THEMATICS as a Tool for Functional Genomics

Mary Jo Ondrechen
mjo@neu.edu

Department of Chemistry and Chemical Biology and the Institute for Complex Scientific Software, Northeastern University, Boston, MA 02115, USA

Keywords: active site, function prediction, protein function, titration curves, THEMATICS

1 Introduction

Structural genomics efforts are determining in high throughput the three-dimensional structures of thousands of gene products. Many of these protein structures are of unknown function. The next major step after genome sequencing and protein structure determination is the determination of the function of the thousands of newly-discovered gene products. This presentation focuses on THEMATICS [3, 4], a simple computational predictor of information about the function of a protein. In particular, THEMATICS predicts the location of, and some chemical characteristics of, the active site of an enzyme. Application of THEMATICS requires only the structure of the subject protein. Thus THEMATICS works for proteins that have no similarity in sequence or in structure to any protein of known function. The accuracy of THEMATICS predictions is verified using proteins whose functions and mechanisms have been well studied experimentally.

2 Method

THEMATICS is a new technique based on established computational procedures. Most of our applications of THEMATICS to date have used experimentally-determined three-dimensional protein structures as input. We have also had success recently with homology model structures. For the structure one calculates the electrical potential function using a finite difference Poisson-Boltzmann procedure [1, 2, 5, 6]. Then a Monte Carlo procedure is used to obtain the mean net charge C, averaged over the ensemble of protein molecules, as a function of pH for each of the ionizable groups in the protein structure. The resulting C(pH) curves represent the theoretical microscopic titration curves for each ionizable species in the protein sequence. The majority of these C(pH) curves in a given protein structure have the typical sigmoidal shape, as predicted by the Henderson-Hasselbalch equation. However, a small fraction (3 - 7%) of the residues in an enzyme show unusual shapes in the predicted C(pH) function, such that partial protonation persists over a wide pH range.

The next part of a THEMATICS calculation is the classification of each of the predicted C(pH) curves into two categories: Typical or Perturbed. Originally, this was done by eye. Curves that were easily discernable as non-sigmoidal were labeled as perturbed and all of the others were considered typical. Any doubtful cases were placed in the “typical” category. Then, the three-dimensional protein structure was viewed with the perturbed residues labeled. A cluster of two or more perturbed residues in physical proximity was found to be a reliable predictor of the active sites of enzymes; we call this THEMATICS prediction a “positive” cluster.

More recently, we have used two different mathematical tests to classify the residues as either perturbed or typical. One of our methods involves fitting the computed C(pH) curves to a parametrized sigmoid function followed by statistical analysis of the curve features and of the fitting results. The second method classifies the computed C(pH) curves using machine learning techniques. Thus we now have fast, systematic, automated procedures for the classification part of the THEMATICS computation.
3 Results and Discussion

For most of the enzymes studied to date, the only cluster of two or more perturbed residues in physical proximity is in the known active site; thus THEMATICS correctly and unambiguously finds the active site. Some examples of such enzymes are: acetylcholine esterase, aldose reductase, apurinic apyrimidinic endonuclease, chorismate mutase, colicin E3, HIV-1 protease, pepsin, phosphomannose isomerase, subtilisin Carlsberg, and triosephosphate isomerase. In these cases, the positive cluster predicted by THEMATICS is in agreement with the evidence available in the literature for each enzyme about the location of the active site and the identification of catalytically active residues. The residues included in the THEMATICS positive cluster tend to be very highly conserved across species [3].

In a small fraction of cases, THEMATICS predicts two or more positive clusters. We are uncertain at this point whether some of these clusters are simply false positives or whether they have some function that has not yet been determined. For instance, for the enzyme phosphoglucone isomerase, THEMATICS predicts three positive clusters. One of these clusters is known to be the catalytically active site for isomerization. This enzyme is known to be multifunctional but it has not yet been determined where any of the other functionally important sites are located.

In reference [3], a human observer determined which C(pH) curves are perturbed. The observer’s visual criteria tend to be more conservative than the mathematical criteria. Therefore, the mathematical analyses generally give more complete lists of the ionizable residues in an active site. There have been two cases to date of enzymes where the human observer was unable to find a positive cluster in the structure: dihydrofolate reductase and human adenosine kinase. For both of these proteins, the mathematical analysis of the predicted titration curves correctly located the active site. The performance of the mathematical tests has been consistently equal to or better than that of the human observer.

THEMATICS has proved to be a successful method for the location and characterization of the active sites of enzymes with a variety of different folds and chemistries. Wider application continues.

4 Acknowledgments

The author thanks the National Science Foundation of the U.S.A. for support of this research through grants MCB-0074574 and MCB-0135303 and for support of the Molecular Modeling Facility and the Advanced Scientific Computation Center at Northeastern University. The support of the Institute for Complex Scientific Software at Northeastern University is also gratefully acknowledged.

References