Toward a New Theoretical Framework for Biology

Tim Otter

Crowley Davis Research
280 South Academy Ave., Suite 140
Eagle, ID 83616

Abstract. To help us understand the nature of information processing in living systems, biology needs a new theoretical framework that is amenable to computational approaches. Recent progress in cell and molecular biology has made it clear that the "central dogma" (DNA → RNA → protein) is an overly simplistic concept. The goals of this paper are: 1) to specify which aspects of the central dogma are useful and valid while identifying what is missing, inaccurate or misleading, and 2) to suggest a strategy for designing computational approaches that consider the hierarchical organization of living systems and that place the central dogma in the context of living cells.

1 Introduction

Recent progress in cellular and developmental biology has been staggering, and yet, cracks have started to appear in the foundations of biology [4]. Specifically, a call has arisen for a new theoretical framework for biology, one that merges empirical with computational approaches [1,2,6,8]. This should not be seen as a failing of biology, but rather, part of a natural progression of scientific thought driven by the need to set the theoretical underpinnings of cell structure and function, gene expression, development, and physiology in a context that makes them amenable to formal computational approaches [7,11].

Tantalizing as this sounds, formulating such a theory is a daunting task. What would it contain or provide? What problems would it address? Not everyone agrees that such a theoretical framework can be constructed, or if it can, how to begin and what steps to follow. I am proposing that we begin this deep and difficult task by cleaning up the central concepts of cellular and molecular biology that emerged in the latter part of the 20th century following the elucidation of the structure of DNA in 1953 [2]. The goal is to specify and retain what is useful and valid while identifying what is missing, inaccurate or misleading. This paper focuses on the "central dogma of molecular biology", in shorthand: DNA → RNA → protein.

2 Problems with the Central Dogma

The central dogma (also called the "flow of genetic information"; Figure 1A) is one of the main theoretical tenets in cell biology and genetics. To be more complete the diagram should include replication of DNA, and the fact that some DNA segments are transcribed into RNA but not further translated into protein (tRNA, rRNA). Adding intermediate steps between DNA and RNA (RNA processing & splicing to remove introns) make the concept a more accurate representation of gene expression in eukaryotes.

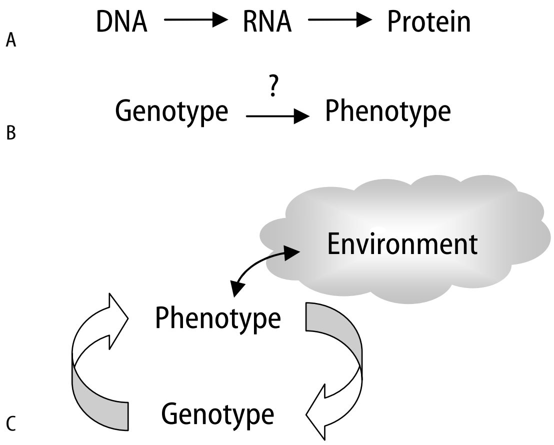

Figure 1. In its classic form (A), the central dogma indicates a linear relationship between genes and traits; this simple 1:1 mapping does not explain the global relationship between genotype and phenotype (B). Revised (C) the relationship is more complex: circular and open to signals from the environment (see text).

Still, even with these features added, the central dogma, though technically correct for a single gene, is incomplete: in the words of Reed [11], "a naïve fiction". As written, the central dogma is inadequate to explain more generally how the phenotype (traits/ appearance) is related to or derived from the genotype (genes) (Fig. 1B). The "one gene-one enzyme" theory[1] of gene expression explains how the genetic code is organized in blocks that each corresponds to a particular protein, when the code is translated during protein synthesis. Each protein carries out a particular function. By extrapolation, then, one might conclude that the phenotype[2], what an organism does and how it appears, is the aggregate of all expressed genes in a cell, and that the phenotype is specified in the genotype (in its DNA). The end point of this line of reasoning has been called "genocentric" or "genetic determinism", a view that every trait in the organism is specified in its genetic code. However, extrapolating from a

[1] In the case of eukaryotic multi-subunit proteins this must be restated as the "one gene-one polypeptide" theory, taking into account that a gene may include several distinct blocks of code (exons) that are spliced together by removal of the intervening sequences (introns). This revision does not alter the fundamental 1-to-1 correspondence between the genetic unit (a gene) and the functional protein unit (a polypeptide).

[2] The term phenotype derives from the Greek root *phenos*, meaning "appearance" (e.g., "phenomenon"). Phenotype includes the organism's physical traits, metabolic state, stage of development, and other discernable characters. Unlike the genotype, phenotype is a complex, spatially-ordered and temporally-defined state.

single gene model to the genome is too simplistic: development is much more complex than a 1-to-1 mapping of genotype onto phenotype.

Because every cell in a multicellular organism[3] expresses only a limited subset of its genes at any given time, we must ask whether the information as to where and when a particular gene is expressed resides in the genome. The on/ off switches themselves, the regulatory machinery, includes segments of DNA that control transcription, but the signaling molecules that turn genes on or off are proteins[4]. The amino acid sequences of these regulatory proteins are, of course, encoded by other genes, but there is no genetic program that specifies directly when or where genes are transcribed (see [4] Chap 4).

This example points out that the phenotype includes regulatory functions and so the phenotype directly controls gene expression. While the genotype evokes the phenotype, genetic information does not by itself determine what an organism becomes or does. In fact, the cell provides the means to express the genetic code. All of the processes involved in the 'classic' central dogma –replication, transcription, splicing, capping, and translation—are carried out by the metagenetic apparatus of a cell, its molecular machinery. Thus, the relationship between genotype and phenotype becomes complex and circular, not linear, and linked to signals from the cell's environment (Figure 1C).

Control of gene expression in a given cell is triggered by its environment –growth factors, other chemicals, and electrical or mechanical signals. In this way, development of a phenotype depends on two kinds of information: genetic and environmental. The information contained in a cell's genetic code is part, but only part, of the information that living cells process. In other words, living cells have access to at least two kinds of information: the hereditary information stored in its DNA and the information the cell gathers (and records) about its environment.

A genotype is a set of possibilities; the phenotype is what develops in a particular set of conditions.

3 The Problem of Levels

The example above is but one of many reasons why the central dogma must be revised. Polanyi [10] has argued that living organisms "… form a hierarchy in which each higher level represents a distinctive principle that harnesses the level below it (while being itself irreducible to its lower principles)…". Accordingly, an organism's phenotype represents a higher level than its genotype, and so in Figure 1C this is reflected by the placement of phenotype above genotype, not at the same level. This

[3] For the sake of argument, this discussion focuses on complex multicellular animals with clearly differentiated tissues and organs, but with appropriate modifications it could apply equally well to plants, fungi, or any organism with a multicellular body.

[4] Recently RNA-based genetic control ("riboswitches") has been identified in bacteria, but its importance in gene regulation among eukaryotes is an open question.

difference is possible because part of the genetic code includes instructions for making sensory devices that can detect either chemical, electrical, or mechanical signals. Once a cell builds and deploys sensors, it begins to collect information (that is not genetically encoded) about its environment. This gives it access to periodically updated or continuous signals that allow a primitive sense of time, and memory (*cf*. current state with previous state) to develop (discussed by Pattee [9]). By recording such signals and learning to recognize patterns in them that are crucial to survival, living organisms gain access to a type of information that is constrained by, but not derived from, both physical laws and genetic code.

In other words, genes determine the types of sensors a cell can make, but genes do not specify the patterns of information a given cell will receive nor the kinds of responses it makes to any given signal. The responses are harnessed by the cell's machinery[5] (to use Polanyi's terms) to useful purposes, survival and reproduction.

With Polanyi's analysis in mind we can construct a provisional hierarchy for living organisms (Figure 2). According to this view, physical and chemical laws are the most fundamental, but alone they are probably not adequate to explain (or derive) life's processes or its organization. The genetic code, the sequence of a cell's DNA, is placed on a higher level because it is subject to additional constraints layered on top of the physical laws. In addition to following chemical and physical laws (e.g., that govern covalent bonds and weak bonds, molecular stability, folding into a helix, etc.), DNA's structure, its sequence of bases, has been subject to constraints placed on it by the selective pressures of evolution, so the nucleotide sequence reflects an organism's evolutionary history. Thus, predicting a cell's DNA sequence simply by applying the laws of chemistry and physics would be difficult, if not impossible. Besides, there is little point in taking on such a challenge when we can determine the genotype empirically.

Above the genetic code but relying upon it are the cell's macromolecular machines (e.g., ribosomes, microtubules, membranes, or complexes of metabolic enzymes) formed by the assembly of several or many components, proteins or segments of RNA that are encoded by different genes. Still higher are the metabolic and regulatory aspects of phenotype, organized as networks, including: 1) processes and molecules involved in cell signaling –production, transmission, and reception of signals; 2) processes and molecules of metabolism, and; 3) the processes and molecules of homeostasis, growth, repair, and development. The molecules of these

[5] The machine is a widely used metaphor for living organisms and their components. All metaphors are equations that literally are false, but interesting enough that for cause we ignore the inequality. In some sense, then, living organisms are machines, if we qualify that they can build, repair and reproduce themselves and seek energy. Machines are examples of composite structures made or assembled from a collection of parts. Each part has its own characteristics, but in the machine the parts interact and bear relationships to one another that are constrained so as to harness the function of the parts toward a specific purpose. In this way, machines are capable of carrying out tasks that the parts cannot do individually or as an unassembled group. Thus, claiming a cell is a machine helps us understand some important aspects of living cells, but when pushed to its limit, the statement loses its value.

regulatory networks are mostly proteins (plus some RNA), and their sequences are specified in the genetic code, but, as far as we know, the structure of any network is not specified in the genetic code[6]. To understand networks, we must map the pathways involved and identify their points of convergence. Likewise, we probably cannot predict how a cell gets energy from its genetic code; we simply need to find out what it eats.

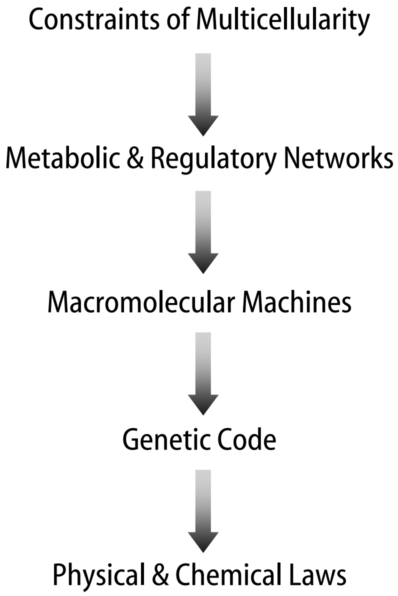
Constraints of Multicellularity
Metabolic & Regulatory Networks
Macromolecular Machines
Genetic Code
Physical & Chemical Laws

Figure 2. A provisional hierarchy for living organisms. Each level constrains the levels below it (after Polanyi [10]).

All of the above components and attributes are found in living cells, whether a unicellular organism or part of a multicellular body.

Further up in the hierarchy are those processes and relationships that constrain individual cells to function as part of a group –tissue, organ, or organism. These are the constraints (developmental, physiological, ecological, or evolutionary (descent by modification of ancestral forms)) placed on cells that are parts of a multicellular organism. It seems implausible to derive these properties from physical laws or the genetic code or any combination of information or rules that govern lower order phenomena. Instead we must measure parameters that describe higher order structure and function. For example, as multicellular bodies develop specialized regions or

[6] Brenner [2] makes a bolder claim: “Genomes do not contain in any explicit form anything at a level higher than the genes. They do not explicitly define networks, cycles or any other cluster of cell functions.”

clusters (e.g., an organ that performs a limited set of tasks), there develops an obligatory dependency on other clusters, and consequently, a heightened requirement for communication and feedback among the specialized clusters. Therefore, knowing what each cluster does (its specialty) and what it needs (what kinds of things, how much, at what rate, etc.) are prerequisite to understanding integrated function.

In considering the apparent hierarchical organization of living systems, are the levels we have described valid measures from the organism's perspective or simply logical constructs of the human mind? While this raises questions that go beyond the scope of this paper, one must recognize that one of the hallmarks of living systems is feedback control, and when such control operates between (perceived) levels, distinguishing between the components of one level or another becomes difficult or impossible (discussed by Huang [6] and by Reed [11]).

4 Emergent Properties

One of the central concepts we must consider is emergence (emergent properties), defined simply as those properties of composite structures that are not evident in the parts (components) or in an arbitrary arrangement of them. Accordingly, the disassembled pieces of a watch lack the emergent property of an intact watch, keeping time[7]. In this example, precise arrangement of the parts constrains their movements so as to produce (emergent) time-keeping.

While few would argue whether emergent properties[8] exist, there is considerable debate as to how they arise. By definition, a composite structure consists of its parts, no more and no less, but when assembled the relationships of the parts to one another are precisely defined and constrained. So emergence must derive from the nature of these constraints and the relationships among the parts in the composite structure.

Polanyi [10], Conrad [3] and Pattee [9] have emphasized significant stumbling blocks associated with emergence in hierarchical systems. One set of problems stems from the indirect relationship between genotype and phenotype, as illustrated below.

[7] Time-keeping also requires a source of energy. Many emergent properties in biological systems are also energy-dependent.

[8] Properties such as self-organization, self-assembly, and self-reproduction are often used to illustrate emergent properties that are inherently biological (see, for example Garibay and Wu [5]). However, this terminology needs to be reexamined in light of the confusion about levels. While it is fair to say that living cells reproduce (themselves), applying similar terminology to DNA can be misleading because the "self" referred to in "self-replication" is unspecified. Replication of DNA requires at least 5 different types of enzymes, and more if one includes the processes involved in DNA repair to ensure that copying mistakes are corrected. In living cells, DNA replication is also tied to the cell cycle and a host of control pathways associated with it.

4.1 Molecular specificity

Specificity of interaction between proteins or between enzymes and their substrates depends on precise molecular fit, complementary shapes that are stabilized by a number of weak (~0.1 -5 kcal/ mole) inter-atomic bonds. How is a protein's shape, and hence its function, encoded? The ultimate, functional shape that the properly folded polypeptide chain will have (its tertiary structure) depends on the sequence of amino acids (its primary structure), which is basically a codon-by-codon readout of the DNA sequence of the gene that encodes it. Folding involves local interactions between neighboring amino acids to produce α-helixes and β-sheets, which associate to form higher order domains. So, in a sense, for at least some proteins, the genetic information (genotype) specifies, through the folding process, a protein's shape and therefore what kinds of complementary shape(s) that protein can bind to (its function or phenotype). Protein folding is often cited as an example of (emergent) self-organization.

However, primary sequence alone is not sufficient to produce the functional, folded shape of many proteins. There are several reasons for this. To fold properly, some proteins require other proteins called molecular chaperones to help them along. In other cases, the translated amino acid sequence actually specifies an inactive precursor (e.g., proinsulin). Activation may involve cutting by an enzyme or covalent modification of one or more amino acids (e.g., phosphorlyation or acetylation) or binding of ligands (e.g., Ca^{2+}, cAMP, ATP or GTP). With any of the above examples we could not hope to infer the function of a protein simply on the basis of its (genetic) sequence. In each example, functionality depends on some other molecule that is not encoded in the gene for the protein of interest. The other necessary molecule may be a protein specified by another gene or it may be a metabolite. Furthermore, some proteins may be part of a stimulatory pathway in one type of cell but the same protein may be inhibitory in another cell. In this case, "function" is not intrinsic to the protein, but it is defined contextually. The bottom line is that for many proteins, one of the most fundamental aspects of phenotype, molecular specificity, cannot be inferred from the information in a single gene. We must also know about the modifications to the protein's structure and function that occur in the living cell, and we must understand the context in which the protein functions.

Conrad [3] refers to molecular specificity as the basis of pattern recognition tasks, a form of computation, carried out constantly, rapidly, and with great fidelity in living cells. In context, then, Conrad observes that there is "…a highly malleable relationship between the level at which the computation is executed (the level of folded shape) and the level at which it is tunable (the level of amino acid sequence)." In other words, a protein's function, its phenotype, is the product of the evolution of that protein's gene sequence plus the evolution and function of any other genes that are required to render that protein fully active. Therefore, to generalize from the simplest case where no such controls happen to apply (the protein folds spontaneously into its active conformation, requires no cofactors for activity, and is not modified after translation) grossly oversimplifies the global relationship between genotype and phenotype.

5 Computational Approaches

Given this hierarchical nature of living systems, our attempts to formulate a theory focus on the relationships between levels, and whether we can understand one level by calculation from the levels below it. This is a difficult question to answer, and approaching it draws upon elements of computational theory and information theory.

The goal of this discussion is not to propose specific computational strategies but rather, to set some general guidelines that are consistent with the analysis presented above, wherein the processes of the central dogma are set in an appropriate context, the living cell. One of the main goals of computation is to be able to predict how living systems will respond to perturbation, to be able to predict what a cell (or tissue or organ) will do without having to run the experiment. As explained by Brenner [2], this goal can best be met by simulation. Brenner writes:

"Building theoretical models of cells would be based not on genes but on their protein products and on the molecules produced by these proteins. …I want the new information embedded into biochemistry and physiology in a theoretical framework, where the properties at one level can be produced by computation from the level below. ... This computational approach is related to Von Neumann's suggestion that very complex behaviours may be explicable only by providing the algorithm that generates that behaviour, that is, explanation by way of <u>*simulation*</u>*.* (emphasis mine) *…A proper simulation must be couched in the machine language of the object, in genes, proteins and cells."*

The obstacles to simulation are both theoretical and practical. How can we calculate in the "upward direction" (against the arrows in Figure 2), from the constituents of lower levels of the hierarchy toward higher levels? The key here, I believe, will be to include information about the constraints operating at the upper level that harness lower level elements toward a specific purpose. This approach, which is based on specifying rules that govern higher order relationships, serves a dual purpose: it provides information about higher level features and makes our calculations more manageable by decreasing the size and complexity of the solution space. We do not need to examine all interactions of the components, but only those that sustain the higher order properties we have specified.

Polanyi [10] and Pattee [9] have argued that "upward" calculation will be difficult, if not impossible. Part of the difficulty lies in determining which higher-level properties may be the best ones. Brenner's [2] insistence on using "the machine language of the object", while reasonable, may be difficult to achieve. The central dilemma is knowing whether the properties one selects are simply convenient measures chosen by the experimenter or whether they are meaningful parameters to the system itself (e.g., a living cell). Whether this issue is framed as a "from-at" problem (Polanyi) or a "problem of observables" (Pattee), the best approach at this stage is to tap the deep well of information on cellular structure and function that has been developed during the last 100 years.

References

1. Bray, D.: Reasoning for results. Nature **412** (2001) 863

2. Brenner, S.: Theoretical Biology in the Third Millennium. Phil. Trans. R. Soc. Lond. B **354** (1999) 1963-1965

3. Conrad, M.: Molecular computing: the lock-key paradigm. IEEE Computer **25**(**11**) (1992) 11-20

4. Fox-Keller, E.: Making Sense of Life. Harvard University Press, Cambridge, MA (2002)

5. Garibay, I., Wu, A.: Cross-fertilization between proteomics and computational synthesis. In: AAAI Spring Symposium series (2003) 67-74

6. Huang, S.: The problems of post-genomic biology. Nature Biotechnology **18** (2000) 471-472

7. Lazebnik, Y.: Can a biologist fix a radio –or, what I learned while studying apoptosis. Cancer Cell **2** (2002) 179-182

8. Nurse, P.: Understanding cells. Nature **424** (2003) 883

9. Pattee, H.H.: The problem of observables in models of biological organizations. In Evolution, Order, and Complexity, Khalil, E.L., Boulding, K.E. (eds.) Routledge, London (1996) 249-264

10. Polanyi, M.: Life’s Irreducible Structure. Science **160** (1968) 1308-1312

11. Reed, M.C.: Why is mathematical biology so hard? Notices Amer. Math. Soc. **51**(**3**) (2004) 338-342