Systems (eds Greenspan, R. J. & Kyriacou, C. P.) 15–28 (John Wiley & Sons, New York, 1994).
11. Pflugfelder, G. O. Genetic lesions in *Drosophila* behavioural mutants. *Behav. Brain Res.* **95**, 3–15 (1998).
12. Shaffer, P. *Amadeus* (Harper & Row, New York, 1980).
13. Lawrence, P. A. *The Making of a Fly* (Blackwell Science, Oxford, 1992).
14. Beadle, G. W. & Tatum, E. L. Genetic control of biochemical reactions in *Neurospora*. *Proc. Natl Acad. Sci. USA* **27**, 499–506 (1941).
15. Wood, W. B. & Edgar, R. S. Building a bacterial virus. *Sci. Am.* **217**, 60–74 (1967).
16. Nusslein-Volhard, C., Frohnhofer, H. G. & Lehmann, R. Determination of anteroposterior polarity in *Drosophila*. *Science* **238**, 1675–1681 (1987).
17. Jan, Y. N. & Jan, L. Y. Genetic control of cell fate specification in *Drosophila* peripheral nervous system. *Annu. Rev. Genet.* **28**, 373–393 (1994).
18. Leptin, M. Gastrulation in *Drosophila*: the logic and the cellular mechanisms. *EMBO J.* **18**, 3187–3192 (1999).
19. Tully, T., Preat, T., Boynton, S. C. & Del Vecchio, M. Genetic dissection of consolidated memory in *Drosophila*. *Cell* **79**, 35–47 (1994).
20. Simon, M. A., Botwell, D. L., Dodson, G. S., Laverty, T. R. & Rubin, G. M. *Ras1* and a putative guanine nucleotide exchange factor perform crucial steps in signalling by the sevenless protein tyrosine kinase. *Cell* **67**, 701–716 (1991).
21. O'Kane, C. & Gehring, W. J. Detection *in situ* of genomic regulatory elements in *Drosophila*. *Proc. Natl Acad. Sci. USA* **84**, 9123–9127 (1987).
22. Rorth, P. A modular misexpression screen in *Drosophila* detecting tissue-specific phenotypes. *Proc. Natl Acad. Sci. USA* **93**, 12418–12422 (1996).
23. Gerlai, R. Gene-targeting studies of mammalian behavior: is it the mutation or the background genotype? *Trends Neurosci.* **19**, 177–181 (1996).
24. de Belle, J. S. & Heisenberg, M. Expression of *Drosophila* mushroom body mutations in alternative genetic backgrounds: a case study of the mushroom *body miniature gene* (*mbm*). *Proc. Natl Acad. Sci. USA* **93**, 9875–9880 (1996).
25. Osborne, K. A. *et al.* Natural behavior polymorphism due to a cGMP-dependent protein kinase of *Drosophila*. *Science* **277**, 834–836 (1997).
26. Greenspan, R. J. A kinder, gentler genetic analysis of behavior: dissection gives way to modulation. *Curr. Opin. Neurobiol.* **7**, 805–811 (1997).
27. Pereira, H. S., Macdonald, D. E., Hilliker, A. J. & Sokolowski, M. B. *Chaser* (*Csr*), a new gene affecting larval foraging behavior in *Drosophila melanogaster*. *Genetics* **141**, 263–270 (1995).
28. Griffith, L. C., Wang, J., Zhong, Y., Wu, C. F. & Greenspan, R. J. Calcium/calmodulin-dependent protein kinase II and potassium channel subunit *eag* similarly affect plasticity in *Drosophila*. *Proc. Natl Acad. Sci. USA* **91**, 10044–10048 (1994).
29. Fedorowicz, G. M., Fry, J. D., Anholt, R. R. & Mackay, T. F. Epistatic interactions between smell-impaired loci in *Drosophila melanogaster*. *Genetics* **148**, 1885–1891 (1998).
30. Clark, A. G. & Wang, L. Epistasis in measured genotypes: *Drosophila* P-element insertions. *Genetics* **147**, 157–163 (1997).
31. Greenspan, R. J. & Tully, T. in *Flexibility and Constraint in Behavioral Systems* (eds Greenspan, R. J. & Kyriacou, C. P.) 65–80 (Dahlem Konferenzen, Berlin, 1994).
32. Cooke, J., Nowak, M. A., Boerlijst, M. & Maynard-Smith, J. Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends Genet.* **13**, 360–362 (1997).
33. Misawa, H. *et al.* Contrasting localizations of MALS/LIN-7 PDZ proteins in brain and molecular compensation in knockout mice. *J. Biol. Chem.* **276**, 8264–9272 (2001).
34. Tononi, G., Sporns, O. & Edelman, G. M. Measures of degeneracy and redundancy in biological networks. *Proc. Natl Acad. Sci. USA* **96**, 3257–3262 (1999).
35. Edelman, G. M. in *The Mindful Brain* (eds Edelman, G. M. & Mountcastle, V. B.) 51–100 (MIT Press, Cambridge, Massachusetts, 1978).
36. Edelman, G. M. *Topobiology* (Basic Books, New York, 1989).
37. Wodicka, L., Dong, H., Mittmann, M., Ho, M.-H. & Lockhart, D. J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.* **15**, 1359–1367 (1997).
38. Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S. & Garrels, J. I. A sampling of the yeast proteome. *Mol. Cell Biol.* **19**, 7357–7368 (1999).
39. Livesey, F. J., Furukawa, T., Steffen, M. A., Church, G. M. & Cepko, C. L. Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene *Crx*. *Curr. Biol.* **10**, 301–310 (2000).
40. Barkai, N & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913–917 (1997).
41. Bhalla, U. S. & Iyengar, R. Emergent properties of networks of biological signaling pathways. *Science* **283**, 381–387 (1999).
42. Harwood, J. *Styles of Scientific Thought: The German Genetics Community 1900–1933* (Chicago Univ. Press, Chicago, 1993).
43. Gould, S. J. & Lewontin, R. C. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond.* B **205**, 581–598 (1979).
44. Farah, M. J. Neuropsychological inference with an interactive brain: a critique of the 'locality' assumption. *Behav. Brain Sci.* **17**, 43–61 (1994).
45. Dudai, Y., Jan, Y.-N., Byers, D., Quinn, W. G. & Benzer, S. *dunce*, a mutant of *Drosophila* deficient in learning. *Proc. Natl Acad. Sci. USA* **73**, 1684–1688 (1976).
46. Byers, D., Davis, R. L. & Kiger, J. A. Jr Defect in cyclic AMP phosphodiesterase due to the *dunce* mutation of learning in *Drosophila melanogaster*. *Nature* **289**, 79–81 (1981).
47. Bellen, H. J., Gregory, B. K., Olsson, C. L. & Kiger, J. A. Jr Two *Drosophila* learning mutants, *dunce* and *rutabaga*, provide evidence of a maternal role for cAMP in embryogenesis. *Dev. Biol.* **121**, 432–444 (1987).
48. Boynton, S. & Tully, T. *latheo*, a new gene involved in associative learning and memory in *Drosophila melanogaster*. *Genetics* **131**, 655–672 (1992).
49. Pinto, S. *et al. latheo* encodes a subunit of the origin recognition complex and disrupts neuronal proliferation and adult olfactory memory when mutant. *Neuron* **23**, 45–54 (1999).
50. Heisenberg, M., Wonneberger, R. & Wolf, R. *optomotor-blind*^{H31} — a *Drosophila* mutant of the lobula plate giant neurons. *J. Comp. Physiol.* A **124**, 287–296 (1978).
51. Pflugfelder, G. O. *et al.* The *lethal(1)optomotor-blind* gene of *Drosophila melanogaster* is a major organizer of optic lobe development: isolation and characterization of the gene. *Proc. Natl Acad. Sci. USA* **89**, 1199–1203 (1992).
52. Kopp, A. & Duncan, I. Control of cell fate and polarity in the adult abdominal segments of *Drosophila* by optomotor-blind. *Development* **124**, 3715–3726 (1997).
53. Wu, C.-F., Ganetzky, B., Jan, L. Y., Jan, Y.-N. & Benzer, S. A *Drosophila* mutant with a temperature-sensitive block in nerve conduction. *Proc. Natl Acad. Sci. USA* **75**, 4047–4051 (1978).
54. Lee, C. G., Chang, K. A., Kuroda, M. I. & Hurwitz, J. The NTPase/helicase activities of *Drosophila maleless*, an essential factor in dosage compensation. *EMBO J.* **16**, 2671–2681 (1997).
55. Kernan, M. J., Kuroda, M. I., Kreber, R., Baker, B. S. & Ganetzky, G. *nap*^{ts}, a mutation affecting sodium channel activity in *Drosophila*, is an allele of *mle*, a regulator of X chromosome transcription. *Cell* **66**, 949–959 (1991).
56. Pak, W. L., Grossfield, W. J. & Arnold, K. S. Mutants of the visual pathway of *Drosophila melanogaster*. *Nature* **227**, 518–520 (1970).
57. Bloomquist, B. T. *et al.* Isolation of a putative phospholipase C gene of *Drosophila*, *norpA*, and its role in phototransduction. *Cell* **54**, 723–733 (1988).
58. Dushay, M. S., Rosbash, M. & Hall, J. C. The disconnected visual system mutations in *Drosophila melanogaster* drastically disrupt circadian rhythms. *J. Biol. Rhythms* **4**, 1–27 (1989).
59. Riesgo-Escovar, J., Raha, D. & Carlson, J. R. Requirement for a phospholipase C in odor response: overlap between olfaction and vision in *Drosophila*. *Proc. Natl Acad. Sci. USA* **92**, 2864–2868 (1995).

Acknowledgements

Helpful comments on the manuscript were provided by C. Deutsch, G. Edelman, J. Gally, J. Hall and F. Jones, for which I am very grateful. I thank K. McCarthy for assistance with Latin grammar.

OPINION

Predicting adaptive evolution

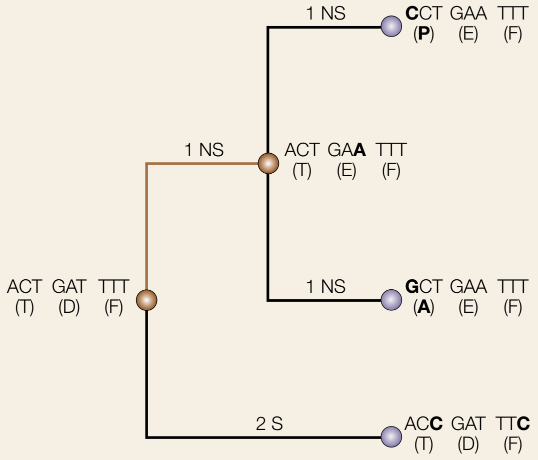
Robin M. Bush

Phylogenetic trees reconstruct past evolution and can provide evidence of past evolutionary pressure on genes and on individual codons. In addition to tracing past evolutionary events, molecular phylogenetics might also be used to predict future evolution. Our ability to verify adaptive hypotheses using phylogenetics has broad implications for vaccine design, genomics and structural biology.

It is well documented that some genes evolve more quickly than others; for instance, in the human species, certain histone genes are highly conserved, whereas immunoglobulin loci are extremely polymorphic[1]. A lack of genetic variation might indicate the occurrence of purifying selection — a force that preserves the adapted condition and that is therefore typically observed in functionally important genes. By contrast, extensive variation in genes indicates that the encoded protein might benefit from undergoing amino-acid replacements. Such positive selection has been recently observed in genes that have an adaptive function. Until now, it has been difficult to link the patterns of molecular variation to the selective pressures responsible for them. However, in some systems, notably in viral species, sufficient sequence data now exist to test adaptive hypotheses directly using phylogenetic analysis.

Phylogenetic trees are a graphic means of reconstructing evolution on the basis of similarity between the characters of the individuals under study; the length of a horizontal branch on the tree reflects the amount of change between an individual and its nearest ancestor (BOX 1). Evolutionary pressure on a gene or codon can be detected by comparing the rates of synonymous (silent) and non-synonymous (amino-acid changing, or non-silent) nucleotide substitutions across the branches of a tree. In the absence of selection, the synonymous and non-synonymous substitution rates should be equal (FIG. 1a). Most coding genes show an

Box 1 | Phylogenetic trees



Phylogenetic trees are a graphic means of representing the relationships between individuals on the basis of their similarities. In the diagram, the sampled individuals are represented by terminal nodes (purple dots) and are connected by branches. Terminal nodes are connected to their inferred ancestors, the internal nodes (brown dots), by terminal branches, whereas internal nodes are connected to one another by internal branches. The length of the horizontal branches represents the genetic distance between individuals. In the example above, the two uppermost nodes are considered each others' nearest relative because they are identical at eight out of nine nucleotide positions, the bottom sequence is evolutionarily more distant as it is identical to the top sequence at only five sites (single letters in brackets refer to amino acids), bold text highlights a nucleotide substitution.

Several methods exist to construct phylogenetic trees, but those most commonly used are known as maximum parsimony, maximum likelihood and neighbour joining. The maximum-parsimony approach to constructing an evolutionary tree operates on the principle that simple solutions are preferred to more complex ones. This means that the preferred tree will be one that requires the smallest number of evolutionary events. In this example, a minimum of five nucleotide substitutions (in bold) are required to reconstruct the evolution history that links the three terminal nodes. Maximum-likelihood methods infer the tree topology (branching sequence) that is most consistent with the observed data. These methods calculate the possibility that any given topology will produce the observed sequences if calculated for all or many possible topologies that have been constructed according to pre-defined evolutionary hypotheses. Neighbour joining is a type of cluster analysis in which pairs of nodes are iteratively combined to form larger and larger trees (starting with the two most closely related nodes) based on the minimal distance between clusters. For a thorough review of these and other phylogenetic methods, see REF. 25. (NS, non-synonymous; S, synonymous.)

excess of synonymous substitutions, which indicates that purifying (stabilizing) selection is operating to preserve the current structure and function of the protein (FIG. 1b). Neutral or conserved substitution patterns provide limited insight into the evolutionary process, because the phylogenetic tree provides no additional information as to why the gene evolved in this manner.

Much more interesting studies are possible when substitution rate analyses indicate the occurrence of positive selection (FIG. 1c). Positive selection is natural selection that favours amino-acid change. Continual positive selection leaves a characteristic pattern on a phylogenetic tree in the form of a greater rate of non-synonymous than synonymous substitution. Potentially, these trees can provide a great deal of additional information about the nature of adaptive change in a system. Typically, only a small number of codons per gene seem to be positively selected. In proteins of known structure, studying the effects of changing these particular residues might lend insight into the functional role of the protein. In proteins of unknown structure, knowing the location of positively selected residues in the two-dimensional structure provides a starting point to determine the three-dimensional structure of the protein, as these residues typically lie in positions exposed to external selective forces. In addition, we can test adaptive hypotheses by correlating change over time (across the tree) at the putative, positively selected codons with changes in phenotype or in fitness. Given a sufficient understanding of how a protein responds to selection in a particular system, it might be possible to predict its response to future selective challenges.

In this article, I outline the theoretical basis for the research into substitution rate analysis and summarize the biological systems in which evidence of positive selection has been detected. I discuss the cases in which predicting evolution might be realistic, along with some of the potential pitfalls encountered in this type of work. Last, I present some practical applications of substitution rate analysis: for epitope identification in vaccine design, for the determination of protein structure, and as a tool for interpreting the results of whole-genome sequencing projects.

Positive selection

The study of adaptive evolution using substitution rate analysis involves two basic steps. The first involves reconstructing the evolutionary history of a gene in the form of a phylogenetic tree. A tree depicts the changes that occur as sequences descend from a common ancestor (BOX 1). In the second step, the tree is used to estimate the non-synonymous and synonymous nucleotide substitution rates over time (BOX 2). A substitution rate might be calculated for the entire gene by summing substitutions across codons; however, with sufficient data, rates might also be estimated for each individual codon. Tests that sum substitutions across codons might fail to identify positively selected genes if high non-synonymous substitution rates occur at only a few codons. Despite this drawback, genic level studies have identified a number of putative, positively selected genes (see REF. 2 for a comprehensive list). Most of these genes fall into two principal groups: pathogen surface proteins, and sperm proteins of aquatic animals that practice external fertilization.

The surface proteins of pathogens must change their three-dimensional structure to avoid recognition by antibodies that are raised in response to previous antigen exposure. Therefore, it is likely that evasion of the host immune system drives repeated amino-acid replacements in surface proteins. Indirect support for this assumption has been found: the codons in genes with a high non-synonymous substitution rate typically code for residues that are exposed on the surface of the pathogen. Examples include the *porB* gene of the gonorrhoea-inducing bacterium *Neisseria gonorrhoea*[3], which encodes protein channels in the lipopolysaccharide layer of Gram-negative bacteria, and the gp120 envelope gene of the human immunodeficiency virus (HIV-1)[4]. The same is true of haemagglutinin (HA), which, along with neuroaminadase (NA), is one of the most antigenic surface proteins of the influenza virus. Here, the positively selected residues lie on the surface of the protein, within known antibody-binding sites[5].

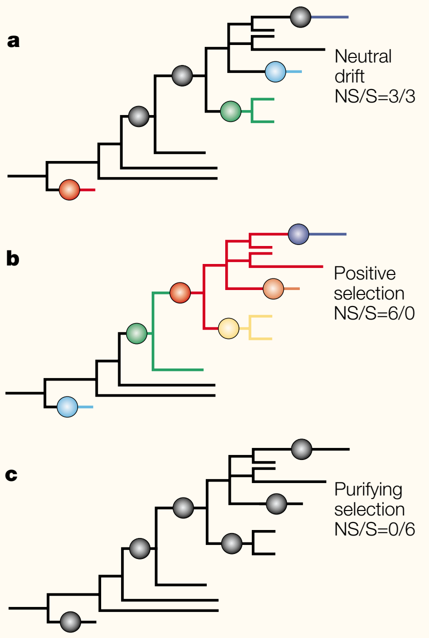


Figure 1 | **Effects of selection on substitution rates.** Non-synonymous (NS) and synonymous (S) nucleotide substitutions that typify three selective regimes: **a** | selective neutrality (NS = S), **b** | purifying (stabilizing) selection, which conserves the present sequence (NS < S), and **c** | positive selection, which favours amino-acid replacement (NS > S). Non-synonymous substitutions are represented by coloured dots and synonymous substitutions are indicated by black dots.

Pathogens exert strong selective pressure on their hosts so, not surprisingly, there is evidence of positive selection from various host-defence systems. Human major histocompatibility complex (MHC) antigen-recognition sites seem to be positively selected[6]. Although plants use resistance genes and chitinases (enzymes that degrade fungal cell walls) for defence rather than the T cells, MHC and antibodies that are used by animals, surface proteins of plant pathogens show signs of positive selection[7,8] as do plant-defence systems[9]. These studies suggest that host–pathogen systems involve exquisite matching between the pathogen and host receptors.

Proteins that are involved in the reproduction of externally fertilizing marine organisms provide another class of genes under positive selection. The two most extensively studied cases are the lysin gene of abalones (a shellfish)[10–13] and the bindin gene in sea urchins[14–17]. Lysin, which is released from sperm at fertilization, dissolves the vitelline coat of the egg in a species-specific manner; bindin is a sperm acrosomal protein that mediates species-specific recognition and binding between the sperm and the egg after the sperm has penetrated the egg jelly. Enforcement of species-specific sperm recognition might be the principal selective force for change in bindin and lysin, as host specificity for sperm recognition requires correct matching of the sperm surface with receptors on the egg. Positively selected codons in abalone lysin lie on the surface of the molecule and are associated with structural features that are thought to be involved in binding[13]. In terms of the need for specific matching, this system shares many similarities with the host–pathogen studies above.

In summary, reasonable adaptive hypotheses have been proposed to explain how certain patterns of genetic change might have been produced by positive selection. But how do we test whether positive selection actually occurred?

Testing adaptive hypotheses

Positive selection produces an excess of non-synonymous substitutions on a phylogenetic tree. However, this excess alone is not sufficient evidence to invoke positive selection. Support for the hypothesis requires an increase in fitness caused by amino-acid replacements at the putative, positively selected sites. So far, there has been only one test of this hypothesis, using the gene for haemagglutinin, the principal surface antigen of the H3N2 subtype of human influenza A (H3N2 refers to the particular HA and NA gene variants that it contains).

Human influenza evolves so rapidly that vaccine strains must be updated almost yearly. Selection favours haemagglutinin variants that escape recognition by the antibodies that are formed in response to past infection or vaccination. New lineages of H3N2 influenza A that differ in their haemagglutinin arise frequently. As shown in FIG. 2, at any given time several closely related lineages co-circulate. For reasons that are not yet understood, all but a single lineage dies out within a few years. Relative fitness, the rate of increase of a genotype relative to other genotypes in a population, is thus unambiguous in this system.

We constructed phylogenetic trees that represented the evolution of the H3 human haemagglutinin gene through 11 successive influenza seasons. We found that lineages undergoing the greatest number of new amino-acid replacements at putative, positively selected codons were fitter than other lineages in 9 out of 11 recent influenza seasons[5,18]; that is, lineages with the most replacements would outcompete all others. These results support the hypothesis that replacement substitutions at positively selected codons more effectively changed the shape of the haemagglutinin with respect to antibody recognition than did substitutions at other codons.

Predicting evolution

In the previous section, I showed that positive selection can explain the high rate of non-synonymous versus synonymous substitution in human influenza haemagglutinin. In essence, these retrospective tests involved going back in time to see whether we could predict subsequent evolution in our 11 years of data. Our predictions were successful in 9 out of 11 years.

Our studies are based on the assumption that the selective pressure on influenza during our study period was directed towards avoiding immune recognition. We also assume that this selective pressure persists today and, based on this assumption, propose that circulating strains with the most additional mutations at these same positively selected codons at present will be the progenitors of future influenza lineages. It remains to be seen how well our hypothesis holds up.

Influenza is perhaps the only natural system at present available in which it is possible to try to predict evolution at the population level. This is due to three main factors. First, haemagglutinin evolves very rapidly, allowing us to observe change easily. Second, haemagglutinin is one of the best-studied genes in terms of positive selection[5,18–24]. One reason for this is the high quality of the available data. Sequences used in our work and that of many other recent studies were generated by the Influenza Branch of the US Centers for Disease Control and Prevention (CDC, see link) as part of the World Health Organization (WHO) influenza surveillance programme (see link). Sequences that date back to the 1968 emergence of the H3N2 subtype in humans are available, along with data on the date of collection and laboratory culture, through the Influenza Sequence Database at Los Alamos National Laboratory (see link). Third, prediction might be limited to influenza because of the unambiguous measure of fitness available in this system, which is assessed by the survival or extinction of a particular lineage.

The only system for which the wealth of sampling data approaches influenza is the human immunodeficiency virus (HIV), which also evolves rapidly. However, in contrast to influenza, many new mutant lineages of HIV survive at the population level, rather than just one. There is no clear means by which to compare the fitness of different HIV isolates in a population on a real-time basis. Linear replacement of HIV-1 strains

Box 2 | Substitution rate estimation

The most popular method for estimating positive selection on the genic level is that of Nei and Gojobori[35], as implemented in the computer program MEGA (Molecular Evolutionary Genetics Analysis, see link). This program determines statistical differences between synonymous and non-synonymous substitution rates calculated from pairwise sequence comparisons. One drawback of this method is that it might fail to find evidence for positive selection when only a few codons in the gene are positively selected.

The simplest method for detecting positive selection on individual codons compares the number of non-synonymous and synonymous substitutions at each codon with binomial expectations based on the average numbers of substitutions across codons[5]. A more sophisticated version of this method has since been developed in which expectations are calculated according to the probability of non-synonymous or synonymous change for the specific residues in each position[22]. These methods require a relatively large data set because each codon has to undergo a substantial number of substitutions to obtain a statistically significant result.

Small data sets can be analysed using maximum-likelihood approaches[36]. These methods test substitution patterns against expectations based on various models of evolution, including positive selection. Maximum-likelihood techniques are computationally intensive, making their use on large data sets problematic[24]. A more crucial problem with these methods is that they might make more assumptions than can be justified given the small size of the typical data set[37].

over time has been seen in individual human hosts[25], so it might be possible to predict its evolution in this limited context. In influenza, prediction has a direct clinical application in terms of vaccine strain selection. Unfortunately, the equivalent for HIV — developing vaccines for individual HIV-infected patients on the basis of evolution of the virus within their bodies — is not possible at this time.

Evolutionary prediction might be feasible in other model systems. Wichman, Bull and collaborators showed that genetic change occurred over time at many of the same positions in two related bacteriophage strains that evolved on *Escherichia coli* hosts in the laboratory[26,27]. Several of these sites showed evidence of positive selection, and these sites made up a disproportionately large share of positions that differed between the two parental phage strains. The positively selected residues were all surface exposed. Site-directed mutagenesis shows that these residues affect host binding, but lie outside of the putative binding site. Whether the positively selected residues are involved directly in binding is, as yet, unknown, but the hypothesis that these residues increase fitness could be tested using direct-competition experiments.

Potential pitfalls

We encountered three problems in our work on influenza A that are rarely, if ever, addressed in other studies of substitution rates[5,18]. The experimental pitfalls that I describe in this section are not specific to influenza and concern errors in phylogenetic reconstruction[5], artefacts caused by laboratory evolution[28] and sampling bias[29].

Phylogenetic uncertainty. Insufficient sampling can cause error in phylogenetic reconstruction, and consequently error in identifying codons that are under positive selection. One way to estimate sampling error is a statistical technique called bootstrap analysis[30]. In this technique, new data sets are created by randomly sampling characters from the original data set. The resulting data sets are the same size as the original, but some characters have been left out and others duplicated. The bootstrap value of a node (a branch division on a tree) is the percentage of times that node is present in the set of trees constructed from the new data sets. A bootstrap value of 95% or higher is typically considered good statistical support for a node. We obtained poor bootstrap support for a large number of nodes in our influenza A haemagglutinin tree. When we examined hundreds of equally plausible trees we found an excess of non-synonymous substitutions at some codons in only a small number of trees[5]. Because we planned further studies based on these results[18], we limited our list of putative, positively selected codons to those present in most trees. I know of only one other study that examined this problem: analyses using the HIV-1 gp120 gene were reported to be robust to error in tree topology[4].

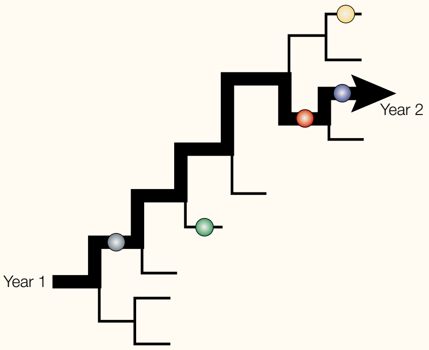


Figure 2 | **Predicting evolution.** This phylogenetic tree represents a simplified view of influenza A haemagglutinin evolution during a single year. New mutant lineages continually arise and then all but one become extinct. Dots indicate amino-acid replacements at codons known to have been positively selected in the past. In our studies, the single lineage that survives, shown here in bold, has typically undergone the greatest number of additional amino-acid replacements at the known positively selected codons.

Laboratory evolution. The study of adaptive evolution focuses on pathogens because of their medical importance and because they evolve quickly enough to be studied in real time. Many pathogens also quickly adapt to laboratory culture conditions; so, sequences obtained from culture might contain artefacts that introduce error into substitution rate analysis. A pertinent example involves human influenza, which is typically cultured inside chicken eggs. Egg-adapted amino-acid replacements are known to occur around the receptor-binding pocket of the haemagglutinin protein. If undetected, these mutations will be assigned as an extra mutation on the terminal branch of a phylogenetic tree (BOX 1). This is simply because the affected sequence is grouped with its nearest relative on the basis of similarity at the hundreds of unaffected codons in addition to the egg-adapted codon. We estimated that about 8% of the amino-acid replacements in our influenza data set were egg-adapted artefacts[28]. To prevent lab artefacts from affecting our analyses, we eliminated all mutations assigned to terminal branches before calculating substitution rates.

Laboratory evolution results in amino-acid replacements in both the gp120 envelope glycoprotein of HIV-1 (REF. 31) and the VP1 capsid protein of foot-and-mouth disease virus[32]. It would be interesting to know whether laboratory artefacts also affect the substitution rate studies of these pathogens[4,23,33].

Sampling bias. Influenza isolates sent to the CDC from the WHO collection centres are screened for antigenic similarity to known circulating strains. Isolates that are antigenically indistinguishable from reference strains on the basis of a haemagglutinin-binding test are typically not sequenced. This purposeful sampling bias increases the number of non-synonymous substitutions in our data even when these substitutions imparted no selective advantage to the virus

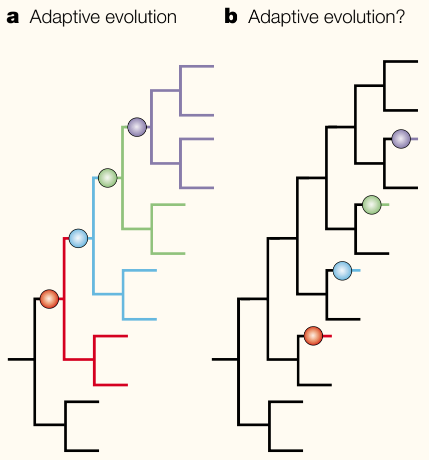


Figure 3 | **Adaptive evolution.** The inference we draw from relative substitution rate analysis should be interpreted in the light of where the substitutions appear on the tree. In this illustration, both trees have four non-synonymous substitutions at a single codon, shown as dots, but no synonymous substitutions. Do the two trees provide equal evidence of positive selection? **a** | Each of the mutations on the tree swept to fixation in the population, implying that they were selectively advantageous. **b** | The same mutations occurred in lineages that quickly became extinct. The pattern seen in **a** is therefore stronger evidence for positive selection. In our work on influenza, we eliminated all terminal mutations in our analyses[5,18].

in nature. This sampling bias is most pronounced in the class of substitutions assigned to the terminal branches of the tree[29]. So, when we eliminated mutations on terminal branches to minimize the effects of laboratory evolution on our analyses, we also reduced the degree to which we overestimated the non-synonymous rate because of sampling bias. I know of no other studies in which the effects of sampling bias have been examined.

Elimination of the mutations that were assigned to the terminal branches of our haemagglutinin tree resulted in a 70% reduction in the number of mutations available for substitution rate analysis. However, the remaining data contained strong evidence for positive selection. Selectively advantageous mutations are, by definition, retained in a population longer than are neutral or deleterious substitutions that occur at the same time. Changes that persist in the population will be assigned to the internal (as opposed to terminal) branches of phylogenetic trees (FIG. 3a). It is difficult to attribute an excess of non-synonymous mutations to positive selection when they occur on lineages that quickly become extinct (FIG. 3b).

Future applications

Although the prospect of predicting evolution is exciting, predictions can be verified only in very rapidly evolving systems. More practical applications of substitution rate analysis include identifying epitopes for vaccine development, constructing theoretical models of protein structure, and interpreting the results of genome sequencing projects.

As described above, all of the putative, positively selected codons of the influenza virus are located in known antibody-binding sites on the exposed surface of the haemagglutinin[5]. If these binding sites had not been previously identified, our analyses would have pointed to their location. Identifying functionally important sites might develop as one of the chief uses of substitution rate analysis. These methods could be particularly helpful in searching for conformational epitopes — antigenic structures composed of non-contiguous residues that lie near one another only when the protein is correctly folded. These might appear in a gene as scattered codons that show similar evidence of positive selection.

Results of positive selection studies can also be used to help guide construction of theoretical protein structure models. For example, the structures of many porins have yet to be resolved using X-ray crystallography. Protein purification seems to destroy bonds between the porin and other components of the cell membrane that are crucial to its three-dimensional conformation. In the *porB* gene of *Neisseria gonorrhoea*[3], the putative, positively selected segments lie on the exposed loops of the porin. We might reasonably expect regions of other porins to show this pattern. In another example, Ishimizu *et al.*[34] identified four regions in the seminal RNase (*S-RNase*) gene that have an excess of non-synonymous substitutions. This gene is associated with the self-incompatibility system in the Rosaceae. Homology searches based on predicted secondary structure indicate that these four regions are exposed on the surface of the style (the portion of the female reproductive organ on which pollen grains attach and germinate) and thus are candidate sites for recognition of self-derived pollen. These results indicate that identifying the surface-exposed segments of a protein using substitution rate analysis could, along with two-dimensional structure prediction, provide a basis for constructing three-dimensional models of proteins that lack amino-acid homology with proteins of known crystal structure.

Whole-genome surveys provide another exciting area to apply these techniques. Open reading frames obtained from genome surveys are screened for possible function using homology searches against sequences already held in GenBank. Simultaneous calculation of the number of synonymous and non-synonymous differences between homologous sequence pairs could help to identify genes that are under intense natural selection. Results from more sophisticated screens could benefit from substitution rate analysis as well. For instance, Intercell (see link), in collaboration with The Institute for Genomic Research (TIGR, see link), recently announced (at the American Society for Microbiology — TIGR 2001 conference on microbial genomes) an antigen identification technique in which the peptide products of shredded genomic DNA from the *Staphylococcus aureus* genome were exposed to human antibodies. The question is whether the peptide regions to which antibodies bind in such a screen are also antigenic in their natural form. One might pursue this question by sequencing the same regions in related organisms and by contrasting non-synonymous and synonymous substitution rates. On the basis of the data reviewed above, a high rate of non-synonymous substitution would provide an excellent reason to suspect that a region binds antibodies in its intact as well as in its shredded form.

> "… practical applications of substitution rate analysis include identifying epitopes for vaccine development, constructing theoretical models of protein structure, and interpreting the results of genome sequencing projects."

In summary, we now have analytical methods to identify genes, gene segments and individual codons that are under selective pressure to change. If the evolutionary forces using this selection can be identified, predicting the future course of evolution might be possible in systems such as rapidly evolving pathogens. The broader application of these methods are exciting and diverse, as they bring a new research tool to vaccine design, genome sequence interpretation and protein structure prediction.

Robin M. Bush is at the Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697, USA. e-mail: rmbush@uci.edu

Links

FURTHER INFORMATION *Neisseria gonorrhoea* | HIV-1 gp120 | foot-and-mouth disease virus | US Centers for Disease Control and Prevention | World Health Organization influenza surveillance programme | Influenza Sequence Database at Los Alamos National Laboratory | X-ray crystallography primer | Rosaceae | American Society for Microbiology — TIGR 2001 conference | *Staphylococcus aureus* genome | Intercell | The Institute for Genomic Research | Molecular Evolutionary Genetics Analysis | Phylogenetic Analysis by Maximum Likelihood | Robin Bush's lab | Walter Fitch's lab

1. Li, W.-H. & Graur, D. *Fundamentals of Molecular Evolution* (Sinauer Associates, Sunderland, Massachusetts, 1991).
2. Yang, Z. H. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496–503 (2000).
3. Smith, N. H., Maynard Smith, J. & Spratt, B. G. Sequence evolution of the *porB* gene of *Neisseria gonorrhoeae* and *Neisseria meningitidis*: evidence of positive Darwinian selection. *Mol. Biol. Evol.* **12**, 363–370 (1995).
4. Yamaguchi-Kabata, Y. & Gojobori, T. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J. Virol.* **74**, 4335–4350 (2000).
5. Bush, R. M., Fitch, W. M., Bender, C. A. & Cox, N. J. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.* **16**, 1457–1465 (1999).
6. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
7. Wang, G. L. *et al. Xa21D* encodes a receptor-like molecule with a leucine-rich repeat domain that determines race-specific recognition and is subject to adaptive evolution. *Plant Cell* **10**, 765–779 (1998).
8. Meyers, B. C., Shen, K. A., Rohani, P., Gaut, B. S. & Michelmore, R. W. Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. *Plant Cell* **10**, 1833–1846 (1998).
9. Bishop, J. G., Dean, A. M. & Mitchell-Olds, T. Rapid evolution in plant chitinases: molecular targets of selection in plant–pathogen coevolution. *Proc. Natl Acad. Sci. USA* **97**, 5322–5327 (2000).
10. Lee, Y. H., Ota, T. & Vacquier, V. D. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.* **12**, 231–238 (1995).
11. Hellberg, M. E. & Vacquier, V. D. Rapid evolution of fertilization selectivity and lysin cDNA sequences in teguline gastropods. *Mol. Biol. Evol.* **16**, 839–848 (1999).
12. Yang, Z. H., Swanson, W. J. & Vacquier, V. D. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol. Biol. Evol.* **17**, 1446–1455 (2000).
13. Kresge, N., Vacquier, V. D. & Stout, C. D. Abalone lysin: the dissolving and evolving sperm protein. *Bioessays* **23**, 95–103 (2001).
14. Metz, E. C. & Palumbi, S. R. Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol. Biol. Evol.* **13**, 397–406 (1996).
15. Vacquier, V. D., Swanson, W. J. & Lee, Y. H. Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? *J. Mol. Evol.* **44**, S15–S22 (1997).
16. Biermann, C. H. The molecular evolution of sperm bindin in six species of sea urchins (Echinoida: Strongylocentrotidae). *Mol. Biol. Evol.* **15**, 1761–1771 (1998).
17. Palumbi, S. R. All males are not created equal: fertility differences depend on gamete recognition polymorphisms in sea urchins. *Proc. Natl Acad. Sci. USA* **96**, 12632–12637 (1999).
18. Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J. & Fitch, W. M. Predicting the evolution of human influenza A. *Science* **286**, 1921–1925 (1999).
19. Fitch, W. M., Leiter, J. M. E., Li, X. Q. & Palese, P. Positive Darwinian evolution in human influenza A viruses. *Proc. Natl Acad. Sci. USA* **88**, 4270–4274 (1991).
20. Ina, Y. & Gojobori, T. Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses. *Proc. Natl Acad. Sci. USA* **91**, 8388–8392 (1994).
21. Fitch, W. M., Bush, R. M., Bender, C. A. & Cox, N. J. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl Acad. Sci. USA* **94**, 7712–7718 (1997).
22. Suzuki, Y. & Gojobori, T. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **16**, 1315–1328 (1999).
23. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449 (2000).
24. Yang, Z. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J. Mol. Evol.* **51**, 423–432 (2000).
25. Strunnikova, N., Ray, S. C., Livingston, R. A., Rubalcaba, E. & Viscidi, R. P. Convergent evolution within the V3 loop domain of human immunodeficiency virus type 1 in association with disease progression. *J. Virol.* **69**, 7548–7558 (1995).
26. Wichman, H. A., Badgett, M. R., Scott, L. A., Boulianne, C. M. & Bull, J. J. Different trajectories of parallel evolution during viral adaptation. *Science* **285**, 422–424 (1999).
27. Wichman, H. A., Scott, L. A., Yarber, C. D. & Bull, J. J. Experimental evolution recapitulates natural evolution. *Phil. Trans. R. Soc. Lond.* B **355**, 1677–1684 (2000).
28. Bush, R. M., Smith, C. B., Cox, N. J. & Fitch, W. M. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proc. Natl Acad. Sci. USA* **97**, 6974–6980 (2000).
29. Bush, R. M., Fitch, W. M., Smith, C. B. & Cox, N. J. in *Options for the Control of Influenza IV* (ed. Osterhaus, A. D. M. E.) (Elsevier, Amsterdam, in the press).
30. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
31. Sullivan, N. *et al.* CD4-induced conformational changes in the human immunodeficiency virus type 1 gp120 glycoprotein: consequences for virus entry and neutralization. *J. Virol.* **72**, 4694–4703 (1998).
32. Martínez, M. A., Verdaguer, N., Mateu, M. G. & Domingo, E. Evolution subverting essentiality: dispensability of the cell attachment Arg-Gly-Asp motif in multiply passaged foot-and-mouth disease virus. *Proc. Natl Acad. Sci. USA* **94**, 6798–6802 (1997).
33. Martín, M. J., Núñez, J. I., Sobrino, F. & Dopazo, J. A procedure for detecting selection in highly variable viral genomes: evidence of positive selection in antigenic regions of capsid protein VP1 of foot-and-mouth disease virus. *J. Virol. Methods* **74**, 215–221 (1998).
34. Ishimizu, T. *et al.* Identification of regions in which positive selection may operate in S-RNase of Rosaceae: implication for S-allele-specific recognition sites in S-RNase. *FEBS Lett.* **440**, 337–342 (1998).
35. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
36. Nielsen, R. & Yang, Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936 (1998).
37. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, Oxford and New York, 2000).

Acknowledgements

I thank N. Cox, C. Bender Smith and W. Fitch for their contributions to our collaborative work on influenza. I am supported by a grant from the US National Institutes of Health and by funds provided by the University of California for the conduct of discretionary research by Los Alamos National Laboratory, conducted under the auspices of the US Department of Energy.

OPINION

Protecting genetic privacy

Patricia A. Roche and George J. Annas

This article outlines the arguments for and against new rules to protect genetic privacy. We explain why genetic information is different to other sensitive medical information, why researchers and biotechnology companies have opposed new rules to protect genetic privacy (and favour anti-discrimination laws instead), and discuss what can be done to protect privacy in relation to genetic-sequence information and to DNA samples themselves.

The simultaneous publication of two versions of the human genome could be an important impetus to take more seriously the legal, ethical and social policy issues at stake in human genome research[1,2]. There are many such issues, and the one that has caused the most public concern is that of genetic privacy. As DNA sequences become understood as information, and as this information becomes easier to use in digitized form, public concerns about internet and e-commerce privacy (regarding the security with which an individual's private details are protected) will merge with concerns about medical record privacy and genetic privacy. In this paper, we outline the key public policy issues at stake in the genetic privacy debate by reviewing generally medical privacy, by asking whether genetic information is like other medical information, and by outlining the current controversies over privacy in genetic research. We conclude with some public policy recommendations.

Privacy

Privacy is a complex concept that involves several different but overlapping personal interests. It encompasses informational privacy (having control over highly personal information about ourselves), relational privacy (determining with whom we have personal, intimate relationships), privacy in decision-