Sampling rare events: Statistics of local sequence alignments

Alexander K. Hartmann*
Department of Physics, University of California, Santa Cruz, California 95064
(Received 11 September 2001; published 15 April 2002)

A method to calculate probability distributions in regions where the events are *very unlikely* (e.g., $p \sim 10^{-40}$) is presented. The basic idea is to map the underlying model on a physical system. The system is simulated at a low temperature, such that preferably configurations with originally low probabilities are generated. Since the distribution of such a physical system is known, the original unbiased distribution can be obtained. As an application, local alignment of protein sequences is studied. The deviation of the distribution $p(S)$ of optimum scores from the extreme-value distribution is quantified. This deviation decreases with growing sequence length.

DOI: 10.1103/PhysRevE.65.056102 PACS number(s): 05.10.−a, 87.10.+e, 87.15.−v

In many fields of physics, such as statistical physics, particle physics, or biophysics, the rare-event tails of probability distributions are studied. Here a method is presented, which allows one to obtain the probabilities down to $p \sim 10^{-40}$ for problems where the distribution is taken over a quenched disorder. As an example, the method is applied to and explained by using a biological problem, which has a high relevance for genome research.

Modern molecular biology, e.g., the *human genome project* [1], relies heavily on the use of large databases [2], where DNA or protein sequences are stored. The basic tool for accessing these databases and comparing different sequences is *sequence alignment*. The result of each comparison is a maximum alignment *score S*. One is interested either in *global* or *local* optimum alignments. For the first case, the score is maximized over all alignments of both complete sequences. The optimum local alignment is the optimum over all global alignments of all possible pairs of contiguous subsequences. To estimate the significance of the result of a comparison, one has to know, based on a random model, the statistical distribution $p(S)$ of scores. For biologically relevant models, e.g., for protein sequences with BLOSUM62 substitution scores [3] and affine gap costs [4], $p(S)$ is not known in the interesting region, where $p(S)$ is small. A number of empirical studies [5,6] for local alignment, in the region where $p(S)$ is large, suggest that $p(S)$ is an *extreme-value* (or Gumbel) distribution [7]

$$p_G(S) = \lambda e^{-\lambda(S-u)} \exp(-e^{-\lambda(S-u)}), \tag{1}$$

where u denotes the maximum of the distribution and λ characterizes the behavior for large values of S, i.e., the tail of the distribution.

In this work, to determine the tail of $p(S)$, a *rare-event simulation* is applied. For dynamical problems, such as investigating queuing systems or studying the reliability of technical components, several techniques [9] have been developed. Related methods have been introduced in physics [10,11].

By simply changing perspective, one can apply these standard techniques to many other problems. Here, the method is applied to sequence alignment. The basic idea is that one uses a physical system, which has a state given by a pair of sequences and is held at temperature T, instead of directly drawing the random sequences. This idea is similar to the simulated annealing approach [12], used to find approximate solutions of hard optimization problems. But the method presented here goes much beyond simulated annealing, because not only the minimum of one system but the whole distribution over all random instances is sampled during one run. The state of the system changes in time, governed by the rules of statistical mechanics. The energy E of the system is defined as $E=-S$. Therefore, at low temperatures the system prefers pairs of sequences with high score value S. Since the thermodynamic properties of the system are known, it is possible to extract the target distribution $p(S)$ from the measured distribution $p^*(S)$ of scores.

To fix the notations, for the alignment problem two sequences $\mathbf{x}=x_1x_2\cdots x_n$ and $\mathbf{y}=y_1y_2\cdots y_m$ over a finite alphabet with r letters are given. For DNA the alphabet has four letters, representing the bases; for protein sequences it has 20 letters, representing the amino acids. Let f_i be the probability for the occurrence of letter i, assuming here that all letters of a sequence are independent. An alignment is a pairing $\{(x_{i_k},y_{j_k})\}$ $(k=1,2,\ldots,K,\ 1\leqslant i_k<i_{k+1}\leqslant n$ and $1 \leqslant j_k<j_{k+1}\leqslant m)$ of letters from the two sequences. Note that some letters may not be aligned, i.e., *gaps* do occur. To each alignment a score is assigned, via a scoring function $s(x,y)$. The total score is the sum of scores of all aligned letters $\Sigma_k s(x_{i_k},y_{j_k})$ plus the costs of all gaps. Here, so called *affine* gap costs (α,β) are considered, i.e., a gap of length l has costs $g(l)=-\alpha-\beta(l-1)$. The optimum *global* alignment $S_g(\mathbf{x},\mathbf{y})$ is obtained by maximizing the score over all values of K and over all possible placements of the gaps. The optimum *local* alignment S is the maximum over all possible contiguous subsequences $\tilde{\mathbf{x}}=x_ix_{i+1}\cdots x_{i+l-1}$, $\tilde{\mathbf{y}}=y_jy_{j+1}\cdots y_{j+k-1}$ of the optima $S_g(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$. Hence, since an alignment of zero length has score zero, the optimum local alignment is always non-negative by definition. For both global and local alignment efficient algorithms [13,4] exist, which calculate an optimum alignment in time $O(nm)$.

*Email address: hartmann@bach.ucsc.edu

Hence, one can easily generate, e.g., $N \approx 10^5$ samples of pairs of sequences according the frequencies f_i, obtain each time the optimum alignment, and calculate a histogram of the optimum scores S. This *simple sampling* allows one to calculate $p(S)$ in the region where the probabilities are large [e.g., $p(S) \approx 10^{-4}$]. Recently, the *island method* [14] was introduced, which allows a speed up of several orders of magnitudes for very long sequences, but still the far end of the distribution is out of reach. Also, please note that biologically relevant protein sequences have lengths of few hundred amino acids, as studied in this paper.

As already sketched, to determine the behavior of $p(S)$ at the rare-event tail [e.g., $p(S) \approx 10^{-40}$], one views each pair $c=(\mathbf{x},\mathbf{y})$ of sequences as the state of physical system, which behaves according the rules of statistical mechanics, with $-S$ being the energy of the system. More precisely, instead of considering many independent pairs of fixed sequences, a Markov chain [15] $c(0) \to c(1) \to c(2) \to \cdots$ of pairs is used to generate the instances. For each instance $c(i)$ the optimum local alignment score S is calculated. Below $p(c \to c')$ denotes the transition probability from state c to state c'. Changing the sequences dynamically is similar to annealed disorder simulations [8]. *But*, while the physics of an annealed system is different from the physics of the related quenched system, here an annealed-disorder-like simulation is used via applying a simple transformation (see below) to obtain the *true* behavior of the quenched system.

The simplest rule for the transition is, to choose randomly a position in one of the sequences with all positions being equiprobable and to choose randomly a new letter from the alphabet, the letters having probabilities f_i, i.e., $p(c \to c') = f_i/(n+m)$ if c,c' differ by at most one letter, and $p(c \to c')=0$ otherwise. With this choice of the transition probabilities, for $t \to \infty$ all possible pairs of sequences have the probability $P(c)=\Pi_i f_{x_i} \Pi_j f_{y_j}$ of occurring. Hence, simple sampling is reproduced.

To increase the efficiency, one can change the sampling distribution [9,16], a standard method for simulating rare events [17], which allows one to concentrate the sampling in small regions in configuration space. A good choice for the transition rule of the sequence-alignment problem is first to change one position of one sequence randomly as above, recalculate the optimum alignment $S(c')$ with a standard algorithm and accept this move $c \to c'$ with the Metropolis probability [16] $\max[1,\exp(\Delta S/T)]$, where $\Delta S = S(c') - S(c)$. This leads to the equilibrium state of a physical system at temperature T with energy $E=-S$, with the distribution weighted by the sequence probabilities $P(c)$. The advantage of this approach is that the equilibrium distribution $Q(c)$ is known from statistical physics [18]: $Q(c) = P(c)\exp[S(c)/T]/Z$ with $Z(T)=\Sigma_c P(c)\exp[S(c)/T]$ being the partition function. Thus, the estimator for the probability to have score S in the biased ensemble is

$$p^*(S) = \frac{\exp(S/T)}{Z(T)} \sum_c{}' P(c), \tag{2}$$

where the sum Σ' runs over all sequences with score S. Thus, from the measured histogram of scores $p^*(S)$ one obtains the estimator for the unbiased distribution through $p(S)=\Sigma_c' P(c) = p^*(S) Z(T) \exp(-S/T)$. $Z(T)$ is unknown *a priori*, but can be determined very easily, as shown below.

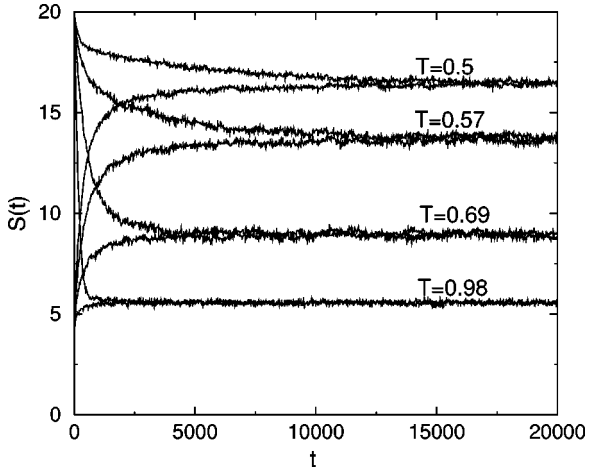

FIG. 1. Average alignment score S as a function of step t for $n,m=20$, 4 letters, local alignment without gaps for different temperatures T. For each temperature, 1000 simulations were started with two random sequences (low scores) and 1000 simulations with two equal sequences (high scores).

Note a striking difference from conventional Monte Carlo (MC) simulations of random systems. For the conventional approach, different samples with quenched disorder are studied by MC simulations, each sample having the same probability. Within the method presented here, a biased simulation is done *on the disorder*, while the behavior of each random sample is determined exactly, resulting finally in the unbiased distribution over the disorder.

To describe the behavior of $p(S)$ over a wide range, the model must be simulated at several temperatures. For this reason, and to increase the efficiency, the model is simulated via the *parallel tempering* method [19]. Using this technique, the system is simulated at N_T different temperatures $T_1 < T_2 < \cdots < T_{N_T}$ in parallel, i.e., with N_T independent pairs $c(T_i)$ of sequences. The main idea of parallel tempering is that from time to time the configurations between neighboring temperatures T_i, T_{i+1} are exchanged according a probabilistic rule [19]. Here, each simulation step consists of one Markov step for each configuration c and one exchange step between one neighboring pair $c(T_i), c(T_{i+1})$.

Next, a simple example is given, illustrating how the method works. Optimum local alignments without gaps for sequences of equal length $m=n=20$ and $r=4$ letters, all having the same probability 1/4, are calculated. For the test the following score is applied: $s(x,y)=1$ if $x=y$ and $s(x,y)=-3$ otherwise. Two types of runs are performed: (a) initially, all pairs of sequences are random, and (b) initially, each pair consists of two equal sequences. Thus, for the first type, initially the score is low, while for the second type the score is initially maximal. This provides a criterion for equilibration: if the average score for both initial configurations agrees within error bars (at time t_0), the simulation is long enough. In Fig. 1 the average optimum score S for the beginning 10% of the running time of 1000 independent runs

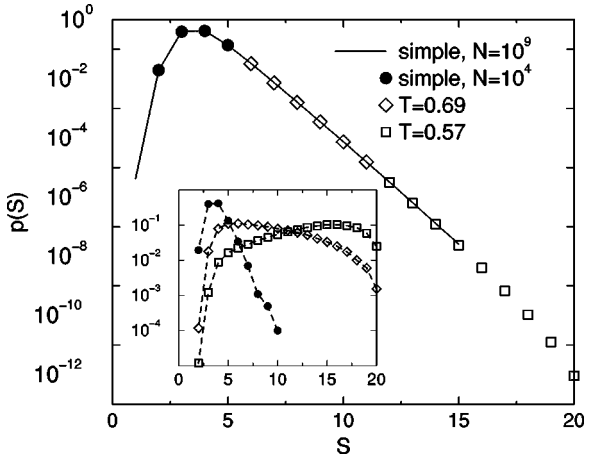

FIG. 2. Rescaled distribution $p(S)$ for the direct simulation and for $T=0.57$, $T=0.69$ for $n,m=20$, 4 letters, local alignment without gaps. The solid line is the result of a large simple sampling simulation with $N=10^9$ samples. Inset: Raw distribution of alignment scores S for the direct simulation and for $T=0.57$ and $T=0.69$.

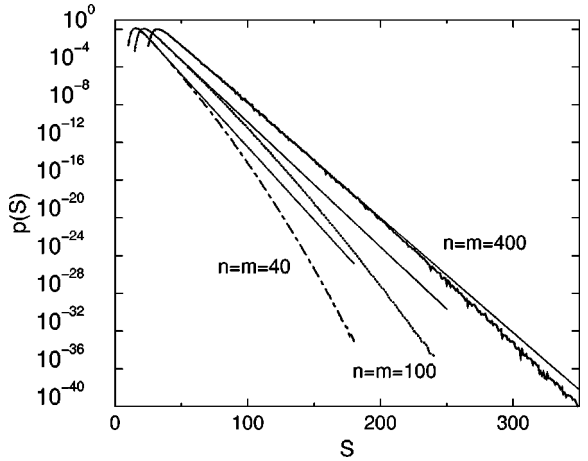

FIG. 3. Distribution of alignment scores S for $L=40,100,400$, BLOSUM62 substitution matrix and affine (12,1) gap costs. The thin solid lines are fits to extreme-value distributions with parameters (λ,u), yielding $(\lambda,u)=(0.355(5),15.35(4))$ $(n,m=40)$, $(0.304(2),21.67(4))$ $(n,m=100)$, and $(0.280(3),32.01(3))$ for $n,m=400$.

and four different temperatures T is shown.

To obtain weakly correlated samples, only values at t_0, $t_0+\tau$, $t_0+2\tau$, etc., are taken, where τ is the characteristic time in which the score-score correlation $c_S(t_0,t)=(\langle S(t_0)S(t)\rangle-\langle S\rangle^2)/(\langle S^2\rangle-\langle S\rangle^2)$ decreases to $1/e$.

In the inset of Fig. 2 the raw distribution of S for two temperatures is shown together with a distribution from a simple sampling of $N=10^4$ realizations. Clearly, with the statistical mechanics approach, the region of high scores is sampled much more frequently.

For low scores, the final distributions obtained from the simple sampling and from the finite-temperature simulation must agree. This can be used to determine the constant $Z(T)$. It is chosen such that the difference in an interval $[S_1,S_2]$ between the simple sampling distribution and the rescaled distribution at T is minimal. In the same way $Z(T)$ at lower temperatures can be obtained by matching to distributions obtained before at higher temperatures. The final distribution is shown in Fig. 2. For each data point, the distribution with the highest accuracy was taken. For comparison, a simple sampling distribution obtained using a huge number of samples ($N=10^9$) is shown. Both results agree very well. Note that the distribution from the finite-T approach spans almost the entire interval $[0,20]$. In principal, the region for very small score S can be investigated also using the given method by simulating at *negative* temperatures. How powerful the given method can be seen by looking at the right border of the interval, where a value $p(20)=9.13(20)\times10^{-13}$ was obtained. This agrees within error bars with the exact result [21] $0.25^{20}\approx9.09\times10^{-13}$. Also the same model with (3,1) gap costs was tried and again a perfect agreement with a huge simple sampling simulation was found. This example illustrates that the method presented here is indeed able to calculate accurately the distribution $p(S)$ of optimum alignment scores in regions where $p(S)$ is very small.

Next, the results for a biologically relevant case are presented. Sequences of amino acids distributed according the background frequencies by Robinson and Robinson [20] are used together with the BLOSUM62 scoring matrix [3] for (12,1) affine gap costs. This type of system has been studied in Ref. [6] in the region where $p(S)$ is large. Here, sequences of length $n=m$ in the range $[40,400]$ were considered. The simulations were performed for $n_T=7$ temperatures $T\in[2\cdots10]$ ($[3.5\cdots10]$ for $n,m=400$), with up to 100 independent runs of lengths up to $t_{\max}=4\times10^5$ steps. To test for equilibration, it was again checked whether simulations starting at random states (low score) and starting at ground states (maximum possible score) converged to the same average energy. For the lowest temperatures it was not possible to equilibrate the system within the given time. The reason is that near $T\approx1/\lambda$ the equilibration times seem to diverge. This indicates a phase transition in the physical system with (probably) a glassy phase at low temperatures. Hence, for the evaluation only data from those temperatures were used, where equilibration could be guaranteed.

In Fig. 3 the distributions $p(S)$ of optimum alignment scores are shown. To obtain the same accuracy with a simple-sampling approach, given a computer that optimizes say 10^6 samples per second, a total simulation time of about 2.5×10^{17} times the age of the universe would be necessary. Also shown in Fig. 3 are fits of the low-score data to Gumbel distributions. The resulting parameters (λ,u) are comparable to the values found [6] before and depend slightly on the sequence length. For high scores, significant deviations from the pure Gumbel behavior are visible, in contrast to the earlier predictions. Since the deviations occur at high score values, they could not be detected before using conventional methods. The reason for the deviations is edge effects: very long alignments cannot start near the end of either of the sequences, so they become even more unlikely. The results found here can be fitted very well to *modified* Gumbel distributions of the form

$$\tilde{p}_G(S) = k\lambda e^{[-\lambda(S-u)-\lambda_2(S-u)^2]}\exp(-e^{-\lambda(S-u)}), \quad (3)$$

with $k \approx 1$, resulting in values for (λ, λ_2, u) of $(0.3277 \pm 0.0003, 8.56 \times 10^{-4} \pm 3 \times 10^{-6}, 15.35 \pm 0.04)$ for $n,m=40$, $(0.2783 \pm 0.0003, 1.72 \times 10^{-4} \pm 1 \times 10^{-6}, 21.67 \pm 0.04)$ for $n,m=100$, and $(0.2733 \pm 0.0004, 6.1 \times 10^{-5} \pm 2 \times 10^{-6}, 32.01 \pm 0.03)$ for $n,m=400$. Anyway, with increasing lengths n,m, on a scale of scores proportional to $u \sim \ln n$, $p(S)$ approaches the Gumbel distribution more and more, i.e., $\lim_{n\to\infty}\lambda_2 = 0$.

To summarize, a method has been presented, which allows one to study rare events in random systems down to regions of *very low* probabilities. The basic idea is to interpret the probability space as the phase space of a physical system. From the distribution of states, the original unbiased distribution can be obtained. The method is applied to a biologically relevant case of the local sequence-alignment problem. The distribution $p(S)$ can be studied in regions where the probability is as small as 10^{-40}, and yet the deviations of the distribution from the theoretical prediction are visible.

The author developed the idea for this method at the workshop ''Statistical Physics of Biological Information'' at the Institute for Theoretical Physics in Santa Barbara during discussions with P. Grassberger and E. Marinari. The author would like to thank A.P. Young and P. Grassberger for critically reading the manuscript and interesting discussions and A.P. Young also for additional support. The simulations were performed on a Beowulf Cluster at the Institut für Theoretische Physik of the Universität Magdeburg with technical support from S. Mertens and H. Bauke. Financial support was obtained from the DFG (Deutsche Forschungsgemeinschaft) under Grant No. Ha 3169/1-1.

[1] International Human Genome Sequencing Consortium, Nature **409**, 860 (2001).
[2] S. M. Brown, *Bioinformatics* (Eaton, Natick, MA, 2000); H. H. Rashidi and L. K. Buehler, *Bioinformatics Basics* (CRC Press, Boca Raton, FL, 2000).
[3] S. Heinkoff and J.G. Heinkoff, Proc. Natl. Acad. Sci. U.S.A. **89**, 10 915 (1992).
[4] O. Gotoh, J. Mol. Biol. **162**, 705 (1982).
[5] T.F. Smith, M.S. Waterman, and C. Burks, Nucleic Acids Res. **13**, 645 (1985); J.F. Collins, A.F.W. Coulson, and A. Lyall, CABIOS, Comput. Appl. Biosci. **4**, 67 (1988); R. Mott, Bull. Math. Biol. **54**, 59 (1992); M.S. Waterman and V. Vingron, Proc. Natl. Acad. Sci. U.S.A. **91**, 4625 (1994); Stat. Sci. **9**, 367 (1994).
[6] S.F. Altschul and W. Gish, Methods Enzymol. **266**, 460 (1996).
[7] E. J. Gumbel, *Statistics of Extremes* (Columbia University Press, New York, 1958).
[8] See, e.g., R. B. Stinchcombe, in *Phase Transitions and Critical Phenomena*, edited by C. Domb and M. S. Green (Academic, New York, 1983), Vol. 7; P.J. Shah and O.G. Mouritsen, Phys. Rev. B **41**, 7003 (1990); R. Voča, Phys. Rev. E **60**, 3516 (1999).
[9] B. D. Ripley, *Stochastic Simulation* (Wiley, New York, 1987).
[10] B.A. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992).
[11] P. Grassberger, Phys. Rev. E **56**, 3682 (1997).
[12] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, Science **220**, 671 (1983).
[13] S.B. Needleman and C.D. Wunsch, J. Mol. Biol. **48**, 443 (1970); T.F. Smith and M.S. Waterman, *ibid.* **147**, 195 (1981).
[14] S.F. Altschul, R. Bundschuh, R. Olsen, and T. Hwa, Nucleic Acids Res. **29**, 351 (2001).
[15] S. Lipschutz and M. Lipson, *Probability* (McGraw-Hill, New York, 2000).
[16] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[17] V. Kalashnikov, *Geometric Sums: Bounds for Rare Events with Applications* (Kluwer Academic, Dordrecht, 1997).
[18] L. E. Reichl, *A Modern Course in Statistical Physics* (Wiley, New York, 1998).
[19] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992); K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).
[20] A.B. Robinson and L.R. Robinson, Proc. Natl. Acad. Sci. U.S.A. **88**, 8880 (1991).
[21] Only when $\mathbf{x}=\mathbf{y}$ the highest posssible score is achieved.