Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W121–W125***
doi:10.1093/nar/gki373

PhyloDome—visualization of taxonomic distributions of domains occurring in eukaryote protein sequence sets

Maria Novatchkova[1,2,*], Michael Wildpaner[2], Dieter Schweizer[1] and Frank Eisenhaber[2]

[1]Gregor-Mendel-Institute (GMI) of the Austrian Academy of Sciences, Dr Ignaz Seipel-Platz 2, A-1010 Vienna, Austria and [2]Research Institute of Molecular Pathology (IMP), Dr Bohr-Gasse 7, A-1030 Vienna, Austria

Received February 9, 2005; Revised and Accepted March 7, 2005

ABSTRACT

The analysis of taxonomic distribution and lineage-specific variation of domains and domain combinations is an important step in the assessment of their functional roles and potential interoperability. In the study of eukaryote sequence sets with many multi-domain proteins, it can become laborious to evaluate the phylogenetic context of the many occurring domains and their mutual relationships. PhyloDome is an answer to that problem. It provides a fast overview on the taxonomic spreading and potential interrelation of domains that are either given as a list of names and PFAM/SMART accessions or derived from a user-defined set of sequences. This taxonomic distribution analysis can be helpful in protein function and interaction assignment as the comparative study of potential Hedgehog pathway members in *C.elegans* shows. An implementation of PhyloDome is accessible for public use as a WWW-Service at http://mendel.imp.univie.ac.at/phylodome/. Software components are available on request.

INTRODUCTION

Globular domains of proteins (in addition to non-globular segments) have been recognized early on as the fundamental building blocks of protein structure and function (1). As a consequence, the evolution of protein complexity has often been rationalized in the context of domain evolution (2). In this respect, it was observed that two phenomena accompany the raise of complexity during organism evolution—an increase of combinatorial diversity through rearrangements of domain architectures (3,4) and the expansion of already existing domain families (5). Domain expansion and the following functional diversification have been recognized as major factors in the extensions of protein functions and the implementation of adaptive tasks (5). Therefore, taxonomic analysis and assessment of lineage-specific variances are important aspects in evaluating the potential functional role of single domains as well as their combinations.

Although the taxonomic distribution of domains occurring in eukaryote multi-domain proteins can provide important indications on function and domain interrelation, this part of the sequence-analytic procedure is not well supported by available tools. Therefore, we developed PhyloDome, which can provide a fast overview on the lineage distribution of domains that are either found in a user-defined set of eukaryote proteins or supplied as a list of names and PFAM (6) or SMART (7) accessions.

DESCRIPTION OF PhyloDome

Methods

In principle, taxonomic distributions can be described in two ways. First, by a phylogenetic distribution: $\boldsymbol{A}(\Delta) = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_i, \ldots, \boldsymbol{a}_N)$, where domain Δ occurs $\boldsymbol{a}_i$-times ($\boldsymbol{a}_i \geqslant 0$) in the proteome derived from genome $\boldsymbol{i}$ ($\boldsymbol{i} = 1, \ldots, \boldsymbol{N}$; $\boldsymbol{N}$ is the number of genomes). Or second, by a phylogenetic profile of domain Δ as $\boldsymbol{sign}(\boldsymbol{A}(\Delta)) = (\alpha_1, \alpha_2, \ldots, \alpha_i, \ldots, \alpha_N)$ where

$$\alpha_i = \begin{cases} 1 & \text{if } 0 < \boldsymbol{a}_i \\ 0 & \text{if } 0 = \boldsymbol{a}_i \end{cases}.$$

Due to the known importance of lineage-specific expansion/contraction of domains and the loss of information in the profile representation, the phylogenetic distribution of a domain is more useful with respect to protein function analysis. Therefore, PhyloDome uses this parameter in most of its tasks. Information on domain occurrences is derived beforehand by inspecting recent proteome releases for the incidence of

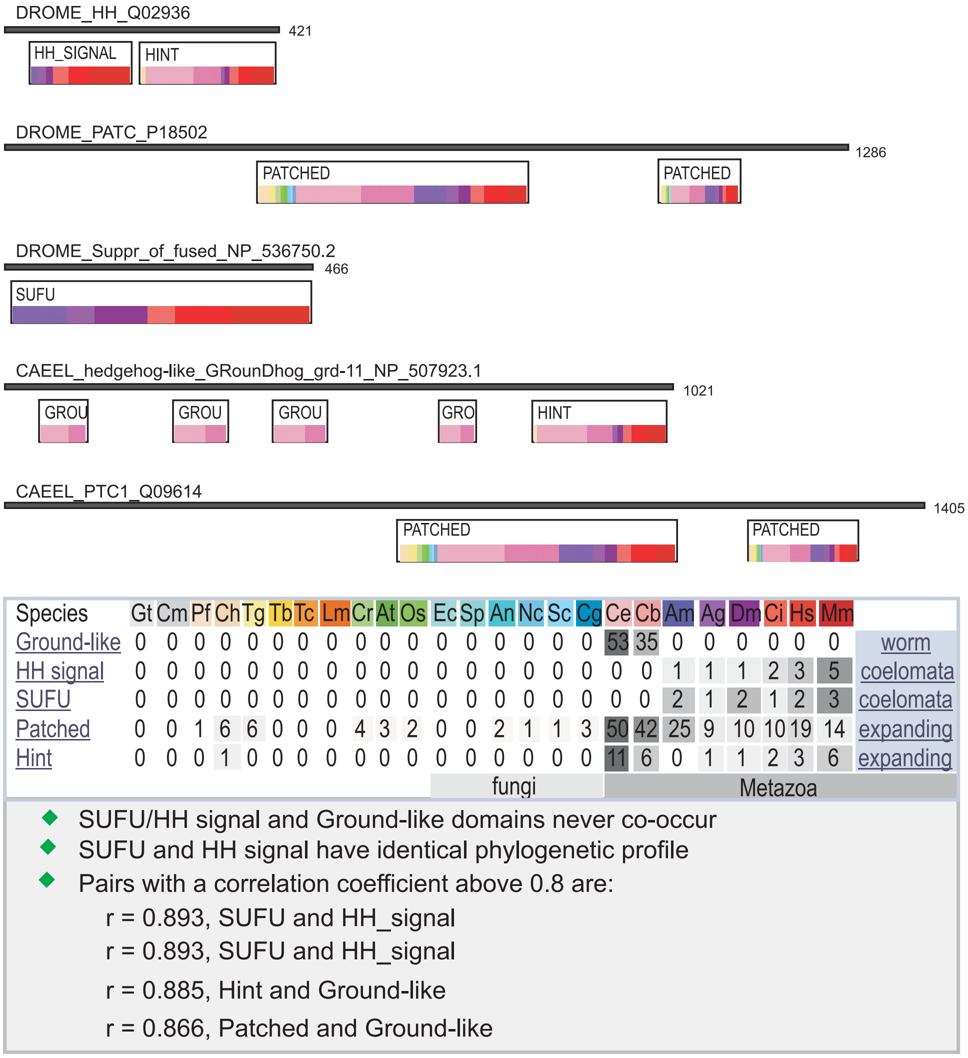
[*] To whom correspondence should be addressed at Mail Research Institute of Molecular Pathology, Dr Bohr-Gasse 7, A-1030 Vienna, Austria. Tel: +43 1 7973 0556; Fax: +43 1 7987 153; Email: novatchkova@imp.univie.ac.at

PFAM (6) or SMART (7) domains using RPS-BLAST with standard parameterization (*E*-value cutoff 0.001; sequences masked for coiled-coils and low complexity). At the time of submission, most sequence data come from the Integr8 (8) and Ensembl (9) genome projects based on December 2004 data mirrors. Please consult the PhyloDome homepage for a complete listing of effective source databases, hyperlinks and mirror dates.

Two modes of visualization (graphical and tabular) are used to present the taxonomic spreading of one or multiple domains across the 25 studied species (Figure 1). In addition to just visualizing taxonomic distribution, PhyloDome also provides some help in the evaluation of results, when viewing both single as well as multiple domains.

Focusing on single domains, PhyloDome aims at classifying these according to evolutionary scenarios that are known to have functional significance. Distinction is made between uniformly distributed (χ^2), lineage-specific expanding (Dixon's outlier test) and several types of lineage-specific domains (e.g. fungi, arthropoda, worm, chordata, etc.). The assignment into these categories allows further interpretations in line with the observation that uniformly distributed domains are mostly involved in basic biological mechanisms, while taxon- and lineage-specific expanding domains are probably serving adaptive functions (5).

Besides single domain classifications, Phylodome also supports the taxonomic comparison between domains. The simplest possible assessment is based on phylogenetic profiles



Species	Gt	Cm	Pf	Ch	Tg	Tb	Tc	Lm	Cr	At	Os	Ec	Sp	An	Nc	Sc	Cg	Ce	Cb	Am	Ag	Dm	Ci	Hs	Mm	
Ground-like	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	53	35	0	0	0	0	0	0	worm
HH signal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	2	3	5	coelomata
SUFU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	2	1	2	3	coelomata
Patched	0	0	1	6	6	0	0	0	4	3	2	0	0	2	1	1	3	50	42	25	9	10	10	19	14	expanding
Hint	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	11	6	0	1	1	2	3	6	expanding
												fungi						Metazoa								

- SUFU/HH signal and Ground-like domains never co-occur
- SUFU and HH signal have identical phylogenetic profile
- Pairs with a correlation coefficient above 0.8 are:

  r = 0.893, SUFU and HH_signal

  r = 0.893, SUFU and HH_signal

  r = 0.885, Hint and Ground-like

  r = 0.866, Patched and Ground-like

Figure 1. PhyloDome representation of sequence sets exemplified by the analysis of Hedgehog pathway proteins in *D.melanogaster* and *C.elegans*. The Hint and Patched domains, which are implicated in cholesterol modification and sensing, show an over-representation in worm, indicated by the magenta fraction in the domain-representation, and are detectable throughout metazoans. In contrast, the Hedgehog signaling (HH signal) and SUFU domains, present in coelomata, and the Ground domain in worm are found in taxonomically exclusive groups. This suggests that not only Hedgehog signaling itself, but also cholesterol signaling is shared between *D.melanogaster* and *C.elegans* Hedgehog and Patched-homologs. Domains are shown as bars divided into colored areas proportional to their occurrence found in proteomes of fully sequenced eukaryote genomes. The color code is repeated in the tabulated phylogenetic profiles: Gt, *Guillardia theta*; Pf, *Plasmodium falciparum*; Ch, *Cryptosporidium hominis*; Tg, *Toxoplasma gondii*; Tb, *Trypanosoma brucei*; Tc, *Trypanosoma cruzi*; Lm, *Leishmania major*; Cm, *Cyanidioschyzon merolae*; Cr, *Chlamydomonas reinhardtii*; At, *Arabidopsis thaliana*; Os, *Oryza sativa*; Ec, *Encephalitozoon cuniculi*; Sp, *Schizosaccharomyces pombe*; En, *Emericella nidulans*; Nc, *Neurospora crassa*; Sc, *Saccharomyces cerevisiae*; Cg, *Candida glabrata*; Ce, *Caenorhabditis elegans*; Cb, *Caenorhabditis briggsae*; Am, *Apis mellifera*; Ag, *Anopheles gambiae*; Dm, *Drosophila melanogaster*; Ci, *Ciona intestinalis*; Mm, *Mus musculus*; Hs, *Homo sapiens*.

and allows very limited reasoning on domain interrelation. The following interpretable scenarios can be distinguished: (i) if domains co-occur in different proteomes, they fulfill the minimal requirement for their functional interaction; (ii) if two domains never occur together in the same taxon, they are, as a rule, not functionally linked. In exceptional cases, two domains with exclusive profiles might represent functional equivalents either corresponding to sequentially very divergent, taxon-specific instances of a domain superfamily or to non-orthologous replacement, a concept that has been previously developed for whole genes of prokaryotes (10,11). In PhyloDome, the user is alerted for the occurrence of overlapping as well as exclusive patterns within a given query set.

As an alternative, phylogenetic distributions of domains (Δ_1 and Δ_2) can also be compared based on the correlation coefficient $\boldsymbol{r}$ between the respective vectors, $\boldsymbol{A}(\Delta_1)$ and $\boldsymbol{A}(\Delta_2)$. In order to confirm that a functional relationship between domains is associated with high correlation of their respective taxonomic distributions, we used two measures for domain interrelatedness: first, physical association of domains in one protein; second, the functional distance between domains based on their Gene Ontology (12) classification.

We observed that physically linked domains also tend to have a high taxonomic correlation coefficient (up to 52% of physical links between domains are found with an $\boldsymbol{r} \geqslant 0.8$ when counted on a sequence basis, Figure 2a). The reverse conclusion is true as a tendency: a high correlation coefficient is indicative for a functional link. The fraction of physically associated domains, among the taxonomically correlating domains (with at least one domain from a multi-domain protein) increases with the correlation coefficient and comprises ~11% for $0.8 \leqslant \boldsymbol{r} \leqslant 1$ regardless of the mathematical form of $\boldsymbol{r}$ (Figure 2b). Among domains with an available Gene Ontology assignment, the average functional distance between domains (counted by the minimal number of separating vertices in the GO tree) tends to drop with increasing taxonomic correlations (Figure 2c).

Input

The input to PhyloDome can be (i) one or a set of sequences entered in a fasta or raw sequence format or (ii) one or a set of domain names and PFAM or SMART accessions. Although it can be used for the analysis of a single domain or protein, PhyloDome unfolds its real potency in the interpretation of sequence sets where functional relationships are the target of interest. Such sets can, for example, contain sequences or domains that are genetically shown to be members of the same pathway or seem to form a complex in a model organism, possibly, together with their homologs in other taxa.

Output

If a sequence set is given as an input, PhyloDome explores the query proteins with RPS-BLAST against the PFAM-A library (13), and derives a list of significantly hitting domains. These domains or the user-supplied domain list are subjected to further analysis. In the results page, a graphical representation of all relevant domains (if appropriate, as domain architecture diagrams of queries) is returned (Figure 1). Color-coded bars reflect the domains' distribution in the (almost complete) proteomes of 25 fully sequenced eukaryotes. For the ease of

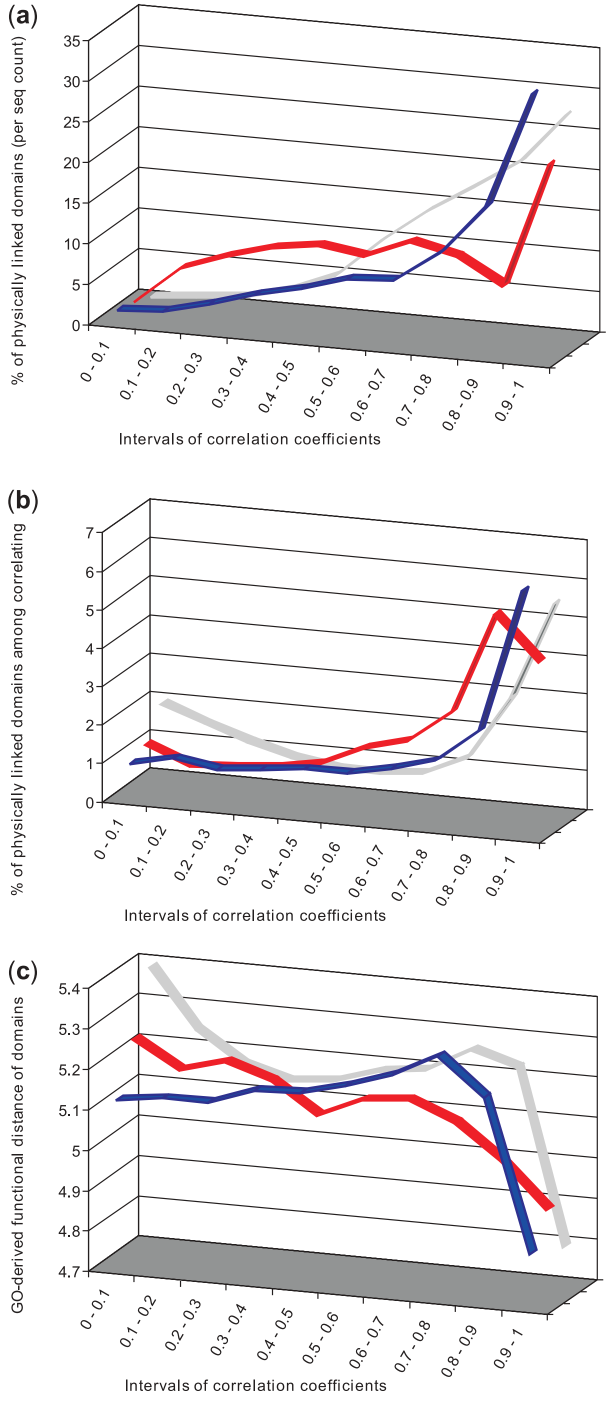


Figure 2. Taxonomic correlation and functional link between domain pairs. (**a**) The distribution of multi-domain proteins with physically linked domain pairs is shown with respect to the taxonomic correlation coefficient (cc) (only reliable physical links between non-homologous domains found in more than three sequences across all species have been considered). (**b**) Diagram showing the fraction of physically associated domain types among the taxonomically correlating domain pairs (with at least one domain from a multi-domain protein). (**c**) Average functional distance between correlating domain pairs estimated by the minimal number of vertices separating them within the GO tree. These data show that a functional relationship between domains is associated with high correlation of their respective taxonomic distributions. Although the performance of the various correlation coefficients is similar, the Pearson cc appears slightly more predictive and is, therefore, used by PhyloDome. (dark-blue, Pearson cc of taxonomic distribution; red, Pearson cc of taxonomic profile; gray, Spearman cc of taxonomic distribution).

interpretation, a phylogenetic tree of eukaryote taxa is supplied with the same color-coding. For each domain, a mouse-over function supplies domain name, PFAM accession number, a link to the PFAM domain annotations and the numerical data for the phylogenetic profile.

In addition to the described graphical display, the evolutionary distribution is tabulated numerically for all domains (Figure 1). Gray background shading varying proportionally in its intensity shows the domain occurrences visually. For pairs of domains, overlaps/exclusions of taxonomic profiles and significant correlation coefficients based on their phylogenetic distributions are reported. With this data, the user can rapidly identify the domains in a protein set evolving in a distinct and correlated fashion, which might be functionally linked.

The following sources of possible errors should be taken into account when interpreting PhyloDome outputs. Of course, the computation results depend on the accuracy of the domain models and the completeness of the domain library. With regard to phylogenetic distributions, the accuracy of measured domain occurrences $\boldsymbol{a}_i$ in proteomes of model organisms is critical. Incomplete genome sequencing, assembly errors and inaccurate gene structure determination will, most often, lead to lower domain occurrences since protein sequences might become (i) absent from the derived proteome, (ii) shortened or (iii) partially substituted by or appended to non-sense sequences. On the other hand, false positive domain assignments (especially for small domains and domains with compositional bias) might artificially increase domain occurrences.

Application example: hedgehog signaling in *Caenorhabditis elegans*?

The occurrence of about a dozen of proteins with Hint (PFAM accession PF01079) (14) and Patched (PF02460) (15) domains encoded in the worm genome has led to the conclusion that a Hedgehog-related pathway might exist in *Caenorhabditis elegans*. Many Hint domain-containing worm sequences (e.g. NP_500347 and NP_501673) are even annotated as Hedgehog-like in GenBank protein database. As studied mainly in fly, Hedgehog is known to be expressed as a precursor protein, and auto-processed via its catalytic C-terminal domain. A thioester intermediate attacked by cholesterol releases the N-terminal signaling peptide modified by cholesterol and the C-terminal cysteine peptidases domain (16). The Patched receptor (Ptc) is thought to sequester Hedgehog by a cholesterol-dependent process. Ptc signals on through Smoothened and a complex including Fused, Costal, Suppressor-of-Fused and Cubitus interruptus (17).

Considering the so-called Hedgehog-related proteins in *C.elegans* (Figure 1) as an example, we show that analysis of the phylogenetic distribution of domains constituting homologous multi-domain proteins in different species helps to correctly transfer functional annotations. For this purpose, the known Hedgehog pathway proteins Hedgehog (Q02936), Suppressor-of-Fused (NP_536750.2), and Patched (P18502) from fly as well as the Hedgehog-like (NP_507923.1) and Patched homolog 1 (Q09614) from worm were submitted to PhyloDome (Figure 1). Focusing on single domain evolutionary scenarios, it becomes clear that cholesterol-based signaling is of enhanced importance in worm. The two domains of the pathway, known to be involved in cholesterol-dependent processes (18) Hint (cholesterol modification) and Patched (sterol sensing) are lineage-specifically expanded in worm. Other domains found in Hedgehog signaling (Hh_signaling/PF01085 and Sufu/PF05076) are not detectable in worm and are, apparently, coelomata specific. These observations indicate that it is not the Hedgehog signaling itself, but only the cholesterol modification and sensing mechanism that is expanded in worm.

Further backing comes from the pairwise co-evolution analysis. Whereas the C-terminal cysteine peptidase domain Hint is shared, the coelomata-specific domains involved in Hedgehog signaling (Hh signaling and Sufu) and the worm-specific Ground-like domain (PF04155) are occurring in taxonomically exclusive groups. Therefore, and because of the absence of significant sequence similarity, the two groups of domains are most likely functionally unrelated. The phylogenetic patterns of the cholesterol-signaling domains Hint and Patched show a high correlation coefficient supporting their potential functional linkage. These data support the following model of protein function evolution in the different taxonomic ranges: it seems that divergent cholesterol-dependent signaling processes have evolved in coelomata and pseudo-coelomata lineages (with a vast expansion in pseudo-coelomata).

CONCLUSION

The visualization tool PhyloDome is thought to facilitate studies of eukaryote protein sets in the context of taxonomic distribution of their domains. This program supports the easy identification of typical evolutionary scenarios of single domains and domain pairs and enhances their indicative value for functional annotation.

ACKNOWLEDGEMENTS

The authors are grateful for the generous support from the Austrian Academy of Science and Boehringer Ingelheim. This project has been partly funded by the Austrian Science Fund (FWF), Austrian Gen-AU bioinformatics integration network sponsored by the Austrian Ministry of Education, Science and Culture (BM-BWK) and by a grant from the Austrian Economy Ministry (BM-WA). Funding to pay the Open Access publication charges for this article was provided by the Research Institute of Molecular Pathology.

Conflict of interest statement. None declared.

REFERENCES

1. Janin,J. and Chothia,C. (1985) Domains in proteins: definitions, location, and structural principles. *Methods Enzymol.*, **115**, 420–430.
2. Koonin,E.V., Wolf,Y.I. and Karev,G.P. (2002) The structure of the protein universe and genome evolution. *Nature*, **420**, 218–223.
3. Apic,G., Gough,J. and Teichmann,S.A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.
4. Vogel,C., Teichmann,S.A. and Pereira-Leal,J. (2005) The relationship between domain duplication and recombination. *J. Mol. Biol.*, **346**, 355–365.
5. Lespinet,O., Wolf,Y.I., Koonin,E.V. and Aravind,L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.*, **12**, 1048–1059.

6. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
7. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.
8. Kersey,P., Bower,L., Morris,L., Horne,A., Petryszak,R., Kanz,C., Kanapin,A., Das,U., Michoud,K., Phan,I. *et al.* (2005) Integr8 and genome reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.*, **33**, D297–D302.
9. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
10. Galperin,M.Y. and Koonin,E.V. (2000) Who's your neighbor? New computational approaches for functional genomics *Nature Biotechnol.*, **18**, 609–613.
11. Koonin,E.V., Mushegian,A.R. and Bork,P. (1996) Non-orthologous gene displacement. *Trends Genet.*, **12**, 334–336.
12. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
13. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.
14. Aspock,G., Kagoshima,H., Niklaus,G. and Burglin,T.R. (1999) *Caenorhabditis elegans* has scores of hedgehog-related genes: sequence and expression analysis. *Genome Res.*, **9**, 909–923.
15. Kuwabara,P.E., Lee,M.H., Schedl,T. and Jefferis,G.S. (2000) A C. elegans patched gene, ptc-1, functions in germ-line cytokinesis. *Genes Dev.*, **14**, 1933–1944.
16. Jeong,J. and McMahon,A.P. (2002) Cholesterol modification of Hedgehog family proteins. *J. Clin. Invest.*, **110**, 591–596.
17. Nybakken,K. and Perrimon,N. (2002) Hedgehog signal transduction: recent findings. *Curr. Opin. Gen. Dev.*, **12**, 503–511.
18. Kuwabara,P.E. and Labouesse,M. (2002) The sterol-sensing domain: multiple families, a unique role? *Trends Genet*, **18**, 193–201.