유전체학과 생물정보 사업

스몰소프트 박기정

1. 서 론

유전정보분석을 위해 컴퓨터와 소프트웨어를 도구로 사용하는 학문 분야를 생물정보학(bioinformatics)이라고 한다. 세계적인 관심 속에서 2001년 인간유전체 프로젝트에 대한 세부 결과가, Celera와 NHGRI를 포함한 미국과 유럽의 두 그룹에 의해 발표되면서, 현대과학계의 혁명적인 사건의 하나로 기록되었는데, 이러한 엄청난 생물학 연구결과의 생산을 위한 방대한 정보 분석의 기반에 바로 생물정보학이 있었다. 실제로 국내에는 인간유전체 연구결과의 발표와 동시에 생물정보학이라는 분야가 일반에 알려지기 시작했지만, 유전정보 분석을 위해 컴퓨터를 활용하고자 하는 시도는 선진국에서 이미 1960년대 말부터 시작되었고, 1980년대 말부터 일반적으로 이용되다가 유전체 프로젝트가 시작된 1990년대 초부터 본격화되었다.

거대 연구과제로서 1990년대 초 미국에서 시작된 유전체 프로젝트(genome project)는 생물의 유전정보를 총체적으로 밝히고자 하는 시도로서, **기존의 유전자 단위의 연구방법이 아닌 한 생명체의 총체적인 정보인 유전체를 단위를 한 실험 구상과 정보처리를 수행하는 유전체학(genomics)이라는 학문을 탄생시켰다.** 생명체의 유전자를 담고 있는 DNA는 염기라고 하는 생물학분자 단위로 구성되는데, 이 염기를 하나의 문자로 볼 때, 유전체 프로젝트의 일차 결과는 이 염기라는 문자로 구성되는 매우 긴 암호 문장으로 표시된다. 이 암호 문장은 대장균의 경우에는 길이가 약 4백 5십만이고, 사람의 경우에는 약 30억에 이른다. 식물의 유전체는 사람보다 훨씬 긴 DNA를 가지고 있는 경우가 많다. 어떤 생물에 대한 유전체의 이러한 염기구성을 밝히는 것이 유전체 프로젝트의 일차적인 목적이다. 수년전에는 세균이나 바이러스의 유전체 프로젝트의 결과가 대단한 연구결과였지만, 불과 몇 년 동안 수백 개의 유전체 프로젝트가 종료되거나 진행되고 있다. 국내에서도 2000년에 최초로 Zymomonas mobilis라는 세균의 유전체 프로젝트 결과 발표 이후, 매년 한 해에만 국내 연구진에 의해 수 개의 유전체 프로젝트 결과가 발표되고 있고, 유전체 프로젝트를 수행할 수 있는 연구팀이나 벤처회사가 십여 개 이상 등장했다.

인간유전체에 대한 유전자지도가 구체화되면서 실제로 인간의 질병과 노화 등에 대한 연구를 위한 획기적인 자료가 갖추어 졌고, 미생물에 대한 유전자 지도들이 완성되어 가면서 이들 미생물을 기반으로 하는 신물질연구나 신약연구가 가속화되고 있다. 생명현상의 규명에서 볼 때, DNA 염기구성과 유전자지도를 기반으로 해서, 각 유전자가 생명 현상을 위해 각각 생산하는 생체물질과 이들의 상호 반응 및 조절 기작에 대한 근본 문제를 풀기까지에는 매우 긴 과정이 남아있지만, 어쨌든 이러한 유전체 프로젝트는 학문적이나 산업적인 응용을 위해 획기적인 결과를 내고 있는 것이다. 국내에서도 한 해 동안만 수 개 이상의 유전체 프로젝트가 종결되고, 매년 월등히 향상된 연구결과가 나오고 있다.

이러한 **유전체 프로젝트와 이를 기반으로 하는 생물학 연구와 산업의 대부분 과정에서 수반되는 방대한 유전정보분석을 위해 생물정보학이 그 역할을 하고 있다.** 어떤 면에서는 생물학과 관련된 산업의 주요 경쟁력의 핵심요소가 생물정보학이 되고 있는 것이다. 한 미생물의 유전체 프로젝트를 위한 생물학 실험결과는 잘 갖추어진 국내 실험실에서 수 주 내에 생산될 수 있지만, 연구결과로 발표하기까지의 시간은 수개월 이상이 필요한데, 이는 유전정보 분석을 위한 처리과정에 소요되는 시간이다. **생물정보학은 유전자 데이터와 정보를 대상으로 하여, 유전자의 기능을 유추하거나 유전자간의 관계에 대한 정보를 추출하는 등, 생명체의 암호를 판독하는 과정에 관련된 전반적 분석 방법을 연구하고 이 방법을 표현하는 프로그램을 개발하는 것이다.** 한편, 이러한 과정에 따르는 데이터와 기존의 엄청난 생물학 데이터를 관리하기 위한 다양한 데이터베이스 처리도 담당한다. 유전체학의 전반에서 소프트웨어를 필요로 하고 있는 상황이다.

2. 유전체 프로젝트와 생물정보학

유전체 정보 분석은, 데이터베이스 관련분야와 분석 도구(analysis tools; analysis programs) 개발 관련 등으로 나눌 수 있다. 하나의 염기를 컴퓨터의 저장 단위인 한 바이트(byte)에 저장하면, 사람의 DNA 서열 만을 위해 약 3 GB 정도가 필요한데, 실제 연구자료 중 이 서열이 차지하는 부분은 일부이므로 이의 수십 배에 달하는 자료가 되며, 이를 컴퓨터에서 관리하고 분석하기 위해서는 다시 이의 몇 배에 해당하는 크기의 자료처리가 된다. 따라서 유전체 자료를 관리하기 위해서는 데이터베이스 시스템의 개발과 관리가 기본적이다. 분석 도구에 대한 연구는 크게 2가지 방향으로 이루어지고 있다. 첫 번째는, 궁극적으로 생명현상을 일으키는 생체 고분자의 구조에 대한 연구로, 물리, 화학적인 모델에 의해 계산해서 예측된 생체 고분자의 구조와 분자간의 상호작용을 그래픽으로 표현하고 이 기법에 의해 신약후보 물질을 설계한다.

두 번째는, 유전자 서열을 대상으로 하여, 유전자의 기능을 유추하거나 서열간의 관계에 대한 정보를 추출하는 등, 서열로 표현되어 있는 생명체의 암호를 판독하는 과정에 관련된 전반적 분석 방법을 연구하고 이 방법을 표현하는 프로그램을 개발하는 것이다. 이 분야 이론에서는 생물학 서열의 특성을 반영하여 생물학 서열의 성질과 관계성을 수학적으로 정의하여 주어진 정의에 의한 최적의 해답을 구하는 방법을 다루게 된다.

그 외에 유전체 정보 분석에서 다루는 분야들은 다양하다. 실험을 통해 얻은 DNA의 서열에서 단백질을 발현하는 유전자가 있는가를 예측하는 것(gene prediction이라고 한다), 미지의 DNA나 단백질의 서열을 기존의 기능이 알려진 DNA나 단백질 서열과 비교하여(homology search라고 한다) 그 미지 서열의 기능을 밝히거나 유추하는 것 등과 같은 방법들은 현재 많이 활용되고 있는 방법이며, 복잡하게 상호기작관계가 밝혀져 있는 유전자들과의 정보(metabolic pathway databases)와 비교하여 미지 서열의 기능을 밝히는 것, 한 생물의 유전체를 다른 생물체의 유전체와 총체적으로 비교하는 것(comparative genomics) 등과 같이 유전체 프로젝트의 종료와 함께 새롭게 개발되는 방법들도 많이 등장하고 있다. 최근 DNA 칩을 활용한 대용량의 유전자분석에서도 생물정보학적인 방법이 데이터해석을 위해 이용되고 있다.

최근에 개발된 여러 이론들이 생물정보 분석을 위해서 시험되고 일부는 활용되고 있다. 유전체연구는 많은 과제가 종료되고 있고, 동시에 다수의 생물체에 대한 유전체 프로젝트가 시작되고 있다. 많은 유전체 분석 전문기업이 미생물의 유전체에 대해서는 단 하루 만에 서열결정 실험을 완료할 수 있는(10년 전에는 수년의 노력이 필요한 과제였다) 장치를 갖추고 있고, 이런 서열 데이터로부터 유용한 정보를 얻기 위해, 수십에서 수백 명의 생물정보 분석 인력이 많은 데이터처리와 이에 필요한 프로그램들을 개발하고 있다. 국내의 유전체연구도 2000년을 시작으로 본격화되었으며, 미생물의 유전체 서열결정을 수주 이내에 완료할 수 있는 시설을 갖추고 있는 벤처기업도 이미 등장하였다. 한편, 이들 과제의 산물인 유전체정보로부터 유용한 학문적 결과와 산업적 결과를 얻기 위해 유전체후속(post-genome) 프로젝트들이 최근에 진행되고 있는데, 방대한 유전체 자료로부터 유전자들의 구체적인 기능을 밝히기 위한 연구의 주된 방법론으로서 생물정보학이 각광을 받고 있다. 미래의 생물학자는 실험보다는 소프트웨어를 활용한 정보 분석에 (극단적으로 In Silico Biology라고 하는) 더 많은 시간을 투자해야 할 것으로 전망하는 견해도 많다.

생물정보학의 분야가 다양하기 때문에 인력도 다양하게 필요하기는 하지만, 절대적으로 인력이 부족한 상황이며, 특히 고급 연구개발 인력은 현재의 수요로 볼 때 부족이 심각한 상황이다. 생물정보학은 하나의 학문이 아닌 생물학과 전산학의 두 학문의 혼합학문으로 발전되어 왔다. 생물정보학을 전공하기 위해서는 기본적으로, 분자생물학에 대한 전문지식과 프로그래머로서의 전산학지식을 갖추어야 한다. 따라서, 전산학에서 고급프로그래밍을 위해 필요한 이론과 실습을 익혀야 하고, 생물학적인 기반지식을 폭넓게 갖추어야 한다. 이를 기반으로 하여 생물정보학 고유의 이론연구와 개발경험을 갖추어야 한다.

3. 유전체 정보 분석시스템

유전체 프로젝트를 수행하는데 있어 일반적으로 요구되는 유전정보분석시스템의 구조는 그림 1과 같다. 유전정보분석시스템은 크게 유전체 서열결정시스템, 유전체 기능분석시스템 그리고 이와 상호연관이 있는 통합 데이터베이스 시스템 등 크게 세 가지로 구성 된다.

유전체 서열결정시스템의 경우, 대용량 sequencing 기계로부터 나오는 염기서열에 대한 정보를 볼 수 있는 base calling, 반복되는 염기서열과 vector 염기서열의 오염을 걸러주는 repeat mask와 vector screening, sequencing 결과로 생기는 contig들의 물리적 위치를 결정해주는 contig assembly와 이것의 결과를 시각적으로 보여주어 연구자가 수정 또는 편집이 가능한 환경

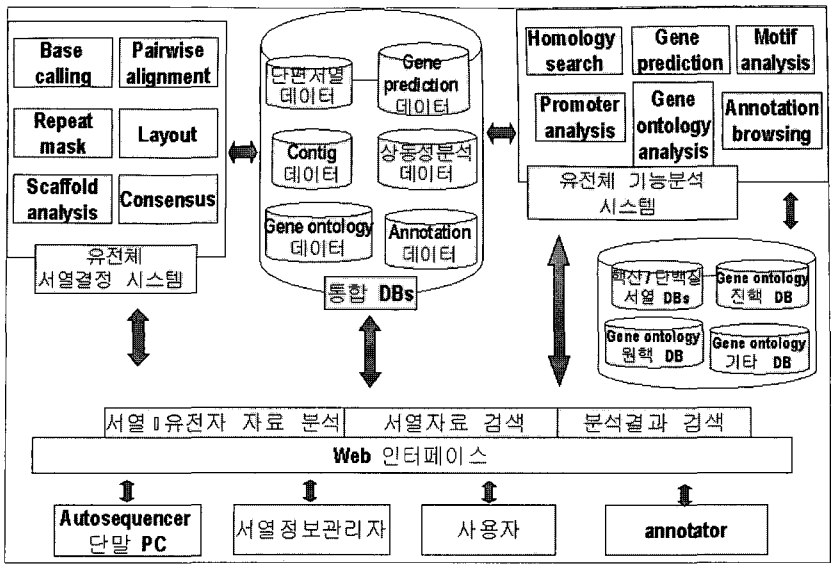


그림 1 유전체 프로젝트 수행을 위한 유전정보분석시스템 구조

을 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 다수의 세부 모듈로 구성되어 있다. 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 계산 시간이 가장 많이 소요되는 step 이다. 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 평가 받고 있다. 이러한 기능 부위는 생명체의 수억 년 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 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 예상 된다.

Annotation browser는 진행 중이거나 완료된 genome project의 annotation의 결과에 대하여 조사하고자 하는 유전자의 유전체내의 물리적인 위치 등을 포함한 해당 유전자가 가지고 있는 모든 유전정보를 그래픽화면으로 가시화 시켜주는 작업을 수행 한다. 수작업으로 annotation이 필요한 유전자의 경우 이것의 정보수정 및 보완작업도 가능케 하는 기능을 가진다(그림 2).

통합데이터베이스 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.

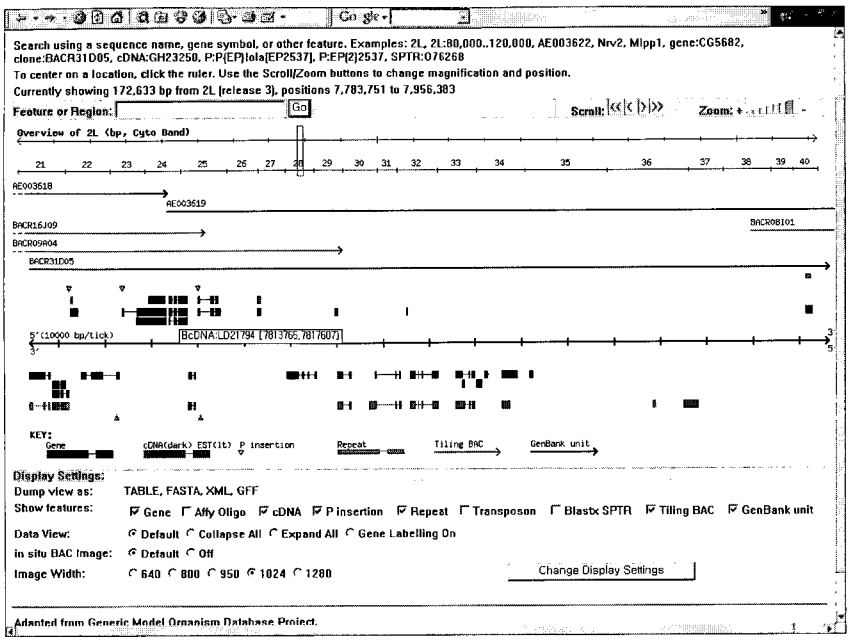


그림 2 초파리(Drosophila melanogaster) 2번 염색체의 annotation browser

3.1 프로모터 예측

원핵생물의 프로모터 예측방법은 이미 알려진 consensus sequence(pattern)나 주어진 염기서열 set로부터 counting 접근방법으로 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 pattern방법의 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 matrix방법 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 matrix방법처럼 프로모터 지역과 비 프로모터(non-promoter)지역에 대한 분별력이 떨어져 다수의 false positive를 생산한다는 문제점이 있다. 1992년 Kanehisa가 neural network기법을 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.

진핵생물의 프로모터 예측은 프로모터 영역과 비 프로모터 영역의 유전자서열의 성분(composition) 차이에 기초하여 조절부위를 찾는 search-by-content algorithm과 앞서 언급한 TATA box 나 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.

Search-by-content와 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 hexamer빈도 등을 이용한 TSSG 가 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 3에 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.

세계적으로 가장 뛰어난 진핵생물의 프로모터 예측 프로그램들의 sensitivity도 현재에는 60%에도 미치지 못하는 저조함을 나타내고 있다.

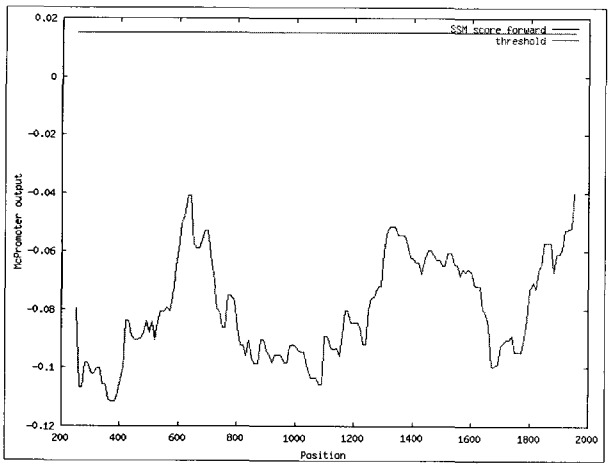


그림 3 인간 12번 염색체 일부지역에 대한 McPromoter System을 이용한 프로모터 검색결과

3.2 유전자 예측

유전체 내의 정확한 유전자 위치를 알아내기 위해 많은 gene prediction 모델들이 개발되어왔다. 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 초기연구는 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 prediction 프로그램들이 등장하기 시작했다. 그 중에서 Fickett의 모델에 근거한 TestCode와, neural network 접근방식으로 여러 가지 구성에 대한 통계적 수치를 적용해 염기서열 단편을 coding region과 non-coding region으로 구분한 GRAIL이 가장 널리 사용되었다.

이전까지 만들어진 gene prediction 모델들은 DNA의 한쪽 strand만을 분석하도록 만들어졌지만, GeneMark가 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 위치 특이적인 non-homogeneous 5th-markov chain을 DNA 양쪽 가닥에 모두 구성하고, 각 markov chain의 상대적인 score에 따라 coding region을 찾아낸다.

GeneMark이후 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. Glimmer를 개선한 Glimmer2에서는 interpolated Markov model 대신 interpolated context model을 사용하는데 markov chain에서는 바로 이전의 염기서열 구성에 따라 확률 값을 계산하는데 비해 이 model에서는 염기서열단편의 상호 의존성을 측정하고 그 의존성에 따라 확률 값을 계산한다.

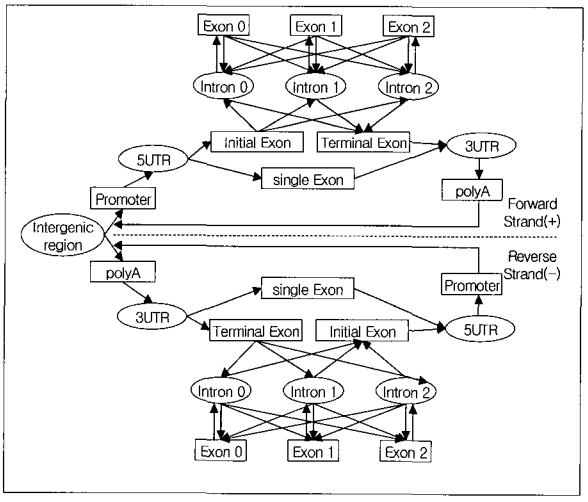


그림 4 Duration HMM을 이용한 gene prediction model

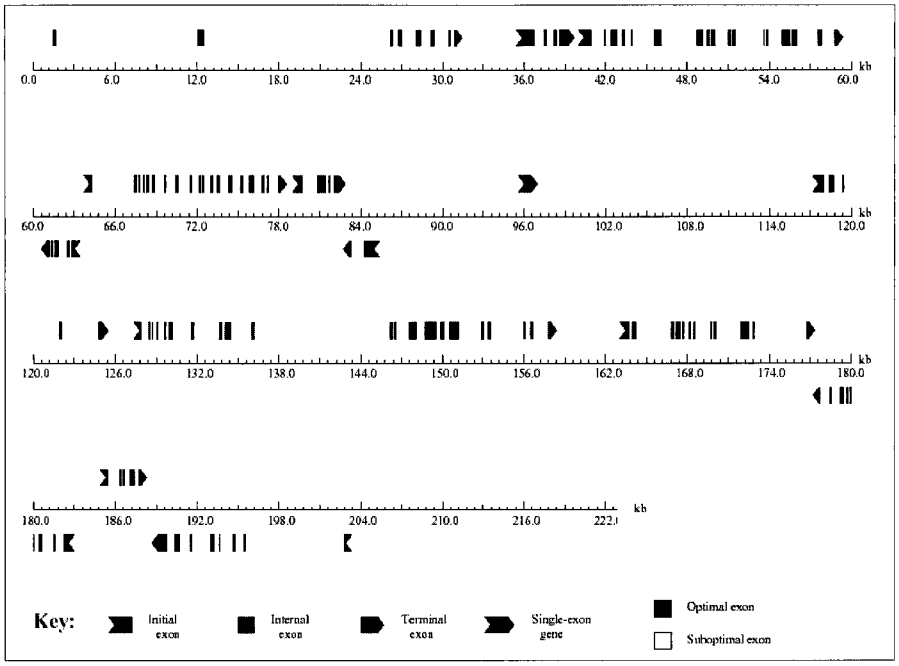


그림 5 GenScan을 이용한 인간 12번 염색체 일부분의(GenBank U47924, 222kb) 염기서열 유전자 예측 결과

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. elegans와 포유동물의 gene prediction을 연구한 Gelfand에 의해서 시작되었다. 이 두 프로그램은 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 모델을 사용한 GeneID, neural network과 dynamic programming을 혼합한 GeneParser, linguistic method를 사용한 GenLang, discriminant analysis를 사용한 FGENEH, decision tree를 사용한 morgan, generalized HMM을 사용한 Genie, 그리고 duration HMM을(그림 4) 사용한 GenScan이 있다.

3.3 통합 유전체 분석시스템

통합 유전체 분석 시스템은 통합된 환경에서 다수의 모듈을 실행하고, 결과를 통합 데이터베이스를 관리하도록 구성하고 있다(그림 6, 그림 7). 사용자 인터페이스도 web 기반으로 관리하고 있는 경우와 Windows 프로그램으로 개발되는 경우의 두 가지가 모두 존재한다. 상용의 프로그램은 대단히 고가이거나, 현재는 각 연구그룹의 시스템 상황과 맞지 않는 경우가 대부분이나, 분석 환경이 표준화되면, 상용시스템으로서의 큰 가치를 가지게 될 것이다.

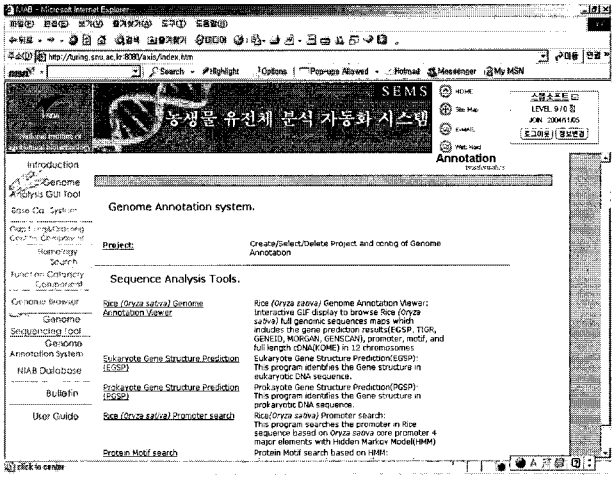


그림 6 Web 기반 유전체 분석 통합시스템의 예

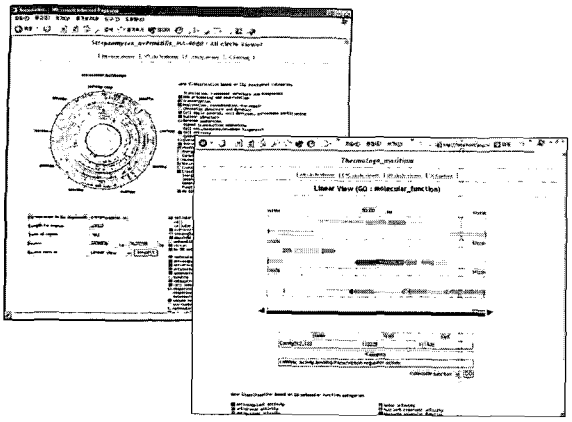


그림 7 Web 기반 유전체 가시화 시스템

4. 유전체 연구 및 관련 소프트웨어 개발의 전망과 비전

영화로도 유명해 진 '쥬라기 공원'의 원작인 마이클 크라이턴의 소설에는, 소설 저작 당시의 첨단 연구였던 몇 가지 생물정보학 기법을 기술하고 있다. 일종의 '공룡 유전체 프로젝트'로서 공룡 유전체의 긴 DNA 염기구성을 결정하기 위해, 작은 DNA 조각의 염기구성을 결정한 후, '그림 맞추기 퍼즐'처럼 각 조각을 맞추고 연결하여 긴 DNA의 염기구성을 완성한다거나(contig 형성), 공룡 DNA의 손상된 부분의 유전정보를 복구하기 위해, 이미 알려져 있는 다른 생물들의 DNA와 비교하고(다중 서열정렬) 유사한 생물의 DNA부분을 계산하여(계통분석), 손상된 DNA 부분을 가장 합리적으로 예측한다. 몇 년 전에 개봉된 '가타카'라는 영화에서는 개인에 대한 유전정보를 이용한 새로운 유전자차별이 이루어지고 있는 미래사회가 그려지고 있는데, 손가락 끝에서 소량을 채혈하여 순식간에 신원을 밝힌다거나, 태어난 아기의 유전정보를 기존 데이터와 비교하여 질병별 발병 가능성을 진단하거나 하는 장면 등, 개인에 대한 유전정보를 실시간으로 분석하는 장면들이 많이 등장하고 있다. 사실 현재의 유전체 프로젝트나 생물학 연구의 발전 속도를 보면, 이들 영화에서 나오는 상황을 가능하게 하는 생물정보학적인 시스템의 구축은 그리 먼 이야기가 아니다. 유전정보에 기반한 바이오 사회는 노화억제나 질병정복이라는 긍정적인 면과 유전자차별과 개인 유전정보 남용 등의 부정적인 면을 모두 가지고 있지만, 한편에서는 피할 수 없는 미래이기도 하다.

5. 유전체 분석소프트웨어와 사업화

유전정보분석을 위해 다루는 분야는 매우 다양하며, 이들 분야의 산물인 소프트웨어 또한 다양하다. 대부분의 생물학연구자나 이러한 소프트웨어와 밀접한 관계가 있으며, 가까운 미래에 일반이 이러한 소프트웨어를 활용해야 한다고 한다면, 그 시장 규모는 엄청나다고 할 수 있다. 유전정보분석을 다루는 국내 생물정보학 벤처기업은 모두 최근 5년 내에 설립된 기업들로, 유전체 사업과 병행하거나 유전체 사업을 수행하는 바이오 벤처기업과 공조하고 있으며, 그 수는 10개 이내이다. 한편에서는 바이오 벤처기업과 유사한 성격과 인맥을 가지면서 그 운명을 같이 한다고 할 수 있지만, 한편에서는 정보기술 벤처기업과 유사한 성격을 가지고 있다. 흔히 BIT(BT와 IT의 혼합)라고 불리는 최근의 유행의 첨단에 서 있는 이들 기업은, 존재나 존재가치에 대한 인식조차 불확실하던 수 년 전으로부터 주목받기 시작하는 현재까지 급격한 변화를 겪고 있다. 한 치 앞을 보기 힘든 요즘의 기업환경에서도 아마도 수 년 간은 많은 변화를 겪게 될 것으로 보인다.

중장기적으로 생물정보학 벤처가 해결해야 할 문제는 결코 가볍지 않다. **생물정보학 벤처의 특성상 기본적으로 R&D 성격을 가져야만 한다는 것과, R&D 투자와 그로 인한 회 수 경험이 적은 국내 투자환경에서 R&D 기업에 대한 투자가 인색할 수밖에 없다는, 두 현실이 양립해 가야한다는 것이, 생물정보학 기업으로서는 매우 부담스러운 기업환경**이라고 할 수 있다. 시장 가치 인식에 대한 차이 역시 만만치 않다. 관련 기업 입장에서는 생물정보학 관련 시장이 분명해 보이는 상황에서도, 현재의 수익성과 시장형성을 기준으로 볼 때 일반적으로 불투명하게 보일 때, 객관적인 자료로서 이를 설득하기는 매우 어려운 일이다. 이것은 모험을 담보로 한 기회를 가지고 있는 벤처기업으로서 짊어지고 가야할 몫으로 보인다. **기업이 생산해야할 제품개발을 위해 당면한 문제는 연구개발 인력의 절대부족이다.** 생물정보학의 분야가 다양하기 때문에 인력도 다양하게 필요하기는 하지만, 절대적으로 인력이 부족한 상황이며, 특히 고급 연구개발 인력은 현재의 수요로 볼 때 부족이 심각한 상황이다. **생물정보학은 하나의 학문이 아닌 생물학과 전산학의 두 학문의 혼합학문으로 발전되어 왔으므로, 이 분야의 연구개발을 수행하기 위해서는, 기본적으로 분자생물학에 대한 전문지식과 프로그래머로서의 전산학 지식을 갖추어야 한다.** 고급프로그래밍을 위해 필요한 이론 및 실습과 생물학적인 지식을 기반으로 하여 생물정보학 고유의 이론연구와 개발경험을 갖추는 것이 요구되는 것이다.

다수의 바이오벤처 기업들이 어려운 여건에서도 기술개발을 계속해 오고 있다. 생물정보학 벤처기업들이 운영을 위한 단기적인 수익성 확보와 함께 포기해서는 안

되는 부분은 기술개발과 독자기술에 의한 제품생산이다. 선진국에서는 이미 당연시 되고 있는 유전정보분석 소프트웨어 시장의 도래에 대한 인식을 가지고, **생물정보학 벤처기업들은 기술과 제품개발을 해나가야 할 것이며, 정책면에서도 이를 위한 기술개발 인력 양성에 대한 교육 투자나 기업정책 수립 등이 이루어져야 할 것이다.**

참고문헌

[1] Vanet A. and Marsan M., "Promoter sequences and algorithmical methods for identifying them," *Res. Microbiol.*, Vol 150, pp.779-199, 1999.

[2] Fickett J.W. and Hatzigeorgiou A.G., 'Eukaryotic promoter recognition', *Genome Research* Vol. 7, pp.861-878, 1997.

[3] 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.*, Vol. 297, pp.599-606, 2000.

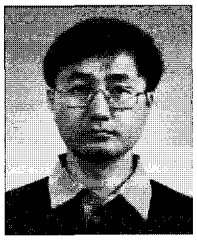
[4] Uberbacher, E. and Mural, J., "Locating protein coding regions in human DNA sequences by a multiple sensor-neural network approach," *Proc. Natl. Acad. Sci.* Vol. 88, pp.11261-11265, 1991.

[5] Borodovsky M. and McIninch, J., "GENMARK: parallel gene recognition for both DNA strands," *Computer & Chemistry*, Vol. 17, No. 2, pp.123-134, 1993.

[6] Delcher A. L., Harmon D., Kasif S., White, O. and Salzberg S.L. "Improved microbial gene identification with GLIMMER," *Nucl. Acids Res.* Vol. 27, No. 23, pp.4636-4641, 1999.

[7] Burge C. and Karlin S., "Prediction of Complete Gene Structures in Human Genomic DNA," *J. Mol. Biol.* Vol. 268, pp.78-94, 1997.

박 기 정



1986. 2 서울대학교(학사)
1989. 2 한국과학기술원(석사)
2002. 2 한국과학기술원(박사)
1989. 2~1998. 6 한국생명공학연구원 (선임연구원)
2000. 3~현재 (주)스몰소프트 대표이사
관심분야 : bioinformatics, computational biology, genomics, algorithm, machine learning
E-mail : kjpark@smallsoft.co.kr