
Journal of Biological Physics **28:** 745–764, 2002.

A 21st Century View of Evolution

J.A. SHAPIRO
Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637, USA

Abstract. Physicists question whether there are 'universals' in biology. One reason is that the prevailing theory of biological evolution postulates a random walk to each new adaptation. In the last 50 years, molecular genetics has revealed features of DNA sequence organization, protein structure and cellular processes of genetic change that suggest evolution by Natural Genetic Engineering. Genomes are hierarchically organized as systems assembled from DNA modules. Each genome is formatted and integrated by repetitive DNA sequence elements that do not code for proteins, much as a computer drive is formatted. These formatting elements constitute codons in multiple genetic codes for distinct functions such as transcription, replication, DNA compaction and genome distribution to daughter cells. Consequently, there is a computation-ready Genome System Architecture for each species. Whole-genome sequencing indicates that rearrangement of genetic modules plus duplication and reuse of existing genomic systems are fundamental events in evolution. Studies of genetic change show that cells possess mobile genetic elements and other natural genetic engineering activities to carry out the necessary DNA reorganizations. Natural genetic engineering functions are sensitive to biological inputs and their non-random operations help explain how novel genome system architectures can arise in evolution.

Key words: cellular computation, DNA rearrangements, genome system architecture, mobile genetic elements, natural genetic engineering, repetitive DNA, signal transduction

1. Complexity, Genetics and Evolution: The Argument in a Nutshell

Our Symposium is entitled 'Life as a Complex System.' In order to understand this topic, we have to ask, 'Why do living organisms use so many different molecules to carry out the basic tasks of survival, growth and reproduction?' With regard to genetics and evolution, I think an overall answer can be summarized in the following points:

- Complexity permits sophisticated information processing. Cells have to deal with literally millions of biochemical reactions during each cell cycle and also with innumerable unpredictable contingencies. They are constantly evaluating multiple interml and external signals and adjusting their activities to continue the basic processes of survival and reproduction. Cells carry out their computations by a process of molecular interactions. More molecules means more powerful computational capacity.

- Genomes integrate into cellular information processing because they are organized as computational storage organelles. That is, DNA serves as a data storage medium. To participate in cellular activities, genomes interact computationally with dynamic cellular complexes composed largely (but not exclusively) of proteins. As we shall see, genomes are built (Lego-like) of hierarchically organized modular systems. Much like the programs stored on a computer drive, genomic systems and subsystems are formatted by generic (i.e. repetitive) signals that provide functional addresses for the data in each module. The formatting is as important as the data (i.e. protein coding sequences) in providing a Genome System Architecture for each organism or species.
- Evolutionary genomic change occurs largely by a process of Natural Genetic Engineering. Systemic genome organization means that new functions arise by the cut-and-splice rearrangement of genetic modules. Living cells possess mobile genetic elements and other biochemical functions which carry out the underlying DNA rearrangements. Cells regulate the activation of natural genetic engineering functions. Thus, cells have a capacity for major genome reorganization in response to evolutionary crisis. Moreover, the fact that natural genetic engineering changes are neither random in nature nor restricted to a single site in the genome means that they can create novel distributed (multilocus) systems and new genome system architectures.

2. Universals in Biology

At a meeting of physicists, it is important to address the issue of whether there are 'Universals' in biology or, as many scientists believe, just many separate examples of specific systems, each evolved to take care of a particular adaptive need. One source of this latter view is the conventional theory that evolution occurs by a random walk through adaptive space and produces a virtually endless series of *sui generis* inventions. One alternative to this conventional view is that there exist design principles and procedures that are used repeatedly in evolution (in other words, evolution occurs as an engineering process). Consistent with this alternative 'evolutionary engineering' view is the fact that there are, indeed, a number of Universals in biology (Table I).

Some of these Universals have been known for a long time, but two of them developed out of the molecular biology revolution in the second half of the 20th Century. The first post-1953 Universal, the idea that cells compute and make decisions, is not new. But its widespread acceptance has only recently emerged from studies of biological regulation and the identification of countless molecules, multimolecular complexes and signaling systems which provide detailed control over the operation of virtually every aspect of cellular function [1, 2]. On a very short time-scale, this computational capacity allows cells to respond appropriately to internal and external signals, to adjust to changing conditions, to detect and cor-

Table I. Some universals in biology

Up to 1953
• Biochemical unity of life (chirality, metabolism, genetics)
• Biological self-organization
– assimilation of energy and matter into more biomass
– growth, development and reproduction
• Cellular organization of living matter
• Hereditary transmission of information (inheritance in cell) lineages, DNA replication and segregation)
• Diversity and evolution (greater resource utilization, enhanced survival) 1953–now
• Cellular computation and decision-making (short-term adaptations)
– Surveillance, sensitivity and signal transduction
– Biochemical and genetic regulation
– Error detection and repair, checkpoints
– Complexity and connectivity (reliability, precision, robustness)
• Built-in natural genetic engineering mechanisms for genome change (long-term adaptations)

rect misfunctions and to coordinate the millions of biochemical events involved in metabolism, growth, morphogenesis, cell division and multicellular development.

The second post-1953 Universal, the recognition that the vast majority of genetic change results from the action of cellular biochemical systems that act on DNA, is far less widely known and its significance is not appreciated outside a small group of specialists [3, 4]. The discovery of built-in natural genetic engineering mechanisms dates back to Barbara McClintock's pioneering cytogenetic studies of the late 1940s and early 1950s [5]. However, the ubiquity of internal systems for genome change only became apparent through molecular studies in bacteria in the 1960s and, with recombinant DNA technology, in eukaryotes in the 1970s and 1980s [6–8]. In terms of a 21th Century view of evolution, the major importance of natural genetic engineering is that this capability removes the process of genome restructuring from the stochastic realm of physical-chemical insults to DNA and replication accidents. Instead, cellular systems for DNA change place the genetic basis for long-term evolutionary adaptation in the context of cell biology where it is subject to cellular control regimes and their computational capabilities [9–11].

3. How Does the Genome fit in the Information Economy of the Cell?

The genome is the long-term storage medium for each species (much like a computer hard disk) and consists of the total information content of the DNA molecules in the cells of that species. Although most genomics researchers focus on the 'coding' regions of the genome that determine the proteins a species can synthesize,

Table II. Different classes of genomic information

• Coding sequences for RNA and protein molecules
• Start and stop sites for transcription
• Processing signals for primary transcripts
• Control signals for level of expression
• Control signals for dynamic access at right time and place. Identifiers for coding sequences that must be coordinately or sequentially expressed
• Signals for chromatin condensation/chromatin remodeling
• Signals for DNA replication
• Signals for distribution to daughter cells (non-random chromosome partitioning)
• Signals for error correction and damage repair
• Signals for reprogramming when necessary
(Specific references can be found in refs. 1, 2, 10 and 11)

genomes are built up of protein-coding and other classes of DNA sequences that are combinatorially formatted to carry out the multiple tasks necessary for overall genome function (Table II). While textbooks call the triplet code for amino acids in proteins '*the* genetic code,' there are in fact many genetic codes for the different aspects of genome coding, packaging, replication, distribution, repair and evolution.

An absolutely fundamental point to appreciate is that genomes only function through interaction with other dynamic information systems in the cell. By itself, DNA is inert. The information stored in DNA molecules requires interpretation by the highly dynamic cellular systems that control DNA packaging, imprinting, replication, transcription, translation, splicing, signal transduction, morphogenesis and so forth. The significance of these essential interactions is that the genome necessarily constitutes part of a computational system in every living cell. The character of an organism is not determined solely by its genome. For example, in species with complex life-cycles, the same genome encodes two quite distinct organisms, such as the caterpillar and butterfly. We also see the cell-dependent nature of genome function in mammalian cloning, where a nucleus is removed from one terminally differentiated cell type that is capable only of a highly restricted range of genome expression and inserted into an egg cell, where the genome can provide the stored information for embryonic development and occasionally encode the formation of a normal individual [12, 13].

4. Cellular Computation and Genome System Architecture

There are many ways to visualize the systemic nature of genomic coding. One that is discussed in other parts of these PROCEEDINGS is the organization of protein molecules as systems of discrete structural and functional domains encoded in evolutionarily mobile DNA modules [14] (these PROCEEDINGS, Symposia

B6 and B7). Here we will examine the basic principles of gene expression and transcriptional regulation and the evolution of our dimensionless concept of the gene into the more complex notion of a genetic locus.

4.1. THE *lac* OPERON: A SIMPLE BUT ILLUSTRATIVE EXAMPLE

Our example is the *lac* operon encoding the capacity for lactose utilization in the bacterium *E. coli*. Like virtually all classical genes, *lac* began existence as a point on a genetic map in the late 1940s, soon after the discovery of genetic exchange in bacteria (Figure 1; see 15, 16, for detailed references to *lac* operon history).

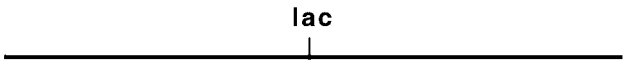

Figure 1. The *lac* gene, site of mutations affecting lactose utilization by *E. coli* (1947).

In the following years, Jacques Monod studied *lac* genetics because he had discovered that *E. coli* could discriminate between glucose and lactose; when given a mixture of the two sugars, the bacteria would invariably consume all the glucose before starting to consume the lactose. In 1961, Monod and his colleagues proposed the operon model. In the operon, *lac* had evolved into a system of 'structural genes' encoding the proteins of lactose transport and metabolism (*lacZYA*), a 'regulator gene' encoding a repressor (*lacI*) and a completely novel type of genetic element, the 'operator' (*lacO*) (Figure 2).

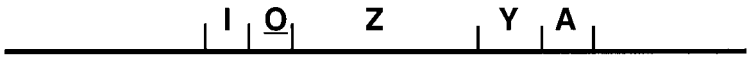

Figure 2. The *lac* operon (1961).

It is important to recognize that *lacO* was not a 'gene' encoding a product. Instead, it was a site on the DNA where the repressor bound to block the initial step of *lacZYA* expression from the adjacent DNA. This conceptual development was revolutionary in its impact on our understanding of genome function. Such *cis*-acting protein binding sites are now recognized as key components of the genome, essential for processes such as replication, transcription and genome distribution to daughter cells.

After the operon model, other scientists discovered additional binding sites in *lac*, including the site of RNA polymerase binding or 'promoter' (*lacP*), a site

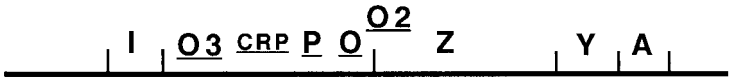

Figure 3. The *lac* operon (1990).

Table III. Computational operations in *lac* operon regulation

Operations involving *lac* operon products	
• LacY + lactose(external) → lactose(internal)	(1)
• LacZ + lactose → allolactose (minor product)	(2)
• LacI + *lacO* → LacI-*lacO* (repressor bound, *lacP* inaccessible)	(3)
• LacI + allolactose → LacI-allolactose (repressor unbound, *lacP* accessible)	(4)
Operations involving glucose transport components and adenylate cyclase	
• IIA^{Glc}-P + glucose(external) → IIA^{Glc} + glucose-6-P(internal)	(5)
• IIA^{Glc}-P + adenylate cyclase(inactive) → adenylate cyclase(active)	(6)
• Adenylate cyclase(active) + ATP → cAMP + P∼P	(7)
Operations involving transcription factors	
• Crp + cAMP → Crp-cAMP	(8)
• Crp-cAMP + *CRP* → Crp-cAMP-*CRP*	(9)
• RNA Pol + *lacP* → unstable complex	(10)
• RNA Pol + *lacP* + Crp-cAMP-*CRP* → stable transcription complex	(11)
Partial computations	
• No lactose → *lacP* inaccessible	(3)
• Lactose + LacZ(basal) + LacY(basal) → *lacP* accessible	(1, 2, 4)
• Glucose → low IIA^{Glc}-P → low cAMP → unstable transcription complex	(5, 6, 7, 10)
• No glucose → high IIA^{Glc}-P → high cAMP → stable transcription complex	(5, 6, 7, 8, 9, 11)

for binding the Crp transcription factor that mediates glucose control of *lacZYA* transcription (*CRP*) and two additional operator sites that permit cooperative binding of the repressor (*lacO2, lacO3*). Thus, by the mid 1980s, molecular genetic analysis had decomposed the dimensionless *lac* gene into a structured system of protein-coding and *cis*-acting regulatory sites (Figure 3).

The importance of the organization of the various *lac* regulatory sites is that they permit the molecular computations that allow *E. coli* to discriminate glucose from lactose – that is, to control expression of the lactose metabolic proteins so that they are only synthesized once glucose is no longer available. The basic biochemical reactions and molecular interactions involved in this computation can be stated as logical propositions that can then be combined into partial computations (Table III). These partial computations illustrate the molecular logic allowing the cell to execute the following overall computation: 'IF lactose present AND glucose not present AND cell can synthesize active LacZ and LacY, THEN transcribe *lacZYA* from *lacP*.'

Two aspects of this particular genomic computation deserve special mention. The first aspect is that the computation involves many molecules and compartments of the cell, not just DNA and DNA binding proteins. For example, the membrane transport proteins LacY and IIA^{Glc} are essential. The second noteworthy aspect is that the computation involves the use of chemical symbolism as information is transmitted. Thus, the presence of allolactose inducer represents the availability

of lactose and the ability of the cell to synthesize functional LacY and LacZ. Similarly, the concentrations of IIA^{Glc}-P and cAMP represent the availability of glucose to the cell. Both whole cell involvement and transient chemical symbols are typical of cellular computation and signal transduction in general.

4.2. ORGANIZATION OF HIGHER ORDER SYSTEMS IN THE GENOME

The *lac* operon is a relatively simple and thoroughly studied example of a genomic system. Despite its simplicity, it illustrates the basic features of how the genome is organized for interaction with other cellular information systems. One way to build higher-level systems in the genome is to use common binding sites at multiple genetic loci [10, 11, 17, 18 and U. Alon, these PROCEEDINGS]. The *CRP* site of the *lac* operon serves in just this way. It is located in the transcriptional regulatory regions of diverse loci in the *E. coli* genome, connecting them into a network of functions regulated by the availability of glucose as a preferred growth substrate [15].

Another complementary way of building up higher-order systems is to complexify the structure of the transcriptional regulatory regions in genetic loci so that they contain many more protein binding sites. This strategy has been adopted in higher eukaryotes that undergo elaborate processes of cellular differentiation and multicellular development. Elaborate transcriptional regulatory architectures permit the formation of developmental expression networks in which multiple genetic loci are coordinately regulated by shared subsets of specific binding site motifs [19]. Moreover, the combinatorial diversity of complex regulatory regions means that the expression of each genetic locus is subject to computationally intricate controls as the intracellular concentrations of different transcription factors change. A particularly well-studied example of such a regulatory region in the sea urchin has been analyzed at both the molecular and computational levels [20].

The transcription factor binding sites that serve to integrate multilocus genetic systems are one class of the highly diverse family of repetitive DNA sequences. With some exceptions, repetitive sequences do not encode proteins and their occurrence at multiple locations in the genome distinguish them from classical genetic loci that map to specific sites. These repetitive sequences are major players in genome organization. In the draft human genome, for example, only 3% of the sequence information is in protein-coding exons, but over 45% is in repeated DNA sequences [21, 22]. Repetitive DNA is far more taxonomically specific than protein-coding DNA and serves as the most reliable indicator of identity for a species or even, as used in forensic analysis, for identifying an individual [23].

Contrary to what most people believe (and to what is often asserted in discussions of genome evolution), the fact is that repetitive DNA elements play critical roles in functional organization of the genome (Table IV). The effects on gene expression are quite dramatic and relate to the role of repetitive DNA in altering the structure of DNA compacted with histones and other DNA binding proteins into an

Table IV. Functions of repetitive DNA sequence elements

- Coordinated expression of unlinked genetic loci: activator regions, silencing regions
- DNA replication origins and chromosome end stability (telomeres)
- Chromosome distribution during cell division: centromeres, chromosome pairing
- Chromatin organization and gene expression in development: chromatin domains, heterochromatin nucleation, insulator elements, position effects

(Specific references can be found in refs. 10, 11)

Table V. Genome system architecture

- Genome as an information storage organelle (cf. disk drive)
- Formatting for coding sequence expression: translation signals, codon bias, splicing signals, etc.
- Formatting for transcription: regulatory regions (5′, 3′ and intron), linkage of genetic loci, chromatin domains
- Formatting for replication: origins, telomeres, chromatin domains
- Formatting for DNA condensation and spatial organization in the nucleus: euchromatin, heterochromatin, laminar attachment sites
- Formatting for genome segregation: centromeres, chiasmata
- Formatting for genome reorganization and natural genetic engineering: transposable and repetitive elements, DNA rearrangement signals

→ 'Systems all the way down' (organized by repetitive signals)

overal structure called 'chromatin.' Repetitive DNA tends to form a densely packed DNA-protein complex called 'heterochromatin' which generally inhibits both expression and replication of the underlying DNA sequences. Classical cytogenetic studies of a phenomenon called 'position effect' have shown that expression of many genetic loci can be dramatically inhibited by proximity to heterochromatic regions of the genome [10, 24, 25]. By forming extended domains of distinct chromatin configuration, the distribution of repetitive DNA elements can coordinately regulate expression of groups of linked genetic loci.

The above considerations and many other studies have led to the concept that genomes have an overall organization (Table V) [10]. In keeping with our computational analogy, we may call this overall organization, unique to each species, its Genome System Architecture.

5. Implications of Genome System Architecture for Evolutionary Change

If it is true, as all sequencing projects indicate, that genomes are composed of Lego-like assemblies of smaller and larger modular genetic elements (segments of protein coding sequences, regulatory sites, repetitive DNA elements, chromatin domains), then it follows that a major source of genetic novelty must be the re-

arrangement of these modular components. For example, a new protein function can be generated by the rearrangement of existing coding modules (often referred to as 'domain- or exon-swapping') [26]. In addition, patterns of gene expression and coordination of multiple genetic loci can arise through the distribution of transcriptional regulatory sites among new groups of genetic loci or by rearranging the combinations of binding sites in complex control regions. Even higher order changes in functional organization can occur by creating new combinations of linked genetic loci and/or altering the boundaries of chromatin domains. This last process can occur through redistributing repetitive DNA elements without changes in the structure of individual genetic loci. In other words, by a process of cut-and-splice genetic engineering leading to the appropriate changes in the arrangements of diverse genetic modules, genomes can be altered locally, regionally or globally and they can be reformatted to acquire distinct system architectures. It may be postulated that such reformattings are more important events in the origin of new species and genera than gradual changes in proteins.

The existing whole genome sequences (in particular, the draft human genome sequence) appear to support this view of genome evolution. There is a pattern of the retention and reuse of genomic systems, which can often be adapted to novel cellular or organismal functions. The following general features have been observed:

- Many important proteins are formed by novel arrangements of conserved domains; new domains appear rarely, but new combinations are common [21, 22].

- Functional modules tend to be amplified; these modules can be domains, entire protein determinants, regulatory sites, entire genetic loci, complex higher order assemblies of multiple genetic loci, such as the homeobox clusters [27], or long chromosome segments [28]. The importance of duplications in evolution was pointed out many years ago by Ohno [29].

- Multiple chromosome rearrangements are involved in the relocation of amplified modules throughout the genome or in reorganizing similar chromosome segments shared between related organisms [21, 30, 31].

- There are major differences in the identity, numbers and locations of repetitive sequence elements between related organisms; often these repetitive elements define specific functional regions, such as the tandem arrays that help define chromosome centromeres [28].

In brief, the existing sequence databases indicate that genomes have undergone multiple instances of major DNA rearrangements. As we shall nee, all living cells possess the cellular systems necessary to carry out these rearrangements and thereby reorganize their genomes when necessary.

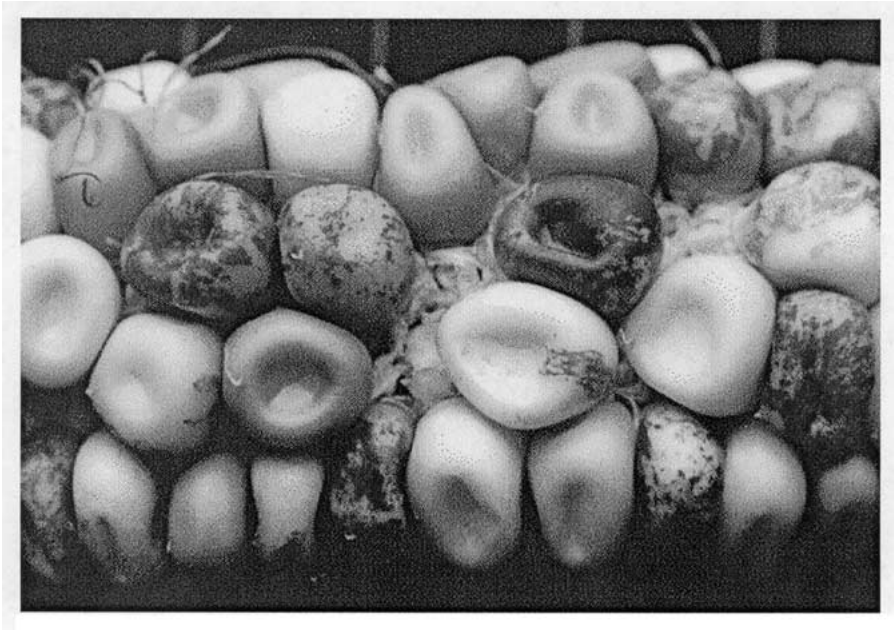

Figure 4. A maize ear illustrating the kind of internally-generated genetic changes studied by McClintock.

6. Where does Genetic Change Really Come From?

The conventional view is that genetic change comes from stochastic, accidental sources: radiation, chemical, or oxidative damage, chemical instabilities in the DNA, or from inevitable errors in the replication process. However, the fact is that DNA proofreading and repair systems are remarkably effective at removing these non-biological sources of mutation. For example, consider that the *E. coli* cell replicates its 4.6 megabase genome every 40 minutes. That is a replication frequency of almost 2 kHz. Yet, due to the action of error-recognition and correction systems in the replication machine and in the cell to catch mistakes in already-replicated DNA, the error rate is reduced below one mistake in every 10^{10} base-pairs duplicated and a similar low value is observed in mammalian cells [32]. That is less than one base change in every 2000 cells, certainly well below the mutation frequencies I have measured in *E. coli* of about four mutations per every 100 to 1000 cells.

In addition to proofreading systems, cells have a wide variety of repair systems to prevent or correct DNA damage from agents that include superoxides, alkylating chemicals and irradiation [33]. Some of these repair systems encode mutator DNA polymerases which are clearly the source of DNA damage-induced mutations and also appear to be the source of so-called 'spontaneous' mutations that appear in the absence of an obvious source of DNA damage [34].

Results illustrating the effectiveness of cellular systems for genome repair and the essential role of enzymes in mutagenesis emphasize the importance of McClintock's revolutionary discovery of internal systems generating genome rearrangements, particularly when an organism has been challenged by a stress affecting genome function (Figure 4) [5].

McClintock recognized that genetic change is a cellular process, subject to regulation and is not dependent on stochastic accidents. The idea of internally-generated, biologically regulated mutation has profound impacts for thinking about the process of evolution. Darwin himself acknowledged this point in later editions of *Origin of Species*, where he wrote about natural 'sports' or '. . . variations which seem to us in our ignorance to arise spontaneously. It appears that I formerly underrated the frequency and value of these latter forms of variation, as leading to permanent modifications of structure independently of natural selection.' (6th edition, Chapter XV, p. 395).

To see the real-world evolutionary importance of built-in biological mechanisms of genetic change, we have only to consider the post-WWII emergence of multiple antibiotic resistance in bacteria. This phenomenon represents the largest and best-documented evolutionary experiment in the molecular biology era. Interestingly, when antibiotic use began, we had a robust theory of how resistance would evolve by modification of existing cell components so that they were no longer antibiotic-sensitive. This theory was confirmed by laboratory experiments. Nonetheless, when the basis of naturally evolving multiple antibiotic resistance was determined, the experimentally-confirmed theory was wrong. Resistance resulted from the presence of new biochemical activities in the bacteria, encoded by new transmissible genetic systems that could accumulate additional DNA encoding these resistance activities [35].

7. Natural Genetic Engineering

7.1. MOBILE GENETIC ELEMENTS

As might be expected from the patterns of genetic reorganization observed in whole genome databases, living cells possess a variety of biochemical systems capable of creating novel DNA structures. Sometimes these systems are individual proteins and create localized changes, such as the mutator DNA polymerases mentioned above, or the site-specific recombinases that join together rare recognition sites (site-specific recombination; see refs. in [9, 35]). Sometimes, much larger multiprotein assemblages are involved, like the apparatus for carrying out homologous genetic recombination or for repairing severed DNA molecules by non-homologous joining of broken ends [36]. Among the most important systems are those called 'transposable elements' (TEs) [7, 8], which make up about 43% of the human genome [21]. These TEs include the mobile 'controlling elements' discovered by McClintock and they comprise integrated systems of proteins and nucleic acids that interact to mobilize DNA to new locations in the genome.

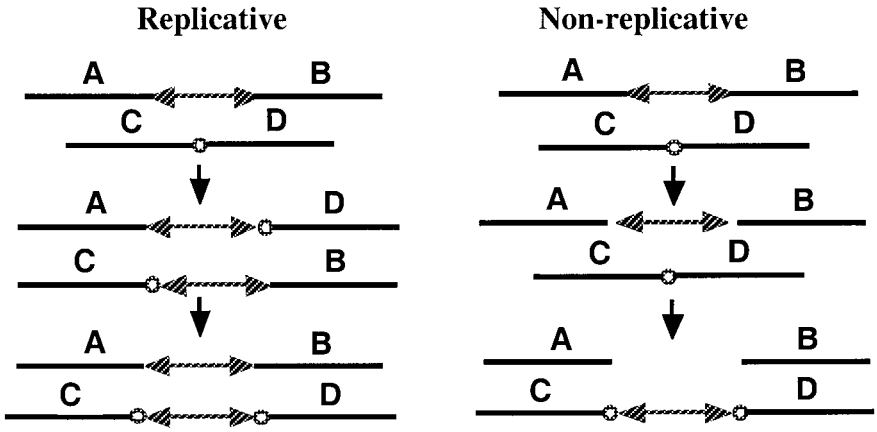

Figure 5. Replicative and non-replicative mechanisms of DNA transposition [37]. The double-headed arrows indicate a DNA transposon. The small circles indicate a short sequence at the target site; the duplication of this target sequence is a halfmark of almost all transposition events [37, 38].

One broad class of TEs are those that act on the genome purely at the DNA level. The TEs that mobilize antibiotic resistance determinants are DNA-based elements celled 'transposons,' and this name has become general for all TEs that operate at the DNA level. Both prokaryotes and eukaryotes possess transposons that have two basic activities. They can mobilize themselves to new sites in the genome by one of two related mechanisms (Figure 5, [37]), and they can rearrange adjacent segments of the genome in a variety of characteristic ways (Figure 6) [38]. What is of basic importance here is the fact that transposons can make any segment of the genome mobile and thus capable of duplication and that they can generate precisely the type of large-scale rearrangements observed in sequenced genomes. There are about 300,000 DNA transposons in the human genome (3% of the genome).

In eukaryotes, there are additional classes of TEs based on the reverse transcription of RNA into DNA copies which can then be inserted into new genomic locations. These elements are called 'retrotransposons,' and they come in two basic varieties [39]. One variety includes retroviruses, such as HIV, whose DNA proviruses can move from the genome of one cell to that of another via infectious particles, and non-infectious retroviral-like elements whose DNA can move from one site to another within the same nucleus. The integrated DNA structures of this retroviral class of retrotransposons are characterized by long terminal repeat or LTR structures at each end. The human genome contains about 450,000 LTR retrotransposons (8% of total DNA). A second class of retrotransposons consists of elements lacking terminal repeats but having a polyA tail at one end, like a synthetic cDNA. These polyA-tailed TEs fall into two major classes: LINEs, or Long Interspersed Nucleotide Elements and SINES, or Short Interspersed Nucle-

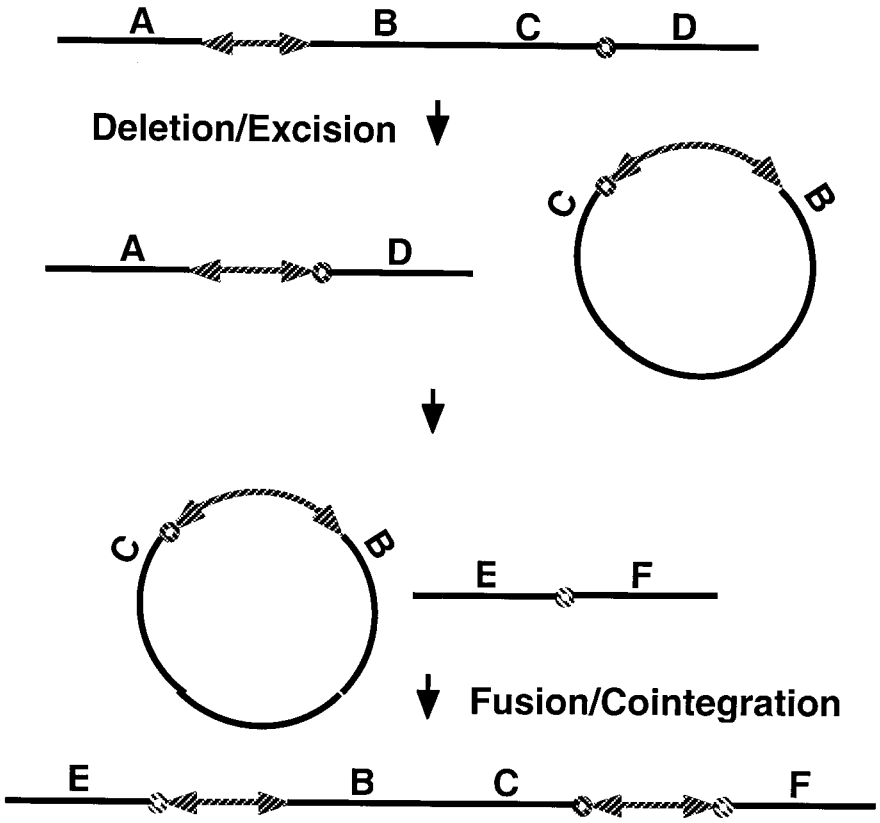

Figure 6. Transposition of a chromosomal segment to a new site in the genome mediated by a replicative transposon (see ref. [38] for details).

otide Elements. There are, respectively, 850,000 and 1,500,000 LINEs and SINEs in the human genome (representing 21% and 13% of total DNA).

Like LTR retrotransposons, LINE elements encode their own reverse transcriptase and integration activities. Thus, once the original retrotransposon has been transcribed, both classes can copy the RNA into DNA and reinsert the DNA copy at a new location. The SINE elements do not encode reverse transcriptase activity and appear to depend upon LINEs for the copying and chromosomal integration of their transcripts. It is quite significant that the unusually abundant SINE elements are often highly specific for particular taxonomic groups. For example, a number of mammalian orders (primates, rodents, artiodactyls, etc.) share LINE and some less abundant SINE elements, but the most abundant SINEs in each order's genome are limited to that order [40]. Thus, a primate cell can be distinguished from a rodent cell simply by examining the SINEs and genetic changes involving these taxonomically-limited SINEs are unique to the group which possesses them.

Retrotransposons are powerful agents of genome restructuring. Because they are so abundant, they can mediate large scale chromosome restructurings by means of homologous recombination between similar elements at distant locations. However, they have other unique capabilities. The LTRs of the retroviral class contain potent signals for initiating and terminating transcription. Thus, their insertion near or within a genetic locus can place it under novel transcriptional controls, as happens when leukemia viruses activate cellular oncogenes. Some SINE elements also have transcription factor binding sites in their sequences and can place adjacent coding sequences under specific new kinds of regulation [41, 42]. Moreover, insertion of SINEs into introns can create novel exons and add sequence motifs to pre-existing proteins [43]. In some cases, the coding sequences for particular proteins are largely derived from SINEs and encode functional molecules that can be found only in the originating taxon [43].

LINEs are master genetic engineers and it has been argued that they are the major force in structuring mammalian genomes [44]. Often LINE transcription does not terminate at its polyA tail but reads through into adjacent DNA; upon reverse transcription and integration into an intron, this adjacent sequence material can form part of a new regulatory region or a coding exon. In fact, LINE-mediated exon shuffling to create new multidomain proteins has been experimentally demonstrated [45]. In addition, LINE element activities can reverse transcribe and insert cellular mRNAs into the genome, creating extra intron-free copies of a coding sequence. This process is the source of 'processed pseudogenes' in the genome and apparently it also has played a major role in the amplification of olfactory receptor proteins that are major adaptive inventions of mammals [46].

7.2. NON-RANDOMNESS AND REGULATION OF NATURAL GENETIC ENGINEERING ACTIVITIES

The foregoing discussion and an extensive literature that cannot be cited here make it clear that TEs and other natural genetic engineering functions have the capacity to reorganize genomes in just the ways needed to reformat modular genome system architectures. This point is increasingly recognized [21, 22]. However, the degree to which these genome reorganization activities are not random is poorly appreciated. Non-randomness is evident at three levels: mechanism, timing and sites of action.

Mechanistically, we can appreciate the very characteristic ways that TEs rearrange DNA from schemes like Figure 6. The ability to create large segments flanked by copies of the TE is built into the process of replicative transposition [38]. Another kind of non-randomness occurs when a proretrovirus inserts in a new genome location. In that case, a whole package of transcriptional and other regulatory signals in the LTRs has been placed next to new sequences, creating the potential for a new genomic system. The action of LTR-containing retrotransposons in yeast (and also certain transposons in bacteria) as mobile activator elements illustrates the functional utility of such 'package deal' rearrangements [7, 9].

As far as timing of natural genetic engineering is concerned, McClintock emphasized the importance of stress events she called 'genome shocks' for activating the built-in systems of DNA rearrangement. Now that these systems have been investigated at the molecular level, we have many examples illustrating how natural genetic engineering can be kept latent during normal proliferation but specifically activated in response to particular signals (see 9, 10, 16 for specific references not given below).

- In repair responses, we know that DNA damage triggers the activation of mutator polymerases and non-homologous end joining activities [34, 36].
- In specialized DNA rearrangement systems, like the ones used in our immune cells to generate an enormous variety of coding sequences for antibodies and T-cell receptors, the necessary genetic engineering functions turn on in response to developmental controls.
- Similarly, a yeast retrotransposon undergoes transcription, reverse transcription and integration in response to mating pheromone.
- In bacteria, the phenomenon of 'adaptive mutation' occurs when cells activate TEs and mutator polymerases in response to long-term starvation signals [16, 47].
- A particularly important source of rapidly-activated natural genetic engineering called 'hybrid dysgenesis' takes place when individuals mate with individuals from a distinct interbreeding group or species [7]. Hybrid dysgenesis results in extraordinarily high rates of mutation and chromosome rearrangements caused by DNA transposons and LINE elements in fruit flies, DNA transposons in nematode worms and retroviruses in mice and wallabies [48].

These examples make it clear that natural genetic engineering occurs episodically and non-randomly in response to stress events that range from DNA damage to the inability to find a suitable mating partner. One important consequence of such episodic activation is that multiple connected genetic changes can occur at different genome locations within a brief period of time. Studies of hybrid dysgenesis in the fruit fly [49] have documented such temporally coordinated changes within a single cell cycle during the mitotic development of the germ line. Since these multiple changes occur several cell divisions before gametes are formed, multiple sperm or eggs (and, consequently, multiple individuals) can be produced which share a constellation of related genome alterations.

In addition to temporal specificity, it turns out that many natural genetic engineering functions show intriguing degrees of selectivity in where they act within the genome. This selectivity appears to be chiefly related to interactions between natural genetic engineering systems and the cellular systems controlling transcription and chromatin formatting. The examples we have of target selection include the action of localized point mutagenesis, retrotransposons and DNA transposons (see 9, 10 for specific references):

- Somatic hypermutation restricted to the antigen-binding domains of antibody coding sequences.
- 50–75% preference for insertion of yeast retrotransposons Tyl-4 upstream of RNA polymerase III transcription start sites; with Ty3, a direct interaction has been demonstrated between the integration complex and PoIIII transcription factors.
- Preference of Ty1 insertion upstream of RNA polymerase II start sites.
- Preference for Ty5 retrotransposon insertion into transcriptionally silenced regions of the yeast genome.
- Preference for P factor insertion into the 5' end of transcribed DNA sequences in fruit flies [50].
- Targeting of genetically engineered P factor vectors containing transcription factor binding sites to short chromosome regions known to bind the corresponding factor.
- Specificity of the HeT-A and TART retrotransposons for chromosome ends in fruit flies.

These few cases of targeting for natural genetic engineering may well be the tip of the iceberg. It is likely that many more instances will be discovered when the target specificity of TEs and mutators are investigated systematically. The indication that target selection for natural genetic engineering can interact with the transcriptional control apparatus and chromatin formatting provides a realistic basis for thinking about molecular mechanisms that can target mobile regulatory modules (e.g. LTRs or transposons) to a series of functionally related genetic loci. This kind of molecular targeting would greatly enhance the potential for creating novel adaptive multilocus genome systems in response to an evolutionary crisis.

8. A 21st Century View of Evolution

Evolution is the history of organisms that have succeeded in adapting to changing circumstances. Over evolutionary time, this means altering the genome – the long-term information storage organelle of all living cells – to provide the functional information needed to survive and reproduce in new conditions. Those organisms that have the most flexible computational capabilities, in particular those that have the best means of altering information stored in the genome, will have an advantage. Thus, it makes sense for organisms to possess crisis-responsive natural genetic engineering functions and we should not be surprised to find them ubiquitous in contemporary organisms, all of whom are evolutionary winners. Indeed, it is now difficult to imagine how organisms that depend upon gradual accumulation of stochastic mutations could persist in the evolutionary rat race.

The last half century has taught us an astonishing amount about how living organisms function at the molecular level, in particular about how they execute cellular computations through molecular interactions and about the systemic, modular, computation-ready organization of the genome. We have come to realize

some of the basic design features that govern genome structure. Combining this knowledge with our understanding of how natural genetic engineering operates, it is possible to formulate the outlines of a new 21st Century vision of evolutionary engineering that postulates a more regular principle-based process of change than the gradual random walk of 19th and 20th Century theories. Such a new vision is not all-encompassing because it cannot provide detailed accounts for major events currently beyond the reach of science, such as the origin of cellular life or the mechanisms of endosymbiotic events underlying the emergence of distinct superkingdoms and kingdoms of life [51, 52]. Nonetheless, a 21st Century view of evolution can help us understand how new taxonomic groups have emerged bearing novel complex adaptations.

As I see it, a 21st Century view of evolution has to include the following features:

- Major evolutionary change to the genome occurs by the amplification and rearrangement of pre-existing modules. Old genomic systems are disassembled and new genomic systems are assembled by natural genetic engineering functions that operate via non-random molecular procesces.
- Major alterations in the content and distribution of repetitive DNA elements results in a reformatting of the genome to function in novel ways – without major alterations of protein coding sequences. These reformattings would be particularly important in adaptive radiations within taxonomic groups that use the same basic materials to make a wide variety of morphologically distinct species (e.g. birds and mammals).
- Large-scale genome-wide reorganizations occur rapidly (potentially within a single generation) following activation of natural genetic engineering systems in response to a major evolutionary challenge. The cellular regulation of natural genetic engineering automatically imposes a punctuated tempo on the process of evolutionary change.
- Targeting of natural genetic engineering processes by cellular control networks to particular regions of the genome enhances the probability of generating useful new multi-locus systems. (Exactly how far the computational capacity of cells can influence complex genome rearrangements needs to be investigated. This area also holds promise for powerful new biotechnologies.)
- Natural selection following genome reorganization eliminates the misfits whose new genetic structures are non-functional. In this sense, natural selection plays an essentially negative role, as postulated by many early thinkers about evolution [53]. Once organisms with functional new genomes appear, however, natural selection may play a positive role in fine-tuning novel genetic systems by the kind of micro-evolutionary processes currently studied in the laboratory.

A more speculative feature of a new evolutionary vision is the idea that much of the creative assembly of complex new systems may proceed prior to expression through rearranging components available in the functionally redundant or

'facultative' part of the genome [54]. This kind of 'experimental' natural genetic engineering process may be considered an activity of the R & D sector of the biological information economy [55].

Molecular genetics has amply confirmed McClintock's discovery that living organisms actively reorganize their genomes [5]. It has also supported her view that the genome can 'sense danger' and respond accordingly [56]. The recognition of the fundamentally *biological* nature of genetic change and of cellular potentials for information processing frees our thinking about evolution. In particular, our conceptual formulations are no longer dependent on the operation of stochastic processes. Thus, we can now envision a role for computational inputs and adaptive feedbacks into the evolution of life as a complex system. Indeed, it is possible that we will eventually see such information-processing capabilities as essential to life itself.

References

1. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D.: *Molecular Biology of the Cell*, 3rd ed., Garland, New York, 1994.
2. Gerhart, J. and Kirschner, M.: *Cells, Embryos and Evolution*, Blackwell, London, 1997.
3. Caporale, L.: *Molecular Strategies for Biological Evolution*, New York Acad. Sci., New York, 1999. (published as *Annal. NY Acad. Sci*. **870**).
4. McDonald, J.F.: *Transposable Elements & Genome Evolution*, Kluwer, Dordrecht, 2000.
5. McClintock, B.: *The Discovery and Characterization of Transposable Elements*, Garland, New York, 1987.
6. Bukhari, A.I., Shapiro, J.A. and Adhya, S.L. (eds.): *DNA Insertion Elements, Episomes and Plasmids*. Cold Spring Harbor Press, Cold Spring Harbor, NY, 1977.
7. Shapiro, J.A. (ed.), *Mobile Genetic Elements*, Academic Press, New York, 1983.
8. Berg, D.E. and Howe, M.M. (eds): *Mobile DNA*, American Society for Microbiology Press, Washington, D.C., 1989.
9. Shapiro, J.A.: Natural Genetic Engineering in Evolution, *Genetica* **86** (1992), 99–111.
10. Shapiro, J.A.: Genome System Architecture and Natural Genetic Engineering in Evolution, *Annal. NY Acad. Sci*. **870** (1999), 23–35.
11. Shapiro, J.A.: Transposable Elements as the Key to a 21st Century View of Evolution, *Genetica* **107** (1999), 171–179.
12. Wilmut, I., Schnieke, A.E., McWhir, J., Kind, A.J. and Campbell, K.H.: Viable Offspring derived from Fetal and Adult Mammalian Cells, *Nature* **385** (1997), 810–813.
13. Rideout III, W.M., Eggan, K. and Jaenisch, R.: Nuclear Cloning and Epigenetic Reprogramming of the Genome, *Science* **293** (2001), 1093–1098.
14. Doolittle, R.F.: The Multiplicity of Domains in Proteins, *Annu. Rev. Biochem*. **64** (1995), 287–314.
15. Reznikoff, W.S.: The Lactose Operon-Controlling Elements: A Complex Paradigm, *Mol. Microbiol*. **64** (1992), 2419–2422.
16. Shapiro, J.A.: Genome Organization, Natural Genetic Engineering, and Adaptive Mutation, *Trends in Genetics* **13** (1997), 98–104.
17. Britten, R.J. and Davidson, E.H.: Gene Regulation for Higher Cells: A Theory, *Science* **165** (1969), 349–357.

18. Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M.G. and Alon, U.: Ordering Genes in a Flagella Pathway by Analysis of Expression Kinetics from Living Bacteria, *Science* **292** (2001), 2080–2083.
19. Arnone, M.I. and Davidson, E.H.: The Hardwiring of Development: Organization and Function of Genomic Regulatory Systems, *Development* **124** (1997), 1851–1864.
20. Yuh, C.H., Bolouri, H. and Davidson, E.H.: Genomic *cis*-regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene, *Science* **279** (1998), 1896–1902.
21. International Human Genome Sequencing Consortium: Initial Sequencing and Analysis of the Human Genome, *Nature* **409** (2001), 860–921.
22. Venter, J.C. et al.: The Sequence of the Human Genome, *Science* **291** (2001), 1304–1351.
23. http://web.uvic.cal/~bioweb/people/choy/dlevin/Forensic/index.htm
24. Bell, A.C., West, A.G. and Felsenfeld, G.: Insulators and Boundaries: Versatile Regulatory Elements in the Eukaryotic Genome, *Science* **291** (2001), 447–450.
25. Misteli, T.: Protein Dynamics: Implications for Nuclear Architecture and Gene Expression, *Science* **291** (2001), 843–847.
26. Gilbert, W. Why genes in pieces? *Nature* **271** (1978), 501.
27. Patel, N.H. and Prince, V.E.: Beyond the Hox complex, *Genome Biology* **1** (2000): reviews 1027.1–1027.4 The electronic version of this article is the complete one: http://genomebiology.com/2000/1/5/reviews/1027 .
28. The Arabidopsis Genome Initiative: Analysis of the Genome Sequence of the Flowering Plant *Arabidopsis thaliana*, *Nature* **408** (2000), 796–815.
29. Ohno, S.: *Evolution by Gene Duplication*, Springer-Verlag, New York, 1970.
30. Graham, M.: Cereal Genome Evolution: Pastoral Pursuits with 'Lego' Genomes, *Curr. Op. in Genet. & Devel.* **5** (1995), 717–724.
31. Dehal, P. et al.: Human Chromosome 19 and Related Regions in Mouse: Conservative and Lineage-Specific Evolution, *Science* **293** (2001), 104–111.
32. Kunkel, T.A. and Bebenek, K.: DNA Replication Fidelity, *Annu. Rev. Biochem.* **69** (2000), 497–529.
33. http://www.nih.gov/sigs/dna-rep/whatis.html
34. Goodman, M.F.: Mutagenesis: Purposeful Mutations, *Nature* **395** (1998), 221–223.
35. Shapiro, J.A.: Natural Genetic Engineering, Adaptive Mutation & Bacterial Evolution, In: E. Rosenberg (ed.), *Microbial Ecology and Infectious Disease*, ASM Press, Washington, 1999, pp. 259–275.
36. Haber, J.E.: Partners and Pathways: Repairing a Double-Strand Break, *Trends Genet.* **16** (2000), 259–264.
37. Craig, N.L.: Unity in Transposition Reactions, *Science* **270** (1995), 253–254.
38. Shapiro, J.A.: A Molecular Model for the Transposition and Replication of Bacteriophage Mu and Other Transposable Elements, *Proc. Nat. Acad. Sci. U.S.A.* **76** (1979), 1933–1937.
39. Boeke, J.D. and Corces, V.G.: Transcription and Reverse Transcription of Retrotransposons, *Annu. Rev. Microbiol.* **43** (1989), 403–434.
40. Deiningen P.L.: SINEs: Short Interspersed Repeat DNA Elements in Higher Eucaryotes, In ref. 8, pp. 619–636.
41. Britten, R.J.: Mobile Elements Inserted in the Distant Past have Taken on Important Functions, *Gene* **205** (1997), 177–182.
42. Brosius, J.: RNAs from all Categories Generate Retrosequences that may be Exapted as Novel Genes or Regulatory Elements, *Gene* **238** (1999), 115–134.
43. Nekrutenko, A. and Li, W.-H.: Transposable Elements are Found in a Large Number of Human Protein Coding Regions, *Trends in Genetics* **17** (2001), 619–625.
44. Kazazian, H.H.: L1 Retrotransposons Shape the Mammalian Genome, *Science* **289** (2000), 1152–1153.

45. Moran, J.V., DeBerardinis, R.J. and Kazazian, Jr, H.H.: Exon Shuffling by L1 Retrotransposition, *Science* **283** (1999), 1530–1534.
46. Brosius, J.: Many G-Protein-Coupled Receptors are Encoded by Retrogenes, *Trends in Genetics* **15** (1999), 304–305.
47. McKenzie, G.J., Harris, R.S., Lee, P.L. and Rosenberg, S.M.: The SOS Response Regulates Adaptive Mutation, *Proc. Natl. Acad. Sci. USA* **97** (2000), 6646–6651.
48. O'Neill, R.J., O'Neill, M.J. and Graves, J.A.: Undermethylation Associated with Retroelement Activation and Chromosome Remodelling in an Interspecific Mammalian Hybrid, *Nature* **393** (1998), 68–72.
49. http://www.wisc.edu/genestest/CATG/engels/Pelements/Pt.html#Abstract
50. Spradling, A.C., Stern, D., Kiss, I., Roote, J., Laverty, T. and Rubin, G.M.: Gene Disruptions Using P Transposable Elements, *Proc. Natl. Acad. Sci. USA* **92** (1995), 10824–10830.
51. http://www.geocities.com/jjmohn/endosymbiosis.htm
52. Woese, C.R., Kandler, O. and Wheelis, M.L.: Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria and Eucarya, *Proc. Natl. Acad. Sci. USA* **87** (1990), 4576–4579.
53. De Vries, H.: *The Mutation Theory*, Open Court, Chicago, 1910.
54. Golubovsky, M.D.: personal communication.
55. Katsenelinboigen, A.: *Evolutionary Change: Toward a Systemic Theory of Development and Maldevelopment*, Gordon and Breach, Amsterdam, 1997.
56. McClintock, B.: Significance of Responses of the Genome to Challenge, *Science* **226** (1984), 792–801.