Biochemical and Biophysical Research Communications 401 (2010) 382–384

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Prediction of midbody, centrosome and kinetochore proteins based on gene ontology information

Wei Chen [a,*], Hao Lin [b,**]

[a] Department of Physics, School of Basic Medical Sciences, North China Coal Medical University, Tangshan 063000, China
[b] 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

ARTICLE INFO

Article history:
Received 2 September 2010
Available online 18 September 2010

Keywords:
Microkit proteins
Support vector machine
Gene Ontology
Localization

ABSTRACT

In the process of cell division, a great deal of proteins is assembled into three distinct organelles, namely midbody, centrosome and kinetochore. Knowing the localization of microkit (midbody, centrosome and kinetochore) proteins will facilitate drug target discovery and provide novel insights into understanding their functions. In this study, a support vector machine (SVM) model, MicekiPred, was presented to predict the localization of microkit proteins based on gene ontology (GO) information. A total accuracy of 77.51% was achieved using the jackknife cross-validation. This result shows that the model will be an effective complementary tool for future experimental study. The prediction model and dataset used in this article can be freely downloaded from http://cobi.uestc.edu.cn/people/hlin/tools/MicekiPred/.

1. Introduction

Cell division, a small segment of a larger cell cycle, is the process by which a parent cell equally and faithfully divides into two daughter cells [1]. The division in eukaryotes is known as mitosis, and leaves the daughter cell capable of dividing again. The corresponding sort of division in prokaryotes is known as binary fission. In another type of cell division present only in eukaryotes, called meiosis, a cell is permanently transformed into a gamete and cannot divide again until fertilization [1]. During cell division, numerous proteins spatially and temporally organize protein super-complexes at the three distinct regions of midbody [2,3], centrosome [4,5] and kinetochore [6,7], and orchestrate the accomplishment of cell division process.

Proteins located in the different organelles (midbody, centrosome and kinetochore) play distinctive roles in various processes. Midbody proteins are indispensable for cytokinesis, asymmetric cell division, and chromosome segregation [3], while centrosomal proteins are involved in fertilization and intracellular trafficking [8]. The kinetochore contains more than 45 different proteins, mediating the attachment and segregation of chromosome through microtubule of mitotic spindles [7,9–11]. Knowing the locations of microkit proteins is essential to understand their functions. Unfortunately, experimentally obtaining localization information is both expensive and time-consuming. Therefore, it is critically important to develop an automated method to reliably and quickly annotate microkit protein localizations.

In the last two decades, a great number of methods have been developed to predict protein localization, most of which were based on terminal signaling peptides [12,13], pseudo amino acid composition [14,15–18], dipeptide composition [19,20], functional domain composition [21,22]. And a number of machine learning approaches have been introduced, such as the Markov chain method [23], discriminate function [20,24,25], SVM [14,26,27], artificial neural network [28,29].

To the best of our knowledge, there exists no theoretical method for microkit protein localization prediction. In this article, a computational model called MicekiPred was developed to predict the localization of microkit proteins according to gene ontology (GO) information. In this model, the sequences of microkit proteins were translated into discrete numbers using GO information and then these numbers were integrated into a vector as the SVM input. In the jackknife cross-validation, MicekiPred yields a total accuracy of 77.51% for predicting the localization of 1005 non-redundant microkit proteins.

2. Materials and methods

2.1. Dataset

A set of 1489 microkit proteins were extracted from MiCroKit 3.0 [30]. 1248 proteins (278 midbody proteins, 570 centrosome proteins and 400 kinetochore proteins) with single localization were selected from the 1489 proteins. To prepare a high quality dataset, the following two procedures were performed. (i) Proteins

* Corresponding author.
** Corresponding author.
E-mail addresses: chenwei_imu@yahoo.com.cn (W. Chen), hlin@uestc.edu.cn (H. Lin).

0006-291X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2010.09.061

with sequence identity greater than 40% to each other were removed using the CD-HIT program [31]. (ii) In order to utilize the GO information, 25 proteins (2 midbody, 16 centrosome and 7 kinetochore proteins) with no GO annotation in MiCroKit 3.0 were removed from the non-redundant dataset. Finally, 236 midbody, 438 centrosome and 331 kinetochore proteins were retained in the high quality dataset.

2.2. Support vector machine

Support Vector Machine (SVM) is an effective method for supervised pattern recognition [32]. The SVM is well founded theoretically and of better-quality in practical applications. It has been widely used in the field of subcellular localization prediction [14,26,27]. The basic idea of SVM is to transform the data into a high dimensional feature space, and then determine the optimal separating hyperplane. Since this work deals with proteins of 3 localizations, this is a multi-class problem. For handling a multi-class problem, "one-versus-one (OVO)", "one-versus-rest (OVR)" are generally applied to extend the traditional SVM. In this work, OVO strategy was employed for making prediction. The implementation of SVM is based on LibSVM 2.84 written by Chang and Lin [33]. A radial basis function (RBF) was chosen as the kernel function. The grid search method is applied to tune the regularization parameter C and the kernel width parameter γ.

2.3. GO information

GO is a controlled vocabulary for uniformly describing gene products in terms of biological processes, cellular components and molecular function in any organism [34]. It has been shown that GO terms can be used to improve the performance of protein subcellular localization prediction [21,35]. Thus, we constructed a feature vector based on GO as the SVM input. The microkit proteins in the high quality dataset cover 1738 different GO entities. Each of the 1738 entities is served as a base to define a protein sample. If there is a hit corresponding to the ith entity, then the ith component of the protein in the 1738-dimensional GO space is assigned 1; otherwise, it is assigned 0 [36,37]. Thus, in the GO space, the protein sequence could be formulated as

$$G = [g_1, g_2, g_3, \ldots\ldots, g_{1738}]^T \tag{1}$$

2.4. PseAA composition

The pseudo amino acid (PseAA) composition proposed by Chou [15] describes both the feature of amino acid composition and the long distance interaction of physicochemical properties between residues. According to the concept of Chou's PseAA [15], the protein sequence could be represented by a $(20 + \lambda)$-dimensional vector defined by $(20 + \lambda)$ discrete numbers.

$$P = [p_1, p_2, \ldots p_{20}, p_{20+1}, \ldots\ldots, p_{20+\lambda}]^T \tag{2}$$

Here the first 20 numbers represent the classic amino acid composition, and the next λ discrete numbers describe sequence correlation factor, which can be calculated on the PseAAC web server (http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/). While using Chou's PseAA, two parameters (weight factor ω and correlation factor λ) should be determined in advance.

2.5. Prediction assessment

The performance of the model was measured in terms of Sensitivity (Sn), Specificity (Sp), Average accuracy (Aac), Total accuracy (Tac) and Matthew's correlation coefficient (MCC) [38].

$$Sn(i) = \frac{TP(i)}{TP(i) + FN(i)} \tag{3}$$

$$Sp(i) = \frac{TN(i)}{TN(i) + FP(i)} \tag{4}$$

$$MCC(i) = \frac{TP(i) \times TN(i) - FP(i) \times FN(i)}{\sqrt{(TP(i) + FN(i)) \times (TN(i) + FP(i)) \times (TP(i) + FP(i)) \times (TN(i) + FN(i))}} \tag{5}$$

$$Aac = \frac{1}{N} \sum_{i=1}^{k} Sn(i) \tag{6}$$

$$Tac = \frac{1}{N} \sum_{i=1}^{k} TP(i) \tag{7}$$

Here k ($k = 3$) is the number of classes, N is the total number of sequences, $TP(i)$, $TN(i)$, $FP(i)$, and $FN(i)$ represent true positive, true negative, false positive and false negative of class i, respectively.

3. Results and discussion

Three cross-validation methods, namely sub-sampling test, independent dataset test and jackknife test are often employed to evaluate the predictive capability of a predictor. Among the three methods, the jackknife test is deemed the most objective and rigorous one [39] that can always yield a unique outcome as demonstrated by a penetrating analysis in a recent comprehensive review [37] and has been widely and increasingly adopted [40–43]. Accordingly, the jackknife test was used to examine the performance of the model proposed in this study. In the jackknife test, each sequence in the training dataset is in turn singled out as an independent "test sample" and all the rule-parameters are calculated without using this one.

Based on GO information, a SVM model was constructed and tested in the high quality dataset containing 1005 proteins. With $C = 32$ and $\gamma = 0.000488$, we obtained an optimized results with a total accuracy of 77.51% for predicting microkit protein localization in the jackknife test (Table 1).

Because there is no other method for microkit protein localization prediction, we compared the performance of the above model to that obtained by using the parameter of PseAA. After a great number of testing, we found that, when $\omega = 0.05$, $\lambda = 8$, $C = 0.5$ and $\gamma = 0.5$, the SVM model based on PseAA yielded a maximum accuracy. The result of jackknife test shows a total accuracy of 54.33%, which is much lower than that achieved by combining SVM with GO information.

To avoid losing any information, we integrated the above two features: GO and PseAA. Then each protein sequence in the dataset was represented by a $1738 + 28 = 1766$-dimensional vector. The first 1738 elements reflecting the GO information and the remains are PseAA. However, as shown in the last column of Table 1, the improvement of the total accuracy was unremarkable compared to that obtained by using GO information alone. Especially, the average accuracy obtained by GO was 75.48% which is slightly better than that (75.04%) obtained by mixed features. In addition, although the mixed features could achieve higher sensitivity for centrosome protein localization prediction, the predictive successful rates of midbody and kinetochore proteins by using mixed features were not better than that by using GO information.

These results imply that GO plays a major role for the prediction of microkit protein localizations, while the information extracted from PseAA is insignificant. Thus, we recommend using the GO information for microkit protein localization. Consequently, by using GO information, we constructed a prediction model, MicekiPred, which

Table 1
Performance of MicekiPred for microkit protein localization prediction.

	GO (C = 32, γ = 0.000488)			PseAA (C = 0.5, γ = 0.5)			GO_PseAA (C = 8.0, γ = 0.00195)		
	Sn (%)	*Sp* (%)	*MCC*	*Sn* (%)	*Sp* (%)	*MCC*	*Sn* (%)	*Sp* (%)	*MCC*
Midbody	64.41	93.86	0.62	22.46	93.73	0.24	59.32	97.58	0.66
Centrosome	84.70	75.28	0.60	77.63	43.74	0.23	90.87	70.42	0.61
Kinetochore	77.34	91.12	0.70	47.52	70.94	0.19	74.92	93.08	0.70
Aac (%)	75.48			49.20			75.04		
Tac (%)	77.51			54.33			78.21		

could be freely downloaded from http://cobi.uestc.edu.cn/people/hlin/tools/MicekiPred/.

4. Conclusion

The accurate localization prediction of microkit proteins will be the foundation of understanding the molecular regulatory mechanisms of midbody, centrosome and kinetochore. By using GO information as the input parameter, we developed a SVM model, MicekiPred, to predict the localization of microkit proteins. The validation of MicekiPred in the high quality dataset showed a total accuracy of 77.51%, demonstrating that microkit protein localization can be accurately predicted by using the information deposited in GO. We hope that MicekiPred will be an effective tool for future experimental procedures in the realm of microkit protein localization annotation.

Acknowledgments

This work was supported by The Scientific Research Startup Foundation of UESTC, the Fundamental Research Funds for the Central Universities (ZYGX2009J081) and The Scientific Research Foundation of Sichuan Province (2009JY0013).

References

[1] D.O. Morgan, The Cell Cycle: Principles of Control, New Science, London, 2007.
[2] M.S. Otegui, K.J. Verbrugghe, A.R. Skop, Midbodies and phragmoplasts: analogous structures involved in cytokinesis, Trends Cell Biol. 15 (2005) 404–413.
[3] A.R. Skop, H. Liu, J.3rd. Yates, et al., Dissection of the mammalian midbody proteome reveals conserved cytokinesis mechanisms, Science 305 (2004) 61–66.
[4] S. Doxsey, D. McCollum, W. Theurkauf, Centrosomes in cellular regulation, Annu. Rev. Cell Dev. Biol. 21 (2005) 411–434.
[5] Z. Yang, J. Loncarek, A. Khodjakov, et al., Extra centrosomes and/or chromosomes prolong mitosis in human cells, Nat. Cell Biol. 10 (2008) 748–751.
[6] T. Sakuno, K. Tada, Y. Watanabe, Kinetochore geometry defined by cohesion within the centromere, Nature 458 (2009) 852–858.
[7] X. Wan, R.P. O'Quinn, H.L. Pierce, et al., Protein architecture of the human kinetochore microtubule attachment site, Cell 137 (2009) 672–684.
[8] S.L. Jaspersen, M. Winey, The budding yeast spindle pole body: structure duplication, and function, Annu. Rev. Cell Dev. Biol. 20 (2004) 1–28.
[9] I.M. Cheeseman, A. Desai, Molecular architecture of the kinetochore-microtubule interface, Nat. Rev. Mol. Cell Biol. 9 (2008) 33–46.
[10] T.U. Tanaka, A. Desai, Kinetochore-microtubule interactions: the means to the end, Curr. Opin. Cell Biol. 20 (2008) 53–63.
[11] S. Westermann, D.G. Drubin, G. Barnes, Structures and functions of yeast kinetochore complexes, Annu. Rev. Biochem. 76 (2007) 563–591.
[12] O. Emanuelsson, H. Nielsen, S. Brunak, et al., Predicting subcellular localization of proteins based on their N-terminal amino acid sequence, J. Mol. Biol. 300 (2000) 1005–1016.
[13] K. Nakai, P. Horton, PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, Trends Biochem. Sci. 24 (1999) 34–35.
[14] Y.D. Cai, X.J. Liu, X.B. Xu, et al., Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect, J. Cell. Biochem. 84 (2002) 343–348.
[15] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins 43 (2001) 246–255.
[16] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[17] H.B. Shen, K.C. Chou, Ensemble classifier for protein fold pattern recognition, Bioinformatics 22 (2006) 1717–1722.
[18] H.B. Shen, K.C. Chou, PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2008) 386–388.
[19] F.M. Li, Q.Z. Li, Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach, Amino Acids 34 (2008) 119–125.
[20] H. Lin, Q.Z. Li, Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant, Biochem. Biophys. Res. Commun. 354 (2007) 548–551.
[21] K.C. Chou, Y.D. Cai, Prediction of protein subcellular locations by GO-FunD-PseAA predictor, Biochem. Biophys. Res. Commun. 320 (2004) 1236–1239.
[22] C. Guda, S. Subramaniam, PTARGET: a new method for predicting protein subcellular localization in eukaryotes, Bioinformatics 21 (2005) 3963–3969.
[23] Z. Yuan, Prediction of protein subcellular location using Markov chain models, FEBS Lett. 451 (1999) 23–26.
[24] K.C. Chou, D.W. Elrod, Using Discriminant Function for Prediction of Subcellular Location of Prokaryotic Proteins, Biochem. Biophys. Res. Commun. 252 (1998) 63–68.
[25] K.C. Chou, D.W. Elrod, Protein subcellular location prediction, Protein Eng. 12 (1999) 107–118.
[26] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769.
[27] S.J. Hua, Z.R. Sun, Support vector machine approach for protein subcellular localization prediction, Bioinformatics 17 (2001) 721–728.
[28] Y.D. Cai, K.C. Chou, Using Neural Networks for Prediction of Subcellular Location of Prokaryotic and Eukaryotic Proteins, Mol. Cell. Biol. Res. Commun. 4 (2000) 172–173.
[29] A. Reinhardt, T. Hubbard, Using neural networks for prediction of the subcellular location of proteins, Nucleic Acids Res. 26 (1998) 2230–2236.
[30] J. Ren, Z.X. Liu, X.J. Gao, et al., MiCroKit 3.0: an integrated database of midbody Centrosome and kinetochore, Nucleic Acids Res. 38 (2010) 155–160.
[31] W.Z. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein database, Bioinformatics 17 (2001) 282–283.
[32] V.N. Vapnik, The Nature of Statistical Learning Theory, Wiley-Interscience, New York, 1998.
[33] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/_cjlin/libsvm.
[34] M. Ashburner, C.A. Ball, J.A. Blake, et al., Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.
[35] K.C. Chou, Y.D. Cai, A new hybrid approach to predict subcellular localization of proteins by incorporating Gene Ontology, Biochem. Biophys. Res. Commun. 311 (2003) 743–747.
[36] K.C. Chou, H.B. Shen, Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic k-nearest neighbor classifiers, J. Proteome Res. 5 (2006) 1888–1897.
[37] K.C. Chou, H.B. Shen, Review: recent progress in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.
[38] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta. 405 (1975) 442–451.
[39] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[40] Y.D. Cai, K.C. Chou, Predicting subcellular localization of proteins in a hybridization space, Bioinformatics 20 (2004) 1151–1156.
[41] D. Zou, Z. He, J. He, β-Hairpin prediction with quadratic discriminant analysis using diversity measure, J. Comput. Chem. 30 (2009) 2277–2284.
[42] K.C. Chou, D.W. Elrod, Prediction of enzyme family classes, J. Proteome Res. 2 (2003) 183–190.
[43] K.C. Chou, H.B. Shen, ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information, Crit. Rev. Biochem. Mol. Biol. 376 (2008) 321–325.