Research Article

Development of a Machine Learning Method to Predict Membrane Protein-Ligand Binding Residues Using Basic Sequence Information

M. Xavier Suresh, 1 M. Michael Gromiha, 2 and Makiko Suwa 3

1Department of Bioinformatics, Sathyabama University, Chennai 600119, India
2Department of Biotechnology, IIT Madras, Chennai 600032, India
3Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan

Correspondence should be addressed to M. Xavier Suresh; xaviersuresh@gmail.com

Received 28 June 2014; Revised 7 January 2015; Accepted 8 January 2015

Academic Editor: Paul Harrison

Copyright © 2015 M. Xavier Suresh et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Locating ligand binding sites and finding the functionally important residues from protein sequences as well as structures became one of the challenges in understanding their function. Hence a Naïve Bayes classifier has been trained to predict whether a given amino acid residue in membrane protein sequence is a ligand binding residue or not using only sequence based information. The input to the classifier consists of the features of the target residue and two sequence neighbors on each side of the target residue. The classifier is trained and evaluated on a nonredundant set of 42 sequences (chains with at least one transmembrane domain) from 31 alpha-helical membrane proteins. The classifier achieves an overall accuracy of 70.7% with 72.5% specificity and 61.1% sensitivity in identifying ligand binding residues from sequence. The classifier performs better when the sequence is encoded by psi-blast generated PSSM profiles. Assessment of the predictions in the context of three-dimensional structures of proteins reveals the effectiveness of this method in identifying ligand binding residues from sequence information. In 83.3% (35 out of 42) of the proteins, the classifier identifies the ligand binding sites by correctly recognizing more than half of the binding residues. This will be useful to protein engineers in exploiting potential residues for functional assessment.

1. Introduction

Membrane proteins are an important class of molecules which play key roles in various biologically important functions such as the maintenance of ionic and proton balance, transport of substrates, ions, energy, and information across the membrane, light harvesting, photosynthesis, and other biological processes [1]. Membrane proteins are classified mainly into two types: (i) formed by bundles of apolar transmembrane α-helices (TMH) and (ii) β-barrels (TMS). It has been estimated that nearly 45% of the drugs on the market target membrane protein receptors [2]. Advancement of high throughput technologies enable the whole genome sequencing of a number of organisms. It has been estimated that, in many genomes, TM proteins comprise 20–35% of all proteins [3, 4] and hence significant progress has been made in recent years in the determination of the structures of membrane proteins. Attempts have been made to determine the complete structures or domains of membrane proteins by crystallographic and solution or solid-state NMR spectroscopy methods [5, 6]. This in turn increases the entries significantly in various structural databases.

The vast majority of signal transduction events begin with the interactions of extracellular signaling molecules (ligands) to their respective membrane-bound receptors. However, identifying functional residues in proteins is a complex issue, even when atomic detailed structures are available [7]. Various approaches are evolved to study the functional residues [8, 9]. Zhang and Grigorov, 2006, studied a hierarchically organized structural relationship among
protein binding sites using similarity networks [10]. Further
the prediction of function from sequence and structural data
has been extensively reviewed by Watson et al. [11]. These
studies showed that almost all the functions of membrane
proteins are mediated by interactions, which have a pivotal
role in biological processes essential to life, and hence under-
standing protein-ligand interactions is of prime importance.
Traditionally protein-ligand interactions are studied through
laboratory experiments, which are often time consuming
and costly [12, 13]. Accordingly computational methods have
 evolved and become increasingly dominant in understanding
protein-ligand interactions.

Protein-ligand interactions have been extensively studied
in recent years for various reasons [14–18] such as carbo-
hydrate recognition, drug interaction, and DNA binding.
Moreover, the prediction of ligand binding sites is an essential
part of the drug discovery process. Knowing the location of
binding sites greatly facilitates the search for hits, the lead
optimization process, the design of site-directed mutagenesis
experiments, and the hunt for structural features that influ-
ence the selectivity of binding in order to minimize the drug's
adverse effects. Several reports throw light on the prediction
as well as design of ligands and ligand binding sites [19–21]
using amino acid residue features and various algorithms of
machine learning [22–28]. Recently Xie and Hwang, 2015,
reviewed the underlying concepts of the methods used by
various tools for predicting protein-ligand binding sites [29].
Prediction of protein functional residues using sequence
conservation and multiple sequence alignments have also
been reported [30–32]. However, our understanding about
the interaction of ligands with membrane proteins is very
limited when compared to the other class of proteins known
as globular proteins.

The recent explosion in the availability of complete
genome sequences has led to the cataloging of tens of
thousands of new proteins and putative proteins. Previous
research focused mainly on prediction of membrane proteins
and their types [33–36]. Due to the absence of intricacies
of structural information the problem of ligand binding
prediction in membrane proteins is ignored for a long time;
however, the growth of the databases and construction of
well-defined dataset paved ways to this study. Hence, in
this work, we started from analyzing a set of nonredundant
membrane protein-ligand complexes and derived several
important sequence descriptors and trained a Naive Bayes
classifier. Bayesian classifiers are probabilistic models, based
on Bayesian theorem, robust to real data noise and missing
values [37]. The Naive Bayes classifier is one of the most effec-
tive and efficient classification algorithms in the literature [38]
showing good performance. With the help of the machine
learning technique we tried to predict the ligand binding
residues in membrane proteins from sequence information
alone.

2. Materials and Methods

2.1. Data Sets. A data set of ligand-binding membrane
proteins was extracted from structures of known membrane
protein-ligand complexes in the Protein Data Bank [39].
The dataset was culled using the list of membrane proteins
obtained from PDBTM [40], TMPDB [41], MPDB [42], and
a large collection of membrane protein structures [43]. The
resulting dataset consists of 31 membrane proteins from
which 42 sequences (nonhomologous chains are taken into
account) were considered in the present study with mutual
sequence identity ≤ 30% using BLASTCLUST program from
NCBI and each protein has at least 50 amino acid residues. All
the structures have resolution better than 3.0 Å and R factor
less than 0.3.

2.2. Ligand Information. Several reports in the literature used
all nonprotein and nonwater molecules as ligands [44]. In this
study, ligand is considered as a molecule that binds with the
proteins that have structural and/or functional role and will
be present within a cut-off distance of 4.5 Å from any of the
protein atoms.

2.3. Definition of Ligand Binding Residues. Any of the atoms
of the ligand is in contact with the any of the atoms of
a particular residue, which is said to be in binding if
the distance between them is lower than the cut-off value
4.5 Å. This definition has also been used in our previous
studies. The 42 proteins sequences (chains with at least one
transmembrane domain from 31 membrane proteins) in the
dataset consist of 10657 residues in total and 1431 of them
(13.43%) are identified as ligand binding residues.

2.4. Description of Naive Bayes Classifier. We used the Naive
Bayes implementation in the Weka package from the Uni-
versity of Waikato, New Zealand [45, 46], for predicting the
ligand binding residues in membrane proteins. For each input
target residue, the classifier produces a Boolean output (with
1 denoting a binding residue and 0 denoting a nonbinding
residue). The Naive Bayes classifier assumes independence
of the attributes given the class. For an input $X = x_1, x_2, \ldots, x_n$,
a Naive Bayes classifier assigns it a class label $c$ by optimizing
the posterior:

$$ c = \arg \max_c P (c \mid X = x_1, x_2, \ldots, x_n) $$

$$ = \arg \max_c P (c) \prod_{i=1}^{n} P (x_i \mid c). \tag{1} $$

In the case of two-class classification ($c \in \{0, 1\}$), this is
equivalent to determining $c$ by comparing the ratio likelihood
with a parameter $\theta$ as in

$$ \frac{P (c = 1 \mid X = x_1, x_2, \ldots, x_n)}{P (c = 0 \mid X = x_1, x_2, \ldots, x_n)} = \frac{\prod_{i=1}^{n} P (x_i \mid c = 1)}{\prod_{i=1}^{n} P (x_i \mid c = 0)} > \theta. \tag{2} $$

$c$ is predicted to be 1 if the ratio likelihood is greater than $\theta$,
and 0 otherwise. $\theta$ takes the value of 1. When a target residue
and its neighbors were encoded using numeric features such
as binding propensity and hydrophobicity, the numerical values were normalized using the normalization filter of Weka. We used leave-one-protein-out cross-validation to validate the classifier. In each round of experiment, all proteins except one were used as the training set and the remaining protein was used to test the classifier.

2.5. Naïve Bayes Classifier Using Sequence Based Parameters as Input. The input to the Naïve Bayes classifier contains the identities of $2n + 1$ residues in the form of $X = (x_{t-n}, x_{t-n+1}, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_{t+n-1}, x_{t+n})$, where $x_t$ is the property of target residue and $x_{t-n}, x_{t-n+1}, \ldots, x_{t-1}$ and $x_{t+1}, x_{t+n-1}, x_{t+n}$ are the identities of $n$ residues on each side of the target residue. Different values of $n$ from 1 to 7 were tried and the best performance was obtained when $n = 2$ (corresponding to a window size of 5). A training example is an ordered pair $(X, c)$, where $c \in \{0, 1\}$. 1 indicates that the target residue (the residue in the center of the input window) is a binding residue and 0 indicates that target residue is not a binding residue. For a test example $X$, the classifier outputs 1 (i.e., $X$ is predicted to be a binding residue) or 0 (i.e., $X$ is predicted to be a nonbinding residue) as the class label of $X$.

2.6. Naïve Bayes Classifier Using PSSM Profiles as Inputs. In the present study, a reference database with known nonredundant membrane protein sequences constructed separately was used for the purpose of generating PSSM profiles. We set parameters of PSI-BLAST [47] using BLOSUM62 substitution matrix, three iteration runs, and exception value 0.001. The other parameters are set using default values. The PSI-BLAST program by querying each protein chain against the nonredundant database is used to generate PSSM profiles which are in the form of $20N$ matrix, where $N$ is the total number of amino acid residues in the query protein sequence. Let the residue $i$ be represented by $a_i = (a_{i1}, \ldots, a_{i20})$ where $1 \leq i \leq N$. Each query residue is represented by a vector of 20 attributes. The input pattern to the Naïve Bayes classifier using the PSSM profile features for the residue $i$ is $x_i = (a_{i1}, \ldots, a_{i1}, \ldots, a_{ik})$ where $k$ is the number of neighborhood residues on either side. We construct a matrix with window size $s = 2k + 1$ centered on the target residue $i$. The used profile $x_i$ is the form of a $20 \times s$ matrix. These profiles are normalized into the range (0, 1) using the normalization option of Weka. Another set of attributes was also generated in such a way that it utilizes the values of BLOSUM62 matrix as features.

2.7. Performance Measures. We utilized the following parameters to evaluate the performance of our prediction method because no single performance measure provides a complete picture of performance of the classifier: accuracy, correlation coefficient (MCC), specificity, and sensitivity. These measures are defined as

\[
\text{Accuracy} (\%) = \frac{TP + TN}{N} \times 100;
\]

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}},
\]

\[
\text{Specificity} (\%) = \frac{TN}{TN + FP} \times 100;
\]

\[
\text{Sensitivity} (\%) = \frac{TP}{TP + FN} \times 100,
\]

where $TP$ is the number of true positives (residues predicted to be binding residues that are in fact binding residues); $FP$ is the number of false positives (residues predicted to be binding residues that are in fact not binding residues); $TN$ is the number of true negatives (residues predicted to be nonligand binding residues that are in fact not ligand binding residues); $FN$ is the number of false negatives (residues predicted to be nonligand binding residues that are in fact ligand binding residues); $N$ is the total number of residues ($TP + TN + FP + FN$).

3. Results and Discussion

In this work, we trained a Naïve Bayes classifier to predict whether a given amino acid residue in a membrane protein sequence is ligand binding or not based on its sequence information. The Naïve Bayes classifier algorithm as implemented in Weka, a machine learning package, is adopted. The Naïve Bayes classifier is adopted for several reasons. The prime advantage of the Bayesian classifiers is that they are probabilistic models, based on Bayesian theorem, robust to real data noise and missing values [37]. The Naïve Bayes classifier assumes independence of the attributes used in classification but it has been tested on several artificial and real datasets, showing good performance even when strong attribute dependence is present. It is one of the most effective and efficient classification algorithms in the literature and is simple to implement and use [38].

3.1. Prediction Results. We used a dataset of 42 nonredundant transmembrane protein sequences (chains with at least one transmembrane domain from 31 membrane proteins) to train the Naïve Bayes classifier. Three methods were used to encode the protein sequence. They are amino acid properties, BLOSUM62 and PSSM profiles based encodings. In the sequence based method 48 important amino acid properties (for more details see [48, 49]) such as hydrophobicity, polarity, molecular weight, and charge, are used as features to encode the protein sequence. It will be noted that some of the parameters are related. In addition it may be advisable to keep a larger input in order to avoid losing useful parameters. And hence we used all the parameters. The best prediction performance measures were obtained for a window size of 5 ($k = 2$), keeping the central residue as the target residue. The overall prediction accuracy obtained by this method was 64% and the other prediction measures are shown in Table 1.

BLOSUM matrices are based on observed alignments. Though BLOSUM62 is tailored for comparisons of moderately distant proteins, it has been used in detecting closer relationships between proteins since they best represent the physicochemical characteristics of the amino acid substitutions. And hence we used this as a feature set for training.
Table 1: The prediction performance of the Naïve Bayes classifier using different encodings of sequence.

<table>
<thead>
<tr>
<th>Number</th>
<th>Type of encoding</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Sequence</td>
<td>63.8%</td>
<td>60.5%</td>
<td>64.4%</td>
</tr>
<tr>
<td>(2)</td>
<td>Blosum62</td>
<td>67.9%</td>
<td>59.0%</td>
<td>69.4%</td>
</tr>
<tr>
<td>(3)</td>
<td>PSSM</td>
<td>70.7%</td>
<td>61.1%</td>
<td>72.2%</td>
</tr>
</tbody>
</table>

The prediction performance of classifier trained using BLOSUM62 elements as input was relatively better than that of the sequence based classifier. The performance measures are shown in Table 1.

Further we have incorporated the evolutionary information in the form of PSSM profiles based encoding. Position specific iterative BLAST (PSI BLAST) is a strong measure of residue conservation in a given location. When a residue is important for biological function it is conserved through cycles of PSI BLAST. The performance measures of the PSSM profiles based classifier are given in Table 1. Interestingly, the prediction accuracy (71%) is higher than the other two classifiers trained with the same dataset. The large predictive power of the evolutionary information as measured in this work may be due to the reason that residue conservation in protein families is directly related to its contribution to protein stability or function. It has been established by several researchers that the prediction of structural properties is significantly enhanced by the use of PSSM profiles compared to predictions based on unique representations of amino acid sequence and its environment. In addition the ligand binding in membrane proteins is largely influenced by their particular structural architecture.

We analyzed the predicted binding residues by the highest performance classifier to understand the reliability of the method. Interestingly, in 83.3% (35 out of 42) of the proteins, the classifier identifies the ligand binding sites by correctly recognizing more than half of the binding residues. In more than 90% of the proteins, the classifier correctly identifies at least 20% of the binding residues suggesting the possibility of using such classifiers to identify potential ligand-binding membrane proteins. The per protein prediction accuracy is given in Table 2. Moreover, those nonbinding residues predicted as binding residues will be in contact if we just increase the cutoff distance about 6–8 Å. Most of the false positive residues are either sequence neighbors or structural neighbors that can influence ligand binding.

3.2. ROC Curve. The receiver operating characteristic curve (ROC curve) is a plot of the "sensitivity" (TP/(TP + FN)) versus the "1-specificity" (FP/(TN + FP)) [50]. It shows the tradeoff between true positive rate and false positive rate when different threshold values are used for the classifier. Figure 1 shows such a plot for the predictor with sequence, BLOSUM62, and PSSM profiles based encoding obtained using Weka. It could be noted from the figure that there is slight improvement while using PSSM profiles as the input features for the classifier.

Table 2: Per protein prediction accuracy.

<table>
<thead>
<tr>
<th>Pdb ID</th>
<th>Number of residues</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ar1_b</td>
<td>252</td>
<td>50.0</td>
<td>83.2</td>
<td>79.4</td>
</tr>
<tr>
<td>1be3_c</td>
<td>196</td>
<td>80.0</td>
<td>89.6</td>
<td>89.0</td>
</tr>
<tr>
<td>1ds8_h</td>
<td>246</td>
<td>40.0</td>
<td>91.6</td>
<td>90.5</td>
</tr>
<tr>
<td>1ds8_l</td>
<td>281</td>
<td>54.8</td>
<td>65.2</td>
<td>62.4</td>
</tr>
<tr>
<td>lehk_a</td>
<td>544</td>
<td>43.5</td>
<td>71.0</td>
<td>66.3</td>
</tr>
<tr>
<td>lehk_b</td>
<td>166</td>
<td>30.0</td>
<td>77.0</td>
<td>74.0</td>
</tr>
<tr>
<td>leys_m</td>
<td>318</td>
<td>52.6</td>
<td>75.5</td>
<td>67.2</td>
</tr>
<tr>
<td>lfs8_a</td>
<td>338</td>
<td>25.5</td>
<td>55.9</td>
<td>50.9</td>
</tr>
<tr>
<td>lfs8_b</td>
<td>254</td>
<td>29.5</td>
<td>68.0</td>
<td>61.2</td>
</tr>
<tr>
<td>liwo_a</td>
<td>994</td>
<td>10.0</td>
<td>82.3</td>
<td>80.8</td>
</tr>
<tr>
<td>lj4n_a</td>
<td>249</td>
<td>50.0</td>
<td>68.2</td>
<td>67.3</td>
</tr>
<tr>
<td>lj95_a</td>
<td>98</td>
<td>60.0</td>
<td>62.9</td>
<td>62.7</td>
</tr>
<tr>
<td>lkb9_c</td>
<td>385</td>
<td>66.7</td>
<td>79.4</td>
<td>74.8</td>
</tr>
<tr>
<td>lkb9_h</td>
<td>93</td>
<td>20.0</td>
<td>73.8</td>
<td>70.7</td>
</tr>
<tr>
<td>lkb9_j</td>
<td>55</td>
<td>100</td>
<td>38.6</td>
<td>47.0</td>
</tr>
<tr>
<td>lkte_c</td>
<td>130</td>
<td>100</td>
<td>19.8</td>
<td>23.0</td>
</tr>
<tr>
<td>lkte_d</td>
<td>119</td>
<td>36.8</td>
<td>62.5</td>
<td>58.2</td>
</tr>
<tr>
<td>lklp_a</td>
<td>430</td>
<td>23.8</td>
<td>72.8</td>
<td>70.4</td>
</tr>
<tr>
<td>lkgf_b</td>
<td>289</td>
<td>47.2</td>
<td>88.4</td>
<td>76.8</td>
</tr>
<tr>
<td>lgha_a</td>
<td>56</td>
<td>79.4</td>
<td>16.7</td>
<td>57.6</td>
</tr>
<tr>
<td>lqns_a</td>
<td>301</td>
<td>0</td>
<td>79.8</td>
<td>78.4</td>
</tr>
<tr>
<td>lm0k_a</td>
<td>222</td>
<td>59.5</td>
<td>68.3</td>
<td>65.1</td>
</tr>
<tr>
<td>lm56_a</td>
<td>547</td>
<td>50.4</td>
<td>69.1</td>
<td>65.19</td>
</tr>
<tr>
<td>lm56_b</td>
<td>260</td>
<td>33.3</td>
<td>79.9</td>
<td>77.7</td>
</tr>
<tr>
<td>lnek_c</td>
<td>129</td>
<td>92.3</td>
<td>27.9</td>
<td>48.0</td>
</tr>
<tr>
<td>lnek_d</td>
<td>113</td>
<td>92.6</td>
<td>3.7</td>
<td>25.6</td>
</tr>
<tr>
<td>lqnc_a</td>
<td>292</td>
<td>27.5</td>
<td>88.5</td>
<td>71.5</td>
</tr>
<tr>
<td>loy9_a</td>
<td>1006</td>
<td>100</td>
<td>84</td>
<td>84.0</td>
</tr>
<tr>
<td>lp49_a</td>
<td>549</td>
<td>44.4</td>
<td>73.1</td>
<td>70.1</td>
</tr>
<tr>
<td>lp7b_a</td>
<td>258</td>
<td>60</td>
<td>76.7</td>
<td>76.38</td>
</tr>
<tr>
<td>lppq_g</td>
<td>75</td>
<td>14.3</td>
<td>64.9</td>
<td>54.9</td>
</tr>
<tr>
<td>lp7v_a</td>
<td>417</td>
<td>37.5</td>
<td>65.7</td>
<td>65.13</td>
</tr>
<tr>
<td>lql6_c</td>
<td>224</td>
<td>57.1</td>
<td>42.9</td>
<td>38.89</td>
</tr>
<tr>
<td>lq90_a</td>
<td>292</td>
<td>35.1</td>
<td>89.6</td>
<td>82.65</td>
</tr>
<tr>
<td>lq90_b</td>
<td>212</td>
<td>50</td>
<td>78.3</td>
<td>66.35</td>
</tr>
<tr>
<td>lqle_c</td>
<td>273</td>
<td>35.5</td>
<td>65.5</td>
<td>62.08</td>
</tr>
<tr>
<td>lrc2_b</td>
<td>231</td>
<td>40</td>
<td>63.2</td>
<td>61.67</td>
</tr>
<tr>
<td>lv54_d</td>
<td>144</td>
<td>84.2</td>
<td>42.1</td>
<td>47.86</td>
</tr>
<tr>
<td>lv54_g</td>
<td>84</td>
<td>60</td>
<td>20</td>
<td>40.0</td>
</tr>
<tr>
<td>lv54_j</td>
<td>58</td>
<td>30</td>
<td>81.8</td>
<td>72.22</td>
</tr>
<tr>
<td>lv59_b</td>
<td>138</td>
<td>42.4</td>
<td>83.2</td>
<td>73.13</td>
</tr>
<tr>
<td>lv55_d</td>
<td>168</td>
<td>61.1</td>
<td>82.9</td>
<td>80.49</td>
</tr>
</tbody>
</table>

Average 50.18 66.04 65.25

3.3. Comparison with Other Algorithms. Though several methods address the issue of protein-ligand interactions [22–27, 30–32], the method reported here is particularly for membrane proteins. Since the features derived are from the dataset of membrane protein sequences, its performance is very poor for globular proteins. However, for comparison of performance, few other algorithms implemented in Weka,
Figure 1: Showing ROC plots of classifiers trained with sequence (thin), BLOSUM62 (grey), and PSSM (dark) based encoding of the amino acid sequence of membrane proteins.

Table 3: The prediction performance of different machine learning algorithms.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy %</th>
<th>Sensitivity %</th>
<th>Specificity %</th>
<th>Net prediction %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>70.9</td>
<td>59.5</td>
<td>72.2</td>
<td>65.9</td>
</tr>
<tr>
<td>SMO</td>
<td>86.5</td>
<td>0</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>RBF network</td>
<td>86.5</td>
<td>0</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>Multilayer perceptron</td>
<td>84.6</td>
<td>29.6</td>
<td>93.1</td>
<td>61.4</td>
</tr>
<tr>
<td>IBk</td>
<td>83.3</td>
<td>31.3</td>
<td>91.4</td>
<td>61.4</td>
</tr>
<tr>
<td>ADTree</td>
<td>86.6</td>
<td>4.5</td>
<td>99.3</td>
<td>51.9</td>
</tr>
<tr>
<td>J48</td>
<td>87.4</td>
<td>28.1</td>
<td>93.5</td>
<td>60.8</td>
</tr>
</tbody>
</table>

3.4. WEB Based Tool. With the optimized parameters during cross-validation the current PSSM profiles based Naïve Bayes predictor has also been implemented as a web based tool which will be freely accessed from following url: http://tmbeta-genome.cbrc.jp/tm-lig/tm-lig.html. The only input to this predictor is the membrane protein sequence. The web server will automatically generate PSSMs of the given sequence against a reference data and use them as the input to the Naïve Bayes classifier trained for predictions of 42 membrane proteins. It requires less than a minute. The results presented include the raw probability scores and annotation of the residues whether ligand binding or not.

3.5. Identification of Binding Residues in Cytochrome BD Oxidase. We used the trained classifier to identify the binding residues for an unknown membrane protein sequence randomly selected from swissprot, a sequence database. The protein is cytochrome bd oxidase from *E. coli*, important for anaerobic oxidation [51]; its structure has not yet been determined. The predicted results were compared with the functional information available from the literature [51–54]. Interestingly, 50% of the residues predicted to be ligand binding are involved in interaction. In addition, few of the predicted residues belong to the segments which were experimentally determined to be functionally important [52]. Since the structure of this protein has not been determined, a homology model has been built (data not shown) and comparative analysis of binding sites with the related structures revealed that nearly 80% of the predicted residues are found to be along the lining of the proposed binding sites. This indeed increases the confidence level of using this predictor prior to planning for any mutagenesis or any functional assessment related experiment with more confidence rather than a random start.

4. Conclusion

Using a well constructed dataset, a Naïve Bayes predictor is trained and tested to predict the ligand binding residues in membrane proteins from amino acid sequence. Several encodings were used to test the performance of the predictor and PSSM profiles based predictor was shown to have better prediction accuracy (71%). With the level of success achieved in this study, putative ligand-binding sites predicted by the classifiers trained using a machine learning approach should be useful for guiding experimental investigations into the role of specific residues of a protein in its interaction with ligand, for example, by localizing candidate residues for mutagenesis. This paves ways for further improvements and predictions based on sequence methods for membrane protein-ligand interactions. Currently, we are investigating other machine learning classification methods to improve the accuracy of the prediction, which warrants further exploration.

Conflict of Interests

The authors declare that they have no conflict of interests related to this paper.

Acknowledgments

The computational facilities at CBRC and the financial assistance from JSPS are gratefully acknowledged. The authors thank the anonymous reviewers for their valuable suggestions.
References


Submit your manuscripts at http://www.hindawi.com