Selection Operating on Protein-coding Genes in the Human Genome

Diogo Meyer, *Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, Brasil*

Eugene E Harris, *City University of New York, New York, USA and Profesor Vistante, Universidade de São Paulo, Brasil*

Advanced article

Article Contents

Online posting date: 14th March 2008

The dominant selective force acting on protein-coding regions throughout the human genome is purifying selection, which removes deleterious mutations. The fraction of substitutions that were positively selected, and can therefore be considered to be adaptive, can be estimated using methods based on comparisons of the relative amounts of change at two classes of sites – sites which if changed produce amino acid changes and sites which if changed do not lead to amino acid changes. Current estimates find very low estimates of adaptive evolution in protein-coding regions during human evolution.

Introduction

The pioneering work of King and Wilson (1975) revealed that the proteins of humans and chimpanzees are remarkably similar, with an identity of approximately 99%. Their study raised the fundamental question as to what are the genetic bases of the phenotypic differences between these two species, and endorsed the perspective that the apparently large differences in morphology and behaviour may be understood in the light of an overall small genetic differentiation if we assume that the genetic changes involved regulatory regions. However, a proper understanding of the role played by the amino acid differences between these species was still lacking.

The entire genome sequences for humans and chimpanzees are now available, as are databases containing a description of polymorphism (between-individual deoxyribonucleic acid (DNA) differences) in humans at millions of nucleotide positions scattered throughout the genome. The divergence between humans and chimpanzees at the nucleotide level is very similar to that predicted by King and Wilson, and implies that these species differ at approximately 60 000 amino acid sites. The population genetic theory developed over the last decades offers a framework by which we can interpret the nature of the mutations and modes of selection that took place as these species diverged. Mutations can be placed in three main categories: (a) those that are selected (either positively or negatively); (b) those that are neutral (i.e. have no effect on fitness) and (c) those that have low selection coefficients, and thus behave as neutral in small populations (where the effects of drift dominate) or are selected in large populations (where the deterministic effects of selection prevail).

By comparing the genomes of humans and chimpanzees, and by analysing polymorphism within humans, we can address some of the fundamental questions of molecular evolution and molecular anthropology. First, we can quantify the proportion of mutations that belong to each of these three categories, and can infer what proportion of all protein-coding changes that have accumulated between these species are the result of positive selection and of genetic drift. Second, we can identify specific genes that have experienced positive selection since humans and chimpanzees diverged, and propose hypotheses about what environmental factors account for such selection. Third, we can compare the rates of molecular evolution along human and chimpanzee lineages in the light of their different demographic histories, since it is known that differences in effective population size will influence the balance of evolutionary forces acting on genetic variation.

Methods for Detecting Selection

At the heart of most methods for testing and quantifying selection in protein-coding regions is the notion that nucleotide changes can be classified as synonymous (those that do not alter the amino acid at that specific codon) or nonsynonymous (those that do alter an amino acid at a codon). It is generally assumed that synonymous changes

ELS subject area: Evolution and Diversity of Life

How to cite:
Meyer, Diogo; and, Harris, Eugene E (March 2008) Selection Operating on Protein-coding Genes in the Human Genome. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0020791

are neutral and that nonsynonymous changes, because they lead to changes in the protein, can be selected.

McDonald and Kreitman (1991) provided a framework for the analysis of coding sequence evolution through the comparison of nonsynonymous and synonymous differences among species (D_n and D_s) with the number of nonsynonymous and synonymous polymorphisms (P_n and P_s) within species. Under neutrality we expect the D_n/D_s will be equal to P_n/P_s, whereas positive selection is expected to increase D_n/D_s relative to polymorphism. Although early applications of this test were directed at individual genes, it is now possible to apply it to datasets on a genomic scale. If we assume that an excess of nonsynonymous divergence is the result of increased fixation of advantageous alleles under positive selection, we can quantify the fraction of positively selected substitutions by comparing how much greater D_n/D_s is than P_n/P_s (Fay *et al*., 2001).

The number of synonymous and nonsynonymous differences between the DNA sequences of two species can also be used to estimate *rates* of nonsynonymous and synonymous substitution per site (d_n and d_s, respectively). Purifying selection against deleterious mutations is expected to result in $d_n < d_s$, whereas sustained positive selection is expected to increase the nonsynonymous substitution rate, resulting in $d_n > d_s$. Under a regime where all mutations have equal fitness effects, regardless of whether they are nonsynonymous or synonymous, we expect $d_n = d_s$. Thus, d_n/d_s ratios provide information about the type of selection that has acted upon a locus.

Purifying Selection

One of the fundamental predictions of the neutral theory (Kimura, 1968) is that a large proportion of the changes that alter protein sequences is deleterious and is removed from populations by purifying selection (or negative selection). Kimura also predicted that proteins would differ in their substitution rates as a function of their degrees of functional constraint. For example, mutations that change proteins in ways that result in a large fitness decrease are quickly removed from populations, whereas those that alter proteins in a way that does not have such strong selective consequences have a greater probability of persisting and becoming fixed (reaching a frequency of 100%), thus resulting in higher substitution rates. Both these expectations have been confirmed by the analysis of data involving diverse species (Kimura, 1983).

Humans and chimpanzees are identical in as many as 29% of proteins, with the majority of proteins (71%) differing by as few as one or two amino acids (Mikkelsen *et al*., 2005). Such high similarity is strong evidence for the action of purifying selection, and the degree to which this force has operated in maintaining this similarity can be quantified by comparing the rates of nonsynonymous and synonymous substitutions. For the human lineage, we typically find that d_n is between 20 and 25% of the d_s (Mikkelsen *et al*., 2005; Bakewell *et al*., 2007). Such a result indicates that purifying selection plays an important role in conserving protein function, with between 75 and 80% of amino acid altering mutations being removed by purifying selection along the human lineage. This finding raises the challenge of explaining the microevolutionary processes that account for the roughly 20–25% of protein-coding differences between these species that were not removed by purifying selection. For instance, what proportion of these differences was positively selected because the differences helped the species adapt to their different environments? Conversely, how many of these differences were neutral, having had little or no impact on fitness, and were fixed by random genetic drift?

Genomic Rates of Positive Selection

Using the McDonald–Kreitman approach, described earlier, estimates of the adaptive rate of genomic evolution have been made for several pairs of closely related species. In comparisons between humans and chimpanzees, based on large datasets, the ratio of D_n/D_s was not greater than P_n/P_s, as would be expected if positive selection were common. In fact, estimates of the fraction of sites evolving under positive selection were not significantly different from zero in two studies (see accompanying table **Table 1**; Zhang and Li, 2005; Mikkelsen *et al*., 2005). In addition, genome-wide studies comparing humans and chimpanzees using d_n/d_s tests found less than 1% of loci indicating positive selection, again suggesting that positive selection on protein-coding genes has been relatively rare in the history of our lineage (Nielsen *et al*., 2005). However, some studies did estimate higher rates of adaptive evolution. An important early study, based on a relatively small set of genes, used the McDonald–Kreitman method and estimated that as many as 35% of amino acid substitutions were adaptive (Fay *et al*., 2001). There are reasons, however, to believe that this may be an overestimate. Since data were limited at the time, polymorphism data derived from a set of candidate disease genes was compared with divergence data derived from a different set of genes. Eyre-Walker (2006) has suggested that if the polymorphism genes were more constrained than the divergence genes, then this would create artifactual evidence for adaptive evolution. In a study by Bustamante *et al*. (2005), using an approach that estimated the selective coefficient of a large set of re-sequenced loci, it was estimated that approximately 6% of loci were positively selected. Gojobori *et al*. (2007) used another approach designed to avoid the effects of possible biases in polymorphism datasets, since such datasets may be overrepresented for nonsynonymous polymorphism that could possibly lead to underestimates of adaptive evolution, and estimated that approximately 10–13% of amino acid substitutions between humans and chimpanzees may be adaptive. Regardless of the specific amount of adaptive evolution, it is apparent that the majority of amino acid substitutions between humans and chimpanzees are the result of neutral mutations that drifted to fixation.

Table 1 Estimates of selective constraint (purifying selection) and adaptive evolution during human evolution based on genome-wide studies of protein-coding genes

Study	Total genes	Data used in analysis	Species compared with human	Adaptive evolution (%)	Purifying selection[a] (%)
Clark *et al*. (2003)	7645	Divergence	Chimpanzee Mouse	0.08[b]	—
Arbiza *et al*. (2006)	9674	Divergence	Chimpanzee Mouse/rat	1.12[b] (5.96)[i]	80[c]
Mikkelsen *et al*. (2005)	13 454	Divergence	Chimpanzee Mouse	~0.0[d]	79[c]
Nielsen *et al*. (2005)	8079	Divergence	Chimpanzee	0.4[b]	—
Gojobori *et al*. (2007)	5008[e] & 5535[f]	Divergence & polymorphism	Chimpanzee	10.4[d] and 12.8[d]	—
Fay *et al*. (2001)	182	Divergence & polymorphism	Old World monkeys	35[d]	80[g]
Bakewell *et al*. (2007)	13 888	Divergence	Chimpanzee Rhesus Macaque	1.1[b] (1.7)[i]	74[c] (75.5)[i]
Zhang and Li (2005)	479	Divergence & polymorphism	Chimpanzee Old World monkeys Mouse	~0.0[d]	—
Bustamante *et al*. (2005)	4916	Divergence & polymorphism	Chimpanzee	~6.0[b]	13.5[h]

[a]Percentage of amino acid altering mutations that were removed.
[b]Percentage of genes that show evidence of positive selection.
[c]Percentage calculated after subtracting the ratio d_n/d_s from 1.0.
[d]These values are the percentage of amino acid substitutions over all genes that show evidence of positive selection estimated using an McDonald–Kreitman approach applied over large sets of genes.
[e]Based on the SNP dataset from Perlegen Biosciences (http://www.perlegen.com) used to obtain the first estimate of adaptive evolution in column five.
[f]Based on the SNP dataset from the International HapMap Project (http://www.hapmap.org) used to obtain the second estimate of adaptive evolution in column five.
[g]This value is estimated on the basis of human polymorphism data by subtracting the fraction of neutral amino acid polymorphism (assumed to be represented by the ratio of nonsynonymous to synonymous mutations within a common frequency class ($>20\%$)) from 100% (all possible amino acid mutations).
[h]This value is based on the percentage of genes showing evidence of negative (or purifying) selection.
[i]This value is for the chimpanzee.

Interestingly, different patterns are emerging from analyses of other species. Contrasts between species of *Drosophila* have suggested that more than 40% of amino acid replacements are adaptive, and analyses of viruses and bacteria yield even higher values of adaptive evolution (Eyre-Walker, 2006).

Even if positive selection in protein-coding genes is relatively rare, it is of interest to determine which genes were positively selected since humans and chimpanzees diverged from their common ancestor. The McDonald–Kreitman test and tests that scan the genome for loci with $d_n/d_s > 1$ have been widely used to address this question. These analyses have generated lists of genes that are candidates for positive selection, and have found functional categories that are enriched for these genes, such as immunity and pathogen-resistance, reproduction (gametogenesis and fertilization), sensory reception (olfactory and auditory), apoptosis and nucleotide metabolism and repair. Methods for detecting positive selection have also been applied to the data on polymorphism for humans (Sabeti *et al*., 2006). Such tests are based on patterns of linkage disequilibrium, levels and patterns of polymorphism and population differentiation. However, since these methods rely on features of polymorphism, they can only detect positive selection that

acted since the origin of modern *Homo sapiens* or since the time human populations began to differentiate.

Slightly Deleterious Mutations

In its original formulation, the neutral theory assumed that deleterious mutations had a sufficiently strong effect that they were immediately removed from the population by natural selection, and thus did not contribute to polymorphism or divergence. However, mutations with small selection coefficients with respect to the population size (i.e. those for which $2N_e s$ is close to 1, where N_e denotes effective population size and s denotes the selection coefficient) may have their changes in frequency determined largely by genetic drift rather than natural selection. With respect to humans, we would like to know how large this class of 'nearly neutral' mutations is.

An extension of the neutral theory was developed by Ohta (1973). In this formulation, it was proposed that a large class of mutations are 'nearly neutral', among which are those mutations that are slightly deleterious. Support for the nearly neutral theory was originally provided by the excess of low-frequency allozyme (protein) variants within human populations, and has since been strengthened by evidence from recent studies of nucleotide polymorphism. For example, nonsynonymous polymorphisms have been found to be less variable on average (as measured by expected heterozygosity) than synonymous polymorphisms (e.g. Hughes *et al*., 2003), consistent with the interpretation that selection is maintaining slightly deleterious mutations at low frequencies. Also, in comparisons of ratios of nonsynonymous and synonymous change, the P_n/P_s ratio for genome-wide surveys of human polymorphism shows a value of 38.42%, which is substantially greater than the D_n/D_s ratio of divergence between humans and chimpanzees, which stands at 23.76% (e.g. Bustamante *et al*., 2005). Furthermore, the P_n/P_s ratio for rare polymorphism (i.e. those polymorphisms in which the minor allele is present at less than 20% in the population) has been found to be considerably higher than the P_n/P_s of common polymorphism (those above 20% frequency). All these results are expected if purifying selection is acting on slightly deleterious mutations, maintaining them at low frequencies within our species and removing them from the population before they can drift to fixation.

When the P_n/P_s ratios for rare and common classes of polymorphism are compared, the difference between them reflects the fraction of slightly deleterious mutations removed from the population before these mutations can attain higher frequencies, and contribute to common polymorphism. Thus, the method represents a way to estimate the fraction of slightly deleterious mutations in the human population. Using this approach, values ranging from 12 to 25% were obtained, depending on which particular dataset and gene class were analysed. In another approach, which consisted in estimating selection coefficients, 13.5% of loci were found to have significantly greater P_n/P_s than D_n/D_s (Bustamante *et al*., 2005), indicating that negative selection is acting on mutations at these loci. Although the exact values are still uncertain, it has become apparent that a class of slightly deleterious mutations contributes to a substantial part of human polymorphism.

The Influence of Effective Population Size on Selection and Genetic Drift

Differences in effective population size (N_e) are expected to influence the effectiveness of natural selection. Thus, we expect smaller populations to have lower rates of adaptive evolution, and also lower rates of purifying selection.

The relationship between N_e and the intensity of purifying selection is supported when the d_n/d_s ratio is compared between the genomes of species having different values for N_e. For example, the d_n/d_s ratio in comparisons between rat and mouse genomes is approximately 0.13, while that between the human and chimpanzee genomes (having considerably smaller N_e values) is nearly 0.20 (Mikkelsen *et al*., 2005), indicating less purifying selection in chimpanzees and humans compared to rats and mice. The same relationship holds when we compare humans and chimpanzees. Humans are estimated to have an N_e several times smaller (approximately 10 000) compared to chimpanzees (52 000–96 000) (Chen and Li, 2001). Bakewell *et al*. (2007) calculated a d_n/d_s ratio nearly 0.24 for chimpanzees, found to be significantly smaller than the ratio in humans (approximately 0.26), indicating less purifying selection during human evolution.

As with purifying selection, positive selection is more effective in larger populations compared to smaller populations, and species with larger N_e are expected to show relatively more adaptive change. The large amount of adaptive change estimated for *Drosophila*, bacteria and viruses (as noted above) may be examples of this relationship given the large population sizes of these organisms. In comparisons of humans with chimpanzees, several studies have estimated that considerably more genes (up to 50% more) have been positively selected along the chimpanzee lineage than along the human lineage (Arbiza *et al*., 2006; Bakewell *et al*., 2007). Thus, the evolution in protein-coding genes in the human lineage appears to be influenced by drift to a greater extent than in lineages of our closest relative or other organisms.

Methodological Problems and Challenges

The accuracy of estimates concerning the amount of adaptive change in humans, and other species, will improve as new data emerge and as researchers discover the factors that can lead to biases in these estimates. Such factors as changes in population size and mutational biases can lead

to biased estimates of adaptive change, and need to be accounted for in analytical methods. Difficulties also arise because adaptive change may have been targeted at specific amino acids within genes, and this signal may be lost when the entire sequence is analysed (see Hughes, 2007). There is also the possibility that much adaptive change has occurred outside the protein-coding fraction of the genome, and though we are only just beginning to survey these regions, initial analyses are yielding encouraging signs. In the following section, these methodological problems and challenges are discussed.

Accounting for slightly deleterious mutations

It appears that a large fraction of polymorphism in humans is contributed by slightly deleterious mutations. Thus, if we assume that population size remains roughly constant over time, P_n/P_s is overestimated, since slightly deleterious mutations are overrepresented in polymorphism over divergence data, leading to an underestimation of adaptive evolution. This difficulty has been addressed by using only the fraction of polymorphisms that are common (e.g. those with frequencies greater than 0.20, and that are therefore less likely to include slightly deleterious variants). On the other hand, if populations experience expansions in size, P_n/P_s will be reduced due to the increased effectiveness of purifying selection in the larger population, leading to overestimates of adaptive evolution. When population bottlenecks occured in the past, slightly deleterious mutations can be fixed at a higher rate than predicted from the polymorphism data, and can also lead to false inferences of positive selection. Such demographic effects must be taken into account when applying these tests especially for human populations that are known to have undergone large size fluctuations during evolutionary history. An approach to dealing with the presence of slightly deleterious mutations is to carry out the McDonald–Kreitman type analyses under models that can explicitly account for mutations with selective effects that are neither very deleterious nor neutral. In this way, Bustamante *et al*. (2005) were able to provide estimates of selection coefficients for each locus.

Making tests based on d_n/d_s more realistic

The requirement of $d_n/d_s > 1$ to infer positive selection is clearly very conservative, since nearly all genes evolve predominantly under purifying selection, with a small subset of codons under positive selection. New versions of the d_n/d_s tests, which allow for variation among codons when testing the null hypothesis of neutrality have been developed, and have the potential to detect positive selection even if a small number of codons are under selection (e.g. Nielsen and Yang, 1998).

In addition, mutational biases can affect the results of d_n/d_s tests. The usual assumption in this test is that the rate of nonsynonymous mutations is equivalent to the rate of sequence divergence at synonymous sites and at sites within noncoding regions. However, Subramanian and Kumar (2006) pointed out that coding sequences are enriched for hypermutable CpG dinucleotide sites, which by inference would indicate that the nonsynonymous mutation rate is greater than the synonymous rate. They argued that if this mutation bias is left uncorrected, it leads to an underestimation in the amount of purifying selection, and possibly an overestimation of adaptive evolution.

Developing tests for adaptive evolution at nonprotein-coding loci and considering regulatory evolution

Although many methods for detecting positive selection were developed for protein-coding genes, new methods are being extended to noncoding regions. One method is a modified McDonald–Kreitman test that compares the ratio of substitutions to polymorphisms at sites in different regions flanking or within coding genes where most mammalian gene expression control sequences are found (e.g. at 5′ intergenic and in 5′ first intron regions) and compares them to a similar ratio at putatively neutral sites within introns. Interestingly, comparing humans and chimpanzees, this method detected no evidence of adaptive evolution in such regions, and furthermore found little or no evidence that these regions were conserved by purifying selection (Keightley *et al*., 2005). Another method has been applied to identify conserved noncoding sequences across the genome and then search within these sequences for evidence of accelerated change in recent lineages (Prabhakar *et al*., 2006). Within the human lineage, results indicated that there are 80% more accelerated sequences than would be expected by chance, with sequences tending to fall near genes in particular categories, for example neuronal cell adhesion. It appears quite possible that recent accelerated change in these sequences could signify changes in gene regulation that underlie important phenotypic adaptations, but much further work is needed. **See also**: Gene evolution and human adaptation; Molecular Evolution: Introduction; Molecular Evolution: Neutral Theory; Purifying Selection: Action on Silent Sites; Selective and Structural Constraints; Synonymous and Nonsynonymous Rates

References

Arbiza L, Dopazo J and Dopazo H (2006) Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Computational Biology* **2**(4): e38.

Bakewell MA, Shi P and Zhang J (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences of the USA* **104**(18): 7489–7494.

Bustamante CD, Fledel-Alon A, Williamson S *et al*. (2005) Natural selection on protein-coding genes in the human genome. *Nature* **437**(7062): 1153–1157.

Chen FC and Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of

the common ancestor of humans and chimpanzees. *American Journal of Human Genetics* **68**(2): 444–456.

Clark AG, Glanowski S, Nielsen R *et al*. (2003) Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. *Science* **302**(5652): 1960–1963.

Eyre-Walker A (2006) The genomic rate of adaptive evolution. *Trends in Ecology and Evolution* **21**(10): 569–575.

Fay JC, Wyckoff GJ and Wu CI (2001) Positive and negative selection on the human genome. *Genetics* **158**(3): 1227–1234.

Gojobori J, Tang H, Akey JM and Wu CI (2007) Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. *Proceedings of the National Academy of Sciences of the USA* **104**(10): 3907–3912.

Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99**(4): 364–373.

Hughes AL, Packer B, Welch R *et al*. (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proceedings of the National Academy of Sciences of the USA* **100**(26): 15754–15757.

Keightley PD, Lercher MJ and Eyre-Walker A (2005) Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biology* **3**(2): e42.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature* **217**(5129): 624–626.

Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

King MC and Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science* **188**(4184): 107–116.

McDonald JH and Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**(6328): 652–654.

Mikkelsen TS, LaDeana WH, Eichler EE *et al*. (Chimpanzee Sequencing and Analysis Consortium) (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**(7055): 69–87.

Nielsen R and Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**(3): 929–936.

Nielsen R, Bustamante C, Clark AG *et al*. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology* **3**(6): e170.

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* **246**: 96–98.

Prabhakar S, Noonan JP, Paabo S and Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**(5800): 786.

Sabeti PC, Schaffner SF, Fry B *et al*. (2006) Positive natural selection in the human lineage. *Science* **312**: 1614–1620.

Subramanian S and Kumar S (2006) Higher intensity of purifying selection on >90% of the human genes revealed by the intrinsic replacement mutation rates. *Molecular Biology and Evolution* **23**(12): 2283–2287.

Zhang L and Li WH (2005) Human SNPs reveal no evidence of frequent positive selection. *Molecular Biology and Evolution* **22**(12): 2504–2507.

Further Reading

Fay JC and Wu CI (2001) The neutral theory in the genomic era. *Current Opinions in Genetics and Development* **11**(6): 642–646.

Fay JC and Wu CI (2003) Sequence divergence, functional constraint, and selection in protein evolution. *Annual Review of Genomics and Human Genetics* **4**: 213–235.