유전자 알고리즘을 이용한 프로모터 영역의 전사인자 결합부위 패턴 탐색

(Pattern Search for Transcription Factor Binding Sites in a Promoter Region using Genetic Algorithm)

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

요 약 유전자 발현에 매우 중요한 신호역할을 하는 프로모터 영역은 여러 전사인자들이 결합하는 특정 부위들을 갖고 있다. 전사인자의 결합부위는 프로모터의 다양한 부위에 위치하며, 진화론적으로 잘 보존된 Consensus 형태의 염기서열 패턴을 띠고 있다. 본 논문은 이러한 최적의 패턴들을 탐색하기 위해 유전자 알고리즘을 기반으로 하면서, 동시에 MEME 알고리즘의 N-occurrence-per-dataset 모델의 가정과 패턴의 길이를 결정할 수 있는 Wataru 방법의 장점을 따르는 새로운 방법을 제시하고 있다. 이러한 탐색 방법은 유전체 연구자들이 임의의 DNA 염기서열 상에서 프로모터 영역을 예측하거나 특정 전사인자의 결합부위를 탐색하는데 적극 활용할 수 있다.

키워드 : 유전자, 프로모터, 전사인자, 유전체, MEME 알고리즘, Wataru 방법, 유전자 알고리즘

Abstract The promoter that plays a very important role in gene expression as a signal part has various binding sites for transcription factors. These binding sites are located on various parts in promoter region and have highly conserved consensus sequence patterns. This paper presents a new method for the consensus pattern search in promoter regions using genetic algorithm, which adopts the assumption of N-occurrence-per-dataset model of MEME algorithm and employs the advantage of Wataru method in determining the pattern length. Our method will be employed by genome researchers who try to predict the promoter region on anonymous DNA sequence and to find out the binding site for a specific transcription factor.

Key words : Promoter, Gene Expression, Transcription Factor, Consensus, Genetic Algorithm, MEME Algorithm, Wataru Method, Genome

1. 서 론

1990년대 초반부터 본격적으로 개시된 인간유전체 프로젝트를 비롯한 각종 유전체 프로젝트들의 결과로 각 생명체의 방대한 유전 정보들이 엄청나게 쏟아지고 있다. 이러한 방대한 유전 정보들을 체계적이고 효율적으로 분석함으로써 보다 더 많은 생물학적 지식을 얻고자 국내뿐만 아니라 전 세계적으로 생물정보학 분야에 많은 투자를 하고 있다. 현재 생물정보학의 근간은 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 DNA(Deoxyribonucleic Acid) 염기서열로부터 해당 단백질의 기능을 예측하고자 하는 시도가 핵심 사항이다. 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 진행되고 있고, 이와 관련해서 다수의 알고리즘[1,2,3] 및 프

† 학생회원 : (주) 스몰포스트 정보기술연구소
kbkim@bioinfo.smallsoft.co.kr
†† 종신회원 : 충남대학교 컴퓨터공학과 교수
keb@ce.cnu.ac.kr
논문접수 : 2002년 9월 9일
심사완료 : 2003년 2월 7일

로그램들이 개발되었다.

유전자들이 발현되기 위해서는 DNA를 주형으로 mRNA (messenger ribonucleic acid)가 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 붕괴되고, 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[4-6], 진핵생물의 프로모터에 대한 연구[7-10]는 그 자체의 복잡한 구조 때문에 상대적으로 많지 않다. 그렇지만 연구의 필요성이 심각하게 대두되면서 최근 활발한 연구가 진행되고 있다. [그림 1]의 프로모터 영역에는 RNA 중합효소 이외에 수많은 전사인자들의 결합 부위가 존재하는데 특히 전사 시작점에서 5′ 방향으로 250 bp 까지의 서열상에 집중적으로 있다[8]. 진핵생물의 전사인자 결합부위들 중에 대표적인 것이 TATA 상자 및 Initiator 등이 있다. 이러한 결합부위들은 특정 유전자 그룹에 따라서 다르게 분포되어 있다고 알려져 있다. 즉, 어떤 기능을 하는 유전자들의 프로모터라든지 특정 조직(Tissue)에서 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.

프로모터 영역을 연구해야 하는 중요성[9]은 첫째, 유전자의 신호역할을 하는 프로모터가 어떤 단백질로 발현하는 유전자인지에 대해 확인할 수 있는 근거를 제시한다. 둘째, 프로모터 부위를 연구함으로써 유전자의 발현을 어떻게 조절하는지 전체적인 조절 네트워크를 밝힐 수 있는 근거를, 제시한다. 셋째, 동시적 또는 계층적으로 작용하는 조절 네트워크에서 특정한 조절인자에 대응하는 유전자의 네트워크를 밝힐 수 있는 근거를 제시한다. 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 연구들이 있다[8]. 두 번째는 염기서열에 포함되어있는 결합부위들을 패턴으로 인식하여 밝혀내는 연구들[5, 6, 11]이 있으며, 세 번째는 프로모터들을 특정 기능과 연관된 그룹으로 분류하여 그룹별로 조절에 관한 특징들을 밝히는 연구들[7]이 있다. 이 세가지 범주는 서로 연관이 깊어 실제로 상호 긴밀하게 다루어지기도 한다. 본 논문은 두 번째 범주를 다루고 있으며, 전사인자의 결합부위에 대한 최적의 패턴들을 탐색하기 위해서 유전자 알고리즘을 기반으로 하는 새로운 방법을 제시하고 있다. 특히 우리가 제안하는 방법은 기존의 MEME(Multiple EM for Motif Elicitation) 알고리즘[12]의 n-occurrence-per-dataset 모델의 가정을 수용함과 동시에 패턴 길이를 결정할 수 있는 Wataru 방법[7]의 장점을 따르고 있다.

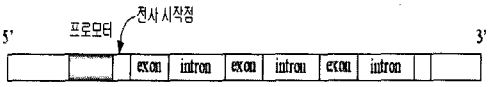

그림 1 진핵생물 유전자의 일반적인 형태

본 논문의 전체 구성을 살펴보면, 2장에서는 본 논문에서 다루게 될 문제에 대한 정의를 다루고, 3장에서는 기존의 생물학적인 consensus 패턴을 탐색하기 위해 사용되는 방법들을 살펴본다. 4장에서는 본 논문이 제시하는 방안인 유전자 알고리즘을 이용한 consensus 패턴 탐색 방법에 대해 논하며, 5장에서는 실험 및 실험결과에 대해서 언급하고, 마지막으로 6장에서는 결론 및 향후 연구계획에 대해서 논한다.

2. 문제 정의

핵산 뉴클레오티드의 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 정도인 염기서열이 20~30개가 있다고

하자. 이러한 염기서열들은 길이가 6~10 bp인 패턴들을 포함한다. 그리고 각 패턴들은 모든 염기서열에 공통적으로 반드시 포함될 필요는 없고, 각 염기서열에 대해 이러한 패턴들이 하나도 존재하지 않을 수도 있고, 하나 이상이 존재할 수도 있다. 염기서열들의 집단에서 유의하게 나타나는 패턴들을 모두 찾는데 주의해야 할 것은 찾아야 할 패턴들이 consensus 형태를 갖는다는 것이다. 즉, 정확히 일치하지는 않고 약간씩 다른 형태이지만 하나의 패턴을 나타내는 경우, 이들의 대표적인 consensus 패턴을 찾아야 한다(그림 2). consensus를 다루는 이유는 하나의 조절인자들이 다소 상이한 여러 인식부위들을 인지하고 결합하기 때문이다. 용어의 간략화를 위해서 이하에서 언급하는 패턴은 consensus를 의미한다.

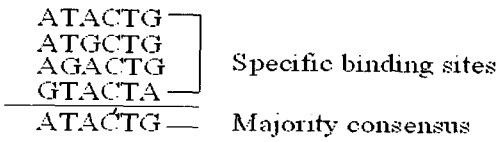

그림 2 Consensus한 부위

본 논문은 위에서 정의한 문제를 다루고 있다. 단순한 문자열 일치 알고리즘으로 찾을 수 없는 consensus 패턴을 찾기 위해 마르코프 연쇄라는 확률적 모델을 적합도 함수로서 적용한 최적화 알고리즘으로 알려진 유전자 알고리즘 [13-16]을 사용한다. 실험 대상 데이타로서 인간의 하우스키핑(Housekeeping) 유전자 프로모터 영역의 염기서열들과 인간의 간(Liver) 조직에서 발현되는 유전자 프로모터 영역의 염기서열들을 EPD(Eukaryotic Promoter Database) 프로모터 데이타베이스[17]와 EMBL (European Molecular Biology Laboratory) 핵산서열 데이타베이스로부터 추출하여 사용한다. 각 염기서열은 전사 시작점으로부터 앞부분의 250 bp 염기서열들이다. 참고로 하우스키핑 유전자는 세포의 기초적인 대사에 필요한 유전자들을 말한다. 결론적으로, 실험 결과로서 얻어진 consensus 패턴들과 전사인자 데이타베이스인 TFD(Transcription Factor Database) [18]에 있는 기존에 알려진 전사인자 결합부위들과의 연관성을 분석한다.

3. 기존의 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 Wataru 방법[7], EM 알고리즘[6] 및 MEME 알고리즘[12] 등에 대해 간단히 소개하고자 한다. 이들은 다음 가정[12]에 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, 나머지 두 방법은 두 번째 모델에 해당된다.

3.1 Wataru 방법

Wataru와 Kanehisa[7] 등이 제시한 패턴 탐색 방법으로써 자율학습 문제를 해결하기 위해 여러 통계적 방법들을 사용하고 있다. 길이가 200 bp인 정렬되지 않은 염기서열들의 집단으로부터 최적의 길이를 가진 패턴들을 찾기 위한 전체적인 방법은 [그림 3]에 나타나 있다. 길이가 L인 N개의 서열로 이루어진 서열집단으로부터 길이가 6인 모든 패턴들을 만들고, 각 패턴이 집단내의 k개의 서열들에서 나타날 확률을 마르코프 연쇄[5]와 이항분포를 이용하여 구한다. 이러한 확률이 f % 보다도 큰 패턴과 이 패턴과 s % 대체를 허용하는 패턴들을 모아서 이들 패턴들의 원래 서열상의 위치를 파악한다. 다른 패턴들과 연이어 나타나는 경우에는 길이를 늘려 [그림 3] 의 2번과 같이 보존된 단편들을 생성한다. 그 다음, 이들로부터 다중정렬과 정보량 [11, 19]의 분석을 사용하여 최적의 길이를 갖는 패턴을 얻는다.

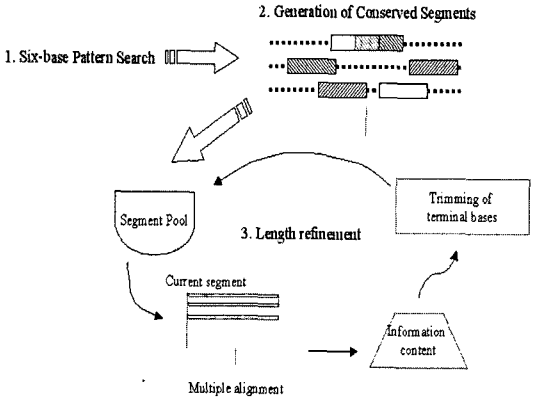

그림 3 Wataru에 의한 방법

알고리즘이 전반적으로 큰 특징이 없으며, 표준적인 통계적 방법만을 사용하였고 진행되는 절차가 복잡하다. 그러나 초기값으로 패턴의 길이를 미리 정해주고 하던 기존의 방법들과는 달리 최적의 패턴을 얻기 위해 길이를 다시 고려한 점은 본 연구에서 장점으로 받아들였다.

3.2 EM(Expectation Maximization) 알고리즘

Cardon과 Stormo[6]에 의해 지도학습 문제를 해결하는 수단으로서 EM 알고리즘이 사용되었다. 입력데이타로서 정렬이 되지 않은 염기서열들의 집단을 사용하고 초기에 길이가 W인 임의의 패턴을 취하여 결국 모든 염기서열들로부터 공유되는 최적의 패턴에 대한 확률적 모델을 반환한다. [그림 4]에 EM 알고리즘에 대한 개요가 나타나 있다.

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

그림 4 EM 알고리즘

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 시작점으로 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 재추정하는 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 서열에서 나타나는 패턴이 여러 개 있을 경우에도 하나 밖에 찾지를 못한다. 셋째, 또한 같은 이유로 패턴이 없는 서열이 과대추정 되고 패턴이 여러 개 나타나는 서열이 과소추정 되는 경우가 생긴다. 이러한 단점들을 극복한 방법이 다음에 소개할 MEME 알고리즘이다.

3.3 MEME(Multiple EM for Motif Elicitation) 알고리즘

EM 알고리즘을 확장한 것으로 자율학습 문제를 해결하고자 Bailey와 Elkan[12]에 의해 사용된 알고리즘이다. 입력 데이타로서 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 잡음에 민감하지 않다. [그림 5]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 5 MEME 알고리즘

안쪽 루프는 EM을 기반으로 한 알고리즘을 선택된 시작점들에 따라 반복적으로 진행이 된다. 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 두었다. 그리고 일단 발견된 패턴은 이후에 고려 대상에서 제외시켜 다른 패턴을 찾는데 잡음이 되지 않도록 하였다. 단점으로는 패턴이 발견될 때까지 EM 알고리즘을 반복적으로 수행함으로써 시간이 많이 걸린다는 것이다. 그리고 패턴의 길이와 발견될 패턴의 수를 근사하게 추측하여 처음부터 정해주어야 한다는 것이다.

4. 유전자 알고리즘을 이용한 consensus 패턴 탐색

본 연구에서는 앞에서 언급한 문제의 대상이 되는 핵산 염기서열들로부터 유의적으로 나타나는 패턴들을 탐색하기 위해 최적화 알고리즘의 하나인 유전자 알고리즘[13-16]을 이용했다. 이는 생물진화의 원리를 모방한 생성 및 검증의 반복 절차에 의해 선택도태로서 결국 최적의 해를 찾는데 효율적이기 때문이다. 염기서열상에서 나타나는 패턴들을 탐색하기 위해 단순한 문자열 일치 알고리즘을 이용하지 않은 이유는 생물학적으로 의미 있는 패턴들이 consensus 패턴을 띠고 있기 때문이다. 즉, 앞에서 언급한 것처럼 하나의 전사인자들이 인식하는 염기서열 패턴이 단일의 것이 아니라 여러 개의 상동성 패턴을 인식하기 때문이다. 따라서 패턴내의 위치별 염기들이 비임의의 특성을 고려하여 마르코프 연쇄라는 확률적 모델을 적합도 함수로서 적용하였다. 찾고자 하는 패턴은 모든 염기서열에 포함되지 않아도 된다는 가정 하에 크게 두 단계에 걸쳐서 패턴을 찾았다. 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 6]과 같다.

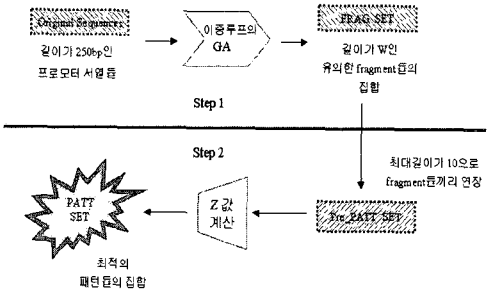

그림 6 전체적인 구현 절차

4.1 유의한 단편들을 찾는 단계(Step 1)

한 집단에 길이가 L이고 개수가 N개인 서열들이 있다고 하자. 그리고 이들 각 서열들은 A, T, G, 및 C 등의 염기로 이루어져 있다. 전사인자 결합부위의 길이가 보통 6~10 bp정도로 알려져 있고, 나중에 패턴의 길이에 대한 절차가 있기 때문에 윈도우 크기인 W를 6으로 하였다. 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. 본 연구에서는 기대되는 유의한 단편의 개수로 30~50으로 보고 M을 50으로 하였다. 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 이용하였다. 적합도 값에 의해 다음 세대에 생존할 개체들을 선택하여 변형된 2점 교차를 수행한다. 그리고 지정한 변이율로써 돌연변이를 수행하고 나면 다음 세대의 집단이 결정된다. 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 7]에 나타나 있다.

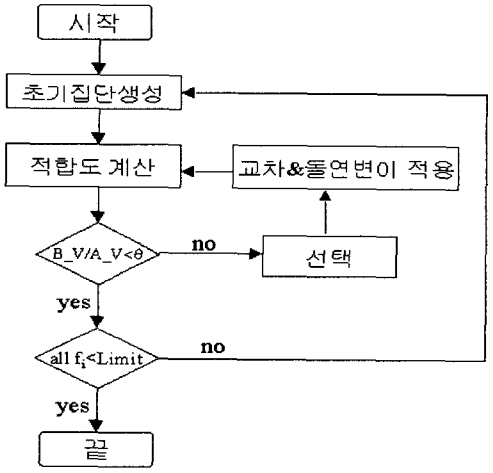

그림 7 이중루프의 유전자 알고리즘

4.1.1 적합도 함수

적합도 함수는 집단에 있는 개체들의 적응도를 평가하기 위한 수단이다. 각 개체의 적합도는 다음 세대에 생존할 개체들을 선택하는 기준이 되고 또한 최고값과 평균값의 비를 이용하여 수렴성을 측정할 수 있는 기준을 제시한다. 이러한 적합도 함수는 서열내의 염기들이 비임의적인 특성을 고려하여 확률적인 모델인 1차 마르코프 연쇄 모델[5]과 포아송 분포를 이용하여 나타내었다.

다음에 나오는 P_i는 i번째 개체, 즉 서열이 $X_1X_2\cdots X_W$인 개체가, 길이가 L인 N개의 원래의 서열들로부터 나올 수 있는 확률이다. 상호 인접한 두 염기간의 의존성을 반영하는 조건부 확률의 곱으로 표현되는 1차 마르코프 연쇄 모델을 적용하였다.

$$P_i = P(X_1)P(X_2 \mid X_1) \cdots P(X_W \mid X_{W-1}) \quad (1)$$

$$p = 1 - \exp\{-P_i \cdot (L - W + 1)\} \quad (2)$$

p는 길이가 L인 하나의 서열에서 i라는 개체가 적어도 한번 나타날 확률이다. 이는 P_i의 확률이 작기 때문에 포아송 분포를 따른다는 가정으로 구하였다. 따라서 i라는 개체의 적합도는 길이가 L인 N개의 서열들로부터 발견되리라 기대되는 서열의 수로 하였다. 개체 i의 적합도는 아래의 f_i로 구한다.

$$f_i = N \cdot P \quad (3)$$

각 개체에 대한 적합도를 구할 때 다음과 같은 제약을 두었다. [그림 7]의 바깥 루프를 한번이라도 돌고 나서 다시 새로운 세대의 집단에 있는 개체들의 적합도를 구하기 위해서 이미 유의한 단편으로 판별이 된 것이 개체로서 존재한다면 그 개체의 적합도를 0으로 하여 더 이상 고려대상이 되지 않도록 하였다.

4.1.2 선택(Selection)

적합도 값이 높은 개체일수록 다음 세대로의 생존 여부에 대한 기회를 더 많이 제공받고, 반대로 적합도 값이 낮은 개체들은 자연도태가 되도록 한다. 또한 최적의 해를 구하기 위해서 교차라든지 돌연변이 같은 조작을 각각의 확률로 수행하여 개체들에 변형을 가한다. 집단의 각 개체들의 적합도를 f_i에 의해 구하고 나서 최량 적합성 값인 B_V와 평균 적합성 값인 A_V를 계산한다. 여기서 B_V는 수렴성 여부를 가릴 때 사용된다. 각 개체의 선택의 수는 A_V/f_i 값으로서 결정한다. 즉, 우선 A_V/f_i 값의 정수만큼의 개수를 선택한다. 그리고 나머지 소수 부분의 확률만큼 더 선택을 한다. 그러나 이렇게 하면 집단의 개수가 기대 이상으로 더 늘어나거나 감소하는 경우가 생긴다. 이러한 것을 방지하기 위하여 소수 부분의 확률만큼 더 선택된 개체들은 임시 집합인 *Temp_Set*에 저장해둔다. 그리고 집단의 크기인 M과 정수만큼 선택된 개체수 사이의 차이만큼 다시 *Temp_Set*으로부터 무작위로 추출하여 항상 집단의 크기가 M이 되도록 한다.

4.1.3 교차(Crossover) 및 돌연변이(Mutation)

교차는 유전자 알고리즘에서 개체들의 변형을 일으키는 주된 조작으로서 변형된 2점 교차를 사용하였다. 방법은 [그림 8]과 같다. 교차를 적용할 두 개체를 선택하면 임의로 0~W의 수 중에서 세 개를 구한다. 첫번째 수는 첫번째 개체에서 교차를 일으킬 시작점을 가리키고 두 번째 수는 두 번째 개체에서 교차를 일으킬 시작점을 가리키며 세 번째 수는 교차 대상의 길이를 가리킨다. 이렇게 변형된 2점 교차를 사용한 이유는 초기 집단 생성시의 개체들로부터 패턴 시작점이라든지 서열내 일부분의 위치를 제대로 찾을 수 있는 기회를 보다 효과적으로 부여하기 위함이다. 교차는 P_c의 확률로서 이루어지는데 0.6~1.0이 적당하다고 알려져 있다. 본 연구에서는 수렴에 근접하는 정도에 따라서 다르게 주었다. 즉, 수렴에 근접했을 경우엔 좀 더 낮은 교차율을 주어 빨리 수렴에 이르도록 하였다.

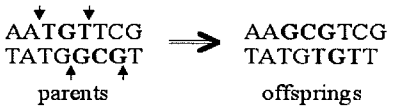

그림 8 변형된 2점 교차

돌연변이는 교차와 더불어 개체들의 변형을 일으키는 조작으로서 핵산염기 단위로 P_m의 확률로 이루어진다. 돌연변이를 너무 큰 변이 확률로 설정하면 임의 탐색으로 변해버리게 되지만 그렇다고 돌연변이가 없는 경우에는 초기 유전자의 조합 이외의 공간을 탐색할 수 없으며, 결국 찾고자 하는 해의 질에도 한계가 드러난다. 따라서 어느 정도의 변이는 필요한데 보편적으로 사용되는 0.001의 확률을 적용했다. 현재 세대의 집단이 선택교차에 이어 돌연변이라는 조작을 거치게 되면 다음 세대의 집단으로 결정된다.

그림 9 점 돌연변이

4.1.4 수렴

최종 수렴에 이르기까지 두 가지의 수렴 기준이 적용된다. [그림 7]에서 보면 안쪽 루프의 수렴여부와 바깥쪽 루프의 수렴여부가 각각 필요하다. 안쪽 루프는 초기 집단으로부터 유의한 단편들을 찾기 위한 과정이고 바깥쪽 루프는 찾아진 것 외에 다른 유의한 단편들을 찾기 위해서 초기 집단을 생성하는 단계로 다시 되돌아가는 과정이다. 안쪽 루프의 수렴여부는 B_V와 A_V의

비로써 결정이 된다. *B_V/A_V*의 값이 보다 크면 수렴하지 않은 경우로써 유전자 알고리즘의 세 가지 조작인 선택, 교차 및 돌연변이를 적용하여 다음 세대의 개체들을 재생산한다. 그리고 그 비가 보다 작으면 수렴한 경우로써 바깥쪽 루프의 과정으로 수행이 된다. 안쪽 루프가 수렴이 되었을 경우 수렴이 된 집단으로부터 유의한 단편들을 추출하여, 패턴이라고 결정하기 전의 유의한 단편들을 모아두는 *FRAG*라는 집단에 저장한다 [그림 6]. 그리고 바깥쪽 루프를 돌면서 유의한 단편들은 계속적으로 이 집단에 저장된다. 이때 단편이 유의한지의 여부를 결정하는 기준이 있어야 한다. 이러한 기준은 단편이 무작위적이라고 가정했을 때 유의수준(α)으로 나올 수 있는 기대값으로 하였다. 이 값을 임계값 (*Limit*)으로 하여 이 값보다 큰 적합도를 갖는 단편들은 유의한 것으로 결정한다. 임계값을 구하는 식은 다음과 같다.

$$p = 1 - \exp\{-(0.25)^{W} \cdot (L - W + 1)\} \quad (4)$$

$$Limit = Z_{\alpha} \cdot \sqrt{N \cdot p \cdot (1-p)} + N \cdot p \quad (5)$$

위의 식에서 임의의 개체 i가 될 확률은 그 개체가 임의의 문자열이라고 가정을 하고 각 핵산 염기가 A, T, G, 및 C 네 가지 중 하나이기 때문에 단순히 0.25의 확률을 W제곱하였다. 그리고 바깥쪽 루프의 수렴여부를 결정하기 위해서 안쪽 루프에서 수렴된 집단에 속하는 개체들의 적합도가 모두 *Limit*값보다 작은지를 살펴본다. 만일 개체들의 적합도 모두가 *Limit*값보다도 작은 경우가 ρ회 이상 계속된다면 알고리즘의 전체적인 수렴이라고 결정하고, 그렇지 않다면 두 가지 루프과정을 계속 진행한다. 본 연구에서는 ρ를 20회로 적용하였다.

4.2 적당한 길이의 패턴들을 찾는 단계(Step 2)

위의 단계를 거치고 나면 *FRAG*라는 집단에는 유의한 단편들이 모이게 된다. 그러나 이러한 단편들이 찾고자 하는 패턴이라고 말하지 못하는 이유는 W라는 길이로 고정을 시킨 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 6]. 여기서 최대 길이를 10 bp로 정한 이유는 대부분의 전사인자 결합부위들이 6~10 bp이라는 점과 Wataru의 실험결과에서 7 bp의 패턴이 최대 길이라는 사실을 참고하여 임의로 정한 것이다. 그리고 실제 실험의 결과에서도 가장 긴 패턴이 7 bp였다. *Pre_PATT* 집단에 있는 서열들의 길이를 고려한 적합도 함수에 의해 적합도 값을 구한다. 이때 *Pre_PATT* 집단의 i번째 염기서열을 한 염기씩 줄여가면서 W이상의 모든 가능한 길이의 내부 염기서열들의 적합도 값을 구한다. 그리고 *Limit*값보다 큰 것들을 따로 모으고 그 중에서 Z값이 가장 큰 염기서열을 i번째 염기서열에 대한 패턴으로 간주하고 *PATT*라는 집단에 중복이 되지 않게 저장을 한다. Z값은 다음과 같이 계산한다.

$$Z = \frac{f_i - N \cdot p}{\sqrt{N \cdot p \cdot (1-p)} / \sqrt{N}} \quad (6)$$

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

5. 실험 및 결과

5.1 실험 데이타 추출

현재 EPD에는 인간 유전자 프로모터부위에 대한 염기서열 데이타가 377개 있으나, 데이타의 개수가 조직이나 기능별로 그룹을 만들어 특이적인 실험 결과를 얻기에는 부족한 형편이다. 하지만 현재 실험적으로 검증된 프로모터를 얻을 수 있는 다른 데이타베이스가 없는 형편이다. 그래서 EPD에서 가장 많이 확보할 수 있는 하우스키핑 유전자 프로모터들과 간에서 발현되는 유전자 프로모터들, 즉 각각 20개와 26개의 두 그룹으로써 실험을 하였다. 간 유전자 그룹의 경우에는 간에서 발현되지만 다른 조직에서도 발현될 수 있는 것들도 포함시켰다. EPD의 각 엔트리 중에서 발현/조절 하부핵심어가 하우스키핑 유전자와 간으로 일치하는 것들 가운데 전사 시작점의 앞부분이 250 bp이상의 길이를 만족하는 것들을 각각 하우스키핑 유전자 프로모터와 간 유전자 프로모터 그룹으로 생성하였다. 실제 핵산 염기서열은 EPD와 하이퍼링크된 EMBL 핵산 데이타베이스로부터 추출하였다. 추출된 데이타들은 양이 많아 본 논문에 싣지 않았다. 이러한 데이타들을 면밀히 조사한 바에 의하면, 하우스키핑 유전자 프로모터의 경우 각 염기의 조성 비율이 A가 0.189, T가 0.169, G가 0.324 그리고 C가 0.318로써 G+C %가 상대적으로 높고, 간 유전자 프로모터의 경우 A가 0.231, T가 0.229, G가 0.272 그리고 C가 0.268로써 모두 비슷한 비율로 나타나는 가운데 TATA 상자가 많이 포함되어 있었다.

5.2 실험 결과

유전자 알고리즘에서 수렴에 이르기까지 하우스키핑 유전자 그룹의 경우 1655세대, 간 유전자 그룹의 경우 2533 세대를 거쳤다. 각 그룹별 수렴추이가 [그림 10]과 [그림 11]에 나타나 있는데 900세대 이후는 생략하였다. 각 그림에서 가로축은 세대의 변이를 의미하고 세로축은 적합도 값을 의미한다. 굵은 선은 임계값을 나타내는데, 임계값 이상의 적합도 값을 갖는 개체들을 유의한 것으로 간주한다. 유전자 알고리즘이 이중 루프이기 때문에 그림에서 구간마다 수렴이 되고 다시 유전자 알고리즘이 반복되는 모습이 보여진다. 수렴의 여부는 평균값과 최고값의 비율로써 정하였는데 거의 1이 되었을 때 수렴한 것으로 판정하였다. 또한 임계값 이상의 최고값이 없는 구간이 20회 이상 반복되면 전체적인 수렴으로 간주하였다.

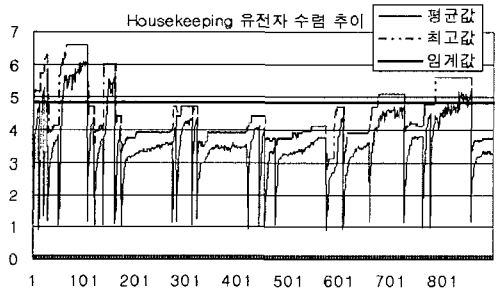

그림 10 하우스키핑 유전자 그룹의 수렴 추이

두 그룹에 대해서 각각 최적의 길이를 갖는 유의한 패턴들 가운데 가장 높은 Z 값을 갖는 10개의 패턴을 결정하였다. [표 1]과 [표 2]는 이들 10개의 패턴들과 이들의 적합도 값, 실제 각 패턴이 20%의 불일치를 허용할 경우 그룹 내에 나타나는 염기서열의 수 그리고 Z 값 등을 보여준다. 각각 확률적으로 계산된 적합도 값에 비해 실제 나타나는 빈도수는 거의 모든 염기서열에서 나타나는 수준이었다. Z 값은 약 4정도 되더라도 0.1%의 유의수준에 해당하는데 모두 이보다 훨씬 큰 두 자리수로써, 결정된 패턴들의 유의성을 검증할 수가 있다. 게다가 위의 각 그룹 내에서 유의하게 나타난 10개의 패턴들과 일치하거나 유사한 TFD내의 포유동물의 전사인자 결합부위 염기서열 및 전사인자 등을 찾아보았다. 유사한 기준은 염기서열 길이의 차이가 2 bp까지 그리고 불일치 정도는 20 %까지 허용하였다. 하우스키핑 유전자의 경우, G+C %가 높아서 예상대로 거의 G 및 C가 포함된 패턴들만이 추출되었고, 간 유전자의 경우는 예상대로 잘 알려진 HNF1(Hepatocyte Nuclear Factor-1)과 LF-A2(Liver Specific Factor-A2) 들이 검출되었다.

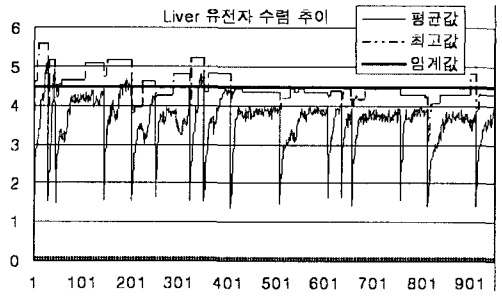

그림 11 간 유전자 그룹의 수렴 추이

표 1 하우스키핑 유전자 프로모터 그룹의 대표적인 10개의 패턴들

Patterns	Fitness value	Frequency	Z value
GGCGCG	6.61	18	23.29
GGCGGGG	2.96	16	22.05
GGCGGC	6.13	17	21.23
GCCGGG	6.09	18	21.06
GAGGCG	6.09	18	21.06
GCGGC	6.03	17	20.81
GCGCAC	5.93	13	20.39
GACGGC	5.77	18	19.72
GCCGTC	5.68	18	19.31
CGCGCC	5.54	17	18.73

표 2 간 유전자 프로모터 그룹의 대표적인 10개의 패턴들

Patterns	Fitness value	Frequency	Z value
CGCCCC	5.70	20	17.93
GAGGGG	5.62	20	17.57
CCTGCCC	2.53	12	16.31
AGAGGT	5.23	21	15.91
CCTCCCC	2.28	14	15.70
CCCTTC	5.17	19	15.65
CCCAAG	5.08	22	15.27
AGGCCC	5.08	21	15.27
CCAGGG	5.07	21	15.22
CGGTTAA	2.20	8	14.97

6. 결론 및 향후 연구계획

핵산 염기인 A, T, G, C의 조합으로 이루어진 DNA 염기서열 상에서 유전자의 발현에 대한 신호역할을 하는 프로모터 영역들을 선택하여, 유전자가 어떤 기능을

하는지 아니면 어떤 특별한 조직에서 특이적으로 발현하는지의 여부를 기반으로 그룹을 만들고, 해당 그룹 내에서 특이적으로 발견되는 패턴들을 찾아 보았다. 프로모터 영역에 존재하는 전사인자들의 결합부위는 각 전사인자들이 잘 인식하고 결합할 수 있도록 진화론적으로 잘 보존되어 있는데, 문자열로 말하자면 유사한 패턴의 형태로 되어있다. 패턴 자체 또는 그것의 일부분으로서 가능성이 높은 단편들을 유전자 알고리즘과 마르코프 연쇄 모델을 이용하여 구하였다. 최적화 알고리즘인 유전자 알고리즘의 특성상 전역적인 탐색을 함으로 국소최대값에 빠질 염려가 없고 또한 *Limit*라는 임계값을 두어 이중 순환을 하였기 때문에 유의한 모든 패턴들의 탐색이 가능하다. 그리고 각 단편들로부터 Z 값을 기준으로 적당한 길이를 갖는 최적의 패턴들을 생성하였다. 이는 MEME 알고리즘[12]의 N-occurrence-per-dataset 모델의 가정을 따르면서 또한 Wataru[7] 방법의 패턴 길이까지 결정할 수 있는 장점을 취하였다. 실험 데이타로서 하우스키핑 유전자의 프로모터 영역과 간 유전자의 프로모터 부위를 개별 그룹으로 하여 EPD와 EMBL 핵산 염기서열 데이타베이스로부터 추출하였다. 이것들은 전사 시작점에서 250 bp 앞부분인 염기서열들이다. 결과적으로 얻어진 모든 패턴들이 TFD에 존재하는 기존에 알려져 있는 전사인자 결합부위들과 같거나 유사하였다. 이로 인해 염기서열상에서 유의적으로 나타나는 패턴들과 결합부위들 간의 연관성이 다분히 존재한다고 결론지을 수 있다. 그러나 같거나 유사한 것으로 검색된 TFD의 결합부위들 사이의 생물학적 의미까지는 고려하지 못하였으며, 하우스키핑 유전자와 간 유전자의 프로모터 그룹 간의 차이점을 판별할 수 있는 모델을 제시하지 못했다. 이는 차후의 연구 과제로서 해결되어야 할 문제라고 본다. 본 연구에서는 염기서열을 단순히 A, T, G, C의 조합으로 이루어진 문자열로만 보고 유전자 알고리즘을 이용하여 통계적 추론을 하였다. 그러나 현재 전사인자 및 이와 상호 작용하는 여러 거대 분자들에 대한 생물학적 정보가 많이 알려져 있지 않아서 그러한 정보까지 고려하지 못했지만 이러한 생물학적 지식들을 충분히 반영한다면 차후에 보다 흥미 있는 연구가 될 것으로 여겨진다. 또한 현재 EPD데이타베이스에 존재하는 실험적으로 검증된 프로모터들의 수가 상당히 부족한 형편이며, 조직 부위별이라든지 기능별로 충분히 정리되어 있지 않은 상태이다. 따라서 신빙성 있는 데이타를 충분히 얻기가 힘들었고 따라서 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.

비록 향후에 해결되어야 될 문제점들이 많이 남아 있지만, 어쨌든 본 연구에서 행한 패턴 탐색을 통하여 유전자의 발현여부에 중요한 영향을 주는 여러 전사인자들의 결합부위를 효율적으로 찾아낼 수 있을 뿐만 아니라 전사조절 메커니즘을 규명하는데도 커다란 기여를 할 수 있을 것이다. 나아가서는 특정 프로모터 영역을 예측함으로써 신규 유전자 발굴 및 유전자의 기능별 분석도 가능하게 되어 엄청난 생물학적 실험 비용 및 시간을 절약함으로써 획기적인 경제적 파급효과를 얻을 수 있을 것으로 본다.

참 고 문 헌

[1] E. Snyder and G. Stormo, Identification of protein coding regions in genomic DNA, Journal of Molecular Biology, Vol. 248, pp 1-18, 1995.

[2] M. Burset and R. Guigo, Evaluation of gene structure prediction programs, Genomics, Vol. 34, pp 353-367, 1996.

[3] C. Burge and S. Karlin, Prediction of complete gene structures in human genomic DNA, Journal of Molecular Biology, Vol. 268, pp 78-94, 1997.

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

[5] 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.

[6] 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.

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

[8] Dan S. Prestridge, "Predicting Pol II Promoter Sequences using Transcription Factor Binding Sites", Journal of Molecular Biology, Vol. 249, pp. 923-932, 1995.

[9] Dan S. Prestridge, SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements, CABIOS, Vol. 7, pp. 203-206, 1991.

[10] James W. Fickett and Artemis G. Hatzigeorgiou,

"REVIEW Eukaryotic Promoter Recognition", Genome Research, Vol. 7, pp. 861-878, 1997.
[11] Thomas D. Schneider, Gary D. Stormo and Larry Gold, Information Content of Binding Sites on Nucleotide Sequences, Journal of Molecular Biology, Vol. 188, pp. 415-431, 1986.
[12] Timothy Bailey and Charles Elkan, "Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization", Machine learning Journal, Vol. 21, pp. 51-83, 1995.
[13] Z. Michalewicz, 유전자 알고리즘, 그린출판사, 1996.
[14] David Beasley, David R. Bull and Ralph R. Martin, An Overview of Genetic Algorithms, University Computing, Vol. 15, No. 2, pp. 58-69, 1993.
[15] Ching Zhang and Andrew K.C.Wong, "A genetic algorithm for multiple molecular sequence alignment", CABIOS, Vol. 13, No. 6, 1997.
[16] Cedric Notredame and Desmond G. Higgins, "SAGA: sequence alignment by genetic algorithm", Nucleic Acids Research, Vol. 24, No. 8, pp. 1515-1524, 1996.
[17] 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.
[18] Ghosh, D., A relational database of transcription factors, Nucleic Acids Research, Vol. 18, pp. 1749-1756, 1990.
[19] Timothy L. Bailey, Likelihood vs. Information in Aligning Biopolymer Sequences, UCSD Technical Report CS93-318, 1993.

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

공 은 배

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