

# SCIENTIFIC REPORTS



OPEN

## Human Protein Subcellular Localization with Integrated Source and Multi-label Ensemble Classifier

Xiaotong Guo<sup>1</sup>, Fulin Liu<sup>1</sup>, Ying Ju<sup>2</sup>, Zhen Wang<sup>2</sup> & Chunyu Wang<sup>3</sup>

Received: 24 February 2016

Accepted: 26 May 2016

Published: 21 June 2016

Predicting protein subcellular location is necessary for understanding cell function. Several machine learning methods have been developed for computational prediction of primary protein sequences because wet experiments are costly and time consuming. However, two problems still exist in state-of-the-art methods. First, several proteins appear in different subcellular structures simultaneously, whereas current methods only predict one protein sequence in one subcellular structure. Second, most software tools are trained with obsolete data and the latest new databases are missed. We proposed a novel multi-label classification algorithm to solve the first problem and integrated several latest databases to improve prediction performance. Experiments proved the effectiveness of the proposed method. The present study would facilitate research on cellular proteomics.

Cells are highly ordered structure and contain various subcellular compartments that ensure the normal operation of the entire cell. These subcellular structures include nuclei, mitochondria, endoplasmic reticulum, Golgi apparatus, cell membrane, and extracellular matrix. The biological function of cells is executed by its unique proteins. Protein synthesized on the ribosome must be transported to its corresponding subcellular structures to play a normal biological function. If protein subcellular localization does not correspond to its position, serious loss of function or disorder occurs in organisms. Researchers found the aberrant protein subcellular localization in some cell lesions (such as cancer cells)<sup>1</sup>. The subcellular location of proteins is an important attribute of proteins, which is useful in determining protein function, revealing the mechanism of molecular interaction, and understanding the complex physiological processes<sup>2</sup>. The subcellular location of proteins is of great significance to cell biology, proteomics, and drug design research<sup>3</sup>.

Using conventional biochemical research methods, such as cell separation method, electronic microscopy, and fluorescence microscopy, to predict protein subcellular localization is expensive, time consuming, and laborious<sup>4</sup>. In today's post-genome era, large amounts of protein sequence provide raw materials for the development of biological information and a stage for machine learning methods' application in the field of life science<sup>5</sup>.

The typical protein subcellular location system based on machine learning methods includes the following four basic steps: (1) establishment of protein data set, (2) protein sequence feature extraction, (3) design of multi-label classification algorithm, and (4) construction of Web server<sup>6</sup>.

Databases for protein subcellular location, include LOCATE<sup>7</sup>, PSORTdb<sup>8</sup>, Arabidopsis Subcellular DB<sup>9</sup>, Yeast Subcellular DB<sup>10</sup>, Plant-PLoc<sup>11</sup>, LOCtarget<sup>12</sup>, LOC3D<sup>13</sup>, DBSubloc<sup>14</sup>, and PA-GOSUB<sup>15</sup>. However, none of the current works on computational protein subcellular localization have integrated these sources. Only part of the protein sequences were employed for training in previous works. In this paper, we collected existing related data sets and integrated a complete data set.

Feature extraction is a key process in various protein classification problems. Feature vectors are sometimes called as fingerprints of proteins. The common features include Chou's PseACC representation<sup>16</sup>, K-mer and K-ship frequencies<sup>17</sup>, Chen's 188D composition and physicochemical characteristics<sup>18</sup>, Wei's secondary structure features<sup>19,20</sup>, and PSSM matrix features<sup>21</sup>. Several web servers were also developed for feature extraction of protein primary sequence, including Pse-in-one<sup>22</sup>, Protrweb<sup>23</sup>, and PseAAC<sup>24</sup>.

Proper classifier can help to improve the prediction performance. Support vector machine (SVM), k-nearest neighbor (kNN), artificial neural network, random forest (RF)<sup>25</sup>, and ensemble learning<sup>26,27</sup> are often employed for special peptide identification. However, subcellular localization of a protein in essence is a multi-label

<sup>1</sup>School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing, China. <sup>2</sup>School of Information Science and Technology, Xiamen University, Xiamen, China. <sup>3</sup>School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. Correspondence and requests for materials should be addressed to C.W. (email: chunyu@hit.edu.cn)

Ordinal	Subcellular location	The number of proteins
1	centrosome	77
2	cytoplasm	817
3	cytoskeleton	79
4	endoplasmic reticulum	229
5	endosome	24
6	extracellular matrix	385
7	Golgi apparatus	161
8	lysosome	77
9	microsome	24
10	mitochondria	364
11	nucleus	1021
12	peroxisome	47
13	cell membrane	354
14	synaptic vesicle	22
Total number of non repeating protein sequences:		3106
Total number of protein Locative sequences:		3681
Multiple marker protein sequence data set $D_{M1}$		526
Single marker protein sequence data set $D_{S1}$		2580

**Table 1.** The protein sequences distribution on 14 subcellular sites.

classification problem, which is different from methods for identifying cellular factors (multi classification learning). Recently, several multi-label classification methods have been employed for subcellular localization in different species, including human<sup>28,29</sup>, plant<sup>30</sup>, virus<sup>31,32</sup>, eukaryote<sup>33,34</sup>, animal<sup>35</sup>. Features were also extracted according to n-gram<sup>36</sup>, Chou's PseAAC representation<sup>37</sup>, and gene ontology<sup>38</sup>. They all focused on the features construction. Only the basic multi-label strategies were employed. Most of their researches have transferred SVM to multi labels. We found that advanced ensemble multi-label learning techniques would further improve the performance.

## Material and Methods

**Integration of multiple protein subcellular localization sources.** In this section, we reconstruct the training set for human protein subcellular localization study. The new data set has a richer source and we further reduce the redundancy with CD-HIT<sup>39</sup>. Meanwhile, we expand the size of data sets, which render the training set data more comprehensive and provide a more convincing database for the multi-label classification learning step. The training set reconstruction will be introduced from two aspects, namely, data sources and data processing. The new dataset contains mainly two sources, which are LOCATE<sup>7</sup> and Hum-mPLoc 2.0<sup>40</sup>.

About 526 ( $480+43+3=526$ ) protein sequences are recorded as multi-label sequences (no repeat), which have two or more types of subcellular sites (the number of sites  $P_1$  is greater than or equals to 1)  $D_{M1}$ . The protein sequence distribution on each subcellular site is shown in Table 1.

The subcellular sites contained in the proteins in Hum-mPLoc 2.0 are scarce, but parts of the protein data contain three or four subcellular sites. Proteins are rich and varied. Therefore, certain superiority is shown in terms of protein function.

From the LOCATE database, we directly obtained the document human.xml of the original XML format about subcellular localization of human. The document accommodates abundant information about human proteins. Our goal is to obtain 64,637 human protein amino acid FASTA sequences and the subcellular sites (site number  $P_2$  is more than or equals to 1) of these sequences. After a rigorous data processing, we obtain the reference data set containing 6776 different protein sequences (no repeat)  $D_2$ . The 6776 protein sequences are distributed in 37 subcellular structures and possess two subcellular locations at most. Among these sequences, 4066 have only one type of subcellular location, which belongs to the single marker sequence data set  $D_{S2}$ . Approximately 2710 protein sequences have two subcellular locations (site number  $P_2$  equals to 1), which belong to the multiple marker sequence data set  $D_{M2}$ . A total of 9486 ( $4066+2710*2=9486$ ) protein sequences (proteins locative, a repetitive protein sequence) correspond to 37 subcellular locations. The protein sequence distribution on each subcellular site is shown in Table 2.

Results of data processing indicate an extremely rich types of proteins and subcellular sites in the LOCATE database. However, the number of protein sequences, which have multiple subcellular sites, is relatively small, especially those belonging to three or more types of subcellular sites. This finding indicates that the protein data in the LOCATE have problems in functional diversity. To compensate for the limitations in the LOCATE database and Shen's basic data set, we combine two types of data and reconstruct basic data sets. By combining Tables 1 and 2, we conclude that the 14 types of subcellular sites in Hum-mPLoc 2.0 are contained entirely in 37 types of subcellular sites in the LOCATE database, which is conducive to our data set reconstruction.

In order to prove the necessary of multi-label classification in the protein subcellular localization, it is required to compare the performances of multi-label and single-label classifiers. However, multi-label dataset cannot be

Ordinal	Subcellular location	The number of proteins
1	Apical Plasma Membrane	44
2	Plasma Tube Basement Membrane	101
3	Cellular Component Unknown	6
4	Centrosome	27
5	Cytoplasm	1044
6	Cytoplasmic Vesicle	176
7	Cell Scaffold	72
8	Early Endosome	147
9	Endoplasmic Reticulum	343
10	Nuclear Body	454
11	ERGIC	7
12	Extracellular Matrix	394
13	Golgi Apparatus	387
14	Golgi Cis Cisterna	27
15	Golgi Trans Cisterna	11
16	Golgi Trans Face	27
17	Mitochondrial Inner Membrane	8
18	Late Endosomes	34
19	Lipid Lowering Granule	1
20	Lysosome	222
21	Medial-Golgi	17
22	Black Body	20
23	Microtubule Organizing Center	8
24	Micro Tube	4
25	Mitochondrion	279
26	Nuclear Membrane	155
27	Nucleolus	810
28	Nucleus	2721
29	Mitochondrial Outer Membrane	2
30	Peroxisome	128
31	Cell Membrane	1711
32	Muscle Fiber Membrane	5
33	Secretory Granules	25
34	Secretory Vesicle	8
35	Synaptic Vesicle	29
36	Tight Junction	26
37	Transport Vesicle	6
Total number of non repeating protein sequences:		6776
Total number of protein Locative sequences:		9486
Multiple marker protein sequence data set $D_{M2}$		2710
Single marker protein sequence data set $D_{S2}$		4066

**Table 2. The protein sequences distribution on 37 subcellular sites in LOCATE.**

used for single-label classifiers. Therefore, the data sets of multi-label protein sequences and single-label protein sequences were reconstructed separately, but they both come from the sources mentioned in the above section. The reconstructed data set was  $D_{RM}$ , and the single labeled data set was  $D_{RS}$ . Therefore,

$$D_{RM} = \text{CD-HIT}[D_{M1} + D_{M2}] \quad (1)$$

$$D_{RS} = \text{CD-HIT}[D_{S1} + D_{S2}] \quad (2)$$

$$D_R = D_{RM} + D_{RS} \quad (3)$$

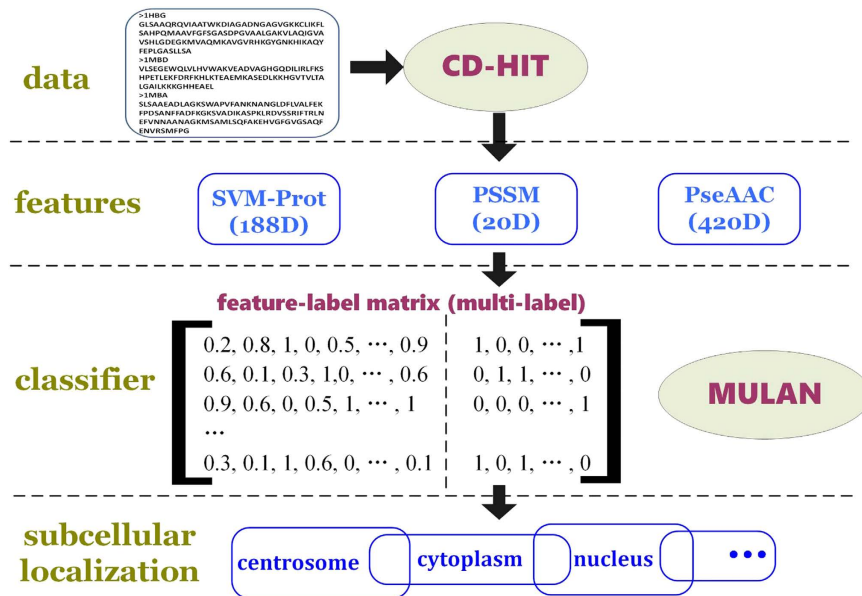
CD-HIT<sup>39</sup> is a software for reducing the similarity of the protein sequences. It can delete the similar sequences from the data set. Here we made the similarity of each pair sequences is less than 40%. Table 3 shows the protein sequences of the reconstructed data set  $D_R$  and the subcellular sites.

The multiple labeled set $D_{RM}$			The single labeled set $D_{RS}$		
Ordinal	Subcellular location	The number of proteins	Ordinal	Subcellular location	The number of proteins
1	Apical Plasma Membrane	16	1	centrosome	44
2	Plasma Tube Basement Membrane	29	2	cytoplasm	508
3	Cellular Component Unknown	4	3	cytoskeleton	46
4	Centrosome	37	4	endoplasmic reticulum	77
5	Cytoplasm	542	5	endosome	163
6	Cytoplasmic Vesicle	29	6	extracellular matrix	419
7	Cell Scaffold	43	7	Golgi apparatus	125
8	Early Endosome	52	8	lysosome	86
9	Endoplasmic Reticulum	43	9	Micro Tube	11
10	Nuclear Body	179	10	mitochondria	355
11	ERGIC	4	11	nucleus	952
12	Extracellular Matrix	68	12	peroxisome	56
13	Golgi Apparatus	147	13	cell membrane	565
14	Golgi Cis Cisterna	7	14	synaptic vesicle	16
15	Golgi Trans Cisterna	3	15	Cytoplasmic Vesicle	17
16	Golgi Trans Face	11	16	Black Body	4
17	Mitochondrial Inner Membrane	4	17	Nuclear Membrane	1
18	Late Endosomes	16	18	Secretory Granules	1
19	Lipid Lowering Granule	1	19	Secretory Vesicle	2
20	Lysosome	39	\	\	\
21	Medial-Golgi	7	\	\	\
22	Black Body	2	\	\	\
23	Microtubule Organizing Center	1	\	\	\
24	Micro Tube	15	\	\	\
25	Mitochondrion	52	\	\	\
26	Nuclear Membrane	46	\	\	\
27	Nucleolus	268	\	\	\
28	Nucleus	768	\	\	\
29	Mitochondrial Outer Membrane	1	\	\	\
30	Peroxisome	11	\	\	\
31	Cell Membrane	271	\	\	\
32	Muscle Fiber Membrane	1	\	\	\
33	Secretory Granules	9	\	\	\
34	Secretory Vesicle	3	\	\	\
35	Synaptic Vesicle	12	\	\	\
36	Tight Junction	9	\	\	\
37	Transport Vesicle	4	\	\	\
The reconstructed multiple labeled set $D_{RM}$					1354
The reconstructed single labeled set $D_{RS}$					3448
The reconstructed protein subcellular localization data set $D_R$					4802

**Table 3. Subcellular sites and protein sequences distribution in  $D_R$ .**

**Features for subcellular localization.** The above section mainly discusses a series of preprocessing with the data set. The reconstructed data set provides a reliable database for the study on the positioning method. This section focuses on specific features of protein subcellular localization based on machine learning.

In this section, three types of feature extraction methods are introduced based on the position-specific scoring matrix (PSSM)<sup>41</sup>, pseudo-amino acid composition<sup>42</sup>. In the long process of evolution, some characteristic genes are not eliminated but are selectively retained. These characteristics can effectively characterize the corresponding protein. Feature extraction methods based on PSSM are conducted to compare the protein sequence and rationally analyze with the invariance. PSSM matrix represents the comparison results between the input protein sequence and its homologous protein sequence in Swiss-Prot database. The multiple sequence alignment tools are HAlign<sup>43</sup> and PSI-BLAST<sup>44</sup> (position-specific initiated BLAST). Each input protein sequence generates a PSSM matrix after multiple sequence alignment. The elements in PSSM matrix characterize homology level between amino acids in some positions in the input protein sequence and the amino acid in the corresponding position in its homologous sequence. A smaller element value indicates higher conservation; lower conservation means



**Figure 1.** Working flow chart for our method.

that the amino acid in the position is prone to mutation. We extracted 20D and 420D features from the PSSM according to different parameters, which are described in detail in the supplementary materials.

The purpose of PseAAC is also to improve the accuracy of protein subcellular localization and the prediction of membrane protein. We extracted 188D features from PseAAC, including 20D features of amino acid compositions, 24D features based on the contents of amino acids with certain physicochemical properties, 24D features of bivalent frequency and 120D features from eight physicochemical properties. It is described in detail in the supplementary materials, too.

**Multi-label classification ensemble learning method.** We employed the ensemble multi-label classification method for improving the prediction performance. There have been no ensemble methods for multi-label classification in bioinformatics so far. Next we described the ensemble voting strategies of our method.

Basic classifiers are denoted as  $C = \{c_i : i = 1, 2, \dots, p\}$ , and the labels are denoted as  $Y = \{\lambda_j : j = 1, 2, \dots, q\}$ .

**MeanEnsemble algorithm.** The prediction result is the probability that the sample is predicted to be  $\lambda_j$  by  $c_i$ . We calculate the average value of each column. Each training sample generates a set of  $q$ -dimensional vector:

$$V_{ME} = \{v_j : 1 \leq j \leq q\}$$

$v_j$  is the probability that the sample belongs to the corresponding class label. If  $0.5 \leq v_j \leq 1$ , the sequence belongs to  $\lambda_j$ . If  $0 \leq v_j < 0.5$ , the sequence does not belong to  $\lambda_j$ .

**MajorityVoteEnsemble algorithm.** Every basic classifier separately predicted a sample. The prediction result is  $S$ ,  $S \in (-1, +1)$ . If  $S = -1$ , the sample is recognized as the counterexample by the base classifier; otherwise, it is identified as a positive example. We calculate the average value of each column, and each training sample generates a set of  $q$ -dimensional vector:

$$V_{MV} = \{v_j : 1 \leq j \leq q\}$$

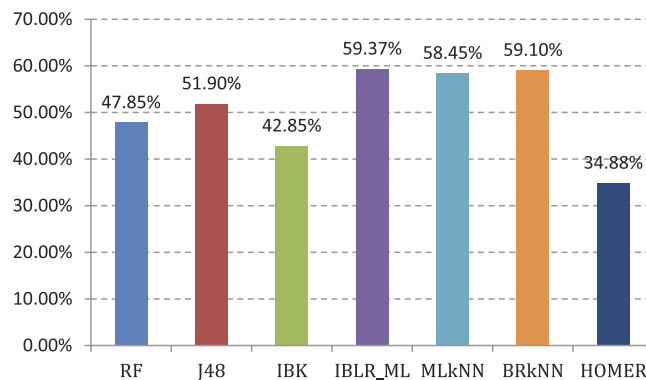
If  $v_j \geq 0$ , the sample belongs to  $\lambda_j$ ; otherwise, it does not.

**TopKEnsemble algorithm.** In each column in the result matrix,  $P$  accuracy values are sorted in descending order and the average of the first  $K$  ( $K$  is determined by  $p$ ) accuracy values is calculated to obtain a set of  $q$ -dimensional vector:

$$V_{TK} = \{v_j : 1 \leq j \leq q\}$$

If  $0.5 \leq v_j < 1$ , the sequence belongs to  $\lambda_j$ . If  $0 \leq v_j < 0.5$ , the sequence does not belong to  $\lambda_j$ .

The work flow of our protein subcellular localization prediction method can be shown in Fig. 1. In the data part, two sources of protein subcellular localization information were integrated. Then we tried three kinds of common features for representing the protein sequences. Multi-label classifier was employed for the prediction. The implementation was done with Mulan<sup>45</sup>, which is an open source machine learning software tool.



**Figure 2.** AP value comparison of 7 different basic classifiers in 188D features.

**Evaluation criteria and measurement.** Average precision (AP)<sup>46</sup>: AP refers to the average accuracy of multi-label classification. This index is positively related to multi-label classification system performance. If AP = 1, the classification effect is the best. The calculation formula of AP is as follows:

$$AP(f) = \frac{1}{N} \sum_{i=1}^N \frac{1}{|y_i|} \sum_{\lambda \in y_i} \frac{|L_i|}{rank_f(x_i, \lambda)} \quad (4)$$

$$L = \{\lambda' | rank_f(X_i, \lambda') \leq rank_f(X_i, \lambda), \lambda' \in y_i\} \quad (5)$$

Here  $N$  is the number of all samples;  $|y_i|$  is the number of the samples with label  $y_i$ ;  $rank(x_i, \lambda)$  means the prediction value (sometimes viewed as probability) of sample  $x_i$  with label  $\lambda$ . We use AP as a primary measure of our comparative experiment.

## Results and Discussion

**Contrast experiments based on 188-dimensional classical features.** Experiment (1): Seven types of multi-labeled base classifiers are used to provide a fivefold cross validation for 188-dimensional feature<sup>18,47</sup> training set. Classification performance is shown in Fig. 2. Detail value is shown in the Table S1 in supplementary materials. We take AP as the main reference indicator, and the AP values of the seven basic classifiers are shown in Fig. 2. The seven types of commonly used base classifiers in the experiment are random forest (RF), decision tree (J48), k nearest neighbor (IBK), logistic regression for multi-label classification (IBLR\_ML)<sup>48</sup>, k nearest neighbor for multi-label classification (MLkNN)<sup>49</sup>, lazy multi-label classification (BRkNN)<sup>50</sup>, and Hierarchy of multi-label learners (HOMER)<sup>51</sup>. The former three classifiers are single-label ones, while the latter four are multi-label classifiers.

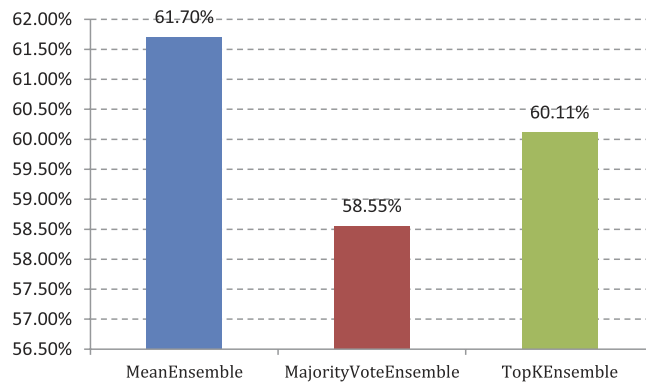
IBLR\_ML achieves the highest AP value of the cross validation (59.37%), whereas HOMER has the lowest value (34.88%). The AP values of RF and IBK are less than 50%. We abandon the above three base classifiers with lower AP values. The four basic classifiers with higher AP values, namely, J48, IBLR\_ML, MLkNN, and BRkNN, are integrated to the classification algorithm in Experiment (2).

Experiment (2): The four basic classifiers retained in Experiment (1) are integrated using our multi-label ensemble classification algorithms. We provide a fivefold cross validation for training sets. The AP values are shown in Fig. 3. Figure 3 demonstrates that the integration effect of MeanEnsemble multi-label ensemble classification algorithm for four types of base classifiers in Experiment (1) is optimal. The AP value is 61.70%.

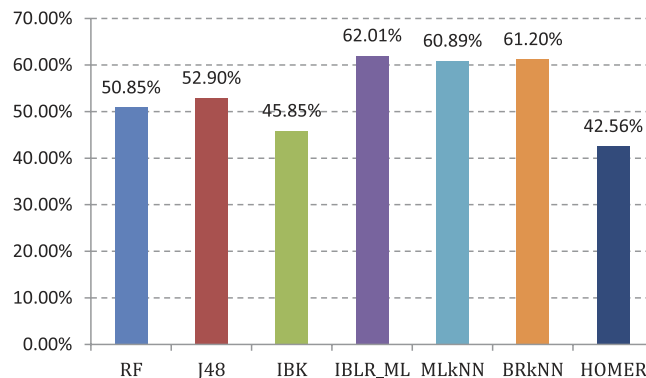
The results of Experiments (1) and (2) show that the ensemble classification algorithm has a significant role in improving the accuracy of protein subcellular localization. We should notice that this is a serious imbalanced classification problem. The classifiers would prefer to the dominating labels. In the Table S4, we showed the detailed performances of individual subcellular locations. In the previous works, all the small classes were combined into a big class. We firstly tried to categorize 37 subcellular structures for prediction. Comparing with previous works, we have applied more subcellular structures and gotten more average accuracy.

**Contrast experiments based on PSSM-20-dimensional feature.** Experiment (3): Seven types of multi-labeled base classifiers are used to provide a fivefold cross validation for PSSM-20-dimensional feature training set. Classification performance is shown in Table S2 in the supplementary materials. Based on Table S2, we conclude that the AP value of fivefold cross validation that corresponds with PSSM-20d is better with better classification results. We still take AP as the main reference indicator, and the AP values of the seven base classifiers are shown in Fig. 4.

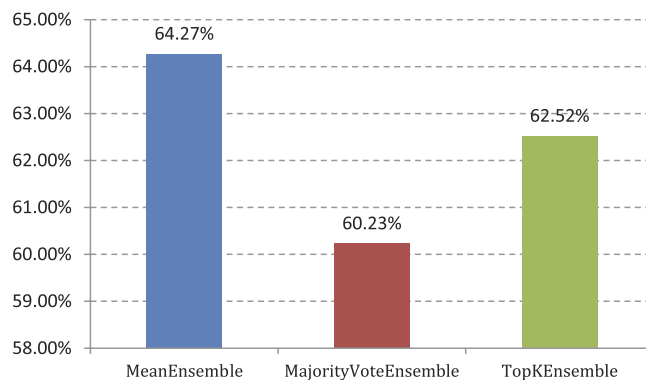
The chart shows that the IBLR\_ML classifier obtains the highest AP value (62.01%). It has improved appropriately compared with the validation result of 188-dimensional feature training set. The rest of the base classifiers' training effects have different degrees of improvement compared with Experiment (1). The four base classifiers with higher AP values, namely, J48, IBLR\_ML, MLkNN, and BRkNN, are integrated to the classification algorithm in Experiment (4).



**Figure 3.** AP value comparison on 3 different ensemble classifier in 188D features.



**Figure 4.** AP value comparison on 7 different basic multi-label classifiers in 20D features.



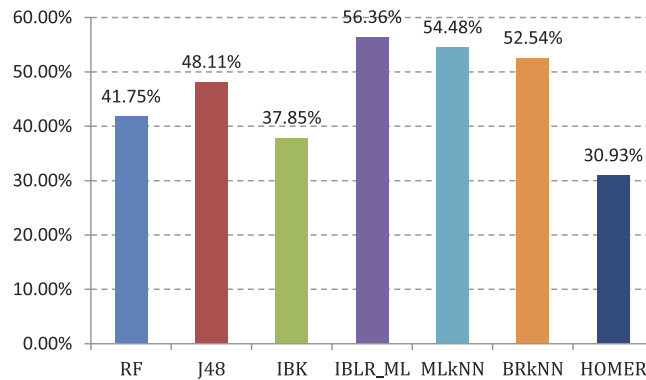
**Figure 5.** AP value comparison on 3 different ensemble multi-label classifiers in 20D features.

Experiment (4): We provide a fivefold cross validation for the training set with the same method as that in Experiment (2). The AP values are shown in Fig. 5.

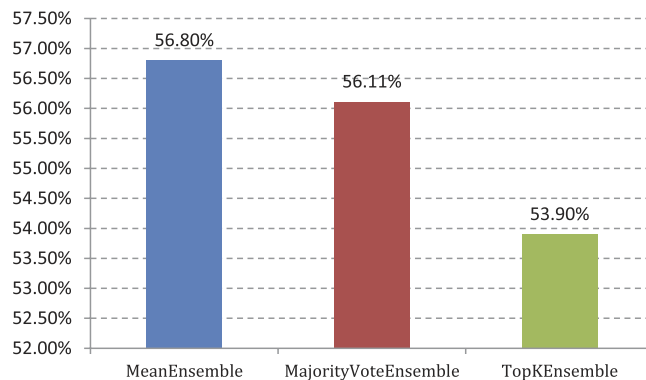
The MeanEnsemble multi-label ensemble classification algorithm is still the best and better than the cross validation results of Experiment (2). The AP value reached 64.27%. TopKEnsemble and MajorityVoteEnsemble algorithms exhibit a larger increase compared with the training results in Experiment (2), but still less than the integrated effect of MeanEnsemble.

The results of Experiments (3) and (4) show that the ensemble classification algorithm has a significant role in improving the accuracy of protein subcellular localization again.

**Contrast experiments based on PseAAC-420-dimensional feature.** Experiment (5): Seven types of multi-labeled base classifiers are used to provide a fivefold cross validation for PseAAC-420-dimensional feature<sup>42</sup> training set. Classification performance is shown in Table S3 in the supplementary materials. From Table S3 we



**Figure 6.** AP value comparison on 7 different basic multi-label classifiers in 420D features.



**Figure 7.** AP value comparison on 3 different ensemble multi-label classifiers in 420D features.

Methods	Average Precision
IMMMLGP	0.5725
Hum-mPLOC 2.0	0.5644
mGOF-loc	0.582
Our method(188D features)	<b>0.5937</b>
Our method(20D features)	<b>0.6201</b>

**Table 4.** Accuracy comparison with state-of-the-art methods.

can see that the AP values of fivefold cross validation that correspond with PseAAC-420d decline compared with 188d. The AP value of IBLR\_ML is 56.36%, which is still the highest. It declines 3.01% and 5.65% compared with Experiments (1) and (3), respectively. The cross validation results are shown in Fig. 6.

The chart shows that the cross validation results of PseAAC-420-dimensional feature training set are the worst. The training results of the seven types of base classifiers decline compared with Experiments (1) and (3).

Experiment (4): We provide a fivefold cross validation for the training set with the same method as that in Experiment (4). The AP values are shown in Fig. 7.

**Comparison with state-of-the-art methods.** In order to prove the performance of our method, we compared with the latest protein subcellular localization web servers, including IMMMLGP<sup>28</sup>, Hum-mPLOC 2.0<sup>40</sup>, mGOF-Loc<sup>52</sup>. The first one is a multi-label classifier, while the other two can only predict as single class. So we employ  $D_{RM}$  for the multi-label classification and  $D_{RS}$  for single-label classification. Since there are both multi-label and single-label classifiers, we cannot compare in the multi-label measurements, including Macro-averaged Precision, Micro-averaged Precision, Macro-averaged F-Measure, and Micro-averaged F-Measure. We just compare the average accuracy in the testing dataset. Table 4 showed the performance comparison in accuracy. From Table 4 we can see that our method outperformed the other latest methods. All of the accuracy rates come from 10-fold cross validation.

Besides that, we also tested our methods on other species, including plant, virus, eukaryote, and animal. Related datasets and performance were shown in Table S5 and S6 in the supplementary materials. We concluded



that our methods can also work on other species. But the performances were all poorer than human dataset. It is due to our integrated human protein subcellular localization dataset is more complete than other species. We will continue to collect the other species protein subcellular localization data in the future.

**Experiments analysis and discussion.** We compare and analyze the training results of Experiments (1), (3), and (5) and Experiments (2), (4), and (6).

First, the seven cross validation results that correspond to PSSM-20-dimensional feature training set are better than the other two feature extraction algorithms. The IBLR\_ML-based classifier shows the best performance, with the highest AP value of 62.01%. The contrast experimental results show that cross validation effects of PSSM-20 dimensional feature training set is the best for the base classifier.

Second, the cross validation results of MeanEnsemble, TopKEnsemble, and MajorityVoteEnsemble on PSSM-20-dimensional feature training set are higher than those of 188d and PseAAC-420d. The advantages of PSSM-20d in multi-label ensemble classification are shown.

By comparing the experimental results of the two groups, we conclude that the 20-dimensional feature extraction algorithm based on the PSSM is the most effective for protein subcellular localization.

Then we compare and analyze the training results of Experiments (3) and (4). Based on the integrated effect, the algorithm MeanEnsemble effect is the best, with an AP value of 64.27%, which is higher than predicting AP of any type of base classifier. The algorithm performance of MajorityVoteEnsemble is the worst, with an AP value fivefold cross training of only 60.23%. This value is lower than the multi-label classification results of the base classifiers IBLR\_ML, BRkNN, and MLkNN with the same background data set, not embodying out the superiority of the integrated thought. It will be time consuming. By comparing the experimental results, we conclude that the multi-label classifier ensemble algorithm MeanEnsemble achieves the best effect for PSSM-20-dimensional feature training set. In the integrated four base classifiers, IBLR\_ML shows the best multi-label learning performance.

## Conclusion

Protein subcellular localization with computational methods is a multi-label classification problem. State-of-the-art prediction methods employ traditional single label machine learning. We proposed novel multi-label ensemble classification techniques with novel hybrid protein features. Experiments proved the effectiveness of our features and the ensemble strategy. Several recent works have proved that ensemble learning<sup>53</sup> and feature reduction<sup>54</sup> can improve the performance of weak learning problems. However, the present work employed the simplest voting strategy and did not conduct any feature reduction techniques. Moreover, class imbalance occurred in protein subcellular localization problems. Imbalance learning for binary classification has been developed and applied in bioinformatics research<sup>55,56</sup>. However, no imbalance learning techniques exist for multi-class and multi-label classification. All these problems and application on large data<sup>57</sup> would be investigated in future work.

## References

1. LaQuaglia, M. J. *et al.* YAP protein expression and subcellular localization in pediatric liver tumors. *CANCER RES* **75**, 2107–2107 (2015).
2. Huh, W.-K. *et al.* Global analysis of protein localization in budding yeast. *NATURE* **425**, 686–691 (2003).
3. Maliepaard, M. *et al.* Subcellular localization and distribution of the breast cancer resistance protein transporter in normal human tissues. *CANCER RES* **61**, 3458–3464 (2001).
4. Camp, R. L., Chung, G. G. & Rimm, D. L. Automated subcellular localization and quantification of protein expression in tissue microarrays. *NAT MED* **8**, 1323–1328 (2002).
5. Gardy, J. L. & Brinkman, F. S. Methods for predicting bacterial protein subcellular localization. *NAT REV MICROBIOL* **4**, 741–751 (2006).
6. Wang, Z., Zou, Q., Jiang, Y., Ju, Y. & Zeng, X. Review of protein subcellular localization prediction. *CURR BIOINFORM* **9**, 331–342 (2014).
7. Sprenger, J. *et al.* LOCATE: a mammalian protein subcellular localization database. *NUCLEIC ACIDS RES* **36**, D230–D233 (2008).
8. Rey, S. *et al.* PSORTdb: a protein subcellular localization database for bacteria. *NUCLEIC ACIDS RES* **33**, D164–D168 (2005).
9. Li, S., Ehrhardt, D. W. & Rhee, S. Y. Systematic analysis of Arabidopsis organelles and a protein localization database for facilitating fluorescent tagging of full-length Arabidopsis proteins. *PLANT PHYSIOL* **141**, 527–539 (2006).
10. Kumar, A. *et al.* Subcellular localization of the yeast proteome. *GENE DEV* **16**, 707–719 (2002).
11. Chou, K. C. & Shen, H. B. Large-scale plant protein subcellular location prediction. *J CELL BIOCHEM* **100**, 665–678 (2007).
12. Nair, R. & Rost, B. LOCnet and LOCtarget: sub-cellular localization for structural genomics targets. *NUCLEIC ACIDS RES* **32**, W517–W521 (2004).
13. Nair, R. & Rost, B. LOC3D: annotate sub-cellular localization for protein structures. *NUCLEIC ACIDS RES* **31**, 3337–3340 (2003).
14. Guo, T., Hua, S., Ji, X. & Sun, Z. DBSubLoc: database of protein subcellular localization. *NUCLEIC ACIDS RES* **32**, D122–D124 (2004).
15. Lu, P. *et al.* PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization. *NUCLEIC ACIDS RES* **33**, D147–D153 (2005).
16. Du, P., Wang, X., Xu, C. & Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *ANAL BIOCHEM* **425**, 117–119 (2012).
17. Liu, B. *et al.* Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *J THEOR BIOL* **385**, 153–159 (2015).
18. Cai, C., Han, L., Ji, Z. L., Chen, X. & Chen, Y. Z. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *NUCLEIC ACIDS RES* **31**, 3692–3697 (2003).
19. Wei, L., Liao, M., Gao, X. & Zou, Q. An Improved Protein Structural Prediction Method by Incorporating Both Sequence and Structure Information. *IEEE T NANOBIOSCI* **14**, 339–349 (2015).
20. Wei, L., Liao, M., Gao, X. & Zou, Q. Enhanced Protein Fold Prediction Method through a Novel Feature Extraction Technique. *IEEE T NANOBIOSCI* **14**, 649–659 (2015).
21. Xu, R. *et al.* Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC SYST BIOL* **9**, S10 (2015).

22. Liu, B. *et al.* Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *NUCLEIC ACIDS RES* **43**, W65–W71 (2015).
23. Xiao, N., Cao, D. S., Zhu, M. F. & Xu, Q. S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *BIOINFORM* **31**, 1857–1859 (2015).
24. Shen, H.-B. & Chou, K.-C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *ANAL BIOCHEM* **373**, 386–388 (2008).
25. Zhao, X., Zou, Q., Liu, B. & Liu, X. Exploratory predicting protein folding model with random forest and hybrid features. *CURR PROTEOMICS* **11**, 289–299 (2014).
26. Zou, Q. *et al.* Improving tRNAscan-SE annotation results via ensemble classifiers. *MOL INFORM* **34**, 761–770 (2015).
27. Wang, C., Hu, L., Guo, M., Liu, X. & Zou, Q. imDC: an ensemble learning method for imbalanced classification with miRNA data. *GENET MOL RES* **14**, 123–133 (2015).
28. He, J., Gu, H. & Liu, W. Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites. *Plos one* **7**, e37155 (2012).
29. Mei, S. Multi-label multi-kernel transfer learning for human protein subcellular localization. *Plos one* **7**, e37716 (2012).
30. Wu, Z.-C., Xiao, X. & Chou, K.-C. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *MOL BIOSYST* **7**, 3287–3297 (2011).
31. Xiao, X., Wu, Z.-C. & Chou, K.-C. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J THEOR BIOL* **284**, 42–51 (2011).
32. Wang, X., Li, G.-Z. & Lu, W.-C. Virus-ECC-mPLoc: a multi-label predictor for predicting the subcellular localization of virus proteins with both single and multiple sites based on a general form of Chou's pseudo amino acid composition. *PROTEIN PEPTIDE LETT* **20**, 309–317 (2013).
33. Chou, K.-C., Wu, Z.-C. & Xiao, X. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *Plos one* **6**, e18258 (2011).
34. Wang, X. & Li, G.-Z. A multi-label predictor for identifying the subcellular locations of singleplex and multiplex eukaryotic proteins. *Plos one* **7**, e36317 (2012).
35. Lin, W.-Z., Fang, J.-A., Xiao, X. & Chou, K.-C. iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *MOL BIOSYST* **9**, 634–644 (2013).
36. Xiao, X., Wu, Z.-C. & Chou, K.-C. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. *Plos one* **6**, e20592 (2011).
37. Mei, S. Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. *J THEOR BIOL* **310**, 80–87 (2012).
38. Wan, S., Ma, M.-W. & Kung, S.-Y. mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC BIOINFORM* **13**, 1 (2012).
39. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *BIOINFORM* **28**, 3150–3152 (2012).
40. Shen, H.-B. & Chou, K.-C. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *ANAL BIOCHEM* **394**, 269–274 (2009).
41. Chou, K.-C. & Shen, H.-B. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *BIOCHEM BIOPH RES CO* **360**, 339–345 (2007).
42. Chou, K. C. Prediction of protein cellular attributes using pseudo-amino acid composition. *PROTEIN: STRUC, FUNC, & BIOINFORM* **43**, 246–255 (2001).
43. Zou, Q., Hu, Q., Guo, M. & Wang, G. HAlign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy. *BIOINFORM* **31**, 2475–2481 (2015).
44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J MOL BIOL* **215**, 403–410 (1990).
45. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J. & Vlahavas, I. MULAN: A Java library for multi-label learning. *J MACH LEARN RES* **12**, 2411–2414 (2011).
46. Zhou, Z.-H., Zhang, M.-L., Huang, S.-J. & Li, Y.-F. Multi-instance multi-label learning. *ARTIF INTELL* **176**, 2291–2320 (2012).
47. Lin, C. *et al.* Hierarchical classification of protein folds using a novel ensemble classifier. *Plos one* **8**, e56499 (2013).
48. Cheng, W. & Hüllermeier, E. Combining instance-based learning and logistic regression for multilabel classification. *MACH LEARN* **76**, 211–225 (2009).
49. Zhang, M.-L. & Zhou, Z.-H. ML-KNN: A lazy learning approach to multi-label learning. *PATTERN RECOGN* **40**, 2038–2048 (2007).
50. Spyromitros, E., Tsoumakas, G. & Vlahavas, I. An empirical study of lazy multilabel classification algorithms. In *AI:THE, MOD & APP* 401–406 (Springer, 2008).
51. Tsoumakas, G., Katakis, I. & Vlahavas, I. Effective and efficient multilabel classification in domains with large number of labels. In *Proc. ECML/PKDD 2008 MMD'08*. 30–44.
52. Wei, L., Liao, M., Gao, X., Wang, J. & Lin, W. *mGOF-Loc: A Novel Ensemble Learning Method for Human Protein Subcellular Localization Prediction*. (2016) Available at: <http://server.malab.cn/mGOF-loc/Index.html> (Accessed: 5th May 2016).
53. Lin, C. *et al.* LibD3C: ensemble classifiers with a clustering and dynamic selection strategy. *NEUROCOMP* **123**, 424–435 (2014).
54. Zou, Q., Zeng, J., Cao, L. & Ji, R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *NEUROCOMP* **173**, 346–354 (2016).
55. Song, L. *et al.* nDNA-prot: Identification of DNA-binding Proteins Based on Unbalanced Classification. *BMC BIOINFORM* **15**, 298 (2014).
56. Zou, Q., Xie, S., Lin, Z., Wu, M. & Ju, Y. Finding the best classification threshold in imbalanced classification. *BIG DATA RES*, doi: 10.1016/j.bdr.2015.12.001 (2016).
57. Zou, Q. *et al.* Survey of MapReduce Frame Operation in Bioinformatics. *BRIEF BIOINFORM* **15**, 637–647 (2014).

## Acknowledgements

This work was supported by the Natural Science Foundation of China (No. 61402132).

## Author Contributions

X.T.G. initially drafted the manuscript and did most of the codes work. F.L.L. helped to collect the protein localization data. Y.J. helped to revise the English. Z.W. participated in the design of the experiments. C.Y.W. guided the whole works and helped to draft the manuscript. All authors read and approved the final manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Guo, X. *et al.* Human Protein Subcellular Localization with Integrated Source and Multi-label Ensemble Classifier. *Sci. Rep.* **6**, 28087; doi: 10.1038/srep28087 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>