EVEREST: A collection of evolutionary conserved protein domains

Elon Portugaly[1]*, Nathan Linial[1], Michal Linial[2]

[1]School of Computer Science & Engineering. [2]Dept. of Biological Chemistry, Inst. of Life Sciences.

The Hebrew University of Jerusalem

Abstract

Protein domains are subunits of proteins that recur throughout the protein world. There are many definitions attempting to capture the essence of a protein domain, and several systems that identify protein domains and classify them into families. EVEREST, recently described in Portugaly *et al.* (2006), *BMC Bioinformatics*, **7**:277, is one such system that performs the task automatically, using protein sequence alone. Herein we describe EVEREST release 2.0, consisting of 20,029 families, each defined by one of more HMMs. The current EVEREST database was constructed by scanning UniProt 8.1 and all PDB sequences (total over 3,000,000 sequences) with each of the EVEREST families. EVEREST annotates 64% of all sequences, and covers 59% of all residues. EVEREST is available at [http://www.everest.cs.huji.ac.il/]. The website provides annotations given by SCOP, CATH and Pfam A. It allows for browsing through the families of each of those sources, graphically visualizing the domain organization of the proteins in the family. The website also provides access to analyzes of relationships between domain families, within and across domain definition systems. Users can upload sequences for analysis by the set of EVEREST families. Finally an advanced search form allows querying for families matching criteria regarding novelty, phylogenetic composition and more.

*To whom correspondence should be addressed. email: elonp@cs.huji.ac.il

1 Introduction

Proteins are comprised of one or several domains. The literature in protein science teems with definitions that attempt to capture the correct notion of a protein domain. Employing a structural point of view, domains are sometimes defined as minimal segments of the protein that will fold to their native shape should they be isolated from the rest of the peptide chain. Other definitions take an evolutionary perspective and define domains as segments of the sequence that recur in different proteins. Based on these definitions, several attempt to define and classifiy domains within protein databases. These systems vary both in the type of data they analyze and in the amount of manual input they incorporate. SCOP (1) and CATH (2) are both classifications of domains that analyze protein structures. SCOP is a manual classification while CATH classification is determined using a combination of automated and manual procedures. The relative scarcity of protein structures has led to the development of protein domain classification systems that take as input only protein sequence information. Databases such as Pfam A (3), BLOCKS (4), SMART (5) offer comprehensive collections of families that were compiled by human experts, with the aid of computational tools (see review in (6, 7)). These methods provide high quality definitions that are most useful for biologists. However, they incorporate a great deal of human labor and expertise and require external information to identify new domain families. Several automatic systems for the identification and classification of domains in a database of protein sequences have been described in the literature. These include the ProDom algorithm (8) that was adopted by Pfam and forms Pfam B, and the more recent ADDA (9). EVEREST is our attempt at creating such an automatic system.

The different definitions for protein domains and for protein domain families do not always agree. In some cases these disagreements are the results of mistakes and inaccuracies. However, in many cases, more than one interpretation of the sequence or structure data is valid. The protein domain world is highly complex. For example, domains are hierarchical in nature, in two different senses. First, one domain may be composed of two or more sub-domains. Second, domain families may be grouped to superfamilies or divided into sub-families. Due to this complexity, several domain definition systems may disagree on the interpretation of a protein, and yet all be correct in some sense. It is therefore important to develop tools for browsing protein domain families and for comparing them, both within and across domain definition systems.

1.1 The EVEREST process

We have developed EVEREST (EVolutionary Ensembles of REcurrent SegmenTs), an automatic computational process identifying protein domains and classifying them into families. The EVEREST process begins by constructing a database of protein segments that emerge in an all vs. all pairwise sequence comparison. It then proceeds to cluster these segments, choosing the best clusters using machine learning techniques, and creating a statistical model for each of the them. This procedure is then iterated: The aforementioned statistical models are used to scan all protein sequences, to recreate a segment database and to cluster them again.

EVEREST has been thoroughly tested and evaluated, and has been shown to reconstruct 56% of Pfam A families and 63% of SCOP families with high accuracy, and to suggest many new domain families. A recently published manuscript describes the EVEREST process and its evaluation in detail (10).

2 The EVEREST database and website

The EVEREST database contains 20,029 families, each defined by one or more HMMs. The current release of the EVEREST database was constructed by scanning UniProt release 8.1 (11) and the sequences of all PDB (12) structures (total over 3 million sequences) with each of the EVEREST families. EVEREST annotates 93% of all Swiss-Prot sequences and 62% of all TrEMBL sequences (64% over all UniProt), and covers 84% of all residues in Swiss-Prot (56% for TrEMBL, 59% over all UniProt). For PDB, 88% of all sequences are annotated, and 84% of all residues are covered.

The EVEREST database of protein domain families can be accessed through the EVEREST website ([http://www.everest.cs.huji.ac.il/]). The web-site allows browsing through EVEREST domain families as well as domain families defined by SCOP, CATH and Pfam A. EVEREST families contain domains on both UniProt and PDB sequences. SCOP and

CATH families only contain domains on PDB sequences and Pfam families only contain domains on UniProt sequences. A family page in the website provides a graphical representation of all proteins containing a domain of the family, and of all domains, as defined by the above four domain definition systems, on these proteins.

EVEREST families are denoted as EV*RR.NNNNN* where *RR* stands for the release number and *NNNNN* stands for the family number within the release.

The website also features analysis of relationship between families and searches for proteins and families on the basis of keywords, family statistics, family phylogenetic profile and more. Finally, the user may upload a sequence to be scanned for EVEREST families and stored for future browsing by that user.

At any stage of the browsing, the user may customize the set of databases used. As a default non-redundant subsets of UniProt and PDB are used. The user may instead select to view the full versions of the sequence databases or to limit the view to the Swiss-Prot subset of UniProt. The user may also select which of the external domain definition systems to show, and at what level of classification (superfamilies or families for SCOP, homologous superfamily or S35 clusters for CATH and clans or families for Pfam).

2.1 Protein page

The protein page is accessible by textual search for keywords, accession numbers and names, as well as through links from domain family pages of all domains on the protein. The main body of the page starts with general information regarding the protein, followed by the sequence of the protein. Below that is a graphical representation of the domains on the protein. Domains are shown for all systems selected for view by the user. Each domain segment serves as a hyper link to the family page of the represented domain's family. For all but EVEREST domains, the segments representing the domains are color coded for family. EVEREST families are color coded by the best score they receive with respect to any reference family in the database (see section 3.2 "**Evaluating domain families using reference systems**"). See Figure 1 for an example.

2.2 Domain family page

A family page can be produced for families of the EVEREST, SCOP, CATH and Pfam systems. The main part of the page contains general information about the family followed by records describing all proteins containing domains of the family.

The general information part contains the family's name and links to the home page of the family for families defined by the external systems, followed by download links for the HMMs defining the family for EVEREST families. Below those is a link to a list of the domains of the family in a tabular form, followed by links to pages describing the scoring of the family by reference families from other systems and to the scoring of families from other systems using this family as a reference. See section 3.2 "**Evaluating domain families using reference systems**" for further details on the scoring of families.

Below the general family information part, each protein record contains textual information about the protein and a schematic representation of all domains on the protein, in the same format as in the protein page, with the exception that EVEREST families are not coded for score. The main family of the page is always color coded red.

At the left of the page is a vertical strip containing links to other parts of the website, followed by a legend for the color coding of the domain families appearing in the page. The legend also provides information about relationships between those families and the main family of the page, as illustrated in Figure 2.

2.3 Relationship between families

Our database describes relationships between domain families, both within and across domain definition systems. These relationships allow for the comparison of families and for browsing the domain family space from one family to related families. We define two dimensions of relations between protein domain families. The first dimension describes the relationship between "typical" domains of the two families. The second dimension describes the relationship between the two domain families in terms of set inclusion. For example, let us review the relationship between EV02.00096 and SCOP family *c.69.1.12: Haloperoxidase.* All 6 *c.69.1.12* domains are super-domains of domains of EV02.00096, but EV02.00096 contains 21 other domains, unrelated to *c.69.1.12* domains. Ascending one level in the SCOP hierarchy, all of EV02.00096 domains are sub-domains of SCOP super-family *c.69.1: alpha/beta-Hydrolases* domains, which in turn contains domains unrelated to EV.00096 domains. Thus *c.69.1.12* is a sub-family of super-domains of EV02.00096 which is a sub-family of sub-domains of *c.69.1.* See Figure 2 for an excerpt from the family page of EV02.00096 describing its relationships with SCOP families. Section 3.3 "**Relationships between domain families**" describes the definitions we use for marking relationships between families.

2.4 Family query page

The website allows querying for domain families by several criteria. The user may select one or more criteria to apply in conjunction. Following are the different criteria types available:

- Textual search in family name.
- Family size limits.
- Average domain size limits.
- Family taxonomical composition as defined by limits on the proportion of the domains in the family in user requested taxa. Taxa from all levels of the phylogenetic tree are available.
- Criteria regarding the novelty of the family as defined by limits on the proportion of domains in the family that are known to other domain definition systems (see section 3.2 "**Evaluating domain families using reference systems**").
- Limits on the scoring of the family by the best matching reference family of user selected reference domain definition systems (see section 3.2 "**Evaluating domain families using reference systems**").

Some criteria definitions, especially those involving phylogenetic profiling, may produce searches that require several minutes to complete. Therefore, users are asked to provide an email address to which we send an email with a hyperlink to the results of the search once it is completed.

For an example of search, suppose we wish to look for a new target for structural determination that might be applicable to medical research. We set the number of domains found on UniProt to be between 50 and 500. We request that the average size of the domain be between 100 and 200 amino acids - the usual range for structural domains. We ask that there would be no domains in PDB, because we want an unknown structure. Furthermore, we request that the proportion of the family covered by Pfam A to be at most 10% since Pfam families are already on the structural genomics target lists. Finally, because we wish for applicability to medical research, we ask that the family contains human proteins and rodent proteins. We set the search in motion. After a few seconds we are asked to be more precise regarding the taxa criteria. Since we knew of the many human viruses taxa, we have asked for "human -virus", so we only have to select "Homo Sapiens" amongst the many human bacteria and other parasites. For "rodent" we select the "Rodentia" order. Because our search contains phylogenetic criteria, it could take a while. Finally, when the search is over we receive an email with hyperlink to the list of 89 families it produced.

2.5 EVEREST annotation of user sequences

Users may also upload their own sequences to be scanned for EVEREST families. The scan takes several minutes to a few hours, and the user is notified by email upon completion. The email contains a hyperlink to a protein page of the uploaded sequence. Furthermore, during sessions starting from the hyperlink in the email, the user's uploaded sequence will show in the family pages of all domains found on this sequence.

2.6 Registration

Users may choose to register to our database. Registration provides the users with a private space in our database, in which the user's searches and uploaded sequences are stored. Thereafter, upon logging in, the user may access lists of all searches they performed and of all sequences uploaded. Furtheremore, all sequences uploaded by the user will show in the family pages of all domains found on those sequences.

2.7 Downloads

The EVEREST database is available for download through the downloads link in the website. Available for download are the HMMs defining the families, in HMMER format, and flat files listing the EVEREST domains found on UniProt and on the PDB sequences.

3 Technical details

3.1 Data sources

Protein sequences were taken from UniProt release 8.1 (11) and PDB (as downloaded from the PDB server on February 2006) (12).

EVEREST release 2.0 family models were generating by applying the EVEREST algorithm to release 49.2 of the Swiss-Prot database (11). These models were then used to identify family members on all sequences in our database.

SCOP domains were taken from ASTRAL release 1.69 (13). CATH release 2.6.0 was used. Pfam A families and clans (14) were taken from the InterPro database, release 12.1 (15).

Phylogenetic tree was downloaded from the NCBI Taxonomy FTP site (16).

3.2 Evaluating domain families using reference systems

The EVEREST system is evaluated by computing its coverage of reference systems and its accuracy when taking those reference systems as gold standards for domain family definitions. To this end we have developed a scoring scheme that enables scoring an evaluated domain family with respect to a reference domain family in the context of a reference system of domain families. A detailed description of the scoring scheme and the results of applying it to EVEREST is given in (10). Briefly, for an evaluated family e, let $\Pi(e)$ be a collection of reference domains given by allowing each domain in the evaluated family to collect those reference domains that significantly intersect with it. Then, when evaluating e with respect to a reference family r, a true positive would be a member of $\Pi(e)$ that is also a member of r, a false positive would be a member of $\Pi(e)$ that is not a member of r, and a false negative would be a member of r that is not a member of $\Pi(e)$. The score of e with respect to r would be the size of the intersection of $\Pi(e)$ and r divided by the size of their union. We have calculated the scores of EVEREST families with respect Pfam families and with respect to SCOP and CATH families. We have also calculated the scores of SCOP families with respect to CATH families and vice versa. Since Pfam families are defined on UniProt sequences, while SCOP and CATH families are defined on PDB sequences, we cannot score Pfam with respect to SCOP and CATH, furthermore, since a-priory, EVEREST is less reliable than SCOP, CATH and Pfam, and Pfam is less reliable than SCOP and CATH, we do not score the latter systems with respect to the former.

3.3 Relationships between domain families

Observing two domain instances on the same protein, we mark five relations, namely *sub-domain*, *super-domain*, *same*, *N-neighbor* and *C-neighbor*, as illustrated in Figure 3. When marking these relations, we allow each pair of domain instances a and b to be either *strongly following*, *possibly following*, *contradicting* or none of the above, with respect to each of the possible relationship types. *Strongly following* is always also *possibly following.* A pair of domain instances can be *possibly following* two different relations, but a pair that is *strongly following* a relation cannot be *possibly following* any other relation.

Let P_a be the proportion of domain a that is covered by domain b and P_b be the proportion of domain b that is covered by domain a. Let C_a be the middle position of domain a and C_b be the middle position of domain b. Table 1 shows the different conditions used for defining *strongly following* and *possibly following* for the different relations. For *N-neighbor* and *C-neighbor relations*, if the pair is not possibly following, it is defined to be contradicting. For *sub-domain*, *super-domain* and *same* relations, if a pair is not *possibly following* the relation and is not *strongly following* either of the two neighbor relations, it is defined to be contradicting. We also note the natural notion of reciprocity of relation. Namely *sub-domain* is reciprocal to *super-domain*, *N-neighbor* is reciprocal to *C-neighbor* and *same* is reciprocal to itself.

Observing two domain families A and B, we count for each of the above five relations the number of domains a of A for which there exist a domain b in B such that the pair a, b is *strongly following*, *possi-*

bly following and *contradicting* the relation. These counts form the basis of the second dimension of the relationship between the families. If all, or nearly all of the domains of A have a certain relation with a domain of B, but a significant number of the domains of B do not have the reciprocal relation with a domain of A, then B is a *super-family* of A with respect to that relation, and A is a *sub-family* of B with respect to that relation. If all, or nearly all of the domains of A have a certain relation with a domain of B and all or nearly all of the domains of B have the reciprocal relation with a domain of A then A and B are *matching* families with respect to that relation. We do not provide exact definitions and thresholds for these terms. Instead we provide, and graphically visualize the counts of the domains in each family sharing the relation, and let the user decide how to name the relationship between the families.

4 Maintenance and future developments

The EVEREST database is designed to handle multiple versions of EVEREST and of all other information sources (sequence database and domain definition systems). In fact, EVEREST families defined by a scan of an older Swiss-Prot version are available by choosing to view EVEREST release 1.0. We will run the EVEREST process at least once a year to define new families and update the database as new releases of UniProt, PDB, SCOP, CATH and Pfam are available.

Storing search results opens many options for combining the results of different searches. We plan to enable more sophisticated searches by adding tools for conjunction and disjunction of result sets, as well as tools for combining result sets via the family relations defined in section 3.3 "**Relationships between domain families**". An example search using such a tool would be to define two sets of SCOP families of two different functions using keyword search, and then to look for EVEREST families that are super-families of members of both SCOP sets.

5 Acknowledgments

We thank Alex Savenok for the design and much of the programming of the EVEREST website. We thank Reuven Abliyev and Menachem Fromer for their support. We also thank the system team in the Hebrew University School of Computer Science and Engineering.

E.P. is supported by an Eshkol fellowship from the Israeli Ministry of Science and by the Sudarsky Center for Computational Biology. This work is partially funded by NoE (Framework VI) BioSapiens consortium.

References

1. Hubbard,T.J., Ailey,B., Brenner,S.E., Murzin,A.G. and Chothia,C. (1999) SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res,* **27**(1), 254–6.
2. Orengo,C.A., Pearl,F.M., Bray,J.E., Todd,A.E., Martin,A.C., Lo Conte,L. and Thornton,J.M. (1999) The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res,* **27**(1), 275–9.
3. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res,* **30**(1), 276–80.
4. Henikoff,S., Henikoff,J.G. and Pietrokovski,S. (Jun, 1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics,* **15**(6), 471–479.
5. Schultz,J., Copley,R.R., Doerks,T., Ponting,C.P. and Bork,P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res,* **28**(1), 231–4.
6. Henikoff,S. (1995) Comparative methods for identifying functional domains in protein sequences. *Biotechnol Annu Rev,* **1**, 129–147.
7. Liu,J. and Rost,B. (2003) Domains, motifs and clusters in the protein universe. *Curr Opin Chem Biol,* **7**(1), 5–11.
8. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief Bioinform,* **3**(3), 246–51.
9. Heger,A. and Holm,L. (2003) Exhaustive enumeration of protein domain families. *J Mol Biol,* **328**(3), 749–67.
10. Portugaly,E., Harel,A., Linial,N. and Linial,M. (2006) EVEREST: automatic identification and classification of protein domains in all protein sequences. *BMC Bioinformatics,* **7**, 277.

11. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., E.,G., Huang,H., Lopez,R., Magrane,M., Martin,M.J., Mazumder,R., O'Donovan,C., Redaschi,N. and Suzek,B. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res,* **34**(Database issue), D187–91.

12. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res,* **28**, 235–42.

13. Chandonia,J.M., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2002) ASTRAL compendium enhancements. *Nucleic Acids Res,* **30**(1), 260–3.

14. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R., Eddy,S.R., Sonnhammer,E.L. and Bateman,A. (Jan, 2006) Pfam: clans, web tools and services. *Nucleic Acids Res,* **34**(Database issue), D247–D251.

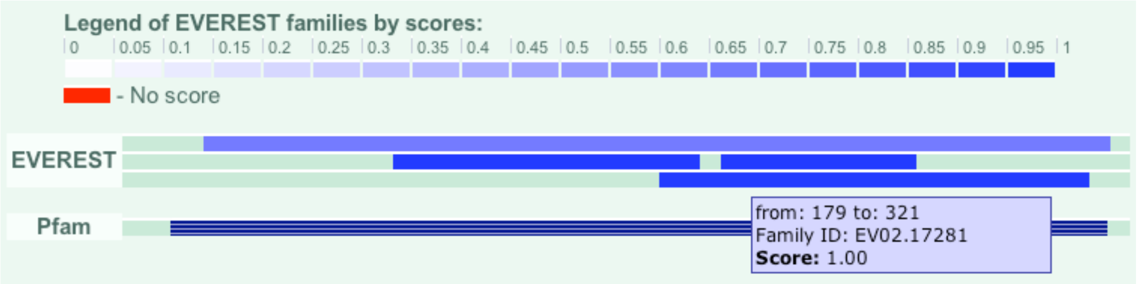
15. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L., Copley,R., Courcelle,E., Das,U., Durbin,R., Fleischmann,W., Gough,J., Haft,D., Harte,N., Hulo,N., Kahn,D., Kanapin,A., Krestyaninova,M., Lonsdale,D., Lopez,R., Letunic,I., Madera,M., Maslen,J., McDowall,J., Mitchell,A., Nikolskaya,A.N., Orchard,S., Pagni,M., Ponting,C.P., Quevillon,E., Selengut,J., Sigrist,C.J., Silventoinen,V., Studholme,D.J., Vaughan,R. and Wu,C.H. (Jan, 2005) InterPro, progress and status in 2005. *Nucleic Acids Res,* **33**(Database issue), D201–D205.

16. Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (Jan, 2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* **28**(1), 10–14.

Tables

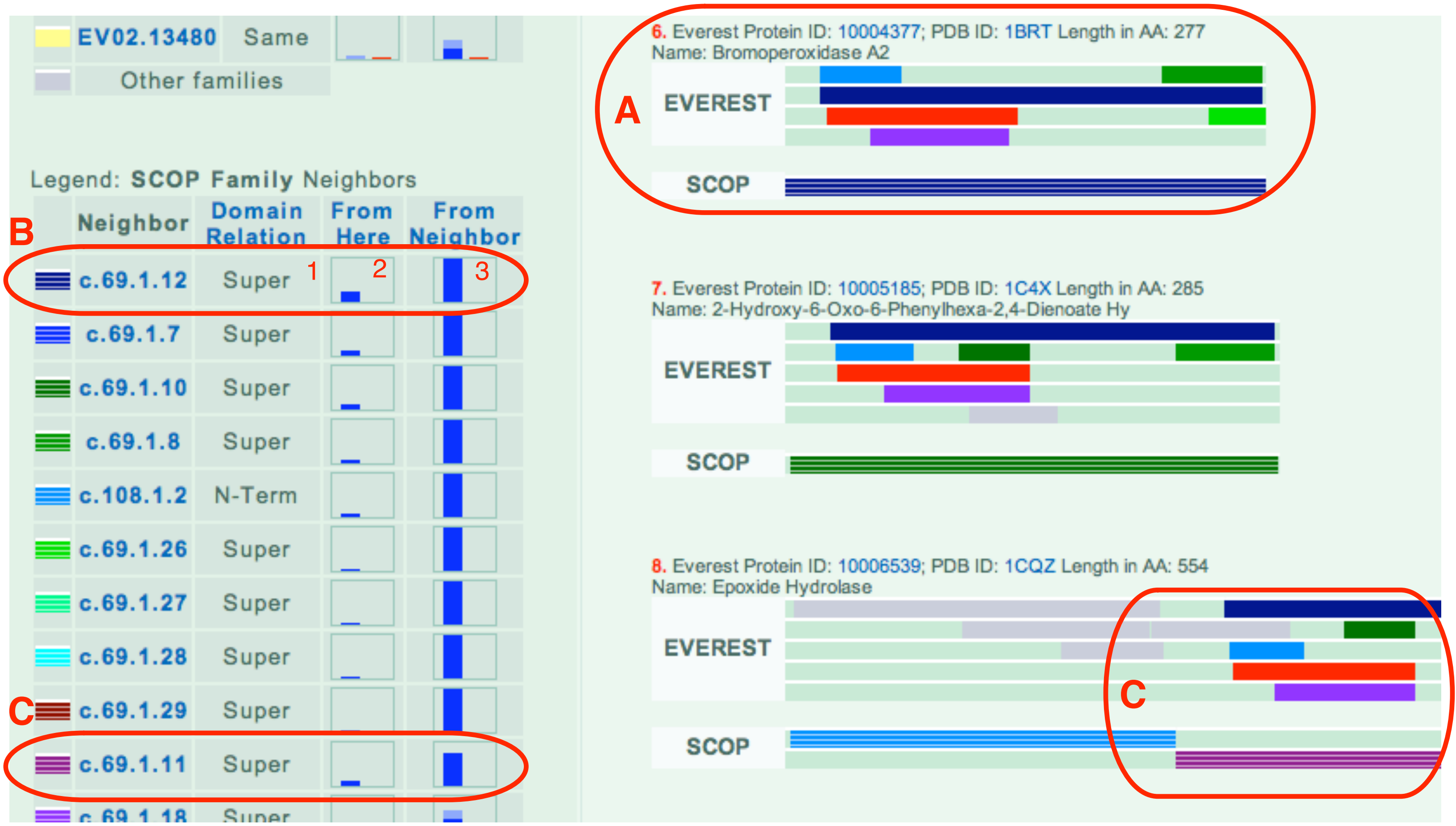
Table 1 - Parameters for defining relations between two domain instances

	Conditions					
Relation	**Strongly following**			**Possibly following**		
sub-domain	$P_a \geq 0.9$	$P_b < 0.65$		$P_a \geq 0.75$	$P_b < 0.8$	
super-domain	$P_a < 0.65$	$P_b \geq 0.9$		$P_a < 0.8$	$P_b \geq 0.75$	
same	$P_a \geq 0.9$	$P_b \geq 0.9$		$P_a \geq 0.75$	$P_b \geq 0.75$	
N-neighbor	$P_a \leq 0.1$	$P_b \leq 0.1$	$C_a < C_b$	$P_a \leq 0.25$	$P_b \leq 0.25$	$C_a < C_b$
C-neighbor	$P_a \leq 0.1$	$P_b \leq 0.1$	$C_a > C_b$	$P_a \leq 0.25$	$P_b \leq 0.25$	$C_a > C_b$



from: 179 to: 321
Family ID: EV02.17281
Score: 1.00

Figure 1: **Example protein record.** Excerpt from the protein page of HMUU_YRPE - “Hemin transport system permease protein hmuU” showing the graphical representation of the domains on the protein. The width of the record is proportional to the length of the protein sequence. Colored segments mark domains found by different systems (here EVEREST and Pfam) on the sequence. EVEREST segments are color coded for the best score their family receives with respect to any reference family in the database. Other segments are color coded for family. A color legend is available in a vertical stripe in the left side of the page.



	EV02.13480	Same		
	Other families			

Legend: SCOP Family Neighbors

	Neighbor	Domain Relation	From Here	From Neighbor
	c.69.1.12	Super		
	c.69.1.7	Super		
	c.69.1.10	Super		
	c.69.1.8	Super		
	c.108.1.2	N-Term		
	c.69.1.26	Super		
	c.69.1.27	Super		
	c.69.1.28	Super		
	c.69.1.29	Super		
	c.69.1.11	Super		
	c.69.1.18	Super		

6. Everest Protein ID: 10004377; PDB ID: 1BRT Length in AA: 277
Name: Bromoperoxidase A2

EVEREST

SCOP

7. Everest Protein ID: 10005185; PDB ID: 1C4X Length in AA: 285
Name: 2-Hydroxy-6-Oxo-6-Phenylhexa-2,4-Dienoate Hy

EVEREST

SCOP

8. Everest Protein ID: 10006539; PDB ID: 1CQZ Length in AA: 554
Name: Epoxide Hydrolase

EVEREST

SCOP

Figure 2: **Relationship between EV02.00096 and SCOP *c.69.1.12*.** Excerpt from the family page of EV02.00096 is shown. **A** Record for PDB sequence 1BRT is highlighted. The EV02.00096 domain, in red, is a sub-domain of the SCOP *c.69.1.12* domain, in striped dark blue. **B** The relationship between EV02.00096 and *c.69.1.12* is described by (**1**) the keyword "Super" indicating that *c.69.1.12* domains are super-domains of EV02.00096 domains, (**2**) the left bar graph, which through the height of the bar indicates that less than a quarter of EV02.00096 domains participate in this relationship, and (**3**) the right bar graph, indicating that all of the domains of *c.69.1.12* participate in this relationship. **C** EV02.00096 is also a super-family of sub-domains of *c.69.1.11*.

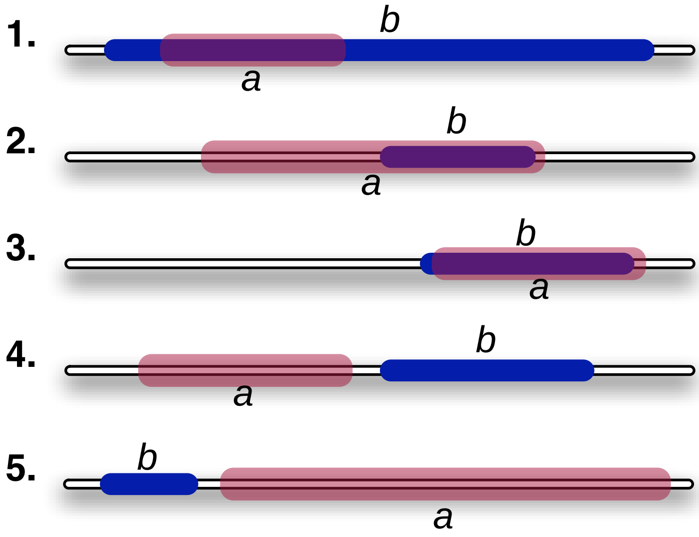


Figure 3: **Five types of relations between domain instances.** Illustration of the five defined relation types between two domain instances on the same protein. **1.** *sub-domain*: domain a is a sub-segment of domain b. **2.** *super-domain*: domain a is a super-segment of domain b. **3.** *same*: domain a is the same segment as domain b. **4.** *N-neighbor*: domain a is N-terminal to domain b. **5.** *C-neighbor*: domain a is C-terminal to domain b.