

ShapeFinder: A software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis

SUZY M. VASA,^{1,6} NICOLAS GUEX,^{2,6} KEVIN A. WILKINSON,³ KEVIN M. WEEKS,³
and MORGAN C. GIDDINGS^{1,4,5}

¹Department of Biomedical Engineering, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA

²Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

³Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA

⁴Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA

⁵Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA

ABSTRACT

Analysis of the long-range architecture of RNA is a challenging experimental and computational problem. Local nucleotide flexibility, which directly reports underlying base pairing and tertiary interactions in an RNA, can be comprehensively assessed at single nucleotide resolution using high-throughput selective 2'-hydroxyl acylation analyzed by primer extension (hSHAPE). hSHAPE resolves structure-sensitive chemical modification information by high-resolution capillary electrophoresis and typically yields quantitative nucleotide flexibility information for 300–650 nucleotides (nt) per experiment. The electropherograms generated in hSHAPE experiments provide a wealth of structural information; however, significant algorithmic analysis steps are required to generate quantitative and interpretable data. We have developed a set of software tools called ShapeFinder to make possible rapid analysis of raw sequencer data from hSHAPE, and most other classes of nucleic acid reactivity experiments. The algorithms in ShapeFinder (1) convert measured fluorescence intensity to quantitative cDNA fragment amounts, (2) correct for signal decay over read lengths extending to 600 nts or more, (3) align reactivity data to the known RNA sequence, and (4) quantify per nucleotide reactivities using whole-channel Gaussian integration. The algorithms and user interface tools implemented in ShapeFinder create new opportunities for tackling ambitious problems involving high-throughput analysis of structure–function relationships in large RNAs.

Keywords: RNA SHAPE chemistry; capillary electrophoresis; high-throughput RNA structure analysis; chemical modification; hydroxyl radical

INTRODUCTION

A critical requirement for a full understanding of the function of any RNA is an accurate picture of its higher order structure. Analysis of in-solution nucleic acid structure information often requires that RNA or DNA fragment lengths be analyzed at single nucleotide resolution. Important examples in this class include chemical modification and cleavage, and modification–interference experi-

ments (Ehresmann et al. 1987; Stern et al. 1988; Strobel 1999; Brenowitz et al. 2002; Tullius and Greenbaum 2005; Wilkinson et al. 2006). These experiments can be used to analyze local nucleotide conformational differences, solvent accessibility, functional group modifications, and interactions with protein and small molecule ligands for RNA and DNA structure. For over three decades, these classes of experiments have been evaluated by resolving nucleic acid fragments on polyacrylamide sequencing gels (Maxam and Gilbert 1980). Gel electrophoresis has the advantages of good nucleotide resolution and low material costs. However, gel electrophoresis is time consuming, typically limited to reads of 80–100 nucleotides (nt) per gel at single nucleotide resolution, and prone to band overlap and compression artifacts.

In contrast, the capillary electrophoresis instruments used for DNA sequencing routinely yield read lengths of 300–1000 positions at single nucleotide resolution with few

⁶These authors contributed equally to this work.

Reprint requests to: Kevin M. Weeks, Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA; e-mail: weeks@unc.edu; fax: (919) 962-2388; or Morgan C. Giddings, Department of Biomedical Engineering, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA; e-mail: giddings@unc.edu; fax: (919) 962-2388.

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.1166808>.

artifacts. The absence of appropriate software algorithms that address the unique quantitative properties of raw electropherograms generated by structure-probing experiments has prevented the use of capillary electrophoresis for nucleotide-resolution analysis of nucleic acid folding, dynamics, and ligand binding.

To address this problem, we have created a new software suite called ShapeFinder that automates the steps required to extract quantitative, single nucleotide resolution reactivity information for 300–650 nt in a single capillary electrophoresis run. We focus here on the analysis of selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) experiments (Merino et al. 2005; Wilkinson et al. 2005, 2006, 2008). However, the algorithms in ShapeFinder also work well for analyzing capillary electrophoresis data from other classes of nucleic acid reactivity experiments, including those that use other chemical modification agents or hydroxyl radicals to map structure and solvent accessibility (data not shown).

RNA structure and hSHAPE chemistry

SHAPE chemistry holds considerable promise for determining the nucleotide-resolution structure of any RNA under diverse, relevant states (Badorrek and Weeks 2005, 2006; Badorrek et al. 2006; Chen et al. 2006; Gherghe and Weeks 2006; Dann et al. 2007; Vicens et al. 2007; Duncan and Weeks 2008; Jones et al. 2008; Stoddard et al. 2008; Wang et al. 2008). SHAPE chemistry measures local backbone flexibility at nearly every position in an RNA by forming sparse 2'-O-adducts using a hydroxyl-selective electrophile (Fig. 1A; Merino et al. 2005). Nucleotides constrained by base-pairing or tertiary interactions are unreactive, while conformationally flexible (and, likely, single-stranded) nucleotides preferentially form 2'-O-adducts (Fig. 1A,B). Sites of modification are located by annealing a 5'-end-labeled primer to the RNA and then extending the primer to the nearest site of modification using reverse transcriptase in an optimized primer extension reaction (Merino et al. 2005; Wilkinson et al. 2006). The products of this experiment are 5'-end-labeled cDNA fragments whose length and amount correspond to the position and degree of modification—and hence local nucleotide flexibility—at every nucleotide in an RNA (Fig. 1C). To assess RNA degradation and position-dependent processivity of the primer extension reaction, a control that omits the reagent is performed in parallel. In addition, one or two dideoxy sequencing reactions are used to map reactivity to the RNA sequence (Fig. 1D).

In high-throughput SHAPE (hSHAPE), each component of a SHAPE experiment uses the same primer sequence labeled with a color-coded fluorophore. The resulting cDNAs are combined and resolved in a single capillary on a capillary electrophoresis sequencing instrument (Mortimer and Weeks 2007; Duncan and Weeks 2008; Wilkinson et al.

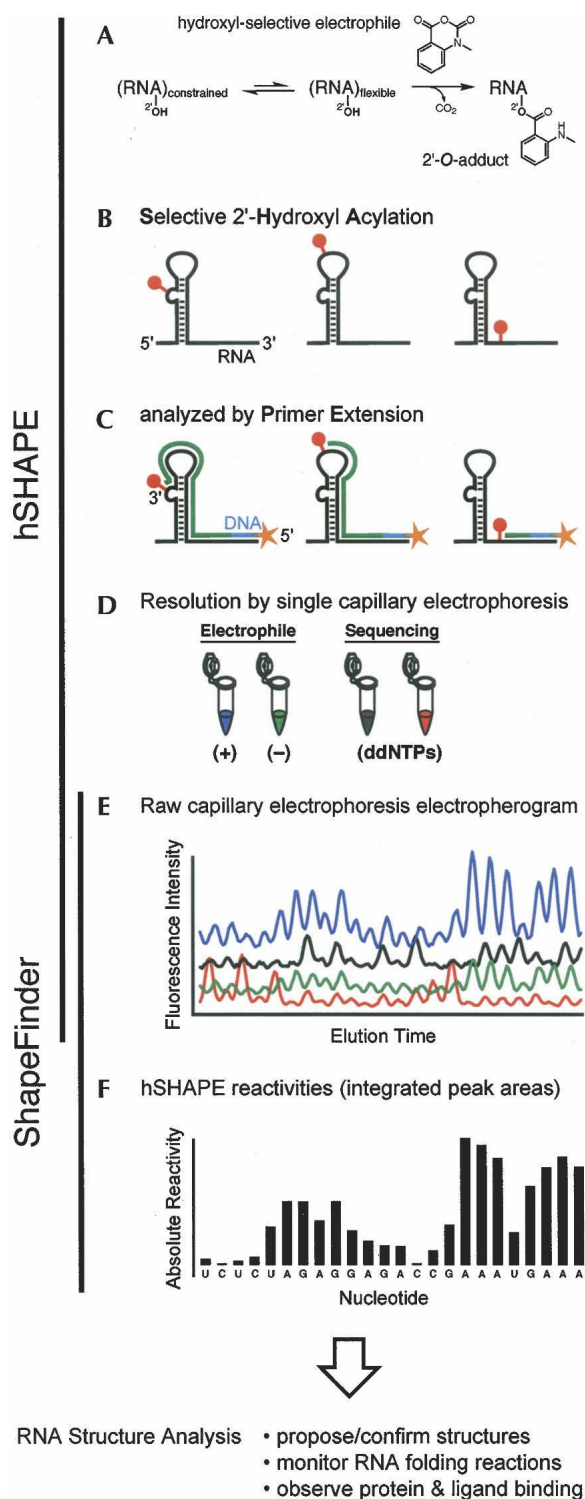


FIGURE 1. Overview of high-throughput selective 2'-hydroxyl acylation analyzed by primer extension (hSHAPE) and data processing using ShapeFinder.

2008). Raw capillary electrophoresis profiles contain 300–650 nt of SHAPE reactivity information and thus provide a comprehensive, nucleotide-resolution view of RNA secondary and tertiary structure in a single experiment.

Algorithmic challenges for nucleic acid structure analysis resolved by capillary electrophoresis

The output of an hSHAPE experiment resolved by capillary electrophoresis is an electropherogram, or trace. A trace contains three to four individual channels that report fluorescence intensity versus elution time information, where each channel corresponds roughly to one of the SHAPE reactions (Fig. 1E). The results of hSHAPE and DNA sequencing experiments resemble each other in that both experiments generate fluorescence intensity versus elution time data and must be processed extensively in order to yield useful nucleotide resolution information. However, extracting reactivity versus nucleotide position information for an hSHAPE experiment, or any other nucleic acid reactivity experiment, requires unique algorithms and data processing strategies.

The first and most important difference is that peak magnitude in DNA sequencing contains little meaning other than to indicate which nucleotide is present. In contrast, *both* peak intensity *and* position are meaningful for all peaks in the (+) and (–) reagent channels in an hSHAPE experiment. Peak intensity spans a dynamic range of 50-fold and reports the structure-sensitive yield of the 2'-O-adduct, and thus local nucleotide flexibility (Fig. 1A). Critically, the processing steps applied to hSHAPE data must not disturb relative intensity or distribution of peak features in the electropherogram.

Second, two to three meaningful peaks in distinct channels [in the (+) and (–) reagent channels, and potentially in one sequencing channel] are observed per nucleotide in an hSHAPE electropherogram, versus one important peak per position in a sequencing experiment. Thus, hSHAPE traces lack alignment cues exploited in DNA sequencing electropherograms, and corresponding peaks must be aligned to each other with greater precision than is required for sequencing.

Third, peaks must be located, aligned, and quantified for every position in the (+) and (–) reagent channel, whereas sequencing only requires locating the most intense peak per position in the four reaction channels. The absence of a peak in the hSHAPE (+) reagent channel conveys significant information because it indicates that a nucleotide is constrained by base pairing or tertiary interactions, so accurate identification and quantitative analysis of positions where peaks are minimal or absent is an essential requirement for hSHAPE. Finally, fully automated analysis of hSHAPE data requires that sparse sequencing trace channels be aligned with a known input sequence, in contrast to DNA sequencing where the goal is to obtain the nucleotide sequence.

RESULTS AND DISCUSSION

ShapeFinder

The initial processing steps required to convert raw capillary electrophoresis profiles into useful reactivity information are similar to those involved in the analysis of DNA sequencing traces. We therefore extended the BaseFinder platform, a framework originally designed for DNA trace processing, analysis, and base-calling (Giddings et al. 1998), to analyze nucleic reactivity information. ShapeFinder is a modular, extensible software package in which each signal-processing algorithm is implemented as a tool. The results of each analysis step are immediately displayed to the user in a straightforward graphical user interface (Fig. 2).

ShapeFinder reads and displays files from most common sequencing platforms, including generic tab-delimited .txt files, the Beckman .esd and .dat files, and the ABI .fsa, .abi, and .ab1 formats. ShapeFinder also uses a new file format (.shape) to store the raw and processed hSHAPE data along with the tool parameters that have been applied to the data set. The .shape file allows for review and re-execution of trace processing steps and facilitates testing different parameter choices.

The net output of ShapeFinder is a table of quantitative reactivity information as a function of position in the nucleotide sequence (Fig. 1F). For hSHAPE experiments, reactivity information has thus far been used to develop models for RNA secondary structures, to monitor RNA folding reactions, and to evaluate protein binding and macromolecular complex formation (Mortimer and Weeks 2007; Duncan and Weeks 2008; Wilkinson et al. 2008).

ShapeFinder tools

ShapeFinder executes its algorithms via a sequence of tools, called a script. Each tool accomplishes a specific data processing step by applying user-definable parameters to the electropherogram. The script is displayed in the Scripting Inspector window in the ShapeFinder user interface (Fig. 2, lower right). Tools are added to a script using the Tool Inspector window, which also displays the parameter values associated with each tool (Fig. 2, upper right). A processing tool is added using the “Append” button; tools already in a script may be changed and rerun by selecting “Replace.” An individual step and its associated parameters are reviewed by selecting the tool entry in the Scripting Inspector window.

Analysis of an hSHAPE raw capillary electrophoresis profile involves three major processing steps. First, the raw electropherogram is subjected to preprocessing to correct for fluorescent background, spectral overlap between the fluorescent channels, mobility shifts imparted by tagging the primers used in the primer extension steps with different dyes, and signal decay at long read lengths. Second, channels are aligned so that all peaks in the (+) and (–)

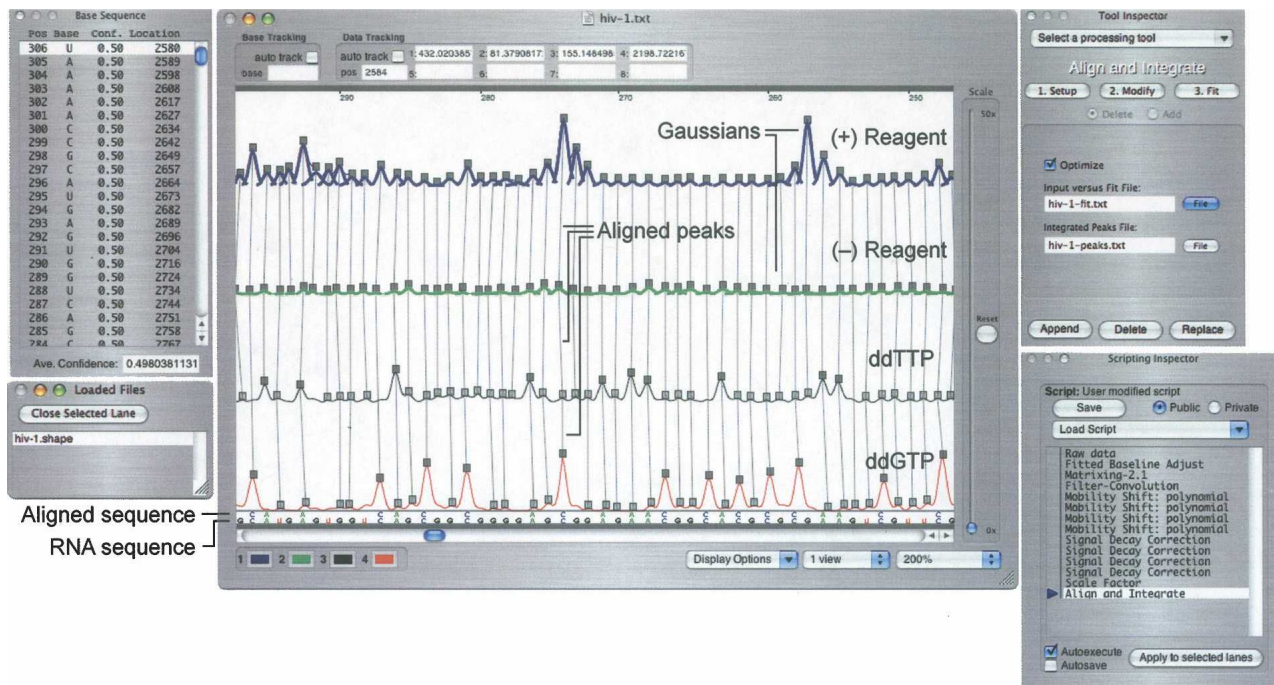


FIGURE 2. ShapeFinder at the Align and Integrate stage. The Data View Window (*center*) provides graphical feedback on each data processing step. The Tool Inspector window (*upper right*) displays the user-definable parameters for the tool selected in the Scripting Inspector. The Scripting Inspector (*lower right*) displays the tools thus far applied to the data; several tools have been applied multiple times using iteratively refined parameters.

reagent channels are identified and linked to the input RNA sequence, including peak positions corresponding to zero reactivity. Finally, quantitative nucleotide reactivities are obtained by performing a whole-channel Gaussian integration for all peaks in the (+) and (-) reagent channels. Subtracting the integrated values for the (-) reagent from the (+) reagent profiles yields the absolute nucleotide-resolution reactivity for every RNA position over read lengths typically spanning 300–650 nt. An experienced individual can perform the data processing steps in 1–2 h.

We will illustrate these processing steps using an experiment performed on a transcript corresponding to the first 976 nt for the NL4-3 strain of the HIV-1 genome. We use a scheme in which blue and green channels represent the (+) reagent and (-) reagent experiments, and black and red represent RNA sequencing ladders (reflecting chain termination by ddGTP or ddTTP, respectively) (Fig. 3). Data are collected from the capillary electrophoresis instrument such that the small fragments representing the 3'-end of the RNA read elute first.

Data preprocessing

Fitted Baseline Adjust, Matrixing, and Smoothing tools

Channels in raw capillary electropherograms are convoluted by detector background, overlapping emission spectra, detector noise, and horizontal offset between channels (Figs. 1E, 3A). Since these traits are common to all electropherogram

data, initial processing of the raw electropherograms is analogous to DNA sequencing.

Detector background imparts an idiosyncratic vertical offset to each channel. The Fitted Baseline Adjust tool adjusts each channel to a common baseline by zeroing each channel over a window of detector readings, typically, 10 times the average peak width.

The fluorescent dyes used to distinguish the channels in a capillary electrophoresis electropherogram excite at similar wavelengths and have overlapping emission spectra. Thus, some dye signals are detected in several fluorescent channels by the instrument detector. For example, in the sample data set, the (+) reagent peaks are detected in both the blue and green channels (Fig. 3A). The Matrixing tool determines the unique quantitative contribution of each fluorophore to signal intensity in each channel (Fig. 3B). Parameters for the Matrixing tool must be calibrated once for each set of dyes (described in the Materials and Methods section). Some commercial instruments implement these steps using instrument-specific software, and these alternative algorithms can be used in place of those in ShapeFinder, provided they correct completely for spectral overlap and do not leave significant residuals in other channels.

Trace data from a DNA sequencer contain fluctuations due to detector noise so that each major peak may have minor peaks and valleys of its own, which complicate downstream peak finding. Smoothing can increase read length and peak detection by ~10% for datasets with low

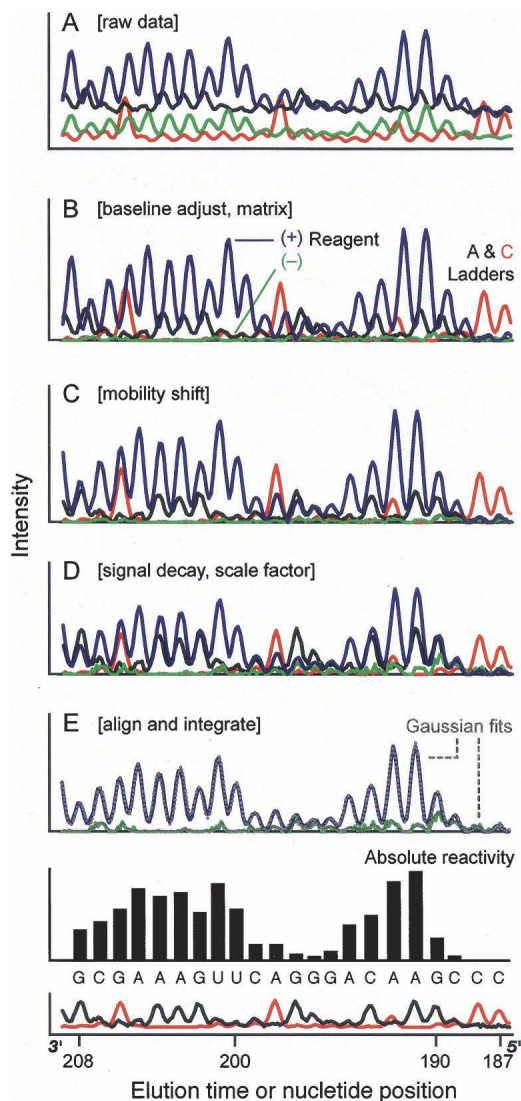


FIGURE 3. Electropherogram analysis as implemented using ShapeFinder tools. (A) Unprocessed capillary electrophoresis electropherogram. (B) Net result after application of the Fitted Baseline Adjust and Matrixing tools. (C) After the Mobility Shift: Polynomial tool (from four serial executions of the tool). (D) After signal decay correction and rescaling. (E) Whole-channel Gaussian integration of the (+) and (-) reagent channels obtained using the Align and Integrate tool. (Solid bars) Absolute SHAPE reactivities after subtracting background. For clarity, sequencing channels are offset from the (+) and (-) reagent channels.

signal to noise ratios. The results of the Baseline Correction, Matrixing, and Smoothing tools on the HIV-1 NL4-3 transcript are shown in Figure 3B.

Mobility shift

In an hSHAPE experiment, each reaction is analyzed using a DNA primer labeled with a different fluorophore (Fig. 1D). The dyes alter electrophoretic migration rates so that cDNAs of the same length have slightly different elution times (Fig. 3B). Because a large peak is not expected at

every position in hSHAPE data, mobility shift correction must be performed more accurately than is required for DNA sequencing to facilitate accurate linking of corresponding peaks between channels. ShapeFinder implements several mobility shift tools that can be combined in serial to account for horizontal offset without significantly altering peak shapes. Two to four serial applications of the Mobility Shift: Cubic tool typically places all channels on a consistent x -axis (Fig. 3C). Parameters for an initial mobility shift must be set once for each set of primers. These parameters can also be fine-tuned on a trace-by-trace basis.

Signal decay correction

Inspection of all of the channels in an hSHAPE experiment indicates that peak intensities decay with increasing read length (best visualized in Fig. 6A, see below). There are three sources of this decay, depending on the reaction channel. (1) Reverse transcriptase is not perfectly processive and fails to elongate at every position with an unmodified 2'-hydroxyl, such that the probability of adding an additional nucleotide to a cDNA is slightly less than 1. (2) The (+) reagent reaction is designed such that, on average, one in every 300 nt is modified. However, some RNAs react two or more times, since modification at one site does not preclude modification at other sites. For RNAs containing multiple adducts, only the first modification is detected, thus favoring short cDNAs. (3) In the sequencing reactions, the population of extending primers decreases by a small factor each time a dideoxynucleotide is incorporated. Thus, signals decay exponentially to zero in all channels at read lengths of 400–650 nt.

Signal decay correction must be performed using a physically meaningful statistical model. We find that signal decay is well modeled as:

$$D(x) = Aq^x + C, \quad (1)$$

where D is the signal intensity as a function of primer elongation, A is the amplitude of the decay, C represents the measured intensity at the end of the channel, and q is the probability of extension at position x (Badorrek and Weeks 2006).

The user sets (1) the range of data points; (2) the channel in which to apply the tool; and (3) a scaling factor to maintain the corrected scale relative to the other channels. The algorithm calculates new values and a properly corrected channel is readily verified by visually inspecting the data: intense peaks in the beginning, middle, and end of the channel should be of uniform height.

Scaling

Experimental variations in performing primer extension, inherent differences in dye intensity, and second-order effects of the ShapeFinder Smoothing and Signal Decay Correction tools can cause the channels to be on different

relative scales. The channels are adjusted manually such that the smallest 5%–10% of the peaks in the (+) channel match corresponding (–) channel intensity (Fig. 3D, cf. blue and green channels). This correction assumes that there are always a few completely unreactive nucleotides in an hSHAPE read whose peak intensities should exactly match the corresponding (–) peak intensities. For ease in data viewing and further analysis, sequencing peaks are set to match moderately intense peaks in the (+) channel (Fig. 3D, cf. red, black channels and blue channels).

After preprocessing, all channels have a baseline set to zero, peaks in different channels corresponding to the same nucleotide have the same elution time, and well-defined peak intensities correspond quantitatively to cDNA amounts (Fig. 3D).

Whole-channel peak alignment and integration

The heart of the new ShapeFinder program is the Align and Integrate tool, which calculates hSHAPE reactivities for every analyzable nucleotide in an electropherogram (Fig. 3E). There are four phases to the algorithm: (1) peak finding and linking; (2) alignment to the RNA sequence; (3) user editing of the alignment by adding or deleting peaks; and (4) quantification of 2'-*O*-adduct formation by Gaussian curve fitting (Fig. 4). The ShapeFinder Align and Integrate tool implements these steps in the Setup, Modify, and Fit panels (Fig. 2, upper right). Phases 1 and 2 are performed using the Setup panel; the Modify panel allows the user to add and remove peaks in phase 3; and the Fit panel is used to manage phase 4. The tool is iterative and alignments are recalculated after each round of user input.

Primer extension stops at the base preceding the nucleotide containing a 2'-*O*-adduct; thus, the (+) and (–) reagent peaks are one nucleotide shorter than the cDNA fragments generated by dideoxy sequencing (Merino et al. 2005). The sequencing alignment is therefore shifted by one nucleotide relative to the (+) and (–) reagent channels. To avoid confusion, the ShapeFinder display shows the sequencing peaks without an offset, but the resulting output files show the offset.

Setup

In the Setup phase, the user assigns each channel to one of the four SHAPE reactions [(+) and (–) reagent, and sequencing ladders] and specifies the region of the trace to be analyzed using either numerical trace positions or by selecting a region of the trace in the main window. A Refine option enables automatic interpolation of peaks in the (+) or (–) reagent channels based on expected peak spacing in a given region of the trace. The Setup phase also reads an ascii text file containing the RNA sequence that is used to align the trace data to the RNA nucleotide position.

A preliminary alignment is initiated after these parameters have been set. The data view window displays the four channels and demarcates identified peaks with squares (Fig. 2). Vertical lines show peaks inferred to correspond to the

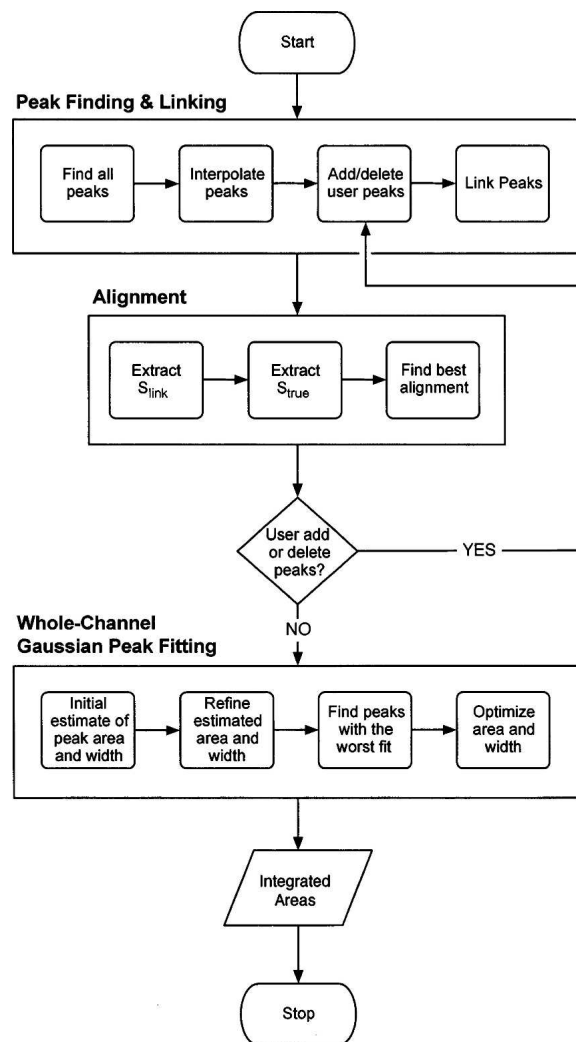


FIGURE 4. Flowchart of the Align and Integrate algorithm, which involves three phases: (1) peak finding and linking of (+) reagent, (–) reagent, and sequencing peaks; (2) peak alignment to the RNA sequence; and (3) calculation of per nucleotide SHAPE reactivities by Gaussian integration.

same RNA position and linked with the input sequence. Light-shaded squares in either the (+) or (–) reagent channels indicate unlinked peaks. For the sequencing channels, light-shaded squares report peaks that were identified but were not accepted as part of the sequencing ladder.

Modify

In portions of a run with strong signal, low noise, and good alignment, the results of the first alignment step are typically satisfactory. However, in some regions, especially near the ends of a read, meaningful peaks may be missed, not aligned, or assigned incorrectly. These regions often contain high-quality and quantitative structural information that can be gleaned with operator supervision. To this end, ShapeFinder allows manual editing and extension of

the automatically generated alignment using the Modify panel (illustrated schematically in Fig. 4).

The ShapeFinder-determined sequence alignment is displayed at the bottom of the data window (Fig. 2). The top sequence is determined from the sequencing channels, while the bottom sequence shows the input sequence. For completely aligned data, letters coincide vertically between the two datasets. When the data are partially misaligned relative to the input sequence, there will be a horizontal offset between the two sequences. The addition or deletion of a peak in the (+) reagent channel usually corrects the alignment.

Peaks are deleted by clicking on the square at the top of each peak; peaks can be added by clicking at the desired position for a new peak in any channel. The Modify panel displays a list of added or deleted peak positions (Fig. 2, left panel). Data ready for Gaussian fitting (Fig. 2, center panel) should be correctly aligned to the sequence and have all (+) and (-) peaks linked to each other and to the sequence, as indicated by filled boxes. Depending on the quality of the data, zero to ~20 peaks may need to be deleted or added at either end of a trace. Particularly difficult or unalignable regions may be removed in the Setup panel by adjusting the Trace Range. For the HIV-1 example data set, three unaligned (+) reagent peaks were deleted at the beginning of the trace. This correction then allowed complete alignment of >400 continuous nucleotides in the RNA.

Fit

Once all peaks have been identified and linked with the input sequence, the Fit phase of the Align and Integrate tool performs whole-channel Gaussian peak integration for the (+) and (-) reagent channels (Fig. 3E). Each peak is fit to

$$y_i(x) = \frac{A_i}{\sqrt{2\pi}\sigma_i} e^{-1/2(x-\mu_i/\sigma_i)^2}, \quad (2)$$

where A_i is the peak area and μ_i and σ_i are the center and width of peak i , respectively. The tool has both fast and optimize modes. The Optimize option provides a more accurate peak fitting, but is more computationally demanding (Fig. 5).

Once fitting is complete, ShapeFinder displays the calculated peaks superimposed upon the (+) and (-) reagent channels in the data view window (Fig. 2, bold lines). The fitted Gaussian curves for all peaks, the calculated peak areas, the net absolute reactivity at every position, the alignment to the input sequence, and identified peak positions are output in text files.

Example of a complete hSHAPE experiment, quantified by ShapeFinder

A complete hSHAPE electropherogram contains structural information for several hundred RNA nucleotides (Fig. 6A). As outlined above, the raw data for the HIV-1 example

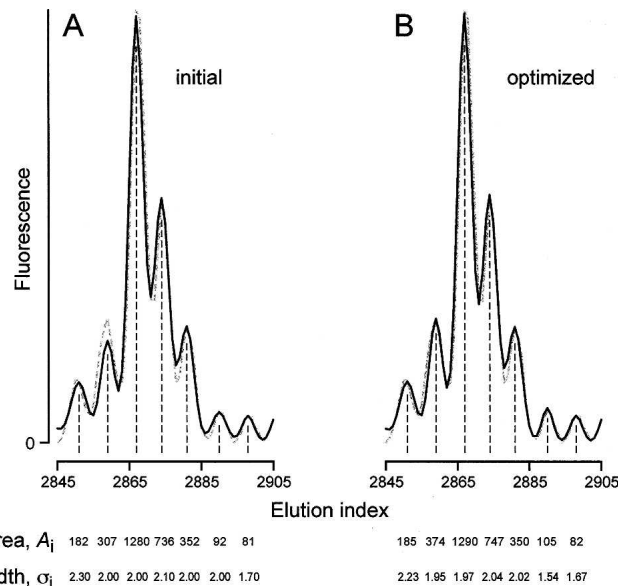


FIGURE 5. Whole-channel peak integration. (A) Preliminary local fit (solid line) to initial values for A_i and σ_i , compared with the electrophoresis data (dashed line). (B) Final optimized fit.

RNA includes all of the typical characteristics of raw electropherogram sequencing data, including baseline offset, fluorescent overlap, channels on different intensity scales (black and red channels, Fig. 6A), mobility offset for cDNAs of the same length, and signal decay such that peaks at the left of the channel are four times more intense as those at the right (blue and red channels, Fig. 6A). After applying the preprocessing tools, all channels have a baseline set to zero, peak intensities correspond quantitatively to cDNA amounts, and overall peak heights are distributed evenly throughout each channel (Fig. 6B).

The Align and Integrate algorithm then (1) aligns the sequencing ladders and the (+) and (-) reagent channels with the input sequence; and (2) calculates the areas under all analyzable peaks in the (+) and (-) channels by whole-trace Gaussian integration. Automatic subtraction of the (-) from the (+) reagent peak areas yields the absolute hSHAPE reactivity for every nucleotide in the capillary electrophoresis electropherogram (Fig. 6C, bars). In this typical experiment, single nucleotide resolution SHAPE reactivities were obtained for positions 491–905 in the HIV-1 transcript, for a total read length of 415 nt.

Analysis of accuracy and the reproducibility of hSHAPE and ShapeFinder

The most important criteria by which to judge ShapeFinder is whether its algorithms yield reproducible and accurate RNA structure information. After performing a SHAPE experiment on the well-studied tRNA^{Asp} molecule, cDNA fragments were resolved either (1) using radiolabeled DNA primers and detection by denaturing gel electrophoresis or

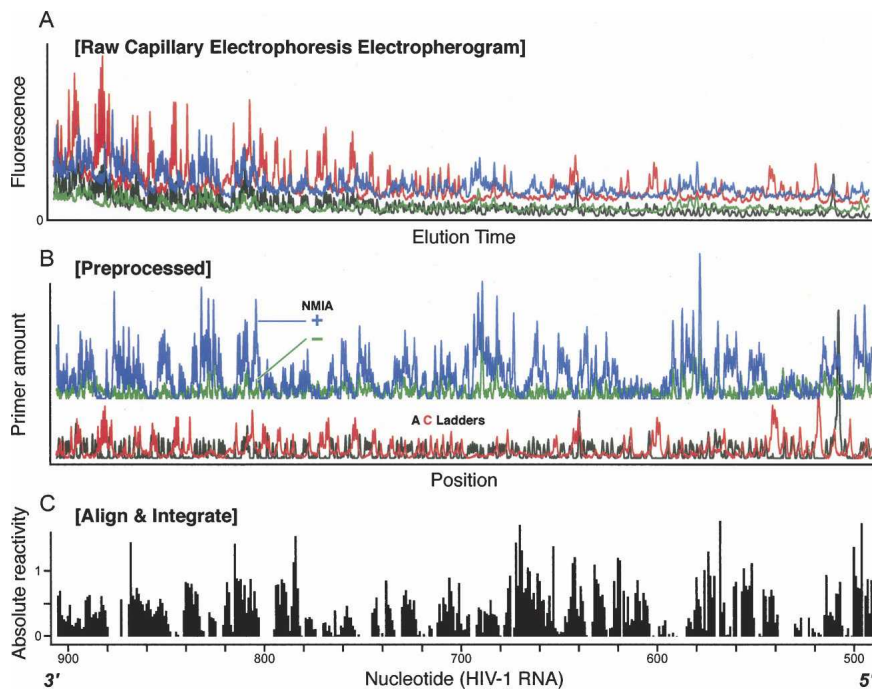


FIGURE 6. Overview of a complete hSHAPE experiment, processed using ShapeFinder and representing a total read length of 415 nt from an HIV-1 transcript RNA. (A) Raw electropherogram from a DNA sequencer. The data consist of four channels of fluorescence intensity information as a function of elution time. (B) Preprocessed SHAPE data. Each channel now represents dye amount, not fluorescence, as a function of elution time for each of the four channels. For clarity, A and C sequencing ladders are offset from the (+) and (–) reagent channels. (C) Sequence alignment and whole-channel Gaussian peak integration using the Align and Integrate tool to calculate absolute SHAPE reactivities.

(2) by capillary electrophoresis and ShapeFinder. The tRNA molecule was imbedded in a previously described structure cassette to facilitate analysis by primer extension (Merino et al. 2005). This RNA is 132-nt long, which is close to the size limit that yields a single-nucleotide resolution banding pattern in routine sequencing gels. cDNAs resolved by gel electrophoresis were quantified using SAFA, which has been independently validated to calculate accurate band intensities (Das et al. 2005). We compare these experimental results to a SHAPE experiment performed under the same conditions on the same RNA, but analyzed using fluorescent primers, capillary electrophoresis, and ShapeFinder. For both datasets, SHAPE reactivity data are normalized to a scale that spans 0 to ~ 1.5 and in which 1.0 is defined as the average intensity of highly reactive positions.

Quantitative analysis of cDNA fragments obtained from a SHAPE analysis of the tRNA^{Asp} transcript yielded nearly identical reactivities at almost all positions, regardless of the separation and analysis platform (Fig. 7A, cf. solid and open columns). The linear correlation coefficient, R , between the two datasets is 0.91, indicating 83% (R^2) of the variability in the hSHAPE data is predicted by the variability of the gel data. This correlation is significant at the $P < 0.0001$ level. Comparison of reactivity differences between

the two datasets yielded a Student t-test P -value of 0.84, again indicating the group reactivities are statistically equivalent.

The only significant differences in measured nucleotide reactivity occurred at positions 29–32. The differences reflect the difficulty in calculating intensities in the context of band compression that occurs when cDNA fragments for this RNA are resolved by gel electrophoresis (Fig. 7A, labeled; Chamberlin and Weeks 2000; Wilkinson et al. 2005); these positions were therefore not included in the correlation analysis. In contrast, positions 29–32 were readily interpretable in the capillary electrophoresis trace. Thus, ShapeFinder yields quantitative values for per nucleotide reactivities that are as accurate as the conventional approach using gel electrophoresis. The primary difference is that capillary electrophoresis is less sensitive to band compression artifacts.

Second, we analyzed the reproducibility of SHAPE reactivities for five independent datasets from the HIV-1 RNA, three corresponding to a primer binding at position 342 and two for a primer binding at position 535 (Fig. 7B). These primers bind 193 nt apart and yield overlapping reads of ~ 200 nt.

The region of overlap corresponds to the 3' portion of one primer read, and the 5'-most end of the second primer read (dashed arrows, Fig. 7B). The overlapping regions therefore also correspond to sets of peaks that have been differentially adjusted by the signal decay correction algorithm.

Correlation coefficients calculated between the 10 possible pairs of the five datasets indicated a very strong correlation between the datasets, with R^2 values of 0.86–0.97 (all P -values < 0.0001). A one-way analysis of variation (ANOVA) performed between the five datasets showed the SHAPE reactivities to be statistically equivalent ($P = 0.77$). Furthermore, Levene's Test indicated constant variance between the five datasets (P -value = 0.26). Finally, we calculated the standard deviation for each measurement in the 181–230 window. A plot of the per position standard deviation as a function of mean SHAPE reactivity is linear ($R = 0.73$; $P < 0.0001$). Linear regression indicates that the average measurement error at any one nucleotide is $0.04 + 0.11 \times$ (per position measurement) in SHAPE units. Thus, for representative low and high SHAPE reactivities of 0.1 and 0.7, measurement errors are expected to be ± 0.05 and ± 0.12 SHAPE units, respectively.

In sum, these statistical tests indicate that SHAPE reactivities as quantified using ShapeFinder (1) are calculated

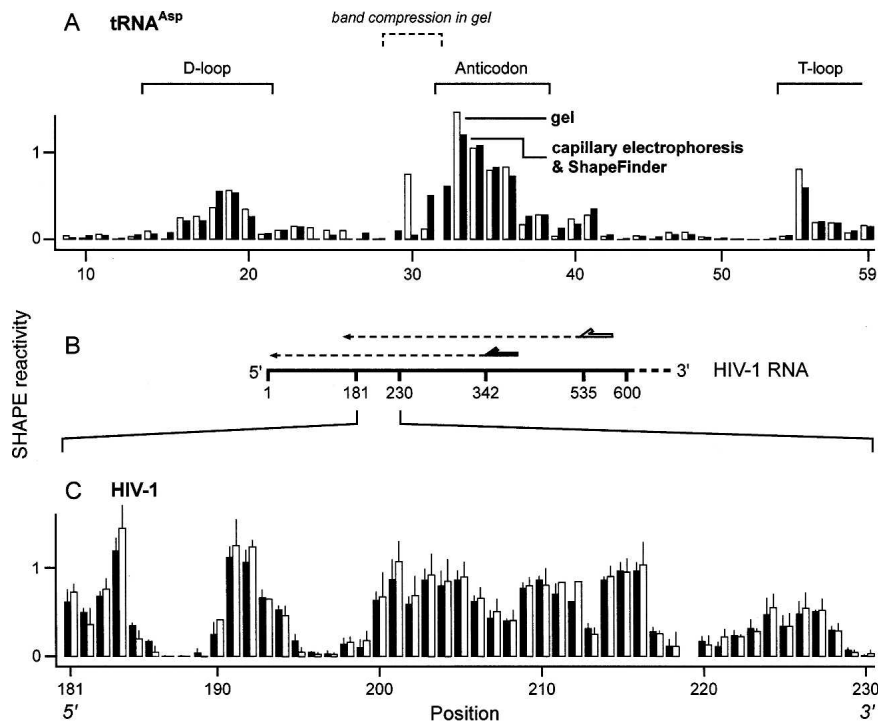


FIGURE 7. Accuracy and reproducibility of hSHAPE and ShapeFinder. (A) Comparison of nucleotide reactivity as quantified by ShapeFinder (solid bars) and denaturing gel electrophoresis (open bars). Loops in tRNA^{Asp} are indicated explicitly. Bands visualized by gel electrophoresis were quantified using SAFA (Das et al. 2005). (B) Overlapping reads for HIV-1 genome transcripts. Primers (solid and open arrows) anneal to the RNA 193 nt apart; reads (dashed lines) therefore overlap by ~200 nt. (C) Mean hSHAPE reactivities and standard deviations calculated from overlapping and replicate reads. Primers annealed at positions 342–363 (solid columns) and 535–555 (open columns). Data shown report three experiments from the 342 primer and two experiments from the 535 primer; whiskers report standard deviations. Due to high background, no data were available at nucleotide 219.

accurately over hundreds of nucleotides; (2) are accurately corrected for signal decay as modeled by Equation 1; and (3) exhibit small absolute measurement errors. Combining quantitative reactivities from individual reads of 300–650 nt can therefore robustly monitor the structures of long RNAs, potentially spanning thousands of nucleotides.

PERSPECTIVE

Experiments that probe nucleotide reactivity and solvent accessibility represent powerful approaches for analyzing conformational changes and protein and ligand binding for RNAs of known structure, and for developing models for RNAs whose structures are not known. A critical limiting step in such analyses has been the use of gel electrophoresis technology to visualize the results of these experiments. In many cases, more effort is required to obtain, manipulate, and quantify information by gel electrophoresis than is spent actually performing an experiment or interpreting its result.

The algorithms implemented in ShapeFinder dramatically lower the barriers to monitoring the structure of large RNAs. Depending on the characteristics of an RNA, we

routinely obtain read lengths of ~400 nt, with reads of up to 650 nt in favorable cases. This means that single nucleotide resolution structure information can now be obtained for entire large catalytic and regulatory RNAs or domains of larger RNAs like ribosomal RNAs in a single experiment.

While ShapeFinder accelerates the ability to interrogate RNA structure in solution at single nucleotide resolution, we are continuing to develop new methods and algorithms to further improve the speed, accuracy, and automation of hSHAPE analysis. Important objectives include improved and automatic mobility shift alignment and algorithmic optimizations to reduce the computational overhead for peak curve fitting. It is our hope that ShapeFinder will make it possible to tackle new classes of problems related to the role of long-range RNA structure in biological function.

MATERIALS AND METHODS

SHAPE data

SHAPE experiments were performed on HIV-1 or tRNA^{Asp} transcripts exactly as described (Merino et al. 2005; Wilkinson et al. 2008). Most of the SHAPE reactivity data on HIV-1 sequences presented here were reported previously (Wilkinson et al. 2008). Briefly, DNA templates encoding the 5'-most 976 nt of the HIV-1 NL4-3 strain (Gen Bank AF324493) or tRNA^{Asp} were generated by PCR. The RNA construct was produced by *in vitro* transcription and purified by gel electrophoresis. The HIV-1 RNA and tRNA^{Asp} (1 pmol) were refolded in 50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), and 5 mM MgCl₂ or 100 mM HEPES (pH 8.0), 100 mM NaCl, 10 mM MgCl₂, respectively, for 30 min at 37°C. (+) and (-) reagent SHAPE reactions were initiated by treating the RNA with N-methylisatoic anhydride (NMIA, 32.5 mM in DMSO) or DMSO, respectively. After the NMIA hydrolyzed completely (60 min) (Merino et al. 2005), the RNA was recovered by ethanol precipitation and mixed with fluorescently labeled DNA primers (Proligo or LI-COR) that annealed either at positions 342–363, 535–555, or 956–976. Primer extension was initiated by addition of Superscript III reverse transcriptase (Invitrogen). Sequencing markers were generated using unmodified RNA by performing primer extension in the presence of dideoxy NTPs. Dyes for the (+) and (-) reagent and sequencing lanes were Cy5, WellRed D3, WellRed D2, and LI-COR IR 800, respectively. cDNA products from the four reactions were mixed, purified, and separated on a Beckman CEQ2000XL capillary electrophoresis DNA sequencer. tRNA^{Asp} experiments were performed both using a 5'-[³²P]

labeled primer and resolving primer extension fragments on electrophoresis gels (Merino et al. 2005; Wilkinson et al. 2005) or using fluorescently labeled (Prologo or LI-COR) primers and capillary electrophoresis, in the same manner as the HIV-1 RNA, except that 130 mM NMIA in DMSO was used. Fluorescence intensity over the four channels was monitored at a rate of 2 Hz and yielded an average of ~ 10 points per peak position. Raw electropherograms were output from the capillary electrophoresis instrument in the Beckman .txt format and read directly into ShapeFinder.

ShapeFinder software

ShapeFinder is an extension of the BaseFinder trace-processing platform and is written in Objective-C (Giddings et al. 1998). It is distributed as a Universal Binary, and runs on Macintosh computers running Mac OS X 10.4 or later. ShapeFinder is freely available for noncommercial use. Both the ShapeFinder software, which includes a detailed help package for new users of hSHAPE, and all HIV-1 data and example scripts used in this work are available at: <http://bioinfo.unc.edu/downloads/>.

Fitted Baseline Adjust

The Fitted Baseline Adjust tool calculates a common baseline for each channel while keeping the experimentally recorded data intact (Giddings et al. 1998). The local minima in a channel are found after dividing the channel into windows representing 5–20 times the average peak width. For the HIV-1 data set, the window size was 200 because peak widths usually are $\sim 10 \pm 5$ data points.

Matrixing

Spectral overlap in each channel was removed using a linear transformation matrix so that each channel represents dye amount as a function of position. The transformation matrix is calibrated using four extension reactions run in separate capillary columns, which need be performed only once per dye set. The extension reactions must generate a series of intense, but not saturating, peaks for each fluorophore, and is most easily achieved by generating sequencing channels. The extension products are resolved in independent capillary runs, such that each electropherogram contains fluorescence from a single dye. The user selects an intense peak for each of the dyes, and ShapeFinder automatically calculates the transformation matrix.

Smoothing

The peak-fitting and alignment algorithm implements an internal smoothing step to reduce high-frequency channel noise, which is sufficient in most cases. This smoothing step is used only to facilitate automatic peak detection; peak integration is always performed on the original, unsmoothed, data. If more control is desired over noise reduction, a separate smoothing step can be applied using the Filter-Convolution tool (Giddings et al. 1998). Recommended parameters are a Gaussian width $\sigma = 1$ and window size of 10.

Mobility shift

Mobility shift parameters are calculated using a sequencing experiment in which all four dye-labeled primers are extended

in the presence of the same dideoxy nucleotide and resolved in a single capillary. Tool parameters are initialized by dragging portions of channels so that all peaks align to a user-chosen reference channel. The algorithm models mobility shifts using a polynomial fit of the data. These parameters need be determined only once for a given primer set, but individual electropherograms may require fine-tuning. Two to three fine-tuning iterations of the Mobility Shift: Cubic tool are usually sufficient for an hSHAPE electropherogram.

Signal decay correction

This tool corrects the decay in peak intensity due to the stochastic nature of 2'-hydroxyl modification, the imperfect processivity of reverse transcriptase, or addition of ddNTPs. At each nucleotide position, there is a probability p that a reagent-modified nucleotide will stop reverse transcriptase. The probability that the reaction will continue, q [$q = (1-p)$], yields the exponential form observed for peak drop-off that we model with Equation 1. The algorithm first identifies peak locations, calculates their height, and removes outliers. Peaks are identified by considering seven consecutive points, calculating the slope of the line connecting each sequential pair of points, and averaging the six consecutive slopes. Peak maxima are identified as the point where derivatives transition from positive to negative. Outlier peaks are identified and excluded using a box plot model in which outliers fall outside 1.5 times the interquartile range of the data (Howell 2002). Second, the algorithm fits (Levenberg 1944; Marquardt 1963) the remaining peak heights to Equation 1 to determine A , C , and q . The probability of extension, q , is ~ 0.999 for most datasets, whereas A and C reflect the arbitrary instrument units that describe fluorescence intensity. Each channel is independently corrected for signal decay as:

$$I_{\text{new}}(x) = N \times I_{\text{old}}(x) / D(x), \quad (3)$$

where $I_{\text{old}}(x)$ is the original measured intensity at position x , $D(x)$ is from Equation 1, N is a user-definable rescaling factor that maintains overall peak intensity relative to the other channels, and $I_{\text{new}}(x)$ is the new, corrected, intensity at position x .

Scale Factor

This simple tool rescales individual channels in a trace using a user-specified, nonzero channel scaling factor. Data should be scaled so that peak heights range above 100 (arbitrary) units to improve the accuracy of subsequent peak fitting.

Align and Integrate

An overview of this algorithm is shown in Figure 4.

(1) Peak finding and linking

This first phase accepts user input that specifies (1) which reactions [(+), (–), or sequencing] correspond to each channel; (2) the region of the preprocessed data to analyze; and (3) the sequence. The sequence is read from an ASCII text file; white space, non-A, G, C, U, T, and N characters, and FASTA headers are ignored. Following a simple smoothing step (Supplemental

Fig. S1, dashed lines), the algorithm identifies peaks in each channel as the highest points in a window of ± 3 neighboring points. Additional peaks are interpolated when the distance between peaks is greater than the most frequent inter-peak distance (Supplemental Fig. S1, initially identified versus interpolated peaks are shown by the open and closed circles in Phase A, respectively). Interpolated peaks are then shifted left or right to correspond to a maximum [for example, Supplemental Fig. S1, position 2845 in the (+) channel in Phase B]. Finally, peaks in the (+) and (-) reagent channels are aligned with each other if they are positioned near each other on the elution time axis. The algorithm matches the perfectly aligned peaks first, and then incrementally relaxes the peak offset from zero to a maximum value $k/2$, where k is the median distance between neighboring peaks, to link the remaining peaks (in Supplemental Fig. S1, illustrated by lines linking the circles).

The Setup phase also implements a Refine option that creates a peak in the (-) reagent channel when a matching peak is not found corresponding to one identified in the (+) reagent channel (Supplemental Fig. S1, Phase C, for example, position 2627). Also, if a matching (+) reagent peak is not found for an identified (-) reagent peak, the (-) reagent peak is automatically deleted. When the Modify option is used to add or delete peaks (see Fig. 4), the algorithm adds these peaks and automatically creates the appropriate new peak links (Supplemental Fig. S1, Phase D).

Peaks in the sequencing channels are identified and aligned in a similar fashion (Supplemental Fig. S2, Phases A and B). Real and background peaks in the sequencing channels are distinguished using a user-definable sensitivity cutoff: Peaks are part of the sequencing ladder only if their height is greater than the median channel intensity multiplied by the sensitivity level. Decreasing the sensitivity causes more sequencing peaks to be identified as part of a sequencing ladder. In the final step of this phase, sequencing peaks are linked to the (+) and (-) reagent peaks if the peak is within ± 2 points on the x -axis of a (+) or (-) reagent peak (Supplemental Fig. S2, Phase C).

(2) Alignment to the RNA sequence

The algorithm then aligns the trace with the RNA sequence (Fig. 4, Alignment steps). A sequence, S_{link} , is derived by correspondence of the sequence ladder from the ddNTP channels to the (+) reagent peaks. An example sequence for a reaction using ddUTP and ddGTP might be NCAANCNNNCNCAC, where N indicates the positions of nonsequenced nucleotides. S_{link} is then compared with S_{true} , corresponding to the known input RNA sequence, by sliding S_{link} along the length of S_{true} . The algorithm accepts the alignment that contains the most matching positions between S_{link} and S_{true} (Supplemental Fig. S2, illustrated in Phase C). In Figure 2, S_{true} is at the bottom of the data window (as the Input sequence) and S_{link} appears immediately above (aligned sequence). Comparing the two sequences readily identifies a correct alignment.

(3) Editing of the alignment

After the initial alignment, mismatches in the alignment of S_{link} with S_{true} may be observed as consistent horizontal offsets in the data window. Identifying the location of a missed or incorrectly added peak is accomplished in a straightforward way by locating

the position where horizontal offset begins. The alignment is edited iteratively by manually adding or deleting peaks using the Modify panel. When adding a (+) reagent peak, adding a corresponding (-) reagent peak is often necessary.

(4) Whole-channel Gaussian peak fitting

Once the alignment is correct, the intensities of each peak in the (+) and (-) reagent channels are quantified by fitting a Gaussian curve to each peak in the entire channel (Equation 2). Peaks are characterized by area, position, and width (A_i , μ_i , and σ_i , respectively). The peak position, μ_i , was determined during the peak finding phase; thus, area, A_i , and width, σ_i , are the remaining unknowns. ShapeFinder uses an exhaustive search algorithm to optimize A_i and σ_i for each peak.

Initial estimates of A_i and σ_i for a given peak are calculated from a local three-peak Gaussian fit consisting of the target peak and its neighboring peaks on each side. Initial values of A_i are taken from $\gamma_i / 2 \leq A_i \leq 10\gamma_i$, where γ_i is the amplitude of the peak fluorescence intensity. σ_i estimates are taken from the range $0.8 \leq \sigma_i \leq 4.5$. The algorithm optimizes A_i first by adjusting the sample space of A in the interval $A_{i,\text{best}} - 0.5A_{i,\text{best}} \leq A_i \leq A_{i,\text{best}} + 0.5A_{i,\text{best}}$, where $A_{i,\text{best}}$ is the area calculation from the previous round that best fit the data. Next, the search algorithm refines estimates for both A_i and σ_i by using a different sample space for σ_i , $0.4\sigma_{\text{med}} \leq \sigma_i \leq \sigma_i + 0.5\sigma_{\text{med}}$, where σ_{med} is the peak width median calculated from all σ_i estimated previously. These steps yield good agreement between the experimental and fit intensities, although a subset of the peak areas is underestimated slightly (see Fig. 5A).

If the Optimize option is enabled, estimates of A_i and σ_i are improved further. New parameters are estimated by sampling $A_i \leq A_{\text{new}} \leq A_i + 10A_i$, and fixing the width, ω , as the minimum σ_i computed thus far; each σ_i is retained as $\sigma_{i,\text{old}}$ for the future. As the initial new ω is the minimum σ_i computed so far, A_i estimates increase to compensate for the smaller ω . In the final phase of the Optimize algorithm, peak widths are improved in two stages. Peak widths are first optimized by sampling a new σ_i from $\omega \leq \sigma_{i,\text{new}} \leq \sigma_{i,\text{old}}$, starting with the peak with the poorest fit and proceeding to the remaining peaks, for three iterations. In the second stage, the peak with the poorest fit is optimized by sampling, $\sigma_i \leq \sigma_{i,\text{new}} \leq \sigma_i + 0.1\sigma_{i,\text{old}}$, provided $\sigma_i + 0.1\sigma_{i,\text{old}} < \sigma_{i,\text{old}}$. The new width is accepted only if it improves the fit. Results of this final optimization step are shown in Figure 5B. Fitting ~ 400 nt of the HIV-1 sample data requires ~ 16 min on a 1.5-GHz Power PC processor versus 3 min with the Optimize option disabled.

The absolute reactivities for all analytical peaks in the trace are then calculated by subtracting the (-) reagent areas from those for the (+) reagent channel and are output in text files. The Input versus Fit File contains the calculated curve fit for the (+) and (-) reagent channels. The Integrated Peaks File lists calculated peak positions, widths, areas, RMS errors for the (+) reagent (RX) and (-) reagent (BG) channels, the alignment to the target RNA sequence, and net absolute hSHAPE reactivities (Fig. 6C). As the lengths of each cDNA fragment in the sequencing channel are one nucleotide longer than the (+) reagent channel, the sequencing alignment is shifted by one nucleotide such that (+) and (-) reagent reactivity information is attributed to the correct nucleotide position (Supplemental Fig. S2, Phase D). Only the Integrated Peaks File reflects this shift; previous processing steps do not account for this offset.

Statistical analyses

Statistical analyses were performed using R (R Development Core Team 2008). If a nucleotide was present in one dataset, but was absent in the others, the nucleotide was removed from the analysis. Pearson's correlation coefficients were computed for each possible pairing of the five HIV-1 data sets, resulting in 10 calculated correlation coefficients per position. One-way ANOVA and Levene's tests were employed for determining mean reactivity differences and differences in reactivity variation among the five HIV datasets, respectively.

SUPPLEMENTAL DATA

Supplemental material can be found at <http://www.rnajournal.org>.

ACKNOWLEDGMENTS

This work was supported by a grant from the U.S. National Institutes of Health (AI068462 to K.M.W. and M.C.G.). We thank many members of the Weeks laboratory for continuous feedback and suggestions during the software development process.

Received May 2, 2008; accepted July 9, 2008.

REFERENCES

- Badorrek, C.S. and Weeks, K.M. 2005. RNA flexibility in the dimerization domain of a γ retrovirus. *Nat. Chem. Biol.* **1**: 104–111.
- Badorrek, C.S. and Weeks, K.M. 2006. Architecture of a γ retroviral genomic RNA dimer. *Biochemistry* **45**: 12664–12672.
- Badorrek, C.S., Gherghe, C.M., and Weeks, K.M. 2006. Structure of an RNA switch that enforces stringent retroviral genomic RNA dimerization. *Proc. Natl. Acad. Sci.* **103**: 13640–13645.
- Brenowitz, M., Chance, M.R., Dhavan, G., and Takamoto, K. 2002. Probing the structural dynamics of nucleic acids by quantitative time-resolved and equilibrium hydroxyl radical "footprinting." *Curr. Opin. Struct. Biol.* **12**: 648–653.
- Chamberlin, S.I. and Weeks, K.M. 2000. Mapping local nucleotide flexibility by selective acylation of 2'-amine substituted RNA. *J. Am. Chem. Soc.* **122**: 216–224.
- Chen, Y., Fender, J., Legassie, J.D., Jarstfer, M.B., Bryan, T.M., and Varani, G. 2006. Structure of stem-loop IV of Tetrahymena telomerase RNA. *EMBO J.* **25**: 3156–3166.
- Dann, C.E., Wakeman, C.A., Sieling, C.L., Baker, S.C., Irnov, I., and Winkler, W.C. 2007. Structure and mechanism of a metal-sensing regulatory RNA. *Cell* **130**: 878–892.
- Das, R., Laederach, A., Pearlman, S.M., Herschlag, D., and Altman, R.B. 2005. SAFA: Semiautomated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. *RNA* **11**: 344–354.
- Duncan, C.D.S. and Weeks, K.M. 2008. SHAPE analysis of long-range interactions reveals extensive and thermodynamically preferred misfolding in a fragile group I intron RNA. *Biochemistry* **47**: 8504–8513.
- Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J.P., and Ehresmann, B. 1987. Probing the structure of RNAs in solution. *Nucleic Acids Res.* **15**: 9109–9128.
- Gherghe, C. and Weeks, K.M. 2006. The SL1-SL2 (stem-loop) domain is the primary determinant for stability of the γ retroviral genomic RNA dimer. *J. Biol. Chem.* **281**: 37952–37961.
- Giddings, M.C., Severin, J., Westphall, M., Wu, J., and Smith, L.M. 1998. A software system for data analysis in automated DNA sequencing. *Genome Res.* **8**: 644–665.
- Howell, D.C. 2002. *Statistical methods for psychology*. Duxbury/Thomson Learning, Pacific Grove, CA.
- Jones, C.N., Wilkinson, K.A., Hung, K.T., Weeks, K.M., and Spremulli, L.L. 2008. Lack of secondary structure characterizes the 5' ends of mammalian mitochondrial mRNAs. *RNA* **14**: 862–871.
- Levenberg, K. 1944. A method for the solution of certain problems in least squares. *Q. Appl. Math.* **2**: 164–168.
- Marquardt, D.W. 1963. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.* **11**: 431–441.
- Maxam, A.M. and Gilbert, W. 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65**: 499–560.
- Merino, E.J., Wilkinson, K.A., Coughlan, J.L., and Weeks, K.M. 2005. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127**: 4223–4231.
- Mortimer, S.A. and Weeks, K.M. 2007. A fast acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.* **129**: 4144–4145.
- R Development Core Team. 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org>.
- Stern, S., Moazed, D., and Noller, H.F. 1988. Structural analysis of RNA using chemical and enzymatic probing monitored by primer extension. *Methods Enzymol.* **164**: 481–489.
- Stoddard, C.D., Gilbert, S.D., and Batey, R.T. 2008. Ligand-dependent folding of the three-way junction in the purine riboswitch. *RNA* **14**: 675–684.
- Strobel, S.A. 1999. A chemogenetic approach to RNA function/structure analysis. *Curr. Opin. Struct. Biol.* **9**: 346–352.
- Tullius, T.D. and Greenbaum, J.A. 2005. Mapping nucleic acid structure by hydroxyl radical cleavage. *Curr. Opin. Chem. Biol.* **9**: 127–134.
- Vicens, Q., Gooding, A.R., Laederach, A., and Cech, T.R. 2007. Local RNA structural changes induced by crystallization are revealed by SHAPE. *RNA* **13**: 536–548.
- Wang, B., Wilkinson, K.A., and Weeks, K.M. 2008. Complex ligand-induced conformational changes in tRNA^{Asp} revealed by single nucleotide resolution SHAPE chemistry. *Biochemistry* **47**: 3454–3461.
- Wilkinson, K.A., Merino, E.J., and Weeks, K.M. 2005. RNA SHAPE chemistry reveals non-hierarchical interactions dominate equilibrium structural transitions in tRNA^{Asp} transcripts. *J. Am. Chem. Soc.* **127**: 4659–4667.
- Wilkinson, K.A., Merino, E.J., and Weeks, K.M. 2006. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): Quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protocols* **1**: 1610–1616.
- 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**: e96. doi: 10.1371/journal.pbio.0060096.

Supporting Information for:

ShapeFinder: A software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis

Suzy M. Vasa¹, Nicolas Guex¹, Kevin A. Wilkinson, Kevin M. Weeks*, Morgan C. Giddings*

¹ These authors contributed equally.

* Correspondence: weeks@unc.edu and giddings@unc.edu

Figure S1. Peak finding. The electropherogram shows the (+) and (–) reagent channels (blue and green, respectively). The (–) reagent intensities have been inverted and are plotted on an expanded scale to facilitate visualization of peak synchronization. Preprocessed channels are shown as solid lines, channels smoothed over a 3-point window are dashed. (A) Identification of peak positions by analysis of (i) signal amplitude versus (ii) interpolation are illustrated by open and closed circles, respectively. (B) Refinement of interpolated peak positions. (C) Automatic addition of missing peaks (blue circles) after comparison of the (+) and (–) reagent channels. (D) Incorporation of peaks added (red circles) or deleted by the user and subsequent refinement of peak positions. Positions of synchronized (+) and (–) reagent peaks that will be used during the integration phase are emphasized with solid lines.

Figure S2. Sequence alignment. Electropherogram showing the (+) reagent (bottom) and the ddNTP (top) channels. Sequencing channels have been inverted to facilitate visualization of peak synchronization with the reagent channel. Reagent peaks assigned in the alignment step (Figure S1) are highlighted with blue dotted lines. Results of the four phases of sequence assignment are shown explicitly. (A,B) Detection of peaks corresponding to the first and second sequencing channels, respectively. Peaks not accepted as valid sequencing positions are shown with black and red filled circles. (C) Assignment of input sequence with the identified sequencing peaks. (D) Complete alignment of the input RNA sequence. This alignment is offset by one nucleotide to reflect that dideoxy sequencing fragments are 1 nucleotide longer than the cDNA fragments that identify 2'-O-adduct sites.

