PBOND: Web server for the prediction of proline and non-proline cis/trans isomerization

Konstantinos P Exarchos\textsuperscript{1,2,3}, Themis P Exarchos\textsuperscript{1,2,3}, Costas Papaloukas\textsuperscript{1,4}, Anastassios N Troganis\textsuperscript{4}, Dimitrios I Fotiadis\textsuperscript{1,2,*}

\textsuperscript{1}Unit of Medical Technology and Intelligent Information Systems, Dept of Computer Science, University of Ioannina, Ioannina, Greece
\textsuperscript{2}Institute of Biomedical Technology, CERETETH, Larissa, Greece
\textsuperscript{3}Dept of Medical Physics, Medical School, University of Ioannina, Ioannina, Greece
\textsuperscript{4}Dept of Biological Applications and Technology, University of Ioannina, Ioannina, Greece

*Corresponding author: Dimitrios I. Fotiadis, Address: Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, P.O. Box 1186, GR 45110 Ioannina, Greece. Fax: +30 26510 98889

\textit{E-mail address: fotiadis@cs.uoi.gr} (D.I. Fotiadis)

\textit{Keywords:} Peptide bond, cis/trans isomerization, support vector machines
Abstract

PBOND is a web server which predicts the conformation of the peptide bond between any two amino acids. PBOND classifies the peptide bonds in one out of four classes, namely, cis imide (cis-Pro), cis amide (cis-nonPro), trans imide (trans-Pro) and trans amide (trans-nonPro). Moreover, for every prediction a reliability index is computed. The underlying structure of the server consists of three stages: i) feature extraction, ii) feature selection and iii) peptide bond classification. PBOND can handle both single sequences as well as multiple sequences for batch processing. The predictions can either be directly downloaded from the web site or returned via email. The PBOND web server is freely available on [http://195.251.198.21/pbond.html](http://195.251.198.21/pbond.html)

Introduction

The peptide bond linking adjacent amino acids in protein structures can adopt either the cis or the trans conformation. The cis conformation occurs rarely in polypeptides because of the higher intrinsic energy compared to the trans conformation. Despite their infrequent occurrence, cis peptide bonds are very important in a variety of biological processes, such as protein folding, regulation, cell signaling and splicing of protein molecules (1). Recent studies have indicated that prolyl cis/trans isomerization can act as a molecular timer, to help control the cellular process, making it a new target for therapeutic interventions (2). Furthermore, cis peptide bonds, especially the ones between non-proline residues, are located near the active sites of proteins, or have roles in the function of the protein molecules (1,3).
In order to predict the proline isomerization, Frömmel et al. (4) extracted patterns, based on physicochemical properties. Wang et al. (5) trained a Support Vector Machine (SVM) using only the primary sequence as input in order to discriminate between the two conformations of proline peptide bonds. COPS algorithm (6) aimed to predict the peptide bond formation between any two amino acids employing an extension of the Chou-Fasman parameters. Song et al. (7) predicted the isomerization of proline peptide bonds using multiple sequence alignment profiles and secondary structure as input.

Most of the aforementioned studies focus only on the proline residues, ignoring the rare but highly important non-proline cis peptide bonds. Here, we make a further distinction of the peptide bonds, into four classes, namely cis-Pro, cis-nonPro, trans-Pro and trans-nonPro. Hence, PBOND not only predicts the peptide bond conformation between any two amino acids, but also designates potential cis-nonPro formations. Furthermore, a reliability index is computed, which represents the confidence assigned to each prediction. A majority voting scheme is also available, which provides consensus prediction of 10 SVM classifiers. PBOND has been developed using 3050 high quality protein sequences, i.e. resolution < 2.0Å, R-factor < 0.25 and sequence identity < 25%.

**Materials and Methods**

The PBOND web server graphical interface, as it is shown in Figure 1, consists of six fields. The numbers, placed next to every field in the figure, follow the same notation as below.

**Figure 1**
1. The user may choose to process either a single sequence or upload multiple sequences for batch processing; there is no upper limit to the number of sequences submitted for batch processing.

2. Amino acid sequences can be provided in FASTA format either by pasting them in the text box or uploading them within a text file. Each input sequence must have a maximum length of 1000 residues.

3. After the sequence or the sequences are uploaded, several features are extracted. More specifically, multiple sequence alignment profiles, in the form of position specific scoring matrices (PSSMs) are obtained after running PSI-BLAST (8) against one of the provided protein databases; the choice of the database highly affects the computational time, whereas, only slight perturbations are expected in terms of performance. Next, the predicted secondary structure of every residue in the query sequence is computed using PSIPRED (9); real valued predictions of solvent accessibility are obtained from RVP-net (10); six widely used physicochemical properties are also employed for every residue (volume, hydrophobicity, polarity, charge, aromatic and aliphatic character). All the above features are extracted using a sliding window with size w=11 (7,11), centered at each residue, whose peptide bond with the preceding amino acid we are trying to predict; outside this range, the influence of the surrounding residues towards the peptide bonds formation decreases. The resulting feature vector consists of 331 attributes.

4. Next, the user may choose either to employ the whole feature vector for the prediction or an optimal reduced set of features (12) identified in (11).
5. The user may choose to invoke either a single SVM model or 10 SVM models, each
one trained with a different dataset. In the latter, each model independently assigns a
label (cis-Pro, cis-nonPro, trans-Pro, trans-nonPro) to every residue in the query
sequence and then a linear time majority voting algorithm calculates the consensus of
the 10 predictions.

6. If a valid email address is supplied, the results are submitted in a compressed file;
otherwise, if the email field is left blank, the prediction results can be downloaded
directly from the web page, where they will be available for 10 days.

The output of the PBOND server consists of a compressed file which contains either a
single text file, or multiple text files, in case of batch processing. These files contain plain
text with the predictions for every sequence uploaded, along with a reliability index for
every prediction. The first and last five peptide bonds of every sequence are labeled as
“n/a” since there are not enough residues in the sliding window to make a prediction. It
should be noted that multiple simultaneous requests can be handled efficiently by the
PBOND server.

**Results and Discussion**

Due to the scarcity of \textit{cis} peptide bonds (both cis-Pro and cis-nonPro), a severe class
imbalance problem emerges, posing a tradeoff between the identification of as many
potential \textit{cis} formations and certain false positive predictions. However, the biological
significance of \textit{cis} formations outweighs possible overpredictions. Hence, special
attention was given during the training and evaluation of the PBOND server so that
important cis formations aren’t neglected. For this purpose, the predictive models of PBOND server have been trained using fully balanced datasets, in which all four classes are equally represented. The evaluation of PBOND has been performed on fully balanced disjoint data segments coming from the initial unbalanced dataset (13,14).

Table 1 presents the performance achieved using the initial feature vector, with and without performing majority voting. Sensitivity and Positive Predictive Value (PPV) are also provided for the two general classes (cis/trans). The performance achieved using the initial input vector is in general quite poor, even though voting slightly improves the results.

**Table 1**

In Table 2 the performance achieved using the optimal reduced set of features is shown, with and without the employment of the majority voting algorithm.

**Table 2**

It is clear that the feature selection improves to a certain extent the classification outcome; a further increment in the results is achieved using the consensus prediction of the 10 models. Furthermore, the reliability index associated with every prediction can be used for post processing the prediction results.

Based on the physicochemical properties of the ±6 surrounding amino acids, Frömmel et al. aimed to predict the peptide bond conformation of proline residues. They extracted 6 patterns which correctly assigned 73% of cis prolines. Although the reported results are promising, such refined dataset (242 proline bonds) diminishes the credibility of the proposed method. The proposed rules were later tested on a larger dataset, yielding inferior results. Wang et al., as well, focused only on the proline residues and employed
single sequence information, coded in binary form in order to predict the conformation of the peptide bond. The prediction accuracy achieved by the method is 70% and 77% when evaluated with independent datasets and the jackknife test, respectively. Song et al. provided as input to an SVM multiple sequence alignment profiles coupled with secondary structure information in order to predict the proline cis/trans isomerization. The overall reported accuracy is 71% after performing 5-fold cross validation. Only Pahlke et al. aimed to predict the peptide bond conformation between any two amino acids, using the secondary structure of amino acid triplets; however the reported results (overall accuracy 66%) are quite unsatisfactory. This could be attributed to the refined length of the sliding window, as well as to the small number of employed features.

A detailed comparison of the available prediction methods in the literature and PBOND is presented in Table 3. Both qualitative and quantitative measures are provided.

Table 3

The performance PBOND compares well with previously published studies, albeit validated on different datasets and using different evaluation methods. Moreover, PBOND is able to identify the scarce but highly important non-proline cis peptide bonds.

Authors’ contributions

KPE designed the study, implemented the server and prepared the manuscript. TPE and CP provided valuable comments and suggestions throughout the study and helped in the manuscript preparation. ANT and DIF supervised the study and provided substantial advice and guidance. All authors have read and approved the final manuscript.
References


Figure Legends

Figure 1: The PBOND web server graphical interface.
## Tables

Table 1. Performance measures using the whole feature vector, with and without performing majority voting.

<table>
<thead>
<tr>
<th></th>
<th>NO VOTING</th>
<th></th>
<th>VOTING</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Se (%)</td>
<td>PPV (%)</td>
<td>Se (%)</td>
<td>PPV (%)</td>
</tr>
<tr>
<td>cis-Pro</td>
<td>62.30</td>
<td>61.95</td>
<td>71.18</td>
<td>67.20</td>
</tr>
<tr>
<td>cis-nonPro</td>
<td>61.05</td>
<td>60.40</td>
<td>64.41</td>
<td>61.79</td>
</tr>
<tr>
<td>cis</td>
<td>61.68</td>
<td>61.18</td>
<td>67.78</td>
<td>64.45</td>
</tr>
<tr>
<td>trans-Pro</td>
<td>61.70</td>
<td>62.06</td>
<td>65.25</td>
<td>69.37</td>
</tr>
<tr>
<td>trans-nonPro</td>
<td>59.90</td>
<td>60.58</td>
<td>60.17</td>
<td>62.83</td>
</tr>
<tr>
<td>trans</td>
<td>60.80</td>
<td>61.32</td>
<td>62.71</td>
<td>66.10</td>
</tr>
<tr>
<td><strong>Overall accuracy</strong></td>
<td><strong>61.24</strong></td>
<td></td>
<td><strong>65.39</strong></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Results obtained using the optimal reduced set of features. Performance measures are shown with and without performing majority voting.

<table>
<thead>
<tr>
<th></th>
<th>NO VOTING</th>
<th>VOTING</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Se (%)</td>
<td>PPV (%)</td>
</tr>
<tr>
<td>cis-Pro</td>
<td>71.55</td>
<td>69.46</td>
</tr>
<tr>
<td>cis-nonPro</td>
<td>77.40</td>
<td>68.08</td>
</tr>
<tr>
<td>cis</td>
<td>74.45</td>
<td>68.77</td>
</tr>
<tr>
<td>trans-Pro</td>
<td>67.75</td>
<td>70.71</td>
</tr>
<tr>
<td>trans-nonPro</td>
<td>64.65</td>
<td>73.92</td>
</tr>
<tr>
<td>trans</td>
<td>66.20</td>
<td>72.32</td>
</tr>
<tr>
<td>Overall accuracy</td>
<td><strong>70.23</strong></td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Comparison of PBOND with available peptide bond conformation prediction methods.

<table>
<thead>
<tr>
<th>Target</th>
<th>Features</th>
<th>Se (%)</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frömmel et al.</td>
<td>Proline</td>
<td>73</td>
<td>86</td>
</tr>
<tr>
<td>Wang et al.</td>
<td>Proline</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>Pahlke et al.</td>
<td>Any amino acid</td>
<td>35</td>
<td>66</td>
</tr>
<tr>
<td>Song et al.</td>
<td>Proline</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>Any amino acid</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>PBOND</td>
<td>Proline</td>
<td>74</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>Non-Proline</td>
<td>76</td>
<td></td>
</tr>
</tbody>
</table>

*aPhCh: Physicochemical properties.
*bSS: Secondary structure.
*cASA: Accessible surface area.