Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W281–W283***
doi:10.1093/nar/gki488

One-Block CYRCA: an automated procedure for identifying multiple-block alignments from single block queries

Milana Frenkel-Morgenstern, Alice Singer[1], Hagit Bronfeld[1] and Shmuel Pietrokovski*

Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel and [1]Bioinformatics undergraduate program, Bar-Ilan University, Ramat-Gan 52900, Israel

Received February 14, 2005; Revised March 3, 2005; Accepted April 20, 2005

ABSTRACT

One-Block CYRCA is an automated procedure for identifying multiple-block alignments from single block queries (http://bioinfo.weizmann.ac.il/blocks/OneCYRCA). It is based on the LAMA and CYRCA block-to-block alignment methods. The procedure identifies whether the query blocks can form new multiple-block alignments (block sets) with blocks from a database or join pre-existing database block sets. Using pre-computed LAMA block alignments and CYRCA sets from the Blocks database reduces the computation time. LAMA and CYRCA are highly sensitive and selective methods that can augment many other sequence analysis approaches.

INTRODUCTION

Comparison of multiple sequence alignments (profiles) with other profiles can identify subtle protein relationships beyond the resolution of sequence-to-sequence or sequence-to-profile comparisons (1–5,7–9). The main advantages of using profiles instead of sequences are better characterization of the compared regions and the possibility for giving more weight to, or using only, conserved regions. Using only conserved regions significantly reduces the search space and avoids possibly spurious hits by non-conserved and misaligned regions. Blocks are local ungapped profiles of the most conserved regions of protein families and domains (2).

LAMA is a profile-to-profile alignment method, previously developed by us, for comparing blocks with each other and for searching databases of blocks with block queries (5). It is a highly sensitive method for detecting sequence similarities that are often not found by other profile-to-profile and sequence-to-profile methods (6). LAMA alignments do not use gaps, since the compared profiles are short and are themselves constructed from ungapped conserved regions.

CYRCA is a method for detecting weak protein sequence similarities by aligning multiple blocks (3). The resulting multiple-block alignments are identified as block sets with consistent and transitive relationships, derived from pairwise block alignments previously found by LAMA. Namely, if blocks A, B and C are aligned to each other in the same phase in overlapping regions, then these blocks are probably genuinely similar to each other, even if each pairwise alignment score is insignificant by itself (Figure 1). CYRCA implements this approach by using graph theory and a bottom-up algorithm. Blocks are represented as graph nodes and their LAMA alignments as the graph edges. The simplest transitive block relationship is a triangle graph (a cycle of three blocks). CYRCA first identifies consistent triangles, joins triangles with common edges and finally adds linear edges that have very high alignment scores. CYRCA sets are, thus, identified from large-scale LAMA comparisons of many blocks with each other, typically using the whole Blocks database. These comparisons take a few days to compute. CYRCA analyses are used to annotate the Blocks database. Analysis of specific blocks has identified biologically significant and genuine relationships (10,11), but it requires manual interventions.

Here we present a procedure and a web server for automatically adding new blocks to previously constructed or constructing new CYRCA sets.

ONE-BLOCK CYRCA ALGORITHM

In the first step of the algorithm, each of the query blocks is compared by LAMA with the database for which CYRCA sets were previously computed (the current version of the Blocks database). All hits above the user-specified score threshold are retained. Next, all of the blocks found to be similar to the query or queries are compared by LAMA with each other. This is

*To whom correspondence should be addressed. Tel: +972 8 934 2747; Fax: +972 8 934 4108; Email: shmuel.pietrokovski@weizmann.ac.il

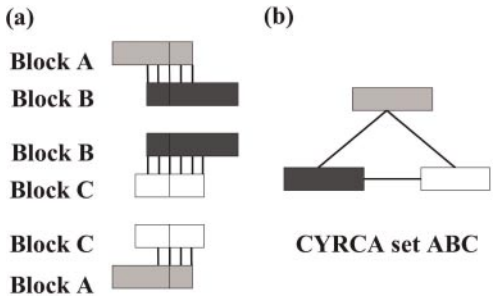

Figure 1. A graphical example of a consistent set of aligned blocks. (**a**) Three consistently aligned pairs of blocks A, B and C are presented. Blocks are presented as rectangles with one position marked by a vertical line. The aligned region is shown for each pair of aligned blocks. (**b**) A consistent CYRCA set obtained from the block alignments shown in (a). Such basic consistent graphs are then joined to form larger consistent sets (3).

actually implemented using pre-computed LAMA results. The resulting query hits (graph edges), together with the edges found within the database, are then examined using the CYRCA algorithm. This can identify new sets or join the block queries to existing CYRCA sets (Figure 2).

The One-Block CYRCA procedure analyzes the relationship of one or a few block queries to a database, whereas the basic CYRCA procedure inter-compares a whole database. We took advantage of this to use a more sensitive and time-consuming value for the CYRCA cycle size parameter. One-Block CYRCA sets are identified by first locating consistent cycles of any size, not just triangular ones as in the basic method of Kunin *et al*. (3). This allows a more in-depth analysis by One-Block CYRCA.

DESCRIPTION OF THE WEB INTERFACE

Input to the server (http://bioinfo.weizmann.ac.il/blocks/OneCYRCA) is one or more blocks supplied by the user. The blocks can be in the Blocks database format (2) (http://blocks.fhcrc.org) or in another commonly used multiple sequence alignment format (multiple FASTA, CLUSTAL or MSF). These latter formats can be found in many multiple alignments databases, such as Pfam (12), CDD (13) and SMART (14), and as the output of multiple alignment programs such as MEME (15), T-COFFEE (16) and DIALIGN (17). From these alignment types only ungapped regions wider than four columns will be used. The user can upload the input from a local file or can paste it into the query window. If an email address is provided, the output will be sent to it.

Parameter default values are supplied but can be changed by the user. The default *Z*-score threshold parameter of the LAMA alignment significance (5.6) corresponds to ~1% significance level. The 'Linear edge' threshold score parameter default value (8.0) is more selective, since it is used for adding to CYRCA sets linear edges whose consistency cannot be checked (3).

The output of the server includes the CYRCA sets found with the query blocks. These can be expanded pre-computed or new sets. If the queries were from the Blocks database, it is possible that the sets will be unchanged pre-computed ones. The sets are shown with the description of their blocks, list of all the set edges (pairwise alignments) and phase alignment of all blocks. There are also links to the block entries in the Blocks database and to interactive graph representations and superimposition of structures present in different blocks of each set (Figure 3).

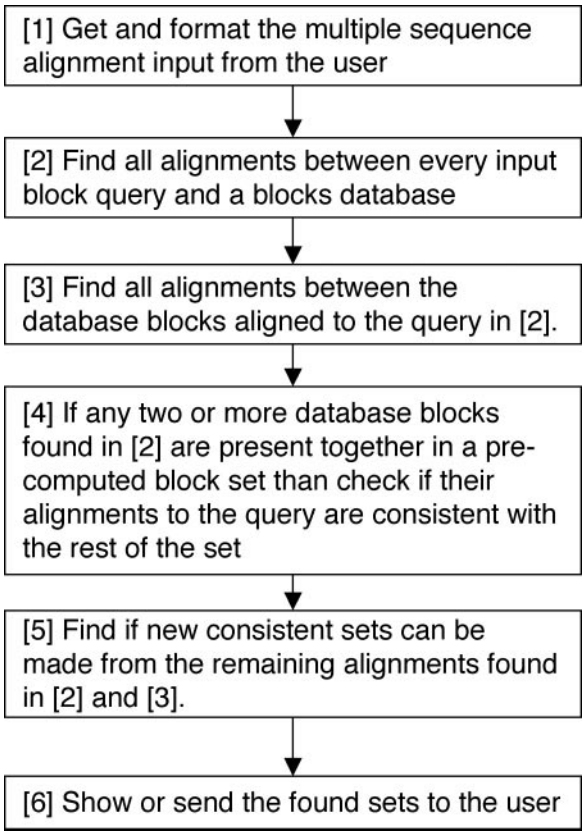

Figure 2. Flow diagram of the One-Block CYRCA procedure.

EXAMPLE

HNH and GIY-YIG nuclease domains are often accompanied by regions with conserved sequence motifs. We have previously shown that these motifs are similar to known DNA binding motifs and probably confer the substrate specificity to the nuclease catalytic regions (11). Our analysis was based on block-to-block alignments found by LAMA and CYRCA. This required careful manual intervention since the nuclease-associated modular DNA-binding domains (NUMODs) blocks we found were not part of the Blocks database. Submitting the NUMOD motifs to the One-Block CYRCA server returned the block sets we used to identify their function (Figure 3).

DISCUSSION

The high selectivity of One-Block CYRCA is derived from the transitive nature of its search. It is not a simple query-against-database search. All hits found by the query are examined further if they are consistently similar to each other. This identifies a set of similar blocks that can form a multiple-block alignment. This CYRCA approach was described by us in Ref. (3).

The One-Block CYRCA method has a novel combination of searching with short queries (corresponding to local sites on proteins) and using the powerful methodology of

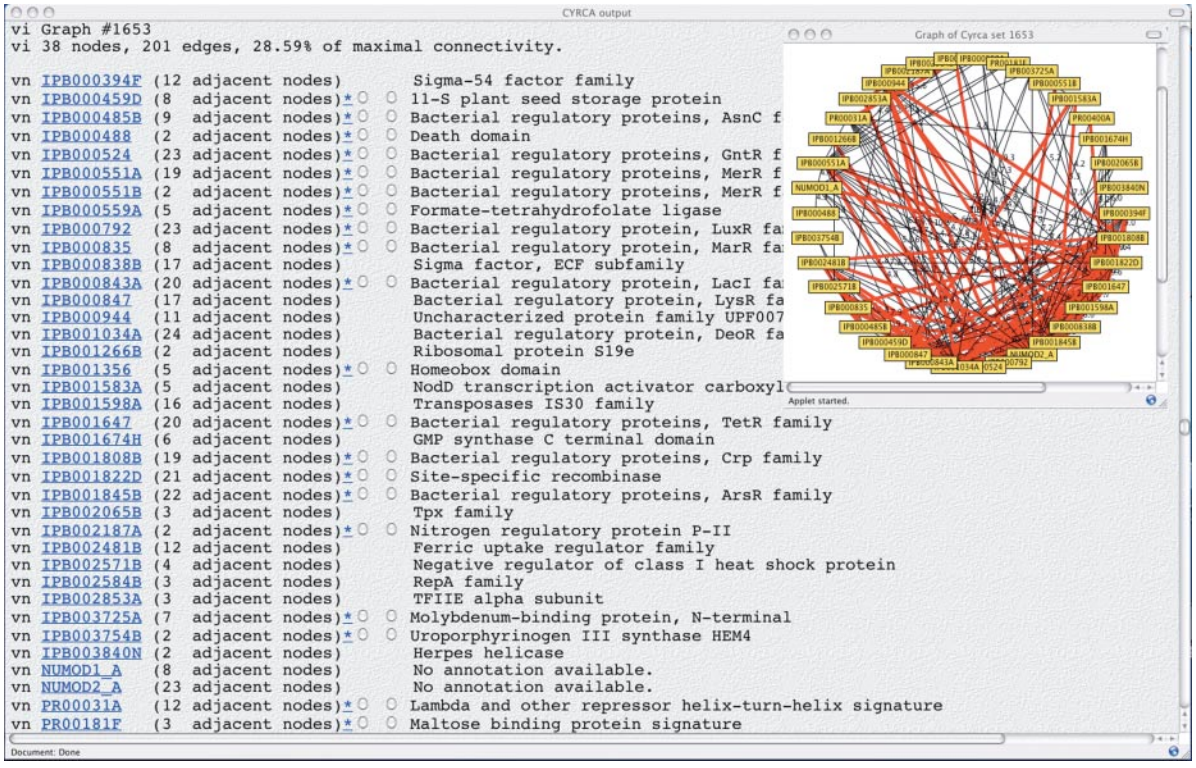

Figure 3. Representative output from the One-Block CYRCA server.

profile-to-profile methods. Other servers and programs either use short queries to compare with sequences or compare long gapped profiles (or HMMs) with other profiles. Our approach allows the identification of weak and localized similarity between proteins embedded in otherwise different contexts.

ACKNOWLEDGEMENTS

S.P. holds the Ronson and Harris Career Development Chair. Funding to pay the Open Access publication charges for this article was provided by the Weizmann Institute of Science Crown Human Genome, and Leon and Julia Forscheimer Center Molecular Genetics centers.

Conflict of interest statement. None declared.

REFERENCES

1. Gotoh,O. (1993) Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Comput. Appl. Biosci.*, **9**, 361–370.
2. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
3. Kunin,V., Chan,B., Sitbon,E., Lithwick,G. and Pietrokovski,S. (2001) Consistency analysis of similarity between multiple alignments: prediction of protein function and fold structure from analysis of local sequence motifs. *J. Mol. Biol.*, **307**, 939–949.
4. Panchenko,A.R. (2003) Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res.*, **31**, 683–689.
5. Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.
6. Frenkel-Morgenstern,M., Voet,H. and Pietrokovski,S. (2005) Enhanced statistics for local alignment of multiple alignments improves prediction of protein function and structure. *Bioinformatics*, doi:10.1093/bioinformatics/bti462.
7. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
8. Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.
9. Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.
10. Amitai,G., Belenkiy,O., Dassa,B., Shainskaya,A. and Pietrokovski,S. (2003) Distribution and function of new bacterial intein-like protein domains. *Mol. Microbiol.*, **47**, 61–73.
11. Sitbon,E. and Pietrokovski,S. (2003) New types of conserved sequence domains in DNA-binding regions of homing endonucleases. *Trends Biochem. Sci.*, **28**, 473–477.
12. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
13. Marchler-Bauer,A., Anderson,J.B., Cherukuri,P.F., DeWeese-Scott,C., Geer,L.Y., Gwadz,M., He,S., Hurwitz,D.I., Jackson,J.D., Ke,Z. *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.*, **33**, D192–D196.
14. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.
15. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
16. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
17. Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.