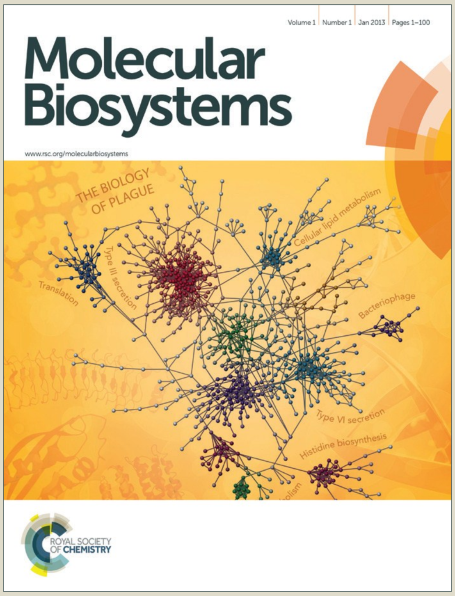
This article can be cited before page numbers have been issued, to do this please use: H. Tang, W. Chen and H. Lin, *Mol. BioSyst.*, 2016, DOI: 10.1039/C5MB00883B.


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Molecular Biosystems
www.rsc.org/molecularbiosystems
THE BIOLOGY OF PLAGUE
Cellular lipid metabolism
Type III secretion
Translation
Bacteriophage
Type VI secretion
Histidine biosynthesis
ROYAL SOCIETY OF CHEMISTRY

View Article Online
DOI: 10.1039/C5MB00883B

Molecular Biosystems

ARTICLE

Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique

Received 18th December 2015,
Accepted 00th

DOI: 10.1039/x0xx00000x

www.rsc.org/

Hua Tang,[a,]* Wei Chen [b,c] and Hao Lin [c,]*

Immunoglobulin, also called antibody, is a group of cell surface proteins which are produced by the immune system in response to the presence of a foreign substance (called antigen). They play key roles in many medical, diagnostic and biotechnological applications. Correct identification of the immunoglobulins is very crucial to the comprehension of humoral immune function. With the avalanche of protein sequences generated in postgenomic age, it is highly desired to develop computational methods to timely identify immunoglobulins. In view of this, we designed a predictor called "IGPred" by formulating protein sequences with the pseudo amino acid composition into which nine physiochemical properties of amino acids were incorporated. Jackknife cross-validated results showed that 96.3% immunoglobulins and 97.5% non-immunoglobulins can be correctly predicted, indicating that IGPred holds very high potential to become a useful tool for antibody analysis. For the convenience of most experimental scientists, a web-server for **IGPred** was established at http://lin.uestc.edu.cn/server/IGPred. We believe that the web-server will become a powerful tool to study immunoglobulins and to guide the related experimental validations.

1. Introduction

Immunoglobulin, called antibody, is a group of cell surface proteins which are involved in the recognition, binding, or adhesion processes of cells. They play crucial roles in the detection of the potentially harmful molecules [1]. When an alien substance enters the body, the immune system is capable of recognizing it as invader and subsequently activates B lymphocytes to secrete the immunoglobulin for attacking antigens. For example, when preventing a toxin from expression, immunoglobulins will neutralize the poison simply by changing its chemical composition. Stabilin-2 can bind to both Gram-positive and Gram-negative bacteria for defending against bacterial infection [2]. Some special immunoglobulins play an important role in the regulation of diseases. For instance, Kim-1 helps T-helper cell development and regulates asthma and allergic diseases. It is the key factor in kidney injury and repair. Kim-1 also acts as a receptor for hepatitis A virus, Dengue virus, ebolavirus and marburg virus by binding exposed phosphatidyl-serine at the surface of virion membrane [3-5]. Due to their special biological activity, immunoglobulins have been applied in many medical, diagnostic and biotechnological fields [6]. Thus, it is necessary to perform a deep study on immunoglobulins for understanding immune system and developing antibacterial-drug. However, the currently avalanche data do not allow us to investigate each protein because of costly and time-consuming biochemical experiments. Thus, development of computational method is a popular strategy to save experiment expenditure.

In fact, the three dimension structure prediction of immunoglobulin has attracted several scientists because it is the fundament of theoretical study on the interaction between ligand and receptor. Marcatili et al. have developed a method to predict the 3D structure of antibodies [7, 8]. The method builds a structural model of an antibody by only a few minutes (~10 min on average). A very satisfactory accuracy can be achieved. Based on the strategy, they constructed a web server called Pigs for the automatic modelling of immunoglobulin variable domains based on the canonical structure method. It allows users to choose templates (for the frameworks and the loops) and modelling strategies in an automatic or manual fashion. The prediction results on the target antibody can be freely downloaded or displayed on-line. There is no limitation on the number of submitted sequences. Thus, it is user-friendly and flexible. Klausen et al. developed another webserver called LYRA for lymphocyte receptor structural modelling [9]. For their benchmark dataset, the average RMSD accuracies of 1.29 and 1.48 Å were obtained for B- and T-cell receptors, respectively.

These tools do provide convenience to most of scholars. However, the first step to reveal the biological function of immunoglobulin is to correctly identify them. To the best of our knowledge, there is no such tool which can accurately judge whether a new protein is immunoglobulin or not. With the appearance of more and more protein data, it is urgent to develop a predictor to recognize immunoglobulins. In the past two decade, lots of predictors for protein structure and function have been developed based on machine learning methods and protein sequence information [10-30]. Encouraging results obtained by these references

[a.] Department of Pathophysiology, Sichuan Medical University, Luzhou 646000, China. E-mail: tanghua771211@aliyun.com
[b.] Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China.
[c.] Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China. E-mail: hlin@uestc.edu.cn

View Article Online
DOI: 10.1039/C5MB00883B

stimulate us to construct a powerful tool to discriminate immunoglobulins from non-immunoglobulins.

Thus, the present study was devoted to develop a computational method to predict immunoglobulins and build a friendly web-server for convenience. This work includes the following four sections: the benchmark dataset construction, description on immunoglobulin sequence, discriminated algorithm and predicted results.

2. Material and methods

2.1. Benchmark datasets

A high quality benchmark dataset can guarantee the reliability and accuracy of predictive model. The immunoglobulin superfamily information derives from the annotation in the Universal Protein Resource (Uniprot) [31]. The sequences of immunoglobulins were also extracted from the UniProt. The immunoglobulins usually locate in cell membrane or outside the cell, thus, to obtain a reliable contrast, we just selected the negative samples from the two subcellular locations: cell membrane and outside the cell. The amino acid sequences of negative samples were also downloaded from the Uniprot. To guarantee the quality of benchmark dataset, we performed the following steps to select proteins. At first, we excluded the protein sequence if it contains ambiguous residues (such as ‘‘X’’, ‘‘B’’ and ‘‘Z’’). Secondly, if a sequence is the fragment of other proteins, the sequence was excluded. Thirdly, we only chose the proteins from human, mouse and rat. Fourthly, to avoid any similarity bias which will result in the overestimate of predicted results, we used the CD-HIT program [32] to remove the highly similar sequences by setting the cutoff of sequence identity as 60%. In fact, if using 25% sequence identity as cutoff, the dataset will be more strict and objective. However, the current data do not allow us to do so; otherwise, the number of proteins would be too few to have statistical significance. After such a screening procedure, we finally obtained 228 samples for the benchmark dataset S as formulated as follows:

$$\mathrm{S} = \mathrm{S}_{\mathrm{im}} \cup \mathrm{S}_{\mathrm{non-im}} \qquad (1)$$

where the subset S_{im} contains 109 immunoglobulin samples, $\mathrm{S}_{\mathrm{non-im}}$ contains 119 non-immunoglobulin samples. The symbol ∪ represents the union in the set theory. The codes of 228 sequences can be freely downloaded from http://lin.uestc.edu.cn/server/IGPred/data.

To further investigate the prediction capability of the proposed model, we collected a test dataset from Uniprot. These proteins were obtained according to the same step of the training data construction. If a protein overlaps with the sample in training set, the protein will be excluded. Thus, the test data is independent from the train data. As a result, total of 20 immunoglobulins and 20 non-immunoglobulins were achieved and can be freely checked and downloaded from http://lin.uestc.edu.cn/server/IGPred/data.

2.2. The representation of peptide samples

How to formulate protein sequences with an effective mathematical expression is a crucial step in immunoglobulin prediction. A straightforward way is to use entire amino acid sequence of protein as formulated by

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\ldots\mathrm{R}_L \qquad (2)$$

where R_1, R_2 and R_L denote the 1st, 2nd and Lth residue of the protein sample **P**. Such formulation is easily utilized by various sequence similarity search tools, such as BLAST and FASTA, to perform statistical prediction. The results are always good for the query sequences which have high similar sequences in searching dataset. Thus, the similar-based method is straightforward and intuitive. However, it is failed to work when the similar sequences of the query sequences are not found in the training dataset.

Thus, using discrete vectors to represent protein samples have been proposed in protein classification. The pseudo amino acid composition (PseAAC) is a widely used method to represent protein sequences because it can not only include amino acid composition, but also contain the correlation of physicochemical properties between two residues [11, 33-46]. Based on the concept of PseAAC, we made an improved revision on PseAAC by replacing amino acid composition with *g*-gap dipeptide composition. Accordingly, each protein in our benchmark dataset can be defined by a 400+*nλ* dimension vector as formulated by:

$$\mathbf{P} = [x_1 \cdots x_{400} x_{400+1} \cdots x_{400+n\lambda}]^T \qquad (3)$$

here

$$x_u = \begin{cases} f_u & (1 \le u \le 400) \\ \tau_u & (400+1 \le u \le 400+n\lambda) \end{cases} \qquad (4)$$

In Eq. (4), the f_u is the normalized frequency of the *g*-gap dipeptides in protein **P**, and can be expressed as:

$$f_u = \frac{n_u}{\sum_u n_u} \qquad (5)$$

where n_u denotes the number of the *u*-th *g*-gap dipeptide in the protein **P**. The τ_j in Eq.(4) is the *j*-tier sequence correlation factor computed by the following formula:

$$\begin{cases} \tau_1 = \frac{1}{L-1}\sum_{k=1}^{L-1} H_{k,k+1}^1 \\ \tau_2 = \frac{1}{L-1}\sum_{k=1}^{L-1} H_{k,k+1}^2 \\ \quad\vdots \\ \tau_n = \frac{1}{L-1}\sum_{k=1}^{L-1} H_{k,k+1}^n \\ \tau_{n+1} = \frac{1}{L-2}\sum_{k=1}^{L-2} H_{k,k+2}^1 \\ \tau_{n+2} = \frac{1}{L-2}\sum_{k=1}^{L-2} H_{k,k+2}^2 \\ \quad\vdots \\ \tau_{n+n} = \frac{1}{L-2}\sum_{k=1}^{L-1} H_{k,k+2}^n \\ \quad\vdots \\ \tau_{n\lambda} = \frac{1}{L-\lambda}\sum_{k=1}^{L-\lambda} H_{k,k+\lambda}^n \end{cases} \quad (\lambda < L) \qquad (6)$$

where $H_{k,k+\lambda}^n$ is the correlation function and can be given by

$$H_{k,k+\lambda}^n = h^n(R_k) \cdot h^n(R_{k+\lambda}) \qquad (7)$$

where $h^n(R_k)$ is the *n*-th kind of the physicochemical values of the amino acid R_k. The values should be converted to standard type by:

$$h^n(R_k) = \frac{h_0^n(R_k) - \langle h_0^n(R_k)\rangle}{SD\langle h_0^n(R_k)\rangle} \qquad (8)$$

where $h_0^n(R_k)$ is the original physicochemical values of the *k*-th amino acid.

According to Eqs (3), each protein can be expressed by 400+*nλ* features. To optimize feature subset, based on the analysis of variance (ANOVA) [13, 47], we used a feature selection technique to rank features defined as:

View Article Online
DOI: 10.1039/C5MB00883B

$$F(u) = \frac{\sum_{i=1}^{2} m_i \left(\frac{\sum_{j=1}^{m_i} x_{u(i,j)}}{m_i} - \frac{\sum_{i=1}^{2}\sum_{j=1}^{m_i} x_{u(i,j)}}{\sum_{i=1}^{2} m_i} \right)^2}{\sum_{i=1}^{2}\sum_{j=1}^{m_i} \left(x_u(i,j) - \frac{\sum_{j=1}^{m_i} x_{u(i,j)}}{m_i} \right)^2 \Big/ (m_1+m_2-2)} \quad (9)$$

where $x_u(i,j)$ denotes the frequency of the *u*-th features of the *j*-th sample in the *i*-th group; m_i denotes the number of samples in the *i*-th group (here m_1=109, m_2=119). It is obviously that the larger the $F(u)$ value, the better discriminative capability the *u*-th feature has.

2.3. Support vector machine

Several methods, such as, fisher discrimination (FD) [14], neural network (NN) [48], ensemble learning [49-53] and k-nearest neighbors (KNN) [20] have been applied in protein classification. In this study, we selected support vector machine (SVM) to perform discrimination as its excellent learning ability, especially for small samples. The basic idea of SVM is to transform the input vector into a high-dimension Hilbert space and to seek a separating hyperplane in this space. The radial basis function (RBF) defined as $K(\vec{p_i}, \vec{p_j}) = exp(-\gamma\|\vec{p_i} - \vec{p_j}\|^2)$ were used in the current study because it is suitable for non-linear classification. We used a grid search method with 5-fold cross-validation test to obtain the best values for the regularization parameter *C* and kernel parameter *γ*. To implement SVM, the soft package LibSVM (version 2.88) (https://www.csie.ntu.edu.tw/~cjlin/ libsvm/) was used.

2.4. Performance evaluation

The following four indexes called sensitivity (*Sn*), specificity (*Sp*), and overall accuracy (*Acc*) were introduced to measure the prediction quality [19, 47, 54].

$$Sn = \frac{TP}{TP+FN} \qquad 0 \le Sn \le 1 \qquad (10)$$

$$Sp = \frac{TN}{TN+FP} \qquad 0 \le Sp \le 1 \qquad (11)$$

$$Acc = \frac{TP+TN}{TP+TN+PF+FN} \qquad 0 \le Acc \le 1 \qquad (12)$$

where *TP*, *TN*, *FP* and *FN* are the true positive, true negative, false positive and false negative, respectively.

We also plotted the receiver operating characteristic (ROC) curves to show the predictive capability of our method across the entire range of SVM decision values. The ROC curve also presents the model behaviour of true positive rate (sensitivity) against false positive rate (1-specificity) in a visual way. The area under the ROC (auROC) was calculated to quantitatively and objectively measure the performance of proposed method. A perfect classifier gives AUC = 1, the random performance gives AUC = 0.5.

3. Results and Discussion

3.1. Physicochemical properties

The physicochemical properties of amino acids do play important roles in protein structure and function. In this work, six widely used properties that are hydrophobicity, hydrophilicity, side chain mass, pK of the α-COOH group, pK of the α-NH_3^+ group and pI at 25°C were utilized in the Eqs. (4-8). We also introduced three new characteristics of amino acids called rigidity, flexibility and irreplaceability.

Table 1. The nine physicochemical properties used in the current study

Amino acids	Hydrophobicity	Hydrophilicity	Mass	pK1	pK2	pI	Rigidity	Flexibility	Irreplaceability
A	0.62	-0.5	15	2.35	9.87	6.11	-1.338	-3.102	0.52
C	0.29	-1	47	1.71	10.78	5.02	-1.511	0.957	1.12
D	-0.9	3	59	1.88	9.6	2.98	-0.204	0.424	0.77
E	-0.74	3	73	2.19	9.67	3.08	-0.365	2.009	0.76
F	1.19	-2.5	91	2.58	9.24	5.91	2.877	-0.466	0.86
G	0.48	0	1	2.34	9.6	6.06	-1.097	-2.746	0.56
H	-0.4	-0.5	82	1.78	8.97	7.64	2.269	-0.223	0.94
I	1.38	-1.8	57	2.32	9.76	6.04	-1.741	0.424	0.65
K	-1.5	3	73	2.2	8.9	9.47	-1.822	3.950	0.81
L	1.06	-1.8	57	2.36	9.6	6.04	-1.741	0.424	0.58
M	0.64	-1.3	75	2.28	9.21	5.74	-1.741	2.484	1.25
N	-0.78	0.2	58	2.18	9.09	10.76	-0.204	0.424	0.79
P	0.12	0	42	1.99	10.6	6.3	1.979	-2.404	0.61
Q	-0.85	0.2	72	2.17	9.13	5.65	-0.365	2.009	0.86
R	-2.53	3	101	2.18	9.09	10.76	1.169	3.060	0.60
S	-0.18	0.3	31	2.21	9.15	5.68	-1.511	0.957	0.64
T	-0.05	-0.4	45	2.15	9.12	5.6	-1.641	-1.339	0.56
V	1.08	-1.5	43	2.29	9.74	6.02	-1.641	-1.339	0.54
W	0.81	-3.4	130	2.38	9.39	5.88	5.913	-1.000	1.82
Y	0.26	-2.3	107	2.2	9.11	5.63	2.714	-0.672	0.98

The flexibility and rigidity of amino acid side chains may comprise important information for the understanding of protein structure and function [55]. The flexibility and rigidity scales were generated by descriptor projection to vectors using principle

View Article Online
DOI: 10.1039/C5MB00883B

component analysis. In the evolution, some residues are easily replaceable, but others are difficult. Thus, averaged mutational deteriorations (AMD) of amino acids can be used to describe their irreplaceability [56]. It shows that the rarer is substitution for a residue, the higher is its value of AMD. The irreplaceability is a response to mutational deterioration in the course of evolution of life. The values of nine physicochemical properties for 20 amino acids were all listed in Table 1.

3.2. Predictive accuracy

Based on above description that nine physicochemical properties were used, we have 400+9λ features, namely n=9 in Eqs.(3-6). Subsequently, we must determine the sequence correlation factor λ. In order to include long-range correlated information as more as possible, meanwhile, without wasting computational source, we set the λ=10. Accordingly, for each g-gap dipeptide, each protein in benchmark dataset is represented by a 490 dimension vector.

For the purpose of obtaining the best predictive performance, we should pick out the optimal features which can produce the maximum Acc. If we investigated all the combinations of features, the best feature subset must be obtained. However, the number of all possible combinations for 490 features is so huge that it is beyond computational capability for most computers. Thus, it is impossible to examine the performance of all feature subsets. To save the computational time, the F score defined in Eq.(9) was used to perform feature selection. We initially ranked all features according to their F scores from large to small. Subsequently, the *Acc* of the first feature with the largest $F(u)$ was investigated by using SVM. Furthermore, we examined the performance of a new feature subset which was produced by adding the feature with the second highest F value into the former feature subset. This process was repeated from the large to the small $F(u)$ value until 490 combinations were examined. The optimal feature subset can be achieved when the best predictive accuracy was observed. On the basis of the feature selection, the high-dimensional data will be projected into a low-dimensional space. The optimal feature subset was used to build the final predictive model.

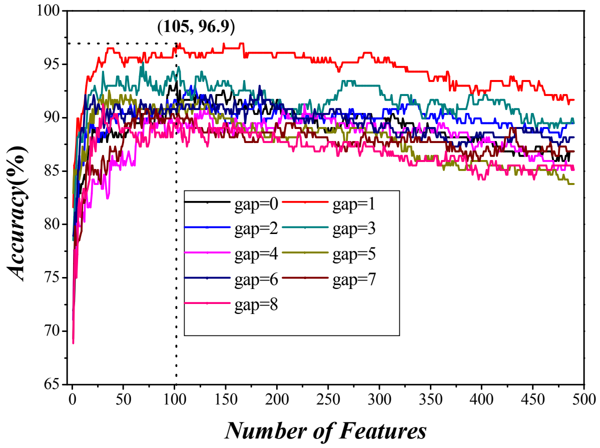


Fig. 1 A plot to show the feature selection results. When the top 105 feature were used to perform prediction, the overall success rate reached its peak of 96.9%.

We varied the parameter g from 0 to 8. Then total of 4420 (490×9) feature subsets should be investigated. As a result, 9 curves were plotted in a 2D Cartesian coordinate system with the number of features as its abscissa and the *Acc* as its ordinate (Figure 1). For statistical predictive test, independent dataset test, n-fold cross-validation test and jackknife test are three widely used strategies [57-62]. Jackknife test can yield unique result for a given benchmark dataset. Thus, it has been widely used to evaluate the performance of the proposed methods in practical application. For time saving, we used 5-fold cross-validation in feature selection. Once the best feature subset was found, the jackknife test was used to measure the performance of the feature subset again. As shown in Fig.1, the maximum Acc is 96.9% when the top ranked 105 features with g=1 were used. The *Sn* and *Sp* are 96.3% and 97.5%, respectively in jackknife test. The false positive rate is only 2.5%. Such high accuracy suggested that our method can correctly identify immunoglobulin.

To describe the performance of our model with 105 features across the entire range of SVM decision values, the ROC curve were also provided in Figure 2. We noticed that the curve is close left and top coordinate axis, demonstrating that the model is very suitable for classification. The auROC is 0.994 in jackknife cross-validation.

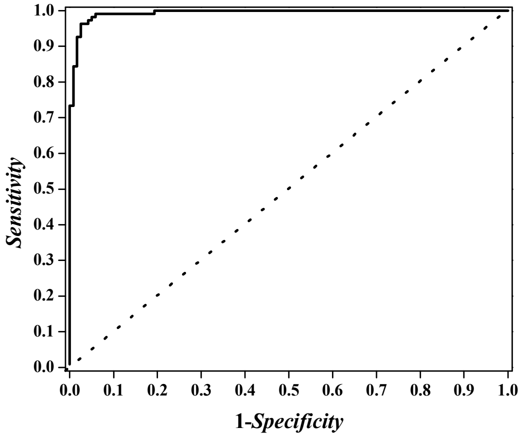


Fig. 2 The ROC curve for the model with 105 optimal 1-gap dipeptides in jackknife cross-validation. The diagonal dot line denotes a random guess with the auROC of 0.5.

We further investigated the performances of three state of-the-art classifiers: Naïve Bayes, BayesNet and RBFNetwork on the same benchmark dataset using the same features. The predicted results are all recorded in Table 1. Comparison in Table 1 demonstrates that the SVM is the best one among all algorithms for immunoglobulin predictions.

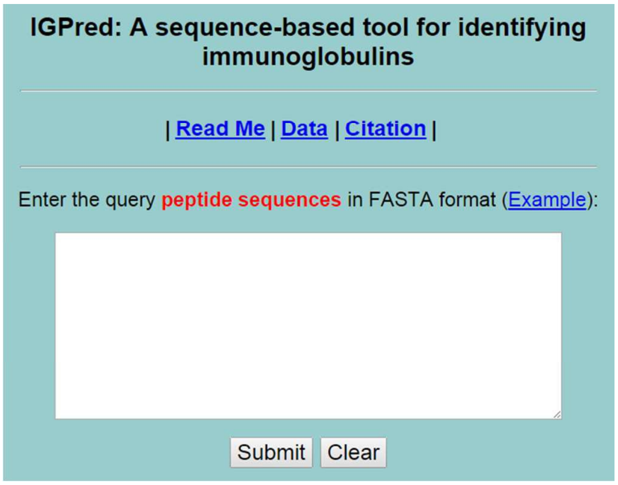
Table 1 Comparing the performance of different algorithms

Algorithm	*Sn*(%)	*Sp*(%)	*Acc*(%)	auROC
SVM	96.3	97.5	96.9	0.994
Naïve Bayes	89.9	90.8	90.4	0.958
BayesNet	92.7	92.4	92.5	0.974
RBFNetwork	88.1	89.9	89.0	0.887

View Article Online
DOI: 10.1039/C5MB00883B

3.3. Web-server construction

According to above calculation, a user-friendly web-server called **IGPred** was constructed as shown in Figure 3. Users may browse the web server at http://lin.uestc.edu.cn/server/IGPred. The Read Me button provides a brief introduction about the predictor and the caveat when using it. The Data button lists a link for downloading the benchmark datasets. The Citation button gives the relevant paper of **IGPred**. The Example button provides example sequences in FASTA format. Users may type or copy/paste the query peptide sequences with FASTA format into the input box at the centre of main page. After submitting protein sequences, results will be shown in a new interface.



IGPred: A sequence-based tool for identifying immunoglobulins

| Read Me | Data | Citation |

Enter the query peptide sequences in FASTA format (Example):

Fig. 3 A semi-screenshot to show the top page of the **IGPred** webserver. Its website address is http://lin.uestc.edu.cn/server/IGPred.

3.4. Further discussion

To examine the prediction capability of the web-server on independent data, we collected 20 immunoglobulins and 20 non-immunoglobulins from Uniprot, which are independent from train data. By submitting these data into the web-server, we noticed that all proteins can be correctly recognized, suggesting that the server is powerful and there is a low false positive rate of **IGPred** for the external validation sets. Of course, the **IGPred** was constructed based on the proteins from cell membrane or outside the cell. Thus, the model just focuses on the proteins from the two locations.

Over-fitting is an inevitable problem in machine learning. Previous studies [12, 63, 64] have shown that protein sequence descriptor-based machine learning models often have high risk of over-fitting. In this study, to reduce the risk of over-fitting, we used cross-validation and independent data to evaluate the performance of the proposed model. High accuracies demonstrate that the model is reasonable.

According to reference [65], the irrelevant, redundant information and inter-correlation of different features can also generate the risk of over-fitting, thus, we used a feature selection technique rank the original protein sequence features. The original high dimensional features of 490 vectors were reduced to 105 vectors. The higher predictive performance was yielded in jackknife cross validation. Low dimension feature can also contribute to avoid the over-fitting. Results showed that the reasonable feature selection technique can improve the predictive accuracies of the predictive model in the cross-validation.

4. Conclusion

Immunoglobulin is the most important component in immune system. The knowledge for immunoglobulin is conductive to the development of anti-disease drugs. Thus, we performed a theoretical work to discriminate immunoglobulins from non-immunoglobulin. A very high accuracy model was obtained. Results demonstrate that the proposed method can efficiently pick out informative features and improve predictive performance. Based on the optimal model, an online predictor **IGPred** was established for identifying immunoglobulins. We are sure that this predictor will become a useful tool for immunoglobulin analysis and further experimental research. Moreover, the method proposed in this study can be generalized to the prediction of other proteomics.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HT and HL conceived and designed the experiments. HT and HL performed the feature selection and the analysis. WC and HL constructed the Web server. HT, WC and HL wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Applied Basic Research Program of Sichuan Province (No. 2015JY0100 and LZ-LY-45), the Scientific Research Foundation of the Education Department of Sichuan Province (11ZB122), the Nature Scientific Foundation of Hebei Province (No.C2013209105) and Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028).

References

1. A. N. Barclay, *Seminars in immunology*, 2003, **15**, 215-223.
2. H. Adachi and M. Tsujimoto, *The Journal of biological chemistry*, 2002, **277**, 34264-34270.
3. D. Feigelstock, P. Thompson, P. Mattoo, Y. Zhang and G. G. Kaplan, *Journal of virology*, 1998, **72**, 6621-6628.
4. A. S. Kondratowicz, N. J. Lennemann, P. L. Sinn, R. A. Davey, C. L. Hunt, S. Moller-Tank, D. K. Meyerholz, P. Rennert, R. F. Mullins, M. Brindley, L. M. Sandersfeld, K. Quinn, M. Weller, P. B. McCray, Jr., J. Chiorini and W. Maury, *Proceedings of the National Academy of Sciences of the United States of America*, 2011, **108**, 8426-8431.
5. L. Meertens, X. Carnec, M. P. Lecoin, R. Ramdasi, F. Guivel-Benhassine, E. Lew, G. Lemke, O. Schwartz and A. Amara, *Cell host & microbe*, 2012, **12**, 544-557.
6. G. A. Nevinsky and V. N. Buneva, *Journal of immunological methods*, 2002, **269**, 235-249.

View Article Online
DOI: 10.1039/C5MB00883B

7. P. Marcatili, P. P. Olimpieri, A. Chailyan and A. Tramontano, *Nature protocols*, 2014, **9**, 2771-2783.
8. P. Marcatili, A. Rosi and A. Tramontano, *Bioinformatics*, 2008, **24**, 1953-1954.
9. M. S. Klausen, M. V. Anderson, M. C. Jespersen, M. Nielsen and P. Marcatili, *Nucleic acids research*, 2015, **43**, W349-355.
10. L. Chen, C. Chu, T. Huang, X. Kong and Y. D. Cai, *Amino acids*, 2015, **47**, 1485-1493.
11. K. C. Chou, *Proteins*, 2001, **43**, 246-255.
12. C. Ding, L. F. Yuan, S. H. Guo, H. Lin and W. Chen, *Journal of proteomics*, 2012, **77**, 321-328.
13. H. Ding and D. Li, *Amino acids*, 2015, **47**, 329-333.
14. H. Ding, L. Luo and H. Lin, *Protein and peptide letters*, 2009, **16**, 351-355.
15. G. L. Fan, X. Y. Zhang, Y. L. Liu, Y. Nang and H. Wang, *Journal of computational chemistry*, 2015, **36**, 2317-2327.
16. Y. E. Feng, *Interdisciplinary sciences, computational life sciences*, 2015.
17. T. A. Holton, G. Pollastri, D. C. Shields and C. Mooney, *Bioinformatics*, 2013, **29**, 3094-3096.
18. X. Ma, J. Guo and X. Sun, *BioMed research international*, 2015, **2015**, 425810.
19. W. S. Sanders, C. I. Johnston, S. M. Bridges, S. C. Burgess and K. O. Willeford, *PLoS computational biology*, 2011, **7**, e1002101.
20. A. Suratanee and K. Plaimas, *Journal of bioinformatics and computational biology*, 2014, **12**, 1450017.
21. P. P. Zhu, W. C. Li, Z. J. Zhong, E. Z. Deng, H. Ding, W. Chen and H. Lin, *Molecular bioSystems*, 2015, **11**, 558-563.
22. X. Y. Cheng, W. J. Huang, S. C. Hu, H. L. Zhang, H. Wang, J. X. Zhang, H. H. Lin, Y. Z. Chen, Q. Zou and Z. L. Ji, *PloS one*, 2012, **7**, e38979.
23. C. Lin, Y. Zou, J. Qin, X. Liu, Y. Jiang, C. Ke and Q. Zou, *PloS one*, 2013, **8**, e56499.
24. E. S. Olson, T. A. Aguilera, T. Jiang, L. G. Ellies, Q. T. Nguyen, E. H. Wong, L. A. Gross and R. Y. Tsien, *Integrative biology : quantitative biosciences from nano to macro*, 2009, **1**, 382-393.
25. B. Liu, J. H. Xu, X. Lan, R. F. Xu, J. Y. Zhou, X. L. Wang and K. C. Chou, *PloS one*, 2014, **9**.
26. B. Liu, D. Y. Zhang, R. F. Xu, J. H. Xu, X. L. Wang, Q. C. Chen, Q. W. Dong and K. C. Chou, *Bioinformatics*, 2014, **30**, 472-479.
27. L. Song, D. Li, X. Zeng, Y. Wu, L. Guo and Q. Zou, *BMC bioinformatics*, 2014, **15**, 298.
28. L. Wei, M. Liao, X. Gao and Q. Zou, *IEEE transactions on nanobioscience*, 2015, **14**, 649-659.
29. P. M. Feng, W. Chen, H. Lin and K. C. Chou, *Analytical biochemistry*, 2013, **442**, 118-125.
30. P. M. Feng, H. Lin and W. Chen, *Computational and mathematical methods in medicine*, 2013, **2013**, 567529.
31. E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A. J. Bridge, S. Poux, L. Bougueleret and I. Xenarios, *Methods in molecular biology*, 2016, **1374**, 23-54.
32. L. Fu, B. Niu, Z. Zhu, S. Wu and W. Li, *Bioinformatics*, 2012, **28**, 3150-3152.
33. S. Ahmad, M. Kabir and M. Hayat, *Comput Meth Prog Bio*, 2015, **122**, 165-174.
34. K. C. Chou, *Bioinformatics*, 2005, **21**, 10-19.
35. G. L. Fan, X. Y. Zhang, Y. L. Liu, Y. Nang and H. Wang, *Journal of computational chemistry*, 2015, **36**, 2317-2327.
36. M. Hayat and N. Iqbal, *Comput Meth Prog Bio*, 2014, **116**, 184-192.
37. M. Hayat and A. Khan, *Protein and peptide letters*, 2012, **19**, 411-421.
38. H. Mohabatkar, M. M. Beigi, K. Abdolahi and S. Mohsenzadeh, *Med Chem*, 2013, **9**, 133-137.
39. M. Mohammad Beigi, M. Behjati and H. Mohabatkar, *Journal of structural and functional genomics*, 2011, **12**, 191-197.
40. L. Nanni, S. Brahnam and A. Lumini, *Amino acids*, 2012, **43**, 657-665.
41. L. Nanni, A. Lumini, D. Gupta and A. Garg, *Ieee Acm T Comput Bi*, 2012, **9**, 467-475.
42. S. S. Sahu and G. Panda, *Comput Biol Chem*, 2010, **34**, 320-327.
43. X. Wang, W. W. Zhang, Q. W. Zhang and G. Z. Li, *Bioinformatics*, 2015, **31**, 2639-2645.
44. X. B. Zhou, C. Chen, Z. C. Li and X. Y. Zou, *Journal of theoretical biology*, 2007, **248**, 546-551.
45. H. Matsui, K. Tomizawa and M. Matsushita, *Nihon yakurigaku zasshi. Folia pharmacologica Japonica*, 2003, **121**, 435-439.
46. M. Lindgren, M. Hallbrink, A. Prochiantz and U. Langel, *Trends in pharmacological sciences*, 2000, **21**, 99-103.
47. H. Ding, P. M. Feng, W. Chen and H. Lin, *Molecular bioSystems*, 2014, **10**, 2229-2235.
48. D. A. Dobchev, I. Mager, I. Tulp, G. Karelson, T. Tamm, K. Tamm, J. Janes, U. Langel and M. Karelson, *Current computer-aided drug design*, 2010.
49. C. Lin, W. Q. Chen, C. Qiu, Y. F. Wu, S. Krishnan and Q. Zou, *Neurocomputing*, 2014, **123**, 424-435.
50. J. S. Orange and M. J. May, *Cellular and molecular life sciences : CMLS*, 2008, **65**, 3564-3591.
51. L. Y. Wei, M. H. Liao, Y. Gao, R. R. Ji, Z. Y. He and Q. Zou, *Ieee Acm T Comput Bi*, 2014, **11**, 192-201.
52. Q. Zou, Z. Wang, X. Guan, B. Liu, Y. Wu and Z. Lin, *BioMed research international*, 2013, **2013**, 686090.
53. Q. Zou, J. C. Zeng, L. J. Cao and R. R. Ji, *Neurocomputing*, 2016, **173**, 346-354.
54. B. Liu, L. Fang, F. Liu, X. Wang and K. C. Chou, *Journal of biomolecular structure & dynamics*, 2016, **34**, 223-235.
55. J. Gottfries and L. Eriksson, *Molecular diversity*, 2010, **14**, 709-718.
56. L. F. Luo, *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life*, 1988, **18**, 65-70.
57. S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, W. Chen and K. C. Chou, *Bioinformatics*, 2014, **30**, 1522-1529.
58. H. Lin and W. Chen, *Journal of microbiological methods*, 2011, **84**, 67-70.
59. H. Lin, E. Z. Deng, H. Ding, W. Chen and K. C. Chou, *Nucleic acids research*, 2014, **42**, 12961-12972.
60. W. Chen, P. Feng, H. Ding, H. Lin and K. C. Chou, *Analytical biochemistry*, 2015, **490**, 26-33.
61. P. Feng, W. Chen and H. Lin, *Genomics*, 2014, **104**, 229-233.
62. H. Ding, L. Liu, F. B. Guo, J. Huang and H. Lin, *Protein and peptide letters*, 2011, **18**, 58-63.

View Article Online
DOI: 10.1039/C5MB00883B

63. F. Cheng, W. Li, G. Liu and Y. Tang, *Current topics in medicinal chemistry*, 2013, **13**, 1273-1289.

64. B. Liu, J. H. Xu, Q. Zou, R. F. Xu, X. L. Wang and Q. C. Chen, *BMC bioinformatics*, 2014, **15**.

65. F. X. Cheng, Y. D. Zhou, J. Li, W. H. Li, G. X. Liu and Y. Tang, *Molecular bioSystems*, 2012, **8**, 2373-2384.