Homology

a personal view on some of the problems

There are many problems relating to defining the terminology used to describe various biological relationships and getting agreement on which definitions are best. Here, I examine 15 terminological problems, all of which are current, and all of which relate to the usage of homology and its associated terms. I suggest a set of definitions that are intended to be totally consistent among themselves and also as consistent as possible with most current usage.

I have frequently been asked about many controversial issues concerning the usage of homology and related terms. I examine some of these below as a set of 15 problems. This is my opinion on how best to maximize clarity on the use of these concepts with as little pain to alternative views as possible. Part of that clarity lies in making sure the definitions are self consistent. There are many alternative definitions for most of these terms and it might seem that we don't need another paper discussing this. But there is so much anguish about best usage, and so much misusage by new investigators to the field, especially by molecular biologists, mathematicians and bioinformatics people, that, if this could help investigators express themselves more clearly and get others to examine their own definitions and keep them within some bounds, it will be worth the effort. I have avoided phrases like 'I would suggest' and 'in my opinion' to save space. Insert them liberally wherever the text seems too dogmatic. Although the examples are largely molecular, the intent is to be as universal as possible, and a glossary is listed in Box 1. This article is an invited follow up on the excellent paper here in 1997 (Ref. 1). Other good discussion of most of these topics can be found in Refs 2 and 3. For a comparison between molecules and morphology, see Ref. 4.

Homology is the relationship of two characters that have descended, usually with divergence, from a common ancestral character. This is important because most of the terminological problems stem from different definitions of homology. Characters can be any genic, structural or behavioral feature of an organism. Analogy is distinguished from homology in that its characters, although similar, have descended convergently from unrelated ancestral characters. The cenancestor is the most recent common ancestor of the taxa being considered[5].

The other homologies problem

Organic chemists consider compounds such as methane, ethane and propane to be an homologous series because each differs from the next by a CH_2 group. Thus, homoserine has one more CH_2 group than serine. Mathematicians have special meanings for the term as well. There is no point in worrying about these differences, except to suggest that molecular biologists, mathematicians and bioinformaticists working in the field of biology learn and adhere to the biological definitions.

The redefinition problem

Homology was first defined in biology with something like its present meaning by Owen in 1843 who characterized homology as 'the same organ under every variety of form and function'[6]. Common ancestry is not mentioned in that definition, which is unsurprising given that these were pre-Darwinian and pre-Mendelian times. Owen's definition of homology emphasizes structure and location rather than ancestry. Some would have us return to Owen's definition, perhaps out of a sense of precedence or some perceived need for unchanging meaning. But that would mean inventing a word to designate common ancestry. The meaning of a word should change if that change is a refinement that increases clarity of present-day thought and exposition, as it does here.

The character/character-state problem

Many systematists, and nearly all molecular evolutionists, distinguish between a character, say amino acid, and its character states, say glycine and phenylalanine. This useful distinction is not universal. Many systematists will, if two character states are not the same, assert that the characters are non-homologous! This is confusing because it implies that the two characters do not have a common ancestor, which, if true, means they should not have been comparing the character states in the first place. Homology resides in the characters, not in their states!

The homology/homoplasy problem

Analogy describes characters whose similarity arises from convergent processes. Homology describes characters, irrespective of their character states, whose similarity arises after divergence from a common ancestral form. Homoplasy is the complement of analogy in that these two categories constitute all known non-random explanations of similarity. Some would make homoplasy, a term introduced by Lankester[7], the complement of homology. But homoplasy is a relation of two character states in a tree, whereas analogy is a relation of two characters, independent of any tree, making homoplasy noncomplementary to homology.

The recognition of homology problem

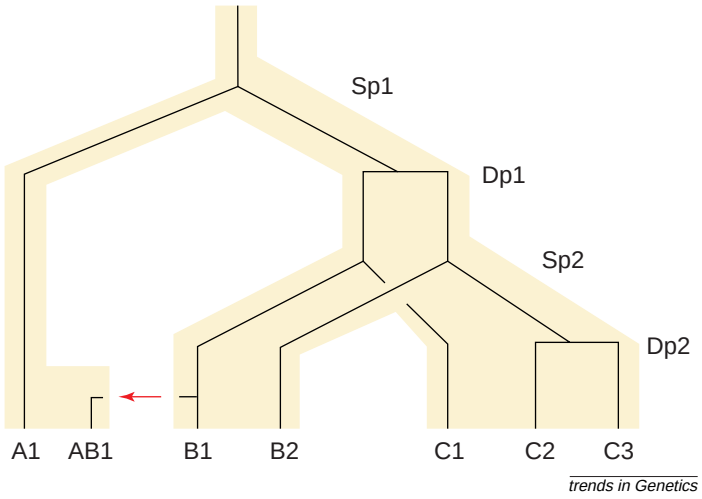
How does one know for sure that two sequences are homologous? One would always 'know' if we defined homology objectively as their having amino acids or nucleotides

Walter M. Fitch
wfitch@uci.edu

Department of Ecology and Evolutionary Biology, University of California, 321 Steinhaus Hall, Irvine, CA 92697, USA.

0168-9525/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(00)02005-9

FIGURE 1. Orthology, paralogy and xenology



The idealized evolution of a gene (lines) is shown from a common ancestor in an ancestral population (the gray background), descending to three populations labelled A, B and C. There are two speciation events (Sp1 and Sp2), each occurring at the junctions shown as an upside down Y. There are also two gene-duplication events (Dp1 and Dp2), depicted by a horizontal bar. Two genes whose common ancestor resides at a Y junction (speciation) are orthologous. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogous. Thus, C2 and C3 are paralogous to each other but are orthologous to B2. Both are paralogous to B1 but orthologous to A1. The red arrow denotes the transfer of the B1 gene from species B to species A. As a result, the AB1 gene is xenologous to all six other genes. All three subtype relationships are reflexive, that is, A1=>B1 implies B1=>A1 where => should be read, for example, as 'is orthologous to.' However, the relationships are not transitive. Thus, C2=>A1=>C3 might be true, but it is not necessarily therefore true that C2=>C3, as indeed it is not in the figure if => is read as 'is orthologous to.' A different non-transitivity occurs for 'is paralogous to' with B2=>C1=>C2.

that are at least X% identical. But what is the appropriate value of X? One should not define homology objectively because: (a) it requires defining homology by an arbitrary amount of identity; (b) it excludes the possibility of analogy; and (c) this still does not solve the problem of our confidence that the characters asserted to be homologous do have a common ancestor. Homology is here an abstraction in that it is a relationship, common ancestry, the nature which we find important to know about, but which we can only infer with more or less certainty.

It is worth repeating here that homology, like pregnancy, is indivisible[8]. You either are homologous (pregnant) or you are not. Thus, if what one means to assert is that 80% of the character states are identical one should speak of 80% identity, and not 80% homology.

The homology subset problems

There are three disjoint subtypes of homology. Orthology is that relationship where sequence divergence follows speciation, that is, where the common ancestor of the two genes lies in the cenancestor of the taxa from which the two sequences were obtained[9]. This gives rise to a set of sequences whose true phylogeny is exactly the same as the true phylogeny of the organisms from which the sequences were obtained. Only orthologous sequences have this property.

Paralogy is defined as that condition where sequence divergence follows gene duplication[9]. Such genes might descend and diverge while existing side by side in the same lineage. Mixing paralogous with orthologous sequences can lead to a tree that has the correct phylogeny for the sequences but not for the taxa from which they derive; a gene tree is not necessarily a species tree.

Xenology is defined as that condition (horizontal transfer) where the history of the gene involves an interspecies transfer of genetic material[12]. It does not include transfer between organelles and the nucleus. It is the only form of homology in which the history has an episode where the descent is not from parent to offspring but, rather, from one organism to another. Unrecognized xenology has the greatest negative impact causing bizarre taxon phylogenies; however, it is that very bizarreness that alerts us to recent xenology. The acquisition of chloroplasts by a eukaryote was a xenologous (in this case, symbiotic) event and, if one constructs trees that mix chloroplast genes with nuclear and prokaryotic homologs, the result will often be a bizarre sister-group relation between plants and cyanobacteria. Nevertheless, all unduplicated chloroplast genes are, presumably, orthologous within the plants, even those that have been relocated into the nucleus. Gogarten has proposed a special term, synology, for those xenologs that arise, not by the transfer of a gene between two species, but by a hybridization of two species[12]. One might then question, given a successful hybrid, whether the two species are not effectively one and this is simply a case of reassortment among alleles at a locus. The subtype relationships, ortholog, paralog and xenolog (illustrated in Fig. 1), should be used whenever that relation is known or assumed, and the term homolog should be reserved for those cases where (a) homology can be inferred but not the subtype, or (b) the assertion is correct for all subtypes.

If there is more than one ortholog, which one is 'correct'? There is a tendency to wish that there could be only one ortholog in an organism. This is frequently not the case. Figure 1 shows a gene tree. The A1 gene has three orthologs in species C. The nature of the subtype relationship depends solely on whether the cenancestral sequence occurs at a speciation or a duplication event. Consider, for example, α and γ hemoglobin from a human (say C1 and C2, respectively, in Fig. 1) and the α of frog (say A1). The two human sequences are paralogous to each other but both are orthologous to the frog α hemoglobin. But what if we replaced the latter with the orangutan α (B1) sequence, thereby reversing the order of the duplication and speciation events? Now, although the three human sequences remain paralogous among themselves, only the human α hemoglobin is orthologous to the orangutan α hemoglobin. Putting all four sequences in the tree, or adding others, cannot change any of those relationships, only our ability to detect them. Note that you get the correct species tree for these sequences if you use A1 plus B1 and C1 or if you use A1 plus B2 and C2 and/or C3. Thus, there can be more than one ortholog and all are correct.

There is another side to this question however. Sometimes one wishes, in the face of multiple genic orthologs, to designate pairs that carry out the same function. For example, the α hemoglobin sequence duplicated at the base of the mammals affords an opportunity for the development of a fetal form, thereby providing the unborn offspring with the ability to extract oxygen from the maternal blood supply. Both mammalian hemoglobins are orthologs of the bird α hemoglobin, but only one is used like the bird α hemoglobin as the adult transporter of oxygen. It is called α hemoglobin while the fetal form is called γ hemoglobin. It would be appropriate, whenever a pair of

BOX 1. Glossary

Analogy
The relationship of any two characters that have descended convergently from unrelated ancestors.

Cenancestor
The most recent common ancestor of the taxa under consideration.

Characters
Any genic, structural or behavioral feature of an organism having at least two forms of the feature called character states, for example: metatarsals, separate (crocodiles) or fused (birds); feather color, red (cardinals) or blue (blue jays); nucleotide, A, C, G or T.

Gene conversion
The replacing of a block of DNA from one gene with the homologous residues in its paralog.

Homology
The relationship of any two characters that have descended, usually with divergence, from a common ancestral character.

Homoplasy
The relationship of any two identical character states that must have arisen independently, given a specific phylogenetic tree.

Indel
A gap in a sequence alignment introduced to account for an insertion or deletion in one or more genes.

Orthology
The relationship of any two homologous characters whose common ancestor lies in the cenancestor of the taxa from which the two sequences were obtained.

Paralogy
The relationship of any two homologous characters arising from a duplication of the gene for that character.

Xenology
The relationship of any two homologous characters whose history, since their common ancestor, involves an interspecies (horizontal) transfer of the genetic material for at least one of those characters.

paralogous genes are both orthologous to a more distant gene – and they diverge such than one form retains the old function while the other acquires a new function – to label the pair of orthologs retaining the same function as isorthologs (from iso, meaning same). Holland has suggested other relationships among paralogs[13].

The gene loss problem

Imagine that a gene duplicated and then, following a subsequent speciation, one lineage lost one gene and the second lineage lost the other gene. What does one call the relationship of the remaining two genes? Paralogy, of course. The definition of the forms of homology does not change by virtue of the known, suspected, or unknown presence of a copy of a gene. This brings up the related problem, can a gap be a homolog? Yes. The alignment of molecules often needs indels (gaps required because there was either an insertion in one sequence or a deletion in the other; see Box. 1). Remembering that characters have states, then one of those states could be 'deleted.' Present/absent might be good systematic character states.

The structure/function problem

The definition of homology is about characters. Examples include genic (molecular), structural (morphological), functional (metabolic, regulatory, and behavioral) characters, and no doubt others. Should all characters be admitted? There are many examples in different organisms where the same structure has different functions or different structures have the same function. A marvelous example is the reptilian articular and quadrate bones of the mandible, which are orthologous to the mammalian malleus and incus bones of the ear[15]. Confusion can occur if homology can apply both to structural and to functional characters. Nevertheless, I would raise no bar to the inclusion of all kinds of characters, provided one is careful to indicate whether the homology is genic, structural, functional or behavioral. The routine use of these adjectives would solve most of 'the problem of levels' in the use of homology raised by Dickinson and others[16].

The bird/bat limbs problem

Are their forelimbs homologous or not? The forelimbs of the bat and the bird are adapted to flight, but the evolution to flight occurred independently in each lineage. Their cenancestral limb is the forelimb of a flightless reptile that is itself the reptilian cenancestor of the birds and mammals. Thus the limbs are (structurally) orthologous. On the other hand, the flight of birds and bats is (functionally) analogous.

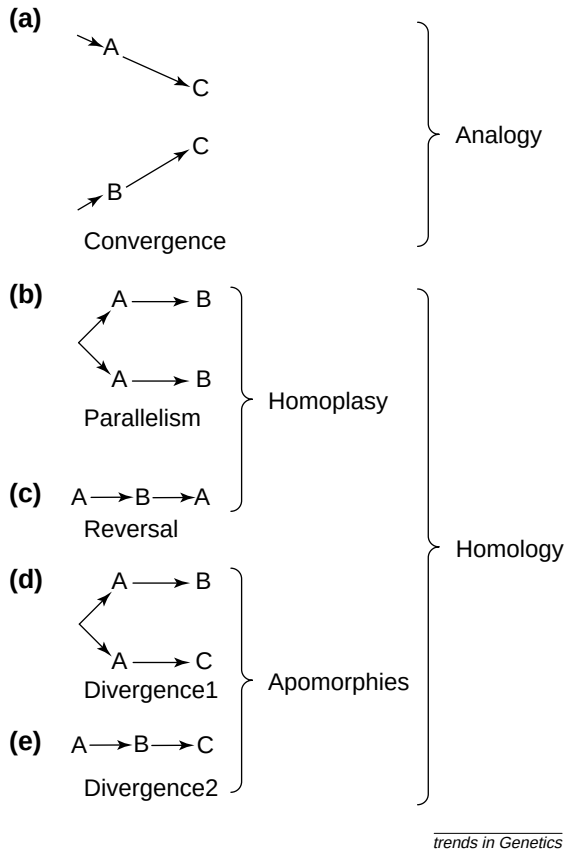
The parallel/convergence problem

Five possible relationships for two changes in a character are shown in Fig. 2. They are given the names that generally conform to the English meanings of the words. However, many use the term convergence for changes that are parallel. Calling convergent that which does not converge can only be another source of confusion and should be resisted.

The homology/analogy problem

There is a tendency to assume that if two characters are significantly similar they must be homologous. This assumption has been proven to be untrue many times when the characters were morphological or behavioral. For nucleotide and amino acid sequences, the situation is different. Most of the time, the degree of similarity is so great that one (including me) will say that convergence could not have caused this much similarity[17]. It is commonly believed that there is no test that can distinguish between homology and analogy. However, with a method that infers ancestral sequences, it is quite possible. For example, are α and β hemoglobin homologous or analogous? One compares the similarity of the ancestral α and β sequences to the similarity of the present-day sequences. If the ancestral sequences are significantly more alike than today's sequences, the genes are homologous. If the reverse is true, the sequences are analogous. In this case the test has been performed and the α and β hemoglobins are homologous[9]. The method is not circular; it does detect analogy when analogy is present.

FIGURE 2. Terminological relations



Five possible relationships (a–e) that two mutations might have, along with the common sense meanings that describe those relationships. The two right-hand columns show a disputed view of the higher relationships.

There are no proven cases of genic analogy. A most interesting case is that of the foregut lysozyme from colobines and artiodactyls. Messier and Stewart[18–20] found a significantly elevated ratio of the non-silent (amino acid replacing) to silent nucleotide substitutions in the single branch leading to the colobines, which they correctly interpreted as demonstrating positive selection. Moreover, they noticed that five of the nine amino acid replacements on that branch also occurred in the orthologous lysozyme in its descent to the ancestral artiodactyl. So what term(s) best describe the relationship between these two lysozymes? There have been five parallel mutations (by the definitions in Fig. 2) in genically orthologous sequences that, even as these parallelisms occur, are nevertheless diverging overall. There is no convergence (as these authors claim) and thus no analogy at sequence level. We do have convergence, and so analogy, at the functional level where both lysozymes independently adapted to processing foregut ferments in the stomach. Thus, we have genic orthology and functional analogy.

Whereas most genic similarities are homologies, some motifs could well be analogous. This might include, for example, the motif for vertebrate initiation of translation, RNNAUGG (Ref. 21). This might be particularly likely where a single message has two initiation sites. But, whereas one might demonstrate that there must be some analogous motifs, it seems insuperable deciding which are homologous and which analogous. Shimeld[22] has suggested that the motif PFSIXNXXS is convergent in the homeobox and fork head genes.

The gene conversion problem

Consider a gene duplication creating paralogs, followed some time later by a speciation event, and then by gene conversions in which copies of blocks of DNA from one gene simply replace the homologous residues in its paralog. This causes the paralogs to look more like each other. Indeed, they might look so much more alike than they otherwise should that, although they continue to look like paralogs within a species, they will appear to have duplicated recently and independently in each species, so that all comparisons between species will appear to be orthologous comparisons. Because the conversion process is destroying the evidence of the early duplication but not preventing the divergence of the two genes between species, the conclusion of orthology between species is, in fact, operationally correct and we are not misled about anything except the recentness of the gene duplication. This can only be detected by reference to the surrounding sequences. Such a case has been observed by Rudikoff *et al.*[23] in the mouse T-lymphocyte antigen receptor. In that case, recent gene conversions made exon-1 look as if it had duplicated recently and separately in each of three different mouse lineages when, in fact, there was only one duplication more than ten times earlier, before their cenancestor.

The recombination problem

Two sequences or domains might have a common ancestor, in which case they are homologous, irrespective of the degree of similarity. A gene can be constructed from the domains of several other different genes. For example, enterokinase has at least five domains in addition to the protease domain[24]. One domain is related to a low-density lipoprotein receptor, another to a metalloprotease of the renal glomerulus, another to the *Drosophila* dorsal–ventral patterning gene and yet another to lymphocyte cell-surface antigens. Wherever this occurs, the terms we are discussing do not apply to the whole gene. What we have is a series of domains, each of which is paralogous to a similar domain of a different gene. We must recognize that not all parts of a gene have the same history and thus, in such cases, that the gene is not the unit to which the terms orthology, paralogy, etcetera apply. In particular, if the domain that is homologous to the low-density lipoprotein receptor constitutes 20% of enterokinase, then enterokinase is only 20% homologous to that lipoprotein receptor, irrespective of its percent identity. If, at the same time, this common domain were half of the lipoprotein receptor, the receptor would be 50% homologous to the enterokinase. The homologies are not the same in both directions if the proteins are of unequal length! This is the only situation where 'percent homology' has a legitimate meaning and, even there, it is dangerous and better called, as Hillis has suggested, partial homology[25].

The tandem repetitive characters problem

Some DNA is composed of chunks that are tandemly repeated, sometimes many times. The chunks are necessarily paralogous but they do represent a special case. Normally, the presence of gaps in an alignment is the result of a simple insertion or deletion (indel) in a gene. In the case of repeat arrays, the gaps can be the result of there

being different numbers of the repeats. They should be treated differently from indels when building phylogenies but there is no nomenclatural problem; they are iterative, tandem or serial paralogs[10,11].

The gene/allele problem

Are two alleles paralogs? Presumably not, given that the definition includes gene duplication. But consider the following. The earliest vertebrates had only one hemoglobin gene and so there was no cooperativity that allowed the more efficient transport of oxygen. Then a mutation arose that permitted some slight but beneficial degree of heterozygous cooperativity. It would be selected for until perhaps the population had about equal frequencies of the two alleles. At this point, only half the members of the population could enjoy this benefit because, at most, half would be heterozygous under random mating. And then another mutation came along (this time a translocation or a duplication of one of the alleles) and the duplication spread through the population because 100 percent of the population might now be functionally 'heterozygous.' In the stroke of a translocation we have converted two alleles into two loci. Was it at this moment that the paralogous nature of these two genic entities was created? I think the answer should be yes but there is now some blurring of the importance of differentiating between alleles and genes.

Conclusion

I recognize, and even accept, that homology has been used by various people with different meanings, even though similarity was a common denominator among these meanings. The two most important of these meanings related homology to similar structures and/or to similar functions. (By structures I mean both molecular sequences and morphology.) Life would have been simple had phylogenetic homology necessarily implied structural homology or either of them necessarily implied functional homology. However, they map onto each other imperfectly and my definition of homology includes all forms of characters. We could reduce confusion by always indicating the kind of homology we are referring to when using the term.

I have covered as many problems as I could in this brief exhortation so that there would be a comprehensive, consistent set of terms and meanings, with the idea that this comprehensiveness would be an argument for using these terms or something closely similar and that any proposed replacement of these terms be at least as encompassing and consistent. By following clear definitions, many of the problems people have raised are simply usage problems and the terms used to describe different kinds of homology can, when used strictly, get across the appropriate, specific meanings involved.

References

1 Abouheif, E. *et al.* (1997) Homology and developmental genes. *Trends Genet.* 13, 432–433
2 Hall, B.K., ed. (1995) *Homology, the hierarchical basis of comparative biology*, Academic Press
3 Bock, G.R. and Cardew, G., eds (1999) *Homology*, John Wiley & Sons
4 Patterson, C. (1988) Homology in classical and molecular biology. *Mol. Biol. Evol.* 5, 603–625
5 Fitch, W.M. and Upper, K. (1988) The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harbor Symp. on Quant. Biol.* LII, 759–767
6 R. Owen (1843) *Lectures on the comparative anatomy and physiology of the invertebrate animals.* Longman, Brown, Green & Longmans
7 Lankester, E.R. (1870) On the use of the term homology in modern zoology. *Ann. Mag. Nat. Hist. Ser.* 6, 34–43
8 Reeck, G.R. *et al.* (1987) "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell* 50, 667
9 Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113
10 Ghiselin, M.T. (1969) The distinction between similarity and homology. *Syst. Zool.* 18, 148–149
11 Owen, R. (1848) On the archtype and homologies of the vertebrate skeleton, Van Voorst
12 Gray, G. and Fitch, W.M. (1983) Evolution of antibiotic resistance genes: The DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. *Mol. Biol. Evol.* 1, 57–66
13 Gogarten, J.P. (1994) Which is the most conserved group of proteins? Homology-Orthology, paralogy, xenology and the fusion of independent lineages. *J. Mol. Evol.* 39, 541–543
14 Holland, P.W.H. (1999) The effect of gene duplication on homology. In *Homology* (Bock, G.R. and Cardew, G., eds), pp. 226–242, John Wiley & Sons
15 Strickberger, M.W. (1996) *Evolution* (2nd edn), Jones and Bartlett
16 Dickinson, W.J. (1995) Molecules and morphology: Where's the homology? *Trends Genet.* 11, 119–121
17 Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453
18 Stewart, C-B. and Wilson, A.C. (1987) Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters. *Cold Spring Harbor Symp. Quant. Biol.* 52, 891–899
19 Swanson, J.W. *et al.* (1991) Stomach lysozyme gene of the langur monkey: Tests for convergence and positive selection. *J. Mol. Evol.* 33, 418–425
20 Messier, W. and Stewart, C-B. (1997) Episodic adaptive evolution of primate lysosymes. *Nature*, 385, 151–154
21 Kozak, M. (1991) An analysis of vertebrate mRNA sequences: intimations of translational control *J. Cell Biol.* 115, 887–903
22 Shimeld, S.M. (1997) A transcription modification motif encoded by homeobox and fork head genes. *FEBS Lett.* 410, 124–125
23 Rudikoff, S. *et al.* (1992) Exon-specific gene correction (conversion) during short evolutionary periods: homogenization in a two-gene family encoding the β chain constant region of the T-lymphocyte antigen receptor. *Mol. Biol. Evol.* 9, 14–26
24 Kitamoto, Y. (1994) Enterokinase, the initiator of intestinal digestion, is a mosaic protease composed of a distinctive assortment of domains. *Proc. Natl. Acad. Sci. U. S. A.* 91, 7588–7592
25 Hillis, D.M. (1994) Homology in molecular biology. In *Homology, the hierarchical basis of comparative biology* (Hall, B.K., ed.), pp. 339–368, Academic Press

Virtual biology in the CAVE

Like any emergent, self-organizing process, embryonic development rapidly proceeds from the seemingly simple (dividing cells, epithelial sheets) to the extremely complex (formation of germ layers, gastrulation, organogenesis). For example, the *Drosophila* embryo develops from a single-cell layered blastoderm that is transformed into a complex, folded and layered gastrula in approximately two hours. The development of complex structures, such as those observed during embryonic development, occurs at all levels of biological organization, including subcellular organellar assembly (e.g. assembly of the Golgi, endoplasmic reticulum), outer nuclear envelope assembly and disassembly, formation of epithelial sheet specializations during organogenesis, trabeculae formation during lung development, and blood vessel branching and anastomosis, to name just a few. Our most recent knowledge of developmental processes comes from genetic and cellular analyses that have revolutionized our understanding of basic developmental processes at the molecular level. As such,

0168-9525/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(00)01996-X