

Introduction to Gene Prediction Using HMM Algorithm¹⁾

Keon-Kyun Kim²⁾ · Eunsik Park³⁾

Abstract

Gene structure prediction, which is to predict protein coding regions in a given nucleotide sequence, is the most important process in annotating genes and greatly affects gene analysis and genome annotation. As eukaryotic genes have more complicated structures in DNA sequences than those of prokaryotic genes, analysis programs for eukaryotic gene structure prediction have more diverse and more complicated computational models. There are Ab Initio method, Similarity-based method, and Ensemble method for gene prediction method for eukaryotic genes. Each Method use various algorithms. This paper introduce how to predict genes using HMM(Hidden Markov Model) algorithm and present the process of gene prediction with well-known gene prediction programs.

Keywords : 유전자 서열 분석, 유전자 예측, HMM

1. 서론

유전체 프로젝트, 즉, 유전자를 해독해 유전자 지도를 작성하고 유전자 배열을 분석하는 작업이 거의 완료되면서 여러 생명체의 유전체에 대한 연구가 활발히 진행되고 그 결과가 데이터베이스에 저장되고 있다. 유전체 프로젝트의 첫 단계라고 할 수 있는 유전체 염기서열 분석의 비율이 증가하면서 유전체 내의 정확한 유전자 위치를 알아내기 위해 많은 유전자 구조 예측 모델들이 개발되었다. 유전자의 위치를 정확하게 밝혀내는 것은 뉴클레오티드 혹은 아미노산 서열에 쓰인 고차원적인 구조와 기능적인

-
- 1) 이 논문은 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(R01-2006-000-11087-0).
 - 2) First Author : Graduate Student, Department of Statistics, Chonnam National University, Gwangju, 500-757, Korea.
E-mail : maryjane1844@naver.com
 - 3) Corresponding Author : Professor, Department of Statistics, Chonnam National University, Gwangju, 500-757, Korea.
E-mail : espark02@chonnam.ac.kr

정보를 밝혀내기 위한 중요한 의미를 지닌다.

생물학에서 사용하는 컴퓨터를 이용한 계산적 방법론들은 물리적으로 존재하지 않는 기본 원리의 방정식을 푸는 것이 아니라 관찰된 데이터로부터 얻는 경험적인 지식을 처리한다. 유전자 서열 분석에 숨어있는 기본적인 아이디어는 문자생물학에서 파생된 경험적인 지식에서 비롯한다. 즉, 두 문자가 유사한 서열을 가지면, 진화적인 관계나 물리 화학적인 제약 때문에 유사한 3차원 구조와 비슷한 생물학적 기능을 갖기 쉽다. 따라서 유전자 서열 분석의 주된 작업은 구조적, 기능적인 속성으로 확장이 가능한 서열 특징을 찾는 것이다.

진핵생물의 유전자는 원핵생물의 유전자보다 구조가 훨씬 복잡하고 유전체 크기에 비해 유전자의 밀도가 훨씬 떨어진다. 원핵생물의 유전자 구조가 프로모터, 캐시코돈, 부호화 영역, 정지코돈, 비부호화 영역등으로 이루어진데 비해 진핵생물의 유전자는 cap, polyA와 같이 전사에 관련된 신호(signal)가 더 존재하며, 부호화 영역도 donor, acceptor signal (Transcription regulation region)에 의해 엑손(exon)과 인트론(intron)으로 나누어진다(태홍석, 박기정, 2003).

진핵생물의 유전자 예측은 크게 세 가지 방법으로 나누어진다. 첫 번째 방법은 염기서열을 통해 유전자를 예측하는 Ab Initio 방법이고(Haussler, 1998) 두 번째 방법은 알려진 단백질의 서열인 EST(Expressed Sequence Tag)와의 유사성을 비교하여 유전자를 예측하는 방법인 Similarity-based 방법이 있다(Gish & States, 1993). 마지막으로 위 두 가지 방법을 통합한 Ensemble 방법이 있다(Haussler, 1998).

본 논문에서는 위 세 가지 방법 중 염기서열을 통해 유전자를 예측하는 방법인 Ab Initio 방법에서 가장 많이 쓰이는 HMM(Hidden Markov Model)을 사용한 유전자 예측 방법을 소개하고, 널리 알려진 HMM을 이용한 유전자 예측 프로그램의 활용법을 제시하고자 한다.

2. 유전자 예측

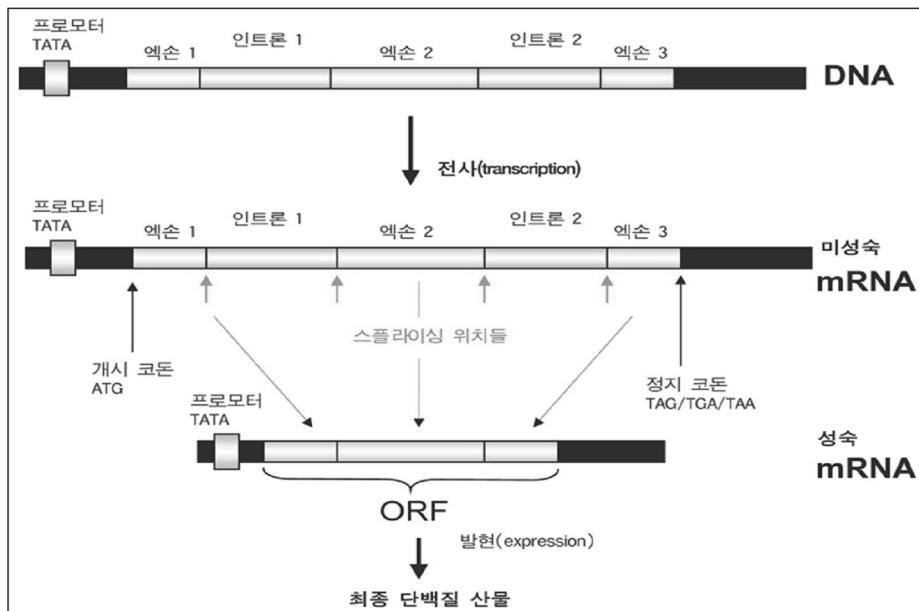
2.1 유전자(Gene)란?

유전자란 한 세대에서부터 다음 세대로 그 개체의 모든 생물학적 정보를 전달해주는 물리적, 기능적인 단위이다. 쉽게 말하자면, 한국인의 피부가 백인이나 흑인과 다르고, 눈동자와 머리색은 까맣다는 특징들이 모두 유전자에 의해서 이루어진다. 즉, 부모에서 자식으로 물려지는 특징, 즉 형질을 만들어 내는 인자로서 유전 정보의 단위의 실체는 생물 세포의 염색체를 구성하는 DNA가 배열된 방식이다. 문자유전학적인 관점에서 살펴보면 유전자의 원천적인 정보는 DNA에 있다고 한다.

DNA에 적혀 있는 유전정보를 mRNA로 옮기는 과정을 전사라고 한다. RNA 중합효소가 이 과정을 맡는다. 기본적으로는 DNA 복제과정 중 한쪽 부분과 유사하나, 전사 과정에서는 한쪽 가닥만을 정보로 삼아 옮겨 적고, RNA가 합성된 이후 DNA는 원상 복구된다.

원핵세포의 경우 전사된 mRNA는 그대로 다음 과정인 번역과정으로 넘어가게 되나, 진핵세포의 경우 중간에 끼어 있는 인트론을 제거하고 엑손만을 남겨야 하므로 만들어진 mRNA를 가공하는 과정을 거친다. 이러한 과정을 거친 후 부호화 영역은 RNA로 변화하며, 이들은 리보솜에서 아미노산을 합성하며, 아미노산이 모여 특정한

역할을 하는 단백질이 되는 것이다(<그림 1>). 따라서, 단백질이 유전자의 최종 산물이며, 인간의 경우 10만여개의 단백질이 어떠한 역할을 하느냐에 따라 그 사람의 형질이 나타나게 되는 것이다. 또한 같은 기능을 하는 단백질이 DNA 서열상의 차이로 인해(이를 SNP: Single Nucleotide Polymorphism) 약간씩 다른 모양과 활성을 가짐으로서 각각 다른 형질을 보이게 된다.



<그림 1> 진핵생물의 유전자구조 (주현, 한진, 2004)

진핵생물의 경우 엑손 부분만이 단백질의 발현 부위에 속하기 때문에, 엑손과 인트론 두 영역을 구별할 수 있는 특별한 방법이 요구된다.

현재까지 알려진 발현 유전자 서열 예측법(혹은 단백질 부호화 영역 판별법) 중 가장 신뢰도가 높은 방법은 Hidden Markov Model(은닉 마코프 알고리즘)을 이용하는 방법이다(주현, 한진, 2004). 이 알고리즘에 대한 더 자세한 설명은 2.3절에서 기술하였다.

2.2 컴퓨터를 이용한 계산적 유전자 서열 분석

인간의 DNA 서열 분석에서 중요한 문제는 유전자의 구조, 조절, 전사에 관련된 요소들의 정보를 담고 있는 기능적인 자리(functional site)를 찾는 것이다. 연구자들이 찾는 요소들은 splice site, 개시코돈, 정지코돈, branch point, 전사에 관여하는 프로모터와 terminator, polyadenylation site, ribosomal binding site, topoisomerase II binding site, topoisomerase I 절단자리와 여러 가지 전사조절인자들의 결합자리 등이다. 이러한 국부적인 위치들을 신호라고 부르며, 이들을 찾아내는 방법은 신호감지장치(signal sensor)라고 부른다. 반면 엑손과 인트론과 같은 확장된, 다양한 길이를 가

지는 영역들은 contents라 불리며 content sensor라고 부르는 방법으로 분석한다.

1) 신호 감지장치

가장 기본적인 신호 감지장치는 간단한 보존 서열(consensus)을 허용 가능한 변이들(allowable variations)로 기술하는 표현이다. 보다 정밀한 감지장치들은 consensus 대신에 가중행렬을 사용하여 고안할 수 있다. 이 때 패턴의 각 위치는 잔기에 대한 일치(match)도 허용하지만 이에 따른 별도 비용이 부과된다. 후보 자리에 가중행렬 감지장치에 의해 회신되는 스코어는 그 자리에 개개의 잔기들이 일치될 때의 비용을 합한 값이다. 이 스코어가 기준치를 상회하면 후보 자리는 일치한다고 예측한다. 회신되는 스코어는 간단한 통계 모델에서 로그우도비에 해당한다. 각 위치는 가능한 잔기들에 걸친 하나의 독립적인 별개의 분포에 의해 결정된다. 신경망(Neural network)보다 정교한 유형의 신호 감지장치들이 많이 사용된다(Haussler, 1998).

2) Content Sensors

가장 중요하면서 연구가 널리 된 content sensor는 부호화 영역을 예측하는 것이다. 원핵생물의 경우 간단하게 긴 ORF(Open Reading frame)를 찾음으로서 유전자들의 위치를 아는 것이 통례이다. 그러나 진핵생물들에게는 이 방법이 적당하지 않다. 진핵생물에서 비부호화 영역들을 부호화 영역들로부터 구분하기 위해서 exon content sensor들은 코돈 구조에 존재하는 뉴클레오티드 빈도와 의존도에 의한 통계적인 모델들을 사용한다. 가장 많이 사용되는 모델은 앞서 언급한 마코프 모델로서 유전자 예측 프로그램인 GeneMark(Borodovsky and McIninch, 1993)로 인하여 유명해지기 시작했다.

여러 단백질 부호화 측정 도구들이 있었지만, 각각 사용될 경우 효과적이지 못하지만, 이들의 조합은 효과적일 수 있는데, 예를 들면 GRAIN exon detector가 그런 경우이다. 다른 content sensor들은 CpG island(유전자의 시작부분에 자주 나타나는 영역들로 dinucleotide CG가 다른 서열에서 보다 빈번하게 나타난다.)들과 사람의 ALU서열들과 같은 반복 DNA에 대한 감지장치들을 포함한다. 후자의 경우는 반복 DNA를 제거하고 나머지 DNA를 분석하기 위한 마스크나 여과장치로 흔히 사용된다(Haussler, 1998).

3) 통합된 유전자예측 프로그램

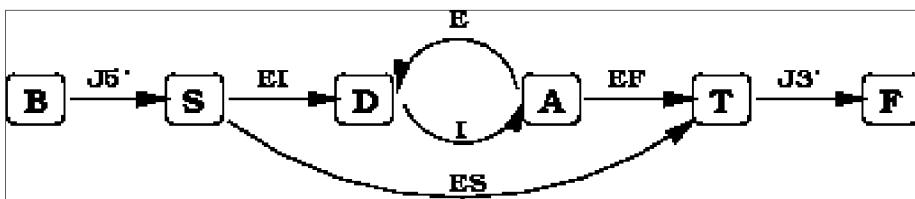
신호감지장치와 content sensor만으로는 유전자 동정이 성공할 수 없다. 이들이 인식하는 통계적인 신호들은 매우 약하고(Agarwal and Bafna, 1998), 신호와 content 사이에는 감지하기 어려운 종속적인 관계가 있는데, splice 자리 강도와 엑손 크기 사이에 존재하는 상관성을 예로 들 수 있다(Zhang, 1998). 지난 5년간 완전한 유전자 구조를 밝히기 위해, 신호와 content들을 취합하는 여러 가지 프로그램이 개발되었다. 이 프로그램들은 유전자 특성을 사이의 보다 복잡한 상호의존도를 다룰 수 있도록 설계되었다. 언어적 은유법을 적용하여, DNA서열을 엑손과 인트론들의 연속인 유전자들로 구분하는 과정은 한 문장을 분석하여 그 구조상 문법적인 부분들로 나누는 과정

과 동일시된다. Searls는 형식적인 문법을 사용하여 언어학적 용어로 유전자 구조를 처음으로 기술하였다(Dong and Searls, 1994). 이 이론에 근거한 GenLang은 가장 초기의 통합된 유전자예측 프로그램이었다. 오늘날 통합된 유전자예측 프로그램들처럼 GenLang은 후보 엑손들과 다른 스코어 영역들과 자리들을 최대 스코어합을 가지는 유전자로 예측을 하기 위해 동적 프로그램 기법을 사용했다. 동적 프로그램 기법에서 성공의 열쇠는 최적화 할 수 있는 올바른 스코어 함수를 개발하는 것이다. 이 방법의 성공은 엑손들에서 코돈의존도에 관련된 변수들을 포함하는 유전자들의 통계적인 모델, splice site들의 특성들, 어떤 기능적인 특성들이 다른 특성들을 뒤따를 것인가에 대한 ‘언어학적’ 법칙을 (<그림 2> 참조) 정의하는 데 있다. 이 모델은 뉴클레오티드의 기능적인 역할이나 위치를 나타내는 각 뉴클레오티드와 연관된 잠재적인 변수를 포함한다. 예를 들어 G 잔기는 GT consensus donor splice site의 부분이거나 개시코돈의 세 번째 위치일 수도 있다. 어떤 기능적인 특징들이 다른 특징들을 뒤따를 것인가에 대한 언어학적 규칙들은 은닉된 변수들에 대한 마코프 과정의 변수들에 의해 표현된다. 이러한 이유 때문에 HMM이라고 불리며 유전자예측 HMM들은 Searls가 사용한 유전자 구조 문법들의 확률론적 해석들로 간주 할 수 있다(Haussler, 1998).

초기의 유전자예측 HMM들은 EcoParse (E. coli 대상) (Krogh et al., 1994)와 Xpound (인간 유전체 대상)(Thomas and Skolnick, 1994)가 있다. 보다 최근의 프로그램들은 GeneMark-HMM (박테리아 유전체 대상)(Lukashin and Borodovsky, 1998), Veil(Henderson et al., 1997), HMMgene (인간유전체 대상)(Krogh, 1997) 등이 있다. Generalized HMMs (GHMMs) 또는 (Hidden) Semi Markov 모델들이라 불리는 보다 일반화된 확률적 모델들은 그 뿌리를 GeneParse(Snyder and Stormo, 1995)에 두고 있으며, Genie(Reese et al., 1997)와 GENSCAN(Burge and Karlin, 1997)에 이르러 보다 완전하게 발전되었다.

유전자예측 프로그램들은 알려진 유전자들과 이에 상응하는 단백질들과의 비교를 이용하기 보다는 유전자들의 일반적인 특성들에 근거하여 유전자 구조를 예측한다. 부수적으로 EST 일치들과 같은 정보에 근거하기도 한다. 단백질 데이터베이스 상동성과 EST 일치들은 유전자 예측을 사후에 정당화하는 방법들로 오랫동안 사용되어 왔다. 그러나 새로운 방법들은 이 정보를 직접 유전자예측 알고리즘 자체에 통합하였다. 최신 유전자예측 프로그램들은 DNA를 모든 가능한 reading frame의 단백질로 번역한 후, 유사한 단백질 서열을 찾아 단백질 데이터베이스를 검색한 후 얻어진 상동성 검색 결과들을 여러 통계적 도구들과 결합하였다.

상동성을 이용한 접근은 Gelfand(1996) 등에 의해 개발된 유전자예측 프로그램에서 극단적인 형태로 이용되었다. Procrustes라 불리는 이 시스템은 사용자로 하여금 예측될 유전자와 가까운 단백질 homolog를 제공하도록 요구한다. 그러면 Smith-Waterman 알고리즘에 유사한 ‘spliced alignment’ 알고리즘이 DNA를 homolog에 배열함으로써 추정하는 유전자구조를 유도하는데 사용된다. 이 방법의 가장 큰 취약점은 예측될 유전자에 가까운 homolog를 요구한다는 것이다. 일반적으로는 homolog들이 알려지지 않았거나 거리가 먼 경우가 종종 있다. 이런 경우는 이 시스템이 부적절할 것이다. 하지만, 매우 가까운 homolog가 있는 경우는 Procrustes는 매우 효과적인 유전자 예측 방법이다.



<그림 2> 다중 엑손 유전자로 구성된 서열을 분석할 때, 어떤 기능적인 특성들이 다른 특성들을 뒤따를 것인가에 대한 언어학적 법칙의 설명 (Kulp et al., 1996)

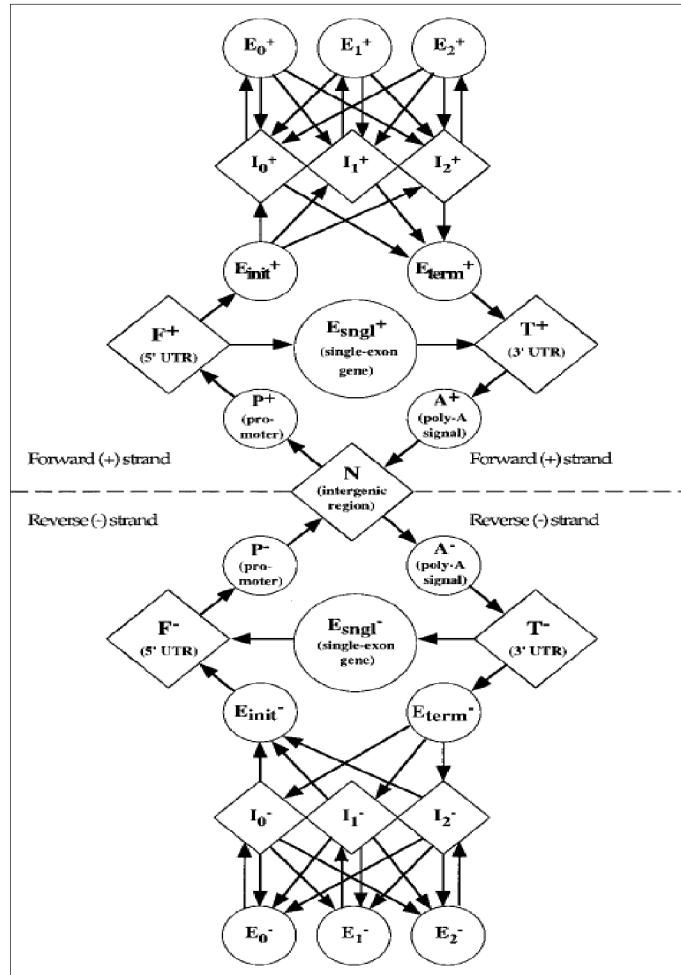
<그림 2>에서 화살표들은 content를 의미하고 노드들은 신호를 의미한다. Content는 5' UTR(J5'), initial exon(EI), exon(E), intron(I), single exon(ES), final exon(EF), 3' UTR(J3')를 나타낸다. 신호들은 begin sequence(B), start translation(S), donor splice site(D), acceptor splice site(A), stop translation(T), end sequence(F)를 나타낸다. Candidate 유전자 구조는 B에서 F로의 경로를 따라서 만들어진다. HMM (또는 hidden semi-Markov model)은 확률론적 모델들을 화살표와 노드들 각각에 적용한다(Kulp et al., 1996).

2.3 진핵생물의 유전자 예측을 위한 HMM

HMM은 초기에 음성 패턴 인식을 위하여 개발되었으나, 그 인식의 정확도와 높은 신뢰도로 인하여 현재는 생물정보학 분야에서 CpG 모티브 분석, 프로모터 결합 부위 탐색, 단백질 부호화 영역 판별 등에 널리 이용되고 있다(주현, 한진, 2004).

HMM은 조건부 확률에 근거해서 바로 이전 상태로부터 현재 상태를 추측해 내는 마코프 연쇄의 확장된 모델이다. HMM은 상태(state)들과 심볼(symbol)들을 기본 요소로 가지고 있으며 관찰되는 심벌들의 출현빈도를 측정해서 조건부 확률에 따라 실제로 가장 근사한 상태들의 구성을 추측해 내는 모델이다. 1990년 초반부터 모티브 도메인 겹침이나 프로모터 예측과 같은 생물학을 위한 연구에서 HMM이 적용되기 시작했다. 특히, DNA 염기서열과 같이 연속적이고 반복되는 형태를 가진 구조에서 HMM을 구성하기가 용이하다.

이러한 유전자예측에 쓰이는 HMM은 계속 변형되어 현재 GENSCAN의 Generalized HMM(GHMM)에 의해 발전되었다. GHMM의 유전자구조 모델은 <그림 3>과 같다.



<그림 3> 유전자 구조 모델 (Burge and Karlin, 1997)

<그림 3> 에서의 은닉 상태는 진핵생물 유전자의 근본적인 기본적인 단위(엑손, 인트론, 유전자간 영역 등)를 나타내며 아래와 같이 정의한 표현들을 사용하여 유전자 구조 모델을 세운다.

N : intergenic region

P : promoter

F : 5'untranslated region (extending from the start of transcription up to the translation initiation signal)

E_{sigl} : single-exon (intronless) gene (translation start \rightarrow stop codon)

E_{init} : initial exon (translation start \rightarrow donor splice site)

$E_k (0 \leq k \leq 2)$: phase k internal exon (acceptor splice site \rightarrow donor splice site)

E_{term} : terminal exon (acceptor splice site \rightarrow stop codon)

- T : 3'untranslated region (extending from just after the stop codon to the polyadenylation signal)
- A : polyadenylation signal
- $I_k (0 \leq k \leq 2)$: phase k intron
- E_k^+ : Forward-strand internal exon(Accept site-> coding region->donar site)
- E_k^- : E_k^+ 의 역순

Semi-Markov 모델의 일종인 이 모델은 Rabiner(1989)가 state duration HMM로 설명하였다. 이 모델은 순서가 있는 상태의 집합 $q = \{q_1, q_2, \dots, q_n\}$ 으로 구성된 “parse” Φ 를 발생한다. 이 때, 이들의 길이 $d = \{d_1, d_2, \dots, d_n\}$ 로 이루어진 집합을 이용하는데, 이는 각 상태 유형들이 따르는 확률 모델들을 사용하여 길이 $L (= \sum_{i=1}^n d_i)$ 의 DNA 서열 S 를 생성한다. 이 서열의 길이 L 에 해당하는 parse의 생성은 다음과 같다.

- 1) 초기 상태 q_1 을 최초 상태의 분포에 따라 선택하며 이때 $\pi_i = P\{q_1 = Q^{(i)}\}$ 가 되며, <그림 3>과 같은 경우 $Q^{(i)} \{i = 1, 2, \dots, 27\}$ 을 갖는다.
- 2) 상태 q_1 에 대응되는 state duration d_1 은 길이 분포 $f_{Q^{(i)}}$ 로부터 생성한다.
- 3) 길이 d_1 의 조각 서열 s_1 은 적절한 서열 모델에 의해 생성한다.
- 4) 1차 마코프 모형에 의해 상태 q_2 의 값은 q_1 에 영향을 받으며 이때의 상태전이 확률은 다음과 같이 나타낸다.

$$T_{ij} = P\{q_{k+1} = Q^{(j)} | q_k = Q^{(i)}\}$$

이러한 과정은 길이 $L = \sum_{i=1}^n d_i$ 까지 반복한다. 그리고 이 모델은 초기확률 $\vec{\pi}$, 상태 전이확률 행렬 T , 길이 분포 f 와 서열 발생 모형의 집합인 P 로 구성된다. 기본적인 아이디어는 유전자 서열과 유사한 확률 모델을 정한 후, 해당서열과 가장 높은 우도를 갖는 것으로 유전자 구조를 결정하는 것이다.

길이가 L 인 서열 S 와 parse ϕ_i 의 결합 확률은 다음과 같다.

$$P\{\phi_i, S\} = \pi_{q1} f_{q1}(d_1) P\{S_i | q_1, d_1\} \times \prod_{k=2}^n T_{q_{k-1}, q_k}(d_k) P\{S_k | q_k, d_k\}$$

이 때, 최적의 parse는 뷔터비 알고리듬에 의해 계산하며 $P(S)$ 는 전전선택법에 의해 계산한다.

3. HMM을 이용한 유전자 예측 프로그램

HMM을 이용한 유전자 예측 프로그램 중에서 가장 널리 쓰이고, 정확도가 높은 몇 개의 프로그램을 선별하여 그 활용법을 제시하고자 한다(Majoros et al., 2004; Stanke et al., 2004). <표 1>에 제시되어 있는 바와 같이 프로그램에 따라 입력 형식과 출력

형식이 동일하지 않고, 프로그램의 이용이 web server, UNIX, 혹은 Linux system 중 하나 혹은 두 가지에서 가능하며, 시스템간 호환성이 없어서 예측 프로그램의 이용에 어려운 점이 있다.

Web server를 이용하는 경우, 서버에 이상이 있을 경우 프로그램을 실행할 수 없는 어려움이 있다. 다른 미리 사이트들도 모두 동일한 주소에 연결되어 있어 대치하여 사용할 수 있는 Web 주소가 없다. Linux 환경에서 실행 가능한 프로그램은 프로그램을 다운로드하여 설치하여야 하며, 어려운 점은 Linux 사용 환경에 익숙해야 하고, 다운로드된 버전의 새로운 사용법을 익혀야 한다는 것이다.

본 논문은 유전자 예측 프로그램의 활용법에 관하여 다룸으로써, 이 분야에 관한 연구를 시작하는 연구자들에게 가이드를 제공하고자 한다.

<표 1> HMM을 이용한 유전자 예측 프로그램

Program (Website)	Organism	알고리즘	입력형식/ 출력형식	방식
HMMgene (http://www.cbs.dtu.dk/services/HMMgene/)	Vertebrates. <i>C.elegans</i>	GHMM	FASTA / GFF	Web server
GENSCAN (http://genes.mit.edu/GSCAN.html)	Vertebrates. <i>Arabidopsis</i> . <i>maize</i> .	GHMM	FASTA / web	Web server
TWINSCAN (http://mblab.wustl.edu/queries.html)	all species	GHMM	FASTA / web, GTF	Web server
GeneZilla (http://www.genezilla.org/)	Human. <i>Arabidopsis</i> . Vertebrates.	GHMM	FASTA / GFF	UNIX/Linux (C++)
AUGUSTUS+ (http://augustus.gobics.de/)	all species	GHMM	FASTA / GTF	Web server UNIX/Linux

3.1 HMMgene

1) 개요

HMMgene은 정확한 예측을 위하여 흔히 쓰이는 1차 마코프모델이 아니라 4차 Inhomogeneous Markov Model을 이용하여 주어진 DNA서열 내의 전체 유전자들을

예측하는 프로그램이다. 또한 상태 서열을 예측할 때 주로 쓰는 비터비 알고리즘을 개선하여 효율성을 높였다. 또한 같은 부위에 대한 복수 예측을 해줌으로써 단일 유전자가 포함된 한 부위에서 일어날 수 있는 가능한 많은 집합을 볼 수 있게 되므로 좀 더 정확한 예측이 가능해진다.

2) HMMgene의 활용

① 형식 : Web server

② URL : <http://www.cbs.dtu.dk/services/HMMgene/>

③ 사용방법

FASTA형식으로 이루어진 유전자 서열 파일을 직접 업로드하거나 직접 web server에 입력하여 예측할 유전자가 속하는 올바른 유기체(organism)를 선택하여 Submit버튼을 클릭한다. 클릭 후 자신의 e-mail계정을 입력하면 이 e-mail계정으로 GFF형식의 예측된 결과가 있는 url을 알려준다. 옵션으로는 예측할 유전자 서열의 일부분을 알고 있을 때 이를 입력하여 그 부분을 제외한 나머지 부분을 예측할 수 있다.

④ 출력결과

HMMgene의 출력결과는 GFF형식으로 이루어져 있다. 이는 매우 단순하여 이 출력결과를 이용하는 perl이나 awk로 개발된 프로그램에 적용하기가 쉽게 되어 있다.

<그림 4> HMMgene의 초기화면

HMMgene result

Explanation of [output format](#)

```

## gff-version 1
## date: Sun Apr 22 09:04:27 2007
## HMMgene1.1a (human model sim10gc.D.bsmod)

# SEQ: HUMDZA2G 14694 (+) A:1833 C:1640 G:1857 T:1599
HUMDZA2G HMMgene1.1a firstex 5313 5388 0.637 + 1 bestparse:cds_1
HUMDZA2G HMMgene1.1a exon_1 9329 9589 0.848 + 1 bestparse:cds_1
HUMDZA2G HMMgene1.1a exon_2 12907 13182 0.602 + 1 bestparse:cds_1
HUMDZA2G HMMgene1.1a lastex 14052 14335 1.001 + 0 bestparse:cds_1
HUMDZA2G HMMgene1.1a CDS 5313 14335 0.324 + . bestparse:cds_1
# SEQ: HUMDZA2G 14694 (-) A:1599 C:1857 G:1640 T:1833

```

<그림 5> HMMgene의 출력결과의 예

3.2 GENSCAN

1) 개요

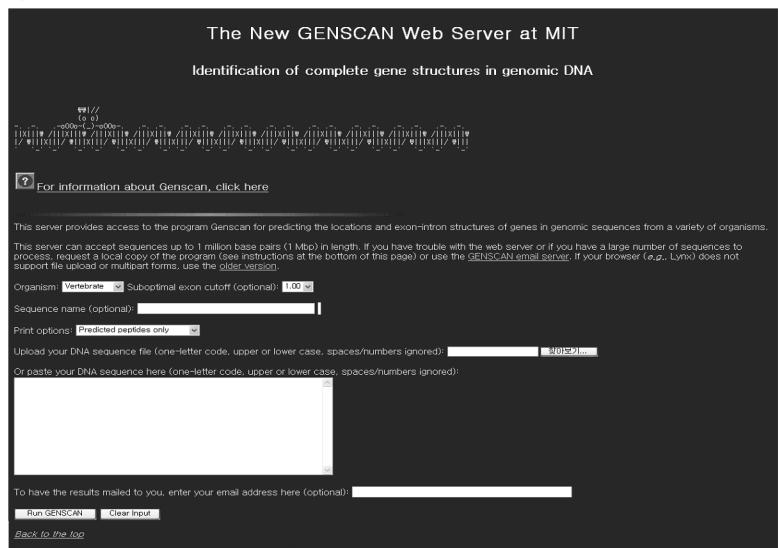
기존의 유전자 서열에서 전체적인 유전자 구조(삽입가능한 엑손들의 집합)를 예측하는 응용 프로그램들은 두 가지 중요한 한계를 가지고 있었다. 그 하나는 입력 서열이 정확하게 하나의 완전한 유전자(부분적으로 유전자의 일부이거나 여러 유전자의 집합이 아닌)이어야 한다는 가정을 해야 했고, 또 한 가지는 독립적인 대조용 유전자들로 측정한 정확성이 처음 생각했던 것보다 현저하게 낮다는 것이다. GENSCAN은 기본적으로 Generalized HMM학률모델을 사용하는 방법들과 유사하지만 위의 한계가 있는 다른 기법들에 비해 몇 가지 차이점을 가지고 있다. 첫째, 이중가닥 유전자 서열 모델을 사용하였다는 것인데 이것은 직접적으로 이중가닥 DNA를 동시에 분석함으로써 가능하다. 둘째, 기존의 방법들이 정확하게 하나의 complete gene이 존재한다고 가정하여 그것에 대해서만 가능했지만, GENSCAN에서 고려된 학률모델은 입력서열이 부분적으로 유전자의 일부인 경우 뿐만 아니라 여러 유전자의 집합에 대해서도 분석이 가능했다. 셋째, MDD (Maximal Dependence Decomposition)를 이용하여 DNA 서열에서 기능적 신호를 모델화 하였다. 이것은 서열의 위치에 따라 핵산들 사이의 의존성이 있음을 가정한 모델이다.

GENSCAN은 엑손, 인트론 그리고 유전자간 영역의 길이의 분포 및 조합의 특징과 함께 전사, 번역, 삽입 신호의 상세 내용을 결합한 유전자 구조를 일반적 학률 모델로 설명하고, 이를 통해 유전체 DNA에서 유전자의 정확한 엑손/인트론 구조를 설명할 수 있다. DNA 서열이 입력으로 주어지면 유전자의 구조적, 구성적인 특성에 관한 학률 모델을 기반으로 가장 잘 분석된 유전자 구조와 근접하는 영역들을 찾아낸다. 이 때 엑손/유전자 이라 추정되는 영역과 더불어 엑손이 전사와 번역을 거쳐서 단백질 서열을 구성 할 때 예상되는 서열을 출력한다. 제약사항으로는 단백질 암호를 갖고 있는 유전자에 대해서만 고려되어지며, 전사단위들이 중복되지 않는다고 가정한다는 것이다.

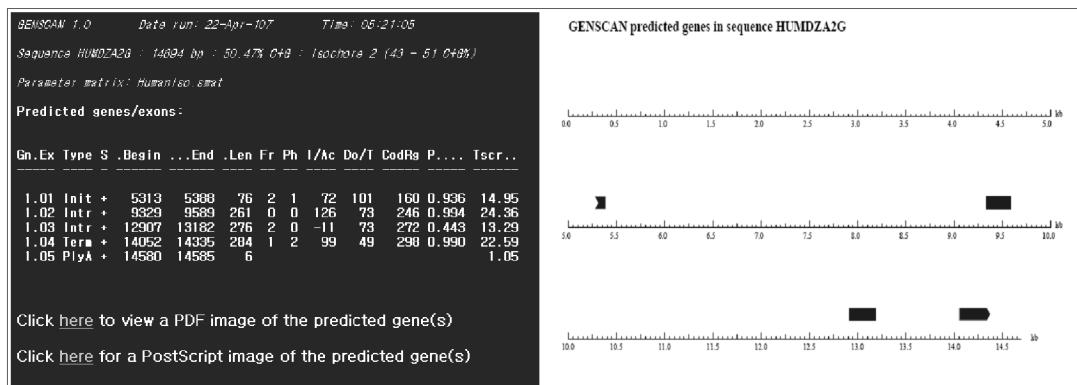
2) GENSCAN의 활용

- ① 형식 : Web server
- ② URL : <http://genes.mit.edu/GENSCAN.html>
- ③ 사용방법

GENSCAN은 HMMgene과 마찬가지로 web server에서 활용할 수 있다. HMMgene과 다른 점이 있다면 GENSCAN에서는 HMM에 의해서 분석과정을 거치는데 그 결과로 나오는 확률값들의 절사값 이상이 되는 엑손들을 모두 출력시키게 된다. 이 옵션은 GENSCAN에 의해서 찾을 수 없었던 엑손들을 찾거나 GENSCAN 자체가 중복을 허용하지 않는 제약을 가지고 있으므로 중복된 엑손 영역들을 찾고자 할 때 사용한다. 또한 GENSCAN은 포스트스크립트 파일을 출력하여 예측된 서열들의 최적의 엑손과 인트론, 차선의 엑손의 정보를 그래픽으로 보여준다.



<그림 6> GENSCAN의 초기화면



<그림 7> GENSCAN의 출력결과의 예

④ 출력결과

GENSCAN의 출력결과는 web server에서 바로 확인할 수 있다. GENSCAN은 또한 예측된 정보뿐 아니라 예측된 유전자의 단백질 서열도 확인할 수 있다. 또한 예측된 서열에 대해 그래픽 정보를 PDF파일로 볼 수 있다.

3.3 TWINSCAN

1) 개요

TWINSCAN이란 엑손/인트론 영역 판별을 위하여 고안된 도구로서 GENSCAN을 확장한 것이다. TWINSCAN은 유전자의 개수를 알 수 없는 대용량 자료를 분석하기 위하여 특별히 고안되었다. 특징은 100~200kb의 작은 크기를 가지는 유전체 서열의 판독률을 높이기 위하여 적계유전자끼리의 정렬 결과를 엑손/인트론 판별에 사용될 HMM의 입력 값으로 사용한다(주현, 한진, 2004). 즉, 입력서열과 이 서열과 가장 밀접한 관계를 가지는 서열, 즉 "Infomat"을 정렬하여 새로운 보존서열을 만들어 유전자를 예측한다.

2) TWINSCAN의 활용

① 형식 : Web server, download for UNIX/Linux

② URL : <http://mblab.wustl.edu/query.html>

③ 사용방법 (Web server)

유전자 예측 프로그램의 web server로 예측하는 방식은 매우 유사하다. TWINSCAN 역시 FASTA형식의 예측할 서열을 직접 입력하거나 파일을 선택하여 유기체를 선택 후 실행 한다. 유기체가 선택될 때 자동으로 Informat 유전자가 선택되어 입력된 서열과 정렬을 실행하게 된다.

④ 출력결과

TWINSCAN의 출력결과는 web server 혹은 e-mail로 확인 할 수 있다. TWINSCAN은 GENSCAN과 비슷한 출력결과를 보여주나 e-mail로 첨부된 파일은 GTF형식으로 확인할 수 있다.

Please enter your e-mail address: (Required)

Organism: (Required) Check existing TWINSCAN human, mouse and rat annotations on the UCSC browser.
 Human Mouse Rat

You can either upload a text file or cut and paste your sequence into the box below.

Case Sensitivity: Upper vs. lower case has no meaning Sequence & lower case masked
Low Complexity Regions: Do not mark low complexity regions Mark low complexity regions

Options

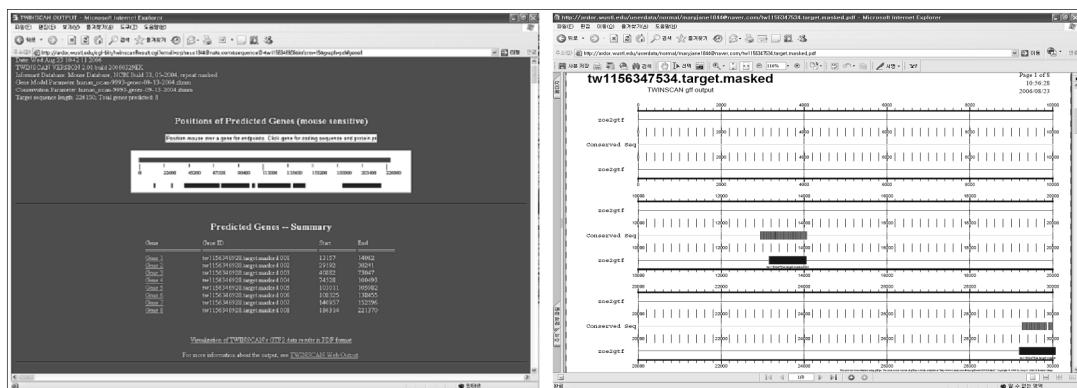
Conservation sequence Coding sequence prediction Protein prediction Graphical output:

Based on the organism you selected above for the target sequence, TWINSCAN will automatically determine the recommended genome against which to compare your sequence. Enabled, the recommended choice is shown below.

Informant Genome:

Please email questions and comments to twinscan@cs.wustl.edu

<그림 8> TWINSCAN의 초기화면



<그림 9> TWINSCAN의 출력결과의 예

3.4 GeneZilla

1) 개요

GeneZilla는 다른 유전자 예측 프로그램과 마찬가지로 GHMM을 기초로 만들어 졌다. GeneZilla는 사용자에 의해 재설정하거나 재훈련 할 수 있는 소프트웨어를 포함하고 있다. 이러한 이유로 GeneZilla는 여타 프로그램과는 다르게 UNIX/Linux 기반의 C++로 작성되었다. 실행속도와 메모리요구량은 유전자서열 길이에 비례하며, 자체 개발한 암호해독 알고리즘을 이용하여 다른 프로그램보다는 일반적으로 예측속도를 단축할 수 있다. GeneZilla의 기본적 모델 기법은 GlimmerHMM과 유사하며, GHMM의

모형에 TATA box state, signal peptides, branch point, CAP sites, polyadenylation signal state, UTR state를 추가한다(Jonathan et al., 2006).

2) GeneZilla의 활용

- ① 형식 : UNIX/Linux
- ② URL : <http://www.genezilla.org/> (다운로드 배포 사이트)
- gcc version 3.3.3이상 (컴파일)
- ③ 구성파일 : genezilla.tar.gz (소스파일, 매뉴얼, 인간과 척추동물 파라메타 파일)
 그 외 각 유기체에 대한 파라메타 파일
- ④ 사용방법 : genezilla.tar.gz의 압축을 풀고 소스코드를 컴파일 한다. 커맨드 라인
 상에서 다음과 같이 사용된다.

```
>mkdir obj (컴파일 할 폴더 생성)
>make genezilla (genezilla로 컴파일, genezilla 실행파일 생성)
GeneZilla는 아래와 같은 커맨드에 의해 실행된다.
>genezilla <*.iso> <*.fasta>
여기서, *.iso는 파라메타파일의 이름이며 예측하고자 하는 유전자가 속한 유기
체를 선택하면 된다. 그 뒤에 바로 FASTA 형식의 유전자서열을 선택하면 된다.
```
- ⑤ 출력결과
GeneZilla의 출력결과는 위 명령을 실행하면 바로 확인할 수 있으며, 여러 유전
자 서열의 FASTA파일을 올리면 GeneZilla는 첫 번째 서열만 예측이 가능하다.

617264 GeneZilla initial-exon 1432 1567 . + . transcript_id=1
617264 GeneZilla internal-exon 1590 1812 . + . transcript_id=1
617264 GeneZilla final-exon 2276 2571 . + . transcript_id=1

<그림 10> GeneZilla의 출력결과의 예

3.5 AUGUSTUS+

1) 개요

AUGUSTUS+는 유전자 예측의 정확도를 높이기 위하여 유전체간 비교, EST나 단백질의 정렬과 같은 외적인 증거를 활용하는데 특징이 있다. 외적인 정보는 서열 자체의 내적 증거와 조화를 이루어야 하는데, 예를 들면 강력한 내적인 증거와 일치하지 않는 외적인 증거를 제외하는 것을 GHMM을 이용한 확률적 모형을 도입하여 해결하였다. AUGUSTUS+는 GeneZilla와 마찬가지로 사용자에 의해 재훈련 될 수 있는 소프트웨어를 포함하고 있다(Stanke et al., 2004).

또한 AUGUSTUS+는 선택적으로 삽입과 전사를 하는 유전자도 예측가능하다. 선택적 삽입은 HMM의 가장 큰 문제점으로 피보나치 수열 증가에 해당하는 다양성을 지니고 있어 예측 불가능한 어려움이 있다. 자연계에서 생물체는 단백질의 다양성을 증가시키기 위하여 이를 선택적 삽입을 이용한다. 현재까지 알려진 질병만 하더라도 40여 가지 이상의 주요 질병들이 이 선택적 삽입에 의하여 발생된다는 점에서 이를 예측하는 문제는 중요하다(주현, 한진, 2004). 마지막으로 AUGUSTUS+는 하나 이

상의 유전자로 구성된 입력 서열에 대해 각각의 유전자 별로 예측이 가능하다.

2) AUGUSTUS+의 활용

① 형식 : Webserver, download for UNIX/Linux

② URL : <http://augustus.gobics.de/>

③ 사용방법(Web server)

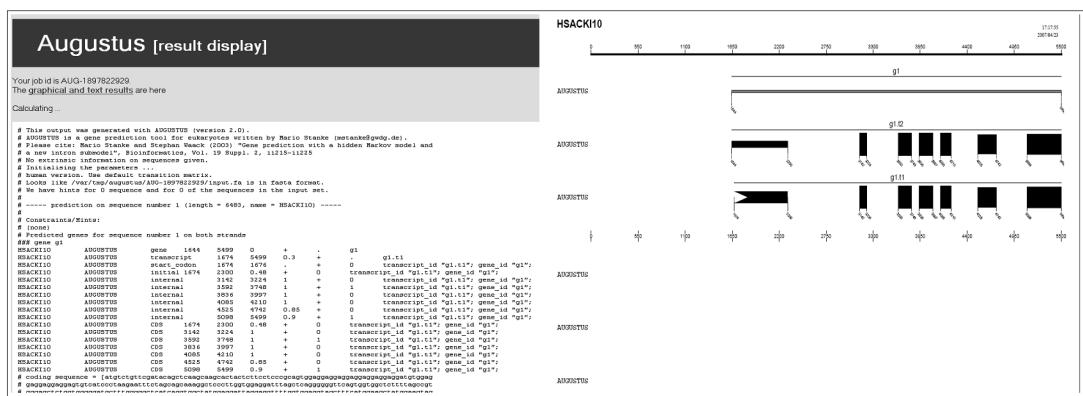
AUGUSTUS+ 역시 사용 방법은 간단하다. 예측하고자 하는 FASTA 형식의 파일을 직접 입력 혹은 입력 파일을 선택하여 실행시키면 된다.

④ 출력결과

AUGUSTUS+의 출력결과는 Web server에서 바로 확인할 수 있으며 출력형식은 GTF 형식이다. 또한 그래프로서 정보를 확인할 수 있다.



<그림 11> AUGUSTUS의 초기화면



<그림 12> AUGUSTUS+의 출력결과의 예

4. 결론 및 토의

지금까지 생명정보학 분야에서 중요한 영역을 차지하는 서열분석과 특정기능과 관련된 자리와 영역을 동정하는 방법에 대해서 살펴보았다. 또한 HMM을 이용한 유전자 예측 프로그램 중 현재까지 가장 널리 활용되고, 성능이 우수한 것을 선택하여 소개하였다. 생물학적 지식이 충분하지 않아 각각의 프로그램에 대해 기본 파라메타를 수정하거나, 다른 모델을 세우는 것과 같은 시도는 해 볼 수 없었으나. 제공되는 형식과 사용법은 간략하게 설명하였다. 서열 분석을 시작하는 사람들에게 도움이 되었으면 한다.

HMM에 기반한 프로그램 뿐만 아니라 여러 다른 기법을 사용하는 프로그램을 함께 평가하여, 예측의 정확도가 낮은 유전자에 대해 확률/통계적인 방법의 효율적인 이용에 관한 연구가 필요하다. 또한, 근본적으로 생명의 현상과 신비를 이해하기 위해서는 DNA레벨 뿐 아니라 단백질의 구조와 기능을 예측하고 동정하는 과정을 필요로 한다. 수학적, 계산적 방법론의 발달로 적합한 모델을 구축하는 관련분야의 연구가 촉진될 것이 기대된다.

참고 문헌

1. 주현, 한진 (2004). 프로테오믹스-단백질에 대한 이해 및 기능 해석의 새로운 접근과 응용, 범문사.
2. 태홍석, 박기정 (2003). Duration HMM을 이용한 진핵생물 유전자 예측 프로그램 개발, *The Korean Journal of Microbiology*, 39, 207-215.
3. Agarwal, P. and Bafna, V. (1998). The ribosome scanning model for translation initiation for gene prediction and full-length cDNA detection. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, 2-7.
4. Allen, J. E., Majoros, W. H., Pertea, M., and Salzberg, S. L. (2006). JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. *Genome Biology*, 7, suppl 1:59.
5. Borodovsky, M. and McIninch, J. (1993). GenMark: parallel gene recognition for both DNA strands. *Journal of Computational Chemistry*, 17, 123-133.
6. Burge, C. and Karlin, S. (1997). Predictions of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268, 78-94.
7. Dong, S. and Searls, D. B. (1994). Gene structure prediction by linguistic methods. *Genomics*, 162, 705-708.
8. Gelfand, M. S., Mironov, A. A. and Pevzner, P. A. (1996). Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences*, 93, 9061-9066.
9. Gish, W. and States, D. J. (1993). Identification of protein-coding regions by database similarity search. *Nature Genetics*, 3, 266-272.

10. Haussler, D. (1998). Computational genefinding. Trends in Biochemical Sciences, *Supplementary Guide to Bioinformatics*, 23, 12–15.
11. Henderson, J., Salzberg, S. and Fasman, K. (1997). Finding genes in human DNA with a hidden Markov model. *Journal of Computational Biology*, 4, 119–126.
12. Krogh, A., Mian, I. S. and Haussler, D. (1994). A Hidden Markov Model that finds genes in *E. coli* DNA. *Nucleic Acids Research*, 22, 4768–4778.
13. Krogh, A. (1997). Two methods for improving performance of an HMM and their application for gene finding. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, 5, 179 – 186.
14. Krogh, A. (1998). *Gene finding: putting the parts together. Guide to Human Genome Computing*, chapter 11, 261–274. Academic Press, 2nd edition
15. Kulp, D., Haussler, D., Reese, M. and Eeckman, F. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. In *ISMB-96*, 134–142.
16. Larimer, F., Mural, R., Shah, M., Subramanian, A., and Uberbacher, E. C. (1998). Microbial GRAIL Gene-Finding Systems, *ASM Conference on Small Genomes*, September
17. Lukashin, A. V. and Borodovsky, M. (1998). Genemark.hmm: new solutions for gene finding. *Nucleic Acids Research*, 26, 1107–1115.
18. Majoros, W.H., Pertea, M., Salzberg, S.L. (2004). TigrScan and Glimmer-HMM: two open source ab initio eukaryotic gene-finders, *Bioinformatics*, 20, 2878–2879.
19. Rabiner, L. R. (1990). A tutorial on hidden Markov models and selected applications in Speech Recognition. *Proceeding of the IEEE*, 77, 257–286.
20. Reese, M. G., Eeckman, F. H., Kulp, D. and Haussler, D. (1997). Improved splice site detection in genie. *Journal of Computational Biology*, 4, 311–323.
21. Snyder, E. and Stormo, G. (1995). Identification of protein coding regions in genomic DNA. *Journal of Molecular Biology*, 248, 1–18.
22. Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. (2004). AUSTUS: a web server for gene finding in eukaryotes, *Nucleic Acids Research*, 32, 309–312.
23. Thomas, A. and Skolnick, M. (1994). A probabilistic model for detecting coding regions in DNA sequences. *IMA Journal of Mathematics Applied in Medicine and Biology*, 11, 149–160.
24. Zhang, M. Q. (1998). Statistical features of human exons and their flanking regions. *Human Molecular Genetics*, 7, 919–932.

[2007년 4월 접수, 2007년 5월 채택]