Application of THEMATICS to Non-Catalytic Ligand-Binding Proteins

Leonel F. Murga[1] Jaeju Ko[2] Mary Jo Ondrechen[1]

[1] *Department of Chemistry and Chemical Biology and Institute for Complex Scientific Software, Northeastern University, Boston, MA, USA*

[2] *Department of Chemistry, Indiana University of Pennsylvania, Indiana, PA, USA*

Email : leonel@neu.edu, chemko@iup.edu, mjo@neu.edu

ABSTRACT: THEMATICS is a simple computational method for predicting functional sites in proteins. The method computes the theoretical titration curves of the ionizable residues of a protein using its 3D structure, determines the residues with perturbed, non-Henderson- Hasselbalch titration behavior, and identifies clusters of these perturbed residues in physical proximity. We have shown previously that this method is highly successful in predicting catalytic sites in enzymes. In the present study, we apply the method to non-catalytic ligand-binding proteins. It is shown that THEMATICS can predict non-catalytic binding sites. The success rate is better than 80 % for a set of 30 non-catalytic, ligand-binding proteins. The application of the method to Glutamine-binding protein from *E. coli* is discussed in detail.

1 INTRODUCTION

Computational methods for detecting functional sites in proteins have become increasingly important, as structural genomics initiatives continue to determine the structures of "hypothetical" proteins whose functions are unknown. The Protein Data Bank (PDB) [1] currently contains on the order of 10^3 such structures. THEMATICS – Theoretical Microscopic Titration Curves – is a computational procedure for predicting interaction sites in proteins [2-5] and we have previously shown that the method is highly effective in predicting the active sites of enzymes [6, 7]. We now apply the same procedure to a set of thirty non-catalytic, ligand-binding proteins.

Predictive methods for searching for functional sites often rely on conserved residues [8-11] or on both sequence and structural information [12-17]. Some methods utilize computed electrostatic energies of the residues [18, 19] or consider structural and sequence information in addition to computed energies [20]. Some predictive methods search for binding pockets by certain geometric criteria [21-26]. A more recent method Q-SiteFinder [27] is based on computed interaction energy between a probe and the protein. The methods based on geometric criteria or computed energies require the three-dimensional structure of the protein as does THEMATICS.

A unique aspect of THEMATICS is that it extracts information from the shapes of the computed titration curves. A small fraction of the ionizable residues in a protein has previously been predicted to exhibit perturbed, non-Henderson-Hasselbalch (H-H) titration behavior [28-31]. We have shown that such residues with perturbed titration behavior tend to occur in the active site of an enzyme with high frequency, and with sufficiently low frequency elsewhere, that they serve as markers of chemical reactivity.

THEMATICS calculates the protonation state of ionizable residues in proteins as a function of pH (*i.e.*, titration curves) using a hybrid Monte Carlo procedure [32, 33] based on the electrostatic potential function obtained from solving linearized Poisson-Boltzmann (PB) equations [34, 35]. It then determines the perturbed residues that deviate the most from H-H titration behavior and identifies the clusters of two or more such residues in physical proximity. It will be shown here that THEMATICS is almost as effective in predicting functional sites for non-catalytic ligand-binding proteins as it is for enzymes.

2 METHODS

2.1 Theoretical Microscopic Titration Curves

A detailed procedure for computing theoretical microscopic titration curves has been given elsewhere [6, 7]. Briefly, the PB calculations were performed using the University of Houston Brownian Dynamics (UHBD) program [36] on each protein; the mean protonation state of each ionizable residue as a function of pH was then computed by the program HYBRID [33]. The protein structures were downloaded from the PDB. The H atoms were added to the structure using TINKER [37] and the OPLS-UA force field [38, 39].

2.2 Statistical Analysis of the Titration Curves

For most of the ionizable residues in a protein, the average protonation state θ as a function of pH can be expressed as:

$$\theta(\mathrm{pH}) = \left(10^{\mathrm{pH-pKa}} + 1\right)^{-1} \qquad (1)$$

This equation is the H-H equation and applies to residues that form a cation upon protonation (Arg, His, Lys, and the N-terminus) and to residues that form anions upon deprotonation (Asp, Cys, Glu, Tyr, and the C-terminus). Figure 1 shows some examples of titration curves for Phospholipase C (PDB code: 1AH7) from *Bacillus cereus* [40]; H142 in Figure 1 exhibits typical, H-H titration behavior. For such a residue, the predicted average protonation state falls sharply in a pH range close to the pK_a.

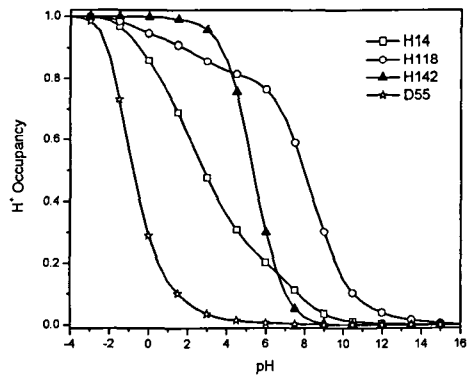

Figure 1: Theoretical titration curves, plotted as ensemble average proton occupation θ as a function of pH, for selected residues in Phospholipase C. The known active-site residues are H14, H118, and D55.

A protein is essentially a polyprotic acid and thus all of its ionizable groups deviate from the H-H equation at least to some degree. However, for most of these groups the deviations are small and the shape of their titration curves follows Eq. 1 very closely. Only a small fraction of groups show a significant deviation from the normal H-H behavior. For example, H14, H118, and D55 in Figure 1 correspond to unusual titration curves and deviate from H-H behavior: the non-asymptotic regions of the titration curves are elongated or step-wise; the D55 curve is asymmetric with elongation on the high pH side. We have previously shown that residues in the active site of proteins frequently exhibit this perturbed behavior [2, 3, 6, 7]. H14, H118, and D55 are indeed in the active site of Phospholipase C [41], whereas H142 is not.

We define the first derivative function f of the θ(pH) curve as:

$$f = -d\theta/d(pH) \qquad (2)$$

The f functions are essentially proton binding capacities [42-44], which measure the change in concentration of a bound proton per unit change in its chemical potential. The binding capacity is also proportional to the well- known Hill coefficient [42].

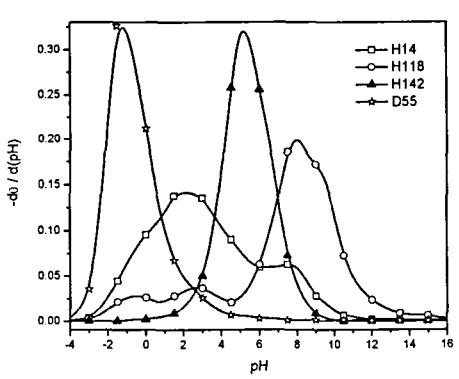

Figure 2: First derivative functions f(pH) of the residues in Figure 1.

Figure 2 shows the f functions of the residues shown in Figure 1. For almost all ionizable residues, the first derivative itself is negative, because increasing pH favors deprotonation. For ideal H-H behavior, f is always greater than or equal to zero. The f function of H142 resembles a Gaussian distribution; however, the f functions of H14, H118, and D55 have multiple peaks or they are asymmetric. Note that these f functions are normalized, *i.e.*, the area under the f curve is unity. This is because θ always runs from 1 to 0 in Eq. 1, so that Δθ is always unity over the full range of θ values; this is true for both perturbed and normal ionizable residues.

The f functions may be treated as distributions and characterized by their moments [7, 42]. Hence we define the n^{th} central moment μ_n as:

$$\mu_n = \int (pH - M_1)^n \cdot f \cdot d(pH) \qquad (3)$$

where M_1 is the first raw moment, defined by the expression for the n^{th} raw moment as:

$$M_n = \int (pH)^n \cdot f \cdot d(pH) \qquad (4)$$

The integrals in equations (3) and (4) are over all space (−∞ to +∞). Equations 2, 3, and 4 for each ionizable residue are evaluated numerically from the theoretical titration curve of each residue.

For a residue exhibiting H-H behavior, 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. However, interactions between ionizable residues in a protein will lead to asymmetry in the f functions and thus the odd-numbered central moments will be non-zero. These interactions will also cause broadening of the f functions and increase the values of the even-numbered moments above those for H-H residues.

We define the Z score for the n^{th} central moment as:

$$Z_n = (|\mu_n| - \langle|\mu_n|\rangle)/\sigma_n \qquad (5)$$

Here $\langle|\mu_n|\rangle$ is the mean of the absolute value of the n^{th} central moment, averaged over all of the ionizable residues in the protein. Strictly, the absolute sign is needed only for the odd moments. σ_n is the standard deviation of the (absolute) values of the n^{th} 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. As illustrated in Figure 2, the f functions are peaked functions, 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.

We previously used the criterion $Z_3>1$ or $Z_4>1$ to select the active-site residues in enzymes; we will use this same criterion to select the binding residues also. This is because the active-site residues tend to have high third and/or fourth central moments and because the $Z_3>1$ or $Z_4>1$ criterion selects most of the known active-site residues [7]. In fact, for most enzymes, the $Z_3>1$ or $Z_4>1$ criterion is more selective than visual identification; often the set of positive residues selected by the statistical criterion is smaller than the set of residues selected by

visual observation. Residues that meet the $Z_3>1$ or $Z_4>1$ test are termed THEMATICS positive residues.

Once the THEMATICS positive residues are selected, we group these residues into clusters based on spatial proximity. The distance between two positive residues is defined as the distance between their charge centers, *e.g.*, the distance between C^{δ}(Glu) and C^{γ}(Asp). For this work, a positive residue is defined as a cluster member 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 are considered predictive and are termed THEMATICS positive clusters. If a THEMATICS positive cluster contains at least one known binding residue, we consider that prediction a success. The $Z_3>1$ or $Z_4>1$ criterion finds only one positive cluster for Phospholipase C: [H14, D55, H69, H118, H128]. All of the residues in this cluster are known to be in the active site and D55 is critical in catalysis [41].

3 RESULTS

The $Z_3>1$ or $Z_4>1$ criterion was applied to a set of 30 non-catalytic ligand-binding proteins in order to predict their ligand-binding sites. The application to Glutamine- binding protein from *E. coli* is discussed in detail. Results for 30 binding proteins follow.

3.1 Glutamine-Binding Protein

Glutamine-binding protein (GlnBP) from *E. coli* is a monomeric protein consisting of two similar globular domains linked by two peptide hinges [45, 46]. It is involved in the active transport of L-glutamine across the cytoplasmic membrane. Calculations were performed on a monomer using the 1.94 Å structure (PDB code: 1WDN) [45]. This protein was crystallized with a bound glutamine; however, the electrostatic potential functions were computed without the bound ligand in the structure.

The binding pocket of GlnBP is composed of D10, F13, F50, A67, G68, T70, R75, K115, T118, G119, H156, D157, and Y185, which account for all hydrogen-bonding and hydrophobic interactions between the ligand and the binding pocket [45]. THEMATICS predicts only one cluster: [**D10**, E17, D28, Y86, **H156**, **D157**, **Y185**, Y213, Y217], where the residues known to be in direct contact with the ligand are indicated in **boldface**. This cluster includes four of the six ionizable residues in the binding pocket. The other five residues in the predicted cluster are not necessarily false positives. E17 and D28, although they do not bind the ligand directly, are located in close proximity to D10 and the ligand: the distance between D10 and E17 is 5 Å and between D10 and D28 9Å as measured by the separation between the charge centers of the residues as discussed above. The three tyrosines, Y86, Y213 and Y217, are located at the entrance of the binding pocket near the surface of the protein. As far as we are aware, there has not been any study to suggest whether these three tyrosines are important in recognition of the ligand. Y86 is one of the "hinge" residues.

Some sample titration curves for GlnBP (1WDN) and the corresponding f functions are shown in Figures 3 and 4. Notice that the perturbations on these residues are not as visually obvious as the perturbed residues in Figures 1 and 2. Weaker perturbations, hence smaller $|\mu_3|$ or μ_4 values, may be more typical for non-catalytic proteins, at least for many of the cases reported here.

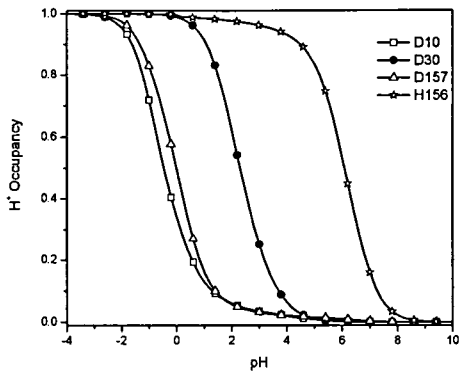

Figure 3: Theoretical titration curves, plotted as ensemble average proton occupation θ as a function of pH, for selected residues in GlnBP (1WDN). D10, D157, and H156 are known to be involved in binding.

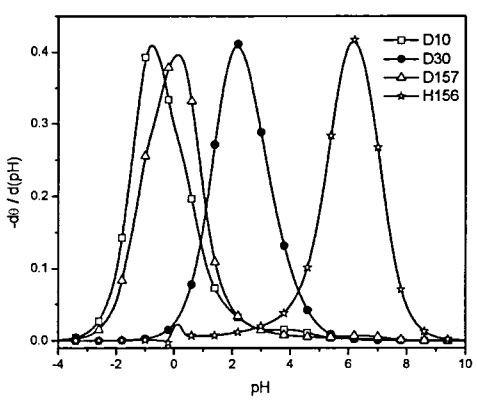

Figure 4: First derivative functions f(pH) of the residues in Figure 3.

Even though Phospholipase C (245 residues) and GlnBP (223 residues) are comparable in size, the perturbations are much more pronounced for the active site residues of Phospholipase C (see Figures 1 and 2). This could be in part because Phospholipase C has more than three times as many ionizable residues as GlnBP: 86 ionizable residues for Phospholipase C and 26 ionizable residues for GlnBP. Therefore, the statistical (rather than visual) selection criterion seems to be more important for non-catalytic proteins than for enzymes. A systematic study is needed before we can generalize about this apparent trend in the magnitude of the moments.

Calculations were also performed on a ligand-free "open" structure of GlnBP using the 2.3 Å resolution structure (PDB code: 1GGG) [46]. The structure of the open conformation is considerably different from that obtained with the bound ligand. In the open conformation, the two globular domains (termed large and small) are physically separate, connected only by two hinges (residues 85-89 and 181-185). In the bound form, these domains close up to form a binding cleft. THEMATICS analysis obtains two predictive clusters for this open structure:[E17, D28] and [Y86, Y213, Y217] as shown in Table 1. It is interesting to note that THEMATICS selects

all of the residues that were identified with the closed structure except for those residues that are directly involved in binding. Although these clusters, particularly [E17, D28], are located in close proximity to the binding pocket, we consider this a failed prediction.

Because the two domains (small = residues 90 to 180, large = residues 1to 84 & 186 to 226) of the open structure are spatially far apart, the electrostatic energy landscape and consequently the magnitude and the range of $|\mu_3|$ and μ_4, are quite different for two domains. The average and standard deviation of $|\mu_3|$ and μ_4 for the small domain were 0.26 ± 0.26 and 3.8 ± 1.1, respectively, while those of the large domain were 0.67 ± 1.6 and 8.7 ± 14.7. So when the criterion is applied to each domain separately, a different set of clusters is predicted: [E17, D28] and [Y86, Y217] from the large domain, [K105, D152], [Y143, **H156**, **D157**] and [Y163, K166] from the small domain. One of the clusters in this two-domain analysis includes two binding residues. The same two-domain statistical analysis applied to the bound form of the structure (1WDN) yielded only one cluster: [**D10**, E17, D28, **H156**, **D157**, **Y185**, Y217], which is nearly identical to the original analysis for this form.

3.2 Summary for 30 Ligand-Binding Proteins

Results for 30 non-catalytic ligand-binding proteins are summarized in Table 1. Here we use **boldface** to indicate residues identified as important for binding function either in the original reference contained in the PDB structure file or in PDBsum site [47]. The thirty proteins in this set were selected by a search of the PDB using the keywords "binding" or "transport". The set was reduced by eliminating mutants and those structures for which no information about the binding site was found. Finally, the structures were then ranked according to the number of missing residues and the thirty most complete were selected. All of these proteins have at least one ionizable residue in each binding site.

The calculations were performed on the monomer for all but seven proteins; for the seven proteins indicated with superscript [(b)] in Table 1, the dimer was analyzed. All hetero-atoms and the ligands were removed from the structure before performing the calculations. Five proteins from Table 1 contain two binding sites per chain or per biological unit; therefore, there are total of 35 binding sites for the entire set. 29 predictive clusters have at least one known binding residue; therefore the success rate is 83% (29/35). The success rate per protein is 80% (24/30). These are less than the 91% success rate for enzymes [7].

Examining the failed cases reveals that most of those proteins have highly hydrophobic binding sites. For example, there is only one ionizable residue in the hormone-binding pocket of both forms of Shbg (1F5F and 1KDM). THEMATICS does identify, for both forms, the ionizable residue D65 as an isolated positive, which is not considered predictive. Likewise, there is only one ionizable residue in the binding pocket of Mosquito sterol carrier protein-2 (1PZ4) and THEMATICS identifies R24 as an isolated positive.

There are several ionizable residues in the vitamin B12-binding pocket of Cobalamine transporter (1NQH): Y229, R497, Y531, and Y579. However, THEMATICS found the Ca-binding site only: [**D193, D195, D230**]. Among these three residues, D230 is slightly exposed to the vitamin B12-binding pocket and connected to A231, which forms a hydrogen-bond with the ligand. Among the four ionizable residues in the binding pocket, only R497 forms a hydrogen bond with the ligand. The four ionizable residues do not have a significant level of electrostatic interaction with each other and their interactions with the ligand are primarily hydrophobic.

While we are able to rationalize the above four failed cases and to some extent, the open form of GlnBP (section 3.1), THEMATICS is unable to identify any of the residues in the binding pocket of Jacalin (1JAC) nor any residues close to the binding pocket. We note that the binding pockets of Jacalin are exposed to the solvent more so than most other binding pockets. The solvation weakens the electrostatic interactions. While the method has found surface interaction sites, predicted clusters are more often found in clefts than on the surface [6].

For Human S100A6 calcium-binding protein (1K9K), the largest THEMATICS positive cluster contains the residues from two Ca-binding pockets. This is because the two Ca ions are only 11.5Å apart and the two binding sites form one THEMATICS cluster.

We note that THEMATICS positive clusters tend to be more localized on the binding site than many pocket search methods [6], with most of the cluster members immediately surrounding the ligand. The maximum distance across the largest predicted cluster in Table 1 is about 27 Å; this corresponds to the binding region of the complex of Actin with Vitamin D-binding protein (1LOT) and reflects the length of that region. Most of the clusters are significantly smaller. THEMATICS also finds 1.9 predictive clusters per biological unit (58/30), reflecting a relatively low rate of false positives.

As shown in Table 1, most THEMATICS positive clusters contain a few residues that are not directly involved in binding of the ligands. These residues should not necessarily be considered as false positives. In most cases, these residues occur in close proximity to bound substrate or may assist in binding or other functions.

We presented a simple computational procedure for the identification of functional sites in proteins; the method can be automated. Our method is applicable to both enzymes and non-catalytic binding proteins, albeit less effective in searching for hydrophobic binding pockets with few or no ionizable residues.

PDB Code	Protein Name	THEMATICS Results [a]
1A99	Putrescine Receptor (Potf) From *E. Coli*	[E66, E184, **E185, D247, D278, Y314**] [H123]
1ABE	L-Arabinose-Binding Protein	[**E14**, E20, **D89, D90**, D206, D235, H259]
1ADL	Adipocyte lipid-binding protein	[**Y19**, R78, **R106, C117, R126, Y128**] [C1]
1BYK	Trehalose Repressor From *E. Coli* [b]	[**R71a**, D73a, **E77a**, E99a, **Y157a**, D159a, D241a, **Y284a**] [D94a, H110b] [Y198a] [H274a] & identical clusters with a and b exchanged
1DK0	Hemophore Hasa From *Serratia Marcescens*	[**H32, Y55, H83, H128, H133**] [D112] [E148]
1DQE	Bombyx Mori Pheromone Binding Protein	[D32] [D63, **H70, H95, E98**] [H123] [[E137]
1EJE	Fmn-Binding Protein from *Methanobacterium thermoautotrophicum*	[D36, E38] [D66, **H67, H68, E105**, D143, **H144**] [E78, E132]
1F5F	The N-Terminal G-Domain Of Shbg In Complex With Zinc [c]	[**D50, E52**, D96, **D162**] [**D65**] [E104, E115] [**D117**]
1FX3	*Haemophilus influenzae* SecB [b]	[Y24a, K26a, **D27a, E31a, E86a, D27b, E31b, E86b**] [E76a, D77a] [Y24b, K26b] [C106a, C111a] [C106b, C111b]
1GGG	*E. Coli* Glutamine Binding Protein (Without Substrate)	[E17, D28] [Y86, Y213, Y217]
1JAC	Jacalin [b]	[D6, Y32, H44] [Y19, D59, E63, K91, Y93,Y96] [K117]
1K9K	Human S100A6 Calcium Binding [c]	[K18, Y19] [**D25**, H27, **E33, D61, D65, E67, E72**]
1KDM	Human Sex Hormone-Binding Globulin [b] (Tetragonal Crystal Form)	[**D65a**] [E77a, E104b, E115b,E120b, R123b, R125b] [D96a] [E104a, E115a, E120a, R123a, R125a, E 77b] [D117a] [**D65b**] [D96b] [D117b]
1KLL	Mrd Protein [b]	[**D52a** , D61a, H71b, Y97b, H100b, D117b, D119b, **D124b**]
1LAH	Periplasmic lysine-, arginine-, ornithine-binding protein	[**D11, Y14, D30**, D85, Y88, D91, **D161**, E162, D211, Y223, Y236, D238]
1LOT	The Complex Of Actin With Vitamin D-Binding Protein [b] [c]	[**Y120a, Y147a, Y151a**, Y162a, Y166a, R187a, K222a, Y297a, Y166b, **D288b, K291b**] [K287a, Y133b, Y143b] [**K213b**, C217b, C257b, **Y306b**] [H265a] [Y394a] [H275b] [C285b]
1N2Z	Btuf, The Vitamin B12 Binding Protein Of *E.* Coli	[E35, D52, **D242**, E245, **R246**] [D108] [Y143, K147, E181, E186, K190] [K153, K156, D266]
1N4A	Periplasmatic Cobalamine Transporter from *E. Coli*	[E13, D30, **D220, E223**] [Y121, K125, E159] [D244]
1NQH	Outer Membrane Cobalamine Transporter from *E. Coli* [c]	[R14, D81, Y118, Y221, Y223] [R47, R48, Y109, R111, E465, D482, K504] [H75, E413] [R84] [**D193, D195, D230**] [K244, Y246]
1OU8	Aaa+ Protease Delivery Protein	[**Y28**] [**Y44, K76**] [D110]
1OZ7	Echicetin From The Venom Of Indian Saw-Scaled Viper [b] [c]	[**E27a, E43a, D81b**] [E48a] [**E71a, E75a, E99a**] [Y14b, K120b] [**D70b, H95b**]
1P28	Pheromone Binding Protein From The Cockroach *Leucophaea Maderae*	[**Y5, R33, Y75, K85**]
1PMP	Lipophilic Binding Protein P2 from *Bos Taurus*	[Y19, R78, **R106**, C117, C124, **R126, Y128**] [K105]
1POT	Spermidine/Putrescine-Binding Protein Complexed With Spermidine from *E. coli*	[**E36**] [E63, Y66, Y86, K89, **D168**, R170, **E171, D257**]
1PW4	Glycerol-3-Phosphate Transporter From *E. Coli*	[**Y38, Y42, R45, K46**, Y76, **Y266, R269, Y270, Y393**] [K284]
1PZ4	Mosquito Sterol carrier protein-2	[D6, E67, D68, D69] [**R24**] [H28, Y30, D47, E61] [E91]
1USK	L-Leucine Binding Protein with Bound L-Leu	[E22, D51, H76] [D121, **E226**] [H145, D146]
1VYF	*Schistosoma Mansoni* Fatty Acid Binding Protein	[K48, K65] [C62, E72, R107, **R127, Y129**] [**D76**]
1WDN	*E. Coli* Glutamine Binding Protein	[**D10**, E17, D28, Y86, **H156, D157, Y185**, Y213, Y217]
2ABH	Phosphate-Binding Protein	[Y16, Y33] [Y104, **R135**, Y193] [Y196, Y206]

Table 1: THEMATICS results for thirty non-catalytic binding proteins [d]

[a] THEMATICS positive clusters for each subunit are reported except when the binding sites are composed of the residues from different subunits; a and b designations distinguish the residues from different subunits; [b] THEMATICS analysis was performed on the dimer unit; [c] there are two binding sites in these proteins; [d] Residues belonging to the same cluster are shown together in square brackets. Residues in **boldface** are known to be involved in ligand binding. An isolated positive is not considered a successful prediction in this study, even when that residue is known to be a binding residue.

4 ACKNOWLEDGMENTS

This work was supported by the U. S. National Science Foundation under grant MCB-0135303 and by the Institute for Complex Scientific Software at Northeastern University.

5 REFERENCES

[1] Westbrook J, Feng Z, Chen L, Yang H, Berman HM, The Protein Data Bank and structural genomics, *Nucleic Acids Res*, **31**: 489-491, 2003.

[2] Murga L, Wei Y, Andre P, Clifton JG, Ringe D, Ondrechen MJ, Physicochemical methods for prediction of functional information for proteins, *Israel Journal of Chemistry*, **44**: 299-308, 2004.

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

[4] Ringe D, Wei Y, Boino KR, Ondrechen MJ, Protein Structure to Function: Insights from Computation, *Cellular Molecular Life Sciences*, **61**: 387-392, 2004.

[5] Shehadi IA, Yang H, Ondrechen MJ, Future directions in protein function prediction, *Mol Biol Reports*, **29**: 329-335, 2002.

[6] Ko J, Murga LF, Wei Y, Ondrechen MJ, Prediction of active sites for protein structures from computed chemical properties, *Bioinformatics*, **21**: i258-i265, 2005.

[7] Ko J, Murga L, Andre P, Yang H, Ondrechen MJ, Williams RJ, Agunwamba A, Budil DE, Statistical Criteria for the Identification of Protein Active Sites Using Theoretical Microscopic Titration Curves, *Proteins: Structure Function Bioinformatics*, **59**: 183-195, 2005.

[8] del Sol Mesa A, Pazos F, Valencia A, Automatic methods for predicting functionally important residues, *J Mol Biol*, **326**: 342-358, 2003.

[9] Lichtarge O, Bourne HR, Cohen FE, An evolutionary trace method defines binding surfaces common to protein families, *J Mol Biol*, **257**: 342-358, 1996.

[10] Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N, Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues, *Bioinformatics*, **18**: S71-S77, 2002.

[11] Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O, An accurate, sensitive, and scalable method to identify functional sites in proteins, *J Mol Biol*, **326**: 255-261, 2003.

[12] 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.

[13] 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.

[14] Aloy P, Querol E, Aviles FX, Sternberg MJE, 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.

[15] 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.

[16] Gutteridge A, Bartlett G, Thornton JM, Using a neural network and spatial clustering to predict the location of active sites in enzymes, *J Mol Biol*, **330**: 719-734, 2003.

[17] Innis CA, Anand AP, Sowdhamini R, Prediction of functional sites in proteins using conserved functional group analysis, *J Mol Biol*, **337**: 1053-1068, 2004.

[18] 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.

[19] Elcock AH, Prediction of functionally important residues based solely on the computed energetics of protein structure, *J Mol Biol*, **312**: 885-896, 2001.

[20] Greaves R, Warwicker J, Active site identification through geometry-based and sequence-profile based calculations: Burial of catalytic clefts, *J Mol Biol*, **349**: 547-557, 2005.

[21] Peters KP, Fauck J, Frommel C, The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria, *J Mol Biol*, **256**: 201-213, 1996.

[22] Binkowski TA, Naghibzadeh S, Liang J, CASTp: computed atlas of surface topography of proteins, *Nucleic Acids Res*, **31**: 3352-3355, 2003.

[23] Venkatachalam CM, Jiang X, Oldfield T, Waldman M, LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites, *J Mol Graph*, **21**: 289-307, 2003.

[24] Brady Jr GP, Stouten PFW, Fast Prediction and Visualization of Protein Binding Pockets With PASS, *Journal of Computer-Aided Molecular Design*, **14**: 383-401, 2000.

[25] Laskowski RA, SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions, *J Mol Graph*, **13**: 323-330, 1995.

[26] Kleywegt GJ, Jones TA, Detection, delineation, measurement and display of cavities in macromolecular structures, *Acta Cryst*, **D50**: 178-185, 1994.

[27] Laurie ATR, Jackson RM, Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites, *Bioinformatics*, **21**: 1908-1916, 2005.

[28] Bashford D, Gerwert K, Electrostatic calculations of the pKa values of ionizable groups in bacteriorhodopsin, *J Mol Biol*, **224**: 473-486, 1992.

[29] Beroza P, Fredkin, D. R., Okamura, M. Y., Feher, G., Electrostatic calculations of amino acid titration and electron transfer, Q-AQB-->QAQ-B, in the reaction center, *Biophys J*, **68**: 2233-2250, 1995.

[30] Carlson HA, J.M. Briggs and J.A. McCammon, Calculation of the pKa values for the ligands and side chains of Escherichia coli D-alanine:D-alanine

ligase, *J Med Chem*, **42**: 109-117, 1999.
[31] Sampogna RV, Honig B, Environmental effects on the protonation states of active site residues in bacteriorhodopsin, *Biophys J*, **66**: 1341-1352, 1994.
[32] Bashford D, Karplus M, Multiple-site Titration Curves of Proteins: An Analysis of Exact and Approximate Methods for Their Calculation, *J Phys Chem*, **95**: 9556-9561, 1991.
[33] Gilson MK, Multiple-site titration and molecular modeling: two rapid methods for computing energies and forces for ionizable groups in proteins, *Proteins: Structure, Function, and Genetics*, **15**: 266-282, 1993.
[34] Bashford D, Karplus M, pKa's of Ionizable Groups in Proteins: Atomic Detail from a Continuum Electrostatic Model, *Biochem*, **29**: 10219-10225, 1990.
[35] Honig B, Nicholls A, Classical electrostatics in biology and chemistry, *Science*, **268**: 1144-1149, 1995.
[36] Madura JD, Briggs JM, Wade RC, Davis ME, Luty BA, Ilin A, Antosiewicz J, Gilson MK, Bagheri B, Scott LR, McCammon JA, Electrostatics and diffusion of molecules in solution - Simulations with the University of Houston Brownian Dynamics program, *Comp Phys Commun*, **91**: 57-95, 1995.
[37] Ren P, Ponder JW, Polarizable atomic multipole water model for molecular mechanics simulation, *J Phys Chem B*, **107**: 5933-5947, 2003.
[38] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML, Comparison of simple potential functions for simulating liquid water, *J Chem Phys*, **79**: 926-935, 1983.
[39] Jorgensen WL, Tirado-Rives J, The OPLS Potential Functions for Proteins. Energy Minimization for Crystals of Cyclic Peptides and Crambin, *J Am Chem Soc*, **110**: 1657-1666, 1988.
[40] Hough E, Hansen LK, Birknes B, Jynge K, Hansen S, Hordvik A, Little C, Dodson E, Derewenda Z, High-resolution(1.5 A) crystal structure of phospholipase C from Bacillus cereus., *Nature*, **338**: 357-360, 1989.
[41] Martin SF, Hergenrother PJ, General base catalysis by the phophatidylcholine-preferring phospholipase C from Bacillus cereus: the role of Glu4 and Asp55., *Biochem*, **37**: 5755-5760, 1998.
[42] Di Cera E, Chen Z-Q, The Binding capacity is a probability density function, *Biophys J*, **65**: 164-170, 1993.
[43] Di Cera E, Gill SJ, Wyman J, Binding Capacity: Cooperativity and buffering in biopolymers, *Proc Natl Acad Sci (USA)*, **85**: 449-452, 1988.
[44] Wyman J, Linked functions and reciprocal effects in hemoglobin: A second look, *Adv Protein Chem*, **19**: 223-286, 1964.
[45] Sun Y-J, Rose J, Wang B-C, Hsiao C-D, The structure of glutamine-binding protein complexed with glutamine at 1.94 A resolution: Comparison with other amino acid binding proteins, *J Mol Biol*, **278**: 219-229, 1998.
[46] Hsiao C-D, Sun Y-J, Rose J, Wang B-C, The crystal structure of glutamine-binding protein from Escherichia coli, *J Mol Biol*, **262**: 225-242, 1996.
[47] Laskowski RA, Chistyakov VV, Thornton JM, PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids, *Nucleic Acids Res*, **33**: D266-D268, 2005.