BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 6 2004, pages 974–975
DOI: 10.1093/bioinformatics/bth032

SQUARE—determining reliable regions in sequence alignments

Michael L. Tress, Osvaldo Graña and Alfonso Valencia*

CNB-CSIC, Calle Darwin, Cantoblanco, 28049 Madrid, Spain

Received on August 6, 2003; Revised on November 6, 2003; accepted on November 7, 2003
Advance Access publication February 5, 2004

ABSTRACT

Summary: The Server for Quick Alignment Reliability Evaluation (SQUARE) is a Web-based version of the method we developed to predict regions of reliably aligned residues in sequence alignments. Given an alignment between a query sequence and a sequence of known structure, SQUARE is able to predict which residues are reliably aligned. The server accesses a database of profiles of sequences of known three-dimensional structures in order to calculate the scores for each residue in the alignment. SQUARE produces a graphical output of the residue profile-derived alignment scores along with an indication of the reliability of the alignment. In addition, the scores can be compared against template secondary structure, conserved residues and important sites.

Availability: http://www.pdg.cnb.uam.es/servers/square

Contact: mtress@gredos.cnb.uam.es

INTRODUCTION

Protein sequence alignments underpin much of protein bioinformatics. They are central to most fold detection, multiple alignment and modelling approaches, yet they are often one of the least considered aspects. Even now when a researcher is faced with an alignment that has been returned from an automatic approach, there is little that they can do to quickly check the validity of the alignment and instead have to resort to laborious 'by hand' methods.

We have recently described an approach that goes some way towards making up for this lack of information (Tress *et al*., 2003). The method assesses the reliability of alignments between unknown query sequences and template sequences of known structure. Alignments are evaluated using scores derived from profiles built around the template sequences. One of the advantages of this technique is that the alignment is evaluated residue by residue, thus helping the researcher decide which parts of an alignment are reliable and which parts are not.

The approach works for pairwise alignments generated by any method, as long as the template sequence has a known structure. Here, we introduce SQUARE (Server for Quick Alignment Reliability Evaluation) the Web-based version of this method.

*To whom correspondence should be addressed.

PRE-PREPARATION

Sequence profiles were generated for the 19 840 non-identical chains present in the current edition of the PDB structural database (Berman *et al*., 2000). Profiles, or position-specific score matrices (PSSMs), were constructed by running PSI-BLAST (Altschul *et al*., 1997) for just four iterations with a local non-redundant (90%) database and a generous *E*-value inclusion cut-off of 0.01. The PSI-BLAST PSSMs were converted into numerical matrix files using the first step of the IMPALA process (Schaffer *et al*., 1999).

METHOD OVERVIEW

SQUARE can calculate the residue-by-residue profile-derived alignment scores given an initial alignment between a target sequence and template of known structure (Fig. 1, an overview of the process).

For each template sequence residue there is a set of 20 associated probability scores in the corresponding numerical IMPALA matrix file one for each amino acid. A value can be extrapolated from the IMPALA matrix for each aligned residue, resulting in a string of profile-derived alignment scores. These alignments scores, a sierra of jagged peaks and troughs on their own, are smoothed by the server using a triangular five-residue window (Tress *et al*., 2003).

The smoothed alignment scores are complemented by an indication of which residues are reliably aligned. Reliably aligned residues are calculated directly from the profile-derived alignment scores and can be modified by three user-defined input options.

The options that define reliably aligned regions have been tested on alignments from the multiple alignment program CLUSTAL (Thompson *et al*., 1994), the sequence profile-based methods, IMPALA, SAM T99 (http://www.cse.ucsc.edu/research/compbio/HMM-apps/) and PSI-BLAST, and the fold-recognition programs GenTHREADER (http://www.psipred.net) and 3DPSSM (http://www.sbg.bio.ic.ac.uk/~3dpssm/). The default values worked

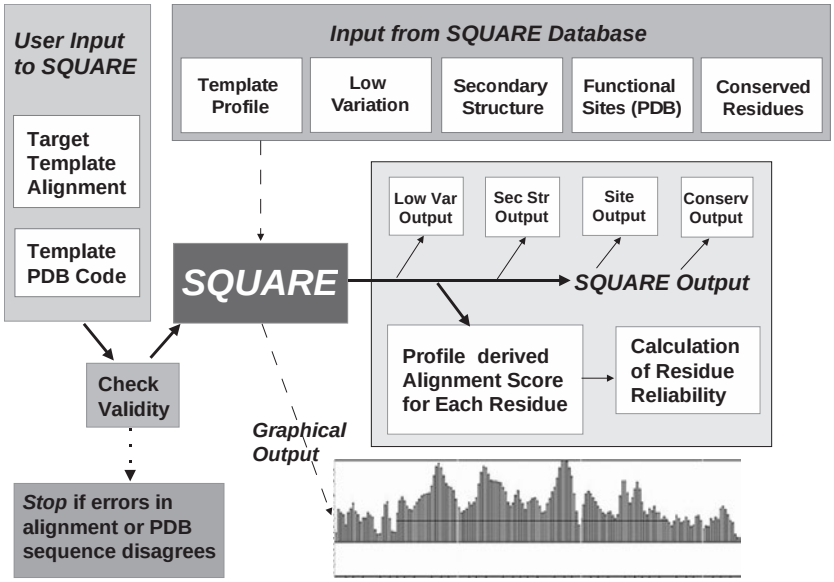

Fig. 1. An overview of the SQUARE process. Includes an example graphical output. Those regions considered to be reliably aligned are marked with a horizontal black line in the graphical output.

for all methods, but the cut-off values for alignments from those methods that also use template sequence PSI-BLAST profiles such as GenTHREADER and IMPALA, could be stricter, while the cut-off values could be relaxed for SAM T99, a method that avoids PSI-BLAST profiles altogether.

Sometimes large portions of alignments, or even whole alignments, may have unexpectedly poor scores. Often the explanation for scores that are sub-par is that there is low sequence variability in the template PSI-BLAST profile. The numerical matrix files were scanned for residues exhibiting symptoms of low sequence variation and this information is also displayed on the output page.

In addition, template secondary structure designations from DSSP files (Kabsch and Sander, 1983), are displayed to allow cross-checking. Another reason for including this information in the output is that predictions made for strand and helix regions tended to be even more reliable than those made in loop regions.

Residues involved in 'important' sites (generally binding sites) are derived from those PDB files that contain a SITES register. If the template chain does not contain site information, it is extracted from an identical PDB sequence. Our study showed that the method is particularly effective at predicting alignment reliability for important site residues, especially when the aligned target and template residues are identical amino acids.

The server also displays a conservation score for each residue in the template sequence. Residue conservation is taken directly from the HSSP database (Sander and Schneider, 1991) and is a further aid to alignment reliability evaluation.

ACKNOWLEDGEMENTS

We acknowledge continuous support and interesting discussions with the Protein Design Group at CNB-CSIC. This work was supported in part by TEMBLOR grant, reference: CE: QLRI-CT-2001-00015.

REFERENCES

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

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,F., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.J. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Sander,C. and Schneider,R. (1991) Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Tress,M.L., Jones,D.T. and Valencia,A. (2003) Predicting reliable regions in protein alignments from sequence profiles. *J. Mol. Biol.*, **330**, 705–718.