BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 6 2012, pages 880–881
doi:10.1093/bioinformatics/bts003

Structural bioinformatics

Advance Access publication January 11, 2012

RIBER/DIBER: a software suite for crystal content analysis in the studies of protein–nucleic acid complexes

Grzegorz Chojnowski[1,*], Janusz M. Bujnicki[1,2] and Matthias Bochtler[1,3]

[1]International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, [2]Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Umultowska 89, 61-614 Poznan and [3]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawinskiego 5a, 02-106 Warsaw, Poland

Associate Editor: Burkhard Rost

ABSTRACT

Summary: Co-crystallization experiments of proteins with nucleic acids do not guarantee that both components are present in the crystal. We have previously developed DIBER to predict crystal content when protein and DNA are present in the crystallization mix. Here, we present RIBER, which should be used when protein and RNA are in the crystallization drop. The combined RIBER/DIBER suite builds on machine learning techniques to make reliable, quantitative predictions of crystal content for non-expert users and high-throughput crystallography.

Availability: The program source code, Linux binaries and a web server are available at http://diber.iimcb.gov.pl/ RIBER/DIBER requires diffraction data to at least 3.0 Å resolution in MTZ or CIF (web server only) format. The RIBER/DIBER code is subject to the GNU Public License.

Contact: gchojnowski@genesilico.pl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 12, 2011; revised on December 6, 2011; accepted on December 27, 2011

1 INTRODUCTION

Protein crystallographers who work on protein–nucleic acid complexes routinely face the problem that crystals that grow in a co-crystallization experiment do not necessarily contain all components present in the solution. The crystal content can be clarified by spectroscopic methods, but the equipment for such measurements is not commonly available. Alternatively, crystals can be washed, dissolved and analyzed by gel electrophoresis with appropriate staining, but this method is labor intensive, destructive and does not always provide a clear-cut answer.

In this work, we present the new program RIBER for detecting the presence of RNA stems in macromolecular crystals based on diffraction data alone. RIBER complements the previously developed program DIBER (Chojnowski and Bochtler, 2010) intended to search for double-stranded B-DNA and not double-stranded A-RNA (Table 1). The two programs are implemented as a stand-alone software suite and a web server RIBER/DIBER providing an easy way to judge nucleic acid content of a crystal based on a diffraction dataset, before the crystal structure is solved.

*To whom correspondence should be addressed.

Table 1. The classification performances of RIBER and DIBER

Crystal content	Correct predictions (%)	
	DIBER (combined mode)	RIBER
DNA	95.0 (90.0)	93.3
Protein–DNA	80.5 (89.2)	52.4
Protein	82.1 (83.9)	**80.9 ± 0.3**
Protein–RNA	65.2 (59.3)	**65.8 ± 0.3**
RNA	44.3 (44.3)	**74.7 ± 0.3**

Bold font represents estimated classifier performance. Regular font values are the correct prediction rates for datasets not used in the classifier training (see Section 2 for details).

The method may help to avoid a laborious phasing procedure when the component or the complex of interest is not present in the crystal.

2 MATERIALS AND METHODS

2.1 Diffraction data used for training and testing the classifiers

Crystal structures solved at 3.0 Å or better were downloaded from the Protein Data Bank (Bernstein, *et al.*, 1977) together with the corresponding experimental diffraction data. All reported calculations are based on experimental diffraction data. Structural information was only used to select and classify datasets according to their macromolecular content. Detailed information about curating the datasets used for training the RIBER classifier and DIBER benchmarks are available in Supplementary Material.

2.2 RIBER performance estimates

The RIBER classification performance was estimated using a repeated subsampling validation procedure. The classifier was trained with equal numbers of randomly selected diffraction datasets from each class (50% of instances of least numerous set of RNA only crystals). The remaining structures were used for testing. The average classification performance from 100 training and testing cycles was used as an estimate of the true classification performance.

2.3 Implementation

The program is written in C/C++ and relies on the CCP4 (Winn, et al., 2011) and Clipper (Cowtan, 2003) libraries for handling diffraction data. The LIBSVM (Chang and Lin, 2011) library is used for decision making.

3 RESULTS

Both RIBER and DIBER extract two parameters from the dataset: the first is a measure of a unit cell size and is primarily used to distinguish nucleic acid crystals from all others. The second parameter is a measure for the largest local average of reflection intensities. A large value for this parameter indicates the presence of very characteristic diffraction signals related to the regular stacking of A-RNA or B-DNA base pairs. A support vector machine (SVM) is used to make a prediction, using either only the two parameters described above or optionally also a third score (combined mode, available for DIBER only), which is calculated with the help of the molecular replacement program PHASER (McCoy *et al.*, 2007) for those users who hold a license (free for academic users). DIBER and RIBER use similar parameterizations of the diffraction data and a SVM to classify crystal content.

3.1 DIBER benchmark

The DIBER program has been benchmarked with the structures which have appeared in the PDB since the stand-alone version of the program was developed in 2009 (as described in the Supplementary Material). Within the error limits, the benchmark results obtained for protein and DNA-only agree with the previously published performance estimates [(Chojnowski and Bochtler, 2010), Table 1 and Supplementary Table S2]. Surprisingly, the currently observed correct classification rate for protein–DNA complexes is higher than previously reported. The discrepancy is due to a change in the composition of the test set. The new sample contains more protein DNA complexes with long helices, which produce strong diffraction signals that are easy to detect (75 versus 23% of molecules with more than 10 Watson–Crick base pairs).

3.2 Judging RNA content of a crystal with RIBER

Unlike DNA, naturally occurring RNA molecules rarely display long, regular double-stranded helices (Saenger, 1984). More often, they form large, complex structures with short double-stranded stems connected by single-stranded loops. However, RNA and DNA crystals share common features. First, similarly to DNA-only crystals, crystals that contain only RNA tend to have smaller unit cells than crystals with both RNA and protein (Supplementary Fig. S1a). Second, the base pairs forming RNA stems are often regularly stacked and produce characteristic diffraction signals analogous to the ones observed for double-stranded B from DNA helices (Supplementary Fig. S1b).

Therefore, the RIBER classifier is based on the parameterization used originally in DIBER to judge DNA content of a crystal. However, the program and SVM parameters were optimized with respect to the classification performance between RNA, protein–RNA and protein-only crystals. The benchmarks of a resulting RIBER classifier are presented in Table 1.

The RIBER performance has also been tested on a set of structures containing single-stranded, but not double-stranded RNA (both alone and in complex with proteins) that were originally rejected from the training set. As could be expected from the paucity of regularly stacked bases, most of these were misclassified as pure proteins (Supplementary Table S1).

4 CONCLUSIONS

In this article, we confirm earlier estimates of a very high performance of DIBER in discriminating between crystals formed by protein alone, double-stranded B-DNA alone and protein–B-DNA complex (Chojnowski and Bochtler, 2010). We also show that DIBER performs poorly for RNA, i.e. it fails to confidently discriminate protein alone versus protein–RNA complex versus RNA alone (it performs well for protein and protein–RNA complexes at the expense of RNA-only crystals, Table 1). RIBER, however, performs for double-stranded RNA much better than DIBER. Hence, RIBER complements DIBER for analyses of crystal content in crystallization trials of protein–nucleic acid complexes. The overall performance of RIBER is noticeably weaker than DIBER which is not surprising. RNA and proteins are alike in terms of structural complexity. This makes their crystals difficult to distinguish based on limited information provided by the diffraction data. Nonetheless, the discriminative power of RIBER is significant and we believe it will be a useful tool, in particular in situations, where the selection of most promising crystals for the diffraction and structure solution is a crucial factor to maximize the success in structure determination.

ACKNOWLEDGEMENTS

Funding: Foundation for Polish Science (TEAM/2009-4/2 to J.M.B., START fellowship to G.Ch.); European Commission (Health-Prot, contract number 229676); National Science Center grants (N N302 654640 and N N301 425038 to M.B.).

Conflict of interest: none declared.

REFERENCES

Bernstein,F.C., *et al.* (1977) Protein Data Bank - computer-based archival file for macromolecular structures. *Eur. J. Biochem.*, **80**, 319–324.

Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 1–27.

Chojnowski,G. and Bochtler,M. (2010) DIBER: protein, DNA or both? *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 643–653.

Cowtan,K. (2003) The Clipper C++ libraries for X-ray crystallography. *IUCr Comput. Comm. Newslett.*, **2**, 4–9.

McCoy,A.J. *et al.* (2007) Phaser crystallographic software. *J. Appl. Crystallogr.*, **40**, 658–674.

Saenger,W. (1984) *Principles of Nucleic Acid Structure*. Springer advanced texts in chemistry. Springer, New York.

Winn,M.D., *et al.* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. D*, **67**, 235–242.