

GENOMIC DNA SEQUENCE VISUALIZATION THROUGH PENTAHEDRON WALKING

Chia H. Yeh¹, Po Y. Sung², Hsuan T. Chang³ & Chung J. Kuo²

¹Department of Electrical-Systems Engineering, University of Southern
California, Los Angeles, CA 90089

²Graduate Institute of Communication Engineering, National Chung Cheng
University, Chiayi, 62107 Taiwan

³National Yunlin University of Science and Technology, Touliu Yunlin, 64002
Taiwan

E-mail : chyeh@sipi.usc.edu, pysung@samlab.ee.ccu.edu.tw,
htchang@yuntech.edu.tw, kuo@ee.ccu.edu.tw

Abstract

Genomic DNA sequences have been represented as long text data composed of alphabets A, C, T and G and these sequences present visualization challenges due to massive amount of discrete and multi-dimensional data. In this paper, we proposed a visual technique called VBP (Visualization by Pentahedrons) algorithm to visualize and analyze similarity of DNA sequences. Here, Markov chain model is employed to calculate the state transition probabilities of DNA sequences and map them into four different pentahedrons. The top pinnacle serves as the VBP walk origin, and the other four bottom points is the destination when the next read symbol is one of the alphabet A, C, T, and G. Therefore, a three-dimensional trajectory can be drawn for visualization. Since this is a first order Markov model, four pentahedrons are totally available for the query sequence. When the target sequence is very close to the query sequence, the VBP walk of the both sequences in the three-dimensional space are very close to each other. While the case when both sequences are far away from each other, the walk traces will be even more apart since four pentahedrons are based only on the query sequence. Simulation results are presented to further demonstrate the efficiency of our VBP algorithm in fast discriminating the difference of the global and local segments difference among DNA sequences.

Keywords : DNA, VBP, Markov chain model, pentahedron, visualization.

1. INTRODUCTION

With the fast development of the Human Genome Project (HGP), people work on important achievements to determine complete DNA sequences in the human genomes that contain long strings of nucleotide sequence (about 3,000,000,000 base pairs). When biologists get these data, the further work is to analyze what meaning (or equivalently the biological function) they stand for. Genomic DNA sequences are represented as long sequential strings of four types nucleotide that are basic unit to determine the hereditary information of biology. Each nucleotide is represented as one of four alphabets A, C, G, and T separately. Finally, the DNA sequences are stored as long text data.

Although DNA sequences can be simply represented as text data, it is still hard for human to analyze the meaning from these long strings text data. More specifically, it's almost impossible for people to compare or discriminate the similarity from these text data directly. Many previous works focus on mapping four alphabets into numeric [1][2][3]. Consequently, the similarity (or relatively difference) between DNA sequences can be analyzed and compared by digital signal processing techniques [4][5]. However, these existing techniques cannot easily visual DNA sequences to discriminate each other.

In this paper, we proposed a technique to visualize DNA sequences directly and discriminate the similarity among DNA sequences [6]. This visualization technique called VBP (Visualization by Pentahedrons) algorithm employs the Markov chain model to calculate the state transition probabilities of DNA sequences and map them into four different pentahedrons. Then, a trajectory in three-dimensional space can be drawn for visualization according to the relationship between top pinnacle and the other four bottom points of each pentahedron. Therefore, the similarity between DNA sequences can be fast discriminated by the VBP algorithm. In addition, the difference of global and local segments among DNA sequences also can be classified.

The rest parts of this paper are organized as follows. Section 2 presents the VBP algorithm how to visualize DNA sequences into the three-dimensional trajectory. Simulation results are presented to demonstrate the efficiency of our VBP algorithm in fast discriminating the difference among DNA sequences in Section 3. Finally, Section 4 briefly concludes this paper.

2. VISUALIZATION BY PENTAHEDRONS ALGORITHM

Proposed VBP algorithm concentrates on the biological sequences visualization by probabilistic model. Here, a first-order Markov chain model is used to describe the transition probability between every alphabet over all DNA sequences. Therefore, the complex information of DNA sequences can be integrated into general and conformable probabilistic model. Then, three-dimensional trajectory can be drawn for analysis. In the next section, we first discuss how to build the Markov chain model for DNA sequence in order to establish corresponding pentahedrons. Then, the three-dimensional trajectory walked by the pentahedrons will be explained.

2.1 Markov Chain Model for DNA Sequence

A DNA sequence composed of four alphabets A, C, G, and T and can be represented as

$$x_i = x_1, x_2, \dots, x_n \quad i = 1, \dots, n \quad x_i \in \{A, C, T, G\}, \quad (1)$$

where n is the length of DNA sequence. Figure 1 shows the state-transition diagram of four-state Markov DNA sequence. Four states correspond to a particular residue (also called nucleotides A, C, G, and T). Each arrow represents the probability of one state following the same state or another state. Therefore, 16 state transition probabilities are available which correspond to the probabilities of A following A, C, G, or T, C following A, C, G, or T, G following A, C, G, or T, and T following A, C, G, or T. The state transition probability a_{st} can be calculated and represented as following equations.

$$a_{st} = P(x_i = t \mid x_{i-1} = s). \quad (2)$$

where

$$s \in \{A, C, T, G\},$$

$$t \in \{A, C, T, G\}.$$

Markov Chain Model is employed to describe the state transition probabilities of DNA sequences. In the next section, we will show how to form four pentahedrons according to these state transition probabilities and its walking rule to transfer a DNA sequence into a three-dimensional trajectory.

2.2. Visualization by Pentahedrons (VBP) algorithm

In Section 2.1, we represent the relation of four alphabets in DNA sequences as the state transition probabilities. Then, we map them into four different pentahedrons shown in Fig. 3. The top pinnacle of these four pentahedrons represents the corresponding states A, C, G, and T, respectively. The other four bottom points are the destination when the next state is one of the alphabets A, C, G, and T. Therefore, the three-dimensional coordinate of top pinnacle and other four bottom points of each pentahedron can be calculated. Figure 2 shows how to calculate the three-dimensional coordinate of the top pinnacle to one bottom point. We define the three-dimensional coordinate (x,y,z) of the top pinnacle in each pentahedron is $(0,0,1)$. Then, the length $\overline{\alpha\beta}$ is known according to the state transition probability $a_{\alpha\beta}$, so the three-dimensional coordinate of points β and r can be obtained.

Because length $\overline{\alpha\beta}$ is equal to $\overline{\beta\gamma}$ due to the isosceles triangle, the coordinates on x -axis and y -axis of point r can be calculated by projecting $\overline{\beta\gamma}$ onto x -axis and y -axis. The other bottom points of pentahedron are calculated by the same method. Bottom points in x - y plane locate on different quadrants. A is located at quadrant-I, bottom points C, G, and T are located at quadrant-II, III, and IV, respectively. In Fig. 3, the angle between edges from top pinnacle to any two neighboring bottom points of each pentahedron itself is the same.

The vectors of top pinnacle to the other four bottom points can be calculated according to their three-dimensional coordinates. Therefore, we can map alphabets A, C, G, and T into three-dimensional vectors $v_1, v_2, v_3, \dots, v_n$. Eventually, a DNA sequence can be represented by the combination of these three-dimensional vectors shown in Equ. 3.

$$S = \{v_1, v_2, v_3, \dots, v_n\}, \quad (3)$$

where n is the length of DNA sequence and . The top pinnacle serves as the VBP walk origin and locates at the origin of the three-dimensional Cartesian coordinate. Consequently, the vector of first alphabet in DNA sequence is $v_1 = (0,0,0)$ and we can get a new vector sequence S' by accumulating these vectors shown in Equ. 4.

$$S' = \{w_1, w_2, w_3, \dots, w_n\}. \quad (4)$$

where

$$\begin{aligned}w_1 &= v_1, \\w_2 &= v_1 + v_2, \\&\vdots \\w_n &= v_1 + v_2 + v_3 + \cdots + v_n.\end{aligned}$$

Hence, a three-dimensional trajectory can be drawn to visualize according to vector sequence S' . Here, we represent the relation of four alphabets in DNA sequences as a Markov chain model and map them into four pentahedrons in order to draw the three-dimensional trajectory. Each DNA sequence has its four corresponding pentahedrons and the three-dimensional trajectory depend on these pentahedrons. Please note that if four pentahedrons of two DNA sequences are similar, these two sequences have similar statistic characteristics. Therefore, the three-dimensional trajectory will show clearer difference than that of using the unity geometric graph. In addition, four pentahedrons can be used as an indexing to preprocess DNA sequences' similarity. If sequences have similar pentahedrons, the detail difference can be shown in the three-dimensional trajectories in order to save the time to draw the trajectory on every sequence in database. In the next section, we will illustrate simulation results of our VBP algorithm.

3. EXPERIMENTS AND RESULTS

Two sets of DNA gene sequences DHFR (Di-Hydro-Folate Reductase) and TGFA (Transforming Growth Factor-Alpha) for human and rat are used for experiments. Homologous sequences for DHFR or TGFA exist high similarity in the biology point of view in different species (human or rat). Tables 1(a) and 1(b) show the results of state transition probabilities of human and rat DHFR gene sequences, respectively. According to these tables, two sets of four corresponding pentahedrons for human and rat for DHFR gene sequence can be obtained in Figs. 3(a) and 3(b). Therefore, the VBP walk of human and rat DHFR gene in the three-dimensional space can be drawn in Figs. 6(a) and 6(b). These two trajectories are very close to each other because they are homogenous gene sequence. Similar results for TGFA gene sequences are shown in Figs. 7(a) and 7(b). Their corresponding state transition probabilities are also shown in Tables 2(a) and 2(b); two sets of four pentahedrons are shown in Figs. 4(a) and 4(b), respectively. Experimental results of two inhomologous sequences, human proactin and trypsin gene sequences are shown in Figs. 8(a) and 8(b). The transition probabilities and sets of four pentahedrons of these two inhomologous sequences are

shown in Tables 3(a), 3(b), Figs. 5(a) and 5(b), respectively. Obviously, we can easily recognize the trajectory similarity between homogenous and inhomogeneous genes by the VBP algorithm.

4. CONCLUSION

In this paper, a new method called VBP algorithm is proposed to visualize DNA sequences. This method employed a Markov chain model to calculate state transition probabilities and map these probabilities into four pentahedrons in order to draw a three-dimensional trajectory visualize. Therefore, biologists can discriminate similarity or difference between DNA sequences according to these trajectories. This method can be applied in fast discrimination when the DNA sequences are numerous.

5. REFERENCES

- [1] E. A. Cheever, D. B. Searls, W. Karunaratne, and G. C. Overton, "Using signal processing techniques for DNA sequence comparison," *Bioengineering Conference, 1989. Proceedings of the 1989 Fifteenth Annual Northeast*, pp. 173-174, 1989.
- [2] D. Anastassiou, "DSP in genomics: processing and frequency -domain analysis of character strings," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 1053-1056, 2001.
- [3] P. Cristea, "Genetic signal analysis," *Sixth International Symposium on Signal Processing and its Applications*, Vol. 2, pp. 703-706, 2001.
- [4] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, Vol. 18, pp. 8-20, 2001.
- [5] W. Wong, and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Transaction on Signal processing*, Vol. 50, pp. 628-634, 2002.
- [6] E. H.-H Chi, P. Barry, E. Shoop, J. V. Carlis, E. Retzel, and J. Ried, "Visualization of biological sequence similarity search results," *IEEE Visualization 95*, pp. 44-51, 1995.

s \ t	A	C	G	T
A	0.2400	0.2286	0.3829	0.1486
C	0.2792	0.3250	0.1208	0.2750
G	0.2047	0.2953	0.2441	0.2559
T	0.0660	0.2386	0.4873	0.2081

(a)

s \ t	A	C	G	T
A	0.2330	0.2273	0.3920	0.1477
C	0.2833	0.3208	0.1126	0.2833
G	0.2046	0.2934	0.2510	0.2510
T	0.0700	0.2350	0.4850	0.2100

(b)

Table 1: State transition probabilities of (a) human and (b) rat DHFR genes

s \ t	A	C	G	T
A	0.3520	0.1620	0.2905	0.1955
C	0.3500	0.2917	0.1083	0.2500
G	0.3910	0.1504	0.2632	0.1954
T	0.1679	0.2748	0.2519	0.3054

(a)

s \ t	A	C	G	T
A	0.3516	0.1648	0.2912	0.1924
C	0.3516	0.3125	0.1016	0.2343
G	0.3806	0.1567	0.2612	0.2015
T	0.1667	0.2803	0.2500	0.3030

(b)

Table 2: State transition probabilities of (a) human and (b) rat TGFA genes

s \ t	A	C	G	T
A	0.3168	0.1832	0.2376	0.2624
C	0.2544	0.3333	0.1096	0.3026
G	0.2934	0.2994	0.2096	0.1976
T	0.1624	0.3299	0.2944	0.2132

(a)

s \ t	A	C	G	T
A	0.2434	0.2540	0.2857	0.2169
C	0.2931	0.3190	0.0733	0.3147
G	0.2525	0.2677	0.2576	0.2222
T	0.1333	0.3222	0.4222	0.1222

(b)

Table 3: State transition probabilities of (a) human proactin and (b) human trypsin genes

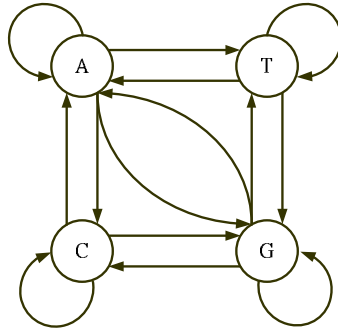


Figure 1: Markov chain model for DNA

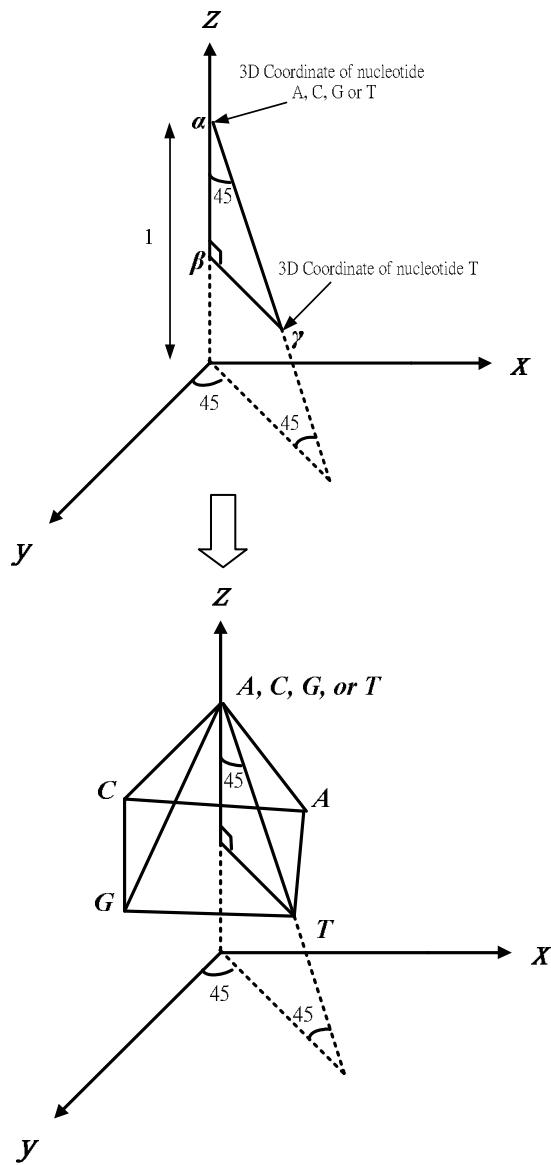


Figure 2: Relationship between top and bottom

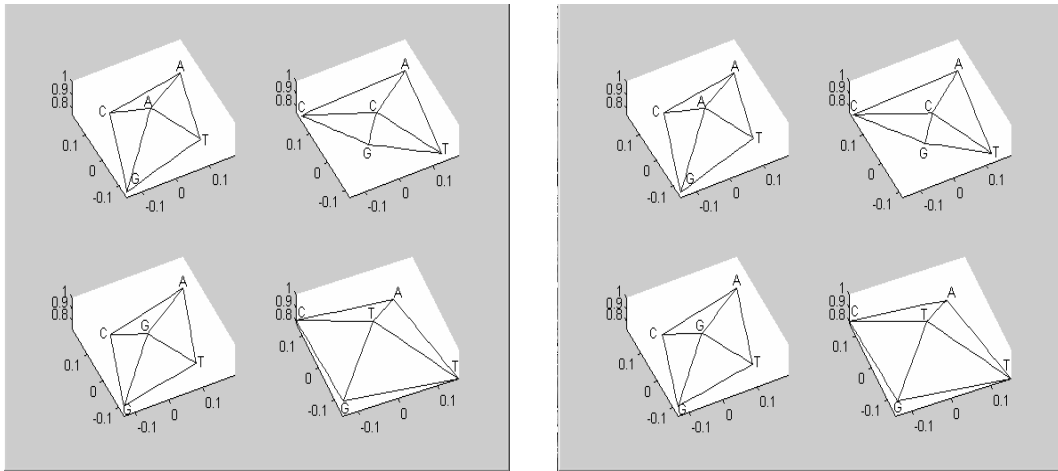


Figure 3. Four pentahedrons of (a) human DHFR gene (b) rat DHFR gene

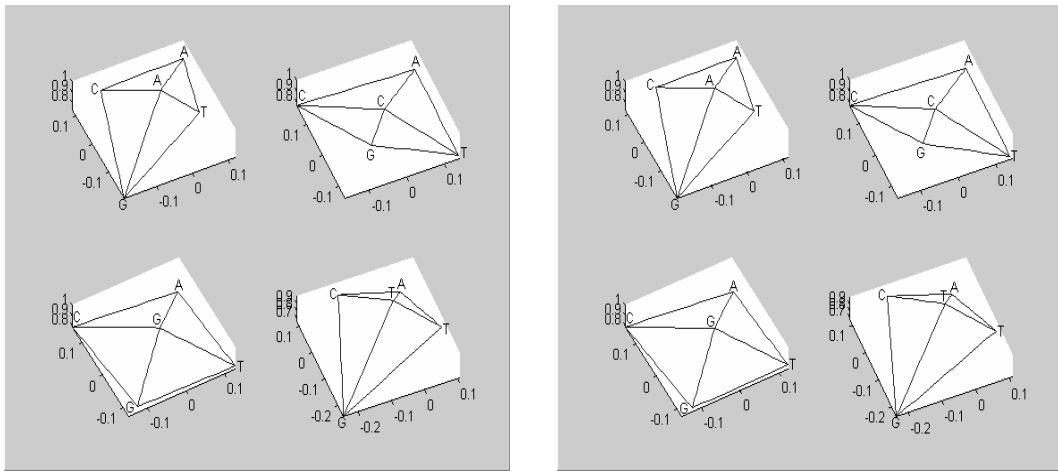


Figure 4. Four pentahedrons of (a) human TGFA gene (b) rat TGFA gene

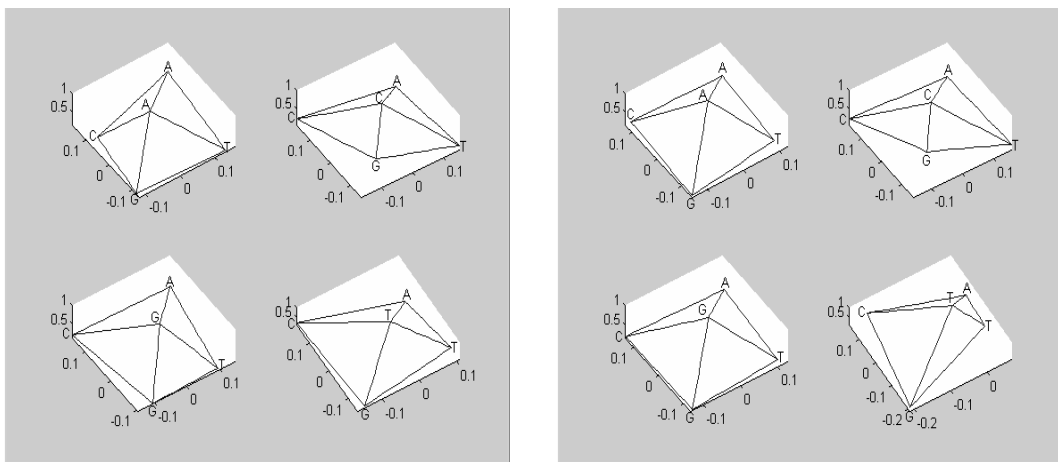


Figure 5. Four pentahedrons of (a) human proactin gene (b) human trypsin gene

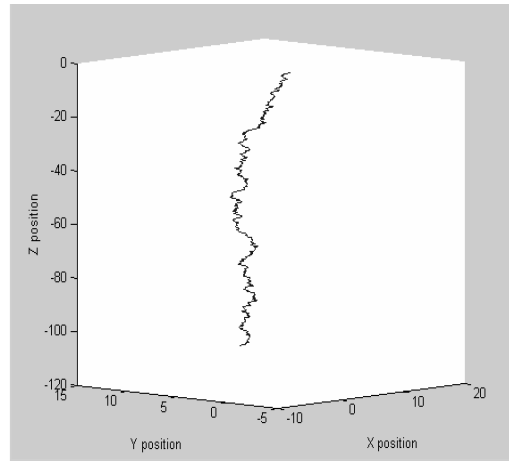
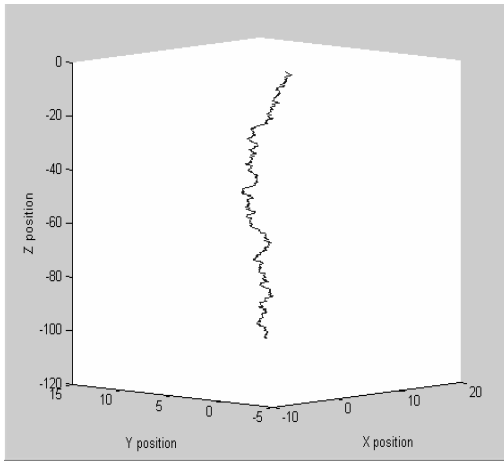


Figure 6. Three-dimensional trajectories of (a) human DHFR gene (b) rat DHFR gene

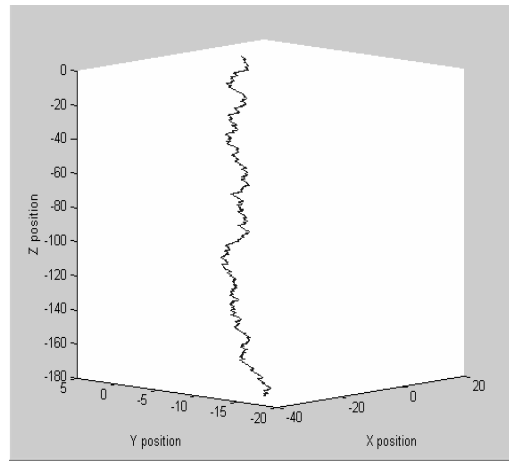
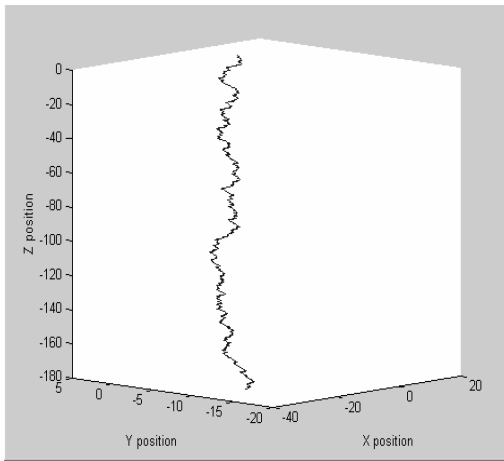


Figure 7. Three-dimensional trajectories of (a) human TGFA gene (b) rat TGFA gene

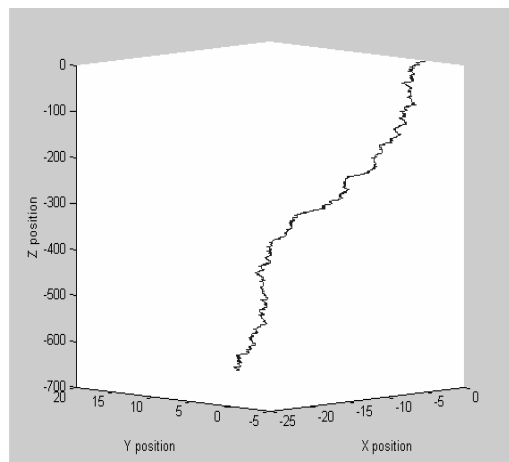
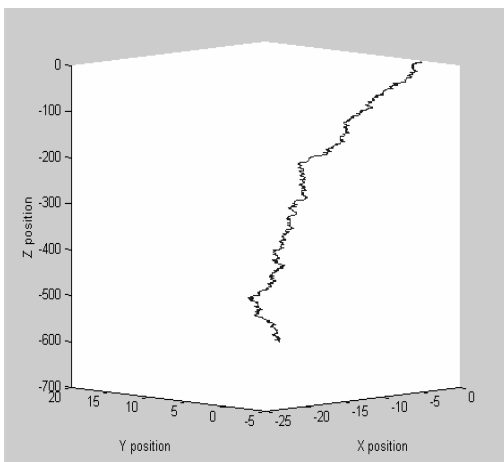


Figure 8. Three-dimensional trajectories of (a) human proactin gene (b) human trypsin gene