Prediction of active sites for protein structures from computed chemical properties

Jaeju Ko¹, Leonel F. Murga², Ying Wei² and Mary Jo Ondrechen²,*

¹Department of Chemistry, Indiana University of Pennsylvania, 975 Oakland Avenue, Indiana, PA 15705 USA and ²Department of Chemistry and Chemical Biology and Institute for Complex Scientific Software, Northeastern University, Boston, MA 02115, USA

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT
Motivation: Identification of functional information for a protein from its three-dimensional (3D) structure is a major challenge in genomics. The power of theoretical microscopic titration curves (THEMA TICS), when coupled with a statistical analysis, provides a method for high-throughput screening for identification of catalytic sites and binding sites with high accuracy and precision. The method requires only the 3D structure of the query protein as input, but it performs as well as other methods that depend on sequence alignments and structural similarities.

Contact: m.ondrechen@neu.edu

1 INTRODUCTION
The Protein Databank (Berman et al., 2000) currently contains on the order of 10³ protein structures annotated as ‘hypothetical’ or ‘unknown function’, and this number is increasing rapidly as a result of structural genomics efforts (Westbrook et al., 2003). Fast, reliable computational methods are needed to predict functional information for these proteins and this is the primary purpose of the work described here. In addition, it is desirable to identify which properties of active residues are required for catalysis and/or recognition, in order to understand how one might engineer these properties into a protein scaffold (Dwyer et al., 2004; Kaplan and DeGrado, 2004; Looger et al., 2003) to design an enzymatic species with a specified type of reactivity.

This paper describes a method for the identification of active site residues in a protein structure, based on the computed acid–base properties of the ionizable residues (Arg, Asp, Cys, Glu, His, Lys, Tyr and the N- and C-termini). Our method, THEMA TICS (Murga et al., 2004; Ondrechen et al., 2001; Ringe et al., 2004; Shehadi et al., 2002), utilizes Poisson–Boltzmann methods to calculate the theoretical microscopic titration curves of the ionizable residues (Antosiewicz et al., 1996; Bashford and Karplus, 1991; Georgescu et al., 2002; Gilson, 1993; Yang et al., 1993). THEMA TICS requires only the three-dimensional (3D) structure of the query protein as input and thus applies to novel folds as well as to familiar folds.

The novel feature of THEMA TICS is that it extracts information from the shapes of the theoretical titration curves calculated for the ionizable residues in a protein. We have argued that residues with significantly perturbed predicted titration behavior tend to occur in the active site of an enzyme and in binding sites of proteins to such a great extent that they serve as revealing markers of reactivity and recognition. THEMA TICS identifies these perturbed residues and searches for clusters of two or more such residues in physical proximity; these clusters are accurate predictors of interaction sites.

Originally, a human observer would identify the residues with anomalous titration curves (Ondrechen et al., 2001). Here statistical criteria are described to select the active residues. This enables the automated identification of the anomalous residues and of active sites. We will show that the statistical criteria are simple to apply and highly accurate. Results are reported for some well-characterized proteins for verification purposes and some new predictions are made for structural genomics proteins.

2 METHODS
2.1 Calculation of theoretical microscopic titration curves
In order to calculate the titration curve of each ionizable residue, finite difference Poisson–Boltzmann calculations were performed using UHBD (Madura et al., 1995) on each protein followed by the program HYBRID (Gilson, 1993), a Monte Carlo method for computing proton occupation as a function of pH. The protein structures used as the input data for these calculations were downloaded from the Protein Data Bank (http://www.rcsb.org/pdb/) and the hydrogen atoms were built into the structure using the program TINKER (Ren and Ponder, 2003) and the OPLS-UA force field (Jorgensen et al., 1983; Jorgensen and Tirado-Rives, 1988). Substrates, water molecules, cofactors and salts that crystallized with the...
proteins were not included in the electrostatic calculations. The internal dielectric constant was assumed to be 20 and the external dielectric was assumed to be 80; other assumptions made in the calculation have been described elsewhere (Ko et al., 2005; Murga et al., 2004).

2.2 Statistical analysis of the curves

Most of the ionizable residues in a protein obey the Henderson–Hasselbalch (H–H) equation, which can be written as:

\[
C(pH) = (\pm 1)/(1 + 10^{\pm (pH - pK_a)}),
\]

(1)

where the plus sign applies for residues that form a cation upon protonation (Arg, His, Lys and the N-terminus) and the minus sign for residues that form an anion upon deprotonation (Asp, Cys, Glu, Tyr and the C-terminus). Note that \( C \) in Equation (1) represents the average charge on a particular residue for a large ensemble of protein molecules. For a monoprotic acid \( C(pH) \) obeys Equation (1) and has the familiar sigmoid shape. Thus, as \( pH \) is increased, the predicted average charge falls sharply in a \( pH \) range close to the \( pK_a \). Note that \( C \) always runs from +1 to 0 for the cation-forming residues and from 0 to −1 for the anion-forming residues, which means that \( \Delta C \) is always unity over the full range of \( C \) values; this is true for all ionizable residues, even those that do not fit the H–H equation.

A protein is a biomacromolecular polyprotic acid and thus all of its ionizable groups deviate from the H–H equation, at least to some extent. However, for most of these groups the deviations are small and the shape of their titration curves follows very closely the one predicted by the H–H equation. Only a small fraction of groups show a prominent deviation from the normal H–H behavior. We have previously shown (Murga et al., 2004; Ondrechen et al., 2001; Shehadi et al., 2002) that residues in the active site of proteins commonly exhibit this perturbed behavior.

A first derivative function \( f \) of the \( C(pH) \) curve is defined as:

\[
f = -dC/d(pH).
\]

(2)

For most ionizable residues the first derivative itself is negative, as charge almost always decreases with increasing \( pH \); for normal H–H behavior \( f \) is positive or zero, always. The \( f \) functions are essentially proton binding capacities (Di Cera et al., 1988; Wyman, 1964), which measure the change in concentration of a bound ligand per unit change in the chemical potential of the ligand. The \( f \) functions resemble Gaussian-like distribution functions. For H–H residues, \( f \) is a sharply peaked function with a peak at \( pH = pK_a \). The \( f \) functions are also normalized, since the range of \( C \) values necessarily implies that the area under the \( f \) curve is unity. The \( f \) curves may be treated as distributions and characterized by their moments. Hence we can define the \( n \)-th central moment \( \mu_n \) as:

\[
\mu_n = \int (pH - M_1)^n \cdot f \cdot d(pH)
\]

(3)

where \( M_1 \) is the first raw moment, defined by the expression for the \( n \)-th raw moment as:

\[
M_n = \int (pH)^n [-dC/d(pH)]d(pH).
\]

(4)

Note that the integrals in Equations (3) and (4) are over all space (−∞ to +∞). Equations (2), (3) and (4) for each residue are evaluated numerically from the theoretical titration curves. We define the Z score for the \( n \)-th central moment as:

\[
Z_n = (\mu_n - \langle \mu_n \rangle)/\sigma_n, \quad n \text{ even}
\]

(5)

\[
Z_n = (|\mu_n| - \langle |\mu_n| \rangle)/\sigma_n, \quad n \text{ odd}
\]

(6)

Here \( \langle \mu_n \rangle \) is the mean value for the \( n \)-th central moment (or its absolute value), averaged over all of the ionizable residues in the protein. \( \sigma_n \) is the standard deviation of the values of the \( n \)-th central moment for the set of all ionizable residues in the protein. The Z score thus represents the deviation from the mean value in units of the standard deviation. The first derivative functions \( f \) are peaked functions even for the non H–H titration curves, 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 theoretical titration curves.

For a residue exhibiting H–H behavior, all the odd-numbered central moments are zero. The second and fourth central moments have the values 0.620 and 1.62 respectively. 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 even-numbered moments above the limiting values given for H–H residues.

The first derivative functions \( f \) for most of the ionizable residues in the active site of a protein exhibit significant asymmetry or broadening and many exhibit both. The criterion we have chosen to identify active site residues is \( Z_3 > 1 \) or \( Z_4 > 1 \). This means that the corresponding central moment is more than one standard deviation greater than the mean. The criterion \( Z_3 > 1 \) or \( Z_4 > 1 \) is selective; the set of residues that it identifies as positive tends to be a little smaller than the set identified by the human observer (Ondrechen et al., 2001). In other words, there may be a few residues with predicted titration curves that look non-sigmoidal to the eye, but the degree of perturbation is not sufficient to meet the \( Z_3 > 1 \) or \( Z_4 > 1 \) threshold. 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 then identify clusters of these residues in physical proximity.
The distance between two positive residues is defined as the distance between the centers of positive or negative charge in their ionized forms. For present purposes, a positive residue is defined as a cluster member if it is within 9 Å of any other positive residue in the cluster. Clusters containing only one member are not considered predictive. Clusters containing two or more positive residues are considered predictive and are termed THEMATICS-positive clusters.

3 RESULTS

We first describe in detail the application of the $Z_3 > 1$ or $Z_4 > 1$ criterion to three proteins, followed by a summary of the results for a set of 12 enzymes that span the major E.C. classes 1–6. Predictions for Structural Genomics proteins and novel folds follow.

3.1 Glutamate racemase

Glutamate racemase catalyzes the synthesis of D-glutamate, an essential building block of the peptidoglycan layer in bacterial cell walls. Calculations were performed on the biologically active dimer from *Aquifex pyrophilus*, using the structure (PDB code 1B73) determined at 2.3 Å resolution (Hwang et al., 1999). Glutamate racemase has 86 ionizable residues in each subunit of 254 amino acids.

The $Z_3$ and $Z_4$ values for all of the ionizable residues were computed and the residues that satisfy the $Z_3 > 1$ or $Z_4 > 1$ criterion were found to be: D7, Y39, C70, Y123, C139, C178, H180, E147, E148, K185 and K189. The known catalytic residues are in boldface. Four THEMATICS-positive clusters in the dimer structure were identified using the 9 Å cutoff distance: [D7, Y39, C70, Y123, C139, C178, H180, E147, E148, K185, K189], and two identical clusters with a and b reversed, where the two subunits of the dimer are labeled as a and b.

All of the known catalytic residues, except S8, belong to one of the two equivalent nine-membered THEMATICS-positive clusters located at the dimer interface. Although THEMATICS does not classify serine residues as either positive or negative, we note that it has identified as positive the nearest ionizable neighbors of the active site S8. This site S8 is very close to D7, Y39 and to C70. The side chain O atom of S8 is 3.8 Å from a carboxylate O atom of D7, 6.9 Å from the phenolic O atom of Y39 and 3.8 Å from the S atom of C70.

As mentioned above, THEMATICS also predicts as positive two equivalent two-membered clusters, each of which contains two lysine residues, K185 and K189. Each one of these clusters is located at the entrance of one of the two catalytic cavities; it is not clear at this time whether these two pairs of lysines play a role in shuttling reactants and products in and out of the active site, or whether they are simply false positives.

3.2 L-Fuculose 1-phosphate aldolase

L-Fuculose 1-phosphate aldolase catalyzes the reversible cleavage of L-fuculose 1-phosphate to dihydroxyacetone phosphate and l-lactaldehyde. Calculations were performed on the structure (PDB code 1FUA) from *Escherichia coli* determined at 1.92 Å resolution. This structure lacks the flexible C-terminal tail (residues 207–215) and thus this portion could not be included in our calculations. A mutagenesis study (Joerger et al., 2000) concludes that E73 acts as an acid and as a base during the catalysis and that it is the most important catalytic residue. This study further concludes that Y113′ and Y209′ (where the prime designates the other subunit) also play a role in catalysis because the replacement of either Y113 or both residues results in, respectively, a 20- and a 500-fold reduction in activity. H92, H94 and H155 are bound to a catalytic Zn ion in the active site. In addition, the study reports that the adjacent nonpolar region of E73 is a factor in substrate selectivity.

L-Fuculose-phosphate-aldolase has 60 ionizable residues in each subunit of 215 amino acids. The $Z_3 > 1$ or $Z_4 > 1$ criterion finds four equivalent predictive clusters in the tetramer: [E73, H77, H92, H155], one for each of the four subunits. It also finds two one-member clusters (of C160 and of E175) that are not predictive. Table 1 shows the $Z$ values for these residues. The $Z$ values for Y113 are <1.0, so it does not meet the criterion. Y209 was not included in the calculation because it is a part of the C-terminal tail that is missing from the PDB structure. It is possible that Y113 is not predicted by our method because of the absence of Y209. Perhaps the interaction between Y113 and Y209 enhances catalysis, as the experiments suggest. The $Z_3 > 1$ or $Z_4 > 1$ criterion locates the remaining active site residues correctly, even though nine residues are missing from the input structure in the electrostatic potential calculations.

| Table 1. Z values for L-fuculose-phosphate-aldolase |
|--------------------------|--------------------------|
| Residue | $Z_3$ | $Z_4$ |
| E73 | 2.2 | 1.0 |
| H77 | 2.0 | 1.1 |
| H92 | 4.3 | 6.4 |
| H155 | −0.56 | 1.8 |
| C160 | 2.3 | 1.4 |
| E175 | 3.7 | 1.9 |

3.3 Glycine amidinotransferase

Glycine amidinotransferase catalyzes a step in creatine biosynthesis by formation of guanidinoacetic acid, the immediate precursor of creatine. Calculations were performed on the crystal structure (PDB code 1JDW) of the recombinant human enzyme determined at 1.9 Å resolution (Humm et al., 1997). Glycine amidinotransferase has 125 ionizable residues in each chain of 423 amino acids.

A structural study by Humm et al. (1997) proposes D254, H303 and C407 as the catalytic triad and D170, M302, H303,
Table 2. THEMATICS positive residue for some representative enzymes

<table>
<thead>
<tr>
<th>PDB ID/name (E.C.)</th>
<th>Result (from Z₃ OR Z₄ rule)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1EBF/homoserine dehydrogenase (1.1.1.3)</td>
<td>[K117, K118, E208, D210, D213, D214, D219, R222, K223], [Y141, H142, E143, D351], [E170], [E262], [D268]</td>
</tr>
<tr>
<td>1PAB/alanine dehydrogenase (1.4.1.1)</td>
<td>[E8, E13, R15, K147, E157, Y193, H195, Y116, E117]</td>
</tr>
<tr>
<td>1JDW/glycine amidinotransferase (2.1.4.1)</td>
<td>[D170, D254, H303, D305, H406], [Y203]</td>
</tr>
<tr>
<td>1AL6/citrate synthase (2.3.3.1)</td>
<td>[Y231, H235, D237, H238, H274, R329, D375, R401], [Y318], Y330, [Y190, Y219], [E86], [D174]</td>
</tr>
<tr>
<td>1CHD/protein-glutamate methyltransferase (3.1.6.1)</td>
<td>[H190, H233, H248, H256, D286]</td>
</tr>
<tr>
<td>5FTI/bis(5′-adenosyl)-triphosphatase (3.6.1.29)</td>
<td>[H35, H94, H96, H98], [E54]</td>
</tr>
<tr>
<td>1FUA/l-fuculose 1-phosphate aldolase (4.1.2.17)</td>
<td>[E73], [H77], [H129], [C160], [E175]</td>
</tr>
<tr>
<td>2PLC/phosphatidylinositol diacylglycerol lyase (4.6.1.13)</td>
<td>[H45], [D46], [D82], [Y71], [K115], [D127]</td>
</tr>
<tr>
<td>1B73/glutamate racemase (5.1.1.3)</td>
<td>[D7], Y39, C70, Y123, C139, C178, H180, E147, E148], [K185, K189]</td>
</tr>
<tr>
<td>1PYP/phosphoenol pyruvate mutase (5.4.2.9)</td>
<td>[D85], [D87], E114, E161, H190], [K102], [C145]</td>
</tr>
<tr>
<td>2HGS/glutathione synthase (6.3.2.3)</td>
<td>[R125], D127, E144, K305, K364, E368, Y375, E425], [R236], Y265, R267, Y270, C294], [E287], [D469]</td>
</tr>
<tr>
<td>1JM/but adducinate synthase (6.3.4.4)</td>
<td>[D13], K16, H41, E221, K267, Y269, R305, K331], [H110], [C328]</td>
</tr>
</tbody>
</table>

Residues belonging to the same cluster are shown together in square brackets. Known catalytic residues are in boldface; other residues in direct contact with a bound ligand are underlined. Clusters with two or more residues are predictive.

D305, R322, S354, S355 and G402 as the residues involved in binding of the substrates. THEMATICS finds one predictive cluster: [D170, D254, H303, D305, H406]. This cluster contains two of the three residues of the proposed catalytic triad, D254 and H303, shown in boldface. Of the four ionizable residues reported to be involved in substrate recognition and binding, the Z₃ > 1 or Z₄ > 1 criterion finds three of them, D170, H303 and D305, where underline designates binding, but misses R322.

3.4 Summary for twelve enzymes

Results for 12 enzymes are shown in Table 2 with their E.C. numbers and with the list of all THEMATICS-positive residues, with clusters shown in square brackets. These enzymes represent a wide spectrum of chemical functions, as reflected in the E.C. classifications. Residues reported to be catalytic are shown in boldface. Other residues in direct contact with a bound ligand in the active site, as determined by the program LPC (Sobolev et al., 1999), are underlined.

The Z₃ > 1 or Z₄ > 1 criterion does quite well in predicting the active sites for these enzymes. Note that this set includes members of the E.C. classes 1–6, illustrating the applicability of the present method across different classes of enzymes. The agreement is excellent between the predictive clusters and the experimentally determined active sites. Active site information is obtained from the Catalytic Residue (CATRES) Dataset (Bartlett et al., 2002), except in the case of phosphoenol pyruvate mutase, where the active site information was obtained from the PDB structure file. The predictive clusters contain some residues involved in recognition, in addition to the catalytic residues. Some of these are identified by LPC (Sobolev et al., 1999) and are underlined. Of the remaining residues predicted, it is not yet clear how many play a supporting role in catalysis and/or binding and how many are false positives.

Using the very simple Z₃ > 1 or Z₄ > 1 rule, THEMATICS finds about three-quarters of all of the residues listed in the literature as catalytic. The exact percentage varies according to the database used (CATRES, the PDB structure SITE field, and protein-specific literature). The success rate in finding the local region of the active site, one that includes most of the important residues, is far higher than this.

3.5 Structural genomics proteins

A summary of the THEMATICS results for seven structural genomics proteins is given in Table 3. The last two are novel folds. Six of the seven are from Thermotoga maritima and one, NADH pyrophosphatase, is from E.coli k12.

NADH pyrophosphatase (PDB code 1VK6) was analyzed using the structure determined at 2.20 Å resolution. This structure was crystallized with one Zn ion and three 2-methyl-2,4-pentanediol (MPD) molecules. C98, C101, Y123, C139, C178, and D254 are of unknown function. TM0935 is a hypothetical protein; its sequence yields only three unique PSI-BLAST (Altschul et al., 1997) hits with an expect value E better than 0.001 (Jones and Swindells, 2002). Here the number of unique PSI-BLAST hits includes the sequence of the subject protein in the count and is based on
THEMATICS-predicted clusters do occur on the surface. For example, in the case of phosphatidylinositol diacylglycerol lyase (PDB ID 2PLC; Table 2), the first predicted cluster [H45, D46, D82], is located in the second-largest cleft in the structure, as determined by CASTp (Liang et al., 1998). The second predictive cluster consists of two residues, Y71 and K115, both of which are highly solvent-exposed and are not located in any cleft. Thus it is possible to find perturbed predicted titration behavior for surface residues, although it more typically occurs in clefts. There is some possibility that such surface sites could have binding or recognition functions.

### 3.6 Application to novel folds

One of the great strengths of THEMATICS is its ability to make predictions for novel folds, since only the structure of the query protein is used as input. TM0160 (PDB ID 1VJL) is a new fold reported by the Joint Center for Structural Genomics. It is a hypothetical protein related to wound-inductive proteins in plants. Its sequence returns only two unique PSI-BLAST hits with an E-value better than 0.1 and only four unique hits with an E-value of better than 1.0. TM1602 (PDB ID 1J5Y) is reported to be a transcriptional regulator of the biotin repressor family, the sequence of which gives only two unique PSI-BLAST hits with $E < 0.01$ and only three with an E-value better than 1.0. These are examples of protein structures that are less accessible to functional site prediction methods based on sequence conservation or evolutionary history because these proteins possess a small number of close homologues.

### 3.7 Clefts and surface sites

Most of the predicted THEMATICS-positive clusters occur in clefts in the protein structure, as is typical of catalytic sites in enzymes. However, a small number of THEMATICS-predicted clusters do occur on the surface. For instance, for phosphatidylinositol diacylglycerol lyase (PDB ID 2PLC; Table 2), the first predicted cluster [H45, D46, D82], is located in the second-largest cleft in the structure, as determined by CASTp (Liang et al., 1998). The second predictive cluster consists of two residues, Y71 and K115, both of which are highly solvent-exposed and are not located in any cleft. Thus it is possible to find perturbed predicted titration behavior for surface residues, although it more typically occurs in clefts. There is some possibility that such surface sites could have binding or recognition functions.

#### Table 3. THEMATICS-positive residues for seven structural genomics proteins

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB code</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>NADH pyrophosphatase</td>
<td>1VK6</td>
<td>[H146, E174, E177, E178, E219]</td>
</tr>
<tr>
<td>Glycyl-tRNA synthetase (TM0216)</td>
<td>1J5W</td>
<td>[D26a, E28a, H212a, Y216a, D234a, K238a, H241a, E28b, H212b, Y216b, K238b, H241b], [H35a], [H35b]</td>
</tr>
<tr>
<td>TM0935</td>
<td>1O50</td>
<td>[R37a, E116a, D129a, R37b, E116b, D129b]</td>
</tr>
<tr>
<td>TM1083</td>
<td>1J5U</td>
<td>[H4a, H6a, E32b, H4b, H6b, E32a], [H13a, H13b]</td>
</tr>
<tr>
<td>TM1602</td>
<td>1J5Y</td>
<td>[H79a, E87a, H146a, H148a] and equivalent clusters in subunits b, c and d</td>
</tr>
<tr>
<td>TM0160</td>
<td>1VJL</td>
<td>[E39, H58, D115], [E142, E145]</td>
</tr>
<tr>
<td>TM0449</td>
<td>1KQ4</td>
<td>[R74a, R78a, R80a, R174a, R165c, Y84d, Y96d, Y133d, R147d] and three equivalent clusters</td>
</tr>
</tbody>
</table>

a search of the comprehensive NCBI non-redundant database. Likewise TM1083 is a hypothetical protein; its sequence is even more remote and returns only one unique PSI-BLAST hit with an $E$-value better than 0.1 and only four unique hits with an $E$-value of better than 1.0. TM1602 (PDB ID 1J5Y) is reported to be a transcriptional regulator of the biotin repressor family, the sequence of which gives only two unique PSI-BLAST hits with $E < 0.01$ and only three with an E-value better than 1.0. These are examples of protein structures that are less accessible to functional site prediction methods based on sequence conservation or evolutionary history because these proteins possess a small number of close homologues.

Indeed THEMATICS-predictive clusters are often located in a region that is a small subset of a large cleft. For instance, citrate synthase, the largest cluster, has eight members (Table 2), five of which are members of the largest cleft, as determined by CASTp (Liang et al., 1998). A total of 77 residues, 22 of which are ionizable, are members of this largest cleft. Similarly THEMATICS predicts one predictive cluster with eight members for adenylosuccinate synthase (Table 2). Most of these eight residues belong to the largest cleft. This largest cleft consists of 77 residues, 23 of which are ionizable. This illustrates how the predictive clusters obtained from THEMATICS tend to be more specific, localized regions than the largest cleft and tend to identify the precise part of the cleft where reactivity occurs.

Furthermore, active sites are not always in the largest clefts. For instance the THEMATICS-predictive cluster shown in Table 2 for phosphoenol pyruvate mutase [D85, D87, E114, E161, H190], occurs in two locations in the dimeric structure, once in each of the two subunits of the dimer. These two clusters constitute the ninth and tenth largest clefts of the dimeric structure.
3.9 Cluster dimension

The clusters predicted by THEMATICS tend to surround the binding pocket of the substrate molecule, so that even if the cluster has several members, it extends over a relatively small area. For instance, homoserine dehydrogenase from Saccharomyces cerevisiae has a relatively large active site that accommodates both the substrate molecule and the dinucleotide cofactor molecule NAD. While the largest predicted THEMATICS cluster has nine members (Table 2), the maximum distance across this predicted cluster is only 17.5 Å; this reflects the size of the combined binding sites for the substrate and for the NAD cofactor (DeLaBarre et al., 2000). Here the maximum distance across the cluster is equal to the largest element of the distance matrix for that cluster, where the distance between residues is defined as the distance between the centers of charge, as above.

Similarly for citrate synthase from chicken, the largest THEMATICS-positive cluster contains eight residues (Table 2). The maximum distance across this cluster is 15.2 Å, again a reasonable size for an enzyme active site.

Of the twelve enzymes shown in Table 2, the one with the largest linear dimension is adenosylsuccinate synthetase from E.coli. The maximum distance across its eight-member cluster is 19.4 Å. This is also an enzyme that binds both a cofactor and a substrate, and the size of the predicted cluster again reflects the actual combined binding pockets of cofactor and substrate molecules. The smallest active site cluster prediction in Table 2 is that of phosphatidylinositol diacylglycerol lyase, which measures 6.5 Å across at its widest.

Most of the THEMATICS-positive clusters contain at least a couple of residues that are not necessarily known catalytic or recognition residues. It is not known at this time whether these additional residues indeed play some important role. Some of these additional residues are immediate neighbors of the bound substrate molecule. Others are located just behind the known catalytic and recognition residues, in the ‘second coordination shell’ of the bound substrate molecule. However, it is clear that the THEMATICS-positive clusters tend to be well localized in space and are of the appropriate size for enzyme active sites.

3.10 Possible application to non-catalytic binding sites

In addition to catalytic sites, non-catalytic recognition sites and other reversible binding sites can be found by THEMATICS, at least in some cases. For instance, the protease Kex2 (Holyoak et al., 2003) is specific for dibasic sites. THEMATICS found the protease catalytic site and also the recognition residues in the S1 and S2 sites (Ringe et al., 2004). Whereas for the non-specific protease subtilisin, THEMATICS found only the catalytic site. S1 and S2 are the two recognition pockets, the specificity determinants for Kex2, for the two basic sites P1 and P2 on the N-terminal side of the reactive amide linkage in the substrate protein molecule.

While the primary focus of THEMATICS analysis has been on enzymes, there are a few examples of application to non-catalytic binding proteins. For example, for adipocyte lipid-binding protein from mouse, THEMATICS analysis was performed on the monomeric structure (PDB ID 1ADL). One predictive cluster was found: [Y19, R78, R106, C117, R126, Y128]. Calorimetric and crystallographic studies indicate that the thermodynamics of binding is dominated by R106, R126 and Y128 (LaLonde et al., 1994). In addition, Y19, R78 and C117 are all located within 5 Å of the bound arachidonic acid molecule in the crystal structure of the complex.

For the putrescine receptor from E.coli (PDB ID 1A99), six residues are determined by X-ray crystallography and mutational analysis to be crucial to putrescine binding (Vassylev et al., 1998). Of these, three residues are ionizable: E185, D247 and D278. THEMATICS finds a positive cluster at the putrescine binding site: [E66, E184, E185, D247, D278, Y314] that includes the three crucial ionizable residues. In addition, Y314 is located about 3 Å away from the bound ligand.

4 DISCUSSION

We have presented a computational method that identifies active sites in proteins. The method is based on the statistical metrics that characterize the theoretical microscopic titration curves (THEMATICS) of the ionizable residues. This method first calculates the theoretical titration curves of each residue from the three-dimensional structure of the query protein. Next, the statistical metrics (µ3 and µ4) are evaluated numerically from the theoretical titration curves of each residue of the protein. The statistical criterion Z3 < 1 or Z4 < 1 is applied to select the THEMATICS-positive residues. Finally, the spatially localized THEMATICS-positive clusters are identified using a distance cutoff criterion.

While the theoretical titration curves are obtained by an approximate electrostatic calculation, the quality of the computed curves is good enough for the purposes of active site prediction. While more accurate methods are available, [e.g. Alexov and Gunner (1997); Baker et al. (2001)], the approximations are necessary in order to perform the analysis in a reasonable amount of time and on large numbers of protein structures. The current method takes anywhere from several minutes for a very small protein to a couple of days for a large, multimeric system with a couple thousand residues; for a protein of average size, the analysis requirement is the order of a few hours. The present method represents a balance between speed and accuracy and it achieves a high rate of success in active site prediction.

We utilize a relatively high value of 20 for the internal dielectric constant of the protein. This helps to offset the tendency of Poisson–Boltzmann calculations to overestimate the
electrostatic interaction energies inside the protein. The value of 20 for the internal dielectric has been shown to yield pK_a values in best agreement with experiments (Antosiewicz et al., 1994). Indeed, predicted perturbed titration behavior is more likely to occur in an environment with low dielectric constant, although THEMATICS does sometimes predict clusters of residues on the surface that are highly solvent-exposed.

As discussed in the Results section, even though the current method is not yet fully optimized, the prediction of the active sites in enzymes is remarkably accurate. This same trend is observed on a larger set of proteins, where THEMATICS finds the correct active site in 91% of the enzymes studied. While THEMATICS also finds non-catalytic binding sites at least some of the time, more examples are needed in order to determine its success rate in the identification of non-catalytic interaction sites.

For the THEMATICS-predictive clusters of two or more residues, there are relatively few false positive clusters. Our method yields 1.7 predictive clusters per subunit. Other methods define clusters differently and so direct comparison is not straightforward, but can give some idea of the precision. For instance, the method of Gutteridge et al. (2003) predicts 7.2 clusters per protein. To achieve an overall accuracy of 83.6%, the method of Innis et al. (2004) predicts 1.98 clusters per protein.

Within a THEMATICS-predictive cluster that corresponds to an enzyme active site, it is possible to identify some of the residues as involved in catalysis and/or recognition, based on literature data. It is not known at this time whether the additional residues in the cluster play some role or whether they are simply false positives. The role of the ‘second shell’ residues is particularly intriguing: are they there to provide electrostatic interaction with the reactive residues and thus assist catalysis and/or recognition, or are they simply subjected to a strongly pH-dependent electric field because of their proximity to the reactive residues? Site-directed mutagenesis experiments are in progress to seek answers to this question.

Our method, although purely computational, has experimental basis. For example, Chivers et al. (1997) interpret the pK_a values of active site residues of *E.coli* thioredoxin by invoking microscopic pK_a’s, because a simple H–H equation does not fit the titration curves of C32 and of C35. Equations that take into account multiple protonation events are fitted to the titration data and two microscopic pK_a values for each Cys are obtained, that is, the titration curves are perturbed. This study (Chivers et al., 1997) concludes that the C32, C35 and D26 (proximal to C32 and C35), are involved in complex equilibria and that each titration curve cannot be explained by a single pK_a. A theoretical framework for decomposing a multi-component titration curve into a weighted sum of H–H curves has been described (Onufriev et al., 2001).

Success rates for THEMATICS are comparable to those of the best structure-to-function methods available (Gutteridge et al., 2003; Innis et al., 2004; Ota et al., 2003). There is one notable difference between our approach and these other methods: THEMATICS is completely independent of sequence alignments, conservation scores, and structural comparisons. Thus THEMATICS applies equally well to novel folds, proteins with few or no known close homologues, engineered structures, proteins of unknown function, as well as familiar folds and sequences.

THEMATICS complements well other methods for identification of interaction sites in proteins for applications in functional genomics.

**ACKNOWLEDGEMENTS**

J.K. thanks the Department of Chemistry and Chemical Biology and the Institute for Complex Scientific Software at Northeastern University for their hospitality during her sabbatical leave. This work was supported by the National Science Foundation under grant MCB-0135303 and by the Institute for Complex Scientific Software at Northeastern University.

**REFERENCES**


DeLaBarre,B., Thompson,P.R., Wright,G.D. and Berghuis,A.M. (2000) Crystal structures of homoserine dehydrogenase suggest a novel catalytic mechanism for oxidoreductases. *Nat. Struct. Biol.*, 7, 238–244.
Prediction of active sites for protein structures


