RNA Modeling Using Gibbs Sampling and Stochastic Context Free Grammars*

Leslie Grate and **Mark Herbster** and **Richard Hughey** and **David Haussler**
Baskin Center for Computer Engineering and Computer and Information Sciences
University of California
Santa Cruz, CA 95064

I. Saira Mian and **Harry Noller**
Sinsheimer Laboratories
University of California
Santa Cruz, CA 95064

Keywords: RNA secondary structure, Gibbs sampler, Expectation Maximization, stochastic context-free grammars, hidden Markov models, tRNA, snRNA, 16S rRNA, linguistic methods

Abstract

A new method of discovering the common secondary structure of a family of homologous RNA sequences using Gibbs sampling and stochastic context-free grammars is proposed. Given an unaligned set of sequences, a Gibbs sampling step simultaneously estimates the secondary structure of each sequence and a set of statistical parameters describing the common secondary structure of the set as a whole. These parameters describe a statistical model of the family. After the Gibbs sampling has produced a crude statistical model for the family, this model is translated into a stochastic context-free grammar, which is then refined by an Expectation Maximization (EM) procedure to produce a more complete model. A prototype implementation of the method is tested on tRNA, pieces of 16S rRNA and on U5 snRNA with good results.

Introduction

Tools for analyzing RNA are becoming increasingly important as *in vitro* evolution and selection techniques produce greater numbers of synthesized RNA families to supplement those related by phylogeny. Two principal methods have been established for predicting RNA secondary structure base pairings. The first technique, phylogenetic analysis of homologous RNA molecules (Fox & Woese 1975; Woese *et al.* 1983; James, Olsen, & Pace 1989), ascertains structural features that are conserved during evolution. The second technique employs thermodynamics to compare the free energy changes predicted for formation of possible secondary structure and relies on finding the structure with the lowest free energy (Tinoco Jr., Uhlenbeck, & Levine 1971; Turner, Sugimoto, & Freier 1988; Gouy 1987; Zuker 1989). When several related sequences are available that all share a common secondary structure, combinations of different approaches have been used to obtain improved results (Waterman 1989; Le & Zuker 1991; Han & Kim 1993; Chiu & Kolodziejczak 1991; Sankoff 1985; Winker *et al.* 1990; Lapedes 1992; Klinger & Brutlag 1993; Gutell *et al.* 1992).

Recent efforts have applied *Stochastic Context-Free Grammars* (SCFGs) to the problems of statistical modeling, multiple alignment, discrimination and prediction of the secondary structure of RNA families (Sakakibara *et al.* 1994; 1993; Eddy & Durbin 1994; Searls 1993). This approach is related to use of Hidden Markov Models (HMMs) to model *E. coli* DNA (Krogh, Mian, & Haussler 1993) and protein families and domains (Krogh *et al.* 1994; White, Stultz, & Smith 1994; Baldi *et al.* 1994). It incorporates elements of both the thermodynamic and phylogenetic approaches, with emphasis on the latter. The method of Sakakibara *et al* (Sakakibara *et al.* 1994; 1993) requires some initial knowledge of the common secondary structure of the sequences in the family. In contrast, Eddy and Durbin (Eddy & Durbin 1994) derive the structure of the grammar directly from unaligned sequences and estimate the probability parameters of the resulting grammar using Expectation Maximization (EM). Here we propose a different method for deriving the structure of the grammar from unaligned sequences which uses Gibbs sampling techniques (Geman & Geman 1984) described in (Lawrence *et al.* 1993; 1994). It is related to the EM methods described in (Neal & Hinton 1993) (incremental EM) and (Meng & Rubin 1992) (*Partitioned Expectation/Conditional Maximization*, or PECM).

The Gibbs sampler we propose simultaneously estimates the secondary structure of each sequence and a statistical model of the family with parameters describing the consensus secondary structure of the set as a whole. In particular, we estimate the number of helices, the length of each helix, the probability that a

*This work was supported in part by NSF grants CDA-9115268 and IRI-9123692, and NIH grant GM17129. Questions or comments should be addressed to haussler@cse.ucsc.edu.

helix is present, and the general nesting pattern of the helices. Furthermore, for each base-pair in each helix, a separate probability distribution over the 16 possible nucleotide pairs that could occur is estimated. The Watson-Crick pairs have much higher *a priori* probabilities in this estimation, but non-Watson-Crick pairs are also allowed, with appropriately small probabilities. Since these probabilities are estimated from the sequences, they are also influenced by phylogenetic relationships among the sequences The phylogenetic relationship guides the development of the statistical model in a large part through these probability parameters. In addition to parameters involving helices, parameters relating to the lengths of loops and other features are also estimated. When the Gibbs sampling has produced a crude statistical model for the family, this model is translated into a SCFG which is then refined by an EM procedure to produce a more complete model, as described elsewhere (Sakakibara *et al.* 1993). A prototype implementation of the method is tested on tRNA, pieces of 16S rRNA, and U5 snRNA.

Methods

Since this work builds on modeling RNA families with SCFGs (Searls 1993; Eddy & Durbin 1994; Sakakibara *et al.* 1993), we provide a review of this method first.

SCFG Overview

A grammar is a set of *productions* or rewrite rules. An example of an RNA grammar is shown in Figure 1 (in practice, a grammar would have many more productions). The symbols S_i are called *nonterminal symbols* and S_0 is the *start symbol*. The letters A,C,G,U are called terminal symbols and each represents a nucleotide. A grammar can be used to *derive* a set of RNA molecules. A molecule is derived by starting with the start symbol, and then repeatedly choosing a nonterminal symbol in the current molecule, finding a production in the grammar that has that symbol on the left hand side, and replacing that symbol in the molecule with the symbols on the right hand side of the production (this is termed *applying* the production), until there are no more nonterminals left in the molecule. A typical derivation is illustrated in Figure 2. When a production is applied, the left hand side nonterminal is shown with lines emanating from it to each of the symbols in the right hand side. The result is called a *derivation tree*, which can be seen by ignoring the dashed line. Ignoring the nonterminals (imagining that they really were replaced), leaves only the derived RNA molecule. The primary structure of the molecule is seen by tracing the letters from left to right along the frontier of the tree (dashed line). The secondary structure can be seen by highlighting the branching links between nucleotides that are derived from productions of the form $S_i \rightarrow \mathtt{X} S_j \mathtt{Y}$, where S_i and S_j are nonterminals and X and Y are nucleotides. These productions define the base-pairing in the molecule. Contiguous sequences of these base-pairs are helices (each

$$
\begin{aligned}
P = \{\ & S_0 \rightarrow S_1, && S_7 \rightarrow \mathtt{G}\ S_8, \\
& S_1 \rightarrow \mathtt{C}\ S_2\ \mathtt{G}, && S_8 \rightarrow \mathtt{G}, \\
& S_1 \rightarrow \mathtt{A}\ S_2\ \mathtt{U}, && S_8 \rightarrow \mathtt{U}, \\
& S_2 \rightarrow \mathtt{A}\ S_3\ \mathtt{U}, && S_9 \rightarrow \mathtt{A}\ S_{10}\ \mathtt{U}, \\
& S_3 \rightarrow S_4\ S_9, && S_{10} \rightarrow \mathtt{C}\ S_{10}\ \mathtt{G}, \\
& S_4 \rightarrow \mathtt{U}\ S_5\ \mathtt{A}, && S_{10} \rightarrow \mathtt{G}\ S_{11}\ \mathtt{C}, \\
& S_5 \rightarrow \mathtt{C}\ S_6\ \mathtt{G}, && S_{11} \rightarrow \mathtt{A}\ S_{12}\ \mathtt{U}, \\
& S_6 \rightarrow \mathtt{A}\ S_7, && S_{12} \rightarrow \mathtt{U}\ S_{13}, \\
& S_7 \rightarrow \mathtt{U}\ S_7, && S_{13} \rightarrow \mathtt{C} \qquad \}
\end{aligned}
$$

Figure 1: This set of productions P generates RNA sequences with a certain restricted structure. $S_0, S_1, \ldots, S_{13}$ are nonterminals; A, U, G and C are terminals representing the four nucleotides.

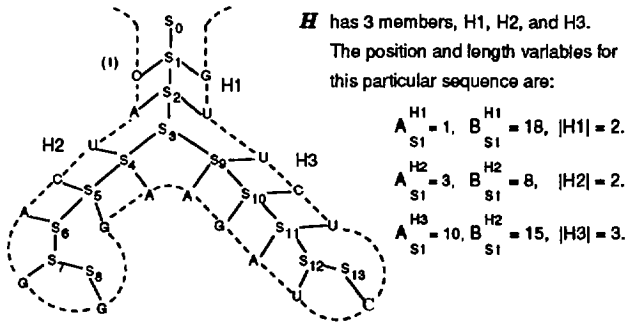

Figure 2: Derivation tree for the RNA sequence CAUCAGGGAAGAUCUCUUG using the grammar whose productions are given in Figure 1. The dashed line shows the primary sequence and $\mathcal{H}$ is the set of helix models. The 3 helix regions are $H1$, $H2$ and $H3$, and the nesting structure is "(()())" with $H1$ enclosing the others.

such helix we will later denote by H). The nesting structure of the helices is apparent in the derivation: reading the sequence left-to-right, the bottom two helices are clearly nested within the topmost helix, giving the nesting "(()())".[1]

For each nonterminal, placing a probability distribution over the productions with that nonterminal on the left hand side, enables the selection of an appropriate production at random every time a rule is applied (nonterminal is replaced). The result is a *Stochastic Context Free Grammar* (SCFG). An SCFG defines a probability distribution over a family of RNA molecules, where here, for simplicity, we identify an RNA "molecule" with its primary sequence and secondary structure. The probability of a molecule is the probability that it will be derived in a random derivation, assuming each production is chosen independently. Hence an SCFG defines a stochastic model for the family. In (Sakakibara *et al.* 1993), SCFG models were used to determine the most likely secondary

[1] Context-free grammars can only represent secondary structure with properly nested helices; they cannot represent pseudoknots. When we model a family with pseudoknots, these are currently ignored.

structure for an RNA sequence from a family, and to discriminate sequences in the family from those not in the family. An Expectation Maximization (EM) algorithm was then used to estimate the probabilities of the productions of an SCFG from unaligned training sequences. We use these same methods here.

Gibbs Sampling for common RNA secondary structure

The Gibbs sampler we use applies a variant of the method of Lawrence and colleagues (Lawrence *et al.* 1994) to locate helices in RNA molecules. Let $\mathcal{S}$ be a set of $|\mathcal{S}|$ sequences (each individual sequence $S \in \mathcal{S}$ has its own length, $|S|$) with similar secondary structure. Let $\mathcal{N}$ be four parameters representing the probabilities of each of the four nucleotides in the RNA family from which the sequences in $\mathcal{S}$ are drawn. We refer to these parameters as the *null model*. Let $\mathcal{H}$ be a set of $|\mathcal{H}|$ helix models and $H \in \mathcal{H}$ denote an individual helix model. Helices H in $\mathcal{H}$ are intended to be specific parametric models of the helices that are common to the sequences in $\mathcal{S}$. Associated with each helix H are the parameters $|H|$, its length in base-pairs, r^H, a matrix of parameters specifying the probability of occurrence for each of the 16 possible pairs of nucleotides that could occur in each base-pair of H, and p^H, the probability that H is present in a sequence in the family. Associated with $\mathcal{H}$ is a tree structure representing the nesting relationships among the helices in $\mathcal{H}$. Collectively, these parameters of $\mathcal{H}$ form a crude statistical model of the sequence family (defined more formally below). The goal is to estimate the structure and parameters of this model from the family of sequences $\mathcal{S}$. $\mathcal{H}$ is then used to create a grammar for this family and a more sophisticated parametric model in the form of a SCFG.

Consider the sequence in Figure 2. If this sequence has the typical secondary structure for sequences in its family, then the $\mathcal{H}$ for this family would have 3 individual helix models, $H1$, $H2$ and $H3$ with the lengths as shown and would specify how they are nested: $H2$ and $H3$ are totally enclosed by $H1$. The locations of these helices in this particular sequence are also shown.

In order to define fully the statistical model represented by $\mathcal{H}$, and to estimate $\mathcal{H}$ from the data $\mathcal{S}$, it is necessary to consider the "missing data" consisting of the set of hidden random variables that define the location of each helix within each sequence S. Let S_i denote the i-th nucleotide of a sequence S and $S_{i..j}$ the subsequence with endpoints S_i and S_j. The location of a helix H within a sequence S requires knowledge about both sides of the helix. These helix location parameters are denoted A_S^H and B_S^H. The first (5') side of a helix maps to the substring $S_{A_S^H \ldots A_S^H + |H| - 1}$ and the other (3') side maps to the substring $S_{B_S^H \ldots B_S^H + |H| - 1}$. Refer to Figure 2 for an example. The variables A_S^H and B_S^H are 0 if the helix H does not occur in sequence S. X_S denotes the set of location variables A_S^H and B_S^H for all helices in the sequence S, and we let $\mathcal{X}$ denote the set of all X_S.

We can formally define the manner in which $\mathcal{H}$ is a (crude) statistical model of a family of RNA sequences using the hidden variables $\mathcal{X}$. For a given sequence S,

$$P(S|\mathcal{H}) = \sum_{X_S} P(S|X_S, \mathcal{H}) P(X_S|\mathcal{H}),$$

where $P(S|X_S, \mathcal{H})$ is the probability of observing the sequence S given the particular placement X_S of the helices in S. Assuming independence, this is simply the product of the probabilities of each of the base-paired pairs of nucleotides in S, calculated using the parameters r^H, times the product of the probabilities of all the remaining nucleotides in S not placed in the helices, calculated using the parameters $\mathcal{N}$ of the null model. The term $P(X_S|\mathcal{H})$ is the prior probability of the placement X_S, which is 0 for any placement that violates the nesting structure of $\mathcal{H}$, and otherwise is proportional to the product of terms that are p^H for each helix H that is placed in S and $1 - p^H$ for each helix not placed.

The aim of our approach is to estimate the parameters in $\mathcal{H}$ from S. The method we use is *Maximum A Posteriori* (MAP) estimation, implemented by a Gibbs sampler/incremental EM method: We seek $\mathcal{H}$ that maximizes $P(\mathcal{H}|\mathcal{S})$. By Bayes rule, this is equivalent to maximizing $P(\mathcal{H})P(\mathcal{S}|\mathcal{H}) = P(\mathcal{H}) \prod_{S \in \mathcal{S}} P(S|\mathcal{H})$. The terms $P(S|\mathcal{H})$ in the latter product are defined above. The prior $P(\mathcal{H})$ we define to be uniform (or "uninformed") in all parameters, except for the parameters r^H that define the frequencies of the 16 possible nucleotide pairs in a given base paired position. For each of these we use a specific Dirichlet prior that we estimated from analysis of 16S rRNA multiple alignments (Sakakibara *et al.* 1993). During estimation, these guide the Gibbs sampler to position helices in appropriate places in the sequences, and greatly improve the results.

Currently we use a simple greedy method, described in the next section, to initially estimate the number of helices, their lengths and their nesting structure. Several candidate structures may be produced. For each of these, the Gibbs sampler is then used to compute an estimate of $\mathcal{H}$, and the best of these solutions is kept.

The sampling method we use maintains a placement X_S of the helices from $\mathcal{H}$ in each sequence $S \in \mathcal{S}$, along with current estimates of all the parameters in $\mathcal{H}$. At each step, a single helix H in a single sequence S is chosen at random, removed from S, and replaced at random back into S. The replacement is done according to the conditional distribution over all possible locations for the replacement given the current values of the parameters in $\mathcal{H}$ and the current positions of the other helices in S. The parameters in $\mathcal{H}$ are then reestimated given the current placement of all helices in all sequences. [2] Apart from it's close similarity with the

[2] In the current implementation, the parameters in $\mathcal{H}$ are

Gibbs sampler of (Lawrence *et al.* 1994) which it was modeled on, this system can also be viewed as a hybrid of the incremental EM method discussed in (Neal & Hinton 1993) and the PECM method of (Meng & Rubin 1992).

In the course of Gibbs sampling, we also obtain estimates for the hidden variables $\mathcal{X}$ giving the locations of the helices in each sequence. The Gibbs sampler can be annealed as in (Geman & Geman 1984) to make $\mathcal{X}$ approach MAP estimates. Hence the Gibbs sampler can be used alone to try to determine the secondary structure of each of the sequences. This is a direct extension to RNA of the Gibbs sampling approach for proteins described in (Lawrence *et al.* 1994). However, the model $\mathcal{H}$ used by the Gibbs sampler does not allow the length of a helix to vary between sequences, nor does it currently have any sophisticated parametric models for the loop regions. In contrast, SCFG models include all the parameters of the Gibbs sampling models plus site specific insertion and deletion probabilities for base-pairs within helices. More importantly, SCFG models have detailed parameteric models of each loop, including conserved nucleotides, average length and site specific insertion and deletion probabilities. Thus we expect to obtain more accurate secondary structure predictions from these SCFG models, at least when proper Bayesian methods are used to avoid overfitting the sequence data.

We have developed a program that translates the parametric model $\mathcal{H}$ produced by the Gibbs sampler to a stochastic context free grammar G. The initial values of the parameters in G that cannot be obtained from $\mathcal{H}$ are set according to an appropriate prior distribution. Then, using the same sequences $\mathcal{S}$, the EM algorithm described in (Sakakibara *et al.* 1993) (called *Tree-grammar EM*) is used to obtain a MAP estimate of the parameters of G using the same prior. Implicit in this estimation is a reestimation of the hidden random variables $\mathcal{X}$ that give the locations of the helices in each sequence. Generally, we obtained improved estimates in this way. However, since the length of a helix varies from sequence to sequence in the placement assigned by the SCFG model, the final locations of the helices cannot be specified by the simple random variables A_S^H and B_S^H defined above.

reestimated according to the mean (not the mode) of their posterior distribution given sufficient statistics (i.e. counts) from the current placement of the helicies and the parameters of the Dirichlet priors. Since the counts change little in each replacement, this calculation is efficient. Counts for the helix that is being replaced are subtracted from the total counts when it is removed, the parameters of the model are reestimated before the conditional distribution on possible replacement locations is calculated, and then the new counts from the replacement added back to the total counts after the replacement, as in (Lawrence *et al.* 1994).

Implementation of the method

The current implementation of our model construction algorithm has three parts: the development of an initial model $\mathcal{H}$, a Gibbs sampling placement of the helices in $\mathcal{H}$ in each sequence and re-estimation of the parameters of $\mathcal{H}$, and the generation and training of a SCFG with Tree-Grammar EM (Sakakibara *et al.* 1993). We use a massively parallel computer to speed the process (a MasPar MP-2204).

Development of an initial model While it is possible in principle for a Gibbs sampler with an appropriate prior and some kind of "model surgery" method similar to that used in (Krogh *et al.* 1994) to arrive at a correct helix model $\mathcal{H}$ starting from an empty model, this method would be time consuming and prone to falling into local minima. For this reason, we set out from the start to develop an alternative method for finding a good initial model $\mathcal{H}$.

Our method works by first finding likely helices in each sequence individually (making use of the prior base pairing statistics drawn from 16S rRNA (Sakakibara *et al.* 1993)) and then selecting a dominant nesting structure among those generated for each sequence in $\mathcal{S}$. The initial per-sequence search is a fast top-down parsing heuristic. In the first step, for all possible helices of length 4-14 and locations of the helix sides, we compute the numbers of bits saved (in the minimal-length encoding sense) by encoding the bases of the helix jointly as base-pairs rather than as independent bases. Our encoding is based on the null model probabilities of bases (denoted $\mathcal{N}$) and the prior base-pairing probabilities (r^H) mentioned above. That is, for a given $H \in \mathcal{H}$, and all possible starting positions A_S^H and B_S^H, we compute the number of bits required to encode the helix positions assuming the helix information

$$\sum_{i=1}^{|H|} -\log_2 P(S_{A_S^H+i-1} \leftrightarrow S_{B_S^H+|H|-i} | r^H), \quad (1)$$

where $\leftrightarrow$ denotes base-pairing, and, as the null model, the number of bits required to encode the two sequence segments independently of each other

$$\sum_{i=1}^{|H|} -\log_2 P(S_{A_S^H+i-1} | \mathcal{N}) - \log_2 P(S_{B_S^H+i-1} | \mathcal{N}), \quad (2)$$

and we take the difference of these two numbers. Figure 3 shows examples of this cost function. The helix position that saves the highest number of bits is the winner, and ties are broken arbitrarily.

Because of our requirement for proper helix nesting, each location of a helix breaks the sequence into two regions where additional helices could be: the outside (the concatenation of $S_{1 \ldots A_S^H-1}$ with $S_{B_S^H+|H| \ldots |S|}$) and the inside ($S_{A_S^H+|H| \ldots B_S^H-1}$) of the current helix (Figure 4). These two regions are then considered independently and recursively, and a possible nesting structure

Figure 3: Cost function, as defined by equation 2, of pairing a specific motif to a sequence. The ordinate is the number of bits saved (or goodness of fit) in arbitrary units, if the motif started at that point in the sequence. A peak indicates a starting location where the motif has a likelihood of pairing. The lefthand side shows pairing the motif GU, and the righthand side CCC, to the sequence shown on the abscissa. GU will like best to pair with CA, second best with CG, third best with UG (as indicated by our Dirchlet priors and as seen in graph). CCC would like best to pair with GGG, and there is only one position in the sequence where there are three Gs in a row.

is constructed. The recursion terminates when either a region is too small (fewer than 4 bases), or when placing a new helix results in no encoding-length savings over not placing that helix.

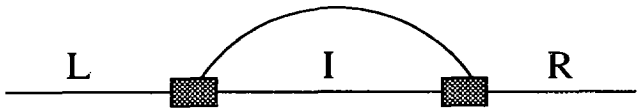

Figure 4: When two parts of a sequence are determined to be base paired (shaded boxes), the inside region I becomes independent from the outside regions L and R. Searching for base pairs then focuses on region I, and the concatenation of L and R.

Structure	Number	Structure	Number
(()()())	43	(()())	16
((()))	9	((())())	8
(((())))	6	()()()())	4
(()(()))	4	()()()	2
(()())()	1	((()())())	1
((()()))	1	((()))()	1
(()((()))	1	((())()())	1
(()(())())	1	()(()())	1

Figure 5: Nesting structures produced by the greedy algorithm for 100 tRNA sequences. The correct cloverleaf tRNA nesting structure, "(()()())", clearly dominates, and the second most common is this structure with one helix removed.

After completing this process on all sequences, the resulting nesting structures are compared, and the number of occurrences of each nesting type is tabulated (Figure 5). From these, a single dominant, or most frequently occuring, nesting structure is chosen for use in the Gibbs sampleing phase.

Refinement with Gibbs Sampling When trying to estimate the secondary structure of the sequences directly from the Gibbs sampler without using a SCFG, the goal is to estimate the locations $\mathcal{X}$ of all helices in the model in all sequences in $\mathcal{S}$. During this phase the nesting structure of the helix model $\mathcal{H}$ is fixed. We explored letting the helix lengths $|H|$ be variable (i.e. re-estimated during sampling) and holding them fixed. We examined re-estimating the helix probability parameters p^H and the base-pair probability matrices r^H and holding them fixed to the mode of the prior. There is a "freezing" option in which as soon as the placement X_S of the helices of the model $\mathcal{H}$ on a sequence S fits the nesting structure, S is removed from future consideration by the Gibbs sampler and X_S is "frozen". In this case we rely on the SCFG produced in a later stage to reliably reestimate the secondary structure of S in cases where its structure is frozen in to a suboptimal configuration.

Given a current placement X_S for the helices on a sequence S, the inner loop of the Gibbs sampling phase for S consists of selecting uniformly at random $H \in \mathcal{H}$ and modifying its position. For each possible repositioning X'_S of H in the sequence S, including the case when H does not occur in S, we calculate[3]

$$\frac{P(S|X'_S,\mathcal{H})P(X'_S|\mathcal{H})}{P(S|X_S,\mathcal{H})P(X_S|\mathcal{H})}. \quad (3)$$

These numbers are then normalized to sum to 1, to generate a posterior distribution over repositionings. This distribution will look very similar to those in Figure 3, where the abscissa indicates a pair of sequence starting positions. A new location for helix H in sequence S is then selected at random from this distribution, and it will probably be one of the highly favored positions that correspond to the peaks in the distribution.

Normally, we only consider repositionings of H that do not overlap the current positions of the other helices in S and are consistent with the nesting structure in $\mathcal{H}$. However, we obtain improved results if a repositioning of H may overlap another helix. In this case this other helix is simply removed, and (3) is calculated for the resulting X'_S. When the parameters of $\mathcal{H}$ are not held fixed during the Gibbs sampling, these parameters are re-estimated after each of the above helix replacements using the same method as in (Lawrence *et al.* 1994), employing the Dirichlet priors in the case of the r^H parameters.

For each $S \in \mathcal{S}$, a phase of the Gibbs sampling consists of performing the inner loop helix replacement a sufficient number of times to allow all helices several chances at repositioning within S (experimentally, $7|\mathcal{H}|$). If the "freezing" option is turned on, se-

[3]Actually, there is a slight adjustment to the parameters in $\mathcal{H}$ when the counts for the position of the helix H in S are decremented due to its removal, as in (Lawrence *et al.* 1994).

quences that match the dominant structure to start with require zero Gibbs phase iterations. The number of phases required to reach the point where few, if any, additional sequences are fit is very data dependent (around 4 phases for tRNA, and 15 or more for U5 snRNA). This is due to the presence of variant structures in the sequences. If some sequences lack some helices or the helices are of widely varying lengths, it is difficult for the sampler to identify where these helices fit in the overall model.

Training and evaluating the SCFG The initial SCFG for this part of the method is formed from the Gibbs sampling model $\mathcal{H}$ by considering only the sequences in $\mathcal{S}$ that match the dominant helix nesting structure. First, the helix productions of the grammar are generated based on the nesting structure and helix lengths found in $\mathcal{H}$. In our current implementation, the initial probabilities for each helix position are taken only from the prior distribution; in the future, we will transfer position-dependent base pairing statistics from the Gibbs sampler to the SCFG. The length of the loop regions between the helices are estimated as the arithmetic mean of the loop regions across the sequences.

The grammar is then trained using the entire sequence set $\mathcal{S}$ (including the ones the Gibbs sampler was unable to match to the dominant nesting structure). The parser associated with the SCFG training algorithm is used to determine the final estimate of the secondary structure of each sequence based on the trained grammar, see (Sakakibara *et al.* 1993).

Software implementation The MasPar MP-2204 at UCSC is a single instruction stream, multiple data stream (SIMD) parallel computer with 4096 32-bit processing elements arranged in a 64 × 64 grid (Nickolls 1990). It provides fast evaluation of all possible helix locations during our greedy model creation step. One processing element can be assigned to each of the $O(n^2)$ possible positions of helices of length l, and all positions are evaluated in parallel in $O(l)$ time. The minimum-cost position, following the greedy paradigm, is then located in $O(\log n)$ time and reported to the controlling program. Gibbs sampling iterations benefit in the same way.

Experimental Results

For tRNA, U5 snRNA, and parts of *E. coli* 16S rRNA, we produced helix nesting models (i.e. $\mathcal{H}$) using the Gibbs sampler. In the case of tRNA, this was translated to a SCFG followed by refinement using Tree-Grammar EM. All experiments used the freezing option and fixed helix lengths in the Gibbs phase (lengths having been determined in the greedy phase).

tRNA In our previous studies of tRNA (Sakakibara *et al.* 1993), we employed a hand generated helix nesting structure that captured fine details of tRNA structure. We would not expect the SCFG generated from the results of the Gibbs sampler to achieve the same level of detail for this or any other RNA family. However, if the nesting structure and estimates of the helix and loop lengths are fairly reasonable, we hoped that the Tree-Grammar EM algorithm would learn sufficient details to produce a sensitive discriminator.

Sequences were taken from a database that includes tRNAs from virus, archaea, bacteria, cyanelle, chloroplast, cytoplasm and mitochondria (Steinberg, Misch, & Sprinzl 1993). We used 10 sequences chosen at random from 1222 tRNAs to produce a helix nesting model which was then translated into a SCFG. The results were essentially insensitive to the number and choice of the sequences ie. similar results were obtained when the data set consisted of 4 to 100 sequences. This Gibbs grammar was compared to the hand generated nesting structure from our earlier work (Sakakibara *et al.* 1993). Both nesting structures were similar cloverleaf structures and corresponding helices had similar lengths (Figure 6). Starting from the Gibbs generated grammar, we repeated our earlier experiments that used the training set `MT10CY10` and Tree-Grammar EM (Sakakibara *et al.* 1993) to produce a discriminator. This discriminator (which was totally automatically produced, the only human input was choosing the initial sequences) was not as sensitive as our previous one, but is still remarkably good (Figure 7).

16S rRNA We examined three regions of 16S rRNA corresponding to nucleotides 588 to 880 in *E. coli* 16S rRNA (Woese *et al.* 1980; Noller & Woese 1981; Woese *et al.* 1983; Gutell *et al.* 1985). For each of the three segments, we used the Gibbs sampler to produce helix nesting models for four sequences (*E. coli* and three archaea) taken from the RDP (Larsen *et al.* 1993). Figure 9 shows the three regions along with the consensus structure produced from their respective Gibbs sampler models. Although Gibbs sampling produced a very good nesting structure, each individual sequence can differ from the standard.

U5 snRNA snRNAs are expected to be more difficult than tRNA and 16S rRNA because the structures change dynamically *in vivo* and strong static secondary structures may not be so prominent (Guthrie & Patterson 1988; McKeown 1993; Wise 1993). We used 34 U5 snRNA sequences (Guthrie, Roha, & Mian 1993) to produce helix nesting models (Figure 8). Gibbs sampling finds the accepted secondary structure (Guthrie & Patterson 1988) with the addition of one extra stem (stem X).

Discussion

We have described a new method to help determine the common secondary structure of homologous RNA molecules from a set of unaligned sequences using Gibbs sampling techniques similar to those that have been applied to proteins (Lawrence *et al.* 1993;

1994). The method is used to estimate the numbers, lengths and nesting structure of RNA helices in related RNA sequences. Previously, we relied on hand generated grammars to derive the structure of the grammar from unaligned primary sequences (Sakakibara *et al.* 1993). The novel aspect of the Gibbs sampling approach is that this task is performed automatically. The resulting crude statistical model is then translated into a SCFG and subsequently refined using EM. Since the model probabilities are computed from the sequences themselves, the phylogenetic relationships guide development of the model. While we have not explored the method completely yet, it has been applied with some success to tRNA, U5 snRNA and regions of 16S rRNA. We had less success in applying it to other snRNA families and to whole 16S sequences.

One of the weaknesses of the present method is in the greedy step to determine the nesting structure of the helices in the model. For families of longer RNA sequences, with more complex nesting structure, and for families with more variant structure, this method will often not produce a clear dominant nesting structure. We performed experiments in which Gibbs sampling is used to simultaneously estimate the locations and parameters of many helices without requiring a rigid common nesting structure. However, the results were not very promising. It appears that without some kind of information about the preferred locations of the helices within the sequence, there is too much "noise" in the form of competing placements for the Gibbs sampler to reliably sort things out. Thus, the method described here does not appear to scale up well to larger RNA modeling problems.

In order to address these problems, we are currently modifying the method to use information from a multiple alignment of the sequences to suggest the locations of potential helices in the model and their nesting structure, and to constrain the search for revised helix locations during Gibbs sampling. We use an HMM to align the sequences initially as in (Krogh *et al.* 1994), and then later add special penalty functions to this HMM to encourage base pairing in the helices we have found.

Acknowledgements

The authors thank Kevin Karplus, Chip Lawrence, Gary Stormo, and Michael Waterman for helpful discussion of these ideas, and Michael Brown, Deirdre Des Jardins, Kimmen Sjölander, and Rebecca Underwood for providing some of the software.

References

Baldi, P.; Chauvin, Y.; Hunkapillar, T.; and McClure, M. 1994. Hidden markov models of biological primary sequence information. *PNAS* 91:1059–1063.

Chiu, D. K., and Kolodziejczak, T. 1991. Inferring consensus structure from nucleic acid sequences. *CABIOS* 7:347–352.

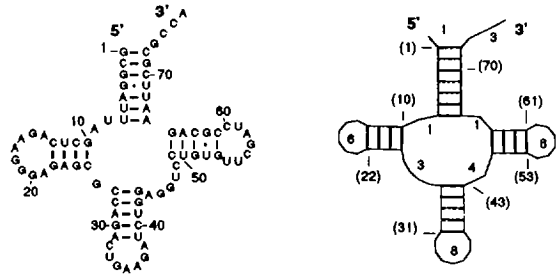

Figure 6: A standard tRNA (left) and tRNA structure produced by the Gibbs generated grammar (right). Only the fine hand tuned details are different: there is a bi-modal length distribution on some of the loops whereas the automatically generated Gibbs grammar structure has a single average loop length. The numbers in parenthesis indicate what position in the predicted structure maps to the real secondary structure. Note that the length of the loop regions is variable, but the lengths of base paired regions are constant. Since the loop lengths in the grammar structure are averaged from the sequences used to create the grammar, there is not a one to one correspondence between the nucleotide positions of the two structures.

Eddy, S. R., and Durbin, R. 1994. RNA sequence analysis using covariance models. *NAR* in press.

Fox, G. E., and Woese, C. R. 1975. 5S RNA secondary structure. *Nature* 256:505–507.

Geman, S., and Geman, D. 1984. Stochastic relaxations, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 6:721–742.

Gouy, M. 1987. Secondary structure prediction of RNA. In Bishop, M. J., and Rawlings, C. R., eds., *Nucleic acid and protein sequence analysis, a practical approach.* Oxford, England: IRL Press. 259–284.

Gutell, R. R.; Weiser, B.; Woese, C. R.; and Noller,

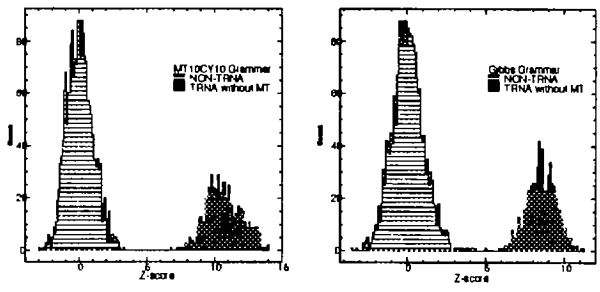

Figure 7: Discrimination histograms for tRNA, using the hand generated grammar (left) (Sakakibara *et al.* 1993), and the Gibbs grammar (right). This technique provides a very sensitive method of determining if a sequence is a member of a particular family.

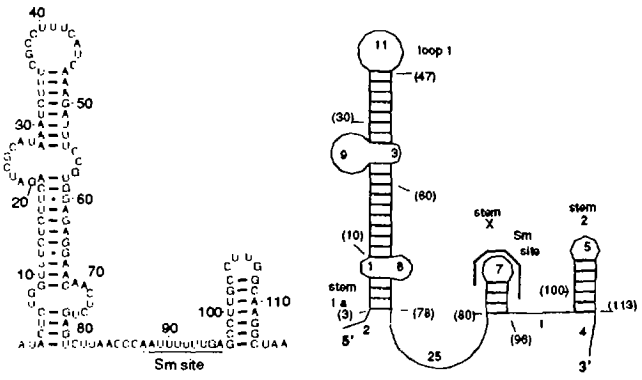

Figure 8: The accepted secondary structure for human U5 snRNA (left), and fit to the helix nesting structure produced by the Gibbs sampler (right). The stem/loop 1 and 2 regions are similar to the generally accepted ones (Guthrie & Patterson 1988). Stem X is in a variable region and frequently forms using part of the U-rich Sm binding site and A residues between stem 1a and the Sm binding site (it is not found in 20%, 7 out of 34, sequences).

H. F. 1985. Comparative anatomy of 16S like ribosomal RNA. volume 32 of *Prog. Nuc. Acids. Res. Mol. Biol.* 155–216.

Gutell, R. R.; Power, A.; Hertz, G. Z.; Putz, E. J.; and Stormo, G. D. 1992. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *NAR* 20:5785–5795.

Guthrie, C., and Patterson, B. 1988. Spliceosomal snRNAs. *ARGen* 22:387–419.

Guthrie, C.; Roha, H.; and Mian, I. S. 1993. Spliceosomal snRNA structure: function. Unpublished.

Han, K., and Kim, H.-J. 1993. Prediction of common folding structures of homologous RNAs. *NARes* 21:1251–1257.

James, B. D.; Olsen, G. J.; and Pace, N. R. 1989. Phylogenetic comparative analysis of RNA secondary structure. *Methods in Enzymology* 180:227–239.

Klinger, T., and Brutlag, D. 1993. Detection of correlations in tRNA sequences with structural implications. In Hunter, L.; Searls, D.; and Shavlik, J., eds., *First International Conference on Intelligent Systems for Molecular Biology.* Menlo Park: AAAI Press.

Krogh, A.; Brown, M.; Mian, I. S.; Sjölander, K.; and Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology* 235:1501–1531.

Krogh, A.; Mian, I. S.; and Haussler, D. 1993. A Hidden Markov Model that finds genes in *E. coli* DNA. Technical Report UCSC-CRL-93-33, University of California at Santa Cruz, Computer and In-

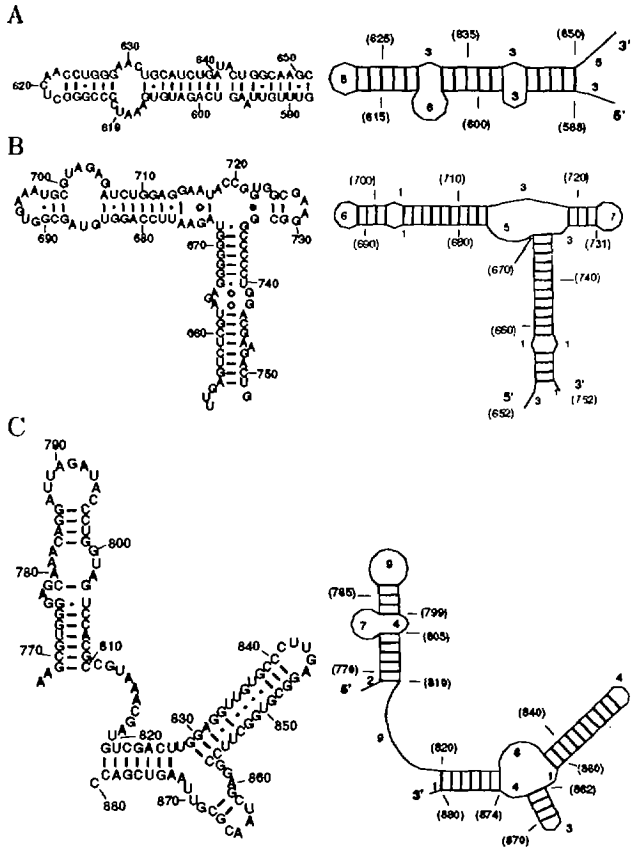

Figure 9: Three different regions of *E. coli* 16S rRNA showing the secondary structure (left) and consensus structure produced by the Gibbs Sampler (right). (There is not a one to one correspondence between the nucleotide positions of the two structures because the loop lengths in the grammar structure are averaged from the sequences used to create the grammar: thus do not match any particular sequence) *A.* Nucleotides 588 to 651; The number of helices is correct, but the rightmost one is too short. *B.* Nucleotides 652 to 752. The base paired region around 680 corresponds very well. In the region between 660 and 670, the loop in the standard secondary structure is incorporated into the long helix. The base paired region around 690 is shifted so that it forms a 4 long base paired region containing 2 CG bonds. Likewise, the region around 720 is shifted so that it forms a 4 long helix containing 3 CG bonds. *C.* 767 to 880. The base paired regions around 770, 785, and 820 are in excellent agreement. The two base paired regions 840 and 862 are both shifted. In the 840 region the standard structure has many weak GU base pairs. The Gibbs sampler finds a slightly more strong pairing incorporating the loop and bases around region 855.

formation Sciences Dept., Santa Cruz, CA 95064. in preparation.

Lapedes, A. 1992. Private communication.

Larsen, N.; Olsen, G. J.; Maidak, B. L.; McCaughey, M. J.; Overbeek, R.; Macke, T. J.; Marsh, T. L.; and Woese, C. R. 1993. The ribosomal database project. *NARes* 21:3021–3023.

Lawrence, C. E.; Altschul, S.; Boguski, M.; Liu, J.; Neuwald, A.; and Wootton, J. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* 262:208–214.

Lawrence, C. E.; Altschul, S.; Wootton, J.; Boguski, M.; Neuwald, A.; and Liu, J. 1994. A gibbs sampler for the detection of subtle motifs in multiple sequences. In *Proceedings of the Hawaii International Conference on System Sciences*, 245–254. Los Alamitos, CA: IEEE Computer Society Press.

Le, S. Y., and Zuker, M. 1991. Predicting common foldings of homologous RNAs. *J. Biomolecular Structure and Dynamics* 8:1027–1044.

McKeown, M. 1993. The role of small nuclear RNAs in RNA splicing. *Current Opinion in Cell Biology* 5:448–454.

Meng, X.-L., and Rubin, D. B. 1992. Recent extensions to the em algorithm. *Bayesian Statistics* 4:307–320.

Neal, R. M., and Hinton, G. E. 1993. A new view of the em algorithm that justifies incremental and other variants. to appear in Biometrika.

Nickolls, J. R. 1990. The design of the Maspar MP-1: A cost effective massively parallel computer. In *Proc. COMPCON Spring 1990*, 25–28. IEEE Computer Society.

Noller, H. F., and Woese, C. R. 1981. Secondary structure of 16S ribosomal RNA. *Science* 212:403–411.

Sakakibara, Y.; Brown, M.; Hughey, R.; Mian, I. S.; Sjölander, K.; Underwood. R.; and Haussler, D. 1993. The application of stochastic context-free grammars to folding, aligning and modeling homologous RNA sequences. Technical Report UCSC-CRL-94-14, University of California, Santa Cruz, Computer and Information Sciences Dept., Santa Cruz, CA 95064.

Sakakibara, Y.; Brown, M.; Mian, I. S.; Underwood, R.; and Haussler, D. 1994. Stochastic context-free grammars for modeling RNA. In *Proceedings of the Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society Press.

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

Searls, D. B. 1993. The computational linguistics of biological sequences. In *Artificial Intelligence and Molecular Biology.* AAAI Press. chapter 2, 47–120.

Steinberg, S.; Misch, A.; and Sprinzl, M. 1993. Compilation of tRNA sequences and sequences of tRNA genes. *NAR* 21:3011–3015.

Tinoco Jr., I.; Uhlenbeck, O. C.; and Levine, M. D. 1971. Estimation of secondary structure in ribonucleic acids. *Nature* 230:363–367.

Turner, D. H.; Sugimoto, N.; and Freier, S. M. 1988. RNA structure prediction. *Annual Review of Biophysics and Biophysical Chemistry* 17:167–192.

Waterman, M. S. 1989. Consensus methods for folding single-stranded nucleic acids. In Waterman, M. S., ed., *Mathematical Methods for DNA Sequences.* CRC Press. chapter 8.

White, J. V.; Stultz, C. M.; and Smith, T. F. 1994. Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. *Mathematical Biosciences* 119:35–75.

Winker, S.; Overbeek, R.; Woese, C.; Olsen, G.; and Pfluger, N. 1990. Structure detection through automated covariance search. *CABIOS* 6:365–371.

Wise, J. A. 1993. Guides to the heart of the spliceosome. *Science* 262:1978–1979.

Woese, C. R.; Magrum, L. J.; Gupta, R.; Siegel, R. B.; Stahl, D. A.; Kop, J.; Crawford, N.; Brosius, J.; Gutell, R.; Hogan, J. J.; and Noller, H. F. 1980. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *NARes* 8:2275–2293.

Woese, C. R.; Gutell, R. R.; Gupta, R.; and Noller, H. F. 1983. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. *Microbiology Reviews* 47(4):621–669.

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