Tạp chí Tin học và Điều khiển học, T.23, S.1 (2007), 1–14

COMPUTATIONAL DETECTION OF NATURAL SELECTION IN GENE FAMILY EXPANSION AND CONTRACTION

CHI NGUYEN[1], NELLO CRISTIANINI [2]

[1] *Department of Computer Science, University of California, Davis, USA*
[2] *Department of Statistics, University of California, Davis, USA*

Abstract. Researchers have generally attributed ostensibly large differences in family size to the effects of natural selection in promoting either the expansion or contraction of families along specific lineages, without a strong statistical basis. Here we use a model of stochastic birth and death for mapping gene family evolution onto a phylogeny, and show that it can be efficiently applied to multi-species comparisons. The model offers the opportunity for researchers to make stronger inferences regarding the role of natural selection and changing duplication and deletion rates in gene family expansion or contraction. The work is performed as a part of the first author's MS thesis under the second author's supervision.

Tóm tắt. Thông thường các nhà nghiên cứu gắn sự khác nhau lớn theo vẻ bề ngoài của kích cỡ gia đình với sự ảnh hưởng của việc lựa chọn tự nhiên trong quá trình thúc đẩy sự mở rộng hay thu hẹp của gia đình trong những thế hệ sau mà không dựa trên một cơ sở thống kê chắc chắn. Trong bài báo này, chúng tôi sử dụng một mô hình thống kê sinh-tử để ánh xạ sự tiến hoá của các họ gien tới một cây phân loài, và chỉ ra rằng mô hình này có thể được áp dụng một cách có hiệu quả cho sự so sánh đa loài. Mô hình đưa ra một cơ hội cho các nhà nghiên cứu thực hiện những suy diễn chính xác hơn liên quan tới vai trò của sự lựa chọn tự nhiên và tốc độ nhân đôi hay loại bỏ trong sự mở rộng và thu hẹp họ gien. Kết quả nghiên cứu này là một phần trong luận văn thạc sỹ của tác giả thứ nhất dưới sự hướng dẫn của tác giả thứ hai.

1. INTRODUCTION

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 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).

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 evolu-

tion—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. Copley et al. 2003; Lander et al. 2001; Lespinet et al. 2002; Oakeshott et al. 1999). While many of these comparisons 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.

We believe that the inability to make statistical inferences about the role of natural selection in the evolution of gene family size is 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, 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 (Bailey 1964; Darwin 1956; Karlin and Taylor 1975), and have also found use in population genetics and phylogenetics (e.g. 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 et al 1998; Lynch and Conery 2000, 2003; Qian et al. 2001). 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). 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 many species. It should be noted that stochastic BD processes are quite different than 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 paper 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 non-randomly. We demonstrate the usefulness of our approach on the whole genomes of five closely related yeast species: S. cerevisiae, S. paradoxus, S. mikatae, S. kudriavzevii, and S. bayanus.

2. BIRTH AND DEATH MODEL OF GENE FAMILY EVOLUTION

Suppose that we have a gene family of individual genes whose total size (number of genes) at time t is given by the discrete random variable $X(t)$, where the probability that $X(t)$ takes the value n is $p_n(t)$ (Bailey 1964). (BD processes are generally described in terms of a population of individuals, but we have merely changed the terminology to be consistent with the biological process under consideration.) 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 due to genetic drift 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:

Chance of one gain $= \lambda X(t)\Delta t + o(\Delta_t)$.

Chance of one loss $= \mu X(t)\Delta t + o(\Delta t)$.

Chance of more than one of these events $= o(\Delta t)$.

Chance of no change $= 1 - (\lambda + \mu)X(t)\Delta t + o(\Delta t)$.

We assume that the probability of two such events occurring, $o(\Delta t)$, is negligible. 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, a, is non-zero ($X(0) = a$).

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

$$p_{n|a}(t) = \sum_{j=0}^{\min(a,n)} \binom{a}{j} \binom{a+n-j-1}{a-1} \alpha^{a+n-2j}(1-2\alpha)^j$$

where α is given by:

$$\alpha = \frac{\lambda t}{1+\lambda t}.$$

The stochastic mean and variance for this case are (Bailey 1964): $m(t) = a$ and $\sigma^2(t) = 2a\lambda t$

Here we find that the expected size of the gene family is simply equal to the initial number, a. 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 on every branch of a tree.

3. LIKELIHOOD CALCULATION OF GENE FAMILY SIZE OVER A PHYLOGENY BY USING PROBABILISTIC GRAPHICAL MODELS

In order to draw statistically motivated conclusions from gene family size data in several related species, we use a probabilistic graphical model (Jordan, in preparation) that represents the probability distribution over the observed gene family data. The graphical models machinery allows us to efficiently compute the likelihood of the given observed data, conditioned on the family size of their common ancestor. In addition, the computation of this conditional

likelihood allows us to calculate the most likely value of ancestral states and eventually a p-value associated with the observed data (see next section).

We now discuss a method to identify gene families that are evolving under selective pressure, the first main algorithmic result of this paper. We assume that for a given set of species we are given the phylogeny along with the branch lengths and the evolution rate parameter λ. The given species correspond to the leaf nodes in the phylogenetic tree; the common ancestors are represented by the internal nodes. Furthermore, the size of a large number of gene families is known in each of the given species. This section will describe how to compute the likelihood of that assignment efficiently, given the BD model and the phylogeny.

More correctly, we will not compute the likelihood of the gene family sizes of the leaf node species, but their likelihood conditioned on gene family size in the root species. The reason is that no probability density for the root species' size is known, nor do we want to impose one artificially: we do not want to assess the likelihood of gene families based on their average size (which is largely determined by the oldest ancestor), but rather on the divergence of this size over the phylogenetic tree. Thus, instead of computing one likelihood, for each family we can compute a conditional likelihood, conditioned on some given root node family size.

We will later use this conditional likelihood to perform a hypothesis test where we estimate the p-value, the probability that an equally (or less) likely gene family evolution arises by chance. Gene families whose size distribution across the taxa cannot be explained by the null model will have a low p-value, and can be assumed to be under some kind of selective pressure.

A graphical model based on the phylogenetic tree

Assuming the BD model as the null model for evolution, the size of a gene family in a given species depends only on the size of the same gene family in the node directly above it in the phylogenetic tree: it is specified fully by the conditional probability $p_{n|a}(t)$ to observe a family size of n given the ancestor has size a. Therefore, given the gene family size of the common ancestor of all taxa (i.e. the root node species), it is possible to compute the probability of the evolution below it as the multiplication of the conditional probabilities of the family size of each of the descendants of the root, given the family size of their direct parent.

However, we are generally not interested in the likelihood of a *complete* phylogenetic tree given the root species' size. We rather want to know the likelihood to observe the gene family sizes for the given taxa corresponding to the *leaf nodes* only, conditioned on the root species' size. This can be computed by averaging over all possible assignments of internal nodes (except for the root node), a process called *marginalization* in the graphical models literature.

While the marginalization step involves averaging over a very large set of internal node assignments, namely exponential in the size of the tree, it is a nontrivial result of the theory of Probabilistic Graphical Models that these computations can be performed in a very efficient way, by resorting to algorithms referred to *as message passing* or *sum-product* [Jordan, 2004]. The complexity of these algorithms is only linear in the size of the phylogenetic tree.

For a practical implementation of the algorithm, we need to make the assumption that the maximal gene family size is limited. However, since the conditional probability distribution associated with the BD model drops of quickly for large values, this assumption is very reasonable for a large enough upper limit. Apart from an efficient method to perform marginalizations, PGMs make it possible to compute the most likely assignment of the unspecified

internal nodes.The algorithm is a variant of the one used to compute the conditional likelihood and is known in the graphical models literature as the max-product algorithm. We do not go into the details here but refer the reader to the relevant literature (see e.g. Jordan, in preparation). We can thus use this method to efficiently compute the most likely gene family sizes at all internal nodes of our tree.

4. INFERRING λ

Thus far we assumed that the evolution parameter λ and the family sizes of the root node species were given. Alternatively we can learn it from the data using Expectation Maximazation (EM). Specifically, we equate λ to that value that maximizes the conditional log likelihood of the complete set of gene families, which is the sum of the conditional log likelihoods of the individual gene families. For each gene families, the conditional log likelihood is conditioned on the root species family size that yields the largest value. As can be seen in Figure 1, the optimal value for λ is 0.002.

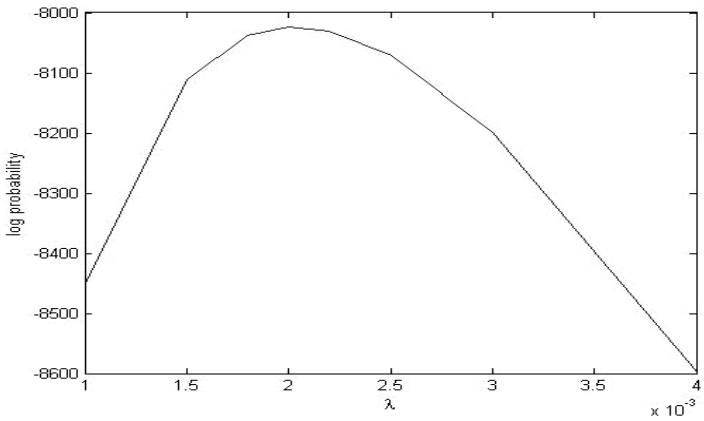

Figure 1. The log likelihood of all gene families as a function of the parameter

5. TESTING HYPOTHESES ABOUT GENE FAMILY EVOLUTIONS

In the previous section we described how, given the family size in the common ancestor, the likelihood for the family sizes of the given species can be computed. Of course in practice we do not know the actual value of the root gene family size. However, we could make the conservative choice to assign that value to it that leads to the largest conditional likelihood.

Still, this is not sufficient to return an interpretable result: a larger root family size will undesirably yield consistently lower likelihoods, since the conditional probability distribution of a child node's family size is more spread out for a larger parent family size (remember that the square of the variance is proportional to the parent family size: $\sigma^2(t) = 2a\lambda t$). For example, for the phylogeny described in the experiments below, the assignment of the family sizes of the leaf node species equal to (2(2(2(22)))) (in the Newick notation) has a maximal conditional likelihood of 0.384 (when the root's family size is equal to the value 2), while the maximal conditional likelihood for the leaf node gene family sizes (20(20(20(2020)))) is

0.002 (when the root's family size is equal to the value 20) , even though both represent a most likely assignment under the BD model, respectively starting from a root family size of 2 and 20. Therefore, in order to obtain an interpretable result, we need to compute p-values corresponding to these the likelihoods. These p-values indicate what the probability is to observe this particular likelihood, given the size of the root species family size.

6. *P*-VALUES AND CONDITIONAL *P*-VALUES FOR GENE FAMILY EVOLUTIONS

Assume for now that the size of the gene family at the root species is given. Then, we can compute the likelihood of the given family sizes of the leaf nodes as explained above. Subsequently, as we will show shortly below, it is possible to estimate the probability to observe the same or a smaller likelihood, given this root family size. If the root family size is given indeed, this is the p-value we are looking for. When this p-value is small this indicates that, given the root family size, it is very improbably to observe taxa with gene family sizes as unlikely as the assignment we are assessing, thus providing evidence that the evolution can not be explained by the chance from the null model.

In practice the root gene family size is not given. Therefore, we compute a so-called conditional p-value for all possible root family sizes, one of which, corresponding to the true root value, is the true p-value for the gene family evolution. Since there is no way to find out which root family size is the true one, we make the conservative choice to pick that value that leads to the largest conditional p-value. As such, an upper bound on the true p-value is obtained. A fortiori, if this value is small, the assignment of the leaf node taxa is unlikely to be a explainable by the null model.

Unfortunately, the conditional p-value can not be computed analytically in a reasonable computation time. Therefore, we use the graphical model defined by the BD model and the phylogenetic tree structure to randomly sample gene family evolutions starting from the given root family size. This can be done efficiently thanks to the tree structure of this graphical model. Subsequently, for each of these random samples, the likelihood given the root family size can be computed. Based on the likelihoods of a large number of random samples (1000 for each root family size, in this paper), the conditional p-value can be reliably and efficiently estimated by counting the proportion of these random samples that have a likelihood lower than or equal to the one for the given gene family.

To assess the tightness of this upper bound, we can use the general fact that the p-values of random samples sampled from the null model are uniformly distributed [Link]. Indeed, the probability to observe a random sample with p-value lower than a given p-value is equal to the probability to observe a sample that is less likely than the given sample, which is the p-value again.

7. IDENTIFYING THE UNLIKELY BRANCH

For the improbable gene family evolutions (i.e. the ones with a low upper bound on the p-value) we further want to identify the branch in the phylogenetic tree that is responsible for the low p-value. There are two ways of doing this.

The first way is by computing an upper bound for the p-value corresponding to the likelihood of the pair of subtrees as obtained by removing a branch in the phylogenetic tree, and this once for each branch. If after any of these branch removals the upper bound on the p-value is larger than a threshold level, this branch may be held responsible for the low p-value of the complete phylogenetic model. Again, the upper bound on the p-value can be computed as the maximal conditional p-value, now conditioned on two root values - one for each subtree - instead of on one. As above, the conditional p-values can be estimated by random sampling. Note that, since one branch is removed at a time, an implicit assumption of this approach is that only one branch in the phylogenetic tree violates the BD model. The second approach makes use of the most likely assignment of the gene family sizes for the species associated with the internal nodes of the phylogenetic tree. This most likely assignment can be computed efficiently with a technique from the probabilistic graphical models literature known as the max-product algorithm, related to the Viterbi algorithm. After the most likely assignment for the internal nodes is computed, one can compute a p-value for the transitions on each of the branches. A low p-value for a certain branch indicates that the BD model is likely to be violated at that branch.

Whereas the first approach is more rigorous when at most one branch violates the BD model (a hypothesis test is carried out for every partial model as obtained by removing one and only one branch), the second approach is computationally faster, and more suitable when several branches do not follow the BD model. In principle, the first approach could be extended however, by removing more than one branch at a time. However, for an increasing number of branch removals this quickly becomes computationally intractable.

8. EXPERIMENTAL RESULTS

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 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 1 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 1 as time, t, in million years. We estimated the evolutionary rate parameter λ as 0:002 per million years .

In the 32 million years since the most recent common ancestor of the five species, 1254 of the 3517 gene families shared among them has 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. As expected from the birth and death process, there was a high correlation between the length

of each branch in millions of years and the number of changes occurring along that branch (Spearman's 1/2=0.82). 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;). 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.

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 less than 0.01 (35 are expected). The unlikely families are summarized in Table 2, 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 most cases (see Table 2).

For the first four families identified in Table 2 the observed gene family sizes are so unlikely that it is hard to determine where any one unlikely event occurred. Two of these gene families are of unknown function, and the other two are 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).

9. DISCUSSION

In this paper 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. Pagel 1999).

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 7X 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 2-3X (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. In cases where we 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 taxation 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 3. We were also able to identify a significant expansion of the flocculin gene family in S. cerevisiae, 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), we detected the signature of adaptive natural selection on the flocculins. 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. One caveat 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 non-homologous pairing, unequal crossing over, and therefore more duplication and eventual taxation 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 non-homogeneities 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.

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 non-mendelian manner, and are often selected against by the organisms they inhabit. Because the parameters for gain and loss of TEs can be quite difierent than those for other gene families (see, e.g.Kidwell 2002; Li 1997), 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 above sections we show how we can estimate the value of that makes the entire dataset 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 S. cerevisiae (0.004 per million years; Lynch and Conery 2003). In the future we hope to extend the model by making it possible to allow to vary along branches of a phylogenetic tree or by allowing the birth and death rates to be unequal on any branch. 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.

Table 1. The number of gene families that showed an expansion, no change, or a contraction along the 8 branches, according to the most likely assignments of the gene family sizes of the ancestors. The first column contains the branch number, along with the length of the branch, t, in millions of years. The last column shows the average gene family expansion among all families along each branch, where a contraction is counted as a negative expansion.

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

Table 2 shows the gene families identified as unlikely under the BD model. 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; two numbers are provided, the first one from the branch deletion method (method 1), the second one from the transition probabilities along each branch (method 2). In most cases both methods give the same answer. Newick numbers in bold indicate the branch identified by method 1. The fourth column gives the resulting p-value after deleting the responsible branch as identified by method 1, and the last column gives the p-value of the least likely branch transition as computed in method 2. Note that for the first four gene families neither method was able to identify one single branch that violates the BD model, and only method 2 was able to identify a branch for the fifth and sixth families listed. The four gene families that were missed by the approximate sampling method are marked with an asterix in the first column.

Table 2

Family name	Family sizes in Newick notation	Pred. branch	Method 1	Method 2
Transposon	(2 (8 (15 (34 83))))	?/?	<0.01	
Unknown	(7 (16 (7 (20 17))))	?/?	<0.01	
Transposon	(17 (14 (15 (1 5))))	?/?	<0.01	
Unknown	(5 (11 (14 (4 2))))	?/?	<0.01	
Stress response	(15 (33 (24 (30 31))))	?/1	<0.01	0.000
Flocculation	(10 (6 (8 (11 14))))	?/2	<0.01	0.002
Amino acid biosynthesis	(3 (8 (6 (6 5))))	1/1	0.137	0.001
*PGM/PMM	(1 (3 (3 (2 1))))	1/3	0.045	0.007
*Ribosomal L1	(1 (4 (1 (1 1))))	2/2	0.661	0.000
Elongation factor	(1 (4 (2 (1 1))))	2/2	0.197	0.003
Chaperone	(1 (4 (2 (2 1))))	2/2	0.112	0.003
Phosphatidylinositol 4-kinase	(2 (9 (4 (2 2))))	2/2	0.064	0.000

Carbamoyl-phosphate synthase	(2 (6 (5 (3 3))))	2/1	0.048	0.003
Alpha/beta hydrolase	(2 (2 (6 (2 2))))	3/3	0.777	0.000
Dihydrouridine synthase	(1 (1 (6 (1 1))))	3/3	0.657	0.000
Type I phosphodiesterase	(1 (1 (4 (1 1))))	3/3	0.657	0.000
Guanine nucleotide exchange factor	(2 (2 (5 (2 3))))	3/3	0.243	0.006
DNA binding domain	(2 (2 (5 (2 1))))	3/3	0.199	0.000
Ankyrin repeat	(1 (2 (7 (1 1))))	3/3	0.195	0.000
-Unknown -Unknown	(1 (2 (4 (1 1))))	3/3	0.195	0.002
Acetate transporter	(2 (4 (5 (2 2))))	3/3	0.118	0.006
*TruD	(1 (1 (3 (1 2))))	3/3	0.115	0.000
*Unknown	(1 (1 (3 (2 1))))	3/3	0.115	0.000
Flavodoxin	(2 (3 (5 (1 1))))	3/7	0.110	0.000
Swi2/Snf2 ATPase	(17 (20 (25 (18 15))))	3/3	0.061	0.000
GTPase-activating protein	(2 (4 (6 (3 2))))	3/1	0.047	0.004
Maltose transport	(4 (7 (8 (5 4))))	3/1	0.043	0.010
Trichothecene pump	(5 (5 (7 (10 6))))	4/4	0.331	0.000
RNA polymerase Rpb1	(4 (3 (5 (7 4))))	4/4	0.252	0.000
ATPase	(1 (1 (2 (3 1))))	4/4	0.122	0.000
MAL transcription factor	(2 (5 (4 (7 4))))	4/4	0.086	0.000
Hydroxymethylpyrimidine synthesis	(3 (5 (2 (7 4))))	4/4	0.015	0.000
Ribosomal protein (60S)	(2 (1 (1 (1 3))))	5/5	0.305	0.000
eIF4E-associated protein	(1 (2 (1 (1 3))))	5/5	0.228	0.000
Hydrolase	(8 (11 (12 (11 7))))	5/5	0.161	0.000
Metal-dependent phosphohydrolases	(1 (1 (2 (1 5))))	5/5	0.122	0.000
Sortilin	(5 (4 (7 (4 8))))	5/5	0.045	0.000
Helicase	(1 (3 (3 (2 34))))	5/5	0.038	0.000
NAD kinase	(3 (1 (1 (2 4))))	5/5	0.038	0.001
Hydroxyisocaproate dehydrogenases	(3 (1 (2 (1 3))))	5/5	0.038	0.002
ABC transporter	(15 (18 (17 (12 8))))	5/5	0.013	0.000
Thiol oxidase	(1 (1 (4 (2 3))))	6/3	0.212	0.002
Leucine rich repeat	(4 (3 (1 (2 1))))	6/1	0.076	0.027
HSP70 Chaperone	(13 (17 (18 (12 13))))	7/7	0.141	0.006
-Transcription factor -PolIII transcription factor-Cytoplasmic protein that binds Tor2p-Ribosomal SSU (40S) -Adenylate cyclase activity, G-protein signaling -RRM1	(1 (3 (3 (1 1))))	7/3	0.124	0.007
Myosin	(5 (9 (9 (5 5))))	7/7	0.068	0.001
Cation transport enzymes	(8 (10 (13 (6 5))))	7/7	0.048	0.000
S-methyltransferase	(2 (5 (5 (1 1))))	7/7	0.037	0.000
-PDRE transcription factor-Component of peripheral vacuolar membrane protein complex	(1 (4 (4 (1 1))))	7/3	0.024	0.002
1,3-beta-D-glucan synthase	(3 (8 (7 (3 3))))	7/7	0.015	0.000

10. CONCLUSION

This paper has attempted to provide the model 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.

REFERENCES

[1] N. T. J. Bailey , *The elements of stochastic processes* John Wiley & Sons, Inc., New York. 1964.

[2] R. R. Copley, L. Goodstadt, and C. Ponting, Eukaryotic domain evolution inferred from genome comparisons, *Current Opinion in Genetics & Development* **13** (2003) 623–628.

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

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

[5] N. R. Friedma, and A. L. Hughes, Gene duplication and the structure of eukaryotic genomes, *Genome Research* **11** (2001) 373–381.

[6] M. A. Huynen, and E. Van Nimwegen, The frequency distribution of gene family sizes in complete genomes, *Molecular Biology and Evolution* **15** (1998) 583–589.

[7] Y. L. Jin, and R. A. Speers, Flocculation of Saccharomyces cerevisiae, *Food Res. Int.* **31**(1998) 421–440.

[8] I. M. Jordan, Graphical models (To appear: Statistical Science 2004 (Special issue on Bayes Statistics).

[9] S. Karlin, and H. M. Taylor, *A first course in stochastic processes*, Academic Press, New York. 1975.

[10] G. P. Karev, Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezovskaya, and E. V. Koonin, Birth and death of protein domains: A simple model of evolution explains power law behavior, *BMC Evolutionary Biology* **2** (2) (2002).

[11] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. Lander, Sequencing and comparison of yeast species to identify genes and regulatory elements, *Nature* (423) (2003) 241–254.

[12] M. G. Kidwell, Transposable elements and the evolution of genome size in eukaryotes, *Genetica* (115) (2002) 49–63.

[13] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody et al., Initial sequencing and analysis of the human genome, *Nature* (409) (2001) 860–921.

[14] O. Lespinet, Y. I. Wolf, E. V. Koonin, and L. Aravind, The role of lineage-specific gene family expansion in the evolution of eukaryotes, *Genome Research* **12** (2002) 1048–1059.

[15] W. H. Li, *Molecular evolution*, Sinauer Associates, Sunderland, Mass. 1997.

[16] M. Lynch, and J. S. Conery, The evolutionary fate and consequences of duplicate genes, *Science* (290) (2000) 1151–1155.

[17] M. Lynch, and J. S. Conery, The evolutionary demography of duplicate genes, *Journal of Structural and Functional Genomics* **3** (2003) 35–44.

[18] M. Nei, X. Gu, and T. Sitnikova. Evolution by the birth-and-death process in multigene families of the vertebrate immune system, *PNAS* **94** (1997) 7799–7806.

[19] J. G. Oakeshott, C. Claudianos, R. J. Russell, and G. C. Robin, Carboxyl/cholinesterases: a case study of the evolution of a successful multigene family, *BioEssays* **21** (1999) 1031–1042.

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

[21] J. Qian, N. M. Luscombe, and M. Gerstein, Protein family and fold occurrence in genomes: Power-law behaviour and evolutionary model, *Journal of Molecular Biology* (313) (2001) 673–681.

[22] W. J. Reed, and B. D. Hughes, A model explaining the size distribution of gene and protein families, *Mathematical Biosciences* **189** (2004) 97–102.

[23] A. Rokas, B. L. Williams, N. King, and S. B. Carroll, Genome-scale approaches to resolving incongruence in molecular phylogenies, *Nature* (425) (2003) 798–804.

[24] H. J. Sims, and K. J. Mcconway, Nonstochastic variation of species-level diversification rates within angiosperms, *Evolution* **57** (2003) 460–479.

[25] B. Snel, P. Bork, and M. A. Huynen, Genomes in flux: The evolution of archaeal and proteobacterial gene content, *Genome Research* **12** (2002) 17–25.

[26] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, A genomic perspective on protein families, *Science* (278) (1997) 631–637.

[27] U. Theissen, M. Hoffmeister, M. Grieshaber, and W. Martin, Single eubacterial origin of eukaryotic sulfide:quinone oxidoreductase, a mitochondrial enzyme conserved from the early evolution of eukaryotes during anoxic and sulfidic times, *Molecular Biology and Evolution* **20** (2003) 1564–1574.

[28] R. L. Wang, A. Stec, J. Hey, L. Lukens, and J. Doebley, The limits of selection during maize domestication, *Nature* (398) (1999) 236–239.

[29] Z. H. Yang, and J. P. Bielawski, Statistical methods for detecting molecular evolution, *Trends in Ecology and Evolution* **15** (2000) 496–503.

[30] Link http://business.clayton.edu/arjomand/business/p-value.htm

Received on March 7 - 2005

Revised on October 15 - 2006