Comparative analysis of prokaryotic and eukaryotic transcription factors using machine-learning techniques

Nilkanta Chowdhury & Angshuman Bagchi*

1Department of Biochemistry and Biophysics, University of Kalyani, Kalyani, Nadia 741235, India. Angshuman Bagchi – E-mail: angshu@klyuniv.ac.in, angshuman_bagchi@yahoo.com; Telephone: +919051948843; Fax: +913325828282; *Corresponding author

Received June 24, 2018; Revised June 28, 2018; Accepted June 28, 2018; Published June 30, 2018

doi:10.6026/97320630014315

Abstract: The DNA-protein interactions play vital roles in the central dogma of molecular biology. Proper interactions between DNA and protein would lead to the onset of various biological phenomena like transcription, translation, and replication. However, the mechanisms of these well-known processes vary between prokaryotic and eukaryotic organisms. The exact molecular mechanisms of these processes are unknown. Therefore, it is of interest to report the comparative estimate of the different properties of the DNA binding proteins from prokaryotic and eukaryotic organisms. We analyzed the different sequence-based features such as the frequency of amino acids and amino acid groups in the proteins of prokaryotes and eukaryotes by statistical measures. The general pattern of differences between the various DNA binding proteins for the development of a prediction system to discriminate between these proteins between prokaryotes and eukaryotes is documented.

Keywords: Prokaryotic and Eukaryotic Organisms; DNA binding proteins; Transcription factors; Distribution of amino acid residues.

Background: DNA protein interactions as in DNA transcription are at the heart of the central dogma of molecular biology. The transcription is the process of transfer of genetic information from DNA molecules. The process is regulated by a set of proteins. These proteins are referred to as the transcription factors (TFs) [1]. The mechanism of the process is a very complex one and is mainly mediated by a complex interplay between the TFs with DNA. However, the mechanism of DNA transcription is different in prokaryotic and eukaryotic organisms [2, 3].

However, the molecular details of the transcription processes in the pro- and eukaryotic organisms are still at its infancy. In this work, we tried to analyze the different aspects of the transcription factors from pro- and eukaryotic organisms. For the comparison purposes, we used the amino acid sequences of the DNA binding proteins (DBPs) and transcription factors (TFs) from UniProt [4].

We compared the TFs using their sequence information only as sequence is more abundant than structure [5]. The main motivation of carrying out the work is to discriminate between the different classes of microorganisms. We, for the first time, put forward some plausible discriminatory features between the TFs from the different branches of organisms. Interestingly, the TFs from the pro- and eukaryotic organisms can be distinctly identified using the amino acid frequency analyzes in the TFs.

We also analyzed the statistical efficacies of the features used in the study to discriminate between the different classes of microorganisms using machine-learning techniques. The ideas regarding these features may further be utilized to come up with a prediction system to discriminate between the different branches of organisms.

Methodology: Data collection: We downloaded the sequences of DNA binding proteins (DBPs) from UniProt [4]. We collected the amino acid sequences of the DNA binding proteins from 1012 prokaryotic organisms and 1425 eukaryotes. We divided our dataset into two groups, the largest group containing the whole DBP data, and a small subgroup...
containing the transcription factor (TF) sequences, which were also present in the DNA binding protein dataset. The data collection process was carried out using an in-house tool written in Python (Figure 1).

**Redundancy check to the dataset:**
The raw dataset may be biased because of having multiple copies of a single sequence. We, therefore, performed a redundancy check, by means of distance matrix calculation. The distance matrix was generated by Hamming distance algorithm [6, 7]. After this redundancy check, we were able to eliminate the redundancy in the dataset and prepared a clean dataset. The clean dataset contained 270 DBP sequences from prokaryotes and 347 DBP sequences from eukaryotes; among them, there were 92 sequences of TF from prokaryotes and 182 sequences of TF from eukaryotes. So the DBP dataset contained 270 prokaryotic and 347 eukaryotic sequences. As the eukaryotic DBP sequences were present in higher number than the prokaryotic DBP sequences, we had split the eukaryotic DBP sequences into two sets. Eukaryotic DBP set 1 contained sequences starting from 1 to 270 and eukaryotic DBP and set 2 contained sequences starting from 78 to 347 so that there were equal numbers of amino acid sequences in the datasets. For the same reason, the eukaryotic TF dataset was split into two sets. TF set 1 contained sequences starting from 1 to 92 and TF set 2 contained sequences starting from 91 to 182. Thus all the datasets were balanced. The distribution of the dataset is shown in Table 1.

The list of UniProt IDs used in these datasets was present in Table S1 (see Supplementary data).

**Frequency Calculation:**
After the preparation of these clean datasets, we performed amino acids and amino acids group frequency calculations. We categorized the amino acid groups into Hydrophobic (HB), Hydrophilic (HI), Charged (CR), Basic (BS) and Acidic (AC) [8]. This frequency calculation was done to normalize the dataset. The entire frequency calculation was done using an in-house python script. We had calculated the frequency of amino acids and amino acid groups separately for the two datasets DBP and TF, and separately for eukaryotic set 1 and eukaryotic set 2.

**Machine learning using WEKA:**
We used the overall amino acid frequencies and amino acids group frequencies of the prokaryotic and eukaryotic organisms as features to distinguish between prokaryotic and eukaryotic DNA binding proteins using the tool WEKA [9]. WEKA is a tool, containing a collection of machine learning algorithms, is commonly used in data mining problems in bioinformatics. We have used the Support vector machine (SVM) algorithm and the SMO classifier [10] with 10 fold cross-validation. The 10 fold cross validation is a kind of default test option of WEKA. It randomly splits the dataset into training and testing datasets and runs the test. It does this operation 10 times with random splitting of the input data into training and testing datasets. We prepared the input dataset for WEKA using data distribution as described in Table 1.

**Results:**
**Amino acids and amino acid group frequency**
A distinguishable difference was found in the frequency patterns between eukaryotic and prokaryotic amino acid sequences in the DNA binding proteins. This distinguishable difference pattern in amino acid and amino acid group frequency can be used to discriminate them. The bar graph (Figure 2) and boxplot (Figure 3 and Figure 4) were used to decipher the patterns of the differences.

**Machine learning results:**
We found that amino acids and amino acid group frequency can be used as features to train a SMO classifier in WEKA to distinguish prokaryotic and eukaryotic DNA binding proteins on other features such as sequence similarity and various other machine learning algorithms.

---

**Table 1:** The distribution of the dataset.

| DNA Binding Protein (DBP) Dataset | Transcription Factor (TF) Dataset |
|----------------------------------|----------------------------------|
| Prokaryote 1 - 270               | Eukaryote Set 1 1 - 270          |
| Eukaryote Set 2 78 - 347         | Eukaryote Set 1 1 - 92           |
|                                  | Eukaryote Set 2 91 - 182         |

The list of UniProt IDs used in these datasets was present in Table S1 (see Supplementary data).

![Flowchart diagram of the in-house python tool.](image)

**Figure 1:** Flowchart diagram of the in-house python tool.
the basis of their amino acid and amino acid group frequency as given in Table 2.

Table 2: Results obtained from WEKA analysis.

|                       | (Transcription Factor Set-1) | (Transcription Factor Set-2) | (DNA Binding Protein Set-1) | (DNA Binding Protein Set-2) |
|-----------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| Total Number of Instances | 184                        | 184                        | 540                        | 540                        |
| Correctly Classified Instances | 94.0217 %                 | 93.4783 %                 | 88.3333 %                 | 90 %                       |
| Incorrectly Classified Instances | 5.9783 %                  | 6.5217 %                  | 11.6667 %                 | 10 %                       |

Detailed Accuracy By Class

| Class     | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class     |
|-----------|---------|---------|-----------|--------|-----------|-----|----------|----------|-----------|
| Prokaryot | 0.924   | 0.043   | 0.955     | 0.924  | 0.939     | 0.881| 0.94     | 0.92     | Prokaryot |
| Eukaryot  | 0.957   | 0.076   | 0.926     | 0.957  | 0.941     | 0.881| 0.94     | 0.908    | Eukaryot  |
| Weighted Avg. | 0.94   | 0.06    | 0.941     | 0.94   | 0.94     | 0.881| 0.94     | 0.914    |

Detailed Accuracy By Class

| Class     | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class     |
|-----------|---------|---------|-----------|--------|-----------|-----|----------|----------|-----------|
| Prokaryot | 0.924   | 0.054   | 0.944     | 0.924  | 0.934     | 0.87 | 0.935    | 0.911    | Prokaryot |
| Eukaryot  | 0.946   | 0.076   | 0.926     | 0.946  | 0.935     | 0.87 | 0.935    | 0.902    | Eukaryot  |
| Weighted Avg. | 0.935  | 0.065   | 0.935     | 0.935  | 0.935     | 0.87 | 0.935    | 0.907    |

Detailed Accuracy By Class

| Class     | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class     |
|-----------|---------|---------|-----------|--------|-----------|-----|----------|----------|-----------|
| Prokaryot | 0.863   | 0.096   | 0.9       | 0.863  | 0.881     | 0.767| 0.883    | 0.845    | Prokaryot |
| Eukaryot  | 0.904   | 0.137   | 0.868     | 0.904  | 0.886     | 0.767| 0.883    | 0.833    | Eukaryot  |
| Weighted Avg. | 0.883  | 0.117   | 0.884     | 0.883  | 0.883     | 0.767| 0.883    | 0.839    |

Detailed Accuracy By Class

| Class     | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class     |
|-----------|---------|---------|-----------|--------|-----------|-----|----------|----------|-----------|
| Prokaryot | 0.904   | 0.104   | 0.897     | 0.904  | 0.9       | 0.8  | 0.9      | 0.859    | Prokaryot |
| Eukaryot  | 0.896   | 0.096   | 0.903     | 0.896  | 0.9       | 0.8  | 0.9      | 0.861    | Eukaryot  |
| Weighted Avg. | 0.9    | 0.1     | 0.9       | 0.9    | 0.9       | 0.8  | 0.9      | 0.86     |

Discussion: Data show that the sequence-based features of the DBPs and TFs could very well be used to distinguish between these classes of organisms. In all our analyses, we obtained an overall accuracy greater than 85% and an AUC value of 0.9. However, we had to use a comparatively small dataset due to paucity of data in the databases. None-the-less, this is the up to date data available till the date mentioned in the manuscript. Available predictors combine both the sequence and structural information for the discrimination purposes. Our predictor uses only sequence information and therefore may be considered a more general one as sequence information is more abundant than structural information. For extraction of the features, we used an in-house script written in python.

Acknowledgment: The authors acknowledge University of Kalyani, Kalyani (W.B.) India and DBT funded Bioinformatics Infrastructure Facility (BIF) for providing infrastructure facilities. NC would like thank theUGC, Govt. of India for financial support. The infrastructural supports from DBT, India (San No. 102/IFD/SAN/1822/2015-2016 dated August 06 2015) and ICMR (BIC/12(02)/2014), Govt. of India are duly acknowledged.
Figure 2: The bar-graph representation of amino acids and amino acid group frequency in prokaryotes and eukaryotes (Blue: Prokaryote; Red: Eukaryote Set-1; Green: Eukaryote Set-2).

Figure 3: Amino acids and amino acid group frequency from TF dataset.
Figure 4: Amino acids and amino acid group frequency from DBP dataset.

References:

[1] Latchman DS. International Journal of Biochemistry and Cell Biology. 1997, 29:1305. [PMID: 9570129]
[2] Spitz & Furlong, Nat. Rev. Genet. 2012, 13:613. [PMID: 22868264]
[3] Bagchi A. Gene. 2016, 586:274 [PMID: 27083770]
[4] UniProt Consortium, Nucleic Acids Res. 2013, 41:D43. [PMID: 23161681]
[5] Al-Shahib A et al. BMC Genomics. 2007, 8:78 [PMID: 17374164]
[6] Hamming RW. Bell Syst. Tech. J. 1950, 29:147.
[7] Blackburne BP and Whelan S. Bioinformatics. 2012, 28:495. [PMID: 22199391]
[8] Nelson DL & Cox MM. Lehninger Principles of Biochemistry. 2005, 4.
[9] Frank E et al. “Weka,” in Data Mining and Knowledge Discovery Handbook, 2005, 1305.
[10] Frank E et al. Bioinformatics. 2004, 20:2479 [PMID: 15073010]

License statement: This is an Open Access article which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. This is distributed under the terms of the Creative Commons Attribution License.
**Table S1:** List of UniProt id of the FASTA files used as dataset

| Prokaryotic TF | Eukaryotic TF | Prokaryotic DBP | Eukaryotic DBP |
|---------------|--------------|----------------|----------------|
| A0A0H2VJZ8    | A0AVK6       | A0A072Z681     | A0AVK6         |
| A0QZ11        | A2D9X4       | A0A0H2VJZ8     | A0JP82         |
| A0R618        | G0SB31       | A0A0H2XIU6     | A2D9X4         |
| A6T8N1        | G4NEJ8       | A0QZ11         | A5J036         |
| B2SU53        | L711M8       | A0R618         | A6ZL36         |
| B8FW11        | O00327       | A3DJ38         | B4F610         |
| C3W947        | O00482       | A3FMN7         | C0JJWR6        |
| D5KM69        | O15350       | A57Y69         | C7SWF3         |
| G3XY4         | O15409       | A6T8N1         | D2W6T1         |
| O34777        | O43435       | B2MU09         | D9IWL3         |
| O36517        | O43524       | B2SU53         | D9J034         |
| O66551        | O54790       | B8FW11         | E0YCK3         |
| O66858        | O94916       | C1D7P6         | F7WD42         |
| O68014        | O95238       | C3W947         | G0SB31         |
| O69245        | P01100       | D4EMQ0         | G4NEJ8         |
| P03023        | P01106       | D5KM69         | L711M8         |
| P03052        | P02340       | D5MXN7         | M1GSK9         |
| P06533        | P02833       | D9N168         | O00327         |
| P06534        | P02836       | E1C9K5         | O00482         |
| P07674        | P03001       | G3XY4          | O13988         |
| P0A0I7        | P03069       | O25100         | O14770         |
| P0A0N4        | P03372       | O25386         | O14862         |
| P0A247        | P04150       | O25758         | O15350         |
| P0A4T9        | P04386       | O25841         | O15409         |
| P0A6X7        | P04637       | O34777         | O15527         |
| P0A881        | P05412       | O38417         | O43435         |
| P0A8U6        | P05554       | O52512         | O43524         |
| P0A8V6        | P05725       | O66551         | O54790         |
| P0AC10        | P06536       | O66659         | O74859         |
| P0AC7         | P06601       | O68588         | O75362         |
| P0AC9         | P06602       | O68014         | O75531         |
| P0ACT         | P07270       | O68557         | O80358         |
| P0AF28        | P07272       | O68847         | O82175         |
| P0AFJ5        | P08046       | O69245         | O94468         |
| P0AG30        | P08368       | O87365         | O95238         |
| P0AGK8        | P09077       | O87963         | O95243         |
| P0CIU6        | P09631       | P00582         | O95551         |
| P0DJL7        | P09956       | P00642         | P00639         |
| P10026        | P0C852       | P00648         | P00734         |
| P17893        | P0CY08       | P02958         | P01100         |
| P21866        | P0CY10       | P03004         | P01106         |
| P22262        | P10037       | P03013         | P01127         |
| P23873        | P10085       | P03018         | P01837         |
| P23874        | P10276       | P03023         | P02263         |
| P25144        | P11473       | P03052         | P02340         |
| P27709        | P11851       | P03067         | P02853         |
| P33905        | P11938       | P03856         | P02856         |
| P39075        | P13297       | P04390         | P03001         |
| P40676        | P13393       | P04395         | P03069         |
| P44558 | P14859 | P04995 | P03372 |
|--------|--------|--------|--------|
| P46428 | P14921 | P05050 | P03870 |
| P68261 | P15036 | P05102 | P03880 |
| P71099 | P15207 | P05327 | P03882 |
| P96711 | P15806 | P05525 | P04150 |
| P9WGZ1 | P16236 | P06134 | P04275 |
| P9WJB7 | P16767 | P06533 | P04386 |
| P9WME9 | P17679 | P06534 | P04637 |
| P9WMF8 | P17789 | P06612 | P05231 |
| P9WMH1 | P18113 | P07013 | P05412 |
| P9WMH3 | P19419 | P07674 | P05554 |
| P9WPY9 | P19544 | P08394 | P05725 |
| Q0F6M2 | P19793 | P09184 | P06401 |
| Q1D4H5 | P19838 | P09546 | P06536 |
| Q2ACK9 | P20153 | P09883 | P06601 |
| Q2FZS6 | P20226 | P09980 | P06602 |
| Q32WH4 | P20263 | P0A017 | P06766 |
| Q3ZD72 | P20393 | P0A0N4 | P06786 |
| Q45782 | P20823 | P0A247 | P07199 |
| Q46731 | P21952 | P0A459 | P07270 |
| Q46864 | P22121 | P0A4T9 | P07272 |
| Q57468 | P22415 | P0A6C1 | P07276 |
| Q5F882 | P22670 | P0A6R3 | P08046 |
| Q5Y812 | P22829 | P0A626 | P08151 |
| Q746J7 | P23511 | P0A7C2 | P08638 |
| Q7AKF2 | P23760 | P0A7G6 | P09077 |
| Q7XOD9 | P23772 | P0A809 | P09631 |
| Q8T3D2 | P24781 | P0A881 | P09651 |
| Q8A8V8 | P25490 | P0A8J2 | P09838 |
| Q8E565 | P25502 | P0A8U6 | P09874 |
| Q8G8H0 | P25799 | P0A8V6 | P09884 |
| Q8NMG3 | P27577 | P0A988 | P09956 |
| Q8YAF1 | P28147 | P0A9H1 | P0C882 |
| Q93Z0 | P28324 | P0A835 | P0CY08 |
| Q9CHR1 | P28347 | P0A5C1 | P0CY10 |
| Q9EZJ8 | P29617 | P0A5C10 | P10037 |
| Q9HUS3 | P31266 | P0A7C8 | P10085 |
| Q9JIS1 | P34707 | P0ACP7 | P10276 |
| Q9KQU8 | P35680 | P0AC52 | P11308 |
| Q9KWU8 | P35869 | P0ACT4 | P11387 |
| Q9S166 | P36956 | P0AD12 | P11473 |
| Q9ZH6 | P38144 | P0AE8 | P11831 |
| P38830 | P0AEK0 | P11938 |
| P38867 | P0AF28 | P12689 |
| P41235 | P0AFJ5 | P12956 |
| P42226 | P0AFY8 | P13051 |
| P42227 | P0AG30 | P13297 |
| P42582 | P0AG74 | P13393 |
| P43680 | P0AGE0 | P13864 |
| P46531 | P0AGK8 | P14585 |
| P47902 | P0CIK6 | P14653 |
| P48436 | P0CI76 | P14736 |
| P49711 | P0D1L7 | P14859 |
| P51608 | P0DJO8 | P14921 |
| P52952 | P11405 | P15036 |
| P53539 | P13920 | P15207 |
| P53762 | P13925 | P15424 |
| Accession Code | Accession Code | Accession Code |
|----------------|----------------|----------------|
| Q8GZB6         | P9W1I3         | P36956         |
| Q8IKJ2         | P9W1J7         | P38144         |
| Q8L7G0         | P9WME9         | P38850         |
| Q8MXE7         | P9WMF8         | P38867         |
| Q8NH1W3        | P9WMH1         | P39748         |
| Q94702         | P9WMH3         | P41235         |
| Q94IF5         | P9WNV3         | P42224         |
| Q95VR4         | P9WPY9         | P42226         |
| Q969G2         | Q031W6         | P42227         |
| Q99551         | Q06B24         | P42582         |
| Q99626         | Q0P6M2         | P43246         |
| Q9C932         | Q1D415         | P43680         |
| Q9H3D4         | Q2ACK9         | P46531         |
| Q9NQV7         | Q2FZ56         | P47902         |
| Q9NUX5         | Q2I6W2         | P48436         |
| Q9UHX1         | Q32WH4         | P49711         |
| Q9UMN6         | Q3ZD72         | P49916         |
| Q9Y5R6         | Q45458         | P50534         |
| Q45488         | P50549         |                |
| Q45782         | P51608         |                |
| Q46731         | P52952         |                |
| Q46864         | P53539         |                |
| Q46896         | P53762         |                |
| Q46944         | P53999         |                |
| Q47112         | P54098         |                |
| Q47152         | P54132         |                |
| Q47155         | P54274         |                |
| Q47673         | P54841         |                |
| Q47P10         | P55265         |                |
| Q4UNB2         | P55318         |                |
| Q53632         | P56178         |                |
| Q56215         | P60896         |                |
| Q57253         | P61244         |                |
| Q57267         | P61823         |                |
| Q57468         | P61978         |                |
| Q5F882         | P62805         |                |
| Q5F9M9         | P63159         |                |
| Q5J66         | P70118         |                |
| Q5KWC1         | P70340         |                |
| Q5L0J3         | P70348         |                |
| Q5T64         | P70512         |                |
| Q5T65         | P83949         |                |
| Q5T65         | P84022         |                |
| Q5Y812         | P87249         |                |
| Q72I39         | P97360         |                |
| Q746J7         | P97471         |                |
| Q746M7         | P98177         |                |
| Q7AKF2         | Q00559         |                |
| Q7CWV1         | Q00403         |                |
| Q7DD47         | Q00422         |                |
| Q7MHK3         | Q00613         |                |
| Q7X0D9         | Q00653         |                |
| Q816E8         | Q00958         |                |
| Q83TD2         | Q01147         |                |
| Q84AF2         | Q01167         |                |
| Q8AAV8         | Q01543         |                |
| Q8DPM2         | Q01663         |                |
| Accession Numbers | Accession Numbers |
|-------------------|-------------------|
| Q8E565            | Q01826            |
| Q8EFJ3            | Q02078            |
| Q8EIX3            | Q02080            |
| Q8EVR5            | Q02486            |
| Q8GGH10           | Q02548            |
| Q8KNP2            | Q02880            |
| Q8NMC3            | Q03164            |
| Q8RT59            | Q03347            |
| Q8RNV5            | Q04049            |
| Q8RNV8            | Q04206            |
| Q8RT53            | Q04207            |
| Q8YAF1            | Q04863            |
| Q8Z2A5            | Q05195            |
| Q8ZG78            | Q05783            |
| Q928V6            | Q06330            |
| Q933Z0            | Q06453            |
| Q93PU6            | Q06831            |
| Q97FM4            | Q07230            |
| Q99U17            | Q08050            |
| Q9AC34            | Q08874            |
| Q9AFI5            | Q08957            |
| Q9AMH9            | Q12778            |
| Q9CHR1            | Q13469            |
| Q9EJZ8            | Q13569            |
| Q9FL6L0           | Q14191            |
| Q9HUS3            | Q14653            |
| Q9HUM3            | Q14863            |
| Q9IIS1            | Q14919            |
| Q9JZ0N            | Q15109            |
| Q9KEJE            | Q15365            |
| Q9KJ88            | Q15366            |
| Q9KQU8            | Q15554            |
| Q9KVD2            | Q15561            |
| Q9KWU8            | Q16254            |
| Q9KROR            | Q16531            |
| Q9RPJ3            | Q16666            |
| Q9RT63            | Q17034            |
| Q9RWI8            | Q25442            |
| Q9RY80            | Q3UPPW2           |
| Q9S166            | Q4PRK9            |
| Q9WY48            | Q4VWW5            |
| Q9WYV0            | Q58HP3            |
| Q9X2H9            | Q5AP80            |
| Q9X4C9            | Q5EAW4            |
| Q9XDH5            | Q5NE14            |
| Q9Z3B4            | Q5XJA0            |
| Q9Z0H6            | Q60793            |
| Q9ZL26            | Q61473            |
| V6F4Q0            | Q64249            |
| Q6E01             | Q6CPM4            |
| Q6MZP7            | Q6N021            |
| Q6N538            | Q6NT76            |
| Q6ZQJ5            | Q7ID3             |

ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 14(6): 315-326 (2018) 324 ©2018
