BIOINFORMATICS ORIGINAL PAPER
Vol. 22 no. 4 2006, pages 445–452
doi:10.1093/bioinformatics/btk008

Sequence analysis

CMfinder—a covariance model based RNA motif finding algorithm

Zizhen Yao[1,*], Zasha Weinberg[1] and Walter L. Ruzzo[1,2]

[1]Department of Computer Science and Engineering and [2]Department of Genome Sciences, University of Washington, Seattle WA 98195-2350, USA

Received on June 9, 2005; revised on December 12, 2005; accepted on December 13, 2005
Advance Access publication December 15, 2005
Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: The recent discoveries of large numbers of non-coding RNAs and computational advances in genome-scale RNA search create a need for tools for automatic, high quality identification and characterization of conserved RNA motifs that can be readily used for database search. Previous tools fall short of this goal.

Results: CMfinder is a new tool to predict RNA motifs in unaligned sequences. It is an expectation maximization algorithm using covariance models for motif description, featuring novel integration of multiple techniques for effective search of motif space, and a Bayesian framework that blends mutual information-based and folding energy-based approaches to predict structure in a principled way.

Extensive tests show that our method works well on datasets with either low or high sequence similarity, is robust to inclusion of lengthy extraneous flanking sequence and/or completely unrelated sequences, and is reasonably fast and scalable. In testing on 19 known ncRNA families, including some difficult cases with poor sequence conservation and large indels, our method demonstrates excellent average per-base-pair accuracy—79% compared with at most 60% for alternative methods. More importantly, the resulting probabilistic model can be directly used for homology search, allowing iterative refinement of structural models based on additional homologs. We have used this approach to obtain highly accurate covariance models of known RNA motifs based on small numbers of related sequences, which identified homologs in deeply-diverged species.

Availability: Results and web server version are available at http://bio.cs.washington.edu/yzizhen/CMfinder/

Contact: yzizhen@cs.washington.edu

Supplementary information: Supplementary technical details are available at http://bio.cs.washington.edu/yzizhen/CMfinder/

1 INTRODUCTION

Non-coding RNAs (ncRNAs) are functional RNA molecules that do not code for proteins. In the last five years, many discoveries of highly structured *cis*-regulatory elements in 5′ or 3′-untranslated regions (5′- or 3′-UTR) of mRNAs point to a variety of biological roles for ncRNAs, including localization, replication, translation, degradation and stabilization of transcripts (Conne *et al*., 2000; Mandal *et al*., 2003; Hentze and Kuhn, 1996). Notable examples are the mRNA elements known as riboswitches—parts of mRNA molecules that can directly bind a small target molecule to regulate their own activity (Mandal *et al*., 2003; Winkler and Breaker, 2003; Mandal *et al*., 2004).

*To whom correspondence should be addressed.

Important computational advances, such as development of the Rfam database (Griffiths-Jones *et al*., 2003) and fast genome-scale covariance model (CM) searches (Eddy and Durbin, 1994; Weinberg and Ruzzo, 2004a,b), aid RNA research significantly. A key problem remaining is to identify conserved secondary structure motifs among related sequences, and characterize them by models that can be used for homology search. For example, identification of such motifs in untranslated regions of orthologous bacterial genes has been critical to the discovery of new riboswitches (Barrick *et al*., 2004), but available techniques require significant manual work. Comparative sequence analysis is generally recognized as the most reliable method of RNA structure prediction, but these methods can fail when sequence conservation is too low (owing to poor alignments) or too high (owing to lack of sequence covariation). Single-sequence structure prediction is inaccurate, and simultaneous multiple sequence alignment and folding is computationally expensive. Finally, these tools generally do not interface smoothly with RNA homology search tools.

In this paper, we present a new algorithm for solving this motif discovery problem. Oversimplifying considerably, it is an expectation maximization (EM) algorithm like MEME (Bailey and Elkan, 1995), but instead of weight matrix models, it captures RNA secondary structures with covariance models (Eddy and Durbin, 1994). Because of the greatly increased complexity of the problem, we applied several techniques to solve scalability and convergence issues. These include use of careful heuristics for choosing a set of candidate structure elements to initialize the EM iteration. To improve the accuracy of consensus secondary structure in the M-step, we have combined mutual information with data-dependent, position-specific priors for base pairing based on a thermodynamic model. The key merits of our solution include:

- Applicable to unaligned input with unrelated sequences, long flanking regions and/or low sequence similarity;
- Reasonably fast and scalable with respect to the number and length of input sequences;
- Producing a motif structural alignment and statistical model that can be directly used for homology search.

The third point is particularly important, because we view this tool as only one component of a discovery pipeline wherein a motif model built from a small dataset will be used to find more instances, allowing the model to be extended and refined iteratively.

Extensive testing demonstrates that these goals are largely met. For most tests, the hand-curated ‘seed’ alignments from Rfam are our gold standard. CMfinder achieved better results than other

methods in 17 of 19 tests, and predicted base pairs with average 77% sensitivity, 81% specificity and 79% accuracy, compared with at most 60% accuracy for the other methods. Most disagreements with Rfam are local perturbations such as small shifts or extra base pairs. Finally, we have very encouraging results from preliminary tests combining our method with other computational tools to construct a pipeline for novel ncRNA family discovery. In particular, integrating with the 'Footprinter' DNA motif discovery tool (Blanchette and Tompa, 2003) and an RNA genome search tool (Weinberg and Ruzzo, 2004a,b), CMfinder produced highly accurate models of several large *cis*-regulatory RNA families, starting from unaligned 5′-UTRs of small sets of orthologous genes. For example, a blind test identified 447 instances of a bacterial motif that turned out to be the T-box leader. These included 87% of its 342 Rfam family members, plus an additional 148 hits, 89 of which are strongly supported by functional annotation. In this and six other datasets the CMfinder models identify members of corresponding Rfam families with 92% specificity and 80% sensitivity on average. We are currently applying this technique to genome scale search for novel *cis*-regulatory RNA motifs in microorganisms.

2 RELATED WORK

A classical approach to find RNA motifs is to construct a consensus RNA secondary structure from a given multiple sequence alignment based on covariation, thermodynamic stability, phylogeny, etc. (Eddy and Durbin, 1994; Hofacker *et al*., 2002; Akmaev *et al*., 2000; Gulko and Haussler, 1996; Knudsen and Hein, 1999). These methods require a good alignment, thus are unreliable for datasets with low sequence similarity. Coventry *et al*. (2004) identify conserved RNA regions based on sequence alignment by searching for correlated reverse-complementary regions. This method is robust to small variations of the alignment, but not major perturbations.

An alternative approach is to fold sequences separately, then find a consensus secondary structure, e.g. using a tree edit algorithm as in Hofacker *et al*. (1994) and Höchsmann *et al*. (2003). A key drawback of this approach is the inaccuracy of secondary structure prediction for a single sequence (Dowell and Eddy, 2004). Sankoff (1985); Gorodkin *et al*. (1997); Havgaard *et al*. (2005); Mathews and Turner (2002) solve this problem by inferring the alignment and folding simultaneously using dynamic programing. However, computational expense limits these approaches to small datasets.

Probabilistic approaches such as EM algorithms and Gibbs sampling have been applied to infer consensus RNA structure from unaligned sequences, based on covariance models or stochastic context-free grammars (Eddy and Durbin, 1994; Sakakibara *et al*., 1994; Grate *et al*., 1994). These methods, however, are not designed for the motif discovery problem in which RNA motifs are only present in local regions of a subset of sequences.

Graph theoretical techniques have been proposed to identify RNA motifs in unaligned sequences by comparison and assembly of stable stems (Touzet and Perriquet, 2004; Ji *et al*., 2004; Bafna *et al*., 2005). See Gardner and Giegerich (2004) for a recent comparative survey on RNA structure prediction.

3 METHODS

The basic idea of CMfinder is to use a CM to model an RNA motif, a finite mixture model to describe motif distribution in sequences, and an EM framework to search the motif space. It is motivated by two previous techniques: the DNA motif finding program MEME (Bailey and Elkan, 1995), and the CM-based RNA analysis tool COVE (Eddy and Durbin, 1994). The combination of these two methods is non-trivial owing to the increased complexity: in MEME, the motif model is an ungapped weight matrix with a relatively short, fixed length window, while RNA motifs are generally much longer with significant secondary structures and frequent insertions/deletions. COVE assumes that each sequence is an instance of the model and performs global alignment rather than the more difficult local alignment we need. In the EM framework, the expanded search space and higher model complexity suggest that it is infeasible to explore all subsequences as MEME does, and having a good starting point is critical. To address these issues, CMfinder chooses motif candidates with potentially stable secondary structures, selects a conserved set across all sequences and aligns them heuristically. This step is loosely similar to Carnac (Touzet and Perriquet, 2004) and ComRNA (Ji *et al*., 2004). The subsequent EM iteration refines the model and alignment. The following section elaborates this idea, and further technical details are available in the Supplementary information.

3.1 Construction of heuristic initial alignment

The goal of this step is to identify the approximate location and structure of the motif efficiently. The key design issue is the trade-off between accuracy and efficiency. As the motif will be refined later, we can tolerate alignment errors provided correctly aligned instances are well-represented, but need robustness to differences in dataset size, sequence similarity and consensus structure.

3.1.1 Candidate selection We first eliminate a large portion of the search space by focusing on strong candidates, i.e. segments with potentially stable secondary structure. For each input sequence, we use Vienna (Hofacker *et al*., 1994) to compute the minimum free energy of all subsequences, sort them according to their free energy scaled by sequence length, then choose the top ranking candidates iteratively. Similar ideas in Carnac and ComRNA use simple stems as candidates, but in our context allowing a richer set of candidates gives better performance.

3.1.2 Candidate comparison and alignment To find the consensus alignment of the candidates, our next step is to compare their predicted secondary structures. We use the tree-edit algorithm of Hofacker *et al*. (1994), modified to compare candidates at the single base or base pair level so that the comparison is sensitive to both sequence and structure. This improves its ability to distinguish between RNAs with relatively simple structures. The complexity of this algorithm is approximately quadratic in the length of the candidates, thus far more efficient than the Sankoff-style algorithms (Sankoff, 1985; Gorodkin *et al*., 1997; Mathews and Turner, 2002). It is a simpler framework for structure and sequence comparison than the the heuristics used in Carnac and ComRNA. Its drawbacks include the potential inaccuracy of secondary structure prediction and the simplified edit distance model. Despite the efficiency of the tree-edit algorithm, pairwise comparisons of numerous candidates can be expensive. To improve efficiency further, we only align two candidates if they are compatible with locally conserved regions found by BLAST search. This heuristic also improves accuracy by preventing obvious misalignments. CARNAC and comRNA also rely on similar anchor-based techniques to reduce their otherwise prohibitive computational cost; it is less critical in our case, but still valuable.

To construct the initial alignment, we want to choose one candidate from each sequence so as to minimize the sum of pairwise scores. We approximately solve this NP-hard problem using a heuristic that selects a central consensus candidate, and structually aligns it to its best match in each sequence. We repeat this process on the unselected candidates to find 1–10 initial alignments as seeds for the EM algorithm.

3.2 Refining alignments via CM-based EM

To improve the quality of the initial alignments, we adopt an iterative EM-like algorithm based on covariance models. The CM (Eddy and Durbin, 1994) is a probabilistic model for RNA families that cleanly describes both the secondary structure and the primary sequence consensus. We apply a finite mixture model to describe a sequence as a mixture of regions that follow the background distribution and regions that follow the motif CM, then use the EM algorithm to estimate the model and motif instances simultaneously. For clarity, we first assume that motif instances only occur among candidates, a restriction we relax in Section 3.2.3. The following notations are used:

- M: the motif CM.
- B: the background distribution.
- $\Gamma = (M,B,\gamma)$: the finite mixture model, where γ is the mixture probability that a sequence contains a motif.
- N: the total number of sequences.
- $S = (s_i)_{1 \le i \le N}$: the input sequences.
- m: the number of candidates in each sequence.
- $C_i = (c_{ij})_{1 \le j \le m}$: the candidate set of sequence s_i.
- $\Pi_i = (\pi_{ij})$: the alignments of candidates C_i with M.
- $X_i = (x_{ij})$: the occurrence of the motif in C_i ($x_{ij} = 1$ if c_{ij} is a motif instance, and $x_{ij} = 0$ otherwise).
- $D = (L_1, L_2, \ldots, L_l)$: the sequence alignment. L_i: a column.
- $\sigma = (\alpha,\beta)$: consensus secondary structure for D. α: a set of (indices of) single-stranded columns. β: a set of (pairs of indices of) base paired columns.

The aim is to find $\Gamma = (M,B,\gamma)$ that maximizes the log likelihood

$$\log P(S|\Gamma) = \log \prod_i \sum_{X_i} \sum_{\Pi_i} P(s_i, X_i, \Pi_i|\Gamma).$$

The E-step estimates the hidden variables X_i and Π_i, while the M-step updates the CM M and γ. The major CM functions (e.g. alignment, scanning, etc.) are adopted from Eddy and Durbin (1994), and will not be explained here.

3.2.1 The E-step For each candidate c_{ij} of sequence s_i, we need to estimate two hidden variables: alignment π_{ij} and motif assignment x_{ij}. Given a motif assignment, the inside–outside algorithm is usually used for CM parameter estimation by summing over all possible alignment paths. However, in order to update the model structure, we need a sequence alignment (see M-step). We use the Viterbi algorithm to compute the optimal alignment π_{ij} of candidate c_{ij}. Assuming zero or one motif occurrence per sequence ($\sum_j x_{ij} \le 1$), the motif assignment probabilities can be estimated based on a mixture model (Bailey and Elkan, 1995):

$$P(x_{ij} = 1) = \frac{\lambda P(c_{ij}|M)/P(c_{ij}|B)}{1 - \gamma + \sum_{k=1}^{m} \lambda P(c_{ik}|M)/P(c_{ik}|B)}, \quad (1)$$

where $\lambda = \gamma/m$. To save the cost of the inside algorithm for $P(c_{ij}|M)$, we use the probability of the optimal alignment $P(\pi_{ij}|M)$ instead. Although this approximation tends to underestimate $P(x_{ij} = 1)$, most real motif instances distinguish themselves from the background so dramatically that this approximation is sufficient. The resulting probability can be interpreted as the probability of a candidate with the suggested structural alignment being a motif instance.

3.2.2 The M-step Here, we update M and γ. The maximum likelihood estimate of γ is simply $(1/N) \sum_{i=1}^{N} \sum_{j=1}^{m} x_{ij}$. To update M, we first need to adjust the structure of M, then infer the transition and emission probabilities. Given the structure, the second issue can be solved using a Bayesian posterior estimate with Dirichlet prior (Eddy and Durbin, 1994), hence, we focus on the first issue. This is easy in MEME because the structure is predetermined. In CMfinder, it is equivalent to finding a consensus secondary structure. We formulate this in the following Bayesian framework.

Our goal is to find $\hat{\sigma} = \arg\max_{\sigma} P(D,\sigma)$. Assuming independence of non-base paired columns, then

$$P(D|\sigma) = \prod_{k \in \alpha} P(L_k) \prod_{(i,j) \in \beta} P(L_i L_j) \quad (2)$$

$$= \prod_{1 \le k \le l} P(L_k) \prod_{(i,j) \in \beta} \frac{P(L_i L_j)}{P(L_i)P(L_j)} \quad (3)$$

$$\text{Let } I_{ij} = \log \frac{P(L_i L_j)}{P(L_i)P(L_j)}.$$

Using maximum likelihood parameter estimates and a multinomial model for each column/column pair, I_{ij} is the mutual information between columns i and j. Without prior information, the optimal structure maximizes $\sum_{(i,j) \in \beta} I_{ij}$, which is the approach adopted by COVE. This method works well in large, phylogenetically diverse datasets, but not when there is insufficient covariance, as would be expected with a few closely related sequences.

We solve the problem by introducing an informative prior on structures. Let q_i be the prior for column i to be single stranded, and p_{ij} the prior for columns i, j to be base paired, then $P(\sigma) = \prod_{k \in \alpha} q_k \prod_{(i,j) \in \beta} p_{ij}$, and $P(D,\sigma)$ can be rewritten as

$$P(D,\sigma) = P(D|\sigma)P(\sigma) = \prod_{1 \le k \le l} P(L_k) q_k \prod_{(i,j) \in \beta} \frac{P(L_i L_j)}{P(L_i)P(L_j)} \frac{p_{ij}}{q_i q_j} \quad (4)$$

$$\text{Let } K_{ij} = \log \left(\frac{P(L_i L_j)}{P(L_i)P(L_j)} \frac{p_{ij}}{q_i q_j} \right) = I_{ij} + \log \frac{p_{ij}}{q_i q_j},$$

then the maximum likelihood structure σ maximizes $\sum_{(i,j) \in \beta} K_{ij}$. We infer a prior on structures based on a thermodynamic model. For each sequence, we calculate the partition function P_{ij} (Hofacker *et al*., 1994; McCaskill, 1990), which estimates the probability of forming base pair i,j, averaged over all possible structures. We estimate the column pairing probabilities p_{ij} by averaging the partition functions of the aligned sequences. The probability that a column is unpaired is estimated as $q_i = 1 - \sum_j p_{ij}$. Note that candidates are weighted based on their probabilities to be motif instances when computing I_{ij}, p_{ij} and q_i. Finally, we use a dynamic programming algorithm to choose a set of compatible base pairs maximizing the sum of K_{ij}. Since p_{ij} and q_i are predicted from the given sequences, they are not 'priors' in a strict sense. However, the mutual information and the partition function look at the same data from different perspectives: the mutual information measures the conservation of base pairs in the particular sequences from a evolutionary point of view, while the partition function is based on a thermodynamic model that is generically applicable to all RNAs. Combining them gives us the power of both approaches: we rely on the energy model when there is little mutual information and use mutual information if the structure is ambiguous based on the energy model. In comparison with our method, RNAalifold (Hofacker *et al*., 2002) uses a linear combination of the energy contribution and mutual information. Our probabilistic integration provides some justification for combining these two seemingly disparate elements.

3.2.3 Adjusting candidates We introduce a second EM phase identical to the first, except that the CM is used to scan each sequence, and the top hits in each are treated as candidates. This helps to discover motif instances missed by initial candidate selection and to refine the alignment.

3.2.4 Adding additional family members The EM framework enables automatic update of the covariance models based on new sequences that may contain the motif. We simply repeat the EM algorithm using the existing motif as seed. This feature is highly effective at discovering large

and diverse RNA families from a small set of related sequences; see Section 4.3.

3.2.5 Combining Motifs In theory, this method can find arbitrarily large motifs, but in practice, it works best with relatively short ones (<100 bases). To overcome this, we apply a greedy algorithm to merge multiple motifs hierarchically. See Supplementary information for details.

3.2.6 Run time The computation time of the EM algorithm is dominated by the Viterbi alignment, with complexity $O(NL^3|M|)$ for each iteration (where L is the maximum sequence length and $|M|$ is the number of states in the CM). The EM algorithm generally converges in <15 iterations. Overall, a typical CMfinder run (on <60 sequences of <1 Kb average length) takes 1–60 min, depending on the number and the complexity of motifs, and is practical for most applications.

4 RESULTS

Rfam is a large collection of multiple sequence alignments and CMs for ncRNA families (Griffiths-Jones *et al*., 2003). It contains a CM built from a hand curated 'seed' alignment for each ncRNA family. Additional homologs are then predicted by searching genome databases with the model. We used Rfam seed alignments to evaluate CMfinder's performance on three increasingly difficult tasks:

(1) Discovery: given unaligned seed members of an Rfam family (together with flanking regions), can CMfinder construct a good alignment and CM to characterize the family?
(2) Robustness: How does the quality of the alignment degrade as more flanking sequence is added and as seed sequences are replaced by unrelated sequences?
(3) Scale-up: Does it scale to plausible genome-wide discovery scenarios, where the initial set of sequences are taken, say, from a small group of orthologous genes? Are the sensitivity and specificity of the CMfinder model adequate when used to scan several gigabases of genome sequence?

In all the three tasks, our results are strong.

4.1 Discovery

We selected 19 families from Rfam (release 6.1, Aug 2004) as our test data, with varying length (26–216 bases), sequence identity (43–81%), and number of family members (9–75). This dataset captures the diversity of known RNA families, while excluding highly conserved ones, and emphasizing *cis*-regulatory elements, especially riboswitches. For each family, we took the seed alignment as the motif, and included 200 bases of genomic sequence flanking the motif, randomly distributed between the 5′ and 3′ sides, to simulate the realistic situation where motif locations are unknown.

We compared CMfinder with RNAalifold (Hofacker *et al*., 2002), Pfold (Knudsen and Hein, 1999), Foldalign (Havgaard *et al*., 2005; Gorodkin *et al*., 1997), ComRNA and Carnac. The accuracy of all predictions are computed at the base pair level relative to Rfam annotation. Let TP be the correctly predicted base pairs, FP the falsely predicted base pairs and FN the true base pairs that are not predicted. The sensitivity is defined as TP/(TP + FN), specificity as TP/(TP + FP) and overall prediction accuracy as their geometric mean. For decent sensitivity and specificity, the latter metric approximates Matthews Correlation (Gorodkin *et al*., 2001). Note that this is a very stringent metric. Small shifts and extra base pairs are counted as false negatives/positives while they are considered neutral in some other work. More importantly, this penalizes incorrect motif boundaries in this local alignment setting.

Fairly benchmarking different programs created with disparate goals is tricky. We are sometimes using these tools for purposes that they were not designed or optimized to do. Nevertheless, we feel that our choices are plausible ones for our goals (automated, genome-scale RNA motif discovery) and constitute a reasonable comparison of these tools for that purpose. The results may also illustrate the consequences of such tool abuse. CMfinder performs well in this context, but the other tools may excel in other contexts (or this one, if more cleverly used). In general, all programs received the same input and were run with default parameters without any per-data-set tuning. As one exception, among the programs tested, only ComRNA can predict pseudoknots, but we chose non-default parameter settings preventing this, which may understate its performance. (Lacking a 'gold standard' for pseudoknots in the test data, all obvious alternatives seemed worse.) As another important exception, both RNAalifold and Pfold require aligned inputs, so we used ClustalW alignments; other programs were given unaligned inputs. For ComRNA, we set a run time constraint comparable with the run time of CMfinder, and chose the motif with the best accuracy for each dataset. For Foldalign, we tried both multiple alignment (Gorodkin *et al*., 1997) and pairwise alignment (Havgaard *et al*., 2005), and report the results for the latter as it is faster and generally gives better results. We summarize its accuracy as the average of all pairwise comparisons. Although pairwise alignments are not directly comparable with multiple alignments, the tool ultimately attempts to achieve the same goal as our tool, and it is interesting to test whether more sequences improve consensus structure prediction accuracy. For CMfinder, we produced up to 10 motifs for each dataset, combined them, and retained the motif with the best average alignment score as the final output.

The descriptions of each Rfam family and prediction accuracies of all tested methods are summarized in Table 1. More detailed comparisons including predicted structures and alignments are included in the Supplementary information. CMfinder achieved the best performance on all families except s2m and RFN. The disagreement with Rfam for s2m is because of a small shift of a helix. For RFN, the motif we produced is partial, and most prediction errors are local, and in regions with great sequence conservation. CMfinder significantly outperformed the other methods on families with low sequence conservation or short motifs such as the SECIS family. This is a difficult test case owing to low sequence similarity, and two conserved non-canonical G–A pairs in the stem–loop. The other four methods tested predict no base pairs, while CMfinder correctly aligns and annotates the region enclosed by the two G–A pairs. RNAalifold, Pfold, Carnac and ComRNA have relatively weak performance for such families, presumably because sequence conservation is insufficient to delineate possible alignments. Inclusion of arbitrary flanking regions makes sequence-based alignment even harder. Although our initial pairwise alignment algorithm is much simpler than pairwise Foldalign, we gained more information by comparing all sequences. We have also tested a set of methods that perform global alignments on Rfam families without flanking regions. Again, CMfinder usually outperforms other methods; see Supplementary information.

To quantify the effectiveness of each component of our algorithm, we compared the performance of CMfinder with its four

Table 1. Summary of Rfam test families and results

ID	Family	Rfam ID	#seqs	%id	length	#hp	CMfinder	CW/Pfold	CW/RNAalifold	Carnac	Foldalign	ComRNA
1	Cobalamin	RF00174	71	49	216	4	**0.59**	0.05	0	X	—	0
2	ctRNA_pGA1	RF00236	17	74	83	2	**0.91**	0.70	0.72	0	0.86	0
3	Entero_CRE	RF00048	56	81	61	1	**0.89**	0.74	0.22	0	—	0
4	Entero_OriR	RF00041	35	77	73	2	**0.94**	0.75	0.76	0.80	0.52	0.52
5	glmS	RF00234	14	58	188	4	**0.83**	0.12	0.18	0	—	0.13
6	Histone3	RF00032	63	77	26	1	**1**	0	0	0	—	0
7	Intron_gpII	RF00029	75	55	92	2	**0.80**	0.30	0	0	—	0
8	IRE	RF00037	30	68	30	1	**0.77**	0.22	0	0	0.38	0
9	let-7	RF00027	9	69	84	1	**0.87**	0.08	0.42	0	0.71	0.78
10	lin-4	RF00052	9	69	72	1	**0.78**	0.51	0.75	0.41	0.65	0.24
11	Lysine	RF00168	48	48	183	4	**0.77**	0.24	0	X	—	0
12	mir-10	RF00104	11	66	75	1	**0.66**	0.59	0.60	0	0.48	0.33
13	Purine	RF00167	29	55	103	2	**0.91**	0.07	0	0	—	0.27
14	RFN	RF00050	47	66	139	4	0.39	**0.68**	0.26	0	—	0
15	Rhino_CRE	RF00220	12	71	86	1	**0.88**	0.52	0.52	0.69	0.41	0.61
16	s2m	RF00164	23	80	43	1	0.67	**0.80**	0.45	0.64	0.63	0.29
17	S_box	RF00162	64	66	112	3	**0.72**	0.11	0	0	—	0
18	SECIS	RF00031	43	43	68	1	**0.73**	0	0	0	—	0
19	Tymo_tRNA-like	RF00233	22	72	86	4	**0.81**	0.33	0.36	0.30	0.80	0.48
				Average accuracy:			**0.79**	0.36	0.28	0.17	0.60	0.19
				Average specificity:			0.81	0.42	0.57	**0.83**	0.60	0.65
				Average sensitivity:			**0.77**	0.36	0.23	0.13	0.61	0.17

#seqs, the number sequences in each family's seed alignment. (For ease of post processing, we only chose one sequence per EMBL ID.) %id, average sequence identity among family members. length, average length of family members (nucleotides). #hp, number of hairpin-loops in the consensus structure. Last six columns, accuracies; bold highlights the best result in each row. CW/Pfold, Pfold using ClustalW alignment. CW/RNAalifold, similar. (X, Carnac terminated abnormally, presumably due to memory problems. —, Foldalign (pairwise) not tested due to the heavy computation cost. RNAalifold, Carnac and ComRNA do not predict any consensus structure in many cases, so the corresponding accuracies are 0.)

variants: (1) heuristic initial alignment before EM iteration (*Ini*); (2) initial alignment with EM iteration based on mutual information only (*Ini_em_mi*); (3) initial alignment with EM iteration based on folding energy only (Ini_em_fe) and (4) ClustalW alignment with EM iteration (*Clustal_em*). To speed the EM iteration, we trimmed regions with >10% gaps from both ends of the ClustalW alignment.

The motif prediction accuracy of the these five methods are shown in Figure 1. First, we observe that the EM iteration improves the prediction accuracy considerably, from an average of 66% in the initial alignments to 83%. Second, the energy-based partition function (Ini_em_fe) generally outperforms mutual information (*Ini_em_mi*) in the EM algorithm, while using mutual information in addition to folding energy yields improvements on SECIS and s2m. Third, CMfinder has better performance than Clustal_em except for RFN and S_box. For RFN, the CMfinder motif is only partial. Meanwhile, CMfinder is far more effective at locating poorly conserved and/or short motifs, such as IRE and SECIS.

This suggests that our heuristics are generally more robust, but ClustalW can be a complementary alternative for constructing initial alignments. Finally, there is significant improvement of Clustal_em over Pfold and RNAalifold. All three are based on ClustalW alignment, yet Clustal_em achieves 61% average prediction accuracy, compared to 36% for Pfold and 27% for RNAalifold. On families where both RNAalifold and Pfold fail, such as S_box, Cobalamin and Purine, Clustal_em has 55–80% prediction accuracy. To summarize, both the initial alignment procedure and the EM module are effective components of CMfinder, which make it reliable on a variety of datasets.

4.2 Robustness

We tested CMfinder on more challenging datasets with larger flanking regions and only a subset of the sequences containing real RNA motifs. To form each dataset, we randomly selected n family members, including a given length of flanking sequence, then permuted the motif regions in k of them (again, randomly selected). The latter sequences serve as control sequences; the rest, with real RNA motifs, are referred to as test sequences. We performed this test on Histone3, IRE and SECIS families. Only the IRE results are presented here; others are available at our supplementary website. Figure 2a shows the scores of the motif instances produced by CMfinder when tested on datasets with flanking regions varying from 50 to 400 bases on both the 5′ and 3′ sides, with 1/4 control sequences, while Figure 2b varies the fraction of control sequences from 1/8 to 1/2, all with 100-base flanking regions. We categorized three types of predicted motif instances: true and false motif instances in the test sequences (Tt and Ft respectively) and those in the control sequences (Fc). Tts and Fts are determined by whether their overlaps with corresponding Rfam motif instances are >10 bases. Most overlaps were shorter than 5 or longer than 25 bases.

Figures 2a and b generally show good score separations between true motif instances (Tt) and false ones (Ft and Fc). As the flanking region increases, candidate selection becomes more difficult owing

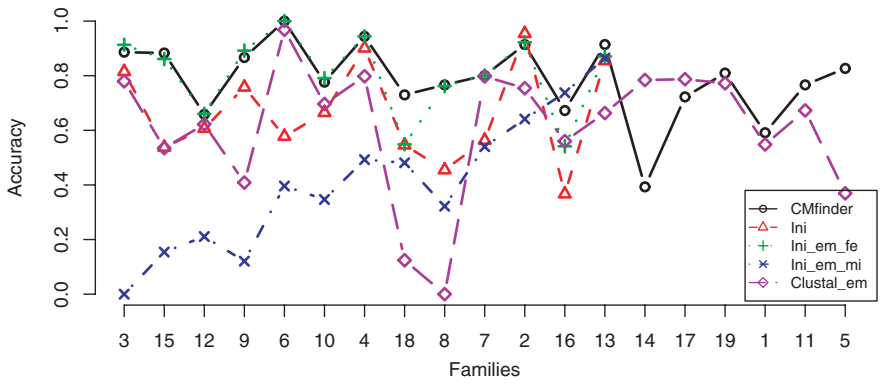

Fig. 1. Comparison of CMfinder with its variants. The initial alignment corresponding to the best output motif is selected for each family. For the rightmost 6, final CMfinder motifs are combinations of multiple motifs, precluding comparison to ini, ini_em_mi and ini_em_fe.

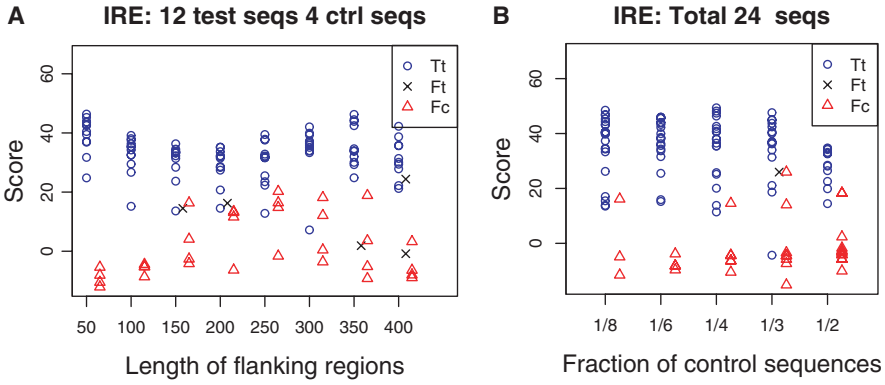

Fig. 2. (a) and (b) Robustness Test. Each column represents a test dataset, each point represents a motif instance predicted in a sequence. Tt, True motif in test seq.; Ft, False motif in test seq.; Fc, False motif in control seq. (see text).

to the larger number of stable local structures (we used the same number of candidates in all the test cases), and good local alignments occur more easily by chance. Although the score differences between the true and false ones tend to decrease as the flanking region increases, the distinction is generally clear enough to differentiate the two types. There are a total of 20 false motif instances in all datasets above the score threshold 10, but closer examination reveals that 11 of them in fact correspond to real IREs (which often occurs in tandem) present in Rfam, but not seed members. The score difference between true and false motifs is even more apparent in Figure 2b. The only Ft turns out to be a real IRE element. In both figures, there is only one Tt with score <10. Overall, even in the presence of significant amounts of extraneous sequence, CMfinder successfully predicted Histone3, IRE and SECIS motifs, which are among the more difficult of our test families.

4.3 Scale-up: a framework for novel ncRNA discovery

CMfinder is applicable to unaligned sequences with low sequence conservation, is robust to noise, produces highly descriptive motif models and permits automatic update based on new sequence data. These features make it an effective tool for novel ncRNA discovery as we demonstrate in a pilot application of this type, sketched below.

Cis-regulatory elements such as riboswitches can be expected to be conserved in upstream regions of orthologous genes. In an ongoing collaboration with Martin Tompa and colleagues at UW, we have begun a systematic search for ncRNAs based on this scenario. In outline, we are doing the following. For each protein coding gene in a specific organism (*Bacillus subtilis*, initially), we find its 19 closest matches in GenBank bacterial genomes based on pairwise BLAST scores of protein sequences. We rank the datasets based on patterns of conservation identified by the DNA phylogenetic footprinting tool Footprinter (Blanchette and Tompa, 2003) in the upstream intergenic regions of each group of 20 orthologs. Finally, we apply CMfinder to analyze the top ranking datasets. The covariance models of promising RNA motifs are used to scan the bacterial genome database. In some cases, hits from this scan were used to refine the RNA motifs, and the genome scan repeated. Second scans used an *E*-value cutoff of 1, but initial scans used an cutoff of 100, as the initial models learned from small datasets tend to be overly specific. Of the initial 115 datasets, we found that over 30 correspond to ribosomal RNAs, 13 contain T_boxes, 22 contain riboswitches, with additional sporadic cases of RNase P, tRNAs, the CIRCE element (Narberhaus, 1999) and other DNA binding sites. We selected seven predicted motifs to analyze more fully and determined their agreement with corresponding Rfam families. The results are summarized in Table 2.

Table 2. Comparison of scan predictions with Rfam results

Gene	#motif[a]	#hits[b]	Rfam	#seed[c]	#full[d]	#spec[e]	#sens[f]
metK	13	150	S_box	71	151	0.97	0.96
ribB	9	106	RFN	48	114	0.92	0.85
folC	9	447	T_box	67	342	0.67	0.87
xpt	14	106	Purine	37	100	0.92	0.97
glmS	16	33	glmS	14	37	1.00	0.89
thiA	16	305	THI	237	366[g]	1.00	0.83
ykoY	10	34	yybP-ykoY	74	127	0.97	0.26

[a]The number of sequences with the predicted motif.
[b]The number of hits in the final scan.
[c,d]The number of seed/full members in the Rfam family.
[e,f]The specificity and sensitivity of predicted members relative to Rfam.
[g]Excluding 16 members in Eukaryotic genomes, which were not scanned.

In one example, from upstream sequences of 20 orthologous folC genes, we found a conserved RNA motif with nine instances. The first and the second genome scan found 234 and 447 hits respectively. Jeffrey Barrick (personal communication) identified this as the T_box leader (Grundy *et al*., 1994), a *cis*-regulatory element that interacts with tRNA to control expression of tRNA processing genes. Our motif model discovered 299 out of 342 Rfam T_box members, and 89 out of 148 additional hits are upstream of and on the same strand as aminoacyl-tRNA synthetase genes, where most T_box leaders are found, and the others are largely in poorly annotated regions. While we started with a small set of sequences from phylogenetically closely related species (Bacillales/Clostridia), the genome scan discovered a diverged homolog in *Symbiobacterium thermophilum* (an Actinobacterium) included in Rfam, and another in *Geobacter sulfurreducens* (a δ Proteobacterium), presumably the one reported in Winkler *et al*. (2001).

We repeated the test for the other families, although for the metK and ribB datasets, we scanned the sequence database only once. Averaged over all seven families, our predicted motifs achieved 92% specificity and 81% sensitivity. Except for the T_box, the specificities range from 92 to 100%. For the T_box family, if the 89 additional hits supported by annotation are counted as true positives, then specificity increases to 87% from 67%. All families except ykoY have sensitivity over 80%. Its poor sensitivity is likely due to the fact that the motif is only learned from sequences upstream of ykoY genes, while the Rfam seed also includes motifs from yybP genes. Six of the seven datasets were picked because they held known riboswitches, to confirm that our methods 'discover' them. This may have unconsciously biased our execution of the experiment, but the T_box example was entirely blind—we did not know there was an RNA motif near folC genes, only discovering its identity after completing the scans. See Supplementary information for details.

In summary, we have automatically learned highly accurate CMs from small automatically constructed datasets. In contrast, the Rfam models are learned from hand curated seed alignments, usually containing many more sequences. Automated model construction should not supplant the high-quality curated Rfam seed alignments, but we are optimistic that it will allow broad-scale screening for new *cis*-regulatory ncRNAs with minimal manual intervention.

5 CONCLUSION

In this paper, we presented an algorithm for finding conserved RNA motifs in a set of unaligned sequences. The key contribution of this paper is to propose an EM framework that integrates various judiciously chosen techniques for secondary structure prediction and alignment which together makes this approach computationally feasible, robust and accurate. In particular, we proposed a principled statistical framework that combines an energy model with mutual information, which significantly improves the performance. In addition, we have presented an effective RNA motif discovery pipeline using CMfinder and other tools to iteratively expand and refine the motifs automatically. We are now applying this discovery pipeline for large-scale screening of RNA elements in microorganisms, and our preliminary results are very promising. Our future work will involve designing motif significance test statistics, incorporating phylogenetic information, and better priors for covariance models.

ACKNOWLEDGEMENTS

We thank the reviewers for constructive comments; J. Barrick for offering his biological expertise; B. Knudsen for sharing Pfold; A. Prakash, S. Neph and M. Tompa for valuable input at several stages and insightful comments on the manuscript. Supported in part by NHLBI 1 P01 HL072262-01, HG-00035 and NIEHS P30ES07033.

Conflict of Interest: none declared.

REFERENCES

Akmaev,V.R. *et al*. (2000) Phylogenetically enhanced statistical tools for RNA structure prediction. *Bioinformatics*, **16**, 501–512.

Bafna,V., Tang,H. and Zhang,S. (2005) Consensus folding of unaligned RNA sequence revisited. In *Proc. Res. Comp. Mol. Biol*., Cambridge, MA, p1.

Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. In *Proc. Intel. Sys. Mol. Biol* . St. Louis, pp. 21–29.

Barrick,J.E. *et al*. (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl Acad. Sci. USA*, **101**, 6421–6426.

Blanchette,M. and Tompa,M. (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res*., **31**, 3840–3842.

Conne,B. *et al*. (2000) The 3′ untranslated region of messenger RNA: A molecular 'hotspot' for pathology? *Nat. Med*., **6**, 637–641.

Coventry,A *et al*. (2004) MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 12102–12107.

Dowell,R.D. and Eddy,S.R. (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 71.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*., **22**, 2079–2088.

Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.

Gorodkin,J. *et al*. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res*., **25**, 3724–3732.

Gorodkin,J. *et al*. (2001) Discovering common stem–loop motifs in unaligned RNA sequence. *Nucleic Acids Res*., **29**, 2135–2144.

Grate,L. *et al*. (1994) RNA modeling using Gibbs sampling and stochastic context free grammars. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **2**, 138–146.

Griffiths-Jones,S. *et al*. (2003) Rfam: an RNA family database. *Nucleic Acids Res*., **31**, 439–441.

Grundy,F. *et al*. (1994) Interaction between the acceptor end of tRNA and the T box stimulates antitermination in the *Bacillus subtilis* tyrS gene: a new role for the discriminator base. *J. Bacteriol*., **176**, 4518–4526.

Gulko,B. and Haussler,D. (1996) Using multiple alignments and phylogenetic trees to detect RNA secondary structure. *Pac Symp Biocomput*., 350–367.

Havgaard,J. *et al*. (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**, 1815–1824.

Hentze,M.W. and Kuhn,L.C. (1996) Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc. Natl Acad. Sci. USA*, **93**, 8175–8182.

Höchsmann,M., Toller,T., Giegerich,R. and Kurtz,S. (2003) Local similarity in RNA secondary structure. In *Proc. Compu. Sys. Bioinfo.,* Stanford, CA, pp. 159–168.

Hofacker,I.L. *et al*. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol*., **319**, 1059–1066.

Hofacker,I.L. *et al*. (1994) Fast folding and comparison of RNA secondary structure. *Chemical Monthly*, **125**, 167–188.

Ji,Y. *et al*. (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, **20**, 1591–1602.

Knudsen,B. and Hein,J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446–454.

Mandal,M. *et al*. (2003) Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell*, **113**, 577–586.

Mandal,M. *et al*. (2004) A glycine-dependent riboswitch that uses cooperative binding to control gene expression [Erratum (2004) *Science*, **306**, 1477]. *Science*, **306**, 275–279.

Mathews,D.H. and Turner,D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol*., **317**, 191–203.

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Narberhaus,F. (1999) Negative regulation of bacterial heat shock genes. *Mol. Microbiol*., **31**, 1–8.

Sakakibara,Y., Brown,M., Mian,I.S., Underwood,R. and Haussler,D. (1994) Stochastic context-free grammars for modeling RNA. In *Proc. Hawaiian Inter. Conf. Sys. Sci*., Hawaii, pp. 284–294,

Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math*., **45**, 810–825.

Touzet,H. and Perriquet,O. (2004) CARNAC: folding families of related RNAs. *Nucleic Acids Res*., **32**, W142–W145.

Weinberg,Z. and Ruzzo,W.L. (2004a) Faster genome annotation of non-coding RNA families without loss of accuracy. In *Proc. Res. Compu. Mol. Bio*., Glasgow, Scotland, pp. 243–251.

Weinberg,Z. and Ruzzo,W.L. (2004b) Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics*, **20** (suppl.1), I334–I341.

Winkler,W. and Breaker,R.R. (2003) Genetic control by metabolite-binding riboswitches. *Chembiochem*., **4**, 1024–1032.

Winkler,W. *et al*. (2001) The GA motif: an RNA element common to bacterial antitermination systems, rRNA, and eukaryotic RNAs. *RNA*, **7**, 1165–1172.