특 집

# 유전정보 분석시스템

이 대 상, 태 홍 석, 박 기 정
(주)스몰소프트

## I. 서 론

유전체(Genome)는 유전자(GENe)와 염색체(chromosOME)의 합성어로, 특정 생명체가 생명체 고유의 기능을 수행하는데 필요한 모든 유전자원의 총합을 일컫는 말이다. 유전체 프로젝트는 genome의 모든 유전자 서열을 밝히고, genome 내에 들어 있는 유전정보 및 그 세부적인 기능과 작용을 알아내는 전반적인 작업을 말한다.

유기체의 genome 염기서열이 모두 해독된 것은 1978년 Sanger 그룹이 박테리아에 감염하는 바이러스의 일종으로, 유전체의 크기가 5 kilobase 정도인 phi-X174의 전체 염기서열을 밝혀낸 것이 최초이다. 그 이후 1995년 미국의 TIGR (The Institute for Genomic Research)의 Fleishmann 연구진이 *Haemophilus influenzae* 라는 미생물의 genome 염기서열을 모두 밝혀내면서 genome 연구가 본격적으로 시작되었다.

인간의 경우, 1988년 Human Genome Project Organization(HUGO)라는 국제연구단체의 출범으로 1개의 염기(base)를 밝히는데 1달러의 경비가 소요될 것으로 예상하여, 30억불의 예산으로 인간유전체 연구가 1990년에 본격적으로 시작되었다. 23쌍의 인간 염색체 가운데 22번 염색체가 1999년에 염기서열 해독이 완료되었으며, 2001년 2월 International Human Genome Sequencing Consortium과 Celera사가 동시에 인간 30억개의 DNA염기서열 초안을 해독했다고 Nature와 Science지에 각각 발표하였다. James Watson과 Francis Crick이 DNA의 이중나선 구조를 발표한지 50주년이 되는, 올해 4월 원래 일정보다 2년 앞당겨 human genome project가 최종 완료되었다.[1]

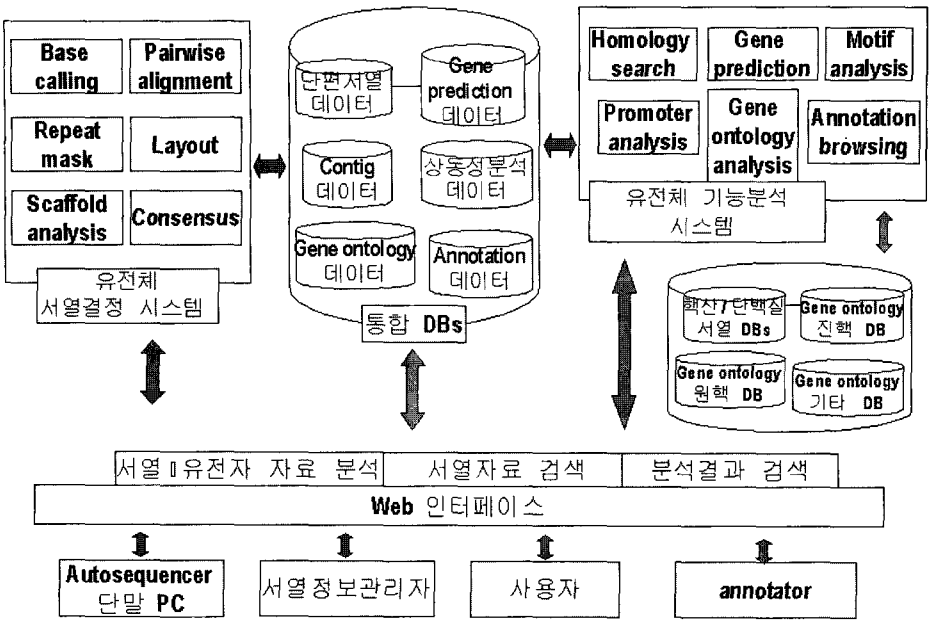
2003년 9월 현재 Genomes Online Database에 의하면 803개의 genome project가 완료 또는 진행 중이며, 이를 세부적으로 살펴보면 genome project가 완료되어 학계에 발표된 것이 160개, 원핵생물의 경우 393개의 genome project가 진핵생물의 경우 250개의 유전체 프로젝트가 진행 중이라고 한다. 이는 2000년 6월 100여종의 생물체에 대한 유전체 프로젝트 수에 비하면 급격한 증가이고 향후 유전체 프로젝트의 수는 계속 증가할 것으로 보인다.

## II. 본 론

### 1. 유전정보분석시스템(genetic information analysis system)

유전체 프로젝트를 수행하는데 있어 일반적으로 요구되는 유전정보분석시스템의 구조는 〈그림 1〉과 같다. 유전정보분석시스템은 크게 유전체 서열결정시스템, 유전체 기능분석시스템 그리고 이와 상호연관이 있는 통합 데이타베이스 시스템 등 크게 세 가지로 구성된다.[2]

유전체 서열결정시스템의 경우, 대용량 sequencing 기계로부터 나오는 염기서열에 대한 정보를 볼 수 있는 base calling, 반복되는 염기서열과 vector 염기서열의 오염을 걸러주는 repeat mask




〈그림 1〉 유전체 프로젝트 수행을 위한 유전정보분석시스템 구조.

와 vector screening, sequecing 결과로 생기는 contig들의 물리적 위치를 결정해주는 contig assembly와 이것의 결과를 시각적으로 보여주어 연구자가 수정 또는 편집이 가능한 환경을 제공해주는 contig visulalization, 관련 contig들을 연결시켰을 때 나타나는 scaffold의 방향이나 전체 genome상에서 차지하게 될 scaffold 상호간의 위치 등을 분석해주는 기능을 담당하는 scaffold analysis 기능을 포함하고 있다. 유전체 서열결정시스템에서 얻어진 유전자의 염기서열이 향후 유전정보분석에 있어 가장 기본이 되는 작업이므로, 이 단계에서의 관련 생명체가 가지고 있는 염기서열을 정확히(오류범위 0.01% 이하) 읽어내는 것이 중요하다.

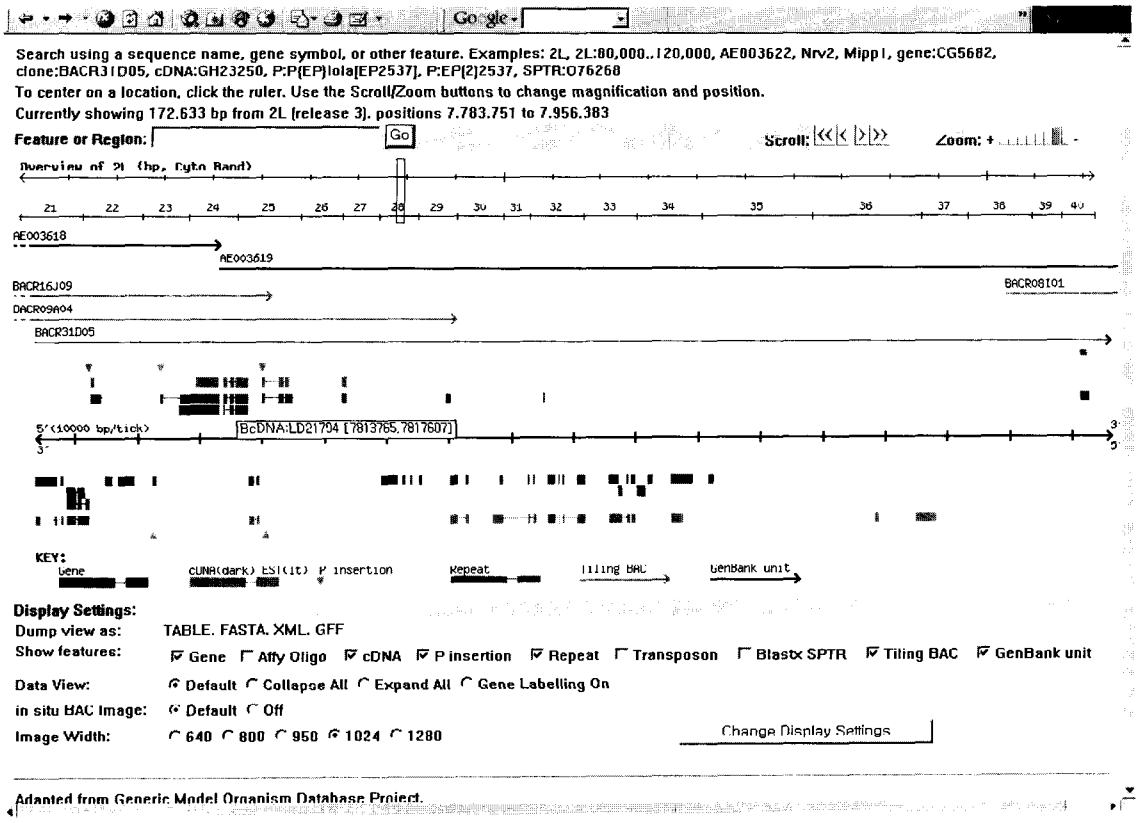
유전체 기능분석시스템은 크게 6개의 세부 모듈로 구성되어 있다.

앞으로 상세히 다룰 유전자 예측(gene prediction)과 프로모터 예측(promoter prediction)이 가장 핵심적인 분석기능이며 유전자 예측의 정확성에 따라서 genome project의 성패가 결정된다고 하여도 과언이 아니다.

Contig 서열자체나 유전자 예측의 결과로 나온 유전자들에 대하여 기존의 핵산 또는 단백질 Database와 상동성 검색을 지원하는 homology search는 현재 유전체 프로젝트의 결과로 기하급수적으로 증가하고 있는 데이터로 인하여 계산시간이 가장 많이 소요되는 late limiting step이다. 이러한 검색에 소요되는 시간문제를 해결하기 위하여 병렬처리기법이나 PC-cluster server와 같은 하드웨어의 아키텍처를 활용하는 등 다양한 시도가 진행되고 있다.

Motif 분석은 유전자 발현의 산물인 단백질이 고유의 기능을 수행하는데 필요한 기능을 담당하는 부위를 분석하는 방법으로 그 중요성은 크게 평가받고 있다. 이러한 기능 부위는 생명체의 수억년 동안의 진화과정에서도 기능은 보존되어질 가능성이 높으므로 PROSITE와 같은 대표적인 단백질 데이터베이스를 기반으로 하여 패턴검색이 주로 이루어져 왔다. 또한 motif 구성을 위한 알고리즘은 information theory나 HMMs (hidden markov models)을 활용한 통계적이나 기계학습 방법으로 비교적 정교한 motif 프로파일을 구성하기 위한 방법들이 개발되고 있다.

유전체 프로젝트의 증가와 아울러 유전체 분석




〈그림 2〉 초파리(*Drosophila melanogaster*) 2번 염색체의 annotation browser.

작업의 최종단계인 Gene Ontology 분석의 경우, 유전체를 구성하는 개별 유전자가 세포의 생체작용에 필요한 개별 분류항목들 가운데 어디에 속하고 있는지 분석하여 세분화된 분류체계에 할당하는 작업을 수행한다. 원핵생물(prokaryotes)과 진핵생물(eukaryotes)의 기능분류체계가 서로 다르고, 진핵생물 내에서도 식물과 동물에 따라 기능분류체계가 서로 다르므로 현재의 분류체계의 개선과 unknown 또는 unidentified 범주의 기능이 새로운 분류항목으로 추가되어 더욱 복잡하고 정교한 분류체계가 마련될 것으로 보인다. 나아가 생명체 각 종(species) 또는 속(genus)에 따른 추가적인 분류가 도입될 것으로 예상된다.[3]

Annotation browser 기능은 진행 중이거나 완료된 genome project의 annotation의 결과에 대하여 조사하고자 하는 유전자의 유전체내의 물리적인 위치 등을 포함한 해당 유전자가 가지고 있는 모든 유전정보를 그래픽화면으로 가시화시켜주는 작업을 수행한다. 수작업으로 annotation이 필요한 유전자의 경우 이것의 정보 수정 및 보완작업도 가능케 하는 기능을 가진다〈그림 2〉.

통합데이타베이스 시스템은 genome project를 수행할 때 마다 생성되는, clone, contig, gene prediction, 상동성 검색 data, annotation data를 실시간으로 처리할 수 있어야 한다. 또한 상동성 검색을 위한 핵산 및 단백질 DB, 미생물 Gene Ontology DB, 식물 및 동물 Gene Ontology DB 등이 필수적으로 요구된다.

이러한 유전정보분석시스템의 기본적인 구조에, 최근 각광받고 있는 서로 다른 종(species)의 생명체들을 유전체 수준에서 비교분석작업을 수행할 수 있는 비교유전체(comparative geno-

mics) 시스템이 추가된다면 이는 보다 더 강력한 유전정보분석시스템이 될 것으로 사료된다.

### 2. 프로모터 예측(promoter prediction)

프로모터는 좁게는 RNA polymerase가 binding하는 DNA의 일부분(core promoter region)을, 넓게는 유전자발현에 영향을 미치는 coding sequence의 5′ upstream 부분을 말한다. 유전자 발현의 첫 번째 단계인 전사의 시작(transcription initiation)은 특정 유전자가 시간과 공간적으로 적절하게 발현되어야 생명체 본연의 기능을 수행할 수 있으므로 유전자의 조절에 있어 매우 중요하다.

프로모터는 CpG island, Transcription factor binding sites, TATA box, CAAT box, initiator site 등의 주요 특성을 보이며, 이들은 원핵생물과 진핵생물에 따라 나타나는 양상이 서로 다르며 진핵생물 내에서도 척추동물과 무척추동물에 따라 나타나는 패턴이 다르다. 이러한 프로모터 부위의 복잡성 때문에 프로모터 예측은 매우 어려운 분야에 속한다.

원핵생물의 프로모터 예측방법은 이미 알려진 consensus sequence(pattern)나 주어진 염기서열 set로부터 counting 접근 방법으로 consensus sequence를 추출해 내는 방법들이 초기에 주로 사용되었다. 그러나 이러한 DNA의 프로모터와 같은 기능성 부위들(functional sites)은 엄격히 보존되지 않았고, positions에 따라서도 각기 다른 보존성 정도를 보이므로 어떤 pattern들은 modeling할 수 없는 단점 때문에 pattern을 이용한 방법들은 DNA binding의 특이성을 나타내기에는 부적절하였다.

이러한 pattern 방법의 문제점을 해결하기 위해 DNA binding site의 특이성을 weight matrix를 사용하여 나타내는 방법이 1980년도에 도입되었다. Mulligan이 log-likelihood matrices로 구한 스코어가 주어진 프로모터 set의 양적인 활성도와 높은 연관성이 있음을 보였으나, 이러한 weight matrix 방법 또한 30~40%의 true promoter 예측에 실패했고, 예측된 프로모터의 45~60% 정도가 false positive로 나올 정도로 성능이 떨어졌다. 이러한 false positive를 줄이려는 다양한 시도가 DNA binding site의 프로모터를 구성하고 있는 다양한 요소들 간의 정확한 조합과 공간적 조직을 포함한 상호의존 관계를 고려하여 시도되고 있다.

Markov chains과 Hidden Markov Models (HMMs)을 이용한 프로모터 예측 방법은 최근에 시도되고 있으나 앞서 언급한 weight matrix 방법처럼 프로모터 지역과 비프로모터(non-promoter) 지역에 대한 분별력이 떨어져 다수의 false positive를 생산한다는 문제점이 있다. 1992년 Kanehisa가 neural network 기법을 이용하였으나 이러한 방법을 genome project에 곧바로 응용하기에는 neural network을 훈련시킬 만큼의 충분한 예가 부족하다는 단점 때문에 널리 활용되지 못하였다. Expectation-maximization(EM), Gibbs sampling algorithm, MEME 등의 방법을 통해 원핵생물의 promoter의 motif를 추출하는 방법들이 현재 시도되고 있다.[4]

진핵생물의 경우 원핵생물과는 달리 RNA polymeraseII라는 단백질이 전사를 담당하고 있으며 여기에 Transcription factors라는 수많은 보조 단백질이 전사 개시에 관여하고 있으므로 원핵생물에 비해 훨씬 복잡한 전사 mechanism을 가지고 있다.

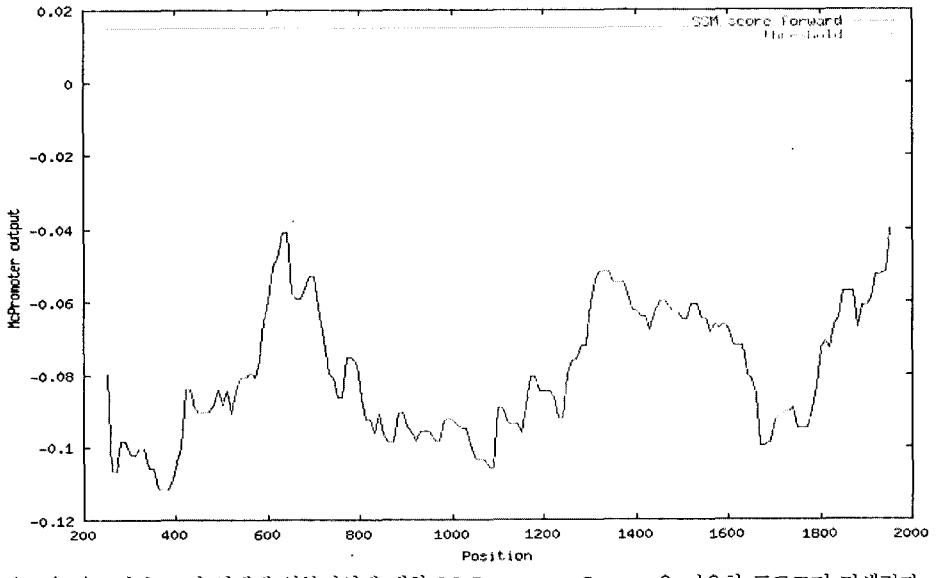
진핵생물의 프로모터 예측은 프로모터 영역과 비프로모터 영역의 유전자서열의 성분(composition) 차이에 기초하여 조절부위를 찾는 search-by-content algorithm과 앞서 언급한 TATA box나 initiator, transcription factor binding sites와 같은 프로모터 구성요소들을 발견에 기초를 둔 search-by-signal algorithms과 이 두 가지 방법을 적절히 조합한 algorithm으로 나눌 수 있다.[5]

Search-by-content algorithm을 사용한 기존의 프로모터 예측 방법들은 1996년도 Huthchinson 프로모터와 단백질 coding region 그리고 noncoding region간의 핵산 6개의 빈도

차이를 측정하는 방법을 이용한 PromFind, 1997년도의 Audic과 Claverie등이 실험적으로 증명된 promoter Database인 EPD(www.epd.isb-sib.ch)의 척추동물 프로모터와 비프로모터 염기서열에 대하여 서로 다른 차수(order)의 markov chain을 훈련시키고 이것을 Bayes' rule을 이용하여 250bp의 슬라이딩 윈도우로 분류한 프로그램, 2000년 Scherf가 무작위로 선택한 프로모터의 training sequence로부터 추출한 genetic context 에 초점을 맞추어 개발한 PromoterInspector[6], 2000년 Zhang이 위치의존적인 5-tuple 측정과 quadratic discriminant analysis(QDA) 방법을 이용한 CorePromoter 등이 대표적인 프로그램들이다.

Search-by-signal algorithm에 기초한 프로모터 예측 프로그램들은 1995년 Prestridge가 TATA box와 특정 transcription factor binding sites의 density를 고려하여 만든 PromoterScan, 1996년 Reese가 delay neural net architecture를 이용하여 TATA box와 loosely conserved initiator region element를 training시켜 만든 NNPP, 1999년 Knudsen이 transcription factor binding site에 대하여 multi-layer perceptron에 기반한 Promoter 1.0, 2000년 Chen이 weight matrices의 collection과 over-represented oligonucleotides에 기초한 PromFD 1.0 등이 있다.

Search-by-content 와 search-by-signal algorithm을 조합한 방법으로는 1997년 Solovyev가 linear discriminant function combining, TATA box score, transcription start site (TSS)주변의 triplet preferences, TSS로부터 1~100, -101~-200, -201~-300 지역의 hexamer 빈도 등을 이용한 TSSG가 있다. Ohler가 interpolated Markov chain을 기반으로 하여 promoter 부분을 크게 upstream 1, upstream 2, TATA , spacer, initiator, downstram의 6개의 state로 나눠 개발한 McPromoter system이 있다. 인간의 12번째 염색체의 일부분을 McPromoter System을 이용하여 프로모터 부분을 예측한 결과를 〈그림 3〉에 표시하였다. 아울러 2002년 Bajic이 nonlinear promoter recognition model, signal processing, artificial neural network과 새로이 개발된 sensor 등의 방법을 조합하여 만든 Dragon Promoter Finder(DPF)가 있다.




〈그림 3〉 인간 12번 염색체 일부지역에 대한 McPromoter System을 이용한 프로모터 검색결과.

〈표 1〉 진핵생물 프로모터 검색 프로그램들의 분류 및 특성

| Search by | 저자 | 프로그램명 | core promoter signal | TF binding site | CpG island 특성 | sensitivity (%) |
|---|---|---|---|---|---|---|
| content | Huthchinson | PromFind | no | no | no | 29 |
| | Audic | - | no | no | no | 24 |
| | Scherf | PromoterInspector | no | no | no | 48 |
| | Zhang | CorePromoter | no | no | no | - |
| signal | Prestridge | PromoterScan | yes | yes | no | - |
| | Reese | NNPP | yes | no | no | 54 |
| | Knudsen | Promoter 1.0 | yes | yes | no | - |
| | Chen | PromFD 1.0 | yes | yes | no | - |
| both | Solovyev | TSSG | yes | no | no | 29 |
| | Ohler | McPromoter | yes | yes | yes | |

Sensitivity는 (true positive)/(true positive+false negative)의 비율을 %로 표시하였으며, 수치가 없는 경우는 나머지 프로그램들과의 성능비교테스트 set이 적절치 않거나 비교가 부적절하여 표기하지 않았다.

진핵생물의 프로모터 예측 프로그램들의 특성을 정리한 〈표 1〉에서 보는 바와 같이 sensitivity는 60%에도 미치지 못하는 저조함을 보였다.

### 3. 유전자 예측(gene prediction)

유전체 내의 정확한 유전자 위치를 알아내기 위해 많은 gene prediction 모델들이 개발되어 왔다. 생물체의 유전체에 존재하는 유전자의 위치를 정확하게 밝혀내는 것은 유전자간의 상호관계, 그 유전자 산물인 단백질들 간의 상호작용, 그리고 나아가서는 비슷한 유전자들을 가지는 생물종간의 연관성을 밝히는데 매우 중요한 의미를 가진다.

1980년대 초에 Shepherd, Fickett, 그리고 Staden과 McLachlan에 의한 gene prediction의 초기 연구는 아미노산 분포와 codon usage의 경향을 통계적으로 측정해서 genome sequence에 존재하는 단백질의 coding region을 밝혀내고자 하였다. 그 후 coding region과 non-coding region에서 k-tuple(oligonucleotide) frequencies autocorrelation, fourier spectra, purine/pyrimidine periodicity, 그리고 local compositional complexity/entropy 등과 같은 구성의 차이가 많이 존재한다는 것이 알려지면서, 이러한 구성들의 차이를 이용하여 유전체에 존재하는 coding region의 정확한 위치를 밝혀내고자 하는 시도가 이루어졌고, 이와 더불어 gene prediction프로그램들이 등장하기 시작했다. 그 중에서 Fickett의 모델에 근거한 TestCode와, neural network접근방식으로 여러 가지 구성에 대한 통계적 수치를 적용해 염기서열 단편을 coding region과 non-coding region으로 구분한 GRAIL[7]이 가장 널리 사용되었다.

이전까지 만들어진 gene prediction 모델들은 DNA의 한쪽 strand만을 분석하도록 만들어졌지만, GeneMark[8]가 등장하면서 DNA foward와 reverse strand를 동시에 분석하여 한쪽 가닥의 coding region에 의해서 다른 가닥에서도 그 위치에서 non-coding region임에도 불구하고 coding region처럼 인식되는 'shadow' coding region 문제를 해결하고자 하였다. GeneMark는 non-coding region에서는 homogeneous 5th-markov chain, coding region에서는 codon의 위치특이적인 non-homogeneous 5th-markov chain을 DNA 양쪽 가닥에 모두

$$P(COD_m|F)=\frac{P(F|COD_m)*P(COD_m)}{\sum_j P(F|COD_j)*P(COD_j)+\sum_j Q(F|COD_j)*Q(COD_j)+P(F|NON)*P(NON)} \quad (1)$$

$$Q(COD_m|F)=\frac{Q(F|COD_m)*Q(COD_m)}{\sum_j P(F|COD_j)*P(COD_j)+\sum_j Q(F|COD_j)*Q(COD_j)+P(F|NON)*P(NON)} \quad (2)$$

$$P(NON|F)=\frac{P(F|NON)*P(NON)}{\sum_j P(F|COD_j)*P(COD_j)+\sum_j Q(F|COD_j)*Q(COD_j)+P(F|NON)*P(NON)} \quad (3)$$

구성하고, 각 markov chain의 상대적인 score에 따라 coding region을 찾아낸다. 아래의 식 (1)은 임의의 염기서열단편 F에 대해 forward strand의 coding region이면서 codon 1, 2, 3 번째 위치 중 m번째 위치에서 시작할 확률을 계산하는 식이다. 그리고 식 (2)는 reverse strand의 coding region에 대한 계산식이고, 식 (3)은 non-coding region에 대한 계산식이다.

GeneMark 이후 원핵생물 유전체에 대한 gene prediction 프로그램들 중 가장 널리 사용되는 널리 사용되고 있는 프로그램은 Glimmer이다. Coding 및 non-coding region에서의 6-tuple의 출현빈도를 측정해서 coding region을 찾는 GeneMark와는 달리 Glimmer에서는 interpolated markov model을 사용하였다. Glimmer에서 사용된 interpolated markov model은 식 (4)에서와 같이 8-tuple의 출현빈도를 측정하고 그 측정값이 신뢰할 만큼의 충분한 값을 가지게 되면 $\lambda_k$를 1로 두고, 그렇지 못하면 그 출현빈도에 따라 $\lambda_k$를 $1>\lambda_k\geq 0$의 값으로 둔 후 7-tuple이나 더 짧은 길이의 출현빈도를 측정하여 markov chain을 계산한다.

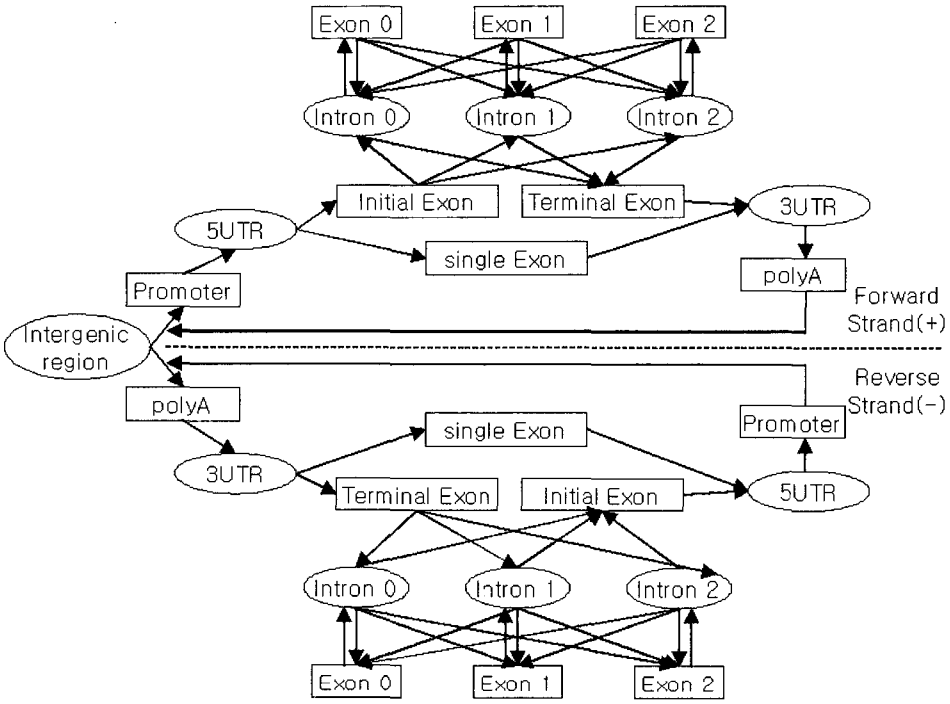
$$\begin{aligned} P(S|M)&=\sum_{x=1}^{n}\mathbf{IMM}_8(S_x)\\ \mathrm{IMM}_k(S_x)&=\lambda_k(S_{x-1})\cdot P_k(S_x)\\ &\quad+[1-\lambda_k(S_{x-1})]\cdot \mathrm{IMM}_{k-1}(S_x) \end{aligned} \quad (4)$$

Glimmer를 개선한 Glimmer2[9]에서는 interpolated Markov model 대신 interpolated context model을 사용하는데 markov chain에서는 바로 이전의 염기서열 구성에 따라 확률값을 계산하는데 비해 이 model에서는 염기서열 단편의 상호 의존성을 측정하고 그 의존성에 따라 확률 값을 계산한다.

원핵생물의 유전자와는 달리 진핵생물의 유전자는 그 구조가 더 복잡하다. 진핵생물의 경우, 유전체 크기에 비해 유전자의 밀도가 원핵생물보다 훨씬 떨어진다. 그리고 원핵생물의 유전자 구조가 Promoter, start codon, coding region, stop codon, non-coding region으로 이루어진데 비해 진핵생물의 유전자는 cap, polyA와 같이 전사에 관련된 signal이 더 존재하며, coding region도 donor, acceptor signal에 따라 exon과 intron으로 나누어진다. 전사를 통해서 만들어진 pre-mRNA에 존재하는 intron들은 splicing이라는 과정을 통해 제거되고, 그 이후 번역이 이루어져서 단백질이 생성된다.

진핵생물에 대한 gene prediction은 선충의 일종인 *C. elegans*와 포유동물의 입력 염기서열로부터 initial exon과 terminal exon을 포함하는 완성된 구조의 유전자를 찾아내고자 하였다.

이 후 이러한 개념에 기초한 gene prediction에 대한 연구가 활발해지면서 많은 진핵생물 gene prediction 알고리즘들이 개발되었다. 그 예로 hierarchical rule을 이용하여 exon의 가능성이 있는 단편에 대해 순위를 계산하는 모델을사용한 GeneID, neural network과 dynamic programming을 혼합한 GeneParser, linguistic method를 사용한 GenLang, discriminant analysis를 사용한 FGENEH, decision tree를 사용한 morgan, generalized HMM을 사용한



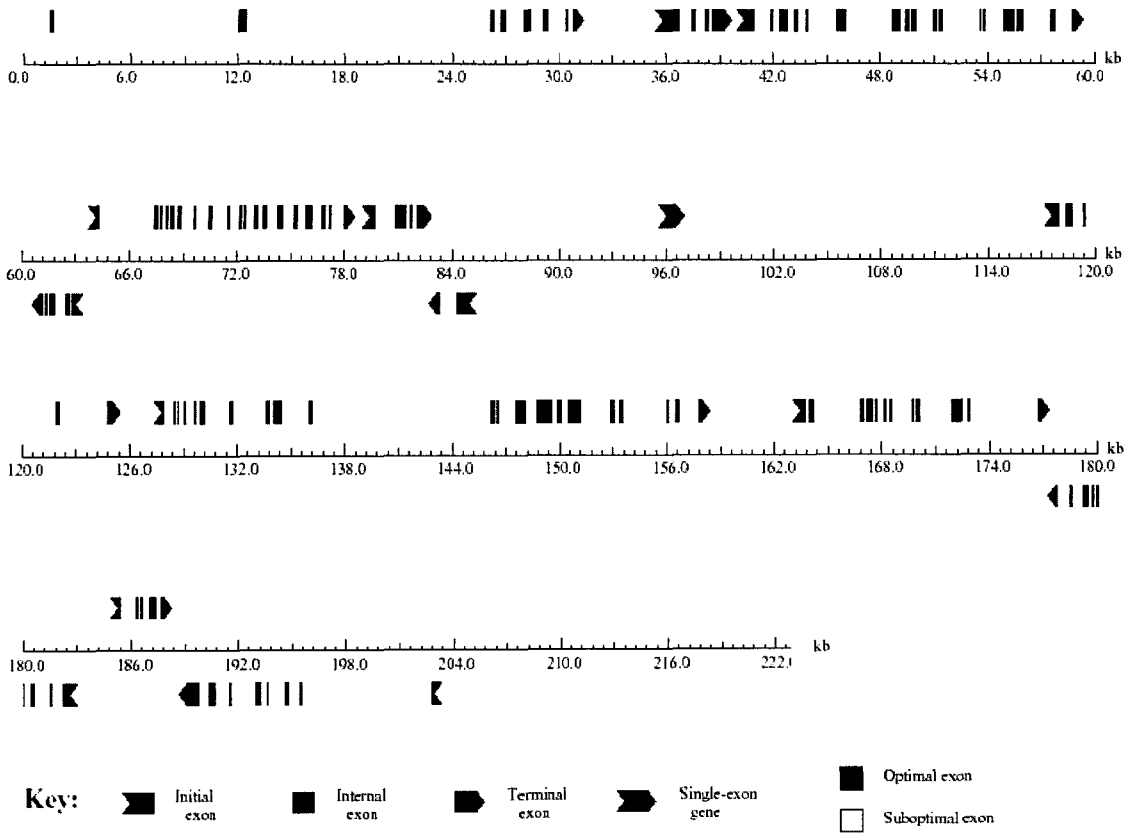

〈그림 4〉 그림 4. Duration HMM을 이용한 gene prediction model.
사각형과 동그라미는 HMM의 state들을 나타낸다. 각 state들 간의 transition은 오직 화살표 방향으로만 가능하다. Intergenic region을 제외한 모든 state들은 forward와 reverse strand에 따라 구분되어 한 쌍씩 존재한다. Forward와 reverse strand의 transition 방향은 반대방향으로 향한다.

Genie, 그리고 duration HMM을 사용한 GenScan[10]이 있다.

이 프로그램들 가운데서 현재 가장 널리 사용되고 있는 프로그램은 GenScan으로 HMM의 응용 model인 duration HMM을 기본 model로 하고 다양한 model을 복합적으로 적용하였다.

GenScan의 HMM을 구성하는 state들로는 유전자 발현의 조절부위를 담당하는 promoter, 전사의 종결부위에 나타나는 polyA, 전사 개시 지점과 start codon의 중간영역인 5′ UTR, stop codon과 polyA의 중간영역인 3′ UTR, intron 없이 하나의 exon만으로 하나의 유전자를 구성하는 single exon, start codon으로 시작하는 initial exon, stop codon을 포함하는 terminal exon, intron과 intron 사이에 있는 internal exon(시작하는 codon의 위치에 따라서 exon 0, exon 1, exon 2로 나눈다.), 그리고 바로 이전 exon이 끝나는 codon의 위치에 따라서 나누어진 intron 0, intron 1, intron 2로 구성되어있다. 이 state들은 forward와 reverse strand에 따라서 구분하며, 유전자 구조에 포함되지 않는 intergenic region까지 모두 27개의 state들로 구성되어 있다〈그림 4〉. GenScan 프로그램을 이용하여 인간 염색체 12번의 일부영역(12p13, GenBank ACCESSION number U 47924)에 대한 유전자 예측 결과는 〈그림 5〉와 같다.

각 state들 간의 transition은 전사에 관련된 signal들인 promoter와 polyA tail, 번역에 관련된 signal인 start codon과 stop codon, splicing에 관련된 signal인 donor와 acceptor의 위치에서만 이루어진다. 이러한 signal의 위치를 예측하기 위해서 weight matrix model과 weight array model 그리고, maximal




〈그림 5〉 GenScan을 이용한 인간 12번 염색체 일부분의(GenBank U47924, 222kb) 염기서열 유전자 예측 결과.

dependence decomposition과 같은 model들이 사용되었다. State들의 길이도 이 model에서는 중요한 의미를 가지며, 실제로 유전체에서 exon이나 intron의 길이는 일정한 분포를 가지는데 duration hmm은 state들의 이러한 길이의 분포를 쉽게 반영할 수 있는 장점을 가지고 있다.

## III. 결 론

유전자분석시스템의 주요 구성요소에 대한 소개와 프로모터와 유전자 예측에 널리 사용되는 주요 프로그램들을 살펴보았다. 프로모터 예측의 경우 가장 큰 문제점으로 대두되는 것이 false positive수가 많다는 것이다. 현재까지 나와 있는 프로모터 예측 프로그램 가운데 민감성(sensitivity)이 54%로 가장 높았던 NNPP의 경우 이것의 프로그램 특이성(specificity)이 460bp마다 한 개의 False positive가 발견되었을 정도로 매우 낮게 나타났다.

최근의 프로모터 예측 프로그램들은 search-by-content나 search-by-signal을 조합하는 방향으로 나아가고 있는 추세이며, 프로모터 예측을 통하여 나타난 후보군을 promoter 부분이 가지는 DNAase I derived bendability, propeller twist, chromatin 구조, nucleosome 위치 등과 같은 물리화학적 특성까지 고려하여 더 정확한 예측을 위한 추가적인 자료로 사용하는 시도가 일부 추진되고 있지만, 현재로는 DNA 이중나선의 bending이나 twisting에 대한의 정

확한 예측자료의 미비로 어려움이 예상된다.

유전자 예측의 경우 초기의 gene prediction 프로그램들에 비해 최근의 프로그램들은 많은 발전을 이루어 상당히 정확한 gene prediction을 수행하지만 아직까지는 해결해야 될 문제점들이 있다.

두 개의 유전자가 겹쳐진 위치에서 발현될 경우, 이러한 유전자들을 찾아내지 못한다. 이는 대부분의 gene prediction model들은 유전자가 겹쳐서 발현하는 것을 고려하지 않았기 때문이다. 진핵생물의 유전자 예측에 가장 널리 이용되고 있는 GenScan의 경우에도 두 개의 state들이 동시에 한 위치에서 나타나는 경우를 고려하지 않아 이러한 유전자를 예측하지 못한다. tRNA, rRNA, snRNA와 같은 유전자의 경우, 생물학적으로 중요한 의미를 지니고 있으며 높은 상동성을 가짐에도 불구하고 기존의 유전자 예측프로그램이 이를 찾아내지 못한다. 진핵생물의 경우 alternative splicing이라는 과정을 통하여 한 유전자로부터 다양한 종류의 단백질이 생성됨에도 불구하고 이러한 점을 고려하지 않아 기존의 프로그램들이 이러한 유전자의 생성물들을 찾지 못하고 있다.

프로모터와 유전자 예측 프로그램의 성능 향상은 유전체 기능분석시스템의 정확성을 높이는 것과 직결되어 있고, 이는 유전체연구를 가속화 시키는데 있어 반드시 해결해야 할 과제로 사료된다.

## 참 고 문 헌


〔1〕 http://www.ornl.gov/TechResources/Human-Genome/project/50yr.html

〔2〕 Lee D., Seo H., Tae H., Nam H., Park K. Development of a Web-based Genome Annotation System. The Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB 2003), Berlin, Germany. *Currents in Computational Molecular Biology 2003* pp. 19-20, 2003.

〔3〕 Seo H., Kim K.-B., Tae H., Park W., and Park K. Development of gene ontology analysis and classification tools for microbial genome annotation. *Currents in Computational Molecular Biology (RECOMB 2002)* Washington p.165, 2002.

〔4〕 Vanet A., Marsan M. Promoter sequences and algorithmical methods for identifying them. *Res. Microbiol.* 150 : 779-199, 1999.

〔5〕 Fickett J.W., Hatzigeorgiou A.G., Eukaryotic promoter recognition, *Genome Research* 7 : 861-878, 1997.

〔6〕 Scherf M., Klingenhoff A., and Werner T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector : a novel context analysis approach. *J Mol Biol.*, 297 : 599-606, 2000.

〔7〕 Uberbacher, E. and Mural, J. Locating protein coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci.* 88 : 11261-11265, 1991.

〔8〕 Borodovsky M. and McIninch, J. "GENMARK : parallel gene recognition for both DNA strands", *Computer & Chemistry*, 17(2) : 123-134, 1993.

〔9〕 Delcher A. L., Harmon D., Kasif S., White, O., and Salzberg S. L. "Improved microbial gene identification with GLIMMER", *Nucl. Acids Res.* 27 (23) : 4636-4641, 1999.

〔10〕 Burge C. and Karlin S. Prediction of Complete Gene Structures in Human Genomic DNA. *J. Mol. Biol.* 268 : 78-94, 1997.

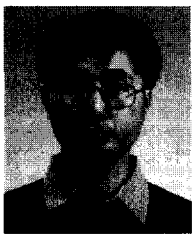
## 저 자 소 개


**이 대 상**

1993년 2월 경북대학교 유전공학과 (이학사), 1995년 2월 한국과학기술원 생명과학과 (이학석사), 2003년 2월 한국과학기술원 생물과학과 (박사수료), 1995년 3월~1999년 7월 : (주)한일합섬 부설 한효과학기술원(연구원), 1999년 10월~2000년 12월 : (주)인바이오넷 (연구원), 2001년 1월~현재 : (주)스몰소프트 (선임연구원), 〈주관심 분야 : Promoter prediction, genome informatics〉

**태 홍 석**

2001년 2월 경북대학교 미생물학과 (이학사), 2003년 2월 경북대학교 미생물학과 (이학석사), 2003년 9월 충남대학교 컴퓨터공학과 (박사과정재학), 2001년 1월~현재 : (주)스몰소프트 (연구원), 〈주관심 분야 : Duration HMM을 이용한 Gene Prediction.〉



**박 기 정**

1986년 2월 서울대학교 컴퓨터공학과 (공학사), 1987년 2월 한국과학기술원 전산학과 (석사과정), 1989년 2월 한국과학기술원 생물과학과 (이학석사), 2002년 2월 한국과학기술원 생물과학과 (이학박사), 1989년 2월~1998년 6월 : 생명공학연구원 (선임연구원), 1998년 6월~2000년 2월 : 한국과학기술원 의과학센터 (연구원), 2000년 3월~현재 : (주)스몰소프트 (대표이사), 〈주관심 분야 : Sequence alignment/multialignment Gene prediction, genome annotation〉