

Gene and Body: Building and Maintaining the Phenotype of Living Organisms

Tim Otter and Robert Davis

Crowley Davis Research
280 South Academy Ave., Suite 140
Eagle, ID 83616
timmo@cdres.com

Abstract

Biological development is a progressive process, building from the fertilized egg via the embryo to the adult body. Development involves coordinated, stage-specific expression of genes, but the unidirectional, 1:1 mapping of genotype onto phenotype implied by the “central dogma” (DNA → RNA → protein) is not adequate to explain what happens during development. This paper explores the global relationship between genotype and phenotype in a manner that places the central dogma in the context of living cells.

Introduction

In simple terms, development is the process of becoming an adult. Modern developmental biology –the study of those phenomena, processes, and mechanisms that lead from fertilized egg via embryo to adult– involves the fusion of two disciplines, embryology and molecular genetics. One of the main goals of developmental biology is to understand how genes shape the adult body.

Recently, as technologies for large-scale sequencing of DNA have come available, the composition and organization of entire genomes (human, mouse, pufferfish (*Fugu*), nematode worm (*Caenorhabditis elegans*), fruit fly, plant (*Arabidopsis thaliana*), and others) have been determined. This broadening of perspective has led, in turn, to pressing needs for computational approaches to handle large (genomic) data sets and for a theoretical framework for biology that can help to model the process of development in quantitative terms.

Merging of biology with computer science is a two-way street: biologists stand to gain more powerful and quantitative modeling tools while computer scientists benefit by incorporating biological principles to produce more robust and fault-tolerant computational strategies. However, until recently biologists have not generally framed their questions in this way, and so bringing these two disciplines together has revealed some problems and inadequacies in the conceptual framework of biology. In the interest of promoting fuller collaboration, this paper explores the global relationship between genotype and phenotype, between the information contained in an organism’s genome and the processes for building and maintaining a multicellular body.

As explained below, the key is understanding the cellular context in which the genes operate. Establishing and maintaining order requires energy, and part of what a cell provides is a supply of energy to support its high degree of organization. The phenotype is a complex, spatially ordered and temporally defined state, but the genotype is essentially a parts list. In the words of Franklin Harold (2001), it is best to think of a cell “...as a spatially structured self-organizing system made of gene-specified elements. Briefly, the genes specify What; the cell as a whole directs Where and When...”.

One Gene, One Protein?

Beadle and Tatum’s classic (1941) studies on metabolic pathways established the fundamental 1:1 correspondence between the genetic unit, a gene, and a phenotypic property, a functional enzyme. Subsequently their “one gene, one enzyme” hypothesis has been modified and updated to allow for multi-subunit enzymes, proteins without enzymatic activity, and splicing together of coding regions by removal of introns, but Beadle and Tatum’s essential concept –the correspondence between gene and enzyme activity– remains.

Soon after Watson and Crick proposed the double helix model of DNA (1953) the genetic code was deciphered, and by 1967 the collinear nature of genes and proteins became apparent. Thus, the correspondence between genes and proteins holds at the level of their building blocks, whereby 1 codon, a triplet of DNA nucleotides, specifies 1 amino acid in the corresponding protein. This relationship is summarized in the central dogma of molecular biology:

DNA → RNA → protein

or more simply,

gene → protein

As long as the encoding gene has the proper sequence of nucleotides, the resulting polypeptide will have the correct sequence of amino acids. The ultimate, functional shape of the properly folded polypeptide chain depends, in turn, on its sequence of amino acids. Folding involves local interactions between neighboring amino acids to produce

α -helices 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 thereby what kinds of complementary shape(s) that protein can bind to (its function or phenotype). Mutations that change the sequence of amino acids so as to alter the folded protein's shape thereby impair (very rarely, enhance) its function.

In sum, at the level of individual genes that code for single-subunit enzymes, there is a 1:1 correspondence between a gene and the function carried out by the encoded protein. By extrapolation, then, one might conclude that the phenotype, which includes the organism's physical traits, metabolic state, stage of development, and other discernable characters, is simply the aggregate of all expressed genes in a cell, and that every trait in the organism is therefore specified in the genome (in its DNA).

However, things are not so simple. As the following examples illustrate, function often derives from complex interaction of proteins encoded by different genes.

For many proteins, the correct amino acid sequence alone is not sufficient to produce the functional, folded shape. There are several reasons for this. In some cases, the translated amino acid sequence actually specifies an inactive precursor (e.g., proinsulin), and activation may involve cutting by an enzyme or covalent modification of one or more amino acids (e.g., phosphorylation or acetylation). As a result, we could not hope to infer the function of such a protein simply on the basis of its (genetic) sequence; functionality depends on a protein specified by another gene.

To fold properly, some proteins require other proteins called molecular chaperones to help them along. Folding of tubulin, for example, is guided by a complex of chaperones called TRiC or CCT (Leroux and Hartl 2000). Tubulin is found only in eukaryotic cells, where its role is to assemble into rod-shaped polymers called microtubules that help to establish cell polarity, maintain cell shape, and move chromosomes during mitosis. Assembly of microtubules requires precise, complementary molecular fit of the tubulin subunits. Ill-formed tubulin does not assemble. That is why, when tubulin genes are cloned in bacteria (prokaryotes), tubulin protein is made but no microtubules form: bacteria lack the chaperones needed to render tubulin assembly-competent.

Thus, the same gene, one that functions properly in its natural eukaryotic setting, becomes functionally bankrupt in a prokaryotic cell. Clearly, tubulin's function is not an exclusive property of the encoding gene, but it is defined in part by the cellular context.

Most proteins do not function in isolation but instead are components of macromolecular machines (e.g., ribosomes, membranes, or complexes of metabolic enzymes) whose function integrates the activity of several different proteins or segments of RNA that are encoded by different genes. The replication fork machine is a macromolecular complex

that performs one of the most fundamental tasks in living cells: it copies the DNA, so that during cell division each daughter cell receives a full set of genetic information.

DNA replication actually involves many steps and processes, all of which must be coordinated so that the machine copies both strands of the DNA. DNA polymerase, the enzyme that copies and corrects its own errors ("proofreads") moves in one direction only along the template strand of DNA. Since the two strands of the double helix run antiparallel, to copy the other strand the replication fork machine makes a series of discontinuous fragments and then links them with another enzyme, DNA ligase. When all is said and done, more than a dozen proteins, several of these with multiple subunits, comprise the replication fork machine (Table 1).

<u>Protein</u>	<u>Function</u>	<u>Subunit Structure</u>
<i>ORC</i>	<i>Recognize start (origin)</i>	<i>multiple monomers</i>
<i>Helicase</i>	<i>Unwinds DNA helix</i>	<i>6-mer</i>
<i>Gyrase</i>	<i>Relieves supercoil tension</i>	<i>4-mer: 2 x 2 subunits</i>
<i>SSB</i>	<i>Stabilizes ssDNA</i>	<i>multiple monomers</i>
<i>Primase</i>	<i>Synthesizes RNA primer</i>	<i>4 subunits</i>
<i>Clamp</i>	<i>Slides DNAPol along helix</i>	<i>ring-shaped dimer</i>
<i>Loader</i>	<i>Loads & positions clamp</i>	<i>5 subunits</i>
<i>DNAPol</i>	<i>Copies & proofreads</i>	<i>3 subunits</i>
<i>Ligase</i>	<i>Seals fragments</i>	<i>1 subunit</i>
<i>RNAase</i>	<i>Removes primer</i>	<i>1 subunit</i>
<i>~5 more</i>	<i>Repair & mitochondrial replication</i>	

Table 1. Molecular composition of a generic machine for replication of DNA. The names of proteins vary from species to species, but their functions (middle column) are conserved.

Thus, to claim that all cells need to copy their DNA is DNA polymerase would be absurd. The cell must first unwind the helix (helicase), relax the resulting supercoil twists (gyrase), stabilize the single stranded DNA (SSBs), copy and proofread (DNA polymerase), start the fragments along the opposite strand with an RNA primer (primase), link the fragments (ligase), and so forth.

The bottom line is that for some proteins, one of the most fundamental aspects of phenotype, activity or function, 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.

In summary, 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, is not modified after translation, and is not a component of a macromolecular machine) grossly oversimplifies the global relationship between genotype and phenotype.

Genome, Genotype, and Phenotype

Consequently, the central dogma is inadequate to explain more generally how the phenotype (traits/ appearance) is related to or derived from the genotype (genes) (Figure 1A). Every cell in a multicellular organism expresses only a limited subset of its genes at any given time, so we must ask whether the instructions for where and when a particular gene should be expressed reside in the genome. It is often stated that the genetic instructions specify the characteristics of an organism, and that development is simply the execution of a program encoded in the genes, but genotype does not ‘produce’ phenotype any more than a list of ingredients produces a pie. The signaling molecules that turn genes on or off are regulatory proteins encoded by other genes, but there is no genetic program that specifies directly when or where genes are transcribed (see Fox-Keller 2002, Ch. 4). As far as we know, the genome does not specify when or where genes are transcribed, nor how proteins assemble into macromolecular machines, nor signaling or metabolic networks, nor any other aspect of cellular function on a higher level of organization. During development the genes are expressed and controlled by the metagenetic apparatus of living cells, the biochemical machinery that carries out molecular processes that are organized in both specific location and direction.

The term ‘genome’ is itself a source of some confusion. Loosely, it means ‘a complete set of genetic information’, but when a more precise definition (‘the genes contained on a single set of chromosomes’) is offered, its meaning becomes less clear, especially when the sufficiency of the genome to specify the phenotype is implied. For example, this definition works fairly well for human females, but not for males where genes on all 22 autosomes plus the X and the Y chromosome are needed. Genome derives from two greek roots –*gene*, meaning race or tribe, and the suffix –*ome*, meaning mass or tumor. The latter connotes a disorganized lump, collection or set, but we know that to function properly each gene must be located near its appropriate control sequence or cis-regulatory region. Translocation can result in significant changes in phenotype or cause disease, and yet ‘genome’ usually refers to a list of the genes themselves, their positions unspecified. This suggests a larger problem, what is lost in the blender, the structural organization of a living cell.

In this paper, ‘genome’ includes all of the information contained in an organism’s DNA, including the protein-coding regions, introns, control sequences, centromeres, telomeres, genes that specify RNA products (e.g., tRNA, rRNA, or spliceosome components), pseudogenes, and

repetitive sequences. Only ~1.5% of the human genome codes for proteins and the remaining 98.5% includes the other components listed above. Genotype refers more specifically to which versions of those coding genes an individual has.

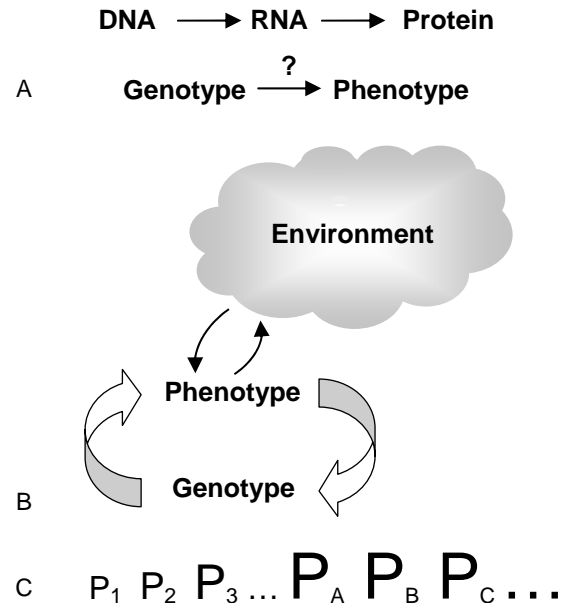


Figure 1. (A) The central dogma outlines the steps involved in expression of a single gene, but this simple 1:1 mapping does not explain the global relationship between genotype and phenotype. Revised, (B) the relationship is more complex: circular and open to signals from the environment (see text). During development the phenotype (P_1, P_2, \dots) changes rapidly as growth and differentiation proceed; during adulthood (P_A, P_B, \dots), phenotype appears static, but repair and replacement continue (C).

‘Phenotype’, derived from the Greek *phenos*, meaning appearance, refers to the organism’s discernable traits. Here again, when one looks more closely, ambiguities arise. Macroscopically, for example, phenotype may appear normal (without symptoms) but a blood test reveals abnormally low levels of a particular enzyme. Should the results of the blood test be considered as part of the phenotype? Within limits lower enzyme levels may not matter, or the body may compensate, but when we describe a phenotype in molecular and quantitative terms, we must specify when, where, and in what amount for each protein.

During development the phenotype changes drastically as the zygote cleaves to form a multicellular embryo, embryonic tissues arise, and differentiation into specialized organs takes place. The developing embryo is constructing itself; part of this process involves shaping an environment that is suitable for development (Figure 1B).

Control of gene expression in a given cell is triggered by its environment –growth factors, other chemicals, and electrical or mechanical signals. While the genotype evokes the phenotype, genetic information does not by itself determine what an organism becomes or does. Instead, development of a phenotype depends on two kinds of information: the hereditary information stored in its DNA and the information the cell gathers (and records) about its environment. In other words, the information contained in a cell's genetic code is part, but only part, of the information that living cells process.

The insufficiency of the DNA for determining what an organism will become –the form and appearance it takes on during development– has been acknowledged for some time, but recent cloning experiments have driven this point home. Successful cloning of a domestic cat from a female nuclear donor with calico coloring (mottled orange, white, and black fur) produced a kitten (clone) whose coat color does not closely resemble that of its nuclear donor “parent” (Shin et al. 2002). This is possible because coat color patterns are controlled during development by both genetic and epigenetic factors. Thus, a genetic clone is not necessarily a phenotypic copy of the donor. Furthermore, many cloned animals survive for only a short period of time. Shorter lifespan apparently results because long-term survival requires that roughly half the genes must be derived from each parent. This suggests that, in addition to the information coded in the DNA, a genome made by joining two haploid sets contains essential information about ancestry, and it emphasizes the continuity from one generation to the next. Parents contribute more than just their DNA code.

Accordingly, an organism's phenotype represents a higher level than its genotype, and so in Figure 1B this is reflected by the placement of phenotype above genotype, not at the same level. This difference is possible because part of the genetic code includes instructions for making sensory devices that can detect either chemical, electrical, or mechanical signals. 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 way it responds to any given signal. 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. By recording such signals and learning to recognize patterns that are crucial to survival and reproduction, living organisms gain access to a type of information that is constrained by, but not derived from, both physical laws and genetic code. Thus, the relationship between genotype and phenotype becomes complex and circular, not linear, and linked to signals from the cell's environment (Figure 1B).

Furthermore, as multicellular bodies develop specialized regions or 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. Accordingly, one of the hallmarks of living systems is feedback control. 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, whether in the developing embryo or in the adult body.

In contrast to a developing embryo, an adult's phenotype appears static, but this appearance is deceiving (Figure 1C). Living organisms are engaged in ongoing processes of repair, replacement, and regeneration. Growth of hair, claws, or nails, sloughing of dead skin, renewal of blood cells, and protein turnover all attest to this dynamic state. Some organisms, such as salamanders, the simple *Hydra*, and most plants have a greater capacity for regeneration than others (e.g., humans) do, but even adult human bodies contain pockets of stem cells with substantial regenerative capacity. Recently, the discovery that zebrafish hearts are capable of regeneration has led biologists to rethink the notion that terminally differentiated tissues are incapable of significant regrowth (Poss et al., 2002).

Given the continuity of evolutionary processes, that evolution proceeds by descent by modification of ancestral forms, it makes sense that the processes that guide development of evolutionarily primitive multicellular organisms (e.g., *Hydra* and many protists) apply, in large measure and with appropriate modification or embellishment, to the development of complex, highly differentiated organisms.

At some fundamental level, then, in terms of gene transcription, synthesis of proteins, and degradation of damaged or obsolete molecules and cells, the processes that build multicellular organisms are very similar to those that maintain and repair the adult body. They involve the same genes, operating in the same cellular context, interacting by the same kinds of feedback controls.

References

- Fox-Keller, E. 2002. *Making Sense of Life*. Cambridge, MA: Harvard University Press.
- Harold, F. 2001. *The Way of the Cell*. New York, NY: Oxford University Press.
- Leroux, M.R. and Hartl, F.U. 2000. Protein Folding: Versatility of the Cytosolic Chaperonin TRiC/CCT. *Current Biol.* 10:R260-R264.
- Poss, K.D., Wilson, L.G., and Keating, M.T. 2002. Heart Regeneration in Zebrafish. *Science* 298:2188-2190.
- Shin, T., Kraemer, D., Pryor, J., Liu, L., Howe, L., Buck, S., Murphy, K., Lyons, L., and Westhusin, M. 2002. A Cat Cloned by Nuclear Transplantation. *Nature* 415: 859.