Author Query Form

Journal: *Bioinformatics*
Article Doi: 10.1093/bioinformatics/btw646
Article Title: MIToS.jl: mutual information tools for protein sequence analysis in the Julia language
First Author: Diego J. Zea
Corr. Author: Cristina Marino-Buslje

AUTHOR QUERIES – TO BE ANSWERED BY THE CORRESPONDING AUTHOR

The following queries have arisen during the typesetting of your manuscript. Please click on each query number and respond by indicating the change required within the text of the article. If no change is needed please add a note saying "No change."

AQ1: Please check that all names have been spelled correctly and appear in the correct order. Please also check that all initials are present. Please check that the author surnames (family name) have been correctly identified by a pink background. If this is incorrect, please identify the full surname of the relevant authors. Occasionally, the distinction between surnames and forenames can be ambiguous, and this is to ensure that the authors' full surnames and forenames are tagged correctly, for accurate indexing online. Please also check all author affiliations.

AQ2: Please check and confirm the affiliations for the authors have been correctly identified, and correct if necessary.

AQ3: Please provide city for first affiliation.

AQ4: Please check whether the short title is OK as set.

AQ5: Please check that the text is complete and that all figures and their legends are included.

AQ6: Figure has been placed as close as possible to its first citation. Please check that it has no missing sections and that the correct figure legend is present.

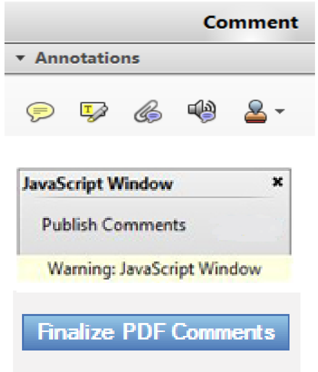
AQ7: Please note that there will be a charge of £100/US$190 per page for extra printed pages above 7 pages for an Original Article, 4 pages for a Discovery Note and 2 pages for an Applications Note. If your article exceeds these lengths, please confirm that you accept the charge.

AQ8: Please note, colour figures in print will be charged £350/$600 per figure. Colour figures online will not be charged. Please confirm which figures should be printed in colour and reword the legend/text to avoid using reference to colour if the figures are to be printed in black and white.

AQ9: Please provide a Funding statement, detailing any funding received. Remember that any funding used while completing this work should be highlighted in a separate Funding section. Please ensure that you use the full official name of the funding body, and if your paper has received funding from any institution, such as NIH, please inform us of the grant number to go into the funding section. We use the institution names to tag NIH-funded articles so they are deposited at PMC. If we already have this information, we will have tagged it and it will appear as coloured text in the funding paragraph. Please check the information is correct.

AQ10: Please check and confirm the edit made in Fig. 1 caption.

Making corrections to your proofs



- Please use the tools in the Annotation and Drawing Markups toolbars to correct your proofs. To access these, press 'Comment' in the top right hand corner.
- If you would like to save your comments and return at a later time, click **Publish Comments**.
- If applicable, to see new comments from other contributors, press **Check for New Comments**
- Once you have finished correcting your article, click **Publish Comments** and then **Finalize PDF comments**. Do not click **Finalize PDF comments** before you have finished correcting your proof.
- The **Publish Comments** option is available as a pop up JavaScript Window, shown right. **Do not close the Javascript window.**

Annotation tools

Insert text at cursor: click to set the cursor location in the text and start typing to add text. You may cut and paste text from another file into the commenting box

Replace (Insert): click and drag the cursor over the text then type in the replacement text. You may cut and paste text from another file into the commenting box

Strikethrough (Delete): click and drag over the text to be deleted then press the delete button on your keyboard for text to be struck through

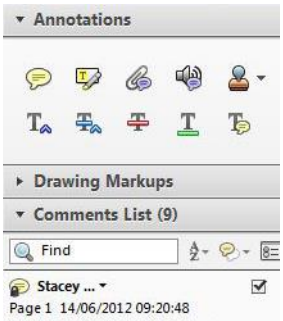
Underline: click to indicate text that requires underlining

Add sticky note: click to add a comment to the page. Useful for layout changes

Add note to text: click to add a note. Useful for layout changes

Highlight text: click to highlight specific text and make a comment. Useful for indicating font problems, bad breaks, and other textual inconsistences

Comments list This provides a list of all comments and corrections made to the article. It can be sorted by date or person.



Drawing tools There is the ability to draw shapes and lines, if required.

 **Attach file, Record audio, and Add stamp** are not in use.

J_ID: BIOINFO Customer A_ID: BTW646 Copyedited by: PR Manuscript Category: Applications Note Cadmus Art: OP

Bioinformatics, 2016, 1–2
doi: 10.1093/bioinformatics/btw646
Applications Note

OXFORD

Sequence analysis

MIToS.jl: mutual information tools for protein sequence analysis in the Julia language

Diego J. Zea[1], Diego Anfossi[1], Morten Nielsen[2,3] and Cristina Marino-Buslje[1,*]

[1]Structural Bioinformatics Unit, Fundación Instituto Leloir, Ciudad Autónoma de Buenos Aires, ~~C1405BWE,~~ Argentina, [2]Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark and [3]Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on March 28, 2016; revised on July 6, 2016; editorial decision on October 8, 2016; accepted on October 13, 2016

Abstract

Motivation: MIToS is an environment for mutual information analysis and a framework for protein multiple sequence alignments (MSAs) and protein structures (PDB) management in Julia language. It integrates sequence and structural information through SIFTS, making Pfam MSAs analysis straightforward. MIToS streamlines the implementation of any measure calculated from residue contingency tables and its optimization and testing in terms of protein contact prediction. As an example, we implemented and tested a BLOSUM62-based pseudo-count strategy in mutual information analysis.

Availability and Implementation: The software is totally implemented in Julia and supported for Linux, OS X and Windows. It's freely available on GitHub under MIT license: http://mitos.leloir.org.ar.

Contacts: diegozea@gmail.com or cmb@leloir.org.ar

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

1 Approach and implementation

The Mutual Information Tools for protein Sequence analysis (MIToS) is an open-source Julia package that allows users to study protein sequences and structures using information measures derived from multiple sequence alignments (MSA). The standard MSA format used by the tool is Stockholm (used by the Pfam database), as it allows to enrich the alignment with annotations. These MSA annotations are managed by MIToS methods. Sequence positions (i.e. residue number in UniProt) are stored as sequence annotations while the columns numbers of the original input MSA are stored as file annotations. Annotations are kept updated when mutating operations are performed on the MSA (i.e. elimination of insert columns). That allows to keep track of sequence positions through a pipeline. The stored residue positions can be easily extracted as vector of numbers. MIToS maps structure and sequence residues using SIFTS as it was shown that aligning PDB sequences (atom-res) with MSA sequences often yields incorrect alignments in those regions flanking missing residues (Velankar *et al.*, 2013).

MIToS encodes amino acids and gaps as integer numbers. This allows fast indexing of residue contingency tables used for counts and probabilities estimations. The residue contingency tables are used ~~for~~ residue ~~frequency counting either by~~ sequence or columns. Also join frequencies (residues pairs, triplets, etc.) can be calculated. These residues frequencies can used for calculating any measure (i.e. entropy, mutual information) based on them.

The PDB module of MIToS allows to read PDB or PDBML files as vectors of residues identified by its residue number (from PDB and PDBe when possible), three letters code (to allow ambiguities and not standard residues), chain and model number. MIToS provides methods to calculate the most used distances between residues (i.e. C alpha, C beta, etc.) and also 10 different types of residues contacts (van der Waals, ionic interactions, etc.). It also performs structural superimposition using the Kabasch algorithm (Kabsch, 1978) and calculates RMSD and RMSF (Root-Mean-Square Deviation and Fluctuation) measures of the superimposed structures.

The Pfam module of MIToS makes more simple and efficient the management of Pfam alignments (integrating SIFTS, PDB and MSA annotations). It was used ~~to~~ parameter optimization of the BLOSUM-based pseudo-count correction for MIp (BLMIp, note that the difference with ZBLMIp is only the Z-score calculation). The Pfam module has most of the needed functions to evaluate the predictive performance in terms of contact prediction ~~and can be used for testing other scores~~ on a Pfam dataset.

2 Case implementation

Mutual Information (MI) derived scores are covariation measures calculated between columns in an MSA, giving potential insight into residues coevolution. MI has proved to be useful for structural contacts and functional sites prediction in proteins (de Juan *et al.*, 2013; Marino Buslje *et al.*, 2010). The starting point of MIToS was the implementation of the corrected Mutual Information score (ZMIp) described in Buslje *et al.* (2009). ZMIp is based on the MIp measure of Dunn *et al.* (2008) with the following corrections: (i) redundancy reduction using a Hobohm I clustering for sequence weighting (Hobohm *et al.*, 1992), (ii) a pseudo-count correction to deal with low number of sequences and (iii) a Z score transformation against a null distribution obtained by calculating MIp on a set of random alignments generated by shuffling the sequences in the MSA. These corrections involve a higher computational cost in comparison with a raw MIp, making important the performance of the language used for its implementation. Here, the Julia language was chosen since it is a high level programming language for scientific computing, easy to use and modify, designed for parallelism with a performance close to C in terms of computing time.

A major problem for a reliable MI calculation is the number of sequences in the MSA (Buslje *et al.*, 2009; Dunn *et al.*, 2008). In a previous study, we found that the set of applied corrections, including a fix uniform pseudo-count correction, improved the performance of MI as a predictor of residue contacts, but also found that the performance decreased for MSA having less than 400 clusters of sequences sharing less than 62% sequence identity (Buslje *et al.*, 2009). We hypothesized that this could be improved by implementing a more biologically relevant pseudo-count correction. In this work, we used the MIToS framework to implement and test a BLOSUM-based pseudo-count correction, that considers the nature and the observed frequency of the amino acids, inspired by Altschul *et al.* (1997) (see supplementary data for a detailed description of the implementation and parameters optimization). MIToS' Information and Pfam modules have all the required tools to make a streamline implementation and testing. We compared this approach to a uniform pseudo-count correction (see Fig. 1). Performance of the methods ~~was~~ evaluated as the area under the receiver operating characteristic curve (AUC) for contact prediction. Residue contacts ~~relative to the used~~ MSA columns and AUC calculation were determined with few functions from the Pfam module.

For families in the testing dataset with less than 400 clusters, the mean AUC improvement was found to be significant ($\Delta AUC = 0.013$, $P < 0.01$, Wilcoxon signed rank test). For MSA with more than 400 clusters, however, the two approaches have comparable performance ($P = 0.82$, Wilcoxon signed rank test) (See Fig. 1). Given this, we recommend to use ZBLMIp only for MSAs with less than 400 clusters, because it is computationally more expensive than ZMIp and it only achieves superior performance in that interval.

In conclusion, we demonstrate the usefulness of MIToS implementing, optimizing and testing a Z score for corrected MIp using BLOSUM62 pseudo frequencies (ZBLMIp) in a comprehensive dataset.

Fig. 1. Mean (solid line) and standard error (shadowed area) of the AUC given the number of cluster at 62% identity. 25% of the families in the testing dataset have less than 66 sequences (first gray vertical line) while 50% have less than 266 sequences (second vertical line). (**A**) Using a sliding window of length 200 clusters, with a step size of 10 cluster. (**B**) Using a sliding window of length 50 clusters, with a step size of 10 cluster until 500 clusters

3 Simple usage

MIToS has a collection of scripts for running common operations from the command line without coding in Julia. Most scripts accept a file or a list of files as input, and the output is written on the same directory of the input with a suffix in the file name before the extension. When using a list of files, the parallel-computing capabilities of Julia can be used for running each file on a different process.

Acknowledgements

We would like to thank Elin Teppa, Javier Iserte, Franco L. Simonetti, Patricio Barletta and Diego Vadell for their invaluable contributions, Jorge Fernández de Cossío Díaz for the Kabsch algorithm implementation and Thomas Breloff for his help with plotting methods.

Funding

C.M.B, D.J.Z and MN are researchers at the Argentinean National Research Council (CONICET). The work was partially supported by Leloir Institute foundation and PICT 2014-1787.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Buslje,C.M. *et al.* (2009) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, **25**, 1125–1131.

de Juan,D. *et al.* (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.

Dunn,S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Hobohm,U. *et al.* (1992) Selection of representative protein data sets. *Protein Sci. Publ. Protein Soc.*, **1**, 409–417.

Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A*, **34**, 827–828.

Marino Buslje,C. *et al.* (2010) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. *PLoS Comput. Biol.*, **6**, e1000978.

Velankar,S. *et al.* (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.