Z. Phys. Chem. **216** (2002) 77–90

Multilevel Evolution: The Fate of Duplicated Genes

By Paulien Hogeweg*

Theoretical Biology and Bioinformatics Grp, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

Dedicated to Prof. Dr. Peter Schuster on the occasion of his 60th birthday

(Received May 2, 2001; accepted June 19, 2001)

Evolution of Complexity / Gene Duplication / Gene Regulation / Gene Expression / Orthologs vs. Paralogs / Genotype-Phenotype Mapping / Neutral vs. Adaptive Evolution

Biological evolution is a multilevel process and should be studied as such. A first, important step in studying evolution in this way has been the work of Peter Schuster and co-workers on RNA evolution. For RNA the genotype-phenotype mapping can be calculated explicitly. The resulting evolutionary dynamics is dominated by neutral paths, and the potential of major change by a single point mutation.

Examining whole genomes, of which about 60 are now available, we see that gene content of genomes is changing relatively rapidly: gene duplication, gene loss and gene generation is ubiquitous. In fact, it seems that point-mutations play a relatively minor role, relative to changes in gene regulation and gene content in adaptive evolution.

Large scale micro-array studies, in which the expression of every gene can be measured simultaneously, give a first glimpse of the 'division of labor' between duplicated genes. A preliminary analysis suggests that differential expression is often the primary event which allows duplicated genes to be maintained in a genome, but alternate routes also exist, most notably on the one hand the mere need of a lot of product, and on the other hand differentiation within multi-protein complexes consisting of homologous genes.

I will discuss these results in terms of multilevel evolution. in particular in terms of information integration and the alternatives of 'individual based' *versus* 'population based' diversity.

* E-mail: P.Hogeweg@bio.uu.nl

1. Introduction

1.1 Multilevel evolution

Many years ago, Maynard Smith urged an audience of population geneticists "to go once a year to the zoo, to seek out the elephant, stand in front of her, greet her, and say: *Elephant, I believe you came about by random mutations*". This, to get the point across how little was understood about genotype-phenotype mapping in general and about the evolution of complexity in particular.
At that time not even the idea of actually 'making something by random mutations' was familiar yet, but it has now become standard practice in both in vitro and in silico 'optimization' procedures (in particular in evolving RNA enzymes, and in evolutionary computation respectively). Moreover, the seminal work of Peter Schuster and his group from [6] 1987 onward on RNA genotype-phenotype mapping, *i.e*. on the one example where we can calculate a phenotype from the genotype, has shed light a.o. on how the non-linearity of this mapping can help rather than hinder evolutionary (optimization) processes. See e.g. [14, 20]. Moreover, this work has finally brought neutral drift and adaptive evolution into one framework such that neutrality in fact helps adaptation [7, 11, 24]. The work on RNA genotype phenotype mapping has also demonstrated that evolutionary processes can have interesting side-effects, e.g. the evolution towards mutational robustness [13, 15] during evolution to a fixed target (whereas the opposite is true in a coevolutionary (red queen) situation). Nevertheless, although the trend towards robustness could have increased complexity as a side-effect, the evolution of complexity in replicationrate mediated evolution is still enigmatic. I think the recommendation to meet the elephant once a year still holds, but also that it is worthwhile to take up the challenge to understand why evolution *sometimes* leads to increased complexity.

A starting point for such an endeavor, is, I think, the realization that, unlike the studied RNA case, the genotype-phenotype mapping of organisms involves (the interaction of) many entities. These entities of which evolving replicators are composed, can themselves be replicators subject to a mutation selection process. Plasmids and endobacteria are straightforward examples, but genes can be considered as such as well in view of within genome duplication and horizontal transfer. Moreover, also non-replicating entities may act as levels of selection, and can enslave the evolution of the replicators of which they are composed. For example, [1] have shown that the dynamic properties of spiral wave patterns which generically form in spatial versions of the classical Hypercycle model of [2] determine the evolutionary fate of the replicating RNA's of which they are composed.

It is crucial to take the multilevel nature of biotic systems into account when we study (the evolution of) complexity [9, 10]. Indeed the multilevelness co-defines the complexity. Moreover, complexity, and the evolution of complexity, should be examined at different levels. It appears that it is through

the interaction of the levels that replicationrate based selection can lead to increased complexity.

1.2 Individual based and population based diversity

In models which incorporate three or four levels of selection, *i.e.* plasmids, cells, spatially localized populations of cells, and in case of the four level models, viruses, we have shown that indeed increase of diversity (complexity) can switch from one level to the other [17, 18]. For example, in the four level model of bacteria, restriction-modification encoding plasmids (RM systems) and bacteriophages in a spatial explicit situation, we see two possible outcomes, which we named 'individual based diversity' and 'population based diversity' respectively. In the case of individual based diversity all bacteria contain all restriction-modification types, and each should therefore be able to defend itself against many bateriophages. In the second case all types of restriction-modification systems survive in the population as well, but each individual bacterium contains just one or no restriction-modification system. The latter 'simple' (*i.e.* containing few RM systems) bacteria reach much higher densities. This is because the phages cannot deal with a very diverse population and get (almost) extinct. On the other hand, the 'complex' bacteria (*i.e.* containing many different RM systems) are 'all complex together', and the population is homogeneous. This leads to the full methylation of the phage population, and the RM systems do not give any protection anymore. The two modes occur for a large range of parameter settings as alternative stable attractors of the system in this four level model, whereas in a similar three level model (of colicines, bacteria, and spatial sub-populations) there is a phase transition between the two modes.

These examples clearly show that complexity can reside at different levels, and that fitness and complexity are not correlated. They also show that complexity nevertheless arises at some levels and determines the fitness at other levels.

1.3 Genome evolution

In this paper we take these insights as a starting point to examine the evolutionary fate of duplicated genes. From studying fully sequenced genomes, of which more than 60 are now available, it has become clear that the turnover of genes is a relatively rapid process. Comparing detectable similarity after evolutionary divergence with respect to protein sequence, gene content and regulation, in mainly prokaryotes, [12] have shown that the sequence is most conserved, while the fastest change occurs at the level of gene regulation. Gene content does contain a strong phylogenetic signal ([21] showed that a respectable 'tree of life' can be contructed from this signal alone), and the flux of genes (loss, duplication, genesis, and recruitment through horizontal transfer) is high along the

branches of the tree (Snel and Huynen submitted). Thus, while most evolutionary conceptualization has focussed on small changes in existing genes, changes in gene content and regulation appears a major driving forces in evolution.

Such studies differentiate between two subsets of homologous genes, *i.e*. so called *orthologous* genes, *i.e*. genes that diverged due to a speciation event, and the so called *paralogous genes* genes which came about by duplication, and diverged within one genome [5].

It is tempting to compare this difference to the population based diversity *versus* the individual based diversity mentioned above. Gene duplication and retention of several 'copies' in the genome results in a complexity increase at the level of the organisms whereas speciation and the divergence of genes due to isolation in different species results (at least in the case of sympatric speciation) in an increase of complexity at the ecosystem level.

Many theoretical and experimental studies have focussed on the mechanisms of organism speciation, *i.e*. on the generation of 'population based diversity'. In this paper I will focus on 'gene speciation', within an organism, *i.e*. how paralogous genes 'earn their keep' so that the can persist in a genome. Thus we address 'individual based diversity'. To this end, we study the relation between homology and gene expression.

2. Expression pattern of homologous genes

Basically there are four possibilities of the relationship between homology and expression patterns: Do homologous genes do the same thing in the same environment, the same thing in different environments, or different things in the same or in different environments. Despite the ill-definedness of 'doing the same thing' we will take these categories as search image to interpret the (dis)similarity between gene expression patterns and homology. We will show that all four possibilities do occur, and we identify and characterize groups of genes in each of the categories.

2.1 The data: yeast

In this pilot study on the evolutionary fate of duplicated genes we focus on expression patterns of homologous genes in yeast.

The basic data we use consist of:

- Gene expression data:
  2467 yeast genes for which annotations are available, 79 conditions (cell cycle, sporulation, diauxic shift) [3]. A subset of 1899 sufficiently differentially expressed to be included in the analysis (taking the (arbitrary) criterion of a difference of 2 on the log scale of the data).
- Homology of these genes:
  Homology was calculated using Smith-Waterman local alignment. Pairs of

genes were included for P-values smaller then 0.01. There are 7319 homologous pairs of genes among the 2467 annotated genes. 1364 genes have one or more homologs. The highly expressed subset of genes has 4598 homologous pairs involving 1012 genes.

- Annotation of these genes:
  The annotation consists of two parts: process identification (e.g. protein synthesis), and protein identification (e.g. RRNA Helicase).

2.2 Methods

We perform analyses starting from each of these datasets and mapping the results onto the other data. Doing so we highlight different aspects, at different scales of resolution between homology, expression and function of the genes.

The following methods and criteria were used:

- Expression similarity is expressed as the cosine of the angle of the expression of both genes in the 79 dimensional space. We use this non-normalized correlation coefficient because the data have a meaningful zero value: *i.e.* expression equal to the control.
- Hierarchical cluster analysis of the expression data is done using average group similarity, *i.e.* the UPGMA method which is also sometimes used in phylogenetic analysis. Here, however, it is simply used as a space conserving clustering method. In most cluster analyses only genes which are differentially expressed by at least a factor 2 were included in the analysis as indicated above.
- Comparison of expression pattern and homology was represented by plotting the cluster membership of the two genes of each homologous pair. Clusters, on the y-axis are ordered along the dendrogram (note that the hierarchical structure of the dendrogram only defines a partial ordering, but this ordering is nevertheless informative). Homologous pairs (x-axis) are ordered according to the cluster membership of the 'first' gene, while the homologous pairs are included twice, with each of the genes once as the 'first' gene. We did all analyses for 50 upto 300 clusters (in steps of 50 clusters). 50 clusters corresponds to an optimal splitting level (according to the criterion of [8]) while 300 clusters corresponds to local maximum for the number of homologs in the same cluster relative to a randomized dataset. The randomization was performed by shuffling of the cluster memberships of the 'second' gene. This conserves the cluster structure.
- Families of homologs were constructed by characterizing each gene by the genes to which it is homologous (ignoring the level of homology). Similarity of genes was expressed in terms of the cooccurrence of homologs: ($S_{i,j} = a/(a+b+c)$ where a is the number of shared homologs, and b and c the number of non-shared homologs of i and j, respectively) and clustering was done by flexible linkage using a slight space contraction. By this method percolation of homologous families was minimized.

- Phylogenetic trees of homologous families were constructed using the Neighbor–Joining method [19] using either percentage identity or the exponent of the *P*-value as similarity criterion.
- Within gene families the comparison between the level of expression similarity and level of homology was done by comparing the tree-structure of the dendrogram of expression similarity and the phylogenetic tree based on homology. This was done by forcing the tree structure of the homology on the similarity data (and vice versa) and comparing the length of the tree with on the one hand the 'native' tree of the data and on the other hand the crosswise tree with randomized labels.

2.3 From expression (dis)similarity to homology and function of genes

2.3.1 Correlation between protein homology and gene expression

When we simply plot the amount of homology (expressed as percentage identity) *versus* similarity in expression profile (expressed as non-normalized correlation) (Fig. 1) we see that only for high levels of homology (> 75%) some correlation is detectable (the line is a fifth-order regression). But also for very high homologies there are genes whose expression is strongly negatively correlated. [23] analyzed this upper homology region and concluded that there is

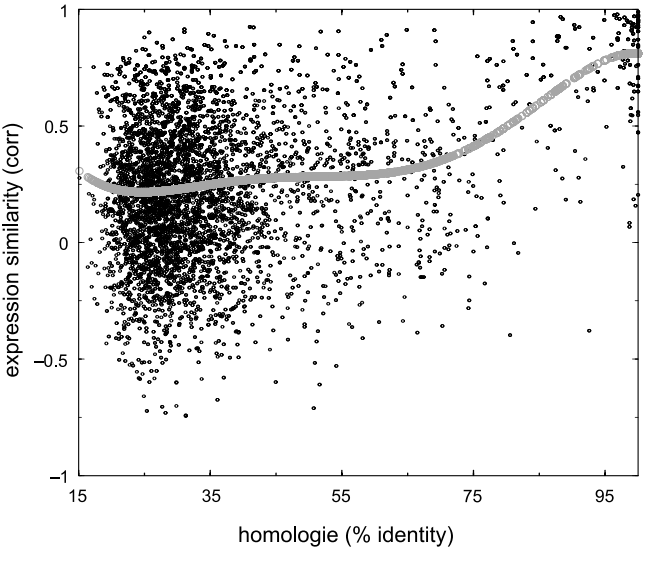


Fig. 1. Scatterplot of homology *versus* expression similarity. From the high-order regression it is seen that only for high homology values there is a weak correlation between homology and expression similarity. Note, however, that strong negative correlation in expression occurs also for very high identity values.

no significant correlation between similarity of gene-expression and homology even among the highly homologous genes. He interpreted this result in terms of the 'neutralist *versus* selectionist debate', and concluded that apparent neutral divergence of proteins may result from adaptive evolution because of changes in expression pattern. We will examine the full set of homologous pairs, and examine the divergence in sequence and expression of different functional classes to get more insight the process of individual-based and population-based divergence.

2.3.2 Expression patterns of homologous pairs

Fig. 2 displays the cluster membership of homologous pairs in 300 clusters. It is clear that, although there is significantly more cluster co-membership than in the randomized dataset, homologous pairs of genes in each cluster are found all over the dendrogram. A similar image is produced for smaller numbers of clusters, when limiting the set to relatively high homology, or to those for which homology is detected over more than 80% of the length of the protein (data not shown). Note that the genes which are excluded from the data (cluster 0) also have homologs in virtually all clusters.

In Fig. 3 we display the set of proteins which do have similar expression. The figure displays the clusters in the same format as Fig. 2, and indicates the annotation of the largest groups of genes which are co-expressed. The size of the protein family of the (first member of the) homologous pair, and how many of these are co-expressed in the same cluster is indicated. Moreover on the negative axis the level of homology of the pairs is plotted (within each cluster the pairs are ordered according to % identity). It is clear that whether we

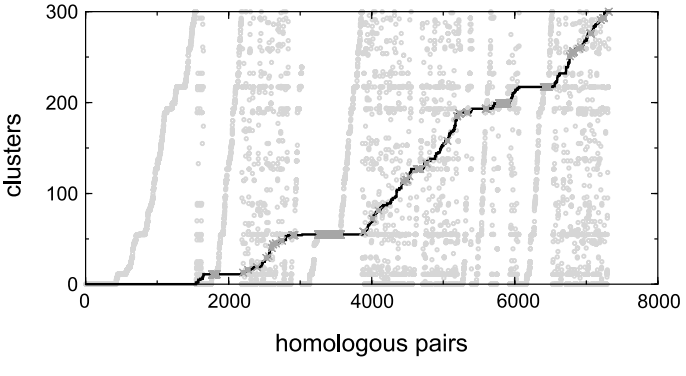


Fig. 2. Cluster membership of homologous pairs. Of the 7319 homologous pairs 786 occur in the same cluster. The black line indicates the cluster membership of the first protein of the pair, the dots the cluster membership of the second one. Grey crosses indicate where both occur in the same cluster. Cluster 0 indicates that the protein is not differentially expressed.

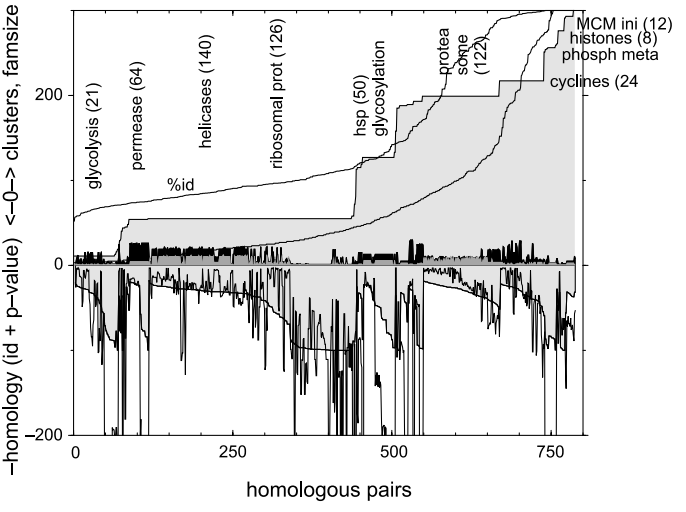
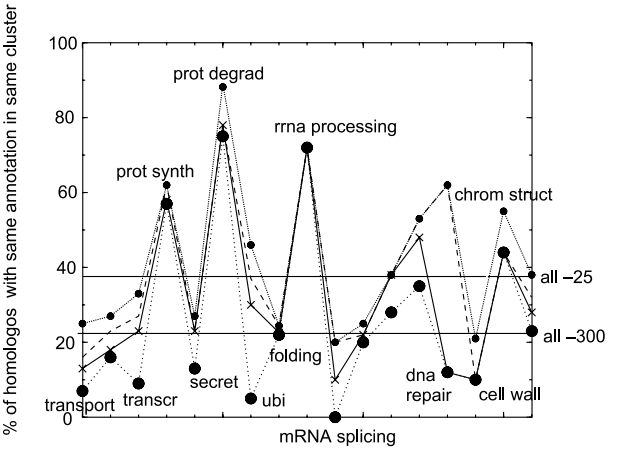


Fig. 3. Overview of homologous pairs with similar expression pattern, *i.e*. those occurring in the same expression cluster. Labels identify the major classes of the proteins belonging to this set. Clusters to which these pairs belong are plotted in the same way as in Fig. 2. Within each cluster the pairs are ordered according to % identity (shaded on the negative axis). p-value is also given (line on negative axis). The lines on the positive site show the distribution of % identity and p-value. The grey and black bars (positive, close to zero) indicate how many homologs each of the proteins has (black) and how many of these occur in the same cluster (grey), and hence are incorporated in this figure.

express homology in terms of % identity or in terms of p-value, co-expression occurs both for high and for low levels of homology. For example, ribosomal proteins occur in almost identical pairs, which are very similarly expressed. In the same cluster (which contains genes highly expressed for translation) a large class of homologous helicases are expressed also. In contrast to the ribosomal proteins these have low homology, and they have homologs at the other end of the dendrogram as well. Proteasome 20S subunits form a tight expression cluster, but they have only low levels of homology. On the other hand some members of the large family of (hexose) permeases are coexpressed, both with low and with high levels of homology, whereas many of the homologous pairs of this family have very divergent expression patterns (Figs. 2 and 7).

2.4 From annotation to expression similarity

We further investigate the relation between functional classes of proteins and the co-expression of homologs by selecting classes of homologous genes with the same process annotation. Clearly this analysis is confined to those functional classes which contain enough homologous genes. Fig. 4 shows



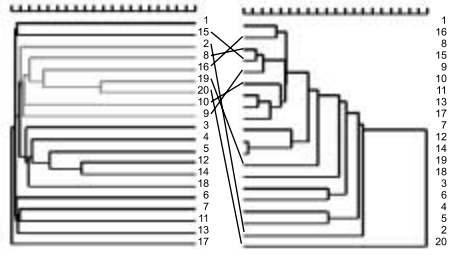
ANNOTATION	TOT	PAIRS	%	CLUS	%
TRANSPORT	1100	877	79	68	7
CELL CYCLE	517	149	28	25	16
TRANSCRIPTION	325	65	20	6	9
PROTEIN SYNTHESIS	411	250	60	145	57
SECRETION	336	59	17	8	13
PROTEIN DEGRADATION	278	149	53	113	75
PROTEIN DEGRADATION, UBI	159	112	70	6	5
PROTEIN FOLDING	121	44	36	10	22
RRNA PROCESSING	241	44	18	32	72
MRNA SPLICING	200	20	10	0	0
PROTEIN GLYCOSYLATION	93	70	75	14	20
CYTOSKELETON	128	21	16	6	28
DNA REPLICATION	123	45	36	16	35
DNA REPAIR	135	16	11	2	12
CELL WALL	190	19	10	2	10
CHROMATIN STRUCTURE	79	18	22	8	44
ALL	4598	2231	48	528	23

Fig. 4. Percentage of protein pairs with the same process annotation which occur in the same expression cluster (for 25, 50, 75 and 300 cluster). Note that for larger number of clusters the pattern of over and under represented groups becomes more pronounced.

the fraction of genes in these classes which are co-expressed for different clustering levels. This compared to the average value of co-expression over all the data. The order is based on the number of occurrences of the label in the full dataset. It is striking that when the number of clusters is increased the pattern becomes more pronounced. Above average co-expression occurs in protein synthesis and protein degradation, whereas below average co-expression occurs in proteins involved in internal regulatory processes (e.g. transcription factors, ubiquitination) or with external communication (e.g. transport). A similar analysis on gene annotation instead of process annotation, leads to similar conclusions: below average co-expression occurs for the various permeases, for protein kinases, Ub.-conjugating enzyme and transcription factors, whereas high co-expression occurs for RNA helicases, 20S and 26S proteasome subunits and translation elongation factors.

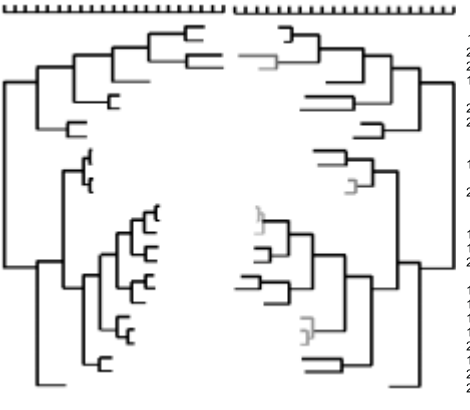
2.5 From homology to expression similarity

In Figs. 5 and 6 two examples are shown which compare phylogenetic tree structure with the clustering of gene expression patterns. In the case of RNA-helicases UPGMA clustering is used for both the phylogeny and the similarity dendrogram to highlight amount of divergence in the phylogeny. We see that the gene with the most divergent expression, codes for one of the pair of most homologous proteins. However the second most homologous pair has closely similar expression. In the case of hexose permeases and close homologs we see similar contrasts. Here we show the Neighbour Joining phylogeny with branch length according to homology (P-values) and according to expression similarity. Again we see that close homologs can have diverged in expression or that similar expression is maintained. Fig. 7 displays the rela-



1 12 YBR142W 20 MAK5 MRNA SPLICING RNA HELICASE DBP1
16 12 YNL112W 22 DBP2 MRNA DECAY RNA HELICASE DBP1
8 12 YGL078C 22 DBP3 RRNA PROCESSING RNA HELICASE DBP1
15 12 YLL008W 20 DRS1 RRNA PROCESSING RNA HELICASE DBP1
9 12 YGL171W 20 ROK1 RRNA PROCESSING RNA HELICASE DBP1
10 12 YHR065C 21 RRP3 RRNA PROCESSING RNA HELICASE DBP1
11 12 YJL033W 20 HCA4 RRNA PROCESSING RNA HELICASE DBP1
13 12 YKR024C 20 DBP7 RIBOSOME ASSEMBLY PUTATIVE RNA HELICASE
17 12 YNR038W 20 DBP6 RRNA PROCESSING RNA HELICASE, PUTATIVE
7 12 YFL002C 20 SPB4 RRNA PROCESSING, 25S RNA HELICASE
12 12 YJL138C 21 TIF2 PROTEIN SYNTHESIS TRANSLATION INITIATION FACTOR
14 12 YKR059W 21 TIF1 PROTEIN SYNTHESIS TRANSLATION INITIATION FACTOR
19 0 YOR204W 21 DED1 RNA PROCESSING ATP-DEPENDENT RNA HELICASE
18 6 YOR046C 20 DBP5 MRNA EXPORT RNA HELICASE DBP1
3 12 YDL084W 20 SUB2 MRNA SPLICING RNA HELICASE DBP1
6 48 YDR194C 20 MSS116 MRNA SPLICING, MITOCHOND RNA HELICASE
4 7 YDL160C 20 DHH1 TRANSCRIPTION RNA HELICASE DBP1 MRNA
5 1 YDR021W 20 FAL1 RRNA PROCESSING RNA HELICASE DBP1
2 42 YBR237W 21 PRP5 MRNA SPLICING RNA HELICASE DBP1
20 36 YPL119C 20 DBP1 MRNA PROCESSING RNA HELICASE DED1

Fig. 5. RNA-Helicases: comparison of homology and expression similarity. Left: dendrogram based on UPGMA of homology (p-value). Right: dendrogram of expression similarity. Of some proteines the correspondence is indicated with connecting lines.



1 21 YBR298C 24 MAL31 TRANSPORT MALTOSE PERMEASE
11 21 YGR289C 23 AGT1 TRANSPORT ALPHA-GLUCOSIDE
21 12 YML123C 22 PHO84 TRANSPORT INORGANIC PHOSPH
27 0 YPR198W 16 SGE1 CRYSTAL VIOLET RESISTANC
16 36 YIL171W 18 HXT12 TRANSPORT HEXOSE PERMEASE
8 12 YDR497C 25 ITR1 TRANSPORT INOSITOL PERMEASE
25 36 YOL103W 25 ITR2 TRANSPORT INOSITOL PERMEASE
20 0 YDL138W 26 RGT2 TRANSPORT GLUCOSE PERMEASE
3 38 YDL194W 25 SNF3 TRANSPORT GLUCOSE PERMEASE
4 33 YDL245C 27 HXT15 TRANSPORT HEXOSE PERMEASE
19 0 YJR158W 27 HXT16 TRANSPORT HEXOSE PERMEASE
9 0 YEL069C 27 HXT13 TRANSPORT HEXOSE PERMEASE
24 0 YNR072W 27 HXT17 TRANSPORT HEXOSE PERMEASE
5 21 YDR342C 27 HXT7 TRANSPORT HEXOSE PERMEASE
6 21 YDR343C 27 HXT6 TRANSPORT HEXOSE PERMEASE
12 21 YHR092C 27 HXT4 TRANSPORT HEXOSE PERMEASE
17 6 YJL214W 27 HXT8 TRANSPORT HEXOSE PERMEASE
20 5 YLR081W 26 GAL2 TRANSPORT GLUCOSE AND GALACT
7 21 YDR345C 27 HXT3 TRANSPORT HEXOSE PERMEASE
13 2 YHR094C 27 HXT1 TRANSPORT HEXOSE PERMEASE
14 33 YHR096C 27 HXT5 TRANSPORT HEXOSE PERMEASE
15 36 YIL170W 25 HXT12 TRANSPORT HEXOSE PERMEASE
18 36 YJL219W 26 HXT9 TRANSPORT HEXOSE PERMEASE
26 36 YOL156W 26 HXT11 TRANSPORT HEXOSE PERMEASE
10 36 YFL011W 27 HXT10 TRANSPORT HEXOSE PERMEASE
22 13 YMR011W 26 HXT2 TRANSPORT HEXOSE PERMEASE
23 36 YNL318C 26 NONE TRANSPORT HEXOSE PERMEASE

Fig. 6. Hexose permeases (and close homologs): comparison of homology and expression similarity. Left: phylogeny based on Neighbour joining method on homology (p-value). Right: same phylogeny topology with branch length based on expression similarity.

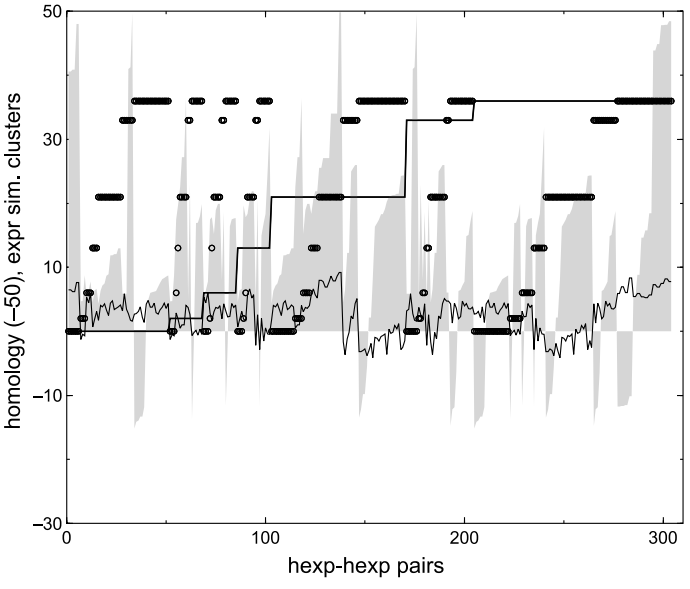


Fig. 7. Hexose perm – hexose perm pairs.

tionship between homology and expression within the hexose permease in a way similar to Figs. 2 and 3. From this representation we also see that co-expression of genes occurs for the full range of degree of homology, and so does differential expression. Intermediate homology is mostly differentially expressed.

From our branch-length statistics (see methods) (and similarly for tree-cutting methods) only protein kinases and hexose permeases emerge as having a higher correlation between homology and expression pattern than random. This result is in apparent contradiction to that obtained from the analysis of classes of genes with the same annotation: both emerged there as having below average co-epression of homologs. The difference is of course that here we look at degree of homology/similarity, whereas that analysis was based on qualitative annotation/homology and similarity. The two protein families are rather different. Whereas hexose permeases have relatively high homology, and homology over the entire sequence, the homology of protein kinases is much lower and the local alighment picks up only a small part of the molecule. In the case of the permeases very highly homologous genes are co-expressed. This seems to reflect recent duplication. Protein kinases evolve mainly by domain shuffling. It is interesting to note that when we base our phylogenetic tree on full alignments [22] the similarity between expression data and homology becomes less apparent.

3. Conclusion and discussion

Our basic question was to what extent homologous genes tend to do the same thing in the same environment, the same thing in different environments, or different things in the same or in different environments. Fig. 8 summarizes our results. All four possibilities occur.

- The 'same-same' category should be the usual one just after gene duplication when the regulatory region is duplicated as well. Moreover this 'undifferentiated' state appears to be sustainable, as mere bulk requirement of the gene product can favor multiple copies of the gene. This mode of 'speciation' has no counterpart in organism level speciation.
- Doing the 'same in different environments', *i.e.* divergence in expression rather than protein structure, appears to be a frequent mode of divergence and occurs especially in genes which communicate with the environment. As such these genes provide a type of 'individual based diversity' (*i.e.* multiple modes of interaction with the environment) which is rather similar to that mentioned above for RM systems.
- Divergence of genes without changing expression pattern is most prominent in homologous subunits of macro-molecular complexes, or otherwise closely cooperating proteins. They 'do the same differently'!
- Most homologs are dissimilar both in sequence and in expression pattern. They can have evolved along both routes, or simultaneously.

The present analysis is only a very crude initial attempt to map the fate of duplicated genes. Many caveats should be kept in mind. The expression data compiled by [3] are used for many initial attempts to analyze expression data. However, they are heavily biased towards cell cycle, with only sporulation and diauxic shift as other processes. The data also are noisy, and so is the clustering (especially when also lowly regulated genes are included). So we will have both missed differentiation in gene expression, and mis-identified noise as differential expression. Moreover, the analysis should be extended in two im-

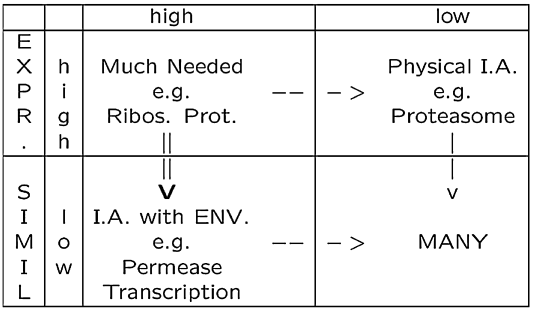


Fig. 8. Homology.

portant ways: including both homologs and expression data of other species, and including an analysis of upstream regions of the genes to identify the mutational 'cause' of differential expression. Obviously extension to multicellular organisms is an important next step: differential expression in that case can include even more 'doing the same same in different circumstances'.

In the introduction we discussed two modes in which organisms cope with a variable environment: individual based and population based diversity. In the first case gene regulation lets individuals cope with changes in the environment, in the latter case simply other individuals can take advantage of the changes. A so called 'red queen' evolution is a third mode: continuous evolutionary, rather than regulatory adaptation to the environment. Simulation models have suggested genomic adaptation to efficient evolutionary change [13, 16], Very interesting results on gene expression show that very few mutations can lead to changes in expression pattern of a whole battery of genes [4]. Moreover these data have shown that the expression changes due to evolutionary adaptation partially mimic expression changes due to regulatory adaptation. A great challenge for evolutionary theory is to understand how the build up of individual complexity, and therewith the potential for regulatory adaptation shapes the genotype-phenotype landscape, and therewith the potential for evolutionary adaptation.

References

1. M. A. Boerlijst and P. Hogeweg, Physica D **48** (1991) 17.
2. M. Eigen and P. Schuster, Naturwissenschaften **64** (1977) 541.
3. M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, Proc. Natl. Acad. Sci. USA **95** (1998) 14863 (see also: http://cmgm.Stanford.EDU/pbrown/).
4. T. L. Ferea, D. Botstein, P. O. Brown and R. F. Rosenzweig, Natl. Acad. Sci. USA **96** (1996) 9721.
5. W. M. Fitch, Annu. Rev. Genet. **7** (1973) 343.
6. W. Fontana and P. Schuster, Biophys. Chem. **26** (1987) 123.
7. W. Fontana and P. Schuster, Science **280** (1998) 1451.
8. P. Hogeweg, Comp. Biol. Med. **6** (1976) 199.
9. P. Hogeweg, Physica D **75** (1994) 275.
10. P. Hogeweg, Artif. Life **6** (2000) 85.
11. M. A. Huynen, J. Mol. Evol. **43** (1996) 165.
12. M. A. Huynen and P. Bork, Proc. Natl. Acad. Sci. USA **95** (1998) 5849.
13. M. A. Huynen and P. Hogeweg, J. Mol. Evol. **39** (1994) 71.
14. M. A. Huynen, P. F. Stadler and W. Fontana, Proc. Natl. Acad. Sci. USA **93** (1996) 397.
15. E. van Nimwegen, J. P. Crutchfield and M. Huynen, Proc. Natl. Acad. Sci. USA **96** (1996) 9716.
16. L. Pagie, Information integration in evolutionary processes, PhD Thesis, Utrecht University (1999).
17. L. Pagie and P. Hogeweg, J. Theor. Biol. **196** (1999) 251.
18. L. Pagie and P. Hogeweg, Bull. Math. Biol. **62** (2000) 759.
19. N. Saitou and M. Nei, Mol. Biol. Evol. **4** (1987) 406.

20. P. Schuster, W. Fontana, P. F. Stadler and I. L. Hofacker, Proc. R. Soc. London, Ser. B **255** (1994) 279.
21. B. Snel, P. Bork and M. A. Huynen, Nat. Genet. **21** (1999) 108.
22. J. D. Thompson, D. G. Higgins and T. J. Gibson, Nucl. Acids Res. **22** (1994) 4673.
23. A. Wagner, Proc. Natl. Acad. Sci. USA **97** (2000) 6579.
24. E. Zuckerkandl, J. Mol. Evol. **44** (1997) 470.