Protein classification into domains of life using Markov chain models

Francisca Zanoguera and Massimo de Francesco
Serono Pharmaceutical Research Institute
14, chemin des Aulx, CH-1228 Switzerland
Francisca.Zanoguera@serono.com

Abstract

It has recently been shown that oligopeptide composition allows clustering proteomes of different organisms into the main domains of life. In this paper, we go a step further by showing that, given a single protein*, it is possible to predict whether it has a bacterial or eukaryotic origin with* 85% *accuracy, and we obtain this result after ensuring that no important homologies exist between the sequences in the test set and the sequences in the training set. To do this, we model the sequence as a Markov chain. A bacterial and an eukaryote model are produced using the training sets. Each input sequence is then classified by calculating the log-odds ratio of the sequence probability for each model. By analyzing the models obtained we extract a set of most discriminant oligopeptides, many of which are part of known functional motifs.*

1. Introduction

Analysis of k-tuple compositional bias is a powerful tool for biological sequence classification. This approach characterizes a sequence or a class of sequences by looking at the relative frequencies of the "words" that compose it. The advantage with respect to other methods is that it does not require alignment of the sequences. On the other hand, sequences that may not seem to bear resemblance from an alignment point of view may be found to be similar in "word" composition. Although much of the work in this area has been carried out without using any explicit mathematical modeling, the introduction of Markov models adds a formal probabilistic framework to compositional bias analysis.

Compositional biases between coding and non coding DNA sequences have been extensively exploited in gene finding (see [5] for a review). Interestingly, biases within the same genome are consistently smaller than biases between different organisms [3, 6].

It has been shown that genome bias translates into a corresponding bias in the amino acid composition [9], implying that oligopeptide composition can also be used for sequence discrimination. This has been demonstrated by Stuart, Moffet and Baker [10] and later by Qi and Hao [8], who have constructed phylogenies using only oligopeptide frequency information that are in good agreement with the SSU rRNA tree of life. The latter find that the branching showing the division of life into the three main domains, bacteria, eukaryota and archaea appears very prominently. A similar result has been found by Pe'er et al. [7]. In these approaches, the proteome for a set of organisms is analyzed to create a vector with either the relative frequencies or the deviations from the expected frequency of each possible oligopeptide of length k. The vectors are then either clustered [8], or analyzed using Principal Component Analysis [7] or Singular Value Decomposition [10], to reach the aforementioned conclusions. However, these methods need a large set of proteins from the same organism in order to be able to determine its domain of life. In this paper, we show that, given a *single protein*, it is possible to predict whether it has a bacterial or eukaryotic origin with 85% accuracy.

In section 2 we introduce the approach used for the classification. Section 3 presents the main results. Finally, in section 4 we give a list of most discriminant oligopeptides.

2. Markov chain-based bayesian classifier

We model the sequence using a Markov chain of order $k-1$. This model makes the assumption that the existence of a certain amino acid at a given position in the sequence depends only on the previous $k-1$ amino acids (and thus there are no longer dependencies), i.e.

$$P(s_i|s_{i-1}s_{i-2}\dots s_0) = P(s_i|s_{i-1}s_{i-2}\dots s_{i-k})$$

where s_i represents the amino acid at position i in the sequence S.

Using this assumption and Bayes' rule, the *a posteriori* probability of the sequence being emitted by a certain model

Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)
0-7695-2194-0/04 $20.00 © 2004 IEEE

can be written as:

$$P^{model}(S) = \prod_{k \leq i < L} P^{model}(s_i | s_{i-1} s_{i-2} \dots s_{i-k})$$

where L is the sequence length.

Every class in the training set is used to estimate the conditional probabilities for that model, as:

$$\hat{P}^{model}(s_i | s_{i-1} \dots s_{i-k}) = \frac{\# s_i s_{i-1} \dots s_{i-k}}{\# s_{i-1} \dots s_{i-k}}$$

where $\hat{P}^{model}$ is the estimated probability for that model and $\# s_i s_{i-1} \dots s_{i-k}$ is the number of times that the k-tuple $s_i s_{i-1} \dots s_{i-k}$ appears in the training set.

A sequence S is assigned to the model (class) for which the *a posteriori* sequence probability is higher.

2.1. Markov chain order

Higher Markov model orders produce a more accurate model of the sequence, as long as the conditional probabilities can be reliably estimated from the training set. Thus, in practice, it is the training set size that determines the maximal Markov model order. In this work, we have used Markov order 3 (tetrapeptides). This corresponds to $20^4 = 160,000$ possible tetrapeptides.

2.2. Dataset

The dataset used in this analysis was the SwissProt database release 42 [4]. We randomly subdivided the sequences into 5 equally sized subsets, one serving as test set, the others forming the set of training sequences. In order to ensure independence between the training and test sets, we filtered the test set to remove any proteins bearing more than 35% identity over 90% of the sequence with any of the proteins of the training set using BLAST [1]. Proteins shorter than 50 amino acids were also removed, as we do not expect the classifier to perform well on very short sequences, which do not contain many different tuples for analysis. This minimum length was arbitrarily chosen.

To estimate the error rate of the classifier, we used a jackknife approach with 5 equally sized subsets.

3. Results

Although the classifier can be used with any number of classes, we begin by presenting classification results with only two classes: eukaryota and bacteria. Archaeal proteins were not considered initially because the smaller size of the training set did not allow to accurately estimate the Markov chain probabilities, which lowered performance. We present nevertheless the results obtained when including archaea in section 3.2.

3.1. Classification into two classes: eukaryota and bacteria

When only two classes are considered, we can use the log-likelihood ratio of the probabilities for each model as a classification score:

$$Score(S) = log \frac{P^{euk}(S)}{P^{bact}(S)}$$

We assign the protein to the eukaryotic class when $Score(S) > 0$, and to the bacterial class when $Score(S) < 0$.

The classifier correctly assigns the superkingdom of origin (eukaryota or bacteria) of an input sequence in 84.8% of the cases. Figure 1 displays the score distribution for the two classes, where the peaks are clearly separated. The scores have been multiplied by 1000 to make them more legible.

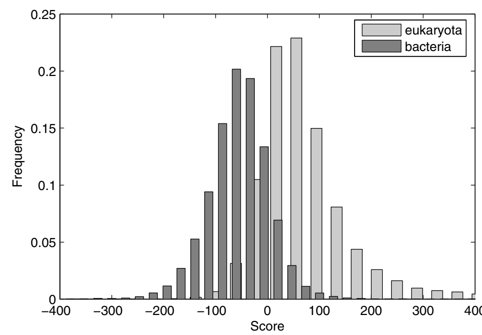

Figure 1. Histogram of log-likelihood ratio for eukaryotic and bacterial proteins.

3.2. Including archaea

The smaller number of archaeal proteins available makes the estimation of the peptide probabilities less accurate. Thus, lower performances are expected when including this group in the classifier. The results when including archaea are the following:

	Correct	Incorrect
Eukaryota	83.51%	16.49%
Bacteria	82.12%	17.88%
Archaea	66.63%	33.37%

We observe a small (expected) decrease in the classifier accuracy for eukaryota and bacteria. This is due to the inclusion of an extra group, which increases the probability of making a mistake. For archaea, the classification accuracy is much lower (66.6%), because the peptide probabilities can only be roughly estimated.

Peptide	$\lvert log(p^{euk}/p^{bact})\rvert$	PRINTS	Description
PYK(C)	4.14	PR00048	Zinc finger motif
FWQ(C)	3.98		
AWY(W)	3.89		
GCY(C)	3.88	PR00389	Phospholipase A2
PYQ(C)	3.82	PR00048	Zinc finger motif
FFI(C)	3.80	PR00153	Cyclophilin peptidyl-prolyl cis-trans isomerase
GPM(C)	3.70	PR00114	Serine/threonine phosphatase family
DRC(C)	3.70	PR00389	Phospholipase A2
MIQ(C)	3.69	PR00389	Phospholipase A2
WWN(N)	3.59		
PFK(C)	3.49	PR00048	Zinc finger motif
WVP(C)	3.49	PR00152	RuBisCO small subunit
QQM(W)	3.48		
HTY(C)	3.47	PR01109	C-C chemokine receptor type 4
WDD(M)	3.45	PR00190	Actin
HQW(Y)	3.42	PR01166	Cytochrome c oxidase subunit
WHH(T)	3.42	PR00190	Actin
MVK(C)	3.40		
ACW(C)	3.38	PR00284	Toxin
ECK(C)	3.36		
SCK(C)	3.34		
PHK(C)	3.33	PR00048	Zinc finger motif
VYP(W)	3.27	PR00814	Beta haemoglobin
MRM(C)	3.17		
PYE(C)	3.13	PR00048	Zinc finger motif
HFF(C)	3.13	PR00245	Olfactory receptor
QWI(W)	3.12		
WGE(C)	3.09	PR00269	Pleiotrophin/midkine family
CWF(P)	3.09		
WWN(W)	3.08		
RGP(P)	3.08	PR01797	Saposin
WFC(N)	3.05	PR00472	Casein kinase II regulatory subunit family
NYG(W)	3.03		
PYM(C)	3.02	PR00048	Zinc finger motif
HRK(C)	3.01		
NAC(W)	3.00	PR00284	Toxin

Table 1. Most discriminant tetrapeptides.

4. Most discriminant tetrapeptides

Table 1 shows the most discriminant tetrapeptides between eukaryota and bacteria, where the notation ABC(D) refers to the probability of amino acid D happening after ABC. Most of them are part of PRINTS [2] signatures, as indicated in the table. It is noticeable that, of the 32 tuples, 22 end in a cysteine, 6 end in tryptophan and only 4 have other endings. Tryptophan and cysteine are the least abundant amino acids, and their over-representation in these motifs seems to underline the importance of their presence.

5. Conclusions

We have shown that, based on its oligopeptide composition only, a protein can be classified into its superkingdom of origin with high accuracy. When only the eukaryotic and bacterial domains are considered, we achieve 85% accuracy. Lower accuracy is obtained when including archaea, due to the smaller size of the training set. However, a similar performance is expected when more proteins will be available from this superkingdom.

This result is based exclusively on oligopeptide composition and is not a consequence of homologies between the training and test set, because homologous proteins were eliminated from the test set.

By looking at the oligopeptides that have higher impact in the classification, we have obtained a list of most discriminant oligopeptides, and have found that most of them are included in known functional motifs. Novel motifs could be found by examining the unmatching oligopeptides.

References

[1] Altschul et al. Basic Local Alignment Search Tool. *J. of Mol. Biol.*, 215:403–410, 1990.

[2] T. Attwood, M. Beck, and D. Parry-Smith. PRINTS–a database of protein motif fingerprints. *Nucleic Acids Res.*, 22(3590-3596), 1994.

[3] B. E. Blaisdell. A measure of the similarity of sets of sequences not requiring sequence alignment. *P. Natl. Acad. Sci. USA*, 83:5155–5159, July 1986.

[4] B. Boeckmann et al. The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, 31:365–370, 2003.

[5] J.-M. Claverie. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.*, 6(10):1735–1744, 1997.

[6] S. Karlin, J. Mrázek, and A. M. Campbell. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.*, 179(12):3899–3913, June 1997.

[7] Pe'er et al. Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. *Proteins*, 54:20–40, 2004.

[8] J. Qi and B. Hao. Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. *J. Mol. Evol.*, 58(1):1–11, 2004.

[9] G. A. Singer and D. A. Hickey. Nucleotide bias causes a genomewide bias in the amino acid composition. *Mol. Biol. Evol.*, 17(11):1581–1588, 2000.

[10] G. W. Stuart, K. Moffet, and J. J. Leader. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol. Biol. Evol.*, 19(4):554–562, 2002.