Identification of Prokaryotic Small Proteins using a Comparative Genomic Approach: Supplementary Material

Josue Samayoa, Fitnat Yildiz, Kevin Karplus

April 25, 2011

1 Approach

Our method combines predictive features derived from structure prediction tools, comparative genomics, and sequence composition analysis (Figure 1). In order to determine whether a given sequence codes for a small protein, we begin with a multiple genome alignment for our target organism. We then use this alignment to generate scores for each sequence based on three categories of analysis. The first analysis we perform is to analyze the observed codon composition for a given sequence according to a log-odds score. We score each sequence for agreement with its genome's codon biases on long protein genes. Second, we analyze each sequence for protein-like conservation patterns in the multiple sequence alignment. We score an alignment of a homologous sequence to the target sequence according to a BLOSUM90 substitution matrix. We then compare this to the score of the target sequence aligned to itself. We expect homologous sequences that code for proteins to have a score similar to the target self-alignment score. Finally, we look for prediction strength and consistency among a set of local structure alphabets. For each sequence we generate three independent predictions for a given local structure alphabet and measure their overall agreement. We hypothesize that sequences coding for a protein will generate more consistent predictions than sequences not coding for a protein. We combine these scores, for a set of positive and negative training examples, to generate a model which we use to predict on new sequences.

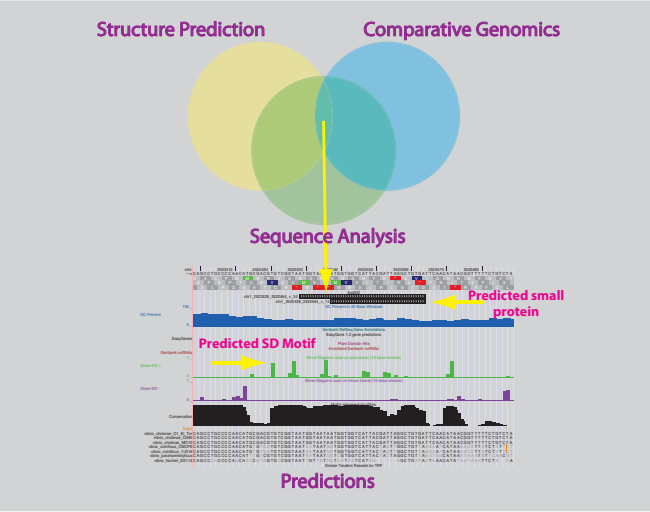

Figure 1: Method overview. Our approach combines predictive features derived from structure prediction, comparative genomics, and sequence analysis to generate a prediction as to whether a given open reading frame (ORF) encodes for a protein. Once we have identified candidate ORFs, we can map these coordinates onto the corresponding genome and visualize the results with the UCSC Microbial Genome Browser. This allows us to look for other hallmarks of protein encoding genes such as the presence of a ribosome binding site (Shine-Dalgarno Motif).

2 Local structure alphabets

We generated predictions for a set of 15 local structure alphabets. Included in this set is the protein blocks alphabet (alphabet composed of 16 average protein fragments of 5 residues in length) [3], an alphabet of ϕ-ψ angles developed by Bystroff [2], and several novel alphabets developed in the Karplus Lab [6, 5, 1]. The set of novel alphabets includes str2 (an extension of the DSSP [4] secondary structure alphabet with an expanded beta sheet class); alpha (an 11-letter alphabet based on the torsion angle, alpha, which is defined for residue i as the angle between C_α atoms of residue $i-1$, i, $i+1$, and $i+2$); o-sep and n-sep (hydrogen bond type classification based on the chain separation between hydrogen donor and acceptor); o-notor and n-notor (similar to o-sep and n-sep, but instead of chain separation looks at the torsion angle between the peptide planes of the hydrogen donor and acceptor); o-notor2 and n-notor2 (based on o-notor and n-notor respectively but also requires a minimum chain separation of 5 and classifies multiple hydrogen bonds into 3 separate categories); strand-sep (classifies beta strands based on the separation between the beta partners); str4 (a

combination of several different alphabets including str2, notor2, and alpha); CB8-sep9, CB-burial-14-7, and near-backbone-11 (burial alphabets designed to classify based on the number of residues observed to occur within imaginary spheres centered around points near each residue).

3 Weights for protein conservation analysis

This section describes and lists the weights used for each homolog genome in the BLSOUM-loss calculations. Each genome was weighted according to how related the homologous species was to the target genome, *E. coli* K12. Weights were calculated by taking the complete genome alignment between *E. coli* K12 and each homologous species genome and calculating the percent identity across all aligned regions. This percent identity was then subtracted from 1 to generate the weight. According to this calculation then, species very similar to *E. coli* K12 were down-weighted while more distantly related species were given a higher weight. Table 1 lists all the weights used in the BLOSUM-loss calculation in descending order.

Species	Weight
Blochmannia floridanus	0.425
Buchnera aphidicola	0.421
Yersinia pestis	0.316
Enterobacter 638	0.235
Salmonella enterica ATCC 9150	0.209
Salmonella enterica CT18	0.209
E. coli UTI89	0.039
E. coli APEC O1	0.038
E. coli CFT073	0.038
E. coli 536	0.037
E. coli 0127 H6 E2348 69	0.036
E. coli SECEC SMS-3-5	0.034
E. coli O157H7 EDL933	0.029
E. coli O157H7 EC4115	0.028
E. coli O157H7	0.028
Shigella flexneri	0.025
E. coli E24377A	0.021
E. coli SE11	0.020
E. coli HS	0.016
E. coli C-ATCC-8739	0.014
E. coli DH10B	0.002
E. coli W3110	0.001

Table 1: Table of weights used for BLOSUM-loss score. Each weight was determined by first calculating the sequence identity in a complete genome alignment between each homologous species and *E. coli* K12. The percent identity was then subtracted from one. Species closely related to *E. coli* K12 were given a lower weight according to this calculation and more distantly related organisms were given higher weight.

4 Distributions of individual scoring features

We looked at how well each individual feature discriminated the two training populations by creating a histogram of each set's scores (Figures 2 through 18). It was clear that by all measures the set of true protein coding sequences had a very distinguishable distribution. This is best evidenced by the scores resulting from our codon bias and BLOSUM-loss calculations, Figure 2 and Figure 3 respectively.

For the set of all GenBank-annotated genes 1000 bases or longer (Positive training set), the resulting codon bias scores indicated better agreement with the genome's codon bias composition model than with the background model based solely on GC-content. On the other hand the set of all non-annotated ORFs (Negative training set) contained a much broader distribution with more ORFs scoring negatively indicating better agreement with the background codon composition model. The BLOSUM-loss measure yielded similar resolution between the positive and negative training examples. As expected the set of true protein codon sequences had, on average, significantly smaller BLOSUM-loss scores than the set of all non-annotated sequences. As with the codon bias measure, the negative data set had a much broader distribution than the positive data set.

Agreement among the local structure predictions also yielded good discrimination between the positive and negative training data. Shown in Figures 4 through 18 are the distributions of prediction consistency scores for 15 local structure alphabets. In 14 out of the 15 alphabets, the predictions for the true protein coding sequences were more consistent than predictions for the set of all non-annotated ORFs.

There was one notable exception: the predictions for the burial alphabet CB-burial-14-7 were more consistent on average for the negative training data (Figure 8). Despite this observed behavior the scores for CB-burial-14-7 alphabet still resolved the negative and positive training data quite well. A possible explanation for the consistency on the negative training data is that the negative training examples contained more predictions for exposed residues. Predictions for buried residues were rare among this set of sequences. We hypothesized that the greater prediction agreement was due in large part to the higher than normal percentage of residues predicted to be solvent exposed. Unfortunately for this hypothesis, the other two burial alphabets (near-backbone-11 and CB8-sep-9) behaved more like the structural alphabets, being more consistent for the positive training data.

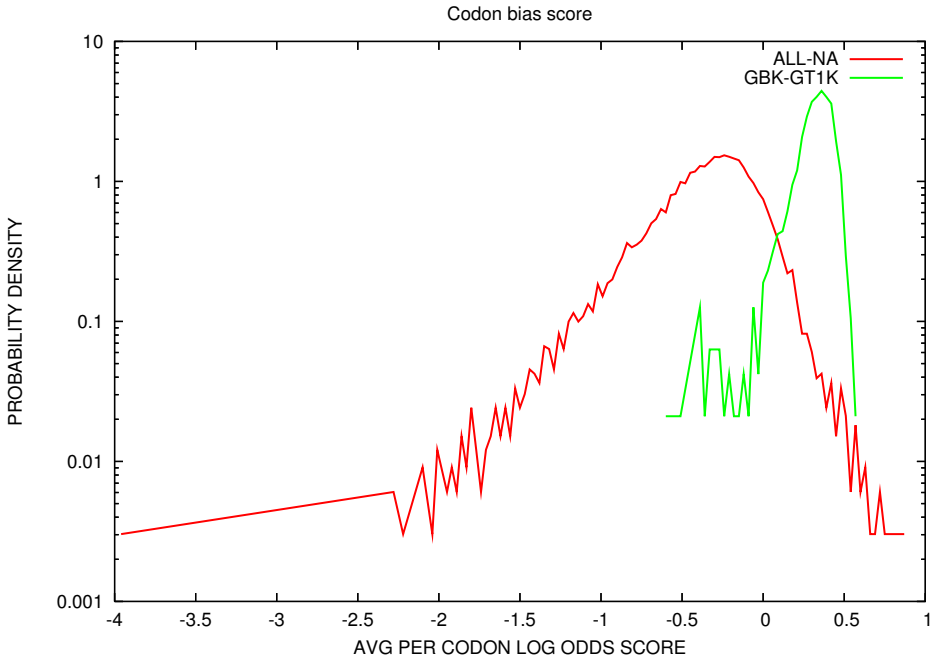

Figure 2: Distribution curves of codon bias scores for two sets of *E. coli* K12 ORFs. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (All-NA, red). The score is a log-odds metric that looks for selection of high expression amino acids in a given genome. Positive scores indicate better agreement with a genome's known codon bias model. Negative scores indicate better agreement with a model based on GC-richness.

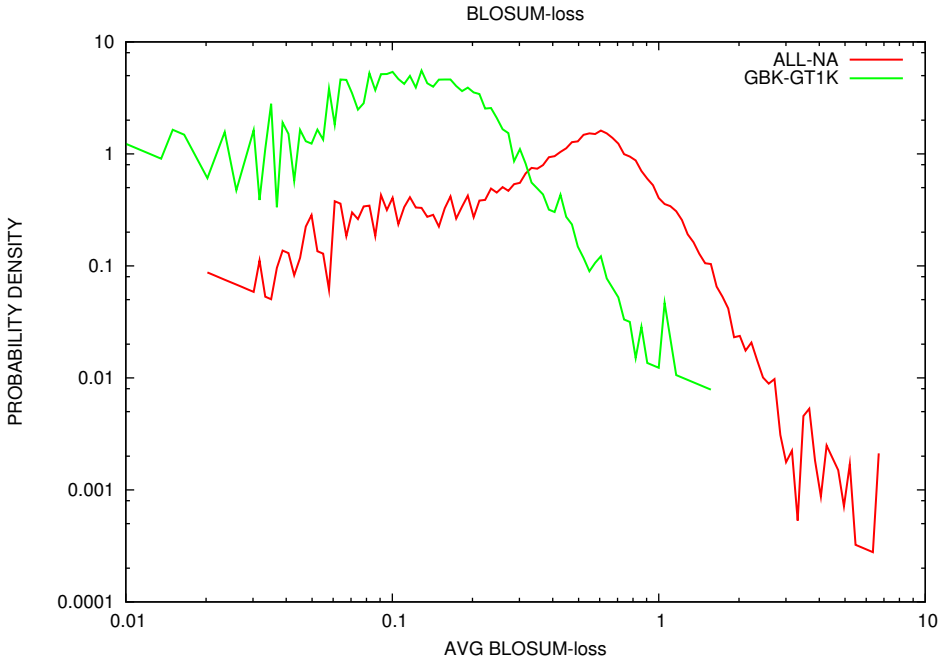

Figure 3: Distribution curves of BLOSUM-loss scores for two sets of *E. coli* K12 ORFs. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (All-NA, red). BLOSUM-loss is measured by looking at the ratio of target-to-homolog BLOSUM90 score vs the target-to-self-BLOSUM90 score. The ratio is averaged across all codons with at least 1 DNA mutation in 1 homologous sequence. The final score is subtracted from 1.01 for plotting purposes. A score of 0.01 represents perfect amino acid conservation while large scores represent little conservation of the amino acids.

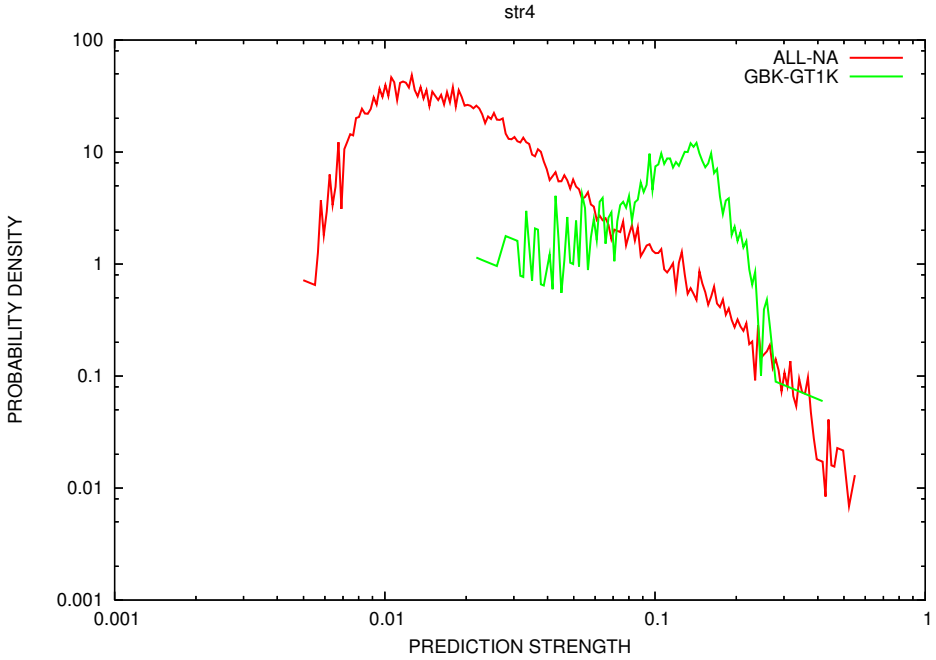

Figure 4: Distribution curves for prediction agreement among the str4 alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

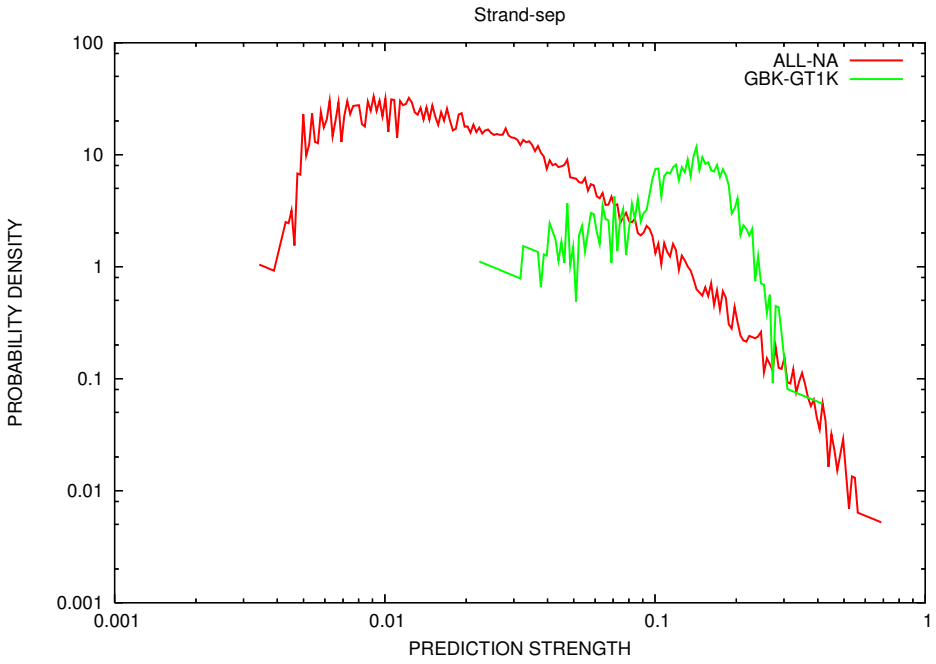

Figure 5: Distribution curves for prediction agreement among the strand-sep alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

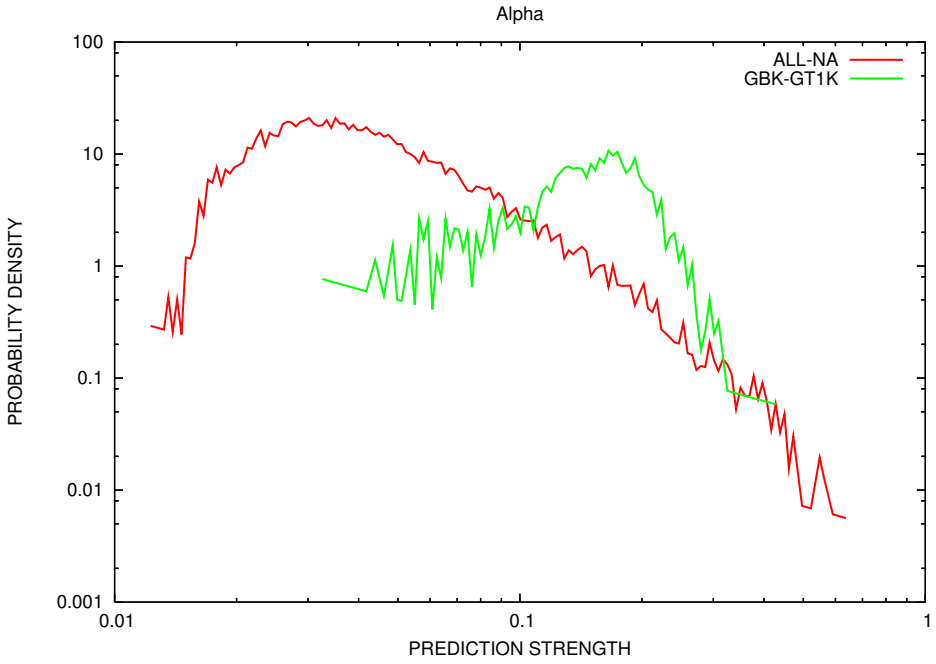

Figure 6: Distribution curves for prediction agreement among the alpha alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

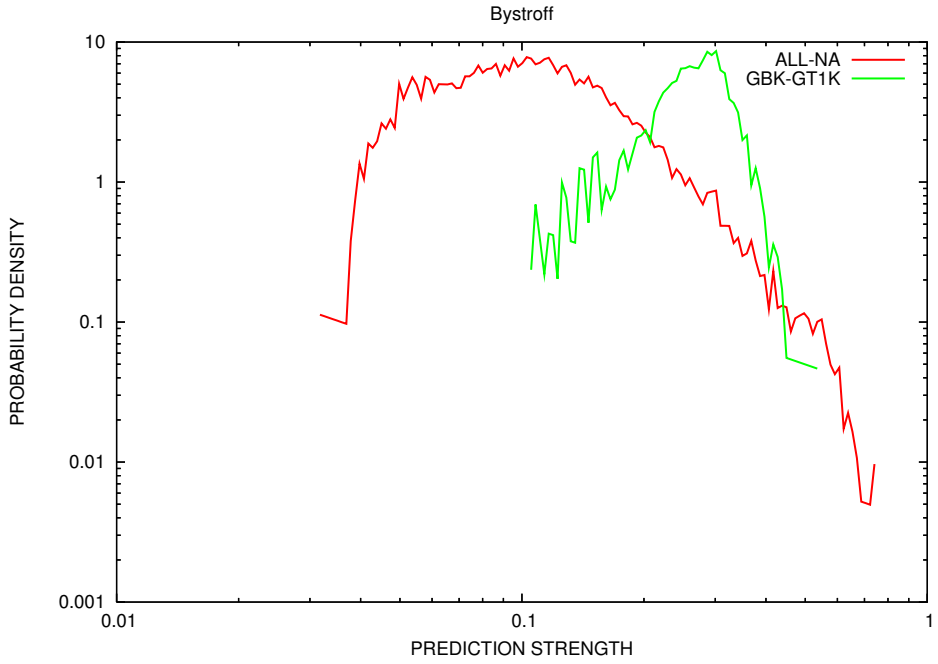

Figure 7: Distribution curves for prediction agreement among Bystroff alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

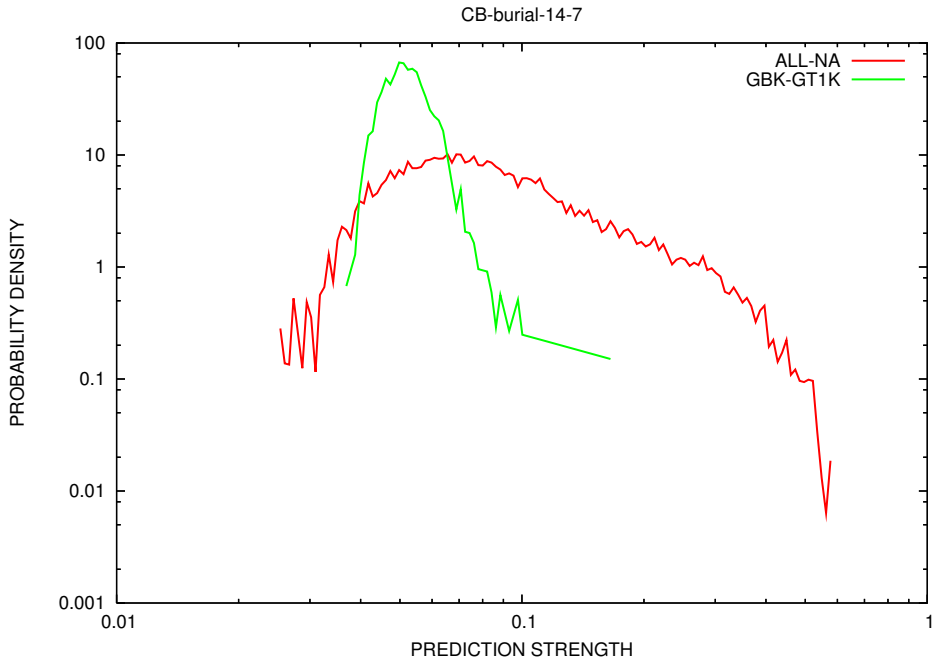

Figure 8: Distribution curves for prediction agreement among CB-burial-14-7 alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions. Surprisingly, the set of true proteins has lower prediction agreement scores than the small ORFs, unlike other local structure alphabets. Even the other burial alphabets, near-backbone-11 in Figure 13 and CB8-sep-9 in Figure 15 do not have this anomalous behavior.

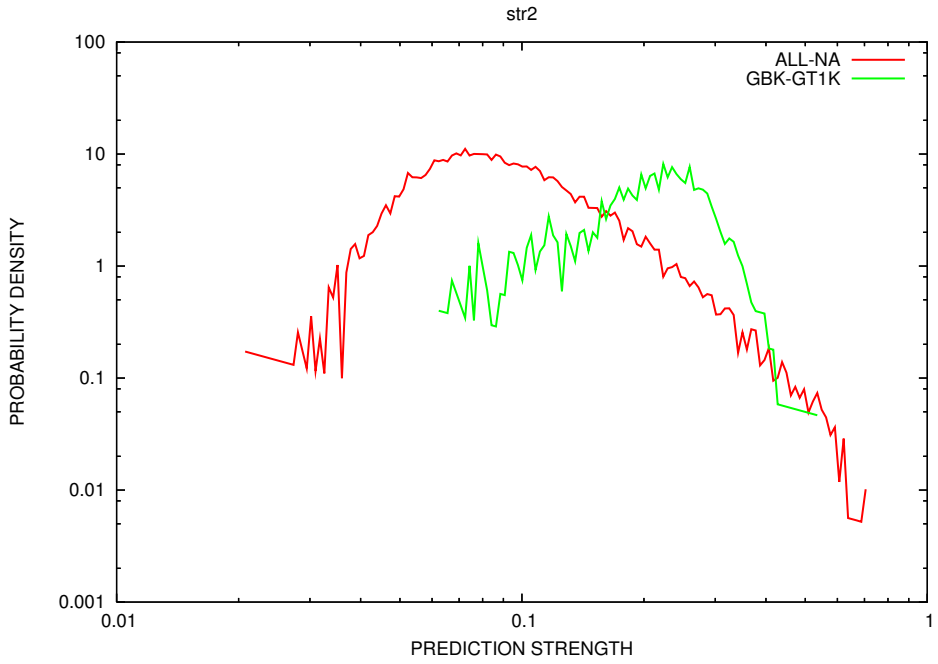

Figure 9: Distribution curves for prediction agreement among the str2 alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

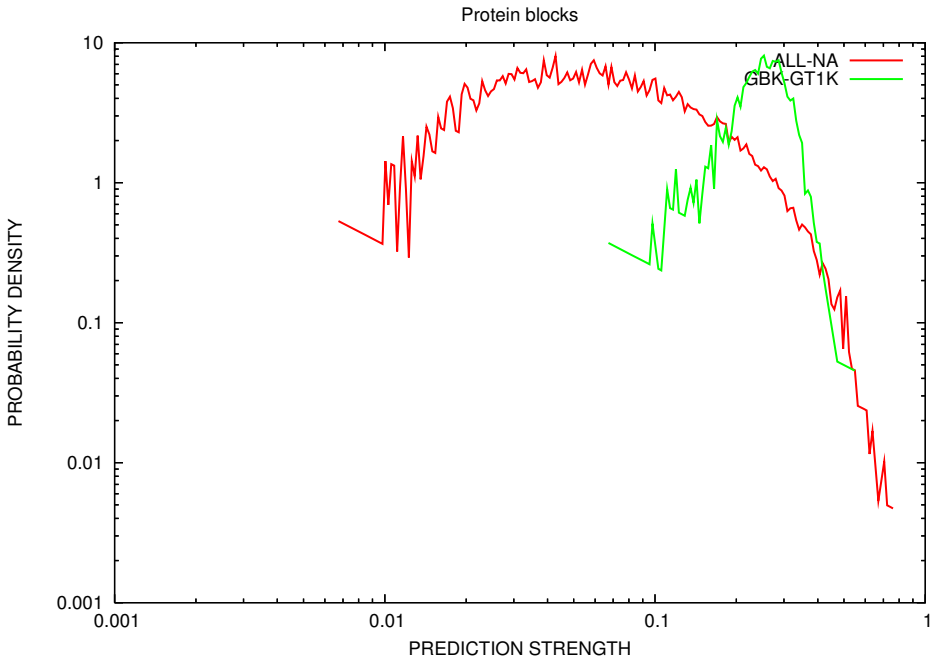

Figure 10: Distribution curves for prediction agreement among the protein blocks alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

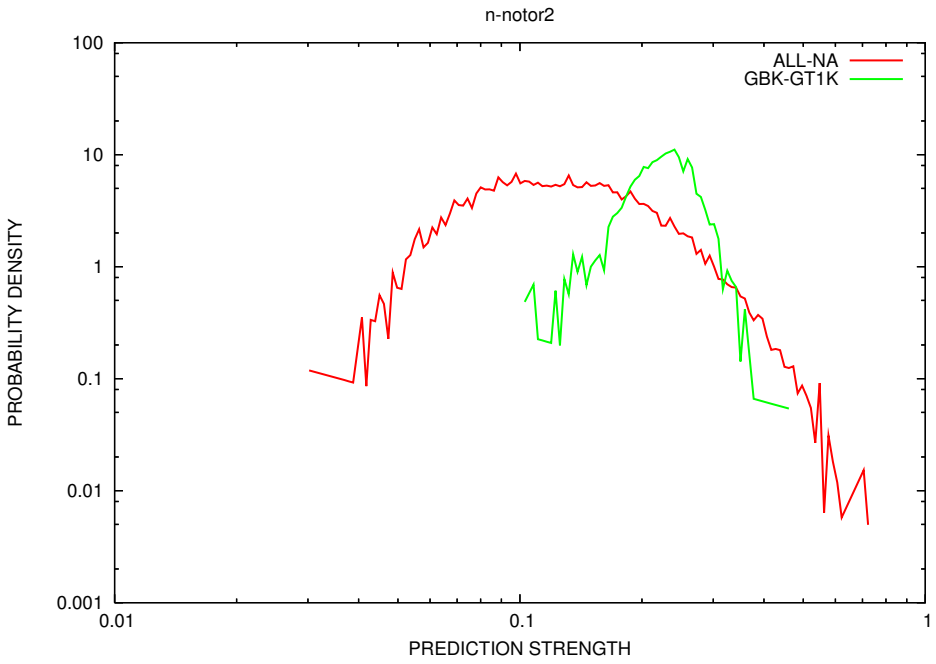

Figure 11: Distribution curves for prediction agreement among the n-notor2 alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

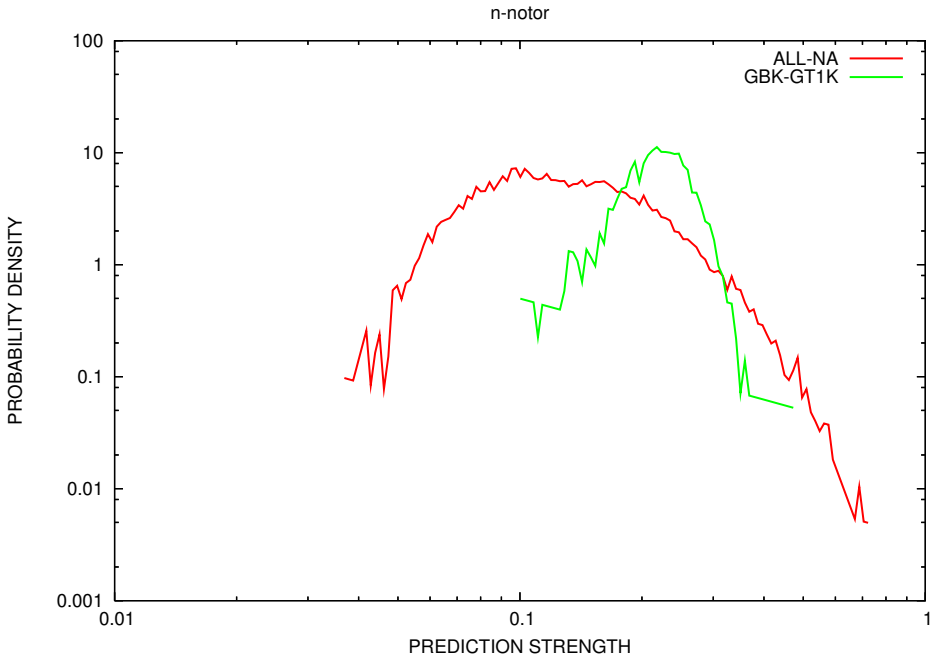

Figure 12: Distribution curves for prediction agreement among the n-notor alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

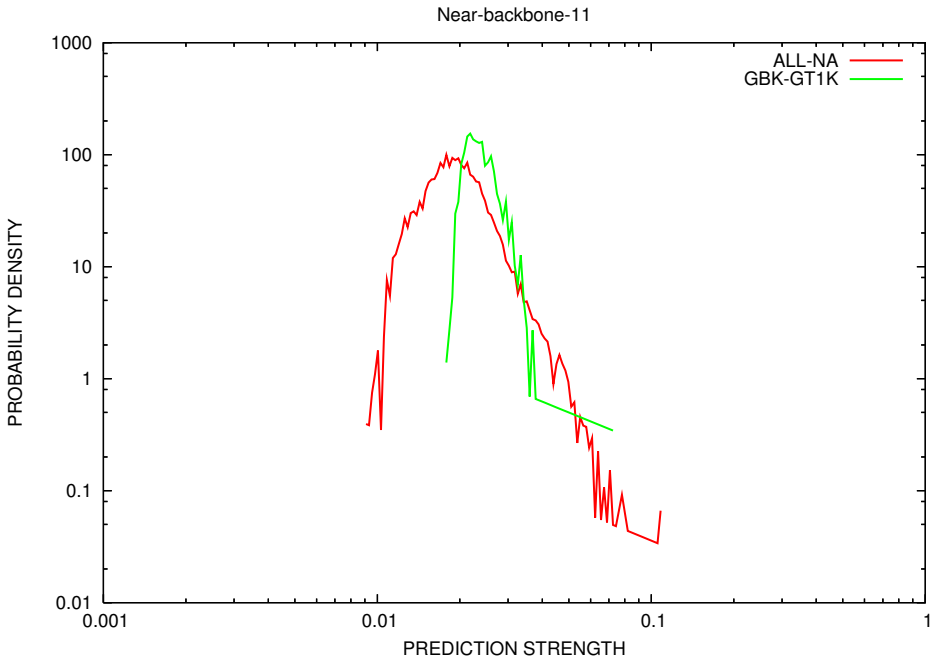

Figure 13: Distribution curves for prediction agreement among near-backbone-11 hydrogen bond alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

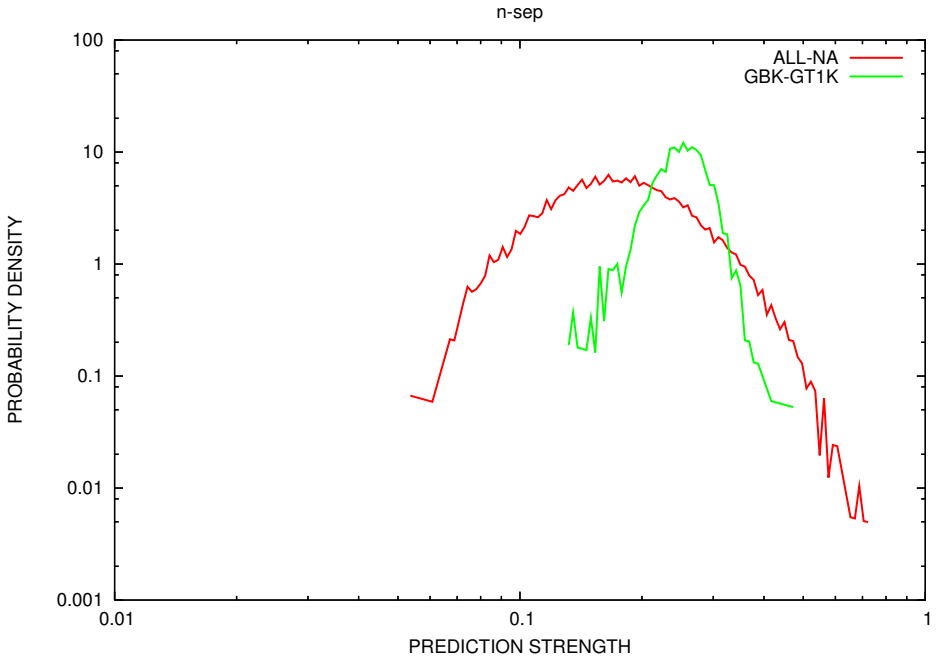

Figure 14: Distribution curves for prediction agreement among the n-sep alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

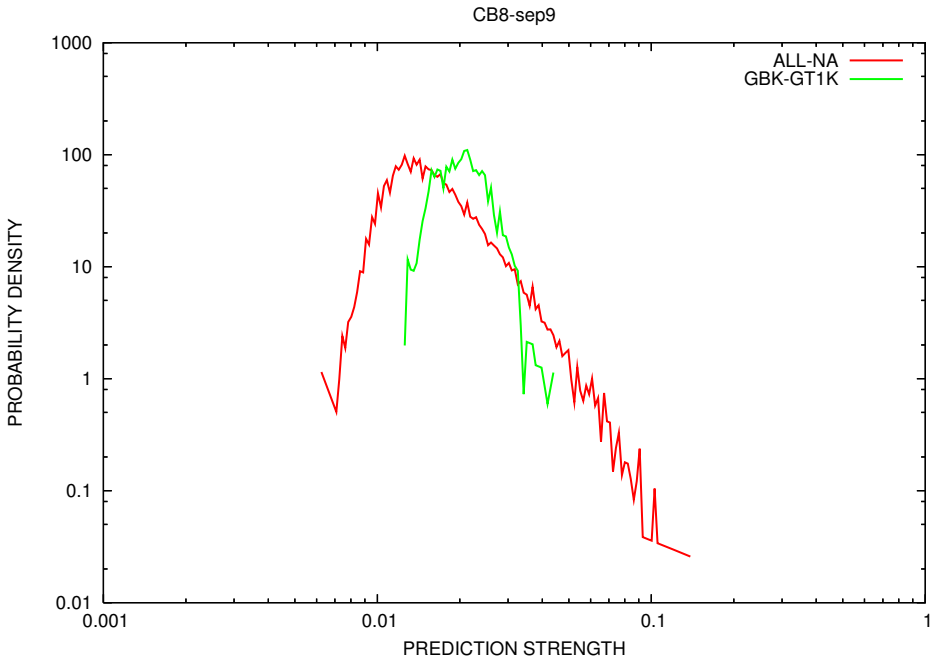

Figure 15: Distribution curves for prediction agreement among CB8-sep9 burial alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

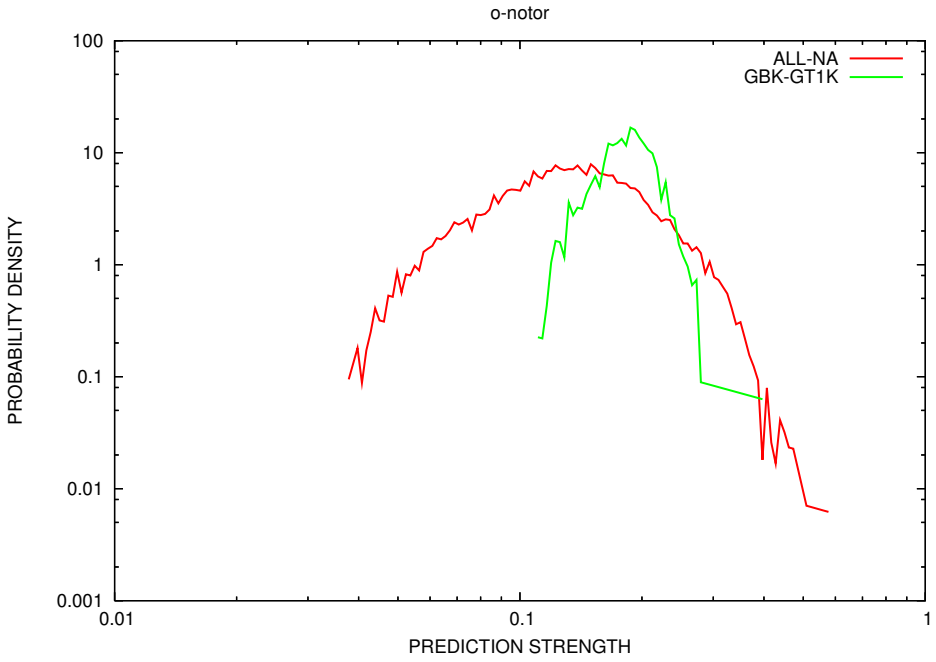

Figure 16: Distribution curves for prediction agreement among o-notor hydrogen bond alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

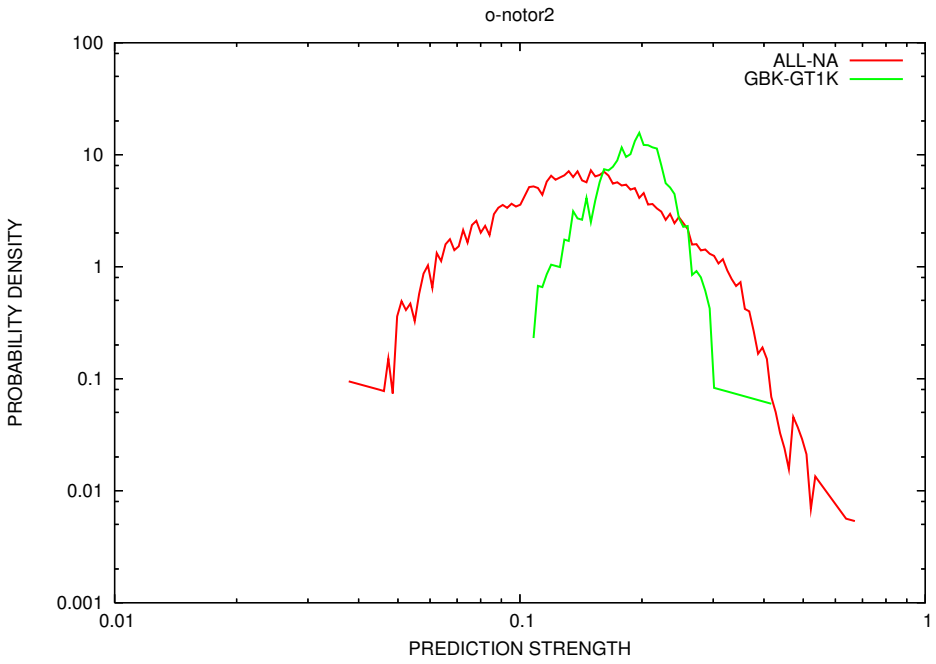

Figure 17: Distribution curves for prediction agreement among o-notor2 hydrogen bond alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (ALL-NA, Red). Higher scores indicate greater agreement among predictions.

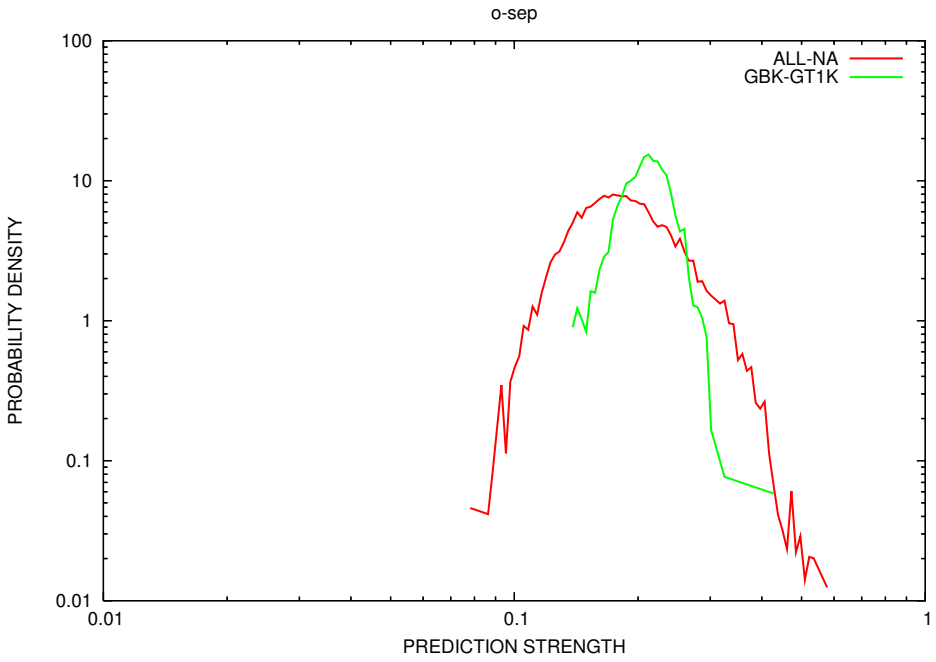

Figure 18: Distribution curves for prediction agreement among o-sep hydrogen bond alphabet predictions for ORFs in the *E. coli* K12 genome. Shown are the set of all GenBank-annotated ORFs 1000 bases or longer (GBK-GT1K, green) and the set of all possible ORFs with no more than 20% of their sequence overlapping with any GenBank annotations (All-NA, red). Higher scores indicate greater agreement among predictions.

Feature	1 Comp	2 Comp	3 Comp	4 Comp	5 Comp	6 Comp	7 Comp
Codon Bias	13.21	14.43	11.62	10.54	10.43	10.56	10.63
CB-burial-14-7		-74.47	-59.76	-105.29	-100.94	-110.73	-113.39
BLOSUM-loss			-5.54	-4.81	-4.56	-4.55	-4.59
CB8-sep-9				247.19	230.57	250.96	254.30
n-notor2					4.51	11.91	1.02
Bystroff						-7.53	-8.05
n-notor							12.07

Table 2: Weights for features from three fold cross validation experiments with multiple genome data. Table of logistic regression coefficients for each component in the optimal 1-, 2-, 3-, 4-, 5-, 6-, and 7-feature model. The final model consisted of codon bias score, CB-burial-14-7, BLOSUM-loss, CB8-sep-9, n-notor2, Bystroff, and n-notor added in that order.

5 Cross training results

The results from the systematic analysis of all 17 features was that the codon bias calculation was the best single feature. We generated true-positive-vs-false-positive curves for each feature. At 800 TPs, roughly half of positive training examples, the codon bias calculation has only 85 FPs. The next best single feature was the BLOSUM-loss score with 394 FPs after the first 800 TPs, a significant drop-off in performance. It seems that selection for efficient expression is stronger than selection to preserve the identities of the amino acids. Figure 19 shows the TP vs. FP curves for the cross-validation tests of the 7 models. We can clearly see the improved performance as we move from one to two, three, four, and even five features. After this point, however, we seem to have saturated our performance. The best 7-feature combination included codon bias, CB-burial-14-7, BLOSUM-loss, CB8-sep9, n-notor2, Bystroff, and n-notor, added in that order. Figure 20 shows the curves for each individual component of the best 7-feature model.

Table 2 lists the regression coefficients for each of the parameters in the optimal 1-, 2-, 3-, 4-, 5-, 6-, and 7-feature logistic regression models. Higher values indicate more emphasis placed on a particular feature. Positive values indicate a direct correlation between outcome and a given feature. Conversely, negative values indicate an inverse relationship between outcome and a given feature. The negative value was expected for BLOSUM-loss, which is minimized for true proteins, but was somewhat surprising for the local structure alphabet CB-burial-14-7

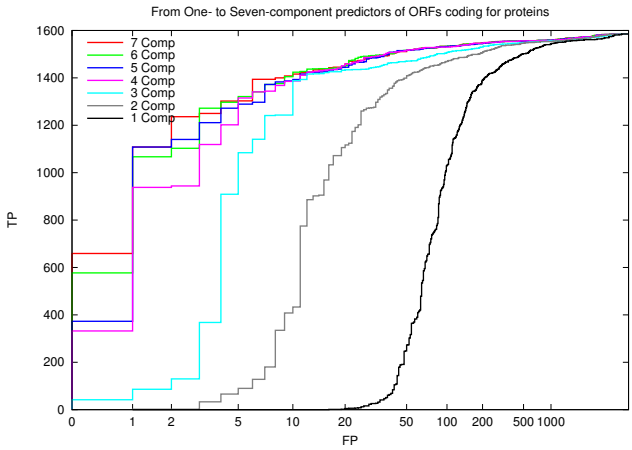

Figure 19: Average cross-validation results for 1-feature through 7-feature logistic regression models. Features were added in the order codon bias, CB-burial-14-7, BLOSUM-loss, CB8-sep9, n-notor2, Bystroff, n-notor. Performance improves only slightly after 5 features have been added.

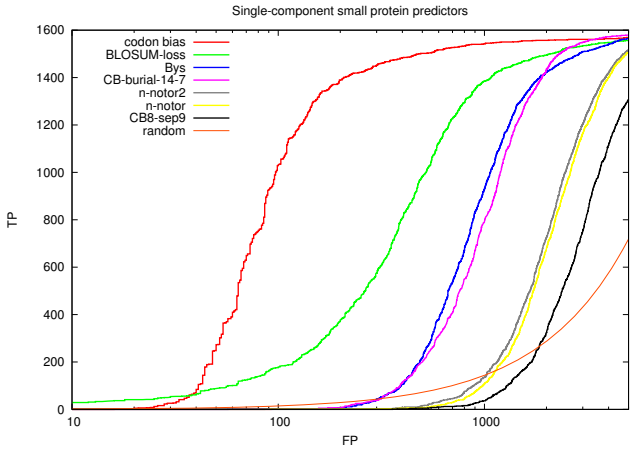

Figure 20: Average cross-validation results for single feature logistic regression models. These are all of the metrics included in the optimal 7-feature logistic regression model. The codon-bias score is clearly the best performing single scoring feature. Note that the order the feature was added does not correspond to their performance order as a single feature. For example, the BLOSUM-loss measure is the second best here but carries much of the same information as codon bias, and so is the 3rd feature added, not the second.

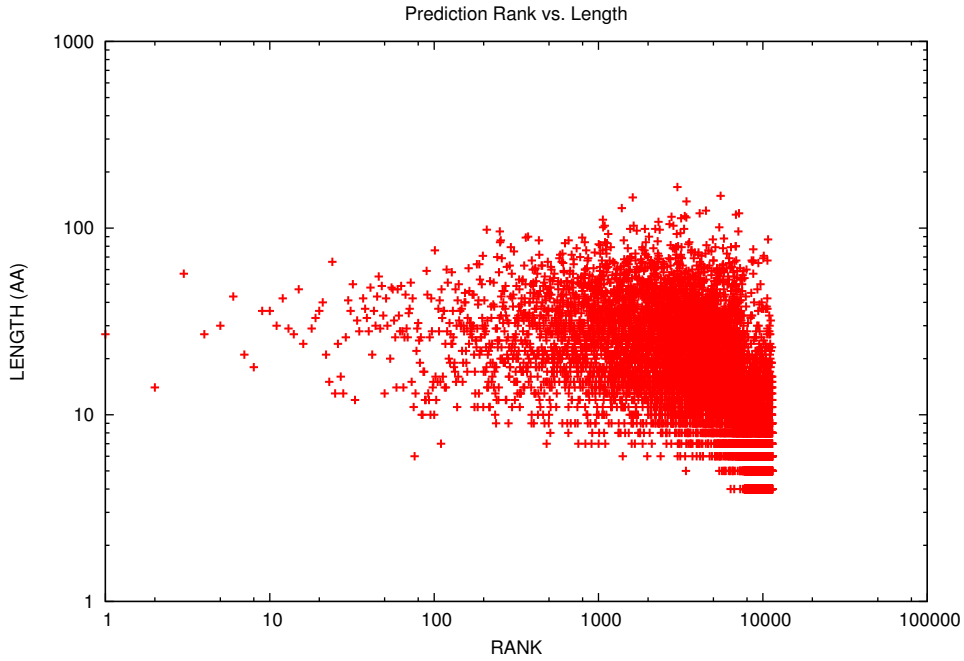

Figure 21: Plot of prediction's rank versus the gene length in number of amino acids. This ensemble consists of all un-annotated ORFs plus the validated small proteins set from *E. coli* K12. There was correlation between rank and ORF size, Spearman's $\rho = -0.71$ and Kendall's $\tau = -0.52$. This was especially true for very small ORFs, i.e. very small ORFs scored poorly.

6 Length bias

We investigated how prediction accuracy correlated with sequence length. ORF length is typically a good predictor of whether or not a given gene is real or an artifact of random chance. The longer the ORF, the less likely it is to have occurred randomly. For this reason we took great care to avoid any length bias. Figure 21 shows the rank of all test predictions (i.e., all un-annotated ORFs plus the validated small proteins set) against the length of each sequence. There is a strong correlation between length and prediction ranks, Spearman's $\rho = -0.71$ and Kendall's $\tau = -0.52$. Specifically, very small ORFs tend to rank poorly. However, if we focus on the top prediction ranks, the area we are most interested in, the correlation goes away. Within the top 1000 predictions, the Spearman's $\rho = 0.02$ and Kendall's $\tau = 0.02$ as well. Looking at the top 200 predictions the Spearman's $\rho = -0.05$ and Kendall's $\tau = -0.03$.

7 Comparison to Glimmer

The default Glimmer protocol correctly identified 16 of the experimentally validated small proteins within the top 200 predictions but was unable to perform better than our method overall. For example, we found that 5 of the top 10, 10 of the top 20, 11 of the top 30, 13 of the top 40, and 26 of the top 100 were from the set of experimentally validated small proteins. Glimmer found 2 of the top 10, 5 of the top 20, 8 of the top 30, 8 of the top 40, and 14 of the top 100 were from the set of experimentally validated small proteins. Our method found a total of 35 validated small proteins within the top 200 predictions compared to only 16 validated small proteins within the top 200 predictions for the default Glimmer protocol (Table 3)

ID	Length	TM pred	Glimmer	Our Rank	ID	Length	TM pred	Glimmer	Our Rank
ilvX	51	No		7983	ldrB	108	Yes		26
yjeV-ORF2.1	54	No		3061	ldrC	108	Yes		39
ibsB	57	No		4013	ldrD	108	Yes		57
ibsD	60	Yes		348	yohO	108	Yes		144
ibsC	60	Yes		372	yshB	111	Yes		49
ibsE	57	Yes		165	ybgT	114	Yes	21	273
ibsA	60	Yes		1721	rpmJ	117	No	34	943
ypfM	60	No		155	yqfG	126	Yes	114	14
yoeI	63	No		4303	ecnA	126	Yes	20	66
yobI	66	No		685	blr	126	Yes		1405
ymjD	66	No		993	ythA	126	Yes		na
ynbG	66	No		4442	ydfB	129	No	11	6
ypdK	72	Yes		34	ymiA	129	Yes	7	59
yoaJ	75	Yes		106	sgrT	132	No	24	91
yphI-shoB	81	Yes		4	yqgB	132	No	121	1161
yqeL	81	No		973	ydaG	135	No	14	560
yrbN	81	No		656	sra	138	No		1037
yohP	81	Yes		1	yqcG	141	No	2	378
ydgU	84	Yes		45	yceO	141	Yes		18
azuC	87	No		15	ylcG	141	No	67	5052
kdpF	90	Yes		12	ykgO	141	No		140
ynhF	90	Yes		169	rpmH	141	No	65	167
tisB	90	Yes		5	yobF	144	No		2940
yccB	93	Yes		200	mgrB	144	Yes	34	73
yncL	96	Yes		1600	ecnB	147	Yes	5	84
yneM	96	Yes		76	hokB	150	Yes		442
yoaK	99	Yes		50	ybhT	150	Yes		42
ykgR	102	Yes		63	hokC	153	Yes		153
yniD	108	Yes		10	hokA-yiaZ	153	Yes	51	96
ldrA	108	Yes		11	hokE-ybdY	153	Yes	73	271

Table 3: Comparison of performance between our method and default Glimmer protocol. Listed are the prediction ranks according to our method for each individual experimentally validated small protein in the side-by-side comparison with Glimmer. Also listed is the length of each protein, in bases, and whether or not it has a predicted trans-membrane domain (TM). The TM predictions are derived from the computational and experimental published results of Hemm et al. The default Glimmer protocol identified 16 of the previously annotated small proteins in *E. coli* K12, while our method placed 35 proteins from this set in the top 200 predictions. The validated small protein, ythA, was not evaluated by our method because it displayed 100% conservation among all species in our multiple alignment. Missing values in the Glimmer ranks represent cases that were either not in the prediction ensemble or were outside the top 200 predictions.

8 Identification of novel small proteins

Several small ORFs not previously annotated as protein encoding were among the highest scoring in the validation experiment. Supplementary table "novel_predictions.txt" lists the chromosome ID, plus strand starting coordinate, plus strand ending coordinate, amino acid length, unique ID, coding strand, prediction rank in validation experiment, and the score of the best match to a Shine-Dalgarno motif within the 15 bp immediately upstream of the ORF's start codon for each ORF with no more than 20% of its sequence overlapping a GenBank annotation in the validation experiment. This represents an ensemble of 11,185 sequences.

References

[1] J. Archie and K. Karplus. Applying undertaker cost functions to model quality assessment. *Proteins*, 75:550–555, May 2009.

[2] C. Bystroff, V. Thorsson, and D. Baker. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.*, 301:173–190, Aug 2000.

[3] A.G. de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41:271–287, Nov 2000.

[4] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, Dec 1983.

[5] R. Karchin, M. Cline, and K. Karplus. Evaluation of local structure alphabets based on residue burial. *Proteins*, 55:508–518, May 2004.

[6] R. Karchin, M. Cline, Y. Mandel-Gutfreund, and K. Karplus. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins*, 51:504–514, Jun 2003.