의존성 반영 분해모델에 의한 유전자의 핵심 프로모터 영역 예측

(Prediction of Core Promoter Region with Dependency-Reflecting Decomposition Model)

김 기 봉 † 박 기 정 † 공 은 배 ††
(Ki-Bong Kim) (Kiejung Park) (Eun Bae Kong)

요 약 다수의 미생물 유전체 프로젝트들이 완료되면서 엄청난 양의 유전체 핵산 염기서열 데이타들이 양산되고 있다. 이러한 상황에서 전산 기법을 이용하여 유전체 DNA 염기서열 상에서 유전자의 프로모터 영역을 규명하는 문제는 최근에 상당한 연구의 관심대상으로 떠오르고 있다. 본 논문에서는 전사조절의 핵심 역할을 하는 -10 영역과 전사개시 부위를 포함한 원핵생물의 핵심 프로모터 영역에 대한 의존성 반영 분해모델 (Dependency-Reflecting Decomposition Model)을 제안한다. 이 모델은 인접한 위치에 존재하는 핵산 염기들 사이의 의존성뿐만 아니라 인접하지 않은 위치의 핵산 염기들간의 의존성까지 고려함으로써 핵산 염기서열 상에 내포되어있는 중요한 생물학적 의존성들을 함축하고 있다. DRDM 모델은 우수한 성능평가 결과를 보였으며, 미생물 유전체 Contig들 상에서 임의의 유전자 프로모터를 예측하는데 효과적으로 이용될 수 있다.

키워드 : 유전체, 유전자, 프로모터, 전사조절, 원핵생물, 염기서열, 의존성 반영 분해모델

Abstract A lot of microbial genome projects have been completed to pour the enormous amount of genomic sequence data. In this context, the problem of identifying promoters in genomic DNA sequences by computational methods has attracted considerable research attention in recent years. In this paper, we propose a new model of prokaryotic core promoter region including the -10 region and transcription initiation site, that is, Dependency-Reflecting Decomposition Model (DRDM), which captures the most significant biological dependencies between positions (allowing for non-adjacent as well as adjacent dependencies). DRDM showed a good result of performance test and it will be employed effectively in predicting promoters in long microbial genomic Contigs.

Key words : Genome, DNA, Prokaryote, Promoter, -10 region, Transcription, Dependency-Reflecting Decomposition Model, Contig

1. 서 론

인간 유전체 프로젝트(Human Genome Project) 개시 이후 모델 유기체 및 각종 미생물들을 대상으로 하는 염기서열 결정 프로젝트들이 활발히 진행되었다. 인간의 초안 유전체 염기서열이 2000년 상반기에 이미 세상에 발표되었고, 완료된 미생물 유전체가 이미 90여종에 이른다. 또한 현재 진행중인 유전체 프로젝트만 해도 700 여건에 이르는 것으로 추정된다. 이러한 유전체 프로젝트들은 엄청난 핵산 염기서열 데이타를 양산하고 있으며, 양산된 막대한 양의 분자 정보들을 해석하고 이해하는데 초점이 맞추어져 있는 전산생물학 (Computation Biology)은 눈부신 발전을 거듭하고 있다. 비록 전산생물학은 추정 유전자와 그에 상응하는 기능들을 규명하는데 많은 기여를 하고 있으나, 현재 이용 가능한 유전체 정보들을 해독하기 위해서 해야 할 일들이 여전히 많이 산적해 있다. 특히 유전자가 자신의 최종 산물을 생성하는 유전자 발현과정과 그러한 과정에서 이루어지는 복잡한 유전자 제어 및 조절에 있어서 더욱 더 그러하다.

† 비 회 원 : (주)스몰소프트
kbkim@bioinfo.smallsoft.co.kr
kpark@bioinfo.smallsoft.co.kr
†† 종신회원 : 충남대학교 컴퓨터공학과 교수
keb@cc.cnu.ac.kr
논문접수 : 2002년 9월 9일
심사완료 : 2003년 2월 4일

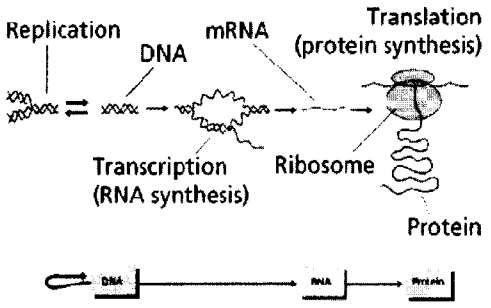

그림 1 Central Dogma : 유전 정보가 흐르는 원칙

그림 1에서 보는 것처럼 유전 정보의 흐름은 유전 정보인 DNA가 자신의 복제판을 만들기 위한 복제(Replication) 과정을 제외하고는 일방통행식이다. 즉, 유전 정보인 DNA 중 특정한 유전자가 전사(Transcription) 과정을 통해서 mRNA를 만들고, 이렇게 생성된 mRNA의 정보가 번역(Translation) 과정을 거쳐 활성이 있는 단백질(Protein)이 만들어진다. 이와 같이 유전자 발현은 DNA로부터 RNA 분자로 그리고 나서 RNA로부터 단백질 분자로의 유전정보의 이동을 말하며, 앞에서 언급한 것처럼 전사와 번역 단계로 나누어 생각할 수 있다. 전사는 유전자 발현의 첫 단계로서 초기에 전반적인 유전자 발현의 흐름을 제어할 수 있는 에너지 효율성 측면뿐만 아니라 생명현상의 경제적 측면에서 유전자 제어의 중요한 기점을 이루고 있다[1]. 즉, 세포내의 유전자 발현의 제어는 대부분 전사단계에서 이루어진다. 전사단계의 유전자 발현을 제어하는데 관여하는 많은 요소들이 있으나 대표적으로 유전자 바로 앞부분에 위치하고 있는 프로모터(Promoter) 영역과 이 영역에 결합하여 실제 전사작업의 개시를 촉발하는 여러 전사관여 효소 및 조절인자(Transcription Factor)들이 존재한다. 전사의 개시는 먼저 RNA 중합효소(Polymerase)가 DNA 주형에 결합함으로써 이루어지며, 이어서 DNA 이중나선의 일부분이 붕괴되고 RNA 중합효소는 전사 개시점을 인지하여 상보적인 염기쌍 처리 과정을 통해서 mRNA을 합성하기 시작한다. RNA 중합효소가 DNA를 따라 이동하면서 mRNA 생성하다가 특정 염기서열을 만나면 전사가 종결되는데, 그 특정 염기서열을 전사 종결자(Transcription Terminator)라 한다. 프로모터는 RNA 중합효소가 DNA 주형에 처음 결합하는 부분이며, 정확한 아미노산 서열의 단백질 형성을 위해서 RNA 중합효소가 결합하는 정확한 위치를 결정해 주며, 자신의 염기서열의 방향성에 의해 DNA 두 가닥 중 전사에 이용될 가닥을 결정해 주는 매우 중요한 전사조절 신호 부위이다. 대부분의 경우 프로모터는 전사 개시점을 기준으로 해서 상위방향(즉, 전사진행 방향의 역방향을 의미함)으로 200~500 bp의 넓은 영역을 이루고 있고, 다양한 많은 전사 조절인자들의 결합부위를 갖고 있다. 이와 같이 프로모터의 핵산 염기서열은 전사 개시점의 위치를 결정하며, 나아가서는 유전자가 전사되는 빈도수에 중요한 요인으로 작용한다(즉, 프로모터의 강도를 의미함).

이러한 문맥에서 전산기법에 의해 유전체 DNA 염기서열에서 특정 전자조절인자 결합부위 및 프로모터 영역을 예측하는 문제는 상당히 중요한 최근의 연구관심사로 대두되고 있으며, 활발한 연구들이 진행되고 있다. 임의의 염기서열로부터 전사 조절인자 결합부위나 프로모터 영역을 예측하는 것은 전사와 번역의 정확한 핵산서열 결정자들을 규정하는 본질적인 생화학적 문제와 상당히 밀접한 관련이 있을 뿐만 아니라, 유전자의 발현양상을 예측하거나 유전자를 규명 하는데 결정적인 단서를 제공할 수 있다. 그러나 어쨌든 유전자의 프로모터를 예측하는 문제는 그 차제로서 분명히 흥미롭고 중요한 연구 관심사임이 틀림없다. 따라서 프로모터 예측은 유전자 규명을 위한 프로그램의 하나의 하부모듈로서 개발되기도 하고, 또한 독립적인 모듈로서 개발되기도 한다.

프로모터 예측에 사용되는 기존의 알고리즘들은 전사조절인자의 결합부위에 해당되는 조절신호 부위의 규명에 근간을 두고 있다. 비록 15년 넘게 이러한 조절신호 부위에 대한 예측 문제가 다루어져 왔으나, 아직도 해결되지 않은 채 개선해야 할 많은 문제점들이 남아 있다. 이러한 문제점들은 대부분 염기서열을 단순한 문자열로 간주함으로써 실제 염기서열 내에 내포되어 있는 생명현상의 상황효과를 제대로 반영하지 못하기 때문에 기인한다고 할 수 있다. 이러한 측면에서 본 논문은 조절신호 부위에 역점을 두고 있는 것이 아니라 넓은 프로모터 영역 중에서 정보량(Information Content)이 상대적으로 많은 신호 부위를 먼저 탐색하고, 그런 부위를 기준으로 하여 위치별 핵산 상호간의 의존성 및 연관성을 잘 함축함으로써 생명현상의 상황효과(Context Effect)를 잘 반영할 수 있는 모델인 의존성 반영 분해 모델(Dependency-Reflecting Decomposition Model : DRDM)을 제안하고 있다. 진핵생물과 원핵생물은 유전자의 구조와 유전자의 발현 양상이 상당히 다르기 때문에 당연히 다르게 취급되어야 한다. 따라서 본 논문에서 제안하고 있는 DRDM는 원핵생물에 초점을 맞추었으

며, 특히 원핵생물 중에서 실험적으로 가장 잘 규명되어 있는 *E. coli*를 대상으로 하고 있다. 실험적으로 입증된 *E. coli*의 프로모터들을 근거로 프로모터 영역이 갖고 있는 신호부위의 염기서열 패턴들의 위치 상호간의 의존성 및 문맥적 특성을 가장 잘 반영할 수 있도록 DRDM을 고안하였다. 이 모델을 사용하여 실제 교차검증을 통해 *E. coli* 프로모터를 예측해본 결과 뛰어난 예측 결과를 보였다. 임의의 *E. coli*나 진화론적으로 가까운 박테리아 유전체 염기서열들 상에서 정확한 프로모터 영역을 우리가 제안하고 있는 DRDM 모델을 이용하여 예측함으로써 유전자 발현 양상 및 신규 유전자 규명에 중요한 단서를 얻을 수 있으므로, 유전체 연구자들의 연구 효율성 및 연구성과를 극대화 시킬 수 있을 것이다.

본 논문의 전체적인 구성을 살펴보면, 2장에서는 기존의 프로모터 예측 방법 및 문제점에 대해서 알아보고, 3장에서는 DRDM 모델을 만드는데 중요한 재료가 된 *E.coli*와 *E. coli*유전자의 전사에 관여하는 RNA 중합효소의 구조 및 특성에 대해 살펴본다. 4장에서 우리가 제안하는 DRDM 모델에 대해서 살펴보고, 5장에서는 실제 실험적으로 규명된 *E. coli* 프로모터를 가지고 DRDM 모델을 적용한 실험결과 및 향후 연구과제에 대해 논의한다.

2. 기존의 프로모터 예측 방법 및 문제점

기존의 프로모터 영역 예측 방법들은 업밀히 말하면 전사 조절인자 결합부위 예측에 초점을 두고 있다고 할 수 있다[2,3]. 즉, 특정 전사 조절인자 결합부위를 탐색함으로써 특정 기능을 띠는 유전자의 프로모터를 규명하는 것이다. 그러한 까닭은 종(Species)별로 그리고 같은 종(Species)이더라도 특정 기능별 유전자에 관여하는 전사 조절인자가 다르고, 그러한 전사 조절인자가 결합하는 프로모터 영역내의 위치 및 인식하는 핵산 염기들의 패턴 양상도 다양하기 때문이다. 이러한 방법들은 특정 전사 조절인자가 인식하는 염기서열 패턴들을 Consensus 서열, 정규표현, 프로파일(Profile) 및 가중치 매트릭스(Weight Matrix), 그리고 은닉 마르코프 모델(Hidden Markov Model) 형태로 표현한다. Consensus 서열을 이용할 경우에는 Consensus 서열과의 유사성(Similarity) 정도를 측정하여 전사 조절인자 결합부위를 예측하며, 정규표현을 이용할 경우에는 일반적인 패턴검색 방법으로 전사 조절인자 결합부위를 예측한다. 이러한 예측 부위를 근거로 포괄적인 프로모터 영역을 추정할 수 있다. 그러나 이들 두 가지 접근 방법은 근본적으로 민감도(Sensitivity)와 선택도(Selectivity) 측면에서 많은 취약점을 갖고 있다. Consensus 염기서열은 각 위치별로 가장 많이 출현되는 염기를 할당하여 표시하는 방법으로 예측 결과에 있어서 False Negative Rate가 상당히 높을 수 밖에 없어 민감도가 상당히 떨어진다. 정규표현의 경우 예외적인 경우에 해당하는 것과 최상의 경우에 해당하는 것(즉, Consensus일 경우)을 차등화할 수 있는 능력이 없어 당연히 False Positive Rate가 상당히 높게 나타나게 되므로 선택도가 낮을 수 밖에 없다. 위치별 염기의 발생빈도에 따라 가중치를 달리 둠으로써 Consensus 서열과 정규표현이 갖고 있는 문제점을 어느 정도 극복할 수 있는 프로파일이나 가중치 매트릭스를 사용하는 경우 Log-odds Score에 의한 스코어 스킴(Score Scheme)으로 전사 조절인자 결합부위를 예측한다. 달리 선택할 경우의 수가 많지 않다면 올바른 부위를 정확히 밝혀낼 수 있다는 점에서 대다수의 경우 단순한 프로파일 기법들이 상당히 잘 작동한다고 밝혀져 있다[4]. 이 점에서 가중치 매트릭스 모델는 간단하고, 직관적으로 쉽게 이해할 수 있고, 이용하기 쉽다는 점에서 중요한 이점을 갖고 있다. 게다가, 상대적으로 적은 수(즉, 10여 개에서 몇 백개 정도까지)의 신호 핵산 염기서열들이 이용될 수 밖에 없을 때, 가중치 매트릭스 모델은 사용하기에 가장 최상의 모델 유형으로 인식되고 있다. 그러나 이러한 가중치 매트릭스 모델의 한계점은 실제 생명현상에서 중요한 요인이 될 수 있는 위치 사이의 의존성을 전혀 반영하지 않고 있다는 것이다. 이러한 측면에서 은닉 마르코프 모델은 인접한 위치간의 의존성을 반영하고 있다는 점에서 나름대로의 장점을 갖고 있어 널리 사용되고 있다. 그러나 이 또한 Gap이 있는 위치들간의 의존성을 반영하지 못한다는 단점을 갖고 있다.

전산생물학적 접근 방법에서 간과하기 쉬운 것은 염기서열을 단순히 하나의 문자열로 취급하는 것이다. 그러나 실제 생명체 내에서는 3차원 구조 내에서 염기서열 내부간의 상호작용 뿐만 아니라 나름대로 많은 다른 분자들과 상호작용을 한다. 즉, 1차원적인 구조에서 멀리 떨어져 있는 위치들이 실제 아주 가까이 인접해 있을 수 있다는 것이다. 게다가 전사에 관여하는 효소들은 거대 분자인 단백질로서 핵산 DNA와 일정 부분 결합하면서 상호작용을 하는데, 이때 결합되는 영역 내의 염기들은 다소 떨어져 위치하는 경우가 많다. 동시에 서로 다른 염기서열에 결합하여 전사조절에 관여하는 인자들끼리 상호 작용하는 경우에는 이들 인자들의 결합부위들 사이에 상당한 연관성 및 의존성을 띠고 있다고 볼

수 있다. 그러나 기존의 방법들은 이러한 DNA의 구조적인 특성이나 서로 다른 조절부위들 간의 상호작용 등을 포함한 전반적인 상황효과(Context Effect)들을 고려하지 않고 있다. 이로 인해 민감도와 선택도 측면에서 개선해야 할 여지가 많다고 할 수 있다.

기존 방법들의 민감도 및 선택도의 취약점을 보완하고, 생명체내의 실제적인 상황효과를 충분히 반영하기 위해 인접한 위치 뿐만 아니라 인접하지 않은 위치들간의 의존성을 충분히 반영하는 새로운 모델을 제안하게 되었다. 즉, 기존의 가중치 매트릭스 모델과 은닉 마르코프 모델의 단점을 보완하고, 민감도와 선택도를 높이기 위해 원핵생물의 핵심 프로모터 영역의 인식 측면에서 관측 가능한 모든 의존성들을 보다 정교하게 반영할 수 있는 훨씬 적합한 모델을 개발하였다. 또한 기존의 방법들은 특정 조절인자 부위를 찾아내어 특정 기능 프로모터, 즉, 특정 유전자의 프로모터를 검색하는 경우에 해당되나, 우리가 제시하는 방안은 대규모 염기서열 결정 프로젝트로부터 쏟아지는 임의의 염기서열에 대해 전사 개시점을 비롯한 핵심 프로모터 영역을 예측하는 것으로써 근본적으로 그 접근 방법이 다르다 하겠다.

3. *E. coli*와 RNA 중합효소

우리가 제안한 DRDM모델을 만들기 위해 사용한 대상 유기체(Target Organism)는 원핵생물인 *Eschericahia coli*로서, 이는 오랫동안 유전자 조절 연구의 모델 시스템으로 유전자 조절 및 물질대사 등의 메커니즘에 대해 가장 잘 밝혀져 있는 유기체이다. 사실 이러한 이유에서 *E. coli*와 관련하여 여러 개의 데이타베이스[5] 뿐만 아니라 유전자조절 모델 및 이론들이 이미 여러 개 개발되어 있다[6]. 게다가, *E. coli*의 DNA 염기 서열상에서 프로모터나 조절부위의 발생을 예측하기 위한 몇몇 전산 알고리즘들이 개발되었다[7, 8]. 특히 k-12를 비롯한 여러 *E. coli* 균주(Strain)들의 염기서열 결정프로젝트가 완료되어 전체 유전체 정보가 발표된 상태이다.

박테리아(Bacteria), 원시세균(Archaea), 및 진핵생물의 RNA 중합효소는 서로 현저한 차이가 있다. 특히 *Escherichia coli*의 프로모터 핵산 염기서열을 인식하는데 관여하는RNA 중합효소는 β, β', α, 및 σ 등의 4개의 단백질 서브유닛(Subunit)으로 구성된 복합체 형태의 $\alpha_2\beta\beta'\sigma$.구조를 갖는다. 소위 이러한 홀로효소(Holo-enzyme)은 두개의 기능적인 구성성분으로 나눌 수 있다. 즉, 핵심효소 부분($\alpha_2\beta\beta'$, 또한 E로 표기되기도 함)과 시그마 인자(Sigma Factor) 부분으로 구분된다. 시그마 인자는 프로모터 핵산 염기서열을 인식하고 중합효소와 DNA간의 복합체 구성에 관여하며, 전사를 시작하는 DNA의 위치를 인지한다. 성공적으로 전사가 개시된 후에는 시그마 인자는 홀로효소로부터 방출되며[9], 대부분의 전사(RNA 합성)는 핵심효소 부분($\alpha_2\beta\beta'$)이 담당하게 된다. 여러 종류의 시그마 인자들이 존재하며, 시그마인자의 종류에 따라 중합효소가 인식하는 프로모터 핵산 염기서열의 패턴그룹이 다르다. 즉, 각기 다른 시그마인자를 갖는 각 RNA 중합효소들은 서로 다른 핵산 염기서열그룹을 갖는 특정 프로모터 부분집합을 인식한다. 이러한 양상의 생물학적 의미는 각 프로모터 그룹은 생리적으로 유사한 조건 하에서 어떤 상황에 의해 동시에 발현될 필요가 있는 유전자들을 함께 조절하고 있다는 것이다.

4. 의존성반영 분해모델(Dependency-Reflecting Decomposition Model)

4.1 모델 생성의 전체적인 절차

우리가 제안하는 인식 모델은 유전자의 핵심 프로모터 영역의 패턴들을 잘 특징 지우기 위한 것으로, 그 기반이 되는 것은 다중정렬, 정보이론, χ^2검증, 결정트리(Decision Tree)의 일종인 최대 의존성 분해개념(Maximal Dependence Decomposition) 및 은닉 마르코프 모델 등을 최대한 활용하고 적용함으로써 생성된다. 전체적인 절차는 [그림 2]에 도식화된 것과 같으며, 개괄적으로 살펴보면 다음과 같다. 먼저 해당 데이타베이스로부터 실험적으로 입증된 프로모터 염기서열을 추출해내고, 전사 개시점을 기준으로 Gap를 허용하지 않는 다중정렬을 한다. 다중정렬된 염기서열들의 위치별 정보량을 조사하여 모델화할 영역을 한정한다. 한정된 영역에 대해 χ^2검증을 통해 각 위치들 사이의 의존성을 파악하고, 의존성에 따라 데이타 집합을 보다 작은 하부

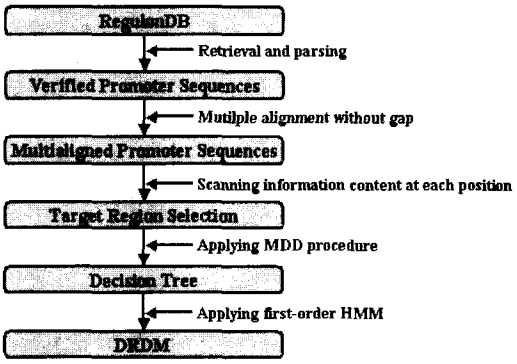

그림 2 전체적인 DRDM 생성 개요

집합으로 순차적으로 분해해서 하나의 결정트리를 구성한다. 구성된 결정트리의 말단 노드에 대해 k-order 은닉 마르코프 모델을 만들면 최종 모델이 완료된다. k값은 말단노드의 데이타 크기를 고려하여 적당히 조절하면 될 것이다. 본 논문에서는 k값을 1로 하였다.

4.2 프로모터의 염기서열 추출

우리가 제시한 인식 모델을 만들기 위해서 먼저 *Escherichia coli*의 전사조절과 오페론(Operon) 구성에 관한 데이터베이스인 RegulonDB 데이타베이스의 'promoter' 테이블로부터 실험적으로 이미 입증된 검색 가능한 핵심 프로모터의 핵산 염기서열들을 607개나 수집했다. 이러한 작업은 네트웍 프로그래밍에 의한 원격 데이타베이스 검색 자동화 및 검색결과의 체계적인 파싱(Parsing)을 통해 이루어졌다. RegulonDB 데이타베이스에 저장되어 있는 각 프로모터 엔트리의 핵산 염기서열의 크기는 동일한 81 bp이며, 전사 개시점을 기준으로 상위부분(Upstream)의 60개의 핵산 염기와 하위부분(Downstream)의 20개의 핵산 염기로 이루어져 있다. 수집한 프로모터 염기서열들을 각 시그마인자별로 분류한 결과에 의하면 607개 중 548개가 σ^{70} 클래스에 속하고, σ^{54}, σ^{38}, σ^{32}, σ^{28} 및 σ^{24} 등의 각 클래스별로 10개, 21개, 22개, 3개 및 3개에 달했고, 구분이 명확하지 않거나 두 개 이상의 클래스에 중복적으로 속하는 것을 기타로 구분하였으며 그 개수는 83개였다.

4.3 정보량(information content)에 의한 모델 대상 영역 결정

다음 단계로 우리는 프로모터 핵산 염기서열들을 전사 개시점([그림3]의 61번 위치)을 기준으로 Gap를 허용하지 않은 다중정렬을 하고, 다양한 프로모터 핵산 염기서열 데이타들의 각 부분집합 내에서 각 위치별 Shannon의 정보량을 조사했다. [그림3]의 위쪽 도표의 실선 a는 실험적으로 검증된 607개의 프로모터 핵산 염기서열 데이타에 대한 각 위치별 정보량을 나타내며, 점선 b는 시그마-70 클래스 계열로 알려진 548개의 프로모터 핵산 염기서열 데이타에 대한 각 위치별 정보량을 나타낸다. [그림 3]의 아랫쪽 도표의 실선 c는 시그마-70 클래스 계열을 제외한 나머지 142개의 프로모터 핵산 염기서열 데이타에 대한 각 위치별 정보량을 나타내며, 점선 d는 시그마-70 클래스의 프로모터 핵산 염기서열 데이타를 퓨린(Purine : A 혹은 G)와 피리미딘(Pyrimidine : C 혹은 T) 관계로 본 경우에 대한 각 위치별 정보량을 보여준다. [그림 3]을 통해 알 수 있듯이 전반적으로 전사개시 부위와 -10영역(50번 위치 주변)가 상대적으로 높은 정보량을 보인다.

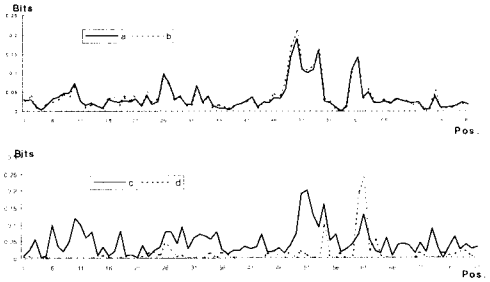

그림 3 다양한 부분집합의 프로모터 핵산 염기서열 엔트리들의 위치별 Shannon 정보량

지금부터 우리는 일반적으로 통용되는 염기서열 위치 협약을 준수하여 전사 개시점을 0으로, 전사 개시점 보다 하나씩 상위로(전사진행 방향의 역방향) 올라갈수록 1씩 감소하고 하위로 내려갈수록 1씩 증가해서 번호를 붙이도록 한다. 우리는 Claude Shannon의 이론을 바탕으로 불확실성 정도를 아래의 식으로 정의한 Tom Schneider의정의를 따랐다[10].

$$H(l) = -\sum_{B=A}^{T} f(B,l)\log_2 f(B,l) \text{ (bits per position)} \quad (1)$$

여기서 $f(B,l)$은 l번 위치에서 핵산 염기 B의 빈도수를 나타내고, B는 핵산 염기 A, C, G, 및T 중의 어느 하나를 의미하며, $H(l)$은 l번 위치에서의 불확실성을 나타낸다. 각 위치별 전체 정보량은 핵산 염기서열이 정렬되어 있을 때 불확실성의 감소량으로 표현된다.

$$R(l) = \log_2 N - H(l) \text{ (bits per position)} \quad (2)$$

여기서 N은 선택 혹은 메시지의 수, 즉, DNA인 경우 4(A, C, G, 또는 T)이고, $R(l)$은 l번 위치에 존재하는 정보의 양이다. 이 경우 주어진 임의의 위치에서 최대 불확실성 값은 2가 된다. 각 위치별 정보량을 기반으로 해서 원래의 폭 넓은 프로모터 영역(81 bp)을 상대적으로 풍부한 정보를 갖고 있는 적합한 한정영역(26 bp)으로 모델 대상영역을 국한했다. 국한된 모델 대상영역은 -21번 위치에서 +4번 위치까지 이며, 우리가 모델을 만드는데 이 부분만을 사용했다.

4.4 MDD(Maximal Dependence Decomposi tion) 절차에 의한 분해과정

인접한 위치의 핵산 염기서열들간의 상호 의존성뿐만 아니라 2개 이상 떨어져 있는 핵산 염기서열들간의 모든 상호 작용들을 모델에 반영하기 위해, 모델 대상 영역 (-21~+4)내의 모든 위치 사이의 핵산 염기들간의 상호 의존성 정도를 고려했다. χ^2 통계치를 이용해서 핵산 변수 N_i와 N_j사이의 의존성을 측정하여, i번 위치에서 어떤 핵산 염기(들)의 발생과 j번 위치에서 어떤

핵산 염기(들)의 발생 사이의 상호 연관성(혹은 의존성)이 있는지 여부를 파악했다. N_i와 N_j는 핵산 염기서열 내의 i번 위치와 j번 위치에서의 임의의 핵산 염기, 즉 A, C, G, 및 T를 의미한다.

표 1 핵산 변수 N_{-7}와 N_{-11}사이의 사건 테이블.

N_{-11}	N_{-7}								
	A		T		G		C		Total
	O	E	O	E	O	E	O	E	O
A	29	63	179	117	25	33	27	46	260
T	66	47	46	87	27	25	54	34	193
G	31	20	23	36	16	10	10	14	80
C	22	18	25	33	16	10	17	13	74
Total	148		273		78		108		607

표 1은 그러한 일례로서 607개의 프로모터 핵산 염기서열로 이루어진 서열 집합에서 -11번 위치와 -7번 위치 사이의 일반적인 4×4 사건 테이블(Contingency Table)을 보여주고 있다(사건의 관찰값은 "O"로 표현하고, 예상값은 "E"로 표시함). 관측된 값 $\chi^2=199.96$은 $p<0.001$ 수준에서 위치들간의 함축적인 의존성 정도를 나타낸다. 사건 테이블 자료를 살펴보면, 이러한 의존성의 대부분은 N_{-11}이 핵산 염기가 A일 때 그에 상응하여 -7번 위치에서 핵산 염기 T의 발생빈도가 높아지는 긍정적 연관에서 기인한다. 그 외에도 -11번 위치에서 핵산 염기 A의 발생과 -7번 위치에서 핵산 염기 A의 발생 사이에 다소 미세한 부정적 연관이 보인다. [표 1]에서 전체적으로 가장 주목할만한 특징은 -7번 위치의 핵산 염기 분포가 N_{-11}이 핵산 염기 A(-11번 위치의 Consen- sus 핵산 염기) 인지 여부에 매우 의존적이란 것이다. 의존성 정도를 척도로 이용하여 앞의 과정에서 이미 확보한 실험적으로 검증된 모든 프로모터 핵산 염기서열 엔트리들을 두개의 새로운 하부그룹으로 분해한다. 이러한 순차적인 분해는 S. Burge와 S. Karlin이 인간 유전자의 Donor Splice Site 예측에 처음 적용한 MDD (Maximal Dependence Decomposition) 절차에 기반을 두고 있다. MDD 절차는 정렬된 신호서열 집합으로부터 인접한 의존성 뿐만 아니라 인접하지 않은 의존성까지 반영하는 모델을 생성하는 것을 목표로 하고 있다[11]. 본질적으로 이러한 모델은 신뢰할 수 있을 정도로 충분한 데이타들이 가용적으로 주어질 경우 비조건부 가중치 매트릭스 모델의 확률값을 적합한 조건부 확률값으로 대체시킴으로써 가능하다.

의존성에 따른 순차적 세부 분해 방법은 다음과 같다. 길이가 k인 N개의 정렬된 핵산 염기서열 엔트리로 이루어진 데이타 집합 D가 주어지면, 첫 번째 단계로 $i\neq j$인 각 i번과 j번 위치의 N_i와 N_j에 대한 χ^2 통계값인 $\chi^2_{i,j}$를 계산한다. 만약 인접한 위치 사이뿐만 아니라 비인접한 위치 사이에 강한 의존성이 있으면 아래의 절차를 차례로 진행한다 :

(1) 각 i번 위치에 대한 핵산 변수 N_i와 i번 위치를 제외한 나머지 모든 위치의 핵산 염기들과의 의존성 양의 척도가 되는 χ^2값들의 합 $S_i=\sum_{i\neq j}\chi^2_{i,j}$을 계산한다(행의 합).

(2) S_{i1}의 값이 최대가 되게 하는 i_1값을 선택하고, 데이타 집합 D를 두개의 부분집합 D_{i1}와 $D_{i1}-$로 분해한다. D_{i1}는 i_1번 위치에서 높은 빈도수를 갖는 Consensus 핵산 염기(들)을 갖고 있는 모든 핵산 염기서열 엔트리로 이루어지고, $D_{i1}-$는 그렇지 않은 모든 핵산 염기서열 엔트리로 이루어진 부분집합이다.

(3) 부분집합 D_{i1}와 $D_{i1}-$에 대해 앞의 두 단계를 반복하고, 그 과정에서 다시 만들어진 부분집합들에 대해 또 다시 위의 분해과정을 같은 식으로 수행하다 보면 최대한 $k-1$ 수준을 갖는 이진트리가 만들어진다(그림 4).

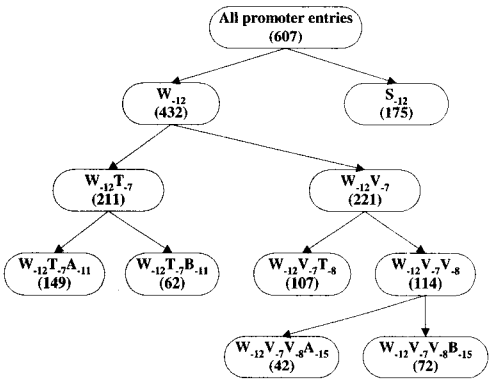

그림 4 MDD절차에 의해 생성된 607개 프로모터 데이타에 대한 이진트리

이러한 반복적인 세부 분해과정이 이진트리의 각 노드에 대해 연속적으로 수행하다가 다음의 3가지 조건 중 어느 하나라도 만족되면 수행을 종결한다.

i) 이진트리의 $k-1$번째 수준에 도달한 경우(더 이상 분해하는 것이 불가능한 상태).

ii) 분해해야 할 부분집합에서 위치별 핵산 염기들 사이에

서 유의수준 이상의 의존성이 감지되지 않은 경우.

iii) 세부 분해과정의 결과로 만들어진 부분집합의 핵산 염기서열 엔트리의 수가 미리 설정해 둔 최소값 M 이하로 떨어질 경우(분해한 후에 데이타의 개수가 너무 작아서 신뢰할 만한 조건부 확률값이 결정되어질 수 없을 경우).

그림 4는 앞에서 언급한 607개의 프로모터 핵산 염기서열 엔트리 데이타 집합에 MDD 절차를 적용한 결과를 보여준다. 최초의 분해는 -12번 위치의 Consensus 핵산 염기 W(핵산 염기 A 혹은 T를 의미함)를 기준으로 이루어졌으며, 그 결과로 432개의 프로모터 핵산 염기서열 엔트리를 갖는 부분집합 W_{-12}와 175개를 갖는 부분집합 S_{-12}(S는 핵산 염기 C 또는 G를 의미)로 나누어졌다. 부분집합 S_{-12}에서 각 위치별 핵산 염기들 사이에 유의수준 이상의 의존성이 파악되지 않으므로 더 이상의 분해는 없다. 그러나, 부분집합 W_{-12}는 데이타 집합의 크기가 충분히 크고, 위치별 핵산 염기들 사이에 유의수준 이상의 의존성이 존재하기 때문에(여기서는 그 데이타를 보여주지 않고 있음), 7번 위치의 Consensus 핵산 염기 T의 유무를 근거로 해서 분해되었으며, 그 결과로 부분집합 $W_{-12}T_{-7}$와 $W_{-12}V_{-7}$(V는 핵산 염기A, C 혹은 G를 의미함)로 분해되었다. 이런 식으로 계속 진행한다. 이와 같이 이 방법의 기본 개념은 먼저 현존하는 가장 유의수준이 높은 의존성을 먼저 고려하고, 그것을 기준으로 데이타의 집합을 분해하고, 분해된 하부 데이타 집합에서 또 다시 현존하는 가장 유의수준이 높은 의존성을 고려하고 타당성이 있을 경우 또 다시 분해하는 방식으로 반복적으로 의존성을 분석한다. 이렇게 최종적으로 만들어진 이진트리의 각 말단 노드들은 학습을 위해 사용된 핵심 프로모터 핵산 염기서열 엔트리들의 부분집합을 갖게 된다.

4.5 은닉 마르코프 모델 적용을 통한 최종 DRDM 생성 완성

MDD 절차에 의해 생성된 이진트리의 각 말단 노드의 염기서열 데이타 집합에 대해 1차 은닉 마르코프 모델을 적용함으로써 우리의 최종적인 복합모델인 의존성 반영 분해모델이 완성된다. 은닉 마르코프 모델을 만들 때 데이타 집합의 크기가 크면 클수록 보다 높은 차수가 선호되나 학습 데이타의 양을 토대로 본 논문에서는 1차수로 결정을 하게 되었다. 문제의 본질상 일반적으로 유전자 예측을 위해 사용하는 복잡한 은닉 마르코프 모델[11]이나 단백질의 특정 도메인 및 모티프를 모델링하기 위해 사용되는 삽입 및 삭제를 허용하는 프로파일 은닉 마르코프 모델[12]등을 사용할 필요가 없다. 본 논문에서 사용한 것은 단순하면서 직관적인 left-to-right 구조에 기반을 두고, 공백상태(gap state)가 없는 은닉 마르코프 모델을 사용한다. 각 말단 노드의 염기서열의 크기, 즉 윈도우 크기가 26(-21~+4)이므로 윈도우내 각 위치가 순차적인 하나의 상태(state)이고 이들 상태간의 전이는 순차적으로 반드시 이루어지므로 전이확률(transition probability)은 1이다. 그리고 각 상태에서의 해당 염기서열 생성확률은 전체 염기서열 데이타들을 프로파일해서 각 위치에서의 해당 염기서열의 발생확률을 바로 앞 위치의 염기에 대한 조건부 확률로 계산하였다. 엄밀히 말하자면 염기서열 분석에서 언급되고 있는 가중치 배열 매트릭스(Weight array matrix)의 형태라 할 수 있다. 즉, 1차수이기 때문에 인접한 바로 앞의 염기와의 모든 가능한 조건부 확률을 계산해야 할 경우의 수는 16이므로, 최종적으로 16×26 매트릭스가 만들어지고 각각의 원소(element)들은 이러한 조건부 확률 값을 갖게 된다. 그리고 각 확률값을 Null 모델일 때의 확률값으로 나누고, 그 결과값에 자연대수를 취함으로써 log-odds로 전환하여 나중에 log-odds 값으로 계산하기 쉽게 하였다. 실제 임의의 염기서열이 입력으로 들어오게 되면 윈도우를 한 염기씩 옆으로 슬라이딩(sliding)하면서 log-odds 값을 계산하여 정해진 cut-off 값을 기준으로 'hit' 여부를 결정하면 된다.

프로모터 핵산 염기서열을 생성하기 위한 우리의 의존성 반영 분해모델(DRDM)은 본질적으로 다음에 기술된 분할 프로시저의 결정체라 할 수 있다.

(1) N_{-12}은 합쳐놓은 모든 프로모터 핵산 염기서열들의 1차 은닉 마르코프 모델로부터 생성된다.

(2a)만약 $N_{-12} \neq W$이면, 프로모터 핵산 염기서열내의 나머지 위치에 있는 핵산 염기를 생성하기 위해 부분집합 S_{-12}의 1차 은닉 마르코프 모델이 이용된다.

(2b)만약 $N_{-12} = W$이면, N_{-7}은 부분집합 W_{-12}를 위한 1차 은닉 마르코프 모델로부터 생성된다.

(3a)만일 $N_{-12} = W$이고 $N_{-7} \neq T$이면 부분집합 $W_{-12}V_{-7}$을 위한 1차 은닉 마르코프 모델이 사용된다.

(3b) $N_{-12} = W$이고 $N_{-7} = T$이면, N_{-11}은 $W_{-12}T_{-7}$을 위한 일차 은닉 마르코프 모델로부터 생성된다.

(4) 이와 같이 계속해서 전체 26 bp 크기의 핵산 염기서열이 생성되어질 때 까지 진행한다.

이와 같이, 우리가 제안하고 있는 의존성 반영 분해모델(DRDM)은 생체 내에서 고분자들의 유기적인 상호작용 및 연계 등 실제적인 생물학적 상황들을 잘 반영할

수 있도록 인접한 핵산들간의 상호 의존성뿐만 아니라 일정한 거리가 떨어져 있는 핵산들간의 상호 의존성도 실제로 잘 반영 할 수 있어 민감도와 선택도 측면에서 그 성능이 뛰어난 것으로 여겨진다.

5. 실험결과 및 향후 연구과제

DRDM의 성능을 검증하기 위해 앞에서 언급한 것과 같이 고정된 길이(-21~+4)를 갖는 실험적으로 입증된 607개의 프로모터 핵산 염기서열에 대해 10배 교차검증(10-fold Cross-Validation) 실험을 했다. 또한 단순한 가중치 매트릭스인 WMM(Weight Matrix Model)과 DRDM과 마찬가지로 각 위치별 의존성을 근거로 MDD 절차를 거쳐 생성되었으나, 이진트리의 말단노드 데이타에 대해 은닉 마르코프 모델을 적용하는 대신에 단순한 가중치 매트릭스 모델을 적용한 의존성 분해 가중치 매트릭스인 DDWMM(Dependence Decomposition Weight Matrix Model) [13]등과도 성능을 비교해 보았다. 이러한 성능을 비교하기 위해 DRDM에 대해 행한 10배 교차검증시와 같은 조건 하에서 WMM와 DDWM에 대해 10배 교차검증을 행했다. 표 2는 같은 조건 하에서 실험적으로 입증된 607개 프로모터 서열 데이타에 대해 행하여진 WMM, DDWMM 및 DRDM 등의 성능실험의 성공 백분율을 보여준다.

표 2 프로모터 서열 데이타에 대한 DRDM, DDWMM, 및WMM 등의 성능 비교

Testing No.	1	2	3	4	5	6	7	8	9	10	Average(%)
DRDM Sucesss(%)	98	97	99	99	98	98	99	97	99	99	98.3
DDWMM sucess(%)	85	75	85	79	84	82	84	82	89	78	82.3
WMM success(%)	59	66	52	46	46	56	49	49	67	50	54

표 2에 나타나 있는 바와 같이 실험 결과에 의하면 우리가 만든 DRDM이 아주 뛰어난 성능을 보였고, WMM 및 DDWMM과 비교해서도 상대적으로 월등히 낳은 성능을 나타냈다. 비록 아주 뛰어난 성능을 보였지만 그래도 몇 가지 개선의 여지를 찾아 볼 수 있다. 하나의 예로서, N_i와 N_j사이의 비교에 있어서 중요한 단점은 엄청나게 편향된 조성을 갖는 i번 j위치와 번 위치에 대해 사건 테이블의 예상값이 매우 작아지게 되면 (즉, 10미만) χ^2 검증의 신뢰성이 매우 떨어지게 된다. 이러한 문제를 극복하기 위해서는 i번 위치의 Consensus 핵산 염기와 N_j간의 비교가 하나의 선택할 수 있는 방법이 될 것이다. 왜냐하면 Consensus 핵산 염기와 N_j 사이의 비교에서는 보다 낮은 빈도수를 나타내는 핵산 염기들이 Consensus 핵산 염기로 합산되어 계산되므로 앞의 문제가 보다 더 완화되기 때문이다. 게다가 우리가 제안한 방법은 관련된 충분한 생물학적 데이타들을 고려하고 있지 않다. 그렇다고 해서 우리의 방법을 개선하기 위해서 생물학적인 실체에 대해 명백한 모델링을 반드시 포함시켜야 한다는 것을 의미하는 것은 아니다. 그것보다는 오히려 예측 방법을 고안할 때 어떤 데이타를 포함시켜야 할지와 무엇을 예측해야 할지를 결정할 때 생물학적인 지식을 고려하는 것이 매우 중요하다는 것을 의미한다. 전산기법에 의한 프로모터 예측은 매우 힘들고 복잡한 문제인데, 크게 두 가지 카테고리로 나누어 질 수 있다. 첫번째 카테고리는 정확한 전사 개시점을 예측하는 것이고, 두 번째는 프로모터가 존재할 법한 일반적인 영역을 예측하는 것이다. 후자의 경우는 비정렬된 DNA 핵산 염기서열에서의 패턴발견과 밀접한 관련이 있다. 본 논문에서는 전자의 경우를 다루었으나, 이들 두가지 문제를 각각 다루고 있는 서로 다른 알고리즘들을 결합하는 것이 아마도 많은 강점을 가질 것이다. 향후에 해볼만한 중요한 일들은 보다 미묘한 생물학적인 특징들을 잘 참작하고 반영할 수 있는 보다 융통성 있고 민감한 염기서열 분석 방법들을 개발하는 것이다. 실험적으로 입증된 Pre-mRNA Splicing, 전사 및 번역 등에 관여하는 핵산 서열 신호 데이타의 수가 보다 더 축적되어지면 그러한 신호의 복잡한 통계적 특성들을 연구할 수 있는 기회가 더 많이 주어질 것이고 결과적으로 앞에서 얘기한 미래의 일들이 가까운 미래에 이루어질 수 있을 것이다.

참 고 문 헌

[1] W. S. Reznikoff, D. A. Siegele, D. W. Cowing, and C. A. Gross, "The regulation of transcription initiation in bacteria", Annu. Rev Genet., Vol. 19, pp. 355-387, 1985.

[2] Harmen Bussemaker, Hao Li, and Eric Siggia, "Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis ", PNAS, Vol. 97, No. 18, pp. 10096-10100, 2000.

[3] Florea, L., Li, M., Riemer, C., Giardine, B., Miller, W., and Hardison, R., "Validating computer programs for functional genomics in gene regulatory regions", Current Genomics, Vol. 1, No. 1, pp. 11-27.

[4] K. Frech, K. Quandt, and T. Werner, "Software

for the analysis of DNA sequence elements of transcription", Comput. Appl. Biosci., Vol 13, pp. 89-97, 1997.
[5] H. Salgado, et. al., "RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12", Nucle. Acids Res., Vol. 29, pp. 72-74, 2001.
[6] J. Collado-Vides, "Grammatical model of the regulation of gene expression", Proc. Natl Acad. USA, Vol. 89, pp. 9405-9409, 1992.
[7] D. Thieffry, H. Salgado, A. M. Huerta and J. Collado-Vides, "Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichiacoli K*-12", Bioinformatics, Vol. 14, pp. 391-400, 1998.
[8] G. Z. Hertz, G. W. Hartzell III and G. D. Stormo, "Identification of consensus patterns in unaligned DNA sequences known to be functionally related", Comput. Applic. Biosci., Vol. 6, pp. 81-92, 1990.
[9] C. A. Gross and M. Lonetto, Bacterial sigma factors, Cold Spring Harbor Laboratory Press, 1992.
[10] T. D. Schneider and R. M. Stephens, "Sequence Logos: A New Way to Display Consensus Sequences", Nucleic Acids Res., Vol. 18, pp. 6097-6100, 1990.
[11] C. Burge and S. Karlin, "Prediction of Complete Gene Structure in Human Genomic DNA", Journal of Molecular Biology, Vol. 268, pp. 78-94, 1997.
[12] S. L. Salzberg, D. B. Searls and S. Kasif, "Computational Methods in Molecular Biology", Elsevier Science B. V., 1998.
[13] K. B. Kim, K. J. Park and E. B. Kong, "Prokaryotic Promoter Recognition with Dependence Decomposition Weight Matrix Method", 2nd International Symposium on Advanced Intelligent Systems Conference Proceedings, Vol. II No. 2 p277-281, 2001.

김 기 봉
1999년 5월~현재 ㈜스몰소프트 대표이사(실장/기술이사 역임) 1994년 3월~1999년 2월 한국과학기술연구원 생명공학연구소 연구원 1998년 3월~2001년 2월 충남대학교 컴퓨터공학과 박사수료 1995년 3월~1997년 2월 경북대학교 미생물학과 석사 1985년 3월~1992년 2월 경북대학교 미생물학과 졸업 관심분야는 생물정보학, 기계학습

박 기 정
2000. 3~현재 (주)스몰소프트 부설 정보기술연구소 소장/책임연구원 1998. 6 ~ 2000. 2 한국과학기술원 의과학센터연구원 1989. 2~1998. 6 한국과학기술연구원 생명공학연구소 (선임)연구원 1991. 3~2002. 2 한국과학기술원 생물과학과이학박사 1987. 3~1989. 2 한국과학기술원 생물과학과이학석사 1986. 3~1987. 2 한국과학기술원 전자계산학과 석사과정 1982. 3~1986. 2 서울대학교 공과대학 컴퓨터공학과 공학학사

공 은 배
1996년~현재충남대학교 컴퓨터공학과 정교수. 1995년 Oregon State Univ.전산학 박사. 1978년~1981년 서울대학교 계산통계학과 석사. 1974년~1978년 서울대학교 계산통계학과 졸업. 관심분야는 암호학, 기계학습, 생물정보학