

# Automated band annotation for RNA structure probing experiments with numerous capillary electrophoresis profiles

Seungmyung Lee<sup>1</sup>, Hanjoo Kim<sup>1</sup>, Siqi Tian<sup>2</sup>, Taehoon Lee<sup>1</sup>,  
Sungroh Yoon<sup>1,3,\*</sup>, Rhiju Das<sup>2,4,\*</sup>

<sup>1</sup>Department of ECE, Seoul National University, Seoul 151-744, Korea <sup>2</sup>Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, USA <sup>3</sup>Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-744, Korea <sup>4</sup>Department of Physics, Stanford University, Stanford, CA 94305, USA

## ABSTRACT

**Motivation:** Capillary electrophoresis (CE) is a powerful approach for structural analysis of nucleic acids, with recent high-throughput variants enabling three-dimensional RNA modeling and the discovery of new rules for RNA structure design. Among the steps composing CE analysis, the process of finding each band in an electrophoretic trace and mapping it to a position in the nucleic acid sequence has required significant manual inspection and remains the most time-consuming and error-prone step. The few available tools seeking to automate this band annotation have achieved limited accuracy and have not taken advantage of information across dozens of profiles routinely acquired in high-throughput measurements.

**Results:** We present a dynamic-programming based approach to automate band annotation for high-throughput capillary electrophoresis. The approach is uniquely able to define and optimize a robust target function that takes into account multiple CE profiles (sequencing ladders, different chemical probes, different mutants) collected for the RNA. Over a large benchmark of multi-profile data sets for biological RNAs and designed RNAs from the EteRNA project, the method outperforms prior tools (QuSHAPE, FAST) significantly in terms of accuracy compared to gold-standard manual annotations. The amount of computation required is reasonable at a few seconds per data set. We also introduce an ‘E-score’ metric to automatically assess the reliability of the band annotation and show it to be practically useful in flagging uncertainties in band annotation for further inspection.

**Availability:** The implementation of the proposed algorithm is included in the HiTRACE software, freely available as an online server and for download at <http://hitrace.stanford.edu>.

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

## 1 INTRODUCTION

RNA molecules play diverse roles in encoding and regulating genetic information, and much of this versatility can be traced to the formation of intricate RNA structures. To this end, chemical probing methodologies provide a general and rapid means to

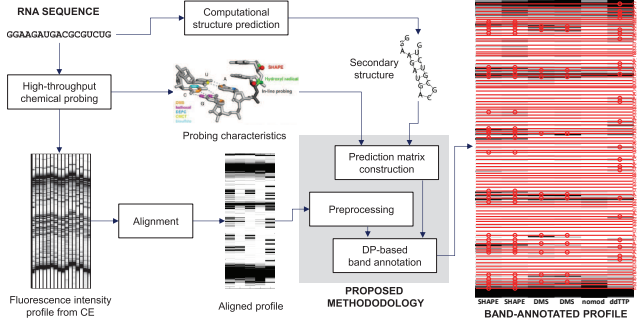
mapping RNA secondary and tertiary structure at single-nucleotide resolution (Weeks, 2010).

There exist many chemical probing techniques, most of which have common experimental procedures, as follows. Given an RNA of interest folded in solution, a chemical reagent modifies the RNA, either cleaving it or forming a covalent adduct with it at a rate correlated with the accessibility of particular moieties at each nucleotide or the frequency at which each nucleotide fluctuates into a conformation activated for chemical reaction. Examples of such chemical reagents, all with distinct mechanisms, include hydroxyl radicals, 2'-OH acylating chemicals (SHAPE), dimethyl sulfate (DMS), and CMCT (Weeks, 2010). Subsequent reverse transcription detects the modification sites as stops to primer extension at nucleotide resolution. The resulting complementary DNA (cDNA) fragments are resolved in sequencing gels followed by individually quantifying band intensities. Prior to the mid-2000s, the bottlenecks were the final steps (gel running and band quantification).

To resolve fragments in a more high-throughput fashion, capillary electrophoresis (CE) was developed and is reaching wide use. CE-based chemical probing can produce hundreds of electrophoretic profiles exhibiting tens of thousands of individual electrophoretic bands from a single experiment, leading to recent breakthroughs in two-dimensional mapping of complex RNA structures (Kladwang *et al.*, 2011) and their excited states (Tian *et al.*, 2014), and extension to large complexes such as entire viruses (Watts *et al.*, 2009) and to RNA design problems (Lee *et al.*, 2014). Further developments in next-generation sequencing readouts are promising but still show biases compared to CE measurements (Lucks *et al.*, 2011; Kladwang *et al.*, 2014).

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 bands embedded in these traces. 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), fast analysis of SHAPE traces (FAST; Pang *et al.*, 2011), and QuShape (Karabiber *et al.*, 2013).

\*To whom correspondence should be addressed



**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) traces. If available or estimated through computational prediction, we also provide the RNA’s secondary structure. From this information and the characteristics of the chemical probing method used, we derive a prediction matrix that stores expected interaction patterns across the residues and traces. Based on the aligned CE traces and prediction matrix, we apply a dynamic-programming approach that finds the optimal selection of the band locations under a well-defined scoring scheme.

A typical high-throughput CE analysis pipeline consists of the following steps (Yoon *et al.*, 2011; Karabiber *et al.*, 2013; Kladwang *et al.*, 2014): 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 an electrophoretic trace to a position in the nucleic acid sequence. For verification, visual inspection in this phase is inevitable to some extent. However, in practice, this band annotation step often takes significant manual efforts in CAFA and SHAPEfinder, for they were designed to focus more on alignment and peak fitting. HiTRACE, QuShape, and FAST have provided improved levels of band annotation support, but band annotation remains still the most time-consuming and error-prone step for large data sets.

This paper describes a dynamic-programming based approach to automated band annotation for such large CE data sets. These data sets involve at least four and up to hundreds of multiple aligned traces for each RNA, based on sequencing ladders for the four different nucleotide types, different chemical modifiers SHAPE, and/or chemical modification under different solution conditions or with different mutations. The central innovations herein are (1) an accurate and well-tested procedure to integrate information across these multiple traces into a single consensus band annotation with accuracy approaching that of manual annotation, and (2) a reliability estimator for this procedure. Figure 1 shows the overview of the proposed methodology.

## 2 METHODS

### 2.1 Problem definition

Given an RNA sequence  $s$  probed at  $N$  nucleotides, assume that we carry out the chemical structure probing of this sequence using  $M$  different treatments, each of which is run in a separate capillary lane. Assume that the fluorescence intensity of each capillary is measured over  $K$  time points.

We define a *profile* (also called a *trace*) as the sequence of intensity values from a capillary. For any particular profile, the reactivity of each nucleotide to the chemical reagent is represented at a specific location in the series of intensity values, and  $N$  such locations are sequentially spread throughout the entire profile. All profiles are assumed to be well aligned using the procedure described in Yoon *et al.* (2011) such that each nucleotide corresponds to the same location across all profiles. The entire CE measurement can then be arranged in a  $K \times M$  matrix  $\mathbf{D}$ . Normally,  $N \ll K$ , i.e. each electrophoretic profile is finely sampled in time. 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  $s$  for each of  $M$  treatments. This prediction can be arranged in a  $N \times M$  matrix  $\mathbf{P}$  called the *prediction matrix* (see below).

The problem of band annotation is therefore 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 is optimized over all possible  $\binom{K}{N}$  possibilities. The selected  $N$  points map to the locations of the nucleotides of the sequence  $s$  in the CE measurement (see Supplemental Fig. S1).

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
- $s \in \{A, C, G, U\}^N$ : the nucleotide sequence

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

### 2.2 Prediction matrix construction

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 that the nucleotide is reactive to the reagent, whereas zero indicates no reactivity. For instance, 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. We allow the use of numerous chemical probing strategies: dimethyl sulfate alkylation [DMS], carbodiimide modification [CMCT], and ‘others’ that can produce bands at all locations, including 2’-OH acylation [the SHAPE strategy] (Kladwang *et al.*, 2014). We also allow input of a secondary structure in dot-parentheses notation. Nucleotides forming base pairs are not expected to show bands in DMS, CMCT, SHAPE, and other structure mapping profiles. Sequencing experiments that terminate reverse transcription of the RNA with ddNTP incorporation produce bands after nucleotides complementary to the terminating nucleotide. Based on this information, 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  $j$ .

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

### 2.3 Initialization of candidate peaks from profiles

The first step is to locate prominent peaks on each profile (each column of  $\mathbf{D}$ ). As discussed in Section 2.1, peaks in CE profiles are the locations where significant reactivities are observed, implying that bands are more likely to exist at the same position. Thus, these peaks are matched with bands afterwards. (Here and below, ‘peak’ refers to a local maximum in each profile, of which there may be many; whereas ‘bands’ refers to the desired  $N$  band locations.) Let  $\mathbf{d}_j$  be the  $j$ -th column vector of  $\mathbf{D}$ ,  $1 \leq j \leq M$ . Briefly, the following procedure is executed.

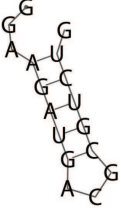
1. Select candidates for the peaks in  $\mathbf{d}_j$  that can be mapped into elements of the sequence  $s$ . These peaks are selected to satisfy the following conditions. First, a peak  $\mathbf{d}_j(k)$  must have a higher intensity (a fundamental property of a peak) than those of its neighbors,  $\mathbf{d}_j(k-1)$  and  $\mathbf{d}_j(k+1)$ . Second, a peak must be with a significant curvature

**A**

	SHAPE			DMS			CMCT			ddTTP	
unpaired	A	1	unpaired	A	1	unpaired	A	0	paired or unpaired	A <sup>a</sup>	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		C	0

**B**

GGAAGAUGACGCGUCUG



**C**

Sequence order	nt	SHAPE	DMS	CMCT	nomod	ddTTP
1	G	1	0	1	0	0
2	G	1	0	1	0	1
3	A	1	1	0	0	1
4	A	0	0	0	0	0
5	G	0	0	0	0	1
6	A	0	0	0	0	0
7	U	0	0	0	0	0
8	G	0	0	0	0	1
9	A	1	1	0	0	0
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. <sup>a</sup>The bands on ddTTP are expected to be at positions right before where As are located (and showing up right immediately afterward in electropherograms of complementary DNA). (b) Example target sequence and its estimated secondary structure, here predicted by the Vienna RNA package (Hofacker, 2003). (c) The prediction matrix for the example in (b).

which can be measured by the second derivative of time series; since the time series given are discrete, the curvature is estimated as follows:

$$\Gamma = \Delta^- - \Delta^+ \quad (1)$$

where

$$\Delta^- = \max(d_j(k) - d_j(k-1), \frac{(d_j(k) - d_j(k-2))}{2})$$

$$\Delta^+ = \min(d_j(k+1) - d_j(k), \frac{(d_j(k+2) - d_j(k))}{2})$$

The  $\Delta^-$  and  $\Delta^+$  in (1) approximate the slope of left and right side of peak respectively, and  $\Gamma$  is the difference between them; thus, the magnitude of  $\Gamma$  represents how abruptly the curve has turned from upwards to downwards. Now we choose  $N_j^{\text{peak}}$  peaks with highest  $\Gamma$  from the points satisfying the first condition, where  $N_j^{\text{peak}}$  is set twice the number of nucleotides reactive to the chemical agent used for the  $j$ -th profile (i.e. the number of ones on the  $j$ -th column of  $\mathbf{P}$ ). Call these candidate peak locations  $A_j^i$  ( $1 \leq i \leq N_j^{\text{peak}}$ ).

- In preparation for the sampling scheme and scorefunction computation below, estimate the ideal separation between bands based on the remaining peak locations:  $\rho \triangleq (\max_j k_j^f - \min_j k_j^f) / (N - 1)$ , where  $k_j^f$  and  $k_j^r$  are the locations of the foremost peak and the rearmost peak respectively on the  $j$ -th profile.
- In preparation for the scorefunction computation below, construct a matrix based on these candidate peak locations called the *bonus matrix*  $\mathbf{B} \in \mathbb{Z}^{K \times M}$ . Let  $\bar{\Gamma}$  be the mean value of  $\Gamma_i$  of the candidate peaks. Initialize  $\mathbf{B}$  to all zero. At each peak  $A_j^i$ , we apply a uniform bonus, supplemented by a stronger bonus at sharp peaks:  $\mathbf{B}(A_j^i, j) = \bar{\Gamma}/2 + \Gamma_i$ .

## 2.4 Formulation as dynamic programming

**2.4.1 Basic motivation** 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  $(1, 2, \dots, N)$  and  $(1, 2, \dots, K)$  without allowing gaps for the latter.

RNA sequence index : -1--2---3...N...-

Measurement index : 123456789.....K

In the example above, the first three bands are located at 2, 5, and 9 time units. In order to find the most probable one among all such alignments, each possible alignment is given a score that represents its 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. More formally, define a matrix  $\mathbf{F}$  indexed by  $n$  and  $k$  ( $1 \leq n \leq N$ ;  $1 \leq k \leq K$ ) where the value  $\mathbf{F}(n, k)$  indicates the maximum score up to the band  $n$  and position  $k$ . (More details on  $\mathbf{F}$  are given below.) The matrix  $\mathbf{F}$  is filled up recursively:

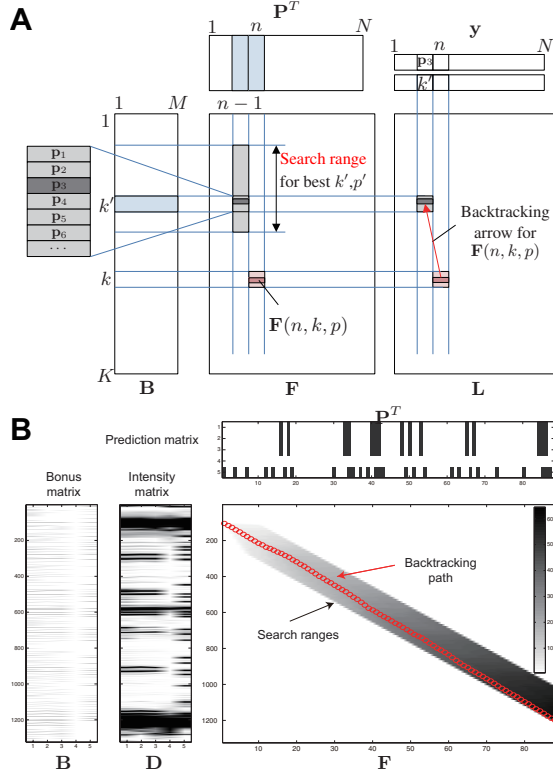
$$\mathbf{F}(n, k) = \max_{k-2.5\rho \leq k' < k} \{ \mathbf{F}(n-1, k') + S(n, k', k) \} \quad (2)$$

where  $S(n, k', k)$  is the score attained by going from position  $k'$  to  $k$  for band  $n$ . As shown in eq (2), the mappings in  $\mathbf{F}(n, k)$  consists of mapping band  $n$  to location  $k$  in addition to the solution for  $\mathbf{F}(n-1, k')$ , where  $k'$  is the  $\text{argmax}$  in (2). The constraint on  $k'$  in (2) implies that a jump from  $k'$  to  $k$  is forward and its width is capped by a reasonable upper bound so that the entire search space can be narrowed down for efficient implementation; it was also confirmed through tests that the existence of upper bound does not affect the outcome.

**2.4.2 Degeneracy breaking and primary profile** In the proposed method, a band is allowed to be matched to a candidate peak even if their positions are slightly off from each other; in other words, an exact positional coincidence is not required for a peak-band matching (see Section 2.5.2 for detail). Thus, the formalization of our problem in the previous section fails to guarantee that two different bands will be matched to distinct closest peaks. If there are two bands close to each other and only one peak is available for matching, both bands might be matched with that single closest peak at the same time. In order to avoid such degeneracies, an additional search variable  $p$  is introduced: the relative position of the matched peak to the band position  $k$ . The tuple  $(n, k, p)$  corresponds to the instance in which the band  $n$  is located at position  $k$ , and matched with the peak at  $k+p$  if there is any; no score bonus if there is no peak at the position. The matrix  $\mathbf{F}$  is now redefined as a 3-dimensional matrix as follows:

$$\mathbf{F}(n, k, p) = \max_{\substack{k-2.5\rho \leq k' < k \\ |p| < \rho/2 \\ k' + p' < k + p}} \{ \mathbf{F}(n-1, k', p') + S(n, k', k, p) \} \quad (3)$$

The constraint  $|p| < \rho/2$  is to restrict bands to be matched only with nearby peaks, and the last constraint  $k' + p' < k + p$  means that two distinct bands cannot share the same peak. One problem that arises with the use of  $p$  is that there should be  $M$  such  $p$ 's for  $M$  profiles, implying that the matrix  $\mathbf{F}$  should not be 3-dimensional but actually  $(M+2)$ -dimensional. However, this would make solving this problem too costly. As a compromise, the problem is simplified by choosing one primary profile among  $M$  profiles so that  $p$  is applied only to it; therefore  $\mathbf{F}$  may remain as a 3-dimensional matrix. Our software automatically determines the primary profile based on the data type with a preference for sequencing ladders. For our data sets, the last profile (a ddTTP ladder) was selected; without loss of generality,  $\mathbf{d}_M$  will be considered as the primary profile in the rest of this paper.



**Fig. 3.** Formulation as dynamic programming. (a)  $F(n, k, p)$  depends on  $F(n-1, k', p')$  in the previous column and the gap bonus  $S(n, k', k, p)$  between them. The best tuple  $(k', p')$  that maximizes  $F(n, k, p)$  is searched for in the range  $k-2.5\rho \leq k' < k$ ;  $k'+p' < k+p$  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.

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

$$\begin{aligned} \mathbf{L}(n, k, p) &= (\mathbf{L}_k(n, k, p), \mathbf{L}_p(n, k, p)) \\ &= \underset{\substack{k-2.5\rho \leq k' < k \\ |p| < \rho/2 \\ k'+p' < k+p}}{\operatorname{argmax}} \{ \mathbf{F}(n-1, k', p') + S(n, k', k, p) \} \end{aligned} \quad (4)$$

and respectively store the position  $k'$  and the relative peak location  $p'$  from which  $\mathbf{F}(n, k, p)$  is derived as in (3). The output array  $\mathbf{y}$  is derived from  $\mathbf{L}_k$  and  $\mathbf{L}_p$  as follows:

$$\begin{aligned} \mathbf{y}(n) &= (\mathbf{y}_k(n), \mathbf{y}_p(n)) \\ &= \begin{cases} \underset{k,p}{\operatorname{argmax}} \{ \mathbf{F}(N, k, p) \}, & \text{if } n = N; \\ \mathbf{L}(n+1, \mathbf{y}_k(n+1), \mathbf{y}_p(n+1)), & 1 \leq n \leq N-1. \end{cases} \end{aligned} \quad (5)$$

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 (3) consists of the following two components:

$$S(n, k', k, p) = S_{\text{dist}}(n, k-k') + w_{\text{peak}} \cdot S_{\text{peak}}(k, p) \cdot \mathbf{P}(n, :) \quad (6)$$

where  $S_{\text{dist}}$  and  $S_{\text{peak}}$  are functions returning vectors of nonnegative elements, and  $\mathbf{P}(n, :)$  is the  $n$ -th row of the prediction matrix  $\mathbf{P}$ . The dot product in the second term is a sum over all lanes  $m$  from 1 to  $M$ . A coefficient  $w_{\text{peak}}$  of 1.0 gave acceptable annotations in initial tests and was not further optimized.

**2.5.1 Distance bonus term** It is empirically supported that the length between consecutive locations,  $k'$  and  $k$ , is quite evenly distributed.  $S_{\text{dist}}$  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 (Mills and Kramer, 1979; Sasaki et al., 1998). These observations lead to the definition of distance bonus term as follows:

$$S_{\text{dist}}(n, d) = \frac{f_{(\rho', \frac{\rho}{2})}(d)}{f_{(\rho', \frac{\rho}{2})}(\rho')} \quad (7)$$

where

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

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  $\rho'$ .

**2.5.2 Peak bonus term** The second score term favors band locations near peaks of the electrophoretic profiles with a significant curvature. As the peak bonus is granted only for the profiles with a band at the location,  $\mathbf{P}(n, :)$  must be referred before actually adding up the bonuses as demonstrated in the equation (6).  $S_{\text{peak}}$  is a function that returns a nonnegative  $M$ -dimensional value where each of its entries represents the peak bonus from each profile:

$$S_{\text{peak}}(k, p) = (S_{\text{peak}}^1(k), \dots, S_{\text{peak}}^{M-1}(k), S_{\text{peak}}^M(k, p)) \quad (8)$$

where  $S_{\text{peak}}^m$  stands for the bonus from matching a peak to a band in  $\mathbf{d}_m$ , assuming such a band exists. The bonus was designed to be boosted for a greater curvature at the peak and the proximity of the peak to the band, so  $S_{\text{peak}}^m$  is defined as the product of a Gaussian density function and an entry of  $\mathbf{B}$  corresponding to the candidate peak closest to location  $k$ :

$$S_{\text{peak}}^m(k) = \max_{|q| < \rho/2} \frac{f_{(0, \frac{\rho}{5})}(q)}{f_{(0, \frac{\rho}{5})}(0)} \cdot \mathbf{B}(k+q, m) \quad (9)$$

for  $m < M$ , and

$$S_{\text{peak}}^M(k, p) = \frac{f_{(0, \frac{\rho}{5})}(p)}{f_{(0, \frac{\rho}{5})}(0)} \cdot \mathbf{B}(k+p, m) \cdot (M-1) \quad (10)$$

As described above, this peak bonus is taken from the primary profile (typically a sequencing ladder) rather than searching for optimal peak/band matches across all profiles to allow degeneracy breaking at reasonable computational expense. [A separate dynamic-programming-based band annotation algorithm was also tested which does not carry out the peak/band degeneracy breaking of eq. (10) and gave slightly worse performance; see Supplemental Figure S2.]

## 2.6 Reliability evaluation

While the presented band annotation method was found to be quite accurate, it was not perfect. We therefore sought a method to assess the reliability of automatically determined band locations prior to practical application. We devised a score to predict the quality of results. The idea behind the score is that when optimization of eq. (6) fails to achieve the desirable



**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.* (2014).

Name	# profiles	# nt	# bands per profile	# total bands
R45 <sup>a</sup>	60	108	88	5280
R46 <sup>a</sup>	80	108	88	7040
R47 <sup>b</sup>	90	112	92	8280
R47B <sup>b</sup>	36	112	92	3312
R48 <sup>b</sup>	96	112	92	8832
R49 <sup>b</sup>	18	112	92	1656
R49B <sup>c</sup>	48	115	95	4560
R50 <sup>c</sup>	54	115	95	5130
R43 <sup>d</sup>	40	98	78	3120

<sup>a</sup> Flavin mononucleotide (FMN) aptamer with single binding site (Lee *et al.*, 2014); <sup>b</sup> FMN aptamer with single binding site II; <sup>c</sup> FMN binding branches; <sup>d</sup> The backwards C

solution, we typically see extraordinarily short or long distances between consecutive locations (little information from  $S_{\text{dist}}$ ) or bands on the primary profile without proper matching to peaks (little information from  $S_{\text{peak}}$ ). The  $E$ -score is defined with the following terms:

- $n_1$ : number of bands on the primary profile without corresponding peak
- $n_2$ : number of gaps with length less than  $\rho/4$ , or greater than  $2\rho$
- $N_M^{\text{peak}}$ : number of bands on the primary profile predicted by  $\mathbf{P}$
- $E = 1 - \max(\frac{n_1}{N_M^{\text{peak}}}, \frac{n_2}{K-1})$

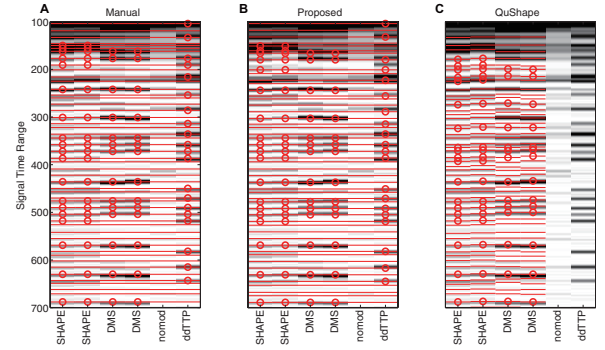
$E$ -score is a value between 0 and 1 and conservatively estimates the number of normalities in the output relative to the number of bands and locations. Greater  $E$ -score means less abnormalities in the output which is believed to result from output digressing from the correct answer, so it can be expected that output with  $E$  closer to 1 would be more reliable than output with smaller  $E$ . The relationship between  $E$ -score and accuracy is presented in the Results section.

### 3 RESULTS

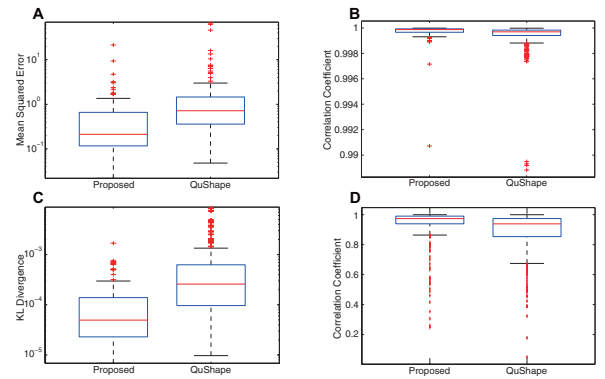
#### 3.1 Robust determination of band positions

Figure 4(a)–(c) shows the electrophoretic profiles annotated with band locations by three different methods: reference, proposed, and QuShape (Karabiber *et al.*, 2013), respectively. The reference annotation was based on expert assignments carried out at the time of data acquisition (Lee *et al.*, 2014). QuShape was chosen as the comparison target for its superior accuracy in band annotation relative to other software we tested, FAST and ShapeFinder (data not shown); nomod and ddTTP profiles were used as references (RXS1, BGS1) while running QuShape. Visual inspection suggests that the proposed method produces annotations more compatible with the reference. In this profile, the annotation determined by QuShape deviates from the reference position, particularly near the beginning of sequence.

To generally and quantitatively assess the accuracy of automated band annotation, we applied the proposed method and QuShape to 95 data sets acquired in the EteRNA project (Table 1). For both methods, we computed the mean squared error (MSE) of the band

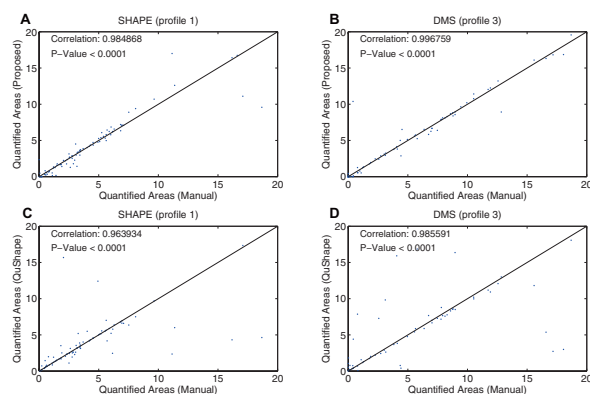


**Fig. 4.** Determination of band locations for data set ‘ViennaRNA design 03.’ (a) Reference (manual) annotation. Red horizontal lines represent all determined band locations corresponding to RNA sequence. Red circles represent the bands reactive to chemical agents for each profile. (b) The band locations determined by the proposed method. (c) The band locations found by QuShape (Karabiber *et al.*, 2013).



**Fig. 5.** Proposed method (left) vs. QuShape (right). Each plot represents each metric's distribution across 95 data sets. (a) Mean squared error (MSE) for band locations (b) Pearson's correlation coefficient for band locations (c) KL divergence for band locations (d) Pearson's correlation coefficients for area quantification. MSE units are normalized so that average distance between band locations is unity.

locations determined by the proposed method with respect to the reference locations, in units of average distance between locations. For a sense of scale, the typical MSE achieved by expert annotation is 0.15, based on comparisons of different experts' annotations with each other and to next-generation-sequencing-based measurements, where sequence annotation is unambiguous (Kladwang *et al.*, 2014); see Supplemental Fig. S3. In our experience, a band annotation result with MSE lower than 0.5 typically requires no or a small number single-click corrections. The box plots in Figure 5(a)–(c) and individual MSE values (Supplemental Tables S1 and S2) reveal that the proposed method outperforms QuShape across the data sets. For example, the median MSE of the proposed method is 0.21, well under our target value of 0.5, compared to 0.72 from QuSHAPE. As separate metrics of accuracy, we measured the Pearson's correlation coefficient and the Kullback-Leibler (KL) divergence between the reference and computationally determined band positions. Again, the average correlation coefficient of the proposed method is 1.68 times closer to 1, and the average



**Fig. 6.** Accuracy of quantifying peak areas for data set ‘FMN Binding Branches’. (a–b) Correlation of the reference and the quantified areas by the proposed method is shown for profiles 1 (SHAPE) and 3 (DMS). (c–d) Correlation of the reference and the areas quantified by QuShape.

KL divergence is 5.84 times smaller. These results quantitatively confirm what we observed qualitatively on using these tools: significantly less manual intervention is needed with the proposed method compared to QuShape.

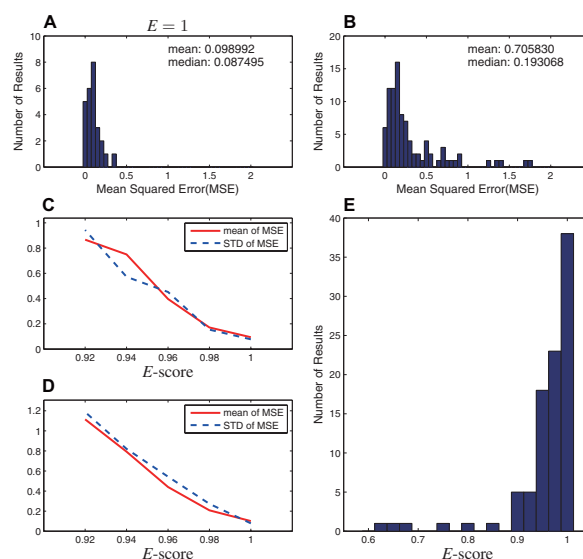
### 3.2 Accurate peak-area quantification

In the RNA structure mapping pipeline, the band annotation is followed by peak deconvolution, which fits each band with a Gaussian curve and outputs the quantified area of the band. To see how these final band quantification results are impacted by the band annotation method, we calculated Pearson’s correlation coefficients between band areas 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 QuShape. For fair comparison, we applied the same peak deconvolution software (HiTRACE; Yoon *et al.*, 2011) to these three methods.

As one example, Figure 6(a)–(b) shows the correlation of results between the proposed method and reference for a specific data set (FMN Binding Branches) for two chemical modification strategies (SHAPE and DMS). Figure 6(c)–(d) shows the correlation between the QuShape and reference results, which is visually worse than the proposed method in both cases. Over all the data sets, Figure 5(d) and Supplemental Table S1 gives the distribution of the Pearson’s correlation coefficients. The median correlation coefficient for the proposed method is 0.976, which is higher than that for QuShape (0.939) 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.

### 3.3 *E*-score reliability metric predicts MSE accuracy

In Section 2.6, we proposed *E*-score to evaluate the quality of results from our method. We assessed the use of *E*-score based on its ability to predict the accuracy of the band annotations compared to gold standard annotations, quantitatively evaluated as mean squared error (MSE). Figure 7(a) show the distribution of the MSEs where (a) only contains results satisfying  $E = 1.0$ ; the MSE values



**Fig. 7.** (a) Distribution of MSE for the results with 1 *E*-score. (b) Distribution of MSE for the whole 95 results. (5 results with MSE > 2 are omitted for better demonstration) (c) Trends of mean and standard deviation of MSE with respect to *E*-score over artificial data generated from a single original data set. (d) Trends of mean and standard deviation of MSE with respect to *E*-score for artificial data generated from the whole 95 data sets. (e) Distribution of *E*-score over 95 data sets.

are substantially smaller than those in (b), which includes all 95 data sets. For example, all 26 results under constraint  $E = 1.0$  have MSE below 0.5 as shown in (a), confirming that a ‘perfect’ *E*-score essentially guarantees high quality of band annotations; furthermore, 50 out of 51 results with  $E > 0.97$  have MSE below 0.5 (even the one exception has MSE less than 1). In addition to this experimental test, artificial data sets were generated based on the original data sets through random convolution in terms of amplitude and interval for further verification. Figure 7(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) further confirm that a lower *E*-score corresponds to MSE values with higher (worse) mean and standard deviation. Figure 7(e) shows the histogram of the *E*-scores over the 95 data sets prepared. Overall, 39% of the data sets have *E*-score equal to 1, and 84% have *E*-score greater than 0.97, suggesting that poor *E*-scores and subsequent detail manual correction will be encountered in a minority of cases.

### 3.4 Results in longer, biological RNA sequences

In an effort to test the proposed method’s compatibility with a wide array of high-throughput RNA structure mapping data sets, we prepared sample experimental data sets of biologically derived RNAs. These additional 21 data sets include Class I ligase (Bagby *et al.*, 2009), the Tetrahymena L-21 ScaI ribozyme (Russell *et al.*, 2006), a four-way junction from the *E. coli* 16S ribosomal RNA (Tian *et al.*, 2014), RNA replicases (C19, tC19 and tC19Z) (Wochner *et al.*, 2011), human Hox transcripts 5’ UTR (Hox5 and Hox9D189) (Xue *et al.*, 2014) and RNA Puzzle entries (#5-10,

and 12) (Cruz *et al.*, 2012). In each data set, complete sets of chemical modifier reactions (nomod, SHAPE, DMS, CMCT) and reference ladders (ddNTPs) are present. In addition, a hepatitis delta virus genomic segment studied previously allowed direct comparison to the FAST software (Supplemental Fig. S4) (Pang *et al.*, 2011). These RNAs had lengths up to 400 nucleotides, significantly longer than the 100-nt EteRNA designs (Table 1). Despite this increase in length, the band annotation results from the proposed method were still highly consistent with the reference expert annotation. Excluding an abnormal result from L-21 caused by an experimental issue that disallowed alignment of sequencing ladders, the maximum of MSE is only 0.68. Furthermore, the two worst MSE values (0.68 and 0.63) and two lowest *E*-scores (0.83 and 0.90) coincide in the results for AdoCbl(noref) and tRNA, confirming *E*-score's utility.

## 4 DISCUSSION

The proposed method for band annotation is unique in its ability to take into account all available CE profiles; prior methods (such as those available in QuShape and FAST) have focused on a single profile at a time with a reference profile if needed. The distinctive robustness of the proposed method is primarily attributed to this capability to integrate information across profiles. The method does require an accurate alignment of all profiles prior to band annotation. Our prior work (Yoon *et al.*, 2011) described a different dynamic programming algorithm to accomplish this preceding alignment based on standards co-loaded with each sample. In well over 100 data sets analyzed here, we saw only one case where inter-profile alignment was problematic (L-21 ScaI group I intron) and required manual intervention. Therefore, our alignment and annotation results herein confirm that all steps, including alignment and annotation, of RNA structure mapping CE analysis can now be routinely achieved through automated algorithms.

To flag cases with uncertain automated band annotation, we have introduced the *E*-score for reliability estimation. According to our experiences, given any data set for CE analysis, the band annotations with  $E > 0.97$  are almost always reliable and can be safely adopted for final steps of band quantitation whereas the results with  $E \leq 0.97$  are less likely to be reliable. Informally, we have encountered data sets in which even expert annotation is ambiguous and has required special additional experiments (such as co-loading sequencing ladders in the same color as the sample) to resolve (Tian *et al.*, 2014). This suggests that automated band annotation cannot improve much further; a valuable development would be reliability estimates for specific subsets of bands rather than a global number. An additional useful development would be use of known band intensities based on prior experiments or on base pair probability estimates, rather than coarse predictions for profiles based on sequence, modifier, and a single secondary structure.

The proposed algorithm has order of  $NK$  time and space complexity, and the practical time demand of band annotation was reasonable in our experiments. The proposed method was implemented in the MATLAB programming environment (The MathWorks, <http://www.mathworks.com>), and 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 4 min (for each data set, mean 2.2837 sec; median 2.2707 sec).

## 5 CONCLUSION

In the analysis of CE profiles, band annotation has remained the most time-consuming and error-prone step, due to the lack of robust computational tools for automating the process. Using a dynamic-programming approach, the proposed algorithm can find an optimal arrangement of bands in a given CE profile, under a scoring scheme suitable for high-throughput CE experiments with multiple profiles. On over 100 CE data sets including designed and biological RNAs, the proposed method identified the band positions matching the reference positions with accuracy sufficiently high as to obviate or significantly reduce manual correction. Finally, the quality of the band positions are well predicted by *E*-score, flagging unreliable annotations to the user.

## 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

- Bagby, S. C. *et al.* (2009). A class I ligase ribozyme with reduced  $Mg^{2+}$  dependence: Selection, sequence analysis, and identification of functional tertiary interactions. *RNA*, **15**(12), 2129–2146.
- Cruz, J. A. *et al.* (2012). RNA-puzzles: A CASP-like evaluation of rna three-dimensional structure prediction. *RNA*, **18**(4), 610–625.
- Hofacker, I. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, **31**(13), 3429–3431.
- Karabiber, F., McGinnis, J. L., Favorov, O. V., and Weeks, K. M. (2013). Qushape: Rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA*, **19**(1), 63–73.
- Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011). A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat Chem*, **3**(12), 954–962. 10.1038/nchem.1176.
- Kladwang, W., Mann, T. H., Becka, A., Tian, S., Kim, H., Yoon, S., and Das, R. (2014). Standardization of rna chemical mapping experiments. *Biochemistry*, **53**(19), 3063–3065.
- Lee, J., Kladwang, W., Lee, M., Cantu, D., Azizyan, M., Kim, H., Limpaecher, A., Yoon, S., Treuille, A., Das, R., and Participants, E. (2014). RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences*, **111**(6), 2122–2127.
- Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., and Arkin, A. P. (2011). Multiplexed rna structure characterization with selective 2-hydroxyl acylation analyzed by primer extension sequencing (shape-seq). *Proceedings of the National Academy of Sciences*, **108**(27), 11063–11068.
- Mills, D. R. and Kramer, F. R. (1979). Structure-independent nucleotide sequence analysis. *Proceedings of the National Academy of Sciences*, **76**(5), 2232–2235.
- 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.

- Russell, R., Das, R., Suh, H., Travers, K., A, L., MA, E., and D, H. (2006). The paradoxical behavior of a highly structured misfolded intermediate in rna folding. *J. Mol. Biol.*, **363**(2), 531–44.
- Sasaki, N., Izawa, M., Sugahara, Y., Tanaka, T., Watahiki, M., Ohara, E., Funaki, H., Yoneda, Y., Ozawa, K., Matsuura, S., et al. (1998). Identification of stable rna hairpins causing band compression in transcriptional sequencing and their elimination by use of inosine triphosphate. *Gene*, **222**(1), 17–24.
- Tian, S., Cordero, P., Kladwang, W., and Das, R. (2014). High-throughput mutate-map-rescue evaluates shape-directed rna structure and uncovers excited states. *RNA*.
- Vasa, S. et al. (2008). ShapeFinder: A software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA*, **14**(10), 1979–1990.
- 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.
- Wochner, A., Attwater, J., Coulson, A., and Holliger, P. (2011). Ribozyme-catalyzed transcription of an active ribozyme. *Science*, **332**(6026), 209–212.
- Xue, S., Tian, S., Fujii, K., Kladwang, W., Das, R., and Barna, M. (2014). RNA regulons in Hox 5' UTRs confer ribosome specificity to gene regulation. *Nature*, page 10.1038/nature14010 (in press).
- 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.