Structure *14*, 1355–1367, September 2006 DOI 10.1016/j.str.2006.06.018

Classification of Protein-DNA Complexes Based on Structural Descriptors

Ponraj Prabakaran,[1,2] Jörg G. Siebers,[1] Shandar Ahmad,[1,3] M. Michael Gromiha,[1,4] Maria G. Singarayan,[1,5] and Akinori Sarai[1,*]
[1] Department of Bioscience and Bioinformatics
Kyushu Institute of Technology
680-4 Kawazu
Iizuka 820-8502
Japan

Summary

We attempt to classify protein-DNA complexes by using a set of 11 descriptors, mainly characterizing protein-DNA interactions, including the number of atomic contacts at major and minor grooves, conformational deviations from standard B- and A-DNA forms, widths of DNA grooves, GC content, specificity measures of direct and indirect readouts, and buried surface area at the complex interface. 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.

Introduction

Proteins that bind to specific DNA target sites play essential roles in all aspects of gene regulation. In recent years, structural and biochemical studies have been devoted to characterizing protein-DNA recognition (Sarai and Takeda, 1989; Pabo and Sauer, 1992; Steitz, 1993; Luisi, 1995). To date, more than 1,000 structures of different protein-DNA complexes have been determined (Deshpande et al., 2005) and have been previously overviewed for classifications (Luscombe et al., 2000). These structures have revealed distinct types of structural motifs of proteins such as helix-turn-helix (HTH) and zinc finger (Zn-finger) motifs. While each class of proteins exhibits a rather similar binding mode with DNA, proteins within some classes are quite diverse in terms of structure, the mode of interactions, and the wide range of recognition sequences. Also, different types of proteins are combined to form a molecular complex in the promoter region to provide highly cooperative regulation of gene expression. Thus, the conservation of a structural motif in proteins may not necessarily dictate the way by which DNA sequences are recognized, and the relationship between the classification of DNA binding proteins based on the protein motif and the functional characteristics of protein-DNA recognition is not so obvious. In particular, it remains to be clarified whether proteins within the same class use a similar mechanism of sequence recognition, and whether different classes of proteins share some common mechanisms.

There have been a number of statistical analyses that have been used to determine the general principles of protein-DNA interaction. With respect to the contacts between protein and DNA, various interactions, including hydrogen bonding (Mandel-Gutfreund et al., 1995; Luscombe et al., 2001) C-H···O (Mandel-Gutfreund et al., 1998), cation-π (Wintjens et al., 2000; Rooman et al., 2002; Gromiha et al., 2004a), and water mediation (Reddy et al., 2001), have been studied. The other analyses include the physicochemical properties, namely, amino acid sequence composition, solvent accessibility, electrostatics (Ahmad et al., 2004; Ahmad and Sarai, 2004), polarity, packing, DNA conformation parameters, interface area (Nadassy et al., 1999; Jones et al., 1999), and spatial relationships at the protein-DNA interface (Pabo and Nekludova, 2000). All such studies focused mostly on a subset of features and thus may not lead to a unified understanding of protein-DNA recognition.

Here, we attempt to classify protein-DNA complexes based on a clustering approach by taking into account most of the key structural parameters involved in the recognition process and by identifying the patterns and relationships existing in different subgroups of the complexes. In the present analysis, we have used 11 parameters to classify a data set of 62 protein-DNA complexes. The computed parameters specify important properties of both protein and DNA upon binding, such as the number of atomic contacts, GC content, DNA conformation and parameters, and buried interface area. We also included statistical energy Z-score values for direct and indirect readouts with which we recently quantified the specificities of protein-DNA interactions (Selvaraj et al., 2002; Gromiha et al., 2004b). Thus, these parameters serve as unique variables or descriptors that form the basis for our cluster analysis. We will compare the classification based on the present analysis with the conventional motif-based classification, and we will 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. We will discuss a possible implication of these results for the mechanism of protein-DNA recognition.

*Correspondence: sarai@bse.kyutech.ac.jp
[2] Present addresses: Protein Interactions Group, CCRNP, National Cancer Institute, National Institutes of Health, Frederick, Maryland 21702.
[3] Present address: Department of Biosciences, Jamia Millia Islamia University, New Delhi-110025, India.
[4] Present address: Computational Biology Research Center (CBRC), AIST 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan.
[5] Present address: Developmental Therapeutics Program, SAIC-Frederick, Inc., National Cancer Institute, Frederick, Maryland 21702.

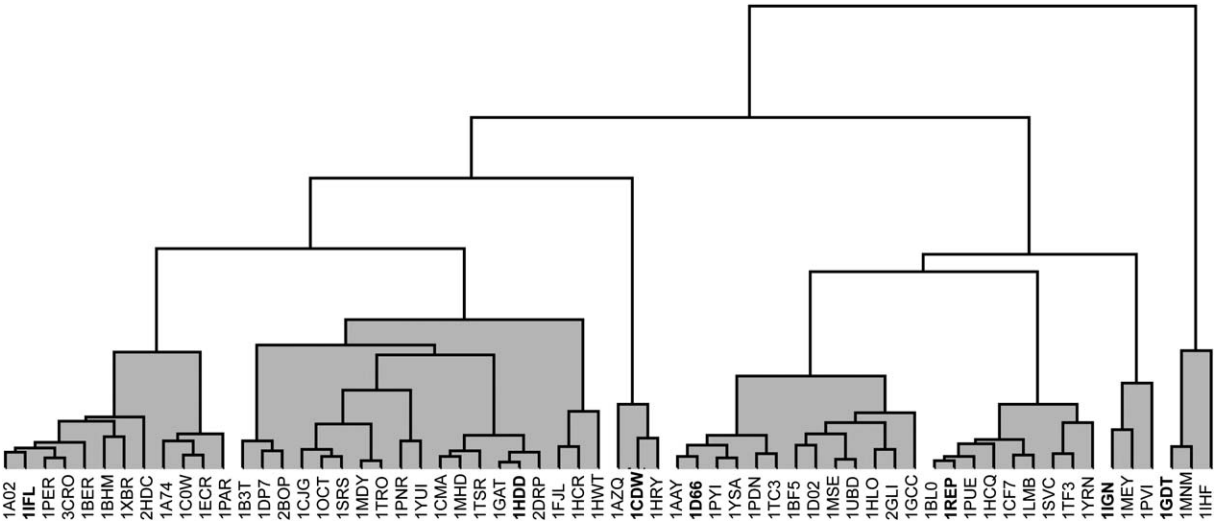

Figure 1. Dendogram of 62 Protein-DNA Complexes Obtained by a Hierarchical Cluster Analysis

A hierarchical tree (obtained by Ward's method) is depicted with seven clusters, and the exemplars are shown in bold type.

Results

Analyzing the Number of Clusters in the Data Set

Figure 1 shows the result of hierarchical clustering for the 62 protein-DNA complexes. The number of clusters in the hierarchical tree was selected as seven after analyzing the fusion coefficients of the clustering tree and the intracluster distances from the *k*-means models. We carried out a *k*-means cluster analysis in a recursive manner with $k = 2–10$, and we found that $k > 8$ produced outliers with a single member. The *k*-mean method takes the input parameter, *k*, and partitions the total number of objects into *k* clusters so that the intracluster similarity is high and the intercluster similarity is very low. In fact, the *k*-means algorithm computes the distance between each object in the clusters and the means of all of the clusters. If an object is found to be closer to a cluster other than the one to which it presently belongs, it is reassigned to its closest cluster. In this way, the means of all clusters were calculated and the partitions were successively reassigned until the squared-error variance converged toward a minimum, while the intercluster distance converged toward a maximum. The squared-error criterion is defined as $E = \Sigma_{i=1,k}\Sigma_{p\in Ci}|p - m_i|^2$, where *E* is the sum of the square-error term for all objects in the data set, *p* is the point in space that represents a given object, and m_i refers to the mean of cluster C_i.

The results of both hierarchical and *k*-means clustering analyses were used to determine the number of clusters in the data set. For the hierarchical case, a bootstrap validation as implemented in ClustanGraphics5 was carried out after performing the Ward's procedure. Here, the bootstrapping process aims to identify partitions that are farthest from random. The sequences of actual fusion values obtained from the data set were compared with hierarchical clustering sequences obtained by generating random trails of 120 trees up to the 25 cluster level. The significance test with a t-statistic value of 2.57 at a departure from randomness was considered to be significant. The results of the bootstrap validation comparing the fusion coefficients for a data set of protein-DNA complexes with these values for randomized subtrees are given in Figure 2A. The dark-gray area at the bottom of the graph shows the fusion values calculated by using the actual data set that were divided into a different number of clusters (*k*). The white band shows the range of fusion values obtained from 120 trials of randomizing the data; in this confidence interval, the central line represents the mean of the fusion values for each number of clusters, obtained from the random trials. The width of the white band is one standard deviation about the mean. The gray zones indicate where the fusion values for the given data depart significantly from random. In the present data set of protein-DNA complexes, the significant departure from random occurs within ten clusters only. We also analyzed the number of clusters in the hierarchical tree by using the best-cut procedure (Mojena and Wishart, 1980) as adopted in ClustanGraphics5, which uses significance tests on the fusion values and identifies the number of clusters in the tree. The best-cut procedure suggested a model with 12 clusters and an outlier, integration host factor complex (PDB code: 1IHF), which had a t-statistic value of 2.65 at the 5% significance level. Further, the model could be truncated into nine clusters without an outlier. The procedure was based on the upper tail method, which uses the fusion values as a series and computes the mean and standard deviation, and a t-statistic as the standardized deviation from the mean. It then computes the standard deviation for each fusion value on this distribution, which is assumed to be normal, and selects the first one as "significant" if its t-value exceeds the 5% level.

Further, we analyzed the results of the *k*-means method obtained by varying the value of *k* from 2 to 10. The mean intracluster distance was calculated for each *k*-means model and plotted in Figure 2B. We found that models with $k > 8$ resulted in outliers with a single member and that a cluster model with $k = 7$ has the lowest average intracluster distance among the models

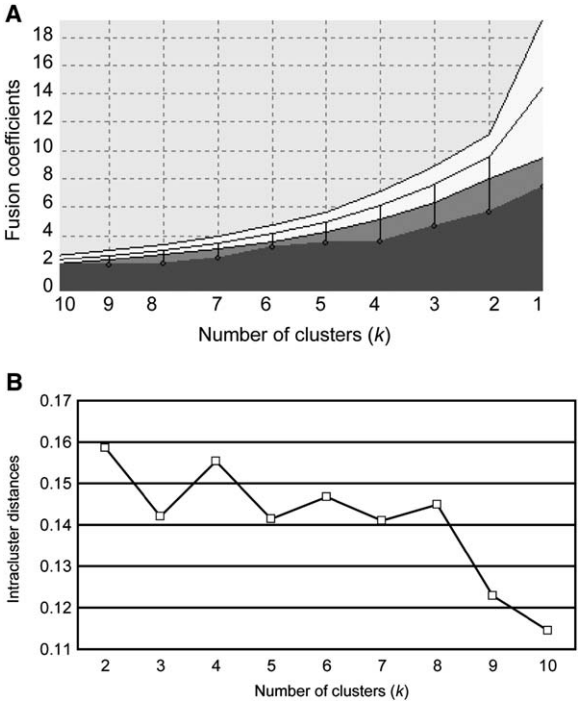

Figure 2. Estimation of the Number of Clusters in the Data Set of Protein-DNA Complexes

(A) Bootstrap validation based on the results of hierarchical cluster analysis. The dark-gray portion shows fusion values obtained from original data as presented, the white band represents the range of fusion values obtained from 120 trials of randomizing the data, and the gray zones indicate where the fusion values for the given data depart significantly from random.

(B) The intracluster distance is the average value of distances measured between cluster centroids and corresponding exemplars for a particular cluster.

with $k \leq 8$. Therefore, we assigned the number of clusters (k) as seven for the data set, and the seven cluster members for hierarchical and k-means models are grouped in Table 1. By comparing the cluster members from hierarchical and k-means models, we could observe more than 70% cluster membership conservation across the six clusters. PDB codes in italics represent matching members in the corresponding six clusters: 1, 2, 3, 4, 5, and 7 from the hierarchical method are comparable to 1, 7, 4, 3, 6, and 2, respectively, from the k-means method (Table 1). Cluster 6 from the hierarchical model and cluster 5 from the k-means model do not have any matching members, as they dispersed with other clusters that have shorter intercluster distances in the k-means model, for example between clusters 3 and 6, 5, and 7 (see Table 4).

Statistical Validations and Intra- and Intercluster Distances

Since we sought to establish a descriptive basis for the classification of protein-DNA complexes and quantitative results by using numerical descriptors, we adhered to the results of k-means clustering for testing statistical validations and interpreting the results. Moreover, there is no provision for a relocation of objects that may have been incorrectly grouped at an early stage of hierarchical clustering. However, the k-means partitioning refines the cluster structure by relocating the members so that each cluster become more compact and its members become tightly situated around the centroid of the cluster. Table 2 shows the list of 62 protein-DNA complexes along with 11 parameters, which are grouped into 7 clusters by using the k-means analysis. To validate the k-means clustering model with $k = 7$, we carried out the k-means analysis n times with $n - 1$ cases, where $n = 62$, excluding one complex from the list at a time, and we calculated the frequency of occurrence of each complex for a particular cluster. This process is referred to as the jackknife procedure and is used to establish the statistical significance. By using the results of jackknife calculations, we quantified the cluster membership for each member in the seven clusters. The conservation of cluster membership (N) and intracluster distance (D) for each protein-DNA complex in the seven clusters are given in Table 3. The intracluster distance is the

Table 1. Cluster Memberships for the Seven Clusters Identified by Hierarchical and k-Means Analyses

Hierarchical Clustering (Ward's Method)	PDB Codes
Cluster 1	*1A02 1IF1 1PER 3CRO 1BER 1BHM 1XBR 2HDC* 1A74 1C0W 1ECR 1PAR
Cluster 2	*1B3T 1DP7 2BOP 1CJG* 1OCT 1SRS *1MDY 1TRO* 1PNR *1YUI 1CMA 1MHD* 1TSR *1GAT 1HDD* 2DRP 1FJL 1HCR 1HWT
Cluster 3	*1AZQ 1CDW 1HRY*
Cluster 4	*1AAY 1D66 1PYI 1YSA* 1PDN 1TC3 *1BF5 1D02 1MSE* 1UBD *1HLO 2GLI 1GCC*
Cluster 5	*1BL0 1REP 1PUE 1HCQ 1CF7 1LMB 1SVC 1TF3 1YRN*
Cluster 6	1IGN 1MEY 1PVI
Cluster 7	*1GDT 1MNM 1IHF*
k-Means Clustering	**PDB Codes**
Cluster 1	*1A02 1BER 1BHM 1IF1 1PER* 1TSR *1XBR 2HDC 3CRO*
Cluster 2	1A74 1C0W 1ECR *1GDT 1IHF 1MNM* 1PAR
Cluster 3	*1AAY 1BF5 1D02 1D66 1GCC 1HLO 1MSE* 1PVI *1PYI 1YSA* 2DRP *2GLI*
Cluster 4	*1AZQ 1CDW 1HRY*
Cluster 5	1FJL 1HCR 1HWT 1OCT 1PNR 1SRS
Cluster 6	*1BL0 1CF7 1HCQ* 1IGN *1LMB* 1MEY *1PUE 1REP 1SVC 1TF3* 1UBD *1YRN*
Cluster 7	*1B3T 1CJG 1CMA 1DP7 1GAT 1HDD 1MDY 1MHD* 1PDN 1TC3 *1TRO 1YUI 2BOP*

PDB codes in italics denote 44 conserved members across 6 clusters in the 2 cluster models.

Table 2. The 7 Clusters of Protein-DNA Complexes with Their DNA Binding Motifs and Calculated Values for the 11 Descriptors

PDB Code	Protein Name	Motif	E/P	M-cont	m-cont	TCD
Cluster 1						
1A02	Fos/Jun/NFAT	LZ	E	24	2	0.68
1BER	Catabolite gene activator protein	HTH	P	19	0	0.63
1BHM	Endonuclease BamHI	LSH	P	25	3	1.06
1IF1	Interferon regulatory factor 1	HTH	E	16	0	0.83
1PER	434 repressor-OR3	HTH	P	24	0	0.85
1TSR	p53 tumor suppressor	LSH	E	10	0	1.16
1XBR	T domain (Brachyury TF)	β barrel	E	4	2	1.07
2HDC	HNF-3/fkh TF Genesis	Winged helix	E	31	8	0.75
3CRO	434 Cro-OR1	HTH	P	16	0	0.55
Cluster 2						
1A74	Homing endonuclease I	β ribbon	E	25	3	0.56
1C0W	Diphtheria toxin repressor	HLH	P	35	0	0.58
1ECR	Replication terminator protein	β ribbon	P	18	9	0.71
1GDT	Recombinase γδ resolvase	HTH	P	29	17	0.33
1IHF	Integration host factor	β ribbon	P	5	15	0.35
1MNM	MATα2/MCM	HTH	E	27	20	0.28
1PAR	Arc repressor	β ribbon	P	48	0	0.31
Cluster 3						
1AAY	Zif268 (three zinc fingers)	ZF	E	28	0	0.29
1BF5	STAT-1	β hairpin	E	10	7	0.31
1D02	Endonuclease MunI	HTH	P	24	0	0.68
1D66	GAL4	ZF	E	19	0	0.46
1GCC	Aterf1-GCC-box binding domain	β ribbon	P	32	0	0.79
1HLO	TF Max	HTH	E	24	0	0.21
1MSE	c-Myb	HTH	E	31	2	0.48
1PVI	PvuII endonuclease	β hairpin	P	62	1	0.49
1PYI	Pyrimidine pathway regulator 1	ZF	P	6	0	0.33
1YSA	GCN4	LZ	E	8	0	0.21
2DRP	Tramtrack protein	ZF	E	19	0	0.38
2GLI	Zinc finger protein GLI1	ZF	E	33	0	0.27
Cluster 4						
1AZQ	Hyperthermophile Sac7d	β barrel	A	0	4	0.74
1CDW	Human TBP core domain	β ribbon	E	0	15	0.81
1HRY	Human SRY	HMG box	E	18	19	0.45
Cluster 5						
1FJL	Paired homeodomain	HTH	E	10	25	0.47
1HCR	Hin recombinase	HTH	P	11	12	0.51
1HWT	HAP1	ZF	P	42	26	0.48
1OCT	Oct-1 POU domain	HTH	E	11	6	0.36
1PNR	Purine repressor	HTH	P	28	14	0.55
1SRS	Serum response factor core	Coiled coil	E	3	11	0.31
Cluster 6						
1BL0	MarA	HTH	P	15	0	0.68
1CF7	Transcription factor E2F-4	Winged helix	E	15	2	0.64
1HCQ	Estrogen receptor	ZF	E	37	0	0.83
1IGN	Rap1	HTH	E	68	12	0.42
1LMB	lambda repressor	HTH	P	35	0	0.65
1MEY	Consensus zinc finger protein	ZF	na	73	0	0.32
1PUE	PU.1 ETS domain	HTH	E	23	0	0.87
1REP	Replication initiation protein	HTH	P	21	3	0.74
1SVC	Transcription factor NF-κB	β barrel	E	38	0	0.61
1TF3	Transcription factor IIIA	ZF	E	31	0	0.31
1UBD	Human YYI	ZF	E	35	0	0.23
1YRN	MATa1/α2	HTH	E	19	9	0.74
Cluster 7						
1B3T	Nuclear protein EBNA1	β helix	E	11	7	0.84
1CJG	lac repressor	HTH	P	3	6	0.61
1CMA	Met repressor-operator	β ribbon	P	8	0	0.67
1DP7	MHC class II TF hRFX1	Winged helix	E	20	6	1.06
1GAT	GATA-1	ZF	E	15	2	0.57
1HDD	Engrailed homeodomain	HTH	E	7	3	0.63
1MDY	MyoD bHLH domain	HTH	E	22	0	0.23
1MHD	Smad3	β hairpin	E	20	0	0.76

Table 2. *Extended*

$Rmsd_{cont}$-B	$Rmsd_{cont}$-A	mgw	Mgw	GC%	PD-Z	DNA-Z	ΔASA
2.62	5.96	5.38	11.28	27.77	−3.36	−1.83	4229
6.76	7.13	3.91	11.21	40.00	−1.95	−0.81	2803
2.46	5.47	6.00	14.13	40.00	−2.88	−1.25	4104
2.92	6.39	5.30	12.26	38.46	−1.06	−1.67	4241
3.29	6.80	5.58	11.85	26.31	−2.45	−1.10	2835
2.74	5.00	6.42	11.31	40.00	−1.42	−1.15	2109
4.52	5.66	6.83	11.79	21.05	−2.04	−2.37	4274
1.68	7.40	3.65	13.73	23.52	−0.27	−0.55	2884
2.99	6.54	5.55	11.71	26.31	−2.04	0.26	2944
6.30	6.92	7.58	13.97	40.00	−1.57	0.74	4339
5.73	6.22	8.21	12.27	38.09	−0.51	−1.50	5226
4.91	6.13	7.95	12.65	28.57	−1.06	−1.09	5122
8.45	7.85	7.16	11.51	34.28	−1.98	−1.72	5671
16.12	11.78	8.46	10.12	38.23	−1.19	−2.26	4911
5.92	8.10	6.14	11.36	32.00	−4.38	−2.96	6605
5.23	7.71	7.12	12.78	33.33	0.59	−1.66	4193
2.06	3.53	7.80	11.29	80.00	−2.96	−1.11	2763
1.57	5.90	5.55	12.32	47.05	−0.27	−0.44	2076
1.99	5.89	7.43	12.54	60.00	1.08	−0.13	2863
2.38	5.20	6.62	11.43	73.68	−1.76	−1.71	2539
1.63	4.25	6.08	11.55	72.72	−0.76	1.33	1605
2.02	3.05	8.43	10.93	63.63	0.10	−1.64	2545
2.57	4.89	5.62	12.84	45.45	−0.36	−2.00	2804
3.56	4.76	8.88	15.11	66.66	1.94	0.08	4365
1.48	5.03	5.74	11.74	57.14	−2.85	−0.65	1923
1.50	3.71	5.94	11.47	63.63	−3.04	−2.10	2283
1.68	3.17	6.52	11.95	38.88	−1.36	−2.28	1738
4.38	4.19	7.37	11.91	60.00	−0.45	0.13	3241
4.51	2.22	11.61	14.08	25.00	−0.90	−0.64	1574
5.60	3.12	10.56	10.91	43.75	−2.24	−0.61	2976
4.48	2.05	12.98	10.56	50.00	−0.16	−0.85	2076
3.26	4.97	6.03	11.25	23.07	−2.70	−1.02	3513
1.30	4.98	6.10	11.05	23.07	−1.76	0.42	2833
2.77	6.97	5.46	10.59	45.00	−3.23	−2.57	2603
2.31	7.85	6.01	12.00	35.71	−1.58	−2.14	3238
4.66	7.06	8.25	10.47	37.50	−0.08	0.69	3665
3.32	7.40	6.19	11.12	44.44	−3.04	−2.37	4036
5.05	6.12	5.56	11.96	45.45	−2.66	−2.49	2443
1.87	7.15	5.55	12.67	46.66	−3.22	−3.69	2718
1.78	6.97	6.39	10.76	58.82	−1.71	−2.48	2553
3.04	7.13	6.20	13.01	66.66	−0.83	−2.20	4298
2.67	7.42	7.13	11.87	47.36	−2.90	−4.34	2978
2.71	4.97	6.62	12.88	50.00	−3.61	−2.22	2556
4.30	5.52	6.16	13.00	53.33	−1.13	−2.72	2057
3.80	6.36	6.83	11.77	52.38	−1.96	−3.15	2707
1.91	4.71	4.30	11.40	72.72	−2.56	−2.19	4118
3.66	5.67	5.52	11.45	60.00	−3.20	−2.32	3252
2.08	6.07	5.09	12.43	50.00	−1.26	−2.12	2863
1.85	4.60	5.93	11.48	47.05	−5.92	−2.87	3427
4.19	5.59	7.87	9.61	55.55	−1.44	−2.10	5386
4.58	6.67	7.65	11.57	45.45	−1.13	−1.44	3400
3.53	5.63	7.65	10.43	44.44	−0.21	−1.57	1769
2.86	5.49	7.98	10.27	50.00	−0.76	−0.73	3512
1.75	4.47	7.91	11.75	25.00	−0.41	−2.49	1985
3.05	4.64	6.99	11.63	30.00	−1.07	−3.69	2759
2.54	5.70	8.06	10.07	42.85	−0.65	−2.48	2676
2.01	4.82	6.78	10.88	38.46	−1.86	−2.20	1933

(*Continued on next page*)

Table 2. *Continued*

PDB Code	Protein Name	Motif	E/P	M-cont	m-cont	TCD
1PDN	Paired domain (prd)	HTH	E	3	2	0.27
1TC3	Transposase	HTH	E	17	8	0.61
1TRO	Trp repressor	HTH	P	19	0	0.25
1YUI	GAGA-factor	ZF	P	23	5	0.31
2BOP	Bovine papilloma virus-1 E2	β barrel	E	20	0	1.12

Abbreviations for motifs: HTH, helix-turn-helix; ZF, zinc finger; HLH, helix-loop-helix; HMG, high-mobility group; LSH, loop-sheet-helix; LZ, leucine zipper. Abbreviations for organisms: E, eukaryote; P, prokaryote. Abbreviations for descriptors: M-cont, number of major groove contacts; m-cont, number of minor groove contacts; TCD, total contact distance; $Rmsd_{cont}$-B, rms deviation from B-DNA (Å); $Rmsd_{cont}$-A, rms deviation from A-DNA (Å); mgw, minor groove width (Å); Mgw, major groove width (Å); GC%, GC content in percent; PD-Z, energy Z-score value for direct interaction; DNA-Z, energy Z-score value for indirect interaction; ΔASA, interface surface area ($Å^2$) of a protein-DNA complex (see Experimental Procedures for the details about each descriptor).

geometric distance between each cluster member and the corresponding cluster's centroid, which is a measure of the tightness of the clusters. The exemplar is one that has the maximum or total conservation (N_{max} = 62) and is closest to the corresponding cluster's center. The average conservation of cluster membership (AVER) ranges from 65% to 98%, which can be calculated as (AVER/N_{max}*100). Using the results of jackknife calculations, a 62 × 62 conservation proximity matrix is derived and displayed by using ClustanGraphics5 as depicted in Figure 3. The proximity matrix diagram illustrates the conservation of cluster members; some of the members within a cluster are much more highly conserved than the rest of the cluster members, which are shown in dark blue. The less conserved complexes are denoted by the proximity matrix representation in light blue. The red color denotes the members of protein-DNA complexes that share cluster membership to more than one cluster. The intercluster distances range from 0.519 to 2.785, as shown in Table 4, while the average intracluster distances range from 0.086 to 0.485 (Table 3). These numeric values suggest that complexes in the same group are tightly clustered and that those between different clusters are separated.

Cluster Profiles Reveal Common Structural/Functional Properties

The cluster profiles obtained from the *k*-means analysis reflect the way in which the descriptors are distributed across the seven clusters (Table 5). Each number shown in Table 5 is the average value of the pertinent descriptor for the corresponding cluster along with the standard deviation value (STDEV). The bold-faced numbers show the maximum and minimum values for the 11 descriptors among the 7 clusters. Most of the profile values of the descriptors have several multiples of their STDEV values. The higher STDEV value might indicate a smaller subset with unusual values for the descriptor. A closer examination of cluster profiles reveals the roles of various parameters in different clusters and their interplay in cluster formation and function. In the following sections, we describe in detail each cluster profile and its role in protein-DNA interactions.

Cluster 1

There are nine members in the cluster. Five of these, including four transcription factors Fos/Jun/NFAT (1A02), interferon regulatory factor IRF-1 (1IF1), Brachyury T domain (1XBR), HNF-3/fkh transcription factor Genesis (2HDC), and a transcriptional activator Catabolite Activator Protein, CAP (1BER), are involved in transcriptional function. These five transcription-related proteins have a variety of DNA binding motifs, including HTH, leucine zipper (LZ), winged α helix, and β barrel (Table 2), which generally bind to the major groove of DNA (Muller, 2001). The other 4 members include 434 repressor-OR3 (1PER), 434 Cro-OR1 (3CRO), endonuclease BamHI (1BHM), and p53 tumor suppressor (1TSR). 434 repressor and 434 Cro, which share the HTH motif, are found to be topologically related by DALI procedure (Holm and Sander, 1993). Endonuclease BamHI and p53 both have a common core motif consisting of a central β sheet with α helices on both sides (LSH). The cluster profile (Table 5) shows that the cluster has the largest average value of the major groove (12.14 Å) as well as the smallest average value of the minor groove (5.4 Å), compared with the corresponding average values of 11.69 and 6.9 Å for the 62 complexes. The average numbers of major and minor groove contacts are 18.8 and 1.7, respectively. These indicate larger major groove recognition via DNA binding motifs than minor groove, where the least numbers of contacts are made. The cluster is also characterized by the lowest average value of GC content (31.5%), which represents the AT-rich segment of DNA targets. Accordingly, the minor groove contacts in the AT-rich region of the binding site appear in the p53 (Cho et al., 1994), 434 repressor, and IRF-1 complexes (Escalante et al., 1998). The other dominant parameter profiling the cluster is TCD, with the highest average value, 0.84. The energy Z-score values of direct and indirect interactions are −1.94 and −1.16, respectively, which may imply prominent direct readouts in protein-DNA recognition.

Cluster 2

The cluster has seven members and includes homing endonuclease I-PpoI (1A74); diphtheria toxin repressor, DtxR (1C0W); replication terminator protein, Tus (1ECR); recombinase γδ resolvase (1GDT); integration host factor, IHF (1IHF); ternary complex containing homeodomain repressor protein MATα2 and the MADS-box transcription factor MCM1 (1MNM); and Arc repressor (1PAR). The main feature of the complexes in this group is the high conformational deformation of bound DNA molecules, which are indicated by the highest average $rmsd_{cont}$-B and $rmsd_{cont}$-A values, 7.52 Å and 7.82 Å, respectively, suggesting neither B form nor A form DNA conformation. The deformation is exhibited through an overall bending of 20°–60° observed in the DNA helical axis in these complexes (Dickerson, 1998); however,

Table 2. *Extended*

$Rmsd_{cont}$-B	$Rmsd_{cont}$-A	mgw	Mgw	GC%	PD-Z	DNA-Z	ΔASA
2.53	5.17	7.94	11.14	61.53	−2.00	−4.34	2592
2.34	5.39	8.58	11.79	57.89	−1.66	−2.22	1957
3.79	6.33	6.89	10.63	33.33	−1.28	−2.72	3162
3.53	8.10	9.15	10.25	33.33	1.27	−3.15	2326
3.64	6.05	6.96	9.63	75.00	−0.85	−2.19	2974

an extreme change is seen for IHF, where DNA executes a U-turn (>160°) as it wraps around the protein (Rice et al., 1996). The cluster has the largest average protein-DNA interface area, ΔASA, (5153 Å^2). So, the larger conformational changes constrain DNA afar from both the B and A forms and cause the formation of a huge protein-DNA interface area. The average energy Z-score value for the direct readout is −1.44, while that value for indirect readout is −1.49. These values imply a role for both the direct readout and the indirect conformational effects of DNA in protein-DNA recognition. This is supported by the significant number of specific contacts at the major and minor grooves, M-cont and m-cont, of 26.7 and 9.1, respectively. There is a predominant mode of β ribbon-DNA interactions in the complexes of I-PpoI, Tus, IHF, and Arc repressor. The HTH-DNA interactions occur at the major groove in the complexes of DtxR, γδ resolvase, and MATα2. The lower average GC content value of 34.9% indicates a TA richness of the DNA binding sequence. Particularly, the AT binding site preference is implicated for endonuclease activity of I-PpoI (Flick et al., 1998) and DtxR-specific operator (Pohl et al., 1999). Most of the group members show slightly widened major and minor grooves, and the calculated average widths of major and minor grooves are 12.09 and 7.52 Å, respectively, for this cluster (11.69 and 6.9 Å are the average values for the 62 complexes in the data set).

Cluster 3

Of the 12 members in the cluster, 5 of them are classical Zn-fingers—Zif268 (1AAY), Gal4 (1D66), pyrimidine pathway regulator 1 (1PYI), tramtrack protein (2DRP), and five-finger Gli (2GLI). The other seven complexes are STAT-1 (1BF5), MunI restriction endonuclease (1D02), GCC-box binding protein (1GCC), max (1HLO), c-Myb (1MSE), PvuII endonuclease (1PVI), and basic leucine zipper GCN4 (1YSA). The cluster is uniquely classified by the highest GC content, 60.7%, and a considerably widened major groove width of 12.09 Å. High prevalence of the G·C sequence in these complexes may favor retaining the B-DNA conformation and is reflected by the lowest average $rmsd_{cont}$-B value, 2.24 Å, among the seven clusters. We find that the common properties of such a widened major groove, GC-rich site and B-DNA-like conformation, clearly group all of the Zn-fingers and the related complexes into this cluster. The finding that Zn-fingers and other complexes, such as the leucine zipper in the cluster, share the B-DNA conformation, containing a widened major groove, is reminiscent of an earlier study on the Zn-finger-DNA and some protein-DNA complexes that revealed the B_{eg}-DNA structures ("eg" stands for enlarged groove) (Nekludova and Pabo, 1994). The present cluster also has a relatively high number of major groove contacts (M-cont of 24.7) and the lowest number of minor groove contacts (m-cont of 0.8). The other characteristic parameter is TCD, with the lowest average value (0.41). The average interface area, ΔASA of 2562 Å^2, is also smaller for the cluster. The average energy Z-score values for the direct and indirect readouts in the cluster are −0.89 and −0.88, respectively.

Cluster 4

This smallest cluster has only three members: Sac7d (1AZQ), TATA-box binding protein (1CDW), and hSRY-HMG (1HRY). The cluster has the largest average minor groove width, 11.7 Å, which is almost the width of the major groove (11.8 Å), and it concomitantly has the lowest $rmsd_{cont}$-A value, 2.5 Å, among the seven clusters. These indicate a close A form DNA conformation for the complexes. The high number of minor groove contacts, 12.7, and the fewest average major groove contacts, 6.0, in the cluster strongly group these proteins into minor groove binders (Bewley et al., 1998). It is worth noting that members of the present cluster have the lowest average protein-DNA interface area, ΔASA (2209 Å^2), whereas the other minor groove binding proteins in cluster 2, such as IHF and γδ resolvase, have a larger average ΔASA value of >5500 Å^2.

Cluster 5

There are six cluster members that are related to homeodomain proteins. The members are paired (PaX) class homeodomain (1FJL); Hin recombinase (1HCR); fungal transcription factor HAP1 heme activator protein (1HWT); Oct-1 POU (1OCT); purine repressor, PurR (1PNR); and human serum response factor, which is a transcription factor belonging to the MADS domain protein family (1SRS). Topology analysis of proteins in the data set as performed by DALI (Holm and Sander, 1993) reveals the structural similarities between the paired class cooperative homeodomain and the POU domain. The overall structure of Hin recombinase resembles both a prototypical bacterial HTH and the eukaryotic homeodomain, and, in many respects, it is an intermediate between these two DNA binding motifs (Feng et al., 1994). Interestingly, HAP1 has a binuclear zinc cluster domain with a sequence of Arg-Lys-Arg-N-Arg in the N-terminal arm that shows homology to the N-terminal arms of homeodomain proteins interacting

Table 3. The Calculated Conservation and Intracluster Distances of Cluster Members for the *k*-Means Model

PDB	N	D
Cluster 1		
1A02	57	0.168
1BER	59	0.219
1BHM	54	0.223
1IF1	55	0.116
1PER	62	0.086
1TSR	40	0.213
1XBR	51	0.213
2HDC	56	0.262
3CRO	54	0.173
AVER	54.2	0.186
STDEV	6.2	0.056
Cluster 2		
1A74	28	0.278
1C0W	28	0.200
1ECR	28	0.197
1GDT	62	0.146
1IHF	60	0.485
1MNM	56	0.328
1PNR	28	0.269
AVER	41.4	0.272
STDEV	16.8	0.112
Cluster 3		
1AAY	53	0.215
1BF5	40	0.202
1D02	25	0.202
1D66	62	0.144
1GCC	44	0.267
1HLO	50	0.194
1MSE	24	0.175
1PVI	21	0.443
1PYI	47	0.202
1YSA	49	0.234
2DRP	42	0.212
2GLI	30	0.159
AVER	40.6	0.221
STDEV	12.9	0.077
Cluster 4		
1AZQ	60	0.248
1CDW	62	0.168
1HRY	60	0.210
AVER	60.7	0.209
STDEV	1.2	0.040
Cluster 5		
1FJL	62	0.186
1HCR	52	0.203
1HWT	56	0.272
1OCT	31	0.200
1PNR	43	0.260
1SRS	60	0.195
AVER	50.7	0.219
STDEV	11.8	0.037
Cluster 6		
1BL0	53	0.171
1CF7	59	0.182
1HCQ	57	0.179
1IGN	36	0.327
1LMB	57	0.178
1MEY	56	0.291
1PUE	54	0.206
1REP	62	0.134
1SVC	56	0.210
1TF3	56	0.144
1UBD	52	0.181
1YRN	51	0.290
AVER	54.1	0.208
STDEV	6.4	0.062
Cluster 7		
1B3T	20	0.271
1CJG	34	0.148
1CMA	62	0.123
1DP7	20	0.212
1GAT	59	0.194
1HDD	52	0.149
1MDY	59	0.173
1MHD	58	0.156
1PDN	51	0.203
1TC3	58	0.179
1TRO	54	0.202
1YUI	48	0.275
2BOP	27	0.286
AVER	46.3	0.198
STDEV	15.5	0.052

PDB, Protein Data Bank code; N, the number of occurrences of a member in a particular cluster as calculated by the jack-knife procedure (N_{max} = 62 for total conservation); D, the distance between a member and the centroid of the corresponding cluster, $(\Sigma_i[x_i - x]^2)^{1/2}$; AVER and STDEV denote the average and standard deviation values, respectively, for each cluster. The PDB code in bold face denotes the exemplar for each cluster.

with a TA-rich DNA minor groove (King et al., 1999). The similar minor groove interactions were also noted at the carboxyl-terminal peptide of the Hin domain (Feng et al., 1994), the hinge helices of PurR (Schumacher et al., 1994), and serum response factor at TA-rich sites (Pellegrini et al., 1995). Therefore, these complexes with common functional interactions are grouped together, and the common properties are reflected by the highest profile value for the number of minor groove contacts, m-cont of 15.7, among the seven clusters and the lower GC content of 34.8% (Table 3), which represents the TA-rich binding sites. Four of the homeodomain-related proteins have an HTH motif and, combined with the other two, make a significant number of major groove interactions, M-cont of 17.5. The average DNA conformation in the cluster is closer to B-DNA, with an average rms deviation value ($rmsd_{cont}$-B) of 2.93 Å. Further, this group has the average energy Z-score values of −2.07 and −1.17 for the direct and indirect readouts, respectively. The experimental observations indicate that these types of homeodomain proteins can specifically bind to DNA not only through direct contact with the major and minor grooves, but also through indirect effects (Wolberger, 1996). The smaller TCD value of 0.45 indicates the compactness of the DNA binding proteins profiling the present cluster.

Cluster 6

The cluster has 12 members. Of these, six complexes have HTH motifs, four have Zn-finger motifs, and other two have a winged helix and a β ribbon. Most of these complexes in the cluster exclusively function as transcription factors (Muller, 2001). Out of the 12 group members, 9 of them are involved in different transcriptional function. These include: transcriptional activator MarA (1BL0), cell cycle transcription factor E2F-DP

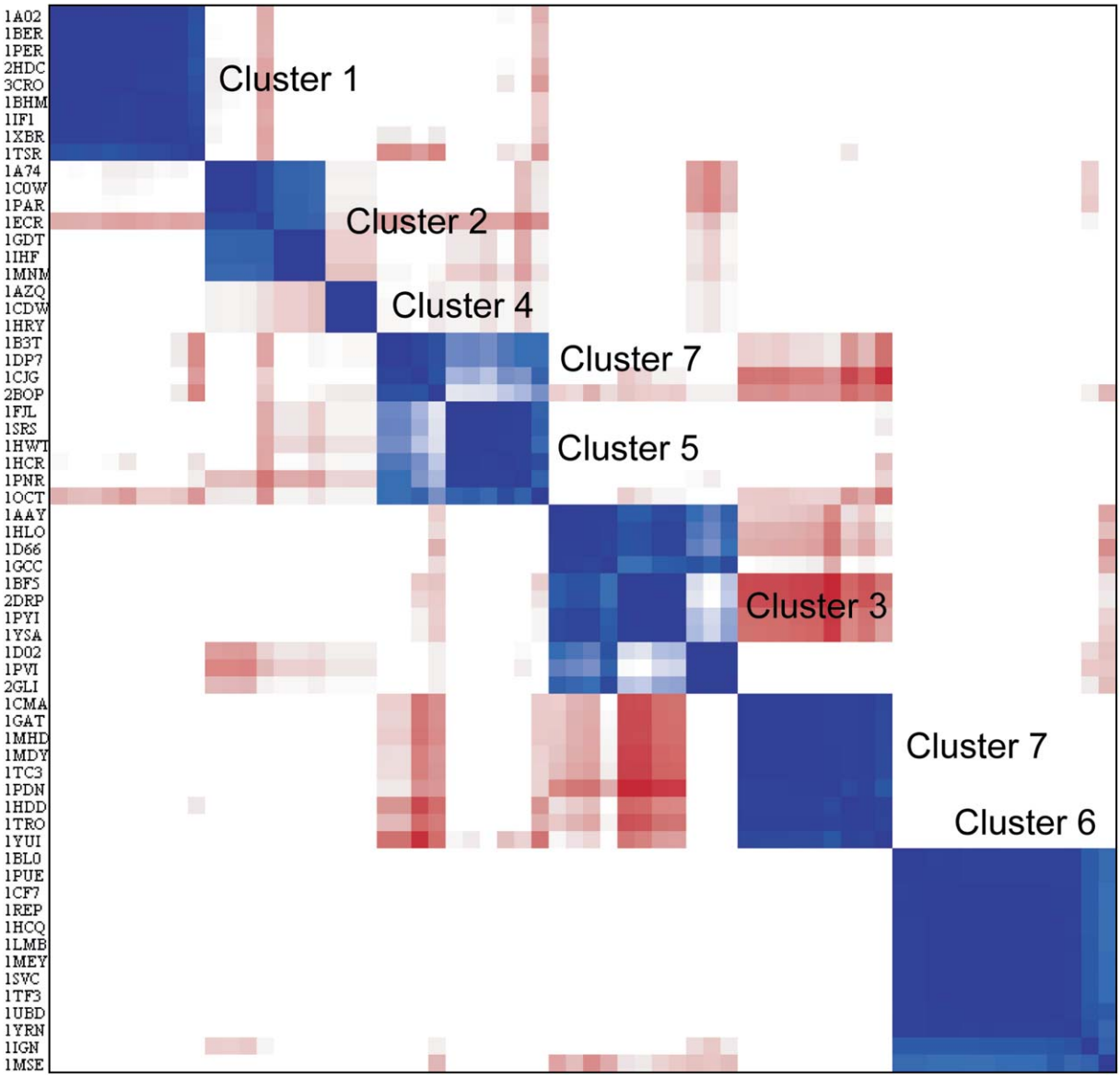

Figure 3. The Proximity Matrix Diagram for Conservation of Cluster Memberships in Seven Clusters

Conservation within clusters is shown in blue, and conservation across clusters is shown in red. The depth of shading indicates the level of conservation.

(1CF7), estrogen receptor from ligand-activated transcription factor family (1HCQ), transcriptional regulator Rap1 (1IGN), PU.1 ETS domain transcription factor (1PUE), replication initiator protein (1REP), transcription factor NF-κB (1SVC), TFIIIA (1TF3), and transcription initiator YY1 (1UBD). The other three complexes are λ repressor (1LMB), Cys_2His_2 consensus Zn-finger (1MEY), and ternary MATa1/α2 homeodomain complex (1YRN). The average values of the major and minor grooves are 12.06 and 5.94 Å, respectively, and those respective average values are 11.7 and 6.9 Å for the 62 complexes in the data set. The cluster has the highest profile value, 34.2, for the major groove contacts (M-cont), where HTH or Zn-finger normally attacks the DNA, and the lower average value, 2.2, for the minor groove contacts (m-cont). The average value of GC content is 54.2%, and this relatively high occurrence of G·C bases favors the B-DNA form, as the average $rmsd_{cont}$-B value is 2.89 Å compared to 6.06 Å for $rmsd_{cont}$-A. The average energy Z-score values for the direct and indirect readouts for

Table 4. Intercluster Distances between Cluster Centroids in the *k*-Means Clustering Model

Cluster	1	2	3	4	5	6	7
1	0	1.205	0.966	2.476	0.76	0.635	0.631
2		0	1.921	2.756	0.986	1.438	1.34
3			0	1.823	1.119	0.638	0.519
4				0	2.002	2.785	1.441
5					0	0.962	0.645
6						0	0.622
7							0

Table 5. The Cluster Profile of the *k*-Means Analysis Is Shown

Cluster	M-cont	m-cont	TCD	$Rmsd_{cont}$-B	$Rmsd_{cont}$-A	mgw	Mgw	GC%	PD-Z	DNA-Z	ΔASA
1	18.8 (8.3)	1.7 (2.7)	**0.84 (0.21)**	3.33 (1.49)	6.26 (0.80)	**5.40 (1.05)**	**12.14 (1.07)**	**31.5 (8.0)**	−1.94 (0.94)	−1.16 (0.77)	3380 (827)
2	26.7 (13.4)	9.1 (8.4)	0.45 (0.17)	**7.52** (3.96)	**7.82** (1.91)	7.52 (0.79)	12.09 (1.23)	34.9 (4.1)	−1.44 (1.54)	−1.49 (1.15)	**5153** (819)
3	24.7 (15.0)	**0.8** (2.0)	**0.41** (0.18)	**2.24** (0.90)	4.47 (0.98)	6.83 (1.13)	12.09 (1.10)	**60.7** (12.2)	**-0.89** (1.58)	−0.88 (1.11)	2562 (752)
4	**6.0** (10.4)	12.7 (7.8)	0.66 (0.19)	4.87 (0.64)	**2.46** (0.58)	**11.72** (1.21)	11.85 (1.94)	39.6 (13.0)	−1.10 (1.05)	**-0.70** (0.13)	**2209** (710)
5	17.5 (14.6)	**15.7** (8.1)	0.45 (0.09)	2.93 (1.12)	6.54 (1.25)	6.34 (0.97)	11.08 (0.55)	34.8 (9.8)	−2.07 (1.18)	−1.17 (1.44)	3315 (534)
6	**34.2** (19.0)	2.2 (4.1)	0.59 (0.21)	2.89 (1.09)	6.06 (0.99)	5.94 (0.79)	12.06 (0.73)	54.2 (8.65)	**-2.58** (1.39)	**-2.73** (0.69)	2997 (672)
7	14.5 (7.2)	3.0 (3.0)	0.61 (0.29)	3.10 (0.86)	5.70 (0.96)	7.72 (0.69)	**10.74** (0.78)	45.6 (14.2)	−0.93 (0.85)	−2.41 (0.94)	2802 (966)

Each number refers to the average value of the pertinent descriptor for the corresponding cluster. The standard deviation values (STDEV) are given in parentheses. The bold-faced numbers show the maximum and minimum values for the 11 descriptors among the 7 clusters. The units for descriptors are as mentioned in Table 2.

this cluster are −2.58 and −2.73, respectively, which are the highest among the seven clusters. The transcription factors in the cluster utilize both readout mechanisms to perform the recognition of the specific target sites.

Cluster 7

This is the biggest cluster among the 7 and has 13 members. The common structural features likely shared by all of the cluster members are (1) a highly compressed major groove, as evident from the lowest average value, 10.74 Å, for major groove width (Mgw); (2) a relatively wide minor groove with the average width (mgw) value of 7.72 Å; and (3) a smaller average protein-DNA interface area (ΔASA) of 2802 $Å^2$. Six of the cluster members use an HTH-motif including lac repressor (1CJG), engrailed homeodomain (1HDD), MyoD (1MDY), paired domain (1PDN), Tc3 transposase (1TC3), and trp repressor (1TRO). Two complexes, erythroid transcription factor Gata-1 (1GAT) and GAGA factor (1YUI), have different Zn domains. The DNA binding modes and orientation of the individual fingers for GATA-1 and GAGA factor in this cluster are quite different from those of Zif268 and other Zn-fingers (Omichinski et al., 1992, 1997) grouped in Cluster 3. Out of five other members, three use a β hairpin for DNA interactions, and these are the complexes with met repressor (1CMA), Smad MH1 (1MHD), and bovine papilloma virus E2 protein (2BOP). Two complexes with Epstein-Barr virus nuclear antigen 1 EBNA1 (1B3T) and regulatory factor RFX1 (1DP7) form α helix-DNA interactions. Among the members of this cluster, we intriguingly find a subgroup that is mainly related to viral function, namely, EBNA1 virus nuclear protein, which activates DNA replication from oriP, the latent origin of DNA replication in Epstein-Barr virus; bovine papilloma virus E2 protein, which is structurally homologous to the core domain of EBNA1 (Bochkarev et al., 1996); and regulatory factor RFX1 (1DP7), which is a transactivator of human hepatitis B virus enhancer I. Many of the proteins in the cluster specifically use β hairpins or loop fragments at the end of helices or strands, and they require compressed major grooves to fit snugly and to make specific contacts. The average energy Z-score values for the direct and indirect readouts in the cluster are −0.93 and −2.41, respectively, suggesting that the proteins largely recognize DNA via an indirect manner. The conclusion about such a dominant, indirect readout effect has been experimentally verified for E2 papillomavirus complex with a series of DNA binding sites by using the cyclization kinetics method (Zhang et al., 2004).

Discussion

We have applied the cluster analysis to protein-DNA interactions by using several important structural parameters as descriptors. The cluster analysis allows us to consider an arbitrary number of equally weighted parameters for the purpose of detecting relevant subgroup clusters and identifying common properties and relationships. We identify 7 clusters in the data set of 62 protein-DNA complexes by using hierarchical and *k*-means analyses. The jackknifing approach validates the clustering results and finds that the cluster memberships are conserved over 80%. The individual clusters have characteristic structural properties in common, implying that the modes of protein-DNA recognition are distinct among different clusters. We find some general trends: for example, homeodomains are associated with TA-rich sequences and the highest profile value for the number of minor groove contacts; Zn-fingers involve GC-rich binding sites with an enlarged B-DNA conformation; and transcription factor complexes with HTH motifs have widened major grooves, whereas viral and other disease-related proteins prefer compressed major grooves in which β hairpin motifs bind. Also, minor groove binders are grouped into two major classes—one has close A-DNA, and the other has highly deformed conformations that have distinctly smaller and larger interface areas respectively. Thus, the cluster analysis takes into account a number of key parameters related to protein-DNA interactions and enables us to mine the data for patterns and relationships in the different groups. These findings observed in different clusters would certainly be hidden for other statistical analyses, which might use a few parameters and focus on the full data set.

DNA binding proteins are usually classified according to the structural motif of proteins. However, the conservation of structural motifs in proteins may not necessarily dictate the way by which DNA sequences are recognized. For example, there are many members in homeodomain, Zn-finger, and HTH families of proteins, but their target DNA sequences, the way by which these target sequences are recognized, and their biological functions are significantly diverse among the members. Thus, it would be more appropriate to classify the DNA binding proteins based on distinct structural descriptors characterizing protein-DNA recognition rather than the properties of proteins alone. 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.

We have started with 22 parameters to describe the characteristics of protein-DNA complex structures, and we reduced this number by half by removing redundancy. The choice of the descriptors is subjective and somewhat arbitrary; thus, it may affect the result of cluster analysis. However, we expect that the inclusion of more parameters would not change the results drastically, since the number of independent parameters would be limited. We have also considered 62 protein-DNA complex structures, but inclusion of more structures may change the present results. This is related to the number of representative DNA binding proteins that can exist in the structurome. It is likely that many DNA binding proteins would be classified into the clusters we identified in the present study, as the present data set covers a wide spectrum of known DNA binding proteins. However, new clusters will emerge as a greater number of unique protein-DNA complexes become available, and these clusters will also refine the rules and improve prediction of protein-DNA binding. We believe that the present approach based on the cluster analysis that uses structural descriptors characterizing protein-DNA interactions provides a potential framework for identifying common properties and rules that govern protein-DNA recognition.

Experimental Procedures

Data Set of Protein-DNA Complexes

We have considered the same set of 62 nonredundant protein-DNA complexes that we used in our previous studies (Gromiha et al., 2004b; Ahmad and Sarai, 2004). Briefly, this data set was created as follows: we constructed a data set containing the structures of 52 protein-DNA complexes solved by X-ray and NMR methods for an earlier study (Kono and Sarai, 1999). For the present study, we initially excluded these 52 structures as well as those for which the resolution data was poor (> 3.2Å), from the protein-DNA complexes available in the Protein Data Bank (PDB) (Deshpande et al., 2005). The remaining set of complexes was individually screened against the previously constructed set of 52 complexes by using FASTA (Pearson and Lipman, 1988). Each complex with less than 25% identity was added to the original 52, eventually resulting in a unique set of 62 protein-DNA complexes.

Descriptors Used for Cluster Analysis

For each protein-DNA complex, we calculated the following parameters by using atomic coordinates from the PDB (Deshpande et al., 2005).

Number of Contacts: Descriptors 1–4

The total number of hydrogen bonds between DNA and protein with the donor-acceptor distance cutoff value of 3.5 Å was calculated with and without backbone atoms. The contacts were further classified into DNA major (M-cont) and minor (m-cont) groove contacts.

Total Contact Distance, TCD: 5

A statistical parameter, described by Zhou and Zhou (2002), involving interresidue contacts within the protein was developed for predicting the folding rate and is used in this study. The TCD parameter represents the average sequence separation by contacting residues within a cutoff distance, is calculated by using protein structural information, and is successfully correlated with the experimentally observed logarithmic values of protein folding rates.

Rmsd-B and Rmsd-A: 6–11

Rmsd-B and rmsd-A are the root-mean-square deviation (rmsd) values calculated by fitting the bound DNA structure to the canonical B and A form DNA structures, respectively. Rmsd calculations were done in three separate ways: by fitting the P atoms ($rmsd_P$-B and $rmsd_P$-A), by fitting all atoms of the backbone ($rmsd_{bb}$-B and $rmsd_{bb}$-A), and by fitting contact atoms ($rmsd_{cont}$-B and $rmsd_{cont}$-A). The contact atoms are defined as the DNA atoms that make contacts to protein within a distance of 3.5 Å in the complex. The canonical B-DNA and A-DNA structures with the same sequence as that of bound DNA were generated by using the biopolymer module of SYBYL6.9 software (Tripos Inc., USA). The ProFit program was used to perform the rmsd calculations, and the superimposition procedure, performed by fitting the atoms in ProFit, was carried out by using the McLachlan algorithm (McLachlan, 1982).

Widths and Depths of DNA Major and Minor Grooves: 12–15

The widths and depths of major (Mgw, Mgd) and minor (mgw, mgd) grooves of the bound DNA were calculated by using the program Curves 5.1 (Lavery and Sklenar, 1989), in which phosphorous atoms were chosen as the reference atoms for measuring the groove width.

Bending: 16

The total bend along the helical axis of the DNA was calculated by using Curves 5.1.

GC Content: 17

The GC content is the percentage value of GC in complexed DNA.

Z-Score, PD-Z: 18–20

Energy Z-score values were derived from the statistical interaction potentials by using protein-DNA complex structures (PD-Z) (Kono and Sarai, 1999; Selvaraj et al., 2002); this procedure quantifies the specificity of direct readout of protein-DNA recognition. The energy Z-score calculations were also performed by replacing the bound DNA with the canonical B-DNA and A-DNA structures with the same sequence, PD_A-Z and PD_B-Z, respectively.

Z-Score, DNA-Z: 21

Z-score values were derived from the sequence-dependent energy function of DNA conformation (Gromiha et al., 2004b), which quantifies the specificity of indirect readout of protein-DNArecognition.

ΔASA: 22

The protein-DNA interface area was calculated by subtracting the solvent-accessible surface area (ASA) of the protein-DNA complex from the sum of the ASA values of the separated protein and DNA molecules. ASA values were calculated by using the web-based program GETAREA1.1 (Fraczkiewicz and Braun, 1998).

A correlation analysis with all of the aforementioned 22 parameters was performed in order to get rid of highly correlated ones. By setting a threshold value of $r = 0.40$, we got the final set of 11 parameters that we chose for performing the clustering analysis, namely, M-cont, m-cont, TCD, $rmsd_{cont}$-B, $rmsd_{cont}$-A, Mgw, mgw, GC%, Z-score (PD-Z), Z-score (DNA-Z), and ΔASA. Even though $rmsd_{cont}$-B, $rmsd_{cont}$-A, and ΔASA showed mutual correlations among themselves, with the "*r*" values ranging from 0.40 to 0.53, we decided to include them, as their values represent independent properties. We combined all of the parameters with equal weight for the cluster analysis in order to analyze their roles within the different groups of protein-DNA complexes.

Clustering Methods and Distance Measures

Clustering normally refers to the identification of groups within a data set such that similarities within the groups are significantly larger than those between groups (Kaufman and Rousseeuw, 1990). Clustering algorithms can be broadly divided into two types: hierarchical and partitioning (Han and Kamber, 2001). Hierarchical clustering techniques proceed as either a series of successive mergers or successive divisions (i.e., they are either agglomerative or divisive). The agglomerative hierarchical approach starts with the individual objects. The most similar objects are grouped first, after which the groups are successively merged based on their similarities. The merging process continues until all groups have been joined into a single cluster. At each stage of the clustering process, the program computes a distance measure of choice, called the fusion coefficient, which indexes the relative distance between the two objects that were linked at each stage. Eventually, all subgroups are fused into a single cluster. The divisive hierarchical method works in the opposite direction, starting with all of the objects in

the same cluster. It then successively divides them up until each object forms a group. Among the partitioning techniques, one uses the *k*-means algorithm, which requires prior specification of the number (*k*) of groups. The algorithm then uses the input parameter, *k*, and randomly partitions a set of objects into *k* clusters. The objects are then continually reassigned between the clusters to optimize an objective partitioning criterion, often called a similarity function, until the clusters are as compact and separate as possible.

Among the available clustering algorithms, which one to use depends both on the type of data to be analyzed and the purpose of the clustering. Also, as cluster analysis is always used as a descriptive or exploratory tool, it is permissible to employ several algorithms with the same data set in order to get the best possible picture of its underlying structure. Since our data set contained variables with different units and ranges of values, we first performed a Z-score transformation that standardized the data. We then used ClustanGraphics5, a cluster analysis software and graphics tool (Wishart, 1999). The initial step in a cluster analysis is to set up a proximity matrix, which we constructed by computing squared Euclidean distances between all pairs of the 62 protein-DNA complexes. We then used the Increase in Sum of Squares method (also called Ward's hierarchical method) as a linkage technique. This method tries to minimize the sum of squares of distances between any two hypothetical clusters formed at each step, and it tends to create clusters of a small size. A hierarchical clustering tree was then constructed by using ClustanGraphics.

Acknowledgments

P.P. thanks the Science and Technology Agency (STA) of Japan for a postdoctoral fellowship award. This work is supported in part by Grants-in-Aid for Scientific Research 16014219 (A.S.) from the Ministry of Education, Culture, Sports, Science and Technology in Japan. Part of the work was done at the RIKEN Institute. We thank the RIKEN Institute for that support. We thank Drs. H. Kono and S. Selvaraj for the information about direct Z-score calculations.

Received: May 31, 2006
Revised: June 27, 2006
Accepted: June 27, 2006
Published: September 12, 2006

References

Ahmad, S., and Sarai, A. (2004). Moment-based prediction of DNA-binding proteins. J. Mol. Biol. *341*, 65–71.

Ahmad, S., Gromiha, M.M., and Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics *20*, 477–486.

Bewley, C.A., Gronenborn, A.M., and Clore, G.M. (1998). Minor groove-binding architectural proteins: structure, function, and DNA recognition. Annu. Rev. Biophys. Biomol. Struct. *27*, 105–131.

Bochkarev, A., Barwell, J.A., Pfuetzner, R.A., Bochkareva, E., Frappier, L., and Edwards, A.M. (1996). Crystal structure of the DNA-binding domain of the Epstein-Barr virus origin-binding protein, EBNA1, bound to DNA. Cell *84*, 791–800.

Cho, Y., Gorina, S., Jeffrey, P.D., and Pavletich, N.P. (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science *265*, 346–355.

Deshpande, N., Addess, K.J., Bluhm, W.F., Merino-Ott, J.C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z., et al. (2005). The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. Nucleic Acids Res. *33*, D233–D237.

Dickerson, R.E. (1998). DNA bending: the prevalence of kinkiness and the virtues of normality. Nucleic Acids Res. *26*, 1906–1926.

Escalante, C.R., Yie, J., Thanos, D., and Aggarwal, A.K. (1998). Structure of IRF-1 with bound DNA reveals determinants of interferon regulation. Nature *391*, 103–106.

Feng, J.A., Johnson, R.C., and Dickerson, R.E. (1994). Hin recombinase bound to DNA: the origin of specificity in major and minor groove interactions. Science *263*, 348–355.

Flick, K.E., Jurica, M.S., Monnat, R.J., Jr., and Stoddard, B.L. (1998). DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI. Nature *394*, 96–101.

Fraczkiewicz, R., and Braun, W. (1998). Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J. Comp. Chem. *19*, 319–333.

Gromiha, M.M., Santhosh, C., and Ahmad, S. (2004a). Structural analysis of cation-pi interactions in DNA binding proteins. Int. J. Biol. Macromol. *34*, 203–211.

Gromiha, M.M., Siebers, J.G., Selvaraj, S., Kono, H., and Sarai, A. (2004b). Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. J. Mol. Biol. *337*, 285–294.

Han, J., and Kamber, M. (2001). Data Mining: Concepts and Techniques (San Francisco: Morgan Kaufmann Publishers).

Holm, L., and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. J. Mol. Biol. *233*, 123–138.

Jones, S., van Heyningen, P., Berman, H.M., and Thornton, J.M. (1999). Protein-DNA interactions: A structural analysis. J. Mol. Biol. *287*, 877–896.

Kaufman, L., and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis (New York: John Wiley & Sons, Inc.).

King, D.A., Zhang, L., Guarente, L., and Marmorstein, R. (1999). Structure of a HAP1-DNA complex reveals dramatically asymmetric DNA binding by a homodimeric protein. Nat. Struct. Biol. *6*, 64–71.

Kono, H., and Sarai, A. (1999). Structure-based prediction of DNA target sites by regulatory proteins. Proteins *35*, 114–131.

Lavery, R., and Sklenar, H. (1989). Defining the structure of irregular nucleic acids: conventions and principles. J. Biomol. Struct. Dyn. *6*, 655–667.

Luisi, B. (1995). DNA-protein interaction at high resolution. In DNA-Protein: Structural Interactions, D.M.J. Lilley, ed. (Oxford, UK: IRL Press), pp. 1–48.

Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. (2000). An overview of the structures of protein-DNA complexes. Genome Biol. *1*, REVIEWS001.

Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. *29*, 2860–2874.

Mandel-Gutfreund, Y., Schueler, O., and Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. J. Mol. Biol. *253*, 370–382.

Mandel-Gutfreund, Y., Margalit, H., Jernigan, R.L., and Zhurkin, V.B. (1998). A role for CH…O interactions in protein-DNA recognition. J. Mol. Biol. *277*, 1129–1140.

McLachlan, A.D. (1982). Rapid comparison of protein structures. Acta Crystallogr. A *38*, 871–873.

Mojena, R., and Wishart, D. (1980). Hierarchical grouping methods and stopping rules: an evaluation. In COMPSTAT 1980 Proceedings (Vienna: Physica-Verlag), pp. 426–432.

Muller, C.W. (2001). Transcription factors: global and detailed views. Curr. Opin. Struct. Biol. *11*, 26–32.

Nadassy, K., Wodak, S.J., and Janin, J. (1999). Structural features of protein-nucleic acid recognition sites. Biochemistry *38*, 1999–2017.

Nekludova, L., and Pabo, C.O. (1994). Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes. Proc. Natl. Acad. Sci. USA *91*, 6948–6952.

Omichinski, J.G., Clore, G.M., Robien, M., Sakaguchi, K., Appella, E., and Gronenborn, A.M. (1992). High-resolution solution structure of the double Cys2His2 zinc finger from the human enhancer binding protein MBP-1. Biochemistry *31*, 3907–3917.

Omichinski, J.G., Pedone, P.V., Felsenfeld, G., Gronenborn, A.M., and Clore, G.M. (1997). The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode. Nat. Struct. Biol. *4*, 122–132.

Pabo, C.O., and Nekludova, L. (2000). Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J. Mol. Biol. *301*, 597–624.

Pabo, C.O., and Sauer, R.T. (1992). Transcription factors: structural families and principles of DNA recognition. Annu. Rev. Biochem. *61*, 1053–1095.

Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA *85*, 2444–2448.

Pellegrini, L., Tan, S., and Richmond, T.J. (1995). Structure of serum response factor core bound to DNA. Nature *376*, 490–498.

Pohl, E., Holmes, R.K., and Hol, W.G. (1999). Crystal structure of a cobalt-activated diphtheria toxin repressor-DNA complex reveals a metal-binding SH3-like domain. J. Mol. Biol. *292*, 653–667.

Reddy, C.K., Das, A., and Jayaram, B. (2001). Do water molecules mediate protein-DNA recognition? J. Mol. Biol. *314*, 619–632.

Rice, P.A., Yang, S., Mizuuchi, K., and Nash, H.A. (1996). Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. Cell *87*, 1295–1306.

Rooman, M., Lievin, J., Buisine, E., and Wintjens, R. (2002). Cation-pi/H-bond stair motifs at protein-DNA interfaces. J. Mol. Biol. *319*, 67–76.

Sarai, A., and Takeda, Y. (1989). Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. Proc. Natl. Acad. Sci. USA *86*, 6513–6517.

Schumacher, M.A., Choi, K.Y., Zalkin, H., and Brennan, R.G. (1994). Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. Science *266*, 763–770.

Selvaraj, S., Kono, H., and Sarai, A. (2002). Specificity of protein-DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. J. Mol. Biol. *322*, 907–915.

Steitz, T.A. (1993). Structural Studies of Protein-Nucleic Acid Interaction (Cambridge, UK: Cambridge University Press).

Wintjens, R., Lievin, J., Rooman, M., and Buisine, E. (2000). Contribution of cation-pi interactions to the stability of protein-DNA complexes. J. Mol. Biol. *302*, 395–410.

Wishart, D. (1999). ClustanGraphics Primer: A Guide to Cluster Analysis (http://www.clustan.com/).

Wolberger, C. (1996). Homeodomain interactions. Curr. Opin. Struct. Biol. *6*, 62–68.

Zhang, Y., Xi, Z., Hegde, R.S., Shakked, Z., and Crothers, D.M. (2004). Predicting indirect readout effects in protein-DNA interactions. Proc. Natl. Acad. Sci. USA *101*, 8337–8341.

Zhou, H., and Zhou, Y. (2002). Folding rate prediction using total contact distance. Biophys. J. *82*, 458–463.