Detecting the Coevolution of Biosequences—An Example of RNA Interaction Prediction

Chen-Hsiang Yeang, Jeremy F. J. Darot,† Harry F. Noller,‡ and David Haussler§*

*Simons Center for Systems Biology, Institute for Advanced Study, Princeton, New Jersey; †EMBL, European Bioinformatics Institute, Cambridge, England, United Kingdom; ‡Center for Molecular Biology of RNA and Department of Molecular, Cell, and Developmental Biology, University of California at Santa Cruz; and §Center for Biomolecular Science and Engineering, University of California at Santa Cruz

A probabilistic graphical model is proposed in order to detect the coevolution between different sites in biological sequences. The model extends the continuous-time Markov process of sequence substitution for single nucleic or amino acids and imposes general constraints regarding simultaneous changes on the substitution rate matrix. Given a multiple sequence alignment for each molecule of interest and a phylogenetic tree, the model can predict potential interactions within or between nucleic acids and proteins. Initial validation of the model is carried out using tRNA and 16S rRNA sequence data. The model accurately identifies the secondary interactions of tRNA as well as several known tertiary interactions. In addition, results on 16S rRNA data indicate this general and simple coevolutionary model outperforms several other parametric and nonparametric methods in predicting secondary interactions. Furthermore, the majority of the putative predictions exhibit either direct contact or proximity of the nucleotide pairs in the 3-dimensional structure of the *Thermus thermophilus* ribosomal small subunit. The results on RNA data suggest a general model of coevolution might be applied to other types of interactions between protein, DNA, and RNA molecules.

Introduction

Understanding the evolution of biological systems at different levels is a central question of biological science. Selective constraints often operate on the functions of the entire molecular system, which requires coordinated interactions of its components. The evolution of those components is thus likely coupled.

Perhaps, the most well-known example of dependent evolution is the coevolution between the components of a molecular apparatus. Examples include the compensatory substitution of nucleic acids in RNA molecules (Noller and Woese 1981; Gutell et al. 1986; Rzhetsky 1995; Hofacker et al. 1998; Knudsen and Hein 1999; Eddy 2001; Rivas et al. 2001; di Bernardo et al. 2003; Coventry et al. 2004; Washietel, Hofacker, and Stadler 2005; Washietl, Hofacker et al. 2005; Pedersen et al. 2006), the coevolution of amino acid residues between ligand receptor pairs (Goh et al. 2000; Ramani and Marcotte 2003), protein–protein interactions (Barker and Pagel 2005), intramolecular interactions (Pollock et al. 1999; Fares and Travers 2006), and the copresence of enzymes in the same metabolic pathways (Bowers et al. 2004). Detecting the coevolution of biosequences is important in determining the structure of protein and RNA molecules, predicting molecular interactions and the functions of genes. At conceptual level, it is the first step toward a comprehensive understanding of the evolution of molecular systems.

Coevolution of genes has been investigated in various previous studies. Some of these have demonstrated the coevolution of genes by correlating their sequence substitution rates with functional properties such as their physiological functions (Wall et al. 2005), the number of interactions (Fraser et al. 2002), their interacting partners (Fraser et al. 2002), and their coexpressed genes (Jordan et al. 2004). Others have applied different correlation metrics to capture the covariation of sequences, including correlation coefficients (e.g., Goh et al. 2000; Dutheil et al. 2005; Fares and Travers 2006), mutual information (e.g., Atchley et al. 2000; Ramani and Marcotte 2003; Gloor et al. 2005), multiple dependency score (Tillier and Lui 2003), and the deviance between marginal and conditional distributions (e.g., Lockless and Ranganathan 1999).

A major drawback of these approaches is that they did not give a quantitative measure of how likely covariation is to arise from neutral evolution. Many authors have thereby extended the continuous-time Markov process to coevolving sequences in the problems of predicting RNA secondary structures (e.g., Knudsen and Hein 1999; Eddy 2001; Rivas et al. 2001; Pedersen et al. 2006), amino acid residue, and protein–protein interactions (e.g., Pollock et al. 1999; Barker and Pagel 2005). However, the number of parameters in those models grows quadratically with the number of possible joint states. For instance, the dimension of a substitution rate matrix of 2 amino acids is 400 × 400. It is computationally expensive to estimate those large numbers of parameters, and the estimated parameters are subject to overfitting limited sequence data. Previous approaches address these problems by reducing the number of states (e.g., Pollock et al. 1999; Barker and Pagel 2005), specifying the rules of interactions in the substitution rate matrix (e.g., Rzhetsky 1995) or restricting to RNA–RNA interactions (e.g., Knudsen and Hein 1999; Eddy 2001; Rivas et al. 2001; Pedersen et al. 2006).

We propose a general continuous-time Markov model to detect coevolution from aligned biomolecular sequences. Sequence substitution of the 2 sites is modeled by a joint continuous-time Markov process. The null (independent) model hypothesizes that 2 sites evolve independently. The alternative (coevolutionary [CO]) model is obtained from the null model by reweighting the independent substitution rate matrix to favor double over single changes. The model hypothesizes that coevolving sites have a positive, fixed rate for double changes and smaller rates for single changes relative to the null model. The spatial dependency of adjacent site pairs is captured by a hidden Markov model (HMM), where the hidden variables are the

Key words: coevolution, continuous-time Markov models, RNA tertiary interactions, RNA secondary interactions.

E-mail: chyeang@soe.ucsc.edu.

Mol. Biol. Evol. 24(9):2119–2131. 2007
doi:10.1093/molbev/msm142
Advance Access publication July 17, 2007

interaction states of the site pairs and the observables are their sequences across species. Similar to other continuous-time Markov models, it incorporates the information of sequence substitution and phylogeny, thus reduces the spurious covariations arising from common phylogeny. Furthermore, it applies a simple reweighting scheme on the substitution rate matrix that requires neither simplification of states nor prior knowledge about interactions. It allows us to detect various types of interactions (e.g., noncanonical RNA tertiary interactions, interactions of amino acid residues, protein–protein, and protein–DNA interactions) without incorporating the complex interaction rules in the model or learning a large number of parameters.

As a proof-of-concept demonstration, we applied the model to predict the secondary and tertiary interactions of 16S rRNA and tRNA molecules. The results indicate a general model of coevolution achieves an accuracy level comparable or superior to specific models encoding the RNA base pairing rules and various nonparametric scores of covariation. Furthermore, it also detects the tertiary interactions of the RNA molecules that do not necessarily follow typical base pairing rules. This is encouraging for the model's applicability to other types of coevolution such as protein–protein and protein–DNA interactions, where even less is known a priori.

Materials and Methods

Overview of the CO Model

The CO model we developed operates on the 2 paired, aligned families of sequences along 2 orthogonal dimensions. The first dimension is time, with a continuous-time Markov process modeling the sequence substitution of the 2 entities considered. This model operates at each of the paired positions across species. The second dimension is space, with an HMM operating along the consecutive paired positions and determining that regions of the 2 entities are coevolving. It belongs to a class of probabilistic models termed graphical models (Jordan et al. 1999), which includes a wide range of models such as Bayesian networks, Markov random fields, HMMs, and so on. Similar graphical models were introduced by Yang (1995), Felsenstein and Churchill (1996), and have been recently adopted for instance by Siepel and Haussler (2004) to detect the conserved regions of DNAs. The inputs of the model consist of 2 families of aligned sequences (one for self interactions), a phylogenetic tree of the species with branch lengths, and the parameters pertaining to the continuous-time Markov process and the HMM. The outputs of the model are the coevolving position pairs.

Sequence Substitution Model of Single Molecular Entities

The sequence substitution of a single site is modeled by a continuous-time Markov process (Yang 1995). Denote by $x(t)$ the sequence composition at time t. For an RNA nucleotide $x(t) \in \{A, C, G, U\}$. The probability vector $\mathbf{P}(x(t))$ follows a Markov process at an infinitesimal time interval:

$$\frac{d\mathbf{P}(x(t))}{dt} = \mathbf{P}(x(t))\mathbf{Q}. \tag{1}$$

where $\mathbf{Q}$ is the substitution rate matrix. It is a 4×4 matrix for nucleic acids. Each row of $\mathbf{Q}$ must sum to 0 in order to make $\mathbf{P}(x(t))$ a valid probability vector. Additional constraints may be applied to $\mathbf{Q}$. In this work, we use the HKY model of nucleotide substitution (Hasegawa et al. 1985). It characterizes $\mathbf{Q}$ by 5 (coupled) parameters: a stationary distribution (π_A, π_C, π_G, π_U) and a transition/transversion ratio κ:

$$\mathbf{Q} = \begin{pmatrix} - & \pi_C & \kappa\pi_G & \pi_U \\ \pi_A & - & \pi_G & \kappa\pi_U \\ \kappa\pi_A & \pi_C & - & \pi_U \\ \pi_A & \kappa\pi_C & \pi_G & - \end{pmatrix} \tag{2}$$

Each diagonal entry is -1 times the sum of the other entries in the same row.

The transition probability $\mathbf{P}(x(t)|x(0))$ at a finite time interval t is given by the matrix exponential $e^{\mathbf{Q}t}$, which is the solution of equation (1):

$$P(x(t)=b|x(0)=a) = e^{\mathbf{Q}t}[a,b]. \tag{3}$$

Sequence Substitution Model of 2 Molecular Entities

The continuous-time Markov model can be extended to the joint states of 2 sites. Define $\mathbf{x}(t) = (x_1(t), x_2(t))$ as the joint state of 2 sites—such as the sequence composition of a nucleotide pair—at time t. The sequence substitution follows the same equation for the single-site evolution (eq. 1), but the dimensions of the probability vector and the substitution rate matrix are much bigger. Here $\mathbf{P}(\mathbf{x}(t))$ is a 1×16 vector and $\mathbf{Q}$ a 16×16 matrix.

We first consider a null model in which the 2 nucleotides evolve independently with an identical rate matrix. The transition probability of the joint state is the product of the transition probabilities of the 2 nucleotides. The transition probability matrix of the joint state is

$$P(\mathrm{x}(t)|\mathrm{x}(0)) = e^{\mathrm{Q}_2 t}. \tag{4}$$

where $\mathbf{Q}_2$ is the substitution rate matrix of 2 independent sites. It can be shown (Pagel 1994) that

$$\mathbf{Q}_2((a_1,a_2),(b_1,b_2)) = \begin{cases} \mathbf{Q}(a_1,b_1) & \text{if } a_2=b_2, \\ \mathbf{Q}(a_2,b_2) & \text{if } a_1=b_1, \\ -\mathbf{Q}(a_1,b_1)-\mathbf{Q}(a_2,b_2) & \text{if } a_1=b_1, a_2=b_2, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

In $\mathbf{Q}_2$, the rate of a single-nucleotide change is equal to the corresponding rate in the single-site rate matrix $\mathbf{Q}$, and the rates of double nucleotide changes are all zero. This is intuitive since only one nucleotide can change within an infinitesimal time interval if 2 sites evolve independently.

A general model of dependent evolution can be obtained by "reweighting" the entries of the independent rate matrix by a "potential matrix" ψ:

$$\mathbf{Q}_2^w = \mathbf{Q}_2 \circ \psi, \tag{6}$$

where ψ is a 16 × 16 matrix and "∘" denotes the following operation:

$$(\mathbf{Q}_2 \circ \psi)(a,b) = \begin{cases} \mathbf{Q}_2(a,b) \cdot \psi(a,b) & \text{if } a \neq b, \mathbf{Q}_2(a,b) > 0, \\ \psi(a,b) & \text{if } a \neq b, \mathbf{Q}_2(a,b) = 0, \\ -\sum_{b' \neq b} \mathbf{Q}_2(a,b') \circ \psi(a,b') & \text{if } a = b. \end{cases} \tag{7}$$

where $a = (a_1, a_2)$ and $b = (b_1, b_2)$. It multiplies an off-diagonal, nonzero entry $\mathbf{Q}_2(a, b)$ by $\psi(a, b)$, sets the value of a zero entry $\mathbf{Q}_2(a, b)$ to $\psi(a, b)$, and normalizes a diagonal entry as -1 times the sum of the other entries in the same row. The reweighted $\mathbf{Q}_2^w$ is a valid substitution rate matrix. ψ specifies the hypotheses of selective constraints on nucleotide pair states. Higher weights reward the transitions to the advantageous states, and lower weights penalize the transitions to the disadvantageous states.

CO Models

Equation (7) can express any valid substitution rate matrix. Instead of learning an unrestricted substitution rate matrix from the data (e.g., Knudsen and Hein 1999; Holmes and Rubin 2002; Pedersen et al. 2006), we parametrize ψ by 2 free parameters. This restricted model alleviates the problems of estimating the parameters over the high dimensional (3D) space and overfitting limited data.

If the sequence composition of interacting pairs is known, we can set ψ to reward the transitions to the interacting pairs and penalize the transitions of the opposite directions. In general, the sequence pairing rules can be complex or unknown. We thus propose a general reweighting scheme which is not tied to specific rules of interactions. We assume there are multiple selectively advantageous sequence pairs that are distinct in each component. Without knowing which states have selective advantages or disadvantages, we only consider whether a transition changes 1 or 2 nucleotides. The model rewards the transitions where both nucleotides change and penalizes the transitions where only one nucleotide changes. We call this model a "simple CO" model. The reweighting scheme becomes:

$$\psi(a,b) = \begin{cases} r & \text{if } a_1 \neq b_1 \text{ and } a_2 \neq b_2, \\ \epsilon & \text{if either } a_1 = b_1 \text{ or } a_2 = b_2, \\ 1 & \text{otherwise.} \end{cases} \tag{8}$$

where $\int < 1$ and $r > 0$ are free parameters of penalty and reward. This model does not require prior knowledge about interactions, hence can be possibly applied to different types of interactions. Moreover, it introduces only 2 extra free parameters $\int$ and r, and thus alleviates the overfitting problem.

As a comparison, we introduce 2 other reweighting schemes that explicitly incorporate the base pairing rules of RNA secondary interactions. The "Watson–Crick coevolution" model—abbreviated as WC model—rewards the single transitions that establish Watson–Crick base pairing from noninteracting pairs and penalizes the single transitions to non–Watson-Crick base pairs:

$$\psi(a,b) = \begin{cases} \frac{1}{\epsilon} & \text{if } a \notin \text{WC and } b \in \text{ WC, single changes,} \\ \epsilon & \text{if } b \notin \text{ WC, single changes,} \\ 0 & \text{ double changes.} \end{cases} \tag{9}$$

where $\int < 1$ and WC denotes the states AU and GC in both orders.

Some GU/UG pairs in RNA molecules form weaker hydrogen bonds (GU wobble). The "Watson–Crick with GU wobble" model—abbreviated as WCW model—rewards the state transitions that establish or maintain Watson–Crick or GU base pairs and penalizes the state transitions that break the extended rule:

$$\psi(a,b) = \begin{cases} \frac{1}{\epsilon} & \text{if } b \in \text{ WCW, single changes,} \\ \epsilon & \text{if } b \notin \text{ WCW, single changes,} \\ 0 & \text{ double changes.} \end{cases} \tag{10}$$

where WCW denotes the states of WC, GU, or UG. The CO models similar to the WC and WCW schemes have been applied in previous studies to predict RNA structures (e.g., Noller and Woese 1981; Rzhetsky 1995; Hofacker et al. 1998; Eddy 2001; Rivas et al. 2001; di Bernardo et al. 2003; Coventry et al. 2004; Washietel, Hofacker and Stadler 2005; Washietl, Hofacker et al. 2005).

Evaluating the Likelihood of Sequence Data

Given the parameters of the continuous-time Markov process and the phylogenetic tree, we want to know how likely the observed sequences are to arise from the underlying process. The observed sequences correspond to the states of the leaves in the phylogenetic tree, and the marginal likelihood of the observed sequences is the joint likelihood summed over all possible states of internal (ancestral) nodes. This marginal likelihood can be efficiently calculated using a dynamic programming algorithm (Felsenstein 1981). Briefly, let u be a node in the tree, v and w be its children, and t_v and t_w be the branch lengths of (u, v) and (u, w). Define $P(L_u|a)$ as the probability of all the leaves below u given that the base assigned to u is a. The algorithm follows the recursion:

$$P(L_u|a) = \begin{cases} \delta(x_u = a) & \text{if } u \text{ is a leaf,} \\ \sum_b e^{\mathbf{Q}t_v} a, b] P(L_v|b) \sum_c e^{\mathbf{Q}t_w} a, c] P(L_w|c) & \text{otherwise.} \end{cases} \tag{11}$$

where δ(.) is an indicator function. Gaps on leaf nodes are treated as missing data such that each nucleotide is given an equal probability.

Capturing Spatial Dependency

RNA secondary interactions often form stems comprising consecutive base pairs. To capture the spatial dependency between adjacent pairs, we define an HMM on the nucleotide pairs along the aligned sequences. We pair an

RNA sequence against itself in the opposite direction, with different offsets specifying the end positions of the pairs. At position s, the hidden variable $y(s) \in \{0, 1\}$ indicates whether coevolution occurs at position s (i.e., $y(s) = 1$) or not (i.e., $y(s) = 0$). The spatial dependency of hidden variables is specified by a homogeneous Markov chain with transition probability $P(y(s+1) = 1|y(s) = 0) = P(y(s+1) = 0|y(s) = 1) = \alpha$. The observed variable $X(s)$ comprises the pairs of sequences at position s across all species. The emission probability $P(X(s)|y(s))$ corresponds to the likelihood of the sequence data, conditioned on the null model of independent evolution or the alternative model of coevolution. Given the transition and emission probabilities, we then apply the Viterbi algorithm to identify the interacting regions of the 2 sequences. Similar approaches have been applied to detect the conserved regions of DNAs (e.g., Yang 1995; Felsenstein and Churchill 1996; Siepel and Haussler 2004).

Detecting Covariation Using Nonparametric Scores

A simple and popular method of detecting the covariation of 2 sites is to calculate the mutual information of their sequences across the sample species. Denote x_1 and x_2 the sequence composition of sites 1 and 2, $P_{12}(x_1, x_2)$ the joint probability mass function of x_1 and x_2, and $P_1(x_1)$ and $P_2(x_2)$ the marginal probability mass functions of x_1 and x_2. The mutual information between x_1 and x_2 is

$$MI(x_1;x_2) = \sum_{x_1,x_2} P_{12}(x_1,x_2)\log\left(\frac{P_{12}(x_1,x_2)}{P_1(x_1)P_2(x_2)}\right). \quad (12)$$

$MI(x_1;x_2)$ is high if x_1 can reliably predict x_2 and vice versa, which is equivalent to the covariation between x_1 and x_2. This method is popular for its low computational cost. Yet it also picks up spurious covariation due to shared phylogeny. Various other methods have been proposed to reduce the spurious covariation. For instance, Atchley et al. (2000) simulated the sequences according to the phylogenetic tree and used the simulated sequences to evaluate the significance of a mutual information. Tillier and Lui (2003) calibrated mutual information scores to avoid selecting a site that correlates with many other sites. Dutheil et al. (2005) calculated a vector of expected numbers of changes between each pair of leaf nodes on the phylogenetic tree based on a sequence substitution model and evaluated the correlation coefficient between 2 vectors (Dutheil et al. 2005).

Data and Preprocessing

Aligned 1542-base 16S rRNA sequences from thousands of species were compiled in the ribosomal RNA database (Ribosomal Database). To maximize the coverage of the phylogeny, we extracted representative sequences from 146 species covering archaea, bacteria, eukaryotes, mitochondria, chloroplast, and cyanobacteria. A phylogenetic tree was derived from these sequences using the parsimony DNAPARS program of the PHYLIP package (PHYLIP). The parameters of the HKY model of single nucleotides and the branch lengths of the tree were estimated using the maximum likelihood methods in the PAML package (PAML). The species names of the 146 16S rRNA sequences and their phylogenetic tree are reported in Supplementary File 1 (Supplementary Material online). We estimated parameters $\int$, r, and α by varying their values over discrete combinations ($\int$ from 0.05 to 0.9, r from 0.1 to 0.9, and α from 0.05 to 0.45) and identified the combination of these values that gave rise to the highest area under the receiver operating characteristic (ROC) curve for false positives ≤ 200. The parameter values are as follows: $\int = 0.1$, $r = 0.12$, $\alpha = 0.3$ for the CO model, $\int = 0.05$, $\alpha = 0.3$ for the WC model, and $\int = 0.9$, $\alpha = 0.3$ for the WCW model.

The 16S rRNA sequence was then paired with itself in the opposite direction in order to evaluate potential coevolution between all possible nucleotide pairs. The first entity in the model was the 16S rRNA sequence itself, and the second entity was the reversed sequence, shifted by a number of nucleotides varying from 1 to 1542 and "rolled over" to match the length of the first entity. Because interactions between adjacent nucleotides are physically infeasible, we only considered the pairs that were at least 3 nucleotides apart.

Aligned tRNA sequences from 60 species were extracted from the *Rfam* database (Rfam) (accession number RF00005). The selection criteria, selected species and their phylogenetic tree, procedures of phylogeny reconstruction, parameter estimation, and data pre-processing closely follow those of the 16S rRNA data and are described in Supplementary File 1 (Supplementary Material online).

Results

We applied the general the CO model to aligned 16S rRNA and methionine tRNA sequences. As a comparison, we also applied 2 other continuous-time Markov models of nucleic acid pairs (the WC and WCW models), mutual information scores with significance (Atchley et al. 2000), multiple dependency score (Tillier and Lui 2003), and the CoMap program (Dutheil et al. 2005) to the same data sets. We gauged the performance of models in terms of the ROC curves of secondary interaction prediction and the capacity of detecting tertiary interactions. On 16S rRNA data, the CO model outperformed all the other methods in detecting the secondary interactions. Moreover, the majority of the putative interactions predicted by the CO model contained the nucleic acid pairs that were in contact or close in the 3D structure of the ribosome complex. We also tested the robustness of the prediction results against alignment and found that the accuracy was sensitive to the quality of aligned sequences. By running the predictions on subsets of the sequences, we found that a wide coverage of sequences on the phylogenetic tree was needed. On tRNA data, the CO model has slightly better performance than WC, WCW, and mutual information scores.

16S rRNA Structure

16S ribosomal RNA is a major part of the small subunit (30S subunit in bacteria) of ribosome. The secondary structure of the *Escherichia coli* 16S rRNA is shown in figure 1. About half of its 1542 bases participate in secondary

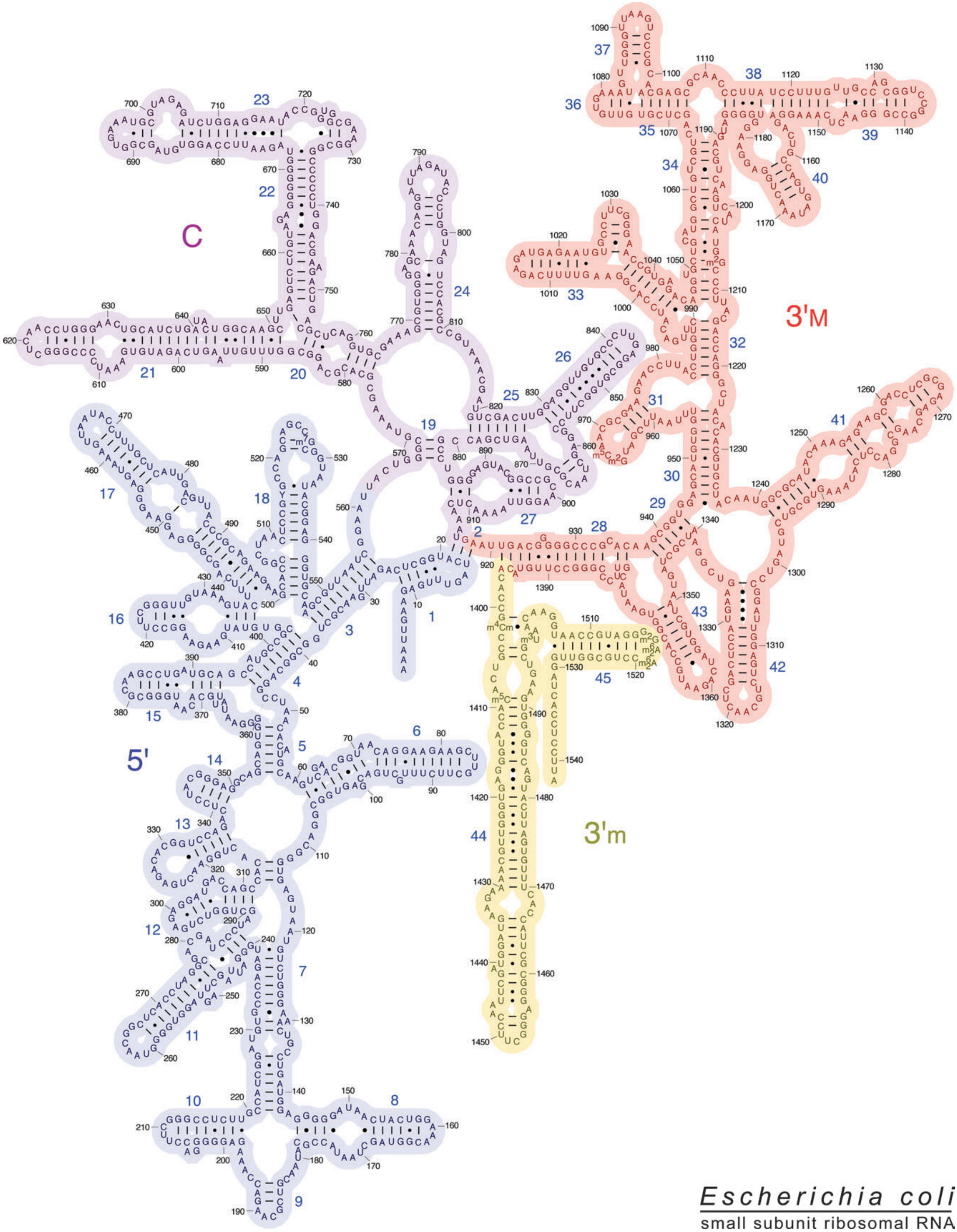

FIG. 1.—16S rRNA secondary interactions, excerpt from Yusupov et al. (2001).

Table 1
Putative Tertiary Interactions of 16S rRNA Predicted by the CO model, Reported in CRW Database

Pair	Covariation	Proximity	Pair	Covariation	Proximity
1450–1453	Strong	Direct contact	−245–283	Strong	Direct contact
−1303–1334	Strong	Direct contact	771–807	Mild	Direct contact
1029–1032	Strong	Direct contact	−722–733	Strong	Direct contact
1515–1521	Strong	Direct contact	70–99	Strong	Direct contact
−570–866	Strong	Direct contact	658–747	Mild	Direct contact
−450–483	Weak	Direct contact	70–98	Weak	Direct contact
1289–1352	Weak	Direct contact	183–193	Weak	Direct contact
618–621	Weak	Direct contact	−440–497	Strong	Close
1007–1021	Mild	Close	1005–1024	Weak	Close
1008–1022	Weak	Close	73–99	Weak	Close
998–1041	Weak	Close	1010–1021	Weak	Close
73–96	Weak	Close	72–97	Weak	Gap
72–98	Weak	Gap	72–75	Weak	Gap
202–215	Weak	Gap	71–98	Weak	Gap
464–467	Weak	Gap	73–100	Weak	Gap
71–99	Weak	Gap	95–98	Weak	Distant
139–1019	Weak	Distant	865–1382	Weak	Distant
74–747	Weak	Distant	245–351	Weak	Distant
1001–1020	Weak	Distant	223–1021	Weak	Distant
602–838	Weak	Distant	617–1289	Weak	Distant
139–1010	Weak	Distant			

interactions (437 secondary interactions). Yet almost all nucleotides in the ribosomal RNA are involved in interactions with either other RNA nucleotides or ribosomal proteins (Noller 2005). Unlike secondary interactions, almost all possible base pairs appear in tertiary interactions (Noller 2005). The available 3-dimensional structures of the 30S subunit (Wimberly et al. 2000) and the entire ribosome complex (Yusupov et al. 2001) from *Thermus thermophilus* and *E. coli* (Schuwirth et al. 2005) provide a standard for the verification of tertiary interaction prediction.

16S rRNA Tertiary Interaction Prediction

We first demonstrate the capacity of the CO model in detecting 16S rRNA tertiary interactions. By setting the log likelihood ratio threshold to be 6.0 (simulation P value $< 10^{-6}$), there were 41 putative predictions that were not secondary interactions. They are the likely candidates for tertiary interactions.

Table 1 lists the nucleotide positions of the 41 putative predictions. The majority of the pairs are separated by between 10 and 60 nucleotides. The relatively short distance along the primary sequence is consistent with the previous observation that interdomain interactions of rRNAs are rare (Yusupov et al. 2001).

Some tertiary interactions exhibit strong compensatory substitutions between Watson–Crick states. For instance, pair 1303–1334 constitutes 98 CGs, 43 GCs, and 2 UAs. Other pairs have strong covariation patterns between non–Watson-Crick or GU states. For instance, pair 245–283 constitutes 68 CCs and 65 UUs. By examining the sequence composition in the phylogenetic tree (fig. 2), we found that these double substitutions occurred multiple times across different lineages, suggesting they were unlikely to arise from neutral mutations. The sequence composition of all the 41 putative predictions is given in Supplementary File 2 (Supplementary Material online).

We verified these putative predictions by examining the 3D coordinates of the nucleotide pairs from the structure data of the ribosome 30S subunit of *T. thermophilus* (Protein Data Bank accession number 1J5E; Wimberly et al. 2000). Strikingly, among the 41 putative predictions, 15 demonstrate direct contact of the nucleic acid pairs (the closest distance between the atoms of the 2 nucleotides ≤ 4 Å), 8 demonstrate proximity of the nucleotide pairs (the closest distance between the 2 nucleotides > 4 Å but ≤ 8 Å). Eight pairs contain gaps in the corresponding positions in *T. thermophilus*. Only 10 predicted pairs are physically distant in the 3D structure. Overall, more than half of the putative predictions are supported by the 3D structure data as likely candidates for tertiary interactions. We also examined the putative predictions from the mutual information scores. By setting the threshold to be 0.78, there were 47 putative predictions. Only 14 of them were less than 8 Å apart in *T. thermophilus* 30S ribosomal subunit, substantially fewer than the CO model predictions. The 47 putative predictions and their sequence composition are shown in Supplementary File 2 (Supplementary Material online).

The 41 putative predictions were compared with the 16S rRNA annotation in the Comparative RNA Web site (CRW, Cannone et al. 2002). Forty-four base pairs were annotated by CRW as tertiary interactions, whereas 17 of them were categorized by us as secondary interactions according to figure 1. Among the 27 remaining tertiary interactions, 6 were overlapped with our predictions (marked in table 1). Despite the small overlap, our predictions were substantially better than Dutheil et al. (2005), where only 3 predictions on 16S rRNA were overlapped with the tertiary interactions in CRW.

We then investigated the *T. thermophilus* ribosomal structure of 9 position pairs in table 1 that demonstrate strong covariation of the sequence composition. We denote a sequence composition strongly covarying if there exists only a few (between 2 and 4) “dominant” states (the nucleotide pair sequences which occur in more than 10 species),

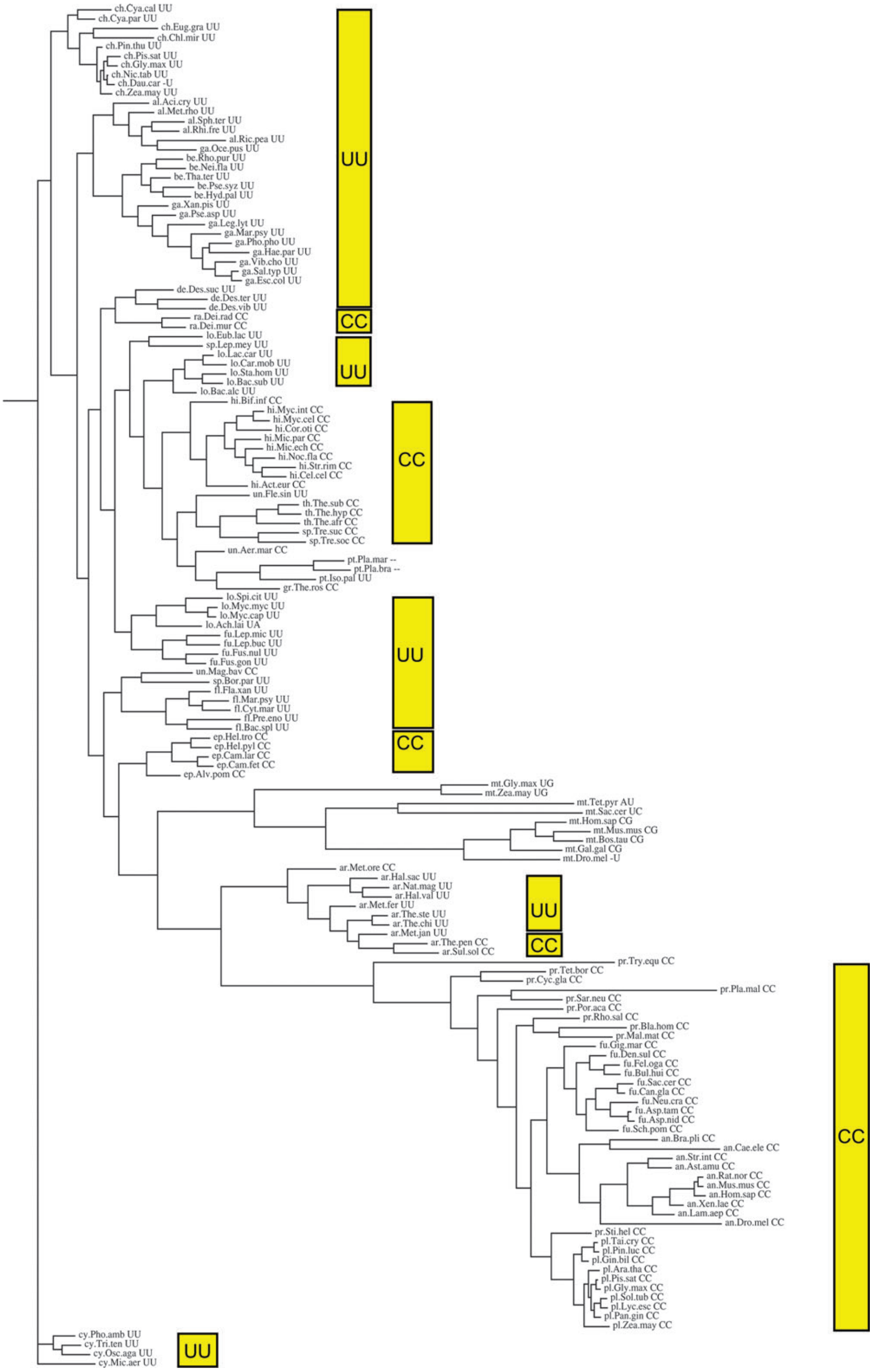

FIG. 2.—Sequence covariation of pair 245–283 detected by CO model; gray: CC pairs, black: UU pairs.

and those dominant states either share no common bases or are Watson–Crick or GU pairs. Most of them possess either hydrogen bonds or other structural constraints subject to coevolution, further corroborating the capacity of our model to detect CO pairs.

Pair 1450–1453 is dominated by GA and UG pairs among the 146 sequences. It is part of the UACG tetraloop in *T. thermophilus* and part of the UUCG tetraloop in *E. coli*.

Pair 1303–1334 is dominated by CG and GC pairs. It is a standard Watson–Crick base pair and belongs to

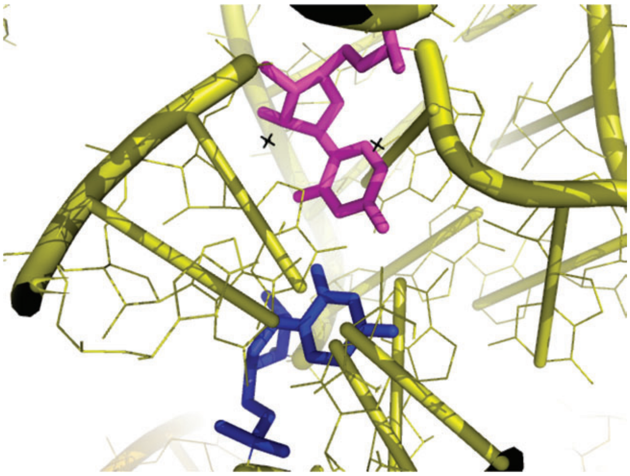

FIG. 3.—Predicted interaction of 16S rRNA, 245–283; top: 283, down: 245.

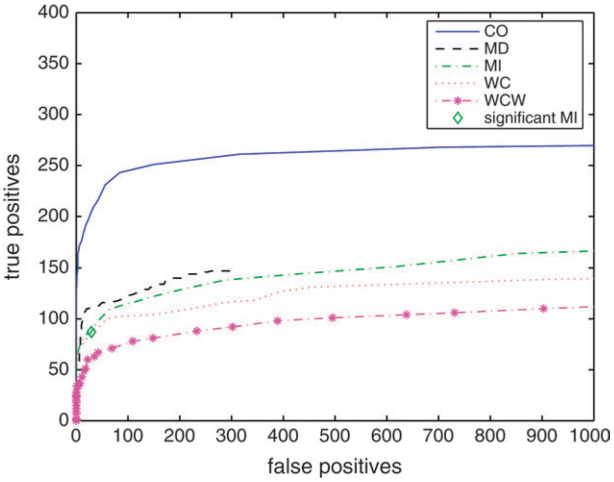

FIG. 4.—ROC curves of 16S rRNA secondary interaction prediction. MD, multiple dependency; MI, mutual information; WC, Watson–Crick reweighting; WCW, Watson-Crick-GU reweighting; significant MI, mutual information score with $P<0.001$.

a stem between positions 1303–1314 and 1334–1323 in the 3′M domain. It is categorized by CRW as a tertiary interaction.

Pair 1029–1032 is dominated by UG and GA pairs. Like pair 1450–1453, it is part of a tetraloop from 1029 to 1032 in the 3′M domain. Despite the lack of hydrogen bonding between the two positions, the tetraloop varies between the standard UUCG and GNRA forms across species (Woese et al. 1990). The tetraloop is UUCG in *E. coli* and GNRA in *T. thermophilus*.

Pair 1515–1521 is dominated by GC, UG, and CG pairs. This is an interesting case because positions 1515 and 1520 (and positions 1514 and 1521) form standard Watson–Crick secondary interactions. In *T. thermophilus* 16S rRNA, nucleotide 1515 seems to pair with 1521, whereas in *E. coli* 16S rRNA nucleotide 1515 pairs with 1520 (data not shown). The electron density in *T. thermophilus* and *E. coli* structures indicates 1515 is located in the plane between 1520 and 1521. This suggests a 3-way interaction between 1515, 1520, and 1521 may occur.

Pair 570–866 is dominated by GC, UA, and CG pairs. It is a previously identified Watson–Crick tertiary interaction (Gutell et al. 1986).

Pair 440–497 is dominated by UU, AC, and CG pairs. In spite of its strong covariation, the 2 bases are 10.8 Å apart, suggesting that their interaction is unlikely. The coevolution may reflect other structural constraints that are not yet clear.

Pair 245–283 is dominated by UU and CC pairs, and pair 722–733 is dominated by AA and GG pairs. They belong to homopyrimidine and homopurine rRNA tertiary interactions, respectively. Pair 245–283 belongs to Saenger XII- or XIV-type interaction (Saenger 1984). Pair 722-733 belongs to Saenger type I noncanonical pair (Saenger 1984). An A-A pair in *T. Thermus* and a G-G pair in *E. coli* should be isosteric with each other as the distances between the C′1 carbons of the 2 bases are conserved in A-A and G-G pairs (Leontis et al. 2002). The C-C and U-U pairs are also isosteric.

Pair 70–99 demonstrates covariation between UG, AU, GC, and CG pairs. It is a likely Watson–Crick secondary interaction in the 5′ domain. It does not appear on the initial list of secondary interactions because the region around this pair is highly variable, and alignment mistakes are likely to occur. In *E. coli*, pair 70–98 is a secondary interaction (see fig. 1). However, multiple position pairs with small offsets from 70 to 98 were detected by the CO model: 70–99, 71–99, 72–98, 72–97, 73–96, 73–99, 73–100. These signals probably correspond to the same base pair in the misaligned sequences.

As an example, figure 3 shows pair 245–283 in *T. thermophilus* 16S rRNA using *PyMOL* (PyMOL). The *PyMOL* visualization of the 33 putative interactions (excluding the 8 pairs with gaps in the *T. thermophlus*) is provided in Supplementary File 3 (Supplementary Material online). Intriguingly, all but one of the pairs demonstrating strong covariation patterns have direct contacts, and almost all the pairs that demonstrate proximity but not direct contact have weaker covariation patterns. Conversely, some pairs that do not have strong covariation patterns also have direct contacts.

16S rRNA Secondary Interaction Prediction Accuracy

Figure 4 shows the ROC curves of 16S rRNA secondary interaction prediction using 5 methods: the CO, WC, and WCW models of continuous-time Markov processes, the mutual information score, and the multiple dependency score (Tillier and Lui 2003). By varying the confidence thresholds on the scores, each model gives different combinations of false-positive and true-positive numbers. For illustrative purposes, the figure only shows the portion of the ROC curves where the number of false positives is less than 1000. The predicted interactions (both real secondary and putative tertiary interactions) of the CO model are reported in Supplementary File 4 (Supplementary Material online).

The CO model substantially outperforms all the other methods. For instance, with 150 false positives, the CO model (solid) can recover 251 secondary interactions (sensitivity rate 57%), whereas the multiple dependency (dashed) and mutual information (dashed dotted) scores with similar false-positive numbers can only recover 131 (sensitivity rate 30%) and 121 (sensitivity rate 28%) interactions, respectively. The WC (dotted) and WCW (solid star) models have an even inferior performance. The multiple dependency score marginally outperforms mutual information, indicating its capacity of removing some false positives attributed to the common phylogeny. Yet the sensitivity is still only half of the CO model.

We also compared the prediction performance to the mutual information score with statistical significance (Atchley et al. 2000) and the CoMap program based on sequence substitution models (Dutheil et al. 2005). Setting a stringent threshold on the p-value of mutual information generated by bootstrap simulation can reduce false positives but also increase false negatives. Although it is useful in setting the cutoff value, the ROC curve of mutual information is unchanged. The significant threshold of mutual information ($P < 0.001$, mutual information 0.56, 29 false positives, 87 true positives) is marked on the mutual information ROC curve of figure 4. The ROC curve of the CoMap program lies below the WCW model (results not reported). We suspected that the poor performance was due to the inaccurate parameter setting. With similar number of bacterial 16S rRNA sequences (79 from Dutheil et al. and 77 from ours), the 2 models reported the same number of predictions (126) and yielded similar prediction accuracy (117 from Dutheil et al. and 118 from ours). In addition, by comparing with the CRW (Cannone et al. 2002), our model detected 6 tertiary interactions and the CoMap program detected only 3 of them.

The poor performance of the WC and WCW models is partly due to the existence of many false-positive pairs following the Watson–Crick or GU pairing rules. For example, pair 1321–1355 is dominated by UG (70) and UA (50) sequences (WCW log likelihood 1.355), which can be explained by several single-nucleotide transitions (UA ↔ UG) along the lineages. The reweighting schemes of WC and WCW models (eqs 8 and 9) also introduce many artifacts. Whereas the transition probability from a non-WC (or GU) state to a WC (or GU) state is rewarded, the probability of staying at a non-WC (or GU) state remains high due to low substitution rate or short branch length. If 2 nodes sharing a common parent possess a WC (or GU) and a non-WC (or GU) state, respectively, then assigning the common parent to the non-WC (or GU) state of one child yields high WC (WCW) scores. For example, pair 384–666 is dominated by GG (66) and CG (25) sequences (WC log likelihood score 13.80). By assigning GG to a common ancestor of a GG group and a CG group, the transition GG–CG is rewarded and the transition GG–GG is not penalized. In addition, mutual information yields the false positives that can be explained by a few neutral mutations along the phylogeny. Pair 585–930 contains covarying base pairs GC (80), UG (39), and CA (10), hence has a high mutual information score (0.784). However, those sequences occur at separate clades of the phylogenetic tree, thus is likely to arise from independent mutations.

The substantial number of false positives comparable to the number of detected interactions seems unsatisfactory. However, considering the large number of pairs calculated, the specificity of each model is extremely high. The total number of nucleotide pairs considered is (1542 × 1539)/2 = 11,86,569. Thus, even the apparent upper limit of the false-positive number in figure 4 (300 false positives) yields a very high specificity rate (99.97%). As described previously, some "false-positive" pairs may reflect tertiary interactions or other structural constraints because figure 4 only considers secondary interactions.

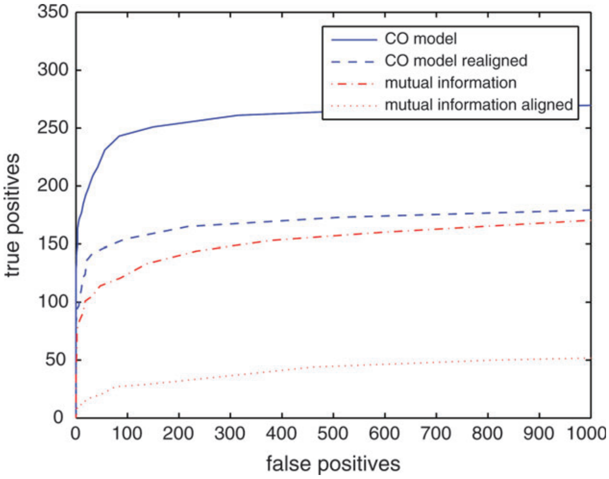

FIG. 5.—Robustness of ROC curves against alignment; solid: CO with original alignment, dashed: CO with realignment, dashed dotted: mutual information with original alignment, dotted: mutual information with realignment.

Sensitivity to Alignment

The credibility of interaction predictions from CO models relies on the quality of sequence alignment across multiple species. Misaligned sequences may break the covariation patterns of interacting positions or introduce spurious covariations of noninteracting positions. The 16S rRNA sequences have reasonably good structural alignments. In general, however, the structure data of an RNA molecule may not be available. To justify the general applicability of CO models, we examined the sensitivity of prediction performance against alignments.

We removed all the gaps in the 16S rRNA data and realigned the sequences using ClustalW (Thompson et al. 1994) and applied the 4 methods to realigned data. Figure 5 shows the ROC curves of each method on structurally aligned data and on the alignment which is based purely on sequences. Clearly, the ROC curve of each method on realigned data is substantially lower than that on the structurally aligned data, indicating the sensitivity to alignment. However, the CO model is the least vulnerable to realignment. On realigned sequences, the ROC curve of the CO model (dashed curve) still lies above those of all other methods (the WC and WCW curves lie below the mutual information curves and are not shown). Moreover, the ROC curve of the CO model on realigned data is still superior to those of all other models on the original (structural) alignment.

Sensitivity to Selected Sequences

The quality of RNA secondary interaction prediction also depends on the proper selection of sequences. Two questions regarding the choice of data arise: how many sequences are needed, and what phylogenetic branches of species are the most informative. To answer these questions, we varied the sequence data based on 2 criteria and evaluated the ROC curves of each method on different datasets.

We first randomly chose subsets of sequences from the original data. As expected, the ROC curves improve as more sequences are incorporated, but the improvement does not grow linearly with the size of the data. The sensitivity improves by 3-fold as the data increases from 10 to 20 species, 2-fold as the data increases from 20 to 50 species, but only improves by 30% as the data grows from 50 to 100 species. The ROC curve on 130 species is close to that on the entire data (146 species). However, the apparent 130-species limit is the result from subsampling the specific 146-species data set. It is not known whether a much larger sample (e.g., 500 species) would substantially improve the prediction accuracy. The ROC curves of the CO model on random subsets of sequences are shown in Supplementary Figure 3 (Supplementary Material online).

The sensitivity with respect to sample size is not independent from the representativeness of sequences on the phylogenetic tree. On the one hand, sequences concentrating on a narrow clade may lack covariation. On the other hand, structural variation of ribosomal RNAs over a wide range of clades may introduce noise in prediction results. To examine the dependency of predictions on representativeness of selected species in the phylogenetic tree, we then compared the ROC curve of the CO model on the original data with 3 subsets of data extracted from different phylogenetic branches: the data excluding mitochondria sequences, the data of eukaryotes and archaea, and the data of bacteria. Figure 6 shows the ROC curves on those data sets. The ROC curve on all-but-mitochondria data (dotted) overlaps with the ROC curve on the original data (solid). The insignificant contribution of mitochondria sequences can be due to its small size (9) compared with the entire data set, its structural variation relative to cytoplasmic rRNAs, or the accelerated evolutionary rates of mitochondria. The bacteria data (dashed) yield better performance than the eukaryotes/archaea data (dash dotted). The difference again can be attributed to either the sample size (77 versus 50) or the representativeness of the sequences. Despite the structural difference of rRNAs in prokaryotes, eukaryotes, and archaes, combining the sequences from three kingdoms substantially improves the accuracy compared to prokaryotes and eukaryotes sequences alone. This suggests that the signal from sequence covariation exceeds the noise from structural variation when widening the coverage of sequences in the kingdom of life.

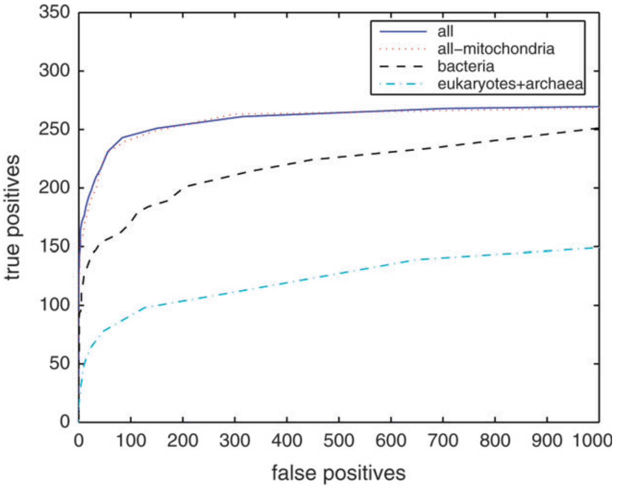

FIG. 6.—ROC curves of 16S rRNA interaction prediction with different clades; solid: the entire data set (146), dotted: nonmitochondria sequences (137), dashed: bacteria sequences (77), dashed dotted: eukaryotes and archaea sequences (50).

tRNA Prediction Accuracy

We applied the 4 models (CO, WC, WCW, mutual information) to predict secondary and tertiary interactions of methionine tRNA molecules. Overall, the false-positive numbers are much smaller than the 16S rRNA due to the shorter length of tRNA sequences. With less than 10 false positives, both CO and WCW models can detect almost all 21 secondary interactions. All the 3 parametric models (CO, WC, and WCW) outperformed the mutual information score in predicting secondary interactions. However, both the CO model and mutual information scores were able to detect several tertiary interactions that were neither Watson–Crick nor GU pairs. Details about the analysis results on tRNA data are reported in Supplementary File 1 (Supplementary Material online).

Programs and Running Time

All the prediction methods are implemented in C and compiled in Linux CentOS 4.4 Operating System. A screening on the 16S rRNA sequences (1,188,111 pairwise comparisons, 146 species) takes about 5 h on a Pentium(R) 4, 3 GHz CPU machine. The C codes and the 16S rRNA inputs and outputs of the programs are provided in Supplementary File 5 (Supplementary Material online).

Discussion

In this study, we propose a probabilistic graphical model to detect coevolution between interacting components. The model incorporates phylogenetic relations between species, sequence substitution rates for neutral mutation, selective constraints of interactions, and the spatial dependency between adjacent sites. The generality and simplicity of our model enable it to detect more complex interactions and alleviates the problem of overfitting the data.

A primary advantage of the CO model is its capacity in predicting RNA tertiary interactions. Among the 41 putative interactions on 16S rRNA, 23 demonstrate either direct contact or proximity in the 30S subunit of *T. thermophilus*. The results suggest that many so-called false positives are likely candidates for tertiary or indirect interactions.

Particularly, almost all the pairs exhibiting strong covariation patterns have direct contact, despite other contacting pairs having weaker covariation scores. This implies that coevolution is a strong indicator but not a necessary consequence of physical interactions.

Several pairs demonstrate strong covariation yet possess no apparent hydrogen bonds. We suspect that coevolution in these cases arises from the structural constraints beyond secondary and tertiary interactions. Examples include pair 1029–1032 in the tetraloop (which covaried between UG and GA) and pair 440–497 (which covaried between UU, AC, and CG). These structural constraints are worth pursuing.

In addition to Watson–Crick and GU pairs, the putative interactions contain the following common base pairs: GA, CC, UU, AC, CA, AA, and GG. The existence of the diverse nucleotide pair configurations confirms the complexity of tertiary interactions reported in previous studies. It also illustrates the power of a general CO model in detecting RNA tertiary interactions.

The simple CO model can successfully detect more than half of the secondary interactions of 16S rRNAs. The CO model is significantly better than the other 4 methods in predicting secondary interactions. The poor performance of WC and WCW models is likely due to the artifacts of not penalizing conserved non-WC (or WCW) states as discussed in Results. Mutual information selects spurious covariation due to common phylogeny. Marking the significance of mutual information reduces false positives but also increase false negatives. The multiple dependency score slightly outperforms mutual information, yet is still inferior to the CO model. In contrast, the CoMap program achieves a similar performance compared with the CO model. The results suggest that phylogenetic information is crucial to improve the detection.

All the methods tested in this work are sensitive to the quality of sequence alignment. As shown in figure 5, the ROC curve on realigned 16S rRNA data that are purely based on sequences (ClustalW) is substantially lower than that based on the structure. The CO model is the least affected by sequence alignment as its ROC curve on realigned data is still higher than all the other models on structurally aligned data. Yet the sensitivity of the CO model is still reduced by 40% on realigned data. The dependency on alignment quality may limit the applicability of any CO model to the RNAs with unknown structure. A more reliable approach is to adopt an iterative process of sequence alignment and structure prediction. Rather than making one-shot prediction on an unreliably aligned data, we can iteratively use the prediction to improve alignment and predict the structure on the improved alignment (Lescoute et al. 2005).

The quality of 16S rRNA interaction prediction also depends on the size of the sequence data and the representativeness in the phylogenetic tree. The marginal advantage of adding new sequences decreases as more sequences are included. In our specific data set, 130 sequences are the saturation limit as adding more sequences does not improve the prediction. In addition, representatives from bacteria, eukaryotes, and archaea are all needed in order to cover sufficient covariation in the secondary interactions. In contrast, mitochondria sequences are dispensable. Given the progress of sequencing technologies, data sets of hundreds of RNA sequences from diverse set of species will be available for the CO models in the near future.

The simple CO model can successfully detect all the secondary interactions and several tertiary interactions of methionine tRNAs (results presented in Supplementary File 1, Supplementary Material online). The 4 methods achieve similar performance in the tRNA data, yet the CO model is better than the other 3 methods when the number of false positives exceeds 5. Both the CO model and mutual information also detect several tertiary interactions that do not follow Watson–Crick or GU pairing rules, such as CA/UC covariation (pair 42–49) and AA/GG covariation (pair 30–82).

In this work, we apply the same substitution rate matrix to all the sites. In reality, the substitution rates can vary drastically across sites. Slowly evolving sites will not be detected by the CO model as it rewards double changes in the rate matrix. Rapidly evolving sites, in contrast, may induce false positives. A possible improvement is to divide the sequence into several regions and apply different rate matrices to different regions.

Another possible artifact of reweighting the substitution rate matrix is that it distorts the stationary marginal distribution of single nucleotides. Because there are only 2 free parameters but eight equalities to satisfy, it is generally infeasible to maintain the stationary marginal distribution. Also, because the CO model imposes the same reward to each double change, it may favor the sequence composition where many base pairs occur. This may explain some false-positive predictions.

We only consider pairwise interactions between 2 nucleotides. Higher order interactions are treated as aggregation of pairwise interactions. An example is a secondary interaction (16–30) and a tertiary interaction (30–82) in tRNA (Supplementary File 1, Supplementary Material online). Each pairwise covariation (16–30, 30–82, 16–82) was detected by the CO model. Screening more general high order interactions is computationally more involved as the size of the joint substitution rate matrix grows exponentially with the dimension.

In this work, we only concern the interactions of nucleic acids within one RNA molecule (tRNA or 16S rRNA). The main reason of choosing these molecules is the abundant sequence data and information about the molecular structures. They serve as a good test case for our model. The CO model, however, is not restricted to intramolecular RNA interactions. In the future, we plan to apply the model to predict the inter molecular RNA/DNA interactions such as between tRNAs and rRNAs, between micro RNAs, and their targets.

Because the CO model requires no knowledge about interaction rules and contains only a few extra free parameters independent of the size of the substitution rate matrix, it is natural to extend the model to other types of interactions such as intraprotein, protein–protein, protein–DNA, and protein–RNA interactions. Unlike RNA structure, these interactions are less well studied; hence, a large training data for parameter estimation is not readily available. Possible ways to resolve this problem include estimating the parameters from limited known training data (e.g., amino acid residues that are physically in contact and undergo compensatory substitution across species) or only reporting the results that are robust against a wide range of parameter settings. In addition to physical interactions, coevolution may arise from functional constraints beyond physical interactions. A general model of coevolution serves as a powerful tool to investigate a wide range of CO phenomena.

Supplementary Material

Supplementary Files 1–5 and Figure 3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Albion Baucom and Jürgen Brosius for valuable discussions and providing information about RNA secondary and tertiary interactions and the 16S rRNA sequence data. We thank Robert Baertsch for technical help with the *PyMOL* visualization software and Jacob Pedersen for discussion about the CO models. The UCSC Center for Genomic Science (National Institutes of Health/National Human Genome Research Institute grant 1 P41 HG02371-02) to C.H.Y.; the Leon Levy Foundation; the Simons Foundation at the Institute for Advanced Study; Microsoft Research scholarship to J.D.

Literature Cited

Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol. 17:164–178.

Barker D, Pagel M. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comp Biol. 1:24–31.

Bowers PM, Cokus SJ, Eisenberg D, Yeates TO. 2004. Use of logic relationship to decipher protein network organization. Science. 306:2246–2249.

Cannone JJ, Subramanian S, Schnare MN, et al. (14 co-authors). 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics. 3:2.

Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrel DM, Bandela AM, Cardenas E, Garrity GM, Tiedje JM. 2007. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Research. 35 (database issue):D169–D172. [Internet]. Available from: http://rdp.cme.msu.edu/.

Coventry A, Kleitman DJ, Berger B. 2004. MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. Proc Natl Acad Sci USA. 101:12102–12107.

di Bernardo D, Down T, Hubbard T. 2003. ddbRNA: detection of conserved secondary structures in multiple alignments. Bioinformatics. 19:1606–1611.

DeLano W. 2002. The PyMOL User's Manual. San Carlos, (CA): DeLano Scientific. [Internet]. Available from: http://pymol.sourceforge.net/.

Dutheil J, Pupko T, Jean-Marie A, Galtier N. 2005. A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol. 22:1919–1928.

Eddy SR. 2001. Non-coding RNA genes and the modern RNA world. Nat Rev Genet. 2:919–929.

Fares M, Travers SAA. 2006. A novel method for detecting intramolecular coevolution: adding a further dimension to select constraints analyses. Genetics. 173:9–23.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 17:368–376.

Felsenstein J, Churchill G. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol. 13:93–104.

Felsenstein J. 1989. PHYLIP-Phylogeny inference package. Cladistics. 5:164–166. [Internet]. Available from: http://evolution.genetics.washington.edu/phylip.html.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary fate in the protein interaction network. Science. 296:750–752.

Gloor GB, Martin LC, Wahl LM, Dunn SD. 2005. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry. 44: 7156–7165.

Goh CS, Bogan AA, Joachmiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with their interaction partners. J Mol Biol. 299:283–293.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. 2003. Rfam: an RNA family database. Nucleic Acids Research 31(1):439–441. Available from: http://www.sanger.ac.uk/cgi-bin/Rfam/getacc?RF00005.

Gutell RR, Noller HF, Woese CR. 1986. Higher order structure in ribosomal RNA. EMBO J. 5:1111–1113.

Hasegawa M, Kishino H, Yano T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22:160–174.

Hofacker IL, Fekete M, Flamm C, Huynen MA, Rauscher S, Stolorz PE, Stadler PF. 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. Nucleic Acids Res. 26:3825–3836.

Holmes I, Rubin GM. 2002. An expectation maximization algorithm for training hidden substitution models. J Comp Biol. 317:753–764.

Jordan MI. 1999. Learning in graphical models. Cambridge (MA): MIT Press.

Jordan IK, Marino-Ramfrez L, Wolf YI, Koonin EV. 2004. Conservation and coevolution in the scale-free human gene coexpression network. Mol Biol Evol. 21:2058–2070.

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

Leontis N, Stombaugh J, Westhof E. 2002. The non-Watson-Circk base pairs and their associated isostericity matrices. Nucleic Acids Res. 30:3497–3531.

Lescoute A, Leontis NB, Massire C, Westhof E. 2005. Recurrent structural RNA motifs, isostericity matrices and sequence alignments. Nucleic Acids Res. 33:2395–2409.

Lockless SW, Ranganathan R. 1999. Evolutionary conserved pathways of energetic connectivity in protein families. Science. 286:295–299.

Noller HF. 2005. RNA structure: reading the ribosome. Science. 309:1508–1514.

Noller HF, Woese CR. 1981. Secondary structure of 16S ribosomal RNA. Science. 212:403–411.

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc Lond B Biol Sci. 255:37–45.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comp Biol. 2:1–12.

Pollock DD, Taylor WR, Goldman N. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol. 287:187–198.

Ramani AK, Marcotte EM. 2003. Exploiting the co-evolution of interacting proteins to discover interaction specificity. J Mol Biol. 327:273–284.

Rivas E, Klein RJ, Jones TA, Eddy SR. 2001. Computational identification of noncoding RNAs in E. coli by comparative genomics. Curr Biol. 11:1369–1373.

Rzhetsky A. 1995. Estimating substitution rates in ribosomal RNA genes. Genetics. 141:771–783.

Saenger W. 1984. Principles of nucleic acid structure. New York: Springer-Verlag.

Schuwirth BS, Borovinskaya MA, Hau CW, Zhang W, Vila-Sanjurio A, Holton JM, Cate JH. 2005. Structures of the bacterial ribosome at 3.5 Å resolution. Science. 310: 827–834.

Siepel A, Haussler D. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. J Comput Biol. 11:413–428.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Tillier ERM, Lui WH. 2003. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics. 19:750–755.

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaver G, Eisen M, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci USA. 102: 5483–5488.

Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol. 23:1383–1390.

Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA. 102:2454–2459.

Wimberly BT, Brodersen DE, Clemons WM, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V. 2000. Structure of the 30S ribosomal subunit. Nature. 407:327–339.

Woese CR, Winker S, Gutell RR. 1990. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops. Proc Natl Acad Sci USA. 87:8467–8471.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics. 139:993–1005.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in BioSciences 13L555-556 (http://abacus.gene.ucl.ac.uk/software/paml.html).

Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JHD, Noller HF. 2001. Crystal structure of the ribosome at 5.5A resolution. Science. 292:883–896.

Peter Lockhart, Associate Editor

Accepted July 12, 2007