POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix

Kana Shimizu1,* , Shuichi Hirose2 and Tamotsu Noguchi1

1Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064 and 2PharmaDesign, Inc., 2-19-8 Hatchobori, Chuo-ku, Tokyo 104-0032, Japan

Received on April 4, 2007; revised on May 2, 2007; accepted on June 16, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Protein disorder is characterized by a lack of a stable 3D structure, and is considered to be involved in a number of important protein functions such as regulatory and signalling events. We developed a web application, the POODLE-S, which predicts the disordered region from amino acid sequences by using physicochemical features and reduced amino acid set of a position-specific scoring matrix.

Availability: POODLE-S is available from http://mbs.cbrc.jp/poodle/poodle-s.html and can be used by both academic and commercial users.

Contact: poodle@cbrc.jp

1 INTRODUCTION

Protein disorder is a widespread phenomenon, in which there is a lack of a stable 3D structure and a high degree of flexibility in the polypeptide chain. This phenomenon is considered to provide essential biological functions because dynamic conformation allows proteins to interact with multiple targets (Dunker et al., 2002). As the primary structure of the disordered regions is different from that of folded regions (Garner et al., 1998), the development of prediction methods based on amino acid sequence analysis has been encouraged (Jones and Ward, 2003; Li et al., 1999; Linding et al., 2003; Obradovic et al., 2003). We focused our attention on two facts. First, amino acid composition has different propensities in the N-term, C-term and internal regions (Shimizu et al., 2005). Second, general physicochemical properties, rather than specific amino acids, are the key factors that contribute to the development of protein disorder (Weathers et al., 2004). Then, we investigated if/how different physicochemical properties are required to characterize disorder in different regions (Shimizu et al., 2005). Our application, POODLE-S, defines a suitable length and position for the N-term and C-term regions for predicting disorder, and provides specific predictions on the basis of these regions by selecting physicochemical features, which are discriminative factors for each region.

2 OUTLINE OF METHODS

We used a $\chi^2$-test to define seven regions on the basis of positions from the N-terminal, so that each data item had similar amino acid composition. The POODLE-S application consists of seven predictors, which use support vector machines. Each predictor is prepared for each region, and selects its own features as follows.

1. The predictor selects specifically discriminative physicochemical features for a region from 10 different physicochemical properties (hydrophilic, hydrophobic, charged, positive, negative, aromatic, aliphatic, tiny, small and polar). Also, amino acids, which do not have any selected physicochemical properties, are selected as features.

2. A position-specific scoring matrix (PSSM) of target sequences via PSI-BLAST is obtained. The PSSMs are divided into sliding windows of size $m$. Each window is a matrix $E_{ij}(i = 1, \ldots, m, j = 1, \ldots, 20)$ (where $j$ represents each of the 20 amino acids).

3. Each feature is calculated as $F_{ic} = \sum_{j \in c} E_{ij}$ ($i = 1, \ldots, m, j \in c$ means that $j$ has the characteristic $c$).

3 PERFORMANCE

We used the dataset of the latest Critical Assessment of Techniques for Protein Structure Prediction (CASP7, http://predictioncenter.org/casp7/Casp7.html) to assess how well the POODLE-S performs. First, POODLE-S was trained on high-resolution single chained X-ray crystal structural data (Shimizu et al., 2005) and the DisProt database (Vucetic et al., 2005). All the data was obtained before the CASP7 prediction season. Therefore, at the time it was trained, the POODLE-S contained no information about CASP7 targets sequences. We used sensitivity $[tp/(tp+fn)]$, specificity $[tn/(tn+fp)]$, selectivity$[tp/(tp+fp)]$, and Matthews’ $\text{cm}$. 


2CASP7 provided 100 valid targets during the prediction season. We evaluated results using 89 targets whose structures are available from Protein Data Bank.
Table 1. Comparison of sensitivity (SEN), specificity (SPC) selectivity (SEL) and Matthews correlation coefficient (MCC) for POODLE-S and three successful groups

<table>
<thead>
<tr>
<th></th>
<th>SEN</th>
<th>SPC</th>
<th>SEL</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>POODLE-S</td>
<td>0.445</td>
<td>0.961</td>
<td>0.341</td>
<td>0.358</td>
</tr>
<tr>
<td>DISOPRED</td>
<td>0.513</td>
<td>0.949</td>
<td>0.313</td>
<td>0.366</td>
</tr>
<tr>
<td>Fais</td>
<td>0.673</td>
<td>0.924</td>
<td>0.286</td>
<td>0.402</td>
</tr>
<tr>
<td>ISTZORAN</td>
<td>0.802</td>
<td>0.833</td>
<td>0.180</td>
<td>0.328</td>
</tr>
</tbody>
</table>

Table 2. Comparison of MCC for POODLE-S and three successful groups, for seven regions

<table>
<thead>
<tr>
<th></th>
<th>NR1</th>
<th>NR2</th>
<th>NR3</th>
<th>IR</th>
<th>CR3</th>
<th>CR2</th>
<th>CR1</th>
</tr>
</thead>
<tbody>
<tr>
<td>POODLE-S</td>
<td>0.395</td>
<td>0.469</td>
<td>0.370</td>
<td>0.195</td>
<td>0.155</td>
<td>0.334</td>
<td>0.446</td>
</tr>
<tr>
<td>DISOPRED</td>
<td>0.414</td>
<td>0.382</td>
<td>0.275</td>
<td>0.151</td>
<td>0.166</td>
<td>0.382</td>
<td>0.483</td>
</tr>
<tr>
<td>Fais</td>
<td>0.362</td>
<td>0.415</td>
<td>0.323</td>
<td>0.258</td>
<td>0.262</td>
<td>0.382</td>
<td>0.397</td>
</tr>
<tr>
<td>ISTZORAN</td>
<td>0.359</td>
<td>0.440</td>
<td>0.265</td>
<td>0.195</td>
<td>0.299</td>
<td>0.368</td>
<td>0.372</td>
</tr>
</tbody>
</table>

POODLE-S defined seven regions: NR1: 0–9, NR2: 5–14, NR3: 10–24, IR: 20–*35, CR3: *39–*15, CR2: *19–*5 and CR1: *9–*0. Position *x is x residues away from the C-terminal. Predictions of overlapped regions are calculated by averaging these two predictions.

The correlation coefficient (MCC) for assessment. This coefficient balances sensitivity and specificity, and is calculated as follows.

\[
\text{MCC} = \frac{(tn \times tp) - (fn \times fp)}{\sqrt{(tp + fp) \times (tn + fn) \times (tp + fn) \times (tn + fp)}}
\]

(tp: true positive, tn: true negative, fp: false positive and fn: false negative).

Table 1 shows the results of the assessment of POODLE-S based on the four different scores in comparison with three other groups successfully participating in CASP7. The predictions of ‘DISOPRED’ (Ward et al., 2004), ‘ISTZORAN’ (Li et al., 1999; Obradovic et al., 2003) and ‘fais’ were downloaded from the CASP7 website. ‘DISOPRED’ is a fully automatic server group, while both ‘ISTZORAN’ and ‘fais’ registered as human expert groups, which can use any combination of computational and human methods. The data indicate that our method is of comparable accuracy (MCC) with the other three top groups. It is characterized by an average lower sensitivity (SEN), which is, however, compensated by a higher specificity (SPC) and selectivity (SEL).

The results of POODLE-S indicate that it performs better on regions NR2 and NR3. Region-specific feature selection appears to be an effective way of predicting protein disorder.

4 THE POODLE-S SERVER

The web server takes a single amino acid sequences as an input. Also, users are required to input an accessible e-mail address where the result of the prediction is sent. The POODLE-S provides both text output and graphical output. The text output style is based on the CASP format. Data in this format are inserted between the MODEL and the END records. Each line consists of a residue code a two-state prediction code and a confidence score. The symbols for the two-state order/disorder prediction are ‘O’ for order and ‘D’ for disorder. The last column should indicate the probability of a residue being in the disordered region. This value is between 0.0 and 1.0. The graphical output, in the form of an interactive line graph, is available from a URL, which is included in the e-mail and is accessible for 2 weeks after a submission. The user can display a position from the N-terminal, an amino acid code and, a probability score by pointing the cursor on the line graph (Fig. 1).

ACKNOWLEDGEMENTS

We would like to thank Yoichi Muraoka from Waseda University and Satoru Kanai from PharmaDesign, Inc. for helpful discussions. We also thank an anonymous reviewer for his/her helpful comments, which improved the manuscript.

Conflict of Interest: none declared.

REFERENCES


