Published online 4 June 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W291–W296***
doi:10.1093/nar/gkn324

E1DS: catalytic site prediction based on 1D signatures of concurrent conservation

Ting-Ying Chien[1], Darby Tien-Hao Chang[2,*], Chien-Yu Chen[3], Yi-Zhong Weng[1] and Chen-Ming Hsu[4]

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, [2]Department of Electrical Engineering, National Cheng Kung University, Tainan 701, [3]Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei 106 and [4]Department of Computer Science and Engineering, Yuan Ze University, Chung-Li 320, Taiwan, ROC

Received February 3, 2008; Revised April 25, 2008; Accepted May 7, 2008

ABSTRACT

Large-scale automatic annotation of protein sequences remains challenging in postgenomics era. E1DS is designed for annotating enzyme sequences based on a repository of 1D signatures. The employed sequence signatures are derived using a novel pattern mining approach that discovers long motifs consisted of several sequential blocks (conserved segments). Each of the sequential blocks is considerably conserved among the protein members of an EC group. Moreover, a signature includes at least three sequential blocks that are concurrently conserved, i.e. frequently observed together in sequences. In other words, a sequence signature is consisted of residues from multiple regions of the protein sequence, which echoes the observation that an enzyme catalytic site is usually constituted of residues that are largely separated in the sequence. E1DS currently contains 5421 sequence signatures that in total cover 932 4-digital EC numbers. E1DS is evaluated based on a collection of enzymes with catalytic sites annotated in Catalytic Site Atlas. When compared to the famous pattern database PROSITE, predictions based on E1DS signatures are considered more sensitive in identifying catalytic sites and the involved residues. E1DS is available at http://e1ds.ee.ncku.edu.tw/ and a mirror site can be found at http://e1ds.csbb.ntu.edu.tw/.

INTRODUCTION

Recent large-scale genome projects have accumulated abundant sequence and structure data with unknown functions, which raises a large demand of automated function inference using computational tools (1–3). Identifying important residues of protein sequences is one of the most important steps in function inference, since many studies have shown that functionally important residues can usually serve as good signatures for function prediction (4–8). There has been many efforts on predicting functional sites based on structural analyses (7,9–15). Jones and Thornton (11) provided a comprehensive review of these methods. However, computational tools that utilize protein structural information are limited, since there is a great quantity of protein sequences without experimentally determined or computationally modeled structure available for learning. This emerges alternative approaches that utilize the sequence information alone. It has been shown that the sequence conservation property so far serves as one of the most powerful indices for detecting functionally important residues in proteins (16–18). Moreover, conservation information is found to be more effective on predicting catalytic sites and residues near ligands than the residues in protein–protein interfaces (18).

A widely used approach for estimating residue conservation is multiple sequence alignment (MSA). Many scoring schemes have been proposed (18,19). When incorporated with phylogenetic information, the evolutionary trace (ET) method identifies sites critical to protein functions by detecting important mutations across subfamilies (20). Another well-known method to identify function-related residues is motif discovery based on a set of homologous sequences (8,21,22). These motif discovery methods usually find short amino acid stretches represented as consecutive regular expressions or profiles. However, short patterns are considered less complete and not specific enough in characterizing the protein function (1) and tend to result in false positives when they are used to detect important residues on sequences (16). Nevertheless, it is favorable if we can find longer sequence motifs that cover the binding sites as complete as possible.

Several databases have been proposed for characterizing important residues of enzymes, most based on sequence

*To whom correspondence should be addressed. Tel: +886 6 2757575 62421; Fax: +886 6 2345482; Email: darby@ee.ncku.edu.tw

and structure conservation and some from literatures (10,13,23). E1DS provides an alternative way to derive useful information about enzyme binding regions by a novel pattern mining algorithm that discovers long sequence motifs (24). The performance evaluation conducted in this study shows that E1DS is capable of delivering favorable sensitivity rates in detecting catalytic sites and residues without using structure information.

METHODS

Figure 1 shows the workflow of E1DS. In 'Signature Construction', a signature database is constructed to expedite the prediction process when a protein is submitted. Then the most appropriate signature is chosen for function inference. E1DS reports the positions of the query sequence that are matched by the signature as the functionally important residues. In this section, we will first describe how the signature database is constructed, including the data collection process and the employed pattern mining algorithm. After that, we illustrate the signature matching procedure that aims at predicting the catalytic sites of the query protein.

Data collection for signature construction

E1DS signatures are constructed based on the protein sequences from Swiss-Prot database (25) release 52.0. A protein is selected as training data of E1DS if it is annotated with exactly one 4-digital EC number. Such sequences are grouped by their EC numbers. The sequence signatures of each EC group are generated using the pattern mining method described as below.

Pattern mining for generating 1D signatures

Sequential pattern mining has been widely used in identifying sequence motifs from biological data (26–28). The derived patterns usually highlight important positions that are conserved either for structural or for functional purposes. For proteins, conserved residues with respect to protein functions are often scattered in the primary structures. This challenges the mining algorithms to distinguish signals (true motifs) from noises. It is observed that insertion and deletion of residues are often found in loose loops, but seldom in the regions close to functional sites of proteins. In this regard, we recently proposed a mining algorithm that considers two types of gap constraints for efficiently discovering conserved regions. These regions are simultaneously conserved during evolution but separated by large wildcard regions with irregular lengths (24). The proposed algorithm, named WildSpan, employs a two-phase mining strategy, where the first step grows sequential blocks and the second step concatenates these conserved blocks with flexible gaps, i.e. successive wildcards of different lengths. WildSpan was first used in the web server MAGIIC-PRO for detecting functional signatures of a query protein along with its homologs (27).

When constructing the signature database of E1DS, the WildSpan package is employed by an iteratively mining strategy that aims at collecting a set of satisfied signatures to serve as diagnostic patterns for each EC group. This is denoted as the 'Signature Mining' procedure in Figure 1. In the first run of WildSpan, the sequence with median length is selected from all the members of the target EC group as the reference protein. At the end of the first mining stage, the signature that matches the most member sequences is picked. If the picked signature is observed in all the members of the target EC, the mining process stops. Otherwise, another median-length sequence is selected from the excluded member sequences as the reference protein for the next call of WildSpan. Here the excluded sequences are those EC members that are not matched by the picked signature (i.e. the picked signature is not present in each of the excluded sequences). In the second run, the signature that matches the most excluded sequences derived in the first run will be picked. This procedure is repeated until the set of picked signatures cover all the members of the target EC or no more signatures can be found.

Prediction of catalytic sites

Given an amino acid sequence, E1DS first tries to identify the possible EC group to which it belongs. This is achieved by invoking three iterations of PSI-BLAST (29) on the query protein against all the training sequences of E1DS.

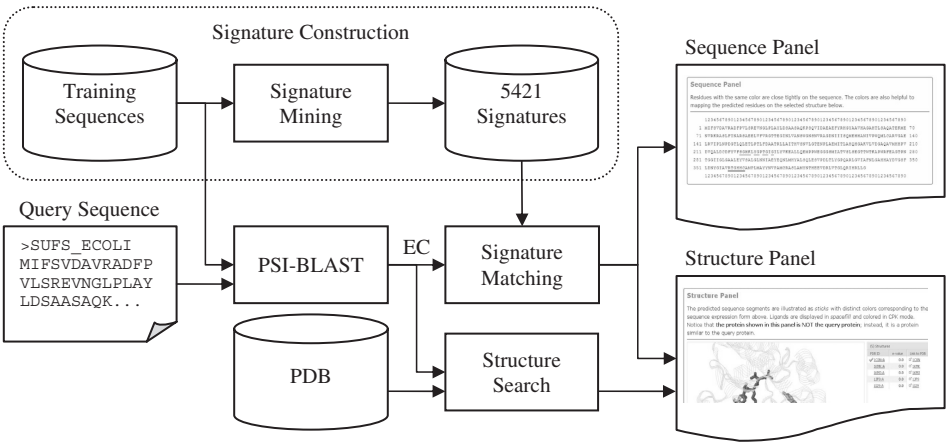

Figure 1. Workflow of the analysis procedures incorporated in E1DS. In this figure, procedures in the 'Signature Construction' are performed only once, while other procedures are performed every time when a new query comes.

The other two important parameter settings for PSI-BLAST, the cutting threshold for output (*e*) and the threshold for inclusion in multipass model (*h*), are set to *e*-values of 10^{-3} and 2×10^{-3}, respectively, following the suggestions of a previous study (30). Among the homology list found by PSI-BLAST, the 4-digital EC number of the training enzyme with the highest bit score is chosen. Since each training sequence has exactly one 4-digital EC number as have been described, one and only one EC number, called the target EC, can be chosen without ambiguity for further signature matching and prediction process.

For each signature in the target EC, ClustalW (31) is employed to align the query sequence with the reference sequence of the signature. This is denoted as the 'Signature Matching' procedure in Figure 1. Figure 2 shows an example of the alignment delivered by ClustalW, in which '∗' indicates identical matches, ':' indicates conserved substitutions and '.' indicates semiconserved substitutions in the alignment. On the reference sequence of the signature, we define that one residue is 'covered' by the signature as long as it can be matched by the sequential blocks in the signature. In Figure 2, the signature shown has two blocks written in regular expression form, 'S-x-H-K-x-x-x-P-x-G-x-G' and 'A-x-x-x-G-x-x-C'. These two blocks are two conserved regions commonly shared by the member sequences of EC 2.8.1.7, where the capital letters stand for residues that are highly conserved and the symbol 'x' is the location where mutations are observed within the EC group. The positions matched by 'x' are weighted equally as those matched with a capital letter, since sometimes important residues are specific only to subfamilies. In Figure 2, the segments of the reference sequence covered by the signature are highlighted in yellow. For the query sequence, a residue is covered by a signature if (i) it is aligned to a residue of the reference sequence with a '∗', ':' or '.' symbol in the consensus line of ClustalW; (ii) the aligned residue of the reference sequence is covered by the signature and (iii) it is not an Ala, Ile, Leu, Pro or Val. Finally, the signature in the target EC that covers the most residues of the query sequence is chosen to make the prediction, and the covered residues of the query sequence by the chosen signature are the predicted residues. In Figure 2, the residues colored in green are reported as functionally important residues in this example.

In case the suggested EC number does not fit the expectation of the users, they can manually select other EC numbers through a candidate list collected from the other homologs found by PSI-BLAST. When a different EC number is specified, E1DS will reperform the prediction process described to adapt the prediction results. This option is, in particular, useful when multiple functions are investigated.

WEB INTERFACE

To use E1DS, the user needs to input the amino acid sequence of the query protein in one-letter codes (FASTA format). Alternatively, UniProt (32) accession numbers and entry names or PDB IDs with chain numbers specified are allowed. After the 'Signature Matching' process, the users can take a look at the predicted catalytic residues highlighted on the query sequence in the region of 'Sequence Panel'. In addition, E1DS will try to collect PDB structures that are similar to the query sequence. This is denoted as the 'Structure Search' procedure in Figure 1. If there are available PDB structures that are similar to the query sequence, a structure panel will be activated automatically as shown in Figure 3. There are two subregions in the E1DS structure panel. The left side is a Jmol plug-in (available at http://www.jmol.org/) for rendering a selected PDB structure. The right side lists available PDB structures and provides an interactive interface for selecting the PDB chain rendered in Jmol.

PERFORMANCE

We evaluate the performance of E1DS using a collection of known catalytic sites. The performance of E1DS is reported in terms of the number of catalytic sites and the number of catalytic residues that can be predicted. The E1DS signatures are compared with existing PROSITE patterns (8) which are designed for characterizing protein functions. Furthermore, we compare the performance of E1DS with a structure-based approach, THEMATICS (15).

Datasets

The catalytic site information is obtained from the Catalytic Site Atlas (CSA) (23), a manually curated database documenting enzyme active sites and catalytic

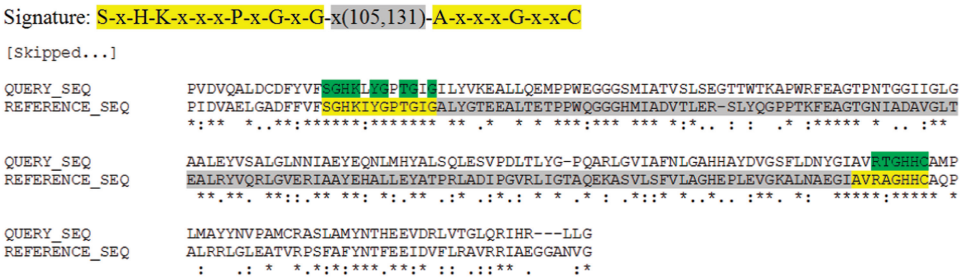

Figure 2. An example to demonstrate the 'Signature Matching' procedure adopted by E1DS. Yellow residues on the reference sequence are 'covered' by the signature. On the query sequence, green residues are those residues aligned with the covered residues of the reference sequence and are not an Ala, Ile, Leu, Pro or Val. The residues marked as green are predicted as functionally important residues of the query sequence based on the signature shown.

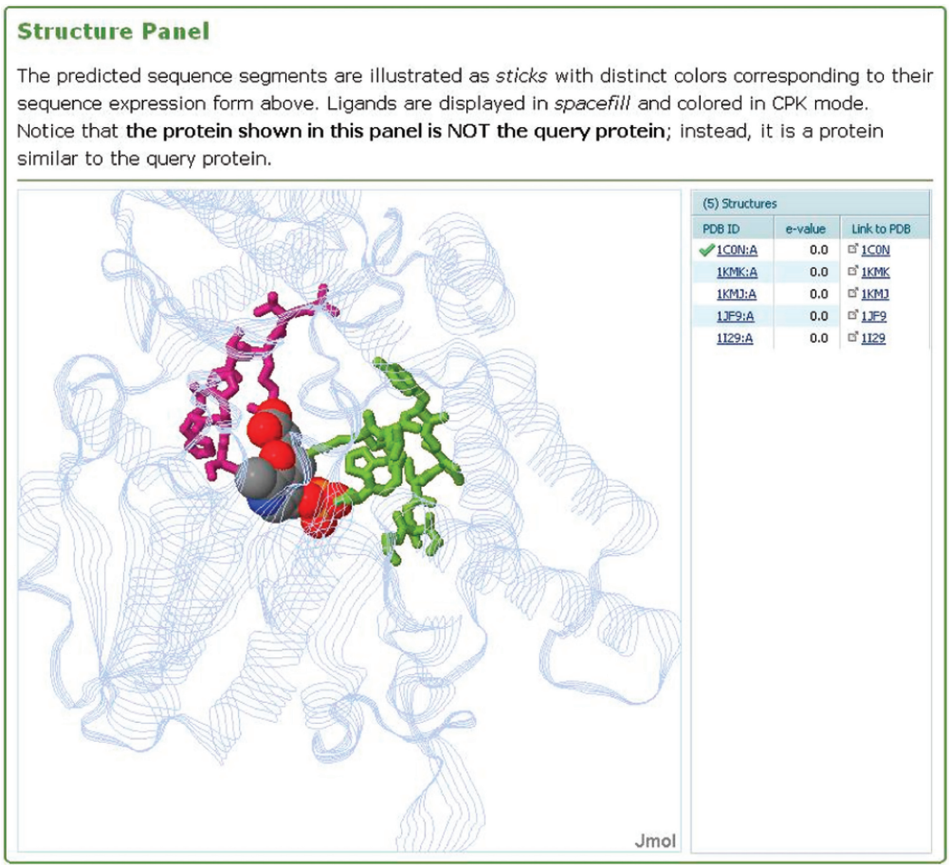

Figure 3. The structure panel of E1DS that provides 3D view of the signature. The list control sitting at the right side provides an interactive interface to select the protein structure for rendering.

residues derived from literatures. In the CSA version of 2.2.8, there are 1882 hand-annotated entries as well as 67 731 homologous entries found by PSI-BLAST alignment (*e*-value <0.00005 to one of the hand-annotated entries). Here, we consider only the hand-annotated entries, since the prediction performance of E1DS on homologous entries of CSA can significantly be affected by a large amount of homologies originated from a small proportion of hand-annotated entries. Sites associated with multiple 4-digital EC numbers or with an obsolete PDB ID (in the PDB release of 19 June 2007) are also excluded.

In this way, a dataset of 831 catalytic sites is created, named CSA831 in the following descriptions. The CSA831 dataset contains 2573 catalytic residues and spans 362 4-digital EC numbers. We observe that for some ECs we do not have E1DS signatures due to lacking sufficient homologs in the pattern mining stage. In the CSA831 dataset, there are 570 sites from 237 ECs that have E1DS signatures and 413 sites from 186 ECs that have PROSITE patterns. To alleviate the interference owing to lack of signatures/patterns, we define the 346 sites that have both E1DS signatures and PROSITE patterns as the second test set, CSA346. The CSA346 dataset spans 146 4-digital EC numbers.

The third test set is extracted from CatRes database (33), containing 178 proteins. The one with PDB code 1A6F is excluded because it has no annotated catalytic residue in CatRes. The resultant set contains 612 catalytic residues and is named CatRes177 that spans 173 4-digital EC numbers. Again, to investigate the performance of E1DS when sequence signatures are available, CatRes177 is refined as set CatRes121 that contains test cases from ECs with at least one E1DS signature. The CatRes121 dataset spans 117 4-digital EC numbers.

Evaluation

We follow some measures employed in previous studies to evaluate the performance of catalytic site and catalytic residue prediction. For catalytic site prediction, we adopt three measures defined in THEMATICS. One prediction is considered as 'correct' if $\geq 50\%$ residues of the target site have been captured by the predictor. One prediction is considered as 'partially correct' if at least one catalytic residue but $<50\%$ residues of the target site have been captured by the predictor. The total success rate is the number of 'correct' plus 'partially correct' predictions divided by the number of test sites. For catalytic residue prediction, two commonly used measures, *sensitivity* and *specificity*, are reported along with the average number of residues predicted. The *sensitivity* is defined as the number of true positives (catalytic residue that is correctly predicted) over all catalytic residues, while the *specificity* is defined as the number of true negatives (non-catalytic residue that is not predicted as a catalytic residue) over all non-catalytic residues.

Table 1. Performance statistics for E1DS and PROSITE on CSA831 and CSA346

	E1DS		PROSITE	
	CSA831	CSA346	CSA831	CSA346
Prediction of catalytic site				
Correct predictions (%)	35.5	51.7	18.9	38.2
Partially correct predictions (%)	14.1	18.2	14.8	31.2
Total success rate (%)	49.6	69.9	33.7	69.4
Prediction of catalytic residue				
Sensitivity (%)	30.0	40.9	16.3	31.6
Specificity (%)	96.7	95.8	98.6	97.2
No. of predicted residues in average	12.7	15.9	5.6	11.0

As shown in Table 1, E1DS delivers the total success rate of 49.6%, ~16% higher than PROSITE on the CSA831 dataset. Moreover, >70% (35.5% divided by 49.6%) of successful predictions of E1DS are correct. This ratio is much higher than the predictions of PROSITE in which correct predictions account for ~56%. It suggests that E1DS signatures are capable of not only identifying more catalytic sites but also providing more comprehensive information of predicted catalytic sites. Similarly for all 2573 catalytic residues in the CSA831 dataset, E1DS successfully captures 30.0% while PROSITE only captures 16.3% catalytic residues. However, PROSITE has slight advantage over E1DS (98.6% versus 96.7%) in terms of *specificity*. This result is reasonable since E1DS signatures are constructed to characterize the function regions as complete as possible, while PROSITE patterns are designed for function inference to achieve both high sensitivity and specificity when performing function prediction. For a single chain, E1DS reports 12.7 putative catalytic residues while PROSITE reports 5.6 putative catalytic residues in average.

As described in the previous section, the CSA831 dataset contain sites that have no E1DS signatures associated with the desired 4-digital EC number. The CSA346 column in Table 1 focuses on those catalytic sites that have both E1DS signatures and PROSITE patterns. In this subset, both E1DS and PROSITE improve the performance in a significant degree. The comparison indicates that the better performance of E1DS in total success rate on the CSA831 dataset might be due to its higher signature coverage of ECs.

Table 2 shows the performance of E1DS on CatRes177. E1DS delivers 41.8% correct and 15.2% partially correct predictions at the site level and 32.9% *sensitivity* and 96.9% *specificity* at the residue level. These statistics are similar to the performance on the CSA831 dataset. THEMATICS was evaluated using the same 178 proteins from CatRes database (15). However, nine sites were excluded because of poor structure quality and/or other structural issues. According to records reported in the paper of THEMATICS, it achieves 48.5% correct and 29.0% partially correct predictions at the site level and 41.1% *sensitivity* at the residue level, when the *Z*-score cutoff was set to 1.0. Among the 177 tested catalytic sites, E1DS only made predictions on 121 sites. E1DS failed to produce predictions on the remaining sites due to lacking of signatures for those EC groups to which those catalytic sites belong. This is one of the limitations of homology-based approaches, and it is expected to be alleviated as the number of homologs increases in sequence databases. With respect to the 121 catalytic sites for which E1DS can find applicable signatures to make predictions, 45.0% correct and 19.2% partially correct predictions at the site level and 39.8% *specificity* and 95.8% *sensitivity* at the residue level can be achieved. In summary, the results shown in Tables 1 and 2 reveal that E1DS signatures provide useful information in the analysis of functionally important residues as long as some homologs of the query sequences are available.

Table 2. Performance statistics for E1DS on CatRes177 and CatRes121

	CatRes177	CatRes121
Prediction of catalytic site		
Correct predictions (%)	41.8	45.0
Partially correct predictions (%)	15.2	19.2
Total success rate (%)	57.0	64.2
Prediction of catalytic residue		
Sensitivity (%)	32.9	39.8
Specificity (%)	96.9	95.8

CONCLUSION AND FUTURE PERSPECTIVES

In this article, we propose the E1DS server that aims at predicting catalytic residues of enzymes from sequence information alone. The experimental results reveal that the precalculated E1DS signatures are capable of providing useful information in the analysis of functional important residues as long as some homologs of the query sequences are available. E1DS will be regularly updated based on the newest release of Swiss-Prot and PDB databases. Furthermore, we would exploit more sequence databases to construct sequence signatures in the future.

ACKNOWLEDGEMENTS

The authors would like to thank National Science Council of Republic of China, Taiwan, for the financial support under the contracts: NSC 96-2627-B-002-003-,

95-3114-P-002-005-Y, 95-2221-E-002-274-MY2, 96-2320-B-006-027-MY2 and 96-2221-E-006-232-MY2. We also thank Dr Shou-De Lin for valuable comments. Funding to pay the Open Access publication charges for this article was provided by National Science Council of Republic of China, Taiwan.

Conflict of interest statement. None declared.

REFERENCES

1. Friedberg,I. (2006) Automated protein function prediction—the genomic challenge. *Brief. Bioinform.*, **7**, 225–242.
2. Chandonia,J.M. and Brenner,S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.
3. Watson,J.D., Laskowski,R.A. and Thornton,J.M. (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.*, **15**, 275–284.
4. George,R.A., Spriggs,R.V., Bartlett,G.J., Gutteridge,A., MacArthur,M.W., Porter,C.T., Al-Lazikani,B., Thornton,J.M. and Swindells,M.B. (2005) Effective function annotation through catalytic residue conservation. *Proc. Natl Acad. Sci. USA*, **102**, 12299–12304.
5. Tian,W.D., Arakaki,A.K. and Skolnick,J. (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res.*, **32**, 6226–6239.
6. Kasuya,A. and Thornton,J.M. (1999) Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.*, **286**, 1673–1691.
7. Torrance,J.W., Bartlett,G.J., Porter,C.T. and Thornton,J.M. (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J. Mol. Biol.*, **347**, 565–581.
8. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., De Castro,E., Langendijk-Genevaux,P.S., Pagni,M. and Sigrist,C.J.A. (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, D227–D230.
9. Cheng,G., Qian,B., Samudrala,R. and Baker,D. (2005) 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.
10. Sheu,S.H., Lancia,D.R., Clodfelter,K.H., Landon,M.R. and Vajda,S. (2005) PRECISE: a database of predicted and consensus interaction sites in enzymes. *Nucleic Acids Res.*, **33**, D206–D211.
11. Jones,S. and Thornton,J.M. (2004) Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.*, **8**, 3–7.
12. Innis,C.A. (2007) siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res.*, **35**, W489–W494.
13. Meng,E.C., Polacco,B.J. and Babbitt,P.C. (2004) Superfamily active site templates. *Proteins-Struct. Funct. Bioinform.*, **55**, 962–976.
14. Dundas,J., Ouyang,Z., Tseng,J., Binkowski,A., Turpaz,Y. and Liang,J. (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.*, **34**, W116–W118.
15. Wei,Y., Ko,J., Murga,L.F. and Ondrechen,M.J. (2007) Selective prediction of interaction sites in protein structures with THEMATICS. *BMC Bioinformatics*, **8**, 119.
16. La,D., Sutch,B. and Livesay,D.R. (2005) Predicting protein functional sites with phylogenetic motifs. *Proteins-Struct. Funct. Bioinform.*, **58**, 309–320.
17. Petrova,N.V. and Wu,C.H. (2006) Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. *BMC Bioinformatics*, **7**, 312.
18. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
19. Valdar,W.S.J. (2002) Scoring residue conservation. *Proteins-Struct. Funct. Genet.*, **48**, 227–241.
20. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
21. Liu,A.H., Zhang,X.M., Stolovitzky,G.A., Califano,A. and Firestein,S.J. (2003) Motif-based construction of a functional map for mammalian olfactory receptors. *Genomics*, **81**, 443–456.
22. Puntervoll,P., Linding,R., Gemund,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M.A., Ausiello,G., Brannetti,B., Costantini,A. *et al.* (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630.
23. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.
24. Hsu,C.-M. (2007) WildSpan: discovery of discontinuous functional motifs from biological sequences using constraint-based sequential pattern mining. *Ph.D. Thesis*. Yuan Ze University, Taoyuan, Taiwan.
25. Bairoch,A., Bougueleret,L., Altairac,S., Amendolia,V., Auchincloss,A., Puy,G.A., Axelsen,K., Baratin,D., Blatter,M.C., Boeckmann,B. *et al.* (2007) The universal protein resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
26. Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.
27. Hsu,C.-M., Chen,C.-Y. and Liu,B.-J. (2006) MAGIIC-PRO: detecting functional signatures by efficient discovery of long patterns in protein sequences. *Nucleic Acids Res.*, **34**, W356–W361.
28. Jonassen,I. (1997) Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.*, **13**, 509–522.
29. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
30. Jones,D.T. and Swindells,M.B. (2002) Getting the most from PSI-BLAST. *Trends Biochem. Sci.*, **27**, 161–164.
31. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Clustal-W—Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
32. Bairoch,A., Bougueleret,L., Altairac,S., Amendolia,V., Auchincloss,A., Puy,G.A., Axelsen,K., Baratin,D., Blatter,M.C., Boeckmann,B. *et al.* (2008) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
33. Bartlett,G.J., Porter,C.T., Borkakoti,N. and Thornton,J.M. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.