전용 서열블라스트 시스템 설계 및 구현

최영상, 최한석
목포대학교 멀티미디어공학전공
e-mail:choi7040@mokpo.ac.kr

Design and Implementation of Personalized Sequence BLAST System

Young-Sang Choi, Han-Suk Choi
Dept of Multimedia Engineering, Mokpo National University

요 약

본 논문에서는 지놈 연구에서 반드시 필요한 유전체 서열정보의 상동성 검색을 위해 해외 데이터베이스에만 의존하고 있는 국내 지놈 연구자들에게 더욱 빠르고 쉽게 접근할 수 있는 서열블라스트를 만들 수 있는 방법을 제공함으로써 국내 지놈연구에 도움이되고자 특정 종류의 유전자정보 만을 모아놓은 지놈 서열블라스트 시스템을 설계하고 구현하는 방법을 제시하였다. 이 시스템을 활용하면 많은 생물 지놈 연구자들의 연구시간을 획기적으로 단축할 수 있을 것이다.

1. 서론

현재 바이오테크놀로지와 관련된 모든 산업의 급속한 발전에 따라 전세계의 많은 생물학자들이 실험을 통해 많은 생물학 데이터를 양산해 놓고 있다. 이러한 데이터를 잘 활용한다면 실험을 훨씬 쉽게 할 수 있을 뿐만 아니라 새로운 정보를 얻을 수 있게 된다. 그러나 이러한 데이터는 점점 대용량이 되어가고 이에 따라 필요한 정보를 얻기 위해서 많은 노력이 필요한 실정이다. 이 대량화된 정보를 처리하기 위해서는 컴퓨터의 이용이 반드시 필요하게 되었다. 이런 이유로 컴퓨터의 이용을 통해 생물정보를 처리하는 생물정보학(bioinformatics) 연구가 활발히 진행되고 있다. 그러나 정보를 생산하는 생물학자들은 이를 효과적으로 활용할 수 있는 방법을 알지 못하여 많은 시간과 노력을 허비하고 있는 실정이다. 이에 본 논문에서는 생물학자의 컴퓨터의 활용에서 가장 많이 사용되고 있는 방법인 BLAST (Basic Local Alignment Search Tool)를 효과적으로 사용할 수 있도록 방법을 제시하고자 한다.

생물학자들은 실험을 통해 산출된 잘 모르고 있는 핵산(DNA 또는 RNA) 서열이나 단백질 서열이 주어지면, 먼저 분석하고자 하는 서열과 관계되는 서열들을 서열 데이터베이스(sequence database)로부터 찾아낸다. 그 다음에 데이터베이스에서 찾아진 서열들로부터 분석하고자 하는 서열의 성질들을 유추해낸다. 이러한 서열의 성질은 많은 정보를 담고 있게 되는데 이러한 정보를 anotation 정보라고 한다. 서열데이터베이스 검색에서 가장 대중적으로 사용되고 있는 방법이 BLAST이다. BLAST를 사용하는 방법에는 두 가지가 있다. 첫 번째는 미국 NCBI 사이트(http://www.ncbi.nlm.nih.gov/blast/)에서 웹을 통해서 BLAST 검색을 하는 것이다. 하지만 분석하고자 하는 서열들의 수가 아주 많을 때는 검색 속도도 늦어지고 또한, 식물서열과 동물서열을 비교하는 오류데이터가 검색된다. 두 번째 방법은 전용서버를 구축해 놓고 NCIB에서 필요한 데이터만 다운로드하여 로컬서버에 저장을 해 놓고 서열을 비교하는 방법이다. 전용 서버를 구축하여 데이터를 저장해놓고 실험데이터와 비교한다면 빠른 시간내에 검색과 연구를 수행 할 수 있을 것이다. 또한 이 annotation 정보를 저장해 놓고 특정서열과 연결해 놓으면 향후 다른 연구자의 유사한 서열정보시 관련 데이터를 더 효과적으로 제공 할 수 있게 된다.

이에 본 논문에서는 특정 유전체의 일반 정보, 기능성 정보등을 포함하여 관련 연구 분야의 발전에 기여할 수 있는 전용서열 블라스트 시스템을 설계하고 프로토타입을 구현하였다.

본 논문의 2절에서는 관련연구동향을 살펴보고 3절에서는 전용서열 블라스트시스템의 프로토타입을 설계 및 구현하고 실험하였다. 4절은 결론 및 향후 연구방향에 대하여 설명한다.

2. 관련 연구

2.1 연구동향

국내 생물학관련 연구자들은 실험결과에 나온 DNA sequencing, Protein sequencing등 각종 실험 결과 비교할 때나, 새로운 기능과 구조등 연구할 때 주로 미국의 NCBI, 유럽의 EBI, ExPASy, 일본의 DDBJ등에서 서비스하고 있는 데이타베이스 및 상동성 검색, Multialignment용 소프트웨어 즉, FASTA(1,2), BLAST(3,4)를 이용하고

있다. 그러나 국내에서는 GenBank(5,6), EMBL 그리고 DDBJ(7)와 같은 database를 운영하지 못하고 분석 관련 S/W개발에 관한 연구도 미진한 실정이다.

미국의 NCBI와 같은 국제적인 기관에서 GenBank를 구축, 관리하고 있음에도 불구하고 유럽, 일본에서 자체 DNA Center를 운영하는 것은 그 만큼 효율성이나 중요성을 인식하고 있기 때문이며, 실제로 막대한 연구비 투자를 하고 있다. 그러나 국내에는 생물학관련 데이터베이스를 운영하고 있는 대표적인 기관이 없고 이에 대한 인식도 아직은 부족한 실정이다. 또 분석 S/W를 비롯한 각종 응용 프로그램들이 데이터베이스를 근간으로 개발되고 있으므로 염기서열 및 단백질 서열 정보들을 서로 공유하면서 연구에 기초 자료가 되는 부분들을 이용하고자 하여도 해외 정보망을 사용하여야 한다. 국내에서 데이터베이스를 운영하지 않으면 분석 소프트웨어 개발 뿐만 아니라 체계적인 생물학 정보 개발에도 후진성을 면하기 어렵고 정보를 기반으로 하는 모든 분야 연구의 발전은 국외 정보에 의존할 수밖에 없는 실정이다.

2.2 PIPELINE

본 시스템에 사용되는 EST데이터들은 원시 EST데이터들을 각 생물에서 얻는다. 이러한 EST데이터들은 분석이 되어 잇지 않아서 본 시스템에 사용이 불가능하다. 이러한 EST데이터를 분석하기 위해서는 PIPELINE이라는 절차를 걸쳐서 사용가능한 데이터가 된다. PIPELINE의 절차를 설명하자면 생물의 조직에서 최초의 ACGT의 형태로 추출하는 작업을 Base Calling이라 한다. Base Calling 작업을 마쳐야지만 비로써 생물정보학적인 데이터가 된 것이다. 이후에는 Vector Trimming을 거치게 된다. 이 과정에서 100bp이하인 데이터들은 삭제가 된다. Low quality region을 제거, 중복성을 줄이기 위한 클러스터링 등이 이 과정에 포함이 된다. 이러한 과정을 통하여 얻어진 EST데이터들 즉 서열들은 클러스터링과 어셈블리 과정을 거치게 되는데 이 과정을 마친 후 생성되는 컨티그(Contig)/싱글레톤(Singleton)서열정보, 어셈블리정보, 클러스터링 정보가 생성되게 되며 이러한 정보들을 데이터베이스에 저장하게 된다.

클러스터링/어셈블리 과정을 수행한 후 생성된 컨티그/싱글레톤 서열 정보를 통하여 알려진 서열 정보를 대상으로 하는 BLAST를 이용하여 알려진 서열에 주석을 다는 Annotation 과정을 거쳐서 본 시스템에 사용되는 원시 데이터가 생성이 된다.

3. 전용 서열 블라스트 시스템

3.1 시스템 설계

전용 서열 블라스트 시스템은 유전체정보, 서열정보, 유용물질 정보로 구성되었고 각 부분의 연관성을 찾아 한 화면에서 유전체 연구에 필요한 모든 자료들을 동시에 볼 수 있어 연구자들이 최소한의 시간을 투자하여 필요한 자료를 검색할 수 있도록 설계하였다. 유전체의 일반정보를 입력하게 되어 그 생물의 기본적인 특성을 알 수 있다. 그 활용범위까지 자세하게 입력을 한다. 이 부분은 실제 DNA와는 상관이 없지만 Sequencing 된 EST 데이터들이 어떤 생물인지 알아야 하기 때문에 입력이 필요하다.

PIPELINE에서 분석된 EST 데이터들은 본 시스템의 데이터베이스에 저장이 된다. 이때 데이터의 양이 엄청 많아서 LOAD DATA INFILE이란 명령어를 사용해서 MySQL에 저장한다. cDNA Library 정보를 브라우징 할 수 있게 하며, NCBI의 BLAST를 로컬서버에 맞게 수정을 하여 서열을 비교하고자 하는 연구자들이 사용할 수 있도록 한다. 유전체 일반정보에서 클릭하여 cDNA Library 정보를 볼 수 있도록 하며, Annotation 정보 페이지에서는 전용 BLAST로 페이지로 이동해서 BLAST 검색을 하거나, EXPASY(http://www.expasy.ch)DB와 검색비교를 할 수 있으며, UniProt (http://www.ebi.uniprot.org)DB와도 비교 검색 할 수 있도록 설계하였다.

3.2 실험환경

3.2.1. 실험환경

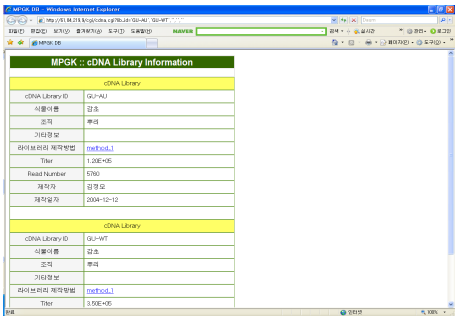
본 시스템은 linux 운영체제를 기반으로 mysql DB를 사용하였으며 PERL, PHP, JAVA script,등을 사용하여 구현하였다. 개발환경은 Linux 시스템이었지만 필수적인 소프트웨어가 지원하는 운영체제라면 어느 시스템에서도 시스템 구축이 가능하도록 하였다. 본 시스템에서 사용되는 모든 데이터에 대한 데이터 객체를 구현하고 실제 응용프로그램에서 각 객체를 생성하여 데이터의 이용을 원할하게 하였다.

3.2.2. 실험내용

본 논문에서 구현된 시스템은 감초를 대상으로 실험을 하였다. 시스템에서 분석된 내용을 웹에서 확인할 수 있다. 감초의 유전체 정보를 볼 수 있다. 그림1은 감초의 일반정보이다.

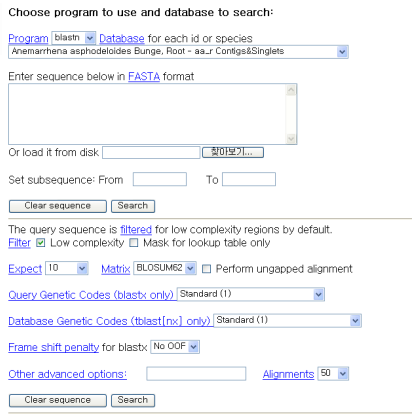
(그림 1) 유전체 일반정보

유전체의 자세한 정보 및 NCBI의Taxomony ID를 링크하여 실시간 검색을 할 수 있게 구성하였으며 anotation 정보를 데이터베이스화하여 효율적으로 검색할 수 있도록 구성하였다. 그림 2는 cDNA 라이브러리이며 그림1에서 클릭하여 정보를 볼 수 있다.

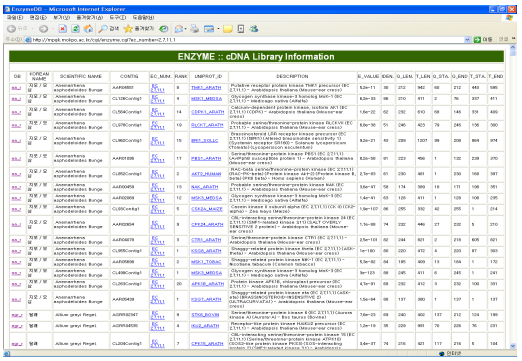


(그림 2) cDNA 라이브러리

그림 3은 전용 서열 불라스트 시스템의 웹 인터페이스 모습이다.



(그림 3) 전용 서열블라스트



(그림 4) Anotation 정보

NCBI에서 제공하는 Web 블라스트를 이용하여 구성된 BLAST는 연구자의 서열정보를 입력하여 시스템의 특정서열정보와 비교할 수 있게 제공되며 외국의 유명 통합정보 시스템과 연결되도록 구성하여 연구자들이 자신의 연구결과물을 국내뿐만 외국에서 제공하는 정보와 비교하여 연구하기 쉽도록 구성되었다. 그림 4는 Annotation 정보이다.

3.3 평가

EST데이터의 활용 방안이 많이 제시되고 있지만 그에 맞는 EST 데이터의 분석 및 서비스 프로그램이 부족한 실정이다. 그래서 특화된 새로운 시스템 개발이 절실히 요구되었다. 본 시스템에서 감초를 대상으로 실험을 해 보았다. 그 결과 본 시스템은 EST 데이터의 분석 및 웹인터이스에서 효율적임이 나타났다.

4. 결론 및 향후 연구방향

전용블라스트 시스템을 구축하게 되면 외국에만 의존하던 유전체 정보를 국내에서도 검색하고 활용할 수 있어 향후 포스트게놈시대에 첨병이 될 수 있으리라 본다. 또한, 연구자들은 빠른 정보를 얻을 수 있어 연구자들의 연구시간을 단축할 수 있게 하였고 유전체 경쟁시대를 대비하여 자체 유전체 정보를 확보하여 막대한 국가적 재산이 외부로 빠져나가는 것을 방지하는 효과가 있다.

향후 연구방향은 연구자에게 연구자가 필요한 다양한 정보를 한 눈에 찾을 수 있게 하여 연구자의 연구시간을 단축하는데 집중하며 각각 구성된 시스템을 연결하여 새로운 정보를 재생산해 낼 수 있는 방법을 연구해야 할 것이다.

참고문헌

[1] Shivashankar H. Nagaraj, Robin B.Gasser and Shoba Ranganathan "A hitchhiker's guide to expressed sequence tag (EST) analysisHill

[2] Hao Xu "EST Pipeline System: Detailed and Automated EST Data Process-ing and Mining"

[3] Apu~~a C. M. "ESTWeb: bioinformatics services for ESTsequencing projects"

[4] Patricia Ayoubi. "PipeOnline 2.0: automated EST processing and functional data sorting"

[5] Viktor Stolc. "A Gene Expression Map for the Euchromatic Genome of Drosophila melanogaster"

[6] 제임스 티스달, 박현석 역, 필로 시작하는 바이오인포매틱스, 한빛미디어, 2002

[7]George Reese, 서환수 역, MySQL 시스템 관리와 프로그래밍, 한빛미디어, 2002.

[8]김태환, 바이오인포매틱스 ESTs 서열분석, 생능출판사, 2003.