Multigene Families: The Problem of Molecular Recapitulation

S. N. Rodin, A. Y. Rzhetsky, and A. A. Zharkikh [1]

Introduction

The multigene families (MF) are known to have been formed in the course of evolution mainly by sequential duplication of ancestor genes. Almost all MFs are characterized by some specified order of homologous gene expression in the course of ontogenesis. The question arises: are the genes expressed in early ontogenetic stages more "ancient" than their ontogenetically later expressed homologues? Zuckerkandl [1] was the first to formulate and study this question with respect to the MFs. Taking into account that divergence of α- and β-subfamilies of globins occurred much earlier than those of β-like genes, he compared human β-globins – namely, γ (fetal) and β (adult) protein sequences – with α-globin. The latter protein sequence was taken as a marker close to the "ancestor". Zuckerkandl supposed that if the fetal β-like globin (γ) was closer to the α-globin that the adult one (β), the former protein could be assumed to be more ancient than the latter one, and thus evidence in favour of molecular recapitulation would be found. Nevertheless, he discovered that both γ- and β-sequences showed the same number of amino acid dissimilarities (55) with the α-globin [1]. This result compromised the idea of molecular recapitulation for a rather long period.

It is a priori evident that if the phenomenon of molecular recapitulation really takes place, it must be caused by the stabilizing natural selection: the earlier a gene is expressed in ontogeny, the wider is the range of possible undesirable consequences of any mutation in the gene. Selection of this kind must preserve the structure of "functional" domains of the gene much more carefully than those which are "subneutral". Thus, it is not unlikely that a large number of subneutral substitutions is masking a smaller number of substitutions located in the functional sites. Therefore, we decided to verify this suggestion using more representative samples of globin nucleotide sequences and more adequate and rigorous methods than Zuckerkandl of phylogenetic analysis and of differentiating the mutations in the globin functional sites from all the others.

Results

All of the sequences employed were taken from the GenBank data base. Trees were constructed by means of the maximum parisimony method of Zharkikh [2] (program UNISUB). A number of other programs from the VOSTORG package were also used [3].

Using the data of Perutz [4], we have divided the amino acid sites of the globins into two groups: "functional" and "nonfunctional" (or "subneutral"). All amino acid sites that participate in some important functional contacts were assigned to the former group. This group includes sites involved in: the α- and β-contacts with haem, the Bohr effect, the α–β bonds between the haemoglobin sub-

[1] Institute of Cytology and Genetics, Siberian Branch of USSR Academy of Sciences, Novosibirsk, 630090, USSR.

Haematology and Blood Transfusion Vol. 35
Modern Trends in Human Leukemia IX
R. Neth et al. (Eds.)

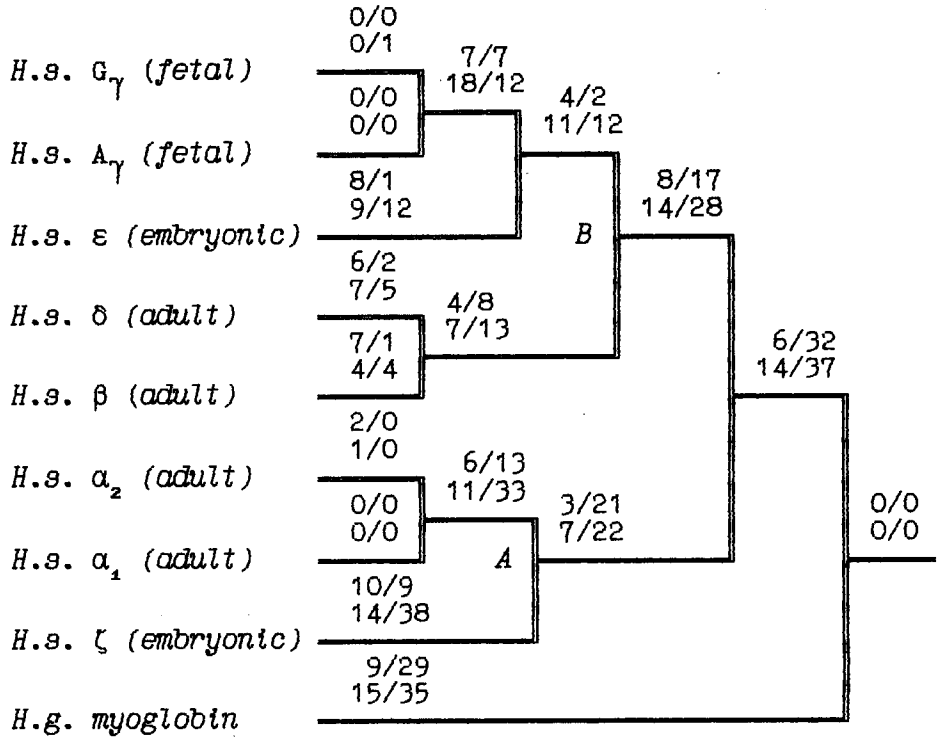

Fig. 1. Phylogenetic tree inferred by maximum parsimony method for eight human globin genes. The *Halichoerus grypus* myoglobin gene was used as a homologous but relatively distant gene to determine the position of the tree root. The total number of substitutions is 593. The ancestral sequences were reconstructed for tree nodes *A* and *B* (for α-like and β-like sequences, respectively).

Abbreviations: *H.s.*: Homo sapiens; *H.g.*: Halichoerus grypus.

Upper line of numbers above each branch, the numbers of synonymous/nonsynonymous reconstructed nucleotide substitutions in the "functional" sites. *Lower line of figures* above each branch, the same for "nonfunctional" sites

units, the binding of 2,3-diphosphoglycerate (for β-like chains) and the salt bridges. The nonfunctional group includes all the other sites.

On the base of the primary DNA sequence alignment, phylogenetic trees for the globin genes of *Homo sapiens* (see Fig. 1), *Capra hircus* and *Xenopus laevis* (not presented) were inferred. In order to determine the position of the tree root, we used the *Halichoerus grypus* myoglobin gene as a homologous but relatively distant gene.

When estimating branch lengths of the trees we sorted the reconstructed nucleotide substitutions in a special way. Each nucleotide substitution was characterized from two points of view: on the one hand, as affecting a functional or nonfunctional site of the protein, and on the other hand, as synonymous or nonsynonymous. Using the estimated branch lengths we computed the distances between the present day sequences and the corresponding ancestor ones reconstructed for each α- and β-gene cluster. The results are presented in Table 1.

Studying both α- and β-like human sequences revealed the same regularity: the number of reconstructed nonsynonymous substitutions fixed in the functional sites of the embryonic genes (ζ and ε) is threefold less than in adult genes (α_1, α_2 in α-cluster and β, δ in β-cluster). The analogous values for the fetal and adult β-like genes are almost equal (about nine substitutions) (see Table 1).

In fact, the same could be said about the *C. hircus* genes. The goat β-cluster consists of three groups of genes [5]: the β^c, β^A and β^F genes (the last one is also often designated γ); they are orthologous to the human β-globin, $\psi\beta^X$, $\psi\beta^Y$ and $\psi\beta^Z$ pseudogenes and the δ-globin gene.

Table 1. Evolutionary distances between the human present-day globin genes and the reconstructed ancestral sequences

Gene	"Functional" sites			"Nonfunctional" sites			Total
	Synonymous	Nonsynonymous	Total	Synonymous	Nonsynonymous	Total	
ε	12	3	15	20	24	44	59
G_γ	11	9	20	29	25	54	74
A_γ	11	9	20	29	24	53	73
δ	10	10	20	14	18	32	52
β	11	9	20	11	17	28	48
α_1	6	13	19	11	33	44	63
α_2	8	13	21	12	33	45	66
ζ	10	9	19	14	38	52	71

In the individual development of a goat, besides embryonic (ε^I and ε^{II} genes), fetal (β^F/γ) and adult stages (β^A) of globin gene expression, an additional "preadult" or "juvenile" stage is found which is characterized by the expression of β^C gene [5].

Thus, the ε^I (goat β-like embryonic) gene appears to be the closest of the β-like genes to the "ancestor" gene (if only the nonsynonymous substitutions in functional sites are considered). Almost negligible regularity $\beta^A > \beta^C > \gamma$ is observed for the other three genes (see Table 2).

As for the goat ε^{II} gene, it was noticed that it exceeds all other β-like genes both in the total number of substitutions and in almost any particular group of distances (see Table 2). Taking into account that this gene

1) significantly differs from the goat ε^I and human ε genes,
2) has accumulated large numbers of nonsynonymous substitutions in the "functional" sites (Table 2), and
3) is orthologous to the primate gene ($\psi\beta_1$) that was proved to be a pseudogene, it is reasonable to suggest that the goat ε^{II} gene is not an active one, but could be involved in some other processes, e.g. regulation of ontogenetic expression of the globins, as proposed by Goodman et al. [6] for the primate $\psi\beta_1$ gene.

Finally, the most significant regularity was found for the *X. laevis* globin genes [7]: the tadpole genes from both α- and β-clusters are approximately twice as close to the corresponding ancestors than the adult ones and it was the class of nonsynonymous substitutions in the functional sites that revealed this difference (see Table 3).

Table 2. Evolutionary distances between the globin genes of the goat and the reconstructed ancestral sequence

Gene	"Functional" sites			"Nonfunctional" sites			Total
	Synonymous	Nonsynonymous	Total	Synonymous	Nonsynonymous	Total	
ε^I	5	4	9	26	18	44	53
ε^{II}	13	14	27	32	23	55	82
γ	7	9	16	10	27	37	53
β^C	8	10	18	14	29	43	61
β^A	7	11	18	11	27	38	56

Table 3. Evolutionary distances globin genes of the clawed frog and the reconstructed ancestral sequences. The tadpole genes are designated as (*t*), and the adult ones as (*a*)

Gene	"Functional" sites			"Nonfunctional" sites			Total
	Synonymous	Nonsynonymous	Total	Synonymous	Nonsynonymous	Total	
$\alpha(t)$	8	8	16	20	34	54	70
$\alpha_1(a)$	21	19	40	26	42	68	108
$\alpha_2(a)$	23	19	42	29	43	72	113
$\beta_1(t)$	15	21	36	30	49	79	115
$\beta_2(t)$	17	20	37	32	49	81	118
$\beta_1(a)$	17	27	44	21	39	60	104
$\beta_2(a)$	17	28	45	21	39	60	105

Summing up, let us note that the effect expected by Zuckerkandl can be clearly seen when embryonic/"ancestor" and adult/"ancestor" distances are compared. It does not hold true when comparing fetal/"ancestor" and adult/"ancestor" distances. The latter conclusion is obviously in agreement with Zuckerkandl's idea: there were no embryonic-stage globins in his sample of amino acid sequences. There are good reasons to consider the fetal-stage globins (and the goat "preadult" globin) as the product of relatively recent gene duplications. Thus, the time span after the last duplication might have been insufficient to accumulate the differences in the degree of evolutionary conservatism of the fetal- and adult-stage globin genes.

It should be emphasized that when analysing phylogenetic relations in some other MFs [immunoglobulin genes of mammals [8], insect chorion protein genes [9], and even homeoboxes of some regulatory genes of *Drosophila melanogaster* responsibel for embryonic morphogenetic gradients, segmentation and differentiation of the segments (S. N. Rodin, unpublished)] we found a tendency resembling that described here for globin genes. For example, the order of duplication of immunoreceptor progenitor genes in the evolutionary past was in good agreement with the order of gene rearrangements and their expression in the course of B- and T-lymphocyte differentiation [8].

Discussion

"Relay-Race" Regime of Molecular Evolution

Any significant increase in the rate of substitution fixation in a particular gene from a multigene family could be explained in two ways. The first explanation implies that the pressure of stabilizing (negative) natural selection is lessened. The second possible cause of the same phenomenon might be the improvement in the gene function that is provided by positive natural selection. In the second case, the higher the rate of adaptive evolution, the larger the substitution load, i.e. Haldane's dilemma must be playing an important role in evolutionary periods of just this kind. These two possible reasons might appear to be combined in the case of globin gene family evolution [10–13]. Although gene multiplications seem to be quite an ordinary event in genome evolution, they far more often give rise to silent pseudogenes than to novel functional genes.

The above may imply that multigene family evolution occurs in this "relay-race" mode, i.e. at any moment, most probably only one gene within the same family is allowed to evolve in an adaptive manner [11].

In fact, the relay-race mode of molecular evolution may be considered as a general theoretical substantiation of a cascade-like pattern of switches in ex-

pression from one structural gene to another in the course of ontogenesis.

Regulation of Development and Anaboly

The majority of authors (see [14]) are unanimous in assuming that ontogenesis is regulated by a number of genes that are organized as a "Bickford fuse" or a "relay-race with a specified time of last participant arrival". This means that the expression of "the right gene in the right time and in the right cell" requires a chain of intermediate regulatory gene activations. The last participant of this relay-race must activate the target gene. This chain of activations must be characterized by strict adherence to the expression timetable. Each regulatory gene might be responsible for multiple gene activations. In turn, a group of regulatory genes is often controlled by a higher order regulatory gene. Thus, the scheme of gene interactions in ontogeny is undoubtedly a hierarchic one.

The mode of terminal addition of new stages (called anaboly by Severtsov [15]) appears to be the least dangerous mode of gaining ontogenetic complexity. The latter does not mean that "nonanabolic" evolutionary rearrangements of individual development are forbidden, but in reality they are likely to occur far more rarely than the anabolic ones.

There are well-studied examples where the prolonged activity of an earlier expressed gene compensated for a malfunction in its later expressed homologues (see [16]), i.e. the earlier expressed gene could be said to recapitulate the ancestral mode of expression. Notably, among all the reported cases of human globin gene malfunctions (thalassaemias) there are no examples of compensating embryonic gene damage by expression of fetal or adult globin genes. Thus, one can conclude that, for example, a normal activation of fetal globins takes place only provided that the embryonic gene was expressed normally etc. Thus, the structural globin genes are also organized into some analogue of the regulatory hierarchy and the later expressed genes are more open to evolutionary changes.

Recapitulation and Selective Strategies

The so-called "biogenetic law" of Haekkel was proved to hold true only in some cases and not in others (see [14]). However, one can explain (and maybe even predict) whether recapitulation will be found in any particular case if the following speculations are valid.

There are two main "poles" of natural selection that are recognized by ecologists [17]. The complexity of any ecological system is thought to be determined, on the one hand, by the quantity of free energy available and, on the other hand, by the stability of the environment.

An environment which is characterized by low probability of intensive disastrous fluctuations is usually most densely populated. Plant and animal communities in these conditions are known to form complex trophic chains that utilize free energy in the most efficient way. The intensive intra- and inter-specific competition that is observed in these cases favours the increase of organism complexity. Selection of this kind is called "K-selection" [17].

When the environment is unstable (large parts of populations are randomly eliminated) the individuals which have more offspring are most successful. This kind of selection is known as "r-selection". A prolonged period of r-selection may cause a drastic reduction in the morphologic and ontogenetic complexity.

It is quite reasonable to suggest that the anabolic complication of ontogenesis must be demonstrated by species evolving under pronounced K-type natural selection. On the other hand, it is unlikely that traces of a recent terminal addition of new stages will be found when typical r-strategy species are considered.

Of course, when real organisms are being dealt with, the picture might appear to be much more complex. First of all,

ancestors of almost any present-day animal surely underwent multiple successions of r- and K-selection. This means that what could be observed a posteriori is a complicated tangle of tendencies. Apart from that, there are a great number of species which could not be definitely classified according to the r/K scheme. Thus, the hypothesis suggested may be applied only to relatively "recent" spans of evolutionary time when the species observed are known to evolve under one kind of selection.

Summary and Conclusions

Multigene families (MF) represent the most promising level of genome organization when studying the molecular basis of both developmental and evolutionary processes. Haldane's cost of selection "allows" almost all MFs to increase their complexity in evolution in a relay-race manner. Each MF is in turn characterized by a strict ontogenetic order of expression of homologous structural genes. According to Zuckerkandl, if any earlier expressed gene resembles in structure the ancestor gene more than its later expressed homologue, this could be considered as a case of molecular recapitulation. We showed here that this phenomenon does occur invarious MFs when comparison is performed only for sites that are known to be involved in selectively important functional bonds. For all other sites, conditionally denoted nonfunctional or subneutral, this regularity is not valid. The dichotomic mode of switches in gene expression, unreciprocity of ontogenetic compensation of human globin gene malfunctions (adult by fetal but not reverse), allelic and isotypic exclusions in expression of immunoglobulin genes clusters are certainly associated with the molecular recapitulation phenomenon.

References

1. Zuckerkandl E (1968) Hemoglobins, Haeckel's "biogenetic law", and molecular aspects of development. In: Rich A, Davidson N (eds) Structural chemistry and molecular biology. Freeman, San Francisco, pp 256–274
2. Zharkikh AA (1977) Algorithms of phylogenetic tree buildinging from amino acid sequences (in Russian). In: Ratner V (ed) Mathematical models of evolution and selection. Institute of Cytology and Genetics, Novosibirsk, pp 5–52
3. Zharkikh AA, Rzhetsky A, Morozov PS, Sitnikova TL, Krushkal JS (1990) VOSTORG: package of a microcomputer program of phylogenetic analysis. Gene (in press)
4. Perutz MF (1972) Nature of haem-haem interaction. Nature 237:495–499
5. Schon EA, Cleary ML, Haynes JR, Lingrel JB (1981) Structure and evolution of goat γ-, β^{C}- and β^{A}-globin genes: three developmentally regulated genes contain inserted elements. Cell 27:359–369
6. Goodman M, Koop BF, Czelusniak J, Weiss ML (1984) The η-globin gene family of mammals. J Mol Biol 180:803–823
7. Knochel W, Meyerhof W, Stadler J, Weber R (1985) Comparative nucleotide sequence analysis of two types of larval β-globin mRNA of *Xenopus laevis*. Nucleic Acids Res 13:7899–7908
8. Rzhetsky A, Rodin SN (1987) Theoretical analysis of relations between an order of evolutionary divergencies and developmental stages (in Russian). Genetics (USSR) 23:2183–2195
9. Rzhetsky A, Rodin SN, Zharkikh AA (1990) "Biogenetic law" and evolution of multigene families (in Russian). Institute of Cytology and Genetics, Novosibirsk, pp 1–60
10. Ratner VA, Rodin SN, Zharkikh AA (1977) Analysis of globin phylogeny by a more precise method (in Russian). In: Ratner VA (ed) Mathematical models of evolution and selection. Institute of Cytology and Genetics, Novosibirsk, pp 53–96
11. Rodin SN (1985) Multigenic families: evolutionary problems (in Russian). Mol Biol (Mosc) 21:198–240
12. Li W-H (1985) Accelerated evolution following gene duplication and its implication for the neutralist-selectionist con-

troversy. In: Ohta T, Aoki K (eds) Population genetics and molecular evolution. Springer, Berlin Heidelberg New York, pp 333–352
13. Goodman M, Moore GW, Matsuda G (1975) Darwinian evolution in the geneology of haemoglobin. Nature 253:603–608
14. Raff RA, Kaufman TC (1983) Embryos, genes and evolution. Macmillan, New York
15. Severtsov AN (1945) Evolution of fins (in Russian). USSR Academy of Sciences, Moscow (Selected works, vol 2)
16. Henthorn PS, Mager DL, Huisman THJ, Smithies O (1986) A gene deletion ending within a complex array of repeated sequences 3′ to the human β-globin gene cluster. Proc Natl Acad Sci USA 83:5194–5198
17. MacArthur RH, Wilson EO (1967) The theory of island biogeography. Princeton University Press, Princeton