Prediction of Mitochondrial Targeting Signals Using Hidden Markov Models

Yukiko Fujiwara [1] **Minoru Asogawa** [1] **Kenta Nakai** [2]
fujiwara@ccm.cl.nec.co.jp asogawa@ccm.cl.nec.co.jp nakai@imcb.osaka-u.ac.jp

[1] Computational Engineering Technology Group, C&C Media Laboratories, NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216, Japan

[2] Institute of Molecular and Cellular Biology, Osaka University
1-3 Yamada-oka, Suita 565, Japan

Abstract

The mitochondrial targeting signal (MTS) is the presequence that directs nascent proteins bearing it to mitochondria. We have developed a hidden Markov model (HMM) that represents various known sequence characteristics of MTSs, such as the length variation, amino acid composition, amphiphilicity, and consensus pattern around the cleavage site. The topology and parameters of this model are automatically determined by the iterative duplication method, in which a small fully-connected HMM is gradually expanded by state splitting. The model can be used to predict the existence of MTSs for given amino acid sequences. Its prediction accuracy was estimated to be 86.9% using the cross validation test. Furthermore, a higher correlation was observed between the HMM score and the in vitro *ATPase activity of MSF, which can be regarded as an experimental measure of signal strength, for various synthetic peptides than was observed with other methods.*

1 Introduction

In this paper, we present a novel algorithm for predicting the existence of mitochondrial targeting sequences (MTSs) from amino acid sequence data. Mitochondria are essential organelles of eukaryotic cells, involved in ATP synthesis through cellular respiration (for a textbook, see [1]). To perform this function, mitochondria are equipped with a number of specific proteins, but most of them are encoded in the nuclear DNA and are transported into it after being synthesized in the cytosol. MTSs, which are usually encoded as N-terminal presequences, are used as the signal for this transport process; more specifically, they are signals to the mitochondrial matrix by default. MTSs are rich in alanines, leucines, arginines, and serines, but they are poor in acidic residues. In addition, it is widely accepted that MTSs can fold into amphiphilic structures; *i.e.,* there is usually about 3.6 residue-periodicity of hydrophobicity, which is the periodicity of the α-helix. There is also a report that another 4.8 residue-periodicity is predominant near the C-terminal-side region of MTSs [16]. Although their precise recognition processes have not been fully clarified, the MTSs are recognized by several kinds of proteins (for a recent review, see [14]). One such factor in the cytosol is mitochondrial import stimulating factor (MSF), which binds specifically to MTSs. This binding is ATP-dependent and MSF acts like a molecular chaperone. After penetrating the mitochondrial outer and inner membranes, MTSs are cleaved off by a specific enzyme, mitochondrial processing peptidase, in the inner membrane. It has been pointed out that there exist some weak consensus patterns around the cleavage site, reflecting its substrate specificity [8]. Since some proteins are further processed by mitochondrial intermediate peptidase, these consensus patterns become more obscure.

The problem of recognizing the existence of MTSs by computer is interesting from both theoretical and practical viewpoints. Nakai and Kanehisa included a subprogram for detecting MTSs in their PSORT program to predict the subcellular localization sites of proteins [13]. Since the use of knowledge of the cleavage-site consensus and the periodicity could not improve the prediction accuracy at that time, only the amino acid composition of N-terminal 20 residues was used for the variable of the discriminant analysis. Claros also developed a useful tool, MitoProt, for characterizing MTSs but it could not perform for automatic prediction [4]. Recently, Claros and Vincens examined the predictability of a number of variables to discriminate MTSs and they presented two prediction methods [5].

We report a new prediction method using hidden Markov models (HMMs). While HMMs [12] are widely used in speech recognition and have been applied to bioinformatics, such as protein modeling and multiple alignment [3, 11], their models were left-to-right ones, which are not suited for representing the periodic nature of MTSs. Therefore, we used a more general model and automatically optimized it for given training data. This optimization is achieved by "iterative duplication", which has been developed by our group [6, 7]. The obtained HMM turned out to represent the known features of MTSs well in terms of both amino acid composition and periodicity. Furthermore, the inclusion of extra three N-terminal residues of the mature portion for training was effective for including the consensus pattern around the cleavage site in the model and it did indeed raise the prediction accuracy. Finally, using experimental data, we confirmed that our prediction score is quantitatively more reasonable than those of other methods.

2 Data and Algorithm

2.1 Sequence data

Amino acid sequence data were taken from SWISS-PROT (Release 34.0 [2]). Simply, sequences annotated to have MTSs with the cleavage site information were used as positive data (547 sequences), while *Saccharomyces cerevisiae* sequences annotated to be localized at any positions other than mitochondria were used as negative data (1,273 sequences). The lengths of MTSs ranged from 8 to 140 and the average was 35.0. The amino acid composition of MTSs was similar to what has been previously reported: rich amino acids were 13.4% for alanine (A), 12.1% for leucine (L), 12.3% for arginine (R), and 11.6% for serine (S), while poor ones were 0.8% for aspartic acid (D) and 1.1% for glutamic acid (E). To train the model, the sequences of MTSs plus the three residues in the following mature portion were extracted from the positive data.

2.2 Construction of the Model

An HMM is characterized by the number of states, number of output symbols per state, initial state distribution, state transition probability, and symbol observation probability in each state [12]. In our model of MTSs, each state corresponds to a certain residue position of MTSs, and each state can output 20 kinds of symbols corresponding to 20 amino acids. The initial state is to be at the first position of the MTS, which is the first residue of the sequence. The state transition probability represents the probability of moving to the next position of MTSs. To allow for a variable length of MTSs and to represent their amphiphilic nature, the network topology used is not the usual left-to-right type but a more general form. The final state corresponds to the position where the signal information ends (*i.e.,* near the cleavage site). The network topology and parameters were determined by the iterative duplication method developed by our group (see below).

The algorithm of the iterative duplication method which optimizes the network topology and parameters is shown in Figure 1. The original algorithm [6, 7] was modified to suit this problem.

First, the training data are evenly divided into estimating data and determining data. The former are used in the parameter estimation phase, in which a fully-connected HMM of three states is iteratively expanded by state splitting. To avoid local maxima, several initial HMMs with randomly-chosen initial parameters are used. For each model, parameters are estimated with the Baum-Welch algorithm and the likelihood for the estimating data is calculated by multiplying the probability of

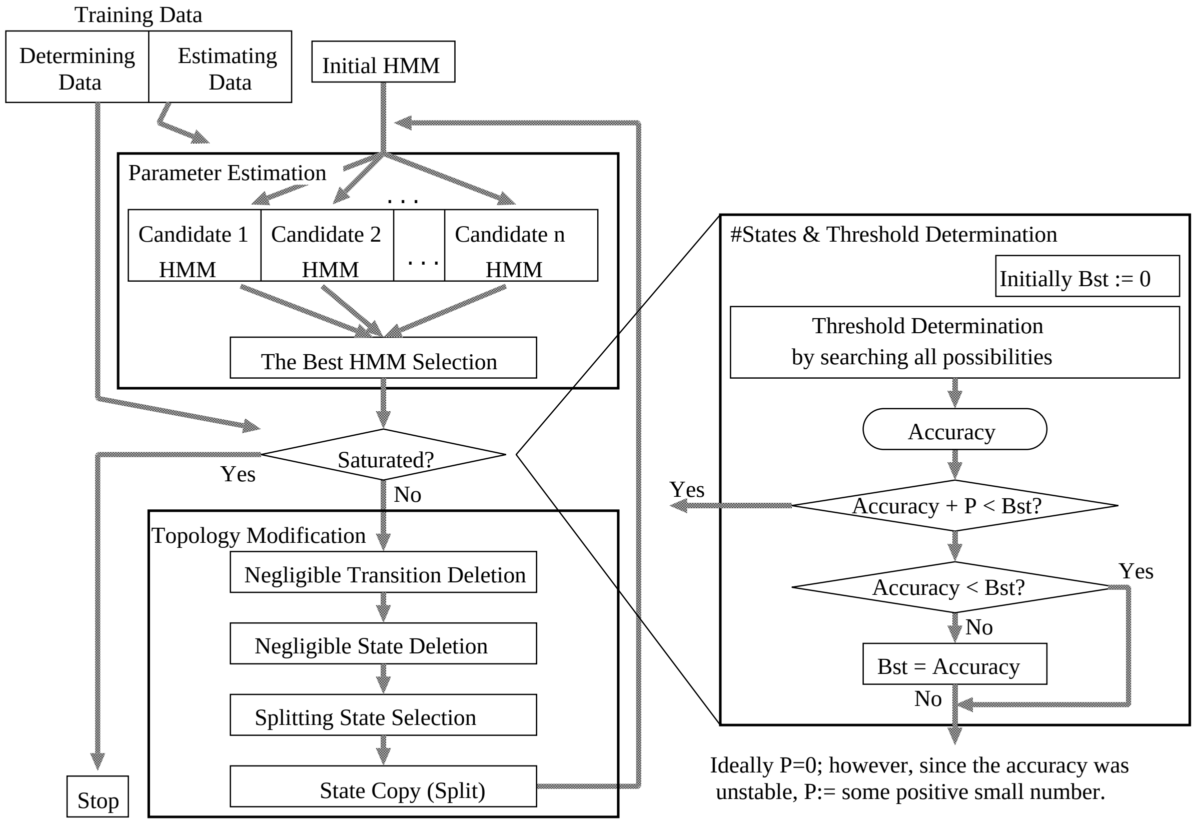


Figure 1: Improved iterative duplication method

observing each MTS in the data set. Then, the one giving the best likelihood is chosen.

In the next step, the topology modification phase, the transitions with negligible transition probabilities are deleted first. Negligible probabilities means that they are less than $\epsilon = \max(\epsilon_1, r)$, where ϵ_1 is a smoothing value and r is a convergence radius [12]. Next, negligible states, that is, ones with negligible initial probability and incoming transition probabilities, are deleted. Next, one of the states is selected and duplicated so that the new state has the same transition probabilities as the original state. If a state having a self-loop is chosen, the generated state is designed to have a self-loop, and the transition from the new state to the original state is considered as well vice versa. The way of selecting a state to be split was described in [7]. Then, the new topology is iteratively used as an initial HMM for the parameter estimation phase.

Once the best HMM has been selected in the parameter estimation phase, the threshold value is optimized by examining all values so that the accuracy for the determining data becomes maximal. The algorithm terminates when this accuracy saturates. In the current implementation, the number of states giving the best accuracy for the determining data is selected in a series of state numbers whose limit is set to an appropriately large number.

2.3 Validation of the model

The derived model was used to predict the existence of MTS for arbitrary given sequences. In this prediction phase, the highest score was selected from the scores for the N-terminal subsequences of lengths from 11 to 143 for both positive and negative data.

We used the three-fold cross validation method to estimate the prediction accuracy of our method:

the data were divided into three, and the two-thirds were used for training while one-third was used for testing. The division was done so that there were no pairs with more than 50% identical residues between the training and testing data.

The ATPase activities of MSF for various synthetic peptides [10] were used as a semi-quantitative measure of MTS strengths. For reference, two other scores were calculated for each sequence of the synthetic peptides. One was the highest alignment score (percentage of identical residues) with our positive data. This was calculated by clustalV [9]. The other was the discriminant score of a subprogram used in PSORT [13]. Since the program requires N-terminal 20 residues, poly-alanines were added for shorter peptide sequences.

3 Results and Discussion

3.1 Constructed models represent general features of MTSs

In the three trials of cross validation, the numbers of states constructed were 26, 24, and 23 for the training data of MTSs with the extra three residues, while they were 29, 25, and 33 for the data of exact MTSs. One of obtained models is shown in Figure 2. The topology of this model is rather complicated, but one can see that various known features of MTSs are coded in it. First, the model represents the compositional tendency of MTSs: the output symbols are rich in alanines (A), serines (S), leucines (L), and arginines (R), but there are no acidic residues (D/E). Second, the model contains several loops. Although their interpretation is difficult, they seem to be amphiphilic loops, which have been pointed out to exist in MTSs. Several pseudo-helical-wheel diagrams are also shown in the figure. In these diagrams, basic residues (R/K) tend to be located at one side and hydrophobic residues (L/V/F/I) tend to be at the other side. Third, the model appears to encode the substrate specificity of mitochondrial processing peptidase. Both sequence analyses [8] and *in-vitro* experiments [15] suggested the importance of an arginine residue frequently observed at the -2 or -3 position from the cleavage site. Since three residues from the mature portion were included as the training sequences, it is reasonable that state S0 corresponds to position +3 and state S10 corresponds to position -3 in Figure 2. Consistent with the above knowledge, states S10 and S11 output symbol R with high probabilities. Moreover, state S12, which corresponds to position -1, outputs Y most frequently, while state S13 outputs A and S frequently. This observation is highly consistent with the R-3 motif, $RXY|(S/A)$ where '|' denotes the cleavage site, proposed by Gavel and von Heijne [8]. There is a path that bypasses the S13 state. If this path is selected, the S10, S11, and S12 states correspond to positions -2, -1, and +1, respectively. In this case, the pattern also seems to be similar to the positions around the intermediate cleavage site in the R-10 motif, $RX|(F/I/L)SX_6|X$. Thus, our model can be interpreted as a degenerated representation of two cleavage-site motifs.

3.2 Prediction accuracy is at a practical level

The results of the cross validation test are summarized in Figure 3. It can be seen that the variances of the accuracy between the three trials became smaller in the data with the cleavage-site information ('ext.MTS') for all cases. Furthermore, the value of the accuracy for detecting positive data significantly improved (from 85.0% to 89.2%) while the total average slightly improved (from 85.5% to 86.9%). Although our model is not designed to predict the exact cleavage site, it can predict that it will probably be at the position between the states S12 and S13 (in some cases between S11 and S12). In fact, in most cases of the prediction phase, subsequences which show the highest score nearly corresponded to the range of presequences (data not shown). We conclude that our model has some practical value for predicting the presence of MTSs from only amino acid sequence data.

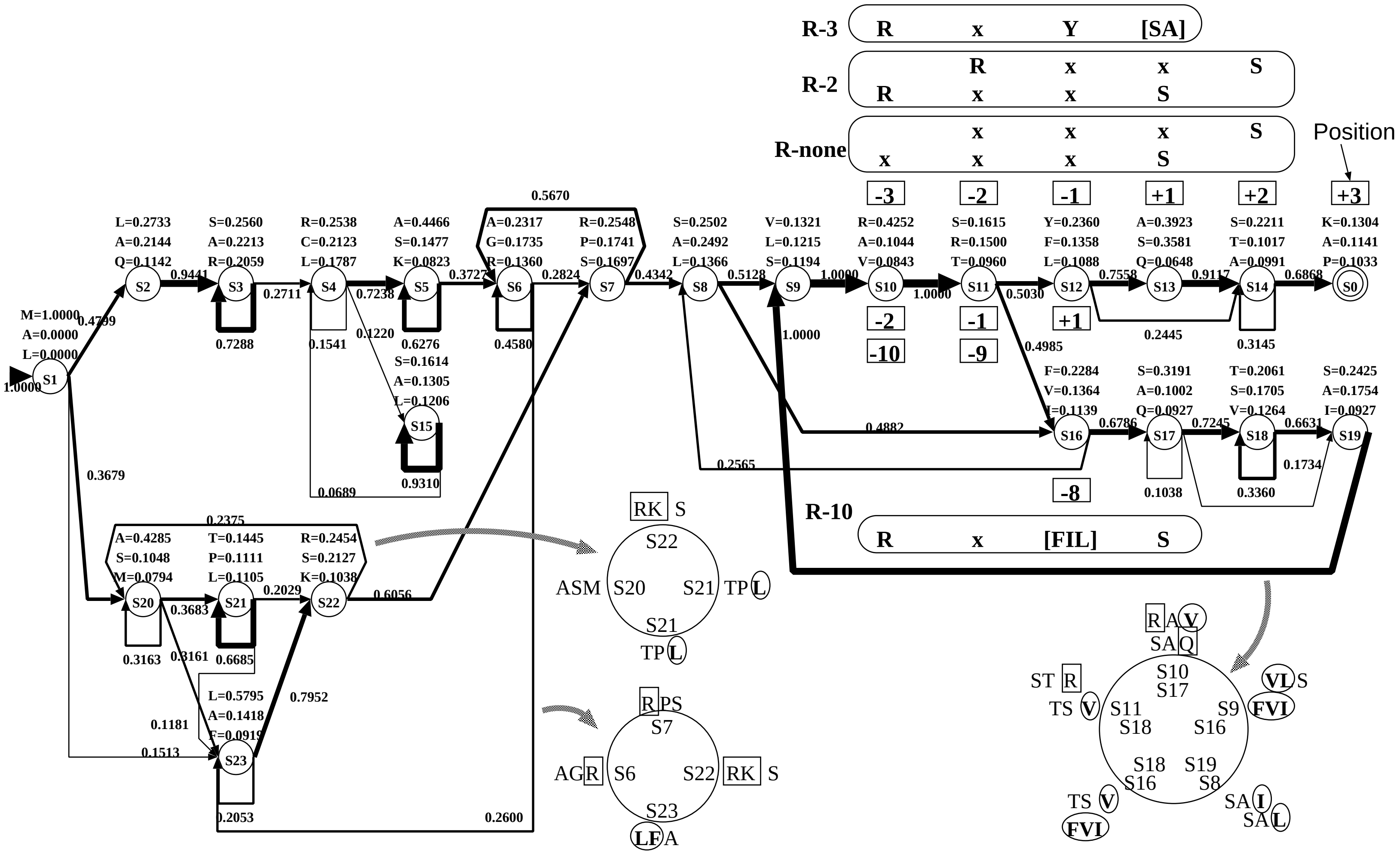
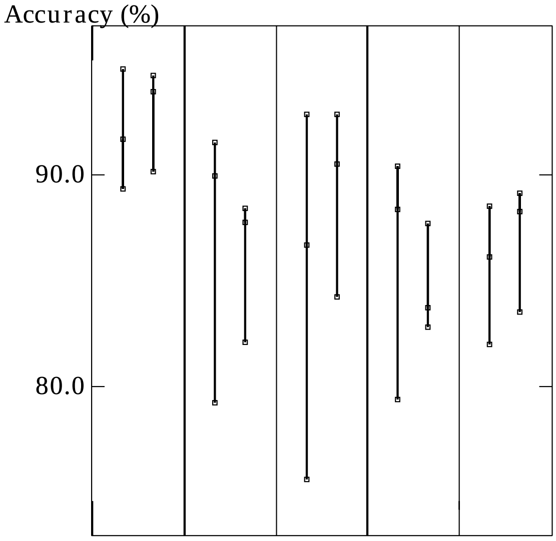


Figure 2: An HMM for ext.MTSs. The initial state is S1 (probability 1.0 arrow) and the final state is S0 (double circle). The three highest observation probabilities at each state are noted and the transitions with negligible probabilities are omitted. Some examples of the observed periodicity are shown by the circles viewed from above, where hydrophobic and hydrophilic amino acids are surrounded by circles and squares.



	Training		Testing		
	Positive	Negative	Positive	Negative	Average
(MTS)	92.0%	87.0%	85.0%	85.9%	85.5%
(Ext.MTS)	93.0%	86.1%	89.2%	84.7%	86.9%

Figure 3: Prediction accuracy of our model. To show the effect of including the cleavage-site motif, both the results for the data of MTSs only (denoted by MTS) and the data of MTSs plus the following three residues (denoted by ext.MTS for 'extended MTSs') are shown. In each block of the above graphs, the left bar shows the range of accuracy in three trials for MTSs and the right bar shows the range of accuracy for the extended MTSs. In the lower table, the averaged accuracy of three trials is indicated.

3.3 HMM scores correlate with the signal strength

To further examine the ability of our method to assess the intracellular signal strength, we plotted the correlation between the experimentally-measured values and the calculated scores in Figure 4. As an experimental measure, we used the ATPase activity of MSF induced by various synthetic peptides (The data were taken from [10]). The data for which ATPase activity was not detectable were treated as having no activity. As theoretical scores, we used the z-score of our HMM, the best homology score (percentage match of identical residues) with our positive data, and the discriminant score involved in PSORT. In the plot of PSORT, open rectangles were used to show sequences shorter than 20, where poly-alanines were added to make the length become 21. The correlation coefficients were 0.66 for our model, 0.63 for the homology (0.61 for its logarithm), and 0.35 for the PSORT certainty. In principle, the alignment scores of the synthetic peptides that are subsequences of known MTSs will be 100%. For example, in the plot (Alignment), pAd-(17-32) and pAd-(43-58) denote residues 17-32 and 43-58 of the pre-adrenodoxin, both of which scored 100%. However, since both of these peptides induce little ATPase activity, this scoring method fails to predict the signal strength quantitatively. In addition, it is quite reasonable that the PSORT cannot show significant correlation with the signal strength because it only considers the amino acid composition of the N-terminal 20 residues, but the synthetic peptides contain some examples that are similar in composition but do not show amphiphilicity. As a result, our current model seems to represent the MTS activity well. Further comparison with the method of Claros and Vincens should also be done.

Acknowledgments

Kenta Nakai is partly supported by a Grant-in-Aid (*Genome Science*) for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports, and Culture of Japan.

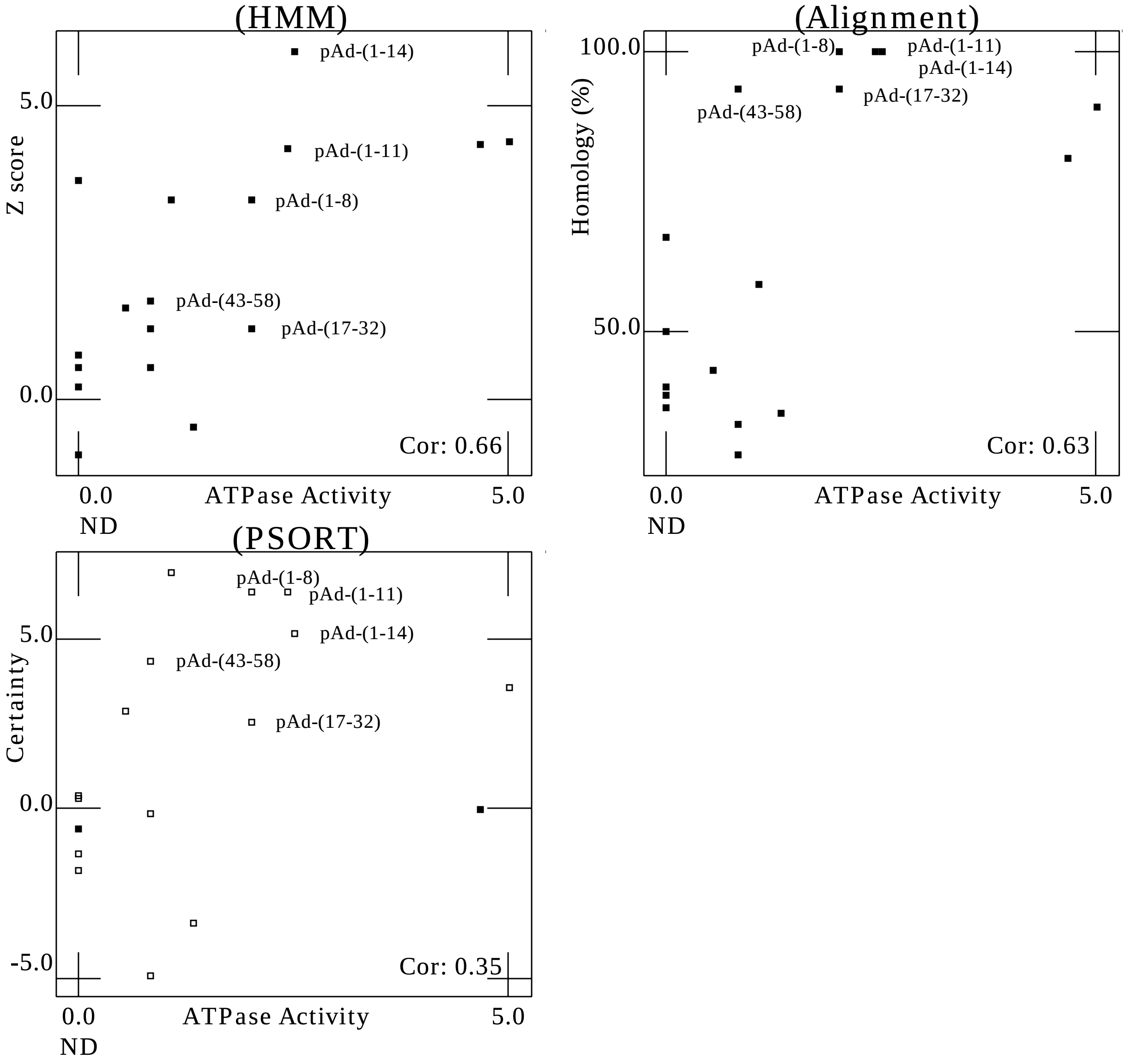


Figure 4: Correlation between the biological signal strength and various scores. In each plot, the longitudinal axis indicates the ATPase activity of MSF (nanomoles per 30 minutes) induced by various peptides, which were measured by Komiya *et al.* The latitudinal axes are: the z score calculated by our model, the homology score (in %) with the positive data, and the discriminant score of the subprogram included in PSORT for the plots, (HMM), (Alignment), and (PSORT), respectively.

References

[1] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D., "Molecular Biology of the Cell (3rd ed.)," Garland, New York, 1994.

[2] Bairoch, A., "The SWISS-PROT protein sequence data bank and its new supplement TREMBL," *Nucleic Acids Res.*, 24, 21–25, 1996.

[3] Baldi, P., Chauvin, Y., Hunkapiller T. and McClure, M.A., "Hidden Markov Models of Biological Primary Sequence Information," *Proc. Natl. Acad. Sci. U.S.A.*, 91 (3), 1059–1063, 1994.

[4] Claros, M.G., "MitoProt, a Macintosh application for studying mitochondrial proteins," *CABIOS*, 11, 4, 441–447, 1995.

[5] Claros, M.G. and Vincens, P., "Computational method to predict mitochondrially imported proteins and their targeting sequences," *Eur. J. Biochem.*, 241, 779–786.

[6] Fujiwara, Y., Asogawa, M. and Konagaya, A., "Stochastic Motif Extraction using Hidden Markov Model," *Proceedings of the Second Intelligent Systems for Molecular Biology*, 121–129, 1994.

[7] Fujiwara, Y., Asogawa, M. and Konagaya, A., "Hidden Markov Model to Extract Leucine Zipper Motif," *NEC Research & Development*, 38(3), 1997.

[8] Gavel, Y. and von Heijne, G., "Cleavage-site motifs in mitochondrial targeting peptides," *Protein Eng.*, 4, 33–37, 1990.

[9] Higgins, D.G. and Sharp, P. M., "Fast and Sensitive Multiple Sequence alignments on a microcomputer," *CABIOS*, 5, 151–153, 1989.

[10] Komiya, T., Hachiya, N., Sakaguchi, M., Omura, T. and Mihara, K., "Recognition of Mitochondria-targeting Signals by a Cytosolic Import Stimulation Factor, MSF," *J. Biol. Chem.*, 269, 49, 30893–30897, 1994.

[11] Krogh, A., Brown, M., Mian, I.S., Sjölander, K. and Haussler, D., "Hidden Markov models in computational biology: Applications method to protein modeling," *J. Mol. Biol.*, 235, 1501–1531, 1994.

[12] Levinson, S.E., Rabiner, L.R. and Sondhi, M.M., "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *Bell Syst. Tech. Journal*, 62 (4), 1035–1074, 1983.

[13] Nakai, K. and Kanehisa, M., "A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells," *Genomics*, 14, 897–991, 1992.

[14] Schatz, G., "The protein import system of mitochondria," *J. Biol. Chem.*, 271, 31763–31766, 1996.

[15] Song, M.-C., Shimokata, K., Kitada, S., Ogishima, T. and Ito, A., "Role of basic amino acids in the cleavage of synthetic peptide substrates by mitochondrial processing peptidase," *J. Biol. Chem.*, 120, 1163–1166, 1996.

[16] Von Heijne, G., Steppuhn, J. and Herrmann, R.G., "Domain structure of mitochondria and chloroplast targeting peptides," *Eur. J. Biochem.*, 180, 535–545, 1989.