BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 13 2011, pages 1798–1805
doi:10.1093/bioinformatics/btr277

Structural bioinformatics

Advance Access publication May 10, 2011

HiTRACE: high-throughput robust analysis for capillary electrophoresis

Sungroh Yoon[1,*], Jinkyu Kim[1], Justine Hum[2], Hanjoo Kim[1], Seunghyun Park[1], Wipapat Kladwang[2] and Rhiju Das[2,3,*]

[1]School of Electrical Engineering, Korea University, Seoul 136-713, Republic of Korea, [2]Department of Biochemistry, Stanford University, Stanford, CA 94305, USA and [3]Department of Physics, Stanford University, Stanford, CA 94305, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Capillary electrophoresis (CE) of nucleic acids is a workhorse technology underlying high-throughput genome analysis and large-scale chemical mapping for nucleic acid structural inference. Despite the wide availability of CE-based instruments, there remain challenges in leveraging their full power for quantitative analysis of RNA and DNA structure, thermodynamics and kinetics. In particular, the slow rate and poor automation of available analysis tools have bottlenecked a new generation of studies involving hundreds of CE profiles per experiment.

Results: We propose a computational method called *high-throughput robust analysis for capillary electrophoresis* (HiTRACE) to automate the key tasks in large-scale nucleic acid CE analysis, including the profile alignment that has heretofore been a rate-limiting step in the highest throughput experiments. We illustrate the application of HiTRACE on 13 datasets representing 4 different RNAs, 3 chemical modification strategies and up to 480 single mutant variants; the largest datasets each include 87 360 bands. By applying a series of robust dynamic programming algorithms, HiTRACE outperforms prior tools in terms of alignment and fitting quality, as assessed by measures including the correlation between quantified band intensities between replicate datasets. Furthermore, while the smallest of these datasets required 7–10 h of manual intervention using prior approaches, HiTRACE quantitation of even the largest datasets herein was achieved in 3–12 min. The HiTRACE method, therefore, resolves a critical barrier to the efficient and accurate analysis of nucleic acid structure in experiments involving tens of thousands of electrophoretic bands.

Availability: HiTRACE is freely available for download at http://hitrace.stanford.edu.

Contact: sryoon@korea.ac.kr; rhiju@stanford.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 7, 2010; revised on April 12, 2011; accepted on April 27, 2011

1 INTRODUCTION

Capillary electrophoresis (CE) is a widely used approach for biochemical analysis. The rapid electrophoretic separation of fluorescently labeled nucleic acid fragments inside electrolyte-filled capillaries significantly accelerated genome sequencing (Ruiz-Martinez *et al.*, 1993; Woolley and Mathies, 1995). A more recent, powerful application of CE enables the high-throughput structure analysis of self-assembling nucleic acid-containing systems (Mitra *et al.*, 2008; Vasa *et al.*, 2008; Das *et al.*, 2010; Kladwang and Das, 2010; Weeks, 2010) as complex as viruses (Watts *et al.*, 2009; Wilkinson *et al.*, 2008) and ribosomes (Deigan *et al.*, 2009) at single-nucleotide resolution.

The CE profiles obtained in this recent generation of 'structure-mapping' experiments present tens of thousands of individual electrophoretic bands; quantifying these data gives detailed portraits of nucleic acid structure, folding thermodynamics and kinetics but requires significant informatics efforts (Mitra *et al.*, 2008). 'Base-calling' software packages can assign sequences to these bands in special four-color experiments (see, e.g. Ewing and Green, 1998; Ewing *et al.*, 1998) but are not applicable to structure mapping experiments, which require more robust sequence annotation and quantitative fits of each profile to a sum of peak shapes. Such quantitative analysis is aided by the design of experiments so that the desired information appears as differences between corresponding bands across profiles [see, e.g. (Das *et al.*, 2005; Kladwang and Das, 2010)]; then, sequence annotation of one profile results in annotation of corresponding bands across the entire data. For these datasets, tools for *alignment of features* or 'rectification' (Das *et al.*, 2005; Laederach *et al.*, 2008) across different profiles resulted in improvements in quantification speed and accuracy, but these tools remain poorly automated. As the experimental steps of large-scale CE measurements continue to accelerate, the bioinformatic task of profile alignment has become a rate-limiting step in carrying out these information-rich structural studies.

Current approaches to aligning and fitting capillary profiles include capillary automated footprinting analysis (CAFA; Mitra *et al.*, 2008) and ShapeFinder (Vasa *et al.*, 2008); we have found these methods difficult to apply to large-scale titration or mutate-and-map datasets (Kladwang and Das, 2010; Kladwang *et al.*, 2011). For instance, CAFA is focused more on peak fitting and has limited alignment capabilities. The ShapeFinder alignment function can align spectrally separated products within a single capillary but not profiles across multiple capillaries with initially poor alignment. Use of these tools requires tedious manual intervention and risks bias or unnecessary errors from such manipulation. Analysis tools for alignment and peak fitting have also been proposed in other domains

*To whom correspondence should be addressed.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from bioinformatics.oxfordjournals.org at Korea University Library on June 27, 2011

such as chromatography (Nielsen *et al.*, 1998; Tomasi *et al.*, 2004), mass spectrometry (Kazmi *et al.*, 2006; Wong *et al.*, 2005) and slab gel electrophoresis (Das *et al.*, 2005; Laederach *et al.*, 2008), but, empirically, these approaches give unsatisfactory performance for CE data.

To address the limitations of existing methods, we have developed *high-throughput robust analysis for capillary electrophoresis* (HiTRACE) to automate the alignment and quantification of nucleic acid structure mapping profiles obtained from hundreds of capillaries. As depicted in Figure 1, the proposed method consists of four major steps: preprocessing (step A), correlation-optimized linear alignment (step B), dynamic programming-based non-linear adjustments (step C), sequence annotation (step D) and peak fitting (step E). After describing the core algorithms that underlie the robust automation of each step, we present quantitative comparisons illustrating the substantial boosts in both accuracy and speed of HiTRACE over previous approaches. With the proposed methodology, the previously rate-limiting step of quantifying high-throughput CE data is now faster than experimental data acquisition times, enabling the investigation of nucleic acid structure at an unprecedented rate.

2 METHODS

2.1 Experimental setup

An experimental protocol that is optimal for HiTRACE alignment and quantification has been developed; complete descriptions of reaction components, purification procedures and sequencing ladder generation have been given previously (Das *et al.*, 2010; Kladwang and Das, 2010; Kladwang *et al.*, 2011). Briefly, RNA samples were chemically modified under the desired solution conditions and then reverse transcribed with primers (labeled at the 5′ ends with the rhodamine green fluorophore) complementary to the 3′ end of the RNA. Because the reverse transcription stops at modified nucleotides, the length distribution of the resulting DNA products encodes the chemical reactivities of the RNA. Length separation of the DNA was carried out on Applied Biosystems ABI 3100 and ABI 3730 sequencers; these instruments permit the single-nucleotide separation of products as long as 500 nt for 16 and 96 samples, respectively. To facilitate HiTRACE alignment, all samples were coloaded with a reference ladder that fluoresces in a different color and provides fiducial markers that are identical between samples. The ladder, prepared in a large batch for many experiments, was derived by reverse transcribing an arbitrary RNA (typically the 202-nt P4-P6 RNA) with a Texas red-labeled primer.

2.2 Assumptions and definitions

CE profiles each contain hundreds of 'bands' (when the data are viewed in gray scale) or 'peaks' (when the intensity is plotted as a function of electrophoresis time) whose intensities or areas report on individual residues of a nucleic acid sequence. In CE experiments that use hundreds of capillaries, profiles are typically obtained in multiple batches of experiments, e.g. with 16 capillaries in an ABI 3100 sequencer, as illustrated in Figure 1. The first profile of each batch is designated the reference to which other profiles of the batch should be aligned. Each profile i represents fluorescence intensity measured at uniformly spaced time points (here, 0.1 s) denoted by $n=1,2,\ldots,N$ with associated intensity values $y_i(n)$. As shown in Figure 1, the horizontal and vertical axes correspond to the profile index and the measurement position in time points, respectively. Fluorescence intensity levels are represented in gray scale, with nucleic acid species of different lengths appearing as separated, dark bands. The desired final output of the proposed methodology is a set of aligned profiles with their quantified band areas.

2.3 Preprocessing (step A)

In a typical profile, the starting and ending regions contain no signal. To accelerate subsequent steps, the user has the option of defining a window that brackets the electrophoretic signals in all the profiles. As another preprocessing step, we subtract an offset, constant within each profile, so as to bring the signal to zero at the boundaries of the window; this step corrects for overall drift in signal baselines that are observed in sequencer detectors. We have also implemented an option to derive and subtract a smooth (but not necessarily linear) baseline from each profile by using a procedure similar to Xi and Rocke (2008). This operation removes smoothly varying backgrounds in fluorescence signal sporadically seen in experimental CE profiles and, empirically, brings independent replicates into closer agreement.

2.4 Alignment by linear transformation (step B.1)

The first alignment step involves a linear scaling and shifting of the time axis based on maximizing the correlation coefficient between each fluorescence profile $y_i(n)$ and the reference profile $y_1(n)$ within each batch:

$$(\Delta_i^*,\sigma_i^*)=\underset{(\Delta_i,\sigma_i)\in D_i\times S_i}{\operatorname{argmax}}\left\{\operatorname{corr}\left[y_1(n),y_i\left(\frac{n}{\sigma_i}-\Delta_i\right)\right]\right\} \quad (1)$$

where D_i and S_i represent the sets of possible values of the shift Δ_i and scale factor σ_i, respectively, and $D_i\times S_i$ denotes their Cartesian product. Based on the values found above, we first time scale each profile $y_i(n)$ by σ_i^* using linear interpolation and then shift it by Δ_i^*. The correlation coefficient was chosen as the optimization target because it is independent of signal offset and scaling and has been widely used in other alignment tasks (Bylund *et al.*, 2002; Nielsen *et al.*, 1998; Pravdova *et al.*, 2002; Tomasi *et al.*, 2004). We carry out the search over shifts (Δ_i) efficiently through a Fast Fourier Transform (FFT; Oppenheim and Schafer, 2009). By default, we carry out the alignment based on the reference ladder pattern that is coloaded with each sample (see above).

2.5 Alignment between batches (step B.2)

Due to variabilities between batches, performing only the intrabatch alignment above produces stratified alignment results, where a number of up-and-down 'stairs' appear. To resolve this problem, we perform an additional *interbatch* alignment.This step constructs a representative profile of each batch by calculating the average of the first, middle and last profiles selected from each batch. We align these representative profiles to the first representative profile by the procedure above. Assume that, for the representative profile from batch b, we have determined Δ_b^* and σ_b^* values. We then realign all the profiles in batch b using these Δ_b^* and σ_b^* values (see Fig. 1B). More details of step B.2 can be found in the Supplementary Material.

2.6 Non-linear alignment (step C)

With current CE equipment, we found that it was not feasible to get complete alignment for profiles with single-band resolution through just linear scaling of the time axis. There are two reasons for this problem. First, the electrophoretic mobilities of the same products in different capillaries, or for the same capillary used at different times, can vary due to temperature differences and geometry differences. As a result, long profiles, containing hundreds of bands measured over tens of minutes, can be aligned well over the initial part of the data (e.g. the first 2 min) or the final part of the data (e.g. the last 2 min), but both parts cannot be simultaneously aligned with a single linear transformation. Second, we often run capillary electrophoresis experiments for structure mapping of molecules with slightly different sequences, e.g. libraries of single-mutation constructs (Kladwang and Das, 2010; Kladwang *et al.*, 2011). This leads to small perturbations in the band mobilities at the site of the mutation and requires a locally non-linear transformation to permit alignment. To correct for both these issues, we perform another round of refining the alignment.

Downloaded from bioinformatics.oxfordjournals.org at Korea University Library on June 27, 2011

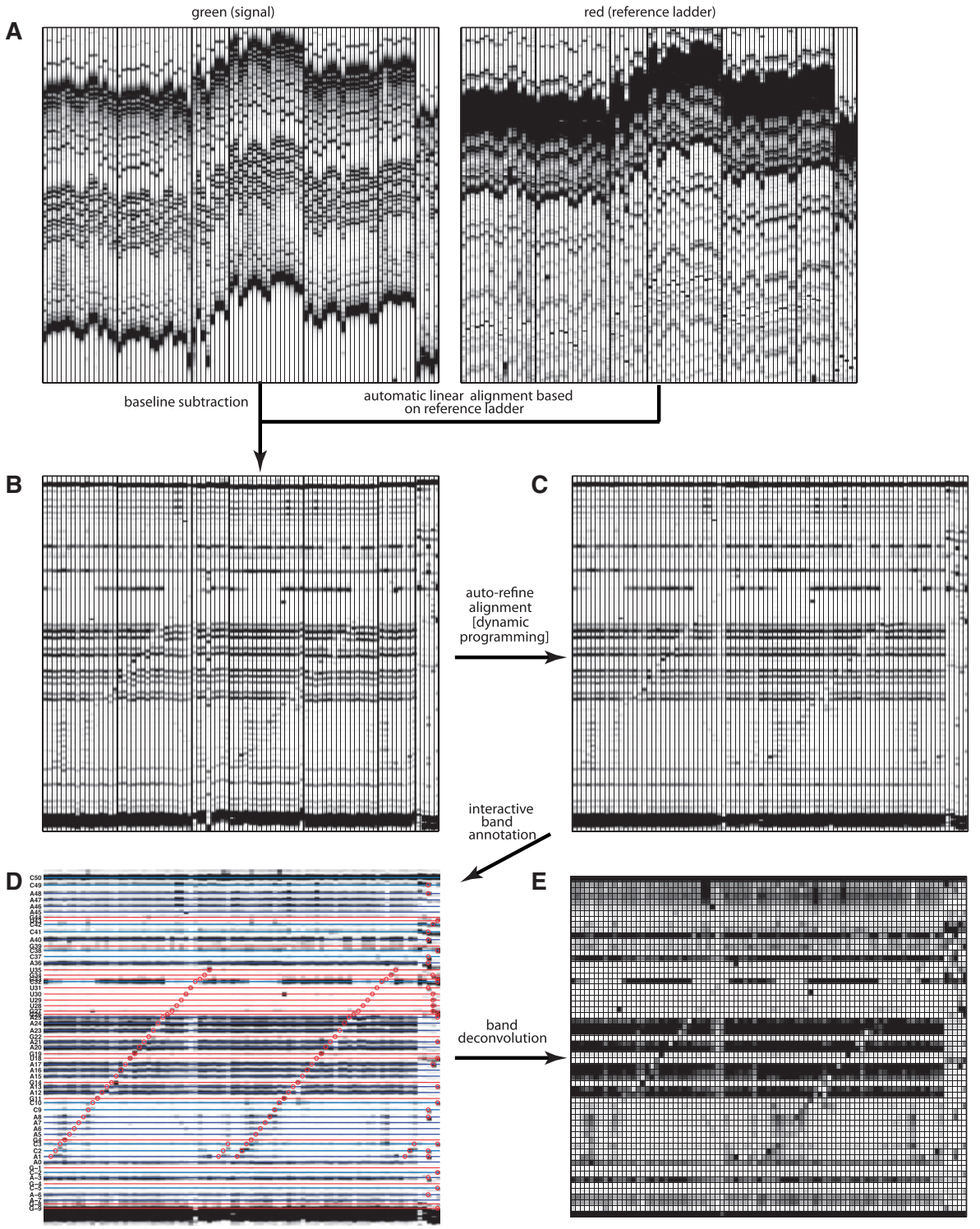

Fig. 1. Overview of the proposed HiTRACE methodology. (**A**) Raw electropherograms for an example dataset. (left) Dimethyl sulfate (DMS) modification of the MedLoop RNA (Kladwang *et al.*, 2011), read out by reverse transcription with rhodamine green-labeled primers followed by DNA separation by capillary electrophoresis; data shown are DMS profiles for 80 (of 120) single nucleotide mutants, 2 replicate controls without chemical modification and sequencing ladders for C, U and G. (right) Electropherograms of the Texas red-labeled DNA ladder that was coloaded with each sample to produce fiducial markers for alignment. (**B**) Profiles after automated preprocessing (baseline subtraction) and correlation-optimized linear alignment. (**C**) Profiles after automated alignment refinement by dynamic programming-based non-linear adjustments. (**D**) Interactive sequence annotation guided by features (red circles) at mutation positions and bands in the sequencing ladder. Blue, cyan, orange and red lines correspond to modifications at A, C, G and U, respectively. (**E**) Quantitated band areas (the final output of HiTRACE). Shorter DNA fragments (higher mobility) are at the top of each panel.

Downloaded from bioinformatics.oxfordjournals.org at Korea University Library on June 27, 2011

The concept underlying the non-linear alignment is depicted in the Supplementary Material, and resembles the warping method presented in Nielsen *et al.* (1998) for chromatographic data. We break the time axis of a non-reference profile into m-pixel windows and then shift each window boundary within a predefined range over the reference profile so as to maximize the correlation between profiles summed over all windows. We assume that the window ordering is preserved during alignment. The number of possible arrangements in this setup is large but can be enumerated efficiently by a dynamic programming (DP) approach (Bylund *et al.*, 2002; Cormen *et al.*, 2009; Nielsen *et al.*, 1998; Robinson *et al.*, 2007) that recursively solves the problem for the first window, then the first two windows, etc. As in steps B.1 and B.2, we accelerate the calculation by computing correlation coefficients through FFT. Example results are shown in Figure 1C. The Supplementary Material includes a graphical description of determining the shift amount for each window edge for aligning two example profiles.

2.7 Sequence annotation (step D)

Each band in a fluorescence profile corresponds to a position in the nucleic acid sequence. Currently, we carry out sequence annotation interactively and manually, as this encourages visual inspection of the data and makes use of expert knowledge to ensure accurate annotation. This step is accelerated compared to prior approaches (Mitra *et al.*, 2008; Vasa *et al.*, 2008) through visual feedback. Sequence assignments are made based on Sanger sequencing ladders included in the experiments; as the user makes assignments, 'guidemarks' appear at expected band positions (one residue longer than the corresponding position of modification, due to dideoxynucleotide incorporation). These guidemark positions can be visually confirmed or adjusted to overlay on experimental bands (see circles in Fig. 1D). In addition, these guidemarks can be set to appear on A and C positions for dimethyl sulfate alkylation experiments (Peattie and Gilbert, 1980; Tijerina *et al.*, 2007), as well as mutated positions in mutate-and-map experiments (Fig. 1D), which typically give visually distinct perturbations in chemical modification. These features provide cross-checks on the sequencing ladder that confirm accuracy. Due to the alignment of traces achieved in previous steps, sequence annotations need to only be provided once and are applicable to all traces. Automated annotation procedures are also being developed and will be incorporated in future versions of HiTRACE.

2.8 Band deconvolution and quantification (step E)

In this last step of HiTRACE, we approximate each profile $y(n)$ for $n = 1,2,\ldots,N$ by a sum $f(n)$ of K Gaussian curves with the form

$$f(n) = \sum_{k=1}^{K} A_k \exp\left[-\frac{(n-\mu_k)^2}{2\sigma_k^2}\right] \quad (2)$$

such that the deviation defined by

$$\sqrt{\frac{1}{N}\sum_{n=1}^{N}[f(n)-y(n)]^2} \quad (3)$$

is minimized. A_k, μ_k and σ_k are the parameters that determine the amplitude, the center location and the width, respectively, of a peak modeled by a Gaussian. We find the optimal values of these parameters by a standard Levenberg–Marquardt optimization technique for least-square minimization (Levenberg, 1944; Marquardt, 1963), and report the area of each peak as the final output.

2.9 Implementation and data preparation

We implemented the proposed HiTRACE methodology in the MATLAB programming environment (The MathWorks, http://www.mathworks.com) and are making it freely available for download at http://hitrace.stanford.edu.

For comparison with HiTRACE, we also prepared the implementations of the five different profile analysis algorithms: CAFA (Mitra *et al.*, 2008), ShapeFinder (Vasa *et al.*, 2008), msalign (Kazmi *et al.*, 2006), SpecAlign (Wong *et al.*, 2005) and COW (Tomasi *et al.*, 2004). We could not apply some methods to all situations due to their intrinsic limitations. For instance, the alignment feature of CAFA and ShapeFinder requires significant manual intervention to handle hundreds of profiles; we did not include ShapeFinder in the alignment result comparison. Similarly, msalign, SpecAlign and COW can align profiles but do not carry out peak fitting. We thus excluded them in fitting result comparisons.

2.10 Criteria for evaluating alignment results

We applied two widely used mathematical criteria—the mean squared error (MSE; Kay, 1993) of aligned peak positions with respect to the reference peaks and the Kullback–Leibler (KL) divergence (Cover and Thomas, 2006) between reference and non-reference profiles.

In MSE computation, we consider the position p of each peak in the reference profile as the true value being estimated, and use the position $\hat{p}$ of the aligned peak on a non-reference profile as the estimator of p. The MSE for the j-th reference peak p_j is then

$$\mathrm{MSE}_j = E\left[(\hat{p}_j - p_j)^2\right] = \frac{1}{L}\sum_{i=1}^{L}(\hat{p}_{ij} - p_{1j})^2 \quad (4)$$

where L is the number of profiles in the dataset used, and p_{1j} and $\hat{p}_{ij}$ represent the positions of the j-th reference peak and the peak on profile i that is aligned to p_j, respectively. For the peak detection step involved in the MSE computation, we used the peak algorithm described by Mitra *et al.* (2008), which is specifically designed for finding peaks in CE profiles and shows satisfactory performance for our purpose.

To evaluate the alignment results from an information-theoretic perspective, without explicitly considering specific peaks or band positions, we utilized the KL divergence. We calculated the KL divergence between the reference profile $y_1(n)$ and a non-reference profile $y_i(n)$ as

$$D_{\mathrm{KL}}(y_1||y_i) = \sum_{n=1}^{N} y_1(n)\log\frac{y_1(n)}{y_i(n)} \quad (5)$$

where N is the number of pixels in each profile. We repeat this calculation for every reference and non-reference pair in a dataset. Before computing KL divergence, intensity values were limited to two SDs above the mean to prevent KL divergence values from being dominated by strong bands at the beginning and end of each profile.

3 RESULTS

3.1 High-throughput RNA structure mapping datasets

To test HiTRACE, we collected 13 nucleic acid structure mapping experiments read out by capillary electrophoresis (Table 1). These datasets were diverse: probed molecules included artificial model systems (the MedLoop RNA and the X20/H20 RNA/DNA system) as well as natural structured RNAs (a conserved domain from the signal recognition particle and the P4-P6 domain of the Tetrahymena group I ribozyme), with lengths between 60 and 202 nt. Three common chemical modification strategies were represented in the data: dimethyl sulfate alkylation (Tijerina *et al.*, 2007), carbodiimide modification (Walczak *et al.*, 1996) and 2′-OH acylation [the SHAPE strategy (Merino *et al.*, 2005)]. In addition, the datasets were challenging in their size. Three experiments each gave 182 bands over 480 electropherograms, for a total of 87 360 bands per dataset. Finally, to test the precision of quantification relative to other sources of error, five experiments were conducted twice by

Downloaded from bioinformatics.oxfordjournals.org at Korea University Library on June 27, 2011

Table 1. High-throughput RNA structure mapping datasets analyzed by HiTRACE

Name	Profiles (n)	Bands per profile (n)	Total bands (n)
X20/H20 DMS-1[a]	98	40	3920
X20/H20 DMS-2[a]	88	40	3520
MedLoop DMS-1[b]	120	60	7200
MedLoop DMS-2[b]	136	60	8160
MedLoop CMCT-1[b]	128	60	7680
MedLoop CMCT-2[b]	120	60	7200
SRP DMS-1[c]	88	60	5280
SRP DMS-2[c]	96	60	5760
SRP CMCT-1[c]	88	60	5280
SRP CMCT-2[c]	88	60	5280
P4-P6 DMS[c]	480	182	87 360
P4-P6 CMCT[c]	480	182	87 360
P4-P6 SHAPE[c]	480	182	87 360

SRP, signal recognition particle conserved domain; P4-P6, P4-P6 domain of the Tetrahymena group I ribozyme; DMS, dimethyl sulfate; CMCT, 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluenesulfonate; SHAPE, selective hydroxyl acylation analyzed by primer extension.
[a]Kladwang and Das, 2010.
[b]Kladwang *et al.*, 2011.
[c]This work.

two independent researchers. Additional datasets were collected to confirm HiTRACE's ability to quantify data for RNAs over 400 nt in length (the L-21 ScaI Tetrahymena group I ribozyme) and to compare overlapping SHAPE data derived from reverse transcription starting at different primers on the same RNA (the P4-P6 domain). Overall, these datasets provide a diverse and challenging benchmark of nucleic acid CE experiments at the large scale permitted by current high-throughput experimental protocols.

3.2 Robust alignment of CE profiles

As the most basic test, we first compared the alignment results of HiTRACE with previously available methodologies by visual inspection (Fig. 2). Prior to alignment, CE experiments gave initially poor alignments of DMS chemical mapping profiles for the 60 band MedLoop RNA and the 182 band P4-P6 RNA ('Raw' in Fig. 2A and B). Application of automated HiTRACE alignment aligns the strong bands across all profiles ('HiTRACE' in Fig. 2A and B; see also Fig. 1A–C). In the alignment results produced by methods other than HiTRACE, profiles within each group tend to be reasonably aligned, whereas profile groups are not well-aligned. We did not observe this 'stratification' problem in the HiTRACE result, mainly due to the interbatch alignment step (B.2) used by HiTRACE. Additionally, comparing HiTRACE results with SpecAlign and CAFA results reveals the effectiveness of the HiTRACE non-linear alignment step, which adapts alignment to weakly varying electrophoretic rates along the profile. In the alignment results produced by SpecAlign and CAFA, some parts of the profiles appear reasonably aligned, but the top (SpecAlign) or bottom (CAFA) portions are not well aligned. For msalign and COW, this problem is much more noticeable.

For more quantitative evaluation of profile alignments, we compared the different methodologies in terms of two mathematical criteria, mean squared error in peak position (MSE) and KL divergence between profiles. Figure 2C and D show the distributions of the average MSE and KL divergence values over the 13 datasets

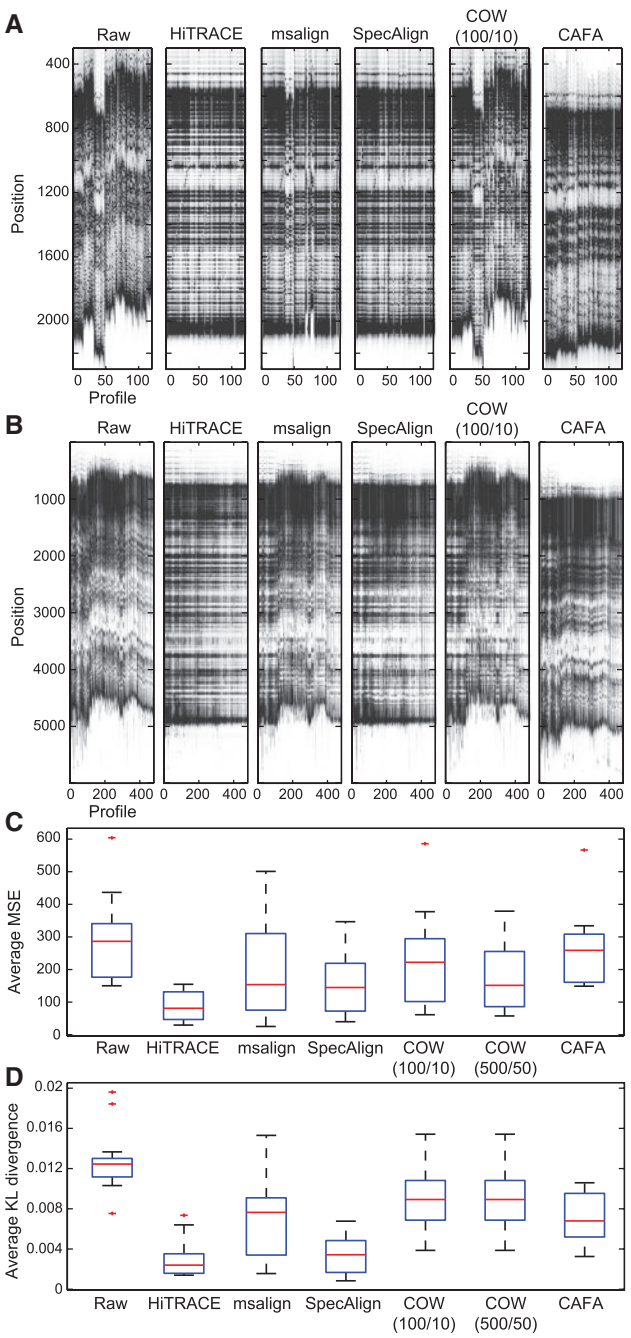

Fig. 2. Comparison of available alignment strategies for nucleic acid CE profiles. (**A**) Comparison of electrophoretic profiles of the 88 profile MedLoop DMS mutate-and-map dataset (Kladwang *et al.*, 2011) (replicate 2) before alignment and after alignment by HiTRACE, msalign (Kazmi *et al.*, 2006), SpecAlign (Wong *et al.*, 2005), COW (Tomasi *et al.*, 2004) and CAFA (Mitra *et al.*, 2008). (**B**) Alignment results for the 480 profile P4-P6 DMS dataset. (**C**) Quantitative comparison of alignment results for all 13 datasets based on MSE (Kay, 1993) of aligned peak positions with respect to the reference peaks. The line in the middle of a box is median value; box boundaries represent 75th and 25th percentiles; error bars represent the most extreme values whose distance from the box is less than 1.5 times the box length; plus symbols are outliers beyond this range. (**D**) Quantitative comparison in terms of the Kullback–Leibler divergence (Cover and Thomas, 2006) between reference and non-reference profiles.

Downloaded from bioinformatics.oxfordjournals.org at Korea University Library on June 27, 2011

used for different algorithms. With respect to HiTRACE, the alternative methodologies produced poorer results, 1.73–3.09 and 1.51–3.94 times higher median MSE and KL divergence values, respectively.

3.3 Leveraging accurate alignments into accurate quantification

To assess the accuracy of the entire quantification procedure, including alignment, sequence annotation and band deconvolution, we compared final quantified results between HiTRACE and previously available software for RNA structure mapping CE data, using two MedLoop DMS mutate-and-map datasets (Kladwang *et al.*, 2011) (see also Supplementary Material for a comparison with the X20/H20 DMS data). Each set contained at least 120 profiles with 60 bands, for a total of 7200 data points per set. (Further comparisons between software packages were precluded by the difficulty of carrying out the analysis with prior software: ShapeFinder gave poor alignment even after several hours of manual intervention, and CAFA analysis required 10 h of manual adjustment for this data set and would have required days for larger data sets.)

The MedLoop sets gave excellent Pearson's correlation coefficients between band intensities quantified with HiTRACE to those quantified with CAFA (r of 0.979 and 0.965; Fig. 3A and B), confirming the lack of any major systematic errors introduced by the HiTRACE method. We hypothesized that the small, residual variance between the methods might stem from user-introduced variation during alignment (CAFA) or sequence assignment of bands (in CAFA and HiTRACE). To test this hypothesis, we carried out replicate quantification of the same datasets; the second independent analysis gave values with correlation coefficient (r) to the first analysis of 0.987 and 0.989 (HiTRACE; Fig. 3C and D) and 0.989 and 0.974 (CAFA; Fig. 3E and F). We conclude that any differences between HiTRACE and CAFA can be explained by imprecision (variance of 1.1–1.3% in HiTRACE and 1.1–2.6% in CAFA) introduced by users; this error is much smaller than variances arising from experimental error, as is discussed in the next section.

3.4 Consistency in band quantification between experimental replicates

A stringent measure of the accuracy of an experiment and its analysis is the correlation of quantified intensities between independent replicates. The goodness of this correlation is determined by experimental factors, including small variations in sample purity, pipetting errors, temperature differences and variable times of each experimental step, and is also sensitive to any uncertainties arising from the data analysis procedure. We compared correlation coefficients between separate independent replicates of the MedLoop DMS mutate-and-map experiments (Kladwang *et al.*, 2011), quantified by both HiTRACE and CAFA (Fig. 4A and B). In both cases, the cross-replicate correlations (0.89–0.90) are significantly lower than the intrareplicate comparisons (0.97–0.99) above, verifying that variances in experimental procedures exceed any variances in the data quantification.

The throughput of HiTRACE quantification enabled us to carry out this cross-replicate comparison for the additional replicate sets (see Supplementary Material) and to explore whether alternative data processing schemes might improve the precision of the HiTRACE quantification. We tested a computationally expensive

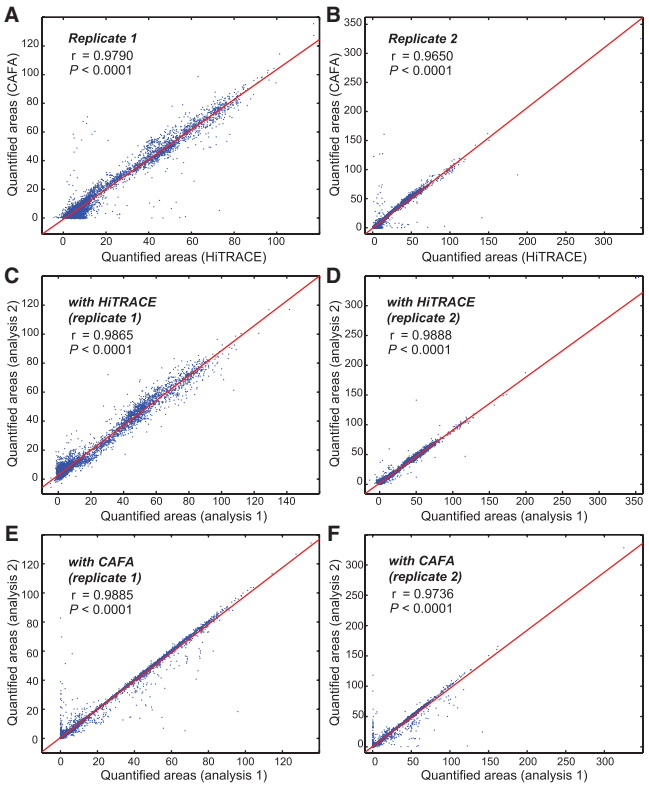

Fig. 3. Quantification accuracy and precision for HiTRACE and CAFA. Correlation of HiTRACE and CAFA results for two MedLoop DMS datasets (**A**) and (**B**) confirms the absence of any systematic deviation between the two approaches. Precision of HiTRACE (**C**) and (**D**) is similar or better than CAFA (**E**) and (**F**), based on independent analyses of the same dataset.

band deconvolution procedure [previously used in SAFA (Das *et al.*, 2005)] that optimized centers of fitted Gaussians for each individual profile. We observed indistinguishable cross-replicate correlation coefficients (Fig. 4C) with this procedure as compared to the the default HiTRACE method (no refinement of band centers). This comparison further validated the high quality of the profile-to-profile alignment in earlier HiTRACE steps, and motivated our choice to make as the HiTRACE default the 10- to 100-fold faster band-deconvolution procedure without band center refinement. We observed similarly invariant or slightly worse correlation coefficients in experiments without the baseline subtraction procedure; with additional alignment steps of 'binarized' profiles; and with other methods to automatically refine band positions in each profile (see Supplementary Material).

3.5 Reduced time demand of quantification

Although HiTRACE relies on multiple steps for accurate analysis, the time demand of quantification by HiTRACE was considerably smaller (a few minutes) than the time required by prior informatic approaches as well as the time involved in preparing and obtaining the CE experiments (a few hours). Figure 5 shows the average running time of HiTRACE for different datasets, along with the breakdown of the running time. The largest dataset (P4-P6 CMCT; 480 profiles and 87 360 bands) took approximately 12 min to

Downloaded from bioinformatics.oxfordjournals.org at Korea University Library on June 27, 2011

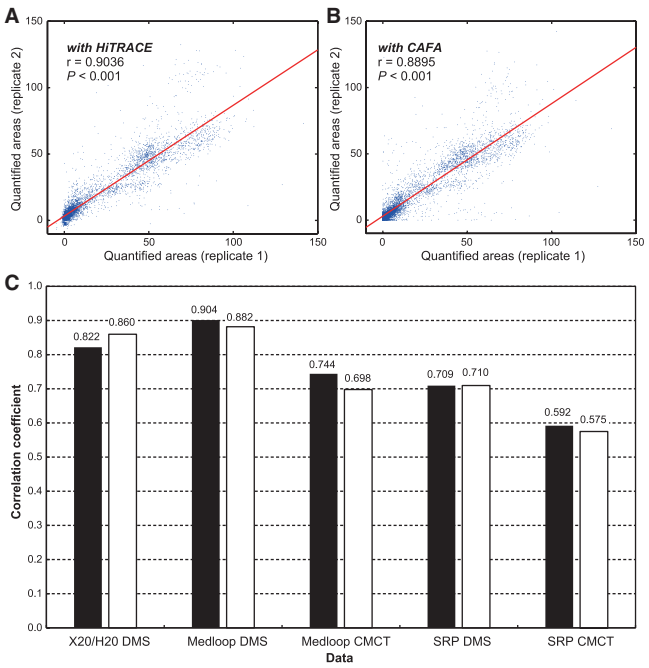

Fig. 4. Correlation of results between experimental replicates for HiTRACE (A) and CAFA (B) for the MedLoop DMS mutate-and-map experiments (Kladwang *et al.*, 2011), and (C) for HiTRACE on five replicate data sets without (black bars) and with (white bars) optimization of Gaussian positions during band deconvolution.

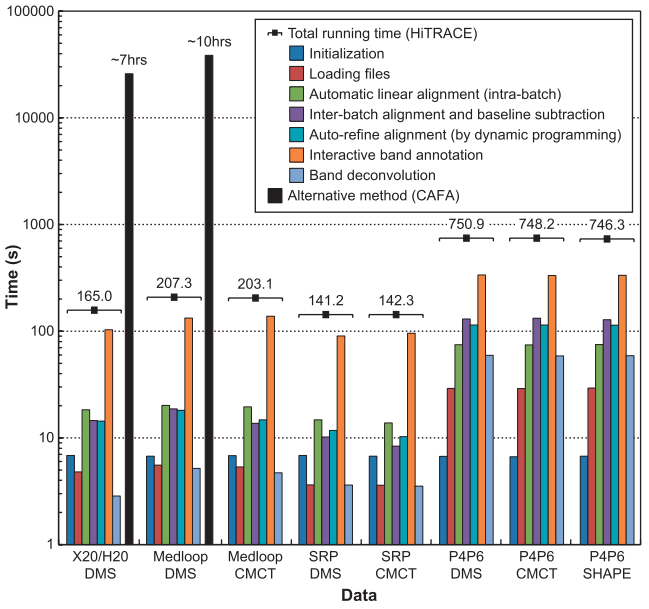

Fig. 5. The average running time of HiTRACE on different datasets. The time was measured on a personal computer system equipped with a 2.66 GHz Core i5 processor (4 cores; multithreading enabled) with 4 GB RAM.

quantify, and the smaller sets (88–136 profiles, 4000–8000 bands) required 3 min or less. Overall, HiTRACE averaged 1.58 s per profile from beginning (raw data load-in) to end (quantified band intensities). For the same datasets, the overall computational time of the tools for alignment only (i.e. msalign, SpecAlign and COW) were between 10 min to 2 h (without peak fitting) depending on the data size. As discussed above, CAFA and ShapeFinder, the previous full suites available for nucleic acid CE quantification, required even more time (hours for the smaller datasets, extrapolated to days or weeks for the larger sets). As shown in Figure 5, the HiTRACE time breakdown is similar for all datasets, except for the P4-P6 RNA datasets, in which later stages are lengthened by increasing the number of bands in each profile (200 residues in the P4-P6 RNA, compared to under 100 residues for the other RNAs). We further used HiTRACE on datasets with longer RNAs (up to 400 nt) and involving reverse transcription with primers complementary to the middle of a long RNA; the HiTRACE procedure was readily applied to these datasets (see Supplementary Material), and, encouragingly, the time demand remained linear with the number of bands.

4 DISCUSSION

HiTRACE employs a series of automated techniques to control the high level of variability in parameters of CE systems and to resolve a key alignment bottleneck of modern nucleic acid structure mapping experiments. Several algorithmic advances are responsible for HiTRACE's accuracy and speed, including dynamic programming strategies that have not been been previously considered in the field. Quantitative comparisons on large experimental datasets demonstrate the utility of a linear time-axis transformation used in globally aligning profiles as well as the importance of a non-linear alignment procedure for resolving further unavoidable variations in elution rates along a capillary. In addition, an interactive band annotation interface increases user convenience and provides accurate starting positions for the subsequent quantification step. These improvements have brought down the overall analysis time of datasets with tens of thousands of electrophoretic bands from days to minutes. The largest time savings of the method are on experiments in which the same RNA sequence is probed under a variety of solution conditions, chemical modifiers, kinetic timepoints or mutations [see, e.g. (Das *et al.*, 2010; Kladwang and Das, 2010; Kladwang *et al.*, 2011; Mitra *et al.*, 2008; Weeks, 2010; Wilkinson *et al.*, 2008)]. Now, the slow step in these and other experiments is interactive band annotation, which takes minutes (Fig. 5). As more automated band assignment methods are developed (R.D., unpublished data; P.Pang, M.Elazar, J.S.Glenn, personal communication), we plan to incorporate them into this interface.

Although we designed HiTRACE primarily for RNA chemical structure mapping, the principles and premises that underlie HiTRACE are general and can easily be modified for use in other types of experimental assays. To enhance the adoption of this tool, we have created a stand-alone version of HiTRACE with a graphical user interface. We are also making the source code freely available to encourage further innovation and incorporation of these algorithms into other laboratories' CE software suites. Beyond the datasets discussed herein, HiTRACE is continuously being used for other studies, totalling over 20 000 profiles (greater than 2 million bands) at the time of submission (W.K., R.D., unpublished data; see also http://rmdb.stanford.edu). Given its accuracy, robustness and efficiency, we expect that HiTRACE will become a valuable tool for nucleic acid experimentalists entering a high-throughput era of structural analysis.

Downloaded from bioinformatics.oxfordjournals.org at Korea University Library on June 27, 2011

ACKNOWLEDGEMENTS

The authors thank Dr Alain Laederach at Wadsworth Center for providing CAFA sample data and members of the Das lab for comments on the manuscript and extensive testing.

Funding: National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. 2011-0009963 and No. 2011-0000158 to S.Y. in part); Burroughs-Wellcome Foundation Career Award at the Scientific Interface (to R.D. for computational work in part).

Conflict of Interest: none declared.

REFERENCES

Bylund,D. *et al.* (2002) Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *J. Chromatogr. A*, **961**, 237–244.

Cormen,T. *et al.* (2009) *Introduction to Algorithms*, 3rd edn. The MIT press, Cambridge, MA, USA.

Cover,T. and Thomas,J. (2006) *Elements of Information Theory*, 2nd edn. John Wiley and Sons, Hoboken, NJ, USA.

Das,R. *et al.* (2005) SAFA: semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. *RNA*, **11**, 344–354.

Das,R. *et al.* (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods*, **7**, 291–294.

Deigan,K. *et al.* (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl Acad. Sci. USA*, **106**, 97.

Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.

Ewing,B. *et al.* (1998) Base-calling of automated sequencer traces Using Phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.

Kay,S. (1993) *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, Upper Saddle River, NJ, USA.

Kazmi,S. *et al.* (2006) Alignment of high resolution mass spectra: development of a heuristic approach for metabolomics. *Metabolomics*, **2**, 75–83.

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**, 7414–7416.

Kladwang,W. *et al.* (2011) A mutate-and-map strategy accurately infers the base pairs of an 35-nucleotide model RNA. *RNA*, **17**, 522–534.

Laederach,A. *et al.* (2008) Semiautomated and rapid quantification of nucleic acid footprinting and structure mapping experiments. *Nat. Protocols*, **3**, 1395–1401.

Levenberg,K. (1944) A method for the solution of certain nonlinear problems in least squares. *Quart. Appl. Math.*, **2**, 164–168.

Marquardt,D. (1963) An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.*, **11**, 431–441.

Merino,E. *et al.* (2005) Advances in RNA structure analysis by chemical probing. *J. Am. Chem. Soc.*, **127**, 4223–4231.

Mitra,S. *et al.* (2008) High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Res.*, **36**, e63.

Nielsen,N. *et al.* (1998) Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A*, **805**, 17–35.

Oppenheim,A.V. and Schafer,R.W. (2009). *Discrete-Time Signal Processing*, 3rd edn. Prentice Hall, Upper Saddle River, NJ, USA.

Peattie,D. and Gilbert,W. (1980) Chemical probes for higher-order structure in RNA. *Proc. Natl Acad. Sci. USA*, **77**, 4679.

Pravdova,V. *et al.* (2002) A comparison of two algorithms for warping of analytical signals. *Anal. Chim. Acta*, **456**, 77–92.

Robinson,M. *et al.* (2007) A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments. *BMC Bioinformatics*, **8**, 419.

Ruiz-Martinez,M. *et al.* (1993) DNA sequencing by capillary electrophoresis with replaceable linear polyacrylamide and laser-induced fluorescence detection. *Anal. Chem.*, **65**, 2851–2858.

Tijerina,P. *et al.* (2007) DMS footprinting of structured RNAs and RNA–protein complexes. *Nat. Protocols*, **2**, 2608–2623.

Tomasi,G. *et al.* (2004) Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemom.*, **18**, 231–241.

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**, 1979–1990.

Walczak,R. *et al.* (1996) A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *Curr. Opin. Struct. Biol.*, **2**, 367–379.

Watts,J.M. *et al.* (2009) Architecture and secondary structure of an entire hiv-1 RNA genome. *Nature*, **460**, 711–716.

Weeks,K. (2010) Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struct. Biol.*, **20**, 295–304.

Wilkinson,K.A. *et al.* (2008) High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol.*, **6**, e960883–e960899.

Wong,J. *et al.* (2005) SpecAlign–processing and alignment of mass spectra datasets. *Bioinformatics*, **21**, 2088–2090.

Woolley,A. and Mathies,R. (1995) Ultra-high-speed DNA sequencing using capillary electrophoresis chips. *Anal. Chem.*, **67**, 3676–3680.

Xi,Y. and Rocke,D. (2008) Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics*, **9**, 324.

Downloaded from bioinformatics.oxfordjournals.org at Korea University Library on June 27, 2011