Estimating the tempo and mode of gene family evolution from comparative genomic data

Matthew W. Hahn,[1,6,7] Tijl De Bie,[4,6] Jason E. Stajich,[5] Chi Nguyen,[2] and Nello Cristianini[3]

[1]*Center for Population Biology,* [2]*Department of Computer Science, and* [3]*Department of Statistics, University of California, Davis, California 95616, USA;* [4]*ISIS Research Group, University of Southampton, Southampton, SO17 1BJ, United Kingdom;* [5]*Department of Molecular Genetics and Microbiology, Duke University, Durham, North Carolina 27708, USA*

Comparison of whole genomes has revealed that changes in the size of gene families among organisms is quite common. However, there are as yet no models of gene family evolution that make it possible to estimate ancestral states or to infer upon which lineages gene families have contracted or expanded. In addition, large differences in family size have generally been attributed to the effects of natural selection, without a strong statistical basis for these conclusions. Here we use a model of stochastic birth and death for gene family evolution and show that it can be efficiently applied to multispecies genome comparisons. This model takes into account the lengths of branches on phylogenetic trees, as well as duplication and deletion rates, and hence provides expectations for divergence in gene family size among lineages. The model offers both the opportunity to identify large-scale patterns in genome evolution and the ability to make stronger inferences regarding the role of natural selection in gene family expansion or contraction. We apply our method to data from the genomes of five yeast species to show its applicability.

[Supplemental material is available online at www.genome.org.]

One of the major goals of evolutionary biology has been to identify the genetic changes underlying phenotypic differences between organisms, and to distinguish the evolutionary forces responsible for these changes. Past studies have necessarily focused on small numbers of nucleotide differences between orthologous genes, largely because of the technical limitations on DNA sequence collection. The recent sequencing of many whole genomes, however, has erased this limitation. Researchers may now focus on large-scale genomic differences between organisms that play an important role in adaptive evolution, including large changes in the size of gene families (e.g., Tatusov et al. 1997; Lander et al. 2001; Snel et al. 2002; Lynch and Conery 2003).

While the newfound ability to observe gene family expansions and contractions has stimulated many new hypotheses, we still lack a statistical framework that would allow for strong inferences regarding gene family evolution. Especially interesting to evolutionary studies are the causes of changes in gene family size. Unlike the analysis of nucleotide sequence evolution—where there are well-accepted methods for testing for the action of natural selection (e.g., Yang and Bielawski 2000)—there are no such methods in the analysis of gene family evolution. Generally, researchers have ascribed large differences in gene family size between genomes to natural selection, without any consideration of the expected difference in size due to random gene gain or loss over long periods of time (e.g., Oakeshott et al. 1999; Garczarek et al. 2000; Lander et al. 2001; Szathmary et al. 2001; Holt et al. 2002; Lespinet et al. 2002; Ranson et al. 2002; Copley et al. 2003; Lutfalla et al. 2003). While many of these differences may certainly be due to natural selection promoting the expansion or contraction of gene family size, most are simple comparisons in which one species is found to have a larger or smaller number of genes.

The inability to make statistical inferences about the role of natural selection in the evolution of gene family size may be due to the lack of a null model. With no expectation for how similar or different in size families are likely to be, researchers are unable to make probabilistic statements about observed disparities. While simple statements about the equivalence of two numbers can be made with tests of homogeneity (such as a χ^2), these tests do not take into account the time since divergence of two taxa. Observing a gene family with 100 members in one taxa and 50 in another is certainly striking if they have diverged for 5 million years, but if they have not shared a common ancestor for 250 million years the biological significance of the difference is less obvious. In addition, when data are available on gene family size in more than two taxa, it would be informative to use phylogenetic relationships among the species to identify lineage- or branch-specific expansions and contractions (e.g., Lespinet et al. 2002). A statistical model of gene family evolution that allows for both hypothesis testing and phylogenetic inference, therefore, would be very useful.

We propose to use the well-studied stochastic birth and death (BD) process as a model for gene family evolution. Birth and death models have been widely studied in statistics (Darwin 1956; Bailey 1964; Feller 1968), and have also found use in population genetics and phylogenetics (e.g., Slatkin and Rannala 1997; Sims and McConway 2003). The observation in multiple genomes that both gene family sizes and gene duplicate ages are approximately Poisson-Dirichlet distributed suggested that they could be explained by a random gain and loss process (Huynen and van Nimwegen 1998; Lynch and Conery 2000, 2003; Yanai et al. 2000; Qian et al. 2001; Karev et al 2002; Gu and Zhang

[6]These authors contributed equally to this work.
[7]Corresponding author.
E-mail mwh@Indiana.edu; fax (812) 855-6705.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3567505.

2004; Zhang and Gu 2004). Indeed, the first use of stochastic birth and death models for studying gene domain duplication and deletion was by Karev et al. (2002), and for studying gene duplication and deletion was by Reed and Hughes (2004) and Gu and Zhang (2004). Karev et al. (2002) showed that a random BD model explained the distribution of gene family sizes within a genome very well. Here we attempt to extend this approach to study divergence in gene families between species. It should be noted that stochastic BD processes are quite different from the conceptual model of gene birth and death used by Nei and colleagues to explain sequence similarity among closely linked gene duplicates (Nei et al. 1997).

In this study we associate the evolution of a gene family over a phylogeny with a probabilistic graphical model (PGM). The use of such a PGM allows for probabilistic inferences on the rate and direction of change in gene family size. Furthermore, we show how this methodology can be used to identify those families and those branches that are evolving nonrandomly. We demonstrate the usefulness of our approach on the whole genomes of five closely related yeast species—*Saccharomyces cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus*.

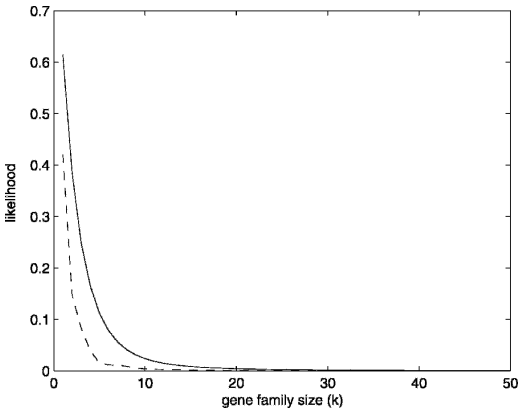

Figure 1. Here, we visually explain the "large family bias" problem. The solid line shows the likelihood of gene families with sizes ($k(k(k(kk))))$ as a function of k, for values of k from 1 to 50. The dashed line shows the average likelihood of gene families evolved from a common ancestor with family size equal to k. The average is computed over 100 random samples for each value of k. Clearly, the likelihoods for large gene families are consistently and significantly smaller.

Results

Calculating the likelihood of gene family data

In order to draw statistically motivated conclusions from gene family-size data in several related species, we use a probabilistic graphical model (PGM) (Lauritzen 1996; M.I. Jordan, in prep.) that represents the probability distribution over the observed gene family data. By specifying a stochastic birth and death (BD) model and a prior distribution for the common ancestor (or root node), the graphical models machinery makes it possible to efficiently compute the likelihood of the observed data by a process called marginalization (see Methods; Supplemental materials). By using this likelihood as a test statistic, a corresponding *P*-value can be computed (see next section). In addition, the PGM approach provides a way to infer the most likely values of ancestral states.

In this study we are interested in assessing the likelihood of gene families with respect to the BD model, independent of the unknown gene family size in the common ancestor. In other words, the prior on the root node value should be noninformative, and a natural choice seems to be the uniform distribution (cf. Felsenstein 1981). Unfortunately, even a uniform prior introduces an undesirable bias here, similar to other cases in phylogenetics (e.g., Zwicki and Holder 2004). In our case the use of such a uniform prior consistently attributes larger likelihoods to smaller gene families (see Fig. 1). In addition, since the *P*-values presented in the next section are computed as the probability that a random gene family has a likelihood smaller than that of the observed gene family, a uniform prior would result in consistently smaller *P*-values for large gene families. Other priors would introduce other biases, most often also favoring small gene families.

An intuitive explanation for the bias we observe, which we refer to as the "large family bias," is that a small family size in the common ancestor generally leads to small family sizes in the leaf species. The number of possible assignments from an ancestrally small gene family is relatively small, such that the likelihood of any individual assignment will be relatively large (since likelihoods sum to one). Gene families that were large in the common ancestor, on the other hand, will give rise to many more outcomes in the leaf nodes, and will thus tend to have smaller likelihoods and *P*-values for any individual outcome.

In principle, one can compensate for this bias by using a prior that heavily favors large family sizes. However, this is hard to do in practice and theoretically unsatisfying (such a prior would have to be improper, meaning that it does not integrate to 1). It may also be undesirable, as it relies on the assumption that small and large gene families evolve in similar ways—an assumption we do not wish to make if it can be avoided. Therefore, we prefer to use an exact, if slightly more involved approach that solely depends on the relative sizes of the gene families in the leaf node species and avoids the use of a prior on the root family size by treating it as a nuisance parameter (see, e.g., Lindsey 1996; Demortier 2003). To achieve this, our method relies on conditional likelihoods: likelihoods that are conditioned on a specific value for the root family size and can be computed just as efficiently by a similar marginalization procedure. In the next section we explain how these conditional likelihoods can be used to calculate conditional *P*-values.

Apart from an efficient method to compute likelihoods and conditional likelihoods, PGMs make it possible to compute the most likely assignment of the unspecified internal nodes; here, the ancestral gene family sizes. The algorithm is a variant of the marginalization procedure and is known in the graphical models literature as the max-product algorithm. For more details, we refer the reader to the relevant literature (e.g., Pearl 1988; M.I. Jordan, in prep.).

Furthermore, our framework also makes it possible to estimate the maximum likelihood value of λ, the birth and death rate parameter for our phylogenetic tree. This parameter describes the probability that any gene will be gained or lost, and hence has a large effect on the rate of gene family evolution. In the Discussion we compare our estimate of λ to the estimate of Lynch and Conery (2003) that was taken from just the *S. cerevisiae* genome sequence, and show that the two are very close.

Testing hypotheses about gene family evolution

Often we wish to know how probable it is to observe gene family data under the null hypothesis of random change. Because the

BD model uses information about the time in the phylogenetic tree and the birth and death rates of genes, it offers an ideal null model for hypothesis testing. Using a BD model in this way makes it possible to identify gene families that have undergone unusual expansions or contractions. This method furthermore enables us to identify the branch in the phylogeny upon which the unlikely change took place.

As argued above, likelihoods or conditional likelihoods cannot directly be used to identify unusual gene families, because larger gene families will by necessity result in lower likelihoods under a stochastic BD process (the "large family bias"). Instead, we can use our conditional likelihoods as test statistics to calculate conditional *P*-values, each one conditioned on one of the possible root-node assignments. Such a conditional *P*-value is defined as the probability that a random gene family (with fixed root family size) has a smaller conditional likelihood than the given gene family. Then, because the true root-node value is unknown, we conservatively pick the largest conditional *P*-value, which we can show to represent a tight upper bound on the true *P*-value in our problem (see Methods; Supplemental material). Such an upper bound on the *P*-value is called a supremum *P*-value in statistics, and it is often used for composite hypothesis testing with one or more nuisance parameters (Lehmann 1959; Demortier 2003). Because of its tightness as an upper bound in our problem, we refer to the supremum *P*-value as simply the *P*-value in the remainder of this study. In the Methods section we show how it can efficiently and accurately be computed using a sampling procedure.

Furthermore, we propose two methods to identify the branch in the phylogeny upon which nonrandom changes occurred (for families with a low *P*-value). Our first method computes a *P*-value corresponding to the observed data after the deletion of one branch in the PGM, and this once for each branch (for each gene family). If, after the deletion of a branch, the resulting *P*-value rises above some threshold *P*-value (0.01 here), then the branch that was cut is implicated in nonrandom evolution. Our second method uses a likelihood ratio test to compare a model allowing the λ parameter to vary along each branch singly to the model with one λ for the whole tree (see Methods; Supplemental materials). It is notable that, in all cases, the branch with the largest likelihood ratio was also the branch that yielded the largest *P*-value after cutting it, as computed by the first method.

Global view of *Saccharomyces* gene family evolution

We used the machinery described above to study the evolution of gene family size in five whole fungal genomes. To our knowledge, the five sequenced *Saccharomyces* genomes are the best example of a closely related group of eukaryotes, where multiple whole genomes have been sequenced and where there is also a well-supported phylogenetic tree with branch lengths.

The consensus phylogenetic tree of the five *Saccharomyces* species (Fig. 2) comes from the study of Rokas et al. (2003) that used 106 orthologous genes from each of the species, singly and by concatenation. The tree had 100% bootstrap support at every node. In Newick notation, the tree in Figure 2 is written (*S. bayanus* (*S. kudriavzevii*(*S. mikatae*(*S. paradoxus S. cerevisiae*)))). Branch lengths were inferred from the data in Rokas et al. (2003) and Kellis et al. (2003). They are indicated in Figure 2 as time, *t*, in million years. We estimated the evolutionary rate parameter λ as 0.002 per million years (see Supplemental materials).

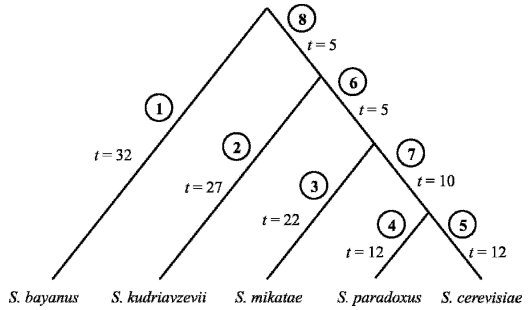

Figure 2. The phylogenetic tree. Branch lengths *t* are given in millions of years. The branch numbers used in this study are shown in circles.

To define gene families, we took all of the genes in all five species together and generated a pairwise matrix of distances among genes (see Supplemental materials). We then clustered genes using the TRIBE-MCL algorithm (Van Dongen 2000; Enright et al. 2002), and counted the number of genes in each family that came from each species. By clustering all of the genes at the same time, we are able to confidently compare the size of families between genomes.

In the 32 million years since the most recent common ancestor of the five species, 1254 of the 3517 gene families shared among them have changed in size; the remaining set are monomorphic across the tree (of course, equal numbers of losses and gains in any single gene family will be unobservable). Using our PGM we were able to infer the most likely ancestral gene family sizes for all of these gene families. This makes it possible to count changes in gene family size on all eight branches of the tree, and enables us to infer their direction by a comparison of the species at the top and bottom of each branch in the tree. Expansions outnumbered contractions on four of the eight branches, and contractions outnumbered expansions on the remaining four. Table 1 shows the number of families that expanded, contracted, or stayed the same on each branch of the tree.

We can see that along branches 2 and 3, leading to *S. kudriavzevii* and *S. mikatae*, many more families have expanded than contracted. Concomitant with this, these two genomes have more genes (7144 and 7236) than any of the other three (6265, 6128, and 6700 for *S. bayanus*, *S. paradoxus*, and *S. cerevisiae*; see

Table 1. The number of gene families that showed an expansion, no change, or a contraction along the eight branches, according to the most likely assignments of the gene family sizes of the ancestors

Branch #	Expansions	No change	Contractions	Average expansion
1 (*t* = 32)	97	3181	239	−0.050
2 (*t* = 27)	383	3032	102	0.095
3 (*t* = 22)	509	2922	86	0.147
4 (*t* = 12)	96	3383	38	0.019
5 (*t* = 12)	44	3426	47	0.021
6 (*t* = 5)	3	3491	23	−0.005
7 (*t* = 10)	10	3313	194	−0.052
8 (*t* = 5)	2	3515	0	0.001

The first column contains the branch number, along with the length of the branch, *t,* in millions of years. The next three columns show how often an expansion, no change, or a contraction occurred along this branch. The last column shows the average gene family expansion among all families along each branch, where a contraction is counted as a negative expansion.

Methods). This correlation is likely due to the expanding gene families and not to some other aspect of genome evolution; the average gene family size is larger in *S. kudriavzevii* and *S. mikatae* than in the other three species (1.56 and 1.61 vs. 1.41, 1.43, and 1.43).

We can also examine the average change in gene family size along each branch (Table 1). Again, we see that branches 2 and 3 have the largest positive changes of any branch, supporting the role of gene family expansion in genome expansion in these species. Though branch 5 (leading to *S. cerevisiae*) has slightly more contractions than expansions, the net change in family size is positive on average (0.021). Examining the data reveals the reason for this apparent contradiction: the RNaseH and helicase gene families have had huge expansions along this lineage (see Discussion). If we remove these two families, the average net change along this branch becomes negative (-0.002). In general, however, most changes in gene family size are quite small, and the resulting average change is correlated with the number of expansions and contractions.

Because an ancient genome duplication most likely preceded the common ancestor of these *Saccharomyces* species (Kellis et al. 2004), it may be that many of the patterns we see are due to differential loss of genes among lineages. If this nonequilibrium condition is correct, then many inferred expansions along one lineage or another may, in fact, be lineage-specific retention of specific families. Nonetheless, our identification of unusually evolving branches is still correct (see next section); it is only the process responsible for these deviations that remains a question.

Identification of unusually evolving gene families in *Saccharomyces*

As explained above, the PGM also allows us to compute *P*-values to identify gene families that are highly unlikely under the random BD process. Of the 1254 gene families that differed in number between genomes, 58 had *P*-values <0.01 (35 are expected). The unlikely families are summarized in Table 2 and in the Supplemental materials, along with the specific branch that is responsible for the violation (when such a branch could be identified). The two methods that we used to identify the offending branch agreed in every case.

Two of the most unlikely gene families in Table 2, where it seems difficult to explain the low *P*-value by just one branch, correspond to transposable elements (TEs). While it is interesting to see these large changes, transposable elements violate the assumptions of the BD model in a number of ways and it can therefore be seen as a validation of our approach that they are identified as unlikely (see Discussion). Regardless of their lack of agreement with our model, the observation that one TE family has expanded and that one has contracted in *S. cerevisiae* suggests that there may be competition between these intracellular parasites (Leonardo and Nuzhdin 2002).

One of the most interesting gene families identified by our method is the significant expansion of the flocculin family in the ancestor of *S. cerevisiae* and *S. paradoxus*, and the continuing expansion in *S. cerevisiae* (both likelihood ratio tests are significant). These genes are involved in yeast flocculation—the manner in which yeast come together in solutions. Flocculation is one of the most important traits that have been selected for in the domestication of the brewer's yeast, *S. cerevisiae* (Jin and Speers 1998). Domesticated yeast fall to the bottom of tanks once all of the sugars have been consumed in any fermenting brew, avoiding the need for any complicated removal; wild yeast species stay suspended in the liquid or flocculate too early (Jin and Speers 1998). These phenotypes are largely affected by the flocculin gene family.

Table 2. List of the most significant gene families identified as unlikely under the BD model

Family name name	Family sizes in Newick notation	Predicted branch	Likelihood ratio
Stress response	(**15** (33 (24 (30 31))))	1	6e6
Amino acid biosynthesis	(**3** (8 (6 (6 5))))	1	36
PGM/PMM	(**1** (3 (3 (2 1))))	1	9
Ribosomal L1	(1 (**4** (1 (1 1))))	2	3e3
Chaperone	(1 (**4** (2 (2 1))))	2	47
α/β hydrolase	(2 (2 (**6** (2 2))))	3	2e4
Dihydrouridine synthase	(1 (1 (**6** (1 1))))	3	6e5
Trichothecene pump	(5 (5 (7 (**10** 6))))	4	6e3
RNA polymerase Rpb1	(4 (3 (5 (**7** 4))))	4	1e3
Transposon	(2 (8 (15 (34 **83**))))	5	4e54
Helicase	(1 (3 (3 (2 **34**))))	5	1e39
Thiol oxidase	(1 (1 (**4** (**2 3**))))	6	1e3
Leucine rich repeat	(4 (3 (**1** (**2 1**))))	6	38
Flocculation	(10 (6 (8 (**11 14**))))	7	85
Transposon	(17 (14 (15 (**1 5**))))	7	6e10
Myosin	(5 (9 (9 (**5 5**))))	7	76

The first column gives the gene family name; the second column describes the gene family size among the five *Saccharomyces* species in Newick notation. The third column gives the branch that is predicted to be responsible for the overall low *p*-value of the family using the likelihood ratio test, and the fourth column gives the corresponding likelihood ratio. Newick numbers in bold indicate the branch identified.

Discussion

In this study we have presented and evaluated a method for studying the evolution of gene families over a phylogeny. Based on data from multiple whole genomes, the method can be used to examine the rates and direction of change in gene family size among taxa. Our method also allows for hypothesis testing: we have shown how we can identify gene families that have had unlikely histories given a model of random gene birth and death. Importantly, the PGM methodology used here scales linearly with the number of new genomes added; the most challenging aspect of future analyses may simply be getting reliable phylogenetic trees for the species considered. This PGM approach is conceptually similar to the maximum-likelihood approach taken by others to study the evolution of phenotypic quantitative characters (e.g., Mooers and Shluter 1998; Pagel 1999). As with simple Brownian motion approximations of the evolution of phenotypic characters, many mechanistic and biological aspects of gene duplication and loss are not directly addressed by the BD model; nonetheless, we think that our method is an important first step.

Our analyses have revealed a large number of changes in gene family size across the *Saccharomyces* tree: 1254 of 3517 families changed in size. Every branch of the phylogeny was inferred to have changes along it, with longer branches having commensurately more changes (Table 1). One concern we had prior to our analysis was that the uneven sequence coverage of these five genomes would affect our results; this did not appear to be the

case. *S. cerevisiae* is, in fact, the only eukaryotic with a fully sequenced genome—all of the other yeast genomes are covered to differing extents. *S. paradoxus* was sequenced to 7× coverage (i.e., shotgun sequencing was done equivalent to seven times the length of the genome), while *S. bayanus*, *S. kudriavzevii*, and *S. mikatae* were sequenced to 2–3× (Cliften et al. 2003). Despite this unevenness among taxa, our results do not seem to have been affected: *S. kudriavzevii* and *S. mikatae* were predicted to have both the largest number of genes and the largest number of gene family expansions. If the lack of sequence coverage had been a problem, we would have expected these genomes to show fewer genes and smaller gene family sizes on average.

As described above, the null BD model can be used to test whether gene families are, on average, diffusing evenly along the tree. This model can be violated when processes such as natural selection give a direction to the expected random walk, causing extreme expansions or contractions to gene family size. We were able to detect such changes on almost every branch of the tree, and on every external branch leading to an extant species (Table 2). In cases where we did not reject the null hypothesis, it does not mean that natural selection is not acting on members of a gene family, only that we cannot detect its role in affecting the differences in size of the family. Natural selection may have played a role in the fixation of a small number of duplicates within a family, but, much like other statistical tests in molecular evolution, we only have the power to detect the repeated occurrence of events.

One of the most extreme examples that we found was in the helicase family, where *S. cerevisiae* has 34 members of this family, while none of the other species have more than three. While we have no firm biological explanation for this pattern, it is highly possible that in the domestication of yeast, increased rates of cell division—and therefore, of DNA replication—were selected for. We were also able to identify an expansion of the flocculin gene family in *S. cerevisiae* (as well as in its common ancestor with *S. paradoxus*), a change that is unsurprising considering the fact that flocculation has been selected for in the domestication of this brewer's yeast (Jin and Speers 1998). Like other genes that have undergone artificial selection during domestication (e.g., Wang et al. 1999), the flocculin gene family may show the signature of adaptive natural selection. This is the first example to our knowledge, however, of selection on gene family size being implicated in domestication.

Any inference of natural selection with our method comes with a number of caveats that must be mentioned. Small *P*-values do not necessarily imply natural selection, only that the data do not fit the null model. One caveat to our null is that we have implicitly assumed that there is no relationship between family size and duplication and deletion rates. It may be, for instance, that large gene families are more likely to undergo nonhomologous pairing, unequal crossing over, and therefore more duplication and eventual fixation due to drift (Li 1997). A homogeneous birth and death model may also not be absolutely correct for small gene families, as under the BD model families will always eventually reach the absorbing state of zero genes. Because many genes appear to be conserved over very long periods of time (e.g., Theissen et al. 2003), there may be a decreased loss rate in small families in order to prevent extinction of required gene functions. The possibility of nonhomogeneities in very large or very small gene families suggests that models incorporating these processes be studied. Karev et al. (2002) found that a random BD model with added parameters for birth and death rates for the largest and smallest families fit the distribution of gene families in a single genome slightly better than a completely homogeneous model. The improved fit to the data, however, was not shown to be significantly better than models without the two extra parameters. The framework we have provided here should allow for the testing of models that include heterogeneous gain and loss rates across gene families. Although large families are expected to show greater change in number between species simply because there are more chances for gain and loss—and the opposite is true for small families—we will in the future be able to test whether the observed changes are more or less than are expected. This approach will also be able to inform us as to whether families with different physical distributions, such as those arrayed in tandem, show inherently different rates of birth and death than more dispersed families.

The issue of gene families having intrinsically different birth and death rates extends beyond the consideration of family size. For example, one family of genes that does not follow this assumption is transposable elements (TEs): they can multiply in number in a nonmendelian manner, and are often selected against by the organisms they inhabit. Because the parameters for gain and loss of TEs can be quite different from those for other gene families (see, e.g., Li 1997; Kidwell 2002), the disparity in TE number between genomes can be due to processes unique to this family. So our finding that TEs are at the top of our list of unusual gene families is not surprising. Results for transposable element families or other genomic parasites using the BD model, therefore, should not be parameterized with gain and loss rates inferred from the majority of protein-coding genes.

In addition to the assumptions of equivalent birth and death mechanisms among families, one other very important aspect of any random-point process is the assumption of independence among individual genes. The BD model assumes that each gene in a family has an independent probability of being duplicated or deleted; any large-scale chromosomal duplication, deletion, or polyploidization may act on multiple members of a family at once. This is potentially a common violation of the model in light of the frequency of larger scale duplications and deletions that include gene duplicates (Friedman and Hughes 2001). As a result, we cannot compare taxa that are separated by a whole-genome duplication in the same manner as has been presented here. This also means that any unusual gene family should be examined in more detail to determine the nature of the changes in gene family size; obvious duplications of large regions containing multiple members of a family, for example, may moderate conclusions about natural selection.

Our hypothesis-testing framework requires an estimate of λ, the birth and death parameter determining the rate of evolution. In the Supplemental materials, we show how we can estimate the value of λ that makes the entire data set maximally likely (using Expectation Maximization); reassuringly, the resulting value we obtained (0.002 per million years) is very close to the previous estimate of λ found using data from only the birth rate in *S. cerevisiae* (0.004 per million years; Lynch and Conery 2003). We believe that the assumption of equal birth and death rates is consistent with the data: if we compare fully sequenced genomes between relatively closely related species that are not separated by a polyploidization event, we find that they contain similar numbers of genes. Even among species whose time to most recent common ancestor (TMRCA) is many millions of years (Myr), we find similar numbers of genes; among mammals (human, mouse: 25,000 genes; TMRCA = 75 Myr; Abril et al. 2002); nema-

todes (*Caenorhabditis elegans* and *C. briggsae*: 19,000 genes; TMRCA = 100 Myr; Stein et al. 2003); and dipterans (mosquito, fruit fly: 13,500 genes; TMRCA = 260 Myr; Holt et al. 2002). These similarities indicate that there is no general trend toward larger or smaller genomes or gene families between these species. In the future, we hope to extend the model by making it possible to allow λ to vary along branches of a phylogenetic tree so that we can estimate λ independently for each branch, and to allow birth and death rates to differ when significant genome expansions are detected. We can also analyze the data under a range of values for the branch lengths, *t*, as the analyses presented here assume that the estimates are accurate. These refinements may then provide a clearer picture of the evolution of gene family size.

Conclusions

This study has attempted to provide the machinery needed to study gene family evolution among multiple whole genomes. The methodology can be used for parameter estimation, inferences on the direction and magnitude of evolutionary change, and hypothesis testing. As more genome sequences become available, we hope that this framework makes it possible to identify the genetic changes that are responsible for the phenotypic diversity found in nature. Correlated changes between families or with environmental conditions can then tell us about the mechanisms and modes of natural selection (Harvey and Pagel 1991).

Methods

Birth and death model of gene family evolution

Suppose that we have a family of individual genes whose total size (number of genes) at time t is given by the discrete random variable $X(t)$. Then, the probability that the random variable $X(t)$ takes the value c, given that $X(0) = s$ will be denoted by $P(X(t) = c|X(0) = s)$ (see Bailey 1964). Let us assume that every gene in the family is equally capable of either being duplicated (birth) or lost via deletion or pseudogenization (death); here, we include both the processes of origination and fixation within the terms "birth" and "death". The probability of any gene being duplicated (and fixed) in time Δt is $\lambda\Delta t$ or being lost (and fixed) is $\mu\Delta t$. It follows that in a family of size $X(t)$ at time t, the possible transitions are:

- Probability of one gain $\lambda X(t)\ \Delta t\ + o(\Delta t)$.
- Probability of one loss $\mu X(t)\ \Delta t\ + o(\Delta t)$.
- Probability of more than one of these events $o(\Delta t)$.
- Probability of no change $1 = (\lambda + \mu)\ X(t)\Delta t + o(\Delta t)$.

The probability of two such events occurring, $o(\Delta t)$, is negligible for Δt very small. As the size of a gene family grows, the probability of there being a gain or loss also grows. If the gene family contains zero members in a particular lineage, then there is no chance of birth or death, and this is considered an absorbing state; we are therefore only concerned with situations in which the initial number of genes in a family, $X(0) = s$, is nonzero.

If we consider the case where $s \geq 1$ with equal gain and loss rates per gene ($\lambda = \mu$), then the transition probabilities are:

$$P(X(t) = c|X(0) = s) = \sum_{j=0}^{\min(s,c)} \binom{s}{j}\binom{s+c-j-1}{s-1}\alpha^{s+c-2j}(1-2\alpha)^j, \qquad (1)$$

where α is given by $\alpha = \frac{\lambda t}{1+\lambda t}$. Then the stochastic mean and variance for $X(t)$ given $X(0) = s$ are (see Bailey 1964):

$$\text{Mean}(X(t)|X(0) = s) = s,$$
$$\text{Var}(X(t)|X(0) = s) = 2s\lambda t. \qquad (2)$$

Here, we find that the expected size of the gene family is simply equal to the initial number, s. This is because, with equal birth and death rates, the gene family is neither consistently expanding nor contracting, so that the probability of either increasing or decreasing is equivalent.

Calculating likelihoods using probabilistic graphical models

Based on the birth–death (BD) model and the structure of a phylogenetic tree, we can construct a probabilistic graphical model (PGM) that parameterizes the probability distribution over the gene family sizes in the tree. The BD model represents the conditional distributions corresponding to the branches. Of course, only the gene family sizes in the tips of the tree are known, so we are interested in the marginal probability of the leaf (tip) nodes, rather than in the probability of a complete assignment of all nodes in the tree. This can be computed by averaging over all possible assignments of unspecified internal nodes (except for the root node, on which we are conditioning), a process called marginalization in the graphical models literature. Because of the large number of possible internal state assignments, efficient algorithms have been developed in the PGM literature to carry out this calculation; these are known as message passing or sum-product algorithms (see, e.g., Felsenstein 1981; Lauritzen 1996; M.I. Jordan, in prep.). See the Supplemental materials for more details.

Testing hypotheses about gene family evolution

As noted before, the root-node gene family size is not known, so a genuine *P*-value for the observed values of the leaf species cannot be computed, even in principle. However, for each gene family we can compute conditional *P*-values as we call them, conditioned on a specific value for the root family size. Such a conditional *P*-value is computed based on the corresponding conditional likelihood as a test statistic, as the probability that a random gene family with the same root family size (on which it is conditioned) has a smaller conditional likelihood. The conditional *P*-value that is computed based on the true (but unknown) root value is equal to the true *P*-value that we are interested in. Since there is no way to find out which root family size is actually the true one, for each gene family we computed all conditional *P*-values, conditioned on all choices for the root family size from one up to 100, and picked the largest. (The conditional *P*-values always show a single sharp peak around a specific root family size, which was well below 100 for all gene families studied in this study.) This maximal conditional *P*-value is referred to as the supremum *P*-value in the literature (e.g., Demortier 2003), and clearly represents an upper bound on the true *P*-value, which is equal to one of the conditional *P*-values. A fortiori, if the supremum *P*-value is small, the observed gene family sizes are unlikely to be explainable by the BD model. A common concern about the use of the supremum *P*-value is its sensitivity, or how tight an upper bound on the *P*-value it represents (see Berger and Boos 1994; Demortier 2003). In the Supplemental materials, we describe a way to assess this; it turns out that it is very tight in our problem, warranting its use as a genuine *P*-value.

We developed two methods to calculate the conditional *P*-values—an analytic method that calculated them exactly, and a sampling method that was much faster. Briefly, the sampling method generated data under the BD model over the phylogenetic tree, conditioned on a root-node size. For each resulting sample, the conditional likelihood was calculated, and doing this

for many samples gave a null distribution of conditional likelihoods. The observed conditional likelihood of the data was then compared with this null distribution to give a conditional *P*-value. The exact and approximate methods agreed completely when 10,000 samples were taken.

Identifying the unlikely branch

For the gene families that we have identified as unlikely under the BD model (i.e., the ones with a low *P*-value), we further want to identify the branch in the phylogenetic tree that is responsible for this violation. We have two ways of doing this, both of which always agreed with one another on the data used in this study (see Supplemental materials for more details). The first method investigates how much the *P*-value improves after allowing "total freedom" along one of the branches. This is done by recomputing the *P*-value for the gene family after deleting that branch in the PGM. If deleting a specific branch yields a large improvement in *P*-value, this implies that the remainder of the branches did, in fact, follow the BD model, and hence the deleted branch is responsible for the low overall *P*-value.

The second method works by allowing each branch, in turn, to have its own value for λ potentially different from the rest of the tree; this value was found by expectation-maximization. The likelihood of the data under this two-parameter model was then compared with the likelihood under a one-parameter model that was constrained to a single λ for all branches in a likelihood ratio test.

Note that because branches are investigated one at a time, an implicit assumption of these approaches is that only one branch in the phylogenetic tree violates the BD model. Table 2 lists the branch with the largest likelihood ratio that is significant, assuming that a likelihood ratio test of the two models is χ^2 distributed. We refer to the Supplemental material for more details.

Acknowledgments

We thank D. Begun, J. Gillespie, R. Glor, C. Jones, A. Kern, J. Kelly, K. McConway, J. Mezey, L. Moyle, B. O'Meara, M. Rockman, M. Sanderson, N. Takebayashi, and three anonymous reviewers for discussion and comments. T.D.B acknowledges support from the Fund for Scientific Research—Flanders (F.W.O.–Vlaanderen) for a research visit to U.C. Davis, from the IST Programme of the European Community under the PASCAL Network of Excellence (IST-2002-506778), and from the EU project LAVA (IST-2001-34405); C.N. is supported by the National Budget of Vietnam (grant no. 322/QD-TTg); J.E.S. is supported by an NSF predoctoral fellowship; M.W.H. is supported by an NSF Interdisciplinary Informatics postdoctoral fellowship.

References

Abril, J.F., Agarwal, P., Alexandersson, M., Antonarakis, S.E., Baertsch, R., Berry, E., Birney, E., Bork, P., Bray, N., Brent, M.R., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Bailey, N. 1964. *The elements of stochastic processes*. John Wiley & Sons, Inc., New York.

Berger, R.L. and Boos, D.D. 1994. P values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89:** 1012–1016.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.

Copley, R., Goodstadt, L., and Ponting, C. 2003. Eukaryotic domain evolution inferred from genome comparisons. *Curr. Opin. Genet. Dev.* **13:** 623–628.

Darwin, J.H. 1956. The behaviour of an estimator for a simple birth and death process. *Biometrika* **43:** 23–31.

Demortier, L. 2003. Constructing ensembles of pseudo-experiments. In *Proceedings of PHYSTAT2003: Statistical problems in particle physics, astrophysics, and cosmology* (eds. L. Lyons, et al.), pp. 256–260. Stanford Linear Accelerator Center, Stanford, CA.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30:** 1575–1584.

Feller, W. 1968. *An introduction to probability theory and its applications*. John Wiley & Sons, Inc., New York.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17:** 368–376.

Friedman, R. and Hughes, A.L. 2001. Gene duplication and the structure of eukaryotic genomes. *Genome Res.* **11:** 373–381.

Garczarek, L., Hess, W., Hotzendorff, J., Staay, G.V.D., and Partensky, F. 2000. Multiplication of antenna genes as a major adaptation to low light in a marine prokaryote. *Proc. Natl. Acad. Sci.* **97:** 4098–4101.

Gu, X. and Zhang, H. 2004. Genome phylogeny inference based on gene contents. *Mol. Biol. Evol.* **21:** 1401–1408.

Harvey, P.H. and Pagel, M.D. 1991. *The comparative method in evolutionary biology*. Oxford University Press, Oxford, UK.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M.C., Wides, R., et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298:** 129–149.

Huynen, M.A. and van Nimwegen, E. 1998. The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* **15:** 583–589.

Jin, Y.L. and Speers, R.A. 1998. Flocculation of *Saccharomyces cerevisiae*. *Food Res. Int.* **31:** 421–440.

Karev, G.P., Wolf, Y.I., Rzhetsky, A.Y., Berezovskaya, F.S., and Koonin, E.V. 2002. Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evol. Biol.* **2:** 18.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B.W., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Kellis, M., Birren, B.W., and Lander, E.S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428:** 617–624.

Kidwell, M.G. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115:** 49–63.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lauritzen, S.L. 1996. *Graphical models*. Clarendon Press, Oxford, UK.

Lehmann, E.L. 1959. *Testing statistical hypotheses*, chap. 3. John Wiley & Sons, Inc., New York.

Leonardo, T.E. and Nuzhdin, S.V. 2002. Intracellular battlegrounds: Conflict and cooperation between transposable elements. *Genet. Res.* **80:** 155–161.

Lespinet, O., Wolf, Y., Koonin, E., and Aravind, L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* **12:** 1048–1059.

Li, W.-H. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, MA.

Lindsey, J.K. 1996. *Parametric statistical inference*. Oxford University Press, Oxford, UK.

Lutfalla, G., Crollius, H.R., Strange-Thomann, N., Jaillon, O., Mogensen, K., and Monneron, D. 2003. Comparative genomic analysis reveals independent expansion of a lineage-specific gene family in vertebrates: The class II cytokine receptors and their ligands in mammals and fish. *BMC Genomics* **4:** 29.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.

———. 2003. The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics* **3:** 35–44.

Mooers, A.O. and Schluter, D. 1998. Fitting macroevolutionary models to phylogenies: An example using vertebrate body sizes. *Contr. Zool.* **68:** 3–18.

Nei, M., Gu, X., and Sitnikova, T. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl. Acad. Sci.* **94:** 7799–7806.

Oakeshott, J., Claudianos, C., Russell, R., and Robin, G. 1999. Carboxyl/cholinesterases: A case study of the evolution of a successful multigene family. *Bioessays* **21:** 1031–1042.

Pagel, M.D. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* **48:** 612–622.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, San Mateo, CA.

Qian, J., Luscombe, N.M., and Gerstein, M. 2001. Protein family and fold occurrence in genomes: Power-law behaviour and evolutionary model. *J. Mol. Biol.* **313:** 673–681.
Ranson, H., Claudianos, C., Ortelli, F., Abgrall, C., Hemingway, J., Sharakhova, M.V., Unger, M.F., Collins, F.H., and Feyereisen, R. 2002. Evolution of supergene families associated with insecticide resistance. *Science* **298:** 179–181.
Reed, W.J. and Hughes, B.D. 2004. A model explaining the size distribution of gene and protein families. *Math. Biosci.* **189:** 97–102.
Rokas, A., Williams, B.L., King, N., and Carroll, S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425:** 798–804.
Sims, H.J. and McConway, K.J. 2003. Nonstochastic variation of species-level diversification rates within angiosperms. *Evolution* **57:** 460–479.
Slatkin, M. and Rannala, B. 1997. Estimating the age of alleles by use of intrallelic variability. *Am. J. Hum. Genet.* **60:** 447–458.
Snel, B., Bork, P., and Huynen, M. 2002. Genomes in flux: The evolution of archaeal and proteobacterial gene content. *Genome Res.* **12:** 17–25.
Stein, L.D., Bao, Z.R., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N.S., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biology* **1:** 166–192.
Szathmary, E., Jordan, F., and Pal, C. 2001. Can genes explain biological complexity? *Science* **292:** 1315–1316.
Tatusov, R., Koonin, E., and Lipman, D. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.
Theissen, U., Hoffmeister, M., Grieshaber, M., and Martin, W. 2003. Single eubacterial origin of eukaryotic sulfide:quinone oxidoreductase, a mitochondrial enzyme conserved from the early evolution of eukaryotes during anoxic and sulfidic times. *Mol. Biol. Evol.* **20:** 1564–1574.
Van Dongen, S. 2000. *A cluster algorithm for graphs*. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam.
Wang, R.L., Stec, A., Hey, J., Lukens, L., and Doebley, J. 1999. The limits of selection during maize domestication. *Nature* **398:** 236–239.
Yanai, I., Camacho, C.J., and DeLisi, C. 2000. Predictions of gene family distributions in microbial genomes: Evolution by gene duplication and modification. *Phys. Rev. Lett.* **85:** 2641–2644.
Yang, Z. and Bielawski, J. 2000. Statistical methods for detecting molecular evolution. *Trends Ecol. Evol.* **15:** 496–503.
Zhang, H. and Gu, X. 2004. Maximum likelihood for genome phylogeny on gene content. *Stat. Appl. Genet. Mol. Biol.* **3:** Article 31.
Zwickl, D.J. and Holder, M.T. 2004. Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. *Systematic Biology* **53:** 877–888.

Received December 15, 2004; revised version accepted May 17, 2005.