博士論文

Deciphering the global landscape of RNA secondary structure based on comprehensive prediction and reproducible high-throughput structure analyses

（網羅的構造予測及び再現性を考慮したハイスループット構造解析による**RNA**二次構造の全体像の解明）

河口　理紗

Risa Kawaguchi

Acknowledgments

This dissertation comprises a doctoral research result of the author during the PhD program in the department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, the University of Tokyo. First of all, the author deeply appreciates countless instructive advices from Prof. Hisanori Kiryu, an associate professor at the department, about a sincere attitude to science and research as a supervisor. In addition, the author would thank him to help the author have various opportunities freely to attend a meeting, collaborate with people outside, and work as a committee of the conferences for young researchers. As a chief examiner, Prof. Kiyoshi Asai, a full professor at the department, gave the author many advices about how to conduct, appeal, and extend ParasoR research besides homemade feast. The author thank Prof. Martin Firth, Prof. Koji Tsuda and Prof. Masahiro Kasahara at the department, and Prof. Tatsuya Akutsu at Bioinformatics Center, Institute for Chemical Research, Kyoto University, for giving thoughtful comments and instructions and accepting the offer of deputy examiner.

For conducting this research, the author is grateful to the previous and current lab members, Dr. Hirotaka Matsumoto, Mr. Yuto Ichikawa, Mr. Yasuhiro Kojima, Mr. Yutaro Konta, and Mr. Hiroshi Miyake, and other members of the student room for daily discussion and encouragement. Moreover, the author received meticulous support and help from Ms. Yoko Nomura. Prof. Michiaki Hamada, Dr. Tsukasa Fukunaga, and Prof. Kengo Sato gave many advices about ParasoR development and analyses. The author also thank Kara Finley, Sean Eddy and Thomas Jones for comments on the ParasoR research. This research was supported by a Grant-in-Aid for JSPS Fellows [Grant Number 14J00402]. Dr. Naoto Kondo and Dr. Michihira Tagami have supported me to analyze NGS data including high-throughput structure analyses. Lastly, the author would like appreciate family help to improve the author's concentration on research.

Preface

This research was published in

- Risa Kawaguchi and Hisanori Kiryu. “Parallel computation of genome-scale RNA secondary structure to detect structural constraints on human genome.” *BMC Bioinformatics*, BioMed Central. **17**:203 (2016).

This research was presented in

- “Computational Analysis of RNA Structure and Function” meeting in Benasque, Spain, Jul. 2015 as oral presentation.
- RiboClub meeting in Quèbec, Canada, Sep. 2013 as oral and poster presentation.
- The RNA Society, 21st annual meeting and the RNA Society of Japan, 18th annual meeting in Kyoto, Japan, Jun. 2016 as poster presentation.
- The International Society for Computational Biology (ICSB)-Asia meeting in Shenzhen, China, Dec. 2012 as poster presentation.
- Bioinformatics Week in Odaiba (BiWO) 2012 in Tokyo, Oct. 2012 as poster presentation.
- Asian Young Researchers Conference on Computational and Omics Biology (AYRCOB), 6th meeting in Shenzhen, China, Dec. 2012 as poster, 7th meeting in Tokyo, Japan, Sep. 2013 as oral and poster, and 8th meeting in Hsinchu, Taiwan, Jan. 2015 as poster presentation.
- EMBL conference series “From Functional Genomics to Systems Biology” meeting in Heidelberg, Germany, Nov. 2014 as poster presentation.

- Informatics in Biology, Medicine and Pharmacology, 5th meeting in Tokyo, Sep. 2016 as oral presentation.
- Japan RNA society, 14th in Sendai, Jul. 2012, 16th meeting in Aichi, Sep. 2014, 17th meeting in Hokkaido, Jul. 2015 as poster presentation.
- The Molecular Biology Society of Japan, 35th meeting in Fukuoka, Dec. 2012 and 37th meeting in Yokohama, Dec. 2014.
- NGS Genba no kai meeting, 3rd meeting in Kobe, Sep. 2013 and 4th meeting in Tsukuba, Jul. 2015 as poster presentation.
- Japanese association of the younger researchers of bioinformatics, 4th meeting in Aichi, Mar. 2013 as oral, 5th meeting in Tokyo, Feb. 2014 as oral, 6th meeting in Kobe, Oct. 2014 as oral, 7th meeting in Yamagata, Oct. 2015 as poster, and 8th meeting in Hokkaido, Oct. 2016 as poster presentation.

ABSTRACT

BACKGROUND

RNA secondary structure is one of the functional features of RNA molecules widely studied by both experimental and computational approaches. Such structure consists of canonical base pairs and highly contributes to the stability of single-stranded RNA in the cell. Previous studies on the functionality of small non-coding RNAs revealed that they form specific secondary structures to interact with RNA binding proteins or other regulatory RNAs. Moreover, recent high-throughput structure analyses suggested that local structures of long RNAs such as mRNA and long non-coding RNA (lncRNA) potentially play an important role to modulate the exposure of binding sites for post-transcriptional regulation, for example, splicing, translation, and degradation.

Comprehensive RNA secondary structure analyses have been promoted mainly by *in silico* analyses because the conventional experiments for structure analyses are extremely costly and time consuming. Many computational algorithms have been developed for comprehensive structure prediction of small RNAs (sRNAs) to reveal the close relationship between structure and function of sRNAs. A thermodynamic folding, which is based on the energy parameters that were experimentally determined, is one of the leading methods to discover the significant RNA structure characteristics and motifs. However, very few studies have examined the structure propensity of long RNAs such as lncRNAs and transcripts, due to the fundamental limitation of thermodynamic folding methods on computational time and resource. To reduce long computational time, previous studies examined the structure propensity only around the tips of functional sites. For example, the structure stability around splicing sites on pre-mRNAs is suggested to affect the pattern of alternative splicing and contribute to produce a diversity of matured transcripts. On the other hand, there is little knowledge about structure propensity deep inside introns so far because of the limitation of existing computational methods.

Compared to *in silico* analyses, the throughput and scalability of experimental structure analyses has drastically improved combined with sequencing technology, which are referred to as high-throughput structure analyses. High-throughput structure analyses, such as PARS and icSHAPE, have been applied to explore a variety of transcriptome-wide landscape of RNA secondary structure, and discovered common structure propensity around regulatory sites or *in vivo* and *in vitro* difference. However, past studies have rarely paid a careful attention to convert sequencing data to structure information. For that reason, they were not to be robust for the systematic biases nor irreproducible random noises produced during library preparation and sequencing processes. As a contamination of low reliable reads can also affect the accuracy of downstream analyses such as computational prediction with the guide of high-through structure analyses, there has still remained a severe problem to decipher the global landscape of RNA secondary structure based on computational and experimental methodologies.

ParasoR: Parallel computation of genome-scale RNA secondary structure to detect structural constraints on human genome

To solve the difficulty of computational prediction of RNA secondary structure for long transcripts, I developed a novel algorithm, *ParasoR*. ParasoR enables the distributed computation of various structural features even for long RNAs based on the Boltzmann ensemble over globally consistent secondary structures. Unlike the previous sliding-window methods that predict locally stable RNA secondary structures within the independent sliding window, ParasoR can exhaustively compute globally consistent structural features such as structural profiles and γ-centroid structures as well as conventional base pairing probabilities. ParasoR divides dynamic programming (DP) matrices into smaller pieces, such that each piece can be computed by a separate computer node without losing the connectivity information between the pieces. ParasoR also directly computes the ratios of DP variables to avoid the reduction of numerical precision caused by the cancellation of quite a large number of Boltzmann factors.

ParasoR was developed for the structure prediction of long RNA sequences, for example, to examine the positional structure propensity of introns. To evaluate the accuracy of ParasoR for long RNAs, conserved secondary structure motifs found in genome and mRNA sequences and dataset of high-throughput structure analysis were applied for the performance validation. I evaluated the prediction accuracy of ParasoR and other tools using CisReg data, which contains high-quality sub-structures within long sequences. As a result, ParasoR is comparable to or better than the state-of-the-art algorithms for the prediction of stable motif structures such as cis-regulatory elements in long RNAs. Next, I investigated the congruence between computational predictions and PARS data, one of the high-throughput structure analyses. Consequently, although all of the prediction methods showed a high consensus with the PARS-based classification, ParasoR had an almost comparable area under the curve (AUC) score around 0.6 to LocalFold and RNAplfold, and showed a slightly higher AUC when the 32-nt averaging was applied to the score calculation. They also attain high AUCs around 0.7-0.8 when ambiguous sites are removed from the PARS data. The differences of AUC scores among these programs are of the order of 0.01 and thus very small for these datasets.

Using ParasoR, I investigated the global structural preferences of transcribed regions in the human genome, particularly for intronic regions. A genome-wide folding simulation indicated that transcribed regions are significantly more structural than intergenic regions after removing repeat sequences and k-mer frequency bias. In particular, a highly significant preference for base pairing was observed over entire intronic regions as compared to their antisense sequences, as well as to intergenic regions. A comparison between pre-mRNAs and mRNAs showed that coding regions become more accessible after splicing, indicating constraints for translational efficiency. Such changes are correlated with gene expression levels, as well as GC content, and are enriched among genes associated with cytoskeleton and kinase functions.

reactIDR: Statistical approach to robust RNA reactivity classification based on reproducible high-throughput structure analyses

Currently more than dozen research studies about a novel high-throughput structure analysis have been published, which do not suffer from scalability problems encompassed in the conventional structure analyses. While high-throughput structure analyses can be practical for quite a few

subjects, estimated reactivity scores tend to be sparse and inconsistent between each structure analysis. This inconsistency is supposed to be the distinctive difference of detectability and systematic biases of individual high-throughput structure analyses. For example, DMS-Seq has a preference of probed nucleotides (enrichment of adenine and cytosine) while PARS can detect both highly single-stranded and double-stranded regions using two different enzymes. Moreover, sequencing read counts are susceptible to be violated by the random noise with regard to the low expression transcripts. Taken together, a comparison of multiple high-throughput structure analyses should be based on reliable reactivity information after correcting the systematic biases to establish common overall landscape of RNA secondary structure.

To establish a statistical methodology for robust structure analyses, I developed a novel pipeline, reactIDR, which is designed to extract reliable structure information from sequencing-based structure analyses. To evaluate the reliability of each reactivity score, the irreproducible discovery rate (IDR) is computed by modeling the joint probability distribution among replicates. IDR estimation can be also carried out considering local consistency of read coverage based on the hidden Markov model so that reactIDR has the potential to extract a larger number of reproducible but low-coverage regions, due to the additional information about local consistency of IDR profiles. reactIDR can also compute p-values based on the null distribution of Poisson and negative binomial distribution considering a different sequencing depth of each transcript.

The efficiency of IDR index for reproducibility was evaluated by comparing IDR-based reactivity classification and stem probability of computational prediction over the entire transcriptome. As a result, IDR-based classification was highly consistent with that of stem probability. Moreover, several machine learning algorithms were applied to evaluate the weight of each feature for correct classification of human 18S rRNA reference structure as well as transcriptome-wide stem probability. Consequently, the accuracy of structure classification using sequencing-based features increased for the 18S rRNA structure from 0.6 to 0.8 when IDR-based filtering was performed, suggesting IDR can be a suitable measurement to exclude unreliable regions from the whole dataset. Filtering of irreproducible regions also increased the accuracy of stem probability classification by several machine learning-based classifiers for human transcriptome. In conclusion, IDR-based filtering can be considered effective to evaluate unreliability of sequencing data, leading the increase of robustness against very sparse high-throughput data. Furthermore, my analyses suggest that a combination of computational and experimental genome-wide structure analyses has a promising potential to infer the global landscape of RNA secondary structure.

CONTENTS

Introduction

In the recent field of biology, scientists can measure multi-dimensional biological features to understand the mechanism of biological regulation and misregulation, following the advancement of measurement methods such as DNA sequencing technology, mass spectrometry, and electron microscopy. While each methodology has a potential to reveal the biological phenomena in the different scale of organization, it must possess specific limitations on the range of determination, detectable targets, and systematic biases depending on the experimental conditions [1, 2]. Also, measuring instruments ofter produce too large amounts of data to analyze by humans alone due to an improvement of their outputs. Owing to decipher the hidden mechanism behind such large data, a computational model is required to consider the process of quantification as well as the activity of biological systems we are interested in.

RNA secondary structure is one of the functional features of RNA molecules which have been widely studied in both of experimental and computational fields. RNA secondary structure consists of base pairs between complementary bases and it can highly contribute to the stability of single-stranded RNA in a cell. Since such structure can change the property of molecules, the existence of RNA secondary structure was originally demonstrated by the evidence that RNA showed changes in absorbance in different temperatures [3, 4]. In the functional analyses of non-coding RNA, RNA secondary structure has been shown to be closely related with the promoting appropriate constructions of high-dimensional structures and selections for binding target of RNAs and proteins [5]. Recent studies have revealed that RNA secondary structure is also influential on the regulation acting on coding RNAs, for example, by modulating splicing efficiency [6], translation efficiency [7, 8], and degradation [9].

To understand the influence of RNA secondary structure for transcriptome, a computational algorithm has been developed for RNA secondary structure prediction referring experimentally validated structures or thermodynamic parameters [10, 11]. This is because the conventional methods of experimental structure analysis, such as gel-electrophoresis after structure probing, electron microscopy, X-ray crystal structure analyses, and nuclear magnetic resonance spectroscopy, are extremely costly and time consuming. Previous computational analyses have discovered common RNA structure motifs and evolutionary conserved motifs that might be functional [12, 13]. Nevertheless, very few studies have examined the whole structure propensity of long RNAs such as mRNA, lncRNA and pre-mRNA [14]. Due to the fundamental limitation on computational time and resource for thermodynamic folding methods, only partial regions such as the tips or surroundings of functional have been studied for such RNAs so far [15, 5]. Although sliding-window-based approaches can be a practical tool for genome-wide structure analyses under the current constraints of computational resources [16, 17], they compute only the structures within each artificial sequence window in the input sequence. That is, they cannot predict the globally consistent secondary structures for the whole sequence. Therefore, it has remained a big challenge to examine the positional structure propensity of long RNAs, such as that deep inside introns, only by previous techniques that cannot handle an ensemble of possible structures.

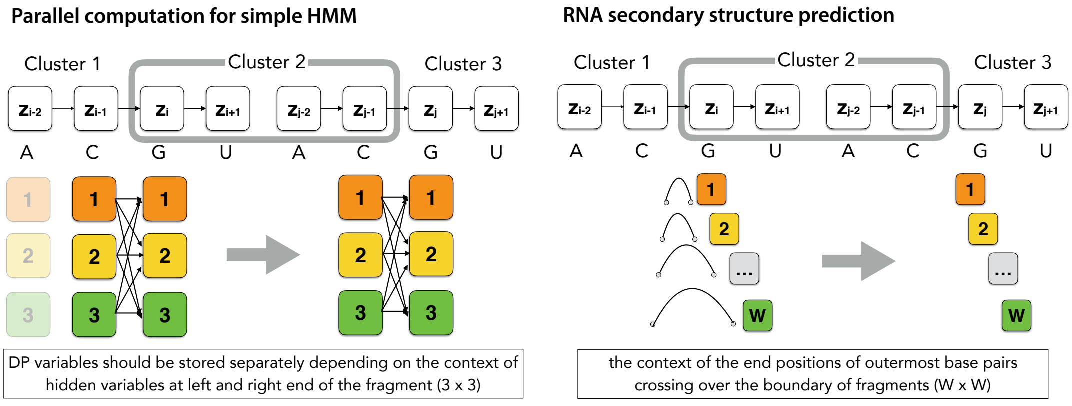
On the other hand, the structure probing technique has been recently improved in terms of scalability by combining with high-throughput sequencing technology to solve the limitation of the conventional experimental analyses. There are more than tens of library construction methods developed for transcriptome-wide detection of single- and double strandedness of RNA molecules at single-base resolution [18]. For example, PARS (parallel analysis of RNA structure) [19] uses two types of enzymes; one is nuclease S1, which recognizes single-stranded regions to cleave, and another is RNase V1, which cleaves double-stranded regions. By sequencing the RNA fragments treated with these two enzymes, PARS can estimate the position of cleavage sites which are likely to be single- or double-stranded regions for thousands of transcripts. Comprehensive structure analyses based on such techniques are referred as RNA structurome, because they enable to observe conformational alterations of transcribed RNAs in various conditions such as *in vitro* and *in vivo*. Actually, the conformational changes of RNA was indicated to affect the efficiency of alternative splicing [20] and the distance of ribosomes during translation process [21] by such high-throughput structure analyses.

However, there is serious a problem that the structure profiles obtained by these analyses are highly inconsistent when the intra-comparison is carried out among different methodologies. This is considered to be derived from not only the variation of actual conformation across each sample but also the systematic biases produced by the different protocols. One of the reasons of systematic biases is a different preferences for cleavage or modification sites, resulting in the variation of rarely mapped regions. Since one of the probing reagents, dimethyl sulfate (DMS) selectively modifies adenine and cytosine residues, sequenced reads tend to be rarely enriched in high-GU content region [22, 23]. Moreover, the efficiency of RNA structure probing and sequencing strongly depends on the expression level of RNA, the length of whole sequence, and density of potentially probed nucleotides. These biases are supposed to be more complicated than previous high-throughput analyses such as RNA-Seq or ChIP-Seq, and thus cannot be solved just by applying existing methods developed for other high-throughput technologies. Consequently, a novel method should be developed for genome-wide structure analyses to solve the ambiguity of structure probing data such as sparseness, systematic biases, and irreproducibility particularly for low expressed genes or intronic regions [24].

While experimental structure analyses can reflect actual base reactivity inside the cell, computational prediction methods can designate highly plausible base pairs or structures that each RNA forms. Thus, to predict what structures RNA forms and what features existent structures have, it is required to develop computational methods which are applicable for genome-wide dataset of high-throughput structure analyses to extract reliable structure information from the mixture of true and spurious signal enrichment.

PARASOR AND REACTIDR

In this study, I present two novel computational approaches for *in silico* structure analyses and *in vitro* and *in vivo* structure analyses. To predict RNA secondary structure for whole transcriptome, I developed an algorithm named *ParasoR*, which carries out parallel computation with direct computation of the ratio of dynamic programming (DP) matrix. Using ParasoR, I examined the landscape of RNA secondary structure for whole transcripts including intronic regions and whole genome sequences. To compare general types of high-throughput structure analyses data and extract reliable structure information, I developed a novel pipeline *reactIDR*. reactIDR has potential to construct the common landscape of RNA secondary structure from structure probing methods considering the specific biases involved in high-throughput structure analyses and the consistency among replicates. Although a comprehensive structure analysis contains many difficulties for cross comparison across individual methodologies, the comparison of reliable information between various conditions and samples is expected to reveal the function of


Parallel computation for simple HMM
Cluster 1
Cluster 2
Cluster 3
z_{i-2}
z_{i-1}
z_i
z_{i+1}
z_{j-2}
z_{j-1}
z_j
z_{j+1}
A C G U A C G U
1 2 3
DP variables should be stored separately depending on the context of hidden variables at left and right end of the fragment (3 x 3)
RNA secondary structure prediction
Cluster 1
Cluster 2
Cluster 3
A C G U A C G U
1 2 ... W
the context of the end positions of outermost base pairs crossing over the boundary of fragments (W x W)

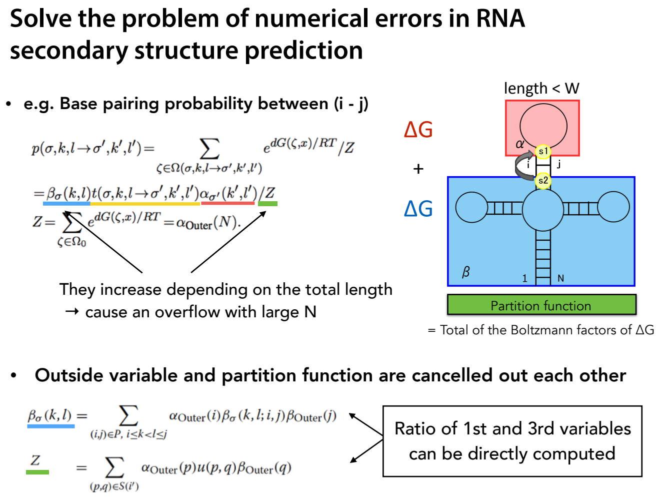
(a) (b)

Figure 1. Example of parallel DP algorithm with 3 clusters based on (a) HMM with 3 types of hidden variable and (b) the inside algorithm of RNA secondary structure prediction with the constraint of maximal span (W). In both of models, computing of DP variables is carried out sequentially in general due to the dependency of proximal regions on the left-hand side of each DP variable. However, parallel computation becomes possible if the context of hidden variables

RNA secondary structure in the wide range of areas, from maturing RNA itself to modulating interaction with other molecules in living cells. Moreover, genome-wide structure analyses based on both computational and experimental approaches might enable to disclose the global RNA landscape and individual differences from such global view.

CONCEPTS OF PARASOR

ParasoR is a novel algorithm that can handle long RNAs for RNA secondary structure prediction with the constraint of maximal span for the distance of base pairing, represented by W. The feasibility of ParasoR for long RNAs consists of two key ideas, parallel computation and a novel calculation algorithm to avoid numerical errors. A technique used in ParasoR is generally applicable for parallel computation of dynamic programming (DP). Let consider the parallelization of hidden Markov model (HMM) with three types of latent variables (Figure 1). 3 clusters can be used simultaneously to calculate DP variables for each partial fragment. However, such distribution cannot produce the same result computed without distribution because the computation of DP variables is sequentially carried out depending on neighboring DP variables in HMM. Not to lose any information by distributed computing, DP variables must be separately stored in each cluster depending on the hidden variable at the left end of fragments. Then, the variables are sequentially multiplied by those of the next cluster only whose hidden variable is matching to the subject to multiply. These products are theoretically same with those computed for the sequence without any fragmentation. For RNA secondary structure prediction, the possible positions of the outermost base pair crossing over the boundary of fragments corresponds to the hidden variable in HMM. Therefore, each DP variable must be separately stored as W different variables for exact computation. In this way, the problem that structure prediction of long RNAs requires long computational time can be solved by making use of a number of computer clusters.



Solve the problem of numerical errors in RNA secondary structure prediction

- e.g. Base pairing probability between (i - j)

$$p(\sigma,k,l \rightarrow \sigma',k',l') = \sum_{\zeta \in \Omega(\sigma,k,l \rightarrow \sigma',k',l')} e^{dG(\zeta,x)/RT}/Z$$

$$= \underline{\beta_\sigma(k,l)}\,\underline{t(\sigma,k,l \rightarrow \sigma',k',l')}\,\underline{\alpha_{\sigma'}(k',l')}\,\underline{/Z}$$

$$Z = \sum_{\zeta \in \Omega_0} e^{dG(\zeta,x)/RT} = \alpha_{\text{Outer}}(N).$$

They increase depending on the total length
→ cause an overflow with large N

- Outside variable and partition function are cancelled out each other

$$\underline{\beta_\sigma(k,l)} = \sum_{(i,j)\in P,\, i \le k < l \le j} \alpha_{\text{Outer}}(i)\beta_\sigma(k,l;i,j)\beta_{\text{Outer}}(j)$$

$$\underline{Z} = \sum_{(p,q)\in S(i')} \alpha_{\text{Outer}}(p)u(p,q)\beta_{\text{Outer}}(q)$$

Ratio of 1st and 3rd variables can be directly computed

Figure 2. Example of expected value computation about secondary structure property based on energy models. An overflow and underflow problem is caused by the increase and subtraction of the partition function and outside variables in accordance with the increase of the sequence length.

On the other hand, the computational prediction of long RNAs also possesses another problem that exact values of DP variables increase depending on the sequence length N. However, since the final product such as base pairing probability is within the range of $[0, 1]$, severe defects of numerical precision may occur through the subtraction by the large partition function. To maintain the precision of DP variables even for long RNAs, ParasoR performs direct computation of the ratio of DP variables 2. Such ratio of DP variables is limited by the magnitude of W and does not suffer the affect of numerical errors. The detailed algorithm is described in the first part of this thesis.

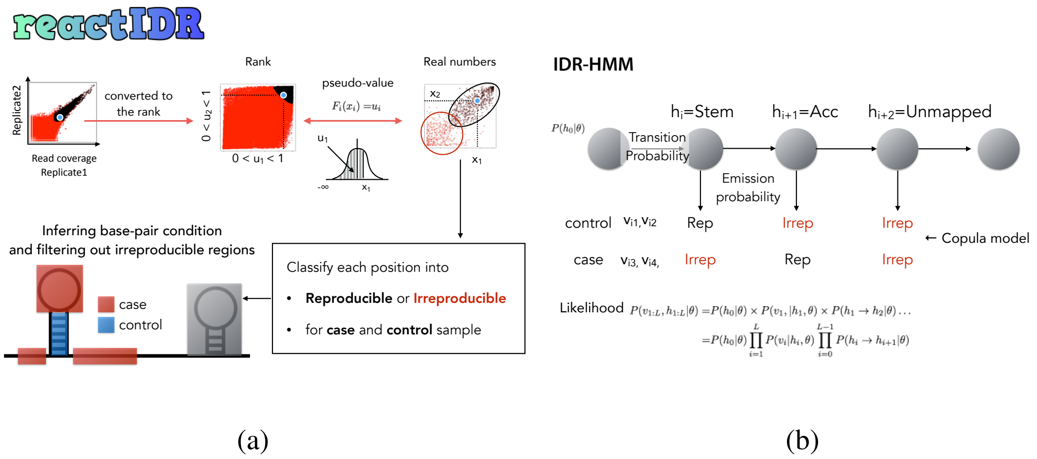

reactIDR
Rank
pseudo-value
Real numbers
Replicate2
converted to the rank
Read coverage Replicate1
0 < u2 < 1
0 < u1 < 1
$F_i(x_i) = u_i$
u1
-∞
x1
x2
Inferring base-pair condition and filtering out irreproducible regions
case
control
Classify each position into
Reproducible or Irreproducible
for case and control sample
IDR-HMM
$P(h_0|\theta)$
Transition Probability
Emission probability
hi=Stem
hi+1=Acc
hi+2=Unmapped
control
vi1,vi2
case
vi3, vi4,
Rep
Irrep
Irrep
Irrep
Rep
Irrep
← Copula model
Likelihood
$P(v_{1:L}, h_{1:L}|\theta) = P(h_0|\theta) \times P(v_1, |h_1, \theta) \times P(h_1 \to h_2|\theta) \dots$
$= P(h_0|\theta) \prod_{i=1}^{L} P(v_i|h_i, \theta) \prod_{i=0}^{L-1} P(h_i \to h_{i+1}|\theta)$
(a)
(b)

Figure 3. (a) An overview of reactIDR. reactIDR estimates IDR considering the similarity of observed read coverage among replicates. Computation of reactivity score and RNA secondary structure prediction are then carried out based on the status of reproducibility of each region in each sample condition such as case and control. (b) Schematic illustration of IDR-HMM.

CONCEPTS OF REACTIDR

reactIDR is designed to extract reliable reactivity data from replicated dataset of high-throughput structure analyses based on their reproducibility. To assess the reproducibility of each position, irreproducible discovery rate (IDR) is estimated in reactIDR for base reactivity measured in several conditions by the mixture copula fitting to infer the joint probability of true and spurious signals. Filtering out of unreliable regions according to IDR must increase the accuracy of downstream analyses such as RNA secondary structure prediction with the assist of reactivity scores. In addition, a novel algorithm, IDR-HMM is implemented in reactIDR to estimate a new type of IDR leveraging local enrichment of reproducible signals, which would detect reliable regions with locally consistent reproducibility. The detailed algorithm is also described in the second part of this thesis.

Part I

Parallel computation of genome-scale RNA secondary structure to detect structural constraints on human genome

1 BACKGROUND

SPLICING EVENTS ARE ASSISTED BY THE RNA SECONDARY STRUCTURE OF INTRONIC REGIONS SURROUNDING SPLICE SITES

The existence of intronic regions is essential for producing the proteomic diversity of eukaryotes through alternative splicing (AS) [25]. To achieve such complex splicing events, most eukaryotes (except an intron-less nucleomorph genome [26]) are equipped with several types of spliceosomes. These complex molecular machines are composed of 5 snRNAs and more than 100 proteins [27]. Spliceosomes recognize splicing motif sites [e.g., two types of splice sites (SSs), donor and acceptor sites, and branch points], so that AS is carried out for introns with a wide range of lengths (from dozens to several tens of thousand nucleotides) in the context of nearly constant exon sizes [28].

Analytical determination of the features that spliceosomes recognize for proper splicing has been an important problem in the field of bioinformatics [29, 30], because AS abnormalities are involved in neuronal disorders and other diseases [27, 31, 32]. Computational approaches have revealed that functional SSs contain characteristic RNA secondary structures around them, in addition to well-known sequence motifs such as flanked GT-AG dinucleotides within introns [30, 33]. In previous research, such characteristics of secondary structure were required to attain a notably high accuracy of AS prediction [34]. An association between splicing and RNA secondary structure has also been validated by several experiments [35, 36, 37]. For instance, homologous *14-3-3ζ* genes of insects were reported to need two types of complementary intronic sequence segments for mutually exclusive splicing, and the alternative exons that were present in the mature mRNA appeared to depend on the stability of their base pairings [38]. Accordingly, explaining the roles of RNA secondary structure in splicing completion and AS regulation is an important endeavor.

COMPUTATIONAL RNA SECONDARY STRUCTURE ANALYSES AROUND SPLICE SITES

Since mutation experiments with structure probing methods such as nuclear magnetic resonance spectroscopy or gel electrophoresis are time-consuming and laborious, very few studies have experimentally validated the complete secondary structures around SSs [36]. High-throughput structure analyses, such as PARS [19], have also been rarely applied to pre-mRNAs because of their paucity of sequencing reads mapped to the intronic regions [24]. Hence, computational prediction has significantly contributed to the comprehensive analyses of RNA secondary structures surrounding SSs.

These studies have revealed that the density of stable base pairs is regulated around SSs in complex ways. Around alternatively spliced exons, stable structures were shown to be overrepresented and conserved relative to constitutive or skipped exons [15, 39]. At the same time, a significant enrichment of single-stranded transcript regions was also observed around splicing enhancer/silencer motifs [40]. This is presumably because splicing enhancer and silencer regions tend to contain binding sites of SR proteins and hnRNPs, which can regulate the splicing efficiency, and the exposure of such regions increases the binding efficiency of these splicing factors [41, 42].

Difficulty in RNA secondary structure prediction of full-length introns

The ends of introns are known to be subject to complex structural constraints; however, little is known about the presence of structural constraints deep inside introns. Although the density of structural motifs of splicing factors will be low compared to the motif around the SSs, it is highly plausible that an intronic region far from the SSs also needs to satisfy various structural requirements for the normal progression of transcription, degradation, and splicing. A detailed structural analysis of intronic sequences would be useful to test the existence of such structural constraints, and would serve as a valuable aid to understanding what makes the introns different from intergenic regions. Nevertheless, very few studies have examined the structure propensity of full-length introns and pre-mRNAs, owing to the prohibitive time complexity of global structure prediction; the original mfold and McCaskill's algorithms require $O(N^3)$ time complexity for input sequence length N [10, 11]. Because it is computationally infeasible to apply the algorithms to long RNAs, some folding programs restrict the allowed sequence distance between base pairing partners to within a given value W [16, 43, 44], which reduces the time complexity to $O(NW^2)$. Even with the maximal-span constraint, the computation time for long transcripts is prohibitive. A more serious problem is that the magnitude of the partition functions grows exponentially with the input length N, which can cause overflow or underflow errors when computing structural properties such as base-pairing probabilities and accessibilities.

To circumvent these problems, sliding-window-based approaches have been developed, in which the folding algorithm is run for each artificial sequence window of length L in the input sequence [43, 45, 17, 46]. Because such algorithms are easily parallelizable and do not cause numerical errors as long as L is not excessively large, they can be a practical tool for genome-wide structure analyses under the current constraints of computational resources. For example, in Ref. [17], the authors used the minimum free energy (MFE) of each sequence window to investigate the structural preferences of transcribed regions. However, since it computes only the energy values of sliding windows and does not predict consistent secondary structures or stochastic structural indicators, detailed structural analyses such as the comparison with experimental data and investigation of the positional specificity of structural constraints were difficult. Other tools for genome-wide MFE-structure prediction using sliding-window approaches have similar problems [43, 45], because they were designed to search for unidentified short structural RNAs whose exact boundaries are unknown but not to analyze the structure propensity of a section of a continuous long RNA. As such, it has remained difficult to examine the positional structure propensity of introns using previous techniques that cannot handle an ensemble of possible structures for long transcripts.

A novel software ParasoR for genome-scale structure analyses

In this study, I developed a novel software, **"ParasoR"**, which enables the distributed computation of various structural features of long RNAs based on the Boltzmann ensemble over globally consistent secondary structures. ParasoR divides dynamic programming (DP) matrices into smaller pieces, such that each piece can be computed by a separate computer node without losing the connectivity information between the pieces. ParasoR avoids the numerical problems of previous algorithms by directly computing the ratios of DP variables whose magnitudes are bounded independently of N. ParasoR can exhaustively compute structural features such as structural profiles [5] and globally consistent γ-centroid structures [47], as well as conventional base pairing probabilities, stem probabilities, and accessibilities. Using ParasoR, I investigated the structural preferences of entire transcribed regions in the human genome. To my knowledge, there is no exhaustive study examining the landscape of the structure stability of human introns using these probabilistic structural indicators. My analyses demonstrate the potential of ParasoR to accelerate large-scale structural analyses performed *in silico*.

2 METHODS

2.1 PARASOR: A PARALLEL SOLUTION FOR LOCAL RNA SECONDARY STRUCTURE ANALYSIS

ParasoR overview

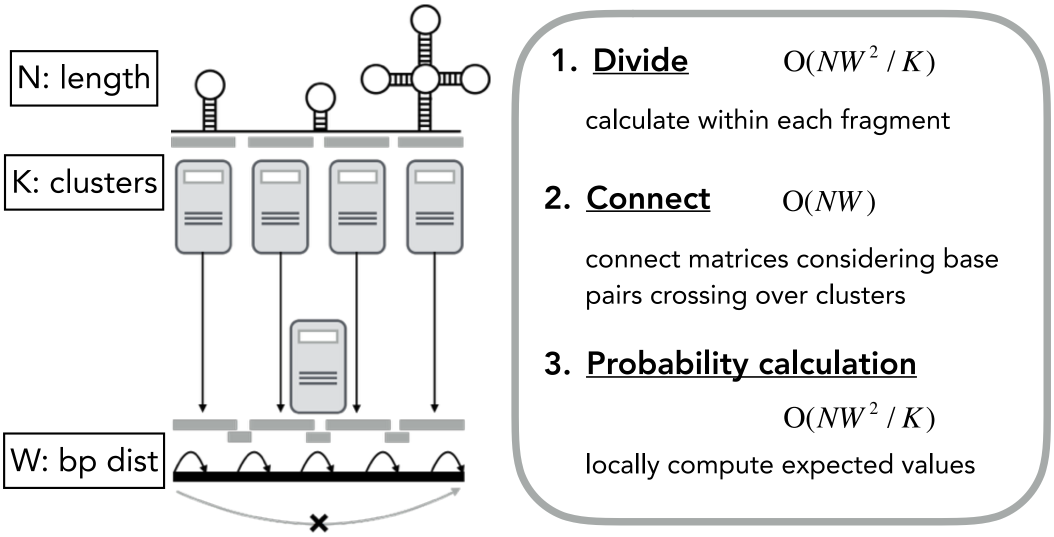

N: length
K: clusters
W: bp dist
1. Divide $O(NW^2/K)$
calculate within each fragment
2. Connect $O(NW)$
connect matrices considering base pairs crossing over clusters
3. Probability calculation
$O(NW^2/K)$
locally compute expected values

Figure 4. A target sequence fragment is assigned to K computational nodes, and $d\alpha_k^h$ is stored in external memory in the Divide procedure to solve the dependency problem that exists around the ends of a given fragment. In the Connect procedure, exact local fold changes $\Delta\alpha$ are computed by the summation of $d\alpha_k^h$ for each pairing pattern at the left end of the assigned fragment. In the computation of expected values, a variety of measures are available using the DP variables whose magnitudes are bounded independently of N, such as $u(k,l)$, $\Delta\alpha$, and $\Delta\beta$.

ParasoR is a novel software application to exactly compute various expected values such as *stem probability* [11, 44] and *accessibility* [9, 48, 49] from the Boltzmann ensemble of global secondary structures, with the constraint of maximal base-pair span. I consider only the structures containing short-range base pairs, since it is well known that the energy model of the secondary structure is inaccurate for predicting distant base pairings [50]. The maximal span constraint limits the structure ensemble to the set of global secondary structures that contain only base pairs with spanning lengths $\leq W$. In Ref.[46], it is shown that the constraint of maximal span for the distance of base pairing can improve the accuracy of structure prediction. This constraint also reduces the computational complexity of structure prediction from $O(N^3)$ to $O(NW^2)$, as described in the Background section.

ParasoR is the only tool developed to date that can make global structure predictions for long RNAs (even for ~3G base sequences). This high scalability of ParasoR is attained by the following two techniques: (1) solving numerical error problems by considering only the ratios of dynamic DP variables, and (2) allowing distributed computation for a computer cluster. Owing to its memory- and disk-saving design, ParasoR is also useful for small-scale studies that use a single computer.

Figure 4 shows a ParasoR's workflow. In ParasoR, the structure prediction is carried out based on the Rfold grammar [44] and the inside-outside algorithm. For a given set of sequences, ParasoR constructs a database of local fold changes of inside and outside DP variables $\Delta\alpha$ and $\Delta\beta$ through the Divide and Connect procedures. From this database, ParasoR computes the following features for any queried region: (i) *base-pairing probability*; (ii) *stem probability*, represented as $p_{\text{stem}}(i)$ at i-th position; (iii) *accessibility*; (iv) *structural profiles* $p_{\delta}(i)$, which represents the probability that the position i is a part of specific loop type δ = *bulge*, *exterior*, *hairpin*, *multi*, or *interior* [5]; and (v) a globally consistent secondary structure of credible base pairs (e.g., γ-centroid structure [47] with $\gamma \leq 1$).

This database can be used repeatedly for the fast structure simulation of similar but different sequences, such as those with point mutations or incomplete RNAs that appear during transcription elongation. ParasoR can also be applicable for the fast simulation of co-transcriptional splicing by using partial DP variables in the database that correspond to partially transcribed RNAs.

Time complexity of ParasoR

For a given RNA sequence, ParasoR exactly computes various expected values from the Boltzmann ensemble of secondary structures under a maximal pair-distance constraint. It avoids numerical errors by dealing with only the ratios of DP variables, which do not change in magnitude as the sequence length N changes. To allow distributed computing, ParasoR divides the DP matrices into smaller pieces without losing their mutual dependencies. The computational complexities are given by either

1. $O(NW^2/K + NW)$ time, $O(N/K + W^2)$ memory for each node, and $O(NW)$ disk space or
2. $O(NW^2/K + KW^2)$ time, $O(N/K + W^2)$ memory for each node, and $O(N + KW^2)$ disk space, which requires less disk space than (i) but twice the computational time $O(NW^2/K)$ for DP matrices construction.

Here, N denotes the input sequence length; W denotes the maximal span of base pairs; and K denotes the number of available computer nodes.

2.2 RFOLD ALGORITHM

To explain ParasoR algorithm, I first introduce Rfold grammar [44], which is based on the conventional energy models subject to the constraint of maximal distance between base pairs. Throughout this article, I use a grammatical formulation of secondary structure developed in the Rfold model. In this section, I describe the detail of Rfold grammar and how to calculate stochastic expected values based on the Boltzmann ensemble.

Rfold grammar

In the Rfold model, the transition rules are expressed as follows.

$$
\begin{aligned}
\text{Outer} &\longrightarrow \epsilon|\text{Outer}\cdot a|\text{Outer}\cdot\text{Stem}\\
\text{Stem} &\longrightarrow b<\cdot\text{Stem}\cdot b>|b<\cdot\text{StemEnd}\cdot b>\\
\text{StemEnd} &\longrightarrow s_n|s_m\cdot\text{Stem}\cdot s_n\ (m+n>0)|\text{Multi}\\
\text{Multi} &\longrightarrow a\cdot\text{Multi}|\text{MultiBif}\\
\text{MultiBif} &\longrightarrow \text{Multi1}\cdot\text{Multi2}\\
\text{Multi1} &\longrightarrow \text{MultiBif}|\text{Multi2}\\
\text{Multi2} &\longrightarrow \text{Multi2}\cdot a|\text{Stem}
\end{aligned}
$$

In these rules, ϵ represents the null terminal symbol; a, an unpaired nucleotide; s_k, an unpaired subsequence whose length is k; and $b<, b>$, a base pair. Rfold enables an enumeration of possible structure patterns following these transition rules. In addition, Rfold sums up the Boltzmann factor of energies of the partial structure that belongs to each state based on the inside-outside algorithm so that Rfold calculates the expected value of certain partial structure or transition based on the Boltzmann ensemble. Each inside and outside variable is sequentially computed using a dynamic programming technique as follows.

$$
\alpha_{\text{Stem}}(i,j)=\sum\begin{cases}\alpha_{\text{Stem}}(i+1,j-1)\cdot t(\text{Stem}\rightarrow\text{Stem})\\ \alpha_{\text{StemEnd}}(i+1,j-1)\cdot t(\text{Stem}\rightarrow\text{StemEnd})\end{cases}
$$

$$
\alpha_{\text{MultiBif}}(i,j)=\sum\begin{cases}\alpha_{\text{Multi1}}(i,k)\cdot\alpha_{\text{Multi2}}(k,j)\cdot t(\text{MultiBif}\rightarrow\text{Multi1}\cdot\text{Multi2})\\ \text{for } i<k<j\end{cases}
$$

$$
\alpha_{\text{Multi2}}(i,j)=\sum\begin{cases}\alpha_{\text{Stem}}(i,j)\cdot t(\text{Multi2}\rightarrow\text{Stem})\\ \alpha_{\text{Multi2}}(i,j-1)\cdot t(\text{Multi2}\rightarrow\text{Multi2})\end{cases}
$$

$$
\alpha_{\text{Multi1}}(i,j)=\sum\begin{cases}\alpha_{\text{Multi2}}(i,j)\cdot t(\text{Multi1}\rightarrow\text{Multi2})\\ \alpha_{\text{MultiBif}}(i,j)\cdot t(\text{Multi}\rightarrow\text{MultiBif})\end{cases}
$$

$$
\alpha_{\text{Multi}}(i,j)=\sum\begin{cases}\alpha_{\text{Multi}}(i+1,j)\cdot t(\text{Multi}\rightarrow\text{Multi})\\ \alpha_{\text{MultiBif}}(i,j)\cdot t(\text{Multi}\rightarrow\text{MultiBif})\end{cases}
$$

$$
\alpha_{\text{StemEnd}}(i,j)=\sum\begin{cases}t(\text{StemEnd}\rightarrow(\text{Hairpin}))\\ \alpha_{\text{Stem}}(i',j')\cdot t(\text{StemEnd}\rightarrow(\text{Interior})\rightarrow\text{Stem})\\ \text{for } i\le i'<j'\le j,\ 0<(j-j')+(i'-i)\le C\\ \alpha_{\text{Multi}}(i,j)\cdot t(\text{StemEnd}\rightarrow\text{Multi})\end{cases}
$$

$$
\alpha_{\text{Outer}}(j)=\sum\begin{cases}1\text{ if } j=0\\ \alpha_{\text{Outer}}(j-1)\cdot t(\text{Outer}\rightarrow\text{Outer})\\ \alpha_{\text{Outer}}(k)\cdot\alpha_{\text{Stem}}(k,j)\cdot t(\text{Outer}\rightarrow\text{Outer}\cdot\text{Stem})\\ \text{f}or\ (j-W-1)\le k<j\end{cases}
$$

$$\beta_{\text{Outer}}(j) = \sum \begin{cases} 1 \text{ if} j = N + 1 \\ \beta_{\text{Outer}}(j+1) \cdot t'(\text{Outer} \rightarrow \text{Outer}) \\ \alpha_{\text{Stem}}(j,k) \cdot \beta_{\text{Outer}}(k) \cdot t'(\text{Outer} \rightarrow \text{Outer} \cdot \text{Stem}) \\ \text{for } j < k \leq (j + W + 1) \end{cases}$$

$$\beta_{\text{StemEnd}}(i,j) = \quad \beta_{\text{Stem}}(i-1, j+1) \cdot t'(\text{Stem} \rightarrow \text{StemEnd})$$

$$\beta_{\text{Multi}}(i,j) = \sum \begin{cases} \beta_{\text{StemEnd}}(i,j) \cdot t'(\text{StemEnd} \rightarrow \text{Multi}) \\ \beta_{\text{Multi}}(i-1,j) \cdot t'(\text{Multi} \rightarrow \text{Multi}) \end{cases}$$

$$\beta_{\text{Multi1}}(i,j) = \sum \begin{cases} \beta_{\text{MultiBif}}(i,k) \cdot \alpha_{\text{Multi2}}(j,k) \cdot t'(\text{MultiBif} \rightarrow \text{Multi1} \cdot \text{Multi2}) \\ \text{for } j < k \leq (i + W + 1) \end{cases}$$

$$\beta_{\text{Multi2}}(i,j) = \sum \begin{cases} \beta_{\text{Multi2}}(i, j+1) \cdot t'(\text{Multi2} \rightarrow \text{Multi2}) \\ \beta_{\text{Multi1}}(i,j) \cdot t'(\text{Multi1} \rightarrow \text{Multi2}) \\ \beta_{\text{MultiBif}}(k,j) \cdot \alpha_{\text{Multi1}}(k,i) \cdot t'(\text{MultiBif} \rightarrow \text{Multi1} \cdot \text{Multi2}) \\ \text{for}(j - W - 1 \leq k < i) \end{cases}$$

$$\beta_{\text{MultiBif}}(i,j) = \sum \begin{cases} \beta_{\text{Multi1}}(i,j) \cdot t'(\text{Multi1} \rightarrow \text{MultiBif}) \\ \beta_{\text{Multi}}(i,j) \cdot t'(\text{Multi} \rightarrow \text{MultiBif}) \end{cases}$$

$$\beta_{\text{Stem}}(i,j) = \sum \begin{cases} \alpha_{\text{Outer}}(i) \cdot \beta_{\text{Outer}}(j) \cdot t'(\text{Outer} \rightarrow \text{Outer} \cdot \text{Stem}) \\ \beta_{\text{StemEnd}}(i',j') \cdot t'(\text{StemEnd} \rightarrow (\text{Interior}) \rightarrow \text{Stem}) \\ \text{for } i' \leq i < j \leq j', 0 < (i - i') + (j' - j) \leq C \\ \beta_{\text{Multi2}}(i,j) \cdot t'(\text{Multi2} \rightarrow \text{Stem}) \\ \beta_{\text{Stem}}(i-1, j+1) \cdot t'(\text{Stem} \rightarrow \text{Stem}) \end{cases}$$

2.2.1 *Stem probability in Rfold algorithm*

Any secondary structure ζ of sequence x is specified by a list of base pairs. In the conventional Turner energy model, a base pair of x_k and x_l must be one of the canonical base pairs {AU, UA, CG, GC, GU, UG}, and the distance between them should satisfy $5 \leq (l - k + 1)$. I designate a position pair (i, j) an outermost pair if (x_{i+1}, x_j) forms a base pair and there is no base pair that encloses (i, j) in ζ. Since the maximal span constraint is imposed in Rfold model, the outermost pair (i, j) also satisfies $(j - i) \leq W$. Then, the structure ζ is uniquely decomposed into the set of non-overlapping substructures that are enclosed by an outermost pair for each and fragments of exterior loops between or flanking them. I define the set of potential outermost pairs of x as $P = \{(i, j) \mid (x_{i+1}, x_j)$ is one of the canonical base pairs and $5 \leq j - i \leq W\}$.

In the Rfold model, there are 6 non-terminal symbols, in which the transition between Outer and Stem state corresponds to the transition from an exterior loop to the outermost base pair.

A partition function Z is calculated by an inside variable of Outer state α_{Outer} and Stem state α_{Stem} as follows.

$$\alpha_{\text{Outer}}(j) = \sum_{\zeta \in \Omega(1,j)} e^{dG(\zeta, x_{1,j})/RT} = \sum \begin{cases} 1 \ \text{ if } j = 0 \\ \alpha_{\text{Outer}}(j-1) \cdot t(\text{Outer} \to \text{Outer}) \\ \alpha_{\text{Outer}}(k) \cdot \alpha_{\text{Stem}}(k,j) \cdot t(\text{Outer} \to \text{Outer} \cdot \text{Stem}) \\ \ \text{for } (j - W) \leq k < j \end{cases}$$

$$\beta_{\text{Outer}}(j) = \sum_{\zeta \in \Omega(j+1,N)} e^{dG(\zeta, x_{j+1,N})/RT}$$

$$Z = \sum_{\zeta \in \Omega_0} e^{dG(\zeta, x_{1,N})/RT} = \alpha_{\text{Outer}}(N) = \beta_{\text{Outer}}(0)$$

Here, $dG(\zeta, x)$ represents the free energy for sequence x with structure ζ; R represents the gas constant; T represents the absolute temperature; t represents the Boltzmann factor for the transition; Ω_0 and $\Omega(k,l)$ represent the set of all secondary structures of sequence x and subsequence $x_{k,l}$ between k and l, respectively.

The expected values in this study such as stem probability are estimated as the sum of the state transition probabilities $p(\sigma, k, l \to \sigma', k', l')$.

$$\begin{aligned} p(\sigma, k, l \to \sigma', k', l') &= \sum_{\zeta \in \Omega(\sigma,k,l \to \sigma',k',l')} e^{dG(\zeta,x)/RT}/Z \\ &= \beta_{\sigma}(k,l) t(\sigma, k, l \to \sigma', k', l') \alpha_{\sigma'}(k', l')/Z \end{aligned} \tag{1}$$

where σ and σ' represent non-terminal symbols of the Rfold grammar; k, l, k', and l' represent sequence positions; $\Omega(\sigma, k, l \to \sigma', k', l')$ represents the set of secondary structures containing the transition $(\sigma, k, l \to \sigma', k', l')$; $\alpha_{\sigma'}(k', l')$ represents the inside variable; and $\beta_{\sigma}(k, l)$ represents the outside variable.

Summation of the state transition probability following this equation permits the computation of several measures of structure such as base pairing probability, accessibility, and structural profiles [49, 5].

A type of expected values called the stem probability was extensively examined in this thesis. The stem probability $p_{\text{stem}}(i)$ (which is equal to $1 - \text{accessibility}(i)$) at sequence position i is the probability that the base at position i is within a stem and is defined as $p_{\text{stem}}(i) = \sum_{j(>i)} p(i,j) + \sum_{j(<i)} p(j,i)$, where $p(i,j)$ represents the base-pairing probability [11].

For long sequences such as mRNA and pre-mRNA, however, the increased number of possible structures makes it almost impossible to calculate Z and β directly even though the increase of α is restricted by the maximal span.

2.3 PARASOR ALGORITHM

2.3.1 *Avoiding numerical problems by using a ratio of DP variable and partition function*

In Equation 1, the magnitudes of α and t do not change with N, since they are computed from the subsequence $x_{k,l}$ whose length does not exceed W. Z and β are, however, the sums of Boltzmann factors for the subsequences of length $O(N)$, and grow exponentially with N. On the other hand, the probability $p(\sigma, k, l \to \sigma', k', l')$ should be between 0 and 1, and so a large cancellation between $\beta_{\sigma}(k, l)$ and Z must occur, which reduces the numerical precision. The

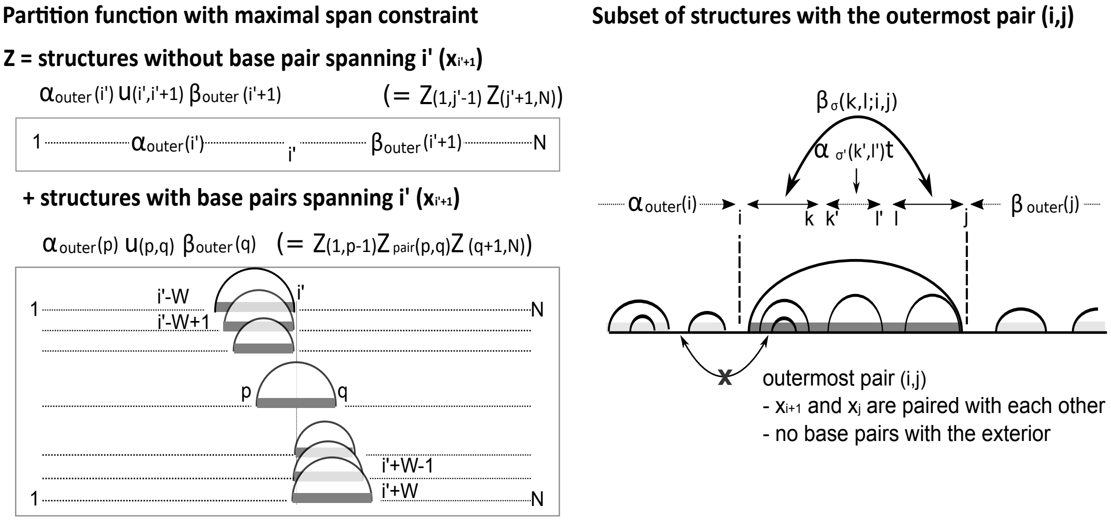

Partition function with maximal span constraint
Z = structures without base pair spanning i' (xi'+1)
αouter(i') u(i',i'+1) βouter(i'+1)
(= Z(1,j'-1) Z(j'+1,N))
1
αouter(i')
i'
βouter(i'+1)
N
+ structures with base pairs spanning i' (xi'+1)
αouter(p) u(p,q) βouter(q)
(= Z(1,p-1)Zpair(p,q)Z(q+1,N))
1
i'-W
i'
N
i'-W+1
p
q
i'+W-1
i'+W
1
N
Subset of structures with the outermost pair (i,j)
βσ(k,l;i,j)
α σ'(k',l')t
αouter(i)
i
k
k'
l'
l
j
β outer(j)
outermost pair (i,j)
- xi+1 and xj are paired with each other
- no base pairs with the exterior

Figure 5. Schematic illustration of the decomposition of partition function and correspondence between variables, their assigned regions, and the structure constraints. One example structure is expressed by the arcs representing the base pairs. By the summation of expected values for structures with each outermost pair, ParasoR can calculate an expected value from the ensemble of global structures.

cancellation is assured because the contributions from the structures far outside of (k, l) are almost the same. This can be seen from the following decompositions.

$$\beta_\sigma(k,l) = \sum_{(i,j)\in P,\ i\le k<l\le j} \alpha_{\text{Outer}}(i)\beta_\sigma(k,l;i,j)\beta_{\text{Outer}}(j) \tag{2}$$

$$\begin{aligned} Z =& \alpha_{\text{Outer}}(i')t(\text{Outer} \to \text{Outer})\beta_{\text{Outer}}(i'+1)+ \\ & \sum_{(i,j)\in P,\ i\le i'<j} \alpha_{\text{Outer}}(i)t(\text{Outer}, i, j \to \text{Outer}\cdot\text{Stem}, i, j)\alpha_{\text{Stem}}(i,j)\beta_{\text{Outer}}(j) \qquad (3) \\ =& \sum_{(p,q)\in S(i')} \alpha_{\text{Outer}}(p)u(p,q)\beta_{\text{Outer}}(q) \end{aligned}$$

$$u(p,q) = \begin{cases} t(\text{Outer} \to \text{Outer}) & \text{if } p+1=q \\ t(\text{Outer}, p, q \to \text{Outer}\cdot\text{Stem}, p, q)\alpha_{\text{Stem}}(p,q) & \text{otherwise} \end{cases}$$

Here, i' can be set to any position, the set $S(i')$ is defined as $\{(i,j) \in P \mid i \le i' < j\} \bigcup \{(i', i'+1)\}$ for position i', and $\beta_\sigma(k,l;i,j)$ are the outside variables for the subsequence located between the outermost pair (i, j), satisfying the initial condition $\beta_{\text{Stem}}(i,j;i,j) = t(\text{Outer}, i, j \to \text{Outer}\cdot\text{Stem}, i, j)$. Equation 2 follows, because the outside variables are the sum of the contributions of all possible patterns of outermost pairs. Equation 3 also follows, because a base represented by the position i' is either within the outermost pair (i, j) or is an exterior base (illustrated in Figure 5). It should be noted that the dynamic range $(i, j) \in P$ in Equations 2 and 3 can be simplified to the set $\{(i,j) \mid i \le j, (j-i) \le W\}$, when the values of $\beta_\sigma(k,l;i,j)$ and $\alpha_{\text{Stem}}(i,j)$ are zero for $(i,j) \notin P$.

For those who are familiar with the partition function algorithms, it is noted that Equation 3 for any position i' is also represented by the decomposition of the partition function Z into the sum of those of smaller subsequences for any nucleotide position j', as below.

$$Z = Z(1, j' - 1)Z(j' + 1, N) + \sum_{(i,j)\in P',\ i\leq j'\leq j} Z(1, i - 1)Z_{\text{pair}}(i, j)Z(j + 1, N)$$

Here, P' is the set $\{(i, j) \mid (x_i, x_j)$ is one of the canonical base pairs, $5 \leq (j - i + 1) \leq W\}$, $Z(k, l)$ is the partition function for subsequence $x_{k,l}$, and $Z_{\text{pair}}(k, l)$ is the partition function of subsequence $x_{k,l}$ with an outermost pair between x_k and x_l (note that $Z(1, i - 1)$ and $Z(j + 1, N)$, etc., actually need to include the contributions of dangling or mismatch scores that depend on the exterior bases outside of the sequence ranges).

Next, I define the ratio of the DP variables and partition function $r(i, j)$ for any position pair (i, j) such that $i \leq j, (j - i) \leq W$:

$$\begin{aligned}
r(i, j) &:= \frac{Z}{\alpha_{\text{Outer}}(i)\beta_{\text{Outer}}(j)} \\
&= \frac{\sum_{p,q\in S(i)} \alpha_{\text{Outer}}(p)u(p, q)\beta_{\text{Outer}}(q)}{\alpha_{\text{Outer}}(i)\beta_{\text{Outer}}(j)} \\
&= \sum_{p,q} \left\{ \frac{\alpha_{\text{Outer}}(p)}{\alpha_{\text{Outer}}(i)}\ u(p, q)\ \frac{\beta_{\text{Outer}}(q)/\beta_{\text{Outer}}(i)}{\beta_{\text{Outer}}(j)/\beta_{\text{Outer}}(i)} \right\} \\
&= \sum_{p,q} \left\{ \frac{\alpha_{\text{Outer}}(p)}{\alpha_{\text{Outer}}(p + 1)} \cdots \frac{\alpha_{\text{Outer}}(i - 1)}{\alpha_{\text{Outer}}(i)}\ u(p, q) \right. \\
&\quad \left. \frac{\beta_{\text{Outer}}(q)}{\beta_{\text{Outer}}(q - 1)} \cdots \frac{\beta_{\text{Outer}}(i + 1)}{\beta_{\text{Outer}}(i)} \frac{\beta_{\text{Outer}}(j - 1)}{\beta_{\text{Outer}}(j)} \cdots \frac{\beta_{\text{Outer}}(i)}{\beta_{\text{Outer}}(i + 1)} \right\} \\
&= \sum_{p,q} \left\{ \prod_{h=p}^{i-1} 1/\Delta\alpha(h) \right\}\ u(p, q)\ \left\{ \prod_{h=i}^{j-1} \Delta\beta(h) \right\} / \left\{ \prod_{h=i}^{q-1} \Delta\beta(h) \right\} \\
\Delta\alpha(h) &:= \alpha_{\text{Outer}}(h + 1)/\alpha_{\text{Outer}}(h) \\
\Delta\beta(h) &:= \beta_{\text{Outer}}(h)/\beta_{\text{Outer}}(h + 1)
\end{aligned}$$

In my implementation, $\Delta\alpha$ and $\Delta\beta$ are stored as logarithmic values; hence, the summations in the above formula are replaced by logsum operations. In the following subsections, I show a DP algorithm that directly computes these values without recourse to α_{Outer} and β_{Outer}. On the other hand, inner variables $u(p, q)$ can be computed without numerical difficulties by using the ordinary inside algorithm. In this manner, ParasoR can avoid the computation of variables that exponentially increase with N for r. Then, the fold change $\beta_\sigma(k, l)/Z$ can be represented by the outside variable for a subsequence between i and j ($|j - i| = O(W)$) and $r(i, j)$, as below.

$$\beta_\sigma(k, l)/Z = \sum_{(i,j)\in P,\ i\leq k<l\leq j} \beta_\sigma(k, l; i, j)/r(i, j)$$

In this way, ParasoR can compute an expected value only by the variables whose absolute values are bounded independently of N.

2.3.2 *ParasoR algorithm and its implementation*

ParasoR is a parallel extension of the Rfold algorithm intended for the prediction of genome-scale sequences, and it is accomplished via three procedures: *Divide*, *Connect*, and *Probability* *calculation* procedure. One of the strongest features in ParasoR is its construction of databases of α_{Outer} and β_{Outer} in the form of their fold changes such as $d\alpha$, $d\beta$, $\Delta\alpha$, and $\Delta\beta$; each element is obtained by comparison with adjacent elements. The algorithm evaluates local increase rations of α_{Outer} and β_{Outer} in logarithmic scale, $\alpha(i) = \log(\alpha_{\text{Outer}}(i))$ and $\beta(i) = \log(\beta_{\text{Outer}}(i))$. Because a logarithmic transformation is applied to all internal variables of ParasoR in order to reduce numerical errors, I use $\alpha(i)$ and $\beta(i)$ hereafter to explain ParasoR algorithm.

When K computational nodes are available to calculate the secondary structure of a sequence of length N and maximal span W, ParasoR requires $O(NW^2/K)$ computational time in the Divide procedure, $O(NW)$ or $O(KW^2 + NW^2/K)$ in the Connect procedure, and $O(NW^2/K)$ in the Probability calculation procedure.

ParasoR is written in the C++ language, and it uses a portion of the source code of the ViennaRNA package [51] for energy parameters and Centroid fold for visualization [47]. The ParasoR source code and a detailed manual are available https://sites.google.com/site/cawatchm/.

In the following sections, I describe how to calculate $d\alpha$, $d\beta$, $\Delta\alpha$, and $\Delta\beta$ to obtain r. Then, calculation of the expectation values is formulated using r to avoid the direct computation of Z and β whose magnitudes grow exponentially with N.

2.3.3 *Divide procedure*

To explain the methodology of the Divide procedure, I focus on the calculation of $d\alpha$ and $d\beta$ for the subsequence from the sth to eth positions after its assignment to a single node in the Divide procedure. First, I describe how to calculate $d\alpha$ using v, which is defined as follows to absorb all required energy during the transition between the Stem and Outer state.

$$v_{i,j,h} := \begin{cases} -\infty & \text{if } i < (s+1) \text{ and } i \neq (s-h) \\ \log(t(\text{Outer} \to \text{Outer})) & \text{else if } j = i+1 \\ \log\left(\alpha_{\text{Stem}}(i,j) \cdot t(\text{Outer} \to \text{Outer} \cdot \text{Stem})\right) & \text{otherwise} \end{cases}$$

Then, α is simply expressed using v.

$$\begin{aligned} \alpha(i) &= \log \alpha_{\text{Outer}}(i) \\ &= \log\left(\sum_{k=i-W}^{i} \exp(\alpha(k) + v_{k,i,i-k})\right) \\ &= \log\left(\sum_{h=0}^{W} \exp(\alpha(s-h) + \alpha_{i,h})\right) \end{aligned} \tag{4}$$

where $\alpha_{i,h}$ is a partial summation of α and recursively obtained as below.

$$\alpha_{i,h} := \begin{cases} 0 & \text{if } i < (s+1) \\ \log\left(\sum_{k=1}^{W} \exp(\alpha_{i-k,h} + v_{i-k,i,h})\right) & \text{else} \end{cases}$$

In Eq. 4, $\alpha_{i,h}$ is calculated from the data available in the same computational node while $\alpha(s-h)$ depends on the results of the previous node. To cancel out the contribution of $\alpha(s-h)$, I consider $d\alpha_{i,h}$, or the local differences of α between two adjacent values.

$$\begin{aligned} d\alpha_{i,h} &:= \alpha_{i,h} - \alpha_{i-1,0} \\ &= \begin{cases} 0 & \text{if } i < (s+1) \\ \log \sum_{k=1}^{W} \exp\left(d\alpha_{i-k,h} + v_{i-k,i,h} - \sum_{j=i-k}^{i-1} d\alpha_{j,0}\right) & \text{if } i \geq (s+1) \end{cases} \end{aligned}$$

In this way, ParasoR computes $d\alpha_{k,h}$ and constructs a partial database of $d\alpha_{k,h}$ using a single node in $O(NW^2/K)$ computational time.

In a similar way, $d\beta$ is also calculated by defining another v whose boundary condition is different from the previous v.

$$v_{i,j,h} := \begin{cases} -\infty & \text{if } e < j \text{ and } j \neq (e+h) \\ \log(t(\text{Outer} \to \text{Outer})) & \text{else if} j = i+1 \\ \log(\alpha_{\text{Stem}}(i,j) \ \cdot t(\text{Outer} \to \text{Outer} \cdot \text{Stem})) & \text{otherwise} \end{cases}$$

$$\begin{aligned} \beta(i) &= \log \beta_{\text{Outer}}(i) \\ &= \log\left(\sum_{k=i+W}^{i} \exp(\beta(k) + v_{i,k,i-k})\right) \\ &= \log\left(\sum_{h=0}^{W} \exp(\beta(e+h) + \beta_{i,h})\right) \end{aligned}$$

$$\beta_{i,h} := \begin{cases} 0 & \text{if } e < i \\ \log\left(\sum_{k=1}^{W} \exp(\beta_{i+k,h} + v_{i,i+h,h})\right) & \text{else} \end{cases}$$

$$\begin{aligned} d\beta_{i,h} &:= \beta_{i,h} - \beta_{i+1,0} \\ &= \begin{cases} 0 \\ \log \sum_{k=1}^{W} \exp\left(d\beta_{i+k,h} + v_{i,i+k,h} - \sum_{j=i+1}^{i+k} d\beta_{j,0}\right) \end{cases} \end{aligned}$$

2.3.4 *Connect procedure*

The variables $d\alpha$ and $d\beta$ in the previous section contain the dependency on the fragmentation pattern. In the Connect procedure, ParasoR integrates these variables to calculate unique local fold changes $\Delta\alpha$ and $\Delta\beta$ for each position as follows.

$$\begin{aligned} \Delta\alpha_j &:= \alpha_j - \alpha_{j-1} \\ &= \log\left(\frac{\sum_{h=0}^{W} \exp(d\alpha_{j,h} + d\alpha_{j-1,0} - \sum_{i=s-h+1}^{s} \Delta\alpha_i)}{\sum_{h=0}^{W} \exp(d\alpha_{j-1,h} - \sum_{i=s-h+1}^{s} \Delta\alpha_i)}\right) \end{aligned} \tag{5}$$

$$\begin{aligned} \Delta\beta_j &:= \beta_j - \beta_{j+1} \\ &= \log\left(\frac{\sum_{h=0}^{W} \exp(d\beta_{j,h} + d\beta_{j+1,0} - \sum_{i=e}^{e+h} \Delta\beta_i)}{\sum_{h=0}^{W} \exp(d\beta_{j+1,h} - \sum_{i=e}^{e+h} \Delta\beta_i)}\right) \end{aligned} \tag{6}$$

where $\Delta\alpha$ and $\Delta\beta$ are computed by the sum of Boltzmann factors of the free energy of the secondary structures inferred within the subsequences of length $W \times$ (constant value) or less. At the same time, the calculation of these values indirectly includes the influences of long flanking regions (discussed in the Result section).

I implemented the Connect procedure in two different ways. In the first way, ParasoR stores the whole matrix of $d\alpha$ and $d\beta$ in the Divide procedure and sequentially computes $\Delta\alpha$ and $\Delta\beta$ in a single node. This implementation enables the removal of redundant calculations at the expense of reducing the size of external files. This procedure requires $O(NW)$ computational

Algorithm 1 Memory-saving algorithm in the Divide and Connect procedures

```
1: Compute $d\alpha_k^h$ $(s \le k \le e)$ and $d\beta_k^h$ $(s \le k \le e)$ in multiple nodes
2: Store only $d\alpha_k^h$ $(e - W \le k \le e)$ and $d\beta_k^h$ $(s \le k \le s + W)$ in the disk
3: Compute and save $\Delta\alpha(k)$ $(e - W \le k \le e)$ and $\Delta\beta(k)$ $(s \le k \le s + W)$ following Eq. 5
   and Eq. 6 for all the divided segments using a single node
4: Compute $d\alpha_k^h$ $(s \le k \le e)$ and $d\beta_k^h$ $(s \le k \le e)$ in multiple nodes again
5: Compute and save $\Delta\alpha$ and $\Delta\beta$ for all sequence positions in multiple nodes using the partial
   $\Delta\alpha$ and $\Delta\beta$ saved previously
```

time to establish databases for $\Delta\alpha$ and $\Delta\beta$ of file size $O(N)$ using $d\alpha$ and $d\beta$ which need a temporal storage file of $O(NW^2)$. This temporal storage becomes, however, very large, up to around 10 TB for the human genome sequence in double (8-byte) precision.

Accordingly, ParasoR also uses another implementation that constructs a part of $d\alpha$ and $d\beta$ to compute a vector of $\Delta\alpha$ and $\Delta\beta$ of length only $(W + 1)$ for the ends of the assigned sequence from the saved $d\alpha$ and $d\beta$. Then, each node recalculates $d\alpha$ and $d\beta$ and completes the databases of $\Delta\alpha$ and $\Delta\beta$ for the assigned region using the partial vector of $\Delta\alpha$ and $\Delta\beta$. A pseudo code for this algorithm is given (Algorithm 1). This procedure only requires the temporary storage for $d\alpha_k^h$ $(e - W \le k \le e)$ and $d\beta_k^h$ $(s \le k \le s + W)$ of size $O(KW^2)$. The time complexity of the first and fourth steps is $O(NW^2/K)$ for each node. The time complexity of the third step is $O(KW^2)$ for a single node. The time complexity of the last step is $O(NW/K)$ for each node.

2.3.5 *Probability calculation procedure*

For whole genome sequences, a partition function Z becomes so large that I cannot compute Z directly. However, Z is required in the form of β/Z in the probability formula, and the increases of β and Z are considered to balance each other because a large number of possible structures would increase both of them. In this section, I show that $r = Z/\beta$ is evaluated only by local variables such as $\Delta\alpha$ and $\Delta\beta$, which are restricted to increase only up to the constant value depending on the maximal span rather than the whole sequence length.

As I described in the subsection “Stem probability in Rfold algorithm”, I characterize the set of the “local” structures by locally maximal, non-overlapping substructures enclosed by a base pair. The pair of bases that encloses each local structure of subsequence $x_{i,j}$ between i and j is referred to as the outermost pair. The sequence length of the local structures $(|j - i|)$ must not exceed W in ParasoR algorithm because base pairs between more distant bases are not allowed by the constraint of maximal span. Thus, all possible structures should consist of non-overlapping local structures and fragments of exterior loops between them. In Equation 1, the magnitudes of α and t do not change with N since they are computed from the subsequence $x_{k,l}$ whose length does not exceed W. Z and β are, however, the sums of Boltzmann factors for the subsequences of length $O(N)$ and grow exponentially with N. On the other hand, the probability $p(\sigma, k, l \to \sigma', k', l')$ should be between 0.0 and 1.0, and so a large cancellation between $\beta_\sigma(k, l)$ and Z must occur, which reduces the numerical precision. The cancellation is assured because the contributions from the structures far outside of (k, l) are almost the same. This can be seen from the decomposition of Equations 2 and 3. Equation 2 follows because the outside variables are the sum of the contributions of all possible patterns of outermost pairs.

Equation 3 also follows because a base at any position i' ($x_{i'+1}$) is either within the outermost pair (p,q) or is an exterior base.

Then I define $r(i,j)$ for the computation of expected values of local structures on subsequence $x_{i+1,j}$ as below.

$$r(i,j) = \log\left(\sum_{p,q}\left\{\prod_{h=p}^{i-1}\exp(-\Delta\alpha_h)\right\}\ u(p,q)\ \left\{\prod_{h=i}^{q-1}\exp(-\Delta\beta_h)\right\}\left\{\prod_{h=i}^{j-1}\Delta\exp(\beta_h)\right\}\right)$$

In the previous subsections, I have already shown the DP algorithm that directly computes $\Delta\alpha$ and $\Delta\beta$ without recourse to α_{Outer} and β_{Outer}. The inner variables $u(p,q)$ can be computed without numerical difficulties by using the ordinary inside algorithm.

In this manner, I can avoid the computation of variables that exponentially increase with N for $r(i,j)$. Using $r(i,j)$, a fold change $\beta_\sigma(k,l)/Z$ is replaced by the outside variable for local structures and r as below.

$$\beta_\sigma(k,l)/Z = \sum_{(i,j)\in P,\ i\le k<l\le j}\beta_\sigma(k,l;i,j)/\exp(r(i,j))$$

In the ParasoR software, it computes $r(x,x)$ for the start position x first, then uses it to calculate r for the outermost pair (i,j) shown as below.

$$\begin{aligned}
r(x,x) &:= \log\left(\frac{Z}{\alpha_{\text{Outer}}(x)\beta_{\text{Outer}}(x)}\right) \\
&\log\left(\sum_{i,j}\left\{\prod_{h=i}^{x-1}\exp(-\Delta\alpha_h)\right\}\ u(i,j)\ \left\{\prod_{h=x}^{j-1}\exp(-\Delta\beta_h)\right\}\right) \\
&= \log\left(\sum_{(i,j)\in S(x)}\exp\left(-\sum_{k=i}^{x-1}\Delta\alpha_k + \log(u(i,j)) - \sum_{k=x}^{j-1}\Delta\beta_k\right)\right) \\
r(i,j) &= \log\left(\frac{Z}{\alpha_{\text{Outer}}(i)\beta_{\text{Outer}}(j)}\right) \\
&= r(x,x) - \sum_{k=i}^{x-1}\Delta\alpha_k - \sum_{k=x}^{j-1}\Delta\beta_k
\end{aligned}$$

I should consider at most W^2 patterns of i and j for r computation and less than $W+1$ additions of energies that depend on subsequences shorter than $2W$. Also, an addition of $\Delta\alpha$ and $\Delta\beta$ should be executed fewer than W times. Therefore, in this way, I can compute an expected value only by variables whose absolute value has an upper bound depending on W.

For the next position $x+1$, the calculation of $r(x+1,x+1)$ from $r(x,x)$ takes $O(1)$ time because the following equation holds.

$$r(x+1,x+1) = r(x,x) + \Delta\alpha_x - \Delta\beta_x$$

Because both a number of these summations and the length of required subsequences are not directly dependent on length N, ParasoR is able to compute the expected values using only "local" variables whose magnitudes are independent of N. I also discuss the numerical precision of computed variables such as $\Delta\alpha$, $\Delta\beta$, and expected variables for actual data in the Result section.

The stability of RNA secondary structure is considerably affected by sequence composition because the secondary structure is comprised of base pairs that differ in strength depending on base type. However, sequence compositions are constrained by multiple biological functions such as coding for proteins, regulating gene expression, and so on. The determination of structural preferences requires a normalization of the influence of sequence composition. GC content is a feature that is commonly used for proper structure evaluation. Previous computational and experimental analyses have shown an apparent correlation between GC content and evaluation index [52]. However, genome sequences are known to have even more complicated sequence biases, for example, AT/GC asymmetry or codon bias. As such, I designed a normalization method for genome-wide comparisons of stem probability using GC content and other, more complex features.

Using python 2.7 and the NumPy library, I implemented a linear regression using the average stem probability $\bar{p}_{\text{stem}}$ with a ridge penalty. A simple implementation is available at my website (https://sites.google.com/site/cawatchm/). I examined the linear regression of stem probabilities with GC content as well as k-mer composition frequencies with $k = 3$ and 4. I used a least-squares method for estimation of a parameter vector w and a regularization term λ, and its error function is formulated as below.

$$\frac{1}{2}\sum_{n=1}^{N}(y_n - w^T x_n)^2 + \frac{\lambda}{2}w^T w$$

In this formula, λ is a constant and w is a parameter vector in the same dimension as x. This equation is differentiable at w, and I obtain w to minimize this error function.

$$w = \left(\sum_n x_n^T x_n + \lambda I\right)^{-1}\left(\sum_n x_n y_n\right)$$

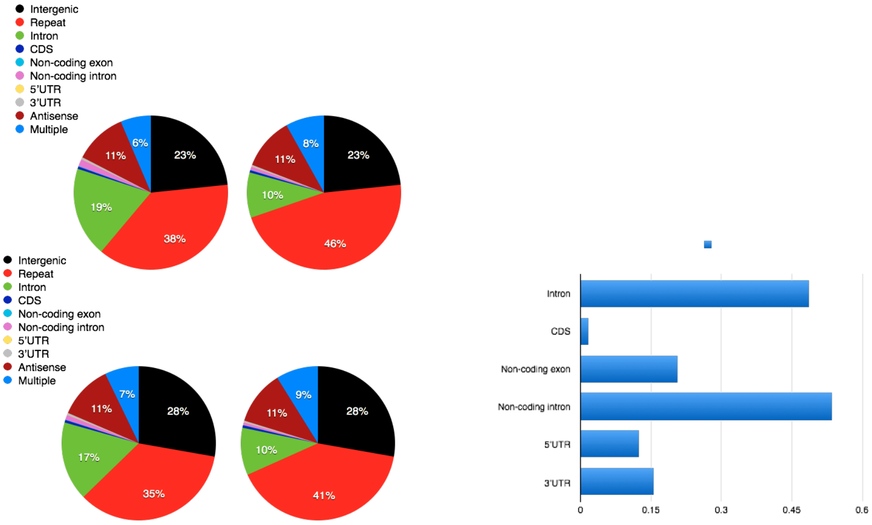
For example, when I selected a 32-mer window size and GC content as a regression feature, I calculate the GC content and average stem probability for each 32-mer fragment. Then, I set x_n and y_n as follows.

$$x_n = \left(\frac{1}{32}, \frac{\text{GC content}}{32}\right)$$
$$y_n = \text{average stem probability}$$

The first element of x_n shifts the mean stem probability to around 0.6. Similarly, x_n for the 3-mer composition regression is formulated as

$$x_n = \left(\frac{1}{32}, \frac{\#\text{AAA}}{32}, \frac{\#\text{AAC}}{32}, \cdots \frac{\#\text{UUU}}{32}\right),$$

where #NNN corresponds to the number of occurrences of NNN in 3-mer sliding windows for the 32-mer fragment. Through this normalization process, 32-mer fragments with more than (window size)/4 ambiguous characters “N” were excluded.



(a)

(b)

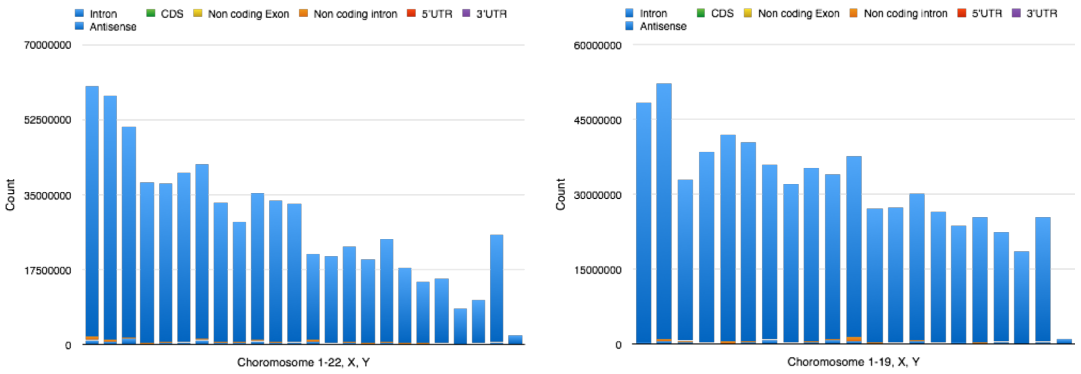
Figure 6. (a) Fractions of 32-mers categorized by the genome annotations. Percentage of 10 annotation groups on human (Top) and mouse (Bottom) genome based on 32-mer fragmentations before (Left) and after removing repetitive elements (Right) from the groups of transcribed regions and Antisense. (b) Ratio of repeat regions included in each annotation group of transcribed regions. Almost half the number of Intronic regions of coding and non-coding RNA is classified into the repetitive elements.

2.5 MANIPULATION OF DATASETS AND DATABASE CONSTRUCTION

Annotation rule

I downloaded assemblies hg19 and GRCm38 of the reference human and mouse genomes, respectively, from the UCSC Genome Bioinformatics Site [53]. I annotated the genomes using the output of RepeatMasker and the RefSeq genes [54], which represent 45, 377 human and 33, 988 mouse genes. They mapped to the same region on the genome sequence as many as 19 times. Where there were overlapping annotations, I prioritized them according to the strength of their biases in base compositions (in the following descending order): *Repeat*, *CDS*, 3′-UTR, 5′-UTR, *Intron*, *Non-coding RNA exon*, *Non-coding RNA intron*, *Antisense*, and *Intergenic* regions. The fractions of each sequence annotation for the human and mouse genome are shown in Figure 6(a). The total sample sizes used for hypothesis testing were as follows: Intergenic 41, 764, 604, Intron 17, 004, 161, CDS 843, 820, 5′-UTR 86, 026, and 3′-UTR 631, 546 for the genomic sequence; Intron 34, 544, 536, CDS 1, 469, 721, 5′-UTR 201, 267, and 3′-UTR 1, 278, 476 for pre-mRNA; CDS 1, 772, 283, 5′-UTR 215, 658, and 3′-UTR 1, 279, 508 for mRNA.

Here, Repeat annotations represent the sense and antisense strands of all the repeat elements reported by RepeatMasker containing retrotransposons, tandem repeats, and so forth. These annotated groups contain repetitive elements in different ratios (Figure 6(b)) and these proportions of repeat sequence probably have a close relationship with the function of each category. I carefully removed these repeat elements because the regression method implemented in this study does not take account of the mutual dependencies of 4-mer sequences in the same 32-mer window. For example, a simple repeat of Adenine (AAAAAAAAAAAA...) cannot bind to the other As in this window, but the linear regression may evaluate the component of AAAA to



(a) (b)

Figure 7. Properties of regions classified into Antisense regions. Bar graphs show nucleotide counts located in antisense strand of the transcribed region for each chromosome of human (a) and mouse (b). Most of antisense regions belong to Antisense group while a small number of nucleotides are annotated as the other annotation groups.

have a weak binding efficiency because AAAA in other contexts can form a stem. Actually, I have shown in Figure 36 that low complexity elements contained in the Repeat category tend to have large $|\Delta\bar{p}_{\mathrm{stem}}|$ (detailed in the Result section), that is, the prediction of stem probabilities by linear regression is systematically poor for such sequences. Since the repeat regions occupy a large part of non-coding regions, such bias will make the hypothesis testing extremely difficult. Therefore, my analyses only studied non-repetitive sequences. I also describe the important detail about the property of repetitive elements in the Result section.

In addition, the antisense strand of the transcribed regions was removed from the intergenic regions and classified as “Antisense” group because of the possible strict constraints derived from the coding genes in the sense strand. Although the antisense sequence of a transcribed gene may be possibly mapped to another transcribed gene, the majority of such regions were classified into the antisense group (Figure 7). The analysis of average stem probabilities for each 32-mer such as Figures 33 and 39 may have suffered from the influence of multiple annotations around the boundary regions. To remove such ambiguous effects, fragments that overlap with multiple annotations were classified into “Multiple” annotation group. Thus, the Intergenic regions contain no Repeat regions, no sense or antisense strands of transcribed regions, and no sequences close to their boundaries.

Shuffling of sequence

To investigate the structural preferences of genome sequences, randomly shuffled sequences were used for comparison. The codon composition ratio is a conserved factor among transcripts. Thus, 3-mer shuffling using the ushuffle module (`http://digital.cs.usu.edu/~mjiang/ushuffle/`) was applied to produce random transcripts for each concatenated mRNA and pre-mRNA sequence. At the same time, I divided whole sequences into each block with ambiguous characters inserted as a boundary between each transcript or found in pre-mRNA as an ambiguous region. For genome analyses, chromosome sequences are too long for the memory usage requirements of ushuffle. Accordingly, I randomly shuffled chromosome 1 for each individual sequence block in the resolution of single nucleotide, again divided by strings of the ambiguous character “N”.

Sequence logo construction

To construct sequence logos in Figures 24 and 25, the RWebLogo package (http://cran.r-project.org/web/packages/RWebLogo/index.html) was applied to partial sequences around motif sites. For sequences around splice sites, however, the original sequence set was too large to apply RWebLogo. Hence, I implemented seqLoGo (https://github.com/carushi/seqLoGo) in Go language, which allows compressing all sequences to a small number of sequences while still expressing the same information with an accuracy rate that depends on the number of sequences retained after compression.

High-throughput structure analyses

To compare ParasoR with high-throughput experimental structural analyses, I used normalized PARS score datasets GSM1226157 and GSM1226158 obtained from renatured samples of GM128678 [20]. PARS scores were calculated for $74,211$ genes from v_i and s_i data, which indicate read coverage normalized by sequence depth of samples after nuclease V1 and S1 degradations, respectively; the maximum scores of v_i and s_i are $216,048$ and $252,848$ respectively, and the average scores are 1.21 for v_i and 1.20 for s_i. To obtain sequence information of transcripts for ParasoR prediction, I extracted $33,603$ genes that are included in PARS dataset and mappable to the RefSeq gene database. Among them, I excluded genes for which there were no nucleotides with s_i or v_i higher than 0. Then, I calculated the PARS score for each position following the equation $\text{PARS} = \log_2(v_i + 5) - \log_2(s_i + 5)$, which is defined in [20]. Moreover, I used raw v_i and s_i scores for further filtering of obscure information. In particular, if a threshold is x, every position with $v_i + s_i$ less than x is excluded from the comparison. Then, I compared the PARS scores with stem probabilities $p_{\text{stem}}(i)$. At the same time, the number of mapped reads was used to filter out inconclusive regions.

Genome-wide structure propensity analyses

To compare the structural preferences of different genomic regions, I first computed stem probabilities for all genomic positions using chromosome sequences as input RNA sequences. As shown in Figure 15, this roughly corresponds to computing the (unaveraged) stem probabilities for sequence windows of $\sim 2,000$ bases in length. I also computed the stem probabilities for RefSeq pre-mRNAs and RefSeq mRNAs with the true boundaries. These probabilities were used for the calculation of the average stem probabilities $\bar{p}_{\text{stem}}(i)$ for non-overlapping 32-nt sequence windows. This length was chosen because the raw stem probabilities exhibit a bimodal distribution with peaks around 0 and 1, while the average stem probabilities exhibit a distribution close to normal when the averaging length is more than 32 bases (Figure 11). The unimodality of $\bar{p}_{\text{stem}}(i)$ is important for the normalization of k-mer frequency bias below, as the linear regression requires unimodal objective variables for its high efficiency. Also, I expect that the distribution of average stem probabilities better represents local structural preferences than does the distribution of single-base stem probabilities. Even though a larger window size could also give a unimodal distribution, too large a window size leads to a highly peaked distribution around 0.5, in which no region-specific structural features will remain. A large window size also causes a reduction in the degrees of freedom for the hypothesis tests and thus reduces the significance of p-values (detailed in the section “Genome-wide simulation by ParasoR”).

Conformational changes after splicing

To investigate structural changes around SSs after splicing, I computed the difference in stem probabilities between mRNA and pre-mRNA as $\Delta q_{\text{stem}}(i) = p_{\text{stem,mRNA}} - p_{\text{stem,pre-mRNA}}$ for each site and each 32-mer sliding window in mRNA. For $\Delta q_{\text{stem}}(i)$ of each site, I then computed the median and median absolute deviation of $\Delta q_{\text{stem}}(i)$ values within a 200-nt window around each SS. I computed the correlations of them with gene expression levels, GC contents around SSs, and intron lengths. For gene expression levels, I used the CAGE promoter FANTOM5 expression data [55, 56]. I used average mRNA expression levels across all tissues and removed tissue-specific mRNAs that satisfy $\log_{10}$(median normalized expression)≤ 0.5. To summarize GC content, I used the GC content of 200 bases around each SS in the mRNA, as well as the averaged GC contents for the 200-nt sequences around the donor and acceptor sites in the pre-mRNA. In the gene set enrichment analysis, I ranked all SSs according to the median of $\Delta q_{\text{stem}}(i)$ for each SS, and the functional enrichment among the top 10 % of the most post-accessible or post-structural genes was analyzed using the DAVID web tool [57].

Computer cluster

For genome-wide computation, I used a super computer system at the Human Genome Center (http://hgc.jp), which consists of Intel Xeon E7 8837, Intel Xeon X5675, and AMD Opteron 6276 CPUs and has a total memory of 2TB.

3 RESULTS

In this chapter, I confirmed the characteristic and validation of ParasoR and applied ParasoR to genome-wide structure simulation to reveal structure constraints on transcriptome and genome sequences. The structural preferences of mRNAs computed by ParasoR shows a high concordance with those determined by high-throughput sequencing analyses. Using ParasoR, I investigated the global structural preferences of transcribed regions in the human genome. A genome-wide folding simulation indicated that transcribed regions are significantly more structural than intergenic regions after removing repeat sequences and k-mer frequency bias. In particular, I observed a highly significant preference for base pairing over entire intronic regions as compared to their antisense sequences, as well as to intergenic regions. A comparison between pre-mRNAs and mRNAs showed that coding regions become more accessible after splicing, indicating constraints for translational efficiency. Such changes are correlated with gene expression levels, as well as GC content, and are enriched among genes associated with cytoskeleton and kinase functions.

3.1 VALIDATION OF PARASOR IMPLEMENTATION

While a method of predicting RNA secondary structure has already been widely developed, to the best of my knowledge, there has not yet been a study that has examined structural properties of genome-scale data with various parameters. In this section, I examined the congruence of the results under a variety of conditions. Moreover, I analyzed properties of the energy model under the maximal span constraint.

Energy models

ParasoR calculates the expected values of secondary structure property based on the energy model in the same way as Rfold [44] and RNAfold [51], which approximates ΔG of a single secondary structure using many energy parameters for the substructures. In this study, I used 4 energy models; Turner (1999), Turner (2004), Andronusce (2007), and Andronescu (2010) energy model. Turner (1999) and Turner (2004) energy model [58] was consructed based on the experimental results of melting temperature for artificual nucleic acids sequences. On the other hand, Andronescu (2007) [59], and Andronescu (2010) [60] energy model were estimated by computational approach using the databes of known secondary structures. A general trend of each type of energy models is that Turner models give higher rewards for base pairing compared to Andronescu energy models. Since there is no previous knowledge about the accuracy of each energy model during the structure prediction of long RNAs, I applied several energy models to extract robust results regardless of the choice of energy models.

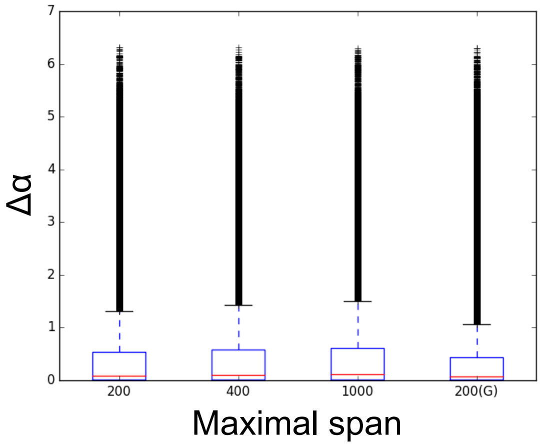


Figure 8. Distribution of the ratio of DP variables. Boxplots show $\Delta\alpha$ calculated for human mRNA sequences with different maximal spans (200, 400, and $1,000$ as shown in x-axis) and human chromosome 1 with $W = 200$ (the rightmost boxplot). $\Delta\alpha$ was randomly sampled once per $1,000$ nt.

Property of energy model and stem probability

The possible patterns of RNA secondary structure generally multiply in accordance with the sequence length N. A MFE, which is one of the measures for evaluating the structure stability of RNA, also decreases with N. Although MFE can be normalized by dividing it by N to compare MFE between RNA sequences of different lengths, a previous investigation has indicated the risk of over normalization by division by N [52]. Because there is insufficient investigation of energy increase such as $\Delta\alpha$, I computed the distributions of $\Delta\alpha$ and $\Delta\beta$ using concatenated transcript sequences with the maximal span $W = 200$, 400, and $1,000$. I randomly sampled one $\Delta\alpha$ per $1,000$ nt and Figure 8 shows that most of the $\Delta\alpha$ distribution was within a range from 0 to 1.6 even though the variance estimates increase with W. Although the length of transcript sequences without ambiguous bases is short, the result of human chromosome 1 sequence was also consistent with that of transcripts sequences.

In addition, because each local structure contributes to the increase of $\Delta\alpha$ and $\Delta\beta$ equally, the increase of $\Delta\beta$ is intrinsically comparable to $\Delta\alpha$ based on existing energy models. In this way I showed that $\Delta\alpha$ and $\Delta\beta$ calculated in ParasoR are small enough to calculate accurate stem probabilities while avoiding numerical errors even for genome sequences. On the other hand, because a partition function Z is calculated by summation of $\Delta\alpha$, Z is expected to be too large value for large N since it can be estimated by $\exp(\overline{\Delta\alpha} \times N)$, where $\overline{\Delta\alpha}$ represents an expected value of $\Delta\alpha$.

The MFE is widely used to evaluate structure stability and extract structural preference of the sequence. However, this measure reflects only the best stabilized structure while ignoring numerous other structures [61]. The main purpose of ParasoR is the analysis of genome-scale RNA sequences, which decreases the quality of MFE estimates owing to an excessive number of possible structures. As such, I used the stem probability $p_{\text{stem}}(i)$ as a measure to evaluate local structural preferences. The stem probability is another typical index that expresses the tendency of structure preferences at a single-nucleotide resolution, and it has the advantage of being able to consider the stability of base pairing among an enormous variety of structures. To determine the general tendency of stem probability $p_{\text{stem}}(i)$, I generated $1,000$ random sequences of $1,000$ nt in length with a 50 % GC content and computed averaged stem probability $\mu_{p_{\text{stem}}}$ of

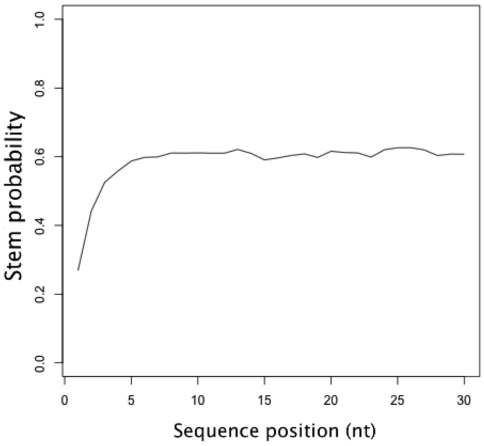


Figure 9. Regional property of average stem probability. $\mu_{p_{\text{stem}}}$ was computed based on Turner energy model (1999) with $W = 1,000$ (which is equal to the sequence length). Several bases at the end of sequence show substantial declines of p_{stem}.

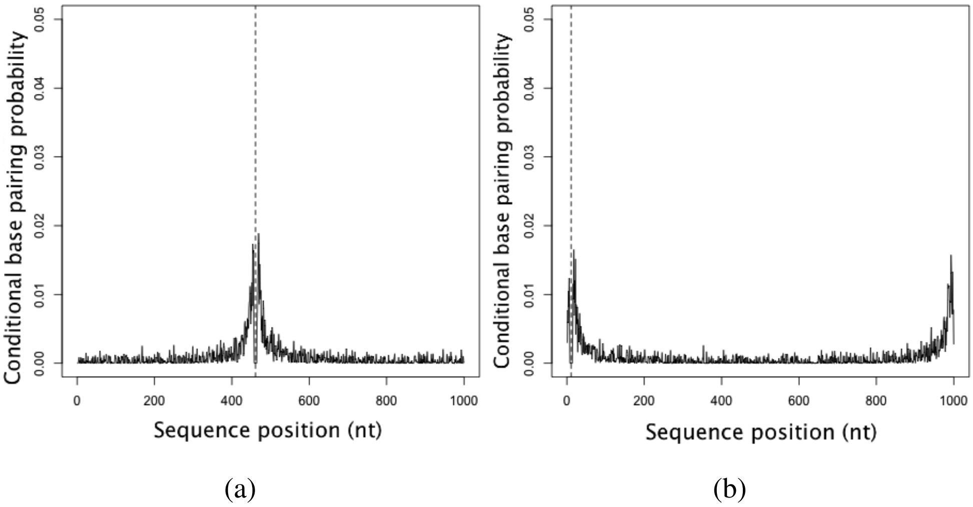
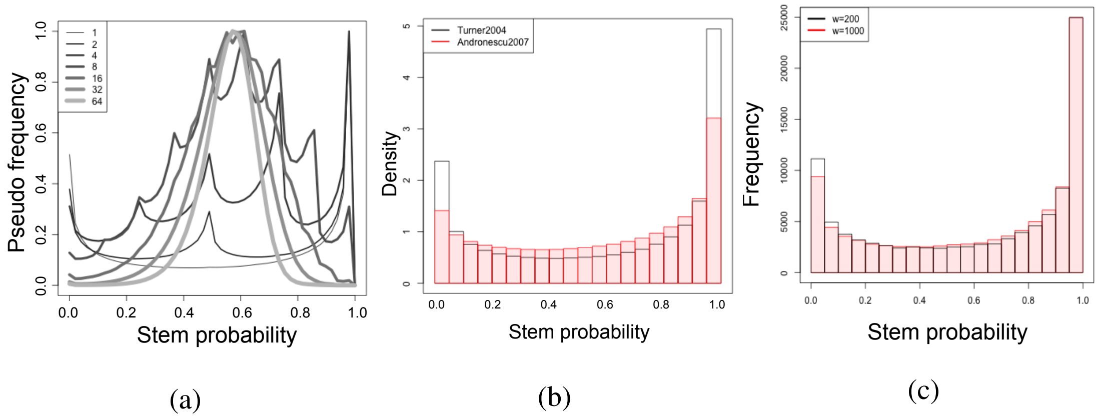


Figure 10. Regional property of Conditional base pairing probability (Cbpp). Cbpp was calculated by ParasoR using Turner energy model (1999) with $W = 1000$ (which is equal to the sequence length). Y-axes show Cbpp between two positions; one is indicated by a dotted line, and another is shown in each x-coordinate.

position i for each sequence without limiting the maximal span, in which p_{stem} is consistent with that of RNAfold (data not shown). In this analysis, $\mu_{p_{\text{stem}}}(i)$ was practically distributed around 0.6 with a notable fall at both ends of sequences, indicating that artificial sequence boundaries potentially cause such bias (Figure 9).

I also calculated a conditional base pairing probability p_{Cbpp} for each base pair to show that such a tendency is produced by the specificity of pairing with other sequence tips. Here, p_{Cbpp} of the jth position among base pairs with the ith base is defined as follows.

$$p^i_{\text{Cbpp}}(j|i) := p_{\text{Bpp}}(i,j)/p_{\text{Bpp}}(i)$$



(a) (b) (c)

Figure 11. Property of stem probability. (a) Average stem probability of each window size (1 – 64 nt) for human genome data. The distribution of average stem probabilities is bimodal for small window sizes, and becomes close to unimodal as the window size increases. (b) Distribution of stem probability computed by Turner (2004) and Andronescu (2007) energy model using the transcript data. (c) Distribution of stem probability for a random sequence of 100, 000-nt length which has 50 % GC content, with the condition of $W = 200$ and 1000.

where $p_{\mathrm{Bpp}}(i, j)$ is the base pairing probability between the ith and jth bases, and $p_{\mathrm{Bpp}}(i)$ is a marginalized base pairing probability for the ith base , which is equal to $p_{stem}(i)$. Hence, p_{Cbpp} is comparable to a likelihood of being partner with ith base. Figure 10 shows an example of p_{Cbpp} profiles for the 461th and 11th bases. While a profile of p_{Cbpp} for a nonterminal base indicates structural preferences for pairing within neighborhoods, such a profile for ends clearly shows that they are likely to form base pairs with their opposite ends.

This result raises an issue about sliding-window methods because it can reduce the calculation time but generate many biases at the ends of each window although the global folding algorithm such as ParasoR may cause them only at the ends of sequences.

Next, I simulated a distribution of stem probabilities p_{stem} and average stem probability $\bar{p}_{\mathrm{stem}}$ for genome sequence data. As shown in Figure 11(a), the distribution of p_{stem} is bimodal at 0 and 1 and becomes nearly unimodal by window averaging after the structure prediction. In particular, the distribution of $\bar{p}_{\mathrm{stem}}$ for each 32-mer or longer unit is almost unimodal. I also investigated the influence of different energy parameters on the distribution of p_{stem}. In Figure 11(b), although both of models show a bimodal distribution with the peak around 0 and 1, the strength of bimodality is different for each energy model. In particular, the Turner (2004) model is shown to have produced more strongly bimodal results than did the Andronescu (2007) model. This is likely because their energy parameters were estimated from different knowledge of secondary structures and thus reflects properties of the datasets. The influence of different maximal span sizes was also tested for various window sizes using structure-prediction data from human chromosome 1. To evaluate an agreement between different conditions, a correlation coefficient was computed for several window sizes ranging from 10 to 400 nt. A comparison between $\bar{p}_{\mathrm{stem}}$ from Turner (2004) and Andronescu (2007) was higher than 0.8, at least for any selected window size (Figure 12). Even accessibility, another typical index for structure evaluation, showed a high correlation with $\bar{p}_{\mathrm{stem}}$. I also compared $\bar{p}_{\mathrm{stem}}$ with $W = 200$ and that of three other maximal spans: 100, 150, and 400. All of them show the correlation coefficient higher than 0.8 for the window size larger than 32-mer (Figure 12). For longer $W (= 1, 000)$, I have also shown a histogram of stem probabilities for $W = 200$ and $W = 1, 000$ in Figure 11(c).

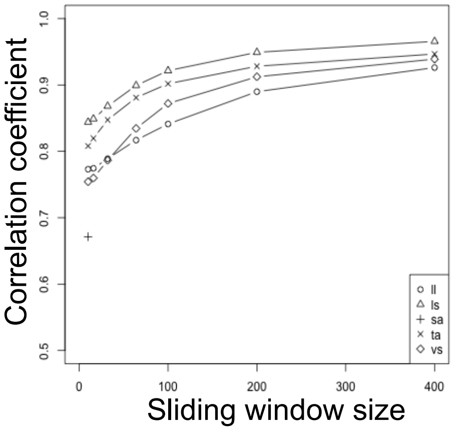


Figure 12. Correlation with sliding window with varying the parameter for structure prediction. Correlation coefficient of averaged p_{stem} (W = 200) was computed with accessibility (sa+), averaged p_{stem} by the different energy model (ta×: Turner (2004) versus Andronescu (2007) energy model), or averaged p_{stem} of the different length of maximal span (vs◇: W = 100, ls△: W = 150, and ll○: W = 400) computed by Turner energy model (2004). X-axes correspond to the window size for averaging of p_{stem}. Although a longer window size increases a value of correlation coefficient, it may reduce a characteristic for each region.

The correlation coefficient of stem probabilities between $W = 200$ and $W = 1,000$ is 0.713. The stem probability gradually increases with W as the number of possible base pairs increases.

For subsequent comparisons among different regions, a proper normalization according to sequence compositions is required. As such, I applied a linear regression method in which a unimodal distribution of $\bar{p}_{\text{stem}}$ is suitable. Although a larger window size produces stronger unimodality, structural preferences specific to each region were also averaged to the mean value. Based on these results, I selected the 32-mer window size in order to produce a concordant unimodal stem probability distribution without excess standardization.

Testing the accuracy of ParasoR by comparison with RNAplfold

To evaluate the robustness of ParasoR against overflow problems, I compared stem probabilities calculated with ParasoR and RNAplfold, which calculates the same values theoretically when the maximum window size and maximal span of base pairing are set to N and W, respectively. For comparison, I generated random sequences of $1,000$, $3,000$, and $5,000$ nt in length, each with a 50 % GC content. Figure 13 shows boxplots of stem probability differences between ParasoR and RNAplfold (ViennaRNA package v2.0.7) for each sequence with varying the maximal span W. There is little difference between these methods among $1,000$-nt sequences. However, critical differences appeared for sequences with lengths of $3,000$ and $5,000$ nt. This discrepancy is due to an overflow of RNAplfold that results in a computed probability exceeding 1, while ParasoR can still produce an appropriate distribution in the range from 0 to 1 for much longer sequences such as genome sequences (Figure 14).

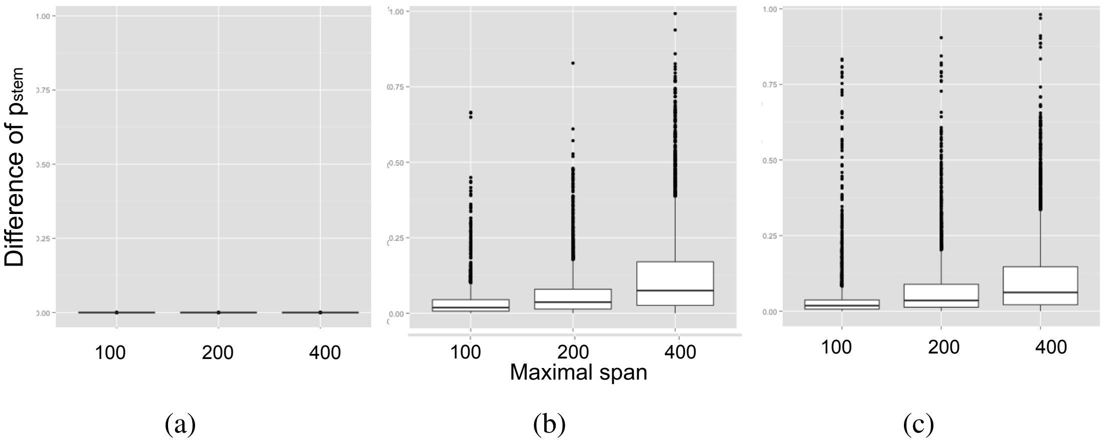


Figure 13. Boxplots for the differences of stem probability between RNAplfold and ParasoR. I used the random sequence of different length ((a) $N = 1,000$, (b) 3,000, and (c) 5,000). Each boxplot corresponds to the result from the different maximal span of base pairings (100, 200, and 400) in ParasoR and RNAplfold (the window size is set to the sequence length).

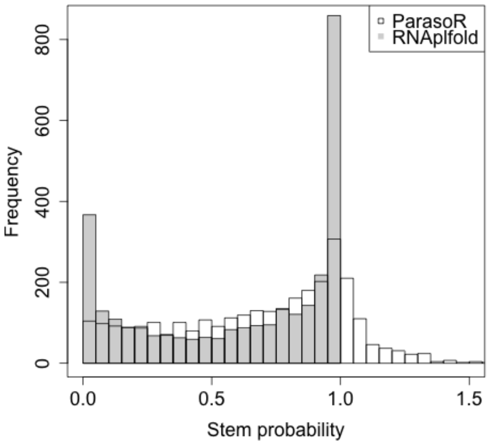
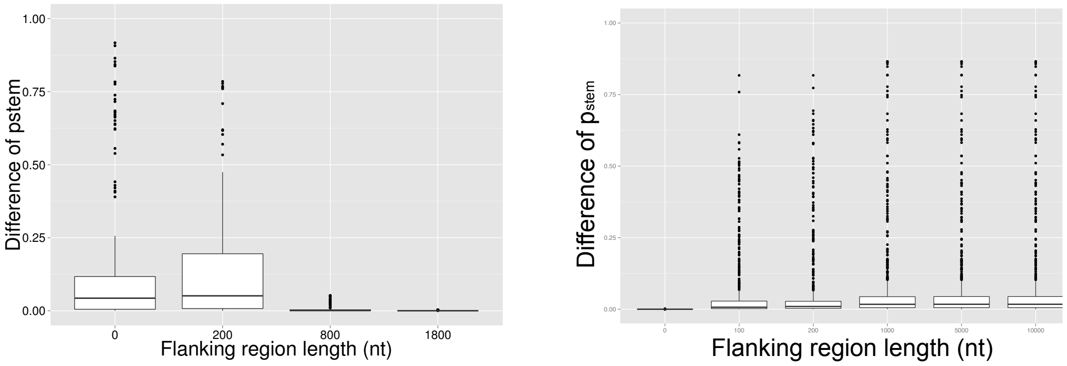


Figure 14. Overflow of stem probability. Histogram of stem probabilities of ParasoR and RNAplfold for a sequence of 3,000-nt length with the maximal span of base pairings 200 (the window size is set to 3,000).

Influence of flanking region

I examined a way to take account of the structural influence of flanking sequences around target sites. To visualize the remodeling of the secondary structure by flanking sequences, the *PTGFR* gene (NM_001039585) was chosen as a subject sequence because it is sufficiently long (> 49,000 nt) and located at chr1:78,956,727-79,006,386, which is not the edge of a sequence block. Using RNAplfold, $p_{\text{stem}}(i)$ was calculated for a 200-nt center region of subsequences that contains a center region and varied lengths of flanking regions at both sides within the range at which no severe overflow problems occur in 13. As shown in Figure 15(a), differences of two stem probabilities converge to 0 as longer flanking regions are added to the RNAplfold computation. This demonstrates that the computed stem probabilities converge when flanking sequences at both ends exceed $2W$ bases, which is consistent with previous findings in Ref.[44, 5]. Thus, the stem probabilities computed for the entire chromosome roughly



(a) (b)

Figure 15. (a) Difference of stem probability between ParasoR and RNAplfold. Boxplots of the differences of p_{stem} between ParasoR and RNAplfold for the specific region of 200-nt length on human chromosome 1 with a different length of flanking region. (b) Difference of stem probability with flanking regions. Boxplots of the differences of p_{stem} between RNAplfold and ParasoR with the flanking region of variable length at both sides. p_{stem} was computed for the sequence ($N = 1,000$) by RNAplfold with $L = 200$ and $W = 1,000$. I then computed stem probability of the same sequence by ParasoR with the flanking region of $0,\ 100,\ 200,\ 1,000,\ 5,000,$ and $10,000$-nt length (represented in the x-axis).

correspond to those for sliding windows with large ($\sim 4W$) margins. ParasoR simulated RNA secondary structure for the whole chromosome sequence while RNAplfold computed stem probability with a part of flanking sequence. The x-axis represents the length of flanking sequences around the region applied for the computation of RNAplfold.

I also calculated the influence of flanking region from an opposite viewpoint by appending additional sequence in order to examine the change in stem probability. For comparison, $p_{\text{stem}}(i)$ was calculated for a sequence with $N = 1,000$ by RNAplfold. Comparing these stem probabilities to those calculated by ParasoR for the same sequence revealed that they have little difference from each other (Figure 15(b)). Then, I appended random flanking regions of $N = 100,\ 200,\ 1,000,\ 5,000,$ and $10,000$ nt to both sides of the sequence, resulting in an increased disparity that eventually converged to a fixed difference of stem probability. This also suggests that flanking regions have enough influence to change structural preferences only around the sites in the range of $\sim$ 1000 bases.

Varying the precisions of floating point numbers

I investigated how the numerical precision of floating point variables affects the ParasoR results. I computed the probabilities with varying precisions of real numbers using float, double, and long double types as variable declaration in the C++ program. double (long double) provides a precision at least as much precision as float (double) [62]. In the x86-64 architecture used to test ParaoR performances, long double is handled as 80-bit x87 floating point type [63]. Figure 16 shows the differences in $p_{\text{stem}}(i)$ computed for 1,000-nt random sequences with increasing the length of flanking sequences using different floating point variables. While the differences in $p_{\text{stem}}(i)$ between float and long double variables increase rapidly, those between double and `long double` variables remain small even for a long sequence of length 10K bases, indicating using double or long double enables to safely analyze long RNAs without degradation of numerical precision. Hence, I concluded that using 64-bit double is sufficient to avoid numerical problems.

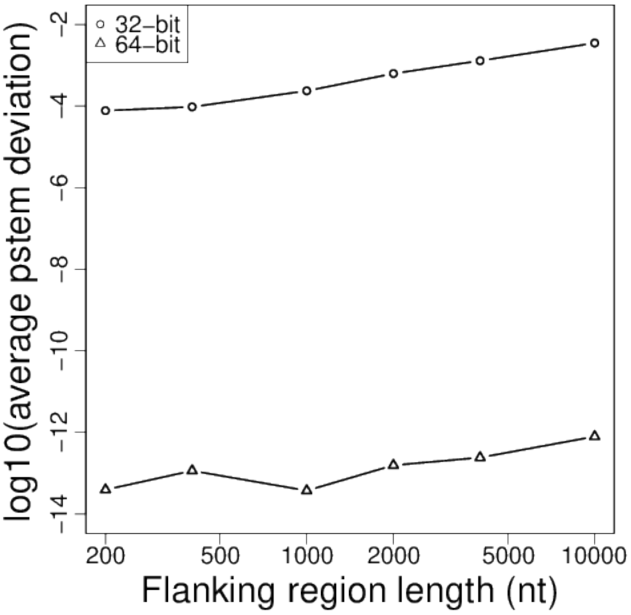


Figure 16. Influence of the different numerical precision. Average absolute deviation of p_{stem} on the different conditions of the numerical precision of floating point variables. p_{stem} was computed for the 1,000 bases located at the center of random sequences with the flanking region of the dfifferent length (shown in x-axis) using `float`, `double`, and `long double` floating variables. The y-axis represents the deviation of stem probabilities that were computed using `float` or `double` floating point variables from those with `long double` variables.

3.2 CONCORDANCE OF PARASOR PREDICTION WITH VALIDATED RFAM STRUCTURES AND A HIGH-THROUGHPUT STRUCTURE ANALYSIS

Since ParasoR was developed for the structure prediction of long RNA sequences, I tested its accuracy with the genome and mRNA sequences, using validated structures from the Rfam database [64] and a high-throughput structure analysis [20]. In this section, I performed a validation of ParasoR and sliding-window prediction for cis-regulatory elements structures listed in Rfam database. Moreover, I present a comparison of ParasoR prediction and other methods with PARS data for human transcripts.

Concordance between ParasoR predictions and Rfam structures

First, to evaluate the performance of structure prediction, I used CisReg data, which was compiled in Ref. [46] and contains high-quality sub-structures within long sequences. To construct the dataset, they searched the Rfam database for structures annotated as *cis*-regulatory elements, and obtained 2, 500 structures, as well as the flanking mRNA or genomic sequences of lengths up to 3, 000 nt on both sides. Then, I predicted secondary structures for these whole RNAs and compared them with known structures only within the region of target *cis*-regulatory elements. As for ParasoR, I used the γ-centroid structure with $\gamma = 1$ [47]. Since RNALfold

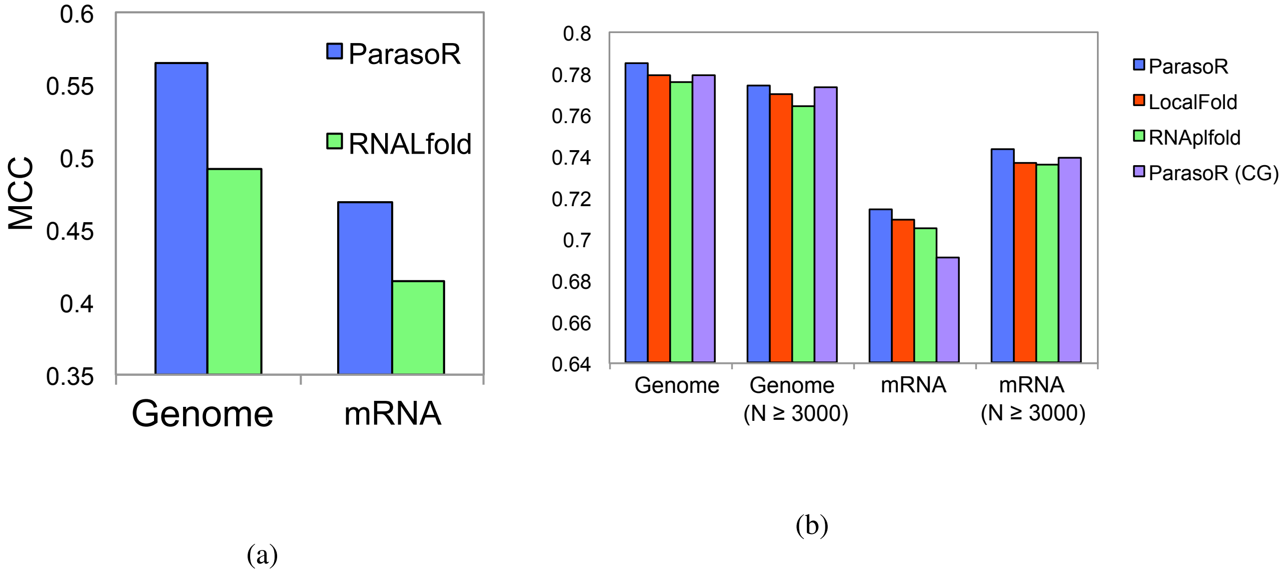


Figure 17. Accuracy comparison of single structure prediction. (a) MCC scores describing the structure predictions of *cis*-regulatory elements in the CisReg genome and mRNA dataset for the performance evaluation of ParasoR and RNALfold. (b) AUC scores of ROC curves between CisReg data and p_{stem}-based classification. Each position was classified into "accessible" or "structured" groups using three tools with Turner (2004) energy model, and ParasoR with Andronescu (2007) (CG) energy model. mRNA dataset is the set of cis-regulatory elements found on the known mRNA sequences, and genome dataset contains the others. The suffix "long" means that the dataset does not contain the sequences whose total length is less than 3,000 nt.

predicts multiple overlapping structures, I extracted the longest structures in ascending order of free energies without any overlap, in accordance with the post-processing described in Ref. [44]. In addition, I used the Matthews correlation coefficient (MCC) scores to evaluate the accuracy of ParasoR and RNALfold for prediction with one condition since RNALfold has no appropriate parameters that control the balance between sensitivity and specificity. MCC score is defined as below

$$MCC = \frac{TP \times TN - FN \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where TP, TN, FP, and FN correspond to the numbers of true positive, true negative, false positive, and false negative predictions, respectively. Figure 17(a) shows MCC scores of ParasoR for mRNA and genome datasets, which are substantially higher than those of RNALfold. It indicates the efficiency of γ-centroid structure prediction for long RNA sequences, as well as short RNA sequences [47], as they predict fewer false positives than the MFE-based method.

I also tested the accuracy of binary classification, which predicts whether each base is structural (base-paired) or accessible (unpaired), based on the stem probability for each position. This kind of problem is more meaningful when the input RNA does not take a single stable structure. For two average-sliding-window methods, RNAplfold (ViennaRNA package v2.0.7) and LocalFold (v1.0), I set the parameter as follows: a maximal span of pairing L to 150, average window size W to 200, and skip size b to 10 (only for LocalFold) according to the previous study [46], in which optimal parameter sets were investigated. For the stem probability p_{stem} computed by ParasoR, LocalFold [46], and RNAplfold [43], I progressively changed a critical p_{stem} threshold and classified each position as structured or accessible, depending on whether

p_{stem} was higher than the threshold or not. In addition, I prepared partial dataset of genome and mRNA sequences whose length are longer than $3,000$ nt (represented as genome_long and mRNA_long). For validation of stem probabilities, I drew the receiver operating characteristic curve (ROC), which plots false positive rates ($\frac{FP}{FP+TN}$) for x-axis and true positive rates ($\frac{TP}{TP+FN}$) for y-axis with varying threshold. Then, I evaluated these classifications with the Rfam reference structure by the area under ROC (AUC) using both of Turner (2004) and Andronescu (2007) energy model. As a result, ParasoR with Turner (2004) model showed the highest accuracy compared to other tools or ParasoR with Andronescu (2010) model regardless of the minimum length of sequences in dataset (Figure 17(b)). In summary, ParasoR is comparable to or better than the state-of-the-art algorithms for the prediction of stable motif structures such as *cis*-regulatory elements in long RNAs. .

Concordance between ParasoR predictions and PARS data

Recently, high-throughput structure analyses such as PARS have gained attention because of their wide application range and high productivity. However, there has been few comprehensive comparison between computational predictions and such analyses because of an inability to produce comparable amount of results via computational methods. Hence, I examined the concordance between PARS scores and two computational prediction measures of ParasoR; *stem probability* p_{stem} of each single nucleotide and *accessibility* of each 10-base unit. Both of these measures exhibited a significant correlation with the PARS score and in particular the stem probability exhibited a higher correlation coefficient than did accessibility. The Pearson's product-moment correlation coefficients were 0.212 ($p < 2.2e - 16$) for stem probability. This seems to be because accessibility is a measure of being accessible for a series of nucleotides at a certain time, despite positional independency of stem probabilities and PARS scores. I also investigated the correlation of the average stem probability for each 32-mer $\bar{p}_{\text{stem}}$ with PARS score, and their correlation coefficient is 0.208 and still high ($p < 2.2e - 16$). Accordingly, I used p_{stem} and $\bar{p}_{\text{stem}}$ as measures in the subsequent analyses.

I tested a prediction regarding long maximal spans to determine the influence of a fixed W. Figure 65 shows the distribution of correlation coefficients calculated between PARS scores and $\bar{p}_{\text{stem}}$ for each 20-mer, which is a measure of consistency used in [20]. A distribution with $W = 200$ apparently contains much more regions of positive correlation between the PARS scores and $\bar{p}_{\text{stem}}$. When I increased W to $1,000$, there was little change to the distribution, though it also introduced a positive bias. According to this, differences between ParasoR prediction and PARS scores are expected to not be caused by ignoring distant base pairs but other factors instead, such as complicated secondary structures, higher dimensional structures, and experimental biases.

As previously described, the distribution of PARS scores is highly concentrated around 0 and has a long tail at both sides. Because sequence reads are likely to be assigned sparsely, almost uncovered and totally obscure positions might contributed to structural tendencies. Accordingly, I examined the influence of such obscure regions by filtering the sequence data with a variable threshold t for the minimum read coverage. Figure 64 shows scatter plots of PARS scores and p_{stem} when $t = 0$ and $t = 40$. Apparently, application of a strict t threshold removed data with PARS scores around 0 and elucidated a concentration of samples with positive PARS scores and $p_{\text{stem}} \approx 0$ or negative PARS scores and $p_{\text{stem}} \approx 1$. Actually, the correlation coefficient between

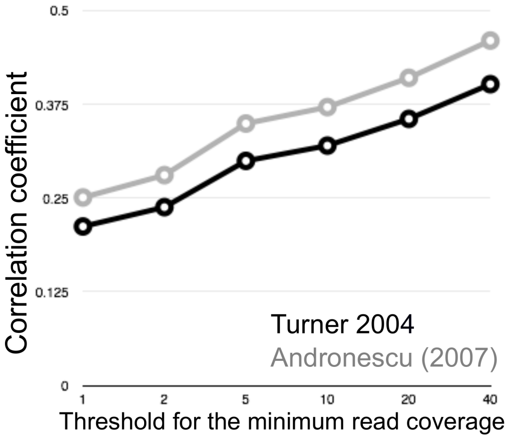


Figure 18. Correlation between ParasoR and PARS scores with filtering. Correlation coefficients between p_{stem} and PARS score filtered by setting the minimum read coverage shown in x-axis. The black and grey line represent correlation coefficients computed for stem probabilities based on Turner (2004) and Andronescu (2007), respectively.

p_{stem} and PARS score at each position increases with threshold even though the two variables differ in sample size (Figure 18). Therefore, it is suggested that positions with rich information of sequence reads is likely to be consistent with ParasoR prediction.

Consistency of thermodynamic prediction methods and PARS score based on stem probability

Then, I investigated the congruence between computational predictions and PARS data in the same way as CisReg. To compare p_{stem} and PARS data from human mRNAs, I divided all nucleotide positions into two groups, accessible and structured, as determined by PARS scores. AUCs of three tools were then computed with a progressively changing p_{stem} threshold for their classification. To calculate p_{stem}, three energy models were applied to this analysis, Turner (2004), Andronescu (2007), and Andronescu (2010) energy model. Among them, Andronescu (2010) model was applied to ParasoR only because RNAplfold, and eventually LocalFold cannot read a parameter file of Andronescu (2010) model. I obtained thresholds of evenly spaced PARS scores between the maximum and minimum values to define true structured and accessible regions. Then for p_{stem}, I set thresholds from 0.0 to 1.0 to categorize each nucleotide as structured or accessible. The accuracy of the predictions was evaluated by setting true structured and accessible regions according to the PARS threshold with prediction of being structured or accessible according to the threshold of p_{stem}. Then I calculated AUCs of ParasoR, and two stochastic sliding-window methods, LocalFold and RNAplfold with PARS scores with

Table 1. AUC values between PARS data and prediction tools for various conditions

Energy model	maximal span	ParasoR	LocalFold	RNAplfold
Turner (2004)	200	0.5977	0.6059	0.6067
Turner (2004)	1,000	0.5969		
Andro (2007)	200	0.6101	0.6178	0.6185
Andro (2010)	200	0.6117		

varying thresholds to evaluate the accuracy of ParasoR in classifying regions as structured or accessible.

Firstly, I fixed a threshold for PARS score to a middle value (-0.39) based on the maximum and minimum PARS score. Figure 19(a) and Table 1 show AUCs which were obtained by varying a threshold for classification of stem probabilities. Consequently, although all of the prediction methods showed a high consensus with the PARS-based classification, ParasoR had an almost comparable AUC score to LocalFold and RNAplfold (0.610 versus 0.618 and 0.619, respectively Figure 19(a,Left)). In addition, when I compared the 32-nt average of p_{stem} with the 32-nt average of PARS scores, ParasoR showed a slightly higher AUC than the other tools (0.581 versus 0.578 and 0.578, respectively Figure 19(a,Right)). These results are important, as I extensively study the distribution of such averaged p_{stem} in the later sections. In addition, since setting a maximal span W to $1,000$ decreased AUC compared to AUC with $W = 200$, 200 is considered to be a proper maximal span for structure prediction of transcripts with reasonable computation time. For p_{stem} computed by three types of energy models, AUC values of ParasoR are slightly less than those of LocalFold and RNAplfold and the result is consistent for three energy models although Andronescu (2010) model, which was optimized using the Boltzmann likelihood algorithm, has a highest accuracy for prediction, followed by Andoronescu (2007), and Turner (2004) model.

Since ParasoR has been developed to analyze genome-wide structure propensity, a consensus between ParasoR and PARS data in terms of structure propensity analyses was also tested for three categories of transcript region, 5′-UTR, 3′-UTR, and CDS. In Ref. [20], an average PARS score was used to estimate the likelihood of being structured for each region, and it was concluded that 5′-UTRs are less structured than 3′-UTRs and CDS. However, I found that the average PARS scores are affected by a small number of outliers with extremely large PARS scores. Figure 20(a-b) shows the distribution of PARS scores for each annotation class. In both of distributions, with or without filtering, 5′-UTR is found to have more structured (> 0) and less accessible (< 0) regions compared to other groups. In [20], however, 5′-UTR was estimated to be more accessible using an average of PARS scores even though this measure is considered strongly influenced by large absolute values. In contrast, ParasoR evaluates each position equally within a range from 0 to 1. As an absolute read coverage at each position is also determined by the number of mapped reads, it is supposedly an inappropriate measure to compare with p_{stem}. Accordingly, I calculated a median PARS score of each annotation class for comparison. In Figure 20(c), the median PARS scores for CDS, 5′-UTR, and 3′-UTR regions were 0.6041, 0.6280, and 0.5083, respectively. These scores agree with the order of average p_{stem} determined by ParasoR, which are 0.5956, 0.6248, and 0.58978, respectively, and furthermore, this order is robust to different filtering thresholds. Both scores consistently indicate that 5′-UTRs have the highest stem density, whereas the stem density of CDS regions is the lowest. This high stem density of 5′-UTRs is mostly explained by their high GC content, as I show in the subsection “Genome-wide simulation to detect structural constraints on transcribed regions”.

Because PARS scores with low-read depths are supposed to be less reliable, I set a threshold for the minimum read depth to filter out less reliable sites for a comparison of stem probability and PARS score. For such a sample dataset, PARS score distributions were obtained for two groups, accessible ($p_{\text{stem}}(i) < 0.5$) and structured ($p_{\text{stem}}(i) > 0.5$) regions according to ParasoR-based p_{stem}. Figure 19(b) shows that the PARS scores are more consistent in structured regions as their median values increase with the strictness of the threshold, while PARS scores

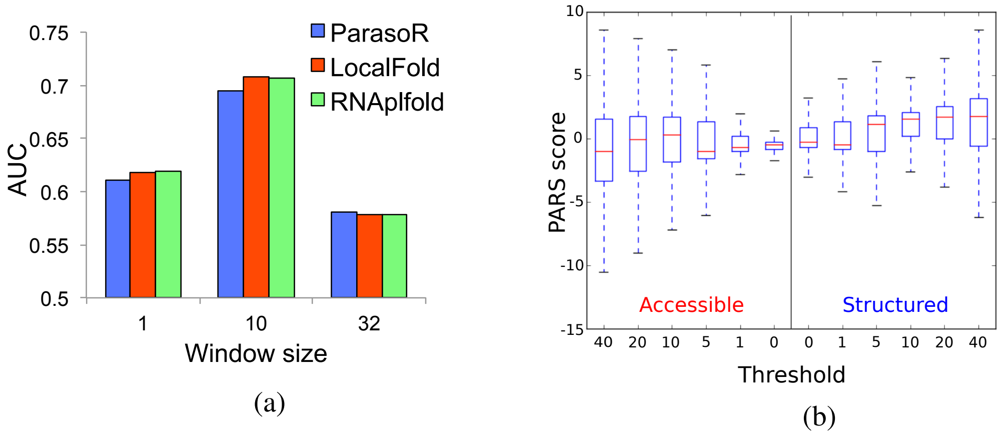


Figure 19. Comparison of stem probabilities and PARS scores. (a) AUC scores describing the prediction of positions with high PARS scores (i.e., structured regions) by stem probabilities for ParasoR and other tools. (b) Distribution of PARS scores for accessible and structured regions with varying read-depth thresholds. Each position was classified into Accessible or Structured depending on the stem probability of ParasoR ($p_{\text{stem}} < 0.5$ or $p_{\text{stem}} > 0.5$) after filtering of the minimum read-depth. Outliers are excluded from each Tukey boxplot.

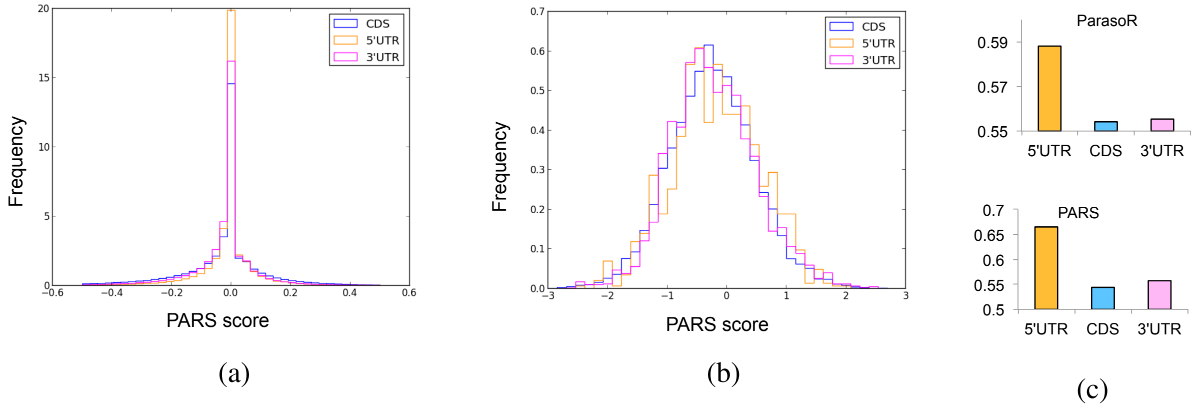


Figure 20. (a-b) Histograms of PARS scores for each annotation (CDS, 5′-UTR, and 3′-UTR). Each figure shows the histograms before (a) and after filtering (b) with the threshold one for the minimum read coverage. (c) Comparison of the average stem probabilities of ParasoR and the median of filtered PARS scores among 5′-UTR, CDS, and 3′-UTR categories.

fluctuate around 0 in accessible regions, although they are consistent for a very strict threshold that requires ≥ 40 read counts to designate a site. Next, I confirmed an accuracy of each tool with different conditions such as thresholds for PARS score and filtering by read depth. Figure 21 shows AUC values about classification of three tools for whole or filtered dataset, and for a different threshold for PARS score. As a result, both filtering of low-read depth region with a liberal threshold for PARS score and no filtering with strict thresholds increased AUC for prediction. Strict thresholds for PARS scores increased AUC of each tool, indicating high accuracy. Although a liberal threshold decreases the AUCs, filtering out rarely mapped regions increases the AUC as well as the accuracy of strict thresholds. Such tendency was common for Turner (2004) and Andronescu (2007) model. Comparing an impact of energy model and software, selecting a different energy model produced a larger change of AUC than selecting a different method.

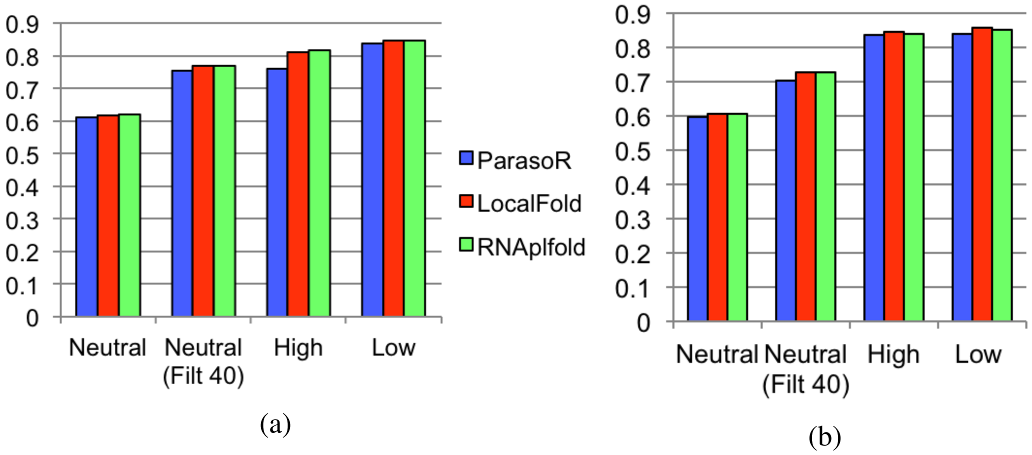


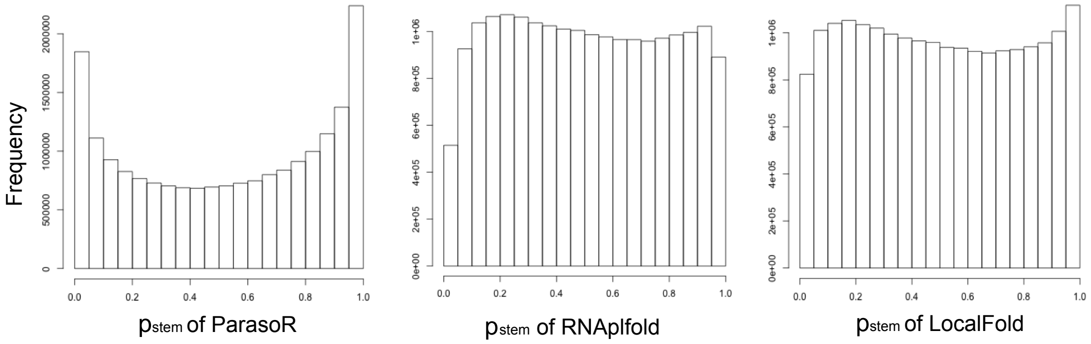
Figure 21. AUC scores for ROC curves between PARS and three tools. Each AUC was calculated for "true" PARS score and "prediction" of three tools (p_{stem}) using (a) Turner (2004) and (b) Andronescu (2007) energy model for all positions, or limited positions where the read coverage of PARS is no less than 40 (used in Filt40). Neutral (−0.39), High (8.28), and Low (−9.06) corresponds to the threshold for PARS score used for classification of "accessible" or "structured" groups with progressively changing a threshold for the p_{stem}-based classification to calculate each AUC value.

These analyses indicate that computational predictions and PARS analyses are highly congruent with each other in their ability to classify sequences as accessible or structured and in their measurement of regions based on functional annotation.

For comparison between p_{stem} and high-throughput structure analyses dataset, p_{stem} with Andronescu (2010) model showed the highest accuracy compared to other tools or ParasoR with Turner (2004) and it is an opposite result to that of Rfam motif prediction. I inferred it is due to the difference of stem probability distributions among different tools. A bimodal distribution of ParasoR is similar to that of RNAfold (Figure 22(a)) and appropriate to predict rigid structures such as those in the CisReg dataset. On the other hand, the distributions of LocalFold and RNAplfold are supposedly more suitable for PARS data, which has a bell-shaped smooth distribution. In fact, I can increase concordance of ParasoR with the PARS data by using the Andronescu model and averaging stem probabilities to reduce the bimodality of the distribution compared to other tools (see the right bar graph of Figure 19(a)). Figure 11(b) shows that Andronescu (2010) model instead of Turner (2004) model decreases the bimodality of stem probability distribution. Figure 22(c) shows that averaging of stem probability also makes a distribution close to unimodal. These indicate the lower accuracy of ParasoR for PARS data as compared to RNAplfold and LocalFold mainly results from the difference of their distributions.

Advantages of ParasoR compared to sliding-window methods

I have shown that the stem probabilities computed by the probabilistic folding methods such as ParasoR, RNAplfold, and LocalFold are highly consistent with the known structures of *cis*-regulatory elements and the PARS data. They also attain AUCs around 0.7-0.8 when ambiguous sites are removed from the PARS data (Figure 21). The differences of AUC scores among these programs are of the order of 0.01 and thus very small for these datasets.



(a) (b) (c)

Figure 22. Each figure shows the histogram of p_{stem} calculated by ParasoR (a), RNAplfold (b), and LocalFold (c) for human matured transcripts.

Although ParasoR has similar accuracy to the other two tools, it is distinctly different from them, because it is a global folding method and the expected values such as stem probability and accessibility are computed from the Boltzmann ensemble of globally consistent secondary structures, as in McCaskill's algorithm and RNAfold[51]. The difference between ParasoR and the latter two global folding algorithms is its scalability to handle long RNAs; ParasoR can compute the structural properties of the longest pre-mRNA in the human genome without any problem, whereas such computation is impossible for the other global algorithms, owing to the numerical errors and high time and space complexities.

In contrast to these global folding algorithms, RNAplfold and LocalFold average the probabilities that are computed from mutually inconsistent local RNA structures on different sliding windows. Although these sliding-window algorithms may capture the effects of structural obstacles such as bound proteins, introducing artificial boundaries at every sequence position may also cause artificial effects on the results. For example, it is known that accessibilities are artificially high close to the window boundaries [46], and both the window ends tend to be paired with each other

Another difference between the sliding-window methods and the global folding methods is that the probability distributions they produce are markedly different. As shown in Figure 22, the distribution of ParasoR has bimodal peaks around probability 0 and 1, while the distributions of the other tools are more even. Furthermore, ParasoR has additional useful options that are not available in the other tools. For example, it can compute globally consistent γ-centroid structures, which are more accurate than MFE structures (Figure 17(a)). The structural profile for each sequence position was also shown to be a very powerful means of understanding the complex structural specificities of RNA-binding proteins [5]. I therefore conclude that ParasoR is currently the most suitable program to analyze the structural properties of a transcriptomic-scale dataset with high confidence.

3.3 PREDICTION OF STRUCTURE PROFILES FOR HUMAN TRANSCRIPT

I performed positional structure propensity analyses for human mRNAs and pre-mRNAs using ParasoR. To extract the common properties of human transcripts, I computed $\mu_p(i)$, the positional profile of probabilities averaged across all human mRNAs or pre-mRNAs. Figure 23 shows $\mu_{p_{\text{stem}}}(i)$, the positional profile of stem probabilities around start codons, the 1st-3rd exon

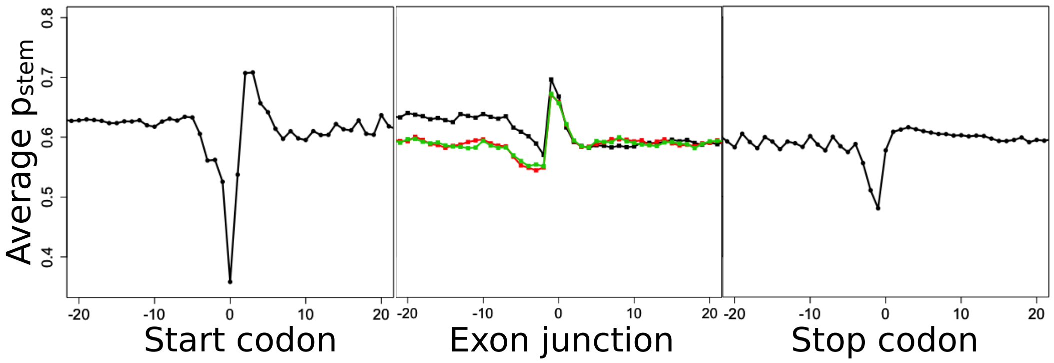


Figure 23. $\mu_{p_{\text{stem}}}(i)$ around start codons (Left), exon junctions (Center), and stop codons (Right). Profiles of the first, second, and third exon junctions are drawn in black, green, and red, respectively.

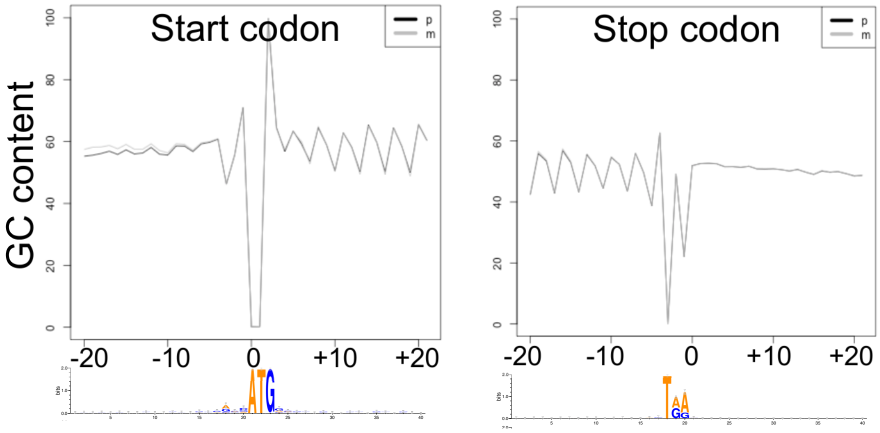


Figure 24. Profiles of GC contents around start codon (Left) and stop codon (Right). Sequence logos were constructed for all of pre-mRNA sequences.

junctions, and the termination codon, which are computed from mRNA sequences. I consistently observed many characteristics reported in previous experimental analyses, such as the sudden fall of stem density before start and termination codons, an increase within start codons, and 3-mer periodicity in coding regions [20]. Figure 23 shows the average stem probability for each position within 40 nt around SSs grouped by the first, second, and third SSs from the front and back (where duplication was allowed). The first SSs clearly tend to be structured, while weaker disparities in stem probabilities are observed around the last SSs. This preference corresponds to high GC contents associated with CpG islands in or around promoter regions (Figure 25). In the profile of average stem probability differences, however, there is little disparity among SSs (Appendix Figure 68).

Next, I analyzed the positional specificity for structural profiles $\mu_{p_\delta}(i)$ ($\delta =$ bulge, exterior, hairpin, or interior) computed from pre-mRNA sequences. Because the magnitude of $\mu_{p_\delta}(i)$ strongly depends on the loop type, I averaged μ_{p_δ} across the 300 nt surrounding each SS on both sides to compute $\overline{\mu}_{p_\delta}$ and normalize the differences among loop types. In Figure 26, $\log(\mu_{p_\delta}(i)/\overline{\mu}_{p_\delta})$ is plotted to show the specific increase of loop probabilities around the donor and acceptor sites. Around donor sites, p_{bulge} and p_{internal} increase at position 1-3 nt, consistently with the two $\mu_{p_{\text{stem}}}$ peaks located at both sides of the donor sites (Figure 25). Previously, several studies have reported the presence of conserved stable stem structures around donor sites

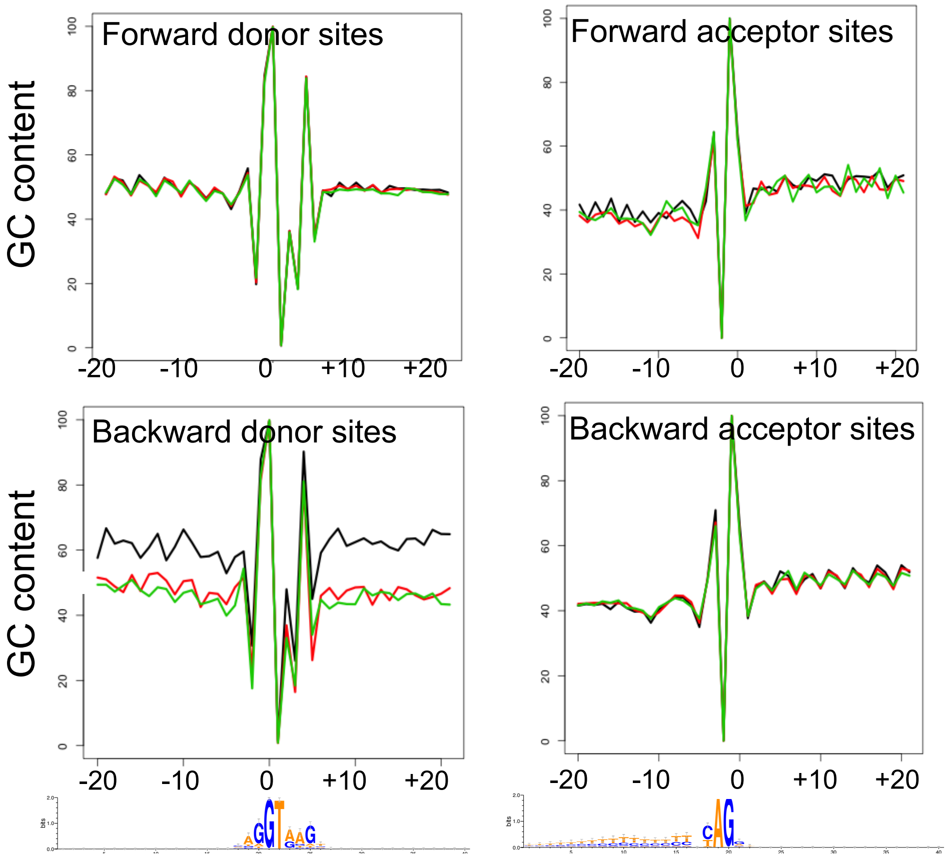


Figure 25. Profiles of GC contents around donor site (Left) and acceptor site (Right) of pre-mRNA averaged for the 1st, 2nd, and 3rd splice site (black, red, and green) from the front (Upper) and from the back (Lower). Sequence logos were constructed for all of pre-mRNA sequences.

[15, 65], and a stem-bulge structure upstream of the donor site is associated with the induction of Rex protein binding in HTLV-2 [66, 67] or the reduction of U1 snRNP binding in exon 10 of *tau* [68]. In contrast, the structural profiles around acceptor sites contain three separate peaks: p_{hairpin} at 3-9 nt upstream of the acceptor site, p_{multi} at 10-30 nt upstream of the acceptor site, and p_{exterior} at 13-40 nt upstream of the acceptor site. These peak locations are roughly within the polypyrimidine tract, which is the known binding site for U2AF and PTB [69]. In a previous study, the existence of loop structures was predicted to change the activity of the neighboring alternative acceptor sites in yeast [70]. Accordingly, such preferences for loop types generated by sequential motifs may help optimize the binding efficiency of constitutive splicing factors. As these preferences for specific loop types around motif sites have not been investigated in previous studies, identifying them and the splicing activity of each site can reveal unknown loop preferences optimized for binding a particular splicing factor.

3.4 GENOME-WIDE SIMULATION BY PARASOR

k-mer frequency linear regression to remove sequence composition biases

I designed a normalization method for genome-wide comparisons of stem probabilities using GC content and other, more complex features as described in the section “Linear regression

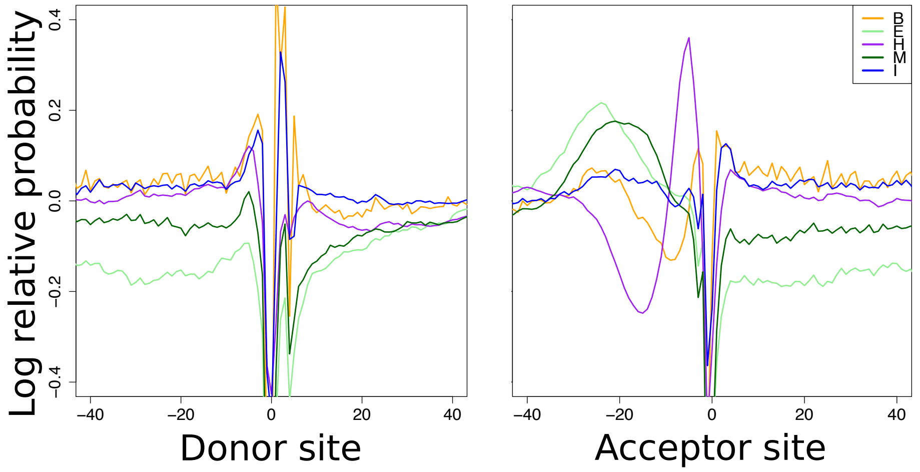
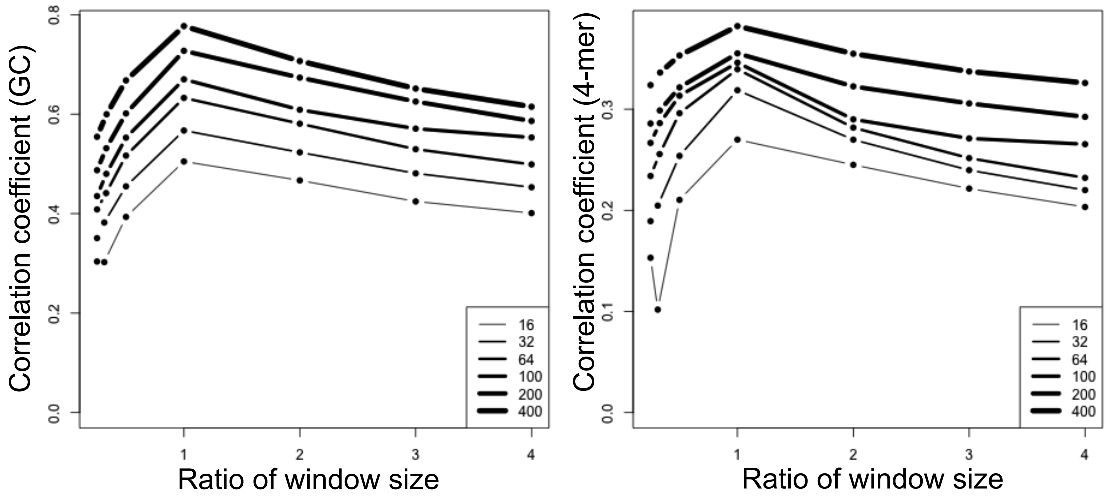


Figure 26. Log relative probability around donor sites (Left) and acceptor sites (Right) for Bulge (B), Exterior (E), Hairpin (H), Multi (M), and Internal (I) loops, which are represented by orange, light green, purple, dark green, and blue lines, respectively. Each position shows $\log(\mu_{p_\delta}(i)/\overline{\mu}_{p_\delta})$ for the loop type δ. A 0 position indicates the starts of introns for donor sites, and the starts of exons for acceptor sites.

of stem probability with sequence composition statistics". Using Python 2.7 and the NumPy library, I implemented a linear regression using the average stem probability $p_{\text{stem}}(i)$ with a ridge penalty. The model was trained with the stem probabilities on both strands of the entire human genome. The maximal span W for computing stem probabilities is basically set to 200 for all the experiments. In the previous section, I have shown a histogram of stem probabilities for $W = 200$ and $W = 1,000$ (Figure 11(c)). Although the stem probability gradually increases with W as the number of possible base pairs increases, the accuracy of structure prediction is not much affected by the value of W. As shown in Table 1, a large maximal span ($W = 1,000$) only slightly decreased the accuracy of prediction against PARS score dataset.

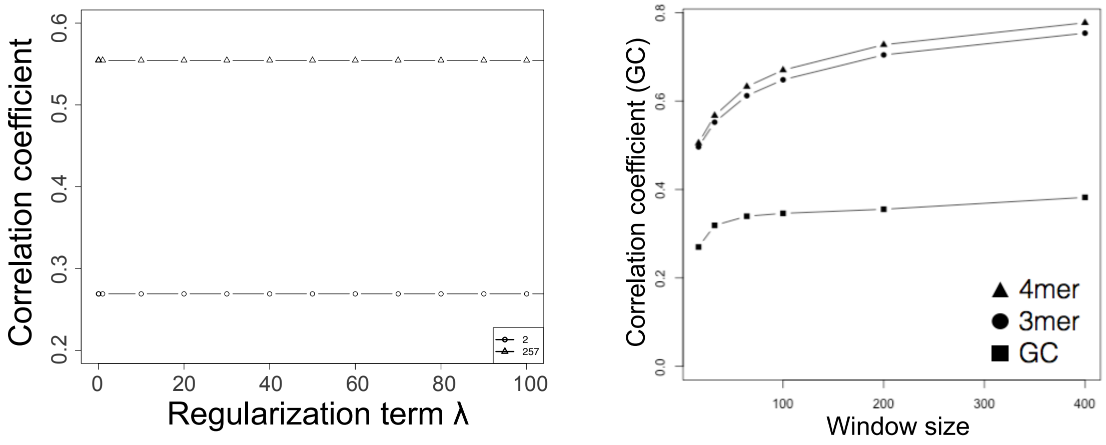
For parameter optimization, I used stem probability data from both strands of human chromosomes and calculated the correlation coefficient between y_n and $w^T x_n$. In this analysis, 4-mer composition exhibited a higher correlation coefficient with stem probabilities than did GC content (Figure 27). Although other window sizes were tested for summarizing features, feature window sizes that were the same as those of average stem probabilities achieved the highest correlation coefficients. Based on this result, I used a 32-mer window size for feature calculation in the subsequent regressions of average stem probability $\bar{p}_{\text{stem}}(i)$. Figure 28(a) shows the result of λ estimation with GC content and 4-mer regression for human chromosome 1. An alteration of λ is observed to have almost no influence on the correlation between y_n and $w^T x_n$, likely because there is sufficient data to avoid overfitting. Hence, I set λ to 0 in subsequent analyses. Figure 28(b) shows the highest correlation coefficients of stem probability with GC content, 3-mer composition, and 4-mer composition. Figure 29 shows scatter plots of these GC content and 4-mer composition regressions with the actual stem probability for each 32-mer fragment $\bar{p}_{\text{stem}}$. Owing to its fewer parameters, GC content was outperformed by 4-mer composition in data normalization. Accordingly, I mainly discuss the result of stem probabilities normalized by 4-mer composition $\Delta\bar{p}_{\text{stem}}$ (computed by ($\bar{p}_{\text{stem}}$ − regression of $\bar{p}_{\text{stem}}$)). However, structure tendency determined by my analyses generally does not change between the GC and 4-mer normalization analyses. This normalization additionally reduces autocorrelation (Figure 30). According to this analysis, normalized stem probability indicates structure ten-



(a)

(b)

Figure 27. Influence of the different feature window size on linear regression. Correlation coefficients of averaged p_{stem} and the regressed value by L2 regression were computed using the feature of (a) GC contents and (b) 4-mer compositions with different window sizes for feature calculation (shown in the legend). The x-axes represent the relative window size for computing averaged p_{stem} and the regressed value to the window size for the feature selection.



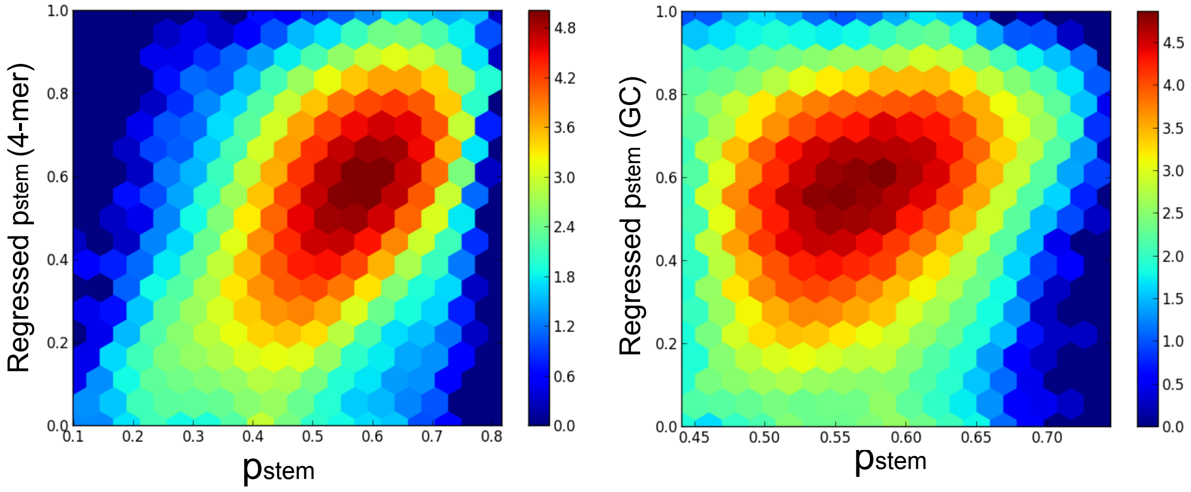
(a)

(b)

Figure 28. Influence of the regularization term and feature selection on linear regression. (a) Correlation coefficients of averaged p_{stem} and the regressed value by L2 regression with varying the regularization term λ. For GC content (represented as 2) and 4-mer composition (257), λ has little influence on normalization efficiency. (b) Correlation coefficients of averaged p_{stem} and the regressed value with the different window size for both of averaging and feature calculation (shown in x-axis).

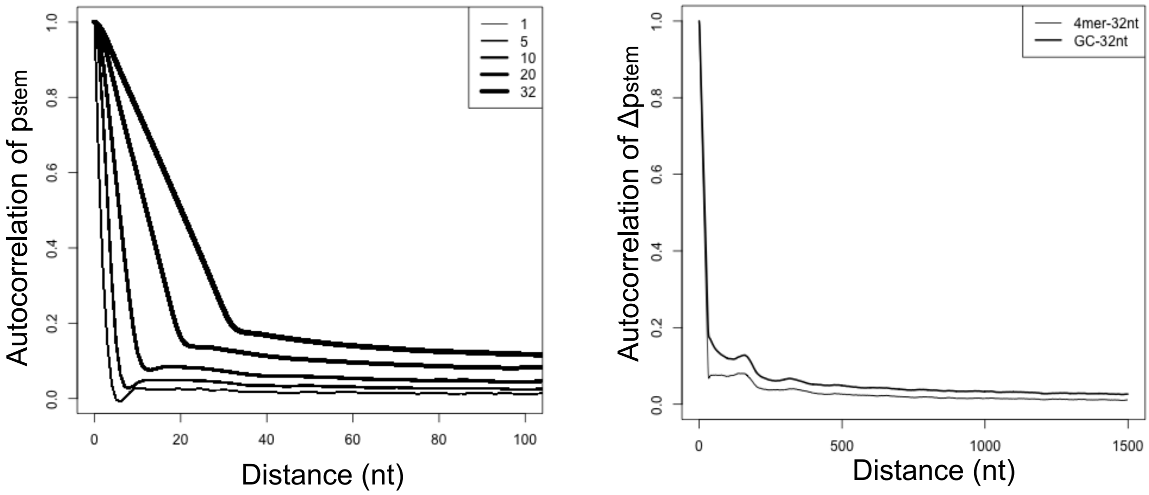
dency within a narrow range compared to raw stem probability, which is highly influenced by regional heterogeneity in sequence composition.

I subtracted the average probability predicted by regression from $\bar{p}_{\text{stem}}(i)$, which is denoted by $\Delta\bar{p}_{\text{stem}}(i)$. In Figure 30, I show that this subtraction greatly reduces the correlation between



(a) (b)

Figure 29. Efficiency of linear regression for stem probability. Hexagonal binning plots for averaged p_{stem} (y-axis) and “predicted” stem probability of regression (x-axis) with features of GC content (a) and 4-mer (b) compositions.



(a) (b)

Figure 30. Autocorrelations of stem probability. Autocorrelations of (a) p_{stem} and (b) $\Delta\bar{p}_{\text{stem}}$ was computed using GC and 4-mer compositions as the feature of regression. The distance between two samples applied for the computation of autocorrelation is shown in x-axis.

neighboring 32-nt windows, ensuring independence between samples. As such, I used $\Delta\bar{p}_{\text{stem}}(i)$ of all non-overlapping 32-nt windows as independent degrees of freedom for hypothesis testing. Also, I re-implemented several hypothesis testing algorithms for cases in which popular statistical tools such as R cannot handle the necessarily large number of data points such as Figure 33 and 39.

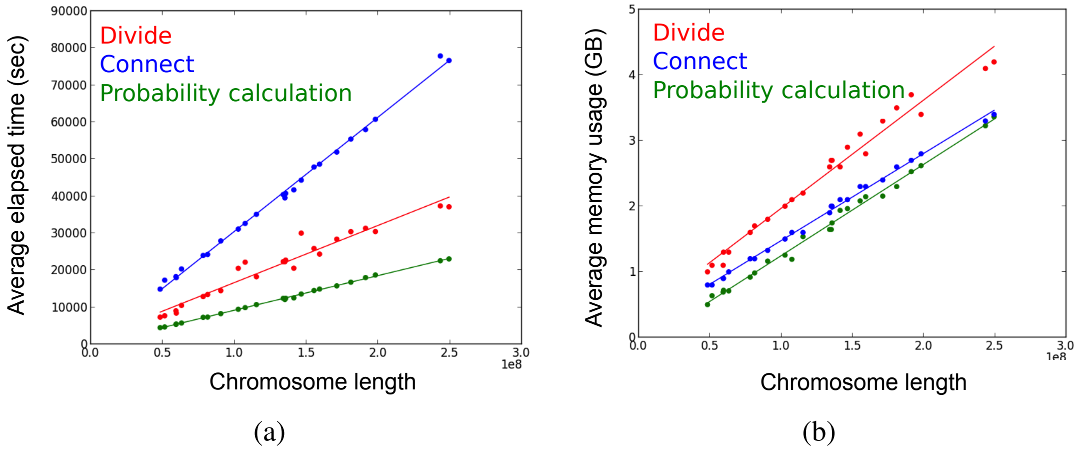


Figure 31. ParasoR performance. Average elapsed times in seconds (Left) and required memory sizes in GB (Right) for the simulation of each chromosome. p_{stem} was computed for the continuous chromosomes using at most 300 nodes in hgc super computer. X-axis represents the chromosome length (nt). The computation time and required memory was averaged for both strands as well as all of nodes during Divide (red), Connect (blue), and Probability calculation procedure (green).

Calculation time and memory for the structure simulation of human genome

I applied ParasoR to human chromosomes and surveyed its resource and time efficiency by measuring the computational time and memory used by the HGC super computer. Three hundred computational nodes were used in parallel to analyze the human chromosomes. Figure 31 shows the time and memory usage averaged for the plus and minus strand of each chromosome. For this genome-wide calculation, ParasoR spent 0.92 days at most for each chromosome, and the computational time is consistent with its linear dependency on sequence length N. Maximum memory usage was approximately 4GB, which is typically available in personal computers. I measured the computational time and memory usage for analyzing a concatenated human genome sequence ($\sim$ 3.1G) with an HGC super computer.

Table 2 shows the maximum computational time of each process in memory-saving mode for the human genome sequence. Using $3,000$ nodes, ParasoR requires at most 3.3 GB of memory for these processes. If I ideally run all processes in parallel, ParasoR can complete its probability calculation of the entire human genome in 18.2 hours.

Table 2. Computational time for human genome sequence analysis

Procedure	Computational time (sec.)
1st Divide	$22,084$
1st Connect	403
2nd Divide	$20,608$
2nd Connect	$5,932$
Probability Calculation	$16,508$

Comparison of calculation time with other tools

For comprehensive mRNA analyses, I measured a running time of two of sliding-window methods, LocalFold and RNAplfold, for the calculation of stem probability or accessibility. My dataset contains totally $140,031,550$ nt of mRNA sequences and I distributed mRNA sequences into 300 multiple FASTA files to contain the same number of bases as much as possible. I computed elapsed times of RNAplfold and LocalFold for these 300 files and obtained an average calculation time (Table 3).

For the sequence of $\sim$ 140M nt, a calculation time of ParasoR was estimated for Divide, Connect, and Probability calculation procedure by the slope and intercept value obtained from the running time of human chromosomes. This suggested that a running time of ParasoR is comparable to that of LocalFold. In addition, it is also possible to accelerate the structure calculation even more when I have already constructed database because ParasoR can skip the step of Divide and Connect procedure.

Random sequences for the simulation of pre-mRNA and intron computation

While ParasoR is the only application which is available for human chromosomes and genome sequence, it is just a simulation and such long RNA is never transcribed in real. To clarify the utility of ParasoR parallelization for real transcripts, I applied ParasoR algorithm of single core and multiple core mode to the random sequences whose length is in the range of human pre-mRNA (from $1,000$ to $1,000,000$ nt) and which have 50 % GC content . Figure 32 shows an average elapsed time of ParasoR in the single core or multiple nodes of HGC super computer for 100 sequences. In ParasoR, the number of computer nodes is automatically reduced to an appropriate size depending on the balance of sequence length and maximal span. Thus, I showed the maximum number of available computing nodes in Figure 32. In addition, only a single sequence was applied to measure the running time of $100,000$-nt and $1,000,000$-nt sequence dataset due to the problem of long calculation time. I plotted a hundredfold running time for these conditions. As a result, the running time proportionally increased according to the sequence length N. The result of ParasoR running time is less than tenth of that of Rfold implementation. Although ParasoR with multiple nodes possesses an overhead cost for distribution and requires an additional process such as database construction, this parallelization achieved a substantial acceleration of the structure prediction for longer sequences as the number of computer nodes increases. In the human pre-mRNA dataset, the longest intron exceeds $1,000,000$ ($1,055,451$)

Table 3. Computational time for human transcripts

Software	Computational time (sec.)
ParasoR	78,839
(Divide)	42,906
(Connect)	22,874
(Probability calculation)	12,958
LocalFold	87,547
RNAplfold	926

For ParasoR, computational time was estimated using a running time for human chromosome.

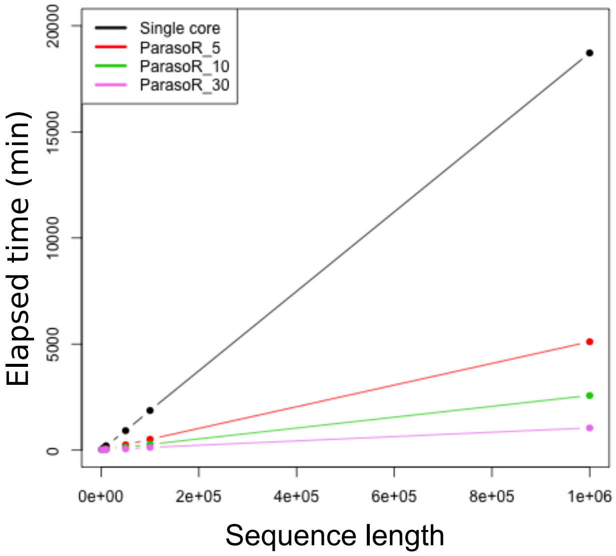


Figure 32. Relationship between elapsed time (minutes) and the number of nodes. Elapsed time of ParasoR was measured with single core or multiple nodes (setting the maximal number of computing nodes to 5, 10, and 30) for 100 sequences with the maximal span 200. X-axis indicates the sequence length of the random sequences.

and ParasoR also showed its efficiency for the sequences of such length. Therefore, I concluded that ParasoR algorithm is still effective for pre-mRNA and full-length intron sequences.

Genome-wide simulation to detect structural constraints on transcribed regions

Since the energy scale of secondary structure folding is high enough to influence the efficient progression of various biological processes, such as transcription elongation and translation, I expect many transcribed regions in the genome to be subject to various structural constraints. To study the structural preferences of transcribed regions relative to untranscribed regions, I computed the distributions of average stem probabilities $\bar{p}_{\mathrm{stem}}(i)$ for 32-nt windows over both the strands of entire human chromosomes, and compared those of the different functional regions. This window size was chosen because it produced distributions that were close to the normal distribution for which statistical analyses are easier. Also, $\bar{p}_{\mathrm{stem}}$ are expected to represent local structural features better than single-base stem probabilities Figure 33(a) shows the distributions of $\bar{p}_{\mathrm{stem}}$ among five types of genomic regions: 5′-UTR, 3′-UTR, CDS, Intron, and Intergenic regions. Here, I removed repeat sequences elements from these regions. Further, the Intergenic regions are defined as the genomic regions that contain no repeat regions, no sense or antisense sequences of transcribed regions, and no sequences close to their boundaries (see the Methods section). All annotation categories exhibit a similar unimodal distribution. The 5′-UTR category apparently has the highest median, which is consistent with the elevated GC content around 5′-UTR regions [20]. The descending order of the stem probability medians (5′-UTR, 3′-UTR, and CDS categories) is the same as that of their structural strengths computed from mRNAs (Figure 20(c)) in the previous section. It also shows all transcribed regions have higher median stem probabilities than those of the Intergenic regions, which may suggest the

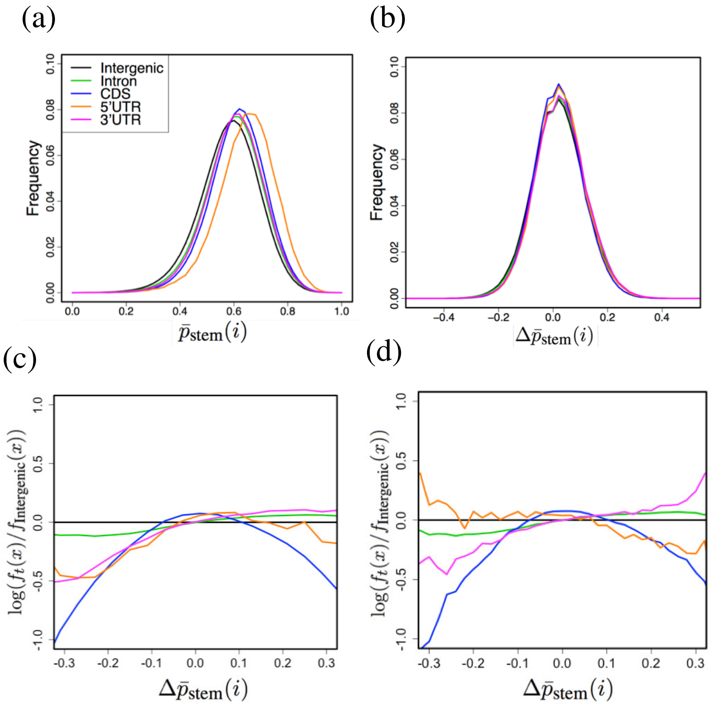
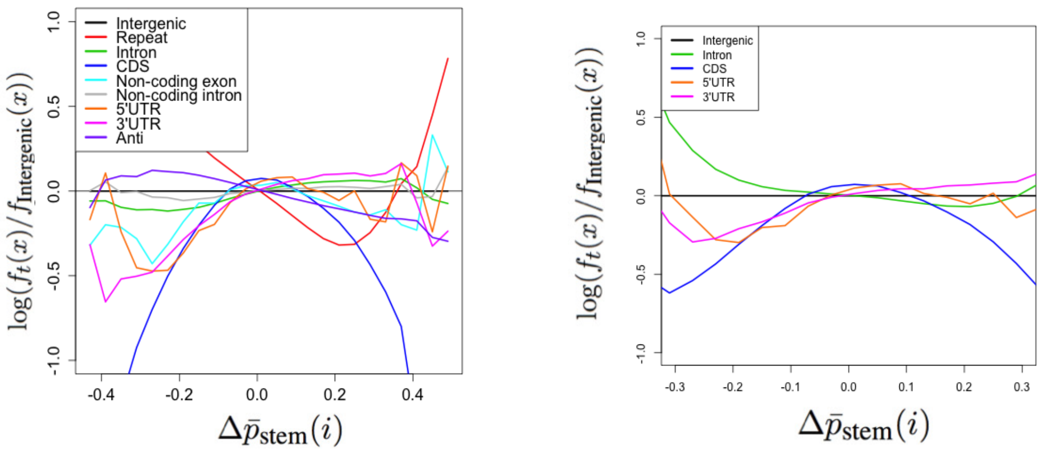


Figure 33. Structure propensity of Intergenic regions, Intron, CDS, 5′-UTR, and 3′-UTR (represented by black, green, blue, orange, and pink, respectively). (a) Distributions of raw $\bar{p}_{\text{stem}}(i)$ for each annotation category. (b) Distributions of normalized average stem probabilities $\Delta\bar{p}_{\text{stem}}(i)$. (c) Log ratios of densities $\log(f_t(x)/f_{\text{Intergenic}}(x))$, where $f_t(x)$ is the probability density of $\Delta\bar{p}_{\text{stem}}(i)$ at x for $t =$ Intron, 5′-UTR, CDS, 3′-UTR. (d) is similar to (c), except that $\Delta\bar{p}_{\text{stem}}(i)$ was computed for the true boundaries of pre-mRNAs.

hypothesis that transcribed regions are constrained by their secondary structure. I confirmed these genome-wide analyses of structure preference using the mouse genome. The raw and normalized stem probabilities agree with those observed across the human genome (Appendix Figure 66). As such, structural features may be shared between even distantly related species.

However, it should be noted that genomic sequences are also subject to various constraints that are unrelated to RNA secondary structure, and various characteristics of stem probabilities in transcribed regions may be side effects of sequence biases caused by such constraints. Therefore, I modeled the influences of sequence biases by training a linear regression model with $\bar{p}_{\text{stem}}(i)$ as targets and 4-mer frequencies as features. I then computed the normalized stem probability $\Delta\bar{p}_{\text{stem}}(i)$, which is the difference between $\bar{p}_{\text{stem}}(i)$ and its regressed value. This normalization mostly eliminated the differences in the medians among annotation groups so that the distinct difference in $\bar{p}_{\text{stem}}$ median values was explained by a sequence bias (Figure 33(b)). Then, I focused on the residual part $\Delta\bar{p}_{\text{stem}}$, because large $\Delta\bar{p}_{\text{stem}}$ values represent structural preferences that are not merely explained by 4-mer frequencies.

To extract a faint structural propensity of each transcribed region in $\Delta\bar{p}_{\text{stem}}(i)$ compared to Intergenic regions, I plotted the log ratios $\log(f_t(x)/f_{\text{Intergenic}}(x))$, where $f_t(x)$ is the probability density of $\Delta\bar{p}_{\text{stem}}(i)$ at x for $t =$ Intron, 5′-UTR, CDS, 3′-UTR (Figure 33(c)). As for the CDS regions, the density of this ratio is more concentrated around the center (Conover test, $p < 10^{-1586}, n \sim 10^5$; the sample size is detailed in the Methods section), which indicates that the structural strengths in the CDS regions are more strongly determined by their base compositions than Intergenic regions. Additionally, introns and 3′-UTRs contain a higher rate of structured regions than that of Intergenic sequences, while 5′-UTRs, 3′-UTRs, and introns all exhibit



(a) (b)

Figure 34. Structural propensity of each annotation group. (Left) Log odds ratios $\log(f(x)/f_{\text{Intergenic}}(x))$ for Δp_{stem} distributions of all of annotation groups on human genome. Repeat region shows the specific tendency to have a peak far from 0. (Right) Log odds ratios for p_{stem} distributions of each annotation without the removal of repetitive regions.

lower rates of accessible regions. Figure 33(d) is similar to Figure 33(c), except that $\Delta\bar{p}_{\text{stem}}$ was calculated for pre-mRNAs, rather than for chromosomes. In this analysis, the distributions of introns, CDS, and 3′-UTR regions are qualitatively similar to those in Figure 33(c). In contrast, 5′-UTRs exhibit increased accessible regions in a wide range ($-0.3 < \Delta\bar{p}_{\text{stem}} < 0.0$) as compared to that shown in Figure 33(c). This is likely explained by boundary effects such as the insertion of ambiguous characters or the phenomena in which the tips of sequences tend to be accessible (described in the subsection "Property of energy model and stem probability"). The structure propensity of pre-mRNAs and mRNAs was more investigated in the next section.

I also examined a distribution of both raw and normalized stem probabilities generated for other types of annotation. In Figure 34(a), non-coding regions are observed to be more similar to introns than to CDS regions. This comparison was repeated for log ratios of densities $\log(f_t(x)/f_{\text{Intergenic}}(x))$ (t = genomic region); x normalized average stem probability $\Delta\bar{p}_{\text{stem}}(i)$). Most of annotation categories in transcribed regions contain more structured fragments than do intergenic regions. Transcribed regions are thus inferred to be more likely to form base pairs than are intergenic and repetitive regions. In addition, the distribution of these log ratios for CDS and repeat regions are notably different from those of the other annotation categories. In particular, repeat regions were unlikely to be sufficiently normalized and can be incorrectly estimated to be less structured. Figure 34(b) shows a similar result for the transcribed regions, though it includes repetitive fragments. Although the distribution of raw stem probabilities is less influenced by this inclusion, negative normalized stem probabilities increased the log ratio of each annotation category. To determine the structural tendency of such regions, it would be necessary to make comparisons among regions with the same repeat sequences in different positions or annotation categories. Unfortunately, because I aimed to compare structure constraints in the scale of functional regions, I removed all repetitive regions from other annotation categories.

I normalized stem probability by GC content of each window, not by k-mer frequencies, and the same tendency was obtained, though GC content did not normalize the unimodal

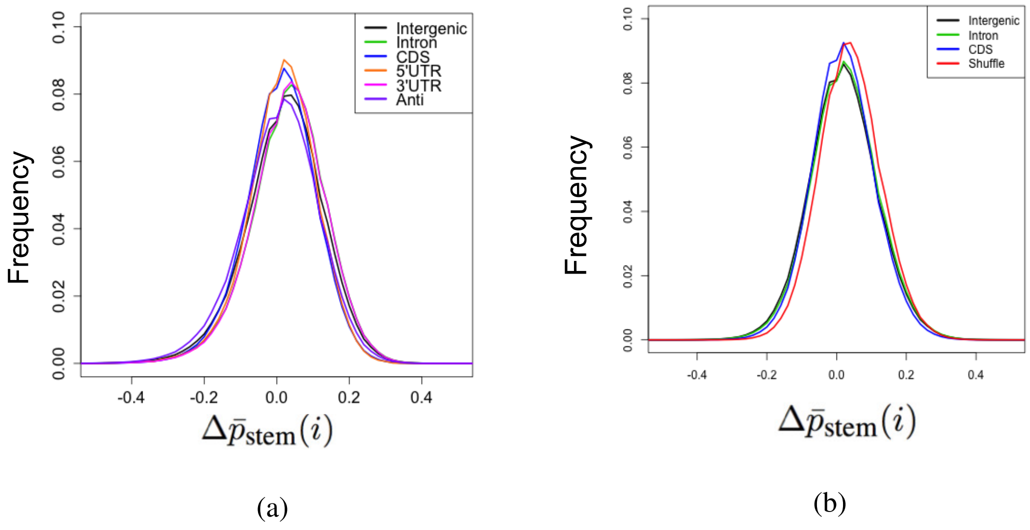


Figure 35. Structural propensity with other controls. (a) Histogram of $\Delta\bar{p}_{\text{stem}}$ computed by GC regression for human genome. (b) Histogram of $\Delta\bar{p}_{\text{stem}}$ using the L2 regression with 4-mer compositions. $\Delta\bar{p}_{\text{stem}}$ of Shuffle region was computed for the sequence of 1-mer shuffled chromosome 1 while those of the others were computed for original genome-sequence.

distribution as well as 4-mer composition did (Figure 35(a)). When using stem probabilities of 1-mer-shuffled human chromosome 1 (detailed in the Methods section), the random sequence was highly concentrated at 0.6 compared to intergenic regions (data not shown); on the other hand, a normalized stem probability distribution of shuffled sequence contains more structured regions (Figure 35(a)). This is likely because repetitive, GC-rich regions contribute to increases of structure across regions. This suggests that each genomic region is less structured than completely randomized sequence because of various biological constraints.

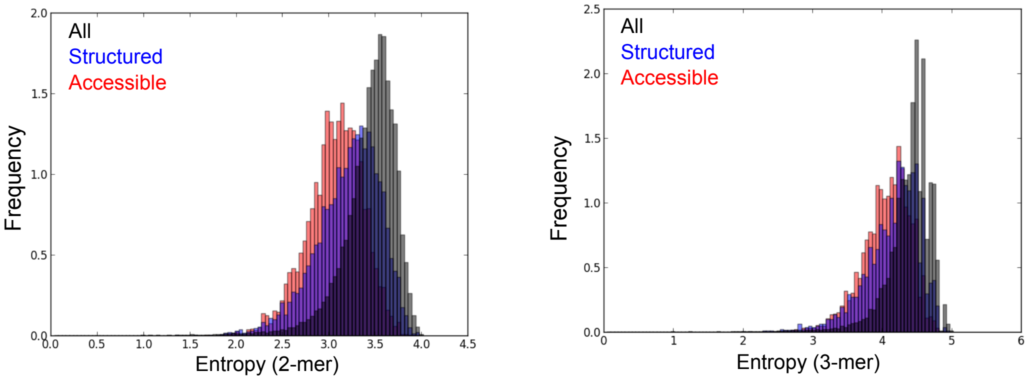
Genomic features in structured and accessible regions after normalization

In previous analyses, structural preferences were detected for each annotation category and compared with intergenic regions. In this subsection, I investigated the correlation between genomic sequence features and structural preferences regardless of genomic annotation.

To determine the cause of these structure tendencies, accessible and structured regions were defined as having normalized stem probabilities of more than 0.3 and less than −0.3, respectively. Then, I computed the entropy of 2-mer and 3-mer frequencies in accessible and structured 32-mer regions as follows.

$$p(s) := \frac{\text{frequency of some k-mer } s}{\sum_{s'} \text{frequency of some k-mer } s'}$$
$$E_{\text{k-mer}} = \sum_{s} -p(s)\log_2 p(s)$$

For comparison, the entropy of chromosome 1 was calculated as a background. Figure 36 shows an example entropy distribution computed for 2-mer and 3-mer fragments on human chromosome 1. As expected, accessible and structured regions are biased toward a low entropy



(a)

(b)

Figure 36. Histograms of the entropy of 2-mer (a) and 3-mer (b) compositions for each 32-mer fragment on human chromosome 1. Both of images exhibit the entropy distribution for whole fragments (black), structured fragments where $\Delta\bar{p}_{\text{stem}} > 0.3$ (blue), and accessible fragments where $\Delta\bar{p}_{\text{stem}} < -0.3$.

in comparison with all of chromosome 1. In particular, accessible regions are likely to have lower entropy than structured regions are.

The relationship between structure tendency and conservation was examined using PhastCons scores [71]. Appendix Figure 67 shows boxplots of normalized stem probability grouped by PhastCons score. Although there is a fluctuation of the median, no correlation was observed between conservation and structure tendency. The Pearson product-moment correlation coefficient was also not significant ($p > 0.05$).

3.5 STRUCTURAL PREFERENCES IN MRNA AND PRE-MRNA

Structural preferences of human and mouse transcripts

In the previous section, specific structural tendency was observed in transcribed regions of genome sequences. In this section, structural features of transcribed region were surveyed in the forms of mRNA and pre-mRNA. First, a distribution of average stem probabilities $\bar{p}_{\text{stem}}$ was calculated for mRNA and pre-mRNA in human and mouse transcriptomes (Figure 37). Analysis of these datasets yields similar distributions as those of full genome sequences, but repeat regions in the mouse genome produce a small peak among extremely unstructured regions, and this may be produced by distinctive repetitive elements in the mouse genome such as SINE B1 and B2 elements. I also calculated log ratios of densities $\log(f_t(x)/f_{\text{Intergenic}}(x))$ where $f_t(x)$ is the probability density of $\Delta\bar{p}_{\text{stem}}(i)$ at x for each annotation group t on mRNA and pre-mRNA sequences in the same way as in the genome analyses. All groups, except for CDS, 5′-UTR, and repeat regions, are likely to be more structured than intergenic regions. While CDS and repeat regions are similar to those of entire genome sequences, 5′-UTRs are more accessible than their expected distribution based on sequence composition. This result is consistent across human and mouse transcriptomes. In the log ratio against intergenic regions, only antisense introns exhibit an opposite tendency compared with their sense strand counterparts (Figure 37(Center)). This is likely because intronic sequences have an AT/CG asymmetry, which forces antisense regions to have a tendency against structure. To confirm such structure tendency in transcribed

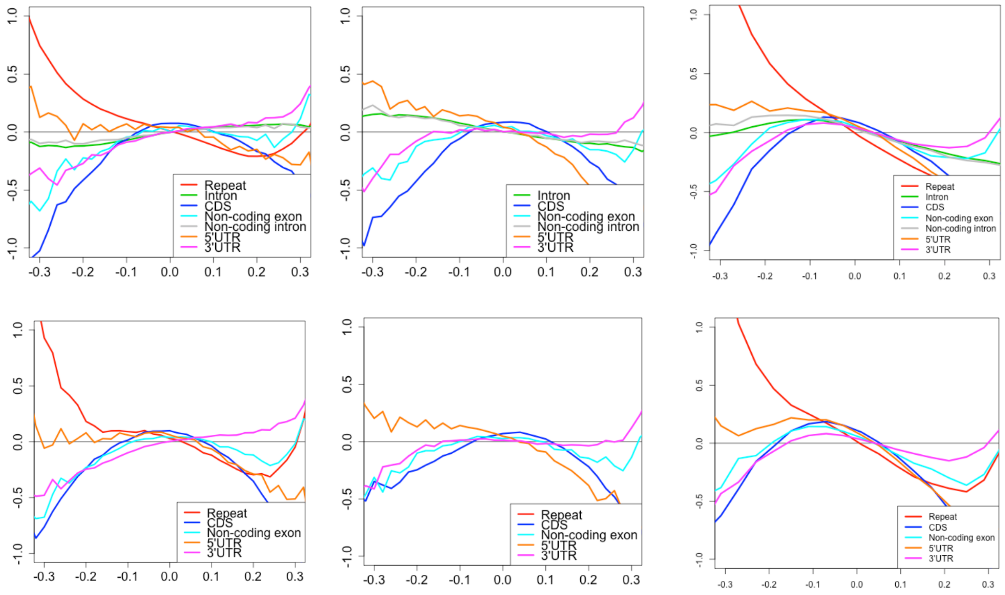


Figure 37. Structure propensity of each annotation group with different conditions. Log odds ratios $\log(f(x)/f_{\text{Sense Intergenic}}(x))$ for $\Delta\bar{p}_{\text{stem}}$ distributions of each annotation on pre-mRNA (Top) and mRNA (Bottom) of sense sequence in human genome (Left), antisense sequence in human genome (Center), and sense sequence in mouse genome (Right). 5′-UTR shows an elevation of log odds ratios at the region of positive x, indicating that the group contains the larger ratio of accessible fragments compared to that of the Intergenic group.

regions, both antisense sequences and shuffled 3-mer sequences were also used as a reference for comparison because they have the same GC content and sequence lengths.

I compared the distribution of sense sequences and these backgrounds using *D* statistics. To compare structural similarity, *D* statistics from the Kolmogorov-Smirnov test between sense strand and background distributions were evaluated using the ks2samp function in the SciPy library. Figure 38 shows *D* statistics of mRNA and pre-mRNA compared with the normalized stem probability distribution of antisense and shuffled 3-mer sequences. I used all of regions of shuffled sequences for comparison although repetitive regions and multi-annotated windows were excluded from sense and antisense sequences. Both intronic and exonic regions of transcripts significantly differed from their antisense (*p*-values: exon mRNA, $9.66e-20$; exon, pre-mRNA ~ 0.0; intron ~ 0.0) and shuffled 3-mer sequences (*p*-value: mRNA, ~ 0.0; pre-mRNA ~ 0.0). All of *D* statistics are positive and this means that the exonic region is unlikely to be structured compared to antisense and shuffled 3-mer sequences.

To estimate the statistical significance of structural preferences of each annotation category relative to Intergenic regions, Wilcoxon's rank sum test was applied to the distribution of $\Delta\bar{p}_{\text{stem}}$ for each annotation group, as well as their antisense sequences. Figure 39 shows the *Z*-scores of Wilcoxon's rank sum tests; a positive (negative) value indicates the region contains a higher (lower) ratio of structured regions than that of Intergenic regions. I observed that the number of structured regions of introns is significantly higher than that of Intergenic regions at a significance level of $p < 10^{-7940}$ ($n \sim 10^7$, calculated from *Z*-score with a one-sided test, hereafter). In contrast, the antisense sequences of introns significantly contain more accessible positions than Intergenic regions ($p < 10^{-13655}$, $n \sim 10^7$). Antisense and sense sequences

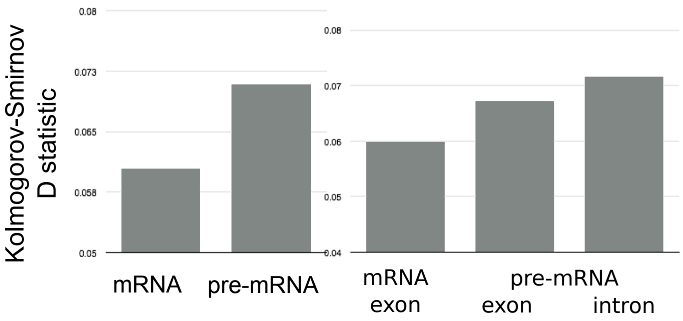
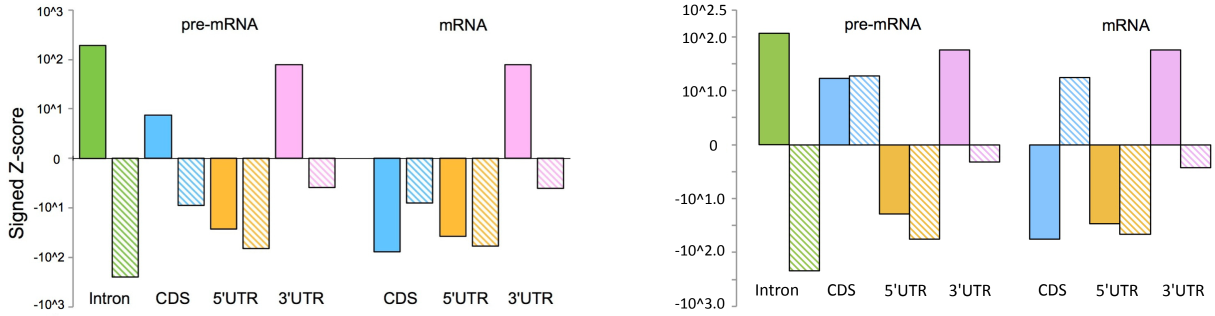


Figure 38. Structure propensity of transcripts with other controls. Kolmogorov-Smirnov D statistics for $\Delta\bar{p}_{\text{stem}}$ distributions of exonic and intronic regions sense and anti-sense transcripts and $\Delta\bar{p}_{\text{stem}}$ distributions of sense and 3-mer shuffled sense transcripts with (pre-matured) or without (matured) introns.



(a)

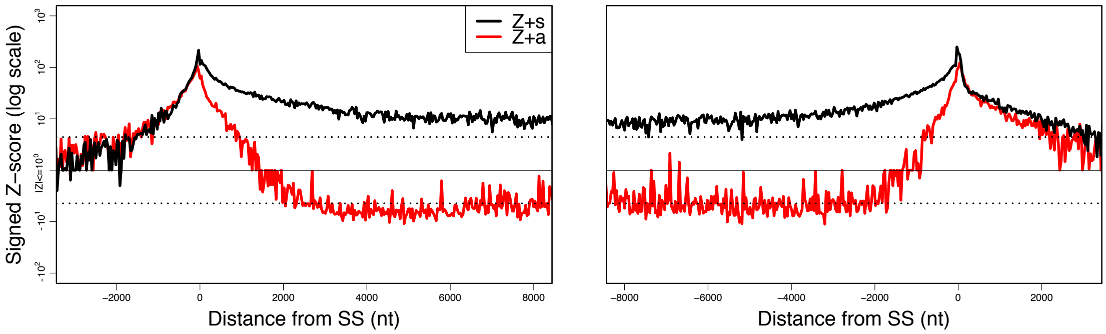
(b)

Figure 39. Z-score of Wilcoxon's rank sum tests used to assess the structural preference for (a) human and (b) mouse. Y-axis value indicates the annotation category has a higher (lower) stem density than that of Intergenic regions. Filled and shaded bars represent Z-scores of sense and antisense sequences of each annotation, respectively.

have the same GC content, but they can show a different strength of 4-mer normalization as well as $\bar{p}_{\text{stem}}$ (detailed in the subsection "Difference of structure propensity between sense and antisense sequence". Hence, such a different tendency of sense and antisense sequences cannot be explained by systematic differences in GC content or other strand-symmetric sequence features between the intron and intergenic regions. Additionally, 3′-UTRs exhibit the same trend, but at a lower significance level (sense: $p < 10^{-151}$, antisense: $p < 10^{-940}$, $n \sim 10^6$). The 5′-UTR sequences possess more accessible positions than do Intergenic regions, which is consistent with Figure 33(d). When pre-mRNA and mRNA are compared, I observe that the Z-scores of CDS change from positive to negative (pre-mRNA: $p < 10^{-13}$, $n \sim 10^6$, mRNA: $p < 10^{-1301}, n \sim 10^6$), while 5′-UTRs and 3′-UTRs do not exhibit notable changes. The increased accessibility after splicing in the part of CDS regions suggests the existence of structural constraints in the particular mRNAs for translational efficiency or resistance to degradation [72].

Positional profiles of structural preferences around splicing sites

To reveal positional profiles of structural preferences, I computed the normalized average stem probability of 32-mer fragments $\Delta\bar{p}_{\text{stem}}$ along the sense and antisense strands. After this,



(a) (b)

Figure 40. Positional profiles of structural preferences around splicing donor (a) and acceptor (b) sites. Z-scores of Wilcoxon's rank sum statistics for the normalized average stem probability are drawn in black for sense and in red for antisense sequences. Dotted lines represent Z-scores that correspond to the Bonferroni-corrected p-values (< 0.05) in a one-sided test.

I computed Z-scores for the Wilcoxon's rank sum test of the distributions of each position surrounding these SSs with the intergenic distribution of such data. To calculate $\Delta\bar{p}_{\text{stem}}$, each pre-mRNA sequence was fragmented into 32-mers from the start or end of each SS. Then, $\Delta\bar{p}_{\text{stem}}$ was determined according to its distance from the nearest SS.

Positional dependency was examined by computing the Z-scores for each 32-nt positional bin around the donor and acceptor sites in pre-mRNAs using Wilcoxon's rank sum statistics (Figure 40). Both sense and antisense strands exhibited a positive peak around the donor ($p < 10^{-10054}$ for $342,755$ SSs) and acceptor ($p < 10^{-13512}$ for $326,618$ SSs) sites. This pattern indicates structural constraints for splicing regulation [73], which is not easily explained by primary sequence biases. Inside exons, both sense and antisense Z-scores approach zero as the distance from SSs increases. In contrast, Z-scores for the sense strand remain positive within introns ($8,000$ nt downstream of the donor site, $p \sim 10^{-28}$ for $26,970$ SSs; $8,000$ nt upstream of the acceptor site, $p \sim 10^{-9}$ for $27,126$ SSs), and those for the antisense strand become negative. As the Z-score for each bin was independently computed, the entire range of introns appears to be subject to structural constraints.

I note that a significant structural preference can be caused by only a small portion of transcribed regions, owing to the large degrees of freedom of the hypothesis tests. For example, my results suggest that the entire intronic regions are dispersed with small intronic elements that tend to be more structured as compared to controls, but this does not necessarily mean that the majority of introns forms highly stable structures. Despite these technical intricacies, I confirmed the same stem preference of intronic regions by different normalization methods, such as regression using GC content instead of 4-mer frequencies and block-wise-shuffled genome sequences instead of Intergenic regions as background (Figures 35 and 38). Furthermore, I also found comparable results for the mouse genome, which implies that these trends are conserved among mammals (Figure 39).

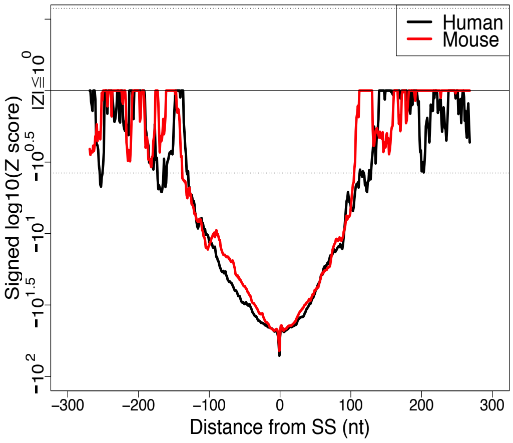


Figure 41. Z-score of Wilcoxon's signed rank test for structural differences caused by splicing events around SSs in human and mouse genomes. The difference of stem probability was averaged for a 32-mer sliding window separately for the upstream and downstream regions of SSs. The dotted line represents a Z-score that corresponds to a significant Bonferroni-corrected p-value (< 0.05) in a one-sided test.

3.6 CONFORMATIONAL CHANGES CAUSED BY SPLICING EVENTS

As described in the previous section, I investigated the structural preferences of transcribed regions by elaborate normalization procedures, such as masking repeat sequences, subtracting contributions from k-mer frequency bias by linear regression, and comparisons of functional regions with Intergenic and antisense regions. In this section, I investigate the structural changes after splicing performed by directly computing the difference of stem probabilities between mRNA and pre-mRNA at upstream and downstream exonic regions from SSs. Although this method cannot analyze intronic sequences, it has the advantage of constancy in the primary sequences for which stem probabilities are compared.

Conformational change after splicing event

I investigate the structural changes after splicing by the difference of stem probabilities between mRNA and pre-mRNA as $\Delta q_{\text{stem}}(i) = p_{\text{stem,mRNA}}(i) - p_{\text{stem,pre-mRNA}}(i)$ for each exonic site individually. Figure 41 shows the positional Z-scores of Wilcoxon's signed rank test for $\Delta q_{\text{stem}}(i)$ of p_{stem} averaged by a 32-nt sliding window, where a negative (positive) Z-score indicates that a nucleotide position changes to be more accessible (structural) after splicing. Both human and mouse analyses show that splicing causes a significant reduction in stem density within approximately 100 bases around the SSs ($p < 10^{-1111}$ at the start position of the left side exon for 343,403 SSs). I also showed a consistent tendency in the case of p_{stem} using a single nucleotide window (Appendix Figure 69).

By calculating p_{cross}, a probability of crossing base pair around SSs, I showed that pre-mRNA tends to form base pairs crossing over both of SSs compared to mRNA (Figure 42). Thus, it is considered that intronic regions around SSs are likely to form base pairing with exonic regions against the exonic regions which are concatenated after splicing.

When I determined $\Delta q_{\text{stem}}(i)$ according to annotation group, $\Delta q_{\text{stem}}(i)$ was significantly more accessible for all annotation groups except 3′-UTRs. However, because the average distance from SSs differs for each annotation type, I filtered out almost unchanged samples with lower

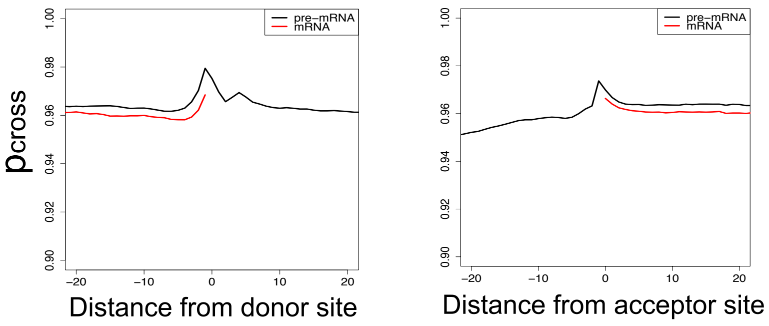
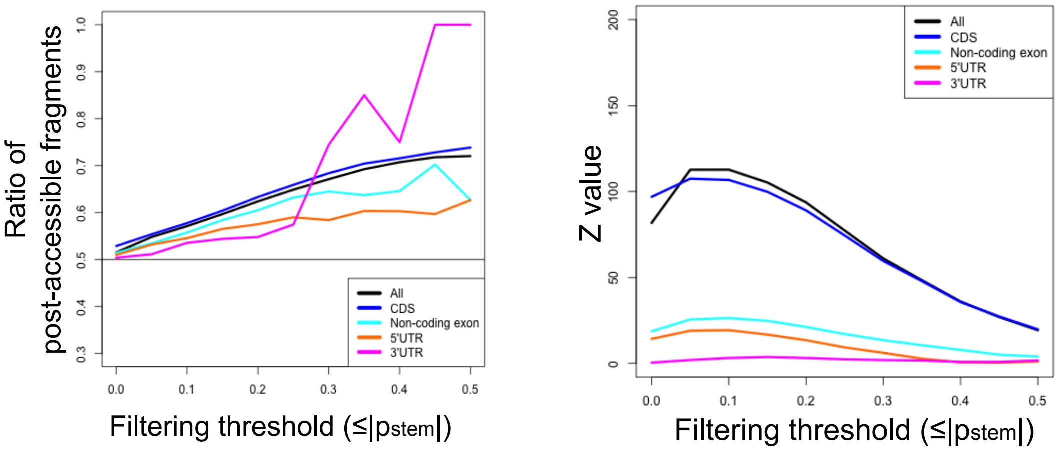


Figure 42. Propensity of structural interaction. Profiles of the probability of crossing base pairs p_{cross} at each position around donor (a) and acceptor site (b). That probability corresponds to (1-probability that neighbor two bases belong to an exterior loop or end of the outermost base pair).



(a)

(b)

Figure 43. Structural difference caused by splicing. (a) Ratios of post-accessible fragments and (b) the absolute values of Z-score computed by Wilcoxon's signed rank test for the difference of p_{stem} on mRNAs or matured non-coding RNAs in 5 annotation groups in human. X-axis represents the threshold for filtering samples where the absolute difference of p_{stem} is lower than the threshold. This filtering process enables to only retain substantially different samples in terms of p_{stem}.

differences than a given threshold so that I could test structure differences only for substantially influenced regions. For each threshold, a number of post-accessible fragments exceeded the post-structural fragments in general (Figure 43(a)). Wilcoxon's signed rank test indicated that CDS and non-coding exon were significantly accessible at every threshold (Figure 43(b) and Appendix Table 10). In contrast, significant influences were not detected in other groups at other thresholds. Therefore, although other annotation groups have a tendency to become more accessible after splicing, their changes are supposed to be not significantly drastic.

Cause of drastic conformational changes

To determine the gene features that are correlated with structural changes, I first computed the median and median absolute deviation of Δq_{stem} computed for each single exonic site within the 200-nt window around each SS. To discover the causes of drastic conformational changes during splicing, I tested the correlation between structure arrangement and four quantitative features: (1) gene expression, (2) GC content around mRNA SSs, (3) GC content around pre-mRNA SSs, and (4) intron length. The gene expression data consisted of CAGE data generated by FANTOM5 [55], and it was used to investigate the correlation of gene expression with structural influences of splicing. Each transcription start site expression count was averaged among different samples and removed as a tissue-specific gene if $\log_{10}$(median expression) ≤ 0.5. Regarding other features, GC content around SSs was calculated for 200-nt windows at exon junctions as well as at 5′- and 3′- SSs.

As measures of conformational change, a median, maximum value, minimum value, and median of the absolute deviation were calculated for the distribution of the disparity in stem probability between mRNA and pre-mRNA $\Delta q_{\text{stem}}(i) = p_{\text{stem, mRNA}}(i) - p_{\text{stem, pre-mRNA}}(i)$ for 200 nt around SSs. To remove the duplicate influence of nearby SSs, each position was exclusively assigned to the nearest SS. SSs with fewer than 100 assigned bases were removed from subsequent analyses.

Median Δq_{stem} values were significantly correlated with both gene expression data and GC content of mRNA and pre-mRNA (Table 4). A negative correlation between gene expression and median Δq_{stem} can be interpreted as a small number of regions becoming structured in highly expressed mRNAs. A median absolute deviation (MAD) of Δq_{stem} was obtained by $\text{median}(|\Delta q_{\text{stem}} - \text{median}(\Delta q_{\text{stem}})|)$ for each SS and used to indicate the magnitude of fluctuation. According to my analyses, the MAD of Δq_{stem} only significantly correlates with mRNA GC content. This suggests that the secondary structures of GC-rich mRNAs are likely to be less altered by the splicing process. In addition, the maximum and minimum Δq_{stem} values are shown in Appendix Table 11. The maximum Δq_{stem} is significantly correlated with all three of these features although the minimum Δq_{stem} is significantly correlated with mRNA GC content only. These results suggest that splicing specifically regulates structured RNA through its effects on translational efficiency.

Gene set enrichment analyses

Finally, I studied enriched functional terms in gene sets that contain the SSs whose structure is dramatically changed through splicing events. I refer to the sites with the median of $\Delta q_{\text{stem}}(i) <$

Table 4. Correlation coefficients between conformational changes and gene features. I tested statistical significance by Pearson's correlation test. (*: $p < 0.05$, **: $p < 1.0 \times 10^{-3}$ after Bonferroni multiple correction) The total number of tested SSs was 108,668 for the features of gene expression correlation coefficient and 261,161 for the other correlation coefficients.

Feature	Median	Median absolute deviation
Gene expression	−0.013**	0.010
mRNA GC %	0.008**	−0.007*
pre-mRNA GC %	0.029*	−0.004
Intron length	−0.004	−0.001

0 and $\Delta q_{\text{stem}}(i) > 0$ as *post-accessible* and *post-structural* sites, respectively; then, I selected the top 10 % of post-accessible and post-structural SSs.

To characterize the relationship between conformational arrangement and biological functions, gene set enrichment analyses were performed for genes that are conformationally changed between mRNA and pre-mRNA. I computed a median $\Delta q_{\text{stem}}(i)$ for each SS and defined a $\Delta\bar{q}_{\text{stem}}(u)$ for a given SS u. I produced subsets of post-accessible and post-structural regions by selecting the top 1 % and 10 % of SSs according to $\Delta\bar{q}_{\text{stem}}(u)$. The top 1 % of post-accessible and post-structural SSs were mapped to 230 and 233 RefSeq genes, respectively, and those of top 10 % were mapped to $2,087$ and $2,195$ RefSeq genes, respectively. Then I checked the distributions of the maximum intron length, average intron length, and number of introns for all and selected genes, respectively. The distributions of the top 1 % and 10 % of genes were significantly different from that of all the genes (Table 5), but the distribution of the top 1 % of genes may have differed as a result of the small sample size.

Then, these genes were remapped to UniProt IDs (208 (post-accessible) and 204 (post-structural) IDs for top 1 % genes, and $1,868$ (post-accessible) and $2,000$ (post-structural) IDs for the top 10 %) for Gene Ontology (GO) enrichment analyses using Ontologizer [74]. Table 6 shows the GO terms that were detected as significantly enriched among influenced genes. There were no enriched GO terms among the top 1 % of genes. However, 23 GO terms were enriched among the top 10 % of post-accessible genes, though no GO terms were enriched among post-structural genes. For post-structural and post-accessible gene sets in the mouse transcriptome, the top 1 % of post-accessible and post-structural genes exhibited 41 and 48 enriched GO terms, respectively. Among the top 10 % of mouse genes, 6 and 10 GO terms were significantly enriched among the post-accessible and post-structural genes, respectively (Appendix Table 12). These GO terms overlapped with those identified in the human analyses-for example, organelle, catalytic activity, and phosphorus metabolic process. Among the top 1 % of genes, however, there was no consensus between human and mouse genes in terms of GO term enrichment, and the distribution of intron features in human and mouse differed among all genes, as mentioned before. Hence, I applied a threshold of 10 % for selecting post-accessible and post-structural genes.

Moreover, functional clustering was performed on the post-accessible gene set with DAVID [57], which eventually established 338 clusters based on this 10 % threshold. Table 7 shows the top three clusters for each test, where the expression analysis systematic explorer (EASE) score represents a significance measure for enrichment and corresponds to the negative log of the average p-value over the functional terms in a cluster. In the human genome, the keywords of the clusters with the highest expression analysis systematic explorer (EASE) score were as follows: cytoskeleton (16.19), kinase (11.90), centrosome (9.18), actin-binding (7.12), and ubiquitin conjugate (6.42). Since the ubiquitin conjugate cluster includes genes related to muscle function such as Titin and to synaptic function such as Synaptoagmin-associated genes,

Table 5. Statistical tests on the distribution of intron features of the top 1 % and 10 % of genes

feature	1 % post-accessible	1 % post-structured	10 % post-accessible	10 % post-structured
maximum intron length	6.26×10^{-15}	2.46×10^{-08}	2.2×10^{-16}	2.2×10^{-16}
average intron length	5.80×10^{-05}	0.042	2.2×10^{-16}	0.0026
the number of intron	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}

I computed Bonferroni-corrected p-values by using Wilcoxon's rank sum test.

this group seems to be involved in the construction and differentiation of muscle or metabolic pathways. I also applied DAVID to post-accessible and post-structural mouse genes, and the pattern is the same in the mouse genome. For post-accessible genes, the extracted functional clusters with the highest EASE score were as follows: cytoskeleton (14.60), ATP-binding (13.11), centrosome (7.33), cell division (5.72), and actin-binding (5.42). For post-structural genes, the extracted functional clusters with the highest EASE scores were as follows: ATP-binding (16.57), kinase (6.19), C2 domain (4.23), and cytoskeleton (4.17). These results may suggest that the genes associated with cytoskeleton, kinase, and other ATP-binding proteins are often post-transcriptionally regulated by the changes to their secondary structures.

Conformational changes of the mRNA that encodes the F-actin binding protein

I also used *NEXN* gene to investigate the difference of γ-centroid structures in sense and antisense strand. Splicing out of the 3rd intron sequence was observed to produce the largest difference of stem probabilities on the peripheral exonic region among all of 12 introns in *NEXN* gene (Figure 44).

Figure 44 shows the gene that has the most post-accessible SS in the cytoskeleton cluster according to the DAVID analysis. This *NEXN* gene (NM_144573) encodes nexilin, which is a filamentous actin-binding protein that functions in cell adhesion and migration. It has 12 SSs and several alternative splicing patterns, such as exon skipping at the 3rd, 6th, and 11th exons (Figure 44(b)) [75]. The analysis of stem probabilities suggests a large increase of accessibility through splicing at the 3rd SS (Figure 44(b)). Figure 44(c) shows the secondary structures around the third and fourth exons of the *NEXN* gene, where I have depicted only the credible base pairs with probability ≥ 0.5. Before splicing, both the donor and acceptor sites form stems with intronic bases, while they are unpaired in the spliced mRNA structure. It is possible that the strong stems between exonic and intronic regions around the 3rd SS have important roles in regulating the observed AS patterns.

Difference of structure propensity between sense and antisense sequence

I suggested the structure propensity of intronic regions differ from each other in sense and antisense strand in the previous section. That tendency was shown only after the normalization by the k-mer composition. In this section, I thus examined the difference of structure propensity between sense and antisense sequence about intact stem probability, partition function, and γ-centroid structure computed by energy models.

To extract the structure propensity bias of the energy model, I produced random sequences of 200-nt long with the specific GC or GU content, and then computed the partition function of them and their complementary sequences. As a result, the partition function of random sequences showed few deviation between sense and its antisense sequence, and the relationships of the partition function with varying GC content is almost linear (Figure 45(a)). On the other hand, the biased GU content produced the non-linear relationships between the partition function of sense and antisense sequences (Figure 45(c)). This is supposed to be mainly caused by the disappearance of G-U base pairs in the complementary sequence. In addition, random sequences having the same GU content still showed the deviation of the partition function, indicating the existence of more complex biases in both of Turner and Andronescu energy model.

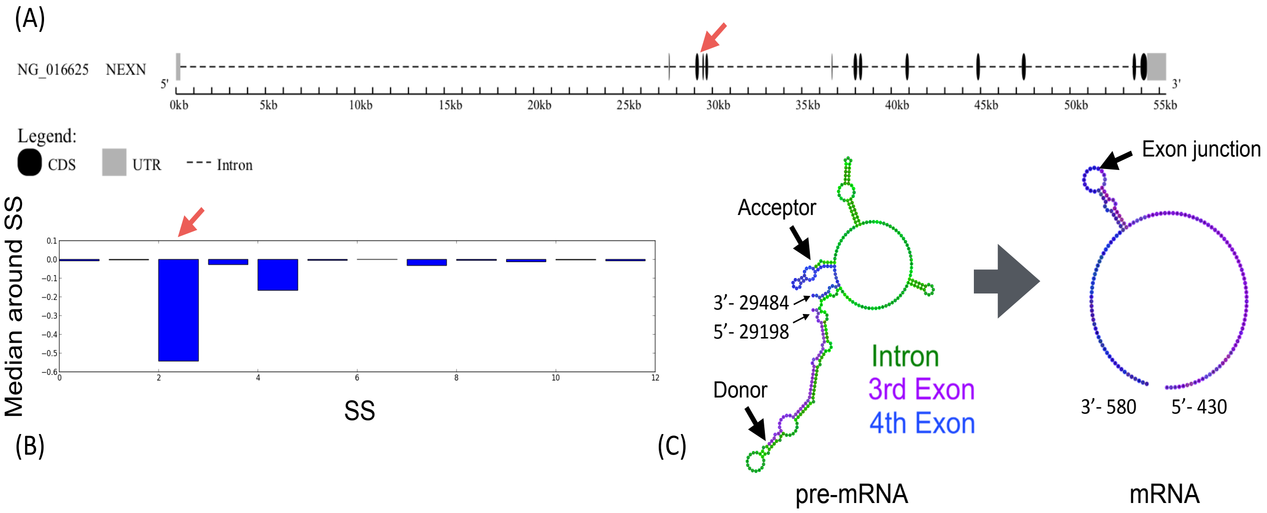
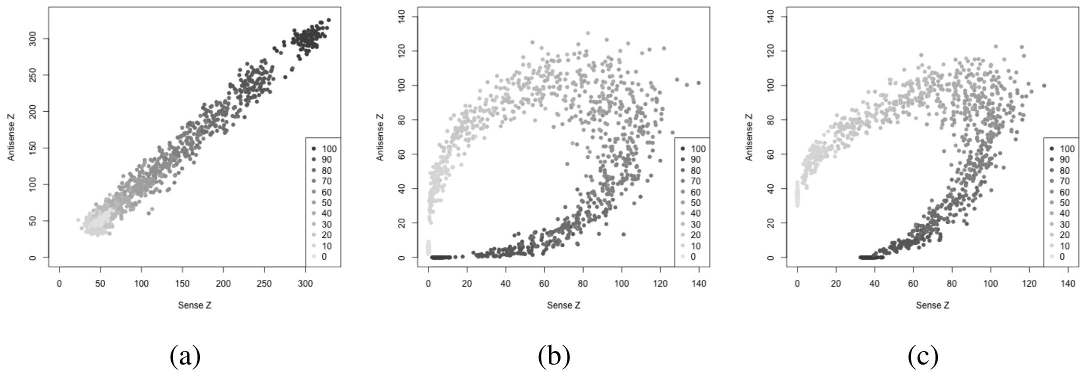


Figure 44. Gene structure and conformational change of the *NEXN* gene. (a) Gene structure of the *NEXN* gene. The GSDS tool has been used for visualization [76]. (b) Median difference of stem probability around each SS. The third SS shown with a red arrow is the most post-accessible SS among the genes in the cytoskeleton cluster. (c) Partial γ-centroid structure of pre-mRNA and mRNA (base pairs whose probability $\geq$ 0.5) around the third SS in the *NEXN* gene. For visualization, I only extracted the region of the substructure enclosed by the outermost pair or exterior loops around the 3rd intron in pre-mRNA and the 3rd exon junction in mRNA.



(a)

(b)

(c)

Figure 45. Structural bias between sense and antisense strand. Partition function of sense and antisense random sequences of 200-nt long using different energy models. To produce random sequences, GC content (Left) and GU content (Center and Right) were altered from 0 to 100 % (shown in legend).

Next, the comparison of stem probability was applied to *NEXN* gene (NM_144573) (Figure 44). To examine the relationship of stem probability between sense and antisense strand, I calculated the difference of stem probability ($p_{\text{stem, mRNA}} - p_{\text{stem, pre-mRNA}}$) for each 32-nt window of sense and reversed antisense sequences. Figure 46 shows the average difference of stem probability and GU content for each window on mRNA and pre-mRNA sequence. The average difference of stem probability is clearly proportional to the GU content of the window in the sense strand. Also, pre-mRNA is longer than mRNA so that it contains more biased positions in terms of the difference of stem probability as well as GU content.

In Figure 47, the partial γ-centroid structures were extracted around the 3rd intron region of *NEXN* gene in sense and antisense strand (visualized using `http://biojs.io/d/drawrnajs`). γ-centroid structures were originally predicted for the whole sequences, but only the partial

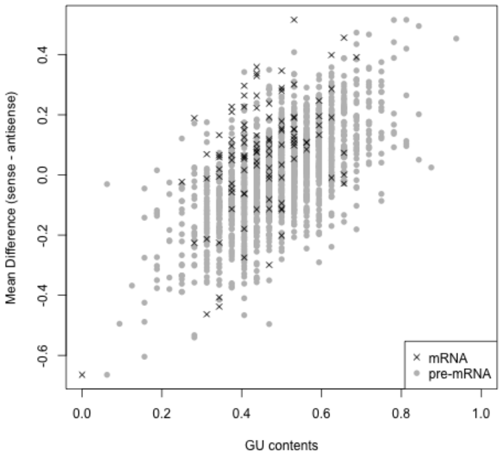


Figure 46. Structural bias of mRNA and pre-mRNA between sense and antisense strand. Mean difference of p_{stem} between sense and antisense structure of NEXN gene (NM_144573) for each window. X-axis represents GU content of each 32-nt window. Y-axis shows the mean difference of stem probability ($p_{\text{stem,sense}} - p_{\text{stem,antisense}}$) of each single base within the window.

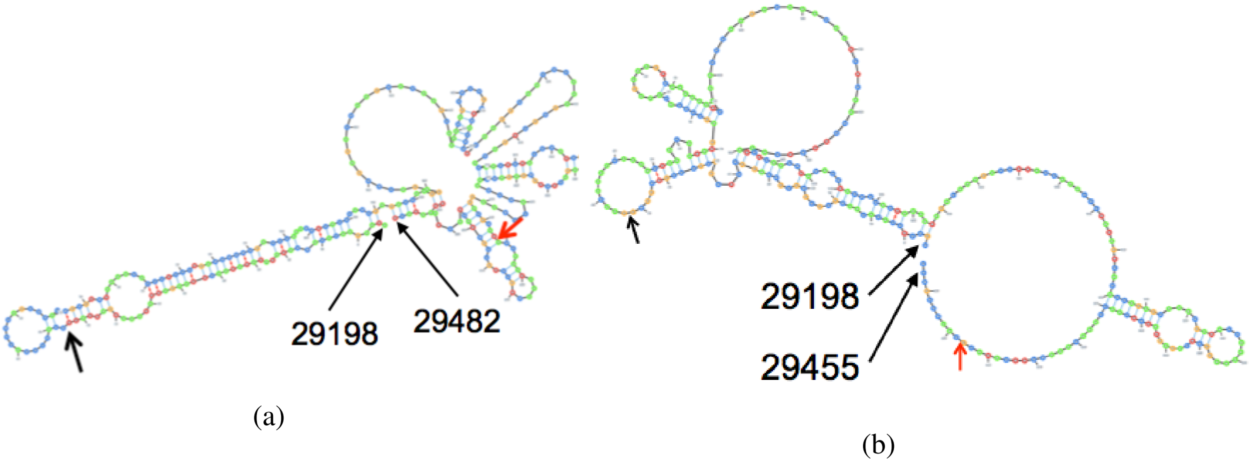


Figure 47. Predicted structure of NEXN in sense and antisense strand. Partial γ-centroid structure of sense (a) and antisense (b) pre-mRNA of NEXN gene (NM_144573) around 3rd intron (29244-29445). Black and red arrows correspond to the donor and acceptor site in sense strand, respectively. For comparison, I reversed the antisense sequence and structure. Hence, each position in antisense strand indicates the position of the complementary base to that of sense strand.

structures around the 3rd intron were extracted for visualization not to have any base pair which has a probability higher than 0.5 with the exterior sequences. Although the longest stem loop is roughly conserved, 2 of 3 stem loops in the right side of the sense structure disappeared from the multi-loop in the antisense structure. The length of stem loop in the left side also became shorter in the antisense structure, resulting the long multi-loop was newly formed in it.

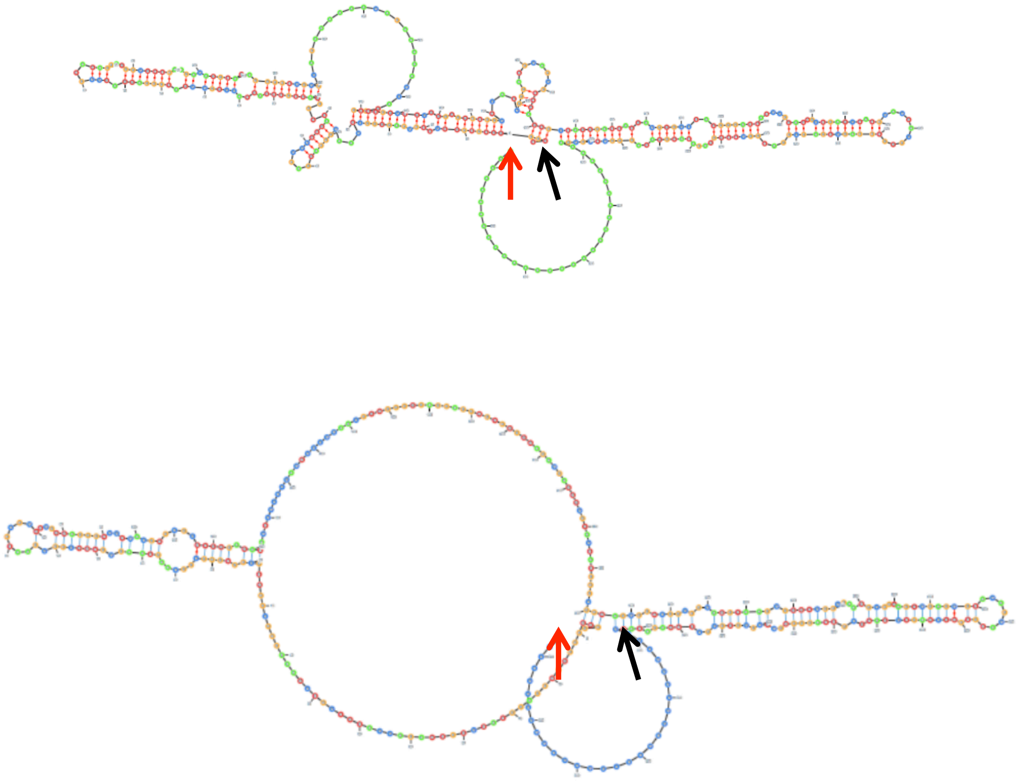


Figure 48. Predicted structure of an Alu sequence in sense and antisense strand.γ-centroid structure of sense (Top) and antisense (Bottom) consensus sequence of the Alu-Jo subfamily. Black and red arrows correspond to the start and end point in sense strand, respectively. As in Figure 47, I reversed the antisense sequence and structure.

Furthermore, the consensus sequence of human Alu repeat, shown as below, was extracted from RepBase [77] for γ-centroid structure prediction.

- Alu-Jo subfamily
  GGCCGGGCGCGGUGGCUCACGCCUGUAAUCCCAGCACUUU
  GGGAGGCCGAGGCGGGAGGAUUGCUUGAGCCCAGGAGUUC
  GAGACCAGCCUGGGCAACAUAGCGAGACCCCGUCUCUACA
  AAAAAUACAAAAAUUAGCCGGGCGUGGUGGCGCGCGCCUG
  UAGUCCCAGCUACUCGGGAGGCUGAGGCAGGAGGAUCGCU
  UGAGCCCAGGAGUUCGAGGCUGCAGUGAGCUAUGAUCGCG
  CCACUGCACUCCAGCCUGGGCGACAGAGCGAGACCCUGUC
  UCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Figure 48 shows the γ-centroid structure ($\gamma = 1$) of the Alu-Jo subfamily sequence in sense and antisense strand. Although the structure of the sense sequence shows two arms of stem loop, antisense one possesses long multi loop as well. Therefore, it is concluded that the predicted structure and structure propensity of sense and antisense sequences possibly show the critical difference depending on their sequence composition.

Table 6. Enriched GO terms among post-accessible genes in human by Ontologizer

GO	term	*p*-value
GO:0044422 (C)	organelle part	0.0011
GO:0044428 (C)	nuclear part	0.0025
GO:0006996 (P)	organelle organization	0.0104
GO:0007059 (P)	chromosome segregation	0.0104
GO:0007049 (P)	cell cycle	0.0104
GO:0003723 (F)	RNA binding	0.0104
GO:0071840 (P)	cellular component organization or biogenesis	0.0104
GO:0032991 (C)	macromolecular complex	0.0104
GO:0051246 (P)	regulation of protein metabolic process	0.0104
GO:0000723 (P)	telomere maintenance	0.0104
GO:1902589 (P)	single-organism organelle organization	0.0113
GO:0031974 (C)	membrane-enclosed lumen	0.0120
GO:0051301 (P)	cell division	0.0160
GO:0036094 (F)	small molecule binding	0.0175
GO:0032268 (P)	regulation of cellular protein metabolic process	0.0216
GO:0033044 (P)	regulation of chromosome organization	0.0216
GO:0005515 (F)	protein binding	0.0301
GO:0032878 (P)	regulation of establishment or maintenance of cell polarity	0.0318
GO:0045595 (P)	regulation of cell differentiation	0.0382
GO:0033043 (P)	regulation of organelle organization	0.0403
GO:0043226 (C)	organelle	0.0403
GO:0043228 (C)	non-membrane-bounded organelle	0.0404
GO:0044427 (C)	chromosomal part	0.0439

Table 7. Top 3 enriched GO terms in post-accessible and post-structural genes.

gene set	EASE score	Keyword and GO term
human	16.2	Cytoskeleton
post-	11.9	Kinase, ATP-binding
accessible	9.2	Centrosome
human	8.2	Serine/threonine protein kinase
post-	5.5	C2 domain
structural	4.3	VWFA domain
mouse	14.6	Cytoskeleton
post-	13.1	ATP-binding
accessible	7.3	Centrosome
mouse	16.6	Kinase, ATP-binding
post-	6.1	Serine/threonine protein kinase
structural	4.2	C2 domain

4 DISCUSSION AND CONCLUSIONS

COMPARISON BETWEEN COMPUTATIONAL STRUCTURE PREDICTION AND EXPERIMENTAL STRUCTURAL ANALYSES

The stem probabilities computed for human mRNAs agreed well with a large-scale experimental structural analysis in terms of both global characteristics (Figures 19(c) and 23(a)) and statistical correlations between the scores (Figure 19(b)). Although there are many reasons why computational folding fails to predict true secondary structures, most disagreement with experimental analyses currently seems to result from insufficient read depths. Thus, I expect more agreement with experiments as sequencing coverage continues to increase. I have also shown that the concordance of the prediction tool can be significantly increased by selecting appropriate averaging-window sizes or energy parameters that are suited to the experimental design and other conditions. It may even be possible to study the effects of pseudoknots or 3D structures by looking at the differences between computational predictions and experimental data that cannot be eliminated by such optimization.

THE INFLUENCE OF SEQUENCE BIASES ON THE ANALYSIS OF STRUCTURAL CONSTRAINTS

A genome-wide comparison of thermodynamic structure stability would clarify the kinds of structural selection that act on the target regions. Simultaneously, such an analysis needs to employ an appropriate normalization scheme to eliminate primary composition biases. For example, the high GC content of CDS regions may lead to erroneous significance of selection pressure toward stable structures over the entire CDS regions, because the stability of RNA secondary structure is generally correlated with the GC content of a target sequence. Although there are several proposed methods to normalize such sequence biases (e.g., shuffling at a 4-fold degenerate site or preserving di-codon counts) [72, 7], they are mostly CDS-specific and could not be used in the present analysis. Intronic sequences are also known to possess several sequence biases, including the asymmetry of A/T and G/C around SSs [78], which does not cancel out by simply normalizing GC contents. As there is no perfect method to accomplish the normalization, I have taken a very conservative approach; I masked repeat regions, removed the contribution of k-mer frequency bias using a regression model, and compared the regions of interest with the Intergenic and antisense sequences. I have shown that the antisense sequences of introns are more different from the sense sequences than Intergenic regions in terms of structural propensity, which implies that the analyses that use only antisense sequences as background would overestimate the selection pressure. To complement this elaborate normalization approach, I have also carried out direct comparison between the same regions of mRNA and pre-mRNA to evaluate structure propensity inside exons, as they trivially do not possess any difference in the sequence composition bias.

I determined that the stem density within CDS regions is better predicted by their sequence compositions (Figures 33(b) and (c)) than are the stem densities of other regions, while introns and 3′-UTRs contain a significantly larger number of regions with higher stem densities than expected (Figure 39). The strand-asymmetric preference for higher stem density persisted over entire intronic regions (Figure 40), which cannot merely be explained by k-mer compositions or strand-symmetric sequence features. Such asymmetric preference is possible due to the strand-asymmetric pairing rules of the Turner energy model and other asymmetric sequence characteristics; G-U base pairing is not conserved in the complementary sequence and the appearance of different pairing patterns in loop regions such as the change from AAA to UUU. In Results section, I investigated the strand asymmetry of partition function and stem probabilities. They show a significant correlation between the strand asymmetry and the "GU" content of sequence (Figures 45 and 46). I have also shown two examples in Figure 47 and 48, in which strong stems in the sense strand are destabilized and decomposed into multi-loops in the antisense strand. The differences of folding energies between sense and antisense strands are also studied in Ref. [17].

As described previously, a significant structural preference can be caused by only a small portion of the transcribed regions, owing to the large degrees of freedom of the hypothesis tests. Therefore, my results suggest that the entire intronic regions are dispersed with small intronic elements that tend to be more structured than Intergenic and antisense sequences, but this does not necessarily mean that the entire regions of introns are highly structured. It should also be noted that I did not investigate the raw stem probabilities but the residual structural preferences remained after removing the sequence bias using linear regression. Therefore, the obvious correlation between stem probability and local GC content is normalized before the main analysis. Thus, the significant p-values supposedly reflect the intrinsic structural preferences beyond the obvious correlation between stem probability and local GC content.

One important future goal will be determining whether the known asymmetric mutation patterns in intronic regions [78] can explain this asymmetric structural preference. It will also be interesting to study various biological causes of the higher stem density within introns. It may prevent stalling of PolII (as in translation [79]), help splicing by shortening the physical distance between the donor and acceptor sites, or prohibit the splicing machine from accessing wrong acceptor sites.

A direct comparison of stem probabilities between mRNA and pre-mRNA showed a clear reduction in stem density around the SSs (Figure 41). My analyses indicate that this reduction is significantly correlated with the strength of gene expression. Together with the observation that SSs exhibit a strong structural preference (Figure 40), these findings suggest that gene expression is mediated by the efficient use of secondary structures that disappear after pre-mRNA splicing.

CONCLUSIONS

Using my novel software "ParasoR" and k-mer regression method, I extracted structure profiles of human transcripts and inferred the genome-wide structure propensity beyond sequence composition biases. The structure profiles predicted by ParasoR showed a high concordance with Rfam structures and high-throughput sequencing analyses. A genome-wide simulation using ParasoR indicated that a structure propensity of transcribed regions is strongly regressed by k-mer composition. By focusing on the residual part of such regression, intronic regions were shown to contain a significantly higher rate of structured regions compared to antisense and intergenic regions, not only around the ends of introns but also throughout entire regions. Furthermore, a comparison between pre-mRNAs and mRNAs suggested that coding regions become more accessible after splicing, presumably because of biological constraints such as translational efficiency.

Part II

Statistical approach to robust RNA reactivity classification based on reproducible high-throughput structure analyses

1 BACKGROUND

GENOME-WIDE RNA SECONDARY STRUCTURE DETERMINATION BY COMPUTATIONAL AND EXPERIMENTAL ANALYSES

RNA secondary structure consists of the combination of canonical base pairs. It plays an important role fin various biological processes as a major contributor the stabilization of interaction with other molecules [9, 80, 81, 82]. Previous studies have revealed that the structure of non-coding RNAs, such as tRNA, rRNA, or snoRNAs, has been evolutionarily conserved, although primary sequences have been less so [14]. Rfam is a database of such conserved structure motifs, which lists 2,474 evolutionary-conserved structure motifs as of April 2016, using structure and sequence similarity as based on a computational algorithm [64]. Further analyses have also shown that secondary structure is influential not just on non-coding RNAs but also on coding RNAs, such as mRNA and pre-mRNA by adjusting the binding of regulators required for the translation, degradation, and localization of RNAs [36] However, crystal structures or other RNA structures determined in extremely different environments from the living cell have been used as a training set for structure prediction. In addition, there would be an intervention of other molecules, such as interaction of RNA binding proteins under *in vivo* conditions. Therefore, there is no guarantee that RNA forms can predict structures *in vivo* at present.

However, conventional experimental analyses have serious limitations in throughput and computational time compared to their computational counterparts. Experimental structure analyses have started with the detection of changes in absorption spectra at different temperatures, indicating the existence of RNA secondary structure [3, 4]. The use of the difference of accessibility between single-stranded and double-stranded regions, the introduction of base modifications specific to accessible regions, or RNA footprinting has been applied for structure analyses. For instance, SHAPE technology [83] uses N-methylisatoic anhydride (NMIA) to induce selective acylation of the ribose 2′-hydroxyl position of flexible nucleotides. Since the primer extension is likely to stop at these acylated nucleotides, the 5′ ends of the chemical-treated RNA fragments must be enriched in single- rather than double-stranded regions. In this way, the accessibility or reactivity profile at each nucleotide can be estimated by the amount of RNA fragments of each length, separated by gel or capillary electrophoresis. This technique has elucidated the structural characteristics of long RNAs, such as rRNAs [84] and the HIV RNA genome [85], combined with capillary electrophoresis. It is, however, barely possible to distinguish base reactivities at distant regions from 3′ while also performing absolute quantification for each nucleotide. In addition, the concentration of target RNA is required to specify which position is indicated by the enrichment of each fragment. For these reasons, it remains difficult to clarify the entire landscape of the RNA secondary structure of human transcriptome, which contains tens of thousands of transcripts.

To solve the scalability problem encompassed in the conventional structure analyses, high-throughput structure analysis methods have been developed to infer the global landscape of RNA secondary structure, which is referred to as RNA structurome [18]. After the development of early methods such as PARS [19], FragSeq [86], and SHAPE-Seq [87], more than a dozen research studies have been carried out for the development of high-throughput structure analyses [18]. The first high-throughput analysis of the PARS method showed that mRNA exhibits specific structure profiles, such as enrichment of base pairs in coding regions, a three-nucleotide periodicity of reactivity in the coding region, and anti-correlation between the efficiency of mRNA translation and structures around the start codon [19]. Additionally, the comparison of high-throughput structure analyses revealed the landscape of RNA secondary structure alteration. In Ref. [20], SNVs that alter RNA structure, known as riboSNitches, were shown to affect the binding efficiency of RBPs such as AGO or LIN28, resulting in the abnormality of post-transcriptional regulation. A comparison of *in vivo* and *in vitro* structure analyses, based on DMS-Seq [22] and icSHAPE [88], also indicated the characteristic tendency of RNA secondary structure that RNA tends to be more accessible *in vivo*, which is compared to *in vitro* supposedly due to the interaction of RBPs and the base modification of epitranscriptome regulation [89]. Accordingly, the integration of high-throughput structure analyses is expected to allow us to further our progress toward an understanding of the disease mechanism caused by the dysfunction of RBP and non-coding RNA.

Difficulty in computational modeling of high-throughput structure analyses

While high-throughput structure analyses can be practical, due to their feasibility of comprehensive RNA identification, the reactivity scores estimated by such analysis tend to be inconsistent with each other (described in the Results section). Because each performs different processes during library preparation and sequencing, estimation of reactivity at each nucleotide might suffer from the influence of various systematic biases specific to each methodology [90]. Hence, such biases may cause overestimation of inconsistency among the cell types or conditions in an integrative comparison of the studies and methods.

As a cause of systematic biases, there are four major differences between typical methods of high-throughput structure analysis (shown in Table 8), which potentially also contribute to the advantages of individual methods. The first is the type of probing, which causes modification or cleavage according to the accessibility of each base. The latter is brought about by a certain nuclease and hydroxyl radical reaction at either structured or accessible nucleotides [19, 91, 92, 93]. Second, the applicability of each reagent is also divergent, for example, dymethyl sulfate (DMS), 1-methyl-7-nitroisatoic anhydride (1M7), and 2-methylnicotinic acid imidazolide-N_3 (NAI-N_3) is a cell-permeable reagent that can observe conformational alterations in various conditions including *in vivo* [94, 22, 95, 96]. Third, the detection process of probed bases has two types: 3′-end enrichment and mismatch enrichment. In general high-throughput structure analyses, the location of probed flexible bases is detected by 3′-end coverage of mapped reads, indicating that each single read corresponds to the single base. That said, a newer technique, mutational profiling (MaP) methodology, detects the reactive regions by noncomplementary nucleotides induced by misreading at the sites of RNA adducts during cDNA synthesis. This method has the advantage that it can infer structural cooperation in multiple structures by cooccurrences of mutation. However, there remain few practical studies about MaP that fulfill the requirement of optimal condition and reverse transcriptase search, in addition to the difficulty of read alignments with a large number of mismatches. The difference in base preferences may also pose a severe problem for the cross-comparison, as structure analyses that use certain specific reagents, such as DMS and CMCT, can modify only partial types of flexible nucleotides (though there is another hypothesis that the probing reaction of DMS occurs substantially at all bases [97]). Besides these differences, each reagent and protocol has a variety of minor dissimilarities such as sequence preferences or the range of the detectable size of RNAs.

However, bioinformatics and statistical techniques for analyzing those datasets beyond these systematic biases have rarely been discussed [102], in spite of many studies on computational structure prediction assisted by structure-probing information [103]. High-throughput structure analyses are considered more susceptible to the sparseness of sequencing reads than RNA-Seq and ChIP-Seq, because reactivity estimation should be carried out for each base position in contrast with sequenced reads that are aggregated for each transcript in RNA-Seq and for each peak around binding sites in ChIP-Seq [104]. Nevertheless, to filter out unreliable regions with low read coverage, most previous researchers set arbitrary formulae or thresholds for reactivity estimation, instead of statistical or probabilistic modeling [19, 98, 88]. To my knowledge, there are a few processing methods based on a statistical or probabilistic framework to obtain reactivity scores without the influence of sparseness [105, 94, 93, 106, 107, 102]. The consistency of reactivity between replicates is assessed in only two of these methods, one being the Cochran-Mantel-Haenszel tests in Mod-seeker [106] and the other being the BUM-HMM model [102]. In particular, BUM-HMM is the first application based on a nonparametric statistical model, and has been shown to increase sensitivity compared to previous pipelines. Yet, both of the above can use neither the consensus information between reagent-treated datasets, nor a dataset with no “control” condition (e.g. PARS and ds/ssRNA-seq compare the outputs in which single- and double-stranded regions must be enriched). In addition, the accuracy of existing methods is not sufficient compared to the existing prediction algorithm, particularly for the prediction of base pairs, as most of the reagents are applicable for the probing of flexible bases.

Table 8. List of high-throughput structure analyses based on sequencing and probing

	Name	Reagent	Base specificity	Detection	Publication
	PARS	nuclease S1 and V1		Cleavage	[19]
	PARTE	RNase V1		Cleavage	[91]
	ds/ssRNA-seq	RNase ONE and V1		Cleavage	[92]
	FragSeq	Nuclease P1		Cleavage	[86]
	SHAPE-Seq	1M7		RT drop-off	[87]
	SHAPE-MaP	1M7,1M6,NMIA		Mutation	[98]
	icSHAPE	NAI-N_3		RT drop-off	[96]
	Map-seq	DMS,CMCT,1M7	(A/C and G/U)	RT drop-off	[95]
	RING-MaP	DMS	A/C	Mutation	[99]
	HRF	Hydroxyl radical		Cleavage	[93]
	DMS-seq	DMS	A/C	RT drop-off	[22]
	DMS-MaPseq	DMS	A/C	Mutation	[100]
	Structure-seq	DMS	A/C	RT drop-off	[94]
	ChemModSeq	DMS,1M7	(A/C)	RT drop-off	[101]
	CIRS-seq	DMS,CMCT	A/C and G/U	Mutation	[23]

DMS, 1M7, 1-methyl-6-nitroisatoic anhydride (1M6), and NAI-N_3 reagents can be applied for *in vivo* structure analyses.

REACTIDR: STATISTICAL APPROACH TO ROBUST RNA REACTIVITY CLASSIFICATION BASED ON REPRODUCIBILITY

In this study, I present a novel pipeline, *reactIDR*, which is designed to extract reliable structure information from general high-throughput structure analyses for robust inference of an RNA secondary structure landscape. To evaluate the reliability of each reactivity score, the irreproducible discovery rate (IDR) is computed by modeling the joint probability distribution among replicates [108]. reactIDR can also estimate locally consistent IDR based on the hidden Markov model (HMM), as well as p-values assuming Poisson or negative binomial distribution as a null distribution of total reads across each transcript. The efficiency of IDR filtering and classification for reproducible structure prediction was evaluated by comparing with the reference structure of 18S rRNA and computational prediction of a whole transcriptome. According to the results, IDR-based classification showed higher consistency with the reference structure and stem probability as calculated by ParasoR, indicating that reactIDR would be a significant assist in extracting the condition-specific difference of secondary structure, with a view to deciphering the global view of RNA secondary structure.

2 METHODS

2.1 REACTIDR

A novel pipeline, reactIDR (developed in https://github.com/carushi/reactIDR), is designed to explore optimal reactivity classification from high-throughput structure analyses considering reproduciblity. reactIDR can evaluate scoring schemes of reactivity by comparing with a reference structure, computational prediction, and other high-throughput structure analyses. Figure 49 shows an overall workflow of reactIDR developed for a general high-throughput structure dataset. This pipeline can run on bed files converted from sam or bam files of mapping results.

The evaluation of high-throughput structure analyses is performed by reactIDR in the following three steps:

1) evaluate the irreproducibility of sequencing data to extract reproducible regions,

2) calculate various structural features of reactivity and read coverage profiles,

3) construct reactivity classifiers based on the selected feature set.

To evaluate irreproducibility, reactIDR computes IDR for each position based on replicate consensus, or HMM-based IDR as another option to take local consistency into account. Based on the IDR computed for raw read coverage or p-values, assuming null hypothesis regarding mapped reads, uncertain areas of the low read coverage can be excluded. reactIDR can explore a wide range of feature sets, such as read coverage profiles and reactivity scores defined in PARS [9] and icSHAPE [88] to apply supervised classification algorithms. Users can also select a type of null distribution and give optional information such as the sequencing depth for score normalization. By fitting of machine learning classifiers for the set of reactivity scores, the optimal feature set is explored for the structure classification of a specific dataset and reference structure. At this time, filtering out unreliable regions would improve the robustness of structure prediction [103].

In the following subsections, I describe the details of IDR estimation, data processing, scoring schemes, and classifier construction.

2.1.1 *Data structure and notation for reactIDR*

In this subsection, I describe data processing, including reactivity scoring and read count normalization using a hmostigh-throughput structure analyses dataset. Most high-throughput structure analyses are designed to measure the flexibility of each nucleotide by observing the coverage of $5'$-end of mapped reads. This is because the base modification induced at flexible nucleotides increases the drop-off ratio of RT on the spot. Hence, the coverage of $5'$ at 1-base downstream is measured as the indicator of reactivity. Cleavage-based structure analyses also measure the same coverage as the strength of single- or double-strandedness, depending

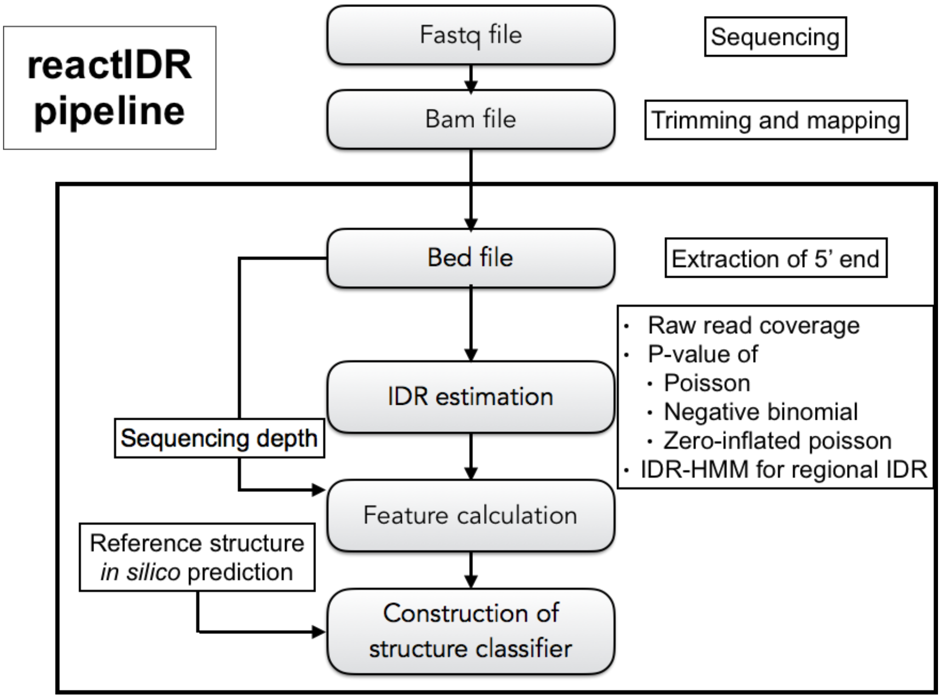


Figure 49. Workflow of reactIDR pipeline. In reactIDR, the input bam files are converted to bed files. For read coverage information at each nucleotide of each transcript, IDR is estimated based on raw read coverage or p-value computed from null distribution. HMM-based IDR can also be calculated to consider local consistency of read coverage. In addition to IDR, many features, including the scores used in the previous studies, are used as a feature for the structure classification, in which the weight of each feature and accuracy is computed.

on treated nuclease. Hence, the input of reactIDR is a read count of the ith nucleotide of the transcript t for the condition, represented as $C_{\text{method\_condition}}(i,t)$. Although MaP analyses deploy another type of read information, meaning that the number of mismatches is considered to be a reactivity, it can also be applicable for IDR estimation in reactIDR if the data on the number of noncomplementary bases are appropriately formated. Note that reactIDR cannot deal with the covariance information between the distant bases.

PARS data consist of two kinds of sequenced samples; one treated with S1 nuclease to measure accessibility, and another treated with RNase V1 to measure the structure stability at each position [9]. To calculate the PARS score, C is first normalized with regards to the number of total mapped reads for each sample (represented as $c_{\text{PARS\_condition}}(i,t)$) [9]. At first, PARS scores (represented as r_{PARS} in this study) were calculated as below:

$$r_{\text{PARS}}(i,t)' = \log_2\left(\frac{c_{\text{PARS\_V1}}(i,t)+1}{c_{\text{PARS\_S1}}(i,t)+1}\right),$$

where $c_{\text{PARS\_V1}}(i,t)$ and $c_{\text{PARS\_S1}}(i,t)$ correspond to the normalized read coverage of a V1 (S1) dataset at the ith nucleotide (detailed in Ref. [9]). The PARS score is then capped to ±7 for scaling. In Ref. [20], however, the PARS score is redefined as a five-base average of the previous PARS scores, as below:

$$r_{\mathrm{PARS}}(i,t) = \log_2\left(\sum_{j=i-2}^{i+2}\frac{c_{\mathrm{PARS\_V1}}(j,t)+5}{5}\right) - \log_2\left(\sum_{j=i-2}^{i+2}\frac{c_{\mathrm{PARS\_S1}}(j,t)+5}{5}\right)$$

In this study, the latter PARS score r_{PARS} was computed using the mean of $c_{\mathrm{PARS\_V1}}$ and $c_{\mathrm{PARS\_S1}}$ to compute a single score from replicates. This is because such averaging can buffer local ambiguity of read mapping in exchange for the detailed resolution. Actually, the consistency of r_{PARS} with computational prediction was already demonstrated in ParasoR research [109].

Another method, icSHAPE [88, 96], detects flexible nucleotides using NAI-N$_3$ reagent by observing the enrichment of RT drop-off compared to the control DMSO sample. In [96], a reactivity score $r_{\mathrm{icSHAPE\_condition}}(i,t)$ is defined as below.

$$\begin{aligned}&r_{\mathrm{icSHAPE\_condition}}(i,t) =\\ &\quad(c_{\mathrm{icSHAPE\_condition}}(i,t) - \alpha c_{\mathrm{icSHAPE\_DMSO}}(i,t))\,/\,c_{\mathrm{icSHAPE\_background,\ DMSO}}(i,t),\end{aligned}$$

where $c_{\mathrm{icSHAPE\_condition}}$ is a normalized read count, $c_{\mathrm{icSHAPE\_background,\ DMSO}}(i,t)$ is the number of read counts that read through the ith nucleotide, and α is a parameter to adjust the strength of background control and set it to 0.25 in this study, as optimized in the previous study [88]. $c_{\mathrm{icSHAPE\_condition}}$ and $c_{\mathrm{icSHAPE\_DMSO}}$ is the normalized read coverage, or the ratio of RT stop read counts against the whole sequencing library. At this time, $c_{\mathrm{icSHAPE\_condition}}$ and $c_{\mathrm{icSHAPE\_DMSO}}$ are scaled so that the mean of 90-95 % most reactive bases across the library are 1. Then, the top 5-95 % of $r_{\mathrm{icSHAPE\_condition}}$ is scaled to the range of $[0,1]$ for each transcript independently by 90 % Winsorization (the bottom and top 5 % bases are set to 0 and 1, respectively). This 90 % Winsorization procedure is performed for all positions without 32 bases at the 5′ and 3′ ends of the transcript, due to the difficulties in appropriate mapping. When $c_{\mathrm{icSHAPE\_background,\ DMSO}}(i,t)$, which does not include cDNA reads that stop at the ith nucleotide, is 0, $r_{\mathrm{icSHAPE\_condition}}(i,t)$ is excluded from the dataset. In addition, due to the bias of read mapping, $r_{\mathrm{icSHAPE\_condition}}$ at the very 5′- and 3′-ends are also excluded. In this study, $c_{\mathrm{icSHAPE\_background,\ DMSO}}$ was simplified as read counts of RT stop at each position.

2.1.2 *IDR: irreproducible discovery rate for high-throughput experiments*

IDR was first developed in Ref. [108] to discover a true signal of protein-binding sites in ChIP-Seq analyses beyond the problem of irreproducible read sampling. To estimate IDR, the mixture copula model is assumed, in which one copula explains the distribution of irreproducible signals, and others do so for true signals with higher correlations of random variables among replicates.

Assuming a two-dimensional mixture copula of true and spurious signals, the dependency between replicates for both signals is involved in a bivariate Gaussian distribution. An indicator K_i, following Bernoulli(π_1), indicates that the ith signal is produced by true (spurious) signals when $K_i = 1$ ($K_i = 0$), where π_i is a proportion of true signals. The distribution of a random variable $z_i = (z_{i,1}, z_{i,2})$ is as follows:

$$\begin{pmatrix} z_{i,1} \\ z_{i,2} \end{pmatrix} \mid K_i = k \sim \mathcal{N}\left(\begin{pmatrix} \mu_k \\ \mu_k \end{pmatrix}, \begin{pmatrix} \sigma_k^2 & \rho_k\sigma_k^2 \\ \rho_k\sigma_k^2 & \sigma_k^2 \end{pmatrix}\right), k = 0, 1, \tag{7}$$

where μ_k, σ_k, ρ_k is the mean, variance, and correlation between replicates. Since the distribution of spurious signals ($k = 0$) should be located around 0 (low coverage regions) with a large variance, it is assumed that $\mu_0 = 0,\ \mu_1 > 0,\ \sigma_0^2 = 1,\ \rho_0 = 0$, and $0 < \rho_1 \geq 1$. In Equation 7, ρ corresponds to the strength of the dependency between replicates. These parameters are then fitted to the ranks of ChIP-Seq peak data based on the pseudo-likelihood method, so that those ranks are derived from the cumulative distribution function of $z_{i,1}$ and $z_{i,2}$. In this way, IDR that is a probability of each peak derived from the irreproducible signal is estimated for the ranks of ChIP-Seq peaks in the order of read coverage.

Although the subject of high-throughput structure analyses is different from ChIP-Seq, a similar tendency, such as irreproducible read counts observed in low coverage regions, is observed due to the systematic bias of sequencing and library preparation. In addition, IDR has the advantage that it can be estimated on the space of joint distribution of random variables that should be only marginally uniformly distributed between $(0, 1)$ for the copula. This means that IDR evaluation is available for a variety of score distributions, such as read coverages, the ratios of control and case samples, and their p-values, even though the information on the magnitude of score ranges is discarded. Hence, reactIDR is considered to have a high capacity to perform robust analyses of high-throughput structure analyses datasets.

2.1.3 *p-value computation based on null distribution*

reactIDR can handle IDR estimation for p-values based on null distribution instead of raw read coverage. The p-values computation can be carried out for three types of distributions in reactIDR; Poisson, negative binomial, and zero-inflated Poisson with parameter optimization based on Markov chain Monte Carlo (MCMC) sampling. This conversion has the potential to detect read enrichments considering the different gene expression levels. These distributions are widely applied in the existing count-based normalization methods to model the read count distribution of individual transcripts in RNA-Seq [110, 111, 112]. This is mainly because the process of selecting sequenced regions can be modeled by those distributions if the selection for the sequenced region is truly random. However, since fitting of the null distribution is required for each transcript in structure analyses, it is not practical to apply MCMC sampling for read count distributions for individual transcripts due to the long computational time. In addition, a mixture of true and spurious signals is expected to appear in such datasets so that there is a possibility that the parameter estimation of null distribution is affected if the large amount of reads counts belongs to true signals. Therefore, I further implemented the modeling of Poisson and negative binomial distribution inferred by the partial datasets which are plausible to be false signals. The detail of modeling is defined in CisGenome, an integrated tool for ChIP-Seq [113]. Using this estimation, it is expected to be more robust for a dataset with a larger number of true signals, because selecting only part of low coverage regions is plausible to be only from spurious signals and suitable for parameter estimation of null distribution.

2.1.4 *Feature calculation and scoring*

reactIDR was developed to find an optimal feature set for the cross-comparison of genome-wide structure analyses. The list of features used in reactivity classification is as follows:

Table 9. List of features used in reactIDR

	Features
1.	Stem probability p_{stem} calculated by ParasoR
2.	The sum of read coverages for replicates of both case and control
3.	IDRs for case and control samples
4.	$r_{\mathrm{PARS}}(i,t)$
5.	$r_{\mathrm{icSHAPE\_condition}}(i,t)$
6.	Means and variances of $C(i,t)$ for each condition among replicates
7.	The sizes of surrounding regions in which all bases have raw coverage greater than 25 %, 50 %, or 75 % of $C(i,t)$ (allowing 1-base gap)
8.	$c_{\mathrm{icSHAPE\_condition}}(i,t)$ and $c_{\mathrm{icSHAPE\_DMSO}}(i,t)$
9.	Five-base averages of $c_{\mathrm{icSHAPE\_condition}}(i,t)$ and $c_{\mathrm{icSHAPE\_DMSO}}(i,t)$ for each condition
10.	$c_{\mathrm{icSHAPE\_background,\ DMSO}}$
11.	The maximum raw read coverage of each transcript, which is used for normalization

Machine learning approach to discover optimally comparable scoring of nucleotide reactivity

I constructed the reactivity classifier for the RNA secondary structure based on machine learning approaches. Random forest (RF), kernel support vector machine (KSVM), neural network (NN), multinomial logistic regression (ML), and naïve bayes (NB) classifiers were applied in this study. These classifiers were implemented in R library (RF: randomForest, KSVM: kernlab, NN: nnet with unit size 2, ML: nnet, and NB: e1071, respectively). The accuracy of each classifier was validated by the measures defined below:

$$Accuracy := (TP+TN)/(TP+FP+FN+TN)$$

$$F_1 := 2 \times \frac{TP/(TP+FP)}{TP/(TP+FP)+TP/(TP+FN)}$$

5-fold cross validation was applied to compute these indexes for each classifier, in which samples of each class (base-paired or loop) were equally divided.

2.2 IDR-HMM

2.2.1 *IDR-HMM: Robust secondary structure classification based on localized reproducible signals of high-throughput structure analyses*

In this section, I describe a novel algorithm, IDR-HMM, which is designed to extract reproducible signals from high-throughput structure analyses, considering local consistency of IDR as on the HMM. The latent variable of IDR-HMM corresponds to the status of the RNA secondary structure. Specifically, stem, accessible, and unmapped regions are inferred for each position as a latent variable. IDR-HMM has the potential to extract a larger number of reproducible but low-coverage regions, due to the additional information about local consistency of IDR profiles.

2.2.2 *Notation of dataset*

An input data of IDR-HMM is the set of read coverage (or ratio of read coverage between different conditions) of $5'$ at one-base downstream for each base of each transcript. A high-throughput structure analyses dataset generally contains samples obtained from multiple conditions. Hence, let there be read coverage data of K samples, in which the 1-k_tth (k_t + 1-Kth) samples are nuclease S1 (V1) treated in PARS dataset, and reagent- (DMSO-) treated in icSHAPE dataset. For the ith position of transcript t, the ranking of the mapped read count ut, i, k is scaled into the range $(0, 1)$ across all positions included in the k-th sample. All read count observation is represented by $v_{t,i}$ as follows:

$$\boldsymbol{v_{t,i}} := \{u_{t,i,1}, \ \ldots, \ u_{t,i,K}\}$$

The enrichment of $v_{t,i,i}$ with $i = 1, \ \ldots, \ k_t$, $i = k_t + 1, \ \ldots, \ K$, and no specific enrichment must be observed at the regions that belong to the accessible, stem, and unmapped classes respectively. Therefore, according to the likelihood of being structured or accessible from the mixture of reliable and spurious signals, IDR-HMM classifies each position into class s, where S = {stem, accessible, and unmapped} and $s \in S$.

2.2.3 *Definition of Gaussian mixture copula*

Here, I will describe the emission probability of read coverages $\boldsymbol{u}$ ($u_1, \ \ldots, \ u_K$) at the ith position of transcript t in IDR-HMM. IDR-HMM is based on Gaussian mixture copula for the joint cumulative distribution of random variable X behind the rankdata $v_{t,i}$. Since the joint distributions of 1-k_tth and k_t + 1-Kth samples should be independent, it is also assumed that $v_{t,i}$ includes only the samples of either of the conditions. In the implementation, the parameters are independently optimized for the two types of datasets.

Let $(X_1, X_2, ...X_K)$ be a random vector distributed as multivariate and absolutely continuous distribution with a cumulative distribution function g, and marginal cumulative distribution function F_i. Since any cumulative distribution function is within the range $[0, 1]$, there exists x_i, such that $F_i(x_i) = u_i$. This x_i is referred to as a pseudo-value of u_i.

A copula of $(X_1, X_2, ...X_K)$ is a distribution function defined as below:

$$\begin{aligned} C(u_1, ..., u_K) =& P[U_1 \leq u_1, \ldots, U_K \leq u_K] \\ =& P[F_1^{-1}(U_1) \leq F_1^{-1}(u_1), \ldots, F_K^{-1}(U_K) \leq F_K^{-1}(u_K)] \\ =& P[X_1 \leq F_1^{-1}(u_1), \ldots, X_K \leq F_K^{-1}(u_K)] \\ =& g(x_1, \ldots x_K) \end{aligned}$$

Then, the emission probability of $\boldsymbol{u}$ is obtained by the differential of the copula:

$$\begin{aligned} P(\boldsymbol{u}|\theta) =& \frac{\partial^k C(u_1, \ldots, u_K))}{\partial u_1 \ldots \partial u_K} \\ =& \frac{\partial^K g(F^{-1}(x_1), \ldots F^{-1}(x_K))}{\partial F_1^{-1} \ldots \partial F_K^{-1}} \frac{\partial F_1^{-1}}{\partial u_1} \cdots \frac{\partial F_K^{-1}}{\partial u_K} \\ =& \frac{\partial^K g(F^{-1}(x_1), \ldots F^{-1}(x_K))}{\partial F_1^{-1} \ldots \partial F_K^{-1}} \prod_{k=1}^{K} \frac{1}{f_k(F_k^{-1}(u_k))}, \end{aligned}$$

where f_k is a differentiation of F_k. Among the many functions that satisfy the condition required to be a copula, a Gaussian mixture copula is applied in IDR-HMM due to its comprehensiveness and applicability. Hereafter, the case of $K = 2$ is considered for simplicity, although the copula can handle high-dimensional correlation. Below, two-dimensional Gaussian copula and its copula are defined:

$$
\begin{aligned}
F_k(x_k) &= \int_{\infty}^{x_k} N(x'_k|\mu, \sigma^2, \rho)\ dx'_k \\
f_{(2)}(\boldsymbol{x}, \boldsymbol{\mu}, \sigma^2, \rho) &= N\left(\left(\begin{array}{c} x'_1 \\ x'_2 \end{array}\right)\middle|\left(\begin{array}{c} \mu \\ \mu \end{array}\right), \left(\begin{array}{cc} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{array}\right)\right) \\
&= \frac{1}{2\pi\sigma^2\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \left(\begin{array}{cc} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{array}\right)^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right\} \\
g(x_1, x_2) &= \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(x'_1, x'_2)\ dx'_1 dx'_2,
\end{aligned}
$$

where $\boldsymbol{\mu}$, σ, and ρ are the parameters to fit the copula to the observation $\boldsymbol{u}$. For the case of the mixture copula, the cumulative joint distribution, F_i, g, and the emission probability of $\boldsymbol{u}$ are formulated as below:

$$
\begin{aligned}
F_k(x_k) &= \int_{\infty}^{x_k} qN(x'_k|\mu, \sigma^2) + (1-q)N(x'_1|0, 1)dx'_1 \\
g(x_1, x_2) &= \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} qN\left(\left(\begin{array}{c} x'_1 \\ x'_2 \end{array}\right)\middle|\left(\begin{array}{c} \mu \\ \mu \end{array}\right), \left(\begin{array}{cc} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{array}\right)\right) \\
&\quad + (1-q)N\left(\left(\begin{array}{c} x'_1 \\ x'_2 \end{array}\right)\middle|\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right)\right) dx'_1 dx'_2
\end{aligned}
$$

$$
\begin{aligned}
P(\boldsymbol{u}|rep, \theta) &= f_{(2)}(\boldsymbol{u}, \boldsymbol{\mu}, \sigma^2, \rho) \prod_{k=1}^{2} \frac{1}{f_k(F_k^{-1}(u_k))} \\
P(\boldsymbol{u}|irep, \theta) &= f_{(2)}(\boldsymbol{u}, \boldsymbol{\mu}, \sigma^2, \rho) \prod_{k=1}^{2} \frac{1}{f_k(F_k^{-1}(u_k))}
\end{aligned}
$$

where $P(\boldsymbol{u}|rep, \theta)$ and $P(\boldsymbol{u}|irep, \theta)$ correspond to the emission probability that u is produced from the distribution of reproducible (rep) and irreproducible ($irep$) signals, respectively.

2.2.4 *Q function*

IDR-HMM estimates the optimal set of the parameters of mixture copula and hidden variables to maximize the likelihood of observed data. Because these parameters and hidden variables cannot be optimized at once, an expectation-maximization (EM) algorithm is applied in IDR-HMM, in which the expected value of the log likelihood function, or Q function, is iteratively maximized.

Let us consider the case of a single transcript t whose length is L, and whose read coverage is represented as $\boldsymbol{v_{1:L}}$. A likelihood of a specific path of latent variables $h_{1:L}$ is formulated as below:

$$
\begin{aligned}
P(v_{1:L}, h_{1:L}|\theta) =& P(h_0|\theta) \prod_{i=1}^{L} P(\boldsymbol{v_i}|h_i, \theta) \prod_{i=0}^{L-1} P(h_{i+1}|h_i, \theta) \\
P(h_0|\theta) :=& 1
\end{aligned}
$$

where h_0 is always *unmapped*, θ consists of two sets of copula parameters θ_1 and θ_2, $P(h_{i+1}|h_i, \theta)$ is a transition probability between h_i and h_{i+1}, and $P(\boldsymbol{v_i}|h_i, \theta)$ is an emission probability defined as follows:

$$
P(\boldsymbol{v_i}|h_i, \theta) = \begin{cases} P(\boldsymbol{v_{i,k_t+1:K}}|rep, \theta_1) P(\boldsymbol{v_{i,1:k_t}}|irep, \theta_2) & \text{if } h_i = \ stem \\ P(\boldsymbol{v_{i,k_t+1:K}}|irep, \theta_1) P(\boldsymbol{v_{i,1:k_t}}|rep, \theta_2) & \text{if } h_i = \ accessible \\ P(\boldsymbol{v_{i,k_t+1:K}}|irep, \theta_1) P(\boldsymbol{v_{i,k_t+1:K}}|rep, \theta_1) & \text{if } h_i = \ unmapped \end{cases} \tag{8}
$$

A transition probability is parametrized by $p(h''|h')$ of and included in θ, which satisfies the probability condition:

$$
\begin{aligned}
P(h''|h', \theta) &= p(h''|h') \\
\sum_{h'} p(h''|h') &= 1
\end{aligned}
$$

Then, a log likelihood ll and Q function Q is obtained as below:

$$
\begin{aligned}
ll(\theta|D, h_{1:L}) =& \log P(\boldsymbol{v_{1:L}}, h_{1:L}|\theta) Q(\theta|\theta') = \quad E_{h|\boldsymbol{v},\theta'}[ll(\theta|D, h)] \\
=& E_{h|\boldsymbol{v},\theta'}[\log P(\boldsymbol{v}, h|\theta)] \\
=& \sum_{h \in H} P(\boldsymbol{h}|\boldsymbol{v}, \theta') \log P(\boldsymbol{v}, h|\theta),
\end{aligned}
$$

where H is the set of possible paths of latent variables and θ' is an old set of parameters obtained in the previous iteration.

2.2.5 *EM algorithm*

Using the EM algorithm, each parameter is iteratively optimized to maximize Q function in IDR-HMM. In E step, a responsibility is computed for the expectation of log likelihood. Then, each parameter is moved to the direction of Q function differentiation in an M step. While the arguments of the maxima are sought in the EM algorithm, the optimization of copula parameters calls for re-computation of the pseudo-values. Due to the long re-computation time of pseudo-values, IDR-HMM actually adopts a strategy to set a limitation on the number of parameter optimizations in a single iteration, referred to as generalized EM [114].

2.2.6 *E step*

In this step, the expectation of log likelihood Q is calculated:

$$
\begin{aligned}
Q(\theta|\theta') &= \sum_{\boldsymbol{h}\in H} P(\boldsymbol{h}|\boldsymbol{v},\theta')\log P(\boldsymbol{v},h|\theta)\\
&= \sum_{\boldsymbol{h}\in H} P(\boldsymbol{h}|\boldsymbol{v},\theta')\log\left(P(h_0|\theta)\prod_{i=1}^{L}P(\boldsymbol{v_i}|h_i,\theta)\prod_{i=0}^{L-1}P(h_{i+1}|h_i,\theta)\right)\\
&= \sum_{\boldsymbol{h}\in H} P(\boldsymbol{h}|\boldsymbol{v},\theta')\sum_{i=1}^{L}\log(P(h_0|\theta)) + \log(P(\boldsymbol{v_i}|h_i,\theta)) + \log(P(h_{i+1}|h_i,\theta))\\
&= \log P(h_0|\theta) + \sum_{\boldsymbol{h}\in H}\sum_{i=1}^{L} P(\boldsymbol{h}|\boldsymbol{v},\theta')\log(P(\boldsymbol{v_i}|h_i,\theta)) + \sum_{\boldsymbol{h}\in H}\sum_{i=0}^{L-1} P(\boldsymbol{h}|\boldsymbol{v},\theta')\log(P(h_{i+1}|h_i,\theta))
\end{aligned}
$$

Of the three terms in the above formula, the first is 0. The second and third term are obtained as follow:

$$
\begin{aligned}
\sum_{\boldsymbol{h}\in H}\sum_{i=1}^{L} P(\boldsymbol{h}|\boldsymbol{v},\theta')\log(P(\boldsymbol{v_i}|h_i,\theta)) &= \sum_{i=1}^{L}\sum_{h'\in S}\gamma(h'_i|v,\theta')\log(P(v_i|h'_i,\theta))\\
\sum_{\boldsymbol{h}\in H}\sum_{i=0}^{L-1} P(\boldsymbol{h}|\boldsymbol{v},\theta')\log(P(h_{i+1}|h_i,\theta)) &= \sum_{h',h''\in S}\gamma(h',h''|\boldsymbol{v},\theta')\log(P(h''|h',\theta)),
\end{aligned}
$$

where a responsibility γ is defined for each position and state as below:

$$
\begin{aligned}
\gamma(h'_i|\boldsymbol{v},\theta') &:= \sum_{\boldsymbol{h}\in H} P(\boldsymbol{h}|\boldsymbol{v},\theta')I(h_i = h'_i)\\
\gamma(h',h''|\boldsymbol{v},\theta') &:= \sum_{\boldsymbol{h}\in H}\sum_{j=0}^{L-1} P(\boldsymbol{h}|\boldsymbol{v},\theta')I(h_j = h')I(h_{j+1} = h''),
\end{aligned}
$$

2.2.7 *M step*

In this step, a new optimal θ is obtained to maximize the expectation of log likelihood, resulting in fitting the model to the observation. The new θ should satisfy the equation $\theta = argmax_\theta Q(\theta|\theta')$ in EM. The posterior for an old parameter θ', such as $\gamma(h'_i|\boldsymbol{v},\theta')$ and $\gamma(h',h''|\boldsymbol{v},\theta')$, can be computed in the previous step after a forward-backward algorithm.

Optimization of transition probability

To obtain the optimal solution of transition probability $p(h''|h')$, Q function is differentiated by $p(h''|h')$, considering the constraint that the sum of $p(h''|h')$ is equal to one by introducing a Lagrange multiplier:

$$
\begin{aligned}
\frac{\partial Q}{\partial p(h''|h')} &= \frac{\partial}{\partial p(h''|h')}\left\{\sum_{h\in H}\sum_{i=1}^{L} P(h|v,\theta')\log P(h_{i+1}|h_i\theta') + \lambda\left(\sum_{h''} p(h''|h') - 1\right)\right\}\\
&= \frac{\partial}{\partial p(h''|h')}\{\gamma(h',h''|\boldsymbol{v},\theta')\log p(h''|h')\} + \lambda\\
&= \sum_{h\in H}\sum_{i=0}^{L-1}\gamma(h',h''|\boldsymbol{v},\theta')/p(h''|h') + \lambda\\
&= N(h',h''|\theta')/p(h''|h') + \lambda = 0,
\end{aligned}
$$

where $N(h''|h')$ is an expectation count of transition between h and h', and λ is a Lagrange multipler. This leads to the following equation regarding λ:

$$\lambda = N(h', h''|\theta')/p(h''|h')$$

Because λ is an only variable, the relationship between $p(h''|h')$ for any h' and h'' is derived as follows:

$$\begin{aligned} p(h'''|h') &= \frac{N(h', h''')}{N(h', h'')} p(h''|h') \\ \sum_{h'''} p(h'''|h') &= \frac{\sum_{h'''} N(h', h'''|\theta')}{N(h', h''|\theta')} p(h''|h') = 1 \\ p(h''|h') &= \frac{N(h', h''|\theta')}{\sum_{h'''} N(h', h'''|\theta')} \end{aligned}$$

Thus, an optimal solution of transition probability is the expectation ratio of transition between each state. In the HMM model, a parameter of irreproducible peaks, q (π in the original paper) should correspond to the ratio of latent variables. In this model, therefore, q is estimated after updating the transition matrix R. When R satisfies the probability condition for each row, it can be converted into the orthogonal matrix of eigenvalues by a matrix P:

$$RP = R(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}) = (\lambda_1 \boldsymbol{x_1}, \lambda_2 \boldsymbol{x_2}, \lambda_3 \boldsymbol{x_3})$$

$$P^{-1}RP = D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix},$$

where x_1, x_2, x_3 are eigenvectors, and $\lambda_1, \lambda_2, \lambda_3$ are eigenvalues of R. The expectation of the duration time $\boldsymbol{d}$ ($d_{\mathrm{Unmapped}}, d_{\mathrm{Stem}}, d_{\mathrm{Accessible}}$) for the $v_{1:L}$ is obtained as follows:

$$\begin{aligned} E[\boldsymbol{d}] &= R + R^2 + \cdots + R^L \\ &= \frac{1}{L}\left(PDP^{-1} + PD^2P^{-1} + \cdots + PD^LP^{-1}\right) \\ &= \frac{1}{L} P \begin{pmatrix} a(\lambda_1) & 0 & 0 \\ 0 & a(\lambda_2) & 0 \\ 0 & 0 & a(\lambda_3) \end{pmatrix} P^{-1} \quad \left(a(\lambda) := \frac{\lambda(1-\lambda^n)}{1-\lambda} \right) \end{aligned}$$

When the sequence length L is so long that the expectation of duration time can be approximated by the equilibrium distribution, it is obtained by the eigenvector, which corresponds to the largest eigenvalue of one. Specifically, the ratios of reproducible samples for *Stem* (d_{Stem}) and *Acc* (d_{Acc}) are obtained as follows, respectively:

$$\begin{aligned} q_{\mathrm{Stem}} &= d_{\mathrm{Stem}} / |\boldsymbol{d}| \\ q_{\mathrm{Acc}} &= d_{\mathrm{Acc}} / |\boldsymbol{d}| \end{aligned}$$

Optimization of copula parameters

The parameters for the mixture copula model μ, σ, and ρ are individually optimized to increase Q function for each sample (θ_1 for $\boldsymbol{v}_{\boldsymbol{i},\boldsymbol{k_t}+\boldsymbol{1}:\boldsymbol{L}}$ and θ_2 for $\boldsymbol{v}_{\boldsymbol{i},\boldsymbol{1}:\boldsymbol{k_t}}$). For the parameters for the reproducible class, a differential of Q function in terms of copula parameters is as below.

$$\begin{aligned}\frac{\partial Q}{\partial \theta} &= \left(\sum_{i=1}^{L} \sum_{h' \in S} \gamma(h'_i | \boldsymbol{v}, \theta') \log(P(\boldsymbol{v_i} | h_i, \theta)) \right)' \\ &= \sum_{i=1}^{L} \sum_{h'_i \in S} \gamma(h'_i | \boldsymbol{v}, \theta') \frac{\partial \log(P(\boldsymbol{v_i} | h'_i, \theta))}{\partial \theta}\end{aligned}$$

In the logarithm form, a differential of $P(\boldsymbol{v_i}|h_i, \theta)$ (defined in Eq. 8) by θ_1 and θ_2 can be considered independently, as follows:

$$\frac{\partial \log P(\boldsymbol{v_i}|h_i,\theta)}{\partial \theta_1} = \begin{cases} \log(P(\boldsymbol{v}_{\boldsymbol{i},\boldsymbol{k_t}+\boldsymbol{1}:\boldsymbol{K}}|rep, \theta_1))' & \text{if } h_i = stem \\ \log(P(\boldsymbol{v}_{\boldsymbol{i},\boldsymbol{k_t}+\boldsymbol{1}:\boldsymbol{K}}|irep, \theta_1))' & \text{otherwise.} \end{cases}$$

$$\frac{\partial \log P(\boldsymbol{v_i}|h_i,\theta)}{\partial \theta_2} = \begin{cases} \log(P(\boldsymbol{v}_{\boldsymbol{i},\boldsymbol{1}:\boldsymbol{k_t}}|rep, \theta_2))' & \text{if } h_i = accessible \\ \log(P(\boldsymbol{v}_{\boldsymbol{i},\boldsymbol{1}:\boldsymbol{k_t}}|irep, \theta_2))' & \text{otherwise.} \end{cases}$$

Since the differential of θ_1 and θ_2 is obtained by similar means, I will consider the case that θ does not contain θ_2 and $k_t = 2$ ($\boldsymbol{v_i} = (u_1, u_2)$). I will introduce several variables for the sake of explanation:

$$\begin{aligned}
&\log R = (\log N_{12r})' \\
&\log I = (\log N_{12i})' \\
&\log S = (\log(qN_{1r} + (1-q)N_{1i}) + \log(qN_{2r} + (1-q)N_{2i})) \\
&\log S_1 = \log f_1 = \log\left(qN_{1r} + (1-q)N_{1i}\right) \\
&\log S_2 = \log f_2 = \log\left(qN_{2r} + (1-q)N_{2i}\right) \\
&N_{12,r} = N\left(\begin{array}{c|c} x'_1 & \mu \\ x'_2 & \mu \end{array}, \begin{pmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{pmatrix} \right) \\
&N_{12,i} = N\left(\begin{array}{c|c} x'_1 & 0 \\ x'_2 & 0 \end{array}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \\
&N_{1,r} = N(x_1|\mu, \sigma^2), N_{1,i} = N(x_1|0,1) \\
&N_{2,r} = N(x_2|\mu, \sigma^2), N_{2,i} = N(x_2|0,1) \\
&s = (x_1-\mu)^2 + (x_2-\mu)^2 - 2\rho(x_1-\mu)(x_2-\mu)
\end{aligned}$$

Then, the emission probability of IDR-HMM, which depends on θ, is a joint probability defined as below:

$$f_{h_i}(F_1^{-1}(u_1), F_2^{-1}(u_2)) := \begin{cases} \log(P(\boldsymbol{v_i}|rep,\theta)) = \log R - \log S & \text{if } h_i = stem \\ \log(P(\boldsymbol{v_i}|irep,\theta)) = \log I - \log S & \text{otherwise.} \end{cases},$$

Since F_1^{-1} and F_2^{-1} also depend on θ, a differentiation of $f_{h_i}(F_1^{-1}(u_1), F_2^{-1}(u_2))$ is obtained as follows (also see Eq. 10 in the Appendix):

$$\frac{\partial \log f_{h_i}(F_1^{-1}(u_1), F_2^{-1}(u_2))}{\partial \theta} = \frac{\partial \log f}{\partial \theta}(F_1^{-1}(u_1), F_2^{-1}(u_2)) + \frac{\partial F_1^{-1}(u_1)}{\partial \theta}\frac{\partial \log f}{\partial x_1} + \frac{\partial F_1^{-1}(u_2)}{\partial \theta}\frac{\partial \log f}{\partial x_2}$$

(9)

To compute this differentiation, it is necessary to obtain three components, the differential of $f(F_1^{-1}(u_1), F_2^{-1}(u_2))$ by θ, and the differential of F_1^{-1} and F_2^{-1} by x_1 and x_2, respectively. $\frac{\partial \log F^{-1}(u)}{\partial \theta}$ is obtained as follows.

$$\begin{aligned}\frac{\partial F_1^{-1}(u_1)}{\partial \theta} &= -\frac{\frac{\partial F_{1,\theta}}{\partial \theta}(F_1^{-1}(u_1))}{\frac{\partial F_1}{\partial x_1}} = -\frac{\frac{\partial F_\theta}{\partial \theta}(F^{-1}(u_1))}{f_1(x_1)}\\ \frac{\partial F_2^{-1}(u_2)}{\partial \theta} &= -\frac{\frac{\partial F_{2,\theta}}{\partial \theta}(F_2^{-1}(u_2))}{\frac{\partial F_2}{\partial x_2}} = -\frac{\frac{\partial F_{2,\theta}}{\partial \theta}(F_2^{-1}(u_2))}{f_2(x_2)},\end{aligned}$$

The numerator of the above equation can be formulated for each parameter, μ, σ^2, and ρ, in the following ways.

$$\begin{aligned}\frac{\partial F_i}{\partial \mu}(F_i^{-1}(u_i)) &= (qN_{ir} + (1-q)N_{ii})'\\ &= -qN_{ir}\\ \frac{\partial F_i}{\partial \sigma^2}(F_i^{-1}(u_i)) &= -\frac{(x_i - \mu_i)}{2(\sigma^2)}qN_{ir}\\ \frac{\partial F_i}{\partial \rho}(F_i^{-1}(u_i)) &= 0,\end{aligned}$$

$\frac{\partial \log f}{\partial x}$ can be calculated as follows:

$$\frac{\partial \log f_{h_i}(F_1^{-1}(u_1), F_2^{-1}(u_2))}{\partial x_i} = \begin{cases} (\log R - \log S)' & \text{if } h_i = stem \\ (\log I - \log S)' & \text{otherwise.}\end{cases}$$

$$\begin{aligned}\frac{\partial}{\partial x_1}\log R &= (\log N_{12r})' = -\frac{(x_1-\mu) - \rho(x_2-\mu)}{\sigma^2(1-\rho^2)}\\ \frac{\partial}{\partial x_1}\log I &= (\log N_{12i})' = -(x_1)\\ \frac{\partial}{\partial x_1}\log S &= (\log(qN_{1r} + (1-q)N_{1i}) + \log(qN_{2r} + (1-q)N_{2i}))'\\ &= -\left(\frac{(x_1-\mu)}{\sigma^2}\frac{qN_{1r}}{qN_{1r} + (1-q)N_{1i}}\right)\end{aligned}$$

$\frac{\partial \log f}{\partial \theta}(F_1^{-1}(u_1), F_2^{-1}(u_2))$ is also obtained as follows:

$$\frac{\partial \log f}{\partial \theta}(F_1^{-1}(u_1), F_2^{-1}(u_2)) = \begin{cases} (\log R - \log S)' & \text{if } h_i = stem \\ (\log I - \log S)' & \text{otherwise}\end{cases},$$

$$\begin{aligned}
\frac{\partial \log R}{\partial \mu} &= (\log N_{12r})' \\
&= \left(-\frac{(x_1-\mu)^2+(x_2-\mu)^2-2\rho(x_1-\mu)(x_2-\mu)}{2\sigma^2(1-\rho^2)}\right)' \\
&= \frac{(x_1-\mu)+(x_2-\mu)+\rho(2\mu-(x_1+x_2))}{\sigma^2(1-\rho^2)} \\
&= \frac{(x_1+x_2-2\mu)}{\sigma^2(1+\rho)} \\
\frac{\partial \log S}{\partial \mu} &= (\log(qN_{1r}+(1-q)N_{1i})+\log(qN_{2r}+(1-q)N_{2i}))' \\
&= \frac{qN_{1r}}{S_1}\frac{x_1-\mu}{\sigma^2}+\frac{qN_{2r}}{S_2}\frac{x_2-\mu}{\sigma^2} \\
\frac{\partial \log R}{\partial \sigma^2} &= (\log N_{12r})' \\
&= \left(-\frac{\log(\sigma^2)}{2}-\frac{(x_1-\mu)^2+(x_2-\mu)^2-2\rho(x_1-\mu)(x_2-\mu)}{2\sigma^2(1-\rho^2)}\right)' \\
&= -\frac{1}{2\sigma^2}+\frac{(x_1-\mu)^2+(x_2-\mu)^2-2\rho(x_1-\mu)(x_2-\mu)}{2\sigma^4(1-\rho^2)} \\
\frac{\partial \log S}{\partial \sigma^2} &= (\log(qN_{1r}+(1-q)N_{1i})+\log(qN_{2r}+(1-q)N_{2i}))' \\
&= \frac{qN_{1r}}{S_1}\left(-\frac{1}{2\sigma^2}+\frac{(x_1-\mu)^2}{2\sigma^4}\right)+\frac{qN_{2r}}{S_2}\left(-\frac{1}{2\sigma^2}+\frac{(x_2-\mu)^2}{2\sigma^4}\right) \\
\frac{\partial \log R}{\partial \rho} &= (\log N_{12r})' \\
&= \left(-\frac{\log(1-\rho^2)}{2}-\frac{s}{2\sigma^2(1-\rho^2)}\right)' \\
&= \frac{\rho}{1-\rho^2}+\frac{(x_1-\mu)(x_2-\mu)(1-\rho^2)-\rho s}{\sigma^2(1-\rho^2)^2} \\
\frac{\partial \log I}{\partial \rho} &= 0 \text{ (const)} \\
\frac{\partial \log S}{\partial \rho} &= 0 \text{ (const)}
\end{aligned}$$

For an actual dataset, the summation of Eq. 9 for all positions of each transcript is performed on θ_1 and θ_2, and the first derivatives of Q function can be computed by multiplying the differential of the emission probability and responsibility. Optimization of the parameters is carried out using the gradient descent method, in which θ is changed from θ' as follows:

$$\theta = \theta' - \alpha\frac{\partial Q}{\partial \theta'},$$

where α is initialized to $5e-12$, multiplied by 10 in each step, then finally stopped when $\alpha > 1e-2$. Although IDR-HMM has not yet been carried out for the transcriptome dataset, it is expected to extract a larger number of reproducible regions from low-coverage nucleotides compared to filtering of the original IDR, due to the additional information of locally consistent signals.

Multiple high-throughput structure methodology comparison and computational prediction

To investigate the consistency between multiple high-throughput structure analyses, I used the Structure Surfer database [90], in which the reactivity scores obtained by icSHAPE [88], DMS-Seq [22], PARS [20], and ds/ssRNA-Seq [90] are mapped to human and mouse genomes and combined with ucsc gene ids. From the entire of dataset, I selected parts of studies analyzing human transcriptome, which consisted of two replicates of PARS (rep1 and rep2), DMS-seq, and the ds/ssRNA-seq dataset. To compare experimental and computational structure analyses, I used ParasoR [109] for prediction of stem probability p_{stem}, which is applicable to long RNAs. Using ParasoR, p_{stem} was computed for each transcript of the hg19 knownGeneMrna dataset downloaded from the UCSC database [53], with a maximal span of $W = 200$. Then, I analyzed the pairwise correlation of the reactivity scores of experimental and computational analyses, only for transcripts that have the same length in both datasets. The transcripts in which less than 100 nucleotides or less than half of the regions are annotated by reactivity scores were excluded from the comparison. After removing the no-score regions, the remaining transcripts were $4,327$ (ParasoR vs PARS rep1), $3,846$ (vs PARS rep2), $48,910$ (vs DMS-seq), $69,248$ (ParasoR vs ds/ssRNA-Seq), $3,716$ (PARS rep1 vs rep2), $4,113$ (vs DMS-Seq), $4,292$ (vs ds/ssRNA-Seq), $3,658$ (PARS rep2 vs DMS-Seq), $3,818$ (PARS rep2 vs ds/ssRNA-Seq), and $47,913$ (DMS-seq vs ds/ssRNA-Seq). Because DMS-seq and ds/ssRNA-Seq detect the reactivity of each nucleotide while other scores correspond to double-strandedness, negative DMS-seq and ds/ssRNA-Seq scores were applied for the calculation of correlation coefficients. To scale each score to $(0,\ 1)$, I computed the normalization factors of the minimum and maximum reactivity scores $smin$ and $smax$ before calculating the normalized score as $\text{Score}' := \frac{\text{Score}-smin}{smax-smin}$.

Reference structure of human rRNA and transcriptome

I established the set of rRNA sequences to map PARS and icSHAPE reads for an evaluation of the accuracy of reactivity classification with IDR filtering. A ribosomal repeating unit of human (NT_167214.1) was extracted from the NCBI database. 5S rRNA sequences were additionally downloaded by Ensembl database API using the settings “gene type” to “rRNA”. Among them, I extracted three unique sequences of 5S rRNA for mapping. As a reference structure of 18S rRNA, a secondary structure and sequence of human 18S rRNA were downloaded from [115]. This sequence has two mismatches with the 18S rRNA sequence of NCBI.

The reference of human transcriptome consisted of UCSC Refseq sequences (as of October 7, 2016) and GENCODE transcript sequences (v12), following the method of constructing the reference sequences described in Ref. [20]. Since there is no reference structure for transcripts, stem probability was computed for all the sequences using ParasoR [109].

Mapping and quantification of PARS and icSHAPE datasets

For whole transcriptome analyses, I downloaded the PARS score dataset for human transcriptome, as measured in the previous study (GEO accession number is GSE50676) [20]. Normalized read counts of GM12878 native deproteinized replicates with nuclease S1 and V1 treatment (hereafter referred to as S1 and V1, respectively) for two replicates. To remove the influence of indiscriminative multi-mapping reads, I realigned sequencing reads of downloaded fastq files for the set of repeat unit sequences and human transcriptome. The detail of the read alignments is described in the Appendix. Read counts were then normalized by total counts mapped to the 18S rRNA and transcriptome.

As a representative of RT stop-based reactivity detection, the icSHAPE dataset was compared with the reference structure and PARS dataset. In this study, the dataset of the previous icSHAPE study for the HEK 293T cell was analyzed for the comparison with the *in vitro* PARS dataset (GEO accession number is GSE74353) [81]. I downloaded fastq files obtained for three conditions with two replicates, which were DMSO-, *in vitro* NAI-N_3-, and *in vivo* NAI-N_3-treated cells. Then, specific barcodes that were 13-nt in length were removed from 5′-ends of RNA fragments after excluding PCR duplications from the sequence library. RT stop counts were measured for the counts of mapped reads using bowtie2 to the human transcriptome reference constructed for PARS analyses. According to the previous study, it is suggested that a large fraction (more than 20 %) of cDNAs tend to have mismatches at the extreme 3′ end due to the terminal transferase activity of the reverse transcriptase [93]. To confirm the abundance of flanking nucleotides observed in icSHAPE and PARS reads, the local alignment of bowtie2 was also performed. The amount of flanking nucleotides of each length was determined by the number of soft-clipped bases in bam files for each 5′ end. The detail of data processing for sequencing reads, such as mapping and trimming, is written in the Appendix.

3 RESULTS

3.1 MULTI-COMPARISON BETWEEN HIGH-THROUGHPUT STRUCTURE PROBING ANALYSES AND *IN SILICO* STRUCTURE ANALYSES

In recent years, more than tens of protocols have been developed for high-throughput structure analyses using a specific reagent or enzyme to recognize flexibility of each nucleotide. They differed from each other in terms of library preparation processes, such as probing reagent, RNA extraction, and existence of barcode addition. Since the development of each method has been conducted by different research group, it is still unknown to what extent each high-throughput analysis can be distinguished from its counterparts in regards to the detection accuracy. In this section, I demonstrate the comparison of reactivity scores of human transcriptome as produced by many different protocols, using the Structure Surfer database [90]. Structure Surfer is a database of transcriptome-wide high-throughput structure analyses for multiple species, and I investigated the basic properties of stored high-throughput structure dataset.

Figure 50 shows the distribution of reactivity for six high-throughput structure analyses. The distributions of reactivity differed between cleavage-based methods (PARS and ds/ssRNA-Seq) and modification-based approaches (DMS-Seq and icSHAPE). In particular, while the score distributions of DMS shows a long tail in one direction, those of PARS and ds/ssRNA-Seq, which is based on the log ratio of the enrichment about single- and double-stranded regions, have a median at the middle of distribution, Hence, this means that the former and latter methods tried to detect the different number of classes of RNA secondary structure, two and single classes, respectively. In addition, handling of reactivity outliers also differed from each other, in particular, the reactivity of DMS was distributed over a wide range (Appendix Figure 70). These differences should be aligned carefully for the cross-comparison between different methodologies, or they might be the cause of spurious inconsistency.

Next, I evaluated the consistency of genome-wide structure analyses on human transcriptome data. I calculated Spearman's correlation coefficients for the pair of vectors of reactivities assigned to the same location across each human transcript after removing defect regions. In addition to reactivities measured by experiments, stem probability p_{stem} was calculated by ParasoR software to confirm the consistency of thethermodynamic folding model and experimental structure analyses. As a result, although more than a half of transcripts showed positive correlation, most correlation coefficients were distributed around 0 between different studies with the two exceptions (Figure 51). One is an apparent higher correlation between PARS replicates, and another is moderate positive correlation between PARS and ds/ssRNA-Seq. The, this tendency was robust to the changes the features for subject extraction, such as a particular length of transcripts and analyzing limited to a specific nucleotide (data not shown). It should be noted that each high-throughput analysis was applied to different types of cell in the different condition and must show cell-to-cell variation. Hence, this result at least indicated that just a direct comparison of reactivity scores over multiple high-throughput analyses exhibited only a weak correlation, despite these studies sharing a common purpose of revealing the structural landscape of human transcriptome. This is a matter of concern because it might be meaningless

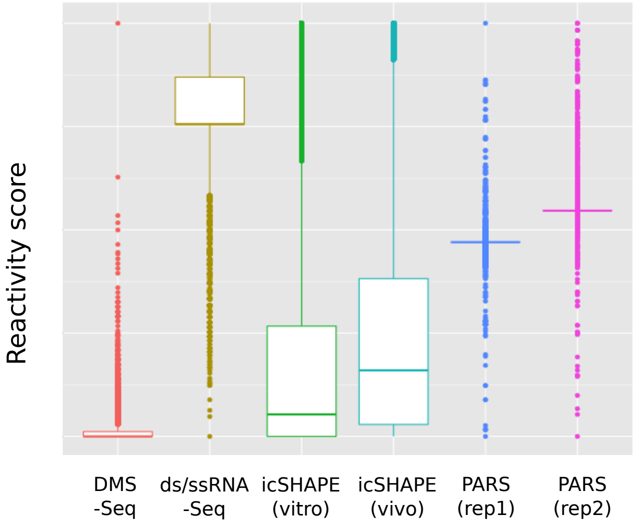


Figure 50. Distribution of reactivity scores for whole transcriptome of 6 types of high-throughput structure analyses archived in Structure Surfer. Each score distribution is scaled to $[0,\ 1]$.

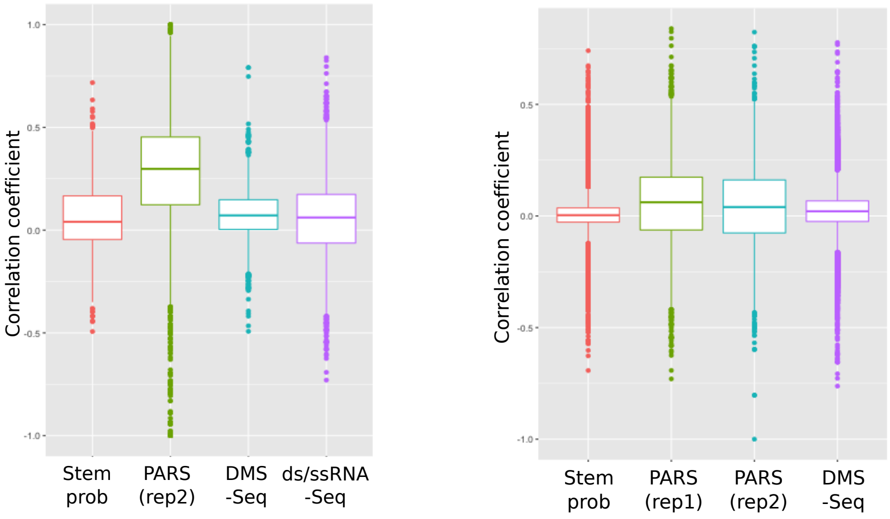
to analyze the whole landscape of RNA secondary structure over transcriptome if the most of transcripts are truly inconsistent.

3.2 PROPENSITY OF SYSTEMATIC ERRORS OBSERVED IN HIGH-THROUGHPUT STRUCTURE ANALYSES

To determine what systematic biases and noises have to be considered to extract only reliable reactivity scores beyond the errors, I examined the properties of replicated genome-wide structure analyses based on cleavage-based methodology (PARS dataset from Ref. [20]) and modification-based methodology (icSHAPE dataset from Ref. [81]).

Computation of PARS score is carried out for the sequencing data of two different libraries treated with the nuclease S1 and V1. Although digestion of RNA by S1 and V1 have been used to infer the secondary structure of RNA, they have individual characteristics, or limitation to recognize base pairs [116]. For example, the frequency of V1 cleavage is known to be lower in addition to that V1 tends to recognize tips of helices rather than the inside of helices. For that reason, I first examined the characteristics of S1- and V1-treated read coverage distribution $c_{\mathrm{PARS\_condition}}$ and checked the consistency of them between the replicates in single-nucleotide resolution. As shown in Figure 52, low-coverage regions, specifically in the range of $[0,\ 10]$, are likely to be irreproducible, compared to the high-coverage regions in a logarithmic scale. This tendency is known as a typical feature observed in sequencing-based quantification such as RNA-Seq [117]. Therefore, it is required to remove the spurious peaks located in low-coverage regions for robust reactivity analyses.

Note that other methodology, such as DMS-Seq and icSHAPE, does not take the ratio of normalized read counts instead of the log ratio. However, the inclusion of such irreproducible regions may produce the order fluctuation of read coverage in even such methodology, resulting in the decrease of rank correlation for the comparison with other high-throughput structure analyses. In addition, such fluctuation may cause a severe effect on downstream analyses, such as classification based on the threshold setting, gene set or GO enrichment analyses, or structure prediction assisted with reactivity.



(a)

(b)

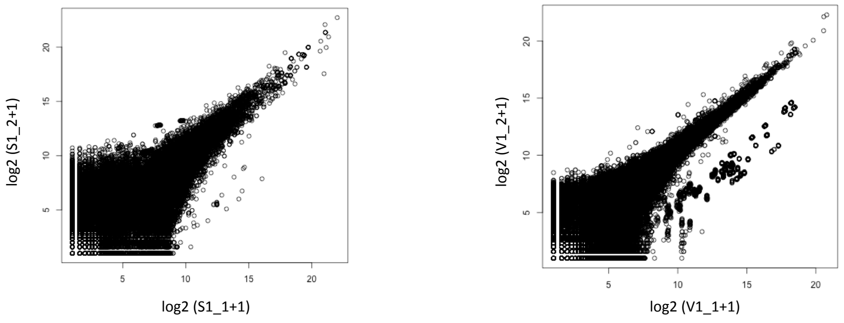
Figure 51. Distribution of correlation coefficients about reactivities of (a) PARS rep1 and (b) ds/ssRNA-Seq with those of other methods across each human transcript.

Moreover, the clusters of potentially irreproducible read counts were observed at the lower right in the V1 distribution in Figure 52. I investigated the cause of inconsistency among those transcripts by extracting the transcripts that have a region meeting the conditions; $c_{\text{PARS\_V1}}$ of rep1 and rep2 are more than 100, and the ratio of $c_{\text{PARS\_V1}}$ between rep1 and rep2 is more than 10. Consequently, 40 unique transcripts were extracted as irreproducible transcripts and they consist of RNase P RNA component, 5S ribosomal, 5S ribosomal pseudogene, U1 small nuclear, and these variants. Because these RNAs are supposedly highly abundant in general conditions [118], the incongruence might be derived from PCR duplication or multiple mapping (which was excluded in the original paper [19] but retained in [20]).

The distributions of means and variances of read counts for each transcript are also plotted in Figure 53. The variances of $c_{\text{PARS\_S1}}$ and $c_{\text{PARS\_V1}}$ were shown to be larger than their means in general, indicating that the read coverage distribution of each transcript does not follow Poisson distribution. However, it is still possible that the read coverage distribution only from random regions can be approximated by Poisson distribution. Thus, the way of p-value computation only from the distribution of low read coverages, as with Ref. [113], would be effective to construct more suitable null hypotheses for PARS score.

As another example, I investigated icSHAPE dataset, which is one of the modification-based structure analyses. Computation of icSHAPE score is carried out for the sequencing data of two different libraries treated with DMSO (control) and NAI-N_3. Using a control sample, false positive sites are excluded from the candidates of flexible nucleotides, such as the enrichment of fragmentation and RT stop by base pairing or modification [119].

I compared the distributions of read coverages, $C_{\text{icSHAPE\_condition}}$ for control, *in vitro*, and *in vivo* samples obtained. Figure 54 shows the scatter plots of $C_{\text{icSHAPE\_condition}}$ only about each nucleotide across human transcriptome with at least single mapped read. As a result, the whole trend of $C_{\text{icSHAPE\_condition}}$ was observed to be biased among replicates, probably due to the difference of total sequencing depth. This suggests the validity of total read count normalization as well as a difficulty of setting a single threshold that works effectively to filter



(a) (b)

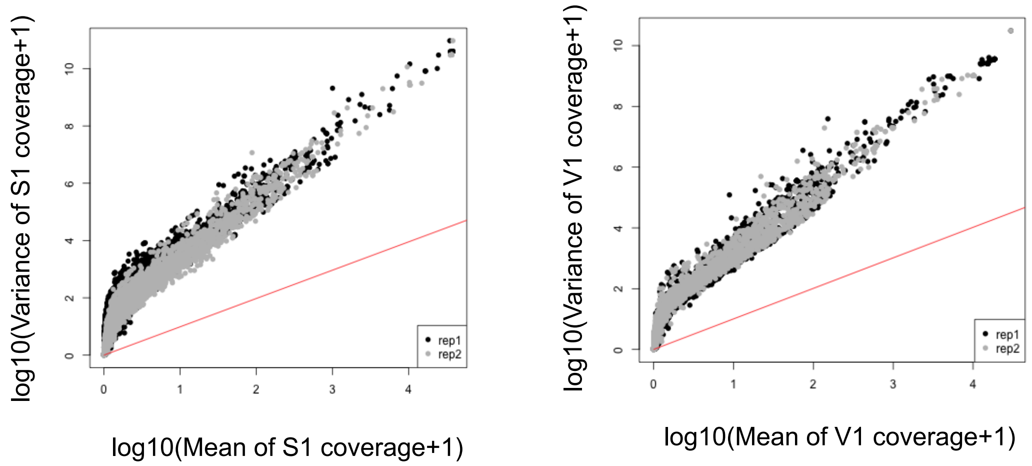
Figure 52. Normalized read coverage distribution $c_{\mathrm{PARS\_condition}}$ of PARS dataset treated with nuclease S1 (a) and V1 (b) for two replicates.

out irreproducible signals. Additionally, the distributions themselves are similar to each other, regardless of treatment of the icSHAPE reagent. Since no specific enrichment was observed in icSHAPE reagent-treated samples, it is indicated that only the ratio difference of read acquisition rate would help to distinguish modification sites.

3.3 CONSISTENCY OF *IN SILICO* STRUCTURE CLASSIFICATION AND REPRODUCIBLE HIGH-THROUGHPUT STRUCTURE ANALYSES BASED ON IDR ESTIMATION

To evaluate the reproducibility of each read coverage, I performed IDR estimation for $c_{\mathrm{PARS\_S1}}$ and $c_{\mathrm{PARS\_V1}}$ using reactIDR. Figure 55 shows the relationship between estimated IDR and $c_{\mathrm{PARS\_condition}}$ for the PARS dataset. Consequently, high-coverage regions were estimated to have apparently lower IDR so that the regions not fluctuated in terms of the ranking are considered as “reproducible”. In the previous study, a threshold for IDR was set to 0.01 for ChIP-Seq analyses. At this time, if the threshold is same, the nucleotides whose $c_{\mathrm{PARS\_condition}}$ is greater than 100 in both of replicates are classified as reproducible peaks. This minimum read coverage is in the same order of that arbitrary set in the previous study [88], indicating that classification of reproducible samples based on IDR is supposed to be reasonable.

Additionally, the surrounding profiles of reproducible and irreproducible $c_{\mathrm{PARS\_condition}}$ were investigated to understand the characteristics of reproducible samples as shown in Figure 56. A representative of reproducible peaks was defined as the region whose $c_{\mathrm{PARS\_condition}}$ is within the range of $[100, 1000]$ in both replicates. That of irreproducible peaks was also defined as the region whose $c_{\mathrm{PARS\_condition}}$ is within the range of $[100, 1000]$ in one dataset, and less than 10 in another. As a result, the peaks of reproducible $c_{\mathrm{PARS\_S1}}$ and $c_{\mathrm{PARS\_V1}}$ were showed to be broad compared to the shallow spike-like forms of irreproducible peaks. It was also observed that the autocorrelation of $c_{\mathrm{PARS\_condition}}$ remained high within several bases, suggesting the positional dependency of $c_{\mathrm{PARS\_condition}}$ enrichment (Appendix Figure 71). Interestingly, $c_{\mathrm{PARS\_V1}}$ of reproducible samples showed another peak at the upstream of several bases. This was most likely to be caused by systematic biases about the enrichment of V1 recognition at the tips of specific structures such as the adjacent nucleotide to short helices.



(a)

(b)

Figure 53. Relationship between mean and variance of $c_{\text{PARS\_S1}}$ (a) and $c_{\text{PARS\_V1}}$ for each transcript. $\log_{10}$ mean and $\log_{10}$ variance of the read count distribution for each transcript is shown in the x- and the y-axis respectively. Red lines show $y = x$, which corresponds to the mean and variance of Poisson distribution.

Next, the consistency between the PARS dataset and computational structure prediction was examined depending on the IDR-based classification. Generally, the correlation between PARS score and p_{stem} is not lower compared to that of high-coverage regions (Appendix Figure 64). Thus, I tested IDR-based classification that a nucleotide is classified into "accessible" and "structured" when estimated IDR was lower than 0.01 for S1 and V1 dataset, respectively. I also defined "neutral" class as the samples included in neither accessible nor structured, or those belonging to both accessible and structured. Figure 57 shows the distribution of p_{stem} for each class based on IDR computed for the S1 and V1 datasets. Consequently, the median of p_{stem} clearly increased in the order of accessible, neutral, and structured. Although there are still a certain number of inconsistent nucleotides between p_{stem} and PARS score, they might be influenced by complex secondary structures, such as long-range base pairs or RBP binding, as well as the failure of computational prediction.

3.4 Accuracy of 18S rRNA structure prediction by reactivity scores with filtering based on IDR

To evaluate the accuracy of reactivity classification with IDR filtering, I performed read mapping for the PARS and icSHAPE datasets, then compared a reference structure of 18S rRNA and reactivity scores. Before comparison of reactivity scores and reference structure, the characteristics of read mapping were investigated, particularly for the flanking nucleotides using bowtie2 with local alignment mode. As a result, a greater number of icSHAPE reads were mapped with soft-clipped bases in local alignment mode than the PARS dataset (Figure 58). This result is consistent with the previous report on flanking nucleotides by the terminal transferase activity [93]. It might also reduce the resolution and precision of structure detection. Accordingly, to avoid such case, I ran bowtie2 with local alignment for read mapping of the icSHAPE dataset

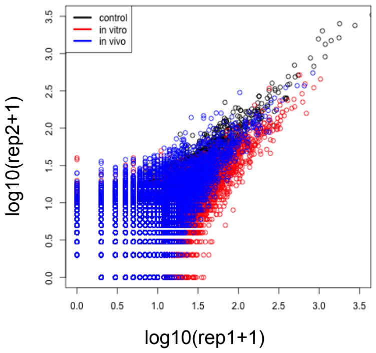
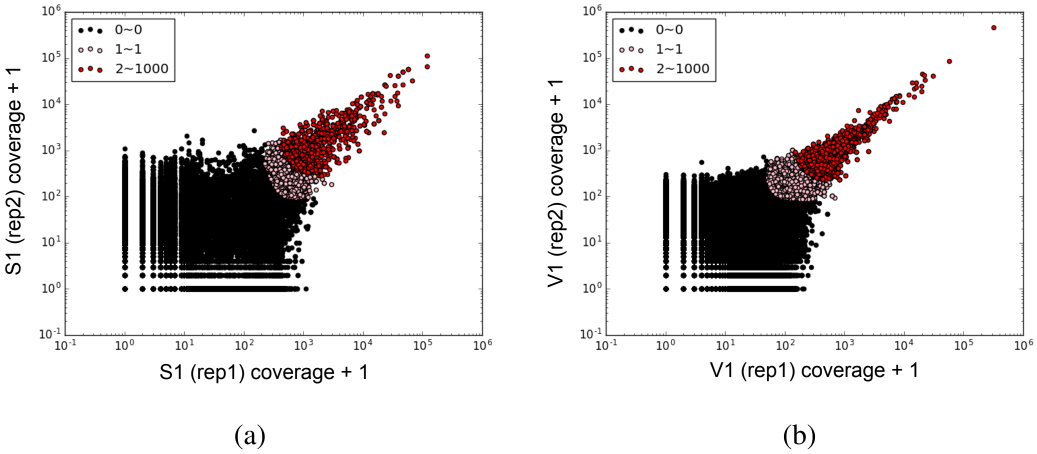


Figure 54. Scatter plots of replicated $C_{\text{icSHAPE\_condition}}$ for each condition; control (DMSO-treated), *in vitro* (NAI-N$_3$-treated *in vitro*), and *in vivo* (NAI-N$_3$-treated *in vitro*). $C_{\text{icSHAPE\_DMSO}}$, $C_{\text{icSHAPE\_}\textit{in vitro}}$, and $C_{\text{icSHAPE\_}\textit{in vivo}}$ at each nucleotide across human transcriptome are shown by black, red, and blue circles, respectively.



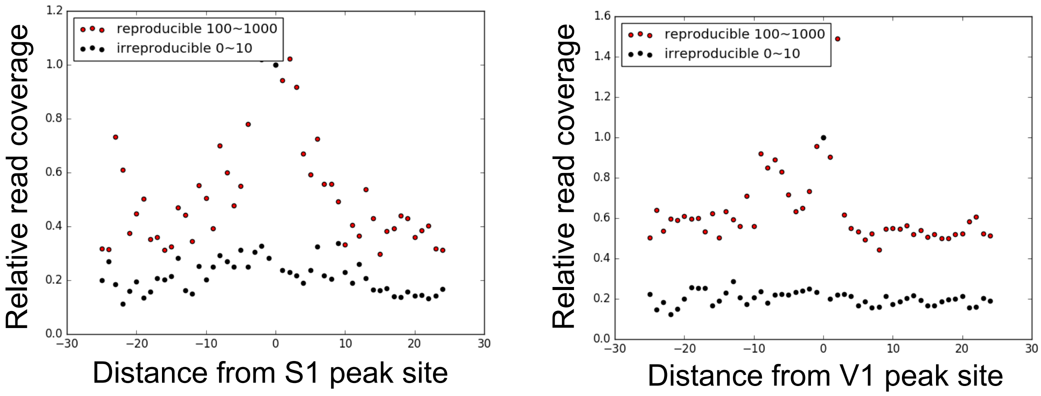
(a)

(b)

Figure 55. Scatter plots for normalized read coverage of PARS S1 $c_{\text{PARS\_S1}}$ (a) and V1 $c_{\text{PARS\_V1}}$ (b). Red, pink, and black points correspond to the nucleotides whose read coverages are highly reproducible, moderately reproducible, and irreproducible, respectively ($-\log_{10}$ IDR is shown in the legends).

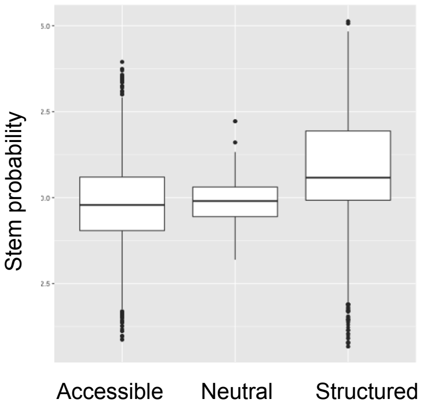
before the truncation of soft-clipped bases. Such exact 5′-end positions after truncation were measured for further analyses.

Next, the consistency between PARS dataset and computational structure prediction was also examined for 18S rRNA by computing correlation between the features of reactivity scores and reference structure converted into a numerical vector consisting of 0, 0.5, and 1.0 at unpaired bases, non-canonical base pairs, and canonical base pairs (Figure 59). As a result, the top 2 largest absolute correlation coefficients were obtained by p_{stem} (0.317, 0.318, 0.319 for PARS, icSHAPE *in vitro*, and *in vivo*, respectively) and icSHAPE score (0.20, -0.23, -0.25) for all nucleotides. Those correlations were increased overall by limiting to the accessible or structure nucleotides based on IDR ((0.41, 0.33, and 0.28) by p_{stem} and (0.34, -0.26, -0.20) by icSHAPE score). This result indicated that the accuracy of computational prediction is substantially higher than those of high-throughput structure analyses without filtering for 18S rRNA, whose structure has been well studied by computational prediction. Also, the difference of accuracy became close after filtering out irreproducible or ambiguous nucleotides in high-throughput structure analyses. Interestingly, a feature of estimated IDR for DMSO-treated samples in



(a) (b)

Figure 56. Positional profiles of $c_{\mathrm{PARS\_S1}}$ (a) and $c_{\mathrm{PARS\_V1}}$ (b) around the reproducible position (red) and irreproducible position (blue). A definition of reproducible and irreproducible signals is that the region whose $c_{\mathrm{PARS\_condition}}$ is within the range of [100, 1000] in both replicates, and the region whose $c_{\mathrm{PARS\_condition}}$ is within the range of [100, 1000] in one dataset, and less than 10 in another, respectively.

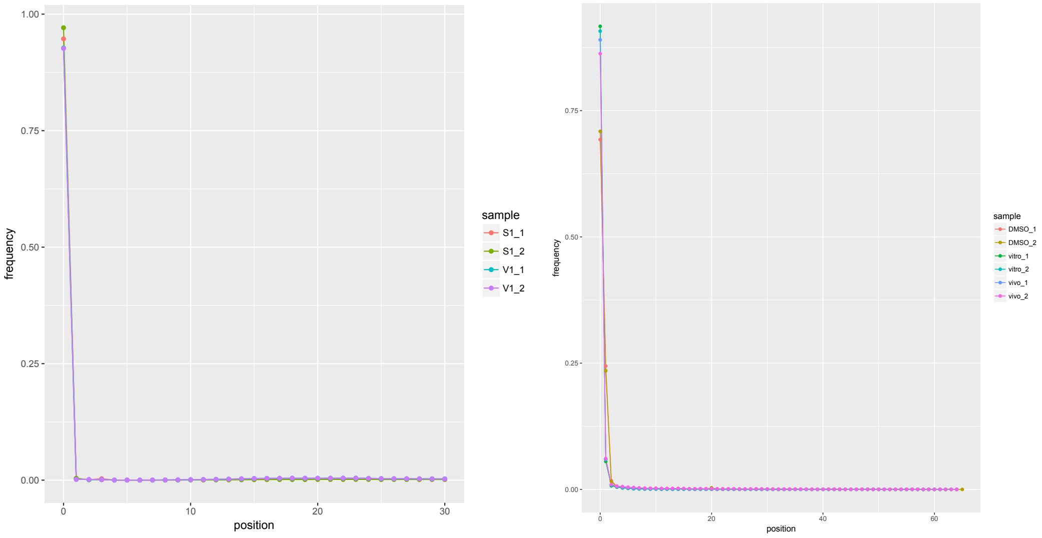


(a)

Figure 57. Distribution of stem probability based on the IDR-based classification into three classes of accessible, neutral, and structured. Each nucleotide is classified as accessible and structured when estimated IDR is lower than 0.01 for S1 and V1 dataset, respectively. After the classification, the samples included in neither accessible nor structured, or those belonging to both accessible and structured were classified into neutral class..

icSHAPE dataset showed positive correlations with base pairing clearly. In icSHAPE and other modification-based structure analyses, the enrichment of DMSO-treated samples has been used only to decrease the significance of the enrichment of reagent-treated samples. However, my result suggests that the control enrichment also contains the available information of base pairs.

Next, I assessed the accuracy of structure classification based on the reactivity scores with IDR filtering. To calculate the accuracy, the threshold for reactivity scores was progressively changed to classify each nucleotide into paired or unpaired. Then, the receiver operating characteristic curves (ROCs) were drawn to calculate an area under ROC (AUC) for 18S rRNA structure prediction and investigate the influence of filtering on its accuracy. The accuracy of icSHAPE and PARS score was similar among both types of dataset though PARS score



(a) (b)

Figure 58. The number of mapped reads to the reference of rRNA repeat unit in PARS (a) and icSHAPE (b) dataset, shown in each number of soft-clipped bases observed at their 5′ end by local alignment.

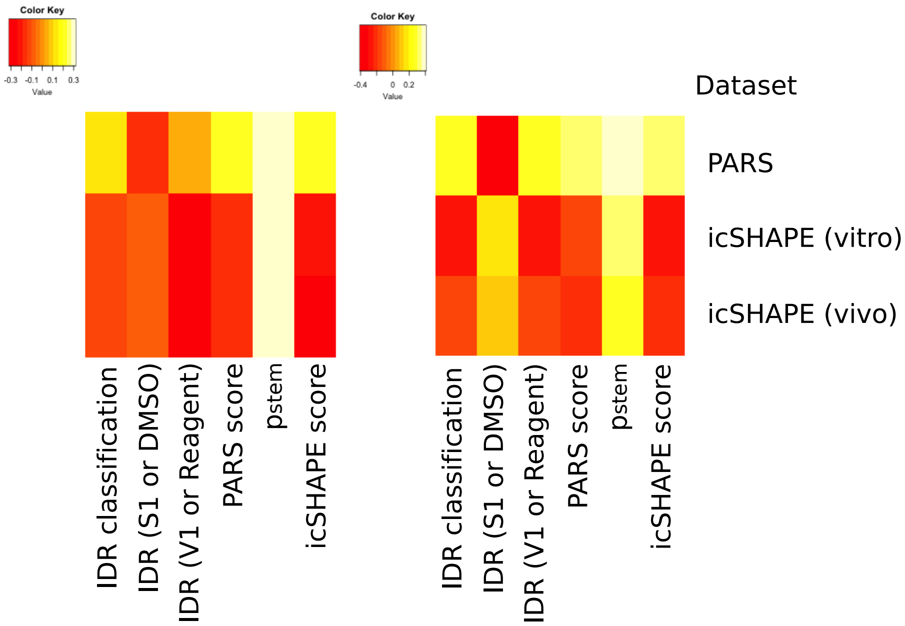
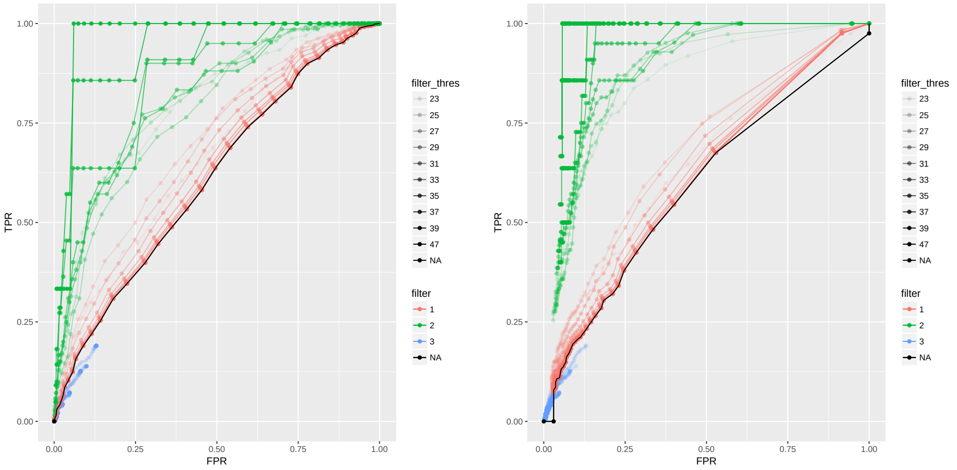


Figure 59. Correlation coefficient between the features of reactivity scores and reference structure of 18S rRNA for all (Left) and partial (Right) nucleotides with IDR filtering for the PARS, icSHAPE *in vitro*, and *in vivo* dataset. As a reference structure, a numerical vector was compared whose elements corresponding to each position are 1 at canonical base pairs, 0.5 at non-canonical base pairs, and 0 at unpaired bases. As IDR classification, each position was evaluated as -1 at accessible, 1 at structured, and 0 at neutral class, according to IDR-based reactivity classification.

exhibited slightly higher AUC in the PARS dataset and icSHAPE score did in icSHAPE dataset (Appendix Figure 72). To evaluate the influence of IDR filtering, I applied three types of filtering using IDRs for S1 (or DMSO-treated) samples and V1 (or reagent-treated) samples. The first filtering is to exclude the base pairs whose IDR for S1 (DMSO-treated) samples is inconsistent to the prediction of base pairing, from the positive set of prediction and reference. The second



(a)

(b)

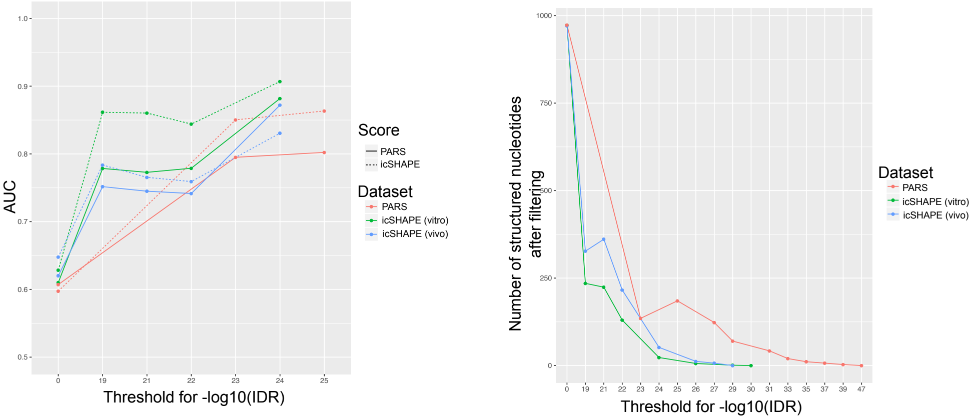
Figure 60. ROC of 18S rRNA structure prediction by the reactivity scores of the PARS dataset with or without IDR filtering. To compute the feature value, a scoring scheme defined in PARS (a) and icSHAPE research (b) was applied. To calculate the accuracy, the threshold for reactivity scores was progressively changed to classify each nucleotide into paired or unpaired (shown in black lines). In addition, three types of IDR filtering were applied to filter spurious signals (shown in red, green, and blue lines, respectively). A color strength corresponds to the magnitude of the threshold for IDR filtering ($-\log_{10}(\text{IDR})$).

is to exclude the base pairs whose at least either IDR for S1 (DMSO-treated) samples or V1 (reagent-treated) samples is inconsistent to the prediction, from the prediction and reference set. The third is to exclude the base pairs detected in the second filtering only from the set of prediction, and retain all of base pairs in the reference set.

Figure 60 shows the ROCs of the PARS and icSHAPE scores computed for the PARS dataset with or without IDR filtering. The first and second filtering drastically increased the accuracy compared to that without any filtering in regard to both of scorings regardless of the type of base pairs, such as canonical or non-canonical (Appendix Figure 73). These filters did not increase the accuracy of p_{stem}-based prediction, indicating that filtering could extract the regions with a reliable sequencing information (Appendix Figures 74, 75, and 76).

However, these filtering could not keep the same number of positive (base pair) dataset and it is difficult to unconditionally compare the accuracy between them. It is evidenced that the third filtering showed almost same or lower accuracy compared to the prediction for complete dataset. Nevertheless, IDR-based filtering can be considered useful to evaluate unreliability of sequencing data, leading the increase of robustness against such very sparse data.

Figure 61 is the summary of AUC changes according to the threshold for the second IDR filtering. Extraction of nucleotides whose IDR is less than 0.01 drastically improved AUCs to around 0.8, with more than a couple of hundreds bases still remained for those high AUC. Considering the AUC 0.73 of BUM-HMM for yeast 18S rRNA, IDR-based filtering is considered to have high performance to extract the regions whose structure can be predicted by reactivity scores well. In addition, the AUCs of *in vitro*-based analyses was generally higher than those of *in vivo*. This tendency is supposedly led from binding of RBPs or systematic biases appeared *in vivo*, resulting inaccurate measures of secondary structure.



(a)

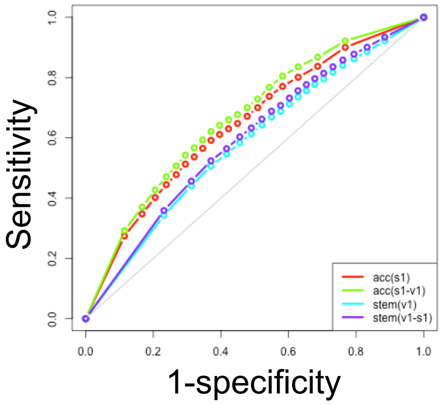
(b)

Figure 61. (a) AUCs computed for the reactivity classification for 18S rRNA with varying the threshold on $-\log_{10}(\mathrm{IDR})$ for the second type filtering. (b) The number of positive (structured) nucleotides in the reference with progressively changed thresholds.

3.5 MACHINE LEARNING-BASED CLASSIFICATION FOR COMPUTATIONAL PREDICTION AND REACTIVITY SCORES ACROSS HUMAN TRANSCRIPTOME

To validate the efficiency of IDR-based filtering to promote consistency between high-throughput structure analyses and computational prediction, I calculated the precision and specificity for the prediction of accessible and structured IDR-based classification by p_{stem}. Figure 62 shows ROCs with varying a threshold on p_{stem} to measure the prediction capability of IDR-based classification. The accuracy of p_{stem}-based classification was higher for the positive set of accessible classes (represented as acc) than structured ones (stem). In addition, the exclusion of neutral samples considered to be inconsistent samples slightly increased the accuracy of prediction for "structured" and "accessible" classes. This tendency of p_{stem} distribution was indicated to agree with the strength of reproducibility of $c_{\mathrm{PARS\_V1}}$ and irreproducibilty of $c_{\mathrm{PARS\_S1}}$. Therefore, it is suggested that IDR-based filtering can extract the regions in which the computational method also predicts fairly-stable structures.

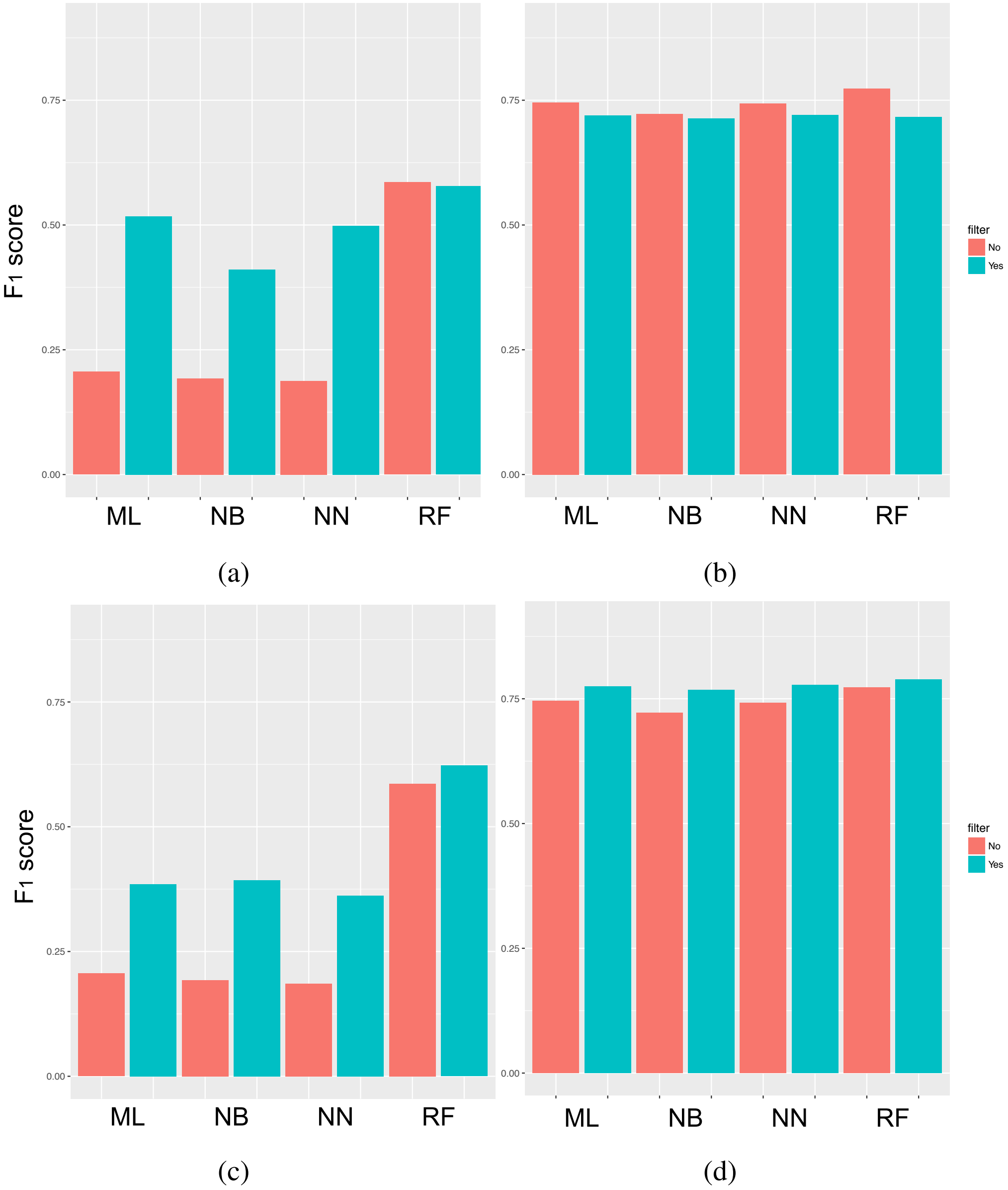
Finally, the consistency of high-throughput structure analyses and computational prediction was examined for the PARS dataset using machine learning-based classification. The accuracy of the predictions was evaluated by five-fold cross validation on the dataset of true structured and accessible regions according to the p_{stem} threshold ($p_{\mathrm{stem}} > 0.9$ or $p_{\mathrm{stem}} > 0.5$ in structured regions) for the four types of classifiers optimized for the test set (detailed in the Methods section). Figure 63 shows F_1 scores of each classifier using two features of PARS scores and icSHAPE scores from the set of possible features. These two features were extracted due to the large correlations with p_{stem} (data not shown). The accuracies of F_1 scores in the case of the modest threshold for p_{stem} are higher (around 0.75) than those with the strict threshold (at most 0.63). However, the accuracies derived from the case of modest thresholds were substantially improved after the second type of IDR-based filtering (described in the previous section). Therefore, these results indicate that IDR-based filtering might be able to remove the regions of irreproducible, but extremely high reactivity scores. Furthermore, as the test set is derived from computational prediction, it is indicated that the sequencing-based features could adequately discriminate the position of highly stable base pairs in theory. Additionally,



(a)

Figure 62. ROCs for the precision accuracy of IDR-based classification by stem probability for human transcriptome. Red, green, blue, and purple lines indicate the prediction accuracy of p_{stem} for "accessible", "accessible and not structured", "structured", and "structured and not accessible" positions, respectively.

IDR estimation based on p-values instead of raw read coverages could increase the precision combined with IDR-based filtering. As human transcriptome consists of a variety of RNAs in terms of their expression levels, computation of p-values may be the assist to avoid the influence of biases.



(a)

(b)

(c)

(d)

Figure 63. F_1 scores for structure classification of human transcriptome based on ML, NN, RF, and NB approach using the features of PARS and icSHAPE score for the PARS dataset with or without filtering. Each figure shows averaged F_1 scores by five-fold cross validation with the different condition. A threshold for p_{stem} to classify each nucleotide into structure or accessible class was set to 0.9 in (a) and (c), and 0.5 in (b) and (c). IDR was computed for $C_{\text{PARS\_condition}}$ in (a) and (b) while it was done for p-values derived from Poisson distribution of $C_{\text{PARS\_condition}}$ in (c) and (d).

4 DISCUSSION AND CONCLUSIONS

MULTI-COMPARISON OF HIGH-THROUGHPUT STRUCTURE ANALYSES

In this study, a surprising result was shown that high-throughput structure analyses for human transcriptome were highly different from each other. However, it is not clear whether such incongruence arises from the heterogeneity of human transcriptome as related to the cellular condition or the systematic biases of each method. In fact, RNA secondary structure is known to be variable depending on environmental elements, such as temperature, pH, or interference of binding factors. For example, it is known that riboswitches change the structure drastically depending on the concentration of substrates in the cell, and thermometers respond to temperature by alternating structures [120]. However, since replicates of high-throughput structure analyses showed substantially high correlation in this study, it is considered that most transcripts form a consensus secondary structure that is stable enough to be detected by probing or cleavage. In addition, it is a typical tendency that the comparison of omics data requires careful normalization to avoid the influence of systematic biases across studies conducted in different laboratories [121]. Therefore, reactIDR can contribute to robust inference of positional reactivity based on only highly reproducible read counts, such as motif findings of stable structures in the living cell.

POTENTIAL FEATURES OF PEAK CLUSTERING FOR READ COVERAGES

For computational prediction of RNA secondary structure, reactivity information of high-throughput analyses is effectively used as a guide to the loop and base pair for each nucleotide to improve prediction accuracy. By assessing the confidence of each reactivity score, it may be possible to remove unreliable guides that may decrease the prediction accuracy. At the same time, an information guide to regional accessibility would be more informative than one on single-nucleotide resolution, because a longer loop may impose strong restriction on possible structures. Actually, Figure 56 shows the tendency of peak-like profiles around reproducible enrichment of PARS data. In addition, regional accessibility might be instructive for computational prediction as the determination of a long loop can enhance the overall accuracies of structure prediction.

An inference of IDR has already been applied for robust peak detection of protein-binding sites in ChIP-Seq and transcription start sites in CAGE-Seq [122, 108]. To my knowledge, this study is the first attempt to extend IDR for the estimation of reproducibility of high-throughput structure analyses. According to my analyses, the profile of read coverage on reproducible positions has a context dependence similar to that of ChIP-Seq and CAGE-Seq. The previous study on BUM-HMM took account of surrounding p-values through HMM modeling to improve the accuracy. Therefore, it may be promising to use clustered read coverage profiles for robust structure analyses in terms of reliability and precision improvements of computational prediction.

PREDICTION ACCURACY OF REACTIDR AND OTHER TOOLS

In this study, I confirmed the efficiency of IDR-based filtering for reliable structure classification. However, it remains to be assessed whether filtering of other features is efficient for increasing classification accuracy. The correlation between multiple features themselves should be concerned not to extract the sufficient minimal feature set for filtering. Another direction for future research is that reactIDR should be compared with the previous methods for robust evaluation of high-throughput structure analyses [105, 107, 102, 123, 106, 96]. Although BUM-HMM is a statistical model which can evaluate replicates' consistency, these previous methods showed only a very low positive ratio in the Ref. [102]. Because reactIDR can handle a combination of features, it is possible to apply additional features that performed well in the previous study. For that reason, reactIDR may constitute a state-of-the-art software tool to analyze genome-wide structure analyses in a flexible manner.

CONCLUSIONS

My novel pipeline reactIDR enables the extraction of reliable reactivities from multiple high-throughput structure analyses. To evaluate the reliability of each reactivity score, the IDR is computed by modeling the joint probability distribution among replicates. reactIDR can also compute p-values based on the null distribution of Poisson or negative binomial distribution, considering a different sequencing depth for each RNA. In addition, construction of a structure classifier is performed for user-defined reference structures, using various features of mapped read coverage profiles. According to my analyses, a combination of computational prediction and IDR-based classification showed higher accuracy for the prediction of reference structure compared to reactivity scores, as used in previous studies. In addition, the result of transcriptome-wide computational prediction was congruence with both IDR-based classification and reactivity scores of high-throughput structure analyses. Therefore, it is suggested that a combination of computational and experimental genome-wide structure analyses has a promising potential to infer the global landscape of RNA secondary structure.

CONCLUSIONS

In this thesis, I present two novel approaches, ParasoR and reactIDR, to integrate *in vivo*, *in vitro*, and *in silico* RNA secondary structure analyses. These approaches have high feasibility for genome-wide structure analyses as well as limited targets such as the family of non-coding RNA.

My novel software "ParasoR" is the first algorithm which can handle genome-scale *in silico* RNA secondary structure analyses, and compute expected values about globally consistent structures with the constraint of maximal span. Using ParasoR and k-mer regression method extracted structure profiles of human transcripts and inferred the genome-wide structure propensity beyond sequence composition biases. In addition, a comparison between pre-mRNAs and mRNAs suggested that coding regions become more accessible after splicing, presumably because of biological constraints such as translational efficiency.

Moreover, my novel pipeline reactIDR enables to extract reliable reactivities from multiple high-throughput structure analyses applied *in vitro* and *in vivo*. To evaluate the reliability of each reactivity score, the irreproducible discovery rate is computed by modeling the joint probability distribution among replicates. Prediction of transcriptome-wide reactivity classification was congruence between IDR-based classification and reactivity scores of high-throughput structure analyses.

As explored in the first part, the structure profiles inferred by ParasoR showed a high concordance with high-throughput sequencing analyses when the threshold for the minimum read depth is set to be strict. However, there still remain significant differences between computational and high-throughput structure analyses for the specific regions. Such differences are also observed between *in vivo* and *in vitro* structure propensity, supposedly due to the binding of other molecules, modification, or any other unknown factors. Therefore, it is suggested that a multi-comparison between computational and experimental genome-wide structure analyses has a promising potential to improve their capability to capture the biological phenomena beyond systematic biases, shed the light on the meaningful differences between the theory and observation, and infer the true landscape of RNA secondary structure.

Appendix

LIST OF ABBREVIATIONS

INTRODUCTION

- DP: dynamic programming
- IDR: irreproducible discovery rate
- RT: reverse transcription
- DMS: dymethyl sulfate

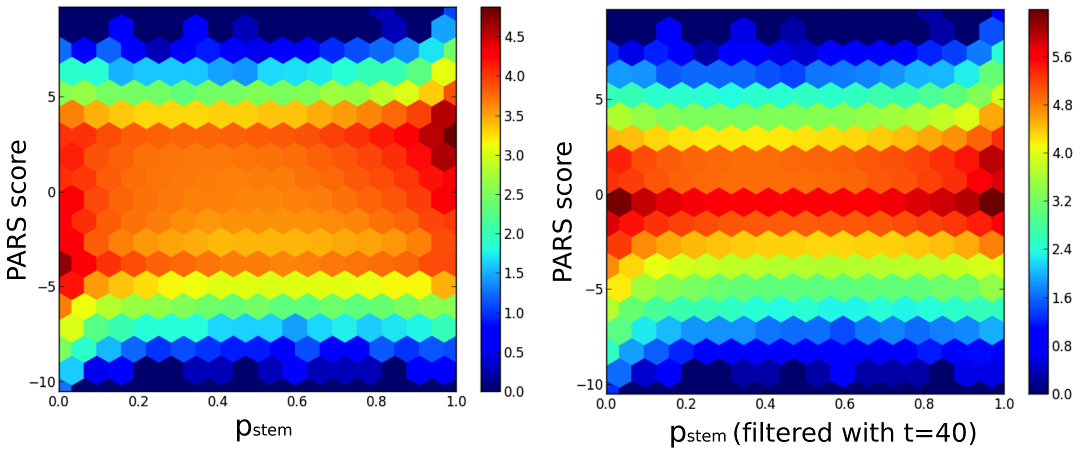
PARALLEL COMPUTATION OF GENOME-SCALE RNA SECONDARY STRUCTURE TO DETECT STRUCTURAL CONSTRAINTS ON HUMAN GENOME

- SS: splice site
- AS: alternative splicing
- DP: dynamic programming
- MFE: minimum free energy
- ROC: receiver operating characteristic curve
- AUC: area under receiver operating characteristic curve
- MCC: Matthews correlation coefficient
- PARS: parallel analysis of RNA structure
- CDS: coding sequence
- UTR: untranslated region

STATISTICAL APPROACH TO ROBUST RNA REACTIVITY CLASSIFICATION BASED ON REPRODUCIBLE HIGH-THROUGHPUT STRUCTURE ANALYSES

- PARS: parallel analysis of RNA structure
- SHAPE: Selective 2′-hydroxyl acylation analyzed by primer extension
- icSHAPE: *in vivo* click selective 2′-hydroxyl acylation and profiling experiment
- DMS: dymethyl sulfate
- 1M7: 1-methyl-7-nitroisatoic anhydride
- NMIA: N-methylisatoic anhydride

- MaP: mutational profiling
- IDR: irreproducible discovery rate
- HMM: hidden Markov model
- MCMC: Markov chain Monte Carlo (MCMC)
- RF: random forest
- KSVM: kernel support vector machine
- NN: neural network
- ML: multinomial logistic regression
- NB: naïve bayes
- EM: expectation-maximization
- ROC: receiver operating characteristic curve
- AUC: area under receiver operating characteristic curve



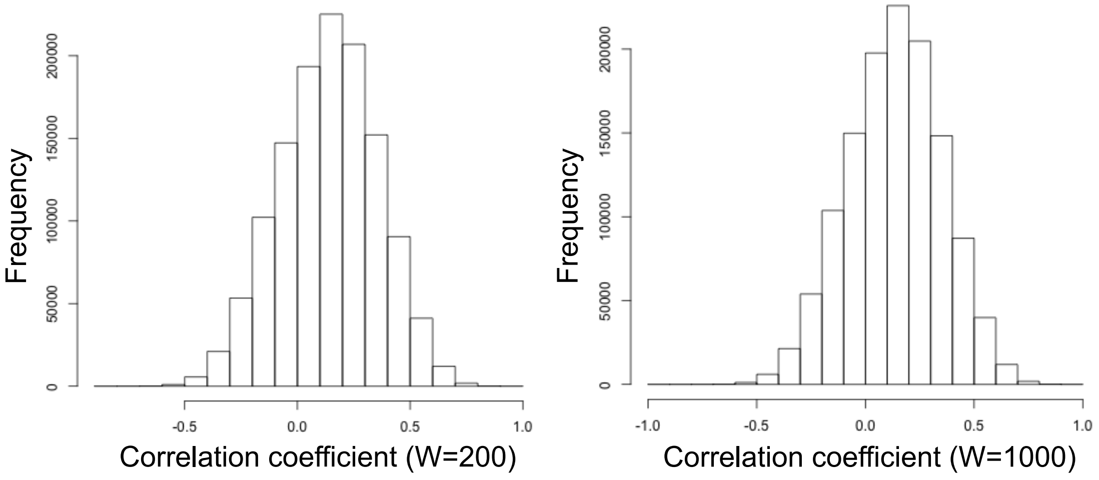
(a) (b)

Figure 64. Relationship between ParasoR and PARS scores with or without filtering. Hexagonal binning plots for PARS scores (shown in y-axis) and p_{stem} (shown in x-axis) of all positions (a) and filtered positions (b) with the threshold $t = 40$ for the minimum read coverage of PARS score.

General tendency of conformational changes between mRNA and pre-mRNA

Table 10. Statistical test of substantial conformational changes for each annotation category in the human and mouse genome

species	human				mouse			
threshold	*p*-value of CDS	non-coding exon	5′-UTR	3′-UTR	*p*-value of CDS	non-coding exon	5′-UTR	3′-UTR
0.00	*0.0	$*2.53\times10^{-244}$	$*3.53\times10^{-147}$	0.42	*0.0	$*2.37\times10^{-104}$	$*1.66\times10^{-108}$	$*1.08\times10^{-8}$
0.05	*0.0	*0.0	$*3.87\times10^{-209}$	*0.00032	*0.0	$*3.92\times10^{-55}$	$*3.84\times10^{-140}$	$*7.60\times10^{-7}$
0.10	*0.0	*0.0	$*2.98\times10^{-189}$	$*1.37\times10^{7}$	*0.0	$*1.96\times10^{62}$	$*1.01\times10^{-114}$	0.00121
0.15	*0.0	$*2.80\times10^{-259}$	$*1.50\times10^{-130}$	$*1.68\times10^{-8}$	*0.0	$*1.71\times10^{55}$	$*4.61\times10^{-84}$	0.00876
0.20	*0.0	$*7.45\times10^{-173}$	$*1.48\times10^{-79}$	0.0017	*0.0	$*8.60\times10^{42}$	$*1.66\times10^{-50}$	0.0149
0.25	*0.0	$*3.55\times10^{-105}$	$*1.22\times10^{-20}$	0.019	*0.0	$*1.26\times10^{26}$	$*1.05\times10^{-27}$	*0.000460
0.30	*0.0	$*1.16\times10^{-61}$	$*8.50\times10^{-10}$	0.054	*0.0	$*2.14\times10^{13}$	$*3.30\times10^{-7}$	*0.000117
0.35	*0.0	$*3.67\times10^{-26}$	0.0061	0.093	$*1.90\times10^{279}$	$*1.78\times10^{7}$	$*4.59\times10^{-5}$	0.00642
0.40	$*4.46\times10^{-279}$	$*4.00\times10^{-15}$	0.51	0.40	$*5.09\times10^{-155}$	$*6.16\times10^{-5}$	0.00873	0.123
0.45	$*7.26\times10^{-164}$	$*4.08\times10^{-7}$	0.69	0.35	$*7.03\times10^{-90}$	0.0107	0.0918	0.109
0.50	$*2.60\times10^{-87}$	$*7.29\times10^{-5}$	0.27	0.11	$*1.25\times10^{-54}$	0.228	0.217	1.0



(a) (b)

Figure 65. Correlation coefficient between ParasoR and PARS scores. Histograms of correlation coefficient between p_{stem} (W = 200 (a) and 1,000 (b)) and PARS score within the non-overlapping window of 20-nt length

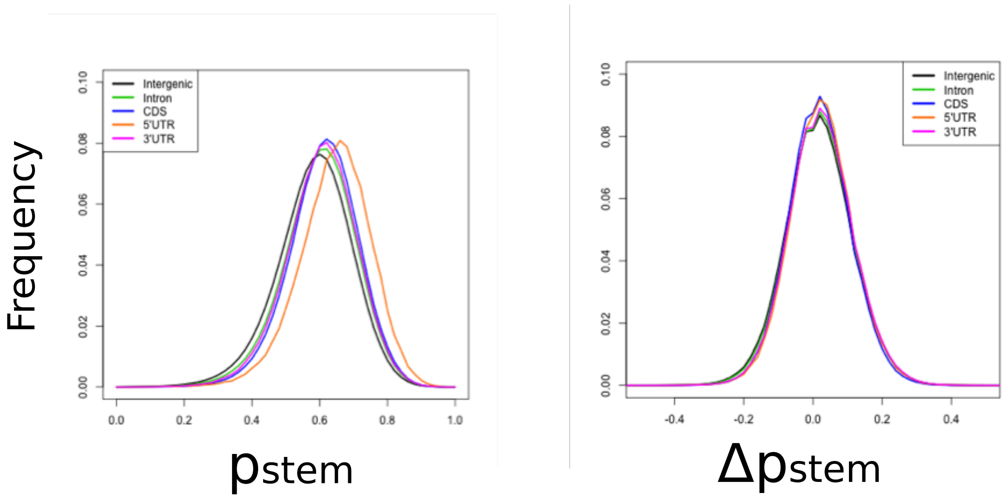


Figure 66. Histograms of raw stem probability (Left) and normalized stem probability (Right) for mouse genome.

Table 11. Correlations between conformational changes and gene features

CC (p-value)	Maximum	Minimum
Expression	−0.018**	0.008
mRNA GC %	0.021**	−0.012**
Intron length	−0.006*	0.004

We tested statistical significance by using Pearson's correlation test. (*: $p < 0.05$, **: $p < 1e-3$ after Bonferroni multiple correction for 6 samples.)

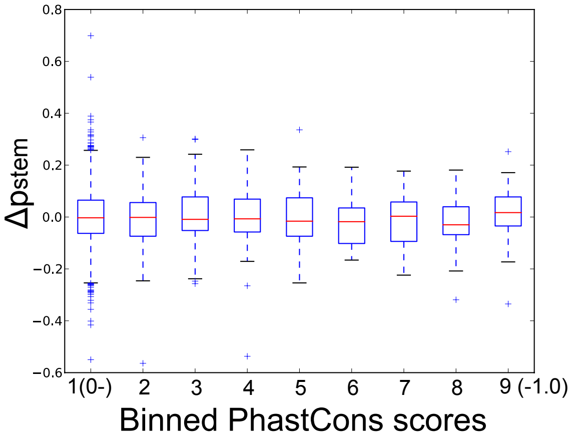


Figure 67. Relationship between structure propensity and conservation score. Boxplots of $\Delta\bar{p}_{\text{stem}}$ binned by the conservation score of each fragment. X-axis shows 9 groups classified by PhastCons score in steps of 1/9 from 0 to 1. Between PhastCons score and $\Delta\bar{p}_{\text{stem}}$, we detected no significant correlation ($p > 0.05$).

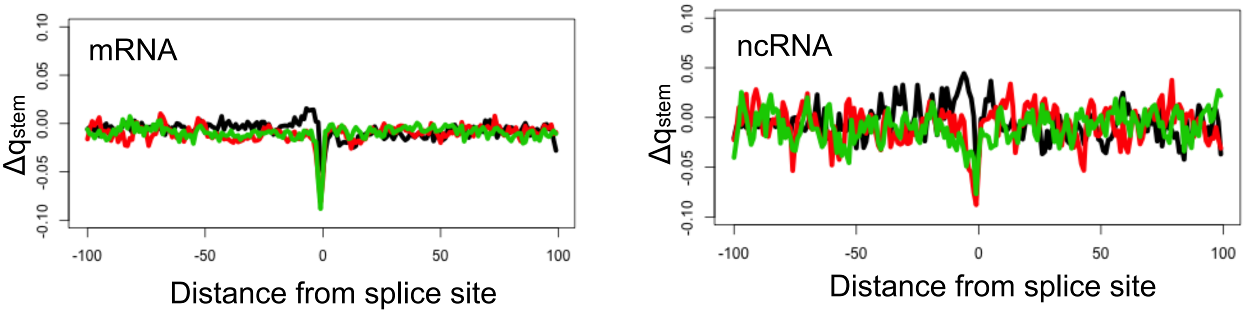


Figure 68. Regional profiles of difference of stem probability. Mean differences of stem probability were computed around the exon junction for mRNA (Left) and non-coding RNA (Right) between pre-mRNA and mRNA for the 1st, 2nd, and 3rd splice site (represented by a black, red, and green line, respectively).

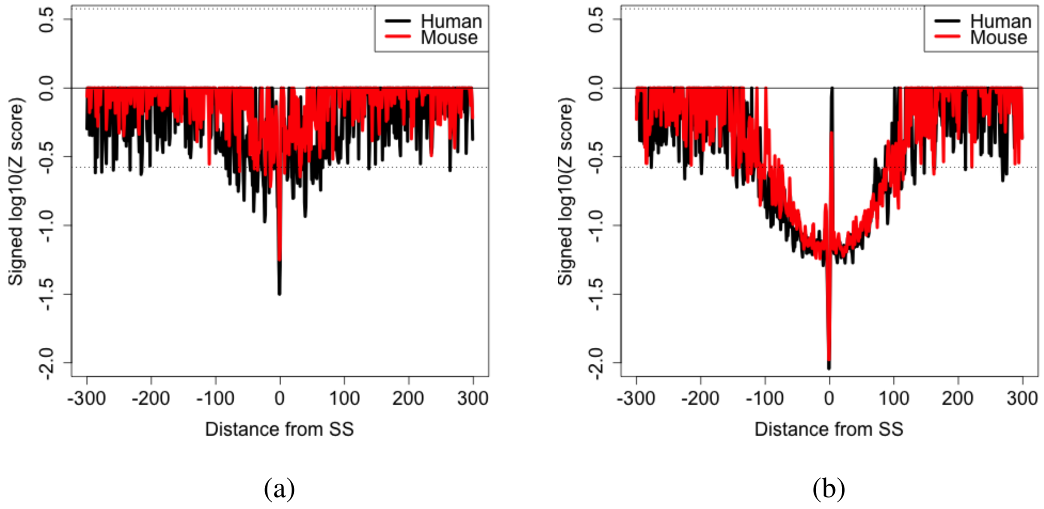


Figure 69. Structure disruption caused by splicing. Z-score profiles by Wilcoxon¨ s signed rank test for difference of stem probability around splice site of mRNA (a) and non-coding RNA (b).

Table 12. Enriched GO terms among post-accessible and post-structural genes in mouse by Ontologizer

GO	term	*p*-value
GO:0005488 (F)	binding	2.14e-5
GO:0006996 (P)	organelle organization	0.024
GO:0044422 (C)	organelle part	0.024
GO:0071840 (P)	cellular component organization or biogenesis	0.034
GO:0043226 (C)	organelle	0.034
GO:0016817 (F)	hydrolase activity, acting on acid anhydrides	0.034
GO	term	*p*-value
GO:0003824 (F)	catalytic activity	0.016
GO:0005524 (F)	ATP binding	0.016
GO:0006793 (P)	phosphorus metabolic process	0.016
GO:0043412 (P)	macromolecule modification	0.016
GO:0032559 (F)	adenyl ribonucleotide binding	0.016
GO:0036094 (F)	small molecule binding	0.016
GO:0030554 (F)	adenyl nucleotide binding	0.016
GO:0019992 (F)	diacylglycerol binding	0.020
GO:0097367 (F)	carbohydrate derivative binding	0.027
GO:0034660 (P)	ncRNA metabolic process	0.028

IDR-HMM ALGORITHM

Computation of pseudo values

To compute cumulative standard distribution, the definition of error function (erf) is used in reactIDR.

$$\begin{aligned}
\mathrm{erf}(x) =& \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \\
=& \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \\
=& \frac{2}{\sqrt{2\pi}} \int_0^{x\sqrt{2}} e^{-\frac{p^2}{2}} dp \\
=& 2\left(\frac{1}{\sqrt{2}} \int_{-\infty}^{x\sqrt{2}} e^{-\frac{p^2}{2}} dp - \int_{-\infty}^{0} e^{-\frac{p^2}{2}} dp\right) = \quad 2\left(\Phi(\sqrt{2}x) - \frac{1}{2}\right) \\
(\mathrm{erf}(x))' =& \frac{2}{\sqrt{\pi}} e^{-x^2} \\
\Phi(x) =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2} dt \\
=& \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right)
\end{aligned}$$

where $\phi(x)$ and $\Phi(x)$ is a standard normal density and standard normal cumulative distribution function, respectively.
$\frac{\partial F_k^{-1}(u_k)}{\partial u_k}$ $\left(\frac{\partial x_k}{\partial u_k}\right)$ can be calculated as follow:

$$\begin{aligned}
F_k(F^{-1}(u_k)) =& u_k \\
\frac{\partial F_k(F^{-1}(u_k))}{\partial u_k} =& \frac{\partial F_k}{\partial x_k}(x_k)\frac{\partial F^{-1}}{\partial u_k} = 1 \\
\frac{\partial F_k^{-1}}{\partial u_k} =& \frac{1}{\partial F_k / \partial x_k} = \frac{1}{f_k(x_k)} \\
=& \frac{1}{f_k(F_k^{-1}(u_k))}
\end{aligned}$$

Differential of joint probability distribution

Using chain rule of differentiation, these equations hold.

$$\frac{df(g(x))}{x} = \left.\frac{\partial f}{\partial x}\right|_{x=g(x)} \frac{\partial g}{\partial x}$$

$$\begin{aligned}\frac{df(F_{1,\theta}^{-1}(u_1), F_{2,\theta}^{-1}(u_2))}{d\theta} &= \lim_{\Delta\theta\to 0} \frac{f_{\theta+\Delta\theta}(F_{1,\theta+\Delta\theta}^{-1}(u_1), F_{2,\theta+\Delta\theta}^{-1}(u_2)) - f_\theta(F_{1,\theta}^{-1}(u_1), F_{2,\theta}^{-1}(u_2))}{\Delta\theta} \\ &= \lim_{\Delta\theta\to 0} \frac{f_{\theta+\Delta\theta}(F_{1,\theta}^{-1}(u_1) + \Delta\theta\frac{\partial F_1^{-1}}{\partial\theta}, F_{2,\theta}^{-1}(u_2) + \Delta\theta\frac{\partial F_2^{-1}}{\partial\theta} - f_\theta(F_{1,\theta}^{-1}(u_1), F_{2,\theta}^{-1}(u_2))}{\Delta\theta} \\ &= \lim_{\Delta\theta\to 0} \frac{\partial f(F_{1,\theta}^{-1}(u_1), F_{2,\theta}^{-1}(u_2))}{\partial\theta} + \frac{\partial F_1^{-1}(u_1)}{\partial\theta}\frac{\partial f}{\partial x_1} + \frac{\partial F_1^{-1}(u_2)}{\partial\theta}\frac{\partial f}{\partial x_2} + O(\Delta\theta) \\ &= \frac{\partial f(F_{1,\theta}^{-1}(u_1), F_{2,\theta}^{-1}(u_2))}{\partial\theta} + \frac{\partial F_1^{-1}(u_1)}{\partial\theta}\frac{\partial f}{\partial x_1} + \frac{\partial F_1^{-1}(u_2)}{\partial\theta}\frac{\partial f}{\partial x_2}\end{aligned} \tag{10}$$

COMMAND LINES FOR PARS READ MAPPING

For the PARS dataset, only the first R1 reads truncated to the first 50 bases were mapped with the allowance of one mismatch using bowtie2. A box below shows an example of command lines for the PARS data processing.

```
cat temp_1.fq \\
| fastx_clipper -a gatcggaagagcggttcagcaggaatgccgag -q33 -m 9 -o
   ↪ temp_trimmed.fq \\
| fastx_trimmer -Q33 -l 50 -i temp.fq -o temp_trimmed_50.fq
bowtie2 -p 8 -N 1 -x (rRNA index) -U temp_trimmed_50.fq | samtools view -
   ↪ Shb -F 4 - > accepted_hits_rRNA.bam
bowtie2 -p 8 -N 1 --local -x (rRNA index) -U temp_trimmed_50.fq | samtools
   ↪ view -Shb -F 4 - > accepted_hits_rRNA_local.bam
bowtie2 -p 8 -N 1 -x (transcriptome index) -U temp_trimmed\_50.fq.gz |
   ↪ samtools view -Sb > accepted_hits.bam
```

For the icSHAPE dataset, the sequencing reads that are completely duplicated were removed as PCR duplications before barcode truncation, because the number of barcodes incorporated in icSHAPE is large enough that the change to two random fragments to be concatenated with the same barcode is very low ($\frac{1}{4^9}$) [96]. Sequenced reads whose first 13 nucleotides were truncated, were mapped to the reference of transcriptome and rRNA repeat unit using local alignment of bowtie2. The number of reads assigned to each position was counted allowing a single flaking nucleotide at their 5′-end. A box below shows an example of command lines used for icSHAPE data processing.

COMMAND LINES FOR ICSHAPE READ MAPPING

```
python read_collapse.py --shell --remove --stdout temp.fq > temp\
   ↪ _multiplexed.fq
java -jar trimmomatic-0.36.jar SE -threads 2 -phred33 temp\_multiplexed.fq
   ↪ temp\_read\_trimmed.fq ILLUMINACLIP:../adapters/TruSeq2-SE.fa:2:30:10
   ↪  TRAILING:20 HEADCROP:13
bowtie2 --local -p 8 -x (transcriptome index) -U temp\_read\_trimmed.fq |
   ↪ samtools view -Shb -F 4 - > accepted\_hits.bam
```

Here is the list of the software tools and their versions used in this study.

1. fastx_toolkit v0.0.13
2. bowtie2 v2.2.9
3. trimmamotic v0.36.
4. samtools v1.3.1
5. bedtools v2.18

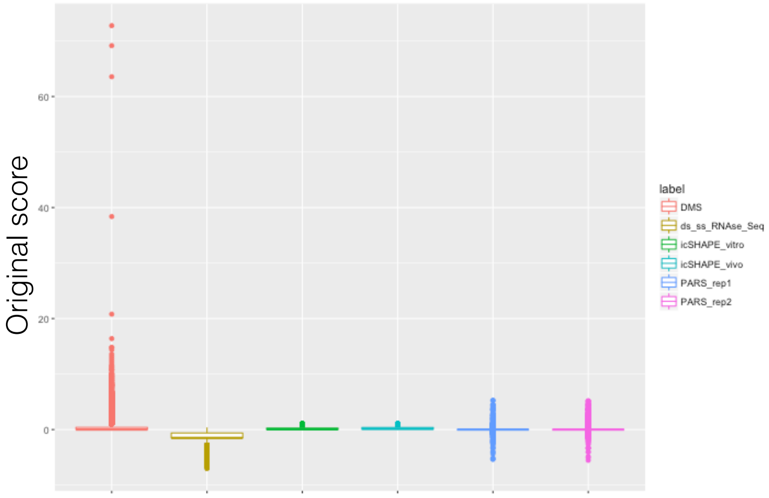
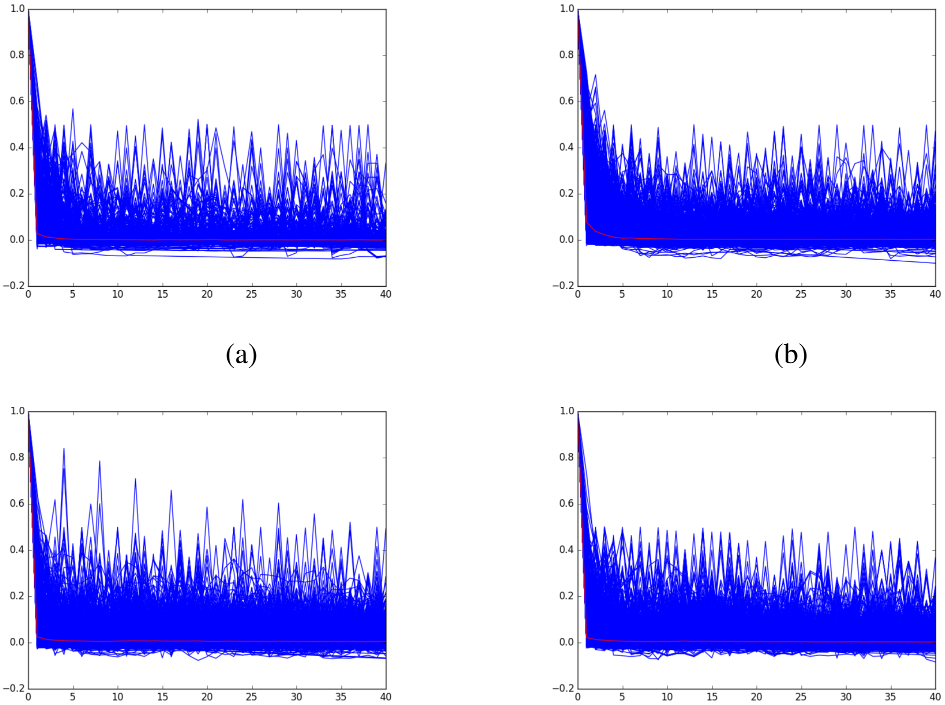


Figure 70. Distribution of reactivity scores achieved in the Structure Surfer [90]. Each distribution is normalized (detailed in the Methods section).



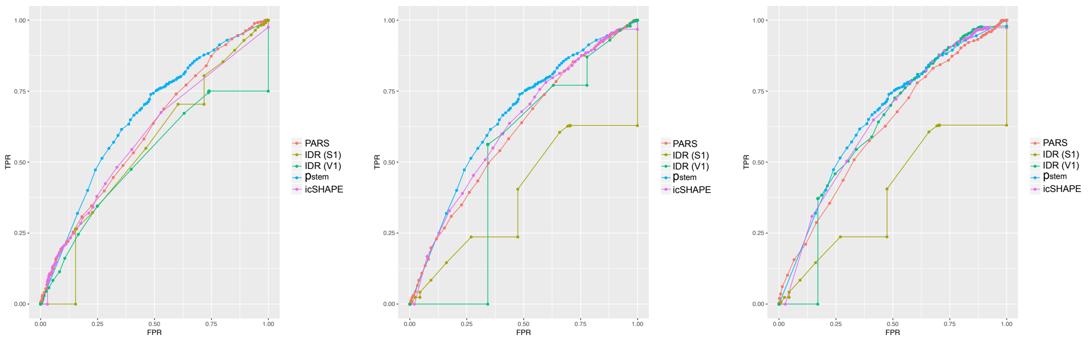
(a)

(b)

(c)

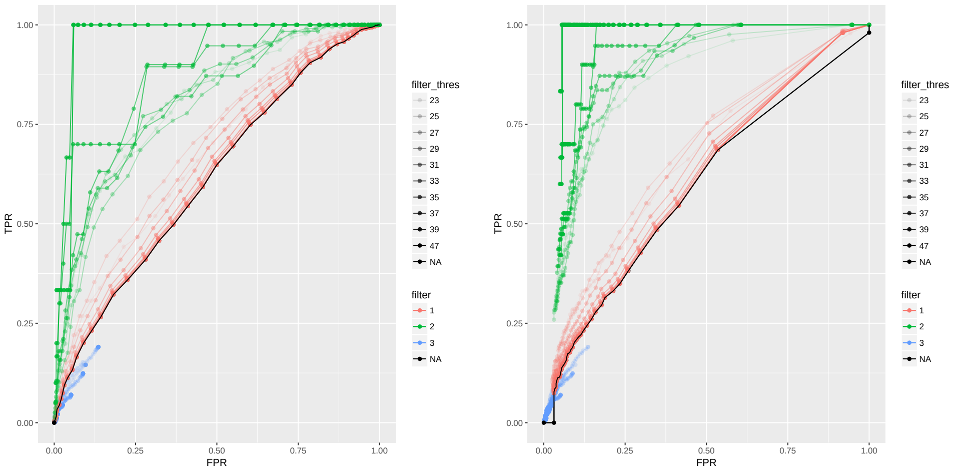
(d)

Figure 71. Autocorrelation of read counts of PARS score for S1 rep1 (a), S1 rep2 (b), V1 rep1 (c), and V1 rep2 (d). Y- and X-axes show autocorrelation and the time lag (distance of nucleotides) used for calculation of autocorrelation, respectively. Blue lines correspond to the autocorrelation of each transcript, and red lines exhibit the mean of autocorrelation for each position.



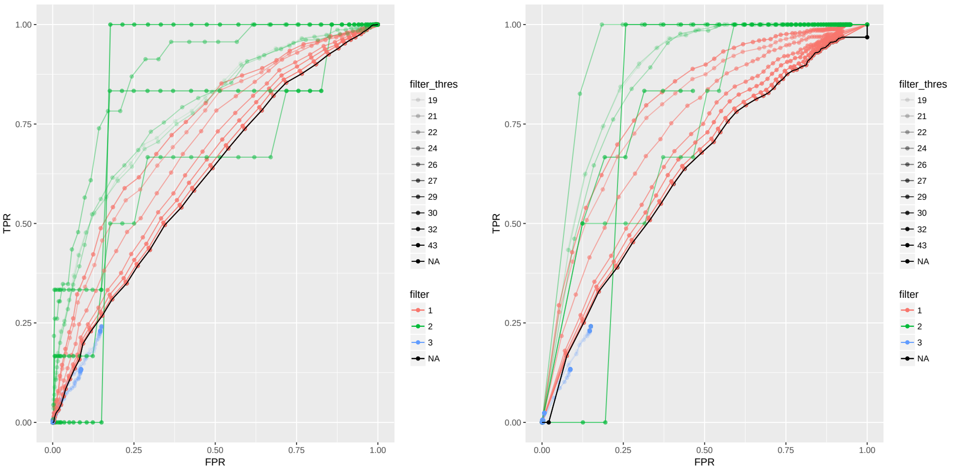
(a) (b) (c)

Figure 72. ROCs of 18S rRNA structure prediction by the reactivity scores of the PARS (a), icSHAPE *in vitro* (b), and icSHAPE *in vivo* (c) dataset. To compute the feature value, six types of scoring schemes were applied: scoring defined in PARS, icSHAPE, IDR for the S1 (DMSO-treated) dataset, IDR for the V1 (reagent-treated) dataset, and p_{stem} computed by ParasoR. To evaluate an overall accuracy, the threshold for reactivity scores was progressively changed to classify each nucleotide into paired or unpaired.



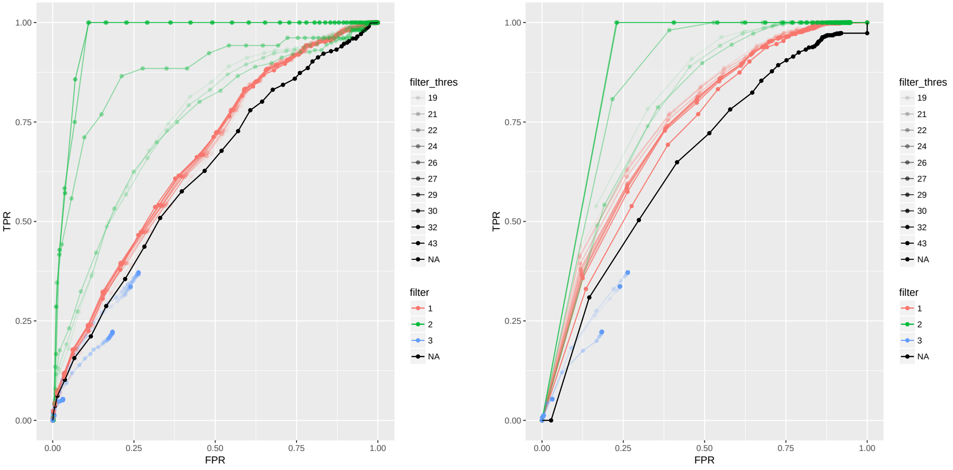
(a) (b)

Figure 73. ROCs of 18S rRNA structure prediction only about canonical bases by the reactivity scores of the PARS dataset with or without IDR filtering, To compute the feature value, a scoring scheme defined in PARS (a) and icSHAPE research (b) was applied. To calculate the accuracy, the threshold for reactivity scores was progressively changed to classify each nucleotide into paired or unpaired (shown in black lines). In addition, three types of IDR filtering were applied to filter spurious signals (shown in red, green, and blue lines, respectively). A color strength corresponds to the magnitude of the threshold for IDR filtering ($-\log_{10}(\text{IDR})$).



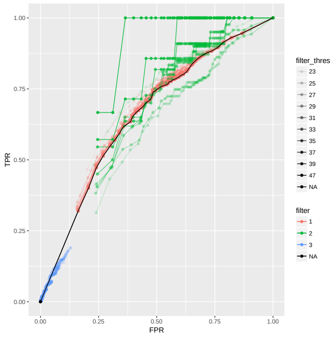
(a) (b)

Figure 74. ROCs of 18S rRNA structure prediction by the reactivity scores of the icSHAPE *in vitro* dataset with or without IDR filtering as done in Figure 73.



(a) (b)

Figure 75. ROCs of 18S rRNA structure prediction by the reactivity scores of the icSHAPE *in vivo* dataset with or without IDR filtering, as done in Figure 74.



(a)

Figure 76. ROCs of 18S rRNA structure prediction by p_{stem} for PARS dataset with or without IDR filtering as done in Figure 75. To calculate the accuracy, the threshold for p_{stem} was progressively changed to classify each nucleotide into paired or unpaired.

[1] S.-I. Consortium *et al.*, “A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium,” *Nature biotechnology*, vol. 32, no. 9, pp. 903–914, 2014.

[2] K. Chandramouli and P.-Y. Qian, “Proteomics: challenges, techniques and possibilities to overcome biological sample complexity,” *Human genomics and proteomics*, vol. 1, no. 1, 2009.

[3] E. P. Geiduschek, J. W. Moohr, and S. B. Weiss, “The secondary structure of complementary RNA,” *Proceedings of the National Academy of Sciences*, vol. 48, no. 6, pp. 1078–1086, 1962.

[4] S. Mitra, M. Enger, and P. Kaesberg, “Physical and chemical properties of RNA from the bacterial virus R17,” *Proceedings of the National Academy of Sciences*, vol. 50, no. 1, pp. 68–75, 1963.

[5] T. Fukunaga, H. Ozaki, G. Terai, K. Asai, W. Iwasaki, and H. Kiryu, “CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data,” *Genome Biol*, vol. 15, no. 1, p. R16, 2014.

[6] N. R. Zearfoss, E. S. Jahnson, and S. P. Ryder, “hnRNP A1 and secondary structure coordinate alternative splicing of Mag,” *RNA*, vol. 19, no. 7, pp. 948–957, 2013.

[7] T. Tuller, Y. Y. Waldman, M. Kupiec, and E. Ruppin, “Translation efficiency is determined by both codon bias and folding energy,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 8, pp. 3645–3650, 2010.

[8] A. E. Borujeni, A. S. Channarasappa, and H. M. Salis, “Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites,” *Nucleic acids research*, p. gkt1139, 2013.

[9] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, “The role of site accessibility in microRNA target recognition,” *Nature genetics*, vol. 39, no. 10, pp. 1278–1284, 2007.

[10] M. Zuker *et al.*, “On finding all suboptimal foldings of an RNA molecule,” *Science*, vol. 244, no. 4900, pp. 48–52, 1989.

[11] J. S. McCaskill, “The equilibrium partition function and base pair binding probabilities for RNA secondary structure,” *Biopolymers*, vol. 29, no. 6-7, pp. 1105–1119, 1990.

[12] S. R. Eddy and R. Durbin, “RNA sequence analysis using covariance models,” *Nucleic acids research*, vol. 22, no. 11, pp. 2079–2088, 1994.

[13] M. Rabani, M. Kertesz, and E. Segal, “Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 39, pp. 14885–14890, 2008.

[14] A. F. Bompfünewerer, C. Flamm, C. Fried, G. Fritzsch, I. L. Hofacker, J. Lehmann, K. Missal, A. Mosig, B. Müller, S. J. Prohaska, *et al.*, "Evolutionary patterns of non-coding RNAs," *Theory in Biosciences*, vol. 123, no. 4, pp. 301–369, 2005.

[15] P. J. Shepard and K. J. Hertel, "Conserved RNA secondary structures promote alternative splicing," *Rna*, vol. 14, no. 8, pp. 1463–1469, 2008.

[16] I. L. Hofacker, B. Priwitzer, and P. F. Stadler, "Prediction of locally stable RNA secondary structures for genome-wide surveys," *Bioinformatics*, vol. 20, no. 2, pp. 186–190, 2004.

[17] Y. Horesh, Y. Wexler, I. Lebenthal, M. Ziv-Ukelson, and R. Unger, "RNAslider: a faster engine for consecutive windows folding and its application to the analysis of genomic folding asymmetry," *BMC bioinformatics*, vol. 10, no. 1, p. 76, 2009.

[18] C. K. Kwok, Y. Tang, S. M. Assmann, and P. C. Bevilacqua, "The RNA structurome: transcriptome-wide structure probing with next-generation sequencing," *Trends in biochemical sciences*, vol. 40, no. 4, pp. 221–232, 2015.

[19] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal, "Genome-wide measurement of RNA secondary structure in yeast," *Nature*, vol. 467, no. 7311, pp. 103–107, 2010.

[20] Y. Wan, K. Qu, Q. C. Zhang, R. A. Flynn, O. Manor, Z. Ouyang, J. Zhang, R. C. Spitale, M. P. Snyder, E. Segal, *et al.*, "Landscape and variation of RNA secondary structure across the human transcriptome," *Nature*, vol. 505, no. 7485, pp. 706–709, 2014.

[21] Y. Mao, H. Liu, Y. Liu, and S. Tao, "Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in Saccharomyces cerevisiae," *Nucleic acids research*, vol. 42, no. 8, pp. 4813–4822, 2014.

[22] S. Rouskin, M. Zubradt, S. Washietl, M. Kellis, and J. S. Weissman, "Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo," *Nature*, vol. 505, no. 7485, pp. 701–705, 2014.

[23] D. Incarnato, F. Neri, F. Anselmi, and S. Oliviero, "Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome," *Genome biology*, vol. 15, no. 10, p. 1, 2014.

[24] S. J. Gosai, S. W. Foley, D. Wang, I. M. Silverman, N. Selamoglu, A. D. Nelson, M. A. Beilstein, F. Daldal, R. B. Deal, and B. D. Gregory, "Global analysis of the RNA-protein interaction and RNA secondary structure landscapes of the Arabidopsis nucleus," *Molecular cell*, vol. 57, no. 2, pp. 376–388, 2015.

[25] J. S. Mattick, "Introns: evolution and function," *Current opinion in genetics & development*, vol. 4, no. 6, pp. 823–831, 1994.

[26] C. E. Lane, K. van den Heuvel, C. Kozera, B. A. Curtis, B. J. Parsons, S. Bowman, and J. M. Archibald, "Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19908–19913, 2007.

[27] N. A. Faustino and T. A. Cooper, “Pre-mRNA splicing and human disease,” *Genes & development*, vol. 17, no. 4, pp. 419–437, 2003.

[28] X. Hong, D. G. Scofield, and M. Lynch, “Intron size, abundance, and distribution within untranslated regions of genes,” *Molecular biology and evolution*, vol. 23, no. 12, pp. 2392–2404, 2006.

[29] M. Burset, I. Seledtsov, and V. Solovyev, “Analysis of canonical and non-canonical splice sites in mammalian genomes,” *Nucleic acids research*, vol. 28, no. 21, pp. 4364–4375, 2000.

[30] S.-A. Marashi, C. Eslahchi, H. Pezeshk, and M. Sadeghi, “Impact of RNA structure on the prediction of donor and acceptor splice sites,” *BMC bioinformatics*, vol. 7, no. 1, p. 297, 2006.

[31] J. Tazi, N. Bakkour, and S. Stamm, “Alternative splicing and disease,” *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1792, no. 1, pp. 14–26, 2009.

[32] M. Polymenidou, C. Lagier-Tourenne, K. R. Hutt, S. C. Huelga, J. Moran, T. Y. Liang, S.-C. Ling, E. Sun, E. Wancewicz, C. Mazur, *et al.*, “Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43,” *Nature neuroscience*, vol. 14, no. 4, pp. 459–468, 2011.

[33] D. J. Patterson, K. Yasuhara, and W. L. Ruzzo, “Pre-mRNA secondary structure prediction aids splice site prediction,” in *Pacific Symposium on Biocomputing*, pp. 223–234, 2001.

[34] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, *et al.*, “The human splicing code reveals new insights into the genetic determinants of disease,” *Science*, vol. 347, no. 6218, p. 1254806, 2015.

[35] O. Gahura, C. Hammann, A. Valentová, F. P
uuta, and P. Folk, “Secondary structure is required for 3′ splice site recognition in yeast,” *Nucleic acids research*, p. gkr662, 2011.

[36] Y. Wan, M. Kertesz, R. C. Spitale, E. Segal, and H. Y. Chang, “Understanding the transcriptome through RNA structure,” *Nature Reviews Genetics*, vol. 12, no. 9, pp. 641–655, 2011.

[37] X. Roca, M. Akerman, H. Gaus, A. Berdeja, C. F. Bennett, and A. R. Krainer, “Widespread recognition of 5′ splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides,” *Genes & development*, vol. 26, no. 10, pp. 1098–1109, 2012.

[38] Y. Yang, L. Zhan, W. Zhang, F. Sun, W. Wang, N. Tian, J. Bi, H. Wang, D. Shi, Y. Jiang, *et al.*, “RNA secondary structure in mutually exclusive splicing,” *Nature structural & molecular biology*, vol. 18, no. 2, pp. 159–168, 2011.

[39] J. Zhang, C. J. Kuo, and L. Chen, “GC content around splice sites affects splicing through pre-mRNA secondary structures,” *BMC genomics*, vol. 12, no. 1, p. 90, 2011.

[40] M. Hiller, Z. Zhang, R. Backofen, and S. Stamm, “Pre-mRNA secondary structures influence exon recognition,” *PLoS Genet*, vol. 3, no. 11, p. e204, 2007.

[41] A. R. Kornblihtt, I. E. Schor, M. Alló, G. Dujardin, E. Petrillo, and M. J. Muñoz, “Alternative splicing: a pivotal step between eukaryotic transcription and translation,” *Nature reviews Molecular cell biology*, vol. 14, no. 3, pp. 153–165, 2013.

[42] N. Lambert, A. Robertson, M. Jangi, S. McGeary, P. A. Sharp, and C. B. Burge, “RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins,” *Molecular cell*, vol. 54, no. 5, pp. 887–900, 2014.

[43] S. H. Bernhart, I. L. Hofacker, and P. F. Stadler, “Local RNA base pairing probabilities in large sequences,” *Bioinformatics*, vol. 22, no. 5, pp. 614–615, 2006.

[44] H. Kiryu, T. Kin, and K. Asai, “Rfold: an exact algorithm for computing local base pairing probabilities,” *Bioinformatics*, vol. 24, no. 3, pp. 367–373, 2008.

[45] X.-F. Wan, G. Lin, and D. Xu, “Rnall: an efficient algorithm for predicting RNA local secondary structural landscape in genomes,” *Journal of bioinformatics and computational biology*, vol. 4, no. 05, pp. 1015–1031, 2006.

[46] S. J. Lange, D. Maticzka, M. Möhl, J. N. Gagnon, C. M. Brown, and R. Backofen, “Global or local? Predicting secondary structure and accessibility in mRNAs,” *Nucleic acids research*, vol. 40, no. 12, pp. 5215–5226, 2012.

[47] M. Hamada, H. Kiryu, K. Sato, T. Mituyama, and K. Asai, “Prediction of RNA secondary structure using generalized centroid estimators,” *Bioinformatics*, vol. 25, no. 4, pp. 465–473, 2009.

[48] H. Tafer, S. L. Ameres, G. Obernosterer, C. A. Gebeshuber, R. Schroeder, J. Martinez, and I. L. Hofacker, “The impact of target site accessibility on the design of effective siRNAs,” *Nature biotechnology*, vol. 26, no. 5, pp. 578–583, 2008.

[49] H. Kiryu, G. Terai, O. Imamura, H. Yoneyama, K. Suzuki, and K. Asai, “A detailed investigation of accessibilities around target sites of siRNAs and miRNAs,” *Bioinformatics*, vol. 27, no. 13, pp. 1788–1797, 2011.

[50] K. J. Doshi, J. J. Cannone, C. W. Cobaugh, and R. R. Gutell, “Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction,” *BMC bioinformatics*, vol. 5, no. 1, p. 1, 2004.

[51] A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neuböck, and I. L. Hofacker, “The vienna RNA websuite,” *Nucleic acids research*, vol. 36, no. suppl 2, pp. W70–W74, 2008.

[52] E. Trotta, “On the normalization of the minimum free energy of RNAs by sequence length,” *PloS one*, vol. 9, no. 11, p. e113380, 2014.

[53] K. R. Rosenbloom, J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, *et al.*, “The UCSC genome browser database: 2015 update,” *Nucleic acids research*, vol. 43, no. D1, pp. D670–D681, 2015.

[54] K. D. Pruitt, G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. M. Farrell, J. Hart, M. J. Landrum, K. M. McGarvey, *et al.*, "RefSeq: an update on mammalian reference sequences," *Nucleic acids research*, vol. 42, no. D1, pp. D756–D763, 2014.

[55] T. F. Consortium *et al.*, "A promoter-level mammalian expression atlas," *Nature*, vol. 507, no. 7493, pp. 462–470, 2014.

[56] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, *et al.*, "Gateways to the FANTOM5 promoter level mammalian expression atlas," *Genome biology*, vol. 16, no. 1, pp. 1–14, 2015.

[57] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[58] D. H. Turner and D. H. Mathews, "NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure," *Nucleic acids research*, p. gkp892, 2009.

[59] M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, and K. P. Murphy, "Efficient parameter estimation for RNA secondary structure prediction," *Bioinformatics*, vol. 23, no. 13, pp. i19–i28, 2007.

[60] M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, and K. P. Murphy, "Computational approaches for RNA energy parameter estimation," *RNA*, vol. 16, no. 12, pp. 2304–2318, 2010.

[61] R. Mori, M. Hamada, and K. Asai, "Efficient calculation of exact probability distributions of integer features on RNA secondary structures," *BMC genomics*, vol. 15, no. Suppl 10, p. S6, 2014.

[62] ISO/IEC, "ISO International Standard ISO/IEC 14882:2011(E) âĂŞ Programming Language C++. [Working draft].," 2011.

[63] "cppreference.com,"

[64] E. P. Nawrocki, S. W. Burge, A. Bateman, J. Daub, R. Y. Eberhardt, S. R. Eddy, E. W. Floden, P. P. Gardner, T. A. Jones, J. Tate, *et al.*, "Rfam 12.0: updates to the RNA families database," *Nucleic acids research*, p. gku1063, 2014.

[65] N. N. Singh, R. N. Singh, and E. J. Androphy, "Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes," *Nucleic acids research*, vol. 35, no. 2, pp. 371–389, 2007.

[66] A. C. Black, C. T. Ruland, M. T. Yip, J. Luo, B. Tran, A. Kalsi, E. Quan, M. Aboud, I. Chen, and J. Rosenblatt, "Human T-cell leukemia virus type II Rex binding and activity require an intact splice donor site and a specific RNA secondary structure.," *Journal of virology*, vol. 65, no. 12, pp. 6645–6653, 1991.

[67] S. Baskerville, M. Zapp, and A. D. Ellington, “Anti-Rex aptamers as mimics of the Rex-binding element,” *Journal of virology*, vol. 73, no. 6, pp. 4962–4971, 1999.

[68] E. Buratti and F. E. Baralle, “Influence of RNA secondary structure on the pre-mRNA splicing process,” *Molecular and cellular biology*, vol. 24, no. 24, pp. 10505–10514, 2004.

[69] K. Sawicka, M. Bushell, K. A. Spriggs, and A. E. Willis, “Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein,” *Biochemical Society Transactions*, vol. 36, no. 4, pp. 641–647, 2008.

[70] M. Plass, C. Codony-Servat, P. G. Ferreira, J. Vilardell, and E. Eyras, “RNA secondary structure mediates alternative $3'$ ss selection in Saccharomyces cerevisiae,” *RNA*, vol. 18, no. 6, pp. 1103–1115, 2012.

[71] A. Siepel and D. Haussler, “Phylogenetic hidden Markov models,” in *Statistical methods in molecular evolution*, pp. 325–351, Springer, 2005.

[72] J. Chamary and L. D. Hurst, “Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals,” *Genome biology*, vol. 6, no. 9, p. R75, 2005.

[73] M. B. Warf and J. A. Berglund, “Role of RNA structure in regulating pre-mRNA splicing,” *Trends in biochemical sciences*, vol. 35, no. 3, pp. 169–178, 2010.

[74] S. Bauer, S. Grossmann, M. Vingron, and P. N. Robinson, “Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration,” *Bioinformatics*, vol. 24, no. 14, pp. 1650–1651, 2008.

[75] P. J. Kersey, J. E. Allen, I. Armean, S. Boddu, B. J. Bolt, D. Carvalho-Silva, M. Christensen, P. Davis, L. J. Falin, C. Grabmueller, *et al.*, “Ensembl Genomes 2016: more genomes, more complexity,” *Nucleic acids research*, vol. 44, no. D1, pp. D574–D580, 2016.

[76] B. Hu, J. Jin, A.-Y. Guo, H. Zhang, J. Luo, and G. Gao, “GSDS 2.0: an upgraded gene feature visualization server,” *Bioinformatics*, vol. 31, no. 8, pp. 1296–1297, 2015.

[77] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz, “Repbase Update, a database of eukaryotic repetitive elements,” *Cytogenetic and genome research*, vol. 110, no. 1-4, pp. 462–467, 2005.

[78] C. Zhang, W.-H. Li, A. R. Krainer, and M. Q. Zhang, “RNA landscape of evolution for optimal exon and intron discrimination,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 15, pp. 5797–5802, 2008.

[79] D. H. Goldman, C. M. Kaiser, A. Milin, M. Righini, I. Tinoco, and C. Bustamante, “Mechanical force releases nascent chain–mediated ribosome arrest in vitro and in vivo,” *Science*, vol. 348, no. 6233, pp. 457–460, 2015.

[80] Y. Sugimoto, A. Vigilante, E. Darbo, A. Zirra, C. Militti, A. D'Ambrogio, N. M. Luscombe, and J. Ule, "hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1," *Nature*, vol. 519, no. 7544, pp. 491–494, 2015.

[81] Z. Lu, Q. C. Zhang, B. Lee, R. A. Flynn, M. A. Smith, J. T. Robinson, C. Davidovich, A. R. Gooding, K. J. Goodrich, J. S. Mattick, *et al.*, "RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure," *Cell*, vol. 165, no. 5, pp. 1267–1279, 2016.

[82] E. Sharma, T. Sterne-Weiler, D. O ' Hanlon, and B. J. Blencowe, "Global Mapping of Human RNA-RNA Interactions," *Molecular cell*, vol. 62, no. 4, pp. 618–626, 2016.

[83] E. J. Merino, K. A. Wilkinson, J. L. Coughlan, and K. M. Weeks, "RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE)," *Journal of the American Chemical Society*, vol. 127, no. 12, pp. 4223–4231, 2005.

[84] K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks, "Accurate SHAPE-directed RNA structure determination," *Proceedings of the National Academy of Sciences*, vol. 106, no. 1, pp. 97–102, 2009.

[85] J. M. Watts, K. K. Dang, R. J. Gorelick, C. W. Leonard, J. W. Bess Jr, R. Swanstrom, C. L. Burch, and K. M. Weeks, "Architecture and secondary structure of an entire HIV-1 RNA genome," *Nature*, vol. 460, no. 7256, pp. 711–716, 2009.

[86] J. G. Underwood, A. V. Uzilov, S. Katzman, C. S. Onodera, J. E. Mainzer, D. H. Mathews, T. M. Lowe, S. R. Salama, and D. Haussler, "FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing," *Nature methods*, vol. 7, no. 12, pp. 995–1001, 2010.

[87] J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin, "Multiplexed RNA structure characterization with selective 2 ′ -hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)," *Proceedings of the National Academy of Sciences*, vol. 108, no. 27, pp. 11063–11068, 2011.

[88] R. C. Spitale, R. A. Flynn, Q. C. Zhang, P. Crisalli, B. Lee, J.-W. Jung, H. Y. Kuchelmeister, P. J. Batista, E. A. Torre, E. T. Kool, *et al.*, "Structural imprints in vivo decode RNA regulatory mechanisms," *Nature*, vol. 519, no. 7544, pp. 486–490, 2015.

[89] M. Frye, S. R. Jaffrey, T. Pan, G. Rechavi, and T. Suzuki, "RNA modifications: what have we learned and where are we headed?," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 365–372, 2016.

[90] N. D. Berkowitz, I. M. Silverman, D. M. Childress, H. Kazan, L.-S. Wang, and B. D. Gregory, "A comprehensive database of high-throughput sequencing-based RNA secondary structure probing data (Structure Surfer)," *BMC bioinformatics*, vol. 17, no. 1, p. 1, 2016.

[91] Y. Wan, K. Qu, Z. Ouyang, M. Kertesz, J. Li, R. Tibshirani, D. L. Makino, R. C. Nutter, E. Segal, and H. Y. Chang, “Genome-wide measurement of RNA folding energies,” *Molecular cell*, vol. 48, no. 2, pp. 169–181, 2012.

[92] F. Li, Q. Zheng, L. E. Vandivier, M. R. Willmann, Y. Chen, and B. D. Gregory, “Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome,” *The Plant Cell*, vol. 24, no. 11, pp. 4346–4359, 2012.

[93] L. J. Kielpinski and J. Vinther, “Massive parallel-sequencing-based hydroxyl radical probing of RNA accessibility,” *Nucleic acids research*, p. gku167, 2014.

[94] Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua, and S. M. Assmann, “In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features,” *Nature*, vol. 505, no. 7485, pp. 696–700, 2014.

[95] M. G. Seetin, W. Kladwang, J. P. Bida, and R. Das, “Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol,” *RNA Folding: Methods and Protocols*, pp. 95–117, 2014.

[96] R. A. Flynn, Q. C. Zhang, R. C. Spitale, B. Lee, M. R. Mumbach, and H. Y. Chang, “Transcriptome-wide interrogation of RNA secondary structure in living cells with ic-SHAPE,” *Nature protocols*, vol. 11, no. 2, pp. 273–290, 2016.

[97] A. Krokhotin, A. M. Mustoe, K. M. Weeks, and N. V. Dokholyan, “DIRECT IDENTIFICATION OF BASE-PAIRED RNA NUCLEOTIDES BY CORRELATED CHEMICAL PROBING,” *RNA*, pp. rna–058586, 2016.

[98] N. A. Siegfried, S. Busan, G. M. Rice, J. A. Nelson, and K. M. Weeks, “RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP),” *Nature methods*, vol. 11, no. 9, pp. 959–965, 2014.

[99] P. J. Homan, O. V. Favorov, C. A. Lavender, O. Kursun, X. Ge, S. Busan, N. V. Dokholyan, and K. M. Weeks, “Single-molecule correlated chemical probing of RNA,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 38, pp. 13858–13863, 2014.

[100] M. Zubradt, P. Gupta, S. Persad, A. M. Lambowitz, J. S. Weissman, and S. Rouskin, “DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo,” *Nature Methods*, 2016.

[101] R. D. Hector, E. Burlacu, S. Aitken, T. Le Bihan, M. Tuijtel, A. Zaplatina, A. G. Cook, and S. Granneman, “Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assembly dynamics at nucleotide resolution,” *Nucleic acids research*, p. gku815, 2014.

[102] A. Selega, C. Sirocchi, I. Iosub, S. Granneman, and G. Sanguinetti, “Robust statistical modeling improves sensitivity of high-throughput RNA structure probing experiments,” *Nature Methods*, 2016.

[103] R. Lorenz, M. T. Wolfinger, A. Tanzer, and I. L. Hofacker, “Predicting RNA secondary structures from sequence and probing data,” *Methods*, 2016.

[104] C. K. Kwok, Y. Ding, Y. Tang, S. M. Assmann, and P. C. Bevilacqua, “Determination of in vivo RNA structure in low-abundance transcripts,” *Nature communications*, vol. 4, 2013.

[105] S. Aviran, C. Trapnell, J. B. Lucks, S. A. Mortimer, S. Luo, G. P. Schroth, J. A. Doudna, A. P. Arkin, and L. Pachter, “Modeling and automation of sequencing-based characterization of RNA structure,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 27, pp. 11069–11074, 2011.

[106] J. Talkish, G. May, Y. Lin, J. L. Woolford, and C. J. McManus, “Mod-seq: high-throughput sequencing for chemical probing of RNA structure,” *RNA*, vol. 20, no. 5, pp. 713–720, 2014.

[107] B. Li, A. Tambe, S. Aviran, and L. Pachter, “Prober: A general toolkit for analyzing sequencing-based ‘toeprinting’ assays,” *bioRxiv*, p. 063107, 2016.

[108] Q. Li, J. B. Brown, H. Huang, and P. J. Bickel, “Measuring reproducibility of high-throughput experiments,” *The annals of applied statistics*, pp. 1752–1779, 2011.

[109] R. Kawaguchi and H. Kiryu, “Parallel computation of genome-scale RNA secondary structure to detect structural constraints on human genome,” *BMC bioinformatics*, vol. 17, no. 1, p. 1, 2016.

[110] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[111] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome biology*, vol. 15, no. 12, p. 1, 2014.

[112] T. N. Vu, Q. F. Wills, K. R. Kalari, N. Niu, L. Wang, M. Rantalainen, and Y. Pawitan, “Beta-Poisson model for single-cell RNA-seq data analyses,” *Bioinformatics*, p. btw202, 2016.

[113] H. Ji, H. Jiang, W. Ma, and W. H. Wong, “Using CisGenome to analyze ChIP-chip and ChIP-seq data,” *Current Protocols in Bioinformatics*, pp. 2–13, 2011.

[114] R. M. Neal and G. E. Hinton, “A view of the EM algorithm that justifies incremental, sparse, and other variants,” in *Learning in graphical models*, pp. 355–368, Springer, 1998.

[115] J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D’Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Müller, *et al.*, “The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs,” *BMC bioinformatics*, vol. 3, no. 1, p. 1, 2002.

[116] W. A. Ziehler and D. R. Engelke, “Probing RNA structure with chemical reagents and enzymes,” *Current protocols in nucleic acid chemistry*, pp. 6–1, 2001.

[117] S. Chatrchyan, V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, *et al.*, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Physics Letters B*, vol. 716, no. 1, pp. 30–61, 2012.

[118] M. Ono, K. Yamada, F. Avolio, M. S. Scott, S. van Koningsbruggen, G. J. Barton, and A. I. Lamond, "Analysis of human small nucleolar RNAs (snoRNA) and the development of snoRNA modulator of gene expression vectors," *Molecular biology of the cell*, vol. 21, no. 9, pp. 1569–1584, 2010.

[119] G. P. Harrison, M. S. Mayo, E. Hunter, and A. M. Lever, "Pausing of reverse transcriptase on retroviral RNA templates is influenced by secondary structures both 5′ and 3′ of the catalytic site," *Nucleic acids research*, vol. 26, no. 14, pp. 3433–3442, 1998.

[120] J. A. Garcia-Martin, I. Dotu, J. Fernandez-Chamorro, G. Lozano, J. Ramajo, E. Martinez-Salas, and P. Clote, "RNAiFold2T: Constraint Programming design of thermo-IRES switches," *arXiv preprint arXiv:1605.04468*, 2016.

[121] G. Csárdi, A. Franks, D. S. Choi, E. M. Airoldi, and D. A. Drummond, "Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast," *PLoS Genet*, vol. 11, no. 5, p. e1005206, 2015.

[122] H. Ohmiya, M. Vitezic, M. C. Frith, M. Itoh, P. Carninci, A. R. Forrest, Y. Hayashizaki, and T. Lassmann, "RECLU: a pipeline to discover reproducible transcriptional start sites and their alternative regulation using capped analysis of gene expression (CAGE)," *BMC genomics*, vol. 15, no. 1, p. 1, 2014.

[123] Y. Tang, E. Bouvier, C. K. Kwok, Y. Ding, A. Nekrutenko, P. C. Bevilacqua, and S. M. Assmann, "StructureFold: genome-wide RNA secondary structure mapping and reconstruction in vivo," *Bioinformatics*, p. btv213, 2015.