BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–8

Automated band annotation for capillary electrophoresis based high-throughput RNA structure probing

ABSTRACT

Motivation: Capillary electrophoresis (CE) is a widely used approach for biochemical analysis. Among many steps composing CE analysis, the process of mapping each band in an electrophoretic trace to a position in the nucleic acid sequence normally is most time-consuming and takes human efforts and manual inspections. There has been a few tools trying to provide automated methods for band annotation only to end up with limited accuracy.

Results: We present a dynamic-programming based approach to automated band annotation for high-throughput capillary electrophoresis. Key to our approach is the transformation of the band annotation problem into an optimization problem to maximize a score defined to indicate the probabilistic likelihood of a set of band locations. Our experiments reveal that our method outperforms prior tools by large in terms of accuracy whereas the increase in the amount of computation required is reasonable. We also introduce a scoring metric under which a set of automatically annotated bands is determined reliable or not. It is also demonstrated by experiments that this metric is practically useful in filtering out results to be rejected for their inaccuracy.

Availability: The implementation of the proposed algorithm is included in the HiTRACE software freely available for download at http://hitrace.stanford.edu.

Contact: sryoon@snu.ac.kr, rhiju@stanford.edu

1 INTRODUCTION

RNA plays diverse roles in encoding and regulating genetic information. We can better understand this versatility of RNA by knowing its higher order structure. To this end, chemical probing methodologies provide a powerful means to mapping RNA secondary and tertiary structure at single-nucleotide resolution (Weeks, 2010).

There exist many chemical probing techniques, most of which have common experimental procedures as follows: Treated with an RNA of interest, a chemical reagent modifies the RNA, either cleaving it or forming a covalent adduct with it. Examples of the reagent include hydroxyl radicals, dimethyl sulfate (DMS), CMCT, kethoxal and bisulfite. Subsequent reverse transcription detects the modification sites as stops to primer extension at nucleotide resolution. Resulting cDNA fragments are resolved in sequencing gels followed by individually quantifying band intensities. Typically, the bottleneck is the final steps (gel running and band quantification).

To resolve fragments in high-throughput fashion, capillary electrophoresis (CE) can be used. CE-based chemical probing can easily produce tens of thousands of individual electrophoretic bands from a single experiment, leading to recent breakthroughs in high-throughput mapping of complex RNA structures (Mitra *et al.*, 2008; Vasa *et al.*, 2008; Weeks, 2010; Das *et al.*, 2010; Kladwang and Das, 2010) such as ribosomes (Deigan *et al.*, 2009), and viruses (Wilkinson *et al.*, 2008; Watts *et al.*, 2009).

Analyzing a large number of electrophoretic traces from a high-throughput structure-mapping experiment is time-consuming and poses a significant informatic challenge, requiring a set of robust signal-processing algorithms for accurate quantification of the structural information embedded in the noisy traces. Current software methods for CE analysis include capillary automated footprinting analysis (CAFA; Mitra *et al.*, 2008), ShapeFinder (Vasa *et al.*, 2008), high-throughput robust analysis for capillary electrophoresis (HiTRACE; Yoon *et al.*, 2011), and fast analysis of SHAPE traces (FAST; Pang *et al.*, 2011).

A typical high-throughput CE analysis pipeline consists of the following steps (Yoon *et al.*, 2011): preprocessing such as normalization and baseline adjustment, alignment, peak detection, band annotation, and peak fitting. Among these, band annotation refers to the process of mapping each band in a electrophoretic trace to a position in the nucleic acid sequence. For verification, visual inspection in this phase is normally inevitable to certain extent. However, in practice, this band annotation step often takes significant human efforts in CAFA and ShapeFinder, for they were designed to focus more on alignment and peak fitting. HiTRACE and FAST provides an improved level of band annotation support, but band annotation remains still the most time-consuming step of a HiTRACE- or FAST-based analysis pipeline.

This paper describes a dynamic-programming based approach to automated band annotation for high-throughput capillary electrophoresis. Figure 1 shows the overview of the proposed methodology for automated band annotation.

2 METHODS

2.1 Problem definition

Given an RNA sequence $\mathbf{s}$ of length N, assume that we carry out the chemical structure probing of this sequence using M different treatments, each of which is run by a separate capillary lane. Assume that the fluorescence intensity of each capillary is measured over K time points. We define a *profile* as the sequence of intensity values from a capillary. The entire CE measurement can then be arranged in a $K \times M$ matrix $\mathbf{D}$. Normally, $N \ll K$. Based on the characteristic of the chemical agent used in each treatment and the secondary structure computationally inferred from the input sequence, we can predict the fluorescence intensity at each position of $\mathbf{s}$ for each of M treatments. This prediction can be arranged in a $N \times M$ matrix $\mathbf{P}$ called the *prediction matrix*.

The problem of band annotation is formulated as selecting N out of the K rows of $\mathbf{D}$ using the information in $\mathbf{P}$ in such a way that a certain objective

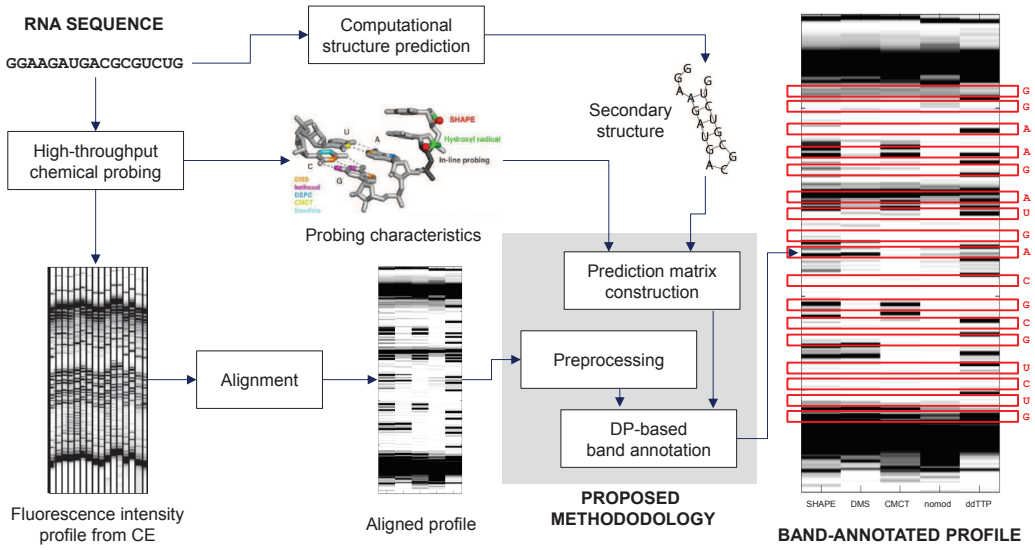


Fig. 1. Overview of the proposed dynamic-programming-based band annotation methodology. Given an RNA sequence, we carry out high-throughput structure-mapping experiments, producing a number of capillary electrophoresis (CE) profiles. Additionally, we computationally predict the secondary structure of the input sequence. From the predicted structure and the characteristics of the chemical probing method used, we drive a prediction matrix that stores expected interaction patterns between the residues. Based on the aligned CE profiles and prediction matrix, we apply a dynamic-programming approach that finds the optimal selection of the band locations under the predetermined scoring scheme.

is optimized over all possible $\binom{K}{N}$ possibilities. The selected N points map to the locations of the nucleotides of the sequence $\mathbf{s}$ in the CE measurement.

The input of the proposed method consists of the following:

- $\mathbf{D} \in \mathbb{R}^{K \times M}$: the fluorescence intensity matrix
- $\mathbf{P} \in \{0,1\}^{N \times M}$: the prediction matrix
- $\mathbf{s} \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}^N$: the nucleotide sequence

and the output is an array $\mathbf{y} \in \mathbb{Z}_+^N$ representing N points selected out of K.

2.2 Prediction matrix construction

Given an RNA sequence, we computationally predict its secondary structure using the Vienna RNA package (Hofacker, 2003). Based on the predicted structure and the properties of the chemical probing used, we construct the prediction matrix $\mathbf{P}$ that stores the expected chemical reactivity for individual residues. The element $p_{ij} \in \mathbf{P}$ indicates such reactivity information of residue i to reagent j.

We assume the use of three chemical probing strategies in this paper: dimethyl sulfate alkylation (Tijerina *et al.*, 2007) [DMS], carbodiimide modification (Walczak *et al.*, 1996) [CMCT], and 2′-OH acylation [the SHAPE strategy (Merino *et al.*, 2005)]. Figure 2(a) defines the expected reactivity of each type of nucleotide to chemical reagents used for chemical probing under the (un)paired condition. The value of one means the reactivity to a reagent (*i.e.*, the existence of a band in the fluorescence profile), whereas zero indicates no reactivity (*i.e.*, no band). For instance, in the unpaired condition, the DMS chemical modifies `A` and `C` but not `U` and `G`, and the entries for `A` and `C` are one, while those for `U` and `G` are zero.

Figure 2(b) shows an example RNA sequence with its secondary structure. Figure 2(c) shows the corresponding prediction matrix $\mathbf{P}$.

2.3 Preprocessing intensity data

Let $\mathbf{d}_j$ be the j-th column vector of $\mathbf{D}, 1 \leq j \leq M$. The following procedure is executed.

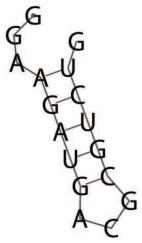
1. Select candidates for the peaks in $\mathbf{d}_j$ that can be mapped into elements of the sequence $\mathbf{s}$. These peaks are selected to satisfy the following conditions. First, a peak $\mathbf{d}_j(k)$ must have a higher intensity (a fundamental property of a peak) than those of its neighbors, $\mathbf{d}_j(k-1)$ and $\mathbf{d}_j(k+1)$. Second, a peak must be with a significant curvature which can be measured by the second derivative of time series; since the time series given are discrete, the second derivative at $\mathbf{d}_j(k)$ is estimated as belows:
$$\Gamma = \Delta^+ - \Delta^-$$
where
$$\Delta^- = \max(\mathbf{d}_j(k) - \mathbf{d}_j(k-1), \frac{(\mathbf{d}_j(k) - \mathbf{d}_j(k-2))}{2})$$
$$\Delta^+ = \max(\mathbf{d}_j(k+1) - \mathbf{d}_j(k), \frac{(\mathbf{d}_j(k+2) - \mathbf{d}_j(k))}{2})$$
The Δ^- and Δ^+ in (1) approximately evaluate the slope of left and right side of peak respectively, and Γ is the difference between them; thus the magnitude of Γ (the gap between two slope measurements) represents how abrubtly the curve has turned from upwards to downwards. For expressional convenience, $\Gamma^- = -\Gamma$ is referred as the curvature of a peak in the rest of this paper because Γ is always negative at a peak. Now we choose n peaks with highest Γ^- from the points satisfying the first condition, while n may vary according to the data set type.
2. In order to remove the influence of noise that every $\mathbf{d}_j$ has in common near the end of the time series, eliminate a portion of tailing peaks as follows: Identify 5 tailing peaks and select one with the hightest intensity among them. Denote i_j be the time index of this peak ($1 \leq i_j \leq K$). Let $i^* = \max_{1 \leq j \leq M} i_j$. For each $\mathbf{d}_j$, remove all peaks appearing after i^*.
3. Based on the remaining peak locations, construct a matrix called the *bonus matrix* $\mathbf{B} \in \mathbb{Z}^{K \times M}$. Let $\bar{\Gamma}$ be the mean value of Γ^- of the remaining peaks. Initialize $\mathbf{B}$ to all zero. If $\mathbf{D}(i,j)$ represents a peak, then we set $\mathbf{B}$ to be the negative gamma of the corresponding peak added by $\bar{\Gamma}/2$.

A

SHAPE			DMS			CMCT			ddTTP		
unpaired	A	1	unpaired	A	1	unpaired	A	0	paired or unpaired	A	1
	U	1		U	0		U	1		U	0
	G	1		G	0		G	1		G	0
	C	1		C	1		C	0		C	0
paired	Any	0	paired	Any	0	paired	Any	0			

B

GGAAGAUGACGCGUCUG



C

Sequence		SHAPE	DMS	CMCT	nomod	ddTTP
order	nt					
1	G	1	0	1	0	0
2	G	1	0	1	0	0
3	A	1	1	0	0	1
4	A	0	0	0	0	1
5	G	0	0	0	0	0
6	A	0	0	0	0	1
7	U	0	0	0	0	0
8	G	0	0	0	0	0
9	A	1	1	0	0	1
10	C	1	1	0	0	0
11	G	1	0	1	0	0
12	C	0	0	0	0	0
13	G	0	0	0	0	0
14	U	0	0	0	0	0
15	C	0	0	0	0	0
16	U	0	0	0	0	0
17	G	1	0	1	0	0

Fig. 2. Prediction matrix. (**a**) Definition of the values appearing in the peak prediction matrix. 1 means that a band is expected in that residue position, whereas 0 means that no band is expected. (**b**) Example target sequence and its structure predicted by the Vienna RNA package (Hofacker, 2003). (**c**) The prediction matrix for the example in (**b**).

4. Determine the ideal separation between bands based on the remaining peak locations: $\rho \triangleq (k_r - k_f)/(N-1)$, where k_f and k_r are the locations of the foremost peak and the rearmost peak respectively. The ideal seperation calculated here is used when shifting the window in dynamic programming.

2.4 Formulation as dynamic programming

In essence, the band annotation problem is to select N out of K points and match them to peak locations (if at all possible) in an optimal way. This is similar to the problem of aligning two sequences $(1, 2, \ldots, N)$ and $(1, 2, \ldots, K)$ without allowing gaps for the latter. With regard to the problem of mapping K points into peak locations, the relative location of the corresponding peak is assigned to each element of the former sequence as shown in the following example:

```
RNA sequence index : -1--2---3...N...-
Measurement index : 123456789.......K
Relative peak location: -2--0---1...0...-
```

The first three bands are located at 2, 5, and 9 time units. The relative peak location assigned to the third band is 1 implying that this band corresponds to the peak located 1 unit behind it, namely at 10 time unit. Each set of alignment and peak location correspondence is given a score that represents its probabilistic likelihood. Dynamic programming can be utilized to find the solution set with the highest score, which in turn leads to the most likely locations of bands.

One difficulty that arises in the dynamic programming approach is that the number of possibilities to be considered grows large as the number of lanes (*i.e.*, profiles) increases; if there are M lanes to consider, $M+2$ sequences must be taken into account for alignment in the dynamic programming, resulting in exponential increase of time complexity according to the

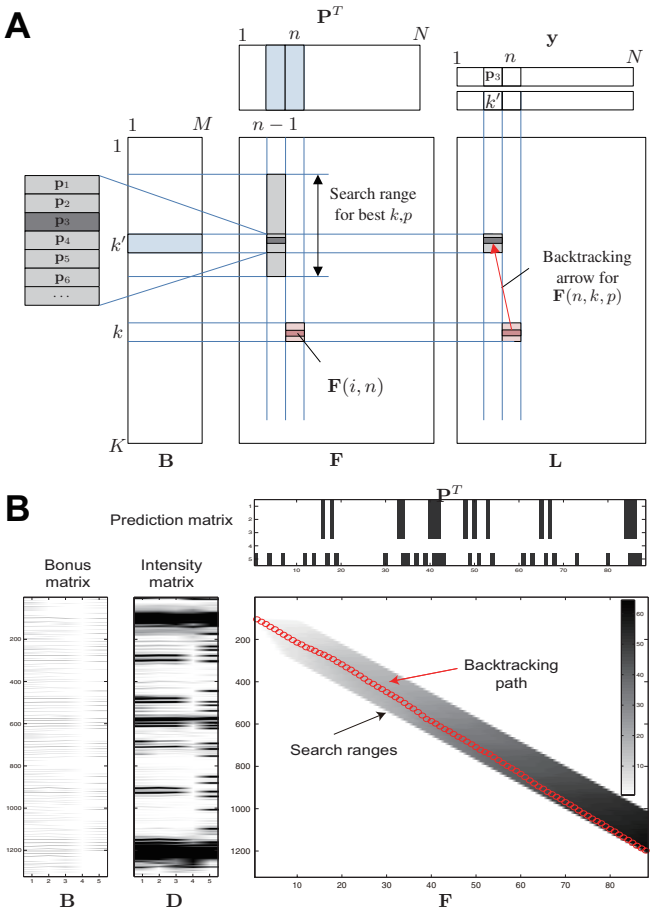


Fig. 3. Formulation as dynamic programming. (**a**) $\mathbf{F}(n,k,p)$ depends on $\mathbf{F}(n-1,k',p')$ in the previous column and the gap bonus $S(n,k',k,n)$ between them. The best tuple (k',p') that maximizes $\mathbf{F}(n,k,p)$ is searched for in the range $k-2.5\rho \le k' < k; k'+p' < k+p$ and is stored in the backtracking matrices $\mathbf{L}_k(n,k,p)$, $\mathbf{L}_p(n,k,p)$. The computation of $S(n,k',k,p)$ is based on the bonus matrix $\mathbf{B}$ and the prediction matrix $\mathbf{P}$ (Section 2.5). (**b**) Example. The data set used is 'FMN Apatamer with single binding site.' $N=88$, $M=5$, $K=1324$. The backtracking path is represented by a series of red circles superimposed on the score matrix $\mathbf{F}$; since $\mathbf{F}$ is 3-dimensional, the figure alternatively represents a reduced matrix $\mathbf{F}'$ defined by $\mathbf{F}'(n,k) = \max_{p'} \mathbf{F}(n,k,p')$. The output array $\mathbf{y}_k$, which stores the position of each circle, indicates the band locations.

increase of M. An alternative approach is not to involve peak matching problem in the dynamic programming and assign peak locations to each band only based on the alignment of the other two sequences and heuristic rules. In this paper, a simple rule is proposed where each band is assigned the nearest peak with acceptably short distance. It is a simple and intuitive rule, but there is no way to prevent a single peak from being simultaneously assigned to two consequent bands under this rule. In order to compromise between fully dynamic programming approach and heuristic rule-based approach, a primary lane is selected and its peaks are assigned to bands by the dynamic programming whereas peaks of the other lanes are assigned to bands by the heuristic rule. The last one (ddTTP) out of M profiles is chosen to be the primary lane for every intensity data set, thus $\mathbf{d}_M$ may be considered as the primary lane in the rest of this paper.

More formally, define a 3-dimensional matrix $\mathbf{F}$ indexed by n, k and p $(1 \le n \le N; 1 \le k \le K; -\rho/2 < p < \rho/2)$ where the value $\mathbf{F}(n,k,p)$

indicates the maximum score up to the band n, position k, and primary peak (peak on the primary lane) at $k+p$. The matrix $\mathbf{F}$ is filled up recursively:

$$\mathbf{F}(n,k,p) = \max_{\substack{k-2.5\rho\le k'<k \\ |p|<\rho/2 \\ k'+p'<k+p}} \left\{\mathbf{F}(n-1,k',p') + S(n,k',k,p)\right\} \quad (1)$$

where $S(n,k',k,p)$ is the score gained by going from position k' to k for band n and primary peak at $k+p$. The first constraint on k' in (1) implies that the jump from k' to k must be forward and the distance between them is capped by a reasonable upper bound so that the entire search space can be narrowed down into a probable area for efficient implementation. The search space is reduced even further into the trace of a moving window as shown in Figure 3, during our implementation. The constraint $k'+p' < k+p$ means that the corresponding primary peak must move forward as well as position k. More details of $S(n,k',k,p)$ is described in the next section.

The backtracking matrices $\mathbf{L}_k$, $\mathbf{L}_p$ for finding the solution itself are given by

$$\begin{aligned}\mathbf{L}(n,k,p) &= (\mathbf{L}_k(n,k,p), \mathbf{L}_p(n,k,p)) \quad (2)\\ &= \operatorname*{argmax}_{\substack{k-2.5\rho\le k'<k \\ |p|<\rho/2 \\ k'+p'<k+p}} \left\{\mathbf{F}(n-1,k',p') + S(n,k',k,p)\right\}\end{aligned}$$

and respectively store the position k and the relative peak location p from which $\mathbf{F}(n,k,p)$ is derived as in (1). The output array $\mathbf{y}$ is derived from $\mathbf{L}_k$ and $\mathbf{L}_p$ as follows:

$$\begin{aligned}\mathbf{y}(n) &= (\mathbf{y}_k(n), \mathbf{y}_p(n)) \quad (3)\\ &= \begin{cases} \operatorname*{argmax}_{k,p} \left\{\mathbf{F}(N,k,p)\right\}, & \text{if } n = N; \\ \mathbf{L}(n+1, \mathbf{y}_k(n+1), \mathbf{y}_p(n+1)), & 1 \le n \le N-1. \end{cases}\end{aligned}$$

The value of $\mathbf{y}_k(n)$ corresponds to the location of the n-th band in the input sequence $\mathbf{s}$. Figure 3 illustrates the proposed dynamic-programming formulation with an example.

2.5 Description of score term

The score term in (1) consists of the following two components:

$$S(n,k',k,p) = S_{\text{dist}}(n, k-k') + w_{\text{peak}} \cdot S_{\text{peak}}(k,p) \cdot \mathbf{P}(n,:) \quad (4)$$

where S_{dist} and S_{peak} are functions returning nonnegative values and $\mathbf{P}(n,:)$ is the n-th row of the prediction matrix $\mathbf{P}$.

2.5.1 Distance bonus term It is empirically supported that the length between consecutive locations, k' and k, is quite evenly distributed. S_{dist} is the bonus term that utilizes this fact and induces the dynamic programming to end up with regularly stretched output. In addition, observations on reference annotations suggest that a gap between two consecutive locations tends to be shorter when the preceding location corresponds to '`G`' in the RNA sequence. These observations lead to the definition of distance bonus term as follows:

$$S_{\text{dist}}(n,d) = \frac{f_{(\rho', \frac{\rho}{2})}(d)}{f_{(\rho', \frac{\rho}{2})}(0)} \quad (5)$$

where

$$\rho' = \begin{cases} \frac{2}{3}\rho, & \text{if } \mathbf{s}(n-1) = \texttt{G}; \\ \rho, & \text{otherwise} \end{cases}$$

and $f_{(\mu,\sigma)}$ is the density function of $N(\mu,\sigma)$. That is, $S_{\text{dist}}(n,d)$ reaches its maximum value 1 when $d = \rho'$ and decreases along a Gaussian curve as d deviates from ρ'.

2.5.2 Peak bonus term Another bonus term used in this method is the peak bonus term which is used to incline our output to locate bands near peaks with a significant curvature. As the peak bonus is granted only for the lanes with a band at the location, $P(n,:)$ must be referred before actually adding up the bonuses as demonstrated in the equation (4). S_{peak} is a function that returns a nonnegative M-dimensional value where each of its entries represents the peak bonus from each lane:

$$S_{\text{peak}} = (S^1_{\text{peak}}, S^2_{\text{peak}}, \ldots, S^M_{\text{peak}}) \quad (6)$$

where S^m_{peak} stands for the bonus from matching a peak to a band in $\mathbf{d}_m$, assuming such a band exists. The bonus is boosted for a greater curvature at the peak and the proximity of the peak to the band, so $S^m_{\text{peak}}(k,p)$ is defined as the product of a Gaussian density function and an entry of $\mathbf{B}$ corresponding to the peak:

$$S^m_{\text{peak}}(k,p) = \frac{f_{(0,\frac{\rho}{5})}(\tilde{p})}{f_{(0,\frac{\rho}{5})}(0)} \cdot \mathbf{B}(k+\tilde{p}, m) \cdot \alpha_m \quad (7)$$

where

$$\alpha_m = \begin{cases} M^{\frac{1}{2}}, & m = M; \\ 1, & \text{otherwise} \end{cases}$$

and

$$\tilde{p} = \begin{cases} p, & \text{if } m = M; \\ \operatorname*{argmin}_{\substack{q\in[-5,5] \\ \mathbf{B}(k+q)\neq 0}} |q| & \text{otherwise} \end{cases}$$

If such $\tilde{p}$ does not exist, set $S^m_{\text{peak}}(k,p) = 0$. As shown above, the peak bonus from the primary lane is given more weight because it plays an important role in finding the optimal solution with dynamic programming.

2.5.3 Peak bonus weight coefficient Two distinct bonus terms are normalized by a linear combination, where an optimal selection of weight coefficient is crucial. An overvalued weight on peak bonus is most likely to result in solutions with exact matches of bands to peaks and uneven distribution of gaps whereas an undervalued weight w_{peak} would bring about output with uniform distances neglecting peak-band matching. An optimal coefficient w_{peak} varies according to dataset and can be discovered through iterative calculations. For the dataset used here, $w_{\text{peak}} = 1.5$ leads to the best output.

2.6 Reliability evaluation

Regardless of how accurately the automated band annotation works, it may always return a misleading output. Thus, it must be supported by a method to assess the reliability of automatically generated band locations prior to practical application. The mean squared error (MSE) may be a good index to the accuracy of our result, but it is measurable only if a right solution (*i.e.*, reference) is provided; in practice, the reliability of the a result needs to be calculated without any reference, so the MSE cannot be counted on. For this reason, an alternative measurement to MSE is proposed to predict the confidence of results based on the number of certain abnormalities. The focus is on the intuition that when the dynamic programming fails to approach the desirable solution for any reason, the balance between two bonus terms is broken because the value of the weight coefficient w_{peak} itself originates from the nature of the data and its correct answer. There are two possible consequences when w_{peak} fails to arbitrate between the distance bonus term and the peak bonus term: extraordinarily short or long distances between consecutive locations, or bands on the primary lane without proper matching to peaks. We devise an indicator for the reliability of output based on the number of such occurrences as follows:

- n_1: number of bands without corresponding peak
- n_2: number of gaps with length less than or equal to $\rho/4$
- n_3: number of gaps with length greater than or equal to 2ρ
- $\mathbf{E} = \max(n_1, n_2, n_3)$

Greater $\mathbf{E}$-score means more abnormalities in the output which is believed to result from output digressing from the correct answer, so it can be expected that output with the smaller $\mathbf{E}$ would be more reliable than output with the greater $\mathbf{E}$. The relationship between the $\mathbf{E}$-score and the actual reliability of output is discussed in the result section.

Table 1. High-throughput RNA structure mapping data sets analyzed by the proposed method (total 522 profiles and 47210 bands). Excluding the last line, there are 95 data sets. More details of these 95 data sets are described in Lee *et al.* (2012). The last data set is from a study on a 187-nt ribozyme.

Name	# profiles	# nt	# bands per profile	# total bands
R45[a]	60	108	88	5280
R46[a]	80	108	88	7040
R47[b]	90	112	92	8280
R47B[b]	36	112	92	3312
R48[b]	96	112	92	8832
R49[b]	18	112	92	1656
R49B[c]	48	115	95	4560
R50[c]	54	115	95	5130
R43[d]	40	98	78	3120
HDV[e]	xx	187	yy	zz

[a] Flavin mononucleotide (FMN) aptamer with single binding site (Lee *et al.*, 2012); [b] FMN aptamer with single binding site II; [c] FMN binding branches; [d] The backwards C; [e] NMIA (SHAPE) modification of the hepatitis delta virus (HDV) ribozyme

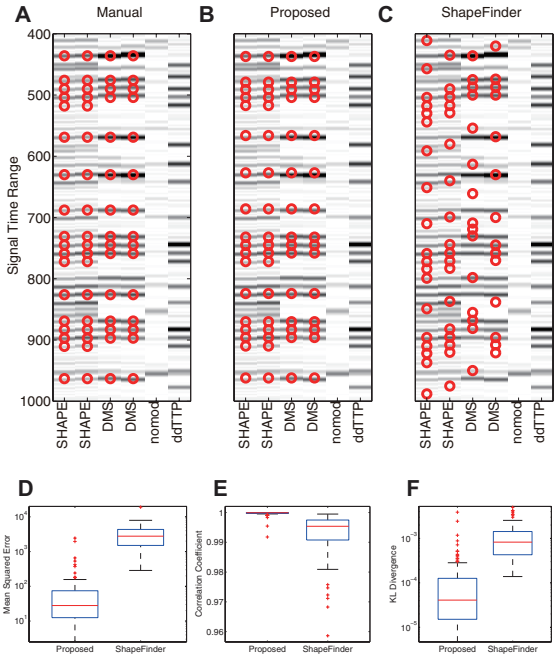


Fig. 4. Determination of band locations. (A) Reference (manual) annotation. Red circles represent band locations. (B) The band locations determined by the proposed method. (C) The band locations found by ShapeFinder (Vasa *et al.*, 2008). (D) Mean squared error (MSE) (E) Pearson's correlation coefficient (F) KL divergence

2.7 Implementation and data preparation

We implemented the proposed method in the MATLAB programming environment (The MathWorks, http://www.mathworks.com) and are making it freely available for download at http://hitrace.stanford.edu. (To be filled with data preparation...)

3 RESULTS

3.1 Robust determination of band positions

Figure 4(a)–(c) shows the electrophoretic profiles annotated with band locations by three different methods: reference, proposed, and ShapeFinder (Vasa *et al.*, 2008), respectively. The reference annotation was prepared manually and verified by experimentalists. Visual inspection suggests that the proposed method produces annotations more compatible with the reference. The annotations determined by ShapeFinder tend to be off from the reference positions downwards. We could observe this trend not only in this profile but in general. In addition, ShapeFinder seems to have difficulties in locating some band locations. For the 95 data sets used in the experiments, ShapeFinder missed 3.77% annotations on average per data set. This limitation originates from the fact that ShapeFinder does not take the number of bands to be annotated as input but determines the band locations based on the intensity and reference ladder. For fair comparison, we considered only those band locations both ShapeFinder and the proposed method identifies, not to penalize ShapeFinder for failing to mark some band locations. Supplementary Material has the band annotation plots for all the profiles used in the experiment.

To assess the accuracy of automated band annotation quantitatively, we computed the mean squared error (MSE) of the band locations determined by the proposed method with respect to the reference locations. In addition, we measured the Pearson's correlation coefficient and the KL divergence between the reference and computationally determined band positions. For comparison, we repeated calculating quantitative metrics for ShapeFinder over the same data. The box plots in Figure 4(d)–(f) reveal that the proposed method outperforms ShapeFinder in all the three metrics by large margin; the average of MSE is 29.49 times smaller, the average correlation coefficient is 21.85 times closer to 1, and the average KL divergence is 5.98 times smaller. According to our experience, a band annotation result with MSE lower than 100 is practically indistinguishable from the reference. For the proposed method, most MSE values were lower than 100 with the median MSE of 27.93. In contrast, nearly all MSE values were above 100 for ShapeFinder (median: 2776.88). This suggests that much more manual intervention would be needed if ShapeFinder were used as the band annotation tool. The MSE and correlation coefficient values for individual profiles are provided in Supplementary Material.

3.2 Accurate peak-area quantification

In the RNA structure mapping pipeline, the band annotation is followed by peak deconvolution, which fits each band with a Gaussian curve and produces the quantified area of the band. To see how the final band quantification results get improved by employing the proposed band annotation method, we calculated Pearson's correlation coefficients between band intensities quantified based on the band annotation found by the proposed method and those quantified based on the reference annotation. We also repeated the calculation with the band intensities quantified by ShapeFinder. For fair comparison, we applied the same peak deconvolution software (HiTRACE; Yoon *et al.*, 2011) to these three methods.

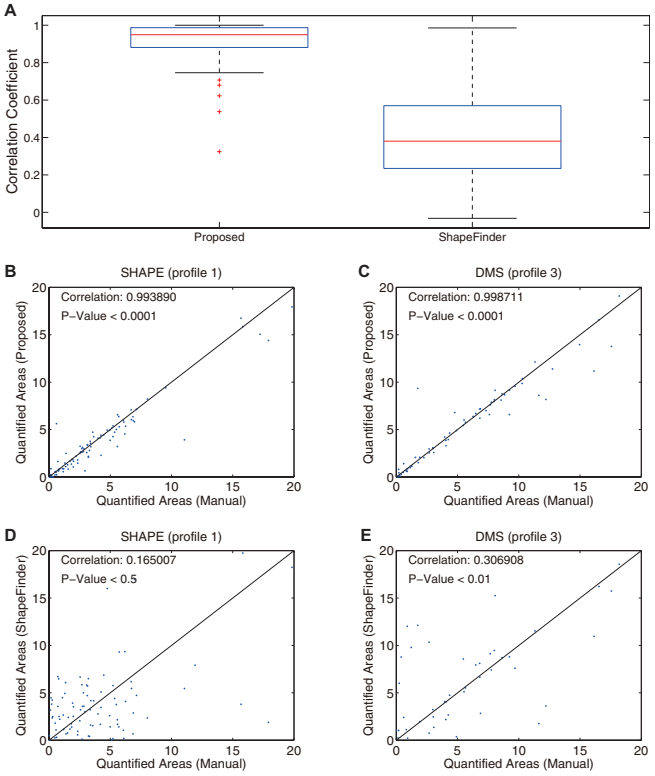


Fig. 5. Accuracy of quantifying peak areas. (**a**) The left box plot represents the distribution of the Pearson's correlation coefficients between the manually quantified areas and those quantified by the proposed method over the 95 data sets. The right box plot shows the distribution of the correlation coefficients between the areas quantified manually and by ShapeFinder. (**b**–**c**) Correlation of the reference and the quantified areas by the proposed method for data set 'FMN Binding Branches' is shown for profiles 1 (SHAPE) and 3 (DMS). (**d**–**e**) Correlation of the reference and the areas quantified by ShapeFinder for the same data set.

Figure 5(a) shows the distribution of the Pearson's correlation coefficients measured using the entire 95 data sets. The median correlation coefficient for the proposed method is 0.95 which is about 2.5 times higher than that for ShapeFinder, 0.38, and the distribution for the proposed method shows smaller variance. This observation suggests that using the proposed band annotation can significantly enhance the accuracy of band quantification. Figure 5(b) and (c) shows the correlation of results between the proposed method and reference for a specific data set (FMN Binding Branches), and Figure 5(d) and (e) shows the correlation between the ShapeFinder and reference results. As expected from Figure 5(a), the proposed method produces results that are more correlated with the reference than ShapeFinder. See Supplementary Material for the results from all individual data sets.

3.3 Results in longer RNAs

To test the proposed method with a longer RNA molecule, we used a chemical mapping data set from a study on the NMIA (SHAPE) modification of the hepatitis delta virus (HDV) ribozyme (187-bp). For this data set, we included the FAST software (Pang *et al.*, 2011)

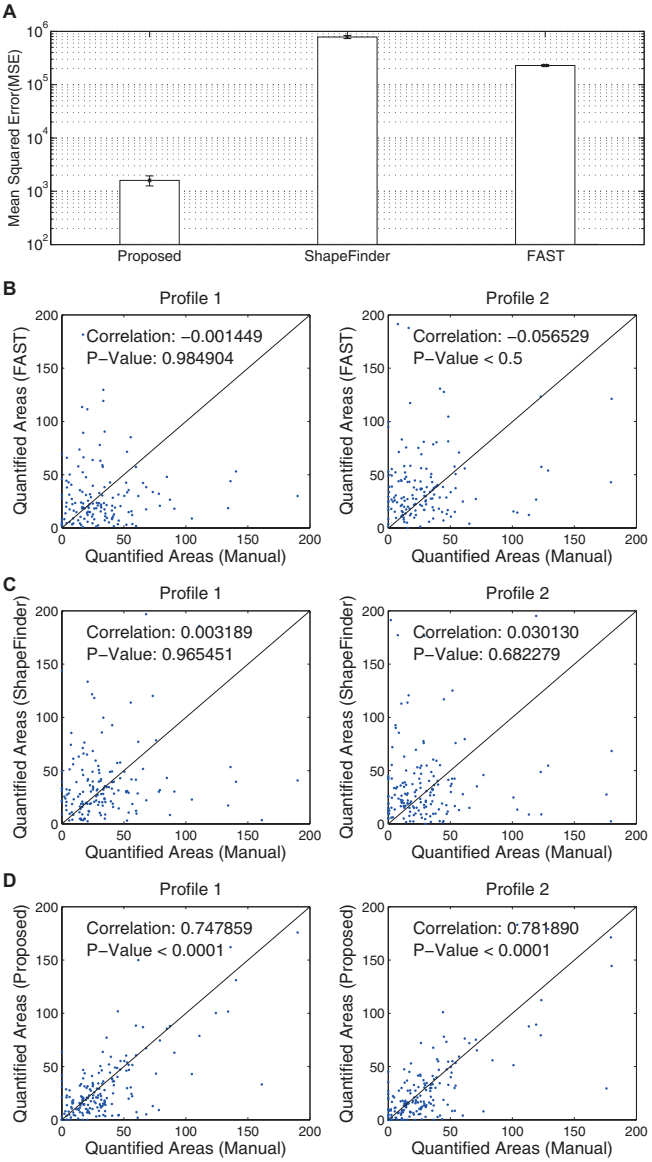


Fig. 6. Assessing band annotation results from 187-nt HDV data set. (**a**) The average MSE between the reference band locations and those determined by different approaches. (**b**–**d**) Correlation between the areas quantified manually and by FAST, ShapeFinder, and the proposed method over two profiles, respectively.

in comparison. FAST requires ddGTP as reference ladder and could not be used for the other 95 data sets presented earlier (they contain ddTTP profiles instead).

We repeated the experiments presented in Sections 3.1 and 3.2 for this HDV data set. Figure 6(a) shows the the average MSE values of band positions determined by the three methods under comparison. Compared to the reference band locations, the proposed method produced the least amount of error, although its average MSE tends to be higher than those for shorter sequences (Figure 4(d)). Still, the proposed method substantially outperformed FAST and ShapeFinder. Figure 6(c)-(d) shows the correlation of the

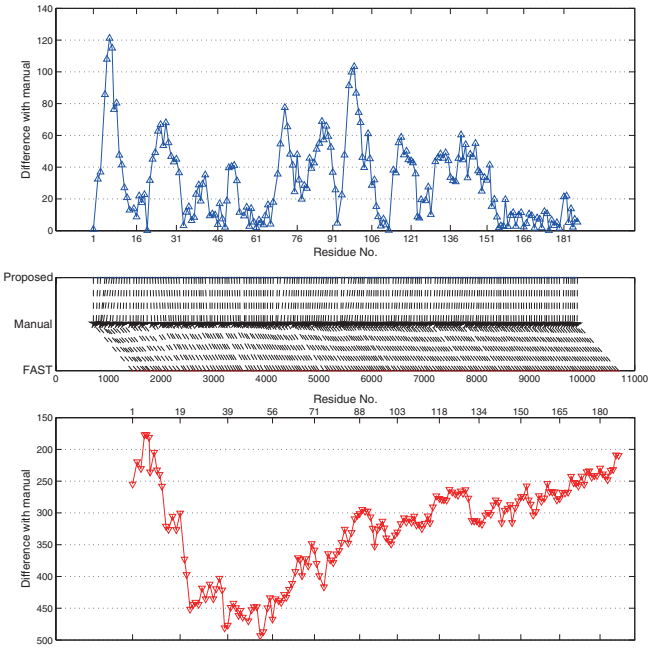


Fig. 7. Error in band positions with respect to the reference band locations for 187-nt HDV data. Upper plot: error over residue positions for the proposed method; middle: mapping between the reference and computationally predicted band locations; lower: error over residue positions for FAST.

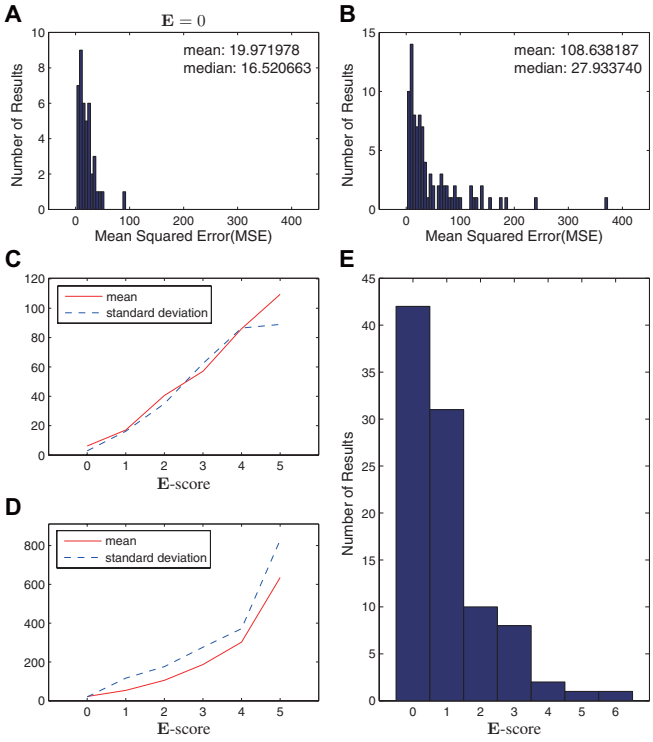


Fig. 8. (**a**) The distribution of MSE for the results with zero **E**-scores. (**b**) The distribution of MSE for the whole results. (**c**) The trends of mean and standard deviation of MSE with respect to **E**-score over artificially generated data sets from a single original data set. (**d**) The trends of mean and standard deviation of MSE with respect to **E**-score for artificially generated data sets from the whole 95 data sets. (**e**) The distribution of **E**-score over 95 data sets.

reference areas with band areas quantified by the proposed method, ShapeFinder, and FAST, respectively. The band areas quantified by the proposed method show a high degree of correlation with the reference ($r = 0.748$ and 0.782), whereas the two other methods give unsatisfactory results ($r < 0.05$).

Figure 7 shows the distribution of errors in predicted band locations over the position of residues in the HDV data set. This is to check if there is any systematic bias in band annotation on the residue position. The middle diagram represents the mapping between the reference band locations and those determined by the proposed method. The upper and lower plots show the position errors for the proposed method and FAST, respectively. For the proposed method, the errors near the start location tends to be larger than those in the middle. According to our experience, there exist very high-intensity bands in the starting and ending portion of a profile, which hinders even the manual band annotation. The larger error near these segments in a profile may be due to these high-intensity bands. The error pattern in the result from FAST shows a different pattern, and the largest errors appear near residues 40–50. Overall, the errors from the proposed method were consistently lower than those from FAST.

3.4 Strong correlation between E-score and MSE

Assuming that mean squared error is a good indicator for the accuracy of a result, an observation on the relationship between **E**-score and mean squared error may lead to the conclusion that **E**-score is a trustworthy index to the confidence of our results. Figure 8(a)–(b) show the histograms of the MSE of the output where (a) only contains results satisfying $\mathbf{E} = 0$ whereas (b) is drawn for the whole 95 data sets; consequently the histogram (a) is a subset of histogram (b). It is obvious from the shapes of figures that most results in (a) are from the left side of (b) implying that the results with high MSE can be filtered out by limiting **E**-score. For example, 41 out of 42 results under the constraint $\mathbf{E} = 0$ have MSE below 60 (even the one exception has MSE less than 100) as shown in (a), reflecting that the constraint $\mathbf{E} = 0$ almost guarantees the high reliability of band annotations. Since 95 data sets may not be sufficient to conclude on a statistically strong correlation, however, artificial data sets are generated based on the original data sets through random convolution in terms of amplitude and interval for further verification. Figure 8(c)-(d) show the trends of mean and standard deviation of MSE with respect to **E**-score, where (c) comes from artificial data generated from a single data set whereas artificial data involved in (d) is generated from the whole 95 data sets. The trends shown in (c) and (d) are consistent with the inference from the comparison of (a) with (b) in the sense that a greater **E**-score is followed by higher and more capricious distribution of mean squared error. In particular, the trends in (c) support the correlation of **E**-score and MSE in a different way from the earlier reasoning by (a)–(b) comparison because only a single data set out of the 95 originals is concerned in (c) unlike in (a) and (b). Figure 8(e) shows the histogram of the **E**-scores over the 95 data sets prepared. Approximately 45 percent of the data sets have zero **E**-score and

about 30 percent of the data sets have $\mathbf{E}$-score one showing that the proposed method gives us output with fairly low $\mathbf{E}$-scores.

4 DISCUSSION

4.1 Use of RNA secondary structure information for band annotation

In the band annotation procedure for an RNA sequence, the proposed method constructs the band prediction matrix $\mathbf{P}$ based on the RNA's secondary structure predicted by the Vienna RNA package (Hofacker, 2003). Although the secondary structure prediction software typically matches most experimental profiles closely, there may exist cases (*e.g.*, complicated pseudoknots) in which the prediction quality is low or even fails. In such cases, using secondary structure information may lead to incorrect band annotation, but through preliminary filtering we can reduce the possibility of inaccurate annotation.

4.2 Algorithm complexity and time demand of band annotation

The proposed algorithm relies on dynamic programming and a simple implementation would have the worst-case time and space complexity cubic to the input size. Even so, the practical time demand of band annotation was tolerable in our experiments. Under the experimental setup used (sequential execution on a Intel core i5 4570 processor with 8-GB main memory), the total time demand of annotating bands in all the 95 data sets did not exceed 5 min (for each data set, mean 2.8517 sec; median 2.8413 sec).

4.3 E-score

Regardless of its success in sifting out unqualified annotations, $\mathbf{E}$-score has limitations in providing various levels of reliability due to its discrete values. According to our experiences, the band annotations with $\mathbf{E} \leq 1$ are always reliable and can be safely adopted for the next process in CE analysis whereas the results with $\mathbf{E} = 2$ are relatively unreliable. Since the reliability levels such as 1.5 is not offered by $\mathbf{E}$-score, the borderline for acceptance cannot be drawn between integers. This incapability of $\mathbf{E}$-score makes it impossible to constrain our results more delicately. In the future, we hope to modify the current $\mathbf{E}$-score into a continuous measure so that the band annotation results can be broken into a far more diverse classes in terms of their reliability.

5 CONCLUSION

In the analysis of CE profiles, the band annotation is one of the most time-consuming steps, due to the lack of robust computational tools for automated annotation. Using a dynamic-programming approach, the proposed algorithm can find an optimal arrangements of bands in a given CE profile, under the carefully crafted scoring scheme suitable for high-throughput CE experiments. According to our experiments with nearly 100 CE data sets representing approximately 47,000 band locations in total, the proposed method identified the band positions matching the reference positions closely. In addition, the proposed reliability index may be consulted while deciding whether or not to adopt the automatically generated annotation.

ACKNOWLEDGMENTS

The authors thank Menashe Elazar at the Glenn Laboratory at Stanford University for providing the HDV ribozyme data.

Funding: This work was supported in part by the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. 2011-0009963 and No. 2011-0000158 to SY) and in part by a Burroughs-Wellcome Foundation Career Award at the Scientific Interface (to RD for computational work).

REFERENCES

Das, R., Karanicolas, J., and Baker, D. (2010). Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods*, **7**(4), 291–294.

Deigan, K., Li, T., Mathews, D., and Weeks, K. (2009). Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, **106**(1), 97.

Hofacker, I. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, **31**(13), 3429–3431.

Kladwang, W. and Das, R. (2010). A mutate-and-map strategy for inferring base pairs in structured nucleic acids: proof of concept on a DNA/RNA helix. *Biochemistry*, **49**(35), 7414–7416.

Lee, J., Kladwang, W., Lee, M., Cantu, D., Azizyan, M., Kim, H., Limpaecher, A., Yoon, S., Treuille, A., and Das, R. (2012). RNA design rules from a massively multiplayer cloud laboratory. *under review*.

Merino, E., Wilkinson, K., Coughlan, J., and Weeks, K. (2005). Advances in RNA structure analysis by chemical probing. *J. Am. Chem. Soc.*, **127**, 4223–4231.

Mitra, S., Shcherbakova, I., Altman, R., Brenowitz, M., and Laederach, A. (2008). High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Research*, **36**(11), e63.

Pang, P. S., Elazar, M., Pham, E. a., and Glenn, J. S. (2011). Simplified RNA secondary structure mapping by automation of SHAPE data analysis. *Nucleic Acids Research*, **39**(22), e151.

Tijerina, P., Mohr, S., and Russell, R. (2007). DMS footprinting of structured RNAs and RNA–protein complexes. *Nature Protocols*, **2**(10), 2608–2623.

Vasa, S., Guex, N., Wilkinson, K., Weeks, K., and Giddings, M. (2008). ShapeFinder: A software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA*, **14**(10), 1979–1990.

Walczak, R., Westhof, E., Carbon, P., and Krol, A. (1996). A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *Current opinion in structural biology*, **2**, 367–379.

Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W., Swanstrom, R., Burch, C. L., and Weeks, K. M. (2009). Architecture and secondary structure of an entire hiv-1 rna genome. *Nature*, **460**(7256), 711–716.

Weeks, K. (2010). Advances in RNA structure analysis by chemical probing. *Current opinion in structural biology*, **20**, 295–304.

Wilkinson, K. A., Gorelick, R. J., Vasa, S. M., Guex, N., Rein, A., Mathews, D. H., Giddings, M. C., and Weeks, K. M. (2008). High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol*, **6**(4), e96+.

Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W., and Das, R. (2011). HiTRACE: high-throughput robust analysis for capillary electrophoresis. *Bioinformatics*, **27**(13), 1798–805.