Score statistics of global sequence alignment from the energy distribution of a modified directed polymer and directed percolation problem

Mihaela E. Sardiu,[1,2] Gelio Alves,[1,2] and Yi-Kuo Yu[2]
[1]*Department of Physics, Florida Atlantic University, Boca Raton, Florida 33431, USA*
[2]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA*

(Received 3 June 2005; revised manuscript received 21 October 2005; published 23 December 2005)

Sequence alignment is one of the most important bioinformatics tools for modern molecular biology. The statistical characterization of gapped alignment scores has been a long-standing problem in sequence alignment research. Using a variant of the directed path in random media model, we investigate the score statistics of global sequence alignment taking into account, in particular, the compositional bias of the sequences compared. Such statistics are used to distinguish accidental similarity due to compositional similarity from biologically significant similarity. To accommodate the compositional bias, we introduce an extra parameter p indicating the probability for positive matching scores to occur. When p is small, a high scoring alignment obviously cannot come from compositional similarity. When p is large, the highest scoring point within a global alignment tends to be close to the end of both sequences, in which case we say the system percolates. By applying finite-size scaling theory on percolating probability functions of various sizes (sequence lengths), the critical p at infinite size is obtained. For alignment of length t, the fact that the score fluctuation grows as $\chi t^{1/3}$ is confirmed upon investigating the scaling form of the alignment score. Using the Kolmogorov-Smirnov statistics test, we show that the random variable χ, if properly scaled, follows the Tracy-Widom distributions: Gaussian orthogonal ensemble for p slightly larger than p_c and Gaussian unitary ensemble for larger p. Although these results deepen our understanding of the distribution of alignment scores, the use of these results in practical applications remains somewhat heuristic and needs to be further developed. Nevertheless, the possibility of characterizing score statistics for modest system size (sequence lengths), via proper reparametrization of alignment scores, is illustrated.

DOI: 10.1103/PhysRevE.72.061917 PACS number(s): 87.10.+e, 05.40.−a, 02.50.−r, 02.50.Fz

I. INTRODUCTION

Seemingly different subjects may share similar underlying mathematical structure. Realization and investigation of such an occurrence often lead to progress in one or both subjects. For example, the recognition and elaboration of the close relationship between the directed polymers/paths in random media (DPRM) problem [1–4] in statistical physics and the score statistics of sequence alignment in bioinformatics has led to fruitful results [5–14]. Recently, it was found [15] that compositional divergence between two related sequences may hinder the detection of their homology. Fortunately, for both global alignment and local alignment this problem can be solved quite effectively [15,16] through deriving a self-consistent scoring scheme. On the other hand, compositional similarity among unrelated sequences may lead to accidental similarity identifications [17]. This will contaminate the statistics and may weaken the efficacy of iterative database search methods such as PSI-BLAST [18]. A rationale was proposed and implemented [17] to improve the accuracy of local alignment score statistics. In this paper we study the corresponding problem in global alignment statistics via a different route: by studying a variant DPRM model that is also related to the percolation problem.

The DPRM problem is one of the best studied systems with quenched disorder. In a $d+1$ dimensional DPRM system, there are d regular spatial dimensions and one timelike dimension that is singled out to specify the elongated direction of the path. The displacement made by the DP, when projected onto the timelike direction, is often identified as the *length* of the DP. Due to the presence of the quenched disorder, the system's free energy depends on the particular realization of the disorder involved. And it is the probability distribution function (pdf) of the free energy that characterizes the statistical properties of the system.

Sequence alignment, on the other hand, is one of the most powerful tools in modern molecular biology. Computer-assisted sequence alignment has become increasingly important due to the rapid growth of DNA and protein databases. The use of sequence alignment ranges from identifying the possible functionality of newly sequenced DNA and protein to the construction of phylogenetic trees [19–21]. Under sequence alignment, the relatedness of two sequences compared is quantified by an alignment score and its associated E value. The latter is the expected number of random hits with the same or even higher score from a given database, and thus provides a meaningful measure of homology detected. Dividing the E value by the database size, one obtains a database independent measure.

Unfortunately, rigorous results relating such database-independent measures to alignment parameters (or scoring function) exist only for gapless alignment, which is less sensitive in detecting distant homology. Concerted efforts [7–14] utilizing the connection to DPRM have been made to better characterize the score statistics of gapped local alignment, a popular tool to find between two sequences the most

homologous segments, one from each sequence. The quantification of the score statistics of local alignment is also approached by employing a more effective sampling method [22], and by computing one of the Gumbel parameters [23] based on a special scoring scheme where the match score is a constant and the mismatch score is twice the gap cost [24].

Despite its straightforward relation to the 1+1 dimensional DPRM problem, the score statistics of gapped global alignment have not yet attracted appropriate attention until recently. When cast in the language of DPRM, global alignment score statistics may be regarded as the (free) energy distribution of the 1+1 dimensional DPRM system with the alignment score identified as the negative of the (free) energy. In 1+1 dimensions, the free energy (or score) of a directed path of length t traversing a random potential energy landscape is known [2] to have free energy

$$F(t) \sim -vt - \chi t^{1/3} \tag{1}$$

[or $S(t) \sim vt + \chi t^{1/3}$] with $-v$ being the average free energy per unit length and χ being a random variable. Although the exponent 1/3 has been known for several decades [25], the distribution of the random variable χ was determined only recently.

By mapping the so-called polynuclear growth (PNG) model to the strong coupling regime (zero temperature limit) of the DPRM and to the longest increasing subsequence (LIS) problem, Prähofer and Spohn [26] pointed out that the probability distribution of the length ℓ from the LIS can be used to characterize the pdf of χ in the DRPM problem. In fact, there are three different subclasses [26–29], each with a different pdf(χ) resulting from different boundary and initial conditions, that all exhibit the same free energy exponent 1/3. This more detailed information will allow for better statistical characterization of gapped global alignment scores of uncorrelated random sequence pairs, once one identifies correctly the corresponding boundary condition in PNG for the version of the DPRM mapped from the sequence alignment problem.

Assuming that each of the c characters in an alphabet occurs with equal probability $1/c$ and by combining several existing results [26,27,30,31] and a change of coordinates, Majumdar and Nechaev [32] obtained the asymptotic χ distribution for the longest common subsequence (LCS) between two character sequences compared. The pdf of χ found in Ref. [32] is one the subclasses obtainable from PNG with a specific boundary condition. The LCS can be interpreted as the global alignment score under a special scoring scheme where the matching score for two identical characters is 1 and the mismatched character pair has a score twice the gap penalty. Generalization to characters with unequal background frequencies, a more general scoring scheme, and accommodation of compositional bias, nevertheless, remains a challenging task and deserves a detailed study.

In this study, we do not impose constraints between gap costs and substitution score. We investigate whether or not other subclasses of pdf of χ obtainable from PNG can be realized in our model of gapped global alignment. To mimic different levels of the compositional bias, we further restrict the proportion of occurrences of a favorable (negative) random potential to be p (see next section for more detail). When the two sequences compared have very dissimilar compositions, the chance for a large alignment score is much smaller than for the case when both sequences have similar compositions. A large p mimics high compositional similarity. When p is small, it is expected that v becomes negative and the alignment score will decrease with length. The scaling behavior of the score near the transition from $v<0$ to $v>0$ is also verified and studied. The critical p, p_c, is defined as the p value that gives rise to $v=0$. Note, however, that this p_c is expected to be very different from the p_c in the standard directed percolation problem [33,34]. This is because, upon increasing p from zero, the condition $v=0$ occurs much earlier before the energetically favorable bonds percolate.

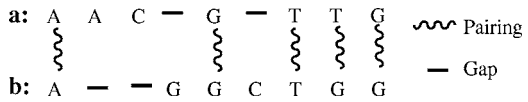

FIG. 1. An example of global alignment between sequences **a** and **b**.

This paper is organized as follows. In the next section, we define the parameter p and establish notation. In the third section, we shall review some previous works that are related to the current study. In the fourth section, the numerics and the associated analysis will be detailed. In particular, the scaling of average score, dependence of p_c on the gap penalty, and a general analysis of the energy distribution will be presented. Although some details will be shown in this section, technical details will be relegated to the Appendix. A summary and some concluding remarks constitute the last section.

II. GLOBAL ALIGNMENT AND THE DPRM

In this section, we will introduce the parameter p to accommodate compositional bias, give a brief introduction to global alignment and the algorithm for global alignment, and finally describe the variant DPRM model that we abstract from the alignment algorithm. For simplicity, throughout this study, we use only a linear gap cost (see below).

Sequence alignment can be used to identify homology between protein or DNA sequences. An alignment between two sequences **a** and **b** is given in Fig. 1. In this particular example, both sequences contain seven characters. **a**=[*AACGTTG*] while **b**=[*AGGCTGG*]. We will use the notation a_i (b_j) to refer to the ith (jth) character of sequence **a** (**b**). Thus a_3 is C, b_5 is T, etc.

The quality of an alignment is usually quantified by the associated alignment score, which is the sum of pairwise substitution scores $s(a_i,b_j)$ and gap penalties $\gamma(i_0,i_f|j_0,j_f)$. Here $s(a_i,b_j)$ denotes the pairwise substitution score when we pair up character a_i from sequence **a** with the character b_j from sequence **b**. Because of its dependence on two characters (indices), a set of substitution scores is often called a *substitution matrix*. A gap is formed when a character from one sequence is not paired with any character from the other sequence, and the function $\gamma(i_0,i_f|j_0,j_f)$ returns the gap penalty when the substrings (of consecutive characters)

$[a_{i_0+1},\ldots,a_{i_f}]$ and $[b_{j_0+1},\ldots,b_{j_f}]$ are not paired with characters from their respective countersequences. Apparently, the case $i_0=i_f$ (or $j_0=j_f$) indicates that the substring $[a_{i_0+1},\ldots,a_{i_f}]$ (or $[b_{j_0+1},\ldots,b_{j_f}]$) contains no characters.

It is a common practice to use the term *scoring function* to denote the combination of the substitution matrix and the gap penalty function used for sequence alignment. Under a given scoring function, the associated alignment score of the example in Fig. 1 will be $s(A,A)-\gamma(1,3|1,2)+s(G,G)-\gamma(4,4|3,4)+s(T,T)+s(T,G)+s(G,G)$, which consists of five pairwise substitution scores and two gap penalties. Although there are many possible alignments, corresponding to different arrangements of gaps and substitutions, between two sequences, one usually refers to the alignment with highest alignment score as the *optimal alignment* and its associated score as the *alignment score*. The alignment example above is termed *global alignment* since the two sequences (**a** and **b**) are aligned from head to toe.

The scoring function used in sequence alignment is often designed by experienced biologists. Different substitution matrices are designed for capturing different types of similarity (or evolutionary distances). The gap penalty function can have many variants. In this study, however, we will only focus on the linear gap function:

$$\gamma(i_0,i_f|j_0,j_f)=[i_f+j_f-i_0-j_0]\delta. \tag{2}$$

The parameter δ is the penalty per unmatched character (gap).

In the following, we will define the parameter p that accommodates the presence of compositional bias, and describe briefly the Needleman-Wunsch [35] algorithm for global alignment and the corresponding DPRM problem. Because the PNG model maps to the zero temperature DPRM problem, we will describe only the optimal alignment algorithm and zero temperature DPRM. The generalization to finite temperature is straightforward and can be found in Ref. [13].

A. Compositional bias and score statistics

The null model assumed in the study of alignment score statistics plays an important role in statistical significance assessment of similarity found between two sequences. One of the most frequently employed null models is the one point random Markov sequence model. Here a sequence $\mathbf{a}=[a_1,a_2,\ldots,a_M]$, with its constituent letters a_i drawn from an alphabet Ω, is assumed to exist with probability

$$P_0(\mathbf{a})=\prod_{i=1}^{M} f(a_i), \tag{3}$$

where $f(a)$ is the background frequency of character a and $\Sigma_{a\in\Omega}f(a)=1$.

The basic idea here is to use the *standard composition* $\{f(a)\}$ as an estimate of the typical character frequencies in a typical sequence. A useful quantity to consider is the average substitution score in a null model,

$$\langle s\rangle_0=\sum_{a,b\in\Omega} s(a,b)f(a)f(b).$$

In order to suppress noise, it is a common practice to require that $\langle s\rangle_0<0$. As we will elaborate later, another useful parameter is the ratio

$$p_{f,f}\equiv\frac{\sum_{a,b}^{+} s(a,b)f(a)f(b)}{\sum_{a,b}^{+} s(a,b)f(a)f(b)+\sum_{a,b}^{-}|s(a,b)|f(a)f(b)}, \tag{4}$$

where $\Sigma^{+(-)}$ means only sum over entries with positive (negative) $s(a,b)$. Consider generating infinitely many letter pairs according to $\{f(a)\}$. The numerator in Eq. (4) can be regarded as the area of the histogram with positive scores, while the denominator can be regarded as the total area of the histograms of all scores. In a coarse-grained view, one may regard the positive score as uniformly distributed between zero and one with probability of occurrence p and the negative score as uniformly distributed between -1 and 0 with probability $1-p$. Aside from its analog to sequence alignment problem, a potential distribution of this sort, albeit in discrete form, has been been used [33,34] in studying the relation between DPRM and directed percolation. Apparently, for $\langle s\rangle_0<0$, one must have $p<1/2$. However, it is possible that two sequences compared are of very similar composition and/or share the same rare amino acids so that the corresponding p can be larger than $1/2$. We therefore allow p in the range $0<p<1$ to accommodate all possibilities.

A compositional bias among two sequences (or from standard composition) occurs when the sequences compared exhibit character compositions significantly different from one another [or from $\{f(a)\}$]. Let $\mathbf{a}=[a_1,a_2,\ldots]$ and $\mathbf{b}=[b_1,b_2,\ldots]$ be two character sequences with their characters a_i and b_j taken from Ω. Let us call $C_{\mathbf{a}(\mathbf{b})}(a)$ the composition frequency of character a in sequence $\mathbf{a}(\mathbf{b})$. In a similar fashion to Eq. (4), one may define a corresponding ratio for the sequence pair **a** and **b**,

$$p_{\mathbf{a},\mathbf{b}}=\frac{\sum_{a,b}^{+} s(a,b)C_{\mathbf{a}}(a)C_{\mathbf{b}}(b)}{\sum_{a,b}^{+} s(a,b)C_{\mathbf{a}}(a)C_{\mathbf{b}}(b)+\sum_{a,b}^{-}|s(a,b)|C_{\mathbf{a}}(a)C_{\mathbf{b}}(b)}.$$

If $p_{\mathbf{a},\mathbf{b}}>p_{f,f}$ ($p_{\mathbf{a},\mathbf{b}}<p_{f,f}$), then the alignment score between two unrelated sequences $(\mathbf{a},\mathbf{b})$ will on average have higher (lower) score than expected from a background model assuming a standard composition $\{f(a)\}$. As one possible way to accommodate such biases, we propose that the background score statistics be studied under different values of p.

B. Algorithms

Let $\mathbf{a}=[a_1,a_2,\ldots,a_M]$ and $\mathbf{b}=[b_1,b_2,\ldots,b_N]$ be two sequences of lengths M and N, respectively, with elements a_i and b_j taken from the alphabet Ω. Under a given scoring function, i.e., a substitution matrix and a gap function, the Needleman-Wunsch algorithm [35] optimizes global alignment by using the dynamic programming method (or the transfer matrix method in statistical physics).

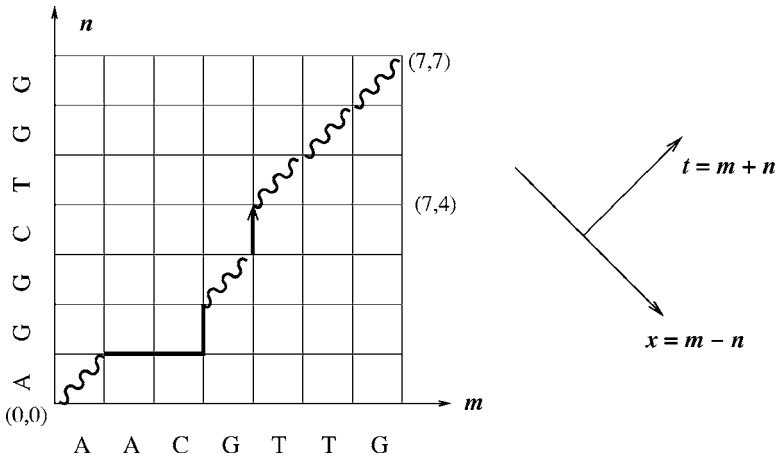

FIG. 2. The alignment lattice. Upon laying sequence **a** along the horizontal axes and sequence **b** along the vertical axes, we note that the directed path here uniquely represents the alignment shown in Fig. 1. The new coordinate system ($x=m-n$, $t=m+n$) is also shown to illustrate the connection between the recursion relation (5) used in sequence alignment and the corresponding one (7) used in DPRM.

For clarity, we introduce the alignment lattice in Fig. 2 with sequence **a** laid along the x direction and sequence **b** laid along the y direction. Note that the alignment example given in Fig. 1 is shown as a (directed) path in the alignment lattice. In fact, each alignment is represented by a unique path and vice versa.

Define the auxiliary quantity $S_{m,n}$ that records the highest global alignment score for alignment paths starting at the origin (0, 0) and terminating at point (m,n). It is not hard to see that for the linear gap case the auxiliary quantity $S_{m,n}$ obeys the following recursion relation:

$$S_{m,n} = \max\begin{Bmatrix} S_{m-1,n-1} + s(a_m,b_n) \\ S_{m-1,n} - \delta,\ S_{m,n-1} - \delta \end{Bmatrix}, \tag{5}$$

with the "boundary conditions"

$$S_{0,n\geqslant 0} = -n\delta \quad \text{and} \quad S_{m\geqslant 0,0} = -m\delta. \tag{6}$$

The alignment score is typically identified as $S_{M,N}$ and the associated optimal alignment is obtained by the *trace-back* method [19]. However, in this study we will use a different definition of score (see below).

C. Variant DPRM

The recursion (5), in fact, is a commonly used approach in statistical physics, i.e., the transfer matrix method. In particular, it is very similar to the transfer matrix used to tackle the zero temperature DPRM problem in 1+1 dimensions. For a detailed review of the DPRM problem, readers are referred to Ref. [4] and references therein. In a 1+1 dimensional DPRM system, each lattice point is labeled by two discrete indices, x and t for space and time, respectively.

To illustrate the connection between DPRM and the sequence alignment problem, we focus on the following variant of DPRM. Using the coordinates defined by $x=m-n$, $t=m+n$, as shown in Fig. 2, a directed path $\mathcal{A}$ starting from the origin ($x=0$, $t=0$) can be regarded as the "world line" of a particle in one dimension. For a given realization of randomness, a random potential $u(x,t)$ is assigned to the bond connecting lattice points $(x,t+1)$ and $(x,t-1)$. There is also a *constant* elastic penalty associated with each bending of the path, e.g., going from (x,t) to $(x-1,t+1)$ instead of to $(x,t+2)$.

At zero temperature, the free energy is the energy of the lowest energy path. Writing the elastic energy as $\bar{\delta}$, one can write down easily the transfer matrix for finding the lowest energy path and its associated energy via the following recursion:

$$E(x,t) = \min\begin{Bmatrix} E(x,t-2) + u(x,t-1) \\ E(x-1,t-1) + \bar{\delta} \\ E(x+1,t-1) + \bar{\delta} \end{Bmatrix}. \tag{7}$$

The lowest energy at time T is then given by

$$\min_x E(x,T). \tag{8}$$

Using $x=m-n$ and $t=m+n$, we can rewrite the recursion (5) in terms of x and t

$$S(x,t) = \max\begin{Bmatrix} S(x,t-2) + s(x,t-1) \\ S(x-1,t-1) - \delta \\ S(x+1,t-1) - \delta \end{Bmatrix} \tag{9}$$

with $s(a_m,b_n)$ rewritten as $s(x,t-1)$. The reason we did not write $s(a_m,b_n)$ as $s(x,t)$ comes from the observation that the letter pair (a_m,b_n) is located at $(m-1/2,\ n-1/2)$, not (m,n), on the alignment lattice (see Fig. 2). Note that if one defines $S(x,t) \equiv -E(x,t)$, then the above recursion is turned into

$$E(x,t) = \min\begin{Bmatrix} E(x,t-2) - s(x,t-1) \\ E(x-1,t-1) + \delta \\ E(x+1,t-1) + \delta \end{Bmatrix}. \tag{10}$$

Therefore the negative of the substitution score plays the role of the potential and the gap cost plays the role of elastic energy.

In our variant model, we take the coarse view from the first subsection and assume that the random potential $u(x,t)$ [or $-s(a_m,b_n)$] is uncorrelated [36] and follows the form

$$u(x,t) = \begin{cases} [-1,0), & \text{with probability } p \\ [0,1), & \text{with probability } (1-p) \end{cases}. \tag{11}$$

Instead of taking the score at the upper-right corner of the lattice in Fig. 2, we look for the maximum score within the alignment lattice,

$$\max_{x,t}\{S(x,t)\} \quad \text{or} \quad \max_{m,n}\{S(a_m,b_n)\}.$$

For a given square lattice of size L, we call the system percolated when the lowest-energy (highest-score) point occurs within $\sqrt{L}$ from the top wall or the right wall, i.e., when the max score point has either its m or n coordinate in the range $[L-\sqrt{L},L]$ [37].

III. RELEVANT TECHNICAL BACKGROUND

Because the maximum height in PNG model can be understood as the LIS of a permutation of N numbers, statistical characterization of the latter can be applied to the former. It is known that in the limit of large N, the LIS of a given permutation has length $\ell_N=2\sqrt{N}$ with probability one [40]. Baik *et al.* [27] recently showed that the fluctuations with respect to the mean value $2\sqrt{N}$ are of the form $\tilde{\chi}N^{1/6}$ and that the distribution of the random variable $\tilde{\chi}$ is characterized by $\text{Prob}(\tilde{\chi}\leqslant x)=F_{\text{GUE}}(x)$, with $F_{\text{GUE}}(x)$ being the Tracy-Widom distribution [41] for the Gaussian Unitary Ensemble. The Tracy-Widom distribution $F_{\text{GUE}}(x)$, being the distribution of the largest eigenvalue of complex Hermitian matrices, is intimately related to the Painlevé II equation

$$u_{xx}=2u^3+xu. \tag{12}$$

This equation admits a globally positive solution with the following asymptotics:

$$u(x)\sim \text{Ai}(x) \quad \text{for } x\to\infty,$$

$$u(x)\sim \sqrt{\frac{-x}{2}} \quad \text{for } x\to-\infty,$$

where $\text{Ai}(x)$ is the airy function of first kind satisfying

$$u_{xx}-xu=0$$

and is related to the modified Bessel function via $\text{Ai}(x)=\sqrt{x/3\pi^2}K_{1/3}(\frac{2}{3}x^{3/2})$. The result of Baik *et al.* then states that

$$\lim_{N\to\infty}\text{Prob}\left(\tilde{\chi}\equiv\frac{\ell_N-2\sqrt{N}}{N^{1/6}}\leqslant x\right)=F_{\text{GUE}}(x) \quad \forall\, x\in R,$$

and the Tracy-Widom distribution $F_{\text{GUE}}(x)$ is related to $u(x)$ via

$$F_{\text{GUE}}(x)=\exp\left[-\int_x^\infty (s-x)u^2(s)ds\right]\equiv\exp[-g(x)],$$

with $g''(x)=u^2(x)$ and $g(x)\to 0$ as $x\to\infty$.

When translating this result into the context of PNG [26], the value N represents the number of nucleation events. Figure 3 illustrates a PNG (with eight nucelation events) and its associated permutation. If one assumes the nucleation events have uniform density, the number of nucleation events will follow a Poisson distribution, have mean $\langle N\rangle$ proportional to the lattice area $t^2/4$, and have a maximum height identified as ℓ_N. Because the maximum height corresponds to the negative of the energy of the lowest energy path of DPRM and

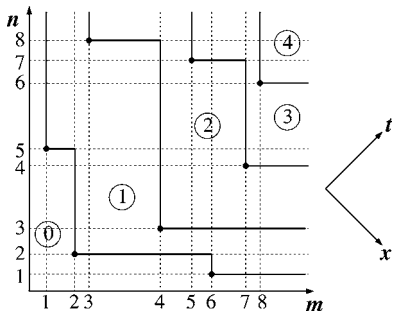

FIG. 3. An example of a polynuclear growth event and its corresponding permutation. A region of height h is labeled by h in a circle. A nucleation point i, labeled by its projection on the m axis, will have a different (permuted) label $\pi(i)$ on the n axis. Therefore the sequence [1,2,3,4,5,6,7,8] is permuted into $[\pi(1),\pi(2),\pi(3),\pi(4),\pi(5),\pi(6),\pi(7),\pi(8)]$ $=[5,2,8,3,7,1,4,6]$ with LIS [2,3,4,6] of length 4, the maximum height of the PNG profile.

because $t\propto N^{1/2}$, we know that the coefficient χ, with appropriate scale, also follows F_{GUE}, the Tracy-Widom distribution for complex Hermitian matrices.

A similar analysis can also be carried out for cases where a flat interface is used as a starting point. In this case, the mapping from PNG to permutation will result in an additional symmetry [26]. This additional symmetry, which was considered by Baik and Rains [28], leads to the prediction that the appropriately scaled random variable χ should follow $F_{\text{GOE}}(x)$, the Tracy-Widom distribution for the Gaussian orthogonal ensemble. As expected, the distribution $F_{\text{GOE}}(x)$ is closely related to $F_{\text{GUE}}(x)$:

$$F_{\text{GOE}}(x)=[F_{\text{GUE}}(x)]^{1/2}\exp[-f(x)/2] \tag{13}$$

with $f(x)=\int_x^\infty u(s)ds$.

Baik and Rains [29] further investigated the case where nucleation events can also happen on the boundary of the lattice. Denote the horizontal axes in Fig. 3 by E_+ and the vertical axes by E_-, one then calls $\alpha_{+(-)}$ the linear nucleation density along $E_{+(-)}$. Let α^2 denote the bulk nucleation event density, the results most revelant to the DPRM are $\alpha_\pm=\alpha$, for which

$$\lim_{t\to\infty}\text{Prob}\left(\tilde{\chi}\equiv\frac{L(t)-2t}{t^{1/3}}\leqslant x\right)=F_0(x), \tag{14}$$

with $L(t)$ identified as $h(\sqrt{2}t)$ in the PNG [26]. This new limiting distribution $F_0(x)$ due to Baik and Rains has not yet been identified with any eigenvalue distribution from random matrices [42]. It should be noted that $F_0(x)$ is also closely related to the Tracy-Widom distribution $F_{\text{GUE}}(x)$. In fact, one has

$$F_0(x)=[1-(x+2f''+2g'')g']\exp[-(g+2f)]. \tag{15}$$

As interpreted by Prähofer and Spohn [26], this distribution corresponds to stationary growth.

In the following section, we will analyze which of the three possible distribution functions—$F_{\mathrm{GUE}}(x)$, $F_{\mathrm{GOE}}(x)$, or $F_0(x)$—is the correct one for the random variable χ in the context of global alignment.

IV. NUMERICS AND ANALYSIS

For the practical use of global alignment, there is little use to keep aligning two sequences if the alignment score decreases with t. Therefore, in our variant model, we look for $\max_{m,n}\{S(a_m,b_n)\}$ instead of just taking the value of $S(a_M,b_N)$. The alignment score may steadily decrease with length aligned when severe compositional biases are present. In principle, such a situation may be alleviated by the compositional adjustment of the scoring matrix [15,16]. However, after the compositional adjustment of the scoring matrix, if the global alignment score still keeps decreasing with length, the two sequences compared cannot be homologous.

In our numerical studies, we first determine for every $L\times L$ lattice the critical concentration $p_c(L)$ of favorable potentials at which the highest score path percolates in the majority of simulations. Recall that the system is deemed to have a percolated path if the highest score point within the lattice has either its m or n coordinate within the range $[L-\sqrt{L},L]$ [37]. The finite size effect is studied in detail in order to extract p_c at inifinite size. However, note that the value of $p_c(L\to\infty)\equiv p_c$ does depend on the gap penalty δ used. Taking p close to p_c and studying the scaling of the ensemble-averaged score, we investigate the relation between v, the average score gain per length, and $|p-p_c|$. We also identify this relation's systematic dependence on δ. We then investigate the distribution of the random variable χ when $p>p_c$, followed by an attempt to characterize the global alignment score statistics at finite sizes. The following subsections are organized according to the order of our study.

A. From finite-size scaling to p_c

To determine $p_c(L\to\infty)$, we first determine $p_c(L)$ for finite L and then extrapolate by assuming that finite-size scaling holds. For an inifinite system near criticality, characterized by some intensive variable T near the critical value T_c, the only relevant length is the correlation length ξ that diverges as $g_{\pm}|T-T_c|^{-\nu}$ when approaching the critical point T_c from above $(+)$ or below $(-)$. Usually, the prefactors g_+ and g_- are not identical. The finite-size effect sets in when ξ grows to be comparable to the system size L. The observables, such as susceptibilities, that diverge as $\xi^{\alpha/\nu}$ can now only have maximum $L^{\alpha/\nu}$. Therefore the observables will have a finite-size scaling form $\xi^{\alpha/\nu}\Lambda_{\pm}(L^{1/\nu}/\xi^{1/\nu})\equiv|T-T_c|^{-\alpha}f_{\pm}(L^{1/\nu}|T-T_c|)$ with asymptotics

$$f_{\pm}(x>0)\simeq\begin{cases}1 & \text{if } x\gg 1\\ x^{\alpha} & \text{if } x\ll 1\end{cases}. \tag{16}$$

Further, because there cannot be a true singularity when the system size is finite, the physical observables must be continuous across T_c. This then requires $(T-T_c)^{-\alpha}f_+(x\to 0)=(T_c-T)^{-\alpha}f_-(x\to 0)$, which comes out of Eq. (16) naturally. One may then define a new scaling function $H(x)$ such that $H(x>0)=f_+(x)$, $H(x<0)=f_-(-x)$, and $H(x)$ is continuous at $x=0$.

For a percolating system, $\xi\sim g_{\pm}|p-p_c|^{-\nu}$ and the argument of the general scaling function H can be written as $L^{1/\nu}(p-p_c)$. Given a finite lattice of size $L\times L$, there is a finite chance, due to statistical fluctuations, for the system to percolate even when $p<p_c$. Similarly, there is also a finite chance for system to remain unpercolated even if $p>p_c$. We define the probability for system to percolate under a given p and L by $\Pi(p,L)$. While $\Pi(p,L)$ is a smooth function of p for finite L, $\Pi(p,L\to\infty)$ will develop into a step function with value zero for $p<p_c$ and value 1 for $p>p_c$. The finite-size scaling idea suggests that one may write $\Pi(p,L)$ as $\Pi[(p-p_c)L^{1/\nu}]$. As L gets larger, we expect that the width of Π gets smaller.

We define $p_c(L)$ by the following integral:

$$p_c(L)=\int p\left(\frac{d\Pi}{dp}\right)dp, \tag{17}$$

which simply means that we calculate the average p weighted by $d\Pi/dp$, a distribution approaching $\delta(p-p_c)$ when $L\to\infty$. Since $\Pi(p,L)$ can be directly measured by Monte Carlo simulations, one can obtain $p_c(L)$ unambiguously with a large enough number of simulations.

Because $d\Pi/dp=L^{1/\nu}\Pi'[(p-p_c)L^{1/\nu}]$ with $[\Pi'(z)\equiv d\Pi(z)/dz]$, we find, when combining with Eq. (17), that

$$\begin{aligned}p_c(L)-p_c&=\int[(p-p_c)L^{1/\nu}]\Pi'[(p-p_c)L^{1/\nu}]dp\\&=L^{-1/\nu}\int z\Pi'(z)dz\propto L^{-1/\nu}\end{aligned} \tag{18}$$

provided that $\Pi'(z)$ is not a symmetric function of z. In the extremely rare case that $\Pi'(z)$ is symmetric around z, $p_c(L)$ will approach p_c even faster. To find simultaneously ν and p_c using Eq. (18), however, requires elaborate trials.

To avoid this tedious procedure, we follow a well-known method, as described in Ref. [43], to compute the width $\Delta(L)$ of $\Pi(p,L)$ via

$$\Delta^2\equiv\int[p-p_c(L)]^2\left(\frac{d\Pi}{dp}\right)dp. \tag{19}$$

Upon writing

$$[p-p_c(L)]^2=(p-p_c)^2+2[p_c-p_c(L)](p-p_c)+[p_c-p_c(L)]^2 \tag{20}$$

and making a simple change of variable to $L^{1/\nu}(p-p_c)$ in place of p on the right-hand side of Eq. (19), one finds that

$$\Delta(L)\sim L^{-1/\nu}, \tag{21}$$

since every one of the three terms on the right-hand side of Eq. (20) is of order $\mathcal{O}(L^{-2/\nu})$. Therefore one has

TABLE I. The values of the width of $\Pi(p)$, the effective p_c, and the location of the maximum of $d\langle S\rangle/dp$ for various system sizes but with gap penalty fixed at δ=0.4. Plotting $\Delta(L)$ against $p_c(L)$ allows one to determine $p_c \equiv p_c(L\to\infty)$, as shown in Fig. 4. The average of the maximum score, elaborated in Sec. IV B, exhibits a weak finite-size effect, and the maximum of its first derivative with respect to p is close to $p_c(L)$.

δ	L	$\Delta(L)$	$p_c(L)$	$p\|_{\max d\langle S\rangle/dp}$
0.4	100	0.0252	0.136	0.138
0.4	200	0.0188	0.128	0.135
0.4	300	0.0161	0.125	0.129
0.4	400	0.0147	0.123	0.127
0.4	500	0.0132	0.121	0.124
0.4	600	0.0123	0.120	0.121

$$p_c(L) - p_c \sim \Delta(L). \tag{22}$$

When plotting $p_c(L)$ as the ordinate and $\Delta(L)$ as the abscissa, one should observe a straight line and the intercept at the y axis should be p_c. We use this method to compute p_c.

Note that $p_c(L)$ does depend on the value of δ used. For each size in L=[100,200,300,400,500,600], the quantiy $\Pi(p,L)$ is obtained by using 250 000 realizations of random potentials and then by taking the ratio of (# percolated)/250 000. Computing $\Pi(p,L)$ from p=0 to p=1 (with increment 0.0048 in p) and using Eq. (17), we deterimine $p_c(L)$ and $\Delta(L)$ for the δ value investigated. This procedure is used for each $\delta \in \{0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8\}$. As an example, we show in Table I the $p_c(L)$ and $\Delta(L)$ values for different L when δ=0.4. Figure 4 then illustrates how we use Eq. (22) to obtain p_c for δ=0.4. Using p_c obtained from Fig. 4, we plot $\ln[p_c(L)-p_c]$ against $\ln(L)$ in Fig. 5 to show that it is indeed straightforward to extract the exponent ν once p_c is determined via Eq. (22). The dependence of ν on δ is documented in Table II. When repeating these procedures for all δ of interest, we obtain the δ dependence of p_c which is also documented in Table II.

B. Scaling of the average of the best score

At $p=p_c$, v (the average score gain per length) is zero. When p is near p_c, we assume that $v \sim \pm|p-p_c|^\gamma$. This assumption will be verified in this section by scaling analysis of the averaged maximum score.

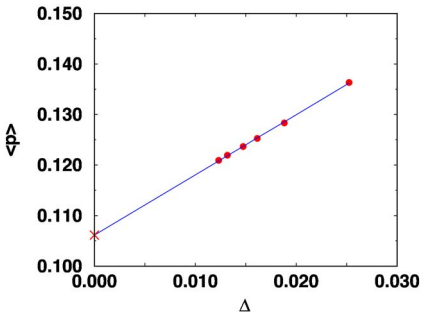

FIG. 4. (Color online) The extrapolation method for obtaining p_c. In this example, the gap penalty is δ=0.4 and the p_c obtained is 0.107.

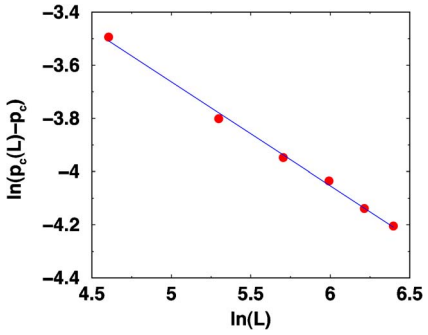

FIG. 5. (Color online) The determination of exponent ν: using the p_c value obtained from Fig. 4 [or equivalently Eq. (22)], one may then use Eq. (21) to obtain the exponent ν from the inverse of the slope of $\ln[p_c(L)-p_c]$ vs $\ln(L)$.

When translating our understanding of the DPRM to global alignment, the score fluctuations for alignment of length t are proportional to $t^{1/3}$ while the average score (over many realizations) is expected to be vt. Right at the critical point v=0, we expect the best scoring path of length t to have alignment score $a(\delta)t^{1/3}$ with the positive constant $a(\delta)$ depending on the gap penalty. Near the critical point, the system cannot quite tell whether it is right at the critical point or not. For $p<p_c$ and $|p-p_c| \ll 1$, the best score will keep growing as $a(\delta)t^{1/3}$ until the linear term $-|p-p_c|^\gamma t$ becomes of comparable size. This defines a saturation of the score for the $p<p_c$ side. Basically, the two quantities become comparable at t_c,

$$t_c = [a(\delta)|p-p_c|^{-\gamma}]^{3/2}, \tag{23}$$

and at this point the best score is expected to saturate at $|p-p_c|^{-\gamma/2}a^{3/2}(\delta)$. When $t<t_c$, the system behaves as if it were still at the critical point, but realizes that it is below p_c when $t>t_c$. Therefore it is the ratio L/t_c that constitutes the argu-

TABLE II. The values of $p|_{\max d^2S/dp^2}$, p_c, ν, and γ for different gap penalties δ. Because the exponent ν is obtained from the inverse value of a slope that itself sensitively depends on the p_c value used, the estimated error bar for ν is the largest. The standard error for each value in the second and third columns is estimated to be ±0.002, while the standard error for each value in the last column is estimated to be ±0.004.

δ	$p\|_{\max d^2S/dp^2}$	p_c	ν	γ
0.1	0.014	0.015	2.03±0.25	0.750
0.2	0.043	0.044	2.34±0.25	0.775
0.3	0.076	0.074	2.65±0.25	0.800
0.4	0.107	0.107	2.58±0.25	0.825
0.5	0.135	0.136	2.55±0.25	0.850
0.6	0.164	0.162	2.56±0.25	0.875
0.7	0.186	0.183	2.82±0.25	0.900
0.8	0.204	0.205	2.50±0.25	0.925

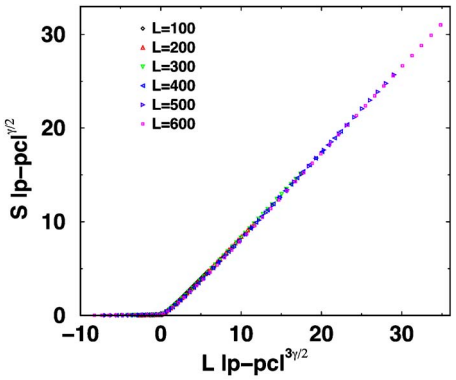

FIG. 6. (Color online) A typical data collapse for $\langle S\rangle$ of various sizes based on the scaling function proposed. The positive abscissa indicates the quantity $L(p-p_c)^{3\gamma/2}$; the negative abscissa records $-L(p_c-p)^{3\gamma/2}$. The gap penalty δ is 0.4. By varying the exponent γ, we find the best γ to be around 0.825.

ment of the scaling function for the averaged score (for $p<p_c$):

$$\langle S\rangle = |p-p_c|^{-\gamma/2} f_-(L^{2/3\gamma}|p-p_c|), \tag{24}$$

with $f_-(x\ll 1)=a(\delta)x^{\gamma/2}$ and $f_-(x\gg 1)=$constant due to the score saturation.

Similarly, when $p>p_c$ and $|p-p_c|\ll 1$, the linear growth term vt becomes dominant only for $t>t_c$ and the score still behaves as if the system were at criticality for $t<t_c$. Therefore for $p>p_c$

$$\langle S\rangle = |p-p_c|^{-\gamma/2} f_+(L^{2/3\gamma}|p-p_c|), \tag{25}$$

with $f_+(x\ll 1)=a(\delta)x^{\gamma/2}$ and $f_+(x\gg 1)=x^{3\gamma/2}$. If we define $f(x>0)=f_+(x)$ and $f(x<0)=f_-(-x)$, then

$$\langle S\rangle = |p-p_c|^{-\gamma/2} f(L^{2/3\gamma}(p-p_c)). \tag{26}$$

The scaling of $\langle S\rangle$ is verified by data collapse using data from many different sizes. Figure 6 shows a typical example of our many data collapse results, each with high quality collapse. The δ dependence of γ is documented in Table II. The trend that γ increases with δ can be understood intuitively. The larger the δ value, the harder it penalizes the gap and the harder it is for high scoring segments to connect through gaps. Consequently, v is expected to vanish faster near p_c.

Apparently, one may regard $t_c(p)$ as the correlation length that diverges near p_c as $|p-p_c|^{-3\gamma/2}$. However, $t_c(p)$ is not the same as the correlation length ξ (the cluster size) that we defined for the percolating probability $\Pi(p,L)$. This is because when $t<t_c$, the previous analysis assumes that the score keeps increasing as $t^{1/3}$. The cluster we defined for percolation, however, may start with one such segment of increasing score, then connect to a bad region where the score slowly decreases, then connect to a segment of increasing score, etc. Therefore our percolation cluster size $\xi(p)$ at a given p is in general larger than $t_c(p)$. The first consequence expected from this argument is that the exponent ν should be larger than $3\gamma/2$ for all δ. As shown in Table II, this is indeed the case.

Further, because $\xi(p)>t_c(p)$, the finite size effect becomes more severe when using $\Pi(p,L)$ as opposed to using $\langle S\rangle_L$. Near critical p, the susceptibility $d^2\langle S\rangle/dp^2$ should show a peak near p_c. When numerically computing this susceptibility, we find that for lattice sizes 300 and larger, there is virtually no difference in terms of peak locations. In Table II, we document the peak location of the susceptibility for various δ in the column headed by $p|_{\max d^2S/dp^2}$. The very small difference between the peak location of the susceptibility and the p_c obtained using Eq. (22) supports the conclusion that using $\langle S\rangle$ leads to a smaller finite-size effect. Another interesting phenomenon, as exemplified in Table I, is that the p value where the maximum of $d\langle S\rangle_L/dp$ occurs agrees reasonably well with $p_c(L)$ obtained using Eq. (17).

C. Score fluctuations for $p>p_c$

Through the numerically verified scaling of the score, elaborated in the previous subsection, we confirmed that the global alignment of length t will have score fluctuations $\chi t^{1/3}$, and obtained systematically the dependence on δ of the exponent γ. In this section, we will examine the probability distribution of the random variable χ. Although Prähofer and Spohn [26] have described for the DPRM three different subclasses that are all consistent with the $t^{1/3}$ score fluctuations, a more detailed study is needed in order to confirm whether any of these three universality subclasses can be applied to the global alignment model we investigated. In order to get the theoretical curves for the three known subclasses, we need the solution $u(x)$ to the Painlevé II equation (12). Because of the nonlinearity, the numerical stability range is very small. We therefore have to perform an asymptotic analysis and use it to obtain an initial value of high enough accuracy to provide a stable solution. The asymptotics are derived in detail in the Appendix.

For a given δ and a p that is greater than $p_c(\delta)$, $N=1\,000\,000$ realizations of the random potential are indexed by i. We may rewrite $S_{\max}(t\equiv m+n)=vt+\chi t^{1/3}$ with their explicit label:

$$S_{\max;i} = vt_i + \chi_i t_i^{1/3}. \tag{27}$$

Note that if our χ were to follow F_{GUE} or F_{GOE}, it would have a nonzero average. This point has to be taken into account explicitly when trying to decide which subclass best fits our random variable χ.

Let us first start with the assumption that our χ variable, after transforming to a proper scale, follows either F_{GUE} or F_{GOE}. Apparently, we have to find out both the scale λ and try to single out the value χ_i for each event i. This is possible because, aside from the average value, the second moment is known for both F_{GUE} and F_{GOE}. We can therefore form a scale independent ratio

$$R = \frac{\langle\chi^2\rangle_c}{\langle\chi\rangle^2} \equiv \frac{\langle\chi^2\rangle - \langle\chi\rangle^2}{\langle\chi\rangle^2}. \tag{28}$$

For F_{GUE}, $R=0.259\,248$; for F_{GOE}, $R=1.104\,454$ [26].

Starting from

$$v t_i^{2/3} = \frac{S_{\max;i}}{t_i^{1/3}} - \chi_i,$$

we have

$$v = \frac{\sum_{i=1}^{N} S_{\max;i}/t_i^{1/3}}{\sum_{i=1}^{N} t_i^{2/3}} - \frac{N\langle\chi\rangle}{\sum_{i=1}^{N} t_i^{2/3}} \equiv v_0 - b'\langle\chi\rangle,$$

with $b' = N/(\Sigma t_i^{2/3})$. Similarly, one may also write

$$\sum_{i=1}^{N} \chi_i^2 = \sum_{i=1}^{N} \left[\frac{S_{\max;i}}{t_i^{1/3}} - v_0 t_i^{2/3} + b'\langle\chi\rangle t_i^{2/3} \right]^2 \equiv N(c + b\langle\chi\rangle + a\langle\chi\rangle^2), \tag{29}$$

with

$$c = \frac{1}{N}\sum_{i=1}^{N} \left[\frac{S_{\max;i}}{t_i^{1/3}} - v_0 t_i^{2/3} \right]^2,$$

$$b = \frac{2}{N}\sum_{i=1}^{N} \left[\frac{S_{\max;i}}{t_i^{1/3}} - v_0 t_i^{2/3} \right] b' t_i^{2/3},$$

$$a = \frac{1}{N}\sum_{i=1}^{N} b'^2 t_i^{4/3}.$$

Equation (29) can then be turned into

$$\langle\chi^2\rangle_c = c + b\langle\chi\rangle + (a-1)\langle\chi\rangle^2, \tag{30}$$

whose left-hand side may be replaced by $R\langle\chi\rangle^2$. One may therefore solve for $\langle\chi\rangle$ via quadrature. Indeed, we have

$$\langle\chi\rangle = \frac{b \pm \sqrt{b^2 + 4(R-a+1)c}}{2(R-a+1)}. \tag{31}$$

We take the negative solution. The ratio between $\langle\chi\rangle$ and the value documented in Ref. [26] dictates the appropriate scale λ. Basically, λ is the multiplicative factor for χ in order to transform it into $\tilde{\chi} \equiv \lambda\chi$, the variable used for these theoretical distributions. Once $\langle\chi\rangle$ is found, we may obtain the individual χ_i by

$$\chi_i = \frac{S_{\max;i}}{t_i^{1/3}} - [v_0 - b'\langle\chi\rangle] t_i^{2/3}$$

and obtain the histogram of $\lambda\chi$.

When assuming F_0 to be the correct distribution function for χ, we cannot use the ratio R to pin down the scale because $\langle\chi\rangle = 0$. However, the scale can still be obtained via $\langle\chi^2\rangle_c = \langle\chi^2\rangle - \langle\chi\rangle^2 = \langle\chi^2\rangle$. Notice that χ_i comes out very simply:

$$\chi_i = \frac{S_{\max;i}}{t_i^{1/3}} - v_0 t_i^{2/3}. \tag{32}$$

We may then compute the scale factor λ via

$$\lambda^2 = \frac{\langle\tilde{\chi}^2\rangle_{F_0}}{\frac{1}{N}\sum_{i=1}^{N} \left(\frac{S_{\max;i}}{t_i^{1/3}} - v_0 t_i^{2/3} \right)^2}, \tag{33}$$

where $\langle\tilde{\chi}^2\rangle_{F_0} = 1.150\,39$ can be taken from Ref. [26]. One then obtains the pdf($\tilde{\chi} \equiv \lambda\chi$) in a very straightforward manner.

Under this protocol, with the scale factor λ determined, there will be *no free parameters* needed to fit our $\lambda\chi$ histogram with the three standardized distributions. As a matter of fact, one will just plot the theoretical curve from a standardized distribution on top of the corresponding normalized histogram of $\lambda\chi$. To be quantitative about the quality of the agreement, we employ the Kolmogorov-Smirnov (KS) statistics test [44] to see whether or not we may regard the numerically obtained distribution of $\lambda\chi$ as generated from the theoretical distribution assumed [45]. Using only the largest cumulative deviation between two distributions, the KS test provides the likelihood of two distributions (one obtained experimentally and numerically, the other given theoretically) being identical. In this study, the largest cumulative deviation and the likelihood of identity for each pair of distributions will be provided especially when comparing among several pairs that are not easily discernible by eye. Further, for a clear visual demonstration, we also plot the cumulative difference as a function of the variable considered.

Within the range $1.2 \leqslant p/p_c \leqslant 1.6$, we find that F_{GOE} gives the best overlap with its corresponding normalized histograms. As evidenced by the plots displaying the cumulative difference between the numerical distribution obtained and the theoretical distribution, histograms corresponding to the F_0 distribution have fewer counts in the large value tail and have more counts in the low value tail. This trend is reversed for histograms corresponding to the F_{GUE} distribution. Typical examples are given in Figs. 7 and 8 for the cases of $\delta = 0.3$ and $\delta = 0.6$, respectively.

This behavior, however, turns out to be transient. As p increases, the pdf of score fluctuations shifts from F_{GOE} to F_{GUE}. And F_{GUE} remains the pdf that best fits its corresponding histograms for even higher p values. Figures 9 and 10 show the overlaps and the cumulative difference between the theoretical curves and their corresponding histograms at $p = 0.5$. As shown by the KS test results and the corresponding plots, one may easily see that F_{GUE} agrees best with its corresponding histograms. Figures 11 and 12 show similar behavior for a higher p value ($p = 0.8$).

In view of random matrix statistics, F_{GUE} and F_{GOE} result from matrices of different symmetry. Therefore the shift of the pdf of χ from F_{GOE} to F_{GUE} merits further investigation, even though a direct connection between the DPRM (or sequence alignment) and random matrix statistics is not yet found. There are several possibilities of such a pdf change to occur: a fortuitous artifact (i.e., the distribution F_{GOE} appears to be a close fit only for a specific size and has no real meaning), a phase transition, or a crossover phenomenon [46] (meaning that F_{GOE} is a critical fixed point whose criti-

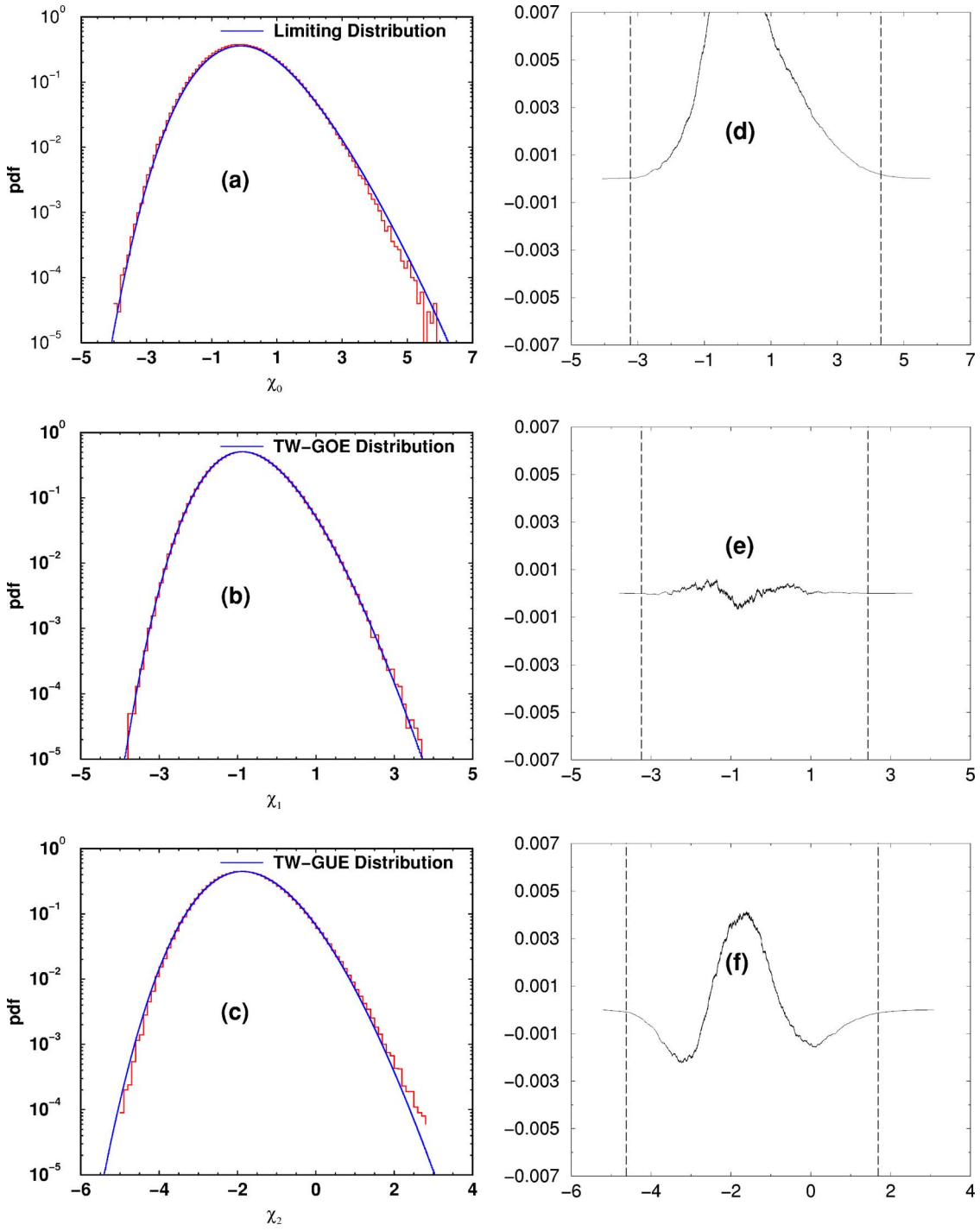

FIG. 7. (Color online) The pdf's of χ and the cumulative deviations between the numerical and theoretical distributions for $p > p_c$ but p still close to p_c. The relevant parameters are as follows: lattice size $L=600$, $p=0.0975$, and gap penalty $\delta=0.3$. With F_0 being the distribution assumed, (a) displays a histogram of $\lambda\chi$ (with $\lambda=3.4552$) and a theoretical curve of the F_0 distribution, while (d) displays the cumulative difference between the numerical distribution and F_0 with the largest absolute deviation being 1.05×10^{-2}; given by the KS statistics test, the likelihood for F_0 to be the correct distribution is 5.69×10^{-10}. With F_{GOE} being the distribution assumed, (b) displays a histogram of $\lambda\chi$ (with $\lambda=2.5680$) and a theoretical curve of the F_{GOE} distribution, while (e) displays the cumulative difference between the numerical distribution and F_{GOE} with the largest absolute deviation being 6.85×10^{-4}; given by the KS statistics test, the likelihood for F_{GOE} to be the correct distribution is 1.0. With F_{GUE} being the distribution assumed, (c) displays a histogram of $\lambda\chi$ (with $\lambda=2.8914$) and a theoretical curve of the F_{GUE} distribution, while (f) displays the cumulative difference between the numerical distribution and F_{GUE} with the largest absolute deviation being 4.14×10^{-3}; given by the KS statistics test, the likelihood for F_{GUE} to be the correct distribution is 6.50×10^{-2}. The pdf's are obtaining by normalizing the histogram properly. In (d)–(f), regions with theoretical pdf values larger than 10^{-3} are sandwiched by two vertical dashed lines.

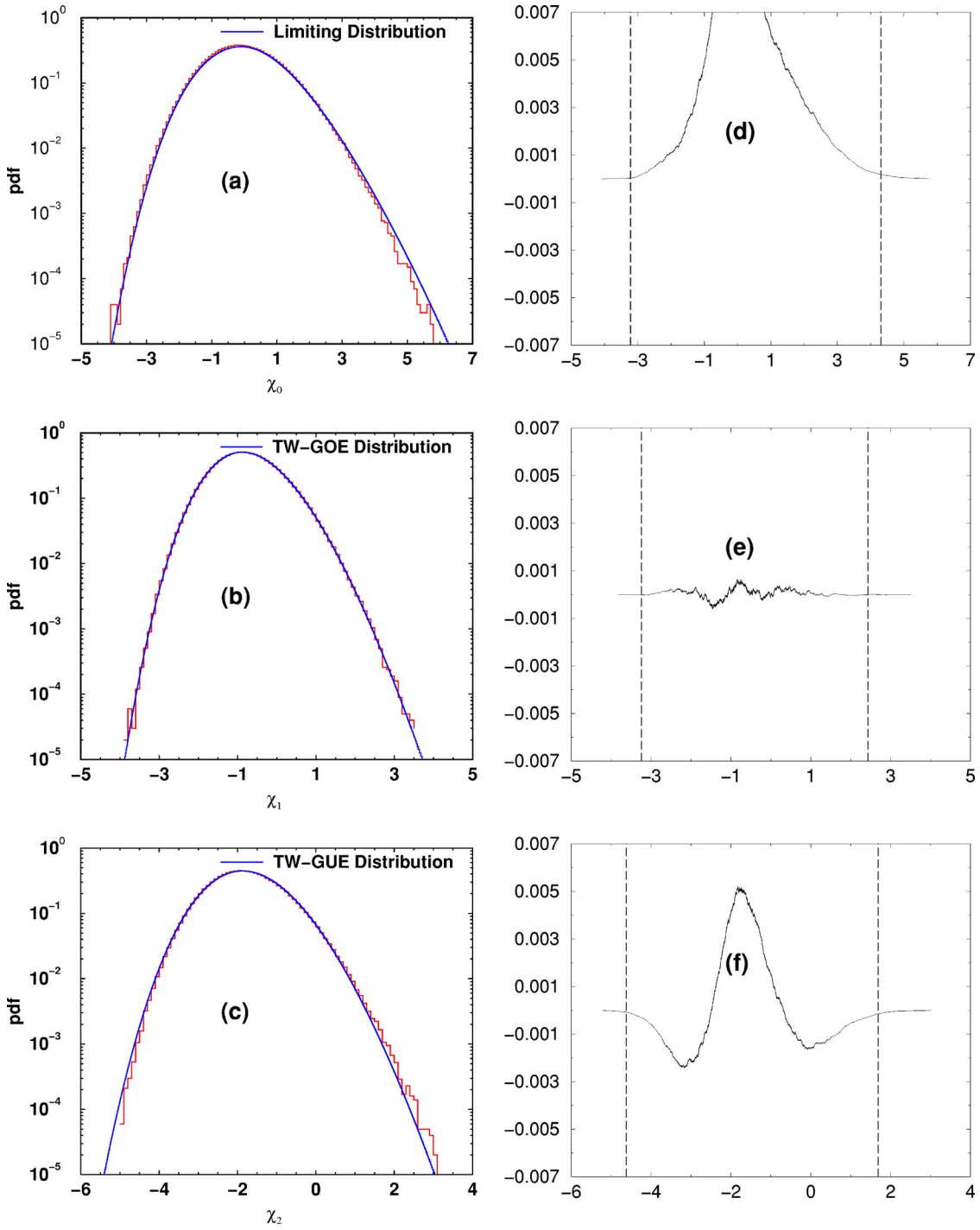

FIG. 8. (Color online) The pdf's of χ and the cumulative deviations between the numerical and theoretical distributions for $p > p_c$ but p still close to p_c. The relevant parameters are as follows: lattice size $L=600$, $p=0.2093$, and gap penalty $\delta=0.6$. With F_0 being the distribution assumed, (a) displays a histogram of $\lambda\chi$ (with $\lambda=2.8373$) and a theoretical curve of the F_0 distribution, while (d) displays the cumulative difference between the numerical distribution and F_0 with the largest absolute deviation being 1.14×10^{-2}; given by the KS statistics test, the likelihood for F_0 to be the correct distribution is 1.03×10^{-11}. With F_{GOE} being the distribution assumed, (b) displays a histogram of $\lambda\chi$ (with $\lambda=2.1109$) and a theoretical curve of the F_{GOE} distribution, while (e) displays the cumulative difference between the numerical distribution and F_{GOE} with the largest absolute deviation being 6.47×10^{-4}; given by the KS statistics test, the likelihood for F_{GOE} to be the correct distribution is 1.0. With F_{GUE} being the distribution assumed, (c) displays a histogram of $\lambda\chi$ (with $\lambda=2.3798$) and a theoretical curve of the F_{GUE} distribution, while (f) displays the cumulative difference between the numerical distribution and F_{GUE} with the largest absolute deviation being 5.23×10^{-3}; given by the KS statistics test, the likelihood for F_{GUE} to be the correct distribution is 8.37×10^{-3}. The pdf's are obtaining by normalizing the histogram properly. In (d)–(f), regions with theoretical pdf values larger than 10^{-3} are sandwiched by two vertical dashed lines.

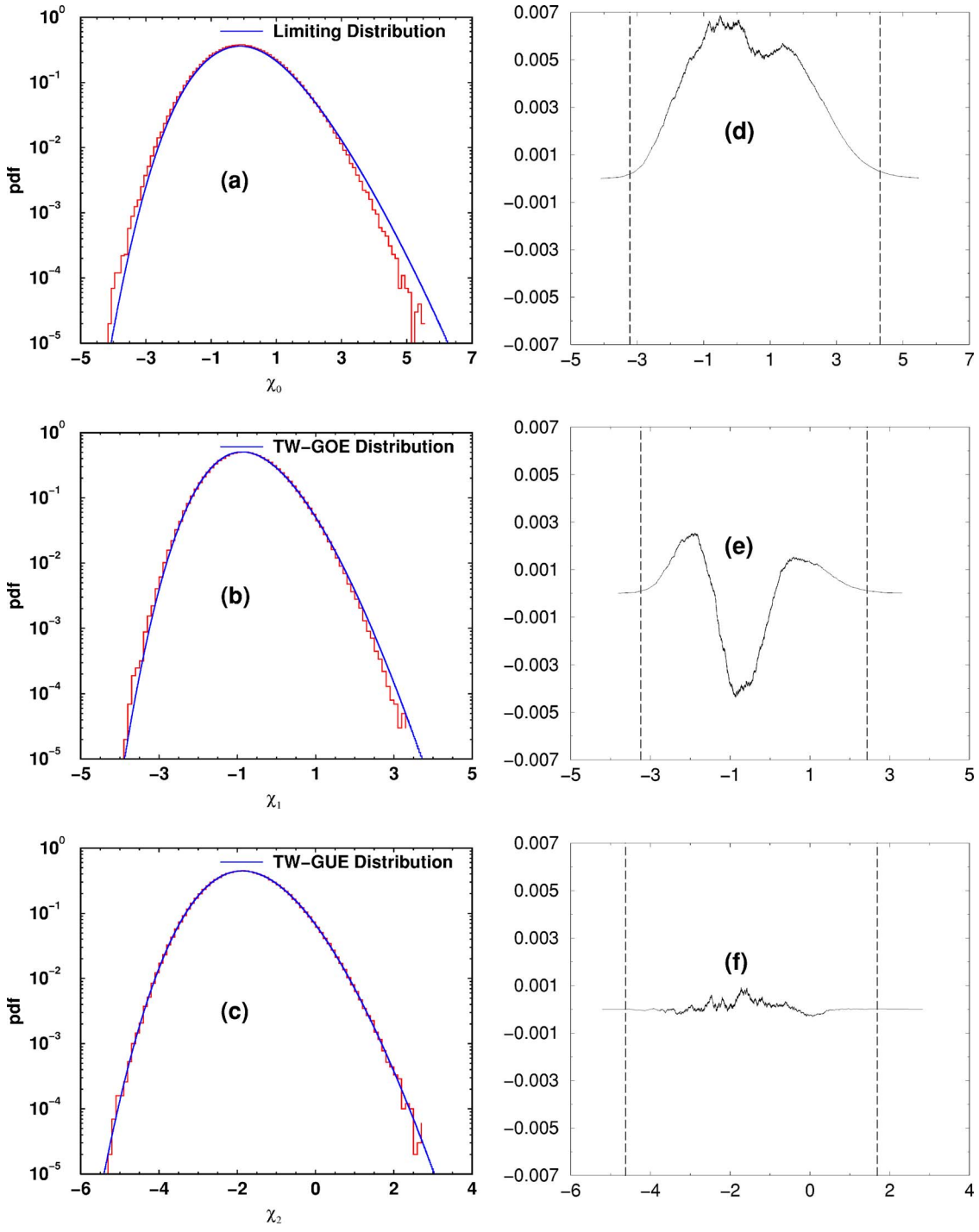

FIG. 9. (Color online) The pdf's of χ and the cumulative deviations between the numerical and theoretical distributions for $p > p_c$ with $p=0.5$. Lattice size 600 and gap penalty $\delta=0.3$ are used. With F_0 being the distribution assumed, (a) displays a histogram of $\lambda\chi$ (with $\lambda=3.3867$) and a theoretical curve of the F_0 distribution, while (d) displays the cumulative difference between the numerical distribution and F_0 with the largest absolute deviation being 6.87×10^{-3}; given by the KS statistics test, the likelihood for F_0 to be the correct distribution is 1.56×10^{-4}. With F_{GOE} being the distribution assumed, (b) displays a histogram of $\lambda\chi$ (with $\lambda=2.5208$) and a theoretical curve of the F_{GOE} distribution, while (e) displays the cumulative difference between the numerical distribution and F_{GOE} with the largest absolute deviation being 4.40×10^{-3}; given by the KS statistics test, the likelihood for F_{GOE} to be the correct distribution is 4.18×10^{-2}. With F_{GUE} being the distribution assumed, (c) displays a histogram of $\lambda\chi$ (with $\lambda=2.8441$) and a theoretical curve of the F_{GUE} distribution, while (f) displays the cumulative difference between the numerical distribution and F_{GUE} with the largest absolute deviation being 9.32×10^{-4}; given by the KS statistics test, the likelihood for F_{GUE} to be the correct distribution is 1.0. The pdf's are obtaining by normalizing the histogram properly. In (d)–(f), regions with theoretical pdf values larger than 10^{-3} are sandwiched by two vertical dashed lines.

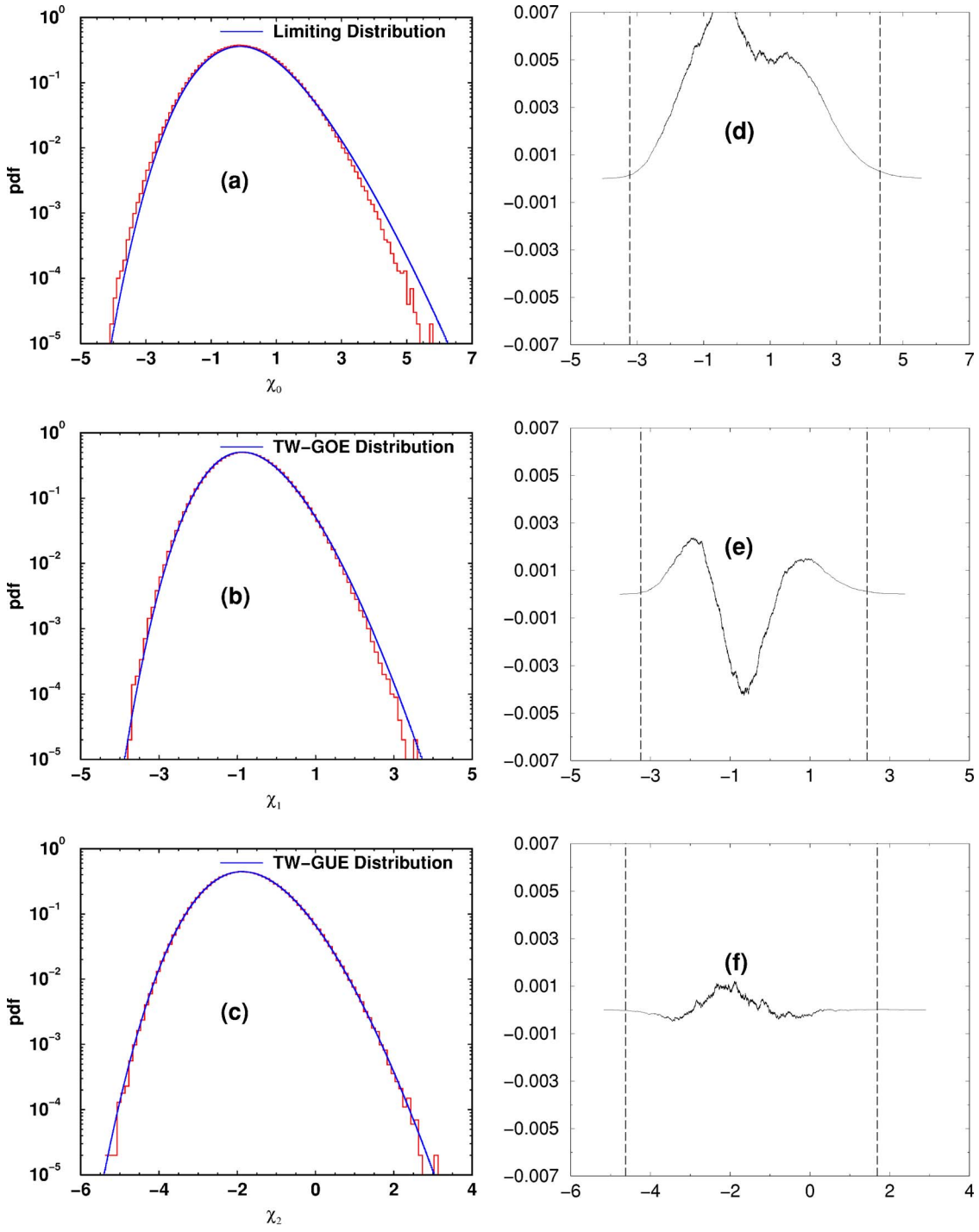

FIG. 10. (Color online) The pdf's of χ and the cumulative deviations between the numerical and theoretical distributions for $p > p_c$ with $p=0.5$. Lattice size 600 and gap penalty $\delta=0.6$ are used. With F_0 being the distribution assumed, (a) displays a histogram of $\lambda\chi$ (with $\lambda=2.8373$) and a theoretical curve of the F_0 distribution, while (d) displays the cumulative difference between the numerical distribution and F_0 with the largest absolute deviation being 7.48×10^{-3}; given by the KS statistics test, the likelihood for F_0 to be the correct distribution is 2.75×10^{-5}. With F_{GOE} being the distribution assumed, (b) displays a histogram of $\lambda\chi$ (with $\lambda=2.1110$) and a theoretical curve of the F_{GOE} distribution, while (e) displays the cumulative difference between the numerical distribution and F_{GOE} with the largest absolute deviation being 4.26×10^{-3}; given by the KS statistics test, the likelihood for F_{GOE} to be the correct distribution is 5.29×10^{-2}. With F_{GUE} being the distribution assumed, (c) displays a histogram of $\lambda\chi$ (with $\lambda=2.3801$) and a theoretical curve of the F_{GUE} distribution, while (f) displays the cumulative difference between the numerical distribution and F_{GUE} with the largest absolute deviation being 1.22×10^{-3}; given by the KS statistics test, the likelihood for F_{GUE} to be the correct distribution is 9.98×10^{-1}. The pdf's are obtaining by normalizing the histogram properly. In (d)–(f), regions with theoretical pdf values larger than 10^{-3} are sandwiched by two vertical dashed lines.

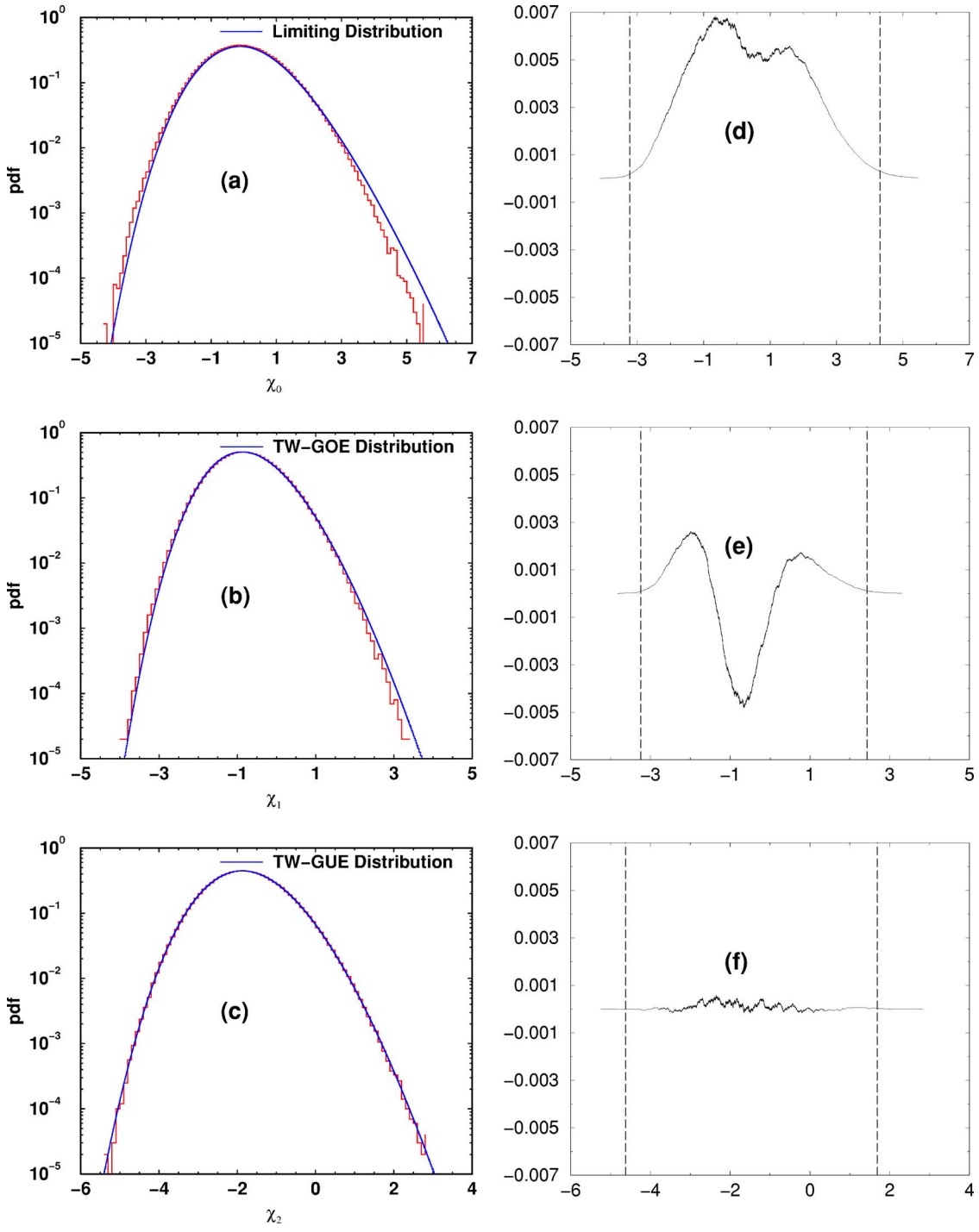

FIG. 11. (Color online) The pdf's of χ and the cumulative deviations between the numerical and theoretical distributions for $p > p_c$ with $p=0.8$. Lattice size 600 and gap penalty $\delta=0.3$ are used. With F_0 being the distribution assumed, (a) displays a histogram of $\lambda\chi$ (with $\lambda=3.9420$) and a theoretical curve of the F_0 distribution, while (d) displays the cumulative difference between the numerical distribution and F_0 with the largest absolute deviation being 6.84×10^{-3}; given by the KS statistics test, the likelihood for F_0 to be the correct distribution is 1.72×10^{-4}. With F_{GOE} being the distribution assumed, (b) displays a histogram of $\lambda\chi$ (with $\lambda=2.9347$) and a theoretical curve of the F_{GOE} distribution, while (e) displays the cumulative difference between the numerical distribution and F_{GOE} with the largest absolute deviation being 4.84×10^{-3}; given by the KS statistics test, the likelihood for F_{GOE} to be the correct distribution is 1.83×10^{-2}. With F_{GUE} being the distribution assumed, (c) displays a histogram of $\lambda\chi$ (with $\lambda=3.3119$) and a theoretical curve of the F_{GUE} distribution, while (f) displays the cumulative difference between the numerical distribution and F_{GUE} with the largest absolute deviation being 5.87×10^{-4}; given by the KS statistics test, the likelihood for F_{GUE} to be the correct distribution is 1.0. The pdf's are obtaining by normalizing the histogram properly. In (d)–(f), regions with theoretical pdf values larger than 10^{-3} are sandwiched by two vertical dashed lines.

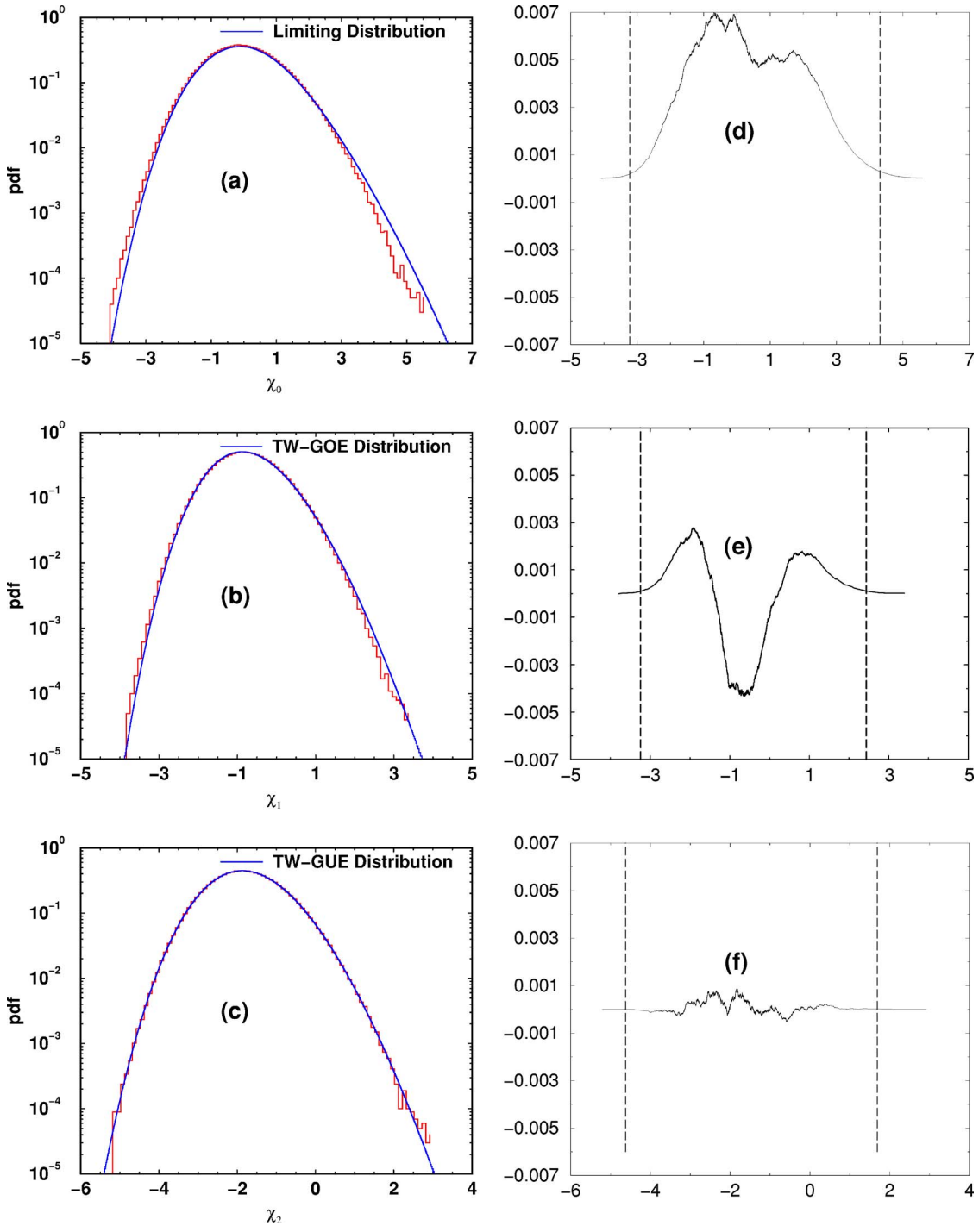

FIG. 12. (Color online) The pdf's of χ and the cumulative deviations between the numerical and theoretical distributions for $p > p_c$ with $p=0.8$. Lattice size 600 and gap penalty $\delta=0.6$ are used. With F_0 being the distribution assumed, (a) displays a histogram of $\lambda\chi$ (with $\lambda=3.2516$) and a theoretical curve of the F_0 distribution, while (d) displays the cumulative difference between the numerical distribution and F_0 with the largest absolute deviation being 7.07×10^{-3}; given by the KS statistics test, the likelihood for F_0 to be the correct distribution is 9.09×10^{-5}. With F_{GOE} being the distribution assumed, (b) displays a histogram of $\lambda\chi$ (with $\lambda=2.4750$) and a theoretical curve of the F_{GOE} distribution, while (e) displays the cumulative difference between the numerical distribution and F_{GOE} with the largest absolute deviation being 4.36×10^{-3}; given by the KS statistics test, the likelihood for F_{GOE} to be the correct distribution is 4.42×10^{-2}. With F_{GUE} being the distribution assumed, (c) displays a histogram of $\lambda\chi$ (with $\lambda=2.7923$) and a theoretical curve of the F_{GUE} distribution, while (f) displays the cumulative difference between the numerical distribution and F_{GUE} with the largest absolute deviation being 8.70×10^{-4}; given by the KS statistics test, the likelihood for F_{GUE} to be the correct distribution is 1.0. The pdf's are obtaining by normalizing the histogram properly. In (d)–(f), regions with theoretical pdf values larger than 10^{-3} are sandwiched by two vertical dashed lines.

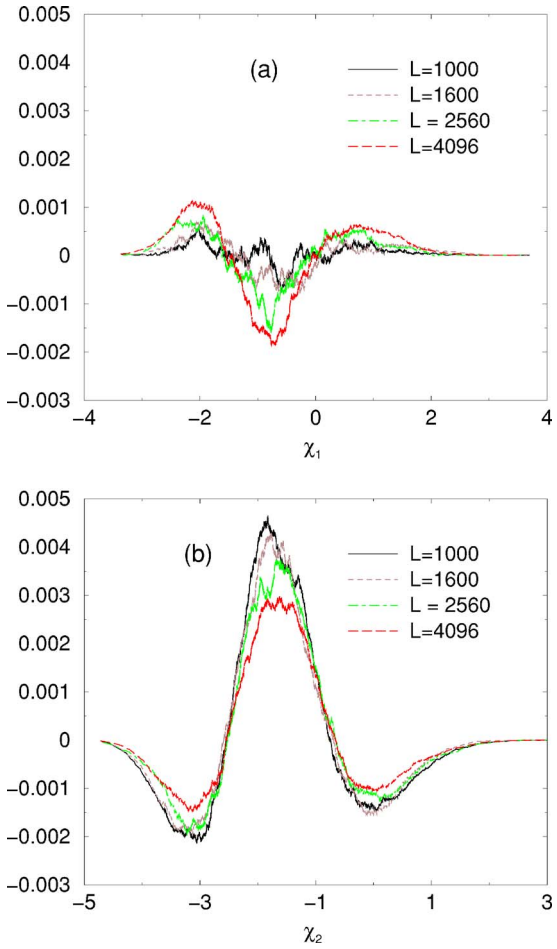

FIG. 13. (Color online) The gradual degradation (improvement) of the agreement between the pdf of χ and F_{GOE} (F_{GUE}) for $\delta=0.3$. System sizes of $L=1000$ (solid line), 1600 (dashed line), 2560 (dot-dashed line), and 4096 (long-dashed line) are studied with $[p-p_c(L)]/p_c(L)\approx 0.1$. Part (a) displays how the amplitude of cumulative difference between the numerical pdf and F_{GOE} gradually increases with size; part (b) displays how this amplitude decreases with size for F_{GUE}. In part (a), the gradual degradation leads to a decrease of the likelihood value (from 100.0 to 87.3 %); in part (b), the gradual improvement leads to an increase of the likelihood value (from 2.7 to 34.0 %).

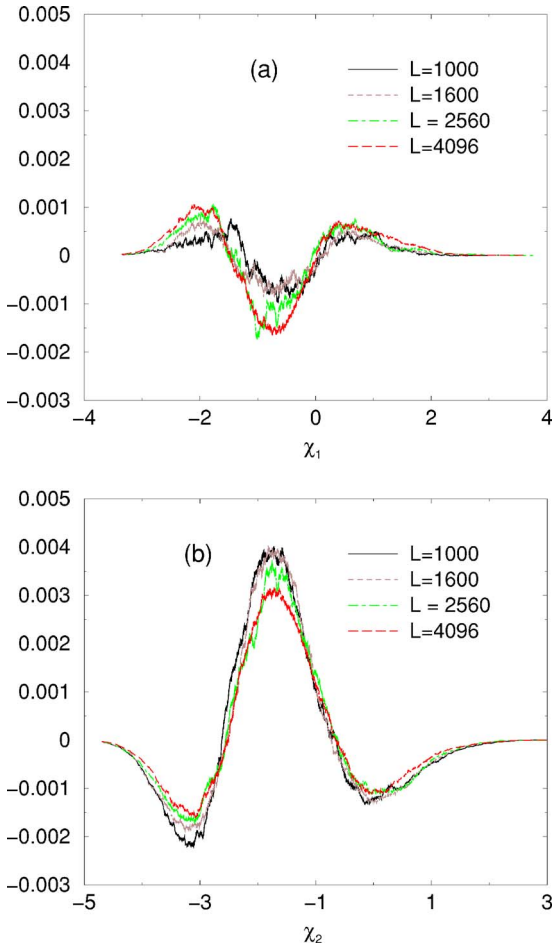

FIG. 14. (Color online) The gradual degradation (improvement) of the agreement between the pdf of χ and F_{GOE} (F_{GUE}) for $\delta=0.6$. System sizes of $L=1000$ (solid line), 1600 (dashed line), 2560 (dot-dashed line), and 4096 (long-dashed line) are studied with $[p-p_c(L)]/p_c(L)\approx 0.1$. Part (a) displays how the amplitude of cumulative difference between the numerical pdf and F_{GOE} gradually increases with size; part (b) displays how this amplitude decreases with size for F_{GUE}. In part (a), the gradual degradation leads to a decrease of the likelihood value (from 100.0 to 92.0 %); in part (b), the gradual improvement leads to an increase of the likelihood value (from 7.80 to 27.0 %).

cal surface is quite close to or even includes the percolation transition point).

The first possibility is unlikely since over a wide range of system size studied at p not too far away from $p_c(L)$, F_{GOE} appears—among the three possible theoretical distributions—to agree best with its corresponding histograms when using the Kolmogorov-Smirnov statistics test. The other two possibilites, a phase transition or a crossover, are more likely. As we point out earlier in this paper, when p just exceeds $p_c(L)$, the highest score path may start with a growth of score, followed by many large scale downs and ups in cumulative score, and finally reach the boundary region of the alignment lattice with highest cumulative score. In this case, when viewed at $t=L$ (half the maximum path length), the statistical averages of score change along the forward and backward directions appear similar. This may resemble the initial condition of flat substrate in the PNG growth [26]. When p (the percentage of positive score bonds) increases, it is expected that along the highest scoring path the number of large scale downs and ups in score will diminish. To investigate whether or not this effect leads to a phase transition (and consequently a change of pdf of χ), we looked for singular behavior of observables, such as $\Pi(p)$, $\langle S_{\max}(p)\rangle_0$, and their derivatives, in the p range where the shift from F_{GOE} to F_{GUE} takes place. No singular behavior was found.

This observation suggests a different explanation, the crossover phenomenon which turns out to match our observation well. As shown by the plots in Figs. 13 and 14 displaying the cumulative deviations for larger system sizes, we see a mild degradation (improvement) of the agreement between the pdf of χ and F_{GOE} (F_{GUE}) as the system size in-

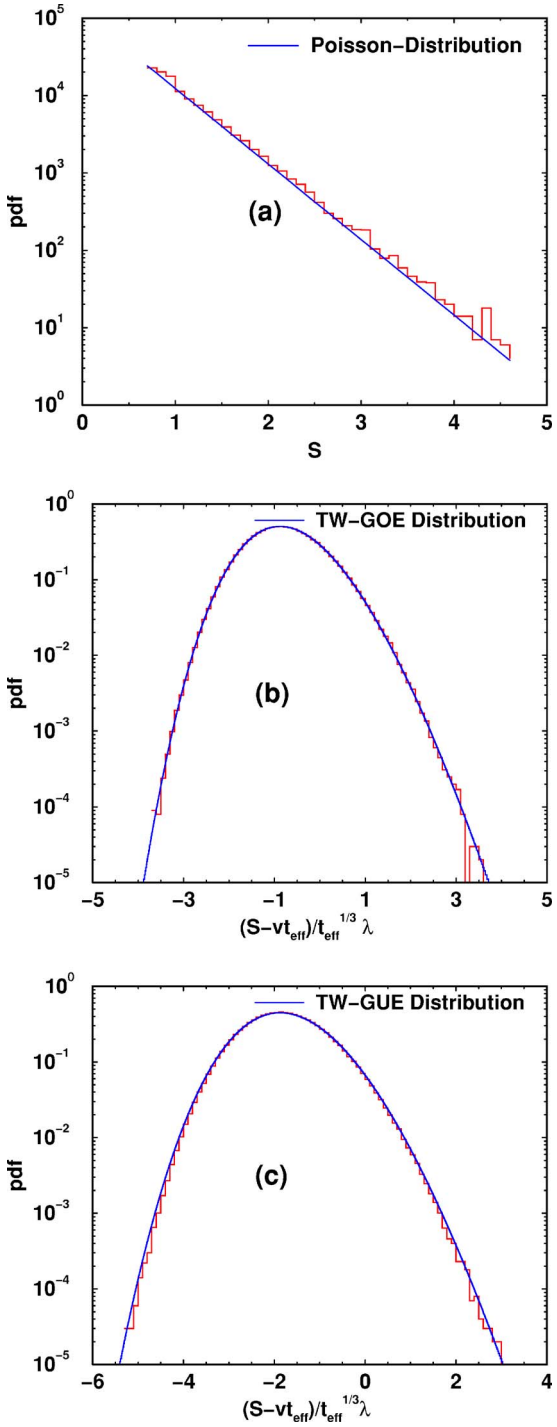

FIG. 15. (Color online) The score pdf for (a) $p<p_c$, (b) $p>p_c$ but close to p_c, and (c) p much larger than p_c. The lattice size used here is $L=600$ and the gap penalty used is $\delta=0.4$. The p values are (a) 0.077, (b) 0.1678, and (c) 0.80. For $\delta=0.4$, the critical p value at infinite size is $p_c=0.107$ while the finite size $p_c(L=600)=0.12$. The effective path lengths t_{eff} are 1168.2 for $p=0.1678$ (b), and 1196.6 for $p=0.8$ (c). The parameter v takes the values 0.046 88 and 0.269 85 respectively for (b) and (c).

creases. For example, for $\delta=0.3$, when the system size increases from $L=1000$ to $L=4096$, at $[p-p_c(L)]/p_c(L)\approx 0.1$, the likelihood of F_{GOE} decreases from 1.0 to 0.7193 while that of F_{GUE} increases from 0.032 to 0.3942; for $\delta=0.6$,

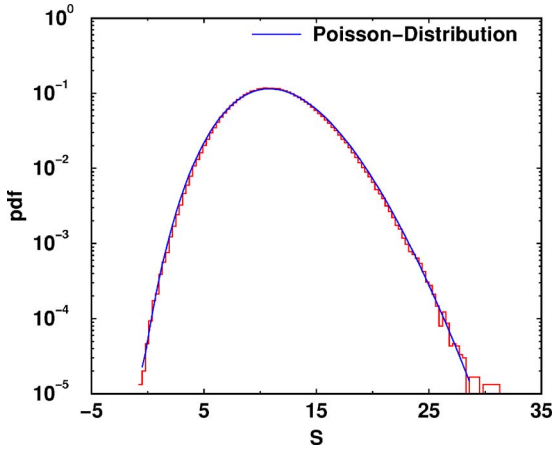

FIG. 16. (Color online) The score pdf for $p=0.1204$ [greater than but *very* close to $p_c(L=600)=0.12$]. The lattice size used here is $L=600$ and the gap penalty used is $\delta=0.4$. As one may see, the score pdf is well fitted by the Poisson distribution with parameters $A_1=0.96$, $A_2=0.5712$, and $\mu=11.46$.

when the system size increases from $L=1000$ to $L=4096$, the likelihood of F_{GOE} decreases from 1.0 to 0.92 while that of F_{GUE} increases from 0.078 to 0.27. This is consistent with a typical crossover scenario. There exists a stable fixed point (F_{GUE}) that can be approached by simply taking a large $p \gg p_c$. Further, there are two or more critical fixed points. The first one, corresponding to the percolation transition point, has two or more relevant scaling fields [47], with one of them flowing (under the coarse-graining procedure of the renormalization group) towards the large p fixed point and the other flowing to the second critical fixed point (F_{GOE}). Near the percolation transition point, if the initial system parameters happen to be very close to but not on the critical surface of the second critical fixed point, the effective system parameters under the coarse-graining procedures will first flow to the vicinity of the second critical fixed point but eventually flow into the global stable fixed point (F_{GUE}). Note that when system size is small, only a few coarse graining steps can be made without encountering the effect of the system boundary. But as the system size grow larger, more such steps can be performed and will bring the system to the global stable fixed point eventually.

It is worth noting that $L=4096$ is considerably larger than the number of amino acid residues in a typical protein. Therefore even though the accuracy of the distribution F_{GOE} will eventually degrade if the system parameters are not right on its corresponding critical surface, for alignments of practical sizes F_{GOE} does represent the distribution of χ when the effective p value of the system is close to $p_c(L)$. It is also interesting to note that in a recent study [30] of the Bernoulli matching (BM) problem, F_{GOE} is the only distribution that is discussed in Ref. [26] but not realizable by the BM model.

D. Score distribution

In real application of global alignment, it is possible that the acutal length is not very large and therefore the score

TABLE III. The asymptotic expansion coefficients for $u(x \ll -1) \equiv \sqrt{-x/2}[1+\Sigma_{i=1}^{\infty}(a_{3i}/x^{3i})]$.

$-a_3$	a_6	$-a_9$	a_{12}	$-a_{15}$	a_{18}	$-a_{21}$
$\frac{1}{8}$	$\frac{73}{128}$	$\frac{10657}{1024}$	$\frac{13912277}{32768}$	$\frac{8045883943}{262144}$	$\frac{14518451390349}{4194304}$	$\frac{18847128706420641}{33554432}$

statistics cannot be inferred accurately just from the distribution of χ. Therefore it might serve a practical purpose if one can characterize the score distribution with modest lattice sizes such as those used in our simulations. This issue is investigated only briefly and empirically in this section.

We found that for $p < p_c(L)$ the probability density of the maximum score exhibits an exponential tail generally fittable by the form

$$y = \exp[-(x - x_0)C]. \tag{34}$$

For $p > p_c(L)$, the majority of events have their path lengths comparable to $2L$. Therefore one may imagine that there is an effective path length t_{eff} (with $t_{\text{eff}} \sim 2L$ for very large p values), and the score distribution is given by

$$f_{\text{pdf}}(S) = dF(\lambda(S - vt_{\text{eff}})/t_{\text{eff}}^{1/3})/dS \tag{35}$$

with F being the distribution function of $\tilde{\chi}$. As we have demonstrated in Figs. 7 and 8, for p greater than but close to p_c, $F(\tilde{\chi})$ is best described by F_{GOE}. For p much greater than p_c, we show—in Figs. 9–12—that F is best described by F_{GUE}. The crossover p value for the score pdf, being similar to that of the score fluctuations, is generally dependent on δ.

The effective length t_{eff} can be determined as follows. Let us define $x \equiv t_{\text{eff}}^{1/3}$. We then have

$$\frac{\lambda}{N}\sum_{i=1}^{N}\left[\frac{S_{\max;i}}{x} - vx^2\right] = \langle\tilde{\chi}\rangle,$$

which then leads to the following cubic equation:

$$x^3 + x\frac{\langle\tilde{\chi}\rangle}{\lambda v} - \frac{\langle S\rangle}{v} = 0$$

that can be solved by elementary methods. The parameters v and λ are obtained via the method proposed in the previous subsection and the value of $\langle\tilde{\chi}\rangle$ is taken from Ref. [26]. As expected, t_{eff} found that this way is very close to the ensemble-averaged path length $\Sigma_{i=1}^{N} t_i/N$.

Figure 15 shows all three possible cases using lattice size 600 and δ=0.4. In part (a), we see that the tail of the pdf of S for $(p=0.077) < p_c(L)$ is well fitted by an exponential. The constants associated with the fitted line are C=2.2466 and x_0=5.190. In part (b), we see that for $p > p_c(L)$ but close to p_c (p=0.1678), the pdf of S is well fitted by $F_{\text{GOE}}(\lambda(S - vt_{\text{eff}})/t_{\text{eff}}^{1/3})$. In part (c), we see that the score distribution for large p value (p=0.8 here) indeed follows the Tracy-Widom GUE distribution $F_{\text{GUE}}(\lambda(S - vt_{\text{eff}})/t_{\text{eff}}^{1/3})$.

To characterize the alignment score statistics using these known distributions, however, one will need to have a good estimate of the parameters v and t_{eff} (or C and x_0) for a given δ. Fortunately, using methods proposed in the previous subsection, these values can be determined with a relatively small number of simulations provided that one knows which theoretical distribution to use. This information presumably can be precomputed via simulations over various ranges of p for different gap costs.

Finally, let us note that for p *very* close to $p_c(L)$ the distribution of the maximum score is well-fitted by the Poisson distribution with the form

$$f_{\text{pdf}}(S) = \frac{A_1}{\sqrt{2\pi}}[e^{(\ln\mu+1)\tilde{S}}\tilde{S}^{-(\tilde{S}+1/2)}e^{-\mu}], \tag{36}$$

where $\tilde{S} \equiv A_1 S + A_2$ represents the appropriately scaled variable and μ is the expectation value of $\langle\tilde{S}\rangle$. An example is shown in Fig. 16. Although at this point we are not able to predict the associated parameters (μ, A_1, A_2) of the Poisson distribution from scoring parameters used, we speculate that the Poisson distribution is the correct pdf for scores right at the critical point. This speculation is supported by the fact that the score pdf is no longer fittable by the Poisson distribution when p increases further where it first turns into F_{GOE} and then into F_{GUE}.

V. SUMMARY AND OUTLOOK

In this paper we investigate the score statistics of global sequence alignment using a variant DPRM model. We introduce a parameter p, the probability of having a positive substitution score at each diagonal bond, to mimic different levels of compositional divergence between two compared sequences. The larger the p value, the more likely it is for the highest score point to be near the boundary of the alignment lattice, similar to a percolating system.

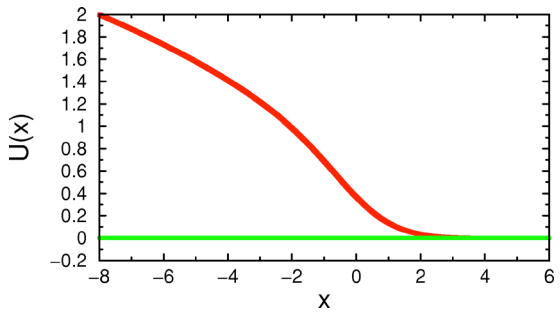

FIG. 17. (Color online) Numerical solution of the Painlevé II equation.

By using finite-size scaling on $\Pi(p,L)$, the probability of percolation at fraction p and size L, we obtain p_c, the critical p at infinite size. We also identify the δ (gap penalty) dependence of the finite size exponent ν. In the context of percolating clusters, the correlation length ξ may be regarded as the largest cluster size (the path length from the origin to the highest score point in the lattice). We also obtain the scaling function for the average maximum score $\langle S\rangle$. The critical length t_c obtained in this analysis represents the length within which the score only increases with length, and is therefore smaller than ξ. We reason and verify that indeed $\langle S\rangle_L$ has a weaker finite-size effect than $\Pi(p,L)$. The dependence of the linear growth velocity v on δ near p_c is also examined and documented.

To investigate which of the three distributions—the limiting distribution F_0, the Tracy-Widom F_{GOE}, or the Tracy-Widom F_{GUE}—our random variable χ follows, we perform an extensive numerical simulation using lattices of linear size up to $L=4096$. Our finding suggests that F_{GUE} is the global stable fixed point for large p while F_{GOE} is a critical fixed point that characterizes well the score statistics when $0<[p-p_c(L)]/p_c(L)\ll 1$. Upon increasing p, the change of pdf of χ from F_{GOE} to F_{GUE} is most likely a crossover phenomena. That is, unless the system parameters are right on the critical surface of F_{GOE}, the pdf of χ will evolve towards F_{GUE} upon the renormalization group flow. Our study suggests that the percolation transition point is quite close to the critical surface of F_{GOE}. However, whether the critical surface of F_{GOE} intersects with the p-δ phase plane or not remains to be studied.

In general it is not known *a priori* how to extract the variable χ from the score and how to scale χ to $\tilde{\chi}$, the variable used for these three standardized distributions. We develop a method to extract the χ variable from each event accurately and with the correct scale factor λ simultaneously determined. To compare with the theoretical predictions, we need an accurate numerical solution to the Painlevé II equation. The prerequisite for this task is to find an initial point x_0 with $u(x_0)$ and $u'(x_0)$ accurately specified. As shown in the Appendix, we accomplish this requirement by a systematic asymptotic expansion of $u(x)$ when x takes negative value but $|x|\gg 1$.

As a cautionary note, we must point out that a systematic application of our theoretical analysis to score statistics is hindered by a few issues. First, the dependence of parameters such as v_{eff} and t_{eff} on gap cost are not yet known analytically; second, the functional realtion between p_c and the gap costs is also not yet known analytically. Fortunately, it is possible to obtain those quantities on a set of selected gap costs using numerical means. This heuristic approach was actually taken in the alignment statistics of BLAST [48]. Once the effective parameters are given, the score statistics follows Eq. (35). The procedure outlined in Sec. IV D allows one to extract the effective parameters using numerical simulations. Through this procedure, the possiblity to characterize the score statistics of modest sequence lengths is illustrated.

Empirically, we find that when $p<p_c$, the pdf of S has an exponential tail. When $p>p_c(L)$ and even for the modest size of $L=600$, the scores—after proper subtraction and rescaling—follow the distribution function of $\tilde{\chi}$. That is to say, when p is slightly greater than $p_c(L)$, the score distribution follows F_{GOE}; when p is much larger than p_c, the score distribution follows F_{GUE}. For p very close to the transition point $p_c(L)$, we find that the score distribution can be well fitted by the Poisson distribution, which we speculate is the correct score pdf at the critical point.

The connection between random matrix statistics and alignment score statistics is interesting and perhaps deserves more investigation. For example, it is not obvious how the symmetry assumed by a certain random matrix ensemble is realized in the sequence alignment problem. It will be interesting to see if any other features of random matrix statistics find their analogs in sequence alignment. For example, it remains to be seen whether the level spacing statistics (characterized by Wigner-Dyson distribution) and/or the universal spectral rigidity have their counterparts in sequence alignment.

ACKNOWLEDGMENTS

We acknowledge the partial support from the NSF through Grant No. DMR-0110903 during the early stage of this research. We also thank the administrative group of the NIH biowulf clusters, where the majority of the computational tasks were carried out. This work was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS.

APPENDIX

To obtain theoretical distribution curves, we need an accurate numerical solution to the Painlevé II equation (12). This appendix shows how we may obtain an accurate $u(x)$ for $x\ll 0$.

The standard method to expand a function at large argument is the asymptotic expansion. Our starting point is to write $u(x)$ as the product of a power series of $(1/x)$ and the asymptotic solution of $u(x)$ at $x\to-\infty$:

$$u(x)=\sqrt{\frac{-x}{2}}\left[1+\sum_{\ell=1}^{\infty}\frac{a_\ell}{x^\ell}\right]\equiv\sqrt{\frac{-x}{2}}+\sum_{\ell=1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}}. \tag{A1}$$

We will show that only a subset of $\{a_\ell\}$ (or equivalently $\{k_\ell\}$) can be nonzero.

Our strategy is to substitute the expression (A1) into both sides of Eq. (12). First we note that

$$u_{xx}=\frac{1}{4\sqrt{2}}(-x)^{-3/2}+\sum_{\ell=1}^{\infty}\frac{\left(\ell^2-\frac{1}{4}\right)k_\ell}{(-x)^{\ell+3/2}}$$

has a leading order term proportional to $(-x)^{-3/2}$. This then demands the cancellation, on the right-hand side of Eq. (12), of terms $(-x)^\alpha$ whose power α is greater than $-3/2$. Now the quantity $2u^3+xu$ can be written as

$$2u^3 + xu = 2\left(\frac{-x}{2}\right)^{3/2} + 3(-x)\sum_{\ell=1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}} + 6\sqrt{\frac{-x}{2}}\left[\sum_{\ell=1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}}\right]^2 + 2\left[\sum_{\ell=1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}}\right]^3 + x\sqrt{\frac{-x}{2}} + x\sum_{\ell=1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}} = 2(-x)\sum_{\ell=1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}} + 6\sqrt{\frac{-x}{2}}\left[\sum_{\ell=1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}}\right]^2 + 2\left[\sum_{\ell=1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}}\right]^3. \tag{A2}$$

Note that the leading powers from the second and third terms are $(-x)^{-1/2}$ and $(-x)^{-3/2}$, respectively. Demanding that the $(-x)^{1/2}$ term have zero coefficient, we find immediately that $k_1=0$. Upon setting $k_1=0$, we see that the second and third terms of Eq. (A2) have leading powers $(-x)^{-5/2}$ and $(-x)^{-9/2}$, respectively. Demanding that the coefficient of $(-x)^{-1/2}$ be zero immediately leads to $k_2=0$. Since the left hand side of the Painlevé II equation does contain an $(-x)^{-3/2}$ term, $k_3\neq 0$. This then tells us that the sum over k_ℓ should start with $\ell=3$. The second derivative of $k_3/(-x)^{3-1/2}$ leads to $(-x)^{-9/2}$ and this is the next leading power on the right hand side of Eq. (12). Similarly, when starting with $\ell=3$, the second and third terms of Eq. (A2) have leading powers $(-x)^{-9/2}$ and $(-x)^{-15/2}$, respectively. Requiring the coefficients of the $(-x)^{-5/2}$ term and the $(-x)^{-7/2}$ term to be zero, we find immediately that $k_4=k_5=0$. It is then a simple matter to use the method of induction to show that $k_{3\ell-2}=k_{3\ell-1}=0$. For the sake of completeness, we will perform such an induction analysis here. Assume that the expansion of $u(x)$ up to $\ell=3j$ with $j\geqslant 2$ is given by

$$u(x) = \sqrt{\frac{-x}{2}} + \sum_{\ell=1}^{j}\frac{k_{3\ell}}{(-x)^{3\ell-1/2}} + \sum_{\ell=3j+1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}}. \tag{A3}$$

We want to show that this implies that

$$u(x) = \sqrt{\frac{-x}{2}} + \sum_{\ell=1}^{j+1}\frac{k_{3\ell}}{(-x)^{3\ell-1/2}} + \sum_{\ell=3(j+1)+1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}}. \tag{A4}$$

By substituting Eq. (A3) into Eq. (A2), we have

$$2u^3 + xu = 2\left(\sqrt{\frac{-x}{2}} + \sum_{\ell=1}^{j}\frac{k_{3\ell}}{(-x)^{3\ell-1/2}}\right)^3 + 6\left(\sqrt{\frac{-x}{2}} + \sum_{\ell=1}^{j}\frac{k_{3\ell}}{(-x)^{3\ell-1/2}}\right)^2\left[\sum_{\ell=3j+1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}}\right] + 6\left(\sqrt{\frac{-x}{2}} + \sum_{\ell=1}^{j}\frac{k_{3\ell}}{(-x)^{3\ell-1/2}}\right)\left[\sum_{\ell=3j+1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}}\right]^2 + 2\left[\sum_{\ell=3j+1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}}\right]^3 + x\left(\sqrt{\frac{-x}{2}} + \sum_{\ell=1}^{j}\frac{k_{3\ell}}{(-x)^{3\ell-1/2}}\right) + x\left[\sum_{\ell=3j+1}^{\infty}\frac{k_\ell}{(-x)^{\ell-1/2}}\right].$$

The second derivative of $u(x)$ contains, for powers larger than $-3j-3/2$, only $\{(-x)^{-3\ell-3/2}\}_{\ell=0}^{j}$. That is, there is no term with $(-x)^{-3j+1/2}$ and no term with $(-x)^{-3j-1/2}$. Checking the right-hand side of Eq. (A5), we see that the first and the fifth terms cannot possibly generate terms with these powers. When $j\geqslant 2$, the third and the fourth terms do not contain the powers $-3j+1/2$ and $-3j+1/2$ either. The only terms that matter are the second and the sixth terms. Demanding that the $(-x)^{-3j+1/2}$ term have zero coefficient, we see that $k_{3j+1}=0$. Similarly, demanding that $(-x)^{-3j-1/2}$ have zero coefficient, we obtain $k_{3j+2}=0$. Therefore we have shown by induction that the asymptotic expansion of $u(x)$ with $x<0$ and $|x|\gg 0$ has the following form:

$$u(x) = \sqrt{\frac{-x}{2}}\left[1 + \sum_{\ell=1}^{\infty}\frac{a_{3\ell}}{x^{3\ell}}\right]. \tag{A5}$$

The coefficients up to order $(-x)^{-41/2}$ were then computed using MATHEMATICA and tabulated in Table III. This asymptotic expansion yields accurate values for $u(-8)$ and $u_x(-8)$, which then allows for numerical integration of $u(x)$ towards positive x. Figure 17 shows the numerical solution obtained through this approach. This numerical solution allows us to calculate $g(x)$ and $f(x)$ which are needed for computing the theoretical distribution functions of χ.

[1] D. A. Huse and C. L. Henley, Phys. Rev. Lett. **54**, 2708 (1985).
[2] M. Kardar, Nucl. Phys. B **290**, 582 (1987).
[3] D. S. Fisher and D. A. Huse, Phys. Rev. B **43**, 10728 (1991).
[4] T. Halpin-Healy and Y.-C. Zhang, Phys. Rep. **254**, 215 (1995).
[5] M. Q. Zhang and T. G. Marr, J. Theor. Biol. **174**, 119 (1995).
[6] T. Hwa and M. Lässig, Phys. Rev. Lett. **76**, 2591 (1996).
[7] T. Hwa and M. Lassig, *RECOMB98* (1998), pp. 109–116.
[8] D. Drasdo, T. Hwa, and M. Lassig, *ISMB98* (1998), pp. 52–58.
[9] M. Kschischo and M. Lassig, Pac. Symp. Biocomput **5**, 621-632 (2000).
[10] Yi-Kuo Yu and T. Hwa, J. Comput. Biol. **8**, 249 (2001).
[11] Yi-Kuo Yu, R. Bundschuh, and T. Hwa, Bioinformatics **18**, 864 (2002).
[12] Yi-Kuo Yu, R. Bundschuh, and T. Hwa, in *Biological Evolution and Statistical Physics*, edited by M. Lässig and A. Valeriani (Springer-Verlag, Berlin, 2002), pp. 3–22.
[13] Yi-Kuo Yu, Phys. Rev. E **69**, 061904 (2004).

[14] M. Kschischo, M. Lässig, and Yi-Kuo Yu, Bull. Math. Biol. **67**, 169 (2005).
[15] Yi-Kuo Yu, J. C. Wootton, and S. F. Altschul, Proc. Natl. Acad. Sci. U.S.A. **100**, 15688 (2003).
[16] Yi-Kuo Yu and S. F. Altschul, Bioinformatics **21**, 902 (2005).
[17] A. A. Schäffer *et al.*, Nucleic Acids Res. **29**, 2994 (2001).
[18] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Nucleic Acids Res. **25**, 3389 (1997).
[19] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis* (Cambridge University Press, New York, 1998).
[20] M. S. Waterman, *Introduction to Computational Biology* (Chapman & Hall, London, UK, 1994).
[21] D. F. Feng and R. F. Doolittle, Methods Enzymol. **266**, 368 (1996).
[22] A. K. Hartmann, Phys. Rev. E **65**, 056102 (2002).
[23] E. J. Gumbel, *Statistics of Extremes* (Columbia University Press, New York, 1958).
[24] R. Bundschuh, Phys. Rev. E **65**, 031911 (2002).
[25] D. Forster, D. R. Nelson, and M. J. Stephen, Phys. Rev. A **16**, 732 (1977).
[26] M. Prähofer and H. Spohn, Phys. Rev. Lett. **84**, 4882 (2000); Physica A **279**, 342 (2000).
[27] J. Baik, P. Deift, and K. Johansson, J. Am. Math. Soc. **12**, 1119 (1999).
[28] J. Baik and E. M. Rains, in *Random Matrix Models and Their Applications*, edited by P. Bleher and A. Its (Cambridge University Press, Cambridge, England, 2001), Vol. 40, pp. 1–19.
[29] J. Baik and E. M. Rains, J. Stat. Phys. **100**, 523 (2000).
[30] S. N. Majumdar and S. Nechaev, Phys. Rev. E **69**, 011103 (2004).
[31] K. Johansson, Commun. Math. Phys. **209**, 437 (2000).
[32] S. N. Majumdar and S. Nechaev, Phys. Rev. E **72**, 020901(R) (2005).
[33] N. I. Levedev and Y.-C. Zhang, J. Phys. A **28**, L1 (1995).
[34] T. Halpin-Healy, Phys. Rev. E **58**, R4096 (1998).
[35] S. B. Needleman and C. D. Wunsch, J. Mol. Biol. **48**, 443 (1970).
[36] Due to the difference in degrees of freedom, there does exist a subtle difference between the sequence comparison and the regular DPRM problems in terms of the noise correlations [38]. Nevertheless, it has been argued [6] and shown numerically [39] that this does not lead to much effect.
[37] In general, one may choose $[L-c\sqrt{L},L]$ as the range for identifying a percolating event with c a reasonable positive constant. Indeed, varying c in the range [0.5, 2] shows no appreciable difference in our final results.
[38] P. DeLosRios and Y.-C. Zhang, Phys. Rev. Lett. **81**, 1023 (1998).
[39] R. Olsen, R. Bundschuh, and T. Hwa, in *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, edited by T. Lengauer *et al.* (AAAI Press, Menlo Park, CA, 1999), pp. 211–222.
[40] A. M. Vershik and S. V. Kerov, Sov. Math. Dokl. **19**, 527 (1977); Funct. Anal. Appl. **19**, 21 (1985).
[41] C. A. Tracy and H. Widom, Commun. Math. Phys. **159**, 151 (1994); **177**, 727 (1996).
[42] For a review on the subject of random matrices, see M. L. Metha, *Random Matrices*, 2nd ed. (Academic Press, San Diego, 1991).
[43] D. Stauffer and A. Aharoni, *Introduction to Percolation Theory*, revised 2nd ed. (Taylor & Francis, Philadelphia, 1994).
[44] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. (Cambridge University Press, New York, 1999).
[45] An important feature of the KS statistics test is that the same likelihood is obtained even if one replaces the random variable x of the distributions considered by any monotonic function of x. For example, the same result is obtained regardless whether one uses x, e^x, or $1/x$ as the random variable [44].
[46] A. Aharony, in *Phase Transitions and Critical Phenomena*, edited by C. Domb and M. S. Green (Academic Press, London, 1976), Vol. 6; for an intuitive and very clear introduction, see J. Cardy, *Scaling and Renormalization in Statistical Physics* (Cambridge University Press, New York, 2002).
[47] F. J. Wegner, in *Phase Transitions and Critical Phenomena*, edited by C. Domb and M. S. Green (Academic Press, London, 1976), Vol. 6.
[48] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, J. Mol. Biol. **215**, 403 (1990); S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Nucleic Acids Res. **25**, 3389 (1997).