Physica A 387 (2008) 6538–6544

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Statistical characterization of a 1D random potential problem—With applications in score statistics of MS-based peptide sequencing

Gelio Alves, Yi-Kuo Yu *

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, United States

ARTICLE INFO

Article history:
Received 5 June 2008
Available online 19 August 2008

Keywords:
Statistical significance
Dynamic programming
Mass spectrometry
Directed paths in random media
Peptide identification

ABSTRACT

We provide a complete thermodynamic solution of a 1D hopping model in the presence of a random potential by obtaining the density of states. Since the partition function is related to the density of states by a Laplace transform, the density of states determines completely the thermodynamic behavior of the system. We have also shown that the transfer matrix technique, or the so-called dynamic programming, used to obtain the density of states in the 1D hopping model may be generalized to tackle a long-standing problem in statistical significance assessment for one of the most important *proteomic* tasks—peptide sequencing using tandem mass spectrometry data.

Published by Elsevier B.V.

1. Introduction

Important in both fundamental science and numerous applications, optimization problems of various degrees of complexity are challenging (see Ref. [1] for an excellent introduction). Optimization conditioned by constraints that may vary from event to event is of especial theoretical and practical importance. As a first example, when dealing with a system under a random potential, each realization of the random potential demands a separate optimization resulting in a different ground state. The thermodynamic behavior of such a system in a quenched random potential crucially depends on the random potential realized. A similar but practical problem may arise in routing passengers at various cities to reach their destinations. In the latter case, the optimal routing depends on the number of passengers at various locations, the costs from one location to the others, which likely to vary from time to time. This type of conditional optimization also occurs in modern proteomics problem, that is, in the mass spectrometry (MS) based peptide sequencing. In this case, each tandem MS (MS^2) spectrum constitute a different condition for optimization which aims to find a database peptide or a *de novo* peptide to best explain the given MS^2 spectrum.

When the cost function of an optimization problem can be expressed as a sum of independent local contributions, the problem usually can be solved using the transfer matrix method that is commonly employed in statistical physics. A well-studied example of this sort in statistical physics is the directed polymer/path in a random medium (DPRM) [2–4]. Even when a small nonlocal energetics is involved, the transfer matrix approach still proves useful [5]. As an example, the close relationship between the DPRM problem and MS-based peptide sequencing, where a small nonlocal energetics is necessary to enhance the peptide identifications, was sketched in an earlier publication [5] and the cost value distribution from many possible solutions other than the optimal one is explored. Indeed, obtaining the cost value distribution from *all* possible solutions in many cases is harder than finding the optimal solution alone. In this paper, we will provide the solution to a generic problem that enables a full characterization of the peptide sequencing score statistics, instead of just the optimal peptide. The 1D problem considered is essentially a hopping model in the presence of a random potential. The solution to this problem may also be useful in other applications such as in routing of passengers and even internet traffic.

* Corresponding author. Fax: +1 301 480 2290.
E-mail address: yyu@ncbi.nlm.nih.gov (Y.-K. Yu).

0378-4371/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.physa.2008.08.024

In what follows, we will first introduce the generic 1D hopping model in a random potential, followed by its transfer matrix (or dynamic programming) solution. We then discuss the utility of this solution in the context of MS-based peptide sequencing, and demonstrate with real example from mass spectrum in real MS-based proteomics experiments. In the discussion section, we will sketch the utility of the transfer matrix solution in other context and then conclude with a few relevant remarks.

2. 1D hopping in random potential

Along the x-axis, let us consider a particle that can hop with a set of prescribed distances $\{m_i\}_{i=1}^{K}$ towards the positive $\hat{x}$ direction. That is, if the particle is currently at location x_0, it can move to location $x_0 + m_1, x_0 + m_2, \ldots x_0 + m_K$ in the next time step. At each hopping step, the particle will accumulate an energy $-s(x)$ from location x that it just visited. The score $s(x)$ (negative of the on-site potential energy) is assumed positive and may only exist at a limited number of locations. For locations that $s(x)$ do not exist, we simply set $s(x) = 0$ there. The energy of a path starting from the origin specified by the sequential hopping events $p \equiv \{m_{h_1}, m_{h_2}, \ldots, m_{h_L}\}$ would have visited locations $\{x_1, x_2, \ldots, x_L\}$ with $x_i \equiv \sum_{j=1}^{i} m_{h_j}$ and has energy

$$E_p(x = x_L) \equiv -\sum_{i=1}^{L-1} s(x_i) \equiv -S_p(x).$$

In general, there can be more than one path terminated at the same point. Treating each path as a state with energy given by E_p, one ends up having the following recursion relation for the partition function $Z(x) \equiv \sum_p \mathrm{e}^{-\beta E_p(x)}$

$$Z(x) = \sum_{i=1}^{K} \mathrm{e}^{\beta s(x-m_i)} Z(x - m_i), \tag{1}$$

where $\beta = 1/T$ plays the role of inverse temperature (with $k_B = 1$ chosen). If one were only interested in the best score terminated at point x, it will be given by the zero temperature limit $\beta \to \infty$ and the recursion relation may be obtained by taking the logarithm on both sides of (1) and divided by β then taking $\beta \to \infty$ limit to reach

$$S_{\mathrm{best}}(x) = \max_{1 \le i \le K} \{s(x - m_i) + S_{\mathrm{best}}(x - m_i)\}, \tag{2}$$

where $S_{\mathrm{best}}(x)$ records the best path score among all paths reaching position x. This update method, also termed dynamic programming, records the lowest energy and lowest energy path reaching a given point x. The lowest energy among all possible at position x is simply $-S_{\mathrm{best}}(x)$ and the associated path can be obtained by tracing backwards the incoming steps. It is interesting to observe that one can also obtain the worst score at each position via dynamical programming

$$S_{\mathrm{worst}}(x) = \min_{1 \le i \le K} \{s(x - m_i) + S_{\mathrm{worst}}(x - m_i)\}. \tag{3}$$

The full thermodynamic characterization demands more information than the ground state energy. In principle, one may obtain the full partition function using Eq. (1) evaluated at various temperatures. This procedure, however, hinders analytical property such as determination of the average energy

$$\langle E \rangle \equiv -\frac{\partial \ln Z}{\partial \beta}.$$

A better starting point may be achieved if one can obtain the density of states $D(E)$. In this case, we have

$$Z \equiv \int \mathrm{d}E \mathrm{e}^{-\beta E} D(E)$$

$$\langle E \rangle = \frac{\int \mathrm{d}E \mathrm{e}^{-\beta E} E D(E)}{\int \mathrm{d}E \mathrm{e}^{-\beta E} D(E)}.$$

Note that if the ground energy E_{grd} of the system is bounded from below, the partition function is simply a Laplace transform of a modified density of states given by

$$Z = \mathrm{e}^{-\beta E_{\mathrm{grd}}} \int_0^\infty \mathrm{d}E \mathrm{e}^{-\beta E} \tilde{D}(E)$$

where $\tilde{D}(E) \equiv D(E - E_{\mathrm{grd}})$ and

$$\langle E \rangle = E_{\mathrm{grd}} + \frac{\int_0^\infty \mathrm{d}E \mathrm{e}^{-\beta E} E \tilde{D}(E)}{\int_0^\infty \mathrm{d}E \mathrm{e}^{-\beta E} \tilde{D}(E)}.$$

This implies that the density of states $D(E)$ together with the ground state energy E_{grd} determine all the thermodynamic behavior of the system. In the next section, we will explain how to obtain the density of states using the dynamical programming technique as well as how to extend this approach to more complicated situations that will be useful in characterizing the score statistics in MS-based peptide sequencing.