Constructing Accurate Contact Maps for Hydroxyl-Radical-Cleavage-Based High-Throughput RNA Structure Inference

Jinkyu Kim, Hanjoo Kim, Hyeyoung Min, and Sungroh Yoon*, *Member, IEEE*

***Abstract*—For rapid ribonucleic acid (RNA) tertiary structure prediction, innovative methods have been proposed that exploit hydroxyl radical cleavage agents in a high-throughput manner. In such techniques, it is critical to determine accurately which residue a specific cleavage agent interacts with, since this information directly reveals the residue–residue interaction points needed for structure inference. Due to lack of effective automated methods, the process of locating contact points has been mostly done manually, becoming a bottleneck of the whole procedure. To address this problem, we propose a novel computational method to determine residue–residue interaction points from 2-D electrophoresis profiles. This method combines the deconvolution method for signal detection and statistical learning techniques for filtering noise, thus boosting specificity and sensitivity in harmony. According to our experiments with over 2000 actual gel profiles, the proposed technique exhibited 56.44%–90.50% higher performance than traditional methods in terms of the accuracy of reproducing manual contact maps measured by the *F*-measure, a widely used evaluation metric. We expect that adopting the proposed technique will significantly accelerate RNA tertiary structure inference, allowing researchers to explore more structures in given time.**

***Index Terms*—Biological signal processing, deconvolution, pattern recognition, ribonucleic acid (RNA), structural bioinformatics.**

I. Introduction

FOR ribonucleic acid (RNA) tertiary structure prediction, several methods have been proposed that exploit hydroxyl radical cleavage agents [1], [2]. By monitoring which residue is cleaved by a specific cleavage agent, we can understand details of residue–residue interactions in a given RNA. A key innovation in this area may be the multiplexed hydroxyl radical (·OH) cleavage analysis (MOHCA) [3], [4] that can perform RNA structure inference in a high-throughput manner. Based on rapid inference of tertiary helical arrangements of RNA molecules, MOHCA gives a possibility to handle large RNA molecules with various RNA folding states in an effective and rapid way.

The purpose of MOHCA is to find out interactions between two or multiple residues in close proximity within a folded RNA molecule. The information on pairs of contacting residues is then translated into structure model constraints for tertiary RNA folding algorithms such as the nucleic acid simulation toolkit (NAST) [7]. The overall procedure of the MOHCA method is illustrated in Fig. 1(a). First, RNAs are transcribed in the presence of modified nucleotide on which the cleavage agent is attached. During transcription, modified nucleotides are randomly incorporated into RNA molecules at the frequency of one modified nucleotide per RNA molecule, and only one type of modified adenosine triphosphate (ATP), guanosine triphosphate (GTP), cytosine triphosphate (CTP), or uridine-5 triphosphate (UTP) is used per transcription reaction. The transcribed RNA is then radiolabeled, folded to the desired structure, and cleaved by the cleavage agent. The cleaved products are first separated by polyacrylamide gel electrophoresis to locate the nucleotide position where cleavage occurs in each RNA molecule. The cleaved RNA molecules are then treated to break their backbone and separated in the orthogonal direction to trace the position of the responsible cleavage agent. Thus, the sequential two-directional electrophoresis determines the cleavage position and the location of the cleavage agent that will form residue–residue pairs whose sugar moieties are present within a certain distance in the structured RNA.

In hydroxyl-radical-based techniques such as MOHCA, the information on residue interactions is contained in the intensity of electrophoresis images. More specifically, as shown in Fig. 1(b), an intensity peak in a gel image corresponds to the location where a cleavage agent reacts physicochemically with its cleavage point(s). A critical step in MOHCA is thus to detect peaks accurately from gel images, and then, construct a contact map that shows all interacting residue pairs. However, due to the high level of noise that coexists with peaks, the major portion of this map construction process in the original MOHCA studies had to be done manually, becoming a bottleneck of the whole procedure. To accelerate the peak detection process in MOHCA, traditional signal processing techniques (such as low-pass filtering followed by zero-crossing detection [6]) and methods developed in the mass spectrometry (MS) community (such as continuous wavelet transform (CWT)-based

Manuscript received September 27, 2010; revised December 15, 2010, January 19, 2011; accepted January 21, 2011. Date of publication January 31, 2011; date of current version April 20, 2011. This work was supported in part by the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology under Grant 2010-0000407 and Grant 2010-0000631. *Asterisk indicates corresponding author.*

J. Kim was with the School of Electrical Engineering, Korea University, Seoul 136-713, Korea. He is now with the R&D Center, LG Display Co., Ltd., Paju-si, Gyeonggi-do 413-779, Korea (e-mail: supmace@korea.ac.kr).

H. Kim is with the School of Electrical Engineering, Korea University, Seoul 136-713, Korea (e-mail: kprotoss@korea.ac.kr).

H. Min is with the College of Pharmacy, Chung-Ang University, Seoul 156-756, Korea (e-mail: hymin@cau.ac.kr).

*S. Yoon is with the School of Electrical Engineering, Korea University, Seoul 136-713, Korea (e-mail: sryoon@korea.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2011.2109716

0018-9294/$26.00 © 2011 IEEE

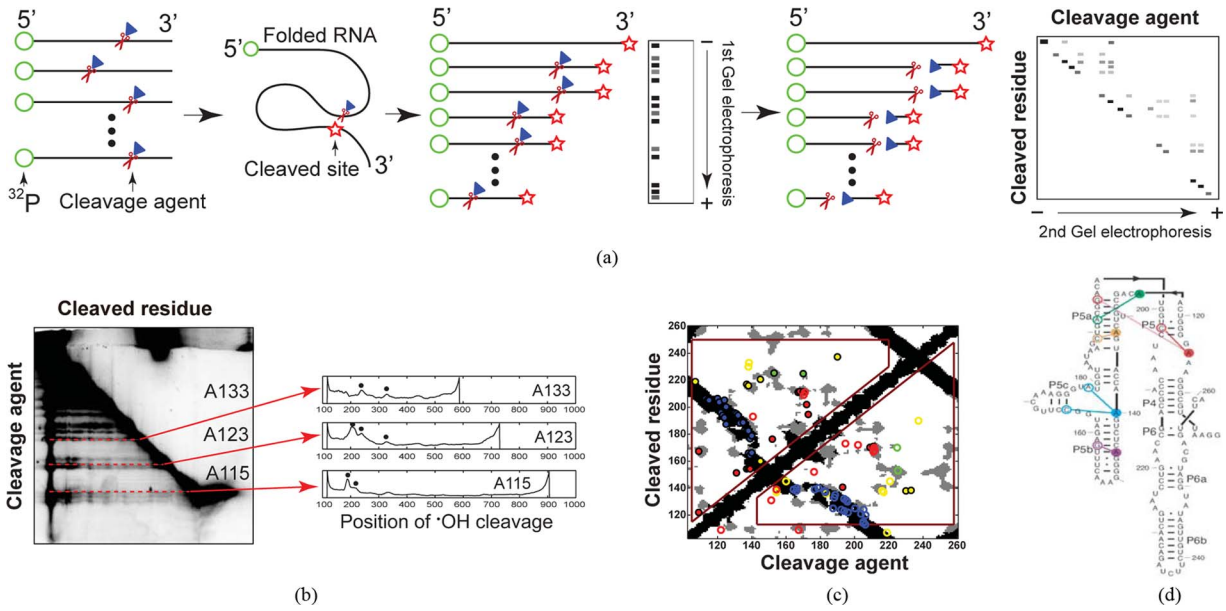


Fig. 1. (a) Overview of MOHCA [3]. (b) Sample 2-D gel electrophoresis image [3] in which the *X*- and *Y*-axes represent the locations of cleaved residues and cleavage agents, respectively. (c) Contact map that shows detailed information on the interactions between cleavage agents and binding sites of the tertiary structure being predicted. More details can be found in Sections II-D and III-B. (d) Native secondary structure with the constraints predicted from the gel in (b) overlaid [3]. This is for the P4–P6 domain of the *Tetrahymena* ribozyme. Cleavage agents (represented by filled circles) are connected to cleaved residues (open circles). Copyright ©2008 R. Das *et al.* All rights reserved.

pattern matching [5]) were initially tried, but these techniques turned out to be of limited value, mainly due to their high false-positive (FP) rates [4]. By combining nonlinear filtering with supervised learning, a more advanced technique was then proposed [4], but this method relies only on partial information available from each gel profile and still cannot construct a contact map accurately.

In this paper, we propose a new computational method to resolve the limitation of the existing techniques. This method combines the peak deconvolution technique developed in gas chromatography [8], [9] for signal detection and statistical learning methods [10] for noise filtering, thus boosting sensitivity and specificity in harmony. Furthermore, by examining multiple batches[1] from various folding states simultaneously, this method can find a contact map that closely resembles the manually constructed one. As shown in Fig. 2, our approach consists of five major steps. First, all potential peaks are detected from individual profiles by deconvolution (step 1), and then, some false findings are filtered out by a binary-classification-based technique that examines multiple profiles within a single batch (step 2). In step 3, the variabilities over replicate batches are handled by peak clustering, and in step 4, peaks from various folding states are considered for interstate analysis. In step 5, a contact map is finally constructed based on the residue interaction information obtained from the previous steps, and the constraints represented in the map can eventually be given to an RNA 3-D structure modeling tool. According to our experimental result, the interbatch analysis (step 3) and the interstate analysis (step 4) primarily contribute to the substantial boosts in accuracy of the proposed method over previous approaches that do not run these two steps.

The rest of this paper is organized as follows. The details of the proposed method can be found in Section II. Section III presents the experimental results we obtained along with discussions on findings. Section IV concludes the paper.

II. Methods

A. Peak Detection by Deconvolution (Step 1)

The first step of the proposed method is to detect peaks from individual profiles by deconvolution. As preprocessing, the baseline of each profile was removed in order to make peaks more salient. In a profile, two windows placed on the left and right sides of the center of each point were set, and the two minimum values within the two windows were calculated. The average of the two minimum values was used as the baseline. For instance, Fig. 3 shows an example profile (in which two true peaks are marked) and its baseline-removed version. Then, we applied to MOHCA profiles the peak deconvolution [8], [9] method that has been widely used in gas chromatography. This method assumes that a single profile consists of multiple separable "bumps" having different center locations and widths. Under this assumption, the peak deconvolution method optimizes its parameters until the difference between the original profile and the superposition of multiple shifted bumps is minimized. The parameter definition depends on the specific peak model that shapes individual bumps. We made the following

[1]A *profile* is a curve that bears the intensity information, and a *batch* refers to a collection of such profiles obtained from a single 2-D gel electrophoresis image. For instance, Fig. 1(b) shows an example batch that contains a number of profiles, three of which are magnified and placed in the right.

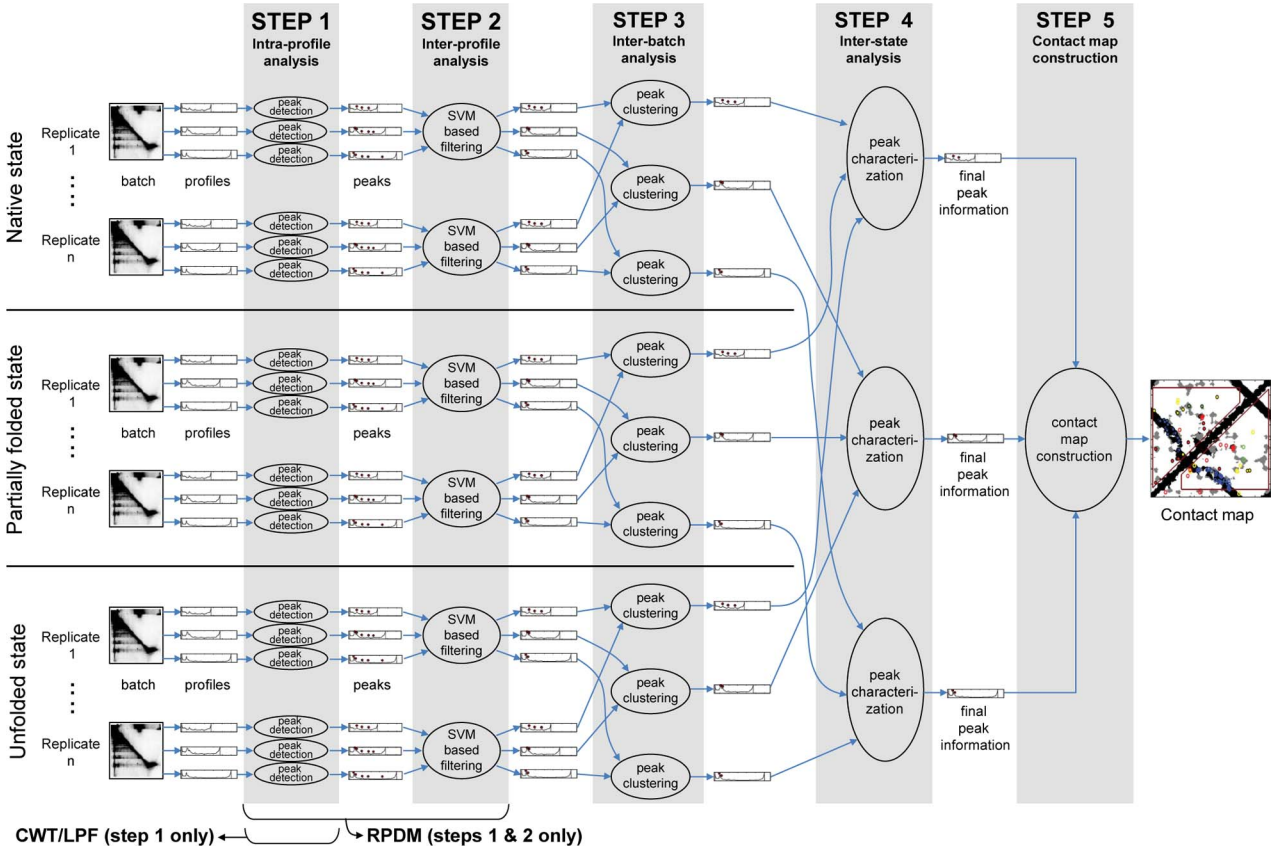


Fig. 2. Overview of the proposed approach. Three folding states exist, namely, unfolded, partially folded, and native states. For generating robust results, several replicate batches are produced in each state. For clarity in visualization, only two replicates per state and three profiles per batch are shown. The CWT-based method [5] and the LPF-based method [6] are applicable to step 1 only. The RPDM method [4] covers only steps 1 and 2.

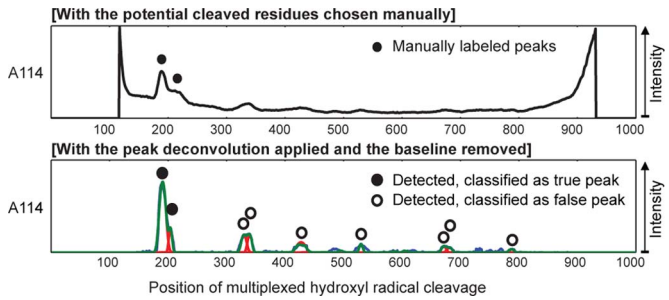


Fig. 3. Applying the peak deconvolution method: original profile (blue), separated profile (red), and accumulated separated profiles (green). In the bottom plot, a filled circle represents a detected peak that is classified as true by the binary classifier; an open circle corresponds to a detected peak that is classified as false.

assumptions on MOHCA profiles when applying the deconvolution method:

A.1) The effects of several cleaved residues are superimposed and appear collectively as a single peak if these residues are overlapping or in close proximity.[2]

[2]In the MOHCA profile images we used, the distance between adjacent bases in an RNA sequence was approximately 3 pixels, and the closest interacting residues we observed were apart by about 10 pixels (i.e., three to four bases).

TABLE I
COMPARISON OF THREE POPULAR PEAK MODELS FOR DECONVOLUTION

	Logistic	Gaussian	Lorentzian
Precision	0.4262 (0.1840)	0.3922 (0.1771)	0.3990 (0.1742)
Recall	0.9632 (0.0732)	0.9594 (0.0806)	0.9572 (0.0812)
F-measure ($\alpha = 0.5$)	0.5682 (0.1745)	0.5338 (0.1746)	0.5411 (0.1718)

A.2) All cleaved residues have the identical distribution defined as a peak model that is centered at its corresponding peak location.

A.3) In a single profile, the number of cleaved residues ranges between 0 and 18.

Generally, the peak deconvolution method consists of two steps, namely, the peak model selection step and the optimization step. To select a peak model adequate to MOHCA profiles, three popular models (Gaussian, Lorentzian, and Logistic) were tried, and the logistic model showed the best performance (see Table I in Section III-C). A bump in the logistic model is defined as

$$f(x) = \frac{2 \times \exp[-(x-c)/(0.477 \times w)]^2}{1 + \exp[-(x-c)/(0.477 \times w)]^2} \quad (1)$$

where $f(x)$ is the intensity (i.e., bump height) at position x and c and w represent the center location and the width of the bump, respectively. As an unconstrained nonlinear optimizer, the Nelder–Mead simplex search algorithm [11] was utilized to optimize parameters such as center locations, peak widths, and the number of peaks. For example, the nine circles in the bottom plot of Fig. 3 represent the nine peaks detected by applying the peak deconvolution method to the baseline-removed profile.

Under the assumption A.1, we found that the peak deconvolution method can separate overlapped peaks that appear as a single peak, better than alternatives such as the robust peak detection method (RPDM) [4], the CWT method [5], and low-pass filter (LPF)-based approaches [6]. Such methods rely on finding local maxima, and in order to avoid finding spurious peaks and noise, typically, assume that the distance between two adjacent peaks is not too close. RPDM further assumes that the magnitude of a valid peak at x_p should be considerably larger than the average of nearby points (i.e., $y(x_p) \gg E[y(x)]$). However, due to noise and other distortions in profiles, which are inherent in gel electrophoresis, these assumptions may not be helpful for true peak detection, but rather tend to make it difficult to distinguish overlapped peaks, thus increasing the number of true peaks undetected.

B. Interprofile Analysis for Removing False Peaks (Step 2)

As previously stated, the deconvolution method was more effective in eliminating false peaks and retaining true ones than the alternative techniques we tried. However, using the deconvolution method alone still turned out to be insufficient for achieving the degree of accuracy needed to fully automate MOHCA. For instance, although the deconvolution method can detect all the true peaks marked in the top plot of Fig. 3, there still remain false peaks in the bottom plot of Fig. 3. Adjusting the parameters of the peak deconvolution could be tried, but this may cause adversary effects on controlling the number of true positives (TPs) and/or false negatives (FNs). Furthermore, there is a need to consider interferences caused by other profiles that require a more sophisticated treatment than simple peak detection.

After the deconvolution step is over, we, therefore, employed a supervised learning technique in order to control peak findings more effectively by reflecting domain knowledge obtained from previous experiments. Since there are a number of labeled reference peaks available from earlier MOHCA studies, we could exploit this statistical learning method that requires enough training examples to be effective.

In particular, we utilized the properties of true MOHCA peaks characterized over multiple profiles within a batch [4]. Based on the procedure proposed in [4], we extracted a 38-D feature vector from each peak candidate. The features used include the intensity information both in time and frequency domains, the number of nearby peaks in other profiles, the type of radical sources (i.e., adenine, uracil, and cytosine), the location of radioactive nucleotides on RNAs (i.e., 5' or 3') and the folding states. According to the procedure outlined in Fig. 4, we then trained a support vector machine (SVM) [10] and used it for filtering out false peaks. See Section III-C for additional details.

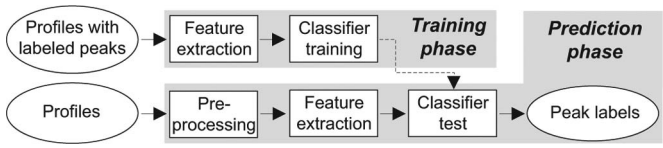


Fig. 4. Interprofile analysis for filtering out false findings (step 2). To incorporate domain knowledge obtained from previous experiments, we use a binary-classification-based approach.

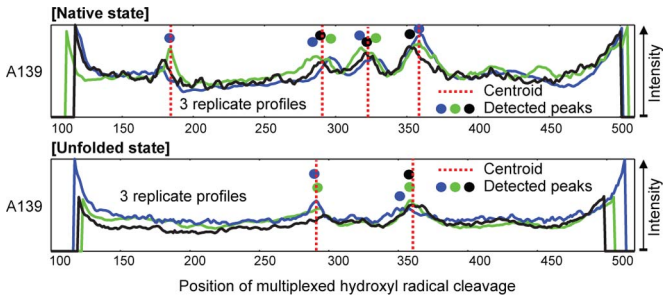


Fig. 5. Peak clustering for handling differences in the location of a peak over replicate profiles. Three replicated profiles are overlapped. The peaks detected in each replicate are represented by filled circles, and the centroid location of merged peaks is marked by a vertical dotted line.

C. Handling Variabilities Over Replicates (Step 3)

After the binary classification step, we performed the k-means clustering [10], [12] of the peaks in each of the three folding states (i.e., native, unfolded, and partially folded). This is to handle the subtle discrepancies of the locations of the same peak over replicate profiles. As depicted in Fig. 5, we took the centroid of each cluster as its representative location (red dotted line). To determine the k parameter of the k-means clustering, we searched for the minimum k that satisfies the following: all elements in a single cluster are within a user-specified range from the centroid and the number of possible clusters is between 1 and the number of peaks.

D. Interstate Analysis by Peak Characterization for Contact Map Construction (Steps 4 and 5)

A *contact map* shows detailed information on the interactions between cleavage agents and binding sites of the tertiary structure being predicted (e.g., see Fig. 7). The information represented by a contact map is later to be used as 3-D structure constraints that are the key ingredients of tertiary folding algorithms such as NAST [7]. Generating an accurate map is thus critical, and to improve the accuracy as much as possible, it is a common practice to generate replicates of a batch and consider multiple batches simultaneously. Although the performance of RPDM [4] was satisfactory as far as peak detection in a single batch is concerned, RPDM (and all the other alternatives) did not consider multiple batches from various folding states for more accurate contact map construction. This accounts for a key difference between the proposed method and existing approaches including RPDM. In terms of the quality of a contact map constructed, the proposed approach thus outperforms RPDM and

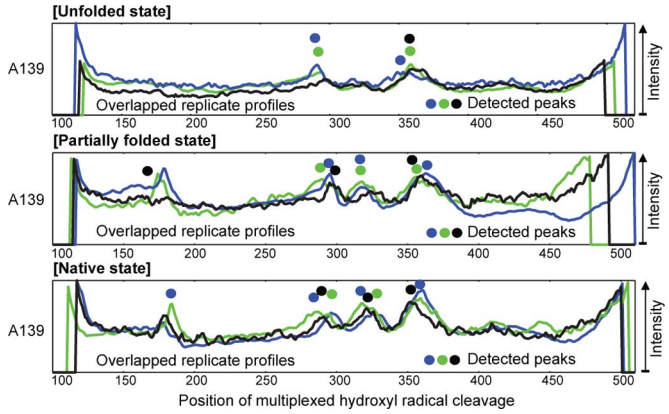


Fig. 6. Examples of cleaved residue profiles obtained from replicates. For each folding state, three replicates and detected peaks therein (filled circles) are overlaid using different colors.

other previous techniques by a large margin, as presented in Section III-B.

The process to build a contact map in this study exploits the following characteristics of true peaks that originate from biological and statistical insights of domain experts:

C.1) True residue pairs appear at least three times out of four replicated profiles. As depicted in Fig. 6, replicate profiles are generated under the identical experimental condition, and statistically significant peaks would, therefore, appear in all the replicates. For each combination of folding states, radical sources, and radiolabel locations, we used four replicates.

C.2) A residue pair observed in the unfolded state has no influence on tertiary structure inference. This kind of residue pair instead gives the binding site information for 2-D RNA structure inference [13].

C.3) The residue pairs observed in an intermediate folding state also appear in "higher degree" folding states (but not vice versa). For example, in Fig. 6, the residue pairs observed in the unfolded state also appear in the partially folded and the native states. Likewise, the residue pairs observed in the partially folded state are also found in the native state.

For accurate tertiary structure prediction, it is important to minimize the number of peaks that do not follow the aforementioned characteristics. After removing those residue pairs that violate the peak characteristics C.1 and C.3, we constructed a contact map based on the interaction information represented by the remaining residue pairs. See Section III-B and Fig. 7 for additional details of drawing a contact map and examples.

Regarding C.3, note that RNA takes hierarchical folding pathway to reach its final folded structure [14], [15]. In an RNA, the primary structure of a linear sequence determines the secondary structure that subsequently defines tertiary structure without much reorganization of the secondary structure. Due to the hierarchical nature of RNA structure, it is simpler to predict RNA folding states rather than those of protein that does not follow the hierarchical and sequential folding rule. In addition, the structure of partially folded states at high monovalent ion concentrations is identified as the folding intermediate [16]. The folding intermediate contains components consisting of secondary structures and short-range tertiary interactions, supporting the formation of native tertiary structure in a sequential order.

E. Data Preparation and Implementation

We tested the proposed technique with 2002 profiles extracted from 79 MOHCA batches. These profiles are from the P4–P6 domain of the *Tetrahymena* ribozyme and cover all three folding states: unfolded, folded, and partially folded states. They also cover all experimentally possible types of radical sources and locations of radioactive nucleotides on RNAs. The code and data used for our experiments are freely available at http://dna.korea.ac.kr/pub/cmap. Additional information (e.g., a list of batches, the types of radical sources, the location of radioactive nucleotides, and other details) is available from the authors upon request.

For performance comparison, we also implemented three alternative techniques. As a simple but popular approach, we implemented an LPF-based method [6] that filters out high-frequency noise and detects zero-crossing points in the derivative as peak locations. The second technique is a CWT-based method [5] that exploits SNR-based peak finding criteria. Lastly, we used the RPDM [4] that combines nonlinear signal processing techniques with a binary classifier.

III. Results and Discussion

A. Performance Evaluation Metrics

To facilitate further discussions, we provide a brief introduction to the performance evaluation metrics we used. In binary hypothesis testing, the outcomes are labeled either *positive* (*P*) or *negative* (*N*). Given that the actual value is *P*, if the outcome by a predictor is *P*, then the outcome is called *TP*; otherwise, it is *FN*. An FP and a true negative (TN) are similarly defined. Recall is defined as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, and precision is defined as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$. We can monitor the tradeoff relationship between precision and recall by using *F*-measure defined as

$$\frac{\mathrm{Precision}\times\mathrm{Recall}}{\alpha\times\mathrm{Recall}+(1-\alpha)\times\mathrm{Precision}} \tag{2}$$

where α is a weighting parameter between 0 and 1. A precision, recall, or *F*-measure value ranges between 0 and 1; an ideal classifier would have the value of 1 for all three. Further details of these metrics can be found in [12] and [17].

B. Accurate Construction of Contact Maps

The primary message we want to deliver is that the proposed method can significantly enhance the quality of constructing contact maps with respect to existing approaches. Fig. 7(a)–(e) shows the contact maps constructed from the P4–P6 domain of the *Tetrahymena* ribozyme by the methods in comparison. As reference map, the manually constructed contact map is also presented. In each map, residue pairs are color-coded according

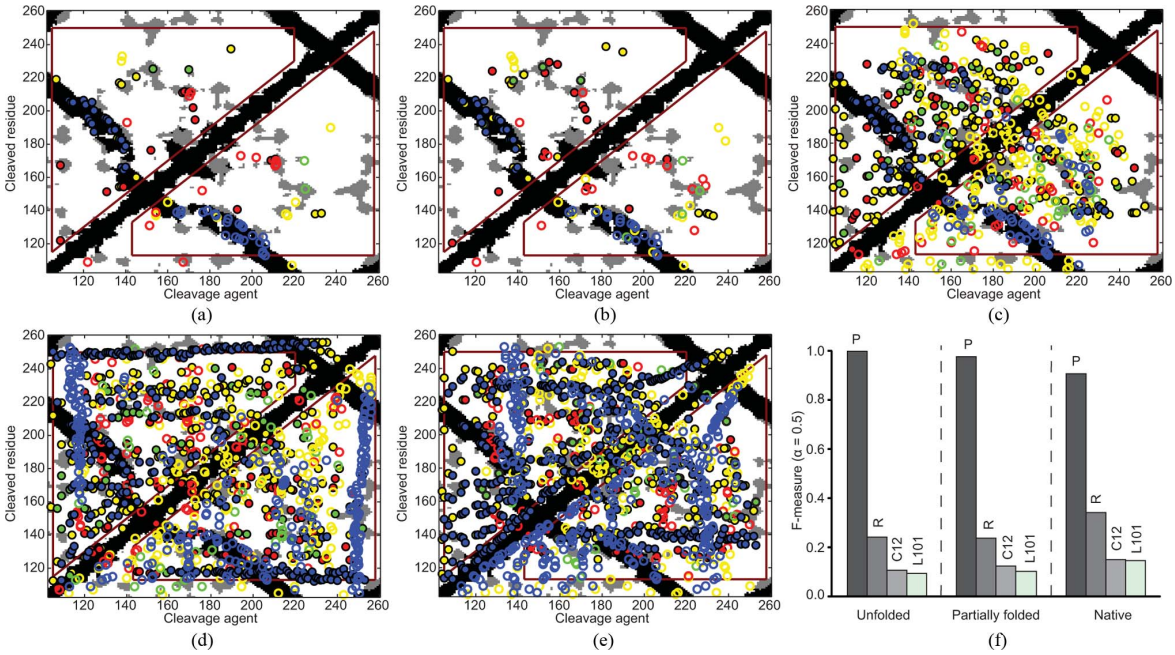


Fig. 7. (a)–(e) Contact maps obtained by different methods. A filled circle represents the predicted residue pair generated by the analysis of profiles, and an open circle indicates a symmetric residue pair of the corresponding filled circle. Each residue pair is color-coded according to the four types (T.1–T.4). The black and gray areas in the background represent the secondary (black) and tertiary (gray) crystal structures that are brought within 25 angstroms. The inner region enclosed by red trapezoidal lines indicates the region where residue pairs are easily distinguished on MOHCA gels. (a) Manual. (b) Proposed. (c) RPDM. (d) CWT (SNR > 12). (e) LPF (101-tap). (f) Comparing the *F*-measure of the competing methods by using the residue pairs in (a) as reference.

to the following classification rule that is based on characteristics C.2 and C.3 described in Section II-D:

- T.1) The residue pairs that appear in every folding state (represented by blue circles).
- T.2) The residue pairs that are found in all folding states except the unfolded state (green circles).
- T.3) The residue pairs that appear only in the native state (red circles).
- T.4) All the other residue pairs (yellow circles).

As seen in this figure, the contact map built by using the proposed method resembles the reference map most closely. For more quantitative comparison, we computed the *F*-measure values of the competing methods by using the residue pairs in the reference map as true values and those in a nonreference map as predicted outcomes, as shown in Fig. 7(f). Evidently, our approach maintains a large performance gap in each folding state with respect to the others, producing 75.59%–90.50%, 74.17%–87.58%, and 56.44%–76.27% higher *F*-measure values (with $\alpha = 0.5$) in the unfolded, the partially folded, and the native states, respectively. As compared with existing methods, the proposed approach performs additional analysis of profiles and batches such as peak clustering and interstate analysis. Our experimental results indicate that these additional procedures primarily contribute to the substantial boosts in accuracy of the proposed method over previous approaches.

C. Comparison of Peak Detection Performance

The secondary message of this paper is that the deconvolution method is more effective than its alternatives in terms of detecting peaks from MOHCA profiles. This is supported by the experimental results, as shown in Fig. 8, which compares the performance of our approach and the other techniques focusing only on the ability to detect true peaks from a batch of MOHCA profiles.

To prepare labeled data for evaluating peak detection performance, we identified 2067 potential peaks from the 79 batches that contain 2002 profiles in total. From these 2067 peaks, we randomly selected some peaks and examined them manually in order to label and use them as the training examples for an SVM-based classifier. This way, 181 positive examples (i.e., true peaks) and 189 negative examples (i.e., false peaks) were prepared and used for training. By using the trained classifier, we predicted the labels of the remaining 1697 ($= 2067 - 181 - 189$) peaks, and then, computed the precision, recall, and *F*-measure values for performance evaluation. In this setup, a TP and an FP mean the number of detected true peaks and the number of detected false peaks, respectively. An FN represents the number of undetected true peaks.

Fig. 8(a)–(c) presents the box-and-whisker plots that compare the peak detection performance of the competing techniques. The proposed method produced 5.0%–76.7% higher median precision values than the alternatives, indicating that our approach detects fewer false peaks. The median recall values of our method were not consistently the best, but this is rather the evidence that the other techniques tend to detect excessive peaks, lowering FN and thus boosting recall. This reasoning is well-supported by examining the *F*-measure that considers both precision and recall in order to reveal the tradeoff between the two. In terms of *F*-measure (with $\alpha = 0.5$), our approach

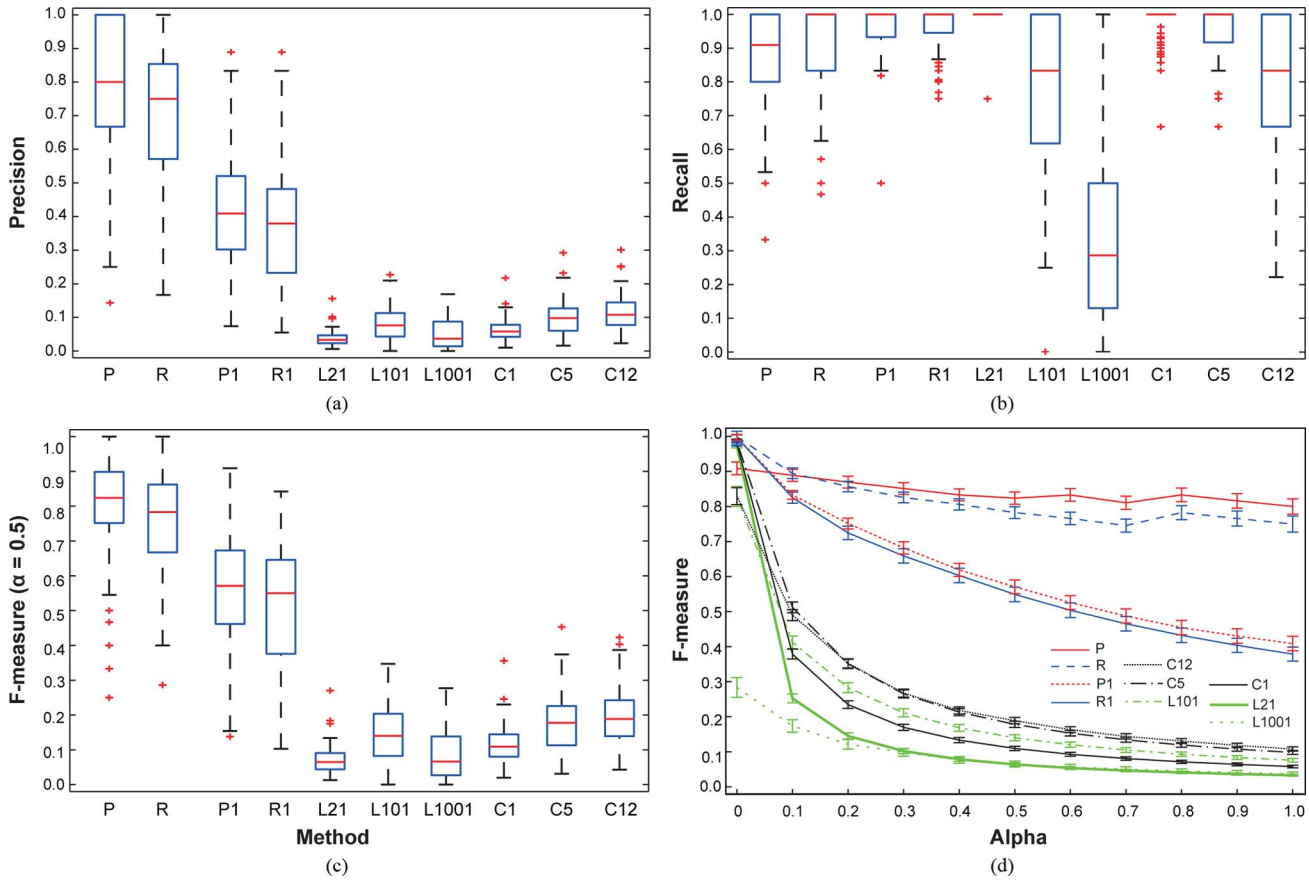


Fig. 8. Comparing the ability to detect peaks. P: proposed method (steps 1 and 2), P1: proposed method (step 1 only), R: RPDM [4] (steps 1 and 2), R1: RPDM (step 1 only), L21: 21-tap LPF [6], L101: 101-tap LPF, L1001: 1001-tap LPF, C1: CWT [5] (SNR $>$ 1), C5: CWT (SNR $>$ 5), C12: CWT (SNR $>$ 12). The normalized passband frequency of all LPFs is 0.01. (a) Precision. (b) Recall. (c) F-measure ($\alpha = 0.5$). (d) F-measure comparison with α varying from 0 to 1.

exhibited the best performance, showing 4.1%–75.9% higher median values. We may root-cause the performance gap as follows. The LPF-based method is inadequate to MOHCA profiles due to the excessive number of zero-crossing points, most of which are due to mere noise. The CWT-based method, which was originally proposed for MS data analysis, showed rather unsatisfactory performance since MS data possess different characteristics and peak formation from MOHCA profiles. For further performance comparison, we computed F-measure by varying the α parameter from 0 to 1, as shown in Fig. 8(d). As is evident in the figure, the proposed technique outperformed the alternatives over most values of α. The other techniques show better performance only when α is near 0 (i.e., when F-measure is basically recall), due to the reasons explained earlier. Note that the proposed method (labeled as P in the figure) and RPDM (labeled as R) perform steps 1 and 2, whereas the others do step 1 only. For fair comparison, we thus measured the performance of P and R without running step 2, and the corresponding results (labeled as P1 and R1, respectively) are also presented in the plots.

We also utilized the labeled data for selecting a peak model for deconvolution. Table I shows the performance of deconvolution-based peak detectors using different models. Shown in the table are the median values with standard deviation in parenthesis, measured over the same data as shown in Fig. 8. The logistic peak model consistently showed the best performance, and thus, we adopted it for implementing our peak detector.

D. Pilot Study on Tertiary Structure Model Construction

We tested the proposed method in terms of its ability to predict tertiary structure models. By using the residue pairs that were predicted as true interaction points by the different techniques used in comparison, the tertiary structure models of the P4–P6 domain of the *Tetrahymena* ribozyme were produced. As golden reference, the corresponding X-ray crystal structure was used. For fair comparison, we predicted over 1000 models with residue pairs generated for each method in the native state. These models were constrained by the NAST [7], a coarse-grained tertiary structure constraining tool. During the simulation, we swept the value of the spring-preferred distance parameter from 1 to 15 and took the best values that show the lowest energy. See [7] and [18] for further information on the parameters of NAST. For runtime optimization, we generated only the coarse-grained models that consist of dummy residue molecules at the C4 position.

To summarize the simulation result quantitatively, we calculated the value of the rms deviation (RMSD) between the predicted models and the (experimentally proven) crystal structure.

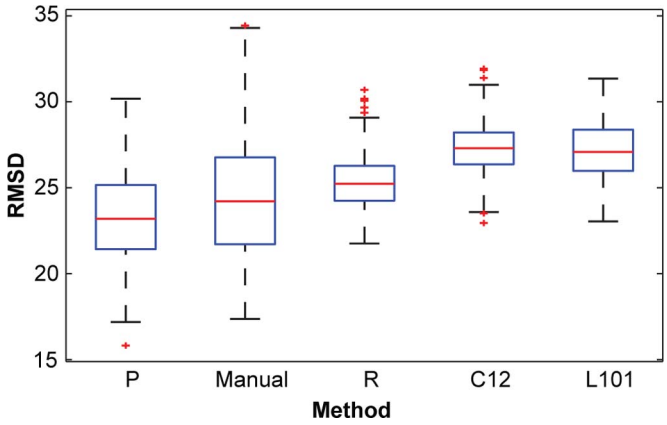


Fig. 9. Distributions of the RMSD values.

Fig. 9 compares the distributions of the RMSD values of the different models. This figure indicates that the proposed technique can produce a tertiary structure that most closely resembles the manually constructed model in this specific example, producing 37.50%, 45.05%, and 45.68% lower minimum RMSD values than RPDM, CWT (SNR $>$ 12), and LPF (101-tap), respectively.

IV. Conclusion

We have proposed a novel computational method for constructing a residue interaction map used in RNA tertiary structure prediction. The proposed approach combines peak modeling for isolating true signals from noisy electrophoresis profiles and supervised classification for further noise reduction. Leveraged by this two-step approach, our algorithm could outperform existing techniques in terms of precision, recall, and F-measure. Most importantly, our approach can produce a contact map that closely resembles the manually built map. We expect that using the proposed method will significantly accelerate RNA tertiary structure inference, allowing researchers to explore more structures in given time.

Acknowledgment

The authors would like to thank the anonymous reviewers of this manuscript for their helpful comments and Prof. R. Das at Stanford University for providing MOHCA data.

References

[1] S. Joseph, M. L. Whirl, D. Kondo, H. F. Noller, and R. B. Altman, "Calculation of the relative geometry of tRNAs in the ribosome from directed hydroxyl-radical probing data," *RNA*, vol. 6, no. 2, pp. 220–232, 2000.

[2] H. Han and P. B. Dervan, "Visualization of rna tertiary structure by RNA-EDTA.Fe(ii) autocleavage: Analysis of tRNA(phe) with uridine-EDTA.Fe(ii) at position 47," *Proc. Nat. Acad. Sci. USA*, vol. 91, no. 11, pp. 4955–4959, 1994.

[3] R. Das, M. Kudaravalli, M. Jonikas, A. Laederach, R. Fong, J. P. Schwans, D. Baker, J. A. Piccirilli, R. B. Altman, and D. Herschlag, "Structural inference of native and partially folded RNA by high-throughput contact mapping," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 11, pp. 4144–4149, 2008.

[4] J. Kim, S. Yu, B. Shim, H. Kim, H. Min, E. Y. Chung, R. Das, and S. Yoon, "A robust peak detection method for RNA structure inference by high-throughput contact mapping," *Bioinformatics*, vol. 25, no. 9, pp. 1137–1144, 2009.

[5] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, 2006.

[6] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[7] M. A. Jonikas, R. J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, and R. B. Altman, "Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters," *RNA*, vol. 15, no. 2, pp. 189–199, 2009.

[8] G. Vivo-Truyols, J. R. Torres-Lapasio, A. M. van Nederkassel, Y. Vander Heyden, and D. L. Massart, "Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part I: Peak detection," *J. Chromatogr. A*, vol. 1096, no. 1/2, pp. 133–145, 2005.

[9] G. Vivo-Truyols, J. R. Torres-Lapasio, A. M. van Nederkassel, Y. Vander Heyden, and D. L. Massart, "Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part II: Peak model and deconvolution algorithms," *J. Chromatogr. A*, vol. 1096, no. 1/2, pp. 146–155, 2005.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[11] J. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, no. 4, pp. 308–313, 1965.

[12] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. (Morgan Kaufmann Series in Data Management Systems). San Francisco, CA: Morgan Kaufman, 2005.

[13] M. Wu and J. Tinoco, I, "RNA folding causes secondary structure rearrangement," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 20, pp. 11555–11560, 1998.

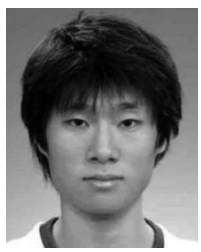
[14] I. Tinoco and C. Bustamante, "How RNA folds," *J. Mol. Biol.*, vol. 293, no. 2, pp. 271–281, 1999.

[15] P. Brion and E. Westhof, "Hierarchy and dynamics of RNA folding," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 26, no. 1, pp. 113–137, 1997.

[16] N. Baird, E. Westhof, H. Qin, T. Pan, and T. Sosnick, "Structure of a folding intermediate reveals the interplay between core and peripheral elements in RNA folding," *J. Mol. Biol.*, vol. 352, no. 3, pp. 712–722, 2005.

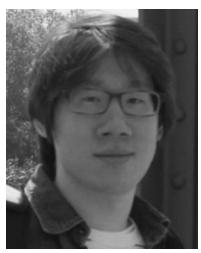
[17] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

[18] R. Das and D. Baker, "Automated de novo prediction of native-like rna tertiary structures," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 37, pp. 14664–14669, 2007.



Jinkyu Kim received the B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, Korea, in 2008 and 2010, respectively.

He is presently a Research Engineer at LG Display Co., Ltd., Paju-si, Gyeonggi-do, Korea. His current research interests include the high-throughput RNA tertiary inference and pattern-recognition-based panel quality enhancements.



Hanjoo Kim received the B.S. degree in electrical engineering in 2009 from Korea University, Seoul, Korea, where he is currently a graduate student.

His current research interests include integrative bioinformatics for microRNA and hardware acceleration of bioinformatics algorithms.

Hyeyoung Min received the B.S. and M.S. degrees in pharmacy from Seoul National University, Seoul, Korea, in 1992 and 1998, respectively, and the Ph.D. degree in cellular and molecular pathology from the University of California, Los Angeles, in 2004, and the M.S. degree in statistics from Stanford University, Stanford, CA, in 2008.

She was a Postdoctoral Research Scientist at Stanford University. She is currently an Assistant Professor in the College of Pharmacy, Chung-Ang University, Seoul, Korea. Her current research interests include bioinformatics and microRNA function in pathogenesis.

Sungroh Yoon (S'99–M'06) received the B.S. degree in electrical engineering from Seoul National University, Seoul, Korea, in 1996, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 2002 and 2006, respectively.

From 2006 to 2007, he was with Intel Corporation, Santa Clara, CA. Previously, he held research positions with Stanford University and Synopsys, Inc., Mountain View, CA. He is currently an Assistant Professor in the School of Electrical Engineering, Korea University, Seoul, Korea. His current research interests include bioinformatics and computational biology.