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/apyrimidinic 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

*To whom correspondence should be addressed.

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.
(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 (Venkataram 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 $\sigma^i_k$ of the values $E^1$ to $E^5$ and by the relative entropy $\mathfrak{R}^i_k$, also referred to as Kullback–Leibler distance (Kullback and Leibler, 1951). The component index $i$ varies from 1 to 5. The quantities $\sigma^i_k$ and $\mathfrak{R}^i_k$ 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 $\mathfrak{R}^i_k$:

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

$Q(X^b)$ 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.

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 $\mathfrak{R}^i_k$ 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 E1–E5 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 (Ei) and the standard deviation σi. If Ei is the magnitude of PCP component i observed in the query sequence, the Z value is calculated:

\[
Z_k^i = \left( \frac{E^i - \langle E^i \rangle}{W \sigma^i_k + \Phi} \right) \quad (2)
\]

\[
S_k^i = \left( \frac{1}{1 + Z_k^i \cdot Z_k^i} \right) \quad (3)
\]

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 σi 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 Sw and a maximum possible score Smax calculated by adding 1 for every significant position:

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

\[
S_{\text{max}} = \sum_{i,k} 1 \quad (5)
\]

The final fractional score for the window is

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

### 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 \(P(X \in \text{APE} \mid S)\) that the query sequence X contains an APE motif for a given score S is given by Bayes theorem:

\[
P(X \in \text{APE} \mid S) = \frac{P(S \mid X \in \text{APE}) \cdot P(\text{APE})}{P(S)} \quad (7)
\]

\(P(S \mid X \in \text{APE})\) 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 \(P(\text{APE})\) 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):

\[
P(S \mid X \in \text{APE}) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left( \frac{S - \bar{S}_{\text{APE}}}{\sigma} \right)^2} \quad (8)
\]

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

\[
P(S) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left( \frac{S - \bar{S}_{\text{AST}}}{\sigma} \right)^2} \quad (9)
\]

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

\[
P(X \in \text{APE} \mid S) = e^{\frac{\Delta \bar{S}_{\text{APE}} - \Delta \bar{S}_{\text{AST}}}{\sigma}} P(\text{APE}) \quad (10)
\]

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 Sm is given by

\[
P(X \in \text{APE} \mid S \leq S_m) = P(\text{APE}) \cdot \int_{0}^{S_m} e^{-\frac{1}{2} \left( \frac{S - \bar{S}_m}{\sigma_m} \right)^2} dS \quad (11)
\]

\[
P(X \in \text{APE} \mid S \leq S_m) = P(\text{APE}) \cdot \left[ \frac{\sigma_m^2}{\Delta \bar{S}_m^2} e^{-\frac{\Delta \bar{S}_m \cdot S_m}{\sigma_m^2}} \right] \cdot \left[ e^{\frac{\Delta \bar{S}_m \cdot S_m}{\sigma_m^2}} - 1 \right] \quad (12)
\]

We compute a total sequence score \(S_X\) over all 12 motifs

\[
S_X = \sum_{m=1}^{12} \log_2 [P(X \in \text{APE} \mid S \leq S_m)] \quad (13)
\]

\[
S_X = \sum_{m=1}^{12} \log_2 \left[ \frac{\sigma_m^2}{\Delta \bar{S}_m^2} e^{-\frac{\Delta \bar{S}_m \cdot S_m}{\sigma_m^2}} \right] + \sum_{m=1}^{12} \log_2 \left[ \frac{\Delta \bar{S}_m \cdot S_m}{\sigma_m^2} - 1 \right] + 12 \log_2 P(\text{APE}) \quad (14)
\]
The first and 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 the third component $E_3$, 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 $E_1$ to $E_5$ is available elsewhere (Venkatarajan and Braun, 2001).

Motifs in the APE family

The APE I protein family consists of apurinic/apyrimidinic endonucleases and exonucleases. The PCP macro of our MASIA program identifies 12 motifs of various lengths with the cutoff values for entropy, $\lambda = 1.25$, for insignificant positions within a motif, $G = 2$, and minimum length, $L = 4$ (Table 1). The most of the motifs are located in the $\beta$-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.

![Table 1. The 12 motifs of APE family defined by MASIA](https://academic.oup.com/bioinformatics/article-abstract/19/11/1381/220412/1384)
Figure 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 dendroogram of the APE family in supplementary material).

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 Zn$^{2+}$ 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$^\alpha$ 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 Zn$^{2+}$ (1QQ9 and 1ATL) ions in their active site, one (1D2N) contained Mg$^{2+}$ and one Ca$^{2+}$. 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, synaptotagmin 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 synaptotagmin 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

<table>
<thead>
<tr>
<th>PDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HD7</td>
</tr>
<tr>
<td>1AKO</td>
</tr>
<tr>
<td>2DNJ</td>
</tr>
<tr>
<td>1BYU</td>
</tr>
<tr>
<td>1B3U</td>
</tr>
<tr>
<td>1MDA</td>
</tr>
<tr>
<td>1MPY</td>
</tr>
<tr>
<td>1YRG</td>
</tr>
<tr>
<td>1QQ9</td>
</tr>
</tbody>
</table>

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).

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 (Venkataraman and Braun, 2001) that our five property vectors represent all known chemical physical 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 superfamiliy members

Despite their low overall sequence identity, APE, DNase-I and IPP families share a similar 3D structure and are members of a superfamiliy with a common SCOP designation (Lo Conte et al., 2002). Our methods rapidly identified members of this superfamiliy 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 superfamiliy members. The structurally equivalent motifs, or molegos, identified by our program are boxed in the FSSP alignment of DNase-I superfamiliy 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 superfamiliy (Schein et al., 2002). No non-APE protein sequence in the ASTRAL40 database scored higher than the average total score ($S_t$) 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 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

<table>
<thead>
<tr>
<th>PDBa</th>
<th>Score in bits (fraction to the highest score)</th>
<th>MOLEGOS found</th>
<th>SCOPb</th>
<th>ECc</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HJ7</td>
<td>1942 (1.00)</td>
<td>1.2,3,4,5,6,7,8,9,10,11,12</td>
<td>d.151.1</td>
<td>4.299.18</td>
<td>APE (Mn/Mg/Pb)</td>
</tr>
<tr>
<td>1AKO</td>
<td>1831 (0.94)</td>
<td>1.2,3,5,6,7,8,9,10,11,12</td>
<td>d.151.1</td>
<td>3.11.12</td>
<td>Exonuclease III</td>
</tr>
<tr>
<td>2DNJ</td>
<td>1072 (0.55)</td>
<td>1.2,5,6,7,8,9,10,12</td>
<td>d.151.1</td>
<td>3.121.1</td>
<td>Deoxyribonuclease I</td>
</tr>
<tr>
<td>1BYQ</td>
<td>971 (0.50)</td>
<td>1.2,5,6,7,8,9,10,12</td>
<td>d.151.1</td>
<td>3.21.12</td>
<td>Phospholipid ethanol phosphate Synaptophosphinase I</td>
</tr>
<tr>
<td>1QQ9</td>
<td>698 (0.36)</td>
<td>5.6,9,10,12</td>
<td>c.56.54</td>
<td>3.14.11.12</td>
<td>Aminopeptidase (Zn, Ca)</td>
</tr>
<tr>
<td>1ATL</td>
<td>633 (0.33)</td>
<td>5.6,9,10,12</td>
<td>d.92.19</td>
<td>3.24.24.24</td>
<td>Snake venom metalloprotease (Zn, Ca)</td>
</tr>
<tr>
<td>1D09</td>
<td>619 (0.32)</td>
<td>5.9,12</td>
<td>d.58.21</td>
<td>2.13.2</td>
<td>Aspartate carboxymethyltransferase (Zn)</td>
</tr>
<tr>
<td>1D2N</td>
<td>613 (0.32)</td>
<td>5.8,9,12</td>
<td>c.37.13</td>
<td>N-ethylmaleimide of sensitive fusion protein (Mg)</td>
<td></td>
</tr>
<tr>
<td>1D0B</td>
<td>579 (0.30)</td>
<td>2.5,9,12</td>
<td>c.10.21</td>
<td>N-ethylmaleimide of sensitive fusion protein (Mg)</td>
<td></td>
</tr>
<tr>
<td>1IEEM</td>
<td>571 (0.29)</td>
<td>5.6,8,12</td>
<td>a.45.11</td>
<td>Glutathione S-transferase</td>
<td></td>
</tr>
</tbody>
</table>

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

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 genome 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


