July 2006

Comparison of Protein Active Site Structures for Functional Annotation of Proteins and Drug Design

Robert Powers
University of Nebraska - Lincoln, rpowers3@unl.edu

Jennifer C. Copeland
University of Nebraska - Lincoln

Katherine Germer
University of Nebraska - Lincoln

Kelly A. Mercier
University of Nebraska - Lincoln

Viswanathan Ramanathan
University of Nebraska - Lincoln

See next page for additional authors

Published in *PROTEINS : Structure , Function and Bioinformatics* 65:1 (2006), pp. 124–135. Submitted October 2005; revised April 2006; accepted April 2006; published online 21 July 2006 in Wiley InterScience (www. interscience.wiley.com). DOI: 10.1002/prot.21092

Comparison of Protein Active Site Structures for Functional Annotation of Proteins and Drug Design

Robert Powers, Jennifer C. Copeland, Katherine Germer, Kelly A. Mercier

Department of Chemistry, University of Nebraska–Lincoln

Viswanathan Ramanathan and Peter Revesz

Department of Computer Science and Engineering, University of Nebraska-Lincoln

Abstract: Rapid and accurate functional assignment of novel proteins is increasing in importance, given the completion of numerous genome sequencing projects and the vastly expanding list of unannotated proteins. Traditionally, global primary-sequence and structure comparisons have been used to determine putative function. These approaches, however, do not emphasize similarities in active site configurations that are fundamental to a protein's activity and highly conserved relative to the global and more variable structural features. The Comparison of Protein Active Site Structures (CPASS) database and software enable the comparison of experimentally identified ligand-binding sites to infer biological function and aid in drug discovery. The CPASS database comprises the ligand-defined active sites identified in the protein data bank, where the CPASS program compares these ligand-defined active sites to determine sequence and structural similarity without maintaining sequence connectivity. CPASS will compare any set of ligand-defined protein active sites, irrespective of the identity of the bound ligand.

Key words: functional annotation, CPASS, hypothetical proteins, ligand-defined active sites

Grant sponsor: Protein Structure Initiative of the National Institutes of Health; Grant number: P50 GM62413; Grant sponsors: Nebraska Tobacco Settlement Biomedical Research Development Funds and NASA Nebraska Space Grant and EPSCoR; Grant sponsor: NIH; Grant number: RR015468-01.

Obtaining the biological function of a protein is essential for determining its potential as a therapeutic target and its utility as part of structure-based drug design effort. Furthermore, understanding the biological function for a protein provides the basis for exploring its cellular activity. An outcome of various genomics efforts has been a vast growth in putative protein sequences that lack any experimental functional annotation.[1,2] Sequence homology has routinely been used as a rapid approach to assign biological function to these hypothetical proteins or proteins of unknown function.[3] This is based on the accepted structural biology paradigm that a similarity in sequence ($\geq$30%) implies a corresponding similarity in both structure and function. At best, sequence homology provides functional assignment for ~50% of the proteins identified in various proteomes.[2,4–6] Structural genomics is augmenting the functional assignment of these hypothetical proteins by determining the corresponding three-dimensional structure.[7] This permits a functional assignment by identifying proteins of known function that exhibit a similar overall fold to the hypothetical protein. Structural homology is a more sensitive approach for assigning function, since there are numerous examples of proteins with similar folds that lack any significant sequence homology.[8,9] This is consistent with the general observation that tertiary structures are significantly more evolutionary stable than protein sequences.[10] Nevertheless, our analyses of the scientific literature for protein structures of hypothetical proteins that are emerging from structural genomics indicate that ~60% of the reported structures correspond to a novel fold or folds that can not be readily assigned to a biological function as determined by the authors.

Sequence and structural homology methods primarily determine ''global'' similarities between the compared proteins.[7] However, the molecular function of a protein is generally restricted to its identified active site, which may involve an interaction with small molecular-weight ligands, nucleic acids, or other proteins. Maintaining the core structural component of the active site is essential for preserving the functional activity of the protein. As a result, protein comparisons that focus on global sequence and structural similarities may miss proteins with conserved active sites but divergent sequences and structures. Thus, a more effective means to infer a biological function of a hypothetical protein would occur through the identification of the protein's active site.

Comparative analysis of protein active sites is also critical for a successful drug discovery program, particularly for elim-

inating potential toxicity pathways. Drug toxicity is a common cause of failure during clinical trials, where undesirable protein–ligand interactions are a plausible mechanism.[11] Efforts to eliminate potential toxicity problems are initially carried-out by screening for drug selectivity against a limited panel, for practical reasons, of very closely related proteins.[12] These protein panels are usually composed of functionally identical proteins with high sequence and structural similarity that are identified by traditional homology methods. Inevitably, this approach will miss proteins that only exhibit similarity in the structural characteristics of the active site. This is particularly problematic for common ligand binding sites, such as ATP, that are drug discovery targets and are present in functionally diverse proteins.[13]

A number of methodologies are being developed to predict the location of active sites in novel protein structures. This is typically accomplished by developing structural descriptors of active sites for defined protein functional classes and then fitting these structural templates to novel folds to identify putative active sites and annotate the hypothetical proteins. A variety of approaches are being applied that include aligning structures to match a few consensus or enzymatic catalytic residues,[14–23] identification of cavities consistent with shapes of known ligands,[24] a sequence independent force field to extract common active site features,[25] theoretical prediction of titration curves,[26] using chemical properties and electrostatic potentials of amino acid residues consistent with active site characteristics,[27,28] neural network analysis of spatial clustering of residues,[29] and conserved residues from multiple sequence alignments (phylogenetic motifs).[20,30]

Nevertheless, direct experimental observation of protein–ligand interactions are a more reliable mechanism for the proper and accurate identification of protein active sites. LigBase is an online database that aligns only active sites present in the protein data bank (PDB) that bind the *identical* ligand, using structure and sequence alignments.[31] Similarly, there are numerous databases that allow searching of the PDB for compounds present in protein–ligand complexes.[32–35] Unfortunately, these databases lack the ability to globally compare an active site identified for a novel protein against the entire structural database, irrespective of the identity of the bound ligand, to determine the relative similarity in the sequence and structure of the active sites.

Towards this end, we have implemented a database and a suite of programs to compare experimentally identified protein active sites to infer biological function (Fig. 1). In this article, we describe the design and application of the Comparison of Protein Active Site Structures (CPASS) database and software that enables both the sequence and structural comparison of ligand-defined active sites to infer functional activity of hypothetical proteins and to aid in the design of drug selectivity.

MATERIALS AND METHODS

Design Philosophy

The main feature that the CPASS program is trying to capture is the similarity in the characteristics of the active site defined by the positions and types of amino acids relative to a

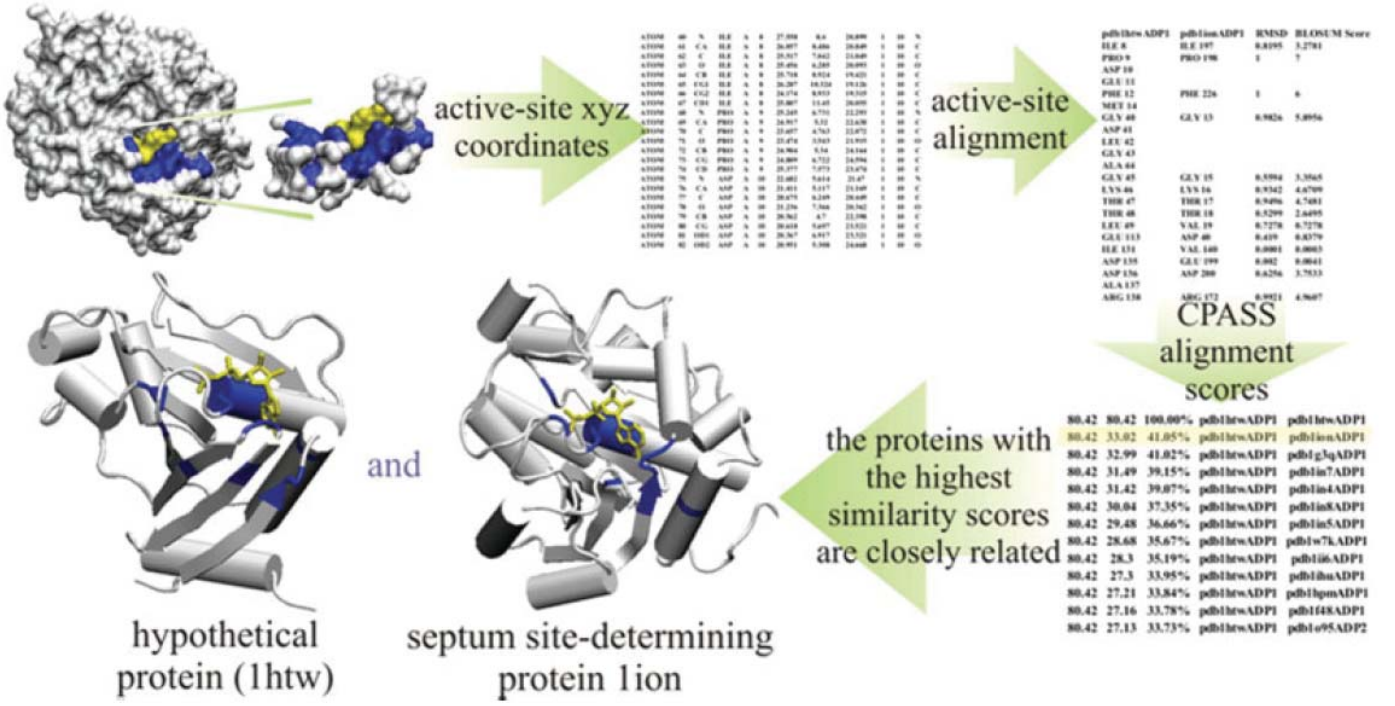

Fig. 1. Schematic diagram of the application of the CPASS database and software to aid in the assignment of biological function to hypothetical or novel proteins. The bound ligand is colored yellow and the active site residues are colored blue. All molecular images were created using VMDXPLOR.[36]

bound ligand. Unlike other approaches, CPASS does not reduce the database to a limited collec tion of consensus templates for each functional family. Similarly, CPASS does not attempt to simulate generic features of active sites by using descriptors mimicking important properties of amino acids. Instead, the CPASS database is composed of ligand-defined protein active site structures culled from the protein data bank (PDB).

A total of ~34,000 X-ray and NMR structures that are currently available in the PDB were analyzed for the presence of a bound ligand. The CPASS database is expected to be routinely updated. Only protein structures that contain a bound ligand are included in the CPASS database. Conversely, structures that do not contain a protein molecule, but only contain a DNA or RNA molecule complexed to a ligand were excluded, since they lack any value in the functional annotation of a protein. The identification of a ligand within a protein PDB file was determined by the presence of either a HET or HETNAM record. Routinely, a single protein PDB file may contain multiple ligands. Each ligand was extracted separately with a uniquely defined active site in the absence of a LINK record in the PDB file. The LINK record identifies bonded atoms from two residue types. If a protein PDB file contains multiple ligands with a LINK record that connects these ligands, then all the ligands are extracted as a single ligand with a single corresponding active site. As an example, consider a PDB file that contains both ATP and an Mg^{2+} ion. In the absence of a LINK record that connects the phosphate group from ATP to the Mg^{2+} ion, two separate ligand coordinate files are extracted—one for ATP and the other for the Mg^{2+} ion. The two ligand coordinate files are then used to identify two separate active sites around ATP and the Mg^{2+} ion, respectively. Conversely, if a LINK record was present in the protein PDB file that indicates a bond between ATP and the Mg^{2+} ion, a single ligand file is extracted from the protein PDB file that contains the coordinates for both ATP and Mg^{2+} ion. This single file that contains both ligands will then be used to determine a single ligand-based active site.

Besides the presence of small molecular-weight ligands defined by the HET and HETNAM records, a number of protein PDB structures contain small peptides, DNA, or RNA sequences complexed to the protein. The CPASS database also includes these small peptides, DNA, and RNA sequences (13 residues) with the corresponding active site defined by these ligands. The presence of a peptide or small nucleic acid chain in the protein PDB file is identified by the SEQRES record, where the total number of residues for a particular chain is ≤13 and a second protein chain is defined with >13 residues.

Currently, ~42,000 protein–ligand binding sites have been identified in the PDB. This list excludes common and abundant buffer reagents, salts, and solvents that generally exhibit nonspecific binding irrelevant to functional activity. A total of 112 ligands are currently excluded from the CPASS database, where the vast majority are common ions (Na^+,Cl^-,SO_4^-), solvents (water, MES, DMSO, 2-mercaptanol, glycerol), and chemical fragments or clusters (acetyl, methyl) (see Supplemental Table 1). Practical considerations required removing these ligands because of the significant increase in the total number of ligand-defined binding-sites in the CPASS database, the negative impact on the CPASS computational time, and the minimal benefit to functional identification. As an example, the isolated calcium ion (PDB Het ID: CA) is present in 2887 structures in the PDB, which results in a total of 7811 binding sites, which by itself is 30% the size of the entire CPASS library. While it would be beneficial to include the functionally relevant calcium binding sites in CPASS, it is not feasible to differentiate between these sites and the numerous irrelevant calcium binding sites present in the various structures. Again, simply including all the calcium binding sites is currently impractical, especially when the 7811 binding sites are combined with other similarly excluded ligands. Additionally, numerous X-ray structures contain redundant copies of essentially identical protein–ligand structures based on the number of structures found within the unit cell. Multiple binding sites within the same structure are identified and only one copy is maintained if the ligand-defined active sites share ≥80% sequence identity and bind the same ligand. Thus, the list may be reduced to ~26,000 ligand-defined binding sites, when these multiple copies from the same PDB coordinate file are eliminated.

The ligands identified from protein–ligand complexes in the PDB are then used to determine ligand-defined active sites within the protein structure. The amino acid residues that comprise an active site are identified by having at least one atom that is ≤6Å from any ligand atom. Thus, the ligand chemical structure and bound conformation determines which amino acids within the protein comprise the active site. Relative changes in the ligand conformation may result in a corresponding change in the composition of the ligand-defined active site. The impact on the active site definition depends on the magnitude of the conformational change and whether this change results in either the complete loss or gain of an interaction with a specific amino acid. In general, ligand conformational changes have minimal impact on the definition of the residues that describe the active site, where residues on the 6-Å peripheral are the most likely to change.

The CPASS active site definition contains the residue types, the corresponding Cα coordinate positions, and the shortest distance from any atom in the residue to any atom in the ligand (d_i). The same active site information is then obtained from a protein–ligand complex for a targeted hypothetical or novel protein from experimental sources. Sequence and structural similarities of ligand-defined active sites for hypothetical or novel proteins are then compared against the entire PDB derived ligand-defined active sites in the CPASS database. Any differences in ligand conformations between the compared active sites will have a minimal impact on the calculated similarity, because the sequence and structure of only the active sites are compared. The ligand structure is not included in the comparison. Again, a ligand conformational change may simply result in the addition or exclusion of amino acid(s) residue in the active site definition. Thus, two similar active site sites would not be missed because of ligand conformation changes, since the remainder of the residues present in each active site would still exhibit the expected similarity in sequence and structure.

Similarity Function

There are two uniquely critical features of the CPASS analysis of ligand-defined protein active sites to identify similarity in structure and function. First, the CPASS analysis is independent of the identity of the bound ligand. Although CPASS allows for the comparison of active sites that contain the same ligand, it is not necessary. The structure of the ligand is not used in the comparison, since it would eliminate any meaning in aligning active sites with distinct but related ligands. In this manner, the ligand-defined active site obtained for the target protein can be compared against the entire CPASS database (~26,000 ligand-defined active sites) or any subset of the database to obtain a meaningful alignment score. Second, the sequence and structure alignment of the ligand-defined database is not dependent on the primary sequence connectivity of the protein. In traditional global sequence or structure homology, the primary sequence connectivity is a fundamental component of the alignment analysis, where the insertion of gaps or deleted regions between the aligned sequences or structures results in a scoring penalty.[37,38] Since the structural organization of a protein active site typically comprises distal sequence regions of the protein coming into close contact as a result of the three-dimensional fold, the primary sequence connectivity is not directly relevant to the sequence and structural alignment of an active site.

Thus, the CPASS program determines the optimal sequence and structural alignment between two compared active sites without maintaining sequence connectivity. The CPASS program determines the alignment of two active sites by maximizing a root-mean-square-difference (rmsd) weighted BLOSUM62[39,40] scoring function (S_{ab}):

$$S_{\mathrm{ab}} = \sum_{i,j=1}^{i=n,j=m} \frac{d_{\min}}{d_i} (e^{-\Delta \mathrm{rmsd}_{i,j}})^2 p_{i,j}$$

$$\Delta \mathrm{rmsd}_{i,j} = \begin{cases} \mathrm{rmsd}_{i,j} - 1 & \mathrm{rmsd}_{i,j} > 1\ \mathrm{A} \\ 0 & \mathrm{rmsd}_{i,j} \leq 1\ \mathrm{A} \end{cases} \quad (1)$$

where active site *a* contains *n* residues and is compared with active site *b* from the CPASS database, which contains *m* residues, $p_{i,j}$ is the BLOSUM62 probability for amino acid replacement for residue *i* from active site *a* with residue *j* from active site *b*, $\Delta\mathrm{rmsd}_{i,j}$ is a corrected root-mean-square-difference in the Cα coordinate positions between residues *i* and *j*, and $d_{\min}/d_i$ is the ratio of the shortest distance to the ligand among all amino acids in the active site, compared with the current amino acid's shortest distance to the ligand. S_{ab} is only summed over the optimal alignment for residue *i* from active site *a* with residue *j* from active site *b*. It is not summed over all possible combinations of *i* and *j*. If the number of residues are not identical between active sites *a* and *b* ($n \neq m$), then the additional residues will not have a corresponding match. Each residue can be used only once in the alignment. If active site *a* contains unmatched residues, then no contribution is made to S_{ab}, which effectively reduces the maximal possible score that can be achieved for active site *a*. As an example, if active site *a* contains an unmatched Ala, a score of 0 is added instead of a possible maximum score of 4 if active site *b* contained an appropriately aligned Ala. The active sites that are being compared are typically in distinct coordinate axes, and so aligning the coordinates in an optimal arrangement without the use of the primary sequence connectivity requires an iterative approximation guided by maximizing this scoring function.

The BLOSUM62 probability matrix was chosen based on the reported evaluation of a number of matrixes, where BLOSUM62 was identified as the best matrix.[40] BLOSUM62 is also widely used to construct sequence alignments and is the default matrix for BLAST.[41]

The calculated rmsd between residues *i* and *j* is corrected by 1 Å ($\Delta\mathrm{rmsd}_{i,j}$) to account for structural variations less than 1 Å that are typically within the experimental accuracy of the two aligned structures. Similarly, squaring the Drmsdi,j weighting function softens the negative impact of larger rmsd values (>2–5 Å) and still allows for a positive (nonzero) contribution to the scoring function. These rmsd values are consistent with generally accepted measures of accuracy for predicted protein– ligand models and imply a potential functional relevance.[42] Thus, a continuous $\Delta\mathrm{rmsd}_{i,j}$ weighting function is created by simply subtracting 1 Å from the observed rmsd value, where a negative value is set to zero. So, an observed $\mathrm{rmsd}_{i,j}$ of 1.3 Å would result in a $\Delta\mathrm{rmsd}_{i,j}$ of 0.3 Å and a resulting 0.741 weighting function on the BLOSUM62 probability. Conversely, an observed $\mathrm{rmsd}_{i,j} < 1.0$ Å would result in a $\Delta\mathrm{rmsd}_{i,j}$ of 0 Å and a resulting 1.0 weighting function on the BLOSUM62 probability.

Since the active site is defined by a strict distance cutoff, relatively large errors may arise in the alignment score due to small structural changes that may occur at the active site boundary. To minimize this effect, the score is also scaled by the shortest distance from an amino acid in the active site to the ligand to de-emphasize amino acids that are at the 6-Å boundary. As an illustration, consider an active site of a targeted protein that contains an alanine where the methyl protons are exactly at the 6-Å limit. The remaining alanine atoms are all beyond the 6-Å limit. The active site of a reference protein does not include this alanine as part of its active site definition because the alanine methyl protons are 6.1 Å from any ligand atom and beyond the 6-Å limit. Thus, because of this 0.1-Å change and the corresponding presence and absence of alanine in the two active site definitions, the similarity scoring function would decrease by 4.0, when these two active sites are compared. Assuming the shortest distance from any atom in the ligand to any atom in the active site is 2 Å , the impact on the similarity score is reduced to 1.33 by using the $d_{\min}/d_i$ (2 Å /6 Å) scaling. Conversely, the distance scaling also places more emphasis on active site amino acids that are closer to the ligand and are presumably more important in both the affinity and selectivity of the bound ligand.

Active Site Similarities

The CPASS program generates two outputs: (i) similarity score and (ii) a file containing the sequence alignment of the two active sites. The similarity score (S) is simply the ratio of the scoring function determined by comparing a protein target active site against a reference active site (S_{ab}) from the CPASS database, with the scoring function of a protein target active site compared against itself (S_{aa}).

$$S = S_{ab} / S_{aa} \times 100 \quad (2)$$

A similarity score is calculated for each comparison. Using the entire CPASS database would result in 26,000 similarity scores. The similarity score is not symmetrical and depends on the order of the comparison. This arises because the scoring function is dependent on the size or the number of amino acids that defines the active site.

Consider comparing a hypothetical or novel protein complexed with adenine against the CPASS database. It is plausible that reference proteins that are complexed with ATP, NAD, or FAD may exhibit a high similarity based on a near complete overlap with the adenine component of their ligand-defined active sites with the adenine complexed to the hypothetical protein. The reverse comparison would yield a significantly smaller similarity score, since a single adenine would only represent a subset of an active site defined by ATP, NAD, or FAD.

To simplify the utility of CPASS and the interpretation of the CPASS output, a web-based interface has been developed that will be accessible through our website http://bionmr-c1.unl.edu (Fig. 2). The CPASS output contains a list of all the aligned active sites, with a similarity score above the cut-off, typically 30%, that is directly linked to a graphical display of the aligned active sites, using Chime.[43,44] Additional information listed is the sequence alignment, the Cα rmsd-weighted function, the rmsd-weighted BLOSUM62 scores, and the protein and ligand identity from the PDB file.

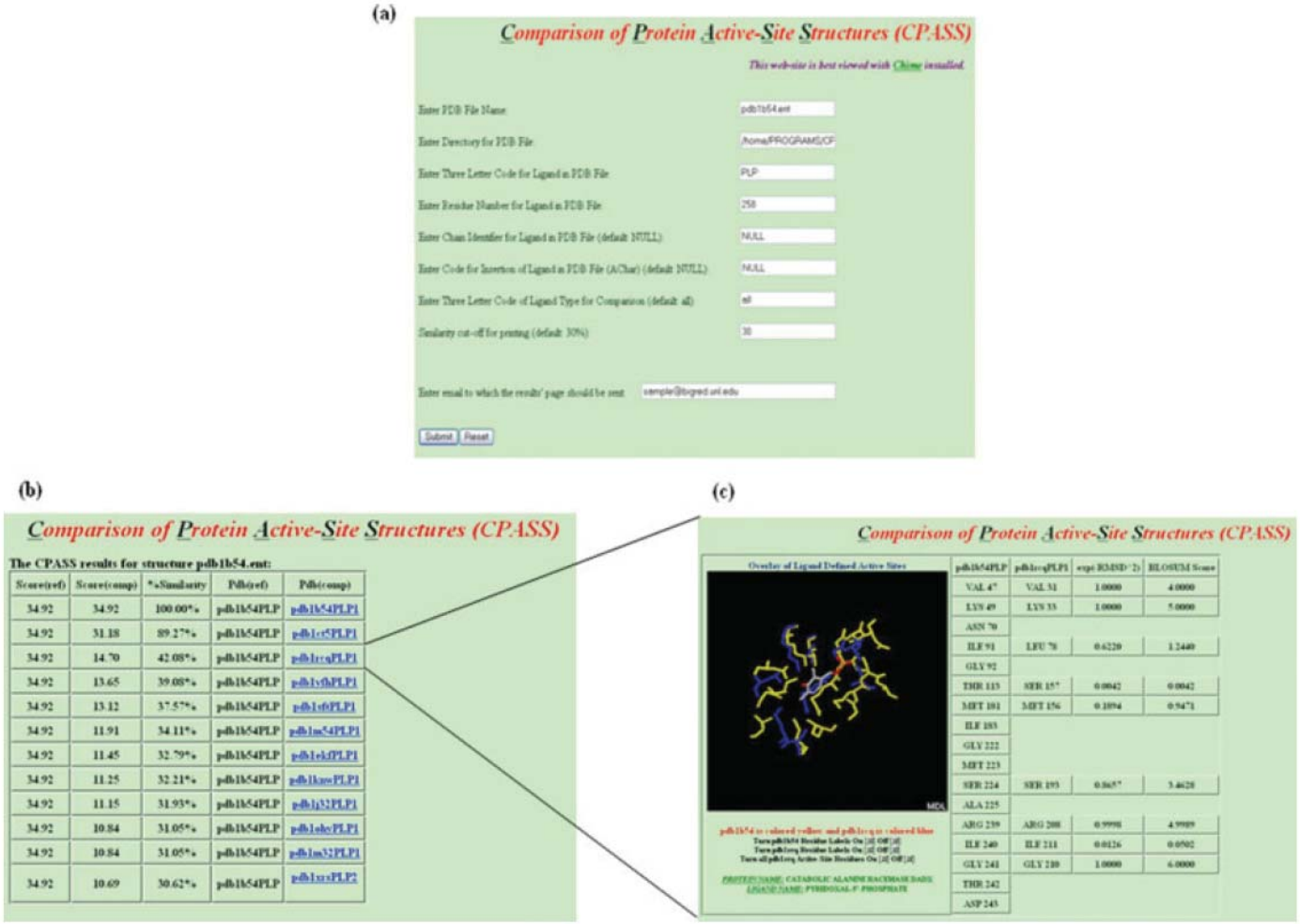

Fig. 2. Screen shots of the web interface to CPASS (a) entry form for active site comparison, (b) list of the active sites with the highest similarity to target protein, (c) graphical display of the aligned active sites' structures, sequence alignments, Ca rmsd weighted function, rmsd-weighted BLOSUM62 scores, and information about the aligned protein and its bound ligand. A hyperlink in the similarity list in (b) links to the display in (c).

RESULTS AND DISCUSSION

Validation of CPASS

The primary application of the CPASS program is to aid in the functional annotation of hypothetical or novel proteins by comparing experimentally-defined ligand based active sites. This is based on the premise that the sequence and structural composition of a protein active site is uniquely defined by the biological function of the protein. It is generally accepted that a global similarity in either sequence or tertiary fold of a protein is correlated to its function.[7] The underlying hypothesis in the application of CPASS is that a biological function may also be assigned to a protein, based on similarities in the characteristics of experimentally defined ligand based active sites in the absence of global sequence or structure homology.[45]

To address this hypothesis and validate the utility of the CPASS program, a general comparison of active site structures with known outcomes was conducted. The resolving power of ligand-defined active sites to identify protein function was ascertained by comparing ATP and pyridoxal 5'-phosphate (PLP) active sites from a variety of functionally distinct proteins. One hundred and seventy six ATP binding sites and 294 PLP binding sites were identified from structures in the PDB. The ATP binding sites were clustered into 19 functional classes based on the enzyme classification in the BRENDA database.[46,47] Similarly, the PLP binding sites were clustered into 20 functional classes. The ATP binding sites were compared with each other for a total of 30,976 comparisons. The PLP binding sites were compared with each other for a total of 86,436 comparisons. The calculations took ~1–2.5 days on a 16-node Beowulf Linux cluster, where each comparison averaged ~40 s. For each protein, the best match for each functional class was identified. Comparisons between proteins with ≥95% sequence similarity were excluded from identifying the best match. As an example, a phospho-transferase (PDB ID:1TQP) from *Archaeoglobus fulgidus* exhibits the highest similarity (52%) to a phosphotransferase (PDB ID:1PHK) from *Oryctolagus cuniculus*. Global sequence alignment of 1TQP with 1PHK using ClustalW[48] yielded an alignment score of only 8%. Conversely, the best match of a phosphotransferase to an alkyltransferase (PDB ID:1G64) is 15%. As anticipated, a higher average similarity score was always seen between proteins of identical function (diagonal) than functionally distinct proteins (off-diagonal) (Fig. 3). The results were independent of the type of ligand (ATP, PLP) or protein function. Nevertheless, the relative range of average similarity scores did vary by the function of the proteins. Comparison of ATP or PLP binding sites from functionally identical proteins resulted in relatively high similarity scores (~40–100%). Conversely, functionally distinct proteins generally yielded relatively low similarity scores despite binding the same ligand. Thus, the highest observed similarity score for a hypothetical protein determined by comparison against the CPASS database would identify the protein(s) that has the highest probability of sharing a similar function with the hypothetical protein.

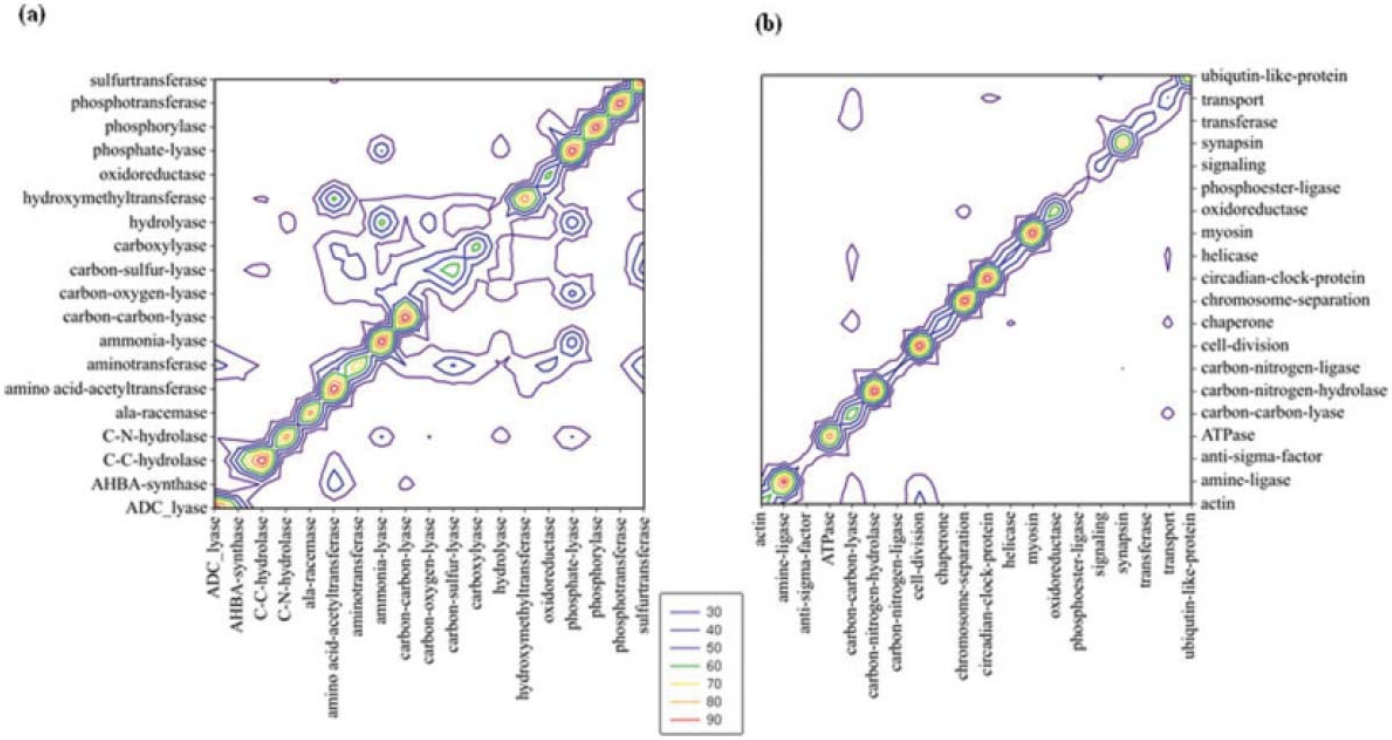

Fig. 3. A contour plot of the percent similarity determined from the CPASS analysis of (a) 294 pyridoxal 5'-phosphate binding sites and (b)176 ATP binding sites are plotted according to protein function. The diagonal compares proteins of identical function. Contours are plotted in 10% increments as indicated by the color chart, where the lowest observed contour is 30%.

Comparison of the ATP and PLP similarity plots (Fig. 3) clearly indicates a difference in the absolute magnitude of some of the off-diagonal peaks. A number of the PLP off-diagonal peaks indicate a 50–60% similarity between different functional classes, whereas the maximum off-diagonal peaks are 30–40% in the ATP plot. Also, more of the off-diagonal peaks are >30% in the PLP plot, where a majority of the ATP off-diagonal peaks are <30%. These observations reflect the relative evolutionary pathways of the ATP and PLP binding sites. Evolutionary analysis of PLP-dependent enzymes indicates only four independent lineages (completely different folds) resulting from two distinct divergent events (reaction specific, substrate specific) from a primordial PLP protein.[49] All PLP-dependent enzymes catalyze amino acid metabolism and as a result share important mechanistic features that include (i) covalent bond between PLP and a lysine residue, (ii) amino acid binding site proximal to PLP for transimination with substrate, (iii) formation of a planer coenzyme-substrate aldimine adduct, and (iv) optimization of noncovalent interaction between the protein and the PLP–substrate complex. These mechanistic requirements suggest that PLP caused evolutionary restraints relative to ATP-dependent proteins. Thus, the evolutionary analysis of PLP-dependent enzymes that indicates close relationships within this protein family is consistent with the high off-diagonal similarities observed in the CPASS analysis, using a narrower functional classification. As an example, CPASS indicates a 63.1% similarity between hydroxymethyl transferase (E.C. 2.1.2) and amino acid acetyl transferase (E.C. 2.3.1), which are both members of the α-family and closely related in the PLP phylogenic map.[49]

Although PLP-dependent enzymes appear to share a common evolutionary pathway, a similar relationship is not expected across the more functionally diverse family of ATP binding proteins. Clearly, the significant differences in function between actin and kinase proteins would imply a very distinct and unrelated evolutionary pathway. In fact, identifying an evolutionary relationship between divergent members of the kinase family alone is challenging.[50] These different functional classes of ATP binding proteins separately and distinctly optimized an ATP binding site specific to the functional needs of the protein. Any similarity in the ATP binding site would result from convergent evolution.[45,51,52] Again, this lack of a strong evolutionary relationship between the various ATP binding proteins is consistent with the relatively low off-diagonal similarity scores observed in the CPASS analysis.

The value of the CPASS analysis is also illustrated by a comparison of the global pair-wise sequence identity determined by ClustalW[48] for the 176 ATP-binding sites with the CPASS similarity score (Fig. 4). A general linear correlation between the CPASS and ClustalW alignment scores is expected and observed. Clearly, as the global sequence identity increases, a corresponding increase in the similarity of the active sites would also occur. This is fundamental to the application of sequence alignment to assign function. The two circled areas in the graph indicate regions that significantly deviate from this linearity.

Region (a) corresponds to CPASS similarity scores that are significantly higher than the corresponding ClustalW scores. This indicates a high similarity in the sequence and structure characteristics of active site for proteins with extremely low (<20%) sequence alignment. These low sequence alignments are not expected to yield a functional annotation, but are consistent with the observation that numerous homologous proteins structures exhibit high global sequence diversity.[53] Again, by emphasizing active site structural alignments with an inherently higher level of conservation relative to global sequence alignments, an increase in the probability of obtaining a functional annotation can be achieved using CPASS.

Region (b) in Figure 4 corresponds to low CPASS similarity for proteins with high sequence alignments. Proteins that have multiple ATP binding sites, which are sequence and structurally distinct, will result in low CPASS scores, when these distinct active sites are compared. This is an expected result and provides a negative control for validating CPASS. Of course, the overall sequence similarity would be high, even though the two ATP binding sites being compared are quite different.

Functional Annotation of Hypothetical Proteins

Further validation of the utility of CPASS to assist in the functional annotation of hypothetical proteins was ascertained by analyzing two structures of hypothetical proteins recently reported in the literature that serendipitously contained a bound ligand. The 2.0 Å X-ray structure of yeast hypothetical protein YBL036C contained a covalently attached pyridoxal 5'-phosphate. CPASS comparison against 294 active sites containing pyridoxal 5'-phosphate indicated that the best match (42% similarity) corresponded to an alanine racemase (Fig. 5).

The function of YBL036C had been tentatively identified as an alanine racemase.[54] Comparison of YBL036C against a structural database identified alanine racemase and ornithine

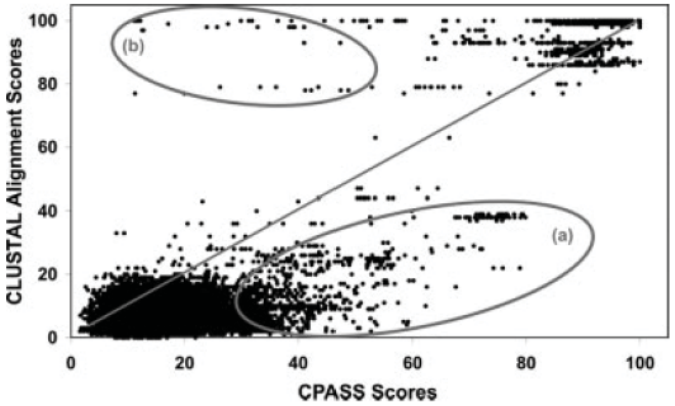

Fig. 4. Comparison of the CPASS active site similarity score and the global percent sequence similarity determined by ClustalW[48] for the 176 ATP binding sites. The circled areas represent significant deviations from a linear relationship between CPASS and ClustalW indicated by the straight line. Region (a) corresponds to high active site CPASS similarity scores for proteins with low global sequence similarity. Region (b) corresponds to proteins with multiple distinct ATP binding sites, where CPASS similarity is expected to be low.

decarboxylase as globally similar in tertiary fold to YBL036C, where all three proteins exhibit a similar TIM-barrel fold. Nevertheless, a poor alignment of YBL036C against alanine racemase was obtained using the entire TIM-barrel fold. Also, the structure of YBL036C could not be determined by molecular replacement using the alanine racemase structure. A brute-force alignment using a subset of the alanine racemase that included the PLP active site resulted in a significant improvement with a 1.72 Å rmsd. Manual comparison of the YBL036C and alanine racemase active sites suggested a significant similarity to justify testing for D-alanine racemase activity. YBL036C was shown to exhibit D-to l-alanine racemase activity. Thus, the CPASS assignment of YBL036C as an alanine racemase is consistent with the previously reported detailed structural and biochemical analysis.

Similarly, a 2.2 Å X-ray structure of hypothetical protein YecO from Haemophilus influenzae contained a bound *S*-adenosyl-L-homocysteine and is amenable to CPASS analysis. CPASS comparison against 46 structures containing *S*-adenosylmethionine and one structure containing *S*-adenosyl-L-homocysteine indicated that the best match (35% similarity) corresponded to a glycine *N*-methyltransferase. This example illustrates the use of CPASS to compare ligand-defined active sites, using related but chemically distinct ligand structures. In this case, *S*-adenosyl-L-homocysteine is a processed cofactor (Fig. 5).

The function of YecO has been identified as a methyl-transferase.[55] Again, this was based primarily on structural comparison using DALI[56] and VAST,[57] along with the presence of *S*-adenosyl-L-homocysteine. Methlytransferase have extremely low sequence homology (3–18%), but most methyltransferase bind the cofactor in a similar manner. Glycine *N*-methyltransferase binds *S*-adenosylmethionine in a drastically different binding mode, compared with other methyltransferase, and was identified as one of the structures most similar to YecO. Again, the CPASS assignment of YecO as a methyltransferase is consistent with the previously reported detailed structural analysis. CPASS identified glycine *N*-methyltransferase as exhibiting a similar active site structure as YecO is also consistent with this previous analysis. These results support the general application of the CPASS database and software to assign a biological function to novel or hypothetical proteins, by comparing experimentally determined active sites.

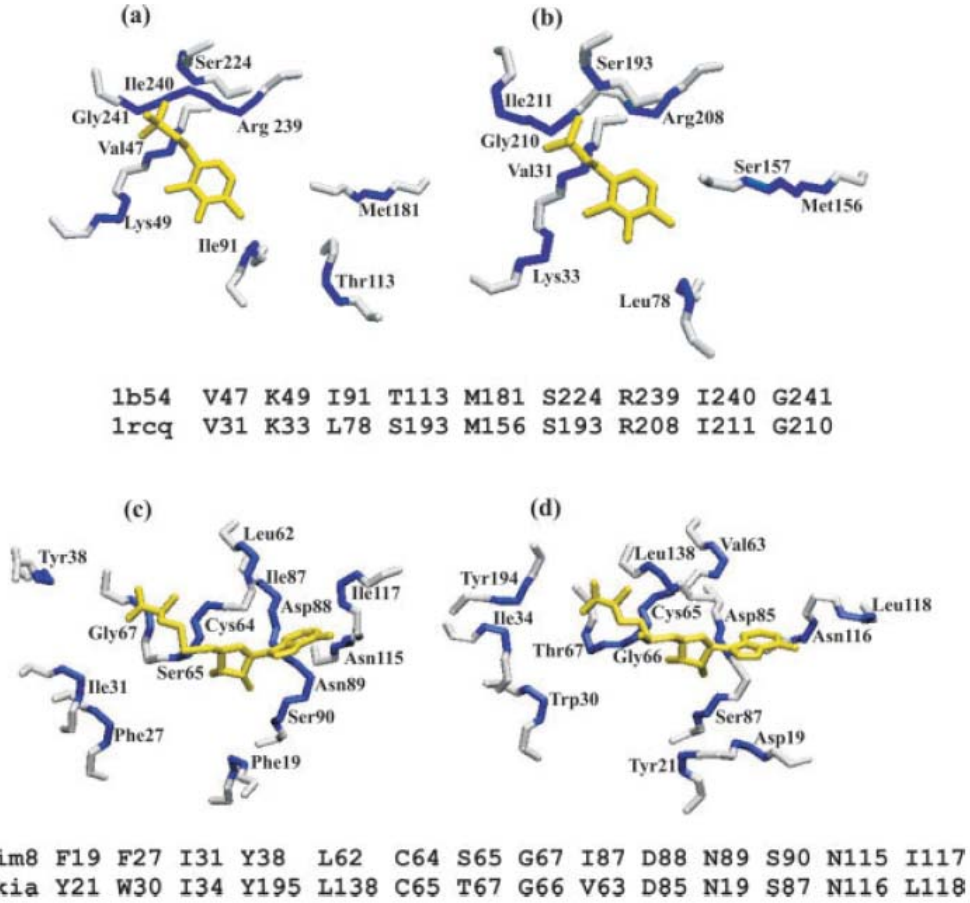

Fig. 5. Top: Comparison of the pyridoxal-5'-phosphate defined active sites for (a) yeast hypothetical protein YBL036C (PDB ID:1B54) and (b) alanine racemase (PDB ID:1RCQ). Bottom: Comparison of the *S*-adenosyl-L-homocysteine defined active site for (c) hypothetical protein YecO from *Haemophilus influenzae* (PDB ID:1IM8) with the (d) *S*-adenosylmethionine defined active site for glycine *N*-methyltransferase (PDB ID:1KIA). The residues aligned by CPASS are colored blue in the structures, and the active site sequence alignments are shown below the structures. Pyridoxal-5'phosphate, *S*-adenosyl-L-homocysteine, and *S*-adenosylmethionine are colored yellow.

Comparison of CPASS to Other Methods

CPASS shares a similarity in concept to other techniques that are being developed to infer function for hypothetical proteins.[14–19,58] Like CPASS, these approaches are using information about the protein's active site to make a correlation with a known protein and assign a function to the unknown protein. Nevertheless, the application and details of the CPASS approach are fundamentally distinct from these other methods. For example, the ''Fuzzy Functional Form'' (FFF) described by Fetrow and Skolnick[59] was developed to provide genome-wide functional annotation, using only the amino acid sequences for each hypothetical protein. The success of the FFF approach is monitored by the number of correctly annotated proteins instead of its ability to correctly annotate a specific protein, which is the objective of CPASS. Thus, the computational speed and broad coverage requirements of FFF result in significant compromises relative to CPASS that included a simplified and limited active site comparison and the complete absence of experimental data. The three-dimensional protein structure, the identity of the active site, and protein–ligand complex structures are unknowns in FFF. In fact, the aim of FFF is to predict the structure and identity of the active site simply from the sequence of the hypothetical protein and a few structure templates for proteins of known function.[17] This is a very laudable but challenging goal. Conversely, CPASS depends on the experimentally determined structure and the unambiguous ligand-defined active site to provide functional information for a single protein.

Briefly, FFF predicts a 3D structure for each hypothetical protein, by threading the sequence into 2–3 structures for proteins of a specific function.[60] Second, FFF uses a consensus active site defined from a sequence alignment of functionally annotated proteins, where an active site residue must be present in $\geq$50% of the aligned sequences. A functional assignment is then made if the threaded sequence is consistent with one of the template structures and if *all* the conserved active site residues overlap with the structural template.

Unlike CPASS, only a few highly conserved amino acids are used to define an active site instead of a complete description for all the active site residues. Conversely, FFF requires that the predicted active site for the hypothetical proteins contain an exact match with the consensus active site, where CPASS provides a similarity score that allows for homologous amino acid substitution. Again, speed dictates this requirement in FFF, but the detailed comparison that is achieved by the precise comparison of ~26,000 ligand-defined active sites with CPASS is lost, potentially resulting in incorrect structural alignments and false assignments. Consider a simple hypothetical example, if a consensus active site contains a conserved aliphatic amino acid (Ala, Ile, Leu, Val), but neither of these residues is consistently present ($\geq$50%) in the aligned sequences, then FFF will not include this descriptor as part of its active site definition. As a result, a hypothetical protein that contains an Arg at this position would equally and probably incorrectly match the consensus active site. There is no differentiation from other hypothetical proteins that correctly contain this conserved amino acid type. Conversely, CPASS utilizes each individual active site for the sequence alignment, where the presence of Arg would result in a negative impact on the CPASS similarity score.

Furthermore, consider large functional families that contain hundreds of members, such as kinases and PTPases. Numerous functional subclasses potentially exist within these large families, where a consensus active site across the entire family is inappropriate, but accurately delineating membership within the subclasses and correctly defining a consensus active site for each subclass may be challenging.[50] The accuracy of the functional assignment for FFF is strongly dependent on the correct description of these consensus active sites. These issues are avoided in CPASS by using the entire ligand-defined active site for comparison (all the individual kinase, PTPase along with other protein active sites are used). CPASS specifically identifies which protein-ligand complexes in the CPASS database and shares a homologous active site with the hypothetical protein. This aspect of CPASS is more computationally intensive relative to FFF, since it requires a comparison of ~26,000 ligand-defined active sites comprising upwards of $\geq$25 amino acids each. But, the structural threading is similarly computationally expensive in the FFF protocol requiring a limited number of structural templates.

Other approaches similar to FFF also attempt to predict function or identify active sites through the use of homology models based on known protein structures.[18,19] These models generally suffer from an abundance of false positives because of the accuracy of the threading procedure. An accurate threaded structure requires 60% of residues in the hypothetical sequence to occupy structurally analogous sites in the target structure.[61] Thus, the sequence for the hypothetical protein needs to share more than 50% sequence identity with the protein structure template.[62] Nevertheless, any sequence can be threaded into a structure template and simply evaluated by an empirical energy function, resulting in incorrect predicted folds. CPASS does not attempt to predict a structure for a hypothetical protein but requires the availability of this structure and avoids the uncertainty generated by a predicted structure. Effectively, FFF and other similar programs are analogous to global sequence alignments, but utilize a structural homology filter to refine the global sequence alignment.

Application of CPASS in Drug Discovery

An important issue in drug discovery is designing selectivity into chemical leads to avoid undesirable activity that may cause toxic side-effects in clinical trials.[63] Improving the affinity of a chemical lead against a defined protein target can be readily quantified, but determining the relative selectivity against *all* potential targets is impractical. The main challenge is in identifying proteins that may be inadvertent targets of the chemical lead. Again, global sequence or structural homology to the protein target is the major method of identifying proteins with a potential affinity to the chemical lead. Unfortunately, this does not yield a thorough analysis of the proteome or a prediction of ligand affinity, since the comparison is not specific to the active

site. CPASS provides an additional approach to identifying potential cross-reactivity between proteins of diverse function by identifying related ligand-biding sites. The ATP and PLP binding-site analysis indicates that the highest observed similarity is between proteins of identical function (Fig. 3). Nevertheless, there are a number of examples where functionally distinct proteins share >30–40% similarity (off-diagonal), such as ATPases and cell division proteins for the ATP binding proteins and hydrolyase and ammonia-lyase for the PLP-dependent enzymes. Again, CPASS will not provide a complete analysis of the entire proteome, since it is limited to the representative protein structures and functions in the PDB. But, CPASS will assist in improving the selectivity of chemical leads by expanding the list of relevant proteins beyond those proteins that are functionally related to the target. Thus, CPASS can identify a broader spectrum of proteins to use in biological assays to test for activity and selectivity against potential drug candidates.

REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Consortium IHGS. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.
2. Venter C, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, George L, Miklos G, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji R-R, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang ZY, Wang A, Wang X, Wang J, Wei M-H, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu SC, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers Y-H, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo´ R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang Y-H, Coyne M, Dahlke C, Mays AD, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. *Science* 2001;291:1304–1351.
3. Xu D, Xu Y, Uberbacher EC. Computational tools for protein modeling. *Curr Protein Pept Sci* 2000;1:1–21.
4. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-H, Smith HO. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* (Washington, DC) 2004;304:66–74.
5. KolkerE, Picone AF, Galperin MY, Romine MF, Higdon R, Makarova KS, Kolker N, Anderson GA, Qiu X, Auberry KJ, Babnigg G, Beliaev AS, Edlefsen P, Elias DA, Gorby YA, Holzman T, Klappenbach JA, Konstantinidis KT, Land ML, Lipton MS, McCue L-A, Monroe M, Pasa-Tolic L, Pinchuk G, Purvine S, Serres MH, Tsapin S, Zakra-

jsek BA, Zhu W, Zhou J, Larimer FW, Lawrence CE, Riley M, Collart FR, Yates JR, III, Smith RD, Giometti CS, Nealson KH, Fredrickson JK, Tiedje JM. Global profiling of *Shewanella oneidensis* MR-1: expression of hypothetical genes and improved functional annotations. *Proc Natl Acad Sci USA* 2005;102:2099–2104.

6. Yakunin AF, Yee AA, Savchenko A, Edwards AM, Arrowsmith CH. Structural proteomics: a tool for genome annotation. *Curr Opin Chem Biol* 2004;8:42–48.
7. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. *Q Rev Biophys* 2003;36:307–340.
8. Overington JP. Comparison of three-dimensional structures of homologous proteins. *Curr Opin Struct Biol* 1992;2:394–401.
9. Powers R, Mirkovic N, Goldsmith-Fischman S, Acton TB, Chiang Y, Huang YJ, Ma L, Rajan PK, Cort JR, Kennedy MA, Liu J, Rost B, Honig B, Murray D, Montelione GT. Solution structure of *Archaeglobus fulgidis* peptidyl-tRNA hydrolase (Pth2) provides evidence for an extensive conserved family of Pth2 enzymes in archea, bacteria, and eukaryotes. *Protein Sci* 2005;14:2849–2861.
10. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol* 1999;291:177–196.
11. Ekins S. Predicting undesirable drug interactions with promiscuous proteins in silico. *Drug Discov Today* 2004;9:276–285.
12. Funk CJ, Davis AS, Hopkins JA, Middleton KM. Development of high-throughput screens for discovery of kinesin adenosine triphosphatase modulators. *Anal Biochem* 2004;329:68–76.
13. Traxler P, Furet P, Mett H, Buchdunger E, Meyer T, Lydon N. Design and synthesis of novel tyrosine kinase inhibitors using a pharmacophore model of the ATP-binding site of the EGF-R. *J Pharm Belg* 1997;52:88–96.
14. Cammer SA, Hoffman BT, Speir JA, Canady MA, Nelson MR, Knutson S, Gallina M, Baxter SM, Fetrow JS. Structure-based active site profiles for genome analysis and functional family subclassification. *J Mol Biol* 2003;334:387–401.
15. Goldman BB, Wipke WT. Quadratic shape descriptors. I. Rapid superposition of dissimilar molecules using geometrically invariant surface descriptors. *J Chem Inf Comput Sci* 2000;40:644–658.
16. Kinoshita K, Nakamura H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci* 2003;12:1589–1595.
17. Fetrow JS, Godzik A, Skolnick J. Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J Mol Biol* 1998; 282:703–711.
18. Guo T, Shi Y, Sun Z. A novel statistical ligand-binding site predictor: application to ATP-binding sites. Protein Eng Des Sel 2005; 18:65–70.
19. Kitson DH, Badretdinov A, Zhu ZY, Velikanov M, Edwards DJ, Olszewski K, Szalma S, Yan L. Functional annotation of proteomic sequences based on consensus of sequence and structural analysis. *Brief Bioinform* 2002;3:32–44.
20. Johnson JM, Church GM. Predicting ligand-binding function in families of bacterial receptors. *Proc Natl Acad Sci USA* 2000;97: 3965–3970.
21. Gutteridge A, Thornton JM. Understanding nature's catalytic toolkit. *Trends Biochem Sci* 2005;30:622–629.
22. Laskowski RA, Watson JD, Thornton JM. Protein function prediction using local 3D templates. *J Mol Biol* 2005;351:614–626.
23. Tian W, Arakaki AK, Skolnick J. EFICAz: A comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res* 2004;32:6226–6239.
24. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol* 2002;323:387–406.
25. Naumann T, Matter H. Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: target family landscapes. *J Med Chem* 2002;45: 2366–2378.
26. Ko J, Murga LF, Wei Y, Ondrechen MJ. Prediction of active sites for protein structures from computed chemical properties. *Bioinformatics* 2005;21(Suppl. 1):i258–i265.
27. Greaves R, Warwicker J. Active site identification through geometry-based and sequence profile-based calculations: burial of catalytic clefts. *J Mol Biol* 2005;349:547–557.
28. Ringe D, Wei Y, Boino KR, Ondrechen MJ. Protein structure to function: insights from computation. *Cell Mol Life Sci* 2004;61: 387–392.
29. Gutteridge A, Bartlett GJ, Thornton JM. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol* 2003;330:719–734.
30. La D, Sutch B, Livesay DR. Predicting protein functional sites with phylogenetic motifs. *Proteins Struct Funct Bioinform* 2004; 58:309–320.
31. Stuart AC, Ilyin VA, Sali A. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics* 2002;18:200,201.
32. Shin J-M, Cho D-H. PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res* 2005;33(Database):D238– D241.
33. Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman HM, Westbrook J. Ligand depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* 2004;20:2153–2155.
34. Hendlich M, Bergner A, Gunther J, Klebe G. Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol* 2003;326:607–620.
35. Kleywegt GJ, Jones TA. Databases in protein crystallography. *Acta Crystallogr D Biol Crystallogr* 1998;D54(6, Part 1):1119–1131.
36. Schwieters CD, Clore GM. The VMD-XPLOR visualization package for NMR structure refinement. *J Magn Reson* 2001;149: 239–244.
37. Wang G, Dunbrack RL, Jr. Scoring profile-to-profile sequence alignments. *Protein Sci* 2004;13:1612–1626.
38. Goonesekere NCW, Lee B. Frequency of gaps observed in a structurally aligned protein pair database suggests a simple gap penalty function. *Nucleic Acids Res* 2004;32:2838–2843.
39. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;89:10915–10919.
40. Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrixes. *Proteins Struct Funct Genet* 1993;17:49–61.
41. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997; 25:3389–3402.
42. Mendez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins Struct Funct Genet* 2003;52: 51–67.
43. Cammer SA. SChiSM: creating interactive web page annotations of molecular structure models using chime. *Bioinformatics* 2000;16:658,659.
44. Pembroke JT. Bio-molecular modelling utilizing RasMol and PDB resources: a tutorial with HEW lysozyme. *Biochem Mol Biol Educ* 2000;28:297–300.

45. Denessiouk KA, Johnson MS. When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families. *Proteins Struct Funct Genet* 2000; 38:310–326.

46. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* 2004;32(Database): D431–D433.

47. Pharkya P, Nikolaev EV, Maranas CD. Review of the BRENDA database. *Metab Eng* 2003;5:71–73.

48. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680.

49. Christen P, Mehta PK. From cofactor to enzymes. The molecular evolution of pyridoxal-5'-phosphate-dependent enzymes. *Chem Rec* 2001;1:436–447.

50. Labesse G, Douguet D, Assairi L, Gilles A-M. Diacylglyceride kinases, sphingosine kinases and NAD kinases: distant relatives of 6-phosphofructokinases. *Trends Biochem Sci* 2002;27:273–275.

51. Denessiouk KA, Lehtonen JV, Johnson MS. Enzyme-mononucleotide interactions: three different folds share common structural elements for ATP recognition. *Protein Sci* 1998;7:1768– 1771.

52. Kabsch W, Holmes KC. Protein motifs. II. The actin fold. *FASEB J* 1995;9:167–174.

53. Aloy P, Oliva B, Querol E, Aviles FX, Russell RB. Structural similarity to link sequence space: new potential superfamilies and implications for structural genomics. *Protein Sci* 2002;11: 1101–1116.

54. Eswaramoorthy S, Gerchman S, Graziano V, Kycia H, Studier FW, Swaminathan S. Structure of a yeast hypothetical protein selected by a structural genomics approach. *Acta Crystallogr D Biol Crystallogr* 2003;59:127–135.

55. Lim K, Zhang H, Tempczyk A, Bonander N, Toedt J, Howard A, Eisenstein E, Herzberg O. Crystal structure of YecO from Haemophilus influenzae (HI0319) reveals a methyltransferase fold and a bound s-adenosylhomocysteine. *Proteins Struct Funct Genet* 2001; 45:397–407.

56. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 1993;233:123–138.

57. Gibrat J-F, Madej T, Bryant SH. Surprising similarities in structure comparison. *Curr Opin Struct Biol* 1996;6:377–385.

58. Pazos F, Sternberg MJE. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci USA* 2004;101:14754–14759.

59. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol* 1998;28:949–968.

60. Xu D, Xu Y, Uberbacher EC. Computational tools for protein modeling. *Curr Protein Pept Sci* 2000;1:1–21.

61. Bryant SH. Evaluation of threading specificity and accuracy. *Proteins Struct Funct Genet* 1996;26:172–185.

62. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* (Washington, DC) 2001;294:93–96.

63. Kubinyi H. Drug research: myths, hype and reality. *Nat Rev Drug Discov* 2003;2:665–668.

Supplemental material follows (2 pp.).

Supplementary Material: " Comparison of Protein Active Site Structures for Functional Annotation of Proteins and Drug Design." Robert Powers, Jennifer C. Copeland, Katherine Germer, Kelly A. Mercier, Viswanathan Ramanathan and Peter Revesz

Table 1S: List of Ligands and PDB HET Labels Excluded from the CPASS Database

2-mercaptanol	BME, SEO	lanthanum	LA
acetate	ACT	lead	PB, PBM
acetic acid	ACY	lithium	LI
acetone	ACN	lutetium	LU
acetonitrile	CCN	magnesium	MG, MO1, MO2, MO3, MO4, MO5, MO6
alcohol	IPA	manganese	MN, MN3, MN5, MW1, MW2, MW3, MH2, MH3, O4M
aluminum	AL	mercury	HG, HG1
amide	AF3,NH2,NH4,NH3, NH	methanol	MOH
antimony	SBO, ND4	methylamine	NME
argon	SB	methyl phosphinic acid	SOM
arsenic	ARS, AST, AR	MES	MES
azide	AZI	molybdenum	MOS, MO7, OMO, MOO, MM4, MO, 4MO, 6MO
barium	BA	nickel	NI, 3NI, NI1, NI2
beryllium	BEF, BF2, BF4	nickel-iron	NFE
bicarbonate	BCT	nitogen dioxide	2NO
borate	BO4	nitrate	NO3
boric acid	BO3	nitrite	NO2
bromine	BR, BRO	nitrogen monoxide	NMO
bromomercury	HG2	nitrogen oxide	NO
cacodylate	CAC	osmium	OS
cadmium	CD, CD1, CD3, CD5	oxygen	O, OX, OEC, O2, OXY, HF5
calcium	CA, OC1, OC2, OC3, OC4, OC5, OC6, OC7, 543	palladium	PD
carbon dioxide	CO2	perchlorate	LCP
carbon monoxide	CMO	peroxide	PEO, PER
cerium	CE	phosphate	2HP, DPO, FPO, PI, IPS, PO4, 3PO
cesium	CS	phosphite	PO3
chlorine	CL, CLO, CFO, LCO	platinum	PCL
cobalt	CO, 3CO, NCO, OCL, OCN, OCM, CO5, OCO, CON	porphyrin	HCO
copper	CU, CU1, ICU, CUO, C2C, C1O, C2O, CUA, CUZ	potassium	K, KO4
copper chloride	CUL	praseodymium	PR
copper-sulfer cluster	CUN, CUM	rhenium	RE, RTC
cyanide	CN, CYN	rhodium	RHD
dimethylformamide	DMF	rithenium	RU
DMSO	DMS	rubidium	RB
ethanol	EOH	samarium	SM
ethylene glycol	EDO, EGL	selenium	SE4, SE, MSE

europium	EU, EU3	silver	AG
fluorine	F, FLO	sodium	NA, NAW, NAO, NA2, NA5, NA6
formic acid	FMT	strontium	SR
gadolinium	GD, GD3	sulfate	SOH, SUL
galluim	GA	sulfite	SO3
glucosamine	NAG	sulfur	S
glycerol	GOL	sulfur dioxide	SO2
gold	AU, AU3, AUC	sulfur oxide	SX
holimium	HO	tantalum	TBR
hydrogen	H	tellerium	TE
hydrosulfuric acid	H2S	terbium	TB
hydroxy	OH, HYD	thallium	TL
hypophosphite	PO2	tungsten	W, WO4, WO5
indium	IN	uranyl	IUM
iodine	IOD, IDO	vanadium	V, V7O, VO3, VO4
iridium	IR, IR3, IRI	water	MTO, DOD, HOH, WAT
iron	OF2, HC1, FCO, FE, FE2, OF3, OF1, 2OF, FEL, OFO, FEA, FEO	xenon	XE
iron-sulfur cluster	WCC, XCC, NFS, CFM, CFN, CLP, FES, F3S, FS3, FS4, SF4, FSO	ytterbium	YB
krypton	KR	yttrium	YT3, Y1
glucosamine	NAG	zinc	ZN, ZN2, ZN3,
various small molecular fragments or clusters	ACE, BUT, CBZ, CO3, CBX, CBM, CM, MCE, CBG, DTN, ETD, ETH, OET, EMC, EOX, FOR, HOA, OHE, OME, 2ME, CH3, TML, MCB, CH2, HDZ, TFH, WO2, OXO, ZRC, CNB, CN1, CNF, OXA, QTR, CYO, OMB, 2PO, PHS, PPM, PVL, SBU, HF3, SFO, SFN, DML, TBU, NTB, ALF, TMA, THJ, SCN, SCC, TFA, MGF, TME, CYA, UNX, UNK, UNL, U1, DIS		