomes. The method we propose is to take in silico biology seriously. By developing in silico models of genomes cells and multicellular systems, we position ourselves to develop a theory of meaning for artificial genomes. Then using that develop a natural semantics of genomes.**

Keywords: Genome semantics; Minimal multicellular system; Central Dogma; Systems biology; Cancer

1. Introduction

This article introduces the new concept of genome semantics to gain an understanding of meaning genomes in the context of multicellular processes including multicellular development. The approach is to use in silico modeling to come up with universal properties of genomes. The artificial genomes have a definite semantics and serve as a basis for understanding their more complex counterparts, the in vivo, natural genomes of real multicellular organisms. The idea is to reverse engineer natural genomes, through the understanding we gain from in silico genomes. Ultimately, in silico artificial genomes and in vivo natural genomes will translate into each other, providing both the possibility of forward and reverse engineering natural genomes.

At present groups are modeling and constructing logic circuits as gene networks. Other groups are using in silico methods to design and reverse engineer single cells. I am involved with research to model and reverse engineer multicellular processes such as cell signaling, chemical gradients, cell division, and the dynamic development of multicellular structure. The modeling of such processes gives fundamental insights into the overall architecture of genomes. Once the minimal cell has been created, we can use this work as a basis for the design of minimal multicellular organisms. These are pre-organisms that mimic the development and behavior of subunits of more complex natural multicellular organs or organisms.

2. The search for meaning

The central problem of genome research is to understand the meaning of genomic regulatory networks that underlie the development and functioning of living systems. A *genome semantics* is a theory of meaning of genomes. Part of that semantics is a *semantic code* that translates genomic sequences into their systemic meanings. Since biological systems are organized in different levels, a semantics of genomes may associate different meanings with a sequence depending on the level of ontology, function and organization. The meta-theory of how we arrive at the semantics of genomes is, explicitly or implicitly, a part of a genomic and proteomic research strategy.

The dominant research strategy for understanding genomes is a bottom-up strategy motivated by the Central Dogma (see below). We believe that the time has come to reconsider the dominant position of this research strategy. Instead of a strictly bottom-up strategy, we urge the consideration of a complementary top-down strategy. We believe a research strategy that integrates higher levels of system information with low-level genomic and proteomic information is necessary in order to decipher the semantic code. We first look at some of the reasons, the Central Dogma is no longer a sufficient organizational paradigm for research on the semantics of genomes. We then look more closely at genome semantics and its relation to in silico multicellular systems biology.

3. Multicellular diseases

Many diseases are inherently multicellular in nature. For example, the etiology and development of cancer involves not just a single cell but also many cells that interact with their neighboring cells. In fact, many cancers are classified into stages by means of criteria such as location, cell differentiation and how they interact with other cells. A stage one cancer is usually very localized and has its own boundaries distinct from other tissues [1]. A stage four cancer is non-local and has invaded other cell tissues, causing secondary cancers (metastasis) in those tissues. To understand such cancers, we need to understand the role of the genome, the interaction of the genome

E-mail address: eric.werner@cellnomica.com (E. Werner).

0014-5793/$30.00 © 2005 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2005.02.011

with the host cell and the interaction of the host cell with other cells.

As is well known by now, the complexity of the regulatory pathways that control the interaction of a genome and its host cell is enormous. Small wonder that the consideration of yet another level of complexity, namely, multicellular interactions may strike the reader as not only premature but virtually impossible at present given that we don't fully understand how a genome functions in a single cell. Indeed, at present most in silico simulations of cells are provisional and are restricted to a single cell or its parts [2]. Yet, a full understanding of the etiology and functioning of cancer requires, we consider at least four levels: the genome, the cell, intercellular interactions and multicellular processes. Indeed, we may need to include organ and system level interactions as well.

In a system level interaction, a coherent system of cells such as a gland interacts with other cellular systems, such as the muscular system by way of cellular communication, using, for example, hormones. To fully understand a genome in the etiology of a disease that is multicellular, we need to understand not only the functioning of the genome within the context of the cell [3], but also how the cell with its genome interacts with other cells. In other words, we need to understand the meaning of the genome in a multicellular context including dynamic multicellular processes. How do we approach such a complex and daunting problem?

4. The Central Dogma and its limits

Many at present are proceeding bottom-up, following a research methodology inspired by the Central Dogma. The *Central Dogma* as originally formulated by Crick is a negative hypothesis that states that information cannot flow downwards from Protein to DNA. Its complement, the *Sequence Hypothesis* is often conflated with the Central Dogma [4,5]. Under it, DNA is transcribed to RNA, and RNA is translated into protein. More abstractly, information flows upward from DNA, to RNA, to proteins, and, by extension, to the cell, and, finally, to multicellular systems. In the ensuing years, many scientists have merged the two hypotheses and refer to them collectively as the Central Dogma. We will use the term in this latter collective, conjunctive sense.

The Central Dogma has been the motivation for a reductionist approach to genome research methodology even if the original authors may not have intended it to be used that way. This reductionist methodology presupposes we must have a theoretical and practical understanding of each lower, more fine-grained level of information and ontology, before we are allowed to proceed to understand the next level of information.

The greatest weakness with a research program that follows the Central Dogma is the staggering complexity. The problem is that the search space for finding a solution is too vast. In computer science problems are often represented in terms of the space of possible paths that may lead to a solution. Such a set of possible paths is called a search space. A solution is a path in such a search space that leads to a solution or goal. Some problems have such vast search spaces that they are practically impossible to solve. Computationally, these are known as NP-complete problems. They are so complex that they cripple our fastest computers. Yet the genomic and cellular networks involve hundreds of interacting parts and appear to involve NP-complete problems.

5. Reducing the semantic search space

In actual scientific practice, however, the researcher forms his research agenda based on higher-level knowledge about higher levels of biological information. They need to be able to see the forest from the trees. Even in the more mundane world of day to day experimental decision and design, the researcher acts in a context of high level, systemic knowledge of phenomena such as the functioning and dynamics of multicellular interactions, multicellular systems, organs, the etiology and dynamics of multicellular diseases, as well as, healthy developmental biological processes. The researcher presupposes this knowledge to give a broad direction to his research and his experimental design. More importantly, it gives the research its significance. This high level, systemic information gives the reason why the research and experiments should be done at all.

Such knowledge is known in the artificial intelligence community as *heuristic knowledge*. Heuristic knowledge is defined as information that reduces a search space. So, in these terms, the scientist uses heuristic, system level biological knowledge to reduce the informal, intuitive, a priori search space that defines his problem. In our case, heuristic information would be used to reduce the semantic search space, the space of possible interpretations of the genomic code.

Is there a way to utilize high-level system information to understand genomes in the context of cellular and multicellular processes? We believe there is. Instead of using the Central Dogma as a paradigm for the process and methodology of discovery, we go in the other direction. We proceed from the system level to the supporting foundational levels.

6. The flow of information

In terms of the flow of information, under the Central Dogma, information cannot flow downward from Protein to RNA to DNA [4]. However, at the system level, information does flow downwards from proteins to DNA. An example is cell signaling. There a series of protein–protein and protein–RNA interactions leads to the activation of DNA transcription. Thus, the Central Dogma describes only one of the informational directions and paths out of many possible informational paths in the cell as informational system. Indeed, there are intracellular informational routes as well as intercellular informational routes. These routes constitute informational routing networks within the cell and between cells. They mediate cellular and extracellular information with the cell's genomic information.

7. Overcoming the Central Dogma

The Central Dogma is not just a hypothesis about the flow of information; it has also been appropriated as a research program. To escape from the constraints of the Central Dogma, we must become conscious of the distinction between the Central Dogma as a scientific hypothesis and the Central Dogma as a research methodology. Otherwise, we may presuppose wrongly that since information does not flow downwards, we cannot move top-down from the higher levels of information to the lower levels. The duality of the flow of information in multicellular, as well as, in single cell systems points the way out of the box of the Central Dogma as research program. It gives us the freedom and provides the scientific legitimacy to take a systemic approach while being consistent with the Central Dogma as a scientific hypothesis in its original restricted form.

8. A systemic approach

In a top-down approach, we simulate multicellular processes at a level of abstraction that allows us to capture many of the system level phenomena that are known from research on the etiology and progression of disease, from research on tissue and limb regeneration, from stem cell research, from cloning experiments, from cell differentiation, from research in microbiology, and from over a century of research in developmental biology. We seek the minimal conditions a genome and its cellular context must satisfy in order to simulate natural multicellular phenomena.

Furthermore, there is nothing to prevent us form using a bottom-up strategy simultaneously. In artificial intelligence, one of the best search strategies is to combine a top-down approach with a bottom-up search, the two searches meeting in the middle to form a solution path.

Note, there is no inherent preferred status to knowledge of biological processes at a lower level of ontology (e.g., biochemistry) over and above other levels of information and ontology. Correct high level information about cellular phenomena (e.g., the orientation of cellular division) and multicellular processes (e.g., cell signaling protocols) will not necessarily be changed by a more detailed, lower ontological view. Often, it is the reverse; the higher-level system knowledge helps us to constrain the search space, and to advance our understanding of lower level processes. Thereby, our understanding of distinct levels of information about a system may change as we gain more knowledge of each.

9. Criteria for in silico systems

Imagine we have a software system that can design artificial genomes in silico and then use that genome to generate an artificial organism in silico. For example, see Fig. 1, where an in silico minimal multicellular organism is shown at a particular stage of dynamic development. How would we know if the in silico genome and organism expresses truths about natural genomes and organisms? Well, first we could see if the in silico system mimics some of the major systemic properties of natural genomes and organisms. For example, is the system able to simulate multicellular development, bilateral symmetry, cell signaling, genome networks, cancer, tissue generation, or cell differentiation? Can we perform mutations on the in silico genome and see effects analogous to what we see in nature, namely, abnormal development, premature death, cancer and homeotic mutations? With each affirmative answer to these questions, we have more confirmation that the in silico system reflects some fundamental properties of natural genomes and multicellular systems. However, we might want to have an even more precise correlation. We may want to translate one genome into the other and see the effects.

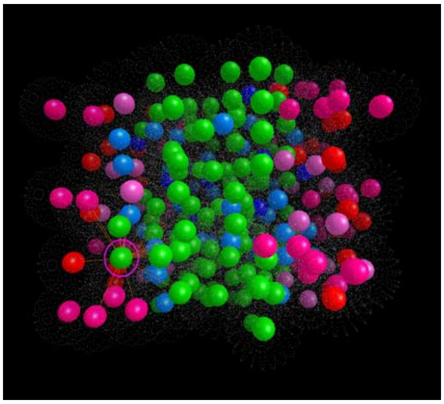

Fig. 1. An in silico minimal multicellular system (mMCO) is captured at a particular stage of 4-dimensional development. In this case we have a bilaterally symmetric 672 cell organism at the 267 cell stage. Cells in different states of differentiation are shown in different colors. The user can choose which cells to view dynamically. The mMCO was developed and simulated using Cellnomica's systems biology mMCO software suite.

10. The semantics of genomes

If we can relate the artificial genome that generates our artificial organism with the natural genome that generates the natural organism, we have the beginnings of a translation of one genome into the other (see Fig. 2). Much like translating English into German, we need to understand what the words are and how they are combined or related into sentences. This is

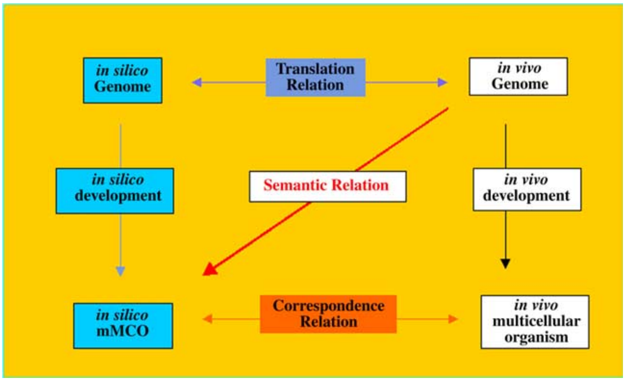

Fig. 2. Shown are the relationships between an in silico mMCO and the in vivo multicellular system in models and emulates. The *translation relation* is a syntactic relationship between the in silico and in vivo genomes. The *semantic relationship* at the center interprets the in vivo system in terms of the in silico mMCO. It relates the syntax of the genome of the natural, in vivo organism with the dynamics of development of the in silico organism. The *correspondence relation* compares both the temporal and morphological development of the systems. The in silico model makes predictions and is corrected via feedback between in silico and in vivo experiments.

called syntax of the language. But first, we need to know the meaning or the semantics of the language. For a translation of one word or sentence or, more generally, sequence is adequate only if the meaning of the two sequences is the same. In logic and linguistics, we call such a theory of meaning a semantics. We need a *semantics of genomes*. A semantic code is more than the regulatory code [6], which is restricted to the logic of gene activation and repression. Genome semantics assigns meaning to a regulatory code or network by way of its function in the cell and the multicellular system [4,7–10].

11. Cracking the semantic code

And, the best way to get at that semantics is to see if we can generate the same structures in our artificial organism when the natural genome is translated into our artificial genome. But it goes both ways. Given we have an artificial genome that generates an organism with a set of systemic properties, if we have the correct semantics can translate that artificial genome into a natural genome. We then insert that genome into a host cell and we should be able to grow in vitro or in vivo a natural organism that has the same or similar systemic properties as the artificial, in silico organism.

12. An in silico Jurassic Park?

If we have such an experimental confirmation, we can then test modifications of the artificial genomes and see if again we have a property isomorphism. Gradually, we would gain significant confidence that our translation was correct and that we had at least a partial semantics of natural genomes. Once, we have that we could design multicellular systems based on those known properties or we could look at new natural genomes and see how they work. More precisely, we could predict how the multicellular system will develop and function in advance of seeing the natural system. Ultimately, we would not need to form a Jurassic Park; we could observe the growth of the animals in silico.

On a more modest, realistic level such a system could reduce or eliminate the need for animal testing. We could do some of the experiments in silico. Tissue design can be made fault tolerant in software. Our understanding of the etiology and dynamics of multicellular diseases could be helped. Our control, for better or worse, of nature would certainly take a grand step forward.

This process of translation gives us a test of adequacy that goes beyond the genome. It places the genome in the context of the cell and then in the context of the development of that cell into a multicellular system. It perhaps is the best way to understand the semantics of genomes.

References

[1] Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. Cell 100, 57–70.

[2] Kitano, H. (2002) Computational systems biology. Nature 420, 206–210.

[3] Huang, S. (2000) The practical problems of post-genomic biology. Nat. Biotechnol. 18, 471–472.

[4] Crick, F. (1958) On protein synthesis. Symp. Soc. Exp. Biol. 12, 138–163.

[5] Crick, F. (1970) Central dogma of molecular biology. Nature 227, 561–563.

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

[7] Werner, E. (1988) Toward a theory of communication and cooperation for multiagent planning, theoretical aspects of reasoning about knowledge in: Proceedings of the Second Conference, pp. 129–143.

[8] Werner, E. (1996) What ants cannot do in: Distributed Software Agents and Applications (Perram, J.W. and Müller, J.P., Eds.), Springer-Verlag, Berlin.

[9] Werner, W. (2005) The future and limits of systems biology. Sci. STKE (to appear).

[10] Werner, E. (2003) In silico multicellular systems biology and minimal genomes. DDT 8, 1121–1127.