Open Access Bioinformatics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Predicting cell adhesion receptors using protein sequence index

This article was published in the following Dove Press journal:
Open Access Bioinformatics
31 May 2011
Number of times this article has been viewed

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 killing[16] 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

Correspondence: Zhijun Wang
Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai 201204, People's Republic of China
Tel +86 21 3393 3192
Fax +86 21 3393 3021
Email wangzhijun@sinap.ac.cn

Jianhua He
Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai 201204, People's Republic of China
Tel +86 21 3393 3256
Fax +86 21 3393 3021
Email hejh@sinap.ac.cn

submit your manuscript | www.dovepress.com
Dovepress
DOI: 10.2147/OAB.S17492

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.[14,19,20] 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).[21] 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.[22] In general, the gene knockout mice offer novel resources for elucidating the molecular basis of CARs,[23] 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.[24]

CARs are primarily characterized by their specific sequence features.[14,25] 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.[26] These applications include protein subcellular location prediction,[27,28] and diagnosis.[29] 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).[31] Cadherins are primarily involved in cell adhesion.[32] 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.[31]
2. Immunoglobulin superfamily-mediated cell adhesion receptors (ISMCAR).[33] ISMCARs are characterized by the presence of varying numbers of Ig-related domains[34] and they have adhesion sandwiches of two β sheets held together by hydrophobic interactions.

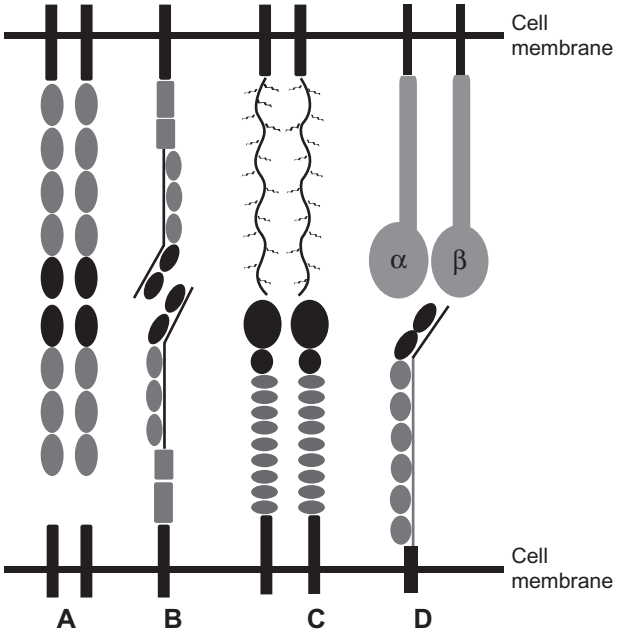

Figure 1 Different types of CARs (cell adhesion receptors): (**A**), cadherin-mediated cell adhesion receptor (CMCAR); (**B**), immunoglobulin superfamily-mediated cell adhesion receptor (ISMCAR); (**C**), selectin-mediated cell adhesion receptor (SMCAR); (**D**), integrin-mediated cell adhesion receptor (IMCAR) (for more information, see reference 30).

3. Selectin-mediated cell adhesion receptors (SMCAR).[35] Selectins show a heterophilic interaction with their counter-receptors. They recognize specific carbohydrate groupings in the counter-receptor or ligand.[36,37]
4. The final major family of CARs is the integrin-mediated cell adhesion receptors (IMCAR).[38–40] Most integrins are predominantly receptors for fibronectins, laminins, and collagens, but a few also play important roles in heterotypic cell adhesion.[30]

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. Dataset

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.

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

Dataset	Number of sequences	
	Dataset A (Training dataset)	Dataset B (Testing dataset)
AP	619	477
NAP	1211	576
Total	1830	1053

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

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:

$$A_1 = [p(a_1),\quad p(a_2),\ \ldots,\ p(a_i),\ \ldots,\quad p(a_{20})]\,, \qquad (1)$$

where A_1 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 i in the protein. Read the 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:

$$A_2 = \begin{bmatrix} a_{1,m} & \cdots & a_{1,m}-1 & \cdots & a_{1,m}-2 & \cdots & a_{1,m}-\kappa & \cdots & 0 \\ a_{2,m} & \cdots & a_{2,m}-1 & \cdots & a_{2,m}-2 & \cdots & a_{2,m}-\kappa & \cdots & 0 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ a_{i,m} & \cdots & a_{i,m}-1 & \cdots & a_{i,m}-2 & \cdots & a_{i,m}-\kappa & \cdots & 0 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ a_{20,m} & \cdots & a_{20,m}-1 & \cdots & a_{20,m}-2 & \cdots & a_{20,m}-\kappa & \cdots & 0 \end{bmatrix}, \quad (1 \le \kappa \le \qquad (2)$$

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_2 using the following Equation 3:

$$p^i(x) = p_1^i x^6 + p_2^i x^5 + \cdots + p_6^i x + p_7^i \qquad (3)$$

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,

$$\gamma = \left[p_1^i,\quad p_2^i,\quad \ldots,\quad p_7^i\right], (\mathrm{i} = 1{:}20) \qquad (4)$$

was calculated, then arrays of 20 different amino acids were assembled into array A_3, and it was expressed as Equation 5:

$$A_3 = [p_1^1,\quad p_2^2,\quad \ldots,\quad p_7^1] \| [p_1^2,\quad p_2^2,\quad \ldots,\quad p_7^2] \| \cdots \| [p_1^{20},\quad p_2^{20},\quad \ldots,\quad p_7^{20}] \qquad (5)$$

where || denotes vector horizontal concatenation, A_3 is a (1×140) 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:

$$A_4 = A_1 \| A_3 \qquad (6)$$

A_4 is a (1×160) 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)

Dataset	Number of sequences	
	Dataset C (Training dataset)	Dataset D (Testing dataset)
NCAR	1211	576
CMCAR	286	228
ISMCAR	59	47
SMCAR	38	20
IMCAR	236	182
Total	1830	1053

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.[26] The proposed fuzzy k-NN classifier assigns the membership values $r_c(Tr)$ of sample Tr to different classes as follows in Equation 7:

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

where c means different classes, and $||Tr - Tr^s||$ is the Euclidean distance between Tr and one of its nearest neighbors Tr^s, 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:

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

Table 3 Calculation results of matrix A_2 (as described in Equation 2) for a specific peptide containing 30 amino acid residues mlsifkpaphkarlpaaeidptyrrlrwqi

Amino acid residue	The result of A_2 matrix for a peptide with sequence as mlsifkpaphkarlpaaeidptyrrlrwqi
S	1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
A	4, 4, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 2, 2, 2, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
L	3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0
V	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
E	1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
P	4, 4, 4, 4, 4, 4, 4, 3, 3, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0
R	4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 1, 1, 0, 0, 0
G	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
I	3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0
N	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
K	2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
D	1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0
T	1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0
F	1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
Y	1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0
M	1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
H	1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
Q	1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0
C	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
W	1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0

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 × 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

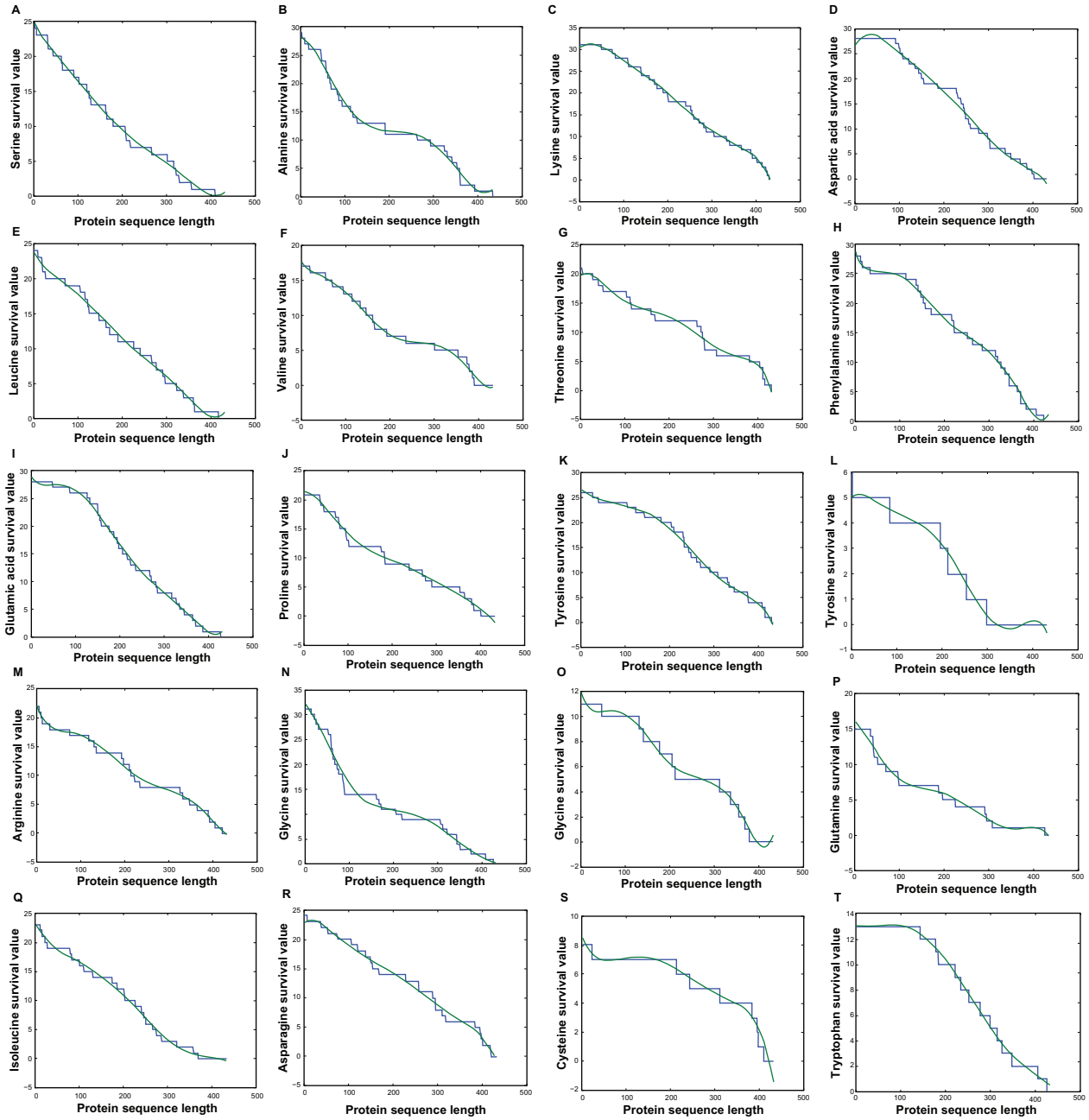

Figure 2 Fitting curves of protein sequence index (PSI) obtained from the protein sequence of mannosyltransferase (GenBanK accession No: XP_721742) using Equation 3. The x-axis represents the protein sequence length of mannosyltransferase, the y-axis represents the amino acid survival value. The blue 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)

Classes	Predictive accuracy in identification			
	Fuzzy k-NN			ANN
	Using PSI		Using amino acid composition	Using PSI
	Self consistence test (Dataset A)	Independent test (Dataset B)	Independent test (Dataset B)	Independent test (Dataset B)
AP	98.5%	94.5%	69.3%	83.4%
NAP	99.7%	94.4%	70.8%	89.9%
Overall	99.3%	94.5%	70.1%	86.9%

Abbreviations: ANN, artificial neural network; NN, nearest neighbor; PSI, Protein Sequence Index.

Table 5 Predictive accuracy in classification of CARs (cell adhesion receptors) into different types

Classes	Predictive accuracy in classification			
	Fuzzy k-NN			ANN
	Using PSI		Using amino acid composition	Using PSI
	Self-consistence test (Dataset C)	Independent test (Dataset D)	Independent test (Dataset D)	Independent test (Dataset D)
NCAR	99.7%	94.4%	73.0%	90.1%
CMCAR	95.4%	92.1%	73.2%	86.8%
ISMCAR	98.3%	95.7%	63.8%	80.6%
SMCAR	97.3%	95.0%	65.0%	85.0%
IMCAR	97.4%	98.9%	71.9%	85.7%
Overall	98.6%	94.7%	72.4%	88.1%

Abbreviations: ANN, artificial neural network; 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; 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

1. Ressom HW, Varghese RS, Zhang Z, Xuan J, Clarke R. Classification algorithms for phenotype prediction in genomics and proteomics. *Front Biosci*. 2008;13:691–708.
2. Sernova NV, Gelfand MS. Identification of replication origins in prokaryotic genomes. *Brief Bioinform*. 2008;9(5):376–391.
3. Aguilar-Ruiz JS, Moore JH, Ritchie MD. Filling the gap between biology and computer science. *BioData Min*. 2008;1(1):1.
4. Cao Y, Liu D, Tavanapong W, Wong J, Oh J, de Groen PC. Computer-aided detection of diagnostic and therapeutic operations in colonoscopy videos. *IEEE Trans Biomed Eng*. 2007;54(7):1268–1279.
5. Kim SH, Lee JM, Lee JG, et al. Computer-aided detection of colonic polyps at CT colonography using a Hessian matrix-based algorithm: preliminary study. *AJR Am J Roentgenol*. 2007;189(1):41–51.

6. Sivashankari S, Shanmughavel P. Functional annotation of hypothetical proteins – A review. *Bioinformation*. 2006;1(8):335–338.
7. Lovering RC, Camon EB, Blake JA, Diehl AD. Access to immunology through the Gene Ontology. *Immunology*. 2008;125(2):154–160.
8. Overbeek R, Bartels D, Vonstein V, Meyer F. Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chem Rev* 2007;107(8):3431–3447.
9. Thomas PD, Mi H, Lewis S. Ontology annotation: mapping genomic regions to biological function. *Curr Opin Chem Biol*. 2007;11(1): 4–11.
10. Yan Q. Bioinformatics databases and tools in virology research: an overview. *In Silico Biol.* 2008;8(2):71–85.
11. Harrill AH, Rusyn I. Systems biology and functional genomics approaches for the identification of cellular responses to drug toxicity. *Expert Opin Drug Metab Toxicol*. 2008;4(11):1379–1389.
12. Sorokin V, Severinov K, Gelfand MS. Systematic prediction of control proteins and their DNA binding sites. *Nucleic Acids Res*. 2009;37(2): 441–451.
13. Sadovskaya NS, Sutormin RA, Gelfand MS. Recognition of transmembrane segments in proteins: review and consistency-based benchmarking of internet servers. *J Bioinform Comput Biol*. 2006;4(5): 1033–1056.
14. Hauck CR. Cell adhesion receptors – signaling capacity and exploitation by bacterial pathogens. *Med Microbiol Immunol*. 2002;191(2):55–62.
15. Stoolman LM, Kaldjian E. Adhesion molecules involved in the trafficking of normal and malignant leukocytes. *Invasion Metastasis*. 1992; 12(2):101–111.
16. Klemm P, Schembri MA. Bacterial adhesins: function and structure. *Int J Med Microbiol*. 2000;290(1):27–35.
17. Wizemann TM, Adamou JE, Langermann S. Adhesins as targets for vaccine development. *Emerg Infect Dis*. 1999;5(3):395–403.
18. Buckley CD, Rainger GE, Bradfield PF, Nasg GB, Simmons DL. Cell adhesion: more than just glue. *Mol Membr Biol*. 1998;15(4): 167–176.
19. Magie CR, Martindale MQ. Cell-cell adhesion in the cnidaria: insights into the evolution of tissue morphogenesis. *Biol Bull*. 2008;214(3): 218–232.
20. Buckley CD, Simmons DL. Cell adhesion: a new target for therapy. *Mol Med Today*. 1997;3(10):449–456.
21. Klaile E, Muller MM, Kannicht C, et al. The cell adhesion receptor carcinoembryonic antigen-related cell adhesion molecule 1 regulates nucleocytoplasmic trafficking of DNA polymerase delta-interacting protein 38. *J Biol Chem*. 2007;282(36):26629–26640.
22. Martin M, Simon-Assmann P, Kedinger M. et al. DCC regulates cell adhesion in human colon cancer derived HT-29 cells and associates with ezrin. *Eur J Cell Biol*. 2006;85(8):769–783.
23. Sivalenka RR, Jessberger R. SWAP-70 regulates c-kit-induced mast cell activation, cell-cell adhesion, and migration. *Mol Cell Biol*. 2004;24(23):10277–10288.
24. Smith SF. Homology search with binary and trinary scoring matrices. *Int J Bioinform Res Appl*. 2006;2(2):119–131.
25. Deller MC, Yvonne Jones E. Cell surface receptors. *Curr Opin Struct Biol*. 2000;10(2):213–219.
26. Keller JM, Gray MR, Givens JA. Fuzzy k-nearest neighbor algorithm. *IEEE Trans Syst Man Cybern*. 1985;15(4):580–585.
27. Tung TQ, Lee D. A method to improve protein subcellular localization prediction by integrating various biological data sources. *BMC Bioinformatics*. 2009;10(Suppl 1):S43.
28. Nasibov E, Kandemir-Cavas C. Protein subcellular location prediction using optimally weighted fuzzy k-NN algorithm. *Comput Biol Chem*. 2008;32(6):448–451.
29. Sengur A. An expert system based on principal component analysis, artificial immune system and fuzzy k-NN for diagnosis of valvular heart diseases. *Comput Biol Med*. 2008;38(3):329–338.
30. Hynes RO. Cell adhesion: old and new questions. *Trends Cell Biol*. 1999;9(12):M33–M37.
31. Wheelock MJ, Shintani Y, Maeda M. Fukumoto Y, Johnson KR. Cadherin switching. *J Cell Sci*. 2008;121(Pt 6):727–735.
32. Boggon TJ, Murray J, Chappuis-Flament S, Wong E, Gumbiner BM, Shapiro L. C-cadherin ectodomain structure and implications for cell adhesion mechanisms. *Science*. 2002;296(5571):1308–1313.
33. Zhang Y, Yeh J, Richardson PM, Bo X. Cell adhesion molecules of the immunoglobulin superfamily in axonal regeneration and neural repair. *Restor Neurol Neurosci*. 2008;26(2–3):81–96.
34. Aricescu AR, Jones EY. Immunoglobulin superfamily cell adhesion molecules: zippers and signals. *Curr Opin Cell Biol*. 2007;19(5): 543–550.
35. Doebis C, Siegmund K, Loddenkemper C, et al. Cellular players and role of selectin ligands in leukocyte recruitment in a T-cell-initiated delayed-type hypersensitivity reaction. *Am J Pathol*. 2008;173(4): 1067–1076.
36. Schreiber T, Ehlers S, Aly S, et al. Selectin ligand-independent priming and maintenance of T cell immunity during airborne tuberculosis. *J Immunol*. 2006;176(2):1131–1140.
37. Smalley DM, Ley K. L-selectin: mechanisms and physiological significance of ectodomain cleavage. *J Cell Mol Med*. 2005;9(2):255–266.
38. Akiyama SK. Integrins in cell adhesion and signaling. *Hum Cell*. 1996;9(3):181–186.
39. Dedhar S. Integrins and signal transduction. *Curr Opin Hematol*. 1999;6(1):37–43.
40. Tuckwell DS. Weston SA, Humphries MJ. Integrins: a review of their structure and mechanisms of ligand binding. *Symp Soc Exp Biol*. 1993; 47:107–136.
41. Haquin S, Oeuillet E, Pajon A, et al. Data management in structural genomics: an overview. *Methods Mol Biol*. 2008;426:49–79.
42. Kazakov AE, Vassieva O, Gelfand MS, Osterman A, Overbeek R. Bioinformatics classification and functional analysis of PhoH homologs. *In Silico Biol*. 2003;3(1–2):3–15.
43. Kurmangaliyev YZ, Gelfand MS. Computational analysis of splicing errors and mutations in human transcripts. *BMC Genomics*. 2008;9:13.
44. Liu H, Wong L. Data mining tools for biological sequences. *J Bioinform Comput Biol*. 2003;1(1):139–167.
45. Serini G, Napione L, Bussolino F. Integrins team up with tyrosine kinase receptors and plexins to control angiogenesis. *Curr Opin Hematol*. 2008;15(3):235–242.
46. Kingsford C, Salzberg SL. What are decision trees? *Nat Biotechnol*. 2008;26(9):1011–1013.
47. Han L, Wang Y, Bryant SH. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem. *BMC Bioinformatics*. 2008;9:401.
48. Ge G, Wong GW. Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics.* 2008;9:275.
49. Levitsky VG, Ignatieva EV, Ananko EA, et al. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. *BMC Bioinformatics* 2007;8:481.
50. Zhang MQ. Discriminant analysis and its application in DNA sequence motif recognition. *Brief Bioinform*. 2000;1(4):331–342.
51. Dai W, Teodoridis JM, Graham J, et al. Methylation Linear Discriminant Analysis (MLDA) for identifying differentially methylated CpG islands. *BMC Bioinformatics*. 2008;9:337.
52. Mereghetti P, Ganadu ML, Papaleo E, Fantucci P, De Gioia L. Validation of protein models by a neural network approach. *BMC Bioinformatics*. 2008;9:66.
53. Yang ZR, Hamer R. Bio-basis function neural networks in protein data mining. *Curr Pharm Des*. 2007;13(14):1403–1413.
54. Jung E, Kim J, Kim M, et al. Artificial neural network models for prediction of intestinal permeability of oligopeptides. *BMC Bioinformatics*. 2007;8:245.

55. Brummendorf T, Lemmon V. Immunoglobulin superfamily receptors: cis-interactions, intracellular adapters and alternative splicing regulate adhesion. *Curr Opin Cell Biol*. 2001;13(5):611–618.
56. Halloran MC, Wolman MA. Repulsion or adhesion: receptors make the call. *Curr Opin Cell Biol*. 2006;18(5):533–540.
57. Hauck CR, Agerer F, Muenzner P, Schmitter T. Cellular adhesion molecules as targets for bacterial infection. *Eur J Cell Biol*. 2006; 85(3–4):235–242.
58. Van Roy F, Berx G. The cell-cell adhesion molecule E-cadherin. *Cell Mol Life Sci*. 2008;65(23):3756–3788.