14. REPPER repeats and their periodicities in fibrous proteins

Markus Gruber, Johannes Soding, and Andrei N. Lupas

Max Planck Institute for Developmental Biology,
Spemannstr.35, 72076 Tubingen, Germany

REPPER (REPeats and their PERiodicities) detects and analyses regions with short gapless repeats in protein sequences or alignments. It finds periodicities by Fourier Transform and internal similarity analysis. Both methods use a sliding window to ensure that different periodic regions within the same protein are detected independently.
Availability: http://protevo.eb.tuebingen.mpg.de/repper.

INTRODUCTION

Many proteins display repeat patterns in their sequences. Most currently available repeat detection tools are homology-based and built to identify divergent, gapped repeats of variable length and spacing in the size range of 20 residues and above (i.e. supersecondary structures and domains). None of these methods is suited to detect repeats shorter than about 20 residues. Thus, these programs are not useful for analyzing one of the largest classes of repetitive proteins, the fibrous proteins, in which the repeat size is typically less than 15 residues.

For fibrous proteins the most commonly used tool for analysis is Fourier Transform (FT). This method has been widely used, particularly in the analysis of coiled coils, and has proven crucial for deducing properties of the tertiary structure, for example supercoil handedness.

In conjunction with programs that predict secondary structure and the occurrence of coiled coils, FT can be very powerful in the analysis of fibrous proteins. In addition, these methods can be usefully complemented by a sequence comparison tool (REPwin), which is conceptually similar to the ones named above, but tailored to detect short consecutive repeats by aligning a sequence to itself, shifted by multiples of a variable offset.

We have therefore built a server that implements new versions of FT (FTwin) and sequence self-comparison (REPwin) and combines their output with that of secondary structure prediction (PSIPRED) and coiled-coil prediction (COILS) into an integrated and detailed overview.

The programs are implemented using a sliding window, so as to show the boundaries of periodic regions and allow the detection of multiple regions with different periodicities in the same protein.

COMPONENT PROGRAMS

FTwin is a Fourier Transform analysis tool that employs a sliding window of user-defined size. A protein sequence is represented by a discrete function of real numbers. Two scales for the analysis of hydrophobic periodicity are provided by the program, one derived from the Kyte Doolittle hydrophobicity scale and the other reflecting a binary weighting of aliphatic residues. In addition, other scales can be set by the user.For a given sequence (or alignment of sequences) the program returns a graph with the significant periodicities as a function of the position in the sequence. The threshold

parameter as well as values for the window size and the periodicity range can be changed via the user interface.

Repeat patterns can also be found by sequence self-commparison. REPwin compares a protein sequence with itself, using the Gonnet similarity matrix and a sliding window of user-defined size. It returns a graph which shows regions of significant self-similarities with their corresponding periodicities. A similarity in the self-alignment is indicative of a region with a periodicity equal to the offset.

RESULTS

YadA from Yersinia enterocolitica contains a left-handed beta-helix with a degenerate periodicity close to 14 and a coiled-coil stalk with an unusual periodicity of 15 residues. FTwin and REPwin clearly identify the two regions with their correct periodicities. The 15-residue periodicity of the coiled coil differs markedly from the canonical seven-residue repeat and COILS therefore only returns intermediate probabilities. In fact, a 15-residue periodicity yields a helical structure with 3.75 residues per turn; it has a right-handed supercoil twist, which is of the same magnitude but opposite handedness to that of canonical coiled coils. This fact was predicted from theoretical considerations and was proven by the crystal structure of a 15-residue periodic protein.

USING PROFILES

In REPPER, the programs FTwin and COILS allow the user to take a multiple sequence alignment as input, and there is also the option to calculate a profile for a given single input sequence using PSI-BLAST with two iterations and an E-value cutoff of 0.001.

This can improve the accuracy of FTwin. The single sequence of the long coiled coil cortexillin does not display the typical periodicity of 3.5, although this coiled coil is regular. If an alignment is used as input, a periodicity of exactly 3.5 is revealed. As many coiled coils have exceptions to the hydrophobic-polar repeat pattern, the FT results get blurred, but as soon as other similar sequences are aligned the pattern becomes more pronounced and therefore more significant in the rsults.

Profiles may also lead to an improvement of COILS. For example, the Bag domain is a 2 three-helix bundle with features typical for coiled coils. In a single sequence analysis, only the first helix obtains high coiled-coil probabilities. The second helix is slightly deformed, which substantially lowers its score. When using a multiple sequence alignment as input, this discontinuity is averaged with many regular sequences, thereby markedly improving the scores for the second helix and yielding a better match to the structure.

CONCLUSION

FTwin and REPwin are two new programs for the prediction of periodic patterns in protein sequences. Although they are aimed primarily at the analysis of fibrous proteins, they can be used for any kind of repetitive sequence provided the following criteria are met: Repeats of the same nature must be consecutive in the sequence, must be of approximately the same size (no major insertions or deletions), and must occur in sufficient number to be detectable by Fourier Transform. This number is a functionofthe sequence similarity between the repeats; whereas nearly identical repeats can even be detected in occurrences of two to five, more degenerate repeatstypically requireat least tenoccurrences

(the sizeof the scanning window in REPPER must be set to reflect this). FTwin and REPwin are complementary, since REPwin searches for repeats in a general way, using a global amino acid replacement matrix, wheras FTwin searches for periodicities of particular, user-defined types (hydrophobic,polar, positively charged, etc.). Their combination with secondary structure and coiled-coil prediction into a single integrated server provides a powerful new tool for the analysis of protein sequences.