Predicting cell adhesion receptors using protein sequence index

Zhijun Wang
Chunyan Xu
Feng Yu
Lin Tang
Jianhua He
Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai, People’s Republic of China

Abstract: Cell adhesion receptors (CARs) play important roles in signaling, regulation, membrane trafficking, immune response, and transport. For a long time, based on their functional and sequence diversity, CARs have been classified into four classes: cadherin-mediated cell adhesion receptors (CMCARs); immunoglobulin superfamily-mediated cell adhesion receptors (ISMCARs); selectin-mediated cell adhesion receptors (SMCARs); and integrin-mediated cell adhesion receptors (IMCARs). Experimental methods suitable to identify and to determine the kind of CARs are time-consuming. It is, therefore, desirable to explore new methods for predicting CARs directly from protein sequence information. This report shows the application of Protein Sequence Index (PSI) as such a method. Two fuzzy k-nearest neighbor (NN) prediction systems were developed to identify adhesion proteins (APs) and classify APs into different CARs with PSI. In the first fuzzy k-NN predicting system, 619 APs and 1211 nonadhesion proteins (NAPs) were used as a training dataset to identify the APs, and they were evaluated by an independent dataset of 477 APs and 576 NAPs. The computed prediction accuracy was 94.5% and 94.4% for the APs and NAPs respectively, using the independent dataset. In the second fuzzy k-NN predicting system, 1211 noncell adhesion receptors (NCARs), 286 CMCARs, 59 ISMCARs, 38 SMCARs, and 236 IMCARs was used as a training dataset to classify CARs into different types, and they were evaluated by an independent testing dataset of 576 NCARs, 228 CMCARs, 47 ISMCARs, 20 SMCARs, and 182 IMCARs. The predicting accuracy was 94.4%, 92.1%, 95.7%, 95.0%, and 98.9%, for NCARs, CMCARs, ISMCARs, SMCARs, and IMCARs, respectively. These findings suggest the usefulness of PSI for facilitating the identification and classification of CARs. A program, ADHEN, was constructed, which can be used to predict the CARs.

Keywords: cell adhesion receptor, protein sequence index, prediction

Introduction

With the rapid increase in protein sequence and functional annotation data from many organisms,1,2 it is essential to analyze various correlations based on large datasets, and to design more reliable analytical and predictive tools.3–9 These tools are crucial for the analysis of biological data and are increasingly used to accelerate progress in biological research.10–13

Cell adhesion receptors (CARs) play important roles in cell signaling,14 protein trafficking,15 virus killing16 and innate immune responses.17 CARs have been found in a variety of pathogenic microbes.18 Prediction of CARs is important for facilitating the study of various biological processes and searching for new vaccine candidates.17 Finding CARs can help scientists to find potential methods to deal with...
infection: for example, abrogation of CARs by either immunizing the host with adhesions or inhibiting the interaction using structural analogs of host cell receptors holds the potential to develop novel preventive strategies. Experimental methods used for characterizing CARs are time-consuming and demand large resources. Typically, CARs are often assayed in vitro either by cell or cell-substrate binding experiments, using transfected cells expressing the molecule of interest, or using monoclonal antibodies or peptides to interfere with the expression of adhesion proteins (APs). In vivo, the function of adhesion is very difficult to determine, and in fact, many functions of CARs are being determined using gene knockout mice. In general, the gene knockout mice offer novel resources for elucidating the molecular basis of CARs, but, these kinds of methods are time-consuming and expensive. It is desirable to develop a computational method for predicting CARs from the protein sequence information.

CARs are primarily characterized by their specific sequence features. Given a protein sequence, the question arises of how to identify this protein as an AP or nonadhesion protein (NAP), and if the protein is an AP, how to classify this AP into one of the different classes of CAR. One might assume that if an accurate, robust, and rapid method for predicting CARs was developed, it could significantly help biologists to reduce the experimental time involved in finding CARs.

A statistical learning method, fuzzy k-nearest neighbor (fuzzy k-NN), has been widely applied in many areas of bioinformatics. These applications include protein subcellular location prediction, and diagnosis. Because fuzzy k-NNs are designed to maximize the margins to separate two classes so that the trained model generalizes well on the datasets, it is thus of interest to explore the use of a fuzzy k-NN as a classifier to predict CARs.

In this report, we explore the application of fuzzy k-NN to develop a prediction system to identify and classify CARs. For a long time, CARs have been divided into four classes (Figure 1):30

1. Cadherin-mediated cell adhesion receptor (CMCAR). Cadherins are primarily involved in cell adhesion. Their extracellular domains contain five characteristic repeats, each comprising a sandwich of β sheets, and they mediate adhesion between cells through the most distal cadherin repeats.

2. Immunoglobulin superfamily-mediated cell adhesion receptors (ISMCAR). ISMCARs are characterized by the presence of varying numbers of Ig-related domains and they have adhesion sandwiches of two β sheets held together by hydrophobic interactions.

3. Selectin-mediated cell adhesion receptors (SMCAR). Selectins show a heterophilic interaction with their counter-receptors. They recognize specific carbohydrate groupings in the counter-receptor or ligand.

4. The final major family of CARs is the integrin-mediated cell adhesion receptors (IMCAR). Most integrins are predominantly receptors for fibronectins, laminins, and collagens, but a few also play important roles in heterotypic cell adhesion.

A novel protein sequence index (PSI) algorithm was designed and applied for training the fuzzy k-NN to identify the CARs. The prediction accuracies of fuzzy k-NN were analyzed. Our results show that the prediction model reveals a high accuracy rate using leave-one-out cross-validation testing and independent dataset testing. It is expected that our method can provide a useful additional technique for finding CARs.

Methods

Datasets

Protein sequence data of APs were retrieved from http://www.ncbi.nih.gov. Any putative or unverified APs were removed from the datasets. NAPs that function within the cell were also retrieved from http://www.ncbi.nih.gov. Four datasets were built as shown in Tables 1 and 2.
A contained 619 APs and 1211 NAPs as a training dataset for identification of APs. Dataset B contained 477 APs and 576 NAPs as a testing dataset for identification of APs (Table 1). Dataset C contained 1211 noncell adhesion receptors (NCARs), 286 CMCARs, 59 ISMCARs, 38 SMCARs, and 236 IMCARs as a training dataset for classification of CARs. Dataset D contains 576 NCARs, 228 CMCARs, 47 ISMCARs, 20 SMCARs, and 182 IMCARs as a testing dataset for classification of CARs (Table 2). These datasets can be downloaded from: http://code.google.com/p/adhen.

### Protein sequence index

The protein sequence index (PSI) was determined according to the following method.

1. Given a particular protein (protein length \( n \)), the number of amino acid residues in this protein is expressed as Equation 1:

   \[
   A_i = [p(a_1), p(a_2), \ldots, p(a_i), \ldots, p(a_{20})],
   \]

   where \( A_i \) means the amino acids composition of the protein, and \( p(a_i) \) means the total number of 20 different amino acid residues (i = 1–20) in the protein, respectively.

2. Construct a \( (20 \times (n + 1)) \) matrix \( A_2 \). The first element \( \alpha_{i,m} \) of each row represents the total number of amino acid residues of the protein from begin \( (j = 1) \) to end \( (j = n) \). If an amino acid residue \( i \) is found in the specific sequence position \( j \), the number of amino acid residues \( \alpha_{i,m} \) is reduced by 1; otherwise, the number of amino acid residues \( \alpha_{i,m} \) remains unchanged.

Matrix \( A_2 \) contains the survival values of 20 different amino acids along the protein sequence for a specific protein. Each row represents the properties of different amino acid residues \( i \) (i = 1:20), and each column indicates the amino acid survival value in the protein. \( A_2 \) is expressed as:

\[
A_2 = \begin{bmatrix}
\alpha_{1,m} & \cdots & \alpha_{1,m}-1 & \cdots & \alpha_{1,m}-2 & \cdots & \alpha_{1,m}-\kappa & \cdots & 0 \\
\alpha_{2,m} & \cdots & \alpha_{2,m}-1 & \cdots & \alpha_{2,m}-2 & \cdots & \alpha_{2,m}-\kappa & \cdots & 0 \\
\vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
\alpha_{i,m} & \cdots & \alpha_{i,m}-1 & \cdots & \alpha_{i,m}-2 & \cdots & \alpha_{i,m}-\kappa & \cdots & 0 \\
\vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
\alpha_{20,m} & \cdots & \alpha_{20,m}-1 & \cdots & \alpha_{20,m}-2 & \cdots & \alpha_{20,m}-\kappa & \cdots & 0
\end{bmatrix}, \quad (1 \leq \kappa \leq \ell)
\]

#### Table 1 Datasets for the identification of APs (protein sequence distance \( P \)-value cutoff = 0.05)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset A</td>
<td>Dataset B</td>
</tr>
<tr>
<td>AP (Training dataset)</td>
<td>AP (Testing dataset)</td>
</tr>
<tr>
<td>619</td>
<td>477</td>
</tr>
<tr>
<td>NAP (Training dataset)</td>
<td>NAP (Testing dataset)</td>
</tr>
<tr>
<td>1211</td>
<td>576</td>
</tr>
<tr>
<td>Total</td>
<td>Total</td>
</tr>
<tr>
<td>1830</td>
<td>1053</td>
</tr>
</tbody>
</table>

**Abbreviations:** AP, adhesion protein; NAP, nonadhesion protein.

Table 3 shows a calculation example of matrix \( A_2 \) for a specific peptide.

3. Fit different amino acid decay information in each row of \( A_1 \) using the following Equation 3:

\[
p^i(x) = p_1^i x^6 + p_2^i x^5 + \cdots + p_{12}^i x + p_{13}^i
\]

A program was designed to find the coefficients of a polynomial function \( p^i(x) \) of degree 6 that fits different amino acid survival values. For a different amino acid \( i \) in the protein, a coefficient array, \( \gamma = [p_1^i, p_2^i, \ldots, p_{12}^i] \), (i = 1:20) was calculated, then arrays of 20 different amino acids were assembled into array \( A_3 \), and it was expressed as Equation 5:

\[
A_3 = \begin{bmatrix}
p_1 x^5 & p_2 x^4 & \cdots & p_{13} x + p_{14} \\
p_{15} x^4 & p_{16} x^3 & \cdots & p_{26} x + p_{27} \\
\vdots & \vdots & \cdots & \vdots \\
p_{109} x & p_{110} & \cdots & p_{120}
\end{bmatrix}
\]

where \( || \) denotes vector horizontal concatenation, \( A_3 \) is a \((1 \times 140)\) vector for each protein. We included the amino acid component array \( A_1 \) of the protein into \( A_3 \) and formed an \( A_4 \) vector, which is expressed as Equation 6:

\[
A_4 = A_1 || A_3
\]

\( A_4 \) is a \((1 \times 160)\) vector, which was used as PSI in the predicting of CAR.
**Table 2** The datasets for the classification of cell adhesion receptor (protein sequence distance P-value cutoff = 0.05)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset C (Training dataset)</td>
<td>Dataset D (Testing dataset)</td>
</tr>
<tr>
<td>NCAR</td>
<td>1211</td>
</tr>
<tr>
<td>CMCAR</td>
<td>286</td>
</tr>
<tr>
<td>ISMCAR</td>
<td>59</td>
</tr>
<tr>
<td>SMCAR</td>
<td>38</td>
</tr>
<tr>
<td>IMCAR</td>
<td>236</td>
</tr>
<tr>
<td>Total</td>
<td>1830</td>
</tr>
</tbody>
</table>

**Abbreviations:** CMCAR, cadherin-mediated cell adhesion receptor; ISMCAR, immunoglobulin superfamily-mediated cell adhesion receptor; SMCAR, selectin-mediated cell adhesion receptor; IMCAR, integrin-mediated cell adhesion receptor; NCAR, noncell adhesion receptors.

**Fuzzy k-NN**

The fuzzy k-NN algorithm can predict the data point by finding its closest neighbors.24 The proposed fuzzy k-NN classifier assigns the membership values $r_c(Tr)$ of sample $Tr$ to different classes as follows in Equation 7:

$$r_c(Tr) = \frac{\sum_{i=1}^{k} r_i(Tr^c)\left(\left\|Tr - Tr^c\right\|^{2(w-1)}\right)}{\sum_{i=1}^{k} \left\|Tr - Tr^c\right\|^{2(w-1)}}, \quad (7)$$

where $c$ means different classes, and $\left\|Tr - Tr^c\right\|$ is the Euclidean distance between $Tr$ and one of its nearest neighbors $Tr^c$, $w$ is the fuzzy strength parameter to determine the weighting of the distance, and $Tr$ was categorized into the class having the highest membership value. Here, we set $k = 1$, and $w = 2$ as default values for our fuzzy k-NN classifier.

**Predictive accuracy**

Various quantitative variables were obtained to measure the effectiveness of the support vector machine (SVM) method: 1) TP, true positives, the number correctly classified; 2) FP, false positives, the number incorrectly classified; 3) TN, true negatives, the number correctly classified; 4) FN, false negatives, the number incorrectly classified. Using the variables above, a series of statistical metrics were computed to measure the effectiveness of the SVM method. To provide an indication of the overall performance of the system, we computed the predictive accuracy (PA) as:

$$PA(\%) = \frac{TP + TN}{TP + FN + TN + FP} \times 100 \quad (8)$$

**Results**

**Datasets**

Fuzzy k-NN classifiers were used for the identification and classification of CARs. To test their capability, the NCBI database was searched to allow the construction of CAR datasets. Highly homologous proteins in the datasets were removed, based on protein sequence distance analysis results. A similarity $P$-distance threshold value of 0.05 was used to ensure the maximum exclusion of homologous proteins. Four datasets of experimentally known CARs from different...
species were prepared by careful examination of literature reports. Any “controversial”, “putative”, “predicted”, or “hypothetical” data were excluded from the datasets. The negative dataset consisted of proteins representing various intracellular enzymes, for example: “dehydratease”, “kinase”, “acyl-CoA synthase”. Finally, we obtained four datasets (A, B, C, D) as shown in Tables 1 and 2. Datasets A and B were used for the identification of APs, and datasets C and D were used for the classification of CARs.

Determination of the PSI
PSIs were computed with the algorithm described in the Methods section. For each protein, a \((1 \times 160)\) vector was constructed. The vector includes the amino acid composition of each protein (initial 20 elements in the PSI) and the fitting coefficients of 20 amino acids for the protein (following 140 elements in the PSI), respectively.

Figure 2 shows the fitting results of 20 amino acids with the PSI of mannosyltransferase (GenBanK accession XP_721742) using Equation 3. The x-axis represents the protein sequence length of mannosyltransferase, the y-axis represents the amino acid survival value. The blue line represents the relationship between amino acid survival value and protein sequence length. The green line represents the fitting curves using Equation 3.

Figure 2 Fitting curves of protein sequence index (PSI) obtained from the protein sequence of mannosyltransferase (GenBanK accession No: XP_721742) using Equation 3. The x-axis represents the protein sequence length of mannosyltransferase, the y-axis represents the amino acid survival value. The blue line represents the relationship between amino acid survival value and protein sequence length. The green line represents the fitting curves using Equation 3.
No: XP_721742) obtained from Equation 2. Our results show that the coefficients of PSI can accurately fit decay efficiencies of different amino acids (Figure 2). After we calculated PSI for each protein, datasets were used to train the fuzzy k-NN classifier.

Determination of predictive accuracy

Identification and classification efficiencies of fuzzy k-NNs for CARs using PSIs were determined. The computation was performed on an ASUS machine (ASUS Computer International, Fremont, CA) with an Intel 2.6 GHz CPU and 2G RAM.

The identification results for APs and NAPs using the first fuzzy k-NN classifier are given in Table 4. When training dataset A is used in self-consistence testing using the leave-one-out cross-validation algorithm, the predictive accuracy in identification of AP and NAP is 98.5% and 99.7%, respectively. When testing dataset B is used in independent testing, the predictive accuracy in identification of AP and NAP is 94.5% and 94.4%, respectively. In identification of APs and NAPs, the fuzzy k-NN classification using PSI can achieve higher accuracy relative to fuzzy k-NN classification using amino acid composition (Table 4). We designed an artificial neural network (ANN) classification to compare with the fuzzy k-NN classification using MATLAB. The ANN classification is a two-layer feed-forward network. ANN has one hidden layer with five tan-sigmoid transfer function (TANSIG) neurons. The second layer has one log-sigmoid transfer function (LOGSIG) neuron, and the epoch parameter is set as 100, and the goal of the training is set as 0.01. In our experimental conditions, after ANN was trained with dataset A, then tested with dataset B, the fuzzy k-NN shows a better predictive accuracy relative to ANN (Table 4).

The classification results for NCARs, CMCARs, ISMCARs, SMCARs, and IMCARs using the second fuzzy k-NN classifier are given in Table 5. When training dataset C is used in self-consistence testing using the leave-one-out cross-validation algorithm, the predictive accuracy in classification of NCARs, CMCARs, ISMCARs, SMCARs, and IMCARs is 99.7%, 95.4%, 98.3%, 97.3%, and 97.4%, respectively. When testing dataset D is used in independent testing, the predictive accuracy in classification of NCARs, CMCARs, ISMCARs, SMCARs, and IMCARs is 94.4%, 92.1%, 95.7%, 95.0%, and 98.9%, respectively. In classification of NCARs, CMCARs, ISMCARs, SMCARs, and IMCARs, the fuzzy k-NN classification using PSI can achieve higher accuracy relative to fuzzy k-NN classification using amino acid composition (Table 5). In our experimental conditions, after ANN was trained with dataset C, then tested with dataset D, the fuzzy k-NN shows a better predictive accuracy relative to ANN (Table 5).

Discussion

With the rapid increase in the size of biological databanks, understanding the data has become critical.41,42 Although laboratory experiment is the most effective method for finding CARs, it is difficult and time-consuming. Therefore, computational tools have been widely used in the fields of classification and cluster analysis of biological data.43–45 There are many computational algorithms available for the classification analysis of biological data, including decision trees,46–48 discriminant analysis,49–51 and neural networks.52–54 Here, we have used the fuzzy k-NN classification in our experiment, because the fuzzy k-NN classifier is fast, easy, and efficient.26

CARs have been receiving much attention in recent years. CARs are essential in almost all aspects of cell development.55, 56 CARs localize on the cell surface and play important roles in different cells. The importance of CARs has been elucidated in many organisms, including bacteria.16,57 The cell adhesion function is highly significant in cell division, cell migration, cell differentiation and apoptosis,58 and therefore, classification of CARs is an important research topic in bioinformatics.30

In this report, we have applied fuzzy k-NNs to identify APs and to classify CARs. We determined PSIs of different

Table 4 Predictive accuracy in identification of AP (adhesion protein) and NAP (nonadhesion protein)

<table>
<thead>
<tr>
<th>Classes</th>
<th>Predictive accuracy in identification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fuzzy k-NN</td>
</tr>
<tr>
<td></td>
<td>Using PSI</td>
</tr>
<tr>
<td></td>
<td>Self consistence test (Dataset A)</td>
</tr>
<tr>
<td></td>
<td>Independent test (Dataset B)</td>
</tr>
<tr>
<td></td>
<td>Using amino acid composition</td>
</tr>
<tr>
<td></td>
<td>Independent test (Dataset B)</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
</tr>
<tr>
<td></td>
<td>Using PSI</td>
</tr>
<tr>
<td></td>
<td>Independent test (Dataset B)</td>
</tr>
<tr>
<td>AP</td>
<td>98.5% 94.5%</td>
</tr>
<tr>
<td>NAP</td>
<td>99.7% 94.4%</td>
</tr>
<tr>
<td>Overall</td>
<td>99.3% 94.5%</td>
</tr>
</tbody>
</table>

Abbreviations: ANN, artificial neural network; NN, nearest neighbor; PSI, Protein Sequence Index.
protein datasets, and PSI values were used to train fuzzy k-NNs. Our results show that the fuzzy k-NNs can identify and classify protein sequences into APs or CARs with high accuracy. These predictors could serve as new leads for further experimental characterization.

In order to improve predictive accuracy, we applied the following conditions for the selection and construction of datasets:

1. To improve the quality of datasets, we used well-annotated preferably experimentally validated data, and avoided sequences with ambiguous annotations, conflicting experimental evidence, or those that were annotated through prediction;
2. To improve the performance of the classifier, we attempted to collect as many sequences as possible to develop an accurate classifier;
3. To avoid redundancy, we removed redundant or highly similar sequences from datasets to avoid biasing the algorithm towards groups of similar sequences, with the protein \( P \)-distance similarity cutoff value set at 0.05 between different protein sequences.

The predictive accuracy of nonmembers appears to be better than that of members. The higher prediction accuracy for nonmembers probably results from the availability of a more diverse set of nonmembers than that of members, which enables a classifier to perform better statistical learning for the recognition of nonmembers. This may partly explain why the prediction accuracy for members is generally lower than that for nonmembers. In our experiments, fuzzy k-NN classification is more efficient than ANN classification in handling unbalanced datasets. Obviously, it is inappropriate to simply reduce the number of nonmembers to artificially match that of members, because this will reduce the diversity to fully represent all nonmembers. In the future, if we can combine the fuzzy k-NN classifier with other classifiers in identification and classification, the predictive accuracy should be significantly improved in the treatment of very large unbalanced datasets.

In summary, prediction of protein sequences with low similarity to specific protein function sequences is a major challenge in computational biology in the postgenomic era. Fuzzy k-NNs with PSI appear to be potentially useful tools for the identification of APs and classification of CARs. It is time for us to produce a proteome level PSI database, so that this algorithm can be applied to genomes. The fuzzy k-NN-derived classification systems with PSI developed in this work can be accessed from http://code.google.com/p/adhen.

**Acknowledgment**

This work was supported by the Natural Science Foundation of Shanghai (Grant Number: 09ZR1436700 and 08JC1422500).

**Disclosure**

The authors report no conflicts of interest in this work.

**References**


