PROTEIN SUBCELLULAR LOCALIZATION PREDICTION WITH WOLF PSORT

PAUL HORTON [A], KEUN-JOON PARK [A,C], TAKESHI OBAYASHI [B,D,E], AND KENTA NAKAI [C]

[a] *Computational Biology Research Center, AIST, Tokyo, Japan*
E-mail: horton-p@aist.go.jp

[c] *Human Genome Center, Institute of Medical Science, University of Tokyo*
E-mail: park-kj@hgc.jp, knakai@ims.u-tokyo.ac.jp

[b] *Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology*
E-mail: toobayas@bio.titech.ac.jp

[d] *Graduate School of Pharmaceutical Sciences, Chiba University*

[e] *Core Research for Evolutional Science and Technology*

We present a new program for predicting protein subcellular localization from amino acid sequence. WoLF PSORT is a major update to the PSORTII program, based on new sequence data and incorporating new features with a feature selection procedure. Following SWISS-PROT, we divided eukaryotes into three groups: fungi, plant, and animal. For the 2113 fungi proteins divided into 14 categories; we found that, combined with BLAST, WoLF PSORT yields a cross-validated accuracy of 83%, eliminating about 1/3 of the errors made when using BLAST alone. For 12771 animal proteins a combined accuracy of 95.6% is obtained, eliminating 1/4 of BLAST alone errors, and for 2333 plant proteins the accuracy can be improved to 86% from 84%.

1. Introduction

Protein localization is a central issue in understanding cells. More than 20 papers[6] have been published in major international journals describing programs for predicting localization from amino acid sequence in eukaryotic cells. Some of these works such as PSORT,[9] PSORTII,[7] the SignalP[11,10] family of programs and others[4] use sorting signal information for prediction. However the majority of prediction programs developed use amino acid content in some form. These methods exploit the longstanding observation[12] that amino acid content correlates strongly with localization site.

In this paper we present WoLF PSORT, a major extension to PSORTII which combines (mostly signal based) features from PSORT and iPSORT[2] with amino acid content. Prediction accuracy is increased by using feature selection while retaining

the simple k nearest neighbor classifier used by PSORTII. The dataset constructed and web service are available at `wolfpsort.org`

2. Methods

2.1. *Dataset*

We prepared the dataset primarily from Swiss-Prot[3] Release 45.0 annotation, ignoring entries with weakening qualifiers such as "by similarity". In addition several hundred *Arabidopsis* entries were added from the Gene Ontology[1] web site (up to the 2004/12/4 release). Entries with evidence codes {TAS, IDA, IMP} were included, with revisions by hand in a few cases.

2.1.1. *Site Definition*

The distribution of localization sites in our database is shown in Table 1. The sites reflect common usage found in localization labeled Swiss-Prot entries. Table 2 shows the corresponding Gene Ontology numbers for the Gene Ontology derived sequences.

Table 1. The distribution of localization sites for each category of organisms.

localization	animal	plant	fungi	localization	animal	plant	fungi
*nucl*ear	2682	433	667	cyto_nucl	245	9	91
*plas*ma membrane	3195	160	220	cyto_mito	18	3	8
*extra*cellular	3130	113	140	cyto_pero	10	0	2
*cyto*sol	1555	452	383	cyts_plas	5	0	0
*mito*chondria	938	200	389	cyto_plas	4	0	0
*chlo*roplast	N/A	744	N/A	cyto_golg	4	0	0
E.R.	425	65	66	E.R._mito	18	0	4
*pero*xisome	217	47	77	E.R._golg	9	0	0
*lyso*some	132	N/A	N/A	extr_plas	19	0	0
*golg*i body	100	25	38	mito_pero	15	0	0
*vacu*ole	16	72	23	mito_nucl	2	0	0
*cyto*skeletal	32	10	5	**sum**	**12771**	**2333**	**2113**

Italics indicate the abbreviated name. Localization names joining two abbreviated names with "_", such as "cyto_nucl", indicate dual localization.

2.2. *WoLF PSORT system*

WoLF is a feature selection program, (the name "WoLF" is loosely inspired by the words "Learning", and "Weighted Features"). WoLF PSORT is the combination of WoLF with a version of PSORTII slightly extended for this purpose. The extended version outputs amino acid features and some iPSORT[2] features as well as the PSORT features. An overview of the system is shown in Figure 1.

Table 2. The correspondence between the localization sites used in our study and the GO numbers for the entries derived from GO annotation.

description	GO numbers	depth	WoLF PSORT site
cytoskeleton	GO:0005856	2	cyts
cytosol	GO:0005829	0	cyto
endoplasmic reticulum	GO:0005783	0	E.R.
extracellular	GO:0005576	0	extr
cell wall	GO:0005618	0	extr
Golgi apparatus	GO:0005794	1	golg
mitochondrion	GO:0005739	0	mito
nucleus	GO:0005634	0	nucl
plasma membrane	GO:0005886	0	plas
peroxisome	GO:0005777	2	pero
vacuolar membrane	GO:0005774	2	vacu
chloroplast	GO:0009507	0	chlo
thylakoid lumen	GO:0009543	0	chlo

Depth indicates the number of levels of "part_of" descendents included with the GO number. For example GO:0005856 and all of its "part_of" children and grandchildren were included in cyts. Note that "extr" and "chlo" are the union of two lines from this table.

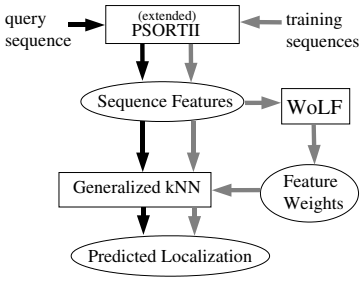

Figure 1. A schematic of the WoLF PSORT system is shown. Rectangles represent programs or procedures and ovals represent computed quantities. The black arrows denote information derived from the query sequence while gray arrows denote information from the training sequences.

2.3. *Classification*

2.3.1. *Candidate Features*

We used fours kinds of features: PSORT[9] features, iPSORT[2] features, amino acid content, and sequence length. Each feature is a deterministic mapping from amino acid sequence to the reals. Since the numerical range of the raw features are not homogeneous, we normalized each feature to its percentile value in the training data. Values observed in test data may not appear in the training data – in which case our programs use linear interpolation (extrapolation) to obtain a pseudo-percentile value.

2.3.2. *Classification Algorithm*

We adopted a weighted version of the *k*NN (*k*-Nearest Neighbors)[5] algorithm for classification. As in standard *k*NN, our method classifies based on the k nearest instances in the dataset. However we slightly extended the distance calculation. In our variation two weights, w_{1i} and w_{2i}, are associated with each feature i. More formally, let F_{ji} and F_{ki} denote the values of feature i in protein instance j and k respectively. The distance $d_W(j,k)$ between j and k is defined as:

$$d_W(j,k) = \sum_i w_{1i}|F_{ji} - F_{ki}| + \sqrt{\sum_i w_{2i}(F_{ji} - F_{ki})^2} \quad (1)$$

combining elements of Euclidean and Manhattan (city block) distances.

2.3.3. *Extensions for Dual Localization Prediction*

The dataset contains some dually localized proteins. We gave partial credit for partially correct predictions as shown Table 3.

Table 3. Example Utility Values.

label	prediction	utility	label	prediction	utility
nucl	nucl	1	nucl_cyto	nucl_mito	0.333
nucl_cyto	nucl_cyto	1	nucl	cyto	0
nucl_cyto	nucl	0.5	mito	nucl_cyto	0
nucl	nucl_cyto	0.5	mito	nucl	0

Examples of the utilities used in this study are shown. "label" denotes the localization site according to the dataset annotation. Note that we chose to use utilities that are symmetric with regard to the label and prediction.

2.3.4. *Feature Selection and Weighting*

We developed a C++ program, WoLF, for selecting the weights in Equation 1. Given a set of candidate features, WoLF selects non-negative integer weights for each feature (a weight of zero is equivalent to exclusion of the feature). WoLF uses a greedy, neighborhood search algorithm to find a locally optimal set of weights. The program uses the jackknife (leave one out cross-validation) utility on the training data to evaluate weight vectors. In the case of ties the simpler weight vector is chosen.

2.3.5. *Reducing Over-reliance on Sequence Similarity*

When training WoLF PSORT we employed a *taboo list* which disallows the use of highly similar sequences (with identical localization sites) as a neighbor when classifying a given instance. We determined the threshold by inspection of the

correlation between best hit eValue and co-localization. The thresholds used were -33, -63, -33.4 $\log_2$ eValue for fungi, plant, and animal respectively.

2.3.6. *Evaluation of WoLF PSORT Accuracy*

5-fold cross-validation was used to estimate the accuracy. For each partition, feature and k value selection were performed using a jackknife test on the training partition.

3. Results

3.1. *Effect of Feature Weighting*

The results of various feature weight vectors, including those selected by WoLF PSORT are shown in Table 4. For those tables the value of k was optimized separately for the taboo and no taboo list cases. The WoLF PSORT weight vectors (one per partition) however were only trained using the taboo list. The confusion matrices for the WoLF PSORT cross-validation for yeast (which had the highest fraction of dual localization annotations) is shown in Table 6.

In addition to cross-validation studies, we also ran the WoLF feature weighting procedure on the complete datasets. The selected features (trained with the taboo list) are shown in table 5.

3.2. *WoLF PSORT Combined with BLAST*

We calculated the utility of combining WoLF PSORT with BLAST in a trivial way; namely using the WoLF PSORT prediction for queries whose best BLAST hit eValue exceeds a given threshold, while predicting the localization of the best BLAST hit otherwise. Ties for the best BLAST hit (especially with eValue=0) were fairly common, in which case we voted amongst the best hits (breaking ties using the overall proportion of each localization in the given dataset). In the rare cases in which no BLAST hit was obtained, the majority classifier was used in lieu of BLAST. The results of this hybrid predictor for the three datasets are shown in Figure 2.

3.3. *WoLF PSORT Server*

The WoLF PSORT server is freely available at `wolfpsort.org`. Detailed information about the features of the query sequence and its k nearest (by Equation 1) neighbors are given. These tables give the user a chance to examine the *evidence* behind the prediction. For example Figure 3 shows a partial screen shot of the detailed page (one click away from the summary page) when the protein AEP_YARLI is used as a query. From the first row in the displayed table one can see that the variable *gvh*, the signal peptide detecting weight matrix score of Gunnar von Heijne,[13] has a very high value (93 percentile), consistent with the prediction of

Table 4. Cross-validated utility with various feature weight vectors.

Fungi Dataset				
weight vector type	#weights	taboo	k	% utility
psortEuclid	31	no	9.2(2.2)	64.7(2.8)
psortEuclid		yes	15.6(6.0)	**61.3(3.1)**
allEuclid	56	no	4.6(3.5)	73.9(2.2)
allEuclid		yes	26.8(11.3)	**69.3(2.9)**
allWeights	112	no	3.2(1.8)	74.3(1.7)
allWeights		yes	31.2(18.0)	**69.4(2.7)**
WoLF PSORT	22.2(4.3)	no	18.2(4.3)	72.6(0.7)
WoLF PSORT		yes	18.2(4.3)	**70.7(1.1)**
Plant Dataset				
weight type	#weights	taboo	k	% utility
psortEuclid	31	no	4.8(1.9)	66.5(2.0)
psortEuclid		yes	41.2(24.3)	**53.6(2.0)**
allEuclid	56	no	1(0)	85.3(1.2)
allEuclid		yes	19(3.8)	**60.0(3.8)**
allWeights	112	no	1(0)	85.2(1.2)
allWeights		yes	20.6(3.2)	**60.0(3.7)**
WoLF PSORT	21.2(1.3)	no	3(2.1)	76.7(2.4)
WoLF PSORT		yes	13.8(7.0)	**65.1(2.6)**
Animal Dataset				
weight type	#weights	taboo	k	% utility
psortEuclid	31	no	1(0)	79.7(0.5)
psortEuclid		yes	36.0(9.3)	**72.2(0.8)**
allEuclid	56	no	1(0)	92.3(0.5)
allEuclid		yes	34.6(6.5)	**77.8(0.6)**
allWeights	112	no	1(0)	93.1(0.6)
allWeights		yes	30.2(7.2)	**79.0(0.7)**
WoLF PSORT	25.8(4.2)	no	39.4(7.2)	83.2(1.2)
WoLF PSORT		yes	39.4(7.2)	**79.7(1.0)**

Utility is given as percent of the maximum possible. The number of (non-zero) weights is omitted when it is the same as the row above. Numerical entries represent averages over 5-fold cross-validation with standard deviations given in parenthesis. The "psortEuclid" weight vector has weight 1 for the quadratic term of each PSORT feature, "allEuclid" has weight 1 for the quadratic term of all features, "allWeights" has weight 1 for all possible terms, and WoLF PSORT is the weight vector selected by WoLF PSORT.

extracellular. One can also see that the value of the *mit* feature is very different between the query and two of its neighbors (PEPF_ASPFU and PEPA_ASPOR). *mit* is a variable designed to discriminate between mitochondrial and non-mitochondrial proteins[9] so this does not seem to weaken the evidence in this case.

4. Discussion

4.1. *Interpretable Results*

WoLF PSORT alone achieves accuracy estimates (with sequence similarity reduction used for training but not evaluation) of 73%, 77%, 83%, on the yeast, plant

Table 5. Selected Features.

feature type	feature name	fungi	plant	animal
	net charge(1, 25)	1	0	1^2
iPSORT	max negative charge(1,20)	0	1	0
	max hydropathy(1,30)	2	1	1,1^2
PSORT	gvh	1	0	1^2
signal peptide	psg	0	2	0
PSORT	mip	0	2	1^2
targeting signal	mit	1	2	1
	alm	1	0	3^2
	m1a	1^2	1	0
PSORT	m1b	1	3	0
membrane	m3a	0	1	0
protein	m3b	1	1	0
related	mNt	1	0	0
	tms	2	1^2	2
	erl	1^2	1^2	0
	leu	1	0	1
PSORT	nuc	1	1	1^2
sorting	pox	1^2	1^2	0
motif	tyr	2	1	1^2
	yqr	1	0	0
	vac	0	0	1
	dna	1^2	1^2	1,4^2
PSORT	rib	2^2	1^2	1^2
non-sorting	myr	1^2	0	1,1^2
motif	rnp	1	1	1^2
	act	0	0	3^2
miscellaneous	length	0	1	1^2

The selected features are shown with their weights. Features with 0 weight for all datasets were omitted. "1^2" and "2^2" indicate the quadratic term rather was selected with a weight of 1 or 2 respectively. The amino acid content features (by one letter code) selected for the three datasets were "ARNDQEGIKMFSWV", "ACQHILSV", and "CIKS" respectively. In each of those cases the weight was 1, the weight type was always linear for the fungi and plant dataset and always quadratic for the animal dataset. "sorting" motifs refer to motifs such as the E.R. retention signal "erl" with a direct causal relationship to localization. Descriptions of the variables can be found on the WoLF PSORT server. For the PSORT variables useful documentation can also be found on the PSORT help page `www.psort.nibb.ac.jp`

and animal datasets, with a small number of features and the trivially simple k nearest neighbors classifier. We do not claim that this will meet the accuracy of sophisticated classifiers such as the popular support vector machine. However the template based nature of the kNN classifier makes individual classifications easier to interpret. The web server provides a tabular display to facilitate this process.

4.2. *Evaluation in the presence of similar sequences*

In this study we included many similar sequences. This makes achieving a high accuracy easy. However by comparing with BLAST we were able to show that our method can be effective even when sequence similarity is low. For example the

Table 6. Confusion Matrix for the Fungi Dataset.

site	nucl	cyto nucl	cyto	cyto	mito	mito nucl	plas	extr	cyto pero	pero	E.R.	golg	vacu	cyts	sum
nucl	**556**	33	37	0	27	2	6	0	0	6	0	0	0	0	667
cyto_nucl	49	**10**	24	0	6	0	0	0	0	2	0	0	0	0	91
cyto	88	16	**232**	3	30	0	1	5	0	7	1	0	0	0	383
cyto_mito	1	0	0	**0**	7	0	0	0	0	0	0	0	0	0	8
mito	38	3	22	0	**299**	3	9	6	0	8	1	0	0	0	389
mito_nucl	1	0	1	0	2	**0**	0	0	0	0	0	0	0	0	4
plas	6	0	3	0	2	0	**205**	2	0	1	1	0	0	0	220
extr	2	2	4	0	2	0	0	**129**	0	0	1	0	0	0	140
cyto_pero	1	0	1	0	0	0	0	0	**0**	0	0	0	0	0	2
pero	16	3	23	0	5	0	6	2	1	**20**	1	0	0	0	77
E.R.	4	2	5	0	2	0	30	11	0	4	**7**	1	0	0	66
golg	10	1	1	0	1	0	10	8	0	0	2	**5**	0	0	38
vacu	2	0	2	0	2	0	10	6	0	0	1	0	**0**	0	23
cyts	4	0	1	0	0	0	0	0	0	0	0	0	0	**0**	5
sum	778	70	356	3	385	5	277	169	1	48	15	6	0	0	2113

The rows represent the dataset labels. The columns represent predictions made by WoLF PSORT

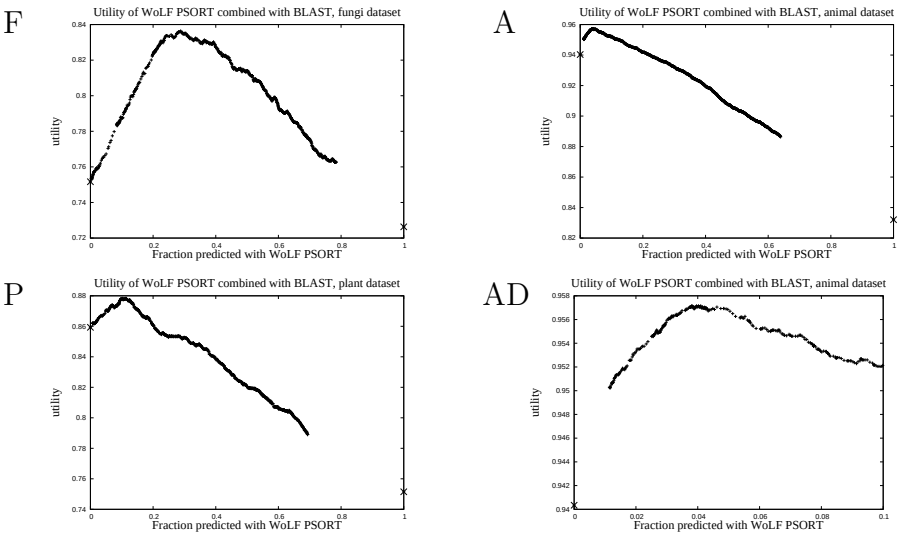

Figure 2. The utility obtained by combining WoLF PSORT and BLAST with a simple eValue threshold is shown for the fungi (F), plant (P), and animal (A) datasets. A detailed view of the animal data (AD) is shown as well. The point at $x = 1$ represents using WoLF PSORT alone. The lack of points between 0.8 and 1 for the fungi data is due to the fact that approximately 20% (30% for plant, 36% for animal) of the proteins have hits with an eValue of "0", a similar (but smaller) gap occurs at the other end due to proteins with no blast hits.

number of errors produced by using BLAST alone on the animal dataset can be reduced by about a quarter (accuracy of $\approx 95.6\%$ *vs* $\approx 94.0\%$, yielding 179 errors) by using WoLF PSORT for proteins without good BLAST hits.

Normalized F

id	site	iPSORT		PSORT Featu									
		-1 25	MxHy1 30	alm	dna	erl	gvh	leu	mla	mlb	m3b	mNt	mit
AEP_YARLI	extr?	55	92	21	46	50	93	47	50	49	48	49	58
AEP_YARLI	extr	55	92	21	46	50	93	47	50	49	48	49	58
PEPF_ASPFU	extr	55	96	17	46	50	96	47	50	49	48	49	83
CAR1_CANPA	extr	55	93	20	46	50	96	47	50	49	48	49	55
PEPA_ASPOR	extr	55	93	21	46	50	93	47	50	49	48	49	94

Figure 3. A screen shot of the server showing the feature table displayed for query and nearest neighbors is shown. The query id, predicted localization, and features are showed aligned with the id, localization, and features of its nearest neighbors. Color is used to attract attention to large differences between the query and nearest neighbors. In the example shown the query is "AEP_YARLI", which also appears on the next line as the nearest neighbor in the dataset.

4.3. *Predicting Dual Localization*

WoLF PSORT was designed with dual localization in mind. The only dual localization category for which SWISS-PROT currently contains a significant number of entries is dual localization to the nucleus and cytosol. This is an important category of proteins, for example some transcription factors are regulated by conditional localization to the nucleus[8].

The prediction results seen in the confusion matrix (Table 6) are mixed. Unfortunately most of the 91 proteins labeled "cyto_nucl" are misclassified as either nuclear or cytosol. On the other hand, perhaps this mistake should not be looked at too harshly – as a "half-right" prediction of a dually localized protein is the *best possible* prediction for existing prediction methods which do not consider multiple localization at all. It seems likely that the current annotation in SWISS-PROT is conservative relative to multiple localization. For example in a recent large scale experiment using GFP fusion proteins to determine localization in yeast, approximately 20% out of a total of over 4000 measured proteins were found to dually localize to the nucleus and cytoplasm.

5. Conclusion

WoLF PSORT achieves a dramatic improvement in prediction accuracy over PSORTII while maintaining the simple, easily understood classifier which has been one reason for PSORTII's widespread use. Indeed by applying a feature selection algorithm, WoLF PSORT is actually *simpler* than PSORTII in the sense that it uses fewer features for classification. WoLF PSORT is also one of the most serious attempts to date to incorporate dually localized proteins into a prediction scheme for eukaryotic cells.

6. Acknowledgement

C.J.Collier designed the initial server task handler. H. Harada helps maintain the current server. Dr. H. Ohta gave valuable advice in preparing the entries extracted from GO. This work was partly supported by MEXT.

References

1. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
2. Hideo Bannai, Yoshinori Tamada, Osamu Maruyama, Kenta Nakai, and Satoru Miyano. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, 18:298–305, 2002.
3. B. Boeckmann, A. Bairoch, R. Apweiler, M. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31:365–370, 2003.
4. M.G. Claros and P. Vincens. Computational method to predict mitochondrially imported proteins and their targeting sequences. *European Journal of Biochemistry*, 241:779–786, 1996.
5. Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.
6. Paul Horton, Yuri Mukai, and Kenta Nakai. Protein localization prediction. In Limsoon Wong, editor, *The Practical Bioinformatician*, pages 193–215. World Scientific, 5 Toh Tuck Link, Singapore 596224, 2004.
7. Paul Horton and Kenta Nakai. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. In *Proceeding of the Fifth International Conference on Intelligent Systems for Molecular Biology*, pages 147–152, Menlo Park, 1997. AAAI Press.
8. José M. Mingot, Eduardo A. Espeso, Eliecer Díez, and Miguel A. Peñalva. Ambient pH signaling regulates nuclear localization of the *aspergillus nidulans* PacC transcription factor. *Molecular and Cellular Biology*, 21(5):1688–1699, March 2001.
9. Kenta Nakai and Minoru Kanehisa. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, 14:897–911, 1992.
10. Henrik Nielsen. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Engineering*, 12(1):3–9, 1999.
11. Henrik Nielsen, Jacob Engelbrecht, Søren Brunak, and Gunnar von Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 10:1–6, 1997.
12. K. Nishikawa and T. Ooi. Correlation of the amino acid composition of a protein to its structural and biological characters. *J. Biochem.*, 91(5):1821–1824, 1982.
13. Gunnar von Heijne. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research*, 14:4683–4690, 1986.