HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots

JIHONG REN, BAHARAK RASTEGARI, ANNE CONDON, and HOLGER H. HOOS
Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

ABSTRACT

We present HotKnots, a new heuristic algorithm for the prediction of RNA secondary structures including pseudoknots. Based on the simple idea of iteratively forming stable stems, our algorithm explores many alternative secondary structures, using a free energy minimization algorithm for pseudoknot free secondary structures to identify promising candidate stems. In an empirical evaluation of the algorithm with 43 sequences taken from the Pseudobase database and from the literature on pseudoknotted structures, we found that overall, in terms of the sensitivity and specificity of predictions, HotKnots outperforms the well-known Pseudoknots algorithm of Rivas and Eddy and the NUPACK algorithm of Dirks and Pierce, both based on dynamic programming approaches for limited classes of pseudoknotted structures. It also outperforms the heuristic Iterated Loop Matching algorithm of Ruan and colleagues, and in many cases gives better results than the genetic algorithm from the STAR package of van Batenburg and colleagues and the recent pknotsRG-mfe algorithm of Reeder and Giegerich. The HotKnots algorithm has been implemented in C/C++ and is available from http://www.cs.ubc.ca/labs/beta/Software/HotKnots.

Keywords: RNA secondary structure prediction; pseudoknots; heuristic algorithms

INTRODUCTION

RNA molecules play diverse roles in the cell: they act as carriers of genetic information, catalysts in cellular processes, and mediators in determining the expression level of genes. The three-dimensional (3D) structure of an RNA molecule is often the key to its function. In turn, the 3D structure of an RNA molecule is significantly shaped by its "secondary structure," which is determined by the collection of hydrogen bonds between pairs of bases in the molecule. The secondary structure of a Hepatitis Delta Virus (HDV) ribozyme sequence is depicted in two different ways in Figure 1. The component stems (or helices) and loops that are formed by the structure are evident in Figure 1A. This structure has a "pseudoknot," indicated by the presence of crossed arcs in the arc representation of the structure, in Figure 1B. Pseudoknots are present in the secondary structure of many RNA molecules, such as ribosomal RNAs, the catalytic core of group I introns, RNase P RNAs, and viral RNAs. In many cases, pseudoknots play functional roles, for example in ribosomal frameshifting (Giedroc et al. 2000), regulation of translation and splicing (Draper et al. 1998), and selinocystein biosynthesis.

Computational prediction of RNA secondary structure is useful not only as a step in determining the 3D structure of RNA molecules, but also in directly inferring and comparing the functions of molecules. The most successful approach for computational RNA structure prediction is comparative sequence analysis, in which covarying residues are identified in a multiple sequence alignment of RNAs with similar structures, but different sequences (Eddy and Durbin 1994). Other comparative methods incorporate evolutionary information (Knudsen and Hein 1999) or are based on probabilistic models, such as stochastic context-free grammars (Durbin et al. 1998). However, these methods can only be used when several related RNA sequences are available and thus are not always applicable. Therefore, computational methods for predicting the secondary structure of a single RNA sequence are in demand, and as new roles for RNA continue to be discovered at a rapid pace, such computational methods are becoming increasingly important. Besides providing a tool for prediction of a single sequence, the study of computational RNA prediction models and methods can help us elucidate the principles governing RNA structure formation. In addition, when

Reprint requests to: Anne Condon, Department of Computer Science, University of British Columbia, 2366 Main Mall, Vancouver, BC, V6T 1Z4, Canada; e-mail: condon@cs.ubc.ca; fax: (604) 822-5485.
Article and publication are at http://www.rnajournal.org/cgi/doi/10.1261/rna.7284905.

RNA (2005), 11:1494–1504. Published by Cold Spring Harbor Laboratory Press.

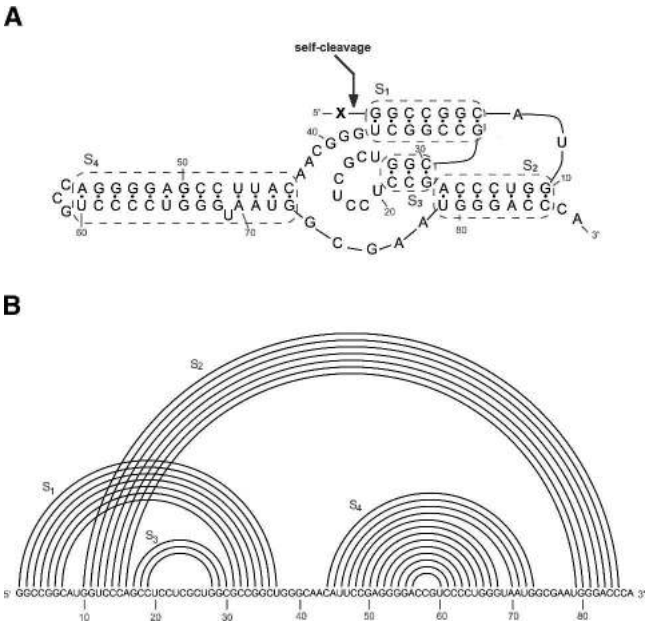

FIGURE 1. Pseudoknotted secondary structure of a Hepatitis Delta Virus ribozyme sequence. (*A*) Graphical view of the secondary structure, with the self-cleavage site indicated with an arrow. (*B*) Arc representation of the same secondary structure. Crossed arcs indicate that the structure is pseudoknotted.

multiple sequences are available, energy-based methods for secondary structure prediction can be combined with covariation methods, yielding better results (Hofacker et al. 2002). When the number of sequences is small or there is a single RNA molecule, prediction of RNA structure based on free energy minimization is the most widely used approach. That is, of the exponentially many possibilities, the goal is to find the structure that has the lowest free energy, where the energy is calculated as a function of thermodynamic free energy terms for the stems as well as conformational entropic terms for the loops. Some of these terms have been obtained experimentally, and others are estimated based on existing databases of naturally occurring structures (Mathews et al. 1999) or are approximated for algorithmic expediency.

Finding the thermodynamically most stable secondary structure for a molecule appears to be computationally intractable (NP-hard) (Akutsu 2000; Lyngsø and Pedersen 2000). Therefore, in practice, algorithms that output the minimum free energy structure for a given input RNA molecule can handle only a restricted class of structures. For example, a widely used algorithm (Zuker and Stiegler 1981), based on the dynamic programming method, finds the minimum free energy structure among those structures that are pseudoknot free. A variant of that algorithm (Lyngsø et al. 1999) has running time $O(n^3)$; throughout n denotes the number of bases of an RNA molecule. The dynamic programming approach has been developed further (McCaskill 1990) to provide the probability that each possible base pair is found in the structure for a given molecule. Further dynamic programming algorithms (Zuker 1989; Wuchty et al. 1998) provide a list of suboptimal structures that have free energy close to that of the minimum free energy structure. All of these developments have significantly enhanced the utility of tools for thermodynamic prediction of pseudoknot-free secondary structures.

Several other researchers have developed dynamic programming algorithms that find the minimum free energy structure from a restricted class that includes certain pseudoknotted structures (Rivas and Eddy 1999; Uemura et al. 1999; Akutsu 2000; Lyngsø and Pedersen 2000; Dirks and Pierce 2003). Of these, the algorithm of Rivas and Eddy can handle the broadest class of structures. In addition to this algorithmic contribution, Rivas and Eddy provide a complete model, along with parameters, for calculating the free energy of pseudoknotted secondary structures. However, the running time of their algorithm is $O(n^6)$, making it feasible to run on small molecules only. Another limitation is that currently the free energy estimates of component pseudoknotted structures used in the algorithm are not optimized. As a result, the minimum free energy prediction is often not correct. Further compounding this problem is the fact that the Rivas and Eddy algorithm only outputs the minimum free energy structure, whereas a valuable feature of tools for prediction of pseudoknot-free structures is that they also provide a list of low-energy suboptimal structures. The pknotsRG-mfe algorithm of Reeder and Giegerich (2004) is another recently developed algorithm which uses a dynamic programming approach, augmented with "canonization rules" that further restrict the class of pseudoknots handled but reduces the running time to $O(n^4)$, which turns out to be good enough to predict structure of sequences with several hundred bases. Dynamic programming algorithms now also provide suboptimal structures (Reeder and Giegerich 2004) and base-pairing probabilities (Dirks and Pierce 2003).

In contrast to the previously discussed approaches based on dynamic programming, heuristic approaches provide no guarantees of finding the minimal energy structure. However, heuristic approaches can be quite fast, thereby having the ability to handle long RNA molecules. Furthermore, they are inherently much less restricted than are dynamic programming algorithms with respect to the complexity of the underlying energy model. In addition, heuristic algorithms are not limited to sampling from a restricted subclass of secondary structures, a feature that becomes more important for longer molecules. In the past decade, there have been significant advances in the development of heuristic algorithms, leading to improvements in solving difficult computational problems in many application areas (see Hoos and Stützle 2004).

Early heuristic algorithms for the prediction of pseudoknotted structures (Abrahams et al. 1990) derived the out-

put structure by step-wise addition of stems, where the stem chosen at each step maximized the decrease in free energy of the structure. A disadvantage of this type of "greedy" approach is that once a stem is added to the structure, it is not possible to later remove the stem. To address this problem, van Batenburg et al. (1995) showed that a genetic algorithm could be promising for prediction of pseudoknotted structures. The same authors described results on a computer simulation of RNA folding pathways using a genetic algorithm for structure prediction (Gultyaev et al. 1995). Their STAR algorithm maintains a list of stems that can be added to a partially formed structure, and a stem is added with probability that depends on the free energy of the stem as well as on the free energy of the loop that is formed when the stem is added. The algorithm also includes a mechanism for removal (disruption) of stems and a crossover mechanism for producing new structures from two "parental" structures. In tests of their algorithm on 10 RNA molecules, the percentage of correctly predicted base pairs ranged from 62% to 87%.

Recently, Ruan et al. (2004) presented a heuristic algorithm called iterative loop matching (ILM) for predicting pseudoknotted RNA secondary structures. Roughly, ILM first uses a dynamic programming algorithm for prediction of pseudoknot-free secondary structures to identify a promising helix (which may contain bulge or internal loops), adds this helix to the structure, removes the bases forming this helix from the sequence, and iterates to find additional helices. Ruan and colleagues (Ruan et al. 2004) also provided a method for determining pseudoknotted secondary structures from multiple homologous sequences.

Isambert and Siggia (2000) used a computer simulation of the folding dynamics of an RNA molecule for structure prediction. In addition to providing candidate minimum free energy structures, their method can provide other useful information, such as identification of kinetically trapped states that may be on the folding pathway of the RNA molecule. We note also that both the algorithms of Gultyaev et al. (1995) and Isambert and Siggia (2000) can also simulate folding during RNA synthesis, providing further insight. The Monte Carlo (probabilistic) simulation of Isambert and Siggia is based on a model that incorporates kinetic as well as free energy principles and provides the equilibrium distribution of structures in the limit. Although this is a potential strength of their method relative to that of Gultyaev et al. (1995), the current implementation handles a limited class of structural topologies for efficiency reasons, and there is currently no analysis of how quickly the Monte Carlo simulation converges to the equilibrium distribution.

Yet another tool for RNA secondary structure prediction, the SAPSSARN software (Gaspin and Westhof 1995) allows the user to dynamically incorporate a chosen set of folding constraints and to compute a series of suboptimal saturated secondary structures satisfying all the given constraints. The approach can handle pseudoknots, but requires interaction by a knowledgeable user, in the form of constraint design.

Here we present a new heuristic algorithm, HotKnots, for prediction of RNA secondary structure, including pseudoknots, which improves on the prediction quality of previous algorithms (see Figs. 2, 3). Like other approaches, our algorithm builds up candidate secondary structures by adding substructures one at a time to partially formed structures. Unlike other approaches, our algorithm maintains multiple partially formed structures, and for each, several different additions of a single substructure are considered, resulting in a tree of candidate structures. Our criterion for determining which substructures to add to partially formed structures at successive levels of the tree is also new, relative to previous algorithms: energetically favorable substructures called "hotspots" are found by a call to Zuker's algorithm, with the constraint that no base already paired may be in the structure. The algorithm uses a standard free energy model (Serra

```
procedure HotKnots
        input: RNA sequence S
        output: list of structures SecStr(S, H_v) for all nodes v in tree T, ranked in increasing
        order of free energy

        generate an initial list L of hotspots h_1, h_2, ..., h_k ;
        create a tree T with single (root) node r and empty hotspot set;
        add k child nodes to r, with the ith child having initial hotspot set H_v = {h_i};
        build each of the k children (as described below);
        output the list of structures SecStr(S, H_v) for all nodes v in tree T, ranked in
                        increasing order of free energy;
end HotKnots


procedure build
        input: node v of tree T with hotspot set H_v
        select good hotspots that don't overlap with those in H_v and add these to list L;
        for each hotspot h in list L do
                if there is no node w of T with H_w=H_v ∪ {h}
                and SecStr(S, H_v ∪ {h}) is promising then
                   create a new node w, make it a child of v, and set H_w tp be H_v ∪ {h};
                   recursively build node w;
                end if
        end for
end build
```

FIGURE 2. Outline of RNA secondary structure prediction algorithm; details are discussed in the text.

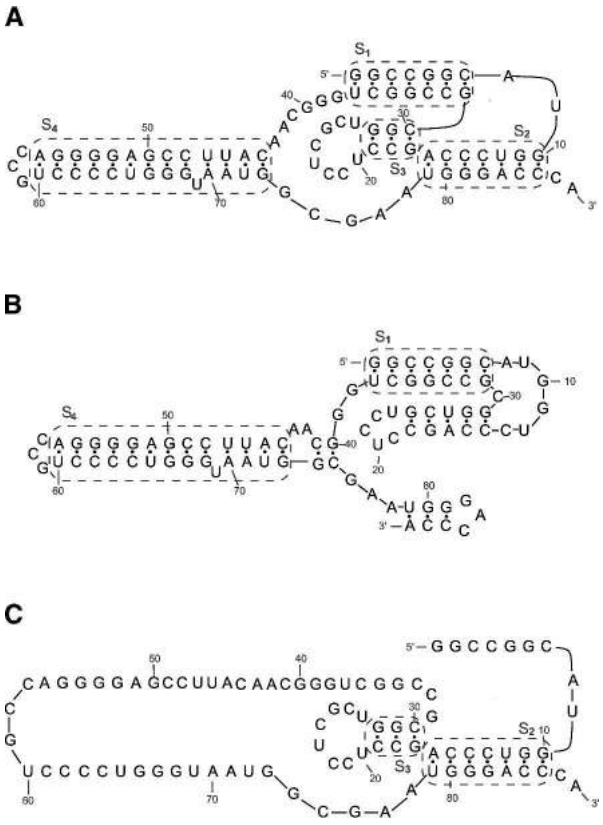

FIGURE 3. Predicting the structure of the HDV ribozyme sequence. (*A*) The known structure. (*B*) Two hotspots chosen initially, as described in the subsection "Generating the initial list of hotspots." These happen to be stems S1 and S4 of the structure in *A*. (*C*) The additional hotspots selected to be added to the structure (see subsection "Selecting good hotspots"). This is substructure S2 of *A*. No further substructures were added by the algorithm, based on our criteria described in the subsection "Selecting good hotspots." Using our method as described in the subsection "The function SecStr (S, H_v)," the set (*S1*, *S2*, *S4*) yields the secondary structure of *A*.

et al. 1995; Mathews et al. 1999), extended to account for pseudoknots (Dirks and Pierce 2003), to determine which structures at nodes of the tree have the lowest free energies, and outputs these. This energy model is also used to determine how to prune the tree of partial structures, so that more alternatives are explored from the most promising (i.e., lowest-energy) partial structures.

An empirical evaluation of HotKnots against five other algorithms for the energy-based prediction of RNA secondary structures including pseudoknots presented later in this paper indicates clearly that our new algorithm reaches state-of-the-art performance in terms of the accuracy of its predictions. In most cases, it performs significantly better than the Pseudoknots algorithm by Rivas and Eddy (1999), the NUPACK algorithm by Dirks and Pierce (2003), and the Iterated Loop Matching algorithm of Ruan et al. (2004). Despite its conceptual simplicity, HotKnots also often performs better than the genetic algorithm from the STAR package of van Batenburg et al. (1995) and the recent pknotsRG-mfe algorithm by Reeder and Giegerich (2004). In addition to establishing HotKnots as a new state-of-the-art algorithm for RNA secondary structures including pseudoknots, our empirical evaluation provides a comprehensive performance comparison of high-performance algorithms for this problem, which should be useful in the context of applications and future algorithm development.

The remainder of this paper is organized as follows. We present results on the performance of our algorithm and two earlier algorithms in the Results section. We discuss these results in the Discussion section, and conclude this section with an outlook on future work. In Materials and Methods, we give a detailed overview of our algorithm, summarize properties of sequences used in our evaluation of the algorithm, and describe our experimental protocol.

RESULTS

We evaluated our algorithm on 43 sequences or sequence fragments, including tRNA, mRNA, tmRNA, HIV-1-RT-ligand RNA, a hepatitis delta virus ribozyme, and viral ribosomal frameshifting signals (see Materials and Methods for details). Table 1 summarizes the sequences used.

Of the 43 sequences, 11 are pseudoknot-free (namely the tRNA sequences and five of the RNaseP sequences); furthermore, 31 of the sequences are relatively short, ranging in length from 28 to 108 nucleotides (nt), whereas the remaining 12 are significantly longer, with lengths ranging from 210 to 400 nt. For each sequence, we measured both the sensitivity and specificity of the lowest free-energy secondary structure predicted by our algorithm. We define the sensitivity to be the ratio of true positives (i.e., base pairs in the predicted structure which are also in the true structure) to the total number of base pairs in the true structure; intuitively, it measures the extent to which the algorithm is able to predict the base pairs that make up the true structure of a given RNA. We define the specificity to be the ratio of the true positives to the total number of base pairs in the predicted structure, which thus indicates the accuracy of the base-pair predictions made by the algorithm. Note that perfect predictions have sensitivity and specificity values of 1.

We also measured the sensitivity and specificity of five other algorithms on each sequence. Three of these are dynamic programming algorithms: Pseudoknot (pknotsRE) (Rivas and Eddy 1999), NUPACK (Dirks and Pierce 2003), and pknotsRG-mfe (Reeder and Giegerich 2004). The other two are heuristic algorithms: the ILM (Ruan et al. 2004) and the STAR algorithm (Gultyaev et al. 1995). The STAR software package implements three algorithms: greedy, stochastic, and genetic. We ran our tests using the genetic algorithm,

TABLE 1. Sequences used in our comparison of algorithms for pseudoknotted secondary structure prediction

Sequence IDs	Sequence type	Reference
DA0260, DA1280, DC0010, DC0262, DD0260, DY4441	tRNA (pseudonot-free)	Sprinzl et al. 1998
RNaseP10058, RNaseP10215, RNaseP9917 EC_RNaseP_P4, RNaseP9955	RNaseP (pseudoknot-free)	Brown 1999
Br-PrP, Ec-S15, Ec-alpha, Hs-PrP, T4-gene32	mRNA	van Batenburg et al. 2001
LP-PK1, Ec-PK1, Ec-PK4, tmRNA 10380	tmRNA	van Batenburg et al. 2001
satRPV, Tt-LSU-P3P7	ribozymes	van Batenburg et al. 2001
HIVRT32, HIVRT332, HIVRT33	HIV-1-RT ligandRNA	Turek et al. 1992
HDV	hepatitis delta virus ribozyme	Isambert and Siggia 2000
MMTV, MMTV-vpk, SRV-1, T2-gene32, BWYV, pKA-A, minimalIBV	viral ribosomal RNA frame-shifting signals	Giedroc et al. 2000
TYMV	viralRNA	Deiman et al. 1997
telo.human	telomerase RNA	Chen et al. 2000
HDV-anti	anti-genomic HDV	Ferre-D'Amareand et al. 1998
TMV.L, TMV.R VDV_IRES, CSFV_IRES, HCV_Ires	viral RNA	van Belkum et al. 1985
A.tum.RNase.P	RNaseP	Brown 1999
EC_rpml	rRNA	van Batenburg et al. 2001

In order, the columns provide (1) sequence ID, as found in the database or paper from which we obtained the sequence; (2) type of sequence; and (3) citation of the database or paper from which we obtained the sequence.

since the STAR website indicates that of the three algorithms, it yields the highest quality predictions.

Table 2, a and b, show the sensitivity and specificity of HotKnots, ILM, pknotsRE, STAR, pknotsRG-mfe, and NUPACK on our sets of 31 short and 12 long test sequences, respectively. When averaged over the set of short test sequences, HotKnots shows the same specificity and sensitivity as pknotsRG-mfe, while surpassing that of all other algorithms. On the set of longer test sequences, HotKnots shows better average specificity and sensitivity than ILM and pknotsRG-mfe, but it is surpassed by STAR. (For almost all of these sequences, pknotsRE and NUPACK failed to produce results in our experiments.) However, on five of the 12 long sequences, HotKnots shows higher sensitivity than STAR, and on four of those also higher specificity.

If we declare an algorithm to dominate on a sequence when both its sensitivity and specificity are at least as good as those of the other algorithms, then HotKnots dominates on 14 of the 31 short sequences, while ILM, pknotsRE, STAR, pknotsRG-mfe, and NUPACK dominate on 9, 13, 11, 16, and 14 sequences, respectively. (Note that more than one algorithm may dominate on the same sequence if their sensitivity and specificity match.) Furthermore, HotKnots dominates on four of the 12 longer sequences, whereas ILM, pknotsRE, STAR, pknotsRG-mfe, and NUPACK dominate on 1, 1, 4, 2, and 0 sequences, respectively.

HotKnots shows good performance on pseudoknotted as well as on pseudoknot-free structures: HotKnots dominates on 14 of the 32 sequences with pseudoknotted structures, while ILM, pknotsRE, STAR, pknotsRG-mfe, and NUPACK dominate on 9, 13, 9, 14, and 13 of these sequences, respectively. Among the 11 pseudoknot-free RNAs, HotKnots dominates on four, whereas ILM, pknotsRE, STAR, pknotsRG-mfe, and NUPACK dominate on 1, 1, 6, 4, and 1 sequences, respectively.

The overall performance advantage of STAR over HotKnots observed on longer sequences is largely due to its higher sensitivity and specificity on four of the five long, pseudoknot-free RNAs.

Figure 4 shows the sensitivity and specificity of the base-pair predictions made by HotKnots versus pknotsRG-mfe and STAR, respectively. The correlation of the performance of STAR versus HotKnots is rather weak across our test sets in terms of both sensitivity and specificity; the same is true for the performance of HotKnots versus ILM and pknotsRE (data not shown). Interestingly, there is a higher correlation between the performance measures for pknotsRG-mfe versus HotKnots, particularly on the longer sequences. This may be due to similarity of the energy models.

The running times of HotKnots, ILM, pknotsRE, pknotsRG-mfe, and NUPACK on each sequence (measured on our reference machine, see Materials and Methods) are given in Table 3. We do not include the running time of STAR since tests of that algorithm were necessarily done on a different machine and operating system, and the program requires user interaction. Roughly, we found that the running time of the STAR algorithm is comparable to that of HotKnots, taking a few seconds on short inputs and several minutes on the long sequences. ILM is significantly faster than either HotKnots or pknotsRE, but as indicated in Table 2, this comes at the cost of poorer-quality predictions. HotKnots is typically significantly faster than NUPACK (which, as previously mentioned, failed to run on the longer sequences), but, particularly for the longer

TABLE 2. Sensitivity and specificity of the predictions of six algorithms on our (a) short and (b) long test sequences

(a)

		Sensitivity						Specificity					
Sequence ID	Length	HotKnots	ILM	pknotsRE	STAR	pknotsRG-mfe	NUPACK	HotKnots	ILM	pknotsRE	STAR	pknotsRG-mfe	NUPACK
BR-PrP	45	0.41	**0.83**	0.5	0.33	0.33	0.41	0.38	**0.76**	0.5	0.26	0.26	0.38
BWYV	28	**1**	0.88	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
DA0260	75	**0.95**	0.68	0.69	0.5	0.77	0.77	0.77	0.68	0.68	0.5	0.85	**0.89**
DA1280	73	**1**	**1**	0.76	**1**	**1**	**1**	**0.95**	0.8	0.69	**0.95**	**0.95**	**0.95**
DC0010	73	**1**	0.9	**1**	0.95	**1**	0.8	**1**	**1**	**1**	0.95	**1**	0.94
DC0262	73	**0.85**	**0.85**	0.61	**0.85**	**0.85**	0.61	**0.78**	0.66	0.52	**0.78**	**0.78**	0.54
DD0260	76	0.28	**0.76**	0.33	0.47	0.28	0.33	0.28	**0.64**	0.29	0.4	0.28	0.26
DY4441	73	0.95	0.76	0.71	**1**	0.19	0.19	0.95	0.69	0.71	**1**	0.16	0.17
Ec-alpha	108	0.45	**0.66**	0.45	0.45	0.45	0.45	0.29	**0.4**	0.29	0.3	0.29	0.3
Ec-PK1	30	**1**	0.36	**1**	0.36	**1**	**1**	**1**	0.44	**1**	0.5	**1**	**1**
Ec-PK4	52	0.68	0.52	0.68	0.68	0.68	**1**	**1**	0.58	0.92	**1**	**1**	**1**
Ec-S15	67	**1**	0.58	0.94	0.58	0.76	0.88	**0.73**	0.47	0.64	0.62	0.68	0.71
HDV	87	0.4	**1**	0.46	0.6	0.96	0.63	0.44	0.88	0.46	0.7	**0.93**	0.61
HDV-anti	91	0.16	**1**	0.41	0.62	0.16	0.41	0.14	**0.66**	0.31	0.6	0.14	0.32
HIVRT32	35	**1**	**1**	**1**	0.9	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
HIVRT322	35	**1**	**1**	**1**	0.9	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
HIVRT33	35	**1**	**1**	**1**	0	**1**	0.9	**1**	**1**	**1**	0	**1**	**1**
Hs-PrP	45	0	**0.27**	0	0	0	0	0	**0.27**	0	0	0	0
LP-PKI	30	0.5	0.5	0.5	0.5	0.5	**0.8**	**1**	0.71	0.83	**1**	**1**	**1**
minimalIBV	45	**0.94**	0.88	**0.94**	0.88	**0.94**	**0.94**	0.88	0.88	**0.94**	0.93	**0.94**	**0.94**
MMTV	34	**1**	0.81	**1**	**1**	**1**	0.45	**0.91**	0.81	**0.91**	**0.91**	**0.91**	0.5
MMTV-vpk	34	**1**	0.54	**1**	**1**	**1**	**1**	0.91	0.54	0.91	0.91	0.91	**1**
pKA-A	36	**1**	**1**	**1**	**1**	**1**	**1**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**
satRPV	73	0.59	0.77	**0.81**	0.59	**0.81**	0.59	0.68	0.68	**0.85**	0.76	**0.85**	0.68
SRV-1	38	**1**	0	**1**	**1**	**1**	**1**	**0.91**	0	**0.91**	**0.91**	**0.91**	**0.91**
T2-gene32	33	**1**	0.58	**1**	**1**	**1**	**1**	**1**	0.7	**1**	**1**	**1**	**1**
T4-gene32	28	0.63	0.63	**1**	**1**	**1**	**1**	0.87	**1**	**1**	**1**	**1**	**1**
TMV.L	84	0.52	**0.8**	0.52	0.64	**0.8**	0.52	0.61	0.76	0.59	0.69	**0.83**	0.61
TMV.R	105	0.67	0.76	0.44	**0.85**	0.67	0.52	0.74	0.7	0.48	**0.96**	0.74	0.54
Tt-LSU-P3P7	65	**0.95**	0.8	0.9	0.6	0.85	**0.95**	**1**	0.69	0.85	0.75	**1**	**1**
TYMV	86	0.72	**0.88**	0.72	**0.88**	0.76	0.44	0.78	0.75	0.78	**0.88**	0.79	0.5
AVERAGE	58	**0.76**	0.74	0.75	0.71	**0.76**	0.72	**0.77**	0.71	0.74	0.74	**0.77**	0.73

(b)

		Sensitivity						Specificity					
Sequence ID	Length	HotKnots	ILM	pknotsRE	STAR	pknotsRG-mfe	NUPACK	HotKnots	ILM	pknotsRE	STAR	pknotsRG-mfe	NUPACK
A.tum.RNase.P	400	**0.77**	0.61	*	0.72	**0.77**	*	0.82	0.62	*	**0.84**	0.82	*
BVDV IRES	239	0.51	**0.85**	*	0.74	0.51	*	0.19	**0.25**	*	0.24	0.17	*
CSFV IRES	235	0.33	**0.74**	*	**0.74**	0	*	0.11	0.23	*	**0.25**	0	*
EC RNaseP p4	353	**0.75**	0.28	*	0.68	**0.75**	*	**0.22**	0.07	*	0.20	**0.22**	*
EC rpml	343	**0.56**	0.46	*	0.50	**0.56**	*	**0.16**	0.11	*	0.15	**0.16**	*
HCV Ires	210	0.36	0.68	**0.72**	0.4	0.36	*	0.11	0.18	**0.22**	0.13	0.11	*
RNaseP10058	342	0.39	0.31	*	**0.57**	0.39	*	0.37	0.27	*	**0.6**	0.39	*
RNase10215	349	0.41	0.40	*	**0.45**	0.42	*	0.38	0.34	*	**0.47**	0.38	*
RNaseP9917	316	0.51	0.60	*	**0.71**	0.51	*	0.45	0.51	*	**0.64**	0.45	*
RNaseP9955	308	0.58	0.47	*	0.63	**0.66**	*	0.56	0.41	*	**0.62**	0.59	*
telo.human	210	**0.7**	0.28	0.48	0.48	0.54	*	**0.55**	0.17	0.32	0.38	0.42	*
tmRNA10380	297	**0.46**	0.34	*	0.25	0.4	*	**0.48**	0.31	*	0.25	0.37	*
AVERAGE	300	0.52	0.50	*	**0.57**	0.48	*	0.36	0.28	*	**0.39**	0.34	*

There is one row of the table per test sequence. Starting from the left, the columns report the sequence ID, sequence length, sensitivity of the HotKnots, ILM, pknotsRE, STAR, pknotsRG-mfe, and NUPACK algorithms, respectively, and the specificity of the HotKnots, ILM, pknotsRE, STAR, pknotsRG-mfe, and NUPACK algorithms. For a given sequence, an algorithm's sensitivity value is marked in bold if the value is at least as great as the sensitivity of the other algorithms, and similarly for specificity.
"*" Indicates we were unable to run the algorithm to completion.

TABLE 3. CPU times for running five of the tested algorithms on our reference machine on our test sequences

Sequences	HotKnots	ILM	pknotsRE	pknotsRG-mfe	NUPACK
Bt-PrP	0.07 Sec	0.03 Sec	7.72 Sec	0.03 Sec	0.68 Sec
BWYV	0.05 Sec	0.001 Sec	0.27 Sec	0.02 Sec	0.06 Sec
DA0260	1.97 Sec	0.05 Sec	5m25.7 Sec	0.06 Sec	10.72 Sec
DA1280	1.4 Sec	0.02 Sec	4m28.99 Sec	0.06 Sec	9.30 Sec
DC0010	0.27 Sec	0.03 Sec	4m31.74 Sec	0.05 Sec	7.70 Sec
DC0262	1.29 Sec	0.02 Sec	6m18.54 Sec	0.06 Sec	11.32 Sec
DD0260	6.24 Sec	0.01 Sec	6m8.08 Sec	0.06 Sec	12.08 Sec
DY4441	1.29 Sec	0.01 Sec	4m36.11 Sec	0.05 Sec	8.98 Sec
Ec-alpha	0.44 Sec	0.01 Sec	71m21.865 Sec	0.15 Sec	13.45 Sec
Ec-PK1	0.04 Sec	0.01 Sec	0.49 Sec	0.02 Sec	0.07 Sec
Ec-PK4	0.07 Sec	0.02 Sec	20.93 Sec	0.03 Sec	1.47 Sec
Ec-S15	0.44 Sec	0.02 Sec	2m18.83 Sec	0.04 Sec	5.62 Sec
HDV	14.3 Sec	0.02 Sec	16m21.8 Sec	0.07 Sec	24.03 Sec
HDV-anti	5.65 Sec	0.01 Sec	21m35.89 Sec	0.08 Sec	28.47 Sec
HIVRT32	0.04 Sec	0.01 Sec	1.2 Sec	0.02 Sec	0.15 Sec
HIVRT322	0.06 Sec	0.02 Sec	1.23 Sec	0.02 Sec	0.14 Sec
HIVRT33	0.04 Sec	0.01 Sec	1.21 Sec	0.02 Sec	0.15 Sec
Hs-PrP	0.06 Sec	0.02 Sec	6.97 Sec	0.03 Sec	0.69 Sec
Lp-PKI	0.04 Sec	0.02 Sec	0.48 Sec	0.02 Sec	0.07 Sec
minimalIBV	0.07 Sec	0.01 Sec	6.89 Sec	0.03 Sec	0.65 Sec
MMTV	0.06 Sec	0.03 Sec	1.0 Sec	0.02 Sec	0.12 Sec
MMTV-vpk	0.03 Sec	0.001 Sec	0.99 Sec	0.02 Sec	0.14 Sec
PKA-A	0.06 Sec	0.01 Sec	1.47 Sec	0.02 Sec	0.19 Sec
satRPV	0.51 Sec	0.001 Sec	4m21.09 Sec	0.04 Sec	7.61 Sec
SRV-1	0.04 Sec	0.02 Sec	2.08 Sec	0.02 Sec	0.21 Sec
T2-gene32	0.04 Sec	0.02 Sec	0.82 Sec	0.03 Sec	0.12 Sec
T4-gene32	0.03 Sec	0.01 Sec	0.27 Sec	0.03 Sec	0.05 Sec
TMV.L	1.44 Sec	0.02 Sec	12m41.7 Sec	0.07 Sec	17.99 Sec
TMV.R	1m57 Sec	0.04 Sec	62m44.18 Sec	0.13 Sec	1m1.78 Sec
Tt-LSU-P3P7	0.61 Sec	0.02 Sec	1m48.02 Sec	0.04 Sec	4.52 Sec
TYMV	18.23 Sec	0.02 Sec	14m33.38 Sec	0.06 Sec	19.20 Sec
A.tum.RNase.P	18m52.05 Sec	0.07 Sec	failed to run	22.05 Sec	failed to run
BVDV IRES	6m6.14 Sec	0.13 Sec	failed to run	2.55 Sec	failed to run
CSFV IRES	6m54.56 Sec	0.13 Sec	failed to run	2.55 Sec	failed to run
EC RNaseP P4	2m31.18 Sec	0.69 Sec	failed to run	12.66 Sec	failed to run
EC rpml	6m36.89 Sec	0.58 Sec	failed to run	11.40 Sec	failed to run
HCV IRES	6m6.14 Sec	0.07 Sec	5424m0.56 Sec	1.61 Sec	failed to run
RNaseP10058	6m54.56 Sec	0.60 Sec	failed to run	11.06 Sec	failed to run
RNaseP10215	2m31.18 Sec	0.74 Sec	failed to run	12.37 Sec	failed to run
RNaseP9917	6m36.89 Sec	0.37 Sec	failed to run	7.93 Sec	failed to run
RNaseP9955	4m19.84 Sec	0.40 Sec	failed to run	6.36 Sec	failed to run
TeloHuman	3m43.8 Sec	0.06 Sec	5373 m 17.43 Sec	1.81 Sec	failed to run
tmRNA 10380	14m43.16 Sec	0.33 Sec	failed to run	6.31 Sec	failed to run

sequences, requires substantially longer run times than pknotsRG-mfe.[1]

With the sole exception of sequence TMV.R, HotKnots' running time on all short sequences is <20 CPU sec on our reference machine, and <1 CPU sec on most of the sequences. On the longer sequences, HotKnots still take <20 CPU min in all cases, and typically <10 CPU min. We note that the code for pknotsRG-mfe was obtained using a compiler that is highly optimized for dynamic programming code, and led to a 90-fold speedup over a non-optimized version of the code.

[1]For the experiments reported in this paper, we used the latest implementation of pknotsRG-mfe available at the time of this writing, which is substantially faster than the one used in Reeder and Giegerich (2004).

DISCUSSION

While HotKnots is a heuristic algorithm that is not guaranteed to find structures that are optimal with respect to the underlying energy model, it perfectly predicts the true secondary structure for seven of our 43 test sequences, whereas the other two heuristic algorithms, ILM and STAR, obtain only three and four perfect predictions, respectively. For comparison, pknotsRE, pknotsRG-mfe,

and NUPACK, which are guaranteed to find structures that are optimal with respect to their energy models, achieve eight perfect predictions, each of which was also obtained by at least one of the three heuristic algorithms. In particular, of the sequences whose secondary structure is perfectly predicted by pknotsRE and pknotsRG-mfe, only one (T4-gene32) is not perfectly predicted by HotKnots. Unlike the heuristic algorithms, pknotsRE, pknotsRG-mfe, and NUPACK support only restricted types of pseudoknots, but this does not seem to be a serious disadvantage in terms of the quality of predictions on the shorter sequences, particularly in the case of pknotsRG. No algorithm gets perfect predictions on any of the 12 long sequences, where the best sensitivity and specificity values are around 0.8.

When considering the best of the 20 lowest-energy suboptimal structures found by HotKnots on the short sequences, an average sensitivity of 0.92 and an average specificity of 0.89 are obtained. For the sequences Ec-PK4, LP-PK1, T4-gene32, and HDV-anti, the secondary structure with the second-lowest energy is in fact the real structure. Other sequences for which one of the suboptimal structures is close to the real structure are: (1) Hs-Prp, for which the suboptimal structure ranked 10th, yields sensitivity 0.76 and specificity 0.8; (2) TYMV, for which the suboptimal structure ranked fourteenth, yields sensitivity 1 and specificity 0.96; (3) satRPV, for which the suboptimal structure ranked seventh, yields sensitivity 0.81 and specificity 0.85; and (4) TMV.L, where sensitivity 0.86 and specificity 0.8 are obtained for the suboptimal structure ranked seventh. When considering the best of the 20 lowest-energy suboptimal structures found by pknotsRG on the short sequences, the sensitivity and specificity are somewhat lower than for HotKnots, at 0.82 and 0.85, respectively.

The sensitivity of the HotKnots predictions is <0.5 on only six of the 31 short sequences, and we investigated some of these sequences in more detail. For the Hs-PrP sequence, the sensitivity of HotKnots is 0, meaning that no base pair in the real structure was predicted; moreover, none of the suboptimal predicted structures contains any real base pairs either. The predicted structure is a long stem with 14 base pairs containing some small bulges and internal loops (and no multiloops or pseudoknots), whereas the real structure is pseudoknotted, with 11 base pairs forming two stems. The free energy of the predicted structure is −21.6kcal/mol; the free energy of the true structure is −6.83kcal/mol. This indicates that the energy model is misleading, in this case overly penalizing the formation of a pseudoknot. The secondary structure obtained for Hs-PRP when GU pairs are not allowed is significantly better, with a sensitivity of 0.52. For the sequence Bt-PrP, a long stem-like structure with internal loops and a large bulge is predicted by HotKnots, having free energy −20.4kcal/mol, whereas the true structure is pseudoknotted, with two stems of length 6, having free energy −12.4kcal/mol, again indicating the weakness of the energy model. In this case, the ILM algorithm is significantly better, perhaps because it favors stems that do not contain loops or bulges—a strategy that, considering ILM's overall performance, does not seem to be effective in general.

The work presented here can be extended in various directions. We believe that there is significant potential to further reduce the running time of HotKnots by using more advanced search techniques. A branch-and-bound method could be used to reduce the size of the tree that is actually generated and searched, reminiscent of an early method proposed by Papanicolaou et al. (1984). In principle, HotKnots is very flexible with respect to its underlying energy model. In particular, by replacing SimFold (the subroutine used within HotKnots for predicting pseudoknot-free structures) with a heuristic procedure for pseudoknot-free structure prediction, HotKnots can be easily modified to support energy models that are not amenable to dynamic programming algorithms. This is relevant in the sense that there is some indication that limitations of the energy model need to be overcome in order to achieve further performance improvements. Finally, in principle it is possible to combine HotKnots with covariation-based secondary structure prediction methods to achieve improved performance in cases where a small number of homologous RNA sequences are available (Hofacker et al. 2002).

MATERIALS AND METHODS

In this section we give detailed descriptions of the HotKnots algorithm and the RNA sequences used in our evaluation, as well as additional information on our computational experiments.

The HotKnots algorithm

Our algorithm is based on the premise that substructures with low energy that can form from the input sequence S are likely to be in the true structure. We focus on simple stem-like substructures that are comprised only of stacked pairs, bulge loops containing one unpaired base, and interior loops with two (opposing) unpaired bases. We call such substructures "hotspots." A set of hotspots is first computed, and each hotspot in the set is used as a basis for expanding a secondary structure for the input sequence S. The algorithm outputs the list of secondary structures corresponding to each hotspot set, sorted by the free energy value that is calculated using the Turner parameters (Serra et al. 1995; Mathews et al. 1999).

Promising sets of hotspots are built up in a tree-like fashion. That is, the algorithm builds a tree T, in which each node v has an associated set of hotspots, H_v. The size of the set of hotspots associated with a node is equal to the distance of the node from the root of the tree, with the root node having no associated

hotspots. As noted above, each set H_v is expanded into a secondary structure, which we denote by SecStr(S, H_v). The algorithm is summarized in Figure 2, and the details, such as how promising hotspots are identified, how SecStr(S, H_v) is expanded from H_v, etc. are described in the following subsections. Figure 3 illustrates how the algorithm works on the HDV ribozyme sequence.

Generating the initial list of hotspots

For each pair of positions i and j in the sequence for which $j - i > 3$, the algorithm finds the minimum energy hotspot that ends with base pair i . j, if any such hotspot exists with energy <0. This is done using a simple local alignment algorithm (Smith and Waterman 1981), in which the two sequences to be aligned are the input sequence S, ordered from 5′ to 3′, and S, ordered from 3′ to 5′, with complementary or G-U pairs being considered a "match." The parameters of the energy model of Turner et al. (Serra et al. 1995; Mathews et al. 1999) are built into the local alignment algorithm, and extra penalties are added for each bulge and interior loop added to the hotspots, in order to keep the total length of hotspots from growing too large.

We use the restriction $j - i > 3$ because the minimum number of unpaired bases in a hairpin loop is three (Mathews et al. 1999; Zuker et al. 1999), and hence the smallest distance between two bases within one base pair should be at least four. Of all the possible hotspots, the first 20 hotspots that have energy lower than −0.4 kcal/mol and have more than two base pairs are chosen for the initial list. A parameter, k, determines the number of hotspots selected. In our experiments, k was set to 20.

Selecting good hotspots

In the "build" procedure, good hotspots are selected in a manner quite different from that used for finding the initial hotspots. The method employs a dynamic programming algorithm for predicting pseudoknot-free secondary structures similar to the mfold algorithm (Zuker and Stiegler 1981; Lyngsø et al. 1999), specifically the SimFold implementation (Andronescu et al. 2005). SimFold can take as input both an RNA sequence and a set of constraints on the output structure. These constraints can include constraints that certain bases must be unpaired in the output structure, and output the pseudoknot-free secondary structure with minimum free energy, among those that satisfy the given input constraints. SimFold also uses the thermodynamic parameters of Turner and colleagues (Serra et al. 1995; Mathews et al. 1999).

To select hotspots of the sequence S that do not overlap (i.e., share bases) with those in a set of hotspots H_v, we use SimFold with input S and the constraints that every base of any hotspot in H_v must remain unpaired. Of those hotspots in the structure output by SimFold, those with free energy below −0.4 kcal/mol are selected. Note that this method for selecting the new hotspots enables the algorithm to find new hotspots that may form pseudoknots with the hotspots already in H_v. This approach is motivated by the idea that if the true structure for sequence S includes the hotspots in H_v, then any new hotspot forming a pseudoknot with those in H_v will be in the secondary structure for S only if that new hotspot has low energy.

The function SecStr(S, H_v)

Next, we describe how the secondary structure SecStr (S, H_v) associated with sequence S and hotspot set H_v is determined. Let $s_1, s_2, \ldots, s_l$ be the segments of sequence S obtained by removing the bases that are in hotspots of H_v. SimFold is used to find the minimum free energy structure for each segment s_i. Then SecStr (S, H_v) is exactly the union of these l secondary structures, plus the hotspots in H_v.

Note that SecStr (S, H_v) contains no pseudoknots other than those implied by the set of hotspots H_v, but if u is a child of v then the secondary structure SecStr (S, H_v) may include new pseudoknots.

Determining if a hotspot set is promising

If the energy of the secondary structure SecStr (S, H_v) is no more than 80% higher than the energy of SecStr (S, H_r) where r is the root node of the tree, and is at most 5kcal/mol, then the set H_v is deemed to be promising. The choices of 80% and 5 kcal/mol were made based on preliminary testing of the code. Again, the Turner parameters (Serra et al. 1995; Mathews et al. 1999) together with those of Dirks and Pierce (2003) for pseudoknotted loops are used to determine the energy of a structure. Note that since H_r is the empty set, SecStr (S, H_r) is in fact the minimum free-energy pseudoknot-free secondary structure for S, as output by SimFold. By only adding nodes to the tree for promising hotspot sets, the search space is reduced, which helps keep the algorithm efficient.

Sequence data set

Table 1 provides references for the sequences used in evaluating the algorithms.

Experimental details

The experiments for HotKnots, ILM, pknotsRE, pknotsRG-mfe, and NUPACK were run on a PC with dual 2-GHz Intel(R) XEON processors (only one of which was used for our experiments) with a 512-KB PCU cache and 4-GB RAM, running SuSE Linux version 9.1 (i586). Running times were measured using the "time" command. The STAR software was run on a similar PC under Windows; it has an interactive interface that does not support accurate measurement of running time, which made precise time comparisons with the other algorithms impossible. STAR runs in rounds; at the end of each round the user is presented with a secondary structure and may indicate whether another round of computation should be performed. In our experiments, we ran the algorithm for between 10 and 20 rounds, depending on the length of the sequence, stopping when the structure output did not change for five rounds.

ACKNOWLEDGMENTS

We thank Dr. Eke van Batenburg (Leiden University) for providing us with the STAR software package for use in our experiments.

Received December 22, 2004; accepted June 24, 2005.

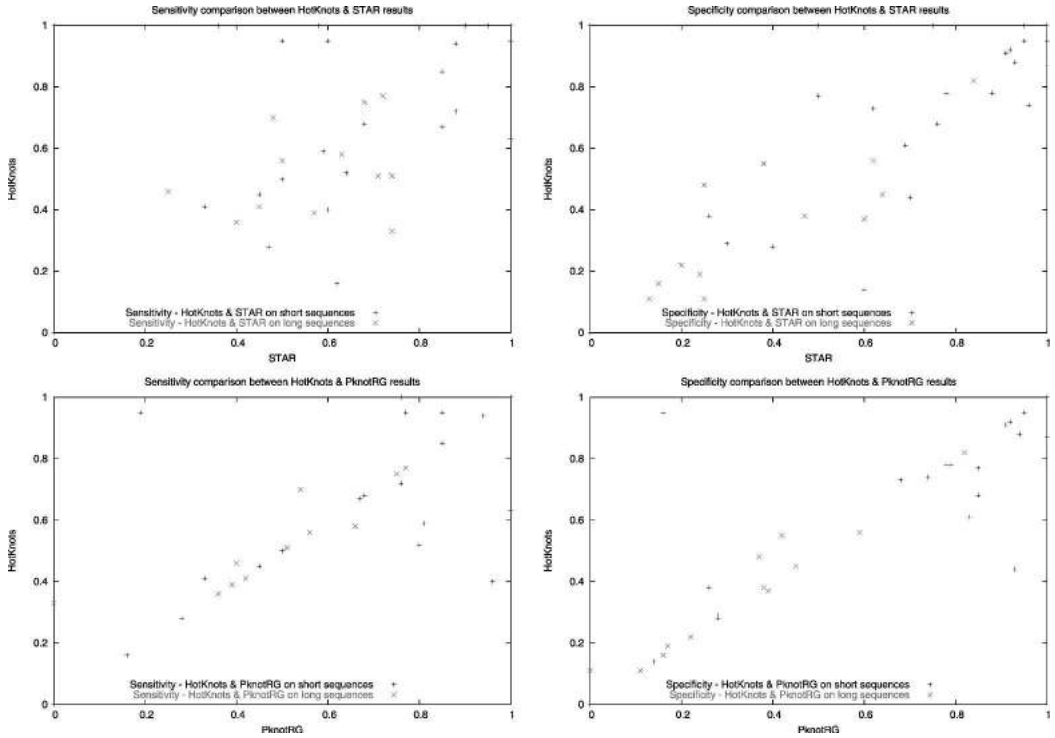

FIGURE 4. Sensitivity (*left*) and specificity (*right*) of the base-pair predictions made by HotKnots vs. STAR (*top*) and pknotsRG-mfe (*bottom*) on the sequences from our test sets. Each data point corresponds to one test sequence.

REFERENCES

Abrahams, J.P., van den Berg, M., van Batenburg, E., and Pleij, C. 1990. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Res.* **18:** 3035–3044.

Akutsu, T. 2000. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math.* **104:** 45–62.

Andronescu, M., Zhang, Z.C., and Condon, A. 2005. Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol.* **345:** 987–1001.

Brown, J.W. 1999. The ribonuclease P database. *Nucleic Acids Res.* **27:** 314.

Chen., J., Blasco, M., and Greider, C. 2000. Secondary structure of vertebrate telomerase RNA. *Cell* **100:** 503–514.

Deiman, B.A., Kortlever, R.M., and Pleij, C.W. 1997. The role of the pseudoknot at the 3′ end of turnip yellow mosaic virus RNA in minus-strand synthesis by the viral RNA-dependent RNA polymerase. *J. Virol.* **71:** 5990–5996.

Dirks, R.M. and Pierce, N.A. 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* **24:** 1664–1677.

Draper, D.E., Gluick, T.C., and Schlax, P.J. 1998. Pseudoknots, RNA folding and translational regulation. In *RNA structure and function* (eds. R.W. Simons and M. Grunberg-Manago), pp. 415–436. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Durbin R., Eddy, S.R., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press, Cambridge, United Kingdom.

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

Ferre-D'Amareand, A., Zhou, K., and Doudna, J. 1998. Crystal structure of a hepatitis delta virus ribozyme. *Nature* **395:** 567–574.

Gaspin, C. and Westhof, E. 1995. An interactive framework for RNA secondary structure prediction with dynamical treatment of constraints. *J. Mol. Biol.* **254:** 163–174.

Giedroc, D.P., Theimer, C.A., and Nixon, P.L. 2000. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.* **298:** 167–185.

Gultyaev, A.P., van Batenburg, F.H.D., and Pleij, C.W.A. 1995. The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.* **250:** 37–51.

Hofacker, I.L., Fekete, M., and Stadler, P.F. 2002. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319:** 155–173.

Hoos, H.H. and Stützle, T. 2004. *Stochastic local search: Foundations and applications.* Morgan Kaufmann, San Francisco, CA.

Isambert, H. and Siggia, E.D. 2000. Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme. *Proc. Natl. Acad. Sci.* **97:** 6515–6520.

Knudsen, B. and Hein, J. 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics.* **15(6):** 446–454.

Lyngsø, R.B. and Pedersen, C.N. 2000. RNA pseudoknot prediction in energy-based models. *J. Comp. Biol.* **7(3):** 409–427.

Lyngsø, R.B., Zuker, M., and Pedersen, C.N.S. 1999. Internal Loops in RNA secondary structure prediction. In *Proceedings of the Third International Conference in Computational Molecular Biology*, pp. 260–267. Association for Computing Machinery, New York.

Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288:** 911–940.

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

Papanicolaou, C., Gouy, M., and Ninio, J. 1984. An energy model that predicts the correct folding of both the tRNA and the 5S RNA molecules. *Nucleic Acids Res.* **12:** 31–44.

Reeder, J. and Giegerich, R. 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics* **5:** 104.

Rivas, E. and Eddy, S.R. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* **285:** 2053–2068.

Ruan, J., Stormo, G.D., and Zhang, W. 2004. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* **20:** 58–66.

Serra, M.J., Turner, D.H., and Freier, S.M. 1995. Predicting thermodynamic properties of RNA. *Meth. Enzymol.* **259:** 243–261.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., and Steinberg, S. 1998. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **26:** 148–153.

Turek, C., MacDougal, S., and Gold, L. 1992. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc. Natl Acad. Sci.* **89:** 6988–6992.

Uemura, Y., Hasegawa, A., Kobayashi, S., and Yokomori, T. 1999. Tree adjoining grammars for RNA structure prediction. *Theor. Comp. Sci.* **210:** 277–303.

van Batenburg, F.H.D., Gultyaev, A.P., and Pleij, C.W.A. 1995. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.* **174:** 269–280.

———. 2001. PseudoBase: Structural information on RNA pseudoknots. *Nucleic Acids Res.* **29:** 194–195.

van Belkum, A., Abrahams, J., Pleij, C., and Bosch, L. 1985. Five pseudoknots are present at the 204 nucleotides long 3′ noncoding region of tobacco mosaic virus RNA. *Nucleic Acids Res.* **13:** 7673–7686.

Wuchty, S., Fontana, W., Hofacker, I.L., and Schuster, P. 1998. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49:** 145–165.

Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* **244:** 48–52.

Zuker, M., and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9:** 133–148.

Zuker, M., Mathews, D.H., and Turner, D.H. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In *RNA biochemistry and biotechnology* (eds. J. Barciszewski and B.F.C. Clark), pp. 11–43. Kluwer Academic Publishers, Dordrecht, The Netherlands.

HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots

JIHONG REN, BAHARAK RASTEGARI, ANNE CONDON, et al.

RNA 2005 11: 1494-1504