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

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
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 varies2. Nevertheless, they are extremely specific. Specificity determinants are not well established and engineering attempts were unsuccessful so far3.

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-TCoffee4. Positions known to be involved in catalysis and fully conserved positions are included in the motif ("physicochemical relaxation"). The classes are: (AVLIMC), (HWFY), (NSQT), (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. Smpl, Nacl, Cfr101I, 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.

c) Type II REase sequences were downloaded from the REBASE database3. This set of sequences is referred to as REset. The secondary structure for REset sequences is predicted using PSIPRED5.

2) 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., [FYH] 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.

Conclusion and Outlook

1. Physicochemical information was used to a) identify the conserved sites b) relax the motifs
2. Structural information used a) to align the sequences (3DTCoffee b) to derive secondary structure constraints (novel for regular expression motif) 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.

Scan2S is available upon request, mm2016@med.cornell.edu, lskrabanek@med.cornell.edu

References

4. Altschul, S.F. et al. (1990) NAR