REVIEW

Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico

Jeroen Raes and Yves Van de Peer

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Ghent, Belgium

Abstract: Duplication of genes increases the amount of genetic material on which evolution can work and has been considered of major importance for the development of biological novelties or to explain important transitions that have occurred during biological evolution. Recently, much research has been devoted to the study of the evolutionary and functional divergence of duplicated genes. Since the majority of genes are part of gene families, there is considerable interest in predicting differences in function between duplicates and assessing the functional redundancy of genes within gene families. In this review, we discuss the strengths and limitations of both older and novel approaches to investigate the evolution of duplicated genes in silico.

Keywords: gene duplication, functional divergence, positive selection, substitution rates

Introduction

In his now classic book *Evolution by Gene Duplication*, published in 1970, Ohno claimed that if evolution had been entirely dependent on natural selection, from a bacterium only numerous forms of bacteria would have emerged, while big leaps in evolution would have been impossible without the creation – through duplication – of many new gene loci with previously nonexistent functions. During the last few decades it became clear that, from an evolutionary point of view, most genes are indeed not unique but are part of larger families of related genes. These gene families have originated by duplication of an ancestral gene, after which these duplicated genes in turn have duplicated. It is now generally believed that extensive gene duplication has been responsible for increased genomic and phenotypic complexity (eg Aburomia et al 2003; Meyer and Van de Peer 2003).

Although there is some evidence that gene duplication is a continuous and very frequently occurring process (Lynch and Conery 2000; Gu et al 2002), more and more genomic data seem to suggest that many duplicates have been formed during some major, large-scale gene duplication events. Entire genome duplication events have been postulated for (members of) the three major eukaryotic kingdoms. On the basis of a genome-wide analysis of the yeast *Saccharomyces cerevisiae*, Wolfe and Shields (1997) postulated a duplication of the entire yeast genome about 100 million years ago, although this event was dated much earlier (200–300 million years ago) by others (Friedman and Hughes 2001). About 13% of the yeast genome still consists of duplicated genes, resulting from this polyploidy event (Seoighe and Wolfe 1999).

For animals, the first indications about large-scale duplications early in the vertebrate lineage were found by the analysis of *Hox* genes (Holland et al 1994). *Hox* genes encode DNA-binding proteins that specify cell fate along the anterior–posterior axis of bilaterian animal embryos and occur in one or more clusters of up to 13 genes per cluster (Gehring 1998). It is thought that the ancestral *Hox* gene cluster arose from a single gene by a number of tandem duplications. The observation that protostome invertebrates, as well as the deuterostome cephalochordate *Amphioxus*, possess a single *Hox* cluster, while Sarcopterygia, a monophyletic group including lobe-finned fish such as the coelacanth and lungfishes, amphibians, reptiles, birds and mammals have four clusters (Holland and Garcia-Fernandez 1996; Holland 1997), supports the hypothesis of 2 rounds (2R) of entire genome duplications early in vertebrate evolution. Additional support comes from the detection and dating of duplicated blocks in the human genome (McLysaght et al 2002), large-scale phylogenetic analysis of gene families (Gu et al 2002) and analysis of gene clusters

Correspondence: Yves Van de Peer, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium; tel +32 9 331 3807; fax +32 9 331 3809; email yves.vandepeer@psb.ugent.be

© 2003 Open Mind Journals Limited. All rights reserved.

such as the major histocompatibility complex (MHC) region (Spring 1997; Abi-Rached et al 2002). However, in general, strong evidence in favour of 2R is hard to find and the 2R hypothesis is still vigorously debated (Furlong and Holland 2002; Larhammar et al 2002; Spring 2002; Friedman and Hughes 2003).

A few years ago, 'extra' *Hox* gene clusters were discovered in fish. Amores and co-authors (1998) described the existence of seven *Hox* clusters in zebrafish (*Danio rerio*), and additional *Hox* clusters have also been described for medaka (*Oryzias latipes*) (Naruse et al 2000), the African cichlid fish *Oreochromis niloticus* (Málaga-Trillo and Meyer 2001) and the pufferfish *Fugu rubripes* (Aparicio et al 1997). All these data strongly point to an additional *Hox* cluster duplication in ray-finned fishes that occurred before the divergence of zebrafish, medaka and pufferfish, at least 100 million years ago (Nelson 1994). Furthermore, mapping data suggest that duplications are not limited to *Hox* clusters, and that large chromosome segments or entire chromosomes are duplicated (Amores et al 1998; Force et al 1999; Postlethwait et al 2000; Woods et al 2000). In the meantime, many other multigene families have been described that have more genes in fish than in other vertebrates (Wittbrodt et al 1998; Postlethwait et al 2000; Taylor, Van de Peer, Braasch et al 2001; Taylor, Van de Peer, Meyer 2001). Moreover, tree topologies clearly support a fish-specific genome duplication that has occurred early in the evolution of ray-finned fishes (Taylor et al 2003; Van de Peer et al 2003).

In plants, early analyses based on the (at that time) unfinished genome sequence of *Arabidopsis thaliana* showed that large-scale gene duplication, probably a complete genome duplication, occurred in the evolution of this model plant (eg Terryn et al 1999; Blanc et al 2000; Paterson et al 2000), an opinion later shared by the Arabidopsis Genome Initiative (AGI 2000). Vision et al (2000) investigated this genome duplication by considering large regions ('blocks') of genes that showed statistically significant colinearity with other regions in the genome. They could reject a single genome duplication event because of the discovery of many overlapping blocks, a phenomenon that can be attributed only to multiple duplication events. By dating these duplicated blocks, these authors postulated up to four different large-scale gene duplication events, ranging from 50 to 220 million years ago. One of these classes, dated approximately 100 million years ago, grouped nearly 50% of all the duplicated blocks, suggesting a complete genome duplication at that time (Vision et al 2000). However, the dating methods used in this study were later criticised (Wolfe 2001; Raes et al 2003). A recent reanalysis of the *Arabidopsis thaliana* genome by Simillion et al (2002) considered heavily degenerated block duplications. These ancient duplicated blocks can no longer be recognised by directly comparing both segments because of differential gene loss, but can still be detected through indirect comparison with other segments. When these so-called hidden duplications are taken into account to describe the duplication landscape in *Arabidopsis*, many homologous genomic regions can be found in five to eight copies, suggesting three polyploidisation events in the evolutionary past of *Arabidopsis thaliana*. Furthermore, about 28% of the genes in *Arabidopsis* are retained duplicates, resulting from these ancient large-scale gene duplication events, the youngest one estimated to have occurred about 75 million years ago (Simillion et al 2002).

Evolution of novel gene functions

Large-scale gene or entire genome duplication events such as those described above have been considered very important for biological evolution because they provide a way to greatly increase the amount of genetic material on which evolution can work (Ohno 1970; Holland et al 1994; Sidow 1996; Prince and Pickett 2002; Holland 2003). Indeed, since duplicated genes are redundant, one of the copies is, at least in theory, freed from functional constraint and can therefore evolve a new function. The classical model, put forward by Ohno (1970), predicts that mutations in the second copy are selectively neutral and will either turn the gene into a non-functional pseudogene or, alternatively, turn the duplicate gene into a gene with a new function, due to a series of non-deleterious random mutations. This model of gene evolution has been widely adopted as an explanation for the evolution of novel genes and gene functions but has been criticised, mainly because little evidence has been found for genes that have obtained novel functions this way. Several alternative models for gene evolution after duplication events have been proposed (Hughes 1994, 1999; Walsh 1995; Nowak et al 1997; Gibson and Spring 1998; Wagner 1998; Force et al 1999). For example, Hughes (1994) and Force et al (1999) argue that when a gene with multiple functions is duplicated, the duplicates are redundant only for as long as each retains the ability to perform all ancestral roles. When one of the duplicates experiences a mutation that prevents it from carrying out one of its ancestral roles, the other duplicate is no longer redundant. According to Force and colleagues' (1999) 'duplication–degeneration–complementation' (DDC) model, degenerative mutations preserve rather than destroy duplicated genes but also change their functions – or at least restrict them – to become more specialised. Gibson and Spring (1998) have argued that alteration of a single domain in a multidomain protein might lead to non-functional complexes that exhibit a so-called 'dominant-negative phenotype'. Their model is based on the observation that, for several genes, point mutations lead to a much more severe phenotype than when the (duplicated) gene is simply knocked out. In this case, one would expect selection against deleterious point mutations resulting in the retention of the gene. As a matter of fact, the gene is not only retained, it is also prevented from diverging too much. Although these models explain gene retention rather than gene evolution, keeping the genes around increases the chance for functional divergence later on, for example, by positive selection (eg Zhang et al 1998; Duda and Palumbi 1999; Hughes et al 2000) or subfunctionalisation (Li 1980; Piatigorsky and Wistow 1991; Hughes 1994; Force et al 1999; Stoltzfus 1999; Wagner 2002). Likewise, processes such as gene conversion might also retard the functional divergence of duplicated genes, while at the same time preventing pseudogenisation of a redundant copy (Li 1997).

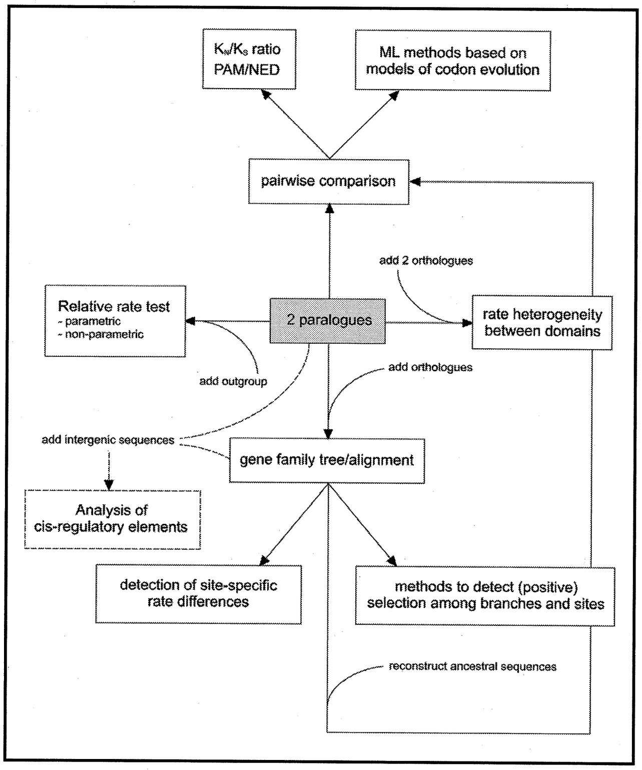
K_N/K_S ratio
PAM/NED
ML methods based on models of codon evolution
pairwise comparison
add 2 orthologues
Relative rate test
- parametric
- non-parametric
2 paralogues
rate heterogeneity between domains
add outgroup
add orthologues
add intergenic sequences
gene family tree/alignment
Analysis of cis-regulatory elements
detection of site-specific rate differences
methods to detect (positive) selection among branches and sites
reconstruct ancestral sequences

Figure 1 Overview of the different in silico approaches to study possible functional divergence at the coding level between two duplicated genes. Simple approaches are often based on the comparison of only two paralogues, while more sophisticated analyses are usually based on a larger collection of sequences. See text for details. ML = maximum likelihood; NED = neutral evolutionary distance; PAM = point accepted mutations.

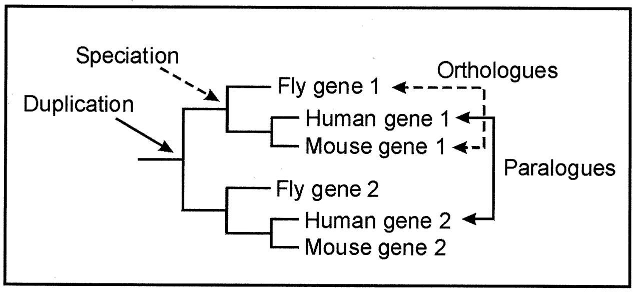

Figure 2 Hypothetical tree depicting a duplication event, followed by speciation. Paralogues arise through duplication (full arrows), while orthologues arise through speciation (dashed arrows).

In this review, we discuss some older and novel in silico approaches to study the evolution of duplicated genes, mainly focusing on the coding part of the gene, in order to find traces that might imply functional divergence after duplication. Figure 1 summarises these different approaches, starting from two paralogues (Figure 2), but extending the set of sequences according to the method used.

Detecting functional divergence

Relative-rate tests

One of the simplest ways to study the evolution of duplicated genes is to investigate whether one of the duplicates has evolved at a faster rate after duplication, compared to a reference or outgroup sequence, using a so-called relative-rate test (Margoliash 1963; Sarich and Wilson 1973). An increase in the rate of evolution could be explained by relaxed functional constraints eventually turning one of the duplicates into a pseudogene, due to accumulation of deleterious mutations. On the other hand, an increase in rate could also point to positive selection by which the gene evolves a new function. In general, relative-rate tests can be divided into two main categories: parametric and non-parametric. Parametric rate tests use a model of evolution to account for multiple substitutions, in order to compute branch lengths more accurately. To this end, many alternatives and improvements have been proposed over the years, using distance (eg Wu and Li 1985; Takezaki et al 1995; Robinson et al 1998) and likelihood (eg Felsenstein 1988; Muse and Weir 1992) approaches. Non-parametric tests have the advantage that they are not influenced by the choice of a, possibly wrong, substitution model (Nei and Kumar 2000). The non-parametric rate test of Tajima (1993) compares two sequences with an outgroup sequence and counts the number of unique substitutions in both lineages. When both genes evolve under the molecular clock model (Zuckerkandl and

Pauling 1965), both genes are expected to have accumulated a similar number of 'unique' substitutions. On the other hand, when one of the duplicates has accumulated a significantly larger number of substitutions, the molecular clock does not apply and one of the paralogues is inferred to have experienced an increased evolutionary rate.

In several studies rate differences between duplicates have been investigated. Hughes and Hughes (1993) did not detect any significant rate differences when investigating 17 recently duplicated genes in the tetraploid frog *Xenopus laevis*. Cronn et al (1999) compared 16 paralogous loci in allotetraploid cotton and did not detect any significant rate difference after duplication, except for one locus, where pseudogenisation of one of the duplicates after the alloploidy event was suspected. In a study of 19 gene families in fish and mammals, Robinson-Rechavi and Laudet (2001) detected four families with a significant rate difference between duplicates. Kondrashov et al (2002) analysed 101 paralogous pairs in prokaryotes and eukaryotes and found about five with a significant rate difference. L Zhang et al (2002) recently compared rates of 105 duplicated gene pairs on chromosomes 2 and 4 of *Arabidopsis thaliana*. Only three of these showed a significant rate difference after duplication at the protein level. In conclusion, according to most of the studies only a very small fraction of the duplicates show an increase in evolutionary rate after duplication, possibly pointing to relaxed functional constraints or positive selection. One of the few studies contradicting this finding was performed by Van de Peer et al (2001), who examined 26 anciently duplicated genes in zebrafish and observed an accelerated rate in about half of the duplicates, using a non-parametric rate test. However, only 2 of 14 duplicated fish genes from the study of Robinson-Rechavi and Laudet showed an accelerated rate. The drawback of both studies is, as with others (see above), the small number of duplicates investigated. Furthermore, the selection of genes might have been biased. For example, the majority of genes investigated by Van de Peer et al (2001) are transcription factors. Whether this bias is responsible for the high fraction of duplicates that evolve at unequal rates remains to be investigated.

Detecting positive selection

A second way to study the evolution of genes after duplication in silico is to compare the rate of nonsynonymous substitutions, ie substitutions leading to amino acid replacements (K_N), with the rate of synonymous substitutions, ie substitutions that do not lead to amino acid replacement (K_S). The ratio of these two values, called ω, provides a measure for the selection pressure on the protein product of a gene. A value of $\omega < 1$ indicates purifying or negative selection that keeps the amino acid sequence from changing, since most amino acid changes are disadvantageous, while $\omega = 1$ indicates neutral evolution (Kimura 1983). When $\omega > 1$, this implies that natural selection favours amino acid replacements, and as a result nonsynonymous substitutions are fixed at a higher rate than synonymous substitutions. A value for ω significantly greater than 1 can thus be an indication for the evolution of the gene towards a new function.

To estimate the number of nonsynonymous and synonymous rates, different approaches exist. In general, these can be divided into two classes: approximate (counting) methods, which estimate K_S and K_N for pairs of sequences; and 'maximum likelihood' methods, which are usually based on an explicit codon-substitution model, using a multiple sequence alignment and a phylogenetic tree. Approximate methods are based on counting the number of observed nonsynonymous and synonymous substitutions per nonsynonymous and synonymous site, after which a correction for multiple substitutions is applied. The simplest methods, such as the one of Nei and Gojobori (1986), assume equal nucleotide frequencies and no bias in the direction of change, while others take into account different rates of transitions and transversions (Li et al 1985; Li 1993; Pamilo and Bianchi 1993; Comeron 1995; Ina 1995). A recently developed method also compensates for codon bias and unequal nucleotide frequencies (Yang and Nielsen 2000).

The first maximum likelihood methods using explicit codon substitution models that allowed estimation of K_N and K_S were developed in 1994 (Goldman and Yang 1994; Muse and Gaut 1994). These methods take into account biases in codon usage, base frequency and transition/transversion ratio. Furthermore, the likelihood framework has the advantage of providing a statistical test to determine whether K_N is significantly higher than K_S. Using a likelihood ratio test (LRT), one can compare the likelihood values under two hypotheses: in this case H_0 where ω is fixed to one, and H_1 where ω is estimated as a free parameter. The rejection of the null model in the LRT, combined with an estimation of $\omega > 1$, indicates positive or adaptive selection (Yang and Bielawski 2000).

Although different methods have been developed to detect positive selection based on ω, it must be noted that the ratio of nonsynonymous over synonymous mutations can be used to detect positive selection only for recently duplicated genes. Once the gene has adapted to its specific function, purifying selection is expected to predominate, allowing the number of synonymous substitutions per site to catch up and eventually exceed the number of nonsynonymous substitutions per site (Hughes 1999; Nei and Kumar 2000; see section titled, The episodic nature of selection).

Using the methods described above, several examples of positive selection have been described in duplicated genes such as the primate ribonuclease (Zhang et al 1998; Zhang J et al 2002), mammalian immunoglobulin (Tanaka and Nei 1989), pregnancy-associated glycoprotein (Hughes et al 2000) and gastropod conotoxin genes (Duda and Palumbi 1999). A more extensive overview of genes (paralogues as well as orthologues) for which positive selection has been detected can be found in Yang and Bielawski (2000).

On the other hand, several large-scale analyses showed that functional divergence through positive selection was not as ubiquitous as previously thought. Hughes and Hughes (1993) detected no positive selection in their analysis of 17 duplicated genes of *Xenopus laevis*, using the method of Nei and Gojobori (1986). Lynch and Conery (2000) observed 328 duplicated pairs with $\omega > 1$ in a maximum likelihood analysis (Goldman and Yang 1994) of 9870 pairs in several different eukaryotes. L Zhang and co-workers (2002), using the same technique, did not detect any genes under positive selection among 242 duplicated gene pairs on chromosomes 2 and 4 in *Arabidopsis thaliana*. Kondrashov and co-workers (2002) found that the large majority of duplicates are under purifying selection, using the method of Pamilo and Bianchi (1993) and Li (1993) in an analysis of 4233 recently duplicated gene pairs in 26 bacterial, 6 archaeal and 7 eukaryotic genomes. Studies looking for positive selection without restricting it to paralogues had also only limited success. Endo et al (1996) applied the Nei and Gojobori (1986) test on 3595 groups of homologous genes and found only 17 groups of genes to have been under positive selection (with $\omega > 1$ for a majority of all pairwise comparisons within a group). Sharp (1997), comparing 363 pairs of genes in mouse and rat, found only one gene, ie interleukin-3, with $\omega > 1$.

The question remains whether positive selection is more rare than expected, or whether the developed methodologies are often incapable of reliably detecting it. At least in one case, the shortcomings of the $\omega > 1$ test to detect positive selection were clearly demonstrated. In a two time-point study on HIV drug resistance, Crandall and co-workers (1999) analysed differences in ω for the protease gene in eight patients using the Nei and Gojobori (1986) method. They showed that in only two cases could positive selection be detected, while parallel adaptive substitutions leading to drug resistance were observed in five of eight patients.

Problems in detecting positive selection

Sequence bias

A first problem in detecting positive selection is that the estimation of K_N and K_S is influenced by sequence composition (eg GC content) and codon biases (Smith 1994). Several analyses discussed above used a simple method that does not compensate for biases in sequence content. More complex methods try to account for these biases and allow for, in general, more accurate estimations of ω (Bielawski et al 2000).

The episodic nature of selection

Another problem is that positive selection is of an episodic nature, which means that, after a period of positive selection, purifying selection usually blurs the substitution pattern indicative of positive selection (Hughes 1999; Nei and Kumar 2000). As a result, positive selection can no longer be detected 30–50 million years after gene duplication by using the ratio of K_N over K_S (Hughes 1999; Hughes et al 2000). To address this problem, three approaches have been used. A first approximate method evaluates whether nonsynonymous mutations occur in such a way as to change protein charge or polarity to a greater extent than is expected under random substitution. This method involves computation of the proportion of radical nonsynonymous differences (p_{NR}) per radical nonsynonymous site versus the proportion of conservative nonsynonymous differences per conservative nonsynonymous site (p_{NC}). When $p_{NR} > p_{NC}$, nonsynonymous differences occur in such a way as to change the property of interest to a greater extent than expected at random (Hughes et al 1990). Since this method looks at nonsynonymous sites only and the resulting amino acid changes, the occurrence of positive selection should be evident for a much longer period. It should be noted though, that this method may be less sensitive to detecting positive selection than looking at the K_N/K_S ratio (Vacquier et al 1997; Hughes 1999). Furthermore, a recent study showed that this measure is heavily influenced by the transition/transversion ratio and amino acid composition of the investigated sequences (Dagan et al 2002). Therefore, inferences on positive selection based on this method should be treated with caution.

The second strategy is based on the reconstruction of ancestral sequences at the internodes of the phylogenetic tree. Given a substitution model and a tree topology, ancestral sequences can be inferred through a variety of parsimony

(Eck and Dayhoff 1966; Fitch 1971; Maddison and Maddison 1992; Swofford 2002), distance (Zhang and Nei 1997), maximum likelihood (Schluter 1995; Yang et al 1995; Koshi and Goldstein 1996; Pagel 1999; Pupko et al 2000, 2002) and hierarchical Bayesian approaches (Huelsenbeck and Bollback 2001). By comparing these ancestral sequences, ω can be measured along a specific branch (between two ancestral nodes, or an ancestral node and an endnode) on the tree, corresponding with a more specific period in evolution. Although not explicitly looking at duplicated genes, Liberles et al (2001) detected about 4% of 8690 chordate and embryophyte gene families investigated to have at least one branch in which $\omega > 1$ using this approach.

A third strategy relies on the above-mentioned maximum likelihood approach using codon models, which allow for ω to vary among branches of the tree. Using an LRT, one can compare the likelihood values under two hypotheses: in this case H_0 where ω is fixed, and H_1 where ω is estimated as a free parameter for a specific branch or branches. If ω is estimated to be >1 for the chosen branch(es) and the LRT gives a significant result, this is indicative for positive selection in that branch (Yang 1998). This technique was successfully applied to duplicated ribonuclease genes, thereby confirming earlier results (Bielawski and Yang 2003).

Positive selection acts locally

Another major reason that might explain the low prevalence of detectable positive selection lies in the fact that, in general, ω is measured as an average over all sites of a gene. This implies that, if only a fraction of sites is under positive selection, their detection is complicated. Not all amino acids of a protein are functionally important and therefore these can evolve in a more neutral way, while others do have important structural and functional roles and are under strong purifying selection. One can imagine that after duplication for example, only the domains involved in substrate binding specificity are under positive selection, while all the other sites retain their original evolutionary rates, obscuring the former sites when looking at the K_N/K_S ratio for the gene as a whole. For example, Hughes and Nei (1988) detected ω values >1 in the antigen recognition region of the MHC, while other regions of the genes had values for ω of less than 1. Endo and co-workers (1996) recognised the possibility of region-restricted positive selection as well, and also used a sliding window method to look for evidence of positive selection, to avoid averaging over the entire gene, an approach also followed by Duda and Palumbi (1999). Fares and co-workers (2002) further improved this kind of approach by estimating the appropriate window size and by detecting saturation at synonymous sites.

Positive selection can also be limited to a few dispersed amino acids. For this reason, methods were developed that allow detection of positive selection at single amino acid sites. One method is based on inferring ancestral sequences for a given tree topology and testing neutrality ($\omega = 1$) for each codon site using the numbers of synonymous and nonsynonymous changes detected throughout the tree. Using this method, positive selection on specific sites of the human leucocyte antigen (HLA) gene was detected, yielding two new putative antigen recognition sites (Suzuki and Gojobori 1999). This method is now also implemented in a publicly available software package for UNIX® called ADAPTSITE (Suzuki et al 2001).[1] Another application of a similar technique was described by Bush et al (1999), who examined positive selection in individual codons for the H3 haemagglutinin gene of the human influenza virus A.

In addition, maximum likelihood models were developed that allow for heterogeneous selection pressure among sites. They also allow hypothesis testing as described above, using classes of sites that have different values of ω. Models implementing discrete as well as continuous (gamma, beta) ω distributions are provided. For example, one can compare (using an LRT) a model in which sites have a continuous distribution of ω values between 0 and 1 with a model having one extra class of sites in which ω is freely estimated. If the LRT is significant and sites in the extra class have an $\omega > 1$, positive selection on a subset of sites is assumed. This method allowed the detection of positive selection in several genes, where earlier methods had failed (Nielsen and Yang 1998; Yang et al 2000). Using a Bayesian approach, the posterior probability for each site to belong to a class of ω values can be calculated, and as a consequence the sites under positive selection can be identified (Nielsen and Yang 1998; Yang et al 2000).

Recently, methods have been developed to combine detection of lineage- and site-specific positive detection (Yang and Nielsen 2002). As in the lineage-specific methods, a branch can be selected, for which positive selection should be tested (the so-called 'foreground' branch). All other branches are referred to as 'background' branches. Two models were developed. The first (referred to as the 'A' model) is based on four classes of sites: namely, two classes containing sites with $\omega_0 = 0$ (class 0) or $\omega_1 = 1$ (class 1), representing sites that are not under positive selection; and two classes allowing (background) sites of the ω_0 and ω_1 class to change to a third (estimated) $\omega_2 > 1$ in the foreground branch, respectively (sites going from purifying to positive selection [$\omega_0 \rightarrow \omega_2$] in class 2 and sites going from neutral evolution to positive selection [$\omega_1 \rightarrow \omega_2$] in class 3). The second (B) model allows also for sites under positive selection in the background lineages, as ω_0 and ω_1 are estimated freely over the entire phylogenetic tree. These models have been applied successfully to detect positive selection after gene duplication in the phytochrome, troponin C and chalcone synthase gene families, for which the previous models did not detect positive selection (Yang and Nielsen 2002; Yang et al 2002; Bielawski and Yang 2003). A new model is currently under development, which is less restrictive and allows a class of sites with two independent estimations of ω for the two branches following the duplication event, in order to model site-specific divergence in selective pressure following duplication. This model further refines the possibilities of the previous ones and has been successfully applied to a number of gene families (Joseph P Bielawski 2002, pers comm, Oct). These recent models, together with the Bayesian identification of sites under positive selection, are very promising and are expected to allow very detailed study of functional divergence after duplication.

All maximum likelihood approaches using codon models described above are implemented in the PAML package (Yang 1997), which is publicly available for UNIX®, Windows® and Apple® Macintosh® operating systems.[2]

One of the most recent developments is the use of 'stand-alone' Bayesian approaches to detect positively selected mutations at specific sites and lineages. Nielsen (2002) and Nielsen and Huelsenbeck (2002) developed a method based on mapping mutations on the phylogenetic tree, which gave similar results to the Yang and Nielsen (2002) maximum likelihood approach. However, this approach allows further exploration of the evolutionary history of the investigated genes. As an example, they showed, rather unexpectedly, that in the influenza haemagglutinin protein, positively selected amino acid changes tended to be mostly conservative, instead of the expected radical substitutions.

Other methods to detect functional divergence

Several methods have been developed to detect functional divergence after duplication on the premise of rate shifts at specific positions or regions of the protein. It is postulated that when new functions are acquired by amino acid substitutions, the selective constraints upon these positions will also change, which in turn will lead to a difference in substitution rate at these sites (the so-called type I functional divergence) (Gu 1999). One of the first methods to detect such rate changes was developed by Gu (1999, 2001) and uses a coefficient of statistical divergence (θ_λ) to measure the functional divergence between two paralogous clusters of a tree. θ_λ is defined as the decrease in rate correlation between the two clusters and was initially estimated using a simple algorithm based on a Poisson model of molecular evolution. Gu also developed a probabilistic model with two possible states for each site: S_0, when the site is 'functional-divergence-unrelated', meaning that the evolutionary rate of that site is the same between two clusters; and S_1 ('functional-divergence-related'), when there is no rate correlation between clusters and altered functional constraints are hypothesised. In this model, θ_λ can be interpreted as the probability $P(S)$ of a site being in the 'functional divergence state'. Using a maximum likelihood approach, θ_λ and the other parameters of the model (the gamma shape parameter α and branch lengths) are estimated, after which an LRT can be used to discern between the null hypothesis that there is no rate difference between the same sites of two clusters ($H_0\colon \theta_\lambda = 0$) and the alternative hypothesis $H_1\colon \theta_\lambda > 0$. The method also allows analysis of three or more clusters at the same time and incorporates a Bayesian approach to predict sites that are likely to be responsible for the functional divergence. It was successfully applied to several vertebrate gene families (for an overview see Gaucher et al 2002). In addition, methods to detect type II functional divergence are proposed. In type II divergence, there is no detectable rate difference between clusters, but sites have functionally diverged shortly after duplication at certain sites, resulting in radical amino acid property differences at these positions between clusters, although the functional constraint (which is reflected by the evolutionary rate) became similar again as soon as these changes had occurred (Gu 2001). The algorithms were recently embedded in a software package called DIVERGE, featuring a graphical user interface for Windows® and Linux® operating systems.[3] This program also allows mapping of these sites on a 3-D structure, if available, to facilitate the understanding of the functional importance of discovered critical sites (Gu and Vander Velden 2002).

Gaucher et al (2001) used statistical quantiles to detect functionally important sites in elongation factors by comparing the bacterial EF-Tu proteins with their eukaryotic (and functionally diverged) EF-1α counterparts. Sites that had a rate difference between the two groups of more than 2 standard deviations from the mean in the distribution of rate differences per site were considered to be candidate sites responsible for the difference in function. Subsequently, they mapped these positions on the known tertiary structure of these proteins. By correlating this position with the known functional divergence of the proteins, they were able to propose putative functions (eg tRNA and cytoskeleton interaction) for these sites.

Liberles (2001) proposed two alternative measures of adaptive evolution. The first method consists of calculating the ratio between the number of point accepted mutations (PAM) and the neutral evolutionary distance (NED) (Peltier et al 2000). The latter distance is based on the proportion of conserved twofold degenerate codons. These codons are chosen because the differences between each of these codons are represented solely by transitions at the third codon position (Peltier et al 2000), making the NED more clock-like than K_S, where transitions and transversions, which occur with different probabilities, are considered. Nevertheless, in general, it is expected that PAM/NED ratios are similar to K_N/K_S ratios, as also observed by Liberles (2001). The second method, the sequence space assessment (SSA) statistic, measures the fraction of amino acid sites that have undergone substitution along a certain branch, compared with the total number of sites that are variable at one or more branches in the tree (normalised for the number of taxa).

Dermitzakis and Clark (2001) modified a method designed by Tang and Lewontin (1999) that measures within-protein rate heterogeneity in duplicated genes. This method, called the paralogue heterogeneity test, was developed particularly to detect subfunctionalisation (see Introduction) at the protein domain level. In other words, it detects whether in one paralogue, one region of the protein has evolved more rapidly than that same region in the other paralogue. The method works by comparing each paralogue to a respective orthologue (Figure 1) using an approach where a Q-value is measured for each site in the alignment. This Q-value is a measure for the density of sequence variability in the region around that site. By comparing the Q-values of both paralogues, regions that differ in variability can be determined. The software tools also contain a script to perform randomisation tests to calculate the significance of the obtained results. The authors applied their method to several mouse and human gene families and detected several cases in which two regions of a protein evolved at a different rate in two paralogues, which may point to subfunctionalisation. A similar method, using user-defined regions, was also described by Marín et al (2001).

Functional divergence at the regulatory level

Although this review focuses on the analysis of the protein-coding part of a gene, novel gene functions arise not only by modification of the coding region, but also by changing its expression. As the expression of genes is, at least partly, dependent on the presence of transcription factor binding sites in regulatory regions, mutations in these elements can alter the expression domain of genes. For example, subfunctionalisation has been proposed to act mainly at the regulatory level, where the reciprocal loss of different regulatory elements can lead to functional divergence through expression in, for example, different organs or stages of development (Force et al 1999). The in silico investigation of promoter regions of duplicated genes should allow the evolution of transcriptional control after duplication to be unravelled. The most straightforward approach would be to align promoters using standard alignment tools, and look for patterns of loss and gain of regulatory motifs. Unfortunately, these alignment methods are rather rigid and when, for example, the motif position or order is changed, or sequences are too divergent, methods based on sequence alignment have serious difficulties aligning homologous regulatory regions. New techniques such as the detection of over-represented motifs by word counting or probabilistic methods and especially methods such as phylogenetic footprinting, which take into account the phylogenetic relationships of genes, do consider this dynamic nature of promoters and allow investigation of whether loss or gain of certain regulatory motifs might have led to the functional divergence of duplicated genes. Nevertheless, although recent developments seem promising, unambiguous identification of regulatory elements is generally far from straightforward. The delineation of promoters is even harder, owing to its complex nature, and in silico promoter prediction is still in its infancy (Rombauts et al 2003).

Conclusions

The function of a gene is usually determined by a rather complex combination of the three-dimensional structure of the protein it encodes and its spatio-temporal expression determined by its *cis*-regulatory elements. In addition, other processes such as post-transcriptional and post-translational modifications, transport and cellular context also play an important role in the definition of a gene's function. Duplicated genes provide an excellent tool to study gene function and functional divergence. After duplication, one gene copy is redundant and, freed from functional constraint, can evolve a new function. Numerous models have been put forward to explain the retention and functional divergence of genes, and the study of these processes, bringing together fundamental evolutionary research and more applied functional genomics, has now become a rapidly growing field of research. Although the in silico determination of functional difference between two duplicated genes is inevitably compromised by the complex nature of what defines a gene's function, much progress has been made in the last few years and many novel approaches have become available, as discussed here, to study the functional diversification of genes. By formulating testable working hypotheses, these in silico methods can speed up and focus research in many different fields of evolutionary and molecular biology.

Acknowledgements

We thank JP Bielawski for helpful discussions and for sharing unpublished results.

Notes

[1] Available at http://www.bio.psu.edu/People/Faculty/Nei/software.htm
[2] Available at http://abacus.gene.ucl.ac.uk/software/paml.html
[3] Available at http://xgu1.zool.iastate.edu/software.html

Glossary

alloploidy fusion of the genomes of two different species

codon bias the preferential usage of certain codon(s) above others coding for the same amino acid; this difference exists between organisms and/or genes

colinearity conserved gene order and content between different genomic segments

differential gene loss reciprocal deletion of genes in duplicated segments

gene conversion a process preventing divergence of homologous loci in a species (also called non-reciprocal recombination)

homologues genes that share a common ancestor

in silico method to study questions in molecular biology using computational means rather than laboratory experiments on cell or tissue cultures (in vitro) or on living organisms (in vivo)

nonsynonymous substitutions nucleotide substitutions that lead to amino acid replacements

orthologues homologous genes that originated through speciation (see Figure 1)

paralogues homologous genes that originated through duplication (see Figure 1)

polyploidy doubling of the copy number of each chromosome in a species

positive selection selection fixing advantageous mutations

pseudogene non-functional gene due to the accumulation of deleterious mutations; the process in which a functional gene becomes a pseudogene is called pseudogenisation

purifying selection selection against deleterious mutations (also called negative selection)

subfunctionalisation process in which duplicated genes divide functions originally exerted by the ancestral gene

synonymous substitutions nucleotide substitutions that do not lead to amino acid replacements

transitions substitution of a purine (A,G) by another purine, or a pyrimidine (C,T) by another pyrimidine

transversions substitution of a purine (A,G) by a pyrimidine (C,T) or vice versa

References

Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H. 2002. Evidence of en bloc duplication in vertebrate genomes. *Nat Genet*, 31:100–5.

Aburomia R, Khaner O, Sidow A. 2003. Functional evolution in the ancestral lineage of vertebrates or when genomic complexity was wagging its morphological tail. *J Struct Funct Genomics*, 3:45–52.

[AGI] Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408:796–815.

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL et al. 1998. Zebrafish Hox clusters and vertebrate genome evolution. *Science*, 282:1711–14.

Aparicio S, Hawker K, Cottage A, Mikawa Y, Zuo L, Venkatesh B, Chen E, Krumlauf R, Brenner S. 1997. Organization of the *Fugu rubripes* Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. *Nat Genet*, 16:79–83.

Bielawski JP, Dunn KA, Yang Z. 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. *Genetics*, 156:1299–308.

Bielawski JP, Yang Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. *J Struct Funct Genomics*, 3:201–12.

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M. 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell*, 12:1093–101.

Bush RM, Fitch WM, Bender CA, Cox NJ. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol*, 16:1457–65.

Comeron JM. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J Mol Evol*, 41:1152–9.

Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol Biol Evol*, 16:372–82.

Cronn RC, Small RL, Wendel JF. 1999. Duplicated genes evolve independently after polyploid formation in cotton. *Proc Natl Acad Sci USA*, 96:14406–11.

Dagan T, Talmor Y, Graur D. 2002. Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. *Mol Biol Evol*, 19:1022–5.

Dermitzakis ET, Clark AG. 2001. Differential selection after duplication in mammalian developmental genes. *Mol Biol Evol*, 18:557–62.

Duda TF Jr, Palumbi SR. 1999. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod Conus. *Proc Natl Acad Sci USA*, 96:6820–3.

Eck RV, Dayhoff MO. 1966. Atlas of protein sequence and structure. Silver Spring, MD: National Biomedical Research Foundation.

Endo T, Ikeo K, Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. *Mol Biol Evol*, 13:685–90.

Fares MA, Elena SF, Ortiz J, Moya A, Barrio E. 2002. A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. *J Mol Evol*, 55:509–21.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet*, 22:521–65.

Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Zool*, 20:406–16.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151:1531–45.

Friedman R, Hughes AL. 2001. Pattern and timing of gene duplication in animal genomes. *Genome Res*, 11:1842–7.

Friedman N, Hughes AL. 2003. The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol Biol Evol*, 20:154–61.

Furlong RF, Holland PW. 2002. Were vertebrates octoploid? *Philos Trans R Soc Lond B Biol Sci*, 357:531–44.

Gaucher EA, Gu X, Miyamoto MM, Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem Sci*, 27:315–21.

Gaucher EA, Miyamoto MM, Benner SA. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. *Proc Natl Acad Sci USA*, 98:548–52.

Gehring WJ. 1998. Master control genes in development and evolution: the homeobox story. New Haven, CT: Yale Univ Pr.

Gibson TJ, Spring J. 1998. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet*, 14:46–9, discussion 49–50.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, 11:725–36.

Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol*, 16:1664–74.

Gu X. 2001. A site-specific measure for rate difference after gene duplication or speciation. *Mol Biol Evol*, 18:2327–30.

Gu X, Vander Velden K. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics*, 18:500–1.

Gu X, Wang Y, Gu J. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet*, 31:205–9.

Holland PW. 1997. Vertebrate evolution: something fishy about Hox genes. *Curr Biol*, 7:R570–2.

Holland PW. 2003. More genes in vertebrates? *J Struct Funct Genomics*, 3:75–84.

Holland PW, Garcia-Fernandez J. 1996. Hox genes and chordate evolution. *Dev Biol*, 173:382–95.

Holland PW, Garcia-Fernandez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Development*, Suppl:125–33.

Huelsenbeck JP, Bollback JP. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. *Syst Biol*, 50:351–66.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc R Soc Lond B Biol Sci*, 256:119–24.

Hughes AL. 1999. Adaptive evolution of genes and genomes. New York: Oxford Univ Pr.

Hughes AL, Green JA, Garbayo JM, Roberts RM. 2000. Adaptive diversification within a large family of recently duplicated, placentally expressed genes. *Proc Natl Acad Sci USA*, 97:3319–23.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335:167–70.

Hughes AL, Ota T, Nei M. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol Biol Evol*, 7:515–24.

Hughes MK, Hughes AL. 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol Biol Evol*, 10:1360–9.

Ina Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J Mol Evol*, 40:190–226.

Kimura H. 1983. The neutral theory of molecular evolution. Cambridge, UK: Cambridge Univ Pr.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome Biol*, 3:research0008.1–9.

Koshi JM, Goldstein RA. 1996. Probabilistic reconstruction of ancestral protein sequences. *J Mol Evol*, 42:313–20.

Larhammar D, Lundin LG, Hallbook F. 2002. The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res*, 12:1910–20.

Li WH. 1980. Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. *Genetics*, 95:237–58.

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol*, 36:96–9.

Li WH. 1997. Molecular evolution. Sunderland, MA: Sinauer.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*, 2:150–74.

Liberles DA. 2001. Evaluation of methods for determination of a reconstructed history of gene sequence evolution. *Mol Biol Evol*, 18:2040–7.

Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA. 2001. The adaptive evolution database (TAED). *Genome Biol*, 2:research0028.1–6.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science*, 290:1151–5.

Maddison WP, Maddison DR. 1992. MacClade: analysis of phylogeny and character evolution [computer program]. Version 3. Sunderland, MA: Sinauer.

Málaga-Trillo E, Meyer A. 2001. Genome duplications and accelerated evolution of Hox genes and cluster architecture in teleost fishes. *Am Zool*, 41:676–86.

Margoliash E. 1963. Primary structure and evolution of cytochrome c. *Proc Natl Acad Sci USA*, 50:672–9.

Marín I, Fares MA, González-Candelas F, Barrio E, Moya A. 2001. Detecting changes in the functional constraints of paralogous genes. *J Mol Evol*, 52:17–28.

McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. *Nat Genet*, 31:200–4.

Meyer A, Van de Peer Y. 2003. 'Natural selection merely modified while redundancy created'– Susumu Ohno's idea of the evolutionary importance of gene and genome duplications. *J Struct Funct Genomics*, 3:7–9.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, 11:715–24.

Muse SV, Weir BS. 1992. Testing for equality of evolutionary rates. *Genetics*, 132:269–76.

Naruse K, Fukamachi S, Mitani H, Kondo M, Matsuoka T, Kondo S, Hanamura N, Morita Y, Hasegawa K, Nishigaki R et al. 2000. A detailed linkage map of medaka, *Oryzias latipes*: comparative genomics and genome evolution. *Genetics*, 154:1773–84.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3:418–26.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford Univ Pr.

Nelson JS. 1994. Fishes of the world. 3 ed. New York: J Wiley.

Nielsen R. 2002. Mapping mutations on phylogenies. *Syst Biol*, 51:729–39.

Nielsen R, Huelsenbeck JP. 2002. Detecting positively selected amino acid sites using posterior predictive P-values. *Pac Symp Biocomput*, 576–88.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148:929–36.

Nowak MA, Boerlijst MC, Cooke J, Smith JM. 1997. Evolution of genetic redundancy. *Nature*, 388:167–71.

Ohno S. 1970. Evolution by gene duplication. Berlin: Springer-Verlag.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature*, 401:877–84.

Pamilo P, Bianchi NO. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol Biol Evol*, 10:271–81.

Paterson AH, Bowers JE, Burow MD, Draye X, Elsik CG, Jiang CX, Katsar CS, Lan TH, Lin YR, Ming R et al. 2000. Comparative genomics of plant chromosomes. *Plant Cell*, 12:1523–40.

Peltier MR, Raley LC, Liberles DA, Benner SA, Hansen PJ. 2000. Evolutionary history of the uterine serpins. *J Exp Zool*, 288:165–74.

Piatigorsky J, Wistow G. 1991. The recruitment of crystallins: new functions precede gene duplication. *Science*, 252:1078–9.

Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan YL, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res*, 10:1890–902.

Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet*, 3:827–37.

Pupko T, Pe'er I, Hasegawa M, Graur D, Friedman N. 2002. A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: application to the evolution of five gene families. *Bioinformatics*, 18:1116–23.

Pupko T, Pe'er I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol*, 17:890–6.

Raes J, Vandepoele K, Simillion C, Saeys Y, Van de Peer Y. 2003. Investigating ancient duplication events in the *Arabidopsis* genome. *J Struct Funct Genomics*, 3:117–29.

Robinson M, Gouy M, Gautier C, Mouchiroud D. 1998. Sensitivity of the relative-rate test to taxonomic sampling. *Mol Biol Evol*, 15:1091–8.

Robinson-Rechavi M, Laudet V. 2001. Evolutionary rates of duplicate genes in fish and mammals. *Mol Biol Evol*, 18:681–3.

Rombauts S, Florquin K, Lescot M, Marchal K, Rouze P, Van de Peer Y. 2003. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol*, 132:1162–76.

Sarich VM, Wilson AC. 1973. Generation time and genomic evolution in primates. *Science*, 179:1144–7.

Schluter D. 1995. Uncertainty in ancient phylogenies. *Nature*, 377:108–10.

Seoighe C, Wolfe KH. 1999. Yeast genome evolution in the post-genome era. *Curr Opin Microbiol*, 2:548–54.

Sharp PM. 1997. In search of molecular Darwinism. *Nature*, 385:111–2.

Sidow A. 1996. Gen(om)e duplications in the evolution of early vertebrates. *Curr Opin Genet Dev*, 6:715–22.

Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci USA*, 99:13627–32.

Smith JM. 1994. Estimating selection by comparing synonymous and substitutional changes. *J Mol Evol*, 39:123–8.

Spring J. 1997. Vertebrate evolution by interspecific hybridisation – are we polyploid? *FEBS Lett*, 400:2–8.

Spring J. 2002. Genome duplication strikes back. *Nat Genet*, 31:128–9.

Stoltzfus A. 1999. On the possibility of constructive neutral evolution. *J Mol Evol*, 49:169–81.

Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol*, 16:1315–28.

Suzuki Y, Gojobori T, Nei M. 2001. ADAPTSITE: detecting natural selection at single amino acid sites. *Bioinformatics*, 17:660–1.

Swofford D. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods) [computer program]. Version 4. Sunderland, MA: Sinauer.

Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, 135:599–607.

Takezaki N, Rzhetsky A, Nei M. 1995. Phylogenetic tests of the molecular clock and linearized trees. *Mol Biol Evol*, 12:823–33.

Tanaka T, Nei M. 1989. Positive Darwinian selection observed at the variable-region genes of immunoglobulins. *Mol Biol Evol*, 6:447–59.

Tang H, Lewontin RC. 1999. Locating regions of differential variability in DNA and protein sequences. *Genetics*, 153:485–95.

Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res*, 13:382–90.

Taylor JS, Van de Peer Y, Braasch I, Meyer A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos Trans R Soc Lond B Biol Sci*, 356:1661–79.

Taylor JS, Van de Peer Y, Meyer A. 2001. Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. *Curr Biol*, 11:R1005–8.

Terryn N, Heijnen L, De Keyser A, Van Asseldonck M, De Clercq R, Verbakel H, Gielen J, Zabeau M, Villarroel R, Jesse T et al. 1999. Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. *FEBS Lett*, 445:237–45.

Vacquier VD, Swanson WJ, Lee YH. 1997. Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? *J Mol Evol*, 44:S15–22.

Van de Peer Y, Taylor JS, Braasch I, Meyer A. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J Mol Evol*, 53:436–46.

Van de Peer Y, Taylor JS, Meyer A. 2003. Are all fishes ancient tetraploids? *J Struct Funct Genomics*, 3:65–73.

Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in *Arabidopsis*. *Science*, 290:2114–17.

Wagner A. 1998. The fate of duplicated genes: loss or new function? *Bioessays*, 20:785–8.

Wagner A. 2002. Asymmetric functional divergence of duplicate genes in yeast. *Mol Biol Evol*, 19:1760–8.

Walsh JB. 1995. How often do duplicated genes evolve new functions? *Genetics*, 139:421–8.

Wittbrodt J, Meyer A, Schartl M. 1998. More genes in fish? *Bioessays*, 20:511–12.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet*, 2:333–41.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–13.

Woods IG, Kelly PD, Chu F, Ngo-Hazelett P, Yan YL, Huang H, Postlethwait JH, Talbot WS. 2000. A comparative map of the zebrafish genome. *Genome Res*, 10:1903–14.

Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc Natl Acad Sci USA*, 82:1741–5.

Yang J, Huang J, Gu H, Zhong Y, Yang Z. 2002. Duplication and adaptive evolution of the chalcone synthase genes of Dendranthema (Asteraceae). *Mol Biol Evol*, 19:1752–9.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13:555–6.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*, 15:568–73.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*, 15:496–503.

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141:1641–50.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, 17:32–43.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, 19:908–17.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–49.

Zhang J, Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol*, 44:S139–46.

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci USA*, 95:3708–13.

Zhang J, Zhang YP, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet*, 30:411–15.

Zhang L, Vision TJ, Gaut BS. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol Biol Evol*, 19:1464–73.

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