POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions

Shuichi Hirose\textsuperscript{1,3,*}, Kana Shimizu\textsuperscript{2}, Satoru Kanai\textsuperscript{1}, Yutaka Kuroda\textsuperscript{3} and Tamotsu Noguchi\textsuperscript{2}

\textsuperscript{1}PharmaDesign, Inc., Tokyo 104-0032, \textsuperscript{2}Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064 and \textsuperscript{3}Department of Biotechnology and Life Science, Graduate School of Engineering, Tokyo University of Agriculture and Technology, Koganei 184-8588, Japan

Received on January 31, 2007; revised on May 29, 2007; accepted on May 30, 2007

Advance Access publication June 1, 2007

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Recent experimental and theoretical studies have revealed several proteins containing sequence segments that are unfolded under physiological conditions. These segments are called disordered regions. They are actively investigated because of their possible involvement in various biological processes, such as cell signaling, transcriptional and translational regulation. Additionally, disordered regions can represent a major obstacle to high-throughput proteome analysis and often need to be removed from experimental targets. The accurate prediction of long disordered regions is thus expected to provide annotations that are useful for a wide range of applications.

Results: We developed Prediction Of Order and Disorder by machine LEarning (POODLE-L: L stands for long), the Support Vector Machines (SVMs) based method for predicting long disordered regions using 10 kinds of simple physico-chemical properties of amino acid. POODLE-L assembles the output of 10 two-level SVM predictors into a final prediction of disordered regions. The performance of POODLE-L for predicting long disordered regions, which exhibited a Matthew's correlation coefficient of 0.658, was the highest when compared with eight well-established publicly available disordered region predictors.

Availability: POODLE-L is freely available at http://mbs.cbrc.jp/poodle/poodle-l.html

Contact: hirose-shuichi@aist.go.jp

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Over the last two decades, many proteins lacking well-defined 3D structures under physiological conditions have been identified using various experimental techniques, such as nuclear magnetic resonance, circular dichroism and X-ray crystallography (Dunker et al., 2001). Those proteins include over 100 proteins or domains that are unfolded over their entire sequences (Uversky et al., 2000), and are called ‘intrinsically disordered proteins’ or ‘intrinsically unstructured proteins (IUPs)’ (Dunker et al., 2001; Tompa, 2002). In addition, even more numerous folded proteins containing unfolded regions, called ‘disordered regions’, have been identified (Romero et al., 1997a). From a structural viewpoint, disordered regions are sequence segments that are not observed on the electron density map in X-ray crystallography (Dunker and Obradovic, 2001). Moreover, theoretical analysis suggests that both IUPs and disordered regions are quite common and are found in proteins from all kinds of organisms, especially in eukaryotic genomes (Dunker et al., 2000; Oldfield et al., 2005a; Ward et al., 2004).

From a functional standpoint, though unstructured, disordered regions play fundamental roles in biological activities (Dunker and Obradovic, 2001; Dunker et al., 2002a, b; Dyson and Wright 2005; Radivojac et al., 2007; Tompa, 2002; Uversky, 2002, 2003) and are associated with diseases (Fink et al., 2005; Iakoucheva et al., 2002). A four-class classification of their functions has recently been proposed: entropic chain, protein modification, molecular assembly/disassembly and molecular recognition (Dunker et al., 2002a). A unique feature when carrying out such functions is their ability to bind multiple partners with high specificity (Dunker et al., 2002a). This unique property is assumed to originate from both the multiple conformations that they can adopt and their binding surfaces, which are typically larger than those of a folded globular protein. Such molecular recognition peculiarities render disordered regions particularly suitable for functions related to signal transduction, cell-regulation and transcription (Dunker et al., 2005; Uversky et al., 2005). Disordered regions have therefore attracted great attention because they might yield a new structure-function paradigm.

Clear sequence differences between ordered and disordered regions have been demonstrated (Romero et al., 1997b; Uversky et al., 2000; Wootton, 1994), thereby stimulating numerous attempts to predict disordered regions solely from amino acid sequences. A potential area of application of disordered region prediction is high-throughput functional/structural proteomics. In such projects, putative disordered regions and IUPs, which often hamper protein analysis, are predicted and removed from experimental targets, as reported in the Center for Eukaryotic Structural Genomics (CESG) target selection (Oldfield et al., 2005b). Hence, the improvements of prediction performances might also have a direct and practical impact on increasing the efficiency of proteomics projects. Additionally, a novel drug discovery strategy, termed disorder-based rational drug

\*To whom correspondence should be addressed.
design, may represent another area of application for disorder region prediction (Cheng et al., 2006).

Methods for predicting disordered regions are classifiable into two groups: the first uses mainly the physico-chemical properties of the amino acids; the second uses evolutionary information. PONDOR (Romero et al., 2001), GlobPlot2 (Linding et al., 2003a), DisEMBL (Linding et al., 2003b), IUPred (Dosztanyi et al., 2005), PreLINK (Coeytaux and Poupon 2005) and FoldUnfold (Galzitskaya et al., 2006a) are included in the first group. For example, PONDOR uses amino acid composition and sequence complexity; GlobPlot2 uses a disorder propensity index; IUPred uses pairwise energy based on a quadratic form in the amino acid composition of protein; PreLINK uses a relationship between amino acid distribution and a putative hydrophobic cluster and FoldUnfold uses the ability to form a sufficient number of contacts in the globular state. Prediction methods included in the second group mainly use profiles generated from PSI-BLAST [e.g. DISOPRED2 (Ward et al., 2004), DISPro (Cheng et al., 2005) and DisPSSMP (Su et al., 2006)] or multiple alignment [e.g. RONN (Yang et al., 2005)].

In spite of those efforts, the critical assessment of techniques for protein structure prediction (CASP) experiment, which has had a category for the prediction of disordered regions since 2002, suggested that substantial room for improvement remains (Obradovic et al., 2005). One promising line of research for improving the prediction of disordered prediction suggests that the sequence characteristics of the disordered regions might depend on their lengths (Peng et al., 2006).

In this study, we investigated whether length dependence is useful for improving the prediction efficiency of disordered regions. We developed Prediction Of Order and Disorder by machine LEarning (POODLE-L; L stands for long), which is an SVM (Support Vector Machine) predictor specifically intended to detect long disordered regions. The prediction performances were improved by extracting the amino acids features specific to long disordered regions. When compared to eight publicly available and well-established disordered region predictors using an independent evaluation dataset of 116 sequences, POODLE-L demonstrated the highest prediction performance.

2 METHODS

2.1 Training and evaluation dataset

Our training dataset (TDS) was constructed as follows. Ordered regions were collected from proteins in the Protein Data Bank (PDB) (Berman et al., 2000) that include no disordered regions or only 'short disordered regions' (shorter than 30aa). Representative sequences with pairwise sequence identity of <30% were collected using PDB-REPRDB (Noguchi and Akiyama, 2003). From that set, we selected protein structures that are monomeric single domains, as defined by SCOP (Murzin et al., 1995), which have resolutions better than 2.0Å, and which are determined with CNS (Brunger et al., 1998), Shelx (Sheldrick, 1997), Refmac (Murshudov et al., 1997) or X-PLOR (Brunger, 1992). Eventually, the ordered regions of TDS included 292 protein sequences (55,784 residues), and contained 93 'short disordered regions' (2% of the total residue number).

Disordered regions of TDS consisted of long disordered regions and IUPs. They were collected from Uversky's (Uversky et al., 2000) and DisProt ver. 2.2 datasets (Vucetic et al., 2005). Disordered regions shorter than 40aa, as well as redundant sequences with sequence identities higher than 90%, were removed according to BLASTClust (Altschul et al., 1990). Consequently, 199 disordered regions (35,428 residues) were collected.

Two prediction performance assessment datasets were prepared. A first dataset, called ADS-1, was used for assessing the prediction performance of the individual predictors. The ADS-1 was constructed according to the same protocol as that used for constructing TDS, except that the required resolution was 2.5Å rather than 2.0 for ordered regions, and was removed all sequences contained in TDS. We used an updated version of DisProt (ver. 3.0) for disordered regions. Disordered regions shorter than 30aa were excluded from the dataset described above. The evaluation dataset comprised 53 ordered regions (11,431 residues) derived from PDB and 63 disordered regions (8,700 residues) derived from DisProt.

A second prediction performance assessment dataset (ADS-2) was used for selecting the descriptor's combinations and was constructed from the PDB according to a protocol similar to that used for constructing TDS. Disordered regions in ADS-2 were defined as a string of 30 or more consecutive residues missing their Cα atomic coordinates. The ADS-2 consisted of 15 sequences with no disordered region and 11 sequences containing one or more long (>30aa) disordered regions, representing 6688 ordered residues, and 564 disordered residues. TDS, ADS-1 and ADS-2 do not overlap with each other, and all sequences are available at http://mbs.cbrc.jp/poodle/poodle-l-datasets.html.

2.2 Two-level SVM prediction

POODLE-L assembles the results of 10 disordered region predictors into a final prediction. Each predictor consists of a two-level SVM prediction, which uses amino acid sequences as input data. The first-level SVM predicts the probability of a 40-residue sequence segment to be disordered. The window size was chosen by comparing, for each descriptor, the classification performances of a 30-residue window with those of a 40-residue window. The sequence was expressed as a 10-D vector, with each component corresponding to a physico-chemical characteristic described in Section 2.3 and used as learning data of the first-level SVM (Fig. 1 Step 1). The disorder probabilities of all 40-residue segments in the protein sequence were calculated using the first-level SVM (Fig. 1 Step 2) and were expressed as real numbers from 0.0 to 1.0 (Fig. 1 Step 3).

The second-level SVM uses the output of the first-level SVM and computes the disorder probability of a single residue. The second-level SVM learning data of each residue were prepared by subdivision into 10 classes using an increment of 0.1 (0.0–0.1, 0.1–0.2, 0.2–0.3, etc.) the disorder probabilities, as calculated with the first-level SVM, of all the windows that include the examined residue (Fig. 1 Step 4). The number of members in each class was normalized to one using the maximum number found within the protein sequence. Consequently, each amino acid residue was expressed as a 10-D vector to be used in the second-level SVM. Note that 'internal residues', which are located more than 40 residues distant from the protein N and C termini, are included in 40 windows, whereas 'termini residues', which are residues located within 39 residues of the protein N and C termini, are included in a smaller number of windows (Fig. 1 Step 3). Thus, the 10-D vectors of termini residues are calculated using a smaller number of first-level SVM outputs than the internal residue 10-D vectors, and the second level SVMs for internal and termini residues were also trained separately. The second-level SVM output was expressed as a real number from 0 to 1, which is called the disorder probability (P),
Residues with disorder probability higher than 0.5 were predicted to be in the disordered state (Fig. 1 Step 5).

2.3 Physico-chemical properties used as descriptors

Ten descriptors were defined as follows and were used in the first-level SVMs:

- The mean hydrophobicity was defined as the average of the modified Kyte and Doolittle's hydrophobic index (Kyte and Doolittle, 1982) of the residues in the window, and normalized to a scale of 0 to 1.

- The hydrophobic cluster value was defined as the length of a hydrophobic cluster divided by 25. A hydrophobic cluster was defined by encoding sequences with a ternary code, i.e. 1 for hydrophobic residues (VILFMYW), 2 for a proline and 0 for the other residues.

- A hydrophobic cluster required that it is constituted by a string of '1's and '0's with a maximum of three consecutive '0's including no '2's. It therefore begins and ends with either '0000' or a '2' (Coeytaux and Poupon, 2005).

- The mean net charge was defined as the absolute value of the average sample means of

\[ P(st)_{ij} = \frac{\sum (F(st)_{ij}) \cdot \sum (F(st))}{\sum (F(st)_{ij})} \]

where \( n_{ij} \) is the occurrence count of amino acid \( i \) in the sequence \( j \), which has a length \( n_j \). In addition, \( F(st) \) was calculated using SWISS-PROT release 47.0. The amino acid frequencies, which are named \( P(\text{ordered}) \) and \( P(\text{disordered}) \), were calculated respectively from the ordered and disordered regions in the TDS. Similarly, \( P(\text{query}) \) was calculated for the query sequence. Next, the Pearson's correlation coefficient between \( P(\text{ordered}) \) and \( P(\text{disordered}) \), and \( P(\text{query}) \), was calculated as:

\[ r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \]

where \( x_i \) and \( y_i \) respectively indicate the value of \( P(\text{query}) \) and \( P(\text{ordered}) \) (or \( P(\text{disordered}) \)), \( \bar{x} \) and \( \bar{y} \) respectively denote the sample means of \( x \) and \( y \).

The secondary structures were described using two descriptors. In the first descriptor, a window containing a residue in an alpha helix

Fig. 1. Schematic representation of a two-level SVM disordered region predictor. The first- and second-level SVMs are shown respectively on the left and right panels. In the first level, a 40-residue window is moved from the N-terminus to C-terminus protein sequence (Step 1). The matrix in the left panel center represents the value to which the sequences are converted using the descriptors, and the name of sequence. For example, seq11 denotes the 40-residue sequence window that starts at residue 11 (Step 1). The first-level SVM is executed using the matrix generated in Step 1 (Step 2). The values in squares represent the disorder probability of the 40-residue sequence segment calculated with the first-level SVM (Step 3).

In the right panel, the square shows the value at which the disorder probability distribution is standardized (Step 4). The second-level SVM is executed using the matrix generated in Step 4 (Step 5).
core region was given a score of 1, and 0 otherwise. A second descriptor was derived similarly, but with beta core residues. The core regions were calculated according to a protocol similar to the Chou–Fasman’s secondary structure prediction method (Chou and Fasman, 1978). In short, a residue was defined as an alpha core region when four or more consecutive residues with alpha helix propensity higher than 1.15 were found with no flanking residues with alpha helix propensity of less than 0.8. Beta sheet core regions were predicted when three or more consecutive residues with beta sheet propensity of greater than 1.20 were found with no flanking residues with beta sheet propensity smaller than 0.8.

The average number of contacts was defined as the average of the expected number of contacts in globular state of all residues in the window (Garbuzynskiy et al., 2004).

2.4 Training method

The SVM used in this study is the libSVM package (Chang and Lin, 2001) with an RBF kernel for the classifiers. The first-level and second-level SVMs were trained using the respective training dataset containing 10-D vectors as described in Physico-chemical properties used as descriptors of section 2.3. Both TDSs were reduced to 1/10th of their original sizes as follows. The TDS was subdivided into clusters, whose number was 1/200th of all data, using a k-means nearest neighbor method. We then randomly selected 20 sequences from each cluster, which yielded the 1/10 reduced dataset. The training parameters (cost and gamma parameter) used in SVM were optimized using a 5-fold cross-validation with the reduced 1/10th dataset.

2.5 Assessment of predictions

The prediction results were assessed on a residue basis: the state of each residue in the protein sequence was predicted and was compared with the experimental state. The prediction results were classified into four categories: $N_{TP}$ is the number of true positives, which is defined as the number of correctly predicted disordered regions. Similarly, $N_{FP}$, $N_{TN}$ and $N_{FN}$ denote the numbers of false positives, which are defined respectively as: ordered residues that were incorrectly predicted as disordered; the number of true negatives, which are defined as correctly predicted ordered residues and the number of false negatives, which are defined as disordered residues incorrectly predicted as ordered.

The first assessment criterion is the receiver operating characteristic (ROC) curve. The ROC curve is obtained by plotting the false positive rate ($R_{FP}(P) = N_{FP}(P)/(N_{TN} + N_{FP})$) against the true positive rate ($R_{TP}(P) = N_{TP}(P)/(N_{TP} + N_{FN})$). The $R_{FP}$ against $R_{TP}$ was plotted while the disorder probability increased from 0 to 1.0 with a 0.01 increment. The larger the area under the ROC curve ($S_{AUC}$), the more robust an algorithm is. An area of 1.00 means a perfect predictor, and an area of 0.50 corresponds to a random guess.

Next, the sensitivity ($S_{sens}$) and specificity ($S_{spec}$), which respectively indicate the fraction of correctly identified disordered regions and ordered regions, were used to evaluate prediction performances. In addition, $S_{sens}$ and $S_{spec}$ were defined as follows.

\[
S_{sens} = \frac{N_{TP}}{N_{TP} + N_{FN}}, \quad S_{spec} = \frac{N_{TN}}{N_{TN} + N_{FP}}
\]

Another commonly used criterion is the Matthew’s correlation coefficient ($S_{MCC}$), which is given as:

\[
S_{MCC} = \frac{N_{TP}N_{TN} - N_{FP}N_{FN}}{\sqrt{(N_{TP} + N_{FN})(N_{TP} + N_{FP})(N_{TN} + N_{FP})(N_{TN} + N_{FN})}}
\]

For a prediction with unequal class frequencies, $S_{MCC}$ favors the correct prediction of small classes. In our case, $S_{MCC}$ will favor the correct prediction of disordered regions over that of ordered regions.

In CASP6, a new criterion, $S_{product}$ was defined to complement $S_{MCC}$ and emphasize the detection of disordered regions (Jin and Dunbrack, 2005). There are considerably fewer disordered residues than ordered residues. Moreover, disordered regions tend to be underpredicted. $S_{product}$ is defined as:

\[
S_{product} = S_{sens}S_{spec} = \frac{N_{TP}N_{TN}}{N_{disorder}N_{order}}
\]

$S_{product}$ ranges from 0 to 1, where 1 indicates a perfect prediction. $S_{product}$ rises much faster than $S_{MCC}$ when the number of correctly predicted disordered residues rises.

3 RESULTS

3.1 Basic two-level disorder predictor

We constructed a basic two-level SVM predictor using all 10 descriptors, and compared its prediction performances with those of an ordinary single-level SVM prediction method, which is identical to the first-level SVM of our basic predictor (Fig. 1). The average prediction performances were estimated by training and evaluating both predictors 20 times using randomly selected data from TDS (Table 1). The $S_{sens}$ of the two-level prediction was 1.7% higher than that of a usual single SVM prediction. In addition, the $S_{MCC}$ and $S_{product}$ were also slightly higher. Furthermore, the two-level SVM prediction algorithm enabled the prediction of disordered regions for residues in terminal regions (see Methods section for further details).

3.2 Combination of descriptors for the first-level SVM

We optimized the first-level SVM by investigating which descriptor combination would yield a better prediction result. However, because of the numerous combinations (1022 patterns), we classified the 10 descriptors into six groups, which reduced the number of combinations to 63 patterns (62 patterns and the basic predictor). The six groups were the hydrophobicity descriptors (mean hydrophobicity, hydrophobic cluster value), the charge descriptors (mean net charge, charge cluster value), the sequence complexity descriptor (sequence complexity), the amino acid composition descriptors (amino acid composition), the secondary structure descriptors (secondary structure) and the average number of contacts descriptor (average number of contacts). We evaluated the prediction performances of all 63 predictors based on their sensitivity and specificity using ADS-2 (Fig. 2). According to this criterion, 9 among the 62 predictors exhibited better performance than the basic predictor; the performance of predictor 34 was the highest (Supplementary Table 1).

3.3 Building POODLE-L

To improve the prediction performance, we assembled the results of the basic predictors and that of the nine predictors with performances higher than that of the basic predictor into a consensus prediction system, POODLE-L. We calculated the average probability of each predictor using 7, 23 and 39-residue windows and attributed the average probabilities to the central residue (Fig. 3 Step 1). For each window and each residue, the two largest and smallest averaged probabilities were omitted (Fig. 3 Step 2). The final disorder probability of
two-level 0.657 /C6 and specificity (learning data. The prediction line performed better than the basic predictor and are identified by their identity numbers. The prediction $S_{\text{sens}}$ and $S_{\text{spec}}$ were averaged over 20 training and evaluation iterations performed using randomly selected learning data.

Table 1. Prediction performance of one-level and two-level SVM prediction

<table>
<thead>
<tr>
<th></th>
<th>$S_{\text{sens}}$</th>
<th>$S_{\text{spec}}$</th>
<th>$S_{\text{MCC}}$</th>
<th>$S_{\text{product}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>one-level</td>
<td>0.640±0.014</td>
<td>0.757±0.016</td>
<td>0.361±0.013</td>
<td>0.485±0.009</td>
</tr>
<tr>
<td>two-level</td>
<td>0.657±0.019</td>
<td>0.751±0.020</td>
<td>0.369±0.013</td>
<td>0.493±0.009</td>
</tr>
</tbody>
</table>

One-level prediction was implemented using a 41aa window size. The result of prediction for each window was assumed to represent the disorder probability of the central amino acid to be disordered. One-level prediction cannot compute probabilities for the 20 amino acids at the N- and C-termini of the sequence. The 20 termini residues were also excluded when assessing the two-level prediction. The performance is measured using the sensitivity ($S_{\text{sens}}$), the specificity ($S_{\text{spec}}$), the Matthew’s correlation coefficient ($S_{\text{MCC}}$), and the CASP weighted score ($S_{\text{product}}$), as defined in Section 2.5.

an amino acid was calculated as the average value over 18 probabilities (Fig. 3 Step 3). The example for the disordered region prediction for two proteins in ADS-1 is shown in Figure 4.

3.4 Comparison with other methods

We compared the prediction performance of POODLE-L was compared to that of eight publicly available disordered region predictors: DISOPRED2, VSL2 (Peng et al., 2006), VL3H (Obradovic et al., 2003), DisEMBL, RONN, IUPred, FoldIndex (Prilusky et al., 2005) and FoldUnfold using ADS-1. According to the ROC curve, POODLE-L performance was the highest among all predictors, especially in the very low false positive rate region. Furthermore, POODLE-L’s prediction performances were also the highest according to $S_{\text{MCC}}, S_{\text{product}}$, which provide a simultaneous evaluation of $S_{\text{sens}}$

Fig. 2. Influence of the descriptor on the predictor’s sensitivity ($S_{\text{sens}}$) and specificity ($S_{\text{spec}}$). The basic predictor is shown by a square. The other predictors are shown with crosses. A performance borderline, where the sum of the sensitivity and specificity is equal to that of the basic predictor, is shown with a dotted line. Predictors upper the dotted line performed better than the basic predictor and are identified by their identity numbers.

Fig. 3. Assembling scheme of the predictors (POODLE-L). pred_15 to pred_53 represent individual predictors shown in Figure 2, and pred_basic represents the basic predictor, which uses all 10 descriptors in the first-level SVM. Ten predictors predict the probabilities of disordered regions from query sequence, respectively (Step 1). The length shows the window length, and the numbers in the dotted line frame are the probabilities calculated with each individual predictor corresponding to an n-th amino acid residue, and used to calculate the POODLE-L’s prediction of disordered regions (Step 2). The mean value is computed using the numbers in the dotted line frame for all residues in a protein (Step 3).

Fig. 4. Two examples of POODLE-L disorder prediction. The residue number is shown on the horizontal axis. The disorder probability of each residue is represented on the vertical axis. Residues with disorder probability higher than a threshold value of 0.5 are predicted as disordered. A horizontal line at a probability of 0.5 shows the threshold. The gray regions show the disordered as defined in DisProt. (A) Suppressor of cytokine signaling 3 (DisProt code: DP00446), (B) Subtilisin E (precursor) (DisProt code: DP00394) from DisProt.
Table 2. Prediction performances of POODLE-L and of eight publicly available disordered region predictors

<table>
<thead>
<tr>
<th>Method</th>
<th>$S_{\text{sens}}$</th>
<th>$S_{\text{spec}}$</th>
<th>$S_{\text{MCC}}$</th>
<th>$S_{\text{product}}$</th>
<th>$S_{\text{AUC}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>POODLE-L</td>
<td>0.669</td>
<td>0.949</td>
<td>0.658</td>
<td>0.635</td>
<td>0.873</td>
</tr>
<tr>
<td>VLSH</td>
<td>0.734</td>
<td>0.858</td>
<td>0.600</td>
<td>0.630</td>
<td>0.856</td>
</tr>
<tr>
<td>VSL2</td>
<td>0.755</td>
<td>0.794</td>
<td>0.547</td>
<td>0.599</td>
<td>0.844</td>
</tr>
<tr>
<td>DISOPRED2</td>
<td>0.635</td>
<td>0.939</td>
<td>0.616</td>
<td>0.596</td>
<td>0.848</td>
</tr>
<tr>
<td>FoldIndex</td>
<td>0.598</td>
<td>0.959</td>
<td>0.614</td>
<td>0.574</td>
<td>0.849</td>
</tr>
<tr>
<td>IUPred$^b$</td>
<td>0.595</td>
<td>0.956</td>
<td>0.607</td>
<td>0.569</td>
<td>0.853</td>
</tr>
<tr>
<td>RONN</td>
<td>0.628</td>
<td>0.837</td>
<td>0.478</td>
<td>0.525</td>
<td>0.797</td>
</tr>
<tr>
<td>FoldIndex</td>
<td>0.622</td>
<td>0.844</td>
<td>0.481</td>
<td>0.525</td>
<td>0.733</td>
</tr>
<tr>
<td>DisEMBL$^c$</td>
<td>0.245</td>
<td>0.964</td>
<td>0.312</td>
<td>0.236</td>
<td>0.733</td>
</tr>
</tbody>
</table>

The prediction performances were measured with the parameters used in Table 1 complemented with the area under the ROC curve ($S_{\text{AUC}}$). All predictions of disordered regions were performed with the respective default options and parameters, except for the following cases.

$^a$FoldUnfold predictions were performed with an average frame of 41 residues, as recommended in the paper (Gulati et al., 2006b).

$^b$IUPred provides three options for predicting disordered regions. The present predictions were performed with the long disorder option.

$^c$DisEMBL provides three options for disordered regions: loops/coils, hot loops (meaning high B factors) and missing coordinates (defined by REMARK465 in PDB). The REMARK465 option was adopted in the definition of disordered region.

The ability of POODLE-L for predicting short disordered regions was low, as anticipated. The assessment dataset for estimating the prediction performance consisted of 15 sequences with short disordered regions (between 5aa and 20aa) from X-ray crystallographic data, in which 3500 residues were classed as ordered and 204 short disordered residues. The $S_{\text{sens}}, S_{\text{spec}}, S_{\text{MCC}}$ and $S_{\text{product}}$ of POODLE-L were 0.353, 0.881, 0.158 and 0.311, respectively, and its performance on several criteria was lower than that of the other predictors (Supplementary Table 2).

4 DISCUSSION

POODLE-L is a disordered region predictor that is especially tuned for predicting long disordered regions. A major feature of POODLE-L is the assembly of multiple individual disordered region predictors into a final integrated predictor. The outputs of 10 two-level SVM predictors are merged using a consensus algorithm. One predictor, the basic predictor, uses all 10 descriptors in the first-level prediction SVM, whereas the other predictors use combinations of descriptors that yielded the better performances. The effectiveness of a consensus prediction algorithm was reported previously for secondary structure prediction (Cuff et al., 1998; Nishikawa and Noguchi, 1991), but to our knowledge this is its first application to the prediction of disordered regions. The prediction performance of POODLE-L was, indeed, 2.6–8.0% higher than that of the individual predictors according to $S_{\text{product}}$ (Table 2 and Supplementary Table 1).

Table 3. List of descriptors used in the first-level SVM that exhibited prediction performances higher than the basic predictor

<table>
<thead>
<tr>
<th>Predictor name</th>
<th>Descriptors$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>predictor15</td>
<td>hydrophobicity, contact</td>
</tr>
<tr>
<td>predictor23</td>
<td>charge, hydrophobicity, composition</td>
</tr>
<tr>
<td>predictor32</td>
<td>hydrophobicity, complexity, composition</td>
</tr>
<tr>
<td>predictor34</td>
<td>hydrophobicity, complexity, contact</td>
</tr>
<tr>
<td>predictor42</td>
<td>charge, hydrophobicity, complexity, composition</td>
</tr>
<tr>
<td>predictor44</td>
<td>charge, hydrophobicity, complexity, contact</td>
</tr>
<tr>
<td>predictor46</td>
<td>charge, hydrophobicity, composition, contact</td>
</tr>
<tr>
<td>predictor49</td>
<td>charge, complexity, composition, contact</td>
</tr>
<tr>
<td>predictor53</td>
<td>hydrophobicity, complexity, composition, contact</td>
</tr>
</tbody>
</table>

The predictor name are the same as those in Figure 2.

$^a$Abbreviations: complexity = sequence complexity, composition = amino acid composition and contact = average number of contacts.

Descriptors that are most useful in POODLE-L for identifying disordered regions can be evaluated from their inclusion in the predictors that produce the best results (Table 3). We find that hydrophobicity descriptors appear in all but one predictor, although the secondary structure is not used in any predictor. That difference does not necessarily mean that disorder and secondary structure propensities are unrelated, but the difference might indicate that the encoding of secondary structure descriptor can be improved. For example, we might use secondary prediction methods such as PSIPRED (McGuffin et al., 2000), which are probably more accurate than our Chou–Fassman-derived prediction scheme. As for the four other descriptor’s groups, one or several of them appear in all predictors, but they appear in various combinations. This interchangeability suggests that the descriptors encode redundant information (Table 3). Overall, though the descriptors defined in this study might not characterize disordered regions exhaustively, hydrophobicity seems likely to be a good indicator for long disordered regions.

POODLE-L exhibited the best prediction performance for long disordered regions when compared with eight publicly available disordered region predictors (Table 2 and Fig. 5), but it was poor at predicting short regions (Supplementary Table 2). Among other reasons, we infer that the high performance of POODLE-L in predicting long disordered regions originates from the large window, which enables the comprehension of information from amino acids that are located distant from each other. This conjecture concurs with our preliminary calculation, indicating that a 40-residue window is more effective than a 30-residue window. Shorter windows are typically used for predicting disordered regions, and IUPred and RONN, e.g. respectively use 21-residue and 19-residue windows.

The performance of POODLE-L for predicting disordered regions shorter than 30 residues was poor, but it was anticipated because it was trained exclusively for predicting long disordered regions. Therefore, the ability of POODLE-L to predict, exclusively and with high reliability, long disordered...
regions suggests that the difference in the physico-chemical properties of long and short disordered regions (Radivojac et al., 2004) can be used to improve the prediction of the disordered region, as suggested by Peng (Peng et al., 2006). In our case, short disordered regions are identified by the method introduced in Shimizu’s work (Shimizu et al., 2005).

5 CONCLUSION

We described POODLE-L, which detects long disordered regions with high accuracy. POODLE-L achieved the best prediction performance among several previously reported prediction methods, according to several criteria. The prediction of long disordered regions appears to stem from POODLE-L’s capability to recognize the sequential differences among long and short disordered regions and ordered regions efficiently. Nevertheless, we believe that the prediction of long disordered regions might be further improved by exploring better descriptors.

ACKNOWLEDGEMENTS

We thank our colleagues in PharmaDesign, Inc. and protein function team in Computational Biological Research Center (CBRC) for helpful advice and discussion. This work was funded by PharmaDesign, Inc. and Advanced Industrial Science and Technology (AIST).

Conflict of Interest: none declared.

REFERENCES


Sheldrick, G.M. (1997) SHELX97, programs for crystal structure analysis (Release 97-2). University of Gottingen, Germany.


