논문 2011-4-16

GenScan을 이용한 진핵생물의 서열 패턴 분석

Anlaysis of Eukaryotic Sequence Pattern using GenScan

정용규*, 임이슬**, 차병헌***

Yong-Gyu Jung, I-Suel Lim, Byung-Heun Cha

요 약 서열 상동성 분석은 생명현상에 관여하는 물질을 정렬, 색인하여 데이터베이스 하는 것으로, 생명정보학의 유용성을 입증하는 분야이다. 본 논문에서는 구조가 복잡한 진핵생물의 서열 패턴을 단백질 서열로 변환하기 위해 은닉마르코프모델을 이용하는 GenScan 프로그램을 이용한다. 서열상동성 분석 중 최소거리 탐색 문제는 문제의 크기가 커지면 계산량이 기하급수적으로 증가하여 정확한 계산이 불가능해진다. 따라서 유사한 아미노산간의 치환과 상이한 아미노산간의 치환 점수를 차등화한 점수표를 적용하고, 은닉마르코프모델 등을 적용해 정교한 전이 확률모델을 적용한다. 변환된 서열을 서열 상동성 분석을 위해 사용되는 blast p를 이용하여, 은닉 마르코프 모델을 도입함으로 인해 단백질 구조 서열로 변환하는 데에 있어서 우수한 기능을 제공함을 알 수 있다.

Abstract Sequence homology analysis in the substances in the phenomenon of life is to create database by sorting and indexing and to demonstrate the usefulness of informatics. In this paper, Markov models are used in GenScan program to convert the pattern of complex eukaryotic protein sequences. It becomes impossible to navigate the minimum distance, complexity increases exponentially as the exact calculation. It is used scorecard in amino acid substitutions between similar amino acid substitutions to have a differential effect score, and is applied the Markov models sophisticated concealment of the transition probability model. As providing superior method to translate sequences homologous sequences in analysis using blast p, Markov models. is secreted protein structure of sequence translations.

Key Words : eukaryotic, Hidden Markov Model, GenScan, blast p

Ⅰ. Introduction

Biological phenomena involved in the sorting and indexing all the materials in the database, the field of bioinformatics, the sequence phase demonstrated that sequence homology are analyzed usefulness. [1] - [2]

Protein key ingredient of life as a plain text string of one-dimensional structure of ownership that the first hypothesis was proposed in 1883 by Curtius. Since then many scientists further mutations and evolutionary process, both historical and more sequence information is encoded. Discovery of life mysteries is the remarkable development of modern molecular biology, biological phenomena are surprisingly 'chemical information', that it was. [3]

In the minimum distance problem, the searching space is increased exponentially by increasing complexity, the exact calculation becomes impossible. Bi-phase sequence analysis Thus, some hermeneutical methods applied to find efficiently an approximation is

*종신회원, 을지대학교 의료IT마케팅학과
**정회원, 을지대학교 의료산업학부 의료전산학전공
***정회원, 을지대학교 임상병리학과, 교신저자
접수일자 2011.6.25, 수정일자 2011.7.29
게재확정일자 2011.8.12

used. Between similar amino acid substitutions and different substitution score between amino acids have differential effect scorecard, or Markov models, and apply the sophisticated concealment or transfer (insertion, deletion, substitution) probability models are applied to more advanced algorithms. [4]

Hidden Markov models with complex relationships, such as intron and exon ongoing life issues in the case of more complex modeling is working properly.

II. Related research

1. prokaryotes and eukaryotes

Depending on the internal structure of cells can be divided into prokaryotes and eukaryotes. Prokaryotes are simple in structure. Which is found only in prokaryotic single-celled or colonial. Three stations of the way biological classification system (archaea, bacteria and eukaryotes classification) approach is the high taxes that bacteria and germs.

The cell membrane of eukaryotic organelles have their own will. Some fungi and unicellular forms, such as Ahmedabad, from bars, plants and animals, such as brown colonies and multicellular forms is still present.

The structure of prokaryotic and eukaryotic cells are shown in Fig. 1. All prokaryotic cells, eukaryotic cells have membrane. The cell membrane separates the internal and external support, and internal and external mass flow control, and also maintains the cell potential. Inner membrane of a salt, accounting for almost all of the cytoplasm. DNA in all cells with the gene and enzyme gene expression and protein and RNA with the information necessary to have. Chain of prokaryotic life in general, compared with its own eukaryotic cells typically are found in multicellular organisms.

Prokaryotic and eukaryotic cells in general larger than 10 times, 1000 times larger sense, by volume. The biggest difference between prokaryotic and eukaryotic cell lines specific to the case of eukaryotic cells are part of the cell membrane is surrounded by [6].

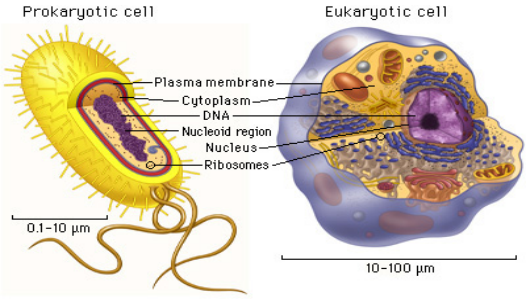


그림 1. 진핵세포와 원핵세포
Fig. 1. Eukaryotic and prokaryotic

2. Hidden Markov Model

Secreted by each state Markov model with transition probability is a finite state machine, each state can also not be observed directly, but instead, each state has a chance to produce their state can be observed. Hidden Markov Model is shown in Table 1 below sets and two state model has a set of three is likely. [5]

표 1. 은닉 마르코프 모델의 집합
Table 1. A set of Hidden Markov Models

observable state	
hidden state	Hidden Markov process is represented by the state of the system
observable state	The visible state of the process
Set of probability	
Π vector	Represents the initial probability of hidden state vector state transition
state transition matrix	Transition probabilities between hidden states, the probability of the transition to a previous state
confusion matrix	Concealment can be observed under certain specific conditions appear likely

III. GenScan

After the first genetic material was separated to determine the sequence of the gene function was seeking to identify the classical paradigm of molecular biology, as opposed genome sequence to decide first

place overall in the mechanical, chemical analysis of the sequence structure information through the genes, regulatory region of genes, RNA, and bioinformatics finding was transformed into a paradigm.

In case of gene finding, the first 6 to decrypt the frame long enough to read the so-called translation initiation frame (Open Reading Frame, ORF), Finding the most popular towing gene begins with a simple algorithm to find. Of course, the case of eukaryotes, complex relationships, such as introns, and exon the problem is more complex modeling.

The important thing to find genes of major issues such as the genomic sequence of the pattern analysis of the structure problem, namely, the problem is that the traditional informatics. Actually GenScan using the Hidden Markov Model is a pattern analysis algorithm [5]. Common local pattern analysis and mathematical analysis of global patterns (Ab Initio) are performed in sequence informatics through virtual sex search sites deduced sequence of the gene may be an inference is probabilistic. In other words, GenScan is a program to turn from the DNA sequence into an amino acid sequence. that turns. Splicing introns are removed, and the promoter alone. Exxon poly (A)-signal and can be found to predict the ORF. Fig. 2 above provided by the MIT GenScan is a web server [7].

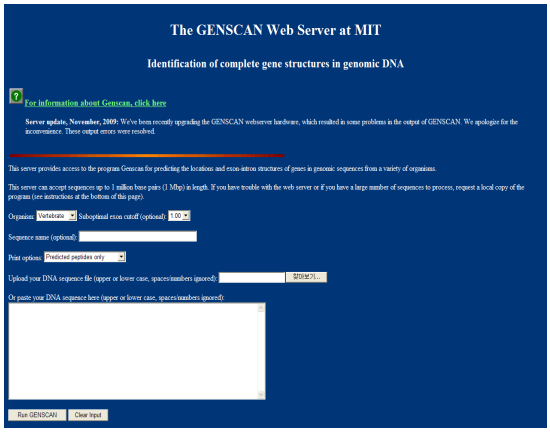


그림 2. The GenScan Web Server at MIT
Fig. 2. The GenScan Web Server at MIT

Ⅳ. Experiments and experimental results

1. Experiments

The algorithm used in this paper GenScan Eukaryotes or Eucaryote Gene Prediction program for the Exxon By Gene are given Intron structure and the predicted location.

NCBI (National Center for Biotechnology Information) is provided by GenBank search of the nucleotide sequence in progress in the state were randomly selected sequences. [8] In this paper, we use a randomly selected sequences CU326341. Fig. 4 as part of this sequence information, sequences, are registered as non-aligned status, and having the sequence of 1 to 178,302 indicates species of Sus scrofa pig. 1-9301 of the randomly selected sequences from sequence to sequence and copy 9412-25141 GenScan to convert the amino acid sequence respectively. Each two sets of experiments, using the blast p program of NCBI will detect whether homology [8].

```
LOCUS       CU326341            178302 bp    DNA     linear   HTG 28-JUL-2009
DEFINITION  Sus scrofa chromosome 11 clone CH242-430A12, WORKING DRAFT
            SEQUENCE, 2 unordered pieces.
ACCESSION   CU326341
VERSION     CU326341.6  GI:254826599
KEYWORDS    HTG; HTGS_PHASE1; HTGS_DRAFT; HTGS_FULLTOP.
SOURCE      Sus scrofa (pig)
  ORGANISM  Sus scrofa
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla; Suina; Suidae;
            Sus.
```

그림 3. CU216341 정보의 일부
Fig. 3. CU216341 Some of the information

2. Experimental results

Randomly dividing Set 1 of a sequence from 1-9301 and Set 2 of sequence from 9412 to 25,141 in protein sequences using the GenScan blast p after switching to the next one on the diagramed put the results shown in Fig. 4.

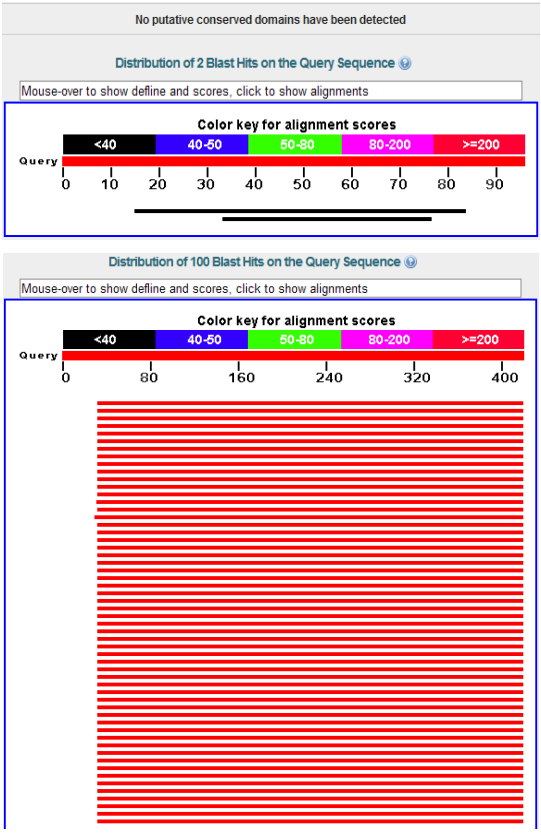
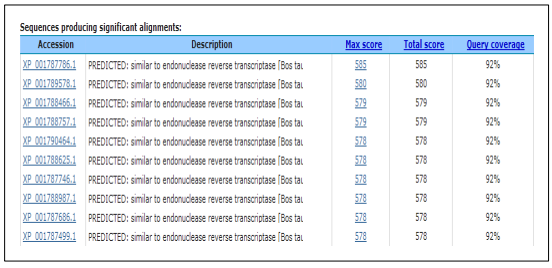


그림 4. 집합 1과 집합 2의 blast p 결과
Fig. 4. blast p results of Set 1 and Set 2

Show homology to the coordinates (a size of peptide) in the top of Fig. 4 (2) are displayed in a black color of the graph the average molecular weight of the protein (a size of peptide) of less than 40 highly homologous to find that no. If the bottom of Fig. 4 (100), all of the colors of the graph is marked in red the average molecular weight of the protein with 200 or more homologous sequences found that the city estimates. Using blast p 1 and Set 2 set of experiments, each the result of homology search through the set, a case of Fig. 5, both of the extracted data, the average molecular weight of the protein peptide 40dalton below query to get the high homologous will not find, except in the case of two sets of data extracted all one hundred or more peptide 200 dalton average molecular weight of proteins to get higher homology can see that the query found.

A set of two blast p search results list of the top 10 dogs are part of the query list is shown as Fig. 5. Among these were the highest homology score XP_001787786.1 the "PREDICTED: similar to endonuclease reverse transcriptase [Bos taurus]" is defined. Converting it into protein sequences before the original DNA sequence data, part of CU216341 XP001787786.1 XP001787786.1 the most similar protein sequence data show the inference.



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage
XP_001787786.1	PREDICTED: similar to endonuclease reverse transcriptase [Bos tau	585	585	92%
XP_001789578.1	PREDICTED: similar to endonuclease reverse transcriptase [Bos tau	580	580	92%
XP_001788466.1	PREDICTED: similar to endonuclease reverse transcriptase [Bos tau	579	579	92%
XP_001788757.1	PREDICTED: similar to endonuclease reverse transcriptase [Bos tau	579	579	92%
XP_001790464.1	PREDICTED: similar to endonuclease reverse transcriptase [Bos tau	578	578	92%
XP_001788625.1	PREDICTED: similar to endonuclease reverse transcriptase [Bos tau	578	578	92%
XP_001787746.1	PREDICTED: similar to endonuclease reverse transcriptase [Bos tau	578	578	92%
XP_001788987.1	PREDICTED: similar to endonuclease reverse transcriptase [Bos tau	578	578	92%
XP_001787686.1	PREDICTED: similar to endonuclease reverse transcriptase [Bos tau	578	578	92%
XP_001787499.1	PREDICTED: similar to endonuclease reverse transcriptase [Bos tau	578	578	92%

그림 5. 집합 2의 blast p 결과의 상위 10개 query
Fig. 5. a set of two of the top 10 query results blast p

V. Results

Using blast p experimental data and with the most similar protein sequences, protein sequences and XP001787786.1 to that used in the experiment of the protein sequence data for the validation of experimental results by comparing the lists look like the following shown in Fig. 6. Query sequence and a subject sequence, Compositional matrix adjust seqencing way, the above will appear similar arrangement referring to Fig. 6, Query sequence 34-419 and the subject of a 886-1272 check point will be similar.

Based on these results, numerical statistics, export data, 387 of the Identities of the 278 matches the import value has a value of approximately 71.8%, Positives value + value of the match, including the import value of 324 has 83.7% . - Gaps around because there are a value of 1 has a value of 0.2%.

```
Query  34    VWY-NTHRQTDQWNRIENPEINPDTYGQLIFDKGGKNIKWEKESLFSKHCWETCTAACKA  92
             VWY +  R  DQWN+IE+PEINP TYG LIFDKGGKNI+W K++LF+K CWE  +  CK
Sbjct  886   VWYWHKDRNIDQWNKIESPEINPRTYGHLIFDKGGKNIQWIKDNLFNKWCWEIWSTTCKR  945

Query  93    MKLEHTLTPCTKINSKWLKDLNIRQDTIKLLEENIGKTLSDINIMNIPSGQSPKAIEIRA  152
             MKLEH LTP TKINSKW+KDLN+R +TIKLLEENIGKTLSDI+   I     P+ +EI+A
Sbjct  946   MKLEHFLTPYTKINSKWIKDLNVRPETIKLLEENIGKTLSDIHHSRILYDPPPRILEIKA  1005

Query  153   KINFWDLIKLKSFCTAKETKKKTKRQLTEWEKIVSNDAMDKSLISRIYKQLIQLNSKKAN  212
             KIN WDLIKLKSFCT+KET  K KRQ +EWEKI++N+A DK LIS+IYKQL+QLNS+K N
Sbjct  1006  KINKWDLIKLKSFCTSKETISKVKRQPSEWEKIIANEATDKQLISKIYKQLLQLNSRKIN  1065

Query  213   QSMEKWAKDLNRHFSKEDTQMASKHMKKCSTSLIIREMQIKTTMRYHLTPVRMAIINKST  272
               ++KWAKDLNRHFSKED QMA+KHMK+CSTSLIIREMQIKTTMRYH TPVRMA+I KST
Sbjct  1066  DPIKKWAKDLNRHFSKEDIQMANKHMKRCSTSLIIREMQIKTTMRYHFTPVRMAVIQKST  1125

Query  273   NGKCWRGCGEKGTLLHCWWEWKLVQPLWRTVWRYFRNLYIDLPYDPAIPLLGIYPDKTLL  332
             N KCWRGCGEKGTLLHCWWE KLVQPLWRTVWR+ + L I+LPYDPAIPLLGI+ ++T
Sbjct  1126  NNKCWRGCGEKGTLLHCWWECKLVQPLWRTVWRFLKKLEIELPYDPAIPLLGIHTEETRR  1185

Query  333   KRDTCTRMFIAALFTIARTWKQPKCPSTDDWIQKRWYIYTMEYYSAIKKDDIMSFAATWM  392
             +RDTCT MFIAALF IARTWKQP+CPS D+WI+K+WYIYTMEYYSAIKKD   S    WM
Sbjct  1186  ERDTCTPMFIAALFIIARTWKQPRCPSADEWIRKQWYIYTMEYYSAIKKDTFESVLMRWM  1245

Query  393   ELENLILSEMSQKDKDKYHMISLITGI  419
             +LE +I SE+SQKDK +Y +++ I GI
Sbjct  1246  KLEPIIQSEVSQKDKHQYSILTHIYGI  1272
```

그림 6. 실험 고찰을 위한 단백질 서열 비교
Fig. 6. Protein sequence comparison for the experimental study

E value expectation of the same size and configuration to match a randomly generated string from the database to measure the potential to occur. E value closer to 0 it becomes less and are more likely to occur by chance. The lower the value that matches E yirwojin is that the better.

E value of less than 1 can be a definite hit. Fig. 6 as a result of this paper, we present the case for the E value 3e-165 (e ≒ 2.71828), so a kind of paper is less than the data used in judging the value of E (Sus crofa: pig) and similar proteins inferred data (Bos taurus: cattle) between the cross can be seen that a high homology.

Ⅵ. Conclusion

CU216341 used in this paper is data in the NCBI GenBank nucleotide sequence data provided by the vast progress of the eukaryotes within the discretion of the vertebrates, is one of the chosen pig process. The complex structures in order to analyze the structure of eukaryotic sequences that are based on concealment Markov model using the MIT GenScan converted to protein sequences was trying to express the probability of NCBI using blast p homology and similar kinds of tests came up with. In this experiment, the structure in complex eukaryotic sequences in the structural analysis of the program algorithm GenScan secreted protein structure due to the introduction of Markov models in sequence translations to provide a superior could see that.

REFERENCES

[1] Hughes TR, Marton MJ, Jones AR, et al. "Functional discovery via a compendium of expression profiles." Cell, 2000

[2] Ideker T, Thorsson V, Ranish JA, et al. "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.", Science, 2001

[3] Curtius T, "Ueber das Glycocoll Chem Ber", 1883

[4] J. Choy and S. B. Cho, “An intrusion detection system with temporal event modeling based on hidden Markov model," Proc. Korea Information Science Society (B), Seoul, pp 306-308, October 1999.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. of the IEEE, vol. 77, no. 2, pp. 257-286, February 1989.

[6] Karlin S and Ladunga I, "Comparison of eukauyotic genomic sequences", Proc. Natl. Acad. Sci, 1994

[7] GenScan, http://genes.mit.edu/GENSCAN.html

[8] NCBI, http://www.ncbi.nlm.nih.gov

[9] NCBI BLAST, http://blast.ncbi.nlm.nih.gov.

[10] Yong-Gyu Jung, Jeong-Seok Kang, Hospital Security System using Biometric Technology, IWIT Vol.11 No.2, 2011

[11] Yong-Gyu Jung, Go-Eun Hur, Ensemble Classification Method for Efficient Medical Diagnostic, IWIT Vol.10 No. 3, p97-102, 2010

[12] DaeSik Ko, JaeCheol Lee , A design of the DNA Scan System using Rotational Axis, Journal of

Korean Institute of. Information Technology, Vol.7 No.1, 2009

[13] HoSeok Chae, JeongAh Kim, MinHee Choi, SungYoung Oh, MinHo Lee and ChiWoo Lee, Reference Model for U-Health Portal System Based on Clinical Decision Supporting Service, Journal of Korean Institute of. Information Technology, Vol.9 No.7, JUL 2011

저자 소개

정 용 규(종신회원)

- 1981년 서울대학교 (이학사)
- 1994년 연세대학교 (공학석사)
- 2003년 경기대학교 (이학박사)
- 1999년~현재 을지대학교 교수
- 2001년~현재 ISO/TC154K위원장

<주관심분야: 임상데이터마이닝, 의료정보시스템, 전자거래표준>

임 이 슬(정회원)

- 2007년~현재 을지대학교 의료산업학부 의료전산학전공

<주관심분야: 의료정보시스템, 데이터마이닝>

차 병 헌(정회원)

- 1981년 조선대학교 (의학사)
- 1991년 한양대학교 (의학석사)
- 1999년 한양대학교 (의학박사)
- 2005년~현재 을지대학교 교수

<주관심분야: 유전학 전공/ 유전자 분석, 노화 기전>