ELSEVIER

Gene 300 (2002) 43–51

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences

Christopher J. Creevey[a], James O. McInerney[a,b,*]

[a]*Bioinformatics and Pharmacogenomics Laboratory, Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland*
[b]*Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK*

Received 18 December 2001; received in revised form 28 June 2002; accepted 18 September 2002

Abstract

Positive selection or adaptive evolution is thought to be responsible, at least some of the time, for the rapid accumulation of advantageous changes in protein-coding genes. The origin of new enzymatic functions, erection of barriers to heterospecific fertilization, and evasion of host response by pathogens, among other things, are thought to be instances of adaptive evolution. Detecting positive selection in protein-coding genes is fraught with difficulties. Saturation for sequence change, codon usage bias, ephemeral selection events and differential selective pressures on amino acids all contribute to the problem. A number of solutions have been proposed with varying degrees of success, however they suffer from limitations of not being accurate enough or being prohibitively computationally intensive. We have developed a character-based method of identifying lineages that undergo positive selection. In our method we assess the possibility that for each internal branch of a phylogenetic tree an event occurred that subsequently gave rise to a greater number of replacement substitutions than might be expected. We classify these replacement substitutions into two categories – whether they subsequently became invariable or changed again in at least one descendent lineage. The former situation indicates that the new character state is under strong selection to preserve its new identity (directional selection), while the latter situation indicates that there is a persistent pressure to change identity (non-directional selection). The method is fast and accurate, easy to implement, sensitive to short-lived selection events and robust with respect to sampling density and proportion of sites under the influence of positive selection.

Keywords: Adaptive evolution; Relative rate test; Method

1. Introduction

Positive Darwinian selection, or adaptive evolution, is a process that encourages the retention of mutations that are beneficial to an individual or population. It may be an ephemeral event (though not always), often leading to the creation of new enzymatic properties in a protein (Stewart et al., 1987; Irwin and Wilson, 1990; Irwin, 1995), alteration of the characteristics of the cell surface (Holmes et al., 1992; Smith et al., 1995), or even assistance in the creation of new species (Lee et al., 1995). Adaptive evolution might be expected in situations where an organism finds itself in a new environment or if its environment is undergoing a significant change.

Adaptive evolution can either be directional or non-directional. Directional selection describes situations where successive amino acid sequence change makes a protein more efficient at performing a particular task. Natural selection approves these modifications and they are preserved in future lineages. Non-directional selection describes a selective regime where the environment is constantly altering and where sequence change is required just to maintain an acceptable level of fitness. In both of these situations, there is selection for replacement substitutions, however the interpretation of these events is quite

Abbreviations: d_n, number of non-synonymous substitutions per non-synonymous site; d_s, number of synonymous substitutions per synonymous site; DNA, deoxyribonucleic acid; MHC, major histocompatibility complex; RI, replacement invariable substitutions; RV, replacement variable substitutions; SI, silent invariable substitutions; SV, silent variable substitutions.

* Corresponding author. Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland. Tel.: +353-1-7083860; fax: +353-1-7083845.
E-mail address: james.o.mcinerney@may.ie (J.O. McInerney).

0141-933/02/$ - see front matter
PII: S0378-1119(02)01039-9

different and it is important to be able to discriminate between them.

Positive selection at the molecular level is most often studied by comparing the number of non-synonymous nucleotide substitutions per non-synonymous site (d_n) to the number of synonymous nucleotide substitutions per synonymous site (d_s) (Li, 1993; Muse and Gaut, 1994; Ina, 1995; Endo et al., 1996; Yang and Nielsen, 2000). The majority of protein-coding regions are thought to be under the influence of negative selection. This is because most protein sequences are well-adapted to carrying out their functions and change would not lead to the creation of a selective advantage (Hughes, 1999). While purifying selection is at work, we expect non-synonymous substitutions to occur very rarely and the value of d_n to be less than that of d_s (Kimura, 1980; Hughes, 1999). It has been assumed that the converse of this situation is indicative of positive selection acting on a protein (Kimura, 1980; Li, 1993). However, if positive selection only acts on a small number of sites (change at other sites being selectively neutral or disadvantageous), then the d_n value is unlikely to ever become larger than the d_s value (Hughes, 1999). In addition, the accuracy of d_s is greatly reduced if there has been substantial sequence change (sequence saturation) since the two last shared a common ancestor. The problem of adequately accounting for superimposed substitutions also affects d_n values although, arguably, not to the same extent (Kimura, 1980; Li, 1993; Hughes, 1999). Finally, pairwise distance methods cannot distinguish between directional and non-directional selection.

The shortcomings of distance based methods (Zhang et al., 1997; Crandall et al., 1999), the extensive computation time and recently uncovered unreliability (Suzuki and Nei, 2001) of likelihood-based methods makes it desirable to find alternative ways of detecting adaptive evolution (Nei and Gojobori, 1986; Nielsen and Yang, 1998; Suzuki and Gojobori, 1999). Here we report a fast and accurate method of detecting adaptive evolution and show that the results of our analyses are more compatible with known biochemical evidence than previous analyses in at least one instance.

In order to assess the usefulness of our method we analyzed the primate lysozyme data-set of Messier and Stewart (1997). They identified positive selection using a method of reconstructing ancestral character states and performing a series of pairwise distance analyses between all combinations of ancestral and contemporary sequences. In doing so, they identified two lineages in which positive selection appeared to be occurring – the lineage leading to colobine monkeys and the lineage leading to hominoids. In a subsequent analysis, Yang (1998) analyzed the same data using a maximum likelihood framework. He concluded that there was evidence for positive selection in the lineage leading to hominoids. The lineage leading to colobines had evidence of a higher replacement to silent substitution rate than the background rate but this rate was not significantly greater than 1. The biochemical evidence for this data-set indicates that the lysozyme in the fore-gut fermenting colobines has acquired new functions (activity at low pH, resistance to cleavage by pepsin) and is the most likely candidate for having undergone an episode of adaptive evolution.

To demonstrate the ability of our method to differentiate between different types of selective pressures, we also present the results of two other data-sets, one under non-directional positive selection, and the other under negative selection.

The MHC is a multigene family whose products are cell-surface glycoproteins that play a key role in the immune system by presenting peptides to T cells (Hughes, 1999). It is also known to be subjected to positive non-directional selection (Hughes and Nei, 1988; Hughes, 1999; Swanson et al., 2001). The data-set used here has been used by previous researchers (Swanson et al., 2001) as a positive control for the purposes of identifying positive selection, and is used here for the purpose of identifying positive non-directional selection.

The third data-set is a group of Carbonic anhydrase I sequences. This has been identified as a housekeeping gene, and there is no evidence of positive selection (Swanson et al., 2001). It will be used here for the purposes of identifying a negative selection result.

2. Materials and methods

2.1. Relative rate ratio test

Our method is a relative rate ratio analysis and proceeds as follows. For any data-set, a phylogenetic tree is constructed and assumed to be correct. This tree is rooted by reference to an appropriate out-group. This effectively converts the data to polarized character types. Hypothetical ancestral sequences are reconstructed at each internal branch through a method that uses the principal of maximum parsimony (Hennig, 1966) applied at the codon level. However, since ambiguities are uninformative, we construct for each data-set substitution matrices based on the types of changes we observe in the data-set at zero-, two- and four-fold degenerate positions. For any ambiguity between two (or more) codons we can then calculate how often we would expect to see each codon, based on the summed changes we observe in the candidates at each zero-, two- and four-fold degenerate site present in the codons. This method only works reliably when there is no ambiguity at the root, however maximum parsimony cannot guarantee this, so in this case the ancestral codon is assigned the same as the ancestor of the out-group. This ensures that there are no ambiguities in the ancestral reconstruction. However, in principle any method that accurately reconstructs ancestral character states may be used.

Using our reconstructed phylogeny, we identify all substitutions that occur across the tree and determine

whether they result in a non-synonymous or synonymous codon change. We then consider each internal branch of the tree and count the changes in the descendent clade described by that internal branch. This results in four values representing those different types of substitutions that have occurred from that internal branch to the tips. The four types of substitutions are classified as replacement-invariable (RI, those replacement substitutions where the new character-state is preserved in all subsequent lineages), replacement-variable (RV, where the replacement substitution is not preserved in all lineages and has changed at least once more in a subsequent lineage), silent-invariable (SI, silent changes that were not observed to have changed again) and silent-variable (SV, silent changes that were observed to have changed again in a subsequent lineage). Using a *G*-test (or Fisher's exact test (Sokal and Rohlf, 1981), when numbers are small) we can compare the ratio of RI substitutions to RV substitutions, with the expectation that this ratio is the same as the ratio of SI to SV substitutions (McDonald and Kreitman, 1991). The ratio of SI to SV is the expected value under the neutral model, since they represent neutrally evolving sites (Kimura, 1983), and the ratio of RI to RV substitutions would be the same as SI to SV under neutrality (McDonald and Kreitman, 1991). In the event of a significant difference in these ratios, it is possible to analyze the result to find out whether there are high numbers of RI substitutions or RV substitutions. The former is indicative of directional selection and the latter is indicative of non-directional selection. Positive selection favours any substitutions that confer an advantage. During an episode of positive directional selection, these advantageous substitutions will occur and subsequently remain invariable in all descendant lineages, at a rate significantly higher than expected from the neutral model (McDonald and Kreitman, 1991). Under non-directional selection the advantageous substitutions will become variable at a rate significantly higher than expected from the neutral model.

2.2. Neutral substitution rate test

Neutral mutations acting on any random genetic sequence would result in about three times more replacement substitutions than silent substitutions. However, due to factors like base composition bias, the transition to transversion ratio and codon usage for any particular data-set, the ratio of replacement to silent changes can differ. We can calculate the expected ratio of replacement to silent changes of any data-set though, by basing them on the total number of replacement or silent sites. The total number of replacement or silent sites in each sequence within each clade was calculated as per the method of Li (1993). This ratio was then used to calculate the expected number of replacement and silent substitutions within each clade. We consider whether the number of replacement substitutions or silent substitutions observed were significantly greater than expected from the neutral model. Using a *G*-test (or Fisher's exact test (Sokal and Rohlf, 1981)), we can determine if any type of substitution occurred more often than expected from neutrality. With this information if there is no significant result using the first test, then we can determine if negative selection was acting or whether the substitutions were appearing at a rate expected under neutrality.

2.3. Simulation of neutrality

Since RI, RV, SI and SV numbers are calculated at each internal branch, this means that we are performing multiple statistical tests on the same data (within any lineage). This increases the probability of obtaining a significant result by chance alone (Roff and Bentzen, 1989). To counter this problem, we used a Monte Carlo technique (Rambaut and Grassly, 1997) to generate 100,000 pseudo data-sets based on each of the three case studies. For each case study, the codon position rate-heterogeneity, base composition bias and the transition/transversion ratio were estimated using maximum likelihood in PAUP* (Swofford, 2002). These variables along with the phylogeny, sequence length and a random coding sequence of correct length were used to generate data-sets in Seq-Gen (Rambaut and Grassly, 1997), which would mimic neutral evolution acting on the data-set. For each of the three case studies, the 100,000 pseudo data-sets were analyzed using the method described in this paper. This generated the expected distribution of *G* (from the *G*-test) at each internal branch if the null hypothesis of neutral evolution were true for the particular case study (Roff and Bentzen, 1989). The number of times that any internal branch showed a significant deviation from expectation represented how often we would expect to have a type 1 error. This distribution was then used to assess the critical level for every internal branch of the particular case study.

3. Results

3.1. Primate lysozyme

We present the results of three data-sets of known evolutionary background to demonstrate the ability of our method in identifying the type of evolutionary pressure acting on a particular locus.

The lysozyme locus appears to have undergone a period of adaptive evolution within primates. A pairwise distance analysis of the sequences using the method of Li (1993) indicates that 198 d_n/d_s ratios are greater than 1, while only 78 of the comparisons have a value less than 1.

In our relative rate ratio analysis of the primate lysozyme sequences, only one lineage shows evidence of a significantly higher number of RI substitutions. This is the lineage leading to the colobine monkeys. Analysis of the lineage leading to the hominoids does not indicate that positive selection is occurring.

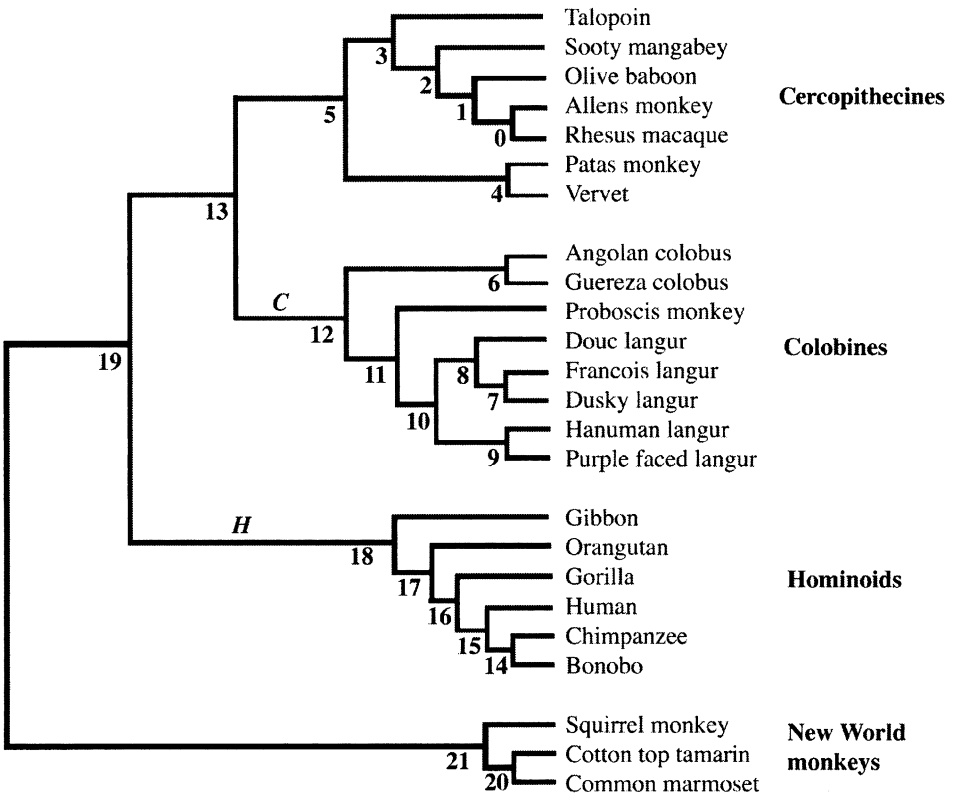

Fig. 1. Phylogenetic tree derived from primate lysozyme genes. The numbers on the internal branches correlate with the numbers for each internal branch in Table 1. The internal branches labelled '*C*' and '*H*' follow the nomenclature of Yang (1998).

Fig. 1 shows the phylogenetic tree obtained from these data. This tree is identical to the one produced by Messier and Stewart (1997) and Yang (1998). The notation used by Yang is also used in this paper with the lineage denoted by '*c*' representing the internal branch leading to the colobines and the internal branch denoted '*h*' representing the internal branch leading to the hominoids. For all internal branches in the tree we count all the substitutions in the descendent clades. This means that the numbers get bigger as we move closer to the root, however, small numbers are adequately dealt with using Fisher's exact test and larger numbers are analyzed using a *G*-test for independence (Sokal and Rohlf, 1981). The internal branches are labelled in Fig. 1 and the numbering scheme is preserved in Table 1. Table 1 indicates the numbers of substitutions that were counted, starting at each internal branch. Branches 11, 12, 13 and 19 showed a significant difference in the ratios of RI/RV versus SI/SV. In this case, these ratios were skewed due to the excessive number of RI events. In the neutral simulation study, of these internal branches, only internal branch 13 resulted in a number of type 1 errors above the critical level of 5%. When the critical level was readjusted to bring the number of type 1 errors to less than 5% the new critical level was 0.0428. Since the *P* value calculated at this internal branch was less than 0.005, the result at this branch was still deemed significant.

Of note is the fact that these four internal branches are connected and they describe a lineage encompassing the common ancestor of Colobines and Cercopithecines. Starting from internal branch 11, we find that subsequently there were six RI sites, one RV site, two SI sites and five SV sites. There are six times as many RI sites than RV sites, however the ratio of SI to SV is seen to be in the opposite direction with over twice as many SV sites. Using the *G*-test, the *P* value is less than 0.05, however using Fisher's exact test, the *P* value is only 0.051282 and is therefore not deemed significant. The results from branch 12 (16 RI, four SI, two RV, five SV) show strong evidence of more RI sites than expected (Fisher's $P = 0.02876$). We would expect to have seen fewer than four RI sites, and instead we see 16.

When the number of replacement or silent substitutions at each internal branch was compared to the expected value under the neutral model, branch 21 (observed replacements = 6, observed silents = 10; expected replacements = 12.39, expected silents = 3.61) was the only one with a significant result (Fisher's $P = 0.02609$). Every other internal branch of the tree showed no significant difference in the number of replacement or silent substitutions than was expected from neutrality (at the $P = 0.05$ level).

Table 1
Primate lysozyme results

#	RI	RV	SI	SV			*P* value
	OR		**OS**		**ER**	**ES**	***P* value**
0	0	3	0	0			
	3		**0**		**2.3**	**0.7**	
1	0	3	0	1			
	3		**1**		**3.1**	**0.9**	
2	1	3	0	1			
	4		**1**		**3.9**	**1.1**	
3	2	4	0	2			
	6		**2**		**6.3**	**1.8**	
4	0	0	0	3			
	0		**3**		**2.3**	**0.7**	
5	5	4	1	5			
	9		**6**		**11.8**	**3.2**	
6	3	0	1	0			
	3		**1**		**3.1**	**0.9**	
7	1	0	0	0			
	1		**0**		**0.8**	**0.2**	
8	1	1	0	3			
	2		**3**		**3.9**	**1.1**	
9	1	0	1	0			
	1		**1**		**1.6**	**0.4**	
10	3	1	1	3			
	4		**4**		**6.3**	**1.7**	
11	6	1	2	5			
	7		**7**		**11.0**	**3.0**	
12	16	3	4	5			*
	19		**9**		**22.0**	**6.0**	
13	24	7	5	10			***
	31		**15**		**36.1**	**9.9**	
14	0	0	0	0			
	0		**0**		**0.0**	**0.0**	
15	1	1	0	3			
	2		**3**		**3.8**	**1.2**	
16	3	1	0	3			
	4		**3**		**5.4**	**1.6**	
17	3	5	0	5			
	8		**5**		**10.1**	**2.9**	
18	11	9	2	5			
	20		**7**		**21.0**	**6.0**	
19	43	20	11	18			***
	63		**29**		**72.1**	**19.9**	
20	1	5	1	4			
	6		**5**		**8.6**	**2.4**	
21	1	5	1	9			
	6		**10**		**12.4**	**3.6**	*

#, internal branch number (see Fig. 1); RI, replacement invariable; RV, replacement variable; SI, silent invariable; SV, silent variable; OR, observed replacement substitutions; OS, observed silent substitutions; ER, expected replacement substitutions; ES, expected silent substitutions. *P* value results: $*P < 0.05$; $**P < 0.01$; $***P < 0.005$.

3.2. Class I MHC glycoproteins

The second protein alignment contains the class I MHC glycoprotein, which is known to be subjected to positive non-directional selection (Hughes and Nei, 1988; Hughes, 1999). This was used by Swanson et al. (2001) as a positive control, and the same data-set which is all Human MHC is used here, but with an added sequence from *Pan troglodytes* (GenBank Accession number: X13116) used as an outgroup. MHC is known to be subject to recombination, which would invalidate the analysis, so this data-set was analyzed using PIST (Worobey, 2001) to test for recombination events in the alignment. The results showed no evidence of recombination. Fig. 2 shows the phylogenetic tree obtained from these data. The numbers given to the internal branches in Fig. 2 are preserved in Table 2, which shows the RI, RV,

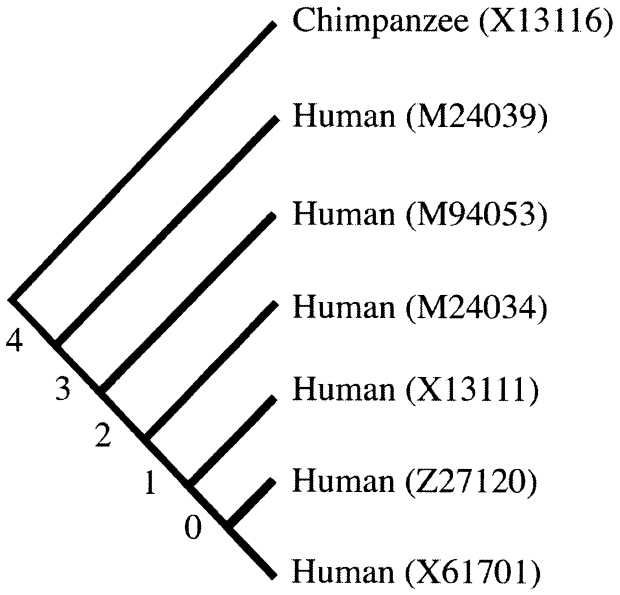

Fig. 2. Phylogenetic tree derived from the class I MHC glycoprotein genes. The numbers on the internal branches correlate with the numbers for each internal branch in Table 2. The accession number of each sequence is contained in parentheses beside the name of its corresponding species.

Table 2
Class I MHC glycoprotein results

#	RI	RV	SI	SV			*P* value
	OR		**OS**		**ER**	**ES**	***P* value**
0	6	28	4	12			
	34		***16***		***37.8***	***12.2***	
1	41	47	34	20			
	88		***54***		***108.2***	***33.8***	***
2	45	62	35	25			*
	107		***60***		***126.3***	***40.7***	*
3	46	66	35	27			
	112		***62***		***131***	***43***	*
4	51	90	38	34			*
	141		***72***		***159.9***	***53.1***	*

#, internal branch number (see Fig. 2); RI, replacement invariable; RV, replacement variable; SI, silent invariable; SV, silent variable; OR, observed replacement substitutions; OS, observed silent substitutions; ER, expected replacement substitutions; ES, expected silent substitutions. *P* value results: $*P < 0.05$; $**P < 0.01$; $***P < 0.005$.

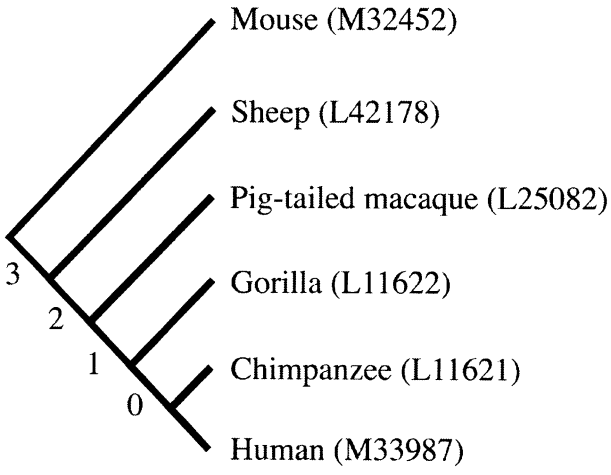

Fig. 3. Phylogenetic tree derived from the Carbonic anhydrase I genes. The numbers on the internal branches correlate with the numbers for each internal branch in Table 3. The accession number of each sequence is contained in parentheses beside the name of its corresponding species.

SI and SV values at each branch along with the appropriate *P* value. A pairwise distance analysis of the sequences using the method of Li (1993) indicates that only four d_n/d_s ratios are greater than 1, while 17 have a value less than 1.

In our analysis while all of the internal branches showed an increased number of RV substitutions, only two (internal branch numbers 2 and 4) were significant at the $P = 0.05$ level. Of these two internal branches only internal branch number 2 resulted in an elevated number of type 1 errors in the neutral simulation study. When the distribution of *G* for this internal branch was taken into account a new critical value of 0.0482 was set. The *P* value calculated for this internal branch was less than the new critical level at 0.0441, so the result at this internal branch was still deemed significant. Analysis of the number of replacement and silent substitutions that occurred within each of the clades showed that four of the clades (defined by internal branch numbers 1, 2, 3, and 4) had significantly higher numbers of silent substitutions (at the $P = 0.05$ level) than was expected under the neutral model. The only clade without a significant result (defined by internal branch number 0) also had an increased number of silent substitutions (see Table 2).

Table 3
Carbonic anhydrase I results

#	RI	RV	SI	SV			*P* value
	OR		**OS**		**ER**	**ES**	***P* value**
0	1	2	1	4			
	3		***5***		***6.2***	***1.8***	
1	8	7	6	6			
	15		***12***		***20.8***	***6.2***	
2	30	15	23	13			
	45		***36***		***62.2***	***18.8***	***
3	57	97	63	71			
	154		***134***		***220.8***	***67.2***	***

#, internal branch number (see Fig. 3); RI, replacement invariable; RV, replacement variable; SI, silent invariable; SV, silent variable; OR, observed replacement substitutions; OS, observed silent substitutions; ER, expected replacement substitutions; ES, expected silent substitutions. *P* value results: $*P < 0.05$; $**P < 0.01$; $***P < 0.005$.

3.3. Carbonic anhydrase I

Carbonic anhydrase I was used by Swanson et al. (2001) as a negative control (representing negative selection) for their study. We use the same sequences here to demonstrate the ability of our method to detect negative selection. A pairwise distance analysis of the sequences using the method of Li (1993) indicates that all d_n/d_s ratios are less than 1. Fig. 3 shows the phylogenetic tree obtained from these data. The numbers given to the internal branches in Fig. 3 are preserved in Table 3.

In our analysis none of the internal branches showed a significant deviation (at the $P = 0.05$ level) in the number of RI or RV substitutions from what was expected under neutrality. Examination of the number of replacement and silent substitutions that occurred within each clade showed that two clades (defined by internal branch numbers 2 and 3) had a significantly greater number of silent substitutions (at the $P = 0.05$ level) than was expected under neutrality. The other two clades (defined by internal branch numbers 0 and 1) had an increased number of silent substitutions (see Table 3) but were not significant at the $P = 0.05$ level.

4. Discussion

4.1. Relative rate ratio test

Our method uses ancestral sequence reconstruction to identify changes that occur within a clade. We analyze all these changes, with the assumption that change in the protein sequence is an ongoing process, with a protein that changes function being poorly adapted in the beginning, but subsequent replacement substitutions making it a better effector of its new role. These replacement substitutions are then preserved in subsequent lineages.

Silent sites are usually thought to be under weak selective pressures. Ignoring translational selection for optimal codons, there are no recognized selective pressures that might influence silent changes in a gene. Therefore, mutational events are thought to play the biggest role in silent site substitutions. A silent site in a phylogenetic tree will be observed to be invariable within a clade, either by random chance or because the clade is shallow and there has been little time to allow a subsequent change. Silent sites within a sequence will be observed to be variable within a clade in a way that is proportional to clade depth. Therefore, calibrating the rate of replacement substitutions by reference to silent substitutions gives an indication of how much

change is not due to the relative depth of the clade combined with the mutational process within the clade. In other words, the ratio of RI to RV sites should be the same as the ratio of SI to SV sites, and any deviation from this is an indication of the presence of some kind of selective pressure.

Positive selection occurs when a character emerges that is so beneficial to an organism that the difference in fitness between novelty and wild-type means that descendants with the novelty have a much greater chance of passing their genes onto their offspring. The individuals with this character flourish, often to the detriment of the others in the population (Hughes, 1999). This character state is rapidly accepted into the population and spends little time as a polymorphism (Hennig, 1966). Since non-synonymous substitutions are the only kind of substitution that will affect the phenotype of an individual, we can trace all these substitutions and identify those that were subsequently fixed within descendant populations.

McDonald and Kreitman (1991) used this idea to detect positive selection at the *Adh* locus in *Drosophila*. Their sampling of multiple individuals from three different species of *Drosophila* implied an interspecies star phylogeny and within each of the three species they found multiple fixations occurring. However, since they did not reconstruct a phylogeny with ancestral sequences, they assumed that the most common genotype was ancestral. This may not always be a valid assumption and represents a situation where our method has an increased discriminating ability. Templeton (1996) expanded on McDonald and Kreitman's approach by dividing the substitutions into further categories, this followed a suggestion by Hudson (1993) that it would be possible to use the relative rate test to compare levels of divergence within and between species for different kinds of nucleotide sites. This increased the discriminating power of the test, however it still could not identify when in the timeframe represented by the data-set the selection for change occurred. By constructing a phylogenetic tree for a data-set and reconstructing the ancestral sequences for this tree we do not assume a star phylogeny, we can infer the ancestral genotypes (Hennig, 1966) and can accurately tell whether or not substitutions were non-synonymous or synonymous. Also, by summing the changes that have occurred since the existence of any hypothetical ancestor, we get an indication of which part of the tree is under positive selection and at which stage of the phylogeny the selection pressure was likely to have changed. This is by comparison to McDonald and Kreitman (and similarly for Templeton) who could only show that within the whole data-set positive selection was acting somewhere (McDonald and Kreitman, 1991).

By counting the substitutions that have occurred anywhere within any clade defined by an internal branch we are determining whether or not that population produced significantly more invariable or variable replacement substitutions since a given time point than we would normally expect under the neutral model. We can test this null hypothesis with appropriate independence tests (Sokal and Rohlf, 1981; McDonald and Kreitman, 1991) or by use of a simulation study to determine the distribution of neutrally evolving data-sets.

One of the main criticisms of the pairwise distances method of detecting positive selection is the failure to account for saturation at the third position of synonymous sites (Hughes, 1999). Synonymous substitutions at the same site will occur independently of whether the substitutions are invariable or not (McDonald and Kreitman, 1991). This means that regardless of how many substitutions have actually occurred we can base our ratios on those synonymous substitutions we observe. Pairwise distances may also show that positive directional selection is occurring when in fact those non-synonymous changes are variable within a population as in the case of the MHC molecule (Hughes and Nei, 1988; Hughes, 1999). A non-synonymous change can occur, but this is not necessarily an adaptive event. If the replacement remains invariable then this is potentially a positive selection event. By basing our results on a phylogenetic tree we can identify when replacements that remain invariable occur, overcoming this problem.

4.2. Neutral simulation study

Although the cumulative distribution of G from the 100,000 pseudo data-sets did not always match the tabulated distribution, it did not lead to an incorrect statistical decision in any of the three case studies used in this paper if the tabulated values were used. The problem of multiple statistical tests increasing the likelihood of obtaining a significant result by chance alone needs to be carefully examined by researchers using this method, however, using a Monte Carlo simulation study as was done here circumvents this problem.

4.3. Primate lysozyme and directional selection

In the primate lysozyme data-set, we identified three internal branches with significantly ($P < 0.05$) increased numbers of replacement invariable substitutions than were expected under neutrality. These internal branches identify the lineage that led to the colobines, the fore-gut fermenters in primates. This result is in agreement with results of other researchers (Messier and Stewart, 1997; Zhang et al., 1997; Yang, 1998). Unlike these researchers, however, we do not identify positive selection in the lineage leading to the hominids. But what we did identify was that all of the Old World Monkeys (Hominoids, Colobines and Cercopithecines) were evolving at the same rate expected under the neutral model. The only exception was branch 21 leading to the New World Monkeys, which was under negative selection. This may explain why using pairwise distances it is possible to identify d_n/d_s ratios greater than 1, even though positive selection was not acting. The results for the

lysozyme data-set show that the replacement substitutions, which may be deleterious, or selectively neutral, appear in the Old World Monkeys at the same rate as expected under the neutral model. As such, the neutral theory (Kimura, 1983) (or the nearly neutral theory (Ohta, 1992)) best describes the evolution occurring at the lysozyme locus in Old World Monkeys. This does not exclude the possibility of positive selection as an evolutionary event, rather it provides a mechanism for evolution in the absence of strong selection (Kimura, 1983). We cannot always trust a positive result using d_n/d_s ratios especially if the locus is evolving at the neutral rate, however we can identify positive selection if we examine what happens after the substitution occurs (RI or RV and SI or SV).

4.4. Class I MHC glycoproteins and non-directional selection

In the class I MHC glycoprotein data-set we successfully identified that non-directional selection was the driving force behind its evolution. Two of our internal branches (numbers 4 and 2) showed significance at the $P = 0.05$ level. These significance levels were due to the increased number of RV substitutions than was expected under neutrality. The more basal branch (number 4) identifies the clade that exclusively represents the human sequences. Unlike with the lysozyme data-set, we do not believe that this is a residual effect from a branch closer to the crown of the tree (branch 2) because there is an internal branch separating the two that has no significance at the $P = 0.05$ level. This leads us to believe that within humans non-directional selection is acting at the MHC locus.

In the examination of the numbers of replacement and silent substitutions within each clade, we see a significantly higher number of silent substitutions than was expected from under the neutral model in four out of the five clades. This agrees with the analysis using the method of Li (1993) which resulted in the majority of the d_n/d_s ratios being less than 1. This is a classic example of where traditional pairwise comparisons may identify negative selection, when in fact positive selection is also occurring. In these situations negative selection acts at most of the sites, but a few adaptive sites contribute to the evolution of the protein. Our analysis successfully identifies the presence of both selective forces.

4.5. Carbonic anhydrase I and negative selection

Finally in the Carbonic anhydrase I data-set we identify negative (purifying) selection. This is in agreement with the findings of other researchers (Swanson et al., 2001). None of the internal branches showed a significant (at the $P = 0.05$ level) increase in the numbers of RI or RV substitutions.

When the numbers of total replacement and silent substitutions within each clade were calculated, two of the branches (numbers 2 and 3) had a significantly higher number of silent substitutions than expected under the neutral model. This is in agreement with the results of the method of Li (1993), which showed that every d_n/d_s ratio was less than 1. The two internal branches that showed no significance had an increased number of silent substitutions, but the calculated P values were greater than 0.05. This is a recurring result, which identifies problems with the power of our statistical tests when small numbers are involved. Further investigation into suitable statistical tests may be required.

We do recognize the areas in which our method may fail. If the sequences are not sufficiently similar, then reconstructing ancestral sequences may be error prone. Additionally, there are problems with the accurate reconstruction of ancestral character states when the contemporary sequences manifest base compositional biases (Cunningham et al., 1998). A full likelihood method of reconstruction (Yang et al., 1995) might be superior and further work may introduce this. Nonetheless, we believe that our reconstructions were ‘not very wrong’, and empirical analysis showed that those sequences that were reconstructed grouped correctly with their descendants (data not shown).

4.6. Concluding remarks

Since the ratios are summed from the tip of the tree to any internal branch at which we are looking, if one internal branch had significantly differing ratios, then a more basal internal branch could show significance also even though it is simply a residual effect of our counting method. In such a case, the internal branch closest to the tip is the most likely to be the one in which positive selection began to occur. We are being careful to outline this potential difficulty with interpreting the results, although if a lineage is identified using our method, then it is almost certain that positive selection occurred on that lineage. Whether or not the internal branch closest to the tips is the one where positive selection began to occur is difficult to say and it is unlikely that a statistical test will resolve this issue.

We assume with this test that selection cannot be strong enough to act on silent sites. The silent sites are our molecular clock against which the replacement substitutions are compared. However, recently researchers (Hurst and Pal, 2001) appear to have found evidence that in some cases there might be purifying selection acting on silent sites. The mechanism or reason has not yet been identified. The question for our method is ‘could selection at silent sites affect the ratio of SI to SV substitutions?’ Since the selection identified thus far is purifying selection, then there would be no extra pressure for either SI or SV substitutions to occur. So we would expect to see a reduction in the number of silent substitutions but no difference in the ratio of SI to SV substitutions.

We have found our method to be reasonably robust to sparse or dense sampling (data not shown). Since we take any internal branch and look at changes all the way to the

tips – not just the change from node to node – sampling becomes less important for the accurate identification of positive selection.

In summary, we have produced a method of detecting positive selection that is superior to other methods, both in terms of speed and accuracy. We can identify internal lineages where positive selection occurred and we can identify positive selection events that occurred over short periods of time.

Acknowledgements

Software to implement this method is available from the authors. We would like to thank Eddie Holmes, Dave McL. Roberts, Richard H. Thomas, Caro-Beth Stewart and Mark Wilkinson for helpful comments on an earlier draft of this manuscript. This project was funded by the Health Research Board of Ireland (Grant number RP 124/2000).

References

Crandall, K.A., Kelsey, C.R., Imamichi, H., Lane, H.C., Salzman, N.P., 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol. Biol. Evol. 16, 372–382.

Cunningham, C.W., Omland, K.E., Oakley, T.H., 1998. Reconstructing ancestral character states: a critical reappraisal. Trends Ecol. Evol. 13, 361–366.

Endo, T., Ikeo, K., Gojobori, T., 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. 13, 685–690.

Hennig, W., 1966. Phylogenetic Systematics, University of Illinois Press, Urbana, IL.

Holmes, E.C., Zhang, L.Q., Simmonds, P., Ludlam, C.A., Brown, A.J.L., 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human-immunodeficiency-virus type-1 within a single infected patient. Proc. Natl. Acad. Sci. USA 89, 4835–4839.

Hudson, R.R., 1993. Levels of DNA polymorphism and divergence yield important insights into evolutionary processes. Proc. Natl. Acad. Sci. USA 90, 7425–7426.

Hughes, A.L., 1999. Adaptive Evolution of Genes and Genomes, Oxford University Press, Oxford.

Hughes, A.L., Nei, M., 1988. Pattern of nucleotide substitution at MHC class I loci reveals overdominant selection. Nature 335, 167–170.

Hurst, L.D., Pal, C., 2001. Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet. 17, 62–65.

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

Irwin, D.M., 1995. Evolution of the bovine lysozyme gene family: changes in gene expression and reversion of function. J. Mol. Evol. 41, 299–312.

Irwin, D.M., Wilson, A.C., 1990. Concerted evolution of ruminant stomach lysozymes. Characterization of lysozyme cDNA clones from sheep and deer. J. Biol. Chem. 265, 4944–4952.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution, Cambridge University Press, Cambridge.

Lee, Y.H., Ota, T., Vacquier, V.D., 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol. Biol. Evol. 12, 231–238.

Li, W.H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36, 96–99.

McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, 652–654.

Messier, W., Stewart, C.B., 1997. Episodic adaptive evolution of primate lysozymes. Nature 385, 151–154.

Muse, S.V., Gaut, B.S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11, 715–724.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Nielsen, R., Yang, Z.H., 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148, 929–936.

Ohta, T., 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23, 263–286.

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.

Roff, D.A., Bentzen, P., 1989. The statistical analysis of mitochondrial DNA polymorphisms: Chi-square and the problem of small samples. Mol. Biol. Evol. 6, 539–545.

Smith, N.H., Smith, J.M., Spratt, B.G., 1995. Sequence evolution of the porB gene of Neisseria gonorrhoeae and Neisseria meningitidis: evidence of positive Darwinian selection. Mol. Biol. Evol. 12, 363–370.

Sokal, R.R., Rohlf, F.J., 1981. Biometry, Freeman, San Francisco, CA.

Stewart, C.B., Schilling, J.W., Wilson, A.C., 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330, 401–404.

Suzuki, Y., Gojobori, T., 1999. A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 16, 1315–1328.

Suzuki, Y., Nei, M., 2001. Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 18, 2179–2185.

Swanson, W.J., Zhang, Z.H., Wolfner, M.F., Aquadro, C.F., 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. USA 98, 2509–2514.

Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4, Sinauer Associates, Sunderland, MA.

Templeton, A.R., 1996. Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. Genetics 144, 1263–1270.

Worobey, M., 2001. A novel approach to detecting and measuring recombination: New insights into evolution in viruses, bacteria, and mitochondria. Mol. Biol. Evol. 18, 1425–1434.

Yang, Z.H., 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15, 568–573.

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

Yang, Z.H., Kumar, S., Nei, M., 1995. A new method of inference of ancestral nucleotide and amino-acid-sequences. Genetics 141, 1641–1650.

Zhang, J.Z., Kumar, S., Nei, M., 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. Mol. Biol. Evol. 14, 1335–1338.