Convergent evolution: the need to be explicit

Russell F. Doolittle

Convergence as a phenomenon in molecular evolution is an issue that confuses many discussions. Often the problem is that not enough care is taken to state exactly what kind of convergence one has in mind. Functional and mechanistic convergence are both common, and some structural convergence has probably occurred, but a convincing case for genuine sequence convergence has yet to be made.

ONE OF THE most frequently asked questions after any lecture on the phylogenetic analysis of amino acid sequences is: 'What about convergence?' Usually, the implication is that some sequence resemblances could be due to convergence rather than divergence and, as a result, the determined phylogeny mistaken. The term 'convergence' is used in many different contexts, however, and much confusion can occur when the subject is raised. As in all matters, a little care taken to define just what is meant can eliminate needless controversy. Here I'd like to categorize the various kinds of convergence that occur in molecular evolution in an effort to dispel some mistaken notions as to how common or uncommon the phenomenon may be.

Convergence implies adaptive change in which lesser related entities come to appear more related than they are. It should not be confused with chance resemblance. We are all familiar with convergence at the morphological level, whereby unrelated or distantly related creatures come to look like each other, usually to the advantage of one or both. Ordinarily the resemblance is superficial and can be detected as such upon careful examination. Some marsupials look very much like certain carnivores, and porpoises and whales have found advantage in being shaped like fish, but the comparative anatomist has no trouble sorting out the marsupials from the placental mammals or the marine mammals from the fish. The game can be more intriguing at the molecular level, however, and the rules are slightly different, as I will try to show.

Functional convergence

Functional convergence is what occurs when some molecular 'functionality' arises independently on more than one occasion. There are numerous examples of enzymes that catalyse the same reactions being concocted independently. For instance, the ability to catalyse the hydrolysis of peptide bonds has evolved many times: there are sulfhydryl proteases, metalloproteases, aspartyl proteases, serine proteases, and some others[1]. Moreover, the serine proteases have evolved on at least three different occasions, as evidenced by the subtilisin, trypsin and α–β type enzymes[2].

Indeed, this kind of functional convergence among enzymes is unexpectedly common. A list of some separately evolved pairs of enzymes that catalyse the same reaction in each case is provided in Box 1. It includes the ubiquitous superoxide dismutases, as well as some glycolytic enzymes such as sugar kinases and aldolases. The phenomenon is not restricted to enzymes. Plants and animals have independently evolved inhibitors of the same proteases and, even more impressive, different bacteria have concocted binding proteins that bind to the very same parts of immunoglobulins[3].

Mechanistic convergence

One of the major surprises to emerge from early X-ray structure determinations of proteins was the observation that in two different serine proteases the sidechains of three amino acids were gathered together into the same geometry from completely different folds and sequence arrangements[4]. Thus, chymotrypsin has a 'catalytic triad' with a histidine at sequence position 57, an aspartate at position 102 and a serine at position 195. The bacterial protease subtilisin, on the other hand, manages the same chemistry with an aspartate at position 32, a histidine at position 64 and a serine at position 221. Moreover, as implied above, the three-dimensional structures of the two types of enzymes are completely different. Since that initial discovery, it has been found that catalytic triads of the general sort [Asp/Glu] His [Ser/Thr] occur in numerous settings[5,6].

Box 1. Some enzymes that have evolved independently on more than one occasion

- Superoxide dismutases
- Aldolases
- Sugar kinases
- Serine proteases
- Alcohol dehydrogenases
- Aminoacyl tRNA synthetases
- Ribonucleotide reductases
- Topoisomerases
- PEP carboxykinases
- Malate dehydrogenases

Structural convergence

It is well known that certain structural motifs, including various barrel and bundle arrangements, are widespread among contemporary proteins. For example, potential α and β segments often occur alternately in a protein sequence and fold into parallel α–β barrels, the three-dimensional structures of which are all remarkably similar, but for which no detectable sequence resemblance exists[7]. The question of whether all these proteins are descended from a common ancestor or whether there has been a convergence to a common structure remains open[8]. Certainly many of them are descended from a common ancestor[9], but their great diversity of function could reflect a general convergence to the same structure, either as a result of the intrinsic stability of these barrels or their ease of formation.

A similar situation exists with regard to β-barrels, a disproportionate number of which have the same strand arrangement. A particularly intriguing example involves the fibronectin type III and immunoglobulin domains, both of which occur in numerous proteins.

R. F. Doolittle is at the Center for Molecular Genetics, University of California, San Diego, La Jolla, CA 92093-0634, USA.

© 1994, Elsevier Science Publishers, (UK) 0968–0004/94/$07.00

Both are all-β sandwiches composed of three- and four-stranded β-sheets[10]. The topology of the seven principal β-strands is identical in both except that the fourth strand is somewhat shifted so that the immunoglobulin domain comprises four versus three β-strands whereas the fibronectin unit comprises three versus four. Still, much of the backbone of the two types of structures is superimposable. There is no detectable sequence similarity. Is this a remarkable structural convergence, or did these two kinds of domain share a common ancestor long ago?

Nucleotide-binding folds

There was a time when it was thought that a single nucleotide-binding fold evolved once and, because of the fundamental importance of this interaction to the evolution of metabolism, radiated throughout the entire biological realm. Today it is realized that not only are there two widely spread 'classical' folds, one for mononucleotides and one for dinucleotides, but there are also other ways of binding the same nucleotides to various proteins[11]. Initially, a well-studied group of small-molecule kinases and guanine-nucleotide-binding proteins was found to have a large anionic hole for binding phosphate groups. The architecture of this hole depends on several adroitly positioned glycines and a key lysine for providing the positive charge. Because these residues are closely spaced in a sequence sense, the members of this family can be identified by the simple motif GXXGXGK. When an X-ray structure was recently determined for a protein kinase[12], a similar large anionic hole was revealed. It was also glycine-dependent and had an essential lysine, but in this case the lysine came from a different part of the chain. The overall fold was completely different, and no sequence resemblance to other mononucleotide-binding proteins was observed. The provision of these similarly functioning holes has been referred to as 'convergent evolution'[11].

Claims of sequence convergence

In none of the cases mentioned so far is there the slightest hint of sequence convergence. To the contrary, common experience shows that sequence divergence is the general rule in protein evolution. If the sequences of a given protein from a variety of species are compared, usually the resemblances between them are roughly proportional to the biological relatedness of the organisms as judged by nonmolecular measures[13–15]. The major driving force behind this divergence is entropy, itself a derivative of neutral replacement. Eventually, sequences diverge so much that no significant resemblance remains, even though the overall shape of the protein has hardly changed. Clearly, there are multitudinous ways of maintaining similar three-dimensional structures with different arrangements of 20 amino acids. Still, anomalous phylogenies based on sequences are not uncommon. Often these are attributable to the statistical vagaries of stochastic events. Beyond that, however, the usual suspects are: (1) unequal rates of change along different lineages, (2) horizontal gene transfers and (3) convergent sequence evolution.

We must keep in mind that the word 'sequence' implies that it is the order of the amino acids that is important. In this regard, we must distinguish long-range sequence similarities, of the sort used to construct phylogenies, from local resemblances that are merely constrained by requirements for certain kinds of sidechains. In theory, similar sequences can be due to chance, convergence or common ancestry. By my definition, convergence differs from chance similarity in that it depends on adaptive replacements that are positively selected. Clearly, in genuine cases of sequence convergence, the adaptive replacements must outnumber those similarities occurring merely by chance. Unfortunately, when people talk about sequence convergence they often have very different degrees of convergence in mind. At one extreme, what is implied is the evolution of similar sequences in completely unrelated proteins, the resemblance being driven by the need for a particular skein of amino acids to satisfy some specific function or structural attribute. At the other extreme is the case in which a small number of adaptive amino acid replacements has occurred in already similar sequences. Between these extremes is the troublesome case of homologous proteins whose sequence resemblances are 'out of the expected order', and there is an inclination to blame 'convergence'. Some illustrations should make the point.

Fibrinopeptides. Many years ago, we reported a strong tendency for the same amino acid replacements to occur along different branches of a phylogenetic tree of the highly variable fibrinopeptides[14]. Glutamate and aspartate seemed to change back and forth with impunity along the divergent lineages. We referred to this kind of parallel evolution as 'conservative variability' and did not use the word convergent. Most of these changes were effectively neutral and not adaptive, but the nature of the changes made some short sequences in diverging creatures look more similar than expected[13]. Other replacements must have occurred at these positions, of course, but presumably they were selected against. Negative selection, which is to say natural rejection, is a mainstay of neutral evolution and keeps it in check. In any event, if comparisons are limited to very short sequences, say five or six residues, it might seem to some observers that convergence had occurred. But when longer stretches are examined, the general trend for divergence overwhelms those chance identities at the locations of conservative variability and puts things in proper perspective.

HIV envelope proteins. Recently, a case of alleged convergence was reported that involves human immunodeficiency virus (HIV) envelope proteins. Specifically, it was concerned with changes occurring during the propagation of the virus in a single patient[16]. The course of change was followed over a period of seven years: one particular hexapeptide sequence varied in different isolates in a way that the authors thought was convergent. Thus, the sequence GPGRAF changed to GPGSAV along two separate lineages, one proceeding via the sequence GPGRAV and the other via GPGSAF. In my view, this kind of variation is similar to that which can be observed in any phylogeny of fibrinopeptides, and sequence convergence based on only two back-and-forth changes seems to be overstating the case.

Visual pigments. It has been claimed that certain visual proteins have arisen in primates and fishes in a convergent manner[17]. Thus, long-wavelength-absorbing opsins thought to correspond to the red and green visual pigments from humans have been found in certain fish, and in each case the red and green types appear to have descended from a common ancestor. Did they gain their new absorption properties as a result of the same adaptive amino acid replacements? As it happens, the red and green human proteins differ in only 15 of their 350 residues, some or all of which must impart the different spectral properties.

In the case of the fish, the two proteins thought to correspond to red- and green-absorbing proteins differ by at least 64 amino acid replacements, only eight of which occur at the locations at which the human red- and green-absorbing proteins differ. In three of these, the amino acid replacements are the same, leading the authors to refer to this as 'convergent evolution'. The three replacements were Ala/Thr, Ala/Ser and Tyr/Phe. Whether or not these replacements actually contribute to the spectral properties of the fish proteins (and there is no independent evidence that they do), these rather modest changes could well fall under the rubric of conservative variability, and the use of the term convergence seems too strong.

Langur digestive enzymes. A frequently mentioned case of adaptive changes giving rise to 'sequence convergence' involves digestive enzymes in leaf-eating columbine monkeys like the langur. These monkeys have a diet not so different from that of ruminants, and their digestive tracts have adapted to the point where they have a compartmentalized stomach and a specialized bacterial flora that can digest cellulose[18]. Like true ruminants, the leaf-eating monkeys have also recruited lysozyme for the disposal of those bacteria[19]. The extraordinary anatomical and physiological convergence of this system led to an investigation of the sequence of langur lysozyme to see if convergence had also occurred at the molecular level[20]. In several instances changes have occurred in the lineage leading to the langur that parallel those found in the bovine enzyme, and some of these may be adaptations to the low pH of the stomach pouch. Unfortunately, these observations are repeatedly referred to as sequence convergence[21,22], leading to the widespread but incorrect notion that the langur lysozyme sequence has converged with that of ruminants to the point where the two sequences cluster together in sequence-based phylogenies. As shown in Fig. 1, computer-generated phylogenies of lysozyme sequences are completely orthodox, in spite of adaptive replacements on the lineages leading to the langur and bovids. As might be expected, the small number of adaptive replacements is set against a backdrop of so many neutral changes that their impact on a sequence-based phylogeny is negligible. In this case there are 31 differences between the bovine and langur lysozyme sequences but only 14 between langur and baboon. Sequence convergence as such simply has not occurred.

A similar situation exists with regard to the langur pancreatic ribonuclease, an enzyme adapted for digesting the large amounts of RNA released by the lysed bacterial flora[23]. Here again, the enzyme appears to have adapted to its digestive role, as reflected by its relatively low isoelectric point, just as has occurred in the pancreatic ribonucleases of ruminants. These studies remind us that not all amino acid replacements in proteins are neutral, but their bearing on sequence-based phylogenies is not very significant.

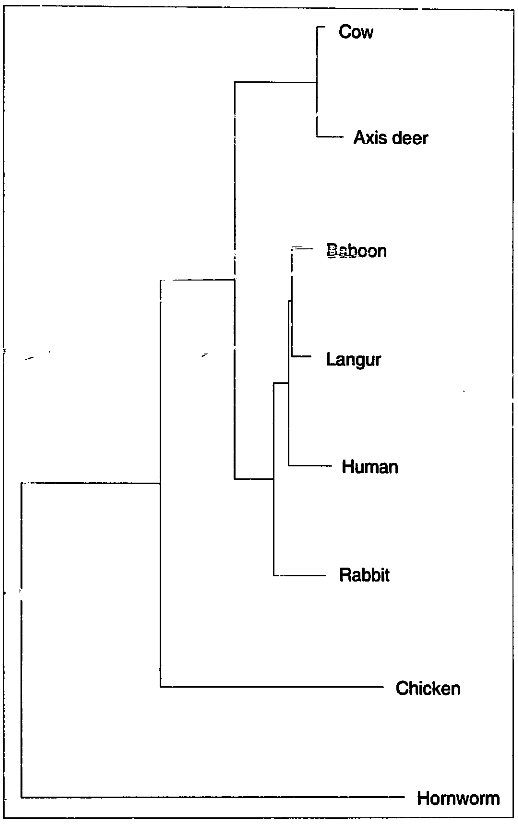

Figure 1
Phylogenetic tree of assorted lysozyme sequences showing that bovine and langur sequences occur in their expected positions. The tree was made by a matrix method after progressive alignment of the sequences[30]. Once the eight sequences were provided, all further operations were conducted automatically. Horizontal distances are proportional to evolutionary distances.

Molecular mimicry

The cases of claimed sequence convergence described above all involve the parallel evolution of sequences that are similar to begin with. This is quite different from evolving similar sequences from completely unrelated ancestors. Does the latter ever occur? What about molecular mimicry?

Molecular mimicry is a term usually reserved for when a parasite or host 'mimics' a protein made by the other, either to avoid an immune response or to interfere with some other vital process[24,25]. There is a large literature on immunological cross-reactivity in this

```
(a)
HIV gp160 (254-525)          ..V V S T Q L L L N G.. (249) ..V G I G A L F L G F L G A..
HLA β (142-215)              ..V V S T P L I R N G.. (51) ...V G G F V L G L L F L G A..

(b)
Circumsporozite (340-364)       ..N S I S T E W S P C S V T C G N G I Q V R I K P G..
TRAP (244-268)                  ..C G V W D E W S P C S V T C G K G T R S R K R E I..

Thrombospondin (382-406)        ..W S L W S T W A P C S V T C S E G S Q L R Y R R C..
Properdin (80-404)              ..W S P W S E W T S C S T S C G N G I Q Q R G R S C..
```

Figure 2
Two possible cases of 'molecular mimicry' in which unrelated proteins of parasite and host have similar sequences. **(a)** The gp160 protein from HIV appears similar to two different regions in human HLA β (Ref. 26). **(b)** Two malarial parasite proteins have segments that are similar to two human blood proteins[27,28]. See text for caveats.

field, but only a few cases where sequence convergence has been claimed. As a case in point, however, it has been claimed that the HIV envelope protein contains sequences that mimic that of part of the human HLA β chain[26]. In particular, there is one stretch of ten residues that has seven identities and another that has eight identities out of 13 (Fig. 2a). This was termed 'convergent evolution' by the authors[26] because no other significant resemblances exist between these two proteins. Still, the case for mimicry is weakened by the fact that one of these stretches is known to be membrane-spanning in the case of HLA, but its correspondent in the HIV protein is not. In any event, the 'sequence convergence' is limited and localized.

In another case, it has been reported that some proteins from malarial parasites have remarkable sequence similarities to the human plasma proteins thrombospondin and properdin[27,28]. Properdin plays a key role in immunity, and it was suggested that mimicry might be involved[28]. In one comparison there is a stretch of 23 residues with 14 identities (Fig. 2b). It has not been ruled out, however, that the protozoan proteins are actually homologs of the mammalian proteins[27], in which case the similarity is not so remarkable. It is necessary to be especially careful in this case because the similar sequences involve an evolutionarily mobile domain that has been shuffled about between properdin and thrombospondin, and perhaps other proteins. On the other hand, if it turns out that the protozoan and mammalian proteins are truly unrelated, this case would have to rate as the nearest thing to sequence convergence yet reported.

Sequence convergence re-assessed

Certain rudiments of protein structure tend to be formed from subsets of amino acids: for instance, membrane-spanning stretches from nonpolar residues, and turns and loops from polar ones. Furthermore, the essence of some fundamental units depends on the sequence: the amphipathic helix tends to an NPPPNPP rhythm and the β-sheet to an NPNPNP pattern (where N and P stand for nonpolar and polar, respectively). Should we expect that the need for an α-helix, for example, would lead to similarities that are more than would be expected by chance and could properly be called convergent? Intuitively, one might think that the need for these structural components could lead to a sufficient number of unidirectional amino acid replacements such that sequence convergence would be evident.

Again, experience argues that sequence convergence is not a major force. Many three-dimensional structures are known that have such complicated but similar folds that they must have descended from a common ancestor, and yet they have no sequence resemblance at all, even though their major secondary structure elements remain in place. By contrast, there are no known examples in which unrelated proteins have sufficiently similar (and sufficiently extensive) sequences to warrant the descriptor 'sequence convergence'.

Having said that, I'd like to qualify my position slightly. Every sequence searcher knows that low-stringency sequence searches of a sample query sequence against a data bank will often identify candidates that have similar secondary structures in part. Many proteins with amphipathic helices will have low-level matches with nematode myosin, for example, since that very long and predominantly helical protein has many different sequence combinations contributing to its helical segments[29]. The issue, however, is whether such matches will stand up to statistical scrutiny upon more sophisticated analysis, and whether the ancestry of the proteins is confounded. I don't know of any cases where either of these is true.

In summary, the availability of 20 amino acids allows many different solutions to the same biochemical problem. Particular amino acids are drawn upon in many different circumstances to provide special properties, including the assembly of catalytic units. Individual adaptive replacements must and do occur. On the other hand, it is common for similar three-dimensional structures to be formed from totally dissimilar sequences. Many of the best examples involve common ancestry, but some three-dimensional barrel folds are so simple and common that they may have evolved independently. By contrast, solving structural problems by generating long strings of amino acids with similar sequences, which is what I think is implied by sequence convergence, has yet to be observed.

References

1 Rawlings, N. D. and Barrett, A. J. (1993) *Biochem. J.* 290, 205–218

2 Ollis, D. L. *et al.* (1992) *Prot. Eng.* 5, 197–211

3 Frick, I-M. *et al.* (1992) *Proc. Natl Acad. Sci. USA* 89, 8532–8536

4 Kraut, J. (1977) *Annu. Rev. Biochem.* 46, 331–358

5 Schrag, J. D., Li, Y., Wu, S. and Cygler, M. (1991) *Nature* 351, 761–764

6 Tai, M-H., Chirala, S. S. and Wakil, S. J. (1993) *Proc. Natl Acad. Sci. USA* 90, 1852–1856

7 Branden, C-I. (1991) *Curr. Opin. Struct. Biol.* 1, 978–983

8 Farber, G. K. (1993) *Curr. Opin. Struct. Biol.* 3, 409–412

9 Wilmanns, M. *et al.* (1991) *Biochemistry* 30, 9161–9169

10 Leahy, D. J., Hendrickson, W. A., Ikramuddin, A. and Erickson, H. P. (1992) *Science* 258, 987–991

11 Schulz, G. E. (1992) *Curr. Opin. Struct. Biol.* 2, 61–67

12 Knighton, D. R. *et al.* (1991) *Science* 253, 407–414

13 Doolittle, R. F. and Blomback, B. (1964) *Acta Chem. Scand.* 17, 1816–1819

14 Mross, G. A. and Doolittle, R. F. (1967) *Arch. Biochem. Biophys.* 122, 674–684

15 Fitch, W. M. and Margoliash, E. (1967) *Science* 155, 279–284

16 Holmes, E. C. *et al.* (1992) *Proc. Natl Acad. Sci. USA* 89, 4835–4839

17 Yokoyama, R. and Yokoyama, S. (1990) *Proc. Natl Acad. Sci. USA* 87, 9315–9318

18 Bauchop, T. and Martucci, R. W. (1968) *Nature* 161, 698–700

19 Dobson, D. E., Prager, E. M. and Wilson, A. C. (1984) *J. Biol. Chem.* 259, 11607–11616

20 Stewart, C-B., Schilling, J. W. and Wilson, A. C. (1987) *Nature* 330, 401–404

21 Stewart, C-B. and Wilson, A. C. (1987) *Cold Spring Harbor Symp. Quant. Biol.* 52, 891–899

22 Stewart, C-B. (1993) *Current Biol.* 3, 158–160

23 Beintema, J. J. (1990) *Mol. Biol. Evol.* 7, 470–477

24 Damian, R. T. (1964) *Am. Nat.* 98, 129–149

25 Damian, R. T. (1987) *Parasitol. Today* 3, 263–266

26 Cordiali, P. *et al.* (1992) *AIDS Res. Hum. Retrovirus* 8, 1561–1565

27 Robson, K. J. H. *et al.* (1988) *Nature* 335, 79–82

28 Goundis, D. and Reid, K. B. M. (1988) *Nature* 335, 82–85

29 Doolittle, R. F. (1987) *Of URFs and ORFs A Primer On How to Analyze Derived Amino Acid Sequences*, University Science Books, Mill Valley, CA

30 Feng, D-F. and Doolittle, R. F. (1990) *Meth. Enzymol.* 183, 375–387

TIBS reference lists

Authors of *TIBS* articles are asked to limit the number of references cited to provide non-specialist readers with a concise list for further reading. It is hoped that the citation of other, more extensive review articles rather than a comprehensive list of original articles enables interested readers to delve more immediately into the topic.