Structural bioinformatics

QUASAR—scoring and ranking of sequence–structure alignments

Fabian Birzele*,†, Jan E. Gewehr† and Ralf Zimmer
Practical Informatics and Bioinformatics Group, Department of Informatics, Ludwig-Maximilians-University, Amalienstrasse 17, D-80333 Munich, Germany

Received on August 19, 2005; revised on September 27, 2005; accepted on October 6, 2005
Advance Access publication October 10, 2005

ABSTRACT
Summary: Sequence–structure alignments are a common means for protein structure prediction in the fields of fold recognition and homology modeling, and there is a broad variety of programs that provide such alignments based on sequence similarity, secondary structure or contact potentials. Nevertheless, finding the best sequence–structure alignment in a pool of alignments remains a difficult problem. QUASAR (quality of sequence–structure alignments ranking) provides a unifying framework for scoring sequence–structure alignments that aids finding well-performing combinations of well-known and custom-made scoring schemes. Those scoring functions can be benchmarked against widely accepted quality scores like MaxSub, TMScore, Touch and APDB, thus enabling users to test their own alignment scores against ‘standard-of-truth’ structure-based scores. Furthermore, individual score combinations can be optimized with respect to benchmark sets based on known structural relationships using QUASAR’s in-built optimization routines.

Availability: The software, examples, the Java documentation and a tutorial are available at http://www.bio.ifl.lmu.de/QUASAR
Contact: fabian.birzele@ifl.lmu.de

1 INTRODUCTION
With the growing gap between the number of known protein sequences in databases like Swiss-Prot/TrEMBL (Boeckmann et al., 2003) and the number of experimentally determined protein structures in the PDB (Berman et al., 2000), automated structure prediction methods have become valuable tools for assigning potential coordinate models to new protein sequences. The first step to building a complete all-atom model is often to align a sequence of unknown structure (the so-called target) to a database of sequences with known structures (so-called templates). On the basis of these alignments and the underlying known template structures, models are built and refined. Since alignment quality determines the model quality, it is desirable to identify good models at the alignment stage in order to get rid of the overhead of producing obviously unsuitable coordinate models. This mainly restricts efforts to sequence– and secondary-structure-based measures (i.e. alignment scores) instead of using structural properties.

The QUASAR (quality of sequence–structure alignments ranking) system has been designed to fit two needs. First, it is a platform-independent and easily extendable software package for scoring and ranking sequence–structure alignments coming from different sources. Second, it aids the process of developing, benchmarking and optimizing new alignment quality measurements. The graphical user interface (GUI) of QUASAR provides quick access to each of the possible use cases and allows for visualization and comparison of the results as well as for configuration of all essential parts. Once configured, QUASAR can also be used directly from the command-line.

2 METHODS
2.1 Scoring alignments
The so-called scoring schemes represent alignment quality scores that require only information that is available from the sequence (e.g. predicted secondary-structure) or that can be directly inferred from the template structure. Scoring schemes provided by the system include several amino acid and secondary-structure-based exchange matrices [like PAM (Dayhoff et al., 1978) and (Luthy et al., 1991)], the two standard secondary-structure fit measures Q3 and SOV (Zemla et al., 1999) as well as two contact-capacity-based scores (Barrera et al., 2003; Singer et al., 2002). The number of available scoring schemes can be easily extended by implementing a Java interface or, in the case of (amino acid exchange) scoring matrices, by adding a text file in a QUASAR specific format that contains the matrix information. This provides a fast connection to matrix collections such as the AAIndex database (Kawashima et al., 1999).

2.2 Combining scores
With the so-called score conductor, the user can integrate several scoring schemes into one scoring function by combining the scores in a weighted sum (assigning user-specified weights for the single scores), i.e. as a linear combination of the individual scores. In addition, by editing the configuration file, experienced users can build more complex, tree-like formulas using further operators like multiplication and division. Therefore, a user can test different combinations of scoring schemes with a minimal amount of extra time and thus improve the ranking quality over the performance of the single scores. The final quality score of every alignment is calculated by combining the single alignment scores according to the formula given in the configuration. Single scores can also be normalized to range between zero and one to combine scores with different magnitudes.

2.3 Benchmarking scores
To help the user find a scoring function that gives the best possible results, QUASAR contains a number of structure-based quality scores like Touch, APDB (O’Sullivan et al., 2003), as well as reimplementations of MaxSub (Siew et al., 2000) and TMScore (Zhang and Skolnick, 2004), both based on

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
3 USE CASES

3.1 Benchmarking and optimization

A first, interactive use case might be as follows: given a new scoring scheme, e.g. a new scoring matrix,

(1) One builds a benchmark set of alignments and loads the data into QUASAR.

(2) In QUASAR, one explores the performance of the new scoring matrix in comparison with and in combination with in-built scores. The evaluation is done with respect to the standard-of-

(3) One further improves the ranking performance by combining well-performing schemes and optimizing their weights using QUASAR’s optimization routines.

(4) Now, one saves the configuration for future use of QUASAR from the command-line.

3.2 Automated alignment ranking

A second, non-interactive use case is the ranking of sequence–structure alignments. Here, one already has an optimized combination of scores together with the corresponding QUASAR configuration at hand. Given a set of different sequence–structure alignments for a target (e.g. to different template structures), one includes the call of QUASAR using the configuration file into the structure prediction process and is thus able to e.g. discard alignments on the basis of the previously optimized alignment score automatically.

4 CONCLUSION

Sequence–structure alignments play an important role in protein structure prediction and analysis. With QUASAR we provide a software that facilitates alignment scoring, comparison of known with user-defined scoring schemes and optimization of score combinations. It is platform-independent and can be used interactively or from the command-line. Future extensions will include new scoring schemes, improved analysis of results and more optimization options. We encourage users to send their own scoring schemes in order to have them included in future releases.

ACKNOWLEDGEMENTS

This work was funded by the German Research Foundation (DFG) under project grant PROSEQ II (ZI616/2).

Conflict of Interest: none declared.

REFERENCES


