A statistical framework for genomic data fusion: Supplementary results

Gert R. G. Lanckriet
Department of Electrical Engineering and Computer Science
University of California, Berkeley

Tijl De Bie
ESAT-SCD
Katholieke Universiteit Leuven, Belgium

Nello Cristianini
Department of Statistics
University of California, Davis

Michael I. Jordan
Division of Computer Science
Department of Statistics
University of California, Berkeley

William Stafford Noble
Department of Genome Sciences
University of Washington

February 10, 2004

Table 2: **Classification performance on the cytoplasmic ribosomal class.** The table lists the percentage true positives at one percent false positives (TP1FP), the ROC score and the test set accuracy for each kernel and each combination of kernels. The column titled "Weight" shows the weight assigned to the kernel by SDP.

Kernel	Combination	TP1FP	ROC	Accuracy	Weight
K_B		72.80 ± 2.19	$.9810 \pm .0002$	$94.59 \pm 0.41\%$	0.45
K_{SW}		86.23 ± 1.70	$.9903 \pm .0012$	$96.77 \pm 0.26\%$	0.58
K_{Pfam}		50.72 ± 3.52	$.9479 \pm .0051$	$93.26 \pm 0.35\%$	0.01
K_E		98.31 ± 0.36	$.9995 \pm .0001$	$99.16 \pm 0.10\%$	4.85
K_{LI}		26.00 ± 2.44	$.8294 \pm .0081$	$91.08 \pm 0.37\%$	0.12
K_D		17.43 ± 1.29	$.8049 \pm .0115$	$88.04 \pm 0.43\%$	0.00
K_{RND1}		1.78 ± 0.59	$.5248 \pm .0092$	$87.55 \pm 0.45\%$	0.00
K_{RND2}		1.13 ± 0.33	$.5004 \pm .0081$	$87.55 \pm 0.45\%$	0.00
K_{RND3}		1.49 ± 0.43	$.5189 \pm .0104$	$87.55 \pm 0.45\%$	0.02
$K_{B,SW,Pfam,E,L,D}$	SDP	99.71 ± 0.17	$.9998 \pm .0000$	$99.29 \pm 0.09\%$	
$K_{B,\ldots,D,RND1}$	SDP	99.57 ± 0.20	$.9998 \pm .0000$	$99.25 \pm 0.11\%$	
$K_{B,\ldots,D,RND1,RND2,RND3}$	SDP	99.57 ± 0.20	$.9998 \pm .0000$	$99.26 \pm 0.09\%$	
$K_{B,\ldots,D}$	unweighted	99.91 ± 0.09	$.9999 \pm .0000$	$99.28 \pm 0.09\%$	
$K_{B,\ldots,D,RND1}$	unweighted	99.39 ± 0.27	$.9997 \pm .0000$	$99.17 \pm 0.10\%$	
$K_{B,\ldots,D,RND1,RND2,RND3}$	unweighted	99.15 ± 0.27	$.9997 \pm .0001$	$99.10 \pm 0.10\%$	

Table 3: **Consistently misclassified proteins: cytoplasmic ribosome.** The table lists proteins that are consistently misclassified by SDP/SVM. The score column lists the mean SVM discriminant across multiple splits.

ORF	Gene	Error	Score	Description
YLR287C-A	RPS30A	FN	-0.097	40S small subunit ribosomal protein
YPL131W	RPL5	FN	-0.162	60S large subunit ribosomal protein L5.e
YGL189C	RPS26A	FN	-0.272	40S small subunit ribosomal protein S26e.c7
YFL034C-A	RPL22B	FN	-0.286	ribosomal protein
YLR406C	RPL31B	FN	-0.313	60S large subunit ribosomal protein L31.e.c12
YIL069C	RPS24B	FN	-0.510	40S small subunit ribosomal protein S24.e
YDL130W	RPP1B	FN	-0.524	60S large subunit acidic ribosomal protein L44prime

Table 4: **Unannotated genes predicted to participate in the cytoplasmic ribosome.** Descriptions that include the phrase "across from" indicate the presence of a ribosomal protein on the opposite strand.

ORF	Gene	Score	Description
YPL142C		1.14964	questionable ORF (across from YPL143W)
YPR044C		0.88675	questionable ORF (across from YPR043W)
YDR417C		0.87126	questionable ORF (across from YDR418W)
YLR062C	BUD28	0.82710	questionable ORF (across from YLR061W)
YGL102C		0.82697	questionable ORF (across from YGL103W)
YLL044W		0.73161	questionable ORF (across from YLR045C)
YLR339C		0.58744	questionable ORF (across from YLR340W)
YJL188C		0.45662	questionable ORF (across from YJL189W)
YNL119W		0.39821	weak similarity to M.jannaschii hypothetical protein MJ1257
YKL056C		0.35834	strong similarity to human IgE-dependent histamine-releasing factor
YLR150W	STM1	0.27035	specific affinity for guanine-rich quadruplex nucleic acids
YLR076C		0.15068	questionable ORF (across from YLR075W)
YML022W	APT1	0.07361	adenine phosphoribosyltransferase
YEL026W	SNU13	0.05930	component of the U4/U6.U5 snRNP

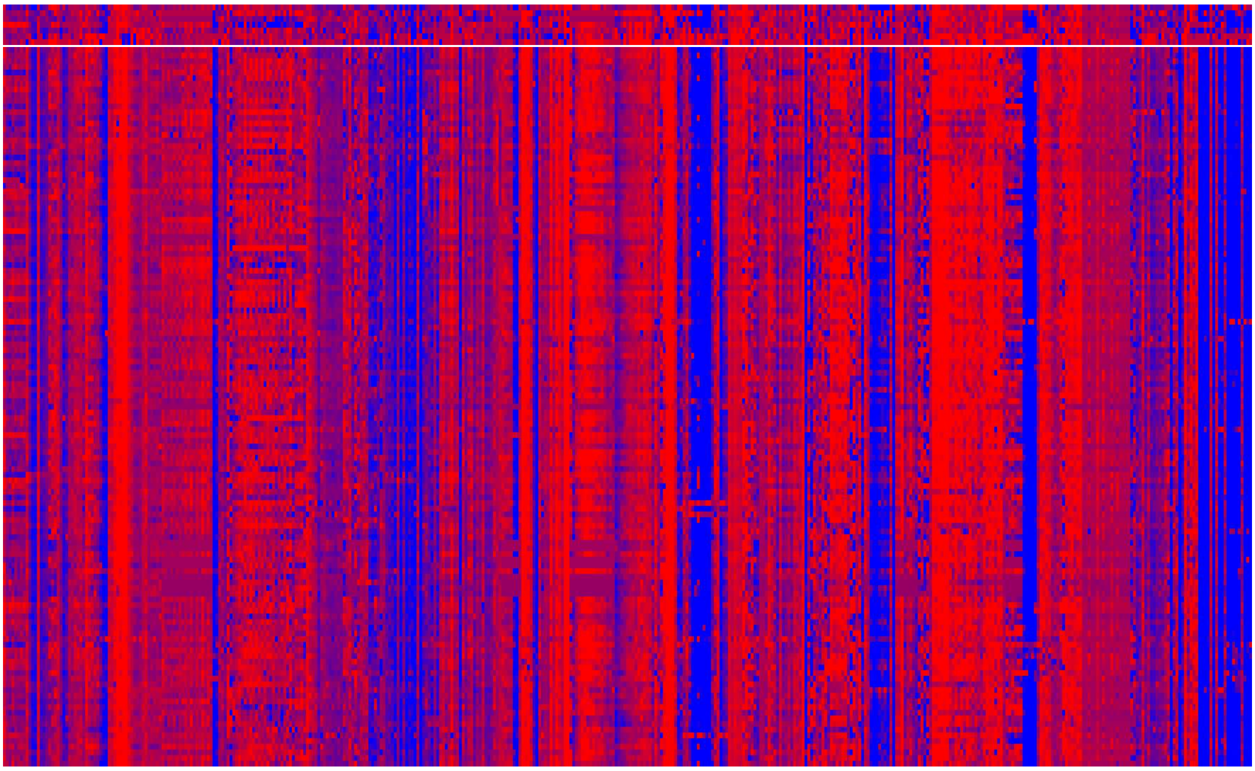

Figure 2: **Expression profiles of the ribosomal genes.** Rows in the matrix correspond to ribosomal genes, and columns correspond to microarray experiments. Each entry in the matrix corresponds to one mRNA expression measurement, with blue corresponding to low values and red corresponding to high values. The profiles of the seven genes that are classified as false negatives by SVM/SDP appear at the top of the picture.

Table 5: **Performance of the SDP/SVM method for membrane protein classification using various combinations of kernels.** Each row in the table corresponds to one experiment, classifying the 497 known yeast membrane proteins versus the 1876 known non-membrane proteins in yeast. The data is split into train and test sets in a ratio of 80/20, and the classifier is a 1-norm soft margin SVM with C=1. The first seven columns indicate the average weight assigned via SDP to each of the seven kernel matrices. A hyphen indicates that the corresponding kernel is not considered in the combination. The rightmost columns list three performance metrics, percentage true positives at one percent false positives (TP1FP), ROC score and test set accuracy (TSA), along with standard deviations computed across 30 randomly generated 80/20 splits.

K_B	K_{SW}	K_{Pfam}	K_{HF}	K_{LI}	K_D	K_E	K_{RND}	TP1FP	ROC	TSA
1.00	–	–	–	–	–	–	–	32.79 ± 1.59%	.8371 ± .0031	83.77 ± 0.27%
–	1.00	–	–	–	–	–	–	23.57 ± 1.67%	.8096 ± .0033	84.94 ± 0.28%
–	–	1.00	–	–	–	–	–	30.15 ± 1.38%	.8382 ± .0038	85.52 ± 0.23%
–	–	–	1.00	–	–	–	–	24.10 ± 0.94%	.7725 ± .0048	83.31 ± 0.27%
–	–	–	–	1.00	–	–	–	15.87 ± 0.76%	.7320 ± .0047	81.21 ± 0.29%
–	–	–	–	–	1.00	–	–	17.15 ± 0.87%	.8487 ± .0039	81.30 ± 0.27%
–	–	–	–	–	–	1.00	–	12.62 ± 1.08%	.7522 ± .0045	80.06 ± 0.30%
–	–	–	–	–	–	–	1.00	1.46 ± 0.24%	.5136 ± .0045	78.38 ± 0.31%
1.41	0.59	–	–	–	–	–	–	34.38 ± 1.87%	.8647 ± .0026	87.26 ± 0.23%
–	–	–	–	0.10	1.90	–	–	17.33 ± 0.98%	.8535 ± .0038	81.24 ± 0.29%
1.56	–	–	–	–	0.44	–	–	37.45 ± 1.60%	.8963 ± .0024	86.65 ± 0.25%
–	1.19	–	–	–	0.81	–	–	28.85 ± 2.05%	.8822 ± .0030	87.35 ± 0.21%
1.92	–	–	–	0.08	–	–	–	35.18 ± 1.25%	.8690 ± .0029	85.71 ± 0.26%
–	1.74	–	–	0.26	–	–	–	25.72 ± 1.76%	.8462 ± .0030	86.11 ± 0.23%
1.55	0.85	–	–	–	0.60	–	–	36.62 ± 2.19%	.9060 ± .0022	88.18 ± 0.22%
2.30	–	–	–	0.01	0.69	–	–	37.07 ± 1.73%	.8952 ± .0024	86.92 ± 0.24%
1.91	0.95	–	–	0.13	–	–	–	34.45 ± 1.85%	.8821 ± .0025	87.64 ± 0.23%
–	1.91	–	–	0.04	1.05	–	–	28.83 ± 2.05%	.8759 ± .0031	87.19 ± 0.21%
–	1.27	0.73	–	–	–	–	–	28.11 ± 1.64%	.8465 ± .0034	86.58 ± 0.23%
–	–	–	0.71	–	1.29	–	–	30.83 ± 1.60%	.8588 ± .0033	85.87 ± 0.24%
–	1.72	0.92	0.37	–	–	–	–	28.15 ± 1.38%	.8434 ± .0035	86.39 ± 0.20%
–	1.42	0.70	–	–	0.88	–	–	32.22 ± 1.80%	.8926 ± .0027	87.74 ± 0.19%
–	1.73	0.87	0.33	–	1.07	–	–	32.33 ± 1.77%	.8920 ± .0028	87.74 ± 0.19%
2.77	1.43	0.54	0.33	0.15	–	0.78	–	34.52 ± 1.91%	.9020 ± .0025	88.09 ± 0.23%
2.54	1.50	0.47	0.33	0.00	1.16	–	–	35.88 ± 2.09%	.9079 ± .0024	88.26 ± 0.23%
2.62	1.52	0.57	0.35	0.00	1.21	0.73	–	36.06 ± 1.95%	.9219 ± .0024	88.66 ± 0.24%
2.97	1.73	0.73	0.42	0.00	1.18	0.86	0.09	35.56 ± 1.89%	.9186 ± .0024	88.36 ± 0.26%
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	36.66 ± 1.83%	.9049 ± .0026	88.43 ± 0.22%

Table 6: **"Distance to uniformity" of the ranking of membrane and non-membrane proteins with signal peptides, as provided by SVM and TMHMM.** Columns in the table correspond respectively to Figures 3, 4 and 5. The "distance to uniformity" for the ranking of non-membrane proteins (DU_{neg}) with signal peptides is obtained by plotting the cumulative absolute value of a given score (NN or HMM) of the below-the-zero-line points, and then computing the normalized 1-norm distance to the cumulative absolute value if the distribution was perfectly uniform, i.e., the line segment connecting the first and last point in the cumulative plot. The distance to uniformity for ranking of the membrane proteins (DU_{pos}) with signal peptides is obtained in a similar way, using the score of the above-the-zero-line points. Bold values indicate better behavior.

Signal Peptide	SVM		T_{ENR}		T_{PH}	
Prediction Method	DU_{neg}	DU_{pos}	DU_{neg}	DU_{pos}	DU_{neg}	DU_{pos}
NN	**0.15**	**0.66**	0.34	**0.69**	0.21	0.49
HMM	**0.17**	**0.64**	0.43	**0.65**	**0.16**	0.47

1 Proteins with Signal Peptides

Figures 3, 4 and 5 and Table 6 illustrate the superior behavior of the SDP/SVM method with respect to proteins that contain signal peptides, as compared to TMHMM.

While the SDP/SVM algorithm is a discriminative method that attempts to find a decision boundary that separates positive and negative instances of membrane proteins, the TMHMM is a generative method that simply attempts to model the membrane proteins. As an illustration of the difference, it is known that the TMHMM tends to yield false positives for sequences containing signal peptides—hydrophobic sequences in the N-terminal regions of proteins. The SDP/SVM approach tends to avoid these false positives, because signal peptides appear among the negative instances in the training set. Indeed, as we show in the online supplement, signal peptides tend to be highly ranked by the TMHMM, and are more uniformly spread within the SDP/SVM rankings.

Signal peptides are identified by the SignalP web server (`www.cbs.dtu.dk/services/SignalP-2.0`). The server provides two types of predictions, based upon a neural network and an HMM. Here, the neural network score (NN) is the sum of the four values output by SignalP. Similarly, the HMM score is the sum of the signal peptide and signal anchor probabilities.

The figures show two complementary effects. First, many non-membrane proteins (points under the zero line) are ranked highly by T_{ENR}, while they are spread more uniformly over the ranking by T_{PH} and the SVM approach. This observation is confirmed by measuring the "distance to uniformity" for the three approaches (Table 6). This effect illustrates the sensitivity of T_{ENR} to signal peptides in non-membrane proteins, yielding false positives. Second, although both SVM and T_{PH} tend to rank the non-membrane proteins with signal peptides about equally uniformly (when using HMM signal peptide predictions), T_{PH} ranks the true membrane proteins with signal peptides quite uniformly as well. This effect, which is also confirmed in Table 6, leads to a high false negative rate for T_{PH}.

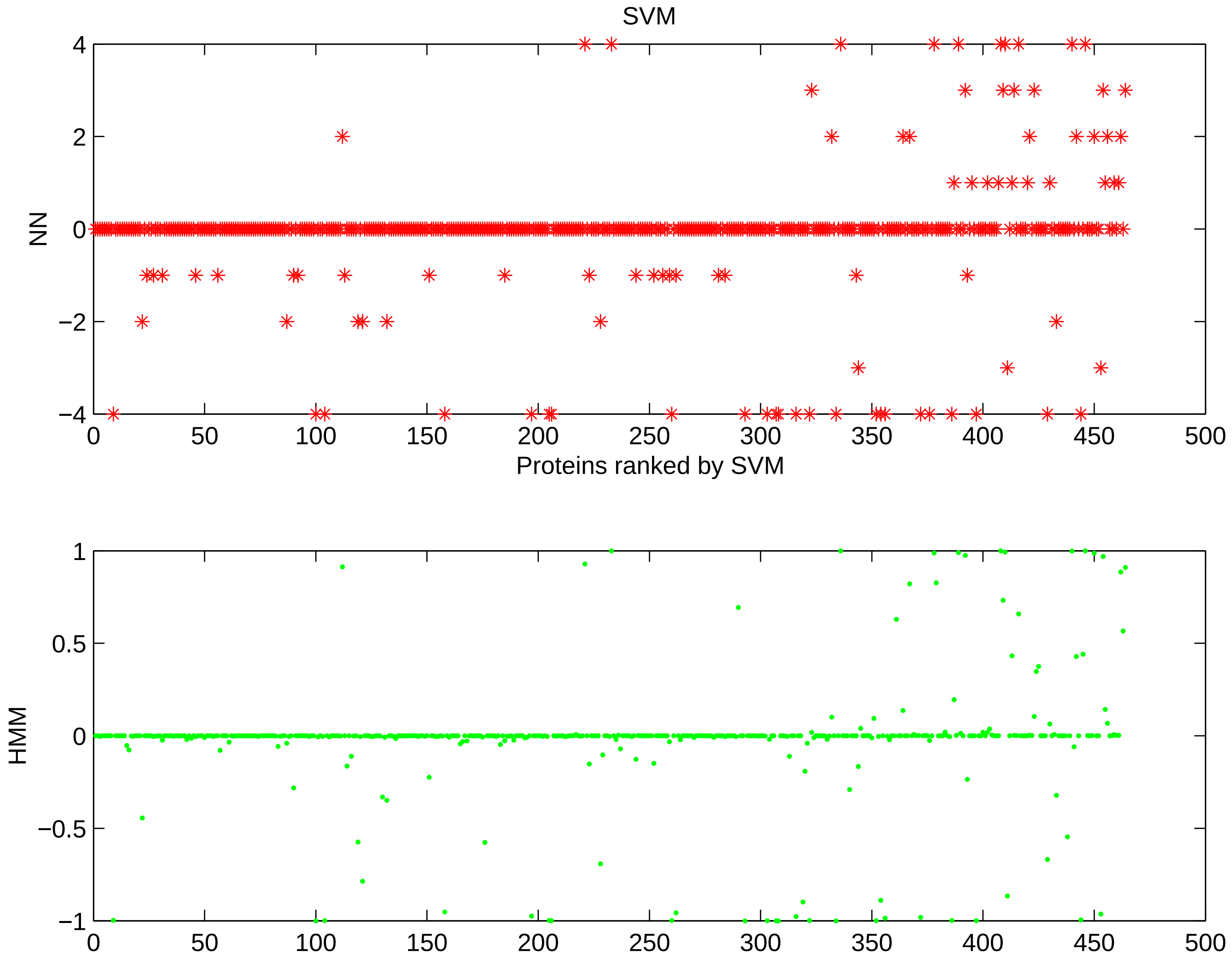

Figure 3: **Ranking of proteins by SVM, highlighting signal peptide properties.** The vertical axis plots the value of the NN and HMM scores multiplied by the true label of the protein (1 or -1). Hence, points below zero correspond to non-membrane proteins, while points above zero correspond to membrane proteins. The horizontal axis is the ranking of proteins induced by the SVM, with predicted membrane proteins on the left.

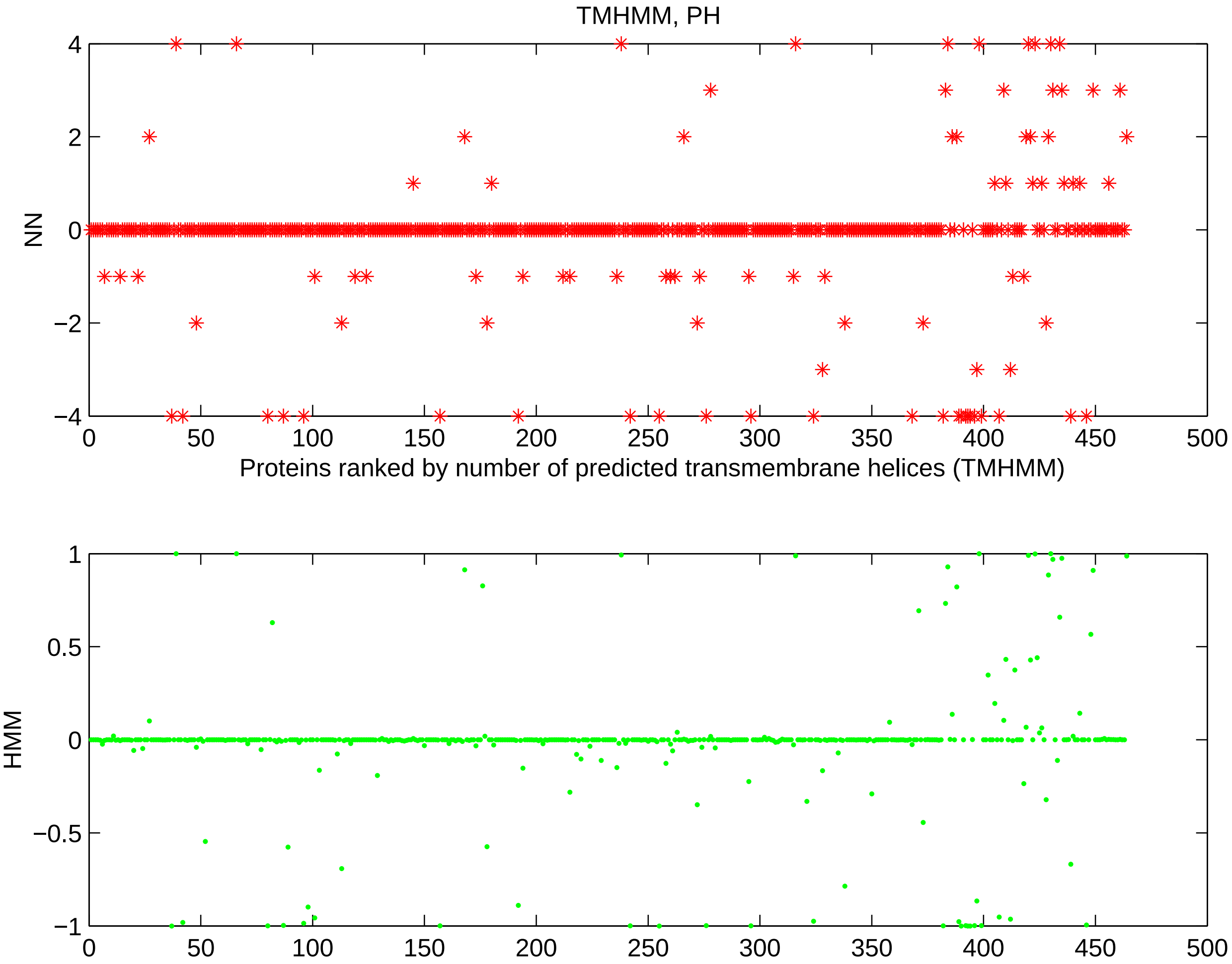

Figure 4: **Ranking of proteins by the number of TMHMM predicted transmembrane helices (T_{PH}), highlighting signal peptide properties.** This plot is similar to Figure 3.

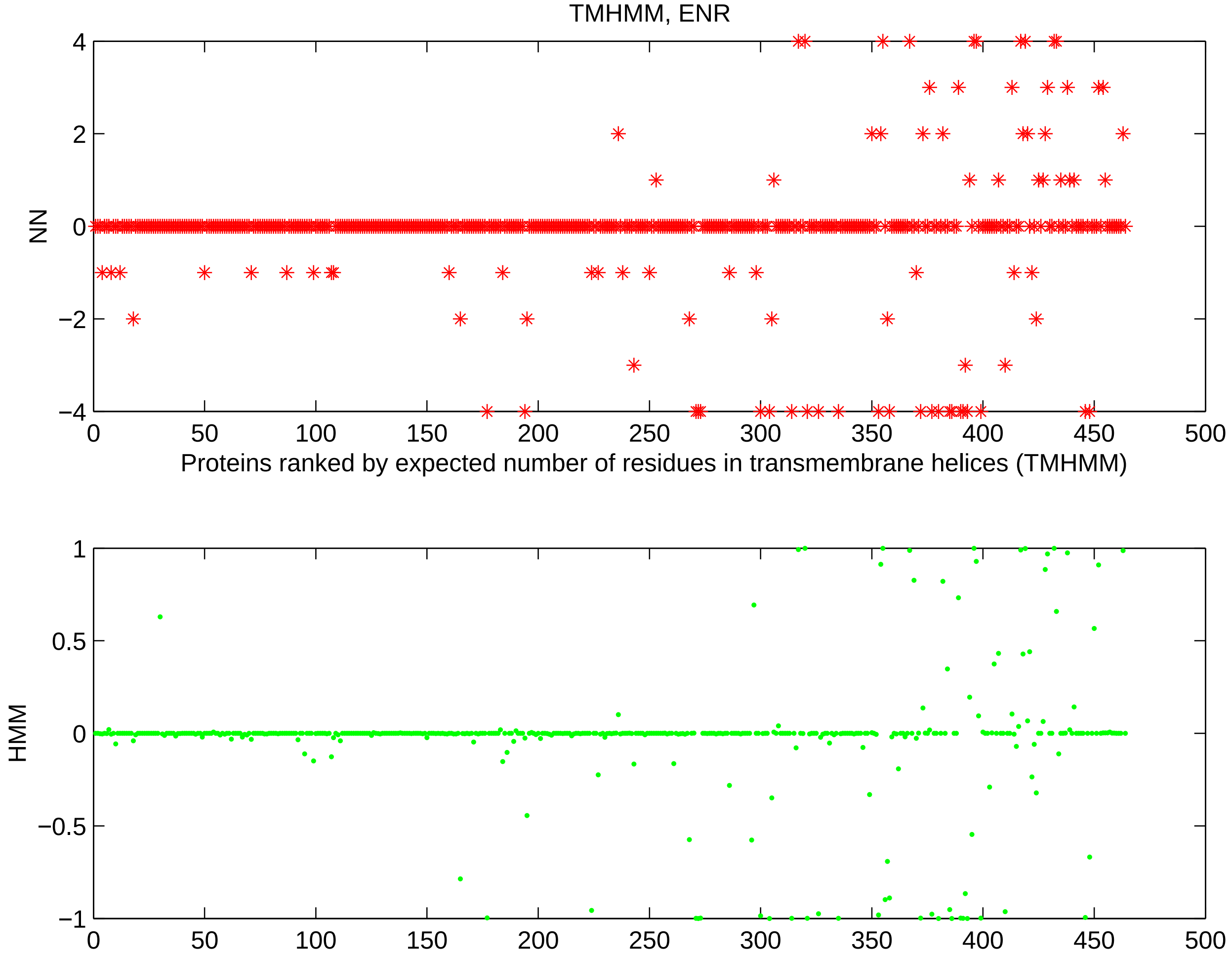

Figure 5: **Ranking of proteins by the TMHMM expected number of residues in transmembrane helices (T_{ENR}), highlighting signal peptide properties.** This plot is similar to Figure 3.