Using the CATH domain database to assign structures and functions to the genome sequences


*Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, U.K., †Helix Research Institute, 1532-3 Yana, Kisarazu-Shi, Chiba, 292, Japan, and †Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, U.K.

Abstract

The CATH database of protein structures contains ~180000 domains organized according to their (C)lass, (A)rchitecture, (T)opology and (H)omologous superfamily [1]. Relationships between evolutionary related structures (homologues) within the database have been used to test the sensitivity of various sequence search methods in order to identify relatives in Genbank and other sequence databases [2]. Subsequent application of the most sensitive and efficient algorithms, gapped blast and the profile based method, Position Specific Iterated Basic Local Alignment Tool (PSI-BLAST) [3], could be used to assign structural data to between 22 and 36% of microbial genomes in order to improve functional annotation and enhance understanding of biological mechanism. However, on a cautionary note, an analysis of functional conservation within fold groups and homologous superfamilies in the CATH database, revealed that whilst function was conserved in nearly 55% of enzyme families, function had diverged considerably, in some highly populated families. In these families, functional properties should be inherited far more cautiously and the probable effects of substitutions in key functional residues carefully assessed.

Introduction

There are nearly 20000 known domain structures in the Protein Databank [4] now held in the Research Collaboratory for Protein Structures at Rutgers University. These data still lag considerably behind the known sequences (~400000 currently in Genbank). However, with the advent of the structure genomic initiatives [5] we can expect the numbers to increase substantially and there are suggestions that we may know all the major folds in nature within the next 5 years. Once their structures have been determined, interest will focus on methods for assigning functional properties to these proteins.

In order to recognize and understand structural and functional relationships between proteins, we have clustered all the known structures into fold groups and evolutionary superfamilies using a combination of automatic and manual

References


Received 1 September 1999
methods. Sequences are first compared to identify close homologues (≥ 35% sequence identity). Methods used include standard pairwise sequence alignment [6] and more recent profile-based alignment algorithms (PSI-BLAST) [3]. Because structure is much more conserved than sequence, very distant homologues are identified using pairwise structure comparisons (SSAP) [7] or three-dimensional profiles (CORA) [8] described below.

Recognized fold groups and families are stored in the CATH database [1], so-called because the organization of the database reflects the hierarchy of protein (C)lass, (A)rchitecture, (T)opology or fold group, and (H)omologous superfamily (see Figure 1). Class depends on the proportion of residues in α-helical or β-strand conformations, whilst architecture simply describes the arrangement of the secondary structures in three dimensions, regardless of their connectivity. The major classes in CATH are mainly-α, mainly-β and α-β, and there are currently 32 different architectures described among these classes.

Because protein domains are considered to be important evolutionary units and are often implicated in gene duplication and swapping, the CATH database and related resources have chosen to cluster proteins on the basis of similarities between their domain folds. This raises the technical difficulty of domain fold recognition and boundary assignment. Although many algorithms have been developed for automatically identifying domains, given a protein’s three-dimensional coordinates, none achieves more than an ~ 80% success rate, which means that most assignments should be checked. For CATH, we apply three independent methods (DETECTIVE [9], PUU

---

Figure 1

CATHerine wheel plot showing the distribution of non-homologous structures [i.e. a single representative from each homologous superfamily (H-level) in CATH] among the different classes (C), architecture (A) and fold families (T) in the CATH database.

Protein classes shown are mainly-α (Non-bundle and Bundle), mainly-β (Roll adjacent to Bundle, Barrel and Sandwich) and αβ (Roll, Barrel, 2-layer Sandwich, 3-layer Sandwich and Complex). Within each class, the angle subtended for a given segment reflects the proportion of structures within the identified architectures (inner circle) or fold families (outer circle). The superfold families are indicated and are illustrated with a MOLSCRIPT [31] drawing of a representative from the family.
Evolution of Sequences, Structures and Genomes

and DOMAK [11]) and if all three agree within reasonable limits, we assign the boundaries automatically [12]. Otherwise, each assignment is checked manually to determine the correct boundaries. Because of the large proportion of multidomain proteins (more than one-third of known structures are multidomain and one-third of these comprise discontiguous segments of the polypeptide chain [12]), this can be a very time-consuming process.

Several other protein structure databases exist, similarly organized by domain fold group and evolutionary family. In the SCOP database in Cambridge [13], structures are compared manually and assigned to homologous families on the basis of common functional or structural properties (e.g. unusual $\beta$-bulges or conserved turns) which suggest an evolutionary link. Similarly, the CAMPASS and related HOMSTRAD databases [14] provide alignments for about 130 protein homologous structural families identified using a combination of manual and automatic methods. By contrast, the DALI Domain Dictionary (DDD) [15] is generated automatically using the DALI structure comparison algorithm [16]. DALI also provides neighbour lists for each known structure which identify related proteins from close homologues through to proteins sharing common structural motifs (e.g. $\beta\alpha\beta$ motifs). A related resource ENTREZ [17] has been set up at the NCBI in the U.S.A.

Below we describe some recently developed methods for improving the assignment of homologous protein structures in the CATH database. Techniques have also been established for identifying relatives in the genome sequences, and these additional data will in turn further improve our detection of remote homologues. Finally, we consider how well functional properties are conserved within different fold groups and homologous families in CATH. The results have implications for the structural genome initiatives which aim to improve the functional annotation of genome sequences by determining structural data for all the major protein families.

**New methods for assigning homologous proteins in the CATH domain database**

Distant structural homologues are assigned to CATH fold groups and superfamilies using the SSAP structure comparison algorithm [7]. This returns a normalized score in the range of 0–100, independent of the sizes of the proteins. Protein pairs scoring above 70 have very similar folds, whilst scores of 80 and above are often found to be homologous, exhibiting common or related functions. Before being assigned to the same homologous superfamily, further evidence of evolutionary origin is sought. Information on protein function is examined, both from the literature and public databases (e.g. SWISS-PROT, Enzyme Database).

**Three-dimensional structural profiles to detect very distant homologues**

We have recently developed a more sensitive procedure for recognizing very distant evolutionary relatives. This used three-dimensional profiles for each homologous superfamily in the CATH database, generated from multiple alignments of selected relatives in each family, using the program CORA [8]. CORA selects the most distant relatives from the family which retain sufficient structural similarity (SSAP similarity score $\geq 75$) to obtain a reliable multiple alignment. Conserved structural positions and their associated structural properties (e.g. average accessibility and distances to other positions in the protein) are identified from the alignment and encoded in a three-dimensional template which can be scanned.

**Figure 2**  
Pie chart showing the result of assigning 2646 recently deposited protein domains into homologous superfamilies in the CATH database

64% of domains were classified using automatic sequence comparison methods (sequence identity $> 35$%), a further 9.8% were identified using PSI-BLAST, 2.9% were classified using combined sequence/structure criteria (SSAP $> 80$, sequence identity $> 25$%), 4.6% of domain were assigned using structural similarity (SSAP $> 70$) and DHS to manually validate functions. 6.9% of folds were new superfamilies and 11.8% novel folds.
against any newly determined structures to identify probable homologues. Tests showed that the templates were significantly more sensitive at recognizing homologues than pairwise comparison methods. Since one template replaces several representatives for many CATH families, considerable increases in speed, of between 20- and 100-fold, were obtained. This should enable CATH to keep pace with the flood of new structures expected from the structure genomic initiatives [S].

Validating homologues by comparison of functional properties

In order to facilitate the process of validating evolutionary relationships and to provide consensus, structural and functional data for each superfamily in CATH; we recently set up a Dictionary of Homologous Superfamilies (DHS). This is a Web-based resource which summarizes functional data for each protein family containing more than two non-identical relatives. Figure 2 shows the proportion of recently determined structures which could be assigned to homologous families on the basis of sequence or structure and functional similarity. Interestingly, only about 11% of the dataset were found to be unique folds, supporting the common hypothesis of a limited number of folds available in nature [18,19].

One of the most powerful methods for validating homologues uses data on protein–ligand interactions. These are identified from the three-dimensional coordinates by the program GROW [20] and can be shown in a diagrammatic representation (DOMPLOT) [21] which highlights sequence motifs extracted from PROSITE [22] often associated with functionally important residues. DOMPLOT can be used to illustrate the position of conserved ligand-interacting residues across a family of related structures. Figure 3 shows an example for a group of flavin-binding TIM barrel structures. Despite having TIM barrel folds, the evolutionary ancestry of these proteins was unclear as their sequence identities...

---

**Figure 3**

DOMPLOT diagram [21] for the output of a CORA [8] multiple structural alignment of five FMN-binding TIM barrels

Abbreviations: Igox, glycolate oxidase; IdorA, dihydro-orotate oxidase; IfcbA, flavocytochrome b2; 2tmdA, trimethylamine dehydrogenase; Ioya, NADPH dehydrogenase. (For references to the individual structures, see their entry in the PDB.) The secondary structure consensus is given at the top of the plot, and E and H correspond to β-strand and α-helix respectively. The diagram is analogous to a multiple sequence alignment, but is derived from structural information, with a gap implying no structural equivalence at that position. Excluding lines in the secondary structure consensus vertical grey lines, with residue numbers directly below, correspond to residues interacting with the cofactor FMN. The ligand interaction consensus indicates the presence of ten completely conserved residue interactions with FMN and suggests a common FMN-binding ancestor for all five proteins.

---

© 2000 Biochemical Society
were below 15\% and their enzyme classifications differed, suggesting unrelated functions. However, GROW analysis showed that substrates always bound in the same place, at the C-terminal ends of the $\beta$-strands in a cavity at the base of the $\beta$-barrel. Furthermore, DOMPLOT revealed ten completely conserved ligand interactions with the common cofactor flavin mononucleotide (FMN). The possibility of both fold similarity and common ligand positions occurring by chance is extremely low and suggests an evolutionary link between these proteins.

**Population of fold groups and evolutionary superfamilies in CATH**

There are 18000 domain structures in the latest release of CATH (release 1.6, June 1999). Figure 1 shows the population of different levels in the hierarchy. Among the 670 fold groups currently recognized, there are 1200 homologous superfamilies. The CATH wheel shown in Figure 1 is shaded according to class and shows that a large proportion of CATH superfamilies (nearly 60\%) belong to one of six simple architectures ($\alpha$-bundle, $\beta$-sandwich, $\beta$-barrel, $\alpha$$\beta$-2 and 3 layer sandwiches, and $\alpha$$\beta$-barrel). Furthermore, some of these architectures contain highly favoured folding arrangements, which we describe as superfolds [19] because of the diversity of sequences and functions they support. More than one-third of superfamilies in CATH belong to the superfolds.

**Assigning structural data to homologous sequences in the genomes**

Despite the exponential increases in the numbers of known protein structures, the number of distinct fold groups has increased much more slowly, with the number currently less than 700. This supports the hypotheses of Chothia [18] and other workers [19] who show that due to the physical constraints on secondary structure packing, there may only be a few thousand folds available in Nature. Bearing this in mind, we may soon have structural representatives for many of the major protein families and it is interesting to see what proportion of genome sequences can currently be assigned to the known structural families. To find out, we must apply highly sensitive sequence search algorithms to recognize as many distant homologues as possible.

Pairwise sequence comparison methods (e.g. [6,23]) are only reliable to about 30\% sequence identity between proteins. More distant relatives are generally recognized using profile-based methods. In order to assess the performance of several sequence search strategies we used a sequence dataset based on CATH superfamilies. This allowed us to identify the most sensitive methods and determine the most reliable thresholds for using them. Interestingly, our results showed that the pairwise gapped blast algorithm was as sensitive as Smith-Waterman for a tenfold increase in speed. All the pairwise methods could be improved by using intermediate sequences [24], whereby two extremely distant homologues scoring below the cut-off can be linked by significant matches to a single intermediate. However, the recently developed PSI-BLAST [3], an interactive profile-based method which generates a position-specific score matrix from a multiple alignment of relatives, was clearly the most sensitive of all the approaches we tested, identifying nearly 70\% of all the homologous pairs in the CATH dataset.

Using the optimal parameters established in these trials, we subsequently developed a protocol for both PSI-BLAST and gapped BLAST to find relatives to all the CATH structural superfamilies in 11 microbial genomes. Gapped BLAST is much faster than PSI-BLAST and can be used in a first-pass strategy to identify close homologues in the genomes, before attempting to capture very remote homologues with PSI-BLAST. Using this approach, we assigned structural data to between 22 and 36\% of the sequences in each genome [2]. Similar statistics have been obtained by other groups [25,26]. Structure prediction techniques, such as one-dimensional-three-dimensional profiles [27] or threading [28], can detect even more distant relationships and assign folds to nearly 50\% of microbial genomes or 26\% of the yeast genome.

The value of mapping sequence to structure data lies in opportunities to inherit or infer functional properties or to improve our understanding of such properties by considering their structural context. However, it is well known in the sequence community that functional assignment by inheritance can often be inaccurate [29]. We decided, therefore, to review the correlation between protein structure and function using the CATH database and associated functional data from the SWISS-PROT and the Enzyme Databases.
Functional conservation in protein folds and evolutionary families

Our analysis was based on enzyme families in CATH, as this allowed us to easily compare functional properties on the basis of their enzyme classification numbers extracted from the Enzyme Database [30]. We found that the majority of fold groups (> 95%) contain protein families with common functional properties. However, the superfolds exhibited considerable functional variation, in particular the TIM barrel folds with more than 40 different observed functions and the Rossmann folds with more than 50 different functions. Considering homologous superfamilies, nearly 55% had completely conserved functions. In a further 20%, function was conserved to the third EC number, which generally meant that the substrate had changed but the active site and catalytic mechanism were conserved. In 10% of families the function had changed completely.

These observations have important implications for the genome projects, and suggest that in some fold groups and families great care must be taken when inheriting functional properties from a structural relative, although it will obviously depend on how distant the relatives are. However, by considering the variation in functional properties across the family, it is possible to assess how reliably functional characteristics might be shared. Furthermore, by considering ligand interaction data, it is sometimes possible to assess whether relatives are using common active sites and hence make reasonable predictions regarding the possible changes in function arising from mutations in key residues within these sites.

Conclusions

The CATH database of protein domain structural families has been set up as a Web-based resource, which can be used for examining structural and functional properties in protein families for which structures have been determined. By incorporating functional data from other public databases (e.g. SWISS-PROT, ENZYME Database) and information on key residues involved in function, we have been able to analyse the correlation between structure and function for enzyme families in CATH. We were also able to identify those families where the structural frameworks tolerate extensive substitutions and support a wide range of different functions. Inheriting functions for genome sequences assigned to these fold groups and families should be done cautiously as possible.

C.A.O. and F.P. acknowledge the support of the Medical Research Council. J. E. B. and A. E. T. acknowledge the support of the Biotechnology and Biological Sciences Research Council and Oxford Molecular Ltd.

References

9 Swindells, M. B. (1993) Protein Sci. 4, 103-112
Origins of AIDS viruses: estimating the time-scale

P. M. Sharp*, E. Bailes†, F. Gao‡, B. E. Beer‡, V. M. Hirsch‡ and B. H. Hahn†

*Institute of Genetics, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH, U.K., †Departments of Medicine and Microbiology, University of Alabama at Birmingham, Birmingham, Alabama 35294, U.S.A., and ‡Laboratory of Molecular Microbiology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, Maryland 20852, U.S.A.

Abstract

The primate lentiviruses comprise SIV strains from various host species, as well as two viruses, HIV-1 and HIV-2, that cause AIDS in humans. The origins of HIV-1 and HIV-2 have been traced to cross-species transmissions from chimpanzees and sooty mangabey monkeys respectively. Two approaches have been taken to estimate the timescale of the evolution of these viruses. Certain groups of SIV strains appear to have evolved in a host-dependent manner, implying a time-scale of many thousands or even millions of years. In stark contrast, molecular clock calculations have previously been used to estimate a time-scale of only tens or hundreds of years. Those calculations largely ignored heterogeneity of evolutionary rates across different sites within sequences. In fact, the distribution of rates at different sites seems extremely skewed in HIV-1, and so the time-depth of the primate lentivirus evolutionary tree may have been underestimated by at least a factor of ten. However, these date estimates still seem to be far too recent to be consistent with host-dependent evolution.

Introduction

Two distinct retroviruses, human immunodeficiency virus types 1 and 2 (HIV-1 and HIV-2), cause the acquired immune deficiency syndrome (AIDS). Together with related simian immunodeficiency viruses (SIVs), found in other primate species, these comprise the primate lentiviruses. These viruses exhibit an extraordinary degree of sequence diversity and provide an interesting and challenging model system in which to study molecular evolution. Furthermore, the results obtained provide insights into the origins of AIDS, and have important practical implications for understanding and tackling the ongoing AIDS pandemic.

Origins of AIDS viruses

Lentiviruses have been isolated from numerous primate species. Phylogenetic analyses indicate that these viruses fall into five major, approximately equidistant, lineages (Figure 1). The two viruses causing AIDS in humans belong to different lineages and represent recent cross-species transmissions from different sources. The vast majority of cases of HIV infection worldwide are due to strains of HIV-1. These viruses cluster with SIVcpz from chimpanzees (Pan troglodytes). Although SIVcpz was first reported in 1990 [1], there remained the question of whether chimpanzees had, like humans, recently acquired the virus from some other species [2,3]. However, we have recently presented evidence strongly suggesting that one particular subspecies of chimpanzee, P. t. troglodytes, does indeed constitute the source of HIV-1 [4].

HIV-2 is only common in west Africa. Strains of HIV-2 cluster with SIVsm from sooty mangabeys (Cercocebus atys), and with SIVmac from macaques (Figure 1). The macaque viruses were among the first SIV strains to be characterized, but it was soon realized that macaques in the wild are not infected with SIV, and that all of the isolates reflect transmission in captivity. In contrast, sooty mangabeys have been shown to be naturally infected with SIV. These monkeys inhabit west Africa, and represent the source of HIV-2.

Both HIV-1 and HIV-2 exhibit considerable genetic diversity. HIV-1 strains have been classified into three groups (M, N and O), with the M group further divided into numerous subtypes (A–J, so far). The phylogenetic interspersion of the SIVcpz and HIV-1 lineages indicates that the