FEBS 25670 FEBS Letters 513 (2002) 129–134

Minireview

Protein domain analysis in the era of complete genomes

Richard R. Copley[a], Tobias Doerks[a,b], Ivica Letunic[a], Peer Bork[a,b,*]

[a] *EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany*
[b] *Max-Delbrück-Center for Molecular Medicine, Berlin-Buch, Germany*

Received 8 November 2001; revised 20 November 2001; accepted 3 December 2001

First published online 20 December 2001

Edited by Gianni Cesareni and Mario Gimona

Abstract Domains present one of the most useful levels at which to understand protein function, and domain family-based analysis has had a profound impact on the study of individual proteins. Protein domain discovery has been progressing steadily over the past 30 years. What are the realistically achievable goals of sequence-based domain analysis, and how far off are they for the sequences encoded in eukaryotic genomes? Here we address some of the issues involved in better coverage of sequence-based domain annotation, and the integration of these results within the wider context of genomes, structures and function.

Key words: Protein domains; Genome analysis; Evolution; Sequence analysis

1. Introduction

Genome sequences may be large, but they are finite. In 1992, Chothia showed that when homology between genes is taken into account by comparison of protein sequence and structure, the overall number of distinct protein families (i.e. folds) is likely to be surprisingly small: a thousand or so, an order of magnitude smaller than the number of genes in a genome [1]. Subsequent work has led to differing estimates (for a recent approach, see e.g. [2]), but the impact of the essential message remains. Now that complete genome sequences are available, we should, within certain limits, be able to delineate all the protein domains that are contained within them. What progress is being made in the journey towards this goal?

The term 'domain' can relate to protein structure or function, but our interest here is in the former sense. Domains are regions of compact protein structure, typically with a hydrophobic core [3]. Beyond this, it is useful to distinguish the term 'domain family' from 'protein family'. The former implies the discrete structural folding unit, whereas the latter refers to a combination of domains that always occur together within the same polypeptide, or to proteins composed of a single domain.

Domains themselves may have evolved from smaller structural units such as repeats [4] or the assembly of small folding motifs into larger structures seen today [5]. There is evidence to suggest that genetically mobile elements like transposable elements may also influence protein structure, e.g. [6]. Even so, whole domains do seem to be the dominant structural unit shaping proteins, and in addition, correlation with experimental results suggests they often represent fundamental functional units. Sequence-based domain definitions are, therefore, one of the most convenient and practically important levels at which to understand the evolution of protein function.

Domain types that are found with different combinations of domains in different proteins are particularly significant. Homologous sequence is assumed to have similar function, whatever the context, so these 'modules' or 'mobile domains' allow the transfer of functional information, such as being involved in a particular kind of interaction, between distinct protein classes. The exact extent to which functional information can be usefully transferred, however, varies greatly and may be difficult to establish a priori, although the higher the proportion of shared domains in two proteins, the more similar their functions [7].

Breaking a protein down into its constituent domain components is evidently a reductionist approach, but one which, to judge from the level of citations of domain discovery papers (see Fig. 1), is of great importance to an understanding of protein function. This is particularly true of metazoan genomes, and human in particular, where multi-domain proteins abound [8]. The identification of orthologues (genes related by speciation events) and paralogues (genes related by an intra-genome duplication event) represents an alternative and complementary approach to tracking the evolution of function, but the many-to-many evolutionary relationships of metazoan proteins, and their multi-domain nature, complicates the application of these concepts, making initial domain-based annotation more appealing [9]. Thus, increasing the numbers of proteins for which domain-based annotation can be provided is an important goal of computational genome analysis.

2. Improvements in sequence database searching

The domain is defined in terms of protein structure, but as protein sequence determines protein structure, and sequence data is much easier (and cheaper) to come by than structural data, much discovery of mobile protein domain families has revolved around the identification of conserved sequences. This means that improvements in sequence database-searching techniques will directly lead to better coverage of protein domains.

*Corresponding author.
E-mail address: bork@embl-heidelberg.de (P. Bork).

0014-5793/02/$22.00
PII: S0014-5793(01)03289-6

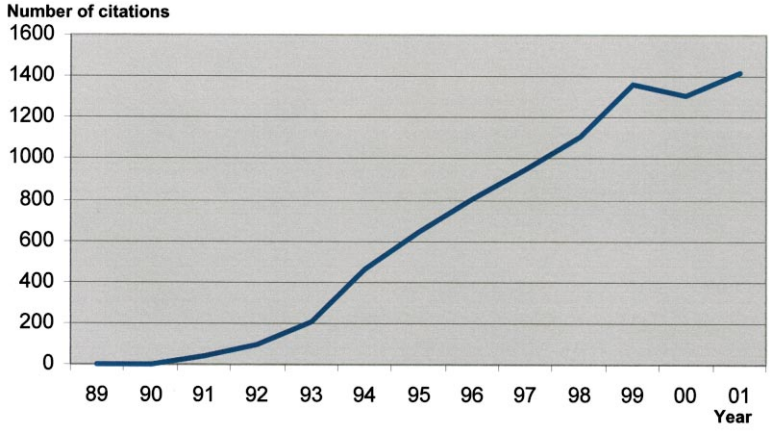

Fig. 1. Plot of total numbers of citations for selected domains (SH2, SH3, PH, CH, BROMO, WW, DH, FHA, PTB, PDZ, WH2, FYVE, EH, VHS, ENTH). The number of citations for 2001 is an estimate – 1/6 of the current numbers for 2001 (i.e. approximately 2 months) were added to provide an approximation to the total for the year. Only articles mentioning the domain names in the title or the abstract were counted.

Techniques for searching sequence databases and determining whether the similarity between two sequences is likely to be due to homology are well established [10], and we have previously described their application to domain discovery [11]. Subtle improvements continue to be made, in particular to the statistical models used to determine significance levels in sequence comparison; new methods have been implemented to better model the effects of biased residue composition of sequences [12,13]. These incremental improvements in specificity (the ability to discriminate between true and false positives) can occasionally lead to more radical improvements in sensitivity, as the correct identification of borderline matches can provide crucial bridges between distinct subfamilies of a given sequence family.

The past 3 decades have seen relatively steady levels of domain discovery (Fig. 2). Although it seems likely that most of the more common mobile protein domains have already been described in the literature, the ever-increasing numbers of sequences in databases provide new sources for domain detection. The value of new sequences comes from two effects. Firstly, they can provide new contextual information for particular regions of homology, making it obvious that a particular homologous sequence family constitutes a mobile module. Secondly, new sequences can lead to the formation of statistically significant bridges between two families that had not previously appeared to be mobile. Additional sequences in a database also reduce the statistical significance of true matches when database searching [14]. In practice, the former two effects appear to more than compensate for this latter phenomenon, so new distant sequence relationships continue to be discovered despite growing database sizes.

3. Automatic approaches to domain detection

It is computationally intensive, but relatively straightforward, to apply database-searching techniques to entire sequence databases (for instance, either a whole genome, or a complete non-redundant sequence database), and thus establish all significant sequence similarities detectable by any given method. These similarities between pairs of sequences can then be clustered into sets of putatively homologous proteins. However, the multi-domain nature of proteins complicates the

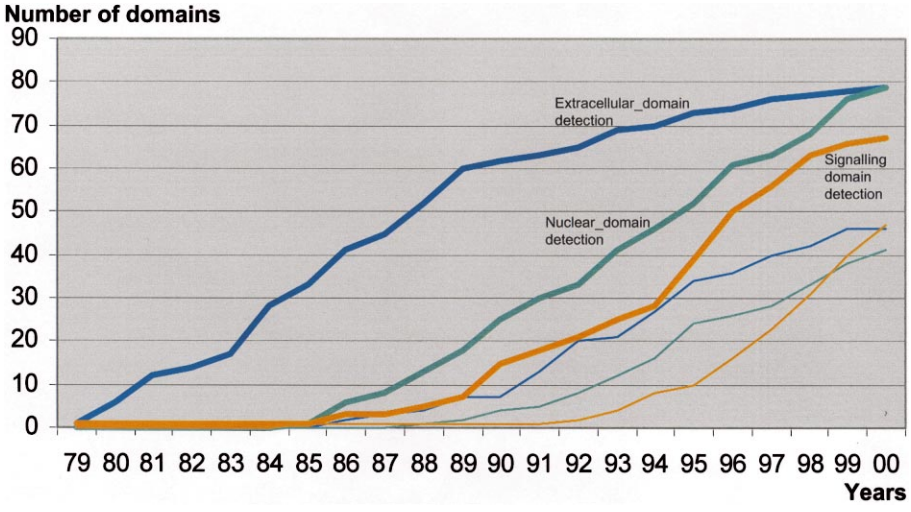

Fig. 2. Plot of cumulative numbers of detected domains or solved structures against year. Bold blue line: number of detected extracellular domains, bold green line: number of detected nuclear domains, bold orange line: number of detected signalling domains, thin blue line: number of solved extracellular domain structures, thin green line: number of solved nuclear domain structures, thin orange line: number of solved signalling domain structures.

clustering procedure. A protein consisting of two domains, A and B, will cause the cluster containing homologs of domain A to be merged with that containing homologs of domain B. A number of automatic techniques have been developed to identify multi-domain proteins and decompose them into their respective domain complements [15–20]; the basic principle of all is that domain boundaries can be inferred by automatic inspection of sequence alignments. In practice, low levels of sequence conservation between members of a domain family can make it difficult to establish domain boundaries, particularly from sets of pairwise sequence comparisons. This, and other problems, such as the difficulty of setting universal thresholds to establish homology between sequences within domain families, and the problem of usefully annotating automatically defined families, reduce the efficacy of these otherwise attractive approaches.

4. Domain databases

Databases such as SMART (http://smart.embl-heidelberg.de), concentrating chiefly on mobile domains [21], and PFAM (http://www.sanger.ac.uk/software/pfam/), which covers protein families and domains [22], make use of hand-edited sequence alignments representing single protein domains. An initial set of homologs is gathered, usually via a database search, although other techniques, such as internal repeat detection can be more efficient [12,43]. The sequences are then automatically aligned and the alignment manually edited to improve quality. This alignment can then be used to perform another search against the sequence database via a profile or hidden Markov model. This process can then be iterated until no new hits are found. Once completed, the model has statistical thresholds associated with it to ensure the correct classification of true and false positive sequence matches. These thresholds are decided by the annotator preparing the specific model, and are not universally applied to all the models in a database.

Various web-based resources exist for searching libraries of the domain models constructed in this way, and thus identifying known domains in protein sequences (see Fig. 3 for an example of SMART output). In addition to the search sites of the databases themselves, 'meta'-sites exist that allow for the searching of multiple domain databases. For instance, the Interpro database, at the EBI (http://www.ebi.ac.uk/interpro/) allows the searching of the Prosite, PFAM, PRINTS, ProDom and SMART model collections [23], and the Conserved Domain Database (CDD) at the NCBI (http://www.ncbi.nlm.nih.gov/structure/cdd/cdd.shtml) allows the searching of profiles derived from SMART and PFAM using a modified version of the Blast algorithm.

5. Increasing coverage: using domain databases as tools in the discovery of new domains

Automatic clustering procedures can be combined with curated domain databases to facilitate the detection of novel domains. Sequence libraries can first be screened for known domains from libraries such as SMART or PFAM. Any se-

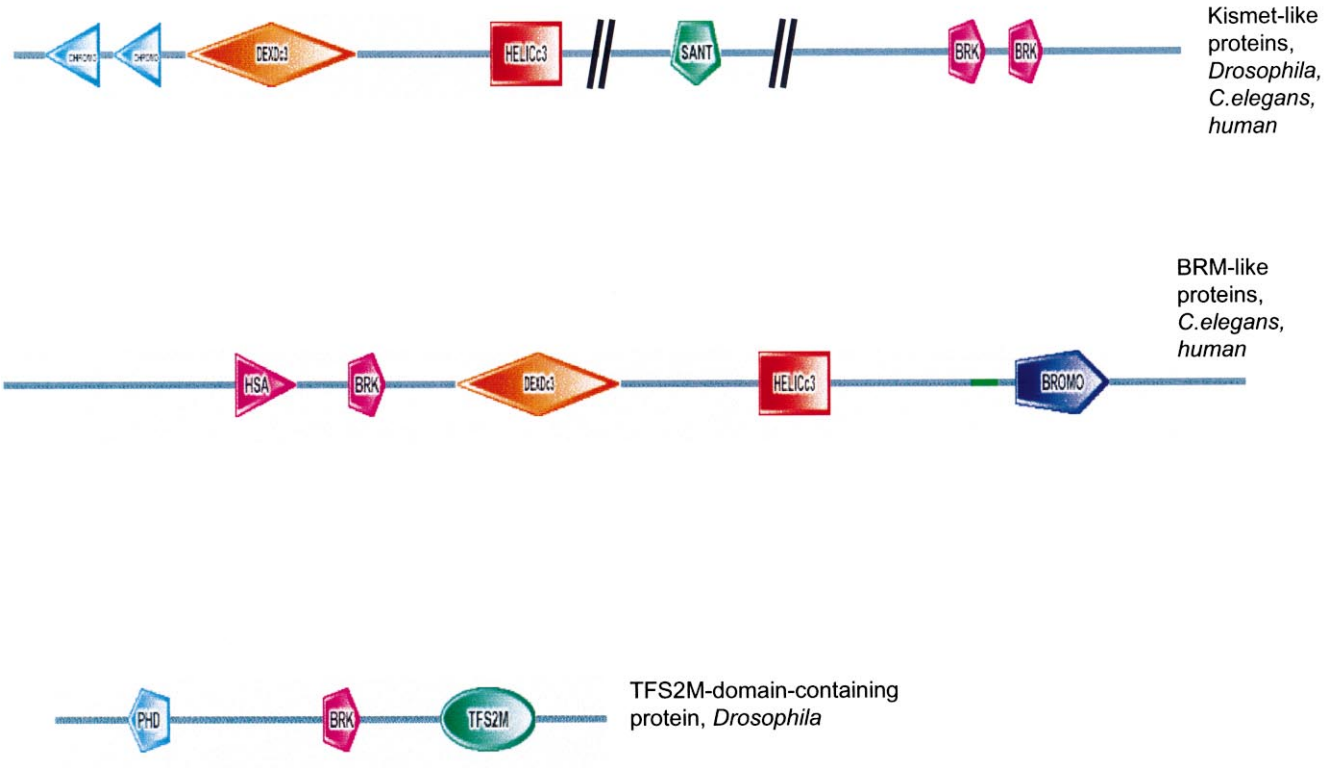

Fig. 3. Domain architectures of proteins containing the BRK domain. The domain is predicted to have a nuclear localization based on co-occurrence with other domains found in nuclear proteins. Only proteins with distinct modular organizations are shown. The domain names are according to the SMART database (http://smart.embl-heidelberg.de). First architecture: Kismet-like protein KIAA1416 (Acc. Nr. Q9P2D1), second architecture: BRM-like protein BRM (Acc. Nr. P51531), third architecture: TFS2M-domain containing protein CG6525 (Acc. Nr. Q9VG78)

quence or region of sequence that does not fall into one of these domain families can then be subjected to automatic homology searching and domain clustering procedures. The ProDom database, for instance, uses the manually curated sequence families contained in PFAM as seeds, and in turn automatically generated ProDom families are used within PFAM, as a source both of additional annotation, and ultimately, new hand-curated models [22,24]. A necessity of such an approach is that close to the full length of known domains is modelled in the curated database (i.e. the database does not simply model a conserved motif within the domain), such that the regions subjected to automatic clustering are distinct from the known domains.

Excluding known protein domains reduces the size of the sequence database that needs to be searched when hunting for new domains, which can serve to highlight hits that were previously of marginal significance. At the same time, if the sequence database searched is cross-referenced to known domains, it is an easier job to identify regions of sequence that are found within different domain contexts. We have recently used such a strategy to search systematically for novel domains associated with nuclear functions leading to the identification of 28 new domains [44]. Although it is possible to automate large parts of the domain detection procedure, and produce a list of candidate domains, large amounts of manual work are involved in distilling useful biological knowledge from the literature, and producing high quality sequence alignments. In particular, automatically identifying distinct domain contexts for homologous regions of sequence (a necessity for the detection of mobile domains) is hampered by problems such as failure to identify all instances of known domains in the source database and erroneous fusions of domains caused by aberrant gene predictions [44].

Systematic large-scale searches to identify novel domains within the completely sequenced genomes of eukaryotes have not been widespread. Hutter and colleagues searched for novel extracellular modules in the predicted set of *Caenorhabditis elegans* proteins, and identified 20 apparently nematode specific motifs [25]. We have recently analyzed all statistically significant repeats within proteins occurring in the *Drosophila* genome, leading to the characterization of 41 repeat or domain families [43]. It is likely that domain-based annotation of eukaryotic genomes would greatly benefit from more such systematic studies.

As long as new sequence is added to databases, it will be necessary to re-visit the annotation of sequences, even within complete genomes, to see if additional domains can be detected by new comparisons. Understanding the domain structure of proteins within completed genomes is vital for a better understanding of the evolutionary forces and emerging functions shaping genomes.

6. Towards complete coverage of domains

At first sight, the obvious goal of computational domain identification might be to ensure that every residue of every sequence is annotated as falling within some domain family. The current release of PFAM (6.6), probably the most comprehensive individual database, contains 3071 models covering 70% of sequences, but only 50% of residues from the sequence database that it is constructed from (Alex Bateman, pers. comm.). Coverage of completed genomes is lower than coverage of non-redundant sequence databases as the contents of non-redundant protein sequence databases are biased towards well-studied proteins [26] (typically around 30% of residues from eukaryotic genomes are currently assigned to PFAM families). The true coverage of globular protein domains may be higher than these numbers suggest. Analysis of the same dataset used to construct PFAM 6.6 shows that 2% of residues are predicted to be coiled-coil regions, and 7% low complexity, with biased amino acid composition. Complete genomes show a similar proportion of non-globular sequence, suggesting that a more realistic target for domain coverage is closer to 90% of residues than 100%.

The fraction of residues that, on the basis of sequence similarity, can be assigned to the structure-based superfamilies of scop [27] ranges from 29% in *C. elegans* to 56% in the parasitic bacteria *Buchnera* [28] (see http://stash.mrc-lmb.cam.ac.uk/). As these numbers are comparable to those found for sequence-based classifications, and as the sequence- and structure-based classifications are partially non-overlapping (i.e. not all sequence-based families have known structures, and not all known structures are used in the construction of sequenced-based families), this suggests that the overall level of possible family-based annotation, if all sources are used, is higher still.

In practice, the maximum realistically achievable level of coverage by sequence-based classifications is likely to be lower than those based on structures. Sequences linking domains together, or found at the N- and C-termini of proteins, are not necessarily distinct globular domains themselves, but can form linker regions specific to particular domain combinations. Sequence boundaries of domains are not necessarily well conserved; the conserved core modelled in a sequence alignment may not represent the full extent of a domain in any given case, again lowering the overall residue coverage. The lack of conservation of these two types of sequence (i.e. domain linker regions and family-specific domain extensions) makes their inclusion in sequence-based classifications problematic. It is difficult to estimate the overall percentage of sequence that is likely to be found in such regions and so get an estimate of the target level of residue coverage of sequence-based classification. Deciding between what is linker and what is domain is only possible on the basis of 3D structure, but clearly 3D structures represent a very biased sample of all protein sequence, and have been chosen specifically for their domain-like properties. One might expect regions of long, unstructured linker between domains to be fairly uncommon in nature – long linker regions would allow relatively independent movement of the domains they linked, in which case there would be few obvious reasons for the two domains to be in the same polypeptide in the first place (simultaneously binding distant sites on DNA or RNA is a possibility, for instance).

7. The values of structures

The interplay between sequence-based analysis of domain families, and the study of 3D structure is great. Protein domains identified by sequence have long been targeted for further structural analyses, and sequence-based identification of conserved domains and their boundaries is often a prelude to structure determination, and hence better understanding of function (Fig. 2). With the advent of structural genomics ini-

tiatives aiming to determine the structures of representatives of all proteins, accurate and thorough sequence-based domain identification will assume a new significance, if overall workload is to be minimized [26].

At the same time, structure determination highlights the limits of sequence-based analysis. The multi-domain nature of many proteins is only recognized after the structures have been solved. For instance, structure determination of the the FERM domain (band 4.1), revealed that the domain was in fact constituted from three distinct structural domains, all with similarity to other independently occurring domains [29]. 3D structure is also crucial for revealing how domains interact, either in the same polypeptide chain, or in complexes. The 3D structure of src tyrosine kinase, for instance, reveals how the SH3, SH2 and tyrosine kinase domains interact via the tail of the tyrosine kinase to inhibit activity [30].

As well as the intrinsic value of structures for rationalizing the function of protein domains, they provide a deeper framework within which to understand domain evolution. Comparison of structures can reveal evolutionary relationships between domains when no significant sequence similarity is detectable. Thus, for examples, the ephrin ectodomain was recently shown to be a distant homolog of plant phytocyanins [31] and the structure of the cytokine interleukin-17 showed it to be a distant homolog of the cystine-knot family [32]. Not all such identifications lead to insights into function, however, the G2 domain of nidogen shows clear similarity to green fluorescent protein, but there do not appear to be any immediate functional consequences of this relationship [33].

Although structural, as opposed to sequence evidence is usually the arbiter in the interpretation of domain evolution, an interesting recent example has shown the opposite. The KH domain, initially identified on the basis of sequence profiles, has recently been demonstrated to adopt two different 3D folds [34]. In general, such results are likely to be unusual, but they have profound consequences for our views of protein evolution [35].

In general, structural data provide a rich complement to sequence analyses, providing information on the evolutionary history of domains, the interactions between them, and detailed information about which residues are responsible for function. Providing better links between the broad brush, but high coverage, of sequence analysis and the fine-grained detail that structure provides will be to the benefit of both approaches.

8. Future directions

The ultimate reason for delineating protein domain families is to better understand protein function. By isolating regions that are conserved in different proteins, we identify common elements that can be valuable for further study. The PX domain, for example, was first described in 1996 [36], but only in the past few months has its function been better characterized (see [37] for a brief review). The basic premise that conserved domains are likely to be functionally important leads to the corollary that less taxonomically widespread will not have such universally important functions. On the other hand, domains found in limited taxonomic ranges are likely to be no less interesting in as much as they will be responsible for organism-specific biology. Given this, it is reasonable to assume that continued efforts will eventually lead to the delineation of all the conserved protein domains and families within completed genomes.

Beyond the identification of members of a particular domain family, much work remains in developing methods that can help make more detailed predictions about the differences in function between family members. Initial function assignments for a domain family are often made based on the known functions of a protein, and the functions of other domains that co-occur with the new domain. At a simple level, for example, domains can be predicted to have a nuclear localization based on co-occurrence with other domains known to be found in proteins localized to the nucleus ([44] and Fig. 3). Use of this kind of contextual information could be formalized and expanded to make more valuable detailed predictions that need apply only to specific members of the family. Another promising approach is the comparison of patterns of residue conservation across different subfamilies within a multiple sequence alignment [38], a procedure that is greatly facilitated by reference to 3D structure [39,40]. Development of these and similar techniques should also benefit from the current focus on large-scale studies of protein–protein interactions [41,42]. Integrating such studies within the hierarchies provided by the results of structural genomics, structural analysis and genome sequencing will provide a rich framework for understanding the diversity and evolution of protein and domain function.

Acknowledgements: We thank Drs. Alex Bateman (The Wellcome Trust Sanger Institute) and Chris Ponting (The MRC Functional Genetics Unit, Oxford) for helpful discussions.

References

[1] Chothia, C. (1992) Nature 357, 543–544.
[2] Wolf, Y.I., Grishin, N.V. and Koonin, E.V. (2000) J. Mol. Biol. 299, 897–905.
[3] Janin, J. and Chothia, C. (1985) Methods Enzymol. 115, 420–430.
[4] Andrade, M.A., Perez-Iratxeta, C. and Ponting, C.P. (2001) J. Struct. Biol. 134, 117–131.
[5] Lupas, A.N., Ponting, C.P. and Russell, R.B. (2001) J. Struct. Biol. 134, 191–203.
[6] Ogata, H., Audic, S., Barbe, V., Artiguenave, F., Fournier, P.E., Raoult, D. and Claverie, J.M. (2000) Science 290, 347–350.
[7] Hegyi, H. and Gerstein, M. (2001) Genome Res. 11, 1632–1640.
[8] Lander, E.S. et al. (2001) Nature 409, 860–921.
[9] Ponting, C.P. and Dickens, N.J. (2001) Genome Biol. 2.
[10] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) Cambridge University Press, Cambridge.
[11] Ponting, C.P., Schultz, J., Copley, R.R., Andrade, M.A. and Bork, P. (2000) Adv. Protein Chem. 54, 185–244.
[12] Mott, R. (2000) J. Mol. Biol. 300, 649–659.
[13] Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. (2001) Nucleic Acids Res. 29, 2994–3005.
[14] Spang, R. and Vingron, M. (2001) Bioinformatics 17, 338–342.
[15] Sonnhammer, E.L. and Kahn, D. (1994) Protein Sci. 3, 482–492.
[16] Park, J. and Teichmann, S.A. (1998) Bioinformatics 14, 144–150.
[17] Guan, X. and Du, L. (1998) Bioinformatics 14, 783–788.
[18] Gouzy, J., Corpet, F. and Kahn, D. (1999) Comp. Chem. 23, 333–340.
[19] Enright, A.J. and Ouzounis, C.A. (2000) Bioinformatics 16, 451–457.
[20] Heger, A. and Holm, L. (2001) Bioinformatics 17, 272–279.
[21] Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P. and Bork, P. (2000) Nucleic Acids Res. 28, 231–234.
[22] Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000) Nucleic Acids Res. 28, 263–266.
[23] Apweiler, R. et al. (2001) Nucleic Acids Res. 29, 37–40.

[24] Corpet, F., Servant, F., Gouzy, J. and Kahn, D. (2000) Nucleic Acids Res. 28, 267–269.
[25] Hutter, H. et al. (2000) Science 287, 989–994.
[26] Vitkup, D., Melamud, E., Moult, J. and Sander, C. (2001) Nat. Struct. Biol. 8, 559–566.
[27] Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G. and Chothia, C. (2000) Nucleic Acids Res. 28, 257–259.
[28] Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001) J. Mol. Biol. 313, 903–919.
[29] Pearson, M.A., Reczek, D., Bretscher, A. and Karplus, P.A. (2000) Cell 101, 259–270.
[30] Xu, W., Harrison, S.C. and Eck, M.J. (1997) Nature 385, 595–602.
[31] Toth, J., Cutforth, T., Gelinas, A.D., Bethoney, K.A., Bard, J. and Harrison, C.J. (2001) Dev. Cell 1, 83–92.
[32] Hymowitz, S.G. et al. (2001) EMBO J. 20, 5332–5341.
[33] Hopf, M., Gohring, W., Ries, A., Timpl, R. and Hohenester, E. (2001) Nat. Struct. Biol. 8, 634–640.
[34] Grishin, N.V. (2001) Nucleic Acids Res. 29, 638–643.
[35] Grishin, N.V. (2001) J. Struct. Biol. 134, 167–185.
[36] Ponting, C.P. (1996) Protein Sci. 5, 2353–2357.
[37] Prehoda, K.E. and Lim, W.A. (2001) Nat. Struct. Biol. 8, 570–572.
[38] Hannenhalli, S.S. and Russell, R.B. (2000) J. Mol. Biol. 303, 61–76.
[39] Lichtarge, O., Bourne, H.R. and Cohen, F.E. (1996) J. Mol. Biol. 257, 342–358.
[40] Aloy, P., Querol, E., Aviles, F.X. and Sternberg, M.J. (2001) J. Mol. Biol. 311, 395–408.
[41] Uetz, P. et al. (2000) Nature 403, 623–627.
[42] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) Proc. Natl. Acad. Sci. USA 98, 4569–4574.
[43] Ponting, C.P., Mott, R., Bork, P. and Copley, R.R. (2001) Genome. Res. 11, 1996–2008.
[44] Doerks, T., Copley, R.R., Ponting, C.P. and Bork, P. (2002) Genome Res., in press.