Title

Pseudo amino acid composition improves antifreeze protein prediction

Sukanta Mondal* and Priyadarshini P. Pai

Department of Biological Sciences, Birla Institute of Technology and Science–Pilani, K.K. Birla Goa Campus, Zuarinagar, Goa 403 726, India

*Corresponding author

Tel: +91-832-258-0149

Fax: +91-832-255-7031

Email: suku@goa.bits-pilani.ac.in; sukanta.mondal@gmail.com

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

Abstract

Antifreeze proteins (AFP) in living organisms play a key role in their tolerance to extremely cold temperatures and have wide range of biotechnological applications. But on account of diversity, their identification has been challenging to biologists. Earlier work explored in this area did not cover introduction of sequence order information, known to represent important properties of various proteins and protein systems for prediction of their attributes. In this study, the effect of Chou's pseudo amino acid composition that presents sequence order of proteins was systematically explored using support vector machines for AFP prediction. Our findings suggest that introduction of sequence order information helps identify AFPs with an accuracy of 84.75% on independent test dataset, outperforming approaches such as AFP-Pred and iAFP. The relative performance calculated using Youden's Index (Sensitivity + Specificity -1) was found to be 0.71 for our predictor (AFP-PseAAC), 0.48 for AFP-Pred and 0.05 for iAFP. We hope this novel prediction approach will aid in AFP based research for biotechnological applications.

Keywords:

Convergent evolution; Support Vector Machines; Sequence order effect; Pseudo amino acid composition; Ten-fold cross-validation.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

1. Introduction

Antifreeze proteins (AFPs), also known as ice-structuring proteins, are a diverse group of polypeptides found in animals, plants, microbes, especially in fish inhabitants of ice-laden sea water, which provide protection from freezing in extremely cold environments [1]. This defense is imparted to the organisms at cellular levels as AFPs have a unique ability to adsorb onto the surface of ice [2]. Depending upon the surrounding, organisms adopt two strategies namely freeze tolerance and freeze avoidance to survive at low and subzero temperatures [3, 4], which may account for the diversity observed among various species.

Analyses of AFPs from fish, insects and plants have shown that there is no consensus sequence or structure for an ice-binding domain. Such an insight, at sequence or structural level, is important for understanding protein-ice interactions and freeze tolerance mechanisms of AFPs. Since these proteins hold a promising scope for wide range of biotechnological applications in industry, medicine, food technology, cell lines and organ preservation, cryosurgery and transgenics, gaining knowledge into their functional mechanisms has become increasingly essential [5].

With the enormous amount of genomic data available today, a rapid, specific and highly precise automated approach is desirable for identification and annotations of AFPs. Researchers, encouraged by the overwhelming success of machine learning methods in protein classification and function prediction [5-14], developed sequence based solutions such as AFP-Pred [5], AFP_PSSM [13] and iAFP approach [14]. These methods explored physicochemical properties [5], evolutionary information [13], n-peptide composition and feature based coding schemes [14]

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

for prediction. However, the scope for truly reflecting the intrinsic correlation of sequence representatives with the object to be predicted by machine learning remained.

In this regard, amino acid composition (AAC) was explored to include sequence order information during prediction of protein attributes. Sequential and Discrete models were formulated to represent various proteins. But using these straightforward models was not so fruitful in preserving the sequence order information. To address this issue, the concept of pseudo amino acid composition (pseAAC) was proposed [15].

In pseAAC, protein sequences are represented as discrete models yet without completely losing the sequence order information. Ever since this idea of pseAAC was proposed, many models for addressing various kinds of problems in proteins and protein related systems have been put forth. Further, different modes of optimal pseAAC composition are known to correspond to different protein attributes. Subsequently, the process of key components selection from the trivial ones for obtaining its pseAAC has become challenging. However, as pseAAC gives important direction for further improvement of the quality of protein attribute prediction, it has captured the interest of biologists [15].

In this study, we have explored the effect of introducing sequence order information in prediction of AFPs by using Chou's pseudo amino acid composition based protein features extracted from AFP dataset [5] followed by classification using Support Vector Machines [16]. Aspects of example selection, influence of large numbers of negative examples and pseAAC modes were searched for development of the AFP predictor. Training was done using ten-fold

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

cross-validation [17] followed by testing on independent test dataset to analyze if sequence order information improved the overall prediction accuracy of AFPs. The findings of our study can facilitate AFP based studies and applications.

2. Materials and Methods

2.1. Dataset

We obtained AFP dataset used for the development of AFP-Pred [5]. Briefly, antifreeze protein sequences were collected from seed proteins of the Pfam database [18], enriched by performing Position Specific Iteration -Basic Local Alignment Search Tool (PSI-BLAST) [19] with a string threshold (E-value) of 0.001, and followed by manual inspection for presence of AFPs. Further, the dataset was free of incomplete sequences and homology bias at ≥ 40% sequence similarity. Altogether comprising of 481 AFPs (positive examples) and 9493 non-AFPs (negative examples).

2.2. Pseudo amino acid composition

Representing protein sequences with sequence order information herein is done using pseudo amino acid composition which is known for its applications in dealing with various kinds of problems in proteins and protein related systems. To develop an effective predictor, a powerful prediction algorithm with an effective mathematical expression, truly representing the protein sequence in correlation with the object to be predicted. Different properties of amino acids in the proteins correspond to different modes of pseAAC composition. The discrete models so derived from the proteins used in our study were based on the following mathematical function described

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

below [9, 15]. The composition of a given protein **P** containing amino acid residues of length L is depicted as:

$$\boldsymbol{P} = R_1 R_2 R_3 R_4 R_5 \ldots R_L \qquad (1)$$

where R_1 represents the 1st residue, R_2 represents the 2nd residue,… and R_L the L-th residue, and they each belong to one of the 20 native amino acids.

In the classical mode (Type 1), amino acid composition is expressed as:

$$\boldsymbol{P} = [p_1, p_2, \ldots, p_{20}, p_{20+1}, \ldots, p_{20+\lambda}]^T, \quad (\lambda < L) \qquad (2)$$

where the 20 + λ components are given by:

$$p_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + \omega \sum_{k=1}^{\lambda} \tau_k}, & (1 \le u \le 20) \\ \frac{\omega \tau_{u-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{k=1}^{\lambda} \tau_k}, & (20+1 \le u \le 20+\lambda) \end{cases} \qquad (3)$$

where ω is the weight factor and τ_k the k-th tier correlation factor that reflects the sequence order correlation between all the k-th most contiguous residues as formulated by:

$$\tau_k = \frac{1}{L-k} \sum_{i=1}^{L-k} \frac{1}{3} \{[H_1(R_{i+k}) - H_1(R_i)]^2 + [H_2(R_{i+k}) - H_2(R_i)]^2 + [M(R_{i+k}) - M(R_i)]^2\} \quad (4)$$

where $H_1(R_i)$, $H_2(R_i)$ and $M(R_i)$ are respectively the normalized hydrophobicity value, hydrophilicity value and the side chain mass for amino acid residue R_i; while $H_1(R_{i+k})$ $H_2(R_{i+k})$ $M(R_{i+k})$ are those for amino acid residue R_{i+k}.

In the amphiphilic mode (Type 2), the given protein **P** can be represented as:

$$\boldsymbol{P} = [p_1, p_2, \ldots, p_{20}, p_{20+1}, \ldots, p_{20+2\lambda}]^T, \quad (\lambda < L) \qquad (5)$$

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

where the 20 + 2λ components are given by:

$$p_v = \begin{cases} \frac{f_v}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{2\lambda} \tau_j}, & (1 \le v \le 20) \\ \frac{\omega \tau_v}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{2\lambda} \tau_j}, & (20+1 \le v \le 20+2\lambda) \end{cases} \quad (6)$$

And τ_j can be represented as follows:

$$\tau_m = \frac{1}{L-l} \sum_{i=1}^{L-l} h^1(R_i).h^1(R_{i+l}) \quad (7)$$

$$\tau_n = \frac{1}{L-l} \sum_{i=1}^{L-l} h^2(R_i).h^2(R_{i+l}) \quad (8)$$

where m = 1, 3, 5, …, (2λ-1); n = 2, 4, 6, …, 2λ and l = 1, 2, 3, 4, …λ (λ< L). The values of $h^1(R_i)$ and $h^2(R_i)$ represent the normalized hydrophobicity and hydrophilicity properties of amino acid residue R_i; while $h^1(R_{i+l})$ and $h^2(R_{i+l})$ are those for amino acid residue R_{i+l}.

Multiple combination of λ in [5, 10, 15, 20, 30, 40] and ω in [0.05, 0.10, 0.30, 0.50, 0.70] were explored using the PseAAC web server available at http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/ for prediction.

2.3. Support vector machines

To discriminate AFPs from non-AFPs using pseAAC features, a classifier popular for solving biological challenges related to prediction, classification and regression [14], the support vector machines (SVMs) was used. SVM is a supervised machine-learning tool based on the structural risk minimization principle of statistics learning theory. It looks for an optimal hyperplane which maximizes the distance between the hyperplane and the nearest samples from each of the two

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

classes. Mathematically, a training vector $x_i \in R^n$, and class values $y_i \in \{-1, 1\}$, $i = 1, \ldots, N$ are used to solve the problems using the following equation:

$$Minimize\ \frac{1}{2} w^T \cdot w + C \sum_{i=1}^{N} \xi_i \qquad (9)$$

$$Subject\ to\ y_i(w^T \cdot x_i + b) \geq 1 - \xi_i\ and\ \xi_i \geq 0 \qquad (10)$$

where w is the normal vector perpendicular to the hyperplane and ξ_i are slake variables for permitting misclassifications. Balancing the trade-off between the margin and the training error is done using C (> 0), the penalty parameter [16]. The user can choose and optimize number of parameters and kernels (e.g. linear, polynomial, radial basis function and sigmoidal) or any user-defined kernel. In this study, we selected radial basis function for AFP prediction with grid searching of influencing parameters, C in [5 – 50] and γ in [0.00001 – 0.1] and developed models using SVM^{light} Version 6.02 package available at http://svmlight.joachims.org/.

2.4. Performance Evaluation

Models of AFPs and non-AFPs were generated in various combinations of pseAAC and SVM parameters on randomly selected 300 positive and 300 negative examples (training dataset) followed by ten-fold cross-validation [17] and performance analysis using the following mathematical formula for sensitivity (recall), specificity, accuracy and Matthew's correlation coefficient (MCC). Since the number of AFPs i.e. 481 positive examples were much smaller than 9423 negative examples, we investigated for bias in identification due to selection process, by keeping the number of positive examples constant to 300. Briefly, we selected 300 negative

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

examples, three times randomly for model generation and evaluated the prediction using the above mentioned mathematical formulae after ten-fold cross-validation. Then, we explored the influence of number of negative examples on the AFP prediction also. We did this by including 900 negative examples instead of the original ratio 1:1 for development of models followed by performance assessment using same evaluation parameters as mentioned above after ten-fold cross-validation. Once we gained insights into the selection bias and influence of negative examples on prediction, the best performing models were selected for prediction on independent test dataset. This comprised of examples in section 2.1 not included in training, i.e., 181 AFPs and 8293 non-AFPs. Subsequently, their performance was evaluated and compared with existing AFP prediction approaches. The mathematical formula for evaluation parameters:

$$Sensitivity = TP/(TP + FN) \quad (11)$$

$$Specificity = TN/(TN + FP) \quad (12)$$

$$Accuracy = (TP + TN)/(TP + FN + TN + FP) \quad (13)$$

$$MCC = ((TP \times TN) - (FP \times FN))/\sqrt{((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN))} \quad (14)$$

where, TP (true positives): Proteins correctly predicted as AFPs, FP (false positives): Proteins incorrectly predicted as AFPs, TN (true negatives): Proteins correctly predicted as non-AFPs and FN (false negatives): Proteins incorrectly predicted as non-AFPs.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

3. Results and Discussion

3.1. Optimal parameters for feature extraction

Prediction of protein attributes requires selection of key representative parameters for improved quality. In this study, various combination of λ, ω and modes of pseudo amino acid composition and SVMs were generated for predictor development by grid searching on 300 positive and 300 negative examples. Since the number of positive examples was limited, we considered the same set of 300 antifreeze proteins as positive examples for all training purposes. Figure 1 shows the performance evaluation (MCC and Accuracy) for various λ, ω and modes of pseAAC. MCC and Accuracy were highest for ω = 0.05. Further, it can be seen that with increase in ω, the quality of prediction is compromised. The best performing model was achieved for ω = 0.05 and λ = 5; in the amphiphilic mode, closely followed by performance with ω = 0.05 and λ = 10; in the amphiphilic mode. Optimal features i.e. ω = 0.05 and lower values of λ in the amphiphilic mode (Type 2) were reserved for development of the predictor.

3.2. Bias during selection of negative examples

To understand if selecting 300 examples from a 9493 proteins biased the prediction, we performed random selection of negative examples three times on non-overlapping datasets. As shown in Table 1 from the prediction performance, the average of mathematical parameters of evaluation mentioned in Section 2.4 for all the generated models suggested the best values for λ = 10. The highest prediction accuracy was 89.69 (0.706) %, MCC = 0.800 (0.0095), sensitivity = 88.89 (1.835) % and specificity = 91.00 (0.330) %. The standard deviation observed in the performance evaluation parameters over prediction performed three times was negligible.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

Therefore, it can be said that the bias associated with selection of negative examples was minimally influential, if at all.

3.3. Bias on account of number of negative examples

After exploring the selection bias, we investigated into the influence of the number of negative examples on prediction. We included 900 non-AFPs (three times higher than the number of positive examples = 300) during the prediction. This showed increase in accuracy from 89.69 % to 91.25 %; specificity from 91.00 % to 96.78 %; decrease in MCC from 0.800 to 0.762; and compromised sensitivity from 88.89 % to 77.67 %. However, if analyzed in relation with positive and negative examples used in the ratio 1:1, the performance of the predictor models was better with balanced dataset, as can be seen in Section 3.2. Therefore, 1:1 ratio of examples in dataset was maintained throughout the predictor development process.

3.4. Performance evaluation and comparison with AFP-Pred and iAFP

After selection of optimal parameters and search for possible bias as mentioned in above sections, the best model was selected for AFP predictor development and named as AFP-PseAAC. Precisely, this included pseAAC parameters $\omega = 0.05$ and $\lambda = 10$ in amphiphilic mode; SVM parameters C = 25 and $\gamma = 0.0005$ and 1:1 ratio of training dataset.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

Upon screening of protein examples excluded from training of the predictor (independent test dataset) an accuracy of 84.75%, sensitivity of 86.19 % and specificity of 84.72 % was obtained as shown in Table 2. These findings encouraged us to compare the performance of our method with that of other previously published methods, if their implementations were readily accessible.

AFP-PseAAC was assessed in relation with AFP-Pred and iAFP to gain insights into its relative prediction power on the independent test dataset. While AFP-PseAAC achieved an accuracy and specificity of (84.75 %, 84.72 %); these values seen for AFP-Pred and iAFP were (69.86 %, 69.67 %) and (95.46%, 97.38 %) respectively. The high accuracy and specificity achieved for iAFP, i.e., above 95%, was notably accompanied with extremely low sensitivity, i.e., 7.18%. On the contrary, AFP-PseAAC showed sensitivity of 86.19%, followed by AFP-Pred which reached close to 78.45%. Clearly, AFP-PseAAC outperformed AFP-Pred and iAFP.

Additionally, to quantify the relative performance of AFP-PseAAC we applied the Youden's Index (J = Sensitivity + Specificity -1) [20] used for gaining insights into the relative performance of tests in general. Youden's index gives the probability of an informed decision and is advantageous as it offers comparison of AFP prediction quality by means of a single informative parameter. AFP-PseAAC showed J = 0.71, followed by AFP-Pred with J = 0.48 and then iAFP with J = 0.05, as shown in Table 2. Since J value for AFP-PseAAC is much higher

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

than the two approaches, it can be suggested that introduction of sequence order effect using Chou's pseAAC enhances AFP identification by SVMs.

4. Conclusion

Diversity renders difficulty towards accurate identification of antifreeze proteins. Earlier work reported sequence based solutions but sequence order effect which is known to improve prediction of protein attributes remained to be explored. Since Chou's pseudo amino acid composition features represents diverse proteins as discrete models yet without entirely losing the sequence order information, we hoped to develop a novel effective approach AFP-PseAAC for prediction of antifreeze proteins using pseAAC and SVMs. The overall performance shown by AFP-PseAAC was better than previous AFP predictors. Interested users may find relevant details for using this approach at https://sites.google.com/site/sukantamondal/software. We anticipate this predictor to facilitate faster and broader applications of AFPs in biotechnology.

Acknowledgement

The authors thank BITS-Pilani, K.K.Birla Goa Campus, for providing the necessary support towards conducting of this research.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

References

1. Se-Kwon, K., 2013. Marine Proteins and Peptides: Biological Activities and Applications, first ed. John Wiley & Sons, United Kingdom.
2. Davies, P.L., Hew, C.L., 1990. Biochemistry of fish antifreeze proteins. FASEB J. 4, 2460–2468.
3. Lewitt, J., 1980. Responses of Plants to Environmental Stresses, vol. 1. Academic Press, New York.
4. Sformo, T., Kohl, F., McIntyre, J., Kerr, P., Duman, J.G., Barnes, B.M., 2009. Simultaneous freeze tolerance and avoidance in individual fungus gnats, Exechia nugatoria. J. Comp. Physiol. B. 179, 897–902.
5. Kandaswamy, K.K., Chou, K.C., Martinez, T., Möller, S., Suganthan, P.N., Sridharan, S., Pugalenthi, G., 2011. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. J.Theor.Biol. 270, 56–62.
6. Anand, A., Pugalenthi, G., Suganthan, P.N., 2008. Predicting protein structural class by SVM with class-wise optimized features and decision probabilities. J. Theor. Biol. 253, 375–380.
7. Cai, Y.D., Ricardo, P.W., Jen, C.H., Chou, K.C., 2004. Application of SVM to predict membrane protein types. J. Theor. Biol. 226, 373–376.
8. Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct. Funct. Genet. 43, 246–255.
9. Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.
10. Chou, K.C., Cai, Y.D., 2005. Prediction of membrane protein types by incorporating amphipathic effects. J. Chem. Inf. Modeling. 45, 407–413.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

11. Chou, K.C., Shen, H.B., 2009. Recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 1, 63–92.

12. Huang, R.B., Du, Q.S., Wei, Y.T., Pang, Z.W., Wei, H., Chou, K.C., 2009. Physics and chemistry-driven artificial neural network for predicting bioactivity of peptides and proteins and their design. J. Theor. Biol. 256, 428–435.

13. Xiaowei, Z., Zhiqiang, M., and Minghao, Y., 2012. Using Support Vector Machine and Evolutionary Profiles to Predict Antifreeze Protein Sequences. Int. J. Mol. Sci. 13, 2196-2207.

14. Yu, C.S., Lu, C.H., 2011. Identification of Antifreeze Proteins and Their Functional Residues by Support Vector Machine and Genetic Algorithms based on n-Peptide Compositions. PLoS ONE. 6, e20445.

15. Chou, K.C., 2009. Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology. Curr. Proteomics. 6, 262-274.

16. Joachims, T., Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

17. Schaffer, C., 1993. Selecting a classification method by cross-validation. Mach. Learn. 13, 135-143.

18. Sonnhammer, E.L., Eddy,S.R., Durbin,R.,1997. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 28, 405–420.

19. Altschul, S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., Lipman,D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

20. Youden, Y.W., 1950. Index for rating diagnostic tests. Cancer. 3, 32-35.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

Table 1

Influence of negative examples during AFP prediction

Dataset	λ	MCC	Accuracy (%)	Sensitivity (%)	Specificity (%)
300 AFPs and 300 non-AFPs[a]	05	0.797 (0.0035)	89.61 (0.344)	88.45 (1.072)	91.00 (0.670)
	10	0.800 (0.0095)	89.69 (0.706)	88.89 (1.835)	91.00 (0.330)
	15	0.796 (0.0148)	89.61 (0.919)	89.22 (1.575)	90.52 (0.790)
	20	0.786 (0.0078)	89.17 (0.440)	88.89 (1.018)	90.33 (0.665)
300 AFPs and 900 non-AFPs	05	0.755	90.92	78.67	96.0
	10	0.762	91.25	77.67	96.78
	15	0.775	91.83	76.67	97.11
	20	0.773	91.83	77.33	97.22

[a]Format: Average evaluation parameter (Standard deviation) upon random selection of negative examples three times.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

Table 2

Performance of AFP-PseAAC compared with AFP-Pred [5] and iAFP [14] on independent test dataset

Predictor	Accuracy (%)	Sensitivity (%)	Specificity (%)	Youden's Index[a]
AFP-PseAAC	84.75	86.19	84.72	0.71
AFP-Pred	69.86	78.45	69.67	0.48
iAFP	95.46	7.18	97.38	0.05

[a]Youden's Index [20] = Sensitivity + Specificity - 1

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

Figure legend

Fig. 1

Exploring optimal pseudo amino acid composition parameters (Type, λ and ω) on the training dataset for the development of the AFP predictor.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014

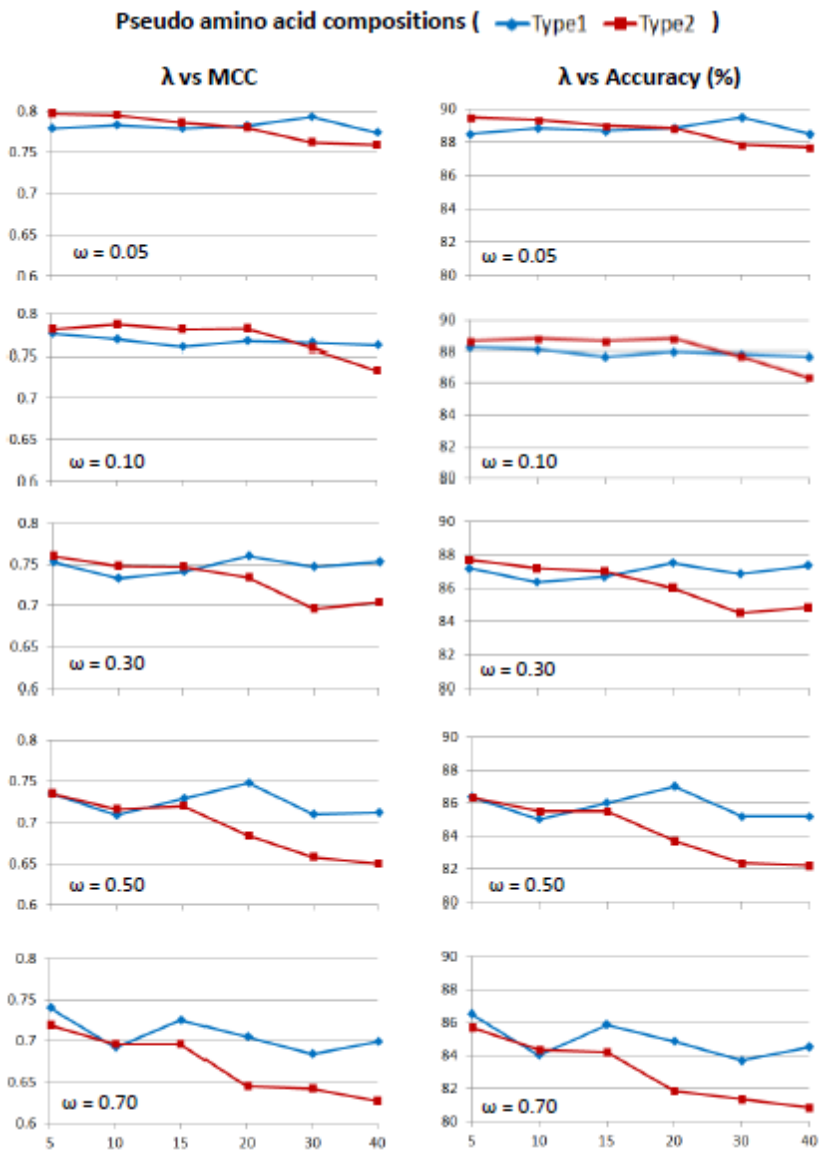

Figure 1

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.224v1 | CC-BY 4.0 Open Access | received: 29 Jan 2014, published: 29 Jan 2014