Molecules and morphology: where's the homology?

W.J. DICKINSON

DEPARTMENT OF BIOLOGY, UNIVERSITY OF UTAH, SALT LAKE CITY, UT 84112, USA.

A few years ago[1], molecular biologists were chastised for sloppy and confusing use of the term 'homology'. Many treated homology as an objective observation rather than an inference, and as a quantitative trait ('percentage homology') rather than a relationship of common evolutionary origin that either does or does not exist (see description of terminology in Box 1). There is another source of confusion that threatens to become increasingly troublesome as the fascinating molecular homologies that lie at the heart of developmental mechanisms are unraveled: there is not necessarily a simple relationship between homology of molecules (or even pathways) and homology of the anatomical features in whose development those components participate. In other words, some recent suggestions notwithstanding[2–7], molecular similarities in the developmental mechanisms that produce specific organs are not, by themselves, strong evidence for homology of those organs.

Levels of homology

The central point of this article is that questions of homology can be examined at multiple levels and that homology between a pair of structures can simultaneously be present at some levels but absent at others. The term 'levels of homology' refers to a nested series of progressively more ancient and inclusive ('deeper') relationships. The classic textbook example of homology, the vertebrate forelimb, conveniently illustrates the point. Considered only as forelimbs, the wings of birds and bats are homologous; considered as wings, they are not. In other words, the last common ancestor of these two groups had forelimbs but not wings. Note that this conclusion is partly based on evidence other than that derived from the direct comparisons of wings: the comparative anatomy of other vertebrate forelimbs; the fossil record; and other anatomical comparisons that reveal, for example, the relationship of bats to other mammals have also been considered (at least implicitly). If doubt remained, sequence data could be collected to confirm that bats descend from wingless mammals.

Now, suppose the molecular mechanisms controlling development in birds and bats are examined. Given the known conservation of mechanisms in vertebrates, homologous molecules and conserved pathways would certainly be found operating in the development of wings in both groups. However, such similarities would not be interpreted as supporting the 'surprising' conclusion that the two sorts of wings are, after all, homologous. Instead, the molecular similarities would be recognized as reflecting homology at a deeper level (forelimbs). In other studies, the danger arises when evidence bearing on homology is less extensive or decisive (or less well known to the average molecular or developmental biologist) than in this example.

Interpreting homeobox gene comparisons

Turning to real molecular examples, the evolving interpretation of comparisons between homeobox genes and clusters in different organisms is instructive. When these were discovered in vertebrates following their initial characterization in insects, early speculations centered on the possibility that insects and vertebrates share a conserved mechanism of segmentation, even though this contradicted the conventional view that the last common ancestor of arthropods and vertebrates was not segmented. However, the discovery of homeobox clusters in unsegmented creatures like *Caenorhabditis elegans* undermined these speculations, particularly since analyses of expression patterns in the worm confirmed that there is no relationship between homeobox genes and reiterated cell lineages that might be regarded as a primitive form of segmentation[8,9]. Homeobox genes are involved in other divergent processes such as limb development in vertebrates[10] and gut differentiation in insects[11]. Thus, the focus of interpreting homeobox gene function shifted progressively from segmentation to anterior–posterior polarity[12] and to axial patterning in general[13]. Homology at even deeper levels, such as positional information *per se* or simply transcriptional regulation, may be most relevant to some homeobox gene comparisons.

A cautious initial interpretation of similarities among insects and vertebrates would have considered all of these possibilities and recognized the need for additional information to distinguish between them. As in the

Box 1. Addendum on terminology

The terminology used in this article to describe relationships is that proposed by Fitch[16], elaborated by Patterson[17], and summarized in the instructions to authors writing for *Molecular Biology and Evolution*. Briefly, features (including molecules) that are similar by virtue of common ancestry are homologous, while those that are similar by convergence are analogous. Among homologous molecules, those produced by gene duplication are paralogous and those separated by speciation are orthologous. It is possible (and useful), as Patterson suggests, to give precise definitions even when there are substantial practical difficulties in deciding which relationship applies in particular cases. There is, however, one problem of definition not dealt with in the cited sources. When duplication produces a paralogous gene set in one species, is the orthologous relationship to homologs in other lineages retained by both, one, or neither of the copies? If it is retained by only one copy, to which copy should the orthologous relationship be assigned? This difficulty does not need to be resolved for the present purpose but it further highlights the complexities of using molecular similarities as evidence for anatomical homology.

example of bird and bat wings mentioned above, more detailed analyses of the relevant systems would not, in isolation, have resolved the question. The progressive interpretation summarized above depended on information about additional species (e.g. *C. elegans*) and other contexts of expression within species (e.g. limbs and guts); in turn, those comparisons depend, at least implicitly, on additional data of various kinds (such as that relevant to phylogeny).

Homologous molecules in analogous organs

A second example highlights the classic problem of convergence, with the deceptive twist that truly homologous molecules may be involved in processes that are only analogous. Products of the *hedgehog* gene in *Drosophila* and of an avian homolog serve strikingly similar functions in wing development[14]. Quite properly, their roles in that context are recognized as analogous, not homologous. Again, *hedgehog* homologs play comparable roles in intercellular signaling in various other developmental contexts in both insects and vertebrates. Undoubtedly, there is deep and interesting homology here but the wing is not the level at which it should be sought.

The probability of encountering such convergence is greatly increased by three well-established features of molecular evolution: (1) even within a single species the same molecule can assume functions in quite different developmental pathways; (2) gene duplication generates paralogous gene families whose members can encompass an even wider range of roles; (3) domain shuffling generates molecules with clear homology in some regions but potentially with quite different overall functions. All of these can be subsumed within the idea of 'levels of homology' adopted here. Molecules with multiple functions could presumably be traced back to some primordial function while paralogous sequences, whether entire molecules or domains, could be traced to a molecule (and function) that existed before some relevant duplication occurred. In either case, the molecular biologist's job (not necessarily simple) would be to determine the context in which an ancestral molecule functioned at the point where paths merge when traced backwards from the current examples under consideration. Such an analysis would identify the level at which the contemporary functions and contexts could usefully be said to be homologous.

Some questionable cases

Molecular similarities have sometimes not been interpreted in an appropriately cautious manner. Based on comparisons of function and expression of the *orthodenticle* gene in *Drosophila* and of homologs in vertebrates, Finkelstein and Boncinelli[2] suggest that, contrary to prevailing opinion, head specialization may have occurred before the ancestral lineages separated. However, the facts permit hypotheses similar to those proposed for interpreting analyses of homeobox genes mentioned above: these *orthodenticle* homologs could be deeply conserved components involved in axial patterning (or another aspect of positional information) not specifically related to cephalization.

Defects caused by *eyeless* in *Drosophila* and a homolog, *Small eye*, in mice have prompted speculation that arthropod and vertebrate eyes are homologous despite fundamental differences in organization[5,7]. This situation may be comparable to that of the *hedgehog* gene in wing development. The roles of these genes in eye development should be termed homologous only if other evidence suggests that an orthologous antecedent of both *eyeless* and *Small eye* functioned in the development of an eye in a common ancestor of arthropods and vertebrates.

Kispert *et al.*[3] suggest homology between the vertebrate notochord and the insect hindgut because the *Brachyury* (*T*) gene and a *T*-related gene (*Trg*), respectively, are required for normal development of those organs. In this case, the molecular homology is confined to a DNA-binding region. This region could have combined with other domains to generate molecules with distinct functions either before or after the separation of vertebrate and arthropod lineages.

Finally, homologous transcription factors seem to play similar roles in regulating some genes in the liver of mammals and the fat body of *Drosophila*, leading to speculations about the homology of these organs[6,15]. As with other developmental regulators, these factors belong to a limited number of families and typically function in a variety of contexts. Again, coincidental similarities between analogous systems are to be expected. This case is also confused by the seemingly interchangeable use of the terms 'homology' and 'analogy' in the discussion.

Laufer and Marigo[4] summarize additional examples in which connections between molecular and anatomical homology have been considered. The issues raised in this article have not always been given adequate attention. It is noteworthy that the majority of 'surprising' anatomical homologies thus far proposed on the basis of molecular data involve comparisons between insects and vertebrates. This certainly reflects the intense effort devoted to molecular analyses of development in these particular systems. As other groups receive more attention, the incidence of convergent examples will surely increase, reinforcing the importance of caution and precision in the interpretation of molecular similarities.

Conclusions

In no case am I arguing that suggested inferences about organ-level homology are definitely wrong; I claim only that the molecular evidence alone is weak and that some authors have been vague or ambiguous with respect to the level of homology suggested. It must be recognized that molecular similarities could reflect homology at any of several levels, that other data must be evaluated to decide which level is most likely in a particular case, and that the level under discussion must be carefully specified in reports of hypotheses and conclusions. Anatomical homology will become a useless concept if it is inferred in all organs in which homologous molecules are found to have similar functions.

Acknowledgements

A stimulating discussion with D. Feener and participants in the Biology 601 seminar focused my thoughts and prompted the writing of this article. J. Seger and G. Herrick made useful comments.

References

1 Reeck, G.R. *et al.* (1987) *Cell* 50, 667
2 Finkelstein, R. and Boncinelli, E. (1994) *Trends Genet.* 10, 310–315

3 Kispert, A., Herrmann, B.G., Leptin, M. and Reuter, R. (1994) *Genes Dev.* 8, 2137–2150
4 Laufer, E. and Marigo, V. (1994) *Trends Genet.* 10, 261–263
5 Quiring, R. *et al.* (1994) *Science* 265, 785–789
6 Sondergaard, L. (1993) *Trends Genet.* 9, 193
7 Zuker, C.S. (1994) *Science* 265, 742–743
8 Kenyon, C. and Wang, B. (1991) *Science* 253, 516–517
9 Salser, S.J. and Kenyon, C. (1994) *Trends Genet.* 10, 159–164
10 Oliver, G. *et al.* (1998) *Cell* 55, 1017–1024
11 Bienz, M. (1994) *Trends Genet.* 19, 22–26
12 Akam, M. (1989) *Cell* 57, 347–349
13 McGinnis, W. and Krumlauf, R. (1992) *Cell* 68, 283–302
14 Ingham, P.W. (1994) *Curr. Biol.* 4, 347–350
15 Abel, T., Bhatt, R. and Maniatis, T. (1992) *Genes Dev.* 6, 466–480
16 Fitch, W.M. (1970) *Syst. Zool.* 19, 99–113
17 Patterson, C. (1988) *Mol. Biol. Evol.* 5, 603–625

Gatecrashers at the catalytic party

W. MARSHALL STARK AND MARTIN R. BOOCOCK

LABORATORY OF GENETICS, INSTITUTE OF BIOMEDICAL AND LIFE SCIENCES, UNIVERSITY OF GLASGOW, CHURCH STREET, GLASGOW, UK G11 5JS.

DNA transposition and site-specific recombination reactions are mediated by assemblies containing several identical protein subunits. How can we establish which subunits are doing what, and where? Attempts to answer these questions in the field of site-specific recombination have come up with surprising and provocative answers.

Catalysis of DNA breakage and rejoining in site-specific recombination is achieved by an assembly of four recombinase subunits. Two subunits bind at a specific 'core' region of about 30 bp in each of the two recombining sites; these core regions usually have identical sequences. There are at least two different ways of exchanging the DNA strands to make recombinants (Fig. 1a,b)[1].

One large family of related recombinases (exemplified by Tn*3* and γδ resolvases) catalyses the breakage ('cleavage') of all four strands before concertedly swapping the ends and rejoining them (Fig. 1a). Cleavage occurs when a DNA phosphodiester bond is attacked by a hydroxyl group on a Ser sidechain from the recombinase, forming a transient protein–DNA covalent linkage. Other recombinases break, exchange and rejoin one pair of homologous strands to make a four-way junction intermediate (Holliday junction); they then complete recombination by breaking, exchanging and rejoining the other pair of strands (Fig. 1b). These enzymes use a Tyr sidechain to provide the hydroxyl nucleophile that cleaves the DNA. This group includes phage λ integrase (Int) and the yeast 2μ plasmid recombinase FLP, and was thought to comprise a second 'Int family' of related recombinases. However, the relatedness of some members of the group has now been questioned.

Recombinases may use a shared active site

Given the above information, a simple (but naive) question can be asked. There are four recombinase subunits, and four bonds in the recombining sites that must be broken. Which subunits catalyse which cleavage reaction (see Fig. 1c)? The question should actually be asked much more carefully because cleavage and rejoining at a particular phosphate might require the intervention of more than one subunit, and a particular subunit might interact with more than one segment of the DNA in the complex. Nevertheless, the primary binding site for a recombinase subunit can be defined using *in vitro* assays such as footprinting, and the functions of specific amino acid residues can also be characterized by, for example, analysing mutant proteins. Therefore, experiments can potentially be devised to establish connections between (1) a DNA bond that is broken, (2) a particular catalytic residue, (3) the subunit of which it is part, and (4) a particular binding site in the reactive complex.

The nucleophilic residue (Tyr or Ser) of the recombinase is an obvious first choice for this sort of analysis, particularly because intermediates with the DNA covalently linked to the recombinase can be trapped *in vitro*[1]. In a thought experiment, one could covalently link an active subunit to its binding site with a chemical crosslinking agent. In the cleaved intermediate, this subunit would be attached to the DNA at the cleaved phosphate and also at its binding site, and the connectivity could be established by conventional electrophoresis techniques. However, the development of efficient methods for site-specific protein–DNA crosslinking is at an early stage. The published work to date has used more indirect methods.

Much interest in these questions has been generated by complementation experiments with the yeast recombinase FLP[2]. Four amino acid residues (two Arg residues, a His, and the nucleophile Tyr) are conserved through the whole Int family of enzymes. Mutation of any of these residues in FLP abolishes recombination activity. However, in a suicide substrate assay (see below), the mutant Y343F, in which Tyr had been changed to Phe, was inactive on its own but could be complemented by FLP mutant in any of the other three residues[3]. This led to the idea of a 'shared active site' in which the Tyr343 nucleophile was provided by one subunit and a 'triad' (Arg-His-Arg) of activating residues by a second subunit (Fig. 1d).

Suicide substrates[4] have been central to the analysis of intermediates in Int family reactions. Many variations on the theme have been used but the principle is the same; the reversal of the DNA cleavage reaction is inhibited because the leaving group