SCOPEC: a database of protein catalytic domains

Richard A. George1,2,∗, Ruth V. Spriggs1,2, Janet M. Thornton2, Bissan Al-Lazikani1 and Mark B. Swindells1

1Inpharmatica Ltd, 60 Charlotte street, London W1T 2NU, UK and 2European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: Domains are the units of protein structure, function and evolution. It is therefore essential to utilize knowledge of domains when studying the evolution of function, or when assigning function to genome sequence data. For this purpose, we have developed a database of catalytic domains, SCOPEC, by combining structural domain information from SCOP, full-length sequence information from Swiss-Prot, and verified functional information from the Enzyme Classification (EC) database. Two major problems need to be overcome to create a database of domain–function relationships; (1) for sequences, EC numbers are typically assigned to whole sequences rather than the functional unit, and (2) The Protein Data Bank (PDB) structures elucidated from a larger multi-domain protein will often have EC annotation although the relevant catalytic domain may lie elsewhere.

Results: SCOPEC entries have high quality enzyme assignments; having passed both computational and manual checks. SCOPEC currently contains entries for 75% of all EC annotations in the PDB. Overall, EC number is fairly well conserved within a superfamily, even when the proteins are distantly related. Initial analysis is encouraging; suggesting that there is a 50:50 chance of conserved function in distant homologues first detected by a third iteration PSI-BLAST search. Therefore, we envisage that a knowledge-based approach to function assignment using the domain–EC relationships in SCOPEC will gain a marked improvement over this base line.

Availability: The SCOPEC database is a valuable resource in the analysis and prediction of protein structure and function. It can be obtained or queried at our website http://www.enzome.com

Contact: richardg@inpharmatica.co.uk

1 INTRODUCTION

New challenges in genome sequence analysis include obtaining functional assignment for the proteome. Such assignment will improve our understanding of many disease processes and assist the development of new drugs. The most popular method for assigning function to a novel sequence requires the identification of homologous sequences with prior functional annotation. However, such methods at best lack precision in their assignment and at worst can be misleading (Iyer et al., 2001). It has been suggested that successful functional assignment is limited to sequences that share a sequence identity >40% (Todd et al., 2001) even though methods such as PSI-BLAST can identify homologues with identities as low as 15% (Altschul et al., 1997).

Knowing the three-dimensional (3D) structure of a protein reveals the molecular details of binding, catalysis and signalling (Thornton et al., 2000). The number of known protein structures in the Protein Data Bank (PDB) (Berman et al., 2000) is currently 23,000 consisting of over 700 unique domain folds (Orengo et al., 1997). Combined with other data, these structures reveal much about the specific molecular mechanisms that underlie biological function (Sheinerman et al., 2003). Structural knowledge has been used to identify and annotate distant relatives. The identification steps can be automated through the application of threading (Jones, 1999) or profile based approaches (Grishkov et al., 1987; Luthy et al., 1994). However, the detailed information derived from 3D structure that illuminates residues important for activity and structural stability, is generally applied in a subsequent manual analysis.

Our focus in this work is the construction of a database of protein structures with known catalytic function that can be utilized by automated methods to robustly assign enzyme function. In this study, our main interest is in the functional information held within the Enzyme Classification (EC) scheme, maintained by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB, 1992). Each enzyme function in the EC scheme is assigned a four-digit code, which represents a hierarchy of functional classification in which the first digit represents a general level of function. Subsequent digits indicate more specific features of the chemical reaction through subclass,

∗To whom correspondence should be addressed.
Domains are the basic units of protein structure, function and evolution (Doolittle, 1995); therefore, it is essential to have a mapping of function to the domain. Furthermore, database search methods are more sensitive when searching at the domain level (George and Heringa, 2002). To understand which structural units are responsible for these enzyme activities, we are using the SCOP classification of structural domains (Murzin et al., 1995) and full-length sequence information from Swiss-Prot (Boeckmann et al., 2003). Like the EC scheme, SCOP is also hierarchical with all domains classified by structural class, followed by fold, superfamily and family. Domains in the same superfamily share distinctive features that suggest a common evolutionary ancestor.

Several problems have to be overcome to achieve our goal of domain–function annotation. First, EC numbers are assigned to whole sequences rather than just the functional unit. Second, single domain PDB structures elucidated from a larger multi-domain protein sequence will often have an EC annotation although the domain may not be the catalytic unit. These problems were overcome to create the SCOPEC database, a comprehensive mapping of SCOP and EC hierarchies.

2 METHODS

All programs used in the construction of the SCOPEC database were written in Perl5. The data are stored in an ORACLE database and can be queried with SQL. Where appropriate, relationship and annotation data were extracted from the Biopendium™ version 14 (Inpharmatica Ltd, London), a proteome scale annotation resource (Swindells et al., 2002).

2.1 Verifying PDB–EC annotation

Figure 1 outlines the construction of the SCOPEC database. To have confidence in the EC annotations found in the PDB, we specified that the same EC annotation must also be found in a Swiss-Prot sequence homologue of similar length. In addition, this verification permitted the correct functional assignment of a protein chain within a multi-chain structure, avoiding the assignment of an EC to a non-enzymatic sequence. The length criterion also prevented errors arising from the inclusion of non-enzymatic single domain structures that have been structurally solved as smaller constructs from larger multi-domain proteins.

All Swiss-Prot proteins with EC assignments were used to query the Biopendium™. We required that PDB sequences returned from PSI-BLAST queries within three iterations had >80% of the PDB sequence aligned to the parent Swiss-Prot sequence, and that there was a difference in length of no more than 70 residues between the two sequences. These criteria ensured that the PDB sequence mapped entirely to the Swiss-Prot sequence: this allows for an average of 35 residues difference at the N- and C-termini and reduces the possibility that the PDB sequence has been truncated. The ‘parent’ Swiss-Prot sequence and its PDB homologue were required to have the same EC assignment up to the second subclass (the first three digits are identical). All resulting PDB–EC mappings were stored in an ORACLE database.
2.2 Identifying catalytic domains in the PDB

Although we now have confidence in the EC assignments of proteins from the PDB, we still have the problem of not knowing which of the domains in a multi-domain protein are catalytic. To address this problem, we first identified reliable assignments by generating a set of single domain proteins (Fig. 1b). All domain information for this work was taken from SCOP version 1.63. We also assigned catalytic domains by using the location of known catalytic residues as described in the Catalytic Site Atlas (CSA) maintained at the European Bioinformatics Institute (Porter et al., 2004) with an ORACLE database version implemented at Inpharmatica Ltd.

Finally, we used the domains in the verified domain set as queries in a database search of the multi-domain protein set. In this way, domains that had been previously identified were used to delineate the catalytic domain(s) of the remaining multi-domain proteins. Relationships were quickly obtained from the Biopendium™, which contains pre-calculated results from iterative database searches initiated from profiles of the domains in the SCOP database. It is important to note that we are not assigning function, but ensuring that domain–EC annotation is correct. Therefore, our criteria were that at least 70% of an annotated domain from the verified domain set must overlap with a region in a homologous protein and that both proteins must share the same first three digits of their EC numbers. Domains identified by this process were then added to the verified set.

2.3 Measuring the success of functional assignment by SCOPEC

A PSI-BLAST search of representative SCOPEC sequences was used to identify homologous Swiss-Prot entries. These Swiss-Prot hits were then used to assess the relationship between sequence and enzyme function conservation. To avoid overestimating success in this benchmark, we used a methodology similar to that proposed by Tian and Skolnick (2003) to remove redundancy in Swiss-Prot sequences. This was achieved by first dividing Swiss-Prot entries into groups containing the same EC number (all four digits). Then, using BLASTCLUST (Altschul et al., 1990) within each group, we retained a set of proteins that had no more than 40% sequence identity.

3 RESULTS

Of the 3758 enzyme activities recorded by the IUBMB (complete four-digit number) only 1884 have a sequence in the Swiss-Prot database. Of these, 976 are represented by known 3D structures in the PDB. We have succeeded in generating a high quality set of catalytic domains for 731 unique EC numbers for the SCOPEC database. SCOPEC also includes 41 distinct ECs where the fourth digit is not assigned; these are likely to cover additional EC activities.

<table>
<thead>
<tr>
<th>SCOP</th>
<th>EC</th>
<th>4th level</th>
<th>3rd level</th>
<th>2nd level</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family x.x.x</td>
<td>ECS</td>
<td>1.7 ± 2.0</td>
<td>1.2 ± 0.6</td>
<td>1.1 ± 0.5</td>
<td>1.1 ± 0.3</td>
</tr>
<tr>
<td></td>
<td>S/EC</td>
<td>1.3 ± 0.2</td>
<td>4.6 ± 5.5</td>
<td>13 ± 15</td>
<td>100 ± 61</td>
</tr>
<tr>
<td>Superfamily x.x</td>
<td>EC/S</td>
<td>3.0 ± 4.6</td>
<td>1.6 ± 1.3</td>
<td>1.5 ± 1.0</td>
<td>1.2 ± 0.5</td>
</tr>
<tr>
<td></td>
<td>S/EC</td>
<td>1.2 ± 0.5</td>
<td>3.5 ± 3.6</td>
<td>9.2 ± 8.9</td>
<td>65 ± 30</td>
</tr>
<tr>
<td>Fold x.x</td>
<td>EC/S</td>
<td>4.0 ± 7.9</td>
<td>2.0 ± 2.6</td>
<td>1.8 ± 1.9</td>
<td>1.4 ± 0.8</td>
</tr>
<tr>
<td></td>
<td>S/EC</td>
<td>1.2 ± 0.5</td>
<td>3.4 ± 3.2</td>
<td>8.5 ± 8.1</td>
<td>55 ± 27</td>
</tr>
<tr>
<td>Class x</td>
<td>EC/S</td>
<td>174 ± 150</td>
<td>55 ± 28</td>
<td>28 ± 9.5</td>
<td>5.6 ± 0.5</td>
</tr>
<tr>
<td></td>
<td>S/EC</td>
<td>1.1 ± 0.4</td>
<td>1.9 ± 1.0</td>
<td>2.8 ± 1.3</td>
<td>4.7 ± 0.8</td>
</tr>
</tbody>
</table>

Two scores are shown: the number of ECs per SCOP category (EC/S) and the number of SCOP categories per EC number (S/EC). Calculations were performed on five SCOP domain classes (α, α+β, α/β, multi-domain) within the SCOPEC database. Only complete ECs were considered when comparing the ECs at the fourth level (x.x.x.x).

The DNA-directed RNA polymerases (EC 2.7.7.6) were seen to have the greatest coverage of structural types: RNA polymerase is a 12 subunit protein complex (Cramer et al., 2001) where each subunit is involved in the overall transcription of DNA into messenger-RNA, but only the RPB1 subunit contains the active site; therefore, all but the catalytic subunits for RNA polymerase were manually removed from SCOPEC.

3.1 SCOP hierarchy versus EC hierarchy

Table 1 shows the average number of unique ECs represented by each level of SCOP and the number of SCOP folds, superfamilies and families within each level of the EC hierarchy. This can be viewed in two ways. Measures of EC per SCOP approximate the effects of divergent evolution, whereas SCOP per EC reports convergent effects. At the most descriptive levels of function and structure, there are 1.73 complete EC numbers (fourth level) per SCOP family and 1.28 SCOP families per EC number. Moving down the EC levels, the number of ECs per SCOP family does not change greatly. This is also true when considering the number of ECs per SCOP superfamily and fold (although the standard deviations increase; reflecting the effects of the superfolds such as TIM barrels). Conversely, the number of SCOP folds/superfamilies/families per EC dramatically increases as the EC level becomes more general. In a similar manner, less specific SCOP levels result in a much larger number of ECs.

Figure 2 plots histograms specifically for the SCOP superfamily level. The majority of functions at the fourth level of EC (Fig. 2a) have a single structural superfamily, but this increases to an average of 3.51 superfamilies per function at the less specific third EC level (Fig. 2b). The average number of ECs (fourth level) per superfamily is 2.96 (Fig. 2c), while at the third level of EC it is 1.58 (Fig. 2d). This means that a structural superfamily is likely to have a single function with multiple substrate specificities; some superfamilies may have a large
number of functions, but they are in the minority. Conversely, a single function can be adopted by many structure types.

### 3.2 Evolution of protein structure and function

The larger number of superfamilies per function (3.51 at the third EC level) versus function per superfamily (1.58) suggests that nature re-invents function (convergent evolution). Following a reinvention, it is likely that modification leads to new specificities of function (divergent evolution); there are 2.96 ECs per superfamily at the more detailed fourth level of function. The likely mechanism of evolution for analogous enzymes is the recruitment of existing functions by a change in specificity or a modified catalytic mechanism (Galperin et al., 1998). However, most superfamilies have a single EC (third level); if a function has significantly changed then the original function has now been lost or is not present in the database. Only 27.6% of superfamilies have more than one function; in this set 28 pairs of superfamilies share two identical functions and 4 pairs share three identical functions.

There is no correlation between popular functions and functional promiscuity in a superfamily; generally, ECs that belong to multiple superfamilies do not belong to superfamilies that have multiple functions. And superfamilies that have multiple ECs do not have ECs that are found in multiple superfamilies.

### 3.3 Functional diversity in SCOP superfamilies

The majority of enzymes belong to the \( \alpha/\beta \) and \( \alpha + \beta \) structural classes. The combination of rigid surfaces formed by \( \beta \)-sheets with the conformational flexibility provided by \( \alpha \)-helices makes these scaffolds most suitable for enzymatic function (Anantharaman et al., 2003). Figure 3a shows the functional coverage of SCOP superfamilies and highlights just a few protein superfamilies with a broad functional
Fig. 3. (a) Functional diversity at the SCOP superfamily level. For each superfamily in the SCOP hierarchy, the number of unique ECs that they cover is calculated at different levels of the EC hierarchy; using the first three digits only (y-axis) and the complete EC number (x-axis). The size of each data point is proportional to the number of EC subclasses covered (EC class coverage is not plotted). For example the NAD(P)-binding Rossmann-fold (c.2.1) covers four unique EC classes, 8 subclasses, 9 sub-subclasses and 38 sub-sub-subclasses (axis: data-point size = 8, y = 9, x = 38). (b) Structural diversity in EC numbers. For each EC number (up to the third level), the coverage of the SCOP hierarchy is calculated; using superfamily level (y-axis) and the family level (x-axis). Calculations were performed on five SCOP domain classes only (all-α, all-β, α + β, α/β, multi-domain). The size of each data-point is correlated to the number of SCOP folds covered by an EC. For example, EC 3.2.1 (glycosidases) covers 5 SCOP classes, 11-folds, 12 superfamilies and 27 families (axis: data-point size = 11, y = 12, x = 27). It should be noted that it is impossible to directly compare the levels of the EC hierarchy between enzyme classes, because, depending on the initial enzymatic class, the subclasses will have alternative meanings.

repertoire. Three superfamilies cover four EC classes; NAD(P)-binding Rossmann-fold (c.2.1), Class I glutamine amidotransferase-like (c.23.16) and the P-loop containing nucleotide triphosphate hydrolases (c.37.1).

Although the amidotransferase-like superfamily covers four EC classes, it only has four unique functions compared with 38 for the NAD(P)-binding Rossmann-fold. In contrast, the trypsin-like serine proteases (b.47.1) have one functional class (hydrolases), one subclass (acting on peptide bonds) and two sub-subclasses (serine or cysteine endopeptidases), but then there are 28 unique EC numbers, representing the large variety of known cleavage sites. Since the fourth level of EC can expand greatly, based on our knowledge of substrate specificity, we will restrict further discussion to the third level of EC. Therefore, the most popular enzymatic structures are the PLP-dependent transferases (c.67.1, 10 functions), α/β hydrolases (c.69.1, 10 functions), P-loop containing nucleotide triphosphate hydrolases (c.37.1, 9 functions) and the NAD(P)-binding Rossmann-fold domains (c.2.1, 9 functions).

The variety in function seen with α/β hydrolases may be due to the observation that these domains vary greatly in size, with only the core and the catalytic triad conserved (Todd et al., 2001). The involvement of NAD(P)-binding Rossmann-fold in many functions might be expected as the fold is involved in binding a co-factor used in varying catalytic mechanisms (Todd et al., 2001). Although the TIM barrel is well known for accommodating many different functions (Nagano et al., 2002), this structure is not discussed here because we are concentrating our analysis at the SCOP superfamily level. The TIM barrel fold has functions in five EC classes, but none of its superfamilies have more than six different ECs (third level).

3.4 Structural diversity in enzymatic function

Most EC functions, 69% (third level), are found in multiple superfamilies. These functions have been invented independently during evolution; the most frequently occurring function is the hydrolases (4.2.1, e.g. carbonate dehydratase) found in 23 different superfamilies. Figure 3b shows the diversity of structure for each EC function, the distribution appears to be more uniform in comparison to the diversity of function in each structural superfamily (Figure 3a). Generally, functions are not tied to a specific fold; some ‘superfunctions’
are seen to be adopted by completely different folds (Hegyi and Gerstein, 1999). For example, cellulase (EC 3.2.1.4) is found in five SCOP folds: α/α toroid, concanavalin A-like lectins/glucanases, cellulases, double psi β-barrel, and TIM barrel. Cellulase is thought to have evolved relatively recently because it has a specific, rather than universal, function (Galperin et al., 1998). Alcohol dehydrogenase (EC 1.1.1.1) and phosphotransferases (EC 2.7.1.37) both show high levels of structural diversity, explained by the relative ambiguity of the EC number, allowing for a broad functional repertoire within a single classification. Overall, EC families do not correlate well with structural families since many functions are represented throughout various structures.

3.5 Conservation of enzyme function in iterative database searching

The focus of this work was to prepare a database for use in structure–function analysis and the assignment of function to novel sequences. In an initial attempt to assess the viability of function assignment, we have taken a representative domain from each SCOPEC enzyme family (a domain taken from the family level of SCOP with a unique EC) and searched a database of Swiss-Prot enzyme families. After the first iteration of PSI-BLAST, only 28.8% of the Swiss-Prot homologues shared EC function at the fourth level. This is not surprising since function description is very specific at this level and many of the homologues will be below 40% sequence identity. However, 85.7% of the homologues returned after the first iteration had the same function to the third EC level. Enzyme class and subclass were conserved with 93.5 and 87.7% of homologues having the same function.

The figures for the proteins that are newly found in the second PSI-BLAST iteration are 76.7, 52.9, 47.9 and 9.2% and for the third iteration are 68.0, 45.8, 39.1 and 2.9% (EC class, subclass, sub-subclass and sub-sub subclass, respectively). These figures are encouraging, since many reports suggest it is very difficult to successfully infer function below 40% sequence identity (Todd et al., 2001; Rost, 2002; Tian and Skolnick, 2003). In contrast, our preliminary work suggests that even for homologues detected by a third iteration PSI-BLAST profile there is a 50:50 chance of assigning a fairly specific three-digit EC number. Confidence in function assignment can be attributed to our confidence in the domain–EC relationships. Our results are not directly comparable with other studies, because we have examined functional conservation at each PSI-BLAST iteration only and not at different percentage sequence identities and PSI-BLAST E-values as observed by others. However, we are in agreement with other studies in that the correct assignment of all four EC digits is limited to very similar sequences.

Calculating the functional conservation separately for each enzyme family after a single database iteration gives an average degree of conservation of 94.5% ± 17.9, 89.8% ± 23.9, 87.4% ± 26.5 and 55.4% ± 41.2, moving down the EC hierarchy. The high levels of SD indicate how enzyme families behave very differently from each other in database searches, suggesting that enzyme families should be treated differently when assigning function.

4 CONCLUSIONS

SCOPEC is a unique resource mapping catalytic function to domain structure. It has representatives for 75% of enzymes with known structure. However, this only accounts for 39% of the functions described in the Swiss-Prot sequence database. This highlights the need for a greater number of solved enzyme structures and much more biochemical analysis to identify a set of sequences with known function.

The analysis of domain–EC relationships in SCOPEC highlights how many functions are represented throughout various structural superfamilies, whereas a structural superfamily is likely to have a single function with multiple substrate specificities. This suggests that nature re-invents function (convergent evolution). Following a reinvention, it is likely that modification leads to new specificities of function (divergent evolution).

In this study, we have examined the conservation of function through successive iterative PSI-BLAST searches of a database of Swiss-Prot enzyme families. In accordance with the above observations: substrate specificity, measured at the fourth EC level, was not conserved within homologues; whereas function, measured at the third EC level, is often conserved. Many enzymes were found to have varying numbers of homologues with different functions, indicating that each family should be treated differently when assigning function to novel sequences. Further information regarding specific residues involved in catalysis, such as those described in the CSA, or substrate specificity will add to our understanding of enzyme evolution and further improve the performance of function assignment.

ACKNOWLEDGEMENTS

We would like to thank Ian Carruthers and Richard Bickerton at Inpharmatica Ltd, and Michael Stevens and Craig Porter at the EBI for helpful discussions. R.A.G. and R.V.S. are grateful for support from the LINK-programme in Applied Genomics funded by the BBSRC.

REFERENCES


