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

Seungmyung Lee¹, Hanjoo Kim¹, Siqi Tian², Taehoon Lee¹, Sungroh Yoon¹,³,* and Rhiju Das²,⁴,*

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

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on November 20, 2014; revised on April 27, 2015; accepted on April 29, 2015

Abstract

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

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

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

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

Supplementary information: Supplementary data are available at Bioinformatics online.
1 Introduction

RNA molecules play diverse roles in encoding and regulating genetic information, and much of this versatility can be traced to the formation of intricate RNA structures. To this end, chemical probing methodologies provide a general and rapid means to mapping RNA secondary and tertiary structure at single-nucleotide resolution (Weeks, 2010).

There exist many chemical probing techniques, most of which have common experimental procedures, as follows. Given an RNA of interest folded in solution, a chemical reagent modifies the RNA, either cleaving it or forming a covalent adduct with it at a rate correlated with the accessibility of particular moieties at each nucleotide or the frequency at which each nucleotide fluctuates into a conformation activated for chemical reaction. Examples of such chemical reagents, all with distinct mechanisms, include hydroxyl radicals, 2'-OH acylating chemicals (SHAPE), dimethyl sulfate (DMS) and 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluene sulphonate (CMCT) (Weeks, 2010). Subsequent reverse transcription detects the modification sites as stops to primer extension at nucleotide resolution. The resulting complementary DNA fragments are resolved in sequencing gels followed by individually quantifying band intensities. Prior to the mid-2000s, the bottleneck was the nucleotide resolution. The resulting complementary DNA fragments are expected interaction patterns across the residues and traces. On the basis of the chemical probing method used, we derive a prediction matrix that stores expected interaction patterns across the residues and traces. On the basis of the aligned CE traces and prediction matrix, we apply a dynamic-programming approach that finds the optimal selection of the band locations under a well-defined scoring scheme.

2 Methods

2.1 Problem definition

Given an RNA sequence $s$ probed at $N$ nucleotides, assume that we carry out the chemical structure probing of this sequence using $M$ different treatments, each of which is run in a separate capillary lane. Assume that the fluorescence intensity of each capillary is measured over $K$ time points. We define a profile (also called a trace) as the sequence of intensity values from a capillary. For any particular profile, the reactivity of each nucleotide to the chemical reagent is represented at a specific location in the series of intensity values, and $N$ such locations are sequentially spread throughout the entire profile. All profiles are assumed to be well aligned using the procedure described in Yoon (2011), such that each nucleotide corresponds to the same location across all profiles. The entire CE measurement can then be arranged in a $K \times M$ matrix $D$. Normally, $N \ll K$, i.e. each electrophoretic profile is finely sampled in time. On the basis of the characteristic of the chemical agent used in each treatment and the secondary structure computationally inferred from the input sequence, we can predict the fluorescence intensity at each position of $s$ for each of $M$ treatments. This prediction can be arranged in an $N \times M$ matrix $P$ called the prediction matrix (see below).

The problem of band annotation is therefore formulated as selecting $N$ out of the $K$ rows of $D$ using the information in $P$ in such a way that a certain objective is optimized over all possible $\binom{K}{N}$ possibilities. The selected $N$ points map to the locations of the nucleotides of the sequence $s$ in the CE measurement (see Supplementary Fig. S1).
The input of the proposed method consists of the following:

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

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

### 2.2 Prediction matrix construction

Figure 2a defines the expected reactivity of each type of nucleotide to chemical reagents used for chemical probing under the (un)paired condition. The value of 1 means that the nucleotide is reactive to the reagent, whereas 0 indicates no reactivity. For instance, the DMS chemical modifies A and C but not U and G, and the entries for A and C are one, while those for U and G are zero. We allow the use of numerous chemical probing strategies: DMS alkylation, CMCT and ‘others’ that can produce bands at all locations, including 2′-OH acylation (the SHAPE strategy) (Kladwang et al., 2014). We also allow input of a secondary structure in dot-parentheses notation. Nucleotides forming base pairs are not expected to show bands in DMS, CMCT, SHAPE and other structure mapping profiles. Sequencing experiments that terminate reverse transcription of the RNA with ddNTP incorporation produce bands after nucleotides complementary to the terminating nucleotide. On the basis of this information, we construct the prediction matrix $P$ that stores the expected chemical reactivity for individual residues. The element $p_{ij} \in P$ indicates such reactivity information of residue $i$ to reagent $j$.

Figure 2b shows an example RNA sequence with its secondary structure. Figure 2c shows the corresponding prediction matrix $P$.

### 2.3 Initialization of candidate peaks from profiles

The first step is to locate prominent peaks on each profile (each column of $D$). Peaks in CE profiles are the locations where significant reactivities are observed, implying that bands are more likely to exist at the same position. Thus, these peaks are matched with bands afterward. (Here and below, ‘peak’ refers to a local maximum in the time series; since the time series given are discrete, the second derivative of time series is estimated as below:

$$ \Gamma = \Delta^- - \Delta^+ $$

where

$$ \Delta^- = \max(d_i(j) - d_{i}(j-1), (d_{i}(j) - d_{i}(j+2)) / 2) $$

$$ \Delta^+ = \min(d_{i}(j+1) - d_{i}(j), (d_{i}(j+2) - d_{i}(j)) / 2) $$

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

2. In preparation for the sampling scheme and score function computation below, estimate the ideal separation between bands based on the remaining peak locations: $\rho = \min(k'_{j} - \min(k'_{j})(N-1) - j)$, where $k'_{j}$ and $k'_{j}$ are the locations of the foremost peak and the rearmost peak respectively on the $j$th profile.

3. In preparation for the score function computation below, construct a matrix based on these candidate peak locations called the bonus matrix $B \in \mathbb{Z}_N^M$. Let $\Gamma$ be the mean value of $\Gamma$ of the candidate peaks. Initialize $B$ to all zero. At each peak $A'_j$, we apply a uniform bonus, supplemented by a stronger bonus at sharp peaks: $B(A'_j) = \Gamma / 2 + \Gamma$.  

### 2.4 Formulation as dynamic programming

#### 2.4.1 Basic motivation

In essence, the band annotation problem is to select $N$ out of $K$ points and match them to peak locations (if at all possible) in an optimal way. This is similar to the problem of aligning two sequences ($1, 2, \ldots, N$) and ($1, 2, \ldots, K$) without allowing gaps for the latter.

**RNA sequence index:** 1--2--3--3...N...
**Measurement index:** 123456789...K

In the example above, the first three bands are located at 2, 5 and 9 time units. To find the most probable one among all such alignments, each possible alignment is given a score that represents its likelihood. Dynamic programming can be utilized to find the
solution set with the highest score, which in turn leads to the most likely locations of bands. More formally, define a matrix $F$ indexed by $n$ and $k$ $(1 \leq n \leq N; 1 \leq k \leq K)$ where the value $F(n, k)$ indicates the maximum score up to the band $n$ and position $k$. (More details on $F$ are given below.) The matrix $F$ is filled up recursively:

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

(2)

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

2.4.2 Degeneracy breaking and primary profile

In the proposed method, a band is allowed to be matched to a candidate peak even if their positions are slightly off from each other; in other words, an exact positional coincidence is not required for a peak-band matching (see Section 2.5.2 for detail). Thus, the formalization of our problem in the previous section allows for an undesired scenario in which two different bands will be matched to the same closest peak (see Supplementary Fig. S2); this can be problematic especially near strong peaks [high $F$ in (1)]. To avoid such degeneracies, an additional search variable $p$ is introduced: the relative position of the matched peak to the band position $k$. The tuple $(n, k, p)$ corresponds to the instance in which the band $n$ is located at position $k$ and matched with the peak at $k + p$ if there is any; there is no score bonus if there is no peak at the position. The matrix $F$ is now redefined as a 3-dimensional matrix as follows:

$$F(n, k, p) = \max_{k-2 \leq k' < k} \left( F(n-1, k', p') + S(n, k', k, p) \right)$$

(3)

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

2.4.3 Backtracking

The backtracking matrices $L_k, L_p$ for finding the solution itself are given by

$$L(n, k, p) = \{L_k(n, k, p), L_p(n, k, p)\}$$

$$= \arg \max_{k-2 \leq k' < k, |p| < \rho/2, k' + p' < k + p} \left( F(n - 1, k', p') + S(n, k', k, p) \right)$$

(4)

and, respectively, store the position $k'$ and the relative peak location $p'$ from which $F(n, k, p)$ is derived as in (3). The output array $y$ is derived from $L_k$ and $L_p$ as follows:

$$y(n) = \left\{ \begin{array}{ll}
y_k(n), & \text{if } n = N; \\
y_p(n), & \text{if } 1 \leq n \leq N - 1.
\end{array} \right.$$

(5)

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

2.5 Description of score term

The score term in (3) consists of the following two components:

$$S(n, k', k) = S_{\text{int}}(n, k - k') + w_{\text{peak}} \cdot s_{\text{peak}}(k, p) \cdot P(n, \cdot)$$

(6)

where $S_{\text{int}}$ and $s_{\text{peak}}$ are functions returning vectors of nonnegative elements, and $P(n, \cdot)$ is the $n$th row of the prediction matrix $P$. The dot product in the second term is a sum over all lanes $m$ from 1 to $M$. A coefficient $w_{\text{peak}}$ of 1.0 gave acceptable annotations in initial tests and was not further optimized.
2.5.1 Distance bonus term
It is empirically supported that the length between consecutive locations, \(k'\) and \(k\), is quite evenly distributed. \(S_{\text{dist}}\) is the bonus term that utilizes this fact and induces the dynamic programming to end up with regularly stretched output. In addition, observations on reference annotations suggest that a gap between two consecutive locations tends to be shorter when the preceding location corresponds to ‘G’ in the RNA sequence (Mills and Kramer, 1979; Sasaki et al., 1998). These observations lead to the definition of distance bonus term as follows:

\[
S_{\text{dist}}(n, d) = \frac{f_{\mu, \sigma}(d)}{f_{\mu, \sigma}(\rho')}
\]

(7)

where

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

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

2.5.2 Peak bonus term
The second score term favors band locations near peaks of the electrophoretic profiles with a significant curvature. \(S_{\text{peak}}\) is a function that returns a nonnegative \(M\)-dimensional value, where each of its entries represents the peak bonus from each profile:

\[
S_{\text{peak}}(k, p) = (S^1_{\text{peak}}(k), \ldots, S^{M-1}_{\text{peak}}(k), S^M_{\text{peak}}(k, p))
\]

(8)

where \(S^m_{\text{peak}}\) stands for the bonus from matching a peak to a band in \(\delta_m\), assuming such a band exists. The bonus was designed to be boosted for a greater curvature at the peak and the proximity of the peak to the bands reactive to chemical agents for each profile. (3) and candidate peaks in other profiles (9). A large number of such failed matches flag an unreliable band annotation, as described next.

2.6 Reliability evaluation
Although the presented band annotation method was found to be quite accurate, it was not perfect. We therefore sought a method to assess the reliability of automatically determined band locations prior to practical application. We devised a score to predict the quality of results. The idea behind the score is that when optimization of Equation (6) fails to achieve the desirable solution, we typically see extraordinarily short or long distances between consecutive locations (little information from \(S_{\text{dist}}\)) or bands on the primary profile without proper matching to peaks (little information from \(S^M_{\text{peak}}\)). The E-score is defined with the following terms:

- \(n_1\): number of bands on the primary profile without corresponding peak.
- \(n_2\): number of gaps with length < \(\rho/4\) or > \(2\rho\).
- \(M^p\): number of bands on the primary profile predicted by \(P\).

\[E = 1 - \max\left(\frac{n_1}{n_1 + n_2}, \frac{M^p}{M}\right)\]

E-score is a value between 0 and 1 and conservatively estimates the fraction of well-annotated bands in the output. The relationship between E-score and accuracy is presented in Section 3.

3 Results
3.1 Robust determination of band positions
Figure 4a–c shows the electrophoretic profiles annotated with band locations by three different methods: reference, proposed and QuShape (Karabiber et al., 2013), respectively. The reference annotation was based on expert assignments carried out at the time of data acquisition (Lee et al., 2014). QuShape was chosen as the comparison target for its superior accuracy in band annotation relative to other software we tested, FAST and ShapeFinder (data not shown); no-modification and ddTTP ladder profiles were used as references (RXS1, BGS1) while running QuShape. Visual inspection suggests that the proposed method produces annotations more compatible with the reference. In this profile, the annotation determined by QuShape deviates from the reference position, particularly near the beginning of the sequence.

To generally and quantitatively assess the accuracy of automated band annotation, we applied the proposed method and QuShape to 95 datasets acquired in the EteRNA project (Table 1). For both methods, we computed the mean squared error (MSE) of the band annotation.
locations determined by the proposed method with respect to the reference locations, in units of average distance between locations. For a sense of scale, the typical MSE achieved by expert annotation is 0.15, based on comparisons of different experts’ annotations with reference locations, in units of average distance between locations.

For a sense of scale, the typical MSE achieved by expert annotation is 0.15, based on comparisons of different experts’ annotations with reference locations, in units of average distance between locations.

Table 1. High-throughput RNA structure mapping datasets analyzed by the proposed method (total 522 profiles and 47,210 bands)

<table>
<thead>
<tr>
<th>Name</th>
<th>No. profiles</th>
<th>No. nt</th>
<th>No. bands per profile</th>
<th>No. total bands</th>
</tr>
</thead>
<tbody>
<tr>
<td>R48</td>
<td>60</td>
<td>108</td>
<td>88</td>
<td>5280</td>
</tr>
<tr>
<td>R46</td>
<td>80</td>
<td>108</td>
<td>88</td>
<td>7040</td>
</tr>
<tr>
<td>R47</td>
<td>90</td>
<td>112</td>
<td>92</td>
<td>8280</td>
</tr>
<tr>
<td>R47b</td>
<td>36</td>
<td>112</td>
<td>92</td>
<td>3312</td>
</tr>
<tr>
<td>R48</td>
<td>96</td>
<td>112</td>
<td>92</td>
<td>8832</td>
</tr>
<tr>
<td>R49</td>
<td>18</td>
<td>112</td>
<td>92</td>
<td>1656</td>
</tr>
<tr>
<td>R49B</td>
<td>48</td>
<td>115</td>
<td>95</td>
<td>4560</td>
</tr>
<tr>
<td>R50</td>
<td>54</td>
<td>115</td>
<td>95</td>
<td>5130</td>
</tr>
<tr>
<td>R43</td>
<td>40</td>
<td>98</td>
<td>78</td>
<td>3120</td>
</tr>
</tbody>
</table>

Excluding the last line, there are 95 datasets. More details of these 95 datasets are described in Lee et al. (2014). FMN, flavin mononucleotide.

Further tests confirmed the utility of using multiple profiles, secondary structure information and peak match degeneracy breaking in producing accurate band annotations (Supplementary Fig. S3).

Fig. 5. Proposed method (left) versus QuShape (right). Each plot represents each metric’s distribution across 95 datasets. (A) MSE for band locations. (B) Pearson’s correlation coefficient r for band locations. (C) KL divergence for band locations. (D) Pearson’s correlation coefficient r for area quantification. MSE units are normalized, so that average distance between band locations is unity.

3.2 Accurate peak-area quantification

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

As one example, Figure 6a and b shows the correlation of results between the proposed method and reference for a specific dataset (flavin mononucleotide binding branches) for two chemical modification strategies (SHAPE and DMS). Figure 6c and d shows the correlation between the QuShape and reference results, which is visually worse than the proposed method in both cases. Over all the datasets, Figure 5d and Supplementary Table S1 gives the distribution of the Pearson’s correlation coefficients. The median correlation coefficient for the proposed method is 0.976, which is higher than that for QuShape (0.939) and the distribution for the proposed method shows smaller variance. This observation suggests that using the proposed band annotation can significantly enhance the accuracy of band quantification.

3.3 E-score reliability metric predicts MSE accuracy

In Section 2.6, we proposed E-score to evaluate the quality of results from our method. We assessed the use of E-score based on its ability to predict the accuracy of the band annotations compared with gold standard annotations, quantitatively evaluated as MSE. Figure 7a shows the distribution of MSE for results satisfying $E = 1.0$; these values are substantially smaller than those in Figure 7b, which includes all 95 datasets. For example, all 26 results under constraint $E = 1.0$ have MSE below 0.5 as shown in Figure 7a, confirming that a
gram of the (worse) mean and standard deviation. Figure 7 shows the histogram of MSE values (0.83 and 0.90) coincide in the results for RNA puzzle 6 (an adenosylcobalamin riboswitch) and rRNA(phe), confirming E-score’s utility.

4 Discussion

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

To flag cases with uncertain automated band annotation, we have introduced the E-score for reliability estimation. According to our results, given any dataset for CE analysis, the band annotations with E > 0.97 are almost always reliable and can be safely adopted for final steps of band quantitation, whereas the results with E ≤ 0.97 are less likely to reliable. Informally, we have encountered datasets in which even expert annotation is ambiguous and has required special additional experiments (such as co-loading sequencing ladders in the same color as the sample) to resolve (Tian et al., 2014). This suggests that automated band annotation cannot improve much further; a valuable development would be reliability estimates for specific subsets of bands rather than a global number. An additional useful development would be use of known band intensities based on prior experiments (Karabiber et al., 2013) or on base pair probability estimates, rather than coarse predictions for profiles based on sequence, modifier and a single secondary structure.

The proposed algorithm has order of NK time and space complexity, and the practical time demand of band annotation was reasonable in our experiments. The proposed method was implemented in the MATLAB programming environment (The MathWorks, http://www.mathworks.com), and under the experimental setup used (sequential execution on a Intel core i5 4570 processor with 8-GB main memory), the total time demand of annotating bands in all the 95 datasets did not exceed 4 min (for each dataset, mean 2.2837 s; median 2.2707 s).

5 Conclusion

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

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

**Funding**
This work was supported in part by the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (grant no. 2011-0009963 and no. 2014M3C9A3063541 to S.Y.) and in part by a Burroughs-Wellcome Foundation Career Award at the Scientific Interface and the U.S. National Institutes of Health (NIGMS R01GM100953 to R.D. for computational work).

**Conflict of Interest**: none declared.

**References**


