Protein J (2009) 28:384–390
DOI 10.1007/s10930-009-9205-0

Multi Label Learning for Prediction of Human Protein Subcellular Localizations

Lin Zhu · Jie Yang · Hong-Bin Shen

Published online: 6 October 2009

Abstract Predicting protein subcellular locations has attracted much attention in the past decade. However, one of the most challenging problems is that many proteins were found simultaneously existing in, or moving between, two or more different cell components in a eukaryotic cell. Seldom previous predictors were able to deal with such multiplex proteins although they have extremely important implications in future drug discovery in terms of their specific subcellular targeting. Approximately 20% of the human proteome consists of such multiplex proteins with multiple sample labels. In order to efficiently handle such multiplex human proteins, we have developed a novel multi-label (ML) learning and prediction framework called ML-PLoc, which decomposes the multi-label prediction problem into multiple independent binary classification problems. ML-PLoc is constructed based on support vector machine (SVM) and sequential evolution information. Experimental results show that ML-PLoc can achieve an overall accuracy 64.6% and recall ratio 67.2% on a benchmark dataset consisting of 14 human subcellular locations, and is very powerful for dealing with multiplex proteins. The current approach represents a new strategy to deal with the multi-label biological problems. ML-PLoc software is freely available for academic use at: http://www.csbio.sjtu.edu.cn/bioinf/ML-PLoc.

Keywords Human protein · Subcellular location · Multiplex protein · Multi-label learning · Support vector machine

L. Zhu · J. Yang · H.-B. Shen (✉)
Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, 800 Dongchuan Road, 200240 Shanghai, China
e-mail: hbshen@sjtu.edu.cn

Abbreviations

SVM	Support vector machine
PSSM	Position specific scoring matrix
PsePSSM	Pseudo position specific scoring matrix
OSH	Optimal separating hyperplane
TR	True positives
TN	True negatives
FP	False positives
FN	False negatives

1 Introduction

Every cell contains numerous protein molecules located in different compartments or organelles, the so-called subcellular locations [1, 2] and one of the fundamental goals in cell biology and proteomics is to identify the functions of proteins in the context of compartments that organize them in the cellular environment. Knowledge of the subcellular localization of proteins is important because it can (1) provide useful insights about their functions, (2) indicate how and in what kind of cellular environments they interact with each other and with other molecules, and (3) help in understanding the intricate pathways that regulate biological processes at the cellular level [3–6]. Of course, protein subcellular locations can be determined by conducting various time-consuming and expensive wet-lab biochemistry experiments. Particularly, with the rapid progress in genome sequencing projects, the number of new protein sequences increased explosively. Facing such a challenge, computational prediction of protein subcellular locations has attracted much attention and much progress has been obtained in recent years [7–17]. However, few of these established methods were specialized for human proteins; timely

annotation of protein subcellular locations is important and urgent because such knowledge directly relates to the practical application of drug discovery for human health and disease.

Two things are especially worthy of further investigation in developing effective algorithms for predicting human protein subcellular locations: (1) Enlarging the application scope. According to the experimental observations, human proteins are located in more than 10 different cell components. Although the algorithm HSLPred developed by Garg et al. [18] was specifically for human proteins, only four subcellular location sites are covered by HSLPred: cytoplasm, mitochondria, nucleus, and plasma membrane. If a query protein was potentially located outside the aforementioned four sites, the predictor would fail to work, or the results thus obtained would not make any sense; (2) Ability to deal with the multiplex proteins. According to our statistics, approximately 20% human proteins are found to be existing at, or moving between, two or more different subcellular locations. To the best of our knowledge, only the recently developed Hum-mPLoc, based on fusion of OET-KNN (optimized evidence-theoretic K nearest neighbor) algorithms, can handle the prediction of multiplex human proteins. Hum-mPLoc constructs an ensemble prediction engine for the entire application space; in that space it determines whether a protein belongs to multiple subcellular locations if the support degrees of these locations are located in a threshold scope optimally derived from the benchmark dataset. In most cases, by fusing multiple classifiers into one framework, the resulting prediction accuracy will be higher than a single independent classifier, which has been demonstrated in various protein classification problems [7, 9, 15, 19–23].

In this study, we propose a novel multi-label learning framework for predicting human multiplex proteins, which is called ML-PLoc. The basic idea of ML-PLoc is to construct independent SVM models for each of the subcellular locations and decompose the multi-label learning problem into multiple independent binary classification problems. The experimental results based on a benchmark dataset consisting of 14 human subcellular locations are quite encouraging. The current approach represents a new strategy to deal with the multi-label biological problems, and hence may become a useful vehicle in the area of bioinformatics and proteomics.

2 Materials and Methods

2.1 Datasets

One of the important steps for constructing an elegant subcellular location predictor is to construct a rigorous benchmark dataset [24, 25]. In this study, protein sequences were collected from the Swiss-Prot database at http://www.ebi.ac.uk/swissprot/ of the version 55.3 released on 29-Apr-2008, where the number of total protein entries is 366,226. The detailed procedures are as following: (1) Only those sequences annotated with "human" in the ID field were collected because the current study was focused on human proteins only, so the number is reduced to 19,293. (2) Sequences annotated with no, ambiguous or uncertain terms, such as "potential", "probable", "probably'", "maybe", or "by similarity", were excluded, where the total remained proteins is reduced to 7,337. As can be seen that more than 62% known human proteins still have no exact subcellular locations in the Swiss-Prot 55.3 database, indicating that the development of automated and accurate human protein subcellular locations very valuable. (3) Sequences annotated with "fragment" were excluded; also, sequences with less than 50 amino acid residues were removed because they might just be fragments. As a result, the number is further reduced to 7,313 with this criterion. (4) In order to collect as much desired information as possible, but meanwhile avoiding any homology bias, a redundancy cutoff was operated by a culling program to winnow those sequences which have $\geq$80% sequence identity to any other in a same subcellular location. After strictly following the aforementioned procedures, we finally obtained 5,909 human protein sequences, which were distributed among the following 14 subcellular locations (Fig. 1), as outlined in Table 1. Among the 5,909

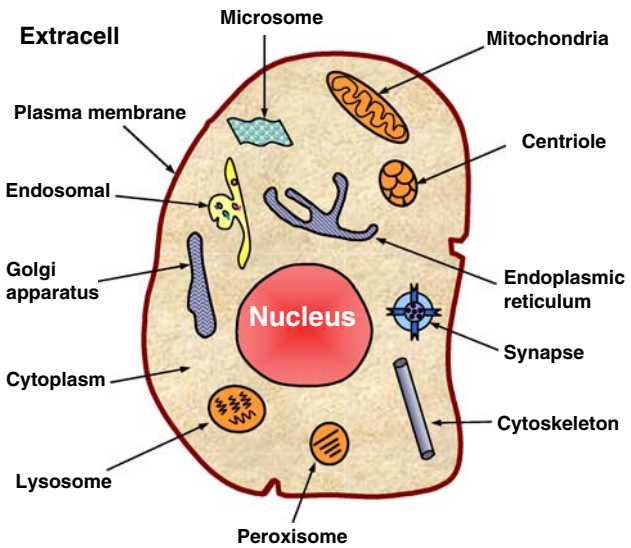

Fig. 1 Schematic illustration to show the fourteen subcellular locations of human proteins: 1 centriole, 2 cytoplasm, 3 cytoskeleton, 4 endoplasmic reticulum, 5 endosome, 6 extracell, 7 Golgi apparatus, 8 lysosome, 9 microsome, 10 mitochondrion, 11 nucleus, 12 peroxisome, 13 plasma membrane, and 14 synapse. (Reproduced from [7] with permission)

Table 1 Breakdown of the human protein benchmark dataset derived from Swiss-Prot database according to the procedures described in Sect. 2.1

Order	Subcellular location	Number of proteins
1	Centriole	99
2	Cytoplasm	1427
3	Cytoskeleton	132
4	Endoplasmic reticulum	330
5	Endosome	29
6	Extracell	804
7	Golgi apparatus	226
8	Lysosome	96
9	Microsome	55
10	Mitochondrion	468
11	Nucleus	2102
12	Peroxisome	58
13	Plasma membrane	1210
14	Synapse	70
Total number of locative proteins		7106
Total number of different proteins		5909[a]

[a] Of the 5,909 different proteins, 4,825 belongs to only 1 location, 986 to 2 locations, 83 to 3 locations and 15 to 4 locations, i.e., total 7,106 locative proteins

proteins, 4,825 belongs to only 1 location, 986 to 2 locations, 83 to 3 locations and 15 to 4 locations. The 5,909 human proteins located into 14 subcellular locations can be represented by:

$$\begin{bmatrix} 1 & \Delta_1^1 & \Delta_1^2 & \cdots & \Delta_1^{14} \\ 2 & \Delta_2^1 & \Delta_2^2 & \cdots & \Delta_2^{14} \\ \vdots & \vdots & \vdots & & \vdots \\ i & \Delta_i^1 & \Delta_i^2 & \cdots & \Delta_i^{14} \\ \vdots & \vdots & & & \vdots \\ 5909 & \Delta_{5909}^1 & \Delta_{5909}^2 & \cdots & \Delta_{5909}^{14} \end{bmatrix} \tag{1}$$

where

$$\Delta_i^j = \begin{cases} 1 & \text{if the } i\text{th protein belongs to the } j\text{th location} \\ 0 & \text{otherwise} \end{cases} \quad (i = 1, 2, \cdots, 5909;\ j = 1, 2, \cdots, 14) \tag{2}$$

According to Eq. 1, we have:

$$\begin{cases} \text{if } \sum_{j=1}^{14} \Delta_i^j > 1, \text{ then the } i\text{th protein is a multiplex protein} \\ \text{if } \sum_{j=1}^{14} \Delta_i^j = 1, \text{then the } i\text{th protein is a single location protein} \end{cases} \tag{3}$$

The sequences of the 5,909 proteins are given in Online Supporting Information A.

2.2 Encode Proteins into Feature Vectors

To develop a powerful method for predicting the subcellular localization of a human protein, one of the most important things is how to represent the sample of a protein by a descriptor that not only contains as much information as possible but also can be handled by a powerful prediction engine.

To incorporate the evolution information of proteins, the PSSM (Position-Specific Scoring Matrix) [26] was used; i.e., according to the concept of PSSM, the sample of a protein **P** can be represented by:

$$\mathbf{P}_{\text{PSSM}} = \begin{bmatrix} E_{1\to1} & E_{1\to2} & \cdots & E_{1\to20} \\ E_{2\to1} & E_{2\to2} & \cdots & E_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ E_{i\to1} & E_{i\to2} & \cdots & E_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ E_{L\to1} & E_{L\to2} & \cdots & E_{L\to20} \end{bmatrix} \tag{4}$$

where $E_{i\to j}$ represents the score of the amino acid residue in the i-th position of the protein sequence being changed to amino acid type j during the evolution process. The positive score means that the corresponding mutation occurs more frequently in the alignment than expected by chance, while the negative one means just the opposite. Here, the numerical codes $1, 2, \cdots, 20$ are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes. The $L \times 20$ scores in Eq. 4 were generated by using PSI-BLAST [26] to search the Swiss-Prot database through three iterations with 0.001 as the E-value cutoff for multiple sequence alignment against the sequence of the protein **P**, followed by a standardization procedure given by:

$$E_{i\to j} = \frac{E_{i\to j}^0 - \bar{E}_i^0}{\text{SD}(E_i^0)} \quad (i = 1, 2, \cdots, L;\ j = 1, 2, \cdots, 20) \tag{5}$$

where $E_{i\to j}^0$ represent the original scores directly created by PSI-BLAST that are generally shown as positive or negative integers; $\bar{E}_i^0$ the mean of $E_{i\to j}^0$over 20 native amino acids; $\text{SD}(E_i^0)$ the standard deviation of $E_{i\to j}^0$. However, according to the PSSM descriptor (Eq. 4), proteins with different lengths will correspond to row-different matrices. To make the PSSM descriptor become a size-uniform matrix, one possible approach is to represent a protein sample **P** by

$$\bar{\mathbf{P}}_{\text{PSSM}} = [\bar{E}_1 \quad \bar{E}_2 \quad \cdots \quad \bar{E}_{20}]^{\text{T}} \tag{6}$$

where

$$\bar{E}_j = \frac{1}{L}\sum_{i=1}^{L} E_{i\to j} \quad (j = 1, 2, \cdots, 20) \tag{7}$$

where $\bar{E}_j$ represents the average score of the amino acid residues in the protein **P** being changed to amino acid type

j during the evolution process. However, if $\bar{\mathbf{P}}_{\mathrm{PSSM}}$ of Eq. 6 was used to represent the protein $\mathbf{P}$, all the sequence-order information during the evolution process would be lost. To avoid complete loss of the sequence-order information, instead of Eq. 6, let us use the pseudo position-specific scoring matrix (PsePSSM) as given by

$$\mathbf{P}^{\xi}_{\mathrm{PsePSSM}} = \left[\begin{array}{cccccccc} \bar{E}_1 & \bar{E}_2 & \cdots & \bar{E}_{20} & \Phi^{\xi}_1 & \Phi^{\xi}_2 & \cdots & \Phi^{\xi}_{20} \end{array} \right]^{\mathrm{T}} \tag{8}$$

to represent the protein $\mathbf{P}$, where

$$\Phi^{\xi}_j = \frac{1}{L-\xi} \sum_{i=1}^{L-\xi} \left[E_{i \to j} - E_{(i+\xi) \to j} \right]^2 \quad (j = 1, 2, \cdots, 20;\ \xi < L) \tag{9}$$

meaning that Φ^1_j is the correlation factor by coupling the most contiguous PSSM scores along the protein chain for the amino acid type j; Φ^2_j that by coupling the second-most contiguous PSSM scores; and so forth. Note that, as mentioned in the Sect. 2, the length of the shortest protein sequence in the benchmark dataset is $L = 50$, and hence the value allowed for ξ in Eq. 9 must be smaller than 50. When $\xi = 0$, Φ^{ξ}_j becomes a naught element and Eq. 8 is degenerated to Eq. 6. Cross-validation experiments show that $\xi = 2$ is an optimized choice in this study, which is adopted in the following experiments.

2.3 Multi-Label Learning and Prediction

Support vector machine (SVM) is a popular machine learning algorithm based on structural risk minimization for pattern classification [27]. The basic idea of SVM is to transform the samples into a high-dimensional feature space and construct an optimal separating hyperplane (OSH) that maximizes its distance from the closest training samples.

For one of the 14 human subcellular location j, we thus can construct a two-class binary classification SVM model, i.e., SVM_j, where the positive (T^+_j) and negative (T^-_j) training samples for SVM_j is as following:

$$\begin{cases} \mathrm{T}^+_j = \bigcup_{i=1}^{5909} (\Delta^j_i = 1) \\ \mathrm{T}^-_j = \bigcup_{i=1}^{5909} (\Delta^j_i = 0) \end{cases} \tag{10}$$

where Δ^j_i is the label information as shown in Eq. 1, $\cup$ is the symbol for union in the set theory. For a query protein $\mathbf{P}$, we have

$$\mathrm{SVM}_j \triangleright \mathrm{P} = y_j \tag{11}$$

where $\triangleright$ represents the prediction operator and $y_j \in \{+1, -1\}$, where $y_j = +1$ means the protein P is predicted belonging to the jth location, otherwise not. Similarly, we can construct 14 independent binary SVM models for the 14 human subcellular locations, i.e.

$$\{\mathrm{SVM}_1, \mathrm{SVM}_2, \cdots, \mathrm{SVM}_{14}\} \tag{12}$$

where SVM_1 is the prediction model for the 1st subcellular location, SVM_2 the 2nd and so forth. From, Eq. 12 we will have 14 independent prediction outputs for a query protein $\mathbf{P}$, i.e.

$$\{y_1, y_2, \cdots, y_{14}\} \tag{13}$$

where $y_j \in \{+1, -1\}$ ($j = 1, 2, \cdots, 14$), indicating whether $\mathbf{P}$ is predicted belonging to the jth location or not. Obviously, if more than 1 prediction output of Eq. 13 is $+1$, then P is predicted to be a multiplex protein. In the case of all the outputs for the 14 binary SVMs of Eq. 13 are -1, then the query protein is inputted into a multi-class SVM classifier trained on the whole dataset consisting of 14 subcellular locations for final prediction. To provide an intuitive picture, a flowchart is provided in Fig. 2 to show the process of how the ML-PLoc classifier works in predicting human protein subcellular localizations.

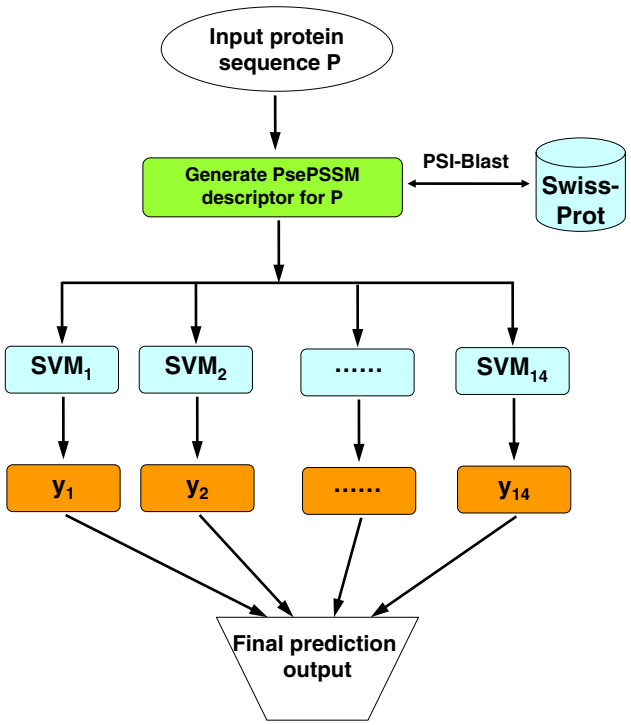

Fig. 2 Figure to show how the multiple label learning ML-PLoc for predicting human protein subcellular locations. SVM_1 is the binary prediction model for the centriole subcellular location, SVM_2 for cytoplasm and so forth

3 Results and Discussions

Jackknife cross-validation test is employed to evaluate the performance of the proposed new prediction algorithm of this study, which has been widely used by investigators to examine the power of various prediction methods. The terms true positives(TR), true negatives(TN), false positives(FP) and false negatives(FN) are used to compare the given prediction of an item (the class label assigned to the item by a classifier) with the desired correct observation (the class the item actually belongs to). The most commonly used performance evaluation measures are the precision and recall, which are also used in current study:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{TF}} \tag{14}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

The LibSVM software package developed by Chang and Lin is used for constructing the SVM model [28]. Two parameters of SVM model should be optimized, i.e., the kernel parameter γ and the regularization parameter C. The parameterization of SVM was then performed through a grid search over γ and C on the benchmark dataset. We considered γ and C values selected both from the range $2^0$to 2^{10}grid and then we tried $11 \times 11 = 121$ different combinations. Finally, we selected the RBF kernel function at $\gamma = 2^1$ and $C = 2^7$ to build the SVM models and this combination of parameters has been demonstrated to yield the best performance in our studies.

Table 2 illustrates the detailed results on the 14 human subcellular locations in the benchmark dataset, as can be

Table 2 Prediction results distributed among the following 14 subcellular locations

Subcellular location	Number of locative proteins	Number of predicted locative protein	Number of correct predicted locative protein	Number of incorrect predicted locative protein	Precision of predicted locative protein	Recall of predicted locative protein
Centriole	99	89	31	58	0.3483	0.3131
Cytoplasm	1,427	1,533	809	724	0.5277	0.5669
Cytoskeleton	132	108	44	64	0.4074	0.3333
Endoplasmic reticulum	330	332	165	167	0.4970	0.5000
Endosome	29	25	3	22	0.1200	0.1034
Extracell	804	858	596	262	0.6946	0.7413
Golgi apparatus	226	217	81	136	0.3733	0.3584
Lysosome	96	79	37	42	0.4684	0.3854
Microsome	55	51	41	10	0.8039	0.7455
Mitochondrion	468	476	269	207	0.5651	0.5748
Nucleus	2,102	2,279	1,632	647	0.7161	0.7764
Peroxisome	58	36	16	20	0.4444	0.2759
Plasma membrane	1,210	1,246	1,000	246	0.8026	0.8264
Synapse	70	62	48	14	0.7742	0.6857
Total	7106	7391	4772	2619	0.6457	0.6715

Table 3 The "exact match"[a] ratio between the prediction outputs and the real-world observations

Number of locations N[b]	Number of different proteins	Number of correct predicted different protein	Prediction accuracy (%)	Prediction accuracy by random predictor[c]
1	4,825	2,901	60.1	$\frac{1}{C_{14}^1} = 7.1\%$
2	986	240	24.3	$\frac{1}{C_{14}^2} = 1.1\%$
3	83	3	3.6	$\frac{1}{C_{14}^3} = 0.27\%$
4	15	1	6.7	$\frac{1}{C_{14}^4} = 0.09\%$

[a] "Exact match" means both the number and the annotations of the subcellular locations are the same

[b] Proteins that simultaneously belong to N subcellular locations

[c] C_{14}^N is the Combination Operator and equal to $\frac{14!}{N!(14-N)!}$

seen from which, the current method can reach an overall prediction accuracy of precision 64.6% and recall 67.2%. As shown in the Table 2, the prediction accuracy varies from different locations. The prediction accuracy is in most cases influenced by two factors given a specific algorithm: (1) the number of training samples in the location; and (2) the multiplex characteristics of proteins in the location. Generally speaking, the more samples in a location and the simpler of their distributions, the higher success rates will be obtained. For example, there are only 29 samples in the endosome location, however, 69% of them are found belonging to two or more locations, i.e., the multiplex proteins; whereas there are 1,210 samples in the plasma membrane location and only about 18% of them are multiplex proteins. That's why we obtain only 12% success rate in endosome whereas 80% in plasma membrane. With more experimental results are obtained in the future, we can further increase the size of the some locations such as endosome, which will of course further enhance the corresponding prediction results.

For the 5,909 different proteins in the benchmark dataset, Table 3 illustrates the "exact match" ratio between the prediction outputs and the real-world observations, where the "exact match" means both the number and the annotations of the subcellular locations are the same. For comparison, the success rates by the random predictor are also shown. As can be seen from Table 3 that the prediction accuracy of current model is significantly higher than the random predictor and it is very promising for handling multiplex proteins. In order to make the readers can understand the output from ML-PLoc much more easily, Table 4 illustrates the prediction results by ML-PLoc for some human proteins in the Swiss-Prot database that do not have subcellular location annotations or are annotated as being uncertain. It is instructive to point out that for the proteins that have only ambiguous annotations in Swiss-

Table 4 The predicted results by ML-PLoc for some human protein without subcellular location annotations available from databanks or annotated as uncertain terms such as "probable", "potential", and "by similarity"

Accession No.	Swiss-Prot code	Annotation in Swiss-Prot database	Identified location by ML-PLoc	Comment[a]
O95340	PAPS2_HUMAN		Mitochondrion	Single
Q92783	STAM1_HUMAN	Cytoplasm (probable)	Cytoplasm	Single
O00213	APBB1_HUMAN		Nucleus	Single
P46926	GNPI_HUMAN	Cytoplasm (by similarity)	Cytoplasm	Single
O00444	PLK4_HUMAN		Cytoplasm	Single
Q9NUI1	DECR2_HUMAN	Peroxisome (by similarity)	Mitochondrion Peroxisome	Multiple
O00570	SOX1_HUMAN	Nucleus (probable)	Nucleus	Single
Q8WWZ3	EDAD_HUMAN	Cytoplasm (probable)	Cytoplasm Nucleus	Multiple
O00762	UBE2C_HUMAN		Cytoplasm Nucleus	Multiple
Q96A47	ISL2_HUMAN	Nucleus (by similarity)	Nucleus	Single
O14529	CUTL2_HUMAN	Nucleus (by similarity)	Nucleus	Single
Q8IVH4	MMAA_HUMAN	Mitochondrion (probable)	Mitochondrion	Single
O14782	KIF3C_HUMAN		Nucleus	Single
O14607	UTY_HUMAN	Nucleus (potential)	Cytoplasm Nucleus	Multiple
O15164	TIF1A_HUMAN	Nucleus (potential)	Nucleus	Single
Q96P15	SPB11_HUMAN	Cytoplasm (by similarity)	Cytoplasm Extracell	Multiple
O43240	KLK10_HUMAN	Secreted protein (probable)	Secreted protein	Single
Q8NCQ2	CS034_HUMAN		Nucleus	Single
O43708	MAAI_HUMAN	Cytoplasm (by similarity)	Cytoplasm	Single
O60291	MGRN1_HUMAN		Cytoplasm	Single

[a] "Single" means protein was predicted belonging to only 1 location, and "Multiple" means protein was predicted belonging to multiple locations

Prot database the predicted outputs from ML-PLoc are in most cases consistent with the uncertain annotations in Swiss-Prot database. However, some are not, for example, the "O14607" is potentially located in nucleus according to Swiss-Prot database, whereas according to ML-PLoc predictor, this protein has a high propensity to be a multiplex protein, moving between both cytoplasm and nucleus, which is very interesting and will sure open a new door for further biological experiments on the protein.

The results obtained by current predictor are very promising considering the following facts: (1) As is well known, the more subcellular locations under its coverage, the more difficult it will be to enhance the overall success rate. Current prediction model has covered 14 human subcellular locations, which is significantly improved compared to HSLPred [18] covering 4 subcellular locations. (2) Inclusion of proteins with multiple location sites will further complicate the difficulty of prediction.

4 Conclusions

Approximately 20% human proteins are multiplex proteins. In this paper, we have proposed a novel multi-label learning framework ML-PLoc for predicting human protein subcellular locations based on SVM algorithm and PsePSSM, which has been demonstrated very powerful for handling the multiplex proteins. To support the people working in the relevant area, this method is freely available as a stand-lone Linux based software at: http://www.csbio.sjtu.edu.cn/bioinf/ML-PLoc. The current approach represents a new strategy to deal with the multi-label biological problems, and hence may become a useful vehicle in the area of bioinformatics and proteomics.

Acknowledgments We gratefully thank two anonymous reviewers for their helpful and constructive comments. This work was supported by the National Natural Science Foundation of China (Grant No. 60704047 and 60805001), Science and Technology Commission of Shanghai Municipality (Grant No. 08ZR1410600, 08JC1410600), sponsored by Shanghai Pujiang Program, Innovation Program of Shanghai Municipal Education Commission (10ZZ17), and supported by Shanghai Leading Academic Discipline Project (Grant No. S30201).

References

1. Chou KC (2004) Curr Med Chem 11:2105–2134
2. Lubec G, Afjehi-Sadat L, Yang JW, John JPP (2005) Prog Neurobiol 77:90–127
3. Ehrlich JS, Hansen MDH, Nelson WJ (2002) Developmental Cell 3:259–270
4. Glory E, Murphy RF (2007) Dev Cell 12:7–16
5. Huang HY, Chang HY, Chou CH, Tseng CP, Ho SY, Yang CD, Ju YW, Huang HD (2009) Nucleic Acids Res 37:D150–D154
6. Lee TY, Hsu JB, Chang WC, Wang TY, Hsu PC, Huang HD (2009) BMC Res Notes 2:111
7. Chou KC, Shen HB (2007) Anal Biochem 370:1–16
8. Chou KC, Shen HB (2008) Nat Protoc 3:153–162
9. Du P, Li Y (2006) BMC Bioinform 7:518
10. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) J Mol Biol 300:1005–1016
11. Gu Q, Ding YS, Jiang XY, Zhang TL (2008) Amino Acids. doi: 10.1007/s00726-008-0209-4
12. Jia P, Qian Z, Zeng Z, Cai Y, Li Y (2007) Biochem Biophys Res Commun 357:366–370
13. Jiang L, Li M, Wen Z, Wang K, Diao Y (2006) Protein J 25:241–249
14. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C, Eisner R (2004) Bioinformatics 20:547–556
15. Nanni L, Lumini A (2008) Amino Acids 34:653–660
16. Yu CS, Lin CJ, Hwang JK (2004) Protein Sci 13:1402–1406
17. Yuan Z (1999) FEBS Lett 451:23–26
18. Garg A, Bhasin M, Raghava GP (2005) J Biol Chem 280:14427–14432
19. Bulashevska A, Eils R (2006) BMC Bioinformatics 7:298
20. Chou KC, Shen HB (2006) J Cell Biochem 99:517–527
21. Gu J, Fu H, Zhang X, Li Y (2007) BMC Bioinform 8:432
22. Nanni L, Lumini A (2007) Pattern Recogn Lett 28:622–630
23. Nanni L, Lumini A (2009) Protein Pept Lett 16:163–167
24. Boschetti E, Lomas L, Citterio A, Righetti PG (2007) J Chromatogr A 1153:277–290
25. Hu L, Ye M, Zou H (2009) Expert Rev Proteomics 6:433–447
26. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF (2001) Nucleic Acids Res 29:2994–3005
27. Vapnik V, Chapelle O (2000) Neural Comput 12:2013–2036
28. Chang CC, Lin CJ (2001). LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm