



Building an automated classification of DNA-binding protein domains

Julia V. Ponomarenko¹, Philip E. Bourne² and Ilya N. Shindyalov^{3,*}

¹Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 10 Lavrentyeva Ave., Novosibirsk, 630090, Russia, ²Department of Pharmacology and San Diego Supercomputer Center and ³San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, CA 92093-0505, La Jolla, USA

Received on April 8, 2002; accepted on June 15, 2002

ABSTRACT

Intensive growth in 3D structure data on DNA-protein complexes as reflected in the Protein Data Bank (PDB) demands new approaches to the annotation and characterization of these data and will lead to a new understanding of critical biological processes involving these data. These data and those from other protein structure classifications will become increasingly important for the modeling of complete proteomes. We propose a fully automated classification of DNA-binding protein domains based on existing 3D-structures from the PDB. The classification, by domain, relies on the Protein Domain Parser (PDP) and the Combinatorial Extension (CE) algorithm for structural alignment. The approach involves the analysis of 3D-interaction patterns in DNA-protein interfaces, assignment of structural domains interacting with DNA, clustering of domains based on structural similarity and DNA-interacting patterns. Comparison with existing resources on describing structural and functional classifications of DNA-binding proteins was used to validate and improve the approach proposed here. In the course of our study we defined a set of criteria and heuristics allowing us to automatically build a biologically meaningful classification and define classes of functionally related protein domains. It was shown that taking into consideration interactions between protein domains and DNA considerably improves the classification accuracy. Our approach provides a high-throughput and up-to-date annotation of DNA-binding protein families which can be found at <http://spdc.sdsc.edu>.

Contact: shindyal@sdsc.edu

INTRODUCTION

The structural analysis of 240 DNA-protein complexes contained in the Protein Data Bank (PDB) was performed

by Luscombe *et al.* (2000). These complexes were classified into eight different structural/functional groups containing a total of 54 structural families. Initial assignment of proteins into eight groups was done manually followed by detailed classification into structural families. The SSAP method of structural alignment (Orengo and Taylor, 1996) was used to assist in proper classification of individual proteins. Even though this classification can be considered a standard, it is now far behind the number of DNA-protein complexes available in the PDB. This implies that the rapid growth of 3D DNA-binding protein structures requires a fully automated approach to their classification as well as faster methodologies than SSAP which uses double dynamic programming and is computationally expensive for many pairwise comparisons.

There are other structural classifications, which are not specifically oriented towards proteins interacting with DNA, but cover all proteins in the PDB. The Structural Classification of Proteins (SCOP) (Murzin *et al.*, 1995; Lo Conte *et al.*, 2000) is one of the most often used. It provides a four level classification of protein domains based on structure and sequence homology. These levels are (i) class, (ii) fold, (iii) superfamily, (iv) family. SCOP combines automated and manual annotation involving human experts. The Dali Domain Dictionary is another classification of protein folds (Dietmann and Holm, 2001; Dietmann *et al.*, 2001). It is automatically built using the Dali algorithm and the concept of attractor regions. They also classify domains into a hierarchy of four levels as follows: (i) supersecondary structural motifs, (ii) topology of the fold, (iii) functional family, (iv) sequence family. Two further structural classifications are CATH (Pearl *et al.*, 2000) and 3Dee (Dengler *et al.*, 2001; Siddiqui *et al.*, 2001). They both combine manual and automated approaches. For clarity we limit our analysis to a comparison of SCOP and the Dali Domain Definition.

*To whom correspondence should be addressed.

These two definitions overlap with CATH and 3Dee in approximately 80% of cases.

Our analysis has shown that fully automated classification approaches do not achieve the same quality of annotation, from a biological standpoint, as manual or combined (manual and automated) approaches. On the other hand, as stated, approaches with a manual component have obvious shortcomings such as: (i) significant backlogs relative to the currently available 3D data; (ii) don't provide a uniform or reproducible classification since a human component is involved; (iii) depend on a highly skilled human expert who is expensive and limited. Thus there is a need for fully automated approaches that are capable of achieving a human expert level of quality. In this work we are aiming to solve this problem on a subset of structure data, namely DNA-binding protein domains. Our approach is based on the analysis of structural similarity and protein-DNA interaction patterns. Our analysis is performed at the level of putative functional domains, which are detected as part of this study. To succeed structural information must be combined with information about protein-DNA interactions. This combination provides significant improvement in the classification of DNA-binding protein domains and indeed approaches the level of quality achievable by a human expert.

Our structural classification was built considering three structural classifications of DNA-binding domains introduced above: (i) classification of DNA-binding protein chains by Luscombe *et al.* (2000), (ii) SCOP classification (Murzin *et al.*, 1995; Lo Conte *et al.*, 2000) and (iii) Dali domain Dictionary classification of protein folds (Dietmann and Holm, 2001; Dietmann *et al.*, 2001). These classifications provided biological reasons (different sometimes) behind particular assignment and helped us to tune our classification algorithm to reflect the best of the knowledge provided in all three classifications resources. This eventually resulted in a high degree of consistency with our classification and demonstrates that automated approaches can provide results comparable to manual classifications.

SYSTEM AND METHODS

The PDB (Berman *et al.*, 2000) of February 13, 2002 with 17 304 entries was used as the source of original structural data.

The overall framework of the approach for building a domain-based dataset is given in Figure 1. We considered only those protein chains which are at least 30 residues long and are not theoretical models. The protein chain was considered as interacting with DNA if the following holds true:

- The DNA fragment size is at least 5 bp long;

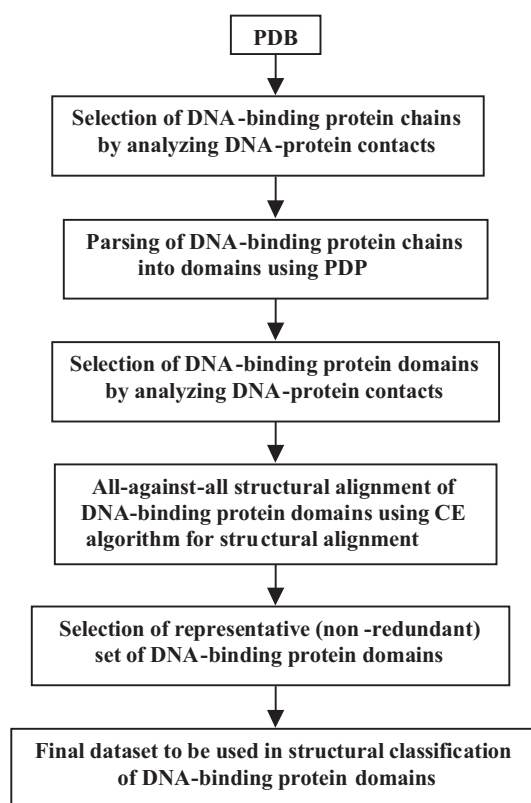


Fig. 1. Overall framework of the approach for building a domain representative dataset.

- At least 5 different protein residues are involved in the interaction with DNA;
- The contact distance cutoff between interacting atoms was $<5\text{\AA}$.

We did not take into account the different types of DNA (A, B, Z) because of the insufficient level of this annotation in the PDB. Also we considered both single-stranded and double-stranded DNA because it is difficult to make an *a priori* decision about the biological significance of the number of strands in a particular case. Besides in some cases only a single strand of double-stranded DNA is present in the crystallized DNA-protein complex. As a result of this procedure we produced a set of protein chains interacting with DNA.

The derived set of protein chains was processed with the Protein Domain Parser (PDP) (Alexandrov and Shindyalov, 2002) software to cut chains into domains in those cases where the chain comprises more than one domain. Domains comprising more than one chain are not considered by this algorithm. Only those domains which maintained interactions with DNA, as defined by the above criteria were kept for further analysis.



Fig. 2. Determination of matched protein-DNA contact pattern for two hypothetical DNA-protein domain complexes *A* and *B* structurally aligned to each other. All residues except those matched to ‘-’ are considered aligned to each other. Stars denote residues involved in protein-DNA interactions. Vertical bars denote matched protein residues involved in interaction with DNA.

Comparison of all DNA-binding protein domains to each other was then performed using the Combinatorial Extension (CE) algorithm for structural alignment (Shindyalov and Bourne, 1998). In this way a set of representative or non-redundant domains was defined. Representatives are different from each other as defined by the following criteria:

- *Rmsd*, root mean squared deviation between two aligned and compared protein domains >2.0 Å;
- *Z-score*, statistical score obtained from CE is <4.5;
- *Rnar*, ratio of the number of aligned residues to the smallest domain length <90%;
- Sequence identity in the alignment <90%.

Structural comparison of the representative DNA-binding protein domains defined above was also performed using the CE algorithm. Two classes of parameters measuring domains similarity used in building the DNA-binding protein domain classification are considered: (1) Parameters measuring structural similarity: *Rmsd*, *Z-score*, *Rnar*; (2) Parameter measuring the match between DNA-protein contact patterns, *Rmat*, can be described in terms of matched protein residues involved in interaction with DNA. For two DNA-protein domain complexes *A* and *B* we consider matched protein residues which are structurally aligned to each other in the structural alignment between *A* and *B* and both interact with DNA (Figure 2). Then $Rmat = \min\{Rmat^A, Rmat^B\}$, where $Rmat^X$ is the ratio of the number of matched residues to the total number of residues involved in contacts with DNA in the DNA-protein complex *X*.

The resulting representative set was used to build a domain classification. Classes consist of domains where the following conditions hold true: (1) For every two domains in the class there is a set of domains from the same class which link these two domains through a DNA-binding domain similarity relationship defined as

follows. Two aligned domains are considered structurally similar (structural neighbors) if the following conditions are satisfied:

$$Rmsd \leq Rmsd_{threshold}, Z - score \geq Z - score_{threshold}, \\ Rnar \geq Rnar_{threshold}, Rmat \geq Rmat_{threshold},$$

where $Rmsd_{threshold}$, $Z - score_{threshold}$, $Rnar_{threshold}$ and $Rmat_{threshold}$ are respective threshold values for the alignment parameters under consideration. The procedure for the choice of thresholds is heuristic and discussed below in the Results section. (2) For every two domains from the different classes there is no such similarity relationship as defined in (1).

In cases where the criterion on matched protein-DNA contact patterns ($Rmat \geq Rmat_{threshold}$) was not satisfied, an iterative procedure of structural alignment optimization was applied. This procedure used dynamic programming alignment with a similarity matrix built for optimal superposition of protein domains from complexes *A* and *B* based on an existing structural alignment. An element of similarity matrix $\{S_{ij}\}$ was defined as follows:

$$S_{ij} = S_{ij}^{dist} + S_{ij}^{cont}, \quad (1)$$

where S_{ij}^{dist} reflects the contribution from the Euclidian distance between C^α atoms:

$$S_{ij}^{dist} = \begin{cases} C_1 - d_{ij}, & \text{if } C_1 - d_{ij} > C_2 \\ C_2, & \text{otherwise} \end{cases}, \quad (2)$$

where C_1 and C_2 are constants determining scaling between d_{ij} and S_{ij} , where d_{ij} is a Euclidian distance between C^α atoms of residue *i* in protein domain from complex *A* and residue *j* in protein domain from complex *B*, and S_{ij}^{cont} reflects the contribution from the matched protein-DNA contact patterns:

$$S_{ij}^{cont} = C_3 \cdot K_i^A \cdot K_j^B, \quad (3)$$

where

$$K_m^X = \begin{cases} 1, & \text{if protein residue is involved in contact} \\ & \text{with DNA} \\ 0, & \text{otherwise} \end{cases}, \quad (4)$$

m—denotes protein residue, *X*—protein-DNA complex; C_3 is a scaling constant.

This procedure allows us to improve the match between protein-DNA contact patterns for protein domains from complexes *A* and *B* also taking into account structural similarity between them.

SCOP version 1.55 (Murzin et al., 1995; Lo Conte et al., 2000) and the Dali Domain Dictionary version 3

(Dietmann and Holm, 2001; Dietmann *et al.*, 2001) were used as the source of alternative classifications for comparison. For comparison of our structural classification of DNA-binding domains with another structural classification standard measures, e.g. Jaccard's, Dice's, Sokal and Sneath's, Simple matching, Rogers and Tanimoto's coefficients (Anderberg, 1973) were used (formulas for coefficients are given in Table 1). The following particular cases were considered: (i) group or family levels in the classification of Luscombe *et al.* (2000); (ii) class, fold, superfamily or family levels in SCOP classification (Murzin *et al.*, 1995; Lo Conte *et al.*, 2000); (iii) attractor, fold topology, functional family or sequence family levels in the Dali Domain Dictionary (Dietmann and Holm, 2001; Dietmann *et al.*, 2001).

RESULTS

805 protein chains involved in the interaction with DNA were selected from 37 119 chains in the PDB. They have been parsed using PDP into 1547 domains, from which 1085 domains were assigned as DNA-binding domains. 338 domains from 269 protein chains (from 217 PDB entries) were selected as representative DNA-binding protein domains. This was done to remove clearly redundant (very similar) domains. 161 representative domains were unique, they did not have any represented domains. 119 representatives were enzymes and the rest were regulatory proteins. 156 domains belong to DNA-protein chains are described in Luscombe *et al.*'s classification, 267 domains from 338 are represented in SCOP, 237 domains are classified in Dali Domain Dictionary. It was shown that the selection of representatives according to the criteria specified above is not likely going to impact resulting classification because representatives and their represented domains have been consistently in agreement with all three classifications, i.e. all represented domains were always in the same family as their representative in all three classifications.

Structural comparison of the 338 representative DNA-binding protein domains to each other was performed using the CE algorithm. The threshold values $Rmsd_{threshold}$, $Z - score_{threshold}$ and $Rnar_{threshold}$ were determined during the process of building an optimal classification with reference to other structural and functional classifications and protein function *per se* as taken from Luscombe *et al.* At $Rmsd_{threshold} = 5.0 \text{ \AA}$, $Z - score_{threshold} = 3.5$ and $Rnar_{threshold} = 70\%$ 98 representative domains do not have structural neighbors, they form so-called singleton classes, the rest, 240 domains form 38 structural classes with multiple members, from which 110 domains forming 4 classes were functionally not correctly classified. All these domains contain a long α -helix but belong to different functional families, therefore their classification is complicated.

Figure 3 illustrates the classification of these domains. For example, the domains of the histone family were mixed with DNA-binding domains of DNA polymerase- β (1BPX:A), replication terminator protein (1ECR:A) and major centromere autoantigen B (1HLV:A). POU-specific DNA-binding domains of Octamer-binding transcription factor 1 (1HF0:A), Pit-1 POU domain factor (1AU7:A, B) and POU-domain transcription factor (1CQT:A, B) were mixed with DNA-binding domains of the 434 Cro protein (3CRO:L) and the phage λ (1LLI:A) and 434 repressors (2OR1:L). The DNA-binding domains of the leucine zipper family were mixed with those from the helix-loop-helix family. 71 domains of different functional families, such as homeodomain, Ets domain, transcription factor TFIID, enzymes and some others, form one class. Further strengthening of clustering thresholds $Rnar_{threshold}$ to 90% (Figure 3) and $Rmsd_{threshold}$ to 2.5 \AA (data not shown) gives a tremendous number of singleton classes, splits the functional classes and leads to a classification where the domains of helix-loop-helix, MADS box, Ets domain and POU-specific domain families become singleton classes. This implies that the analysis of additional structural features of domains is desirable. Our response to this observation involved the detection and evaluation of 3D-interaction patterns in DNA-protein interfaces, their use in optimization of the alignment between two proteins and finally in domain classification.

A new criterion $Rmat$ for a match in protein-DNA contact patterns in two domains was used in addition to the three parameters for domain structural similarity $Rmsd$, $Z - score$ and $Rnar$. Further, an alignment optimization procedure was introduced which takes into account residues involved in protein-DNA interaction and realigns proteins using a criteria combining structural similarity and matching between residues interacting with DNA from two proteins. The procedure is applied when the $Rmat$ value was less than some cutoff value $Rmat_{threshold}$. This procedure allowed us to improve the match between protein-DNA contact patterns for protein domains which are functionally close without significant impact on structural similarity as measured by $Rmsd$, $Z - score$ and $Rnar$. Conversely, it does not improve the match between protein-DNA contact patterns for protein domains which are functionally different. In the second case the values $Rmsd$, $Z - score$, $Rnar$ and $Rmat$ change significantly beyond the established thresholds. After re-alignment the values of $Rmsd$, $Z - score$, $Rnar$ and $Rmat$ were evaluated again using threshold values. Thus DNA-binding domain similarity was defined as follows:

- If $Rmsd > 5.0 \text{ \AA}$ or $Rnar < 70\%$ or $Z - score < 3.5$, then domains are not considered as similar;
- If $Rmsd \leq 3.0 \text{ \AA}$ and $Rnar \geq 80\%$, then domains are

Table 1. Comparison of the classification for all 338 DNA-binding domain representatives with SCOP and Dali classifications at various threshold parameters defining the structural similarity of domains

		N	Jaccard's	Dice's	Sokal&Sncath's I	Sokal&Sncath's II	simple matching	Rogers&Tanimoto's	Yule's
Our classification versus SCOP family level classification for domains represented in SCOP									
<i>Rmat_{threshold}</i>	60%	263	0.306	0.468	0.180	0.985	0.971	0.943	0.975
(3.0Å < <i>Rmsd</i> ≤ 5.0Å	65%	263	0.323	0.488	0.192	0.987	0.974	0.949	0.977
or	70%	263	0.332	0.498	0.199	0.988	0.976	0.954	0.978
70% ≤ <i>Rnar</i> < 80%	75%	263	0.402	0.574	0.252	0.991	0.983	0.966	0.987
<i>Z</i> – score ≥ 3.5)	80%	263	0.456	0.626	0.295	0.993	0.987	0.973	0.993
	50%	263	0.095	0.173	0.050	0.923	0.858	0.751	0.895
<i>Rnar_{threshold}</i>	55%	263	0.099	0.180	0.052	0.930	0.868	0.768	0.894
	60%	263	0.120	0.214	0.064	0.946	0.898	0.815	0.909
(<i>Rmsd</i> ≤ 5.0 Å,	65%	263	0.162	0.279	0.088	0.964	0.930	0.869	0.936
<i>Z</i> – score ≥ 3.5)	70%	263	0.206	0.342	0.115	0.974	0.949	0.903	0.953
	75%	263	0.286	0.444	0.167	0.984	0.969	0.939	0.971
no accounting contacts	80%	263	0.315	0.479	0.187	0.988	0.976	0.952	0.975
	90%	263	0.315	0.479	0.187	0.993	0.986	0.973	1.000
Our classification versus SCOP superfamily level classification for domains represented in SCOP									
<i>Rmat_{threshold}</i>	60%	263	0.170	0.290	0.093	0.932	0.873	0.775	0.814
(3.0Å < <i>Rmsd</i> ≤ 5.0Å	65%	263	0.149	0.259	0.080	0.926	0.861	0.756	0.768
or	70%	263	0.126	0.223	0.067	0.918	0.848	0.736	0.697
70% ≤ <i>Rnar</i> < 80%,	75%	263	0.121	0.216	0.065	0.915	0.843	0.729	0.682
<i>Z</i> – score ≥ 3.5)	80%	263	0.117	0.209	0.062	0.910	0.834	0.715	0.668
	50%	263	0.148	0.258	0.080	0.904	0.824	0.701	0.802
<i>Rnar_{threshold}</i>	55%	263	0.151	0.262	0.082	0.907	0.830	0.709	0.806
	60%	263	0.153	0.266	0.083	0.917	0.847	0.735	0.795
(<i>Rmsd</i> ≤ 5.0 Å,	65%	263	0.176	0.299	0.096	0.932	0.872	0.773	0.829
<i>Z</i> – score ≥ 3.5)	70%	263	0.189	0.319	0.105	0.938	0.883	0.791	0.847
	75%	263	0.186	0.314	0.103	0.939	0.884	0.793	0.840
no accounting contacts	80%	263	0.121	0.215	0.064	0.904	0.824	0.701	0.691
	90%	263	0.056	0.107	0.029	0.780	0.639	0.469	0.217
SCOP family level classification versus Dali functional family level classification for domains represented in SCOP and Dali									
		200	0.555	0.714	0.384	0.992	0.984	0.969	0.994
SCOP superfamily level classification versus Dali functional family level classification for domains represented in SCOP and Dali									
		200	0.585	0.738	0.413	0.990	0.979	0.960	1.000
Our classification vs SCOP family level classification for domains represented in SCOP and Dali									
<i>Rmat_{threshold}</i>	60%	200	0.333	0.500	0.200	0.982	0.964	0.931	0.976
(3.0Å < <i>Rmsd</i> ≤ 5.0 Å	65%	200	0.347	0.516	0.210	0.983	0.967	0.936	0.977
or	70%	200	0.364	0.534	0.222	0.986	0.971	0.944	0.978
70% ≤ <i>Rnar</i> < 80%,	75%	200	0.459	0.629	0.298	0.990	0.981	0.962	0.988
<i>Z</i> – score ≥ 3.5)	80%	200	0.525	0.689	0.356	0.993	0.986	0.972	0.994
	50%	200	0.106	0.192	0.056	0.906	0.827	0.706	0.908
<i>Rnar_{threshold}</i>	55%	200	0.109	0.196	0.057	0.910	0.836	0.718	0.905
	60%	200	0.124	0.220	0.066	0.927	0.864	0.761	0.908
(<i>Rmsd</i> ≤ 5.0 Å,	65%	200	0.167	0.285	0.091	0.951	0.907	0.830	0.933
<i>Z</i> – score ≥ 3.5)	70%	200	0.218	0.358	0.122	0.966	0.934	0.875	0.954
	75%	200	0.320	0.485	0.191	0.981	0.963	0.928	0.973
no accounting contacts	80%	200	0.340	0.508	0.205	0.984	0.969	0.940	0.974
	90%	200	0.347	0.515	0.210	0.992	0.984	0.968	1.000

Table 1 continued . . .

	N	Jaccard's	Dice's	Sokal&Sncath's I	Sokal&Sncath's II	simple matching	Rogers&Tanimoto's	Yule's	
Our classification vs SCOP superfamily level for domains represented in SCOP and Dali									
<i>Rmat_{threshold}</i>	60%	200	0.206	0.342	0.115	0.936	0.880	0.786	0.868
(3.0 Å < <i>Rmsd</i> ≤ 5.0 Å	65%	200	0.186	0.313	0.102	0.931	0.871	0.772	0.837
or	70%	200	0.150	0.261	0.081	0.922	0.855	0.746	0.760
70% ≤ <i>Rnar</i> < 80%,	75%	200	0.145	0.254	0.078	0.922	0.855	0.747	0.745
<i>Z</i> - score ≥ 3.5)	80%	200	0.139	0.244	0.075	0.917	0.847	0.735	0.729
	50%	200	0.157	0.271	0.085	0.892	0.805	0.674	0.842
<i>Rnar_{threshold}</i>	55%	200	0.159	0.274	0.086	0.895	0.810	0.680	0.841
	60%	200	0.160	0.276	0.087	0.907	0.830	0.709	0.817
(<i>Rmsd</i> ≤ 5.0 Å,	65%	200	0.194	0.326	0.108	0.928	0.867	0.764	0.861
<i>Z</i> - score ≥ 3.5)	70%	200	0.219	0.359	0.123	0.939	0.885	0.793	0.884
	75%	200	0.225	0.368	0.127	0.945	0.895	0.810	0.883
No accounting contacts	80%	200	0.156	0.270	0.085	0.916	0.845	0.731	0.786
	90%	200	0.064	0.120	0.033	0.798	0.664	0.497	0.279
Our classification vs Dali functional family classification for domains represented in SCOP and Dali									
<i>Rmat_{threshold}</i>	60%	200	0.499	0.666	0.332	0.987	0.975	0.951	0.993
(3.0 Å < <i>Rmsd</i> ≤ 5.0 Å	65%	200	0.492	0.659	0.326	0.988	0.976	0.952	0.991
or	70%	200	0.444	0.615	0.286	0.987	0.975	0.951	0.985
70% ≤ <i>Rnar</i> < 80%,	75%	200	0.553	0.712	0.382	0.992	0.984	0.968	0.993
<i>Z</i> - score ≥ 3.5)	80%	200	0.641	0.781	0.471	0.995	0.989	0.978	0.999
	50%	200	0.141	0.246	0.076	0.911	0.836	0.718	0.965
<i>Rnar_{threshold}</i>	55%	200	0.147	0.257	0.080	0.916	0.845	0.732	0.968
	60%	200	0.177	0.301	0.097	0.934	0.876	0.779	0.974
(<i>Rmsd</i> ≤ 5.0 Å,	65%	200	0.242	0.390	0.138	0.957	0.918	0.849	0.980
<i>Z</i> - score ≥ 3.5)	70%	200	0.321	0.486	0.191	0.972	0.945	0.895	0.987
	75%	200	0.475	0.644	0.311	0.986	0.973	0.947	0.991
no accounting contacts	80%	200	0.468	0.638	0.306	0.988	0.976	0.953	0.988
	90%	200	0.293	0.453	0.172	0.990	0.979	0.959	0.999

Note: **N**, the number of domains considered; **Jaccard's coefficient** (also known as Gower's general similarity coefficient) is calculated as $A/(A+B+C)$; **Dice's coefficient** (also known as Soren's or Czekanowski's or Nei and Lei's coefficient) is calculated as $2A/(2A+C+B)$; **Sokal and Sncath's coefficient I** is calculated as $A/(A+2(B+C))$; **Sokal and Sncath's coefficient II** is calculated as $2(A+D)/(2(A+D)+B+C)$; **simple matching coefficient** is calculated as $(A+D)/(A+B+C+D)$; **Rogers and Tanimoto's coefficient** is calculated as $(A+D)/(A+D+2(B+C))$; **Yule's coefficient** (also known as the coefficient of colligation) is calculated as $(AD-BC)/(AD+BC)$; **A**, the quantity of true_true; **B**, the quantity of true_false; **C**, the quantity of false_true; **D**, the quantity of false_false.

considered as similar;

- If $Rmat \geq Rmat_{threshold}$ and either: $3.0 \text{ \AA} < Rmsd \leq 5.0 \text{ \AA}$ and $Rnar \geq 70\%$ or $70\% \leq Rnar < 80\%$ and $Rmsd \leq 5.0 \text{ \AA}$, then domains are considered similar.

Figure 3 illustrates the tree classification of 110 domains containing α -helix considered above when patterns of contacts are considered at various values of $Rmat_{threshold}$. If $Rmat < Rmat_{threshold}$ the pair of domains is re-aligned using a new procedure and starting from the initial structural alignment. If the new alignment does not satisfy the conditions $Rmsd \leq 5.0 \text{ \AA}$, $Rnar \geq 70\%$, $Z - score \geq 3.5$, $Rmat \geq Rmat_{threshold}$, the pair of domains is not considered similar. At an $Rmat_{threshold}$ of 60% the domains of helix-loop-helix separate from

leucine zipper domains, histones and transcription factor TFIID form separate classes and helix-turn-helix domains are not mixed with enzyme DNA-binding domains. At an $Rmat_{threshold}$ of 70% POU-specific domains separate from DNA-binding domains of the 434 Cro protein and the phage λ and 434 repressors. At an $Rmat_{threshold}$ of 75% Ets-domains and homeodomains form separate classes. An $Rmat_{threshold}$ of 80% divides two superfamilies of domains containing helix-turn-helix motif (SCOP: a.4.1. and a.4.5.; Dali: DC_3_155 and DC_3_153) and histones H3 and H2A.

The accounting of $Rmat_{threshold}$ starting from 65% allowed us to achieve a better classification of DNA-binding domains than the classification at any $Rnar_{threshold}$ in the range from 50% to 85% for the SCOP family level and for

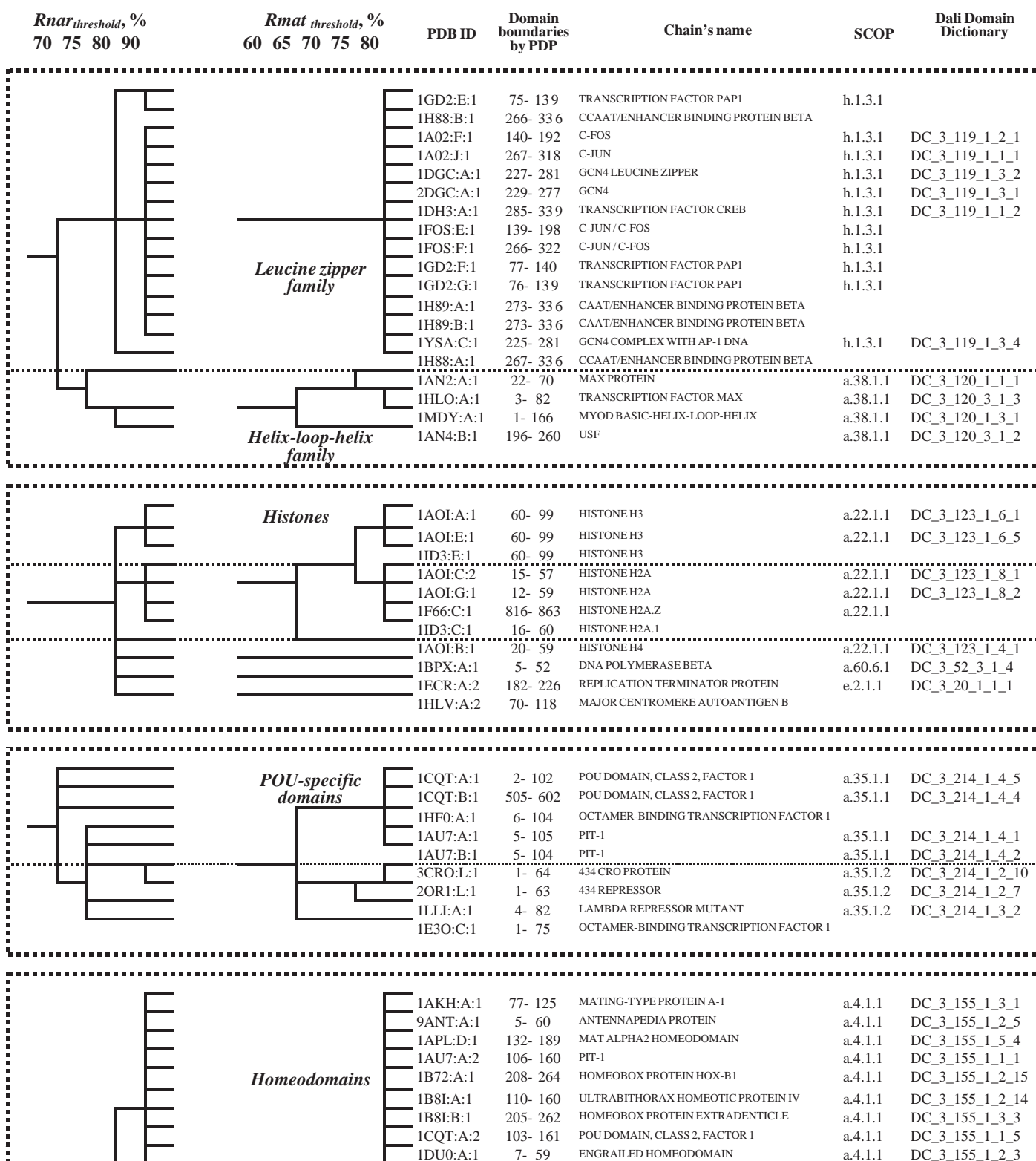


Fig. 3. Final classification of DNA-binding protein domains. (cont.)

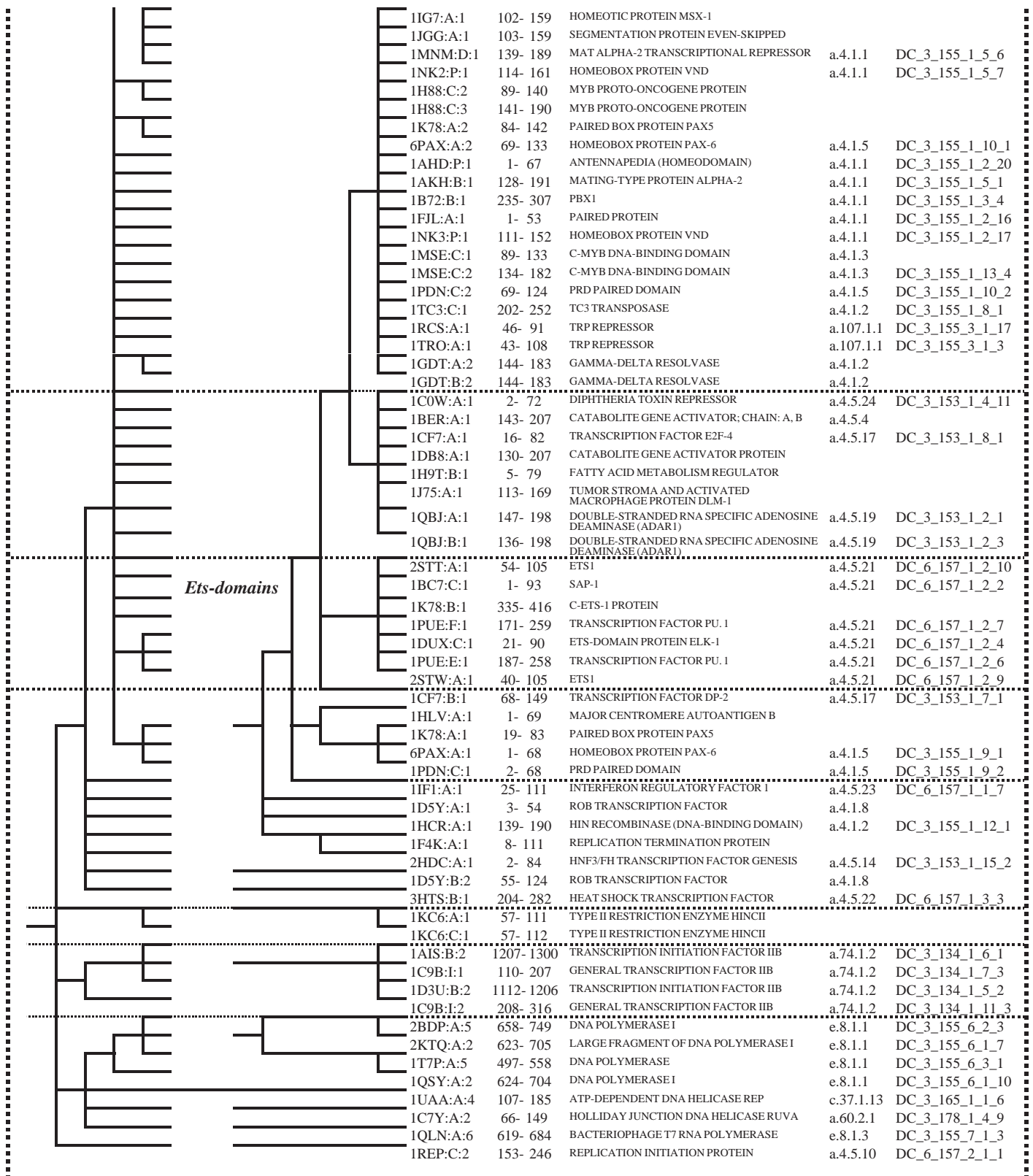


Fig. 3. (cont.)

the Dali functional family level. Table 1 illustrates these results. The boundaries of 263 domains from a total of 267 represented in SCOP are nearly identical. Similarly, 200 domains from 237 domains representing in the Dali Domain Dictionary also have nearly identical domain boundaries with the Dali domain definition and are also represented in SCOP. Thus we have compared our classifications with SCOP on 263 domains and also with SCOP and Dali on 200 domains. One can see from Table 1 that 7 different coefficients, which each measure the match between our structural classifications of DNA-binding domains and those of SCOP and Dali, give values for classifications which are significantly higher than those which do not account for protein-DNA contacts. Further, the coefficients for classifications accounting for contacts are comparable to the coefficients measuring the SCOP classification at the family level and Dali at the functional family level and better than those corresponding to the SCOP superfamily level.

Table 1 also shows that the threshold parameter $R_{mat_{threshold}}$ of 80% gives the best classification when compared with the SCOP family and Dali functional family levels.

Using these parameters our program of automated classification classified 338 representative DNA-binding domains into 45 multimember (containing 2 or more domains) classes and 143 singleton (containing one domain) classes. Taking into consideration the interaction patterns between protein domains and DNA allowed us to improve the classification with respect to the one based on structural parameters only.

DISCUSSION

Our analysis has shown that fully automated protein structure classification approaches, at least for DNA-protein complexes, do not achieve the same quality of annotation, as defined by the biology, as manual or combined (manual and automated) approaches. However, it is known that approaches containing a manual component have obvious shortcomings such as: (i) significant backlog relative to currently available 3D data; (ii) limited throughput likely to be insufficient to handle the growing dataflow bought about by structural genomics; (iii) don't provide uniform or reproducible classification since a human and inconsistent component is involved; and (iv) heavily depend on highly skilled human experts for which the availability is limited. Thus there is a need for fully automated approaches capable of achieving human expert level quality, but which are consistent. In this work we are aiming to meet this need on a subset of structure data, specifically DNA-binding protein domains. Our approach is based on the analysis of structural similarity and protein-DNA interaction patterns. Also our analysis is performed at the level of potential functional domains, which are detected

as part of this study. The addition of protein-DNA interaction data added to purely structural information to yield a better classification is the highlight of this paper. The classification so produced is the most complete and current structural classification of DNA-binding protein domains available.

This is not the first attempt to enhance a purely sequence and/or structure-based classification with additional knowledge (see review by Ponting and Russell (2002)). Notable is the classification of adenine-binding proteins based on the properties of the ligand-binding sites (Cappello *et al.*, 2002) and classification of protein domains based on the distribution of $C^\alpha-C^\alpha$ distances between residues (Carugo and Pongor, 2002). Our results confirm this type of approach. Thus, taking into account additional information, such as interaction patterns improves the classification towards providing more functional meaningful information. Stated in a converse way, by only taking into consideration structural information leads to homologous domains assigned to different functional families.

This analysis produced a complete and current classification of DNA-binding protein domains from available PDB data. It is available from the URL <http://spdc.sdsc.edu>. The future development of this approach to protein classification will be aimed at the following: (i) building multiple alignments of protein domains related within our classification; (ii) enriching existing set of protein domains with other related proteins based on structure and sequence criteria; and (iii) using these enriched datasets in building recognition methods for DNA-binding domains which can be used in genome annotation. The approach used here for DNA-binding protein domains could be extended to the characterization of protein-protein and protein-ligand interactions.

ACKNOWLEDGMENT

This work was supported by the National Partnership for Advanced Computational Infrastructure from the National Science Foundation, and NSF Grants DBI 9808706, DBI 0111710, and NIH NIGMS Grant GM63208-01A1.

REFERENCES

- Alexandrov, N.N. and Shindyalov, I.N. (2002) PDP: Protein Domain Parser. *Bioinformatics*, submitted.
- Anderberg, M.R. (1973) *Cluster Analysis for Applications*. Academic Press.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The protein data bank. *Nucleic Acid Res.*, **28**, 235–242.
- Cappello, V., Tramontano, A. and Koch, U. (2002) Classification of proteins based on the properties of the ligand-binding site: the case of adenine-binding proteins. *Proteins*, **47**, 106–115.

- Carugo, O. and Pongor, S. (2002) Protein fold similarity estimated by a probabilistic approach based on $C^\alpha-C^\alpha$ distance comparison. *J. Mol. Biol.*, **315**, 887–898.
- Dengler, U., Siddiqui, A.S. and Barton, G.J. (2001) Protein structural domains: analysis of the 3Dee domains database. *Proteins*, **42**, 332–344.
- Dietmann, S. and Holm, L. (2001) Identification of homology in protein structure classification. *Nat. Struct. Biol.*, **8**, 953–957.
- Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M. and Holm, L. (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.*, **29**, 55–57.
- Lo Conte, L., Ailey, B., Hubbard, T.J.P., Brenner, S.E., Murzin, A.G. and Chothia, C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.
- 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**, 1–37.
- Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
- Orengo, C.A. and Taylor, W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.*, **266**, 617–634.
- Pearl, F.M.G., Lee, D., Bray, J.E., Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M. and Orengo, C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282.
- Ponting, C.P. and Russell, R.R. (2002) The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 45–71.
- Shindyalov, I.N. and Bourne, P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
- Siddiqui, A.S., Dengler, U. and Barton, G.J. (2001) 3Dee: a database of protein structural domains. *Bioinformatics*, **17**, 200–201.