PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Identification of GATC- and CCGG-recognizing Type II REases and their putative specificity-determining positions using Scan2S—a novel motif scan algorithm with optional secondary structure constraints

Masha Y. Niv,[1]*† Lucy Skrabanek,[1,2] Richard J. Roberts,[3] Harold A. Scheraga,[4] and Harel Weinstein[1,2]

[1] Department of Physiology and Biophysics, Weill Medical College of Cornell University, 1300 York Ave., New York, New York 10021

[2] HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Medical College of Cornell University, 1305 York Ave., New York, New York 10021

[3] New England Biolabs, 240 County Road, Ipswich, Massachusetts 01938-2723

[4] Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301

ABSTRACT

Restriction endonucleases (REases) are DNA-cleaving enzymes that have become indispensable tools in molecular biology. Type II REases are highly divergent in sequence despite their common structural core, function and, in some cases, common specificities towards DNA sequences. This makes it difficult to identify and classify them functionally based on sequence, and has hampered the efforts of specificity-engineering. Here, we define novel REase sequence motifs, which extend beyond the PD-(D/E)XK hallmark, and incorporate secondary structure information. The automated search using these motifs is carried out with a newly developed fast regular expression matching algorithm that accommodates long patterns with optional secondary structure constraints. Using this new tool, named Scan2S, motifs derived from REases with specificity towards GATC- and CGGG-containing DNA sequences successfully identify REases of the same specificity. Notably, some of these sequences are not identified by standard sequence detection tools. The new motifs highlight potential specificity-determining positions that do not fully overlap for the GATC- and the CCGG-recognizing REases and are candidates for specificity re-engineering.

Proteins 2008; 71:631–640.

Key words: secondary structure; protein motif; physicochemical properties; restriction endonucleases; regular expression; specificity-determining positions.

INTRODUCTION

Restriction endonucleases (REases) are components of restriction modification systems that protect bacteria and archaea against invading foreign DNA. Bacteria initially resist infections by new viruses because REases within the cell destroy foreign DNA molecules by hydrolyzing the ester bonds of the sugar-phosphate backbone at a particular recognition sequence. Bacterial DNA is protected from cleavage by REases by methylation (by the corresponding bacterial methylase) of the same sequence.

The restriction-modification (R-M) systems have been classified into Types I through IV, depending on the number and organization of their functional subunits (restriction, modification, and specificity).[1] The Type II REases are the most common among the biochemically characterized REases. Type II REases recognize specific unmethylated DNA sequences and cleave at invariant positions, at or close to the recognition sequence to produce 5′-phosphates and 3′-hydroxyls.[1–3] The specificity of Type II REases has made them indispensable tools in recombinant DNA technologies.[3,4]

A PD-(D/E)XK motif, identified in most of the characterized Type II REases, was shown to be conserved in many enzymes involved in DNA recombination and repair,[5,6] which are now known as the PD-(D/E)XK superfamily. The detection of Type II REase subfamilies that are specific for a particular DNA sequence is challenging

Grant sponsor: NIH; Grant number: GM-14312; Grant sponsor: Cornell University/Weill Medical College.
*Current address: Institute of Biochemistry, Food Science and Nutrition, Faculty of Agricultural, Food and Environmental Quality Sciences, The Hebrew University of Jerusalem, PO Box 12, Rehovot 76100, Israel.
†Correspondence to: Dr. Masha Niv, Institute of Biochemistry, Food Science and Nutrition, Faculty of Agricultural, Food and Environmental Quality Sciences, The Hebrew University of Jerusalem, P.O. Box 12, Rehovot 76100, Israel. E-mail: niv@agri.huji.ac.il
Received 23 May 2007; Accepted 13 August 2007
Published online 30 October 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21777

because of their low sequence identity (15% and below), despite their common function. Furthermore, altering the specificities of these restriction enzymes, for example, by single-site and cassette mutagenesis using insights from known structures, has often been unsuccessful (see Townson *et al.*[7]). It is our goal here to develop protein motifs that can detect Type II REases of a particular specificity, and to highlight potential specificity-determining residues.

We first probe the performance of several commonly used techniques to detect Type II REases that recognize particular DNA sequences. As the recall, or sensitivity, ([true positives]/([true positives + false negatives]) of these methods is very low, we present a new complementary bioinformatics approach, Scan2S.[8] ***Scan*** stands for sequence ***scan***ning for detection of motifs (regular expression patterns), and ***2S*** indicates that the motifs may include secondary (***2***) structure (***S***) information. Implemented in several tools (see for example Refs. 9–12), regular expression patterns are often scanned against sequences of unknown function for homology detection and function prediction.[13] However, to the best of our knowledge, Scan2S is the first regular expression-scanning algorithm that enables the straightforward use of secondary structure constraints in protein patterns. The Scan2S program uses the Java 5.0 regex (regular expression) package. It is fast and supports long and flexible query motifs combined with inclusion of secondary structure constraints.

The new approach described here consists of the following steps: (1) Derivation of the query motif that is identified from positions in the sequence alignment that conserve biochemical function, residue identity, physicochemical property of the residues, or secondary structure. (2) Prediction of the secondary structure for the set of sequences that are being queried using established prediction methods.[14] (3) The Scan2S step, which carries out the search for the motifs derived in step 1 in the datasets prepared in step 2.

In illustrating the application of Scan2S, we show that the use of Scan2S with motifs derived from REases with specificity towards GATC- or CCGG-containing DNA sequences can successfully identify REases of the corresponding specificities in the dataset of all Type II REases.[3] The performance of the method in terms of precision, or positive predictive value (PPV), and recall (sensitivity), is similar to that of BLASTP (basic local alignment search tool for protein sequences) search[15] and better than of additional methods that we have tested as described. Notably, the sets of REases retrieved by the different methods do not overlap fully, and Scan2S provides true positives not found by the other methods. The useful motifs highlight potential specificity-determining positions that can serve as candidates for specificity re-engineering. Interestingly, these sites do not completely coincide for the GATC- and the CCGG-recognizing REases.

METHODS

Motif derivation

Sequence alignments

Structure-based sequence alignments were obtained from the "align structures" option of the TCoffee server http://igs-server.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi.[16] The GATC-specific REases sequence alignment was obtained by aligning the structures of BamHI (2BAM, G^GATCC recognition sequence, where ^ indicates the cleavage site), BstYI (1VRR, R^GATCY) and BglII (1DFM, A^GATCT) from the Protein Data Bank.[17] The CCGG-specific REases sequence alignment was obtained by aligning the X-ray structures for MspI (1SA3, C^CGG), NaeI (1IAW, GCC^GGC) Cfr10I1(1CFR, R^CCGGY), Bse634I (1KNV, R^CCGGY), and NgoMIV (1FIU, G^CCGGC). The alignment of all eight structures was used to identify corresponding positions in these two families.

GATC-specific motif generation

Positions known to be involved in catalysis and all fully conserved positions in the structure-based sequence alignment (except G178) were included in this motif. Position 178 (BamHI numbering) was not included, because even allowing [P,G] in this position resulted in a motif that matched only the three original sequences. Amino acids are grouped into six physicochemical classes, following Mirny and Shakhnovich.[18,19] The classes are: aliphatic [AVLIMC], aromatic [HWYF], polar [NQST], negatively charged [ED], positively charged [KR], and special conformation [GP]. At conserved positions, all amino acid residues with a similar physicochemical property as the residues seen in the alignment are allowed in the motif. Catalytic sites and sites within 5 Å of the DNA are exempt from the "relaxation" treatment, that is, only the residues found in the original alignment are allowed at these sites. Some of the conserved sites lie in conserved secondary structure elements identifiable in the 3D-structures that were used in the original alignment [secondary structure was assigned by DSSP[20]]. This information is included in the motif definition in the form of secondary structure constraints. The GATC-specific motif contains four secondary structure constraints: one for each sequence-conserved site where the secondary structure is identical in all of the aligned structures. For example, the constraint at the second site in the motif (site 28 in BamHI numbering) is "not Extended," which means that this site may not be found in an Extended strand element, because it was not found in a strand in any of the structures in the original alignment. "Extended" stands for a site in an Extended strand, "Helix" for a site in a Helix, and "not Helix" for sites never on a Helix.

Table I
GATC Motif Summary

BamHI no.	Occurrence	Allowed AA	Secondary constraints
14	EEE	DE	
28	EEE	DE	Not extended
58	VVV	V (contact)	
61	KN	KN (putative catalytic)	Helix
68	LLL	AVLIMC	
74	WWW	WFYH	
84	KKK	RK	
94	DDD	D (catalytic)	
97	KKK	RK	
111	EEE	E (catalytic)	Extended
113	EQ	ENQ (catalytic)	
136	III	AVLIMC	
160	EEE	DE	Not extended
173	PPP	PG	
178	GGG	not included	

"Occurrence" indicates the residues present at that site (numbered using BamHI) in the three available PDB structures for GATC-specific REases. "Allowed AA" indicates the amino acids residues allowed in each site. "Secondary constraints" indicates the allowed secondary structure element at that position.

The Scan2S GATC motif is summarized in Table I. The conserved and the catalytic sites are indicated in BamHI numbering. Only residues that occurred in the structure-based alignment are allowed for the catalytic sites 94 and 111, the putative catalytic site 61 and the DNA-contacting (V58) site. E, N, and Q were allowed at catalytic site 113, while the whole physicochemical class is allowed in the rest of the conserved sites. "Secondary constraints" were derived based on the secondary elements of the conserved sites in the three structures.

CCGG-specific motif derivation

The motif derivation is similar to that for GATC, except that the sites are considered conserved if the physicochemical class (rather than the individual residue) is fully conserved in the five aligned sequences. Structures 1FIU, 1SA3, and 1IAW were used in the analysis of protein/DNA contacts (the structures of Bse634I and CFR10I have not yet been solved in complex with DNA). The resulting motif is summarized in Table II.

The REase database

Type II REase sequences were downloaded from the REBASE database.[3] This set of sequences is referred to as REset. There are 1357 REases in the set, 729 of them with known specificities towards DNA sequences. 111 REases in this set recognize GATC-containing DNA sequences and 45 recognize CCGG-containing DNA sequences (referred to as GATC and CCGG REases, respectively). Secondary structure predictions for the REset sequences were obtained using PSIPRED,[14] a two-stage neural network for prediction of protein secondary structure based on the position specific scoring matrices generated by PSI-BLAST (Position specific iterative BLAST). PSIPRED is evaluated as one of the best secondary structure prediction methods and has ~78% precision.[21]

Regular expression match

We have developed Scan2S, a regular expression-based motif-scanning algorithm.[8] Scan2S is designed to find a motif in a protein sequence while also satisfying secondary structure constraints (e.g., a certain residue of the motif sequence must be located on a particular secondary structure element). Each element of a Scan2S motif contains the residue(s) allowed at that position, followed by the secondary structure constraint expected at that position. Motifs can be constructed by using all the conventions recognized by the Java 5.0 regex package (details are given in http://java.sun.com/j2se/1.5.0/docs/api/java/util/regex/package-summary.html). In Scan2S syntax, the following nomenclature is used: each position in the motif is followed by its secondary structure constraint, for example, [FY]H means that phenylalanine or tyrosine must be located in a helix. If there are multiple residues allowed at a position, those residues are bracketed. Similarly, if there is more than one character required to describe the secondary structure constraint, the secondary structure constraint for that position is also bracketed. One can also use the "not" operator to indicate that a residue may not be found in a certain secondary structure element, for example, P[^H] means that the proline in this motif must not lie on a helix. Where there are no secondary structure constraints, this is indicated by a period, for example, [ILV]. means that the residue at that position can be an isoleucine, leucine, or valine, and that there is no secondary structure constraint imposed.

Table II
CCGG Motif Summary

MspI numbering	BamHI numbering	Occurrence	Allowed AA	Secondary constraints
31	57	GGGGG	G (contact)	Not extended
35	61	EEEEE	E (putative catalytic)	Not extended
38	64	ILIIC	AVLIMC	
99	94	DDDDD	D (catalytic)	Not helix
102	97	IIIIV	AVLIMC	
116	110	LVLVI	AVLIMC	Not helix
117	111	SNGD	SNGD (catalytic)	Not helix
118	112	ICVLC	AVLIMC	Not helix
119	113	KKKKK	K (catalytic)	Not helix
121	115	SSTST	ST (contact)	
205	140	ILAAV	AVLIMC	Not helix

"Occurrence" indicates the residues present at that site (based on MspI numbering) in the five available PDB structures for CCGG-specific REases. "Allowed AA" indicates the amino acids residues allowed in each site. "Secondary constraints" indicates the constraint on the allowed secondary structure element.

Since the motif contains sequence and structure constraints, the protein datasets that are being queried must be constructed in the same way, that is, both the sequence and structure information are taken as input. The sequence and structure information for each protein in REset is combined, such that each residue is followed by the secondary structure predicted for that site.

The Scan2S program is available for download at http://physiology.med.cornell.edu/go/scan2s.

Other methods

Sequence similarity detection

BLASTP[15] was used against the REBASE sequences, fvia http://tools.neb.com/~vincze/blast/index.php (with the default cutoffs). This is the only method tested here that does not utilize the structure-based sequence alignment information.

The PSI-BLAST[22] implementation in the MPI toolkit[23] http://toolkit.tuebingen.mpg.de/psi_blast was applied to the alignments of GATC and CCGG REases versus the nonredundant bacterial dataset. Only restriction endonuclease hits were counted in order to compare to the other results that were obtained for REset.

HHPred[24] http://toolkit.tuebingen.mpg.de/hhpred builds a profile hidden markov model (HMM) from a query sequence and compares it with a database of HMMs representing annotated protein families. We ran HHpred using the structure-based alignments as queries against the PfamA dataset. Again, only REase hits were recorded to compare with the other methods.

MAGIIC-PRO[25] http://biominer.bime.ntu.edu.tw/magiicpro/ and PRATT[26] http://expasy.org/tools/pratt/, were used for automated motif derivation from the structure-based multiple sequence alignments of the GATC and the CCGG REases. The resulting motifs were translated into Scan2S syntax and scanned against REset.

Prediction of specificity-determining residues

SDPpred[27] http://math.genebee.msu.ru/~psn/index.htm predicts residues that determine differences in functional specificity of homologous proteins by searching for sites that are well conserved within specificity groups but differ between them. The SDPpred predictions for a structure-based alignment of the three GATC-recognizing and five CCGG-recognizing REases were compared to our own Scan2S-based predictions of specificity determining residues.

RESULTS

Type II REases typically exhibit a pairwise identity below 15% and belong to the "midnight zone" of sequence similarity where homology can be detected only via structural information.[28] Preliminary testing with several automated sequence alignment methods (such as those described in Refs. 16,29–32) confirmed that only structure-based methods provide reliable multiple structure alignments, as assessed by alignment of biochemically known catalytic sites. We studied the 111 Type II REases that recognize GATC-containing DNA sequences (referred to as GATC REases) and the 45 Type II REases that recognize CCGG-containing DNA sequences (referred to as CCGG REases), the only two groups for which at least three protein structures were solved (see Niv *et al.*[33] for a recent survey and analysis of Type II REase structures).

Detection of REases with commonly used methods and with Scan2S

We have used several established methods to detect GATC- and CCGG-recognizing REases in REset (the curated set of Type II REases in REBASE[3]), as described in Materials and Methods section. The number of true positives (REases hits with the correct specificity) and of false positives (REase hits with a known but different specificity) found in the REset dataset are shown in Table III and described later. Table III also reports precision (defined as [true positives/(true positives + false positives)]), and recall (defined as [true positives/(true positives + false negatives)]).

A BLASTP search using sequences of the three GATC REases of known structure as queries retrieved a total of nine Type II REases, all of them having recognition sequences that include GATC. Using the CCGG REases of known structure as queries, BLASTP retrieved a total of 14 CCGG REases, two Type II REases of unknown specificity, and one REase of a different specificity.

Using the GATC multiple sequence alignment query with PSI-BLAST[22] against the nonredundant bacterial genome dataset retrieved the three original GATC sequences, and two additional GATC REases, that were also found by BLASTP. The CCGG multiple sequence alignment query retrieved the five original sequences and four additional CCGG sequences, all of which were found by BLASTP.

A state-of-the-art HMM method HHpred[24] scanned against the PfamA database retrieved only the original Type II REases used in the construction of each of the alignments.

The best ranking motifs identified by the automated motif derivation method PRATT[26] for the GATC and CCGG REases were translated to Scan2S syntax (without adding secondary structure constraints), and scanned against REset. The PRATT-derived GATC motifs matched only the three original sequences from which they were derived. The PRATT-derived CCGG motif matched four of the Type II REases from which it was derived, and two additional CCGG-recognizing REases.

Table III
Methods Comparison

	Recall (%)	Precision (%)	TP	FP	FN	Unknown specificity	TP not found by BLASTP	TP not found by PRATT
Scan2S-GATC	13	100	14	0	97	2	10	12
BLASTP-GATC	9	100	9	0	102	0	0	6
PRATT-GATC	3	100	3	0	108	0	0	0
MAGIIC-PRO	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
PSI-BLAST (MPI)	5	100	5	0	106	0	0	2
HHPred versus pfamA	2	100	2	0	109	0	0	0
Scan2S-CCGG	31	88	14	2	31	1	6	10
BLASTP-CCGG	31	93	14	1	31	2	0	8
PRATT-CCGG	20	100	6	0	24	0	0	0
MAGIIC-PRO	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
PSI-BLAST (MPI)	20	100	8	0	22	0	0	2
HHPred versus pfamA	7	100	3	0	42	0	0	0

A comparison of the numbers of true positive (TP), false positive (FP), and false negative (FN) matches found by using Scan2S and alternative methods.

MAGIIC-PRO automated motif derivation method[25] detected no motifs in the GATC or CCGG structure-based sequence alignments using default parameters.

The Scan2S GATC and CCGG motifs were derived from the GATC and CCGG structure-based multiple sequence alignments, respectively, as described in Materials and Methods section and discussed further in the next sections. The Scan2S GATC motif, which combines sequence and structural data, retrieves 16 sequences from REset, 14 of which are GATC REases and two of which (BjaORF865P and EsaNPORF65P) have unknown recognition sequences. Because the motif is specific (100% precision, 13% recall), we suspect these may be as-yet-unidentified GATC REases. Ten of the Scan2S GATC motif true positive hits were not found by BLASTP, the best performing of the commonly used methods we have tested, and 12 were not found by the PRATT motif, the best automatically found motif we obtained (Table III).

The Scan2S CCGG motif retrieves 17 REases, 16 of which have known recognition sequences, and 14 of those are CCGG-containing sequences (recall 31%, precision 88%). Six of the true positives found by the Scan2S CCGG motif were not found by BLASTP and 10 were not found by PRATT.

The results described earlier indicate that the Type II REases present a significant challenge for all sequence analysis techniques that we have tested, as indicated by the low recall (3–31%). Scan2S performs better than all of the other methods except BLASTP in terms of a combination of significant recall and high precision. Importantly, because the hits obtained by BLASTP and by Scan2S overlap only partially, Scan2S provides a nontrivial addition to the bioinformatics toolbox. Furthermore, the positions participating in the motifs may be important for understanding the function of these proteins. We, therefore, proceed to describe in detail the positions that constitute the Scan2S GATC and CCGG motifs.

Scan2S GATC-specific motif

The structure-based multiple sequence alignment of GATC-recognizing (as well as of CCGG-recognizing) REases is shown in Figure 1.

The pairwise identity of GATC REases included in the multiple sequence alignment is 15% and below, and out of the four positions of the PD-(D/E)XK pattern, two are not conserved even in this small set of three REases acting on similar substrates: amino acid residues I, I, T populate the Proline site of the PD-(D/E)XK pattern (I93 in BamHI) and residues E or Q populate the Lysine site of the pattern (E113 in BamHI). Instead, other sites are conserved (highlighted in light cyan in Fig. 1). The conserved sites were mapped onto the experimental structures of the GATC-recognizing REases with their cognate DNA substrates (see Materials and Methods section), to identify residues likely to be involved in DNA recognition. These can be classified into two groups as described below using BamHI numbering and shown in Figure 2(A). The first group includes a conserved spatial cluster of residues that do not contact DNA, consisting of E28 (E22 in 1DFM, E26 in 1VRR), L68 (L61 in 1DFM, L68 in 1VRR), W74 (W66 in 1DFM, W73 in 1VRR), and K97 (K87 in 1DFM, K122 in 1VRR). The second group includes three sites within 5 Å of the DNA strand that are not part of the catalytic triad (94,111,113 in BamHI numbering). These are positions V58, K61, and K84 [not shown in Fig. 2(A)]. V58 in 2BAM and the corresponding V51 in 1DFM, V58 in 1VRR, are within 5 Å of the ***A***6 and ***T***7 nucleotides. These nucleotides are part of the G***AT***C-containing recognition site, identifying V58 site as a potential novel specificity determinant. K61

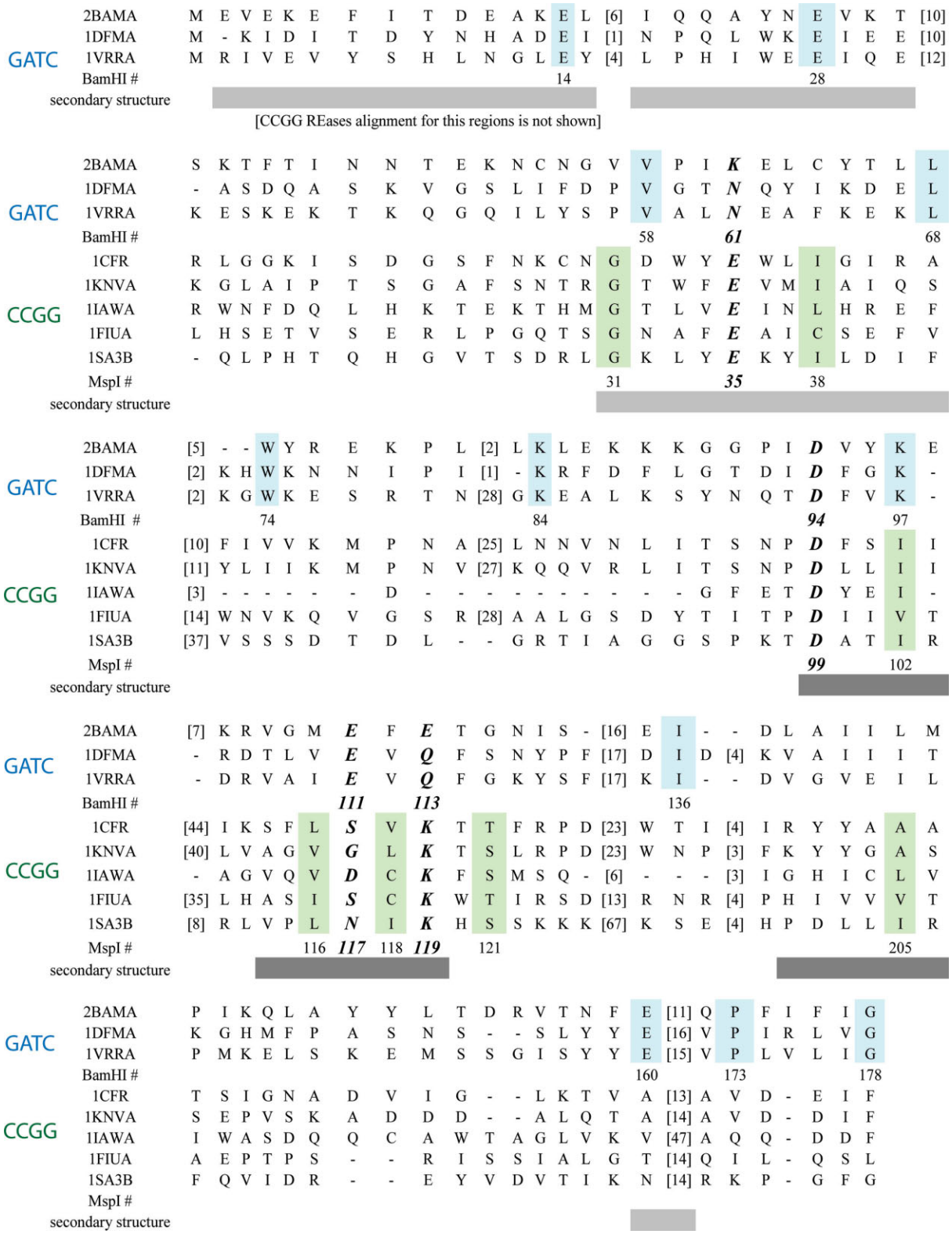

Figure 1

Structure-based sequence alignment of GATC and CCGG REases. The catalytic residues [including the putative catalytic site 61 (Niv et al., unpublished results)] are shown in bold italics. Positions with fully conserved residues in the GATC REases are highlighted in light cyan and indicated in BamHI numbering. Positions with physicochemical properties conserved in the CCGG REases are highlighted in light green and indicated in MspI numbering. Conserved regions with predominantly helical (extended strand) secondary structure are indicated by a light (dark) gray stretch.

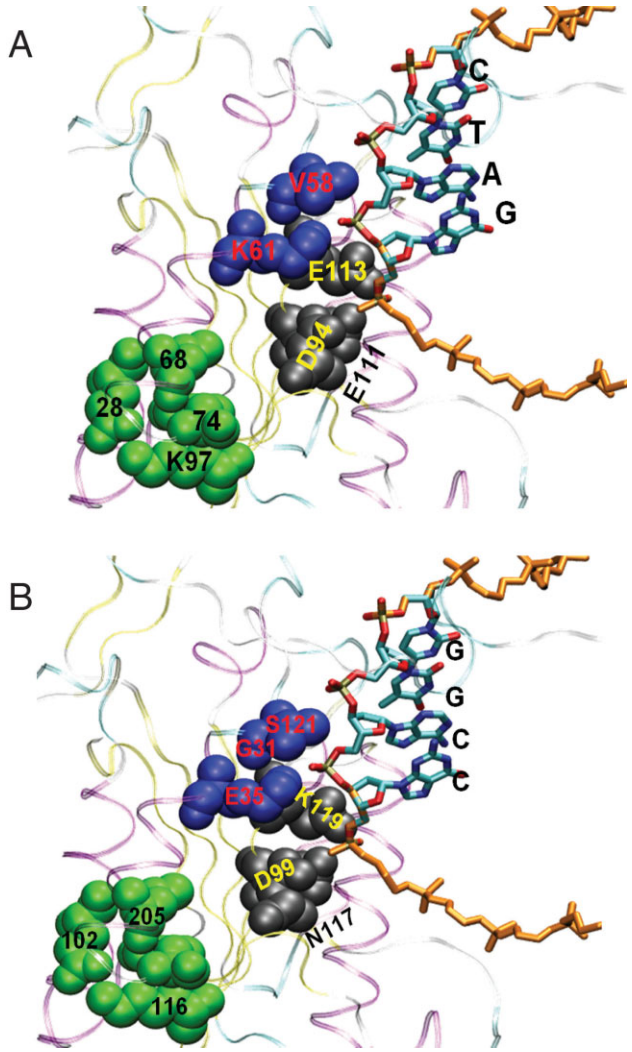

Figure 2

*Conserved patches. (**A**) The protein monomer (chain A in 2BAM.pdb) is shown in ribbon representation colored by secondary structure. The backbone of the DNA strand (chain D in 2BAM.pdb) and the GATC nucleotide bases are shown. The catalytic residues 94, 111, 113 in van der Waals representation are colored grey. The conserved spatial cluster residues (28, 68, 74, and 97) are colored green. The novel GATC-family conserved residues within 5 Å from the DNA strand (58 and 61) are colored blue. (**B**) The protein monomer (chain A 1SA3.pdb) is shown in ribbon representation. The hydrophobic cluster (sites 102, 116, 205) is colored green. The catalytic residues 99, 117, and 119 are colored gray. The novel CCGG-family conserved residues within 5 Å from the DNA strand (sites 31, 35, and 121) are colored blue. The figure was prepared using VMD (Visual Molecular Dynamics) software.*[34]

is a new putative catalytic position (Niv *et al.*, unpublished results). K84 is within 5 Å of the A2 and T3 nucleotides, upstream of the recognition sequence in 2BAM, but the corresponding K74 and K109 residues (in 1DFM and 1VRR, respectively) do not interact directly with the DNA. This site is therefore less likely to be a specificity determinant.

The importance of each component in motif derivation was probed as follows: (1) Allow only for the residue that occurs in the original alignment in the conserved sites. In this case, the motif recalls only the three original sequences (3% recall, 100% precision); exclusion of the secondary constraints results in matching one additional GATC sequence (4% recall, 100% precision). (2) Exclude the secondary structure constraints. In this case the precision drops to 60%, (recall is 14%).

Scan2S CCGG-specific motif

The CCGG motif includes the catalytic residues (MspI numbering used): D99, N117, and K119 as well as the putative catalytic E35 (K61 in BamHI, Niv *et al.*, unpublished results) and sites populated by residues of one physicochemical class only [highlighted in light green in Fig. 1 and shown as spheres in Fig. 2(B)].

In the CCGG-recognizing subgroup of REases, the PD-(D/E)XK motif is not strongly conserved, as the P site (98 in MspI numbering) is occupied by either T or P and the D/E site (117 in MspI numbering) is occupied by N, D, S, or G. Using experimental structures, the sites can be classified into the following two groups. The first group includes conserved residues distant from the DNA strand: I102, L116, and I205. These constitute a hydrophobic cluster [see Fig. 2(B)]. The second group includes conserved residues within 5 Å from the DNA that are not part of the catalytic triad (99, 117, and 119): G31 (57 in BamHI numbering), E35 (putative catalytic (Niv *et al.*, unpublished results) corresponding to K61 in BamHI numbering), S121 (115 in BamHI numbering) and I118 [112 in BamHI numbering, not shown in Fig. 2(B)]. The sidechains of I118 in 1SA3 and the corresponding residues in 1IAW and 1FIU point away from the DNA strand, suggesting that this site is less likely to be involved in recognition than sites 31 and 121.

The importance of each component in motif derivation was probed as follows: (1) Basing the motif only on the sites with identical residues in all five CCGG REases (sites 31, 35 [putative catalytic], 99 and 119 [catalytic] in MspI numbering) without secondary structure constraints results in a promiscuous motif that finds 961 matches in REset, 828 of known specificity, of which 36 are CCGG-specific REases (recall of 80%, but a very low precision of 4%). Adding the secondary structure constraints at the fully conserved positions included in this motif does not change the recall and precision levels significantly. (2) Using the same motif as described in Table II, but allowing only residues that occur in the alignment, results in 10 hits, all of them true positive (100% precision, but only 22% recall). (3) Excluding the secondary structure constraints from the motif described in Table II results in 54% precision and 50% recall.

Our results indicate that secondary structure constraints are important for motif specificity, in agreement with our analysis of secondary structure augmented PROSITE motifs,[8] while the relaxation of the motif to

include residues of conserved physicochemical property is important for better recall.

Notably, the DNA-contacting conserved sites do not fully coincide for the GATC and the CCGG REases. The conserved structural cluster in CCGG REases (I102, L116, and I205 in MSPI numbering, corresponding to K97, M110, and I140 in BamHI numbering) is also different from the conserved structural cluster in GATC REases (E28, L68, W74, and K97 in BamHI numbering): only the K97 site participates in both clusters [see Figs. 1 and 2].

To compare our proposed specificity determinants with other predictions, we used the SDPpred server.[27] SDPpred predicts residues that determine differences in the functional specificity of homologous proteins by searching for sites that are well conserved within specificity groups but differ between them. This approach, therefore, implies that the sites of specificity-determining residues are identical for different specificities. The structure-based sequence alignment for the three GATC-recognizing REases and the five CCGG recognizing REases (derived using TCoffee[16]) was subjected to the SDPpred algorithm.[27] The resulting SDP predictions are G31, E35, I102, and K119 in MspI numbering, corresponding to V57, K61, K97, and E113 in BamHI numbering. Thus Scan2S has found unique potential specificity determining sites (V58 for GATC-recognizing and S121 for CCGG-recognizing REases) and also a unique structural cluster for each group in addition to the SDP predictions. We conclude, therefore, that subfamily-specific positions may be an important mechanism for achieving specificity in protein–protein interactions, as recently discussed by Pirovano *et al.*[35] This concept augments the more established notion of "persistent" positions that are conserved across super-families and folds.[27,36,37]

DISCUSSION

Motivated by the important role of Type II REases in molecular biology, we set out to analyze the sequence/structure/function relations of these proteins. Type II REases are highly divergent in their sequences despite having a common structural core and function and in some cases common DNA specificity. It is important to note that the analysis of the Type II REase subfamilies presented here is complicated by the fact that the multiple sequence alignments were limited to the small number of members that have experimental 3D structures, too few to allow a rigorous statistical analysis. We have, therefore, addressed some of the biological questions and computational challenges by deriving subfamily-specific motifs and highlighting potential specificity determinants. The general applicability of the Scan2S method, and the trade-off between precision and recall upon refining protein patterns using secondary structure constraints was recently shown for PROSITE motifs[8] and is currently being evaluated further for additional sequence-dissimilar protein families.

Here we have focused on two groups of such enzymes, namely the CCGG-recognizing and the GATC-recognizing Type II REases, and identified sites that have conserved physicochemical properties, some of which reside in well-defined secondary structure elements. We have used our novel regular expression matching method, Scan2S, which enables the search of sequence databases using long flexible motifs with optional secondary structure constraints, to detect REases of GATC and CCGG specificities.

The role of motif components in the sequence search

Physicochemical properties

It has been shown that conservation is higher on the level of physicochemical properties than on the level of individual amino acids.[19,38–40] Coarse-grained, or reduced, alphabet approaches that represent the physicochemical properties of amino acids have been applied to pattern recognition, generation of consensus sequences from multiple alignments, protein folding, and protein structure prediction.[41] Different approaches to grouping amino acids according to their physicochemical properties exist in the literature.[19,39,42,43] Here, we used the physicochemical classes of Mirny and Shakhnovich,[19] though coarse graining using parameters from Kidera *et al.*[44] leads to qualitatively similar results (not shown). The physicochemical classes of amino acids were used to identify conserved sites in the CCGG multiple sequence alignment and to relax both the GATC and the CCGG motifs by allowing all amino acids of the dominant physicochemical property at the conserved sites. A related idea has been explored for refining protein prenylation motifs by penalizing deviations from physical property requirements on the sequence,[45] and for derivation of motifs for low sequence similarity DNase-1 related endonucleases.[46] Importantly, we find that relaxation of motifs using physicochemical properties is crucial for improving the motifs recall. However, these properties were not sufficient for obtaining a high specificity motif, and the structural component was utilized as well.

Structural information

Experimental structural information was used in three ways in this study of sequence-dissimilar enzymes: (a) First, it was used to obtain structure-based sequence alignments using the 3D-TCoffee server[16]; (b) Second, the structures were used to identify conserved secondary structure elements which were then included as constraints in the motifs. The inclusion of secondary structure information has been shown to improve similarity

detection by sequence profiles and HMM methods.[47–50] To the best of our knowledge, Scan2S is the first implementation of secondary structure information for refining protein motif and has already been shown to improve the precision of PROSITE motifs[8]; (c) Lastly, in order to identify potential specificity determinants for DNA binding, structures of REase/DNA complexes were used to identify sites that are found at the interaction interface, to restrict the allowed residues at these sites.

Applicability of the method

Structures have been shown to be more conserved than sequences.[51–53] The commonalities obtained using structural data can shed light on members of the protein families for which no structural information exists yet and can highlight putative functional sites. The approach described in this article is suitable for analysis of functionally related, sequence-dissimilar proteins for which several structural representatives are obtained. The main drawback of the application as described here is the laboriousness in the motif derivation stage. We are currently exploring ways to automate the procedure.

CONCLUSIONS

We have derived novel motifs and have used Scan2S (motif scan with optional secondary structure constraints) for detection of GATC- and CCGG-recognizing Type II REases. The specific implementation of Scan2S and other bioinformatics methods reveal that detection of sequence similarity in subfamilies of Type II REases presents a formidable challenge for all the methods tested, as indicated by the low (3–31%) recall levels. Notably, the sets of REases retrieved by the different methods do not overlap fully, and Scan2S provides true positives not found by the other methods. Thus, Scan2S constitutes a novel approach for searches against REset that is *complementary* to BLASTP. The Scan2S program is available for download at http://physiology.med.cornell.edu/go/scan2s. The predictive capabilities of the motifs implemented in Scan2S suggest that the matches to REases of heretofore unknown specificity may have the same specificity as those from which the motif was derived. The motifs highlight potential specificity-determining positions. These positions, which do not coincide fully for the GATC and the CCGG families, offer promising candidates for re-engineering specificity in this biotechnologically important class of DNA processing enzymes.

ACKNOWLEDGMENTS

The authors thank Dr. Daniel Ripoll, Prof. Aneel K. Aggarwal, and Prof. Eva S. Vanamee for helpful discussions.

REFERENCES

1. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev S, Dryden DT, Dybvig K, Firman K, Gromova ES, Gumport RI, Halford SE, Hattman S, Heitman J, Hornby DP, Janulaitis A, Jeltsch A, Josephsen J, Kiss A, Klaenhammer TR, Kobayashi I, Kong H, Kruger DH, Lacks S, Marinus MG, Miyahara M, Morgan RD, Murray NE, Nagaraja V, Piekarowicz A, Pingoud A, Raleigh E, Rao DN, Reich N, Repin VE, Selker EU, Shaw PC, Stein DC, Stoddard BL, Szybalski W, Trautner TA, Van Etten JL, Vitor JM, Wilson GG, Xu SY. A nomenclature for restriction enzymes. DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Res 2003;311805–311812.
2. Pingoud A, Fuxreiter M, Pingoud V, Wende W. Type II restriction endonucleases: structure and mechanism. Cell Mol Life Sci 2005; 62:685–707.
3. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE–enzymes and genes for DNA restriction and modification. Nucl Acids Res 2007; 35(suppl_1):D269–D270.
4. Roberts RJ. How restriction enzymes became the workhorses of molecular biology. Proc Natl Acad Sci USA 2005;102:5905–5908.
5. Bujnicki JM, Rychlewski L. Grouping together highly diverged PD-(D/E)XK nucleases and identification of novel superfamily members using structure-guided alignment of sequence profiles. J Mol Microbiol Biotechnol 2001;3:69–72.
6. Kosinski J, Feder M, Bujnicki JM. The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. BMC Bioinformatics 2005;6:172.
7. Townson SA, Samuelson JC, Xu SY, Aggarwal AK. Implications for switching restriction enzyme specificities from the structure of BstYI bound to a BglII DNA sequence. Structure 2005;13:791–801.
8. Skrabanek L, Niv MY. Scan2S: Increasing precision of PROSITE pattern motifs using secondary structure constraints. Bioinformatics, in press.
9. Gutman R, Berezin C, Wollman R, Rosenberg Y, Ben-Tal N. QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns. Nucleic Acids Res 2005;33:W255–W261(Web Server issue).
10. Salwinski L, Eisenberg D. Motif-based fold assignment. Prot Sci 2001;10:2460–2469.
11. Chakrabarti S, Anand AP, Bhardwaj N, Pugalenthi G, Sowdhamini R. SCANMOT: searching for similar sequences using a simultaneous scan of multiple sequence motifs. Nucleic Acids Res 2005;33: W274–W276 (Web Server issue).
12. Gattiker A, Gasteiger E, Bairoch A. ScanProsite: a reference implementation of a PROSITE scanning tool. Appl Bioinformatics 2002;1:107–108.
13. Bork P, Koonin EV. Protein sequence motifs. Curr Opin Struct Biol 1996;6:366–376.
14. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
15. Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
16. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-coffee. Nucleic Acids Res 2006;34:W604–W608 (Web Server issue).
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
18. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999;291:177–196.
19. Mirny L, Shakhnovich E. Evolutionary conservation of the folding nucleus. J Mol Biol 2001;308:123–129.

20. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
21. Rost B, Eyrich VA. EVA: large-scale analysis of secondary structure prediction. Proteins 2001;45 (Suppl 5):192–199.
22. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
23. Biegert A, Mayer C, Remmert M, Soding J, Lupas AN. The MPI Bioinformatics Toolkit for protein sequence analysis. Nucleic Acids Res 2006;34:W335–W339 (Web Server issue).
24. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33:W244–W248 (Web Server issue).
25. Hsu C, Chen C, Liu B. MAGIIC-PRO: detecting functional signatures by efficient discovery of long patterns in protein sequences. Nucleic Acids Res 2006;34:356–361 (Web Server).
26. Jonassen I. Efficient discovery of conserved patterns using a pattern graph. Comput Appl Biosci 1997;13:509–522.
27. Kalinina OV, Novichkov PS, Mironov AA, Gelfand MS, Rakhmaninova AB. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. Nucleic Acids Res 2004;32:W424–W428 (Web Server issue).
28. Bujnicki JM. Crystallographic and bioinformatic studies on restriction endonucleases: inference of evolutionary relationships in the "midnight zone" of homology. Curr Protein Pept Sci 2003;4:327–337.
29. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
30. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–217.
31. Edgar RC. MUSCLE. Multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 2004;32:1792–1797.
32. Simossis VA, Heringa J. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res 2005;33:W289–W294 (Web Server issue).
33. Niv MY, Ripoll D, Vila JA, Liwo A, Vanamee ES, Aggarwal AK, Weinstein H, Scheraga HA. Topology of type II REases revisited; structural classes and the common conserved core. Nucl Acids Res 2007;35:2227–2237.
34. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14:33–38.
35. Pirovano W, Feenstra KA, Heringa J. Sequence comparison by sequence harmony identifies subtype-specific functional sites. Nucl Acids Res 2006;34:6540–6548.
36. Friedberg I, Margalit H. PeCoP: automatic determination of persistently conserved positions in protein families. Bioinformatics 2002;18:1276–1277.
37. Donald JE, Hubner IA, Rotemberg VM, Shakhnovich EI, Mirny LA. CoC: a database of universally conserved residues in protein folds. Bioinformatics 2005;21:2539–2540.
38. Kidera A, Konishi Y, Ooi T, Scheraga HA. Relation between sequence similarity and structural similarity in proteins—role of important properties of amino-acids. J Prot Chem 1985;4:265–297.
39. Glasser L, Scheraga HA. Investigation of a physical basis for conformational similarity in proteins. J Prot Chem 1991;10:273–285.
40. Grigoriev IV, Kim SH. Detection of protein fold similarity based on correlation of amino acid properties. Proc Natl Acad Sci USA 1999;96:14318–14323.
41. Melo F, Marti-Renom MA. Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. Proteins 2006;63:986–995.
42. Venkatarajan MS, Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. J Mol Model 2001;7:445–453.
43. Solis AD, Rackovsky S. Property-based sequence representations do not adequately encode local protein folding information. Prot Struct Funct Bioinform 2007;67:785–788.
44. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical-analysis of the physical-properties of the 20 naturally-occurring amino-acids. J Prot Chem 1985;4:23–55.
45. Maurer-Stroh S, Eisenhaber F. Refinement and prediction of protein prenylation motifs. Genome Biol 2005;6:R55.
46. Mathura VS, Schein CH, Braun W. Identifying property based sequence motifs in protein families and superfamilies: application to DNase-1 related endonucleases. Bioinformatics 2003;19:1381–1390.
47. Ginalski K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L. ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 2003;31:3804–3807.
48. Grigoriev IV, Zhang C, Kim SH. Sequence-based detection of distantly related proteins with the same fold. Prot Eng 2001;14:455–458.
49. Ginalski K, von Grotthuss M, Grishin NV, Rychlewski L. Detecting distant homology with meta-BASIC. Nucleic Acids Res 2004;32:W576–W581 (Web Server issue).
50. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–960.
51. Lesk AM, Chothia C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 1980;136:225–270.
52. Lesk AM, Chothia C. Evolution of proteins formed by beta-sheets. II. The core of the immunoglobulin domains J Mol Biol 1982;160:325–342.
53. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.