IT 기반 단백질 서열 분석

글 _ 황 미 녕 · 생명정보시스템개발실 · mnhwang@kisti.re.kr

1. 도입

게놈(genome)이란 유전자를 의미하는 'gene'이란 단어와 염색체라는 뜻의 'chromosome'이란 단어를 합성한 신생어로 생명체의 모든 DNA를 일컫는 말이다. DNA란 유전정보를 담고 있는 물질로 모든 생명 현상을 조절하는 중심추 역할을 수행한다. DNA는 아데닌(Adenine), 티민(Thymine), 시토신(Cytosine), 구아닌(Guanine) 등 네 종류의 염기들로 구성되어 있고 이들의 배열 순서에 따라 인종, 체형, 성격, 특정 질병에 걸릴 요인 등 모든 유전형질이 결정된다. 인간의 DNA는 35억여 개의 염기로 구성되어 있고 이를 신문으로 인쇄할 경우 15만여 페이지에 달하는 방대한 양이 된다.

인간게놈 프로젝트는 세포 속에 존재하는 35억 개의 DNA 염기들이 어떤 순서로 배열되어 있는 지를 모두 밝혀내고자 하는 국제 프로젝트로 10년여 간에 걸쳐 약 30억 달러가 투입되어 6개국 1000여 명의 연구진들이 참여해 인간의 DNA 염기서열을 밝혀냈다. 게놈 프로젝트의 의미는 쉽게 비유하면 생명현상의 비밀을 닮고 있는 책을 해독하기 위한 첫 단계로 한글의 자모의 배열을 알아냈다고 생각하면 된다.

생명체의 내부에서는 단순히 염기 서열만으로는 어떠한 작용도 발생하지 않는다. 기능을 할 수 있는 유전자들이 모여서 단백질을 형성해야만 가능하다. 이런 단백질을 만들 수 있는 인간의 기능 유전자는 3만여 개 정도로 추정된다. 이는 전체 30억여 개의 염기 중 약 3% 정도에 불과하며 나머지 97%가 왜 존재하는지는 아직 밝혀지지 않았다. 추정되는 3만여 개의 기능 유전자 중 그 기능이 밝혀진 유전자는 9000개에 불과하다. 또한 DNA 염기 구성은 개인별, 인종별로 특징적인 차이를 보이는데 이를 단기 염기 다형성(SNP: Single Nucleotide Polymorphism) 이라고 부른다. 완성된 유전자 지도를 토대로 3만여 개로 추정되는 기능 유전자가 어떠한 단백질을 만들어 내는지를 연구하여 이를 바탕으로 난치병과 불치병 그리고 장수의 비밀을 밝혀내게 될 것이다. 또한 개인, 인종, 생물 간의 게놈 정보를 비교하여 염기 구조의 차이점을 밝히는 비교 유전자의 연구를 통하여 어떠한 염기 구조의 차이가 개인 간, 인종 간 형질의 차이를 야기하는지를 알아내어 특정 염기 구조의 차이가 질병 발생의 요인이 되는지, 유전자의 특정 부위 돌연변이가 어떠한 기전으로 유전 질환의 발생을 야기하는지를 연구할 수 있을 것이다.

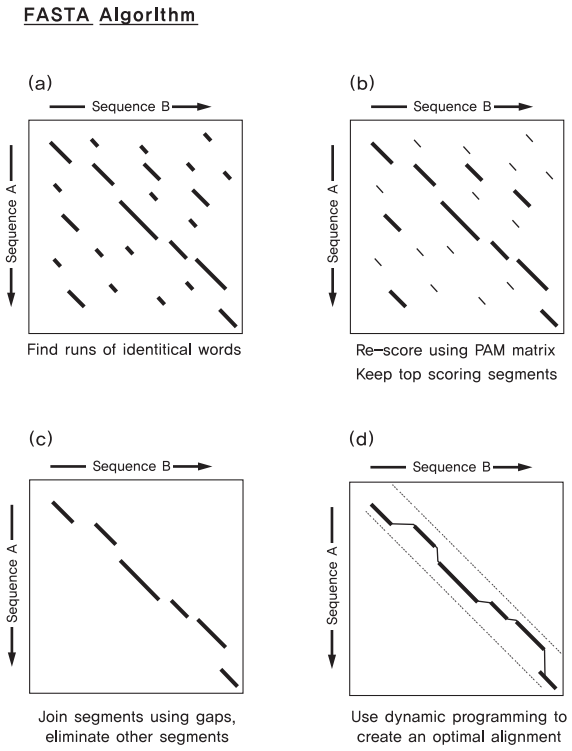
단백질과 DNA의 서열 해독이 늘어남에 따라, 새로이 밝혀진 서열과 유사하거나 상동성이 높은 서열을 기존의 데이터베이스에서 검색해 내는 일의 중요성이 부각되고 있다. 서열 사이의 상동성을 찾아내는 일은 단백질이나 DNA의 기능을 밝히는 중요한 단서가 되기도 한다. (여기서 서열의 상동성이 매우 높다면, 그 기능도 유사할 것이라는 가정이 내포되어 있다.) 현재 많이 사용되는 유전자 및 단백질 서열에 대한 검색시스템은 질의 서열과 데이터베이스에 있는 각각의 서열 간의 부분 정렬 점수(local alignment score)를 계산하여 이 점수를 기준으로 결과를 보여준다. 이런 소모적인 검색 방법의 시간 복잡도는 질의어와 검색 대상이 되는 데이터베이스 서열의 총 개수 뿐만아니라 각각 서열의 길이에도 큰 영향을 받기 때문에 계산량이 매우 커지게 된다. 따라서 사용자가 인내할 수 있는 시간 내에 검색 결과를 보여줄 수 있는 새로운 방법이 필요하다. 우리는 단백질 서열 분석에 있어서 생물학적 접근 방법이 아닌 IT 기반의 접근 방법으로 분석하고자 한다.

2. 단백질 분석 사이트

단백질 서열의 분석은 검색하고자 하는 질의 서열과 데이터베이스에 있는 각각의 서열 간의 유사도 측정에서부터 시작된다. 가장 기본적인 방법은 두 서열의 1:1 정렬(pairwise alignment)로 상동성을 적용하는 범위에 따라 부분 정렬(local alignment)과 전체 정렬(global alignment)을 사용하게 된다. 상동성은 생물학적으로 진화적인 연관성이 있음을 나타내는 것인데, 서열 정렬에서 구할 수 있는 상동성은 정렬 계산의 결과인 상동성 값으로 표현된다. 각 알고리즘은 스코링 매트릭스(scoring matrix)와 갭비용(gap cost)의 파라메터를 사용하는데, 스코링 매트릭스는 모든 아미노산 쌍과 염기쌍에 대해서 생물학적인 모델과 통계적인 방법에 근거하여 연관성 정도를 수치로 표현한 것이고, 갭은 진화상의 삭제나 삽입에 해당하는데, 정렬에서 한 서열에서 생략되어 있거나 다른 서열에서 추가되어 있는 부분을 나타내는 것이다. 정렬 알고리즘에서는 아미노산 쌍이나 염기쌍에 대해 일정한 값을 주듯이, 이들 갭이 발생하는 경우에도 일정한 값을 주어 적절한 형태의 정렬이 구해지도록 한다. 일반적으로 사용되는 스코링 매트릭스는 PAM 시리즈나 BLOSUM 시리즈가 있다. PAM은 1969년 M. Dayhoff 가 고안한 것으로 생물의 계통 관계 분석을 통하여 발표된 것이고, BLOSUM은 1992년 S. Henikoff가 고안한 것으로 BLAST 등에서 기본값으로 채용되어 널리 사용되고 있다.

초창기에 고안된 단백질 서열의 분석 및 검색 방법으로는 FASTA 알고리즘이 유명하다. <그림 1>에서처럼 FASTA는 데이터베이스에 있는 모든 서열과 점 행렬(점 행렬: 두 서열을 가로와 세로로 놓고 일치하는 곳에 점을 찍으면 동일한 서열의 경우 대각선으로 선이 그려지게 된다)을 그리지 않고 '단어'를 기반으로 하여 점 행렬 기법을 사용한다. 입력 서열에서 한 개 혹은 두 개의 단백질 서열로 이루어진 '단어'들의 조합을 만든 후 데이터베이스의 임의의 한 서열에서 각 단어들과 일치하는 단어를 찾아내어 각각의 단어들을 연결하는 대각선을 만들어 각 부분들의 값을 스코링 매트릭스를 이용하여 재계산한다. 가장 큰 값을 가진 부분을 '단위'로 하여 이 부분을 확장하면서 가장 큰 값을 가지는 대각선을 그린다. 각 대각선을 연결하는 것으로 유사성이 높은 서열을 찾아내게 된다. 현재 FASTA는 European Bioinformaitcs Institute에서 http://www.ebi.ac.uk/fasta33으로 웹 서비스되고 있다.

'단어' 단위 검색 방법의 보다 발전된 형태가 1990년에 Karlin과 Altschul가 발표한 BLAST(Basic Local Alignment Search Tool)로 현재 단백질과 아미노산 서열 검색에 가장 널리 사용되고 있다. BLAST는 부분 정렬(local alignment)에 중점을 두기 때문에 속도도 FASTA에 비해서 빠른 편이다. BLAST는 현재 NCBI(National Center for Biotechnology Information)에서 http://www.ncbi.nlm.nih.gov/BLAST/ 웹 서비스되고 있다.
이 외에도 기존에 밝혀진 단백질의 패밀리, 도메인, 기능적인 위치 정보들을 가진 데이터베이스들을 대상으로 서열 검색을 지원해주는 InterPro(http://www.ebi.ac.uk/interpro/) 등이 있다.



<그림 1> FASTA 알고리즘

3. IT 기반의 단백질 서열 분석

전산학 분야에서, "정보 검색" 시스템은 대용량의 문서 데이터베이스를 대상으로 하는 효율적이고 적합한 검색 방법론으로 인정받아 왔으며 원문(full-text) 색인 시스템과 인터넷 정보 검색 분야에 성공적으로 적용되었다. 반면 생물학적 데이터베이스 검색 도구들은 생물 서열 데이터베이스에서 질의 서열과 가장 유사한 서열들을 찾아내는 것이 목표이다. 서열 데이터베이스에서 유사 서열을 찾는 것과 문서 데이터베이스에서 유사한 문서를 찾아내는 것은 그 개념과 방법론에서 보면 매우 유사한 작업이 될 수 있고, 실제로 CAFE 시스템의 예에서 그 가능성을 엿볼 수 있다. 생물학적 서열을 DNA 염기나 아미노산 부호로 쓰여진 문서로 간주하고, 알맞은 색인 방법을 사용하면, 정보검색 방법론을 생물학적 데이터베이스의 검색 방법에 적용할 수 있다. KISTI 바이오인포매틱스센터에서는 N-gram 방법론을 기반으로 하는 단백질 서열의 색인 기법을 제시하고, 이 방법론을 이용하여 단백질 서열 데이터베이스의 검색을 수행하는 시스템을 구현하였다.

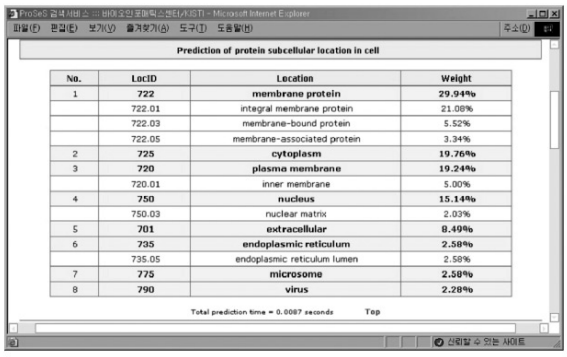
서양어와 같은 문서는 단어나 어구로 구성되는데, 일반적으로 공백(띄어쓰기)에 의해 구분이 가능하다. 이런 특성을 이용하여, 정보 검색 시스템에서는 일반적으로 문서에서 각 단어 또는 어구를 추출하여 역파일(inverted file)에 저장한다. 그러나 DNA나 단백질 서열과 같은 생물학적인 서열은 공백이 없는 문자열이며, 이를 서양어의 단어나 어구에 해당하는 의미 있는 단위로 구분하는 것은 거의 불가능한 일이다. 이런 상황에 맞는 경험적인 색인 방법이 몇 가지 있는데, 이 중 하나가 n-gram 토큰 방법론이다. N-gram을 이용한 색인 방법론은 단어의 경계가 없거나 모호한 중국어 문서에 잘 적용된다. 즉, 생물학적인 서열은 어구의 경계가 없는 문서로 볼 수 있고, 이런 문서의 색인에 잘 적용되는 n-gram 색인 기법을 여기에 적용하는 것이 가능하다. N-gram은 k-tuple이나 w-mer 등의 다른 용어로 표현되기도 하는데, 이것은 서열 안의 일정한 간격을 의미하며, 이 간격은 길이가 n으로 고정된 부분 문자열로 겹치게 된다. 예를 들어 ACEPITCH의 단백질 서열에서 n이 4라면, 최종적으로 나오는 n-gram은 ACEP, CEPI, EPIT, PITC, ITCH가 된다. 단백질 서열 데이터베이스 내의 모든 서열을 색인화하여 n-gram들은 각각 DB 저장 시스템에 검색키와 저장 위치 리스트의 정보를 가지고 저장되게 된다.

서열 간 유사성의 측정은 기존 시스템에서 사용하는 부분 정렬 점수를 사용하지 않고, 정보 검색 분야에서 질의어와 대상 문서의 유사성을 측정하기 위한 모델들 중 벡터 공간 모델을 사용한다.

단백질 서열 분석 시스템을 만들기에 앞서서 단백질 서열과 관련 정보를 담고 있는 사용이 빈번한 데이터베이스로부터 통합 데이터를 구축하였다. 조지타운 대학의 PIR-NREF 단백질 서열 데이터를 기본으로 하여 각 단백질의 패밀리 정보는 iProClass, INTERPRO, BLOCKS, PRINTS, PFAM, METAFAM, COG 등의 단백질 관련 데이터베이스에서 추출하고 세포내 위치 정보는 SwissProt에서 추출하여 통합 데이터를 만들었다. PIR-NREF에서 제공하는 단백질 서열 데이터는 2004년 8월 30일자로 178만 여건 중 단백질의 세포 내의 위치 정보를 제공해주는 SwissProt의 경우는 15만 여건 정도이다.

이렇게 만들어진 단백질 통합 데이터로 KISTI에서 개발된 KRISTAL-2002 정보검색시스템을 이용하여 검색 시스템을 개발하였다. 기본 검색 화면은 〈그림 2〉에서 보듯이 단백질 서열의 검색과 서열 외의 정보를 검색할 수 있도록 구성되어 있다.

단백질 서열 분석 사이트인 ProSeS(Protein Sequence Search)의 서비스 내용은 다음과 같다.

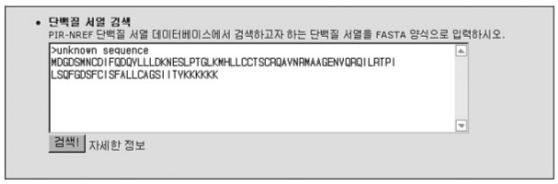


〈그림 2〉 단백질 서열 분석(ProSeS) 사이트

1. 단백질 유사 서열 검색 질의 서열에 대해서 데이터베이스에서 가장 상동성이 높은(가장 유사한) 서열을 찾아 사용자에게 제공한다.
2. 텍스트 검색단백질의 서열 정보 외에 이름과 종에 관련된 정보를 통합 검색하는 것이 가능하다.
3. 세포 내 단백질 위치 예측검색 결과의 단백질 서열

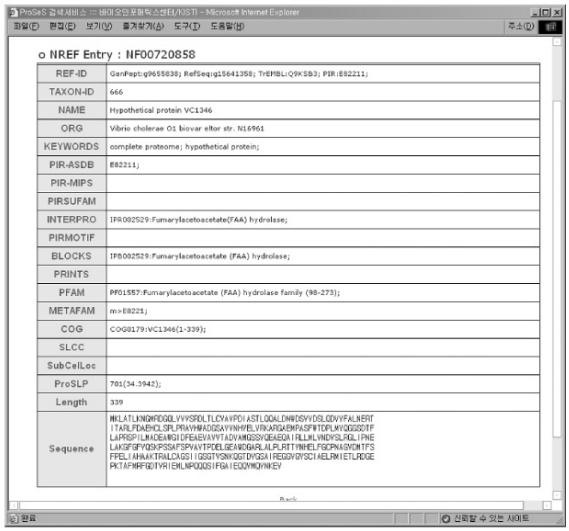
들이 가지는 정보로부터 질의 단백질 서열의 세포내 위치를 예측하여 그 결과를 보여준다.

4. 단백질 패밀리 분류 서비스 검색 결과의 단백질 서열들의 패밀리 분류 정보로부터 질의 서열의 패밀리 분류를 예측한다.
5. 유사 서열 검색을 위해서 질의어를 〈그림 3〉과 같이 FASTA 형식으로 입력한다.



〈그림 3〉 단백질 서열 분석(ProSeS) 사이트

〈그림 4-a〉 유사 사열 검색 결과

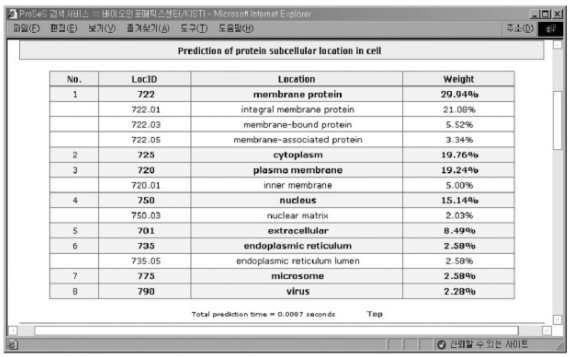


〈그림 4-b〉 NF00720858의 상세 정보 보기

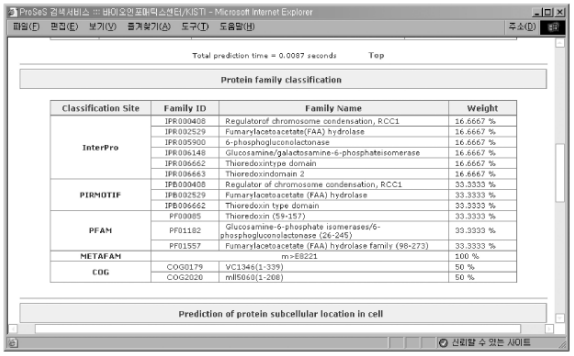
유사 서열 검색을 하게 되면 결과는 〈그림 4〉와 같이 상동성이 높은 서열의 내림차순으로 보여준다. 각각 서열의 ID를 선택하게 되면 상세 정보를 볼 수 있다. 여기에서, 유사 서열 검색 결과는 BLAST에서 동일하게 검색을 수행했을 때의 결과와 같은 결과를 보여 주지는 않는다. 서열 간의 유사성 측정 방법이 틀리기 때문에 동일하게 결과로 나온 서열의 경우에도 순위가 틀리기도 한다. 생물학자들이 가장 신뢰하는 BLAST와 비교하여 상이한 결과의 유효성 평가와 이를 통한 최적의 유사성 측정 방법을 찾아내는 것이 향후 과제이다.

유사한 서열 검색 외에 추가로 〈그림 5〉에서와 같이 세포 내의 위치 예측 결과와 패밀리 분류 결과도 볼 수 있다. 단백질 통합 데이터에는 각각 서열 데이터마다 본 실에서 개발한 세포 내의 위치 예측 시스템인 ProSLP(Protein Subcelluarl Localization Prediction)의 예측 정보를 저장해 놓아 유사 서열 검색 결과 중 상동성이 높은 상위 데이터들로부터 위치 예측 결과를 분석하여 보여 준다. 또한 각각 패밀리 분류 서비스의 분류 정보도 포함하고 있기 때문에, 검색 결과 내의 패밀리 분류 정보를 분석하여 보여 준다.

검색 속도는 질의 서열의 길이가 50 아미노산에서 500 아미노산의 범위에서 비교 검색 실험을 한 결과 ProSeS가 BLAST에 비해 평균적으로 약 6.5배 빠르게 나타났다.



〈그림 5-a〉 세포내의 위치 예측 결과



〈그림 5-b〉 패밀리 분류 서비스

4. 결론

본 생명정보시스템개발실에서는 단백질 서열의 분석에 있어서 생물학적 접근 방법이 아닌 IT 기반의 정보검색 측면에서 접근하여 분석 시스템을 개발하였다. 단백질 서열 분석 서비스인 ProSeS는 개발의 마무리 공정 상태에 있으며 개발이 마무리 되는 대로 CCBB 홈페이지에서 서비스될 예정이다. 현재 CCBB에서 서비스 중인 ProSeS의 서비스는 이전 버전으로 단백질 통합 데이터가 아닌 PIR-NREF 단백질 서열 데이터를 기본 데이터로 사용하고 있다.

현재, 시간 소모적인 비교 방법을 사용하는 BLAST 등은 검색 시간 측면에서 한계에 직면하고 있다. 생물학적 데이터베이스가 날이 갈수록 기하급수적으로 커지는 미래에는 이런 시스템을 보기 힘들지도 모른다. 본 실에서 개발하는 IT 기반의 서열 분석 시스템인 ProSeS가 기존의 생물학적 데이터베이스 검색 도구를 대체 또는 보완하는 실질적인 하나의 대안이 될 수 있을 것이다. KISTI