Monte Carlo Studies of Finite Temperature Sequence Alignment

B.Alver, O.Ozcan

Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA

We report results from a Monte Carlo study of gapped local sequence alignment of two polymers at "finite temperature." The dependance of the "free energy" score on temperature and degree of homology between the polymers is investigated. We observe a phase transition and an increased sensitivity of the free energy on the degree of homology slightly above the phase transition point. The signifance of the match is compared to a global alignment algorithm for globally homologous proteins.

I. INTRODUCTION

Standard algorithms of sequence alignment work on the principle of maximization of a score, calculated with a given scoring matrix [1–4]. There are also various methods developed to calculate the significance of a given match, which is related to the inverse probability of the matching proteins being random. Maximum score alignments are very successful in identifying closely related sequence pairs. However, for sequence pairs with less correlations, these algorithms become sensitive to the selection of scoring matrix, which is difficult to optimize.

Finding the maximum score for given sequences is analogous to finding the minimum energy state of directed polymers [5]. Directed polymers have also been studied at finite temperatures [6]. This problem can also be translated to a problem of sequence alignment at finite temperature [7–9].

It has been argued that finite temperature sequence alignment is less sensitive to the exact scoring parameters [9]. Finite temperature alinment allows estimation of single element pair reliability [7] and of relative significance of different high-scoring alignments [7, 9]. It has been suggested that the analogies to physical systems such as electro-optic circuits may be used to perform finite temperature alignment calculations very quickly [8].

In this paper, we investigate an algorithm based on finite temperature sequence alignment. We test the ability of the algorithm on Monte Carlo generated sequences to differentiate between random and correlated pairs.

II. ALGORITHMS

We use the standard local alignments scoring scheme. We place each of the protein sequences on the axes of a grid and interprete each alignment as a directed path on this grid. The score for a path $\mathcal{A}$ is defined as

$$\mathcal{S}(\mathcal{A}) = \alpha N_+ + \beta N_- + \gamma N_g, \tag{1}$$

where N_+, N_-, and N_g are the numbers of matches, mismatches and gaps respectively.

Similar to the method described in [7], we convert the score of each directed path to a Boltzman factor, and construct a free energy for the ensemble of directed paths as a function of temperature. Thus, the free energy is defined as

$$F(T) = T \ln[\sum_{allpaths \mathcal{A}} e^{\frac{\mathcal{S}(\mathcal{A})}{T}}]. \tag{2}$$

It can be verified that at the $T = 0$ limit, this free energy converges to the maximal score since the maximal score has the highest Boltzman weight.

To accurately take into account contributions from each path and for time efficiency concerns, we use a recursive method similar to the one described in [6]. This leads us to take into account some null paths including only gaps or ones that have many gaps at either end of the alignment. Although exponentially suppressed by the gap punishment, their effect on the calculation, due to the large combinatorics, is hard to estimate, especially at high temperatures. In a recursive calculation, it is nontrivial to remove paths that end with gaps. However, it is possible to take into account paths that start without a gap. Using two different methods, one including, the other excluding the paths that start with gaps allows us to understand the effects of null paths.

A. Method I

Let $p_1 = a_i$, and $p_2 = b_j$ be two proteins with amino-acid sequences indexed by i, and j respectively. We define the partition function at each vertex

$$\begin{aligned} Z(k,k) = Z(k,-k) &= 0, \forall k \\ Z(r,t) &= e^{s(i,j)}[Z(r-2,t)+1] \\ &\quad + e^{\gamma} Z(r-1,t-1) \\ &\quad + e^{\gamma} Z(r-1,t+1)] \end{aligned} \tag{3}$$

where $r = j + i$, $t = j - i$, and $s(i,j)$ is given as

$$s(i,j) = \begin{cases} \alpha & \text{if } a_i = b_j \\ \beta & \text{else (mismatch)} \end{cases} \tag{4}$$

$Z(r,t)$ is the sum of Boltzman factors associated with all the paths that start with a match or mismatch from

$r' < r$ and end at (r, t). 1 is added into the recursion to account for the paths that start at $(r-2, t)$.

Thus, summing local partition function, (3), over the available r, and t values, we can find the total partition function,

$$Z = \sum_{r,t} Z(r,t). \tag{5}$$

B. Method II

For the second method, we define the local partition function

$$\begin{aligned} Z(k,k) &= Z(k,-k) = 1, \forall k \\ Z(r,t) &= 1 + e^{s(i,j)} Z(r-2,t) \\ &\quad + e^{\gamma} Z(r-1,t-1) \\ &\quad + e^{\gamma} Z(r-1,t+1)] \end{aligned} \tag{6}$$

where r, t, and $s(i,j)$ are defined as the ones in the first method.

Again partition function at (r,t), (3), is the sum of Boltzman factors associated with all the paths that start from $r' \leq r$ and end at (r,t). The term 1 is now added into the recursion to account for all the paths that start at (r,t) including ones starting with a gap.

In this case, total partition function is a slightly modified sum:

$$Z = \sum_{r,t} [Z(r,t) - 1]. \tag{7}$$

By subtracting 1 from each local partition function, we excluded the contribution of the paths of length 0, i.e., the paths that start and end at the same point.

The difference between free energy values obtained with the two different methods can not be distinguished from the differences of different alignments in either of the methods. Furthermore there is no observable difference in the distribution over many pairs in the temperature range that we describe below. Therefore we report results with only one of the algorithms, method I.

III. SIMULATIONS

One possible question that our algorithm should be able to answer is, given two sequences, whether we can identify if they are correlated or if they are completely random. We simulate a "sequence" as an array of random integers between 1 and 20. We use sequences of length 1000. Then we simulate a second correlated sequence, where with a probability, p, the content is copied from the the first sequence and with a probability $1-p$, the content is completely random. Furthermore, at each site, an amino acid is removed with a probability g or a random one is added with a probability g, leading to a gap probability of $2g$ in either sequence. The length of second sequence is also adjusted to be 1000 by trimming or random addition. We have generated 25000 pairs for each (p, g) set and evaluated their free energy at 25 different T values ranging from 0.1 to 1.9. In our numerical simulations, we use $\alpha = 1$, $\beta = -1$, $\gamma = -2$.

For different values of the parameters p and g that we have investigated, we observe a phase transition in the free energy versus temperature plots. The phase transition appears to take place at a similar, $T = T_C$, for different values of the parameters. Below this phase transition, free energy, F, increases exponentially with T, while for $T > T_C$, the behavior is linear. (See Fig. 1)

Comparing results for different values of the parameters p and g, given with different colors in Fig. 1, more significant difference is observed at and slightly above T_C. In these plots, black data points are for completely random proteins. The standard deviation of F over the 25000 pair sample is also shown.

We quantify the significance of the measurement as the distance of F for finite (p, g) from the F observed at $p = 0$, namely F_0, normalized by the standard deviation, σ_0. In Fig. 1 $d) - h)$, we plot this quantity for different values of p and g. On these figures, are also shown the standard deviation of F for finite (p, g), normalized by σ_0. We observe huge fluctuations around $T = 1.3$ and maximal significance at slightly higher temperatures.

To answer the question of global similarity between two proteins, which we by construction know have global correlations, one could simply use global alignment at zero temperature. We have compared the significance of our algorithm with the significance of a maximal score global alignment calculated in a similar fashion. The results are shown in Fig. 1 $d) - h)$, with smaller size symbols at $T = 0$.

While mutations are equally detectable for finite temperature local alignment and zero temperature global alignment, the sensitivity to evolution with insertion-deletion seems to be enhanced in finite-temperature alignment. This point is even more strongly observed in Fig. 2, where we plot the distribution of F at fixed values of temperature. Note the change in the position of the peaks colored in orange and red, from plot $d)$ to $c)$.

IV. CONCLUSION

We have studied a finite temperature gapped local sequence algorithm. The temperature dependence of free energy shows the existence of a phase transition. We have quantified the significance of an alignment of correlated pairs as the difference from alignment scores from random pairs in standard deviations. With this measure, we observe high fluctuations in significance at the phase transition temperature. The significance is maximum slightly above this point.

Finite-temperature alignment gives roughly similar sig-

nificance as zero temperature global alignment method. On the other hand, on different types of evolution algorithms, it has the potential to better the maximum score alignment. This can be tested on structurally similar proteins with no apparent sequence similarity. It can be checked whether the same phenomenon as we observed for insertion/deletion evolution applies to this situation. In order to establish this, one can give structurally similar proteins to BLAST and compare the significance with that of the finite-temperature alignment method, or find structurally similar proteins with average significance, and see if our method can differentiate the alignment from a random one.

Most of the significance tests are highly sensitive to the scoring matrix, therefore a more detailed examination of scoring variables might yield interesting results. On the other hand, we believe that the characteristics of finite-temperature alignments, like phase transition and significance enhancement for regularized mutations will persist for most of the reasonable scoring variables. It is still an interesting question if the scoring matrix element can be optimized for finite-temperature alignments.

In this paper, we focused on the comparison of pairs at fixed values of temperatures. More information may be extracted from the temperature dependence of F for a given sequence pair. For example, a higher significance might be obtained by integrating $(F - F_0)/\sigma_0$ in some region above the phase transition. Or the exact location of the phase transition for a given pair might depend on the homology of the sequences.

We conclude that finite temperature sequence alignment is a rich tool that may turn very useful for specific questions in the area of sequence alignment.

[1] S.B. Needleman and C.D. Wunsch *J. Mol. Biol.* 48, 443-453 (1970)

[2] T.F. Smith and M.S. Waterman *J. Mol. Biol.* 147, 195-197 (1981)

[3] Samuel Karlin, Stephen F. Altschul, *Proc. Natl. Acad. Sci. USA* Vol. 87, pp. 2264-2268, March 1990, Evolution

[4] Stephen F Altschul *et al.*, *Nucleic Acids Research*, 1997, Vol. 25, No. 17 3389-3402

[5] M. Kardar and Y.C. Zhang, *Phys. Rev. Lett.* 58, 2087 (1987)

[6] J.M.Kim, A.J. Bray and M.A. Moore, *Phys. Rev. A* 44, 8 (1991)

[7] Maik Kschischo, Michael Lässig, *Finite-temperature Sequence Alignment*, Pacific Symposium on Biocomputing 5:621-632 (2000)

[8] M.Q. Zhang and T.G. Marr *J. Theo. Biol.* 174, 119-29 (1995)

[9] S. Miyazawa, *Protein Eng.* 8, 999-1009, (1996)

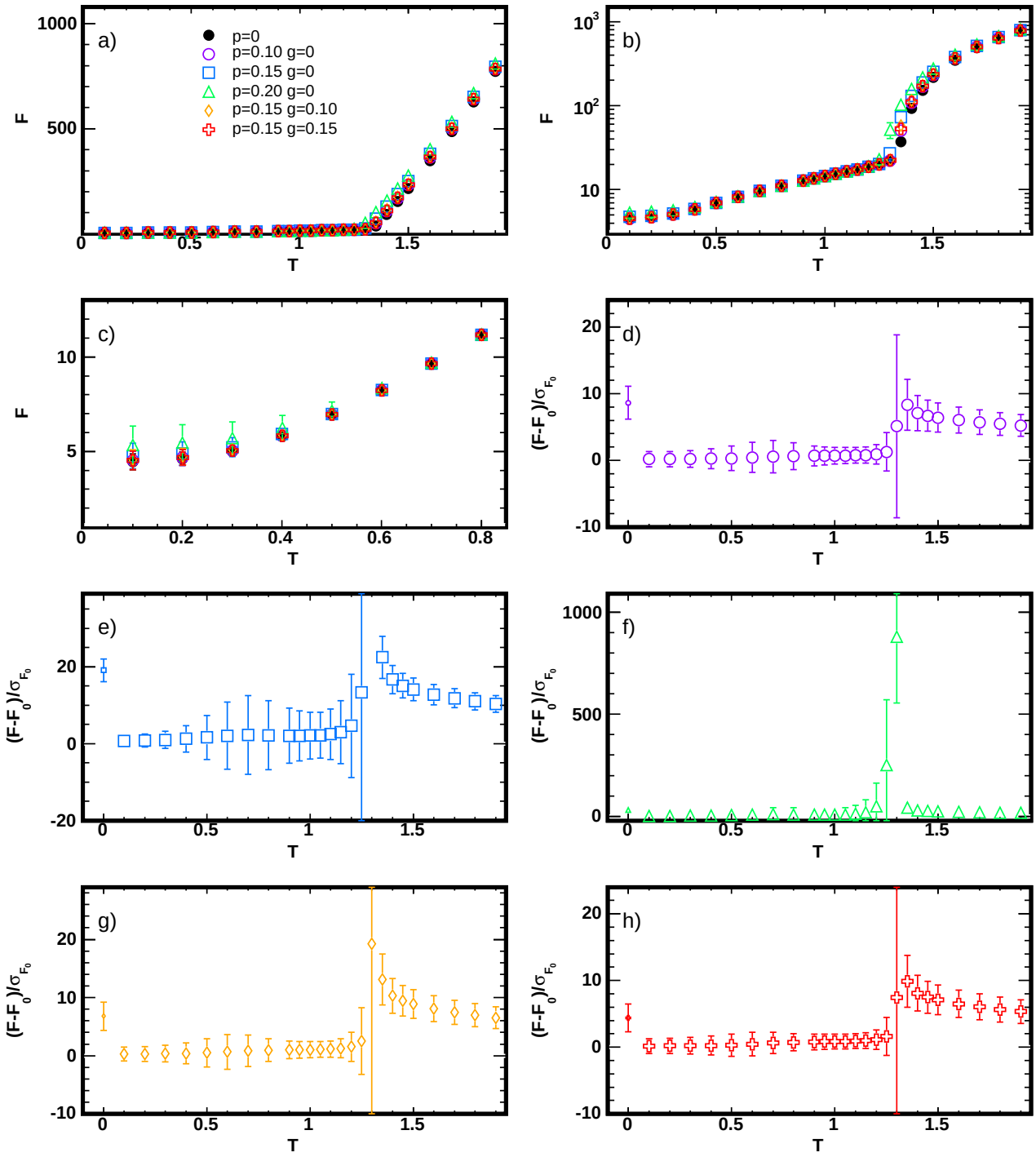

FIG. 1: **a:** Large scale behavior of $F(T)$ in linear scale. Each case is done with 25000 random protein pairs. Error bars reflect one sigma standard deviation. **b:** Same plot in logarithmic scale. **c:** Low T behavior of F. **d-h:** Difference of $F(T)$ for $p = 0.10$, $p = 0.15$, $p = 0.20$, $(p, g) = (0.15, 0.10)$ and $(p, g) = (0.15, 0.15)$, from $F(T)$ for $p = 0$, normalized by the standard deviation for $p = 0$. $T = 0$ data is for global alignment and $T > 0$ points are for local alignment.

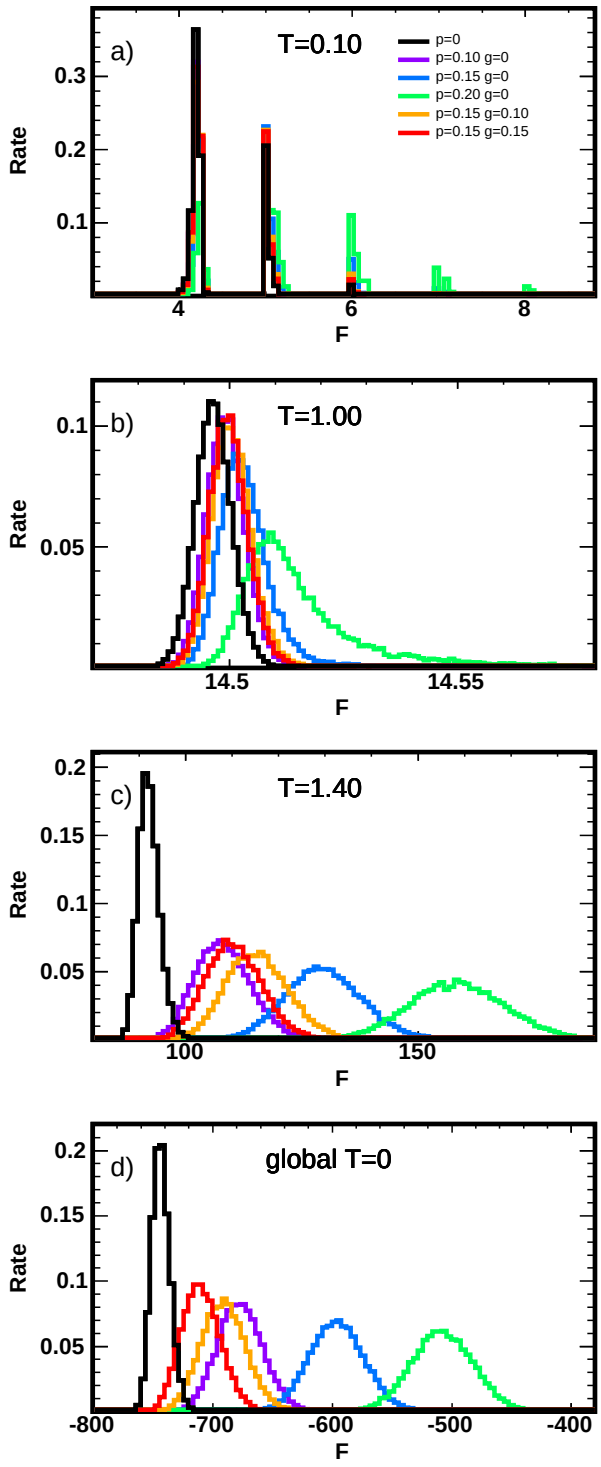

FIG. 2: **a-c:** Distribution of $F(T = 0.10)$, $F(T = 1)$ and $F(T = 1.4)$ for different homology runs. **d)** Distribution of $s = F(T = 0)$ for global alignment.