Protein Engineering vol.14 no.7 pp.465–472, 2001

A Web-based classification system of DNA-binding protein families

M.Karmirantzou and S.J.Hamodrakas[1]

Faculty of Biology, Department of Cell Biology and Biophysics, University of Athens, Panepistimiopolis, Athens 157 01, Greece

[1]To whom correspondence should be addressed. E-mail: shamodr@cc.uoa.gr

Rational classification of proteins encoded in sequenced genomes is critical for making the genome sequences maximally useful for functional and evolutionary studies. The family of DNA-binding proteins is one of the most populated and studied amongst the various genomes of bacteria, archaea and eukaryotes and the Web-based system presented here is an approach to their classification. The DnaProt resource is an annotated and searchable collection of protein sequences for the families of DNA-binding proteins. The database contains 3238 full-length sequences (retrieved from the SWISS-PROT database, release 38) that include, at least, a DNA-binding domain. Sequence entries are organized into families defined by PROSITE patterns, PRINTS motifs and *de novo* excised signatures. Combining global similarities and functional motifs into a single classification scheme, DNA-binding proteins are classified into 33 unique classes, which helps to reveal comprehensive family relationships. To maximize family information retrieval, DnaProt contains a collection of multiple alignments for each DNA-binding family while the recognized motifs can be used as diagnostically functional fingerprints. All available structural class representatives have been referenced. The resource was developed as a Web-based management system for online free access of customized data sets. Entries are fully hyperlinked to facilitate easy retrieval of the original records from the source databases while functional and phylogenetic annotation will be applied to newly sequenced genomes. The database is freely available for online search of a library containing specific patterns of the identified DNA-binding protein classes and retrieval of individual entries from our WWW server (http://kronos.biol.uoa.gr/~mariak/dbDNA.html).

Keywords: DNA-binding proteins/pattern identification/protein family classification/WWW

Introduction

The recent advances in genome sequencing led to a rapid enrichment of protein databases with an unprecedented variety of deduced protein sequences, several of them without a documented functional role (Bork *et al.*, 1994).

Advanced and specialized databases are needed to facilitate the retrieval of relevant information from the deluge of sequence data and to provide insight into the protein structure and function. Further, it is clear that rational classification of proteins encoded in sequenced genomes is critical for making the genome sequences maximally useful for functional and evolutionary studies (Wu *et al.*, 1996).

Computational biology applying fast and sensitive algorithms strives to extract the maximum possible information from these sequences by classifying them according to their homologous relationships, predicting their likely biochemical activities and/or cellular functions, three-dimensional structures and evolutionary origin. This challenge is daunting, given that even in *Escherichia coli*, arguably the best-studied organism (Neidhardt *et al.*, 1996), only 40% of the gene products have been characterized experimentally (Koonin, 1997).

The family of DNA-binding proteins is one of the most populated and studied amongst the various genomes of bacteria, archea and eukaryotes. Most of these proteins, such as the eukaryotic and prokaryotic transcription factors, contain independently folded units (domains) in order to accomplish their recognition with the contours of DNA. It is now clear that the majority of these DNA-binding scaffolds which are in general relatively small, less than 100 amino acid residues, belong to a large number of structural families with characteristic sequences and three-dimensional designs or conformations (Branden and Tooze, 1999).

The DNA-Binding Proteins Database (DnaProt) has been designed as an attempt to classify protein sequences from completely sequenced genomes and more specifically those that characteristically recognize specific DNA chains. It is a secondary, value-added database that organizes non-redundant SWISS-PROT protein sequences on the basis of binding recognition concept, i.e. the sequence-specific protein–DNA

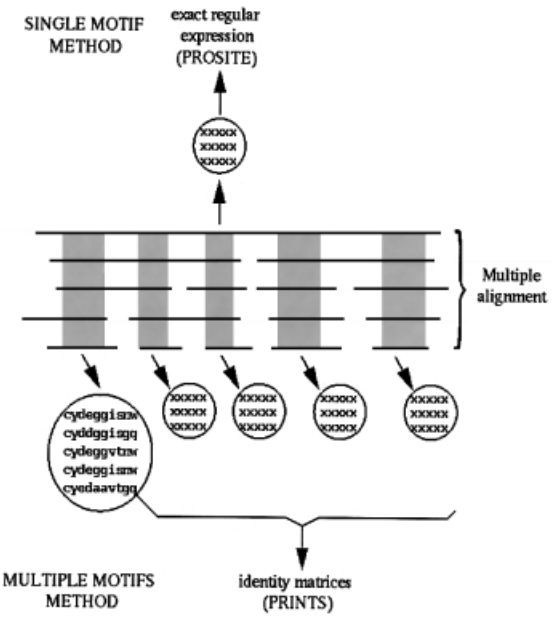

Fig. 1. Illustration of the two principal methods for building pattern databases, i.e. using single motifs and multiple motifs. In the former method, the sequence information can be translated to a single consensus expression, e.g. C–Y–x(2)–[DG]–G–x–[ST], where residues within square brackets are allowed at that position and x denotes any residue. By contrast, in the multiple motif method, a group of motifs called a fingerprint is used to retain all residue information.

code that relates the primary structure of proteins to preferred binding sites of DNA (Choo and Klug, 1997). To our knowledge, this is the first Web-based resource that explicitly classifies DNA-binding proteins of known sequence into specific families and allows queries against a library containing specific patterns of identified classes.

Methods

Collection of DNA-binding proteins

An annotated collection of protein sequences for the families of DNA-binding proteins comprises a fully comprehensive resource that contains 3238 full-length sequences (retrieved from SWISS-PROT database, release 38) (Bairoch and Apweiler, 1998) that include at least a DNA-binding domain (fragments are excluded from the current release of the database).

The selected sequence entries are organized into families according to family relationships defined collectively by PROSITE (Bairoch *et al.*, 1997) and PRINTS (Attwood *et al.*, 1998) patterns that usually tend to correspond to the core structural or functional elements of the protein family. Their conserved nature allows them to be used to diagnose family membership (Figure 1). Initially the protein DNA-binding family catalogue is constructed according to the list of related patterns of the PROSITE database (Bairoch *et al.*, 1997). Protein sequences that have not been identified in the former database are incorporated by the collective homologous entries of the PRINTS resource (Attwood *et al.*, 1998) and reflected in subsequent updates of our pattern library. The DNA-binding protein collection currently includes 3238 entries that can all be identified in the above secondary databases and pattern resources.

Construction of the database–classification system

Combining global similarities and functional motifs into a single classification system, a comprehensive collection of DNA-binding protein sequence motifs was derived, which classifies the selected sequence entries into related recognition families. Newly identified DNA-binding families were included in the taxonomy system of the database and unclassified sequence members were incorporated into the existing classes following the subsequent steps:

1. an all-against-all sequence comparison analysis of the protein sequences, members of a particular DNA-recognition family, using the gapped BLAST program (Altschul *et al.*, 1990) after masking (a) low-complexity and (b) predicted coiled-coil regions (regions containing short-periodicity internal repeats) (Promponas *et al.*, 2000) was

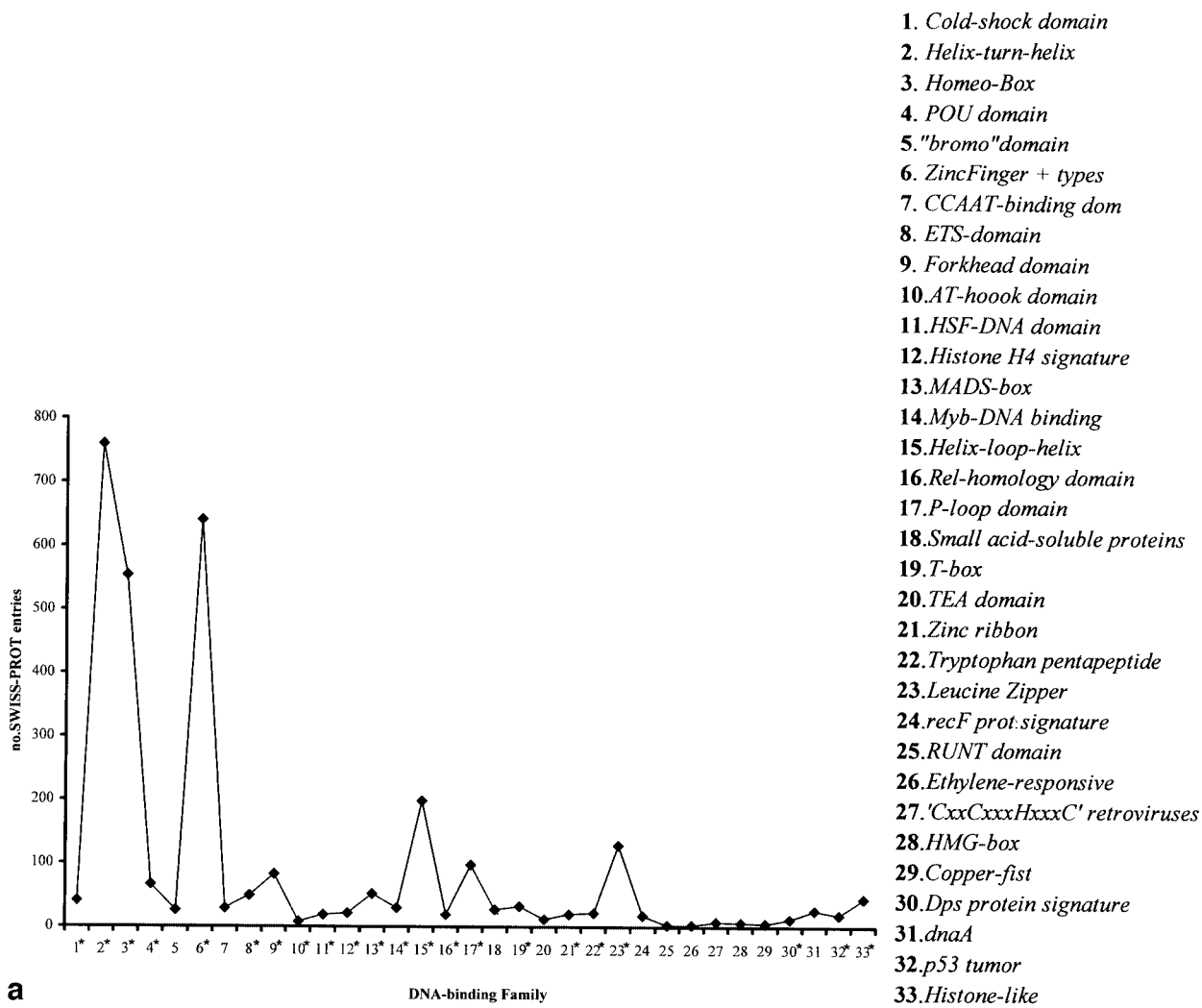

Fig. 2.

carried out;
2. identification of plausible functional motifs, detailed characterization of all conserved amino acid expressions defined by PROSITE and PRINTS and recognition of *de novo* sequence elements that correspond to specific DNA-binding classes was done; and
3. an analysis of DNA-binding protein sequences conserved regions was also performed.

By the design of this procedure, 33 distinct DNA-binding classes were identified (Table I) that may provide simple notions for the comprehensive analysis of DNA-binding family members at the evolutionary and phylogenetic level.

Information at the family level

To maximize family information retrieval and to verify significance of the relationships, the database provides an up-to-date collection of sequence alignments for all members of each DNA-binding family. A case-by-case analysis of each family was enhanced by coloured representations of the salient conserved features of its members using the ClustalW program (Thompson *et al.*, 1994). Invariably, closely related sequences show little or no variation that give rise to conservation blocks of matched amino acids whereas distant related sequences produce 'islands' of conservation that disrupt the correct register of the family alignment.

The motifs that are characteristic for the different protein–DNA recognition classes are short, typically around 10–30 amino acids in length, and tend to correspond to the core structural or functional elements of the proteins. Owing to their conserved nature, they may be used as diagnostic signatures of family membership. The database motif collection currently includes all PROSITE patterns (i.e. regular expressions of DNA-binding protein sequences) and contains additional family fingerprints as maintained in the PRINTS resource. Together, the DNA-binding database contains 3238 sequence entries and 94 recognition patterns.

Characteristics of the Web-based classification scheme

In general, the level of sequence similarity between proteins, members of the same family, is relatively high. A case-by-case examination of sequence alignments revealed that the

Table I. Statistics of the DNA-binding proteins database

	No. of index
Total number of sequences	3491
Full-length, non-redundant sequences	3238
Three-dimensional structures	204
Classes of DNA-binding motifs	33

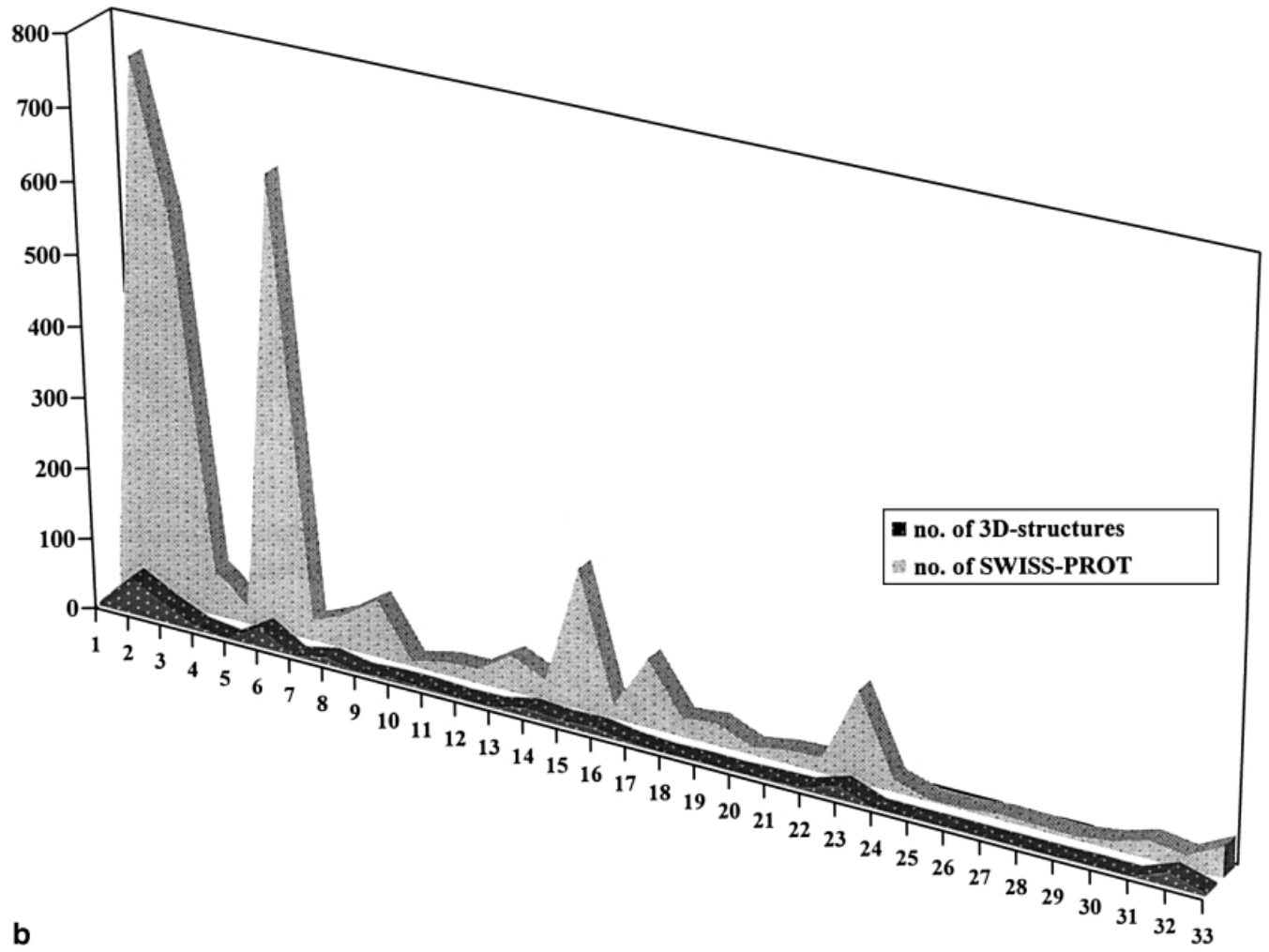

Fig. 2. (**a**) Distribution of DNA-binding protein sequences by the number of identified DNA-recognition category. Families with an experimental structure deposited in PDB denoted by an asterisk. (**b**) Three-dimensional structures and DNA-binding protein sequences per class.

conservation fingerprint extends further than the protein–DNA recognition regions verifying the significance of the relationships within the family members. The functional interpretation of these extended regions is still under investigation. First indications reinforced the hypothesis that such distinct conservation templates can be used for the prediction of functions and may serve as a convenient platform for a variety of evolutionary-oriented analyses of protein families.

The classification of DNA-binding proteins into 33 sequence-specific recognition categories loosely follows the structural taxonomy system introduced by Harrison (Harrison, 1991). While the number of families (22) that have been

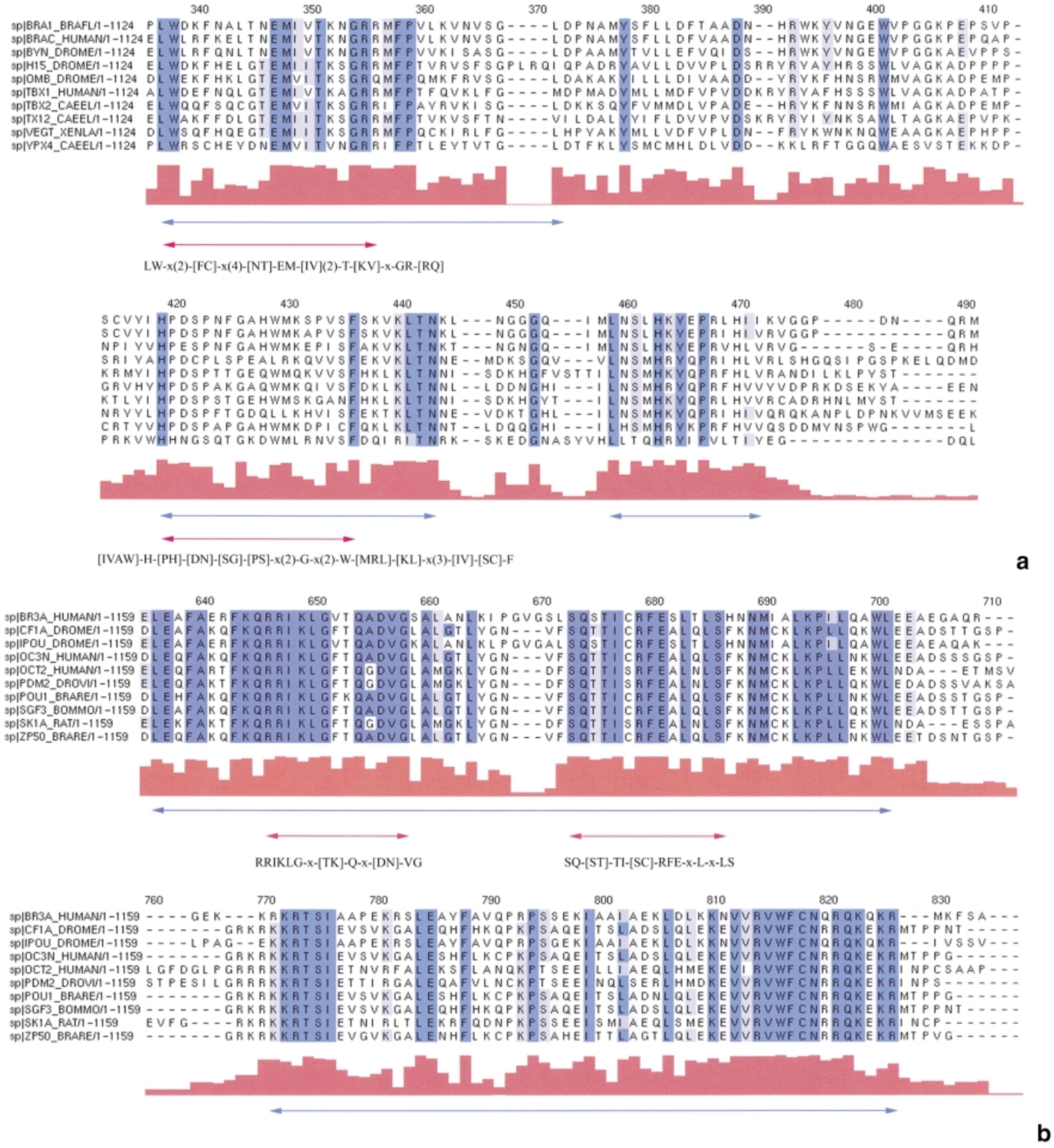

Fig. 3.

structurally characterized is noteworthy (Figure 2a), the current release of the database has also classified entries with no structural representatives (Figure 2b).

Results

Applications of the DNA-binding protein classification system

The most straightforward application of DnaProt is for the prediction of functional sites of individual proteins or protein sets including those from newly completed genomes. Performing a global sequence search against the library of classified DNA-binding patterns contained in the current release allows the diagnosis of a probable family relationship. Given that with the increase of the number of genomes being sequenced and the continued growth of specialized databases, the data serve as a safeguard against the propagation of errors that might be present in the widely used family resources. In particular, it is possible to identify sequence entries or extended conserved regions that might have been missed out during genome annotation or to detect alternative cognates of the given functional patterns among the family members.

The Web-based system of DnaProt also provides opportunities for more detailed analyses of specific DNA-binding protein families. This information might be utilized to detect all the classified entries with a particular 'phylogenetic' pattern, for example those that are found only in specific species, e.g. bacteria. Thus, the site offers a convenient platform for an evolutionary-oriented analysis of DNA-binding protein families.

Three major representative cases

Here, we present three distinct examples of DNA-binding classes. In each case, the analysis of family members and its structural representatives allows one to infer a common feature of family resources, i.e. the caution required to avoid overlying functional predictions and misinterpreted family relationships. The information distilled from these examples both sheds light on the weaknesses of pattern databases and family-specific resources reflecting on their inconsistencies that are endemic in their data sources and presents the major strength of this database, i.e. highly specific family relationships are assigned.

T-box domain

The T-box gene family encodes a variety of transcriptional regulators that have been further identified in invertebrates and vertebrates, including humans (Papaioannou and Silver, 1998; Wattler *et al.*, 1998). The gene products were first uncovered in Brachyury mice on the basis of similarity to its DNA binding domain (T) gene product, which gives a characteristic phenotype to the family. They are essential in tissue specification, morphogenesis and organogenesis and share a well studied and characteristic 170–190 amino acid residue domain named the T-box domain which binds to DNA. Reports on the X-ray structure of the protein product bound to the contours of DNA provide detailed views on the sequence-specific protein–DNA recognition architecture.

As signature patterns for the T-domain, two conserved regions are diagnosed by the PROSITE pattern identification system, reflecting its underlying motif-recognition technique. The first region encoded by PROSITE pattern TBOX_1 (PS01283) corresponds to the N-terminus of the domain and the second one which is encoded by PROSITE pattern TBOX_2 (PS01264) to the central part (Figure 3a). However, the construction of a multiple alignment of the 41 sequences, all T-box family members, housed in the database demonstrates that the local sequence similarity within the family is higher than that underlying the derived regular expressions of PROSITE. Visual inspection of the conserved regions reveals that the complete family relationships cannot be characterized effectively by the individual and originally extracted PROSITE patterns. In fact, our manual analysis revealed that extended, as well as additional, conserved parts of the alignment not only can be used as key discriminators of the specific family, but also can encode sequence-specific information on the protein–DNA recognition code missed by the PROSITE pattern-recognition technique. In particular, as the single crystallographic structure of the T domain–DNA complex [PDBcode, 1XBR (Bernstein *et al.*, 1977); SWISS-PROT code, BRAC_XENLA] revealed that the protein makes contacts with the DNA in both the major and minor grooves (Muller and Herrmann, 1997). Important contacts are made by identical

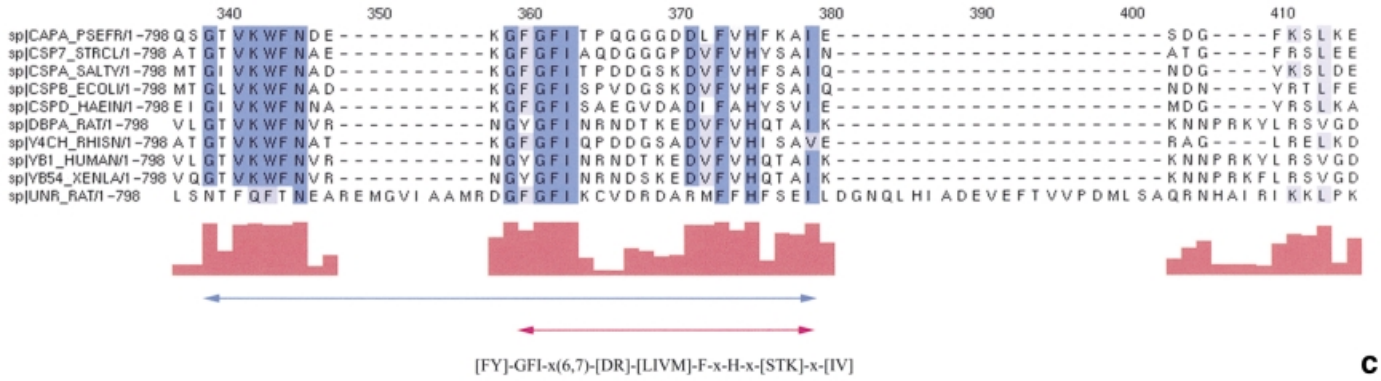

Fig. 3. ClustalW multiple alignments (Thompson *et al.*, 1994) of 10 indicative sequences from three major DNA-binding families [(**a**) T-box, (**b**) POU and (**c**) cold-shock domains]. Handling of the alignments was performed with Jalview (M.Clamp, unpublished data) using the BLOSUM62 colouring scheme, where in conserved sites the dominant residue type has the deepest mid-blue colour and the least has the palest. The identifiers of the sequences according to SWISS-PROT nomenclature are shown on the left-hand side of the N-terminal of each conserved block. Numbers to the right of each ID refer to the start and end residue of the longest sequence in the alignment; the start and end positions do not include gaps. A conservation histogram (Livingstone and Barton, 1993) underneath the alignment highlights the conserved residues. The actual PROSITE DNA-binding patterns are depicted below the sequences and are marked with solid red arrows. Dark blue arrows indicate the extended family signatures emerged from our analysis. For clarity, only 10 randomly selected sequences from the whole family are presented in each case [T-box (41 members), POU (66 members) and cold-shock (45)]. Full family alignments can be viewed after querying the database with the selected family members.

and conservatively substituted residues protruding from loops and strands of the two small β-sheets and from residues of the two C-terminal helices (H3 and H4) (Figure 3a). This sequence-specific recognition signature, which is spread over the whole family alignment, rendered a T-box set of motifs maintained in the database and stored in the library of DNA-binding motifs. When the above selected patterns were used to search SWISS-PROT, the set of sequence entries returned were only correct matches, i.e. members of the T-box family and not a single unrelated sequence.

POU domain

The POU domain, a DNA-binding domain that characterizes a family of eukaryotic transcription factors, has a novel modular structure that contains a 70–75 amino acid specific region (POU_S) (Herr *et al.*, 1988). In some of these regulatory proteins, the selective presence of a variable linker of 15–30 residues and a juxtaposed 60 amino acid POU-type homeo domain (POU_H), always found upstream of the POU_S, form a bipartite DNA-binding POU domain. Such proteins bind to specific DNA sequences to cause temporal and spatial regulation the expression of genes, many of which are involved in the regulation of neuronal development in the central nervous system of mammals (Johnson and Hirsh, 1990). It is thought to confer high-affinity site-specific DNA-binding and to mediate cooperative protein–protein interactions on DNA via specific homodimer or heterodimer formation. Both elements of the POU domain are required for high-affinity sequence-specific DNA binding. The domain may also be involved in protein–protein interactions. Since 1988, when the features of the POU domain were first discovered, several 3D structures of the POU domain have been determined by multi-dimensional NMR (Assa-Munt *et al.*, 1993) and X-ray crystallography up to 2.30 Å resolution (Klemm *et al.*, 1994).

The PROSITE pattern-diagnostic tool derived two signature patterns for the POU domain. The first spans positions 15–27 of the domain, the second positions 42–55 [cf. PROSITE patterns POU_1 (PS00035) and POU_2 (PS00465)] (Figure 3b). The PRINTS resource provides a group of five aligned, unweighted sequence elements that have been manually excised and builds the diagnostic signatures of the POU family. However, when the protein sequences of all family members (66) were gathered together in a multiple alignment using ClustalW, analysis of the most conserved regions within the alignment indicated that a substantial similarity exists between the constituent and phylogenetically divergent sequences. It appears that the population of DNA-binding POU-domain proteins consist of two physically linked regions that are either invariant or conservatively made, while the region of greatest variability is near the middle segment of the proteins. An iterative fine-tuning process that commenced with visual inspection of the alignment and selected excision of conserved sequence-specific elements was capable of building a highly discriminating family signature.

Although our POU-discriminating extended patterns can be regarded as a supplement to either PROSITE or PRINTS signatures, full potency and a better diagnostic performance are gained from the mutual context provided by neighbouring conservatively substituted residues. Two conserved 'islands' encoding residues 466–521 and 544–598 (of the alignment) which satellite the PROSITE conservation marks (Figure 3b) were translated into two single consensus expressions and were included in our collection of DNA-binding motifs. The latter fragments not only allow specific family identification but also capture the structural information embedded within the sequence data. Indeed, key DNA contacts from a number of well-characterized residues [SWISS-PROT code, OCT1_HUMAN; PDB code, 1OCT (Herr *et al.*, 1988)] contribute to our final composite POU signature. For example, the amino acids of H3 DNA recognition helix (section 480–499 of the alignment), an extensive set of residues involved in phosphate contacts (Johnson and Hirsh, 1990), and a number of adjacent residues are required in protein dimerization (Johnson and Hirsh, 1990).

Finally, searching the data source of SWISS-PROT with our composite POU signatures provides the opportunity to validate their diagnostic performance. The designed family discriminators were able to harvest the whole set of sequences (66) that contain both the POU_S- and POU_H-specific domains, while leaving the sub-family members with only one of the domains.

Cold-shock domain

In *E.coli*, an environmental temperature downshift from 37 to 10–15°C results in a 4–5 h lag phase, after which growth is resumed at a reduced rate (Obokata *et al.*, 1991). During this lag period, changes in intracellular protein production and function in response to cold-stress conditions lead to the expression of around 13 proteins, which contain specific DNA-binding regions (Tafuri and Wolffe, 1990). These so-called 'cold-shock' proteins are thought to be a specific physiological mechanism of many organisms for acclimating to thermal stress, possibly by condensation of the chromosome and organization of the prokaryotic nucleoid (Obokata *et al.*, 1991). The most abundant of these factors now found in both prokaryotes and eukaryotes contain a conserved domain of about 70 amino acids, known as the 'cold-shock domain' (CSD), responsible for the specific binding to DNA (Wistow, 1990; Landsman, 1992). Only recently, several crystal structures of CSD proteins have been determined to uncover a similar β-barrel-like architecture. In addition, functional annotation of their structural elements allowed the location of single DNA-binding sites to be established (Feng *et al.*, 1998).

As a signature pattern for the CSD domain, a conserved region, which is located in its N-terminal section, was extracted from PROSITE pattern library encoding by COLD_SHOCK regular expression (PS00352) (Figure 3c). A three-element fingerprint from the PRINTS resource provides another signature for the cold-shock proteins; the first two motifs span the region encoded by PROSITE.

From a multiple alignment of 41 sequences collated from SWISS-PROT, we selected the three most highly invariant segments. Building on the idea of exploiting the maximum information retained in the complete family alignment, we merged the three 'islands' of conservation into a composite pattern (Figure 3c). The latter was translated into a consensus expression during a manual amalgamation process. The extended seed motif was used to search the source database (SWISS-PROT); the query returned revealed seven more sequences to be members of the CSD family. Two of them were excluded as fragments, while the remaining five full-length sequences were incorporated into the family data comprising a final true set of 46 members. Significantly, the order of sequence motifs along the amino acid sequence is distinct for the whole set of cold-shock binding proteins.

Supporting evidence for the high potency of our manually

designed CSD signature comes from structural analyses of the cold-shock DNA complexes. The results demonstrated that a single DNA-binding epitope is localized within three consecutive β-strands while surface-exposed aromatic or basic groups of residues are required for the formation of cross-linked CSD protein–DNA complexes (SWISS-PROT code, CSPA_ECOLI; PDB code, 3MEF) (Schnuchel *et al.*, 1993; Feng *et al.*, 1998). Invariably, the antiparallel β-pleated sheet contains the three conserved sequence motifs that comprise our composite CSD family signature (SWISS-PROT code, YB1_HUMAN). In particular, Lys282, Trp283 and Phe284 (in the β-1-strand) of the first conserved segment, Lys288, Phe290 and Phe292 (in the β2-strand) of the second invariable block and Phe303, Phe304 and His305 (in the β3-strand) constituting the third conserved cluster (alignment numbering) form part of the cold-shock nucleic acid-binding epitope.

In the light of the three different DNA-binding classes discussed above, it is remarkable that conserved residues have conserved roles when they appear at corresponding positions of a particular DNA-binding motif. Although there is no general code for DNA sequence recognition, there may be context-dependent codes for particular DNA-binding motifs.

The database WWW site, data presentation, access and usage

A WWW online server (at http://kronos.biol.uoa.gr/~mariak/dbDNA) has been set up for the distribution of the DNA-binding proteins' resource- and family-specific search system. The database can be accessed in two different modes: (i) keyword search and (ii) DNA-binding pattern recognition. Individual database records or family lists can be retrieved based on keyword search at http://kronos.biol.uoa.gr/~mariak/DB-dna/dbprocess_form.cgi. In particular, the database can be searched using sequence identifiers of SWISS-PROT (code or accession number) – the sequence (in FASTA format) (Pearson and Lipman, 1988) will be extracted from the current version of the DNA-binding sequence library stored in our server. The search results are returned to users as HTML documents. Database records and family information can also be obtained based on the result of a pattern-recognition tool. The program searches the user-supplied query sequence against a collection of DNA-binding sequence-specific motifs and returns family classification with links to DnaProt resource.

In more detail, the DNA-binding proteins' server presents the family classification of any DNA-binding protein sequence as a coloured and hyperlinked bead for each principal type of data. Starting from the top of the Web-based entry page, the identified DNA-binding class name is given, while a number of individual database files provide condensed information and detailed reports for the sequence entry with fields such as unique protein and family identifiers, as well as attributes such as accession number and sequence title.

In the following section of the Web page, sequence data are displayed in FASTA format together with a hyperlinked 'motif flag'. With a simple click on the relevant link, a new window appears presenting a pictorial representation of the sequence entry annotated by its constituent motif or motifs. A finer interpretation of the DNA-binding pattern composition and architecture is thus possible. Following the description of the primary data type there is a list of a few, randomly selected, members of the respective DNA-binding class, named by their accession number and SWISS-PROT identification codes and hyperlinked to other major sequence resources.

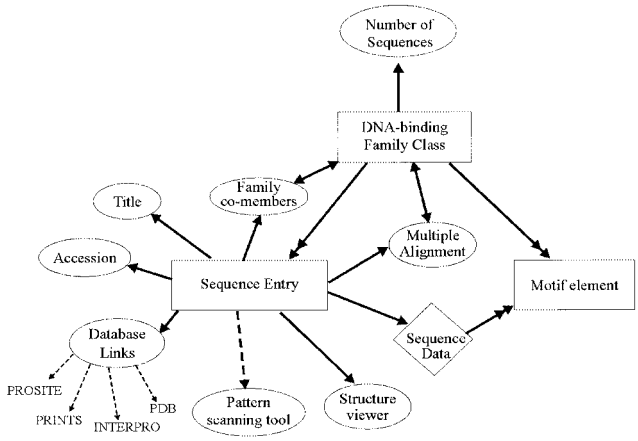

Fig. 4. The DnaProt entity relationship diagram. This diagram depicts the resource data space modelled using three entities: DNA-binding family class, motif element and sequence entry (rectangular boxes). Each entity has a relationship with another, as represented by a connecting arrow. A single arrow denotes a 'single relationship' and a double arrow-head denotes a 'many relationship', i.e. a sequence has more than one DNA-binding motif. Dashed arrows indicate 'external' cross-links of DnaProt with other related databases. The diamond represents a specific assignment given to an entity, in this case 'Sequence Data'. Ellipses denote specific entity attributes.

Currently there are cross-references that link the database to SWISS-PROT, and also to the pattern databases of PROSITE and PRINTS. We also started to add cross-references from the database to INTERPRO, a newly launched resource aiming to provide a central compendium of family and domain descriptions around which the high-level documentation from PRINTS and PROSITE will be united (Apweiler *et al.*, 2000). Where available, the PDB code of the corresponding protein is provided, together with a thumbnail picture of its three-dimensional structure extracted from a standard structure-image library (IMB Jena) (Reichert *et al.*, 2000). Seeking a better dissemination of information on 3D structures and with emphasis on visualization and structural–functional annotation, we also utilize a 3D software tool for the inspection of single structure conformations. Using the SCAR viewer (G.Palaios and S.J.Hamodrakas, unpublished work), the user can visualize the structure concealed in the associated PDB file while obtaining an interactive view of the structural information: number of protein chains, ligands, metal ions, secondary structural elements, fold cartoons, principal interactions, etc.

Further enhancements of database utility are the direct 'ScanProsite' submission and the disposal of multiple alignment family records. Generally, the ScanProsite suite (http://www.expasy.ch/sprot/scnpsite.html) allows the user to scan a query sequence (from SWISS-PROT or provided by the user) for the occurrence of patterns stored in the PROSITE database. To evaluate the one-to-many relationships within the members of each DNA-binding family, a Postscript (ps) version (Adobe Systems, 1985) of their multiple alignment outputs is available. Each ps alignment record has been constructed automatically using the ClustalW program (Thompson *et al.*, 1994).

To gain an overview of the different members included in a family, it is also possible to list all family entries, hyperlinked for subsequent annotation, in one view – the page also provides summary information and count of family members. The underlying model, which constitutes the core of DnaProt resource, is illustrated in Figure 4.

It should be emphasized that the present resource is not yet a sophisticated database. A few new features will be

implemented in the subsequent database releases including a search tool to allow sequence similarity queries against our collection of DNA-binding protein sequences, a JAVA applet to provide an interactive graphical interface for the pictorial representation of protein sequence functional elements and a novel automated tool to allow the prediction of DNA-binding proteins from sequence alone. Lastly, we are planning to develop a database management system to allow rapid, automated and precise data processing aiming to incorporate newly identified DNA-binding entries into our classification scheme.

Discussion

The major objectives of the DNA-binding protein Web-based resource are to maximize family information retrieval and help reveal the relationships within the various functional DNA-binding classes. The classification system, being implemented in a Web-based management structure, is available for online search against a pattern library containing 'signatures' specific for each family of DNA-binding proteins and retrieval of individual entries from a WWW server. Our initial targets include the ability of free data access, the ease of data retrieval from other existing related databases and frequent updates in accordance with other major family database releases.

Continuing efforts will be made to establish more rapid, automated and precise data processing procedures, in order to retain the high level of data accuracy in the database whilst keeping pace with the accelerating rate of global genome sequence discovery. We shall endeavour to maintain as much manual review of records as is reasonably possible, since we find this is a key to pruning out inconsistencies and standard record details and that are endemic in many data sources.

To support full-scale genomic annotation efforts, DnaProt can be used for direct searching against DNA-binding classes and for motif detection. Clearly, a significant majority of the newly characterized genome sequences could be assigned to one of the 33 well-defined functional categories. However, our need for a more detailed and systematic classification of protein sequences collected in advanced resources is apparent from the unprecedented variety of data flow with no functional or structural annotation, hence our insight into several phylogenetic and evolutionary paths still remains hidden.

Conclusions

Building an extremely high-quality and ultimately fully comprehensive catalogue of DNA-binding protein sequences is useful as a principal tool to aid the study of family classification and its bewildering range of functional variations and evolutionary phenotypes.

In an era of information overload, where the use of computational tools is essential, diagnostic family fingerprints are required to design protein family databases and facilitate the retrieval of relevant information providing insights into protein structure and function. However, in view of the fallibility of the different pattern databases and given the overload of sequence information housed in different resources and the erroneous automatically derived functional predictions, reliable family memberships should amalgamate as much sequence information as possible.

Acknowledgements

We thank Vasilis Promponas and Theodore Liakopoulos for helpful discussions and exchange of information. This work was supported by the Greek General Secretariat of Research and Technology (Grant EKBAN 1.3.4).

References

Adobe Systems, Inc. (1985) *PostScript Language Reference Manual.* Addison-Wesley Press, Reading, MA.
Altschul,S.F., Gish,W., Miller,M., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
Apweiler,R. *et al.* (2000) *Bioinformatics*, **16**, 1145–1150.
Assa-Munt,N., Mortishire-Smith,R.J., Aurora,R., Herr,W. and Wright,P.E. (1993) *Cell*, **73**, 193–205.
Attwood,T.K., Beck,M.E., Flower,D.R., Scordis,P. and Selley,J. (1998) *Nucleic Acids Res.*, **26**, 304–308.
Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, **24**, 217–221.
Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.E.,Jr, Brice,M.D, Rodgers,J.R, Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Bork,P., Ouzounis,C. and Sander C. (1994) *Curr. Opin. Struct. Biol.*, **4**, 393–403.
Branden,C.-I. and Tooze,J. (1999) *Introduction to Protein Structure*. 2nd edn. Garland Publishing, Levittown, PA.
Choo,Y. and Klug,A. (1997) *Curr. Opin. Struct. Biol.*, **7**, 117–125.
Feng,W., Tejero,D.E., Zimmerman,M., Inouye,G.T. and Montelione,E. (1998) *Biochemistry*, **37**, 10881–10896.
Harrison,S.C. (1991) *Nature*, **353**, 715–719.
Herr,W. *et al.* (1988) *Genes Dev.*, **2**, 1513–1516.
Johnson,W.A. and Hirsh,J. (1990) *Nature*, **343**, 467–470.
Klemm,J.D., Rould,M.A., Aurora,R., Herr,W. and Pabo,C.O. (1994) *Cell*, **77**, 21–32.
Koonin,E.V. (1997) *Curr. Biol.*, **7**, R656–R659.
Landsman D. (1992) *Nucleic Acids Res.*, **11**, 2861–2864.
Livingstone,C.D. and Barton,G.J. (1993) *Comput. Appl. Biosci.*, **9**, 745–756.
Muller,C.W. and Herrmann,B.G. (1997) *Nature*, **389**, 884–888.
Neidhardt,F.C., Curtiss,R.,III, Ingraham,J.L., Lin,E.C.C., Low,K.B., Magasanik,B., Reznikoff,W.S., Riley,M., Schaechter,M. and Umbarger,H.E. (eds) (1996) *Escherichia Coli and Salmonella. Cellular and Molecular Biology.* 2nd edn. ASM Press, Washington, DC.
Obokata,J., Ohme,M. and Hayashida,N. (1991) *Plant Mol. Biol.*, **17**, 953–955.
Papaioannou,V.E. and Silver,L.M. (1998) *Bioessays*, **20**, 9–19.
Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
Promponas,V.J., Enright,A.J., Tsoka,S., Kreil,D., Leroy,C., Hamodrakas,S.J., Sander,C. and Ouzounis,C.A. (2000) *Bioinformatics*, **16**, 915–922.
Reichert,J., Jabs,A., Slickers,P. and Suhnel,J. (2000) *Nucleic Acids Res.*, **1**, 246–249.
Schnuchel,A., Wiltscheck,R., Czischzi,M., Herrler,M., Willmsky,G., Graumann,P., Marahiel,M.A. and Holak,T.A. (1993) *Nature*, **364**, 169–171.
Tafuri,S.R. and Wolffe,A.P. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 9028–9032.
Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
Wattler,S., Russ,A., Evans,M. and Nehls,M. (1998) *Genomics*, **48**, 24–33.
Wistow G. (1990) *Nature*, **26**, 823–824.
Wu,C. H, Zhao S. and Chen,H.L. (1996) *J. Comput. Biol.*, **3**, 547–561.

Received October 17, 2000; revised February 23, 2001; accepted May 8, 2001