Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

OXFORD

Structural bioinformatics

PhyloFold: Precise and Swift Prediction of RNA Secondary Structures to Incorporate Phylogeny among Homologs

Masaki Tagashira [1,2,*]

[1] Department of Computational Biology and Medical Sciences, University of Tokyo, Chiba, 277-8561, Japan and
[2] Artificial Intelligence Research Center, AIST, Tokyo, 135-0064, Japan.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The *simultaneous consideration of sequence alignment and RNA secondary structure*, *structural alignment*, has been known to help predict more accurate secondary structures of homologs. However, the consideration is heavy and has been done roughly to decompose structural alignments.
Results: The PhyloFold method, which predicts secondary structures of homologs considering likely pairwise structual alignments between the homologs, is invented in this study. The method shows the best prediction accuracy while demanding comparable running time compared with competitors.
Availability: The source code of the programs implemented in this study is available on "https://github.com/heartsh/phylofold" and "https://github.com/heartsh/phyloalifold".
Contact: "tagashira_masaki_17@stu-cbms.k.u-tokyo.ac.jp".
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A large quantity of functional *ncRNA*s have been discovered through sequencing technology including high-throughput sequencing technology (Maxam and Gilbert, 1977; Bentley *et al.*, 2008). The RNAs are related to various biological processes such as epigenetic silencing (Pasmant *et al.*, 2011), splicing regulation (Ji *et al.*, 2003), translational control (Long and Caceres, 2009), apoptosis regulation, and cell cycle control (Kino *et al.*, 2010). However, the functions of a large number of RNAs have been not yet uncovered. Secondary structures of the RNAs are a considerable key to uncovering the functions of the RNAs because the structures play roles in biological processes (Wu *et al.*, 1991) and secondary structures are often conserved among homologs even in the case where the sequences of the homologs are unconserved. (Klein and Eddy, 2003)

To predict more correct secondary structures, the simultaneous consideration of sequence alignment and secondary structure, *structural alignment* (Sankoff, 1985; Hofacker *et al.*, 2004; Havgaard *et al.*, 2007), has been taken into account. (Hamada *et al.*, 2009c, 2011; Tan *et al.*, 2017) However, the consideration has been proven intractable (Sankoff, 1985) and thus has been conducted roughly to decompose structural alignments. (Hamada *et al.*, 2009c,a; Sato *et al.*, 2012; Tan *et al.*, 2017) To accomplish exactly the consideration, the PhyloFold method (Figure 1) is invented in this study. The method predicts secondary structures of homologs considering likely pairwise structual alignments between the homologs by a *sparsification* technique.

2 Methods

2.1 Pairwise structural alignment

Let $\boldsymbol{SS_{RNA}}, \boldsymbol{SA}_{\mathbb{RNA}}, \mathbb{STA}_{\mathbb{RNA}}$ be the (single) RNA secondary structure of the RNA sequence $\boldsymbol{RNA}$, (pairwise) sequence alignment between the pair of RNA sequences $\mathbb{RNA}$, and structural alignment between the pair $\mathbb{RNA}$ respectively. An alignment $\mathbb{STA}_{\mathbb{RNA}}$ is composed of the alignment and structures $\boldsymbol{SA}_{\mathbb{RNA}}, \boldsymbol{SS_{RNA}}, \boldsymbol{SS_{RNA'}}$, i.e. $\mathbb{STA}_{\mathbb{RNA}} = (\boldsymbol{SS_{RNA}}, \boldsymbol{SS_{RNA'}}, \boldsymbol{SA}_{\mathbb{RNA}})$ where $\mathbb{RNA} = (\boldsymbol{RNA}, \boldsymbol{RNA'})$. (Figure 2) Assume that a pair of positions must be pairing if it is aligned with another pair of pairing positions. Also assume secondary structures without any *pseudoknot*s, *collinear* sequence alignments, and *indels of 2-loop*s (Supplementary section 1.1) because their considerations lead to larger computational complexities (Sato *et al.*, 2011; Bradley *et al.*, 2009; Sankoff, 1985). Two pairs of positions are said

© The Author 2018.

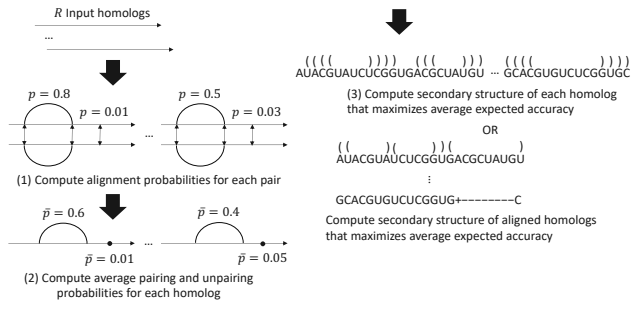

Fig. 1. The workflow of the PhyloFold method proposed in this study, which predicts single and consensus RNA secondary structures. (1) The method computes pair alignment probabilities on pairwise structural alignment for each pair of input homologs. (2) Then, the method calculates average pairing probabilities between a target homolog and multiple support homologs to incorporate more than one support homologs into the subsequent predictions of secondary structures. (3) Finally, the method predicts the single secondary structure of each input homolog or the consensus secondary structure of the input sequence alignment.

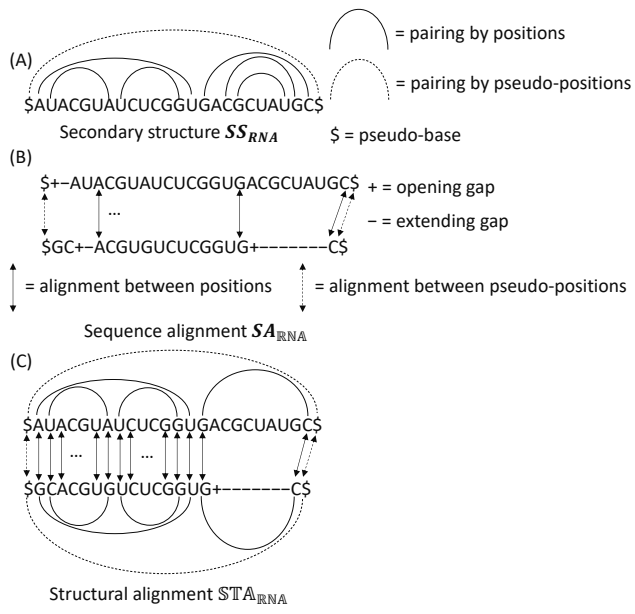

Fig. 2. Examples of (A) a secondary structure SS_{RNA}, (B) a sequence alignment $SA_{\mathbb{RNA}}$, and (C) a structural alignment $\mathbb{STA}_{\mathbb{RNA}}$.

to be *pair-aligned* if they are pairing and aligned. Two positions are said to be *loop-aligned* if they are unpaired and aligned.

2.2 Posterior *pair alignment probability* matrix

Let $\mathfrak{STA}_{\mathbb{RNA}}, fe_{\mathbb{STA}}$ be a set of all possible alignments $\mathbb{STA}_{\mathbb{RNA}}$ and the free energy (or inverse of the score) of the alignment $\mathbb{STA}$ respectively. Assume that the probability of any alignment $\mathbb{STA} \in \mathfrak{STA}_{\mathbb{RNA}}$, $p_{\mathbb{STA}}$, obeys a *Boltzmann (probability) distribution*, i.e. $p_{\mathbb{STA}} = \frac{\exp(-\beta fe_{\mathbb{STA}})}{Z}$ where $Z = \sum_{\mathbb{STA}} \exp(-\beta fe_{\mathbb{STA}})$ and β is a positive real number. Z is called a *partition function*. β scales any energy $fe_{\mathbb{STA}}$.

Let $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$ be the *pair alignment probability* matrix given the pair $\mathbb{RNA}$. Let $I^{\mathrm{PA}}_{\mathbb{STA}ijkl}$ be 1 if the pairs $(i,j),(k,l)$ are pair-aligned in the alignment $\mathbb{STA}$ and 0 otherwise where i,j are two positions in the sequence $\boldsymbol{RNA}$, k,l are two positions in the other sequence $\boldsymbol{RNA'}$, and $i<j, k<l$. The matrix $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$ can be written by the probabilities $p_{\mathbb{STA}}$: $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}} = (p^{\mathrm{PA},\mathbb{RNA}}_{ijkl})$ where $p^{\mathrm{PA},\mathbb{RNA}}_{ijkl} = \sum_{\mathbb{STA}|I^{\mathrm{PA},\mathbb{STA}}_{ijkl}=1} p_{\mathbb{STA}}$. $p^{\mathrm{PA},\mathbb{RNA}}_{ijkl}$ is the probability that the pairs $(i,j),(k,l)$ are pair-aligned.

Any computation methods of a matrix $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$ have not been proposed although the concept of it $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$ is proposed in Hamada *et al.* (2009c). Therefore, a computation method of the matrix $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$ and an efficient sparse version of the method are proposed in this study.

2.3 Composition of structural alignment free energy $fe_{\mathbb{STA}}$

Let $fe_{\boldsymbol{SS}}$ be the free energy of the structure $\boldsymbol{SS}$. The energy $fe_{\mathbb{STA}_{\mathbb{RNA}}}$ is decomposed into additional components: $fe_{\mathbb{STA}_{\mathbb{RNA}}} = fe_{\boldsymbol{SS_{RNA}}} + fe_{\boldsymbol{SS_{RNA'}}} + fe^{\mathrm{PA}} + fe^{\mathrm{LA}} + fe^{\mathrm{gap}}$. Here, $fe^{\mathrm{PA}}, fe^{\mathrm{LA}}$ are the sum free energy of the pair and loop alignments in the alignment $\mathbb{STA}_{\mathbb{RNA}}$ respectively. Also, fe^{gap} is the sum free energy (or penalty) of affine gaps in the alignment $\mathbb{STA}_{\mathbb{RNA}}$.

The components $fe_{\boldsymbol{SS_{RNA}}}, fe_{\boldsymbol{SS_{RNA'}}}$ can be computed by the estimated parameters of the Turner (nearest neighbor) model, which approximates the free energy of RNA secondary structure on thermodynamics. (Turner and Mathews, 2010) As the components, conventional methods based on structural alignment employ posterior pairing probability matrices on single secondary structure, predicted by *inside-outside algorithm*s such as the McCaskill algorithm (McCaskill, 1990), to simplify computations though the suitability of the matrices has not been discussed. (Hofacker *et al.*, 2004; Do *et al.*, 2008). Hence, the Turner model, instead of the matrices, is adopted in this study to prevent the accuracy of a matrix $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$ from losing.

The components $fe^{\mathrm{PA}}, fe^{\mathrm{LA}}$ can be computed by the RIBOSUM score matrices, predicted from learning datasets of validated structural alignments. (Klein and Eddy, 2003) In this study, the RIBOSUM score 80-65 matrices, which are the most popular, are adopted.

In the next section, an efficient (however nevertheless impractical) method that computes a matrix $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$ in the framework of inside-outside algorithm is proposed.

2.4 Inside-outside algorithm that computes pair alignment probability matrix $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$

A probability $p^{\mathrm{PA},\mathbb{RNA}}_{ijkl}$ can be written with two "inside" and one "outside" partition functions $\alpha^{\mathrm{EL,for}}_{NM}, \alpha^{\mathrm{PA}}_{ijkl}, \beta^{\mathrm{PA}}_{ijkl}$: $p^{\mathrm{PA},\mathbb{RNA}}_{ijkl} = \frac{\alpha^{\mathrm{PA}}_{ijkl}\beta^{\mathrm{PA}}_{ijkl}}{\alpha^{\mathrm{EL,for}}_{NM}}$ where N, M are the lengths of the sequences $\boldsymbol{RNA}, \boldsymbol{RNA'}$ respectively. Inside and outside partition functions can be computed with those of shorter and longer substrings respectively and are stored in dynamic programming memory for remaining computation. A matrix $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$ can be computed by Algorithm 1 with the time and space complexities $O(N^4M^4), O(N^3M^3)$. Variants of Algorithm 1 with the time and space complexities $O(N^3M^3), O(N^2M^2)$ can be considered though the sparsification proposed in this study is not effective to the variants.

Algorithm 1 is the "simultaneous" solution of the Durbin (*forward-backward*) algorithm, which predicts posterior alignment probability matrices on pairwise sequence alignment (Durbin *et al.*, 1998), and the McCaskill algorithm (McCaskill, 1990), as expected. Algorithm 1 is also an inside-outside algorithm version of the Sankoff algorithm, which predicts pairwise structural alignments whose free energy is minimum (Sankoff, 1985), as expected. The time and space complexities of Algorithm 1 are preferably around $O(NM)$ to deal with long ncRNAs. In the next section, a solution to the heavy complexity, sparsifying all possible structural alignments $\mathfrak{STA}_{\mathbb{RNA}}$, is introduced.

Algorithm 1 An inside-outside algorithm that computes a pair alignment probability matrix $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$.

1: **function** inOutAlgo($\mathbb{RNA}$)
2: Compute inside partition functions $\{\alpha^{x}_{ijkl}, \alpha^{y}_{uv}\}$ on dynamic programming by Supplementary section 1.3 // Inside step.
3: Compute outside partition functions $\{\beta^{x}_{ijkl}\}$ on the method by Supplementary section 1.4 // Outside step.
4: Compute pair alignment probabilities $\{p^{\mathrm{PA},\mathbb{RNA}}_{ijkl}\}$ by $p^{\mathrm{PA},\mathbb{RNA}}_{ijkl} = \frac{\alpha^{\mathrm{PA}}_{ijkl}\beta^{\mathrm{PA}}_{ijkl}}{\alpha^{\mathrm{EL,for}}_{NM}}$ // Final step.
5: **return** $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$

2.5 Sparsifying pair alignment probability matrix $\boldsymbol{P}^{\mathrm{PA}}_{\mathbb{RNA}}$

From all the possible alignments $\mathfrak{STA}_{\mathbb{RNA}}$, *sparsification* can pick favorable those (e.g. with adequately low free energy) out. It can allow Algorithm 1 to compute only the functions $\{\alpha^{x}_{ijkl}, \alpha^{y}_{uv}, \beta^{x}_{ijkl}\}$ that satisfy sparsification conditions. Let $p^{\mathrm{pair},\boldsymbol{RNA}}_{ij}$ be the pairing probability of the positions i, j given the sequence $\boldsymbol{RNA}$. In this study, the following sparsification conditions are introduced:

- $|u - v| \leq \delta, |(M - u) - (N - v)| \leq \delta$ for any positions u, v
- $|(i - j) - (k - l)| \leq \delta$ for any pairs $(i, j), (k, l)$
- $p^{\mathrm{pair},\boldsymbol{RNA}}_{ij} \geq \epsilon$ for any pairing positions i, j and any sequence $\boldsymbol{RNA}$

where δ, ϵ are sparsification parameters. The first two conditions let Algorithm 1 not consider all the alignment $\mathfrak{STA}_{\mathbb{RNA}}$ with many gaps. The last condition makes Algorithm 1 not consider all the alignments $\mathfrak{STA}_{\mathbb{RNA}}$ with base-pairings difficult to predict (e.g. distant). Algorithm 1 with the above conditions are ideally with the time and space complexities $O(L^2), O(L^2)$ where $L = \max(N, M)$ if the parameters δ, ϵ take sufficiently small and large values respectively.

2.6 *Probabilistic consistency transformation*

Probabilistic consistency transformation is the technique that converts a probability between a target homolog and each support homolog into a metric that summarizes the phylogeny among all the homologs. (Do *et al.*, 2005) The transformation has been required because computational complexities in order to compute posterior probabilities among all the homologs is NP-complete as with multiple precise alignment. (Do *et al.*, 2005; Hamada *et al.*, 2009c; Tan *et al.*, 2017) In this study, the transformations that transform probabilities $\{p^{\mathrm{PA},\mathbb{RNA}}_{ijkl}\}$ are proposed. To average probabilities $p^{\mathrm{PA},(\boldsymbol{RNA},\boldsymbol{RNA}')}_{ijkl}$ between the target homolog and each support homolog $\boldsymbol{RNA}, \{\boldsymbol{RNA}'\}$, the average pairing probability $p^{\mathrm{pair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}_{ij}$ is gained:

$$p^{\mathrm{pair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}_{ij} = \frac{\sum_{\boldsymbol{RNA}'kl} p^{\mathrm{PA},(\boldsymbol{RNA},\boldsymbol{RNA}')}_{ijkl}}{|\{\boldsymbol{RNA}'\}|}$$

To sum probabilities $\{p^{\mathrm{pair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}_{ij}\}$, an average unpairing probability $p^{\mathrm{unpair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}_{i}$ is obtained:

$$p^{\mathrm{unpair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}_{ij} = 1 - \sum_{j:j<i} p^{\mathrm{pair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}_{ji} - \sum_{j} p^{\mathrm{pair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}_{ij}$$

Obviously, the difference between proposed and existing probabilities $p^{\mathrm{pair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}_{ij}, p^{\mathrm{pair},\boldsymbol{RNA}}_{ij}$ is that multiple support homologs $\{\boldsymbol{RNA}'\}$ are considered or not.

2.7 Secondary structure prediction that maximizes average expected accuracy incorporating multiple support homologs

The *accuracy* of a predicted structure $\boldsymbol{SS}$ against a reference alignment $\mathbb{STA}$ is measured based on in terms of positive and negative predictions:

$$a(\boldsymbol{SS}, \mathbb{STA}) = \alpha_1 TP + \alpha_2 TN - \alpha_3 FP - \alpha_4 FN$$

where TP, TN, FP, FN are the numbers of true positive, true negative, false positive, and false negative predictions respectively and $\{\alpha_h\}$ are their scale parameters. In this study, the counts TP, TN, FP, FN are configured as:

$$\begin{cases} TP = \sum_{ij} I^{\mathrm{pair},\boldsymbol{SS}}_{ij} I^{\mathrm{pair},\mathbb{STA}}_{ij} \\ TN = \sum_{i} I^{\mathrm{unpair},\boldsymbol{SS}}_{i} I^{\mathrm{unpair},\mathbb{STA}}_{i} \\ FP = \sum_{i} I^{\mathrm{pair},\boldsymbol{SS}}_{i} I^{\mathrm{unpair},\mathbb{STA}}_{i} \\ FN = \sum_{i} I^{\mathrm{unpair},\boldsymbol{SS}}_{i} I^{\mathrm{pair},\mathbb{STA}}_{i} \end{cases}.$$

Here, $I^{\mathrm{pair},\boldsymbol{SS}}_{ij}$ is 1 if the positions i, j are pairing in the structure $\boldsymbol{SS}$ and 0 otherwise, $I^{\mathrm{pair},\mathbb{STA}}_{ij}$ is 1 if the positions i, j are pairing in the alignment $\mathbb{STA}_{ij}$ and 0 otherwise, $I^{\mathrm{unpair},\boldsymbol{SS}}_{i}$ is 1 if the position i is unpaired in the structure $\boldsymbol{SS}$ and 0 otherwise, $I^{\mathrm{unpair},\mathbb{STA}}_{i}$ is 1 if the position i is unpaired in the alignment $\mathbb{STA}$ and 0 otherwise, and $I^{\mathrm{pair},\boldsymbol{SS}}_{i} = 1 - I^{\mathrm{unpair},\boldsymbol{SS}}_{i}, I^{\mathrm{pair},\mathbb{STA}}_{i} = 1 - I^{\mathrm{unpair},\mathbb{STA}}_{i}$.

Because the accuracy and γ-dependent accuracy

$$a(\boldsymbol{SS}, \mathbb{STA}), a^{\gamma}(\boldsymbol{SS}, \mathbb{STA}) = \gamma TP + TN$$

are equivalent, the expected accuracy to be maximized is gained:

$$\mathbb{E}_{\mathbb{STA}_{\mathbb{RNA}}}[a^{\gamma}(\boldsymbol{SS}_{\boldsymbol{RNA}}, \mathbb{STA}_{\mathbb{RNA}})] = \gamma \sum_{ij|I^{\mathrm{pair},\boldsymbol{SS}_{\boldsymbol{RNA}}}_{ij}=1} p^{\mathrm{pair},\mathbb{RNA}}_{ij} + \sum_{i|I^{\mathrm{unpair},\boldsymbol{SS}_{\boldsymbol{RNA}}}_{i}=1} p^{\mathrm{unpair},\mathbb{RNA}}_{i}$$

where $\gamma = \frac{\alpha_1 + 2\alpha_4}{\alpha_2 + \alpha_3}, p^{\mathrm{pair},\mathbb{RNA}}_{ij} = \sum_{kl} p^{\mathrm{PA},\mathbb{RNA}}_{ijkl}, p^{\mathrm{unpair},\mathbb{RNA}}_{i} = 1 - \sum_{j:j<i} p^{\mathrm{pair},\mathbb{RNA}}_{ji} - \sum_{j} p^{\mathrm{pair},\mathbb{RNA}}_{ij}$. The predicted structure that maximizes expected accuracy $\mathbb{E}_{\mathbb{STA}_{\mathbb{RNA}}}[a^{\gamma}(\boldsymbol{SS}_{\boldsymbol{RNA}}, \mathbb{STA}_{\mathbb{RNA}})]$ can be computed to conduct the Nussinov-type dynamic programming (Nussinov *et al.*, 1978) based on the following recursion:

$$mea_{ij} = \begin{cases} p^{\mathrm{unpair},\mathbb{RNA}}_{i} + mea_{i+1,j} \\ mea_{i,j-1} + p^{\mathrm{unpair},\mathbb{RNA}}_{j} \\ \gamma p^{\mathrm{pair},\mathbb{RNA}}_{ij} + mea_{i+1,j-1} \\ \max_{i \leq k < j}(mea_{ik} + mea_{k+1,j}) \end{cases}.$$

In order to consider more than one support homologs, it is sufficient only that probabilities $p^{\mathrm{pair},\mathbb{RNA}}_{ij}, p^{\mathrm{unpair},\mathbb{RNA}}_{i}$ are replaced with

Table 1. The used parameter values of the PhyloFold method.

Parameter	Value
Sparsification parameter δ	$\lvert N-M\rvert$
Sparsification parameter ϵ	0.005
Opening gap penalty p^{OG}	0
Extending gap penalty p^{EG}	0

Penalties $p^{\mathrm{OG}}, p^{\mathrm{EG}}$ are set to zero because they $p^{\mathrm{OG}}, p^{\mathrm{EG}}$ exert a bad influence on the prediction of secondary structures.

the probabilities $p_{ij}^{\mathrm{pair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}, p_{i}^{\mathrm{unpair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}$ on the above recursion.

Secondary structure prediction is extended to that of sequence alignment to view positions $\{i\}$ on a sequence $\boldsymbol{RNA}$ as columns on a sequence alignment among sequences $\{\boldsymbol{RNA}\}$, $\boldsymbol{SA}_{\{\boldsymbol{RNA}\}}$, on the above recursion. It is known that pairing probabilities of columns i, j given an alignment $\boldsymbol{SA}_{\{\boldsymbol{RNA}\}}$, $p_{ij}^{\mathrm{pair},\boldsymbol{SA}_{\{\boldsymbol{RNA}\}}}$, improves the prediction accuracy of consensus secondary structure. (Hamada *et al.*, 2011; Bernhart *et al.*, 2008) Thus, the mixture of probabilities $\{p_{ij}^{\mathrm{pair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}\}, p_{ij}^{\mathrm{pair},\boldsymbol{SA}_{\{\boldsymbol{RNA}\}}}$ is used on the above recursion to predict the structure of an alignment $\boldsymbol{SA}_{\{\boldsymbol{RNA}\}}$:

$$p_{ij}^{\mathrm{pair,mix},\boldsymbol{SA}_{\{\boldsymbol{RNA}\}}} = \frac{\sum_{\boldsymbol{RNA}'} p_{i^*j^*}^{\mathrm{pair},\boldsymbol{RNA}',\{\boldsymbol{RNA}\}\setminus\boldsymbol{RNA}'}}{2|\{\boldsymbol{RNA}^*\}|} + \frac{p_{ij}^{\mathrm{pair},\boldsymbol{SA}_{\{\boldsymbol{RNA}\}}}}{2}$$

where $\boldsymbol{RNA}' \in \{\boldsymbol{RNA}\}$, $\{\boldsymbol{RNA}^*\}$ is the subset of the sequences $\{\boldsymbol{RNA}\}$ that does not mapped to gaps on the columns i, j in the alignment $\boldsymbol{SA}_{\{\boldsymbol{RNA}\}}$, and i^*, j^* are the positions mapped to the columns i, j in the alignment $\boldsymbol{SA}_{\{\boldsymbol{RNA}\}}$ respectively. A matrix $(p_{ij}^{\mathrm{pair},\boldsymbol{SA}_{\{\boldsymbol{RNA}\}}})$ can be computed by the RNAalipfold algorithm. (Bernhart *et al.*, 2008)

3 Results and discussions

3.1 Implementations and benchmark environments

The PhyloFold method proposed in this study was implemented on the Rust programming language as the version 0.1.0 of the PhyloFold program ("https://github.com/heartsh/phylofold") and the PhyloAliFold program ("https://github.com/heartsh/phyloalifold"). The programs employ multi-threading to achieve running time as fast as possible. Programs were run on a computer composed of the "AMD EPYC 7501" CPU with 128 threads and the clock rate 2 GHz and 128 GB of RAM unless environments for running programs are specified.

3.2 Benchmark of PhyloFold method with competitors

Positive predictive value, sensitivity, and false positive rate are calculated from the numbers of true and false positives and negatives:

$$\begin{cases} PPV = \frac{TP}{TP+FP} \\ SENS = \frac{TP}{TP+FN} \\ FPR = \frac{FP}{TN+FP} \end{cases}.$$

The PhyloFold program performed the best trade-off of the above metrices (Figure 3) among the PhyloFold, CentroidFold (Hamada *et al.*, 2009b), CentroidHomFold (Hamada *et al.*, 2009c), and TurboFold-smp (Tan *et al.*, 2017) programs (Table 2) on test set "unaligned" (Table 3) while demanding comparable running time (Table 4). Ten percent of ncRNA families whose reference seed structural alignments have at most 200 columns and whose numbers of homologs are less than 11 are sampled from the Rfam database (Kalvari *et al.*, 2018). The reference seed structural alignments of the 147 sampled families are compiled as test set "aligned". The PhyloAliFold program demonstrated the superior trade-off of the above metrices (Figure 4) to the CentroidAliFold program (Hamada *et al.*, 2011) on the set keeping comparable running time (Table 5).

Table 2. The profile of γ-dependent programs that predict secondary structures.

Program (version)	Input	Probability type
PhyloFold	Unaligned	Precise alignment
TurboFold-smp	Unaligned	Decomposed alignment
CentroidHomFold	Unaligned	Decomposed alignment
CentroidFold	Unaligned	Single sequence
PhyloAliFold	Aligned	Precise alignment + consensus structure
CentroidAlifold	Aligned	Single sequence + consensus structure

The rightmost column shows the type of probabilities used to predict secondary structures. Precise alignment type is proposed in this study. Decomposed alignment type considers structural alignment to decompose an alignment $\mathbb{STA}_{\mathbb{RNA}}$ into the structures and alignment $\boldsymbol{SS}_{\boldsymbol{RNA}}, \boldsymbol{SS}_{\boldsymbol{RNA}'}, \boldsymbol{SA}_{\mathbb{RNA}}$. Single sequence type does not take support homologs into account. "+ consensus structure" means that probabilities on consensus secondary structure given input sequence alignment are mixed. Precise alignment (+ consensus structure) type is expected to display the highest prediction accuracy.

Table 3. The profile of test set "unaligned".

Family	# homologs	Range nt
RNAIII	4	489–517
Hammerhead ribozyme (type III)	10	40–58
Hammerhead ribozyme (type I)	10	45–118
Bicoid 3 prime-UTR regulartory element	4	543–551
R2 RNA element	10	202–237
Gammaretrovirus core encapsidation signal	10	100–102
RNase MRP	5	246–305
Ciliate telomerase RNA	10	160–215
Coronavirus 3 prime stem-loop…	10	43
Vimentin 3 prime UTR…	10	65–77
RNase E 5 prime UTR element	6	337–339

Unaligned homologs of 11 ncRNA families and the reference secondary structures of the homologs were collected from the RNA STRAND database (Andronescu *et al.*, 2008) and compiled as test set "unaligned". The Rfam database was not used to build test set "unaligned" because this database does not register reference single secondary structures. Ten homologs and the reference secondary structures of them were sampled from each ncRNA family when the family has 11 or more homologs. The middle and rightmost columns show the number of the homologs and the range of lengths on each ncRNA family respectively. "Coronavirus 3 prime stem-loop…" and "Vimentin 3 prime UTR" stand for "Coronavirus 3 prime stem-loop II-like motif (s2m)" and "Vimentin 3 prime UTR protein-binding region" respectively.

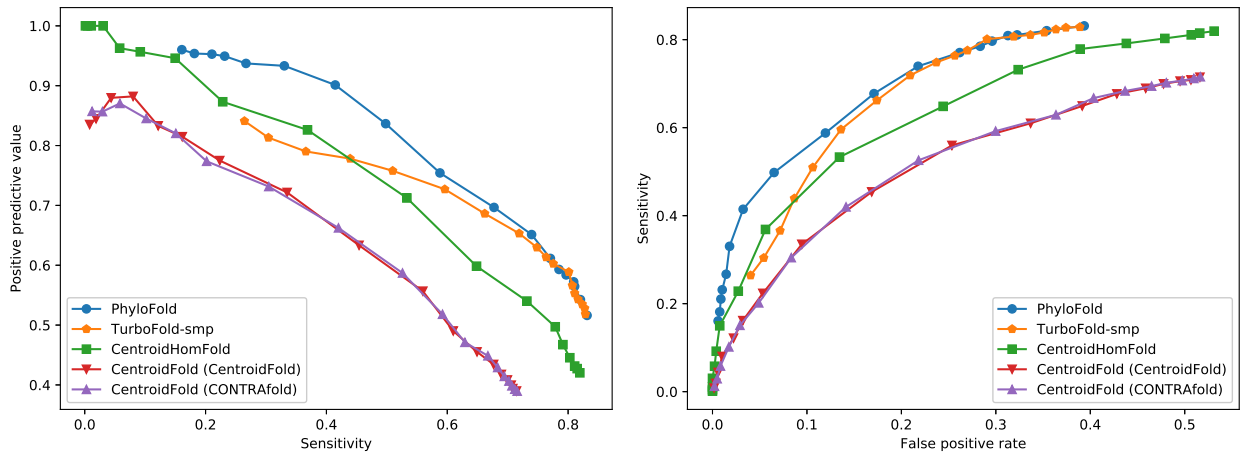

Fig. 3. The benchmark of γ-dependent programs that predict single secondary structures on prediction accuracy using test set "unaligned". The trade-off curves are composed of (left) pairs $\{(SENS, PPV)\}$ and (right) pairs $\{(FPR, SENS)\}$ at each parameter $\gamma = 2^g : g \in \{-7, \ldots, 10\}$ respectively. The PhyloFold method specified its parameter values as in Table 1. The CentroidFold and CentroidHomFold programs were run with their default parameter values. The TurboFold-smp was given only the number of its iterations $\eta = 1$. The different algorithms were shown in the parentheses.

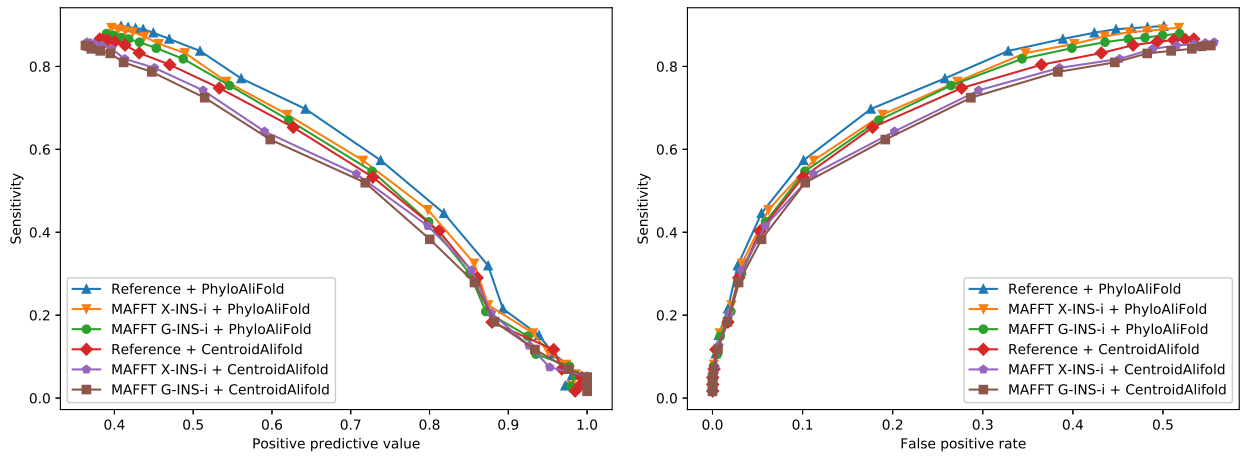

Fig. 4. The benchmark of γ-dependent programs that predict consensus secondary structures on prediction accuracy using test set "aligned". The trade-off curves are composed of (left) pairs $\{(PPV, SENS)\}$ and (right) pairs $\{(FPR, SENS)\}$ at each parameter $\gamma = 2^g : g \in \{-7, \ldots, 10\}$ respectively. The parameter values of the PhyloFold method are the same as Figure 3. The CentroidAlifold program was run with its default parameter values. The MAFFT G-INS-i and MAFFT X-INS-i programs (Katoh and Standley, 2013) were used to compute the sequence alignments to input into the PhyloAliFold program. The MAFFT G-INS-i executes only classic multiple sequence alignment whereas the MAFFT X-INS-i considers structural alignments during multiple sequence alignment to decompose them. The reference sequence alignments of the ncRNA families in test set "aligned" are also used as the input of the PhyloAliFold.

Table 4. The benchmark of γ-dependent programs that predict single secondary structures on running time using test set "unaligned".

Program (algorithm)	Running time	Time complexity
PhyloFold	8.34 m	$O(R^2N^2 + RN^3)$
TurboFold-smp	6.63 m	$O(R^2N^2 + RN^3)$
CentroidHomFold	1.46 m	$O(R^2N^3)$
CentroidFold (CentroidFold)	7.21 s	$O(RN^3)$
CentroidFold (CONTRAfold)	8.42 s	$O(RN^3)$

The rightmost column shows the ideal time complexity of each program where R is the number of homologs and N is the longest length of the homologs. The running time of the programs were measured when $\gamma = 2^0 = 1$ across the test set. All the parameter values of the programs were same as Figure 3. The used computer has the "Intel Xeon CPU" CPU with 64 threads and the clock rate 2.30 GHz and 240 GB of RAM. The PhyloFold and TurboFold-smp programs were run with multi-threading while the CentroidFold and CentroidHomFold programs, which are not implemented on multi-threading, were run with multi-processing to make the benchmark fair.

Table 5. The benchmark of γ-dependent programs that predict consensus secondary structures on running time using test set "aligned".

Program	Running time	Time complexity
PhyloAliFold	7.00 m	$O(R^2N^2 + RC^2 + C^3)$
CentroidAlifold	0.947 s	$O(RN^3 + RC^2 + C^3)$

The rightmost column shows the ideal time complexity of each program where R is the number of homologs, N is the longest length of the homologs, C is the number of columns in an input sequence alignment. The running time of the programs were measured when $\gamma = 2^0 = 1$ across the test set using the MAFFT X-INS-i program to predict input sequence alignments. All the parameter values of the programs were same as Figure 4. The used computer is the same as Table 4. The PhyloFold program was run with multi-threading and multi-processing while the CentroidAlifold program, which is not implemented on multi-threading, was run with multi-processing to make the benchmark fair.

```
(A)
GCCCGGGUGGUGAAAUC--GGUAGACACGCAGGACUUAAAAUCCUGUGGCAU---AAAAGCCAUGUCGGUUCAAGUCCGACCCCGGGCA
GCCGAAAUAGCUCAGUU--GGGAGA-GCGUUAGACUGAAGAUC----------UAAAAGGUCCCUGGUUCGAUCCCGGGUUUCGGCA
GACUCGUUAGCUCAGCC--GGUAGA-GCAACUGGCUUUUAA-C------------CAGUGGGUCCGGGGUUCGAAUCCCCGACGAGUCA
GGCUUUUUAGCUCAGCA--GGUAGA-GCAACCGGCUGUUAA-C------------CGGUUUGUCACAGGUUCGAGCCCUGUAAAAGCCG
---UCUUUCUAGUACUA--AGGAGU-AUAAGUGGCUUCCAA--------------CCACACGGUCUUGGUUAGAGUCCAAGGAAAGAU-
GGAGGAAUACCCAAGUCUGGCUGAAGGGAUCGGUCUUGAAAACCGACAGGGUGUCAAAGCCCGCGGGGGUUCGAAUCCCUCUUCCUCCG
(((((((....((((((.............))))))..........................(((((.......)))))))))))).
(B)
GCCCGGGUGGUGAAAUC-GGUAGA-CACGCAGGACUUAAA-AUCCUGUGG-CAUAAAAGCCA-UGUCGGUUCAAGUCCGACCCCGGGCA
GCCGAAAUAGCUCAGUU-GGGAGA--GCGUUAGACUGAAG-AUCUAAAAG--------GUC---CCUGGUUCGAUCCCGGGUUUCGGCA
GACUCGUUAGCUCAGCC-GGUAGA--GCAACUGGCUUUUA-ACC-AGUGG--------GUC---CGGGGUUCGAAUCCCCGACGAGUCA
GGCUUUUUAGCUCAGCA-GGUAGA--GCAACCGGCUGUUA-ACC-GGUUU--------GUC---ACAGGUUCGAGCCCUGUAAAAGCCG
-UCUUUCUAGUAC--UA-AGGAGU--AUAAGUGGCUUCCA-ACC-ACACG--------GUC----UUGGUUAGAGUCCAA-GGAAAGAU
GGAGGAAUACCCAAGUCUGGCUGAAGGGAUCGGUCUUGAAAACCGACAGGGUGUCAAAGCCCGCGGGGGUUCGAAUCCCUCUUCCUCCG
(((((((..((((.........))..)).(((((........)).))).................(((((.......)))))))))))).
(C)
GCCCGGGUGGUGAAAUC-GGUA-GACACGCAGGACUUAAAAUCCUGU-GGCA--UAAA--AGCCA--UGUCGGUUCAAGUCCGACCCCGGGCA
GCCGAAAUAGCUCAGUU-GGG--AGAGCGUUAGACUGAAGAUCUAAAA----------------GGUCCCUGGUUCGAUCCCGGGUUUCGGCA
GACUCGUUAGCUCAGCC-GGU--AGAGCAACUGGCUUUUAACCAGUGG----------------GU-CCGGGGUUCGAAUCCCCGACGAGUCA
GGCUUUUUAGCUCAGCA-GGU--AGAGCAACCGGCUGUUAACCGGUUU----------------GU-CACAGGUUCGAGCCCUGUAAAAGCCG
-UCUUUCUAGUACUAA--GG---AGUAUAAGUGGCUUCCAACCACACG----------------G--UCUUGGUUAGAGUCCAAGGAAAGAU-
GGAGGAAUACCCAAGUCUGGCUGAAGGGAUCGGUCUUGAAAACCGAC-AGGGUGUCAAAGCCCG---CGGGGGUUCGAAUCCCUCUUCCUCCG
(((((((..((((.........)))).(((((.......)))))).....................(((((.......)))))))))))).
p >= 5.00e-01, p >= 2.50e-01, p >= 1.25e-01, p >= 6.25e-02, p < 6.25e-02
```

Fig. 5. Consensus secondary structures color-coded with their reliability. The structures are of the sequence alignment among homologs sampled from the tRNA family registered in the Rfam database and were computed by (A) the MAFFT G-INS-i and (B) MAFFT X-INS-i programs. The consensus secondary structures were computed by the RNAalifold program (Bernhart et al., 2008). (C) The reference consensus secondary structures of the homologs derived from the database. The reliability was assessed to average mean pairing probabilities $\{p_{ij}^{\mathrm{pair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}\}$ per column pair. The reliability was improved when the quality of the input sequence alignment increased.

4 Conclusions

The PhyloFold method, which predicts secondary structures considering the phylogeny among homologs, has been devised. The method demonstrated superior prediction accuracy to competitors while demanding comparable running time on the prediction of both single and consensus secondary structures. The method has the potential to be improved by secondary structure probing methods such as icSHAPE (Spitale *et al.*, 2015) and SHAPE-MaP (Siegfried *et al.*, 2014) as demonstrated in the SuperFold (Siegfried *et al.*, 2014) and TurboFold methods.

Probabilities $\{p_{ij}^{\mathrm{pair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}\}$ can be used to measure the reliability of predicted secondary structures as a part of postprocessing. (Figure 5) Also, the probabilities $\{p_{ij}^{\mathrm{pair},\boldsymbol{RNA},\{\boldsymbol{RNA}'\}}\}$ can be utilized to improve the prediction accuracy of existing methods that predict RNA-RNA interactions, alignments, etc. such as the IntaRNA (Mann *et al.*, 2017), CentroidAlign (Hamada *et al.*, 2009a), and DAFS (Sato *et al.*, 2012) programs, which employ the decomposition of structural alignments or do not incorporate structural alignments.

Acknowledgements

I thank the members of the Asai and Frith laboratories for discussing this study with me for years. Also I thank Dr. Risa Kawaguchi to share the valuable information about the scaling and logsumexp methods on exact probability computations with me. Most computations were performed on the NIG supercomputer at the ROIS National Institute of Genetics.

Author contributions

M.T. conceived of the proposed method, implemented it to the presented programs, experimented with them and competitors, and wrote this paper.

References

Andronescu, M. *et al.* (2008). RNA STRAND: The RNA Secondary Structure and Statistical Analysis Database. *BMC Bioinformatics*, **9**(1), 340.

Bentley, D. R. *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**(7218), 53–59.

Bernhart, S. H. *et al.* (2008). RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**(1), 474.

Bradley, R. K. *et al.* (2009). Fast statistical alignment. *PLoS Computational Biology*, **5**(5).

Do, C. B. *et al.* (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research*, **15**(2), 330–40.

Do, C. B. *et al.* (2008). A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, **24**(13), i68–i76.

Durbin, R. *et al.* (1998). Biological sequence analysis.

Hamada, M. *et al.* (2009a). CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score. *Bioinformatics (Oxford, England)*, **25**(24), 3236–43.

Hamada, M. *et al.* (2009b). Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**(4), 465–473.

Hamada, M. *et al.* (2009c). Predictions of RNA secondary structure by combining homologous sequence information. *Bioinformatics*, **25**(12), i330–i338.

Hamada, M. *et al.* (2011). Improving the accuracy of predicting secondary structure for aligned RNA sequences. *Nucleic Acids Research*, **39**(2), 393–402.

Havgaard, J. H. *et al.* (2007). Fast Pairwise Structural RNA Alignments by Pruning of the Dynamical Programming Matrix. *PLoS Computational Biology*, **3**(10), e193.

Hofacker, I. L. *et al.* (2004). Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**(14), 2222–2227.

Ji, P. *et al.* (2003). MALAT-1, a novel noncoding RNA and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*, **22**(39), 8031–8041.

Kalvari, I. *et al.* (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, **46**(D1), D335–D342.

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, **30**(4), 772–780.

Kino, T. *et al.* (2010). Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Science signaling*, **3**(107), ra8.

Klein, R. J. and Eddy, S. R. (2003). RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**(1), 44.

Long, J. C. and Caceres, J. F. (2009). The SR protein family of splicing factors: master regulators of gene expression. *The Biochemical journal*, **417**(1), 15–27.

Mann, M., Wright, P. R., and Backofen, R. (2017). IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Research*, **45**, 435–439.

Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA (DNA chenistry/dimethyl sulfate cleavage/hydrazine/piperidine). *Biochemistry*, **74**(2), 560–564.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**(6-7), 1105–1119.

Nussinov, R., Pieczenik, G., Griggs, J. R., and Kleitman, D. J. (1978). Algorithms for Loop Matchings. *SIAM Journal on Applied Mathematics*, **35**(1), 68–82.

Pasmant, E. *et al.* (2011). ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, **25**(2), 444–8.

Sankoff, D. (1985). Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM J. Appl. Math.*, **45**(5), 810–825.

Sato, K. *et al.* (2011). IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**(13), i85–i93.

Sato, K. *et al.* (2012). DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition. *Bioinformatics*, **28**(24), 3218–3224.

Siegfried, N. A. *et al.* (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods*, **11**(9), 959–965.

Spitale, R. C. *et al.* (2015). Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, **519**(7544), 486–490.

Tan, Z. *et al.* (2017). TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Research*, **45**(20), 11570–11581.

Turner, D. H. and Mathews, D. H. (2010). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, **38**(suppl_1), D280–D282.

Wu, F. *et al.* (1991). tat regulates binding of the human immunodeficiency virus trans-activating region RNA loop-binding protein TRP-185. *Genes & development*, **5**(11), 2128–40.