Biological Domain Identification Based in Codon Usage by Means of Rule and Tree Induction

Antonio Neme[1] and Pedro Miramontes[2]

[1] IIMAS, UNAM, México
neme@uxmcc2.iimas.unam.mx
[2] Facultad de Ciencias, UNAM, México

Abstract. There are three domains in living nature: archaea, bacteria and eukarya. It has been shown, trough a number of multivariate tools, that codon usage, a 64 dimensional vector that stablishes how often a given organism makes use of each codon, is related to domain. Another method is proposed here based in rule and tree induction from codon usage of several organisms. It is shown that domain can be identified trough codon usage and a simple set of rules. Two methods were applied, $CN2$ and $C4.5$. Obtained rules describe data better than other methods, in the sense that are topological interpretable and have phenomenological meaning.

1 Introduction

Codon usage is the preference shown by organisms to use a certain synonymous codon to code amino acids. 18 out of 20 amino acids are coded by more than one synonymous codon and the fact some organisms (or genes) prefer a given codon to code for a certain amino acid is known as *codon bias* [5].

Organisms may be tought of as points in $64-$dimensional space, accordingly to codon usage data. The distribution shown by them, thus, codon bias, has been a subject of intense research in molecular biology. Codon bias has not been explained. Several theories have been proposed but there is not a general explanation for it [11, 8]. Each organism may be represented by its codon usage vector, that contains the frequency per ten thousand of each codon.

Grantham used principal componet analysis in [3] to show that codon usage is related to biological domain. Using a self-organizing map, more evidence has been given to show that, in general, codon usage is related to biological domain, with a few counterexamples of special organisms (like *Th. Maritima* and *U. Urealiticum*) that does not seem to follow the expected pattern [7].

A set of understandable rules that classifies properly a group of organisms may be a better tool for molecular biologists to explain codon bias on domain basis, because important variables (frequency of each codon) and its relationships are explicitly settled. In this work, we obtain a set of rules that properly identifies the domain an organism belongs to.

V. Danos and V. Schachter (Eds.): CMSB 2004, LNBI 3082, pp. 221–224, 2005.

2 Methods

Rule and tree inference are part of an artificial intelligence field named *machine learning.* They have been extensively applied in data analysis because they are more transparent and easier to interpret than other methods, as for example, trainned neural networks or a regression models [2]. The goal is to find a function (here in form of rules) f that properly classify a set of examples X. For each individual sample vector X_i, it is associated a label or class (domain in this work), c_i. Thus, $f(X_i) = c_i$ means that applying the set of rules f to X_i the proper class will be identified.

Rule induction may be seen as a search problem: it finds a set of rules that are coherent (no classification errors) and complete (all organisms are classified). There are several algorithms for rule induction [1, 4], but the one applied here is the so called $CN2$. For a detailed explanation of this algorithm, see [2].

The tree induction method applied in this work is $C4.5$, propossed by Quinlan [9]. On it, a set of decisions is found so that each partition maximizes a *gain* criteria, based on information content. At every step, the variable that maximizes information (the number of objects correctly identified) is chosen.

3 Results

Rules obtained by applying CN2 to codon usage vector, obtained from the *Kazusa data bank* [6] of 159 organisms (most of them completely sequenced) are shown in table 1. There are 28 *archaea*, 68 bacteria and 63 eukarya. Codons

Table 1. Rules obtaided by CN2. Numbers bewteen squared brackets identify the number of organisms that satisfies conditions in the rule and belongs to domain *archaea* (first), *bacteria* (second) and *eukarya* (third)

IF UUU<0.42 and CUU<0.25 and CGU<0.02 THEN domain=archaea [13 0 0]
IF UAC>0.18 and CUC>0.35 THEN domain = archaea [7 0 0]
IF UUU<0.27 and CUG<0.29 and AGG>0.16 THEN domain=archaea [12 0 0]

IF UUU>0.18 and CAA>0.28 THEN domain = bacteria [0 22 0]
IF UAC<0.21 and GCC>0.50 THEN domain = bacteria [0 18 0]
IF CCA>0.12 and AAC<0.14 THEN domain = bacteria [0 10 0]
IF UUU>0.11 and CCA<0.09 and GGG<0.15 THEN domain = bacteria [0 13 0]
IF UGU<0.07 and CGU>0.13 THEN domains = bacteria [0 22 0]
IF 0.11<UUC<0.13 THEN domain = bacteria [0 6 0]

IF UCC>0.04 and UGC>0.1 THEN domain = eukarya [0 0 36]
IF UCA>0.15 and CCA>0.19 THEN domain = eukarya [0 0 13]
IF UUC>0.10 and UCU>0.19 and CCA> 0.10 THEN domain = eukarya [0 0 12]
IF AAC>0.77 THEN domain = eukarya [0 0 2]
IF UCC>0.2 THEN domain = eukarya [0 0 6]
IF UUG>0.19 and CGA>0.09 THEN domain = eukarya [0 0 3]

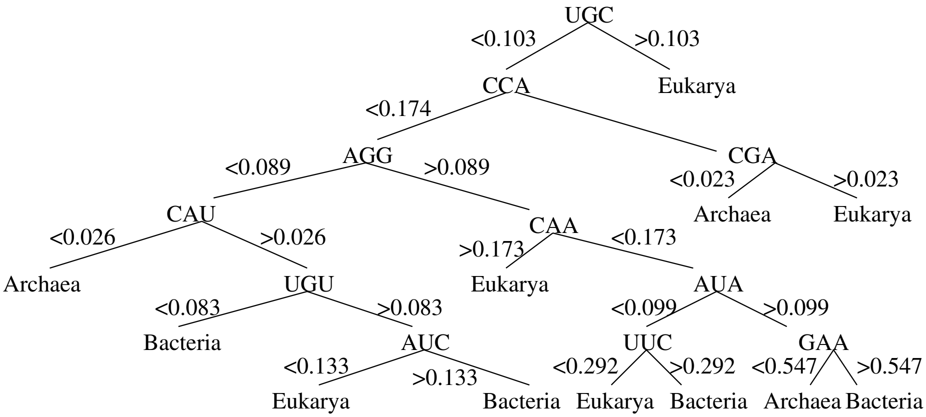

Fig. 1. Tree obtained by C4.5

are represented by its nucleotide sequence, such as AGG, meaning Adenine followed by Guanine followed by Guanine.

There are 18 rules and no classification errors were found. 22 codons, coding for 12 amino acids, were required to classify organisms by CN2 and, as is common in that algorithm [2], some examples are included in more than one rule.

The tree obtained by $C4.5$ for the same organisms is shown in Figure 1. In this tree, only 11 codons, coding for 9 amino acids, were required. There are, however, six misclassified organisms (3.8% error): three archaeas: *Archaeoglobus profundus*, identified as an eukarya, and *Methanococcus voltae* and *Methanosarcina mazei*, both identified as bacterias. Two eukaryas were misclassified: *Ostrinia nubilalis* and *Fusarium sporotrichioides*, both identified as bacteria. The only misclassified bacteria was *Buchnera aphidicola*, identified as eukarya. The rules obtained by CN2 and the tree obtained by $C4.5$ are more evidence to show that codon bias is related to biological domain. This was already known [3, 10], but what we do here is to give basis of explanation of that fact, because the rules and tree are interpretable information.

4 Conclusions

Analysis of biological data in a structured way is easier than doing so for data in form of tables or even in bidimensional maps, as those obtained by multivariate analysis. Here, we applied two formal methods to obtain structure in data for the problem of codon bias. More evidence that codon bias is affected by domains is given, but the difference is that we show a set of rules and an identification tree that may be intrepreted by specialists to explain it with more basis and with readable information. The applied methods were $CN2$ and $C4.5$.

For biologists, it may be of interest to find a pattern in codon bias dictated by domain. It is easier to look for that pattern if information is expressed in form of rules instead of looking at a graph, mainly because the variables (codons) appearing in the relations (the rules) could be interpreted on the light of the studied phenomena (codon bias).

The fact that not all codons are important for domain identification reduce the space of possible explanations. Organisms in codon usage space are not randomly distributed, but biased by biological domain. An explanation based only in those codons that differentiate among domains could be given, by analysing the frequency of use of those codons as well as relationships among them, such that reflects the evolutionary history of life, from the perspective of codon usage.

References

1. Clark, P. and Boswell, R. Rule induction with CN2: some recent improvements. Proceedings of the fifth European Working Session on Learning. Springer. (1992).
2. Flach, P. and Lavrac, N. Rule Induction, in *Intelligent Data Analysis*. Springer. 2003.
3. Grantham R, Gautier C, Gouy M, Mercier R, Pave A.. Codon catalog usage and the genome hypothesis. Nucl. Ac. Res. **1** (1980) 43-74.
4. Lavrâc, P. Flach, B. Adapting classification rule learning to subgroup discovery. Proceedings of the IEEE International Conference on Data Mining. (2002).
5. Lewin, B. Genes VII. Oxford University Press. 2000.
6. Nakamura, Y., Gojobori, T. and Ikemura, T. Codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nucl. Ac. Res. **28** (2000) 292.
7. Neme, A. Codon usage and self-organizing maps. Proceedings of the Mexican Mathematical Society Meeting. 2003.
8. Powell J., Sezzi E., Moriyama E., Gleason J., and Caccone A. Analysis of a shift in codon usage in Drosophila. J. Mol. Ev. **57** (2003) Suppl. 1. 214-225.
9. Quinlan, R. C4.5 : programs for machine learning. Morgan Kaufmann. 1993.
10. Rowe G., Szabo V. and Trainor, L. Cluster analysis of genes in codon space. J. Mol. Ev. **2** (1984) 167-174.
11. Vinogradov, A. DNA helix: the importance of being *GC* rich. Nucl. Ac. Res. **7** (2003) 1838-1845.