Classification of Protein-DNA Complexes
Based on Structural Descriptors

Ponraj Prabakaran\textsuperscript{1} \hfill Shandar Ahmad\textsuperscript{2} \hfill M. Michael Gromiha\textsuperscript{3}
npraba@ncifcrf.gov \hfill shandan@bse.kyutech.ac.jp \hfill michael-gromiha@aist.go.jp

Maria G. Singarayan\textsuperscript{4} \hfill Akinori Sarai\textsuperscript{5}
singarayan@dtpax2.ncifcrf.gov \hfill sarai@bse.kyutech.ac.jp

\textsuperscript{1} Protein Interactions Group, CCRNP, National Cancer Institute, NIH, Frederick, MD 21702, USA
\textsuperscript{2} Department of Biosciences, Jamia Millia Islamia University, New Delhi-110025, India
\textsuperscript{3} Computational Biology Research Center (CBRC), AIST 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan
\textsuperscript{4} Developmental Therapeutics Program, SAIC-Frederick, Inc., National Cancer Institute, Frederick, MD 21702, USA
\textsuperscript{5} Dept. Bioscience and Bioinfo., Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, 820-8502, Japan

Keywords: aberrant splicing, database, point mutation, scanning model

1 Introduction

DNA-binding proteins play essential roles in gene regulation. So far, there are more than 1,000 known structures of different protein-DNA complexes. DNA-binding proteins are usually classified according to the structural motif of proteins such as helix-turn-helix and Zinc finger motifs [1]. However, the conservation of structural motif in proteins may not necessarily dictate the way by which DNA sequences are recognized. Proteins within some classes are quite diversified in terms of structure, the mode of interactions and wide range of recognition sequences. Here, we adopt a different approach to classify protein-DNA complexes, which uses a set of structural descriptors mainly characterizing protein-DNA interactions. The cluster analyses were carried out for a unique set of 62 complexes including a variety of protein motifs, and 7 distinct clusters were revealed from the analyses. We found that some proteins with the same motif are classified into different clusters whereas different proteins with distinct motifs are classified into the same cluster. These results suggest that the conventional motif-based classification of DNA-binding proteins may not necessarily correspond to structural and functional properties of protein-DNA complexes, and that the present classification will help to identify common properties and rules that govern protein-DNA recognition.

2 Methods

2.1 Dataset

We have considered the same set of 62 non-redundant protein-DNA complexes, which we used in our previous studies [2].

2.2 Descriptors

We considered a set of structural descriptors mainly characterizing protein-DNA interactions. We started with 22 structural parameters and got rid of highly correlated ones. We selected the final set of 11 parameters that were used for the clustering. These parameters include the number of atomic contacts at major and minor grooves of DNA, conformational deviations from standard B- and A-DNA forms, widths of DNA grooves, GC-content, specificity measures (Z-scores) of direct and indirect readouts, and buried surface area at complex interface.
2.3 Clustering Methods

Since our dataset contained variables having different units and ranges of values, we first performed a Z-score transformation that standardized the data. We constructed a proximity matrix by computing squared Euclidean distances between all pairs of the 62 protein-DNA complexes. We used hierarchical and $k$-means methods for the clustering by using ClustanGraphics5 [3]. We used Ward’s hierarchical method as a linkage technique. For the hierarchical case, a bootstrap validation was carried out after performing the Ward’s procedure. We performed the $k$-means method by varying the value of $k$ from 2 to 10, and validated the clustering model by carrying out the $k$-means analysis $n$ times with $n-1$ cases, where $n = 62$, excluding one complex from the list at a time, and calculated the frequency of occurrence of each complex for a particular cluster.

3 Results and Discussion

We obtained the result of hierarchical clustering for the 62 protein-DNA complexes, and the number of clusters in the hierarchical tree was selected as 7 after analyzing the fusion coefficients of clustering tree and intracluster distances from the $k$-means models. We compared the results of the two approaches to examine the robustness of the cluster assignments and conserved group memberships. The average conservation of cluster membership ranges from 65 to 98%. We calculated the cluster profiles obtained from the $k$-means analysis. The profiles reflect the way in which the descriptors are distributed across the seven clusters, and a closer examination of cluster profiles reveals the roles of various parameters in different clusters and their interplay in the cluster formation and function. The present results show that some proteins with the same motif are classified into different clusters whereas different proteins with distinct motifs are classified into the same cluster, suggesting that the motif-based classification of DNA-binding proteins may not necessarily correspond to structural and functional properties characterizing protein-DNA recognition. The present approach using cluster analysis with structural descriptors would provide a potential framework for identifying common properties and rules that govern protein-DNA recognition.

References

