Scan2S, a novel regular expression scan with secondary structure constraints applied to Type II Restriction Endonucleases

Cornell University
Weill Medical College

Masha Y. Niv (1), Lucy Skrabanek (1), Richard J. Roberts (3), Harold Scheraga (2), Harel Weinstein (1)

(1) Weill Medical College of Cornell University; (2) Cornell University; (3) New England Biolabs

Abstract

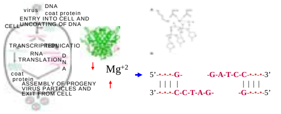
Motivation

Restriction endonucleases (REases) are DNA-cleaving enzymes that have become indispensable tools in molecular biology. REases exhibit structural and functional similarity and, in some cases, specificity for the same DNA sequences, despite dramatically dissimilar sequences. This makes it difficult to identify them in genomes and to classify them functionally based on sequence, and has hampered the efforts of specificity-engineering.

Results

We describe the derivation of novel REase sequence motifs, which extend beyond the PD-(D/E)XK hallmark and incorporate secondary-structure information. To enable automated searches using these novel motifs we developed a fast regular expression matching algorithm, that accommodate long patterns with optional secondary structure constraints. Using this new tool, Scan2S, motifs derived from REases with specificity towards particular DNA sequences (GATC and CCGG) are shown to identify REases of the same specificity. Notably, these motifs highlight potential specificity-determining residues, which can serve as candidates for specificity re-engineering.

Background


virus
DNA
coat protein
ENTRY INTO CELL AND UNCOATING OF DNA
CELL
TRANSCRIPTION
REPLICATIO
RNA
TRANSLATION
DNA
coat protein
ASSEMBLY OF PROGENY VIRUS PARTICLES AND EXIT FROM CELL
Mg+2
5'-•-•-•-G-   -G-A-T-C-C-•-•-•-3'
3'-•-•-•-C-C-T-A-G-   -G-•-•-•-5'

Type II Restriction endonucleases (REases)

•Bacterial defense against viruses

•Lab usage - recombinant DNA

Type II REases are very variable: length varies from 240 to over 1400 aa, sequence identities can be below 10%[1], and even structural topology varies[2]. Nevertheless, they are extremely specific. Specificity determinants are not well established and engineering attempts were unsuccessful so far[3].

Current computational tools detect only a small fraction of REases, and alignments usually need to be structure-assisted.

Aims

Identify motifs common to Type II REases that recognize specific DNA sequences
Use these motifs to detect REases with the same specificity
Transform the motifs into guides for protein engineering of specificity-determining sites

Methods

1) Motif generation

a)GATC-specific motif. Structure-based sequence alignment of GATC-recognizing Type II REases BamHI, BstYI and BglII was obtained with 3D-TCoffee[4]. Positions known to be involved in catalysis and fully conserved positions are included in the motif. In sites that are not in contact with the DNA, amino acid residues of the same physicochemical class as the conserved residue are allowed in the motif ("**physicochemical relaxation**"). The classes are: {AVLIMC}, {HWYF}, {NQST}, {ED}, {KR} and {GP}. **Secondary-structure constraints** are included in the motif for sites that reside in the same secondary elements in all the structures.

b) CCGG-specific motif. Structure-based sequence alignment was obtained for the CCGG-recognizing REases. MspI, NaeI, Cfr10I1, Bse634i and NgomIV. The motif derivation is as above, but the positions are considered conserved if the physicochemical class (rather than individual residue) is fully conserved in the five aligned sequences.

2)Type II REase sequences were downloaded from the REBASE database[3]. This set of sequences is referred to as REset. The secondary structure for REset sequences is predicted using PSIPRED[5].

3). The **Scan2S** step, which performs the search for the motifs derived in step 1 in the datasets prepared in step 2. The Scan2S program uses the Java 5.0 regex (regular expression) package which enables it to support long and flexible motifs, is fast, and enables it to include **secondary-structure constraints** in the query motif. Each position in the motif is followed by its secondary structure constraint, e.g., [FY]H means that a phenylalanine or tyrosine must be found in a helix. [ILV]. means that the residue at that position can be an isoleucine, leucine or valine, and that there is no secondary structure constraint imposed.

Results

Table 1. GATC motif

BamHI #	occurrence	allowed aa	secondary constraints
14	EEE	DE	
28	EEE	DE	not E
58	VVV	V(contact)	
61	KN	KN(putative catalytic)	H
68	LLL	AVLIMC	
74	WWW	WFYH	
84	KKK	RK	
94	DDD	D (catalytic)	
97	KKK	RK	
111	EEE	DE	E
113	ENQ	ENQ(catalytic)	
136	III	AVLIMC	
160	EEE	DE	not E
173	PPP	PG	

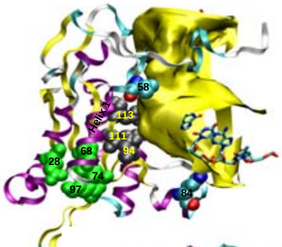


Figure 1. Conserved patches in GATC-recognizing Type II REases in 2BAM.pdb The catalytic residues 94, 111, 113 are colored grey. The conserved structural cluster (residues 28,68,74 and 97) are colored green. The non-catalytic conserved residues within 5Å from the DNA strand are colored by atom type.

Table 2. CCGG motif

BamHI numbering	MspI numbering	Observed Residues	CCGG	secondary constraints
57	31	GGGGG	G (contact)	not E
61	35	EEEEE	E (putative catalytic)	not E
64	38	ILIIC	AVLIMC	
94	99	DDDDD	D (catalytic)	not H
97	102	IIIIV	AVLIMC	
110	116	LVLVI	AVLIMC	not H
111	117	SNGD	SNGD (catalytic)	not H
112	118	ICVLC	AVLIMC	not H
113	119	KKKKK	K (catalytic)	not H
115	121	SSTST	ST(contact)	
140	205	ILAAV	AVLIMC	not H

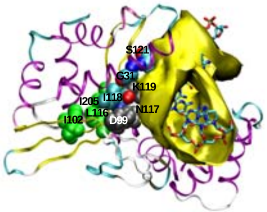


Figure 2. **Conserved residues in CCGG-recognizing Type II REases shown for 1SA3.pdb** The hydrophobic cluster (sites 101, 116, 205) is colored green. The catalytic residues 99, 117 and 119 are colored grey. The non-catalytic conserved residues 5A from the DNA strand (sites 31, 118 121) are colored by atom type.

GATC motif retrieves 13 GATC-recognizing REases and 1 unknown recognition specificity REase (100% precision [true positives / (true positives + false positives)] and 14% recall [true positives / (true positives + false negatives)]); without "physicochemical relaxation", the motif recalls only the three original sequences (3% recall, 100% precision); if the secondary-structure constraints are not included, the precision drops to 60%, (recall is still 14%). **CCGG motif** retrieves 11 CCGG REases, 2 non-CCGG and 1 unknown specificity (recall 36%, precision 85%). Exclusion of the secondary-structure constraints results in 50% recall but only 54% precision. Exclusion of "physicochemical relaxation" results in recall of 60%, but precision of 8%.

Comparison to performance of other programs

	recall	precision	TP	FP	FN	Unknown specificity	true hits not found by BLASTP	true hits not found by PRATT
Scan2S-GATC	14%	100%	13	0	78	1	10	11
PRATT-GATC	3%	100%	3	0	88	0	0	0
BLASTP-GATC	10%	100%	10	0	81	0	0	7
Scan2S-CCGG	36%	85%	11	2	19	1	4	7
PRATT-CCGG	20%	100%	6	0	24	0	0	0
BLASTP-CCGG	50%	100%	15	0	20	1	0	8

Table 3. Comparison of performance to other mehods
TP-True Positives, FP – False Positives, FN-False Negatives
recall [true hits / (true hits + false negatives)]
precision [true hits / (true hits + false positives)]
PRATT is an auotomated sequence pattern derivatiion method[6]

a. Challenge for all the methods we tested as indicated by low (3%-50%) recall.
b. True positive hits of different methods do not fully overlap. Thus Scan2S provides a ***complementary*** approach to BLASTP for searches against REset.
c. Hits of unknown specificity may have the specificity of the REases for which the motif was derived.

Conclusion and Outlook

1. Physicochemical information was used to a) identify the conserved sites b) relax the motifs

2. Structural information was used a) to align the sequences (3DTCoffee) b) derive secondary structure constraints (novel for regular expression motifs!) c) identify protein/DNA contacts

3. Scan2S motifs for GATC and CCGG REases specifically retrieves true hits not found by other methods.

4. Novel specificity-determining sites are a) subfamily-specific b) candidates for specificity re-engineering.

5. Scan2S improves precision of PROSITE motifs by inclusion of secondary structure constraints – work in progress, Skrabanek and Niv.

6. Scan2S is available upon request, man2016@med.cornell.edu,las2017@med.cornell.edu

References

1. Roberts, R.J., Vincze, T., Posfai, J. and Macelis, D. (2005) NAR
2. Niv, M.Y., Rippol, D., Vila, J., Liwo, A., Vanamee, E.S., Aggarwal, A.K., Weinstein, H. and Scheraga, H.A. (2006) submitted to *NAR.*
3. Townson, S.A., Samuelson, J.C., Xu, S.Y. and Aggarwal, A.K. (2005) *Structure*
4. *Armougom F. et al., (2006) NAR*
5. *Jones D.T. (1999) JMB*
6. Jonassen, I. (1997). *Computer Applications in the Biosciences*