유전자 구조분석을 위한 계산학적 방법

충남대학교 공은배*

1. 서 론

Human Genome Project를 통해 30억 base pairs(bp)에 이르는 인간의 염색체 서열해독이 거의 완성단계에 있다. 이 서열을 분석하여 서열의 어느 부분이 어떤 단백질을 코딩하고 있고, 그로부터 그 단백질의 구조와 기능을 알아내야 할 것이다 인간 genome의 90% 정도는 유전정보를 갖고있지 않는 non-coding 부분이고, 약 10% 만이 단백질을 지정하는 유전정보를 갖고 있는 coding 부분으로 알려져 있다. non-coding 부분의 기능은 아직 정확히 알려지지는 않았지만 단백질 합성을 조절하는 것 등과 같은 중요한 기능을 수행하리라 생각된다. 서열분석의 첫 번째 단계로써 단백질을 코딩하는 부분, 즉 유전자를 알아내는 것은 매우 중요하다.

염색체 DNA 서열에서 계산학적인 방법으로 유전자를 찾아내려는 연구가 최근 활발히 진행되고 있다[1~5]. 30억 bp나 되는 서열을 실험적인 방법으로 분석하여 유전자를 알아내는 것은 시간과 비용 면에서 비현실적이다. 자동이든 반자동이든 계산학적인 방법을 활용하여, 유전자일 가능성이 높은 부분을 알아내어 이를 실험적으로 확인해나가는 방법이 유일한 대안이 될 것이다

계산학적인 방법은 통계적인 방법과 진화상의 상동 관계(homology)까지를 고려한 방법으로 대별할 수 있다. 통계적 방법에서는 단백질 발현에 관여하는 프로모터, splice sites 등과 같은 여러 시그널간의 규칙성을 이용하여 유전자 모델을 만들고, 새로운 DNA sequence가 주어지면 이 모델을 바탕으로 분석하여 새로운 유전자를 찾아낸다 본 논문에서는 DNA sequence를 분석하여 새로운 유전자를 찾아내는 계산학적인 방법중 주로 통계적 방법 중심으로 이 분야의 최근 발전을 살펴보고자 한다.

본 논문은 다음과 같이 구성되었다. 2장에서는 단백질 발현에 관한 기초적인 생물학적 지식을 검토한다. 3장에서는 개개의 exon을 찾아내는 간단한 계산학적인 방법들을 살펴보고, 4장에서는 다양한 기능적 시그널 부위를 밝혀내고 개개의 exon을 조합하여 완벽한 유전자 구조 모델을 만들어 내는 방법을 살펴본다. 마지막으로, 5장에서는 앞으로의 연구방향을 간략히 알아본다.

2. 생물학적 기초

이 장에서는 유전자의 구조분석을 이해하기 위해 필요한 생물학적 기초 지식을 검토하겠다[6~9]

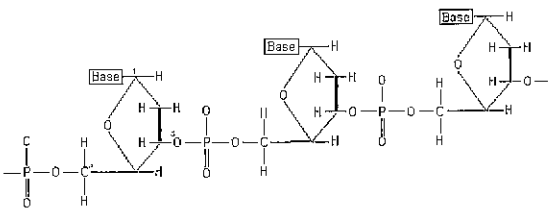

그림 1 DNA 사슬의 구조

2.1 단백질

단백질은 우리 몸의 가장 많은 부분을 차지하는 물질로서 조직과 같은 구조적 역할이나 생화

*종신회원

학 반응에서 촉매의 역할과 같은 중요한 기능을 수행한다. 생명을 유지하기 위해서는 매우 다양한 단백질을 만들어 낼 수 있어야 한다. 단백질은 20가지의 아미노산의 펩티드 결합에 의해 연결된 폴리펩티드이다. 긴 단백질은 4500개나 되는 아미노산으로 구성되는 경우도 있어, 가능한 전체 단백질의 숫자는 2^{4500} 즉 10^{5850}으로 매우 크다. 또한, 단백질은 3차원 상에서 특정한 모습으로 접히는데, 이 모습이 단백질의 화학적인 특징을 결정짓는다. 단백질의 3차원 모습은 1차원의 아미노산 서열로부터 완벽하게 결정지을 수 있는데, 어떻게 결정되는지에 관한 자세한 내용은 아직 해결되지 않은 생물학의 중요한 문제 중의 하나로 남아있다(protein folding problem) [10].

2.2 DNA

DNA도 단백질과 마찬가지로 단순한 분자들의 체인이다. DNA는 잘 알려진 바와 같이 두 줄의 strand로 이중 나선 형태를 취하고 있다. 체인의 각 줄은 변하지 않는 부분인 backbone과 변화가 있는 base로 구성되어 있다. Backbone은 phosphate group과 연결된 deoxyribose로 이루어진다. Deoxyribose는 1′부터 5′까지 번호가 붙은 5개의 탄소를 갖고 있는데, backbone은 한 unit의 3′ 탄소와 그 다음 unit의 5′ 탄소간의 phosphodiester bridge에 의해서 만들어진다. DNA의 변하는 부분은 1′ 탄소에 연결된 A(아데닌), T(티민), G(구아닌), C(시토신) 4개의 뉴클레오티드 base의 서열이다. 그림 1은 DNA 체인의 구조를 보여주고 있다.

DNA 체인은 방향성을 갖는다. 체인에는 다른 뉴클레오티드에 연결되지 않은 양끝이 있는데, 한쪽 끝에는 5′ 탄소가 다른 쪽 끝에는 3′ 탄소가 있다(그림 1 참조). 따라서, base 서열은 항상 5′ --> 3′ 방향으로 쓰여진다.

앞서 말한바와 같이 DNA는 두 줄의 strand이다. 한 strand의 각 base는 다른 strand의 complementary base와 쌍을 이룬다. 뉴클레오티드 A는 T와, C는 G와 항상 complementary base pair를 이룬다. 이런 점에서 DNA는 A, T, C, G를 알파벳으로 갖는 string으로 간주될 수 있다. 각각의 strand는 자신의 방향성을 유지하는데, 두 방향은 서로 반대이다(그림 2 참조).

```
5′ ... TACTGAA ... 3′
3′ ... ATGACTT ... 5′
```

그림 2 DNA 사슬의 방향성

DNA의 주요한 역할은 단백질의 구조에 관한 청사진을 전달하는 것이다 3개의 뉴클레오티드 한 조를 codon이라 하고, 각각의 codon은 하나의 아미노산을 나타낸다. 64개의 서로 다른 codon이 있을 수 있는데, 아미노산은 20가지만 있으므로 다른 codon이 같은 아미노산을 나타내는 경우가 있다. 3개의 codon(UGA, UAG, UAA)은 아미노산을 나타내지 않고 유전자의 끝(stop codon)을 나타내는데 사용된다.

2.3 중심원리

DNA에 담겨있는 정보를 이용하여 어떻게 단백질이 만들어지는가? 이 절에서는 유전자의 구조를 중심으로 DNA 정보가 세포 내에서 단백질 합성에 활용되는 과정을 간략히 살펴보겠다. DNA 서열에서 유전자가 시작되는 부위 앞에는 프로모터라는 일정한 시그널이 있어 앞으로 유전자를 coding하는 부위가 올 것을 미리 알린다. 유전자의 시작부위를 인식한 후, 유전자의 복사본이 RNA로 만들어진다. 이 과정을 전사(transcription)라 한다(그림 3 참조). 이 과정에서 T는 U(우라실)로 바뀐다.

이미 알고 있는 바와 같이 DNA는 double-strand이고 RNA는 single-strand이다 RNA를 만들기 위해서는 DNA의 두 strand 중 하나를 선택하여야 하는데, 항상 5′에서 3′ 방향으로 프로모터를 인식하고 RNA를 5′에서 3′ 방향으로 만든다. 그렇다고 유전자가 항상 같은 strand에서 나와야 한다는 것은 아니다. 예를 들어, 유전자 A는 한쪽 strand에서 유전자 B는 다른 쪽 strand에서 나올 수 있다.

유핵생물(Eukaryotes)의 경우 유전자는 intron과 exon이 번갈아 반복되는 경우가 많다. Intron은 단백질을 코딩하지 않는 부분이고, exon은 단백질 합성에 필요한 정보를 코딩하는 부분이다. 전사가 끝난 후 intron 부분은 떨어져

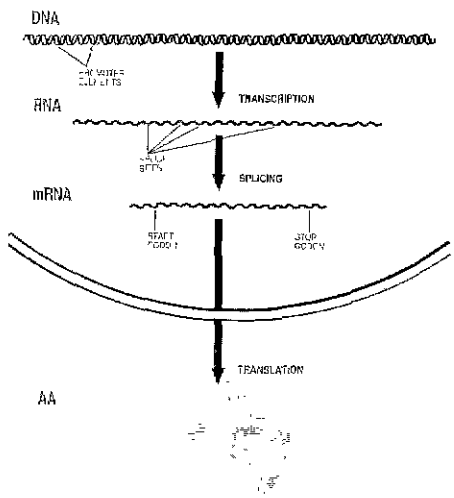

그림 3 중심원리

가고 exon 부분만 남아 mRNA가 된다(그림 3 참조).

리보좀은 mRNA의 정보를 이용하여 단백질을 만든다. 이 과정을 번역(translation)이라 한다. 전사, 번역 등의 과정을 중심원리(central dogma)라 한다.

프로모터라는 서열이 유전자의 시작을 알리는 것과 마찬가지로 유전자의 끝에는 전사과정의 종료를 알리는 일정한 시그널이 있다. 이 시그널을 polyadenylation 시그널이라 한다.

3. 코딩영역 발견

세 개의 base 한 조가 모여 codon을 이루기 때문에, DNA 서열은 코딩을 어디서부터 시작하는가에 따라 세 가지로 해석될 수 있다. 이 세 가지로 묶는 방식을 reading frame이라 한다. 일반적으로, 같은 서열이라도 다른 reading frame에서는 전혀 다른 단백질을 코딩할 것이다. Open Reading Frame(ORF)란 stop codon이 없는 codon의 sequence를 말한다.

이 장에서는 reading frame과 코딩영역을 알아내는 간단한 방법들을 살펴보겠다.

3.1 긴 ORF를 찾는 방법

코딩영역과 비 코딩영역을 구별하는 하나의 방법은 stop codon의 빈도를 살펴보는 것이다. 모든 codon이 균등하게 사용된다고 가정하면 21개의 codon마다 한 개의 stop codon이 있어야한다. 왜냐하면 64개의 가능한 codon중 세 개의 다른 stop codon이 있기 때문이다. 그러나 정상적인 단백질은 보통 1000 base이상의 길이를 갖는다. 그리고 각각의 코딩영역은 끝에 한 개의 stop codon이 있을 뿐이다. 그러므로 코딩영역의 특징 중 하나는 stop codon이 없는 긴 codon의 sequence, 즉 긴 ORF, 라는 것이다.

이런 특징을 이용하여 세 가지의 reading frame으로 DNA 서열을 살펴보아, 그 중 긴 ORF를 찾는다. 이 방법으로는 짧은 코딩영역을 찾아낼 수 없다.

3.2 codon의 규칙성

코딩영역을 찾아내는 또 다른 방법은 codon의 사용빈도를 살펴보는 것이다. 예를 들어, Leucine, Alanine, Tryptophan은 6, 4, 1개의 다른 codon에 의해 각각 코딩된다. DNA 서열에서 codon이 균등하게 사용된다면 Leu, Ala, Try는 6:4:1의 비율로 관찰되어야한다. 그러나 실제로는 단백질에서 6.9:6.5:1의 비율로 관찰된다. 즉, 코딩영역에서 codon은 무작위로 균등하게 사용되지 않는다. 또, A나 T는 코딩영역 codon의 세 번째 자리에 90% 정도 나타나는데 C나 G는 10%만이 나타난다. 이런 특성을 이용하여 코딩영역을 알아낼 수 있다.

GenBank 등과 같은 데이터베이스에서 유전자를 실제로 코딩하는데 사용된 codon의 빈도통계를 이용하여 reading frame을 다음과 같이 결정할 수 있다. Reading frame을 모르는 코딩 서열 $a_1, b_1, c_1, a_2, b_2, c_2, \ldots, a_{n+1}, b_{n+1}, c_{n+1}$이 있다고 하자. 어떤 codon abc의 빈도를 f_{abc}로 나타낸다. codon과 그 다음의 codon이 데이터베이스에서 조사된 빈도 f_{abc}를 따라 독립적으로 이어진다는 가정 하에 reading frame이 $a_1b_1c_1$으로 시작할 비율을 다음과 같이 계산한다.

$$r_1 = f_{a_1 b_1 c_1} * f_{a_2 b_2 c_2} * \cdots\cdots * f_{a_n b_n c_n}$$

마찬가지로 $b_1c_1a_2$와 $c_1a_2b_2$로 reading frame이 시작할 비율을 각각 계산한다.

$$r_2 = f_{b_1 c_1 a_2} * f_{b_2 c_2 a_3} * \cdots\cdots * f_{b_n c_n a_{n+1}}$$

$$r_3 = f_{c_1 a_2 b_2} * f_{c_2 a_3 b_2} * \cdots\cdots * f_{c_n a_{n+1} b_{n+1}}$$

i번째 reading frame이 코딩영역일 확률을 다음과 같이 계산한다.

$$P_i = \frac{r_i}{r_1 + r_2 + r_3}$$

4. 유전자 구조의 확률적 모델

유전자 분석을 위한 초기 연구는 프로모터 영역[11,12], splice site[13], 3장에서 살펴본 코딩영역을 알아내는 연구들과 같이 기능적인 영역들을 개별적으로 예측하는데 초점을 맞췄다. 최근에는 이러한 다양한 기능적 영역의 시그널을 종합하려는 연구가 활발히 진행되고있고 중요한 진전을 보이고 있다. 신경망[14], decision tree, dynamic programming, Markov 모델과 같은 매우 다양한 계산학적 방법들이 사용되고 있다. 본 장에서는 지면의 제약 때문에 Markov 모델을 이용한 유전자의 확률적 모델만을 살펴보겠다.

4.1 Hidden Markov Model

우리가 모르는 어떤 숨어있는 확률과정이 프로모터, 5′ UTR, exon, intron, 3′ UTR 등과 같은 기능적 시그널의 상태를 생성하고, 이 상태들은 다시 각각의 확률모델에 따라 DNA 서열을 생성한다고 생각할 수 있다. 주어진 DNA 서열을 분석하기 위해서는, 그 서열을 가장 초래했을 법한 일련의 기능적 시그널 상태들을 알아내어 유전자를 찾아내야 한다. DNA 서열은 직접 관찰할 수 있지만, 상태는 직접 관찰할 수가 없다.

Hidden Markov Model(HMM)[15]은 상태를 직접 관찰할 수 없는 Markov 체인이다. 대신 상태의 출력은 관찰할 수 있다. 각 상태에서의 출력은 상태별로 다른 확률분포에 따라 한 개의 기호만이 출력된다. DNA의 기능적 시그널들은 한 개의 뉴클레오티드로만 구성되는 것이 아니기 때문에 단순한 HMM은 DNA 분석에 적합하지 않다.

Generalized HMM(GHMM)[16]은 HMM을 일반화한 것으로, GHMM에서는 HMM과 달리 상태에서의 출력이 기호 한 개로 제한되지 않고 임의 길이의 string일 수 있다. 각 상태별로 출력

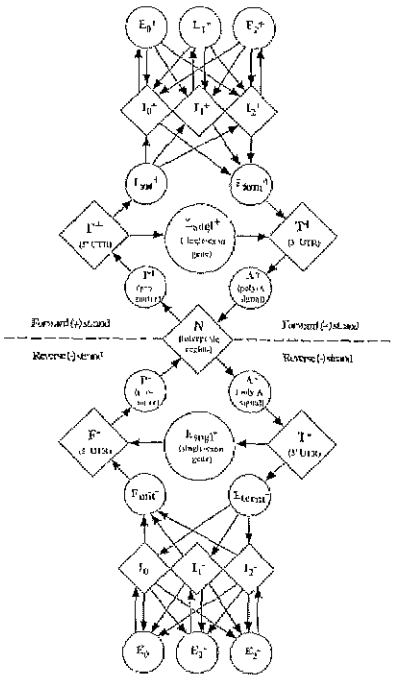

그림 4 GENSCAN의 유전자 모델

string 자체뿐만 아니라 string의 길이까지도 확률분포에 따라 생성된다. GHMM은 다음 4개의 매개 변수에 의해서 표현된다.

- 초기 상태 확률 분포 π
- 상태 전이 확률 매트릭스 T
- 상태에서의 string 길이 분포 f
- 상태에서 string을 생성하는 확률 모델 P

4.2 GENSCAN

GENSCAN[4]은 GHMM를 기본으로 한 유전자 구조에 관한 확률적 모델이다(그림 4 참조). 이 절에서는 GENSCAN을 중심으로 GHMM을 이용한 유전자 구조 분석 방법을 살펴보겠다. GENSCAN 모델에서(숨겨진) 상태는 프로모터, exon, intron 등과 같은 유전자의 기본적인 기능적 단위이다. Intron과 내부 exon은 reading frame과 관련이 있는 세 단계로 나누어져있다. 모델에서의 상태 전이는 생물학적으로 모순이 없는 순서이다.

또한, 모델은 위와 아래 두 부분으로 나누어져

있다. 위쪽 부분은 5′에서 3′으로의 forward strand의 유전자를 아래쪽 부분은 3′에서 5′으로의 backward strand의 유전자를 모델링하고 있다.

이 모델에 의해서 상태의 순서와 그에 따른 길이 L의 DNA 서열이 발생되는 과정은 다음과 같다.

(1) 상태의 초기 분포 π 에 따라 초기 상태 q_1이 선택된다.

(2) 상태 q_1의 string 길이 분포 f 에 따라 길이 d_1이 정해진다.

(3) 상태 q_1의 서열 생성 모델 P 에 따라 길이 d_1의 서열 마디가 생성된다.

(4) 상태 전이 매트릭스 T 에 따라 후속 상태 q_2가 선정된다.

위와 같은 과정을 생성된 서열 마디의 합이 L 이상 될 때까지 반복한다.

반대로 위의 모델은 길이 L의 DNA 서열 S가 주어졌을 때, 매개 변수들이 적절히 결정되었다면, 가장 사실일 것 같은 상태 순서 ϕ_i를 결정하여 유전자의 구조를 분석하는데 사용될 수 있다. Bayes 법칙을 이용하여

$$P(\phi_i, S) = \frac{P(\phi_i, S)}{P(S)} = \frac{P(\phi_i, S)}{\sum_{\phi_i \in \Phi} P(\phi_i, S)}$$

으로 계산할 수 있고, $P(\phi_i, S)$는

$$P(\phi_i, S) = \pi_{q_1} f_{q_1}(d_1) P_{q_1}(S_1|q_1, d_1) * \prod_{k=2}^{n} T_{q_{k-1}, q_k} f_{q_k}(d_k) P_{q_k}(S_k)$$

로 계산할 수 있다. ϕ_i는 상태 순서 $q_1, q_2, \ldots\ldots, q_n$이고, 서열 마디 $s_1, s_2, \ldots\ldots, s_n$은 각각 $d_1, d_2, \ldots\ldots, d_n$의 길이를 갖는다. Viterbi 알고리즘[17,15]을 이용하여 최적의 상태 순서를 결정할 수 있다.

GHMM을 바탕으로 한 GENSCAN의 성능은 상당히 고무적이다. GENSCAN의 정확도는 GRAIL, GeneID+, GeneParser3 등과 같은 다른 프로그램보다 훨씬 좋다.

5. 결 론

염기서열 중 유전자를 알아내는 일은 그것의 기능을 밝히기 위한 극히 중요한 첫 단계작업으로 계산학적인 많은 연구가 이루어지고 있다. 이 문제는 유핵생물의 경우 복잡한 intron-exon 구조 때문에 상당히 어려운 문제이다. 본 논문에서는 GHMM을 바탕으로 한 확률적인 모델을 이용하여 유전자의 구조를 분석하는 방법을 살펴보았다. GENSCAN은 75%에서 80% 정도의 exon을 정확하게 예측하는데, 정확성을 더 높이기 위한 연구가 필요하다.

그런 연구 중 뉴클레오티드나 아미노산 서열 데이터베이스와의 상동 관계를 이용한 최근의 연구[18,19] 결과는 상당히 고무적이다. 이와 같이 상동 관계를 이용하는 것과 5′이나 3′ 경계에서 분석의 정확성을 높이는 것이 앞으로 남은 중요한 연구과제가 될 것이다.

참고문헌

[1] M. Gelfand(1995), "Prediction of Function in DNA Sequence Analysis," Journal of Computational Biology, 1:87-115.

[2] J. Fickett(1996), "Finding Genes by Computer: the State of the Art," Trends Genet.,12(8):316-320.

[3] J. Claverie(1997), "Computational Methods for the Identification of Genes in Vertebrate Genomic Sequences," Human Molecular Gentics, 6(10):1735-1744.

[4] C. Burge & Karlin, S.(1997), "Prediction of Complete Gene Structures in Human Genomic DNA," Journal of Molecular Biology, 268:78-94.

[5] S. L. Salzberg, Pihaela M., Delcher A., Gardner M., & Tettelin H.(1999), "Interpolated Markov Models for Eukaryotic Gene Finding," Genomics, 59:24-31.

[6] B. Lewin(1997), "Genes VI," Oxford University Press.

[7] L. Stryer(1995), "Biochemistry," Freeman.

[8] M. Waterman(1995), "Introduction to Computational Biology," Chapman & Hall.

[9] J. Setubal & Meidanis J.(1997), "Introduction to Computational Molecular

Biology," PWS Publishing Company.
[10] F. Richards(1991), "The Protein Folding Problem," Scientific American, 264(1): 54-63.
[11] J. Fickett & Hatzigeorgiou A.(1997), "Eukaryotic Promoter Recognition," Genome Research(7):861-878.
[12] M. Tompa(1999), "An Exact Method for Finding Short Motifs in Sequences, with Application to the Ribosome Binding Site Problem," ISMB'99:262-271.
[13] M. Reese, Eeckman F., Kulp D., & Haussler D.(1997), "Improved Splice Site Detection in Genie," Journal of Computational Biology(4):311-323.
[14] P. Baldi & Brunak S.(1998), "Bioinformatics: The Machine Learning Approach," MIT Press.
[15] L. Rabiner(1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, 77(2):257-286.
[16] D. Kulp, Haussler D., Reese M., & Eeckman F.(1996), "A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA," ISMB '96:134-142.
[17] A. Viterbi(1967), "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm," IEEE Trans. Informat. Theory, IT-13, 260-269.
[18] O. Gotoh(1999), "Homology-based Gene Structure Prediction: Simplified Matching Algorithm using a Translated codon (tron) and Improved accuracy by Allowing Long Gaps," Bioinformatics 15:190-202.
[19] L. Milanesi, D'Angelo D., Rogozin I. (1999), "GeneBuilder: Interactive in Silico Prediction of Gene Structure," Bioinformatics 15:612-621.

공 은 배

1988~현재 충남대학교 컴퓨터공학과 부교수
1995 Oregon State Univ 전산학 박사
1981 서울대학교 계산통계학과 석사
1978 서울대학교 계산통계학과 졸업
관심분야:기계학습, 암호학, Bioinformatics
E-mail keb@cc.cnu.ac.kr