Published online 2 May 2013 *Nucleic Acids Research, 2012, Vol. 41, Web Server issue* ***W499–W503***
doi:10.1093/nar/gkt329

CMCompare webserver: comparing RNA families via covariance models

Florian Eggenhofer[1],*, Ivo L. Hofacker[1,2] and Christian Höner zu Siederdissen[1],*

[1]Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria and [2]Bioinformatics and Computational Biology Research Group, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria

Received January 31, 2013; Revised April 2, 2013; Accepted April 9, 2013

ABSTRACT

A standard method for the identification of novel non-coding RNAs is homology search by covariance models. Covariance models are constructed for specific RNA families with common sequence and structure (e.g. transfer RNAs). Currently, there are models for 2208 families available from Rfam. Before being included into a database, a proposed family should be tested for specificity (finding only true homolog sequences), sensitivity (finding remote homologs) and uniqueness. The CMCompare webserver (CMCws) compares Infernal RNA family models to (i) identify models with poor specificity and (ii) explore the relationship between models. The CMCws provides options to compare new models against all existing models in the current Rfam database to avoid the construction of duplicate models for the same non-coding RNA family. In addition, the user can explore the relationship between two or more models, including whole sets of user-created family models. Visualization of family relationships provides help in evaluating candidates for clusters of biologically related families, called clans. The CMCws is freely available, without any login requirements, at http://rna.tbi.univie.ac.at/cmcws, and the underlying software is available under the GPL-3 license.

INTRODUCTION

In the past years, and especially with the development of high-throughput methods like RNA sequencing, the scientific community became more and more aware of the importance of non-coding RNAs. These transcripts are found in all domains of life and regulate essential pathways and cellular processes.

Homologs of known RNA sequences can be detected in genomes using a number of methods. For close homologs, sequence-based methods like Blast (1) provide an extremely efficient search method. More remote homologs accumulate mutations on the sequence level, whereas the structure tends to be conserved. In structural non-coding RNAs, most of the statistical information appears to be available with the sequence and secondary structure. Methods like Infernal (2,3) can be used to transform the structural alignment of an RNA family of related sequences into a stochastic model called a covariance model.

RNA family models allow one to find new homolog family members by considering the structure and sequence features of this family. The number of covariance models, which is available from databases like Rfam (4,5), is constantly increasing.

Putative homologs discovered in a genome should, in principle, show strong affinity to only a single RNA family or, by extension, covariance model. In practice, some RNA families [e.g. RNaseP, rRNA (SSU)] have been intentionally split along kingdoms to preserve statistical signals owing to diverse sequence mutations and structural changes.

The CMCompare webserver (CMCws) provides an easy-to-use interface to check the discriminatory power of newly proposed RNA family models. This makes it possible to check that a similar model does not already exist in the database or that a set of existing or newly proposed models is not too closely related to each other in terms of the sequences they accept as putative homologs.

DESCRIPTION OF THE WEBSERVER

Functionality of CMCws

For newly constructed covariance models, it is useful to check what other models are already available in Rfam and compare them with each other. The CMCws is

*To whom correspondence should be addressed. Tel: +43 1 4277 52731; Fax: +43 1 4277 52793; Email: egg@tbi.univie.ac.at
Correspondence may also be addressed to Christian Höner zu Siederdissen. Tel: +43 1 4277 52737; Fax: +43 1 4277 52793; Email: choener@tbi.univie.ac.at

based on 'CMCompare' (6), which returns a Link score for every pair of models checked. Link sequences and their associated Link scores are sequences giving high scores in both models simultaneously. A sequence with a Link score of, say 20 bits, scores at least 20 bits in each of the models. The Link sequence is the sequence with highest overall Link score (6). A high Link score can be an indicator for the following:

(1) A model for the same RNA family is already present in the database. Using a curated model from Rfam avoids repetitive model construction and fine tuning. Also, improvements and extensions can be easier shared by finding and using a common set of models. Detection of a similar model by CMCws allows one to use this model instead.
(2) At least one of the models lacks specificity, meaning that both score high for the same sequence. A model should detect only homologs belonging to the RNA family it represents, but not of member sequences of other families. During model construction, more members belonging to the RNA family are added to ensure detection over bigger phylogenetic distance, which can expand the space of detected sequences and associated structures to overlap with other families. By highlighting these overlaps, CMCws makes it possible to address this lack of specificity.
(3) A biological relationship exists between the models that explain the overlap. Families derived from a common ancestor can share sequence and structure features. Rfam groups families related in this way as clans (7), which has been done up to now in a manual process. CMCws would allow Rfam to find possible candidates for clan members.

Input

After choosing the mode of comparison, the web server accepts a file upload containing one or more Infernal covariance models (Infernal 1.0 or later, Rfam 9 or later) or structural alignments using the Stockholm format as input. Stockholm alignments are internally converted to covariance models for further processing.

Processing

The web server relies on CMCompare (6), which is the first published tool for comparison of covariance models and has already been used in other projects (8,9). CMCompare has been expanded to also compare models created with Infernal 1.1 since publication.

Two modes of processing are available. The first mode allows one to compare the input models against all available models in Rfam or all models of specified subtype (micro RNAs, tRNAs) thereof, which reduces computation time. Alternatively, the set of uploaded models can be compared against each other.

Output

The first mode provides the user with a table of pairwise comparisons against Rfam models, as shown in Figure 1.

The result list, computed by CMCompare, can be filtered by model name, Link score and number of models. Each of the columns can be sorted. These filtering options allow one to easily extract similar models. A weighted graph representation visualizes selected models as nodes, and their Link scores as edges to simplify evaluation, see Figure 2b. By clicking on the edges or the magnifying glass icon, each pairwise comparison can also be viewed in detail, providing the common highest scoring sequence (Link sequence), corresponding structure and further information.

Models of interest from the result list, or a set of models that have been uploaded by the user, can be analyzed with the second mode. This mode returns all pairwise comparisons, which can also be sorted and filtered by the name of a second model. Exploring this list is especially useful to identify groups of models that are closely related and pose potential candidates for clans.

The output is visualized as a graph, as well as a matrix, which gives all pairwise Link scores, simplifying the identification of relevant links, see Figure 2. Comparing models cannot always capture the biological relationship between models, e.g. in the RNase P clan. Although the two different models for bacterial RNaseP are linked with each other, one of them is strongly linked with the corresponding RNaseP model for archae, and the other one is not. By using a graph representation, we are still indirectly able to identify potential clan members.

As noted before, Rfam clans are constructed entirely manually. We believe that CMCws can significantly facilitate this process.

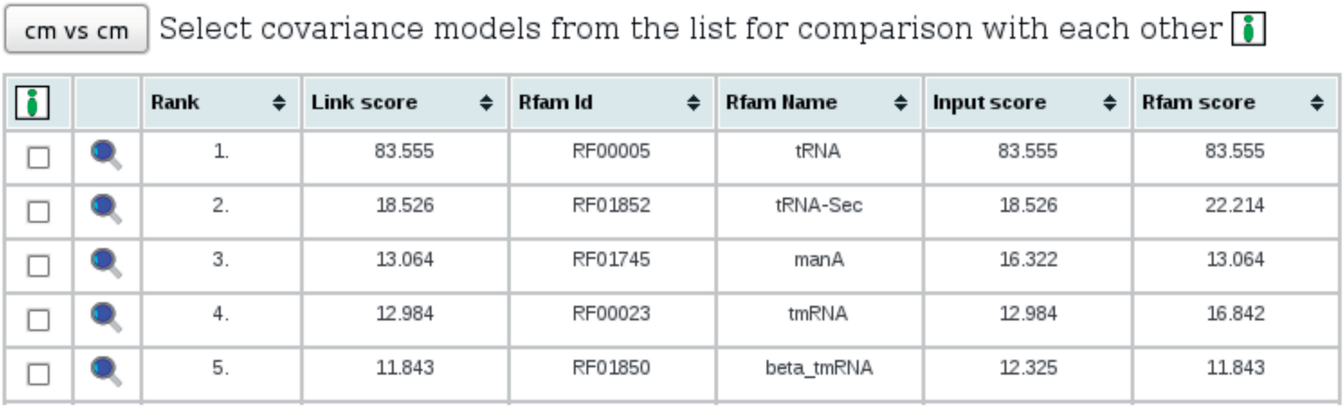
Usage example

Assume we are interested in RNA families related with tRNAs. For this usage example, which follows Figures 1 and 2, we use as input the tRNA model (RF00005) from Rfam.

The first step is to select the comparison versus Rfam mode and upload the model to check for similar models already available from the database.

The top five resulting hits are shown in Figure 1, starting with tRNA having the maximal score possible with this model, compared with itself. The next models have Link scores between 10 and 20, indicating a moderate overlap between them.

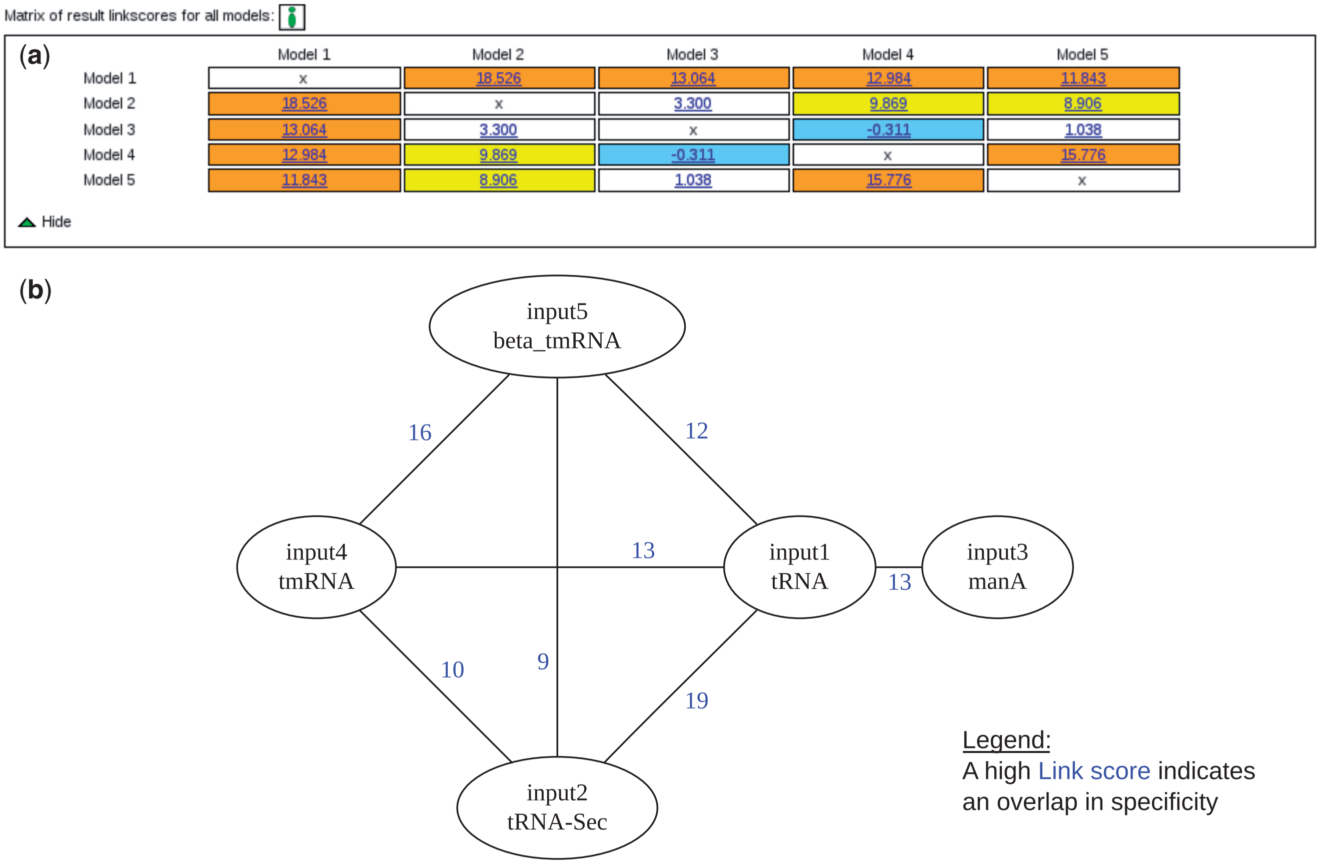
For each of these models, one should investigate the reason for the high Link score, with potential reasons given previously as points 1–3. In decreasing order of Link score, we first consider the tRNA-Sec RNA family. Careful comparison of both secondary structures yields notable differences, including an additional stem in tRNA-Sec, but also some commonality. Based on commonalities and differences in biological action in the cell, as well as the differences in the structural alignment, one will probably not want to join both tRNA and tRNA-Sec, as a single family, but the commonalities are large enough to suggest a common Rfam clan, which for



Select covariance models from the list for comparison with each other

		Rank	Link score	Rfam Id	Rfam Name	Input score	Rfam score
□		1.	83.555	RF00005	tRNA	83.555	83.555
□		2.	18.526	RF01852	tRNA-Sec	18.526	22.214
□		3.	13.064	RF01745	manA	16.322	13.064
□		4.	12.984	RF00023	tmRNA	12.984	16.842
□		5.	11.843	RF01850	beta_tmRNA	12.325	11.843

Figure 1. List of results: contains comparison results corresponding to the current filtering options. The list is sortable by all column names. The magnifying glass links to a detailed view of each comparison. The checkboxes on the right allow to select the models for a comparison with each other. CMCompare computes a score for the Input model (Input score) and for the Rfam model (Rfam score). The lower one is the Link score.


Matrix of result linkscores for all models:

(a)

	Model 1	Model 2	Model 3	Model 4	Model 5
Model 1	x	18.526	13.064	12.984	11.843
Model 2	18.526	x	3.300	9.869	8.906
Model 3	13.064	3.300	x	-0.311	1.038
Model 4	12.984	9.869	-0.311	x	15.776
Model 5	11.843	8.906	1.038	15.776	x

Figure 2. Visualizations: simplifying identification of relevant similarities between models by giving different representations of the pairwise result Link scores. (**a**) Link score matrix containing the similarity between all provided models and highlighting them by color. Clicking the Link score shows a detailed view of the comparison. (**b**) Weighted graph representation of linked models. The nodes indicate the models and contain their identifier. In contrast to the matrix representation, the shown edges correspond with the applied filtering options and redirect to a detailed view of the comparison on clicking. The comparisons against Rfam only show edges between the input and Rfam models. The shown input models 1, 2 and 4, 5 are members of the tRNA-clan, whereas ManA is presumably a false link.

Rfam is true. Incidentally, the CMCompare algorithm proposes a consensus secondary structure of both RNA families for the link sequence, which contains a total of three stems, with one tRNA and two tRNA-Sec stems deleted in the consensus.

The next two models in the list tmRNA and beta_tmRNA have a significantly lower Link score than the tRNA compared with itself but capture the similarity between the models. As an aside, both tmRNAs have a higher Link score between each other than to the tRNA model.

The final model flagged by the CMCws is the manA RNA motif family. The Link score is low (13 bits) so that no immediate action is warranted.

However, the nature of the manA and its secondary structure (the CMCompare algorithm proposes a low-scored cloverleaf consensus structure between the tRNA and manA families) makes it a candidate for further investigation. According to Rfam, this is a computationally identified RNA family that occurs often adjacent to tRNAs (10).

Among the first five hits of the list, we can find three of the five other members of the tRNA clan. To get a better idea about their relationship with each other, we can select and resubmit them to a cm versus cm comparison. Figure 2b shows the result for the submission of the top five models. The matrix representation gives an overview over all comparisons between the submitted models, whereas the weighted graph only shows RNA family models as nodes and linkscores as edges. As expected, we can see that there is a strong connection between the members of this clan and especially between the tmRNA models. The manA is only linked with the tRNA model, but not with the other clan members. The combination of these two comparison modes simplifies finding candidates for clan construction.

Following these conclusions, the tRNA family would be submitted for inclusion in the Rfam database, pointing out it is possible biological relationship with the tRNA-Sec family.

Implementation details

CMCws was implemented in Perl 5 using CGI.pm and the template toolkit. It relies on the jQuery library to allow sortable result tables. The underlying CMCompare algorithm (6) is implemented in Haskell (11). The conversion of input Stockholm-format alignments is done with cmbuild from the Infernal package (3).

The weighted graph representations of the output are created with dot from the graphviz (12) toolset.

The current version of the CMCompare algorithm has a quadratic runtime. With n and m the number of states (roughly the number of columns) in each covariance model, and c a fairly large constant, the runtime is $O(cnm)$. Wall-clock runtimes are from <1 s for small models to ≈30 s for comparisons between members of the RNaseP clan. We plan to improve on these runtimes in the near future to facilitate large-scale comparisons.

Other tools

To our knowledge, there are no algorithms available other than CMCompare that compare RNA family models with each other. Other classes of biopolymers like DNA or Proteins families can be modeled by profile hidden Markov models (HMMs) (13). General work has been done on comparing HMMs (14) with other HMMs. Also comparisons of HMMs with stochastic context free grammars (15), which provide the underlying principles of covariance models, have been investigated, but in both cases, no available tools originated from this work.

DISCUSSION

CMCws simplifies dealing with an increasing number of RNA family models. Covariance models designed for essentially the same structural RNA family can be detected, as can those that capture a sub- or super-set of the structural features. Covariance models with inferior discriminatory power are easily detected by a large number of high Link scores to other RNA family models. Potential clans can be discovered by looking for a small set of CMs with higher Link scores to each other but low Link scores to all other families.

Challenges remain in identifying the cause of non-specificity among covariance models and how to defuse it. Suggestions how to split RNA families into more specific subfamilies and use of meta-families to pool them again could be a first step into this direction. Also, the construction of clans in an entirely unsupervised manner is a goal for the future.

Promising avenues for expanding functionality of CMCompare are other stochastic grammars such as HMMs used in Pfam (16).

This would allow expanding CMCws in the future to provide a comprehensive web server for comparing and analyzing different kinds of databases of stochastic sequence families.

FUNDING

Austrian FWF, project ‘Doktoratskolleg RNA Biology W1207-B09’ (to F.E.) and project ‘SFB F43 RNA regulation of the transcriptome’ (CHzS). Funding for open access charge: Austrian FWF—SFB F43 RNA regulation of the transcriptome.

Conflict of interest statement: None declared.

REFERENCES

1. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
2. Eddy,S.R. and Durbin,R. (1994) Rna sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
3. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
4. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.

5. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
6. Höner zu Siederdissen,C. and Hofacker,I.L. (2010) Discriminatory power of RNA family models. *Bioinformatics*, **26**, 453–459.
7. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2011) Rfam: wikipedia, clans and the ‘‘decimal’’ release. *Nucleic Acids Res.*, **39**, D141–D145.
8. Lange,S.J., Maticzka,D., Möhl,M., Gagnon,J.N., Brown,C.M. and Backofen,R. (2012) Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.*, **40**, 5215–5226.
9. Chen,A. and Brown,C. (2012) Distinct families of cis-acting RNA replication elements epsilon from hepatitis B viruses. *RNA Biol.*, **9**, 1–7.
10. Weinberg,Z., Wang,J.X., Bogue,J., Yang,J., Corbino,K., Moy,R.H. and Breaker,R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.*, **11**, R31.
11. The GHC Team. (1989–2013) *The Glasgow Haskell Compiler (GHC)*. http://www.haskell.org/ghc/.
12. Ellson,J., Gansner,E., Koutsofios,L., North,S.C. and Woodhull,G. (2002) Graphviz open source graph drawing tools. In: Mutzel,P., Jnger,M. and Leipert,S. (eds), *Graph Drawing,* Vol. 2265 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 483–484.
13. Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.
14. Lyngsø,R.B., Pedersen,C.N. and Nielsen,H. (1999) Metrics and similarity measures for hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 178–186.
15. Jagota,A., Lyngsø,R.B. and Pedersen,C.N.S. (2001) Comparing a hidden markov model and a stochastic context-free grammar. In: Gascuel,O. and Moret,B.M.E. (eds), *Algorithms in Bioinformatics,* Vol. 2149 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 69–84.
16. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.