論 文 の 内 容 の 要 旨

論文題目

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）

氏　　名　　河口　理紗

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.