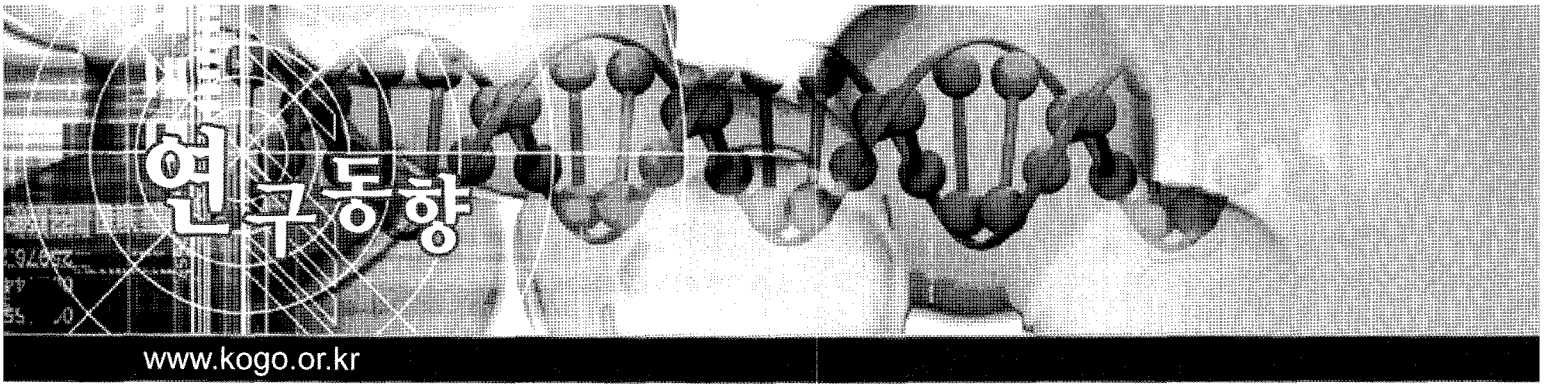
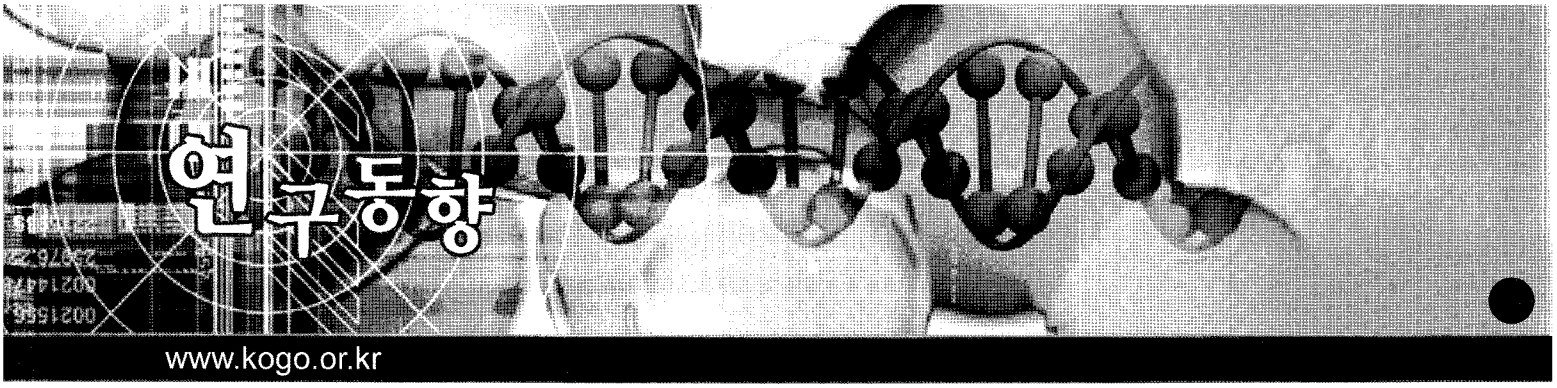


프로모터 영역의 전사인자 결합부위 Consensus 패턴 탐색 방법

김 기 봉(상명대학교 생명정보공학과)

근래의 생물정보학 분야는 인간유전체 프로젝트의 종결과 더불어 신규 유전자의 발굴 및 기능분석에 초점을 두고 있지만, 궁극적으로 단순히 개별 유전자의 기능을 밝히는 것이라기보다 생체내의 분자 네트워크에 대한 총체적인 이해를 목적으로 하고 있다. 이러한 목적을 달성하기 위한 시작점으로 유전자의 기능 및 특성 파악에 초점을 두고 있다. 즉, 일차적인 DNA염기서열로부터 해당 단백질의 기능을 예측하고자 하는 시도가 오래 전부터 핵심 사항이었다. 핵산 염기서열로부터 단백질의 기능을 예측한다는 것은, 구조 및 기능에 대해 이미 밝혀진 기존의 단백질들을 대상으로 서열상의 상동성 검색을 함으로써 해당 염기서열의 구조 및 기능을 역으로 알아내고자 하는 것이다. 따라서 일차적인 DNA염기서열을 가지고 어떤 유전자인지를 분석하는 연구가 많이 진행되어 왔고, 이와 관련해서 다수의 알고리즘 및 프로그램들이 개발되었다. 유전자들이 발현되기 위해서는 DNA를 주형으로 mRNA가 생성되는 전사과정과 mRNA를 주형으로 단백질이 만들어지는 번역과정을 거친다. 전사는 유전자 발현과정의 첫 단계이자 전체적인 유전자 발현 과정을 제어하는 중요한 역할을 한다. 전사가 일어나게끔 조절을 하고 또 그 과정에 관여하여 촉매 역할을 하는 여러 효소 및 전사인자들이 존재하는데, RNA 중합효소가 그 중의 하나로써 핵심적인 역할을 담당한다. 즉, RNA중합효소가 프로모터(Promoter)라 불리어지는 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 종결자까지가 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 그 자체의 복잡한 구조 때문에 상대적으로 많지 않다. 그렇지만 연구의 필요성이 심각하게 대두되면서 최근 활발한 연구가 진행되고 있다. 프로모터 영역에는 RNA 중합효소 이외에 수많은 전사인자들의 결합 부위가 존재하는데 특히 전사 시작점에서 5'방향으로 250 bp 까지의 서열상에 집중적으로 존재한다. 진핵생물의 전사인자 결합부위들 중에 대표적인 것이 TATA 상자 및 Initiator 등이 있다. 이러한 결합부위들은 특정 유전자 그룹에 따라서 다양한 위치에 다양한 결합부위들이 분포되어 있는 것으로 알려져 있다. 즉, 어떤 기능을 수행하는 유전자군의 프로모터인지 혹은 어떤 특정 조직 (Tissue)에서 특이적으로 발현되는 유전자들의 프로모터인지에 따라서 존재하는 결합부위들이 다르다는 것이다. 따라서 종(Species)이나 조직별로 프로모터의 차이점을 탐지하는 것은 유전자 발현의 메커니즘을 이해하는 단서를 제공할 수 있다.

생물정보학 차원에서의 프로모터 연구의 중요성을 들자면 첫째, 특정 프로모터의 제어 하에 있는 해당 유전자가 어떤 단백질로 발현될 것인지에 대한 단초를 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 유전자의 인식에 관한 연구의 한 부분일 수도 있지만 이 같은 중요성으로 인하여 독립적인 연구 대상이 되고 있다. 프로모터 영역을 분석하기 위해서는 전사인자의 결합부위들을 효율적으로 밝혀내는 것이 중요하다. 이러한 결합부위들은 각 종 및 특정 유전자군별로 상이함을 띠고 있지만, 전체적으로 각 전사인자들이 쉽게 인식하고 결합할 수 있도록 나름대로 특정 염기서열들로 잘 보존되어 있다. 전사인자 결합부위들을 규명함으로써 프로모터 영역을 분석하는 범주로는 첫번째, 결합부위들의 위치와 특히 전사 시작점을 탐색하고, 궁극적으로 프로모터 영역을 예측하는 연구방법들이 있다. 두 번째는 염기서열에 포함되어있는 결합부위들을 패턴으로 인식하여 밝혀내는 연구 방법들이 있으며, 세 번째는 프로모터들을 특정 기능과 연관된 그룹으로 분류하여 그룹별로 조절에 관한 특징들을 밝히는 연구 방법들이 있다. 이 세가지 범주는 서로 연관이 깊어 실제로 상호 긴밀하게 다루어지기도 한다. 여기서는 두 번째 범주에 해당되는 전사인자 결합부위 Consensus 패턴 탐색 기술에 대해 소재하고자 한다.

1. 문제 정의

프로모터 영역 내에 존재하는 전사인자 결합부위와 프로모터를 Consensus 패턴 탐색 기술에 대입시키기 위해 문제정의를 해보면 다음과 같다. 프로모터는 핵산 뉴클레오타이드의 각 염기인 A (염기 아데닌의 약어), T (염기 티민의 약어), G (염기 구아닌의 약어), 및 C (염기 시토신의 약어)의 조합으로 이루어져 있고, 그 크기가 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 이상이 존재할 수도 있다. 프로모터 염기서열들의 집단에서 유의하게 나타나는 패턴들을 모두 찾는데 주의해야 할 점은 찾아야 할 패턴들이 Consensus 형태를 갖는다는 것이다. 즉 정확히 일치하지는 않고 약간씩 다른 형태이지만 하나의 패턴을 나타내는 경우, 이들의 대표적인 Consensus패턴을 찾아야 한다는 것이다. Consensus를 다루는 이유는 하나의 조절인자들이 다소 상이한 여러 인식부위들을 인지하고 결합하기 때문이다 (즉, degenerancy 특성을 갖고 있다). 용어의 간략화를 위해서 이하에서 언급하는 패턴은 Consensus를 의미한다.

2. Consensus 패턴 탐색 방법들

정렬되지 않은 서열들의 집단으로부터 각 서열에 존재하면서 생물학적으로 의미가 있는 패턴들을 찾고자 하는 연구들이 많이 수행되었다. 그 중에서 프로파일을 이용한 탐색 방법인 Wataru방법, EM 알고리즘, MEME 알고리즘 및 확률적 모델을 적합도 함수로써 적용한 최적화 알고리즘으로 알려진 유전자 알고리즘 등에 대해 언급하고자 한다. 이들은 다음 가정에 의해 두 가지로 나누어 생각할 수 있다. 첫째, One-occurrence-per-sequence 모델로서 집단의 각 서열은 공유하는 패턴을 반드시 포함하고 있다고 가정하는 지도학습 도구와 둘째, N-occurrence-per-dataset 모델로서 집단의 각 서열은 공유하는 패턴을 전혀 포함하지 않을 수 도 있고, 또한 하나 이상 포함할 수 있다고 가정하는 자율학습 도구가 그것이다. 자율학습 도구는 가능한 최적의 패턴들의 공간이 지도학습

도구의 것보다 훨씬 크기 때문에 탐색에 어려움이 있지만 실제로 더욱 유용한 도구이다. EM 알고리즘은 첫번째 모델에 해당되고, 나머지는 두 번째 모델에 해당된다.

2.1 Wataru 방법

Wataru와 Kanehisa 등이 제시한 패턴 탐색 방법으로써 자율학습 문제를 해결하기 위해 여러 통계적 방법들을 사용하고 있다. 길이가 200 bp인 정렬되지 않은 프로모터 염기서열들의 집단으로부터 최적의 길이를 가진 패턴들을 찾기 위한 전체적인 방법은 [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 % 보다도 큰 패턴과 이 패턴과 s % 대체를 허용하는 패턴들을 모아서 이들 패턴들의 원래 프로모터 서열상의 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.

2.2 EM (Expectation Maximization) 알고리즘

Cardon과 Stormo에 의해 지도학습 문제를 해결하는 수단으로서 EM 알고리즘이 전사인자 결합부위 예측에 사용되었다. 입력데이터로서 정렬이 되지 않은 프로모터 염기서열들의 집단을 사용하고 초기에 길이가 W인 임의의 패턴을 취하여 결국 모든 프로모터 염기서열들로부터 공유되는 최적의 패턴에 대한 확률적 모델을 반환한다. [그림 2]에 EM 알고리즘에 대한 개요가 나타나있다.

이 알고리즘에서는 행렬이 두개 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 시작점

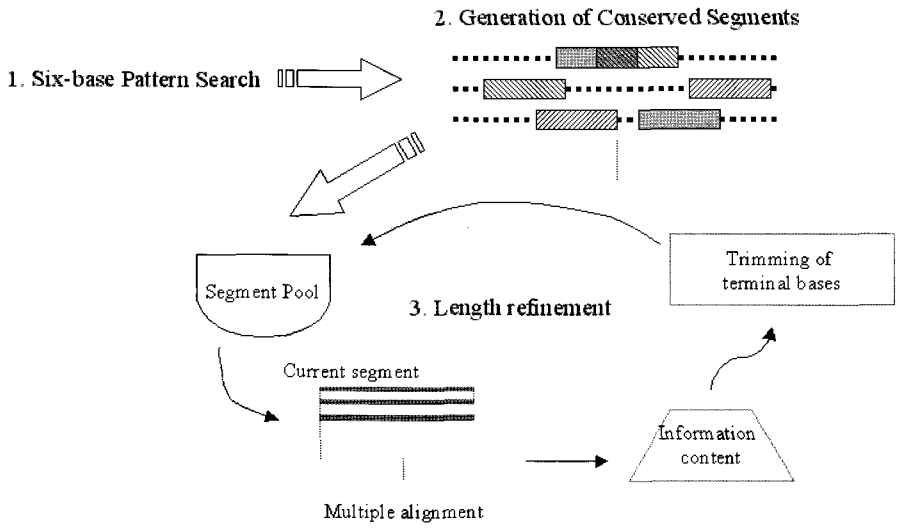


그림 1 전반적인 Wataru 방법의 절차 및 구성

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

그림 2 EM 알고리즘의 개요

으로 될 확률을 나타내는 행렬이다. 시작점으로서 패턴에 대한 ρ를 임의로 선택한다. 그리고서 이 ρ로부터 베이지안 방법을 사용하여 z를 추정하는 과정, 즉 기대 과정과 다시 이 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 염려가 있다. 둘째, One-occurrence-per-sequence 모델이기 때문에 주어진 집단의 프로모터 서열에서 나타나는 패턴이 여러 개 있을 경우에도 하나 밖에 찾지를 못한다. 셋째, 또한 같은 이유로 패턴이 없는 프로모터 서열이 과대추정 되고 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. 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. 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 나타내고 있다.

```
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 진행된다. 사용자 임의로 패턴들이 얼마나 나올 것인지 예상하여 *NSITES*를 결정하고 이 개수가 나올 때까지 계속하게 된다. 바깥 루프는 더 발견될 수 있는 패턴을 찾기 위해서 존재한다. 그리고 일단 발견된 패턴은 이후에 고려대상에서 제외시켜 다른 패턴을 찾는데 잡음이 되지 않도록 한다. 단점으로는 패턴이 발견될 때까지 EM 알고리즘을 반복적으로 수행함으로써 시간이 많이 걸린다는 것이다. 그리고 패턴의 길이와 발견될 패턴

의 수를 근사하게 추측하여 처음부터 정해주어야 한다는 것이다.

2. 4 유전자 알고리즘을 이용한 Consensus 패턴 탐색

서열들로부터 유의적으로 나타나는 패턴들을 탐색하기 위해 최적화 알고리즘의 하나인 유전자 알고리즘을 이용한다. 이는 생물진화의 원리를 모방한 생성 및 검증의 반복 절차에 의해 선택도태로서 결국 최적의 해를 찾는데 효율적이기 때문이다. 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 함수로서 적용한다. 찾고자 하는 패턴은 모든 염기서열에 포함되지 않아도 된다는 가정 하에 크게 두 단계에 걸쳐서 패턴을 찾을 수 있을 것이다. 첫번째 단계는 유전자 알고리즘 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]과 같다.

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 변이율로써 돌연변이를 수행하고 나면 다음 세대의 집단이 결정이 된다. 여기서 수렴 여부를 결정해서 수렴하지 않으면 적응도를 평가

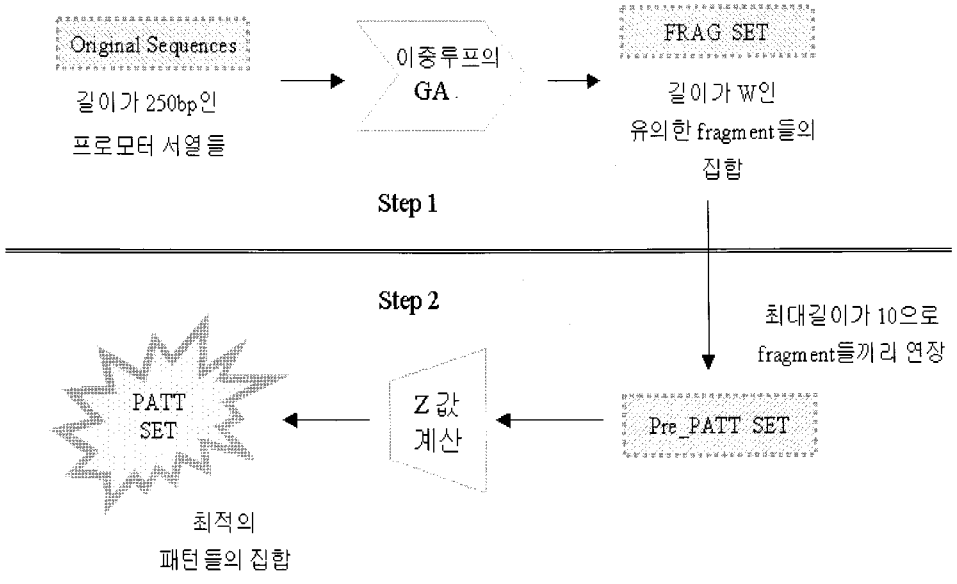


그림 4 유전자 알고리즘을 이용한 패턴 탐색의 전체적인 구현 절차

하는 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]에 나타나 있다.

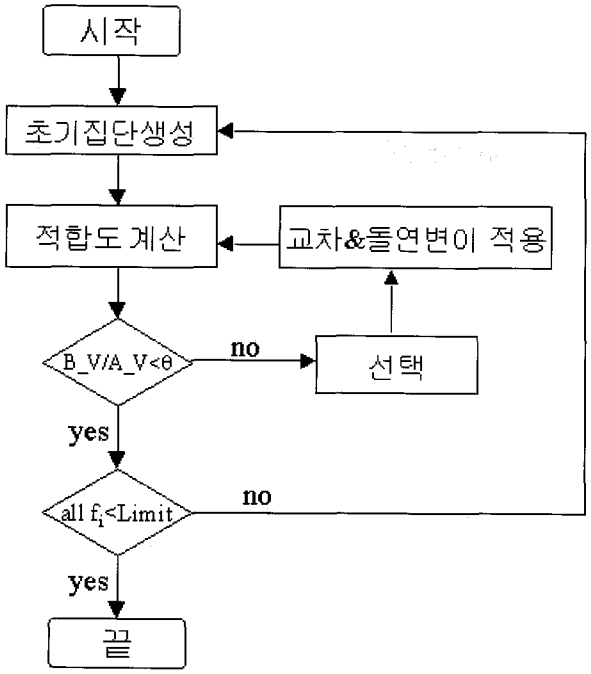


그림 5 이중루프의 유전자 알고리즘

위의 단계를 거치고 나면 *FRAG*라는 집단에는 유의한 단편들이 모이게 된다. 그러나 이러한 단편들이 찾고자 하는 패턴이라고 말하지 못하는 이유는 *W*라는 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 *i*번째 염기서열을 한 염기씩 줄여가면서 *W*이상의 모든 가능한 길이의 내부 염기서열들의 적합도 값을 구한다. 그리고 *Limit*값보다 큰 것들을 따로 모아서 이중 *Z*값이 가장 큰 염기서열을 *i*번째 염기서열에 대한 패턴으로 간주하고 *PATT*라는 집단에 중복이 되지 않게 저장을 한다. *Z* 값은 다음과 같이 계산한다.

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

여기서 *Z*값은 또한 *PATT* 집단에 있는 염기서열들의 최적의 패턴이 되는 순위의 기준이 된다.

맺음말

여기서 언급한 Consensus 패턴 탐색기법들은 단지 프로모터와 전사인자 결합부위에만 국한되는 문제가 아니라 이 외에도 생물학적으로 의미가 있는 다양한 신호부위 및 패턴 등에 적용될 수 있는 기술이다. 그리고 무조건 복잡하고 어려운 알고리즘만이 고급스럽고 좋다고 생각하는 것은 금물이다. 무엇보다도 분석 대상 데이터의 특성에 적합한 알고리즘을 적용하는 것이 중요하며 이러한 계열의 알고리즘들은 매우 선험적인 경향이 있어 사전에 그 결과를 예측하기 매우 힘들다는 사실을 명심할 필요가 있다. 게다가 한편으론 다양한 최적화의 여지가 많기 때문에 개발단

계에서 여러 새로운 시도와 노력이 요구된다. 이러한 측면에서 최근의 연구동향을 보면 새로운 알고리즘 개발에 박차를 가하면서 동시에 기존의 방법들을 재구성 및 변형하여 성능향상을기한다든지 아니면 이종의 기법간의 융합을 통해 새로운 돌파구를 모색하고 있는 추세이다. 게다가 신빙성 있는 데이터를 충분히 확보하여 특정 그룹간 유전자들의 프로모터 영역의 패턴 특이성에 관한 연구, 즉 특정 그룹에서 나타나는 패턴들을 사용하여 임의의 프로모터 염기서열이 어느 그룹에 속하는지 예측하는 내용의 연구라든지 그룹간 차이를 체계적이고 통계적인 방법으로 분석하는 연구 등이 향후에 필요할 것으로 본다.

참고문헌

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

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

S. Sinha and M. Tompa, "A statistical method for finding transcription factor binding sites" 2000, 8:344-354.

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.

Tim Bailey and William E. Hart, "Learning Consensus Patterns in Unaligned DNA Sequences Using a Genetic Algorithm", Sandia Laboratories Tech Report SAND95-2293.

Pesole G., Prunella N., Liuni S., Attimonelli M., and Saccone C., "WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences", Nucleic Acids Research, Vol. 20, pp. 2871-2875, 1992.

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

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

Cavin Perier, R., Junier, T., Bonnard, C. and Bucher, P. "The Eukaryotic Promoter Database EPD: Recent Developments", Nucleic Acids Research, Vol. 27, pp. 307-309, 1999.

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.