A PUBLICATION OF THE RNA SOCIETY

SAFA: Semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments

RHIJU DAS, ALAIN LAEDERACH, SAMUEL M. PEARLMAN, et al.

References	This article cites 38 articles, 12 of which can be accessed free at: **http://rnajournal.cshlp.org/content/11/3/344.full.html#ref-list-1** Article cited in: **http://rnajournal.cshlp.org/content/11/3/344.full.html#related-urls**
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here**

To subscribe to *RNA* go to:
http://rnajournal.cshlp.org/subscriptions

METHOD

SAFA: Semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments

RHIJU DAS,[1,2,4], ALAIN LAEDERACH[3,4] SAMUEL M. PEARLMAN,[4] DANIEL HERSCHLAG,[2] and RUSS B. ALTMAN[3]

[1]Department of Physics, [2]Department of Biochemistry, [3]Department of Genetics, and [4]Department of Computer Science, Stanford University, Stanford, California 94305, USA

ABSTRACT

Footprinting is a powerful and widely used tool for characterizing the structure, thermodynamics, and kinetics of nucleic acid folding and ligand binding reactions. However, quantitative analysis of the gel images produced by footprinting experiments is tedious and time-consuming, due to the absence of informatics tools specifically designed for footprinting analysis. We have developed SAFA, a semi-automated footprinting analysis software package that achieves accurate gel quantification while reducing the time to analyze a gel from several hours to 15 min or less. The increase in analysis speed is achieved through a graphical user interface that implements a novel methodology for lane and band assignment, called "gel rectification," and an optimized band deconvolution algorithm. The SAFA software yields results that are consistent with published methodologies and reduces the investigator-dependent variability compared to less automated methods. These software developments simplify the analysis procedure for a footprinting gel and can therefore facilitate the use of quantitative footprinting techniques in nucleic acid laboratories that otherwise might not have considered their use. Further, the increased throughput provided by SAFA may allow a more comprehensive understanding of molecular interactions. The software and documentation are freely available for download at http://safa.stanford.edu.

Keywords: electrophoresis; footprinting; gel; hydroxyl radical; RNA folding; software

INTRODUCTION

'Footprinting' describes a family of techniques that utilize protection of DNA and RNA from cleavage or modification to report on structure and ligand binding with as fine as single-nucleotide resolution. Footprinting studies have provided a wealth of structural and functional information on nucleic acid behavior, and continue to be widely used in both qualitative and quantitative investigations. Among the numerous applications of footprinting approaches that probe secondary and tertiary structure are the mapping of DNA structure (Sun et al. 1999; Guo and Tullius 2003), thermodynamic dissection of fundamental DNA-protein interactions (Ramesh and Nagaraja 1996; Gross et al. 1998a,b; Brenowitz et al. 2002), portraits of RNA folding pathways at millisecond time resolution (Sclavi et al. 1997, 1998a,b; Deras et al. 2000; Brenowitz et al. 2002; Uchida et al. 2003), and high-resolution structural pictures of RNA-protein interactions in the ribosome (Heilek et al. 1995; Wilson and Noller 1998; Joseph et al. 2000).

The typical readouts of nucleic acid footprinting experiments are images of end-labeled nucleic acid fragments sorted electrophoretically by length using polyacrylamide gels. While visual inspection is frequently sufficient for a qualitative analysis, quantification of the relative proportions of the reaction products is essential for quantitative structural, thermodynamic, and kinetic studies. For gels imaged by methods with linear response (e.g., phosphor storage imaging for ^{32}P-labeled RNA), quantified gel band intensities corresponding to each nucleotide are extracted for a probed molecule and can be used to determine the midpoints and cooperativity of thermodynamic titrations (Johnson et al. 1979; Ackers et al. 1982; Celander and Cech 1990, 1991), the timescales for folding events (Sclavi et al.

1997, 1998a,b), and the changes in solvent accessibility upon folding or binding (Latham and Cech 1989; Felden et al. 1994, 1996; Loizos 2004). Footprinting probes such as the hydroxyl radical (Shadle et al. 1997) coupled to gel electrophoresis report information for every nucleotide within a molecule with up to many hundreds of residues in length. Thus, even a single gel can generate thousands of data points. A typical quantitative study requires the acquisition of tens of gels, potentially overwhelming the investigator with data to analyze. Footprinting can therefore greatly benefit from high-throughput methods for quantifying gel electrophoretograms. Efficient software for base calling, along with improvements in electrophoretic technology, has significantly contributed to the throughput of DNA sequencing (Ewing and Green 1998; Ewing et al. 1998; Dolnik 1999). In an analogous manner, software specialized for the analysis of footprinting electrophoretograms has the potential to increase the throughput of footprinting techniques.

Historically, quantitative footprinting studies were developed using a quantification method termed 'boxing' in which an investigator manually draws rectangles or polygons surrounding a band or group of bands whose intensity is then integrated (Brenowitz et al. 1986a,b). In addition to being tedious, this method is susceptible to variability in different researchers' boxing choices and neglects the overlap of neighboring bands. While valuable when large groups of bands are integrated (such as in DNase I footprint titrations of protein binding), boxing is a poor choice for the individual analysis of a plethora of single bands.

An alternative approach is to generate lane profiles that report the changes in band intensity in the direction of electrophoresis (Shadle et al. 1997). These lane profiles can be viewed as 'footprinting spectra' composed of overlapping peaks. The development of band shape models allow the deconvolution of band overlap, provided proper initial guesses for the band centers and widths are given (Strahs and Brenowitz 1994; Shadle et al. 1997; Pastor et al. 2000; Takamoto et al. 2004a,b). These advances, along with the introduction of a well defined normalization procedure (Takamoto et al. 2004a), have significantly contributed to the accuracy, speed, and reproducibility of the analysis of footprinting gels, but the procedures have not become widespread, possibly due to their complexity. An experienced user still needs an hour or longer to assign band centers, and several more hours of computational time to apply the deconvolution procedure to a typical gel. Indeed, the time required for these computational steps is larger than the typical time spent in preparing the samples and carrying out gel electrophoresis. To access the full potential of footprinting as a high-throughput technique, an automated, rapid, and straightforward procedure for the analysis of gels is needed.

To accomplish this goal, we developed software specifically designed to accelerate single band fitting analysis while preserving or enhancing accuracy and precision in the quantification. The software package SAFA (Semi-Automated Footprinting Analysis) allows a footprinting gel to be exhaustively quantified in as little as 15 min. Four elements contribute to SAFA's speed and accuracy. First, a graphical user interface guides the user sequentially through the analysis (Fig. 1), from cropping the gel image to generating the numerical output of the band intensities. The natural progression of operations (see buttons in Fig. 1) allows researchers to quickly become proficient with SAFA. Keyboard shortcuts make it possible for experienced users to further accelerate their analysis. Second, all of the steps involved in gel analysis (such as lane definition) have been automated; the user immediately visualizes the result of the automated procedure at each step and is always given the option to manually override software missteps. These steps include the automated correction of geometric distortions in the gel, a widespread practical problem. Third, peak-fitting algorithms have been streamlined and accelerated. Fourth, a standard file format is implemented that stores each step of the gel analysis so that users can easily exchange footprinting data, revise the quantification procedure in later sessions, and utilize the resolved band intensities. In addition, a normalization routine developed by Takamoto et al. (2004a) was implemented within this software package for convenience and to minimize processing time. Testing of the SAFA software in our and several other laboratories has demonstrated the ability of this program to achieve the goals of speed, accuracy, and precision. It is our hope that the ease with which inexperienced researchers can use SAFA will stimulate laboratories that use footprinting as a qualitative tool to pursue the quantitative information that is inherent to properly designed footprinting experiments.

METHOD

Footprinting experiments result in gels with several lanes that contain hundreds of bands, many of which display significant overlap with neighboring bands directly above and below (see Fig. 2A; for the purposes of this report, the running direction of the gel is presented as vertically downward, as this is the most common alignment in the biochemical literature). Analysis of such gels can be divided into two parts: a procedure that assigns each band of a gel with a lane number and a residue number (numbers in Fig. 2A), and a quantification procedure that integrates the counts in each band, deconvolving any overlap between neighboring bands. The following sections describe the algorithms implemented for each of these procedures and follow the data flow in SAFA.

Lane and band assignment by "gel rectification"

The first step in gel analysis is the assignment of the thousands of bands within the gel image. This poses the major
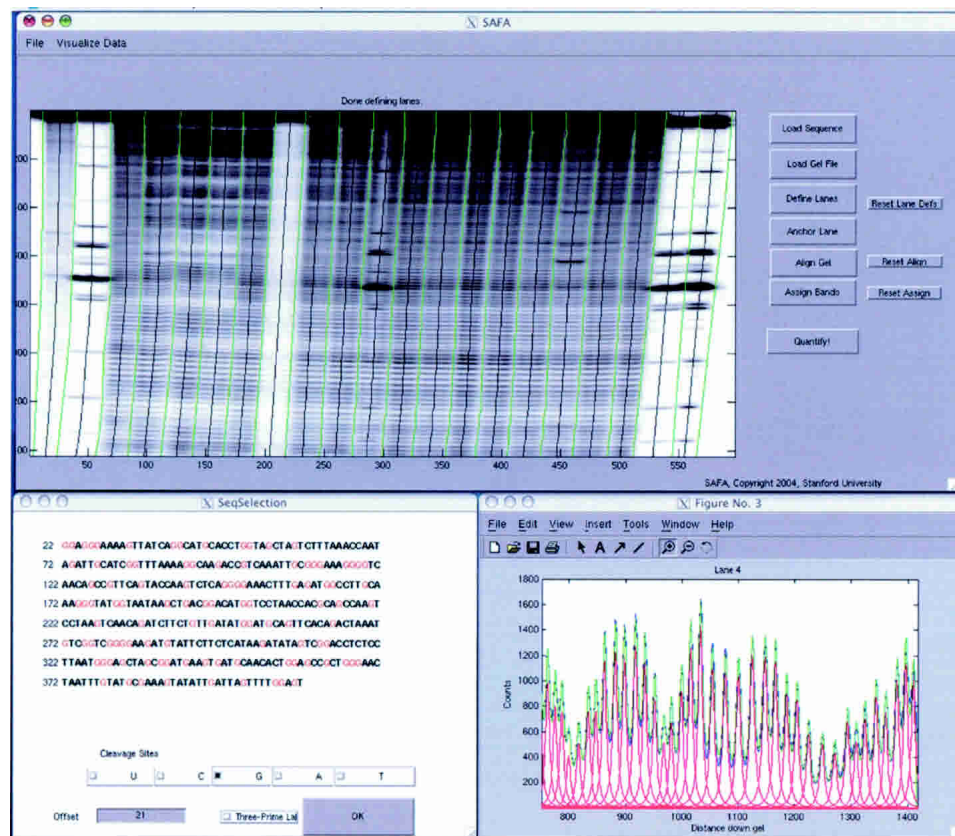
FIGURE 1. Screen capture of Semi-Automated Footprinting Analysis (SAFA) software displaying the loaded image of an RNA footprinting gel (*upper* window), the sequence browser (*lower left* window), and a peak-fitting viewer (*lower right* window). The buttons on the *right* side of the *upper* window sequentially take the user through the different steps of gel analysis, from loading a gel file to quantifying the data. The sequence browser allows the user to define any possible cleavage pattern for the reference lane.

bottleneck for the quantification procedure. In particular, the boundaries between such bands are not straight horizontal and vertical lines but are often curved due to imperfect wells or well loading, to lane widening, and to variation of band mobilities across the gel (Fig. 2A). In principle these effects can be minimized by careful attention to gel loading and by minimizing salt and temperature inhomogeneities. In practice, however, most gels have some imperfections, and much time and data would be lost if imperfect lanes or gels were not quantified. The accurate assignment of each complex boundary line for a gel image can be a time-consuming task for researchers and a complicated task for an automated software algorithm.

The central innovation of SAFA is a simple, semi-automated procedure for band assignment. After assignment of a few of the lane and band boundaries by the researcher, the software fills in guesses for the remaining assignments that can, in turn, be further refined by the researcher. To help the researcher visualize the progress of the band assignment at each step, the two-dimensional gel image is updated by a transformation that maps each (curved) lane or band boundary onto a vertical or horizontal line, respectively. At

the end of this gel image "rectification," the researcher sees a straightened gel image that represents the hypothetical result of an ideal gel electrophoresis experiment, with perfectly straight lanes and horizontal rows of product bands (Fig. 2B). The image transformation bears similarities to the "spatial warping" procedure designed for comparing protein two-dimensional gel images (Lemkin 1997, 1999; Lemkin and Thornwall 1999), but the SAFA rectification is optimized for footprinting gels and preserves the number of counts on the gel to allow accurate quantification. Although the rectification procedure may appear to distort the raw data, we emphasize that the mapping transformation is mathematically reversible, so that the act of defining of each straight line on the rectified image is completely equivalent to drawing of a curved lane or band boundary on the original, untransformed image. Further actions on the rectified gel, including the band deconvolution algorithm, can be similarly mapped back to equivalent actions on the original image. The gel rectification technique implemented in SAFA is therefore similar to previous methods (Takamoto et al. 2004a), but requires less user input and time as many of the steps have been automated and are easier to visualize.
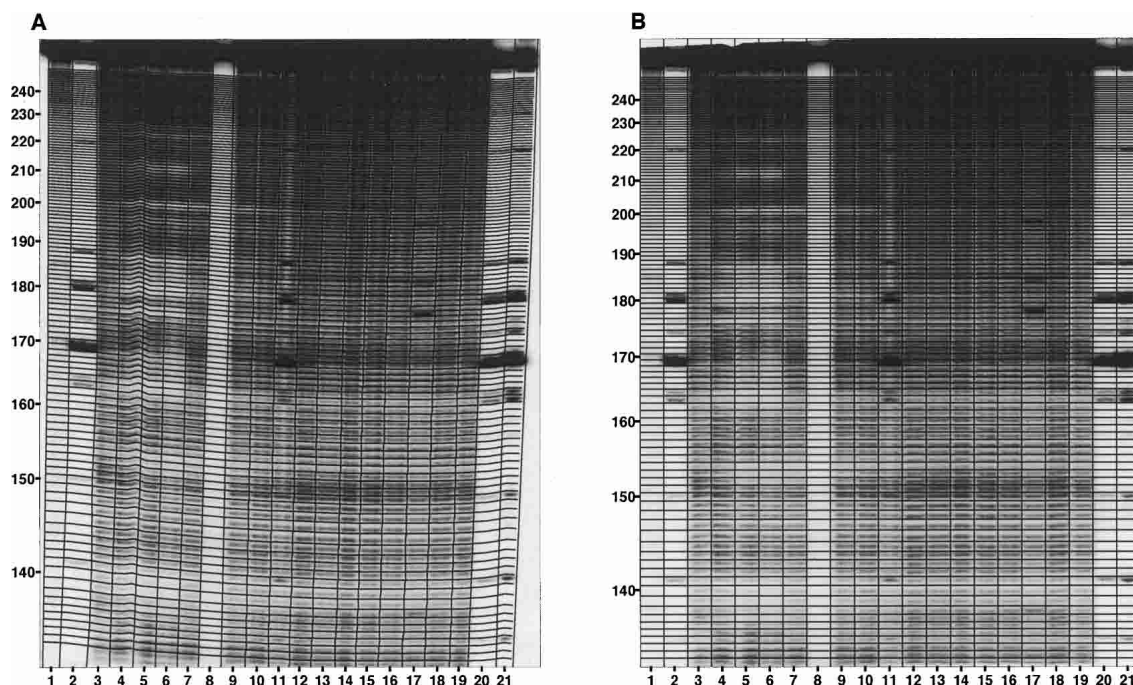
FIGURE 2. The gel rectification procedure used in SAFA to help visualize lane and band boundary assignments. (*A*) A region of a nucleic acid footprinting gel mildly distorted by salt effects and temperature variations; the lane and band boundaries (black lines) are not straight lines, making their assignment difficult. (*B*) Final visualization of the same gel after a rectifying transformation that maps lane and band boundaries to vertical and horizontal lines. The gel image is taken from a hydroxyl radical footprinting gel for the P4-P6 RNA domain from the *Tetrahymena* group I ribozyme (Takamoto et al.). The lanes contain an untreated sample in *1*; ribonuclease T1 digests as reference ladders in *2*, *20*, *11*, and *21*; and footprinting reactions in the presence of 10 mM sodium cacodylate, pH 7.0, and 0.1 mM EDTA plus the following ion concentrations in lanes *3–19*: none, 1 mM MgCl$_2$, 10 mM MgCl$_2$, 100 mM MgCl$_2$, 2 M NaCl, 1.5 M NaCl, 1 M NaCl, 0.7 M NaCl, none, 0.5 M NaCl, 0.3 M NaCl, 0.15 M NaCl, 0.10 M NaCl, 0.08 M NaCl, 0.04 M NaCl, 0.02 M NaCl, none.

The rectification procedure is separated into three sequential steps: (1) lane definition, (2) horizontal gel alignment, and (3) assignment of residue numbers to bands.

(1) Lane definition

In the previous methodology of Takamoto et al., definition of lanes along the vertical axis was accomplished by manually drawing lines down the center of each lane (Takamoto et al. 2004a). The previous procedure was difficult to automate since lane centers do not necessarily correspond to maxima in the horizontal direction. Generally, however, boundaries between lanes are easily defined as regularly spaced continuous minima in the gel image. In SAFA, initial estimates for the curves that delineate the boundaries between lanes are guessed by the software and optionally refined by the user (Fig. 3A). After lane boundaries are drawn, the pixel intensities within each of the lanes are binned so as to map to a rectangle with a user-defined number of columns (default of 10 columns per lane). The resulting transformation rectifies the gel image vertically by aligning all lane boundaries to the vertical axis (straight lines in Fig. 3B).

(2) Horizontal gel alignment

After the lane definition, the gel image typically still appears distorted because the gel "smiles" or "frowns", i.e., bands that correspond to the same nucleic acid product in different lanes do not run at exactly the same vertical position. The next alignment procedure accounts for these residual distortions. The researcher draws several "anchor lines" across the gel image connecting corresponding bands in the gel, e.g., across the uncleaved major band in each lane of a hydroxyl radical footprinting gel and well defined bands at the bottom of the gel (Fig. 3C). A transformation of the image is carried out to map the anchor lines to straight horizontal lines (Fig. 3D). The regions between anchor lines are mapped by linear interpolation, with intensities scaled to preserve the total number of counts in these regions. Further lines can be added or erased by the researcher until the gel is completely aligned.

(3) Assignment of residue number

A procedure for annotation of band numbers on the rectified gel is the final semi-automated step of SAFA. Assignment of residue numbers for bands in each lane is carried out by comparison to a known band pattern in a reference lane. For example, an enzymatic digest by ribonuclease T1 at guanosine sites is typically used in RNA footprinting experiments to generate a cleavage ladder that allows clear residue assignments. A sequence browser (lower left window in Fig. 1) allows the researcher to specify the residue

numbering scheme and the cleavage pattern used in the reference lane. The researcher then identifies the reference bands in the reference lane (e.g., for each guanosine for a ribonuclease T1 ladder) by clicking on the gel image. The positions of all bands between reference bands in the gel are automatically interpolated and can be refined by the researcher (Fig. 3E). The band positions are used as initial estimates in the fitting procedure (y_i^{start}; see below).

Following gel rectification and band assignment, the pixel intensities along the horizontal direction of each rectangular lane are integrated to generate a profile of counts down each lane (Fig. 4A,B, dark curves). Each well resolved band in the gel appears as a peak in a lane profile. The band assignments (Fig. 3E) are stored and passed to the fitting algorithm as starting estimates for a least-squares optimization, described next. The entire gel rectification and band assignment procedure is scalable in the sense that large gels may be analyzed with little additional user input. Finally, given that the user may always override the automated guesses, even gels with major imperfections can be analyzed with SAFA.

Band deconvolution

Once the gel rectification is complete and the band positions defined, the quantification of band intensities is automated and rapid. The quantification procedure in SAFA fits each lane profile $P(y)$ (y is the vertical position down the lane) to a model profile $P^{\text{pred}}(y)$ that is a sum of Lorentzian peaks $L_i(y)$:

$$p^{\text{pred}}(y) = \sum_i L_i(y)$$

$$L_i(y) = \frac{A_i}{4\left(\dfrac{y - y_i}{W_i}\right)^2 + 1} \tag{1}$$

This functional form has been shown empirically to provide a good description of band shapes (Shadle et al. 1997; Takamoto et al. 2004a). There are three parameters to describe each peak: the amplitude A_i, the full width at half-maximum W_i, and center y_i. The number of counts in each band is given by the Lorentzian peak area $(\pi/2)A_iW_i$.

As described in a previous study by Takamoto et al. (2004a), a fit that allows each of these parameters to vary independently generally leads to nonphysical results; some peaks are assigned unusually large areas and widths at the expense of neighboring peak areas, especially in regions with significantly overlap between bands. To prevent such artifacts, the peak width W_i is constrained to vary smoothly, by forcing it to be linearly dependent on the average distance to the two adjacent peaks:

$$W_i = W_0 + \frac{\delta}{2}\,|\,y_{i-1} - y_{i+1}| \tag{2}$$

In this peak width model with two parameters W_0 and δ, bands of lower mobility (and smaller separation between adjacent bands) have smaller widths, asymptotically approaching a minimum width W_0. The constraint equation (2) is slightly different from the empirical equation used in the methodology of Takamoto et al. (2004a), which asymptotically approaches a zero width for bands of lowest mobility.

As a final condition, SAFA requires that the peaks have positive amplitudes ($A_I > 0$). This constraint prevents nonphysical solutions with negative peak areas occasionally observed in lanes where the relative intensity of two adjacent bands is different by an order of magnitude (e.g., in ribonuclease T1 footprinting).

Given an experimental profile (obtained by integrating intensity across a lane as described above), fitting of the model described in equations 1 and 2 involves finding the peak centers (y_i), amplitudes (A_i), *and the width model parameters* W_0 and δ such that the square of the differences between the predicted profile $P(y)$ and experimental profile $P^{\text{pred}}(y)$ is minimized. A standard Levenberg-Marquardt algorithm is used for the least-squares minimization (Teukolsky et al. 1992), and the algorithm's progress can be monitored in real time by the researcher (lower right window in Fig. 1). Starting values for the peak centers y_i are the band positions y_i^{start} assigned by the user; for the amplitudes A_i are $P(y_i^{\text{start}})$; and for the W_0 and δ parameters are 3 pixel widths and 0.1, respectively. The Levenberg-Marquardt fitting algorithm converged to identical solutions with initial values of the peak centers varied by ±40% of the peak widths. Variation of twofold in the initial values of A_i, W_0, and δ led to the same converged solution (data not shown).

Software implementation

We used the MATLAB programming environment (The MathWorks, http://www.mathworks.com) for the development of SAFA. The software includes a straightforward graphical user interface and requires no knowledge of MATLAB command syntax. The SAFA analysis procedure is greatly facilitated by tools for zooming into the gel, toggling between grayscale and pseudocolor representations of the image, adjusting the image contrast, and undoing and refining user-chosen lane boundaries and band assignments. The software has been tested on several platforms (Macintosh, Windows, Linux), allowing widespread use. The most current SAFA executable and source code may be downloaded free of charge from http://safa.stanford.edu.

As input, SAFA accepts an 8-bit or 16-bit TIFF or a *.gel* file produced by the standard scanner software from Molecular Dynamics and readable in ImageQuant. The output of quantified peak areas is a text file that can be further analyzed in general data visualization/fitting applications like Excel (Microsoft) and Kaleidagraph (Synergy Software), in addition to MATLAB.
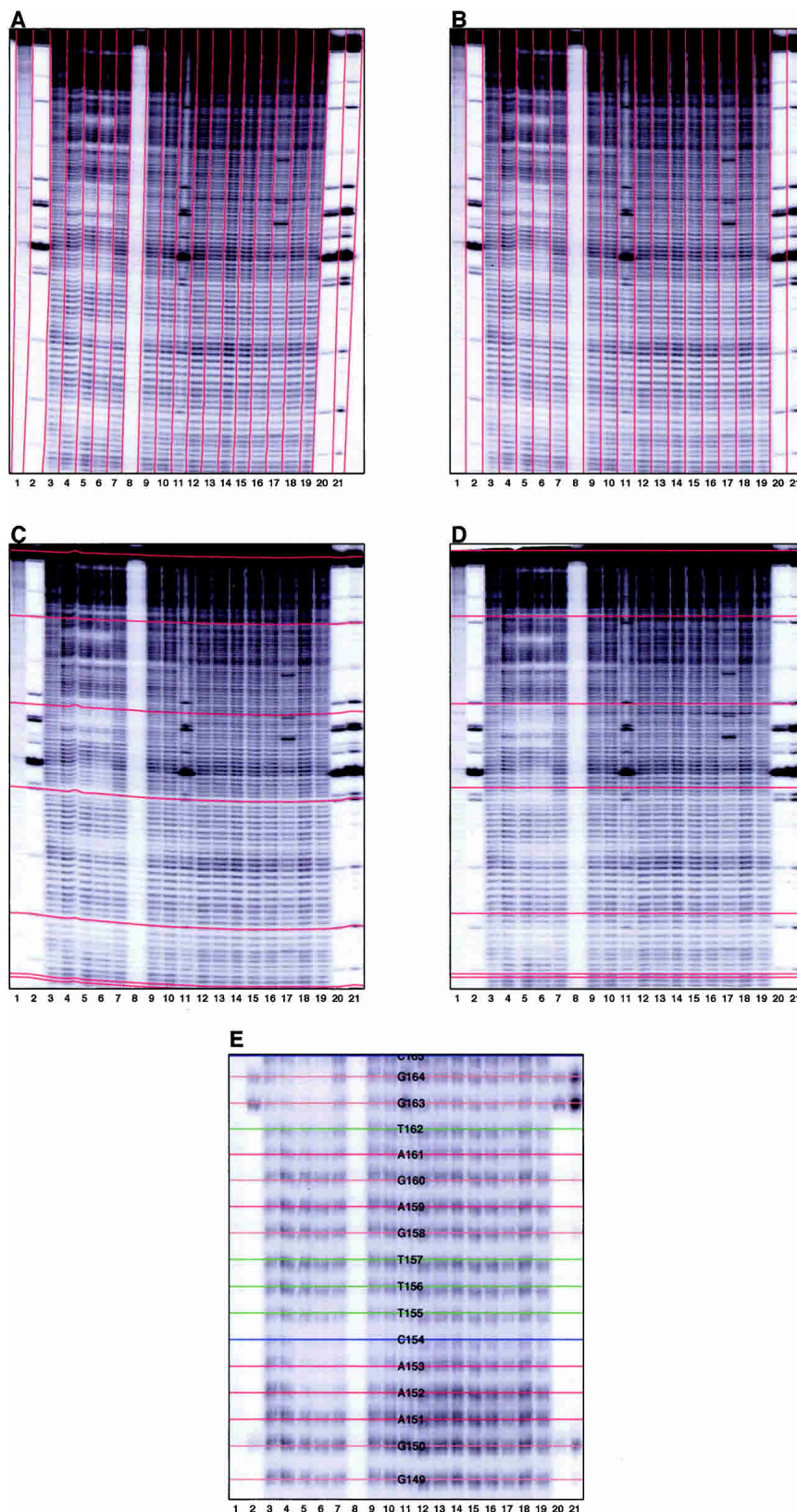
FIGURE 3. Illustration of the individual steps in the gel rectification procedure for rapid definition of lane and band boundaries. (*A*) Cropped gel image with lane boundaries (in red) drawn by the SAFA software and refined by researcher. (*B*) Gel image from *A*, rectified, to align lane boundaries (in red) to vertical axis. (*C*) Anchor lines (in red) connecting corresponding product bands across the different lanes of the gel, defined by the researcher. (*D*) Gel image rectified to align anchor lines (in red) to horizontal axis. (*E*) Assignment of residue numbers to bands in a magnified portion of the rectified gel image based on reference ladders in lanes *2*, *11*, *20*, and *21*.

For thermodynamic titrations and kinetic timecourses, the standardization and normalization procedures of Takamoto et al. (2004a) are particularly useful steps to correct for differences in loading amounts in the quantified footprinting data. An automated utility to carry out these steps is included in the SAFA package.

Software testing

Prior to publication, two SAFA releases were distributed to several laboratories that routinely carry out nucleic acid footprinting experiments. Aside from identifying bugs in the program, the pre-release testing of the software helped to identify the features necessary for making the software generally applicable to a wide range of nucleic acid molecules and footprinting protocols. For example, in the second pre-release, a sequence browser was added that allows the user to define arbitrary sequence numbering and reference lane cleavage patterns.

The software has been tested on titration and structure mapping data for over 150 gels for RNAs including the *Tetrahymena* group I ribozyme and mutants (Latham and Cech 1989; Celander and Cech 1991), the P4-P6 and P5abc subdomains of the *Tetrahymena* ribozyme (Cech 1990; Doherty and Doudna 2000), the *Azoarcus* ribozyme (Adams et al. 2004), and several model hairpin systems (R. Das. and D. Herschlag, unpubl.). To assess user-introduced variability, a hydroxyl radical footprinting gel for the P4-P6 RNA (Fig. 2) was independently analyzed by each of the five authors of the present study. The standard deviations of the quantified peak areas were compared to those obtained from repeated independent quantifications of the same gel by the band-boxing procedure (ImageQuant, Molecular Dynamics) and by the methodology of Takamoto et al. (2004a) (three analyses each).

RESULTS

The new methodology presented herein allows rapid quantification of nucleic

acid footprinting gels by improved two-dimensional image manipulation and nonlinear fitting algorithms. Throughout the development of the SAFA software, we have kept in mind that a state-of-the-art software package must minimally satisfy three basic criteria: (1) The software should yield results that quantitatively agree with previous well tested methodologies; (2) as the software is not fully automated, uncertainties in the quantification due to user-introduced variability must be significantly smaller than uncertainties from other sources of experimental error; and (3) the software should be easier and faster to use than previous methodologies. SAFA satisfies all three criteria, as is described below.

Quantitative comparison to previous methodologies

The most basic test of a new software package is consistency with previous, well tested algorithms. The quantification results of SAFA have therefore been compared to two previous methodologies, the traditional band-boxing procedure and the previous single-band peak-fitting methodology of Takamoto et al. (2004a). Figure 5 illustrates two of these comparisons, for hydroxyl radical footprinting patterns of the P4-P6 RNA from the *Tetrahymena* group I intron in unfolded and Mg^{2+}-folded states. All three procedures give similar results for overall band intensities and identify regions protected upon Mg^{2+}-induced folding. The band areas computed using the band-boxing methodology (gray bars in Fig. 5) show a systematic bias towards increased peak areas at the higher residue numbers that correspond to bands with greater overlap. However, the two peak-fitting methodologies (the method of Takamoto et al. and SAFA, white and black bars, respectively) show excellent agreement with each other (5% and 4% root mean squared deviation between methods for the two lanes analyzed in Fig. 5), despite the differences in the lane definition procedure and peak-width model.

For many footprinting experiments, the absolute band intensities for each lane are less important than the *ratios* of band intensities in different lanes. For example, the data for a residue whose hydroxyl radical reactivity changes during a folding transition is usually plotted relative to the band intensity of a low-salt reference state (Deras et al. 2000; Brenowitz et al. 2002; Uchida et al. 2002). Such normalization can remove systematic errors from the quantification procedure that affect each band similarly in all lanes, e.g., the incorrect deconvolution of overlapping bands in the band-boxing procedure or background cleavage. To assess the residual errors on such ratios, the three quantification methodologies (SAFA, Takamoto et al., and band boxing) were applied to 13 lanes that correspond to a Na^+ titration of the P4-P6 RNA (Uchida et al. 2002). To correct for lane-to-lane loading variation, the data were normalized to invariant band intensities (Takamoto et al. 2004a; the normalization methodology is included in the SAFA software
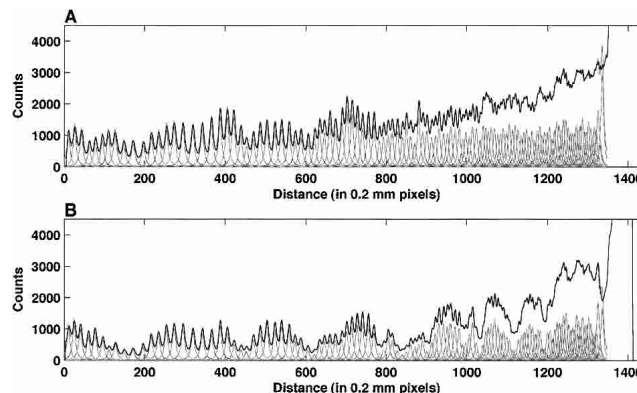

FIGURE 4. The band deconvolution procedure fits the count profiles of the P4-P6 RNA in unfolded (*A*) and Mg^{2+}-folded (*B*) states to a sum of Lorentzian peak shapes and returns the fitted peak areas for the respective lane profiles. The unfolded and folded profiles correspond to lanes *18* and *5*, respectively, in Figure 2. Lane profiles of data in *A* and *B* are shown as solid black lines, the fitted profile as a solid gray line (indistinguishable from data over region of fit), and Lorentzian components of the fit as gray peak shapes. The gel running direction is from *right* to *left* in these depictions.

package). All three quantification methodologies showed a similar scatter of fraction-folded values at each Na^+ concentration (Fig. 6A–C). Standard fits to the Hill relation (Uchida et al. 2002) gave indistinguishable thermodynamic parameters for the band-boxing, Takamoto et al., and SAFA methodologies: midpoints of 820±40 mM, 840±40 mM, and 820±40 mM, respectively, and apparent Hill coefficients of 3.1±0.5, 3.1±0.5, and 3.3±0.5, respectively, with errors estimated from the scatter in the data by bootstrapping (Efron and Tibshirani 1994).

Assessment of user-introduced variability

A fully automated procedure for gel analysis would, by definition, result in zero user-introduced variability. All available quantification procedures are, however, semi-automated, requiring researcher input at several steps. We have assessed the user-introduced variability of the band-boxing procedure, the method of Takamoto et al., and SAFA by carrying out multiple independent analyses of the same gel with each method (Table 1).

The root mean squared deviation between different users' results of the band-boxing procedure is ~6% for well resolved regions. In comparison to the band-boxing procedure, the user-introduced variability of the peak-fitting methodologies is significantly smaller in the well resolved portion of the gel (~1% for both the Takamoto et al. and SAFA procedures; Table 1). These user-to-user variabilities are also shown as error bars on the titration data of Figure 6A–C; for the peak-fitting methodologies, they are mostly obscured by the symbols and are significantly smaller than the scatter of the data points. The primary source of un-
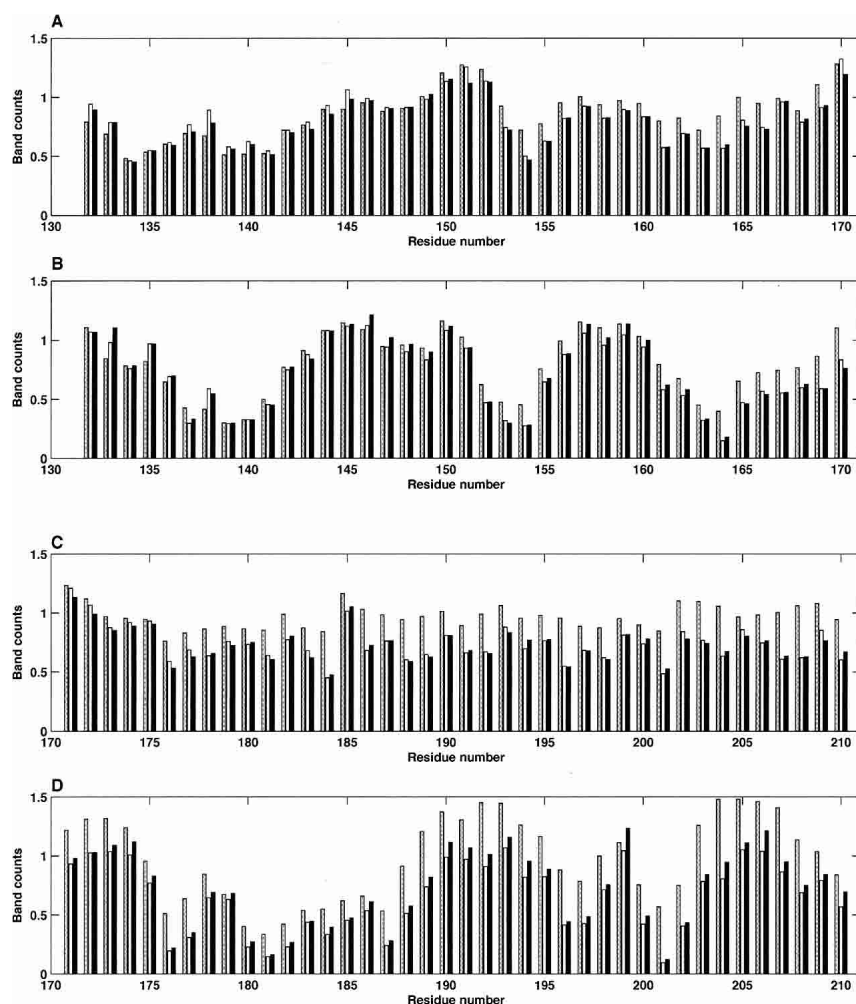
FIGURE 5. Comparison of SAFA analysis (black bars) to the traditional band-boxing procedure (gray bars) and the previous peak-fitting methodology of Takamoto et al. (white bars). To avoid clutter, residues 132–170 are plotted in *A* and *B*, and residues 171–210 in *C* and *D*. To aid visual comparison, the counts have been normalized to the mean value at residues 147–150 for each methodology. The quantified band intensities in (*A,C*) and (*B,D*) correspond to the same footprinting lanes, *18* and *5*, respectively, as shown in Figure 4 for the unfolded and folded states of the P4-P6 RNA.

single-band peak-fitting methodology of Takamoto et al. As a benchmark, careful band-boxing with ImageQuant software takes the researcher ~10 to 15 min per lane, resulting in an overall time demand of 3–5 h to analyze a gel with 20 lanes (Fig. 2). For the same gel, the more accurate and precise peak-fitting methodology of Takamoto et al. requires the researcher to spend 1–3 h to define lanes and to assign initial peak centers and the peak width model. In general, excellent quality gels will require less time to analyze. Further computational time of 5–10 min per lane (Takamoto et al. 2004a) is required to carry out peak fits in the Origin v6.1 software on a current processor (e.g., 1.5 GHz Pentium 4), leading to an overall time of 2.5–6 h for quantification.

The SAFA software accelerates both the band assignment procedure and the peak-fitting algorithm. For an experienced user, the time spent on SAFA's lane assignment, gel rectification, and band assignment procedure *for the entire gel image* ranges from 5 to 10 min, depending on the quality of the gel. The peak-fitting algorithm requires 5 to 10 sec per lane on current processors. This speed increase, relative to the commercial routine in the Takamoto et al. procedure, is likely due to the incorporation of the analytic partial derivatives of equation 1 with respect to the fitting parameters. The overall fitting process is further accelerated because the parameters of the peak width model (equation 2) are fitted simultaneously with the peak centers and amplitudes, rather than adjusted by the researcher in multiple iterations (Takamoto et al. 2004a). The overall time demand for the SAFA procedure is 10–15 min for the gel shown in Figure 2.

certainty in the presented titration data therefore does not appear to be variability introduced by the researcher in carrying out the peak-fitting procedure. Rather, the 5%–10% scatter seen in the titration plots are likely due to sources of experimental error outside the software, such as weak nuclease contaminants and variability in sample handling (pipetting the Fenton reagent, quenching, precipitation, and gel loading) that are not corrected by the normalization procedure.

Time demand of quantification

In addition to maintaining the precision of gel quantification, the SAFA software has been designed to reduce the time demand of gel quantification relative to the traditional band-boxing procedure and to the recently developed

DISCUSSION

Accelerating the slowest step in footprinting experiments

The SAFA software package speeds the quantification and normalization of footprinting gels, from a time demand of a few hours to a few minutes, without sacrificing precision or accuracy (Figs. 5, 6; Table 1). The major time-saving step, relative to the previous single-band peak-fitting methodology of Takamoto et al. (2004a), is the new gel rectification procedure for lane and band assignment. Further

improvements include the two-dimensional visualization of the gel throughout the analysis, rapid revision of band assignments at each step of the analysis, an optimized band fitting algorithm, and a file format that allows researchers to exchange and modify analyses of any gel image. In addition, the normalization procedure of Takamoto et al. (2004a) can substantially improve data precision and is included in SAFA.

The gel quantification procedure is no longer the slowest step in footprinting experiments; over the last year, SAFA has allowed us to obtain quantitative data from over 150 hydroxyl radical footprinting gels for RNAs with lengths ranging from 20 to 400 nucleotides, while developing the software (R. Das and D. Herschlag, unpubl.).

Sources of error in footprinting studies

Biophysical measurements are only meaningful when presented with estimates of their uncertainties, which can arise from variation in sample preparation and from systematic errors in data treatment. These two errors are best appraised by repeated measurements and by repeated data analyses, respectively. For the first error, a single footprinting titration experiment allows a straightforward estimate of measurement error because it intrinsically contains repeat measurements: each gel generally contains several residues that report on a given transition, and their variance yields an estimate of overall error. For the gel presented in Figure 2, the observed scatter in titration data between residues that report on the Na^+ folding transition of the P4-P6 RNA is 5%–10% (Fig. 6). For comparison, the systematic error from data treatment has been estimated by multiple independent analyses with the SAFA software. User-introduced variability was found to be negligible (less than ~1%; Table

TABLE 1. User-to-user variability for the three tested footprinting gel quantification methodologies

Residue number	$\sigma_{user,}$ band boxing	$\sigma_{user,}$ Takamoto et al.	$\sigma_{user,}$ SAFA
132–150	6.6%	1.6%	0.8%
151–170	5.6%	1.3%	0.8%
171–190	8.0%	1.5%	1.0%
191–210	13.0%	4.1%	1.4%
211–230	n.d.	n.d.	1.5%
231–250	n.d.	n.d.	4.1%

Root mean squared deviations (σ_{user}) were calculated between band intensity results from independent quantifications of the same footprinting gel, shown in Fig. 2. Band boxing and the Takamoto et al. method were performed three times each, and SAFA was performed five times, once by each author. The σ_{user} values have been averaged over the specified residues and over the footprinting lanes *5* and *18* shown in Fig. 4. Results were not determined (n.d.) for residue ranges 211–250 for the band-boxing and Takamoto et al. methods.

1) in comparison to the scatter observed in the titration plots (Fig. 6). The primary sources of error for this footprinting experiment are therefore likely to be small variations in the sample handling and small nuclease contaminants.

In gel regions with well separated bands, the accuracy of manual band boxing and SAFA are similar, although SAFA greatly speeds the analysis, as noted above. Further, by deconvolving overlap between closely spaced bands, SAFA expands the region of the gel that can be analyzed with confidence (Table 1). For the most poorly resolved portions of the gel, any quantitative fit to these regions with severe band overlap is suspect (see, e.g., the rightmost fitted peaks in Fig. 4). However, SAFA yields useful qualitative information for these regions that can guide the researcher in deciding whether further experiments are worth pursuing to resolve these regions with, for example, longer electrophoresis times or radioactive labels at the opposite end of the molecules.

Applicability to multiple footprinting techniques

Our primary use of the SAFA software has been for quantifying footprinting experiments in which RNAs are cleaved by solution hydroxyl radicals; these experiments result in dense gels (Fig. 2) in which nearly all bands overlap to some degree with comparably intense neighbors and require the application of band deconvolution algorithms (Shadle et al. 1997; Takamoto et al. 2004a). In contrast, other footprinting techniques result in strong bands that are better isolated, e.g., with hits only at guanosine residues in ribonuclease T1 protection assays, and are thus expected to be less challenging for gel quantification. Indeed, we have been able to apply SAFA without modification to quantify RNA gels of ribonuclease T1 digests (cleavage at guanosines) and
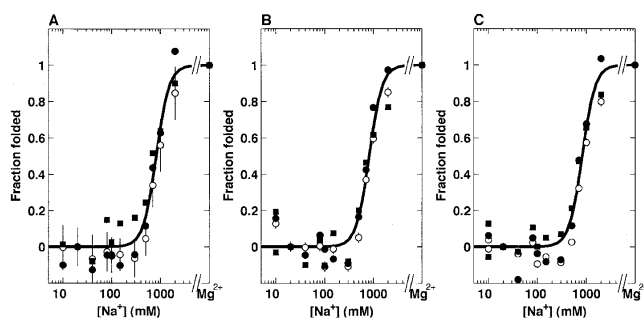

FIGURE 6. Analysis of the P4-P6 Na^+-induced folding transition by three methodologies: band boxing (*A*), the previous peak-fitting methodology of Takamoto et al. (*B*), and the SAFA software (*C*). Sets of quantification results are plotted for residues 153 (○), 154 (●), and 201 (■). The data have been scaled (Uchida et al. 2002) so that a reference unfolded state with low ionic strength (lane *18*) is zero and a reference folded state in 100 mM $MgCl_2$ (lane *5*) is unity (shown as symbols on *right* sides of graphs). The dark curve in each plot represents a Hill fit with midpoint 820 mM and Hill coefficient 3.3. Error bars indicate root mean squared user-to-user variability on residue 153 and, in *B* and *C*, are largely obscured by the symbols.

DNA gels of reverse transcriptase products from dimethyl-sulfate-modified RNAs (modification at adenosines and cytosines) (R. Das and D. Herschlag, unpubl.). In addition, the SAFA software was recently used to quantify gels from tethered radical cleavage experiments (R. Das and D. Herschag, unpubl.) that result in strong cleavage intensities at clusters of nucleotides that are well separated from each other in primary sequence (Joseph et al. 2000). SAFA is expected to have general utility in improving and accelerating the analysis of gel images from a wide variety of footprinting techniques.

ACKNOWLEDGMENTS

REFERENCES

Ackers, G.K., Johnson, A.D., and Shea, M.A. 1982. Quantitative model for gene regulation by λ phage repressor. *Proc. Natl. Acad. Sci.* **79:** 1129–1133.

Adams, P.L., Stahley, M.R., Kosek, A.B., Wang, J., and Strobel, S.A. 2004. Crystal structure of a self-splicing group I intron with both exons. *Nature* **430:** 45–50.

Brenowitz, M., Senear, D.F., Shea, M.A., and Ackers, G.K. 1986a. "Footprint" titrations yield valid thermodynamic isotherms. *Proc. Natl. Acad. Sci.* **83:** 8462–8466.

———. 1986b. Quantitative DNase footprint titration: A method for studying protein-DNA interactions. *Methods Enzymol.* **130:** 132–181.

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.

Cech, T.R. 1990. Self-splicing of group I introns. *Annu. Rev. BioChem.* **59:** 543–568.

Celander, D.W. and Cech, T.R. 1990. Iron(II)-ethylenediaminetetra-acetic acid catalyzed cleavage of RNA and DNA oligonucleotides: Similar reactivity toward single- and double-stranded forms. *Biochemistry* **29:** 1355–1361.

———. 1991. Visualizing the higher order folding of a catalytic RNA molecule. *Science* **251:** 401–407.

Deras, M.L., Brenowitz, M., Ralston, C.Y., Chance, M.R., and Woodson, S.A. 2000. Folding mechanism of the Tetrahymena ribozyme P4-P6 domain. *Biochemistry* **39:** 10975–10985.

Doherty, E.A. and Doudna, J.A. 2000. Ribozyme structures and mechanisms. *Annu. Rev. BioChem.* **69:** 597–615.

Dolnik, V. 1999. DNA sequencing by capillary electrophoresis (review). *J. Biochem. Biophys. Methods* **41:** 103–119.

Efron, B. and Tibshirani, R.J. 1994. *An introduction to the bootstrap.* Chapman and Hall, New York.

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

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Felden, B., Florentz, C., Giege, R., and Westhof, E. 1994. Solution structure of the 3′-end of brome mosaic virus genomic RNAs. Conformational mimicry with canonical tRNAs. *J. Mol. Biol.* **235:** 508–531.

Felden, B., Florentz, C., Westhof, E., and Giege, R. 1996. Usefulness of functional and structural solution data for the modeling of tRNA-like structures. *Pharm. Acta Helv.* **71:** 3–9.

Gross, P., Arrowsmith, C.H., and Macgregor Jr., R.B. 1998a. Hydroxyl radical footprinting of DNA complexes of the ets domain of PU.1 and its comparison to the crystal structure. *Biochemistry* **37:** 5129–5135.

Gross, P., Yee, A.A., Arrowsmith, C.H., and Macgregor, R.B, Jr. 1998b. Quantitative hydroxyl radical footprinting reveals cooperative interactions between DNA-binding subdomains of PU.1 and IRF4. *Biochemistry* **37:** 9802–9811.

Guo, H. and Tullius, T.D. 2003. Gapped DNA is anisotropically bent. *Proc. Natl. Acad. Sci.* **100:** 3743–3747.

Heilek, G.M., Marusak, R., Meares, C.F., and Noller, H.F. 1995. Directed hydroxyl radical probing of 16S rRNA using Fe(II) tethered to ribosomal protein S4. *Proc. Natl. Acad. Sci.* **92:** 1113–1116.

Johnson, A.D., Meyer, B.J., and Ptashne, M. 1979. Interactions between DNA-bound repressors govern regulation by the λ phage repressor. *Proc. Natl. Acad. Sci.* **76:** 5061–5065.

Joseph, S., Whirl, M.L., Kondo, D., Noller, H.F., and Altman, R.B. 2000. Calculation of the relative geometry of tRNAs in the ribosome from directed hydroxyl-radical probing data. *RNA* **6:** 220–232.

Latham, J.A. and Cech, T.R. 1989. Defining the inside and outside of a catalytic RNA molecule. *Science* **245:** 276–282.

Lemkin, P.F. 1997. Comparing two-dimensional electrophoretic gel images across the Internet. *Electrophoresis* **18:** 461–470.

Lemkin, P.F. and Thornwall, G. 1999. Comparing 2-D electrophoretic gels across Internet databases. *Methods Mol. Biol.* **112:** 393–410.

———. 1999. Flicker image comparison of 2-D gel images for putative protein identification using the 2DWG meta-database. *Mol. Biotechnol.* **12:** 159–172.

Loizos, N. 2004. Mapping protein-ligand interactions by hydroxyl-radical protein footprinting. *Methods Mol. Biol.* **261:** 199–210.

Pastor, N., Weinstein, H., Jamison, E., and Brenowitz, M. 2000. A detailed interpretation of OH radical footprints in a TBP-DNA complex reveals the role of dynamics in the mechanism of sequence-specific binding. *J. Mol. Biol.* **304:** 55–68.

Ramesh, V. and Nagaraja, V. 1996. Sequence-specific DNA binding of the phage Mu C protein: Footprinting analysis reveals altered DNA conformation upon protein binding. *J. Mol. Biol.* **260:** 22–33.

Sclavi, B., Woodson, S., Sullivan, M., Chance, M.R., and Brenowitz, M. 1997. Time-resolved synchrotron X-ray "footprinting," a new approach to the study of nucleic acid structure and function: Application to protein-DNA interactions and RNA folding. *J. Mol. Biol.* **266:** 144–159.

Sclavi, B., Sullivan, M., Chance, M.R., Brenowitz, M., and Woodson, S.A. 1998a. RNA folding at millisecond intervals by synchrotron hydroxyl radical footprinting. *Science* **279:** 1940–1943.

Sclavi, B., Woodson, S., Sullivan, M., Chance, M., and Brenowitz, M. 1998b. Following the folding of RNA with time-resolved synchrotron X-ray footprinting. *Methods Enzymol.* **295:** 379–402.

Shadle, S.E., Allen, D.F., Guo, H., Pogozelski, W.K., Bashkin, J.S., and Tullius, T.D. 1997. Quantitative analysis of electrophoresis data: Novel curve fitting methodology and its application to the determination of a protein-DNA binding constant. *Nucleic Acids Res.* **25:** 850–860.

Strahs, D. and Brenowitz, M. 1994. DNA conformational changes associated with the cooperative binding of cI-repressor of bacteriophage λ to OR. *J. Mol. Biol.* **244:** 494–510.

Sun, W., Mao, C., Iwasaki, H., Kemper, B., and Seeman, N.C. 1999. No braiding of Holliday junctions in positively supercoiled DNA molecules. *J. Mol. Biol.* **294:** 683–699.

Takamoto, K., Chance, M.R., and Brenowitz, M. 2004a. Semi-automated, single-band peak-fitting analysis of hydroxyl radical nucleic acid footprint autoradiograms for the quantitative analysis of transitions. *Nucleic Acids Res.* **32(15):** E119.

Takamoto, K., Das, R., He, Q., Doniach, S., Brenowitz, M., Herschlag, D., Chance, M.R. 2004b. Principles of RNA compaction: Insights from the equilibrium folding pathway of the P4-P6 RNA domain in monovalent cations. *J. Mol. Biol.* **343(5):** 1195–1206.

Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. 1992. *Numerical recipes in C: The art of scientific computing.* Cambridge University Press, United Kingdom.

Uchida, T., He, Q., Ralston, C.Y., Brenowitz, M., and Chance, M.R. 2002. Linkage of monovalent and divalent ion binding in the folding of the P4-P6 domain of the *Tetrahymena* ribozyme. *Biochemistry* **41:** 5799–5806.

Uchida, T., Takamoto, K., He, Q., Chance, M.R., and Brenowitz, M. 2003. Multiple monovalent ion-dependent pathways for the folding of the L-21 *Tetrahymena thermophila* ribozyme. *J. Mol. Biol.* **328:** 463–478.

Wilson, K.S. and Noller, H.F. 1998. Mapping the position of translational elongation factor EF-G in the ribosome by directed hydroxyl radical probing. *Cell* **92:** 131–139.