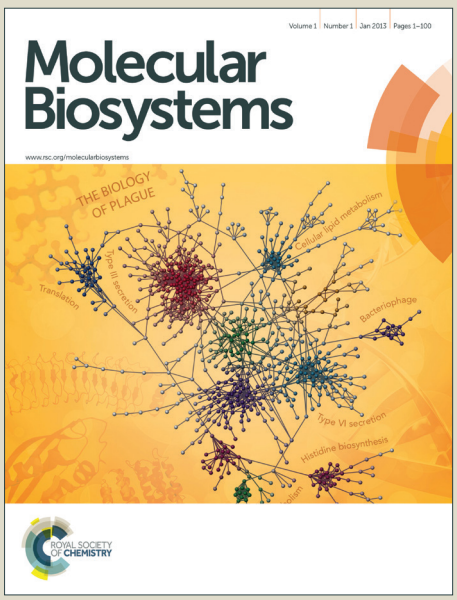
Molecular BioSystems

Accepted Manuscript



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Molecular Biosystems Molecular BioSystems Dynamic Article Links ►

Cite this: DOI:

www.rsc.org/xxxxxx

PAPER

Predicting subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Pan-Pan Zhu,[a] Wen-Chao Li,[a] Zhe-Jin Zhong,[a] En-Ze Deng,[a] Hui Ding,*[a] Wei Chen*[b] and Hao Lin*[a]

[a] *Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China. E-mail: hlin@uestc.edu.cn* (Hao Lin)*; hding@uestc.edu.cn* (Hui Ding)*; Fax: 28 8320 8238; Tel: 28 8320 2351;*

[b] *Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China. E-mail: chenweiimu@gmail.com;*

Mycobacterium tuberculosis is a bacterium that causes tuberculosis, one of the most prevalent infectious diseases. Predicating the subcellular localization of mycobacterial proteins in this bacterium may provide vital clues for the prediction of protein function as well as for drug discovery and design. Therefore, a computational method that can predict the subcellular localization of mycobacterial proteins with high precision is highly desirable. We propose a computational method to predict the subcellular localization of mycobacterial proteins. An objective and strict benchmark dataset was constructed after collecting 272 non-redundant proteins from the universal protein resource (UniProt database). Subsequently, a novel feature selection strategy based on binomial distribution was used to optimize the feature vector. Finally, a subset containing 219 chosen tripeptide features was imported into a support vector machine-based method to estimate the performance of the dataset in accurately and sensitively identifying these proteins. We found that the proposed method gave a maximum overall accuracy of 89.71% with an average accuracy of 81.12% in the jackknife cross-validation. The results indicate that our prediction method gave an efficient and powerful performance when compared with other published methods. We made the proposed method available on a purpose built Web server called MycoSub that is freely accessible at http://lin.uestc.edu.cn/server/MycoSub. We anticipate that MycoSub will become a useful tool for studying the functions of mycobacterial proteins and for designing and developing anti-mycobacterium drugs.

1. Introduction

Mycobacterium tuberculosis is an extraordinarily successful pathogenic bacterium that causes tuberculosis. It has been reported that approximately 9 million people are infected annually[1, 2]. *Mycobacterium tuberculosis* attacks the lung, but it can also infect other organs and systems. It can change its morphology, colony characteristics, virulence, resistance, and immunogenicity. Because of its unique cell wall, studies on the subcellular localization of mycobacterial proteins may provide useful insights about their

Molecular BioSystems Accepted Manuscript

functions and may help in understanding the intricate pathways that regulate the biological processes at the cellular level. Therefore, it has been suggested that the accurate identification of the subcellular localization of mycobacterial proteins may be very important for elucidating protein functions involved in various cellular processes [3].

Wet experiments are an objective approach that could be used to recognize the subcellular localization of mycobacterial proteins; however, biochemistry-based methods are often time consuming and costly. Therefore, the development of computational approaches has attracted a lot of attention. Many methods such as the support vector machine (SVM) [4-7], artificial neural network [8], K-nearest neighbor [9-12], Bayesian classifier [13, 14], increment of diversity [15], covariant discriminant [16, 17] and ensemble learning [18-20] algorithms have been developed to predict the subcellular localization of proteins. Features such as amino acid composition [21], pseudo amino acid composition [22], physico-chemical properties [23], N-peptide composition [24, 25], pseudo-average chemical shift [26], and motifs [27, 28] have been proposed to represent protein sequences.

Although previous studies on the subcellular localization of eukaryotic and prokaryotic proteins have produced encouraging results, few studies have focused on predicting the subcellular localization of mycobacterial proteins. The recent accumulation of proteomics data has triggered the development of computational methods to predict the subcellular localization of mycobacterial proteins. Rashid et al. [29] used SVM-based method to predict 852 mycobacterial proteins and obtained a maximum overall accuracy of 86.8% in a five-fold cross-validation. However, the sequence identity of the database was so high that the performance of the proposed methods may have been overestimated. Lin et al. [30] constructed a non-redundant dataset based on a 852 dataset, which had a sequence identity of only 30%. In a jackknife cross-validation, they reported an overall accuracy of 90% and an average accuracy of 73.9%. Subsequently, Li et al. [28] improved the overall accuracy of their dataset to 95.5% with an average accuracy of 76.5% in a jackknife cross-validation. Fan et al. [26] developed a pseudo-average chemical shift method to predict the subcellular localization of mycobacterial proteins. Although these methods can achieve high overall accuracies, the average accuracy needs to be improved.

Here, we proposed a SVM-based model to identify the subcellular localization of mycobacterial proteins. A binomial distribution method was used to select informative tripeptides and the jackknife cross-validation showed that our model obtained an overall accuracy of 89.71% with an average accuracy of 81.12% in a benchmark dataset containing 272 mycobacterial proteins with sequence identities of no more than 25%. We also compared the performance of our method with the performances of previously published methods. In addition to predicting the subcellular localization of the proteins, our method provided useful information about local sequences, which may have broad applications in areas from protein function to drug design research.

The current study was devoted to enhance the prediction po wer and quality in Predicting the subcellular localization of my cobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition.

According to recent publications [31-37] and a comprehensive review [38], the rest of the papers are organized as follows: (i) construct a valid benchmark dataset to train and test the predictor; (ii) formulate the samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) select a powerful machine learning method to operate the prediction; (iv) perform cross-validation tests to objectively evaluate the anticipated prediction accuracy of the predictor; (v) provide a web-server for the prediction method.

2. Methods

2. 1. Dataset

The data that we designed for predicting subcellular localization of mycobacterial proteins were extracted from the Universal Protein Resource (UniProt) [39]. To guarantee a high-quality and well-defined dataset, we selected the protein sequences as follows: (i) protein sequences that have been reviewed and annotated by experts were chosen; (ii) protein sequences that were fragments of other proteins were removed; (iii) sequences for which the existence of the protein was uncertain or that were predicted were eliminated; and (iv) sequences that were inferred from homologous proteins were eliminated. Generally, if a designed dataset contains highly similar sequences, misleading results with overestimated accuracies will be obtained and the generalization ability of the proposed model will be reduced. To avoid such overestimations, the PISCES program [40] was employed with 25% as the sequence identity cutoff to remove redundant sequences.

As is well known, proteins may simultaneously exist at, or move between, two or more different subcellular localizations. Some web servers such as iLoc-Euk [41], iLoc-Hum [42], iLoc-Plant [43], iLoc-Gpos [44], iLoc-Gneg [45], and iLoc-Virus [46] have been developed to cope with

Molecular BioSystems Accepted Manuscript

the multiple localization problems. However, in this study we didn't consider the case of multiplex proteins because the number of multiplex proteins in the existing mycobacterial protein database is not large enough to construct a statistically meaningful benchmark dataset for studying the case of multiple localizations. As a result, we obtained 272 mycobacterial proteins sequences which are distributed among 3 subcellular localizations (Figure 1) and can be formulated by

$$S=S_1 \cup S_2 \cup S_3 \quad (1)$$

where S_1 contains 153 cytoplasmic proteins, S_2 contains 18 secretory proteins, and S_3 contains 101 membrane proteins. The 272 sequences can be downloaded from the MycoSub.

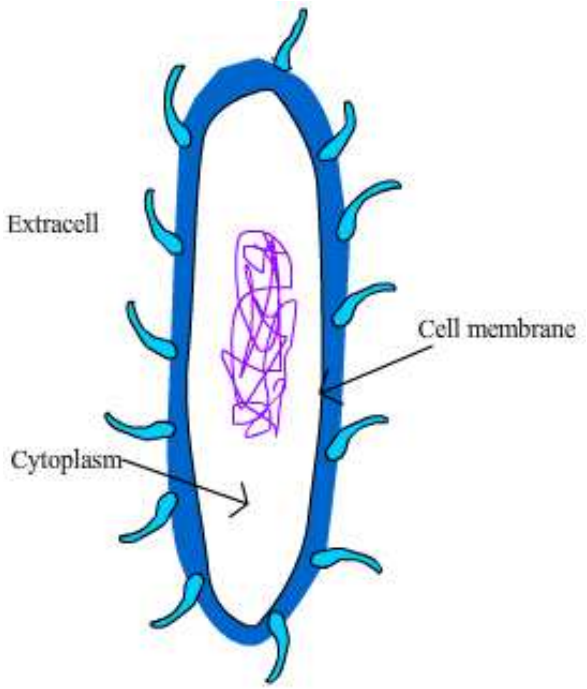


Figure 1. Schematic illustration to show the three subcellular localizations of *M. tuberculosis*: cell membrane, cytoplasm and extracell.

2. 2. Tripeptide compositions

After building the benchmark dataset, we extracted a set of informative parameters that was used to develop the predictor for the subcellular localization of mycobacterial proteins. One of the most important criteria was to formulate an effective mathematical expression that truly reflected the correlation between the intrinsic features of the sequences and the protein types to be predicted. Wang et al. [47] described a protein sequence, which is generally made from various combinations of 20 amino acids. To classify the proteins sequences, the n-gram features were extracted and used as the input signals for the classifier. The *n*-gram features have been used as modulators of biological function [48], to predict plausible structures for oligopeptides, and for denovo protein design [49]. In this work, the tripeptide compositions (3-gram features) were used to represent a protein sequence by an 8000 dimensional vector as:

$$\boldsymbol{F}_{8000} = [\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_i, ..., \boldsymbol{f}_{8000}]^{\mathrm{T}} \quad (2)$$

where the symbol **T** denotes the transposition of a vector and f_i is the frequency of the *i*-th tripeptide in a protein sequence, which can be calculated as:

$$f_i = n_i \Big/ \sum_{i=1}^{8000} n_i = n_i / (L-2) \quad (3)$$

where n_i and L denote the number of the *i*-th tripeptide and the length of the protein sequence, respectively.

2.3. Feature selection

Inclusion of redundant and noisy information can result in poor predictive results and is time costly. To economize on run-time and computational resources, feature selection is a smart technique that can be used to deal with this problem. Currently, many techniques have been proposed to optimize feature sets. They include principal component analysis [50, 51], diffusion maps [52], minimal-redundancy-maximal-relevance [53, 54], analysis of variance [55, 56], local linear discriminant analysis [57], and geometry preserving projections [58].

In the present study, a binomial distribution technique was applied to pick out informative tripeptides [59]. According to Eq. (2), the three types of mycobacterial proteins in the benchmark dataset can contain up to 8000 types of tripeptides. That a particular type of tripeptides appears in one type of protein may be a stochastic event. Thus, it is necessary to judge whether the occurrence of a particular tripeptide in one type of protein is a stochastic event or not.

We initially defined the prior probability q_j as:

$$q_j = m_j / M \quad (4)$$

where m_j is the number of tripeptides that appeared in the *j*-th type of protein and M is the total occurrence frequency of all tripeptides in the benchmark dataset.

The probability of the *i*-th tripeptide occurring in the *j*-th type of protein can be defined as:

Molecular BioSystems Accepted Manuscript

$$p(n_{ij}) = \sum_{m=n_{ij}}^{N_i} \frac{N_i!}{m!(N_i - m)!} q_j^m (1 - q_j)^{N_i - m} \quad (5)$$

where N_i represents the total number of the i-th tripeptides in the benchmark dataset, n_{ij} represents the number of occurrences of the i-th tripeptide in the j-th type of protein, and the sum is taken from n_{ij} to N_i.

If $p(n_{ij})$ is a small value, it indicates the observation of the i-th tripeptide in the j-th type of protein is not a random event. The *CL* of the i-th tripeptide in the j-th type of protein is defined as:

$$CL_{ij} = 1 - p(n_{ij}) \quad (6)$$

The *CLs* of the 8000 tripeptides in the j-th type of protein can be calculated based on Eqs. (4–6). We ranked the 8000 tripeptides in descending order according to their *CLs*. Because three types of proteins were considered in this study, each of the tripeptides had three different rank indexes denoted as r^m, r^s, r^c (r^m, r^s, $r^c \in (1, 8000)$). Thus, we defined the rank of these tripeptides in the benchmark dataset as:

$$\boldsymbol{r_i} = \min\{\boldsymbol{r_i^m}, \boldsymbol{r_i^s}, \boldsymbol{r_i^c}\} \quad (7)$$

where r_i represents the rank of the i-th tripeptide in the benchmark dataset, the superscripts 'm', 's', and 'c' indicate membrane proteins, secretory proteins, and cytoplasmic proteins, respectively.

Subsequently, incremental feature selection was used to determine the optimal number of features. First, the feature subset started from features with the lowest r value in the ranked feature set. Then, a new feature subset was produced when the features with the second lowest r value were added. This two-step process was repeated from the lower r values to the higher r values until all the candidate features were added. The resulting feature subset with k tripeptides can be described as:

$$\boldsymbol{F_k} = \left[\boldsymbol{f_1'}, \boldsymbol{f_2'}, \ldots, \boldsymbol{f_i'}, \ldots, \boldsymbol{f_k'}\right]^{\mathrm{T}} \quad (8)$$

For each of the feature sets, a SVM-based model was constructed and a 5-fold cross-validation test was used to investigate its accuracy. The optimal feature set can be considered to be obtained when the overall accuracy was the maximum. The final predictor was built using the k optimal features.

2.4. Support vector machine

The SVM approach is a supervised machine learning method based on statistical learning theory, which has been used successfully in many bioinformatics applications [36, 40, 60-63]. The basic idea of SVM is to map samples into a high-dimension Hilbert space and to seek a separating hyperplane in this space. To handle multi-class problems, "one-versus-one" and "one-versus-rest" strategies are generally applied to extend the traditional SVM. In this study, the one-versus-one strategy was used. The radial basis function was chosen as the kernel function. The LibSVM2.83 software [60] was used to implement SVM.

2.5. Criteria definitions

Jackknife cross-validation always yields a unique result for a given benchmark dataset and has been widely and increasingly adopted [31, 61, 62, 64-67]. We used the jackknife cross-validation to evaluate the performance of our method. In the jackknife cross-validation, each protein in the dataset was singled out in turn as an independent test sample and all the rule parameters were calculated based on the remaining proteins excluding the one being identified. Furthermore, to reduce the computational time, the 5-fold cross-validation was used to select the C and γ parameters in the SVM.

To provide a more intuitive and easier-to-understand method to measure the prediction quality, we used the following five parameters: sensitivity (*Sn*), specificity (*Sp*), Matthew's correlation coefficient (*MCC*), overall accuracy (*OA*), and average accuracy (*AA*), which were defined as follows:

$$Sn(i) = 1 - \frac{N_-^+(i)}{N^+(i)} \quad (9)$$

$$Sp(i) = 1 - \frac{N_+^-(i)}{N^-(i)} \quad (10)$$

$$MCC(i) = \frac{1 - \left(\frac{N_-^+(i)}{N^+(i)} + \frac{N_+^-(i)}{N^-(i)}\right)}{\sqrt{\left(1 + \frac{N_+^-(i) - N_-^+(i)}{N^+(i)}\right)\left(1 + \frac{N_-^+(i) - N_+^-(i)}{N^-(i)}\right)}} \quad (11)$$

$$OA = \frac{1}{\delta}\sum_{i=1}^{\zeta}\left[N^+(i) - N_-^+(i)\right] \quad (12)$$

$$AA = \frac{\sum_{i=1}^{\zeta}\left(1 - \frac{N_-^+(i)}{N^+(i)}\right)}{\zeta} \quad (13)$$

where $N^+(i)$ is the total number of the investigated mycobacterial proteins samples in the subset S_i, $N_-^+(i)$ is the number of mycobacterial proteins samples in S_i that were incorrectly predicted belonging to the other subsets, $N^-(i)$ is the total number of the mycobacterial proteins samples in all of the other subsets, $N_+^-(i)$ is the number of the mycobacterial proteins samples that were

Molecular BioSystems Accepted Manuscript

incorrectly predicted belonging to S_i, ζ is the number of subsets (Eq.(1)) and δ is the number of the total samples in S.

It should be noted that the set of metrics is valid only for the single-label systems. For the multi-label systems whose existence has become more frequent in system biology [41, 42] and system medicine [68, 69], a completely different set of metrics as defined in [70] is needed.

3. Results and discussion

3.1. Predictive performance

The subcellular localization of mycobacterial proteins predicted using SVM can help economize the resources and computational time required. Following previous studies [64, 71-73], the regularization (C) and kernel (γ) parameters of the SVM were optimized by a grid search with 5-fold cross-validation.

Generally, the high dimension features can lead not only to an over-fitting problem but can also cause information redundancy or noise, which would result in low capability in the generalization of a predictor and poor prediction in the cross-validation. To overcome the high-dimension problems, we selected informative sequence features using binomial distribution. If we selected tripeptides with high *CLs*, the results were robust and credible. The selected tripeptides were also informative; however, the number of these tripeptides was too small to reflect enough information about the proteins. For example, the overall accuracy was only 71.3% when 27 tripeptides were used to predict the subcellular localization of three types of mycobacterial proteins in 5-fold cross-validations. In general, higher numbers of tripeptides can better represent the proteins. However, when too many tripeptides were used, the model would become overfitted, which reduced its robustness and predictive accuracy. For instance, we found that 6281 tripeptides produced an overall accuracy of only 68.4% in the 5-fold cross-validations. Therefore, to build a robust model with high accuracy, it is very important to choose the appropriate number of features.

Because our dataset contained 272 mycobacterial protein sequences, to avoid the overtraining problem we set the maximum number of selected tripeptide features to 272. By selecting the appropriate *CL*, we found that a feature subset containing 219 tripeptides achieved a maximum overall accuracy of 89.71% and an average accuracy of 81.12%. The sensitivities were 88.12%, 61.11%, and 94.12% for membrane proteins, secretory proteins, and cytoplasmic proteins, respectively. With the detailed information is shown in table 1. Furthermore, we also calculated the *OA* achieved by completely random guess (CRG). Obviously, the *OA* achieved by CRG is 50.00%. If considering the weight or prior probability, the *OA* is $[101 \times (101/272)+18 \times (18/272)+153 \times (153/272)]/272 = 45.87\%$. These results demonstrate that our method is superior to random guess and suggest that binomial distribution is a powerful technique for selecting optimal features.

Table 1 The jackknife test results on the dataset M272.

Location	Sn	Sp	MCC
Membrane	88.11	94.15	0.826
Secretory	61.11	1.00	0.771
Cytoplasm	94.12	84.87	0.798
OA(%)	89.71		
AA(%)	81.12		

3.2. Comparison with other methods

Two published datasets, M330 and M638, have been used previously to investigate the performance of prediction methods for the subcellular localization of mycobacterial proteins. Thus, we also tested the predictive capability of our method using these two datasets.

For the M330 dataset, an overall accuracy of 90% and an average accuracy of 73.9% has been reported using pseudo amino acid composition and reduced amino acids [30]. Li et al. [28] improved the overall accuracy to 95.5% and the average accuracy to 76.5%. Fan et al. [26] obtained an overall accuracy of 93.33% with an average accuracy of 78.27% using pseudo-average chemical shift. We found that although the overall accuracy of our method was lower than that of the other methods, the average accuracy of our method was much higher than that of the other methods (Table 2). Notably, our method correctly recognized 75% of the secretory proteins, which is ~30% higher than was reported for the other existing methods.

In the M330 dataset, the numbers of integral membrane proteins (176) and cytoplasmic proteins (111) were much more than the numbers of membrane-attached proteins by lipid anchor proteins (27) and secretory proteins (16). Thus, the overall accuracies of the predictions were influenced significantly by the success rate for the predictions of integral membrane proteins and cytoplasmic proteins. For example, if all integral membrane and cytoplasmic proteins were predicted correctly while the other two types of proteins were all incorrectly predicted, the overall accuracy would be 86.97%, but the average accuracy would be only 50%. Thus, our method is more practical than the other methods tested.

Molecular BioSystems Accepted Manuscript

Table 2 Comparison of our proposed method with other methods on the M330 dataset.

Predictor	Sn (%)				OA (%)	AA (%)
	Integral membrane	Attached to the membrane	Secretory	Cytoplasm		
Our method	90.91	77.78	75.00	86.49	87.56	82.54
Fan et al. [26]	96.63	80.77	40.00	95.69	93.33	78.27
Li et al. [28]	100.00	72.00	46.70	87.30	95.50	76.50
Lin et al. [30]	93.80	74.10	31.30	96.40	90.00	73.90

The comparative results with the M638 dataset are listed in Table 3. Fan et al. [26] obtained an overall accuracy of 87.77% with an average accuracy of 70.37%, while we obtained an overall accuracy of 86.21% with an average accuracy of 73.25% using our method. Although the overall accuracy of our method was slightly lower than that of Fan et al. [26] (about 1.5%), the average accuracy of our method was higher (about 2.9%). In particular, the accuracies for the secretory proteins and attached to the membrane proteins were about 10% higher than the accuracies reported by Fan et al. [26]. These results show that our method was superior to the other existing methods tested, especially for predicting the secretory proteins and attached to the membrane proteins.

Table 3 Comparison of our predictor method with other methods with the M638 dataset.

Location	Our method			Fan et al.[26]		
	Sn	Sp	MCC	Sn	Sp	MCC
Integral membrane	88.85	87.65	0.765	90.45	87.96	0.784
Attached to the membrane	73.33	99.84	0.831	63.33	99.18	0.695
Secretory	41.38	99.84	0.608	34.48	99.34	0.481
Cytoplasm	89.43	87.67	0.765	93.21	91.96	0.847
OA(%)	86.21			87.77		
AA(%)	73.25			70.37		

4. Conclusions

In this work, we developed a promising method to predict the subcellular localization of mycobacterial proteins. A binomial distribution-based feature selection technique was developed to select over-represented tripeptides. In the jackknife test, our proposed model achieved an overall accuracy of 89.71% with an average accuracy of 81.12% for the subcellular localization of mycobacterial proteins. Although our results are encouraging, there is clearly considerable room for improvement. The accuracy of the subcellular localization of mycobacterial proteins was higher using our method compared with the accuracies reported using previous methods, indicating that our method is a much more powerful one. Our method is available on a purpose-built Web server called MycoSub (http://lin.uestc.edu.cn/server/MycoSub).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

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

Acknowledgments

This work was supported by the National Nature Scientific Foundation of China (Nos. 61202256, 61301260, and 61100092), the Nature Scientific Foundation of Hebei Province, China (No.C2013209105), Funding of top-notch talents of Hebei Education Department (BJ2014028) and the Fundamental Research Funds for the Central Universities, China (Nos. ZYGX2012J113, ZYGX2013J102).

References

1. M. Raviglione, B. Marais, K. Floyd, K. Lonnroth, H. Getahun,

Molecular BioSystems Accepted Manuscript

G. B. Migliori, A. D. Harries, P. Nunn, C. Lienhardt, S. Graham, J. Chakaya, K. Weyer, S. Cole, S. H. Kaufmann and A. Zumla, Lancet, 2012, 379, 1902-1913.
2. H. S. Whitworth, M. Scott, D. W. Connell, B. Donges and A. Lalvani, Methods, 2013, 61, 52-62.
3. K. C. Chou and H. B. Shen, Analytical biochemistry, 2007, 370, 1-16.
4. K. C. Chou and Y. D. Cai, J Biol Chem, 2002, 277, 45765-45769.
5. S. Hua and Z. Sun, Bioinformatics, 2001, 17, 721-728.
6. A. C. Lorena and A. C. de Carvalho, Computers in biology and medicine, 2007, 37, 115-125.
7. H. B. Shen, J. Yang and K. C. Chou, Amino acids, 2007, 33, 57-67.
8. A. Reinhardt and T. Hubbard, Nucleic acids research, 1998, 26, 2230-2236.
9. K. C. Chou and H. B. Shen, Journal of proteome research, 2006, 5, 1888-1897.
10. K. C. Chou and H. B. Shen, Biochemical and biophysical research communications, 2006, 347, 150-157.
11. P. Jia, Z. Qian, Z. Zeng, Y. Cai and Y. Li, Biochemical and biophysical research communications, 2007, 357, 366-370.
12. J. Y. Shi, S. W. Zhang, Q. Pan, Y. M. Cheng and J. Xie, Amino acids, 2007, 33, 69-74.
13. A. Bulashevska and R. Eils, BMC bioinformatics, 2006, 7, 298.
14. M. S. Scott, D. Y. Thomas and M. T. Hallett, Genome research, 2004, 14, 1957-1966.
15. Y. L. Chen and Q. Z. Li, Journal of theoretical biology, 2007, 245, 775-783.
16. K. C. Chou, Proteins, 2001, 43, 246-255.
17. K. C. Chou and D. W. Elrod, Protein engineering, 1999, 12, 107-118.
18. C. Lin, Y. Zou, J. Qin, X. Liu, Y. Jiang, C. Ke and Q. Zou, PloS one, 2013, 8, e56499.
19. L. Song, D. Li, X. Zeng, Y. Wu, L. Guo and Q. Zou, BMC bioinformatics, 2014, 15, 298.
20. Q. Zou, X. B. Li, Y. Jiang, Y. M. Zhao and G. H. Wang, Curr Proteomics, 2013, 10, 2-9.
21. A. Garg, M. Bhasin and G. P. Raghava, The Journal of biological chemistry, 2005, 280, 14427-14432.
22. Q. B. Gao, Z. Z. Wang, C. Yan and Y. H. Du, FEBS letters, 2005, 579, 3444-3448.
23. Z. H. Zhang, Z. H. Wang, Z. R. Zhang and Y. X. Wang, FEBS letters, 2006, 580, 6169-6174.
24. Y. Chen, P. Yu, J. Luo and Y. Jiang, Mammalian genome : official journal of the International Mammalian Genome Society, 2003, 14, 859-865.
25. K. Nakai and P. Horton, Trends in biochemical sciences, 1999, 24, 34-36.
26. G. L. Fan and Q. Z. Li, Journal of theoretical biology, 2012, 304, 88-95.
27. P. Horton, K. J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier and K. Nakai, Nucleic acids research, 2007, 35, W585-587.
28. S. N. Tang, J. M. Sun, W. W. Xiong, P. S. Cong and T. H. Li, Biochimie, 2012, 94, 847-853.
29. M. Rashid, S. Saha and G. P. Raghava, BMC bioinformatics, 2007, 8, 337.
30. H. Lin, H. Ding, F. B. Guo and J. Huang, Molecular diversity, 2010, 14, 667-671.
31. W. Chen, P. M. Feng, H. Lin and K. C. Chou, Nucleic acids research, 2013, 41, e68.
32. S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, W. Chen and K. C. Chou, Bioinformatics, 2014, 30, 1522-1529.
33. B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, Q. Dong and K. C. Chou, Bioinformatics, 2014, 30, 472-479.
34. W. Chen, P. M. Feng, E. Z. Deng, H. Lin and K. C. Chou, Analytical biochemistry, 2014, 462, 76-83.
35. H. Ding, E. Z. Deng, L. F. Yuan, L. Liu, H. Lin, W. Chen and K. C. Chou, BioMed research international, 2014, 2014, 286419.
36. H. Lin, E. Z. Deng, H. Ding, W. Chen and K. C. Chou, Nucleic acids research, 2014, doi: 10.1093/nar/gku1019.
37. B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang and K. C. Chou, PloS one, 2014, 9, e106691.
38. K. C. Chou, Journal of theoretical biology, 2011, 273, 236-247.
39. M. Magrane and U. Consortium, Database : the journal of biological databases and curation, 2011, 2011, bar009.
40. G. Wang and R. L. Dunbrack, Jr., Bioinformatics, 2003, 19, 1589-1591.
41. K. C. Chou, Z. C. Wu and X. Xiao, PloS one, 2011, 6, e18258.
42. K. C. Chou, Z. C. Wu and X. Xiao, Molecular bioSystems, 2012, 8, 629-641.
43. Z. C. Wu, X. Xiao and K. C. Chou, Molecular bioSystems, 2011, 7, 3287-3297.

Molecular BioSystems Accepted Manuscript

44. Z. C. Wu, X. Xiao and K. C. Chou, Protein and peptide letters, 2012, 19, 4-14.
45. X. Xiao, Z. C. Wu and K. C. Chou, PloS one, 2011, 6, e20592.
46. X. Xiao, Z. C. Wu and K. C. Chou, Journal of theoretical biology, 2011, 284, 42-51.
47. D. H. Wang, N. K. Lee, T. S. Dillon and N. J. Hoogenraad, Iconip'02: Proceedings Of the 9th International Conference on Neural Information Processing, 2002, 764-768.
48. P. Ung and D. A. Winkler, Journal of medicinal chemistry, 2011, 54, 1111-1125.
49. S. Anishetty, G. Pennathur and R. Anishetty, BMC structural biology, 2002, 2, 9.
50. J. Ma and H. Gu, BMB reports, 2010, 43, 670-676.
51. I. Olivier and T. Loots du, Journal of microbiological methods, 2012, 88, 419-426.
52. J. B. Yin, Y. X. Fan and H. B. Shen, Current protein & peptide science, 2011, 12, 580-588.
53. T. Huang, Z. Xu, L. Chen, Y. D. Cai and X. Kong, PloS one, 2011, 6, e17291.
54. P. Jia, Z. Qian, K. Feng, W. Lu, Y. Li and Y. Cai, Journal of proteome research, 2008, 7, 1131-1137.
55. H. Lin and W. Chen, Journal of microbiological methods, 2011, 84, 67-70.
56. H. Lin and H. Ding, J Theor Biol, 2011, 269, 64-69.
57. T. Wang, J. Yang, H. B. Shen and K. C. Chou, Protein and peptide letters, 2008, 15, 915-921.
58. T. Wang, T. Xia and X. M. Hu, Journal of theoretical biology, 2010, 262, 208-213.
59. Y. Feng and L. Luo, Amino acids, 2008, 35, 607-614.
60. C. C. Chang and C. J. Lin, Acm T Intel Syst Tec, 2011, 2.
61. Z. P. Feng, In silico biology, 2002, 2, 291-303.
62. B. Liu, X. Wang, L. Lin, Q. Dong and X. Wang, BMC bioinformatics, 2008, 9, 510.
63. L. Wei, M. Liao, Y. Gao, R. Ji, Z. He and Q. Zou, IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM, 2013.
64. B. Liu, X. Wang, L. Lin, Q. Dong and X. Wang, Computational biology and chemistry, 2009, 33, 303-311.
65. Y. Zhang, B. Liu, Q. Dong and V. X. Jin, Protein and peptide letters, 2011, 18, 7-16.
66. W. Chen, P. Feng and H. Lin, J Ind Microbiol Biotechnol, 2012, 39, 579-584.
67. P. Feng, H. Lin, W. Chen and Y. Zuo, BioMed research international, 2014, 2014, 935719.
68. L. Chen, W. M. Zeng, Y. D. Cai, K. Y. Feng and K. C. Chou, PloS one, 2012, 7, e35254.
69. X. Xiao, P. Wang, W. Z. Lin, J. H. Jia and K. C. Chou, Analytical biochemistry, 2013, 436, 168-177.
70. K. C. Chou, Molecular bioSystems, 2013, 9, 1092-1100.
71. B. Liu, X. Wang, Q. Zou, Q. Dong and Q. Chen, Molecular Informatics, 2013, 32, 775-782.
72. B. Liu, X. Wang, L. Lin, B. Tang, Q. Dong and X. Wang, BMC bioinformatics, 2009, 10, 381.
73. B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou and X. Wang, Molecular Informatics, 2014, doi: 10.1002/minf.201400025.

Molecular BioSystems Accepted Manuscript