行政院國家科學委員會專題研究計畫 期中進度報告

虛擬篩選與嶄新藥物設計之整合計算平台--（子計畫二）針對虛擬藥物篩選設計先進機器學習演算法(2/3)

期中進度報告(精簡版)

計畫類別：整合型
計畫編號：NSC 95-2627-B-002-009-
執行期間：95年08月01日至96年07月31日
執行單位：國立臺灣大學資訊工程學系暨研究所

計畫主持人：歐陽彥正
共同主持人：陳倩瑜

處理方式：期中報告不提供公開查詢

中華民國 96年05月31日

行政院國家科學委員會補助專題研究計畫 期中進度報告

虛擬篩選與嶄新藥物設計之整合計算平台-(子計畫二)
針對虛擬藥物篩選設計先進機器學習演算法

計畫類別：□個別型計畫 ■ 整合型計畫
計畫編號：NSC 95－2627－B－002－009
執行期間： 2006 年 8 月 1日至 2007 年 7 月 31 日

計畫主持人：歐陽彥正 教授
共同主持人：陳倩瑜 助理教授
計畫參與人員：

成果報告類型(依經費核定清單規定繳交)：■精簡報告 □完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
■涉及專利或其他智慧財產權，■一年□二年後可公開查詢

執行單位：國立臺灣大學資訊工程學系暨研究所

中 華 民 國 96 年 5 月 30 日

可供推廣之研發成果資料表

■ 可申請專利　　□ 可技術移轉　　　　　　　　　　日期：96 年 5 月 30 日

國科會補助計畫	計畫名稱：虛擬篩選與嶄新藥物設計之整合計算平台-(子計畫二) 針對虛擬藥物篩選設計先進機器學習演算法 計畫主持人： 歐陽彥正教授 計畫編號：NSC 95－2627－B－002 － 009
技術/創作名稱	iPDA：蛋白質功能區結構不規則性之預測
發明人/創作人	蘇中才、陳倩瑜、歐陽彥正
技術說明	中文：iPDA 網頁服務旨於提供生化學家一個容易使用且準確度高的「蛋白質功能區之不規則性分析軟體」，其整合本實驗室所開法的蛋白質不穩定區域與蛋白質功能區兩種預測工具，以及多種現今最常使用之蛋白質序列分析軟體。所提供之資訊包括：胺基酸保留性、二級結構預測、疏水性群聚等蛋白質序列之重要特性，將有利於分析蛋白質功能區之結構不規則現象。 英文：iPDA aims to predict the disordered regions of a query protein. Automatic prediction of disordered regions from protein sequences is an important problem in the study of structural biology. The proposed predictor, DisPSSMP2, is different from several existing disorder packages by its employment of position specific scoring matrices with respect to physicochemical properties (PSSMP), where the physicochemical properties adopted here especially take the disorder propensity of amino acids into account. The web server iPDA integrates DisPSSMP2 with several other sequence predictors in order to investigate the functional role of the detected disordered region. The predicted information includes sequence conservation, secondary structure, low complexity, and hydrophobic clusters. Furthermore, a pattern mining package for detecting concurrent sequence conservation is embedded in iPDA for discovering potential binding regions of the query protein, which is really helpful to uncovering the relationship between protein function and its primary sequence. The web service is available at http://biominer.bime.ntu.edu.tw/ipda.
可利用之產業 及 可開發之產品	可將 iPDA 軟體發展成為生物資訊套裝軟體。
技術特點	iPDA 使用本實驗所提出之新的序列特徵進行蛋白質不穩定區域之預測，並搭配本實驗室所研發之序列特徵探勘軟體進行蛋白質功能區之域測。
推廣及運用的價值	蛋白質功能區之不穩定性分析於設計可靠之鍵結能量方程式扮演重要角色，將有助於蛋白質配體嵌合預測工具之開發。

English Abstracts

The primary objective of the integrated project is to develop a novel molecular docking approach that achieves the level of efficiency required for carrying out virtual screening and de novo design and is able to overcome the obstacle of protein flexibility, which is difficult to deal with the conventional approaches. The basis of the novel approach is to exploit the machine learning mechanisms efficiently. In the second year of this project, we have focused on development of a protein disorder predictor that facilitates investigation of the structural flexibility of protein functional regions. This study first examines the effect of a condensed position specific scoring matrix with respect to physicochemical properties (PSSMP) on the prediction accuracy, where the PSSMP is derived by merging several amino acid columns of a PSSM belonging to a certain property into a single column. Next, each conventional physicochemical property of amino acids is decomposed into two disjoint groups which have a propensity for order and disorder respectively. It will be shown by experiments that some of the new properties perform better than their parent properties in predicting protein disorder. In order to get an effective and compact feature set on this problem, a hybrid feature selection method is proposed, which inherits the efficiency of uni-variant analysis and the effectiveness of the stepwise feature selection that explores combinations of multiple features. Finally, the proposed method is integrated in a web server named iPDA with several other sequence predictors, in order to investigate the functional role of the detected disordered region. The predicted information includes sequence conservation, secondary structure, low complexity, and hydrophobic clusters.

Keywords

Protein structure, machine learning, protein interaction.

Chinese Abstracts

此整合型計畫的目標是開發一個新的分子嵌合(molecular docking)方法，以設計一套具備高效能處理虛擬藥物篩選(virtual screening)與起始式藥物設計(de novo)能力的軟體。而此新方法的基礎在於有效地運用機器學習演算法(machine learning algorithms)。在本年度的計畫中，我們致力於開發一個命名為iPDA的高準確度「蛋白質功能區之不規則性分析軟體」。有了iPDA，生化學家便能準備預測蛋白質不規則性的區塊(Disordered regions)，而這些區塊許多都與蛋白質功能相關。iPDA整合本實驗室所開法的蛋白質不穩定區域與蛋白質功能區兩種預測工具，以及多種現今最常使用之蛋白質序列分析軟體。所提供之資訊包括：胺基酸保留性、二級結構預測、疏水性群聚等蛋白質序列之重要特性，將有利於分析蛋白質功能區之結構不規則現象。

Keywords

分子嵌合，機器學習，蛋白質互動

目錄

English Abstract…..iii
Chinese Abstract…..iv
目錄..v
一、前言..1
二、研究目的..1
三、文獻探討..1
四、研究方法..3
五、成果與討論...4
References..8

一、前言

Structural flexibility plays an important role on the development of accurate regression functions for modeling the energy state variations of protein-ligand interactions. The goal of the second year aims at developing a protein disorder predictor that helps to investigate the structural flexibility of protein functional regions. The study in the second year first examines the effect of a condensed position specific scoring matrix with respect to physicochemical properties (PSSMP) on the prediction accuracy, where the PSSMP is derived by merging several amino acid columns of a PSSM belonging to a certain property into a single column. Next, each conventional physicochemical property of amino acids is decomposed into two disjoint groups which have a propensity for order and disorder respectively. It will be shown by experiments that some of the new properties perform better than their parent properties in predicting protein disorder. In order to get an effective and compact feature set on this problem, a hybrid feature selection method is proposed, which inherits the efficiency of uni-variant analysis and the effectiveness of the stepwise feature selection that explores combinations of multiple features. Finally, the proposed method is integrated in a web server named iPDA with several other sequence predictors, in order to investigate the functional role of the detected disordered region. The predicted information includes sequence conservation, secondary structure, low complexity, and hydrophobic clusters.

二、研究目的

Intrinsically disordered proteins or protein regions exhibit unstable and changeable three dimensional structures under physiological conditions [1]. Although lacking fixed structures, protein disorder has been identified to carry out important functions in many biological processes [1,2]. In addition, it is observed that the absence of a rigid structure allows disordered binding regions to interact with several different targets [3,4]. These regions, sometimes called “molecular recognition elements”, usually undergo a disorder-to-order transition when binding to their targets [5,6]. In this regard, predicting protein disorder and investigating its potential for induced folding is a necessary preliminary procedure in understanding protein structure and function [7].

In our recent work DisPSSMP, it is demonstrated that the accuracy of protein disorder prediction can be greatly improved if the disorder propensity of amino acids is considered when generating the condensed PSSM (position specific scoring matrix) features [8]. For iPDA, we implement a two-stage classifier of Radial Basis Function Networks (RBFN) to further enhance the predicting power of DisPSSMP. As unbalanced datasets, a large amount of ordered residues over disordered residues, are employed when training the classifier DisPSSMP2, an alternative decision function is newly adopted and the cutting threshold is dynamically determined by the proportion of predicted secondary structure in the query protein [9].

三、文獻探討

It has been shown in many studies that protein disorders can be predicted from their primary sequences [3, 17, 18, 19, 20, 21, 22]. The prediction methods developed in recent years initiate the possibility of identifying such disordered binding sites automatically [18, 21]. A more general concept is that all the necessary information for the correct folding of a protein is included in its amino acid sequence [23]. Disordered regions are comprised of a category of amino acids distinct from that of ordered protein structures [24]. For example, amino acids of aromatic hydrophobic groups are known to be good for the general stabilization of order, and thus are less found in the disordered regions [17]. Incorporating information of the biased amino acid composition in a neural network predicts the locations of disorder with accuracy better than random guesses [17]. In 1998, Romero *et al.* showed that more than 15,000 proteins in the Swiss-Prot database contain long disordered segments (40 or more residues) based on their predictions [18, 25]. Studies on some of these disordered regions reveal that they are evolutionarily conserved and possess biological functions [3].

Several machine learning approaches, such as neural networks (NNs) [14, 3, 17, 26], logistic regression (LR) [26, 27], discriminant analysis (DA) [27], ordinary least squares (OLS) [26], and support vector machines (SVM) [12, 27, 28] have been introduced to protein disorder prediction. Since different classifiers deliver similar prediction accuracy based on the same feature set [27], extracting more useful features with biological insights to improve the quality of prediction attracts more attention in recent studies [3, 28]. As amino acid composition has been demonstrated as a useful feature for detecting disordered regions, Jones *et al.* showed in their paper that using the position specific scoring matrices (PSSMs) within a specific length of window centred at a given residue can improve the accuracy of predicting its disorder attribute [3]. The values in a position specific scoring matrix indicate the level of conservation of a position and the properties of the substituted residues, which can be derived directly from executing PSI-BLAST for each target protein sequence. PSSMs have been demonstrated to be powerful in constructing feature sets for prediction of single-residue properties from an amino acid sequence, such as category of secondary structures or solvent accessibility [3]. The evolutionary information summarized in a PSSM table generalizes the attribute of each position in a protein sequence, and thus improves the sensitivity of the predicting model.

A development of this approach employs a condensed position specific scoring matrix with respect to physicochemical properties (it will be called PSSMP in this paper) in predicting protein disorder, where the PSSMP is derived by merging several amino acid columns of a PSSM that belong to a certain property into a single column [28]. As a PSSM brings in the evolutionary information on each position, a PSSMP summarizes this information as property attributes. The improvement achieved by PSSMP demonstrates that property attributes are more informative than amino acid attributes in distinguishing ordered/disordered regions. A more comprehensive study conducted in this paper reveals that PSSMPs outperform PSSMs especially when the employed window size is large.

When employing PSSMP tables as the feature set in protein disorder prediction, two questions arise: (1) if all the amino acids in one physicochemical property group contribute to the predicting power; and (2) if all the amino acids in one physicochemical property group result in consistent effect on prediction. It has been widely studied in previous works that the propensity for order or disorder of several amino acids is clear. Hydrophobic amino acids are more frequently observed in ordered regions than disordered regions [22, 23]. Among them the aromatic amino acids are present in different locations to the aliphatic amino acids [29]. On the other hand, the amino acids with charge imbalance are often present in disordered regions. In this paper, we argue that the propensity for order or disorder of each amino acid should be considered when constructing PSSMP. After examining the statistics derived by comparing the sequence segments in ordered regions and disordered regions, we observe that not all the hydrophobic amino acids possess a propensity for order. Thus we suggest that each conventional physicochemical property should be divided into two smaller groups with propensities for order and disorder respectively, such as hydrophobic with an order propensity ($Hydrophobic_O$) and hydrophobic with a disorder propensity ($Hydrophobic_D$). The experiments conducted in this work reveal that some newly derived properties provide more accurate information regarding order or disorder.

Incorporating the propensity for order/disorder with physicochemical properties in PSSMP produces informative features for protein disorder prediction. However, the number of candidate features becomes larger than in the case when only twenty amino acids are considered. The size of the feature set gets even larger when a large window size is employed, which might cause the performance of the learning algorithms to be degraded due to abundant noise. Thus, we present in this paper a feature selection mechanism that considers both the size and effectiveness issues when determining a feature set on protein disorder prediction. A wrapper approach of feature selection is employed during training period that invokes the adopted Radial Basis Function Networks (RBFN) classifier to evaluate the predicting power of a candidate feature set. A cluster-based redundancy removal procedure is incorporated to speed up the stepwise feature selection process, where two levels of redundancy among features are considered.

As far as the experimental materials are concerned, a new dataset PDB693 organized from the Protein Data Bank (PDB) [30] database is coined in this work to benefit the study on protein disorder. PDB693 and another dataset D184 collected from Database of Protein Disorder (DisProt) [31] constitute the training data of our classifier DisPSSMP. The performance of DisPSSMP is compared with twelve existing disorder prediction packages, where the blind testing data comes from a recent study [14]. The experimental results demonstrate that the selected property features are informative in protein disorder prediction and can be used to derive discriminating patterns for order and disorder classification.

四、研究方法

In this section, we provide the details about the procedures of constructing PSSMPs, calculating the propensities for order/disorder of an amino acid, training a predicting model, and selecting useful feature sets respectively.

For each protein in the training and testing data, we employ the PSI-BLAST program [46] to construct its position specific scoring matrix (PSSM). More specifically, the derived PSSM table is a position specific scoring matrix of 20 amino acids, which provides the evolutionary information about the protein at the level of residue types. We name the feature set created based on PSSMs FS-PSSM, which is considered as the baseline of employing evolutionary information in protein disorder prediction. The values in PSSMs, which each represents the likelihood of a particular residue substitution at specific position, are first rescaled to be within 0 and 1 using the logistic function as suggested in [47]:

$$f(x) = \frac{1}{1+\exp(-x)}, \tag{2}$$

where x is the raw value in profile matrices and $f(x)$ is the rescaled value of x. After that, the rescaled profiles are organized into a number of $w \times 20$ dimensional vectors, each of which serves as the feature vector of a target residue as the learning or predicting instances. When w is odd, which is always the case in our experiments, the sliding window of size w for acquiring the feature set of a given residue is centred at the target residue.

We next construct the feature set of PSSMP as follows. First, columns in the original PSSMs are grouped by the user defined property groups and the raw values from different columns are summed up as a new feature column. In a PSSMP table, the entry y_{ik} of position i for property k is defined as follows:

$$y_{ik} = \sum_{j=1}^{20} A_{kj} \times x_{ij} \ , \tag{3}$$

where (1) i is the index of a position; (2) $A_{kj} = 1$, if the j-th type of amino acid belong in the k-th property, and $A_{kj} = 0$, otherwise; (3) x_{ij} is the raw value of the j-th type of amino acid in the position i of the PSSM.

We call the derived feature set as FS-PSSMP. As will be explained the following subsection, different FS-PSSMPs can be generated when different property groups are specified in constructing FS-PSSMP.

Table 1 lists the ten physicochemical groups that are widely used in analyzing protein sequences [13, 48]. This paper proposes considering the propensity for order or disorder of each amino acid when designing a property group in construction of PSSMPs. The propensities for order/disorder of different amino acids have been widely discussed in the previous studies [17, 21, 23, 25, 26, 7, 43, 44, 45]. Some of them specifically provide a measure of propensity based on the occurrences of each amino acid in different regions of the datasets they collected [26, 7, 44]. In this work, we recalculate the propensity for each amino acid based our training data. The propensity $P(AA_i)$ of an amino acid i toward ordered or disordered regions is defined as follows:

$$P(AA_i) = \frac{F_O(AA_i) - F_D(AA_i)}{F(AA_i)}, \tag{4}$$

where $F(AA_i)$ is the frequency of amino acid i in the training data and $F_O(AA_i)$ and $F_D(AA_i)$ are the frequencies of amino acid i in the ordered and disordered regions of the training data. We say amino acid i has a propensity for order if $P(AA_i) > 0$, and verse visa.

Based on the frequencies calculated based on the training dataset, each physicochemical property in Table 1 can be decomposed into two disjoint set as new order/disorder-based property features, as shown in Table 2. Three exceptions are: *Aliphatic* has only three types of amino acid toward ordered regions, and *Negative* and *Proline* have only the disorder type of amino acids. It is noticed that there are some new properties which only comprise a single type of amino acid, such as $Aromatic_D$, $Positive_O$, $Charged_O$, and $Tiny_O$. All the property features will be considered in constructing the PSSMP feature set for protein disorder prediction. Since the size of the feature set is quite large and we do not expect all the property features are useful in predicting protein disorder, a feature selection method will be conducted to find a combination of property features that benefits protein disorder prediction.

In this study, the Radial Basis Function Network (RBFN) is used as the classifier for predicting protein disorder. The employed QuickRBF package [49] is an efficient tool for constructing RBFN classifiers, which uses the Cholesky decomposition technique to resolve the least mean square error optimization problem when constructing a RBFN classifier. We rely on the efficiency of QuickRBF such that a wrapped method of feature selection can be used in constructing our predictor DisPSSMP, where the 'wrapped' means that the classifier is employed in feature selection process for evaluating the predicting power of the candidate feature set [50].

According to the statistics provided in Table 3, the ratios of disorder residues to order residues in datasets PDB693 and D184 are 1:3.83 and 1:2.03, respectively. In order to tackle the problem of the skewed datasets with unbalanced number of positive and negative instances, equal quantity of residues from ordered and disordered regions is used in constructing the classifier. In other words, the same amount of ordered residues as that of the disordered residues in the training sets is randomly selected and the others are removed before the training process.

It is doubted that all of the properties described in Table 1 and Table 2 are useful in the problem of disorder prediction. Thus, it is suggested to conduct a procedure of feature selection on the training data to find a combination of features that perform the best in this problem. Feature selection is a common optimization problem for finding the smallest subset of the features with the best classification performance [51]. However, finding the optimal feature subset is not easy, since there are 2^n possible combinations when given *n* features. The algorithm of evaluating all subsets such as exhaustive search is impractical for large *n*. Therefore, an alternative stepwise feature selection is presented in this paper that takes the characteristics of features into account to improve the computational efficiency. Three factors are frequently used in evaluating the performance of a feature selection approach: classification accuracy, size of the subset, and computational efficiency [51]. The proposed hybrid method employs the efficient uni-variant analysis first, and uses a cluster-based redundancy removal procedure to speed up the tedious stepwise feature selection that explores the predicting power of combinations of multiple features simultaneously.

The proposed feature selection mechanism is described as follows. First, uni-variant analyses are performed by conducting cross-validation on the training data using PSSMPs with one property at a time. After that the dependency analysis is executed in two levels. Since the members of properties listed in Table 1 and Table 2 are clearly specified, it is easy to put the related features in one cluster. The first level considers the dependency between the child properties in Table 2 and their parent properties in Table 1, and the second level considers the hierarchy dependency between the physicochemical properties listed in Table 1. After the dependency analysis, the redundancy removal step brings one feature or the combination of two features that performs the best in each cluster to the next step, stepwise feature selection. The representative properties from each cluster are sorted by their performance, and the final subset is constructed by adding property features one by one until the performance of cross validation on the training data cannot be improved.

Table 1 Conventional Amino Acid Properties (Parent Properties)

Property	I	L	V	C	A	G	M	F	Y	W	H	K	R	E	Q	D	N	S	T	P
Hydrophobic	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y							Y	
Polar									Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
Small			Y	Y	Y	Y										Y	Y	Y	Y	Y
Aliphatic	Y	Y	Y																	
Aromatic								Y	Y	Y	Y									
Positive											Y	Y	Y							
Negative														Y		Y				
Proline																				Y
Charged											Y	Y	Y	Y		Y				
Tiny				Y	Y	Y												Y		

Table 2 Order/Disorder-based Amino Acid Properties (Child Properties)

Property	I	L	V	C	A	G	M	F	Y	W	H	K	R	E	Q	D	N	S	T	P
Hydrophobic$_O$	Y	Y	Y	Y				Y	Y	Y									Y	
Hydrophobic$_D$					Y	Y	Y				Y	Y								
Polar$_O$									Y	Y			Y				Y		Y	
Polar$_D$											Y	Y		Y	Y	Y		Y		
Small$_O$			Y	Y													Y		Y	
Small$_D$					Y	Y										Y		Y		Y
Aliphatic$_O$*	Y	Y	Y																	
Aromatic$_O$								Y	Y	Y										
Aromatic$_D^{\#}$											Y									
Positive$_O^{\#}$													Y							
Positive$_D$											Y	Y								
Negative$_D$*														Y		Y				
Proline$_D$*$^{\#}$																				Y
Charged$_O^{\#}$													Y							
Charged$_D$											Y	Y		Y		Y				
Tiny$_O^{\#}$				Y																
Tiny$_D$					Y	Y												Y		

* *Aliphatic*$_O$, *Negative*$_D$, and *Proline*$_D$ are equivalent to *Aliphatic*, *Negative*, and *Proline* in Table 1, respectively.

# *Aromatic*$_D$, *Positive*$_O$, *Proline*$_D$, *Charged*$_O$, and *Tiny*$_O$ each comprises only a single type of amino acid.

五、成果與討論

In this section, we first describe how the datasets have been prepared and how the performance of prediction is evaluated. Next, we show the results of the feature selection after conducting cross-validation on the training data. At this stage, we also discuss the effect of the window size employed in constructing PSSMP. After that, the resultant feature set is employed in constructing the final predicting model DisPSSMP. Finally, the testing results are evaluated based on the bind testing data, and are compared with other existing packages performing similar tasks. At the end of the section, we show the derived property sets can be used to discover patterns that distinguish ordered and disordered regions.

In this study, five datasets have been collected or newly created for training and validating processes. The detailed statistics about each dataset are provided in Table 3, including the number of chains, ordered/disordered regions, and residues in ordered/disordered regions. The training data used in constructing the predictor DisPSSMP is composed of datasets PDB693 and D184, which are respectively organized from PDB and DisProt database based on the procedures described in the following paragraphs. Meanwhile, three datasets named R80, U79, and P80, which are taken from two related studies [14, 23], constitute an independent testing data. This blind dataset serves as a platform for comparing the performance of the proposed method with some other existing packages performing protein disorder prediction.

Table 3 Summary of the datasets employed in this study

	Training data		Testing data		
	PDB693	D184	R80	U79	P80
Number of chains	693	184	80	79	80
Number of ordered regions	1357	257	151	0	80
Number of disordered regions	1739	274	183	79	0
Number of residues in the ordered regions	201937	55164	29909	0	16568
Number of residues in the disordered regions	52663	27116	3649	14462	0
Total residues in the dataset	254600	82280	33558	14462	16568

The dataset PDB693 contains 693 partially disordered proteins, where the locations of disordered regions are identified by looking for the missing residues in a protein structure from PDB database (28-Aug-2005 version). There are originally 32204 structures in this version of PDB database, and those structures are filtered by a clustering program Cd-Hit [32, 33] such that the resultant nonredundant set containing no pair of protein sequences with similarity identity of more than 70%. The so-called missing residues are those present in the SEQRES records but not in the ATOM records with their alpha-carbon coordinates. A protein sequence is considered in this study only if it contains at least one disordered region with more than 30 consecutive residues. Furthermore, protein sequences of similarity identity of more than 70% against any protein sequence in the independent testing data have been removed, resulting in 693 protein sequences in the PDB693 dataset.

Another training set D184 is extracted from DisProt database. DisProt is a curated database that provides information about proteins that entirely or partially lack a fixed three-dimensional structure under putatively native conditions [31]. The DisProt release 2.2 consists of 202 proteins, including 431 distinct disordered regions in total. Among the 202 proteins, there are 157 proteins that contain at least one disordered region longer than 30 consecutive residues. There are more than 50 wholly disordered proteins in DisProt database which are annotated as serving certain functions. D184 is also filtered by Cd-Hit to remove redundant proteins which have more than 70% identity with some other proteins inside it or in any of the three testing datasets.

The dataset R80 was prepared by Yang *et al.* in 2005 [14]. The 80 protein chains in this dataset are collected from the PDB database, and each protein chain contains a region of at least 21 consecutive disordered residues. Additionally, the dataset U79 organized by Uversky *et al.* in 2000 [23] and the dataset P80 provided by PONDR® web site (retrieved in February 2003) are also compiled into the blind testing set, where the dataset U79 contains 79 totally disordered proteins and the dataset P80 contains 80 completely ordered proteins. By using Cd-Hit again, we observed that two sequences in P80 are subsequences of a protein in R80 and a pair of proteins in U79 have identity of 73%. Like Yang *et al.* did in their paper [14], these three datasets are employed as a platform for comparison of our approach to some other present packages targeting at protein disorder prediction. Thus, we did not change the contents of these three datasets such that the comparison can be carried out. In particular, the datasets U79 (fully disordered proteins) and P80 (globular proteins) together suggest whether the proposed method is under- or over-predicting protein disorder.

Predicting a residue in the given protein sequence as order or disorder is a binary classification problem, and many measures have been introduced for validation issues [34, 35]. Table 4 lists four widely used indices defined by previous related works [14, 22, 28, 34, 35, 36]. We employ these measures in this study to evaluate the performance of different feature sets or different packages. *Sensitivity* represents the fraction of disordered residues correctly identified in a prediction method, while *specificity* indicates the fraction of ordered residues correctly identified. The *Matthews' correlation coefficient* is a popular measure in many bioinformatics problems [37, 38, 39]. However, *sensitivity*, *specificity*, and the *Matthews' correlation coefficient* are seriously affected by the relative frequency of the target class. Therefore, the above three measures are not suitable for evaluating the performance in isolation. The *probability excess* is independent of the relative class frequency, and this measure can be reduced to *sensitivity* + *specificity* – 1 concisely [14]. In addition, some other indices including the *CASP S score*, *product*, and *probability excess* are recommended and advised by CASP6 [35] and Yang *et al*. [14] for evaluating the performance of prediction. Since these three measures have the same tendency with *probability excess*, we adopt only the *probability excess* in this study for simplicity and show the results of other measures in the supplementary.

Table 4 The definition of measures employed in this study

Measure	Abbreviation	Equation *
Sensitivity (recall)	*Sens.*	TP/(TP+FN)
Specificity	*Spec.*	TN/(TN+FP)
Matthews' correlation coefficient	*MCC*	(TP×TN-FP×FN)/sqrt((TP+FP)×(TN+FN)×(TP+FN)×(TN+FP))
Probability excess	*Prob. Excess*	(TP×TN-FP×FN)/((TP+FN)×(TN+FP))

* The definition of the abbreviations used: TP is the number of correctly classified disordered residues; FP is the number of ordered residues incorrectly classified as disordered; TN is the number of correctly classified ordered residues; and FN is the number of disordered residues incorrectly classified as ordered.

In order to conduct a five-fold cross validation, the chains in datasets PDB693 and D184 are randomly split into five subsets of approximately equal size. The results of uni-variant analysis on each property feature are shown in Table 5, in which the properties oriented from the same physicochemical group are put together for the following dependency analysis. The dependency analysis of feature selection aims to answer if a subset of a property group performs better than the original one.

It is observed in Table 5 that the performance of some physicochemical properties has been improved after they are split into order/disorder-based properties. In other words, purifying the physicochemical properties by considering the propensity for order or disorder contributes to the predicting power of the classifier. *Hydrophobic*$_O$ is the best property among all of them and gets an explicit improvement over *Hydrophobic*. On the other hand, neither *Polar*$_D$ nor *Polar*$_O$ get a better performance than *Polar*. In summary, the decomposition of some conventional properties by considering the order/disorder propensity brings explicit benefit in terms of the uni-variant analysis.

Table 5 The performance of each property in the uni-variant analysis on training data

Property	*Sens.*	*Spec.*	*MCC*	*Prob. Excess*
Hydrophobic	0.633	0.717	0.309	0.350
***Hydrophobic*$_O$**	**0.640**	**0.751**	**0.350**	**0.391**
Hydrophobic$_D$	0.519	0.723	0.217	0.241
Polar	**0.616**	**0.734**	**0.312**	**0.350**
Polar$_O$	0.603	0.703	0.269	0.306
Polar$_D$	0.604	0.731	0.299	0.335
Small	0.553	0.742	0.268	0.295
Small$_O$	0.555	0.688	0.214	0.243
***Small*$_D$**	**0.579**	**0.759**	**0.308**	**0.338**
Aliphatic	**0.601**	**0.748**	**0.314**	**0.349**
Aromatic	0.604	0.720	0.288	0.324
***Aromatic*$_O$**	**0.602**	**0.732**	**0.298**	**0.334**
Aromatic$_D$	0.538	0.660	0.173	0.198
Positive	**0.599**	**0.678**	**0.242**	**0.277**
Positive$_O$	0.573	0.662	0.204	0.235
Positive$_D$	0.583	0.667	0.218	0.250
Negative	**0.586**	**0.696**	**0.248**	**0.282**
Proline	**0.564**	**0.684**	**0.218**	**0.248**
Charged	**0.614**	**0.707**	**0.282**	**0.320**
Charged$_O$	0.571	0.664	0.204	0.235
Charged$_D$	0.603	0.706	0.272	0.309
Tiny	0.528	0.732	0.234	0.259
Tiny$_O$	0.577	0.675	0.220	0.252
***Tiny*$_D$**	**0.553**	**0.748**	**0.274**	**0.301**

The best performance among each property group is highlighted with bold font.

Table 6 Performance evaluation on *Hydrophobic*, *Aliphatic*, and *Aromatic*

Property	*Prob. Excess*
Hydrophobic$_O$	0.391
Aliphatic	0.349
Aromatic$_O$	0.334
Aliphatic + Aromatic$_O$	**0.413**

The best performance is highlighted with bold font.

Table 7 Performance evaluation on *Polar*, *Charged*, *Positive*, and *Negative*

Property	*Prob. Excess*
Polar	**0.350**
Charged	0.320
Positive	0.277
Negative	0.282
Positive + Negative	0.321

The best performance is highlighted with bold font.

Table 8 The result of the stepwise feature selection

Property	*Prob. Excess*
Aliphatic+Aromatic$_O$	0.412
Aliphatic+Aromatic$_O$+Polar	0.430
Aliphatic+Aromatic$_O$+Polar+Small$_D$	**0.437**
Aliphatic+Aromatic$_O$+Polar+Small$_D$+Proline	0.435

The best performance is highlighted with bold font.

After the best property for each group has been determined, a second level of dependency analysis is performed by considering the relations between physicochemical properties. The relation of these features is derived by incorporating the inheritance relationships between the child properties and their parent properties. That is, *Aliphatic* and *Aromatic$_O$* are subsets of *Hydrophobic$_O$*, *Tiny* is a subset of *Small*, *Positive* and *Negative* are subsets of *Charged*, which recursively is a subset of *Polar*. Based on these hierarchies, we aim to investigate if a combination of two subproperties performs better than the original one. According to the results shown in Table 6 and Table 7, property features *Aliphatic+Aromatic$_O$* performs better than *Hydrophobic$_O$*, but *Positive+Negative* is not superior to *Polar*.

After the dependency analysis, the redundancy removal step selects the best property from each cluster for the next step of feature selection. The selected representative properties are sorted by their performance in the uni-variant analysis, resulting in the following order: *Aliphatic+Aromatic$_O$*, *Polar*, *Small$_D$*, and *Proline*. The stepwise feature selection is preformed by adding one candidate property in each iteration until the predicting performance cannot be improved. The results of the stepwise feature selection are shown in Table 8, indicating that the four properties, *Aliphatic*, *Aromatic$_O$*, *Polar*, and *Small$_D$*, will be used in constructing the final RBFN classifier for predicting protein disorder. We name the final feature set of PSSMP with four properties as FS-PSSMP-4, the feature set of ten conventional physicochemical properties as FS-PSSMP-10, and the feature set employing the original PSSM as FS-PSSM.

All the experiments described above have been conducted with different window sizes of 11, 35, and 59, and the resulted feature sets are the same as reported. However, though they turn out the same result on feature selection, it is observed that larger window sizes such as 35 and 59 are favourable when prediction accuracy is considered. The current version of DisPSSMP adopts a window size of 47 and thus employs in total $4 \times 47 = 188$ attributes in the feature vector for a query residue.

In this subsection, we compare the performance of three feature sets, FS-PSSM, FS-PSSMP-10, and the proposed FS-PSSMP-4 on the independent testing data, which consists of three datasets, R80, U79, and P80. In the following discussions, the results on datasets U79 and P80 are always combined when they are reported, because U79 contains only fully disordered proteins and P80 comprises only completely ordered proteins. The results reveal that larger window sizes deliver better performance for all the feature sets on the dataset R80, and the performance of FS-PSSMP-4 and FS-PSSMP-10 are generally better than FS-PSSM.

In this subsection we investigate the performance of twelve web servers or packages in protein disorder prediction, some of which were included in comparison with the work of Yang *et al.* in their paper [14]. The predictors for comparison here are RONN [14], IUPred(short) [40, 41], IUPred(long) [40, 41], DISpro [42], DISOPRED2 [12, 36], PONDR® [25], DisEMBL(hot) [7], DisEMBL(465) [7], FoldIndex [23, 43], PreLink [22], GlobPlot [44], and DisEMBL(coils) [7]. DISOPRED2 has a limit of

1000 residues per protein, so 1HN0, 1FO4, and 1PS3 in dataset R80 and the u15 protein in dataset U79 have been removed from the blind testing data when testing DISOPRED2. IUPred provides two choices of predicting short or long disordered regions, and DisEMBL provides three choices: DisEMBL(hot), DisEMBL(465), and DisEMBL(coils). The results show the results in the way of *specificity* versus *sensitivity*, and the plots are rotated anticlockwise by 45° to be equivalent to the plot of *probability excess* = *sensitivity* + *specificity* – 1.

When compared with the other packages, DisPSSMP performs the best when *probability excess* is considered (with a *probability excess* of 0.60). DisPSSMP shows its ability in identifying the boundaries of ordered and disordered regions. The predictors IUPred(long), DISpro, DISOPRED2, DisEMBL(465), and PreLink have a *specificity* of more than 95% but a *sensitivity* of less than 50%, which show the tendency of predicting order more than disorder. In contrast, the predictor DisEMBL(coils) with a *sensitivity* of less than 50% but a *specificity* of more than 70% has the tendency of predicting disorder more than order. It depends on the users to select the predictors IUPred(long), DISpro, DISOPRED2, DisEMBL(465), and PreLink for under-prediction of disorder and DisEMBL(coils) for over-prediction.

The main purpose of the experiment on datasets U79 and P80 is checking whether a method is under-predicting or over-predicting protein disorder. The results of all the methods except IUPred(long) and FoldIndex are similar to that in the main blind testing dataset R80. The *sensitivity* of IUPred(long) and FoldIndex have an improvement of more then 20% in this experiment, and they are ranked as the first and the fourth among all methods. Since IUPred(long) has been designed for predicting context-independent global disorder that encompasses at least 30 consecutive residues in the predicted disordered regions and adopts a large window size like 101 [40, 41], it is suitable for the recognition of the fully globular proteins and the totally unstructured proteins. On the other hand, the training data of FoldIndex contains 91 totally unfolded proteins and 275 globular proteins, resulting in its good performance in discriminating fully ordered proteins from fully disordered proteins [23, 43]. Nevertheless, due to the lack of information about the boundaries between ordered regions and disordered regions, FoldIndex does not have a good performance in R80.

References:

1. Wright PE, Dyson HJ: **Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.** *J. Mol. Biol.* 1999, **293(2):**321-331.
2. Fink AL: **Natively unfolded proteins.** *Current Opinion in Structural Biology* 2005, **15:**35-41.
3. Jones DT, Ward JJ: **Prediction of disordered regions in proteins from position specific scoring matrices.** *Proteins* 2003, **53:**573-578.
4. Dunker AK, Garner E, Guilliot S, Romero P, Albercht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE: **Protein disorder and the evolution of molecular recognition: theory, predictions and observations.** *Pac Symp Biocomput* 1998, **3:**473-484.
5. Romero P, Obradovic Zoran, Dunker AK: **Sequence data analysis for long disordered regions prediction in the Calcineurin family.** *Genome Informatics* 1997; **8**:110-124.
6. Ferron, F., Longhi, S., Canard, B. and Karlin, D. (2006) A practical overview of protein disorder prediction methods. Proteins, 65, 1-14.
7. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB: **Protein disorder prediction: implications of structural proteomics.** *Structure (Camb)* 2003, **11:**1453-1459.
8. Su, C.T., Chen, C.Y. and Ou, Y.Y. (2006) Protein disorder prediction by condensed PSSM considering propensity for order or disorder. BMC Bioinformatics, 7, 319.
9. Su, C.-T. and Chen, C.-Y. (2006) A Two-stage RBFN Classifier for Protein Disorder Prediction. International Symposium on Biomedical Engineering (ISOBME).
10. Dunker AK, Obradovic Z, Romero P, Kissinger C, Villafrance E: **On the importance of being disordered.** *PDB Newsletter* 1997, **81:**3-5.
11. Plaxco KW, Gross M. Cell biology: **The importance of being unfolded.** *Nature* 1997, **386(6626):**657-659.
12. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT: **Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.** *J Mol Biol* 2004, **337:**635-645.
13. Lise S, Jones DT: **Sequence patterns associated with disordered regions in proteins.** *Proteins* 2005, **58:**144-150.
14. Yang ZR, Thomson R, McNeil P, Esnouf RM: **RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins.** *Bioinformatics Advance Access Published* June 9, 2005.
15. Schulz GE: **Nucleotide binding proteins.** *Molecular*

Mechanism of Biological Recognition, Elsevier/North-Holland Biomedical Press. 79-94.

16. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z: **Intrinsic disorder and protein function.** *Biochemistry* 2002, **41:**6573-6582.
17. Romero P, Obradovic Z, Kissinger C, Villafranca JE, Dunker AK: **Identifying disordered regions in proteins from amino acid sequence.** *Proc. IEEE Int. Conf. Neural Networks.* 1997, **1:**90-95.
18. Romero P, Obradovic Z, Kissinger C, Villafranca JE, Garner E, Guilliiot S, Dunker AK: **Thousands of proteins likely to have long disordered regions.** *Pac Symp Biocomput* 1998, **3:**437-448.
19. Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK: **Predicting intrinsic disorder from amino acid sequence.** *Proteins* 2003, **53:**566-572.
20. Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradovic Z: **Optimizing long intrinsic disorder predictiors with protein evolutionary information.** *Journal of Bioinformatics and Computational Biology* 2005, **3:**35-60.
21. Garner E, Romero P, Dunker AK, Brown C, Obradovic Z: **Predicting binding regions within disordered proteins.** *Genome Informatics* 1999, **10:**41-50.
22. Coeytaux K, Poupon A. **Prediction of unfolded segments in a protein sequence based on amino acid composition.** *Bioinformatics* 2005, **21:**1891-1900.
23. Uversky VN, Gillespie JR, Fink AL: **Why are “natively unfolded” proteins unstructured under physiologic conditions?** *Proteins* 2000, **41:**415-427.
24. Garner E, Cannon P, Romero P, Obradovic Z, Dunker AK: **Predicting disordered regions for amino acid sequence: common themes despite differing structural characterization.** *Genome Inform. Ser. Workshop Genome Inform.* 1998, **9:**201-213.
25. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK: **Sequence complexity of disordered protein.** *Proteins* 2001, **42:**38-48.
26. Vucetic S, Brown CJ, Dunker AK, Obradovic Z: **Flavors of protein disorder.** *Proteins* 2003, **52:**573-584.
27. Li X, Romero P, Rani M, Dunker AK, Obradovic Z: **Predicting protein disorder for N-, C-, and internal regions.** *Genome Inform Ser Workshop Genome Infor* 1999, **10:**30-40.
28. Shimizu K, Hirose S, Noguchi T, Muraoka Y: **Predicting the protein disordered region using modified position specific scoring matrix.** *Genome Informatics* 2004, P150.
29. Radivojac P, Obradovic Z, Brown CJ, Dunker AK: **Prediction of boundaries between intrinsically ordered and disordered protein regions.** *Pac. Symp. Biocomput.* 2003, **8**:216-227.
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucl. Acids Res.* 2000, **28:**235-242.
31. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Lakoucheva LM, Cortese MS, Lawson JD, Brown CJ, Sikes JG, Newton CD, Dunker AK: **DisProt: a database of protein disorder.** *Bioinformatics* 2005, **21:**137-140.
32. Li W, Jaroszewski L, Godzik A: **Clustering of highly homologous sequences to reduce the size of large protein databases.** *Bioinformatics* 2001, **17:**282-283.
33. Li W, Jaroszewski L, Godzik A: **Tolerating some redundancy significantly speeds up clustering of large proteins databases.** *Bioinformatics* 2002, **18:**77-82.
34. Melamud E, Moult J: **Evaluation of disorder predictions in CASP5.** *Proteins* 2003, **53**:561-565.
35. Jin Y, Dunbrack RL: **Assessment of disorder predictions in CASP6.** *Proteins* 2005; **Early View**
36. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT: **The DISOPRED server for the prediction of protein disorder.** *Bioinformatics* 2004, **20:**2138-2139.
37. Zhang Q, Yoon S, Welsh WJ: **Improved method for predicting β–turn using support vector machine.** *Proteins* 2005, **21:**2370-2374.
38. Chen YC, Lin YS, Line CJ, Hwang JK: **Prediction of the bonding states of cysteines Using the support vector machines based on multiple feature vectors and cysteine state sequences.** *Proteins* 2004, **55:**1036-1042.
39. Natt NK, Kaur H, Raghava GPS: **Prediction of transmembrane regions of β–barrel proteins using ANN- and SVM-based methods.** *Proteins* 2004, **56:**11-18.
40. Dosztányi Z, Csizmok V, Tompa, Simon I: **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.** *Bioinformatics* 2005, **21:**3433-3434.
41. Dosztányi Z, Csizmok V, Tompa, Simon I: **The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins.** *J Mol Biol* 2005, **347:**827-839.
42. Cheng J, Sweredoski MJ, Baldi P: **Accurate prediction of protein disordered regions by mining protein structure data.** *Data Mining and Knowledge Discovery* 2005, **11:**213-222.
43. Prilusky J, Felder CE, Zeev-BenMordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Summan JL: **FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded.** *Bioinformatics* 2005, **21:**3435-3438.
44. Linding R, Russell RB, Neduva V, Gibson TJ: **GlobPlot: exploring protein sequences for globularity and disorder.** *Nucl. Acids Res.* 2003, **31:**3701-3708.

45. Dunker AK, Brown CJ, Obradovic Z: **Identification and functions of usefully disordered proteins.** *Adv. Protein Chem.* 2002, **62:**25-49.

46. Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.** *Nucl. Acids Res.* 1997, **25(7):**3389-3402.

47. Zhang Q, Yoon S, Welsh WJ: **Improved method for predicting β-turn using support vector machine.** *Bioinformatics* 2005, **21**:2370-2374.

48. Chemical classifications http://prowl.rockefeller.edu/aainfo/pchem.htm.

49. **QuickRBF** http://muse.csie.ntu.edu.tw/~yien/quickrbf/index.php.

50. John GH, Kohavi R, Pfleger K: **Irrelevant features and the subset selection problem.** In *Machine Learning: Proceedings of the Eleventh International Conference, Morgan Kaufmann* 1994, P121-129.

51. Boz O: **Feature subset selection by using sorted feature relevance.** *ICMLA International Conference on Machine Learning and Application, Las Vegas City, USA*, June, 2002, P147-153.