DNA SEQUENCE ALIGNMENT AND CRITICAL PHENOMENA

Dirk DRASDO[(1)], TERENCE HWA[(2)], and MICHAEL LÄSSIG[(1)]
[(1)] Max-Planck Institut für Kolloid- und Grenzflächenforschung, 14513 Teltow, Germany
[(2)] Department of Physics, University of California at San Diego, La Jolla, CA 92093-0319

Abstract

Alignment algorithms are commonly used to detect and quantify similarities between DNA sequences. We study these algorithms in the framework of a recent theory viewing similarity detection as a geometrical critical phenomenon of directed random walks. We show that the *roughness* of these random walks governs the *fidelity* of an alignment, i.e., its ability to capture the correlations between the sequences compared. Criteria for the optimization of alignment algorithms emerge from this theory.

Introduction

The explosion of genetic information has made statistical sequence alignment an indispensable tool in molecular biology [1]. The identity of new genes and relationships about known genes are routinely analyzed by aligning sequences on a computer. This underlies, for example, the retrieval of ancestorial relationships in the history of evolution.

In a typical algorithm [2, 3, 4], each alignment of sequences is assigned a score specified by a set of parameters. Maximization of this score is then used to select the *optimal alignment*, which depends, of course, strongly on the parameters used to define the score. What are then *optimal alignment parameters* making the algorithm most sensitive to the similarities of the sequences? This important problem has so far been solved mostly by trial and error, despite some recent efforts to establish a more solid empirical footing [5, 6].

In this paper, we study the parameter optimization problem using a recent analytical approach to sequence alignment introduced by two of us [7]. The approach is based on a geometrical formulation of sequence alignment [2] and focuses on the morphology of *alignment paths*. This provides a fruitful link (see also Ref. [8]) to various well-studied problems in the statistical physics of critical phenomena.

In a divergent evolution process, similarities between sequences stem from a common ancestor sequence and are gradually destroyed in the course of time. We use a simplified model of evolution: Sequences are altered by a stochastic process of local substitutions, insertions, and deletions. In this model, the mutual similarities between daughter sequences inherited from their common ancestor can be identified uniquely. Hence, we can quantify in an unambiguous way the *fidelity* of an alignment algorithm [7], i.e., its ability to retrieve the inherited similarities from the knowledge of the daughter sequences alone. Then we analyze the dependence of the fidelity on the evolution parameters and the alignment parameters. Maximizing the fidelity defines optimal alignment parameters for given evolution parameters. Conversely, unknown evolution parameters can be reconstructed from alignment data.

In the sequel, we introduce the evolution model and the alignment algorithm used in this work, derive geometrical properties of alignment paths, and discuss how they govern the alignment fidelity and its parameter dependence. In particular, optimal alignment parameters are seen to follow from a simple geometric criterion. Further details of our work are reported in a forthcoming publication [9].

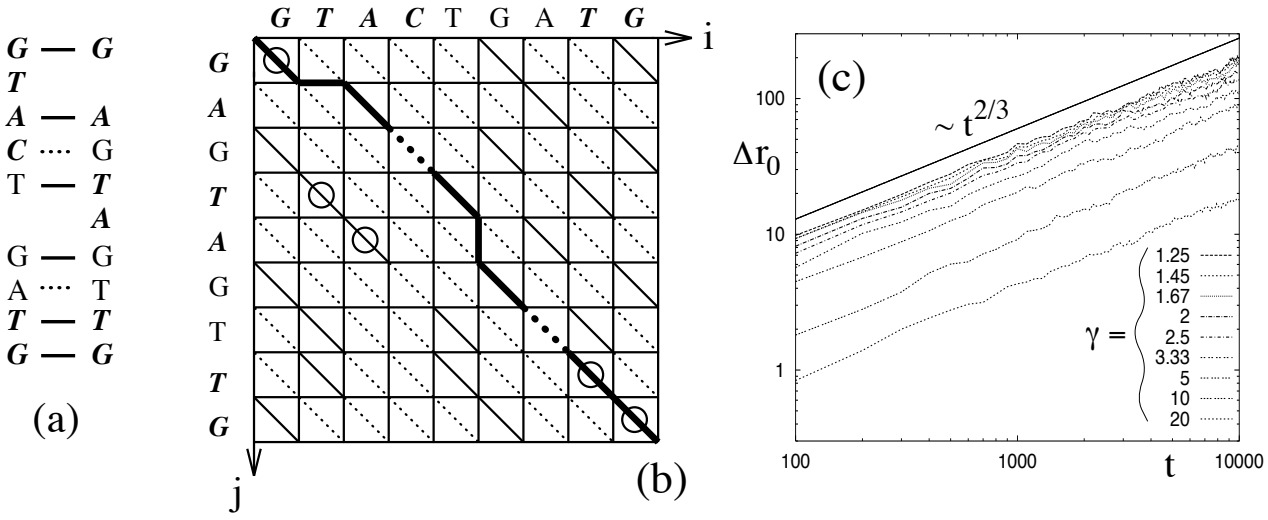

FIGURE 1: (a) One possible alignment of the sequences $\mathcal{D} = \{G,T,A,C,\mathrm{T},\mathrm{G},\mathrm{A},T,G\}$ and $\mathcal{D}' = \{G,A,\mathrm{G},T,A,\mathrm{G},\mathrm{T},T,G\}$; elements conserved from a common ancestor are shown in italics. The alignment has six matches (solid lines), two mismatches (dotted lines), and two gaps. (b) Representation on the alignment grid. Horizontal and vertical bonds represent gaps, solid (dotted) diagonal bonds represent matches (mismatches). Matches corresponding to native pairs are marked with a circle. The directed path $r(t)$ corresponding to the alignment in (a) is shown as a thick line. Its fidelity is 3/5. (c) The mean square displacement $\Delta r_0(t)$ of the optimal alignment path, obtained from a sample of 200 mutually uncorrelated sequence pairs for each value of γ.

Evolution Model and Alignment Algorithm

The simplified stochastic evolution process used in this paper generates two daughter sequences $\mathcal{D}$ and $\mathcal{D}'$ from a common ancestor sequence $\mathcal{A} = \{\mathcal{A}_k\}$ taken to be a random sequence of length $N \gg 1$. Each element $\mathcal{A}_k$ is with probability 1/4 one of the four different nucleotides A, C, G, T; we neglect any correlations within the ancestor sequence. A daughter sequence $\mathcal{D}$ is generated according to the following rules [7] (see also [10, 11]): (a) Each element $\mathcal{A}_k$ is *deleted* with probability $\tilde{p}/2$. (b) Each element $\mathcal{A}_k$ is *substituted* with probability $(1-\tilde{p}/2)p$ by a randomly chosen nucleotide. (c) If an element $\mathcal{A}_k$ is not deleted, an additional random nucleotide is *inserted* immediately to its right with probability $\tilde{p}/2$. If a random element has been inserted, another random nucleotide is inserted immediately to its right with probability $\tilde{p}/2$, etc.

An ancestor element $\mathcal{A}_k$ that is conserved in the evolution process (i.e., not deleted or substituted at any point) gets shifted to a position $i(k)$ due to the insertions and deletions of other elements, and appears as daughter element $\mathcal{D}_{i(k)}$. Since the daughter sequences $\mathcal{D}$ and $\mathcal{D}'$ are generated by independent realizations of the evolution process, a fraction $(1-\tilde{p}/2)^2(1-p)^2$ of the ancestor sequence elements is conserved in $\mathcal{D}$ *and* $\mathcal{D}'$. Each such element $\mathcal{A}_k$ defines a unique pair of daughter elements $(\mathcal{D}_{i(k)} = \mathcal{D}'_{j(k)})$ called a *native pair*.

Alignment algorithms are designed to find the native pairs from the knowledge of the daughter sequences $\mathcal{D}$ and $\mathcal{D}'$ alone. A global alignment of the two sequences is defined as an ordered set of pairings $(\mathcal{D}_i, \mathcal{D}'_j)$ and of gaps $(\mathcal{D}_i, -)$ and $(-, \mathcal{D}'_j)$, each letter $\mathcal{D}_i$ and $\mathcal{D}'_j$ belonging to exactly one pairing or gap (see Fig. 1 (a,b)) [2]. (It is clear that gaps are necessary to account for the shifts due to insertions and deletions and to allow the native pairs to be matched.) We define the *fidelity* of an alignment as the fraction of native pairs $(\mathcal{D}_i, \mathcal{D}'_j)$ that are correctly matched.

In this paper, we use the simplest version of the classic Needleman-Wunsch algorithm [2] to align the sequences $\mathcal{D}$ and $\mathcal{D}'$. An alignment is assigned a score

$$\Sigma = \sqrt{3}\, N_+ - \frac{1}{\sqrt{3}}\, N_- - \gamma N_g \tag{1}$$

given in terms of its total number N_+ of matches ($\mathcal{D}_i = \mathcal{D}'_j$), the total number N_- of mismatches ($\mathcal{D}_i \neq \mathcal{D}'_j$), and the total number N_g of gaps. The scoring function (1) has a single adjustable parameter γ, the effective gap penalty. Without loss of generality, the score contributions of matches and mismatches have been chosen in such a way that a pairing of two independent random elements has the average score 0 and the score variance 1. Maximizing the total score Σ defines the optimal alignment of the sequences $\mathcal{D}$ and $\mathcal{D}'$ for a given value of γ. (From a physicist's point of view, $-\Sigma$ is an energy function that has to be minimized.)

The following geometrical representation of global alignment will prove very useful. Fig. 1(b) shows a two-dimensional grid whose cells are labeled by the index pair (i, j). A given alignment of $\mathcal{D}$ and $\mathcal{D}'$ uniquely defines a *directed path* on the grid [2]: A diagonal bond in cell (i, j) represents the pairing of elements $(\mathcal{D}_i, \mathcal{D}'_j)$. A horizontal bond between cells (i, j) and $(i, j+1)$ represents a gap $(\mathcal{D}_i, -)$ located on sequence $\mathcal{D}'$ between the elements $\mathcal{D}'_j$ and $\mathcal{D}'_{j+1}$. Similarly, a vertical bond between cells (i, j) and $(i+1, j)$ represents a gap located on sequence $\mathcal{D}$ between the elements $\mathcal{D}_i$ and $\mathcal{D}_{i+1}$. Using the rotated coordinates $r \equiv i - j$ and $t \equiv i + j$, this alignment path is described by a single-valued function $r(t)$ measuring its displacement from the diagonal of the alignment grid. The path of the optimal alignment is denoted by $r_0(t)$. The set of native pairs resulting from a given evolutionary history corresponds to a set of special diagonal bonds marked by circles in Fig. 1(b). The fidelity of an alignment is the fraction of native bonds that lie on the alignment path $r(t)$. Any shortest trajectory through *all* native bonds defines again a directed path $R(t)$ on the alignment grid called *target path*.

Alignment Statistics and Roughness

The representation on the alignment grid enables us to express the statistics of the evolution process and of sequence alignments in terms of displacement fluctuations measuring the *roughness* of the directed paths $R(t)$ and $r_0(t)$.

Since insertions and deletions are assumed to be independent of each other, the target path $R(t)$ is just a free random walk on the alignment grid; its mean square displacement $(\Delta R(t_1 - t_2))^2 \equiv \overline{(R(t_1) - R(t_2))^2}$ is given by

$$\Delta R(t) = (\tilde{p}\,|t|)^{1/2} \ . \tag{2}$$

Here and throughout the paper, an overbar is used to denote the average over an ensemble of evolution processes for given p and $\tilde{p}$.

At first glance, the optimal alignment path may appear to be a free random walk as well. However, this is generally *not* the case. Consider first the optimal alignment of a pair of random sequences with no mutual correlations (i.e., in the limit $p = 1$). As pointed out in Ref. [7], the mean square displacement $(\Delta r_0(t_1 - t_2))^2 \equiv \overline{(r_0(t_1) - r_0(t_2))^2}$ of the optimal alignment path follows the scaling law

$$\Delta r_0(t) = A(\gamma)\,|t|^{2/3} \tag{3}$$

in the asymptotic regime $|t| \gg A^{-3/2}(\gamma)$, describing a *correlated* random walk.[1] As expected from the theory of critical phenomena, the entire parameter dependence of the displacement $\Delta r_0(t)$ is contained in the amplitude $A(\gamma)$. The exponent 2/3 of the power law is universal (i.e., independent of γ), just as the exponent 1/2 is universal for free random walks. We have verified this numerically (Fig. 1(c)) and have determined the amplitude function $A(\gamma)$ (for details, see [9]). Over the relevant range, $A(\gamma)$ is a monotonically decreasing function of γ, with $A(\gamma) \propto \gamma^{-4/3}$ in the biologically relevant regime $\gamma \gg 1$.

We can compare the roughness of the free random walk $R(t)$ and the correlated random walk $r_0(t)$. Equating the rms. displacements (2) and (3) defines the *roughness matching* scales

$$\tilde{t}(\gamma,\tilde{p}) = \tilde{p}^3/A^6(\gamma)\ , \qquad \tilde{r}(\gamma,\tilde{p}) = \tilde{p}^2/A^3(\gamma)\ . \tag{4}$$

For $|t| < \tilde{t}(\gamma,\tilde{p})$, the displacement of $R(t)$ exceeds that of $r_0(t)$. For $|t| > \tilde{t}(\gamma,\tilde{p})$, the displacement of the alignment path becomes dominant since the cost of gaps is outweighed by the gain in score from regions of the grid with an excess number of random matches.

Roughness Matching and Fidelity

For daughter sequences with nonvanishing mutual correlations (i.e., for $p < 1$), the statistics of matches and mismatches on the alignment grid differs from the case of uncorrelated sequences: along the target path $R(t)$, there are $U(p,\tilde{p}) \equiv (1-p)^2(1-\tilde{p}/2)^2$ extra matches per unit of t due to the native pairs. The optimal alignment path contains a finite fraction $\mathcal{F}$ of these extra matches, thereby increasing its score. Hence, it has no longer the displacement (3) and remains confined to the vicinity of the target path $R(t)$. The confinement length r_c is defined by the relative displacement of the two paths, $r_c^2(\gamma,p,\tilde{p}) \equiv \overline{(r_0(t)-R(t))^2}$. It is uniquely related to the average fidelity: $\overline{\mathcal{F}}$ decreases with increasing r_c, and $\overline{\mathcal{F}} \sim r_c^{-1}$ for small $\overline{\mathcal{F}}$ [9].

The behavior of the fidelity is well established for an evolution process without insertions and deletions ($\tilde{p} = 0$) [7]. In this limit, $\overline{\mathcal{F}}$ is found to be a monotonically increasing function of γ, reaching its maximum $\mathcal{F}^* = 1$ for $\gamma \to \infty$. This is not surprising since there is no need for any gaps in an alignment if the evolution process has no insertions and deletions.

For small values of $\overline{\mathcal{F}}$, the fidelity has the asymptotic form $\overline{\mathcal{F}} \sim \exp(-C(\gamma)/U(p,0))$, where $C(\gamma) \approx 2A^{3/4}$ is another amplitude function. This form is supported by our numerics [9]. For $U \to 0$, the fidelity approaches zero and the confinement length $r_c \sim \overline{\mathcal{F}}^{-1}$ diverges. This singularity marks a continuous phase transition at $U = 0$, which we call the *detectability transition*. Positive correlations between sequences ($U > 0$) are recovered with a finite fidelity $\overline{\mathcal{F}}(\gamma,U)$, while anticorrelations ($U < 0$) cannot be detected (i.e., $\mathcal{F} = 0$). For the alignment path $r_0(t)$, this is a critical (de-)localization transition[2] between the confined regime ($r_c < \infty$) for $U > 0$ and the regime of correlated displacement fluctuations (3) for $U \geq 0$.

---

[1] This class of random walks is in fact well known to physicists as *directed polymers* in a random medium ([12], see also [13] and references therein). Perhaps the most prominent example occurs in the theory of magnetic superconductors. If these materials are placed into a magnetic field, they develop tubular regions of normal magnetic conductance. These *flux lines* are directed parallel to the applied field (the t direction) and can be described by a displacement vector $r(t)$. In addition, there is often a distribution of point impurities. These act on the flux lines just like a random distribution of matches and mismatches on the alignment grid, causing large displacements of the lowest-energy path $r_0(t)$.

[2] The properties of this phase transition are known from the physics of a magnetic flux line interacting with an attractive linear defect $R(t) = 0$ [14, 15].

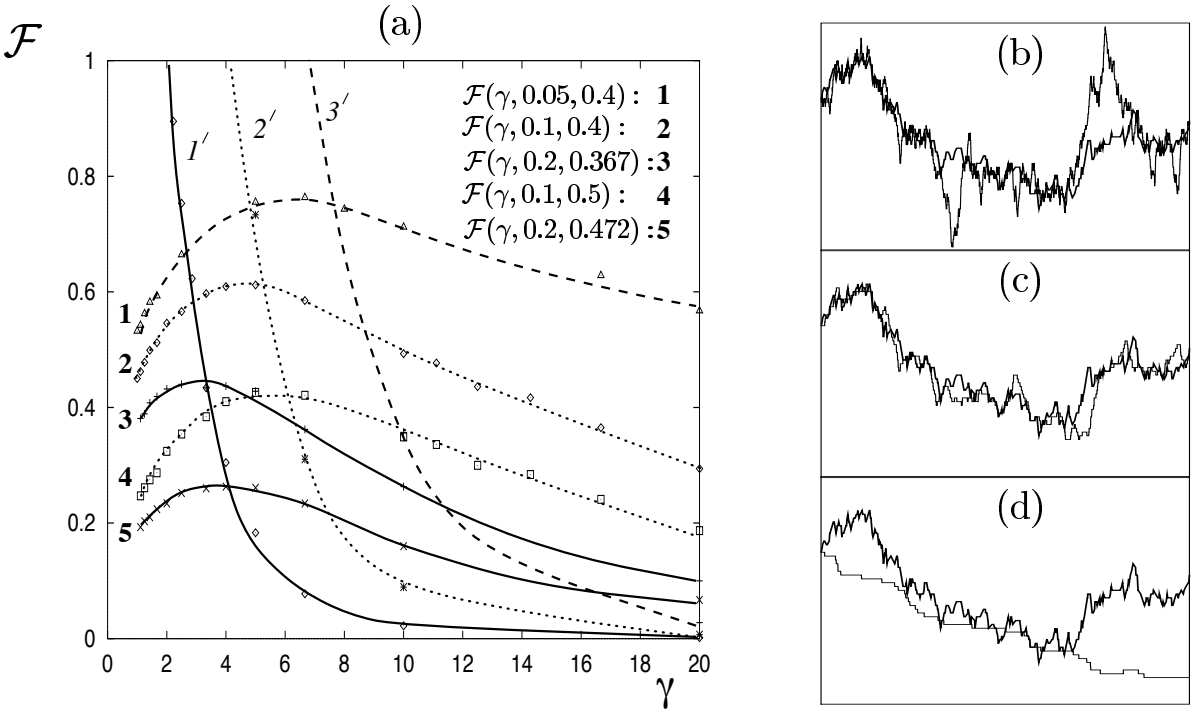

FIGURE 2: (a) The average alignment fidelity $\overline{\mathcal{F}}(\gamma; p, \tilde{p})$ obtained from a sample of 100 sequence pairs (lines 1-5), and the inverse roughness matching scale $\tilde{r}^{-1}(\gamma; \tilde{p})$ (lines *1'-3'*) for several values of p and $\tilde{p}$. The curves belonging to the same value of $\tilde{p}$ are shown in the same line style (solid: $\tilde{p} = 0.2$, dotted: $\tilde{p} = 0.1$, dashed: $\tilde{p} = 0.05$). The maximum of a fidelity curve is close to its intersection point with the roughness matching curve for the same value of $\tilde{p}$. (b-d) The optimal alignment path $r_0(t)$ (thin line) for the same sequence pair and the same target path $R(t)$ (thick line) at three different values of γ: (b) in the random fluctuation regime, (c) at the optimal value $\gamma^*(p, \tilde{p})$, and (d) in the shortcut regime.

For finite insertion/deletion rate $\tilde{p}$, it is clear that $\mathcal{F}$ should decrease to zero for sufficiently large values of γ, since a high gap penalty prevents the alignment path from following a fluctuating target path $R(t)$. The behavior of the fidelity $\overline{\mathcal{F}}(\gamma; p, \tilde{p})$ is rather complex. Fig. 2(a) shows the dependence of $\overline{\mathcal{F}}$ on γ for several values of p and $\tilde{p}$. Unlike for $\tilde{p} = 0$, these curves have their (single) maximum at a finite value $\gamma^*(p, \tilde{p})$. Alignment patterns for $\gamma < \gamma^*(p, \tilde{p})$ and $\gamma > \gamma^*(p, \tilde{p})$ are clearly distinguished by the roughness of the optimal path $r_0(t)$, as one recognizes from the examples of Fig. 2(b-d): (b) For $\gamma < \gamma^*(p, \tilde{p})$, the displacement fluctuations of $r_0(t)$ exceed those of the target path $R(t)$. This can be expressed by the relation $r_c > \tilde{r}$ with $\tilde{r}(\gamma, \tilde{p})$ given by Eq. (4). We call this regime of γ the *random fluctuation regime*. (c) For $\gamma = \gamma^*(p, \tilde{p})$, the displacement fluctuations of boths paths are seen to be of the same size, i.e., $r_c \sim \tilde{r}$. (d) In the *shortcut regime* for $\gamma > \gamma^*(p, \tilde{p})$, the dominant fluctuations are those of the target path, while the alignment path $r_0(t)$ has large straight patches with negligible intrinsic roughness.

One can show that in the asymptotic random fluctuation regime and shortcut regime, the fidelity is a monotonically increasing (decreasing) function of γ, respectively. Hence, the fidelity maximum $\mathcal{F}^*$ should obey the *roughness matching condition*

$$\mathcal{F}^* \approx \tilde{r}^{-1}(\gamma^*, \tilde{p}) \; . \tag{5}$$

As one verifies in Fig. 2(a), the maxima of the curves $\overline{\mathcal{F}}(\gamma; p, \tilde{p})$ are indeed close to the intersection points with the curves $\tilde{r}^{-1}(\gamma; \tilde{p})$ given by Eq. (4) with $A(\gamma)$ taken from Fig. 1(b).

In order to determine the optimal alignment parameter $\gamma^*(p,\tilde{p})$ from this relation, we can approximate the l.h.s. by its value in the random fluctuation regime, $\mathcal{F}^* \sim \exp(-C(\gamma^*)/U(p,\tilde{p}))$, and solve the resulting equation numerically [9]. In the biologically relevant case of a low insertion/deletion rate, where $\mathcal{F}^* \sim 1$, one obtains γ^* by expanding (5) in powers of $\tilde{p}$:

$$\gamma^*(p,\tilde{p}) \sim \tilde{p}^{-1/2}[1 + O(\tilde{p}/(1-p)^2)] \; . \tag{6}$$

In the shortcut regime, the fidelity rapidly decreases with increasing γ, making the (practical) detection of the similarities impossible for sufficiently large γ. The singularities at the theoretical detectability transition turn out to be different from the case $\tilde{p} = 0$ as well [9].

Acknowledgments

The authors are grateful to Stephen Altschul, Steven Benner, Charles Elkan, Walter Fitch, Jeff Thorne, and Michael Waterman for conversations and suggestions. TH acknowledges the financial support of an A. P. Sloan Research Fellowship, an Office of Naval Research Young Investigator Award, and the hospitality of the Max-Planck Institute at Teltow where much of the work was carried out.

References

[1] See review articles in *Met. Enz.* **183**, (1990).
[2] S.B. Needleman and C.D. Wunsch, *J. Mol. Biol.*, **48**, 444 (1970).
[3] T.F. Smith and M.S. Waterman, *Adv. Appl. Math.* **2**, 482 (1981).
[4] For a survey of recent developments, see M.S. Waterman, in *Mathematical Methods for DNA Sequences*, M.S. Waterman ed., CRC Press (1989); and M.S. Waterman, *Introduction to Computational Biology*, Chapman & Hall (1994).
[5] S.A. Benner, M.A. Cohen and G.H. Gonnet, *J. Mol. Biol.* **229**, 1065 (1993).
[6] M. Vingron and M.S. Waterman, *J. Mol. Biol* **235**, 1 (1994).
[7] T. Hwa and M. Lässig, *Phys. Rev. Lett.* **76**, 2591 (1996).
[8] M.Q. Zhang and T.G. Marr, *J. Theo. Biol.* **174**, 119 (1995).
[9] D. Drasdo, T. Hwa, and M. Lässig, Optimal Detection of Similarities in DNA Sequences, preprint (1996).
[10] M.J. Bishop and E.A. Thompson, *J. Mol. Biol.* **190**, 159 (1986).
[11] J.L. Thorne, H. Kishino, and J. Felsenstein, *J. Mol. Evol.* **33**, 114 (1991).
[12] M. Kardar, *Nucl. Phys. B* **290**, 582 (1987).
[13] T. Hwa and D.S. Fisher *Phys. Rev. B* **49**, 3136 (1994).
[14] T. Hwa and T. Nattermann, *Phys. Rev. B* **51**, 455 (1995).
[15] H. Kinzelbach and M. Lässig, *J. Phys. A* **28**, 6535 (1995).