유전체 분석 초기 단계에서 유전자 리스트 작성을 위한 방법론

A Gene-list Identification Methology on the Initial Stage of Genome Project

오정수, 안명상, 조완섭, 권해룡*, 김영창*
충북대학교 경영정보학과
충북대학교 미생물학과*

Oh Jeong-Su, Ahn Myoung-Sang, Cho Wan-Sup,
Kwon Hae-Ryong*, Kim Young-Chang*
Dept. of Management Information System,
Dept. of Microbiology, Chungbuk National University*

요약

나날이 그중요성이 증대되고 있는 생물정보학(Bioinformatics) 분야에서 이미 오래전에 유전자를 예측하고 분석할 수 있는 여러 가지 기법들이나 소프트웨어도구 들이 개발되어 있는 상태이며 현재 많은 생물학자들은 이러한 분석 도구들을 이용해 연구를 수행하고 있다. 비록 기존의 실험적인 방법 없이 여러 생물정보학 분석 도구를 이용해 효율적으로 유전자를 규명하고 기능을 분석하는 것이 가능해졌지만 아직까지 유전체 사업(Genome project)은 많은 시간과 비용이 드는 까다로운 작업임에는 틀림없다. 본 논문은 기존의 유전자 DB와 분석 도구들을 이용하여 유전체 사업 초기 단계에서 구체적인 유전자의 목록을 작성 할 수 있는 방법을 제안한다. 유전체 사업의 초기단계에서는 구체적인 유전자 정보를 얻기가 어려운 상황이나 제안된 기법을 사용하면 구체적인 유전자 정보를 초기에 파악 할 수 있게 된다.

Abstract

To predict and analyze genes, many methods and tools are already developed in Bioinformatics field which is being more important in future. And many biologists have now performed the research with them. Although it is possible to identify gene and to analyze its function efficiently without experimental methods, it is still hard work. In this paper, we propose a method that make gene list on the initial stage of Genome project. It is difficult to obtain detailed gene list in the initial stage of Genome project. but proposed system provides gene information as much as possible even in the initial stage.

I. 서론

분자생물학에서 가장 기초적인 단계는 유전체 분석이라고 할 수 있다. 현재 유전체 분석을 하는데 있어 이전에 비해 여러 기술의 발전으로 인해 실험적으로 분석하던 것보다 생물학자들이 더 효율적으로 유전체 사업(Genome project)을 진행할 수 있게 되었다. 특히 정보기술의 발전은 분자생물학에 있어 획기적인 진전을 가져다 주었는데 유전체사업도 그에 따라 많은 진전을 이루게 되었다. 그러나 아직까지도 유전체사업은 여러 단계를 거쳐야 하는 까다로운 과정이다. 이러한 유전체 분석 과정의 현재 보편 적인 방법으로는 콘티그(contig)를 제작하여 이를 글리머(Glimmer)와 같은 유전자 예측 프로그램에 넣어서 ORF(open reding frame)를 찾고 찾은 ORF에 대해 기능 예측 프로그램을 통해 그 유전자의 기능을 예측하는 것이다. 그런데 이러한 방법은 2kb 이상의 어느 정도 긴 서열의 유전자를 입력하여야 유전자 예측이 가능하다. 따라서 상대적으로 짧은 단편(fragment)들을 결합편집(assembly) 과정을 거쳐 2kb 이상의 콘티그를 제작해야 하는데 이는 많은 시간과 비용이 드는 작업일 뿐만 아니라 유전체사업이 어느 정도 진행된 상태에서나 가능한 것이다. 왜냐하면 콘티그 제작을 위해서는 전체 유전체를 다 해독해야 하며 또한 전체 크기의 7~10배 정도를 해독해야 하기 때문이다. 비록 이런 방법은 비교적 정확하게 유전자를 예측해 주

지만 유전체사업의 초기에는 유전자를 예측하기가 어렵고 유전자에 대한 구체적인 정보를 알기가 어렵다.

따라서 본 논문은 이러한 점을 보완하여 단편서열을 가지고 유전체 사업 초기에도 유전자 분석할 수 있는 새로운 방법을 제안하고자 한다.

제안된 방법은 크게 3가지 목적을 갖고 있다.

첫째. 유전체 사업 초기단계에서 구체적인 유전자 목록을 작성할 수 있다.

둘째. Cog[1][2][3] 데이터베이스와 비교하여 유전자들의 오솔로그(Orthologous) 관계 파악 및 필수 유전자와 종 특이적 유전자를 구별 할 수 있다.

셋째. 산업적으로 유용한 종 특이적 유전자 목록을 통해 후속 연구를 앞당긴다.

본 논문에서는 이와 같은 목적에 부합되도록 관련 유전자 데이터베이스 및 프로그래밍 기법들을 활용한 시스템과 방법을 제안한다. 또한 전체적인 시스템의 구조를 설명하고 시스템 상에서 어떤 방법으로 유전자 목록을 작성 하는지 자세히 설명한다.

Ⅱ. 유전자 예측 시스템

여기서는 제안된 유전자 예측 시스템의 구조와 데이터베이스 및 예측 방법을 소개한다.

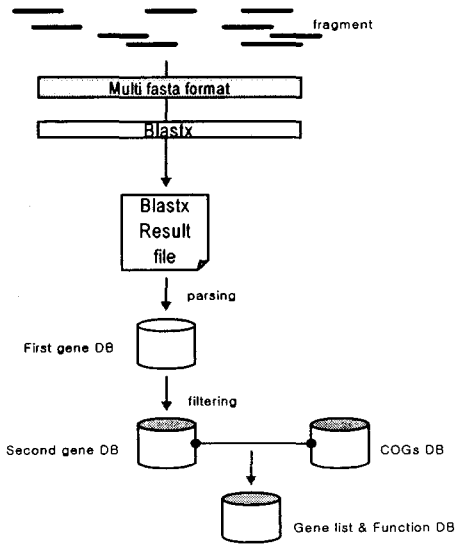
1. 시스템 구조

Gene list 작성을 위한 전체적인 시스템에 대해 알아보도록 하자. 그림 1.은 제안된 유전자 예측시스템의 구조를 보여주고 있다. Blastx라는 프로그램에 여러 개의 단편을 입력, 실행하여 나온 결과 값을 데이터베이스로 구축한다. 그리고 적절한 filtering 조건을 통해 2차 데이터베이스를 구축하고 COG 데이터베이스와 비교하여 유전자 목록을 작성한다.

다음에서 각 단계별로 상세히 설명한다.

1.1 유전자 동정 및 기능예측

먼저 어떤 유전체의 DNA 단편 서열을 만들고 이 단편 서열을 Blastx[4][5] 라는 프로그램을 사용해 전 세계에 있는 모든 protein sequence database 즉 nr (non-redundant) database와 상동성 검색(homology search)을 한다. Blast는 현재 상동성 검색의 가장 대표적으로 사용되고 있는 소프트웨어로 Blastx는 입력받은 DNA fragment sequence전체를 프레임에 따라 단백질 서열로 변환해주고 그것을 기존의 protein database와 비교하여 기능예측 과정인 주해(annotation) 과정을 해주는 프로그램이다. 이렇게 Blastx를 통해 나온 Result File을 파싱하여 데이터베이스에 1차적으로 저장한다. 이 데이터베이스는 Result File의 모든 내용을 추출하여 저장한다.



▶▶ 그림 1. 시스템 개요

1.2 Filtering을 통한 2차 DB구축

다음으로 해야 할 작업은 Blastx의 Result File을 저장한 1차 DB를 적절한 조건에 의해 Filtering하여 2차 DB를 구축하는 것이다.

첫 번째 Filtering 조건은 상동성 검색에 있어 가장 상동성이 높다고 판단된 것만을 가져오는 것이다. 왜냐하면 상동성이 가장 높다고 판단된 것은 그 만큼 유전자일 가능성이 크기 때문이다. 여기서 판단 기준은 expect 값을 기준으로 한다. Blastx에서 결과 값은 expect 값을 기준으로 하여 오름차순 정렬인데 이는 expect 값이 낮

을 수 록 그 만큼 결과에 대한 신뢰도가 높다는 뜻이기 때문이다. 따라서 expect 값을 기준으로 제일 낮은 값은 유전자일 가능성이 큰 것으로 볼 수 있다.

두 번째 Filtering 조건은 결과 값으로 나온 것 중에 Hypothetical~이거나 Unknown 이거나 Unnamed 인 것들은 제거 하는 것이다. 이것들은 아직 까지 밝혀지지 않은 기능들이거나 연구 중인 것들이기 때문이다.

1.3 COG 데이터베이스와 비교

COG(Clusters of Orthologous Group) 데이터베이스는 미국의 NCBI(National Center for Biotechnology Information) 에서 제공하는 것으로 서열이 완전히 밝혀진 Genome gene family 간의 비슷한 기능을 하는 유전자를 그룹핑 해놓은 것으로 현재 완전히 밝혀진 66개의 게놈의 단백질 서열들을 비교하여 각 생물 종에서 서로 유사한 기능을 하는 유전자들을 몇 가지 그룹으로 분류한 것이다. 그룹의 종류는 정보 축적 및 처리 기능 관련 유전자군 ([J], [k], [L]), 세포 생리 기능 관련 유전자군 ([D], [O], [M], [N], [P], [T]),대사 기능 관련 유전자군 ([G], [C], [E], [F], [H], [I]),기능 미확인 ([R], [S]) 군 등으로 크게 구분된다. 따라서 COG 데이터베이스와 비교하게 되면 그 유전자의 오솔로그(orthologous) 관계 파악 및 필수 유전자와 종 특이적 유전자를 구별 할 수 있다. 왜냐 하면 만약 COG 데이터베이스와 비교 했을 때 COG에 포함 되는 것이면 그것은 필수유전자로 볼 수 있는 것이고 포함되지 않는 것이면 이종에만 나타나는 종 특이적 유전자로 볼 수 있기 때문이다.

COG 데이터베이스와의 비교를 위해서는 COG 데이터베이스를 local로 구축하는 것이 필요하다. 비교하려 하는 유전자 들을 COG 웹사이트에서 COGnitor 등을 통해 비교하는 것은 시간이 많이 걸리고 수작업으로 입력하여 Search를 해야 하는 번거로운 작업이기 때문이다.

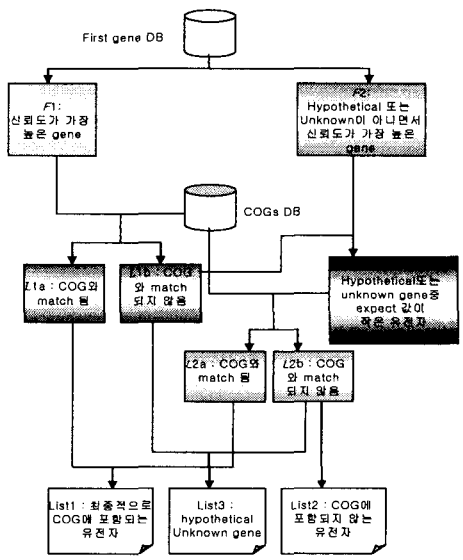
2. 유전자 목록 작성 방법

이제 구성된 시스템 하에서 유전자 목록을 작성하는 방법에 대해 알아보자. 그림 2는 유전자 목록 작성과정을 보여준다.

2.1 Gene list 작성 방법

Blastx 결과 값 에서 모든 정보를 파싱하여 저장한 데이터베이스를 D 라고 하고 D에서 가장 신뢰도가 높은 값만은 추출한 것을 F1, local로 구축된 COG 데이터베이스를 C 라 하자. 신뢰도가 높은 유전자 중에서 COG와 비교하여 1차 리스트를 작성한다. 여기서 COG와 match되는 것들의 리스트를 L1a라 하고 match 되지 않는 것들의 리스트를 L2a라 한다. L1a는 가장 신뢰성이 높은 결과라고 할 수 있다. 왜냐하면 expect 값이 가장 높은 것들 중에서 COG 와 비교하여 COG에 존재하는 것들만 추출한 결과이기 때문이다.

다음으로 D에서 Hypothetical~이거나 Unknown 이거나 Unnamed 인 것들을 제거한 것들 중에서 expect 값이 낮은 값만을 추출한 것을 L1a와 비교해서 L1a에 포함되는 Gene들을 제거한다. 그리고 그 제거한 것들을 F2라 하고 COG와 비교하여 match 되는 유전자 리스트를 L2a, match 되지 않는 것들을 L2b라 한다. 이렇게 하는 이유는 Fragment가 비교적 짧은 서열이기 때문에 신뢰도가 가장 높게 나온 것들이 Hypothetical~ 이거나 Unknown 이거나 Unnamed 이라 할 때 두 번째로 신뢰도가 높은 값들도 유전자일 가능성이 있기 때문이다. 위와 같은 방법으로 하면 두 번째로 신뢰도가 높은 값들을 COG와 비교 할 수 있다.



▶▶ 그림 2. Gene list 작성

2.2 Gene list 작성

Gene list 작성방법에 따라 나온 결과에 대해 Gene list를 작성 한다. 크게 list를 세 가지로 분류할 수 있는데 먼저 L1a와 L1b의 결과를 합해서 최종적으로 COG에 포함된 리스트를 작성한다. 이 리스트는 COG와 비교함으로서 유전자의 오솔로그 관계를 파악 및 필수 유전자를 파악할 수 있는 자료이다. 두 번째로 L2b를 바탕으로 리스트를 작성한다. 이 리스트는 COG에 포함이 되지 않는 유전자로 종 특이적 유전자를 파악할 수 있는 자료이다. 마지막으로 hypothetical~ 이거나 Unknown 이거나 Unnamed인 것들의 리스트를 작성하는 것이다. 이는 L1b와 L2b를 비교하여 추출 할 수 있다. 이 리스트들은 아직 기능이 밝혀지지는 않았지만 후에 연구를 통해 밝혀내야 하는 것들이다.

Ⅲ. 결론

본 논문에서는 Genome project의 초기 단계에서도 구체적인 유전자 정보를 알 수 있고 유전자 리스트를 작성 할 수 있는 방법을 제안하고 실제로 구현하였다. 우리는 우리가 제안한 방법과 시스템이 목표와 부합되는지에 대해 실제로 구현하여 실행해 본 결과 만족할 만한 유전자 목록을 얻을 수 있었다. 제안한 기법은 비교적 쉽게 유전자를 예측할 수 있도록 도와주며, 후속 연구를 앞당기는데 기여 할 것이다.

추후 본 시스템과 방법을 개량하고 다양한 데이터에 대하여 실험하여 편의성과 정확성을 높이고, 다양한 분석이 가능하도록 할 것이다.

▌참고문헌▐

[1] Tatusov, R.L., et al., The COG database: now developments in phylogenetic classification of proteins from complete genomes, Nucleic Acids Res., 29(1)22-28.

[2] T.L. Tatuson, E.V. Koonin, D.A. Natale, and M. Y. Galperin. Using the cog database to improve gene recognition in complete genomes, GENETICA - THE HAGUE - v. 108: 9-17, 2000

[3] http://www.ncbi.nlm.nih.gov/COG

[4] Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 25:3389-3402

[5] Altschul, S.F., et al., Basic local alignment search tool, J. Mol. Biol., 215:403-410.