Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Molecular Graphics and Modelling 24 (2006) 426–433
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Clustering of domains of functionally related enzymes in the interaction database PRECISE by the generation of primary sequence patterns

Melissa R. Landon [a], David R. Lancia Jr.[b], Karl H. Clodfelter [a], Sandor Vajda [a,b,*]

[a] *Graduate Program in Bioinformatics and Systems Biology, Boston University, Boston 02215, MA, USA*
[b] *Department of Biomedical Engineering, Boston University, 44 Cummington Street, Boston 02215, MA, USA*

Received 5 July 2005; received in revised form 3 August 2005; accepted 16 August 2005
Available online 10 October 2005

Abstract

The PRECISE database was developed by our laboratory to allow for the systematic study of the ligand interactions common to a set of functionally related enzymes, where an interaction site is defined broadly as any residue(s) that interact with a ligand. During the construction of PRECISE, enzyme chains are extracted from the protein data bank (PDB) and clustered according to functional homology as defined by the enzyme commission (EC) nomenclature system. A sequence representative is chosen from each cluster based on the criterion set forth by the non-redundant PDB set, and pair-wise alignments of each cluster member to the representative are performed. Atom-based residue–ligand interactions are calculated for each cluster member, and the summation of ligand interactions for all cluster members at each aligned position is determined. Although we were able to successfully align most clusters using a simple dynamic programming algorithm, several cluster created exhibited poor pair-wise alignments of each cluster member to its sequence representative. We hypothesized that the observed alignment problems were, in most cases, due to the incorrect separation and alignment of different domains in multi-domain proteins, a mistake that frequently causes error proliferation in functional annotation. Here we present the results of generating primary sequence patterns for each poorly aligned cluster in PRECISE to assess the extent to which multi-domain proteins that are incorrectly aligned contributes to poor pair-wise alignments of each cluster member to its representative. This requires the use of an iterative locally optimal pair-wise alignment algorithm to build a hierarchical similarity-based sequence pattern for a set of functionally related enzymes. Our results show that poor alignments in PRECISE are caused most frequently by the misalignment of multi-domain proteins, and that the generation of primary sequence patterns for the assignment of sequence family membership yields better alignments for the functionally related enzyme clusters in PRECISE than our original alignment algorithm.

Keywords: Enzyme–ligand interactions; Binding site determination; Enzyme annotation; Pattern-induced multiple sequence alignment; Alignment of multi-domain proteins

1. Introduction

Probably the most frequently occurring association between proteins and their ligands is the binding of substrates by enzymes that catalyze specific biochemical reactions involved in metabolism, signal transduction, and genetic regulation. Nevertheless, most enzyme studies and databases focus on the catalytic properties and thus on the chemistry of enzyme action rather than the recognition of substrates, an important physical step preceding catalysis. For example, the extensive database BRENDA [1] lists reaction specificity, functional parameters, substrates, products, and inhibitors, but does not provide any information on how the various ligands interact with the enzymes. The databases PROCAT [2] and the more recent CSA [3] provide catalytic residue annotation; however, catalytic sites generally only consist of a few, highly conserved amino acid residues, whereas there are usually 10–20 amino acid positions in an enzyme that directly interact with ligands. Unlike the catalytic residues that are highly conserved, residues that participate in ligand binding but do not directly contribute to the catalytic activity can change through evolution, and are responsible for changes in substrate specificity. Thus, information on the entire ligand binding site is required for understanding the energetic contributions to substrate binding, for developing modified

* Corresponding author. Tel.: +1 617 353 4842; fax: +1 617 353 6766.
E-mail address: vajda@bu.edu (S. Vajda).

1093-3263/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2005.08.004

enzymes using methods of protein engineering or directed evolution, and for the design of enzyme inhibitors.

We have recently developed the database PRECISE (Predicted and Consensus Interaction Sites in Enzymes) to provide information on enzyme–ligand interactions [4] by extracting and organizing interaction data for all relevant complexes in the protein data bank (PDB [5]). Interaction sites are defined broadly as any residue(s) that interacts with a ligand, encompassing both catalytic and non-catalytic binding residues. The combination of catalytic and non-catalytic residues in the active site determines substrate specificity. In order to study systematically the residues involved in ligand interactions, enzyme–ligand complexes are collected from the PDB, followed by clustering of the individual enzyme chains in accordance with the functional taxonomy developed by the enzyme commission (EC) of the nomenclature committee of the International Union of Biochemistry and Molecular Biology [6]. Groups are further sub-clustered based on sequence similarity, whereupon a sequence from each cluster is chosen as the representative, and a pair-wise alignment between each member of the cluster and the representative is performed. Following the clustering of all chains within each functional class and performance of pair-wise sequence alignments to the cluster representative, we construct the consensus binding site, i.e., identify all residue positions that contribute to ligand binding in any of the structures, ranked on the basis of the frequency of such interactions found for the residues at each position. The database also includes interactions predicted by solvent mapping, a novel method for determining protein-binding sites based on their structure [7].

In this paper we briefly describe the PRECISE database and the enhancements that have been introduced since its original release in January 2005 [4]. In particular, we focus on the improvements in the alignment of enzyme sequences. As will be described in Section 2, due to the high sequence similarity within the selected clusters of enzymes, in Version 1.0 of PRECISE we have employed a semi-global dynamic programming algorithm to perform pair-wise local alignments of each cluster member to its representative. While this simple approach resulted in good alignments for the majority of functional enzyme clusters in PRECISE, poor alignments were observed in a few cases, indicated by large numbers of gaps and insertions in the sequences, and a surprisingly low apparent sequence identity.

We hypothesized that the observed alignment problems were, in most cases, due to the incorrect separation and alignment of different domains in multi-domain proteins, a mistake that frequently causes error proliferation in functional annotation [8]. Here we present the results of generating primary sequence patterns for each poorly aligned cluster in PRECISE to assess the extent to which the incorrect alignment of multi-domain proteins contributes to poor pair-wise alignments of each cluster member to its representative. Developed by Smith and Smith [9], the creation of primary sequence patterns to illuminate conserved sequence elements prevalent in a set of functionally related proteins has been shown to accurately detect sequence family membership despite low sequence similarity. This requires the use of an iterative locally optimal pair-wise alignment algorithm to build a hierarchical similarity-based sequence pattern. The methodology also allows for the establishment of an information content threshold below which additional sequence patterns will be generated, in effect sub-clustering each family based on the presence or absence of large sequence motifs. The results show that poor alignments in PRECISE are most frequently caused by the misalignment of multi-domain proteins versus the assumption of simple sequence divergence of homologs. Similarity-based sequence patterns have been shown to be a better tool for assessing the extent of structural, and consequently, functional conservation for a set of related proteins. Our results indicate that this approach yields better alignments for the functionally related enzyme clusters in PRECISE than our original alignment algorithm.

2. Materials and methods

2.1. Construction of the PRECISE database

As mentioned in the introduction, PRECISE provides a summary of interactions between the amino acid residues of an enzyme and its various ligands (substrate and transition state analogues, cofactors, inhibitors, and products). In the current version this information is extracted primarily from the enzyme–ligand complexes in the PDB by performing a number of steps as follows.

1. *Clustering homologous enzyme chains*. Although enzymes with the same EC number have the same function, they may substantially differ in terms of sequence and/or structure. We consider chains rather than the entire protein, since enzymes may have chains of different functions and/or non-homologous chains. Since consensus-binding sites can be defined only for enzyme chains with appropriate overlap of their sequences and structures, we have clustered the enzyme chains in the PDB such that in each cluster the proteins have the same EC number, and all chains are sequence-similar. The clustering is based on the non-redundant PDB chain set (http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html), also known as the nrPDB database, maintained by NCBI. The clusters in nrPDB have been constructed by comparing all chains available from the PDB with each other using the BLAST algorithm [10]. The chains are then clustered into groups of sequence-similar entries using a single-linkage clustering procedure. We have used the BLAST p-value of 10^{-40}.
2. *Selecting a representative for each homologous cluster*. The chains within a sequence-similar cluster thus derived are ranked according to the precision and completeness of their structural data. The measures of the structural quality are adopted from the NCBI's non-redundant PDB chain set (nrPDB), and are as follows in the order of priority: (a) lower percentage of residues with unknown amino acid type; (b) lower percentage of residues with incomplete coordinate

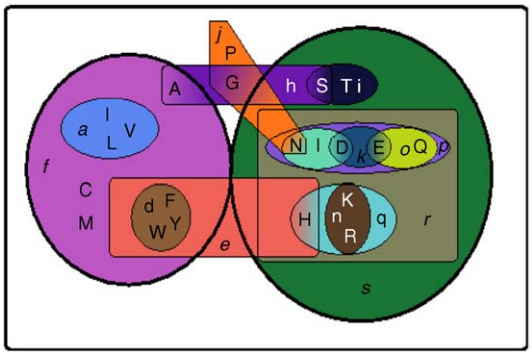
data; (c) lower percentage of residues whose coordinate data are missing; (d) lower percentage of residues with incomplete side-chain coordinate data; (e) higher resolution; (f) larger number of chains (subunits) contained in the PDB entry; (g) larger number of heterogens contained in the PDB entry; (h) larger number of different types of heterogens; (i) larger number of residues; and (j) alphanumerical order of their PDB codes. The top-ranked chain is generally chosen as the representative of the group. In some cases, however, a lower-ranked chain was chosen. For example, if the top-ranked chain was a mutant protein and there was a native protein with reasonably comparable structural quality, then that lower-ranked native protein might replace the mutant as the representative.

3. *Pair-wise alignment of each cluster member to its representative*. Once a sequence representative is chosen for each functional cluster in PRECISE, a pair-wise alignment of each cluster member to its representative is performed. A semi-global dynamic programming algorithm was implemented using the GONNET similarity matrix [11], a gap-opening penalty of 8, and a gap extension penalty of 0.25. Each pair-wise alignment in the cluster is then stacked upon each other, and the total number of interactions for all members at each aligned position is determined.
4. *Selecting ligand type*. For all PDB files containing a ligand(s), the ligand(s) is classified as (a) peptide, (b) nucleotide, (c) cofactor, (d) metal ion, (e) inorganic ion, or (f) ''other''. Most of the ligands of interest (i.e., substrate and transition state analogues, inhibitors, and products) are in the last category, and we plan to provide further classification that will provide detailed information on the particular role of the molecule. However, this type of information is not uniformly contained in databases and hence requires manual curation.
5. *Calculation of hydrogen bonds and non-bonded interactions*. The receptor–ligand interactions for each member of a cluster are determined using the program HBPLUS by Thornton and coworkers (http://www.biochem.ucl.ac.uk/bsm/hbplus/home.html). The results are later parsed and extracted into residue-based formats, giving separate files for each residue that is interacting with a ligand.
6. *Summing all interactions within a cluster for the aligned residues*. The number of ''hits'', or interactions, for a given residue is calculated automatically by summing all of the interactions for each aligned residue of the cluster members. Again, the hits are atom based; two interacting residues can have a hit greater than 1 if several atoms are involved in the interaction.
7. *Final output*. If there are non-homologous chains present in the query PDB file, a subsequent page will let users to specify which chain to display. The output page will show the sequence of the representative of the cluster along with different color codes for each residue representing the number of hits. The blue to red color-scheme indicates the residues that belong to the binding site, as well as the total number of interactions found at each amino acid position in all chains of the cluster. This is the most important information provided by the page. Clicking on any ''colored'' residue (i.e., on a residue that has at least one interaction) displays a separate panel with a detailed list of interactions for the selected residue. For each interaction, the list shows the PDB code and chain identifier of the protein; the name, heteroatom code, and type of the particular ligand; the interacting residue and atom in the protein, and the type of the interaction (non-bonded or hydrogen bonds). It is important that the list shows both the ''interaction position'', i.e., the original sequence number of the interacting residue in the PDB file, and the ''aligned position'', which is the sequence number of the same residue in the alignment of sequences for the entire cluster. There are two additional options available from this page. The first is visualization of the interactions for any of the ligands in the particular PDB file. Clicking on the selected interactions opens a separate page that shows the ligand and the nearby amino acid residues using the Java-based molecular graphics program JMOL. The second option is a link to the appropriate entry in the PDBSum database, providing summary information both on the protein and the ligands. On the main output page, an additional button is provided to open a new window with the alignment. On the right of the sequence, separate panels show all PDB codes and chain identifiers of the entries that form the cluster. The user may select any subset of these entries and recalculate the list of interactions. Additional panels permit the users to restrict the set of interactions to selected interaction types (i.e., non-bonded or hydrogen bond) and to selected ligand types (i.e., peptides, nucleotides, cofactors, metal ions, other inorganic ions, or ''others''). Again, any subset of these can be selected to produce the list of interactions.

The online version of PRECISE is available at http://precise.bu.edu/. The current version (Version 1.4, June 2005) contains 25,373 enzyme chains, extracted from 13,123 structure files of enzymes in the PDB. Enzyme chains belong to 1258 unique EC numbers. Clustering chains with the same EC number and pair-wise BLAST p-value of 10^{-40} or less yields 2434 clusters, i.e., on the average, 1.93 clusters per enzyme number, with 10.47 chains in each cluster. The average percent identity within a cluster is about 98%.

2.2. Generation of primary sequence patterns for poorly aligned clusters in PRECISE

Several clusters in PRECISE Version 1.0 demonstrated overall poor pair-wise alignment quality of cluster members to their representatives. To assess the extent to which this could be attributed to either the sequence divergence of family members or the incorrect alignment of protein domains, we created primary sequence patterns [9] for each family in PRECISE where the average sequence identity of the aligned positions fell below 50%. Studies have shown that while overall sequence similarity can be minimal among a set of functional homologs, 'core domains' typically maintain a disproportionate amount of sequence similarity [12]. Thus, the



Class	Members	Class	Members
a	ILV	n	KR
d	FWY	o	EQ
e	FWYH	o	DENQ
f	AILVMFWYC	q	HKR
h	AGS	r	DENQKHR
i	ST	s	DENQHKRST
j	GNP	X	ACDEFGHIKLMNP QRSTVWY
k	DE	g	gap
l	DN	-	gap

Fig. 1. Hierarchical similarity schema for the generation of primary sequence patterns. A sequence pattern consists of a single line of symbols generated by classifying aligned amino acid positions according to this overlapping schema.

generation of covering sequence patterns that are capable of delineating conserved sequence motifs for protein families of low sequence similarity is a valuable tool for the assignment of sequence family membership; this is achieved through the use of a hierarchical similarity-based system for classification of each aligned position, allowing for a more comprehensive analysis of similarity than normal alignment algorithms are capable of performing. Furthermore, since an information content threshold is employed in the creation of each sequence pattern, allowing for the creation of multiple patterns for a set of related sequences, this method can be used to distinguish functional homologs that may have inserted or deleted domains with respect to each other as multiple patterns may be generated to represent each domain present in the set. To utilize this method, we performed locally optimal pair-wise alignments based on the Smith–Waterman algorithm [13] with a modified gap-weighting scheme [9] for each set of sequences. The sequences are then clustered according to their pair-wise similarity scores, resulting in a binary dendogram. As two sequence nodes are joined, a pattern is created to represent the similarity at each position of the pair-wise alignment; if the amino acid is conserved, that amino acid is entered into the pattern, otherwise if the two amino acids differ, the minimally inclusive class of amino acid is determined, and a symbol representing that class is inserted into the pattern at that position. A schema for hierarchical system of sequence similarity developed is shown in Fig. 1 with the symbols used to represent each class. Gapped positions in the alignments automatically result in the placement of a gap character in the pattern at the aligned position. The tree is reduced by replacing node pairs with increasingly complex patterns that account for the joining of multiple sequences, eventually resulting in a single encompassing pattern determined for the cluster. Multiple patterns can be determined for one set of sequences by placing a minimal threshold on the relative information content of the resulting pattern. The patterns generated for each cluster were used to assess the necessity of further sub-clustering of sequences of multi-domain enzymes in PRECISE. Additionally, the resulting iterative pair-wise alignments were used in replacement of previously generated pair-wise alignments, requiring new calculations of interaction frequencies at each aligned position for incorporation into PRECISE.

3. Results

3.1. Generation of primary sequence patterns for poorly aligned clusters in PRECISE

We examined the 67 clusters in PRECISE that exhibited an average pair-wise sequence identity of less than 50%, with a minimum value of 22.4% and a mean of 41%. The majority of these alignments also contained large gapped regions, as demonstrated by the histogram in Fig. 2 depicting the distribution of the percentages of gapped aligned positions for each sequence in the cluster. A maximal gapped percentage of 72.4% was observed in a kinase cluster consisting of 184 sequences. Table 1 shows the top six most poorly aligned clusters, consisting of known multi-domain enzymes such as kinases, phosphodiesterases, and phosphatases. For each of the 67 clusters in PRECISE of poor alignment quality, we generated primary sequence patterns using the Smith–Smith algorithm and the hierarchical similarity classes shown in Fig. 1.

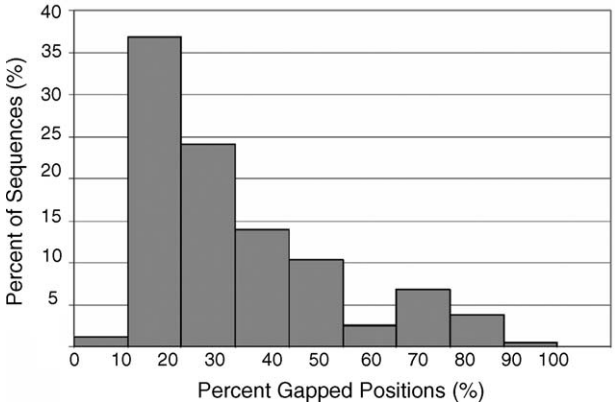


Fig. 2. Distribution of the percentage of gapped alignment positions for the clusters in PRECISE exhibiting a sequence identity of less than 50%. Sixty-seven clusters in PRECISE demonstrated an average sequence identity of each cluster members to its representative of less than 50%. This histogram depicts the percentage of gapped alignment positions for each cluster member.

Table 1
Poorest cluster alignments in PRECISE

EC number	Function	PDB representative, chain	Number of sequences	Percent identity	Percent gapped positions
2.7.1.112	Tyrosine kinase	2abl	184	23.9	45.2
3.2.1.8	Xylanase	1gny, A	70	22.42	48.2
1.1.1.205	IMP dehydrogenase	1b3o, B	27	48.1	34
2.7.1.37	Kinase	1jkk, A	51	37.1	26.2
3.1.4.17	Phosphodiesterase	1rkp, A	92	36.6	32.6
3.3.1.48	Tyrosine phosphatase	1aya, A	91	23.9	49.4

3.2. Improved alignment for functionally related enzymes

Creation of multiple sequence patterns for the 67 poorly aligned clusters of functionally related enzymes in PRECISE resulted in greatly improved alignment quality of each cluster and yielded interesting implications for the analysis of multi-domain proteins. We determined the percent sequence similarity of aligned positions in these clusters based on the frequency of either an amino acid or a similarity character at each position of the sequence patterns generated for each cluster; Fig. 3 demonstrates the distribution of percent similarity for each cluster. The average percent sequence similarity for the 67 clusters analyzed was 76.9%, with a maximal value of 100% and a minimal value of 36%. As a direct comparison to the previous metric used for pair-wise alignment of each cluster member to its cluster representative, we determined the distribution of gapped positions in each sequence of the iterative pair-wise alignments generated; here we observe a striking improvement as the gap percentage is improved from 41% of each alignment on average to an average of 1.45%. Results of the analysis of gapped alignment positions are summarized by the distribution in Fig. 4. A synopsis of the results obtained for the six clusters exhibiting the poorest pair-wise alignments previously, defined in Table 1, is provided in Table 2; in each case the quality of the alignment is improved significantly by the generation of sequence patterns, as indicated by the marked increase in percent similarity and the large decrease in percent gapped positions. Interestingly, the greatest improvements were made to the kinase and xylanase families, where in both instances multiple patterns were created to account for the multiple functional domains found in these families, discussed further in Section 3.3. Overall, results of creating primary sequence patterns for poorly aligned clusters in PRECISE indicates that, for our purposes, this methodology can be used to successfully align functionally related proteins that may exhibit minimal global sequence similarity but more highly conserved sequence motifs in domains that serve common functions.

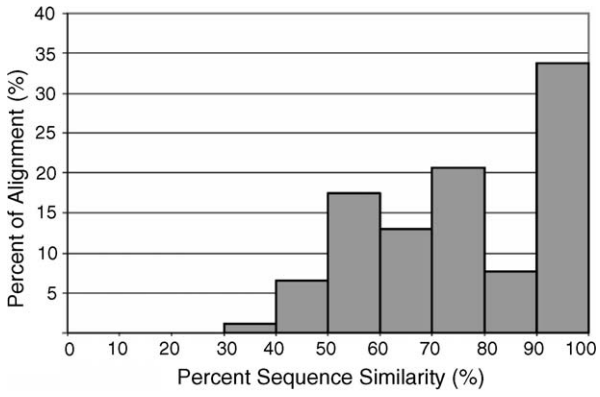


Fig. 3. Distribution of percent sequence similarity of primary sequence patterns generated for the poorly aligned clusters in PRECISE. The percent sequence similarity of each pattern was calculated to be the percent of positions in the pattern for which a gap or 'X' symbol was not present (see Fig. 1). The average sequence similarity for all patterns generated was 76.9%.

3.3. PRECISE clusters with multiple patterns generated

Multiple primary sequence patterns are created when large enough sequence divergence is present between two sequences such that a sequence similarity pattern of high information content cannot be determined. The presence of multiple sequence patterns serves as a good indication of the necessity for sub-clustering functional families based on sequence divergence. Of the 67 clusters in PRECISE for which primary sequence patterns were created, four clusters exhibited large enough sequence divergence to result in multiple patterns generated for these clusters, shown in Table 3. Each of these enzyme families consists of multiple domains that may or may not be catalytic (e.g., protein kinases consisting of SH2 and

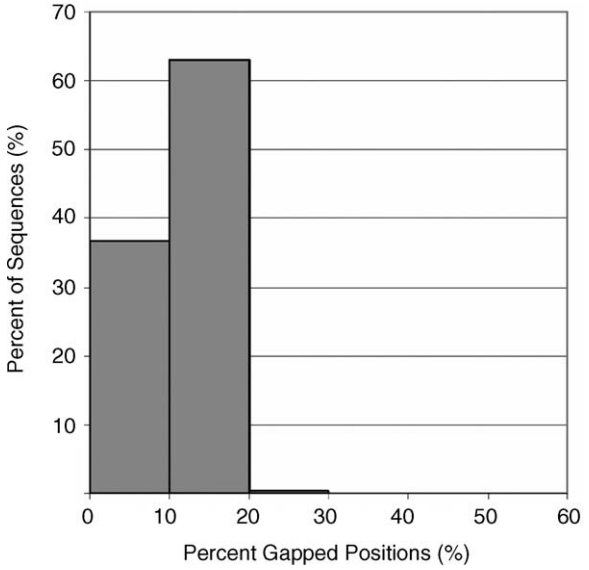


Fig. 4. Distribution of gapped alignment positions for the iterative pair-wise alignments determined by the generation of primary sequence patterns. Since no gap percentage exceeds 40%, the histogram is truncated at 60% gapped percentage.

Table 2
Results of sequence patterns generated for poorest PRECISE clusters

EC number	Function	PDB representative, chain	Number of sequences	Percent similarity	Percent gapped positions
2.7.1.112	Tyrosine kinase	1lck, A; 1jeg, A; 1opl, A[a]	184	58.8, 66.7, 36.1	7.4, 0, 12.0
3.2.1.8	Xylanase	1knl, A; 1us2, A[a]	70	100.0, 43.3	0, 10.0
1.1.1.205	IMP dehydrogenase	1jcn, A	27	58.1	3.9
2.7.1.37	Kinase	1j3h, A	51	42.2	6.9
3.1.4.17	Phosphodiesterase	1xlx, A	92	55.1	7.6
3.3.1.48	Tyrosine phosphatase	1aya, A; 2shp, A[a]	91	100.0, 36.2	0, 15.4

[a] Multiple sequence patterns were generated.

SH3 binding domains as well as a kinase catalytic domain). Because we are interested in the study of common structural interactions within a set of functionally related enzymes, it is necessary to identify and sub-cluster each domain of these proteins if an alignment algorithm is incapable of aligning each domain correctly. The pattern generated for the catalytic domain of protein kinases is shown in Fig. 5a. Similar regions are color-coded to indicate the type of similarity exhibited at each aligned position. This pattern demonstrates the strength of the hierarchical method for defining sequence similarity in accurately aligning protein domains that may only contain a few positions of absolute amino acid conservation but share a relatively higher level of sequence similarity.

For the purposes set forth in the creation of PRECISE, namely the study of substrate interactions common to a set of functionally related enzymes, it is essential to identity functional clusters that contain multi-domain proteins for the further analysis of alignment integrity. However, it may not be necessary to separate multi-domain proteins into multiple sequence sub-clusters in order to prevent the misalignment of functional domains if sufficient sequence similarity is demonstrated amongst family members to ensure the quality of an alignment. An example of this can be seen in the phosphodiesterase (PDE) family, consisting of PDEs 3–5, where a single sequence pattern of high information content was determined for a set of 92 sequences. Previous publications have asserted that while PDEs can exhibit large sequence divergence in looped regions, the catalytic core is maintained. The sequence pattern generated for the PDE family as shown in Fig. 5b demonstrates the agreement of our results of using a pattern-based approach to sequence family assignment with current literature where other approaches were utilized, such as phylogenetic analysis, to create sequence families.

Table 3
Enzyme clusters for which multiple sequence patterns were generated

EC number	Function	Number of sub-clusters created	Domains sub-clustered
2.7.1.112	Tyrosine kinase	3	SH2, SH3, kinase
3.2.1.8	Xylanase	2	Binding, catalytic
3.3.1.48	Tyrosine phosphatase	2	SH2, catalytic
2.7.7.7	DNA-directed polymerase	2	N-terminal, catalytic

4. Discussion

We greatly improved the quality of the alignment data currently available in PRECISE by creating primary sequence patterns for clusters of functionally related enzymes for which simple pair-wise alignments is not adequate for the alignment of homologs. This approach led us to the identification and correction of misalignments of functionally related enzymes that either exhibit low global sequence similarity or contain multiple domains that may not be present in all family members. Since the majority of previously poorly aligned clusters were corrected simply by the use of an iterative pair-wise alignment that takes into account more complex sequence similarity patterns, we concluded from this study that a pattern-based approach improves the alignments of families where the sequence similarity as defined by a simple metric is not high amongst family members. Based on our initial results, we plan to use PIMA for the generation of pattern representations and multiple sequence alignments for each functional family in PRECISE, even those where the previously determined alignment from the nrPDB was sufficient to achieve a high percent sequence identity of each cluster to its representative. In particular we will focus on the families existing in PRECISE where multiple sequence sub-clusters currently exist to determine if sequence similarity may be sufficient to define a single or reduced number of patterns for these families. We plan to perform comparisons of sequence patterns amongst functional families to yield greater insight into the relationship between structure and function, as well as the elucidation of substrate specificity within a functional class. Moreover, ligand interaction patterns common to a set of functionally related enzymes uncovered via these studies may elucidate binding properties important for rational drug design efforts.

In addition to the improvement of alignment quality of functional families in PRECISE, a number of changes to the database have been made both to improve the visualization of interactions and to collect the most recent structures available in the PDB. PRECISE now includes an automatic update process to account for additions and deletions of enzyme structures in the PDB. To reduce time and computation complexity, updates only affect those clusters with new or removed chains in the PDB. This is a substantial improvement over our previous method requiring the recreation of the entire database for each monthly update of structures from the PDB.

Fig. 5. Primary sequence patterns generated for the catalytic domain of kinase family members and members of the phosphodiesterase (PDE) family. (a) Three sequence patterns were generated for a cluster consisting of 192 kinases, accounting for the SH2, SH3, and catalytic domains. Conserved positions are color-coded based on the type of similarity exhibited at each aligned position in the pattern, where red indicates absolute amino acid conservation. (b) The alignment of the PDE family illustrates the high level of sequence similarity maintained across PDEs 3–5.

Visualization improvements have also been made to PRECISE. We have added a rotatable 3D representation of the binding site using the Java-based JMOL molecular viewer. JMOL allows the user to view a specific ligand and a predefined set of atoms from the selected chain. The atom set selected for initial display includes all atoms of the ligand(s) chosen and the side chain atoms of all amino acids that interact with any of the ligands for the given cluster.

We have also developed a high-throughput pipeline for the inclusion of interactions predicted by computational small molecule mapping [7]. PRECISE currently contains the results from mapping over 50 different enzyme structures. The high-throughput pipeline will continue to add predicted small molecule interactions until at least one member of every cluster in PRECISE has been mapped. These predicted interactions are stored separately within the database and can be filtered independently from the observed interactions. This enables the comparison of our predicted interactions and the observed interactions by using the available interface.

Acknowledgements

We would like to thank Temple Smith and Sean Quinlan for their assistance in generation sequence patterns and very useful discussions concerning the alignment of multi-domain proteins. This work was funded by grant no. DBI 0213832 from the National Science Foundation and GM64700 from the National Institute of Health.

References

[1] I. Schomburg, A. Chang, D. Schomburg, BRENDA, enzyme data and metabolic information, Nucl. Acids Res. 30 (1) (2002) 47–49.
[2] A.C. Wallace, R.A. Laskowski, J.M. Thornton, Derivation of 3D coordinate templates for searching structural databases: application to the Ser-His-Asp catalytic triads of the serine proteases and lipases, Protein Sci. 5 (1996) 1001–1013.
[3] C.T. Porter, G.J. Bartlett, J.M. Thornton, The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data, Nucl. Acids Res. 32 (Database issue) (2004) D129–D133.
[4] S.H. Sheu, et al. PRECISE: a database of predicted and consensus interaction sites in enzymes, Nucl. Acids Res. 33 (Database issue) (2005) D206–D211.
[5] J. Westbrook, et al. The protein data bank: unifying the archive, Nucl. Acids Res. 30 (1) (2002) 245–248.
[6] IUBMB, Enyzme Nomenclature: Recommendations of the Nomenclature Committee of the International Unit of Biochemistry and Molecular Biology, Academic Press, San Diego, CA, 1992.
[7] S. Dennis, T. Kortvelyesi, S. Vajda, Computational mapping identifies the binding sites of organic solvents on proteins, Proc. Natl. Acad. Sci. U.S.A. 99 (7) (2002) 4290–4295.
[8] T.F. Smith, X. Zhang, The challenges of genome sequence annotation or ''the devil is in the details'', Nat. Biotechnol. 15 (12) (1997) 1222–1223.
[9] R.F. Smith, T.F. Smith, Automatic generation of primary sequence patterns from sets of related protein sequences, Proc. Natl. Acad. Sci. U.S.A. 87 (1) (1990) 118–122.
[10] S.F. Altschul, et al. Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.
[11] G.H. Gonnet, M.A. Cohen, S.A. Benner, Exhaustive matching of the entire protein sequence database, Science 256 (5062) (1992) 1443–1445.
[12] J. Greer, Proteins 7 (1990) 317–334.
[13] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1) (1981) 195–197.