Chapter 4

Genomics, Proteomics, and Beyond

SAHOTRA SARKAR

1. Introduction

The term "molecular biology" was introduced by Warren Weaver in 1938 in an internal report of the Rockefeller Foundation: "And gradually there is coming into being a new branch of science – molecular biology – . . . in which delicate modern techniques are being used to investigate ever more minute details of certain life processes."[1] Weaver probably only dimly foresaw that these new techniques would ultimately transform the practice of biology in a way comparable only to the emergence of the theory of evolution in the previous century. By the beginning of the twenty-first century molecular biology has become most of biology, either *constitutively*, insofar as biological structures are characterized at the molecular level as a prelude for further study, or at least *methodologically*, as molecular techniques have become a preferred mode of experimental investigation of a domain. Recent biological work at the organismic and lower levels of organization – cytology, development, neurobiology, physiology, etc. – increasingly fall under the former rubric. Work in demography, epidemiology, and ecology falls under the latter, with ecology perhaps being the sub-discipline within biology which has most resisted molecularization. Work in evolution falls under both: constitutively, when the evolution of molecules and molecular structures forming organisms is studied for its own sake, and methodologically, when molecular techniques (most notably, DNA sequencing) are used to reconstruct evolutionary history.

This chapter traces the conceptual shifts that have marked the development of molecular biology during the past half-century with an emphasis on epistemological issues raised by the more recent changes. Section 2 provides the background of classical molecular biology. Section 3 moves on to the genomic and post-genomic era. Section 4 analyzes the prospects for proteomics. Section 5 turns to the nascent project of systems biology. Finally Section 6 turns to the philosophical implications of these developments, namely, the status of reductionism, of the informational interpretation of molecular biology, and the prospect that systems biology will finally reintroduce dynamical considerations in molecular biology. Section 7 invites readers to pursue

1 As quoted by Olby (1974, p.442).

more philosophical exploration of the issues raised by molecular biology which have, until recently, often been ignored by philosophers.

2. Classical Molecular Biology

During the decade following Weaver's introduction of "molecular biology" experimental work showed that the hereditary substance – specifying "genes" [See Gene Concepts] – was deoxyribonucleic acid (DNA). Attention then focused on deciphering the physical structure of DNA, a problem that was solved by Watson and Crick (1953) with their double helix model from 1953. The construction of this model and its subsequent confirmation was a development of signal importance for modern biology.[2] It ushered in the "classical" age of molecular biology with an intriguing informational interpretation of biology [See Biological Information]. Important conceptual innovation also came from Monod and Jacob in the early 1960s, who constructed the "allostery" model to explain cooperative behavior in proteins and the "operon" model of gene regulation.[3] Genes were interpreted as DNA sequences either specifying proteins (the *structural* genes) or controlling the action of other genes (the *regulatory* genes). Perhaps the most important development in classical molecular biology was the establishment of a genetic "code" delineating the relation of DNA sequences to amino acid residue sequences in proteins.[4] Gene *expression* took place by the *transcription* of DNA to ribonucleic acid (RNA) at the chromosomes (in the nucleus), and the *translation* of these transcripts into protein at the ribosomes (in the cytoplasm). The one gene–one enzyme credo of classical genetics was transformed into the one DNA segment–one protein chain credo of molecular biology.

Crucial to the program of molecularizing biology was the expectation – first explicitly stated by Waddington (1962) – that gene regulation explained tissue differentiation and, ultimately, morphogenesis in complex organisms. Genetic reductionism, the thesis that genes alone can explain organismic features, long predates molecular biology (Sarkar, 1998). However, the molecular interpretation of the gene allowed the general explanatory success of molecular biology to be co-opted as a success of molecular genetics. In such a context, Waddington's thesis was positively received and helped usher in an era dominated by *developmental genetics* according to which organismic development was to be understood through the action of genes. Mayr (1961) and Jacob and Monod (1961) independently introduced the metaphor of the genetic program to characterize the putative relation between genomic DNA and organismic development. As molecular genetics began to dominate the research agenda of molecular biology in the 1970s, the emergence of organismic features came to be viewed as determined by "master control genes" (Gehring, 1998). This view was initially supported by the demonstration that some DNA sequences (such as the "homeobox") were conserved across a wide variety of species. DNA came to be viewed as the molecule "defining" life, a view that

2 Sarkar (2005, ch. 1) argues this point in detail.

3 See below, and Monod (1971) and Jacob (1973).

4 Both DNA and protein are linear molecules in the sense that they consist of units connected in a chain through strong (covalent) chemical bonds.

helped initiate the massive genome sequencing projects of the 1990s, which were supposed to produce a gene-based complete biology that delivered on all the promises of molecular developmental genetics. In general, because of the presumed primacy of DNA in influencing organismic features, starting in the early 1960s, molecular genetics began to dominate research in molecular biology.

Thus, genetics and development were the earliest biological sub-disciplines to be reconstituted by molecular biology. In the case of evolutionary biology, as early as the 1950s, Crick (1958) pointed out that the genotype–phenotype relation could be reinterpreted as the relation between DNA and protein, with proteins constituting the subtlest form of the expression of a phenotype of an organism. Consequently, the evolution of proteins (and, later, DNA sequences), especially the question of what maintained their diversity within a population, became a topic of investigation. In the 1960s, these studies led to the neutralist challenge to the received view of evolution [See Molecular Evolution]. More importantly, changes at the level of DNA sequences, provided that these were selectively neutral, permitted the construction of a "molecular clock" that can arguably be used to reconstruct evolutionary history more accurately than what can be achieved by traditional morphological methods (even though such reconstructions have on occasion proved to be controversial).

Meanwhile, biochemistry and immunology also fell under the spell of the new molecular biology. That enzyme interactions and specificity would be explained in molecular terms was no surprise. However, immunological specificity was also believed to be explainable by the same mechanism. This model of immune action was coupled to a selectionist theory of cell proliferation to generate the clonal theory of antibody formation, which combined molecular and cellular mechanisms in a novel fashion [See Self and Nonself]. In both biochemistry and immunology, what was largely at stake was the development of models that could explain the observed specificity of interactions: enzymes reacted only with very few substrates; antibodies were highly specific to their antigens.

Classical molecular biology can be viewed in continuity with both the genetics and the biochemistry of the era that preceded it. From biochemistry – in particular, the study of enzymes in the 1920s and 1930s – it inherited the proposed mechanism that the function or behavior of biological molecules is "determined" by its structure.[5] In the 1950s, structural modeling of biological macromolecules, especially proteins, was pioneered by Pauling and his collaborators using data from *x*-ray crystallography (e.g., Pauling & Corey, 1950). By the early 1960s a handful of such structures were fully solved. These structures, along with the structure of DNA, seemed to confirm the hypothesis that structure explains behavior. Perhaps more surprisingly, it was found that structural interactions seemed to be mediated entirely by the shape of active sites on molecules and that the sensitive details of structure and shape were maintained by very weak interactions.

These experimental observations led to four seemingly innocuous rules about the behavior of biological macromolecules which, in the 1960s and 1970s, formed the theoretical core of molecular biology:[6]

---

5 This idea is of earlier vintage, going back to Ehrlich's "side-chain" theory in the late nineteenth century.

6 For details, see Sarkar (1998, pp.149–50).

(i) the *weak interactions* rule – the interactions that are critical in molecular interactions are very weak;
(ii) the *structure-function* rule – the behavior of biological macromolecules can be explained from their structure as determined by techniques such as crystallography;
(iii) the *molecular shape* rule – these structures, in turn, can be characterized entirely by molecular size and, especially, external shape, and some general properties (such as the hydrophobicity) of the different regions of the surfaces;
(iv) the *lock-and-key fit* rule – in molecular interactions, molecules interact only when there is a lock-and-key fit between the two molecular surfaces. There is no interaction when these fits are destroyed.

Such a lock-and-key fit, based on shape, achieves what is called "stereospecificity," thus resolving the critical problem for classical molecular biology, which was to explain how structure specified behavior. Of the four rules introduced above, the molecular shape and lock-and-key fit rules are the most important because they are the ones that are most intimately involved in the explanation of specificity. In what follows, these four rules will be called the rules of classical molecular biology.

In the 1960s and 1970s these rules were deployed with remarkable success. As noted earlier, enzymatic and immunological interactions were among those that were immediately brought under the molecular aegis. Two other cases are even more philosophically interesting: (i) the allostery model explains why some molecules such as hemoglobin show *cooperative* behavior. In the case of hemoglobin, there is a nonlinear increase in the binding of oxygen after binding is first initiated. This is explained by conformational – shape – changes in the molecular subunits of hemoglobin as the first oxygen molecules begin to bind to them; and (ii) the operon model of gene expression explains *feedback*-mediated gene regulation in prokaryotes. This model explains how the presence of a substrate activates the production of a protein that interacts with it, and its absence inhibits that production.[7] Section 6 will emphasize the philosophical significance of the success of such structural explanation in molecular biology.

However, the 1950s also saw the elaboration of a radically different model of biological specificity, one based on the concept of information, which was only introduced in genetics in 1953 (Sarkar, 1996). This concept soon came to play a foundational role in molecular genetics. DNA was supposed to be the repository of biological information, a genetic "program" was supposed to convert this information into the adult organism, and new information was supposed to result from random mutation (when such mutations were maintained by selection). Information was never incorporated from the environment into the genome. Crick (1958, p.153) enshrined these assumptions in what he called the "Central Dogma" of molecular biology: "This states that once 'information' has passed into protein *it cannot get out again*. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible." Information, in Crick's model, was defined by the sequence of nucleotide bases

7 See Monod (1971) for an accessible accurate account of these two examples and a conceptual summary of theoretical reasoning in early molecular biology.

in DNA or the sequence of amino acid residue in protein molecules. Note the contrast here with the stereospecific physical model of specificity: specificity comes from the combinatorial order or arrangement of subunits in DNA and protein, and not from the physical shape. The Central Dogma has continued to be an important regulative principle of molecular biology in the sense that it is presumed for further theoretical reasoning. Whether it survives recent developments will be discussed later in this chapter.

By the late 1970s it became clear that the simplicity of the picture of genetics inherited from the 1960s was being lost. The initial picture was generated from an exploration of the genomes of prokaryotes (single-celled organisms without a nucleus), especially the bacterium, *Escherichia coli*. In prokaryotes, every piece of DNA has a structural or regulatory function. In the 1970s, it was discovered that the genetics of eukaryotes (organisms with cells with nuclei) turns out to have an unexpected complexity. In particular, large parts of the genomic DNA sequences apparently had no function: these segments of "junk" DNA were interspersed between genes on chromosomes and also within genes. After RNA transcription, non-coding segments within genes were *spliced* out before translation. Gene regulation in eukaryotes was qualitatively different and more complicated than in prokaryotes. Some organisms used nonstandard genetic codes, etc.[8]

Subsequent work in molecular biology has only enhanced the complexity of this picture, so much so that it is reasonable to suggest that the classical picture is breaking down. RNA transcripts are subject to *alternative splicing*, with the same DNA gene corresponding to several proteins. RNA is edited, with bases added and removed, before translation at the ribosome, sometimes to such an extent that it is difficult to maintain that some gene codes for a given protein. There is also no obvious relation between the amount of DNA in an organism and its morphological or behavioral complexity, an observation that is sometimes called the *C*-value paradox (Cavalier-Smith, 1978). Most importantly, it now appears that a fair amount of the so-called junk DNA is transcribed into RNA, though not translated. Thus, presumably, much of the so-called junk DNA is functional, though the nature of these functions remains controversial (see Section 3).

The complexities of eukaryotic genetics, as discovered in the 1970s and 1980s, already begin to challenge the Central Dogma.[9] Much of this work was made possible by the development of technologies based on the polymerase chain reaction (PCR) in the 1980s. There were five salient discoveries that challenged the simple picture inherited from prokaryotic genetics:[10]

(i) the genetic code is not fully universal, the most extensive variation being found in mitochondrial DNA in eukaryotes. However, there is also some variation across taxa (Fox, 1987);
(ii) DNA sequences are not always read sequentially in blocks. There are overlapping genes, genes within genes, and so on (Barrell et al., 1976). Thus, two or more

8 See Sarkar (1996) for a detailed account.
9 Thiéffry and Sarkar (1998) give a history of several earlier challenges. Even in the 1960s there was no unanimity about the status of the Central Dogma.
10 For details, see Sarkar (2005), chapter 8.

different proteins could be specified by the same "gene." Once again the Central Dogma is under challenge since the genome alone does not seem to contain all the information necessary to determine which protein is encoded by the "gene" in question;

(iii) as noted earlier, not all DNA in the genome is functional. Intervening sequences – within and between structural genes – must be spliced out from transcripts (Berget et al., 1977; Chow et al., 1977). This discovery helped resolve the *C*-value paradox mentioned earlier, that is, the absence of any obvious correlation between the size of the genome and morphological and behavioral complexity of an organism;

(iv) the same transcript may be spliced in different ways (Berk & Sharp, 1978). One consequence of such *alternative splicing* is that, as with overlapping genes, two or more different proteins could be specified by the same "gene";

(v) besides splicing, RNA is sometimes subject to extensive editing before translation at the genome (reviewed by Cattaneo, 1991).

Both points (iv) and (v) challenge the Central Dogma for the same reason as point (ii), These developments have led to increasing skepticism of the relevance of the coding model of the DNA–protein relationship and, especially, of the informational model of specificity (see Sarkar, 2005, and Section 6). It is no longer even clear that there is a coherent concept of information in molecular biology (see, however, BIOLOGICAL INFORMATION). Though philosophers – and some biologists – have been slow to recognize this, the one DNA segment–one protein chain credo has long become irrelevant in molecular biology. These developments in eukaryotic genetics paved the way to a reconceptualization of heredity in the emerging field of genomics.

3. Genomics and Post-Genomics

Genomics was ushered in by the decision to sequence the entire human genome as an organized project (the Human Genome Project [HGP]), involving a large number of laboratories in the late 1980s. Subsequently, similar projects were established to sequence the genome of many other species. To date, genomes of over 150 species have been sequenced. Almost every month sees the announcement of the completion of sequencing for a new species. The sheer volume of sequence information that has been produced has spawned a new discipline of "bioinformatics" dedicated to computerized analyses of biological data.

When the HGP was first proposed, there was considerable controversy among biologists about its wisdom (Tauber & Sarkar, 1992; Cook-Deegan, 1994). There were: (i) doubts about its ability to deliver on the bloated promises made by proponents of its scientific and, especially, its medical benefits; (ii) questions whether such organized "Big Biology" projects were wise science policy because of their potential effect on the ethos of biological research; and (iii) worries that society would be legally and medically ill-prepared to cope with the results of rapid sequencing, rather than the normal slower accumulation of human genomic sequence information. It was feared that legislation

protecting genetic privacy and preventing genetic discrimination would not be in place; there would be a shortage of genetic counselors; and so on.

In one important aspect, the critics were correct: there have been few immediate medical benefits from the HGP and no significant such innovation seems forthcoming. Instead, recent work underscores the importance of gene–environment interactions that critics had routinely invoked to criticize the claims of the HGP [See Heredity and Heritability]. However, in another sense, even the most acerbic critics should now accept that the scientific results of the sequencing projects, taken together, have been breathtaking.

Contrary to the expectations of the HGP's proponents, few successful and interesting predictions about organismic development have come from sequence information alone (Stephens, 1998). However, as the following list shows, genomic research is persistently throwing up surprises:

(i) the most important surprise from the HGP was that there are probably only about 30,000 genes in the human genome compared to an estimate of 140,000 as late as 1994 (Hahn & Wray, 2002).[11] In general, plant genomes are expected to contain many more genes than the human genome. Morphological or behavioral complexity is not correlated with the number of genes that an organism has. This has been called the *G*-value paradox (Hahn & Wray, 2002);

(ii) the number of genes is also not correlated with the size of the genome, as measured by the number of base pairs. The fruit-fly, *Drosophila melanogaster*, has 120 million base pairs but only 14,000 genes; the worm, *Caenorhabditis elegans* has 97 million base pairs but 19,000 genes; the mustard weed, *Arabidopsis thaliana* has only 125 million base pairs and 26, 000 genes, while humans have 29,000 million base pairs and 30,000 genes (Hahn & Wray, 2002);

(iii) at least in humans, the distribution of genes on chromosomes is highly uneven. Most of the genes occur in highly clustered sites. Most genes that occur in such clusters are those that are expressed in many tissues – the so-called "housekeeping" genes (Lercher et al., 2002). However, the spatial distribution of cluster sites appears to be random across the chromosomes. (Cluster sites tend to be rich in *C* and *G*, whereas gene-poor regions are rich in *A* and *T*.) In contrast, the genomes of arguably less complex organisms, including *D. melanogaster*, *C. elegans*, and *A. thaliana*, do not have such pronounced clustering;

(iv) only 2 percent of the human genome codes for proteins while 50 percent of the genome is composed of repeated units. Coding regions are interspersed by large areas of non-coding DNA. However, some functional regions, such as *HOX* gene clusters, do not contain such intervening sequences;

(v) scores of genes appear to have been horizontally transferred from bacteria to humans and other vertebrates, though apparently not to other eukaryotes. However, this issue remains highly controversial;

(vi) once attention shifts from the genome to the proteome (the protein complement of a cell – see Section 4), a strikingly different pattern emerges. The human

11 If past trends are at all indicative of the future, all estimates of the number of genes in "higher" animals will decline even further.

proteome is far more complex than the proteomes of the other organisms for which the genomes have so far been sequenced. According to some estimates, about 59 percent of the human genes undergo alternative splicing, and there are at least 69,000 distinct protein sequences in the human proteome. In contrast, the proteome of *C. elegans* has at most 25,000 protein sequences (Hahn & Wray, 2002);

(vii) it now appears that non-coding DNA is routinely transcribed into RNA but not translated in complex organisms (Mattick, 2003). It seems that these RNA transcripts form regulatory networks that are critical to development. Interestingly, the amount of non-coding DNA sequences in organisms appears to grow monotonically with the morphological complexity of organisms;

(viii) at least in *A. thaliana*, there is evidence of genome-wide non-Mendelian inheritance during which specifications from the grandparental, rather than parental, generation are transmitted to descendants (Lolle et al., 2005).

An important task of modern molecular biology is to make sense of these disparate unexpected discoveries. One conclusion seems unavoidable: any concept of the gene reasonably close to that in classical genetics will be irrelevant to the molecular biology of the future [See Gene Concepts].

4. Proteomics

The term "proteome" was introduced only in 1994 to describe the total protein content of a cell produced from its genome (Williams & Hochstrasser, 1997). Unlike the genome, the proteome is not even approximately a fixed feature of a cell (let alone an organism) because it changes over time during development as different genes are expressed. Deciphering the proteome, and following its temporal development during the life cycle of each tissue of an organism, has emerged as the major challenge for molecular biology in the post-genomic era. This project has been encouraged by the discovery of unexpected universality of developmental processes at the level of cells and proteins (Gerhart & Kirschner, 1997). For instance, even though hundreds of genes are known to specify molecules involved in transport across cellular membranes, there are only about twenty transport mechanisms in all living systems. The emergence of proteomics in the wake of the various sequencing projects signals an acceptance of the position that studying processes entirely, or even largely, at the DNA level will not suffice to explain phenomena at the cellular and higher levels of biological organization, including organismic development. Even genomics did not go far enough; a sharper break with the past will be necessary.

Nevertheless, in one very important sense, the emergence of proteomics recaptures the spirit of early molecular biology, when all molecular types, but especially proteins, were foci of interest, and the deification of DNA had not replaced a pluralist vision of the molecular basis for life. In the late 1960s, Brenner and Crick proposed "Project K" which was supposed to be "the complete solution of *E. coli*." *E. coli* (strain K-12) was selected as a model organism because of its simplicity (as a unicellular prokaryote) and ease of laboratory manipulation. Project K included: (i) a "detailed test-tube study of

the structure and chemical action of biological molecules (especially proteins)" (Crick, 1973); (ii) completion of the models of protein synthesis; (iii) work on the structure and function of cell membranes; (iv) the study of control mechanisms at every level of organization; and (v) the study of the behavior of natural populations, including population genetics. Once *E. coli* was solved, biology was supposed to move on to more complex organisms.

Notice that in this project: (i) DNA receives no preferential attention at the expense of other molecular components; and (ii) the centrality of proteins as the most important active molecules in a cell is recognized. Project K accepts that there is much more to the cell than DNA; it accepts that no simple solution of the cell's behavior can be read from the genomic sequence. After a generation of infatuation with DNA and genetic reductionism, the aims of proteomics return in part to the vision of biology incorporated in Project K. However, at least in one important way, that project went beyond proteomics as currently understood: it emphasized all levels of organization whereas the explicit aims of proteomics are limited to the protein level. To understand the biology of organisms, the future will probably require even further expansion – see Section 5.

Meanwhile, work on proteins has also generated unexpected challenges. In particular, the four rules of classical molecular biology have not survived intact and at least the last three will require some modification. It now appears – though the essential idea goes back to the 1960s – that the fit between interacting sites of protein molecules is more dynamic than in the classical model, with the active site often "inducing" an appropriate fit.[12] It also appears that a more complicated model than the original allostery model will be required to account for many cases of cooperativity.

5. Towards a Systems Biology?

Over a half-century ago, Wiener (1948) suggested that living organisms be viewed as systems governed by feedback control. Wiener attempted to found a new discipline – "cybernetics" – for the study of such systems. In spite of Wiener's proselytization on behalf of the new discipline, cybernetics did not amount to much. It generated some excitement in the social sciences in the 1950s and then fizzled out (Heims, 1991). Engineers occasionally referred to cybernetic concepts (especially feedback) but, by the 1980s, that was about all the attention it received. In biology, especially in the emerging field of molecular biology, cybernetics contributed nothing of substance in spite of many attempts to use it (Sarkar, 1996).

Unexpectedly, at the beginning of the twenty-first century, Wiener's vision has returned to the forefront of attention in contemporary molecular biology. The context of Wiener's return is the new "systems biology" approach to the organism. As one of the proponents of the new approach, Kitano (2002), puts it: "Since the days of Norbert Wiener system-level understanding has been a recurrent theme in biological science." Kitano is partly right: ecosystem ecology, also going back to the 1950s, and large-scale studies of the immune system, starting in the 1960s, have both been important parts of biology even though Wiener's direct influence is hard to discern. But, in the new

12 See, for example, Koshland and Hamadani (2002).

molecular biology that came to dominate most of biological research, starting in the 1960s (as discussed in the earlier sections of this chapter), systems thinking was irrelevant. Research was dominated by what will be called "reductionism" in Section 6: trying to explain wholes by constructing them out of smaller and smaller parts (Sarkar, 1998).

Systems biology claims to be the culmination of the move from genetics to genomics to proteomics. Its aim is to study cells and larger units within organisms as composite systems described in terms of both the structures within them and the processes that occur in these structures (Ideker et al., 2001; Weston & Hood, 2004). Almost all advocates of systems biology endorse a collaborative technology-driven enterprise. Biologists, engineers, and computer scientists (among others) are supposed to collaborate to set up the necessary technological infrastructure to track all relevant processes within the cell and record the massive amounts of data that are produced. Integration at all levels – intellectual disciplines, conceptual frameworks, technology creation, and research culture – is expected to be critical to the success of this approach.

The most important innovation of systems biology is its explicit reintroduction of considerations of time into molecular biology – see Section 6 for further reflection on this point. One of the peculiar characteristics of molecular biology has been its avoidance of explicit reference to time: flows of information between nucleic acids, and from them to proteins, control of gene expression through negative feedback and switches – these mechanisms all replace explicit discussion of how the chemical composition of cells change over time. This is one of the salient features that make molecular biology look so different from the biochemistry that preceded it. Systems biology seems to be returning to the older biochemical view, worrying about processes, and how they change over time, but with a radical expansion of scale: in systems biology, thousands of reactants are potentially tracked over time rather than the ten or so which were the limit of classical biochemistry. Systems biology presents a much more dynamic view of biology than traditional molecular biology or even genomics. It promises both conceptual and technological innovations. If it leads to a successful model of even a single cell, it will already have justified the massive spending of the genome sequencing projects.

6. Philosophical Implications

It is time to draw some philosophical implications, first about reductionism which has long been of interest to philosophers, next about the notion of biological information which has recently seen a rapid growth of philosophical attention,[13] and finally about the return of temporal considerations in molecular biology.

6.1. Beyond reductionism?

One of the few philosophical issues in molecular biology that have routinely been discussed is that of reductionism [See Reductionism]. Here, reduction will be construed as

13 That is, relative to other issues in molecular biology; no area of molecular biology has received the philosophical attention it deserves, as Section 7 will note.

the explanation of wholes by parts, that is, reductionist explanations are those in which the weight of a putative explanation is borne by properties of the parts alone.[14] The wholes are biological entities, from cellular organelles to entire organisms. The parts are macromolecular and other components of the cell (and the extra-cellular matrix). Reductionism is the (empirical) thesis that explanations in some discipline will continue to be reductionist. The four rules of classical molecular biology embrace such reductionism and the remarkable success of classical molecular biology marks one of the most important triumphs of reductionism in the history of science (Sarkar, 1998). From the perspective of a reductionist, perhaps the most satisfactory aspect of this success is that cooperative behavior (in the case of allostery) and feedback regulation (in the case of the operon) were accommodated under the reductionist rubric in spite of being important exemplars from the traditional holists' repertoire.[15]

Moreover, the fact that the four rules of classical molecular biology are being challenged (recall the end Section 4), at least to some extent, is not reason enough to generate any new skepticism about the reductionist interpretation of explanation in molecular biology. They do not bring the physical explanation of wholes by parts into question. Rather, they show that the physical rules needed to explain macromolecular behavior are more complicated than previously thought, for instance, by an enzyme's active site inducing a fit with a reactant rather than merely responding to it. In contrast, if RNA-based (or other) regulatory networks turn out to be crucial to explaining development (and evolution, as Mattick [2003] argues – see Section 4), the reductionist interpretation *may* be in trouble. If network-based explanations are ubiquitous, it is quite likely that what will often bear the explanatory weight in such explanations is the topology of the network rather than the specific entities of which it is composed.[16]

Topological explanations have not received the kind of attention from philosophers they deserve even though networks have lately entered the center stage of scientific attention (Mattick & Gagen, 2005). Here "topology" refers to the connectivity properties of systems such as networks which, without loss of generality, can be modeled as directed graphs. The vertices of such a graph represent components of a system, and edges (between vertices), with appropriate directionality and weights, represent interactions between such vertices. How topological an explanation is becomes a matter of degree: the more an explanation depends on individual properties of a vertex, the closer an explanation comes to traditional reduction. The components matter more than the structure. Conversely, the more an explanation is independent of individual properties of a vertex, the less reductionist it becomes. In the latter case, if explanations invoke properties of a graph that measure its connectivity, then these are topological explanations. Such connectivity measures include the number of edges in the graph, the distribution of edge degree between vertices (the "degree" of a vertex being the number of edges incident on it), and so on.[17]

---

14 This is what Sarkar (1998) has called "strong" reduction – for a more carefully characterized treatment of varieties of reduction and reductionism, consult that work.

15 Recall the discussion of Section 2; for more detail, see Sarkar (1998).

16 Some classical phenomena such as dominance have already been interpreted to resist straightforward reductionist explanation (Sarkar, 1998).

17 For a review of network theory, see Newman (2003).

If topological explanations become necessary in molecular biology, it will mark a serious philosophical break with the reductionist classical era, though one that is not completely unexpected. Sarkar (1998) noted how the phenomenon of dominance had no straightforward structural explanation at the molecular level. Rather, the best molecular explanation of dominance involved complex reaction networks, the topological structure of which accounted for why one allele rather than the other was expressed at the phenotypic level.[18] This model predicts that dominance would be ubiquitous because such networks are common. Such an explanation depends very little on exactly what molecules comprise a network. If such network-based models begin to thrive in the post-genomic era, the reductionist interpretation of molecular biology will be seriously threatened.

Finally, systems biologists also reject reductionism – see, for instance, Aderem (2005) – even though the project of system biology emerged from the large-scale genome sequencing projects that had taken reductionism to its limits within biology (Tauber & Sarkar, 1992). As noted earlier (Section 3), contrary to most expectations, the results of sequencing only showed how little functional biology can be read off from sequences alone. Some systems biologists explicitly abandon reductionism to endorse philosophical doctrines such as emergence, according to which properties of wholes cannot be predicted or explained from the properties and organization of parts (Aderem, 2005). Few philosophers who defend reductionism will accept emergence easily, but the question can only be decided when the holists have specific examples in which properties of composite systems have deep explanations but none in terms of their parts. It will be a while before systems biology models get to that stage.

6.2. Beyond DNA information?

As noted earlier (Section 3), it is no longer clear that an informational account is appropriate for molecular biology. Even in the context of an informational account, the developments within eukaryotic genetics and, especially, genomics strongly suggest the view that DNA is the *sole* carrier of information. However "information" is explicated, such a view of DNA cannot be sustained for organisms more complicated than prokaryotes. Most of the critical interactions that determine the future behavior of a cell seem to occur at the level of RNA: splicing, RNA editing, and so on. Because of this feature of cellular interactions, Sarkar (2005, ch. 14) has speculated that the DNA genome consists of a relatively static set of sequestered modular templates (resulting in the "SMT" model of the genome), far from the classical view of the genome coding a program for development. The failure of the sequence hypothesis for many proteins only increases skepticism about the classical picture.

The routine generation of untranslated RNA transcripts from the genome also suggests that, should cellular processes be viewed informationally, RNA networks form a parallel information-processing system partly independent from the genomic DNA (Mattick, 2003). At present, it is unclear whether such information must be viewed *semiotically*, as in the case of DNA, where there is a symbolic coding relation. Similarly,

18 The original model goes back to Kacser and Burns (1981).

the discovery of ubiquitous non-Mendelian genetic specification in *A. thaliana* (Lolle et al., 2005) also suggests that there is yet another parallel system of heredity that can also perhaps be viewed informationally and, once again, is not specified through DNA. However, it is also possible that all such phenomena are best interpreted not informationally but using the more traditional – generally structural – conceptual apparatus of physics and chemistry. However, the distinction between the two frameworks becomes blurred in the case of RNA because the relation between the sequence and three-dimensional conformation seems to be relatively straightforward, at least much more so than in the case of proteins.

Note, however, that in these discussions of biological information, two issues should be distinguished: (i) whether an informational framework for molecular biology is of any use; and (ii) whether, within any such framework, DNA (or, more restrictively, genomic DNA) is the sole repository of that information. The problems mentioned here provide an argument against the second claim, leaving open the status of the first.

6.3. The return of time?

One of the peculiar characteristics of molecular biology has been its avoidance of explicit reference to the temporal dimension of the biological processes going on inside the cell and at other levels. The problem with informational interpretations of molecular biology is that these have always been static: flows of information between nucleic acids, and from them to proteins, control of gene expression through negative feedback and switches – these mechanisms all replace explicit discussion of how the chemical composition of cells change dynamically. Time does not enter explicitly into these accounts of biology though, implicitly, such transfer must take place during some time interval. Systems biology seems to be returning to the older biochemical view, worrying about processes and how they change over time. Systems biology thus presents a much more dynamic view of biology than traditional molecular biology. If systems biology lives up to its promise, the end result will be radically different from the classical molecular biology (discussed in Section 1).

However, even if the nascent project of systems biology fails to develop into anything substantive, proteomics also brings back considerations of time to molecular biology. Recall that the proteome is not a static feature of the cell, let alone the organism: proteomics requires a commitment to the characterization of cellular and organismic change over time. Moreover, the recent discoveries of potentially ubiquitous RNA network-based regulation also underscore the importance of dynamic accounts explicitly taking time into account. Moreover, new micro-array techniques and their extensions are increasingly making temporal stages of cellular changes empirically accessible. The challenge remains to develop a theoretical framework to interpret the empirical information. Any such framework can begin with either a physicalist or an informational characterization of cellular processes or a mixture of both, though prospects for a physicalist account do not seem particularly promising because of the sheer complexity of the molecular networks involved (Sarkar, 2005, ch. 10). But a dynamic informational account also leads to uncharted territory.

In retrospect, what seems surprising is how successful the static framework for classical molecular biology has been given that organisms are obviously dynamic entities

undergoing development over time. It is hard not to predict a future in which molecular biology has an explicit temporal dimension in its models.

7. Conclusions: An Invitation

With perhaps the exception of the question of reductionism, molecular biology has not received the extent of philosophical attention it deserves, and the little that it has received has been limited to the classical period. There are at least two reasons why philosophers should invest more work on the subject: (i) without at least a partial methodological commitment to molecular concepts and techniques, any sub-discipline within biology will likely soon be relegated to irrelevance. Philosophy of biology that does not take molecular biology fully into account will remain incomplete; and (ii) modern molecular biology raises fundamentally new epistemological questions, especially about the relevance of physical versus semiotic or informational accounts that have both dominated discussions of biology for the last century and lived in uneasy tension with each other. The deployment of philosophical techniques – particularly formal techniques – may contribute significantly to the advancement of the field.

The most important task in the philosophy of biology for the next few decades will be to conceptualize the functional role of DNA within the cell so as to explain the surprising organization and other properties of the genome that were discussed earlier. Philosophers will also probably be faced with new problems that arise as molecular biology becomes a dynamic discipline (that is, one in which models have a temporal component to them), whether or not the program of systems biology flourishes. The extent to which the biological sciences are similar to and different from the physical sciences will then have to be reassessed. It also remains an open question whether the new molecular biology will finally be able to explain most, preferably all, facets of organismic development and perhaps help to integrate development with evolution [See Development and Evolution]. In all these areas physical and informational accounts will probably have to interact in order to create a consistent satisfactory picture. As Section 6 indicates, any such attempt must necessarily begin with a clearer account than what is currently available of what "information" must mean in a biological context. This is probably where philosophers have most to contribute to the future of molecular biology [See Biological Information]. Perhaps techniques from formal epistemology or semantics will enable progress where traditional biological tools have largely failed.

References

Aderem, A. (2005). Systems biology: its practice and challenges. *Cell*, 121, 511–13.

Barrell, B. G., Air, G. M., & Hutchison III, C. A. (1976). Overlapping genes in bacteriophage PhiX174. *Nature*, 264, 34–41.

Berget, S., Moore, C., & Sharp, P. (1977). Spliced segments at the 5′ terminus of Adenovirus 2 Late mRNA. *Proceedings of the National Academy of Sciences, USA*, 74, 3171–75.

Berk, A., & Sharp, P. (1978). Structure of the Adenovirus 2 Early mRNAs. *Cell*, 14, 695–711.

Cattaneo, R. (1991). Different types of messenger RNA editing. *Annual Review of Genetics*, 25, 71–88.

Cavalier-Smith, T. (1978). Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *Journal of Cell Science*, 34, 247–78.

Chow, L., Gelinas, R., Broker, T., & Roberts, R. (1977). An amazing sequence arrangement at the 5' ends of Adenovirus 2 messenger RNA. *Cell*, 12, 1–18.

Cook-Deegan, R. (1994). *The gene wars.* New York: Norton.

Crick, F. H. C. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 12, 138–63.

Crick, F. H. C. (1973). Project K: "The Complete Solution of *E. coli*." *Perspectives in Biology and Medicine*, 17, 67–70.

Gehring, W. (1998). *Master control genes in development and evolution: the homeobox story.* New Haven: Yale University Press.

Gerhart, J., & Kirschner, M. (1997). *Cells, embryos, and evolution.* Oxford: Blackwell Science.

Fox, T. D. (1987). Natural variation in the genetic code. *Annual Review of Genetics*, 21, 67–91.

Hahn, M. W., & Wray, G. A. (2002). The G-value paradox. *Evolution & Development*, 4, 73–5.

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics*, 2, 343–72.

Jacob, F. (1973). *The logic of life: a history of heredity.* New York: Pantheon.

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3, 318–56.

Kitano, H. (2002). Systems biology: a brief overview. *Science*, 295, 1662–4.

Koshland, D. E., Jr., & Hamadani, K. (2002). Proteomics and models for enzyme cooperativity. *Journal of Biological Chemistry*, 277, 46841–4.

Lercher, M. J., Urrutia, A. O., & Hurst, L. D. (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics*, 31, 180–3.

Lolle, S. J., Victor, J. L., Young, J. M., & Pruitt, R. H. (2005). Genome-wide non-Mendelian inheritance of extra-genomic information in *Arabidopsis*. *Nature*, 434, 505–9.

Mattick, J. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays*, 25, 930–9.

Mattick, J., & Gagen, M. J. (2005). Accelerating networks. *Science*, 307, 856–7.

Mayr, E. (1961). Cause and effect in biology. *Science*, 134, 1501–6.

Monod, J. (1971). *Chance and necessity: an essay on the natural philosophy of modern biology.* New York: Knopf.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45, 167–256.

Olby, R. C. (1974). *The path to the double helix.* Seattle: University of Washington Press.

Pauling, L., & Corey, R. B. (1950). Two hydrogen-bonded spiral configurations of the polypeptide chains. *Journal of the American Chemical Society*, 71, 5349.

Sarkar, S. (1996). Biological information: a skeptical look at some central dogmas of molecular biology. In S. Sarkar (Ed.). *The philosophy and history of molecular biology: new perspectives* (pp. 187–231). Dordrecht: Kluwer.

Sarkar, S. (1998). *Genetics and reductionism.* New York: Cambridge University Press.

Sarkar, S. (2005). *Molecular models of life: philosophical papers on molecular biology.* Cambridge, MA: MIT Press.

Stephens, C. (1998). Bacterial sporulation: a question of commitment? *Current Biology*, 8, R45–8.

Tauber, A. I., & Sarkar, S. (1992). The Human Genome Project: has blind reductionism gone too far? *Perspectives on Biology and Medicine*, 35(2), 220–35.

Thiéffry, D., & Sarkar, S. (1998). Forty years under the Central Dogma. *Trends in Biochemical Sciences*, 32, 312–16.

Weston, A. D., & Hood, L. (2004). Systems biology, proteomics, and the future of health care: towards predictive, preventative, and personalized medicine. *Journal of Proteome Research*, 3, 179–96.

Wiener, N. (1948). *Cybernetics.* Cambridge, MA: MIT Press.

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids – a structure for desoxyribose nucleic acid. *Nature*, 171, 737–8.

Waddington, C. H. (1962). *New patterns in genetics and development.* New York: Columbia University Press.

Williams, K. L., & Hochstrasser, D. F. (1997). Introduction to the proteome. In M. R. Wilkins, K. L. Williams, R. D. Appel, & D. F. Hochstrasser (Eds). *Proteome research: new frontiers in functional genomics* (pp. 1–12). Berlin: Springer.

Yockey, H. P. (1992). *Information theory and molecular biology.* Cambridge: Cambridge University Press.