ABSTRACT

Summary: We present an automated web server for partial order optimum likelihood (POOL), a machine learning application that combines computed electrostatic and geometric information for high-performance prediction of catalytic residues from 3D structures. Input features consist of THEMATICS electrostatics data and pocket information from ConCavity. THEMATICS measures deviation from typical, sigmoidal titration behavior to identify functionally important residues and ConCavity identifies binding pockets by analyzing the surface geometry of protein structures. Both THEMATICS and ConCavity (structure only) do not require the query protein to have any sequence or structure similarity to other proteins. Hence, POOL is applicable to proteins with novel folds and engineered proteins. As an additional option for cases where sequence homologues are available, users can include evolutionary information from INTREPID for enhanced accuracy in site prediction.

Availability: The web site is free and open to all users with no login requirements at http://www.pool.neu.edu.

Contact: m.ondrechen@neu.edu

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on February 5, 2012; revised on May 24, 2012; accepted on May 26, 2012

1 INTRODUCTION

Over the past decade, Structural Genomics (SG) projects have accumulated structural data for over 11,000 proteins, but most of them are of unknown or uncertain function. Thus, there is high demand for computational methods to predict function from structure. Computational site predictors provide valuable information for function annotation and they are also useful to guide and accelerate mechanistic, ligand-binding and protein engineering studies. A variety of sequence-based methods exist but these often suffer from poor precision compared with structure-based methods. Modern methods use both sequence and structural information to enhance the performance of active site prediction. Recently (Tong et al., 2009), we have reported a new machine learning method, partial order optimum likelihood (POOL), which uses input features from THEMATICS and outperforms many of the best prior methods. THEMATICS, for Theoretical Microscopic Anomalous Titration Curve Shapes (Wei et al., 2007), uses computed electrostatic properties and identifies functionally important ionizable residues based on their deviation from Henderson–Hasselbalch (H–H) titration behavior. In addition, we have also shown an integration of other structure and sequence features along with THEMATICS data can further boost the performance of POOL (Somarowthu et al., 2011b). With the combined input of electrostatic information from THEMATICS, evolutionary information from INTREPID (Sankararaman et al., 2009) and pocket information from ConCavity (Capra et al., 2009), POOL achieves 86.7, 92.5 and 93.8% recall of annotated functional residues at 5, 8 and 10% false-positive rates, respectively, on a standard test set of 100 unique, well-characterized enzymes (Somarowthu et al., 2011b). Using the top 8% of POOL-ranked residues, the functionally important residues are predicted with 89.8% recall and 92.8% specificity. The top 10% of the POOL-ranked residues yields a prediction with 93.5% recall and 90.9% specificity. Furthermore, information about the verification and performance of the POOL method is provided in the Supplementary Material.

Herein, we describe a web server for POOL. The user submits a protein structure and the server automatically performs THEMATICS and ConCavity calculations to obtain the input features for POOL calculations. INTREPID is a separate web server and hence cannot be integrated automatically but users are provided with the option to obtain the results from INTREPID and include them in the submission to the POOL server. The POOL server returns a list of all residues, rank ordered according to their probability of functional importance. The top-ranked residues constitute the functional site prediction.

2 METHODS

2.1 Overview

The POOL method has been described previously (Tong et al., 2009). The following input features are implemented in the current version (Somarowthu et al., 2011b).

2.2 Features

2.2.1 3D Structure-based features

THEMATICS electrostatics features: using only the structure of the query protein as input, THEMATICS (Wei et al., 2007) identifies residues that deviate from H–H titration behavior as active site residues. Briefly, the Poisson–Boltzmann equations are solved, followed by Monte Carlo sampling using HYBRID to compute theoretical microscopic titration curves for all the ionizable residues. Currently, POOL uses the fourth central moment and the theoretical buffer range, which are shown to be good metrics to measure the degree of deviation from H–H titration behavior (Somarowthu et al., 2011b). POOL also generates
environment features for all residues, based on the THEMATICS features of neighboring ionizable residues; thus, POOL predicts all 20 amino acid types.

ConCavity Pocket features: ConCavity (Capra et al., 2009) identifies binding pockets by analyzing the 3D structure of proteins and the method scores each residue on its likelihood of ligand binding, using surface geometric properties. These residue scores are used as one of the input features for POOL.

2.2.2 Sequence features The INTREPID server (Sankararaman et al., 2009) uses phylogenetic tree traversal to identify the functional residues in a protein. Given the sequence, INTREPID assigns a score to all residues according to functional importance. A residue with a higher score is thus predicted to be more functionally important than a residue with a lower score.

3 RESULTS
3.1 Input
The main input for a POOL/THEMATICS calculation is a protein 3D structure in PDB format. This is an advantage because the method needs no further information about a protein. The protein does not have to have any similarity in sequence or in structure to any other protein. However, one must make sure that the input structure is complete and of sufficient quality. For cases where a structure is not available, users can submit a homology model but the accuracy of the prediction depends on the quality of the model.

3.2 Processing
The input structure is pre-processed using YASARA (Krieger et al., 2002) to add any missing atoms. THEMATICS calculations are performed as described before (Wei et al., 2007). The electrical potential is computed by a Poisson–Boltzmann procedure; this is based on a set of atomic charges and molecular surface generated by the atomic radii assigned to the atoms in the input structure. These values are taken from a standard forcefield (CHARMM19). Thus, at the present, we are only able to include standard amino acids in a THEMATICS calculation. The current version of the system will delete any records from a PDB file marked as HETATM.

3.3 Submitting a job and checking on its progress
Jobs can be submitted either with a PDB ID or an uploaded structure file in PDB format. The status and results can be accessed in three ways: (1) at the time of submission, an HTML link is provided; when the job is finished, results appear on that web page. (2) a unique Job ID appears on the submission page; this Job ID can be used to check status or access results using the ‘check status’ window on the home page and (3) if the user wishes to provide an e-mail address, then results are e-mailed when the job is completed. The real time required for a POOL analysis depends on the size of the protein. A small protein of 100 residues takes about 1 min. A more typical-sized protein of 300 residues takes about 10 min. A large protein (1000 or more residues) can take hours. A more typical-sized protein of 300 residues takes about 10 min.

3.4 Output
The output HTML page contains (1) a Jmol Java applet with interactive 3D representation of protein structure and the top 10 residues in the predicted active site and (2) downloadable result file for offline analysis. Typically, the top 8–10% of the ranked residues in this result file are taken to be the predicted set of functionally important residues. The user should download the result file to obtain these top-ranked residues. Confidence information in the form of average recall and specificity rates is provided under the web server’s Help tab and also in the Supplementary Material.

4 APPLICATIONS
Examples of applications of POOL predictions of functionally important residues are in studies of the role of remote residues in enzyme function (Brodkin et al., 2011; Somarowthu et al., 2011a) and in the prediction of the function of SG proteins of unknown function (Han et al., 2011; Parasuram et al., 2010).

Funding: NSF under grants number MCB-0843603 and MCB-1158176 is gratefully acknowledged.

Conflict of Interest: none declared.

REFERENCES
Wei, Y. et al. (2007) Selective prediction of interaction sites in protein structures with THEMATICS. BMC Bioinformatics, 8, 119.