REVIEW

Innovation from reduction: gene loss, domain loss and sequence divergence in genome evolution

Edward L Braun

Department of Zoology, University of Florida, Gainesville, FL, USA

Abstract: Analyses of genome sequences have revealed a surprisingly variable distribution of genes, reflecting the generation of novel genes, lateral gene transfer and gene loss. The impact of gene loss on organisms has been difficult to examine, but the loss of protein coding genes, the loss of domains within proteins and the divergence of genes have made surprising contributions to the differences among organisms. This paper reviews surveys of gene loss and divergence in fungal and archaeal genomes that indicate suites of functionally related genes tend to undergo loss and divergence. Instances of fungal gene loss highlighted here suggest that specific cellular systems have changed, such as Ca^{2+} biology in *Saccharomyces cerevisiae* and peroxisome function in *Schizosaccharomyces pombe*. Analyses of loss and divergence can provide specific predictions regarding protein–protein interactions, and the relationship between networks of protein interactions and loss may form a part of a parametric model of genome evolution.

Keywords: comparative genomics, computational genetics, proteomics, scale-free networks, ancestral state reconstruction

Introduction

Sequence analyses such as database homology searches have proven to be an extremely powerful method for predicting gene function. The novel discipline of comparative genomics examines these sequence comparisons in an evolutionary framework (Koonin et al 2000). Since the amount of information obtained using sequence comparisons can be increased by placing the results into this comparative context, results from comparative genomics represent a major part of the integrative discipline of computational genetics (Marcotte 2000). Ultimately, the goals of computational genetics are to explain differences among organisms in genome content and structure, establish the functional context of gene products and improve the functional annotation of genome sequences using these data. Functional annotations can be used to examine phenotypic differences among organisms with different genes and to propose testable biochemical hypotheses.

The distribution of genes in organisms reflects the combination of several distinct types of evolutionary change (Figure 1). The presence of related genes in two or more organisms can be explained either by the gene originating in the common ancestor of the organisms or by lateral (horizontal) gene transfer. Likewise, the absence of a specific gene in an organism can reflect the appearance of the gene in a distinct lineage or the loss of a gene that was present in an ancestor of the organism. The presence or absence of a gene in a set of organisms has been called the 'phylogenetic profile of the gene (Pellegrini et al 1999). As long as a sufficient number of organisms is used to construct phylogenetic profiles, a strong correlation between the phylogenetic profile of the gene and the function of the encoded gene products is evident (Gaasterland and Ragan 1998b; Pellegrini et al 1999; Marcotte et al 2000).

Despite the usefulness of phylogenetic profiles for the prediction of gene function, the data structures used in these analyses (Figure 1) have two potential limitations for the prediction of gene function. First, closely related organisms will contain similar sets of genes owing to shared evolutionary history, so data in phylogenetic profiles are not independent. This problem can be avoided by using methods developed for comparison of other types of data in a phylogenetic framework (reviewed by Harvey and Pagel 1991). In fact, a recent study (Liberles et al 2002) used parsimony ancestral state reconstruction for the analysis of phylogenetic profiles to address this issue. Second, phylogenetic profiles of genes can be identical when distinct evolutionary processes are responsible for the distribution of genes in extant organisms (eg Figures 1b and 1c). Rigorous methods of reconstructing the evolutionary history of genomes could reduce these confounding aspects of

Correspondence: Edward L Braun, Department of Zoology, University of Florida, PO Box 118525, Gainesville, FL 32611-8525, USA; tel +1 352 846 1124; fax +1 352 392 3704; email ebraun@zoo.ufl.edu

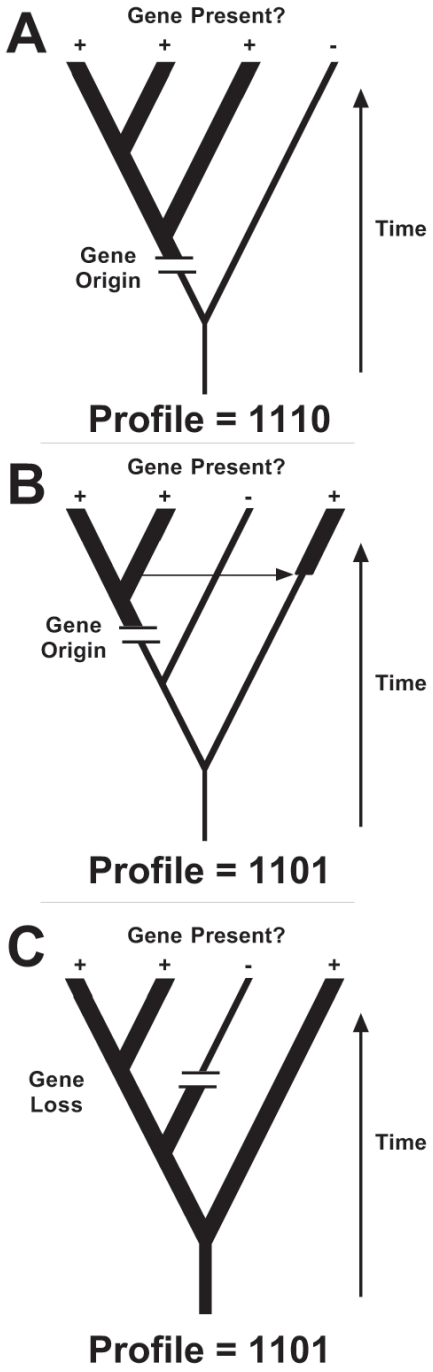

Figure 1 Types of evolutionary change that can result in differences between genomes in gene content. Presence of the gene of interest is indicated by thick lines and breaks in the tree indicate the origin of novel genes and gene loss. **(A)** Presence or absence of the gene reflects the origin of a gene by some mechanism (eg gene duplication or domain rearrangements) in a specific lineage. The origin of the gene divides the organisms into an ingroup with the gene and an outgroup that diverged prior to the origin of the gene. The phylogenetic profiles of the genes are shown below the trees, with the presence or absence of the gene providing one bit of information (see Pellegrini et al 1999). Other phylogenetic profile methods (eg Gaasterland and Ragan 1998a; Marcotte et al 2000) are similar, but they also include information about the significance of the hit. **(B)** Presence of the gene in an outgroup lineage due to lateral gene transfer (indicated with an arrow). **(C)** Absence of the gene in an ingroup lineage due to gene loss. Note that the phylogenetic profiles for (B) and (C) are identical although the evolutionary processes responsible for the distribution differ.

evolutionary history, revealing more accurately the specific types of genomic changes that have occurred.

This review focuses on a surprisingly neglected pattern of genomic change: the loss and divergence of genes during evolution. Large-scale analyses of fungal genomes have shown that functionally related sets of genes have undergone loss and divergence in the yeast *Saccharomyces cerevisiae* (Aravind et al 2000; Braun et al 2000). These results have been extended to the archaeal genus *Pyrococcus* in a more recent survey of gene loss (Ettema et al 2001). Prospects for the use of information on gene loss and divergence to make inferences about gene function, especially inferences regarding protein–protein interactions, are detailed. Related patterns of genomic change, such as domain loss (Braun and Grotewold 2001; Copley et al 2002) and lateral gene transfer (Doolittle 1999), are also discussed in the context of gene loss and divergence. Finally, the potential for examining the loss and divergence of genes in a parametric framework is discussed.

The biological consequences of gene loss

There are two distinct outcomes that can occur after gene loss: the biological function specified by the gene can be either lost or retained. Loss of a biological function is likely to reflect a change in the interaction between the organism and its environment that relaxes the selective pressure that had maintained the function. Alternatively, changes in selective pressures may make retention of specific biological functions disadvantageous, resulting in selection for loss of genes associated with that function. Retention of a biological function despite gene loss can reflect the presence of another protein that can provide the same function (non-orthologous displacement; see Koonin et al 1996) or inherent robustness of the biochemical network (eg Barkai and Leibler 1997). The robustness of most biological networks is unclear, although analyses of connections among metabolites (Albert et al 2000) and protein–protein interactions (Jeong et al 2001) indicate these biological networks exhibit highly heterogeneous scale-free topologies (similar to the example in Figure 2). This suggests biological networks will exhibit robustness with respect to random errors similar to other types of networks with similar topologies (Albert et al 2000).

It is possible to find specific examples of gene loss reflecting each of the potential outcomes described above in previous studies of loss in the *S. cerevisiae* lineage. In some cases non-orthologous displacement represents a likely explanation, since non-orthologous *S. cerevisiae* genes with

Table 1 Examples of biological processes that are absent or altered in *S. cerevisiae* and are predicted to involve protein coding genes that have been lost

Process absent in S. cerevisiae
Cytosine methylation in DNA
Light sensing and signal transduction
Post-transcriptional gene silencing
Process altered in S. cerevisiae
Ion homeostasis (especially Ca^{2+} signalling)
mRNA splicing
Translation initiation

the same function as the genes that were lost have been identified (eg Braun et al 2000 reported loss of a phosphoglycerate mutase in *S. cerevisiae*, which has a known phosphoglycerate mutase encoded by the *GPM1* gene). Although additional cases of non-orthologous displacement probably exist, it is clear that some genes predicted to play a role in biological processes known to differ between *S. cerevisiae* and other eukaryotes have been lost (Table 1). Unfortunately, the precise impact of the genes that have been lost is often unclear. For example, several *S. cerevisiae* Ca^{2+}-binding proteins exhibit functional differences from orthologues in other eukaryotes (evidence reviewed by Braun et al 2000) and several genes encoding proteins involved in ion sensing or transport have been lost in *S. cerevisiae* (Braun et al 1998; Braun et al 2000). However, it is not clear how many of the genes that have been lost interact directly with genes exhibiting altered functions. Studies, using different

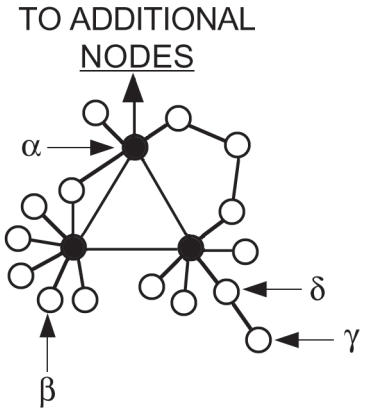

Figure 2 Portion of a hypothetical biological network, showing the heterogeneous number of connections made by different gene products. Highly connected nodes are in black. A variety of biological networks exhibit topologies similar to this figure, described as scale-free networks in which the probability of finding a node with k connections follows a power law, $P(k) \approx k^{-\gamma}$, where γ is a constant. The nature of connections between gene products should be viewed broadly, as protein–protein interactions, metabolic connections, transcriptional regulation and other possible interactions (reviewed by Oltvai and Barabási 2002).

model systems, to examine connections among the products of genes that have been lost and proteins with altered functions should establish the biological impact of these changes more clearly.

Despite the limitations of the available empirical evidence, there are reasons to suspect that gene loss may help generate novel variants that would be favoured under specific circumstances. Since the number of connections among nodes in many biological networks is highly heterogeneous (eg Figure 2), the impact of gene loss will depend on the number of connections involving the gene. Loss of a highly connected gene (eg node α in Figure 2) is likely have a large effect while loss of a gene with a limited number of connections (eg nodes β or γ in Figure 2) is likely to have a more limited impact. This more limited impact reflects the ability of the network topology to insulate other biological processes that involve gene products elsewhere in the network from the impact of loss. Under these circumstances, the potential for loss to generate a novel phenotype by altering a limited set of processes is extremely plausible. The most interesting possibility in this context would be the generation of a novel phenotype that results from the emergent properties of the altered network, as opposed to the relaxation of selection for a specific biological activity.

The probability that a gene loss will ultimately result in genetic innovation is unclear, although it is likely to be lower than the probability that innovation will result from gene duplication. However, the potential for innovation due to gene loss should not be ignored. Provocative evidence that gene loss can result in evolutionary innovations is provided by loss of the gene encoding CMP-*N*-acetylneuraminic acid hydroxylase (CMAH) in humans. CMAH synthesises *N*-glycolylneuraminic acid (Neu5Gc), a sialic acid found in many mammals, including great apes. The CMAH gene was inactivated in the human lineage approximately 2.8 million years ago, shortly before significant increases in brain size relative to body size occurred in the human lineage (Chou et al 2002). Brain tissue of other mammals shows lower levels of mRNA accumulation for CMAH and less Neu5Gc than other tissues (Kawano et al 1995). These observations have led to the hypothesis that low levels of residual Neu5Gc in the brains in other mammals limited brain expansion and that the loss of the CMAH gene released our ancestors from this constraint (Chou et al 2002). This hypothesis is consistent with the existence of additional genetic changes in sialic acid biology specific to the human lineage (Angata et al 2001). Although it is premature to conclude that loss of the CMAH

gene played a direct role in the origin of human-specific features, these observations provide a direction for future research on the genetic basis of the differences between humans and great apes. The possibility that some phenotypic differences between humans and great apes might reflect gene loss emphasises the potential importance of gene loss to biological innovation.

Performing large-scale surveys of gene loss in eukaryotes

There are two potential problems with examining gene loss after the divergence of closely related organisms, such as humans and chimpanzees. First, pseudogene sequences can be retained after the loss of gene function and distinguishing pseudogenes from functional but incorrectly annotated genes can be very difficult in some eukaryotes. In fact, an expressed CMAH pseudogene is present in humans (Chou et al 1998), although the pseudogene does not encode a functional protein. Second, the number of differences in genome content for closely related organisms will be fairly limited, necessitating the use of relatively large amounts of sequence data before instances of gene loss are evident. For these reasons, large-scale surveys of gene loss using relatively divergent organisms are expected to be more useful for computational genetics than comparisons using closely related organisms.

Candidate genes that have undergone loss or divergence in a specific lineage can be detected by conducting comparative database searches using programmes such as BLAST (Altschul et al 1997) or FASTA (Pearson 2000) in an appropriate framework. The set of phylogenetic trees estimated for all genes in the genome of an organism has been called the phylome (Sicheritz-Pontén and Andersson 2001), and relationships between these trees and the phylogeny of the organisms provides an appropriate framework for examining gene loss (Figure 3). A reasonable and appropriate null hypothesis for eukaryotic genes is orthology without lateral transfer or gene loss (see null hypothesis, Figure 3a). This null hypothesis is tested by using sequences from a query organism to search two databases: (1) a 'complete' set of coding sequence data from an organism related to the query organism, which I will call the 'reference organism'; and (2) sets of coding sequence data from one or more organisms that are more distantly related to the query organism than the reference organism, hereafter called the 'outgroup'. Results from these database searches that are incompatible with the null hypothesis are used to highlight sequences that have been lost in the reference organism lineage.

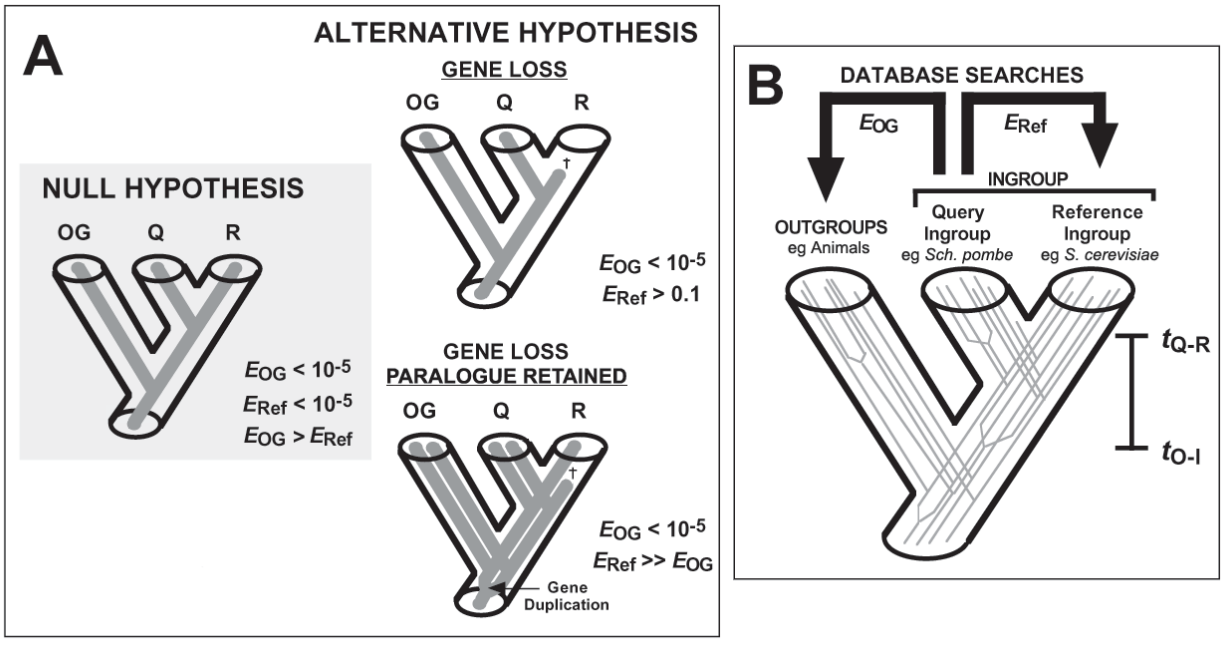

Figure 3 General approach for conducting large-scale screens of gene loss. **(A)** The purpose of the analyses is to identify genes that are inconsistent with the null hypothesis, in which the gene is present in all lineages. The simplest case for gene loss in the reference organism ('R') would be supported by a query sequence ('Q') that detects a homologue in the outgroup ('OG') but not in the reference organism (top figure). Alternatively, gene loss with retention of a paralogue that resulted from an ancient duplication would be supported by a query sequence that detects a homologue in the outgroup that is more closely related to the query than the homologue in the reference organism (bottom figure). **(B)** Schematic showing the use of database searches to identify genes with evolutionary histories that are inconsistent with the null hypothesis. Gene trees are shown nested within the species tree, and the initial set of database (eg BLASTP or TBLASTN) searches used to calculate *E*-values are shown.

The query organism and reference organism must form a well-supported clade, and a complete genome sequence must be available for the reference organism. Query organism sequences are used to search the databases of a reference organism and of outgroup organisms (Figure 3b), and the divergence of the best hit in each database is determined. The *E*-value of the BLAST searches used in this step provides a reasonable measure of sequence divergence, although other measures may be desirable (eg maximum likelihood distance estimates). The expected result under the null hypothesis is for orthologues of the reference organism to exhibit less divergence from the query organism than the outgroup does ($E_{Ref} < E_{OG}$). Although a number of biological phenomena might cause deviations from this null hypothesis, this review will focus on gene loss, and any sequences that have an E_{OG} value substantially less than the E_{Ref} value will be considered candidates for gene loss.

There are two important factors for the analysis of gene loss: appropriate choice of sequence databases and appropriate selection of *E*-values (ie choosing ones that are indicative of homology). The selection of *E*-values indicative of homology for surveys of gene loss must consider that the results of two or more independent database searches will be compared. This makes it problematic to use a single critical value indicative of homology (eg $E < 10^{-5}$), because the use of a single critical value might suggest that gene loss has occurred when E_{OG} is significant but similar to a slightly higher E_{Ref} value that is not significant (eg $E_{OG} = 10^{-6}$ and $E_{Ref} = 10^{-4}$). This problem can be solved by using more stringent critical values for E_{OG} than for E_{Ref} (eg $E_{OG} < 10^{-5}$ and $E_{Ref} > 0.1$, used by Braun et al 2000).

There are multiple reasons why the appropriate selection of sequence databases is critical. The identification of protein coding genes in eukaryotic genomes can be difficult (reviewed by Harrison et al 2002), so it is important to supplement searches of annotated proteins from the reference organism with searches of DNA databases (preferably both expressed sequence tag (EST) and genomic sequence data). Although factors such as the intron–exon structure of genes will have an impact on the *E*-values of homologues identified, these can highlight improperly annotated proteins. The observation that a homologue of the query sequence cannot be detected in annotated proteins, genomic DNA, or EST databases strengthens the conclusion that the gene is absent. If a homologue to the query sequence is present, these results would indicate that (1) it would have to fall in the small regions of 'complete' eukaryotic genomes that are not sequenced (primarily heterochromatin) and (2) that it is absent from EST data because of limited gene expression and/or other factors that limit the representation of specific genes in cDNA libraries. Searches of EST data have the potential to be especially informative, since one might be able to estimate the expected number of ESTs if mRNA accumulation and other factors that determine representation in cDNA libraries are similar to that of orthologues in other organisms. This approach should limit the potential for errors, since the likelihood that multiple functionally related genes will be undetectable in all of the databases is extremely low.

These analyses are also complicated by the potential of the reference organism to retain a paralogue (Figure 3a). Although these cases may seem less interesting than those in which reference organism sequences are completely absent, appropriate choice of databases can allow the detection of biologically interesting instances of gene loss with retention of a paralogue. The relevant factor is the amount of time between the divergence of the outgroup and ingroup ($t_{O\text{-}I}$ in Figure 3b) and the divergence of the query organism and the reference organism ($t_{Q\text{-}R}$ in Figure 3b). The difference between $t_{Q\text{-}R}$ and $t_{O\text{-}I}$ represents the minimum amount of time that both members of the gene family have been retained after duplication. If this amount of time is sufficiently long, any paralogues retained are unlikely to exhibit substantial functional redundancy (see below for examples).

Previous surveys of gene loss in *S. cerevisiae* (Aravind et al 2000; Braun et al 2000) should have detected instances of gene loss with retention of a paralogue that had arisen prior to the divergence of the fungi from other eukaryotes. If the query organism sequence is a paralogue of the top hit in the reference organism, both of these sequences would have to have been retained until the early divergences among ascomycetes, approximately 500 million years later (divergence time estimates are from Feng et al 1997 and Taylor et al 1999), so the genes are expected to show substantial functional divergence. Although the degree of functional divergence will differ among gene families, the well-characterised *rab* gene family provides an informative example of functional divergence typical for ancient gene families. *rab* genes encode monomeric GTPases that regulate intracellular transport (Bock et al 2001), acting between specific intracellular compartments (eg endoplasmic reticulum to Golgi apparatus). Members of each group of *rab* genes encode proteins with different cellular locations and functions in intracellular transport (Bock et al 2001; Gupta and Heath 2002). Despite the functional divergence in the *rab* gene family, BLASTP searches with *rab* queries

detect paralogues with highly significant *E*-values (typically $E < 10^{-20}$). Detecting gene loss with retention of paralogues that exhibit functional divergence similar to the distinct groups of *rab* genes is clearly desirable.

Since the query organism and reference organism are chosen because they form a well-supported clade, sequences from the query organism are expected to show a higher degree of sequence identity to reference organism orthologues than to orthologues in outgroup organisms. However, the degree of bias toward the reference organism will depend upon the time since the divergence of the ingroup organisms (from $t_{Q\text{-}R}$ to present) and the rate of molecular evolution in the ingroup organisms. Thus, examining the distribution of *E*-values graphically (eg Figure 4) to determine the relationship between E_{Ref} and E_{OG} is useful. Although the plots used by Braun et al (2000) and Aravind et al (2000) differ in specifics, both illustrate the general relationship, such as the fact that comparisons of *S. cerevisiae* and distantly related query ascomycetes typically show values of $E_{Ref} \approx E_{OG}$ (Figure 4). This probably reflects the ancient divergences among basal ascomycetes and slightly accelerated molecular evolution relative to other groups of eukaryotes (eg Feng et al 1997).

Aravind et al (2000) and Braun et al (2000) both identified candidates for loss in the reference organism lineage using a 10-log criterion ($\log_{10}[E_{Ref}] + \log_{10}[E_{OG}] > 10$). This criterion was able to identify loss of a *rab* gene in the *S. cerevisiae* lineage (Braun et al 2000), despite the retention of six other groups of *rab* genes. Detailed evaluations of fungal *rab* genes have confirmed the loss of the *rab4* orthologue, which encodes a protein that functions in early transport from endosome to plasma membrane, in *S. cerevisiae* (Gupta and Heath 2002). This result indicates that use of the 10-log criterion for studies of distantly related fungi provides sufficient power to detect the loss of some genes within multigene families. Manual evaluation of genes identified using the 10-log criterion also suggests that it results in limited type I error (query sequences incorrectly thought to be lost in *S. cerevisiae*). However, the expected difference between E_{Ref} and E_{OG} will depend upon the organisms compared, so it is necessary to determine the appropriate *E*-values difference to use for each study. The simplest approach to this problem is to graph *E*-values from database searches using a set of proteins in the query organism that are known to have reference organism and outgroup orthologues.

Sequence divergence and lateral transfer can confound surveys of gene loss

In addition to gene loss, two additional biological phenomena can result in values of E_{OG} that are substantially better than E_{Ref} despite the closer relationship between the query and reference organisms. These phenomena include accelerated divergence in the reference lineage and lateral gene transfer

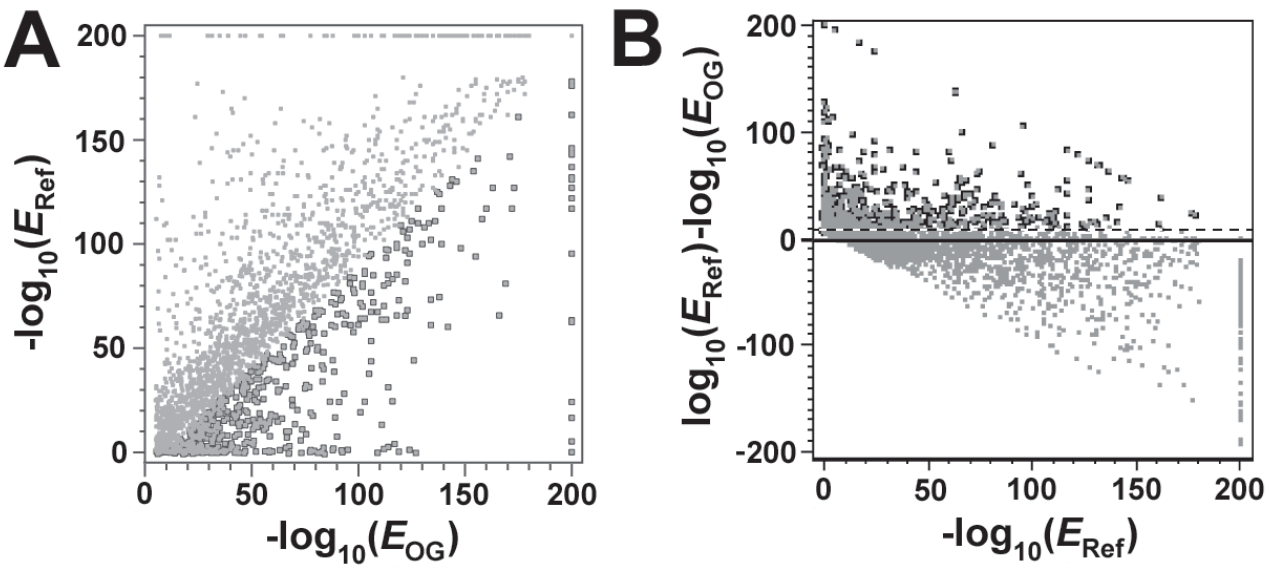

Figure 4 Results obtained using a non-redundant set of annotated proteins from *Sch. pombe* (Wood et al 2002) as queries and *S. cerevisiae* as the reference organism. Outgroup databases corresponded to annotated proteins from the complete and extensively annotated non-fungal eukaryotic genomes available in October 2002 (*Homo sapiens*, *D. melanogaster*, *C. elegans* and *A. thaliana*). **(A)** Data plotted as in Braun et al (2000), with queries that meet the 10-log criterion outlined in black. Note that most points fall along a line of $y \approx x$. **(B)** Data plotted as in Aravind et al (2000), with queries that meet the 10-log criterion outlined in black and the 10-log cutoff emphasised with a dashed line.

involving the query lineage and an outgroup (Figure 5). Gene loss with retention of a paralogue can be differentiated from an increase in the rate of sequence evolution using the top hit in the reference organism as a query in a search of the outgroup database. In contrast, differentiating gene loss from lateral gene transfer is more difficult, although some useful heuristics can be applied.

Differentiating gene loss from sequence divergence

Reference organism sequences that are divergent orthologues of the query sequence are expected to exhibit the smallest amount of divergence from the top outgroup hit of the query sequence when used as a query to search the outgroup database. For example, using the protein encoded by the *ryh1* gene of the *Schizosaccharomyces pombe rab* superfamily as a query sequence assigns a lower *E*-value to the human orthologue of this protein (*rab6B*, $E = 10^{-82}$) than does the most closely related protein in *S. cerevisiae* (*ypt6*, $E = 8 \times 10^{-69}$). However, human *rab6B* is also the top hit when searches using the product of the *ypt6* gene of the *S. cerevisiae rab* superfamily as a query are conducted ($E = 6 \times 10^{-67}$), suggesting that *ypt6* is a divergent *rab6* orthologue. In contrast, using the product of the *ypt4I* gene of the *Sch. pombe rab* superfamily also has a much better outgroup hit ($E = 10^{-54}$) than *S. cerevisiae* hit ($E = 2 \times 10^{-39}$), but the top *S. cerevisiae* hit detects a different outgroup protein as its top hit. This suggests the orthologue of the *Sch. pombe ypt4* (*rab4*) gene has been lost in *S. cerevisiae*, a fact confirmed by the analyses conducted by Gupta and Heath (2002). Since the top hit of the sequences from the query organism and reference organism may differ if there are lineage-specific gene duplications in the outgroup, requiring that both ingroup queries have identical top hits is likely to be too restrictive. However, database searches using the reference organism sequence should assign an *E*-value of the top query organism hit that is close to the *E*-value of the top hit. Braun et al (2000) scored reference organism sequences as divergent orthologues if the top hit of the reference organism sequence had an *E*-value no more than five logs worse than the best hit in the outgroup database.

Although the use of the top reference organism hit as a query in additional searches represents a simple way to differentiate between divergent orthologues and instances of gene loss, several additional factors should be considered when using this strategy. Independent losses in the same gene family in multiple lineages can confound the analysis. For example, the search results of the *rab* superfamily database, described above, that supported an orthologous relationship between *Sch. pombe ryh1* and *S. cerevisiae ypt6* would be similar if those genes were actually ancient paralogues but the *S. cerevisiae ryh1* orthologue and the human *ypt6*

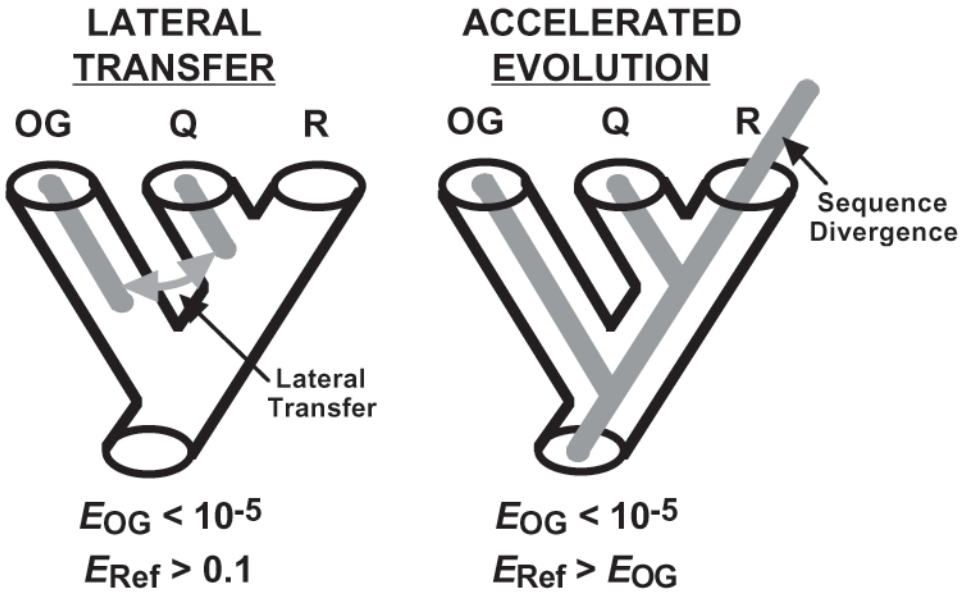

Figure 5 Alternative biological processes that support rejection of the null hypothesis. Both lateral gene transfer between the query lineage and an outgroup lineage and an unusually high rate of sequence divergence produce database search results that appear similar to those of gene loss or loss with retention of a paralogue.

orthologues had been lost. This question could be resolved by using the top human hit of the query organism sequence (*rab6B* in this example) to search the query organism database if the query organism database was also complete. Although a complete genome sequence is available for *Sch. pombe* (Wood et al 2002), the ability to conduct surveys of loss using incomplete query organism databases (eg sets of genes from individual chromosomes or chromosome regions, random sequence tag (RST) data or EST data) is desirable.

Since the use of incomplete query organism databases is desirable, an appropriate strategy is to use an outgroup database containing sequences from many organisms. In fact, instances of gene loss in *S. cerevisiae* that are highlighted by the analyses presented in this manuscript are unlikely to reflect multiple instances of loss in the same gene family because a total of four outgroup databases (underlined in Figure 6) were searched. If we consider the loss and extreme divergence of genes as events that are equally likely, explaining the *rab6* (*ryh1*/*ypt6*) data by invoking an accelerated rate of genome change requires that a single event occurred. In contrast, explaining the observed results of the database searches without invoking accelerated molecular evolution requires several additional instances of ingroup and outgroup loss in the same gene family. In the *rab* superfamily example, three instances of gene loss (loss of the *S. cerevisiae ryh1* orthologue and independent losses of metazoan and plant *ypt6* orthologues) represent the minimum genetic change necessary to explain the database search results.

Incorrect explanations of data indicative of gene loss by invoking divergence at an elevated rate will not inflate type I error. Instead, instances of loss with retention of a paralogue will incorrectly be attributed to rate differences. Although it is desirable to limit all types of error in studies of genome evolution, limiting the number of instances in which a specific null hypothesis is rejected incorrectly may be appropriate for specific studies. Different criteria may be appropriate for large-scale surveys of gene loss and genes that have undergone unusual divergence. The criteria suggested here might limit error in surveys of gene loss, but phylogenetic analyses to establish orthology followed by tests of the molecular clock (eg Wu and Li 1985; Yoder and Yang 2000)

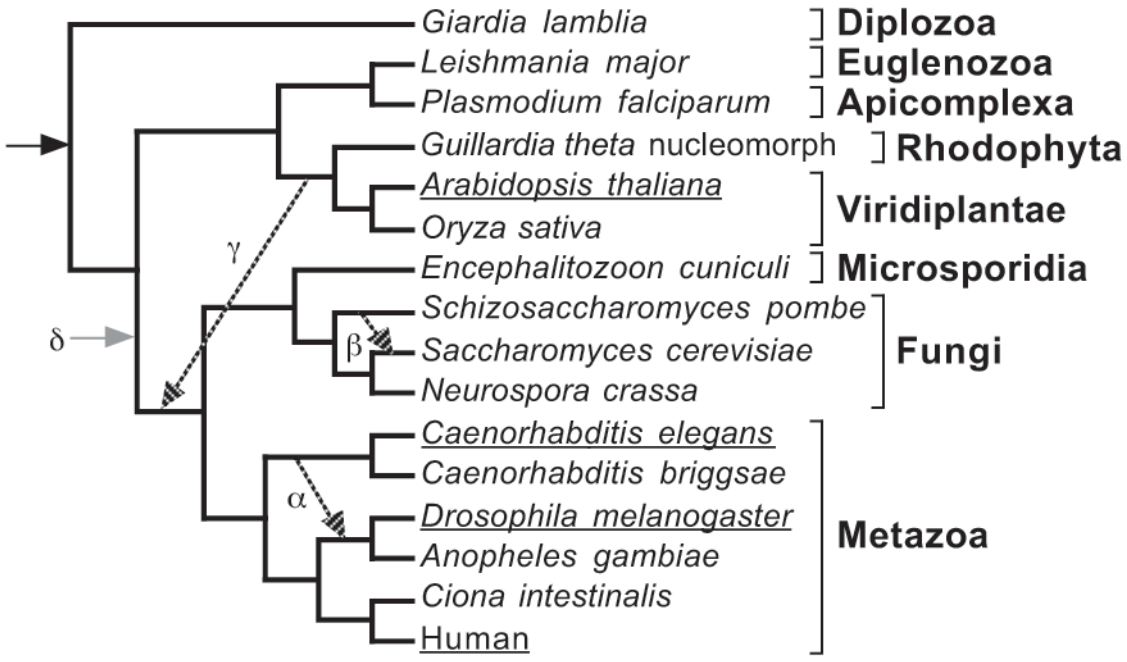

Figure 6 Cladogram showing an estimate of eukaryote phylogeny, emphasising organisms with abundant sequence data. Outgroup databases used for the reanalyses of *S. cerevisiae* and *Sch. pombe* sequences reported in this manuscript are underlined. This figure is based upon Baldauf et al (2000), with information from additional sources (Bruns et al 1992; Liu et al 1999; Lutzoni et al 2001) for the fungi. Controversial relationships are indicated with arrows. Arrow α: many analyses (eg Nielsen 1995; Aguinaldo et al 1997; Mushegian et al 1998; Giribet et al 2000) support a nematode–arthropod clade excluding deuterostomes (eg humans and *Ciona intestinalis*) while others support a coelomate clade with the nematodes basal (eg Hausdorf 2000; Blair et al 2002). Arrow β: some analyses support a *Sch. pombe*–*S. cerevisiae* clade that excludes *N. crassa* (eg Baldauf et al 2000) while others support a basal position for *Sch. pombe* (see above). Arrow γ: some analyses support an animal–plant–fungi clade with other eukaryotes basal (eg Kuma et al 1995; Bapteste et al 2002). Grey root arrow δ: the rooting of the eukaryotic tree (black arrow) may reflect rate differences among eukaryotes (see Stiller and Hall 1999) and the distribution of an enzyme with fused dihydrofolate reductase and thymidylate synthase activities supports this position for the root (Stechmann and Cavalier-Smith 2002).

might be more appropriate for surveys of sequences that have undergone divergence.

Differentiating gene loss from lateral transfer

It is more difficult to differentiate between lateral gene transfer and gene loss than it is to distinguish between gene loss and unequal rates of sequence divergence, although the potential for lateral gene transfer to confound large-scale analyses of gene loss is unclear. Aravind et al (2000, p 11319) asserted that the 'likelihood of [lateral] gene transfer between eukaryotes, at least multicellular ones, is low because, for a gene to be laterally transferred, it must enter the germ line.' In fact, phylogenetic analyses of many candidates for lateral transfer from prokaryotes to animals have supported loss from other groups of eukaryotes rather than lateral transfer (Roelofs and Van Haastert 2001; Stanhope et al 2001). However, some instances of prokaryote-to-animal transfer are supported by phylogenetic analyses (eg Wolf and Koonin 2001), suggesting that lateral transfer into animals is possible.

The distinction between the germ line and somatic cells is less defined outside of animals, so this barrier is unlikely to limit the impact of lateral gene transfer in other groups of eukaryotes. However, other barriers to lateral transfer from prokaryotes to eukaryotes may exist. In particular, the prokaryotes and eukaryotes exhibit differences in the transcriptional 'ground state' of DNA, with prokaryotes exhibiting a non-restrictive ground state while the packing of eukaryotic DNA into chromatin results in a restrictive ground state (Struhl 1999). Thus, genes acquired by lateral transfer are expected to have a higher probability of being expressed in prokaryotes than in eukaryotes.

Since examples of lateral transfer from prokaryotes to eukaryotes exist (eg Wolf and Koonin 2001), Braun et al (2000) explicitly examined lateral transfer into the fungal query organism used (*Neurospora crassa*). Query organism sequences that have homologues in a very divergent outgroup but lack homologues in more closely related organisms require the assumption of either multiple independent losses or lateral gene transfer. *Neurospora crassa* genes that only have prokaryotic homologues (and lack homologues in other eukaryotes) fall into this category. Three independent losses are necessary to explain the distribution of genes present in *N. crassa* but absent from *S. cerevisiae*, animals and plants. Thus, lateral transfer is a more parsimonious explanation when homologues in other eukaryotes are absent. Even if some genes present only in *N. crassa* and prokaryotes are explained by loss rather than lateral gene transfer, the set of sequences lacking homologues in non-fungal eukaryotes should contain a higher proportion of genes reflecting bona fide instances of lateral transfer. Braun et al (2000) identified 13 candidates for lateral transfer into the *N. crassa* lineage, less than one-third of the number of candidates for gene loss (46 candidates for gene loss from 396 *N. crassa* queries with $E_{OG} < 10^{-5}$).

The candidates for lateral transfer into the *N. crassa* lineage identified by Braun et al (2000) could include two additional classes of genes: (1) genes that were present in the common ancestor of *N. crassa* and *S. cerevisiae* and were lost in multiple eukaryotic lineages (see Salzberg et al 2001 for additional discussion); and (2) genes that arose in the *N. crassa* lineage and underwent lateral transfer to prokaryotes. Thus, the number of candidates for lateral transfer actually represents an upper limit for the impact of lateral transfer upon this type of analysis. Therefore, it is unlikely that estimates of the numbers of genes lost in the *S. cerevisiae* lineage are substantially biased upward by lateral gene transfer into the query lineage. As additional groups are examined, it will be imperative to conduct similar analyses to obtain better estimates of the probability of lateral gene transfer into various eukaryotic groups.

Considering all of these possible patterns of genome change, it is possible to obtain estimates for the impact of gene loss upon the *S. cerevisiae* genome. Braun et al (2000), using EST queries corresponding to 10%–15% of *N. crassa* genes, found 46 instances of gene loss. Thus, 300–460 genes present in the common ancestor of *N. crassa* and *S. cerevisiae* were lost in the *S. cerevisiae* lineage. Aravind et al (2000) found 215 candidates for gene loss using a set of annotated *Sch. pombe* proteins that reflects almost 90% of the genes present in the complete *Sch. pombe* based genome sequence (Wood et al 2002). This implies the loss of ~ 240 genes in the *S. cerevisiae* lineage. In fact, reanalysis of gene loss in *S. cerevisiae* using a non-redundant set of annotated proteins from *Sch. pombe* (these data are presented in Figure 4) resulted in the identification of exactly 240 instances of gene loss in the *S. cerevisiae* lineage, indicating that the *Sch. pombe* data available to Aravind et al (2000) were sufficiently complete to provide excellent estimates of the impact of gene loss upon the *S. cerevisiae* genome.

Phylogenetic approaches to infer gene loss and related phenomena

The comparative database searches used by Braun et al (2000) and Aravind et al (2000) must be conducted in an appropriate phylogenetic framework (eg Figure 3). However,

using *E*-values from database searches in this framework also creates problems, since *E*-values are expected to show saturation as sequences diverge as well as dependence upon the length of the matching region in the sequences. The length dependence of *E*-values would reduce the power to reject the null hypothesis for short sequences (or for proteins with short conserved regions) while saturation of the *E*-value parameter would tend to reduce power when rapidly evolving sequences are examined. The latter might represent a major problem if the query organism exhibits a global acceleration in the rate of sequence evolution.

A potential solution for these problems would be the inference of gene loss directly from phylogenetic trees. Parsimony approaches for fitting gene trees to species trees have existed for some time (eg Goodman et al 1979) and efficient algorithmic solutions to this problem have been implemented (eg Zmasek and Eddy 2001). In fact, there have been large-scale analyses using the reconciled gene tree approach to annotate orthologues (eg Storm and Sonnhammer 2002; Zmasek and Eddy 2002). In principle, using this approach to infer numbers of gene loss events is relatively straightforward. However, the large-scale studies of reconciled trees used the neighbour joining method to estimate gene trees and fitted these gene trees to fixed species trees in a framework that considered only duplication and loss. A better approach would be to estimate gene trees using an optimality criterion, instead of an approximate method like neighbour joining, and to reconcile trees in a framework that considers factors such as lateral gene transfer and uncertainty in the species tree (eg the uncertainty evident in Figure 6).

There are approaches to improve these aspects of current large-scale studies of reconciled trees. Novel methods of phylogenetic estimation using the likelihood criterion, such as simulated annealing and Markov chain Monte Carlo (Salter and Pearl 2001; Huelsenbeck et al 2002), are more computationally feasible than standard heuristic search strategies. Approaches developed for reconciling parasite and host phylogenies (eg Page 1994) can be used to examine patterns of gene duplication and loss while considering the possibility of lateral gene transfer, and algorithms for estimating numbers of duplications, losses and lateral transfer events given the gene trees and species trees are available (Charleston 1998; Ronquist 1998). Despite the promise of these approaches, the need to align sequences for phylogenetic analyses is likely to limit their reliability in large-scale studies that may include incorrectly annotated sequences. In contrast, comparative database searches can be used with incomplete query sequences (eg EST or RST data) and they can be relatively resilient to problems associated with annotation if searches of nucleotide databases with six-frame translations are conducted. Furthermore, genes that exhibit an unusually high degree of divergence will not be highlighted using reconciled tree approaches. For these reasons, reconciled tree methods and comparative database searches both have the potential to provide useful and possibly complementary information about patterns of gene loss and other types of genomic change.

Sequence divergence and gene loss are related phenomena

A potential problem for studies of gene loss using database searches is the fact that the best available methods for database searches miss a substantial proportion of true homologues (Brenner et al 1998). Profile search tools, such as PSI-BLAST and HMMER, are more sensitive than standard database search tools but they still miss many divergent homologues (Madera and Gough 2002). Thus, divergent orthologues of the query sequence present in the reference organism might escape detection. In these cases, one would infer gene loss when an outgroup orthologue of the query sequence was identified, despite the presence of the divergent orthologue in the reference organisms. Although this is an inherent source of error for this approach, the failure of database searches to detect any reference organism homologue implies the reference organism sequence must be more divergent than the orthologue present in the outgroup. Thus, results that are interpreted as loss actually imply either loss or extreme sequence divergence in the reference organism.

This raises the question of precisely how distinct gene loss and extreme sequence divergence are, from a biological standpoint. The difficulty associated with distinguishing bona fide instances of gene loss from extreme sequence divergence prompted Aravind et al (2000) to acknowledge that the search results reflected both phenomena. However, the authors added (Aravind et al 2000, p 11324) that they 'believe that [loss and divergence] form a continuum, gene loss being, in many cases, the ultimate case of divergence.' This position is logical when we consider the potential for genes to have multiple functions. Genes that have a single function (or a very limited set of functions) might be subject to loss when selective pressures change. However, genes with a larger number of functions are more likely to be retained despite changes in selective pressures, although rates of sequence evolution at individual sites might exhibit substantial

changes. If a sufficient number of sites show increased rates, there will be more sequence divergence overall (a nonstationary covarion model may provide the best framework to understand this phenomenon, see Penny et al 2001 for discussion).

However, the relationship between gene loss and extreme sequence divergence may be somewhat more complex. It is possible to predict several possibilities if the loss of biological functions is viewed as the breaking of connections in a biological network (eg Figure 2). The connections in biological networks should be viewed broadly, ranging from the production of a metabolite used by a second enzyme to direct physical interactions among proteins. Although a change in selective pressures might allow a connection in the network to disappear, allowing either gene loss or extreme divergence, the network model suggests a direct relationship between gene loss and sequence divergence. In this model, selective pressures constraining specific sites in the more connected gene (eg node δ in Figure 2) would be related directly to maintaining the interaction with the less connected gene (eg node γ in Figure 2). This could be the basis for the modularity in gene loss and divergence noted by Aravind et al (2000) and Braun et al (2000).

To examine the connections among proteins encoded by genes that underwent loss and divergence in the *S. cerevisiae* lineage, I conducted a literature search for protein–protein interactions involving Ca^{2+}-binding proteins that had been lost in the *S. cerevisiae* lineage. Calcium-binding proteins were chosen because most divergent orthologues identified by Braun et al (2000) were involved in Ca^{2+} biology and some of these divergent orthologues are known to exhibit functional divergence from their orthologues in other eukaryotes. Human annexin A11 interacts physically with ALG-2 (Satoh et al 2002), a protein that has a divergent *S. cerevisiae* orthologue encoded by YGR058w. Annexin is one of the Ca^{2+}-binding proteins that had been lost in the *S. cerevisiae* lineage (Braun et al 1998). Strikingly, Satoh et al (2002) found that the amino-terminal 'GYP' extension of annexin A11 is necessary and sufficient for interaction with ALG-2. The domain is present in a subset of annexins, including the *N. crassa* annexin (Braun et al 1998), which is consistent with the possibility that fungal annexin and ALG-2 homologues form a complex.

The loss of annexin and retention of a divergent ALG-2 orthologue in *S. cerevisiae* suggests a model in which selective constraints upon some sites in ALG-2 were released owing to the loss of annexin, but the ALG-2 protein was retained owing to additional functions (Figure 7). The

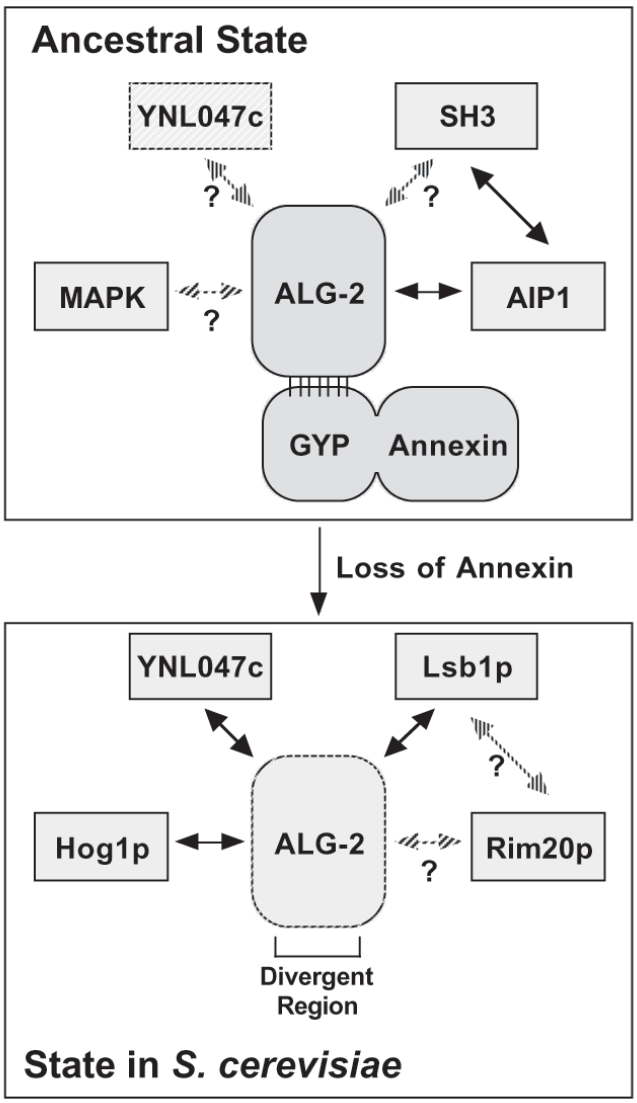

Figure 7 Model for evolutionary changes in protein–protein interactions involving ALG-2 in the *S. cerevisiae* lineage. Evidence for the protein–protein interactions in *S. cerevisiae* were obtained by considering large-scale protein–protein interaction datasets from the DIP database (Xenarios et al 2002) in light of phenotypic information from Giaever et al (2002) as well as other methods for the validation of protein–protein interactions (Deane et al 2002). Evidence for protein–protein interactions in the ancestor is from experiments with mammalian proteins (see Vito et al 1999; Chen et al 2000; Satoh et al 2002). The data necessary for a rigorous ancestral state reconstruction for protein–protein interactions are not available at present, so this model should simply be viewed as a heuristic tool.

function of fungal ALG-2 homologues is unclear, although the *S. cerevisiae* ALG-2 shows modest sensitivity to osmotic stress (data from Giaever et al 2002). Large-scale two-hybrid data (Uetz et al 2000) also indicate a physical interaction with a central regulator of the osmotic stress response (the Hog1p MAP kinase, see O'Rourke et al 2002 for review). Although these data are consistent with the hypothesis that the divergent orthologue of ALG-2 in *S. cerevisiae* was retained because of selection for a function mediated by a complex that contains Hog1p and ALG-2 (YGR058w), the

S. cerevisiae ALG-2 orthologue also appears to interact with additional proteins. Another interesting candidate for protein–protein interactions for ALG-2 (YGR058w) is the product of the *S. cerevisiae RIM20* gene, which is orthologous to the mammalian ALG-2 interacting protein AIP1 (Vito et al 1999). Although a Rim20p–YGR058w interaction has not been identified in large-scale protein–protein interaction screens, *rim20* mutants also show an osmotic stress phenotype.

Although the protein–protein interactions involving ALG-2 orthologues in different organisms have not been examined in detail, I present a model of evolutionary changes in these interactions to illustrate the types of predictions possible when data on gene loss and divergence are integrated with functional information (Figure 7). Ultimately, interactions among proteins in the query organism (in this example, *N. crassa*) must be examined to test predictions from these analyses. Many additional genes that encode physically interacting proteins are evident in the set of proteins that underwent loss or divergence in *S. cerevisiae*, providing a number of targets for analyses in other organisms.

Multiprotein complexes containing gene products that have undergone loss or divergence in *S. cerevisiae* include the splicosome, translation initiation factor eIF3, and the spindle pole body (Aravind et al 2000; Braun et al 2000; Braun 2002). In fact, both members of 61 pairs of interacting proteins from the set of protein–protein interactions validated by Deane et al (2002) were present in the set of 318 *S. cerevisiae* protein coding genes corresponding to divergent orthologues of *Sch. pombe* genes. This is a significantly greater number of pairs than expected by chance in sets of *S. cerevisiae* open reading frames (ORFs) with *Sch. pombe* homologues ($P < 0.001$; 95% confidence interval = 12–36 pairs, calculated by random sampling of ORF names without replacement).

Taken as a whole, these data are consistent with the close relationship between gene loss and sequence divergence and further suggest that it may be possible to examine protein–protein interactions using this type of information. Surveys of gene loss and divergence analyses highlight those parts of the genome that have undergone the greatest amount of change. Thus, genes that are identified in surveys of gene loss and divergence are more likely to exhibit functional difference from their orthologues in other groups of organisms, and researchers should be cautious when extrapolating functional data from systems in model organisms that have undergone an unusually high degree of change. Information from these surveys could allow researchers to focus data collection on sets of genes that are most likely to differ among organisms, rather than completely replicating large-scale datasets from model organisms (eg the large-scale datasets of protein–protein interactions and phenotype on *S. cerevisiae*; see Uetz et al 2000; Giaever et al 2002).

Gene loss and divergence in other eukaryotes

Salzberg et al (2001) emphasised the potentially universal impact of gene loss upon the evolution of eukaryotic genomes. In fact, they suggested that the common ancestor of the 'crown eukaryotes' (a group that includes animals, plants and fungi; see Knoll 1992 for additional details) may have had as many as 10 000 protein coding genes, of which as many as 30% were lost in different lineages. Although certain aspects of the model proposed by Salzberg et al (2001) are clearly unrealistic, the model predicted a number of genes lost in four independent eukaryotic lineages comparable to the number of putative prokaryote-to-vertebrate lateral transfers observed in the initial human genome drafts. In reality, the probability of loss is expected to differ both among genes and among eukaryotic lineages for a variety of reasons. However, this raises the question of precisely what those probabilities are for a variety of eukaryotic lineages.

As one might expect if gene loss has had a fairly universal impact upon eukaryotic genome evolution, specific instances of gene loss are evident in several complete genome sequences such as *Caenorhabditis elegans* and *Drosophila melanogaster* (Steele et al 1999; Aravind et al 2000; Braun 2002). However, it is currently thought that the fungi have experienced more gene loss than complex multicellular lineages. This hypothesis is consistent with estimates of fungal phylogeny indicating that unicellular fungi arose from multicellular ancestors (see Braun et al 2000 for discussion) and the fact that microsporidia, intracellular parasites related to fungi, possess extremely reduced gene complements (reviewed by Keeling and Fast 2002). To examine the impact of gene loss on another fungal lineage, I used *Sch. pombe* as the reference organism and *S. cerevisiae* as the query organism, using a non-redundant set of *S. cerevisiae* queries to conduct the searches of the reference organism (*Sch. pombe*) and the outgroup databases (underlined in Figure 6). This analysis revealed a total of 147 protein coding genes present in *S. cerevisiae* that have been lost in the *Sch. pombe* lineage and 54 *Sch. pombe* genes corresponding to divergent orthologues of *S. cerevisiae* proteins. Although the *S. cerevisiae* lineage lost a larger number of genes, both

unicellular fungi have lost a number of genes that have been retained in the other lineage. Comparisons with more complete sequence information from multicellular fungi such as *N. crassa* should prove interesting.

The candidates for loss in *Sch. pombe* also exhibit functional similarities like those evident in *S. cerevisiae*. One of the most striking is the loss of the Pxa1p and Pxa2p peroxisomal ABC transporter subunits. The *S. cerevisiae* gene that encodes Pxa1p has attracted substantial interest, because mutations in the human orthologue of this gene are responsible for adrenoleukodystrophy (Shani and Valle 1996), a rare inherited metabolic disorder that was the focus of the 1993 film 'Lorenzo's oil'. In fact, a number of other proteins involved in peroxisome function were evident in the list of genes that were lost in *Sch. pombe* (Table 2). A total of 12 *S. cerevisiae* genes annotated with gene ontology terms (The Gene Ontology Consortium 2000) related to 'peroxisome' were lost in the *Sch. pombe* lineage, a number significantly greater than expected based upon the total number of genes lost in the *Sch. pombe* lineage ($P = 2.5 \times 10^{-9}$, binomial test). It is likely that additional analyses, using other fungal query organisms, will reveal the loss of additional sets of functionally related genes in the *Sch. pombe* lineage.

Table 2 Proteins involved in peroxisome function that have been lost in *Sch. pombe*

Gene[a]	*S. cerevisiae (open reading frame)*	*Sch. pombe (E-value)*[b]	*Non-fungal (E-value)*[c]
FAT1	YBR041W	3×10^{-14}	10^{-73}
FAA2	YER015W	7×10^{-48}	8×10^{-87}
POX1	YGL205W	–[d]	10^{-73}
POT1	YIL160C	10^{-47}	10^{-100}
TES1	YJR019C	0.053	10^{-47}
PXA2	YKL188C	8×10^{-6}	6×10^{-86}
FOX2	YKR009C	4×10^{-18}	3×10^{-98}
ECI1	YLR284C	–	8×10^{-19}
CAT2	YML042W	–	5×10^{-75}
MLS1	YNL117W	–	10^{-135}
SPS19	YNL202W	2×10^{-23}	10^{-46}
PXA1	YPL147W	6×10^{-6}	2×10^{-99}

[a] Gene names, protein sequences, open reading frame names and annotations used to identify proteins localised to peroxisomes are from the *Saccharomyces* genome database at http://genome-www.stanford.edu/Saccharomyces

[b] Minimum *E*-value from a BLASTP search of annotated proteins from *Sch. pombe* and TBLASTN searches of genomic DNA and ESTs from *Sch. pombe*. Databases are from http://www.sanger.ac.uk/Projects/S_pombe (annotated proteins and genomic DNA) and http://www.ncbi.nlm.nih.gov/dbEST (ESTs).

[c] Minimum *E*-value from BLASTP searches of annotated proteins from *A. thaliana*, *C. elegans*, *D. melanogaster* and two different annotations of the human genome. Databases are from http://www.arabidopsis.org, http://www.sanger.ac.uk/Projects/C_elegans, http://www.fruitfly.org and ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/protein.

[d] No *Sch. pombe* database had a hit with $E \leq 0.01$.

The use of fungal queries from complex multicellular fungi, such as *N. crassa* and the basidiomycete *Phanerochaete chrysosporium*, have the potential to reveal the number of genes lost in both yeast lineages. In fact, one of the most interesting aspects of this pilot survey of gene loss in *Sch. pombe* was the absence of some genes known to have undergone loss or divergence in *S. cerevisiae* (eg the genes for annexin and ALG-2). These genes probably reflect independent loss in both yeast lineages, since most estimates of fungal phylogeny (eg Bruns et al 1992; Liu et al 1999; Lutzoni et al 2001) (see Figure 6) support the existence of an *N. crassa–S. cerevisiae* clade that excludes *Sch. pombe*. The existence of genes that are more likely to undergo loss is predicted by the hypothesis that the probability of gene loss is related to the number of connections involving each gene in biological networks (eg Figure 2). Additional factors, including the potential for similar biochemical functions to be mediated by distinct proteins (eg the ability of the products of the unrelated *pdxA–pdxJ* and *PDX1* [*SNZ*] genes to mediate pyridoxal phosphate synthesis; see Ehrenshaft et al 1999), will influence the probability of non-orthologous displacement, and hence the probability of gene loss as well. However, broader sampling of genome sequences will be necessary to examine the variance among genes in the probability of gene loss.

Moving these types of studies from fungi to other groups of eukaryotes is likely to prove challenging, despite the important role of loss in the evolution of genome content of these lineages (see Salzberg et al 2001). The large number of introns and complex patterns of alternative splicing evident in many plant and animal genomes complicates the annotation of protein sequences. In fact, analyses suggesting that genes have been retained in organisms with large genomes should be viewed with caution, since pseudogenes may be retained without function (also see the example of the human gene encoding CMAH that was described above). In fact, alignments to processed pseudogenes in translated database searches are likely to have higher *E*-values than alignments to a real gene because the processed pseudogene alignment will be uninterrupted. For this reason, it would be wise to examine alignments to translated genomic sequences that seem unusual based upon the absence of introns or limit searches to EST databases. Even when translation searches are limited to EST databases, the retention of expressed pseudogenes has the potential to be problematic (eg Chou et al 1998). Despite these problems, it is important to remember that the identification of a subset of protein coding genes that have been lost still has the potential to reveal interesting

patterns. Thus, screens for gene loss should adopt a practical viewpoint, attempting to identify a large proportion of genes that have undergone loss while limiting the number of genes incorrectly assigned to the set of lost genes, rather than striving for perfection.

The prospects and perils of domain loss and protein fission

Many proteins are made up of multiple domains (reviewed by Doolittle 1995; Henikoff et al 1997), where domains are defined as distinct segments of proteins capable of folding independently. Multidomain proteins present many problems for assigning orthology in large-scale genome comparisons. Although orthologues are defined as homologous genes related by speciation (Fitch 1970), evidence that two genes are orthologues is most useful in comparative genomics when genes show complete structural and functional correspondence (ie when orthologues represent the 'same' gene in different organisms) (see Koonin et al 2000 for details). However, proteins may be linked by the presence of a single orthologous domain but differ in the overall structure of the protein. Since individual domains often mediate specific biological functions, changes in domain structure are expected to be associated with functional changes.

High-throughput methods to identify orthologues, such as the 'bidirectional best-hit' database search criterion (see Overbeek et al 1999), can be problematic because proteins in multicellular eukaryotes often have a larger number of distinct domains than related proteins in unicellular eukaryotes (Koonin et al 2000). This accretion of additional domains in proteins appears to show a correlation with developmental complexity, but there is no reason to expect changes in domain structure to be unidirectional. Indeed, eukaryotic J domain proteins show evidence of both domain accretion and domain loss during evolution (Figure 8). The contribution of domain loss to the diversity of protein structures is presently unclear, but the possibility of domain loss raises the question of whether some putative gene loss events identified by large-scale surveys reflect domain loss instead (see Braun and Grotewold 2001).

Domain loss may be especially problematic when partial sequences are used for the query organism. Although collecting partial sequence data (EST or RST data) is less expensive than obtaining complete genome sequence data, incomplete queries can suggest gene loss if the sequence tag covers a single domain and domain loss has occurred. Despite this problem, Bean et al (2001) emphasised that the larger problem associated with using incomplete queries in gene loss surveys is likely to be an underestimation of gene loss.

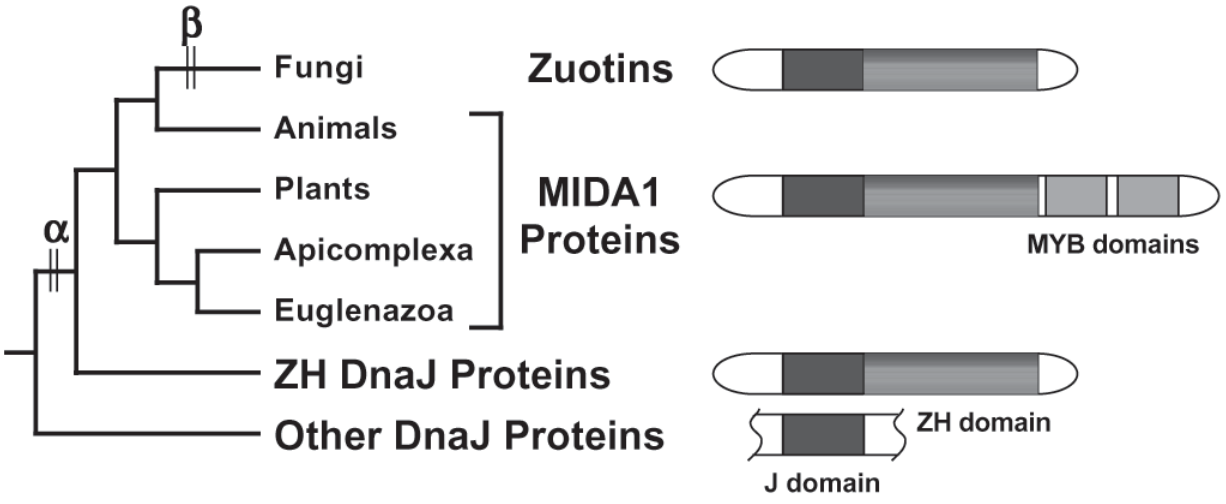

Figure 8 Complete domain loss associated with the origin of fungal Zuotins from MIDA1-like factors. Fungal Zuotins are characterised by the presence of a J domain and a region rich in charged residues called the ZH domain (Braun 2002). The ZH domain is shared by Zuotins, MIDA1-like factors and a second group of J domain proteins (including the J domain proteins YNL227c and At1g74250) and it was probably gained along branch α in this cladogram. MIDA1 proteins are similar to Zuotins, but they also contain two carboxyl-terminal MYB domains. Although a number of proteins containing MYB domains are present in fungi (Lipsick 1996), none of the other fungal MYB domains are closely related to those found in MIDA1 proteins. The phylogenetic analyses of Braun and Grotewold (2001) support a model in which Zuotins are related to MIDA1 proteins by complete loss of the MYB domains along branch β.

For example, the putative helix-turn-helix transcription factor (ORF G6G8.4) annotated by Bean et al (2001) would have had a high probability of being missed in a survey of gene loss that used EST queries. Only its DNA-binding domain is sufficiently conserved to permit the identification of orthologues in searches of databases from distantly related ascomycetes. Since many genes have relatively short conserved regions, it is likely that the estimate of the numbers of genes lost in the *S. cerevisiae* lineage obtained using EST data by Braun et al (2000) is an underestimate. If this is so, then the similarity of previous estimates of the number of genes lost in the *S. cerevisiae* lineage (from Braun et al 2000 and Aravind et al 2000) could reflect the fact that both surveys produced underestimates.

Distinguishing domain loss from gene loss is especially important because there are distinct pathways for domain loss. A conservative pathway for domain loss is defined as one that maintains the function of the original multidomain protein. One conservative pathway is protein fission (Snel et al 2000), which yields two or more distinct proteins that contain the set of domains present in the original protein. The conservative nature of this process reflects the fact that the independent proteins generated by protein fission are thought to form a multiprotein complex with biological activities similar to the ancestral multidomain protein. In fact, the ancestral protein before fission corresponds to a 'Rosetta stone' protein used to predict protein–protein interactions in computational genetics (Marcotte et al 1999). A second conservative pathway would be domain loss after gene duplication, as postulated for the plant *R2R3* MYB proteins (see Braun and Grotewold 1999; Dias et al 2003). In this example, domain loss might alter the function of one duplicate, but the other duplicate would retain the function of the ancestral protein.

Complete loss of domains might involve a pathway comparable to gene loss, in which no detectable orthologue of a domain that was ancestrally present has been retained. This appears to be the pathway that resulted in the fungal Zuotin proteins (Braun and Grotewold 2001; Copley et al 2002), which arose from MIDA1-like factors by complete loss of two MYB domains (Figure 8). As expected for domain loss, the origin of the Zuotins appears to be correlated with the evolution of a novel function. MIDA1-like factors appear to play a role in the regulation of cell division (reviewed by Braun and Grotewold 2001), possibly by binding transcriptional regulators (Shoji et al 1995). In contrast, Zuotin is associated with the ribosome and plays a role in fungal translation. Many J domain proteins are chaperones that cooperate with Hsp70 proteins (Kelley 1998), and Hsp70's that interact functionally with Zuotin include Ssb1/2p and Ssz1p (for details see Michimoto et al 2000; Gautschi et al 2002). Provocatively, Ssb1/2p and Ssz1p are divergent cytoplasmic Hsp70 proteins, encoded by genes that arose by duplication within the fungi (Braun 2002). Thus, an apparently novel function in fungal translation reflects modification of a J domain protein by domain loss and modification of Hsp70 proteins by sequence divergence after gene duplication.

To examine the potential impact of domain loss upon surveys of gene loss, several proteins that exhibit differences in domain structure similar to those observed for Zuotin were identified (Table 3). All proteins identified as candidates for domain loss have specific domains that are present in animals and plants, but absent in fungi. Human query sequences that have undergone domain loss in the fungi were used to search databases of annotated proteins from *Arabidopsis thaliana*, *S. cerevisiae* and *Sch. pombe*. Since top fungal hits of the animal and plant proteins were identified using a bidirectional

Table 3 Domain loss in the evolution of eukaryotic proteins and the impact of domain loss on the identification of divergent orthologues

		E-values[b]		
Protein	*Domain lost in fungi*[a]	*A. thaliana*	*S. cerevisiae*	*Sch. pombe*
MIDA1/Zuotin	MYB (PF00249)	3×10^{-46}	10^{-38}	2×10^{-53}
eIF4G	MA3 (PF02847)[c]	2×10^{-59}	**10^{-33}**	**2×10^{-43}**
TAFii250	Bromodomain (PF00439)	10^{-62}	2×10^{-55}	3×10^{-61}
Spt5p	KOW (PF00467)[d]	10^{-122}	**10^{-68}**	10^{-117}

[a] Pfam accession numbers for the domains that have been lost in fungi are listed.
[b] *E*-value from a BLASTP search using a human query sequence. Fungal *E*-values are presented in bold if they are 10-logs worse than the top hit in *A. thaliana*.
[c] Animal eIF4G orthologues also have a W2 domain (PF02020) after the MA3 domain. A weak hit to the W2 domain is evident in the At3g60240 open reading frame encoding eIF4G in *A. thaliana*, but it is found before the MIF4G domain (PF02854) present in all eIF4G orthologues. If this domain is related to the animal W2 domain by a rearrangement within the gene, the absence of a W2 domain in fungi could reflect domain loss as well.
[d] One of six KOW domains present in Spt5p orthologues is absent in fungi, along with some of the flanking protein sequence. This KOW domain is relatively divergent from the KOW consensus, but it is conserved between plants and animals.

best-hit criterion (data not shown), any candidates for loss or divergence would be classified as divergent orthologues. Using the 10-log criterion, one of these queries would have been identified as a divergent orthologue of the query sequence in both reference organisms and a second would have been identified as a divergent orthologue of the query sequence in *S. cerevisiae* (eIF4G and Spt5p).

Classification of proteins that have undergone domain loss as divergent orthologues of the query sequences may be desirable, since functional divergence may be associated with domain loss. However, since protein fission often represents a conservative pathway of domain loss (see above), distinguishing fission from the complete loss of a protein domain is important. One solution to this problem would be to compare putative divergent orthologues to sets of 'Rosetta stone' proteins from sets of domain linkage data (eg Enright et al 1999; Marcotte et al 1999). Ultimately, the difficulties associated with domain loss stress the importance of extending the approaches that have been used to highlight the loss and divergence of specific protein coding genes to allow the examination of additional types of genomic change.

Performing large-scale surveys of gene loss in prokaryotes

The identification of protein coding genes in prokaryotes has become fairly reliable, owing to the absence of splicosomal introns and the availability of excellent algorithms for gene identification (eg Delcher et al 1999; but see Cambillau and Claverie 2000 for an example of problems with prokaryotic gene annotation). Despite the greater reliability of annotation in prokaryotes, surveys of gene loss are actually more difficult owing to the relatively high rate of lateral gene transfer in prokaryotes (Doolittle 1999). Thus, examining patterns of gene loss in prokaryotes requires assumptions regarding the relative probability of gene loss versus lateral gene transfer.

The functional inferences that can be made regarding gene loss in prokaryotes may also be more limited. For example, the observation of 'modular' gene loss involving functionally related genes may reflect the fact that prokaryotes often cluster functionally related genes into operons (Dandekar et al 1998). Although loss of an operon is likely to reflect changes in selective pressure, just like the loss of multiple genes in eukaryotes, the observation that all genes in an operon were lost adds a limited amount of information to the original observation that the genes form an operon.

Despite the challenges posed by lateral gene transfer, heuristic approaches to assess whether loss or lateral transfer represents a better explanation for the distribution of specific prokaryotic genes are possible. For example, Henrissat et al (2002) suggested that genes associated with glycogen metabolism were lost in a variety of parasitic bacteria because of the broad distribution of these genes and their absence in parasitic bacteria. Indeed, other lines of evidence for extensive independent gene loss in parasitic bacteria (reviewed by Andersson and Andersson 1999) strongly suggest that the absence of these genes reflects loss. Another general pattern of gene loss evident in parasitic bacteria corresponds to the genes involved in DNA repair and recombination. For example, Moran and Mira (2001) highlighted 13 genes involved in DNA repair that were lost in the obligate endosymbiont *Buchnera aphidicola*. However, it is clear that distinct sets of DNA repair genes have been lost in different species of parasitic bacteria. For example, the *recA* gene has been lost in *B. aphidicola* while this species has retained the *recBCD* genes (Moran and Mira 2001). However, *recA* has been retained and the *recBCD* genes have been lost in other parasitic bacterial species, including another *Buchnera* species (Shingenobu et al 2000).

The existence of distinct patterns of gene loss in different parasitic or symbiotic prokaryotes is consistent with our model of biological systems as complex interacting sets of biological molecules (eg Figure 2). Since both *recA* and *recBCD* function in DNA repair and recombination, loss of one probably constrains loss of the other. Expanded sampling of prokaryotic genome sequences has the potential to provide more examples where loss of one gene limits the loss of a second, but it is unclear how many genomes it will be necessary to examine before this approach is broadly useful in computational genetics.

Surveys of gene acquisition and loss have also been conducted for free-living prokaryotes, and the larger number of distinct selective pressures faced by these organisms may ultimately provide more information. For example, a survey of genome change in the archael genus *Pyrococcus* (Ettema et al 2001) scored genes as having been lost if a gene was absent in one or more *Pyrococcus* species but present in the crenarchaeote *Aeropyrum pernix* and at least one euryarchaeote. This criterion is similar to the heuristic applied to fungal gene loss by Braun et al (2000), which assumed relatively limited lateral gene transfer into the query organism lineage. However, a more quantitative framework for the evaluation of gene loss in prokaryotes is clearly desirable.

Since the estimates of the number of genes that were lost depend upon the relative likelihood of gene loss and lateral gene transfer, applying different weights to these events when reconstructing ancestral genome content should be informative. Snel et al (2002) examined this issue using data from archaeal and proteobacterial genomes. As expected, applying higher costs to lateral gene transfer than to gene loss causes the number of inferred gene loss events to increase. However, an appealing aspect of the study by Snel et al (2002) was the fact that a modest penalty for lateral gene transfer (double or three times that of gene loss) was sufficient to move estimates of ancestral genome content for archaea and proteobacteria into the range of extant members of the same groups. In fact, this weight for lateral gene transfer is consistent with the weight implicit in the criterion used by Braun et al (2000) and Aravind et al (2000) for scoring loss in *S. cerevisiae*. Neither study examined the issue rigorously; weighting lateral transfer (gain) double that of gene loss is sufficient for parsimony methods to reconstruct the state for the ancestor of *S. cerevisiae* as presence of the gene when the observed distribution of genes met the criteria for loss in those studies (Braun 2002).

Despite the similarity of these weights to values implicit in previous surveys of gene loss and the appeal of the estimates of ancestral genome size that were obtained, attempts to examine gene loss in prokaryote genomes should be viewed with substantial caution. The approach of parsimony ancestral state reconstruction was used by Mirkin et al (2003) to examine the appropriate weight for lateral transfer relative to loss. Mirkin et al (2003) used different weights to reconstruct sets of genes present in the last universal common ancestor (LUCA) of extant organisms, and concluded that the set of genes reconstructed as present in the LUCA was largely consistent with a minimal gene complement necessary for life when lateral transfers and losses were weighted equally. This would suggest the rate of lateral transfer is much higher and the rate of gene loss lower than implied by the Snel et al (2002) study. Likewise, phylogenetic analyses of genes present in four archaeal genomes conducted by Nesbø et al (2001) suggested a high rate of lateral gene transfer within the archaea, and that these lateral transfers even included the 'core' of genes proposed to be refractory to lateral transfer by Makarova et al (1999). Both of these studies suggest that lateral gene transfer may have had an overwhelming impact upon prokaryotic genome evolution, a model supported by additional lines of evidence reviewed by Lawrence (2001) and Gogarten et al (2002). If lateral transfer occurs as frequently as suggested by these studies, the role for gene loss in shaping prokaryotic genome content is likely to be more limited than implied by Snel et al (2002) and accurate estimates of the impact of gene loss on prokaryotic genomes will be difficult to obtain.

However, the topological conflict among trees estimated using quartets of protein sequences such as those used by Nesbø et al (2001) may not necessarily reflect lateral gene transfer. Some conflict among estimates of phylogenetic trees is expected even when all genes have identical evolutionary histories (Penny et al 1982) and these sampling errors may be exacerbated by the use of limited taxon samples such as quartets (see Pollock 2002 for a review of the issues associated with analyses of limited taxon samples). Likewise, the heuristic for choosing the weight to apply to lateral transfers in the analyses conducted by Mirkin et al (2003) may be misleading. The complex nature of biological networks (eg Figure 2) is likely to make predictions of the minimal sets of proteins necessary for a minimal organism very difficult. Furthermore, estimating the set of genes necessary for viability of the LUCA may be complicated by the fact that a number of functions mediated by proteins in extant organisms may have been mediated by catalytic RNA or RNA–protein complexes in the LUCA (Penny and Poole 1999). Thus, it is unclear how to interpret the fact that equal weighting of lateral transfer and gene loss allowed reconstruction of a LUCA gene set similar to a minimal gene set necessary for viability. Nonetheless, the estimates of archaeal phylogeny for individual genes in Nesbø et al (2001) appear to show more conflict than phylogenetic trees estimated using different genes in animals, fungi and plants (eg the summary of previous analyses presented by Baldauf et al 2000). These results are consistent with the consensus in the evolutionary genomics community that prokaryotes exhibit a higher rate of lateral gene transfer than eukaryotes. Unfortunately, we are now left with the fact that obtaining better estimates for the rates and impact of gene loss in prokaryotic lineages will require accurate estimates of the rates of lateral gene transfer, and that these estimates will be difficult to obtain using available methods.

Despite the problems associated with estimating the probabilities of gene loss, it should still be clear that gene loss has had an important impact upon prokaryotic genome evolution. Most of the orthologue groups in the COG database (Tatusov et al 2001) have a limited number of sequences from each genome, so the acquisition of novel genes by lateral transfer from distinct lineages must be associated with the loss of orthologues present in the recipient genomes. Thus, higher estimates of the rate of lateral gene

transfer might actually be associated with higher rates of gene loss, supporting an overall higher rate of genomic turnover (see Lawrence 2001 for details). If the rate of genomic turnover is very high, parsimony reconstruction of ancestral character states (such as those used by Liberles et al 2002 and Mirkin et al 2003) may be unable to reconstruct the presence or absence of genes in ancestral nodes accurately. Although maximum likelihood reconstruction of ancestral character states should not be subject to similar problems owing to high rates of genomic flux, the differences in genome size among prokaryotes suggest that maximum likelihood estimates of ancestral character states will require a nonstationary model of genome change.

Improving large-scale surveys of gene loss and divergence

Some improvements to the large-scale surveys of gene loss were implemented in the novel comparisons presented in this review (eg Figure 4). Paralogues that reflect lineage-specific duplications in the query organism were removed prior to conducting the database searches, by sorting the sequences by size and adding sequences to a query set only if the sequence had a better hit in a non-fungal database than the query set. Ideally, all sequences would be retained in the query set but sequences that are more closely related to each other than to any sequences in the other databases could be linked in a manner analogous to the linking of ESTs into discontigs that was used in Braun et al (2000). Large-scale datasets on lineage-specific duplications available from studies such as that by Lespinet et al (2002) might facilitate this approach.

However, there is substantial room for improvements to the underlying philosophy of these surveys of gene loss. All of the surveys described here that used comparative database searches employed specific critical values for *E*-values to place genes in distinct categories. A more appropriate framework would examine the fit of multiple models for gene history (eg orthology with constant rates, orthology with unexpected divergence, loss with retention of a paralogue, etc) to the data and describe the model uncertainty in a direct manner. This approach would solve the problems associated with genes that have *E*-values close to critical values, and might allow the use of more realistic models (eg orthology with equal expected numbers of substitutions but an overdispersed model of evolution, like the models of evolution suggested by Gillespie 1991). Regardless of the full set of models tested, the finding that both loss with retention of a paralogue and accelerated sequence divergence fall within a credible set of models, while orthology with a constant rate of sequence evolution does not, would still be informative.

Another important factor to consider in this more flexible analytical framework is uncertainty in the species tree (see Figure 6). In principle, it should be possible to weight results that are dependent upon specific topologies by the probability that the relevant topology is correct. In fact, this approach of weighting ancestral state reconstructions by the posterior probability of the relevant topology has been used in a number of studies that reconstructed changes in whole organism traits (eg Lutzoni et al 2001). It may be more straightforward to implement this weighting in a reconciled tree framework, although there are a number of challenges associated with implementing the reconciled tree approach in a high-throughput framework (see above). Regardless of the approach used to identify candidates for loss, divergence or other types of genomic change, one should confirm the hypothesis using rigorous phylogenetic analyses of the relevant sequences.

Ultimately, the goal of this research would be to move the examination of gene loss and sequence divergence into a rigorous parametric framework. Although parametric models of genome evolution may be difficult to develop, they have the potential to be very informative. For example, the observation that two uncharacterised genes have been lost in a lineage that has undergone substantial gene loss, such as the highly reduced microsporidia (see Keeling and Fast 2002), should be viewed as providing limited information connecting the uncharacterised genes. However, a similar observation in a lineage that has lost few genes should be viewed as much more informative. Likewise, the overall rate of sequence evolution in an organism should be incorporated into these analyses, and parametric models of genome evolution could provide a natural framework for conducting these analyses. These models may require a large number of parameters, but it may be possible to deal with this problem by incorporating information from a variety of genomes in the form of informative priors for maximum a posteriori estimates of ancestral states.

Perhaps one of the most interesting aspects of gene loss is the possibility that the probability of loss is related to the number of connections involving each gene. This may be helpful for models that include variance among genes in the probability of loss, similar to models developed to accommodate differences in the rate at which specific sites

in proteins accept substitutions (see Penny et al 2001 for details). The observation that fungal annexin homologues are likely to have been lost independently in the *S. cerevisiae* and *Sch. pombe* lineages suggests that annexins are relatively prone to loss, at least in the fungal lineage. The fact that the number of connections involving various gene products (eg Figure 2) typically exhibit a good fit to a power law suggests a possible model. However, it will be necessary to propose some functions that relate the probability of gene loss to the number and type of connections in the network, and the most appropriate approach to this problem is unclear at this time. Regardless of the specific models that will eventually be developed, it will be important for parametric models of genome evolution to incorporate various types of biochemical information. In fact, developing these models in a framework that explicitly considers the biochemical nature of the changes that are being reconstructed may be necessary to extract useful information for computational genetic studies.

Discussion

The availability of complete genome sequences has allowed researchers to determine which genes are absent from an organism in addition to establishing the set of genes present in a genome. The absence of a gene can reflect loss of a gene that was present in an ancestor of the organism or the fact that the gene originated in a distinct lineage and was never present in the lineage examined. Thus, the study of gene loss should be viewed as problem in ancestral state reconstruction, since estimates of ancestral states ultimately provide the information about the directionality of change. Combining heuristics such as the results database searches (eg Aravind et al 2000; Braun et al 2000) with other approaches, such as the use of reconciled trees (reviewed by Page and Charleston 1998) and ancestral state reconstruction (eg Liberles et al 2002; Mirkin et al 2003), has the potential to reveal this directionality in genomic change.

Although tools for establishing that gene loss has occurred are available, it is difficult to examine the impact of gene absence on species that lack a certain gene. In this case, standard genetic tools such as the construction of mutants lacking the gene of interest are clearly inappropriate. Despite the difficulties associated with establishing the precise functional implications of specific instances of gene loss, there is substantial evidence that genes subject to loss and divergence in specific lineages are functionally related. However, the extent to which these changes have contributed to phenotypic differences among organisms is unclear, although there is a clear temporal correlation between gene loss and the evolution of specific features in certain organisms (eg the loss and divergence of specific Ca^{2+}-binding proteins in *S. cerevisiae* and characterised biochemical changes in some *S. cerevisiae* Ca^{2+}-binding proteins). Additional examples of changes in *S. cerevisiae* and *Sch. pombe* that may be correlated with gene loss were also reviewed in this paper, along with the provocative evidence that the loss of a specific gene involved in sialic acid biosynthesis may have played an important role in human evolution.

Regardless of the impact that specific gene loss events might have, the observation that functionally related genes tend to undergo loss or divergence is sufficient to allow computational geneticists to exploit information from surveys of gene loss and divergence for the annotation of uncharacterised genes. Although the computational approaches to this problem can be improved by incorporating additional types of information, the methods that are presently available have already revealed a number of interesting patterns. Since the relationship between gene loss, sequence divergence and protein–protein interactions suggests specific biochemical approaches to testing the hypotheses inferred, there may be many possibilities for moving these studies from the computer to the bench.

Acknowledgements

I am grateful to Rebecca Kimball for critical reading and constructive comments on this manuscript. The manuscript was greatly improved by comments from two anonymous reviewers. Todd Matulnik noted the apparent loss of the last KOW domain in Spt5p in the fungi and pointed it out to me. I would also like to express my gratitude to Rebecca Kimball, Don Natvig, Aaron Halpern, Erich Grotewold, Maggie Werner-Washburne and all of my colleagues at the University of New Mexico, Ohio State University and the University of Florida for encouragement and helpful discussions while I was developing the ideas reviewed in this manuscript.

References

Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387:489–93.

Albert R, Jeong H, Barabási A-L. 2000. Error and attack tolerance of complex networks. *Nature*, 406:378–82.

Altschul SF, Madden TL, Schäffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–402.

Andersson JO, Andersson SGE. 1999. Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev*, 9:664–71.

Angata T, Varki NM, Varki A. 2001. A second uniquely human mutation affecting sialic acid Biology. *J Biol Chem*, 276:40282–7.

Aravind L, Watanabe H, Lipman DJ, Koonin E. 2000. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci USA*, 97:11319–24.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, 290:972–7.

Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Duruflé L, Gaasterland T, Lopez P, Müller M et al. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc Natl Acad Sci USA*, 99:1414–19.

Barkai N, Leibler S. 1997. Robustness in simple biochemical networks. *Nature*, 387:913–17.

Bean LE, Dvorachek WH, Braun EL, Errett A, Saenz GS, Giles MD, Werner-Washburne M, Nelson MA, Natvig DO. 2001. Analysis of the *pdx-1* (*snz-1*/*sno-1*) region of the *Neurospora crassa* genome: correlation of pyridoxine-requiring phenotypes with mutations in two structural genes. *Genetics*, 157:1067–75.

Bock JB, Matern HT, Peden AA, Scheller RH. 2001. A genomic perspective on membrane compartment organization. *Nature*, 409:839–41.

Braun EL. 2002. Unpublished analyses of loss, duplication, and divergence during evolution. In possession of the author.

Braun EL, Grotewold E. 1999. Newly discovered plant c-*myb*-like genes rewrite the evolution of the plant *myb* gene family. *Plant Physiol*, 121:21–4.

Braun EL, Grotewold E. 2001. Fungal Zuotin proteins evolved from MIDA1-like factors by lineage-specific loss of MYB domains. *Mol Biol Evol*, 18:1401–12.

Braun EL, Halpern AL, Nelson MA, Natvig DO. 2000. Large-scale comparison of fungal sequence information: mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae*. *Genome Res*, 10:416–30.

Braun EL, Kang S, Nelson MA, Natvig DO. 1998. Identification of the first fungal annexin: analysis of annexin gene duplications and implications for eukaryotic evolution. *J Mol Evol*, 47:531–43.

Brenner SE, Chothia C, Hubbard TJ. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci USA*, 95:6073–8.

Bruns TD, Vilgalys R, Barns SM, Gonzalez D, Hibbett DS, Lane DJ, Simon L, Stickel S, Szaro T, Weisburg WG. 1992. Evolutionary relationships within the fungi: analyses of nuclear small subunit rRNA sequences. *Mol Phylogenet Evol*, 1:231–41.

Cambillau C, Claverie J-M. 2000. Structural and genomic correlates of hyperthermostability. *J Biol Chem*, 275:32383–6.

Charleston MA. 1998. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Math Biosci*, 149:191–223.

Chen B, Borinstein SC, Gillis J, Sykes VW, Bogler O. 2000. The glioma-associated protein SETA interacts with AIP1/Alix and ALG-2 and modulates apoptosis in astrocytes. *J Biol Chem*, 275:19275–81.

Chou H-H, Hayakawa T, Diaz S, Krings M, Indriati E, Leakey M, Pääbo S, Satta Y, Takahata N, Varki A. 2002. Inactivation of CMP-*N*-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proc Natl Acad Sci USA*, 99:11736–41.

Chou H-H, Takematsu H, Diaz S, Iber J, Nickerson E, Wright KL, Muchmore EA, Nelson DL, Warren ST, Varki A. 1998. A mutation in human CMP-sialic acid hydroxylase occurred after the *Homo-Pan* divergence. *Proc Natl Acad Sci USA*, 95:11751–6.

Copley RR, Letunic I, Bork P. 2002. Genome and protein evolution in eukaryotes. *Curr Opin Chem Biol*, 6:39–45.

Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23:324–8.

Deane CM, Salwinski L, Xenarios I, Eisenberg D. 2002. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1:349–56.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*, 27:4636–41.

Dias AP, Braun EL, McMullen MD, Grotewold E. 2003. Recently duplicated maize *R2R3 Myb* genes provide evidence for distinct mechanisms of evolutionary divergence after duplication. *Plant Physiol*, 131:610–20.

Doolittle RF. 1995. The multiplicity of domains in proteins. *Annu Rev Biochem*, 64:287–314.

Doolittle WF. 1999. Lateral genomics. *Trends Genet*, 15:M5–M8.

Ehrenshaft M, Bilski P, Li MY, Chignell CF, Daub ME. 1999. A highly conserved sequence is a novel gene involved in de novo vitamin B6 biosynthesis. *Proc Natl Acad Sci USA*, 96:9374–8.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90.

Ettema T, van der Oost J, Huynen M. 2001. Modularity in the gain and loss of genes: applications for function prediction. *Trends Genet*, 17:485–7.

Feng D-F, Cho G, Doolittle RF. 1997. Determining divergence times with a protein clock: update and reevaluation. *Proc Natl Acad Sci USA*, 94:13028–33.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst Zool*, 19:99–113.

Gaasterland T, Ragan MA. 1998a. Constructing multigenome views of whole microbial genomes. *Microb Comp Genomics*, 3:177–92.

Gaasterland T, Ragan MA. 1998b. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb Comp Genomics*, 3:199–217.

Gautschi M, Mun A, Ross S, Rospert S. 2002. A functional chaperone triad on the yeast ribosome. *Proc Natl Acad Sci USA*, 99:4209–14.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418:387–91.

Gillespie JH. 1991. The causes of molecular evolution. Oxford: Oxford Univ Pr.

Giribet G, Distel DL, Polz M, Sterrer W, Wheeler WC. 2000. Triploblastic relationships with emphasis on the acoelomates and the position of Gnathostomulida, Cycliophora, Platyhelminthes, and Chaetognatha: a combined approach of 18S rDNA sequences and morphology. *Syst Biol*, 49:539–62.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol*, 19:2226–38.

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Zool*, 28:132–63.

Gupta GD, Heath IB. 2002. Predicting the distribution, conservation, and functions of SNAREs and related proteins in fungi. *Fungal Genet Biol*, 36:1–21.

Harrison PM, Kumar A, Lang N, Snyder M, Gerstein M. 2002. A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res*, 30:1083–90.

Harvey PH, Pagel MD. 1991. The comparative method in evolutionary biology. Oxford: Oxford Univ Pr.

Henikoff S, Greene EA, Pietrokowski S, Bork P, Attwood TK, Hood L. 1997. Gene families: the taxonomy of protein paralogs and chimeras. *Science*, 278:609–14.

Henrissat B, Deleury E, Coutinho PM. 2002. Glycogen metabolism loss: a common marker of parasitic behaviour in bacteria? *Trends Genet*, 18:437–40.

Huelsenbeck JP, Larget B, Miller RE, Ronquist F. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst Biol*, 51:673–88.

Jeong H, Mason SP, Barabási A-L, Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature*, 411:41–2.

Kawano T, Koyama S, Takematsu H, Kozutsumi Y, Kawasaki H, Kawashima S, Kawasaki T, Suzuki A. 1995. Molecular cloning of cytidine monophospho-*N*-acetylneuraminic acid hydroxylase. Regulation of species- and tissue-specific expression of *N*-glycolylneuraminic acid. *J Biol Chem*, 270:16458–63.

Keeling PJ, Fast NM. 2002. Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annu Rev Microbiol*, 56:93–116.

Kelley WL. 1998. The J-domain family and the recruitment of chaperone power. *Trends Biochem Sci*, 23:222–7.

Knoll AH. 1992. The early evolution of eukaryotes: a geological perspective. *Science*, 256:622–7.

Koonin EV, Aravind L, Kondrashov AS. 2000. The impact of comparative genomics on our understanding of evolution. *Cell*, 101:573–6.

Koonin EV, Mushegian AR, Bork P. 1996. Non-orthologous gene displacement. *Trends Genet*, 12:334–6.

Kuma K, Nikoh N, Iwabe N, Miyata T. 1995. Phylogenetic position of *Dictyostelium* inferred from multiple protein data sets. *J Mol Evol*, 41:238–46.

Lawrence JG. 2001. Catalyzing bacterial speciation: correlating lateral transfer with genetic headroom. *Syst Biol*, 50:479–96.

Lespinet O, Wolf YI, Koonin EV, Aravind L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res*, 12:1048–59.

Liberles DA, Thorén A, von Heijne G, Elofsson A. 2002. The use of phylogenetic profiles for gene predictions. *Curr Genomics*, 3:131–7.

Lipsick JS. 1996. One billion years of Myb. *Oncogene*, 13:23–35.

Liu YJJ, Whelen S, Benjamin DH. 1999. Phylogenetic relationships among ascomycetes: evidence from an RNA polymerase II subunit. *Mol Biol Evol*, 16:1799–808.

Lutzoni F, Pagel M, Reeb V. 2001. Major fungal lineages are derived from lichen symbiotic ancestors. *Nature*, 411:937–40.

Madera M, Gough J. 2002. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res*, 30:4321–8.

Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV. 1999. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res*, 9:608–28.

Marcotte EM. 2000. Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol*, 10:359–65.

Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285:751–3.

Marcotte EM, Xenarios I, van der Bliek AM, Eisenberg D. 2000. Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci USA*, 97:12115–20.

Michimoto T, Aoki T, Toh-e A, Kikuchi Y. 2000. Yeast Pdr13p and Zuo1p molecular chaperones are new functional Hsp70 and Hsp40 partners. *Gene*, 257:131–7.

Mirkin BG, Fenner TI, Galperin MY, Koonin EV. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol*, 3:2.

Moran NA, Mira A. 2001. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol*, 2:research0054.1–0054.12.

Nesbø CL, Boucher Y, Doolittle WF. 2001. Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. *J Mol Evol*, 53:340–50.

Nielsen C. 1995. Animal evolution. Oxford: Oxford Univ Pr.

Oltvai ZN, Barabási A-L. 2002. Life's complexity pyramid. *Science*, 298:763–4.

O'Rourke SM, Herskowitz I, O'Shea EK. 2002. Yeast go the whole HOG for the hyperosmotic response. *Trends Genet*, 18:405–12.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA*, 96:2896–901.

Page RDM. 1994. Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics*, 10:155–73.

Page RDM, Charleston MA. 1998. Trees within trees: phylogeny and historical associations. *Trends Ecol Evol*, 13:356–9.

Pearson WR. 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol*, 132:185–219.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA*, 96:4285–8.

Penny D, Foulds LR, Hendy MD. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature*, 297:197–200.

Penny D, McComish BJ, Charleston MA, Hendy MD. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J Mol Evol*, 53:711–23.

Penny D, Poole A. 1999. The nature of the last universal common ancestor. *Curr Opin Genet Dev*, 9:672–7.

Pollock DD. 2002. Genomic biodiversity, phylogenetics and coevolution in proteins. *Appl Bioinform*, 1:1–12.

Roelofs J, Van Haastert PJ. 2001. Genes lost during evolution. *Nature*, 411:1013–14.

Ronquist F. 1998. Three-dimensional cost-matrix optimization and maximum cospeciation. *Cladistics*, 14:167–72.

Salter LA, Pearl DK. 2001. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Syst Biol*, 50:7–17.

Salzberg SL, White O, Peterson J, Eisen JA. 2001. Microbial genes in the human genome: lateral transfer or gene loss? *Science*, 292:1903–6.

Satoh H, Shibata H, Nakano Y, Kitaura Y, Maki M. 2002. ALG-2 interacts with the amino-terminal domain of annexin XI in a Ca^{2+}-dependent manner. *Biochem Biophys Res Commun*, 291:1166–72.

Shani N, Valle D. 1996. A *Saccharomyces cerevisiae* homolog of the human adrenoleukodystrophy transporter is a heterodimer of two half ATP-binding cassette transporters. *Proc Natl Acad Sci USA*, 93:11901–6.

Shingenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, 407:81–6.

Shoji W, Inoue T, Yamamoto T, Obinata M. 1995. MIDA1, a protein associated with Id, regulates cell growth. *J Biol Chem*, 270:24818–25.

Sicheritz-Pontén T, Andersson SGE. 2001. A phylogenomic approach to microbial evolution. *Nucleic Acids Res*, 29:545–52.

Snel B, Bork P, Huynen M. 2000. Genome evolution. Gene fusion vs. gene fission. *Trends Genet*, 16:9–11.

Snel B, Bork P, Huynen MA. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res*, 12:17–25.

Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR. 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*, 411:940–4.

Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. *Science*, 297:89–91.

Steele RE, Stover NA, Sakaguchi M. 1999. Appearance and disappearance of *Syk* family protein-tyrosine kinase genes during metazoan evolution. *Gene*, 239:91–7.

Stiller W, Hall BD. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. *Mol Biol Evol*, 16:1270–9.

Storm CEV, Sonnhammer ELL. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18:92–9.

Struhl K. 1999. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, 98:1–4.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, 29:22–8.

Taylor TN, Hass T, Kerp H. 1999. The oldest fossil ascomycetes. *Nature*, 399:648.

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Genet*, 25:25–9.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P et al. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–7.

Vito P, Pellegrini L, Guiet C, D'Adamio L. 1999. Cloning of AIP1, a novel protein that associates with the apoptosis-linked gene ALG-2 in a Ca^{2+}-dependent reaction. *J Biol Chem*, 274:1533–40.

Wolf YI, Koonin EV. 2001. Origin of an animal mitochondrial DNA polymerase subunit via lineage-specific acquisition of a glycyl-tRNA synthetase from bacteria of the *Thermus–Deinococcus* group. *Trends Genet*, 17:431–3.

Wood V, Gwilliam R, Rajandream M-A, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature*, 415:871–80.

Wu C-I, Li W-H. 1985. Evidence for higher rates of nucleotide substitutions in rodents than in man. *Proc Natl Acad Sci USA*, 82:1741–5.

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30:303–5.

Yoder AD, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. *Mol Biol Evol*, 17:1081–90.

Zmasek CM, Eddy SR. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, 17:821–6.

Zmasek CM, Eddy SR. 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinform*, 3:14.