프로모터 염기서열 분석을 위한 데이터 마이닝 기법

김 정 자*, 이 도 헌*

Data Mining Techniques for Analyzing Promoter Sequences

Jung-Ja kim, Doheon Lee

요 약

최근 지놈(Genome) 프로젝트를 통해 DNA 염기서열에 대한 정보가 밝혀짐에 따라 분자 수준의 유전자 정보를 다루는 기법이 활발히 연구되고 있다. 그리고 밝혀진 서열정보들의 방대함으로 미루어볼 때 이들 정보를 데이터베이스화하고 효과적인 분석을 행하기 위한 새로운 컴퓨터의 알고리즘의 개발 또한 시급한 일이다. 이러한 측면에서 본 논문에서는 분자생물학에서 매우 중요한 연구 대상으로 삼고있는 프로모터 서열과 유전자 간의 연관성으로 발현되는 특징을 알아내기 위한 연관 규칙 탐사 알고리즘을 연구한다. 기존의 탐사 알고리즘은 트랜잭션 데이터를 대상으로 하지만 본 논문에서는 생물학적 데이터를 대상으로 하였기 때문에 데이터의 형태와 생물학적인 특성을 수용하는 변형된 연관규칙 알고리즘을 설계한다. 본 연구를 통하여 얻어진 결과는 실제 생물학적 실험대상의 후보조합을 최소화 하므로써 많은 시간과 노력 비용을 절감할 수 있다.

Abstract

As DNA sequences have been known through the Genome project the techniques for dealing with molecule-level gene information are being made researches briskly. It is also urgent to develop new computer algorithms for making databases and analyzing it efficiently considering the vastness of the information for known sequences. In this respect, this paper studies the association rule search algorithms for finding out the characteristics shown by means of the association between promoter sequences and genes, which is one of the important research areas in molecular biology. This paper treat biological data, while previous search algorithms used transaction data. So, we design a transformed association rule algorithm that covers data types and biological properties. These research results will contribute to reducing the time and the cost for biological experiments by minimizing their candidates.

* 전남대학교

Ⅰ. 서 론

바이오인포메틱스(Bioinformatics)란 컴퓨터를 이용하여 생물학 연구에 관련된 제반 정보를 저장, 검색, 분석, 가공하는 연구 분야를 통칭한다. 특히, 최근 지놈(Genome) 프로젝트를 통해 DNA (Deoxyribonucleic acid) 염기 서열에 대한 정보가 밝혀짐에 따라, 분자 수준의 유전자 정보를 다루는 기법이 활발히 연구되고 있다. 발견된 유전자 정보는 주제별로 다수의 생물학 데이터베이스에 저장되어 웹을 통하여 제공되고 있으며 현재 여러 전산 기술을 이용하여 데이터베이스 통합 검색을 위한 시도가 이루어지고 있다[1]. 또한 대량의 지놈 데이터들이 산출됨에 따라 기존의 저장 및 분석 방식으로는 대량의 유전자 서열 정보 및 새로운 형태의 생물학 자료(서열, 이미지)의 처리가 불가능하기 때문에 새로운 분석 도구의 개발을 요구하고 있다.

데이터 마이닝은 방대한 양의 데이터로부터 흥미 있는 패턴이나 특성을 발견하는 기법이다. 연구 정보의 효율적인 저장과 검색을 위해 데이터 웨어하우스와 같은 데이터 저장소의 구축은 필수적이며 서열의 특성 및 진화 적 관계를 파악하기 위해 데이터 마이닝을 통한 새로운 생물학 적 지식들을 발견 할 수 있어야 할 것이다[2][3].

데이터 마이닝의 탐사기법중의 하나인 연관규칙 탐사 기법은 특정 사건이 발생하면 동시에 혹은 일정시간간격 사이에 다른 사건이 일어나는 관련성을 탐사하는 것을 의미한다. 연관규칙 이란 동시에 발생하는 사건들을 규칙의 형태로 표현 한 것으로 전제부에 해당하는 조건항목이 결론부에 해당하는 항목들을 야기한다고 정의한다. 이를 유전자 연구에 적용한다면 특정 종에서 특정 유전자의 염기 서열이 반드시 다른 유전자의 발현을 초래한다는 사실로 대변되어질 수 있다. 분석을 통하여 이들 유전자들로 인하여 나타나는 여러 발현 현상 등을 미리 알 수 있다면 다방면의 연구 분야에 유용한 정보를 제공 할 수 있을 것이다. 본 논문에서는 프로모터 염기 서열 분석을 통해 유전자간의 발현 특성에 대한 연관성을 탐사하고자 한다. 마이닝에서 연구되어진 연관규칙 탐사 기법을 수정. 보완하고, 이를 기반으로 프로모터의 염기 서열을 분석하여 유전자 발현특성을 예측하고 유전자 관련성 분석에 관한 연구를 수행한다. 발견 정보의 분석을 토대로 새로운 정보를 유추함으로써 실질적인 실험에 소요되는 많은 시간, 노력, 비용을 절감할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 본 연구에 필요한 생물학적 배경지식과 연관규칙 탐사기법을 논의한다. 3장에서는 프로모터 염기 서열 분석과정을 논의하며, 결론에서 향후 연구되어야 할 문제점을 논의한다.

Ⅱ. 관련연구

1. 단백질 합성

생물의 형질은 염색체내의 DNA 염기 서열에 따라 결정된다. DNA는 A(Adenin), T(Timen), G (Guanin), C(Cytosin) 의 염기로 구성되는데, 이 염기들이 어떤 순서로 배열되었느냐에 따라 다양한 형질이 발현(Expression)된다. 그림 1은 세포내의 유전 정보의 흐름을 도식하고 있다.

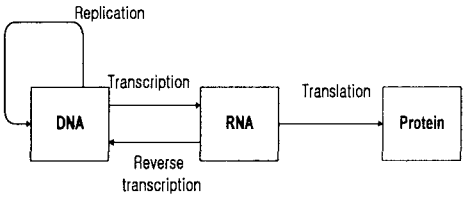

그림 1 세포내의 유전정보의 흐름

DNA는 유전과 관련된 정보를 지니고 있으며 자기 복제의 성질에 의하여 하나의 DNA를 증식하고 자신의 유전정보를 전달하는 전사 (Transcription) 과정을 거쳐 RNA를 합성한다. 합성된 RNA의 염기 배열 은 차례로 3개씩 읽혀지는 번역(Translation) 과정을 거쳐 리보소옴(ribosome)내에서 해당 암호에 따르는 아미노산을 만들며 이 아미노산의 체인이 결국 단백질을 구성하는 일련의 합성단계를 거친다[4].

2. 프로모터

사람의 유전체는 약 10만개의 유전자로 구성되어 있다고 한다. 특정 사람의 표현형은 이러한 모든 유전자의 발현(Expression)에 의하여 결정되며 각각의 기관 및 조직에서 나타나는 기능들은 이들에게서 특이적으로 발현되는 유전자의 세트에 의해 조절된다. DNA 에는 유전자에 해당하는 염기서열과 프로모터에 해당하는 염기서열이 존재한다. 프로모터 영역은 DNA 내에 RNA로의 전사를 조절하는 유전자 앞부분으로 실제 RNA의 합성에 직접 작용하는 조절영역이다. 프로모터의 특정 부위의 염기 서열의 조합이 특정 유전자를 결정짓고 그 유전자는 어떤 특성을 발현할 것이다. 결국 유사한 발현특성을 가진 유전자들의 프로모터에 존재하는 공통된 서열부분을 찾을 수 있다면 프로모터의 동작을 이해하는 중요한 단서를 얻을 수 있다. 현재 유전자에 관한 정보는 상당부분 밝혀진 상태이지만 유전자의 전사를 조절하는 프로모터에 대한 정보는 잘 알려져 있지 않은 상황이다. 주로 경험적 지식에 기반한 실험을 통해서 서열 위치를 찾고 있지만, 고려할 방대한 서열조합으로 인하여 많은 시간과 노력, 비용이 소요되고 있다[4].

3. 연관규칙 탐사

연관규칙 이란 동시에 발생하는 사건들을 규칙의 형태로 표현 한 것으로 특정 사건이 발생하면 동시에 혹은 일정한 시간 간격 사이에 다른 사건이 일어나는 관련성을 의미한다. 전제부에 해당하는 조건항목이 결론부에 해당하는 항목들을 야기한다고 정의하며 조건부를 X 결론부를 Y라할 때 "X->Y"로 표현된다. 가장 널리 알려진 Apriori 연관규칙 탐사 알고리즘은 다음과 같다. 사건 항목들의 전체 집합에서 그 부분 집합들을 트랜잭션으로 정의하여 각 단계마다 임계치 이상의 후보 아이템들만을 선정해 나가면서 최종 결정된 결과 항목의 집합으로 연관 규칙을 추출한다. 이때 적용되는 임계치는 전체 항목중에 연관규칙 X->Y를 지지하는 비율을 의미하는 지지도와 X의 모든 항목을 포함하고 있는 트랜잭션의 개수중에서 Y까지 포함하는 트랜잭션의 비율을 의미하는 신뢰도이다 [5][6].

Ⅲ. 프로모터 염기 서열 분석을 통한 유전자의 관련성 탐사

본 연구에서는 프로모터 염기 서열분석을 통한 유전자 발현 특성을 알아내기 위해 연관규칙 탐사 기법을 적용한다. 전체적인 탐사과정은 그림 2와 같고 세부 과정은 다음의 단계별로 수행하고자 한다.

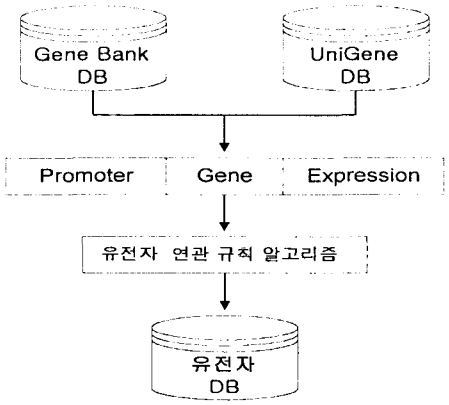

그림 2 유전자 연관성 탐사 단계

(단계 1) 연구 대상 종을 결정하고 유전자에 대한 검색을 수행한다. 이 과정에서 한 개 이상의 데이터베이스(GenBank, UniGene등)가 검색될 수 있다. 관심 주제별 데이터의 검색결과는 데이터 베이스 화한다.

(단계 2) 단계 1에서 구해진 데이터로부터 관심 대상인 유전자의 염기 서열 분석을 위해 유전자와 프로모터 부분으로 세그멘테이션(segmentation)한다.

(단계 3) 단계2에서 구해진 염기 서열에 변형된 연관규칙 알고리즘(유전자 연관규칙 알고리즘)을 적용한다

(단계 4) 단계 3에서 구해진 규칙들을 유전자 베이스에 첨가한다.

1. 단계 1

· 효모의 경우 유전자의 수가 많지 않고 실험 데이터가 정확하므로 본 연구에서는 이들 종을 대상으로 유전자 분석을 수행한다. 유전자 염기 서열

은 미국 NCBI Genbank를 이용하여 해당 종에 대한 유전자를 검색하였다. 그림 3는 효모 종의 하나인 Saccharomyces 의 유전자들에 관한 검색화면이다. 검색어 란에 해당 종(Saccharomyces)과 유전자(genomic)를 입력하면 그림 4의 결과를 제공해준다.

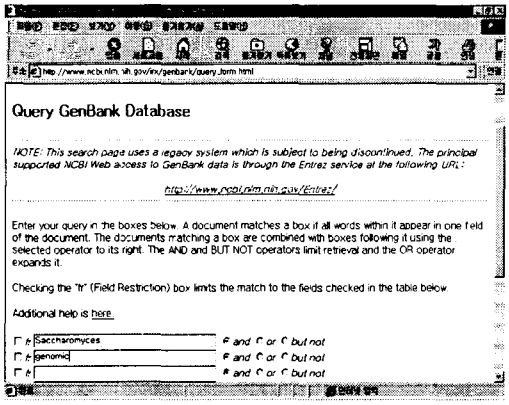

그림 3 GenBank Database 검색예

```
     promoter        645..658                /gene="PHO88"
                     /note="TATA-like sequence"
                     /db_xref="SGD:S0005361"     promoter        678..683
                     /gene="PHO88"
                     /note="TATA-like sequence"
                     /db_xref="SGD:S0005361"     CDS             751..>846
                     /gene="PHO88"
                     /note="PHO88 regulatory protein (AA 1-32)"
                     /codon_start=1                      /db_xref="SGD:S0005361"
                     /protein_id="CAA68454.1"
                     /db_xref="GI:4167"
                     /db_xref="SWISS-PROT:P28852"
                     /translation="MESTSGERSEMIHEDQGIPKVILPADFMKCSR"
BASE COUNT     258 a    162 c    140 g    286 t ORIGIN
        1 gagctcaaca acggaagacc gcctatttaac cccgagtgtc tgttttttag tatactcgaa
       61 gcaaacaaga aaagtttgtt acttaatttc tggctaaaac aaggtagtgc taccctagag
      121 gaagtacgat atcagtacac ttttgtatttt ttaacgaatt atggtattct ttaaaatatt
      181 taacatgcta tcgagttttt ctttctaaaa tgcggtttca aaatttgagt ccgctttcta
      241 tggaaaatat gaatgaggcg aacgtgcttg taccatagtc agattttaac ccgtaggtga
      301 gagcctgagg cttctgtaaa ccacctttgga aacacatacg ggttataaga ttttaacaaa
      361 caatcagtct ccatactcat cctaggatct cttcttccatt tgatgaataa aggattcaaa
      421 cttcagaaag ttattcagcg tatattggct ttccttttaat ctaatgcccc aagccatcat
      481 aaatagccat attactttga aagtgctcat ccgtatacac tgcaacttt ctttttttct
      541 ctttagccga ttatcagctcc taccttgtttt ttttttcttc ttaagattat ttatattcca
      601 attttatcat ccgtattgtt tccaatagcg ttccagtgat cagttatata cttcgagagg
      661 cagataacct ttcttttttat aagtgtttat caaatttaag tctgcaagct atcataagac
      721 gaggatatcc tttggagact catagaaatc atggaaagca catcaggaga acgttccgaa
      781 aatatacatg aggatcaagg gatacccaaaa gtaattctgc ccgctgattt taataaatgc
      841 tctaga//
```

그림 4 [YSCCHR1RAA]Saccharomyces cerevisiae chromosome I right arm sequence.

· UniGene은 인간(human), 생쥐(mouse), 쥐(rat)로부터 유래한 GenBank의 염기 서열을 유전자를 기준으로 서로 중복되지 않도록 클러스터화한 데이터 베이스이다. 각 UniGene 클러스터는 해당 유전자를 대표하는 염기 서열과, 발현되는 조직, 염색체상의 위치등에 대한 정보를 포함하고 있다. 그림 5는 UniGene 데이터 베이스의 사람의 유전자에 대한 검색예 이다.

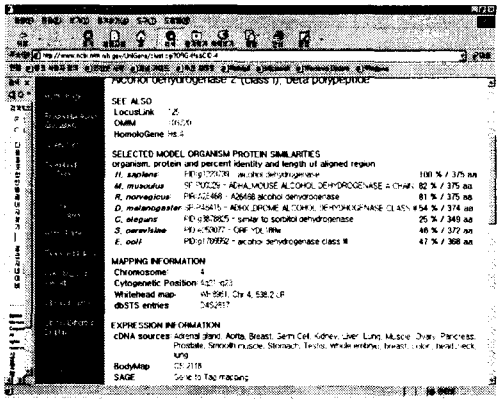

그림 5 UniGene Database 검색 예

2. 단계 2

· 단계1에서 구해진 데이터 그림 4 로부터 관심 대상인 유전자의 염기 서열을 분석하기 위해 유전자와, 프로모터 부분으로 세그멘테이션(segmentation)한다. 세그멘테이션은 그림 4의 GenBank 데이터 상단의 설명부분을 참조하여 대상 유전자, 프로모터로 구분하고 그에 따른 발현특성은 UniGene 데이터베이스 정보를 이용하여 조합한다.

3. 단계 3

· 단계2에서 구해진 염기 서열에 연관규칙 알고리즘을 적용한다. 본 연구에서는 위의 연관규칙 탐사 기법을 그대로 적용할 수는 없다. 첫째, 기존에 연구되어진 연관규칙 탐사 알고리즘들은 트랜잭션 데이터를 대상으로 하였기 때문에 데이터의 형태가 다르다. 도출된 연관규칙의 형태도 맞지 않을것이므로 변형된 알고리즘을 설계한다. 둘째, 생물학적으로, 염기 서열이 완전히 동일하지는 않지만 유사한 서열을 가졌음에도 발현되는 특성이 같으면 동일한 유전자로 취급한다는 점이다. 예를들어 생물학에서 TATA box, CAT box, GC box라는 염기 서열 상자들이 있다. TATA 상자라는 명칭은 첫 네 개의 염기로부터 유래한 것으로 염기 서열의 조합이 T-A-T-A 의 서열 순과 유사하게 나타나는 조합을 의미하며 특정 염기 서열의 조합이 이와 유사하다면 같은 염기 서열로 간주하는 것을 의미한다. 이 문제

또한 규칙 탐사단계에서 조절 파라메터 로서 제한해야 한다.

단계 3을 그림 6의 예제로 논의한다.

· 특정 종의 유전자의 전체적인 염기 서열이 그림 5라 하자. 유전자 앞에 존재하는 일정 영역의 염기 서열이 유전자의 발현을 조절하는 프로모터 부분이다. 연관규칙 정의(X->Y)에 의해 조건부에 나타나는 X의 요소 프로모터를 P1, P2, P3, P4로 정의하고 Y요소는 발현특성 노란머리, 파란눈 으로 정의한다.

P1 : tattagcctt tgtttcgtaa tccttactgt taatggttg ttcaataccg: Gen1 : 노란머리
P2 : tgttagcctt tttatggtac tatttacttt tcgcaatttc agcatcaata: Gen2 : 노란머리
P3 : ccttgacgag aagatgtaca ttgacttctg ggttaacaga ttgcaatacg: Gen3 : 파란 눈
P4 : ccttttcgga aggatgacga acaacagtga cttcgtcgat ggccttgatc: Gen4 : 파란눈

그림 6 염기 서열

그림 6의 염기서열을 분석해보면 다음과 같은 정보를 밝혀낼수 있다.

(1) tattagcctt, tact 서열이 음영되어진 Gen1과 Gen2의 프로모터에 공통적으로 나타나고 이때 발현되는 특징은 노란머리를 갖는다는 것을 알수 있다.

(2) cctt, gatg 서열이 음영되어진 Gen3과 Gen4의 프로모터에 공통적으로 나타나고 이때 발현되는 특징은 파란눈을 갖는다는 것을 알 수 있다.

(1), (2)의결과에 대한 연관 규칙은 아래와 같이 정의된다. 선택된 프로모터들은 다음과 같은 규칙의 형태로 정의되며, 일정 신뢰도 이상의 프로모터들을 ∧ 조합으로 구성한다.

If P1 ∧ P2 ..... then "**********", ∨
If P1 ∧ P2 ..... then "**********", ∨

.

그림 6의 결과로 도출될 수 있는 연관규칙은 아래와 같다.

· If tattagcctt ∧ tact then expression="노란머리" : 프로모터의 염기서열에 tattagcctt, tact 이 존재하면 노란머리를 갖는다.

· If cctt ∧ gatg then expression="파란눈" : 프로모터의 염기 서열에 cctt, gatg 이 존재하면 파란눈을 갖는다.

발견 규칙으로부터 프로모터와 같은 염기 서열을 갖는 다른 프로모터들은 특정 유전자를 결정지을 수 있고 이 유전자를 통해 발현되어지는 특징은 유사한 서열을 갖는 다른 유전자에서도 발현되리라는 것을 예측할 수 있다.

4. 단계 4

· 연관규칙 알고리즘에 의하여 단계3에서 구하여진 규칙들을 유전자 베이스에 첨가한다.

Ⅳ. 결론 및 향후 연구

본 논문에서는 분자 생물학에서 매우 중요한 연구 대상으로 삼고 있는 프로모터 서열과 유전자간의 연관성의 연구에 연관규칙 탐사기법을 사용하여 그들간의 발현되는 특징을 탐사하였다. 본 논문의 결과는 실제 생물학적 실험대상의 후보 조합을 최소화할 수 있으며, 탐사된 규칙으로 분자 생물학 실험실에서는 발견된 서열이 어떤 생물학적 메카니즘을 통해 발현을 조절하는지 실험을 계획할 수 있다. 또한 관련 분야의 실험실에서 효과적인 생물정보의 분석을 가능하게 할 것이며, 실험 결과를 미리 유추하는데 많은 도움을 줄 것이다.

향후 연구 과제로서는 생물학적으로 프로모터의 기작이 생물 종 에 따라 여러 가지 다양성을 수반하기 때문에 이를 최대한 수용할 수 있도록 하고 언급된 여러 가지 염기 서열 상자를 고려한 변형된 연관규칙 탐사 알고리즘을 제시해야 한다. 또한 특정 종에 대한 범위 결정과 발현 특성을 정의하는 문제 또한 연구 하고자 한다.

참고문헌

[1] R. Hofestaedt, "Computer science and biology-the German Conference on Bioinformatics (GCB'96), BioSystems 43 (1997) 69-71

[2] Brachman, R. J. and Anand T. (1996), The Process of Knowledge Discovery in Databases. Advance in knowledge Discovery in Database

and Data Mining(37-57). Menlo Park:AAAI/ MIT Press

[3] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D, "GeneCards:a novel functional genomics compendium with automated data mining and query reformulation support, Bioinformatics, V.14 N.8, pp.656-664, 19980801

[4] Setubal J, Meidanis J. Introduction to Computational Molecular Biology. Boston, MA:PWS Publishing Company, 1997[7].

[5] R. Agrawal, T. Imielinski and A. Swami. "Mining Association Rules between Sets of Items in Large Database", Proc, ACM SIGMOD, 1993, pp.207-216

[6] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules", Proc, VLDB, 1994, pp.487-499

김 정 자

1985 전남대학교 자연과학대학 계산통계학과 (이학사)

1988 전남대학교 자연과학대학 계산통계학과 (이학석사)

1997.3 - 1999.2 전남대학교 자연과학대학 전산통계학과 박사수료

관심분야 : 데이터 마이닝, OLAP, 데이터 웨어하우스, 바이오 인포매틱스

E-mail : jjkim@dbcore.chonnam.ac.kr

이 도 헌

1990 한국과학기술원 과학기술대학 전산학과(공학사)

1992 한국과학기술원 전산학과 (공학석사)

1995 한국과학기술원 전산학과 (공학박사)

1995년 9월- 1995년 12월 인공지능연구센터 연구원

1996년~ 현재 전남대학교 전산학과 조교수

관심분야 : 데이터 마이닝, 데이터 웨어하우스, 워크 플로우 관리, 바이오 인포매틱스

E-mail : dhlee@dbcore.chonnam.ac.kr