doi:10.1016/j.jmb.2008.12.045 J. Mol. Biol. (2009) 387, 416–430

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Domain-Based and Family-Specific Sequence Identity Thresholds Increase the Levels of Reliable Protein Function Transfer

Sarah Addou†, Robert Rentzsch†*, David Lee and Christine A. Orengo

Institute of Structural and Molecular Biology, University College London, London WC1E 6BT, UK

Received 13 August 2008; received in revised form 12 December 2008; accepted 17 December 2008
Available online
25 December 2008

Divergence in function of homologous proteins is based on both sequence and structural changes. Overall enzyme function has been reported to diverge earlier (50% sequence identity) than overall structure (35%). We herein study the functional conservation of enzymes and non-enzyme sequences using the protein domain families in CATH-Gene3D. Despite the rapid increase in sequence data since the last comprehensive study by Tian and Skolnick, our findings suggest that generic thresholds of 40% and 60% aligned sequence identity are still sufficient to safely inherit third-level and full Enzyme Commission numbers, respectively. This increases to 50% and 70% on the domain level, unless the multi-domain architecture matches. Assignments from the Kyoto Encyclopedia of Genes and Genomes and the Munich Information Center for Protein Sequences Functional Catalogue seem to be less conserved with sequence, probably due to a more pathway-centric view: 80% domain sequence identity is required for safe function transfer. Comparing domains (more pairwise relationships) and the use of family-specific thresholds (varying evolutionary speeds) yields the highest coverage rates when transferring functions to model proteomes. An average twofold increase in enzyme annotations is seen for 523 proteomes in Gene3D. As simple 'rules of thumb', sequence identity thresholds do not require a bioinformatics background. We will provide and update this information with future releases of CATH-Gene3D.

Keywords: sequence identity thresholds; domain-based transfer of protein function; genome functional annotation; enzyme classification; KEGG Orthology

Edited by B. Honig

Introduction

Currently, much functional characterisation of genes and proteins relies essentially on sequence-similarity-based approaches, in which experimentally determined functions are transferred to uncharacterised sequence relatives with various levels of confidence. Because BLAST and PSI-BLAST[1] are amenable to large-scale database searches, they are the most widely used automated homology detection methods applied to functionally annotate uncharacterised sequences on a genome-wide scale.

Depending on the organism, only about 5% to 30% of annotations are, at this point, directly experimentally verified, with *Escherichia coli* being an outstanding exception (67%).[2] This means that the majority of databases contain large amounts of inherited annotation. Aside from the errors caused by the inconsistency of experimental methods, inferring annotations electronically between sequences can cause serious problems with the accuracy of the data.[3] Incorrect functional assignment can easily propagate to new database sequence entries and undermine the value of genome annotation. The number of database errors is known to grow over time, as both the number of entries and errors tend to accumulate at an exponential rate.[4,5] Estimates say that as much as 30% of all database annotations may be wrong.[6,7]

*Corresponding author. E-mail address: rentzsch@biochem.ucl.ac.uk.

† Joint first authors.

Abbreviations used: EC, Enzyme Commission; SCOP, Structural Classification of Proteins; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; FunCat, Functional Catalogue; MDA, multi-domain architecture; KO, KEGG Orthology.

Several studies have considered the accuracy of sequence-similarity-based methods for functional annotation transfer. These systematic analyses identified sequence identity thresholds above which consistent function conservation levels are observed between enzyme relatives. A first study[8] examined the relationship between pairwise sequence identity taken from structural alignments in the Families of Structurally Similar Proteins database and several functional descriptors including Enzyme Commission (EC) digits, active sites, and keywords found in SwissProt and the Protein Data Bank. It was found that EC digits are better conserved than other types of functional descriptions and that above 50% sequence identity, all four digits of an EC number are well conserved.

Wilson *et al.* assessed function conservation between enzymatic and non-enzymatic sequences in pairs of structural domains from the Structural Classification of Proteins (SCOP) database and concluded that conservation of the full EC digits can be obtained with 40% sequence identity between two proteins.[9] They also showed that although probabilistic scores outperform sequence identity measures for recognising highly divergent sequences (near the twilight zone), sequence identity is better calibrated for identifying homologues with conserved functions.

In a comparable study,[10] Todd *et al.* investigated the functional variation of homologous enzyme superfamilies in the CATH database and extended their analysis to domain sequence relatives in SwissProt using PSI-BLAST. Their findings suggested that >40% and >60% sequence identity was required for conservation of full EC numbers in single- and multiple-domain proteins, respectively. When functional conservation was measured by the conservation of the first three EC digits, which describe the reaction chemistry of the enzyme, then 30% and 40% identity was required in order to safely infer functional annotation in single- and multi-domain proteins, respectively.

The reliability of all the estimated threshold values stated was later contested by Rost,[11] who pointed out that none of the studies had taken account of the compositional bias in (not only) the SwissProt database. Any all-against-all comparison to study function conservation is, according to this, necessarily biased towards those enzyme families dominating the functional annotations in the database. The relationship between sequence and function conservation, however, may not be the same for all enzyme families.

Various approaches were adopted to reduce this bias. Rost himself selected representatives from homologous enzyme superfamilies so that any trends detected were not biased towards a particular (highly conserved) subfamily. For each 'normalised' superfamily, the percentage of representatives with similar enzymatic function at a particular sequence identity level was calculated, showing that divergence of both the first digit and all four digits of the

Table 1. Summary of earlier studies exploring sequence identity thresholds for the conservation of enzyme functions in sequence relatives

Authors	Dataset size (number of protein sequences)	Dataset composition	Functional annotation source(s)	Sequence alignment tool(s)	Conclusions
Devos and Valencia[8]	2338	Homologous pairs of Protein Data Bank structural domains or sequences	SwissProt for enzyme annotations, keywords and information on binding sites	Derived from structural alignments in the Families of Structurally Similar Proteins database	50% identity is required for conservation of all four EC digits
Wilson et al.[9]	n/a	29,454 representative pairs of structural domains from SCOP. Comparisons at various levels of the SCOP hierarchy (family, superfamily, fold)	The EC scheme and an augmented version of FlyBase to obtain non-enzyme annotations	Smith/Waterman	40% identity is required for conservation of all four EC digits
Todd et al.[10]	65,303	Homologous pairs of structural domains from CATH and their sequence relatives from SwissProt and GenBank	SwissProt and GenBank for enzyme and non-enzyme annotations	Needleman/Wunsch	40% and 60% identity is required for conservation of all four EC digits in single- and multi-domain proteins
Rost[11]	26,243	Whole protein sequences with an EC annotation	SwissProt for enzyme annotations	BLAST	Below 70% identity, both the first and the fourth EC digits start to diverge
Tian and Skolnick[12]	22,645	Whole protein sequence homologues including enzymes and non-enzymes (derived by PSI-BLAST)	SwissProt for enzyme and non-enzyme annotations	Align0 (Myers/Miller global alignment)	40% (60%) identity is required for conservation of the first three (all four) EC digits