COMMENTARY

GENOMICS

Five stages of the Human Genome Project

Richard C. Strohman

Here I describe the Human Genome Project (HGP) in terms of five overlapping stages defined mostly in terms of their biomedical goals. My somewhat arbitrary staging is informed by what I perceive as a steady turnover of ideas having to do with a "program for life," what it means, and where it is located. Initially the program was assumed to be located in the genome and to be isomorphic with phenotype, but this location is gradually being transferred, through intermediate stages, to the level of the organism itself. Each succeeding stage of the HGP is driven by results not anticipated in the prior stage.

While the HGP is not yet complete, enough data have been collected from it and from other genome projects (mouse, worm, fly) to allow some tentative conclusions: (1) There is not sufficient information in genomic databases to provide explanations for complex functional attributes of cells and organisms. (2) Therefore, there must be other informational systems and operating rules that complement genomic systems. (3) Epigenesis is identified as one such system. (4) Program rules by which regulation is produced are extragenomic and are most likely to be found not in molecular mechanisms per se, but in their integration into complex gene circuits and, more peripherally, into their connectedness with regulatory networks (metabolic and other) of cellular dimensions.

Evidence establishing epigenetic networks as sites of control over cellular phenotype, including control over covalent marking of DNA and chromatin (the phenotype of the genotype)[1], has shifted attention from a narrow focus on DNA to the more complex dynamics of gene circuits, and their integration into larger, environmentally open networks of cellular dimensions[2]. This change in emphasis from linear "causal" molecules to the regulatory dynamics of molecular networks is increasingly perceived by a growing number of molecular biologists and geneticists working within the various genome projects[3], as well as in biochemistry[4], integrative biology and physiology[5], developmental biology[6], and medical[7] and behavioral genetics[8].

The following paragraphs will summarize the five stages of the HGP. Stages I (monogenic causality) and II (polygenic causality) deal, respectively, with rare monogenic diseases, and with polygenic diseases associated with an unknown complex of genes coupled to individual experience and environmental history. Because monogenic diseases account for only a small fraction of noninfectious diseases (2%)[8], the emphasis has shifted more to stage II, in which the goal is to focus on candidate genes or small clusters of key genes among the many (possibly tens, hundreds, or thousands) involved in producing a chronic disease or any other complex phenotype. Gene maps for each physiologic function and disease will begin to close the gap between genetic information and functional outcome in cells and organisms. However, as noted previously, these maps mostly assume additive and dominant effects and do not include dynamic rules governing deployment, interaction (epistasis), redundancy (pleiotropy), and connectedness of these genes.

The transition to stage III, analysis of the proteome (the entire protein complement of a genome), focuses on expressed genes (proteins) thereby avoiding some of the problems associated with genomic complexity. However, as in stage II, problems still exist in that there is a continued reliance on describing large numbers of additive agents (proteins) without recourse to rules of interaction, redundancy, and connectedness.

Transgenic analysis (stage IV), in acknowledging the many problems just outlined for the earlier stages, will rely on the normal dynamics of the organism in conjunction with gene transfer between species to produce "novel" phenotypes and thereby reveal programmatic aspects of morphogenetic processes. But here, too, there are unexplored areas, as with pleiotropic genes and proteins, in which developmental and other higher levels of organization remain to be taken into account[4]. Therefore, with this strategy, although detailed genetic maps for a variety of cellular functions will be established, the nature of the processes being perturbed by gene manipulation remains a black box.

The fifth and final stage, complexity, is the logical though unpredictable extension of stage IV, and perhaps represents a new approach recognizing that higher levels of cellular organization and regulation impose constraints on the genome, and that genes and environments are inseparably integrated. Epigenetic regulation of the genome[1] is seen as the most proximal of a hierarchy of constraints extending outward from DNA structure to the cell boundary and beyond. This stage of the project is now engaged with describing the molecular events involved in DNA and chromatin marking and seeks to understand, among other things, how marking constrains and orders patterns of gene expression. While these studies remain descriptive, the next logical stage is already under way. Connectedness is restated in terms of gene circuits[2] or metabolic networks[4], so that the large amount of information inherent in systems of genetic or biochemical activity may be collapsed into a logic of circuits and networks.

The main message here for a biotechnology devoted to finding specific causes and cures for complex diseases is that the desired specificity is severely compromised by a profound genetic, molecular, and informational complexity. For example, coronary artery disease involves several hundred genes. A complex disease like colon cancer is now acknowledged to include not only large-scale mutation but also profound changes in patterns of gene expression[9]. Genetic instability in the forms of loss of heterozygosity[10] and aneuploidy[11] also complicate the simple single or even multiple gene mutation theories of cancer[12]. Considering in addition the classical but mostly unrecognized uncertainties inherent in widespread epistasis and pleiotropy, the present emphasis on dominant gene effects and on single-gene or protein-based diagnosis and therapy for common human diseases must be seen as unrealistic. The gene–protein circuits and network logic studies cited earlier represent some starting points for the development of new understanding and new technologies for complex human phenotypes.

Richard C. Strohman is a professor emeritus in the Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720-3206 (strohman@uclink4.berkeley.edu).

1. *Cell. Mol. Life Sci.* 1998. **54** (entire volume devoted to epigenetic control of transcription).
2. Kauffman, S.A. 1969. *J.Theor. Biol.* **22:**437–467, and Kauffman, S.A. 1993. *The origins of order,* Oxford Univ. Press, New York; Thomas, R. 1998. *Int. J. Dev. Biol.* **42:**479–485.
3. Miklos, G.L.G. and Rubin, G. 1996. *Cell* **86**:521–529.
4. Fell, D. 1997. Understanding the control of metabolism. Portland Press, London; Veech, R.L. and Fell, D.A. *Cell Biochem. Funct.* **229:**236
5. Savageau, M.A. 1991. *New Biologist* **3:**190–197.
6. Gilbert, S., Opitz, J.M. and Raff, R.A. 1996. *Dev. Biol*. **173:**357–372; Goodwin, B.C., Kauffman, S.A. and Murray, J.D. 1993. *J. Theor. Biol*. **163:**135–144; Webster, G. and Goodwin, B. 1996. *Form and transformation: Generative and relational principles in biology*. Cambridge University Press, Cambridge.
7. Sing, C.F, Haviland, M.B and Reilly, S.L 1996. *Ciba Found. Symp.* **197:**211–232.
8. Wahlsten, D. 1999. Single-gene influences on brain and behavior. *Ann. Rev. Psychol.* **50:**599–624.
9. Strohman, R.C. 1994. *Bio/Technology* **12:**156–164.
10. Zhang, L. et al. 1997. *Science* **276:**1268–1272.
11. Vogelstein, B. et al. 1989. *Science* **244:**207–211.
12. Duesberg, P., Rausch, C., Rasnick, D., and Hehlmann, R. 1998. *Proc. Natl. Acad. Sci. USA* **95:**13692–13697.