Evolutionary sequence divergence predicts protein sub-cellular localization signals

Yoshinori Fukasawa
Department of Computational Biology
Graduate School of Frontier Sciences
University of Tokyo
Kashiwa, Japan

Ross KK Leung
Hong Kong Bioinformatics Centre
and School of Biomedical Sciences
Chinese University of Hong Kong
Shatin, China

Stephen KW Tsui
Hong Kong Bioinformatics Centre
and School of Biomedical Sciences
Chinese University of Hong Kong
Shatin, China

Paul Horton*
Computational Biology Research Center
Advanced Industrial Science and Technology
Tokyo, Japan
Department of Computational Biology
Graduate School of Frontier Sciences
University of Tokyo
Kashiwa, Japan

Abstract—**Protein sub-cellular localization is a central problem in understanding cell biology and has been the focus of intense research. In order to predict localization from amino acid sequence a myriad of features have been tried: including amino acid composition, sequence similarity, the presence of certain motifs or domains, and many others.**

Surprisingly, sequence conservation of sorting motifs has not yet been employed, despite its extensive use for tasks such as the prediction of transcription factor binding sites.

Here, we flip the problem around, and present a proof of concept for the idea that the *lack* of sequence conservation can be a useful feature for localization prediction.

I. Introduction

Since proper sub-cellular localization is a prerequisite for protein function, there is a high demand for accurate and complete localization annotation of all proteins [1]. Although proteomics data has allowed large scale determination of protein localization for several model organisms[2], [3], no experimental evidence is available for the vast majority of organisms. Strong sequence similarity is a good indicator of identical localization site, but distant similarity is not [4], and thus for many proteins we must rely on computer prediction.

In cells, the localization of proteins is largely determined by “zip-code” like sorting signals, encoded in their amino acid sequence. Unfortunately these sorting signals seem to be only very loosely determined, allowing very diverse sequences, subject to some constraints on their physico-chemical properties [5].

Among those signals, the most well-known sorting signal is the signal peptide of secretory path proteins. A typical signal peptide spans 15-30 amino acids near the N-terminus. Signal peptides typically show three distinct blocks: the n-region containing positively charged residues, the h-region mainly consisting of hydrophobic residues, and the c-region which includes polar uncharged residues and a weakly conserved cleavage motif [6].

Similarly, the targeting signals of mitochondria and chloroplast are also N-terminally coded [5], and cleaved after import to their final location. Like signal peptides, these signals are often poorly conserved and difficult to align properly between orthologs. Although some consensus motif has been reported for mitochondrial targeting signals [7], [8], it is information poor and produces too many false positives to be used for reliable prediction.

To date, an impressive number of methods have been developed for protein sorting (in 2004 a survey already listed dozens of methods employing fifteen broad categories of features [9], from commonly used ones such as amino acid composition [10], [11], [12]. (and many more) to rare categories such as sequence periodicity [13] and mRNA expression level [14]. Of these features amino acid composition, first proposed by Nakashima & Nishikawa [10] is attractive due to simplicity. The significant correlation between amino acid composition and sub-cellular location is partially causative and partially due to indirect effects such as adaption of surface residues to the pH of the protein's localization site [15].

The one feature conspicuously missing from this list has been evolutionary sequence conservation, despite the fact that it has seen extensive use in sequence analysis from the prediction of transcription factor binding sites [16], to functional RNA [17]. Although the conservation of amino acid composition has been employed [18], sequence conservation per se has not – presumably because sorting signals are indeed not well conserved at the sequence level. Here, we propose that instead of looking for sequence conservation of sorting signals, a more effective approach is to exploit their high evolutionary sequence *divergence*.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

In this paper we first describe our dataset of yeast proteins and their orthologs, the divergence and other features we used for classification, and the classifiers we employed. Then, we present a simple statistical feature analysis followed by the performance of the localization prediction for various combinations of features, classifiers and data. Finally, we discuss the limitations of our work and conclude.

II. Dataset

A. Proteins and their localization classes

This study focuses on the prediction of N-terminal sorting signals in the budding yeast *Saccharomyces cerevisiae* (hereafter "*S.cere.*") – the eukaryotic organism with the most complete annotation available regarding protein sub-cellular localization. We focused on the two most common N-terminal sorting signals, the "signal peptide" (which we abbreviate as "SP"), targeting proteins to the endoplasmic reticulum and the "MTS" (Matrix Targeting Signal) which targets proteins to the matrix (inner compartment) of the mitochondria. Although both of these signals reside near the N-terminus, they are thought to be mutually exclusive, with different properties that are effectively discriminated by the cell. Although other types of N-terminal sorting signals exist, for example the PTS2 signal targeting the proteins to the peroxisome [19], the number of proteins using such signals is much smaller than those using the SP or MTS signals.

In this study we choose to leave these less common signals to future work and instead concentrate on three broad localization classes for proteins in *S.cere.*: 1) with SP's, 2) with MTS's, and 3) N-signal-less; of which we gathered 54, 182, and 462 examples respectively. We used UniprotKB/Swiss-Prot ([20]) to assign localization class labels, augmented by MTS containing proteins determined in the proteomics experiment of Vögtle et al. [21]. Because only a small number of SP's have been directly confirmed experimentally, we also included proteins whose SP is inferred by a combination of their localization site and prediction by SignalP [22] (see Discussion for a justification of using prediction results in our dataset). For N-signal-less proteins we used proteins which localize to the cytosol or nucleus (according to UniprotKB/Swiss-Prot annotation).

1) Removing redundant sequences: To avoid a bias in training and accuracy estimation, we used Blastclust 2.2.22 (http://www.ncbinlmnih.gov/BLAST/) to removed redundant sequences with a setting of 20% identity.

B. Orthologs and multiple alignment

We extracted orthologs from the Yeast Genome Order Browser [23]. YGOB includes curated ortholog sets from 11 fungi genomes (*S.cere.*, *S. castellii*, *S. kluyveri*, *K. waltii*, *A. gossypii*, *C.glabrata*, *K. lactis*, *Z. rouxii*, *K. thermotolerans*, *S. bayanus* and *K. polysporus*). For each *S.cere.* protein in our dataset, we obtained its *ortholog multiple sequence alignment* (orthoMSA) by aligning it to its orthologs with the MAFFT program [24], using "LINSI", its most accurate mode. For this study we only included proteins for which an ortholog is listed for each of the 11 species.

Feature name	Quantity
$\mathrm{LD}(i)$	$\bar{H}_{i-10,i+10}$
$N_{\mathrm{raw}}20$	$\bar{H}_{1,20}$
$N_{\mathrm{raw}}40$	$\bar{H}_{1,40}$
$N_{\mathrm{raw}}80\text{-}99$	$\bar{H}_{80,99}$
μ_w	Average of $\bar{H}_{\mathrm{window}}$ for all length w windows
σ_w	Standard deviation of $\bar{H}_{\mathrm{window}}$ for all length w windows
NCdiff	$N_{\mathrm{raw}}20 - N_{\mathrm{raw}}80\text{-}99$
$N20$	$\frac{(N_{\mathrm{raw}}20-\mu_{20})}{\sigma_{20}}$ (z-score normalized)
$N40$	$\frac{(N_{\mathrm{raw}}40-\mu_{40})}{\sigma_{40}}$ (z-score normalized)
$N80\text{-}99$	$\frac{(N_{\mathrm{raw}}80\text{-}99-\mu_{20})}{\sigma_{20}}$ (z-score normalized)

TABLE I
Smoothed entropy derived features are listed. Quantities shaded in grey were not used directly as features.

III. Features for classification

A. Sequence evolutionary divergence score

Our study required assigning a divergence score to each position of each *S.cere.* protein, based on its orthoMSA.

1) Column entropy score: Several measures have been suggested for scoring evolutionary sequence conservation (or conversely divergence) [25], [26]. Here we adopt a simple Shannon entropy based score. The Shannon entropy $H(i)$ of the ith column of the an orthoMSA is defined as:

$$H(i) = -\sum_{j \in A} F(i,j) \lg F(i,j). \tag{1}$$

where A denotes the set of 20 amino acid characters plus gap characters, and $F(i,j)$ denotes the frequency of character j in the i column of an orthoMSA. Note that when multiple gap characters present in a column, we consider each to be a unique character. For example, the entropy of an orthoMSA column '{L, L, I, -, -}' is computed as one character (the 'L') with frequency 0.4 and three characters with frequency 0.2, because we treat the two '-' characters as distinct. We adopted this treatment of gap characters so that the divergence of orthoMSA columns with many gaps would be considered high. Since we use 11 species, the range of our column divergence score runs from 0 (perfect conservation) to 3.46 bits (maximally diverged).

2) Smoothed entropy score: For many orthoMSA's, the entropy often varies widely from column to column, therefore as a measure of divergence, we adopted a smoothed entropy score, $\bar{H}_{i,j}$, defined as the average entropy score for columns in the interval $[i,j]$.

3) Divergence based features: We employed several smoothed entropy score based features such as the "local divergence" of a position i, which we define as $\mathrm{LD}(i) \equiv \bar{H}_{i-10,i+10}$. These features are summarized in table I.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

B. *Physico-chemical propensities*

To explore the possibility of combining sequence divergence with standard features used in protein localization prediction, we defined three features computed from the first 30 N-terminal residues of each *S.cere.* protein: 1) the number of positively charged residues (#pos), 2) the number of negatively charged residues (#neg), and 3) the average hydrophobicity as measured by the Kyte-Doolittle [27] index (Hphob).

IV. Classifiers

A. *Majority Class Classifier*

The majority class classifier unconditionally predicts all examples to belong to the most common class. Its accuracy is equal to the fraction of examples belonging to the most common class.

B. *J48*

J48 is a version of the C4.5 decision tree induction algorithm of Quinlan [28], implemented in the Weka software package [29]. We used the default value of 0.25 for the confidence factor, which controls the complexity of the induced tree.

C. *Support Vector Machine*

The SVM [30] is perhaps the most popular classifier in current bioinformatics work. In its basic form it is a linear, binary classifier, but it has been extended to non-linear, multiclass classification. In this project, we used the LIBSVM implementation [31]. We used the Gaussian radial basis kernel function with default γ value (1.0 / # number of features). We also used the default value (1.0) for the SVM cost parameter C. In our study we conducted binary and 3-class classification. For multiclass discrimination LIBSVM adopts the "one-versus-one" method, in which a separate SVM is learned for each pair of classes, and majority voting amongst those SVM's is used when classifying examples.

D. *Quantifying feature importance*

We used the so called "F-score" to quantify the importance of each features. The F-score [32] is a simple measure of the predictive power of a feature in isolation (i.e. without consideration of its relationship to other features), defined as:

$$\frac{(\bar{x}^{(+)}-\bar{x})^2+(\bar{x}^{(-)}-\bar{x})^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_k^{(+)}-\bar{x}^{(+)})^2+\frac{1}{n_--1}\sum_{k=1}^{n_-}(x_k^{(-)}-\bar{x}^{(-)})^2} \qquad (2)$$

where $\bar{x}^{(+)}$, $\bar{x}^{(-)}$, and $\bar{x}$ are the mean values of the feature for the positive, negative and combined examples respectively; while $x_k^{(+)}$ and $x_k^{(-)}$ denote the value of the kth positive and negative examples respectively. A larger F-score indicates greater predictive power.

E. *Classification performance evaluation*

Accuracy is not always an effective measure of performance for skewed datasets (i.e. datasets with a very uneven number of examples from different classes) [33]. Therefore we report several measures in addition to accuracy.

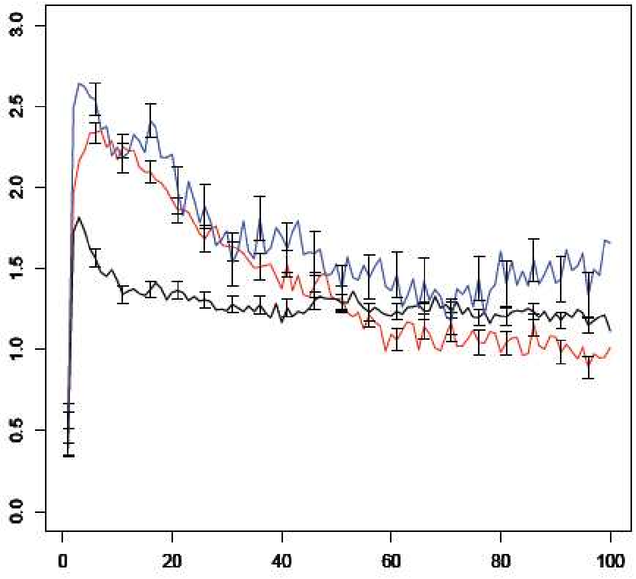

Fig. 2. Local divergence scores are shown for the 100 residue N-terminal region for MTS containing (red), SP containing (blue), and N-signal-less (black) proteins. The error bars denote the standard error. For clarity, error bars are only shown for every fifth position.

1) Matthews correlation coefficient: The Matthews correlation coefficient, MCC [34], is a measure of performance for binary classification defined as follows:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \qquad (3)$$

where 'T' and 'F' stand for "true" and "false", while "N" and "P" stand for "negative" and "positive". Equivalently MCC can be defined as the Pearson's correlation coefficient of the binary vector of class labels compared to the binary vector of predicted class labels. MCC ranges from 1.0 for perfect prediction to -1.0 for perfect inverse prediction. Note that the MCC for the majority class classifier is identically zero, as is the expected value of MCC for random prediction.

V. Results

A. *Feature Analysis*

1) N-terminal sorting signals are evolutionary divergent: It is well known that sorting signals, especially signal peptides, have very low sequence conservation [35]. As shown in Figure 1, this phenomenon is particularly clear for the mitochondrial heat shock protein, mtHSP70, in which main part of the protein is highly conserved but the N-terminal region is highly divergent. Figure 2 quantifies this trend for the proteins in our dataset.

2) Estimate of importance of each feature: As a rough estimate of feature importance, we computed the F-score for each feature (Figure 3). The two highest scoring features are the physico-chemical features #neg and Hphob, but the LD

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

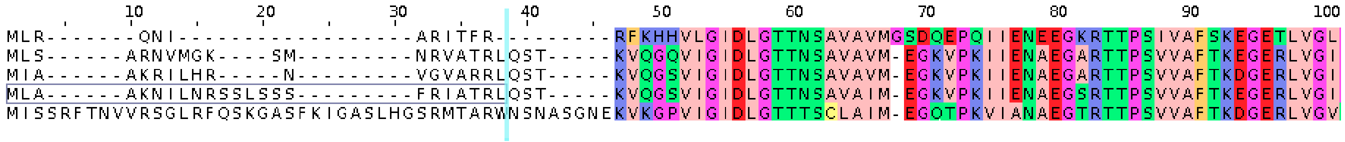
10
20
30
40
50
60
70
80
90
100
MLR-------QNI------------------ARITFR---------RFKHHVLGIDLGTTNSAVAVMGSDQEPQIIENEEGKRTTPSIVAFSKEGETLVGL
MLS------ARNVMGK----SM---------NRVATRLQST-----KVQGQVIGIDLGTTNSAVAVM-EGKVPKIIENAEGARTTPSVVAFTKEGERLVGI
MIA------AKRILHR-----N---------VGVARRLQST-----KVQGSVIGIDLGTTNSAVAVM-EGKVPKIIENAEGARTTPSVVAFTKDGERLVGI
MLA------AKNILNRSSLSSS---------FRIATRLQST-----KVQGSVIGIDLGTTNSAVAIM-EGKVPKIIENAEGSRTTPSVVAFTKEGERLVGI
MISSRFTNVVRSGLRFQSKGASFKIGASLHGSRMTARWNSNASGNEKVKGPVIGIDLGTTTSCLAIM-EGQTPKVIANAEGTRTTPSVVAFTKDGERLVGV

Fig. 1. A multiple sequence alignment of the protein mtHSP70 (*S.cere.* Uniprot accession P12393) from five species of fungi. The light blue line shows the MPP cleavage site located at the end of the MTS. The conserved region is colored by Jalview.

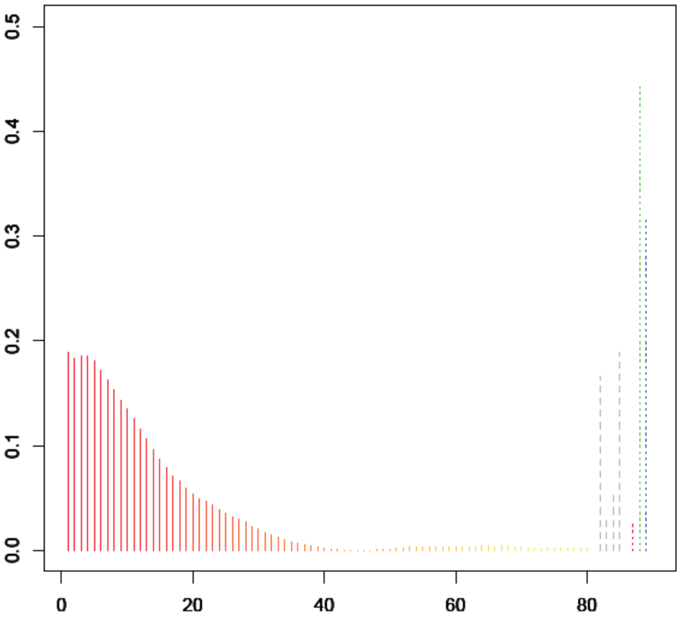

Fig. 3. Importance of each attribute as estimated by F-score is shown. At left, the LD value for each position is shown by solid and heat colored lines. Gray dash lines denote $N20$, $N40$, $N80$-99 and $NCdiff$. Colored and dotted lines denote the N-terminal physico-chemical properties #pos, #neg and Hphob, respectively.

features near the N-terminus also show F-scores significantly greater than zero.

3) Sequence divergence is not redundant to physico-chemical trends: To be promising as a feature for prediction, it is desirable that evolutionary sequence diversity not be perfectly correlated with other useful features. To investigate this we plotted LD(13), the divergence feature with the highest F-score, against the two highest scoring physico-chemical features (Figure 4). Although it is difficult to discern the exact relationship, one can see that the feature pairs do not appear highly correlated.

B. Divergence predicts presence of N-terminal signal

We tested whether sequence divergence can be used to distinguish between proteins with an N-terminal localization signal (MTS or SP) and those with none. As shown in Table II, for this binary classification task, sequence divergence *alone* allows for significantly higher prediction accuracy than randomized control experiments or the majority class fraction (66.2%).

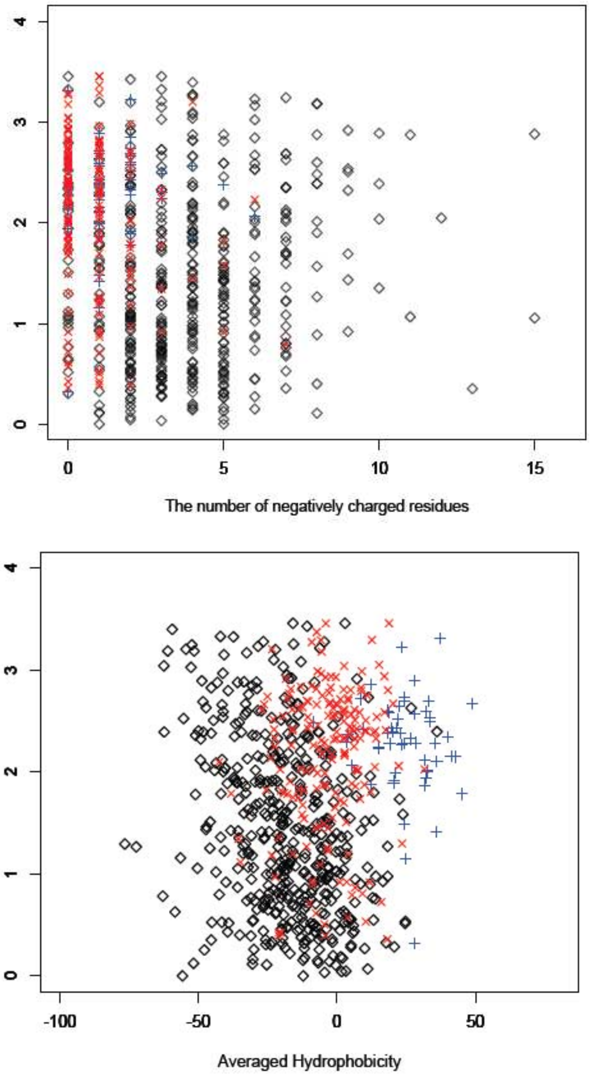

Fig. 4. The scatter plot of LD(13) on the vertical axis *vs.* #neg (top) and Hphob (bottom) on the horizontal axis is shown. MTS, SP, and N-signal-less proteins are represented by red, blue and black dots, respectively.

C. Divergence distinguishes SP vs. MTS vs. N-signal-less

Although the sequence divergence profile of SP's and MTS's appear similar when averaged over proteins containing each signal (Figure 2), we found that sequence divergence is still somewhat effective for the three-way classification of SP *vs.* MTS *vs.* N-signal-free. As shown in Table III the performance with divergence features is slightly better than

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

	mean accuracy	mean AUC	mean MCC
J48	72.49 ± 3.30	**0.68** ± 0.09	**0.40** ± 0.09
- (randomized)	65.85 ± 0.66	0.50 ± 0.01	0.00 ± 0.03
SVM	**74.64** ± 2.38	**0.68** ± 0.03	**0.40** ± 0.06
- (randomized)	66.19 ± 0.09	0.50 ± 0.00	0.00 ± 0.01

TABLE II
THREE CLASSIFICATION PERFORMANCE MEASURES ARE SHOWN FOR THE DISCRIMINATION OF N-SIGNAL CONTAINING AND N-SIGNAL-LESS PROTEINS. AUC DENOTES THE AREA UNDER THE ROC CURVES. (RANDOMIZED) INDICATES THE VALUES OBTAINED WITH THE LOCALIZATION CLASS LABELS RANDOMLY SHUFFLED 100 TIMES. FOR EACH MEASURE THE AVERAGE AND STANDARD DEVIATION IS SHOWN OVER THE 5 FOLDS OF THE CROSS-VALIDATION, OR 500 (5 × 100 TRIALS) FOLDS IN THE CASE OF THE RANDOMIZED DATA.

the majority class fraction (66.2%) and also slightly improves the performance when added to the physico-chemical features.

The ratio of examples in our dataset is 8.56:3.37:1, for N-signal-less, MTS and SP containing proteins respectively. Skewed datasets are known to complicate both learning and performance evaluation [33]. Therefore we also measured performance on a dataset with uniform class occupancy, created by randomly discarding all but 54 proteins from each class. As shown in Table IV, in this experiment the difference between the divergence feature only performance is much higher than the majority class fraction (0.33%) and the divergence features also contribute more to the performance when combined with the physico-chemical features.

VI. DISCUSSION

A. Limitations of our work

We have not attempted to create a state-of-the art predictor. This work must be considered as a proof of concept only with many limitations.

1) Measure of divergence: Many sophisticated measures have been proposed to quantify the degree of sequence conservation [26]. Here we only present results using a simple entropy based measure which ignores the phylogenetic relationship of the species involved.

2) Features used: For non-divergence features we used only three, reasonable but simple, physico-chemical based features.

3) Organism evaluated: We only evaluated our predictions on the well-studied fungi *S.cere.*. Although the mechanisms of sub-cellular localization are similar in principle in animals and plants (chloroplasts also import proteins via N-terminal signals), the details can be different [36], [37].

4) Localization signals/sites: Although many predictors discriminate between 10 or more localization sites (e.g. WoLF PSORT [38]), we focused on only two of the most common sorting signals.

5) Appropriateness of dataset: One weakness in our work, is that many of our SP proteins are not experimentally validated, but in fact partially annotated as SP proteins due to *prediction from amino acid sequence* with SignalP [22]. This unfortunate circularity (predicting predictions) is unavoidable because: 1) only a handful of SP's have been experimentally verified, and 2) the presence of SP's cannot be reliably inferred exclusively from localization site for most *S.cere.* proteins. It may be reasonable to assume that secreted proteins all have SP's, but *S.cere.* secretes very few proteins (the SWISS-PROT derived WoLF PSORT [38] dataset lists only six). Other SP containing proteins generally localize to the E.R. or Golgi body – but proteins annotated to localize to the E.R. or Golgi include non-SP containing proteins such as peripheral membrane proteins which localize to the outside of these organelles.

However, the risk of incorrect conclusion resulted from employing non-verified SP data is small. First, this problem only applies to the SP class, as recent proteomics data has provided direct measurement of many MTS's [21]. Second, given the intense study of *S.cere.* and the continued scrutiny of UniprotKB/Swiss-Prot by the research community, we find it unlikely that a large fraction of the SP proteins in our dataset are incorrectly labeled. Third, our argument is not really very circular. SignalP prediction is based on physico-chemical features but not divergence (or conservation) for prediction, and the results shown in Figure 4 suggest that the features used by SignalP probably do not correlate very closely with sequence divergence.

B. Conclusion

We find it rather remarkable that the accuracy of balanced 3-way prediction can be improved to nearly 60% just by using simply defined sequence divergence features, *while otherwise completely hiding the amino acid sequence of the protein!*

Although we readily admit the limited scope of this work, it is the first to quantitatively show that sequence divergence is a promising feature for localization prediction. We feel confident that our observation will stand the test of time, through the more exhaustive exploration that we expect to follow in the future.

We provide the first quantatitive evidence that evolutionary sequence divergence can be used to predict protein sub-cellular localization.

ACKNOWLEDGMENT

We thank Raymond Wan, Szu-Chin Fu, Junko Tsuji and Kenichiro Imai for help in preparing the proteomics based MTS data.

REFERENCES

[1] F. Eisenhaber and P. Bork, “Wanted: subcellular localization of proteins based on sequence,” *Trends Cell Biol.*, vol. 8, pp. 169–170, 1998.

[2] A. Kumar *et al.*, “Subcellular localization of the yeast proteome,” *Genes & Development*, vol. 16, pp. 707–719, 2002.

[3] W.-K. Huh *et al.*, “Global analysis of protein localization in budding yeast,” *Nature*, vol. 425, pp. 686–691, October 2003.

[4] R. Nair and B. Rost, “Sequence conserved for subcellular localization.” *Protein Sci*, vol. 11, no. 12, pp. 2836–2847, 2002.

[5] G. Schatz and B. Dobberstein, “Common principles of protein translocation across membranes.” *Science*, vol. 271, no. 5255, pp. 1519–1526, 1996.

[6] G. von Heijne, “Patterns of amino acids near signal-sequence cleavage sites,” *Eur. J. Biochem*, vol. 133, pp. 17–21, June 1983.

[7] T. Saitoh *et al.*, “Tom20 recognizes mitochondrial presequences through dynamic equilibrium among multiple bound states.” *EMBO J*, vol. 26, no. 22, pp. 4777–4787, 2007.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

	Divergence		Physico-chemical features		Combination	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.65 ± 0.01	0.34 ± 0.03	0.81 ± 0.05	0.67 ± 0.09	$\mathbf{0.82} \pm 0.04$	$\mathbf{0.68} \pm 0.08$
SP	0.50 ± 0.00	0.00 ± 0.00	0.72 ± 0.05	0.57 ± 0.07	$\mathbf{0.86} \pm 0.06$	$\mathbf{0.72} \pm 0.08$
N-signal-free	0.64 ± 0.02	0.34 ± 0.06	0.79 ± 0.05	0.63 ± 0.10	$\mathbf{0.85} \pm 0.04$	$\mathbf{0.73} \pm 0.08$
% accuracy	71.06 ± 1.57		83.11 ± 3.44		$\mathbf{86.25} \pm 3.56$	

TABLE III

THE 5-FOLD CROSS-VALIDATION PERFORMANCE OF AN SVM CLASSIFIER USING: DIVERGENCE FEATURES ONLY, PHYSICO-CHEMICAL FEATURES ONLY, AND THE TWO COMBINED; IS SHOWN FOR THREE-WAY CLASSIFICATION ON OUR ENTIRE DATASET.

	Profile		Physical features		Combination	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.65 ± 0.10	0.30 ± 0.19	$\mathbf{0.85} \pm 0.05$	$\mathbf{0.74} \pm 0.09$	0.81 ± 0.07	0.62 ± 0.12
SP	0.69 ± 0.05	0.40 ± 0.13	0.78 ± 0.09	0.59 ± 0.13	$\mathbf{0.90} \pm 0.04$	$\mathbf{0.82} \pm 0.07$
N-signal-less	0.73 ± 0.05	0.47 ± 0.11	0.77 ± 0.05	0.53 ± 0.10	$\mathbf{0.87} \pm 0.05$	$\mathbf{0.74} \pm 0.11$
% accuracy	58.56 ± 8.40		73.48 ± 4.41		$\mathbf{81.37} \pm 5.95$	

TABLE IV

THE 5-FOLD CROSS-VALIDATION PERFORMANCE OF AN SVM CLASSIFIER USING: DIVERGENCE FEATURES ONLY, PHYSICO-CHEMICAL FEATURES ONLY, AND THE TWO COMBINED; IS SHOWN FOR THREE-WAY CLASSIFICATION ON A BALANCED DATASET (54 PROTEINS IN EACH CLASS).

[8] H. Yamamoto *et al.*, "Dual role of the receptor tom20 in specificity and efficiency of protein import into mitochondria." *Proc Natl Acad Sci U S A*, vol. 108, no. 1, pp. 91–96, 2011.

[9] P. Horton, Y. Mukai, and K. Nakai, "Protein localization prediction," in *The Practical Bioinformatician*, L. Wong, Ed. 5 Toh Tuck Link, Singapore 596224: World Scientific, 2004, ch. 9, pp. 193–215.

[10] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies." *J Mol Biol*, vol. 238, no. 1, pp. 54–61, 1994.

[11] Z. Yuan, "Prediction of protein subcellular locations using markov chain models." *FEBS Lett*, vol. 451, no. 1, pp. 23–26, 1999.

[12] Q. B. Gao, Z. Z. Wang, C. Yan, and Y. H. Du, "Prediction of protein subcellular location using a combined feature of sequence." *FEBS Lett*, vol. 579, no. 16, pp. 3444–3448, 2005.

[13] N. Sakiyama, K. Runcong, R. Sawada, M. Sonoyama, and S. Mitaku, "Nuclear localization of proteins with a charge periodicity of 28 residues," *Chem-BioInformatics Journal*, vol. 7, pp. 35–48, 2007.

[14] A. Drawid and M. Gerstein, "A bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome," *JMB*, vol. 301, pp. 1059–175, 2000.

[15] M. A. Andrade, S. I. O'Donoghue, and B. Rost, "Adaptation of protein surfaces to subcellular location." *J Mol Biol*, vol. 276, no. 2, pp. 517–525, 1998.

[16] L. McCue *et al.*, "Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes." *Nucleic Acids Res*, vol. 29, no. 3, pp. 774–782, 2001.

[17] L. Martinsen, A. Johnsen, F. Venanzetti, and L. Bachmann, "Phylogenetic footprinting of non-coding RNA: hammerhead ribozyme sequences in a satellite DNA family of dolichopoda cave crickets (Orthoptera, Rhaphidophoridae)." *BMC Evol Biol*, vol. 10, p. 3, 2010.

[18] R. Nair and B. Rost, "Better prediction of sub-cellular localization by combining evolutionary and structural information." *Proteins*, vol. 53, no. 4, pp. 917–930, 2003.

[19] T. Tsukamoto *et al.*, "Characterization of the signal peptide at the amino terminus of the rat peroxisomal 3-ketoacyl-coa thiolase precursor." *J Biol Chem*, vol. 269, no. 8, pp. 6001–6010, 1994.

[20] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, and A. Bairoch, "UniprotKB/Swiss-Prot." *Methods Mol Biol*, vol. 406, pp. 89–112, 2007.

[21] F. N. Vögtle *et al.*, "Global analysis of the mitochondrial n-proteome identifies a processing peptidase critical for protein stability." *Cell*, vol. 139, no. 2, pp. 428–439, 2009.

[22] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, "Improved prediction of signal peptides: Signalp 3.0." *J Mol Biol*, vol. 340, no. 4, pp. 783–795, 2004.

[23] K. P. Byrne and K. H. Wolfe, "The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species." *Genome Res*, vol. 15, no. 10, pp. 1456–1461, 2005.

[24] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform." *Nucleic Acids Res*, vol. 30, no. 14, pp. 3059–3066, 2002.

[25] I. Mayrose, D. Graur, N. Ben-Tal, and T. Pupko, "Comparison of site-specific rate-inference methods for protein sequences: empirical bayesian methods are superior." *Mol Biol Evol*, vol. 21, no. 9, pp. 1781–1791, 2004.

[26] F. Johansson and H. Toh, "A comparative study of conservation and variation scores." *BMC Bioinformatics*, vol. 11, p. 388, 2010.

[27] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein." *J Mol Biol*, vol. 157, no. 1, pp. 105–132, 1982.

[28] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, p. 10, 2009.

[30] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag New York, Inc, 1999.

[31] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[32] Y.-W. Chen and C.-J. Lin, "Combining svms with various feature selection strategies," Available from http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf, 2005.

[33] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. on Knowl. and Data Eng.*, vol. 21, pp. 1263–1284, September 2009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1591901.1592322

[34] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview." *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[35] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, "Locating proteins in the cell using targetp, signalp and related tools." *Nat Protoc*, vol. 2, no. 4, pp. 953–971, 2007.

[36] G. Schneider, S. Sjöling, E. Wallin, P. Wrede, E. Glaser, and G. von Heijne, "Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides," *PROTEINS*, vol. 30, pp. 49–60, 1998.

[37] M. Edman, T. Jarhede, M. Sjöström, and Å. Wieslander, "Different sequence patterns in signal peptides from mycoplasmas, other gram-positive bacteria, and escherichia coli: a multivariate data analysis," *PROTEINS: Structure, Function, and Genetics*, vol. 35, pp. 195–205, 1999.

[38] P. Horton *et al.*, "WoLF PSORT: protein localization predictor." *Nucleic Acids Res*, vol. 35, no. Web Server issue, pp. W585–W587, 2007.