The power of phylogenetic comparison in revealing protein function

Ziheng Yang*

Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, England

Positive selection, that is, fixation of advantageous mutations driven by natural selection, has been an exciting topic to evolutionary biologists, because adaptive changes in genes and genomes are ultimately responsible for evolutionary innovations and species differences. In recent years, detecting signals of natural selection has also become a powerful approach for molecular biologists, biochemists, and virologists to understand the functions of new genes. In a recent issue of PNAS, Sawyer *et al.* (1) described a remarkable study in which phylogenetic sequence comparison identified a small segment of the primate TRIM5α protein to be under positive selection, and functional analysis using mutagenesis confirmed the importance of the segment in species-specific retroviral inhibition.

TRIM5α is a protein in the cellular antiviral defense system in primates and can restrict retroviruses such as HIV-1 and simian immunodeficiency virus in a species-specific manner. The protein was identified recently because the rhesus monkey TRIM5α restricts reverse transcription of HIV-1, whereas the native human TRIM5α does not (2). Sawyer *et al.* (1) sequenced the *TRIM5α* gene from a number of primate species, including hominoids and Old and New World monkeys. Their phylogenetic analysis estimating evolutionary rates identified amino acid positions in the protein at which natural selection appears to have actively promoted amino acid substitutions. In particular, a 13-aa "patch" in the SPRY domain had a concentration of positively selected sites, implicating it as an antiviral interface. By creating chimeric *TRIM5α* genes, Sawyer *et al.* demonstrated that this patch is responsible for most of the species-specific antiretroviral activity. Previous studies using phylogenetic approaches have identified a number of genes under positive selection, especially genes involved in host–pathogen interactions. However, this study is unique in that there was no *a priori* information, structural or otherwise, concerning which part of the protein acts as the interaction interface for viral restriction, and that the adaptive significance of the SPRY patch was entirely a computational prediction, which was validated by further experimental analysis.

In contrast to site-directed mutagenesis, in which mutations are artificially introduced and their effects on function are assayed in the laboratory, evolution may be viewed as Nature's grand experiment, conducted much more slowly but over a vast expanse of geologic time. In this experiment, mutations occur in the genes and genomes, and boom or bust through random genetic drift. Natural selection acts as a filter, weeding out lethal or deleterious mutations while driving advantageous mutations to fixation. Therefore, genetic differences between species we observe today are products of this complex process. Thus, if a gene is highly divergent among species, there are two main explanations for this divergence: (*i*) a high mutation rate or weakened purifying selection (known as relaxed selective constraint) and (*ii*) positive natural selection favoring changes. In the former case, the gene is essentially free to vary because it has no fitness or functional significance. In the latter, the high variability is promoted by natural selection and the gene has extremely important functions. To establish the action of positive selection at the molecular level, one has to rule out the alternative interpretation of relaxed selective constraint, which is typically very difficult (3).

A segment of TRIM5α is under positive selection and is important in species-specific retroviral inhibition.

Comparison of Silent and Replacement Rates Provides an Effective Approach to Detecting Selection on the Protein

For protein-coding genes, an effective approach is to contrast the rates at which synonymous (silent) and nonsynonymous (replacement, amino acid altering) mutations are fixed in the population. The silent rate d_S provides a benchmark against which we can decide whether the replacement rate d_N is accelerated or diminished by natural selection acting on the protein (4). Thus, $d_N < d_S$, $d_N = d_S$, and $d_N > d_S$ represent negative purifying selection, neutral evolution, and positive Darwinian selection, respectively. A problem with this criterion is its lack of power. Most proteins have highly conserved regions or amino acid residues where replacement mutations are not tolerated and d_N is essentially 0. Furthermore, adaptive evolution may occur in an episodic fashion and only in a narrow window of evolutionary time (5). Comparison of a pair of genes, averaging the d_N and d_S rates over all sites in the protein and over the whole time period separating the two sequences, typically fails to infer positive selection, because the signal of positive selection is overwhelmed by the ubiquitous purifying selection. To boost the power of the detection method, recent work has focused on detecting selection that affects individual sites (as opposed to the whole protein) (6–8) or particular lineages (as opposed to the whole phylogeny) (9–11). Those improved detection methods have been much more successful.

Many genes have been detected to be under positive selection by comparing silent and replacement substitution rates (12). Most of these genes fall into three major categories. The first category includes host genes involved in defense or immunity against viral, bacterial, fungal, or parasite attacks, as well as viral or pathogen genes involved in evading host defense. The former includes the major histocompatibility complex (13, 14), lymphocyte protein CD45 (15), plant R-genes involved in pathogen recognition (16), and plant chitinases which confer disease resistance by attacking fungal cell walls (17), to name a few recent examples. The latter includes, among many others, viral surface or capsid proteins (18, 19), *Plasmodium* membrane antigens (20), and polygalacturonases produced by plant enemies,

See companion article on page 2832 in issue 8 of volume 102.

*E-mail: z.yang@ucl.ac.uk.

such as bacteria, fungi, oomycetes, nematodes, and insects (21). One may expect that it is to the pathogen's advantage to mutate into new forms unrecognizable by the host defense system, while it is to the host's advantage to adapt and recognize the pathogen. Thus, an evolutionary arms race ensues, driving new replacement mutations to fixation. The *TRIM5α* protein studied by Sawyer *et al.* (1) is a retroviral inhibitor and belongs to this category. Toxins in snake or scorpion venoms are used to subdue prey and often evolve under similar positive selective pressures (22, 23). The second main group includes proteins or pheromones involved in reproduction (24). It is best for the sperm to recognize and fertilize the egg as soon as possible. However, the egg is a substantial investment and it is best for the egg recognition protein to evolve to avoid fertilization by multiple sperm. The conflicts of interests between the two sexes espouse a genetic battle. A third group of proteins include those that acquired new functions after gene duplications, such as the pancreatic ribonuclease genes in leaf-eating monkeys (25) and xanthine dehydrogenase genes (26).

Other genes have been detected to be under positive selection as well, but they are not as numerous as those involved in evolutionary arms race, such as the host–pathogen antagonism, or reproduction. This pattern appears to be due, in part, to the limitation of the detection methods. These methods rely on excessive replacement substitutions relative to silent substitutions and may not be able to detect one-off adaptive evolution in which an advantageous mutation arose and was fixed in the population quickly, followed by purifying selection. This expectation was confirmed by computer simulations (27). On one hand, focusing on particular lineages or individual sites increases the chance of detecting episodic and local adaptation. On the other hand, the narrowed window reduces opportunities for multiple substitutions, making it difficult for such tests to achieve statistical significance.

Such an Approach of Combining Phylogenetic Analysis with a Well Designed Experiment May Be Applicable to Many Other Systems

Two recent studies took a similar approach to that of Sawyer *et al.* (1). Ivarsson *et al.* (28) identified positively selected amino acid residues in glutathione transferase, multifunctional enzymes that provide cellular defense against toxic electrophiles of both exogenous and endogenous origins. They then used site-directed mutagenesis to confirm that those mutations were capable of driving functional diversification in substrate specificities. The evolutionary comparison thus provided a novel approach to designing new proteins. Bielawski *et al.* (29) detected positively selected amino acid sites in proteorhodopsin, a retinal-binding membrane protein in marine bacteria that functions as a light-driven proton pump. Site-directed mutagenesis and functional assay demonstrated that those sites were responsible for fine-tuning the light absorption sensitivity of the protein to different light intensities in the ocean. In these studies, the comparative analysis has played the role of generating biological hypotheses for validation in the laboratory. The success of these studies suggest that such an approach combining phylogenetic analysis with well designed experiment may be applicable to many other systems. In particular, it may be applied to large-scale comparisons of genes from whole genomes.

Current studies in comparative genomics have mostly relied on similarity matches using BLAST, which are effective in recognizing distant but related sequences. The strategy makes use of negative purifying selection to infer functionally conserved regions in the genome, under the premise that sequence regions conserved across distantly related species are most likely to be functionally important. Protein-coding genes, RNA genes, and regulatory elements have different levels of sequence conservation and can be identified by comparison of species at different evolutionary distances (see ref. 30). As more genomes are sequenced from closely related species, statistical modeling and comparative analysis should make it possible to infer positive Darwinian selection. Clark *et al.* (31) recently performed this kind of study, comparing human and chimpanzee genes with the mouse used as the outgroup. They identified a collection of genes under positive selection along the human lineage, including a few involved in olfaction and speech or underlying known Mendelian disorders, which might be responsible for the differences between the human and the chimpanzee. The study by Sawyer *et al.* (1) and a small handful of similar studies have demonstrated the great potential of phylogenetic methods for detecting molecular adaptation in generating interesting hypotheses to be verified through laboratory experiment.

1. Sawyer, S. L., Wu, L. I., Emerman, M. & Malik, H. S. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 2832–2837.
2. Stremlau, M., Owens, C. M., Perron, M. J., Kiessling, M., Autissier, P. & Sodroski, J. (2004) *Nature* **427,** 848–853.
3. Akashi, H. (1999) *Gene* **238,** 39–51.
4. Miyata, T. & Yasunaga, T. (1980) *J. Mol. Evol.* **16,** 23–36.
5. Gillespie, J. H. (1991) *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford).
6. Nielsen, R. & Yang, Z. (1998) *Genetics* **148,** 929–936.
7. Suzuki, Y. & Gojobori, T. (1999) *Mol. Biol. Evol.* **16,** 1315–1328.
8. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. (2000) *Genetics* **155,** 431–449.
9. Messier, W. & Stewart, C.-B. (1997) *Nature* **385,** 151–154.
10. Yang, Z. (1998) *Mol. Biol. Evol.* **15,** 568–573.
11. Zhang, J., Rosenberg, H. F. & Nei, M. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 3708–3713.
12. Yang, Z. & Bielawski, J. P. (2000) *Trends Ecol. Evol.* **15,** 496–503.
13. Hughes, A. L. & Nei, M. (1988) *Nature* **335,** 167–170.
14. Yang, Z. & Swanson, W. J. (2002) *Mol. Biol. Evol.* **19,** 49–57.
15. Filip, L. C. & Mundy, N. I. (2004) *Mol. Biol. Evol.* **21,** 1504–1511.
16. Lehmann, P. (2002) *J. Appl. Genet.* **43,** 403–414.
17. Bishop, J. G., Dean, A. M. & Mitchell-Olds, T. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 5322–5327.
18. Haydon, D. T., Bastos, A. D., Knowles, N. J. & Samuel, A. R. (2001) *Genetics* **157,** 7–15.
19. Shackelton, L. A., Parrish, C. R., Truyen, U. & Holmes, E. C. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 379–384.
20. Polley, S. D. & Conway, D. J. (2001) *Genetics* **158,** 1505–1512.
21. Götesson, A., Marshall, J. S., Jones, D. A. & Hardham, A. R. (2002) *Mol. Plant Microbe Interact.* **15,** 907–921.
22. Duda, T. F. & Palumbi, S. R. (2000) *Mol. Biol. Evol.* **17,** 1286–1293.
23. Zhu, S., Bosmans, F. & Tytgat, J. (2004) *J. Mol. Evol.* **58,** 145–153.
24. Swanson, W. J. & Vacquier, V. D. (2002) *Nat. Rev. Genet.* **3,** 137–144.
25. Zhang, J., Zhang, Y. P. & Rosenberg, H. F. (2002) *Nat. Genet.* **30,** 411–415.
26. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 13413–13417.
27. Anisimova, M., Bielawski, J. P. & Yang, Z. (2002) *Mol. Biol. Evol.* **19,** 950–958.
28. Ivarsson, Y., Mackey, A. J., Edalat, M., Pearson, W. R. & Mannervik, B. (2002) *J. Biol. Chem.* **278,** 8733–8738.
29. Bielawski, J. P., Dunn, K. A., Sabehi, G. & Beja, O. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 14824–14829.
30. Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., Blanchette, M., Siepel, A. C., Thomas, P. J., McDowell, J. C., *et al.* (2003) *Nature* **424,** 788–793.
31. Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., *et al.* (2003) *Science* **302,** 1960–1963.