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:

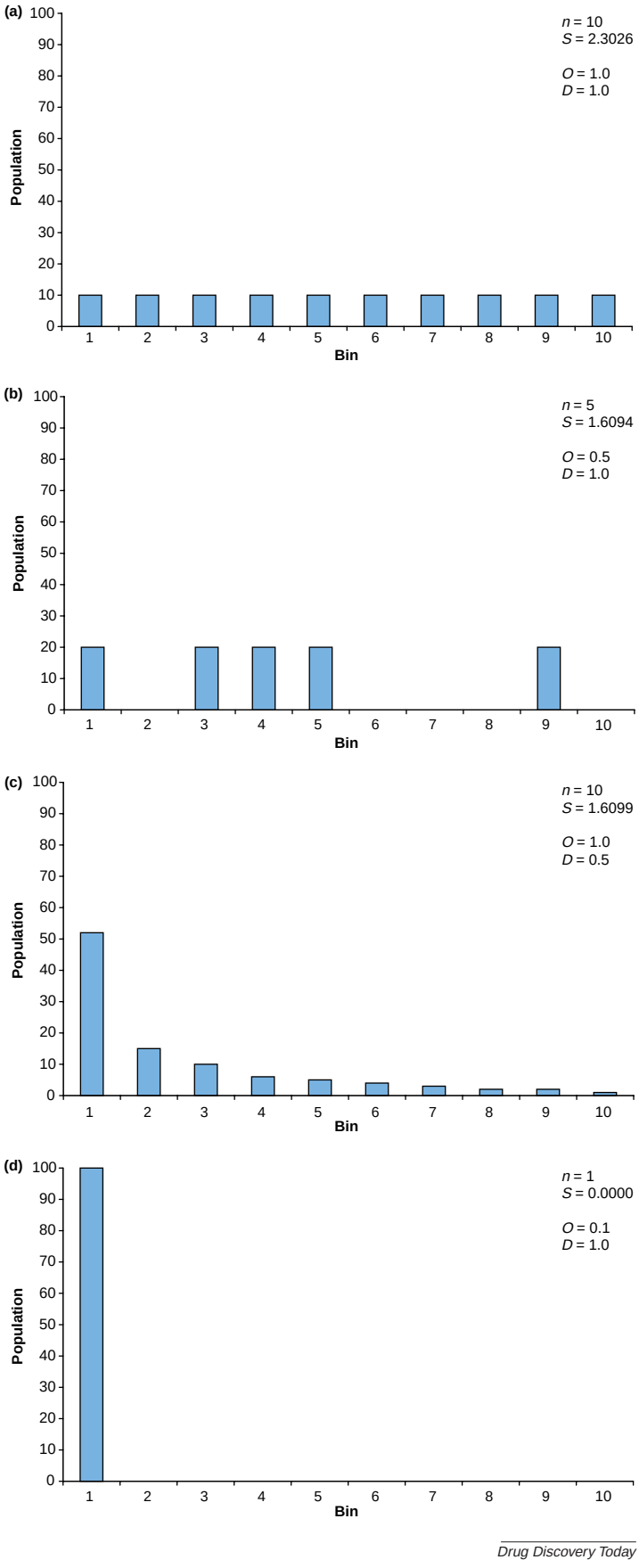

FIGURE 1

Four examples to illustrate the meaning of the different parameters and indices used to quantify the variability of distributions. In all cases, a population of 100 entries has been distributed among the ten protein members of a family. The parameters *n* and *S* are the number of occupied bins (representing each protein member) and the Shannon entropy, respectively, and the indices *O* and *D* are the occupancy and distribution, respectively. In **(a)**, an evenly distributed population among all proteins gives rise to the situation of maximum representativity, $O = D = 1.0$. **(b)** provides an example of a distribution with medium occupancy, $O = 0.5$, but maximum distribution among all the occupied protein members, $D = 1.0$. By contrast, **(c)** shows a situation of maximum occupancy, $O = 1.0$, with an unevenly distributed population within the family, $D = 0.5$. Finally, **(d)** reflects the situation of minimum representativity, $O = 1/N$, in which the entire population is concentrated in a single protein member of the family.

$$S = -\sum_{i=1}^{n} \rho_i \cdot \ln \rho_i \qquad ; \qquad \rho_i = p_i / P$$

where ρ_i and p_i are the probability and the population at each protein *i* of the distribution, respectively. The values of *S* range between 0, reflecting the situation of all population being concentrated in a single protein, and a maximum number, $S_{max} = \ln N$, reflecting the situation of a uniformly distributed population among all protein members of the family. To normalize *S* values for families having a different number of members, a protein family distribution index, *D*, is defined as $D = e^S / n$, with values in the range of $[1/n, 1]$. Should there be no structure in the PDB for a given family, $P = 0$, then *D* will be assigned to $D = 0$. Plotting occupancy versus distribution maps the representativity of enzyme families in the PDB. To illustrate the meaning of the values for the parameters and indices in different distributions, several examples are provided in Figure 1.

Representativity of experimentally determined enzyme structures

The overall content of the PDB has been growing at an almost exponential rate over the years. The growth of enzyme entries in the PDB has followed a similar trend. The question is whether the expansion observed in the number of enzyme structures is translated into an increasing occupancy at all levels of the enzyme nomenclature system. In this respect, Table 1 presents a summary of the number of subclasses, sub-subclasses and enzymes per class currently being populated in the PDB, compared with their corresponding total number in the enzyme nomenclature system. As can be observed, the distribution of the structural population per class given in Table 1 does not match the distribution of the enzyme population per class based on EC numbers. For example, a comparison of the numbers obtained for enzyme classes 1 and 3 reveals that, although both classes contain a similar number of enzymes (1043 and 1050, respectively) and thus contribute almost equally to the total number of enzymes in the classification system (27.4% and 27.6%, respectively), the number of enzyme entries for class 1 is almost 2.5 times lower than for class 3 (2404 and 5950, respectively) and

TABLE 1

Number of subclasses, sub-subclasses, and enzymes structurally represented in the PDB (SR) compared with the corresponding number in the enzyme nomenclature system (ENS) per class. The total number of enzyme structures (population) per class is also added.

Class	Subclasses		Sub-subclasses		Enzymes		Population
	SR	ENS	SR	ENS	SR	ENS	
1	21	22	63	97	240	1043	2404
2	8	9	24	28	277	1095	3124
3	9	13	43	55	384	1050	5950
4	6	7	13	15	115	326	979
5	6	6	16	17	59	155	624
6	6	6	10	10	59	135	386
Total	56	63	169	222	1134	3804	13467

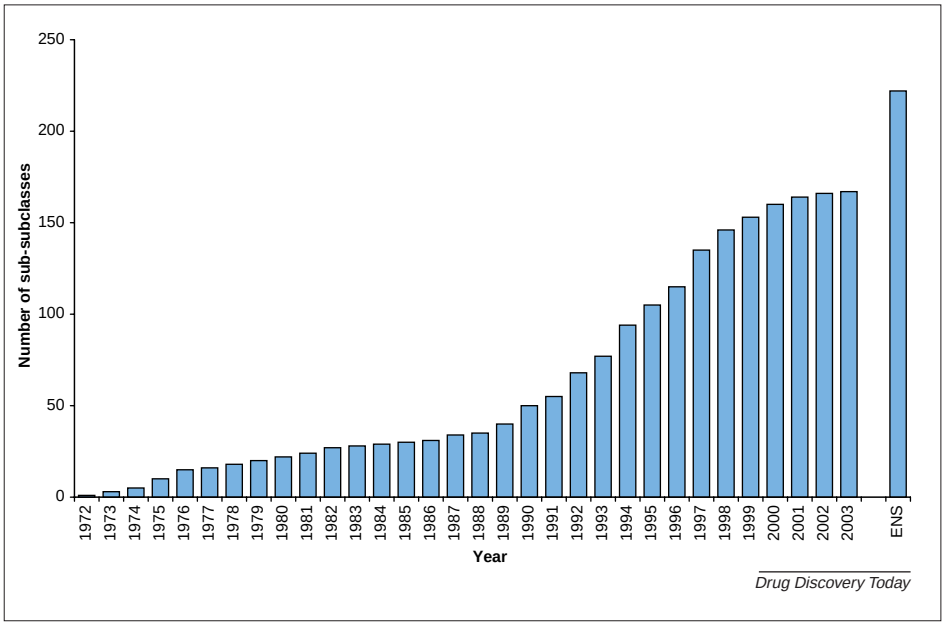

FIGURE 2

Growth in the number of sub-subclasses structurally represented in the PDB. For comparison, the total number of sub-subclasses currently defined in the enzyme nomenclature system (ENS) is also provided.

thus their contribution to the total number of enzyme entries currently available in the PDB is not similarly balanced (17.9% and 44.2%, respectively). This illustrates the present lack of balance between the number of enzyme entries in the PDB and their actual distribution among the classes, subclasses, sub-subclasses and particular enzymes of the enzyme nomenclature system. In terms of occupancy it is quite comforting to realize that overall 88.9% and 76.1% of all enzyme subclasses and sub-subclasses, respectively, are populated to some extent in the PDB. However, less than 30% of all enzymes contain a representative structure in the PDB, despite the total number of enzyme entries being over 3.5 times the number of enzymes catalogued in the enzyme nomenclature system. Therefore, a detailed quantitative analysis is required to assess the current occupancy levels and structural representativities of enzyme families in the PDB and to identify the potential sources for the general unbalance observed between the enzyme nomenclature system and the number of enzyme entries found in the PDB.

Growth in structures versus progress in occupancy

Despite the almost exponential growth in the number of enzyme entries being deposited in the PDB, progress in achieving complete occupancy at all levels of the enzyme nomenclature system appears to be relatively slow. This trend is illustrated in Figure 2 by the growth in the number of new enzyme sub-subclasses having at least one representative structure in the PDB. As can be observed, progress towards full occupancy of enzyme sub-subclasses has dramatically relented in recent years. At present, seven subclasses and 53 sub-subclasses still remain devoid of experimentally determined structures in the PDB (see Table 1). In addition, the number of new enzymes, for which representative structures are deposited in the PDB, has only increased linearly in the last five years with, on average, 97 new enzymes per year. With over 70% of the enzymes still remaining deserted of experimentally determined structures in the PDB, achieving full occupancy of all the enzymes defined in the classification system appears to be 20 years off.

A snapshot of the present structural occupancy in the PDB at the sub-subclass level of the enzyme nomenclature system is provided in Figure 3. The dashed line indicates the situation of full sub-subclass occupancy, which is all enzymes within a sub-subclass having at least one representative structure in the PDB. The size of the circle is proportional to the number of enzyme sub-subclass entries in the PDB, with a maximum value of 1561 entries corresponding to sub-subclass 3.2.1 (glycosidases). In this respect, by comparing sub-subclasses containing a similar number of enzyme members, it is observed that progress towards

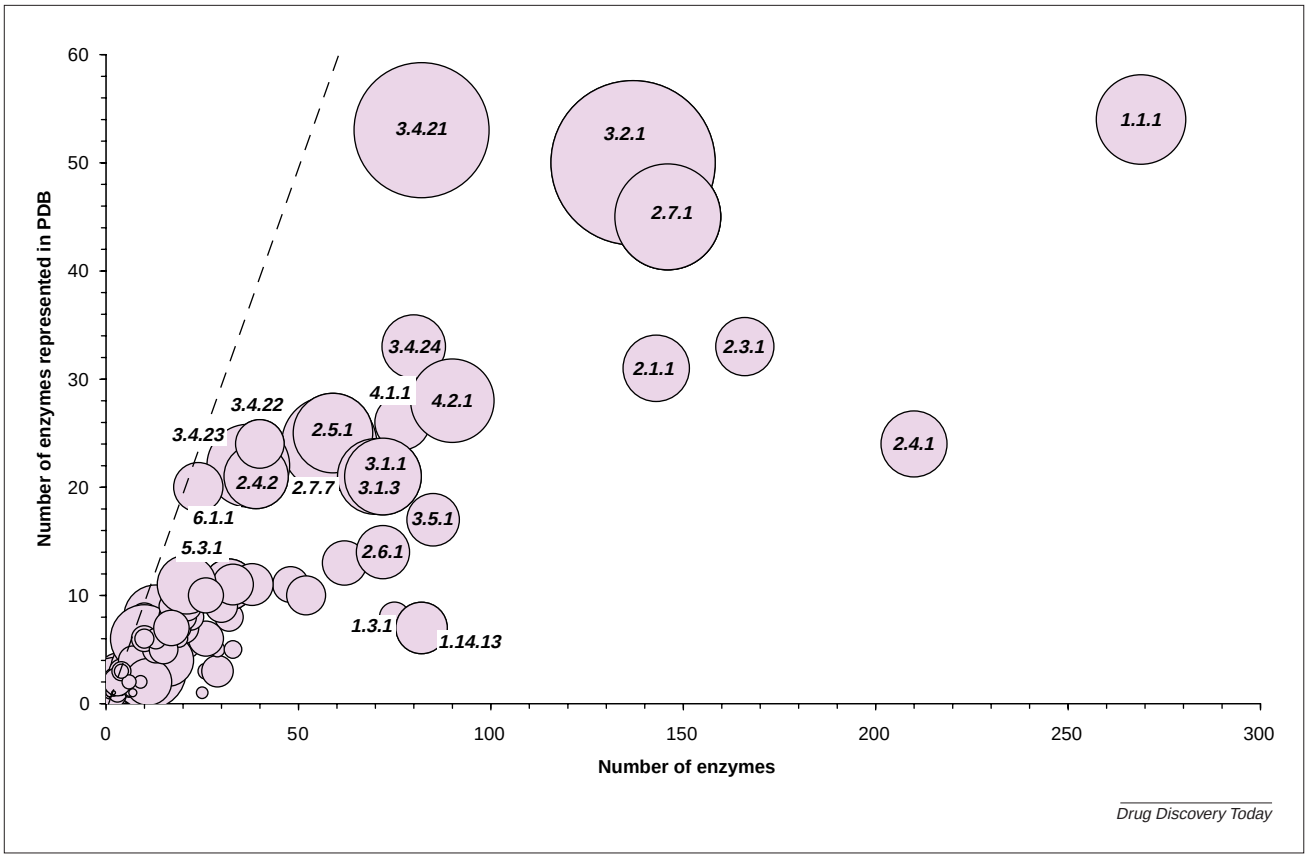

FIGURE 3

Structural occupancy of enzyme sub-subclasses in the PDB.

full occupancy correlates with the number of entries present in the PDB. For example, enzyme sub-subclasses 1.14.13 (oxidoreductases acting on paired donors with NADH or NADPH as one donor and incorporation of one atom), 3.5.1 (hydrolases acting in linear amides), 3.4.24 (metalloendopeptidases) and 3.4.21 (serine endopeptidases) contain 82, 85, 80, and 82 enzyme members, respectively. The corresponding numbers of PDB entries for these sub-subclasses are 155, 162, 235, and 1061, respectively, which correlates well with the 7, 17, 33, and 53 enzyme members within each sub-subclass being populated in the PDB. By contrast, when comparing sub-subclasses with a similar number of enzyme entries in the PDB, a tendency to deviate further from full occupancy is generally observed, as the size of the sub-subclass increases. This can be explained by the fact that the larger the number of enzyme members, the more difficult it is for the sub-subclass to achieve full structural occupancy, as a larger minimum number of representative structures is necessary to populate all enzymes of the sub-subclass. For example, one of the most poorly structurally occupied sub-subclasses is 2.4.1 (hexosyltransferases). With a total of 255 entries in the PDB, they populate only 24 of their 210 enzyme members. By contrast, sub-subclass 2.4.2 (pentosyltransferases), with a similar amount of representative structures in the PDB (241) and number of enzymes populated (21), is much closer to full occupancy of its 39 enzymes. The use of the occupancy and distribution indices defined above will help transforming all these figures into normalized measures that allow for quantitatively assessing and comparing the current experimentally determined structural representativity of enzyme families in the PDB.

Representativity of enzyme families: occupancy versus distribution

So far, only occupancy of the enzyme members within a given sub-subclass has been considered. At this stage, the distribution of the population of PDB entries among the enzyme members of the sub-subclass will be also taken into account. To this end, occupancy and distribution indices were obtained for each enzyme sub-subclass, as described above. The results are graphically depicted in Figure 4, organizing all enzyme sub-subclasses accordingly to the current structural representativity in the PDB. As in Figure 3, the size of the circle is proportional to the number of enzyme sub-subclass entries in the PDB. For the sake of discussion, the representativity map has been arbitrarily divided into four quadrants. Sub-subclasses with high and low structural representativity are located in the top-right and bottom-left quadrants, respectively. The distribution of the 169 structurally represented enzyme sub-subclasses among the quadrants is as follows: the top-right quadrant, with high occupancy and high distribution, collects 61 enzyme families; the top-left quadrant, with low

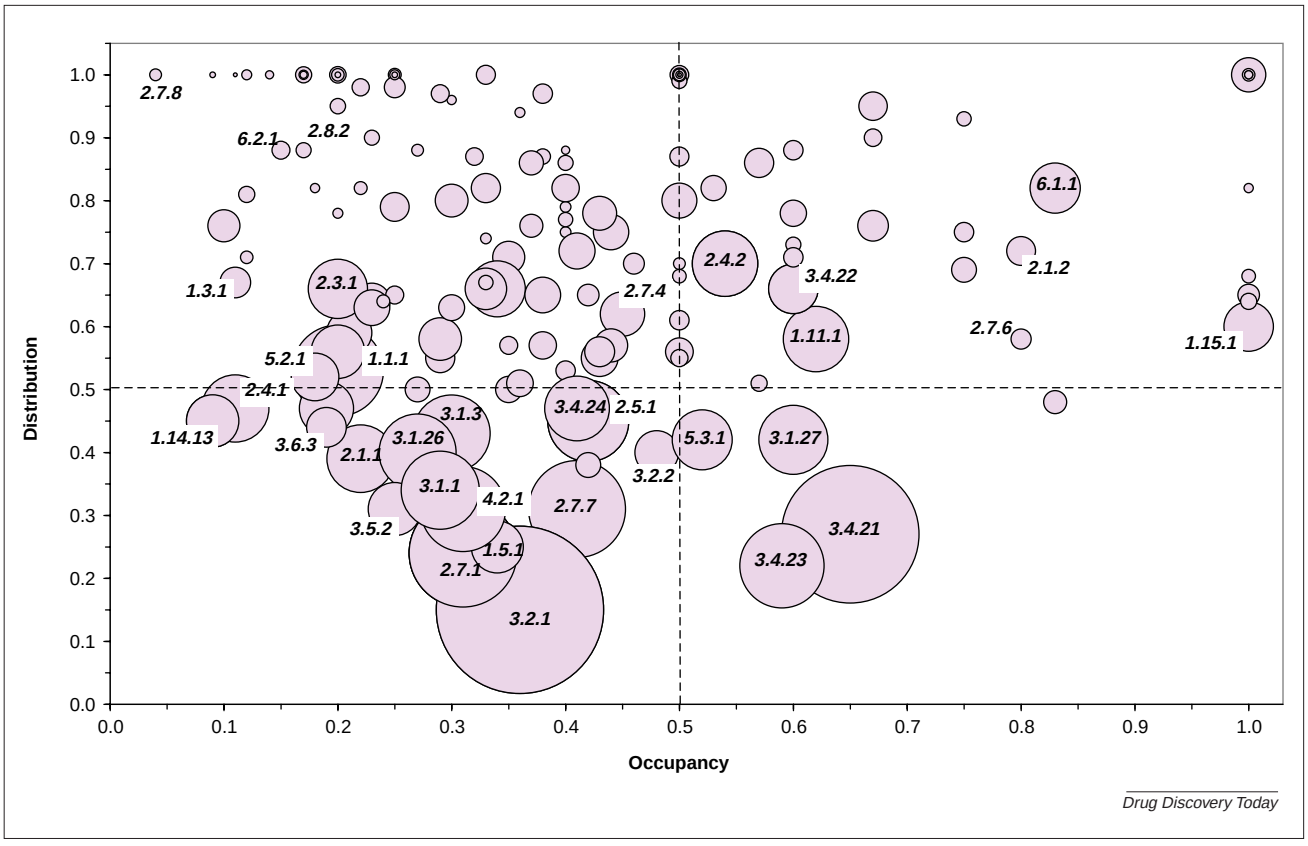

FIGURE 4

Structural representativity of enzyme sub-subclasses in the PDB.

occupancy and high distribution, compiles 84 enzymes families; the bottom-right quadrant, with high occupancy and low distribution, accumulates 5 enzyme families; and the bottom-left quadrant, with low occupancy and low distribution, gathers 19 enzyme families.

Two sub-subclasses containing a similar number of enzyme entries and members but different values for the occupancy and distribution indices will be compared. From the top-right quadrant, sub-subclass 6.1.1 (ligases forming aminoacyl-tRNA and related compounds) is a good example of a well structurally represented sub-subclass. With 142 entries in the PDB, it achieves values for the occupancy and distribution indices of 0.83 and 0.82, respectively. This means that of the 24 enzymes forming this sub-subclass, 83% have at least one representative structure in the PDB and that the variability in the distribution of the number of structures among the occupied enzymes is such as if 82% of those enzymes were equally populated. By contrast, 3.2.2 (enzymes hydrolyzing N-glycosyl compounds) provides an example of a sub-subclass in the bottom-left quadrant, with much lower representativity of experimentally determined structures than 6.1.1. The spread of the 108 entries found in the PDB results in occupancy and distribution indices of 0.48 and 0.40, respectively, meaning that 11 of its 23 enzyme members have at least one representative structure in the PDB, with a variability of the population as if only 40% of those were uniformly populated. In fact, over 60% of all entries for 3.2.2 are concentrated in a single enzyme (3.2.2.22, rRNA N-glycosylase). The use of occupancy and distribution indices and their two-dimensional graphical representation provides a means for detecting this difference in an illustrative yet quantitative manner.

Representativity of target families: the focus on therapeutic relevance

We will now concentrate on the set of 24 enzyme sub-subclasses in the low-distribution side of the structural representativity map to find the source of their low-distribution values despite their relatively large population in the PDB. The list of these 24 sub-subclasses has been compiled in Table 2, ordered according to their occupancy index (see Figure 4). These 24 enzyme sub-subclasses represent only 10.8% of all the sub-subclasses catalogued in the enzyme nomenclature system. However, they contain 40.6% of all the enzymes characterized with an EC number. Altogether, they collect 8118 entries, which represent 60.3% of all enzyme entries in the PDB. All these figures provide an indication of the relative importance that this reduced set of enzyme sub-subclasses has over the others. For this reason, they will be referred to as enzyme target families.

Close inspection of the distribution of PDB entries among their respective enzyme members reveals that

TABLE 2

List of the 24 low-distribution target families (D<0.5) together with the 34 most populated target enzymes identified within these families

Target Family[a]				Target Enzymes[a]		
EC	**Name**	**N**[b]	**P_{PDB}**[b]	**EC**	**Name**	**%P_{PDB}**[c]
1.7.99	Oxidoreductases acting on other nitrogenous compounds as donors with other acceptors	6	30	1.7.99.1	Hydroxylamine reductase	73.3%
3.4.21	Serine endopeptidases	82	1061	3.4.21.4	Trypsin	30.4%
				3.4.21.5	Thrombin	17.5%
3.1.27	Endoribonucleases producing other than 5′-phosphomonoesters	10	268	3.1.27.5	Pancreatic ribonuclease	58.2%
				3.1.27.3	Ribonuclease T(1)	35.4%
3.4.23	Aspartic endopeptidases	37	398	3.4.23.16	HIV-1 retropepsin	62.6%
5.3.1	Intramolecular oxidoreductases interconverting aldoses and ketoses	21	203	5.3.1.1	Triose-phosphate isomerase	36.5%
				5.3.1.5	Xylose isomerase	36.5%
3.2.2	Glycosylases hydrolysing N-glycosyl compounds	23	108	3.2.2.22	rRNA N-glycosylase	60.2%
2.5.1	Not defined	59	367	2.5.1.18	Glutathione transferase	36.5%
3.4.17	Metallocarboxypeptidases	19	35	3.4.17.1	Carboxypeptidase A	71.4%
3.4.24	Metalloendopeptidases	80	235	3.4.24.27	Thermolysin	23.8%
				3.4.24.17	Stromelysin 1	15.3%
2.7.7	Nucleotidyltransferases	58	528	2.7.7.7	DNA-directed DNA polymerase	46.0%
3.2.1	Glycosidases	137	1561	3.2.1.17	Lysozyme	54.3%
1.5.1	Oxidoreductases acting on the CH–NH group of donors with NAD^+ or $NADP^+$ as acceptor	32	152	1.5.1.3	Dihydrofolate reductase	75.7%
4.2.1	Hydrolyases	90	406	4.2.1.1	Carbonate dehydratase	47.5%
2.7.1	Phosphotransferases with OH as acceptor	146	653	2.7.1.112	Protein-tyrosine kinase	35.4%
				2.7.1.37	Protein kinase	26.0%
3.1.3	Phosphoric monoester hydrolases	70	334	3.1.3.48	Protein-tyrosine phosphatase	32.3%
				3.1.3.11	Fructose-biphosphatase	16.2%
3.1.1	Carboxylic ester hydrolases	72	342	3.1.1.4	Phospholipase A2	38.9%
				3.1.1.7	Acetylcholinesterase	17.8%
3.1.26	Endoribonucleases producing 5′-phosphomonoesters	11	334	3.1.26.4	Ribonuclease H	96.1%
3.5.2	Hydrolases acting on C–N bonds, other than peptide bonds, in cyclic amides	16	158	3.5.2.6	β-Lactamase	95.6%
2.1.1	Methyltransferases	143	258	2.1.1.45	Thymidylate synthase	39.9%
2.3.2	Aminoacyltransferases	15	23	2.3.2.13	Protein-glutamine γ-glutamyltransferase	91.3%
2.6.1	Transaminases	72	165	2.6.1.1	Aspartate transaminase	50.3%
3.6.3	Hydrolases acting on acid anhydrides catalyzing transmembrane movement	52	89	3.6.3.14	H^+-transporting two-sector ATPase	53.9%
				3.6.3.4	Copper-exporting ATPase	19.1%
2.4.1	Hexosyltransferases	210	255	2.4.1.1	Phosphorylase	30.2%
				2.4.1.19	Cyclomaltodextrin glucanotransferase	18.8%
1.14.13	Oxidoreductases acting on paired donors, one being NADH or NADPH, with incorporation of one atom	82	155	1.14.13.39	Nitric-oxide synthase	60.0%
				1.14.13.2	4-hydroxybenzoate 3-monooxygenase	21.9%

[a]The order of the target families and enzymes reflects their respective level of occupancy (see Figure 4).

[b]The parameters N and P_{PDB} are, respectively, the number of enzyme members with an EC number and the population in the PDB of a given target family.

[c]Percentage of population of a target family concentrated on a given target enzyme.

these 24 enzyme families have in common the fact that they contain one or two enzymes collecting a significant proportion of all entries available for the family, which explains the low-distribution values obtained for these sub-subclasses (see Figure 4). Enzymes within families that contain over 15% of all the family entries will be referred to as target enzymes. The list of 34 target enzymes identified within the 24 target families is gathered in Table 2. Overall, it is remarkable that, despite representing only 0.9% of all enzymes catalogued in the enzyme nomenclature system, these target enzymes have attracted 4647 enzyme entries, which is 34.5% of all enzyme entries in the PDB.

The list of target families and enzymes collected in Table 2 confirms that the vast majority of them have a well-recognized therapeutic relevance. For example, target family 3.4.21 (serine proteases) is with 1061 structures the second most populated enzyme sub-subclass in the PDB. However, two of its 82 enzyme members compile a

significant percentage of those entries. These are 3.4.21.4, trypsin, and 3.4.21.5, thrombin, altogether representing 47.9% of all the family entries. Thrombin plays a key role in blood coagulation and hemostasis and thus has been historically an attractive target in antithrombotic drug research. Because of its high homology, trypsin has been often used as a protein surrogate for thrombin. Sub-subclass 3.4.23 (aspartic proteases) is also well represented in the PDB with 398 entries, of which 62.6% are concentrated on the enzyme 3.4.23.16, HIV-1 protease, an important target in the fight against HIV infections. Target family 3.4.24 (metalloendopeptidases) is known to play important roles in biological systems. The PDB contains 235 entries, 39.1% of which are concentrated in only two of its 80 enzyme members. These are 3.4.24.27, thermolysin, a bacterial enzyme often considered as a model for the related human enzymes in the family, and 3.4.24.17, stromelysin 1, a potentially important therapeutic target for the treatment of cancer. The PDB contains also 406 entries for sub-subclass 4.2.1 (hydrolyases), 47.5% of which are collected by a single enzyme out of its 90 enzyme members. This is 4.2.1.1, carbonate dehydratase (also known as carbonic anhydrase), the inhibitors of which are widely used pharmacological agents for the treatment or prevention of a variety of diseases, such as glaucoma, cystoid macular edema, diabetic retinopathy, epilepsy, neurological and neuromuscular disorders, obstructive pulmonary disease, sleep apnea, osteoporosis and cancer. Sub-subclass 2.7.1 (phosphotransferases with OH as acceptor) is a large family comprising 146 enzymes, many of which have been shown to be involved in cell signal transduction and are now viewed as potential therapeutic targets for various diseases. Not surprisingly, with 653 entries it is currently the third most populated enzyme sub-subclass in the PDB, 61.4% of them being concentrated in enzymes 2.7.1.112, protein-tyrosine kinase, and 2.7.1.37, protein kinase. These examples provide compelling evidence of the influence that enzymes with potential therapeutic value have had on global protein structural determination efforts.

Concluding remarks

The analysis presented in this study has uncovered the fact that structural determination efforts have historically focused on a small number of enzymes of potential therapeutic relevance. In particular, it was found that 34.5% of all enzyme entries in the PDB correspond to only 34 enzymes, which is 0.9% of all enzymes currently characterized with an EC number. This is not surprising considering that, once the therapeutic value of an enzyme has been established and the initial technical difficulties solved, obtaining additional structures of the same enzyme forming a complex with different ligands is a natural way of capitalizing on the investments made in time and resources. In fact, having multiple structures for a given enzyme should not be regarded necessarily as producing redundant structural information, because the binding of different ligands can induce conformational changes in the residues forming the protein cavity that might help providing a better understanding of the relative importance and potential implications of protein flexibility for drug design, as well as on the mechanism and specificity of that protein. Therefore, the main issue about the bias observed in the current content of the PDB is not so much about the generation of a large number of structures for a reduced set of enzymes, but about the lack of experimentally determined structural information for over 70% of all enzymes and the low pace at which representative structures for new enzymes are being obtained. In this respect, high expectations are put on structural genomics initiatives [28,29] to increase the rate at which new enzymes become structurally represented in the PDB and on homology modeling techniques [30,31] to complement current representativity levels of experimentally determined structures with computationally derived structural models.

Progress towards achieving full enzyme representativity in the PDB should come from recognizing that augmenting the structural representativity within target families provides true added value to family-directed strategies in drug discovery, with an impact from the structure-based selection of compounds for targeted screening campaigns to the structure-based design of targeted chemical libraries with tailored selectivity profiles. In fact, today's irrelevant enzyme might become tomorrow's most desired target and vice versa. In either case, the structural information gained on a particular enzyme, irrespective of its therapeutic relevance, will create exploitable knowledge on its entire family.

Supplementary material

Data on the representativity of target families in the Protein Data Bank are available at http://cgl.imim.es/pdbrtf/ through a web-based tool called PDB_{RTF}.

Acknowledgements

This research was supported by a grant from the Instituto de Salud Carlos III (Ministerio de Sanidad y Consumo), research project reference number 02/3051.

References

1 Caron, P.R. *et al.* (2001) Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* 5, 464–470
2 Bleicher, K.H. (2002) Chemogenomics: bridging a drug discovery gap. *Curr. Med. Chem.* 9, 2077–2084
3 Bredel, M. and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* 5, 262–275
4 Mestres, J. (2004) Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Curr. Opin. Drug Discov. Devel.* 7, 304–313
5 Haggarty, S.J. *et al.* (2003) Chemical genomic profiling of biological networks using graph theory and combinations of small molecule perturbations. *J. Am. Chem. Soc.* 125, 10543–10545
6 Giaever, G. *et al.* (2004) Chemogenomics profiling: identifying the functional interactions of small molecules in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 101, 793–798

7 Bleicher, K.H. *et al.* (2003) Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* 2, 369–378
8 Koch, M.A. and Waldmann, H. (2005) Protein structure similarity clustering and natural product structure as guiding principles in drug discovery. *Drug Discov. Today* 10, 471–483
9 Sali, A. *et al.* (2003) From words to literature in structural proteomics. *Nature* 422, 216–225
10 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242
11 Fradera, X. and Mestres, J. (2004) Guided docking approaches to structure-based design and screening. *Curr. Top. Med. Chem.* 4, 687–700
12 Kitchen, D.B. *et al.* (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* 3, 935–949
13 Shoichet, B.K. (2004) Virtual screening of chemical libraries. *Nature* 432, 862–865
14 Toledo-Sherman, L.M. and Chen, D. (2002) High-throughput virtual screening for drug discovery in parallel. *Curr. Opin. Drug Discov. Devel.* 5, 414–421
15 Lamb, M.L. *et al.* (2001) Design, docking, and evaluation of multiple libraries against multiple targets. *Proteins* 42, 296–318
16 Aronov, A.M. *et al.* (2001) Virtual screening of combinatorial libraries across a gene family: in search of inhibitors of *Giardia lamblia* guanine phosphoribosyltransferase. *Antimicrob. Agents Chemother.* 45, 2571–2576
17 Schapira, M. *et al.* (2003) Nuclear hormone receptor targeted virtual screening. *J. Med. Chem.* 46, 3045–3059
18 Kastenholz, M.A. *et al.* (2000) GRID/CPCA: a new computational tool to design selective ligands. *J. Med. Chem.* 43, 3033–3044
19 Naumann, T. and Matter, H. (2002) Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: target family landscapes. *J. Med. Chem.* 45, 2366–2378
20 Terp, G.E. *et al.* (2002) Structural differences of matrix metalloproteinases with potential implications for inhibitor selectivity examined by GRID/CPCA approach. *J. Med. Chem.* 45, 2675–2684
21 Pirard, B. (2003) Peroxisome proliferator-activated receptors target family landscape: a chemometrical approach to ligand selectivity based on protein binding site analysis. *J. Comput. Aided Mol. Des.* 17, 785–796
22 Ji, H. *et al.* (2003) Computer modeling of selective regions in the active site of nitric oxide synthases: implication for the design of isoform-selective inhibitors. *J. Med. Chem.* 46, 5700–5711
23 Tipton, K. and Boyce, S. (2000) History of the enzyme nomenclature system. *Bioinformatics* 16, 34–40
24 Schomburg, I. *et al.* (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* 32, D431–D433
25 Webb, E.C. Editor (1992) *Enzyme nomenclature*. Academic Press, San Diego. Regularly updated web version: http://www.chem.qmw.ac.uk/iubmb/enzyme
26 Laskowski, R.A. (2001) PDBsum: Summaries and analyses of PDB structures. *Nucleic Acids Res.* 29, 221–222. Weekly updated enzyme structures database: www.ebi.ac.uk/thornton-srv/databases/enzymes
27 Shannon, C.E. and Weaver, W. (1949) *The Mathematical Theory of Communication*, University of Illinois Press
28 O'Toole, N. *et al.* (2003) Coverage of protein sequence space by current structural genomics targets. *J. Struct. Funct. Genomics* 4, 47–55
29 Todd, A.E. *et al.* (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J. Mol. Biol.* 348, 1235–1260
30 Pieper, U. *et al.* (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* 32, D217–D222
31 Hillisch, A. *et al.* (2004) Utility of homology models in the drug discovery process. *Drug Discov. Today* 9, 659–669