Bioinformatics, 34(15), 2018, 2676–2678
doi: 10.1093/bioinformatics/bty158
Advance Access Publication Date: 15 March 2018
Applications Note

Structural bioinformatics

CMV: visualization for RNA and protein family models and their comparisons

Florian Eggenhofer[1,2,*], Ivo L. Hofacker[2,3], Rolf Backofen[1,4] and Christian Höner zu Siederdissen[2,5,6,*]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, 79110 Freiburg, Germany, [2]Institute for Theoretical Chemistry, [3]Bioinformatics and Computational Biology Research Group, University of Vienna, A-1090 Vienna, Austria, [4]Centre for Biological Signalling Studies (BIOSS), University of Freiburg, 79110 Freiburg, Germany, [5]Bioinformatics Group, Department of Computer Science and [6]Interdisciplinary Center for Bioinformatics, University of Leipzig, D-04107 Leipzig, Germany

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on June 21, 2017; revised on February 26, 2018; editorial decision on March 12, 2018; accepted on March 13, 2018

Abstract

Summary: A standard method for the identification of novel *RNAs* or *proteins* is homology search via probabilistic models. One approach relies on the definition of families, which can be encoded as covariance models (*CM*s) or *Hidden Markov Models* (*HMM*s). While being powerful tools, their complexity makes it tedious to investigate them in their (default) tabulated form. This specifically applies to the interpretation of comparisons between multiple models as in family clans. The *Covariance model visualization tools (CMV)* visualize *CM*s or *HMM*s to: I) Obtain an easily interpretable representation of *HMM*s and *CM*s; II) Put them in context with the structural sequence alignments they have been created from; III) Investigate results of model comparisons and highlight regions of interest.

Availability and implementation: Source code (http://www.github.com/eggzilla/cmv), web-service (http://rna.informatik.uni-freiburg.de/CMVS).

Contact: egg@informatik.uni-freiburg.de or choener@bioinf.uni-leipzig.de

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

1 Introduction

Probabilistic models are constructed for specific *RNA* and protein families sharing a common ancestor and a biological function. The most prominent instances are the *HMM* architecture as used by *HMMER3* (Eddy, 2011) and the *CM*s utilized by *INFERNAL* (Nawrocki and Eddy, 2013). Currently there are 2686 *RNA* families available from the *Rfam* (Burge *et al.*, 2012; Kalvari *et al.*, 2017; Nawrocki *et al.*, 2015) database and 16 712 from *Pfam* (Finn *et al.*, 2016). Visualization of the models provides an overview over whole regions and allows to directly inspect states, nodes and probabilities. A *HMM* visualization tool exists as part of SAM (Krogh *et al.*, 1994), while for *CM*s, as far as we are aware, no automatic solution exists.

2 Approach

Each tool of *CMV* accepts one or more models (*INFERNAL*, *HMMER3* format) and optionally one or more corresponding alignments (*Stockholm* format) as input. The tools for comparison visualization require inputs in *CMCompare* (Eggenhofer *et al.*, 2013; Höner zu Siederdissen and Hofacker, 2010) format. Additional parameters can be set that control the level of detail of the visualization. In the minimal setting only the index for each node is shown, while full details provide states and probabilities. Moreover it is possible to select if emission probabilities should be displayed as numerical values or using a graphical representation. The number of entries in the alignment, the image size and the output format (svg, png, eps, pdf) can also be defined via options.

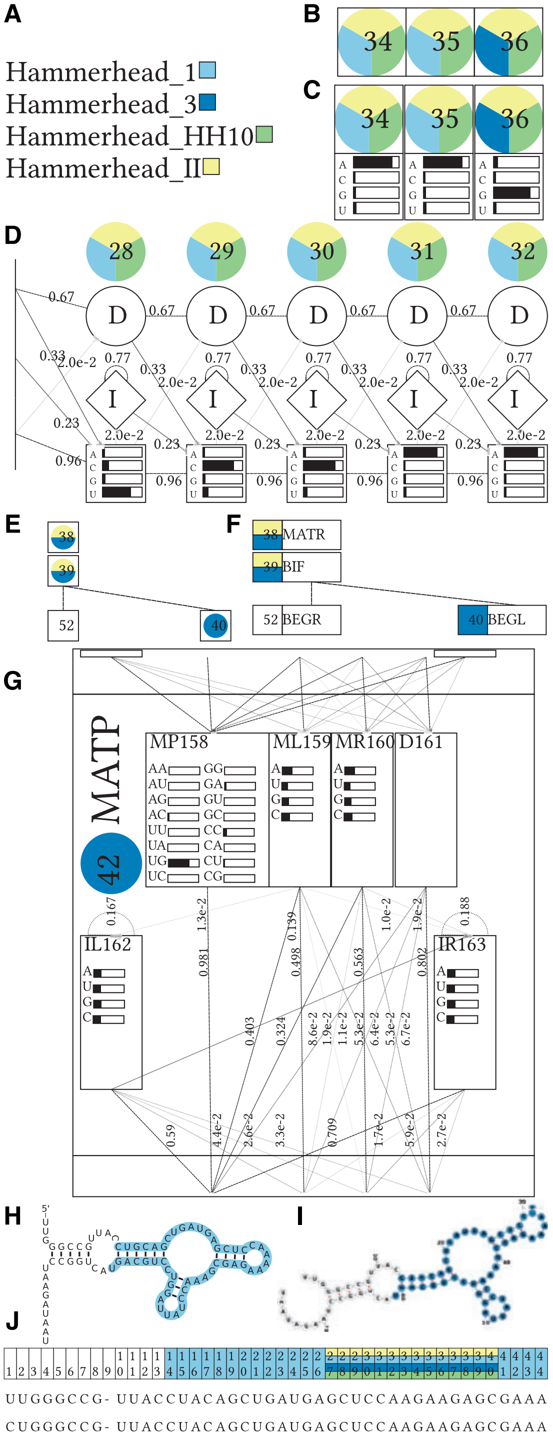

Fig. 1. Visualization of *HMM* (**B**, **C**, **D**) and *CM* (**E**, **F**, **G**) consensus secondary structure (**H**, **I**) and Stockholm Alignment (**J**) for the *Hammerhead RNA_HH9* in comparison with families from the Hammerhead *RNA* family clan (**A**). Color labels indicate to which other model an alignment column or node has been linked via *CMCompare* (Complete figures are shown in Supplementary Material). A: Color Legend for the compared models; B: minimal *HMM* details show nodes with indices, C: simple *HMM* details show emission probabilities as well; D: detailed *HMM* view shows states with emission and transition probabilities; E: minimal *CM* details show nodes with indices; F: simple *CM* details add node type information; G: detailed *CM* view shows nodes with states and emission and transition probabilities; H and I show secondary structure visualization via *R2R* and *forna*; J shows a slice of input alignment, each line corresponds to one family member. Numbers on top of the columns represent the column index stored in the corresponding *CM* node

The tools have been written using the *diagrams* library with a *cairo* back-end for visualization. Processing takes on average, for the first 100 *Rfam* models, 13 s for a model with detailed output (see Supplementary Table 1).

The tools create one visualization output file per input model. If the Stockholm alignment for the family was provided, then a second output file is generated per alignment.

It is possible to select from three levels of visualization detail (minimal, simple, detailed) for family models and, exclusively for *CM*s, linear or tree layout. The minimal detail setting shows each node (roughly corresponding to paired nucleotides or single aminoacids or nucleotides) of the model as a box labeled with the index of the node. When the detail level is set to simple, emission probabilities are included in the visualization for each node in case of *HMM*s and the node type in case of *CM*s. The detailed level shows the individual states (encoding match, insertion and deletion options) per node, with emission and transition probabilities (see Fig. 1B–G). Emission probabilities are either shown as numerical values (score, probability) or as graphical bars. Transition probabilities are visualized as arrows between states, with probabilities indicated by increasing opacity, as well as text labels. For more information and figures see the Supplementary Material.

Results of model comparison are visualized by labeling nodes with colors encoding the linked models (see Fig. 1A). Since the alignment columns corresponding to a node are known via the column index, the comparison information is also annotated in the alignment visualization (see Fig. 1J).

In the case of (structured) *RNA*s this comparative information can be mapped back to the consensus secondary structure of the family, thus enabling the identification of specific motifs or regions that are linked. This is done via labeling a secondary structure visualization of *R2R* (Weinberg and Breaker, 2011) or alternatively an input file for *forna* (Kerpedjiev *et al.*, 2015) (see Fig. 1H and I).

The tool also is available as a web-service, along with documentation and precomputed examples in three detail levels for all available models in the *Rfam* database and the first 1500 models of the *Pfam* database.

3 Conclusion

We provide an open-source tool and web-service for the visualization of *HMM*s, *CM*s, their alignments and, for *RNA*, their consensus secondary structure. The visualizations can supplement models in the *Pfam* and *Rfam* databases and enable convenient inspection of newly constructed models with *RNAlien* (Eggenhofer *et al.*, 2016), *RNAscClust* (Miladi *et al.*, 2017), or the *RNA* workbench (Backofen *et al.*, 2017; Grüning *et al.*, 2017). Nodes linked by comparison to other models are highlighted in the visualization, which allows to investigate sequence and structure elements shared among family clans. This simplifies the identification of domains, respectively secondary structure elements, with potentially related biological functionality.

Funding

This project was funded, in part, by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF), project *Doktoratskolleg RNA Biology W1207-B09*, project *SFB F43 RNA regulation of the transcriptome*, Deutsche Forschungs Gesellschaft (DFG) grant *BA 2168/3-3* and DFG BA 2168/16-1. The open access fee was covered by FWF F 4305-B09. We thank the anonymous reviewer for constructive comments, which helped us to improve the tools and the manuscript.

Conflict of Interest: none declared.

References

Backofen,R. *et al*. (2017) RNA-bioinformatics: tools, services and databases for the analysis of RNA-based regulation. *J. Biotechnol*., **261**, 76–84.

Burge,S.W. *et al*. (2012) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res*., **41**, D226–D232.

Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol*., **7**, e1002195.

Eggenhofer,F. *et al*. (2013) CMCompare webserver: comparing RNA families via covariance models. *Nucleic Acids Res*., **41**, W499.

Eggenhofer,F. *et al*. (2016) RNAlien - unsupervised RNA family model construction. *Nucleic Acids Res*., **44**, 8433.

Finn,R.D. *et al*. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*., **44**, D279–D285.

Grüning,B.A. *et al*. (2017) The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in galaxy. *Nucleic Acids Res*., **45**, W560–W566.

Kalvari,I. *et al*. (2017) Rfam 13.0: shifting to a genome-centric resource for non-coding rna families. *Nucleic Acids Res*., **46**, D335–D342.

Kerpedjiev,P. *et al*. (2015) Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams. *Bioinformatics*., **31**, 3377–3379.

Krogh,A. *et al*. (1994) Hidden markov models in computational biology: applications to protein modeling. *J. Mol. Biol*., **235**, 1501–1531.

Miladi,M. *et al*. (2017) Rnascclust: clustering rna sequences using structure conservation and graph based motifs. *Bioinformatics*, **33**, 2089–2096.

Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Nawrocki,E.P. *et al*. (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*., **43**, D130–D137.

Siederdissen,C. *et al*. (2010) Discriminatory power of RNA family models. *Bioinformatics*, **26**, i453–i459.

Weinberg,Z. and Breaker,R.R. (2011) R2R - software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, **12**, 3.