Bioinformatics Advance Access published November 9, 2010

Generation of a Consensus Protein Domain Dictionary

R. Dustin Schaeffer[1], Amanda L. Jonsson[2], Andrew M. Simms[3], and Valerie Daggett[1,2,3,*]

[1]Biomolecular Structure and Design Program, [2]Department of Bioengineering,
and [3]Biomedical and Health Informatics, University of Washington, Seattle WA, 98195-5013

Associate Editor: Prof. Anna Tramontano

Downloaded from bioinformatics.oxfordjournals.org at Universidade Cat?lica de Bras?lia on November 18, 2010

ABSTRACT

Motivation: The discovery of new protein folds is a relatively rare occurrence even as the rate of protein structure determination increases. This rarity reinforces the concept of folds as reusable units of structure and function shared by diverse proteins. If the folding mechanism of proteins is largely determined by their topology, then the folding pathways of members of existing folds could encompass the full set used by globular protein domains.

Results: We have used recent versions of three common protein domain dictionaries (SCOP, CATH, and Dali) to generate a consensus domain dictionary (CDD). Surprisingly, 40% of the metafolds in the CDD are not composed of autonomous structural domains, i.e. they aren't plausible independent folding units. This finding has serious ramifications for bioinformatics studies mining these domain dictionaries for globular protein properties. However, our main purpose in deriving this consensus domain dictionary was to generate an updated CDD to choose targets for MD simulation as part of our Dynameomics effort, which aims to simulate the native and unfolding pathways of representatives of all globular protein consensus folds (metafolds). Consequently, we also compiled a list of representative protein targets of each metafold in the CDD.

Availability and Implementation: This domain dictionary is available at www.dynameomics.org.

1 INTRODUCTION

Structurally similar proteins need not share significant sequence identity. The early observation of structurally and functionally similar proteins (such as hemoglobin and myoglobin) led to the partition of different sets of structurally similar proteins into folds (Kendrew *et al.*, 1960; Perutz *et al.*, 1960). However, as more structures were determined and more folds discovered, it became clear that not all members of a fold are necessarily linked by a common function (Nagano *et al.*, 2002). Also, the determination of structures with conserved structural cores surrounded by variable regions complicated the classification of new structures into existing folds. Different approaches to resolving this heterogeneity of fold classification have been reviewed elsewhere (Schaeffer *et al.,* 2010). What degree of structural variation is tolerable between a domain and a potential cousin before they no longer can be considered to belong to the same fold?

The inconsistencies of analyzing and generating protein domain dictionaries are one component of the vigorous discussion surrounding the properties of protein 'fold space' (Csaba *et al.*, 2009; Pascual-Garcia *et al.*, 2009; Sam *et al.*, 2006). Distinct folds can contain regions of shared structural similarity (Grishin, 2001). Folds are both populated to different degrees and structurally heterogeneous (Coulson and Moult, 2002; Majumdar *et al.*, 2009; Wolf *et al.*, 2000). This heterogeneity complicates estimates of the size and 'shape' of fold space, and is likely responsible for the wide range of the estimated number of protein folds. The presence of unclear domain boundaries in regions of fold space have led some to question the utility of a hierarchical definition (Kolodny *et al.*, 2006). Furthermore, fold assignment is also dependent on the problem of domain detection (Holland *et al.*, 2006; Majumdar *et al.*, 2009).

The gold standards among domain dictionaries, SCOP (Structural Classification of Proteins) (Murzin *et al.*, 1995) and CATH (Class, Architecture, Topology, Homology) (Orengo, *et al.*, 1997), have been the subject of many detailed comparisons (Day *et al.*, 2003; Hadley and Jones, 1999; Jefferson *et al.*, 2008; Pascual-Garcia *et al.*, 2009; Veretnik *et al.*, 1997). In general, both dictionaries weigh potential functional and evolutionary relationships between fold members with different strengths at different levels of their hierarchies. The presence of shared fragments between differing folds and/or regions of "conserved" structure have been well documented and are one reason for the development of different empirical classification methodologies, as more knowledge of protein structural evolution emerges, hope remains that an evolutionary classification will be derived (Valas *et al.*, 2009). In their early formulations, these domain dictionaries represented different design methodologies. Whereas SCOP was hand curated by experts, CATH was maintained by a combination of automated process and expert curation. However, SCOP has assumed more automated pre-classification of new structures in response to the increasing rate of structure determination, diluting this methodological distinction (Andreeva *et al.*, 2008).

[*]To whom correspondence should be addressed.

© The Author (2010). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

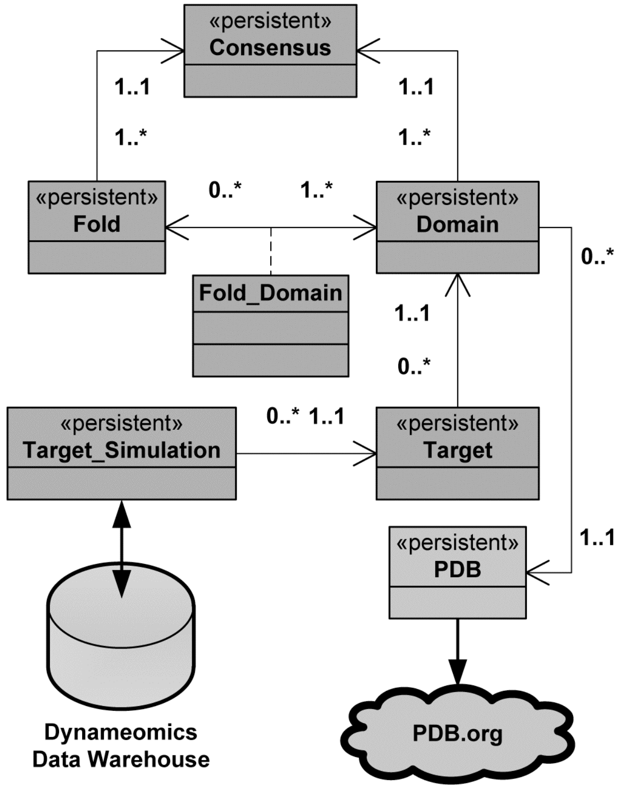


Figure 1. Target Selection and Preparation ('Prep') database schema modified to account for multiple consensus sets and simulation associations. UML schema describes one-to-one (1..1) and one-to-many(1..*) relationships.

Although individual domain dictionaries may contain their own biases, we can minimize the effect of those differences by extracting a consensus from a group of such dictionaries. We previously demonstrated the application of this method to SCOP, CATH, and the Dali Domain Dictionary (Dietmann *et al.,* 2001) to generate a consensus domain dictionary (CDD 2003 version, v2003) (Day *et al.*, 2003). This domain dictionary was the basis of our initial high-throughput survey of native dynamics (Beck *et al.*, 2008). Additionally, the concept of the metafold that we introduced in the v2003 CDD was further developed in a study of 'cradle-loop' structures (Alva *et al.*, 2008). A subset of the representative domains from our v2003 CDD was used to conduct benchmark simulations of standard molecular dynamics (MD) force fields (Rueda *et al.*, 2007).

Here we present an updated CDD (v2009) derived using recent versions of the input domain dictionaries, which incorporate many new structures determined since the v2003 CDD was created. The CDD is the backbone of our high-throughput molecular dynamics initiative, Dynameomics (Beck *et al.*, 2008; van der Kamp *et al.*, 2010). This project seeks to simulate the native and unfolding behavior of representatives of all protein folds. Consequently, we need an objective basis for selection of simulation targets. Therefore, it is important that the CDD be monitored so that we can identify novel topologies as they are classified and observe potential splits within, and mergers between, our metafolds as classifications shift. It is important that we identify domains that appear to be autonomous units, since we use the contents of the CDD as potential targets for simulation of folding/unfolding pathways. The selection process was complicated by the discovery that roughly a third of the consensus folds (metafolds) in the CDD are not autonomous structural units, but instead are dependent components of multi-domain or complex structures (or are small structural motifs).

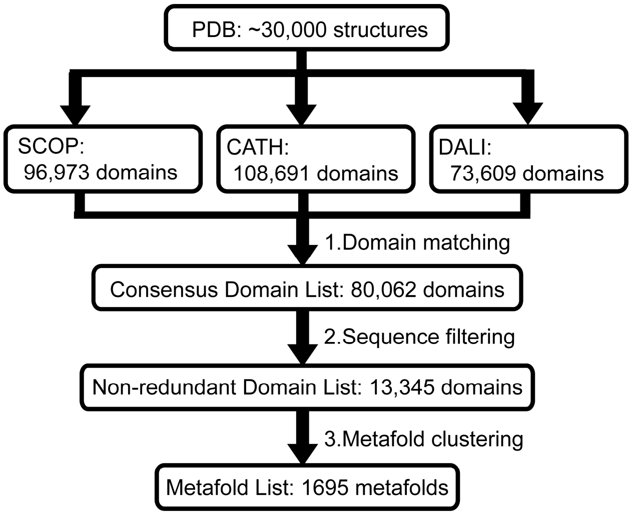


Figure 2. Overview of the consensus domain dictionary (CDD) generation process. Consensus domains are first found between pairs of input dictionaries. The resulting domain list is filtered for sequence identity. The resulting non-redundant domain list is clustered into a list of metafolds. The collected domain lists and metafold list are the contents of the CDD.

Here we present our data model for representing domains and their metafolds over time in a relational database (Simms *et al.*, 2008). We discuss the use of this data model to map domains and their annotations from older versions of our dictionary to the newer one (v2003 → v2009). We present the full v2009 CDD consisting of 1695 metafolds. We then filter the set to remove metafolds that do not represent autonomous units or cannot be simulated for other reasons, which yields 807 metafolds. In addition to being of use to our Dynameomics efforts, the filtered 807 target list is more appropriate for bioinformatics studies investigating globular protein properties than the full consensus domain dictionary or the three parent domain dictionaries by removing folds that do not represent autonomous folded structures.

2 METHODS

2.1 Relational model for consensus set data

The relational schema for the 'Target Selection and Preparation' (or 'Prep') database, which houses our CDD, is shown in Figure 1 in a universal markup language representation (Simms *et al.*, 2008). Consensus domains are stored in the *Domain* table consisting of an identifier, PDB code, and fold identifiers from the SCOP, CATH, and Dali domain dictionaries. A domain must contain fold identifiers from at least two of the three input domain dictionaries. Metafold data are stored in the *Fold* table, which contains a metafold identifier, name, and the metafold's rank (based on domain population). Note that the *Fold* table is, in fact, a table of metafolds. There may be multiple versions of the same domain

Downloaded from bioinformatics.oxfordjournals.org at Universidade Cat?lica de Bras?lia on November 18, 2010

in the *Domain* table (due to multiple CDD versions), and these differing versions may link to multiple metafolds (also due to multiple CDD versions). The many-to-many relationship between *Fold* and *Domain* is implemented via the *Fold_Domain* table. Metafold representatives chosen for simulation are captured in the *Target* table.

As previously stated, domain classifications evolve over time, which can cause changes in the CDD. To capture these changes, the *Fold*, *Fold_Domain*, *Domain*, and *Target* tables include a consensus set identifier to allow multiple versions of metafold and domain definitions to be stored in the same primary tables. To facilitate cross-consensus set queries, fold identifiers are maintained across consensus set releases where this is meaningful. It is possible for new identifiers to be introduced and existing identifiers to be removed in subsequent releases.

Domains and targets are both linked to external data sources. The *Domain* table contains a field for the PDB code, and we populate a local cache table (*PDB*) with specific information synthesized from a given structure's PDBml (Westbrook *et al.*, 2005). Examples include a structure title, dates, methods, and source organism. These fields facilitate local searches and analysis. The *Target_Simulation* table links targets in the Prep database to simulations contained in the Dynameomics data warehouse (Simms *et al.*, 2008).

2.2 Generation of the v2009 CDD

The v2009 CDD was generated as described by Day *et al*. (2003). To generate the CDD, we integrate recent versions of three major domain dictionaries: SCOP (Andreeva *et al.*, 2008), CATH (Cuff *et al.*, 2009), and Dali (Dietmann *et al*., 2001). SCOP v1.73, CATH v3.2, and a March 2005 download of the Dali Domain Dictionary were used as input for consensus generation. CDD generation is a two-step process: First, consensus domains are generated by pairwise comparison between domain dictionaries of residue ranges from the same chain. Where a significant overlap between input domains is detected, a consensus domain is assigned. Second; the set of consensus domains is filtered for sequence similarity and then clustered into a set of metafolds based on their composite fold identifiers. The set of consensus domains and metafolds comprise our CDD. The workflow of this process is outlined in Figure 2.

Our domain matching procedure follows the criteria specified by Dietmann and Holm (2001). A given domain in one input dictionary is compared against analogous domains in the other domain dictionaries. Where the given domain and an analogous domain both overlap to a significant extent (80%) a consensus domain pair is assigned. If a given domain matches domains from both other input dictionaries, the three resulting domain pairs are collapsed into a single consensus domain spanning analogous domains from all three input dictionaries. If a domain from any single domain dictionary has no consensus with any domain from either of the remaining domain dictionaries, it is discarded. Each consensus domain preserves the source data from its input dictionaries (PDB, chain, residue range, and fold identifier). This list is loaded into our database to assist with metafold representative selection and report generation. The schema is illustrated in Figure 1.

The full domain list is filtered by sequence using the SCOP ASTRAL95 sequence-filtered domain list and the CATH 'SOLID' sequence identifiers (Chandonia *et al.*, 2004; Greene *et al.*, 2007). The non-redundant domain list produced by the sequence filter is used as the basis for generation of metafolds. Each domain contains a composite fold identifier derived from its input domain definitions. SCOP and CATH are hierarchal classifications, for SCOP we chose the 'Fold' level to cluster, for CATH we chose the 'Topology' level. Domains whose composite fold identifiers share two of three elements are clustered together into a metafold.

2.3 Mapping between CDD versions

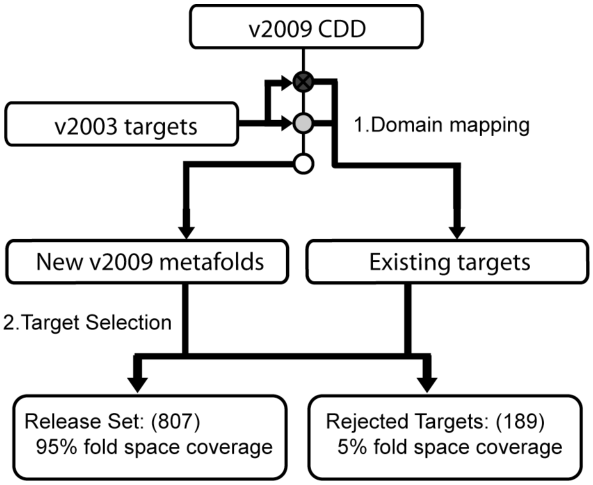


Figure 3. Overview of the mapping and target selection process. Existing v2003 targets are mapped to the v2009 CDD. (1)Where a mapped domain was selected or rejected in the v2003 CDD, this status is maintained in the v2009 CDD. (2)Where a new metafold is observed, targets are selected from available domains in that metafold. Of the 1695 metafolds in the v2009 CDD, 699 were rejected for being not autonomous or membrane proteins. Considering the remaining 189 rejected metafolds with the 807 selected, the domains from the selected metafolds represent 95% of the non-autonomous, non-membrane domains in the CDD.

The CDD is a product of clustering across input domain dictionaries. As these input dictionaries change with the release of new versions, so should the CDD. However, without a detailed description of the changes made, it can be difficult to assign equivalence between two domains from CDDs generated from different inputs. A mapping between the v2003 and v2009 CDD was generated based on domain identifier and fold identifier equivalence (Figure 3). Changes in fold representation in new versions of both CATH and Dali motivated the mapping criteria. Between the release of CATH v2.4 and v3.0, "working" CATH classes [6-9] were no longer included in production releases (Greene *et al.*, 2007). Since the v2003 CDD included these classes, criteria were chosen such that v2003 domains could be reassigned to regular (1-4) CATH classes. Since fold identifiers do not persist between v3.1β and the March 2005 version of Dali, identity between these versions could not be used as the basis for a mapping. In the period of time since we acquired this version of Dali, this domain dictionary has been discontinued (Holm *et al.*, 2008).

Four mapping criteria were defined based on the mapping classification a domain possessed in the v2003 CDD. A v2003 domain possessing composite SCOP, CATH, and Dali fold identifiers is mapped to a v2009 domain if both the SCOP and CATH composite chain and domain (PDB6) and the v2009 Dali PDB6 is defined (though not necessarily equivalent). A v2003 domain possessing only SCOP and CATH fold identifiers is mapped to a v2009 domain if both the SCOP PDB6 identifier and CATH PDB6 identifier are equivalent. A v2003 domain possessing only CATH and Dali fold identifiers is mapped to a v2009 domain if the CATH PDB6 and fold identifiers are equivalent and the Dali fold identifier and PDB6 is defined. A v2003 domain possessing only SCOP and Dali identifiers is mapped to a v2009 domain if the SCOP PDB6 and fold identifiers are equivalent and the Dali fold identifier and PDB6 is defined.

2.4 Selection of domains as metafold representatives

Mapped metafolds are sorted and ranked by their non-redundant population (Figure 4). We examined domains within each metafold to assess

Downloaded from bioinformatics.oxfordjournals.org at Universidade Cat?lica de Bras?lia on November 18, 2010

their suitability as a simulation target. We chose targets that were self-contained domains in a single protein chain that were less than 450 residues in length. Where the structure was determined by X-ray crystallography, we only chose crystal structures with resolutions higher than 3.0 Å. Domains with obligate cofactors (other than Zn^{2+}, Ca^{2+}, and heme) were rejected. Also, domains with multiple Zn^{2+}, Ca^{2+}, and heme sites were rejected, with a few exceptions (e.g. calbindin). Many of the domains rejected for this reason are chains where the cofactor is a major structural element. Domains with a single Zn^{2+}, Ca^{2+}, or heme were selected (i.e. myoglobin) regardless of whether folding information was available regarding the role of the cofactor. These particular cofactors and proteins were included because they have been the subject of numerous biophysical, biochemical, and folding studies. When multiple domains within a metafold met our selection criteria, we preferred domains with biomedical relevance or with experimental folding studies available for comparison. The workflow for target selection where targets exist from a previous CDD is outlined in Figure 3.

The determination of whether a given domain was self-contained was primarily determined by manual inspection. Several factors could lead to the rejection of a domain as not self-contained; these factors could occur either in isolation or in concert with one another. Where a domain was a component of a multi-domain structure, we used a simple “sheet of paper” test to determine whether there was a clean interface between the domain of interest and the rest of the protein. Where a domain could not be cleanly separated from the remainder of the protein, it was rejected because of its convoluted interface. In addition, we examined the proposed biological unit from the deposited transform. Structures with extensive domain swapping or crystal contacts were rejected. Furthermore, structures that were ‘irregular’ (those that possessed little to no structure or hydrophobic core) could also be rejected for being not self-contained. This range of factors led to a broad spectrum of possible buried surface area in rejected metafolds (10% - 60%). Furthermore, where the domain boundary occurred in the middle of a significant secondary structure element (helix or beta sheet), this disruption could be used as a reason for rejection as not self-contained. This was not used as a basis for rejection where the secondary structure was a linking region and could be safely truncated to the previous loop region and where that truncation would not expose significant hydrophobic surface area. Examples of non-autonomous domains are presented in Figure 5.

Where a single suitable domain was selected as a target for simulation it was designated a representative for its metafold. If, after examining all domains within a metafold, a suitable domain could not be found, a domain was chosen as a fold representative and the reasons for its rejection were annotated. Once a domain was selected as a metafold representative, we chose a residue range to simulate that incorporated the input domain definitions such that we avoided disrupting secondary structure elements while removing long, unstructured tails (many of which are cloning artifacts). The distribution by rank of rejected and accepted metafolds is illustrated in Figure 4E.

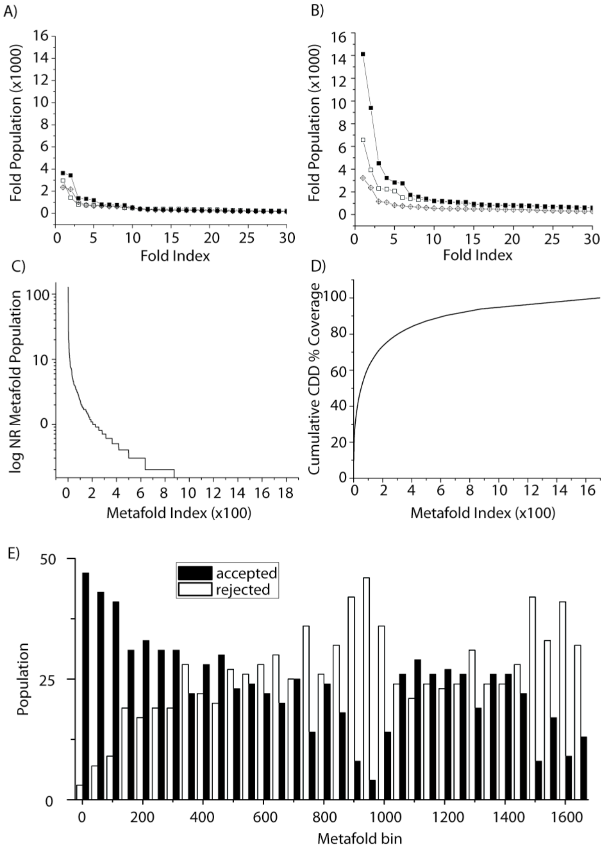


Figure 4. Distribution of domain populations between folds and metafolds. A) Population distribution of top 30 most populated folds in the SCOP (filled squares), CATH (open squares), and DALI (crossed diamond) dictionaries for the v2003 CDD. B) Population distribution of top 30 most populated folds in the SCOP, CATH, and DALI dictionaries for the v2009 CDD. C) Non-redundant population distribution of the 1695 metafolds in the v2009 CDD. D) Cumulative percentage of domains represented by metafold rank. The most populated metafolds account for a large percentage of the domains in the CDD. E) Ranked metafolds binned into 50-rank bins. For example, in the first 50 metafolds, 47 were selected and 3 were rejected.

Table 1. Summary statistics of the SCOP, CATH, and Dali domain dictionaries used in the v2003 and v2009 CDD.

	Chains (C)	Domains (D)	Folds[a]	D/C[b]
v2003				
SCOP	27,308	35,095	783	1.29
CATH	25,622	36,480	1,453	1.42
Dali	21,493	35,492	1,088	1.65
v2009				
SCOP	74,608	96,973	1,280	1.29
CATH	74,240	108,691	1,110	1.46
Dali	52,740	73,609	2,783	1.39

[a] Number of unique folds at the chosen level within each domain dictionary

[b] Number of distinct domains (D) per distinct chain (C)

3 RESULTS

3.1 v2009 Consensus Domain Dictionary

The CDD consists of a set of consensus domains and a list of consensus fold identifiers binding these domains together into metafolds (Figure 2). Consensus domains were identified between pairs of domain dictionaries (SCOP/CATH, SCOP/DALI, Dali/CATH). Summary statistics from each of the domain dictionaries are presented in Table 1. The agreement between domain dictionaries was measured as the fraction of shared con-

Downloaded from bioinformatics.oxfordjournals.org at Universidade Cat?lica de Bras?lia on November 18, 2010

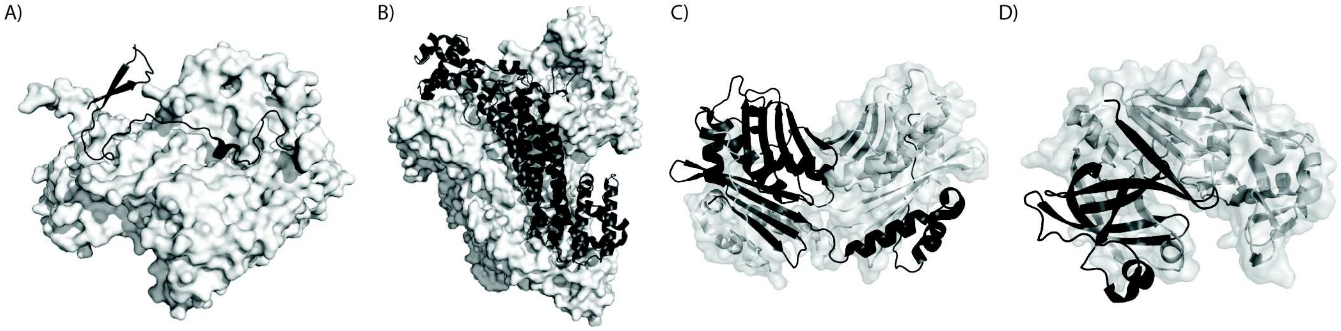


Figure 5. Example metafolds rejected for not being autonomous units. A) Metafold #232, chain 4 of P1/Mahoney poliovirus mutant (1AL2). B) Metafold #2232, Chain A of δ-crystallin I (1I0A). C) Metafold #489, chain B of HSP33 (1HW7). D) Metafold #172, Chain C of cathepsin D (1LYA).

sensus domains divided by the total number of domains originating from structures shared between the two dictionaries.

We reduced the effect of differing release dates (and thus different numbers of structures) by considering only shared structures. CATH and Dali have the highest agreement, with 96% of CATH domains and 90% of Dali domains included in the CATH/Dali consensus domain set. SCOP and CATH have the next highest agreement, with 79% of SCOP domains and 82% of CATH domains in the SCOP domains in the SCOP/CATH consensus domain set. Finally, SCOP and Dali had the lowest agreement, with 65% of SCOP domains and 61% of Dali domains included in the SCOP/Dali consensus domain set. A consensus domain need not exist solely between a single pair of domain dictionaries. Where a consensus domain was determined by each of the three pairwise comparisons, it was collapsed into a single triple consensus domain in the CDD. Thus, four classes of consensus domains were created, SCOP/CATH, SCOP/Dali, Dali/CATH, and SCOP/CATH/Dali.

The v2009 CDD is composed of 80,062 domains, originating from 27,140 PDB structures. The total number of PDB structures considered is lower than the structures available due to the lag between PDB and domain dictionary releases. The domains of the CDD were distributed among the aforementioned classes as follows: 51% SCOP/CATH/Dali, 30% SCOP/CATH, 10% CATH/Dali, and 9% Dali/CATH. To generate the metafold list, the CDD was first filtered by sequence identity (Figure 2). The nrCDD was composed of 13,345 domains. The domains in the CDD clustered into 1695 metafolds. On the whole, these metafolds incorporate 4217 unique consensus fold identifiers derived from 971 unique SCOP folds, 923 unique CATH topologies, and 2362 Dali folds. The distribution of domains per fold for the input domain dictionaries is shown in Figure 4A,B.

3.2 Comparison of v2009 and v2003 CDDs

Both the residue range of a domain and its fold classification can change over time. These changes affect the output of the metafold clustering and the domain contents of the CDD. Since our Dynameomics simulations are indexed against the CDD, it is necessary to track domains across multiple dictionary versions so that information about our simulated domains is current. Where possible, we generated a map between domains in our v2003 and v2009 CDD based on their fold identifiers. There were 31,141 domains in the v2003 CDD. From this dictionary, 4,693 domains could not be mapped forward from v2003 to v2009 and are considered obsolete (discussed below). 26,448 domains were mapped from v2003 to v2009. There are 53,614 new domains in the v2009 dictionary that are not in the v2003 dictionary.

The domains that were not mapped from the v2003 CDD can be broadly partitioned into three categories: (1) domains from structures that were dropped from consideration in one of the input domain dictionaries, (2) domains whose boundaries changed significantly in one of the input domain dictionaries, and (3) domains that were split into multiple domains or merged into a single domain. From each of our input dictionaries used in v2003 CDD, 95% of the structures considered also had at least one domain in the input dictionaries used in the v2009 CDD. The ~5% of structures that were in the v2003 CDD but not in the v2009 CDD had the following properties: the structure was deemed obsolete by the PDB, the structure consisted primarily of nucleic acids, or the structure was a purely computational model. Of those chains that were removed from consideration that were not part of the aforementioned dropped structure set, the majority are rare cases arising from the presence of synthetic linkers and/or multi-chain domains arising from viral capsid structures. In some cases where neither the chain nor structure containing a domain was dropped, but it could still not be mapped, the domain boundaries in the structure were significantly altered. Alternatively a domain was split into multiple domains or merged with other domains. Although we can observe these transitions, we prefer to treat the resulting domain(s) as new. The 4,393 dropped domains from the v2003 → v2009 CDD mapping originated from 2,198 PDB structures. 3,314 of those v2003 domains originate from PDB structures that still contain domains in the v2009 CDD. There are 1,379 v2003 domains originating from 608 PDB structures not found in the v2009 CDD. 319 of these v2003 domains originate from structures that were superseded by newer structures in the PDB. The remaining 1,060 domains are dropped either because they were removed from one of the input dictionaries, or because the domain definition was changed in one or more of the input dictionaries, breaking the original v2003 consensus.

Domains that were mapped from v2003 to v2009 met specific criteria for their particular class (SCOP/CATH, SCOP/Dali, etc.). Of the 26,448 mapped domains, 15,735 were mapped us-

Downloaded from bioinformatics.oxfordjournals.org at Universidade Cat?lica de Bras?lia on November 18, 2010

ing the SCOP/CATH/Dali class, 7,736 were mapped using the SCOP/CATH class, 1,734 were mapped using the SCOP/Dali class, and 995 were mapped using the CATH/Dali class. A majority of the domains in our CDD could be mapped based on their SCOP and CATH identifiers alone. The mapped domains originated from 11,896 PDB structures, leading to an average of 2.23 mapped domains per PDB structure. The mapped domains originate from 857 metafolds in the v2003 CDD and are mapped into 719 metafolds in the v2009 CDD, indicating that some v2003 metafolds and their domain contents were merged into larger v2009 metafolds. Many domains were also folded into larger metafolds as they gained a third input fold identifier. 6,613 mapped domains with defined SCOP, CATH, and Dali domain identifiers in the v2009 CDD contained only two fold identifiers in the v2003 CDD.

'New' domains are those that exist in the v2009 CDD and did not exist in the v2003 CDD. The 53,614 new domains originate from 17,949 PDB structures. These new domains fall into 1565 metafolds. There were 976 metafolds in the v2009 CDD that consisted entirely of new domains, 589 metafolds composed of a mix of mapped and new domains, and 130 metafolds that consist entirely of mapped domains. A majority of the new v2009 domains were placed into metafolds with other mapped domains. 8,401 v2009 domains fell into metafolds composed solely of new domains. The domain population was less than five for 633 of the new v2009 metafolds. The drop off in population with increasing rank and the large number of singleton folds is shown in Figure 4C, which leads to greater coverage of fold space for the top ranked folds and limited additional coverage provided by the low populated high ranks (Figure 4D).

3.3 SCOP and CATH in the v2009 CDD

The consensus generation process can separate an input fold into multiple metafolds or merge multiple input folds into a single metafold. We examined the location of input folds from SCOP and CATH within the CDD closely because it indirectly addresses the continuity of fold space. This analysis also serves as an internal check of the consistency of our metafold clustering method. The domains of an input fold can be distributed into multiple metafolds and/or combined into a metafold with domains from other input folds. To quantify this effect, we analyzed the number of metafolds into which an input fold and its domains are distributed. An input fold can be distributed over many metafolds and yet the vast majority of that fold's domains can still be assigned to a single metafold. Thus, we are primarily interested in the fractional domain population of the metafold containing the majority of an input fold's domains, or the 'most populated metafold.' The net effect of this treatment is that outliers within a fold are partitioned into their own poorly populated or singleton metafolds (metafolds containing only a single domain).

Certain structurally variable topologies (such as the Rossmann folds) are split more evenly across a number of metafolds. The 860 input SCOP folds were spread over 815 CDD metafolds. 12 of these metafolds contained multiple SCOP input folds. The metafold containing the most SCOP input folds was metafold #2 (consisting of a number of Rossmann folds), followed by metafold #16 (consisting of parallel α-helical bundles), and metafold #1 (consisting of IgG-like β-sandwiches). These SCOP folds are bound together by highly populated CATH topologies. A full listing of merged SCOP folds is provided in Table S1. 815 metafolds contained only a single SCOP fold. Of these 815 metafolds, 290 also contained only a single non-redundant domain. We also examined those SCOP folds where the most populated metafold contained a diminished fraction of the total domains, indicating that the SCOP fold was distributed across multiple metafolds. 112 of the input SCOP folds had a fractional population within the most populated metafold of 80% or less. The significance of this fraction can vary, however, if the input fold is poorly populated or if the input fold was not a child of one of the 4 main structural classes (all-α, all-β, α+β, or α/β).

The 892 input CATH folds were distributed over 862 metafolds. 26 metafolds contained domains from multiple CATH folds. The most populated metafold, consisting of IgG-like β-sandwiches, contained four CATH folds. Metafolds #16 and #46 contained three CATH folds. The remaining 23 metafolds each contained two CATH folds. The most populated metafold of the 30 most populated CATH folds is presented in Table S2. Of the 866 metafolds containing only a single CATH fold, 277 also contained only a single nonredundant domain, signifying singleton metafolds. The CATH Rossmann fold (3.40.50) was the most populated of the CATH folds that were significantly distributed over multiple metafolds. This fold was distributed over 42 metafolds, and the most populated metafold of these (#2) contained only 49% of the input fold.

The v2009 CDD has 881 unique SCOP folds from the 11 different SCOP classes (all-α, all-β, α+β, α/β, multidomain α and β, membrane and cell surface, small proteins, coiled coil, low resolution, peptides, and designed) There were 434 SCOP folds that only appeared in metafolds with a simulated metafold representative and 332 SCOP folds that were only found in rejected metafolds. The rejected SCOP folds represent about a third of the folds from each of the top four classes (all-α, all-β, α+β, α/β) found in our CDD, between 27 to 38% of each class. We rejected approximately 70% of each of the multidomain and membrane classes in our set. Similarly, there are 894 CATH topologies in our domain dictionary from the four CATH classes: mainly-α, mainly-β, mixed α-β, and irregular/few secondary structures. The majority (77%) of the irregular class CATH topologies are only found in rejected metafolds. The other three CATH classes all had between 36-47% of topologies found only in rejected metafolds. These classes had a similar number of topologies found only in selected metafolds (40-55%). This analysis of the SCOP and CATH folds reveals that we have not biased our set of selected metafolds towards any fold class or systematically rejected any class, except for unstructured peptides and membrane proteins.

3.4 Selection of Metafold Representatives

Our primary purpose in creating the CDD was to facilitate the simulation of both the native state dynamics and the unfolding behavior of at least one domain from each metafold. As such, we examined domains from each metafold to find a high quality structure suitable for simulation. Such a domain was then selected as a 'metafold representative', or target, of that metafold and prepared for simulation. If no suitable domain could be found we chose one domain from the metafold to represent the reason that `the metafold was rejected (see Figure 5, Table 2).

Downloaded from bioinformatics.oxfordjournals.org at Universidade Cat?lica de Bras?lia on November 18, 2010

The selected representatives for the Top 30 most populated metafolds are presented in Figure 6, the full target set is provided in Table S3. Selected representatives could come from a variety of structural contexts; 387 representatives were the full contents of their PDB structure deposition, 165 representatives were a full chain from a multi-chain deposition, and 165 representatives were excised domains where a chain was chopped to select the domain.

We identified at least one domain suitable for simulation from 807 of 1695 metafolds in the v2009 CDD. Of the remaining 888 metafolds, 585 metafolds consisted of domains that were not self-contained and 87 metafolds consisted of domains that were irregular. Of these 672 metafolds, none were autonomous units (75% of the rejected metafolds or 40% of the total number of metafolds). A summary of the reasons a domain from a metafold was rejected is presented in Table 2. These rejected domains fell into three categories: domain-swapped dimers, domains with a large buried interface in the experimentally determined structure of a complex, and domains with secondary structure elements that continue into other domains of the protein (Figure 5). There was no significant bias in the rejected metafolds with respect to major fold class (all α, all β, mixed α/β). In 11 metafolds, no domains of less than 450 residues were present so the metafold was rejected for reasons of size. In 27 cases, the domains of the metafolds in question were contained a transmembrane region. There were 54 metafolds whose domains required an obligate cofactor. There were 85 metafolds where each of the domains contained a large (greater than 7 residue) gap and were rejected. In 87 cases, the metafold consisted of domains that lacked regular secondary elements and/or were unstructured peptides. In 20 metafolds, all domains had a resolution lower than 3.0 Å. Finally there were two singleton metafolds that were rejected because their domains were of disputed structural validity at the time of writing (Murthy *et al.*, 2009).

In 19 cases, we selected a domain but the resulting native state simulation was not stable and the metafold was rejected. Interestingly, the 19 starting structures were all determined by NMR. For 5 of the 19 targets alternate fold representatives in the form of high-resolution crystal structures were available; the resulting 'replacement' simulations were stable. In the other 14 cases, it was either the only structure for the metafold (with rank 633 or higher) or the alternatives were older PDB entries of equal or lesser quality. In other words for these 'rejected by simulation' cases, no alternative replacement could be found from their respective metafolds, but it would appear to be a problem with the starting structures and not necessarily the simulations (See van der Kamp *et al.*, 2010 for more details).

4 DISCUSSION

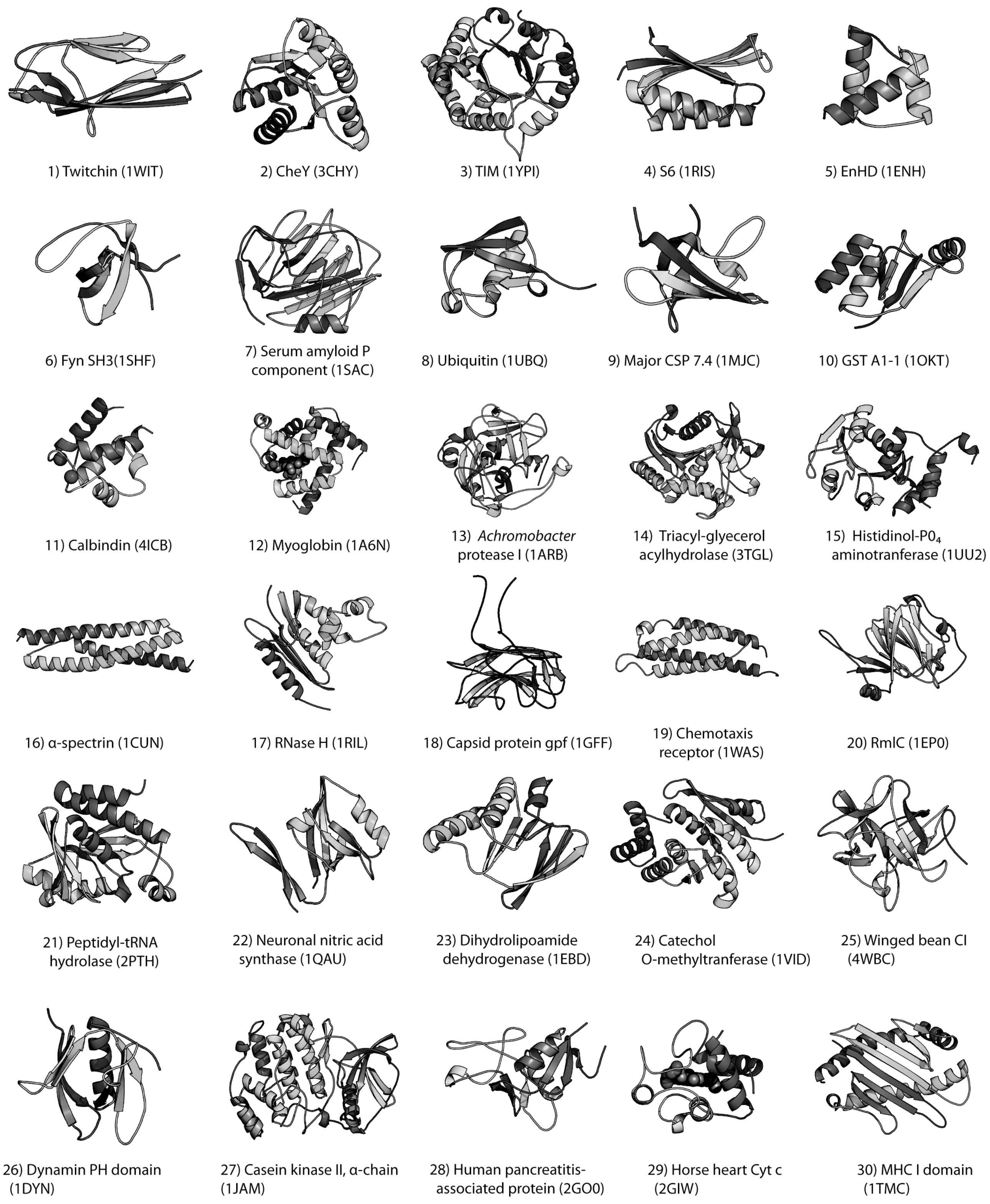
The recognition of spatially distinct motifs and structural patterns is a long-standing component of structural protein studies (Phillips, 1967; Wetlaufer, 1973). The understanding of the term 'domain' to denote an autonomous, structurally cohesive unit is similarly well established (Levitt and Chothia, 1976). However, the multiple extant definitions for 'domain' do not always converge (Majumdar *et al.*, 2009; Sowdhamini and Blundell, 1995). A spatially distinct region within a structure may not coincide with an autonomous, stable unit. Our interest in domain dictionaries is to establish a systematic, broad sampling of topologies that satisfy our autonomy criterion. The single most striking conclusion from this endeavor was that a significant fraction of metafolds generated by our consensus method contained no suitable for simulation. This occurred due to a variety of factors, but the single largest reason for rejection was that the domain was not self-contained. Identification of protein domains can be split into two problems: the partition of a chain into multiple domains, and the separation of domains into folds. The difficulty of partitioning a chain into domains has been well studied (Holland *et al.*, 2006; Veretnik *et al.*, 2004). The separation of domains into fold has been similarly examined. Both problems share similar elements. It may be that the smallest repeating structural element observed between two structures is not necessarily a shared domain. For example, if chain discontinuity is allowed within a domain to increase structural similarity of the domains in a fold, then the structural integrity of the excised region may be sacrificed. The problem becomes more complex when considering domains that are solely observed in the context of multimeric structures or in complexes. In our opinion, one must be very careful to consider the effect inadvertent inclusion of such domains may have on bioinformatics studies; they are not independent, globular structures. We note that the distribution of autonomous and non-autonomous domains is not necessarily related to the dependent or independent folding of these domains in nature. Indeed, characterizing the unfolding behavior of the autonomous domains is one of the primary goals of the simulations we have performed of these domains.

Table 2. Justifications for rejection of 888 metafolds in the v2009 CDD.

Reject Reason	Definition	Metafolds
Not an autonomous domain	Poor interface, continuation of secondary structure into other domains, small with little secondary structure	672
Large gaps	Backbone gap of more than 7 residues	85
Non-parameterized co-factors or structural ions	Structurally necessary non-protein molecules have not been parameterized	57
Membrane	Domain penetrates membrane	27
Size	Larger than 450 residues	11
Resolution	Resolution lower than 3.0 Å	20
Rejected by simulation	Did not pass native (298 K) simulation quality control	14
Other	Structures are in dispute[a]	2

[a] Structure 1BEF was retracted from the PDB, causing rejection of domains 1BEFA01 and 1BEFA02. (Murthy *et al.*, 2009)

Downloaded from bioinformatics.oxfordjournals.org at Universidade Cat?lica de Bras?lia on November 18, 2010



1) Twitchin (1WIT)

2) CheY (3CHY)

3) TIM (1YPI)

4) S6 (1RIS)

5) EnHD (1ENH)

6) Fyn SH3(1SHF)

7) Serum amyloid P component (1SAC)

8) Ubiquitin (1UBQ)

9) Major CSP 7.4 (1MJC)

10) GST A1-1 (1OKT)

11) Calbindin (4ICB)

12) Myoglobin (1A6N)

13) *Achromobacter* protease I (1ARB)

14) Triacyl-glyecerol acylhydrolase (3TGL)

15) Histidinol-P0$_4$ aminotranferase (1UU2)

16) α-spectrin (1CUN)

17) RNase H (1RIL)

18) Capsid protein gpf (1GFF)

19) Chemotaxis receptor (1WAS)

20) RmlC (1EP0)

21) Peptidyl-tRNA hydrolase (2PTH)

22) Neuronal nitric acid synthase (1QAU)

23) Dihydrolipoamide dehydrogenase (1EBD)

24) Catechol O-methyltranferase (1VID)

25) Winged bean CI (4WBC)

26) Dynamin PH domain (1DYN)

27) Casein kinase II, α-chain (1JAM)

28) Human pancreatitis-associated protein (2GO0)

29) Horse heart Cyt c (2GIW)

30) MHC I domain (1TMC)

Figure 6. Structures of the representative domains of the 30 most populated metafolds in the 2009 CDD (Top 30). Domains are named based on their source structure, where a domain was an excised chain or domain, it is named according to the PDB-deposited name for its chain. A color version of this figure is provided as supplementary information.

Downloaded from bioinformatics.oxfordjournals.org at Universidade Cat?lica de Bras?lia on November 18, 2010

We have generated a consensus domain dictionary from three major domain dictionaries. This CDD contains 1695 metafolds. We have inspected each metafold and selected a representative. These representatives constitute our release set, which consists of 807 'simulatable' domains. These 807 metafolds represent 81% (64,700) of the domains in our CDD, or 95% of the known autonomous protein folds (Table 2). This set of domains is the basis for our high-throughput MD simulation of representatives of all globular protein folds (Beck *et al.*, 2008; van der Kamp *et al.*, 2010). Also, to reduce artifacts, we would suggest that the reduced list of 807 metafolds be used for bioinformatics studies, not the full CDD, nor the domain dictionaries from which they were derived.

ACKNOWLEDGEMENTS

We are grateful for financial support provided by the National Institutes of Health [GM50789 to V.D., TG 3 T15 LM007442-04S1 to A.S.] and Microsoft through the External Research Program at Microsoft Research, www.microsoft.com/science

REFERENCES

Alva V., *et al.* (2008) Cradle-loop barrels and the concept of metafolds in protein classification by natural descent, *Curr. Opin. Struct. Biol.*, **18**, 358-365.

Andreeva, A., *et al.* (2008) Data growth and its impact on the SCOP database: new developments, *Nucleic Acids Res.*, **36**, D419-425.

Beck, D.A., *et al.* (2008) Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations, *Protein Eng. Des. Sel.*, **21**, 353-368.

Chandonia, J.M., *et al.* (2004) The ASTRAL Compendium in 2004, *Nucleic Acids Res.*, **32**, D189-192.

Coulson, A.F. and Moult, J. (2002) A unifold, mesofold, and superfold model of protein fold use, *Proteins*, **46**, 61-71.

Csaba, G., *et al.* (2009) Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis, *BMC Struct. Biol.*, **9**, 23.

Cuff, A.L., *et al.* (2009) The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies, *Nucleic Acids Res.*, **37**, D310-314.

Day, R., *et al.* (2003) A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary, *Protein Sci.*, **12**, 2150-2160.

Dietmann, S., *et al.* (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3, *Nucleic Acids Res.*, **29**, 55-57.

Greene, L.H., *et al.* (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution, *Nucleic Acids Res*, **35**, D291-297.

Grishin, N.V. (2001) Fold change in evolution of protein structures., *J. Struct. Biol.*,**134**, 167-185.

Hadley, C. and Jones, D.T. (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP, *Structure*, **7**, 1099-1112.

Holland, T.A., *et al.* (2006) Partitioning protein structures into domains: why is it so difficult?, *J. Mol. Biol.*, **361**, 562-590.

Holm, L., *et al.* (2008) Searching protein structure databases with DaliLite v.3, *Bioinformatics*, **24**, 2780-2781.

Jefferson, E.R., *et al.* (2008) A comparison of SCOP and CATH with respect to domain-domain interactions, *Proteins*, **70**, 54-62.

Kendrew, J.C., *et al.* (1960) Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution, *Nature*, **185**, 422-427.

Kolodny, R., *et al.* (2006) Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction, *Curr. Opin. Struct. Biol.*, **16**, 393-398.

Levitt, M. and Chothia, C. (1976) Structural patterns in globular proteins, *Nature*, **261**, 552-558.

Majumdar, I., *et al.* (2009) A database of domain definitions for proteins with complex interdomain geometry, *PLoS One*, **4**, e5084.

Krishna Murthy, H. M., *et al.* (2009). Dengue virus NS3 serine protease. Crystal structure and insights into interaction of the active site with substrates by molecular modeling and structural analysis of mutational effects. *J. Biol. Chem.* **284**, 34468-34468.

Murzin, A.G., *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, **247**, 536-540.

Nagano, N., *et al.* (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions, *J. Mol. Biol.*, **321**, 741-765.

Orengo, C.A., *et al.* (1997) CATH--a hierarchic classification of protein domain structures, *Structure*, **5**, 1093-1108.

Pascual-Garcia, A., *et al.* (2009) Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures, *PLoS Comput. Biol.*, **5**, e1000331.

Perutz, M.F., *et al.* (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis, *Nature*, **185**, 416-422.

Phillips, D.C. (1967) The Hen-White Lysozyme Molecule, *Proc. Natl. Acad. Sci.*, **57**, 483-495.

Rueda, M., *et al.* (2007) A consensus view of protein dynamics, *Proc. Natl. Acad. Sci.*, **104**, 796-801.

Sam, V., *et al.* (2006) ROC and confusion analysis of structure comparison methods identify the main causes of divergence from manual protein classification, *BMC Bioinformatics*, **7**, 206.

Schaeffer, R.D., *et al.* (2010) Protein folds and protein folding, *Prot. Eng. Des. Sel.*, (2010), in press.

Simms, A.M., *et al.* (2008) Dynameomics: design of a computational lab workflow and scientific data repository for protein simulations, *Protein Eng. Des. Sel.*, **21**, 369-377.

Sowdhamini, R. and Blundell, T.L. (1995) An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins, *Protein Sci*, **4**, 506-520.

Valas, R.E., *et al.* (2009) Nothing about protein structure makes sense except in the light of evolution, *Curr. Opin. Struct. Biol.*, **19**, 329-334.

van der Kamp, M.W., *et al.* (2010) Dynameomics: A comprehensive database of protein dynamics, *Structure,* **18**, 423-435.

Veretnik, S., *et al.* (2004) Toward consistent assignment of structural domains in proteins, *J. Mol. Biol.*, **339**, 647-678.

Westbrook, J., *et. al.*. (2005) PDBML: the representation of archival macromolecular structure data in XML, *Bioinformatics*, **21**, 988-992.

Wetlaufer, D.B. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins, *Proc. Natl. Acad. Sci.*, **70**, 697-701.

Wolf, Y.I., *et al.* (2000) Estimating the number of protein folds and families from complete genome data, *J. Mol. Biol.*, **299**, 897-905.

Downloaded from bioinformatics.oxfordjournals.org at Universidade Cat?lica de Bras?lia on November 18, 2010