BMC Bioinformatics

Methodology article

ConStruct: Improved construction of RNA consensus structures

Andreas Wilm, Kornelia Linnenbrink and Gerhard Steger*

Address: Heinrich-Heine-Universität Düsseldorf, Institut für Physikalische Biologie, Universitätsstr. 1, D-40225 Düsseldorf, Germany

Email: Andreas Wilm - wilm@biophys.uni-duesseldorf.de; Kornelia Linnenbrink - linnenbr@biophys.uni-duesseldorf.de; Gerhard Steger* - steger@biophys.uni-duesseldorf.de

* Corresponding author

Published: 28 April 2008

BMC Bioinformatics 2008, **9**:219 doi:10.1186/1471-2105-9-219

Received: 23 October 2007
Accepted: 28 April 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/219

Abstract

Background: Aligning homologous non-coding RNAs (ncRNAs) correctly in terms of sequence and structure is an unresolved problem, due to both mathematical complexity and imperfect scoring functions. High quality alignments, however, are a prerequisite for most consensus structure prediction approaches, homology searches, and tools for phylogeny inference. Automatically created ncRNA alignments often need manual corrections, yet this manual refinement is tedious and error-prone.

Results: We present an extended version of CONSTRUCT, a semi-automatic, graphical tool suitable for creating RNA alignments correct in terms of both consensus sequence and consensus structure. To this purpose CONSTRUCT combines sequence alignment, thermodynamic data and various measures of covariation.

One important feature is that the user is guided during the alignment correction step by a consensus dotplot, which displays all thermodynamically optimal base pairs and the corresponding covariation. Once the initial alignment is corrected, optimal and suboptimal secondary structures as well as tertiary interaction can be predicted. We demonstrate CONSTRUCT's ability to guide the user in correcting an initial alignment, and show an example for optimal secondary consensus structure prediction on very hard to align SECIS elements. Moreover we use CONSTRUCT to predict tertiary interactions from sequences of the internal ribosome entry site of CrP-like viruses. In addition we show that alignments specifically designed for benchmarking can be easily be optimized using CONSTRUCT, although they share very little sequence identity.

Conclusion: CONSTRUCT's graphical interface allows for an easy alignment correction based on and guided by predicted and known structural constraints. It combines several algorithms for prediction of secondary consensus structure and even tertiary interactions. The CONSTRUCT package can be downloaded from the URL listed in the Availability and requirements section of this article.

Background

Prediction of RNA structure as well as searches for homologues in large genomic sequence databases play a prominent role in the era of non-coding RNAs (ncRNAs). Structure prediction may provide insight into RNA function, and pattern-based database searches [1,2] may reveal new homologues, without the need for time-consuming experiments. Prerequisite for these predictions and

searches as well as for inference of phylogeny [3-5] is the existence of an alignment of RNA homologues correct in terms of both sequence and structure. Sequence alignment tools like CLUSTALW [6] often fail to align ncRNA sequences correctly, especially when sequence homology drops below 60 % [7]. One reason is that ncRNA sequences evolve by compensatory base pair changes and ncRNA homologues are more conserved in structure than in sequence. For example, structural elements like thermodynamically extrastable tetraloops (UNCG, GNRA) share no sequence similarity and therefore cannot be correctly aligned by pure sequence alignment programs. Even structure alignment programs (e. g. DYNALIGN [8], FOLDALIGN [9], PMMULTI [10] or STEMLOC [11]) do not necessarily produce high-quality alignments under all conditions [7]. Moreover, these approaches are computationally extremely demanding, not only because they are based on simplified versions of the Sankoff algorithm [12]. Thus, automatically generated alignments often need to be corrected or refined by hand, which is a complex and tedious task. To ease this task a few sophisticated RNA alignment editors exist, e. g. 4SALE [13], SARSE [14] or S2S [15]. One of these tools is CONSTRUCT (**construc**tion of RNA **con**sensus **struc**tures; [16]), which is not only an RNA alignment editor but also allows for a variety of consensus structure predictions.

Here we present the completely revised and largely extended version of this tool and demonstrate some of its new features. CONSTRUCT allows for generation of RNA alignments, which are correct in terms of sequence and structure, by combining thermodynamic RNA structure prediction, several measures for covariation, and any alignment method. By applying this combination, typical shortcomings inherent to the single methods are eliminated; that is, the need of covariation for many, sufficiently divergent sequences is reduced, and the quality of thermodynamic predictions is enhanced. In contrast to tools, which predict a consensus structure automatically from a fixed alignment (e. g. RNAALIFOLD [17] or ILM [18]), CONSTRUCT allows for an interactive modification and optimization of the alignment. The user is able to modify the alignment similar to other RNA alignment editors [19,13,14]; the consequences of any alignment modifications are, however, immediately visible in a dotplot showing the probability of all base pairs of all RNA structures of the alignment; i. e., the user is guided during the alignment correction. In addition the user can account for sequence and structure constraints during the correction process. Afterwards optimal and suboptimal consensus structures and tertiary interaction can be predicted using a variety of built-in methods and displayed in several ways.

ConStruct's approach to consensus structure prediction

In the following we will describe the basic approach of CONSTRUCT (see Fig. 1).

1. First, an initial sequence alignment needs to be created by means of an alignment program of the user's choice (e. g. by a pure sequence alignment program like MAFFT [20], by a sequence+structure alignment program like STRAL [21] or STEMLOC [11], or by a pure structure alignment program like PROFILE-DYNALIGN [22]).

2. "Thermodynamic base pairing probability matrices" of all sequences in the alignment are automatically generated by means of a front-end program (CS_FOLD) to RNAfOLD [23]. An alternative to these thermodynamic approaches is creation of dotplots [24] with a minimum length of helices and thermodynamic weighting of their base-pair composition [25].

3. Gaps from the initial alignment (step 1) are inserted into the matrices (step 2), resulting in identically sized matrices that are superimposed, thus building a consensus matrix. Ideally, homologous base pairs should now possess identical positions. The base pair matrices for each single sequence as well as the consensus matrix are displayed in a graphical user interface (GUI; see green and blue dots in upper triangle of the consensus dotplots shown in Fig. 1 and 2). The probability of a thermodynamic consensus base pair (see red dots in the consensus dotplots) is calculated such that noise from individual base pairs not representing the consensus is reduced and over-representation of sequence families is avoided (for details see [16]). For an overview of colors used in CONSTRUCT see Table S4 in Additional file 1.

4. The new version of CONSTRUCT now allows to compute either the mutual information content (MI; [26]) or the RNAALIFOLD covariation score [17] to measure the amount of covarying positions (joined nucleotide substitutions, compensatory base pair changes). The results are displayed in the GUI (lower triangle of consensus dotplots in Fig. 1 and Fig. 2), can be filtered and normalized (see below), and are later on used in conjunction with the thermodynamic base pair matrices as the basis for consensus structure prediction. The MI at two aligned nucleotide positions i and j is defined as:

$$\mathrm{MI}_{ij} = \sum_{\mathrm{X,Y}} f_{ij}(\mathrm{XY}) \log_b \frac{f_{ij}(\mathrm{XY})}{f_i(\mathrm{X}) \cdot f_j(\mathrm{Y})}$$

where f_i(X) and f_j(Y) are the frequencies of the nucleotide types X ∈ {A,U,G,C}and Y ∈ {A,U,G,C} at aligned positions i and j, and f_{ij}(XY) is the joint frequency of finding X at i and Y at j. In addition, he user may apply a normaliza-

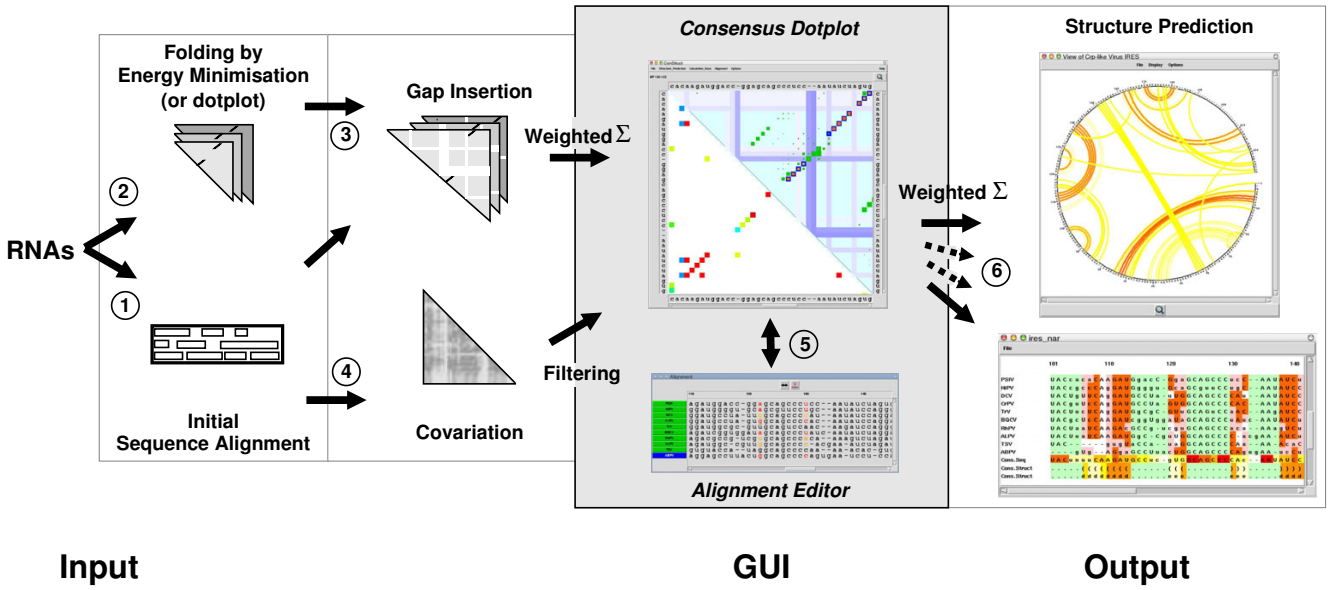

Figure 1

Flowchart and graphical user interface of ConStruct. Steps are numbered as in the text. The graphical user interface (grey part) shows results of a structural alignment for IRES regions of CrP-like viruses [45]; for a full view and further details of this alignment see Fig. 3. The main windows of CONSTRUCT are the "Consensus Dotplot" and the "Alignment Editor". The top-right triangle in the consensus dotplot shows thermodynamic base pairing probability of individual sequences (blue/green) and thermodynamic consensus matrix (red), the horizontal and vertical bars denote gaps; the lower-left triangle shows the MI (as a measure of covariance) normalized by pair entropy and a threshold of t_{CV} = 50 % applied. Predicted structures may be displayed in several representations and formats. On the right side, two possible representations are shown. The Circles plot (upper window) shows the consensus structure as predicted by maximum weighted matching (MWM); consensus pairing probability is color-coded from white to red. The crossing arcs represent pseudoknots. Below the "Structural Alignment Output" is shown. From top to bottom: ten sequences [with background colors green for loops, red for consensus base pairs, pink for consensus base pair changes (covarying pairs), and white for non-base pairs in paired regions], the consensus sequence, and the consensus structure in bracket-dot notation and character-encoded (both with background colors from white to red proportional to sequence conservation resp. pairing probability). For an overview of colors used in CONSTRUCT see Table S4 in Additional file 1.

tion method [27], which enhances separation of truly correlated positions from background correlations. That is done by dividing the MI by the joint entropy

$$h_{ij} = \sum_{X,Y} f_{ij}(XY) \log_b f_{ij}(XY),$$

the upper bound of the MI. For statistical analysis of the MI, maximum likelihood or unbiased probability estimation [28] in nits (b = e) [26] or bits (b = 2) [29] are available.

In comparison to the MI, the covariation score implemented in RNAALIFOLD measures compensations in Watson-Crick and wobble base-pairs [17] only, which is advantageous during search for helices. The meaningfulness of this score can be further improved by taking stacking into account (as shown in [30]), which is also a built-in option of CONSTRUCT. For a comprehensive description of the RNAALIFOLD covariation measure we refer to [17] and [30].

5. The alignment of the sequences is displayed in a separate window (see alignment editor in Fig. 1). Position of base pairs from the dot plots is coupled with the position of the corresponding nucleotides in the alignment; i. e., pointing with the mouse to a consensus base pair highlights the corresponding base pairs in the alignment with a color from white to red according to the individual base pairing probabilities (see also Fig. 2); pointing to a base-paired nucleotide in the alignment changes the color of the corresponding base pair in the dot plot. A selected region of a single sequence or multiple sequences, with a gap at either side, may be moved with the mouse towards the gap, and the dot plots are updated correspondingly. Helices not superimposed are easily detectable. Thus, the user is guided during the alignment correction. These functions of the GUI are extremely helpful while correct-

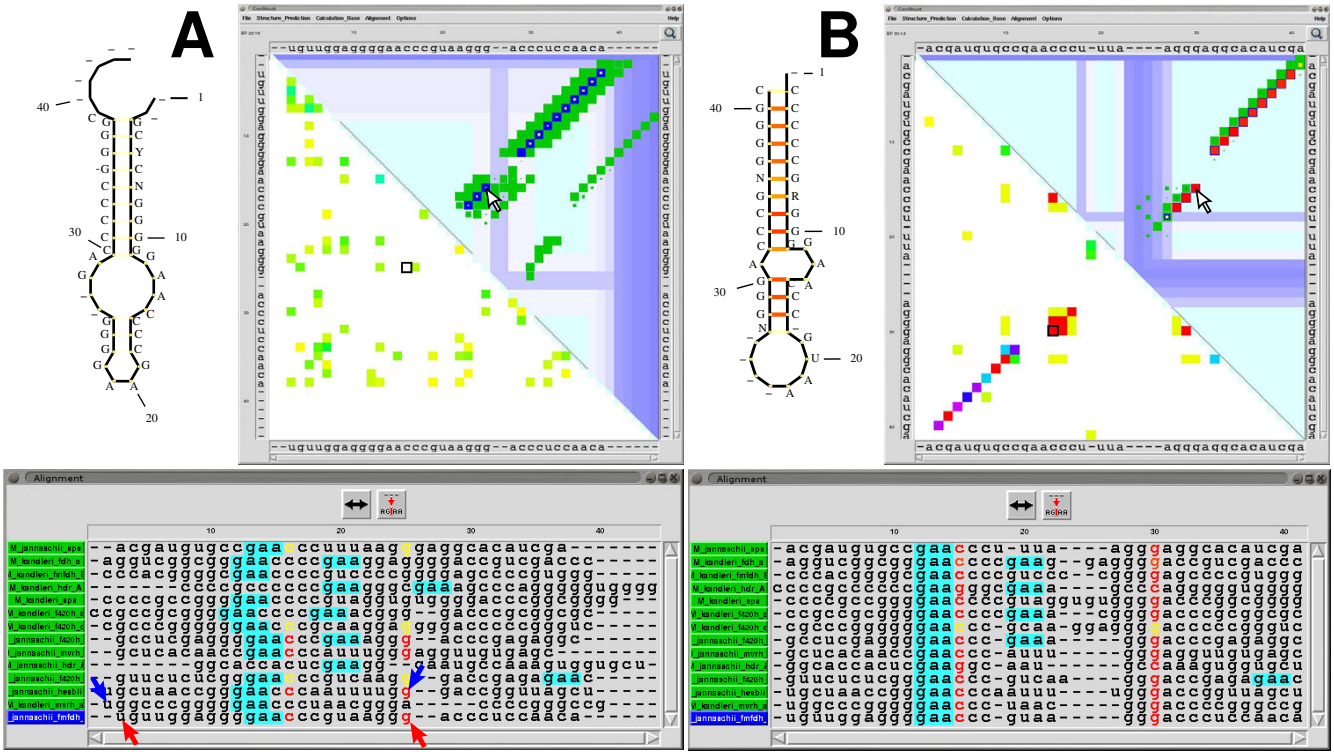

Figure 2

Visualization of alignments by ConStruct. An alignment of SECIS elements created by CLUSTALW (**A**) and after manual optimization/correction using CONSTRUCT (**B**). In both cases predicted consensus structures and CONSTRUCT's GUI are shown. For an overview of colors used in CONSTRUCT see Table S4 in Additional file 1. **Top left:** Corresponding drawings of consensus structures (annotated with the consensus sequence) generated by CONSTRUCT; consensus base pairing probability is color-coded from white to red. **Top right:** Corresponding dotplots: the base pairing probability of individual sequences (dark blue for the selected sequence M_janaschii_sps and green for others) is shown top-right in CONSTRUCT's main window; yellow to red dots show the consensus pairing probability; white to light blue bars denote gaps. The lower-left triangle shows the MI normalized by pair entropy with a threshold of t_{CV} = 30 % in rainbow-colors from yellow to red. The cursors in A and B (arrow in thermodynamics part and black square in MI part) point to a similar position. **Bottom:** Corresponding alignment windows. Nucleotides participating in a base pair to which the cursor points in the dotplot are automatically highlighted [colored by pairing probability from p = 0 (black) to p = 1 (red)]. The motif **GAA** (turquoise background), which is conserved in the internal loop, has been highlighted using the built-in regular expression search. Clicking with a mouse button to position 3 and to position 25 of the last sequence (M_jannaschii_fmfdh_B; see red cursors) in the alignment editor selects this subsequence; clicking once with left or right mouse button to the double-headed arrow moves the subsequence towards 5' or 3' end, respectively, by one position; in the top-right dotplot the corresponding base pairs are automatically positioned. Similarly, clicking to a 5' and a 3' nucleotide of two different sequences (for an example see blue cursors) selects all corresponding subsequences from the sequence range; if none of the subsequences ends in a gap and all are followed by a gap, the subsequence range is moved towards the gap by clicking to the double-headed arrow.

ing positions of structural elements, which were misaligned during the initial sequence alignment (step 1).

6. Consensus structure prediction is now based upon the weighted and filtered summation of the thermodynamic consensus dotplot (step 2) and the covariation dotplot (step 4), whereas the previous version of CONSTRUCT used the unfiltered thermodynamic consensus dotplot alone.

The probability p_c of a thermodynamic consensus base pair at positions i and j is given by

$$p_c(i,j) = \left(\frac{\sum_{s=1}^{N} w_s \cdot p_s(i,j)^{1/3}}{\sum_{s=1}^{N} w_s} \right)^3$$

where N is the total number of sequences and $0 \le w_s \le 1$ is the user-defined weight of sequence s. This weighting can be used to avoid over-representation of a closely related

sequence family in comparison to other sequences. The exponentiation helps to reduce low pairing probabilities from individual sequences.

The linear combination of the thermodynamic and the covariation pairing probabilities

$$P_c(i,j) = w_{\mathrm{TD}} \cdot \begin{cases} p_c(i,j) & \text{if } p_c(i,j) > t_{\mathrm{TD}} \\ 0 & \text{otherwise} \end{cases} + w_{\mathrm{CV}} \cdot \begin{cases} \mathrm{CV}_{i,j} & \text{if } \mathrm{CV}_{i,j} > t_{\mathrm{CV}} \\ 0 & \text{otherwise} \end{cases}$$

allows for thresholds t and a relative weighting ($w_{\mathrm{TD}} + w_{\mathrm{CV}} = 1$) of thermodynamics and covariation. The thresholds serve to further reduce the statistical noise and to suppress false positive base pairs and can be adjusted by the user. According to our experience, use of only thermodynamics with $w_{\mathrm{TD}} = 1.0$ and $t_{\mathrm{TD}} = 0.03$ already results in sensitive and specific predictions of secondary structures. The MI will only give additional information when the alignment contains many and divergent sequences. For tertiary structure predictions or detection of non-canonical base pairs [31], however, the MI must be reasonably high, since thermodynamic data alone are not sufficient.

Prediction of secondary structure is performed by dynamic programming [32] maximizing the weighted combination of the thermodynamic and covariation pairing probability. The new version of CONSTRUCT also allows to predict suboptimal structures [33,34]. More importantly, routines for predicting tertiary interactions (pseudoknots, triple pairs) by maximum weighted matching (MWM) [35] are now built into CONSTRUCT. Examples for both prediction types are given in the Results section.

The predicted consensus structures can be viewed directly or be stored in several formats (RNAVIZ [36], CONNECT [37] or RNAML [38]). Another newly added feature is the support for structure logos [39], which can be directly requested from within CONSTRUCT. Three different graphical representations are supported:

• The first representation is basically an alignment of the sequences, where the background of the nucleotides is colored according to the nucleotide's structural features (for examples see bottom of "Structure Prediction" in Figs. 1 and Fig. 3C). An additional text output describes the structural alignment in numerical form (numbers of base pairs, consensus base pair changes, mismatches, consensus base pairing probability, MI per helical position, and statistical significance of MI values by χ^2 test).

• The consensus structure–annotated by an individual sequence or the consensus sequence with or without alignment gaps–can be displayed in different, SQUIGGLES-like ways (for examples see structures in Fig. 2 left). Overlapping of helical regions may be avoided by user interaction; each helix is selectable by the mouse and may be rotated around the upstream loop. Base pair connections can be color annotated according to their probability.

• The consensus structure–annotated by an individual sequence or the consensus sequence–can be viewed as a circular graph (circles plot) with nucleotides as edges and base pairs connected by probability-annotated arcs (for examples see "Structure Prediction" in Fig. 1 and Fig. 3B). Such a plot allows for representation of tertiary interactions; i. e., crossing arcs denote pseudoknots. If chemical or enzymatic mapping data are available, the accessibility of nucleotides can be marked with small triangles.

The two time consuming steps 1 and 2 are executed only once, whereas the remaining steps are computed on the fly. Handling of less than 100 sequences of length below 500 nucleotides is done fluently on a standard desktop PC.

Results and Discussion

The tool CONSTRUCT combines thermodynamic and statistical methods to predict the consensus structure of a set of homologous RNA molecules. In this respect it is similar to, for example, RNAALIFOLD [17] or ILM [18]. Yet, these programs use fixed alignments as input for structure prediction, whereas CONSTRUCT also allows to correct potentially incorrect alignments beforehand. CONSTRUCT's graphical user interface (GUI) guides the user in optimizing/correcting the alignment with respect to a consensus structure by displaying all base pair probabilities corresponding to the alignment in a consensus dotplot.

Most functions used in CONSTRUCT to extract a consensus structure from a given alignment are well known from the literature [32,33,26,35,17,30]. Given a reasonably good structural alignment CONSTRUCT is able to extract the correct consensus structure even without user intervention. Usually almost optimal results can be obtained by using thermodynamic pairing probabilities alone. If an alignment contains 5–10 sequences with an average pairwise sequence identity below ~70 %, then sensitivity as well as specificity [40] for secondary structure predictions are above 80 % (see Fig. S5 and Table S3 in Additional file 1; see also [30]). If additional information either from MI, normalized MI or RNAALIFOLD covariation is used, the mean accuracy of secondary structure predictions is not increased (see Additional file 1). A bonus is, however, the

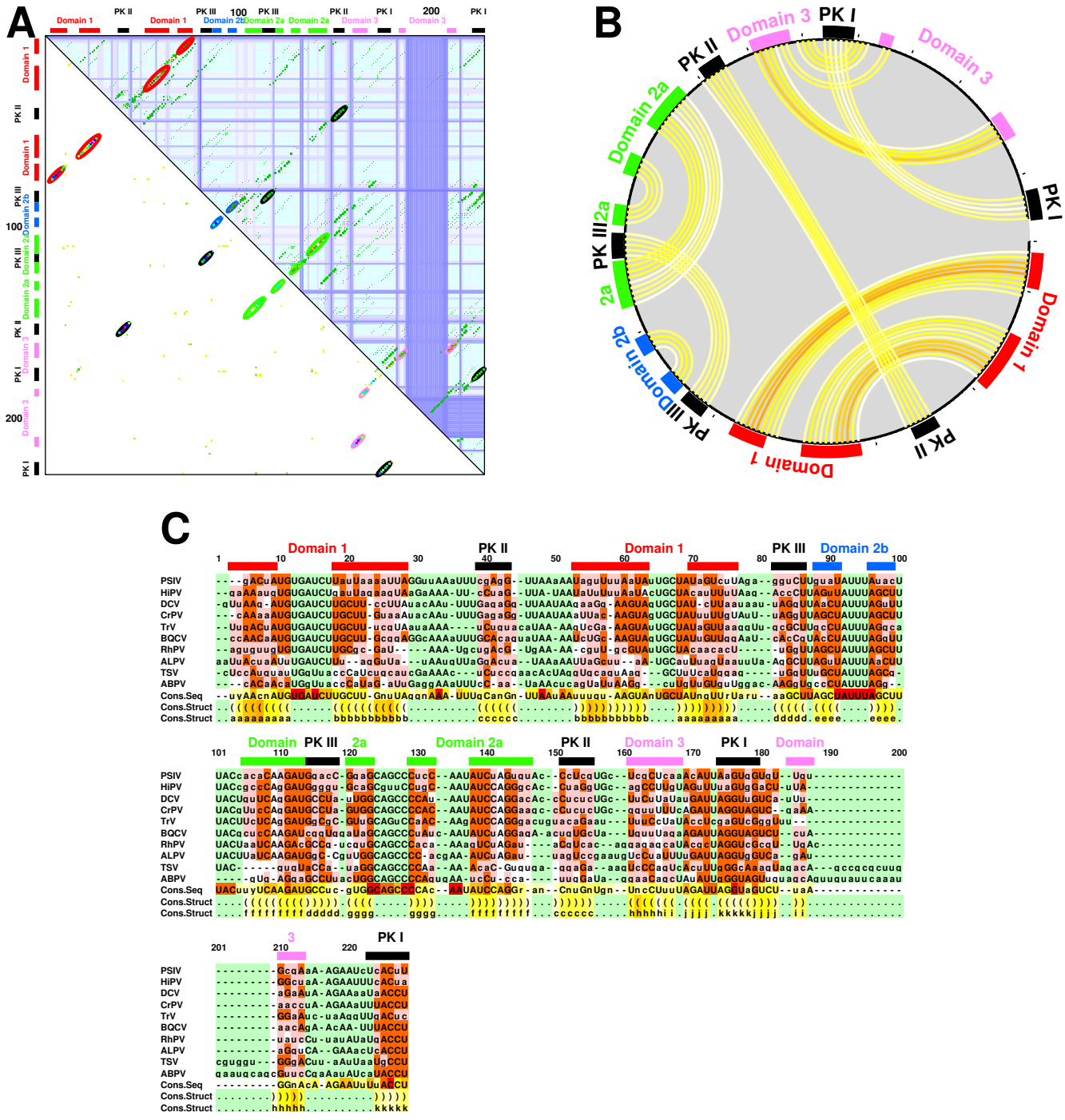

Figure 3

Alignment of internal ribosomal entry sites (IRES) from CrP-like viruses [45]. Subfigures A-C are created by means of CONSTRUCT using the RNAALIFOLD covariation score including stacking and parameters w_{TD} = 0.5, w_{CV} = 0.5, t_{TD} = 0.03, and t_{CV} = 0.15. Colored bars and labels are added in a graphics program according to the nomenclature given in [45]. The used color-coding is explained in Table S4 in Additional file 1. Sensitivity and specificity are above 90 % compared to the structures given in [47] and [46]; falsely predicted are only a few additional, non-contradictory base pairs, for example those labelled by "j" in figure part **C**. **A:** Dotplot; note that base pairs, which give rise to the pseudoknots (PK I-III), are present not only in the covariation plot (lower triangle) but also, with low probability, in the thermodynamics plot (upper triangle); i. e., they are part of suboptimal secondary structures. **B:** Circles plot. **C:** Structural alignment output.

validation of predicted pairings by the statistical information. Nonetheless, covariation scores increase prediction quality for alignments with many sequences, especially when predicting unusual base pairs, and the MI is essential when predicting tertiary interactions (see Example 2 below).

Summary of new features

Since CONSTRUCT's last release 9 years ago [16] the most striking new features are the display and use of several mutual information scores and the ability to predict tertiary interactions. The following types of measures for base pair covariation are now supported (see step 4 above): Mutual Information with optional pair-entropy normalization [26,27] or RNAALIFOLD with optional stacking [17,30], which are an essential requirement for the newly added prediction of tertiary interactions (see below). In previous versions, covariation was only used for a χ^2 test to verify predicted base pair positions. Now both types of covariation are displayed in the GUI. As covariation scores usually suffer from statistical noise and alignment errors, proper filtering and weighting is important when using it as a basis for structure prediction. CONSTRUCT now allows the user to adjust filter and weighting factors according to the displayed data and thus to make optimal use of it. Instead of using only thermodynamic data for structure prediction and validate this using statistics, we now combine the chosen covariance measure with the thermodynamic prediction and apply user-defined weights and thresholds. This filtered and weighted combination of both terms (see step 6 above) builds the basis for the following structure prediction step. On top of the standard Nussinov-style prediction of secondary structure prediction [32], we added the abilities to predict suboptimal consensus structures [33,34] using dynamic programming and tertiary interactions using the maximum weighted matching (MWM) procedures imatch and bmatch [35]; see also Figure 3. Some of these features are demonstrated in the following section.

On top of that, several new convenience features have been added, for example:

• input of alignments in most sequence formats *via* Eddy's SQUID package [41],

• color-coded regular expression searches (see e. g. the colored GAA-motif in the alignment window of Fig. 2),

• the built-in option to request RNA structure logos [39],

• removal of gaps from consensus structure drawings,

• a paned editor window to allow for simultaneous viewing and editing of 5' and 3' ends, and

• support for RNAplfold [42] to fold very long sequences locally.

CONSTRUCT version 3 is a reimplementation of the prior versions. Many time consuming Tcl functions–e. g. those responsible for update of the GUI (consensus dotplot) after alignment modification–have been ported to C and built into the custom Tcl interpreter to speed up the application. This GUI update step after alignment modification could take up to approximately 20 seconds for a big alignment of 450 sequences (moving 15 nucleotides at once); by using C-code (built into the interpreter) instead of interpreted Tcl-code the execution is speedup is roughly 15-fold.

User optimization of a given alignment

Example 1: SECIS

"Selenocysteine insertion sequence" (SECIS) RNA elements from methanogenic organisms [43] form a stem-loop structure characterized by a relatively low degree of sequence conservation in the terminal helix (about 20 nt) and a higher degree of sequence conservation in the remaining part (about 14 nt). Accordingly, alignments created by standard sequence alignment programs are far from structurally correct. However, displaying such an alignment in CONSTRUCT (see dotplot in Fig. 2A for a ClustalW alignment) the correct consensus structure is readily identifiable: note the small yellow dots inside the blue squares and the "helix clustering" visible as a close accumulation of green diagonals in the upper triangle of the dotplot, which are not superimposed, i. e. not correctly aligned. Here, five of the 14 sequences are already superimposed in their structure (note the colored nucleotides in the alignment window). Furthermore, from the dotplot it is already obvious that most other, not-superimposed structures can be aligned with those by mainly horizontal adjustment of base-pair positions. A major shift is necessary only for the two hdr_A sequences (see the off-diagonal helices in the dotplot). The user is guided during this adjustment process–i. e. which of the sequences have to be selected, which nucleotides have to be moved in the alignment, etc.–by the direct interconnection between base pairs in the dotplot and corresponding nucleotides in the alignment editor. Additionally, the possibility to highlight certain nucleotides or motifs in the alignment window by means of regular expressions might be of help during the manual refinement stage. In case of the SECIS elements this is the conserved GAA in the internal loop (see for example the turquoise colored motif in the alignment windows in Fig. 2).

After the correction process (for a step-by-step example see Fig. S6 of Additional file 1) from the sequence alignment in Fig. 2A to the corrected alignment shown in Fig. 2B, the mean thermodynamic pairing probability of the

terminal helix (see yellow to red dots in upper triangles of dotplots in Fig. 2) rises from 0.07 to 0.673, and the mean MI (see lower triangles of dotplots in Fig. 2) from 0.46 nits, which has a low significance according to χ^2 tests, to 0.86 nits, which is highly significant (all data computed inside CONSTRUCT). The alignment length is reduced by three nucleotides and all helices (green diagonals) except one are superimposed, thus building a consensus helix (red diagonal). The exceptional helix belongs to sequence M_kandleri_hdr_A, for which at least two different alignments are possible; details of these alternatives are shown in Additional file 1 (Fig. S2). A decision about such cases is left to the user and/or further (experimental) input.

Example 2: Tertiary interactions in CrP-like viruses

The second example shows prediction of a consensus structure including pseudoknots. In general the prediction of pseudoknots, triple base pairs, or any non-Watson-Crick pairs is difficult in comparison to prediction of a secondary consensus structure because thermodynamic predictions are usually limited to pseudoknot-free structures. Tertiary helices are, however, quite often part of suboptimal secondary structures included in the partition function prediction. Furthermore, base pairs in tertiary structural elements are quite often more conserved in sequence than the isosteric Watson-Crick and wobble base pairs [31] in secondary structural elements. This higher degree of sequence conservation leads to a better (sequence) alignment in corresponding regions. Anyway, pairings predicted by the partition function and/or helix dotplots (step 2) in combination with MI or RNAALIFOLD's covariation measure (step 4) followed by MWM prediction (step 6) results–similar to secondary structure prediction–in prediction with about 80 % sensitivity and specificity (for a detailed analysis see [25]).

Cricket paralysis-like viruses use an internal ribosomal entry site (IRES) for an 5'-end-independent pathway of translation initiation [44-48]. This IRES (with lengths up to 200 nt) contains three pseudo-knots, which are noticeable as crossing lines in the circle plot (Fig. 3B). The alignment contains 10 sequences with an average pairwise sequence identity of about 50 % and is slightly modified by means of CONSTRUCT from that given in [49]. The pseudoknot helices (encircled by black lines in Fig. 3A, top right triangle) are already visible in the aligned "thermodynamic base pairing probability matrices", but are much more prominent in covariation plots (Fig. 3A, left bottom triangle). Summation of pairing probabilities from thermodynamics predicted matrices and covariation plots followed by MWM structure prediction leads to a consensus structure depicted as circular graph in Fig. 3B and as structure-annotated alignment in Fig. 3C. Most predicted pairings are in accordance with those given in the literature [46,47]; sensitivity and specificity are 93 and 90 %, respectively, compared to the structure given in [47] and 92 % compared to the structure given in [46] (computed with compare.pl [25]). CONSTRUCT predicts only additional, non-contradictory base pairs; examples are two additional pairs in the proximal helix of Domain 1 and three to four additional pairs in the hairpin of Domain 3.

Application to reference alignments of BRAliBase

The first comprehensive RNA alignment benchmark (BRALIBASE II; [7]) used reference alignments created from four alignments taken from the RNA family database Rfam [50,51]. The reference alignments were compiled in such a way that each contained five sequences, which were equally distributed across the available range of sequence identity. For each of these four RNA families we took the reference alignment which is hardest to align–i. e., it has the lowest sequence identity–and optimized it using CONSTRUCT. Reference independent quality measures of the original BRALIBASE alignment and the alignment corrected with CONSTRUCT are listed in Table 1. While the sequence identities of the original and optimized/corrected alignment remain almost the same, the structural conservation is increased clearly during the optimization. The structurally misaligned regions are easily identified in CONSTRUCT's consensus dotplot display (see Table S1) and can easily be corrected in a few steps (see also SECIS example in previous section and Fig. 2).

As the BRALIBASE alignments were re-compiled automatically from bigger Rfam alignments they naturally contain some errors. We wish to note however that the BRALIBASE alignments are generally of good quality. A recent publication of another sophisticated editor, SARSE [14],

Table 1: Optimizing reference alignments of BRAliBase.

Alignment	BRALIBASE			CONSTRUCT		
	APSI	SCI		APSI	SCI	
		cov	w/o cov		cov	w/o cov
g2intron/aln51	0.46	0.55	0.42	0.45	0.75	0.58
rRNA/aln74	0.49	0.68	0.53	0.50	0.81	0.61
tRNA/aln27	0.35	1.19	0.73	0.36	1.22	0.76
U5-RNA/aln41	0.50	0.61	0.42	0.50	0.65	0.49

For each RNA family used in BRALIBASE [7] we extracted the hardest alignment, i. e. the alignment with lowest sequence identity (measured as average pairwise sequence identity; APSI), and corrected it using CONSTRUCT. Quality measures for the original BRALIBASE alignments and the alignment optimized with CONSTRUCT are shown: sequence identity is measured as APSI, structural conservation as structure conservation index (SCI) with (cov) and without (w/o cov) covariation term, respectively. By using CONSTRUCT, the structural conservation of the alignments is clearly enhanced while maintaining sequence conservation. Dotplots of the alignments are shown in Table S1 of Additional file 1.

showed that more than 10 % of all entries in the Rfam database are either misaligned (for an example see Fig. S8) or their structure is inconsistently annotated. Thus one should be aware of the limitations when using those alignments as a reference for alignment benchmarking. A benchmark consisting of hand curated alignments supported by predicted and/or known structures has yet to be compiled. RNA alignment editors like SARSE, S2S and CONSTRUCT would be crucial for this task.

Conclusion

RNA alignment and consensus structure prediction is still a circular problem: A consensus structure needs to be known to create a high-quality alignment, but a high-quality alignment is prerequisite for consensus structure prediction. Thus, despite recent advances on this field [52-55] the need for RNA alignment editors which allow manual refinement based on structural properties is still there. These editors are still widely used and become increasingly sophisticated (see e. g. [13] and [14]). Automatically created RNA alignments and corresponding consensus structure prediction can be optimized in most cases (see this manuscript and [14]).

Our tool, CONSTRUCT, guides the user in correcting structurally misaligned regions. Once the initial alignment is refined, CONSTRUCT is able to predict secondary as well as tertiary consensus structures with high sensitivity and specificity. CONSTRUCT has already been described as an effective and "most elegant" [56] tool for structure alignment generation and RNA structure prediction. One of its strength is the "elaborate GUI" [56] that allows for easy identification and correction of structurally misaligned regions, guides the user in correcting an initial RNA sequence alignment, and allows for setting proper weight and threshold parameters for consensus structure prediction. Structurally misaligned regions are readily identifiable in the thermodynamic consensus dotplot and can be corrected by means of the built-in alignment editor. The example shown in Fig. 2 is typical in the sense that with a pure sequence alignment the "correct" consensus structure is already detectable in the CONSTRUCT dotplot and necessary corrections of the initial alignment are quite obvious.

The gold standard approach for RNA consensus-structure prediction–Comparative Sequence Analysis using covariation and MI [28,57]–needs many sequences that have to be nearly perfectly aligned, which in turn is almost impossible for most sequence sets. Even given the perfect and large alignment, predictions only based on MI often suffer from non-informative columns (due to either too high sequence conservation or too many gaps) in the alignment. Purely thermodynamic based prediction methods are usually fairly reliable and allow for structure prediction from a few (in the extreme case one) sequences, and gain specificity and sensitivity when more sequences are added. Yet (standard) thermodynamic approaches alone cannot detect tertiary interaction or non-canonical base pairs. By using the (pair-entropy normalized) MI, which makes explicitly no use of base pairing rules, or the RNAALIFOLD covariation function (including stacking), which acknowledges consistent base pair mutations, CONSTRUCT is also able to predict non-canonical base pairs and tertiary interactions (for example see Fig. 3C). The prediction of tertiary interactions or at least certain types of pseudoknots could in principle be enhanced by including data into CONSTRUCT from structure-prediction programs other than RNAfold (for a review of alternatives see [58]).

From the results presented here–and our experience over the last years using CONSTRUCT–we propose the following approach for building up an RNA alignment for consensus structure prediction by means of CONSTRUCT:

1. Usually few sequences are initially available.

2. By pure sequence search (like BLAST) one could try to find more homologues of the sequence(s) from step 1. (Due to the sequence search the found homologues will be closely related to the already known sequences. For an overview and benchmark of selected RNA search tools see [59].)

3. Create an alignment of the sequences using an alignment program of your choice and depending on length and number of sequence; for example MAFFT [20], STRAL [21] or STEMLOC [11]; see [7,60] for benchmarks. This preliminary consensus structure should be checked for consistency by means of CONSTRUCT using only thermodynamics.

4. With help of the preliminary consensus structure, creation of either a pattern or a covariance model (CM) is possible. Both allow to search–for patterns with programs like PatScan [61] or HyPa [1] and for CMs with programs like infernal [62] or RSEARCH [2]–more specifically for further members of the RNA group under inspection.

Alternatively, reiterate from step 2.

5. Check the refined model for consistency with CONSTRUCT using thermodynamics and covariation analysis. If this gives new information–especially in terms of tertiary interactions and/or base triples–reiterate from step 4, otherwise this final model could be refined further by verification from wet lab experiments.

In case of additional structural knowledge, for example from chemical or enzymatic mapping [63,64], the initial structure prediction by RNAFOLD can accordingly be constrained (see RNAfold manual and [65]) and thus incorporated into CONSTRUCT. If even information on the three-dimensional structure of one of the sequences from the set is available from X-ray or NMR analysis, the use of S2S [15] in addition to CONSTRUCT is advantageous.

Methods

Sensitivity and Specificity

Given a reference and a consensus secondary structure predicted by CONSTRUCT, we use the script compare_ct.pl [40] to compute sensitivity ("hit rate") and specificity (selectivity). In case of tertiary structures we use the corresponding script compare.pl [25].

Alignment Scores

For computation of the structure conservation index (SCI) we used scif [60] and for the computation of the average pairwise identity we used alistat from Sean Eddy's SQUID package [41].

Availability and requirements

CONSTRUCT version 3 is based on a previously published version [16] and has been rewritten and largely extended. The underlying interpreter (Tcl/Tk 8.4) was extended to speed up the application; the installation process uses the GNU autotools. We successfully tested CONSTRUCT under several Linux distributions (Ubuntu, Debian, SuSE, Red Hat Fedora) as well as under Mac OS X (using Fink). Input of sequences is format independent due to use of the SEQIO package [66]. Graphics output is produced in PostScript. Various output formats for structures and further processing are supported (e. g. RNAML, connect etc.). The CONSTRUCT package (source and Debian package) including a manual can be downloaded at http://www.biophys.uni-duesseldorf.de/construct3/.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed to the described tool and approved the final manuscript.

Additional material

Additional file 1

For supplementary material see accompanying PDF file, which is also available at http://www.biophys.uni-duesseldorf.de/construct3/index.html#paper2007suppl

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-9-219-S1.pdf]

Acknowledgements

A.W. was supported by a grant from the German National Academic Foundation. Acknowledgement is made to Dr. M. Schmitz for critical reading of the manuscript. We thank J. Riks and G. Erlenkamp for earlier implementations of several procedures.

References

1. Gräf S, Strothmann D, Kurtz S, Steger G: **A computational approach to search for non-coding RNAs in large genomic data.** In *Small RNAs: Analysis and Regulatory functions of Nucleic Acids and Molecular Biology Series Volume 17*. Edited by: Nellen W, Hammann C. Springer Verlag; 2006:57-74.
2. Klein R, Eddy S: **RSEARCH: finding homologs of single structured RNA sequences.** *BMC Bioinf* 2003, **4:**44.
3. Schöniger M, von Haeseler A: **Toward assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models.** *J Mol Evol* 1999, **49:**691-698.
4. Wolf M, Achtziger M, Schultz J, Dandekar T, Müller T: **Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures.** *RNA* 2005, **11:**1616-1623.
5. Caetano-Anolles G: **Grass evolution inferred from chromosomal rearrangements and geometrical and statistical features in RNA structure.** *J Mol Evol* 2005, **60:**635-652.
6. Thompson J, Higgins D, Gibson T: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
7. Gardner PP, Wilm A, Washietl S: **A benchmark of multiple sequence alignment programs upon structural RNAs.** *Nucleic Acids Res* 2005, **33:**2433-2439.
8. Mathews D, Turner D: **Dynalign: an algorithm for finding the secondary structure common to two RNA sequences.** *J Mol Biol* 2002, **317:**191-203.
9. Havgaard J, Lyngso R, Gorodkin J: **The foldalign web server for pairwise structural RNA alignment and mutual motif search.** *Nucleic Acids Res* 2005, **33:**W650-653.
10. Hofacker I, Bernhart S, Stadler P: **Alignment of RNA base pairing probability matrices.** *Bioinformatics* 2004, **20:**2222-2227.
11. Holmes I: **Accelerated probabilistic inference of RNA structure evolution.** *BMC Bioinformatics* 2005, **6:**73.
12. Sankoff D: **Simultaneous solution of the RNA folding, alignment and protosequence problems.** *SIAM J Appl Math* 1985, **45:**810-825.
13. Seibel P, Müller T, Dandekar T, Schultz J, Wolf M: **4SALE-a tool for synchronous RNA sequence and secondary structure alignment and editing.** *BMC Bioinformatics* 2006, **7:**498.
14. Andersen ES, Lind-Thomsen A, Knudsen B, Kristensen SE, Havgaard JH, Torarinsson E, Larsen N, Zwieb C, Sestoft P, Kjems J, Gorodkin J: **Semiautomated improvement of RNA alignments.** *RNA* 2007, **13(11):**1850-1859.
15. Jossinet F, Westhof E: **Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure.** *Bioinformatics* 2005, **21:**3320-3321.
16. Lück R, Gräf S, Steger G: **CONSTRUCT: a tool for thermodynamic controlled prediction of conserved secondary structure.** *Nucleic Acids Res* 1999, **27:**4208-4217.
17. Hofacker IL, Fekete M, Stadler PF: **Secondary structure prediction for aligned RNA sequences.** *J Mol Biol* 2002, **319:**1059-1066.
18. Ruan J, Stormo GD, Zhang W: **An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots.** *Bioinformatics* 2004, **20:**58-66.
19. Griffths-Jones S: **RALEE-RNA ALignment editor in Emacs.** *Bioinformatics* 2005, **21:**257-259.
20. Katoh K, Kuma Ki, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res* 2005, **33:**511-518.
21. Dalli D, Wilm A, Mainz I, Steger G: **StrAl: Progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time.** *Bioinformatics* 2006, **22:**1593-1599.
22. Bellamy-Royds A, Turcotte M: **Can Clustal-style progressive pairwise alignment of multiple sequences be used in RNA secondary structure prediction?** *BMC Bioinformatics* 2007, **8:**190.
23. Hofacker I: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31:**3429-3431.

24. Tinoco I Jr, Uhlenbeck O, Levine M: **Estimation of secondary structure in ribonucleic acids.** *Nature* 1971, **230:**362-367.
25. Witwer C, Hofacker I, Stadler P: **Prediction of consensus RNA secondary structures including pseu-doknots.** *IEEE/ACM Trans Comput Biol Bioinform* 2004, **1:**66-77.
26. Chiu D, Kolodziejczak T: **Inferring consensus structure from nucleic acid sequences.** *Comp Appl Biosci* 1991, **7:**347-352.
27. Martin LC, Gloor GB, Dunn SD, Wahl LM: **Using information theory to search for co-evolving residues in proteins.** *Bioinformatics* 2005, **21:**4116-4124.
28. Gutell R, Power A, Hertz G, Putz E, Stormo G: **Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods.** *Nucleic Acids Res* 1992, **20:**5785-5795.
29. Schneider T, Stormo G, Gold L, Ehrenfeucht A: **Information content of binding sites on nucleotide sequences.** *J Mol Biol* 1986, **188:**415-431.
30. Lindgreen S, Gardner P, Krogh A: **Measuring covariation in RNA alignments: physical realism improves information measures.** *Bioinformatics* 2006, **22:**2988-2995.
31. Lescoute A, Leontis N, Massire C, Westhof E: **Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments.** *Nucleic Acids Res* 2005, **33:**2395-2409.
32. Nussinov R, Pieczenik G, Griggs J, Kleitman D: **Algorithms for loop matchings.** *SIAM J Appl Math* 1978, **35:**68-82.
33. Steger G, Hofmann H, Förtsch J, Gross H, Randles J, Sänger H, Riesner D: **Conformational transitions in viroids and virusoids: Comparison of results from energy minimization algorithm and from experimental data.** *J Biomol Struct Dyn* 1984, **2:**543-571.
34. Zuker M: **On finding all suboptimal foldings of an RNA molecule.** *Science* 1989, **244:**48-52.
35. Tabaska J, Cary R, Gabow H, Stormo G: **An RNA folding method capable of identifying pseudoknots and base triples.** *Bioinformatics* 1998, **14:**691-699.
36. De Rijk P, De Wachter R: **RnaViz, a program for the visualisation of RNA secondary structure.** *Nucleic Acids Res* 1997, **25:**4679-4684.
37. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31:**3406-3415.
38. Waugh A, Gendron P, Altman R, Brown JW, Case D, Gautheret D, Harvey SC, Leontis N, Westbrook J, Westhof E, Zuker M, Major F: **RNAML: a standard syntax for exchanging RNA information.** *RNA* 2002, **8:**707-717.
39. Gorodkin J, Heyer L, Brunak S, Stormo G: **Displaying the information contents of structural RNA alignments: the structure logos.** *Comp Appl Biosci/Bioinformatics* 1997, **13:**583-586.
40. Gardner P, Giegerich R: **A comprehensive comparison of comparative RNA structure prediction approaches.** *BMC Bioinformatics* 2004, **5:**140.
41. Eddy SR: **SQUID – C function library for sequence analysis.** 2005 [http://selab.janelia.org/software.html].
42. Bernhart SH, Hofacker IL, Stadler PF: **Local RNA base pairing probabilities in large sequences.** *Bioin-formatics* 2006, **22:**614-615.
43. Kryukov GV, Gladyshev VN: **The prokaryotic selenoproteome.** *EMBO Rep* 2004, **5:**538-543.
44. Hellen C, Sarnow P: **Internal ribosome entry sites in eukaryotic mRNA molecules.** *Genes Dev* 2001, **15:**1593-1612.
45. Nishiyama T, Yamamoto H, Shibuya N, Hatakeyama Y, Hachimori A, Uchiumi T, Nakashima N: **Structural elements in the internal ribosome entry site of *Plautia stali intestine virus* responsible for binding with ribosomes.** *Nucleic Acids Res* 2003, **31:**2434-2442.
46. Spahn C, Jan E, Mulder A, Grassucci R, Sarnow P, Frank J: **Cryo-EM visualization of a viral internal ribosome entry site bound to human ribosomes: the IRES functions as an RNA-based translation factor.** *Cell* 2004, **118:**465-475.
47. Schüler M, Connell S, Lescoute A, Giesebrecht J, Dabrowski M, Schroeer B, Mielke T, Penczek P, Westhof E, Spahn C: **Structure of the ribosome-bound cricket paralysis virus IRES RNA.** *Nat Struct Mol Biol* 2006, **13:**1092-1096.
48. Pfingsten J, Costantino D, Kieft J: **Structural basis for ribosome recruitment and manipulation by a viral IRES RNA.** *Science* 2006, **314:**1450-1454.
49. Kanamori Y, Nakashima N: **A tertiary structure model of the internal ribosome entry site (IRES) for methionine-independent initiation of translation.** *RNA* 2001, **7:**266-274.
50. Griffths-Jones S, Bateman A, Marshall M, Khanna A, Eddy S: **Rfam: an RNA family database.** *Nucleic Acids Res* 2003, **31:**439-441.
51. Griffths-Jones S, Moxon S, Marshall M, Khanna A, Eddy S, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Res* 2005, **33:**D121-D124.
52. Meyer I, Miklós I: **SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework.** *PLoS Comput Biol* 2007, **3:**e149.
53. Will S, Reiche K, Hofacker I, Stadler P, Backofen R: **Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering.** *PLoS Comput Biol* 2007, **3:**e65.
54. Torarinsson E, Havgaard JH, Gorodkin J: **Multiple structural alignment and clustering of RNA sequences.** *Bioinformatics* 2007, **23:**926-932.
55. Kiryu H, Tabei Y, Kin T, Asai K: **Murlet: a practical multiple alignment tool for structural RNA sequences.** *Bioinformatics* 2007, **23:**1588-1598.
56. Zuker M: **Calculating nucleic acid secondary structure.** *Curr Opin Struct Biol* 2000, **10:**303-310.
57. Pace N, Thomas B, Woese C: **Probing RNA structure, function, and history by comparative analysis.** In *The RNA World* Edited by: Gesteland R, Cech T, Atkins J. New York: Cold Spring Harbor Laboratory Press; 1999:113-141.
58. Reeder J, Höchsmann M, Rehmsmeier M, Voss B, Giegerich R: **Beyond Mfold: Recent advances in RNA bioinformatics.** *J Biotech* 2006, **124:**41-55.
59. Freyhult E, Bollback J, Gardner P: **Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA.** *Genome Res* 2007, **17:**117-125.
60. Wilm A, Mainz I, Steger G: **An enhanced RNA alignment benchmark for sequence alignment programs.** *Algorithms Mol Biol* 2006, **1:**19.
61. Dsouza M, Larsen N, Overbeek R: **Searching for patterns in genomic data.** *Trends Genet* 1997, **13:**497-498.
62. Nawrocki E, Eddy S: **Query-dependent banding (QDB) for faster RNA similarity searches.** *PLoS Comput Biol* 2007, **3:**e56.
63. Ehresmann C, Baudin F, Mougel M, Romby P, Ebel J, Ehresmann B: **Probing the structure of RNAs in solution.** *Nucleic Acids Res* 1987, **15:**9109-9128.
64. Tullius T, Greenbaum J: **Mapping nucleic acid structure by hydroxyl radical cleavage.** *Curr Opin Chem Biol* 2005, **9:**127-134.
65. Steger G: **Secondary Structure Prediction.** In *Handbook of RNA Biochemistry* Edited by: Bindereif A, Hartmann R, Schön A, Westhof E. Wiley-VCH; 2004:513-535.
66. Knight J: **SEQIO: A C package for reading and writing sequences.** 1996.