Evaluation of Database Annotation to Determine Human Mitochondrial Proteins

Katsuhiko Murakami[1*], Masaharu Sugita[2]

[1] Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.
[2] School of Bioscience and Biotechnology, Tokyo University of Technology, 1404-1 Katakuramachi, Tokyo 192-0982, Japan.

* Corresponding author. Tel.: +81-3-5449-5131; email: murakami.ktk@gmail.com
Manuscript submitted January 9, 2018; accepted February 28, 2018.
doi: 10.17706/ijbbb.2018.8.4.210-217

Abstract: Subcellular localization can be a helpful indication of the function of an unknown protein. Among the reported human mitochondrial proteins, hundreds of proteins have still not been functionally confirmed. To date, several databases for such proteins have been developed; however, their annotations overlap incompletely. A key issue in the completion of a reliable catalog of mitochondrial proteins is the integration of all this information, and the evaluation of the influence of different forms of evidence is also important. Here, we integrated various pieces of evidence (features) from both experimental and computational analyses. Linear and nonlinear prediction models were examined to predict human mitochondrial proteins. By employing a random forest model, an F-score of 0.929 was achieved by cross validation. The contributions of individual features toward the accurate prediction of localization were evaluated. We found only minor differences in importance among different features, with accurate prediction requiring the combination of many features; however, evidence from mass spectrometry experiments emerged as a prominent feature. Focusing on human mitochondrial proteins, we have constructed a high-accuracy prediction model that utilizes many weak features. Evaluation of the importance of individual features provides insights into what information is most valuable for the confirmation of protein localization.

Key words: Database, machine learning, mitochondria, localization, random forest.

1. Introduction

The subcellular localization of a protein can be an important clue towards understanding its function. To determine whether a protein localizes to a specific organelle, one can use both *in silico* and experimental information. Results from both types of studies are annotated in databases such as Uniprot [1]. However, single pieces of evidence are insufficient to determine whether the protein is really localized to an organelle, as both approaches can produce errors, such as false negatives and experimental contamination. Therefore, users must be careful when they interpret localization annotation.

Many computational methods for the prediction of subcellular localization have been developed [2]–[4], with recent studies focusing on multiple subcellular localizations [4]–[10]. However, in many cases, the prediction is based solely on sequence information. Some tools use additional Gene Ontology (GO) data regarding similar proteins, obtained by searching for similarity with the input sequence [11]. Only a few studies [12] have evaluated how experimental results affect the reliability of subcellular localization prediction. These experimental data include not only strong evidence that proves the protein to be

mitochondrial but also weak evidence that does not suffice as the proof. This inconclusive situation is especially true for the data obtained from high-throughput experiments. Therefore, assessing the reliability of the subcellular localization of a protein with such different levels of evidence is problematic. This is often the case with many proteins that are not well studied.

The mitochondrion is an important organelle that hosts several biological processes, including energy generation, cell cycle control, and apoptosis. Mutations in proteins localized in the mitochondria can cause mitochondrial dysfunction and are often related to severe diseases [13]. Therefore, mitochondria have been identified as "hot spots" for dysfunctions, and have gained the attention of many researchers. However, many mitochondrial proteins have yet to be conclusively identified. An estimated 1,500 human proteins are mitochondrial [14]. Several recent efforts to develop databases of the human mitochondrial proteome include MitoCarta [15], [16], MitoProteome [17], and MitoMiner [18]–[20].

MitoCarta [15] is an inventory of mammalian mitochondrial proteins, containing data from humans and mice. For its reference set, it integrates information from literature curation, engineered ascorbate peroxidase (APEX)-based mass spectrometry, and green fluorescent protein (GFP)-tagging/microscopy. In addition, the database includes information obtained by seven different methods, such as tandem mass spectrometry (MS/MS)-based proteomics, yeast homology, co-expression, protein domains, targeting signals from the TargetP predictor [2], ancestry information, and induction signals. The reference set of human mitochondrial proteins was constructed by compiling a conservative set, and adding newly predicted proteins using a naïve Bayes classifier. The naïve Bayes model showed 79% sensitivity and 99.7% specificity at a 5% false discovery rate (FDR) threshold. The naïve Bayes model assumes that all features (pieces of evidence) are independent after conditioning of the class, such as proteins being imported into mitochondria; the model does not consider the dependence of related features. To achieve more accurate prediction using more complex models, nonlinear prediction models are required to explore feature codependency.

MitoProteome is a database that includes human mitochondrial protein sequences from public databases with literature curation, MS/MS studies, Entrez, the Kyoto Encyclopedia of Genes and Genomes (KEGG), Online Mendelian Inheritance in Man (OMIM), the Molecular INTeraction database (MINT), the Database of Interacting Proteins (DIP), PFAM, InterPro, and PRINTS, with a total of 1,705 genes and 3,625 proteins as of 2016. It provides a reference set but lacks many authentic mitochondrial proteins.

MitoMinor integrates various resources, including MitoCarta, and relevant annotation from HomoloGene, GO, MS/MS data, GFP data, KEGG, and OMIM. Version 3.1 also includes the Human Protein Atlas [21], which uses antibodies to annotate protein localization by immunofluorescence. This method can suffer from cross reactivity and staining failure, and therefore, it does not provide complete information on subcellular localization. Several computational prediction tools based on protein sequence were also included: iPSORT [22], TargetP [2], [17] MitoProt [23], and MitoFates [24]. MitoMinor does not provide a mitochondrial protein reference set. As alternatives, it contains sets from the Integrated Mitochondrial Protein Index (IMPI), although the method to determine the set is not described.

These data sources need to be integrated in such a way that each individual piece of evidence contributes to the determination of mitochondrial proteins. Although some of the databases provide integrative scores, such as the IMPI score in MitoCarta, some relevant features are omitted.

Here, we have investigated the possibility of predicting mitochondrial proteins using individually weak pieces of evidence from both *in silico* and experimental results. We show how, in combination, these weak pieces of evidence are useful for accurate discrimination. The results indicate which types of analyses should be conducted to increase the reliability of localization annotation. In this study, we refer to pieces of evidence as "features" according to the practice of the machine learning research field.

2. Results

2.1. Comparison of Several Datasets of Human Mitochondrial Proteins

Of all the human proteins, how many are reliably identified as mitochondrial proteins? To answer this question, we examined the overlap between proteins annotated in different datasets. We collected reference datasets of human mitochondrial proteins from MitoMiner (IMPI), MitoCarta, and MitoProteome. Figure 1 shows a Venn diagram of the reference protein sets.

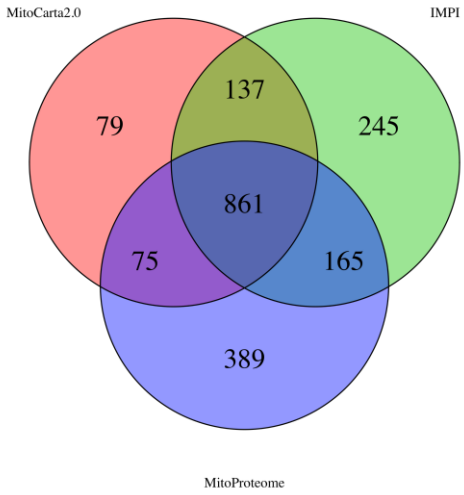

Fig. 1. Venn diagram of human mitochondrial genes in the three datasets.

While the union of the datasets includes 1,951 proteins, only 861 (approximately 44%) were shared amongst the three datasets. The other 66% of the proteins are not commonly recognized as mitochondrial proteins. This might be because most of them have only weak evidence.

2.2. Construction of a Dataset with Various Features

To build an integrated dataset with many features, we selected MitoMiner and MitoCarta 2.0, both of which contain many types of raw data obtained by different experimental and *in silico* methods. In addition, these datasets were downloaded, while others required extensive Web browsing. MitoMiner includes partial information from MitoCarta; these overlapping features were deleted from the integrated data. In combining the two datasets, we joined them with the Ensembl gene ID, which was assigned to relatively more genes than the other IDs were, such as NCBI Gene ID and HGNC. We also discarded irrelevant features that are inappropriate for prediction, such as extra IDs, ambiguous scores, and integrated scores from the developers. Table 1 lists the 34 features used for prediction. The features with ID numbers 7–27 are related to mass spectrometry. The features with ID numbers 29–33 are associated with predictive scores from sequence analysis programs.

After data cleaning for missing values and categorical data conversion, all values were changed to numerical values. We used the Tmito (mitochondrial protein positive, 960 proteins) and Tnonmito (mitochondrial protein negative, 17,468 proteins) datasets in MitoCarta, which were used as training sets for the Maestro score in MitoCarta.

Table 1. List of Features Used in the Study

No.	Category	Feature name	Source
1	Ancestry	RickettsiaHomolog	MitoCarta
2	Coexpression	coExpression	MitoCarta
3	Domain	Domain	MitoCarta

4	GFP	mmGFP	MitoMiner
5	Immuno fluorescence	mmHPAtlas	MitoMiner
6	Induction	PGC	MitoCarta
7	MS spectroscopy score	mmMassEx	MitoMiner
8	MS spectroscopy score	adiposePeak	MitoCarta
9	MS spectroscopy score	brainstemPeak	MitoCarta
10	MS spectroscopy score	cerebellumPeak	MitoCarta
11	MS spectroscopy score	cerebrumPeak	MitoCarta
12	MS spectroscopy score	heartPeak	MitoCarta
13	MS spectroscopy score	kidneyPeak	MitoCarta
14	MS spectroscopy score	largeintestinePeak	MitoCarta
15	MS spectroscopy score	liverPeak	MitoCarta
16	MS spectroscopy score	MSNumPep	MitoCarta
17	MS spectroscopy score	MSNumSpectra	MitoCarta
18	MS spectroscopy score	MSNumTissue	MitoCarta
19	MS spectroscopy score	MSPercent	MitoCarta
20	MS spectroscopy score	MSTotIntensity	MitoCarta
21	MS spectroscopy score	placentaPeak	MitoCarta
22	MS spectroscopy score	skeletalmusclePeak	MitoCarta
23	MS spectroscopy score	smallintestinePeak	MitoCarta
24	MS spectroscopy score	spinalcordPeak	MitoCarta
25	MS spectroscopy score	stomachPeak	MitoCarta
26	MS spectroscopy score	testisPeak	MitoCarta
27	MS spectroscopy score	MS	MitoCarta
28	Sequence length	ProtLen	MitoCarta
29	Sequence prediction	mcTargetP	MitoCarta
30	Sequence prediction	mmIpsort	MitoMiner
31	Sequence prediction	mmMitofates	MitoMiner
32	Sequence prediction	mmMitoprot	MitoMiner
33	Sequence prediction	mmTargetP	MitoMiner
34	Yeast Homolog	YeastHomolog	MitoCarta

2.3. Prediction from Various Features

Many models could be examined to accurately predict protein localization. However, we also aimed to evaluate individual features (variables or pieces of evidence) for prediction. For this, we investigated a random forest model [25] and a logistic regression model with regularization (or penalty). After building the models, we evaluated both the accuracy of each model and the importance of individual features in each model simultaneously. Table 2 shows the performance of each prediction model.

Table 2. Performances of the Prediction Models

Models	Precision	Recall	F-score
Random forest	0.983	0.881	0.929
LR with L1	0.975	0.875	0.922
LR with L2	0.975	0.877	0.924
MitoCarta 2.0 (2016), naïve Bayes	0.997	0.79	0.885

The random forest model achieved an F-score of 0.929 (precision, 0.983; recall, 0.881). For logistic regression, the F-score was 0.922 (precision, 0.975; recall, 0.875) for a model with an L1 penalty, and 0.924 for a model with an L2 penalty. The different regularization methods (L1, L2) showed negligible differences in performance. One reason for this was that the penalty weights became small at the optimization by the grid search stage. We compared the performance of our models with the naïve Bayes model used to provide an integrated score by the MitoCarta database. The naïve Bayes model had an F-score of 0.885 [16], suggesting that our models outperformed the MitoCarta model.

2.4. Feature Importance Analysis

We investigated which features contributed the most to the predictions. Fig. 2 shows the importance of individual features in the random forest model. The importance was calculated as the increase in the error rate if the feature value was randomly permutated between cases. Then, the differences for all the features were normalized. With this model, "MS" (mass spectroscopy-based score) was the prominent feature for prediction.

Fig. 3 shows the coefficients of the features in the logistic regression models. Not surprisingly, the highest ranked features were similar to those observed with the random forest model. Interestingly, the top-ranked feature "Domain" ranked 13th in the random forest model (Fig. 2).

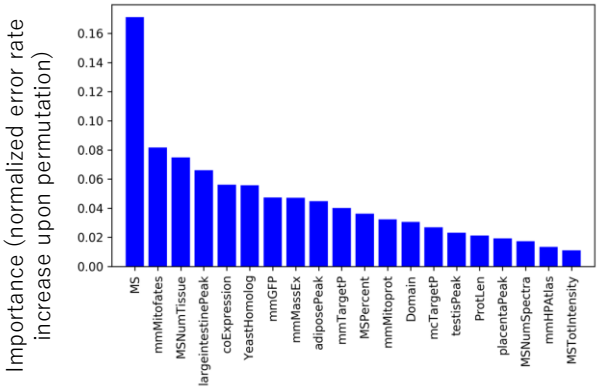

Fig. 2. Importance of the individual features in the random forest model.

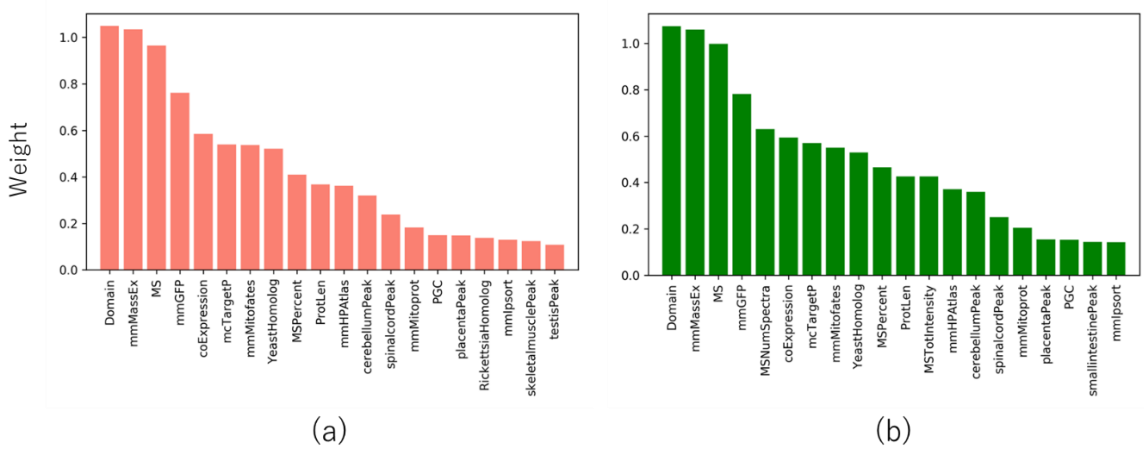

Fig. 3. Coefficients of the features in the logistic regression model: (a) the model with L1 penalty, and (b) the model with L2 penalty.

3. Discussion

In this study, we have attempted to integrate disparate data in multiple datasets to predict a reliable list of human mitochondrial proteins. Both our models performed well compared to MitoCarta. The random forest model, which can better handle dependency among features, performed the best. The possible reasons for the improvement are twofold: the information richness by the variety of features integrated from multiple sources, and the goodness of the model that can combine the dependent features. The two logistic regression models were also better than MitoCarta but slightly worse than the random forest model.

Although the logistic regression model is actually a nonlinear model, the judgement by the model substantially depends on the sum of weighted variables. Therefore, the variables with higher coefficients of the model are expected to have a more positive correlation with the dependent variable or with the probability of a protein being a mitochondrial protein.

The prediction models provide some important suggestions. First, as shown in Fig. 2, the information of mass spectrometry "MS" was the most important among the various features in the random forest model. The second-best feature was the *in silico* sequence-based prediction "mmMitofates." From this feature, the importance decreased gradually, suggesting that these features would not provide stronger support to a protein being a mitochondrial protein than the first "MS" feature. For the logistic regression model in Fig. 3, there are three top features, ranking much higher than the others. The first was Pfam "Domain" information, which was detected from sequence analysis. Of the three, two were mass spectrometry-based features (the second "mmMassEx" and the third "MS"), indicating the increased effectiveness of proteomic experiments for subcellular localization compared to other methods, at least for mitochondria. Accordingly, experimental analysts are encouraged to use datasets related to mass spectrometry for future data collection. However, the differences in importance between features were minimal, and weak features can contribute greatly in combination, as indicated by the accuracy of our prediction models.

A limitation of this study relates to the processing of missing values. The current dataset contains some items that were not measured, resulting in missing values. For instance, for the feature "PGC induction score," 1,996 genes have values, but 17,247 genes do not. If we had deleted all proteins with at least one missing value from the total dataset, the remaining dataset would have contained only 127 genes in total. Therefore, it is not appropriate to delete all proteins with any missing values. Instead, we replaced them with their average.

4. Conclusions

To identify useful information for the identification of human mitochondrial proteins, we explored prediction models combining various features from several mitochondria-specific databases. We achieved high prediction performance with both logistic regression and random forest models. Further analysis of the importance of individual features suggests that accurate prediction depends not on individual features but on the combination of many weak features. However, we did identify comparatively important features. "MS" from MitoCarta was the most useful information, and features related to mass spectrometry contributed more than other types. Therefore, experimental research collecting proteomics data is encouraged. Importantly, our method is also applicable to other subcellular localizations.

5. Materials and Methods

5.1. Data Acquisition and Processing

The raw data in MitoMiner and MitoCarta were downloaded as Excel files. For MitoProteome, the list of proteins was manually searched using a Web browser. Most of the categorical variables were ordinal. The ordinal values were converted to appropriate integers according to the orders. There were missing values in the raw data; these were imputed with the averages, as is standard procedure. All the numerical variables were normalized. To integrate table data from multiple sources, we conducted inner join using Ensemble Gene ID as the key.

5.2. Model Construction

To construct predictive models, we utilized the scikit-learn library and associated libraries for the Python programming language and Python version 3. For the random forest model, the number of trees was set to

100. Increasing the number of trees did not significantly improve the accuracy. For the logistic regression model, the weight of the regularization penalty was optimized (C = 100) using 10-fold cross validation. The dataset is available at the URL https://github.com/katsumk/mitochondrialProtein.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 26330342.

References

[1] Legge, D., *et al*. (2015). UniProt: A hub for protein information. Nucleic Acids Res., *43(D1)*, D204–D212.

[2] Emanuelsson, O., Nielsen, H., & Brunak, S. (2000). Predicting subcellular localization of proteins based on their n-terminal amino acid sequence.

[3] Horton, P., *et al*. (2007). WoLF PSORT: Protein localization predictor. *Nucleic Acids Res*., *35(Web Server issue)*, W585-7.

[4] Wan, S., *et al*. (2015). MPLR-Loc: An adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction. *Anal. Biochem.*, *473*, 14–27.

[5] Wan, S., & Zou, Q. (2017). HPSLPred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source.

[6] Simha, R., *et al*. (2015). Protein (multi-)location prediction: Utilizing interdependencies via a generative model. *Bioinformatics*, *31(12)*, i365–i374.

[7] Simha, R., & Shatkay, H. (2013). Protein (multi-)location prediction: Using location inter-dependencies in a probabilistic framework.

[8] Wan, S., *et al*. (2016). Sparse regressions for predicting and interpreting subcellular localization of multi-label proteins. *BMC Bioinformatics*, *17(1)*, 97.

[9] Shen, H. B., & Chou, K. C. (2007). Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochem. Biophys. Res. Commun*., *355(4)*, 1006–1011.

[10] Lin, W.-Z., *et al*. (2013). iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. Biosyst.*, *9(4)*, 634–644.

[11] Wan, S., *et al*. (2013). GOASVM : A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou' s pseudo-amino acid composition. *J. Theor. Biol*., *323*, 40–48.

[12] Calvo, S., *et al*. (2006). Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet*, *38(5)*, 576–582.

[13] Jiang, Y., & Wang, X. (2012). Comparative mitochondrial proteomics : perspective in human diseases. 1–13.

[14] Taylor, S. W., *et al*. (2003). Characterization of the human heart mitochondrial proteome. *J. Proteome Res.*, *21(3)*, 281–286.

[15] Pagliarini, D. J., *et al*. (2008). A mitochondrial protein compendium elucidates complex i disease biology. *Cell*, *134(1)*, 112–123.

[16] Calvo, S. E., *et al*. (2016). MitoCarta2.0: An updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res*., *44(D1)*, D1251–D1257.

[17] Cotter, D., *et al*. (2004). MitoProteome: mitochondrial protein sequence database and annotation system. *Nucleic Acids Res*., *32(Database issue)*, D463–D467.

[18] Smith, A. C., & Robinson, A. J. (2009). MitoMiner, an integrated database for the storage and analysis of mitochondrial proteomics data. *Mol. Cell. Proteomics*, *8(6)*, 1324–1337.

[19] Smith, A. C., *et al*. "MitoMiner: A data warehouse for mitochondrial proteomics data. *Nucleic Acids Res*.,

40(Database issue), D1160-7. (Jan. 2012).
[20] Smith, A. C., & Robinson, A. J. (2016). MitoMiner v3.1, an update on the mitochondrial proteomics database. *Nucleic Acids Res.*, *44(D1)*, D1258-61.
[21] Uhlén, M., *et al*. (2015). Tissue-based map of the human proteome. *Science (80-.).*, *347(6220)*, 1260419–1260419.
[22] Bannai, H., *et al*. (2002). Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, *18(2)*, 298–305.
[23] Claros, M. G., & Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.*, *241(3)*, 779–786.
[24] Fukasawa, Y., *et al*. (2015). MitoFates : Improved prediction of mitochondrial targeting sequences and their cleavage. 1113–1126.
[25] Breiman, L. (2001). Random forests. *Mach. Learn.*, *45(1)*, 5–32.

Katsuhiko Murakami received his B.S. and M.S. from Kyushu University, Fukuoka, Japan, in 1990 and 1992, respectively, and the Ph.D. from Tokyo University in 2008. He worked at the Tokyo University of Technology as an associated professor. He is currently a project researcher at the Institute of Medical Science, University of Tokyo. His research interests include genomics and databases and datamining in medical and biological data.

Masaharu Sugita received his B.S. from the Tokyo University of Technology, Tokyo, Japan, in 2017. His major research interests include machine learning and bioinformatics.