BIOINFORMATICS

Vol. 17 no. 0 2001
Pages 1–3

Mocca: semi-automatic method for domain hunting

Cédric Notredame

Information Genetique et Structurale, CNRS-UMR 1889, 31 Ch. Joseph Aiguier, 13 402 Marseille, France

Received on Month xx, 2000; revised and accepted on Month xx, 2000

ABSTRACT

Motivation: Multiple OCCurrences Analysis (Mocca) is a new method for repeat extraction. It is based on the T-Coffee package (Notredame *et al.*, *JMB*, **302**, 205–217, 2000). Given a sequence or a set of sequences, and a library of local alignments, Mocca extracts every segment of sequence homologous to a pre-specified master. The implementation is meant for domain hunting and makes it fast and easy to test for new boundaries or extend known repeats in an interactive manner. Mocca is designed to deal with highly divergent protein repeats (less than 30% amino acid identity) of more than 30 amino acids.

Availability: Mocca is available on request (cedric.notredame@europe.com). The software is free of charge and comes along with complete documentation.

Q: Please supply recevied date

Q: Applications Note?

Q: There are some differences in the electronic version and hardcopy. We follow electronic version

INTRODUCTION

Many proteins consist of separately evolved, independent structural units called modules or domains. The great diversity of protein functions is partly due to the vast number of possibilities to arrange a finite number of those basic units (Chothia, 1992). It is generally agreed that a domain is a self-folding unit made of a minimum of 25 amino acids (Bairoch *et al.*, 1997; Corpet *et al.*, 1998). Many of these domains appear as homologous subsequences repeated within a sequence or within a set of sequences, hence the importance of repeats identification in the course of domain hunting.

Many tools exist for discovering and extracting these repeats and without being exhaustive, one can cite PSi-Blast (Altschul *et al.*, 1997), Dot matrices (Junier and Pagni, 2000); Repro (Heringa and Argos, 1993) and the Gibbs Sampler (Lawrence *et al.*, 1993). More recently, Heger and Holm developed a method meant to scan databases for repeats without manual intervention (Heger and Holm, 2000). These automatic methods all share the same drawback: while none of them is 100% accurate, they give the user little scope for testing his own hypothesis in a seamless manner. Multiple OCCurrences Alignment (Mocca) addresses that specific problem. Given some approximate information concerning the whereabouts of one of the repeats (master repeat), it allows the user to tune the parameters describing the repeat family (i.e. start position, length of the master repeat and stringency of the search), and extract other occurrences of that repeat within the dataset. The procedure is fast and simple.

METHODS

Mocca uses a pair-wise sequence alignment algorithm (Durbin *et al.*, 1998). The cost associated with the alignment of each pair of residues uses the '*library extension*' developed for T-Coffee (Notredame *et al.*, 1998, 2000). Figure 1 outlines the strategy used to generate the T-Coffee scoring scheme. Firstly, a primary library is compiled; it contains a series of local alignments obtained using Lalign, an implementation of the Sim algorithm (Huang and Miller, 1991). Given two sequences, Lalign extracts the N top scoring non-overlapping local alignments. We used a modified version that compares two sequences (or a sequence with itself), and extracts every top scoring alignment having a length longer than ten residues and an average level of identity higher than 30%. Lalign reports each alignment along with a score that indicates its statistical significance. In our primary library, such local alignments appear as a series of pairs of residues where each pair receives a weight equal to the score of the alignment it comes from. Given a set of N sequences, the library contains the result of all the possible pair-wise comparisons (including the self-comparisons). This library is fed into T-Coffee to generate the position specific scoring scheme using the '*library extension*' algorithm (Notredame *et al.*, 2000).

In Mocca, a pre-requisite to repeat extraction is the estimation of at least one basic unit repeat among the sequences being analysed (master repeat). In the context of this work, we made the estimation using dotlet, a Java-based dot matrix method (Junier and Pagni, 2000). The master repeat is a sub-string selected within the sequence(s) used to build the library. Mocca extracts every sub-string homologous to the master in a single pass over the target sequences. It is the library extension that

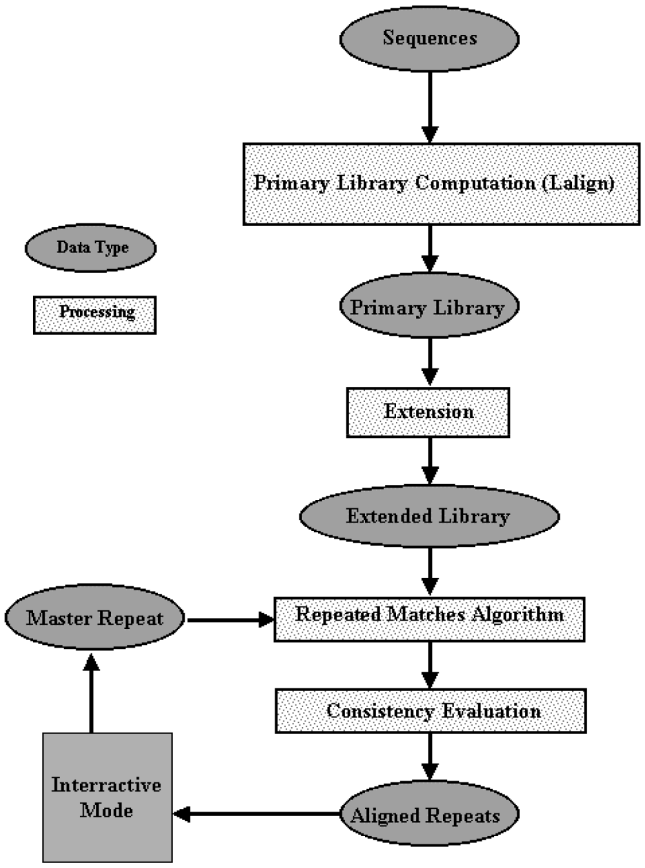

Fig. 1. Layout of the Mocca strategy. The main steps required to extract a repeat with Mocca method are shown. Square blocks designate procedures while rounded blocks indicate data structures.

makes it possible for a single repeat to '*recognize*' each of its homologues (even the distant ones). The extraction process relies on a very efficient dynamic programming procedure known as repeated matches (Durbin *et al.*, 1998). This algorithm reports a series of non-overlapping sub-strings each of them having an alignment to the master associated with a score higher than some pre-specified threshold Th. Th is empirically set to be a function of the maser repeat length (L):

$$Th = S * L$$

S has a value between 0 and 1. By default, $S = 0.05$, but its value can be modified interactively. Two other parameters can also be modified to increase sensitivity and accuracy: the gap opening penalty and the gap extension.

Mocca is part of the T-Coffee package. It is written in Perl and ANSI C. It runs on any UNIX or LINUX platform. It is available free of charge along with documentation. Copies can be obtained on request by sending an e-mail to cedric.notredame@europe.com. The main computational requirement is the Lalign library $O(N^2L^2)$, the motif extraction itself only requires little time (12 s on an IRIX O^2 station for 20 sequences totalling 5000 residues). If the position of one of the repeats is known, the procedure can also be run automatically from the command line. It is recommended to use Mocca in conjunction with other means for the initial estimation of the repeat boundaries (PSi-Blast, Altschul *et al.*, 1997; Dotlet, Junier and Pagni, 2000; Dotter, Sonnhammer and Durbin, 1995;...). Our tests show that Mocca can properly deal with sets of repeats whose multiple alignment indicate less than 15% average identity.

While we currently use Lalign as a source of local information, any other sensible source could be considered. For instance, structural information could easily be added to our procedure, using off the shelf libraries of local structural similarities such as the Dali Domain Dictionary (Holm and Sander, 1998). The input format of Mocca is straightforward and well documented. Mocca is a refinement tool for the discovery and the establishment of new domains. If the master repeat is replaced with a profile or a collection of known characterized repeats, Mocca could also be used to improve the model of a given repeat family and extend the predictive power of its profiles.

ACKNOWLEDGEMENTS

The author wishes to thank the following people: Des Higgins for very helpful comments. Jaap Heringa, Philipp Bucher and Kay Hoffmann for useful discussions and advice at an early stage of the project, Hiroyuki Ogata for helpful comments on the program.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.*, **25**, 217–221.

Chothia,C. (1992) Proteins: 1000 families for the molecular biologist. *Nature*, **357**, 543–544.

Corpet,F., Gouzy,J. and Kahn,D. (1998) The ProDom database of protein domain families. *Nucleic Acids Res.*, **26**, 323–326.

Durbin,R., Eddy,S., Krogh,A. and Mitchinson,G. (1998) *Biological Sequence Analysis*. 1 vols, Cambridge University Press, Cambridge.

Heger,A. and Holm,L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins*, **41**, 224–237.

Heringa,J. and Argos,P. (1993) A method to recognise distant repeats in protein sequences. *Proteins: Struct. Funct. Genet.*, **17**, 391–411.

Holm,L. and Sander,C. (1998) Dictionary of recurrent domains in protein structures. *Proteins*, **33**, 88–96.

Huang,X. and Miller,W. (1991) A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.*, **12**, 337–357.

Junier,T. and Pagni,M. (2000) Dotlet: diagonal plots in a web browser. *Bioinformatics*, **16**, 178–179.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Notredame,C., Holm,L. and Higgins,D.G. (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, **14**, 407–422.

Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel algorithm for multiple sequence alignment. *JMB*, **302**, 205–217.

Sonnhammer,E.L. and Durbin,R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1-10.

To be balanced at final stage