DIMACS Technical Report 99-56
November 1999

An Approach of Information Theory to Multiple Sequence Comparison [1]

by

Weiwu Fang[2]
DIMACS Center, Rutgers University
IAM-CAS and APORC

Fred S. Roberts[3]
DIMACS Center, Rutgers University

Zhengrong Ma
Waksman Institute for Microbiology
Rutgers University

[1] Weiwu Fang's work has been supported by United States NSF grants STC-91-19999 and BIR 94-12594 to Rutgers University and Chinese NSF grants 69772036 and 39830070; Fred Roberts' work has been supported by the United States NSF grants cited above plus grant SBR-97-09134.
[2] IAM-CAS: Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing
APORC: Asian-Pacific Operations Research Center
[3] DIMACS Center: The Center for Discrete Mathematics and Theoretical Computer Science

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs, Bell Labs, NEC Research Institute, and Telcordia Technologies (formerly Bellcore).

DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.

ABSTRACT

Multiple sequence comparison is a basic problem for molecular biology and other sciences. In this paper, we introduce the concept of complete information set and some measurement principles for measuring multiple sequence discrepancy. Based on them, we present a new measurement method satisfying the principles for comparing multiple sequences. We show that this method can effectively distinguish different random sequences or DNA sequences, for example, distinguish DNA sequences of length 8000 by comparisons of 6-8 symbol strings or protein sequences of length 8000 by comparisons of 3-4 strings. It can also measure slight changes of a sequence, e.g., insertion or deletion of a symbol. We apply it in the study of molecular evolution; the results show that there is a hierarchic relationship among the cytochrome C protein sequences of different species, much as that in taxonomy; moreover, these results are consistent with previous studies.

Key Words: multiple sequence comparison, entropy, DNA, measurement

An Approach of Information Theory to Multiple Sequence Comparison

1. Introduction

The comparison of multiple sequences is a basic problem in the study of DNA sequence, communication, linguistics, and other areas. There are many approaches which are designed to make comparison of multiple sequences in molecular biology, but none is considered totally satisfactory, and it is still a challenge for molecular biology and other sciences to find useful and fruitful comparison methods. Some of the main approaches to multiple sequence comparison are as follows (Chan, S.C., Wong, A.K. & Chiu, D.K. 1992):

(1). Alignment approaches (e.g., Needleman and Wunsch's method, 1970).
(2). Heuristic approaches and statistical methods.
(3). Information theory approaches (e.g., Gatlin's approach, 1972).

The alignment approaches are most widely used in the study of DNA sequences. They involve some subjective factors, such as determining optimal criteria, assigning scores, and so on. As Pevzner and Waterman (1995) said, "there is a considerable disagreement about how to weight matches, mismatches, insertions/deletions and gaps in sequence alignment. Ideally, we would like to compute optimal alignments for the infinite set of all possible parameter values". This is a really difficult task. Moreover, for many alignment algorithms of multiple sequence comparison, different results will be produced when a designed comparison order is changed. In addition, it seems that heuristic approaches and statistical methods haven't shown more advantages than alignment approaches so far; moreover, some statistical methods require that DNA sequences should be stationary and ergodic, but these requirements are not supported by experiments (Farach, Noordewier, Savari, Shepp, Wyner and Ziv, 1994, Loewenstern and Yianilos, 1996).

As to the information theory approaches, they only depend on the original information about sequences; no other subjective factors are involved. Unfortunately, most information-theoretic attempts have had little or no success in the study of DNA sequences (Farach, Noordewier, Savari, Shepp, Wyner and Ziv, 1995), and some research has even denied the utility of information theory (Hariri, Weber, and Olmsted III, 1990).

There are two basic approaches based on information theory for comparing DNA sequences. These approaches originate from Shannon's work (1949).

(1). Conditional entropy approaches.

One example is the approach of Gatlin (1972). In his paper, information is classified into two categories: *"potential information"* H_{max}, which is the maximum uncertainty, and *"stored information"* H_m, which is associated with the ordering process. A total divergence D_m is defined as

$$D_m = H_{max} - H_m$$

where

$$H_m = -N \sum_i \sum_j \sum_k \cdots \sum_m p_i p_{ij} p_{ijk} \cdots p_{ijk...m} \log p_{ijk...m} \quad \text{and}$$

the $p_{ij...}$'s are conditional probabilities. This divergence was designed to compare sequences.

(2). The entropy estimation approaches.

Since all DNA sequences are very similar to a random sequence, it is difficult to distinguish them by a simple entropy estimation. So, some researchers try to find a good entropy estimation method (Farach, Noordewier, Savari, Shepp, Wyner and Ziv, 1994, Loewenstern and Yianilos, 1996). Further developments based on these methods might be used for making comparison between sequences.

These approaches are not very effective when they are applied to some sequence comparisons. For example, the (conditional) probability distributions $(p_1, p_2, \cdots, p_n)$ and $(p_2, p_1, \cdots, p_n)$ have the same entropy value, but they represent different information about the sequences.

In this paper, we introduce the concept of complete information set and some measurement conditions for measuring multiple sequence discrepancy. Based on them we present a new measurement method satisfying these conditions for comparing multiple sequences. We show that this method can effectively distinguish different random sequences or DNA sequences based on comparison of 6-8 symbol strings for DNA sequences or 3-4 for protein sequences with length 8000 symbols. It can also measure slight changes of a sequence, e.g., insertion or deletion of a symbol. We apply it in the study of molecular evolution; the results show there is a hierarchic relationship among the cytochrome C protein sequences of living things, much as that in taxonomy, and the results are consistent with previous studies.

2. Complete information set and principles of measurement

"Multiple sequence comparison refers to the search for similarity in three or more sequences" (S.C.Chan, A.K. Wong, and D.K. Chiu, 1992). In order to establish a reasonable basis to measure similarity (or discrepancy) among multiple sequences, we present the concept of complete information set and some principles of measurement.

1. The complete information set of a sequence.

Let $\Sigma = \{a_1, a_2, \cdots, a_m\}$ be an *alphabet* of m symbols, and suppose $S = \{S_1, S_2, \cdots, S_s\}$ is a set of sequences formed from the symbol set Σ. We denote the set of all different sequences formed from Σ with length l by Θ^l; then the number $m(l)$ of all sequences of Θ^l equals m^l. For a sequence $S_k \in \mathcal{S}$, let L_k be its length and n^l_{ik} denote the number of contiguous subsequences in S_k which match the i-th sequence of Θ^l, $l \leq L_k$. It is easy to see that

$$\sum_{i=1}^{m(l)} n^l_{ik} = L_k - l + 1 \quad \text{for each } l \leq L_k \text{ and } k.$$

Letting $p^l_{ik} = n^l_{ik}/(L_k - l + 1)$, we obtain a distribution

$$U^l_k \;:=\; (p^l_{1k},\ p^l_{2k},\ \cdots,\ p^l_{m(l)k})^T \quad \text{where } \sum_{i=1}^{m(l)} p^l_{ik} = 1$$

for each k and $l \leq L_k$. Let Γ^l denote the set of all distributions satisfying $\sum_{i=1}^{m(l)} p^l_{ik} = 1$, i.e.,

$$\Gamma^l := \{(p^l_{1k}, p^l_{2k}, \cdots, p^l_{m(l)k})^T \mid \sum_{i=1}^{m(l)} p^l_{ik} = 1 \text{ and } p^l_{ik} \geq 0\}, \quad (l = 2, 3, \cdots).$$

Thus, for each sequence S_k, we can get a unique set of distributions

$$(U^1_k, U^2_k, \cdots, U^{L_k}_k) \text{ where } U^1_k \in \Gamma^1, U^2_k \in \Gamma^2, \cdots, \text{ and } U^{L_k}_k \in \Gamma^{L_k}.$$

This set contains all structure information of a sequence; in particular, $U^{L_k}_k$ uniquely determines the original sequence, so we call this set $(U^1_k, U^2_k, \cdots, U^{L_k}_k)$ *a complete information set of the sequence* S_k. Any two different sequences have different complete information sets (and vice versa). Discrepancies among these sets mean discrepancies of all structure information among sequences, so a reasonable approach to measurement of discrepancy (similarity) is to find an effective and efficient measurement method which is based on comparisons of complete information sets of multiple sequences.

Due to the limitation of computer power, it is impossible to compare all information contained in complete information sets. Fortunately, there is some kind of connection among $U^1_k, U^2_k, \cdots, U^{L_k}_k$ and a systematic investigation based on features of these sets may allow us find some regularities and obtain useful results for real applications. In fact, there is a lot of information in the beginning information subsets, which determine some basic characteristics of a sequence set and can be used to analyze the similarity of sequences.

In order to develop a discrepancy (similarity) measure, we first introduce some conditions of measurement and experimental criteria as the basic principles which can be used to evaluate such measures.

2. Some measurement principles on discrepancy of multiple sequences.

From Section 2.1, a *complete information set* contains all structure information of a sequence, so a measure of discrepancy (similarity) among distributions of *complete information sets* of multiple sequences can be considered as a measurement of discrepancy among different sequences. As is also mentioned in Section 2.1, the distribution U_k^l of a sequence S_k is obtained by comparing the subsequences of length l of this sequence with sequences in Θ^l. A proper measure of discrepancy of multiple sequences should be based on the same comparison procedures, i.e, should be defined on comparisons of distributions with the same dimension $m(l)$. This means that a measure of discrepancy among $U_1^l, U_2^l, \cdots, U_s^l$ of s sequences is required.

Let $D(U_1^l, U_2^l, \cdots, U_s^l)$ denote a discrepancy measure with respect to length l defined on the distribution set $\{(U_1^l, U_2^l, \cdots, U_s^l) | U_k^l \in \Gamma^l, k = 1, 2, \cdots, s\}$. We present some conditions and criteria for measurement of discrepancy. The following conditions seem to be reasonable requirements.

Condition A.

$D(U_1^l, \cdots, U_s^l)$ *is real, continuous, bounded, and non-negative with respect to all* $U_k^l \in \Gamma^l$.

Condition B.

$D(U_1^l, \cdots, U_s^l)$ *is symmetric with respect to any permutation of* K *or* I, *where* $K = \{1, 2, \cdots, s\}$, $I = \{1, 2, \cdots, m(l)\}$, *and a permutation of* K *corresponds to reordering all of the sequences in* S *and a permutation of* I *corresponds to reordering all sequences in* Θ^l

Condition C.

$D(U_1^l, \cdots, U_s^l) = 0$ *if and only if all* $U_1^l, \cdots, U_s^l$ *are identical to each other.*

Based on the observations in sequence comparisons, we find several features of complete information sets that lead us to further conditions on measurement of discrepancy.

The first feature is called the **"feature of discrepancy augment"**: with the increase of compared length, the number of matched subsequences may decrease and never increase. For example, for sequences AGCT and AACTG, there is one common subsequence CT when the compared length equals 2, but there is no common subsequence when the length is 3. This means their discrepancy in Γ^3 should not be less than that in Γ^2. When l is the greatest, different sequences can be uniquely recognized; that means the similarity of different sequences becomes the smallest. So we have

Condition D (Augment property).

For any group of distributions $U_1, \cdots, U_s$, $D(U_1^l, \cdots, U_s^l)$ *is non-decreasing, i.e.,* $D(U_1^l, \cdots, U_s^l) \leq D(U_1^n, \cdots, U_s^n)$ *if* $l \leq n \leq \min_{k \in K} L_k$.

The second feature is called the **"stability of discrepancy heredity"**: if there doesn't exist any common subsequence of length l in a group of sequences, neither does any common subsequence of length $n > l$ in this group of sequences. In this situation, discrepancy measurements based on comparisons of subsequences of length n or l should have no difference. Thus we present

Condition E (Heredity property).

If there doesn't exist any common subsequence of length l *in a group of* s *sequences, then* $D(U_1^{n+1}, U_2^{n+1}, \cdots, U_s^{n+1}) = D(U_1^n, U_2^n, \cdots, U_s^n)$ *for* $n \geq l$.

This feature also hints to us that it is not necessary to compare all elements of information sets in this situation.

Based on $D(U_1^l, \cdots, U_s^l)$, we can also define another measure *(discrepancy ratio)* according to the augment and bounded property of $D(U_1^l, \cdots, U_s^l)$:

$$R(U_1^l, \cdots, U_s^l) = D(U_1^l, \cdots, U_s^l)/MaxD,$$

where $MaxD$ is the *maximum discrepancy of* s *sequences,* i.e.,

$$MaxD := Max\{MaxD^l | l \leq min_{k \in K} L_k\} \text{ and}$$

$$MaxD^l := Max\{D(U_1^l, \cdots, U_s^l) | U_k^l \in \Gamma^l, k = 1, 2, \cdots, s\}.$$

The third feature is called the **"feature of exponential divergence of different random sequences"**. (Hereafter, the term "different random sequences" means they are independent and identically distributed sequences of i.i.d. symbols). This feature is based on the following fact: the probability that a random sequence contains a specific subsequence of length l is less than $(L_k - l + 1)/m^l$ (when $L_k - l + 1 \leq m^l$), which implies that the number of common subsequences possessed by different sequences will exponentially decrease with the increase of l. This also implies that the discrepancy of different random sequences should become greater quickly with increase of l, i.e., the $R(U_1^l, \cdots, U_s^l)$ value of different random sequences should tend to 1 "quickly"(we purposely leave this vague) with increase of l. This feature can be considered an experimental criterion to evaluate whether a measure is efficient for real applications.

Efficiency Criterion.

For different random sequences, the value $R(U_1^l, \cdots, U_s^l)$ *should tend to 1 "quickly" with the increase in* l.

3. A new approach based on information theory

As is mentioned in Section 2, a *complete information set* contains all structure information of a sequence, so a measurement of discrepancy (similarity), satisfying the conditions in Section 2.2, among distributions of *complete information sets* of multiple sequences can be considered as a measurement of discrepancy among different sequences.

At first, consider two distributions $P(p_1, p_2, \cdots, p_t)$ and $Q(q_1, q_2, \cdots, q_t)$, where $\sum_{i=1}^t p_i = 1$ and $\sum_{i=1}^t q_i = 1$. The Kullback-Leiber entropy and cross entropy (Kullback, 1959),

$$H(P,Q) = \sum_{i=1}^t p_i \log \frac{p_i}{q_i} \quad \text{and} \quad H(P,Q) = \sum_{i=1}^t p_i \log \frac{p_i}{q_i} + \sum_{i=1}^t q_i \log \frac{q_i}{p_i},$$

have been thought to be measures of discrepancy between two such distributions; for example, Kullback (1959) calls the K-L measure "a coefficient of divergence of one distribution from another" and Kapur and Kesavan (1990) call them "directed divergence of a probability distribution P from another probability distribution Q". But, these measures cannot be used in measuring the discrepancy among distributions of *complete information sets* because in most cases there are many zero-entries in distributions of *complete information sets* (these measures cannot be applied in mutually singular cases); moreover, they don't satisfy Conditions D and E; the K-L measure doesn't meet Condition B. In addition, if the Shannon or Shannon/Gatlin approach (entropy or conditional entropy) was used in comparison of distributions of multiple information sets, they wouldn't satisfy Conditions C, D and E. Further, none of them is efficient relative to the Efficiency Criterion (Hariri, B. Weber and J. Olmsted III, 1990).

Fang (1994, 1997a) introduces a function for measuring information discrepancy contained in multiple information sources. This function has many good properties, such as non-negativity, identity, symmetry, boundedness, uniform continuity, monotonicity, linear homogeneity, maximum, invariance on average, recursiveness, and so on. This function under a reduced form can satisfy all the requirements in Section 2 even for mutually singular cases. The function B is defined as follows.

Suppose we are given a group $G(t,s,r)$ of s matrices that are $t \times r$, u_{ikj} is an entry of the k-th matrix, and $\sum_{i=1}^t \sum_{j=1}^r u_{ikj} = 1$, $k = 1, 2, \cdots, s$; further, a $G(t,s,r)$ is called *a matrix group with symmetry structure* if $s \equiv 0 \pmod{t}$, or $n \neq 0$ but $n \cdot r \equiv 0 \pmod{t}$ or $n \cdot r \equiv 0 \pmod{s}$. Let $\mathcal{G}(\mathcal{T}, \mathcal{S}, \mathcal{R})$ denote the set of all $G(t,s,r)$. The measurement function $B(t,s,r)$ of discrepancy among multiple information sources is defined by Fang on $\mathcal{G}(\mathcal{T}, \mathcal{S}, \mathcal{R})$. We do not define $B(t,s,r)$ in generality here.

When $r = 1$, a $G(t,s,r)$ reduces to a group of distributions, i.e., $G(t,s,r) := (U_1, U_2, \cdots, U_s)$ where $U_k := (u_{1k}, u_{2k}, ..., u_{tk})$ and $u_{1k} + u_{2k} + ... + u_{tk} = 1$. In the meantime, the formula $B(t,s,r)$ reduces to the formula

$$B(U_1, U_2, \cdots, U_s) = \sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik} \log \frac{u_{ik}}{\sum_{k=1}^{s} u_{ik}/s},$$

where $0 \cdot \log \frac{0}{0}$ is defined as 0 as in the K-L entropy.

The following lemmas were obtained.

Lemma 1A. (Fang, 1994) *Given $t, s,$ and r and a group $G(t,s,r)$ of matrices, then an upper bound of $B(t,s,r)$ (i.e., $B(U_1, U_2, \cdots, U_s)$) is $s \cdot \log t$.*

Lemma 1B. (Fang, 1997A) *For any $t, s,$ and r and a group $G(t,s,r)$ of matrices, the maximum of $B(t,s,r)$ is 1. $s \cdot \log\ s$ if $s \leq t$;*

2. $s \cdot \log\ t$ if $s > t$ and $G(t,s,r)$ is a group of matrices with symmetry structure.

(These lemmas also imply that the upper bound and maximum are independent of r).

If we measure the discrepancy among $U_1^l, \cdots, U_s^l$ of multiple complete information sets by

$$B(U_1^l, \cdots, U_s^l) = \sum_{k=1}^{s}\sum_{i=1}^{m(l)} p_{ik}^l \log \frac{p_{ik}^l}{\sum_{k=1}^{s} p_{ik}^l/s}, \tag{1}$$

where $U_k^l := (p_{1k}^l, p_{2k}^l, \cdots, p_{m(l)k}^l) \in \Gamma^l$, the following theorems hold.

Theorem 1.

The maximum of $B(U_1^l, \cdots, U_s^l)$ is $s \log s$ for $s \leq m(l)$; the upper bound is $s \cdot \log m(l)$ for $s > m(l)$.

This theorem can be obtained directly from Lemma 1A and Lemma 1B.

Remark. $s\ \leq m(l)$ almost always holds for DNA sequence comparisons; even for $s > m(l)$, the maximum of $B(t,s,r)$ equals or tends to $s \cdot \log m(l)$ (Fang, 1997A).

Theorem 2.

$B(U_1^l, \cdots, U_s^l)$ satisfies Conditions A, B, C, D, and E.

The proof of this theorem is in the Appendix.

For this new measure, the *maximum discrepancy MaxB of s sequences* is:

$$MaxB := Max\{MaxB^l | l \leq min_{k \in K} L_k\} \text{ where}$$

$$MaxB^l := Max\{B(U_1^l, \cdots, U_s^l) | U_k^l \in \Gamma^l, k = 1, 2, \cdots, s\}.$$

For both $s \leq m(l)$ and $s > m(l)$, from Theorem 1 it is easy to see that $MaxB$ is less than or equal to $s \cdot \log s$, so a corollary is:

Corollary 1. *For any group S of s sequences, its maximum discrepancy $MaxB$ is less than or equal to $s \cdot \log s$.*

Thus, we have

$$R(U_1^l, \cdots, U_s^l) = \frac{B(U_1^l, \cdots, U_s^l)}{s \cdot \log s} = \frac{\sum_{k=1}^{s} \sum_{i=1}^{m(l)} p_{ik}^l \log \frac{p_{ik}^l}{\sum_{k=1}^{s} p_{ik}^l / s}}{s \cdot \log s} \leq 1 \tag{2}$$

from Corollary 1 and Section 2.

According to Theorem 2, we see that multiple sequences are different if their $B(U_1^l, \cdots, U_s^l)$ or $R(U_1^l, \cdots, U_s^l)$ is not zero for any l. Moreover, the measured discrepancy will augment with increase of l. Due to the fact that subsequences of independent random sequences tend to be different "exponentially", Theorem 2 also implies that $R(U_1^l, U_2^l, \cdots, U_s^l)$ among random sequences (or DNA sequences) will "quickly" tend to 1 with increase of l, i.e., the new approach possesses the feature of "exponential divergence of different random sequences" as well as the feature of "stability of discrepancy heredity".

Fang (1994, 1997b) defines a measure $B_k(U_1, U_2, \cdots, U_s)$ as follows:

$$B_k(U_1, \cdots, U_s) = \sum_{i=1}^{t} u_{ik} \log \frac{u_{ik}}{\sum_{k=1}^{s} u_{ik}/s}. \tag{3}$$

Fang proves that B_k has similar properties to $B(U_1, U_2, \cdots, U_s)$, so a *discrepancy ratio of the k-th sequence relative to s sequences* can be defined using B_k:

$$R_k(U_1^l, \cdots, U_s^l) = \frac{B_k(U_1^l, \cdots, U_s^l)}{s \log s} = \frac{\sum_{i=1}^{m(l)} p_{ik}^l \log \frac{p_{ik}^l}{\sum_{k=1}^{s} p_{ik}^l / s}}{s \log s}. \tag{4}$$

The efficiency of the new measure.

Computer simulations illustrate that comparisons using these new measures are very efficient. Some computational results for two independent random sequences are shown as the following table.

m (the number of symbols)	sequence length	length l of subsequence	$R(U_1^l,\ U_2^l)$ value
4	3000	9	0.992
20	3000	4	0.991
4	8000	9	0.97
20	8000	4	0.965

In real applications, it may be better to calculate R or B values with a properly smaller l (see Sections 4 and 5).

4. The example of Hariri, Web, and Olmstead about DNA sequence comparison.

In the paper by Hariri, Weber, and Olmsted III (1990), the authors said: "we therefore conclude, albeit reluctantly, that Shannon/Gatlin type calculations are not productive". In this section, a similar conclusion to theirs is found, but our new approach appears to be very productive for this kind of examples.

We obtained the DNA sequences of the examples in the paper of Hariri et al. from the Genbank Database. The DNA sequences and their deduced amino acid sequences have been used. The sequence data were grouped into three classes based on their protein structure as globular, fibrous, and membrane protein groups: chicken egg-white lysozyme (CHKLYS), chicken cytochrome C (CHKCYTC), dog chymotrypsin (DOGCTRP), globular-T4E bacteriophage lysozyme (T4E), mouse alpha-i type I procollagen (MUSPRCL), chicken sacle keratin (CHKKER), chicken muscle tropomyosin (CHKTOP), membrane-bacterio-opsin (BACOSPIN), E.coli H^+ -ATPase F_0 subunit c (ECOFFSC), and bovine cytochrome C oxidase subunit IV (BOVCYTCX).

We made
(1) comparisons of DNA sequences in the same classes,
(2) comparisons of DNA sequences with random sequences,
(3) comparisons of DNA sequences in different classes, and
(4) comparisons of DNA sequences with small changes (e.g., change one symbol).

The following conclusions were obtained based on computational results. Some of them are also shown in Figures 1 to 5.

(1). The new approach is very effective for DNA sequence comparison in the same class.

All DNA sequences in the same class can be distinguished by the new approach. The comparisons based on 6 to 8 symbol strings (l =6, 7 or 8) for DNA sequences (3 to 4 symbol string for protein sequences) are enough to distinguish sequences in the same classes: $R(U_1^l, \cdots, U_s^l)$'s of the new method are greater than 0.982 for $l = 8$ and greater than 0.996 for $l = 9$. (Recall that $R = 0$ if they are identical sequences; $R = 1$ is the maximum of different sequences). This convergence rate is much faster than that of the (conditional) entropy method, which is shown in Figure 1. In this figure, sequences DOGCTRP and T4E were used (s=2); the vertical coordinate represents the value of $R(U_1^l, \cdots, U_s^l)$, or the *Shannon ratio SR of a distribution,* where

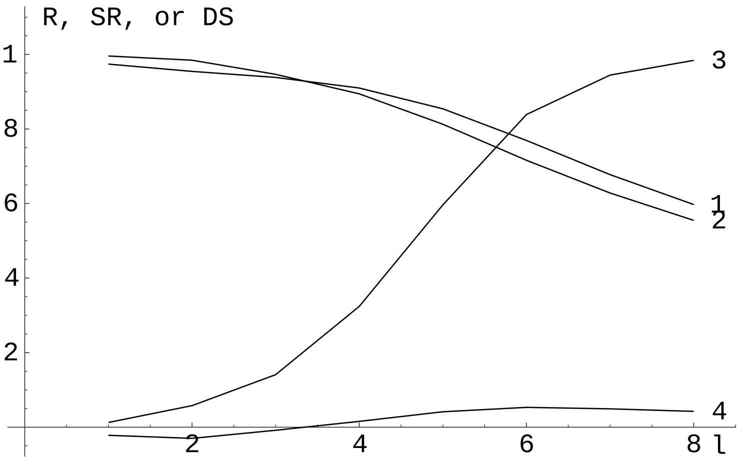

Figure 1: The comparison of convergence rate of R with Shannon ratio, where 1 is the Shannon ratios SR of DOGCTRP, 2 is the Shannon ratios SR of T4E, 3 is the values R of DOGCTRP and T4E, and 4 is the difference DS of two Shannon ratios.

$$SR := \frac{\text{the value of Shannon entropy of the distribution}}{\text{the maximum of Shannon entropy in the same dimension}},$$

or the difference DS of two Shannon ratios; the horizontal coordinate represents the value of l. When the R value of the new approach is greater than 0.982, the Shannon ratio SR is greater than 0.61. The latter value is less than the half way to zero (Shannon entropy and conditional entropy should tend to 0), and it is difficult to continue to calculate values of Shannon entropy and conditional entropy for $l > 9$ because of the limit of computer power. Even though their values could be convergent to 0, the values of Shannon entropy (or conditional entropy) of different sequences are almost equal to each other for any l. In Figure 1, the curve 4 shows the difference of the Shannon ratios of two sequences; it always closes to zero for all l, and sometimes is negative.

(2). All DNA sequences can be distinguished from random sequences by comparisons based on 6 to 8 symbol strings ($l = 6$, 7 or 8). In Fig. 2, the comparison of the sequence CHKLYS with a random sequence of length 1000 is shown.

(3). The new approach is also effective for DNA sequence comparison of different classes.

These results are shown in Fig. 3 by a comparison among three sequences (s=3), which are chosen from the globular, fibrous, and membrane classes respectively. In Fig. 3, the vertical coordinates represent values of $B_k(U_1^l, \cdots, U_s^l)$, $k = 1, 2, 3$; the sum of the values of $B_k(U_1^l, \cdots, U_s^l)$ $(k = 1, 2, 3)$, which is B, equals $0.985 \cdot (3 \cdot \log 3)$ or 98.5% of the maximum $s \log s$ for $l = 8$.

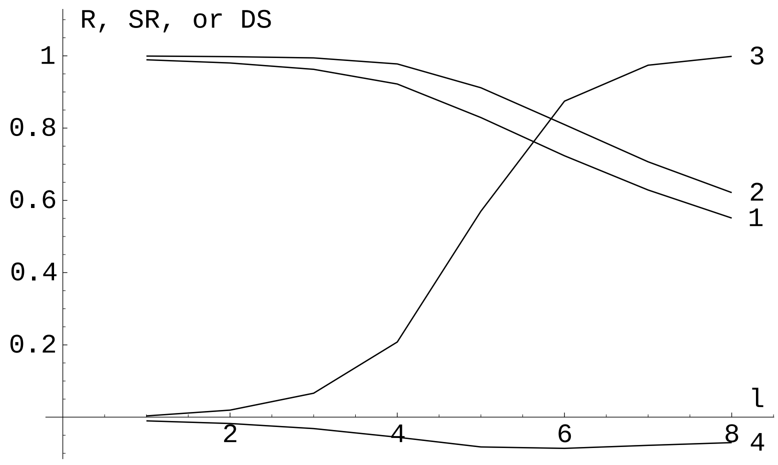

Figure 2: The comparison of the sequence CHKLYS with a random sequence of length 1000, where 1 is the Shannon ratios SR of CHKLYS, 2 is the Shannon ratios SR of the random sequence, 3 is the values R of CHKLYS and the random sequence, and 4 is the difference DS of two Shannon ratios.

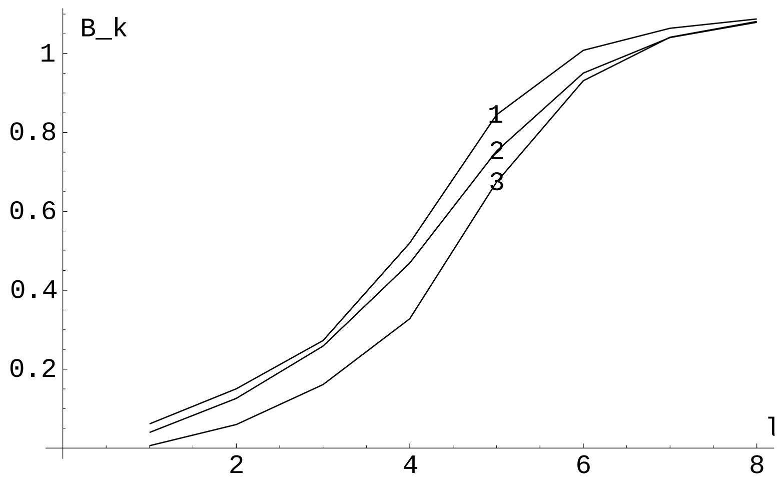

Figure 3: The comparison among three sequences from the globular, fibrous, and membrane classes, where 1 is the values B_k of CHKCYTC, 2 is B_k of MUSPRCL, and 3 is B_k of BACOSPIN.

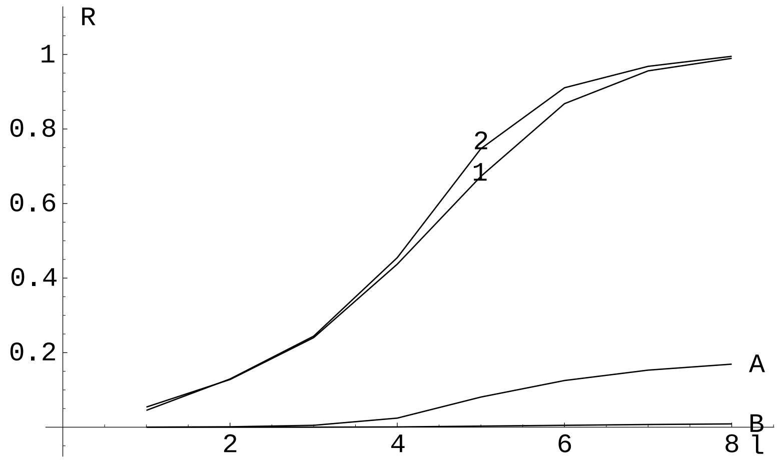

Figure 4: The pairwise comparisons of CHKTOP with other sequences in the fibrous group and with S_1 and S_2 sequences, where the curve 1 represents the values R of CHKTOP and MUSPRCL, 2 is R of CHKTOP and CHKKER, A is R of CHKTOP and the sequence S_1, and B is R of CHKTOP and S_2.

(4). The new approach is sensitive. Even a small change of an sequence (e.g., one symbol is randomly added or deleted) can be measured, and the discrepancy of similar sequences (e.g., chicken smooth-muscle alpha-tropomyosin gene and chicken skeletal muscle beta-tropomyosin gene) can be also measured.

We added two sequences S_1 and S_2 in the fibrous group (MUSPRCL, CHKKER, CHKTOP), where Sequence S_1, the chicken skeletal muscle beta-tropomyosin sequence (CHKTRP), is similar to the sequence of chicken smooth muscle tropomyosin (CHKTOP), and Sequence S_2 is a variation of the chicken smooth muscle tropomyosin sequence (CHKTOP) obtained by randomly inserting a symbol in CHKTOP. We made pairwise comparisons of the chicken smooth muscle tropomyosin sequence (CHKTOP) with each sequence of this new group. The results are shown in Figure 4, and Figure 5 is an enlargement of a part of Figure 4.

These results imply that the new approach not only can be used to distinguish genes of different species, but can also be used to measure variations of a gene.

5. Some elementary results in the study of molecular evolution.

Here we give an elementary example of applications of this new measure. Further investigations are needed. The cytochrome C protein is used in this study. It is an important protein involved in the cellular bioenergetic system in almost

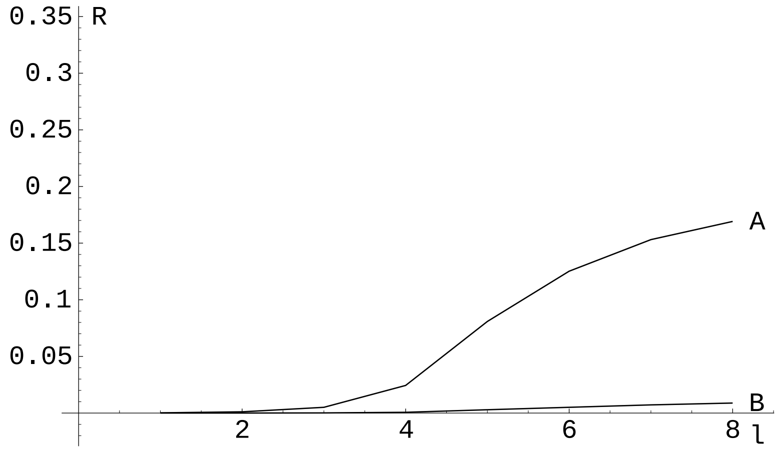

Figure 5: The enlargement of some part of Figure 4 – the pairwise comparisons of CHKTOP with the similar sequence S_1 and the slightly changed sequence S_2, where A is the values R of CHKTOP and S_1, and B is R of CHKTOP and S_2.

all sorts of organisms with cell structure. The cytochrome C is highly conserved in amino acid sequences and is a model in the study of molecular evolution; in Moore and Pettigrew's book (1990), a connection between cytochrome C sequences and biological taxonomy has been made by a complex analysis.

We made pairwise comparisons of the human cytochrome C protein sequence with those of different animals and random sequences by our approach; all sequence data used are from the book mentioned above and the following results were obtained.

(1). Pairwise comparisons of the cytochrome C protein sequence of human with those of different animals.

In Table 1, we list some values R between the human cytochrome C sequence and those of other animals for $l = 3$ and $s = 2$; More values R for different l are shown in Fig. 6.

From the elementary analysis of data above, one can see that the farther the species is from the human being in taxonomy, the bigger the R value in Table 1 is, and this coincides with Moore and Pettigrew's results. It should be noted that sequence of the cytochrome C of the rattlesnake is an oddity compared to others. It is closer to the human sequence even than that of some mammals. The same oddity also exists in the results obtained by other comparison methods like the maximum parsimony method, and the reason for this is not clear; some authors think the oddity for the snake might be the result of gene duplication and parallel evolution (Baba, Darga, Goodman, and Czelusniak, 1981).

Phylum	Class (order)	Animal name	R value compared with human gene ($l = 3$)
vertebrate	mammalia (primates)	rhesus monkey	0.00294
vertebrate	mammalia (primates)	spider monkey	0.0882
vertebrate	mammalia (rodent)	mouse	0.2255
vertebrate	mammalia	rabbit	0.2255
vertebrate	mammalia	kangaroo	0.2353
vertebrate	mammalia	dog	0.2605
vertebrate	peptilia	rattlesnake	0.2610
vertebrate	mammalia	bat	0.2647
vertebrate	mammalia	zebra	0.2647
vertebrate	mammalia	horse	0.2745
vertebrate	mammalia	elephant seal	0.2899
vertebrate	mammalia	hippopotamus	0.2941
vertebrate	aves	chicken	0.3002
vertebrate	aves	ostrich	0.3002
vertebrate	mammalia	pig	0.3122
vertebrate	mammalia	great whale	0.3380
vertebrate	peptilia	turtle	0.3529
vertebrate	pisces	carp	0.3691
vertebrate	amphibia	bullfrog	0.4301
vertebrate	pisces	pacific lamprey	0.4589
vertebrate	pisces	dogfish	0.4889
vertebrate	pisces	tuna	0.5135
arthropoda		freshwater prawn	0.5209
vertebrate	pisces	bonito	0.5234
arthropoda	insecta	fruit fly	0.5254
arthropoda	insecta	honeybee	0.5350
annelida		common brandling worm	0.6189
protozoa		neurospora	0.7673
protozoa		enteromorpha intestinalis	0.8014

Table 1. Comparison of human cytochrome C with those of animals (s=2).

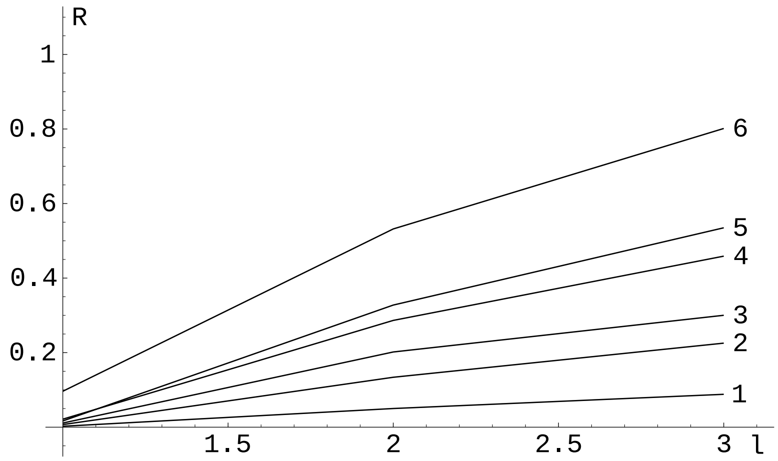

Figure 6: Pairwise comparisons of the human cytochrome C with those of animals, where 1 is that with spider monkey, 2 is with mouse, 3 is with ostrich, 4 is with pacific lamprey, 5 is with honeybee, and 6 is with enteromorpha intestinalis.

(2). Comparisons of the cytochrome C sequences of different animals with random sequences.

We made pairwise comparisons of the cytochrome C sequence of each animal with nine independent random sequences of length 1200, and a mean $\bar{R}$ was obtained by averaging nine values R between each cytochrome C sequence and the nine random sequences ($s = 2$). The results are listed in Table 2 ($l = 3$). The results indicate that the $\bar{R}$'s of all cytochrome C sequences compared with the random sequences are very close to each other (0.9714 ± 0.0021), i.e., all of them have a close degree of similarity compared with random sequences; this means we cannot say which one is more ordered or more disordered relative to random sequences; moreover, the mean R of each cytochrome C sequence compared with the nine random sequences is very close to the $\bar{R}$ (0.9630 ± 0.0057) of other random sequences 1,2,3 and 4 compared with the nine random sequences.

name	The average $\bar{R}$ of 9 discrepancy values R between a gene and 9 random sequences ($s = 2$ and $l = 3$)
human	0.9716
spider monkey	0.9716
mouse	0.9722
dog	0.9720
rattlesnake	0.9716
chicken	0.9717
turtle	0.9715
pacific lamprey	0.9712
fruit fly	0.9721
honeybee	0.9693
mung bean	0.9699
rice	0.9713
wheat germ	0.9719
other random sequences:	
random sequence 1	0.9612
random sequence 2	0.9627
random sequence 3	0.9594
random sequence 4	0.9687

Table 2. The pairwise comparisons of each cytochrome C sequence with 9 random sequences and the comparisons of other random sequences with the nine random sequences.

(3). From (1) and (2) of this section, we can obtain the following observation of interest:

each cytochrome C sequence is almost a random sequence and none of them is more ordered relative to random sequences; however, in contrast to ordinary random sequences, there exist a close hierarchic relationship between all sequences of living things as that in taxonomy.

Finally, based on the observations described in this section and the previous one, we can conclude that this new approach to measuring discrepancy of sequences seems to be sensitive, objective, effective and efficient.

One problem that might be worth investigating in the future is this: there is somewhat big gap between the primate and other mammals in Table 1; what caused the gap? Was it lack of data ? Is there a different explanation?

APPENDIX

We give some definitions and lemmas at first. A subsequence of length l is written as *an l-subsequence;* after attached by a symbol in Σ, an l-subsequence becomes a new $(l+1)$-subsequence; the set of all l-dimensional distributions is denoted by Γ^l, and Γ^{l+1} is defined similarly. The construction procedure of $U_1^l, U_2^l, \cdots, U_s^l$ (or $U_1^{l+1}, U_2^{l+1}, \cdots, U_s^{l+1}$) of a sequence is a matching procedure by comparing the l(or $l+1$)-subsequences of the sequence with sequences in Θ^l (or Θ^{l+1}). The variations from $U_1^l, U_2^l, \cdots, U_s^l$ to $U_1^{l+1}, U_2^{l+1}, \cdots, U_s^{l+1}$ correspond the changes of the number of matched subsequences in construction procedures.

Definition 1. *A suffix $(l+1)$-subsequence of an l-subsequence.*

A *suffix $(l+1)$-subsequence of an l-subsequence* is a subsequence obtained by attaching a symbol in Σ to the end of the l-subsequence.

Obviously, there are m *different suffix $(l+1)$-subsequences* for an l-subsequence. Further, assuming that an l-subsequence is the i-th sequence in Θ^l, we denote the index set of its suffix subsequences in Θ^{l+1} by $s(i)$.

Definition 2. *The critical l-subsequence of a sequence*

Given a sequence, the last contiguous l-subsequence in the sequence is called *the critical l-subsequence of the sequence.*

It is easy to get the following lemmas.

Lemma 2. *If two subsequences are different, then their suffix subsequences are also different to each other.*

A *suffix $(l+1)$-subsequence of an l-subsequence* may not be found in given sequences, but it can be found in Θ^{l+1}.

Lemma 3. *For all $l < L_k$, if a sequence contains n of the same l-subsequences and they are different from the critical l-subsequence of the sequence, then the sequence contains n suffix $(l+1)$-subsequences of these l-subsequences; if a sequence contains n of the same subsequences in which the critical l-subsequence also appears, then this sequence contains $n-1$ suffix $(l+1)$-subsequences of these l-subsequences.*

From the results of Fang 1994 and 1997B, one can get

Lemma 4 *(Expansible property, Fang 1997B).*

$$B(V_1, V_2, \cdots, V_{m(l)}, \mathbf{0}) = B(V_1, V_2, \cdots, V_{m(l)})$$

where V_i denotes the i-th row entries of all distributions U_k $(k = 1, 2, \cdots, s)$, i.e.,

$$V_i := (p_{i1}, p_{i2}, \cdots, p_{is}), \quad i = 1, 2, \cdots, m(l), \quad \textit{and} \quad \mathbf{0} = (0, 0, \cdots, 0).$$

Remark. It has to be noted that the difinitions of U_k and V_i are different. We call $\mathbf{0} = (0, 0, \cdots, 0)$ a *0-row of distributions,* and it can be inserted in any two rows of distributions due to the symmetry of this function.

Lemma 5 *(Fang,1994).*

$$\textit{If} \ \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = a > 0, \quad p_i > 0 \ \textit{and} \ \ q_i > 0, \ \textit{then} \ \ \sum_{i=1}^{n} p_i \log(p_i/q_i) \geq 0$$

with equality iff $p_i = q_i$ for all i.

If we define $0 \cdot \log 0 = 0$ and $0 \cdot \log(0/0) = 0$, then the condition of this lemma can be weakened to $\sum_{i=1}^{n} p_i = a$ and $\sum_{i=1}^{n} q_i = a$ where $p_i \geq 0$ and $q_i > 0$, or $p_i = 0$ and $q_i = 0$.

The proof of Theorem 2. From Fang(1994), we know that $B(U_1^l, \cdots, U_s^l)$ satisfies Conditions A, B, and C, so we only need proofs of Condition D and E.

The proof of Condition D.

First of all, if $U_1^l = U_2^l = \cdots = U_s^l$, then $B(U_1^l, \cdots, U_s^l) = 0$ from Condition C and $B(U_1^{l+1}, \cdots, U_s^{l+1}) \geq 0$ from Condition A; these imply that $B(U_1^{l+1}, \cdots, U_s^{l+1}) \geq B(U_1^l, \cdots, U_s^l)$. So we assume that $U_1^l = U_2^l = \cdots = U_s^l$ doesn't hold in the following proof. In addition, it is natural to assume that s distributions in a group are irrelevant (independent) to each other; moreover, we also assume that $\{U_k^l\}$ and $\{U_k^{l+1}\}$ can be expanded to any dimension by Lemma 4.

Let $\Psi = B(U_1^{l+1}, \cdots, U_s^{l+1}) - B(U_1^l, \cdots, U_s^l)$ where $U_k^l \in \Gamma^l, U_k^{l+1} \in \Gamma^{l+1}$, and $l \leq (\min_{k \in K} L_k) - 1$. If the minimum of Ψ is greater than or equal to zero, the conclusion follows immediately, i.e., it is enough to know what kind of distribution sets $\{U_k^l\}$ and $\{U_k^{l+1}\}$ let Ψ take a minimum and what this minimum is. Thus, the basic idea of this proof is: we first find some necessary conditions of taking the minimum of Ψ, then find a unique subset in Γ^l which satisfies the necessary conditions, finally prove that any distribution set in this subset will let $\Psi \geq 0$.

Let $a = L_k - l$; from the definition of complete information sets in Section 2, we have

$$U_k^l : (p_{1k}^l, p_{2k}^l, \cdots, p_{m(l)k}^l) \ \text{ and } U_k^{l+1} : (p_{1k}^{l+1}, p_{2k}^{l+1}, \cdots, p_{m(l+1)k}^{l+1})$$

where

$$p_{ik}^l = \frac{n_{ik}^l}{(L_k - l + 1)} = \frac{n_{ik}^l}{a + 1} \ \text{ and } \ p_{ik}^{l+1} = \frac{n_{ik}^{l+1}}{(L_k - l)} = \frac{n_{ik}^{l+1}}{a}, \quad i = 1, \cdots, m(l).$$

If the critical l-subsequence is not in the set of n^l_{ik} subsequences, then

$$n^l_{ik} = \sum_{j\in s(i)} n^{l+1}_{jk},$$

from Lemma 3; further,

$$n^l_{ik} - 1 = \sum_{j\in s(i)} n^{l+1}_{jk},$$

if the critical subsequence is one of n^l_{ik} subsequences.

1. Using distribution sets $\{U^l_k\}$ and $\{U^{l+1}_k\}$, we can construct a new distribution set $\{Q^{l+1}_1, Q^{l+1}_2, \cdots, Q^{l+1}_s\}$:

$$Q^{l+1}_k : (q^{l+1}_{1k}, q^{l+1}_{2k}, \cdots, q^{l+1}_{m(l+1)k}), \quad k = 1, 2, \cdots, s,$$

where $q^{l+1}_{jk} = m^{l+1}_{jk}/(a+1)$ and m^{l+1}_{jk} is defined as follows: given a $k \in K$, for the case that $n^l_{ik} - 1 = \sum_{j\in s(i)} n^{l+1}_{jk}$, we choose an index number $r \in s(i)$ with $n^{l+1}_{rk} < a$, and let $m^{l+1}_{rk} = n^{l+1}_{rk} + 1$; for other $j \neq r$, let $m^{l+1}_{jk} = n^{l+1}_{jk}$. It is obvious that $1 \leq m^{l+1}_{rk} < a+1$ and $0 < q^{l+1}_{rk} < 1$. It should be mentioned that the specific number r of a sequence may be different from those of other sequences, and we denote the set of these numbers of s sequences by C_1; the row r of the distribution Q^{l+1}_k is called *the specific row r of the distribution Q^{l+1}_k*. Thus one gets that

$$p^l_{ik} = \frac{n^l_{ik}}{a+1} = \sum_{j\in s(i)} \frac{m^{l+1}_{jk}}{a+1} = \sum_{j\in s(i)} q^{l+1}_{jk} \tag{5}$$

holds for all $i \in \{1, 2, \cdots, m(l)\}$ and $k = 1, 2, \cdots, s$.

According to the recursive property of $B(U_1, U_2, \cdots, U_s)$ (Fang 1997B), one directly gets

$$B(Q^{l+1}_1, Q^{l+1}_2, \cdots, Q^{l+1}_s) \ \geq \ B(U^l_1, U^l_2, \cdots, U^l_s), \tag{6}$$

but we can prove (6) as follows.

For each $i \in \{1, 2, \cdots, m(l)\}$, we have

$$B_i(U^l_1, U^l_2, \cdots, U^l_s) = \sum_{k=1}^{s} p^l_{ik} \log \frac{s \cdot p^l_{ik}}{\sum_{k=1}^{s} p^l_{ik}} = \sum_{k=1}^{s} (\sum_{j\in s(i)} q^{l+1}_{jk}) \log \frac{s \cdot \sum_{j\in s(i)} q^{l+1}_{jk}}{\sum_{k=1}^{s}\sum_{j\in s(i)} q^{l+1}_{jk}}$$

from (5); further, for the index set $s(i) \subset \{1, 2, \cdots, m(l+1)\}$, one gets

$$B_{s(i)}(Q^{l+1}_1, Q^{l+1}_2, \cdots, Q^{l+1}_s) = \sum_{j\in s(i)} \sum_{k=1}^{s} q^{l+1}_{jk} \log \frac{s \cdot q^{l+1}_{jk}}{\sum_{k=1}^{s} q^{l+1}_{jk}}$$

according to (1). Using the last two formulae, we obtain

$$\begin{aligned}\Omega_i := \; & B_{s(i)}(Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1}) - B_i(U_1^l, U_2^l, \cdots, U_s^l) \\ & = \sum_{j\in s(i)} \sum_{k=1}^s q_{jk}^{l+1} (\log \tfrac{s\cdot q_{jk}^{l+1}}{\sum_{k=1}^s q_{jk}^{l+1}} - \log \tfrac{s\cdot \sum_{j\in s(i)} q_{jk}^{l+1}}{\sum_{k=1}^s \sum_{j\in s(i)} q_{jk}^{l+1}}) \\ & = \sum_{j\in s(i)} \sum_{k=1}^s q_{jk}^{l+1} \log \tfrac{q_{jk}^{l+1} \cdot \sum_{k=1}^s \sum_{j\in s(i)} q_{jk}^{l+1}}{\sum_{k=1}^s q_{jk}^{l+1} \cdot \sum_{j\in s(i)} q_{jk}^{l+1}}.\end{aligned} \tag{7}$$

From

$$\begin{aligned}& \sum_{j\in s(i)} \sum_{k=1}^s \tfrac{\sum_{k=1}^s q_{jk}^{l+1} \cdot \sum_{j\in s(i)} q_{jk}^{l+1}}{\sum_{k=1}^s \sum_{j\in s(i)} q_{jk}^{l+1}} = \sum_{j\in s(i)} (\sum_{k=1}^s q_{jk}^{l+1}) \tfrac{\sum_{k=1}^s \sum_{j\in s(i)} q_{jk}^{l+1}}{\sum_{k=1}^s \sum_{j\in s(i)} q_{jk}^{l+1}} \\ & = \sum_{j\in s(i)} \sum_{k=1}^s q_{jk}^{l+1}\end{aligned}$$

and Lemma 5, one gets $B_{s(i)}(Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1}) - B_i(U_1^l, U_2^l \cdots, U_s^l) \geq 0$ for each i, i.e., (6) holds.

2. Now we discuss $\Delta = B(U_1^{l+1}, \cdots, U_s^{l+1}) - B(Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1})$; from the continuity and boundedness of the function B, Δ will have a minimum; we are interested in investigating what kind of distribution sets $\{U_k^{l+1}\}$ and $\{Q_k^{l+1}\}$ may let Δ take the minimum.

(1). We construct another distribution set $\{W_1^1, W_2^1, \cdots, W_s^1\}$, where

$$W_k^1 : (w_{1k}^{(1)}, w_{2k}^{(1)}, \cdots, w_{m(l+1)k}^{(1)}), \quad k = 1, 2, \cdots, s,$$

from $\{Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1}\}$ by changing exactly one distribution in $\{Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1}\}$, say Q_v^{l+1}, where v is an arbitrary index in $\{1, 2, \cdots, s\}$; that is, let $W_v^1 = U_v^{l+1}$ and $W_k^1 = Q_k^{l+1}$, $k = 1, 2, \cdots, s$ and $k \neq v$; it means that $w_{ik}^{(1)} = q_{ik}^{l+1}$ for $i = 1, 2, \cdots, m(l+1)$, $k = 1, 2, \cdots, s$ and $k \neq v$; We postulate that the number of matched subsequences corresponding to $w_{ik}^{(1)}$ is $h_{ik}^{(1)}$. Thus, $h_{iv}^{(1)} = n_{iv}^{l+1}$ for $i = 1, 2, \cdots, m(l+1)$. For the *specific row* r of the distribution Q_v^1, it is easy to see that $h_{rv}^{(1)} = n_{rv}^{l+1} = m_{rv}^{l+1} - 1$ (refer to Part A.1 of this proof). So we have: for $k = v$,

$$h_{rv}^{(1)} = m_{rv}^{l+1} - 1 = n_{rv}^{l+1} \text{ for } r \in C_1,$$

$$h_{iv}^{(1)} = m_{iv}^{l+1} = n_{iv}^{l+1} \text{ for } i = 1, \cdots, m(l+1) \text{ and } i \neq r,$$

$$w_{iv}^{(1)} = h_{iv}^{(1)}/a = n_{iv}^{l+1}/a = m_{iv}^{l+1}/a = (a+1)q_{iv}^{l+1}/a \text{ for } i \neq r, \text{ and}$$

$$w_{rv}^{(1)} = h_{rv}^{(1)}/a = n_{rv}^{l+1}/a = (m_{rv}^{l+1} - 1)/a = ((a+1)q_{rv}^{l+1} - 1)/a \quad \text{for } r \in C_1.$$

For clarity of presentation, we sometimes use the subscript variable n instead of k in the following deductions, so we have

$$\sum_{n=1}^s w_{in}^{(1)} = \sum_{n=1}^s q_{in}^{l+1} + q_{iv}^{l+1}/a \text{ when } i \neq r, \text{ and}$$

$$\sum_{n=1}^{s} w_{rn}^{(1)} = \sum_{n=1}^{s} q_{rn}^{l+1} + q_{rv}^{l+1}/a - 1/a \text{ when } i = r.$$

Let us investigate the difference $\Delta^1 = B(W_1^1, W_2^1, \cdots, W_s^1) - B(Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1})$, where Δ^1 equals the sum of the following 4 parts:
the part (a), $k = v$ and $i \neq r$,

$$\begin{aligned}
\Delta_1^1 &= \sum_{\substack{i=1 \\ i\neq r}}^{m(l+1)} \Big(w_{iv}^{(1)} \log(s \cdot w_{iv}^{(1)} / \textstyle\sum_{n=1}^{s} w_{in}^{(1)}) - q_{iv}^{l+1} \log(s \cdot q_{iv}^{l+1} / \textstyle\sum_{n=1}^{s} q_{in}^{l+1})\Big) \\
&= \sum_{\substack{i=1 \\ i\neq r}}^{m(l+1)} \Big((q_{iv}^{l+1} \cdot (a+1)/a) \log(s \cdot (q_{iv}^{l+1} \cdot (a+1)/a)/(\textstyle\sum_{n=1}^{s} q_{in}^{l+1} + q_{iv}^{l+1}/a)) - \\
&\quad - q_{iv}^{l+1} \log(s \cdot q_{iv}^{l+1} / \textstyle\sum_{n=1}^{s} q_{in}^{l+1})\Big);
\end{aligned}$$

the part (b), $k = v$ and $i = r$,

$$\begin{aligned}
\Delta_2^1 &= w_{rv}^{(1)} \log(s \cdot w_{rv}^{(1)} / \textstyle\sum_{n=1}^{s} w_{rn}^{(1)}) - q_{rv}^{l+1} \log(s \cdot q_{rv}^{l+1} / \textstyle\sum_{n=1}^{s} q_{rn}^{l+1}) \\
&= (((a+1)q_{rv}^{l+1} - 1)/a) \log(s \cdot (((a+1)q_{rv}^{l+1} - 1)/a)/(\textstyle\sum_{n=1}^{s} q_{rn}^{l+1} + q_{rv}^{l+1}/a - 1/a)) - \\
&\quad - q_{rv}^{l+1} \log(s \cdot q_{rv}^{l+1} / \textstyle\sum_{n=1}^{s} q_{rn}^{l+1});
\end{aligned}$$

the part (c), $k \neq v$ and $i \neq r$,

$$\begin{aligned}
\Delta_3^1 &= \sum_{\substack{k=1 \\ k\neq v}}^{s} \sum_{\substack{i=1 \\ i\neq r}}^{m(l+1)} \Big(w_{ik}^{(1)} \log(s \cdot w_{ik}^{(1)} / \textstyle\sum_{n=1}^{s} w_{in}^{(1)}) - q_{ik}^{l+1} \log(s \cdot q_{ik}^{l+1} / \textstyle\sum_{n=1}^{s} q_{in}^{l+1})\Big) \\
&= \sum_{\substack{k=1 \\ k\neq v}}^{s} \sum_{\substack{i=1 \\ i\neq r}}^{m(l+1)} \Big(q_{ik}^{l+1} \log(s \cdot q_{ik}^{l+1} / (\textstyle\sum_{n=1}^{s} q_{in}^{l+1} + q_{iv}^{l+1}/a)) - \\
&\quad - q_{ik}^{l+1} \log(s \cdot q_{ik}^{l+1} / \textstyle\sum_{n=1}^{s} q_{in}^{l+1})\Big) \\
&= \sum_{\substack{k=1 \\ k\neq v}}^{s} \sum_{\substack{i=1 \\ i\neq r}}^{m(l+1)} q_{ik}^{l+1} \log((a \cdot \textstyle\sum_{n=1}^{s} q_{in}^{l+1})/(a \cdot \textstyle\sum_{n=1}^{s} q_{in}^{l+1} + q_{iv}^{l+1})\Big) \\
&= \sum_{\substack{i=1 \\ i\neq r}}^{m(l+1)} \log((a \cdot \textstyle\sum_{n=1}^{s} q_{in}^{l+1})/(a \cdot \textstyle\sum_{n=1}^{s} q_{in}^{l+1} + q_{iv}^{l+1}))(\displaystyle\sum_{\substack{k=1 \\ k\neq v}}^{s} q_{ik}^{l+1});
\end{aligned}$$

the part (d), $k \neq v$ and $i = r$,

$$\begin{aligned}
\Delta_4^1 &= \sum_{\substack{k=1\\k\neq v}}^{s} \Big(w_{rk}^{(1)} \log(s \cdot w_{rk}^{(1)} / \textstyle\sum_{n=1}^{s} w_{rn}^{(1)}) - q_{rk}^{l+1} \log(s \cdot q_{rk}^{l+1} / \textstyle\sum_{n=1}^{s} q_{rn}^{l+1})\Big) \\
&= \sum_{\substack{k=1\\k\neq v}}^{s} \big(q_{rk}^{l+1} \log(s \cdot q_{rk}^{l+1} / (\textstyle\sum_{n=1}^{s} q_{rn}^{l+1} + q_{rv}^{l+1}/a - 1/a)) - \\
&\quad - q_{rk}^{l+1} \log(s \cdot q_{rk}^{l+1} / \textstyle\sum_{n=1}^{s} q_{rn}^{l+1})\big) \\
&= \sum_{\substack{k=1\\k\neq v}}^{s} q_{rk}^{l+1} \log(a \cdot (\textstyle\sum_{n=1}^{s} q_{rn}^{l+1}) / (a \cdot \textstyle\sum_{n=1}^{s} q_{rn}^{l+1} + q_{rv}^{l+1} - 1)) \\
&= (\sum_{\substack{k=1\\k\neq v}}^{s} q_{rk}^{l+1}) \cdot \log(a \cdot (\textstyle\sum_{n=1}^{s} q_{rn}^{l+1}) / (a \cdot \textstyle\sum_{n=1}^{s} q_{rn}^{l+1} + q_{rv}^{l+1} - 1)).
\end{aligned}$$

Let

$$q_i = \sum_{\substack{k=1\\k\neq v}}^{s} q_{ik}^{l+1}, \quad i = 1, 2, \cdots, m(l+1).$$

Then

$$\sum_{n=1}^{s} q_{in}^{l+1} = q_i + q_{iv}^{l+1}, \quad i = 1, 2, \cdots, m(l+1).$$

Δ^1 can be rewritten as

$$\begin{aligned}
\Delta^1 &= \Delta_1^1 + \Delta_2^1 + \Delta_3^1 + \Delta_4^1 \\
&= \sum_{\substack{i=1\\i\neq r}}^{m(l+1)} \Big((q_{iv}^{l+1} \cdot (a+1)/a) \log(s \cdot (q_{iv}^{l+1} \cdot (a+1)/a) / (q_i + q_{iv}^{l+1} + q_{iv}^{l+1}/a)) - \\
&\quad - q_{iv}^{l+1} \log(s \cdot q_{iv}^{l+1} / (q_i + q_{iv}^{l+1}))\Big) + \\
&\quad (((a+1) q_{rv}^{l+1} - 1)/a) \log(s \cdot (((a+1) q_{rv}^{l+1} - 1)/a) / (q_r + q_{rv}^{l+1} + q_{rv}^{l+1}/a - 1/a)) - \\
&\quad - q_{rv}^{l+1} \log(s \cdot q_{rv}^{l+1} / (q_r + q_{rv}^{l+1})) + \\
&\quad \sum_{\substack{i=1\\i\neq r}}^{m(l+1)} q_i \cdot \log(a \cdot (q_i + q_{iv}^{l+1}) / (a \cdot (q_i + q_{iv}^{l+1}) + q_{iv}^{l+1})) + \\
&\quad q_r \cdot \log(a \cdot (q_r + q_{rv}^{l+1}) / (a \cdot (q_r + q_{rv}^{l+1}) + q_{rv}^{l+1} - 1)).
\end{aligned} \tag{8}$$

For the cases that both q_i and q_{iv}^{l+1} equal to zero ($i \in \{1, 2, \cdots, m(l+1)\}$), the variation in Δ^1 is nothing, so we take no account of them in the following proof. Differentiating (8) with respect to q_i ($i = 1, \cdots, m(l+1)$ and $i \neq r$) and q_r respectively, we have

$$\frac{\partial \Delta^1}{\partial q_i} = \log \frac{a \cdot (q_i + q_{iv}^{l+1})}{q_{iv}^{l+1} + a \cdot (q_i + q_{iv}^{l+1})} \leq 0 \quad \text{and} \tag{9}$$

$$\frac{\partial \Delta^1}{\partial q_r} = \log \frac{a \cdot (q_r + q_{rv}^{l+1})}{-1 + q_{rv}^{l+1} + a \cdot (q_r + q_{rv}^{l+1})} \; > \; 0 \tag{10}$$

due to that $a \cdot (q_r + q_{rv}^{l+1}) - (-1 + q_{rv}^{l+1} + a \cdot (q_r + q_{rv}^{l+1})) = 1 - q_{rv}^{l+1} > 0$, $1 > q_{rv}^{l+1} > 0$, and $(a+1) \cdot q_{rv}^{l+1} \geq 1$ (see the Part 1 of this proof).

The variable q_r or q_i is an independent variable relative to q_{rv}^{l+1} and q_{iv}^{l+1}; although q_r and $\sum_{\substack{i=1\\ i\neq r}}^{m(l+1)} q_i$ are dependent, $q_r + \sum_{\substack{i=1\\ i\neq r}}^{m(l+1)} q_i = \sum_{i=1}^{m(l+1)} q_i = \sum_{i=1}^{m(l+1)} \sum_{\substack{k=1\\ k\neq v}}^{s} q_{ik}^{l+1} = s-1$, so Δ^1 is an increasing function w.r.t. q_r and a non-increasing function w.r.t. q_i ($\sum_{\substack{i=1\\ i\neq r}}^{m(l+1)} q_i$) from (9) and (10). In addition, Δ^1 is a continuous function of q_r and q_i. All of them mean that $q_r = 0$ if Δ^1 takes a minimum value, i.e., one of the necessary conditions of Δ^1's taking the minimum is

$$q_r = 0 \quad \text{when} \quad q_{rv}^{l+1} \neq 0. \tag{11}$$

This condition also implies that

$$q_{rv}^{l+1} = \sum_{k=1}^{s} q_{rk}^{l+1} \qquad \text{when } q_{rv}^{l+1} \neq 0.$$

Further, q_i and q_{iv}^{l+1} are independent variables to each other and

$$\frac{\partial^2 \Delta^1}{\partial q_i^2} = \frac{q_{iv}^{l+1}}{(q_{iv}^{l+1} + a \cdot (q_i + q_{iv}^{l+1}))(q_i + q_{iv}^{l+1})} \geq 0,$$

this implies that another necessary condition of Δ^1's taking the minimum is $\frac{\partial \Delta^1}{\partial q_i} = 0$, i.e.,

$$q_{iv}^{l+1} = 0 \qquad \text{if } q_i \neq 0 \quad \text{and} \quad i \neq r$$

from (9); the last condition equals

$$q_i = 0 \qquad \text{if } q_{iv}^{l+1} \neq 0, \quad \text{and} \quad i \neq r. \tag{12}$$

From (12) and (11), we obtain the following necessary condition

$$q_i = 0 \quad \text{if } q_{iv}^{l+1} \neq 0, \quad i = 1, 2, \cdots, m(l). \tag{13}$$

The condition (13) is called the *condition I_0 of the distribution Q_v^{l+1};* this condition means that the corresponding row elements of Q_k^{l+1}, $k = 1, 2, \cdots, s$ and $k \neq v$ to the non-zero row element of Q_v^{l+1} must equal zero.

(2) we construct a distribution set $\{W_1^2, W_2^2, \cdots, W_s^2\}$ from $\{W_1^1, W_2^1, \cdots, W_s^1\}$ by changing exactly one distribution in $\{W_1^1, W_2^1, \cdots, W_s^1\}$, say W_u^1, where u is an arbitrary index in $\{1, 2, \cdots, s\}$ such that $u \neq v$; that is, let $W_u^2 = U_u^{l+1}$ and $W_k^2 = W_k^1$, $k = 1, 2, \cdots, s$ and $k \neq v$; it means that $w_{ik}^{(1)} = q_{ik}^{l+1}$ for $i = 1, 2, \cdots, m(l+1)$, $k = 1, 2, \cdots, s$ and $k \neq v$. We postulate that the number of matched subsequences corresponding to w_{ik}^2 is $h_{ik}^{(2)}$. In this situation, $w_{ik}^{(2)} = w_{ik}^{(1)}$ for $i = 1, 2, \cdots, m(l+1)$, $k = 1, 2, \cdots, s$ and $k \neq u$; for $k = u$,

$$h_{ru}^{(2)} = m_{ru}^{l+1} - 1 = n_{ru}^{l+1} \text{ for } i = r \text{ and } r \in C_1$$

$$h_{iu}^{(2)} = m_{iu}^{l+1} = n_{iu}^{l+1} \text{ for } i = 1, \cdots, m(l+1) \text{ and } i \neq r$$

$$w_{iu}^{(2)} = h_{iu}^{(2)}/a = n_{iu}^{l+1}/a = m_{iu}^{l+1}/a = (a+1)w_{iu}^{(1)}/a \text{ for } i \neq r, \text{ and}$$

$$w_{ru}^{(2)} = h_{iu}^{(2)}/a = n_{iu}^{l+1}/a = (m_{ru}^{l+1}-1)/a = ((a+1)w_{ru}^{(1)}-1)/a \quad \text{for } i = r \text{ and } r \in C_1$$

These mean that we can obtain similar conclusions to (9), (10), (11) and (12) for $\Delta^2 = B(W_1^2, W_2^2, \cdots, W_s^2) - B(W_1^1, W_2^1, \cdots, W_s^1)$.

Continuing to do this, we finally obtain the distribution set $W_1^s, W_2^s, \cdots, W_s^s$, in which $W_k^s = U_k^{l+1}$, $k = 1, 2, \cdots, s$, and

$$\begin{aligned} \Delta &= \Delta^1 + \Delta^2 + \cdots + \Delta^s \\ &= B(W_1^s, W_2^s, \cdots, W_s^s) - B(Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1}) \\ &= B(U_1^{l+1}, U_2^{l+1}, \cdots, U_s^{l+1}) - B(Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1}). \end{aligned}$$

From the last formula, we know that the necessary conditions of $\Delta's$ taking the minimum are a collection of necessary conditions of all Δ^i; this also means that all non-zero entries of all distributions Q_k^{l+1} (or U_k^{l+1}), $k = 1, 2, \cdots, s$ must be located in different rows if Δ takes the minimum. If all distributions of a distribution set satisfy the I_0 condition, then this distribution set is called *an I_0-distribution set;* such distribution sets constitute a unique subset Γ_0^{l+1} in Γ^{l+1}. (In fact, we can easily construct such an I_0-distribution set from any a distribution set $\{U_k^{l+1}\}$ by expanding its dimension and by changing places of some non-zero entries in each distribution).

Finally, Δ is bounded and continuous, so it has a minimum. Since there exists only one subset satisfying the necessary conditions of Δ's taking the minimum, then minimum solutions are certainly in this subset.

3. Let us investigate the difference $\Omega = B(Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1})$ - $B(U_1^l, U_2^l, \cdots, U_s^l)$. According to the construction procedure of $Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1}$, the variation from $U_1^l, U_2^l, \cdots, U_s^l$ to $Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1}$ will lead to a non-negative increase of Ω from (6), and an increase 0 will be the smallest increase of Ω. For an *I_0-distribution set* $\{Q_k^{l+1}\}$, it is easy to see that

$$q_{jk}^{l+1} = \sum_{n=1}^{s} q_{jn}^{l+1} \quad \text{if } q_{jk}^{l+1} \neq 0; \tag{14}$$

in this case, one gets

$$
\begin{aligned}
\Omega &= \sum_{i=1}^{m(l)} \Omega_i = \sum_{i=1}^{m(l)} B_{s(i)}(Q_1^{l+1}, Q_2^{l+1}, \cdots, Q_s^{l+1}) - B_i(U_1^l, U_2^l, \cdots, U_s^l) \\
&= \sum_{i=1}^{m(l)} \sum_{j\in s(i)} \sum_{k=1}^{s} q_{jk}^{l+1} \log \frac{q_{jk}^{l+1} \cdot \sum_{k=1}^{s} \sum_{j\in s(i)} q_{jk}^{l+1}}{\sum_{k=1}^{s} q_{jk}^{l+1} \cdot \sum_{j\in s(i)} q_{jk}^{l+1}} \\
&= \sum_{i=1}^{m(l)} \sum_{j\in s(i)} \sum_{k=1}^{s} q_{jk}^{l+1} \log \frac{\sum_{k=1}^{s} \sum_{j\in s(i)} q_{jk}^{l+1}}{\sum_{j\in s(i)} q_{jk}^{l+1}} \\
&= \sum_{i=1}^{m(l)} \sum_{j\in s(i)} \sum_{k=1}^{s} q_{jk}^{l+1} \log \frac{\sum_{k=1}^{s} p_{ik}^l}{p_{ik}^l}
\end{aligned} \tag{15}
$$

using (5), (7) and (14). According to Lemma 5 and (15), $\Omega = 0$ iff

$$
q_{jk}^{l+1} \cdot \sum_{k=1}^{s} \sum_{j\in s(i)} q_{jk}^{l+1} = \sum_{k=1}^{s} q_{jk}^{l+1} \cdot \sum_{j\in s(i)} q_{jk}^{l+1} \tag{16}
$$

where $q_{jk}^{l+1} \neq 0, \quad j = 1, 2, \cdots, m(l+1)$ and $k = 1, 2, \cdots, s$. Since (14) is the condition letting $\Omega = 0$ for the set $\{Q_k^{l+1}\}$, so is the following condition for the set $\{U_k^l\}$

$$
\sum_{k=1}^{s} p_{ik}^l = p_{ik}^l \text{ when } p_{ik}^l \neq 0, \quad i = 1, 2, \cdots, m(l) \quad \text{and} \quad k = 1, 2, \cdots, s
$$

according to (5), (14) and (16).

4. If a distribution set $\{U_k^l\}$ is *an I_0-distribution set,* it implies that all subsequences of length l are different to each other, and that $\{U_k^{l+1}\}$ and $\{Q_k^{l+1}\}$ obtained from this set $\{U_k^l\}$ are also I_0-distribution sets by Lemma 2.

5. Considering the first, second, third and fourth parts of this proof together, we have the following conclusion: *Under the assumption that $U_1^l = U_2^l = \cdots = U_s^l$ doesn't hold, it is only when a distribution set $\{U_k^l\}$ is an I_0-distribution set that its Ψ may take the minimum value.*

For any I_0-distribution set, we can calculate its Ψ_0 by

$$
\Psi_0 = \Omega_0 + \Delta_0 \quad \text{ or directly using}
$$

$$
\begin{aligned}
\Psi_0 &= B(U_1^{l+1}, U_2^{l+1}, \cdots, U_s^{l+1}) - B(U_1^l, U_2^l, \cdots, U_s^l) \\
&= \sum_{k=1}^{s} \sum_{i=1}^{m(l+1)} p_{ik}^{l+1} \log \frac{s \cdot p_{ik}^{l+1}}{p_{ik}^{l+1}} - \sum_{k=1}^{s} \sum_{i=1}^{m(l)} p_{ik}^l \log \frac{s \cdot p_{ik}^l}{p_{ik}^l} \\
&= s \cdot \log s - s \cdot \log s \\
&= 0.
\end{aligned} \tag{17}
$$

Since the difference Ψ_0 of any distribution set of all distribution sets satisfying the necessary conditions equals zero, the minimum of Ψ is equal to 0, i.e., $B(U_1^{l+1}, U_2^{l+1}, \cdots, U_s^{l+1}) \geq B(U_1^l, U_2^l, \cdots, U_s^l)$.

The proof of Condition E. If there doesn't exist any common subsequence of length l in a group of sequences, then all non-zero entries of all l-dimensional distributions of their complete information sets are located in different rows. This property will be valid for all $n > l$. By a similar calculation to (17), the conclusion is obtained immediately.

References

[1] S.F. Altschul and D.J. Lipman, Trees, stars and multiple sequence alignment, Journal of applied mathematics, 49(1) : 197-209, 1989.

[2] M.L. Baba, L. Darga, L. Goodman, Evolution of cytochrome C investigated by the maximum parsimony method, Journal of molecular evolution, 17: 197-213, 1981.

[3] D.J. Bacon and W. F. Anderson, Multiple sequence alignment, Journal of molecular biology, 191: 153-161 , 1986.

[4] G.J. Barton and M.J.E. Sternberg, Evolution and improvements in the automatic alignment of protein sequences, Protein engineering 1: 89-94, 1987.

[5] S. C. Chan , A. K.C. Wong and D. K. Y. Chiu, A survey of multiple sequence comparison methods, Bulletin of mathematical biology, 54(4): 563-598 1992.

[6] D. Durand and M. Farach, On the design of optimization criteria for multiple sequence alignment, DIMACS technical report 96-53, Jan. 1997.

[7] Weiwu Fang, The disagreement degree of multi-person judgments in additive structure, Mathematical social sciences, 28 (2): 85-111, 1994.

[8] Weiwu Fang, On a global optimization problem in the study of information discrepancy, Journal of global optimization, 11: 387-408, 1997A.

[9] Weiwu Fang , FDOD function and the information discrepancy contained in multiple probability distributions, DIMACS technical report 97-36, Aug, 1997B. (or Lecture Notes in Operations Research 3, 231-243, 1998.)

[10] M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner, and J. Ziv, On the entropy of DNA : Algorithm and measurements based on memory and rapid convergence, in Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithm, 1994.

[11] L. Gatlin, Information theory and the living system, New York Columbia University Press, 1978.

[12] A. Hariri, B. Weber and J. Olmsted III, On the validity of Shannon-information calculations for molecular biological sequences, Journal of theoretical biology, 147: 235-254, 1990.

[13] J.N.Kapur and H.K.Kesavan, Entropy Optimization Principles with Applications. Academic Press, San Diego, 1992.

[14] S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.

[15] D.J. Lipman, S.F. Altschul and J.D. Kececioglu, A tool for multiple sequence alignment, Proc. Natl. Acad. Sci. USA, 86: 4412-4415, 1989.

[16] D.M. Loewenstern and P.N. Yianilos, Significantly lower entropy estimates for natural DNA sequences, DIMACS Technical Report 96-51, 1996.

[17] A.V. Lukashin, J. Engelbrecht, and S. Brunak, Multiple alignment using simulated annealing, Nucl. Acids Research, 20: 2511-2516, 1992.

[18] M.A. McClure , T.K. Vasi and W. M. Fitch, Comparative analysis of multiple protein-sequence alignment methods, Mol. Biol. Evol., 11: 571-592, 1994.

[19] G. R. Moore and G. W. Pettigrew, Cytochrome C evolutionary, structural and physicochemical aspects, Springer-Verlag, 1990.

[20] S.B. Needleman and C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins , J. molec. evol. 19: 153-170, 1972.

[21] P. A. Pevzner and M. S. Waterman, Open combinatorial problems in computational molecular biology, Proceedings of the third symposium on theory of computing and systems, Tel Aviv, Israel, 1995.

[22] F. S. Roberts and R.D. Luce, Axiomatic thermodynamics and extensive measurement, Syntheses 18: 311-326, 1968.

[23] B. Mirkin and F. S. Roberts, Consensus functions and patterns in molecular sequences, Bulletin of mathematical biology, 55(4): 695 -713, 1994.

[24] F. S. Roberts, Measurement theory with applications to decision making, utility, and the social sciences, Addison-Wesley, 1979.

[25] C.E. Shannon, A mathematical theory of communication, Bell System Tech. 27: 379-423, 623-656, 1948.

[26] L. Wang and T. Jiang, On the complexity of multiple sequence alignment, Journal of computational biology, 1(4): 337-348, 1994.

[27] M. S. Waterman , M. Eggert, and E. Lander, Parametric sequence comparisons, Proc. Natl. Acad. Sci. USA, 89 : 6090-6093, 1992.