Genomic Sequence Analysis using Electron-Ion Interaction Potential

Masumi Kobayashi s1090085 Supervised by Hiroshi Toyoizumi

Abstract

This paper proposes two new methods in bioinfomatics for similarity comparison and gene finding. We evaluate similarity comparison and gene finding using Lindley equation and Electron-Ion Interaction Potential (EIIP)[1] by genome data. The first is the similarity comparison method of two sequences that shortens the processing time. The second is the gene finding method. In order to verify these methods, we use the Lindley equation of queuing theory and EIIP.

1 Introduction

Genome-sequencing projects of various creatures began at the end of the 1980s, including the Human Genome Project. The gene information is becoming clearer from these projects. As a result, practical value for medicine, pharmaceuticals, and agriculture have been found. The complete analysis of a human genome was released in the beginning of 2001. The calculation technique used required a computer for the sequence determination and analysis of the human genome. Now, although we face difficult problems of understanding the meaning of these sequences, the results of these research are very useful for humans. As genome research advances, bioinfomatics is becoming more and more important[2].

Sequence comparison is the most important primitive operation in bioinfomatics, serving as a basis for many other, more complex manipulations. Roughly speaking, sequence comparison consists of finding which parts of the sequences are alike and which parts differ. The most famous sequence comparison method is *dynamic programming*[3]. However, the problem with this method is that the calculation time is long. To address this problem, we use a sequence comparison technique with Lindley equation of queuing theory. By using this equation, calculation time can be shortened.

A DNA sequence consists of gene coding regions and meaningless regions without gene information, which are called junk region. Genes are certain contiguous regions of the chromosome, but they do not cover the entire DNA. Finding the gene coding regions is important operation, too. Toyoizumi and Tuchiya showed a technique to find gene coding regions by using Lindley equation[4]. But there is a problem. The determination of score required for Lindley equation is artificial. Therefore, in this paper, we decide the theoretical score by using EIIP, which describes the average energy states of all valance electrons in particular amino acids.

2 Important Basic Knowledge

2.1 Basic Knowledge in DNA

DNA has all the information for building a living body and it is expressed by the base sequence consisting of four nucleotide: Adenine(A), Thymine(T), Guanine(G), and Cytosine(C). This base sequence has a double helix structure, two sequences rolled mutually.

A DNA sequence consists of a row of four nucleotides, and each nucleotide triplet is called a codon. A codon corresponds to an amino acid. There are 64 possible nucleotide triplets, but there are only 20 amino acids. Therefore, two or more triplets correspond to one amino acid. For example, both TTT and TTC code for threonine. On the other hand, three codons, TAA, TAG, and TGA are called stop codon. They do not code for any amino acid and are used as signals at the end of a gene[5]. Table 1 shows the mapping of codons to amino acids.

A DNA sequence can be translated into an amino acid sequence. The sequence which changes the nucleotide triplet of a DNA sequence into a corresponding amino acid is called amino acid sequence.

2.2 Lindley Equation

In order to use Lindley equation, we need to describe the relation between the waiting time of the customer of queuing theory and a DNA sequence. A score is given for the similarity of the amino acid of two target gene sequences, and the sum of score is made to correspond to waiting time of queuing theory. In order to make it correspond with the waiting time of queuing system, W_n and S_k are defined as follows. W_n is the sum of the score to the n-th letter. S_k is amino acid score of the k-th letter.

$$W_n = max_{1 \le i \le n+1} \sum_{k=i}^{n} S_k, \tag{1}$$

where, the value is set to 0 if the range of $\sum$ is an empty set. W_n is the Loynes variable well known for queuing theory, and Lindley equation is known as well.

Table 1: Codon Table

TTT	F	TCT	S	TAT	Y	TGT	C
TTC		TCC		TAC		TGC	
TTA	L	TCA		TAA	STOP	TGA	STOP
TTG		TCG		TAG		TGG	W
CTT	L	CCT	P	CAT	H	CGT	R
CTC		CCC		CAC		CGC	
CTA		CCA		CAA	Q	CGA	
CTG		CCG		CAG		CGG	
ATT	I	ACT	T	AAT	N	AGT	S
ATC		ACC		AAC		AGC	
ATA		ACA		AAA	K	AGA	R
ATG	M	ACG		AAG		AGG	
GTT	V	GCT	A	GAT	D	GGT	G
GTC		GCC		GAC		GGC	
GTA		GCA		GAA	E	GGA	
GTG		GCG		GAG		GGG	

Theorem 1. *(Lindley Equation of Amino Acid Score) The amino acid group score W_n of a letter sequence satisfies the following equation.*

$$W_n = max\{W_{n-1} + S_n, 0\}. \tag{2}$$

However, $W_0 = 0$.

2.3 EIIP

We use the value of electron-ion interaction potential (EIIP) to decide each amino acid score and a stop codon score theoretically. Each amino acid is represented by the EIIP value, which describes the average energy states of all valance electrons in particular amino acids. The EIIP values for each amino acid were calculated using the following general model pseudo-potential. q is the change of momentum k of the delocalized electron in the interaction with potential w. Z_i is the number of valence electrons of the i-th component of each amino acid. N is the total number of i-th amino acid. Table 2 shows EIIP value of each amino acid.

$$< k + q|w|k >= \frac{0.25Z sin(\pi \times 1.04Z)}{2\pi}, \tag{3}$$

where

$$Z = \frac{\sum Z_i}{N}. \tag{4}$$

Table 2: Electron-Ion Interaction Potential(EIIP) values for Amino Acids.

Amino Acid	EIIP	Amino Acid	EIIP
Leu (L)	0	Tyr (Y)	0.0561
Ile (I)	0	Trp (W)	0.0548
Asn (N)	0.0036	Gln (Q)	0.0761
Gly (G)	0.0050	Met (M)	0.0823
Val (V)	0.0057	Ser (S)	0.0829
Glu (E)	0.0058	Cys (C)	0.0829
Pro (P)	0.0198	Thr (T)	0.0941
His (H)	0.0242	Phe (F)	0.0946
Lys (K)	0.0371	Arg (R)	0.0959
Ala (A)	0.0373	Asp (D)	0.1263

3 Similarity Comparison Experiment

First, we describe a technique of the sequence similarity comparison using Lindley equation and EIIP. The genome data that we use is a gene coding region of human α- and β-Hemoglobins[6].

3.1 Hemoglobin

Hemoglobin is contained in erythrocyte. Hemoglobin is the protein that consists of two kinds of subunits to which the structure α and β are similar. It consists of a "hem" containing iron, and a "globin" which is protein, and has the important role of carrying oxygen inside of the body[7].

Table 3: Sequences of Human α- and β-Hemoglobins.

A gene coding region of Human α-Hemoglobin	A gene coding region of Human β-Hemoglobin
VHLTPEEKSAVTALWG	VLSPADKTNVKAAWG
KVNVDEVGGEALGRL	KVGAHAGEYGAEALE
LVVYPWTQRFFESFGD	EMFLSFPTTKTYFPHFD
LSTPDAVMGNPKVKA	LSHGSAQVKGHGKKV
HGKKVLGAFSDGLAH	ADALTNAVAHVDDMP
LDNLKGTFATLSELHC	NALSALSDLHAHKLRV
DKLHVDPENFRLLGN	DPVNFKLLSHCLLVTL
VLVCVLAHHFGKEFT	AAHLPAEFTPAVHASL
KFLASVSTVLTSKYR	DKFLASVSTVLTSKYR

3.2 Amino Acid Scores and the Stop Codon Score by EIIP

We decide each amino acid score and the stop codon score by using the EIIP value. If all scores are positive values, W_n only increases, it is meaningless. We have to determine each amino acid score, which will become a negative or a positive score.

First, we decided each amino acid score by subtracting 0.111 from the EIIP, in order to make only the largest value of the EIIP positive and make all other values negative. Next, we decided each amino acid score by subtracting 0.0953 from the EIIP, in order to make two EIIP values from the largest EIIP value to the second positive and make all other values negative. We decided the amino acid score in this way, and when we decided each amino acid score by subtracting 0.0532 from the EIIP, in order to make nine EIIP values from the largest EIIP value to the ninth positive and make all other values negative, we obtained the best experiment result. And we determined a stop codon score by doubling the lowest amino acid score. Table 3 shows each score of amino acid and the stop codon.

Table 4: Amino Acid and the Stop Codon Score of α-Hemoglobin and β-Hemoglobin.

Amino Acid	Score 1 EIIP - 0.0532	Amino Acid	Score 1 EIIP - 0.0532
Leu (L)	-0.0532	Tyr (Y)	-0.0016
Ile (I)	-0.0532	Trp (W)	0.0016
Asn (N)	-0.0496	Gln (Q)	0.0229
Gly (G)	-0.0482	Met (M)	0.0291
Val (V)	-0.0475	Ser (S)	0.0297
Glu (E)	-0.0474	Cys (C)	0.0297
Pro (P)	-0.0334	Thr (T)	0.0409
His (H)	-0.0290	Phe (F)	0.0414
Lys (K)	-0.0161	Arg (R)	0.0427
Ala (A)	-0.0159	Asp (D)	0.0731
Stop Codon		-0.1064	

3.3 Experiment Result

The results of the calculated scores W_n of α-Hemoglobin and β-Hemoglobin using Lindley equation are shown in Figure 1 and Figure 2. We observe the domain where W_n is high, these results are very similar.

The difference (absolute value) between Figure 1 and Figure 2 is shown with the solid line in Figure 3 and the difference (absolute value) between α-Hemoglobin and β-Hemoglobin, which was rearranged at random,

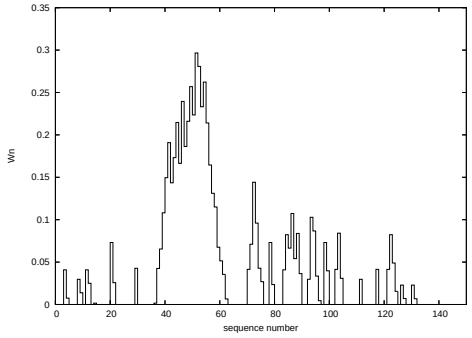

Figure 1: Sample path of α-Hemoglobin using Score 1.

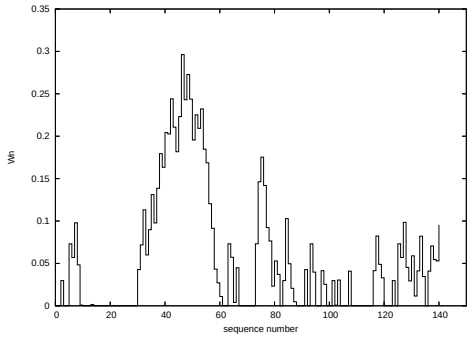

Figure 2: Sample path of β-Hemoglobin using Score 1.

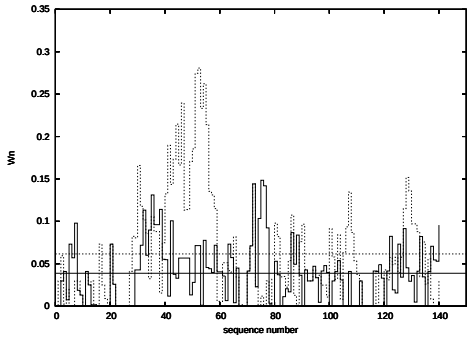

Figure 3: Sample difference of Figure 1 and Figure 2, and α-Hemoglobin and random sequence of β-Hemoglobin using Score 1. The solid line shows difference of Figure 1 and Figure 2. The dotted line shows the difference of α-Hemoglobin and random sequence of β-Hemoglobin. The average of difference of Figure 1 and Figure 2 is 0.038744(the solid line) and the average of α-Hemoglobin and the random sequence of β-Hemoglobin is 0.061567(the dotted line).

is shown with the dotted line in Figure 3. The average of difference between Figure 1 and Figure 2 is 0.038744 and the average between α-Hemoglobin and the random sequence of β-Hemoglobin is 0.061567. The difference between Figure 1 and Figure 2 is about half of α-Hemoglobin and the random sequence of β-Hemoglobin. The difference between Figure 1 and Figure 2 is very small. So, we could show the similarity of α-Hemoglobin and β-Hemoglobin. That is, we showed similarity comparison by using Lindley equation and EIIP.

4 Gene Finding Experiment

Next, we describe a technique for finding the gene coding regions using Lindley equation and EIIP. We use the Escherichia coil O157:H7 Sakai[8] genome data for our experiment.

4.1 Escherichia coil O157:H7 Sakai

Escherichia coil O157:H7 Sakai is a major food-born infection pathogen that causes diarrhea, coilitis, and hemolytic uremia syndrome.

4.2 Amino Acid Scores and the Stop Codon Score by EIIP

In this experiment, we use two kinds of scores. First, we decided the amino acid score like the section 3.2. And, when we decided each amino acid score by subtracting 0.0445 from the EIIP, in order to make ten EIIP values from the largest EIIP value to the tenth positive and make all other values negative, we obtained the best experiment result.

In Score 2, we decided each amino acid score by subtracting 0.0445 from the EIIP, and the stop codon score by doubling the lowest amino acid score. In Score 3, we do not change the amino acid score but change the score of the stop codon from two times to four times. Table 4 shows each score of amino acid and the stop codon.

4.3 Threshold

It is important whether W_n is judged as a gene when the sum of an amino acid score is above how much. Because W_n may become high by chance in the regions that is meaningless at an amino acid sequence.

First, we consider what kind of distribution the appearance probability of the character of an amino acid sequence becomes. The appearance probability of the character of an amino acid sequence assumes that independent and identical distribution is followed. This is equivalent to a junk region without biological information. In this case, the amino acid score W_n of a sequence can be considered to be GI/GI/1 queuing system. It is known that by using Chernoff Bound that the maximum and minimum (Kingman Bound) will be obtained[9].

Table 5: Each Amino Acid and the Stop Codon Score for Escherichia coil O157:H7 Sakai. Each Amino Acid Score of Score 2 and Score 3 is same and the Stop Codon Score is difference.

Amino Acid	Score 2, Score 3 EIIP - 0.0445	Amino Acid	Score 2, Score 3 EIIP - 0.0445
Leu (L)	-0.0445	Tyr (Y)	0.0071
Ile (I)	-0.0445	Trp (W)	0.0103
Asn (N)	-0.0409	Gln (Q)	0.0316
Gly (G)	-0.0395	Met (M)	0.0378
Val (V)	-0.0388	Ser (S)	0.0384
Glu (E)	-0.0387	Cys (C)	0.0384
Pro (P)	-0.0247	Thr (T)	0.0496
His (H)	-0.0203	Phe (F)	0.0501
Lys (K)	-0.0074	Arg (R)	0.0514
Ala (A)	-0.0072	Asp (D)	0.0818
Stop Codon	Score 2 -0.0890	Stop Codon	Score 3 -0.1780

Theorem 2. *(Kingman Bound) The waiting time of GI/GI/1 queuing system fills the following inequality.*

$$\gamma e^{-\eta x} \leq P[W_n > x] \leq e^{-\eta x} \tag{5}$$

Here, γ is a positive and below 1. η is called decay rate of W_n and can be obtained as follows.

$$\eta = sup\{s > 0 : E[e^{sS_n}] \leq 1\}. \tag{6}$$

Remark 1. It is known in bioinfomatics that the distribution of W_n can be similar to $e^{-\eta x}$[10].

Corollary 1. *For any $p \in (0,1)$, let $w_0 = \frac{-logp}{\eta}$, then*

$$P[W_n > w_0] \leq p. \tag{7}$$

If w_0 is estimated like Corollary 1, the probability that W_n of an amino acid score will exceed the threshold w_0 is below p. That is, the probability of judging the portion to be a gene accidentally in the case of a score with bigger W_n than w_0 can be assumed p.

4.4 Calculation Result of Threshold

We calculated the threshold of the amino acid sequence of Escherichia coil O157:H7 Sakai. $f(s) = E[e^{sS_n}]$ is calculated using the appearance probability of each amino acid. Table 5 shows the frequency of appearance of each amino acid and a stop codon. Figure 4 shows

Table 6: Appearance probability of Amino Acid and the Stop Codon.

Amino Acid	Appearance Probability	Amino Acid	Appearance Probability
Leu (L)	0.0822	Tyr (Y)	0.0255
Ile (I)	0.0507	Trp (W)	0.0184
Asn (N)	0.0358	Gln (Q)	0.0389
Gly (G)	0.0590	Met (M)	0.0167
Val (V)	0.0551	Ser (S)	0.0859
Glu (E)	0.0276	Cys (C)	0.0332
Pro (P)	0.0582	Thr (T)	0.0554
His (H)	0.0308	Phe (F)	0.0420
Lys (K)	0.0376	Arg (R)	0.0962
Ala (A)	0.0812	Asp (D)	0.0304
Stop Codon		0.0392	

$f(s)$. From this figure, we can see that *decay rate* η is 2.8. Therefore, we can regard $w_0 = 1.0699$ as a threshold of score similarity, from Corollary 1, and we can see $P[W_n > w_0] < 0.05$, so the probability which exceeds w_0 in junk regions is 0.05.

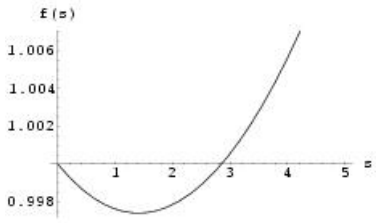

Figure 4: $f(s) = E[e^{sS_n}]$

4.5 Experiment Result

The result of the calculated score W_n of Escherichia coil O157:H7 Sakai is shown in Figure 5 and Figure 6 using the Lindley Equation and EIIP. These two figures show the calculation results of the same portion of the same sequence. We observe the domain where W_n is high. In Score 2, W_n becomes too high, so it is difficult to distinguish the region. On the other hand, in Score 3 which enlarged the value of stop codon, the domain where W_n becomes extremely high is seen. And, the threshold is shown with the dotted line in Figure 6. Table 6 shows the gene information[11] of Escherichia coil O157:H7 Sakai. We were able to find the gene in the portion exceeding threshold.

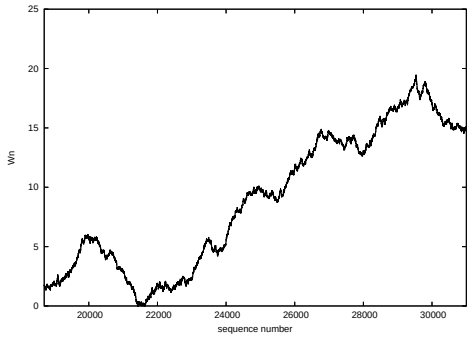

Figure 5: Sample path of Escherichia coil O157:H7 Sakai using Score 2.

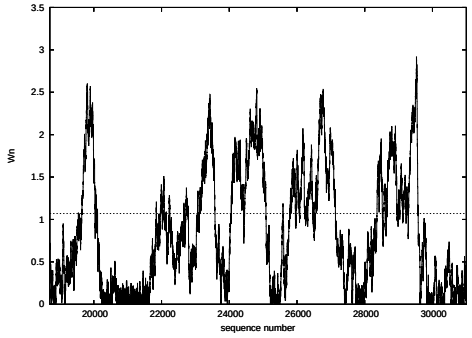

Figure 6: Sample path of Escherichia coil O157:H7 Sakai using Score 3 and the Threshold. The dotted line is Threshold = 1.0699.

Table 7: Gene information of Escherichia coil O157:H7 Sakai.

Gene	Length(bp)	Start	Stop	Product
ECs0058	1287	19464	19892	survival protein
ECs0063	2907	21763	22732	probable ATP-dependent RNA helicase
ECs0064	2352	22787	23570	DNA polymerase II
ECs0066	1503	23894	24394	L-arabinose isomerase
ECs0067	1701	24398	24965	L-ribulokinase
ECs0070	699	25683	25915	putative ATP-binding component of a transport system
ECs0071	1611	25910	26447	putative Transport system permease protein
ECs0072	984	26439	26766	transmin-binding protein
ECs0073	570	26890	27080	hypothetical protein
ECs0076	1401	28024	28490	3-isopropylmalate isomerase (dehydrate subunit)
ECs0077	1095	28491	18856	3-isopropylmalate dehydrogenase
ECs0078	1572	28855	29379	2-isopropylmalate synthase
ECs0079	87	29410	29438	sleu operon leader peptide

5 Conclusion

In this research, we showed two techniques for similarity comparison and gene finding.

First, in the similarity comparison method of two sequences, we can say that the similarity of the two sequences was shown by using Lindley equation and EIIP. So we showed a technique of sequence similarity comparison which shortened the processing time.

Next, in the gene finding method, we showed a technique for finding the gene coding regions from the DNA sequence that consisted of gene coding regions and junk regions by using Lindley equation and EIIP.

6 Acknowledgment

I would like to thank Prof. Hiroshi Toyoizumi for supervising this research and thorough advice and Prof. Nancy Sullivan for her help in improving my English writing.

References

[1] ”Protein sequence comparison based on the wavelet transform approach”, Protein Engineering, vol.15, no.3, pp.193-203, 2002, http://peds.oupjournals.org/cgi/content/full/15/3/193.

[2] J.C.Setubal and J.Meidanis, ”Introduction to Computational Molecular Biology”.

[3] J.C.Setubal and J.Meidanis, ”Introduction to Computational Molecular Biology”, pp.49-80.

[4] H.Toyoizumi and D.Tuchiya, http://www.u-aizu.ac.jp/ toyo/papers/bioinformatics/queue_in_genome.pdf

[5] J.C.Setubal and J.Meidanis, ”Introduction to Computational Molecular Biology”, pp.9-10.

[6] Protein sequence comparison based on the wavelet transform approach http://peds.oupjournals.org/cgi/content/full/15/3/193/F1

[7] PCN, ”Health click”, http://www2.health.ne.jp/word/d6021.html

[8] Genome Information Research Center, Osaka Univ. http://genome.gen-info.osaka-u.ac.jp/bacteria/

[9] Leonard Kleinrock, ”Queuing Systems”, Vol.2, John Wiley and Sons, 1976.

[10] S.Marlin and S.Altschul, ”Methods for assessing the statical significance of molecular sequence features by using general scoring schemes”, Proc.Natl Acad.Sci USA, Vol.87, pp.2264-2268, 1990.

[11] Whole Genome Viewer for E.coli O157:H7 Sakai http://genome.gen-info.osaka-u.ac.jp/cgi-bin/o157/list.pl?table=o157&page=0/