Title

Detecting the dependent evolution of biosequences

Journal

RESEARCH IN COMPUTATIONAL MOLECULAR BIOLOGY, PROCEEDINGS, 3909

ISSN

0302-9743

Authors

Darot, Jeremy
Yeang, Chen-Hsiang H
Haussler, David

Publication Date

2006

Peer reviewed

Detecting the Dependent Evolution of Biosequences

Jeremy Darot[2,3†] Chen-Hsiang Yeang[1†] David Haussler[1]

[†]contributed equally to this work

[1] Center for Biomolecular Science and Engineering, UC Santa Cruz
[2] Department of Applied Mathematics and Theoretical Physics, University of Cambridge
[3] EMBL - European Bioinformatics Institute

Abstract. A probabilistic graphical model is developed in order to detect the dependent evolution between different sites in biological sequences. 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 sequence data. The model is able to accurately identify the secondary structure of tRNA as well as several known tertiary interactions.

1 Introduction

Recent advances in systems biology and comparative genomics are providing new tools to study evolution from a systems perspective. Selective constraints often operate on a system composed of multiple components, such that these components evolve in a coordinated way. We use the term dependent evolution to denote the dependency of sequence evolution between multiple molecular entities. A molecular entity can be a protein, a non-coding RNA, a DNA promoter, or a single nucleotide or residue. Dependent evolution is prevalent in many biomolecular systems. Instances include neo-functionalization and pseudogene formation [1, 2], co-evolution of ligand-receptor pairs [3, 4], protein-protein interactions [5], residues contributing to the tertiary structure of proteins [6], and RNA secondary structure [7]. Understanding dependent evolution helps to predict the physical interactions and functions of biomolecules, reconstruct their evolutionary history, and further understand the relation between evolution and function.

In this work, we develop a computational method for detecting and characterizing dependent evolution in orthologous sequences of multiple species. Continuous-time Markov models of sequence substitutions encoding the dependent or independent evolution of two molecular entities are constructed. The spatial dependency of adjacent sites in the sequence is captured by a hidden Markov model (HMM) specifying the interaction states of sites. As a proof-of-concept demonstration, we apply the model to tRNA sequences and show that the method can identify their secondary and tertiary structure.

Models of co-evolution have been investigated in many previous studies. Some of these have demonstrated that the sequence substitution rates of proteins are correlated with their function [8] and relationships with other proteins, such as the number of interactions [5], their interacting partners [5], and their co-expressed genes [9]. The compensatory substitutions of RNA sequences have been used to predict RNA secondary structure [10–16, 7, 17]. Other studies have attempted to predict protein-protein interactions at the residue or whole-protein levels by using co-evolutionary models [3, 4, 6, 18]. We use a framework of continuous-time Markov models resembling those in [6, 18], although the assumptions and mathematical approaches are significantly different.

2 Methods

In this study we use both general and specific evolutionary models to detect the secondary and tertiary structure of tRNAs. These are well suited to a proof-of-concept demonstration since nucleotide pairs have fewer joint states than residue pairs ($4 \times 4 = 16$ compared to $20 \times 20 = 400$), their interaction rules are relatively simple (primarily Watson-Crick base pairing), the secondary and tertiary interactions of tRNAs are already mapped, and a large number of aligned tRNA sequences across many species are available.

The typical structure of the tRNA encoding methionine is shown in Fig. 1. It comprises four stems, three major loops and one variable loop. Each stem contains several nucleotide pairs forming hydrogen bonds (black bars in Fig. 1). Those base pairs typically conform with the Watson-Crick complementary rule (AU or GC). Several GU pairs also form weaker hydrogen bonds (GU wobble). In addition, nucleotide pairs that are distant in the secondary structure may also form tertiary interactions (dotted lines in Fig. 1). Unlike secondary interactions, tertiary interactions do not necessarily conform with the Watson-Crick rules or GU wobble.

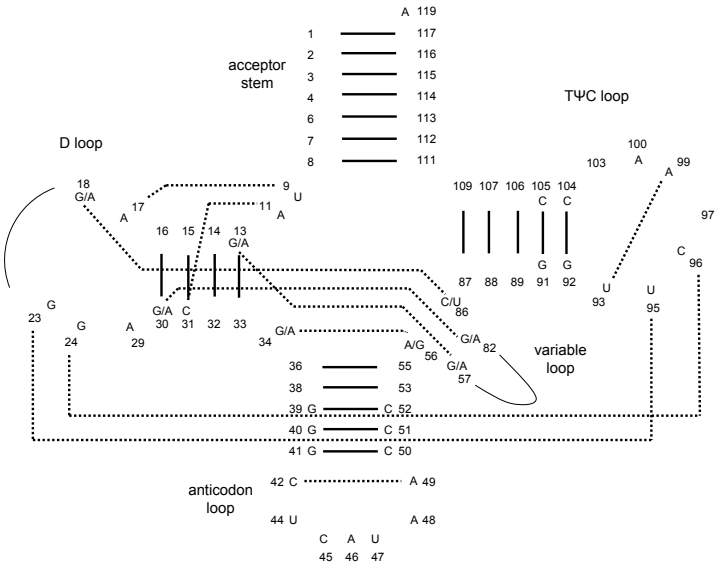
The co-evolutionary model that we developed is a probabilistic graphical model, operating on a given alignment of families of sequences for two molecular entities, along two orthogonal dimensions. The first dimension is time, with a continuous-time Markov process modeling the potentially coupled evolution of the two entities considered. This model operates at each position in the alignment, along a given phylogenetic tree. The second dimension is space, with an HMM operating along the sequence alignment and determining which regions of the two entities are co-evolving. Such graphical models were introduced by [19, 20] and have been recently adopted for instance by [21] to model the evolution of single molecular entities.

Consider first the sequence evolution model of a single nucleotide. It is a continuous-time Markov process with a substitution rate matrix $\mathbf{Q}$:

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

where $x(t)$ denotes the sequence at time t and $\mathbf{P}(x(t))$ a 1×4 probability vector of $x(t)$ being each nucleotide. $\mathbf{Q}$ is a 4×4 matrix with each row summed to

Fig. 1. tRNA secondary and tertiary structure



zero. Different rate matrices have been developed in the literature of molecular evolution. In this work we use the HKY model [22], which characterizes $\mathbf{Q}$ by a stationary distribution $\boldsymbol{\pi}$ and a transition/transversion ratio κ:

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

Each diagonal entry is the opposite of the sum of the other entries in the same row. The transition probability matrix $P(x(t)|x(0))$ is an entry of the matrix exponential of $\mathbf{Q}t$:

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

Given a phylogenetic tree and the length of its branches, the marginal likelihood of the observed sequence data at the leaves 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 [23]. Briefly, let u be a node in the tree, v and w its children, and t_v, t_w the branch lengths of $(u, v), (u, w)$. 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 is then defined by the recursion:

$$P(L_u|a) = \begin{cases} I(x_u = a) & \text{if } u \text{ is a leaf,} \\ \sum_b e^{\mathbf{Q}t_v}[a, b]P(L_v|b) \sum_c e^{\mathbf{Q}t_w}[a, c]P(L_w|c) & \text{otherwise.} \end{cases} \tag{4}$$

where $I(.)$ is the indicator function.

Now consider the sequence evolution model of a nucleotide pair. Define $\mathbf{x}(t) = (x_1(t), x_2(t))$ as the joint state of the nucleotide pair at time t. There are 16 possible joint states. The null model assumes that each nucleotide evolves independently with an identical substitution rate matrix. Therefore, the transition probability matrix is:

$$P(\mathbf{x}(t)|\mathbf{x}(0)) = (e^{\mathbf{Q}} \otimes e^{\mathbf{Q}})^t. \tag{5}$$

where $e^{\mathbf{Q}} \otimes e^{\mathbf{Q}}$ is the tensor product of two identical 4×4 matrices $e^{\mathbf{Q}}$. The outcome is a 16×16 matrix, specifying the transition probability of the joint state in a unit time. Each entry is the product of the corresponding entries in the single nucleotide substitution matrices. For instance,

$$\begin{aligned} & P(\mathbf{x}(1) = (C, G)|\mathbf{x}(0) = (A, U)) \\ = & P(x_1(1) = C|x_1(0) = A)P(x_2(1) = G|x_2(0) = U) \\ = & e^{\mathbf{Q}}[A, C] \cdot e^{\mathbf{Q}}[U, G]. \end{aligned} \tag{6}$$

The substitution rate $\mathbf{Q}_2 = log(e^{\mathbf{Q}} \otimes e^{\mathbf{Q}})$ of the nucleotide pair transitions in (5) is a 16×16 matrix, with the rates of single nucleotide changes identical to those in (2) and zero rates on double nucleotide changes. More precisely, if we denote (a, b) the joint state of a nucleotide pair:

$$\mathbf{Q}_2((a_1, a_2), (b_1, b_2)) = \begin{cases} \mathbf{Q}(a_1, b_1) & \text{if } a_2 = b_2, \\ \mathbf{Q}(a_2, b_2) & \text{if } a_1 = b_1, \\ -\mathbf{Q}(a_1, b_1) - \mathbf{Q}(a_2, b_2) & \text{if } a_1 = b_1, a_2 = b_2, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Equations (5) and (7) are equivalent, and the latter is discussed in [24]. Intuitively, if two nucleotides evolve independently, then during an infinitesimal time only one nucleotide can change, and the rate is identical to the single nucleotide transition rate.

The alternative model assumes that the evolution of the two nucleotides is coupled. One way to express their dependent evolution is to "reweight" the entries of the substitution rate matrix by a potential term ψ:

$$\mathbf{Q}_2^a = \mathbf{Q}_2 \circ \psi. \tag{8}$$

where ψ is a 16×16 matrix and $\circ$ denotes the following operation:

$$\mathbf{Q}_2(a, b) \circ \psi(a, b) = \begin{cases} \mathbf{Q}_2(a, b) \cdot \psi(a, b) & \text{if } a \neq b, \mathbf{Q}_2(a, b) > 0, \\ \psi(a, b) & \text{if } a \neq b, \mathbf{Q}_2(a, b) = 0, \\ -\sum_{b' \neq b} \mathbf{Q}_2(a, b') \circ \psi(a, b') & \text{if } a = b. \end{cases} \tag{9}$$

It multiplies an off-diagonal, nonzero entry $\mathbf{Q}_2(a, b)$ by $\psi(a, b)$, sets the value of a zero entry $\mathbf{Q}_2(a, b)$ as $\psi(a, b)$, and normalizes a diagonal entry as the opposite of the sum of the other entries in the same row. $\mathbf{Q}_2^a$ is a valid substitution rate matrix, thus its exponential induces a valid transition probability matrix.

We give (8) a mechanistic interpretation. The sequence substitution pattern of a co-evolving pair is the composite effect of neutral mutations, which occur independently at each nucleotide, and a selective constraint, which operates on the joint state. The potential term ψ rewards the state transitions that denote co-evolution and penalizes the others. We set the ratio between penalty and neutrality at ϵ, and the reward for simultaneous changes as r.

The choice of rewarded and penalized states is crucial. Here, we apply three different criteria to reweight the joint states. The first criterion rewards the state transitions that establish Watson-Crick base pairing from non-interacting pairs, penalizes the state transitions which break it, and is neutral for all other state transitions. We call it the "Watson-Crick co-evolution" or WC model. Specifically, the potential term is:

$$\psi(\mathbf{x}(0),\mathbf{x}(1)) = \begin{cases} \frac{1}{\epsilon} & \text{if } \mathbf{x}(0) \text{ is not WC and } \mathbf{x}(1) \text{ is WC,} \\ \epsilon & \text{if } \mathbf{x}(0) \text{ is WC and } \mathbf{x}(1) \text{ is not WC,} \\ 1 & \text{otherwise.} \end{cases} \tag{10}$$

The second criterion includes the GU/UG pairs (denoted GU since the order does not matter here) in the rewarded states. It thus rewards the state transitions that establish Watson-Crick or GU wobble base pairs, penalizes the state transitions which break the extended rule, and is neutral for all other state transitions. We call it the "Watson-Crick co-evolution with GU wobble" or WCW model. Specifically,

$$\psi(\mathbf{x}(0),\mathbf{x}(1)) = \begin{cases} \frac{1}{\epsilon} & \text{if } \mathbf{x}(0) \text{ is not WC or GU and } \mathbf{x}(1) \text{ is WC or GU,} \\ \epsilon & \text{if } \mathbf{x}(0) \text{ is WC or GU and } \mathbf{x}(1) \text{ is not WC or GU,} \\ 1 & \text{otherwise.} \end{cases} \tag{11}$$

The third criterion does not use prior knowledge of Watson-Crick base pairing and GU wobble and only considers the simultaneous changes of the two nucleotides ("simple co-evolution" or CO model). It rewards the state transitions where both nucleotides change, and penalizes the state transitions where only one nucleotide changes. Recall that the rates of simultaneous changes in the independent model are zero. Therefore, we reward these transitions not by reweighting their entries in $\mathbf{Q}_2$, but by giving them a positive rate r. Specifically,

$$\psi(\mathbf{x}(0),\mathbf{x}(1)) = \begin{cases} r & \text{if } x_1(1) \neq x_1(0) \text{ and } x_2(1) \neq x_2(0), \\ \epsilon & \text{if either } x_1(1) = x_1(0) \text{ or } x_2(1) = x_2(0), \\ 1 & \text{otherwise.} \end{cases} \tag{12}$$

The CO model assumes that the interacting nucleotide pairs maintain stable states. In order to transition from one stable state to another, both nucleotides must change. We introduce this general model in order to capture tertiary interactions for which pairing rules are complex or unknown. Moreover, since this general model incorporates no knowledge about nucleotide interactions and has only two extra free parameters (ϵ and r), it can be directly extended to more complicated problems such as protein-protein interactions or multi-way interactions.

We apply the dynamic programming algorithm described in (4) to evaluate the marginal likelihood of the nucleotide pair data. Specifically, a, b and c are the joint states of nucleotide pairs and $e^{\mathbf{Q}t}$ is defined as in (5) for the null model and as the exponential of (8) times t for the alternative model.

In order to incorporate the spatial dimension of the nucleotide sequence into the model, we define an HMM for the "interaction states" of the aligned sequences. Suppose that the sequences of two molecular entities are aligned (e.g., a tRNA sequence is aligned with itself in the opposite direction) across all species. We define the "interaction state" $y(s)$ of the sequence pair at alignment position s as a binary random variable, indicating whether co-evolution occurs at position s (i.e., $y(s) = 1$) or not ($y(s) = 0$). The $y(s)$'s are the hidden variables of the HMM. Their transitions are 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 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 co-evolution. The likelihoods are evaluated by the aforementioned dynamic programming algorithm. Given the transition and emission probabilities, we apply the Viterbi algorithm to identify the interacting regions of the two sequences.

Issues arise when there are gaps in the aligned sequences. If "sparse" gaps appear at scattered positions in a few species, we treat them as missing data, by giving an equal probability to each nucleotide. If there are consistent gaps appearing in consecutive regions over many species, we ignore those regions when calculating the likelihood scores.

In order to quantify the confidence of the inferred interaction states, we used the log-likelihood ratio (LLR) between the co-evolutionary model and the null model, at each position within the Viterbi algorithm. Pollock et al. [6] have pointed out that a χ^2 distribution is not appropriate for such co-evolutionary models. For this reason, we have not reported the p-values that might have otherwise been calculated from a χ^2 distribution with one (WC and WCW models) or two (CO model) extra degrees of freedom.

3 Results

We applied our model to the methionine tRNA sequences of 60 species covering the three superkingdoms of life. Three different criteria were used to reward and penalize the joint state transitions in the model of dependent evolution: Watson-Crick base pairing, Watson-Crick base pairing with GU wobble, simultaneous changes. We compared the performance of each model in detecting secondary and tertiary interactions, and further investigated false positives and false negatives.

3.1 Data and Pre-processing

Aligned tRNA sequences were downloaded from the Rfam database [25]. Unique sequences for the methionine tRNA (tRNA-Met, ATG codon) were extracted for

60 species, including archea, bacteria, eukaryotes and their organelles (mitochondria and chloroplast). The length of the complete sequence alignment including gaps was $lseq = 119$ nucleotides. A phylogenetic tree was derived from these sequences using a Metropolis-coupled Markov chain Monte-Carlo (MC^3) simulation implemented in the *MrBayes* program [26]. The resulting tree was found to be robust and consistent with the tree topologies obtained by parsimony using the *DNAPARS* program of the *PHYLIP* package [27]. The phylogenetic tree of the tRNA data is reported in the supplementary materials.

The tRNA sequence was then paired with itself in the opposite direction in order to evaluate potential co-evolution between all possible nucleotide pairs. The first entity in the model was the tRNA sequence itself, and the second entity was the reversed sequence, shifted by a number of nucleotides varying from 1 to $lseq$, and "rolled over" to match the length of the first entity. The co-evolutionary signal, which is the Viterbi path of the HMM, was then plotted as a $lseq \times lseq$ matrix, where the x-axis represents the position in the sequence, and the y-axis the offset. As an example, the expected signal for the structure depicted in Fig.1 is shown in Fig. 2. The figure comprises four symmetric patterns, which correspond to the four stems of the tRNA secondary structure (in yellow): acceptor stem at offset 2 and 3, anticodon stem at offset 30, TΨC stem at offset 44 and 45, and D stem at offset 75. The tertiary structure appears as symmetric isolated nucleotide pairs (in green). The patterns are not symmetric with respect to the diagonal line due to a gap between positions 60 and 80 covering padding to the variable loop.

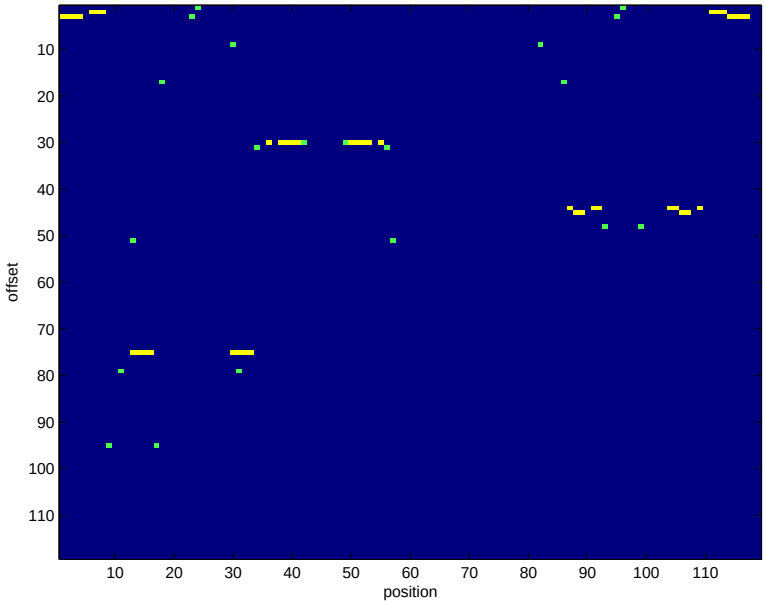
3.2 Sensitivity Analysis

A sensitivity analysis was carried out, varying ϵ from 10^{-3} to 0.90, r from 0 to 0.5, and α from 0.05 to 0.45 (results not shown). It was found that the performance of the different methods depends on a reasonable choice of parameter values. Indeed, the co-evolutionary models merge with the independent model for $\epsilon = 1$ and $r = 0$, therefore no signal can be detected for these parameter values. Conversely, excessively small values for ϵ and large values for r compromise the performance of the analysis. The parameter α can be seen as a spatial "smoothing" factor, which tends to eliminate isolated hits as its value decreases. This can help to eliminate isolated false positives from the contiguous secondary structure signal, but can also prevent the identification of isolated tertiary interactions. We henceforth report the results for $\epsilon = 0.5$, $r = 0.05$ and $\alpha = 0.2$.

3.3 Watson-Crick Co-evolution

The co-evolutionary signal detected by the WC model is shown as a ROC curve and at a particular cutoff LLR value of 5.0 in Fig. 3. At this level of significance, 20 out of 21 secondary interactions were identified (in orange), and 4 out of 10 tertiary interactions (in red), resulting in 22 false positives (in light blue). The "missing" secondary interaction, between nucleotides 36 and 55, shows evidence

Fig. 2. Expected signal for the tRNA secondary and tertiary structure



of GU wobble, which can be contrasted with the purely Watson-Crick base pairing of the true positive pair 39-52 (Table 1). The WC model is not suited to the detection of such an interaction, though it is eventually picked up at a much lower significance level (Fig. 3(a)).

Table 1. Dinucleotide composition of one (a) true positive (b) false negative secondary interaction, WC model

39-52	A	C	G	U
A	0	0	0	14
C	0	0	5	0
G	0	39	0	0
U	2	0	0	0

36-55	A	C	G	U
A	1	0	0	2
C	0	0	18	3
G	0	1	0	2
U	25	0	8	0

As expected, the four tertiary interactions identified by the WC model (Table 2) are mainly Watson-Crick, even though pairs 24-96 and 93-99, which are detected at a comparatively lower significance level, have some non-negligible terms off the second diagonal.

Many of the false positives seem to be vertically aligned in Fig. 3(b). A closer examination reveals that these are composed of nucleotides which are highly conserved individually, and appear to form a Watson-Crick pair without physically

Fig. 3. Results from the WC model (a) ROC curves (b) signal at a $LLR = 5.0$ cutoff

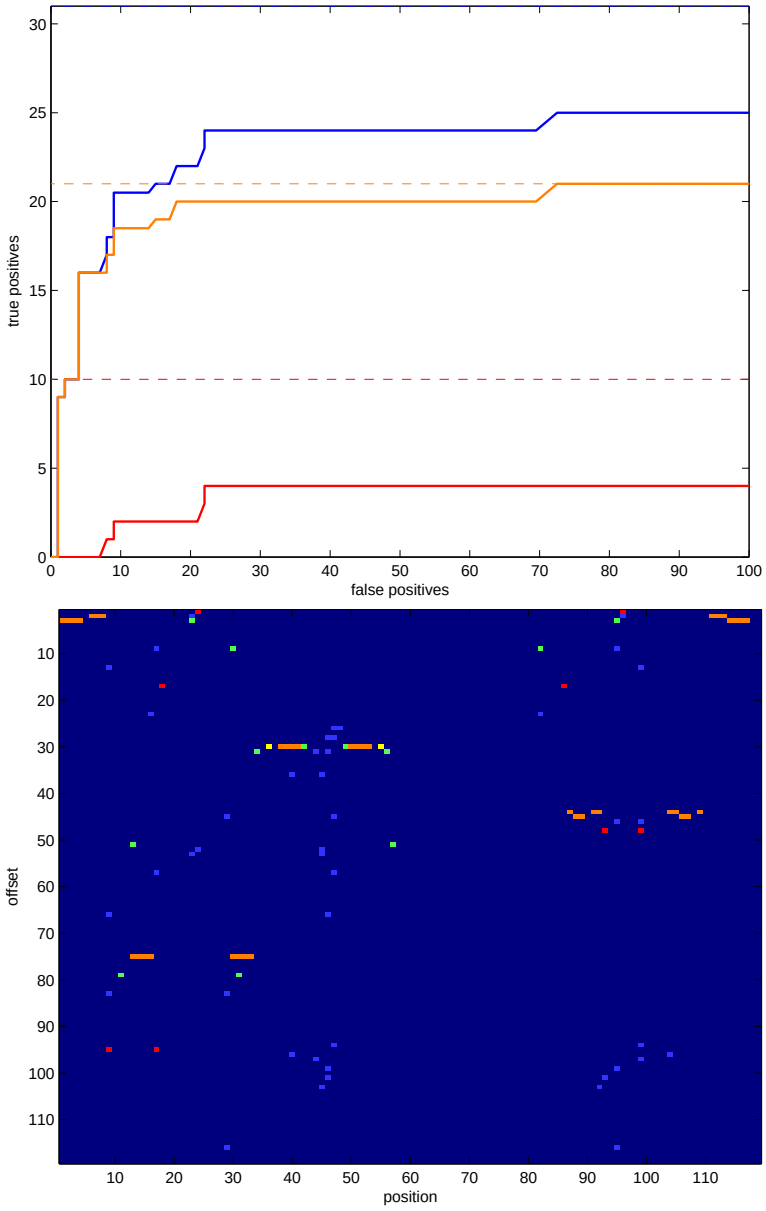


Table 2. Dinucleotide composition of detected tertiary interactions, WC model

9-17	A	C	G	U
A	0	0	0	0
C	0	0	0	0
G	0	0	0	2
U	58	0	0	0

18-86	A	C	G	U
A	0	1	1	24
C	0	0	0	0
G	0	30	0	1
U	1	0	0	0

24-96	A	C	G	U
A	2	0	0	0
C	2	0	0	0
G	0	44	0	0
U	6	0	0	4

93-99	A	C	G	U
A	11	0	0	0
C	0	0	0	0
G	0	0	0	0
U	48	0	0	1

interacting. In particular, the constant nucleotides of the CAU anticodon at positions 45-47 form spurious Watson-Crick base pairs with other highly conserved nucleotides in the different loops of the tRNA structure.

3.4 Watson-Crick Co-evolution with GU Wobble

The co-evolutionary signal detected by the WCW model is shown as a ROC curve and at a particular cutoff LLR value of 5.8 in Fig. 4. At this level of significance, 19 out of 21 secondary interactions were identified, and 1 out of 10 tertiary interactions, for only 2 false positives. The much steeper ROC curve for secondary interactions demonstrates the benefit of incorporating additional biochemical knowledge into the model. Indeed, as many secondary interactions involve some degree of GU wobble, they are detected earlier by the WCW model than they were by the WC model. In contrast, the identification of tertiary interactions does not benefit from the refined model, because those rarely involve GU wobble. The only exception is the 23-95 pair, which involves GU wobble, but it is only detected for more than 200 false positives (beyond the boundaries of Fig. 4(a)).

3.5 Simple Co-evolution Model

The co-evolutionary signal detected by the CO model is shown as a ROC curve and at a particular cutoff LLR value of 0.8 in Fig. 5. At this level of significance, all secondary interactions were identified, and 3 out of 10 tertiary interactions, yielding 25 false positives. The tertiary interactions detected by the CO model include the pairs 9-17 and 18-86, which were also identified by the WC and WCW models. However, neither the 24-96, 93-99 nor 23-95 interactions were identified, as for those pairs one nucleotide often varies while the other remains constant (Table 2). Additionally, the 42-49 interaction was identified by the CO model, which had not been detected by the WC and WCW models because it consists mainly of C-A and U-C pairs.

3.6 Detection of Tertiary Interactions: Summary

Figure 6 highlights the tRNA-Met tertiary interactions that have been detected using one of the three co-evolutionary models. Among the ten annotated tertiary interactions, six were identified by at least one of the models: 9-17 and 18-86 (all three models, solid blue), 24-96 and 93-99 (WC and WCW models, solid

Fig. 4. Results from the WCW model (a) ROC curves (b) signal at a $LLR = 5.8$ cutoff

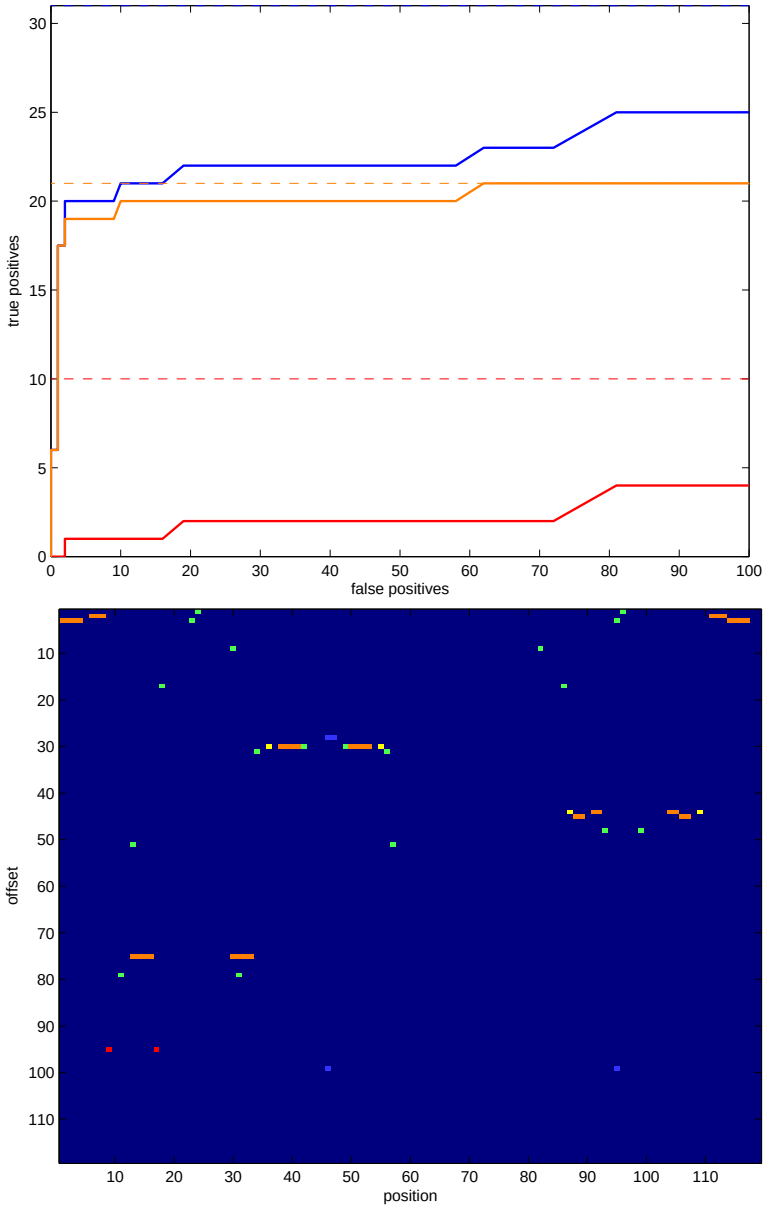


Fig. 5. Results from the CO model (a) ROC curves (b) signal at a $LLR = 0.8$ cutoff

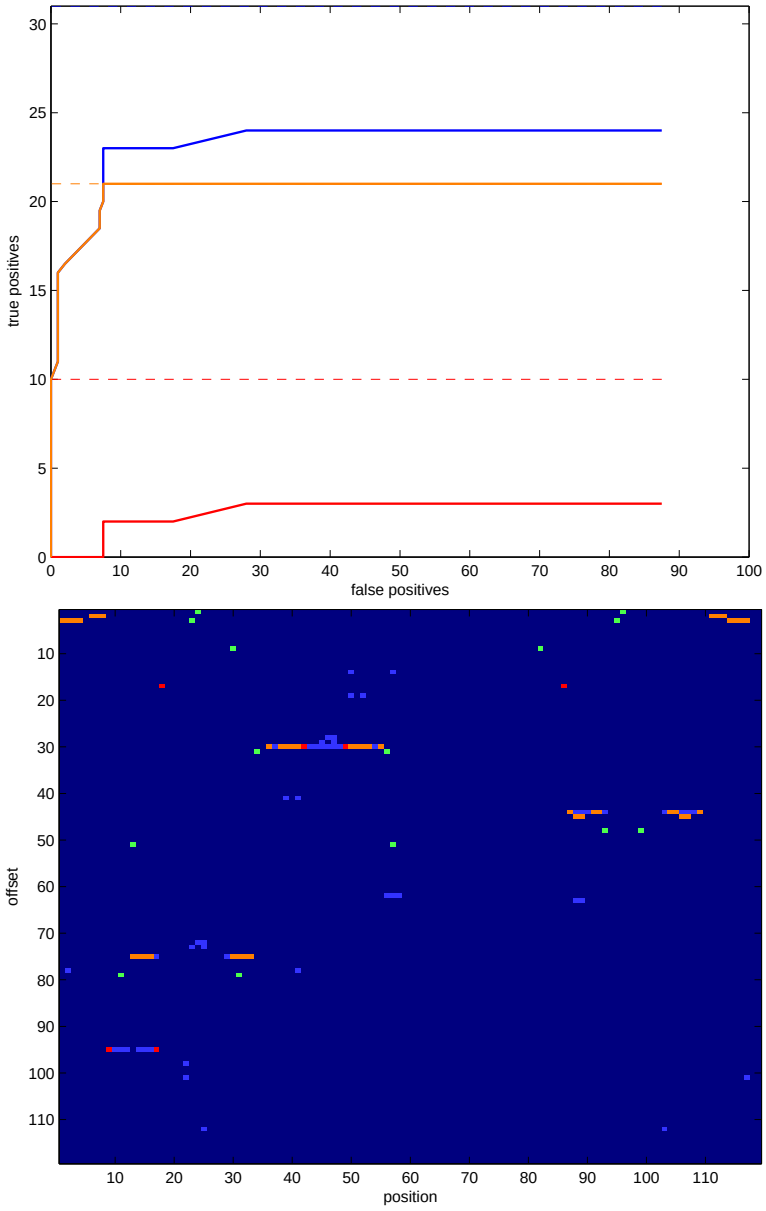
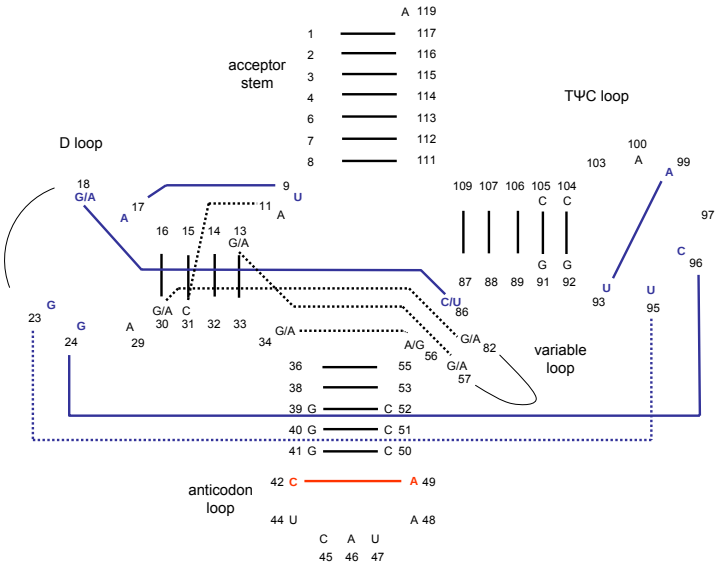


Fig. 6. Detection of tertiary interactions



blue), 23-95 (WCW model, dotted blue), 42-49 (CO model, solid red). The four remaining interactions (dotted black) were not detected by any model. With the possible exception of 30-82, none of those pairs shows a particular bias towards Watson-Crick base pairing or simultaneous evolution in their dinucleotide composition, so the failure to detect them using the aforementioned models is not surprising.

4 Discussion

We have shown that a probabilistic graphical model incorporating neutral mutations, selective constraints and sequence adjacency can successfully identify the secondary and tertiary interactions in a tRNA structure.

The comparison of the results of the WC and WCW models indicates a trade-off between generality and performance. Indeed, increasing the specificity of the model by incorporating more biological knowledge significantly improves the detection of the secondary structure. However, the increased specificity of the WC and WCW models causes them to miss a non-Watson Crick tertiary interaction, which is detected by the much more general CO model. Given this trade-off, the performance of the CO model turns out to be surprisingly good for both secondary and tertiary interactions, and suggests that rewarding non-specific simultaneous changes is a simple, yet powerful approach. This result is encouraging when one considers using such probabilistic graphical models to

investigate the co-evolution of more complex molecular systems, for which the interaction rules are not well characterized and the number of joint states is much larger, e.g., between proteins and nucleic acids.

Currently the parameters of the models – ϵ, r, α and the LLR cutoff – are set empirically. A more systematic way of estimating them from the data and testing the model in cross-validation would be a useful extension of this work.

Some scenarios beyond co-evolution may also be captured by this modeling framework. For instance, instead of rewarding simultaneous changes and penalizing unilateral changes, we can invert the potential term to reward unilateral changes and penalize simultaneous changes. A possible interpretation of this scenario is that the two entities are complementary in function, such as paralogous genes after their duplication. The conservation of one gene allows the evolution of the other, which can acquire a new function. A change in both genes, however, is likely to be detrimental to their original functions and thereby reduces the fitness.

Supplementary Materials

The phylogenetic tree of the tRNA data across 60 species and the statistics of dinucleotide composition of all the tertiary interactions are reported in http://www.soe.ucsc.edu/~chyeang/RECOMB06/.

Acknowledgements

We thank Harry Noller for helpful discussions and providing information about tRNA secondary and tertiary interactions. CHY is sponsored by an NIH/NHGRI grant of UCSC Center for Genomic Science (1 P41 HG02371-02) and JD was sponsored by a Microsoft Research scholarship.

References

1. Ohno, S.: Evolution by gene duplication. Springer-Verlag, Heidelberg, Germany, 1970
2. Lynch, M., Conery, J.S.: The evolutionary fate and consequences of duplicated genes. Science **290** (2000) 1151–1155
3. Goh, C.S., Bogan, A.A., Joachmiak, M., Walther, D., Cohen, F.E.: Co-evolution of proteins with their interaction partners. J. Mol. Biol. **299** (2000) 283–293
4. Ramani, A.K., Marcotte, E.M.: Exploiting the co-evolution of interacting proteins to discover interaction specificity. J. Mol. Biol. **327** (2003) 273–284
5. Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., Feldman, M.W.: Evolutionary fate in the protein interaction network. Science **296** (2002) 750–752
6. Pollock, D.D., Taylor, W.R., Goldman, N.: Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. **287** (1999) 187–198
7. Washietl, S., Hofacker, I.L., Stadler, P.F.: Fast and reliable prediction of noncoding RNAs. PNAS **102** (2005) 2454–2459

8. Wall, D.P., Hirsh, A.E., Fraser, H.B., Kumm, J., Giaver, G., Eisen, M., Feldman, M.W.: Functional genomic analysis of the rates of protein evolution. PNAS **102** (2005) 5483–5488
9. Jordan, I.K., Marino-Ramfrez, L., Wolf, Y.I., Koonin, E.V.: Conservation and coevolution in the scale-free human gene coexpression network. Mol. Biol. Evol. **21** (2004) 2058–2070
10. Noller, H.F., Woese, C.R.: Secondary structure of 16S ribosomal RNA. Science **212** (1981) 403–411
11. Hofacker, I.L., Fekete M., Flamm, C., Huynen, M.A., Rauscher, S., Stolorz, P.E., Stadler, P.F: Automatic detection of conserved RNA structure elements in complete RNA virus genomes. Nucleic Acids Res. **26** (1998) 3825-3836
12. Eddy, S.R.: Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet. **2 (2001)** 919–929
13. Rivas, E., Klein, R.J., Jones, T.A., Eddy, S.R.: Computational identification of noncoding RNAs in E. coli by comparative genomics. Curr. Biol. **11** (2001) 1369-1373
14. di Bernardo, D., Down T., Hubbard, T.: ddbRNA: detection of conserved secondary structures in multiple alignments. Bioinformatics **19** (2003) 1606-1611
15. Coventry, A., Kleitman D.J., Berger, B.: MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. PNAS **101** (2004) 12102-12107
16. Pedersen, J.S., Meyer, I.M., Forsberg, R., Simmonds, P., Hein, J.: A comparative method for finding and folding RNA secondary structures within protein-coding regions. Nucleic Acids Res. **32** (2004) 4925–4936
17. Washietl, S., Hofacker I.L., Lukasser, M., Huttenhofer, A., Stadler, P.F.: Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat. Biotechnol. **23** (2005) 1383–1390
18. Barker, D., Pagel, M.: Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comp. Biol. **1** (2005) 24–31
19. Yang, Z.: A space-time process model for the evolution of DNA sequences. Genetics **139** (1995) 993–1005
20. Felsenstein, J., Churchill, G.: A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13** (1996) 93–104
21. Siepel, A., Haussler, D.: Combining phylogenetic and hidden Markov models in biosequence analysis. JCB **11** (2004) 413–428
22. Hasegawa, M., Kishino, H., Yano, T.: Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22** (1985) 160–174
23. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17** (1981) 368–376
24. Pagel, M.: Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society in London, series B, **255** (1994) 37–45.
25. RNA families database. http://www.sanger.ac.uk/cgi-bin/Rfam/getacc?RF00005
26. MrBayes: Bayesian inference of phylogeny. http://mrbayes.csit.fsu.edu/index.php
27. http://evolution.genetics.washington.edu/phylip.html