Phase Transition in Sequence Unique Reconstruction

Chan Zhou[1*] and Li Xia[2*]

[1]James D. Waston Institute of Genome Sciences, Zhejiang University, Hangzhou 310008, China
zhouchan99@zju.edu.cn

[2]T-Life Research Center, Fudan University, Shanghai 200433, China
lixia@fudan.edu.cn

Written September 11, 2005; Revised September 15, 2005

Abstract. In this paper, we show a kind of phase transition phenomenon in the probability of uniquely reconstructable sequences (URS) under equal-probability independently and identically distributed (eiid) model and non-equal-probability independently and identically distributed (niid) model, respectively. This URS probability is calculated, with the relative error approximate 1%, by doing Monte Carlo experiments. In the Monte Carlo experiments, we use a deterministic finite automaton (DFA) to determine whether a symbolic sequence can be uniquely reconstructed or not from all its substrings of length K (called K-tuples). Furthermore, we compare our experiment results with the real protein sequences to identify a possible biological implication.

Keywords: Phase transition, protein sequence, unique reconstruction, probability, SBH

1. Introduction

Given a symbolic sequence S over an alphabet Σ of length L, we can easily obtain a multiset of its all K-tuples[†] by sliding a K-sized window one letter by one letter. But can we uniquely reconstruct the original sequence from this multiset? A simple question but defies any intuitive answer.

Take the following 10bp long DNA sequence for example: S =TGTGT ATGTC. The multiset of all 3-tuple for this sequence is {TGT, GTG, TGT, GTA, TAT, ATG, TGT, GTC}. After carefully examination, one may construct "TGTAT GTGTC" from the above multiset. It's a different one which share the same 3-tuples with the original sequence. But how about $K = 4$ or $K = 5$? The answer is that "TGTGT ATGTC" can be uniquely reconstructed from its 5-tuples multiset, but not for $K = 4$. In contrast, another sequence "AAAAA AAAAA" is uniquely determined by the multiset of its K-tuples at any integer K value, as long as K is no more than its length 10.

As a matter of fact, how to check whether a sequence can be uniquely reconstructed by its all K-tuples is the core question of sequencing by hybirdization (SBH)[‡] [8,10]. Hence, it has caught many researchers' attention. Pevzner [8] reduced this question to the Eulerian path problem. Consequently, the Best theorem [3,4], an formula for the number of Eulerian cycles in a directed graph, can tell us the number of reconstructed sequence from a K-tuples multiset after transforming the original sequence into an Eulerian graph (ref. to [4] for more details). Obviously, if and only if the number of reconstructed sequence equals one, the original sequence could be uniquely reconstructed. Actually, the Best formula gives much more information other than whether a sequence is uniquely reconstructed or not. In addition, Pevzner [8] and Kontorovich [5] showed the sufficient and necessary condition for a uniquely reconstructable sequence, from different views. Along this line, Li and Xie [6] recently proposed an effective algorithm of a deterministic finite automata (DFA) to examine if a sequence can be uniquely reconstructed or not.

---

[*] These two authors contributed to this work equally. [†] In this paper, any substring with length K is referred as a K-tuple. [‡] SBH is a method to sequence DNA.

Easy to see from the previous examples that not all sequences can be uniquely reconstructed for a given L and K. Naturally, we are interested in the probability that a random sequence of length L can be uniquely reconstructed from the multiset of its all K-tuples, under independently and identically distributed model (iid[§]). Or, in other words, what's the proportion of the uniquely reconstructable sequences (URS) chosen uniformly (i.e. eiid) or nonuniformly (i.e. niid) at random from Σ^L? Dyer [2] and Arratia [1] have proved the asymptotic limiting probability as $L \to \infty$ for this problem. The rub is that their result is acceptable only for very large L, and the error bound is not stable, sometimes rather large.

Instead, we implement a DFA to check if a symbolic sequence is a URS or not, no matter how long the sequence is and regardless of the given alphabet Σ. Then we do Monte Carlo experiments to compute the probability of uniquely reconstructable sequences, with a stable relative error bound which approximate 1% (see Section 2 for more details). For this reason, it's more reasonable to figure out the uniquely reconstructable probability by our methods for any sequence length L, especially efficient for L no more than thousands, which is much closer to real proteins.

Since proteins play an important role in organisms. Here we focus our work on amino acid sequences, in comparison to previous works, which are mainly done on DNA sequences. Most interestingly, we observe a phase transition like phenomenon, regarding to the K value, in URS probability problem. At the same time, we propose formulas to estimate the critical point for the phase transition phenomenon.

In addition, we compare our experiment results with the real protein sequences to discover a possible biological meaning.

All the results mentioned above are shown in Section 3 and discussed in Section 4.

2. Method

2.1. Deterministic Finite Automaton (DFA)

We implement a DFA which will accept and only accept the uniquely reconstructable sequences, according to the algorithm described in [6]. This DFA can work on any symbolic sequence, regardless of the alphabet Σ, sequence length L and the tuple size K. It reads through the input sequence letter by letter until it meets a certain kind of spatial pattern which lead the sequence to be non-uniquely reconstructable. Loosely speaking, this spatial pattern is an interleaved pair of repeated $(K-1)$-tuples (i.e. $\cdots a \cdots b \cdots a \cdots b$, where a, b denote $(K-1)$-tuples) or triple repeats (i.e. $\cdots a \cdots a \cdots a \cdots$, where a denotes a $(K-1)$-tuple) [1,4,6,8].

2.2. Monte Carlo (MC) experiments

We do Monte Carlo experiments to compute the probability of uniquely reconstructable sequences under eiid model and niid model, respectively. 100 Monte Carlo experiments are carried out to calculate one probability for a specified K, L. Among each Monte Carlo experiment, our C++ program samples 1000 sequences from the sample space Σ^L, according to the distribution of Σ^L. 1000 sample points for each experiment are enough, since under this condition the relative error is about 1%, which is acceptable. (e.g. see Figure 1). We also draw the absolute error bar in the probability vs. K graphs (Figure 3).

[§] The iid model includes both eiid model and niid model.

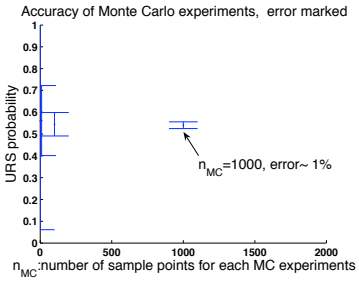


Figure 1: The bar represents the absolute error for probability of uniquely reconstructable sequences (URS) at $K = 5, L = 1000$, under eiid model, where 100 Monte Carlo experiments are taken. As n_{MC}=1000, the relative error reduces to 1%, which is acceptable.

2.3. Least Square Method

We use least square methods (LSM) [7] to derive the relationship between the critical point of K (defined rigorously in Section 3.2) and sequence length L.

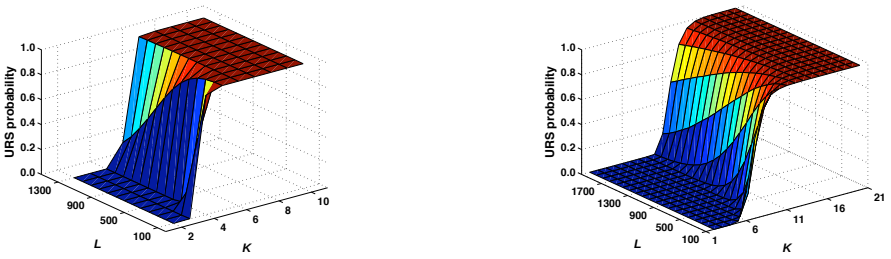
3. Results

3.1. Phase Transition Phenomenon

As mentioned previously, we pay attention to protein sequences. Thus, we choose the 20 amino acids as our alphabet:

$$\Sigma = \{A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V\}$$

Though our methods are applicable to any finite symbolic alphabet.

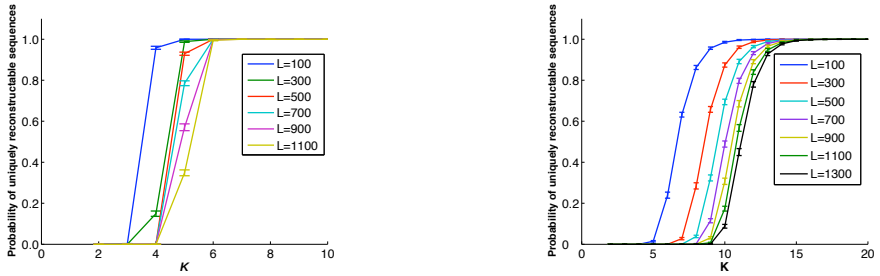


(a) For eiid model (b) For niid mdel

Figure 2: The probability of uniquely reconstructable sequences VS. sequence length L and tuple size K. Both two pictures display a kind of phase transition phenomenon.

Firstly, we carry out Monte Carlo experiments for eiid model, in which the sequence length L range from hundreds to thousands, and K increases from one. The outcome is depicted in Figure 2(a). To clarify the relationship between the URS probability

and K value, we draw the corresponding 2-dimensional graph (Fig. 3(a)) for different sequence lengths L. The bars in Figure 3(a) represent the absolute error coming from Monte Carlo experiments. All errors deviate their corresponding average values about 1%, as shown in Figure 1.



(a) For eiid model (b) For niid model

Figure 3: URS probability VS. tuple size K (sequence length L=constant). Here, the bars stand for the absolute error caused by Monte Carlo experiments. These two 2-dimensional graphs are fetched from several sections of 3-dimensional graphs (Fig. 2). They show a clear cut phase transition phenomenon for different sequence length L.

Similarly, we construct both the 3-dimensional and 2-dimensional graphs for niid model, where weights (Table 1) are adjusted according to the natural abundance of amino acids gotten from NCBI online course [11].

Amino Acids	A	R	N	D	C	Q	E	G	H	I
Frequency(%)	8.3	5.5	4.2	5.3	1.3	3.9	6.3	6.9	2.2	6.0
Amino Acids	L	K	M	F	P	S	T	W	Y	V
Frequency(%)	9.9	5.6	2.4	4.1	4.7	6.8	5.4	1.2	3.1	6.7

Table 1: The frequencies[‖] of 20 amino acids from NCBI online course [11], reflecting the natural abundance of amino acids.

Both 3-dimensional graphs (Figure 2) and 2-dimensional graphs (Figure 3) show a kind of "Phase Transition" phenomenon. That is, the probability jumps suddenly from a low value phase (e.g.< 0.1) to a high value phase (e.g. > 0.9), as K changes a little compared with the probability. For instance, in Figure 3(a), the URS probability approximates zero at $K = 4, L = 1100$ under eiid model, but the probability increases rapidly to a value lager than 0.9 as K increases to 6.

Here, two point are worth noticing. One is that, the phase transition phenomenon of eiid model is more obvious than that of niid model. Namely, the curves in Figure 3(b), though still very sharp, are a little smoother than those in Figure 3(a). We will continue to analyze this aspect in the Discussion. Another is that, all the curves in

[‖] The frequencies are round to integer (xx%) in our experiments. e.g. treating 8.3% as 8%.

Figure 3(b) look like each other. This perhaps contain some scaling information among these curves. But how about the curves in Figure 3(a) ?

3.2. Critical Point of K

So far, we have observed phase transition phenomenon both in the 3-dimensional graphs (Figure 2) and in the 2-dimensional graphs (Figure 3). But what is the turning point of this phase transition phenomenon? To elucidate it rigorously, we define the mininal K value to be the critical point, at which the probability of uniquely reconstructable sequences reaches a value larger than 0.9 for the first time. From now on, $K_{0.9}$ is referred to the critical points of K. We summarize some critical points of K, for different sequence length L from 100 to 10000 in Table 2, under eiid model and niid model, respectively.

Table 2: Critical points $K_{0.9}$ of phase transition phenomenon VS. sequence length L, for eiid model and niid model, respectively.

L	100	200	300	400	500	600	700	800	900	1000
$K_{0.9}$ (eiid)	4	5	5	5	5	6	6	6	6	6
$K_{0.9}$ (niid)	9	10	11	11	12	12	12	12	13	13

L	2000	3000	4000	5000	6000	7000	8000	9000	10000
$K_{0.9}$ (eiid)	6	7	7	7	7	7	7	7	7
$K_{0.9}$ (niid)	14	15	15	16	16	16	16	17	17

Furthermore, we derive the following formula to estimate the critical points of K, by Least Square Method (LSM):

$$K_{0.9} = [1.4490 logL + 1.2121] \ , \quad \text{for eiid model} \tag{3.1}$$

$$K_{0.9} = [3.9749 logL + 0.9490] \ , \quad \text{for niid model} \tag{3.2}$$

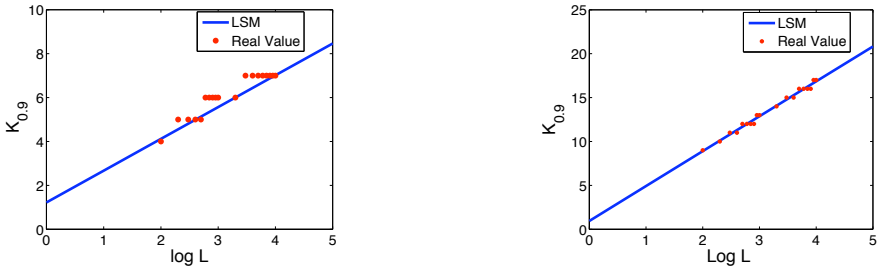
where the brackets in equation (3.1) and equation (3.2) mean round to integer, since it's nonsense for K to be a decimal.

Figure 4 (a) and (b) portrays the error of formula (3.1) and formula (3.2), respectively. In both figures, the red points denote the critical points $K_{0.9}$ gotten from Monte Carlo experiments (Table 2), and the blue lines stand for the results of LSM. As for eiid model, the error arose from formula (3.1) is no more than One, regarding for integer K, while the estimated critical points fit the real K value much better under niid model. In fact, the error is zero for integer K under niid model.

4. Discussion

4.1. Comparing the results of eiid model and that of niid model

We notice that the phase transition phenomenon under eiid model is more obvious than under niid model. Why this happens? Of course, the probability distribution of the



(a) For eiid model (b)

Figure 4: Fitting the critical points $K_{0.9}$ (red) with $logL$ lines (blue), which are obtained by least square method (LSM).

alphabet Σ affect it. Although we do experiments only for one kind of parameter for the niid model, we conjecture that the phase transition phenomenon prevalently exists for any parameter of iid model. We also speculate that the phase transiton phenomenon is sharpest for eiid model, rather than any other niid model, respecting to the same alphabet Σ. Since most extreme value, such as the maximal or minimal value always emerges under the uniform condition.

In addition, the critical points $K_{0.9}$ of phase transition phenomenon are always larger for niid model than that for eiid model, at the same L. The chief reason lies that, due to the biased probability distribution of sample points (i.e. sequences) in the phase space Σ^L under niid model, these typical sequences of niid model are more likely to have long interleaved pair of repeated or triple repeated $(K-1)$-tuples, which cause the sequence cannot be uniquely reconstructed (see. e.g. [1,4,6,8]).

We get the critical point $K_{0.9}$ (vs. L) formula by LSM, but how to study mathematically the exact properties of the critical point is still an open problem.

4.2. About the Alphabet Σ

From our research, what the alphabet is composed of does not influence the URS probability, but the size of alphabet does. Hence, how to derive the exact formula for the probability of uniquely reconstructable sequences in term of sequence L, tuple size K, alphabet size $|\Sigma|$ and the letter distribution over Σ, all of which will affect the URS probability, is worth studying. Although Dyer [2] and Arrtia [1] have obtained the asymptotic formula as $L \to \infty$. It's still meaningful to get an exact one, respecting to L, so that one can investigate the configuration of the URS probability analytically for different sequence length L.

Another relative question is that, how does the alphabet size $|\Sigma|$ affect the phase transition phenomenon of the probability of uniquely reconstructable sequences? We conjecture that there exists the phase transition phenomenon for any finite alphabet. How to investigate this problem analytically instead of doing Monte Carlo experiments is still unknown, while the computer experiments can not exhaust all the possibilities.

Besides, easy to see, a sequence over a small alphabet is likely to form a short

interleaved pair of repeats or short triple repeats. Thus, the smaller the alphabet size $|\Sigma|$ is, the larger the critical point $K_{0.9}$ is, with respect to a constant L. Since these repeats are cause of non-unique reconstruction.

4.3. About Real Proteins

Using our DFA to scan the proteins in PDB.SEQ file [12] which is a collection of SWISS-PORT entries. Above 90% entries of proteins in PDB.SEQ can be uniquely reconstructed at $K = 6$, which consists with the result of Hao et al. [4] gotten by a different method.

Surprisingly, that observation is close to the eiid model (ref. to Fig. 3(a) and Table 2), under which the uniquely reconstructed probability exceeds 0.9 for different L, ranging 100 to 1100, at $K = 6$. One possible explanation is that, in the primordial soup, the amino acids which make up of proteins as we see nowadays may be uniformly distributed, rather than nonuniformly distributed. And yet, in the evolution history, proteins are shaped by the natural selection so heavily that mutations took place in proteins, but still kept the original repeat patterns[**] which probably have important biological function. Anyway, the fact that an majority of real proteins do have a unique reconstruction at $K = 6$, to some extent, supports the compositional distance approach to infer prokaryote phylogeny tree [9].

Another thing is also worth noticing. As indicated by Hao et al. [4], a small group of proteins can only be uniquely reconstruction at large K value (e.g. SRTX_ARTEN protein cannot be uniquely reconstructable until $K = 101$), which should be very very rare according to our experiment results (Figure and Table 2). That implies these proteins may have potential biological functions which are probably relevant to their interleaved pair of $(K-1)$-tuple repeats or triple repeats. These repeats can be found out by modifying our DFA, to further investigate these proteins' functions.

5. Conclusion

Anyhow, to the best of our knowledge, the phase transition phenomenon and the methods (esp. DFA) we employ have not been reported so far.

Last but not least, the phase transition phenomenon in sequence unique reconstruction possibly has profound implications which are still waiting to be explored.

Acknowledgement

This work was accomplished during the Complex System Summer School (Beijing, 2005), supported by SFI and CAS. We would like to thank Prof. Bailin Hao for invaluable discussion and for providing the special protein sequences which have big reconstructable K values. Chan Zhou thanks Qiang Li for pointing [1] to her.

[**] Here the repeat pattern refers to the interleaved pair of repeats and triple repeats.

References

1. R.Arratia, D.Martin, G.Reinert and M.S.Waterman, Poisson process approximation for sequence repeats and sequecing by hybridization, J. of Computational Biology. **3**(1996)425-463.
2. M.Dyer, A.Frieze and S.Suen, The probability of unique solutions of sequencing by hybridization, J. of Computational Biology. **1**(1994)105-110.
3. H.Fleischner, Eulerian graphs and related topics, Part 1, Vol. 2. Annals of Discrete Mathematics 45, Elsevier Science Publishers B. V., Amsterdam-New York, 1991.
4. B.-L.Hao, H.-M.Xie and S.Zhang, Compositional representation of protein sequences and the number of Eulerian loops, http://arxiv.org/abs/physics/0103028, 2001.
5. L.Kontorovich, Uniquely decodable n-gram embeddings, Theoretical Computer Science. **329**(2004)271-284.
6. Q.Li, H.-M.Xie, Finite automata for testing uniqueness of Eulerian trails, http://arxiv.org/abs/cs.CC/0507052, 2005.
7. J.H.Mathews and K.D.Fink, Numerical Methods Using Matlab ,4th Edition, Prentice Hall, New Jersey, 2003.
8. P.A.Pevzner, Computational Molecular Bioloy: An Algorithmic Approach, MIT Press, Cambridge, 2000.
9. J.Qi, B.Wang and B.-L.Hao, Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach, J. of Molecular Evolution. **58**(2004)1-11.
10. M.S.Waterman, Introduction to Computational Biology, Chapman & Hall, London, 1995.
11. The frequencies of 20 amino acids are available from NCBI(National Center for Biotechnology Informaton) website: http://www.ncbi.nlm.nih.gov/Class/Structure/aa/aa_explorer.cgi
12. The PDB.SEQ file we used is downloaded from the following ftp: ftp://ftp.es.embnet.org/pub/databases/swiss-prot/special_selections/pdb.seq