World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:3, No:5, 2009

Introducing Sequence-Order Constraint into Prediction of Protein Binding Sites with Automatically Extracted Templates

Yi-Zhong Weng, Chien-Kang Huang, Yu-Feng Huang, Chi-Yuan Yu, and Darby Tien-Hao Chang

***Abstract*—**Search for a tertiary substructure that geometrically matches the 3D pattern of the binding site of a well-studied protein provides a solution to predict protein functions. In our previous work, a web server has been built to predict protein-ligand binding sites based on automatically extracted templates. However, a drawback of such templates is that the web server was prone to resulting in many false positive matches. In this study, we present a sequence-order constraint to reduce the false positive matches of using automatically extracted templates to predict protein-ligand binding sites. The binding site predictor comprises i) an automatically constructed template library and ii) a local structure alignment algorithm for querying the library. The sequence-order constraint is employed to identify the inconsistency between the local regions of the query protein and the templates. Experimental results reveal that the sequence-order constraint can largely reduce the false positive matches and is effective for template-based binding site prediction.

***Keywords*—**Protein structure, binding site, functional prediction.

I. Introduction

FUNCTION prediction of new proteins is a critical issue in life science [1]. As protein structures become increasingly available, structure analysis of proteins has been widely adopted to extract information alongside protein sequence analysis [2]. In this respect, search for a tertiary substructure that geometrically matches the 3D pattern of the binding site of a well-studied protein provides a solution to predict protein functions [3-6]. Thornton and colleagues constructed a well-know library of structural templates based on expert knowledge and literature searches [7], which were then equipped with the JESS [8] template matching algorithm to construct the Catalytic Site Search (CSS) web server for catalytic site prediction [6]. In our previous work, we showed that a manually curated template library may contain insufficient quantity of entries for making accurate prediction, and developed a web server (Protemot) to predict protein-ligand binding sites based on automatically extracted geometrical templates [9]. However, template matching based on automatically extracted templates is prone to returning many false positive matches [2]. According to the statistics reported in Protemot, it resulted in 564 false positive matches among 972 predictions.

In this study, we present a mechanism to reduce the false positive matches of using automatically extracted templates to predict protein-ligand binding sites. This is done by introducing a sequence-order constraint into the template matching algorithm of Protemot. An experiment of predicting binding location is conducted to evaluate the distance between the actual and the predicted protein-ligand binding regions by the present method. The results indicate that the introduced sequence-order constraint can help to locate the protein binding sites. Additionally, the present method is also evaluated by an experiment of predicting enzyme class. This experiment is conducted to compare the performances of Protemot with and without the proposed sequence-order constraint. Experimental results reveal that the introduced sequence-order constraint can largely reduce the false positive matches. Consequently, the present method can reduce false positive matches of using automatically extracted templates, which is essential for creating and maintaining a comprehensive template library that timely accommodates to the new release of Protein Data Bank (PDB) [10] as the number of entries continues to grow rapidly.

Yi-Zhong Weng is with the department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan, R.O.C. (e-mail: d95032@csie.ntu.edu.tw).

Chien-Kang Huang is with the department of Engineering Science and Ocean Engineering, National Taiwan University, Taipei 106, Taiwan, R.O.C. (e-mail: ckhuang@ntu.edu.tw).

Yu-Feng Huang is with the graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 106, Taiwan, R.O.C. (e-mail: yfhuang@csie.ntu.edu.tw).

Darby Tien-Hao Chang is with the department of Electrical Engineering, National Cheng Kung University, Tainan 70101, Taiwan, R.O.C. (corresponding author to provide phone: +886 6 2757575 EXT 62421; fax: +886 6 2345482; email: darby@ee.ncku.edu.tw).

II. Materials and Methods

This section first describes the automatic mechanism for extracting structural templates of protein-ligand binding sites from PDB, and the local structure alignment algorithm employed to query the template library. The introduced constraints of the alignment procedure are then described in detail.

A. Template Construction

This study builds the template library from protein-ligand complexes in PDB. Each template comprises a number of contact residues. A residue in the crystal structure of a protein-ligand complex of PDB is said to be one of the contact

residues, if it contains one or more heavy atoms that are less than 6.5Å away from the heavy atoms of the ligand [11, 12]. The template extraction process employs several filters to improve the quality of extracted templates. First, PDBsum [13] is queried to obtain all ligand names within each PDB file. A single PDB file containing multiple ligands (e.g. 1A82) results in multiple templates. Subsequently, 'pseudo-ligands', e.g. counter ions, metal ions or molecules used for setting up proper crystallization conditions, are filtered out. Finally, templates with fewer than three contact residues are filtered out.

The version of PDB released on June 19, 2007 contains 12825 protein-ligand complexes out of 44476 crystal structures. The automatic extraction mechanism identified 3507 templates of protein-ligand binding sites from 2947 PDB files. Of these 2947 crystal structures, 476 can derive multiple templates.

B. Template Matching

Fig. 1 depicts the workflow of the present method to match the tertiary structure of the query protein with the templates in the library. A major difference between this work and Protemot is the enhanced screening process, marked with an asterisk in Fig. 1, which is intended to reduce the false positive matches from automatically extracted templates.

The procedure shown in Fig. 1 begins with a cavity identification process to extract those residues in the proximity of a cavity of the query protein. The cavity identification process is based on the recently proposed kernel density estimation algorithm (RVKDE) [3, 14, 15]. The cavity identification process effectively reduces the number of coordinate systems to be examined in the next step. Thus, the cavity identification process in the alignment algorithm narrows down its search on the residues identified as in the proximity of a cavity, instead of all residues in the query protein.

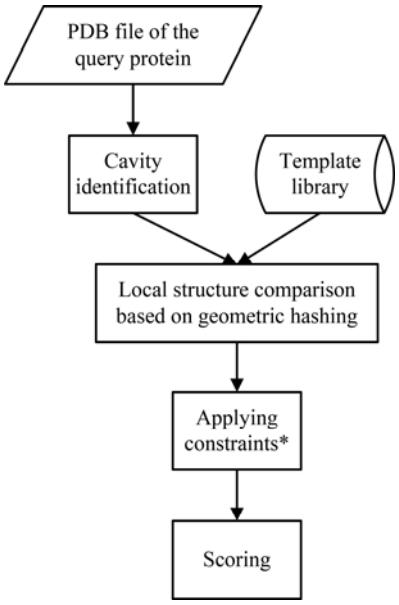

Fig. 1 The workflow of the template matching process incorporated in the present method. The sequence-order constraint proposed in this study is implemented in the step marked with an asterisk

The geometric hashing algorithm in computer graphics [16] is then invoked to compare the crucial substructures in the proximity of a cavity of the query protein with the templates in the library. The alignment frames examined by the geometric hashing algorithm are defined by the two backbone bonds connected to the alpha carbon of each residue. This definition has been widely used when applying geometric hashing algorithm on protein structure alignments [17-19]. This study regards a residue in the query protein as being successfully aligned with one residue in the template, if the distance between this pair of residues in the alignment frame of the coordinate system is $\leq$ 3Å. The likelihood of residue substitution is also considered: a pair of residues must correspond to an entry in the PAM 250 matrix [20, 21] that is ≥ 2. Accordingly, the time complexity of the alignment process is $O(n_1 n_2 (n_1 + n_2))$, where n_1 denotes the number of residues in the template, and n_2 denotes the number of residues in the query protein that pass the cavity identification process.

After the comparison step, several constraints are applied before the conventional scoring process. These constraints are elaborated in the next section. Only the coordinate systems that passed all the constraints in screening can participate in scoring process. We adopt the TM-score [22], a measure of the similarity of topologies of two protein structures as the scoring function. TM-score is more sensitive than the root-mean-square-deviation (RMSD) of the aligned alpha carbons in accessing the quality of structure alignment.

C. Matching Constraints

The design of Protemot includes three constraints to alleviate the false positive matches of using automatically extracted templates: i) ratio constraint, ii) RMSD constraint and iii) orientation constraint. Additionally, a new constraint, iv) sequence-order constraint, is introduced in this study. The first constraint ensures that $\geq$ 20% residues in the template are successfully aligned with the residues in the query protein. The second constraint states that the RMSD of the aligned alpha carbons must be $\leq$ 2.5Å. The third constraint requires that the angle between the opening directions of the predicted binding site and the template must be $\leq$ 30 degrees.

The fourth constraint, which is the major improvement of this work over Protemot, is based on the sequence information within the local region identified by local structure alignment. The local regions by structure alignment may have aligned residues with inconsistent sequence orders. Fig. 2 illustrates an example in which a template is aligned with a query protein with five aligned residue pairs, but with at most three aligned residue pairs sequentially in order. In this case, allowing the order mismatches among aligned residues is one advantage of structure alignment approaches based on geometric hashing over those based on dynamic programming. However, too many order mismatches may imply that this alignment is less preferred, especially when another candidate alignment with comparable score exists.

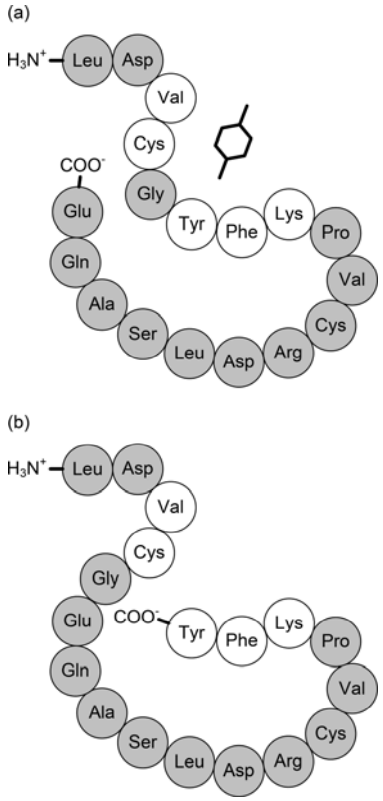

Fig. 2 An example in which a template (a) is aligned onto a query protein (b) with five aligned residue pairs, of which only three are aligned correctly while considering the sequence order from the N-terminal to the C-terminal in the template and the query protein

In Fig. 2, assume that the aligned residues of the query protein are $s = \{\text{Val}, \text{Cys}, \text{Tyr}, \text{Phe}, \text{Lys}\}$, based on the sequence order from the N-terminal to the C-terminal of the template. Then we will first re-order s to $s' = \{\text{Val}, \text{Cys}, \text{Lys}, \text{Phe}, \text{Tyr}\}$ based on the sequence order from the N-terminal to the C-terminal of the query protein. The next step is to employ the longest common subsequence (LCS) algorithm [23] to obtain the longest common subsequence between s and s', which could be $\{\text{Val}, \text{Cys}, \text{Tyr}\}$, $\{\text{Val}, \text{Cys}, \text{Tyr}\}$, or $\{\text{Val}, \text{Cys}, \text{Lys}\}$ in Fig. 2 In our implementation, the ratio of the length of the longest common subsequence to s is defined as sequence-order-conservation (SOC) ratio. Based on SOC ratio, another measure is defined as follows:

$$\text{sRMSD} = \sqrt{\text{SOC} \times \text{RMSD}^2 + (1 - \text{SOC}) \times \text{MAX\_RMSD}^2}$$

where MAX_RMSD denotes the maximum distance between two aligned alpha carbons (*i.e.* the maximum of RMSD) and MAX_RMSD = 3Å in this study; sRMSD represents a heuristic measure to penalize order mismatches by assigning them larger RMSD values. The fourth constraint ensures that SOC ≥ 0.37 and sRMSD ≤ 1.6Å of the aligned alpha carbons.

III. Results and Discussion

Two experiments are conducted to evaluate the effects of introducing the sequence-order constraint. The first experiment predicts the binding location. This experiment is relatively small-scale, because it requires both the training and testing proteins to have at least one protein-ligand complex available in PDB. The second experiment follows the experimental design in the Protemot paper and provides a large-scale evaluation of the performances of Protemot with and without the proposed sequence-order constraint. Since the sequence-order constraint is designed to reduce the false positive matches owing to the automatically extracted templates, it is of interests to include the CSS, of which the template library is manually curated, in this experiment. Finally, two testing examples are presented to illustrate how the proposed sequence-order constraint helps the prediction.

A. Prediction of the Location of Protein-Ligand Binding Sites

1) Experimental design

In the first experiment, all protein-ligand complexes in PDB are grouped into pairs according to their ligand names. From another point of view, with our automatic extraction procedure (see 'Template Construction' for details), the paired complexes can generate two structural templates associated with the same ligand. This experiment examines 1429 pairs of protein-ligand complexes from the version of PDB released on June 19, 2007. These pairs are collected by first scanning all the protein-ligand complexes in PDB. We partition these protein-ligand complexes into 1429 groups according to their ligand names. Finally, a pair is randomly selected from each of the 1429 groups. In summary, the set of 1429 pairs covers all ligands with at least two protein-ligand complexes in PDB spanning 522 SCOP families and 376 SCOP superfamilies [24].

Assume that a pair contains two proteins *A* and *B*, where $\{a_1, a_2, \ldots, a_m\}$ and $\{b_1, b_2, \ldots, b_n\}$ denote the template residues of proteins *A* and *B* extracted by the template construction process. In this experiment, the template matching algorithm (see 'Template Matching' for details) is invoked to first align $\{a_1, a_2, \ldots, a_m\}$ onto protein *B* and then align $\{b_1, b_2, \ldots, b_n\}$ onto protein *A*, i.e. two alignments are derived from one pair of complexes. In the former case, $\{a_1, a_2, \ldots, a_m\}$ represents the template, *B* represents the query protein, and $\{b_1, b_2, \ldots, b_n\}$ represents the answer. The set of residues of protein *B* successfully aligned with $\{a_1, a_2, \ldots, a_m\}$, denoted by $\{b_1', b_2', \ldots, b_m'\}$, represents the predicted binding site. This experiment focuses on the closeness between the predicted binding site and the answer.

2) Evaluation indices

The closeness between the predicted binding site and the real one is measured in two ways. The first measure, which estimates the distance between the centers of the predicted and real protein-ligand binding sites, is the distance from the geometric center of the aligned residues of the query protein ($\{b_1', b_2', \ldots, b_m'\}$) to the geometric center of the template residues of the query protein ($\{b_1, b_2, \ldots, b_n\}$). The second measure, which

estimates the distance between the center of the predicted binding site and the ligand, is the distance from the geometric center of the aligned residues of the query protein to its closest ligand atom in the query protein-ligand complex.

By assigning appropriate thresholds to these two measures, each predicted alignment can be categorized into three conditions. First, a prediction is said to be correct when the distance between the predicted binding site and the answer is shorter than a distance threshold. Second, a prediction is said to be incorrect when the distance between the predicted binding site and the answer is longer than the threshold. Third, no prediction is given if the query protein contains no substructures that can pass the matching criteria elaborated in 'Matching Constraints' when compared with the template. In this experiment, the performance is evaluated by the false positive ratio defined as follows:

$$\text{false positive ratio} = \frac{\text{number of incorrect predictions}}{\text{number of predictions}},$$

where the 'number of predictions' denotes the number of the alignments passing the matching criteria.

3) Experimental results

Table I and Table II show the performance of Protemot and this work in locating binding regions. The false positive ratio of the present method in hitting the proximity of the protein-ligand binding site is in the range 14.9%–54.9%, depending on the distance threshold for defining the 'proximity'. As stated in 'Template Construction', the adopted local structure alignment requires that the distance between the alpha carbons of two aligned residues be less than 3Å. This condition suggests that ≤ 3Å could be a reasonable threshold for Table I. Similarly, the template construction exploits the information from the region within 6.5Å to a ligand and an appropriate threshold for Table II should be ≤ 6.5Å. Based on the two thresholds, the false positive ratios of the present method in locating the binding site are 22.2% and 15.2% in Table I and Table II, respectively. Although the first measure seems to be more sensitive, the preferred measure should depend on the requirement of the application. Nevertheless, from Table I and Table II, the proposed sequence-order constraint can reduce the false positive ratio in hitting the proximity of protein-ligand binding sites, regardless of the measure or threshold.

TABLE I
PERFORMANCE OF LOCATING PROTEIN BINDING REGIONS IN TERMS OF THE DISTANCE BETWEEN THE PREDICTED AND REAL PROTEIN-LIGAND BINDING SITES

Distance threshold	False positive ratio of Protemot1	False positive ratio of this work1
≤ 1Å	67.1%	54.9%
≤ 2Å	49.9%	31.6%
≤ 3Å	42.7%	22.2%
≤ 4Å	39.1%	18.5%
≤ 5Å	37.3%	16.5%

[1]A protein-ligand binding site is considered to be successfully predicted if the distance between the geometric centers of the predicted and the real binding sites is within the distance threshold.

TABLE II
PERFORMANCE OF LOCATING PROTEIN BINDING REGIONS IN TERMS OF THE DISTANCE BETWEEN THE PREDICTED PROTEIN-LIGAND BINDING SITE AND THE LIGAND

Distance threshold	False positive ratio of Protemot[1]	False positive ratio of this work[1]
≤ 3.5Å	33.1%	16.2%
≤ 4.5Å	30.1%	15.4%
≤ 5.5Å	28.9%	15.2%
≤ 6.5Å	27.5%	15.2%
≤ 7.5Å	26.3%	15.1%
≤ 8.5Å	25.1%	15.0%
≤ 9.5Å	24.3%	14.9%

[1]A protein-ligand binding site is considered to be successfully predicted if the distance from the geometric center of the predicted binding site to its closest ligand atom is within the distance threshold.

Furthermore, the false positive ratios of Protemot and this work are analyzed based on the sequence identity between the two proteins of each pair. Fig. 3 shows the prediction results based on sequence identity [25] between the pair. The thresholds are ≤ 3Å for the first measure and ≤ 6.5Å for the second measure. As shown in Fig. 3, the present method achieves superior false positive ratios than Protemot, especially on the pairs with low sequence identity.

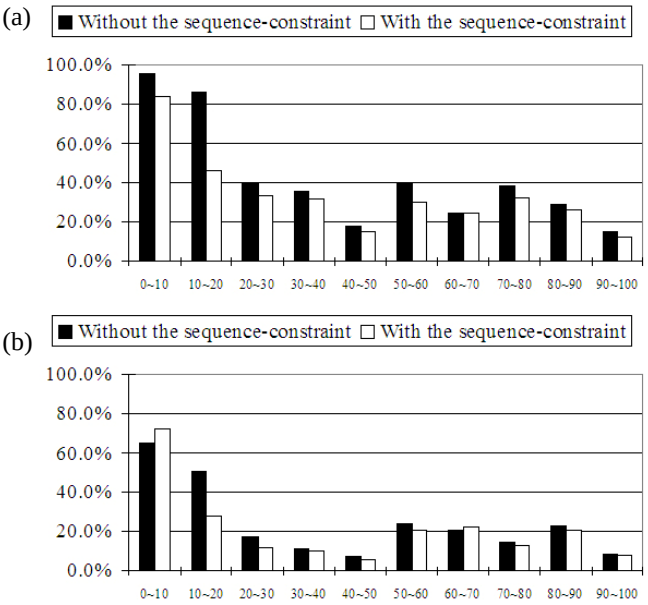

Fig. 2 Comparison between Protemot and this work in locating protein binding regions by sequence identity between the query protein and the template, where *x*-axis is the identity between the template and the query protein and *y*-axis is false positive ratio. (a) A protein-ligand binding site is considered to be successfully predicted if the distance between the geometric centers of the predicted and the real binding sites is within 3Å. (b) A protein-ligand binding site is considered to be successfully predicted if the distance from the geometric center of the predicted binding site to its closest ligand atom is within 6.5Å.

B. Prediction of the Enzyme Class Based On Protein-Ligand Binding Sites

1) Experimental design

This section reports the experiment conducted to evaluate the prediction power of Protemot with and without the proposed sequence-order constraint as well as the prediction power of CSS in predicting enzyme class, which is the main evaluation index in the Protemot paper. The template library of Protemot contains a total of 1051 entries. Conversely, the web version of CSS uses only 147 templates derived from CSA entries. As shown in Table III, for the majority of the templates in the CSS template library, we can find entries in the Protemot template library belonging to similar group in terms of EC numbers, SCOP families, and SCOP superfamilies. The testing dataset contains a total of 1000 non-redundant, randomly selected enzyme structures distributed over 587 four-digit EC numbers. Care has been taken to ensure that the testing dataset does not contain any of those enzyme structures present in both libraries.

In this experiment, each of the 1000 testing enzymes is aligned against all templates in the library. The alignments are ranked based on the TM-score described in 'Template Matching'. The predicted enzyme class is the EC number associated with the template that is highest ranked. The EC number associated with the testing enzyme represents the answer. This experiment determines whether the predicted enzyme class is correct.

2) Evaluation indices

In this experiment, each prediction is categorized into one of three conditions. A correct prediction is one in which the predicted enzyme class matched the answer. An incorrect prediction is one in which the predicted enzyme class did not match the answer. No prediction is made if the testing enzyme contained no substructures passing the matching criteria when compared with all the templates in the template library. The predictor fails to make a correct prediction in some cases due to the lack of a template associated with the answer enzyme class in the library. Thus, each prediction result can be further categorized into five conditions as shown in Table IV and Table V that makes the analysis more complicated. In the Protemot paper, two indexes and their harmonic mean were used to evaluate the performance of the predictors. In this study, we provide a more comprehensive analysis by introducing more indexes, especially on those 'no predictions'. The 'Prediction rate' in Table IV shows the probability that the method could success complete the prediction while processing a new testing enzyme. The 'False positive ratio' in Table IV reveals the prediction performance once the method made a prediction. Otherwise, the accuracy represents an overall index of the prediction performance. Considering the lack of a template associated with the answer enzyme class in the library, whether a 'no prediction' is correct becomes a matter of opinion. The conventional accuracy ('Accuracy' in Table IV) regards such a 'no prediction' as incorrect since each of the 1000 testing enzymes is associated with an EC number. From an active view, these no predictions are correct ('Active accuracy' in Table IV). There is another view on these no predictions is that they are neither correct nor incorrect ('Passive accuracy' in Table IV).

TABLE III
THE OVERLAP BETWEEN PROTEMOT TEMPLATES AND CSA TEMPLATES USED IN THIS ARTICLE

	Protemot	CSS
Number of spanned PDB codes	1051	147
Overlap[1]	14	14
Overlap ratio[2]	1.3%	9.5%
Number of spanned four-digit EC numbers	635	145
Overlap	108	108
Overlap ratio	17.0%	74.5%
Number of spanned three-digit EC numbers	146	74
Overlap	68	68
Overlap ratio	46.6%	91.9%
Number of spanned SCOP families	749	191
Overlap	146	146
Overlap ratio	19.5%	76.4%
Number of spanned SCOP superfamilies	447	145
Overlap	117	117
Overlap ratio	26.2%	80.7%

[1]Number of entries included in both template libraries. [2]Ratio of the overlap entries to all entries in a template library. For instance, there are 1.3% of PDB codes in the Protemot template library that can be found in the CSS template library.

TABLE IV
COMPARISON BETWEEN PROTEMOT AND THIS WORK BASED ON FOUR-DIGIT EC NUMBERS

	Protemot	This work	CSS-L[1]	CSS-H[2]
Total[3]	1000	1000	1000	1000
Cover rate[4]	73.2%	73.2%	14.6%	14.6%
Covered performance				
I. Correct prediction[5]	408	515	81	75
II. Incorrect prediction[6]	310	90	61	8
III. No prediction[7]	14	127	4	63
Uncovered performance				
IV. No prediction[8]	14	123	65	777
V. Incorrect prediction[9]	254	145	789	77
Prediction rate[10]	**97.2%**	75.0%	93.1%	16.0%
False positive ratio[11]	58.0%	**31.3%**	91.3%	53.1%
Accuracy[12]	40.8%	**51.5%**	8.1%	7.5%
Active accuracy[13]	42.2%	63.8%	14.6%	**85.2%**
Passive accuracy[14]	41.4%	**58.7%**	8.7%	33.6%

[1]CSS with lower confidence, with the predictions classified by CSS as 'unlikely' treated as 'no prediction'. [2]CSS with higher confidence, with the predictions by CSS as 'unlikely' or 'possible' treated as 'no prediction' [3]Number of testing enzymes. [4]A testing enzyme is 'covered' if the template library contains a template extracted from a protein-ligand complex structure with the same four-digit EC number as the testing enzyme. [5]The testing enzyme is covered, and the algorithm makes a correct prediction. [6]The testing enzyme is covered, but the algorithm makes an incorrect prediction. [7]The testing enzyme is covered, but the algorithm makes no prediction. [8]The testing enzyme is not covered, and the algorithm makes no prediction. [9]The testing enzyme is not covered, but the algorithm makes a prediction, which is certainly incorrect. [10]Prediction rate = (I+II+V)/Total. [11]False positive ratio = (II+V)/(I+II+V). [12]Accuracy = I/Total. [13]Active accuracy = (I+IV)/Total. [14]Passive accuracy = I/(Total–IV).

TABLE V
COMPARISON BETWEEN PROTEMOT AND THIS WORK BASED ON THREE-DIGIT EC NUMBERS

	Protemot	This work	CSA–L	CSA–S
Total	1000	1000	1000	1000
Cover rate[1]	98.7%	98.7%	72.1%	72.1%
Covered performance				
I. Correct prediction	514	615	143	118
II. Incorrect prediction	447	125	531	28
III. No prediction	26	247	47	575
Uncovered performance				
IV. No prediction	2	3	22	265
V. Incorrect prediction	11	10	257	14
Prediction rate	**97.2%**	75.0%	93.1%	16.0%
False positive ratio	47.1%	**18.0%**	84.6%	26.3%
Accuracy	51.4%	**61.5%**	14.3%	11.8%
Active accuracy	51.6%	**61.8%**	16.5%	38.3%
Passive accuracy	51.5%	**61.7%**	14.6%	16.1%

[1]A testing enzyme is 'covered' if the template library contains a template extracted from a protein-ligand complex structure with the same three-digit EC number as the testing enzyme.

3) Experimental results

Table IV reports the experimental results when the four-digit EC numbers are used as the answers, while Table V reports the experimental results when the three-digit EC numbers are used as the answers. Table IV and Table V clearly reveal that the sequence-order constraint reduces the false positive ratio by 26.7% in Table IV and 29.1% in Table V. As expected, the prediction rate drops (from 97.2% to 75.0%), since that this work involves an additional constraint. As shown in Table VI, most of the 'no predictions' filtered by the sequence-order constraint either result from the lack of a template associated with the answer enzyme class in the library (meaning that they have to be reported as 'no predictions'), or are innocent to the predictor. In summary, the sequence-order constraint only decreases the effective prediction rate by 1.2% (12 incorrectly filtered predictions / 1000) and 1.7% (17 / 1000) in Table IV and Table V. When considering the overall prediction performance, the present method outperforms Protemot no matter which index of 'Accuracy', 'Active accuracy' or 'Passive accuracy' is used in these two tables. These results indicate that the proposed sequence-order constraint can improve protein binding site prediction based on automatically extracted templates.

TABLE VI
STATISTICS OF THE 'NO PREDICTIONS' FILTERED BY THE INTRODUCED SEQUENCE-ORDER CONSTRAINT

	Based on four-digit EC numbers	Based on three-digit EC numbers
Protemot makes a prediction, while this work makes no prediction	222	222
The testing enzyme is covered, and only Protemot makes a correct prediction, while this work makes no prediction[1]	12	17
The testing enzyme is covered, and Protemot makes an incorrect prediction, while this work makes no prediction[2]	101	204
The testing enzyme is uncovered, and Protemot makes an incorrect prediction, while this work makes no prediction[3]	109	1

[1]The 'no predictions' reported by this work were incorrectly filtered by the sequence-order constraint. [2]The 'no predictions' reported by this work were innocent to the predictor, since it delivered an incorrect prediction without the sequence-order constraint. [3]The 'no predictions' reported by this work were correctly filtered, since the testing enzyme is uncovered.

C. Case Study

This section presents two testing examples to illustrate the effect of the sequence-order constraint. One testing case is chosen from the experiment of predicting binding location, and the other is chosen from the experiment of predicting enzyme class. Both cases are chosen from the weak homology region (sequence identity <25%) of the two experiments.

The first case is the alignment between a phosphoserine aminotransferase from *bacillus alcalophilus* (PDB: 1W23) and a ptpase from *bos taurus* (PDB: 1DG9). The two complex structures have a common ligand, 4-(2-hydroxyethyl)-1-piperazine ethanesulfonic acid (HEPES). However, HEPES plays different roles in the two complexes: a buffer in 1W23, and a competitor of vanadate (which acts as an analogue of the transition state of the cleavage reaction) in 1DG9. For this case, Protemot reports an alignment with a RMSD of 1.66. However, this alignment has only three pairs sequentially in order out of five aligned residue pairs, resulting in a sRMSD of 2.07. Therefore, introducing the sequence-order constraint filters this alignment.

The second case is the alignment between two lipases (PDB: 1KU0 and 1HQD) having the same four-digit EC number (3.1.1.3). Accordingly, a predictor for functional inference should report that the two enzymes are highly likely to have the same function. PSI-BLAST is invoked to analyze the two proteins, and finds no significant similarity. Conversely, the present method successfully predicts this case, because it focuses on the local spatial region of the template built according to the (Rp,sp)-o-(2r)-(1-phenoxybut-2-yl)-methylphosphonic acid chloride in 1HQD. Protemot predicts that 1KU0 is more similar to another enzyme (PDB: 3CPA) in the template library over 1HQD, resulting in an incorrect prediction. Like the previous example, the alignment between 1KU0 and 3CPA does not pass the sequence-order constraint. Thus, the sequence-order constraint helps the present method to filter the alignment between 1KU0 and 3CPA and to deliver a correct prediction.

IV. CONCLUSION

This study is motivated by the high false positive ratio in predicting protein-ligand binding sites with automatically extracted templates. A sequence-order constraint is employed to identify the inconsistencies between the local regions of the query protein and the templates. Two experiments are conducted to evaluate the prediction performance of the present method. In the first experiment, the introduced sequence-order constraint can reduce the false positive ratio in locating

protein-ligand binding sites, especially on the pairs with low sequence identity. Results of the second experiment indicate that the sequence-order constraint is helpful in predicting enzyme class. Experimental results reveal that the present method can yield low false positive ratio while using automatically extracted templates, which is essential for creating and maintaining a comprehensive template library as the number of entries in PDB grows exponentially.

Although the present method has been able to provide good performance to predict protein-ligand binding sites, further improvements could be made with respect to the constraint-based design. In this regard, we currently hypothesize that the priority of constraint is higher than the scoring function, meaning that an alignment that fails any constraint is never taken into account, even if it has a promising score. Effort is being made to embed the constraint concept into the scoring step to minimize the loss of possible matches.

ACKNOWLEDGMENT

The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC 96-2320-B-006-027-MY2 and NSC 96-2221-E-006-232-MY2. Ted Knoy and Christine Wang are appreciated for their editorial assistance.

REFERENCES

[1] S. E. Brenner, "A tour of structural genomics," Nature Reviews Genetics, vol. 2, pp. 801-809, Oct 2001.
[2] J. D. Watson, R. A. Laskowski, and J. M. Thornton, "Predicting protein function from sequence and structural data," Current Opinion in Structural Biology, vol. 15, pp. 275-284, Jun 2005.
[3] D. T. H. Chang, C. Y. Chen, W. C. Chung, Y. J. Oyang, H. F. Juan, and H. C. Huang, "ProteMiner-SSM: a web server for efficient analysis of similar protein tertiary substructures," Nucleic Acids Research, vol. 32, pp. W76-W82, Jul 1 2004.
[4] A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson, "Recognition of functional sites in protein structures," Journal of Molecular Biology, vol. 339, pp. 607-633, Jun 4 2004.
[5] F. Ferre, G. Ausiello, A. Zanzoni, and M. Helmer-Citterich, "Functional annotation by identification of local surface similarities: a novel tool for structural genomics," BMC Bioinformatics, vol. 6, p. 194, Aug 2 2005.
[6] J. W. Torrance, G. J. Bartlett, C. T. Porter, and J. M. Thornton, "Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families," Journal of Molecular Biology, vol. 347, pp. 565-581, Apr 1 2005.
[7] C. T. Porter, G. J. Bartlett, and J. M. Thornton, "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data," Nucleic Acids Research, vol. 32, pp. D129-D133, Jan 1 2004.
[8] J. A. Barker and J. M. Thornton, "An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis," Bioinformatics, vol. 19, pp. 1644-1649, Sep 1 2003.
[9] D. T.-H. Chang, Y.-Z. Weng, J.-H. Lin, M.-J. Hwang, and Y.-J. Oyang, "Protemot: prediction of protein binding sites with automatically extracted geometrical templates," Nucleic Acids Research, vol. 34, pp. W303-W309, 2006.
[10] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," Nucleic Acids Research, vol. 28, pp. 235-242, Jan 1 2000.
[11] B. P. Pandey, C. Zhang, X. Z. Yuan, J. Zi, and Y. Q. Zhou, "Protein flexibility prediction by an all-atom mean-field statistical theory," Protein Science, vol. 14, pp. 1772-1777, Jul 2005.
[12] I. Bahar, A. R. Atilgan, and B. Erman, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential," Folding & Design, vol. 2, pp. 173-181, 1997.
[13] R. A. Laskowski, V. V. Chistyakov, and J. M. Thornton, "PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids," Nucleic Acids Research, vol. 33, pp. D266-D268, Jan 1 2005.
[14] Y. J. Oyang, S. C. Hwang, Y. Y. Ou, C. Y. Chen, and Z. W. Chen, "Data classification with radial basis function networks based on a novel kernel density estimation algorithm," IEEE Transactions on Neural Networks, vol. 16, pp. 225-236, Jan 2005.
[15] Y.-J. Oyang, D. T.-H. Chang, Y.-Y. Ou, H.-G. Hung, C.-P. Wu, and C.-Y. Chen, "Supervised Machine Learning with a Novel Kernel Density Estimator," 2007, p. arXiv:stat.ML/0709.2760.
[16] H. J. Wolfson and I. Rigoutsos, "Geometric hashing: An overview," Ieee Computational Science & Engineering, vol. 4, pp. 10-21, Oct-Dec 1997.
[17] C. A. Orengo and W. R. Taylor, "SSAP: Sequential structure alignment program for protein structure comparison," Computer Methods for Macromolecular Sequence Analysis, vol. 266, pp. 617-635, 1996.
[18] X. Pennec and N. Ayache, "A geometric algorithm to find small but highly similar 3D substructures in proteins," Bioinformatics, vol. 14, pp. 516-522, 1998.
[19] N. S. Boutonnet, M. J. Rooman, M. E. Ochagavia, J. Richelle, and S. J. Wodak, "Optimal Protein-Structure Alignments by Multiple Linkage Clustering - Application to Distantly Related Proteins," Protein Engineering, vol. 8, pp. 647-662, Jul 1995.
[20] D. E. Krane and M. L. Raymer, Fundamental concepts of bioinformatics. San Francisco: Benjamin Cummings, 2002.
[21] S. F. Altschul, "Amino-Acid Substitution Matrices from an Information Theoretic Perspective," Journal of Molecular Biology, vol. 219, pp. 555-565, Jun 5 1991.
[22] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," Proteins-Structure Function and Bioinformatics, vol. 57, pp. 702-710, Dec 1 2004.
[23] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, Introduction to Algorithms, Second Edition: The MIT Press, 2001.
[24] A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "Data growth and its impact on the SCOP database: new developments," Nucleic Acids Research, vol. 36, pp. D419-D425, Jan 2008.
[25] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," J Mol Biol, vol. 48, pp. 443-453, 1970.