J Mol Evol (2001) 53:539–554
DOI: 10.1007/s002390010244

JOURNAL OF MOLECULAR EVOLUTION

Intrinsically Driven Changes in Gene Interaction Complexity. I. Growth of Regulatory Complexes and Increase in Number of Genes

Emile Zuckerkandl

Institute of Molecular Medical Sciences, P.O. Box 20452, Stanford, CA 94309, USA

Received: 4 January 2001 / Accepted: 20 May 2001

Abstract. A two-step process, previously considered in the literature, and here named coadaptational drive, is deemed to be largely responsible for both increases in the complexity of transcriptional control and increases in the total gene number, along lines of descent leading to more complex organisms. Coadaptational drive consists in a succession of modifications in the interaction among informational macromolecules, namely, structural decay spread by genetic drift and repair spread by selection. Increased genetic complexity, drawing on the opportunities offered by gene duplication, may be considered to be a secondary effect of such processes of decay and repair. The evolution toward higher regulatory complexity is thus considered to be obligatorily founded in part on random genetic drift. Increases in this complexity would represent primarily a trend intrinsic to the internal molecular environment, with the external environment having only to concur. Direct selection of mutations that increase complexity without the intervention of a phase of genetic drift is acknowledged likely to be a significant process as well, but it is claimed that a sequence of events of direct selection cannot be unlimited and will eventually stall, and that the roots of such a sequence ultimately are to be traced to an episode of coadaptational drive. Controller gene diseases, mostly mild, therefore seem to be essential for the evolution of increased biological complexity. The attempt is made to show or to confirm that (i) a conservative force (repair) provides a mechanism for the generation of novelty, (ii) a prominent part of selection, counterpart to Darwinian selection, originates from the internal environment and derives from the mechanics of genomic processes, and (iii) this selection is at times directional and leads to increases in complexity.

Key words: Regulatory networks — Regulomics — Controller node complexity — Compensatory mutations — Environment-responsive amino acid sites — Spatiotemporal relocalizations of function — Phyletic limitations of complexity

We don't know why there are so many different factors regulating the transcription of these genes.

Scott F. Gilbert (1997, p. 441)

Introduction

The present paper and one to follow suggest that evolutionary increases in complexity at the level of the transcriptional control of gene expression result, to an important extent, from a singular combination of neutral drift and natural selection. This combination, which has been examined before, may be referred to by the phrase coadaptational drive and may be considered to represent a major biological process. In coadaptational drive, a mutation introduces a deadaptation in the fit of a functional molecular complex and the defect, though not exactly reversed, is repaired. The element of drive is produced by the unavoidable and ever-recurring coupling of damage and repair, with the end state comparable in fitness to the initial state, though structurally different.

Coadaptational drive consists in the organism's adaptation to random heritable internal change instead of to changes in the external environment. The two adaptational processes, internally and environmentally condi-

Correspondence to: Emile Zuckerkandl; *email:* EmileIMMS@aol.com

tioned, seem to work in opposite directions: in adaptation to internal change, selection tends to *preserve* the effects of interactions among deteriorating informational macromolecules, whereas in adaptation to environmental change, selection tends to *modify* the organism's operations, including those of interacting informational macromolecules (semantides).[1] Coadaptational drive, namely, the organism's functionally *conservative* structural response to mutations in semantides, paradoxically, is likely, at times, to have side effects leading in fact to significant evolutionary change. In particular, coadaptational drive produces increases in the complexity of the gene interaction system, and these increases can be considered as frequently elicited by processes intrinsic to the organisms.

In the present set of papers, it is argued that gene and protein interaction complexity need not result from direct positive selection but, in many cases, may merely be by-products hitchhiking on processes of selection that respond to other effects of mutational change (Zuckerkandl 1997). The conclusions eventually reached involve the term *controller node*, which is the regulatory "kit" (defined below) specifically allocated to each gene. Controller node complexity is expected to increase decisively through a form of coadaptational drive. As explained in due course, increased controller node complexity leads to increased controller node connectivity and thereby increased gene interaction complexity; this results in a fitness-lowering excess in deleterious pleiotropic effects of mutations, warranting the adoption of genetic devices for isolating regulatory subsystems, which in turn is achieved to a significant degree by the evolutionary development of cell types. Coadaptational drive comes under two guises—increase in controller node complexity and duplication of gene interaction pathways. Under both forms, coadaptational drive may be held largely responsible for the observed net increase in the number of functional genes in higher organisms, an increase that occurs despite probable continued gene losses that are part of genome tectonics. Among reduplicated regulatory pathways or components of pathways, each pathway copy is, through differential regulation, given a spatial or temporal assignment restricted in a particular way. This matter must be considered here, though it is given special attention elsewhere. First, the focus is on increases in controller node complexity.

The spontaneity and simplicity of many processes of complexity increase seem to be inscribed in the structure of the gene regulatory system and in its twin tendencies to conserve function and to distribute functions over organismal stages and sites. For this reason of spontaneity and simplicity, the projected evolutionary implications of coadaptational drive may be considered probable, even before quantitative evaluation. The rate of the presumed evolutionary processes is expected to be low, but so is the rate of any complexity increase in living systems.

At all levels, including the organismal level, a measurement of complexity involves the counting of components. Increases in gene and protein interaction complexity depend upon increasing numbers of regulatory factors being shared by increasing numbers of different genes. In higher metazoa, not only the number of factors, but the size of factor complexes is greater than in bacteria and (Mark Ptashne, personal communication, 2001), also, greater than in yeast. In mammals, for instance, the same factors reappear in the regulatory batteries of many genes, for example, the heat shock factor (HSF) and, more impressively, the CREB-binding protein (CBP) (see Latchman 1998; Locker 2001). It is not a simple matter, either experimentally or conceptually, to measure the complexity of gene/protein and protein/protein regulatory interactions (or, in one phrase, gene/protein/protein interactions) in a way that authorizes quantitative and qualitative comparisons among all genes of a given organism and among genes (homologous or not) in even very distantly related organisms.

The Controller Node

To provide a basis for such comparisons, one may define the regulatory kit of each gene—in fact, of each cDNA, taking into account genes with multiple promoters—as the gene's controller node (Zuckerkandl 1979, 1994, 1997). The controller node is a unit module of specific gene/protein/protein and other specific interactions as allocated to each gene. As one of its limits, among protein factors, only those that either directly bind to DNA or directly interact with factors that bind to DNA are included. (A protein may be considered as "directly interacting with a factor that binds to DNA" as long as it is part of a complex of proteins that interacts with a factor that binds to DNA.) The gene's cis-acting regulatory elements are included in the controller node; the coding sequence and its protein (or RNA) product—as targets of all the regulation—are not included. The controller node represents the regulatory machinery as it relates to the expression of this coding sequence, during a particular phase of this expression (e.g., transcription). When the gene product is itself a regulatory factor or cofactor,[2] this product participates in another controller node or in a number of them.

Conceived in this fashion, the regulatory units of individual genes can in principle be handled independently

[1] "Semantide" is a synonym for "informational macromolecule" that was not generally adopted. It is used again here, because some have expressed fondness for the term.

[2] For present general use, one may define as a protein cofactor any protein factor that does not bind DNA directly but is part of a complex that binds to DNA.

or fitted together, with due attention to their multiply overlapping parts, which represent one of the system's most significant features. Left out, in this version of the controller node, are the third codon positions, which are known to affect transcription rate under certain conditions.) Within the area of any one type of control (e.g., transcriptional control), several superimposed levels of control have to be integrated into any one controller node. Such levels are represented by local and sectorial involvement of chromatin, "local" referring to promoters and enhancers and "sectorial" to *cis*-acting elements beyond promoters and enhancers. Increases in *controller node complexity* express increasing counts in the number of specific regulatory components that participate in the regulation of the expression of a single gene, namely, of factors, cofactors, low molecular weight cofactors of proteins, modifying enzymes such as kinases, acetylases, and methylases, and cis-acting DNA and RNA sequences.

It may thus be assumed that the regulatory complexity of individual genes as well as of specified sets of interacting genes can be quantitated approximately, on the basis of a number of conventions. For various purposes, and with the help of future databases and software, particular types of controller node components (e.g., protein factors in the inclusive sense of the word) and types of control (transcriptional, processing, translational, replicational, recombinational) may be assessed independently. For example, processing controller nodes would be those for genes whose protein products interact in different versions of spliceosomes (an ugly word that one might prefer not to have to use).

Any notion of a "complete" controller node would, however, be bound to remain inapplicable, notably because a controller node's limits shift as a sensitive function of developmental and environmental variations. Whether a specific factor participates in the transcriptional (or any other type) controller node of a gene depends on the factor's concentration. Moreover, the participation of a factor in the control of a gene depends on the concentration—and preferential cellular location—of that factor's *active* form. A factor can have more than one active form besides inactive forms. Despite such thickets of complexities, it would not be productive to abandon this line of analysis. Instead, one may hold on to the assumption that, on the basis of a set of conventions, the number of factors and cofactors that intervene typically and effectively in the regulation of a gene will be approximately determinable and the variations in that number will be approximately determinable. In addition, an opening toward lessened ambiguity is offered by the distinction to be made between actual and synthetic virtual controller nodes (Fig. 1).

The controller node concept offers only one way to sort out phenomena of gene regulation, but it offers the opportunity to carry out comparisons among genes of a given organism and among genes (homologous or not) in even very distantly related organisms. Furthermore, the comparisons should lead to classifying genes according to regulatory factor commonalities and differences. Such commonalities of factors imply commonality of control, even when one and the same factor is activating here and repressing there. The same approach would seem to provide an appropriate framework for studying the evolution of these gene interactions and of gene interaction pathways, as well as their developmental transformations. One problem to be tackled in this way is the evolution of gene interaction complexity.

Coregulated groups of genes participating in common functions have been called syntagms by Garcia Bellido (e.g., 1986, 1994, 1997). As I understand syntagms, they are programs of gene action geared to a particular function and characterized by the reuse of the same genes or their similar paralogs in various combinations. The experimental analysis of syntagms in terms of controller node interaction patterns has begun to be approached tangentially under the banner of proteomics (Jeong et al. 2001; Hasty and Collins 2001).

In Their Control of Individual Genes, Why Do Regulatory Complexes Sometimes Tend to Become More Complex?

We focus on transcriptional regulation, the level at which control of gene expression is apparently most diverse. Why, in the course of evolution toward higher organisms,[3] do complexes of transcription factors and cofactors sometimes tend to become larger?

The trend can be given a simple tentative explanation. An already well-explored two-step mutational pathway would, from time to time, spontaneously lead to complexity increases in the form of increases in number of

[3] The term "higher" as applied to organisms can be criticized on the grounds that it suggests an *advantage* associated with higher complexity, and that the idea of such an advantage is merely the expression of nonscientific anthropomorphic subjectivity. Indeed, although statements about higher complexity certainly do express objective observations, it is suggested here that, in the usual biological sense, there is no pervasive advantage in increasing complexity, namely, no generally recurring increased fitness attributable to complexity per se. However, the concept of higher organisms is legitimate, if science is permitted, in principle, without any a priori restrictions, to deal with all that *is*, however difficult it may be to connect with everything else and to understand. If science is given that license, then higher organisms may be defined as those whose evolution has tended toward developing in all individuals of a species the experience of being "I," and toward generating the correlates of this experience in terms of mental functions of various degrees of sophistication. The existence of this aspect of evolution is not in question; what is in question is whether the aspect is important enough for viewing its absence as a truncation of evolution and of its inherent possibilities. If the aspect is important enough, one may qualify an organism as higher, not because it is us, but because it is.

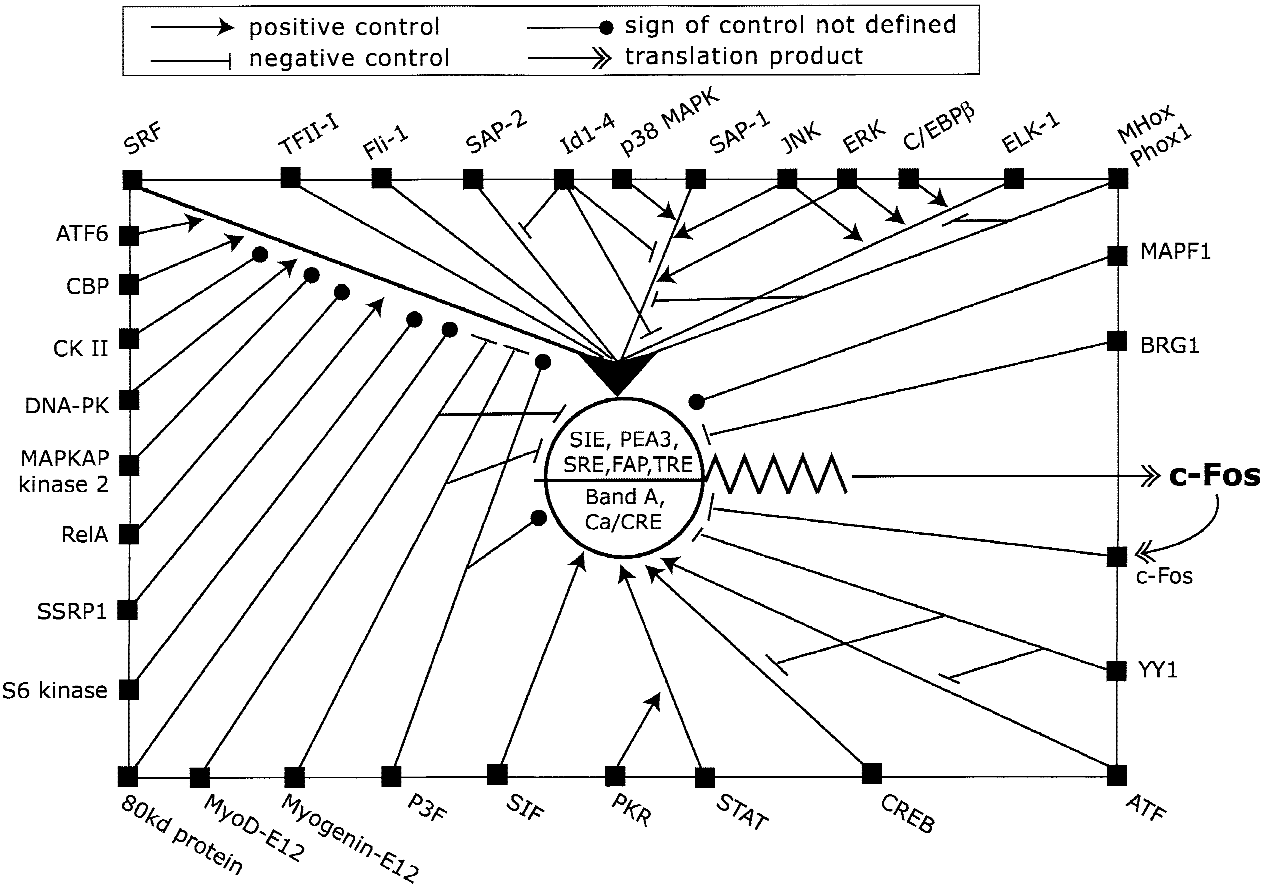

Fig. 1. Synthetic transcriptional controller node for the c-*fos* gene. The representation is limited to factors and cofactors designated at the periphery of the rectangle and *cis*-acting DNA elements listed in no particular order within the central circle. The line traversing this circle stands for the c-*fos* promoter, its zigzag extension, for the coding sequence (which is not considered part of the controller node, c.n.). c-*fos* is not only the c.n.'s gene product, but also one of the factors acting negatively on the c-*fos* promoter. The list of c.n. components is incomplete. Lines that converge at the large arrowhead pointed toward the promoter indicate that the factors to which they connect partake in complexes forming at the SRF-binding element (serum response factor element, SRE) or include neighboring DNA. It is not implied that these factors all interact simultaneously. Synthetic, virtual c.n.s such as this need to be distinguished from actual ones, and are in fact easier to determine. Actual c.n.s are those realized in particular cell types and under particular circumstances. For actual c.n.s, the availability of factors, their quantity, molecular state, access to DNA, and often order of appearance are paramount. On the other hand, a network of virtual c.n.s can in principle be drawn in silico as a synthetic gene interaction blueprint for organisms. It transcends what is actually happening in any cell type, at any particular time, and in any particular organismal location. As soon as a factor/factor/gene interaction is noted as specific and effective in the control of a gene in any cell type, at any time, in any location, under a normal range of environmental conditions, and is judged to be a normal phenomenon, the interaction represents an additional component of the synthetic, virtual c.n. interaction chart. The chart will reflect the circuitry potential of the regulatory system as a whole. It will presumably prove to be species-specific, provide a basis for evolutionary comparisons, and represent an aspect of gene interaction complexity. The sources for the figure, in addition to those indicated for Fig. 1 in Zuckerkandl 1994, are: *CBP* (Ramirez et al. 1997), *STAT* (Wehinger et al. 1996), *Band A* (Omoike et al. 1999), *SSRP1* (Spencer et al. 1999), *BRG1* (Murphy et al. 1999), *ERK* (Frost et al. 1997), *JNK* (Janknecht and Hunter 1997; Lee et al. 1998), *SAP-2* (Price et al. 1995), *Fli-1* (Dalgleish and Sharrocks 2000), *C/EBP*β (Hanlon et al. 2000; Sealy et al. 1997), *TFII-I* (Grueneberg et al. 1997), *ATF6* (Zhu et al. 1997), *CKII* (Manak and Prywes 1993), *PKR* (Deb et al. 2001), *80kd protein* (Drewett et al. 2001), *MyoD* (Groisman et al. 1996; Trouche et al. 1993), *Myogenin* (Groisman et al. 1996; Trouche et al. 1995), *RelA* (Yang et al. 1999), *S6 kinase* (Rivera et al. 1993), *YY1* (Zhou et al. 1995), *SIF* (Wagner et al. 1990), *Mhox/Phox1* (Simon et al. 1997), *p38 MAPK* (Whitmarsh et al. 1997), *SAP-1* (Mo et al. 1998; Price et al. 1995), *Id1-4* (Yates et al. 1999), *ELK-1* (Ling et al. 1998; Shore and Sharrocks 1994), *DNA-PK* (Liu et al. 1993), *MAPKAP kinase 2* (Heidenreich et al. 1999).

factors per controller node. The first mutation is slightly deleterious and, therefore, can spread by random genetic drift (Ohta and Tachida, 1990)—a condition for making a second mutation likely while preserving the first. This first mutation weakens the interaction between a factor and a DNA element or between a factor and another factor, or does both, thereby interfering to some extent with the expression of the gene. A second mutation then restores the lost strength of the regulatory complex by incorporating a further factor into this complex. The second mutation modifies either the structure or the spatiotemporal control of the additional factor—which was already present in the organism. By virtue of now binding to the regulatory complex, the additional factor restabilizes that complex, despite the persistence of the first mutation in either *cis*-acting DNA or in the coding se-

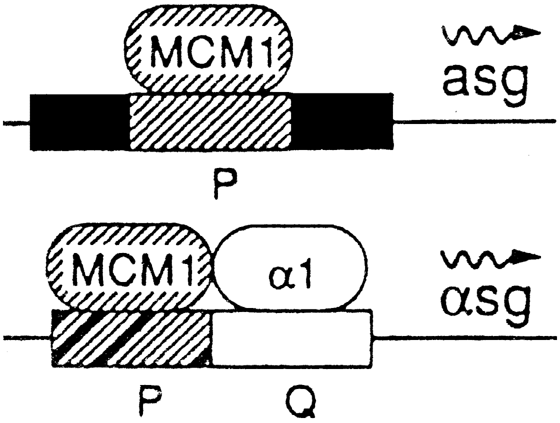

Fig. 2. The binding of an additional factor compensates for a structural alteration in a DNA element. An alteration in DNA element P is compensated by the introduction into the complex of an additional factor, α1. **Top:** Transcription of genes specific for cells of haploid mating type **a** in budding yest (**a**-specific genes; asg) requires the binding of factor MCM1 to DNA element P. **Bottom:** A differently structured P element can still bind MCM1 provided that transcription factor α1, binding to neighboring element Q, helps stabilize the complex, and thereby leads to the transcription of α-specific (αsg) genes that characterize haploid cells of mating type α. [Adapted from Herskowitz (1989).]

quence for one of the original factors (Zuckerkandl 1997, Fig. 1). The second mutation is conceived as positively selected but, at times, might, like the first, spread in the population by genetic drift.

That an additional factor can indeed restabilize such a complex has been shown by an example in yeast, which involves alternative differentiation patterns within a unicellular organism. Herskowitz (1989) found over ten years ago that factor α1 helps factor MCM1 bind to an adulterated P-box (Fig. 2). In such a process, the additional factor α1 increases the complexity of the gene's controller node by one unit.

Turning to multicellular organisms, suppose that a newly introduced factor has a spatiotemporal habitus that differs from what used to be, up to this point, the expression pattern of a given target gene. If there is such a difference in regulatory habitus, then the additional factor, if its spatiotemporal coordinates are indeed restricted, will in turn restrict in its image the spatiotemporal coordinates of the target gene. For example, if the added factor is transcribed and translated during only one stage of development, say, the fetal stage in a mammal, the target gene's transcription will henceforth be limited to the fetal stage.[4]

One example of a switch of a gene active in the adult to a fetal gene might be offered by the fetal non-α-hemoglobin chain gene of cattle. A phylogenetic tree (Fig. 3) showed the fetal chain of cattle probably to be derived from an adult chain (the reverse would be less likely) (Zuckerkandl and Pauling 1965). (This seems to be the first molecular phylogenetic tree ever published. It was based on the assumption of an evolutionary molecular clock, with no corrections introduced besides averaging.) The inference regarding the descent of the contemporary fetal cattle non-α-globin chain is based not only on the number of sequence differences among the chains, but also on the fact that, in the ancestry of the adult and fetal cattle chains, an amino acid *site* was lost (or not added) (Babin et al. 1967), a circumstance that confines the cattle fetal chain to the adult cattle non-α-chain line of descent. In humans—but we do not know the situation in cattle in this regard—the expression of the fetal non-α-globin gene, the γ gene, requires the nuclear factor NF-E4, complexed with the ubiquitous transcription factor CP2 as part of a complex called the stage selector protein (SSP) (Zhou et al. 2000). NF-E4 is a tissue-restricted component of SSP, namely, restricted to the erythroid cell line, and is essential for the transcription of the human γ chain. Thus, the factor NF-E4, at least in humans, would seem to be an additional factor whose action led to the expression of a non-α-chain paralog during fetal life.[5]

Because the evolution of high-order regulatory complexes has not been studied extensively, and because establishing the succession of events is especially difficult for very ancient evolution, available probable examples of the two-step process described may be few at present.

In a plausible model analyzed by Ohta (1987), the nearly-neutral mutation representing the first step of the process is likely to go to fixation when the effective population size is small and when environmental diversity is large. Fixation may also occur when the effective population size is relatively large, because "the larger the population size, the more heterogeneous the environment" is expected to be.

More than two steps are, by the way, involved in the "two-step" process described if there is, initially, more than one slightly deleterious mutation spreading by genetic drift. In any event, a protein preexisting in the organism adapts structurally or *by a switch in spatiotemporal or merely quantitative regulation* in a way such as to be able to restabilize the destabilized assemblage of factors and DNA. The process thus is part of a built-in coadaptational drive, a permanent trend of mutations to modify the mutual structural fit among interacting semantides, while maintaining the necessary level of fit. In

[4] On page S4 of Zuckerkandl (1997), an uncorrected typo changes a rather critical statement into its opposite. The sentence ". . . the full activity of the duplicate gene would not be limited to the embryo" should read ". . . the full activity of the duplicate gene would *now* be limited to the embryo."

[5] The expression of the fetal gene cannot be enforced by the SSP complex at late developmental stages of the erythroid line, a time when higher-order chromatin structure around the fetal gene precludes any effective interaction with factors (Zhou et al. 2000).

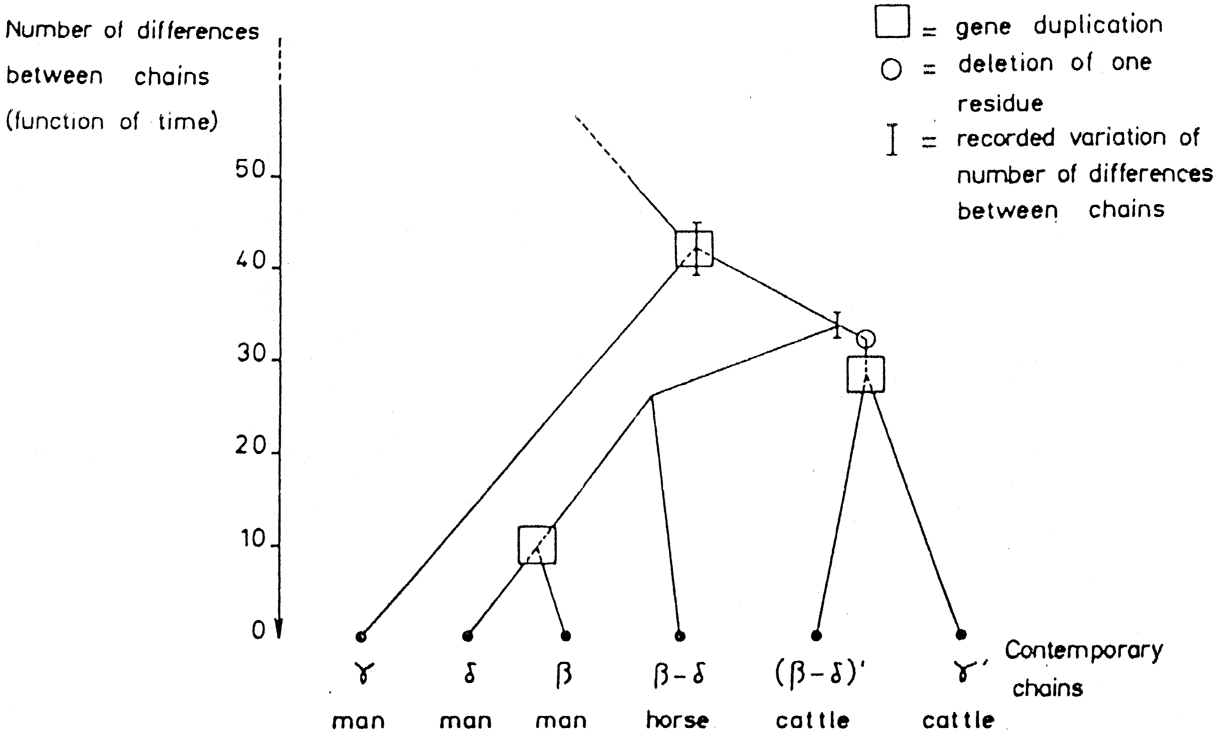

Fig. 3. The bovine fetal non-α chain is more closely related to the bovine adult non-α chain than to the human fetal non-α chain. [From Zuckerkandl and Pauling (1965).]

the spirit of Whyte (1965), this drive is taken here to represent a major factor in evolution.

The two-step mutation mechanism outlined led to alternative differentiation patterns within the same cell or to a temporal shift in gene expression. The same evolutionary scheme could very well apply also to a *spatial* shift in gene expression, particularly to a change in the distribution over cell types of this expression; and I am so bold as to say that, if the scheme can so apply, it will—especially in view of the almost-unlimited resourcefulness of the regulatory system (Latchman 1998; Locker 2001). When a gene has been active in a number of cell types and undergoes duplication, the duplicate may become more cell type specific through two mutations. The first impairs the expression of the duplicate. The impairment will, of course, be limited to the cell types in which the duplicate is expressed. It may be expressed in a few cell types or in a number, if not in all. A second mutation then restores the gene's normal expression rate through the intervention of an additional factor. However, the synthesis of this factor may be specific to, or sufficient in, only one particular cell type or in a few. Suppose that this specificity applies. If, subsequently, the duplicate gene diverges functionally, or if it has already so diverged, the new function, or the new variety of the old function, will be limited to that cell type or those few cell types (Fig. 4).

Moderate damage done by regulatory mutations may be mitigated mostly through other pathways. However, when an appropriate additional factor, one with the appropriate temporal or cell-type specificity, is not found, two such factors might collaborate in creating among themselves the favorable regulatory specificity. This they could do by an overlap in their spatiotemporal ranges. Indeed, the active species of many factors are present in several cell types and at several developmental stages, in an overlapping distribution. If such factors are integrated into the regulatory complex (most likely successively), a narrow spatiotemporal overlap in their expression could limit the expression of the gene considered to a particular developmental phase or cell type (Fig. 5).

Such considerations provide a possible clue as to mechanism of the evolutionary growth of multifactorial protein complexes that bind to DNA. The complexes may be conjectured to be, to a significant extent, remnants of a succession of mutational events, of which the first are slightly deleterious and nearly neutral, and the second advantageous—the lasting legacy of repeated processes of regulatory *restorations,* a legacy with potential implications for future evolution. Regulatory restorations may use a number of other pathways as well, no doubt predominantly; but this one is open, and therefore it will, at times, be used. Progressive evolution is made of comparatively rare events.

The probable impact of regulatory restorations emphasizes the important role that molecular diseases presumably played in evolution (Zuckerkandl and Pauling 1962) or, rather, in the case of gene regulatory systems, the role that *controller gene diseases* (Zuckerkandl 1964) presumably played in evolutionary increases in biologi-

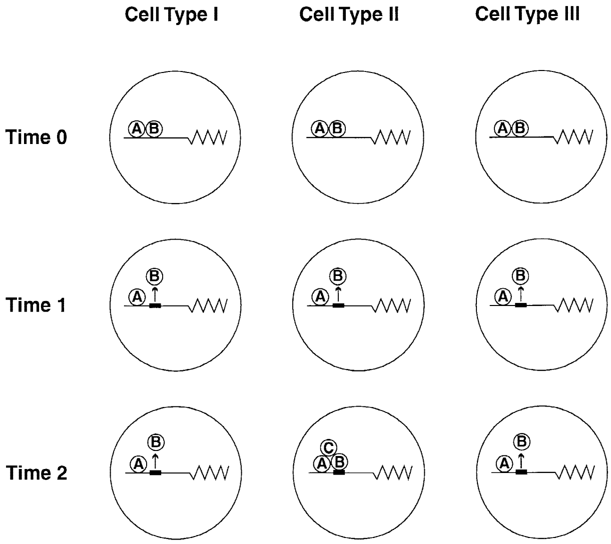

Fig. 4. An additional factor restricts to a single cell type the activity of a gene originally expressed in several cell types. Three cell types are considered, at three evolutionary times. Time 0 shows the wild type. Two transcriptional factors A and B bind to cis-acting DNA and to one another. The *zigzag line* represents a coding sequence. At time 1 a mutation occurs (either in a DNA element, as represented, or in the genes for one of the factors) that results in a slightly deleterious decrease in the stability of the DNA/factor A/factor B complex. At time 2, a mutation causes an additional factor C to restabilize the complex. Assume that such a mutation affecting factor C in either its structure or its control is effective in only one cell type, cell type II in the figure, because, say, factor C is produced in sufficient amounts in that cell type only; then the gene subject to control will henceforth be normally expressed in cell type II only, instead of being normally expressed as heretofore in all three cell types. The differential between the cell types thus introduced may be further increased during subsequent evolution. Even as it is being restricted to a particular cell type, increased complexity, here, is hitchhiking on selection for adequate gene regulation.

cal complexity. Molecular diseases may be considered to be those that result from alterations in the structure of proteins, whereas controller gene diseases result from changes in quantity of proteins (of their active species) without changes in their structure (strictly speaking, from changes in the quantity or cellular distribution of the active species of regulatory protein, from structural changes in those proteins, and from mutations in cis-acting DNA elements). In its medical dimension, regulomics will deal with controller gene diseases.[6]

Theoretical Antecedents of the Two-Step Mutational System

It had been proposed (an unpopular proposal at the time) that selection coefficients be considered for individual amino acid sites (Zuckerkandl 1976a). It had been pointed out that any amino acid substitution at any site—be the substitution selected for or practically neutral—has a chance of having negative pleiotropic effects at other sites of the same protein molecule or on the interaction among protein molecules. (Polynucleotides are to be included as partners.) It had been further assumed that these negative effects are often small enough to be compatible with the fixation of the mutation by random genetic drift. An accumulation of such fitness-lowering mutations characterized by very small negative selection coefficients would eventually lead to the positive selection of a compensatory mutation, so that a system of interacting semantides would evolve continuously by selection, in addition to drift, without actually changing in function (Zuckerkandl 1976a, a paper whose apparently paradoxical aspect is reflected in a definitely paradoxical title; Zuckerkandl 1987, p.41). Ohta (1973, 1988, 1997) analyzed the case of evolving protein molecules in which a second mutation compensates for a slightly deleterious mutation. She stated (1988), “Compensatory evolution proceeds through an intermediate deleterious state.” Selection at individual amino acid sites has since been dealt with quantitatively, most notably, recently, by Susuki and Gojobori (1999).

Hartl and Taubes (1996) pointed out, like Zuckerkandl (1976a) and Ohta (1988, 1997), that “in the long run, the deterioration of fitness resulting from the accumulation of [slightly] detrimental mutations must be balanced by the fixation of compensatory mutations.” Hartl and Taubes’ statement, “In this sense, the continuation of natural selection at the molecular level depends on random genetic drift,” very well summarizes part of the argument of Zuckerkandl (1976a). So does the statement, “On this treadmill, each molecule evolves adaptively but does not improve.” [The phrase “fitness treadmill” is indeed telling (Zuckerkandl 1978, 1986, 1992).] Hartl and Taubes (1996) further state, “. . . On a long enough time scale, most genes undergo selectively driven nucleotide substitutions, though not owing to adaptations to external conditions but rather to compensation for deleterious mutations previously incorporated into the gene” (see also Zuckerkandl 1978, 1987).

The two-step mutation system discussed here fits in with this theoretical position. In concordance with it, Dean et al. (1988) found that most newly arising amino acid replacements in the β-galactosidase enzyme in *E. coli* have rather small effects on fitness. More generally (cf. Eanes 1999), many random amino acid changes generated in the laboratory have very small effects on structural stability and function, and many amino acid polymorphisms occurring in nature may be only slightly

[6] The regulome encompasses the totality of specific molecular interactions that determine gene expression in an organism and includes the topological (circuitry) characteristics of the interaction networks as well as the quantitative aspects of the relations among their components. Regulomics deals with the regulomes of all organisms. It brings to light and systematically records in databases the modules of gene control and their cognate gene/protein and protein/protein interactions (with RNA participating when applicable). It explores the programs of gene deployment that underpin the interaction networks as well as the chromosomal, nuclear, and cellular correlates of the interaction programs.

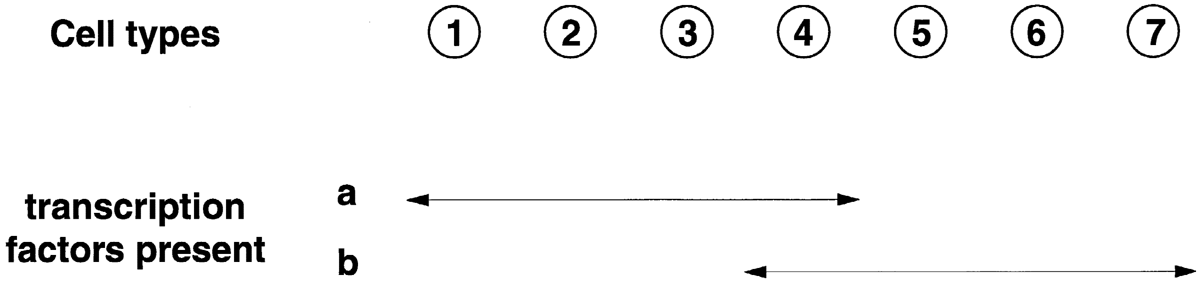

Fig. 5. Confinement of the expression of a gene to a cell type through the action of two factors that differ yet overlap in their distribution over cell types. Cell types 1 to 7 are considered. The *double arrows* indicate in which cell types, respectively, the factors a and b are expressed. The two double arrows overlap in the case of cell type 4, which is thus the only one to express both factors, permitting a particular gene to be activated. (Many variants of this situation are of course possible; e.g., the resulting cell type specificity of the target gene could be broader.)

deleterious. These findings encourage the assumption that, within the gene regulatory system also—in fact, especially within this system, given an expected buffering capacity of interacting controller nodes—in higher as well as lower organisms, a substantive fraction of the mutations will not be strongly deleterious and, therefore, can spread in many populations by genetic drift. On the other hand, compensatory adaptative mutations are also frequent in *E. coli,* in support of the model (Moore et al. 2000).

Coadaptational Drive and Gene Duplication

When an additional factor is incorporated into the transcription complex, the particular version of the complex may often be temporally or spatially restricted; but doesn't the organism need its original gene to continue to function within its previously often wider, and at any rate different, space–time framework?

We may suppose that, much of the time, the answer is yes. In fact, though, the number of cell types in which the original gene was active will presumably tend to shrink as well. The (expressionwise) widely distributed old paralog may well share in the fate of new paralogs. In organisms of increasing complexity, any paralogs will tend to receive assignments of various breaths in their distribution over different developmental phases, periods of the cell cycle, location in the cell, or cell types.

Gene duplication permits the organism to eat its cake and have it too. A major benefit of gene duplication is indeed that the regulatory mode of an original gene may be maintained, and that of its duplicate modified [as apparently first suggested by Zuckerkandl and Pauling (1962), Zuckerkandl (1963), and Ohno (1970)]. One of two similar paralogs would probably have been lost had not the expression of the other become limited to a particular spatial or temporal location. Failing to lose paralogs (because of coadaptational drive) obviously amounts to increasing the number of genes—one of the possible directions in genome tectonics.

According to traditional views, duplicates identical in both their coding functions and *cis*-regulatory functions would not be included in this complexity count. Nondiverged duplicated or multiplied sequences may contribute to *quantitative* features of gene expression, but, one might think, not to gene interaction complexity. Gene duplicates partake in increases in gene interaction complexity only when they harbor some *functional* divergence, whether on a structural or a regulatory basis. However, the concept of functional divergence probably needs to be widened. A change in the spatiotemporal coordinates of the expression of a gene may imply that it functions in a novel way, which is not distinguishable from saying that it exerts a new function. This may hold frequently, but is rather obvious when regulatory mutations lead to heterochronies in which identical genes interact with other genes that they could not communicate with before, even as other interactions might be discontinued (Zuckerkandl 1983).

Complexity increases in gene interaction networks comprise three basic processes: an increase in the number of regulatory connections among preexisting genes; an increase in the number of genes [arriving on the scene with their equipment of regulatory connections, of which some may thus go back in time as far as the origin of the genes themselves (Zuckerkandl 1983)]; and an increase in spatiotemporal restrictions in the distribution of expressed paralogs.

All three closely correlated processes depend on gene duplication, which is the major avenue to an increase in gene number as protected by purifying selection.[7] The first two processes—number of regulatory connections and number of genes—may be considered to result primarily from coadaptational drive. The third process often results in regulatory beyond spatiotemporal isolation of

[7] Not dealt with here are other foundations for increases in gene number, such as partial gene duplicates that are functional and the reshuffling of protein domains, events that may be less frequent than those ascribable to coadaptational drive.

the expression of paralogs, and may be endorsed by selection *over the long haul* as a response of the system to having to contend with deleterious implications of increases in the number of freely intercommunicating gene interactions. The wider perspectives of selection as it intervenes in the evolutionary duplicate gene deployment will be addressed in the follow-up paper. These include probable selection against deleterious pleiotropic effects that are presumably caused by increases in gene interaction complexity, the consequent selection pressure toward a relative autonomy of subprograms of gene interaction provided by the differentiation of cell types, and the orderly, which implies insulated, execution of such subprograms, as well as, eventually, selection against poor adaptability of organisms (i.e., selection for more adaptable competing organisms).

The differential assignments of paralogs may obviously be ratified by selection *over the short term* when fitness-increasing functional differentiation of paralogs happens to occur faster than gene loss (Petit and Zuckerkandl 1976). Yet, the assignments might also be selected in the absence of functional differentiations in the proteins themselves. The selective advantage of an additional spatiotemporal localization may in itself be only slight and thus, in fact, may become fixed in a species by neutral drift as well as by selection. The point is that the spatiotemporal differentiation of the sites at which a function is exercized may be perceived by the organism as functional differentiation and innovation. The localizations in gene expression can entail (i) a temporal assignment of a paralog in *cell development* or the *cell cycle* and (ii) a spatial assignment of a paralog in *cell types* or *cell compartments.* When producing spatiotemporal shifts in expression, regulatory differences alone can make for functional differences.

In a plausibly selectable version of the process, the widespread expression in the organism of an original paralog has been lost and been replaced by a series of more narrowly defined expression assignments in a number of paralogs, as also proposed by Lynch and Force (2000). In that situation, a premium is put on the conservation of the set of multiple narrower expression patterns. Their narrowing probably tends to be advantageous in the presence of spatiotemporally novel tasks to be performed, generated by complexity itself. Lynch and Force (2000) had already observed that “a common fate of the members of duplicate gene pairs is the partitioning of tissue-specific patterns of expression of the ancestral gene,” and that the expression of gene duplicates is frequently altered with respect to timing and tissue specificity. The points made here are in accord with the authors’ inference that “degenerative mutation may be the predominant mechanism that drives the accumulation of gene duplicates in developmentally complex organisms.” Lynch and Force do not bring restoration processes into the picture.

In the version of coadaptational drive that has been principally discussed here, the restoration process consists in the addition to a regulatory complex of a supplementary factor. A more general version of coadaptational drive must be briefly discussed, lest the case presented remain truncated. Indeed, the temporal and spatial localizations of gene expression are not limited to individual genes, as in the situations evoked, but—and this is to be examined elsewhere—often comprise duplicate gene interaction chains if not even networks.

Such interaction chains or networks can in principle be resolved into components represented each by an interacting pair of semantides, though ternary complexes of semantides might frequently have to be considered as units. In their mutual interactions, the members of each pair, if they are functional, are subject to coadaptational drive. Indeed, the interactions tend to decay through the effect of mutations, the damage may spread in populations by genetic drift, and an episode of selection may eventually restore the partly compromised mututal fit of the macromolecules, often in the form of a new structural configuration. Coadaptation might occur, for instance, between a signaling protein and a cell surface receptor protein, as analyzed by Fryxell (1996). Figure 6 (from Zuckerkandl 1983) represents schematically such events as they must be supposed to develop when the pair of semantides considered is a regulatory factor and a DNA element.

The precise succession of evolutionary phases may be variable, but one plausible way to begin the process, depicted in Fig. 6, is through the duplication (complete with regulatory dependencies) of a target gene s.g.—a target, here, of transcriptional control—followed by the duplication of the gene for one of the factors that regulate the target genes. The eventual result of coadaptational drive, in this case, is that the duplicate target genes—they might still be functionally identical—will soon be controlled by two distinct regulatory factors that are no longer significantly cross-binding to the duplicate promoters. They then no longer will belong to the same controller node. Thus, not only does the genome now contain two target genes instead of one, set for (functionally possibly limited) divergence, and two diverging factors; in addition, the way is paved, as well, for a divergence, and thus “duplication,” of the controller node for the target gene duplicates, which hitherto shared the same.

Pathways no doubt exist whereby controller node complexity and controller node connectivity (the sharing of factors by controller nodes) are increased in the absence of mutational damage as a first step, and in the absence of gene duplication. Namely, an appropriately located DNA element of a target gene can accidentally develop an affinity for a factor that happens to be present in relatively large amounts in a given cell type. The expression of the target gene, perhaps also controlling a

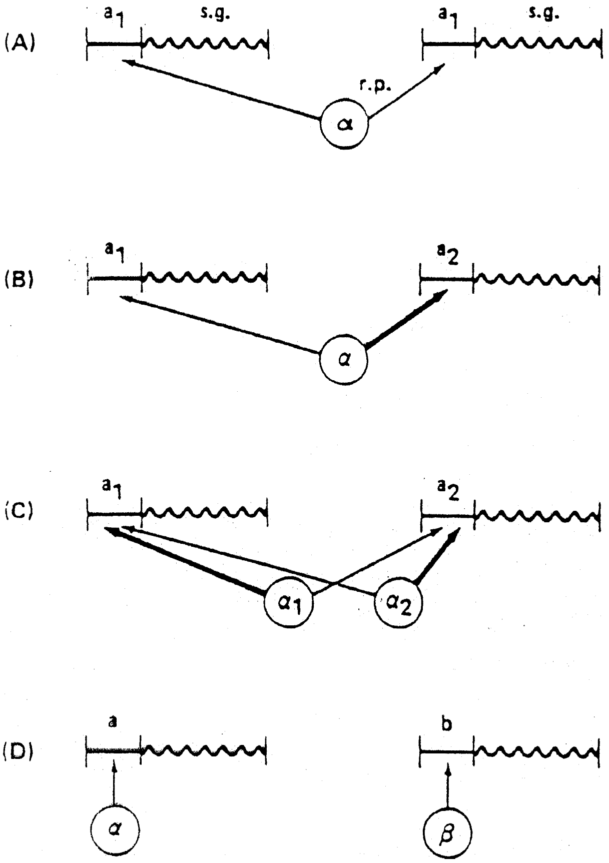

Fig. 6. One of the pathways toward the regulatory segregation of gene duplicates. (A) A structural gene s.g. ("target gene") has duplicated. The coding sequence as well as its *cis*-acting DNA element a_1 are still identical. The same transcription factor α binds to both duplicates. (B) A mutation in one of the *cis*-acting elements can lead to a differential in the affinity of α for the duplicated DNA elements. (C) Now the gene for factor α has also duplicated. The factors controlled by the duplicate factor genes differentiate in regard to their respective affinities for the two *cis*-acting elements of the duplicate target genes. (D) This differentiation progresses to the point where each of the duplicate factors, now labeled α and β, are to be considered as interacting with one of the duplicate target genes only, which can then be said to be independently controlled. (From Zuckerkandl 1983.)

factor, will thereby be increased or decreased. If nothing further happens, the case in point is one of increase in controller node connectivity and gene interaction complexity, at constant gene number. The factor added to the target gene's controller node is indeed not a new paralog. As the evolution of regulatory relationships is further elucidated in its details, examples of such situations should come to light.

Frequently, paralogs that are functionally only slightly diverged tend to be reduced in their level of expression (Stoltzfus 1999), or to be turned into pseudogenes that are not expressed at all (Zuckerkandl and Pauling 1962; Zuckerkandl 1972), or to be bodily eliminated from the genome (Petit and Zuckerkandl 1976), as part of genome tectonics. Coadaptational drive is expected to rescue paralogs from these fates and thereby to increase the number of genes in the genome. Thus, genetic drift would appear to be instrumental in bringing about not only increases in complexity under the guise of increases in controller node complexity and controller node connectivity, but also the drive toward the type of complexity increases represented by increases in gene number.

As an organism's overall complexity increases (a quantity not defined here), the number of regulatory connections among genes is expected to grow faster than the number of genes, since connections can be established in addition to those that each duplicate gene contributes, and since the probability of increases in the number of such connections is expected to be a power function of the number of genes. This might explain why the number of genes does not have to increase greatly when the complexity of the organism increases greatly—at least in some of its parts.

Growth of Regulatory Complexes by Direct Selection

Notwithstanding the likelihood of the two-step process described occurring from time to time, it is to be expected that, in other instances, an additional factor is selected, not for restabilizing a current function, but for adding a new function to the complex. Such growth of regulatory complexes would be adaptive directly and would occur without the help of a prior spreading by genetic drift of slightly deleterious mutations. During the evolution, for example, of the very complex RNA polymerase holoenzyme (Cramer et al. 2000), it may well seem implausible that individual factors that frequently play obviously essential and diverse functional roles have in general arrived on the scene only under the pretext of repairing damage inflicted on the holoenzyme in its previous incarnations.

Roger Kornberg and his associates assess that the polymerase II complex from yeast, counting the different separate but interacting subcomplexes, comprises 57 subunits (mentioned by R. Kornberg in a seminar at Stanford, Feb. 28, 2001), and observe that the subunit composition of polymerase II in humans seems to be virtually identical to that of yeast (Cramer et al. 2000). The subunit compositions of polymerases I and III in HeLa cells show important similarities to that of polymerase II despite the differences (Sentenac et al. 1992; Chedin et al. 1998). There is also a great similarity between yeast and murine Mediator, a complex participating in the polymerase II holoenzyme (Asturias et al. 1999). Such a high degree of conservation indicates that many of the formative stages of these complexes must date back to perhaps even a rather remote ancestor *of* the common ancestor of yeast and human.

In itself such an old age of assemblages of factors and cofactors presents no difficulty for the argument about the growth of regulatory complexes. There can be no doubt that the presumably much more than a billion-year-old phase of evolutionary stability was preceded by a formative phase. The two-step mutational process described would apply to some parts of that formative phase.

Puzzling questions remain, however. If, more than a billion years ago, the progressive accretion of factors and cofactors made it to 57 of them, why did the number not increase since to, say, 62 or decrease to 52? A usual answer would be that the complex grew until it functioned optimally. When it had reached that stage, it quit changing; an improvement in fitness had occurred with the arrival of each participating factor; there was no room for alternating episodes of drift and selection; mutation and selection account for the whole process.

If this process is indeed to be described in classical terms, environmental change is a component not to be left out. Probably the most common interpretation of direct selection of an increase in factor complexity would tend to be that it occurs in response to some environmental change. This view is not plausible, however. Clearly, regulatory complexes as found during a given evolutionary period, and presumably often stable over the long haul, were able to cope effectively with a great variety of environmental changes. In fact, the holoenzyme having remained evolutionarily as stable as it has, environmental change *cannot* have been important in its evolution: As its stability testifies, the holoenzyme protein complex, in various evolutionary eras, must have confronted very diverse environments, without changing significantly in its composition.

Under these conditions, it may be proposed that the motor of direct successive selections of complexity-increasing mutations, *without* intervention of a phase of neutral drift, can run for a time but eventually is expected to stall. This view is based on Ohta's (1973, 1988) paradigm of nearly neutral and of compensatory mutations and on related considerations of Zuckerkandl (1976a, 1987). At the same time it applies the paradigm of Hartl et al. (1985) regarding the exhaustion of possibilities of strong selection after a protein has been honed over long evolutionary periods. It has then reached a stage where further amino acid replacements can provide only a small improvement. When, as often observed as well as calculated, environmental change has little effect on molecular fitness (Hartl et al. 1985)—there is reason to believe that such must be the case of RNA polymerases—the theories of Ohta and of Hartl et al. imply that the improvement offered by favorable mutations would be slight enough that it would spread mostly only through genetic drift. It would seem, then, that for well-established transcriptional complexes, the prospects for further positive selection are bleak in the absence of a preceding propagation of mutational damage by genetic drift. (An objection to this view is discussed below.) In adaptation to a new but constant environment, the rate of adaptive mutations in *populations* of *E. coli* also approaches zero after an initial period of rapid increases in fitness (Lenski and Travisano 1994).

Nevertheless, at each phase of its evolution, the holoenzyme complex would have been competed out by a more elaborate regulatory complex that proved to be more efficient and versatile, had such a more successful complex appeared. In the absence of such competition, the existing complexes proved by their survival through many environments that they were, at the time, "reproductively sufficient" (Zuckerkandl 1992).

If in fact an effective competitor did arrive on the scene at various times to turn smaller holoenzyme complexes into bigger ones, how had *it* conjured up selective forces in its favor, when, according to the demonstration by Hartl et al (1985), in circumstances of achieved functional effectiveness and solidly inferred insensitivity to environmental change, neutral drift rather than selection would most likely intervene? The competitor only displaces the problem. Now *it* has to be explained.

The following explanation may be considered. Conditions for *renewed* direct interventions of positive selection when structure–function relationships are already near-optimal could be created by the generation of new functional opportunities. These would arise through a structural change based on coadaptational drive, in the form of a nearly neutral structural innovation hitchhiking on some particular form of repair process. Once the germ of novel controls of novel molecular interactions has been planted, the fitness treadmill (Zuckerkandl 1978; Hartl and Taubes 1996) can start functioning from a new platform. Coadaptational drive's living symbol, Alice in Wonderland's Red Queen, makes a leap to the side. She had kept running in the same spot, and still does, but, this time, after changing place. The already partly effective rudiment of an additional function has created a "selection vacuum" (Zuckerkandl, 1976a) in the holoenzyme, namely, a situation in which functional effectiveness now is less than optimal. In that situation, structural changes have a chance of being selectable, a chance presumably as great as at any time during the previous history of the molecular complex. The time for positive selection had been more or less over. Yet an "unexpected" selection vacuum appeared, independent of changes in the external environment. Through a particular structural event, attributable to genetic drift or, perhaps, to indirect "hitchhiking" effects of an episode of selection, the system was rejuvenated, as it were, in that opportunities for further selection became available.

Thus, increases in complexity of some assemblages of regulatory factors can be viewed as creating new conditions for further increases in complexity by direct positive selection, at a time when few such opportunities would have seemed to be left. Here, again, the complexity increases would appear to be internally determined by the organism, more than externally conditioned. The drive toward higher organisms would be primarily intrinsic, with the environment having only to concur.

One may, tentatively, project into the past the formation from rudiments, by strings of selected improvements, of contemporary full-grown functions. Yet, even if one postulated a continuous series of episodes of selection in which genetic drift would play little role, that series must be broken somewhere—it must have had an origin. For reasons furnished by the Ohta and Hartl–Dykhuizen–Dean paradigms, one cannot easily place at this origin, once again, an event of direct selection for a new aspect of the complex's regulatory function. The first link in a chain of events of direct selection for new aspects of function would, on the contrary, seem likely to have been mutational damage and the restoration of a complex to its *former* functionality. Indeed, to emphasize it once more, *before* a new structural/functional opportunity had arisen, the previously extant regulatory complex would be improved mostly only via random genetic drift. Thus, if there has been a series of directly selected increases in controller node complexity, random drift still has to be placed at the origin of the series.

If this is so, a form of coadaptational drive is at the origin of even the most sophisticated functional developments, and one might tentatively conclude that the evolution toward higher organisms would never have taken place in the absence of random genetic drift.

A "constructive" evolutionary effect of neutral mutations is particularly well layed out and documented by the novel analyses of Stoltzfus (1999).

Yet, at some point in evolution, the repertory of possible, or sufficiently likely, additional novel kinds of subfunctions to be inserted by additional factors may in fact be exhausted. In support of this concept, Lenski and Travisano (1994), analyzing the results of their adaptation experiment with populations of *E. coli* over 10,000 generations, were forced into giving weight to the view that "the organisms have 'run out of ways' to become much better adapted to their environment." In addition, beyond a threshold of interaction complexity, higher-order factor complexes may be gelled, especially when the system comprises multiple subcomplexes that are interlocked. Factors and subcomplexes of factors are part of an overall orchestration of structural and functional interactions. For example, nucleosomal core histones have to fit in with other histones, histone complexes have to fit in with DNA to organize a chromatin-like structure, and the chromatin structure has to fit in with all steric conditions and functional processes that have used preexisting structures in order to build up, say, transcriptional control. The evolutionary stability of, notably, the ribosome and the nucleosome no doubt have a background of the same kind. A rationale similar in its generality is applicable to the standard genetic code, and a similar type of rationale indeed presumably applies to RNA polymerases (see Hampsey and Reinberg 1999). The enormous structural complexes designated RNA polymerases are only the kernel of a great many more coadapted structures and functions. Before the last common ancestor of yeast and human, too large a number of processes had appearently already evolved and adapted to then existing RNA polymerase complexes for the latter to remain evolutionarily flexible.

For a different, complementary discussion of the evolutionary freezing of individual proteins, a discussion based on the ratio of amino acid residues engaged in specific functions and those available for general functions, see Zuckerkandl (1976b).

It has been reported that "perfect" enzymes such as superoxide dismutase can still be improved (Getzoff et al. 1992; Eanes 1999; Watt and Dean 2000). No surprise here, and no problem for the Hartl/Dykhuizen/Dean paradigm. Space is necessarily left in every protein for functional optimization when the group of realistic optimization criteria is reduced to a single one, such as enzymatic reaction rate. In nature, the "best" protein (and there often surely are quite a few practically equivalent "best" structures) will have accomodated a number of specific trade-offs [e.g., kinetics versus thermal stability in phosphoglucose isomerase (Watt 1995; Watt and Dean 2000)]. No macromolecule can exist that is optimal simultaneously in all respects (Zuckerkandl 1976a).[8]

The overall effects of a changed *external* environment on the evolution of *E. coli* has been measured over 10,000 (at present 28,000!) generations by Lenski and associated groups (Papadopoulos et al. 1999). Effects on the complexity of regulatory systems were not investigated, nor would any be expected to be observed over a short evolutionary period, especially not in bacteria. It should be noted, however, that in this case, as under the Ohta and the Hartl/Dykhuizen/Dean paradigms, after an initial phase of significant adaptation to the new environment, adaptation appears to be essentially completed, and (strongly) positively selected mutations appear to become very rare.

Is the Hartl, Dykhuizen, and Dean model of "evolution to the neutral limit" generally applicable to eukaryotes and diploid organisms? Environment-responsive alleles—and the same should apply to paralogs—do not suggest it, although, as Watt and Dean (2000) have emphasized, the matter depends on the nature of the gene(s) in question and on the organism–environment interactions as they are expressed in these genes. The fact is, different allozymic alleles and combinations of alleles,

[8] It is also possible, of course, that the functional optimum of human superoxide dismutase has not yet been reached, or that it is difficult to realize in nature because of a combinatorial requirement involving a number of amino acid sites, or that it has in the past been "proposed" to natural selection but been passed over because of a lack of significant selective advantage.

say, of phosphoglucose isomerase (Watt 1991), are even nowadays selectively preferred in different environments.

Yet these preferred allozyme alleles and mixtures of allozyme alleles seem to represent a limited set, characterized by limited types of amino acid substitutions at a limited number of molecular sites. One might conceive of these sites as the set of environment-responsive sites (ERSs, to follow the acronym fad). The ERSs may be considered a small group of sites that remained and remain subject to selection by the external environment, even as the the rest of the molecule, already in the rather distant past, approached its structural and functional near-optimum in the Hartl/Dykhuisen/Dean sense, and only fluctuated around it.

It may be deemed arbitrary to carve out two spaces in an enzyme, a large space that has evolved to its "neutral limit" and a small space that remains subject to strong selection as a function of environmental change. Yet such a view may be potentially robust, in that it would not necessarily be contradicted if environmentally adaptive substitutions in a paralog were accompanied by substitutions at a number of molecular sites that are *not* obviously involved in the enzymatic activity. (Such substitutions might be internally coadaptive or practically neutral.) Sequence differences between the two paralogs—say, two isozymes—at amino acid sites other than those whose occupancy is determined by the environment, may very well represent different approaches to the same near-optimum of protein structure. Thus the model of Hartl et al. could indeed be applied to most parts of the isozyme molecules. In those parts, amino acid substitutions would usually spread by genetic drift, although drift might be followed by intramolecular *coadaptational* (or "coadaptative") *selection.* Substantial parts of many proteins presumably always remain open to internally and mutually coadaptive selection, expected mostly to be based on moderate selection coefficients. The partition of proteins into environmentally sensitive and insensitive groups of sites, with the insensitive group conforming to the model of Hartl et al., might be of wide applicability when allele frequencies change as a function of environmental conditions.

Phyletically Limited Complexity Increases

Is there at least a beginning of an explanation why complexity increases occurred along certain lines of evolutionary descent only and not in the descent of most organisms? When all manifestations of coadaptational drive are considered, this drive appears as a frequent and pervasive phenomenon. Indeed, *the fit among coadapted interacting macromolecules is disturbed rather regularly—as regularly as the molecular clock ticks.* This fit will then often be reestablished in a different way, as shown by Fryxell (1996).

On the other hand, it was pointed out, the form of coadaptational drive that leads to the introduction of an additional factor into a controller node is expected to be relatively rare. What might bring about the apparent *grouping* along certain lineages of such rarer events?

Advances toward higher complexities might continue in those lineages in which an event of coadaptational drive happens to have introduced "inadvertently" the rudiments of a new function, or of a novel way to satisfy the requirements of an old function. These rudiments, as mentioned, are then open to being further developed by direct selection. The new function—in morphological terms, let us say, the formation of an efficient wing—may require the build-up of new regulatory connections that produce a new coordinated regulatory subsystem. Such a subsystem would be required for the full development of the new overall function. As the species multiply that have reached this new complexity level in their morphology, physiology, and, foreseeably, sectors of the underlying gene interaction patterns, the chances increase for some of these species to become the venue for an additional cascade of complexity increases. Such cascades may each time be focused on a particular physiological subsystem, with its morphological correlates when applicable. In higher primates, complexity increases in aspects of the central nervous system would be an example.

Thus, once it has been reached, higher complexity *and only higher complexity* would offer opportunities for yet higher complexity to arise. Apparently, steps of this ladder cannot be skipped. This concept would seem to be supported by findings of Richard Lenski and his colleagues in the course of their experiments with "digital organisms," self-replicating, "mutation"-prone computer programs required to "evolve" by solving specific problems. These experiments in silico were reported by Lenski in a seminar at Stanford (May 8, 2001). When "rewards" (additional replication cycles, a limited resource) were meted out for complex functions only, the digital organisms never achieved the complex functions. Simpler functions were an obligatory springboard to the more complex functions.

Hence, perhaps, from time to time, a cascade of complexity increases in particular paraphyletic lines of descent. At each complexity level, the complexity of most taxa stays put; a very few climb another step of the ladder and then produce a number of taxa at this next higher level.

Organisms may drop some of their acquired complexity when, in parasitism and symbiosis, and generally as a function of available energy sources and nutrients, their environment warrants losses of functions. In metazoan taxa, however, complexity mostly remains approximately stable, or in rare cases increases, as it must have done in the ancestors of the Cephalopod molluscs. Results of the intrinsic drive toward increased complexity of genetic systems seem to become "locked in," be it with regard to gene number or with regard to controller node complexity and connectivity. This drive might be actively opposed by selection, as apparently to a large extent in bacteria, or it may not find its appropriate "niche" within types of organisms that are not built to accommodate further functional differentiation. Certain types of organisms might be structurally refractory to further complexity breakthroughs on account of constraints imposed by past evolution, and, within these constraints, have apparently reached their complexity ceiling; or they may be prematurely eliminated by a catastrophe affecting the biosphere (an asteroid or a humanity).

Concluding Remarks

We saw that an obvious potential mechanism for spontaneous increases in gene interaction complexity seems to be inscribed in the very structure of the gene interaction system. Because of this circumstance and of the basic simplicity of the mechanism considered (coadaptational drive under its special as well as general form), the recurrence of such events is to be considered probable, and its evolutionary traces can be observed (Fryxell 1996). The generation of novelty appears to be dependent in critical ways on conservative forces (repair). Thus, a creative and prominent part of selection, counterpart to Darwinian selection, *originates from the internal environment and derives from the mechanics of the genome.*

"Integrons" says Ernst Mayr (1997) (by that is meant levels of biological integration, namely, a certain type of biological unit formed by certain types of subunits) "evolve through natural selection, and at every level they are adapted systems, because they contribute to the fitness of an individual." Yet simpler ancestors, present on earth for many millions of years, must be considered as having no doubt been fit enough. In the process described here of mutational "cures" for controller gene diseases (some such diseases so slight that they spread by genetic drift), what contributes to fitness is not the newly increased complexity, but (as proximate cause) the restoration or preservation of proper gene regulation. The establishment of a higher level of hierarchical integration likely never gets to be selected as such—never passes through the sieve of fitnesses—because segregating additional hierarchical levels of organization is gradual, occurs over long periods of evolutionary time, and never "happens" at any particular moment. Additional hierarchical levels are presumably invisible to positive selection, since selection of fitter organisms operates within much smaller slices of time.

The establishment of additional levels of hierarchical organization may thus be considered incidental and adventitious in terms of the proposed principal molecular mechanism that gives rise to them. At the same time, the spontaneous generation of these organizational levels had been latent in the established biological order—foreshadowed in the constraints and cryptic opportunities defined by the organization already acquired and by the general *types* of tolerable environments (gravity, water, atmosphere, temperature, predators, preys, mates, etc.), such as are in part present as conditions of the evolution of life, in part generated secondarily by life itself, wherever local environmental parameters are conducive to the development of living systems.

Each new spontaneously evolved level of hierarchical organization provides a new field of action for the "fitness treadmill," one absent from earlier fitness games. Organizational spontaneity along an intrinsically and extrinsically limited number of permissible pathways thus would seem to be at the root of evolution. Circumstantial molecular and other evidence may be brought to bear on this view (e.g., the independent evolution of similar forms of mammals in marsupials and eutherians; and Zuckerkandl 1980, 1982). Such evidence seems to point to a probable trend toward parallel evolution (Zuckerkandl and Villet 1988; Zuckerkandl 1994) among living systems anywhere in the universe, a trend of a generality that so far has not been widely envisaged. It would imply organizational commonalities among any possible genetic systems. If the proof is only in the pudding—a pudding is not available. Yet, it might be possible to predict one.

The inferred spontaneous processes of increase in regulatory complexity represent a systematically directional evolutionary parameter. Darwinism has trouble accommodating this kind of parameter, as Wicken (1984) remarked. In fact, introducing this directional parameter is a development that need not pass for an "ism" of any kind, since the parameter is purely mechanistic (which is its only "ism," and not one of ideology). Ernst Mayr (1997) stated, "A modern evolutionist would say that the formation of a more complex system, representing the emergence of a new higher level, is strictly a matter of genetic variation and selection." A more elaborate formulation might begin the sentence as Ernst Mayr did but continue it with the words ". . . is a matter of genetic variation, selection, drift, and spontaneous further accretions of complexity at particular levels of hierarchical integration and in particular sectors of the organism." The presumed evolutionary quasi-independence of localized increases in complexity is a topic awaiting future development.

Acknowledgments. I warmly thank Karl Fryxell for extraordinary help, Marcus Feldman for encouragement and thought-provoking discussions, Tomoko Ohta for repeated kind and thoughtful advice, Ward Watt for insightful criticism and comment, Ernst Mayr for a careful critical reading of the manuscript and remarks that will inspire the article to follow, Mark Ptashne for an informative conversation, and Jiong Ma as well as Susan Quan for highly appreciated assistance in preparing Fig. 1. The present work has been supported in part by grant No. 97-14968 from the National Science Foundation.

References

Asturias FJ, Jiang YW, Myers LC, Gustafsson CM, Kornberg R (1999) Conserved structures of Mediator and RNA polymerase II holoenzyme. Science 283:985–987

Babin DR, Schroeder WA, Shelton JR, Shelton JB, Robberson B (1966) The amino acid sequence of the gamma chain of bovine fetal hemoglobin. Biochem 5:1297–1310

Chedin S, Ferri ML, Peyroche G, Andrau JC, Jourdain S, Lefebvre O, Werner M, Carles C, Sentenac A (1998) The yeast RNA polymerase III transcription machinery: a paradigm for eukaryotic gene activation. Cold Spring Harbor Symp Quant Biol 63:381–389

Cramer P, Bushness DA, Fu J, Gnatt AL, Maier-Davids B, Thompson NE, Burgess RR, Edwards AM, David PR, Kornberg RD (2000) Architecture of RNA polymerase II and implications for the transcription mechanism. Science 288:640–649

Dalgleish P, Sharrocks AD (2000) The mechanism of complex formation between Fli-1 and SRF transcription factors. Nucleic Acids Res 28:560–569

Dean AM, Dykhuizen DR, Hartl DL (1988) Fitness effects of amino acid replacements in the beta-galactosidase of *Escherichia coli.* Mol Biol Evol 5:469–485

Deb A, Zamanian-Daryoush M, Xu Z, Kadereit S, Williams BR (2001) Protein kinase PKR is required for platelet-derived growth factor signaling of *c-fos* gene expression via Erks and Stat3. Embo J 20: 2487–2496

Drewett V, Molina H, Millar A, Muller S, von Hesler F, Shaw PE (2001) DNA-bound transcription factor complexes analysed by mass-spectrometry: binding of novel proteins to the human *c-fos* SRE and related sequences. Nucleic Acids Res 29:479–487

Eanes WF (1999) Analysis opf selection on enzyme polymorphisms. Annu Rev Ecol Syst 30:301–326

Frost JA, Steen H, Shapiro P, Lewis T, Ahn N, Shaw PE, Cobb MH (1997) Cross-cascade activation of ERKs and ternary complex factors by Rho family proteins. EMBO J 16:6426–6438

Fryxell K (1996) The coevolution of gene family trees. Trends Genet 12:364–369

Garcia-Bellido A (1986) Genetic analysis of morphogenesis. In: Gustafson JP, Stebbins GD, Ayala FJ (eds) Genetics, development, and evolution. Plenum, pp 187–208

Garcia-Bellido A (1994) Genética del desarollo y de la evolución. Arbor 147:97–110

Garcia-Bellido A (1997) Progress in biological evolution. In: Burgen A, McLaughlin P, Mittelstraβ J (eds) The idea of progress. Walter de Gruyter, Berlin, pp 175–200

Getzoff ED, Cabelli DE, Fisher CL, Parge HE, Viezzoli MS, Bnci L, Hallewell RA (1992) Faster superoxide dismutase mutants designed by enhancing electrostatic guidance. Nature 358:347–351

Gilbert SF (1997) Developmental biology, fifth ed, Sinauer Associates, Sunderland, MA

Groisman R, Masutani H, Leibovitch MP, Robin P, Soudant I, Trouche D, Harel-Bellan A (1996) Physical interaction between the mitogen-responsive serum response factor and myogenic basic-helix-loop-helix proteins. J Biol Chem 271:5258–5264

Grueneberg DA, Henry RW, Brauer A, Novina CD, Cheriyath V, Roy AL, Gilman M (1997) A multifunctional DNA-binding protein that promotes the formation of serum response factor/homeodomain complexes: identity to TFII-I. Genes Dev 11:2482–2493

Hampsey M, Reinberg D (1999) RNA polymerase II as a control panel for multiple coactivator complexes. Curr Opin Genet Dev 9:132–139

Hanlon M, Bundy LM, Sealy L (2000) C/EBPBeta and Elk-1 synergistically transactivate the *c-fos* serum response element. BMC Cell Biol 1:2

Hartl DL, Dykhuizen DE, Dean AM (1985) Limits of adaptation: the evolution of selective neutrality. Genetics 111:655–674

Hartl DL, Taubes CH (1996) Compensatory nearly neutral mutations: selection without adaptation. J Theor Biol 182:303–309

Hasty J, Collins JJ (2001) Unspinning the web. Nature 411:30–31

Heidenreich O, Neininger A, Schratt G, Zinck R, Cahill MA, Engel K, Kotlyarov A, Kraft R, Kostka S, Gaestel M, Nordheim A (1999) MAPKAP kinase 2 phosphorylates serum response factor in vitro and in vivo. J Biol Chem 274:14434–14443

Herskowitz I (1989) A regulatory hierarchy for cell specialization in yeast. Nature 342:749–757

Janknecht R, Hunter T (1997) Convergence of MAP kinase pathways on the ternary complex factor Sap-1a. EMBO J 16:1620–1627

Jeong H, Mason SP, Barabási A-L, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–42

Latchman D (1998) Eukaryotic transcription factors, third ed, Academic Press

Lee HY, Chaudhary J, Walsh GL, Hong WK, Kurie JM (1998) Suppression of c-Fos gene transcription with malignant transformation of human bonchial epithelial cells. Oncogene 16:3039–3046

Lenski RE, Travisano M (1994) Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. Proc Natl Acad Sci USA 91:6808–6814

Ling Y, West AG, Roberts EC, Lakey JH, Sharrocks AD (1998) Interaction of transcription factors with serum response factor: identification of the Elk-1 binding surface. J Biol Chem 273: 10506–10514

Liu SH, Ma JT, Yueh AY, Lees-Miller SP, Anderson CW, Ng SY (1993) The carboxyl-terminal transactivation domain of human serum response factor contains DNA-activated protein kinase phosphorylation sites. J Biol Chem 268:21147–21154

Locker J (2001) Transcription factors. Academic Press, San Diego

Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473

Manak JR, Prywes R (1993) Phosphorylation of serum response factor by casein kinase. II: evidence against a role in growth factor regulation of *fos* expression. Oncogene 8:703–711

Mayr E (1997) This is biology. Belknapp Press of Harvard University Press

Mo Y, Vaessen B, Johnston K, Marmorstein R (1998) Structures of SAP-1 bound to DNA targets from the E74 and *c-fos* promoters: insights into DNA sequence discrimination by Ets proteins. Mol Cell 2:201–212

Moore FB, Rozen DE, Lenski RE (2000) Pervasive compensatory adaptation in *Escherichia coli.* Proc Roy Soc Lond B Biol Sci 267: 515–522

Murphy DJ, Hardy S, Engel DA (1999) Human SWI-SNF component BRG1 represses transcription of the *c-fos* gene. Mol Cell Biol 19: 2724–2733

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, Berlin

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246:96–98

Ohta T (1987) Very slightly deleterious mutations and the molecular clock. J Mol Evol 26:1–6

Ohta T (1988) Evolution by gene duplication and compensatory advantageous mutations. Genetics 120:841–847

Ohta T (1997) Role of random genetic drift in the evolution of interactive systems. J Mol Evol 44:S9–S14

Ohta T, Tashida H (1990) Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. Genetics 126:219–229

Omoike OI, Benson BA, Chan MA, Benedict SH (1999) Sequences at the 3′ side of the *c-fos* SRE mediate gene expression via an Sob1-dependent, TCF-independent pathway. Biochem Biophys Res Commun 262:523–529

Papadopoulos D, Schneider D, Meier-Eiss J, Arber W, Lenski RE, Blot M (1999) Genomic evolution during a 10,000-generation experiment with bacteria. Proc Natl Acad Sci USA 96:3807–3812

Petit C, Zuckerkandl E (1976) Evolution. Génétique des populations, évolution moléculaire. Hermann, Paris

Price MA, Rogers AE, Treisman R (1995) Comparative analysis of the ternary complex factors Elk-1, SAP-1a and SAP-2 (ERP/NET). EMBO J 14:2589–2601

Ramirez S, Ait-Si-Ali S, Robin P, Trouche D, Harel-Bellan A, Ait-Si-Ali S (1997) The CREB-binding protein (CBP) cooperates with the serum response factor for transactivation of the *c-fos* serum response element. J Biol Chem 272:31016–31021

Rivera VM, Miranti CK, Misra RP, Ginty DD, Chen RH, Blenis J, Greenberg ME (1993) A growth factor-induced kinase phosphorylates the serum response factor at a site that regulates its DNA-binding activity. Mol Cell Biol 13:6260–6273

Sealy L, Malone D, Pawlak M (1997) Regulation of the *c-fos* serum response element by C/EBPbeta. Mol Cell Biol 17:1744–1755

Sentenac A, Riva M, Thuriaux P, Buhler JM, Treich I, Carles C, Werner M, Ruet A, Huet J, Mann C, Chiannilkulchai N, Stettler S, Mariotte S (1992) Yeast RNA polymerase subunits and genes. In: McKnight SL, Yamamoto K (eds) Transcriptional regulation. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp 27–54

Shore P, Sharrocks AD (1994) The transcription factors Elk-1 and serum response factor interact by direct protein–protein contacts mediated by a short region of Elk-1. Mol Cell Biol 14:3283–3291

Simon KJ, Grueneberg DA, Gilman M (1997) Protein and DNA contact surfaces that mediate the selective action of the Phox1 homeodomain at the *c-fos* serum response element. Mol Cell Biol 17:6653–6662

Spencer JA, Baron MH, Olson EN (1999) Cooperative transcriptional activation by serum response factor and the high mobility group protein SSRP1. J Biol Chem 274:15686–15693

Stoltzfus A (1999) On the possibility of constructive neutral evolution. J Mol Evol 49:169–181

Suzuki Y, Gojobori T (1999) Method for detecting positive selection at single amino acid sites. Mol Biol Evol 16:1315–1329

Trouche D, Grigoriev M, Lenormand JL, Robin P, Leibovitch SA, Sassone-Corsi P, Harel-Bellan A (1993) Repression of *c-fos* promoter by MyoD on muscle cell differentiation. Nature 363:79–82

Trouche D, Masutani H, Groisman R, Robin P, Lenormand JL, Harel-Bellan A (1995) Myogenin binds to and represses *c-fos* promoter. FEBS Lett 361:140–144

Wagner BJ, Hayes TE, Hoban CJ, Cochran BH (1990) The SIF binding element confers sis/PDGF inducibility onto the *c-fos* promoter. EMBO J 9:4477–4484

Watt WB (1991) Biochemistry, physiological ecology, and population genetics: the mechanistic tools of evolutionary biology. Funct Ecol 5:145–154

Watt WB (1995) Allozymes in evolutionary genetics: beyond the twin pitfalls of "neutralism" and "selectionism." Rev Suisse Zool 102: 869–882

Watt WB, Dean AM (2000) Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. Annu Rev Genet 34:593–622

Wehinger J, Gouilleux F, Groner B, Finke J, Mertelsmann R, Weber-Nordt RM (1996) IL-10 induces DNA binding activity of three STAT proteins (Stat1, Stat3, and Stat5) and their distinct combinatorial assembly in the promoters of selected genes. FEBS Lett 394:365–370

Whitmarsh AJ, Yang SH, Su MS, Sharrocks AD, Davis RJ (1997) Role of p38 and JNK mitogen-activated protein kinases in the activation of ternary complex factors. Mol Cell Biol 17:2360–2371

Whyte LL (1965) Internal factors in evolution. Tavistock, Publications, London

Wicken JS (1984) On the increase in complexity during evolution. In: Beyond neo-Darwinism. Ho MW, Saunders PT (eds) Academic Press, New York, pp 89–112

Yang X, Chen Y, Gabuzda D (1999) ERK MAP kinase links cytokine signals to activation of latent HIV-1 infection by stimulating a cooperative interaction of AP-1 and NF-kappaB. J Biol Chem 274: 27981–27988

Yates PR, Atherton GT, Deed RW, Norton JD, Sharrocks AD (1999) Id helix-loop-helix proteins inhibit nucleoprotein complex formation by the TCF ETS-domain transcription factors. EMBO J 18:968–976

Zhou Q, Gedrich RW, Engel DA (1995) Transcriptional repression of the *c-fos* gene by YY1 is mediated by a direct interaction with ATF/CREB. J Virol 69:4323–4330

Zhou W, Clouston DR, Wang X, Cerruti L, Cunningham JM, Jane SM (2000) Induction of human fetal globin gene expression by a novel erythroid factor, NF-E4. Mol Cell Biol 20:7662–7672

Zhu C, Johansen FE, Prywes R (1997) Interaction of ATF6 and serum response factor. Mol Cell Biol 17:4957–4966

Zuckerkandl E (1963) Perspectives in molecular anthropology. In: Washburn SL (ed) Classification and human evolution. Aldine Publishing, pp 243–272

Zuckerkandl E (1964) Controller gene diseases: the operon model as applied to beta-thalassemia, familial fetal hemoglobinemia, and the normal switch from the production of fetal hemoglobin to that of adult hemoglobin. J Mol Biol 8:128–147

Zuckerkandl E (1972) Some aspects of protein evolution. Biochimie 54:1095–1102

Zuckerkandl E (1976a) Evolutionary processes and evolutionary noise. II. A selectionist model for random fixations in proteins. J Mol Evol 7:269–311

Zuckerkandl E (1976b) Evolutionary processes and evolutionary noise at the molecular level, I. Functional density in proteins. J Mol Evol 7:167–183

Zuckerkandl E (1978) Multilocus enzymes, gene regulation, and genetic sufficiency. J Mol Evol 12:57–89

Zuckerkandl E (1979) Controller node complexity: a measure of the degree of gene coordination. J Mol Evol 14:311–321

Zuckerkandl E (1980) Sequences, phenotypes, and directional evolution (abstract). Second Intern Congr of Systematic and Evolutionary Biology, University of British Columbia, Vancouver, Canada, p 124

Zuckerkandl E (1982) Molecular bases for directional evolution (abstract). In: Modalités, rythmes, mécanismes de l'évolution biologique. J Chaline (ed), Colloques Internationaux du Centre National de la Recherche Scientifique, No 330, Paris, France

Zuckerkandl E (1983) Topological and quantitative relationships in evolving genomes. In: Structure, dynamics, interactions, and evolution of biological macromolecules, C Hélène (ed), D. Reidel, pp 395–412

Zuckerkandl E (1987) On the molecular evolutionary clock. J Mol Evol 26:34–46

Zuckerkandl E (1992) Revisiting junk DNA. J Mol Evol 34:259–271

Zuckerkandl E (1994) Molecular pathways to parallel evolution: I. Gene nexuses and their morphological correlates. J Mol Evol 39: 661–678

Zuckerkandl E (1997) Neutral and nonneutral mutations: the creative mix—evolution of complexity in gene interaction systems. J Mol Evol 44(Suppl 1):S2–S8

Zuckerkandl E, Pauling L (1962) Molecular disease, evolution, and genic heterogeneity. In: Kasha M, Pullman B (eds) Horizons in Biochemistry. Academic Press, pp 189–225

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Evolving genes and proteins, V Bryson, HJ Vogel (eds), New York: Academic Press, pp 97–166

Zuckerkandl E, Villet R (1988) Concentration-affinity equivalence in gene regulation: convergence of genetic and environmental effects. Proc Natl Acad Sci USA 85:4784–4788