Sequence analysis

CMfinder—a covariance model based RNA motif finding algorithm

Zizhen Yao1,*, Zasha Weinberg1 and Walter L. Ruzzo1,2
1Department of Computer Science and Engineering and 2Department 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.
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; Gulkis 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); Havgard 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 Perriguet, 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 unaggregated 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 sub-sequences 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 in Carnac (Touzet and Perriguet, 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 compute 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 approximatively solve this NP-hard problem using a heuristic that selects a central consensus candidate, and structurally 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 $\gamma$ is the mixture probability that a sequence contains a motif.
- $N$: the total number of sequences.
- $S = (s_i)_{1 \leq i \leq N}$: the input sequences.
- $m$: the number of candidates in each sequence.
- $C_i = (c_{ij})_{1 \leq j \leq m}$: 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 $D$. $\alpha$: a set of (indices of) single-stranded columns. $\beta$: a set of (pairs of indices of) base paired columns.

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

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

The E-step estimates the hidden variables $X_i$ and $\Pi_i$, while the M-step updates the CM $M$ and $\gamma$. 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 $\pi_{ij}$ 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 $\pi_{ij}$ of candidate $c_{ij}$. Assuming zero or one motif occurrence per sequence ($\sum x_{ij} \leq 1$), the motif assignment probabilities can be estimated based on a mixture model (Bailey and Elkan, 1995):

$$P(x_{ij} = 1) = \frac{\lambda P(c_{ij}|M) P(c_{ij}|B)}{1 + \sum_{a=1}^{\infty} \lambda P(c_{ij}|M) P(c_{ij}|B)},$$

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 $\gamma$. The maximum likelihood estimate of $\gamma$ 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 $\sigma = \arg \max_{\sigma} P(D, \sigma)$. Assuming independence of non-base paired columns, then

$$P(D|\sigma) = \prod_{i \leq w} P(L_i) \prod_{(i,j) \in \beta} P(L_i, L_j)$$

$$= \prod_{i \leq w} P(L_i) \prod_{(i,j) \in \beta} P(L_i|L_j)$$

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/pair column, $I_{ij}$ is the mutual information between columns $i$ and $j$. Without prior information, the optimal structure maximizes $\sum_{i,j} I_{ij}$, 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 $P(\sigma) = \prod_{i \leq w} q_i \prod_{(i,j) \in \beta} p_{ij}$, and $P(D, \sigma)$ can be rewritten as

$$P(D, \sigma) = P(D|\sigma) P(\sigma)$$

$$= \sum_{i \leq w} q_i \prod_{(i,j) \in \beta} p_{ij}$$

Let $K_{ij} = \log \frac{P(L_i|L_j)}{P(L_i) P(L_j)} = I_{ij} + \log \frac{p_{ij}}{q_i}$, then the maximum likelihood structure $\sigma$ maximizes $\sum_{i,j} K_{ij}$. 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 $ij$, 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 $q_i = 1 - \sum p_{ij}$. 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 an 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 $O(NL^2M)$ 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), Foldalgin (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 Foldalgin, 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 Foldalgin, 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
To summarize, both the initial alignment procedure 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 SECIS and IRE.

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 ClustalW over Pfold and RNAalifold. All three are based on ClustalW alignment, yet ClustalW 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, ClustalW 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 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

Table 1. Summary of Rfam test families and results

<table>
<thead>
<tr>
<th>ID</th>
<th>Family</th>
<th>Rfam ID</th>
<th>#seqs</th>
<th>%id</th>
<th>length</th>
<th>#hp</th>
<th>CMfinder</th>
<th>CW/Pfold</th>
<th>CW/RNAalifold</th>
<th>Carnac</th>
<th>Foldalign</th>
<th>ComRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cobalamin</td>
<td>RF00174</td>
<td>71</td>
<td>49</td>
<td>216</td>
<td>4</td>
<td>0.59</td>
<td>0.05</td>
<td>0</td>
<td>X</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>cIRNA_pGAI</td>
<td>RF00236</td>
<td>17</td>
<td>74</td>
<td>83</td>
<td>2</td>
<td>0.91</td>
<td>0.70</td>
<td>0.72</td>
<td>0</td>
<td>0.86</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Entero_CRE</td>
<td>RF00484</td>
<td>56</td>
<td>81</td>
<td>61</td>
<td>1</td>
<td>0.89</td>
<td>0.74</td>
<td>0.22</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Entero_OriR</td>
<td>RF00041</td>
<td>35</td>
<td>77</td>
<td>73</td>
<td>2</td>
<td>0.94</td>
<td>0.75</td>
<td>0.76</td>
<td>0.80</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>5</td>
<td>glmS</td>
<td>RF00234</td>
<td>14</td>
<td>58</td>
<td>188</td>
<td>4</td>
<td>0.83</td>
<td>0.12</td>
<td>0.18</td>
<td>0</td>
<td>—</td>
<td>0.13</td>
</tr>
<tr>
<td>6</td>
<td>Histone3</td>
<td>RF00032</td>
<td>63</td>
<td>77</td>
<td>26</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>Intron_pPH</td>
<td>RF00229</td>
<td>75</td>
<td>55</td>
<td>92</td>
<td>2</td>
<td>0.80</td>
<td>0.30</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>IRE</td>
<td>RF00037</td>
<td>30</td>
<td>68</td>
<td>30</td>
<td>1</td>
<td>0.77</td>
<td>0.22</td>
<td>0</td>
<td>0</td>
<td>0.38</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>let-7</td>
<td>RF00027</td>
<td>9</td>
<td>69</td>
<td>84</td>
<td>1</td>
<td>0.87</td>
<td>0.08</td>
<td>0.42</td>
<td>0</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td>10</td>
<td>lin-4</td>
<td>RF00552</td>
<td>9</td>
<td>69</td>
<td>72</td>
<td>1</td>
<td>0.78</td>
<td>0.51</td>
<td>0.75</td>
<td>0.41</td>
<td>0.65</td>
<td>0.24</td>
</tr>
<tr>
<td>11</td>
<td>Lysinated</td>
<td>RF00168</td>
<td>48</td>
<td>48</td>
<td>183</td>
<td>4</td>
<td>0.77</td>
<td>0.24</td>
<td>0</td>
<td>X</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>mir-10</td>
<td>RF00104</td>
<td>11</td>
<td>66</td>
<td>75</td>
<td>1</td>
<td>0.66</td>
<td>0.59</td>
<td>0.60</td>
<td>0</td>
<td>0.48</td>
<td>0.33</td>
</tr>
<tr>
<td>13</td>
<td>Purine</td>
<td>RF00167</td>
<td>29</td>
<td>55</td>
<td>103</td>
<td>2</td>
<td>0.91</td>
<td>0.07</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td>0.27</td>
</tr>
<tr>
<td>14</td>
<td>RFN</td>
<td>RF00050</td>
<td>47</td>
<td>66</td>
<td>139</td>
<td>4</td>
<td>0.39</td>
<td>0.68</td>
<td>0.26</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>Rhino_CRE</td>
<td>RF00220</td>
<td>12</td>
<td>71</td>
<td>86</td>
<td>1</td>
<td>0.88</td>
<td>0.52</td>
<td>0.52</td>
<td>0.69</td>
<td>0.41</td>
<td>0.61</td>
</tr>
<tr>
<td>16</td>
<td>s2m</td>
<td>RF00164</td>
<td>23</td>
<td>80</td>
<td>43</td>
<td>1</td>
<td>0.67</td>
<td>0.80</td>
<td>0.45</td>
<td>0.64</td>
<td>0.63</td>
<td>0.29</td>
</tr>
<tr>
<td>17</td>
<td>S_box</td>
<td>RF00162</td>
<td>64</td>
<td>66</td>
<td>112</td>
<td>3</td>
<td>0.72</td>
<td>0.11</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>SECIS</td>
<td>RF00031</td>
<td>43</td>
<td>43</td>
<td>68</td>
<td>1</td>
<td>0.73</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>Tymo_tRNA-like</td>
<td>RF00233</td>
<td>22</td>
<td>72</td>
<td>86</td>
<td>4</td>
<td>0.81</td>
<td>0.33</td>
<td>0.36</td>
<td>0.30</td>
<td>0.80</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Average accuracy: 0.79 0.76 0.80 0.52 0.52
Average specificity: 0.81 0.42 0.57 0.83 0.60
Average sensitivity: 0.77 0.36 0.23 0.13 0.61

Summary of Rfam test families and results

#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.)
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 occur 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.
The specificity and sensitivity of predicted members relative to Rfam.
The number of seed/full members in the Rfam family.
The number of hits in the final scan.

Table 2. Comparison of scan predictions with Rfam results

<table>
<thead>
<tr>
<th>Gene</th>
<th>#motifs</th>
<th>#hits</th>
<th>Rfam</th>
<th>#seed</th>
<th>#full</th>
<th>#spec</th>
<th>#sens</th>
</tr>
</thead>
<tbody>
<tr>
<td>metK</td>
<td>13</td>
<td>150</td>
<td>S_box</td>
<td>71</td>
<td>151</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>ribB</td>
<td>9</td>
<td>106</td>
<td>RFN</td>
<td>48</td>
<td>114</td>
<td>0.92</td>
<td>0.85</td>
</tr>
<tr>
<td>folC</td>
<td>9</td>
<td>447</td>
<td>T_box</td>
<td>67</td>
<td>342</td>
<td>0.67</td>
<td>0.87</td>
</tr>
<tr>
<td>xpt</td>
<td>14</td>
<td>106</td>
<td>Purine</td>
<td>37</td>
<td>100</td>
<td>0.92</td>
<td>0.97</td>
</tr>
<tr>
<td>glmS</td>
<td>16</td>
<td>33</td>
<td>glmS</td>
<td>14</td>
<td>37</td>
<td>1.00</td>
<td>0.89</td>
</tr>
<tr>
<td>thiA</td>
<td>16</td>
<td>305</td>
<td>THI</td>
<td>237</td>
<td>366</td>
<td>1.00</td>
<td>0.83</td>
</tr>
<tr>
<td>ykoY</td>
<td>10</td>
<td>34</td>
<td>yybP-ykoY</td>
<td>74</td>
<td>127</td>
<td>0.97</td>
<td>0.26</td>
</tr>
</tbody>
</table>

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

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 PfPöld; A. Prakash, S. Nephi and M. Tompa for valuable input at several stages and insightful comments on the manuscript. Supported in part by NHIILB 1 P01 HL072262-01, HG-00035 and NIEHS P30ES07033.

Conflict of Interest: none declared.

REFERENCES

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


