Chapter 14

# From Genes as Determinants to DNA as Resource: Historical Notes on Development, Genetics, and Evolution

Sahotra Sarkar

## 14.1. **Introduction**.

H. J. Muller's 1926 address to a symposium on "The Gene" at the International Congress of Plant Sciences was titled "The Gene as the Basis of Life."[1] For Muller, genes were capable of self-reproduction; consequently, they must have autocatalytic properties. On this basis, Muller argued, the "gene . . . arose coincidentally with growth and 'life' itself."[2] Not only were genes thus constitutive of life, Muller went on, but all of evolution must be explained from a genetic basis: "in all probability all specific, generic, and phyletic differences, of every order, between the highest and lowest organisms, the most diverse metaphyta and metazoa, are ultimately referable to changes in . . . genes."[3] The same year, Muller's mentor, T. H. Morgan, published *The Theory of the Gene*. The book summarized fifteen years of research, primarily on the fruit-fly (*Drosophila melanogaster*), that established the hegemony of genetics in twentieth-century biology. The purpose of this concluding chapter is to reflect on the history of genetics during that century, sketch how it came to dominate discussions of both development and evolution (helping to maintain their long divorce), and finally to speculate how the emergence of genomics and proteomics may be leading to a radically different agenda for biology.

Trained as a turn-of-the-century embryologist, Morgan had denied the full significance of both Darwinism and Mendelism at least until 1910 when he discovered sex-limited Mendelian inheritance of a trait (the mutant white eye in *D. melanogaster*).[4] That discovery spawned a path-breaking research program in genetics. By 1926, Morgan and his laboratory had investigated about 400 mutant characters of *D. melanogaster*. Through the systematic use of linkage mapping (invented by Morgan's student, A. H. Sturtevant, in 1913) these characters were partitioned into four linkage groups corresponding to the four chromosome pairs of Drosophila. The publication of The *Theory of the Gene* marked the completion of one of the most innovative research programs of twentieth-century biology.

That nothing was known about the developmental genesis of these traits at the level of cell, tissue, or organ, had not in any way impeded these investigations. Thus, by 1926, Morgan had not only come to accept and insist upon Mendelism as the theory of heredity, he was ready to demand a sharp divorce of genetics from development:

> "Between the characters, that furnish the data for the [Mendelian] theory and the postulated genes, to which the characters are referred, lies the whole field of embryonic development. The theory of the gene, as here formulated, states nothing with respect to the way in which the genes are connected with the end-product or character. The absence

> of information relating to this interval does not mean that the process of embryonic development is not of interest for genetics . . . but the fact remains that the sorting out of the characters in successive generations can be explained at present without reference to the way in which the gene affects the developmental process."[5]

Morgan was not the first to suggest such a strategy of genetic analysis; in 1914, William Bateson, in his Presidential Address to the British Association for the Advancement of Science, had also noted that the possibility of this separation is the characteristic feature of the new Mendelian genetics.[6]

Meanwhile, genes were slowly acquiring the material reality that most skeptics of Mendelism had long demanded of them. Muller's successes at inducing mutations through physical processes, particularly X-rays, added confidence to the position that genes were associated with definite material objects.[7] The physical interpretation of Mendelism helped establish what came to be called classical (transmission) genetics. In the 1930s and early 1940s, genes were thought to be composed of protein; nucleic acids composed of only four nucleotide bases (*A*: adenine; *C*: cytosine; *G*: guanine; and *T*: thymine) were believed not to be complex enough to provide the variability required to specify the several hundred known genes. However, in 1944, Avery, MacLeod, and McCarty experimentally demonstrated that, at least in bacteria, genes were composed of DNA.[8] The same year, the physicist, Erwin Schrödinger, in a book called *What is Life*?, produced an ingenious combinatorial argument showing that even composites from a small number of building blocks can have more than the amount of variety required of genes.[9] While the significance of this argument was largely unrecognized in the 1940s, Schrödinger's book played a key role in encouraging physical scientists to tackle biological problems, leading to the rapid expansion of molecular biology in the 1950s.

The most crucial development in physical studies of the gene was the decipherment of the structure of DNA by Watson and Crick in 1953.[10] While what generally gets emphasized is the double-helical structure of the model, what is critical to its eventual role in biology is the model of genetic specificity it incorporates. The term "specificity" was introduced in a genetic context by H. A. Timoféeff-Ressovsky and N. W. Timoféeff-Ressovsky only in 1926.[11] However, the specificity of gene action was a presumption of genetics from its inception. Originally proposed as a one-to-one correspondence between gene and trait, the idea survived in an increasingly mitigated form throughout the twentieth century. The double helix provided a model of specificity entirely new in biology: specificity was achieved by the order of arrangement of nucleotide bases, on the possibility of which only Schrödinger had speculated. This model ushered in the age of biological information: information interpreted as a sequence or arrangement of bases became the model of specificity for genetics.[12] Most importantly, it led to the view that genes were the sole purveyors of biological information. Crick summarized the view 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."[13]

The contrast here is with the older physical model of specificity, stereospecificity, dating back to the immunologist Paul Ehrlich's side-chain theory from the 1880s, which had begun to dominate structural studies in biology in the 1920s and 1930s.[14] The rise of the informational perspective also reified the view, articulated by Muller, that genes as determinants of biological features were special, different from the other resources used by organisms during development. The specificity of the gene-gene product (nucleic acid or protein) relationship was informational and thus different from specificity at every other level of biological organization, which remained physical (or stereospecific). Thus arose the view of DNA as the master molecule in charge of development—see Section 14.3. Section 14.2, meanwhile, discusses the displacement of other views of development during the process of establishing the hegemony of genetics.

14.2. **Evocators of Development**.

Around 1900, for biologists in the field and in the laboratory, it was far from obvious that organismic traits could be inherited through discrete units like Mendel's factors. There were two problems:

(i) discrete Mendelizing traits were rare. Most traits varied continuously (or were "quantitative") and were often normally distributed around a population mean, as hypothesized by the biometricians.[15] Moreover, their inheritance seemed to follow rules such as the biometricians' Law of Ancestral Inheritance; and

(ii) developmentally, not only was there no one-to-one correspondence between traits and hereditary factors, there was also ample evidence that the relation between them was not even determinate.

For instance, the German zoologist, R. Woltereck, studied morphologically distinct strains of Daphnia and Hyalodaphnia species from different lakes. These were pure lines which maintained their form through several generations of parthenogenesis. Woltereck focused on continuous traits such as head-height at varying nutrient levels. For both genera, the phenotypes varied between different pure lines, were affected by some environmental factors such as nutrient levels, were almost independent of others such as the ambient temperature, and showed cyclical variation with factors such as seasonality. Moreover, the response of a phenotype to the same environmental change was not identical in different pure lines. Woltereck drew "phenotype curves" to depict this phenomenon. These curves changed for every new variable that was considered. There were thus potentially an almost infinite number of them and Woltereck coined the term "*Reaktionsnorm*" to indicate the totality of the relationships embodied in them.[16] (It was only later that Woltereck's individual phenotype curves came to be called norms of reaction [or reaction norms].)

Woltereck argued that what was inherited was this *Reaktionsnorm* and that hereditary change consisted of a modification of that norm. Even W. Johannsen, who first explicitly made a sharp distinction between genotype and phenotype, endorsed the concept of the reaction norm, which he thought to be "nearly synonymous" with "genotype."[17] Only slightly later, H. Nilsson-Ehle coined the term "plasticity" to describe the non-unique relation of the genotype to the phenotype and argued that this has general adaptive significance.[18] This view found resonance in the Soviet Union where the norm of reaction (understood as what Woltereck had described as individual phenotypic curve) emerged as a concept of central importance in genetics. An avoidance of genetic determinism was clearly concordant with the Soviet program of producing an interpretation of science based on dialectical materialism; phenotypic plasticity, as modeled by variable reaction norms, furthered that project. However, in the West (that is, the United States and Europe outside the Soviet Union), where Johannsen's sharp genotype-phenotype distinction became part of the standard picture of genetics, the subsequent decades witnessed a general trend to emphasize the constancy and causal efficacy of the genotype at the expense of the complexity of its interactions. The norm of reaction remained a relatively ignored concept during this period.[19]

Ironically, the conceptual reticulation of classical genetics that helped maintain the primacy of the gene also emerged from developments in the Soviet Union. There, in the 1920s, an active genetical research group formed around the pioneering population geneticist, S. Chetverikov.[20] In 1922, one member of the group, D. D. Romashoff, discovered the *Abdomen abnormalis* mutation in *Drosophila funebris* which resulted in the degeneration of abdominal stripes.[21] There was individual variability in the mutant phenotype which Romashoff interpreted as a difference in the strength of the mutation's effect. The manifestation of the mutation depended on environmental factors, in particular, on the dryness and liquid content of food, but Romashoff could not rule out the possible influence of other loci. Another member of that group, N. W. Timoféeff-Ressowsky studied the recessive *Radius incompletus* mutation of *D. funebris*.[22] In mutant flies, the second longitudinal vein did not reach the end of the wing. Timoféeff created different pure lines, each homozygous for this mutation. Descendants included phenotypically normal flies. The proportion of normals was fixed for each pure line but varied between lines. External factors had little influence; the differences between the lines were apparently under the control of genotypic factors. Some lines gave a large proportion of mutants but manifested the mutation weakly; in others, the converse was realized. There were many intermediate lines.

The Soviet work was carefully followed by the German neuroanatomist, O. Vogt, who was a frequent visitor to Moscow because of a project to dissect Lenin's brain to demonstrate his genius.[23] Vogt, long committed to a genetic interpretation of psychoses, introduced two new concepts to describe Timoféeff's results: a mutation's "expressivity" was the extent of its manifestation; its "penetrance" was the proportion of individuals carrying it which manifested any effect at all. The differences between different lines were entirely ignored in Vogt's definitions. Expressivity and penetrance became properties of the gene rather than a property of a mutation relative to a constant genetic background. Yet, for historical reasons that are not entirely clear, Timoféeff enthusiastically endorsed the new concepts.[24] What the original results of Romashoff and Timoféeff had shown was a *predictable complexity* in the genotype-environment interaction. Both data sets permitted the construction of norms or reaction though Vogt's reinterpretation made such a move moot. Two related aspects of that reinterpretation deserve emphasis: (i) Vogt ignored the *systematic* differences between pure lines; and (ii) he explicitly introduced expressivity and penetrance as properties of genes on par with, though different from, dominance.

The introduction of expressivity and penetrance constituted a convoluted reticulation of the structure of Mendelian genetics by an *ad hoc* complication introduced to the concept of the gene. Besides having their standard transmission properties, genes were no longer only recessive or dominant (or displaying varying degrees of dominance); they also had degrees of expressivity and penetrance. There was no clear distinction between expressivity and dominance: expressivity, as defined by Vogt, is indistinguishable from the degree of dominance. In retrospect, the purpose that the new concepts served was to maintain a genetic etiology in the face of recalcitrant phenotypic plasticity induced by the complexity of genotype-environment interactions. Variability in the phenotypic manifestation of a trait became a result of a gene's expressivity and (indirectly) its penetrance. If the presence of a gene *for a trait* nevertheless failed to produce the trait, a genetic etiology for the trait was still maintained by simply positing that the gene had incomplete penetrance. If the presence of that gene led to the presence of the trait, but only to some variable degree, the gene was still responsible for the trait but had variable expressivity. The terms “penetrance” and “expressivity” were introduced into the English literature by C. H. Waddington in his *Introduction to Modern Genetics* where they were incorrectly attributed to Timoféeff.[25] Waddington's book, along with Timoféeff's growing prominence within Western genetics, made the terms common currency by the 1950s. Phenotypic plasticity—an almost inevitable outcome if development is the result of a suite of different factors, rather than only of the genotype—was relegated to irrelevance by mystifying the concept of the gene.

Waddington’s role in this story is curious. Though trained primarily as an embryologist, Waddington came to recognize the significance of the new genetics very early. In 1924, Spemann and Mangold had discovered the “organizer,” a region of the early embryo (at the gastrula stage) that seemed to direct subsequent development.[26] This led to an active research agenda by many embryologists to identify the “active principle” of the organizer. Committed reductionists believed this to be a chemical; Spemann, himself, had more holistic leanings. Waddington was among those to demonstrate that dead “organizers” could induce cell differentiation. By 1938 he had come to view organizers as “evocators” of development: “[t]he factor which, in the development of vertebrates, decides which of the alternative modes of development shall be followed is the organiser, or, more specifically, the active chemical substance of the organiser which has been called the evocator."[27] Waddington argued that changes of this sort are discrete, that is, there are definite developmental pathways with no intermediates between them. Because genes were also discrete, Waddington argued that "genes . . . act in a way formally like . . . evocators, in that they *control* the choice of alternative."[28]

For Waddington, the *aristopedia* class of alleles (*aristopedia*; *aristopedia*-*Spencer*; and *aristopedia-Bridges*) at the spineless locus of the third chromosome of *Drosophila melanogaster* provided an apposite example. The presence of the first two alleles from this class (*aristopedia* and *aristopedia*-*Spencer*) led to the transformation of the arista into a tarsus. In the case of the third (*aristopedia-Bridges*), the change was less marked but, even in this case, there was no true intermediate. Rather, a smaller number of segments were altered thus showing that a discrete change had taken place. Waddington's invocation of the language of "control" would be of critical significance after the advent of molecular biology—see Section 14.3. What is critical here is that his work marks the first serious attempt to synthesize genetics and development and it presumes, without argument, the primacy of the gene. Following through on this assessment of the importance of genes, in the 1940s, Waddington shifted the focus of his research from classical embryology to the genetic control of tissue differentiation in Drosophila.[29]

If Morgan had merely argued for a divorce of genetics from development, Waddington, in effect, demanded the subjugation of the former to the latter. A quote, though from a later period, emphasizes this point: “we know that genes determine the specific nature of many chemical substances, cell types, and organ configurations; and we have every reason to believe that they ultimately control all of them.”[30] Given the dominance of developmental genetics in developmental biology since the 1960s, Waddington’s choice of “control” hardly seems unusual today. But, in the embryology of the 1920s and 1930s (and earlier periods), reproduction was an important component of development: a full developmental cycle included reproduction. From a developmental perspective, the one from which Waddington emerged, it makes just as much, if not more, sense to explicate and emphasize the developmental determination of genetics through the control of reproduction, rather than to stipulate the genetic control of development. Nevertheless, Waddington made that fateful move with far-reaching consequences for the study of development in the twentieth century.

### 14.3. **The Age of the Master Molecule**.

As noted in Section 14.1, the construction of the double helix model for DNA and the informational model of biological specificity in 1953 radically altered the conceptual terrain of biology, at least at the organismic and lower levels of organization. Schrödinger had already speculated on the existence of a "hereditary code-script" in 1943; starting in 1954, another physicist, George Gamow, began an explicit program of deciphering the "genetic code."[31] The hope was to discover substantive properties of the code from simple formal rules incorporating functional assumptions about the efficiency and fidelity of information storage and transmission. As the mathematician, S. W. Golomb, put it: "[i]t will be interesting to see how much of the final solution [of the coding problem] will be proposed by the mathematicians before the experimentalists find it, and how much the experimenters will be ahead of the mathematicians."[32] As is often the case, biology was not kind to the mathematicians: the theoretical program of deciphering the code was an unmitigated failure. The code that was experimentally deciphered in the early 1960s had none of the elegance envisioned by the theorists. In spite of this failure, this theoretical research program had one lasting consequence: it helped bring to prominence the idea that the genome should be construed as a computer program. The emergence of this idea was encouraged by the context in which it occurred: this was the period that saw the beginning of large-scale digital computation.[33]

Two papers from 1961, with radically different agendas, explicitly introduced the idea of the genome as a blueprint and a program to be interpreted during development. In their classic paper laying out the details of the operon model for gene regulation, Jacob and Monod concluded:

> "The discovery of regulator and operator genes, and of repressive regulation of the activity of structural genes, reveals that the genome contains not only a series of blueprints, but a co-ordinated program of protein synthesis and the means of controlling its execution."[34]

What is critical about this passage is that agency resides in the genome: it controls the execution of the instructions in it. The fact that these instructions were already being interpreted as information gave credence to the metaphor of a genomic program. The operon model solved the decade-old problem of enzymatic adaptation through gene regulation. Later it became the standard model of gene regulation for most prokaryotic genes as discussed below.

A much more extended and careful discussion of programming and computation came in a paper the major purpose of which was to delimit the domain of molecular biology, that is, prevent its intrusion into organismic biology. In "Cause and Effect in Biology," Mayr notoriously distinguished "proximate" causes investigated by molecular biology from "ultimate" causes that are only provided by evolutionary biology. Evolution is the programmer producing a code that plays itself out in an individual, allowing individual behavior to be purposive:

> "An individual who—to use the language of the computer—has been 'programmed' can act purposefully. . . . . Natural selection does its best to favor the production of codes guaranteeing behavior that increases fitness. . . . . The purposive action of an individual, insofar as it is based on the properties of its genetic code, therefore is no more nor less purposive than the actions of a computer that has been programmed to respond appropriately to various inputs."[35]

Once again, agency resides in the genome, but because of natural selection and, in contrast to Jacob and Monod’s interpretation of the operon, not because of physical or chemical mechanisms.

The critical feature of the operon model was that the regulation of gene activity apparently occurred at the genetic level. This was an unexpected development: while the problem of gene regulation was recognized as being critical to understanding development since a pioneering paper by Haldane in 1932, it was generally believed that the mechanism of control would operate from the cellular level.[36] (Developmental holists believed that the mechanism would operate from even higher levels, for instance, from that of the tissue or organ.) In 1962, in *New Patterns in Genetics and Development*, Waddington seized upon the operon model to argue that regulation at the genetic level provides an explanation of tissue differentiation.[37] Differentiation was thus a matter of switching genes on or off. Even more controversially, Waddington interpreted other spatial developmental phenomena—histogenesis, morphogenesis, pattern formation, *etc.*—as special cases of differentiation.[38] Thus begun the program of a developmental genetics, of explaining development from a genetic basis, which took over the study of development in the 1970s.

A detailed history of developmental genetics is yet to be constructed. From the perspective of that sub-domain, the crucial developments were the discoveries of the homeobox sequence and *HOX* genes in the 1980s which were supposed to control much of morphogenesis.[39] *HOX* and similar genes do have significant regulatory roles in many species. Nevertheless, the confidence of geneticists in “master control genes” for development went far beyond what the data justified. This confidence was reflected in the initiation of the Human Genome Project (HGP) (and, later, other sequencing projects) in the

1990s. Blind sequencing of genomes was supposed to reveal the mechanisms by which biological processes operated at all levels of organization.

The trouble is that, by the late 1980s, there was ample reason to believe that DNA sequences alone would reveal little about biology even at the cellular level, let alone at higher levels. The informational model for DNA sequences as functional genes worked well provided that two conditions were satisfied:

(i) a *sufficiency* condition—inspection of the presence of a DNA sequence in a cell is sufficient to infer a capacity to produce the encoded protein; and

(ii) a *uniqueness* condition—a single DNA sequence produces exactly the encoded protein.

If these two conditions are satisfied the genetic code can be used—as a look-up table—to predict the amino acid sequence of the encoded protein. For prokaryotes, these conditions are satisfied: all DNA sequences, besides regulatory ones and sequences specifying transfer or ribosomal RNA (tRNA and rRNA), code for proteins and do so uniquely.

However, for eukaryotes, this picture begins to unravel.[40] Besides the standard genetic code, mitochondrial DNA and even nuclear DNA in some taxa use variant codes. The extent of such variation is at present unknown. Coding and regulatory regions of DNA are interspersed with long strands of DNA with no identifiable function.[41] These non-functional regions, when occurring within structural (or coding) genes, are transcribed into mRNA only to be spliced out before translation at the ribosome. (Such non-coding regions are called "introns"; coding regions are "exons."[42]) mRNA is also routinely edited through a variety of other mechanisms; bases are added and removed, sometimes in the hundreds. Perhaps the most surprising—and, in retrospect, the most important (see Section 14.4)—discovery was that of alternative splicing: the same mRNA transcript can be spliced in a variety of ways, leading to a set of different proteins. There is no evidence to suppose that the control of alternative splicing can be brought under the aegis of any simple genetic model such as the operon.

Of late it has even become controversial that, without significant modification of the concept of biological information, any informational model of biological specificity can survive. The few attempts to rescue that model deny any claim that genes are the solve purveyors of biological information (see, for instance, Chapter 10). But if they are not, developmental genetics, by itself, has no prospect of providing an adequate model of development. There is more to the phenotype than what can be specified by the genotype.

### 14.4. **After the Human Genome Project**.

The initiation of the Human Genome Project (HGP) was perhaps the most contentious episode in the history of science policy in biology to date. If the HGP is judged by the explicit promises that its proponents made in the late 1980s and 1990s to secure public support (and funding), it has been an unmitigated failure, the most colossal misuse ever of scarce resources for biological research. In 1992, Walter Gilbert claimed:

> "I think there will be a change in our philosophical understanding of ourselves. . . . Three billion bases [of a human DNA sequence] can be put on a single compact disc (CD), and one will be able to pull a CD out of one's pocket and say, 'Here's a human being; it's me!'"[43]

Today the claim seems laughable. None of the promises of Gilbert's radical genetic reductionism has been borne out. Proponents of the HGP promised enormous immediate medical benefits. Arguably, at least, there has been none. Gilbert routinely promised the birth of a new theoretical biology. Instead, emphasis now is on informatics: the design of computational tools to store and retrieve sequence information efficiently and reliably, with little expectation that any great theoretical insight is forthcoming. Commenting on the complexity of sporulation choice by an organism no more complex than *Bacillus subtilis*, C. Stephens recently pointed out:

> "Despite the explosive rate at which sequence databases are growing, and the concomitant increase in computing power available for sifting through them, sequence gazing alone cannot predict with confidence the precise functions of the multitude of coding regions in even a simple genome! Experimental analysis of gene function is still critical, a thought that brings with it the realization that the era of genomic analysis represents a new beginning, not the beginning of the end, for experimental biology."[44]

In one sense, from a perspective that takes the social responsibility of science seriously, to the extent that basic research should provide tangible immediate social benefits, this failure of the HGP is no doubt unfortunate. However, it is not unexpected: in the late 1980s and early 1990s, scientific skeptics of the HGP routinely pointed out that it would not deliver on its promises.[45] More importantly, social skeptics worried abut the use of DNA sequences for discrimination in health care and employment as well as social stigmatization. The failure of the HGP to deliver on its explicit promises provides an argument against the rationale for such uses of DNA and thus assuages some of these social worries, *provided that the failure is publicly recognized*. It must even have been abundantly clear to the proponents of the HGP that their original promises were unrealistic, leaving them vulnerable to charges of fraud in their presentations to public funding bodies.

Nevertheless, no biologist, including those who were initially skeptical of the project on scientific grounds, should any longer denounce the scientific results of the HGP. At the very least, the HGP has killed the facile genetic reductionism of the heyday of developmental genetics. There is little reason any more to suspect that claims of straightforward and irrevocable genetic determination of complex human traits will ever again be credible. It may even spell the extinction of molecular genetics itself, first transforming it into genomics, and then replacing it by proteomics.

The reason that the HGP may have such radical implications for biology is because of the startling properties discovered of the human genome sequence when compared to other species' sequences:

(i) the most important surprise from the HGP was that there are probably only about 31 000 genes in the human genome compared to an estimate of 140 000 as late as 1994.[46] Among the eukaryote genomes that have been fully sequenced, the human estimate remains the highest (in 2001), but not by much. Plant genomes are expected to contain many more genes than in the human sequence. It is already known that the mustard weed, *Arabidopsis thaliana*, has 26 000 genes, almost as many as the human. Morphological or behavioral complexity is not correlated with the number of genes that an organism has. This has been called the G-value paradox[47];

(ii) the number of genes is also not correlated with the size of the genome, as measured by the number of base pairs. *D. melanogaster* has 120 million base pairs but only 14 000 genes; the worm, *Caenorhabditis elegans* has 97 million base pairs but 19 000 genes; *Arabidopsis thaliana* has only 125 million base pairs while humans have 2 9000 million base pairs[48];

(iii) at least in humans, the distribution of genes on chromosomes is highly uneven. Most of the genes occur in highly clustered sites.[49] Most genes that occur in such clusters are those that are expressed in many tissues—the so-called "housekeeping" genes.[50] 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 % of the human genome codes for proteins while 50 % of the genome is composed of repeated units. Coding and other functional regions (including regulatory regions) are interspersed by large areas of “junk” DNA of no known function. However, some functional regions, such as *HOX* gene clusters, do not contain such junk sequences;

(v) hundreds of genes appear to have been horizontally transferred from bacteria to humans and other vertebrates, though apparently not to other eukaryotes;[51]

(vi) once attention shifts from the genome to the proteome, a strikingly different pattern emerges. The human 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 % 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 sequences.[52]

At the very least, except in rare cases, the presence of a particular DNA sequence allows very little to be inferred about what happens in the proteome, let alone at higher levels of organization. At most, that piece of DNA is a potential resource for use during development. Dethroned DNA must find its place among other developmental resources. Some of these other resources are transferred inter-generationally through the material continuity of reproduction (for instance, through the maternal cytoplasm in most “higher” animals). Others are acquired from the environment (for instance, by accretion by some marine animals). Nevertheless, DNA may be special in many ways; as will be argued in Section 14.5, there is a strong case to be made for disparity between DNA and other molecular constituents of cells. All the same, DNA and *ipso facto*, the gene, can no longer be the locus of agency responsible for the structural and behavioral repertoire of living forms including their remarkable diversity.

### 14.5. **Concluding Remarks**.

In the proteomics age, the most important problem in the philosophy of biology is to conceptualize the functional role of DNA within the cell so as to explain the organization and other properties of the genome. This chapter will end with a preliminary attempt to do so by explicating one speculative model which makes some novel predictions, though these have yet to be fully operationally disambiguated from predictions of other more traditional models. The new model will tentatively be called the sequestered modular template (SMT) model of the cell. The construction of this model begins with the observation that the cell is probably the first spatially delimited living structure to have evolved. As biochemists realized in the 1910s and 1920s, the cell's functions are primarily carried out by proteins, mainly enzymes. There are two types of such functions: those that maintain structural and behavioral integrity, and those that encourage reproductive proliferation. Evolutionary biology puts an emphasis on the latter type of function. But the former are as, perhaps even more, important for at least two reasons[53]: (i) without the maintenance of structural and behavioral integrity at least up to reproductive age, there is no question of reproduction; and (ii) in many organisms, especially sexually-reproducing organisms, cellular functions continue beyond reproductive age.

Maintenance of integrity, as well as reproduction, requires the production of replacement parts. Enzymes wear out (in spite of being catalysts); transport-enabling molecular moieties on cell membranes get damaged, as do the membranes themselves. They must be replaced. There are two obvious ways to carry out replacement part production: (a) directly, by growth and fission of the relevant type; and (b) indirectly, using a template. Whether or not, during evolution, the second strategy originally arose and got fixed entirely by accident rather than selection, it has at least two advantages[54]:

(i) suppose that cellular processes are based on a small repertoire of basic chemical mechanisms (as is true in contemporary organisms). Then the direct process of growth and fission would be catholic: the conditions under which one molecular type gets produced will very likely lead to the production of many other molecular types. Indirect reproduction permits preferential control; and

(ii) templates can be sequestered from environmental insults in a way that the active molecules cannot. The latter must necessarily interact with the environment to maintain cellular functions.

For the cell, it makes sense to have templates and, then, to sequester them. There is thus a critical disparity between the templates and the product molecules: DNA and genes are thus special compared to the other molecular constituents of the cell. It makes even more sense to make these templates as physically stable as possible. It is again probably entirely an accident that the first templates were structurally simple molecules: most likely, RNA, the variation in which was entirely combinatorial (that is, in sequence). But template integrity was better protected by a switch to a more stable form: DNA. (For instance, the base, uracil (*U*), is easily transformed to *C* by deamination; DNA uses the more stable *T* instead of *U*.) Enclosing templates by a membrane helps protection: eukaryotes achieve it by producing nuclei (and also enclosing some genes in mitochondria and plastids). After enclosure, further tinkering to increase template protection would be evolutionarily advantageous. Thus, it makes sense to cluster genes when possible: protecting clustered sites is easier than protecting widely dispersed sites. Clustering happens in humans, as noted in Section 14.4. The puzzle is why it does not seem to occur, or occur as much, in the other genomes that have so far been sequenced. A possible resolution of this puzzle is that these genomes are smaller in size resulting in less scope for clustering.

It also makes sense that genes used as templates for many functions, and those that are critical resources in early development, should receive the most protection. *HOX* genes deserve and get such attention. From the perspective of the SMT model, resources should thus be preferentially deployed to protect such genes from mutation. Here, the SMT model makes a prediction partly in variance with the received gene-based evolutionary model. That model would explain the evolutionary conservation of such genes by the deleterious selective effects of such mutations. The SMT model claims that, in addition, repair mechanisms preferentially target such sites. Turning this qualitative claim of differential prediction of the two models into an exact claim will require a quantitative analysis of the SMT model.

Modularity enters this model at two levels: (i) modularity of the genes themselves; and (ii) modularity of functional sub-regions (exons) of genes. At the genetic level, modularity is achieved because, by and large, genes are non-overlapping and, much more importantly, they are separated from each other by long strands of non-functional DNA which helps prevent gene disruption during recombination, a presumably physical inevitability of chromosome duplication. There is obviously a trade-off between this benefit and that of clustering. At the sub-genic level, the benefits of modularity were clearly articulated by Gilbert in 1978.[55] Recombination in introns allows the combinatorial production of new proteins that are still likely to be at least partly functional because component parts have not lost their structural integrity. Gilbert also argued that point mutations at intron-exon boundaries can potentially alter splicing patterns and generate radically different proteins. According to the SMT model, this would be undesirable. The SMT model is inherently conservative: it predicts that such mutations are rare. Moreover, with respect to this mechanism of generating diversity, the SMT model is consistent with the strategy used by the immune systems of mammals, in which recombination rather than somatic mutation (anywhere, and not just at intron-exon boundaries) is the preferred mode of the generation of diversity (though both processes are known to occur).

If modules are being carefully protected—and, therefore, evolutionarily conserved—it makes sense to use the same modules for a variety of purposes. From this perspective, alternative splicing makes sense as a way to utilize templates efficiently. Two predictions of the SMT model about alternative splicing are:

(i) that there is an inverse correlation between genome-wide mutation rates over evolutionary time scales and the degree of alternative splicing in taxa; and

(ii) this is a result of mechanisms at the cellular level.

A low number of genes and a high level of alternative splicing implies that organisms so constructed rely on the use of a large segment of the available combinatorial space at the level of modules within proteins. That organisms of this sort appear to be more structurally and behaviorally complex than others suggests a strong correlation between modularity at the proteomic level and evolvability. These arguments show the fallacy of any attempt at reading an organism off from its DNA sequence alone.

There are several other arguments for the SMT model:

(i) the same gene is often “co-opted” for different functions during the course of evolution. Typically, co-option follows duplication. For instance, the aggregation of the cellular slime mold, *Dictyostelium discoideum*, during times of stress uses a 495-residue long cellular adhesion molecule (CAM). The evolution of multicellularity is believed to have involved use of multiple copies of the corresponding DNA. Eventually, these copies diverged and were then co-opted to produce variants such as N-CAM for neuronal aggregation and H-CAM for hepatic aggregation.[56] From the perspective of the SMT model, it makes sense to transform and use redundant copies of a template;

(ii) transfer RNA (tRNA) and ribosomal RNA (rRNA) are obviously critical to the function of a cell. DNA specific to such RNA (and not to proteins) forms a tiny fraction of the total genome. The number of complete sets of such DNA sequences is correlated with the size of the genome. For instance, for tRNA genes, the human mitochondrial genome has 1 complete set, the bacterium, *Escherichia coli*, has 100 such sets, and the human nuclear genome has 1 000 complete sets. The SMT model predicts that the correlation of genome size and the number of copies reflects the number of such sequences that are likely to be simultaneously necessary: it makes sense to have exactly the optimal amount of some resource. It is unclear whether the received gene-based evolutionary model would make the same prediction. Most importantly, if the last argument is correct, then these different copies should not evolve independently of each other. Rather, their evolution should be concerted as, indeed, appears to be the case[57];

(iii) the various types of RNA-editing systems that have so far been observed, usually classified as insertional or substitutional, rely on a highly heterogeneous class of mechanisms. Consequently they must have arisen independently in different lineages. Because RNA editing often corrects errors in transcription, the SMT model predicts that any available mechanism should be recruited for this purpose, and this would be encouraged throughout evolutionary history.[58] Thus, it makes sense that it evolved independently several times, resulting in a heterogeneous class of editing mechanisms.

These and other similar arguments suggest that the SMT model merits further exploration in future work.

There is, however, one central unresolved issue: the model, as sketched above, assumes that the cell is the locus of agency, that is, the level at which it is appropriate to model and tabulate benefits, costs, and accidents. But, is what is good, bad, or neutral, for the cell also the same at higher levels of organization in multicellular organisms? Cancer trivially shows that it is not always so. Buss and many others have noted the possibility of conflicts of interest between different levels of biological organization.[59] It is far from clear that all the arguments given above will carry over to higher levels of organization than the cell. Finally, there is no reason to suppose that agency resides at exactly one level of organization. If the SMT model is to be successful, it must be able to cope with the possibility and likelihood of distributed agency.

## Acknowledgments

This work was supported by the US National Science Foundation, Grant No. SES-0090036, 2002 –2003. For comments on an earlier version, thanks are due to Kelly McConnell and Alex Moffett.

## References

Adams, M. B. 1980. “Sergei Chetverikov, the Kol'tsov Institute, and the Evolutionary Synthesis.” In Mayr, E. and Provine, W. B. Eds. *The Evolutionary Synthesis: Perspectives on the Unification of Biology*. Cambridge, MA: Harvard University Press, pp. 242 -278.

Avery, O. T., MacLeod, C. M., and McCarty, M. 1944. “Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus III.” *Journal of Experimental Medicine* **79**: 137 -157.

Bateson, W. 1914. “Address of the President of the British Association for the Advancement of Science.” *Science* **40**: 287 –302.

Buss, L. 1987. *The Evolution of Individuality*. Princeton: Princeton University Press.

Cavalier-Smith, T. Ed. 1985. *The Evolution of Genome Size*. Chichester: Wiley.

Covello, P. S. and Gray, M. W. 1993. “ On the Evolution of RNA Editing.” *Trends in Genetics* **9**: 265 -268.

Crick, F. H. C. 1958. “On Protein Synthesis.” *Symposia of the Society for Experimental Biology* **12**: 138 -163.

Dobzhansky, T. 1937. *Genetics and the Origin of Species*. New York: Columbia University Press.

Falk, R. and Sarkar, S. 1992. “Harmony from Discord.” *Biology and Philosophy* **7**: 463 -472.

Gamow, G. 1954a. “Possible Mathematical Relation Between Deoxyribonucleic Acid and Proteins.” *Biologiske Meddelelser udviket af Det Kongelige Danske Videnskabernes Selskab* **22**(3): 1 -11.

Gamow, G. 1954b. “Possible Relation Between Deoxyribonucleic Acid and Protein Structures.” *Nature* **173**: 318.

Gilbert, W. 1978. “Why Genes in Pieces?” *Nature* **271**: 501.

Gilbert, W. 1992. “A Vision of the Grail.” In Kevles, D. J. and Hood, L. Eds. *The Code of Codes*. Cambridge, MA: Harvard University Press, pp. 83 - 97.

Golomb, S. W. 1962. “Efficient Coding for the Desoxyribonucleic Acid Channel.” *Proceedings of the Symposium for Applied Mathematics* **14**: 87 -100.

Gray, M. W. 2000. “RNA Editing: Evolutionary Implications.” In: *Nature Encyclopedia of Life Sciences*. London: Nature Publishing Group. http://www.els.net/ [doi:10.1038/npg.els.0003069]

Hahn, M. W. and Wray, G. A. 2002. “The G-Value Paradox.” *Evolution & Development* **4**: 73 -75.

Haldane, J. B. S. 1932. “The Time of Action of Genes, and Its Bearing on Some Evolutionary Problems.” *American Naturalist* **66**: 5 -24.

Haldane,J. B. S. 1942. *New Paths in Genetics*. New York: Random House.

International Human Genome Sequencing Consortium. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* **409**: 860 -921.

Jacob, F. and Monod, J. 1961. “Genetic Regulatory Mechanisms in the Synthesis of Proteins.” *Journal of Molecular Biology* **3**: 318 –356.

Johannsen, W. 1911. “The Genotype Conception of Heredity.” *American Naturalist* **45**: 129 -159.

Kay, L. E. 2000. *Who Wrote the Book of Life? A History of the Genetic Code*. Stanford: Stanford University Press.

Laubichler, M. and Sarkar, S. 2002. “ Flies, Genes, and Brains: Oskar Vogt, Nikolai Timoféeff-Ressovsky, and the Origin of the Concepts of Penetrance and Expressivity.” In Parker, L. S. and Ankeny, R. Eds. *Medical Genetics, Conceptual Foundations and Classic Questions*. Dordrecht: Kluwer, pp. 63 -85.

Lercher, M. J., Urrutia, A. O., and Hurst, L. D. 2002. “Clustering of Housekeeping Genes Provides a Unified Model of Gene Order in the Human Genome.” *Nature Genetics* **31**: 180 -183.

Mayr, E. 1961. “Cause and Effect in Biology.” *Science* **134**: 1501 -1506.

McGinnis, W. 1994. “A Century of Homeosis, A Decade of Homeoboxes.” *Genetics* **137**: 607 -611.

Morgan, T. H. 1910. “Sex Limited Inheritance in *Drosophila*.” *Science* **32**: 120 –122.

Morgan, T. H. 1926. *The Theory of the Gene*. New Haven: Yale University Press.

Muller, H. J. 1927. “Artificial Transmutation of the Gene.” *Science* **66**: 84 –87.

Muller, H. J. 1962. *Studies in Genetics*. Bloomington: Indiana University Press.

Nilsson-Ehle, H. 1914. “Vilka erfarenheter hava hittills vunnits rörande möjligheten av växters acklimatisering?” *Kunglig Landtbruksakadamiens Handlingar och Tidskrift* **53**: 537 -572.

Ohno, S. and Holmquist, G. P. .2001. “Evolutionary Developmental Biology: Gene Duplication, Divergence and Co-Option.” In: *Nature Encyclopedia of Life Sciences*. London: Nature. http://www.els.net/ [doi:10.1038/npg.els.0001062].

Provine, W. B. 1971. *The Origins of Theoretical Population Genetics*. Chicago: University of Chicago Press.

Romaschoff, D. D. 1925. “Über die Variabilität in der Manifestierung eines erblichen Merkmales (Abdomen abnormalis) bei *Drosophila funebris* F.”

*Journal für Psychologie und Neurologie* **31**: 323 -325.

Sarkar, S. 1991. "*What is Life?* Revisited." *BioScience* **41**(9): 631 -634.

Sarkar, S. 1998. *Genetics and Reductionism*. New York: Cambridge University Press.

Sarkar, S. 1999. "From the *Reaktionsnorm* to the Adaptive Norm: The Norm of Reaction, 1909 -1960." *Biology and Philosophy* **14**: 235 -252.

Sarkar, S. and Tauber, A. I. 1991. "Fallacious Claims for HGP." *Nature* **353**: 691.

Schrödinger, E. 1944. *What is Life? The Physical Aspect of the Living Cell*. Cambridge: Cambridge University Press.

Silverstein, A. M. 1989. *A History of Immunology*. San Diego: Academic Press.

Spemann, H. and Mangold, H. 1924. "Über induktion von embryonalanlagen durch implantation artfremder organisatoren." *Wilhelm Roux' Archiv für Entwicklungsmechanik der Organismen* **100**: 599 -638.

Stephens, C. 1998. "Bacterial Sporulation: A Question of Commitment?" *Current Biology* **8**: R45 –R48.

Tauber, A. I. and Sarkar, S. 1992. "The Human Genome Project: Has Blind Reductionism Gone Too Far?" *Perspectives on Biology and Medicine* **35**(2): 220 –235.

Tauber, A. I. and Sarkar, S. 1993. “The Ideology of the Human Genome Project.” *Journal of the Royal Society of Medicine* **86**: 537 -540.

Thiéffry, D. and Sarkar, S. 1998. “Forty Years under the Central Dogma.” *Trends in Biochemical Sciences* **32**: 312 -316.

Timoféeff-Ressovsky, H. A. and Timoféeff-Ressovsky, N. W. 1926. “Über das phänotypische Manifestieren des Genotyps. II. Über idio-somatische Variationsgruppen bei Drosophila funebris.” *Wilhelm Roux' Archiv für Entwicklungsmechanik der Organismen* **108**: 146 -170.

Timoféeff-Ressowsky, N. W. 1925. “Über den Einfluss des Genotypus auf das phänotypen Auftreten eines einzelnes Gens.” *Journal für Psychologie und Neurologie* **31**: 305 -310.

Waddington, C. H. 1938. *An Introduction to Modern Genetics*. London: George Allen & Unwin Ltd.

Waddington, C. H. 1939. “Genes as Evocators in Development.” *Growth* **1**: S37 -S44.

Waddington, C. H. 1940. *Organisers and Genes*. Cambridge, UK: Cambridge University Press.

Waddington, C. H. 1962. *New Patterns in Genetics and Development*. New York: Columbia University Press.

Wagner, A. 2002. “Gene Duplication and Redundancy.” In *Nature Encyclopedia of Life Sciences*. London: Nature.

http://www.els.net/ [doi:10.1038/npg.els.0001163].

Watson, J. D. and Crick, F. H. C. 1953a. “Molecular Structure of Nucleic Acids--A Structure for Deoxyribose Nucleic Acid.” *Nature* **171**: 737 -738.

Watson, J. D. and Crick, F. H. C. 1953b. “Genetical Implications of the Structure of Deoxyribonucleic Acid.” *Nature* **171**: 964 -967.

Woltereck, R. 1909. “Weitere experimentelle Untersuchungen über Artveränderung, speziell über das Wesen quantitativer Artunterschiede bei Daphnien.” *Verhandlungen der deutschen zoologischen Gesellschaft* **19**: 110 -173.

[1] Muller (1962), pp. 188 -204.

[2] Muller (1962), p. 200.

[3] Muller (1962), p. 195. A decade later, Dobzhansky (1937, p. 11) would echo the same sentiment, defining evolution to be a "change in the genetic [allelic] composition of populations."

[4] Morgan (1910).

[5] Morgan (1926), p. 26.

[6] Bateson (1914).

[7] Muller (1927).

[8] Avery, MacLeod and McCarty (1944).

[9] Schrödinger (1944). Sarkar (1991) provides an assessment of Schrödinger’s conceptual achievements.

[10] Watson and Crick (1953a). See Chapter 1 (§ 1.3.1) for a discussion of the significance of the double helix model.

[11] See Timoféeff-Ressovsky and Timoféeff-Ressovsky (1926).

[12] This started with Watson and Crick (1953a, b); see Chapters 8 -9 for details.

[13] Crick (1958), p153; emphasis in the original. Thiéffry and Sarkar (1998) provide a critical history of the central dogma.

[14] See Silverstein (1989).

[15] See Sarkar (1998) for a discussion of biometry. Provine (1971) provides a detailed history of this dispute.

[16] See Woltereck (1909), p. 135. For historical detail, see Sarkar (1999).

[17] See Johannsen (1911), p. 133.

[18] Nilsson-Ehle (1914).

[19] See Sarkar (1999) for further details.

[20] Adams (1980).

[21] Romashoff (1925).

[22] Timoféeff-Ressowsky (1925).

[23] For this curious history, see Laubichler and Sarkar (2002).

[24] See  Timoféeff-Ressovsky and Timoféeff-Ressovsky (1926).

[25] See Waddington (1938).

[26] See Spemann and Mangold (1924).

[27] See Waddington (1939), p. 37, elaborated in Waddington (1940).

[28] Waddington (1939), p. 37; emphasis added.

[29] Waddington (1962), p. 14.

[30] Waddington (1962), p. 4; this aspect of New Patterns in Genetics and Development will be further discussed in § 14.3.

[31] Gamow (1954a, b). For a history, see Chapter 9.

[32] Golomb (1962), p. 100.

[33] For more on this history, see Kay (2000).

[34] Jacob and Monod (1961), p. 354.

[35] Mayr (1961), pp. 1503 –1504.

[36] See Haldane (1932).

[37] Waddington (1962), pp. 20, 23.

[38] Waddington (1962), pp. 1 -3.

[39] For a balanced analysis, noting both the importance of, and the interpretive excesses about, these discoveries, see McGinnis (1994).

[40] For details of the  processes mentioned in this paragraph, as well as a guide to the literature, see Chapters 8 and 9.

[41] The existence of such DNA initially came as a relief since it resolved the C-value paradox, that genome size was not correlated with organismic complexity (see Cavalier-Smith [1985]).

[42] This terminology was introduced by Gilbert (1978).

[43] Gilbert (1992), p. 96.

[44] Stephens (1998), p. R47.

[45] See, for instance, Sarkar and Tauber (1991), Tauber and Sarkar (1992, 1993).

[46] See Hahn and Wray (2002) and the references therein.

[47] Hahn and Wray (2002).

[48] Hahn and Wray (2002).

[49] These have often been likened to "urban centers" while the gene-poor regions have been likened to "deserts." See http://www.ornl.gov/sci/techresources/Human_Genome/project/journals/insights.html.

[50] Lercher *et al*. (2002).

[51] International Human Genome Sequencing Consortium (2001). (This remains controversial.)

[52] Hahn and Wray (2002).

[53] See, in this context, Chapter 1 (§ 1.2.1).

[54] Haldane (1942) was the first to note the value of template-based reproduction and presciently suggested it as a model for gene duplication.

[55] See Gilbert (1978).

[56] See Ohno and Holmquist (2001).

[57] See Wagner (2002).

[58] See Covello and Gray (1993) for a traditional evolutionary interpretation of RNA editing; see, also, Gray (2000) for a recent review.

[59] Buss (1987), in particular, has argued for the relevance of such conflicts for the evolution of patterns of development, especially germ-line sequestration. See, also, Falk and Sarkar (1992).