Discriminating Microbial Species Using Protein Sequence Properties and Machine Learning

Ali Al-Shahib[1,2], David Gilbert[2], and Rainer Breitling[3]

[1] Biomedical Informatics Signals and Systems Research Laboratory, Department of Electronic, Electrical and Computer Engineering, The University of Birmingham, Birmingham, UK
a.alshahib@bham.ac.uk
[2] Bioinformatics Research Centre, Department of Computing Science, University of Glasgow, Glasgow, UK
[3] Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Haren, The Netherlands

Abstract. Much work has been done to identify species-specific proteins in sequenced genomes and hence to determine their function. We assumed that such proteins have specific physico-chemical properties that will discriminate them from proteins in other species. In this paper, we examine the validity of this assumption by comparing proteins and their properties from different bacterial species using Support Vector Machines (SVM). We show that by training on selected protein sequence properties, SVMs can successfully discriminate between proteins of different species. This finding takes us a step closer to inferring the functional characteristics of these proteins.

1 Introduction

Species divergence is mainly caused by variation in gene and protein sequences but also by differences in the set of genes that is present in a particular species. Proteins that are specific for a particular species may be responsible for its adapted phenotype, e.g. its ability to act as a pathogen or its resistance to a certain drug. Identifying species-specific proteins is thus a relevant aim, and here we make a small contribution towards its achievement.

In this paper, we have compared the proteins of seven different bacterial species by extracting numerous protein sequence properties using state-of-the-art Support Vector Machines. To our surprise, we find that proteins of different species are significantly dissimilar and can be distinguished based on sequence properties selected prior to classification. This discrimination does not rely on any homology criteria but is based only on the biophysical characteristics encoded in the sequence. We have also constructed a phylogenetic tree based on the results of the comparisons, and compared it to the well-documented 16S rRNA dendrogram of the same bacteria. Interestingly, there is no detectable similarity between the two dendrograms.

H. Yin et al. (Eds.): IDEAL 2007, LNCS 4881, pp. 890–897, 2007.

2 Methodology

Seven Sexually Transmitted Disease-causing bacteria were used for this study. The protein sequences were obtained from the Los Alamos National Laboratory [4]. Table 1 shows the total number of proteins of each species in our bacterial database.

Table 1. Total number of proteins of each species in our database

Name	**Total number of proteins**
Chlamydia trachomatis (CT)	902
Haemophilus ducreyi (HD)	1830
Mycoplasma genitalium (MG)	485
Neisseria gonorrhoeae (NG)	2188
Streptococcus agalactiae (SAG)	2177
Treponema pallidum (TP)	1051
Ureaplasma urealyticum (UU)	614

All the functionally known proteins of every genome and their sequences were collected and 2579 sequence properties for every protein were extracted. These include some global properties (e.g. isoelectric point and molecular weight), the frequency and total number of each amino acid, the frequency and total number of certain sets of amino acids (e.g. hydrophobic, charged, polar), the number and size of continuous stretches of each amino acid or amino acid set, secondary structure predictions obtained using the Prof algorithm [5], the position of putative transmembrane helices predicted using TMHMM [6], and that of disordered regions obtained using DisEMBL [7]. A full list of the properties is available at `http://www.dcs.gla.ac.uk/∼alshahib/features.pdf`.

Normalization of the features was performed for all features of all proteins at once, instead of normalization for every individual genome. This is important as performing normalization of the features for every individual species will cause slight differences in the scales of the features when combined with the normalized feature protein values of any other species. These slight differences will cause unjustified discrimination of the species. Thus it was appropriate to rescale each feature by its mean and variance [15] with respect to all proteins from all species.

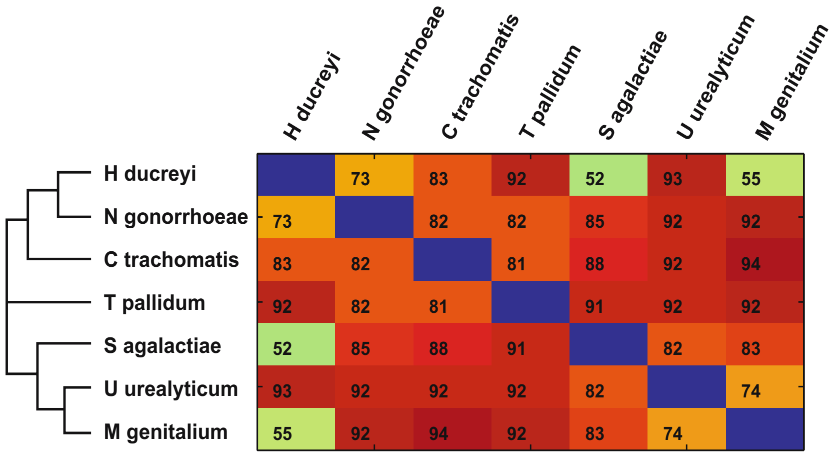
Proteins of every species were combined with every other species (pairwise). A total of 21 pairwise species comparisons were performed. For every comparison, undersampling of the negative examples to equal the positive examples [8] was performed. Five training and test sets were then generated for every comparison and each was homology-corrected as described in [10]. This is vital because once the training and test sets are divided for training and testing, one must make sure that similarity between proteins on both sets is minimal. In other words, predictions must not be made based on homology of proteins in the training and test sets but rather on non-homologous proteins. We have thus implemented a recursive BLAST strategy to assign proteins that show significant sequence similarity to each other to the same set (either test or training). For details see [10].

For every training set of every pairwise species comparison, feature selection was performed using our FrankSum method [9]. Finally, Support Vector Machine classification was performed for every pairwise species comparison and the AUCs were recorded. A polynomial kernel of order 3 with a C value of 1 was used for SVM classification. The WEKA machine learning package was used for this task [11].

3 Results

For every pairwise species comparison, five AUCs (obtained on 5 test set–training set combinations) were recorded. Figure 1 shows the performances achieved when comparing proteins from different source species.

All species pairs can be readily discriminated, except *H. ducreyi* and *S. agalactiae*, and *H. ducreyi* and *M. genitalium*. The median discrimination performance is as high as 91% for the proteins of *Treponema pallidum* and *Ureaplasma urealyticum* compared against all other species. The worst performance, for *Haemophilus ducreyi*, is still a surprising 83%. For any randomly selected pair of proteins from two species, a correct assignment to its species of origin will be possible in 85% of cases. This is achieved based solely on the sequence properties described above. It is all the more unexpected, as the bacteria that we analyse



	H ducreyi	N gonorrhoeae	C trachomatis	T pallidum	S agalactiae	U urealyticum	M genitalium
H ducreyi		73	83	92	52	93	55
N gonorrhoeae	73		82	82	85	92	92
C trachomatis	83	82		81	88	92	94
T pallidum	92	82	81		91	92	92
S agalactiae	52	85	88	91		82	83
U urealyticum	93	92	92	92	82		74
M genitalium	55	92	94	92	83	74	

Fig. 1. Pairwise species discrimination performances. The AUC for classifiers trained to distinguish between proteins from each species pair (median of five replicates). The unrooted tree to the left shows the phylogenetic relationships of the seven bacterial species, based on 16S rRNA analysis.

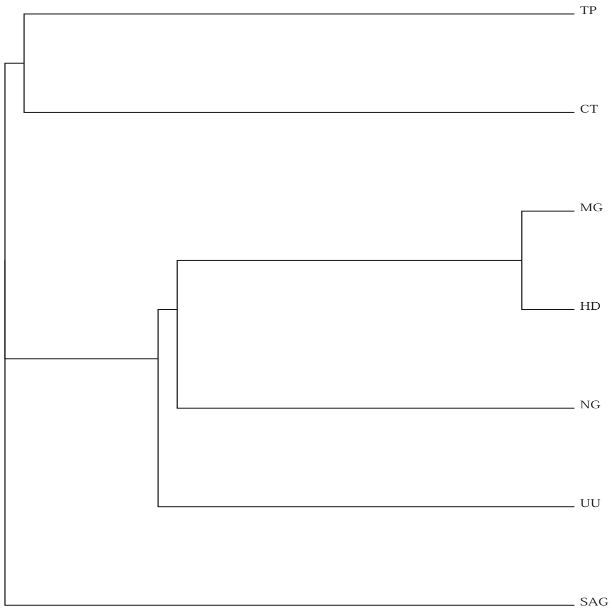


Fig. 2. Dendrogram resulting from pairwise species discrimination using SVMs. The median of the five AUCs of each pairwise species comparison was used as a distance measure between the species. See Table 1 for an explanation of the abbreviations.

are biologically very similar, they all occur in the same uniform environment, inside the urogenital tract of the human body. They naturally differ in their mechanisms of pathogenicity, but nonetheless their general biology should be the same and should make use of very similar molecular structures. The fact that we can identify general species-specific "sequence signatures" is therefore particularly striking.

In addition to the species–species discrimination, it was interesting to explore whether using sequence features to discriminate between bacterial species by machine learning will provide an accurate phylogenetic relationship between the species. The tree could then be compared to the 16S rRNA phylogenetic tree of the STD bacteria.

The median of the five AUCs of each pairwise species comparisons was used as a distance measure between the species in the phylogenetic tree. The OC [12] hierarchical cluster program was used to construct the dendrogram. For the 16S rRNA tree (Table 3), 16S rRNA sequences were obtained from GenBank for representative members of each bacterial genus in the dataset, as well as for three diverse Archaea for use as an outgroup. Sequences were aligned by ClustalW [16], positions containing gaps were removed, and the remaining alignment subjected to phylogenetic analysis using Maximum Likelihood (DNAml and Fast DNAml), Maximum Parsimony (DNAPars) and Neighbor Joining (DNADist+Neighbor),

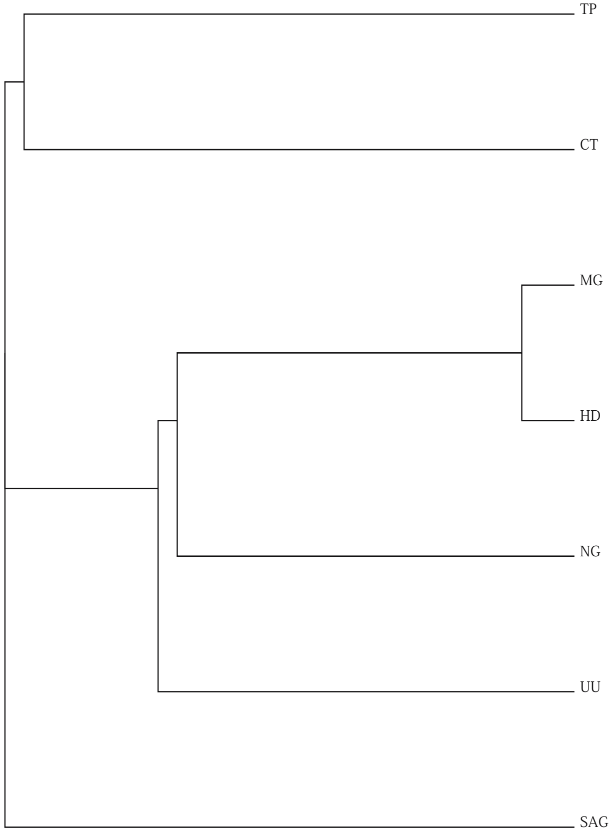


Fig. 3. 16S rRNA Dendrogram resulting from 16S rRNA sequences for the seven bacteria in our database as well as three Archaea outgroups. New species introduced as other members of the bacterial genus are NM = *Neisseria meningitidis* (as an alternative to NG) TD = *Treponema denticola* (as an alternative to TP) and US = *Ureaplasma parvum* serovar (as an alternative to UU). The Archaea include AF = *Archaeoglobus fulgidus*, HS =*Halobacterium salinarum* and PH = *Pyrococcus horikoshii.* see Table 1 for explanations of the other abbreviations.

using BioEdit. The resulting trees were identical except for a single change in the NJ tree. A majority-vote consensus tree was generated and rooted using the Archaea as outgroup.

From Figure 2 we can see that the SVM-based tree shows little similarity to the phylogenetic tree. For instance, the closely related species pairs *Haemophilus ducreyi*/*Neisseria gonorrhoeae* and *Mycoplasma genitalium*/*Ureaplasma urealyticum* are not identified correctly. Only *Chlamydia trachomatis* and *Treponema pallidum* are detected as outliers. However, we have generated random trees and found that the tree in Figure 2 is indeed closer to the 16S RNA tree than the majority of random trees, although this similarity is not obvious based on visual inspection. This indicates that there is some useful phylogenetic signal contained in the SVM results. Perhaps more discriminatory sequence properties could be used for a more accurate construction of the dendogram.

4 Discussion

Comparing protein sequence properties of every species might outline the natural difference of the species and how evolution has played an essential role in their divergence. In this paper, we have used a wide range of sequence properties and have used Support Vector Machines in an attempt to discriminate between proteins of different species. Interestingly, the discrimination performances was as high as 85% AUC (median of all species–species discriminations performed). This is of course of great biological interest. By extracting useful information from the sequence, we hope to shed more light on this variation. At the DNA level, one of the discriminating species-specific features is the varying GC content. Guanine-cytosine (GC) content has been shown to be a biologically important attribute in prokaryotes [13]. It is known to be fairly balanced and tightly controlled across the genome, thus providing high specificity for genome identification. The Percentage GC content in bacteria can range from 25% to 75%. According to Bentley and Parkhill [13], the GC content of prokaryotes depends on the genome size. The correlation between genome size and GC content shows larger genomes tend to have higher GC content than smaller genomes which are AT-rich.

At the amino acid level, we expect preferred amino acids to be different as a result of varying GC contents at the DNA level. This is supported by earlier reports [17] which showed that the amino acid composition of 59 bacterial species was greatly influenced by varying genomic G+C content.

To further elaborate on this, we have recorded the highest 10 selected features when comparing the (GC-rich) β-proteobacterium *Neisseria gonorrhoeae* with the three (GC-poor) Firmicutes species (*Mycoplasma*, *Streptococcus*, and *Ureaplasma*). This is shown in Table 2.

The three amino acids alanine, arginine and proline have high GC content in most of their codons. Our features selected for classification of the proteins agree with the GC-based prediction, in that four features with enriched GC bases were selected amongst the top 10 discriminatory features. This is statistically highly significant and indicates the relevance of varying GC content for our successful species discrimination.

Further analysis of Table 2 shows the frequent occurrence of the amino acids lysine (Lys) and arginine (Arg) as relevant features. These amino acids have been reported earlier to be significantly overrepresented in proteins of particular functional categories (transcription, translation), indicating the importance of our selected features for protein function [18].

Finally, our method has demonstrated the discriminatory power of Support Vector Machine classification as it can use sequence features to discriminate proteins from different species with high reliability and accuracy.

Table 2. Top 10 selected features using the *Neisseria gonorrhoeae* genome. The top 10 selected features for comparing the β-proteobacteria *Neisseria gonorrhoeae* with the three Firmicutes species are shown. The amino acids that contain GC-rich codons are highlighted in bold. Abbreviations: AAs = amino acids, Qt = quarter, div = divided and no. = number. See text for discussion.

MG vs. NG	**NG vs. SAG**	**NG vs. UU**
The no. of amide AAs div by the length of the protein	The no. of mean blocks of charged AAs	The no. of tiny AAs div by the length of the protein
The no. of ile AAs div by the length of the protein	The no. of polar AAs div by the length of the protein	The no. of **pro** AAs from the 0% to 50% region of the protein div by the length of the protein
The no. of lys AAs	The no. of ile AAs div by the length of the protein	The no. of **arg** AAs in the protein
The no. of lys AAs div by the length of the protein	The no. of lys AAs in the 1st Qt of the protein div by the length of the protein	The no. of ile blocks in the 4th Qt of the protein div by the length of the protein
The no. of **arg** AAs in the 1st Qt div by the length of the protein	The no. of **ala** AAs div by the length of the protein	The no. of lys AAs in the longest lys block of the protein
The no. of mean blocks of **ala** AAs in the 3rd Qt of the protein	The no. of **pro** AAs div by the length of the protein	The no. of +ve charged AAs in the 1st Qt of the protein div by the length of the protein
The no. of **pro** blocks in the 25% to 75% region of the protein div by the length of the protein	The no. of gly AAs div by the length of the protein	The no. of **pro** AAs from the 50% to 100% region of the protein div by the length of the protein
The no. of mean blocks of **ala** AAs in the 2nd Qt of the protein	The no. of **arg** AAs in the 4th Qt of the protein div by the length of the protein	The no. of amide blocks in the 4th Qt of the protein div by the length of the protein
The no. of mean blocks of lys AAs in the 1st Qt of the protein	The no. of **arg** AAs in the protein	The no. of **pro** mean blocks in the protein
The no. of mean blocks of lys AAs in the 2nd Qt of the protein	The no. of cys AAs div by the length of the protein	The no. of ile mean blocks in the protein

We hope that this work can be extended by exploring further sequence properties as well as more diverse organisms, to elucidate the underlying biophysical and evolutionary mechanisms.

References

1. Zuckerkandl, E., Pauling, L.: Evolutionary divergence and convergence in proteins. In: Evolving Genes and Proteins, pp. 97–166. Academic Press, New York (1965)
2. Robichaux, R.H., Purugganan, M.D.: Accelerated regulatory gene evolution in an adaptive radiation. Proc. Natl. Acad. Sci. USA 98, 10208–10213 (2001)
3. Gollery, M., Harper, J., Cushman, J., Mittler, T., Girke, T., Zhu, J.-K., Bailey-Serres, J., Mittler, R.: What makes species unique? The contribution of proteins with obscure features. Genome Biology 7, R57 (2006)

4. STDGEN, Los Alamos National Laboratory Bioscience Division STD Sequence Databases, http://www.stdgen.lanl.gov
5. Ouali, M., King, R.D.: Cascaded multiple classifiers for secondary structure prediction. Prot. Sci. 9, 1162–1176 (2000)
6. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.L.: Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. J. Mol. Biol. 305, 567–580 (2001)
7. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., Russell, R.B.: Protein disorder prediction: implications for structural proteomics. Structure 11, 1453–1459 (2003)
8. Al-Shahib, A., Breitling, R., Gilbert, D.: Feature Selection and the Class Imbalance Problem in Predicting Protein Function from Sequence. Applied Bioinformatics 4, 195–203 (2005)
9. Al-Shahib, A., Breitling, R., Gilbert, D.: FrankSum: new feature selection method for protein function prediction. Int. J. Neural Syst. 15, 259–275 (2005)
10. Al-Shahib, A., Breitling, R., Gilbert, D.: Predicting protein function by machine learning on amino acid sequences – a critical evaluation. BMC Genomics 8, 78 (2007)
11. WEKA machine learning package, http://www.cs.waikato.ac.nz/ml/weka
12. Barton, G.: A cluster analysis program (1993), http://www.compbio.dundee.ac.uk/Software/OC/oc.html
13. Bentley, S.D., Parkhill, J.: Comparative Genomic Structure of Prokaryotes. Annual Review of Genetics 38, 771–791 (2004)
14. Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., Ishikawa, H.: Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature 407, 81–86 (2000)
15. Bishop, M.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1993)
16. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D.: Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31, 3497–3500 (2003)
17. Lobry, J.R.: Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene. 205, 309–316 (1997)
18. Bharanidharan, D., Gautham, N.: Amino acid variation in cellular processes in 108 bacterial proteomes. Arch. Microbiol. 184, 168–174 (2005)