Electrostatic Properties for Protein Functional Site Prediction

Joslynn S. Lee and Mary Jo Ondrechen

Abstract The development of computational tools for the prediction of protein function from the three-dimensional structure is a very important problem in the post-genomic era. To date there are over 9,900 structural genomics protein structures in the Protein Data Bank and most of these are of unknown or uncertain function. Methods for the identification of the residues in a protein structure that participate in the biochemical function provide key information about the function of the protein. We and others have developed computational methods for the prediction of functionally important residues in proteins. The focus of this chapter is on protein function at the atomic level, *i.e.* catalysis and recognition. Methods that utilize computed electrostatic properties, specifically THEMATICS and POOL, are described.

Introduction

The development of computational tools for the prediction of protein function from the three-dimensional structure is a very important problem in the post-genomic era. To date there are over 9,900 structural genomics protein structures in the Protein Data Bank [1–2] and most of these are of unknown or uncertain function. Methods for the identification of the residues in a protein structure that participate in the biochemical function provide key information about the function of the protein. We and others have developed computational methods for the prediction of functionally important residues in proteins. The focus of this chapter is on *protein function at the atomic level*, i.e. *catalysis and recognition*, and on methods that utilize computed electrostatic properties.

Computed electrostatic properties bring special advantages to the quest for functional information about a protein structure. First of all, *they require only the structure of the query protein as input*. Thus they return predictions even for novel

M.J. Ondrechen (✉)
Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115, USA
e-mail: M.Ondrechen@neu.edu

D. Kihara (ed.), *Protein Function Prediction for Omics Era*,
DOI 10.1007/978-94-007-0881-5_10, © Springer Science+Business Media B.V. 2011

folds and engineered structures, as well as for proteins with orphan sequences or with few sequence homologues. Furthermore, these predictions are just as reliable for these difficult cases as they are for the well-characterized proteins in the benchmark sets used for the testing and verification of the methods. Second, *these properties are directly related to the chemistry of individual residues* and thus are well suited to the identification of residues with special catalytic or binding properties. Finally, *electrostatics-based methods are orthogonal to the more common sequence-based methods* that rely on sequence alignments and phylogenetic trees; thus, when information about sequence conservation or evolutionary history is available, combination of the methods can, at least in principle, lead to significant enhancement in the quality of the predictions. Indeed, electrostatics-based methods have proved to be powerful tools for functional site prediction.

In the prediction of functionally important residues, there is always a trade-off between sensitivity (the ability to predict the maximum number of truly important residues) and selectivity (the ability to predict only the truly important residues and not the unimportant residues). The goal is to *maximize sensitivity while minimizing false positives*.

In order to test the performance of predictors of functionally important residues, an annotated dataset is needed as a benchmark. Typically the Catalytic Site Atlas (CSA) [3–4], a referenced compilation of catalytically active residues previously identified in the literature for hundreds of enzymes, is used to obtain the validation set for functional site prediction methods. While no listing of catalytically active residues can possibly be complete, as not all residues have been tested experimentally and reported, the CSA represents the best available compilation of known catalytic residues.

Performance in catalytic residue prediction is defined in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Positives and negatives are defined using the CSA as the reference set. The recall rate for catalytic residue prediction is defined as:

$$\text{Recall} = \text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \tag{1}$$

The false positive rate is defined as:

$$\text{False positive rate} = \text{FP}/(\text{TN} + \text{FP}) \tag{2}$$

Finally the specificity is defined as:

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \tag{3}$$

The specificity is related to the false positive rate (FPR) as:

$$\text{Specificity} = 1 - \text{FPR} \tag{4}$$

Previously our group has reported on THEMATICS (for Theoretical Microscopic Titration Curves), an electrostatics-based method for the prediction of functionally important residues in protein 3D structures [5–7]. THEMATICS has been shown to predict functionally important residues with good sensitivity and a low false positive rate [7]. More recently, Partial Order Optimum Likelihood (POOL) [8] utilizes THEMATICS and other input features in a new, monotonicity-constrained maximum likelihood machine learning method, for enhanced performance in prediction of catalytic and binding residues.

Methods

THEMATICS

In the application of THEMATICS, the electrical potential function of the protein structure is first computed using a finite difference Poisson-Boltzmann procedure. Then a hybrid method [9] is used to compute theoretical titration curves for each of the ionizable residues. These titration curves take the form of the proton occupation for each residue as a function of the pH. The shapes of the titration curves are evaluated by an automated procedure, using the curve shape metrics described by Ko et al. [6] to quantify the degree of deviation from the typical Henderson-Hasselbalch (H-H) titration behavior. These curve shape metrics are subjected to statistical analysis in order to identify the residues that deviate most from the ideal H-H curve shape. Note that THEMATICS predictions are based on the *shapes* of the computed titration curves and not on the computed pK_a shifts, although THEMATICS has sometimes been described incorrectly as a pK_a shift method [10]. While pK_a shifts are common in active sites, they also occur too frequently in other parts of protein structures, e.g. salt bridges, to give precise active site predictions.

THEMATICS has been established as a successful, top performing site predictor across a wide range of enzymes from all functional classes [7]. In order to verify its effectiveness in catalytic site prediction, THEMATICS was applied to the entire original, manually curated set of 170 enzymes in the CSA [3–4]. THEMATICS was shown to identify, with high selectivity, all or some of the residues in known interaction sites in 93% of enzymes [7]. When performance in the prediction of annotated catalytic residues was compared with that of other 3D-structure-based methods, THEMATICS showed better sensitivity with much lower false positive rates, as demonstrated by the ROC (Receiver Operator Characteristic – i.e. true positive rate versus false positive rate) curves [7]. A very important characteristic of THEMATICS performance is its selectivity – it predicts precise, highly localized sites [7].

A key feature of THEMATICS is that the query protein does not have to have any similarity, in sequence or in structure, to any other protein. Originally function prediction was based primarily on sequence analysis, although sequence similarity does not always imply functional similarity [11–13]. Other methods use structural relationships in conjunction with sequence analysis [14–32] for improved performance.

There are presently a few approaches in addition to THEMATICS that are based solely on the structure of the query protein and some of these also employ electrostatic properties. Elcock [33] reported that likely functional residues could be identified by their electrostatic folding free energies obtained from solution of the Poisson-Boltzmann equations. Bate and Warwicker [34] later identified a point near the active site using the peak of the electrostatic potential in the solvent space above the protein structure. A graph theoretic approach predicts candidate active site residues based on their closeness of interaction with the other residues in the structure [35]. Another method uses purely geometric features of the protein structure [36]. More recently, ligand binding sites have been predicted through the computational identification of regions where interactions cause a large change in protein conformation distribution [37]. Ligand binding sites can also be detected with the mapping of small solvent-like molecules onto the protein surface, either experimentally [38] or with the corresponding computational docking method [39]. The method of Laurie and Jackson [40] is of this type, but uses only a single van der Waals probe.

THEMATICS, which requires no sequence alignments, has been shown to match performance, or even outperform, the best methods that predict functional sites from sequence alignments *and* the 3D structure. However, it is important to note that the performance of the methods that require a sequence alignment is expected to degrade [24–25] when applied to Structural Genomics (SG) proteins that have fewer, or less diverse, sequence homologues than the well studied proteins in the verification sets. On the other hand, *THEMATICS performance on SG proteins in principle should match its performance on the verification set* because it requires only the 3D structure of the query protein and it treats all input structures in the same fashion; it does not depend on any prior knowledge or relationships to other proteins.

THEMATICS predictions are freely available via the pfweb server: http://pfweb.chem.neu.edu/thematics/submit.html.

Users can either upload a protein structure file in pdb format, or alternatively give the PDB ID for the structure of interest. THEMATICS calculations on the server utilize the optimum statistical and distance cut-offs determined by Wei et al. [7]; these values return the highest Matthews Correlation Coefficient (MCC), as measured using the CSA annotations. The maximum MCC reflects a balance between sensitivity and specificity. Results are returned to the user via e-mail. Results take the form of one or more clusters.

For instance, for the dimer structure with PDB ID 2qe8, an uncharacterized structural genomics protein from *Anabaena variabilis,* THEMATICS returns two clusters for each of the two subunits of the dimer, a seven-member cluster [D123, K246, C249, D250, D293, D306, R342] and a one-member cluster [D202]. Only clusters with two or more residues are considered predictive; thus the seven-member cluster constitutes the functional site prediction and the single-member cluster [D202] is not a part of the predicted active site.

POOL

A new machine learning approach, called Partial Order Optimum Likelihood (POOL) was designed [8] to make significant enhancements in site prediction capability. Originally POOL was applied using THEMATICS input features and later was expanded to include other types of input features. In principle, POOL can use any input feature, provided the probability of the functional importance of residues depends monotonically on that feature.

POOL, a multidimensional, monotonicity-constrained maximum likelihood technique, starts with the hypothesis that the larger the THEMATICS metrics for a given residue, the higher the probability that the residue is important for function. These features consist of two computed properties, called μ_3 and μ_4 [6], of ionizable residues that describe titration curve shape. Extension of the POOL method to include predictions of non-ionizable residues is achieved through the introduction of *environment variables*. While THEMATICS features apply only to the ionizable residues (Arg, Asp, CysH, Glu, His, Lys, Tyr, and the N- and C- termini), the environment variables $\mu_3{}^{\text{env}}$ and $\mu_4{}^{\text{env}}$ measure the magnitude of the THEMATICS features of the ionizable residues that are spatially close to the residue in question. Note that $\mu_3{}^{\text{env}}$ and $\mu_4{}^{\text{env}}$ are properties of all residues, not just ionizable residues. Thus the THEMATICS input feature for POOL is the four-dimensional vector (μ_3, μ_4, $\mu_3{}^{\text{env}}$, $\mu_4{}^{\text{env}}$) for the seven residue types that are ionizable and the two-dimensional vector ($\mu_3{}^{\text{env}}$, $\mu_4{}^{\text{env}}$) for all of the non-ionizable residue types. This extension to include non-ionizable residues results in even better performance than with the original THEMATICS features alone and constitutes to date the best functional site predictor based on 3D structure only, achieving performance that is as good or nearly as good as methods that use both 3D structure and sequence alignment data [8].

It is interesting to note that the THEMATICS features μ_3 and μ_4 are derived from a function that is related to the binding capacity [41] for protons; μ_3 and μ_4 are also related to the coefficients in the proton binding polynomial [42].

These electrostatics features from THEMATICS are combined with multidimensional isotonic regression to form maximum likelihood estimates of probabilities that specific residues belong to an active site. This allows likelihood ranking of all ionizable residues in a given protein based on THEMATICS features. The corresponding ROC curves and statistical significance tests demonstrate that this method outperforms prior THEMATICS based methods, which in turn have been shown previously [7] to outperform other 3D-structure based methods for identifying active site residues.

POOL generates a value for each residue that is proportional to the probability that the residue is functionally important. One of the advantages of POOL is that it can incorporate any residue-based input feature upon which the probability of functional importance depends monotonically.

One such feature is the cleft size rank, an integer that represents the ordinal size of the surface cleft to which a given residue belongs. Previous studies have shown that active site residues tend to be located in one of the largest clefts in a protein structure [43–45]. Indeed it has been reported that in 83% of single-chain enzymes, the active site is located in the largest cleft [44]. Nearly all active sites are principally located in one of the five largest clefts of a protein structure, with the largest cleft containing the active site for the highest fraction of enzymes and with the fractions decreasing as the size rank progresses to smaller clefts [46]. The cleft size rank is a geometric feature that can be quickly computed for each residue in any protein structure. Although the cleft size rank alone does not perform very well for active residue prediction, its inclusion as input to POOL, as an addition to the THEMATICS input features, does lead to small but statistically significant improvement in site prediction performance [8].

Similarly, POOL easily incorporates sequence conservation scores, for those cases where there are a sufficient number of homologues. When this information is included, the resulting method has been shown to outperform the best methods that use any combination of sequence alignments and 3D structures [8]. It is further demonstrated that when THEMATICS features, cleft size rank, and alignment-based conservation scores are used individually or in combination, THEMATICS features represent the single most important component of such classifiers [8]. The POOL method we have developed is general and is a viable machine learning approach to any problem where a predicted outcome depends monotonically on each of the input variables. Most importantly, POOL is a top-performing site predictor and it *enables THEMATICS to be used to predict all residues, not just the ionizable ones*.

POOL output consists of a list of all residues, rank-ordered according to the probability of functional importance. The top-ranking residues constitute the POOL prediction. The cut-off point in the rank-ordered list may be set according to the intended application. The cut-off value is generally set to select the top 5–8% of all residues, as this returns good sensitivity with excellent specificity. A 5% false positive rate, which corresponds to 95% specificity, returns a recall rate of 70%, which is good enough to characterize a functional site. Full recall (100% sensitivity) is achieved with only a 17% false positive rate. This performance compares quite favorably with other methods, for instance INTREPID achieves 93% sensitivity with a 20% false positive rate [28] on a similar test set. However, false positive rates in the 17–20% range may be too high to be useful, as discussed below. We prefer to work with a little lower sensitivity but much better specificity; this combination is achievable with THEMATICS and POOL.

Discussion

What Is the Basis for the Success of THEMATICS?

As a standalone functional site predictor, THEMATICS has been shown to perform very well [7]. Its performance was measured on the original, manually curated set of 170 proteins in the Catalytic Site Atlas [3–4], where catalytic residues are labeled based on experimental literature citations. The THEMATICS success rate was found to be equal to or better than that of other 3D-structure-based methods, but with better precision and lower false positive rates [7]. This was all achieved with only one type of input, namely the computed titration curve shape metrics.

We attribute the success of the method, in particular its ability to predict highly localized, precise active sites, to its reliance on computed chemical properties. Chemically active residues are predicted with information about their chemistry, specifically their proton binding properties. While there is some error associated with the titration curves computed by electrostatics methods, the statistical [6–7] and machine learning [8] analyses on the curve shape metrics have proved to be highly successful in selecting the outliers, i.e. those residues with titration curve shapes that deviate most from typical Henderson-Hasselbalch behavior.

We have argued [47] that the anomalous titration behavior enables a residue, in a large ensemble of protein molecules, to exist in both protonation states with appreciable population over a wide pH range. This is in contrast to a typical Henderson-Hasselbalch weak acid or base, which is protonated at pH values less than the pK_a and deprotonated at pH values greater than the pK_a, with a very narrow pH range around the pK_a where both protonation states are populated in an ensemble of molecules. For the residues with anomalous titration behavior, this pH range is expanded significantly. This type of non-Henderson-Hasselbalch titration behavior is common for polyprotic acids and a protein is in fact a macromolecular polyprotic system.

Furthermore, for an active site residue, this ability to have both protonation states populated over a wider pH range is an advantage in catalysis [47]. First of all, by definition of a catalyst, a catalytic Brønsted-Lowry acid or base must be able to act as both acid and base because it must regenerate itself for the next turnover cycle. Thus a residue that donates a proton as part of the catalytic mechanism must also accept a proton before the end of each cycle. The anomalous titration behavior also enables catalytic residues to have the correct mix of properties. Consider for example one common first step in an enzyme-catalyzed reaction, the abstraction of a proton from an alpha carbon atom, a reaction that requires a strong base. Suppose that the enzyme in question operates in vivo at pH 7. Suppose that the conjugate acid of the catalytic base has a pK_a of 13 and that it obeys the Henderson-Hasselbalch equation. Such a base may not be strong enough to abstract a proton from a carbon atom, but even if it were, it would not be able to react because at pH 7, it is essentially fully protonated. Only one in one million protein molecules in the ensemble would have this residue deprotonated at pH 7. On the other hand, a base with anomalous titration behavior can be a strong base and at the same time have significant population of the deprotonated state at neutrality. Thus *the anomalous titration behavior helps to facilitate catalysis for active site residues.*

It is our working hypothesis that nature builds enzyme active sites with clusters of neighboring ionizable residues with similar pK_a values, so that there is strong interaction between their protonation events. This strong interaction gives rise to anomalous titration curve shapes and promotes catalysis. The deviations in the titration curve shape are measured by the features computed in a THEMATICS analysis.

The enhanced performance afforded by POOL using THEMATICS input features only is attributed to the ability of *POOL to extract more information from these features than the earlier statistical and machine learning analyses; this leads to better quality predictions of functionally important residues.* First POOL was applied with just THEMATICS features as input, using features similar to those used previously by our Support Vector Machine (SVM) classifier [48] and by our statistical selection [6–7]. Tong et al. showed in 2009 [8] that the POOL analysis outperforms all of these earlier THEMATICS analyses with no cleaning of the training data and no clustering after the classification. This suggests that the underlying monotonicity assumptions of POOL enable better use of the THEMATICS input metrics.

Another obvious reason for the success of POOL is its ability to *predict all residues, not just the ionizable residues.* In the previous statistical versions of THEMATICS, only seven types of residues are predicted: Arg, Asp, CysH, Glu, His, Lys, and Tyr. The N- and C- termini are also included in the original THEMATICS analysis, although these residues are only rarely involved in catalysis. Serine is excluded from the original THEMATICS analysis because its pK_a is generally too high for its deprotonation equilibrium to have significant interactions with those of other residues; attempts to include serine in the original THEMATICS analysis lead to lower quality predictions and thus serine has not been considered an ionizable residue for purposes of THEMATICS analyses. In spite of this, THEMATICS has still performed well compared to other 3D-structure-based methods [7]. This is in part because the seven residue types predicted by THEMATICS are the seven most prevalent catalytic residues. Among the literature-annotated catalytic residues analyzed by Bartlett et al. [3], the most common residue types, in order starting with the most common, are: His, Asp, Arg, Glu, Lys, CysH, and Tyr. Together these seven residue types constitute about 75% of all annotated catalytic residues [3, 7]. However this means that THEMATICS has a maximum residue recall rate, or sensitivity, of 75%, since by its nature it cannot predict the remaining 25% of catalytic residues. POOL is advantageous because it can predict all residue types. For instance, POOL predicts all three residues of the catalytic triad of serine proteases such as subtilisin, including the serine, whereas THEMATICS only predicts the Asp and His residues.

POOL is able to take advantage of a variety of input features, in addition to THEMATICS features. Any property of the residues in a protein structure can be a POOL input variable, provided the probability that a residue is catalytically important is a monotonic function of that variable. The current version of POOL incorporates a geometric feature, the cleft size rank, and the sequence conservation scores.

Table 1 summarizes POOL performance with and without conservation scores. The average specificities achieved at 90, 80, and 70% recall, together with the average recall rates achieved at 95, 90, and 80% specificity are shown. Specificity and recall are reported using all three input features, THEMATICS, geometric, and conservation scores (T, G, and C) and using the 3D-structure-based features (T and G) only, as measured on a 160 protein test set [8]. The figures of merit in Table 1 represent outstanding performance; see for example Table 1 of Sankararaman and

Table 1 POOL performance with and without sequence conservation data

Input features	T, G, and C	T and G only
POOL performance		
Specificity at 90% recall (%)	91	89
Specificity at 80% recall (%)	92	91
Specificity at 70% recall (%)	95	93
Recall at 95% specificity (%)	70	60
Recall at 90% specificity (%)	91	87
Recall at 80% specificity (%)	100	100

Input features: T = THEMATICS, G = geometry (cleft size rank), C = conservation scores. Performance data are for a test set of 160 annotated proteins [8]

Sjölander [28]. Table 1 shows that, even in the absence of sequence conservation information, POOL is able to make good predictions of catalytic residues with input features computed solely from the 3D structure of the query protein.

Applications

Prediction of protein functional residues is a first step toward functional annotation of a protein. One specific application has been to functional assignment within superfamilies, which consist of sets of proteins with similar 3D structure but often with significant functional diversity. Wei et al. have shown [49] that, for the small DJ-1 superfamily, placement of the predicted functional residues onto a 3D structural alignment reveals patterns characteristic of biochemical function; this enables one to sort the superfamily into subclasses according to their function.

Other applications of functional site prediction from electrostatic properties include better understanding ligand binding [50–52] and inhibitor design. These applications all require that the functional residues are predicted with both sensitivity and precision.

Precision

While many functional residue prediction methods boast high recall rates of annotated catalytic residues, often these also correspond to high false positive rates. In some cases, the measures of selectivity, such as precision, specificity, or false positive rates, are not reported at all [30]. Of course, the value of a high-recall prediction is significantly diminished if the corresponding false positive rate is high. While not all electrostatics based methods are capable of good selectivity, those that utilize titration curve shapes are able to return very low false positive rates with good recall.

The CSA-100, a non-redundant subset of 100 enzymes from the CSA, is often used for verification purposes [28]. This set of enzymes consists of a total of 36,230 residues, of which 314 are annotated as functionally important. Thus approximately

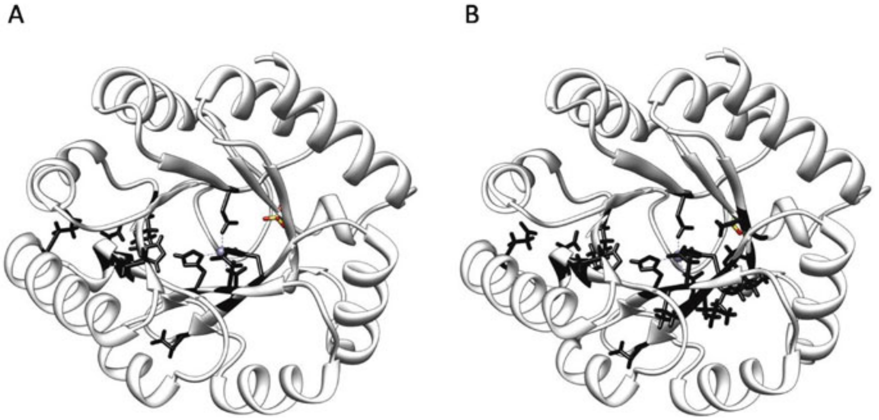

Fig. 1 Predictions for the structural genomics protein Pfal009167. (**a**) 5% POOL cut-off value; (**b**) 8% POOL cut-off value

1% of all residues currently are considered functionally important. While this represents a lower bound, as not all residues have been tested and thus some important residues are not listed in the CSA, it gives a rough idea of the fraction of total residues that should be returned by a site prediction method.

For application purposes, we have found that generally it is less important to predict every single functional residue than it is to predict most of the functional residues with few false positives. We have observed that the fraction of total residues predicted should be in the range of about 5–8% or less. Predictions that return higher fractions of residues are not particularly useful for application purposes, as the predicted region of the protein surface is too large.

This is illustrated in Figs. 1 and 2. Figure 1 depicts typical POOL predictions for the structural genomics protein Pfal00167 (PDB ID 1TQX) [53], a putative D-ribulose 5-phosphate 3-epimerase from *P. falciparum,* a member of the ribulose phosphate binding barrel superfamily [54]. The backbone is shown as a ribbon and the side chains of the predicted residues are shown as dark sticks. The prediction consisting of the top 5% of all residues is shown in Fig. 1a and that of the top 8% of all residues in Fig. 1b. The prediction of Fig. 1a is superimposable on the known active sites of previously characterized D-ribulose 5-phosphate 3-epimerases [54–55] and contains four known catalytic residues H36, D38, H70, and D179. Although this prediction misses one known active site residue, Q177, the similarity of the predicted site to the known binding sites of the well-studied structures with PDB IDs 1RPX [55] and 2FLI [54] is sufficient to confirm the putative functional annotation.

Figure 2 shows the POOL predictions for the same structural genomics protein if higher cut-off values are used. Figure 2a depicts the top 15% of all residues and 2b depicts the top 20% of all residues. These predictions constitute a large fraction of the protein surface area and are less useful.

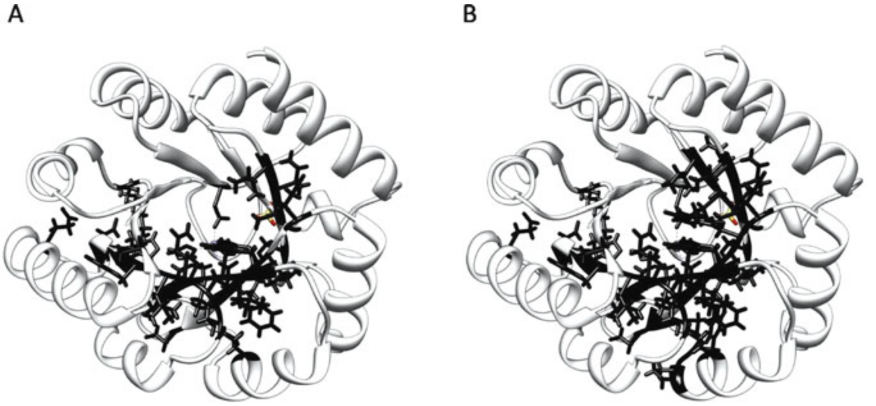

Fig. 2 Predictions for the structural genomics protein Pfal009167. (**a**) 15% POOL cut-off value; (**b**) 20% POOL cut-off value

Future Directions

While POOL in its present form shows excellent performance as a catalytic residue predictor, there are some additional features that could be built in to enhance its performance, including information about evolutionary history obtained from a phylogenetic tree [28, 56].

Acknowledgments This work was supported in part by the National Science Foundation under grant MCB-0843603 awarded to Mary Jo Ondrechen and an IGERT Traineeship and an NSF Graduate Research Fellowship awarded to Joslynn S. Lee.

References

1. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. The Protein Data Bank. Nucleic Acids Res. **28**(1): 235–242 (2000).
2. Westbrook, J., Feng, Z., Chen, L., Yang, H., Berman, H.M. The Protein Data Bank and structural genomics. Nucleic Acids Res. **31**: 489–491 (2003).
3. Bartlett, G.J., Porter, C.T., Borkakoti, N., Thornton, J.M. Analysis of catalytic residues in enzyme active sites. J. Mol. Biol. **324**: 105–121 (2002).
4. Porter, C.T., Bartlett, G.J., Thornton, J.M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.Nucleic Acids Res. **32**(Suppl 1): D129–133 (2004).
5. Ondrechen, M.J., Clifton, J.G., Ringe, D. THEMATICS: a simple computational predictor of enzyme function from structure. Proc. Natl. Acad. Sci. USA **98**: 12473–12478 (2001).
6. Ko, J., Murga, L.F., Andre, P., Yang, H., Ondrechen, M.J., Williams, R.J., Agunwamba, A., Budil, D.E. Statistical Criteria for the identification of protein active sites using theoretical microscopic titration curves. Proteins Struct. Funct. Bioinform. **59**: 183–195 (2005).
7. Wei, Y., Ko, J., Murga, L.F., Ondrechen, M.J. Selective prediction of interaction sites in protein structures with THEMATICS. BMC Bioinformatics **8**: 119 (2007).
8. Tong, W., Wei, Y., Murga, L.F., Ondrechen, M.J., Williams, R.J. Partial order optimum likelihood (POOL): maximum likelihood prediction of protein active site residues using 3D structure and sequence properties. PLoS Comput. Biol. **5**(1): e1000266 (2009).

9. Gilson, M.K. Multiple-site titration and molecular modeling: two rapid methods for computing energies and forces for ionizable groups in proteins. Proteins **15**(3): 266–282 (1993).
10. Gherardini, P.F., Helmer-Citterich, M. Structure-based function prediction: approaches and applications. Brief. Funct. Genomic. Proteomic. (2008).
11. Karp, P.D. What we do not know about sequence analysis and sequence databases. Bioinformatics **14**: 753–754 (1998).
12. Devos, D., Valencia, A. Practical limits of function prediction. Proteins Struct. Funct. Genet. **4**: 98–107 (2000).
13. Wilson, C.A., Kreychman, J., Gerstein, M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J. Mol. Biol. **297**: 233–249 (2000).
14. Landgraf, R., Xenarios, I., Eisenberg, D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J. Mol. Biol. **307**: 487–502 (2001).
15. de Rinaldis, M., Ausiello, G., Cesareni, G., Helmer-Citterich, M. Three-dimensional profiles: a new tool to identify protein surface similarities. J. Mol. Biol. **284**: 1211–1221 (1998).
16. Aloy, P., E.Querol, Aviles, F.X., Sternberg, M.J.E. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J. Mol. Biol. **311**: 395–408 (2001).
17. Ota, M., Kinoshita, K., Nishikawa, K. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. J. Mol. Biol. **327**: 1053–1064 (2003).
18. Gutteridge, A., Bartlett, G., Thornton, J.M. Using a neural network and spatial clustering to predict the location of active sites in enzymes. J. Mol. Biol. **330**: 719–734 (2003).
19. Innis, C.A., Anand, A.P., Sowdhamini, R. Prediction of functional sites in proteins using conserved functional group analysis. J. Mol. Biol. **337**: 1053–1068 (2004).
20. Carter, C.W., LeFebvre, B.C., Cammer, S.A., Tropsha, A., Edgell, M.H. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J. Mol. Biol. **311**(4): 625–638 (2001).
21. Meng, E.C., Polacco, B.J., Babbitt, P.C. Superfamily active site templates. Proteins **55**: 962–976 (2004).
22. Pazos, F., Sternberg, M.J.E. Automated prediction of protein function and detection of functional sites from structure. Proc. Natl. Acad. Sci. USA **101**: 14754–14759 (2004).
23. Cheng, G., Qian, B., Samudrala, R., Baker, D. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family. Nucleic Acids Res. **33**(18): 5861–5867 (2005).
24. Petrova, N., Wu, C. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. BMC Bioinformatics **7**(1): 312 (2006).
25. Youn, E., Peters, B., Radivojac, P., Mooney, S.D. Evaluation of features for catalytic residue prediction in novel folds. Protein Sci. **16**: 216–226 (2007).
26. Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., Ben-Tal, N. ConSurf: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res. **33**(Web Server issue): W299–302 (2005).
27. Innis, C. siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins. Nucleic Acids Res. **35**: W489–W494 (2007).
28. Sankararaman, S., Sjolander, K. INTREPID: INformation-theoretic TREe traversal for protein functional site identification. Bioinformatics **24**: 2445–2452 (2008).
29. Tang, Y.-R., Sheng, Z.-Y., Chen, Y.-Z., Zhang, Z. An improved prediction of catalytic residues in enzyme structures. Protein Eng. Des. Sel. **21**: 295–302 (2008).
30. Bray, T., Chan, P., Bougouffa, S., Greaves, R., Doig, A., Warwicker, J. SitesIdentify: a protein functional site prediction tool. BMC Bioinformatics **10**:379 (2009).
31. Sankararaman, S., Sha, F., Kirsch, J., Jordan, M., K. Sjölander. Active site prediction using evolutionary and structural information. Bioinformatics **26**(5): 617–624 (2010).

32. Wilkins, A., Lua, R., Erdin, S., Ward, R., Lichtarge, O. Sequence and structure continuity of evolutionary importance improves protein functional site discovery and annotation. Protein Sci. **19**: 1296–1311 (2010).
33. Elcock, A.H. Prediction of functionally important residues based solely on the computed energetics of protein structure. J. Mol. Biol. **312**: 885–896 (2001).
34. Bate, P., Warwicker, J. Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. J. Mol. Biol. **340**: 263–276 (2004).
35. Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I., Pietrokovski, S. Network analysis of protein structures identifies functional residues. J. Mol. Biol. **344**: 1135–1146 (2004).
36. Xie, L., Bourne, P.E. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. BMC Bioinformatics **8**: s4–s9 (2007).
37. Ming, D., Cohn, J.D., Wall, M.E. Fast dynamics perturbation analysis for prediction of protein functional sites. BMC Struct. Biol. **8**(5) (2008).
38. Mattos, C., Ringe, D. Locating and characterizing binding sites on proteins. Nat. Biotechnol. **14**(5): 595–599 (1996).
39. Silberstein, M., Dennis, S., Brown, L., Kortvelyesi, T., Clodfelter, K., Vajda, S. Identification of substrate binding sites in enzymes by computational solvent mapping. J. Mol. Biol. **332**: 1095–1113 (2003).
40. Laurie, A.T.R., Jackson, R.M. Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites. Bioinformatics **21**: 1908–1916 (2005).
41. Di Cera, E., Gill, S.J., Wyman, J. Binding capacity: cooperativity and buffering in biopolymers. Proc. Natl. Acad. Sci. USA **85**: 449–452 (1988).
42. Di Cera, E., Chen, Z.-Q. The binding capacity is a probability density function. Biophys. J. **65**: 164–170 (1993).
43. Laskowski, R.A. SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. J. Mol. Graph. **13**: 323–330 (1995).
44. Laskowski, R.A., Luscombe, N.M., Swindells, M.B., Thornton, J.M. Protein clefts in molecular recognition and function. Protein Sci. **5**: 2438–2452 (1996).
45. Liang, J., Edelsbrunner, H., Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. **7**: 1884–1897 (1998).
46. Wei, Y. Computed electrostatic properties of protein 3D structure for functional annotation and biomedical application. Boston: Ph.D. Dissertation, Northeastern University, p. 236 (2007).
47. Shehadi, I.A., Yang, H., Ondrechen, M.J. Future directions in protein function prediction. Mol. Biol. Rep. **29**: 329–335 (2002).
48. Tong, W., Williams, R.J., Wei, Y., Murga, L.F., Ko, J., Ondrechen, M.J. Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. Protein Sci. **17**: 333–341 (2008).
49. Wei, Y., Ringe, D., Wilson, M.A., Ondrechen, M.J. Identification of functional subclasses in the DJ-1 superfamily proteins. PLoS Comput. Biol. **3**(e10): 120–126 (2007).
50. Chan, C.S., Winstone, T.M., Chang, L., Stevens, C.M., Workentine, M.L., Li, H., Wei, Y., Ondrechen, M.J., Paetzel, M., Turner, R.J. Identification of residues in DmsD for twin-arginine leader peptide binding, defined through random and bioinformatics-directed mutagenesis. Biochemistry **47**(9): 2749–2759 (2008).
51. Murga, L.F., Ondrechen, M.J., Ringe, D. Prediction of interaction sites from Apo 3D structures when the holo conformation is different. Proteins **72**(3): 980–992 (2008).
52. Relloso, M., Cheng, T.Y., Im, J.S., Parisini, E., Roura-Mir, C., DeBono, C., Zajonc, D.M., Murga, L.F., Ondrechen, M.J., Wilson, I.A., et al. pH-dependent interdomain tethers of CD1b regulate its antigen capture. Immunity **28**(6): 774–786 (2008).

53. Caruthers, J., Bosch, J., Buckner, F., Voorhis, W.V., Myler, P., Worthey, E., Mehlin, C., Boni, E., DeTitta, G., Luft, J., et al. Structure of a ribulose 5-phosphate 3-epimerase from Plasmodium falciparum. Proteins Struct. Funct. Bioinform. **62**(2): 338–342 (2006).
54. Akana, J., Fedorov, A.A., Fedorov, E., Novak, W.R.P., Babbitt, P.C., Almo, S.C., Gerlt, J.A. d-Ribulose 5-Phosphate 3-Epimerase: functional and structural relationships to members of the ribulose-phosphate binding (β/α)8-barrel superfamily. Biochemistry **45**(8): 2493–2503 (2006).
55. Kopp, J., Kopriva, S., K.-H. Süss, Schulz, G.E. Structure and mechanism of the amphibolic enzyme -ribulose-5-phosphate 3-epimerase from potato chloroplasts. J. Mol. Biol. **287**(4): 761–771 (1999).
56. Lichtarge, O., Bourne, H.R., Cohen, F.E. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. **257**(2): 342–358 (1996).