RNA SECONDARY STRUCTURE PREDICTION USING APPLICATIONS OF THE PARTITION FUNCTION

by

Timothy Wake

Boston College

2003

BOSTON COLLEGE

ABSTRACT

RNA SECONDARY STRUCTURE
PREDICTION USING
APPLICATIONS OF TH E
PARTITION FUNCTION

by Timothy Wake

Advisor: Professor Peter Clote
Department of Computer Science

A dynamic programming algorithm is presented for calculating the partition function and the pairwise base-pairing probabilities over all secondary structures for a given RNA nucleotide sequence, and the calculation of the pairwise base-pairing probabilities; the algorithm is an application of the approach used by McCaskill to accomplish this for nested secondary structures to the class of structures inclusive of pseudo-knots, using a technique due to Eddy et. al.

TABLE OF CONTENTS

Introduction

RNA is a single stranded nucleotide sequence that can form Watson-Crick (A-U, C-G) and the weaker G-U hydrogen bond-pairs with itself, thus forming a complicated secondary and tertiary structure, consisting of certain well-defined substructures, such as hairpin loops, stacked base pairs, bulges and interior loops, multi-loops, and pseudo-knots.

Conventionally, nested secondary structure prediction is the standard, where in a sequence of nucleotides each base pairs with at most one other base, and the overall structures correspond to matched parentheses structures; hence if i , j base-pair, and i<k<j , then if k, ℓ base-pair, then i< ℓ <j. The cases where ℓ <i or ℓ >j are pseudo-knots and where nested structures can be represented in a linear fashion with a single set of parentheses, e.g. (. . (. .) . .), a structure containing a pseudo-knot may require an arbitrary number of symbols to represent, as an example (. . { . .) . . }.

RNA secondary structure prediction is a computationally feasible and broadly studied problem, with a number of approaches available in the literature. It is a problem of interest, as RNA performs certain catalytic functions and triggers retranslation events which depend on its complex three dimensional structure, which is to a large extent determined by the secondary structure of the RNA.

The types of algorithms that attempt to recognize or predict RNA secondary structure are of two broad classes: those that apply knowledge of the structures of other similar sequences to predict probable correspondence, and those that predict the folding of the sequence solely on the basis of the thermodynamic contributions of its substructures; the most common of which predict the single secondary structure that has the optimal (minimum) free energy of all the possible conformations the sequence may take on.

In the next section, we examine an algorithm of the former type that was created over the course of the thesis, as well as one of the applications with which we attempted to ascertain its usefulness. We examine its shortcomings and assess the representation decisions that made them necessary.

In the remainder of the thesis, we present an algorithm of significantly greater complexity, based upon evaluation of the thermodynamics of the ensemble of possible secondary structures. The great majority of the algorithms dealing with secondary structure that are based on structure thermodynamics examine only nested structures; that the region interior to a given base-pair interacts solely with itself and nothing exterior to it lends itself to reasonable and straightforward dynamic programming algorithms.

The nested model, however, is an oversimplification of the secondary structure model; there are instances where pseudo-knotted structures are known to occur, and perform biologically significant functions. Additionally, optimal prediction algorithms such as Zuker's or Eddy-Rivas predict only the individual structure whose energy is the minimum. RNA secondary structure is known to be dynamic over the regions of lowest energy. Our technique, which was originally applied by McCaskill, accounts for this by giving base-pairing probabilities for all structures derived from the statistical mechanical model, which utilizes the Boltzmann probability distribution, thus encompassing an ensemble of likely structures. Our initial application of these base-pairing probabilities is via a maximum weight matching of bases, accounting only for those base-pairing probabilities above a given threshold.

Profile PatScan:
A filter based on PSSMs and 0^{th} and 1^{st} order Markov Models

Expanding upon a computational screening approach used to identify likely SECIS elements by both Kryukov and Lescure, the algorithm consists of a several step process. Given a quantity of known positive examples of a given secondary structure, we are interested in rapidly searching full genomes, applying what amounts to a 'filter', which loosely constrains the resulting possibilities. Having done so, we can reasonably expect to have a set of data which contains the majority of true unknowns within the data examined. The resulting set is ranked according to its similarity to known positives, by applying PSSMs , relative frequencies and 1^{st} order Markov models, according to the constraints imposed by the initial filter.

It was found, however, that this approach failed to capture the necessary information to correctly distinguish unknown positives from negatives, likely due to its failure to capture dependencies between the scored regions. Approaches for capturing covariation and mutual information are frequently effective at resolving these dependencies, as they depend on the pair-wise relative frequencies of all nucleotides within a sequence, or the way that dependencies are displayed by mutual pointwise changes in sequences in the known examples. The difficulties in these approaches is that the motif in question is likely to contain some regions of variable length, which must be fit to the fixed length PSSM developed in a sensible fashion, by using sequence alignment methods or other means, lest variability in inter-sequence length cause incorrect predictions.

SECIS elements are a conserved stem-loop structure present in eukaryotic RNA that codes for selenocysteine. Selenocysteine is incorporated by a retranslation event, mapping an in-frame UGA codon to selenocysteine in the translational process. To prevent the UGA codon from being interpreted as a stop codon, as is normally the case, requires the presence of a selenocysteine insertion sequence (SECIS element) residing in the downstream untranslated portion of the mRNA.

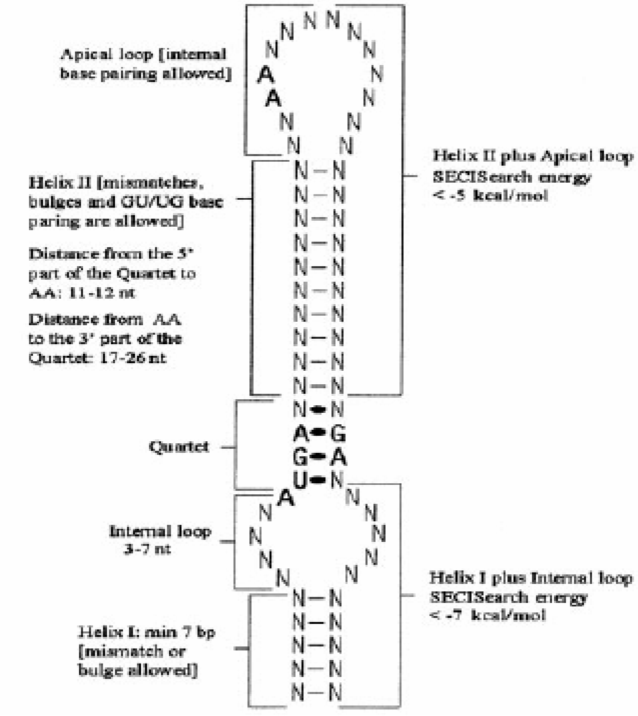
Apical loop [internal base pairing allowed]
Helix II [mismatches, bulges and GU/UG base paring are allowed]
Distance from the 5' part of the Quartet to AA: 11-12 nt
Distance from AA to the 3' part of the Quartet: 17-26 nt
Quartet
Internal loop 3-7 nt
Helix I: min 7 bp [mismatch or bulge allowed]
Helix II plus Apical loop SECISearch energy < -5 kcal/mol
Helix I plus Internal loop SECISearch energy < -7 kcal/mol
A
A
A
G
G
A
U

FIG 1: The conserved SECIS motif from Kryukov

The initial filtering process uses a regular-expression based filter called PatScan, written by Ross Overbeek and similar in function to the tools used by Kryukov and Lescure in the initial phases of their screening techniques. It allows the following constraints to be enforced:

1. Specific sub-expressions may be of variable or fixed length
2. A sub-expression may have a specific nucleotide sequence, or may be a nucleotide drawn from a set of such
3. A sub-expression may be constrained such that it must be possible to base-pair, helically, with another sub-region given a set of base-pairing possibilities (Watson-Crick and potentially G-U)
4. The constraints in (3) may be loosened by allowing specific gaps or mismatches in the helical regions

The designations for variable, fixed-length regions, helices, and invariant sequences imposed in constructing a profile for PatScan require a multiple sequence alignment of the input sequences whose correspondence to each of these basic types (invariant regions, fixed length regions, and variable length regions) allows the results to be examined differently. Where multiple examples of a sought secondary structure exist, and where there is a sensible multiple alignment to correspond to this structure, then PatScan should meaningfully filter out those sequences that cannot reasonably correspond to the consensus secondary structure because they fail to allow for the required helical hydrogen bonds between nucleotides.

Our approach was a refinement on other computational screens such as Lescure's that loosely screened data and then examined them with successive levels of refinement to garner results of increasingly likely correspondence. As an example, an additional layer of refinement in Lescure was to apply a minimum free energy algorithm to assess the comparative energetic stability of the possible hits to that of known SECIS elements, while Kryukov used open reading frame analysis to search for upstream UGA codons.

The method utilized position-specific scoring matrices (PSSM) to measure similarity in fixed length regions. PSSMs are maximum likelihood estimators for the known examples which are used to construct them. They are derived from pseudocounts of the relative frequencies of nucleotide occurrence in a specific position, and hence are the probabilistic model most likely to generate the sequences from which they are derived. For the variable length regions, a first order markov model was constructed from the input sequences, and this was used to judge with what probability a sequence was generated by that model. Invariant sequences, which are required to be conserved according to the restriction imposed by the initial filter, were not measured for the purpose of ranking the hits that passed the screen.

It became apparent with use of this application, however, that the models constructed from the known SECIS elements were insufficiently constrictive. Application of the Vienna-RNA package to the results judged most likely upon running the screening technique upon various EST databases rendered results that failed to correspond in both predicted secondary structure and thermodynamic stability to the stem-loop structure for which we were scanning.

The failure of the ranking algorithm is at heart a conceptual one; the approach breaks up the structure into regions that are scored independently, and doing so prevents proper assessment of covariation between individual nucleotides, which is known to occur when random mutation alters bases but the pairing within the secondary structure is conserved. The only way in which the algorithm measures covariation is in the initial restriction of the regular-expression based filter, that constrains helical base-pairing. Additionally, the imprecision is introduced into the scoring of the variable length regions by making it dependent upon solely its 0th and 1st order compositional frequencies; the PSSMs have the advantage that relative position in the structure is maintained, while the variable length regions may unpredictably interact with other portions of the global structure.

Minimum Free Energy Prediction: Nussinov and Zuker

In the absence of a large body of known examples of a given secondary structure motif, the only available technique for prediction is to infer structure based on the thermodynamic contributions of possible structures. The Nussinov-Jacobson and Zuker-Sankoff minimum free energy prediction algorithms predict the secondary structure of a sequence to be the conformation that has the minimum free energy according to thermodynamic parameters approximated using experimental measurement. This is not entirely accurate, though: the folded conformation of an RNA sequence exists in flux around the region of minimum energy, and thus does not necessarily correspond to the optimal structure.

Of the two, Nussinov-Jacobson is the simpler: It uses a dynamic programming algorithm to maximize the number of base-pairs in the output structure. This aim does not reflect current understanding of thermodynamic contributions within structures; outermost base-pairs have no inherent stabilizing contribution, which is dependent upon the conformation interior to the base pair. As classified below, stacked base pairs and unpaired bases exterior and interior to base-pairs tend to have stabilizing contributions, while interior loops, bulges, hairpins, and multi-loops have destabilizing contributions. Still, the maximization of base-pairs serves as a useful introduction to the dynamic programming techniques of the more sophisticated algorithms. The recursions are as follows:

Where S is the sequence of n nucleotides let S= $S_{0,}\, S_1 \ldots S_{n-1}$
Let bp(i,j) =0 if S_i and S_j cannot basepair, and 1 otherwise.
Let P is the contribution of a single base-pair and let Q be the contribution of an unpaired base.
Where i<=j let wx(i,j) be the minimum free energy of the sequence from S_i to S_j and wx(i,i)=0
Then wx(i,j)= min{ bp(i,j) * (wx(i+1, j-1) + P) ,
wx(i, j-1) + Q , wx(i+1, j) +Q,
min of for i<k<j-1 bp(k+1, j) * (wx(i,k) + wx(k+1,j) + P) }

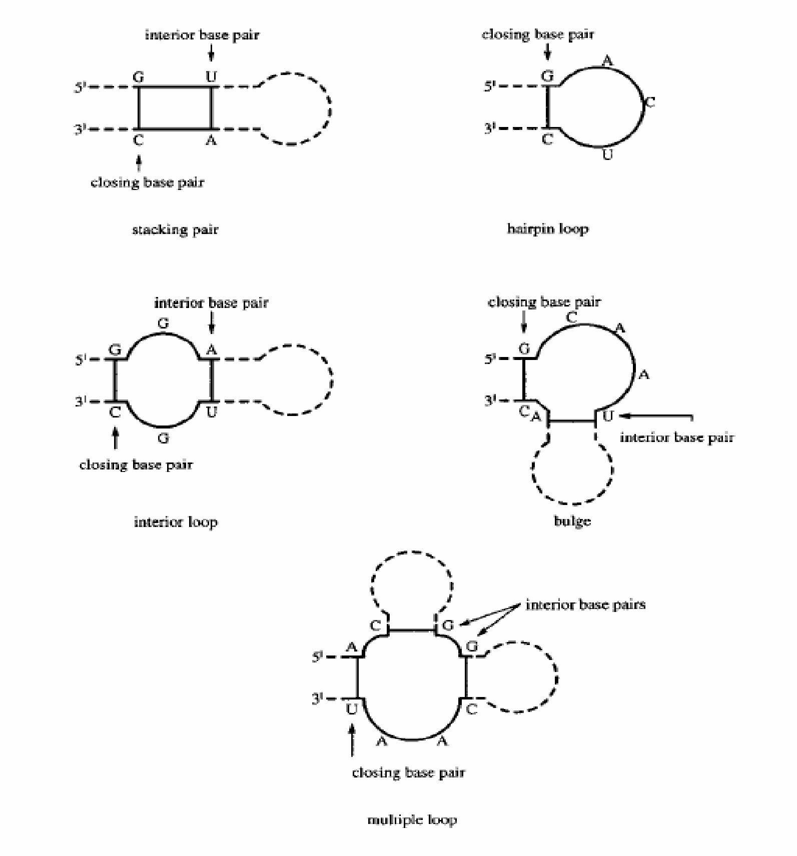

FIG 2: Energetically contributing substructures, from Wuchty et al.

In the recursion, the first case corresponds to the jth base bonding with the ith, the next two cases are dangles, and the last case accounts for when the jth base bonds with some base between i and j.

Revising Nussinov Jacobson to account for the actual thermodynamic contributions of substructures requires several cases:
Hairpin loops interior to a base-pair i,j; bulges, stacked base-pairs and interior loops which are determined by two base-pairs, as seen in figure 2; and multi-loops which are enclosed by an outermost base pair and have several independent

substructures contained by it. In all cases, the outermost base-pair does not have an energy contribution.

Hence Zuker-Sankoff is:

$$vx(i,j) = \text{optimal}\left\{\begin{array}{lll} \mathrm{EIS}^1(i,j) &] & \mathrm{IS}^1 \\ \mathrm{EIS}^2(i,j:k,l) + vx(k,l) &] & \mathrm{IS}^2 \\ P_I + M + wx_I(i+1,k) + wx_I(k+1,j-1) &] & \text{multiloop} \end{array}\right.$$

$$[\forall k, l \quad i \leqslant k \leqslant l \leqslant j]$$

$$wx(i,j) = \text{optimal}\left\{\begin{array}{lll} P + vx(i,j) &] & \text{paired} \\ \begin{array}{l} Q + wx(i+1,j) \\ Q + wx(i,j-1) \end{array} & \Big] & \text{single-stranded} \\ wx(i,k) + wx(k+1,j) \quad [\forall k, \quad i \leqslant k \leqslant j]. &] & \text{bifurcation} \end{array}\right.$$

vx(i,j) is defined to be the energy of the secondary structures from i to j given that i, j base-pair. EIS^1 corresponds to Hairpin loops, and EIS^2 to those structures dependent on two base-pairs for their determination. M is a penalty for initiating a multi-loop.

This algorithm is the standard one, implemented in both mfold and the Vienna RNA package. Its recursions are designed around the nested convention, and as previously noted, predict only the single structure considered to be optimal out of the ensemble of conformations around the region of minimum free energy.

McCaskill's algorithm for Nested Structures

McCaskill's algorithm has a distinct advantage over the minimum free energy structure prediction algorithms, in that it captures the entire ensemble of secondary structures, rather than restricting its output to a single optimal structure. This probabilistic ensemble is normally output in terms of the probabilities of individual base-pairs, and this functionality is implemented for the standard nested model as described by Zuker and Sankoff in the Vienna RNA package.

McCaskill's algorithm utilizes the statistical mechanical model to predict probabilities of individual secondary structures' occurrence, and as an extension of this, the probability that a given pair i,j will base-pair. This is accomplished by means of computing the partition function. Where K is the gas constant, T is the temperature, and S is a given secondary structure, the partition function is the sum of e^ (-Energy(S)/(K*T)) over all structures. Since there are exponentially many secondary structures, McCaskill exploits the fact that additivity of energy for secondary structures implies multiplicativity of the previous term.

In order to count each structure's energy contribution once and only once in the calculation of the partition function, McCaskill introduces several additional restricted matrices.

Q_{ij} is the partition function from i to j. Q^{b}_{ij} is the partition function given that i,j base-pair. Q^{m}_{ij} is the contribution of multi-loops over i,j. In order to calculate these, two additional auxiliary matrices are computed: Q^{1}_{ij}, which is the sum over i<=h<=j of Q^{b}_{ih} and Q^{m1}_{ij} is the sum over i<h<=j of Q^{b}_{ih}*e^(Penalty) where the Penalty is a function of the number of unpaired bases in the multi-loop.

Then the recurrences for the remainder are:

Q^{m}_{ij}= the sum over $i<h<=j$ of (e^(Penalty) + $Q^{m}_{i,h-1}$)*Q^{m1}_{hj}*e^(base) where Penalty is a function of the unpaired bases in the left hand portion of the multi-loop and base is the contribution of an additional base-pair inside a multi-loop.

Q^{b}_{ij}=e^ (-EIS1(i,j)/KT) + for $i<h<\ell<j$ the sum of e^(-EIS(i,j;h,ℓ)/KT)
+ for $i<h<j$ the sum of $Q^{m}_{i+1,h-1}$ * $Q^{m1}_{h,j-1}$*e^(-(M +base)/KT)

Q_{ij} =1.0 + for $i<=h<=j$ the sum of $Q_{i,h-1}$ * $Q^{1}_{h,j}$

Having introduced these matrices, the entry in $Q_{0,n-1}$ will be the sum over all secondary structures. The probability of any one secondary structure's occurrence is then, according to the statistical mechanical model, (e^(-E(S)/KT)/$Q_{0,n-1}$. From this, the individual base-pairing probabilities can be derived in cases of two varieties: where the base-pair is exterior to all others, and where the base-pair is included in a secondary structure interior to another base-pair. We will examine how to calculate these values for the non-nested model later in the thesis.

Eddy-Rivas Algorithm: Adding Hole Matrices

Pseudoknots are functionally important in a number of RNA sequences. They are conserved in ribosomal RNAs, and are apparently used to mimic tRNAs by some viruses. Eddy-Rivas is an algorithm that finds the minimum free energy structure for RNA sequences including pseudoknots in O(n^6) worst case running time. The violation of the nested convention causes the recurrence relation strategy of Zuker-Sankoff to break down, but Eddy-Rivas introduces a new variety of matrix that is quartic in n, called a 'hole'-matrix, to account for structures that have interactions that violate the nested convention.

The premise is that by combining such matrices, that within them contain combinations of other holed matrices, arbitrary pseudoknotted structures can be compared optimally. A simple example combination of the holed matrices is displayed in figure 3.

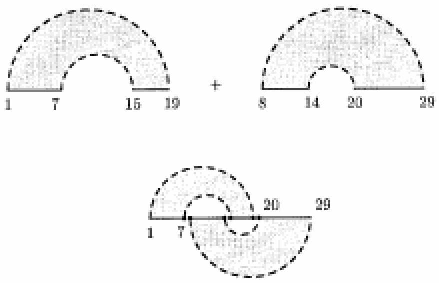

FIG 3: a combination of two holed matrices.

The actual secondary structure corresponding to a simple pseudoknot occurs when a region up or downstream of a loop containing unpaired bases pairs folds back and hydrogen-bonds to bases within the loop, as displayed in figure 4.

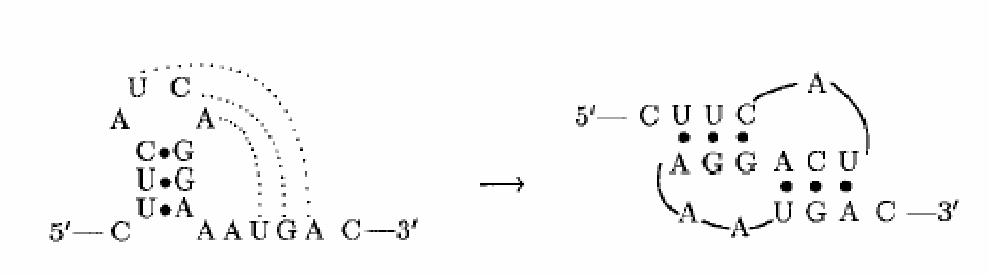

FIG 4: An elementary pseudoknot

The maximization relies upon some untested thermodynamic parameters, due to the relative infrequency of the occurrence of psedoknots, which were chosen for their ability to reproduce experimentally determined results without unduly over-predicting pseudoknotted structures. The parameters punish construction of pseudoknots within pseudoknots, and negative (stabilizing) secondary structure contributions within pseudoknots are lessened by a scaling factor. Eddy-Rivas also implements stabilizing energetic contributions for coaxial stacking, where if i and k base-pair, and l and j base-pair, and l=k+1, then there is a stabilizing energy contribution dependent upon the types of bases in the two respective pairings. Regrettably, in designing and implementing the partition and base-pairing probabilities for RNA secondary structure containing pseudoknots, we were unable to account for the contributions due to coaxial stacking.

The broad specifications for the hole matrices are:

Where i<k<l<j:
whx(i,j;k,l) is the energy of the optimal secondary structure for the region between i and j, excluding the region between k and l.

zhx(i,j;k,l) is the energy of the optimal secondary structure for the region between i and j, excluding the region between k and l, given that i and j base-pair.

yhx(i,j;k,l) is the energy of the optimal secondary structure for the region between i and j, excluding the region between k and l, given that k and l base-pair.

vhx(i,j;k,l) is the same, given that both i and j base-pair, and k and l base-pair.

The recurrences for these matrices recurse on themselves and on the unholed matrices, which are preserved largely intact from the Zuker-Sankoff Algorithm. Notably, the recurrences contain bifurcating cases that allow for a holed matrix element to contain contributions from the sum of two other holed matrices.
In figure 5 we see non-nested bifurcations that contribute to whx.

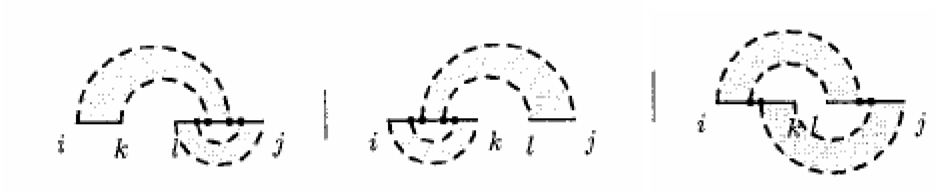

FIG 5: Non-nested contributions to whx(i,j;k,l)

These non-nested bifurcations allow arbitrary, unrestricted pseudoknots to be produced. In figure 6 we see how to construct a pseudoknot of the form (. . [. . { . .) . .] . . } and thus can produce k-ary pseudoknots, requiring an arbitrary number of sets of parentheses is ultimately accounted for in producing the optimal secondary structure.

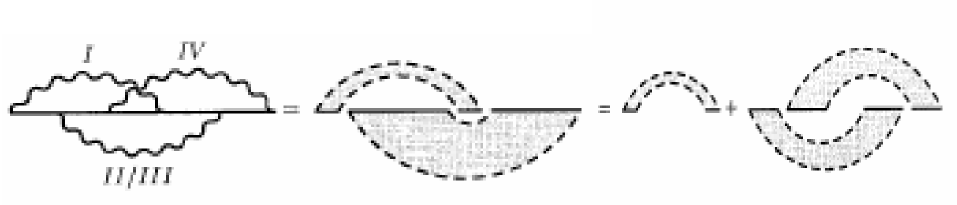

FIG 6: Pseudoknot Construction

In Eddy-Rivas, the final computation is truncated by combining only two hole matrices at the top level. As a result, inclusion of certain types of knots, such as those of a parallel beta-sheet, cannot be determined.

Calculating the Partition Function and Base-Pairing Probabilities with Pseudo-knots

The algorithm presented in this thesis calculates the partition function and base-pairing probabilities for RNA secondary structure including pseudoknots. This is a reasonably straightforward extension of the approach McCaskill used to calculate the partition function for Zuker-Sankoff. The algorithm restricts itself to a subclass of the pseudoknots examined by Eddy-Rivas, being 2-ary pseudoknots which require only two sets of parentheses to represent with a matched parentheses structure. This is equivalent to running Eddy-Rivas in approximation, and restricts pseudoknot varieties to the elementary class. This represents a large improvement in running time over the case where all pseudoknots covered by Eddy-Rivas were examined in the calculation of the partition function, but was in fact a compromise required by the extreme difficulty of accurately incorporating the contributions of pseudoknots crossing pseudoknots without over- or under-counting their contribution. At this juncture, the algorithm suffers from overwhelming time complexity, but I have hopes that several optimizations will allow a reduction in running time.

Several fundamental simplifications of Eddy's algorithm were made, in order to reduce the recurrences to manageable size and complexity. In addition to the failure of the algorithm to examine non-nested bifurcations within pseuodknots, it was also necessary to ignore thermodynamic contributions from coaxial stacking. In the nested case, this was an advantage, in as much as Vienna RNA package, the principal point of comparison, did not implement coaxial stacking. Ignoring coaxial stacking contributions was motivated by the difficulty of introducing subcases that were mutually exclusive and allowed for positive identification of the contiguity of two stem-loops.

As in Eddy-Rivas, we have the unholed matrices wx and vx, and the holed matrices whx, vhx, and zhx. Notably, the matrix yhx has been eliminated. Rather than being the energy of the optimal structure over the region i to j, or i to j excluding k to l, these matrices contain the sum of e^(- Energy(S)/KT) over all secondary structures S that the respective region can take on, with wx roughly corresponding to the term Q in McCaskill's algorithm and vx corresponding to Q^b . Additionally, we have wx1 equating to Q^1, and the additional matrices wm and wm1, which are Q^m and Q^{m1} from McCaskill.

The recurrences for these expressions are defined in precisely the same fashion as McCaskill, but it is notable that the various matrices are not close analogues to those in Eddy-Rivas. The optimization algorithm was unconstrained in comparison to our adaptation: given that only the minimum free energy structure of a given conformation would be traced back for a given entry in the matrices, the Eddy-Rivas algorithm can examine the same substructure multiple times, treating it as a different case in each instance, relying on maximization to choose only the optimal of the examined structures.

Tracing through the Eddy-Rivas recursions renders any number of duplications, all of which are necessarily eliminated in a correct calculation of the partition. The most important of these alterations is that the holed matrices whx, vhx, and zhx are constrained to have at least one base pair such that g,h base-pair, where i<=g<=k and l<=h<=j. This requirement is trivial for the matrices vhx and zhx, which already constrain at least one pairing, but is necessary to avoid recounting nested structures that would result from combinations of whx and other holed matrices. Hence we have that any contribution from the combination of two hole matrices is a contribution from a structure containing a pseudoknot, which was not the case in Eddy-Rivas, which depended upon the weighting and penalizing of combinations of hole matrices to prevent their being chosen as optimal when there was an available nested structure in the same conformation.

Thus the vx(i,j) is the partition function over all nested and non-nested structures between i and j given that i,j base-pair, and wx(i,j) is the partition function for all nested and non-nested structures from i to j.

The recursions for these two functions are as follows:

```
With wx(i,i-1)=1.0, wx(i,i)=1.0
wx(i,j) = 1.0+ the sum for i<=h<=j of wx(i,h-1) * wx1(h,j) (1)
               for all  i<=a<=k< m-1 <m<=l <j
                + wx(i,a-1)*vhx(a,l;k,m)                   (2)
                  *whx(k+1,j;l+1, m-1)* 3P10P*P11
```

Notably, recursion (1) is identical to the nested recursion from McCaskill; though the contributing elements of wx1 (a sum over vx) can themselves contain pseudoknots, wx1 contains no explicit bifurcation term. Since whx is constrained to contain a base pair across the hole, if (2) addresses all of the pseudoknot structure contributions once and only once, then the recursion is correct. Note that the loop bounds allow vhx to be decreased in size to only one base on the left hand side, and one base on the right. This case is undefined in Eddy-Rivas for vhx, but is defined in our vhx. Additionally, where we write energy parameter names or functions as they are defined in Eddy-Rivas e.g. P10P or EIS[1], they actually correspond to e^(P10P/KT) where the initial value of P10P is already the negation of the energy contribution of its corresponding structure element.

```
vx(i,j) = EIS1(i,j) + the sum for i<g<h<j of
                        EIS2(i,j;g,h)                         (1)
                    + the sum for i<h<j of
                        wm(i+1,h-1)*wm1(h,j-1)*P5             (2)
                    + wx(i,a-1)*vhx(a,l;k,m)
                       *whx(k+1,j;m+1,l-1)*3P10P*P11          (3)
                  for all  i<a<=k< m-1 <m<=l <j-1
```

Again, as matrices wm and wm1 are identical to their counterparts in McCaskill, the topmost recursions (1) and (2) are identical to the nested case while (3) attempts to uniquely account for the pseudoknot contributions.

There are then three additional recurrences in the calculation of the partition: whx, vhx, and zhx. Of the three, zhx is the most constrained. It has been redefined

```
zhx(i,j;k,l)= wx(i+1, k) * wx(l,j-1) +                       (1)
              for i<h<=k and l<g<=j the sum of
                 if (h,g can basepair):
                     vx(i,j;h,g)*zhx(h,g;k,l)                 (2)
                          else: 0
```

The first case handles structures interior to the base-pair, and the second recurses on vhx.

```
vhx(i,j;k,l)=
       if(i,j can't basepair or k,l can't basepair): 0       (1)
                   else if(i=k and l=j): 1.0                  (2)
                   else if(i=k or l=j): 0                     (3)
                   else
                     wx(i+1,k-1)*wx(l+1,j-1) +                (4)
                     EIS$^2$(i,j;k,l) +                       (5)
                     for i<g<k and l<h<j the sum of
                        (wx(i+1,g-1)*wx(h+1,j-1) + EIS$^2$
                         (i,j;g,h) ) * vhx(g,h; k,l)          (6)
```

Equation (1) and (3) are base cases for vhx; they prevent the two base-pairs required to satisfy the constraint. Equation (2) is a base case that is necessary to maintain the recursion on vhx in wx and vx. Equation (4) is the case where no additional pairs cross the hole, equation (5) is the contribution of vhx as a stack, bulge, or interior loop. Equation (6) recurses inwards on stacked vhx's internal to this one.

Finally:

```
whx(i,j;k,l)=
                     for i<=g<=k and l<=h<=j the sum of
                       wx(i,g-1)*zhx(g,h;k,l)*wx(h+1,j)     (1)
```

Equation (1) recurses on zhx, upon having enountered an exterior base-pair. This completes the recursions. In total, they primarily differ in effect from Eddy-Rivas in that they do not count non-nested bifurcations within pseudoknots.

Correctness of Probability Backtracking

The central premise of the backtracking to produce base-pairing probabilities is that the energy of an individual base-pair is equivalent to the sum of the probabilities of all of the secondary structures in which it appears. Again, the difficulty lies in setting up the recurrences in such a fashion as to measure the contributions consistently.

In my algorithm, which in this particular is an outright approximation, the probability matrix is developed from largest entry to smallest, as the innermost probabilities are dependent upon the occurrence of secondary structures exterior to them.

Probability recurrences:

```
Pr(h,l)= wx(0,h-1) *vx(h,l) * wx(l+1,n-1)/wx(0,n-1) +      (1)
             For i<h<l<j:
             += Pr(i,j)* vx(h,l) *EIS2(i,j,h,l)/vx(i,j) +   (2)
              Pr(i,j) vx(h,l)* (  wm(l+1,j-1) + wm(i+1,h-1) (3)
                              + wm(l+1,j-1) * wm(i+1,h-1) )
                                /vx(i,j)
             For l+1<k<m <n and 0<=a<h-1
             +=wx(0,a-1)*vhx(a,m-1;h-1,k)*vx(h,l)           (4)
                        *whx(k+1,n-1;l+1,m)/wx(0,n-1)
             For 0<=k<m<h and l<b<=n
               += whx(0,h-1;k,m) *vhx(k+1,b;m-1,l+1)        (5)
                      *vx(h,l)*wx(b,n-1)/(wx(0,n-1)
             For h<=m<k<=l
               +=wx(0,h-1)*vhx(h,l;m,k)                     (6)
                  *whx(m+1,n-1;k-1,l+1)/wx(0,n-1)
               +=whx(0,k-1;h-1,m+1)*vhx(h,l;m,k)            (7)
                  *wx(l+1,n-1)/wx(0,n-1)
                For  0<=a<=i<h<l<=j<=b<=n:
                +=wx(0,a-1)*vhx(i,j;h,l)                    (8)
                      *vhx(h,l;m,k)
                      *whx(m+1,n-1;k-1,j+1)/wx(0,n-1)
                +=whx(0,k-1;i-1,m+1)*vhx(i,j;h,l)           (9)
                      *vhx(h,l;m,k)*wx(j+1,b)/wx(0,n-1)
```

Equations (1), (2), and (3) are drawn from McCaskill's algorithm for nested structures, and are functionally identical to those. They account for the appearance of a given base in an outermost structure, in a bulge, stack, interior loop or multi-loop.

The remaining equations, (6) - (9), attempt to account for a base-pairs occurrence within a pseudoknot. Equations (4) and (5) attempt to account for the occurrence of a base-pair within the interior of a multi-loop but within a nested structure; cases exterior are handled by (1). Equations 4 and 5 are notably in error in this approximation: they constrain unduly the contiguity of the helices at h-1 and h. In these equations, there is a requirement that for correctness, the vhx term be substituted with zhx, and the whx term with yhx, which is defined symmetrically to zhx but recurses outwards from a base-pair on the interior of a holed structure. It is regrettable, but due to time-constraints, it is not possible to incorporate this change at this time.

Equations (6) and (7) account for the instances where the considered base-pair is participating in a bifurcation, and is the outermost such base-pair crossing the hole.

Equations (8) and (9) account for cases where the base-pair is participating in a bifurcation, but is not an outermost base-pair.

Application of Maximum Weight Matching

Withstanding the remaining under-counts recognized in the formulation, the majority of the contributions from elementary pseudoknots are accounted for in the probability back-trace. In order to determine to what extent they could be used to predict pseudoknotted structures, maximum weight matching was applied to the output probability matrices from running the program on a short database of pseudoknotted structures. The structures in question were a well-formed subset of those available in PseudoBase, and included instances of non-elementary structures.

The output probability matrices were mapped into weighted graphs, to which a well known O(n^3) maximum weight matching algorithm was applied, thresholding the edge weights at a threshold of probability .05 to avoid outputting trivial probabilities. Of the 1615 base-pairs, our algorithm identified 56.7%; incorrectly predicting 734. Some pseudoknot was predicted the great majority of the time, predicting ones occurrence all but 3 of the instances were one occurred. Running the same routine on the data sample produced 57.5% correct but incorrectly predicted 808 base-pairs. At a .10 threshold, our algorithm correctly predicted 53.7%, with 663 incorrect base-pairs, while the nested version of McCaskill at the same threshold predicted 55% correctly with 744 incorrectly predicted.

No figures are as yet available as to the extent to which the algorithm overpredicts pseudo-knots using maximum weight matching, but it is likely to be higher than Eddy-Rivas at threshold values that are not-prohibitive, since even with pseudoknots unconstrained, nested contributions that are mutually exclusive may be large.

Additional test runs to answer these questions in particular must wait for a provably correct formulation.

Conclusions

Construction of the algorithm to handle elementary pseudoknots was not wholly without problem; the difficulties of correctly handling the energetic contributions of the restricted set of structures uniquely was impressive. There are certain parameter uses that are omitted from the recurrences for the sake of simplicity-among them are the relative over-counts for dangle contribution. The dangle handling adopted a convention from Vienna RNA packages handling of dangle contributions. The convention is that at any point where a dangle could be placed, that is where a base-pair abuts an unconstrained structure, the base-pair is treated as having a dangle contribution for all secondary structures within the unconstrained region, including those that contain coaxial stacking with the base-pair in question, which should ideally preclude the dangle contribution.

This serves as an example of the magnitude of the difficulties of correctly handling the contribution of thermodynamic parameters with respect to Zuker-Sankoff, a substantially simpler algorithm. It should be noted that my original intent was to handle all classes of pseudoknots handled by Eddy-Rivas, but it was eventually necessary to omit those cases that contained non-nested bifurcations crossing holed structures.

The resulting approximation algorithm, while provably not optimal with respect to a few cases, is not without its merits. Besides providing a foundation for an optimal algorithm, its results in base-pair and pseudoknot prediction are by no means poor. It correctly incorporates most of the energetic contributions of the class of pseudoknots it handles, and produces meaningful probabilities that at their worst account for more of the energetic contributions to the bases in question than does the McCaskill implementation in Vienna RNA package.

Further work includes finding an optimal solution to the problem of dangles based upon mutually exclusive sub-cases of various matrices that additionally will allow the incorporation of coaxial stacking contributions. Additionally, the correction of the undercounts in the probability back-tracking are necessary, as previously noted. Finally, the probability algorithm should be redesigned in such a way as to be able to incorporate information about how one base-pair binding effects the probabilities of the other bindings.

LIST OF FIGURES

BIBLIOGRAPHY

Batenburg,F. H. D. van; Gultyaev, A.P; Pleij, C.W.A; Ng, J; Oliehoek, J; (2000). Pseudobase: a database with RNA pseudoknots. *Nucl. Acids Res.* **28,1**, 201-204

Cary, R. B; Stormo, G. D. (1995) *Graph-Theoretic approach to RNA modeling using comparative data.* ISBM-95, Eds.: Rawlings, C. and others. AAAI Press. 75-80.

Dsouza, M; Larsen, N; Overbeek, R. (1997) *Searching for patterns in genomic data.* Trends Genet. 12:497-8.

Gabow, H. N. (1974) *Implementation of algorithms for maximum matching on nonbipartite graphs.* Ph.D. dissertation, Dept. of Computer Science, Stanford Univ., Stanford, Calif.

Grate, Leslie. (1998) *Potential SECIS Elements in HIV-1 Strain HXB2.* Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology. 17:(5)

Hofacker, I.L; Fontana, W; Stadler, P.F; Bonhoe er, L.S. (1994) *Fast folding and comparison of rna secondary structures.* Monatshefte fur Chemie, 125:167-188.

Huynen M; Gutell R; Konings D. (1997) *Assessing the reliability of RNA folding using statistical mechanics.* Journal of Molecular Biology. 267: 1104-1112

Lescure, A., Gautheret, D., Carbon, P. and Krol, A. (1999) Novel selenoproteins identified *in silico* and *in vivo* by using a conserved RNA structural motif. *J. Biol. Chem.*, **274**, 38147–38154.

Kryukov, GV; Kryukov, VM; Gladyshev, VN. (1999) *New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements.* J Biol Chem. 274(48):33888-97

Mathews, David H; Sabina, Jeffrey; Zuker, Michael; Turner, Douglas H. (1999) *Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure.* Journal of Molecular Biology. *288*: 911-940.

McCaskill, J.S. (1990) *The equilibrium partition function and base pair binding probabilities for RNA secondary structure.* Biopolymers, 29, 1105-19.

Nussinov, R; Jacobson, A. B. (1980) *Fast algorithm for predicting the secondary structure of single-stranded RNA*. Proceedings of the National Academy of Sciences of the United States of America, 77(11):6309--6313.

Rivas, Elena; Eddy, Sean R. (1999) *A dynamic programming algorithm for RNA structure prediction including pseudoknots*. Journal of Molecular Biology, 285:2053.

Wuchty, S; Fontana, W; Hofacker, I.L; Schuster, P.(1999) *Complete suboptimal folding of RNA and the stability of secondary structures*. Biopolymers, 49(2):145-165

Zuker, M; Sankoff, D. (1984) *Rna secondary structures and their prediction*. Bull. Math. Biol., 46:591-621.