Representativity of target families in the Protein Data Bank: impact for family-directed structure-based drug discovery

Jordi Mestres

Analysis of the population of enzyme structures in the Protein Data Bank across all levels of the functional classification based on enzyme commission (EC) numbers reveals that, in spite of the almost exponential growth in the number of structures deposited, progress in achieving complete occupancy at all EC levels is relatively slow. Moreover, inspection of the distribution of the population among the members of the different enzyme families uncovers a strong bias towards enzymes widely recognized as therapeutically relevant targets. The low representativity levels identified in some target families warn on the current scope and applicability of structure-based approaches to family-directed strategies in drug discovery.

The completion of the human genome sequencing, in conjunction with the establishment of classification schemes for the main therapeutically relevant protein families, has opened an avenue towards more systematic strategies to drug discovery [1–4]. In the post-genomic era, the classical approach of screening a collection of molecules on a single target for a particular therapeutic area is evolving into novel chemogenomic approaches based on the profiling of compound libraries on entire target families potentially associated with a variety of therapeutic areas [5,6]. The adoption of this new paradigm is expected to make global drug discovery efforts more efficient through the gain of knowledge within target families and its exploitation in lead generation and optimization processes [7,8].

An important part of the knowledge generated within target families comes from the availability of experimentally determined protein structures. Recent advances in high-throughput methods for protein expression and production, NMR spectroscopy, and X-ray crystallography have led to a significant rise in the number of protein structures solved [9]. Many of these structures are ultimately deposited and made publicly accessible in the Protein Data Bank (PDB), currently containing over 27,000 entries and its size continuing to increase annually at an almost exponential rate [10]. The availability of protein structures has motivated the development of computational methods capable of suggesting the mode of binding of individual ligands into a protein cavity with reasonable accuracy at relatively low cost [11]. Traditionally, these methods have been applied to the virtual screening of large chemical libraries against a particular protein of interest [12,13]. More recently they have been adapted to the virtual profiling of compound databases on multiple family-related proteins [14–17]. As the number of protein family members with representative structures in the PDB expands, it will become increasingly feasible to make family-wide binding-site comparisons to extract commonalities and differences that can then be translated into potential privileged and selective protein–ligand interactions, respectively [18–22].

Jordi Mestres
Chemogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, 08003 Barcelona (Catalonia), Spain
e-mail: jmestres@imim.es

1359-6446/05/$ – see front matter PII: S1359-6446(05)03593-2

However, the immediate applicability of structure-based approaches to entire protein families will be determined not only by the total number of structures available in the PDB for the family but also by the precise distribution of these structures among the different protein members of the family. For example, for a family composed of 20 proteins with a population of 100 structures in the PDB, the degree of structural representativity of the protein family will not be the same if there are five structures available for each one of the 20 protein members of the family than if all 100 structures are variants of the same protein. In the latter case, the strong bias towards a single protein member would imply that an important contribution from homology modeling techniques is required, before undertaking any structure-based activities on the entire family. A quantitative means for assessing the structural representativity of protein families in the PDB should be able to identify potential unbalances in the distribution of structures among the members of a protein family.

Unfortunately, primarily because of technical difficulties, not all of the therapeutically relevant protein superfamilies, namely enzymes, nuclear receptors, ligand-gated ion channels and G protein-coupled receptors, are at present equally represented in the PDB. With over 13,000 entries, enzymes are the most populated family in the PDB. By contrast, around 150 structures are available for nuclear receptors and only a handful has been resolved for G protein-coupled receptors. In view of the significant amount of structural information available for enzymes, this review focuses on analyzing the current structural representativity of enzyme families in the PDB.

Classification and annotation: prerequisites to assessing representativity

The general adoption of hierarchical classification schemes for proteins is an essential aspect for assessing quantitatively the structural representativity of protein families in the PDB. In this respect, the lack of existence of a unified standard classification scheme for all existing proteins remains an open issue in this field, with several classification schemes currently coexisting for many protein families. Upon adoption of a classification scheme, existing protein structures in the PDB can be assigned to a given code within the scheme, a process usually referred to as annotation. The classification scheme for enzymes and its use for the annotation of structures in the PDB are described below, together with details on assessing representativity by means of quantitative measures of the occupancy and distribution of annotations among the complete enzyme classification scheme.

Classification of enzymes

Enzymes constitute a large superfamily of proteins well characterized and classified, with a classification scheme that has prevailed for decades [23,24]. Enzymes are classified according to the type of reaction catalyzed using a four-digit identifier, usually referred to as the enzyme commission (EC) number [25]. The first digit specifies the class of enzyme. There are six different enzyme classes, namely oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases, which are assigned to EC numbers from one to six in that particular order. The second digit specifies the enzyme subclass according to a compound or group involved in the reaction being catalyzed. The third digit specifies the enzyme sub-subclass defining the type of reaction in a more concrete manner. And the fourth digit is a number specifying the individual enzyme within a sub-subclass. As of October 2004, the list of enzymes in the EC classification scheme amounted to 4199. The classification scheme was then processed to take care of all enzyme codes marked as 'deleted' or 'transferred', following the recommendations and annual supplements for the nomenclature and classification of enzymes [25], because their inclusion could interfere with the structural representativity analysis. A total of 395 superseded enzymes were found, resulting in a final list of 3804 enzymes, 222 sub-subclasses and 63 subclasses.

Annotation of enzyme structures

Having defined the classification scheme for enzymes, the next step is the identification of enzyme structures in the PDB and their annotation using that scheme. For this task, data were extracted directly from the PDBsum database [26], a web-based repository that contained 13,467 enzyme entries (as of October 2004), involving 12,854 separate PDB files, some files having more than one EC number associated with them. Some of these original entries corresponded to enzymes that had been assigned to another enzyme code by the Enzyme Nomenclature Committee [25]. Therefore, their populations were transferred to the newly assigned EC codes accordingly. This process affected a population of 80 enzyme entries in the PDB.

Quantitative assessment of representativity

When attempting to analyze the structural representativity of protein families in the PDB, it is important to consider the number of protein members within a family, for which at least one structure exists in the PDB (i.e. occupancy), but also the relative allocation of the number of structures among the protein members of a given family (i.e. distribution). Whereas the former is straightforward to obtain, the latter requires the use of a quantitative means for measuring the variability of distributions.

Given a family of N protein members, with $n \leq N$ of them having at least one structure in the PDB, a protein family occupation index, O, will be defined as $O = n / N$, with values in the range of [0,1]. By contrast, to assess quantitatively the variability of the total number of structures in the PDB for a given family (i.e. population), measures derived from information theory will be used [27]. Accordingly, the entropy, S, of a population $P > 0$, distributed among a number of protein members of a given family, n, is given by: