行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※
※
※　統計自由能偶核對應法之發展及應用
※
※
※
※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：■個別型計畫　　□整合型計畫
計畫編號：NSC 89-2113-M-007-050-
執行期間：89年08月01日至90年07月31日

計畫主持人：黃鎮剛
共同主持人：

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

執行單位：國立交通大學生物科技系

中　華　民　國　90　年　5　月　25　日

Analysis of Structural Information Content in Peptide Fragments

Chen-hsiung Chan[1], Jenn-Kang Hwang[2]*

[1]Department of Life Science
National Tsing Hua University
Hsinchu, Taiwan 30013

[2]Department of Biological Science & Technology &
Institute of Bioinformatics
National Chao Tung University
Hsinchu, Taiwan

* To whom correspondence should be addressed.

Keywords:
2nd structure, protein sequence, distance distribution function

Abstract

We have developed a novel approach to analyze structural information contents in protein fragments. This approach can give quantitative measure of non-randomness of sequence fragment in the conformational space. In this report, we analyze the relationship between protein sequence and its structural information content. We also suggest that the "structural unit" of proteins could be of a optimal length of 6 residues.

Introduction

In the last decades, scientists have different opinions over the non-randomness of protein sequence. It is a generally accepted view (Ptitsyn, 1991) that protein sequences are “slightly edited” random sequences, but it is still not clear how to quantify the degree of non-randomness. To account for the randomness or non-randomness of protein sequences, people have developed various approaches that are based on the Fourier transformation (Berman *et al.*, 1994), information theory (Weiss *et al.*, 2000) and other techniques, yet the results are still inconclusive. Recently, Keefe and Szostak (2001) successfully produced functional protein sequences from random sequence library, experimentally showing that the functionality of a protein could result from an almost random sequence, of which only a small fraction is “slightly edited”. It is then natural to treat protein sequence as an ensemble of peptide fragments or ‘units’, which carry varying degrees of randomness. The concept of sequence “unit” is tempting, but the definition of which remains unclear. There are attempts to identify the sequence “unit”, for example, Kabsch and Sander (1984) showed that an identical pentapeptide fragment adopts different conformations in different proteins. Argos (1987) later made extensive analysis of peptide conformations and concluded that peptide fragments have different structural preferences in different protein environments. Macchiato and coworkers (1985) showed

that protein sequence has a correlation order of 3 or 4, which is close to the smallest possible secondary structure element. Rackovsky (1998) found in TIM barrels a periodicity that could be roughly mapped to strands; People (?) has developed methods to map separated peptide fragments to physically meaningful units. Using a simplified spin-glass-like model, Saito et al (1997) showed that constraints on local configuration increase the foldability of proteins and concluded that peptide fragments may carry variable amounts of structural information.

Methods

For a given set of sequence fragment x, we have a associated vector P_x,

$$\mathbf{P}_x = \left(p_x^B, p_x^E, ..., p_x^U\right) \qquad (1)$$

where p_x^t is the probability of t type secondary structure elements in the sequence x, and $t \in \{B, E, ,..., U\}$. The definition of the secondary structure designators, i.e., $B, E, ,..., U$, follows that of Kabsch & Sander (1983) and is given in Table 1. The distance between $\mathbf{P}_x$ and $\mathbf{P}_y$ is defined as $D_{xy} = |\mathbf{P}_x - \mathbf{P}_y|$. It is convenient to define a reference set $\mathbf{P}_0$ that can consist of all entries in Protein Data Bank (PDB). But it should be noted that the reference state could also consist of a group of proteins characterized by certain properties. The distance between $\mathbf{P}_x$ and $\mathbf{P}_0$ is defined by,

$$D_x = |\mathbf{P}_x - \mathbf{P}_0| \qquad (2)$$

which, as will be shown below, gives the measure of the relative amount of secondary structural information contained in a given peptide sequence x. The distance distribution function (*ddf*) is given by

$$R_X(d) = \sum_{D_x} \delta(d - D_x)$$

where $¢(d)=1$, if $d=0$ and $¢(d)=0$, if $d\neq 0$, and $x\in X$, which is a set of specific sequence fragments. The function $R_X(d)$ gives a complete profile of secondary structure of the sequence elements belonging to the set X. Our formulation is rather general and can be applied to any set consisting of the sequence fragments of a single sequence chain of a protein, or those of a protein family, as long as the sequence elements share a common property such as a fixed sequence length, a specific sequence pattern or other structural characterizations. In this study, we will study the sets that are given as a collection of identical peptide fragments.

Implementation

The secondary structure assignment was taken from the DSSP database (Kabsch and Sander, 1983). The definition of each token in secondary structure designation is listed in Table 1. The reference sequence set contains all non-redundant entries from Protein Data Bank. All programs used in this study were written in Perl and shell script. These programs are portable, and should be able to run on most computing platforms without further modifications. Most data generated in this study has been inserted into a SQL based database for fast look up and cross-reference. We construct X by scanning the distribution of secondary structure over the sequence fragments in the reference set using a sliding window of size l. The sizes of the sliding windows are ranging from 1 to 16 amino acids. It should be noted that, while the construction of the set X depends on the length l, the distance D_x defined by Eq. 2 does not. Hence, D_x offers a convenient measure of structural information contained in a set of sequence fragments.

Results and Discussion

The distribution of secondary structure of the reference set S_0 is shown in Figure 1. The distribution is similar to the result of previous work. The most prominent secondary structure elements are H, α-helix, and E, the extended strand. It is interesting to note the third highest peak is U, which is the unassigned secondary structure and indicates the existence of a rather large portion of un-structured sequences in the PDB Data Bank. As an illustration to the meaning of D_x in Eq. 2, we compare the distributions of S_x (Fig. 2a), where $x = \text{KSELKEL}$, and S_y (Fig. 2b), where $y = \text{GKAKYKA}$, with that of the reference state S_0. Most elements in S_x assume one helical conformation, while those in S_y adopts a variety of secondary structure elements. Table 2. lists some typical examples of these two sets. The calculated values of D_x and D_y are 0.76 and 0.04, respectively. The small value D_y is due to the fact that S_y has an essential identical distribution of secondary structure to that of S_0. The value of D_x offers a quantitative measure to the number of possible conformers that could be adopted by a given peptide fragment x; in other word, the value of D_x indicates the non-randomness of the structure of the peptide fragment. The larger the value of D is, the less random the peptide fragment will be in the conformational space.

Fig. 3 shows the distance profile of the set composed of peptide sequences with lengths ranging from 1 to 16 (solid line), and that of the randomized data set (dashed line). The *ddf* of the former set is basically bell-shaped except for two peaks at a distance of 0.77 and 0.88, which corresponds to ~-helix and extended -strands, respectively. The *ddf* of the randomized data set is quite different. The entire distribution shifts to left and the two peaks for ~-helix and -strands disappear. The distances of the peptide fragments of the randomized data set are significantly lower as expected, agreeing our previous observation that smaller distance implies more randomness of a give sequence fragment in the adopted conformations.

In Figure 4 we show the *ddf* of tri-peptides (solid line), hexa-peptides (dotted line) and 16-peptides (dashed line), respectively. It is interesting note that while the *ddf* of hexa-peptides significantly shifts to the right, the *ddf* of 16-peptides shifts back to the left. These results indicate that hexa-peptides have more definite 2nd structures than 16-peptides; in other word, the structural information of hexa-peptides appears more non-random than that of 16-peptides. We did compare all *ddf*s of peptides of a length ranging from 1 to 16, and found that *ddf* keeps shifting to the right until reaches the length of 6, and then the *ddf* will start to shift back to the left in the distance. These results suggest a tempting idea of a basic "structural unit" in the protein sequences, and the length of which can well be set to the length of 6 residues.

We also applied our approach to a whole protein chain instead of a small peptide fragment. The result is shown in Fig. 5. It could be seen that the *ddf* of a protein chain is much smaller than that of a peptide fragment in general, and that it is rather close to that of randomize peptide fragments (dashed line in Fig. 3).

References

Argos, P. (1987) Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. *J. Mol. Biol.*, **197**, 331-348.

Berman, A. L., Kolker, E., and Trifonov, E. N. (1994) Underlying order in protein sequence organization. *Proc. Natl. Acad. Sci. USA*, **91**, 4044-4047.

Bystroff, C. and Baker, D. (1998) Prediction of local structure in proteins using a library of sequence structure motifs. *J. Mol. Biol.*, **281**, 565-577.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Kabsch, W. and Sander, C. (1984) On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci. USA*, **81**, 1075-1078.

Keefe, A. D. and Szostak, J. W. (2001) Functional proteins from a random-sequence library. *Nature*, **410**, 715-718.

Macchiato, M. F., Cuomo, V., and Tramontano, A. (1985) Determination of the autocorrelation orders of proteins. *Eur. J. Biochem.*, **149**, 375-379.

Rackovsky, S. (1998) "Hidden" sequence periodicities and protein architecture. *Proc. Natl. Acad. Sci. USA*, **95**, 8580-8584.

Rahman, R. S. and Rackovsky, S. (1995) Protein sequence randomness and sequence/structure correlations. *Biophys. J.*, **68**, 1531-1539.

Saito, S., Sasai, M., and Yomo, T. (1997) Evolution of the folding ability of proteins through functional selection. *Proc. Natl. Acad. Sci. USA*, **94**, 11324-11328.

Spang, R. and Vingron, M. (2001) Limits of homology detection by pairwise sequence comparison. *Bioinformatics*, **17**, 338-342.

Weiss, O., Jim\264enez-Montano, M. A., and Herzel, H. (2000) Information content of protein sequences. *J. theor. Biol.*, **206**, 379-386.

White, S. H. and Jacobs, R. E. (1993) The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. *J. Mol. Evol.*, **36**, 79-95.
Table 1: The definition of tokens in secondary structure designation follows that of Kabsch and Sander (1983).

Table 1. Table 1: The definition of tokens in secondary structure designation follows that of Kabsch and Sander (1983).

Token	Definition
B	isolated β-bridge
E	extended strand
G	3_{10}-helix
H	α-helix
I	π-helix
S	bend
T	H-bonded turn
U	Others

Table 2. Some typical examples of the different secondary structures adopted by sequences KSELKEL and GKAKYKA. The letter after the PDB code is the designator of the chain to which where the sequence belongs. .hree peptide fragments with their secondary structure assignment and Id of the PDB entries where they could be found. While the sequence KSELKEL contains mostly helical structure, the sequence GKAKYKA contains a variety of different secondary structurse. The calculated values of these two sequences are 0.76 and 0.04, respectively.

Sequence	Secondary structure	PDB code
KSELKEL	HHHHHHH	1b4c a
	HHHHHHH	1cfp a
	HHHHHHH	1dt7 a
	HHHHHHH	1mho d
	HHHHHHH	1qlk a
	HHHHHHH	1sym a
	THHHHHH	1uwo a
GKAKYKA	SEEEEET	1bw8 a
	SEEEEET	1bxx a
	HHHHHUU	1bzy a
	TTTSSUU	1d6n a
	HHHHSUU	1hmp a
	SEEEEEG	1i31 a

Figure 1. The distribution of secondary structure elements distribution of the reference set.

Figure 2. (a) The distribution of secondary structure elements of peptide fragment *KSELKEL* (filled) and the reference set (open). (b) The distribution of secondary structure elements of peptide fragment *GKAKYKA* (filled) and the reference set (open)..

Figure 3. The distance distribution function of peptide fragments of a length ranging from 1 to 16 residues (solid line) and that of a randomized data set (dashed line).

Figure 4. The distance distribution function of tri-peptide (solid line), penta-peptide (dotted line) and 16-peptide (dashed line).

Figure 5. The distance distribution function of chains of Protein Data Bank.

Fig 1.

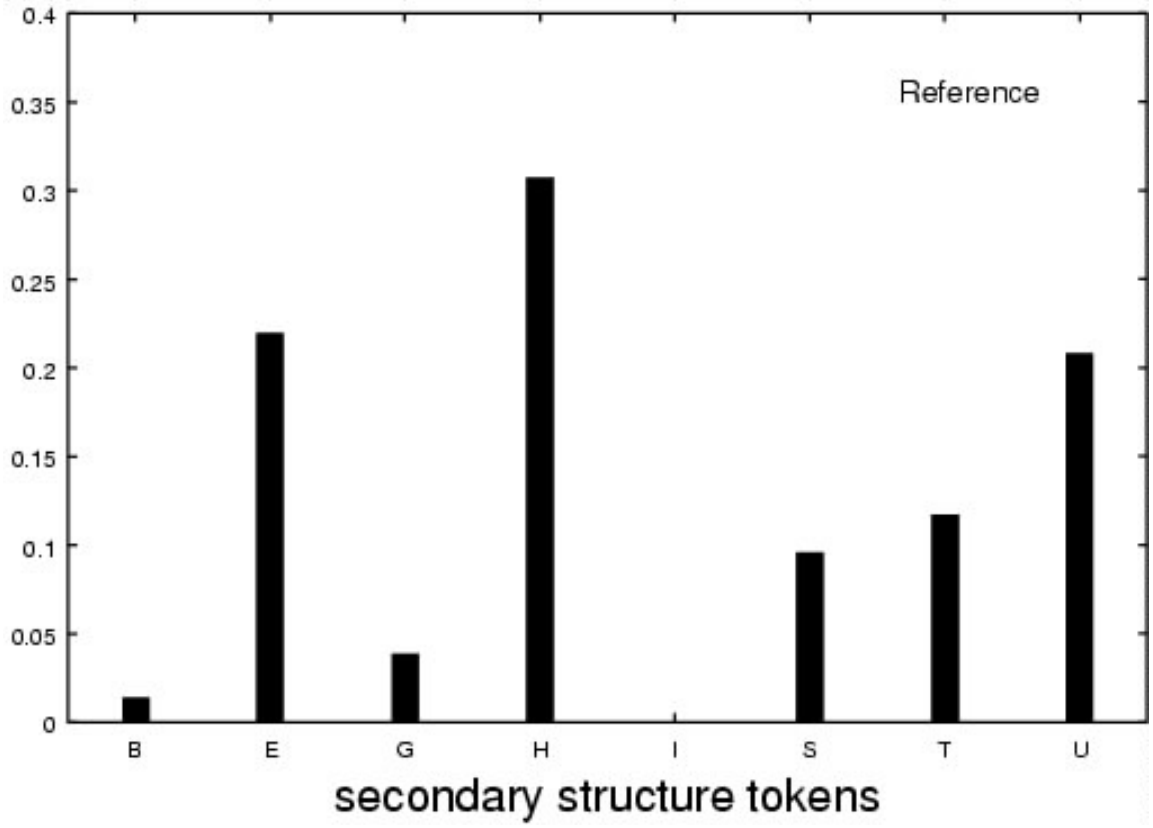
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Reference
B
E
G
H
I
S
T
U
secondary structure tokens

Fig 2.

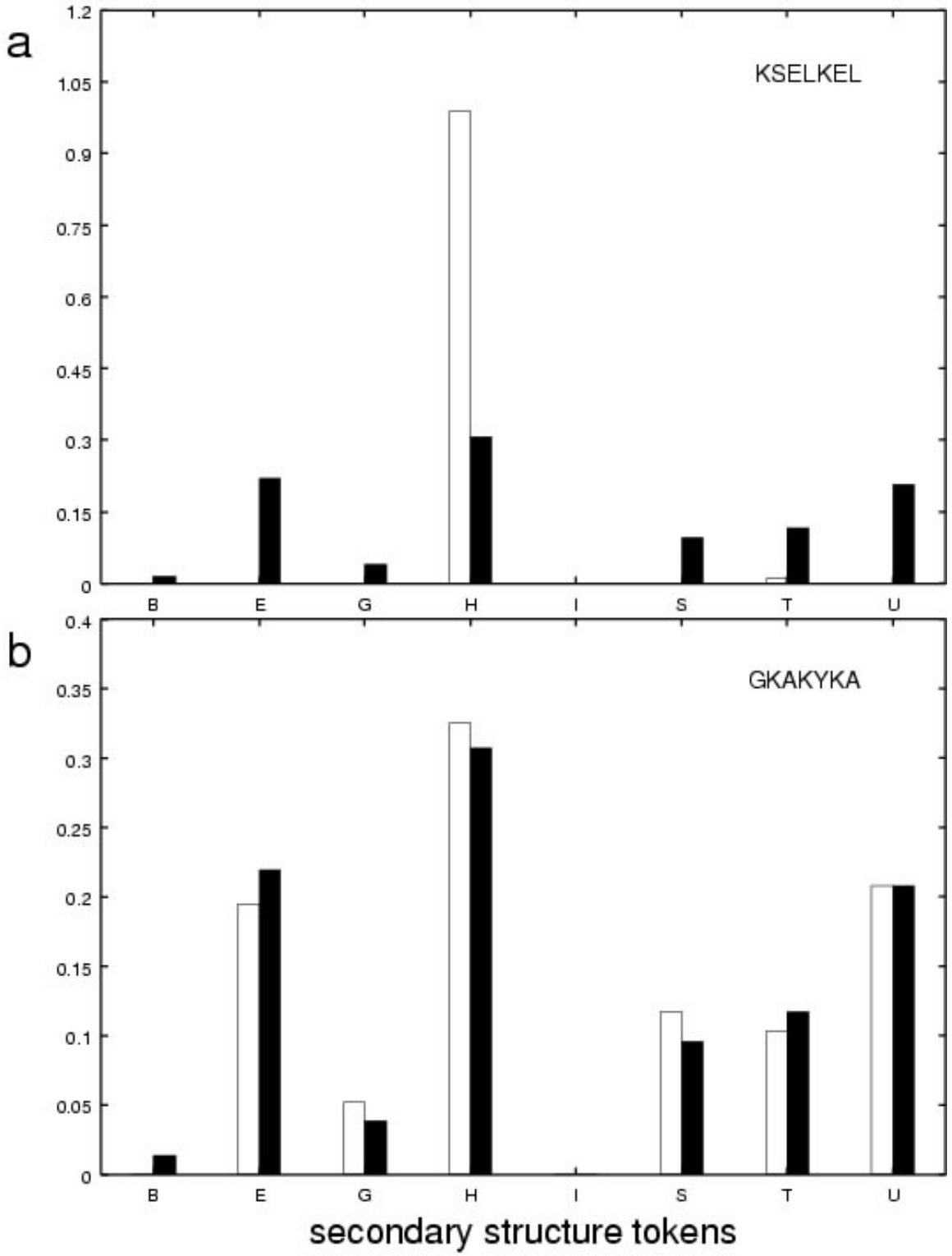
a
KSELKEL
1.2
1.05
0.9
0.75
0.6
0.45
0.3
0.15
0
B
E
G
H
I
S
T
U
b
GKAKYKA
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
B
E
G
H
I
S
T
U
secondary structure tokens

Fig 3.

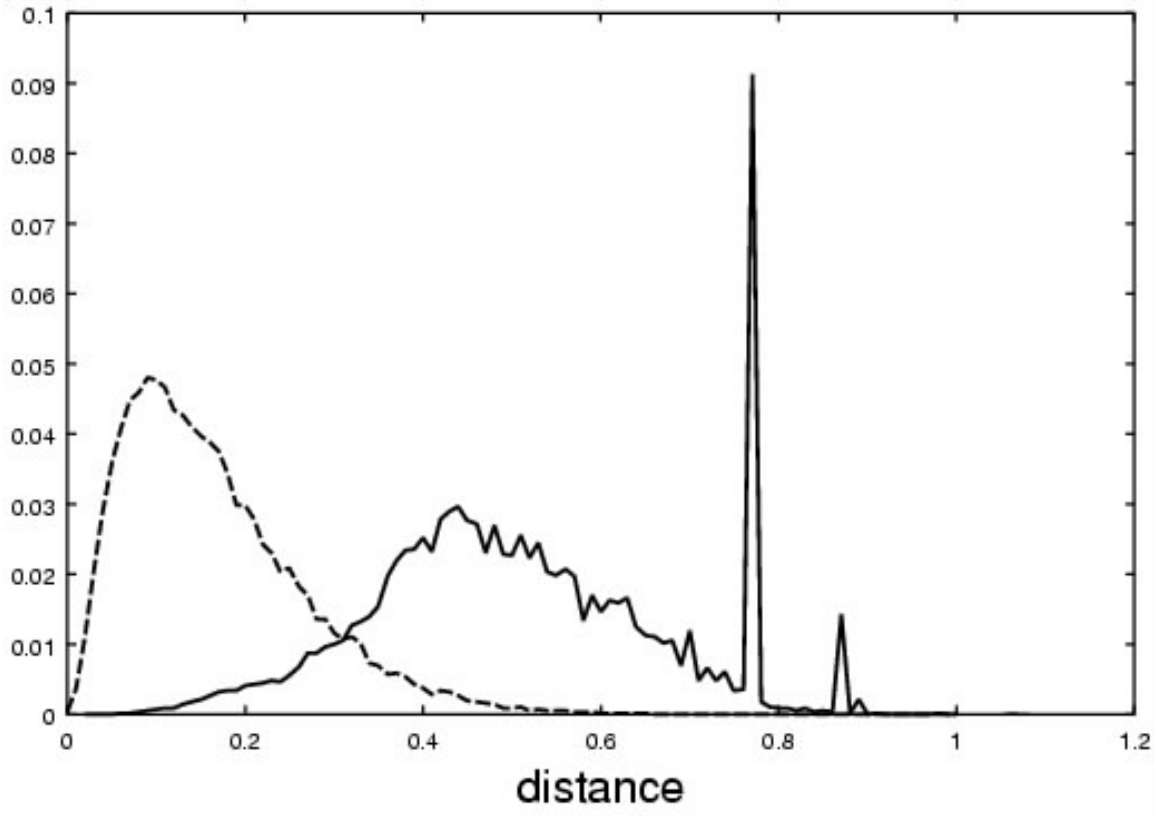
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
0.2
0.4
0.6
0.8
1
1.2
distance

Fig 4.

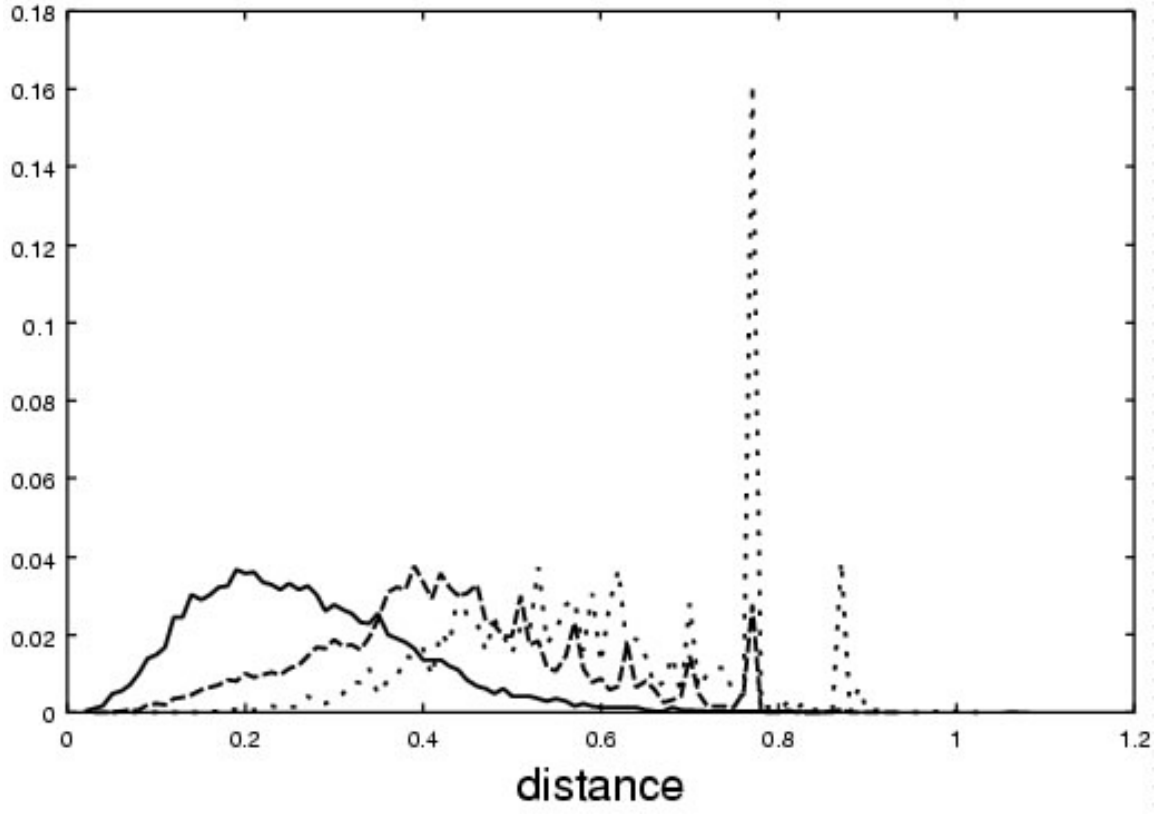

Fig 5.