Study of Accessible Motifs and RNA Folding Complexity

Ydo Wexler[1,⋆,⋆⋆], Chaya Zilberstein[1,⋆], and Michal Ziv-Ukelson[2,⋆]

of Computer Science, Technion - Israel Institute of Technology, Haifa 32000, Israel
{ywex, chaya}@cs.technion.ac.il
School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel
michaluz@post.tau.ac.il

stract. mRNA molecules are folded in the cells and therefore many of their strings may actually be inaccessible to protein and microRNA binding. The d to apply an accessability criterion to the task of genome-wide mRNA motif covery raises the challenge of overcoming the core $O(n^3)$ factor imposed by time complexity of the currently best known algorithms for RNA secondary ucture prediction [24, 25, 43].

We speed up the dynamic programming algorithms that are standard for RNA ding prediction. Our new approach significantly reduces the computations hout sacrificing the optimality of the results, yielding an expected time com- xity of $O(n^2\psi(n))$, where $\psi(n)$ is shown to be constant on average under ndard polymer folding models. Benchmark analysis confirms that in practice runtime ratio between the previous approach and the new algorithm indeed ws linearly with increasing sequence size.

The fast new RNA folding algorithm is utilized for genome-wide discovery accessible cis-regulatory motifs in data sets of ribosomal densities and decay es of *S. cerevisiae* genes and to the mining of exposed binding sites of tissue- cific microRNAs in *A. Thaliana*.

Further details, including additional figures and proofs to all lemmas, be found at: http://www.cs.tau.ac.il/∼michaluz/ adraticRNAFold.pdf

oduction

"lives" of messenger RNAs (mRNAs) begin with transcription and ultimately gradation. During their "lives", mRNAs are translated into proteins. This cess is regulated in a highly organized fashion to ensure that specific genes are at the appropriate times and levels in response to various genetic and environ- muli [11, 35]. It is well-known that mRNA decay and translation are affected ulatory motifs within mRNAs. These motifs serve as binding sites for trans- proteins and microRNAs[1]. Several cis-regulatory RNA motifs were previ- overed experimentally, such as AREs (AU-Rich Elements) [28, 40], which

A destabilizing elements involved in mRNA decay, and TOPs [13, 36], which e translation of ribosomal proteins and elongation factors.

ly, new and interesting data has become available which measures, on a vide scale, the ribosomal densities of mRNAs which reflect translation rates dditional data that measures mRNA decay rates [37]. The results of these mea-, if incorporated with genome-wide mRNA sequences, may reveal a wealth cis-regulatory elements underlying both processes. However, since RNA el- e characterized by both *primary sequence* and higher order *structural con-*he identification of RNA elements is more complicated than identification of ments. During the last decade, many computational efforts have been made to ools for the identification of RNA elements that are common to a group of ly or evolutionarily related genes. Some of these methods rely on a first step ves multiple alignment [2] and require that the sequences be highly similar with, while other methods can detect locally conserved RNA sequence and elements in a subset of unaligned sequences [16, 26]. However, the complex-e methods makes their application impractical for handling the large number ces involved in eukaryotic genome-wide analysis. Nevertheless, it turns out of the RNA regulatory motifs discovered so far are simple stem and loop with a consensus motif residing in the loop area (e.g. IRES) [13, 36].

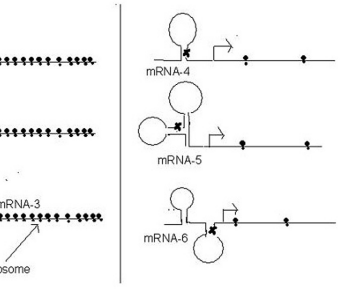
r note that the focus on *local* 2D structural conservation ignores the *global* ion of whether or not the primary sequence sites are indeed accessible to nding. In order to allow the binding between the target cis-regulatory motif ans-regulatory proteins or microRNAs, the base pairs in the motif must be free er chemical bond. This is due to the fact that the chemical recognition is based raction between amino acids residing in the protein and the corresponding es in the cis-regulatory motif residing in the mRNA [6], or on base pairing he microRNA sequence and the motif nucleotides.

ove requirement for chemical availability of motifs to protein binding calls malization of an accessability criterion:

n 1 ("accessible" substring). *Let S be a sequence and s a region i.e. sub-S. We say that s is* **accessible** iff *the following two conditions apply:*

exists a 2D structure of S with predicted free energy G_1 in which none of the otides of s is engaged in base pairing.
$G_0 \leq \delta$, where δ is a user defined threshold parameter, and G_0 is the optimal g free energy of the full string S.

per we suggest a novel approach to the genome-wide discovery of RNA cis- motifs. In our framework, motifs are scored according to their statistical ce when applying the above accessibility criterion. In order to accommodate put mRNA sequences are first filtered according to Definition 1. This is done reduce the noise created by motifs which are not exposed to trans-regulatory



plying the accessibility criterion to
ide motif discovery in mRNA ri-
lensity data. The motif X may be
to differentiate between the set of
ith high density (left) and the set
lensity (right) since its occurrences
1,4,5 and 6 are inaccessible.

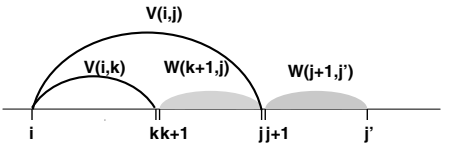


Fig. 2. The competition between candidate $V(i,k)$ and candidate $V(i,j)$ for the minimal $W(i,j')$. Candidate $V(i,k)$ has an advantage over candidate $V(i,j)$ in the additional potential cost for segment $s_{j+1} \ldots s_{j'}$ since it has a wider left-scope for combining this segment in a structure with $W(k+1,j)$. Therefore, if $V(i,k) + W(k+1,j) \leq V(i,j)$ then by triangle inequality $V(i,k) + W(k+1,j') \leq V(i,j) + W(j+1,j')$.

get, the mRNA corresponding to the gene which is targeted for "knock out", be scanned for accessible sites. For this task, the current RNA folding pre-ols are sufficient. However, such tools could not be practically scaled up to ole genome motif discovery, where thousands of mRNAs need to be mined ible sites, without raising severe efficiency problems: the complexity of RNA prediction allowing multiple loops but no pseudoknots is $O(n^3)$ to begin with, s the size of an RNA sequence (typically ~ 2000). This complexity is further to $O(n^3 \cdot m)$ by the need to exhaustively run a sliding window across the uences, where $m = O(n)$ is the number of different starting positions of ac-egions that need to be considered in each gene. Note that the sliding window onal challenge is not addressed by Robins et al. [27], where the computation ed by the fact that only a single optimal folding is computed per gene. Thus, f mining accessible sites for genome-wide motif discovery creates a heavy $m)$ bottleneck in terms of computational complexity, where g is the number n the genome under study (typically in the thousands).

actical considerations raised by such a complexity are exemplified as follows: he genome under study contains 6000 mRNA sequences, of size ~ 2000 nu-each, in which we need to consider all potential sites obtained by sliding a f size $k << 2000$. Given that the folding prediction computation for each takes about twenty seconds[2]: the total time needed for the computation of all ccessible sites in this case would be $6000 \cdot 2000 \cdot 20$ seconds ≈ 7.61 years! ote that even if we confine our search to ~ 300 windows in the UTR regions, eeded still sums up to more than a year. This example demonstrates the need nt folding algorithms, especially when dealing with whole-genome scale data.

classical $O(n^3)$ algorithms for RNA secondary structure prediction [25, 43], e been heavily used by the bioinformatics community in the last two decades, bstantially sped up? Furthermore, could such a speed up be implemented via , low-constant algorithm?

mportant challenges are addressed in the rest of this paper, where we describe amic programming algorithm that exploits the combination of two properties p RNA secondary structure prediction: one is the observed *triangle inequality* f the matrices commonly used in RNA secondary structure prediction (Sec- and the other is the *polymer-zeta* behavior of RNA folding with respect to sequence size (Section 2.4). These observations are utilized here via a simple list algorithm, called Algorithm CANDIDATEFOLD (Section 2.3), which sig- reduces the computations without sacrificing the optimality of the results (no are used). The expected time complexity of Algorithm CANDIDATEFOLD is)) instead of the previously known $O(n^3)$, where $\psi(n)$ is shown to converge ant under models previously described for RNA folding and re-validated by ations (see Section 2.5). Furthermore, due to the simplicity of Algorithm CAN- OLD, it is indeed much faster than the classical algorithm in practice, as sup- experimental performance results in Section 3. Clearly, this new algorithm for up RNA folding prediction is applicable to a wide range of additional biolog- cations, especially to those that require a substantial amount of RNA folding ons.

on the efficient new RNA folding algorithm CANDIDATEFOLD, we conducted hich examines the contribution of the "accessible site" criterion to the discov- A motifs that would otherwise be obscured by noise. The new approach was quantitative data sets of ribosomal densities and decay rates of almost all 00) *S. cerevisiae* genes. By applying our approach, some biologically inter- statistically significant motifs were discovered (Section 5). For example, the the motif $AGCKTTA$ in the decay rates data was $5 \cdot 10^{-7}$. This p-value was fact that the average half-life (i.e. $log(2)$/decay rate) of 24 genes that were contain this motif in an accessible substring was 26 days, while the half-life kground population was 15 days. Relaxing the accessibility criterion lowered cance of the motif by raising its p-value to 0.008.

o employed the "accessible target" criterion to analyze microRNAs regulating cific processes in *A. Thaliana*. Interesting tissue specific microRNAs were d (see Fig. 4).

Accessible Site Prediction Engine

iminaries of RNA Folding Prediction Via Minimum Energy

ypically produced as a single stranded molecule which then folds intra- ly to form a number of short base-paired stems. This base-paired structure is

s which are standard for RNA structure prediction do not deal with pseudo-
s is done mostly in order to simplify the problem and is justified by the fact
pseudoknots do not contribute much to the overall energy and long pseudo-
kinetically difficult to form [20]. Therefore, in this paper we assume that no
ross, however multiple loops are indeed allowed.
the above assumptions, a model was proposed in Tinoco et al. [32] to cal-
stability (in terms of free energy) of a folded RNA molecule by adding in-
contributions from base pair stacking and loop-destabilizing terms from the
structure. This model has proven to be a good approximation of the forces
RNA structure formation, thus allowing fair predictions of real structures by
ng the most stable structures in the model of a given sequence. Based on this
gorithms for computing the most stable structures have been proposed (Nussi-
acobson, 1980 [25]; Zuker and Steigler, 1981 [43]), and various tools for
ondary structure prediction were developed. The tools commonly used today
D [42], Vienna Package [14] and FOLDRNA [41].
ermodynamic parameters used by our accessible site prediction engine are
ntally derived and are identical to those used by the RNA folding tools listed
here the following four recursions are combined to model RNA secondary
folding. Note that the recursions depend on the nature of the energy rules
where $eh(i,j)$ is the energy of the hairpin loop closed by the base pair i,j,
the energy of the stacked pair i,j and $i+1,j-1$ and $ebi(i,j,i',j')$ is
of a bulge or an interior loop closed by i,j with i',j' accessible from i,j.
the boundary conditions $W(i,j) = V(i,j) = +\infty$ if $j-i < 4$. More
ecursions, based on the ones given here, take into consideration exterior base
43]. These are not elaborated here for the sake of simplicity of presentation,
he same reasoning applies to this extension as well. The recursion equations
ated below:

$$= \min\{V(i,j), W(i+1,j), W(i,j-1), \min_{i\le k<j}\{W(i,k)+W(k+1,j)\} \quad (1)$$

nputes the optimal folding of substring $s_i,\ldots,s_j$, which is the value of the
ow i and column j of the main, upper-triangular DP table W. The computa-
s table involves the matrix V whose entries are computed via the following

$$,j) = \min\{eh(i,j), es(i,j)+V(i+1,j-1), VBI(i,j), VM(i,j)\} \quad (2)$$

putes the optimal folding energy of a substring $s_i \ldots s_j$ in which s_i base pairs

$$VBI(i,j) = \min_{i<i'<j'<j}\{ebi(i,j,i',j')+V(i,j)\} \quad (3)$$

nputes the score of an optimal folding of substring $s_i,\ldots,s_j$ given that there
nal loop formed at indices (i,i',j',j).

alysis of the Classical RNA Folding Prediction Engine. The above recur-
mplemented by maintaining four tables of size $O(n^2)$ each. Eq. 1 is clearly
ven the values computed for Eq. 1, the values for Eq. 4 can be computed in
e and space via direct look-up of the minima values previously computed for
2 is also $O(n^2)$.

or the computation of internal loop size energies is naively $O(n^4)$. Practi-
standard to assume that RNA interior loop size is bounded by a constant (15
n temperature and up to 30 nt in extreme heat). The program RNAFOLD in
ckage [14] as well as the MFOLD program [42] use constant gap size in both
to reduce the complexity of Eq. 3 to $O(n^2)$. Lygnso *et. al.* [22] show how
the complexity of this equation to $O(n^3)$ without binding the gap size. On
tical front, Waterman and Smith showed how to compute internal loops in
suming that the loop penalty is a function of its size [34]. Eppstein, Galil and
[7, 9] considered loop destabilizing functions satisfying certain convexity or
conditions, and developed an $O(n^2 \log^2 n)$ algorithm for this case. This was
oved to $O(n^2 \log n)$ [1], and finally to $O(n^2\alpha(n))$ (where α is the inverse of
's function) for logarithmically growing destabilizing functions [19].

on 1. *The $O(n^3)$ bottleneck to RNA Folding Prediction complexity is based on
tation of the minimization term* $\min_{i \le k < j}\{W(i,k)+W(k+1,j)\}$ *in Eq. 1.*

the $O(n^3)$ bound applies to both the *worst case* and the *expected case* time
ies of the classical RNA folding algorithm, since Eq. 1 is called $O(n^2)$ times
all involves the computation of the minimum over $O(n)$ elements on average.

ngle Inequality in the Context of Dynamic Programming

tion we formalize the *triangle inequality* property in the context of dynamic
ing tables and show that the main matrix W, which is the final output of the
ing recursions given in the previous section, obeys this property. Let M be a
rix in which each entry $M(i,j)$ $(i \le j)$ is computed by the following formula:

$$M(i,j) = \min_{i<k\le j}\{M(i,k) + M(k+1,j)\}$$

known inverse quadrangle inequality property [10] is defined as follows.

n 2. *A matrix M obeys the* **inverse quadrangle inequality** *condition iff*

$$i < i' < j < j' \qquad M(i,j') \le M(i,j) + M(i',j') - M(j',j)$$

quadrangle and the inverse quadrangle inequalities have previously been used
p dynamic programming [5, 10]. However, both the quadrangle inequality and
e quadrangle inequality are strong constraints on the input behavior, and do not

n 3. *A matrix M obeys the* **triangle inequality** *property iff*

$$\forall\ i < j < j' \qquad M(i,j') \leq M(i,j) + M(j+1,j').$$

mple 1D Candidate List Approach to the Construction of W

$s_1 \dots s_n$ denote a given RNA sequence. The next two definitions describe olding concepts that will be used in the description of the new algorithm.

n 4 (Structure). *A* **structure** *over a sequence $s_i \dots s_j$ is a folding in which irs with s_j.*

n 5 (Partition Point). *A* **partition point** *in a given folding of $S = s_1 \dots s_n$ is , such that there is no structure over $s_i \dots s_j$ in this folding, where $1 \leq i \leq k$ $\leq n$.*

ction we describe an alternative approach to the computation of W, which . 1. Similarly to the standard algorithm, the new algorithm computes the val- row by row, in bottom-up order (decreasing row index). For each row i of W, $W(i,j)$ is computed in left-to-right order (increasing column index). How- uggested new algorithm, called CANDIDATEFOLD, differs from the original application of Eq. 1 to the computation of $W(i,j)$. In a given row i, instead ering $O(n)$ possible partition points for each column j in Eq. 1, the new al- nly considers a list of candidate partition points, which are maintained in the simple candidate list. In the following sections we show that the expected ize of this candidate list for an n-sized sequence, denoted $\psi(n)$, is constant. r to clearly define the properties that make a potential partition point a quali- date, we first need to simplify Eq. 1. Note that, if the main diagonal $W(r,r)$ zero, then the two terms $W(i+1,j)$ and $W(i,j-1)$ in Eq. 1 could be into the minimization term as special cases. $W(i+1,j)$ would then be ob- special case $k = i$ to yield the sum $W(i,i) + W(i+1,j)$ which is exactly $j)$; similarly, $W(i,j-1)$ would be obtained as the special case $k = j-1$ e sum $W(i,j-1) + W(j,j)$ which is exactly $W(i,j-1)$. However, the s that setting $W(r,r) = 0$ would contradict the boundary conditions set by Stiegler [43], which assume that $W(r,r) = \infty$.

ore, we add two auxiliary matrices, denoted W' and V', computed via the as given below, where Eq. 7 replaces the previous Eq. 1. Note that the matrix ed in order to get around the above boundary condition problem, while matrix to simplify the presentation of the algorithm which is described in the next

$$W(i,j) = W'(i,j) \ \ \forall j \geq i+4 \tag{5}$$

$$V'(i,j) = V(i,j) \ \ \forall j \geq i+4 \tag{6}$$

$$W'(i,j) = \min\{V'(i,j), \min_{i \leq k < j}\{W'(i,k) + W'(k+1,j)\}\} \tag{7}$$

he values of $W(i,j)$ and $V(i,j)$, as computed via Eqs. 2-7, are identical to ined when using Eqs. 1-4.

claim is immediate from Definition 2 and Eq. 7.

he matrix W', as computed by Eq. 7, obeys the triangle inequality.

e claim is used in the next lemma to show that any sum which yields the of Eq. 7 can be reformulated as a corresponding, equal-scoring sum, in which m is a structure (see Definition 4).

. *Consider Eq. 7. For every entry $W'(i,j)$, if there exists an index k, $i \leq k < j$, $W'(i,j) = W'(i,k)+W'(k+1,j)$, then $W'(i,k') = V'(i,k')$ for some index*

g to Lemma 1, Eq. 7 can be reformulated as follows.

$$W'(i,j) = \min\{V'(i,j), \min_{i \leq k < j}\{V'(i,k) + W'(k+1,j)\}\} \tag{8}$$

y, after the transformation to Eq. 8, there are still n candidate partition points mpete for the optimal score in the minimization term. However, the next theo- es a dominance relationship between these candidates (see Figure 2).

1. *If $V'(i,j) \geq V'(i,k) + W'(k+1,j)$ for some $i < k < j$. Then,*

$$j' > j \qquad V'(i,j) + W'(j+1,j') \geq V'(i,k) + W'(k+1,j').$$

1 exposes redundancies in the $O(n)$ computation of Eq. 8, which could be y maintaining a list of only those candidates that are not dominated by others.

n 6 (candidate). *A column index j is a* **candidate** *in a row $i \leq j$ iff $V'(i,j)$ $) + W'(k+1,j)$ $\forall\, i \leq k < j$.*

e definition can be applied to speed up the computation of $W'(i,j)$, as fol- er than considering all possible n partition point indices for the computation one could query the list that contains only partition points that satisfy the can- terion according to Definition 6. This is formalized in the following equation,

$$V'(i,j) = \min\{V'(i,j), \min_{\forall k \in candidate\_list}\{V'(i,k) + W'(k+1,j)\} \tag{9}$$

s implemented via a candidate list that is empty at the start of each row and d throughout the left-to-right computation of row i by appending only those oints which are candidates by Definition 6. Each partition point is consid- andidacy once per row, when its column is reached. The psuedo-code for the for computing Eq. 7, denoted *Algorithm* CANDIDATEFOLD, is given below.

$W'(i,j) \leftarrow \min_{\forall k \in candidate\_list}\{V'(i,k) + W'(k+1,j)\}$
if $(V'(i,j) < W'(i,j))$ *then*
$W'(i,j) \leftarrow V'(i,j)$
Append j to the $candidate\_list$

Case Time Analysis of the Improved RNA Folding Prediction Engine. denote the expected maximal size of the candidate list in a sequence of size n. CANDIDATEFOLD computes each entry in the n^2-sized energy-matrix W'. calculation requires the computation of Eq. 9, where the major work is that ting the minimum among $O(\psi(n))$ candidates. All other recursions remain d. Therefore, the overall average time complexity is $O(n^2 \cdot \psi(n))$ if the stan- d on interior loop size is followed, or otherwise $O(n^2 \cdot \max\{\psi(n), \alpha(n)\})$, $n)$ is the inverse ackerman function.

xt sections we analyze the expected growth of the candidate list size with increasing sequence size and assert the surprising fact that $\psi(n)$ converges ant. This leads to the conclusion that Algorithm CANDIDATEFOLD improves rd $O(n^3)$ classical algorithm (analyzed in section 2.1) by a linear factor on

Polymer-Zeta Property of RNA Folding

ner-zeta property is defined as follows.

n 7. *Let $P(i,j)$ denote the probability of a structure over the substring nder a given set Λ of folding rules, where $j - i = m$. We say that Λ fol-* **olymer-zeta** *property if $P(i,j) = b/m^c$ for some constants $b, c > 0$.*

work shows that RNA, which folds like other polymers, obeys the polymer- rty, namely, the probability that a structure is formed over the subsequence wo positions distant m monomers apart is $P(m) = b/m^c$ where $b = 1$ and , 18]. This fact is explained by modeling the 2D folding of a polymer chain voiding random walk (SAW) in a 2D lattice [33]. In this model the spacial f every nucleotide in the original polymer corresponds to a random step in where edges of the lattice represent possible transition directions. Since this polymer folding also ignores pseudoknots, the walk is called "self avoiding", umption is followed that the walk does not intersect the prefix of the chain. The nterest here is the probability that the m^{th} step in the self avoiding random upies the same vertex in the lattice as the origin. The theoretical exponent o dimensional SAW model is known to be $c = 1.5$ [8]. This is supported in y simulations for collapsing polymers of sequence size up to 3200, as reported ese simulations exhibited an exponent of 1.375 at low temperatures and 1.571 emperatures.

single structure formation probabilities in polymer folding, which were found e polymer-zeta property. We used $50,000$ mRNA sequences with an average 1992 nucleotides from the NCBI databases and found that the probability that al folding forms a structure over $s_i \dots s_j$, where $m = j - i$, is estimated to $m^{-1.47}$. The degree exponent c was estimated in our study to be ~ 1.47 by standard statistical procedures (approximating the MLE parameter followed g "Kolmogorov-Smirnov" and "chi-square" goodness-of-fit tests, using the R analysis package, *http://www.r-project.org*).

nds on $\psi(n)$

nalyze $\psi(n)$ based on our findings. The following observation is immediate ma 1.

ion 1. *A new candidate j is added to the candidate list, in step 6 of Algorithm* TEFOLD*, iff the optimal predicted folding of substring $s_i \dots s_j$ forms a single from index i to index j. The only exception to this case is the boundary candidate i, which is always added as a "virtual" structure to the list.*

t the probability for a new candidate situated m bases away from the start of nce is $b \cdot m^{-c}$, the expected number of candidates in a sequence of length n $b \sum_{i=1}^{n} i^{-c}$. This summation could assume one of three values, according to ted c:

alues $c \geq 1$ this series is a partial sum of the *Riemann Zeta function* defined $\sum_{i=1}^{\infty} i^{-c}$.

$c > 1$, this series is known to converge and thus, $\psi(n) = O(1)$.

$c = 1$, we get a partial sum of the first n elements of the Harmonic series, hich is known to be less or equal to $1 + \ln(n)$ and thus $\psi(n) = O(\log n)$.

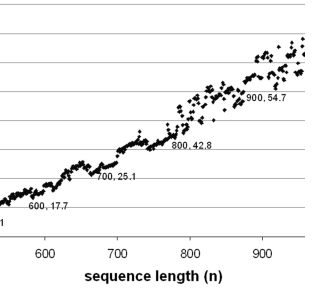
1, we use the power means inequality to obtain the bound $\psi(n) = O(n^{1-c})^c)$.

2. *Applying Algorithm* CANDIDATEFOLD *to the folding of a polymer chain hat obeys the polymer-zeta property with $c > 1$, requires an average of $O(n^2)$ s.*

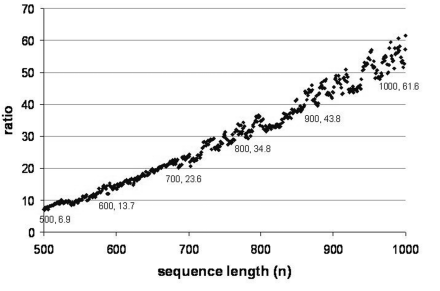
t our simulations estimate c to be 1.47, which implies that $\psi(n) \sim 2.11 \cdot$ 7, which is a constant. Therefore, applying Algorithm CANDIDATEFOLD to g of an RNA sequence of size n takes $O(n^2)$ time on average.

Performance of the New RNA Folding Engine

strate the power of algorithm CANDIDATEFOLD in practice we ran it against



) Random mRNA subsequences



(b) Simulated RNA sequences

average measured run-time ratio of naive/CANDIDATEFOLD as a function of increas-
ce size

demonstrates that the average run time ratio (computed by dividing the run the classical algorithm with ours) is linear in the sequence length n, re- g our time complexity analysis. In Figure 3(a), the analysis was done for 100 for each possible size in the range 500-1000, which were extracted as ran- osen subsequences from 50,000 complete mRNA sequences taken from NCBI The analysis shown in Figure 3(b) was done for 100 sequences of each size he range, which were generated using a Markov-model imitating software. ence-simulation program takes a set of sequences to imitate and a Markovian nput, and generates an output of random sequences according to a Markov- the desired order. The input consisted of 50,000 complete mRNA sequences ed from the NCBI database and the Markovian order parameter was set to 6. results emerged when using the remaining 50,000 mRNA sequences as input order Markovian model simulator.

hods for Mining Accessible Cis and Trans Regulatory Motifs

od for discovering novel cis-regulatory motifs incorporates large scale decay bosomal density measurements, combined with the information from mRNA of the genome under study. It can be formulated as follows. Given a set of $= S_1 \dots S_g$, a parameter k denoting motif window size (could be slightly n the motif residing in the window), and a pre-defined energy threshold δ, we following simple two-stage approach:

Process the sequence set G to extract all "accessible" windows by running a ndow of size k across the mRNA sequence and testing each window for com- th Definition 1. For each shifted window this testing is conducted by masking otides inside the window in order to prevent their engagement in base pairing.

This stage takes as input the accessible substrings, extracted in the first stage, regulatory motifs residing in the data. Two statistical techniques are applied nding on whether the sought motif is cis or trans regulatory:

atory motifs: Enumerate all motifs up to a given size k over the IUPAC al-3]. For each motif use the new data created in stage 1 instead of the original equences, to compute a t-score [12] reflecting the functionality of that motif. alue associated with the computed t-score is small enough, report the motif. e can be efficiently executed by using a variation of the algorithm of Sagot combined with the statistical computation of the t-score [38] and adapted to new "accessible window" data.

ulatory Signals (microRNAs). The search for microRNAs is similar to that of overy, except for the following difference: instead of considering accessible otifs, we considered accessible sites that were predicted to hybridize well with t microRNAs, as described in [39].

ological Study of Accessible Regulatory RNA Elements

cted a study in order to test our novel approach, which applies the "acces-riterion to RNA motif discovery. Using various data sets, significant motifs overed, including some cis-regulatory degradation and translation motifs and cific microRNAs.

of the conducted experiments, two data sets were studied: a set containing essible" substrings, according to Definition 1, and a "control" set which in-original complete mRNA sequences. A comparison of the results obtained f the two sets repeatedly confirms the contribution of the "accessibility" cri-powerful filter for masking out noise associated with inaccessible motifs and significance score of otherwise invisible motifs.

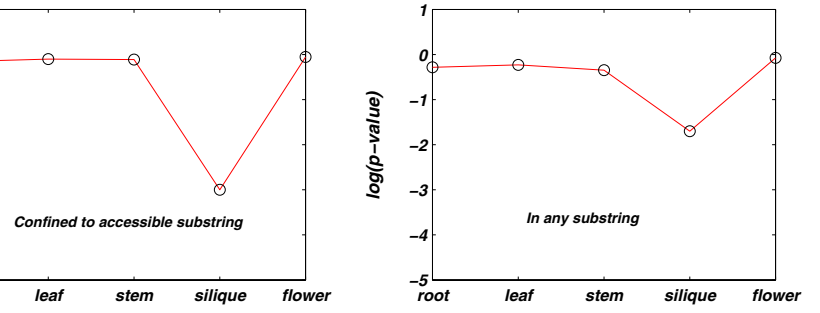
on Related Motifs. Arava *et al.* [3] measured the ribosomal densities of al-he mRNAs of the yeast *S. cerevisiae* under normal cell conditions, using the method. First, mRNAs are extracted from the cells and separated by veloc-entation. Then, each fraction across the gradient is analyzed by microarray s for its mRNA content. Based on this, a fraction is assigned to each mRNA: this fraction is, the higher the mRNA's ribosomal density is. We applied our to this data in order to detect translation cis-regulatory elements within 5' ed region (5'UTR)[3]. A few novel potential cis-regulatory elements were dis-hat may affect translational efficiency (see Table 1). In particular, the average density of the set of mRNAs containing the motif $AGSNNK$ in accessible was low in comparison to the background. Thus, $AGSNNK$ seems to be a repressor.

otifs potentially regulating mRNA translations. The accessible substring criterion was h window size 10 and $\delta = 2Kcal$. The average ribosomal density without the motif ted based on ~ 5000 different genes.

	Number of occurrences	*Average density with the motif*	*Average density without the motif*	*p-value confined to accessible substrings*	*p-value in any substring*	*Hypothesized function*
T	14	1.7	0.7	10^{-18}	10^{-4}	Translation enhancer
K	1292	0.6	0.7	10^{-11}	10^{-3}	Translation repressor

otifs potentially regulating mRNA degradations. The first 3 columns refer to the case le substring with window size 10 and $\delta = 2Kcal$. The average half life without the computed based on ~ 5000 different genes.

	Number of occurrences	*Average half-life with the motif*	*Average half-life without the motif*	*p-value confined to accessible substrings*	*p-value in any substring*	*Hypothesized function*
A	24	26.54	15.46	$4.83 \cdot 10^{-7}$	0.0083	Stabilizer
R	5	57.75	15.5	$2.76 \cdot 10^{-9}$	0.0081	Stabilizer
T	4	42.75	15.49	$4.84 \cdot 10^{-7}$	0.01198	Stabilizer



-161 and it's p-values in different plant tissues. The accessible substring criterion was h window size 25 and $\delta = 6Kcal$.

egulating elements within 3' UTRs[4]. We successfully identified some novel cis-regulatory motifs that may affect mRNA stability (see Table 2). For ex- e average half-lives (i.e. $log(2)$/Decay rate) of the set of mRNAs containing C motif $AGCKTTA$ in accessible substrings was high in comparison to the nd. Thus, $AGCKTTA$ seems to be a strong mRNA stabilizer. Table 2 also tes that, when relieving the accessibility criterion, the significance of the p- stantially dropped.

ecific microRNAs. In order discover microRNAs, which are potential trans- fluencing mRNA stabilities, we collected the genome-wide expression

ere discovered[5]. These microRNAs showed a significant p-value for binding he tissues and non-significant p-values in the rest of the tissues. For example, RNA miR-161, represented in Figure 4, is specific to silique tissue. Interest- figure demonstrates that in most of the tissues the $p - values$ corresponding t (accessible substring) and second (control) input sets are almost similar. in the silique tissue, where the microRNA miR-161 seems to be active, the between the two input sets becomes conspicuous.

edgments. Many thanks to Yoav Arava for inspiration and data, as well as l discussions. We thank Micheal Zuker for very helpful advice. The authors rateful to Ron Shamir, Ron Y. Pinter, Dan Geiger, Zohar Yakhini, Jeannette Christos Faloutsos, Eleazar Eskin and Firas Swidan for helpful discussions nents. The research of Michal Ziv-Ukelson was supported in part by the Aly Post Doctoral Fellowship.

ces

garawal and J. Park. Notes on searching in multidimensional monotone arrays. *Proc. EEE Symp. on Foundations of Computer Science*, 497–512, 1988.
naev, S. Kelley, and G. Stormo. A phylogenetic approach to RNA structure prediction. *nt Conf Intell Syst Mol Biol*, 235:10–17, 1999.
va, Y. Wang, J. Storey, C. Liu, P. Brown, and D. Herschlag. Genome-wide analysis of translation profiles in saccharomyces cerevisiae. *PNAS*, 100:3889–3894, 2003.
istofferson et al. Application of computational technologies to ribozyme biotechnol- oducts. *J.Molecular Struct.(Theochem)*, 311:273, 1994.
chemore, G. Landau, B. Schieber, and M. Ziv-Ukelson. *Re-Use Dynamic Program- or Sequence Alignment:An Algorithmic Toolkit. String Algorithmics*. KCL Press, 2005.
per. Themes in RNA-protein recognition. *J Mol Biol*, 293(2):255–270, 1999.
stein, Z. Galil, and R. Giancarlo. Speeding up dynamic programming. *Proc. 29th Symp. on Foundations of Computer Science*, 488–496, 1988.
ner. Shape of a self-avoiding walk or polymer chain. *JCP*, 44:616–622, 1966.
l and R. Giancarlo. Speeding up dynamic programming with applications to molecular y. *Theoretical Computer Science*, 64:107–118, 1989.
ncarlo. *Dynamic Programming: Special Cases.* In Pattern Matching Algorithms, A. olico and Z. Galil eds., Oxford University Press, 1997.
dwin, P. Okkema, T. C. Evans, and J. Kimble. Translational regulation of tra-2 by its anslated region controls sexual identity in c. elegans. *Cell*, 75:329–339, 1993.
lden. *Methods of Statistical Analysis*. New York: Wiley, 2 edition, 1956.
y and M. Wickens. *Annu Rev Cell Dev Biol*, 14:399–458, 1998.
ofacker. Vienna RNA secondary structure server. *NAR*, (13):3429–3431, 2003.
araman and S.P.Walton. Rational selection and quantitative evaluation of antisense ucleotides. *Biochim.Biophys. Acta*, 1520:105, 2001.
. Xu, and G. Stormo. *Bioinformatics*, 20:1591–1602, 2004.
akcioglu and A. Stella. A scale-free network hidden in the collapsing polymer. *ArXiv ased Matter e-prints*, Sept. 2004.

ri, D. Mukamel, and L. Peliti. Why is the dna denaturation transition first order? *al Review Letters*, 85:4988–4991, 2000.
more and B. Schieber. On-line dynamic programming with applications to the predic- RNA secondary structure. *J. Algorithms*, 12(3):490–515, 1991.
and R. Bundschuh. Quantification of the differences between quenched and annealed ing for RNA secondary structures. *ArXiv Physics e-prints*, Apr. 2005.
ve et al. Cleavage of scarecrow-like mRNA targets directed by a class of arabidopsis A. *Science*, 297:2053–2056, 2002.
Lyngsø, M. Zuker, and C. N. S. Pedersen. An improved algorithm for RNA secondary re prediction. Technical Report RS-99-15, brics, 1999.
thews et al. *RNA*, 5, 1458-1469, 1999.
thews, J. Sabina, M. Zuker, and D. Turner. *JMB*, 288:911, 1999.
sinov and A. Jacobson. Fast algorithm for predicting the secondary structure of single- d RNA. *PNAS*, 77(11):6309–6313, 1980.
esi et al. an algorithm for finding conserved secondary structure motifs in unaligned equences. *NAR*, 32:3258–3269, 2004.
et al. *PNAS*, 102:4006–4009, 2005.
. mRNA stability in mammalian cells. *Microbiol Rev*, 59(3):423–450, 1995.
got. Spelling approximate or repeated motifs using a suffix tree. *LNCS*,111-127, 1998.
th et al. *Eur. J. Pharm. Sci.*, 11:191, 2000.
g et al. framework for RNA silencing in plants. *Genes Dev*, 17:49–63, 2003.
co et al. *Nature New Biology*, 246:40–41, 1973.
nderzande. *Lattice Models of Polymers (Cambridge Lecture Notes in Physics 11)*. idge University Press, 1998.
terman and T. Smith. Rapid dynamic programming algorithms for RNA secondary re. *Adv. Appl. Math.*, 7:455–464, 1986.
lsh, N. Scherberg, R. Gilmore, and D. Steiner. Translational control of insulin biosyn- *Biochem. J.*, 235:459–467, 1986.
kie, K. Dickson, and N. Gray. Regulation of mRNA translation by 5'- and 3'-utr- g factors. *Trends Biochem Sci*, 28:182–188, 2003.
g et al. Decay rates of human mRNAs: correlation with functional characteristics and ce attributes. *Genome Res*, 13:1863–1872, 2003.
erstein, E. Eskin, and Z. Yakhini. Sequence motifs in ranked expression data. In *The ECOMB Satellite Workshop on Regulatory Genomics*, 2004.
erstein, M. Ziv-Ukelson, R. Y. Pinter, and Z. Yakhini. A high-throughput approach ociating microRNAs with their activity conditions. In *RECOMB*, 133–151, 2005.
iaga, J. Belasco, and M. Greenberg. The nonamer uuauuuauu is the key au-rich se- motif that mediates mRNA degradation. *Mol.Cell.Biol.*, 15:2219–2230, 1995.
ker. Computer prediction of RNA structure. *Methods Enzymol.*, 180:262–288, 1989.
ker. *NAR*, (13):3406–15, 2003.
ker and P. Stiegler. Optimal computer folding of large RNA sequences using thermo- ics and auxiliary information. *NAR*, 9(1):133–148, 1981.