BIOINFORMATICS

Vol. 17 no. 4 2001
Pages 343–348

NIFAS: visual analysis of domain evolution in proteins

Christian E. V. Storm and Erik L. L. Sonnhammer

Center for Genomics Research, Karolinska Institutet, 171 77 Stockholm, Sweden

Received on October 11, 2000; revised on December 8, 2000; accepted on December 13, 2000

ABSTRACT

Motivation: Multi-domain proteins have evolved by insertions or deletions of distinct protein domains. Tracing the history of a certain domain combination can be important for functional annotation of multi-domain proteins, and for understanding the function of individual domains. In order to analyze the evolutionary history of the domains in modular proteins it is desirable to inspect a phylogenetic tree based on sequence divergence with the modular architecture of the sequences superimposed on the tree.

Result: A Java applet, NIFAS, that integrates graphical domain schematics for each sequence in an evolutionary tree was developed. NIFAS retrieves domain information from the Pfam database and uses CLUSTAL W to calculate a tree for a given Pfam domain. The tree can be displayed with symbolic bootstrap values, and to allow the user to focus on a part of the tree, the layout can be altered by swapping nodes, changing the outgroup, and showing/collapsing subtrees. NIFAS is integrated with the Pfam database and is accessible over the internet (http://www.cgr.ki.se/Pfam). As an example, we use NIFAS to analyze the evolution of domains in Protein Kinases C.

Contact: christian.storm@cgr.ki.se

INTRODUCTION

Phylogenetic methods are widely used to analyze the evolutionary history of protein sequences (Lake and Moore, 1998). However, many proteins consist of multiple independently evolving domains (Hegyi and Bork, 1997). Such proteins often contain different numbers of domains in different orders; therefore they are not directly amenable to traditional phylogenetic analysis of the entire sequence. Instead it is necessary to isolate the separate domains and carry out the phylogenetic analysis on each domain separately. However, such a reductionistic approach does not lead to understanding how a given combination of domains has evolved from simpler modules, and what the functional implications of this evolution are. In many cases, cassettes of domains have been preserved in a large set of proteins, while in other cases domains have been inserted or deleted more recently.

The goal of this work is to provide a tool that can reveal how domain combinations have evolved in protein sequences. This is especially important when analyzing orthologous relationships between proteins in different organisms. The relationship may seem orthologous in one domain but not in another; in such cases one needs to be careful in predicting function.

We use the Pfam (Bateman *et al.*, 2000) database as the source for protein domain definitions. The Pfam database is a database of protein domain families with manually annotated multiple sequence alignments of high quality. For each Pfam family a profile hidden Markov model (HMM) (Krogh *et al.*, 1994) is calculated from the alignment. The HMMs are used to find all family members in public databases, and as a library of HMMs to search a query sequence against. If available, functional annotation, literature references and database links are included in the family annotation. Pfam version 5.4 (June 2000) consists of 2290 families. Pfam can be accessed over the internet and is currently mirrored at three different sites around the world. Different tools are already available that allow an effective domain analysis. For example a graphical representation of the domain structure of all members in a family can be displayed, and it is possible to search Pfam for proteins that have a given domain architecture or is similar in domain architecture to another protein.

We here introduce a novel analysis method that combines the domain analysis in Pfam with phylogenetic tree analysis. This is achieved by a Java applet named NIFAS, which calls the tree-calculating program CLUSTAL W (Thompson *et al.*, 1994a) and the Pfam Web server for a given domain, and displays the information in a combined display. We here present the features of NIFAS and show how it can be used to analyze the evolution of domain combinations in Protein Kinases C.

METHODS AND IMPLEMENTATION

NIFAS is implemented as a Java 1.1 applet that forms an integral part of the Pfam Web server system. It utilizes a module in the Pfam Web server that calculates graphical representations of domain architectures. As shown in

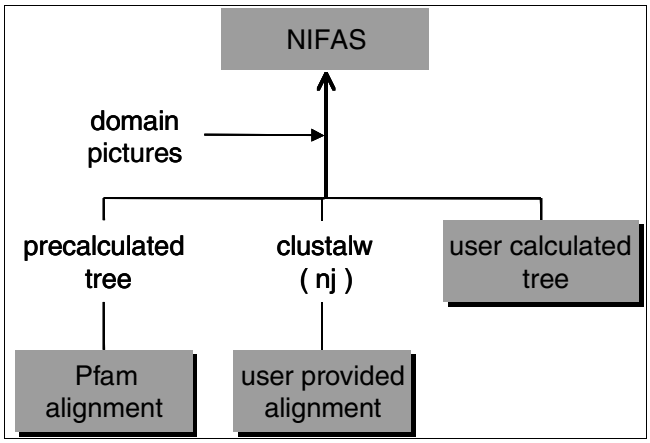

Fig. 1. The different ways of accessing NIFAS. The fastest way is to use the link from a Pfam domain web page. Other possibilities are to upload an alignment or a tree. The Pfam server in Sweden uses precalculated trees; trees for user provided alignments are calculated with CLUSTAL W using the neighbor joining method. The tree is then connected with graphical representations of the protein domain architecture and finally displayed with NIFAS.

Figure 1 these images are then connected with a tree. The tree can be provided in three different ways: the first and easiest way is to use the NIFAS link on a Pfam web page for a given domain family. To allow fast access, the Stockholm Pfam database stores pre-calculated trees for families of up to 500 members (250 for bootstrapped trees). These trees were calculated with CLUSTAL W from the multiple alignment for each Pfam family using 100 bootstraps.

The second way to run NIFAS, by uploading an alignment (Figure 1) on the NIFAS web page (http://www.cgr.ki.se/Pfam/nifas.html), allows more flexibility. Here the user can provide an alternative alignment, and optionally select Kimura (Kimura and Ota, 1972) correction of the distances and the number of bootstraps, up to 1000. The sequence names must be present in the Pfam database, and the alignment must be 'aligned Fasta' format. It is important to provide start and end coordinates after the sequence name (name/start-end) so NIFAS can locate the domain in question.

The third way to run NIFAS is by uploading a tree in Newick format (Felsenstein *et al.*, 2000) on the NIFAS web page. This way any tree calculating method can be used. At the moment it is not possible to upload the user's own domain definitions; only the Pfam definitions can be used.

The two latter methods allow NIFAS analysis of families with more sequences than the limit for the precalculated trees (max 500, or 250 with bootstrapping).

THE GRAPHICAL USER INTERFACE

In Figure 2 an example screenshot of NIFAS is shown. The domains corresponding to the actual sequence segments in the multiple alignment are marked with a tiny tree icon. These can be seen as 'active' domains, while the other domains have no influence on the tree. If a protein has multiple repeats of the active domain it is present multiple times in the tree.

The full description of the domains and sequences are displayed in the information box on top when these objects are clicked. On-line help is provided for these functions. If bootstrapping was selected for the tree calculation, the tree nodes (except the root node) are labeled with colored boxes: green >90%, yellow 90–75%, white 75–50%, and no box 50–0% bootstrap support.

The tree layout can be changed by collapsing/expanding nodes, swapping the branches of a node or by choosing a new outgroup. It is possible to show a 'subtree' only, starting at any node. NIFAS can also interact with the Web browser that called NIFAS. By clicking on a domain or protein the corresponding Pfam Web page is displayed in the browser.

By default one pixel of a domain picture represents two amino acids; this and the scaling of the tree can be changed. The display of domains can be toggled between 3-D look and a plain look, which scrolls faster. These features are accessed via buttons on the navigation bar on top (see Figure 2), which can also be torn off or 'undocked' to allow larger windows of the tree.

The example in Figure 2 shows a selection of proteins in the PTS-HPr family. From this simple view, we can already infer that the two domains in PTFA_HAEIN are more similar to each other than to PTFA_ECOLI and PTFA_SALTY. This observation supports an evolutionary history with an ancestral gene containing one PTS-HPr domain, which was duplicated in the *Haemophilus* lineage. Alternatively, the ancestor did have two domains, and one was lost in the *E.coli* and *Salmonella* lineage, but this is less likely since no other protein is known to contain two PTS-HPr domains.

In many cases one is interested in whether two different domains have evolved in a congruent way, i.e. the domain architecture is ancestral. For example, one could run NIFAS on the other domain in Figure 2, PTS_EIIA_2, and inspect whether the proteins appear in the same tree topology as in PTS-HPr. This is indeed the case (data not shown), and we can infer that the ancestor to PTFA_HAEIN, PTFA_ECOLI, and PTFA_SALTY had one PTS_EIIA_2 domain and one PTS_HPr domain. If the topology were different, we would infer that the ancestor only had one of the domains and that the other domain was inserted independently. For this type of comparative analysis, it is necessary to open two or more

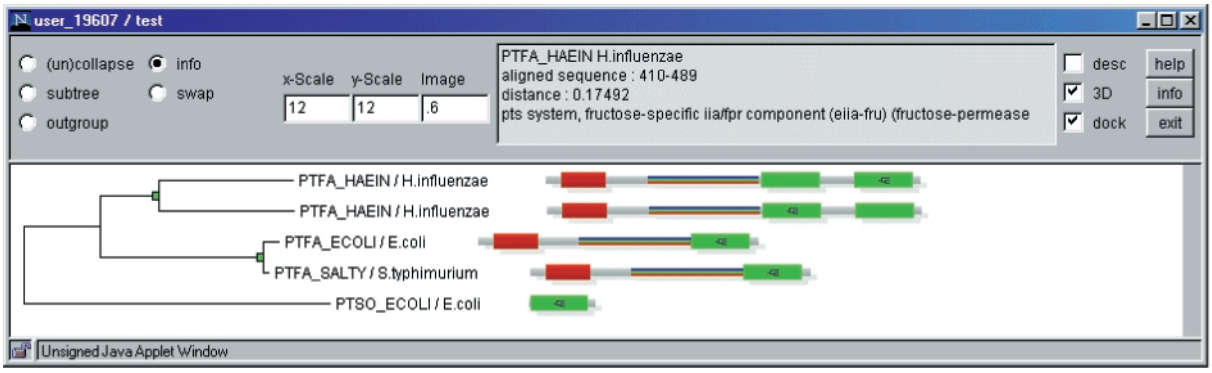

Fig. 2. The NIFAS display. Selected members of the family Pfam: PTS-HPr (PF00381) are shown. The green domains are PTS-HPr domains while the red domains are Pfam: PTS_EIIA_2 domains (PF00359). The tree was calculated using the Pfam alignment of the domains marked with a tiny tree icon. The small node boxes are colored green, indicating a bootstrap support >90%. In this example it is clear that the two PTS-HPr domains in the *Haemophilus* protein (PTFA_HAEIN) are more similar to each other than to the *E.coli* and *Salmonella* PTS-HPr domains. It is thus likely that the duplication of the PTS-HPr domain is unique to the *Haemophilus* lineage.

NIFAS windows simultaneously. In order to fit this much information on the screen it is advisable to 'undock' the NIFAS navigation bar.

ANALYZING THE EVOLUTION OF C1 AND C2 DOMAINS IN PROTEIN KINASES C

A more elaborate evolutionary analysis can be made if the proteins contain more than two domains, take for instance human protein kinases. Protein kinases C that contain the C2 domain and phorbol esters/diacylglycerol binding domains (also known as C1 domain) (Medkova and Cho, 1999; Thomas *et al.*, 1999) can be found in a wide range of species. Although these kinases are closely related, the domain order often differs. In *Homo sapiens* one can find four groups of kinases with different domain structure that have at least the eukaryotic protein kinase domain and one C1 domain (see Figure 3):

—KPCE and KPCL: one kinase domain, two C1 domains, and one C2 domain;

—KPCA, KPCG, KPC1 and KPC2: same as KPCE and KPCL, but the domain order is swapped;

—KPCD and KPCT: one kinase domain, two C1 domains. The Swissprot annotation of these proteins says that they contain an N-terminal C2 domain, but it belongs to a divergent subfamily that is not found by the C2 domain classifiers in Pfam 5.4 or Prosite. This type of C2 domain is also found in kinases similar to KPCD/T in other species, e.g. in *C.elegans*. Because it is not recognized in Pfam, it is not visible in Figure 3. We have however marked it in Figure 5 as a distinct type of C2 domain.

—KPCZ and KPCI: one kinase domain, one C1 domain, and one Octicosapeptide region.

In this analysis we assume that the kinase domain is ancestral and was passed on vertically to its descendants, because it possesses the catalytic activity.

The C1 domain

The tree shown in Figure 3 suggests that the ten proteins in the four groups mentioned above have a common ancestor that contained the kinase domain, the kinase C-terminal domain, and probably at least one C1 domain. But did this ancestor have two C1 domains, one of which was lost in KPCZ/I, or did it have one C1 domain that was duplicated in the other groups?

In NIFAS we can view the same proteins but with the tree calculated from the C1 domains, see Figure 4. Here one can see that the C1 domain of KPCZ/I is more closely related to the first C1 than to the second C1 domain in the other sequences. This topology is in principle compatible with both evolutionary models. However, if the ancestor had one C1 domain, one would need to postulate that after duplication, the second C1 domain evolved at a faster rate after the duplication, or that KPCZ/I evolved in the same direction as the first C1 domain. This is of course possible, but we favor the model with two ancestral domains since it does not require assuming unequal divergence rates.

The C2 domain

Two of the four groups have a C2 domain, but in different locations. Within each group, the tree based on the C2 domain follows the topology of the kinase domain. It is therefore likely that the C2 domain was gained twice: once in the ancestor of KPCE/L and once in the ancestor of KPCA/G/1/2.

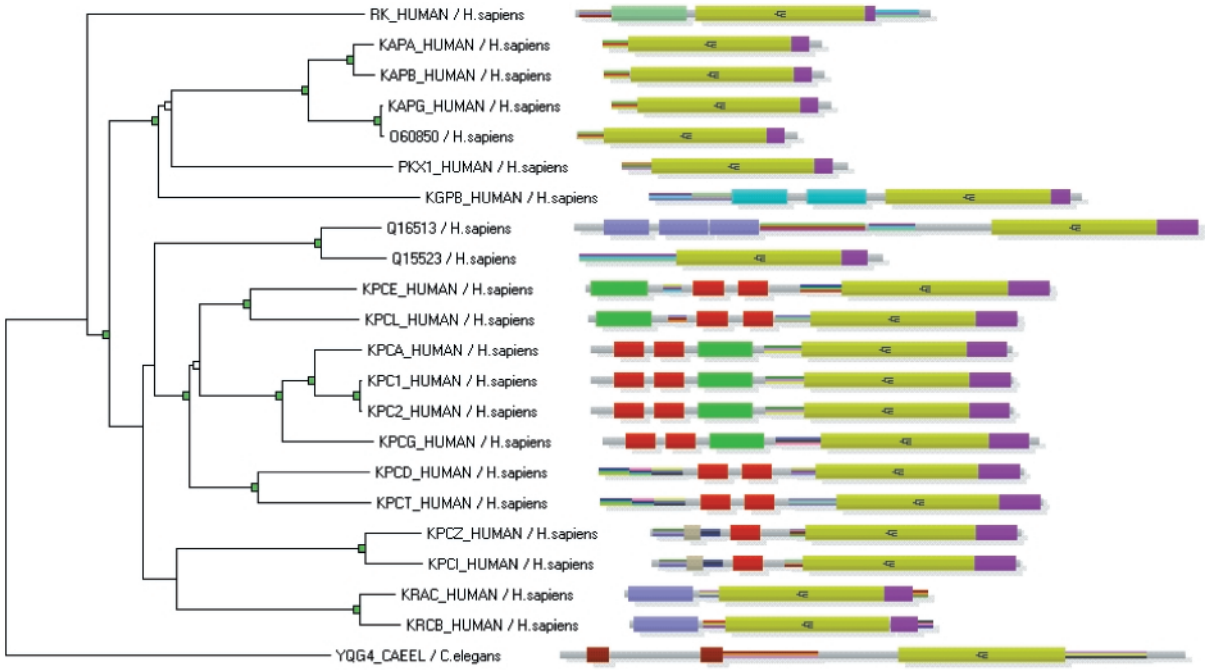

Fig. 3. NIFAS view of human proteins containing the protein kinase domain. The *C.elegans* protein YQG4 is used as outgroup and the tree is calculated from the Pfam alignment for the kinase domain (PF00069: pkinase, yellow), as indicated by the tree icon. Other domains are C1 (PF00130: DAG_PE-bin, red), C2 (PF00168: C2, green), protein kinase C-terminal domain (PF00433: pkinase_C, purple), and octicosapeptide region (PF00564: OPR, light brown).

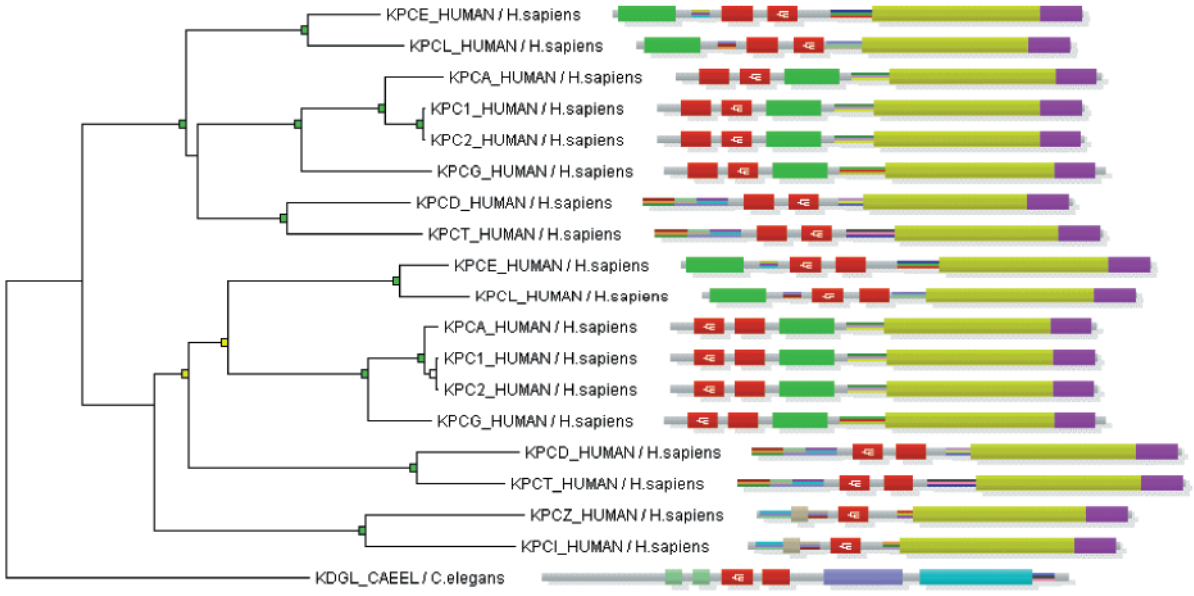

Fig. 4. NIFAS view of human protein kinases with the tree calculated for the C1 domains (red boxes marked with tree icons). A domain in the *C.elegans* protein KDGL is used as outgroup. Domain colors as in Figure 3.

All alternative scenarios seem less likely. It is possible that the ancestor of the KPCE/L/A/G/1/2 branch contained one C2 domain but no C1 domains, and the C1 domains were later inserted twice. This faces the problem that KPCD/T have two C1 domains. Such a scenario would thus postulate that the C1 domains in

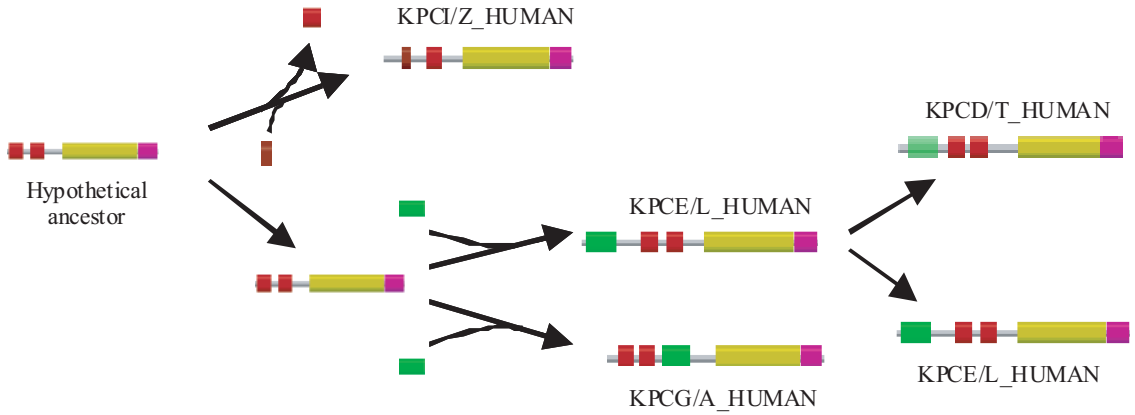

Fig. 5. Reconstruction of events in the evolution of protein kinases C derived from the NIFAS analysis. Domains are colored as in Figure 3. Next to each graphical representation of domain architectures a sample protein is named that represents the domain architecture (only Pfam A domains are shown). Note that no protein like the hypothetical ancestor is known in human. The transparent green box in KPCD/T_HUMAN indicates that it is a divergent type of C2 domain that is not recognised by the Pfam HMM (see text).

KPCE/L/A/G/1/2/D/T were gained in three independent events, which is less parsimonious than the two events necessary for inserting the C2 domains in two different places in the KPCE/L/A/G/1/2 branch.

Summary

An evolutionary model of domain insertion and loss can be constructed from the above observations. A graphical illustration of how the most likely domain rearrangement events occurred in chronological order is found in Figure 5. We refrain from dating these events precisely, but it is clear that they must have happened in early mammalian evolution, since they are shared among all present day mammals but not with lower organisms. The 'hypothetical ancestor' protein with two C1 domains and no C2 domain however appears to predate the mammals. Protein kinases with tandem C1 domains are also found in fungi, suggesting that this duplication goes back to an ancient ancestor, thus lending support to the hypothesis of an early C1 duplication and multiple independent C2 domain insertions. None of the fungal kinases have a single C1 domain. Further support comes from the fact that although many protein kinases have C1 domains but no C2 domain, not a single case is known with a C2 domain but no C1 domain.

DISCUSSION

The analysis of the protein kinases shows that examining phylogenetic relationships of domains rather than of whole proteins can yield insights into the constraints and mechanisms of protein evolution. While any approach of analyzing the phylogenetic relationships of these proteins based on whole sequences rather than on single domains is likely to fail because of the domain shuffling, NIFAS allows a correct and detailed analysis. By comparing the trees derived for the different domains it was possible to develop a model for the evolution of the protein kinases C based on different recombination events.

A problem is the absence of a well-founded method for rooting trees. While the UPGMA-method for building trees gives the position of the root, UPGMA fails to reconstruct a tree correctly if the sequences evolved with different evolutionary rates. The Neighbor-Joining method (Saitou and Nei, 1987), does not fail in these cases, but it does not calculate the position of the root.

Some neighbor-joining programs, e.g. CLUSTAL W, place the root at the center-point of the tree (Thompson *et al.*, 1994b), but this is rather ad-hoc. Ideally a tree is rooted with an outgroup sequence, usually a distant homolog. Finding a correct outgroup sequence is however not without pitfalls. With NIFAS, the additional information of the domain order of a protein can be included when choosing an outgroup sequence.

The protein domain databases are a rich resource of information. So far the phylogenetic analysis of domain architectures has been difficult due to the lack of an intuitive and easy to use tool. NIFAS tries to fill this gap. Although it is connected to Pfam in a way that it uses it domain definitions and multiple alignments, it could easily be altered to work with any domain database.

ACKNOWLEDGEMENTS

We thank Mats Jonsson for developing programs that calculate the domain images and Kevin Howe for installing NIFAS on the Sanger Centre Pfam web server.

REFERENCES

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Felsenstein,J., Archie,J., Day,W.H.E., Maddinson,W., Meacham,C., Rohlf,F.J. and Swofford,D. (2000) The Newick tree format. WWW URL: http://www.evolution.genetics.washington.edu/phylip/newicktree.html.

Hegyi,H. and Bork,P. (1997) On the classification and evolution of protein modules. *J. Protein Chem.*, **16**, 545–551.

Kimura,M. and Ota,T. (1972) On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.*, **2**, 87–90.

Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Lake,J.A. and Moore,J.E. (1998) Phylogenetic analysis & comparativegenomics. In Brenner,S. and Lewitter,F. (eds), *Trends Guide to Bioinformatics.* Elsevier, Amsterdam, pp. 22–23.

Medkova,M. and Cho,W. (1999) Interplay of C1 and C2 domains of protein kinase C-alpha in its membrane binding and activation. *J. Biol. Chem.*, **274**, 19 852–19 861.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Thomas,D.M., Ferguson,G.D., Herschman,H.R. and Elferink,L.A. (1999) Functional and biochemical analysis of the C2 domains of synaptotagmin IV. *Mol. Biol. Cell*, **10**, 2285–2295.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994a) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994b) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.*, **10**, 19–29.