한국산학기술학회논문지
Vol. 9, No. 5, pp. 1218-1224, 2008

프로모터 영역의 전사인자 결합부위 Consensus 패턴 탐색 방법

김기봉[1*]

Search Method for Consensus Pattern of Transcription Factor Binding Sites in Promoter Region

Ki-Bong Kim[1*]

요 약 유전자의 상위부분에 위치하면서 해당 유전자의 발현을 제어하는 신호부위 역할을 하는 프로모터 영역은 다양한 전사인자들이 결합하는 특정 신호부위들을 갖고 있다. 이러한 전사인자 결합부위들은 프로모터 영역 내의 매우 다양한 위치에 자리잡고 있으며, 진화론적으로 잘 보존된 Consensus 형태의 염기서열 패턴을 띠고 있다. 본 논문은 이러한 Consensus 패턴 탐색에 사용되는 Wataru 방법, EM 알고리즘, MEME 알고리즘, 유전자 알고리즘 및 Phylogenetic Footprinting 기법 등에 대해 소개하고, 향후 연구방향에 대한 전망을 제시하고자 한다.

Abstract Located on the upstream of a gene, the promoter region that plays a very important role in the control of gene expression as a signal part has various binding sites for transcription factors. These binding sites are present in various parts of the promoter region and assume an aspect of highly conserved consensus sequence pattern. This paper deals with the introductions of search methods for consensus pattern, including Wataru method, EM algorithm, MEME algorithm, Genetic algorithm and Phylogenetic Footprinting method, and intends to give future prospects of research on this field.

Key Words : Wataru Method, EM Algorithm, MEME Algorithm, Genetic Algorithm, Phylogenetic Footprinting

1. 서론

인간 유전체 프로젝트(Human Genome Project)의 종료가 2003년 4월 공식적으로 선언되었고, 다양한 생명체 종(species)들에 대한 유전체 프로젝트가 성공적으로 완료되었거나 진행 중에 있음에 따라 생물정보학 분야는 기본적으로 신규 유전자의 발굴 및 기능분석에 초점을 두고 있으며, 궁극적으로 단순히 개별 유전자의 기능이나 개별 단백질의 기능과 구조를 규명하는데 머무르지 않고 생체 내의 분자 네트워크를 총체적으로 밝히는데 초점을 맞추고 있다.

이러한 목표를 달성하기 위한 출발점으로 유전자의 기능 및 특성 파악에 초점을 두고 있다. 즉, 유전정보의 실체에 해당하는 일차원적인 DNA 염기서열 내에 암호화된 단백질의 기능과 구조를 예측하고자 하는 시도는 생물정보학 분야의 핵심 사안이었다. DNA 염기서열로부터 단백질의 기능을 예측한다는 것은, 구조 및 기능에 대해 이미 밝혀진 기존 단백질들을 대상으로 서열들 간의 상동성 검색(homology search)을 통해 해당 서열의 구조와 기능을 역으로 추론하고자 하는 것이다. 따라서 일차원적인 DNA 염기서열을 갖고 어떤 유전자인지 분석하는 연구가 활발히 진행되어 왔고, 이와 관련해서 다수의 알고리즘과 프로그램들이 개발되었다[1,2].

유전자가 발현되기 위해서는 유전정보의 실체인 DNA를 주형으로 유전암호 전달 매개체인 mRNA를 생성하는 전사과정(transcription)과 mRNA를 주형으로 유전암호를 해독하여 최종산물인 단백질을 생성하는 번역과정(translation) 등의 일련의 단계를 거친다. 전사는 유전자 발현과정의 첫 단계이자 전체 발현과정을 제어하는 중요한 역할을 한다. 전사를 제어하고 촉매하는 여러 효소와 전사조절 인자들이 존재하며, 그 중에서

[1]상명대학교 공과대학 생명정보공학과
*교신저자: 김기봉(kbkim@smu.ac.kr)
접수일 08년 06월 22일 수정일 08년 08월 31일 게재확정일 08년 10월 `16일

RNA 중합효소가 핵심적인 역할을 담당한다. 즉, RNA 중합효소가 프로모터(promoter)라 불려지는 특정 DNA 염기서열 영역을 인식하고 결합함으로써 전사가 개시된다. 이러한 결합에 이어서 DNA 이중나선의 일부분이 해리되고, RNA 중합효소는 상보적인 염기쌍 처리과정을 통해서 mRNA를 합성하기 시작한다. 프로모터의 염기서열 영역은 전사 시작점의 위치를 결정하며, 전사 시작점으로부터 종결지점까지가 하나의 전사단위가 된다. 실제로 RNA 중합효소 외에도 많은 전사인자들이 RNA 중합효소와 프로모터 영역에 작용하여 발현을 활성화하거나 억제한다. 비교적 단순한 구조를 띠는 원핵생물의 프로모터에 대한 연구는 많은 성과를 거두었으나, 진핵생물의 프로모터에 대한 연구는 본질적인 복잡한 구조 특성 때문에 상대적으로 미진한 편이다[1,2,3]. 그렇지만 연구의 필요성이 심각하게 대두되면서 최근 활발한 연구가 진행되고 있다. 프로모터 영역에는 RNA 중합효소 이외에 수많은 전사인자들의 결합 부위가 존재하는데 특히 전사 시작점에서 상위부분에 해당하는 5'방향으로 250 bp 까지의 서열상에 집중적으로 존재한다. 진핵생물의 전사인자 결합부위들 중에 대표적인 것이 TATA box, BRE, MTE, DPE, DCE 및 Initiator 등이 있다. 이러한 결합부위들은 특정 유전자 그룹에 따라서 다양한 위치에 다양한 종류의 결합부위들이 분포되어 있는 것으로 알려져 있다[3]. 즉, 어떤 기능을 수행하는 유전자군의 프로모터인지 혹은 어떤 특정 조직(tissue)에서 특이적으로 발현되는 유전자의 프로모터인지에 따라서 존재하는 결합부위들이 다르다는 것이다. 따라서 종(species)이나 조직별로 프로모터의 차이점을 탐지하는 것은 유전자 발현의 메커니즘을 이해하는 중요한 단서를 제공할 수 있다.

생물정보학 차원에서의 프로모터 연구의 중요성을 들자면 첫째, 특정 프로모터의 제어 하에 있는 해당 유전자가 어떤 단백질로 발현될 것인지에 대한 단서를 제공한다. 둘째, 프로모터 부위를 연구함으로써 유전자 발현이 어떻게 조절되는지 전체적인 조절 네트워크를 규명할 수 있는 근거를 찾을 수 있다. 셋째, 동시적 또는 계층적으로 작용하는 조절 네트워크에서 특정한 조절인자에 대응하는 유전자의 네트워크를 밝힐 수 있는 근거를 얻을 수 있다. 프로모터 영역의 연구가 유전자 예측 연구의 한 부분일 수도 있지만 위에서 언급한 바와 같은 중요성으로 인하여 독자적인 연구분야로서 자리매김하고 있다. 프로모터 영역을 분석하기 위해서는 전사인자들의 결합부위를 효율적으로 밝혀내는 것이 중요하다. 이러한 결합부위들은 각 종(species) 및 특정 유전자군별로 상이함을 띠고 있지만, 전체적으로 각 전사인자가 특이적으로 인식하고 결합할 수 있는 특정 염기서열들로 잘 보존되어 있다. 전사인자 결합부위 서열정보를 활용하여 프로모터 영역을 예측하고 분석하는 연구분야는 크게 3가지 범주로 나눠볼 수 있다. 첫째, 전사 개시부위(TSS : Transcription Start Site)와 더불어 전사인자 결합부위들을 탐색함으로써 궁극적으로 프로모터 영역을 예측하고자 하는 분야이다. 이 경우 TFD[4], TRANSFAC[5], RegulonDB[6] 등과 같은 공개된 전사인자관련 데이터베이스들을 활용하여 서열기반의 예측모델을 생성하고, 그러한 예측 모델로 전사인자 결합부위와 프로모터 영역을 예측한다. 둘째, 프로모터 서열 집단이 주어졌을 경우, 집단내의 개별 서열들이 갖고 있을 수 있는 모든 전사인자 결합부위들을 탐색하는 분야이다. 즉, 군집 서열 내에서 생물학적으로 의미 있는 패턴을 탐색하는 것이라 할 수 있다. 셋째, 프로모터 서열들을 다양한 클러스터링 알고리즘 (clustering algorithm)을 이용하여 기능별로 적합한 클러스터들을 생성하고 각 클러스터별 유전자 조절의 특성들을 밝히는 연구분야이다. 이 세가지 범주는 상호 연관성이 깊어 상호간에 유기적이고 긴밀하게 다뤄지기도 한다. 본 논문에서는 두 번째 범주에 해당되는 전사인자 결합부위 Consensus 패턴 탐색 기술에 대해 소재하고자 한다.

2. 문제 정의

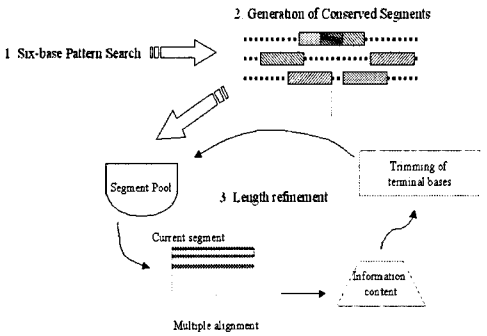
프로모터 영역 내에 존재하는 전사인자 결합부위와 프로모터를 Consensus 패턴 탐색 기술에 대입시키기 위해 문제 정의를 해보면 다음과 같다. 프로모터는 4종류의 뉴클레오티드로 이루어져 있고, 염기이름의 영문 머리글자인 A (adenine), C (cytosine), G (guanine), T (thymine) 등으로 각각의 뉴클레오티드를 표기한다. 즉, 프로모터는 A, C, G, T의 조합으로 이루어져 있고, 그 길이는 250 bp 정도이며 이러한 염기서열이 20~30개가 있다고 하자. 이러한 염기서열들은 길이가 6~10 bp인 전사인자 결합부위, 즉 생물학적으로 의미 있는 패턴들을 포함하고 있다. 그리고 각 패턴들은 모든 프로모터 염기서열에 공통적으로 반드시 포함될 필요는 없고, 각 프로모터 염기서열에 대해 이러한 패턴들이 하나도 존재하지 않을 수도 있고, 하나 이상 존재할 수도 있다. 프로모터 염기서열들의 집단에서 유의하게 나타나는 패턴들을 모두 찾아야 하며, 유의해야 할 점은 검색대상 패턴들이 Consensus형태를 갖는다는 것이다. 즉 정확히

일치하지는 않고 약간씩 다른 형태이지만 하나의 패턴을 나타내는 경우, 이들의 대표적인 Consensus 패턴을 찾아야 한다는 것이다. Consensus를 다루는 이유는 하나의 조절인자들이 다소 상이한 여러 인식부위들을 인지하고 결합하기 때문이다 (즉, degeneracy 특성을 갖고 있음). 용어의 간소화를 위해서 이하에서 언급하는 패턴은 Consensus를 의미한다.

2.1 Wataru 방법

Wataru와 Kanehisa 등이 제시한 패턴 탐색 방법[7,8]으로써 자율학습 문제를 해결하기 위해 여러 통계적 방법들을 사용하고 있다. 길이가 200 bp인 정렬되지 않은 프로모터 염기서열들의 집단으로부터 최적의 길이를 가진 패턴들을 찾기 위한 Wataru 방법의 전체적인 절차 및 구성은 [그림 1]에 나타난 바와 같다. 길이가 L인 N개의 서열로 이루어진 프로모터 서열집단으로부터 길이가 6인 모든 패턴들을 만들고, 각 패턴이 집단내의 k개의 프로모터 서열들에서 나타날 확률을 마르코프 연쇄와 이항분포를 이용하여 구한다. 이러한 확률이 f% 보다 큰 패턴들과 이들에 대해 s% 대체를 허용하는 패턴들을 생성하여 패턴 집단을 만든다. 그리고 프로모터 서열 집단 내의 각 서열에서 이들 패턴들의 위치를 파악한다. 다른 패턴들과 연이어 나타나는 경우에는 길이를 늘려 [그림 1]의 2번과 같이 보존된 단편들을 생성한다. 그 다음, 이들로부터 다중정렬과 정보량 분석을 통해 최적의 길이를 갖는 패턴을 구한다.

알고리즘이 전반적으로 뚜렷한 특징이 없으며, 표준적인 통계적 방법만이 이용되며 진행 절차가 다소 복잡하다. 그러나 초기값으로 패턴의 길이를 미리 정해주고 수행했던 과거의 방법들과는 달리 최적의 패턴을 얻기 위해 길이를 다시 고려한 점은 장점으로 간주할 수 있다.



[그림 1] Wataru 방법의 전반적인 절차 및 구성

2.2 EM (Expectation Maximization) 알고리즘

Cardon과 Stormo에 의해 지도학습 문제를 해결하는 수단으로서 EM 알고리즘이 전사인자 결합부위 예측에 사용되었다[9]. 입력서열 데이터로서 정렬되지 않은 프로모터 염기서열들의 집단을 사용하고 초기에 길이가 W인 임의의 패턴을 취하여 결국 모든 프로모터 염기서열들로부터 공유되는 최적의 패턴에 대한 확률적 모델을 반환한다. [그림 2]는 EM 알고리즘에 대한 기본적인 개요를 나타낸다.

```
EM (dataset, W) {
    choose starting point (ρ )
    do {
        reestimate z from ρ
        reestimate ρ from z
    } until (△ρ ⟨ε )
    return
}
```

[그림 2] EM 알고리즘의 개요

이 알고리즘에서는 행렬이 두 개 필요하다. 하나는 $\rho(=\rho_{lc})$로서 열 $c(1 \le c \le W)$의 위치에 문자가 l(=A, T, G 또는 C)인 확률을 나타내는 행렬이고, 또 하나는 $z(=z_{ij})$로서 i번째 서열의 j번째 위치가 패턴의 시작점으로 될 확률을 나타내는 행렬이다. 시작점으로서 패턴에 대한 를 임의로 선택한다. 그리고 나서 ρ로부터 베이지안 방법을 사용하여 z를 추정하는 과정, 즉 기대 과정과 다시 z로부터 가능도가 최대가 되도록 ρ를 다시 추정하는 과정, 즉 최대화 과정을 수렴에 이를 때까지 반복하여 최적화된 패턴을 얻는다. EM 알고리즘의 단점으로 첫째, 시작점을 어떻게 선택해야 하는지에 대한 규칙이 없기 때문에 선택 여부에 따라 최적의 패턴이 달라질 수 있는 국소최대에 빠질 우려가 있다. 둘째, One-occurrence-per-sequence 모델이기 때문에 주어진 집단의 프로모터 서열에서 나타나는 패턴이 여러 개 있을 경우에도 하나 밖에 찾지 못한다. 셋째, 또한 같은 이유로 패턴이 없는 프로모터 서열이 과대 추정되고 패턴이 여러 개 나타나는 프로모터 서열이 과소추정 되는 경우가 생긴다. 이러한 단점들을 극복한 방법이 다음에 소개할 MEME 알고리즘이다.

2.3 MEME (Multiple EM for Motif Elicitation) 알고리즘

EM 알고리즘을 확장한 것으로 자율학습 문제를 해결하고자 Bailey와 Elkan에 의해 사용된

알고리즘이다[10,11]. 입력 데이터로서 정렬이 되지 않은 서열들의 집단을 사용하고, 서열들로부터의 모든 부분서열들을 출발점으로 하여 최적의 모든 패턴들에 대한 확률적 모델을 반환한다. MEME 알고리즘이 EM 알고리즘의 세가지 단점을 극복한 방법으로 첫째, 서열에서 실제 발생하는 부분서열들을 시작점으로 선택하여 대역적인 최적의 패턴을 찾을 확률을 높였다[12]. 둘째, One-occurrence-per-sequence라는 가정을 배제한 N-occurrence-per-dataset 모델로서 하나 이상의 패턴이 존재하여도 모두 찾을 수가 있다. 셋째, 같은 이유로 하나의 서열에 여러 개의 패턴이 존재하더라도 문제가 되지 않고 또한 패턴이 없는 서열인 경우는 무시되므로 잡음(noise)에 민감하지 않다. [그림 3]은 MEME 알고리즘에 대한 개요를 나타낸다.

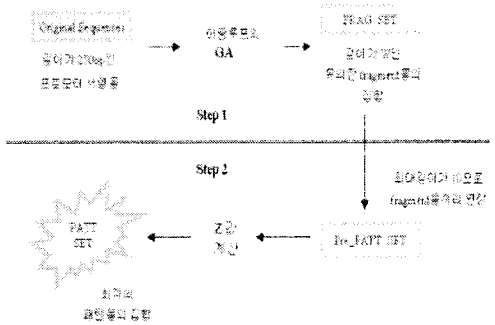
```
MEME (dataset, W, NSITES, PASSES) {
  for i=1 to PASSES {
    for each subsequence in dataset {
      run EM for 1 iteration with starting
        point derived from this subsequence
      choose model of shared motif
        with highest likelihood
      run EM to convergence from starting
        point which generated that model
      print converged model of shared motif
      erase appearances of shared motif
        from dataset
    }
  }
}
```

[그림 3] MEME 알고리즘의 개요

안쪽 루프는 EM을 기반으로 한 알고리즘을 선택된 시작점들에 따라 반복적으로 수행한다. 사용자 임의로 패턴들이 얼마나 나올 것인지 예상하여 NSITES를 결정하고 이 개수가 나올 때까지 계속하게 된다. 바깥 루프는 더 발견될 수 있는 패턴을 찾기 위해서 존재한다. 그리고 일단 발견된 패턴은 이후에 고려 대상에서 제외시켜 다른 패턴을 찾는데 잡음이 되지 않도록 한다. 단점으로는 패턴이 발견될 때까지 EM 알고리즘을 반복적으로 수행함으로써 시간이 많이 걸린다는 것이다. 그리고 패턴의 길이와 발견될 패턴의 수를 근사하게 추측하여 처음부터 정해주어야 한다는 것이다.

2.4 유전자 알고리즘

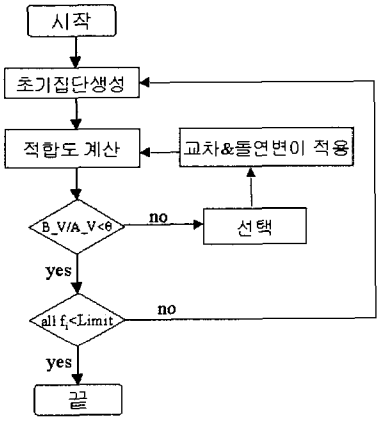
생물정보학 분야에서 염기서열들로부터 유의적으로 나타나는 패턴들을 탐색하는데 사용되는 최적화 알고리즘 중의 하나가 유전자 알고리즘이다[13,14,15]. 이는 생물진화의 원리를 모방한 생성 및 검증의 반복 절차에 의해 선택도태로서 결국 최적의 해를 찾는데 효율적이기 때문이다. 염기 서열상에서 나타나는 패턴들을 탐색하기 위해 단순한 문자열 일치 알고리즘을 이용하지 않은 이유는 앞에서 언급한 바와 같이 생물학적으로 의미 있는 패턴들이 Consensus 패턴을 띠고 있기 때문이다. 즉, 앞에서 언급한 것처럼 하나의 전사인자들이 인식하는 서열 패턴이 단일의 것이 아니라 여러 개의 상동성 패턴을 인식하기 때문이다. 따라서 패턴내의 위치별 핵산들의 비임의적인 특성을 고려하여 마르코프 연쇄라는 확률적 모델을 적합도 함수로 적용한다. 찾고자 하는 패턴은 모든 염기서열에 포함되지 않아도 된다는 가정 하에 크게 두 단계에 걸쳐서 패턴을 찾을 수 있을 것이다. 첫번째 단계는 유전자 알고리즘을 이용하여 유의적으로 나타나는 크기가 W인 단편들을 찾는 것이고, 두 번째 단계는 찾아진 단편들을 가지고 적당한 길이의 패턴들로 결정하는 것이다. 이러한 두 단계를 통한 구현방법의 전체적인 도식은 [그림 4]와 같다.



[그림 4] 유전자 알고리즘을 이용한 패턴 탐색의 전체적인 구현 절차

유전자 알고리즘이 적용될 집단은 길이가 W인 단편들로 집단의 크기인 M개 만큼 구성한다. 집단의 크기에 대해서는 특별한 제약이 없다. 단 너무 크면 시간이나 비용 면에서 비효율적이고, 너무 작으면 최적의 해를 구하기 힘들기 때문에 적당한 수로 정해주어야 한다. 기대되는 유의한 단편의 개수를 개략적으로 추정해서 M값을 정한다. 초기 집단의 각 개체들의 염기는 무작위로 생성한다. 그리고 이러한 각각의 개체들에 대한 적응도를 평가하기 위한 적합도 함수는 N개의 서열들로부터 마르코프 연쇄라는 확률적 모델과 포아송 분포를 이용한다. 적합도 값에 의해 다음 세대에

생존할 개체들을 선택하여 교차 연산을 수행한다. 그리고 지정한 변이율로 돌연변이를 수행하고 나면 다음 세대의 집단이 결정이 된다. 여기서 수렴 여부를 결정해서 수렴하지 않으면 적응도를 평가하는 것부터 반복적으로 다시 수행한다. 한 가지 더 고려해야 할 것은 이렇게 수렴에 이른 집단의 개체들로만 유의한 단편들을 결정하게 되면 그 외에 더 있을지도 모를 유의한 단편들을 놓칠 수가 있다. 그래서 루프를 바깥쪽으로 하나 더 두어 유의한 단편들이 나오지 않을 때까지 초기 집단을 구하는 단계부터 반복적으로 수행한다. 이러한 첫번째 단계에 대한 전체적인 구현 절차는 [그림 5]와 같다.



[그림 5] 이중루프의 유전자 알고리즘

A. For each gene on the input list:
  (a) Search homologous gene to identify the orthologous genes for this gene.
  (b) If this gene has orthologous for every species on the input set of species:
    I. For each orthologous gene:
      • Search the gene location on the genomic sequence of its species.
      • Search the promoter sequences for this gene.
    ii. Process the multiple alignment of the promoter sequences.
    iii. Extract the well-conserved patterns from the alignment.
B. Process the clustering of the patterns.
C. Compare to the known patterns and assign the function to the clusters.

[그림 6] Phylogenetic Footprinting 알고리즘의 개요

위의 단계를 거치고 나면 FRAG라는 집단에는 유의한 단편들이 모이게 된다. 그러나 이러한 단편들이 찾고자 하는 패턴이라고 말하지 못하는 이유는 W라는 길이로 고정시킨 핵산 염기서열들이기 때문이다. 따라서 이런 단편들은 실제 패턴의 일부분일 수가 있으므로 최적 길이의 패턴을 구할 방법이 있어야 한다. FRAG 집단에 모아진 단편들 중에는 서로 중첩되는 것들이 있다. 이는 이러한 단편들이 실제 패턴의 일부분이라는 이유가 될 것이다. FRAG 집단의 단편들끼리 연속적으로 5 bp씩 중첩시켰을 때 1 bp 정도 불일치하는 것을 허용하여 최대 길이가 10 bp이 되도록 가능한 모든 염기서열들을 모아 Pre_PATT 집단을 생성한다[그림 4]. 여기서 최대 길이는 Wataru의 실험결과에서 7 bp의 패턴이 최대 길이라는 점과 대부분의 전사인자 결합부위들이 6~10 bp이라는 점을 감안하여 정한다. Pre_PATT 집단에 있는 서열들의 길이를 고려한 적합도 함수에 의해 적합도 값을 구한다. 이때 Pre_PATT 집단의 i번째 염기서열을 한 염기씩 줄여가면서 W 이상의 모든 가능한 길이의 내부 염기서열들의 적합도 값을 구한다. 그리고 Limit값보다 큰 것들을 따로 모아서 이들 중에서 Z값이 가장 큰 염기서열을 i번째 염기서열에 대한 패턴으로 간주하고 PATT라는 집단에 중복이 되지 않게 저장을 한다. Z 값은 다음과 같이 계산한다.

$$Z = \frac{f_i - N \cdot p}{\sqrt{N \cdot p \cdot (1-p)} / \sqrt{N}} \qquad \text{(수식1)}$$

여기서 p는 길이가 L인 하나의 서열에서 i라는 개체가 적어도 한번 나타날 확률이고, f_i는 개체 i의 적합도를 나타낸다. Z값은 또한 PATT 집단에 있는 염기서열들의 최적의 패턴이 되는 순위의 기준이 된다.

2.5 Phylogenetic Footprinting 방법

수많은 생명체 종들(species)에 대한 유전체 프로젝트가 완성되면서 각광받고 있는 Phylogenetic Footprinting 기법[16,17,18]은 유전자들의 기능부위에 비해 서열 특이적 기능을 갖지 않는 다른 서열 영역 내에서 돌연변이가 발생하고 축적될 확률이 훨씬 높다는 사실을 기반으로 한다. 이러한 맥락에서 프로모터 내의 전사인자 결합부위들은 선택압력(selective pressure)을 받고 있으며, 이로 인해 기능을 띠지 않는 다른 지역의 서열들에 비해 훨씬 느린 속도로 진화하고 있다고 할 수 있다. 이러한 측면을 고려해보면, 전사조절 인자들은 상대적으로 가까운 유연관계를 갖는 서로 다른 종(species)들 간에 진화론적으로 잘 보존되어 있어야 한다. 이러한 사실을 기반으로 Phylogenetic Footprinting은 공통의 조상으로부터 유래된 서로 다른 종들의 유전자인 orthologous 유전자들의 프로모터 영역 내에서 잘 보존된 서열 패턴을 찾는 방법이다. 즉, 여러

종들의 orthologous non-coding 서열 집단으로부터 잘 보존된 영역을 탐색함으로써 전자인자 결합부위를 찾는 기법이다. 이를 위해서는 서로 다른 종들간의 유전체 비교를 위한 다중정렬(multiple alignment)이 필요하며, 게다가 탐색된 패턴들에 대한 그룹별 기능분류를 위해서 클러스터링(clustering) 과정이 요구된다. 다중서열 정렬시에는 로컬 정렬(local alignment) 알고리즘을 사용해야 국부적으로 잘 보존된 공통의 패턴을 탐색할 수 있다. [그림 6]은 Phylogenetic Footprinting 알고리즘에 대한 개요를 나타낸다. 진화 과정에서의 선택압력으로 말미암은 서열 유사성은 많은 생물정보학적 기법의 근간이 되고 있다.

3. 맺음말

본 논문에서 언급한 consensus 패턴 탐색기법들은 프로모터 및 전사인자 결합부위 예측뿐만 아니라 생물학적으로 의미 있는 다양한 신호부위 및 패턴 탐색 등에 적용될 수 있는 방법들이다. 각각의 기법들은 분석 대상 바이오 데이터의 특성과 문제 해결을 위한 전제 조건 등에 매우 민감하기 때문에 어느 알고리즘이 보다 더 성능이 뛰어난지 단언하기 매우 힘들다. 특히, 이러한 알고리즘은 휴리스틱(heuristic)하고 경험적인(empirical) 성향을 갖고 있어 사전에 결과를 예측한다든지, 혹은 결과에 대한 인과관계를 설명하는 것이 매우 힘든 경우가 많다. 그에 반해서 다양한 최적화의 여지가 많기 때문에 개발단계에서 여러 새로운 시도와 노력이 요구된다. 이러한 측면에서 최근의 연구동향을 보면 새로운 알고리즘 개발에 박차를 가하면서 동시에 기존의 방법들을 재구성 및 변형하여 성능향상을 기한다든지 아니면 서로 다른 기법들 간의 융합을 통해 새로운 돌파구를 모색하고 있는 추세이다[16]. 게다가 본문에서도 다루었지만, 다양한 유기체들에 대한 유전체 프로젝트의 결실로 서로 다른 종들 간의 유전체 비교분석을 통해 전사조절 인자 등을 찾고자 하는 Phylogenetic Footprinting 기법이 널리 사용되고 있는 추세이다[17,18]. 향후에는 실험적으로 검증된 양질의 데이터를 충분히 확보하여 특정 그룹간의 프로모터 영역의 패턴 특이성을 연구하는 분야, 즉, 특정 그룹에서 나타나는 패턴들을 사용하여 임의의 프로모터 염기서열이 어느 그룹에 속하는지 예측하고, 패턴 특이성과 유전자 기능과의 연관성을 찾는 연구 분야가 매우 각광받을 것으로 여겨진다.

4. 참고문헌

[1] http://www.nslij-genetics.org/gene/

[2] G. Yi, S. H. Sze and M. R. Thon, "Identifying Clusters of Functionally Related Genes in Genomes", *Bioinformatics* 23(9), pp. 1053-1060, 2007.

[3] Michael Q. Zhan, "Computational Analyses of Eukaryotic Promoters", *BMC Bioinformatics*, 8:S3, 2007.

[4] D. Ghosh, "Object-oriented Transcription Factors Dabase(ooTFD)", *Nucleic Acids Res.* 28(1), pp. 308-310, 2000.

[5] V. Matys, *et al.*, "TRNASFAC: Transcriptional Regulation, from Patterns to Profiles", *Nucleic Acids Res.* 31(1), pp. 374-378, 2003.

[6] S. Gama-Castro, *et al.*, "RegulonDB(version 6.0): Gene Regulation Model of Escherichia coli K-12 beyond Transcription, Active(Experimental) Annotated Promoters and Textpresso Navigation", *Nucleic Acids Res.*, D:120-124, 2008.

[7] Fujibuchi Wataru and Minoru Kanehisa, "Prediction of Gene Expression Specificity by Promoter Sequence Patterns", *DNA Research* 4, pp. 81-90, 1997.

[8] P. Horton and F. Wataru, "An Upper Bound on the Hardness of Exact Matrix Based Motif Discovery", *CPM*, pp.219-228, 2005.

[9] Lon R. Cardon and Gary D. Stormo, "Expectation Maximization Algorithm for Identifying Protein-binding Sites with Variable Lengths from Unaligned DNA Fragments", *Journal of Molecular Biology*, Vol. 223, pp. 159-170, 1992.

[10] Timothy Bailey and Charles Elkan, "Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization", *Machine learning Journal*, Vol. 21, pp. 51-83, 1995.

[11] T. Bailey, N. Williams, C. Misleh, and W. Li, "MEME: Discovering and Analyzing DNA and Protein Sequence Motifs", Nucleic Acids Research, 34:W369-W373, 2006.

[12] Jiang Liu, "A Combinatorial Approach for Motif Discovery in Unaligned DNA Sequences", *Thesis for the degree of master of mathematics*, Univ. of Waterloo, Canada, 2004.

[13] David Beasley, David R. Bull and Ralph R. Martin, "An Overview of Genetic Algorithms", *University Computing*, Vol. 15, No. 2, pp. 58-69, 1993.

[14] 김기봉, 공은배, "유전자 알고리즘을 이용한 프로모터 영역의 전사인자 결합부위 패턴 탐색", *정보과학회논문지(소프트웨어 및 응용)*, 제30권, 제5.6호, pp. 487-496,

2003.

[15] M. R. Berthold, H. J. Lenz, E. Bradley, R. Kruse and C. Borgelt, "Advances in Intelligent Data Analysis V", *Springer Press*, 2003.

[16] Wyeth W. Wasserman and ALbin Sandelin, "Applied Bioinformatics for the Identification of Regulatory Elements", *Nature Review Genetics*, pp. 276-287, 2004.

[17] Boris Lenhard, Albin Sandelin, Luis Mendoza, Par Engstrom, Niclas Jareborg and Wyeth W. Wasserman, "Identification of Conserved Regulatory Elements by Comparative Genome Analysis", *Journal of Biology*, Vol. 2, pp. 13, 2003.

[18] Alona Sosinsky, Barry Honing, Richard S. Mann and Andrea Califano, "Discovering Transcriptional Regulatory Regions in Drosophila by a Nonalignment Method for Phylogenetic Footprinting", *Proc. Natl. Acad. Sci.* U.S.A., 104(15), pp. 6305-6310, 2007.

김 기 봉(Ki-Bong Kim) **[정회원]**

- 1992년 2월: 경북대학교 미생물학과 (이학사)
- 1997년 2월: 경북대학교 미생물학과 (이학석사)
- 2003년 3월: 충남대학교 컴퓨터공학과(공학박사)
- 1999.04~2003.08 ㈜스몰소프트 대표이사/기술이사/연구소장
- 2003.09 - 현재 상명대학교 생명정보공학과 조교수

<관심분야>
바이오데이터 마이닝, 기계학습, 생체조절 네트워크, 유전정보 분석