COMPUTED PROTONATION PROPERTIES: UNIQUE CAPABILITIES FOR PROTEIN FUNCTIONAL SITE PREDICTION

LEONEL F. MURGA[1,2] YING WEI[1]
leonel@brandeis.edu wei.y@neu.edu

MARY JO ONDRECHEN[1]
mjo@neu.edu

[1] *Department of Chemistry & Chemical Biology and Institute for Complex Scientific Software, Northeastern University, Boston, MA 02115 USA*
[2] *Present Address: Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454 USA*

Prediction of protein functional sites from 3D structure is an important problem, particularly as structural genomics projects produce hundreds of structures of unknown function, including novel folds and the structures of orphan sequences. The present paper shows how computed protonation properties provide unique and powerful capabilities for the prediction of catalytic sites from the 3D structure alone. These protonation properties of the ionizable residues in a protein may be computed from the 3D structure using the calculated electrical potential function. In particular, the shapes of the theoretical microscopic titration curves (THEMATICS) enable selection of the residues involved in catalysis or small molecule recognition with good sensitivity and precision. Results are shown for 169 annotated enzymes in the Catalytic Site Atlas (CSA). Performance, as measured by residue recall and precision, is clearly better than that of other 3D-structure-based methods. When compared with methods based on sequence alignments and structural comparisons, THEMATICS performance is competitive for well-characterized enzymes. However THEMATICS performance does not degrade in the absence of similarity, as do the alignment-based methods, even if there are few or no sequence homologues or few or no proteins of similar structure. It is further shown that the protonation properties perform well on open, unbound structures, even if there is substantial conformational change upon ligand binding.

Keywords: function prediction; site prediction; structural genomics; THEMATICS; electrostatics

1. Introduction

Structural genomics efforts are revealing the 3D structures of hundreds of proteins of unknown function, including the structures of orphan sequences and structures with novel features or novel folds. A recent search of the Protein Data Bank (PDB) [4] returned 3108 structures listed as “hypothetical” or “unknown function” or “putative.” This number is constantly increasing at a rapid pace as techniques for high-throughput expression and crystallization are further developed and refined. However, knowledge of the three-dimensional structure does not necessarily imply knowledge of function. In fact, the inference of functional information from the 3D structure has proved to be far more difficult than anticipated. Computational methodologies for the prediction of functional information about these proteins are therefore important and timely.

An additional challenge posed by structural genomics proteins is that nearly all of them are unbound structures, since their natural substrates and other ligands are generally not present or even known. Ligand binding is usually accompanied by some structural change. While these binding-induced structural changes are often small and involve only rotation of the side chains of adjacent residues, significant change in backbone conformation does occur for some proteins. Thus it is desirable to develop predictive structure-to-function methods that are successful for these unbound, apo structures.

Many methods have been reported to date for the prediction of functional sites in proteins. Some of these methods utilize only the 3D structure of the query protein as input [1, 3, 12, 13, 16], and thus are applicable to proteins with few or no sequence homologues, while other methods require both the 3D structure and a sequence alignment [6, 15, 17, 18, 23]. These methods report high success rates in functional site prediction, although often the precision is low, with only a small fraction of the selected residues corresponding to annotated functional residues. Achieving high recall with good precision is a challenge but yields more useful results.

The reliability of computed protonation properties for the identification of interaction sites in protein 3D structures was first reported in 2001 [16]. The method was named THEMATICS (Theoretical Microscopic Titration Curves) and has since been automated using statistical metrics [12, 21]. In the present paper it is shown how computed protonation properties can predict the interaction sites in open, unbound structures for systems undergoing a large conformational change upon ligand binding. It is also shown that these properties alone return low false positive rates, in addition to high site success rates and good residue recall, for the enzymes in an annotated dataset. Examples of the precise, highly localized, predictions are given, including cases with large apo-holo conformational change.

2. Method

Protein structure coordinate files are downloaded from the PDB. The 3D structure files are pre-processed as described by Wei [21]. From the atomic coordinates, the electrical potential function of the protein is calculated with a Finite Difference Poisson–Boltzmann (FDPB) procedure. The FDPB component of the University of Houston Brownian Dynamics (UHBD) program [14] is used for this purpose. The theoretical titration curves are calculated for each ionizable species in the protein (each Arg, Asp, Cys, Glu, His, Lys, and Tyr, plus the N- and C- termini) using a hybrid procedure [10]. Each of these chemical groups in a protein structure can gain or lose an H^+ ion, a proton. For each of these ionizable groups, the hybrid procedure computes the fraction of all the protein molecules in a large ensemble that have this group occupied by a proton at a given pH. These data are expressed as the proton occupation O as a function of the pH. These O(pH) curves constitute theoretical analogues of an experimental titration curve. We have argued that catalytic residues have special properties in their proton transfer chemistry that can be observed in the shapes of the computed O(pH) curves [16].

An extension of Ko's analysis [12] is used to select the residues that are most likely to be involved in catalysis and/or recognition. We define the first derivative function f of the O(pH) curve as:

$$f = -\,dO/d(pH) \tag{1}$$

These f functions are essentially proton binding capacities [8, 9, 22], which measure the change in concentration of a bound proton per unit change in its chemical potential. Note that these f functions are automatically normalized so that the area under the f curve is unity. This is because O always runs from 1 to 0, so that $-\Delta O$ is always unity over the full range of pH values; this is true for both perturbed and normal ionizable residues.

The f functions may be treated as distributions and characterized by their moments [8, 12]. Hence we define the nth central moment μ_n as:

$$\mu_n = \int (pH - M_1)^n \cdot f \cdot d(pH) \tag{2}$$

where M_1 is the first raw moment, defined by the expression for the nth raw moment as:

$$M_n = \int (pH)^n \cdot f \cdot d(pH) \tag{3}$$

Integrals in (2) and (3) are over all space ($-\infty$ to $+\infty$). Equations 1-3 for each ionizable residue are evaluated numerically from the computed theoretical titration curve of each residue.

If there were only one ionizable residue in the protein, that residue would obey the Henderson-Hasselbalch (H-H) equation. For such a residue that obeys the H-H equation, the first raw moment is the pK_a and all the odd-numbered central moments are zero. The second and fourth central moments have the values 0.620 and 1.62, respectively, for an H-H acid or base. However, interactions between ionizable residues in a protein will lead to asymmetry and broadening of the f functions and thus the odd-numbered central moments will be non-zero and the even-numbered moments will be larger. The underlying premise of THEMATICS is that these interactions are strongest for the active site residues and therefore the active site residues are identifiable as those with the most deviant curve shapes and particularly the largest third and fourth moments [12].

We define the Z score for the nth central moment as:

$$Z_n = (|\mu_n| - <|\mu_n|>)/\sigma_n \tag{4}$$

Here $<|\mu_n|>$ is the mean of the absolute value of the nth central moment, averaged over all of the ionizable residues in the protein. Generally only the odd-numbered moments can be negative, so the absolute value sign really is needed only for the odd moments. σ_n is the standard deviation of the (absolute) values of the nth central moment for the set of all ionizable residues in the protein. The Z score represents the deviation from the mean value in units of the standard deviation. The f functions are peaked functions for ordinary residues, but the active site residues deviate the most from the sharply peaked H-H form. The central moments are natural metrics to characterize the width and the shape of these peaked functions and their Z scores provide a way to identify the most deviant curves.

The mean and the standard deviation in Equation (4) are obtained using a specified cut-off fraction of the complete set of ionizable residues in the protein of interest. In particular, the set of ionizable residues in the protein is rank-ordered according to the nth central moment, then the only the lowest, specified fraction is used to obtain the mean and the standard deviation. Thus extremely large values for the central moments, which can have large impact on the mean and standard deviation, are excluded for this part of the calculation. For instance, if the cut-off fraction is 0.98, all of the central moment values that are above the 98th percentile are excluded when the mean and standard deviation are computed. A cut-off fraction of 1.0 is therefore equivalent to Ko's analysis [12]. Z scores are computed for all of the ionizable residues, but the mean and the standard deviation are computed using the smaller, cut-off population.

The criterion $Z_3>1$ or $Z_4>1$ is used to select the positive residues, since most active site residues have either a large third central moment or a large fourth central moment.

Once the THEMATICS positive residues are identified, these residues are grouped into clusters based on spatial proximity. The distance between two positive residues is defined as the distance between their charge centers. A residue is placed into a cluster if it is within 9Å of any other positive residue in the cluster. A one-member cluster is called an isolated positive and it is not considered predictive. Clusters containing two or more positive residues constitute predictions and are termed THEMATICS predictive clusters.

3. Results

3.1 *Performance in Site Prediction*

The method was tested on 169 annotated enzymes, essentially the entire CatRes database, with updated annotations from the Catalytic Site Atlas (CSA) database [2, 19]. Performance is measured by the recall (fraction of residues annotated in the database as catalytically important that are identified by the method), precision (fraction of identified residues that are annotated in the database as catalytically important), and false positive (FP) rate. We note that precision rates should be considered as lower bounds, because the database annotations are incomplete. In other words, not all of the important residues are annotated as such in the database. Recall, precision, FP rate, and Matthews Correlation Coefficient (MCC) for 169 CatRes enzymes, computed using CatRes/CSA annotations only, are shown in Table 1 for different values for the cut-off fraction. Recall rates for residues annotated as catalytically important range from 41% to 63%. Nominal precision rates range from 19% to 14%; this represents the fraction of predicted residues that are annotated in the database as catalytically important. FP rates are low, ranging from 2.0% to 4.7%, depending on the cut-off. Note that the lower cut-off values result in more residues selected, so as the cut-off value is reduced the residue recall rate rises but at some expense in precision.

While Table 1 presents performance in the prediction of catalytic residues, the success rate for the prediction of catalytic sites is higher. Table 2 shows the percentages of the 169 test proteins for which correct or partially correct predictions are made by THEMATICS. Success in site prediction is defined according to designations used in previous work [11]. A site prediction is considered *correct* if it includes half or more of the annotated catalytic residues. A prediction is considered *partially correct* if it contains at least one, but less than half, of the annotated catalytic residues. The total success rate for the prediction of sites is the sum of the correct plus partially correct predictions. In columns 2 through 4 of Table 2, the correct and partially correct predictions are determined with the CatRes/CSA annotations only. In column 5, an expanded set of annotations, including information from the SITE fields of the PDB files and from protein-specific literature references, is used. For cases where a bound structure is available, the residues in direct contact with the bound ligand(s) were added to the list. The CatRes/CSA annotations thus constitute a subset of this expanded list. All methods, including THEMATICS, perform better against this expanded list.

Table 1. Average recall, precision, false positive rate, and MCC for THEMATICS predictions of catalytic residues as functions of the cut-off fraction for the test set of 169 enzymes. Here only the CSA annotations are used as the reference set.

Cut-off	**Recall**	**Precision**	**FP rate**	**MCC**
1.00	41.1%	19.4%	2.07%	0.255
0.99	50.4%	17.9%	2.76%	0.272
0.98	54.2%	16.4%	3.24%	0.270
0.97	58.0%	15.5%	3.74%	0.270
0.96	61.0%	14.6%	4.23%	0.268
0.95	62.8%	13.6%	4.72%	0.262

Table 2. THEMATICS success rates for site prediction for the 169 enzymes in the test set. Success rate is expressed as correct, partially correct, and total, using a cut-off of 1.00, 0.99 and 0.98. In columns 2-4, only the CSA annotations are used as the reference set. In column 5, the total success rate is obtained using an expanded set of annotations.

Cut-off	**Correct Site Rate versus CSA**	**Partially Correct Site Rate versus CSA**	**Total Success Rate versus CSA**	**Total Success Rate versus Expanded Set**
1.00	48.5%	29.0%	77.5%	89.9%
0.99	59.8%	26.0%	85.8%	92.9%
0.98	66.9%	21.3%	88.2%	94.1%

For cut-off values of 1.0, 0.99, and 0.98, THEMATICS makes correct site predictions for 49%, 60%, and 64%, respectively, of the proteins in the CatRes test set, according the CSA annotations only. Total success rates, the sum of the correct and partially correct rates for site prediction, are 90%, 93%, and 94%, for the same three respective cut-off values, according to the expanded annotation set.

These performance data shown in Tables 1 and 2 compare very favorably with other methods that are based solely on the 3D structure of the query protein. Structural Analysis of Residue Interaction Graphs (SARIG) [1] is a graph theoretic approach that calculates residue contacts and identifies the residues that have the highest closeness scores to all other residues. SARIG successfully predicts 46.5% of the annotated catalytic residues for the enzymes in the CatRes database. The reported precision, however, is low; only 9.4% of the predicted residues are known catalytic residues [1]. Thus compared to THEMATICS with a cut-off of 0.99, the residue recall rates are similar but the THEMATICS precision rate is about two-fold better.

Another approach to the prediction of sites from the structure alone involves docking of small solvent molecules onto the protein surface and searching for clusters of energy minima for these molecules [7, 20]. Q-SiteFinder is a simple, fast version of this method developed by Laurie [13] and uses only a methyl group as a probe. For 90% of proteins in the test set, Q-SiteFinder returns a correct site prediction within its top three predicted sites. While precision was not reported, selectivity clearly is low. We estimate that the precision rate for residue prediction by Q-SiteFinder is only about 5%, corresponding to an average of about 60 predicted residues per protein; these estimates are based on a combination of the top three sites as the prediction, which was the basis for the reported success rates [13]. Thus THEMATICS gives comparable success rates in site prediction but with substantially better precision and lower false positive rates.

THEMATICS also performs quite well compared with methods that require sequence alignments and structural comparisons. While there are variations in the annotated sets, one can get some idea of the relative performance. One method based on sequence and structural alignments reports a catalytic residue recall rate of 47% with an FP rate of 5% [17]; THEMATICS with a cut-off of 0.99 has a similar recall rate (53%) but the FP rate is lower by almost one half (2.8%). Another study using Support Vector Machines (SVM) to predict catalytic residues from sequence conservation and structural properties reports an MCC of 0.23 [18], slightly less than the MCC for THEMATICS (0.27). Another very recent paper [23], also using SVM with features obtained from sequence alignments and structural properties, reports 57.0% catalytic residue recall with 18.5% precision, slightly better than THEMATICS; however these data were obtained on a test set of enzymes that possess sequence homologues and structural family members. As the authors point out, there will be a cost in the recall and in the precision when applied to novel folds or remote sequences. THEMATICS performance is roughly the same as that of the SVM-based sequence/structure methods, but without any sequence or structural alignments needed.

3.2 *Predicting Holo Binding Sites from Apo Structures*

For proteins that undergo a small conformational change upon ligand binding, it has been shown already that THEMATICS performs equally well for apo (unbound) structures as it does for holo (bound) structures [21]. In this section we examine pairs of structures that exhibit significant change in backbone conformation going from the apo to the holo form. For such a pair of structures, the Root Mean Square Deviation (RMSD) of the alpha carbon framework may be calculated using the expression:

$$RMSD = \sqrt{\frac{\sum_{i}^{N} d_i^2}{N}} \tag{6}$$

Where d_i represents the distance between the alpha carbon atoms of equivalent residues in the apo and holo forms and N is the number of residues in both structures.

The RMSD itself is not a good measure of the relative change in conformation when one compares pairs of proteins with a wide range of sizes because it depends on the number of residues. To obtain a better sense of the degree of conformation change within the set of proteins studied, we also include the RMSD100 [5] value defined as:

$$RMSD100 = \frac{RMSD}{1 + \ln\left(\sqrt{\frac{N}{100}}\right)} \tag{7}$$

Where RMSD is given in equation (6). RMSD100 is a size-normalized RMSD that allows a better comparison across sets with large variations in size.

To assess how these changes affect the region specifically around the binding site, the RMSD and RMSD100 values for the alpha carbon atoms of "core residues" are reported. These are defined as the set of residues with any atom located within 8 Å of any bound ligand in the holo form. Comparison of the RMSD values for the whole protein and for core residues provides a quantitative description of the effect of the conformational change on the binding site. Ten illustrative examples are given in Table 3.

These are examples from a set of 24 proteins for which apo and holo structures are available for the same species and for which the apo-holo RMSD is 1.5A or greater. RMSD values range from 1.5 to 14.7 in the full set and from 3.7 to 14.7 in the Table 3 examples. The first column gives the species and protein name. The number of residues N in the protein is given in the second column, with the number of residues in the active site core, as defined above, in parentheses. The next two columns give the RMSD and RMSD100 values for the apo-holo pair, with the corresponding values for the core residues shown in parentheses.

THEMATICS predictions for the apo and holo structures for the Table 3 examples are given in Table 4. In Table 4, the PDB code and the THEMATICS predictions are given for the apo form first for each protein, then for the holo form. Residues that belong to the same cluster are shown together in square brackets. Clusters of two or more residues constitute predictions. Clusters with only one member are reported but are not considered predictive. Known catalytic residues and residues in direct contact with the bound ligand of the holo form are shown in **boldface**.

Table 4 demonstrates how THEMATICS is able to predict binding sites in the unbound apo structures. In some cases, for example yeast guanylate kinase, *E. coli* dipeptide binding protein, and human serum transferrin, the binding residues are divided into different, spatially separated clusters in the apo structure. Upon ligand binding, these clusters come together to form a single cluster in the holo form; these predicted residues surround the bound ligand in the holo structure.

Table 3. Ten proteins with significant conformational change upon ligand binding. Number of residues N in the protein and the apo-holo RMSD and RMSD100 are given. Corresponding values for the active site core residues are shown in parentheses.

Species and Protein Name	**N** **(N Core)**	**RMSD** **(Core)**	**RMSD100** **(Core)**
Yeast Guanylate Kinase	186 (45)	4.4 (3.4)	5.5 (9.2)
E. coli Dipeptide Binding Protein	507 (51)	12.3 (10.3)	6.8 (15.6)
Human Serum Transferrin	328 (33)	14.7 (6.7)	9.3 (15.1)
Human Glucokinase	424 (58)	10.9 (9.5)	6.3 (13.0)
Human Lactoferrin	691 (73)	8.2 (5.8)	4.2 (6.9)
E. coli L-Leucine Binding Protein	345 (51)	14.4 (9.3)	8.9 (14.1)
E. coli D-Ribose Binding Protein	271 (47)	8.7 (5.6)	5.6 (9.0)
Emericella nidulans 3-Dehydroquinate Synthase	762 (189)	3.7 (2.3)	1.8 (1.7)
E. coli Ribokinase	610 (115)	9.8 (6.4)	5.1 (5.9)
Limulus polyphemus Arginine Kinase	344 (71)	3.8 (3.1)	2.4 (3.7)

Figure 1 shows the active site region of the aligned apo and holo structures of human serum transferrin with the THEMATICS predictions for both structures. Transferrin is the generic label for a group of proteins found in a wide range of organisms. Their main function is the sequestration of iron either for storage or for transport to the cell interior. The backbones are shown as ribbons and the THEMATICS predicted residues are shown explicitly in stick form. The apo form is in light gray, the holo form dark. Upon binding of iron, there is a large displacement of the binding residue H249. The two clusters predicted by THEMATICS for the unbound apo structure, [E83, **H249**] and [**Y95**, **Y188**, **K206**] come together to form a single cluster, [Y85, **Y95**, **Y188**, **K206**, **H249**, D292, **K296**], in the holo structure. Note that the predicted residues surround the bound iron in the holo form, and that THEMATICS is able to identify them even in the apo form.

In the full set of 24 pairs, THEMATICS predicts correct sites for 23 of them (96%). For one case, threonine synthase from *Thermus thermophyllus*, the correct site is predicted for the holo form but not for the apo form. Somewhat surprisingly, this particular case does not have one of the larger conformational changes (RMSD=2.6;

RMSD100=1.6). Three annotated catalytic residues, [**K61**, **K116**, **R160**], fall above the cut-off for the holo form but fall short of the cut-off for the apo form.

Table 4. THEMATICS predictions for apo and holo structures of the 10 proteins listed in Table 3. The PDB codes for the apo and holo forms are given in the second column, with the apo form given first. Predictions are given in clusters, with members of the same cluster shown together in square brackets. The known binding residues are shown in **boldface**. Clusters of two or more residues constitute predictions.

Species and Protein Name	**PDB ID** **Apo** Holo	**THEMATICS Predicted Clusters**
Yeast Guanylate Kinase	1EX6	[Y25, Y175, K179] [**Y50**, **Y78**] [**E69**, **D100**, **D98**, E153, H162, Y156] [C95] [R135] [D170]
	1EX7	[**K14**, **Y50**, **D98**, E153, **R38**, **R41**, **Y78**, **D100**, H162, **R146**] [Y175, K179]
E. coli Dipeptide Binding Protein	1DPE	[D26, E38, D153, E482, D494, **K498**, H499, H500] [D89] [E143, D149, **D408**, D411, D413] [Y269] [**Y357**, **Y431**]
	1DPP	[Y25, D26, **Y114**, D153, Y239, **R355**, **Y357**, **D408**, H499] [D149, D413]
Human Serum Transferrin	1BP5	[E83 , **H249**] [**Y95** , **Y188** , **K206**]
	1A8E	[Y85 , **Y95** , **Y188** , **K206** , **H249** , D292 , **K296**]
Human Glucokinase	1V4T	[E40A] [**C213A** , **C220A** , **C233A** , **C252A**] [**C230A** , **C382A**] [D274A] [E339A]
	1V4S	[**C213A** , **C220A** , **C233A** , **C252A**] [**C230A** , **C382A**]
Human Lactoferrin	1CB6	[**Y92**, **Y192**, R210] [Y93, R249] [E413 , **Y435** , **Y528** , K546, **H597**]
	1LFG	[Y82 , **Y92** , **Y192** , R210 , **H253**] [Y398] [**Y435** , **Y528** , **H597**]
E. coli L-Leucine Binding Protein	1USG	[E22 , D51 , H76 , Y89] [Y72] [E90 , H113] [D121, **Y150** , Y198 , **E226**] [H145 , D146 , E152 , E205] [K250 , Y252 , D330] [D321 , D325]
	1USK	[E22, D51, H76] [Y72] [E90, H113] [D121, H145, D146, **Y150**, E152, E205, **E226**] [D321, D325]
E. coli D-Ribose Binding Protein	1URP	[E26] [**D89**, H100, **D215**, E140, D191, E192]
	2DRI	[**D89**, **D215,** E140, D191, E192]
Emericella nidulans 3-Dehydroquinate Synthase	1NUA	[Y25, C34, K172]* [**K84A**, K161A, **E194A**, **K197A**, **K250A**, **H271A**, E278A, **H287A**, **R130B**]* [K89A, K89B] [Y134A]* [D176A] [E181B]
	1NVF	[K89A] [**R130A**, **K152B**, **K250B**, **H271B**, **H275B**, E278B, **H287B**, **K356B**] [**K152A**, **K250A**, **K356A**, **R130B**, **H271A**, E278 , **H287A**]
E. coli Ribokinase	1RKA	[**D16A**, H17A, **D67A**, H113A, **D16B**, H17B, **D67B**, H113B] [**E143A** , **D255A**]*
	1RKS	[**D16A**, H17A, **E143A**, **D255A**, **D16B**, H17B, **E143B**, **D255B**] [E188A]* [**E190A**]*
Limulus polyphemus Arginine Kinase	1M80	[**Y68**, **R126**, **C127**, Y134, Y145, K151, R208, E224, **E225**, D226, H227, **R229**, **C271**, **R280**, R330] [D71]
	1BG0	[Y89, **R126**, **C127**, R208, E224, **E225**, D226, **R229**, **R28**, **R309**, **E314**, H315, R330, E335] [H185]

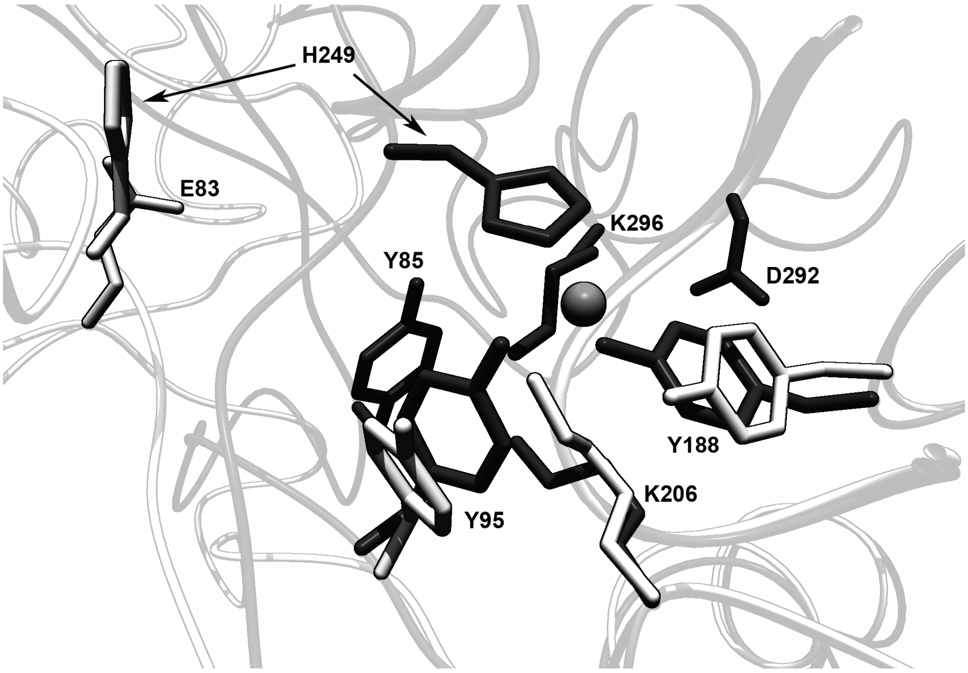

Figure 1. Detail of the active site of human serum transferrin showing the rearrangement that occurs upon binding of iron (unlabeled sphere). The apo form (PDB 1BP5) is shown in light gray; the holo form (PDB 1A8E) is in dark gray. Side chains of the THEMATICS predictions for the unbound apo form (in white), [E83, **H249**] and [**Y95**, **Y188**, **K206**] and for the bound holo form (in black), [Y85, **Y95**, **Y188**, **K206**, **H249**, D292, **K296**] are shown in stick form. Note that the two clusters in the apo structure become a single cluster in the holo form, reflecting the large displacement of H249. Figure prepared with Yasara.

4. Discussion and Conclusions

THEMATICS requires only the 3D structure of the query protein as input and therefore its performance is not affected by the degree of sequence or structural similarity to other proteins. THEMATICS performs significantly better than other 3D-structure-based methods. THEMATICS, using only the protonation properties computed from the structure, also performs quite well against the latest SVM-based methods that do utilize sequence alignments and structural similarity; THEMATICS returns competitive recall and precision with no cost for lack of similarity. THEMATICS thus holds advantages and is effective for novel folds, for engineered structures, and for proteins that do not possess a sufficient set of homologues to obtain meaningful conservation scores.

THEMATICS predicts the binding residues in apo structures, even when there is substantial change in the backbone conformation upon ligand binding and even when these binding residues are physically separated in the apo form. THEMATICS is effective because it utilizes the special electrostatic properties of active site residues. At these sites, there tends to be large interaction between protonation events on the ionizable species. These special electrostatic properties of the catalytic and binding residues are sufficiently preserved in the open form, although diminished compared to the closed form, such that the residues that constitute the two halves of the interaction site may be identified by statistical analysis in an open, unbound structure.

One of the difficulties in catalytic site prediction is that not all of the important residues are included in the annotated database and some have not been studied. Hence some of the "false positives" really are not false. Actual precision and MCC values for each method therefore are really higher and actual false positive rates are really lower than reported. The true quality and effectiveness of THEMATICS and the other top-performing methods is better than that reflected in the precision and MCC values; the relative performance of the different methods is also apparent from these values.

While computed protonation properties give excellent performance when used alone, they show tremendous potential in combination with other, complementary methods. A beta version of a simplified form of the THEMATICS site predictor is now available at: http://pfweb.chem.neu.edu/thematics/submit.html .

Acknowledgments

This work is supported by the National Science Foundation grant MCB-0517292.

References

[1] Amitai, G., A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski, Network analysis of protein structures identifies functional residues, *J Mol Biol*, 344:1135-1146, 2004.

[2] Bartlett, G.J., C.T. Porter, N. Borkakoti, and J.M. Thornton, Analysis of Catalytic Residues in Enzyme Active Sites, *J Mol Biol*, 324:105-121, 2002.

[3] Ben-Shimon, A. and Eisenstein, M., Looking at Enzymes from the Inside out: The Proximity of Catalytic Residues to the Molecular Centroid can be used for Detection of Active Sites and Enzyme-Ligand Interfaces, *Journal of Molecular Biology* 351:309-326, 2005.

[4] 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*, 28:235-242, 2000.

[5] Carugo, O., and S. Pongor, A normalized root-mean-square distance for comparing protein three-dimensional structures, *Protein Sci*, 10:1470-1473, 2001.

[6] Cheng, G., B. Qian, R. Samudrala, and D. Baker, Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design, *Nucleic Acids Res*, 33:5861-5867, 2005.

[7] Clodfelter, K.H., Waxman, D.J. and Vajda, S., Computational Solvent Mapping Reveals the Importance of Local Conformational Changes for Broad Substrate Specificity in Mammalian Cytochromes P450, *Biochemistry*, 45:9393-9407, 2006.

[8] Di Cera, E., and Z.-Q. Chen, The Binding capacity is a probability density function, *Biophys J*, 65:164-170, 1993.

[9] Di Cera, E., S.J. Gill, and J. Wyman, Binding Capacity: Cooperativity and buffering in biopolymers, *Proc Natl Acad Sci U S A*, 85:449-452, 1988.

[10] Gilson, M.K., Multiple-site titration and molecular modeling: two rapid methods for computing energies and forces for ionizable groups in proteins, *Proteins*, 15:266-282, 1993.

[11] Gutteridge, A., G. Bartlett, and J.M. Thornton, Using a neural network and spatial clustering to predict the location of active sites in enzymes, *Journal of Molecular Biology*, 330:719-734, 2003.

[12] Ko, J., L.F. Murga, P. Andre, H. Yang, M.J. Ondrechen, R.J. Williams, A. Agunwamba, and D.E. Budil, Statistical Criteria for the Identification of Protein Active Sites Using Theoretical Microscopic Titration Curves, *Proteins: Structure Function Bioinformatics*, 59:183-195, 2005.

[13] Laurie, A.T.R., and R.M. Jackson, Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites, *Bioinformatics*, 21:1908-1916, 2005.

[14] Madura, J.D., J.M. Briggs, R.C. Wade, M.E. Davis, B.A. Luty, A. Ilin, J. Antosiewicz, M.K. Gilson, B. Bagheri, L.R. Scott, & J.A. McCammon, Electrostatics and diffusion of molecules in solution - Simulations with the University of Houston Brownian Dynamics program, *Comp Phys Commun*, 91:57-95, 1995.

[15] Nimrod, G., F. Glaser, D. Steinberg, N. Ben-Tal, and T. Pupko, In silico identification of functional regions in proteins, *Bioinformatics*, 21 Suppl 1:i328-i337, 2005.

[16] Ondrechen, M.J., J.G. Clifton and D. Ringe, THEMATICS: A simple computational predictor of enzyme function from structure, *Proc. Natl. Acad. Sci. (USA)*, 98:12473-12478, 2001.

[17] Panchenko, A.R., Kondrashov, F. and Bryant, S., Prediction of functional sites by analysis of sequence and structure conservation, *Protein Sci*, 13:884-892, 2004.

[18] Petrova, N. and Wu, C., Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties, *BMC Bioinformatics*, 7:312, 2006.

[19] Porter, C.T., Bartlett, G.J. and Thornton, J.M., The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucl. Acids Res.*, 32:D129-133, 2004.

[20] Silberstein, M., S. Dennis, L. Brown, T. Kortvelyesi, K. Clodfelter, and S. Vajda, Identification of substrate binding sites in enzymes by computational solvent mapping, *J Mol Biol*, 332:1095-1113, 2003.

[21] Wei, Y., J. Ko., L.F. Murga, and M.J. Ondrechen, Selective prediction of interaction sites in protein structures with THEMATICS, *BMC Bioinformatics*, 8:119, 2007.

[22] Wyman, J., Linked functions and reciprocal effects in hemoglobin: A second look, *Adv Protein Chem*, 19:223-286, 1964.

[23] Youn, E., B. Peters, P. Radivojac, and S. D. Mooney, Evaluation of features for catalytic residue prediction in novel folds, *Protein Sci*, 16:216-226, 2007.