Genetics as Explanation: Limits to the Human Genome Project

Irun R Cohen, *The Weizmann Institute of Science, Rehovot, Israel*

Henri Atlan, *Hadassah University Hospital, Jerusalem, Israel and Ecole des Hautes Etudes en Sciences Sociales, Paris, France*

The genome has been likened metaphorically to a computer program; but an inappropriate metaphor can mislead. The genome does not directly program the organism.

Introductory article

Article contents

doi:10.1002/9780470015902.a0005881

Definitions

Genetics refers to the structure and function of genes in living organisms. Genes can be defined in various ways and at various scales of interest: consider evolution, populations, species, organisms, cells and molecules, or hereditary transmission, embryonic development and life management. These are quite diverse subjects, and the people who study them would seem to use the term gene in distinctly different ways. But genetics as a whole is organized by a single unifying principle, the deoxyribonucleic acid (DNA) code; all would agree that the information borne by a gene is linked to particular sequences of DNA. At the chemical level, we can define a gene as a sequence (or combination of sequences) of DNA that ultimately encodes a protein. The genome refers to the germ-line DNA that an organism has inherited from its progenitors. The genome includes DNA genes along with DNA sequences that do not appear to encode proteins. (***See*** Gene Structure and Organization; Genome Organization of Vertebrates; Protein Coding.)

Now we can define the Human Genome Project: the genome project is a translation project. Its objective is to translate the chemical sequence information borne by the genome into the verbal information of human language and thought; the aim is to translate DNA sequences into words and ideas that can develop and spread among human minds. What we can manage to do with this information depends on how well we understand the functions of genomic DNA within the organism.

Metaphors and Programs

Most minds use metaphors to understand and explain; we grasp the essence of the unfamiliar (or the complex) by seeing its likeness to the familiar (or the simple). Metaphors are not merely literary devices; metaphors, which also include mathematical models, can aid precise thinking. What metaphor is suitable for explaining the function of the genome?

Metaphorically, the genome is often likened to a computer program; just as the computer reads and executes the instructions of its program, the body is proposed to read and execute the instructions borne by the genome. The body, from this point of view, is mere hardware. The genome is the boss.

The computer program metaphor is often extended to explain evolution: evolution is thought to improve DNA programs. Diversification of genomes by random mutation combined with the selection of the most successful variants (survival of the fittest) leads, it is claimed, to the continuous upgrading of existing DNA programs. The evolution of genomic DNA is automatic but costly – the death of the less fit drives the process. (***See*** Evolutionary History of the Human Genome.)

Metaphors and Expectations

The computer program metaphor fosters high expectations of the Human Genome Project. Theoretically, if you know all the information borne by a computer program, you can expect to know how a computer using that program will operate; you can understand the computer's present behavior and can predict its future behavior with a high degree of accuracy. You would even be able to repair mistakes in the program, if that program were simple enough.

Metaphorically then, if the genome is really like a computer program, the genome project will empower us to understand the organism, predict its response to the changing environment and provide a key to the cure of its maladies. Or so many would have wished to believe.

Here we shall discuss what a program means to most people and then test whether the genome actually fits the bill. We shall see that the program metaphor is a misleading way to describe the genome; knowing the genome will not explain the organism. We might do well to consider other metaphors for the genome.

Genome is not a Simple Program

The *Oxford English Dictionary* (second edition, 1989) defines a computer program as 'A series of coded instructions which when fed into a computer will automatically direct its operation in carrying out a specific task'. A computer program is usually written intentionally by a computer programmer; the DNA program, by contrast, is written by evolution, without intention. But irrespective of who or what writes a program, at the very least, a program is a plan for a sequence of events. So most people would like a program to be unambiguous, coherent and definite. The program's task should be inherent in the program itself; the information in a program should be sufficient for the job. A program, like a blueprint, is a type of representation. But the genome, as every working biologist knows, is ambiguous, incoherent and indefinite. Most debilitating to the genetic program metaphor, the genome is not autonomous or complete. Consider the following examples:

- A DNA sequence that encodes a protein in a multicellular organism is usually discontinuous and is interrupted by chains of meaningless DNA (introns). The gene transcript (messenger ribonucleic acid (mRNA)) has to be spliced together by proteins that cut out the introns. Thus most DNA sequences are not intrinsically coherent.
- Many DNA coding sequences, perhaps as many as a third, can undergo alternative splicing to produce different proteins. In other words, a single DNA sequence can give rise to more than one species of protein. Moreover, the way the DNA actually gets spliced is not governed by the DNA sequence itself; proteins actually determine the gene – the spliced DNA sequence that is expressed in particular circumstances. Thus, the information encoded in many DNA sequences is intrinsically ambiguous until realized by the action of proteins.
- The protein products encoded by a gene may also be indefinite: a single protein may assume several functionally different conformations, and so the gene that gives rise to the protein may be said to function in more than one way. Moreover, the protein encoded by the gene can (and does) undergo chemical modifications (enzymatic cleavage, aggregation with other molecules, phosphorylation, glycosylation and so forth) to carry out further functions independent of the gene. The protein glyceraldehyde-3-phosphate dehydrogenase first discovered as an enzyme, for example, is now known to have a role in membrane fusion, microtubule bundling, RNA export, DNA replication and repair, apoptosis, cancer, viral infection and neural degeneration. The protein's gene is obviously not the program of the protein.
- The sets of genes expressed at a particular time are determined by molecules external to the genome; the previous history of the DNA can be overridden. For example, the sheep Dolly was cloned by transplanting a nucleus from an udder cell into an ovum. The molecular environment of the udder cell activated the milk genes of the nucleus; after the nucleus was transplanted to the ovum, the genes needed for making a new sheep became activated. The cellular environment 'reprograms' the genome, epigenetically.
- A single protein can function in very different ways during prenatal development and later in life after development is completed. Thus the gene encoding the protein can be seen to perform different functions at different times; the meaning of the gene varies with the stage of development.
- Some genes can be removed from the genomes of experimental animals (knocked-out genes) without producing an overt change in the form or behavior of the animal. Knocking out other genes, in contrast, can lead to severe and unexpected effects. Scientists who knock out genes are not infrequently surprised by the resulting phenotype of the animal. In other words, the impact of a gene on an organism is not readily deducible from knowledge of its DNA sequence.
- The immune system exploits the genome to create novel genes. Each clone of lymphocytes in the immune system constructs its unique antigen receptor by recombining otherwise unexpressed minigene elements inherited in the germ line. The immune system thus manufactures millions of different genes that are not encoded as such in the genome. The immune system functions to heal the organism and protect it from foreign invaders, and is also a key factor in causing autoimmune diseases. Yet the ability of the immune system to recognize antigens, a major determinant of health or disease, is a property removed from the germ line.

In their summation, these and other facts well known to biologists lead to the conclusion that the meaning of the information encoded in the genome is variable and conditional; the meaning of a DNA sequence cannot be derived from the sequence itself. Thus the genome does not encode a coherent plan for a sequence of events. (***See*** Alternative Splicing: Evolution; Alternative Processing: Neuronal Nitric Oxide Synthase; Epigenetic Factors and Chromosome Organization; Vertebrate Immune System: Evolution.)

One may argue that the genome, despite its lack of intrinsic meaning, is still a set of instructions, albeit with many possible branching points. Even so, the extragenomic environment and the history of the organism determine the path through which the genome is executed. Since the given state of an actual person is not determined by the person's genome, the genome is not a representation of the person. For this reason, the master-program metaphor does not clarify the role of the genome, but rather obscures it. The mere encoding of amino acid sequences within DNA nucleotide sequences is not programming. On the contrary, the organism uses, manipulates and, in the case of the immune system, creates genes. The genome acts as the organism's servant, not as its master. Why then have knowledgeable people likened the genome to a master program?

Meaning: Line or Loop?

The concept of the genome-as-program is associated with the idea that the connection between a gene and its meaning is linear: DNA → messenger RNA → protein → meaning.

A specific DNA sequence was seen as the plan for making, through the agency of messenger RNA molecules, a particular protein. The protein (for example, an enzyme that builds or degrades molecules, or a transcription factor that activates genes) is the agent that carries out a defined activity. Since the DNA encodes the protein, the meaning of the information borne by the DNA is transformed ultimately into the precise action performed by the protein as an enzyme, transcription factor or other agent. A one-to-one relationship was envisioned: one gene for each protein, and one protein for each function. Thus the activities of the proteins – the meaning – were held to be inherent in the gene – the information.

But, in reality, the living system is not a linear progression from DNA information to protein function; the system is a recursive loop. Proteins, as we have discussed above, are required to make sense out of the DNA sequence; the proteins are required to activate and even to manufacture the very genes that encode the proteins. This way of drawing the connection is closer to reality (**Figure 1**).

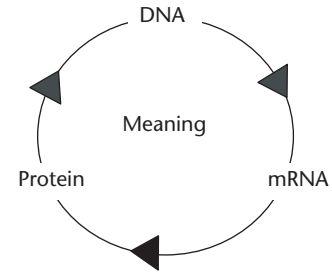

Figure 1

A circle has no beginning and no end: the information actually expressed by DNA is formatted by proteins recursively generated in the process. There is no linear transformation of information (DNA) into meaning (protein action). Genetic information itself is one of the products of protein action; meaning (proteins), as it were, can be said to generate information (legible DNA) in a loop. We might even add the environment to the loop; the influence of proteins on genes is modulated by intracellular and extracellular factors. There is no fixed hierarchy, no one-to-one relationship. The living system is not transformational. The living system is an ongoing process. The meaning of the process that connects DNA and protein is not an outcome of the process; the meaning of the process is the process itself.

Self-organization and Program

Scientists had hoped that the genome might function as a simple program because people, especially scientists, think programmatically. Planning is a characteristic of the human mind. We have intentions and goals; we scheme and we plot. We implement programs, so we take programs for granted; every building has to have an architect; a watch implies a watchmaker.

But we also know of many complex natural phenomena that organize themselves without recourse to a master plan; the world is filled with them. A colony of ants or a hive of bees seem wonderfully organized, yet no single ant or bee, not even the queen, has an idea in mind of what a colony or hive should look like. (Queens are just egg-laying machines.) Each ant and each bee only responds mindlessly to what it senses. What seems to us to be a master plan actualized by each insect colony or hive emerges from the combined actions of the insects themselves, each insect autonomous and entirely ignorant of a world beyond its own sensations.

Similarly, an organism is built and operates with the help of its genome; but the genome is only one element in a recursive process. The iterating cycle of genes that form proteins that form genes is the self-organizing process from which the organism emerges. If there be a genetic program, then such a program writes itself collectively. The action, as it were, precedes the plan. But how can that be?

Complexity, Reduction and Emergence

Physics is the paradigm of sciences; the others try to emulate physics. Physicists explain the behavior of matter by reducing material phenomena to the basic laws of matter and energy. Underlying the physical world are fundamental laws that account for what we see; the material world is explainable by these laws, and so is reducible to these laws. Reduction is done by analyzing the data of sense and experiment to uncover the underlying elements (laws or component parts) that give rise to or 'cause' the data.

Biologists, noting the success of reduction in physics, have attempted to reduce the phenomena of living organisms to the DNA code. Unfortunately, it does not work; life is far too complex to be explained entirely by genomes. (***See*** Systems Biology: Genomics Aspects.)

We do not mean to say that reduction should not be done in biology. On the contrary, scientific reduction has been the key to the identification and characterization of the elements – the cells and molecules – that constitute living organisms. The power of modern biology must be credited to reductive analysis. Our point is that reduction to component parts is only the beginning of wisdom. The essence of biology, like that of other complex systems, is the emergence of high-level complexity created by the interactions of component parts. (***See*** Systems Biology: Genomics Aspects.)

Emergence is not a mystical concept. A physical basis for the emergence of self-organization has been established in studies of nonequilibrium thermodynamics: open systems that exchange matter and energy with their surroundings can maintain themselves in steady states far from equilibrium. The decrease in internal entropy in such systems can be offset by increased entropy in the surroundings; this makes it possible for macroscopic organization to emerge from the coupling of multiple microscopic reactions. Certain coupled chemical reactions exemplify such processes experimentally. Computer simulations of networks of automata have also provided examples of the emergence of high-level nonprogrammed functions created by the interactions of component parts. But these simple examples only illustrate the bare principle; present models of emergence will need upgrading to deal with the complexity of actual biological systems.

Emergence in biology is difficult to study because we have not yet devised a mathematical language suitable for modeling and simulating the generation of high-level complexity out of simple parts. Fortunately, the Human Genome Project, with its need for advanced bioinformatic technology, has invigorated collaborations between biologists, mathematicians, physicists and computer scientists. New ways to model and study the emergence of complexity are already emerging from these activities. But until biology and the informatic sciences develop a common language, we shall have to make do with examples; fortunately examples of emergence abound. Think of your mind. The mind emerges not from neurons but from the interactions of neurons; all the neurons may be intact and alive, but there will be no mind unless the neurons interact. The mind is not reducible to neurons; the mind emerges from the ongoing interactions of neurons. The interactions create a new entity: the mind. Emergent entities, like your mind, are not mere abstractions; they work. (***See*** Information Theories in Molecular Biology and Genomics.)

Genetic Program Metaphor Revisited

The discovery that DNA sequences encode proteins led to the hope that the inherited DNA, the genome, would embody the programs needed to generate and operate the organism; the genome was seen as a representation of the organism. But now biologists have learned that the genome, standing alone, is not an independent program. The organism develops as an emergent process, and not by way of a preexisting plan. The genome is not a representation. What looks like a program is a process. Representation emerges from process.

So we have two alternatives: we can drop the term program and come up with another word for the emergent process we call the program. Or we can continue to use the term program, but understand it to be a metaphor. In either case, the genome by itself is not the master program either in fact or in metaphor. (***See*** Biological Complexity: Beyond the Genome.)

Evolution

Evolution, we should note, does write genomes, but does not improve genomes, even metaphorically. Genomes, at the level of the species, develop from

the processes that adapt an organism to its world. Now a bacterium is no less adapted to its environment than is a human being to its environment. A bacterium as a form of life might, in fact, enjoy a more robust future than the fragile and pugnacious human species. The life and survival of a bacterium would not be improved by making the bacterium more like a human. Self-consciousness would not help a bacterium. Improvement is relative to one's point of view; people like to see themselves as superior to bacteria.

So what does evolution accomplish, if improvement is spurious? Evolution leads to accumulating complexity; humans are objectively more complex than are bacteria. Evolution is a process that, rather than generating improvement, generates new information. But that issue is beyond the scope of this discussion.

Genetic Causality: The Case of Sickle Cell Disease

Detailed knowledge of the DNA sequence, the outcome of the genome project, will not suffice to explain health or disease. We shall have to look to the activation of genes and the dynamic functions of proteins and other molecules involved in the processes of life. Biological and cultural evolution, and the environment too, have their place in the action. Take, for example, the case of sickle cell disease. (***See*** Sickle Cell Disease as a Multifactorial Condition.)

Sickle cell disease is a deficiency of red blood cells (anemia) characterized by an abnormality in the hemoglobin molecule, such that the affected red blood cells assume an elongated shape (like a sickle) at low oxygen tension. The sickle-shaped red blood cells stick together and are destroyed, producing the anemia. Small blood vessels get clogged by the clumps of sickled red cells and tissues suffer from the lack of blood flow. The disease, untreated, results in an early death. What is the cause of sickle cell disease?

The answer is deceptively simple. Sickle cell disease is a genetic disease; the hemoglobin molecule is abnormal because of a mutation in the gene encoding the β chain of the molecule: a single glutamic acid in the protein is replaced by the amino acid valine – this abnormal hemoglobin is called hemoglobin S. Persons who have inherited the gene for hemoglobin S from both parents can make only hemoglobin S and so manifest the disease. Persons who have inherited one hemoglobin S gene and one normal gene (heterozygotes) are essentially free of the disease. Thus we could define the disease as caused by having inherited two copies of the hemoglobin S gene. But this is not the whole story. (***See*** Genetic Variation: Polymorphisms and Mutations.)

The hemoglobin S gene is present mostly in populations of people who have originated in equatorial Africa, and the incidence of the gene is much higher than would be expected from the spontaneous mutation rate of standard hemoglobin to hemoglobin S. The relatively high frequency of a potentially lethal mutation suggests that the mutated gene must have some selective advantage, must contribute to fitness. Well, it turns out that children infected with a certain type of malaria (and who are untreated) will die of the infection if their genome contains only the standard hemoglobin gene. The children who carry one gene for hemoglobin S (heterozyotes), however, are relatively resistant to malaria, and so do not die of that disease. The heterozygous children also do not die of sickle cell disease. Of course, children whose genomes include two of the hemoglobin S genes (homozygotes) die of sickle cell disease. Thus, we might say that hemoglobin S is an advantageous adaptation to malaria, and the gene is maintained in the population, despite the loss of homozyzgous children, at a rate that reflects the selective pressure exerted by the rate and severity of malaria infection. The death of homozygous individuals is the price paid by the population for heterozygous resistance to malaria. So we could say that the high frequency of hemoglobin S (and sickle cell disease) is caused by malaria. By this reasoning, we might say that sickle cell disease has value as a trade-off in exchange for malaria. In environments free of malaria, hemoglobin S provides no advantage.

Should we now conclude that one of the causes of sickle cell disease as a disease (rather than as a trade-off) is the absence of malaria? Sickle cell disease is a serious health problem in the African Americans whose ancestors were taken to America as slaves from West Africa. Should we include the slave trade among the causes of sickle cell disease in an African-American child? Or has the disease been caused by two heterozygotes falling in love?

Persons who have inherited two hemoglobin S genes do much better clinically if they continue to produce fetal hemoglobin after birth; the fetal hemoglobin makes the affected red cell more resistant to the deleterious effects of hemoglobin S. Indeed, homozygous persons are now treated with a drug that induces the production of fetal hemoglobin after birth. Are we to conclude that the normal termination of fetal hemoglobin production is a causal factor in sickle cell disease?

In short, a 'simple' genetic disease like sickle cell disease, which is associated with a defined mutation resulting in a defined molecular abnormality, presents us with a complex causal chain of events. How much more complex are the possible genetic explanations for diseases such as type 1 diabetes or multiple sclerosis, diseases that have been associated with many different

susceptibility genes. Indeed, inheriting susceptibility genes does not make the disease inevitable. Take for example identical twins, who bear identical genomic DNA; if one twin develops type 1 diabetes or multiple sclerosis, the other twin will develop the disease in only about a third of the pairs. Having susceptible DNA does not suffice to explain the disease. Indeed, prevalent genes that are associated with susceptibility to complex diseases are probably advantageous trade-offs. (***See*** Complex Genetic Systems and Diseases.)

The Environment

We can summarize the limitations of the genome most easily by repeating what has already been said many, many times: one's genes are only an incomplete explanation for one's being; the present environment and its history, at the scales of the person, the group and the biosphere, interact with the genome to determine its expressions and effects.

Genome Metaphors

What metaphor might be generally useful for appreciating the function of the genome? Genomic DNA is a reservoir of raw information that, suitably processed, can be translated into the amino acid sequence of functional proteins. Perhaps we could think of the genome as akin to a list of words, a vocabulary book, that can be used to build sentences. One may argue about the meaning or range of potential meanings inherent in any word standing alone; but it is clear that a word gains its fullest, most particular meaning when the word is used as part of a sentence. DNA sequences, like words in natural language, are essentially passive; they may be spliced into different genes that give rise to proteins that function in different ways under different circumstances. Fragments of genomic DNA, like words, acquire different meanings in different contexts. They can be used artfully to tell different stories. The genome, like a vocabulary, is information, transmissible from generation to generation, that is available for processing into meaning. The process itself, as we have discussed, is the story.

We might extend the language metaphor and consider that particular three-member sets of bases – the codons – form the alphabet sequences of the metaphorical DNA words; 'stop', start, 'splice' and other codon signals may be said to encode syntax.

If you would prefer a computer metaphor, consider this one: the genome may be viewed as a compacted substrate for storing information; protein synthesis involves the processing and amplification of genomic DNA into a multitude of RNA molecules and then into an amino acid sequence expressed in the millions of proteins produced using that DNA. From this point of view, genomic DNA is like a database stored in computer memory. Genomic data is special because it is transmitted from generation to generation; such transmission can take place because the genome serves as a template for its own replication between generations. Genomic data is used by the individual organism for development and for maintenance. The organism, of course, also uses data obtained from its evolving self and from the environment. The master program for using this data, as we discussed above, emerges from the living process itself.

Conclusions

The Human Genome Project, like putting a man on the moon, is a costly undertaking of great technical virtuosity. It is good that the project has been done for the daring of it and because it has already provided much important information about genetics and the organism. No less important, the genome project has spawned powerful technologies and has opened biology to the age of informatics. Biology has learned that it is an informatic science. Finally, the very success of the genome project has dispelled the simplistic illusion of the genetic program; biology is now aware of its true complexity. The genome project, wittingly or not, has built the foundation for deeper probes into the complexity of life. The limitations of the project are only the limitations of the genome itself.

See also

Biological Complexity: Beyond the Genome
Complex Genetic Systems and Diseases
Genetics, Reductionism, and Autopoiesis
Systems Biology: Genomics Aspects

Further Reading

Atlan H (1983) Information theory: basic elements and recent developments. In: Trappl R (ed.) *Cybernetics: Theory and Applications*, pp. 9–41. New York, NY: Hemisphere Publications/Springer.

Atlan H (1987) Self-creation of meaning. *Physica Scripta* **36**: 563–576.

Atlan H and Cohen IR (1998) Immune information, self-organization and meaning. *International Immunology* **10**: 711–717.

Atlan H and Koppel M (1990) The cellular computer DNA: program or data? *Bulletin of Mathematical Biology* **52**: 3335–3348.

Cohen IR (2000) *Tending Adam's Garden: Evolving the Cognitive Immune Self*. San Diego, CA: Academic Press.

Cohen IR (2000) Discrimination and dialogue in the immune system. *Seminars in Immunology* **12**: 215–219, 269–271, 321–323.

Chong L and Ray LB (2002) Whole-istic biology. *Science* **295**: 1661.

Fox Keller E (1995) *Refiguring Life. Metaphors of Twentieth-century Biology*. New York, NY: Columbia University Press.

Kam N, Cohen IR and Harel D (2003) The immune system as a reactive system: modeling T cell activation. *Bulletin of Mathematical Biology* (in press).

Kono T (1997) Nuclear transfer and reprogramming. *Reviews of Reproduction* **2**: 74–80.

Louzoun Y, Solomon S, Atlan H and Cohen IR (2001) Modeling complexity in biology. *Physica A* **297**: 242–252.

Nicolis G and Prigogine I (1977) *Self-organization in Nonequilibrium Systems: From Dissipative Structures to Order through Fluctuations*. New York, NY: Wiley.

Science (1999) Complex systems [special issue]. *Science* **284**: 79–109.

Strohman RC (1997) Epigenesis and complexity: the coming Kuhnian revolution in biology. *Nature Biotechnology* **15**: 194–200.