



Identifying property based sequence motifs in protein families and superfamilies: application to DNase-1 related endonucleases

Venkatarajan S. Mathura, Catherine H. Schein and Werner Braun*

Sealy Center for Structural Biology, Department of Human Biological Chemistry and Genetics, University of Texas Medical Branch, Galveston, TX 77555-1157, USA

Received on September 23, 2002; revised on January 2, 2003; accepted on January 28, 2003

ABSTRACT

Motivation: Identification of short conserved sequence motifs common to a protein family or superfamily can be more useful than overall sequence similarity in suggesting the function of novel gene products. Locating motifs still requires expert knowledge, as automated methods using stringent criteria may not differentiate subtle similarities from statistical noise.

Results: We have developed a novel automatic method, based on patterns of conservation of 237 physical–chemical properties of amino acids in aligned protein sequences, to find related motifs in proteins with little or no overall sequence similarity. As an application, our web-server MASIA identified 12 property-based motifs in the apurinic/apryrimidinic endonuclease (APE) family of DNA-repair enzymes of the DNase-I superfamily. Searching with these motifs located distantly related representatives of the DNase-I superfamily, such as Inositol 5'-polyphosphate phosphatases in the ASTRAL40 database, using a Bayesian scoring function. Other proteins containing APE motifs had no overall sequence or structural similarity. However, all were phosphatases and/or had a metal ion binding active site. Thus our automated method can identify discrete elements in distantly related proteins that define local structure and aspects of function. We anticipate that our method will complement existing ones to functionally annotate novel protein sequences from genomic projects.

Availability: MASIA WEB site: <http://www.scsb.utmb.edu/masia/masia.html>

Contact: werner@newton.utmb.edu

Supplementary information: The dendrogram of 42 APE sequences used to derive motifs is available on http://www.scsb.utmb.edu/comp_biol.html/DNA_repair/publication.html

INTRODUCTION

One of the most challenging goals of the genome sequencing projects is to functionally annotate novel gene products (Kelley *et al.*, 2000; Nagano *et al.*, 2002; Norin and Sundstrom, 2002; Rison *et al.*, 2000; Urushihara, 2002; Waterston *et al.*, 2002). A sequence can be recognized as a homologue of a known protein if the pair-wise sequence identity/similarity exceeds a statistically derived threshold (e.g. more than 30% sequence identity or an *E*-value less than 0.001) (Chothia and Lesk, 1986). These global criteria identify only a small fraction of proteins known to be functionally related, as amino acids patterns are differently conserved. Sequence profile searches (Bowie *et al.*, 1991; Gribskov and Veretnik, 1996; Mehta *et al.*, 1999; Rychlewski *et al.*, 2000; Schaffer *et al.*, 1999; Yona and Levitt, 2002) and hidden Markov models (Eddy, 1998; Gough and Chothia, 2002; Martelli *et al.*, 2002) generate position-specific fingerprints of the amino acid sequences in protein families and can identify distantly related proteins. However, the optimal choice of parameters for high-sensitivity/specificity depends on the expert user. A further complication is that enzymes often combine functional elements to create a specific catalytic center. These elements, due to crossover events, may not occur in the same linear fashion in the sequence of related proteins and are not found with global profiles. Statistically derived matrices based on allowed substitution of amino acids are not designed to detect conservation of physical–chemical properties (Falquet *et al.*, 2002). Their insensitivity is demonstrated by the failure of PSI-BLAST to identify either known homologue of APE1, bovine DNase-I or synaptojanin, a member of the Inositol 5'-polyphosphate phosphatase (IPP) family sequence, in the ASTRAL40 structural database (Brenner *et al.*, 2000; Chandonia *et al.*, 2002) with default parameters.

We implemented a method to automatically identify physical–chemical property-based motifs (PCP motifs) (Venkatarajan and Braun, 2001) in our MASIA program

*To whom correspondence should be addressed.

(Zhu *et al.*, 2000; <http://www.scsb.utmb.edu/masia/masia.html>), that can be used to detect common elements in proteins that share no global sequence identity. We used this improved MASIA tool to generate PCP motifs in the DNA repair protein family APE and then searched for proteins in the ASTRAL40 database containing similar sequences. The highest scoring proteins were in the DNase-I like SCOP-superfamily of APE, demonstrating that our method can find non-trivial relationships between distantly related members within superfamilies. Other high-scoring proteins were from different SCOP classifications (Lo Conte *et al.*, 2002), but shared functions with the APE/DNase-I/IPP superfamily, including phosphatase activity and/or metal ion binding. Details of the structural and functional roles of the MASIA motifs of the APE family are described elsewhere (Schein *et al.*, 2002).

SYSTEMS AND METHODS

Conservation of the physical–chemical components

Quantitative descriptors E^1 to E^5 for amino acid properties and their physical interpretation were deduced from a comprehensive list of 237 physical–chemical properties (Venkatarajan and Braun, 2001). The five components E^1 to E^5 will in general be differently conserved at each position of the protein family. We measure the conservation by standard deviations σ_k^i of the values E^1 to E^5 and by the relative entropy \mathfrak{R}_k^i , also referred to as Kullback–Leibler distance (Kullback and Leibler, 1951). The component index i varies from 1 to 5. The quantities σ_k^i and \mathfrak{R}_k^i are calculated for every residue position k in the multiple sequence alignment of the protein family.

Five equally spaced bins characterize the distributions of the E^1 to E^5 values for each of the components. The difference between the observed distribution for the component E^i at position k and the background distribution can be calculated by the relative entropy \mathfrak{R}_k^i :

$$\mathfrak{R}_k^i = \sum_{b=1}^5 Q(X^b) \log_2 \left(\frac{Q(X^b)}{P(X^b)} \right). \quad (1)$$

$Q(X^b)$ is the observed fraction of the component i in the bin b and $P(X^b)$ is the corresponding background frequency. For significantly conserved properties the distributions of the component values E^1 to E^5 are narrower than the background distributions derived from the *a priori* occurrences of amino acids; i.e. we expect low standard deviations and high relative entropy values. If the distributions of the observed frequencies of the components are equal to that of the background frequencies, then \mathfrak{R} will be zero, otherwise it will be positive. High relative entropy values indicate a significant difference between the observed frequencies of distributions in a column and the *a priori* background distribution.

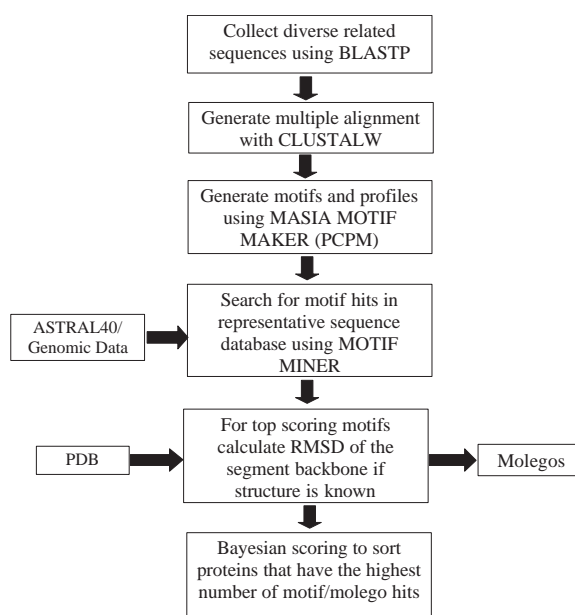


Fig. 1. A flow chart showing steps involved in defining motifs and molegos for a protein family. These can be used to locate proteins with areas of homology in protein databases.

Identification of motifs in a protein family

We define motifs as contiguous positions in a multiple alignment where residues are significantly conserved according to one of the principal components E^i , that is those sites k , where the relative entropy \mathfrak{R}_k^i of at least one component E^i is above a \mathfrak{R} -cutoff value. We introduced a minimum length cutoff (L -cutoff) to define sequence motifs of sufficient length, and a G -cutoff parameter that restricts the maximum number of insignificant positions between two significant positions in a motif. Default values of the parameters \mathfrak{R} -cutoff, L -cutoff and G -cutoff were determined empirically.

Generating MASIA motifs for the APE protein family

Homologues of human APE1 with E -values less than 0.001 were identified in the NCBI protein sequence database using the BLASTP search engine (Altschul *et al.*, 1997). Sequences from 42 organisms ranging from prokaryotes to eukaryotes were selected (see supplementary material) after discarding hypothetical APE-like proteins. The taxonomic classification was used to avoid excessive redundancy. Sequences were aligned with CLUSTALW release 1.8 (Higgins and Taylor, 2000) using the GONNET similarity matrix (Benner *et al.*, 1994), with an opening gap penalty of 10.0 and gap extension penalty of 0.2. The sequence alignment was used as the input for the MASIA program for motif identification, as shown in the flow chart in Figure 1.

Each motif identified by MASIA is quantitatively expressed as a profile, consisting of the average values, standard deviations and the relative entropies for each vector E^1-E^5 at each position (column in the initial alignment) in the motif. This profile was used to search for similar motifs in the ASTRAL40 sequence database (Brenner *et al.*, 2000; Chandonia *et al.*, 2002), which consists of representative sequences corresponding to the SEQRES record of PDB files and classified in SCOP (Lo Conte *et al.*, 2002). The pair-wise identity among the 3635 sequences is less than 40%.

Scoring method for matching motifs in a protein sequence

A Lorentzian based scoring scheme is used to measure the quality of fit for a query sequence to a profile at position k for the component vector i . The motif profile at a significant position k (defined by the relative-entropy cutoff) consists of average of component magnitudes $\langle E_k^i \rangle$ and the standard deviation σ_k^i . If E^i is the magnitude of PCP component i observed in the query sequence, the Z value is calculated:

$$Z_k^i = \left(\frac{E^i - \langle E_k^i \rangle}{W\sigma_k^i + \Phi} \right) \quad (2)$$

$$S_k^i = \left(\frac{1}{1 + Z_k^i * Z_k^i} \right) \quad (3)$$

Where W is the weight for standard deviation (set to 1.5 in the current calculations) and Φ is a small positive shift (set to 0.001) added to the denominator to prevent overflow during calculation when σ_k^i is zero. The individual score for each component was then added for significantly conserved property components along the length of the motifs to obtain a window score S_w and a maximum possible score S_{\max} calculated by adding 1 for every significant position:

$$S_w = \sum_{i,k} S_k^i \quad (4)$$

$$S_{\max} = \sum_{i,k} 1 \quad (5)$$

The final fractional score for the window is

$$S_{\text{fraction}} = \frac{S_w}{S_{\max}} \quad (6)$$

A Bayesian method to score proteins based on PCP motif similarity

We apply the Bayesian method to decide if a given score S for a segment in an arbitrary query sequence is a sufficient match to an APE motif. The conditional

probability $P(X \in APE | S)$ that the query sequence X contains an APE motif for a given score S is given by Bayes theorem:

$$P(X \in APE | S) = \frac{P(S | X \in APE) \bullet P(APE)}{P(S)} \quad (7)$$

$P(S | X \in APE)$ is the probability of finding a motif with score S in the APE family. $P(S)$ is the probability of finding a motif with similar score in all proteins in the ASTRAL40 database, and $P(APE)$ is the probability of finding APE sequences in the ASTRAL40 database. The empirical distributions of scores in the APE family and ASTRAL40 are approximated as a Gaussian distribution (see results for the distribution of scores in ASTRAL40):

$$P(S | X \in APE) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(S-\bar{S}_{APE})^2} \quad (8)$$

Here \bar{S}_{APE} is the average score for the motif in APE and σ is the corresponding standard deviation. The probability of finding a motif with similar score in the ASTRAL40 database is

$$P(S) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(S-\bar{S}_{AST})^2} \quad (9)$$

Substituting Equations (8) and (9) into Equation (7), and simplifying by assuming that the standard deviation of scores for APE and ASTRAL40 are similar (from Table 1), we obtain

$$P(X \in APE | S) = e^{\frac{\Delta\bar{S} \bullet (S-\bar{S})}{\sigma^2}} P(APE) \quad (10)$$

Here \bar{S} is the average and $\Delta\bar{S}$ is the difference between average scores of a motif in APE and ASTRAL40 database. The conditional probability that an APE motif m was found in a query sequence X with a score less than or equal to an observed score S_m is given by

$$\begin{aligned} P(X \in APE | S \leq S_m) \\ = P(APE) \bullet \int_0^{S_m} e^{\frac{\Delta\bar{S}_m \bullet (S-\bar{S}_m)}{\sigma_m^2}} dS \end{aligned} \quad (11)$$

$$\begin{aligned} P(X \in APE | S \leq S_m) \\ = P(APE) \bullet \left[\frac{\sigma_m^2}{\Delta\bar{S}_m} e^{\frac{-\Delta\bar{S}_m \bullet \bar{S}_m}{\sigma_m^2}} \right] \bullet \left[e^{\frac{\Delta\bar{S}_m \bullet S_m}{\sigma_m^2}} - 1 \right] \end{aligned} \quad (12)$$

We compute a total sequence score S_x over all 12 motifs

$$S_x = \sum_{m=1}^{12} \log_2 [P(X \in APE | S \leq S_m)] \quad (13)$$

$$\begin{aligned} S_x = \sum_{m=1}^{12} \log_2 \left[\frac{\sigma_m^2}{\Delta\bar{S}_m} e^{\frac{-\Delta\bar{S}_m \bullet \bar{S}_m}{\sigma_m^2}} \right] + \sum_{m=1}^{12} \log_2 \left[e^{\frac{\Delta\bar{S}_m \bullet S_m}{\sigma_m^2}} - 1 \right] \\ + 12 \log_2 P(APE) \end{aligned} \quad (14)$$

Table 1. The 12 motifs of APE family defined by MASIA

Motif No.	Query sequence (human APE)	(S1)	(S2)	(S3)	(S4)	42 APE score ave. \pm std. dev	ASTRAL40 ave. \pm std. dev
1	62 LKICSWNVVDGLRA 74	0.91*	0.90*	0.63*	0.73*	0.87 \pm 0.05	0.56 \pm 0.05
2	89 PDILCLQETK 98	0.96*	0.93*	0.84*	0.70*	0.92 \pm 0.04	0.61 \pm 0.07
3	125 KEGYSGVGLLSRQCP 139	0.91*	0.86*	0.60	0.66	0.85 \pm 0.06	0.64 \pm 0.05
4	145 GIGDEEHDQEGRVIVAEFDSFVL 169	0.94*	0.77*	0.71	0.81	0.84 \pm 0.09	0.67 \pm 0.07
5	171 YVPNA 175	0.96*	0.96*	0.68	0.86	0.94 \pm 0.06	0.68 \pm 0.13
6	181 RLEYRQRW 188	0.80*	0.70*	0.78	0.77	0.74 \pm 0.06	0.67 \pm 0.05
7	204 PLVLCGDLNVAH 215	0.96*	0.88*	0.82*	0.78*	0.90 \pm 0.04	0.55 \pm 0.08
8	231 GFTPQERQGFEL 243	0.96*	0.91*	0.78	0.73	0.87 \pm 0.09	0.70 \pm 0.07
9	247 VPLADSF 254	0.96*	0.93*	0.70	0.83	0.91 \pm 0.08	0.74 \pm 0.11
10	264 YTFWTYM 270	0.86*	0.77*	0.61	0.70	0.84 \pm 0.08	0.61 \pm 0.06
11	274 RSKNVGWRLDYFLLSHSL 291	0.92*	0.89*	0.56	0.64	0.90 \pm 0.04	0.54 \pm 0.07
12	306 GSDHCPI 312	0.93*	0.94*	0.88*	0.83*	0.92 \pm 0.03	0.52 \pm 0.09

The 12 motifs of APE family defined by MASIA were used as query sequences to locate the best scoring windows in the sequences of human APE (1hd7.pdb; S1), *E.coli* exonuclease (1ako.pdb; S2), bovine DNase I (2dnj.pdb, S3) and the IPPP domain of synaptojanin from yeast (1i9y.pdb, S4). These S values are compared to the distribution of values for the highest scoring window over all sequences in the APE family belonging to APE and in the ASTRAL40 database. *Structurally equivalent motifs, based on FSSP/DALI alignments for the four PDB files

The first and the third terms are constant for a given database so we use only the middle term for our final scoring and ranking of sequences.

Comparison of the property based motif search to PSI-BLAST and BLOCKS

We compared the sensitivity of our method to find proteins related to the APEs in the ASTRAL40 database with that of two versions of PSI-BLAST with default parameters (E -value of 0.005), one locally installed (v 2.2.1) and the other on the web at NCBI (Altschul *et al.*, 1997; Schaffer *et al.*, 2001). To enhance the ability of PSI-BLAST to build a profile, the 42 sequences from the APE family, used in the construction of the motifs, were added to the 3635 sequences from the ASTRAL40 database. We used the human APE sequence as query and ran up to five iterations of PSI-BLAST. We also searched for APE related sequences in the BLOCKS database (Henikoff *et al.*, 2000, 1999) with the default search engine (Henikoff and Henikoff, 1994).

RESULTS

The *a priori* distributions of the property components E^1 to E^5

We compared the distribution of the 20 amino acids according to each of the five property components based on an *a priori* distribution derived from their relative occurrence in the SWISSPROT database (Bairoch and Apweiler, 2000). The distributions for the five vector components were not uniform, as illustrated for E^1 to E^4 in Figure 2. For example, the distribution for the component E^1 , which correlates well with most hydrophobicity

scales (Venkatarajan and Braun, 2001), has the most populated bins at the extreme positive and negative values. If the E^1 values in a given column of a multiple alignment are concentrated in a narrow range, especially towards the middle range, the distribution of E^1 values would differ from the *a priori* distribution and a high relative entropy value is calculated. A physical interpretation is that the residue hydrophobicity is constrained at that position of the protein family during evolution.

In contrast, the component values for E^2 , which correlates best with the size/molecular weight of the residues side chains, has a different distribution, with most residues concentrated in the third bin ($5.2 \geq E^2 > 0.2$), and for the fourth vector, which correlates with the natural frequency of occurrence of amino acids (codon degeneracy), bin occupancy decreases as the value of the component E^4 increases. A more detailed discussion on the physical interpretation and importance of the vector components E^1 – E^5 is available elsewhere (Venkatarajan and Braun, 2001).

Motifs in the APE family

The APE 1 protein family consists of apurinic/aprimidinic endonucleases and exonucleases. The PCP macro of our MASIA program identifies 12 motifs of various lengths with the cutoff values for entropy, $\mathfrak{N} = 1.25$, for insignificant positions within a motif, $G = 2$, and minimum length, $L = 4$ (Table 1). Most of the motifs are located in the β -strands in the core of the protein. All residues known to be involved in metal ion binding of APE1, including 68N, 96E, 210D, 212N, 308D and 309H are part of the 12 motifs. The three PROSITE motifs (Falquet *et al.*, 2002) defined for APE correspond to the motifs 2, 9, 10 and 11 defined by MASIA.

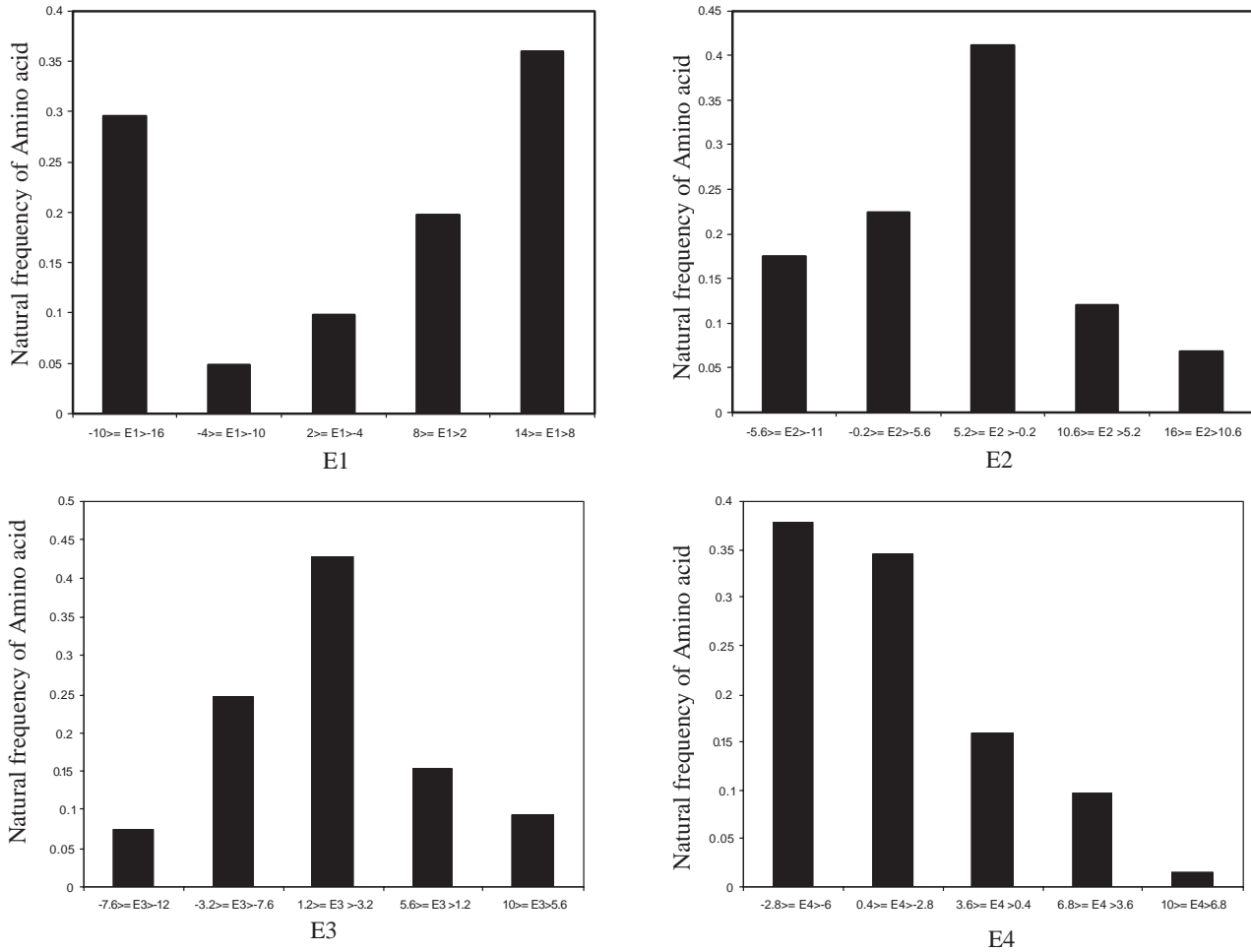


Fig. 2. Distribution of the values for each vector for the naturally occurring amino acids. The frequency of occurrence of each amino acid in SWISSPROT (release 40) was used.

Figure 3 shows a qualitative description of a motif, where + or - indicates significant components at the given position with the average values and * means an insignificant component.

The 12 motifs are differentially conserved in the APE family

Scanning the individual APE sequences with the motif profiles identifies most of the motifs in an equivalent position as in the multiple sequence alignment. Motifs 1, 2, 7, 11 and 12 are particularly well conserved and have consistently high score (*S*) values. However, motifs 4, 6 and 10 score lower in enteric pathogenic bacteria such as *H.influenza*, *E.coli*, *S.enterica*, *Y.pestis* and *V.cholerae*, which might be related to the observed differences in the activity compared to eukaryotes (Schein *et al.*, 2002) (see dendrogram of the APE family in supplementary material).

	P	D	I	L	C	L	Q	E	T	K
E1	*	+	-	-	-	-	+	+	*	*
E2	*	-	*	*	*	*	-	-	*	-
E3	+	*	*	*	*	*	*	-	*	*
E4	*	+	*	*	+	*	+	+	*	-
E5	*	+	*	*	*	*	-	+	*	-

Fig. 3. Qualitative representation of motif 2. The magnitude of a vector is shown as a positive or negative symbol depending on the average value at that position in the multiple alignment. A ‘*’ indicates the relative entropy is less than 1.25, and the position will not be scored.

Distribution of scores in ASTRAL40

Figure 4 compares distribution of the highest scores for all motifs in each of the 3635 sequences in the ASTRAL40 database (solid line) to the target scores in the APE family

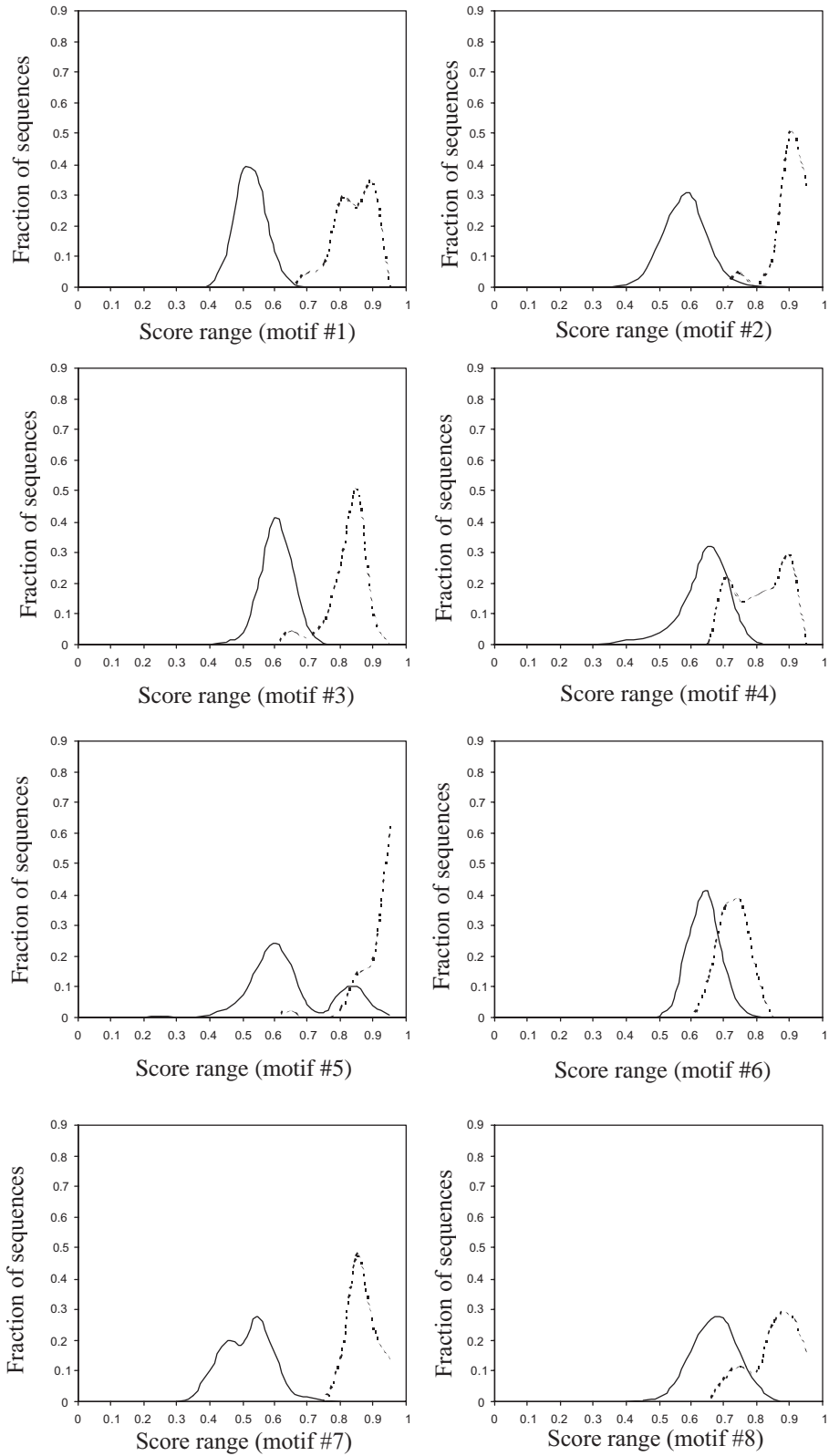


Fig. 4. Distribution of the highest scoring window for each motif in the 3635 sequences in a specific and non-specific database. The dark line shows the distribution of scores from the ASTRAL40 database and the dotted lines are the scores for the 42 APE sequences.

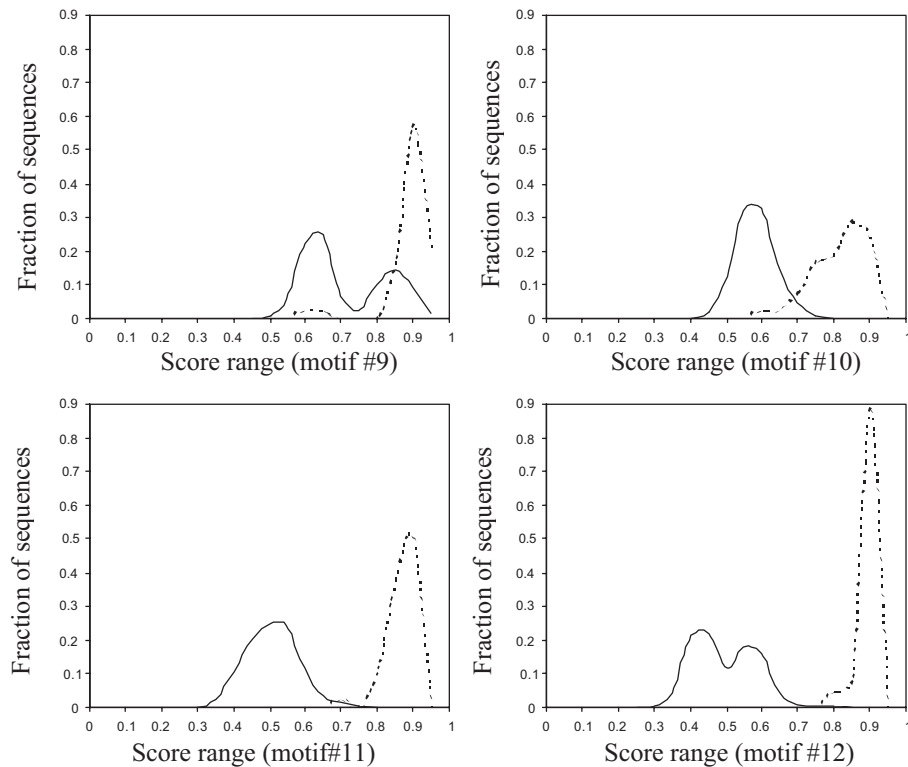


Fig. 4. Continued.

(dashed line). Both distributions can be approximated as a Gaussian function. The scores for motifs 1, 2, 3, 5, 7, 11 and 12, most specific to the APE family, are clearly distinguished.

Ranking all proteins in ASTRAL40 according to the overall score for APE similarity (Equation (14)), we find all the known members of the DNase-like superfamily at the top (Table 2). The table indicates that close homologues should have an overall score greater than about 0.5. All sequences with scores between 0.35 and 0.5 were either phosphatases and/or contained a metal ion binding site. Two of the proteins (1MDA and 1EKM) have catalytic centers containing Cu (II), one has two Zn²⁺ ions (1QQ9), and another contains Fe (II) (1MPY).

Identification of APE protein family and superfamily using Molegos

If a 3D structure of the protein is available, we can combine sequence and 3D structural motifs to find related proteins (Schein *et al.*, 2002). We defined structurally related motifs, or molegos, as those segments with a fractional window score greater than 0.6 and a RMSD value less than 2.5 Å for C^α atoms (Table 3). As with the scoring method based only on sequence, the highest ranking sequences with scores >0.5 were known members of the

DNase-I/APE/IPP superfamily. Six proteins with no overall structural or sequence similarity to this superfamily had scores between 0.3 and 0.5. Of these, three contained one (1D09) or two Zn²⁺ (1QQ9 and 1ATL) ions in their active site, one (1D2N) contained Mg²⁺ and one Ca²⁺. We are now investigating the details of these similarities.

PSI-BLAST and BLOCKS search for APE family and DNase-I superfamily members

Identification of members of a superfamily using the currently available sequence profile methods is difficult. For example, PSI-BLAST searching, using a local program or NCBI web-based version with default parameters (*E*-value 0.005), detected members of the APE family in the non-redundant sequence database. However, neither version revealed DNase-I or IPP sequences even after several iterations. When the *E*-value was increased to 0.1, synaptojanin was revealed within the first iterations, but bovine DNase-I was only detected after four iterations along with more than 500 additional entries. PSI-BLAST also failed to recognize DNase-I or synaptojanin in the ASTRAL40 database, even when we added APE sequences to allow it to form a profile. The BLOCKS search engine did not recognize homology to DNase-I even when the *E*-value cutoff was extended to 100. In

Table 2. APE related sequences in the ASTRAL40 database

PDB ^a	Score in bits (fraction to the highest score)	Motifs found	SCOP ^b	EC ^c	Description
1HD7	1942 (1.00)	1,2,3,4,5,6,7,8,9,10,11,12	d.151.1.1	4.2.99.18	APE
1AKO	1861 (0.96)	1,2,3,4,5,7,8,9,10,11,12	d.151.1.1	3.1.11.2	Exonuclease III
2DNJ	1094 (0.56)	2,6,7,12	d.151.1.1	0.0.0.0	Deoxyribonuclease I
1I9Y	1056 (0.54)	1,4,5,6,7,9,12	d.151.1.2	0.0.0.0	Phosphatidylinositol phosphate Synaptojanin
1B3U	840 (0.43)	5,7,9,12	a.118.1.2	0.0.0.0	Regulatory domain of protein phosphatase
1MDA	814 (0.42)	6,9,11,12	b.69.2.1	1.4.99.3	Methylamine dehydrogenase
1MPY	797 (0.41)	7,9,12	d.32.1.3	1.13.11.2	Catechol 2,3-dioxygenase
1EKM	792 (0.41)	6,7,12	b.30.2.1	1.4.3.6	Copper amine oxidase
1YRG	737 (0.38)	2,9,12	c.10.1.2	0.0.0.0	GTPase RNA1
1QQ9	698 (0.36)	5,6,12	c.56.5.4	3.4.11.-	Amino peptidase

Motifs that scored higher than their average scores in the database were considered as hits and the sequences were ranked according to the bit score obtained for all motif hits (see Methods).

^aPDB code of the protein

^bSCOP code and d.151.1 is the DNaseI superfamily code

^cEnzyme commission classification number

contrast, our method identified these proteins clearly in the structural database (Tables 2 and 3).

DISCUSSION

Identifying and using PCP motifs and molegos

A search method was developed to locate elements with similar physical chemical motifs in distantly related proteins. We demonstrated that the PCP motifs generated by the new macro of our MASIA program from a multiple alignment of APE sequences correlated well with motifs identified by other methods, and our database search method efficiently located known homologues. Our Bayesian scoring method discriminated members of the DNase-I superfamily from the bulk of sequences in the representative ASTRAL40 database (Table 2). We also show that combining sequence and structural data effectively discriminates proteins, with no overall similarity that share partial function (metal binding) with the APE family.

E^1-E^5 vectors provide an alternative scoring method for evaluating homology

All the commonly used methods for genome sequence searching rely on similar, statistically derived scoring matrices (Altschul *et al.*, 1997; Kostich *et al.*, 2002). Frequently, the same scoring matrix is used to search for related sequences (for example, with BLAST), prepare a multiple alignment to analyze sequence conservation and to locate distant relatives of the family according to motif conservation. We have previously shown (Venkatarajan and Braun, 2001) that our five property vectors represent all known physical chemical properties, and provide an alternative to using the amino acid alphabet (Rigoutsos *et al.*, 2002) or selected physical-chemical properties

(Dubchak *et al.*, 1999) to identify homology. Our PCP-motifs complement existing methods for functional cross networking of protein families (Marcotte *et al.*, 1999; Marcotte, 2000; Overbeek *et al.*, 1999).

Shared PCP motifs and molegos identify DNase-I superfamily members

Despite their low overall sequence identity, APE, DNase-I and IPP families share a similar 3D structure and are members of a superfamily with a common SCOP designation (Lo Conte *et al.*, 2002). Our methods rapidly identified members of this superfamily based on their shared PCP motifs. The most conserved motifs, 1, 2, 7 and 12, were found in structurally equivalent regions as defined by FSSP/DALI (Holm and Sander, 1996) in an alignment for the DNase-I superfamily members. The structurally equivalent motifs, or molegos, identified by our program are boxed in the FSSP alignment of DNase-I superfamily in Figure 5. These motifs dictate the formation of a β -strand core that serves as the supporting architecture for metal ion binding and phosphorolysis by members of the APE/DNase-I/IPP superfamily (Schein *et al.*, 2002). No non-APE protein sequence in the ASTRAL40 database scored higher than the average total score (S_x) of 1772 bits calculated for the 42 APE sequences. Thus our method is highly sensitive and specific for detecting APE related sequences.

Proteins with highest scores relative to APE bind metal ions

We found a high proportion of metal ion binding proteins as the highest scoring proteins by searching for proteins that have similar sequences and 3D structure motifs (Table 3). We did not set out to locate only areas of the protein that dictated metal binding but simply to identify

Table 3. APE related sequences in the ASTRAL40 database

PDB ^a	Score in bits (fraction to the highest score)	MOLEGOS found	SCOP ^b	EC ^c	Description
1HD7	1942 (1.00)	1,2,3,4,5,6,7,8,9,10,11,12	d.151.1.1	4.2.99.18	APE (Mn/Mg/Pb)
1AKO	1831 (0.94)	1,2,3,5,6,7,8,9,10,11,12	d.151.1.1	3.1.11.2	Exonuclease III
2DNJ	1072 (0.55)	1,2,5,6,7,9,10,12	d.151.1.1	3.1.21.1	Deoxyribonuclease I
1I9Y	971 (0.50)	1,2,5,6,7,9,10,12	d.151.1.2		Phosphatidylinositol phosphate Synaptojanin
1QQ9	698 (0.36)	5,6,9,10,12	c.56.5.4	3.4.11.-	Aminopeptidase (Zn, Ca)
1ATL	633 (0.33)	5,6,9,10,12	d.92.1.9	3.4.24.42	Snake venom metalloprotease (Zn, Ca)
1D09	619 (0.32)	5,9,12	d.58.2.1	2.1.3.2	Aspartate carbamoyltransferase (Zn)
1D2N	613 (0.32)	5,6,8,9,12	c.37.1.13		N-ethylmaleimide of sensitive fusion protein (Mg)
1D0B	579 (0.30)	2,5,9,12	c.10.2.1		InternalinB LRR domain (Ca)
1EEM	571 (0.29)	5,6,8,12	a.45.1.1		Glutathione S-transferase

Motifs scoring with a fractional window score greater than 0.6 and RMSD less than or equal 2.5 Å were considered to be molegos. Sequences were ranked according to the bit scored obtained for all Molego hits. Metal ions present in the protein structures are indicated in brackets.

^aPDB code of the protein

^bSCOP code and d.151.1 is the DNaseI superfamily code

^cEnzyme commission classification number



Fig. 5. FSSP/DALI alignment of DNase-I family sequences with motifs defined by MASIA for the APE family underlined. Motif areas that are structurally equivalent (as defined by DALI) are boxed.

distant homologues of the APEs. However, as Table 3 indicates, all of the highest scoring proteins identified in the ASTRAL40 structural database, that is those that contained molegos most similar to those of the APE superfamily, use metal ion based catalysis.

CONCLUSION

We have developed and tested a new automated method for identifying protein motifs that are conserved according to physical–chemical properties in aligned sequences. PCP-profiles of these motifs can be used to locate distantly related proteins in sequence database. The 12 motifs identified by MASIA in the APE family include the signatures in the PROSITE database and all amino acids shown previously to be essential for function. The motif profiles successfully identified likely homologues of APE in a database, including several with no overall sequence or structural similarity. We also showed that combining sequence and structural data can locate proteins that share functional similarities. We believe that PCP motifs can play an important role in the functional annotation of proteins in genomic sequence projects.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Department of Energy (DE-FG-00ER63041), a Research Development Grant (#2535-01) of the John Sealy Memorial Endowment Fund, the U.S. Food and Drug Administration (FDA-U-002249-01) and the Advanced Research Program of the Texas Higher Education Coordinating Board. We thank Dr Numan Oezguen and Dr Tadahide Izumi for fruitful discussions and Ms Cynthia Orlea for assistance in preparing the manuscript.

REFERENCES

- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

- Bairoch,A. and Apweiler,R. (2000) The protein information resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, **28**, 45–48.
- Benner,S.A., Cohen,M.A. and Gonnet,G.H. (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.*, **7**, 1323–1332.
- Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.
- Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Res.*, **28**, 254–256.
- Chandonia,J.M., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2002) ASTRAL compendium enhancements. *Nucleic Acids Res.*, **30**, 260–263.
- Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
- Dubchak,I., Muchnik,I., Mayor,C., Dralyuk,I. and Kim,S.H. (1999) Recognition of a protein fold in the context of the structural classification of proteins (SCOP). *Proteins*, **35**, 401–407.
- Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
- Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
- Gough,J. and Chothia,C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.*, **30**, 268–272.
- Gribskov,M. and Veretnik,S. (1996) Identification of sequence pattern with profile analysis. *Meth. Enzymol.*, **266**, 198–212.
- Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
- Henikoff,S., Henikoff,J. and Pietrokovski,S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.
- Henikoff,S. and Henikoff,J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics*, **19**, 97–107.
- Higgins,D.G. and Taylor,W.R. (2000) Multiple sequence alignment. *Meth. Mol. Biol.*, **143**, 1–18.
- Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.
- Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.
- Kostich,M., English,J., Madison,V., Gheyas,F. and Wang,L. (2002) Human members of the eukaryotic protein kinase family. *Genome Biol.*, **3**, 43.
- Kullback,S. and Leibler,R.A. (1951) On information sufficiency. *Ann. Math. Stat.*, **22**, 79–86.
- Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
- Marcotte,E., Pellegrini,M., Thompson,M., Yeates,T. and Eisenberg, (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 25–26.
- Marcotte,E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.*, **10**, 359–365.
- Martelli,P.L., Fariselli,P., Krogh,A. and Casadio,R. (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, **18** (Suppl. 1), S46–S53.
- Mehta,P.K., Argos,P., Barbour,A.D. and Christen,P. (1999) Recognizing very distant sequence relationships among proteins by family profile analysis. *Proteins*, **35**, 387–400.
- Nagano,N., Orengo,C.A. and Thornton,J.M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.*, **30**, 741–765.
- Norin,M. and Sundstrom,M. (2002) Structural proteomics: developments in structure-to-function predictions. *Trends Biotechnol.*, **20**, 79–84.
- Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
- Rigoutsos,I., Huynh,T., Floratos,A., Parida,L. and Platt,D. (2002) Dictionary-driven protein annotation. *Nucleic Acids Res.*, **30**, 3901–3916.
- Rison,S.C., Hodgman,T.C. and Thornton,J.M. (2000) Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics.*, **1**, 56–69.
- Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles: strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
- Schaffer,A., Wolf,Y., Ponting,C., Koonin,E., Aravind,L. and Altschul,S. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
- Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
- Schein,C.H., Izumi,T., Oezguen,N., Feng,Y.L. and Braun,W. (2002) Total sequence decomposition and genomic cross-networking distinguishes functional modules in apurinic/aprimidinic endonucleases. *BMC Bioinformatics*, **3**, 37.
- Urushihara,H. (2002) Functional genomics of the social amoebae, *Dictyostelium discoideum*. *Mol. Cells*, **13**, 1–4.
- Venkatarajan,M.S. and Braun,W. (2001) New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *J. Mol. Model.*, **7**, 445–453.
- Waterston,R.H., Lander,E.S. and Sulston,J.E. (2002) On the sequencing of the human genome. *Proc. Natl Acad. Sci. USA*, **99**, 3712–3716.
- Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.
- Zhu,H., Schein,C.H. and Braun,W. (2000) MASIA: recognition of common patterns and properties in multiple aligned protein sequences. *Bioinformatics*, **16**, 950–951.