Journal of Life Science 2008 Vol. 18. No. 4. 565~571 ©JLS / ISSN 1225-9918

다중 관계 그래프를 이용한 유전체 보존영역의 계층적 시각화와 개략적 전사 annotation 도구

이 도 훈*

부산대학교 정보컴퓨터공학부

Received April 7, 2008 / Accepted April 24, 2008

Rough Computational Annotation and Hierarchical Conserved Area Viewing Tool for Genomes Using Multiple Relation Graph. DoHoon Lee*. *School of Computer Science and Engineering Pusan National University Jangjeon-Dong, Keumjeong-Ku 609-735, Korea* - Due to rapid development of bio-informatics technologies, various biological data have been produced in silico. So now days complicated and large scale biodata are used to accomplish requirement of researcher. Developing visualization and annotation tool using them is still hot issues although those have been studied for a decade. However, diversity and various requirements of users make us hard to develop general purpose tool. In this paper, I propose a novel system, Genome Viewer and Annotation tool (GenoVA), to annotate and visualize among genomes using known information and multiple relation graph. There are several multiple alignment tools but they lose conserved area for complexity of its constrains. The GenoVA extracts all associated information between all pair genomes by extending pairwise alignment. High frequency conserved area and high BLAST score make a block node of relation graph. To represent multiple relation graph, the system connects among associated block nodes. Also the system shows the known information, COG, gene and hierarchical path of block node. In this case, the system can annotates missed area and unknown gene by navigating the special block node's clustering. I experimented ten bacteria genomes for extracting the feature to visualize and annotate among them. GenoVA also supports simple and rough computational annotation of new genome.

Key words : *Relation graph, annotation, visualization, comparative genome, multiple alignment*

서 론

생물정보학의 발달은 생물정보나 관련 연구결과를 대량으로 양산시키고 있다. 그럼에도 불구하고 이들을 효율적으로 관리하고 지원하는 범용시스템이 존재하기 힘든 것은 자료의 방대함은 물론이고 분야의 다양함과 그 분야에 요구되는 자료나 정보가 각각 다르기 때문이다. 이는 범용보다는 각 응용분야에 맞는 시각화나 사용자 편의의 도구가 더 필요하다는 뜻이다. 이러한 도구 개발은 다소 미흡한 결과라 할지라도 연구에 중요한 보조 역할을 하고 있다. 어떤 면에서 보면 새로운 발견의 계기가 될 뿐 만 아니라 그들 정보들 사이의 복잡한 관계를 시각적으로 보여주어 그 정보간의 관계를 이해하는 데 기여한다.

DNA나 단백질 서열의 비교 문제는 생물정보학에서 아주 오래된 문제이다. 이 기본적인 방법론의 연구는 단순한 두 서열간의 문제에서 끝나는 것이 아니라 다중 서열에 대한 정렬 및 다중 서열간의 공통점을 찾아내기 위한 시도의 출발점이 된다. 2000년 이후에 다중서열 정렬과 더불어 이를 보다 직관적으로 해석하고자 하는 도구 개발도 많이 이루어졌다[12]. 현재 활발한 연구 영역을 가지고 있는 annotation에 대한 연구와 제공되는 도구들이 꾸준히 발표되고 있다. GlimmerHMM [7]는 은닉마코브모델(Hidden Markov Model)에 기반한 유전자를 찾는 도구이다. 유전자 찾기는 미생물 유전체 annotation의 첫 번째 단계 중에 하나이고 유전자의 추후 annotation과 기능 발견을 위해 중요하다. Artemis [10]는 유전자를 보여주는 도구이고 서열자료의 시각화와 서열분석의 결과를 위한 annotation 도구이다. Lynn 등[6]은 GeneScan을 이용하여 코딩된 영역을 찾아 그것의 반복성을 알아보기 위해 푸리에변환을 사용하였다. 이 결과와 BLAST를 통해 annotation을 행하였다. 이외에도 Dnault 등[5]은 개선된 계층도 profile을 활용하여 박테리아 유전체 annotation하는 방법을 제안했고 Zhao 등[13]은 예측된 경로와 그 템플레이트 사이의 기능적, 구조적 일치도와 전체적으로 높은 서열 유사도를 이용하여 경로 annotation을 예측하는 방법을 제안했다. Stothard와 Wishart [12]는 현재 온라인에서 제공하는 박테리아 유전체 annotation 관련 소프트웨어와 DB들을 종합적으로 소개하였다. 최근에 McCauley 등[8]은 바이러스 유전체, 4종의 HIV2 서열, 3종의 Hepatitis B 서열을 다중 정렬하여 코딩영역을 annotation하는 방법을 제안하였다.

Artemis Comparison Tool (ACT) [1]는 완전한 유전체 서

*Corresponding author
Tel : +82-51-510-2491, Fax : +82-51-515-2208
E-mail : dohoon@pusan.ac.kr

열들과 관련된 annotaton간의 비교를 상호작용적으로 보여준다. 이 때 사용되는 비교 자료는 서로 다른 프로그램에서 얻은 자료를 사용할 수 있다. BLASTN, TBLASTX, Mummer을 이용하여 구현하였다. K-BROWSER [2]은 다중 정렬 유전체에 서로 얽혀있는 생물자료 정보를 직관적으로 보여준다. 특히 annotation되어 있고 관련 특징들이 예측되고, 또한 전체 관계를 다른 유전체에 다중 정렬된 불특정 여러 유전체를 동시에 보여준다. GenAlyzer [3]는 DNA와 단백질 서열간의 대칭되는 서열들을 상호작호하게 시각화하는 소프트웨어 도구이다. 이는 단계적으로 전체에서 부분 영역을 보여줄 수 있도록 지원하며 종국에서는 부분 서열을 정렬하는 영역을 보여줄 수 있다. 이는 매우 큰 dataset를 조절할 수 있고 많은 영역들의 서열간의 매칭을 보여준다. Mauve [4]는 재할당(rearrangement)과 수평 전달를 표현하는데 있어서 보존된 유전 DNA의 인식과 정렬을 위한 소프트웨어 패키지이다. Rasco [9]는 BLAST Score Ratio (BSR) 방법으로 세 가지 유전체에 있는 모든 추정 펩타이드를 BLAST 점수의 비를 기준으로 하여 유사도를 바탕으로 분류하였다. 이 분석의 결과는 모든 세 유전체사이의 단백질 유사도 정도를 전체적으로 보여줄 수 있다. 부가적인 결과로 부가된 각 유전체 쌍 간의 보존된 유전자 순서를 보여준다. 이런 일련의 정보들을 색깔을 부여하여 유사성이나 차원들을 표현하였다. 다종 간의 DNA 서열에 대한 유사도 측정을 시각화하는 프로그램이다. Phylo-VISTA [11]는 아주 큰 자료를 보여주기 위해 자료를 선택하고 다양한 형태의 해상도로 보여준다.

본 논문에서는 다중 유전체간 비교를 통해 새로운 유전체의 개략적 annotation을 전사적으로 수행하는 도구, GenoVA (Genome Viewer and Annotation)를 소개한다. 여러 유전체간의 보존영역을 찾기 위해 두 서열간의 정렬을 반복적으로 모든 주어진 유전체 쌍에 대하여 실행한다. 그 결과에서 보존되는 영역을 정점으로 하는 다중 관계그래프를 생성하고 생성된 그래프와 연결된 정점들이 클러스터를 만든다. 이 클러스터에 포함된 영역을 비교 분석함으로써 새로운 영역의 annotation이나 의미 있는 영역을 발견하는 도구를 제공한다.

재료 및 방법

기본 전략

본 논문에서 주장하는 방법의 기본적인 생각은 두 유전체간의 유사성에서 출발한다. 참조 유전체에 다른 주어진 $n-1$ 개의 유전체들을 연속적으로 비교하여 $n-1$개의 비교결과를 조합하여 중복되는 영역을 계산한다. 이와 같이 비교되었던 다른 유전체들을 차례로 참조유전체로 하여 각각에 대한 중복 영역을 계산한다. Fig. 1은 그 개념을 보여준다. 각 유전체의 어떤 부분이 다른 유전체의 특정 부분과 얼마만큼 중복되는 지를 직관적으로 알 수 기본 전략을 표현한 것이

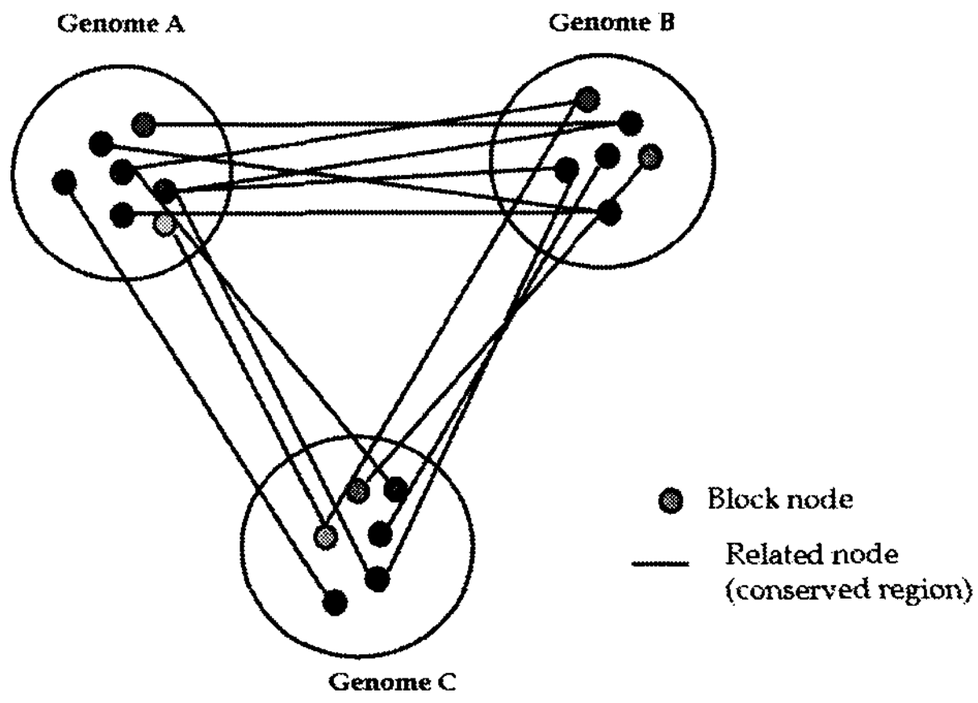


Fig. 1. The relationship of extended alignment among genomes. Conserved area has high score in alignment and the areas are called block node or vertex in multi-graph. The transitive traversal of multi-graph can give us finding unknown area.

다. 중복된 영역이 주어진 임계값 보다 큰 영역, 누계영역을 블록 혹은 블록노드라 한다. Fig. 2는 유전체 *E. coli K12* (NC_000913)을 참조 유전체로 했을때 다른 9개의 다른 유전체와 BLAST 점수 3,000 이상의 영역이 누계로 표현되는 영역이 발생되는 것을 볼 수 있다. 이 블럭은 다중 관계 그래프의 정점으로 설정된다. 이 블럭을 중심으로 다른 유전체와 관련된 블럭들을 계산하고 이들 관계를 시각적으로 보여준다. 연관된 관계를 그래프로 표현하고 이 블록들이 보존되는 영역들을 그래프를 이용해 추이적(transitive)으로 보여준다.

전처리 과정과 다중 관계 그래프 생성

본 논문에서 제시하는 관계 그래프를 구성하기 위해 전처리 과정을 거쳐야 한다. 기본적으로 두 개의 유전체를 먼저 BLAST를 이용하여 유사성을 검사한다. 이에 대해 $n-1$쌍의

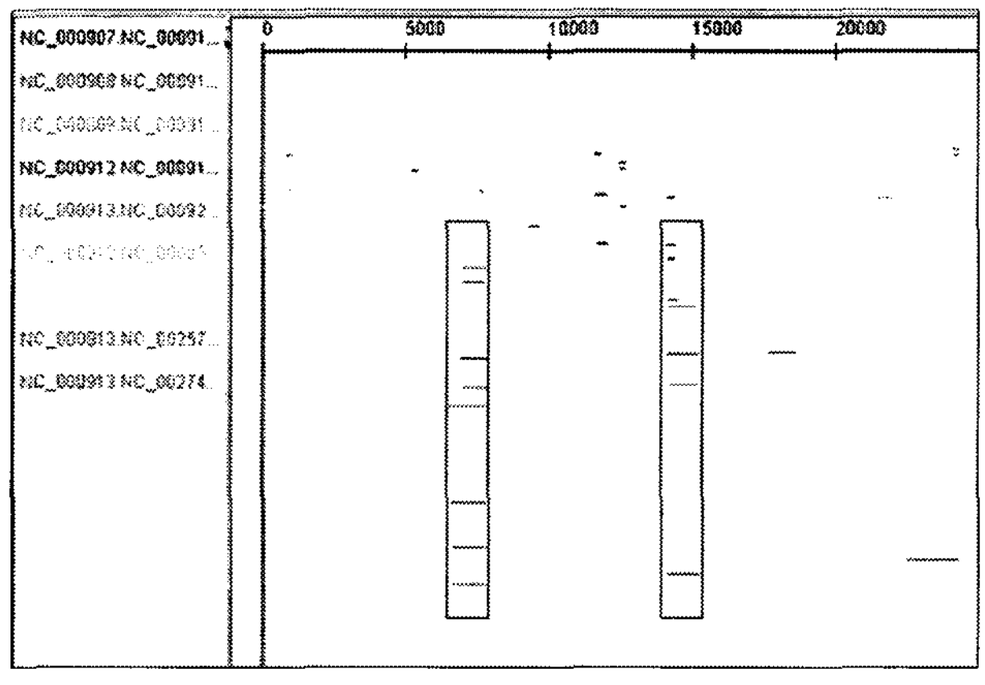


Fig. 2. Example of accumulating overlapped areas. The areas are over BLAST score 3,000 among, 9 genomes by reference genome *E. coli K12* (NC_000913).

유전체 유사성 검사를 한다. 결국 주어진 n개의 유전체에 대하여 $C(n,2)$의 비교를 한다. 이 같은 방법은 시간이 많이 걸리는 문제는 있지만 두 유전체간 보존영역을 가장 잘 알 수 있는 방법이다. 이 영역들을 처리하는데 유전체 크기 때문에 실시간 처리는 거의 불가능하며 관계 그래프를 생성하기 위해 전처리 과정을 거친다. 다중 관계 그래프를 생성하기 위한 전처리 과정은 Fig. 3에서 보이고 있다 그 과정은 정리하면 다음과 같다.

단계1. BLAST와 같은 잘 알려진 정렬 프로그램을 이용하여 정렬한다.

단계2. 한 참조 유전체를 기준으로 정렬을 n개의 유전체로 확대한다.

단계3. 다른 $n-1$개 참조 유전체에 대하여 단계2를 이용하여 정렬한다.

단계4. 일정 점수 이상의 영역을 발췌하여 서로 중복되는 영역을 누계하여 각 영역의 빈도수를 구한다(Fig. 2에서 특정 영역에 중복되는 영역이 많이 나타나고 있는 것을 알 수 있다.).

단계5. 구해진 영역의 빈도수가 일정 임계값을 넘으면 그 지역을 블럭노드로 지정한다. 연관된 블럭노드는 간선을 연결하여 다중 그래프를 생성한다.

다중 관계 그래프를 위한 정보는 전처리 단계에서 모두 생성된다. Fig. 4는 생성된 다중 관계 그래프의 블록노드와 다중으로 연결된 것을 개념적으로 보여준다. 같은 색깔은 보존된 영역이고 생성된 관계 그래프 정보는 특정 블록이 지정되면 그 블록노드에 연결된 모든 다른 유전체의 블록노드, 즉 클러스터링된 블록노드를 보여준다. 각 블럭노드는 각각

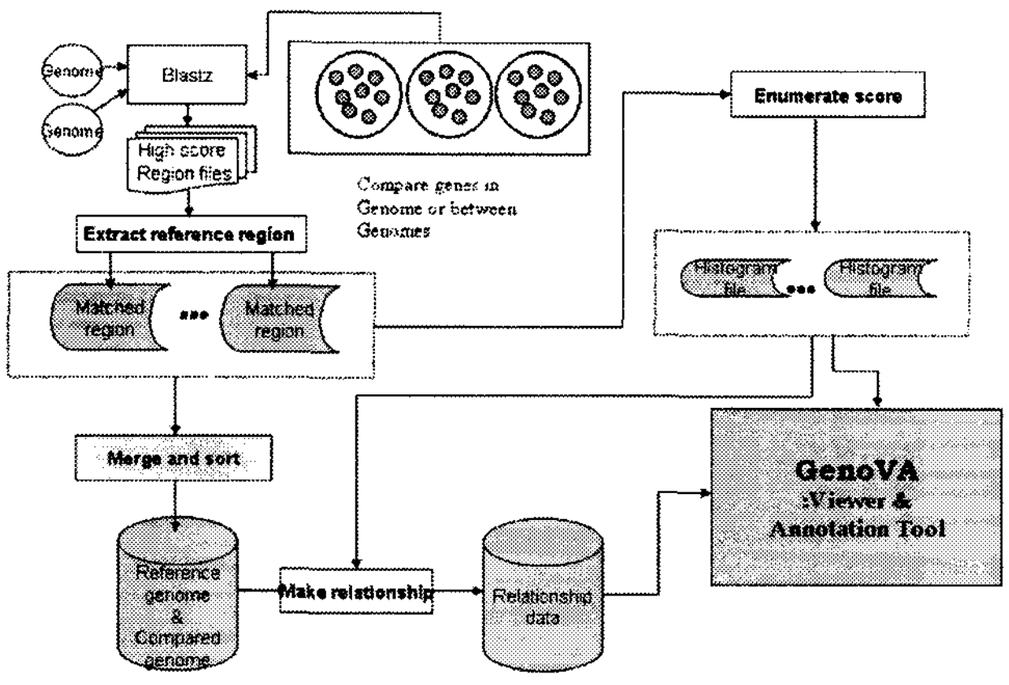


Fig. 3. Procedure of GenoVA. Generating block node of in a genome or among genomes and preprocess procedure for generating multi-graph using the block nodes. High frequency regions are extracted from comparison of all pairs. Then the results are merged and sorted for making relation graph. The relationship data and enumerated score are supported for GenoVA.

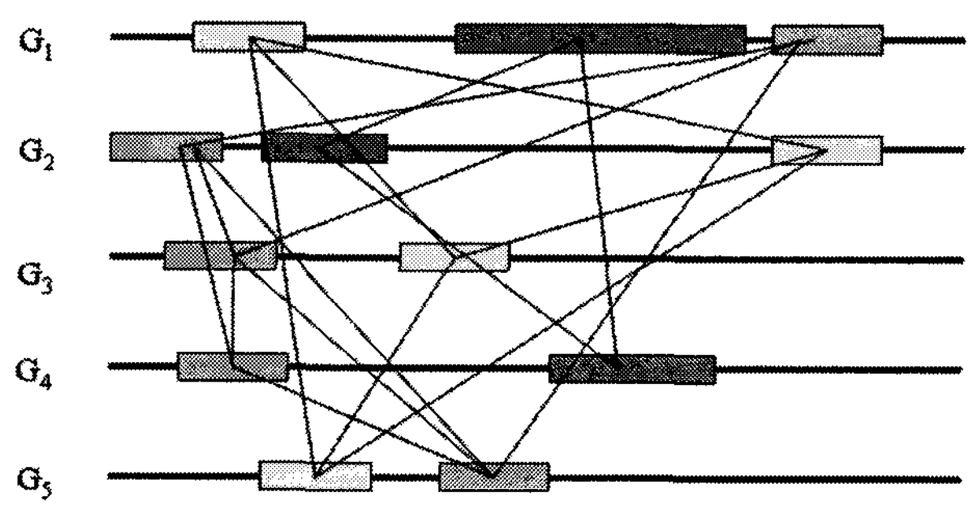


Fig. 4. Example of multi-graph $G(V,E)$ among genomes. Block node (vertex) V has many edges for representing conserved areas. For example, we can show that orange block node in reference genome G_1 conserved genome G_2 and G_4. Same colored vertices are considered as clustering.

에 대응되는 유전체나 유전자의 정보를 이용하여 GenoVA가 실행될 때 관계 그래프를 생성한다. 생성된 그래프를 모두 보이는 것은 너무 많은 간선을 가지고 있으므로 효율적이지 못하다. 따라서 GenoVA에서는 특정 블록노드와 연관된 정보, 그 블록노드의 클러스터링된 정보만을 보인다.

Genome Viewer and Annotation (GenoVA) 시스템 기능

GenoVA는 다음과 같은 기능을 갖는다. (1) 특정 영역이 자주 나타나는 것을 보존된 영역이라 간주하고 이를 블록 혹은 블록노드라 하고 연관된 블록노드를 연결하는 다중 관계 그래프 생성한다. (2) 의미있는 영역, 즉 블록노드를 각 유전체별로 시각화한다. (3) COG와 GenBank와 같이 이미 알려진 정보를 이용하여 각각의 위치에 annotation한다. (4) 특정 블럭의 클러스터링에서 누락되었거나 발견되지 않은 보존영역을 직관적 도출할 수 있다. 이 자료는 생물학적 확인을 위한 실험 자료로 활용할 수 있도록 한다. (5) 새로운 유전체 서열에 대해 개략적인 전사적 annotation한다. (6) 평균빈도수, 중복도, COG수, 관련 유전체간의 통계자료를 일관적으로 제공한다. (7) 참조 유전체를 중심으로 annotation된 자료를 제공한다. (8) 특정 블록의 클러스터링을 보여주고 그에 대응되는 서열정보 및 기존 ClustalW 실행 결과를 동시에 제공한다. (9) 블록노드 생성 때 BLAST 점수를 통해 필터링할 수 있는 기능을 제공한다. 빈도수가 높다든지 직관적으로 의미를 알 수 있는 정보는 대부분 연구를 통해 알려져 있다. 따라서 매우 애매한 영역에서 아직 발견하지 못한 영역을 도출하는 것이 생물학적으로 의미가 있다고 판단하기 때문이다.

Computational annotation

GenoVA의 블록노드는 생물학적 의미가 없는 빈도수와 점수에 의해 만들어진 영역이다. 새로운 유전체를 다른 다수의 유전체와 동시에 참조함으로써 직관적이고 매우 빠른 an-

notation 정보를 얻을 수 있다. 다중 관계 그래프로부터 보존된 영역들을 알 수 있고 그 블록에 대한 정보는 COG, GenBank를 활용하고 서열정보를 얻기 위해 FASTA 파일을 참조하여 annotation한다. Fig. 5는 이런 일련의 시스템 흐름도를 보이고 있다. Fig. 8은 실제로 한 유전체를 새로운 유전체라 가정하고 다른 유전체와 다중 관계 그래프를 활용한 개략적인 계산적 annotation한 예를 보이고 있다. 새로운 유전체를 NC_000000이라 하고 참조된 블록노드와 연결된 정보를 활용해 보다 정확한 annotation를 수행한다. 이 방법은 비록 블록노드가 생물학적인 의미가 없다고 하나 대부분 유전자와 의미 있는 영역이 대부분 블록노드에 포함되고 여러 유전체의 정보를 동시에 활용할 수 있기 때문에 보다 정확한 annotation를 보장한다.

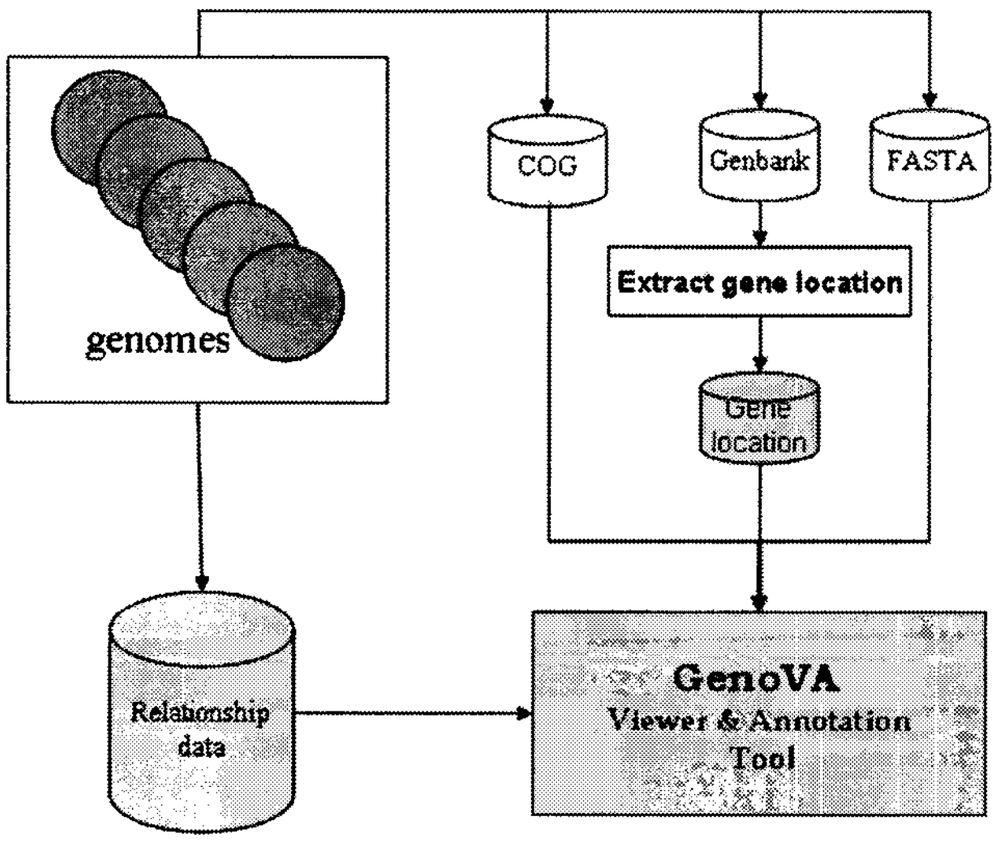


Fig. 5. Flow chart of annotating. The procedure shows annotation steps using multiple relation graph explained previous section. For more biological annotation, GenoVA use known information such as GenBank, COG, and FASTA.

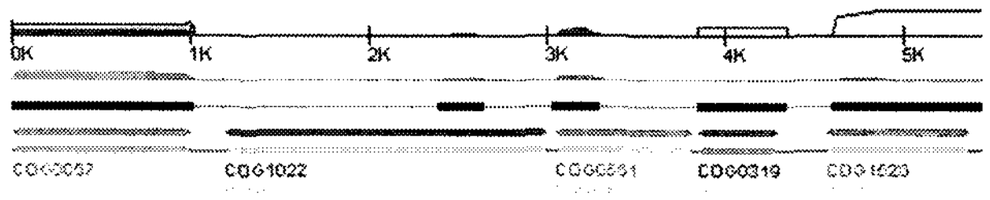
결 과

GenoVA의 구현언어는 JAVA를 사용하였으며 PC 펜티엄 CPU 2.8GHz, 512MB RAM이고 Microsoft XP 운영체제 환경에서 구현하였다. 비록 구현 환경이 PC와 XP환경이나 JAVA언어의 특징으로 GenoVA는 하드웨어 및 운영체제에 독립적인 도구이다. Table 1은 본 논문에서 실험한 10개의 박테리아 유전체와 크기, 그리고 간단한 통계들을 보이고 있다. 유전체의 크기, 보존영역의 비율, 블록노드의 비율, COG의 수, 유전자 수들이 나타난다. 이를 활용한다면 또 다른 형태의 계통도를 만들 수 있다. Fig. 6(a)와 (b)는 각 유전체에 비교되어 나타나는 빈도수, 블록노드, COG와 유전자가 annotation되어 있다. Fig. 6(c)는 참조 유전체 NC_000907의 한 블록노드의 클러스터링 중의 한 경로를 계층적으로 보이고 있다. NC_000913의 한 노드와 연결되고 그 노드는 다시 NC_000909의 노드에 추이적으로 연결되었다. 이 노드는 다시 참조 유전체의 다른 노드에 연결된다. 블록노드에 공통적으로 COG0513이 나타남으로써 각 노드는 family에 속한다는 것을 알 수 있다. 각 블록노드는 생물학적인 의미가 아니라 빈도수와 BLAST 점수에 의해 생성된 영역이므로 생물학적으로 정확히 일치하지 않는다. 이와 같은 블록노드에서 보다 자세한 정보를 위하여 FASTA 양식의 DNA서열을 볼 수 있으며 각각을 ClustalW로 다중 정렬 정보를 얻을 수 있다. Fig. 6(d)가 이를 위한 예를 보이고 있다.

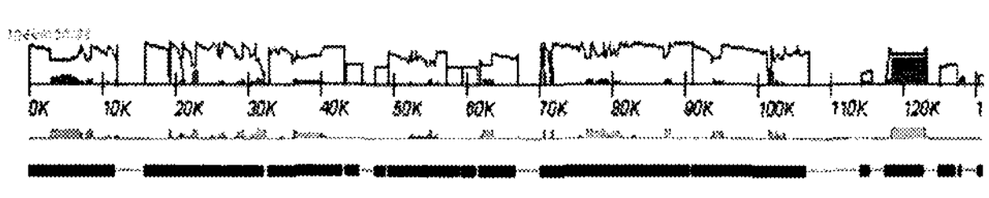
유전체간 보존 영역의 빈도가 매우 높거나 BLAST 점수가 높다는 것은 보존성이 높을 뿐만 아니라 그 영역에 대한 생물학적인 분석이 대부분 잘 되어 있다는 것을 의미한다. 그래서 많은 생물학자들은 그런 영역보다는 빈도수가 높지 않고 BLAST 점수가 높지 않지만 관련 있는 영역을 찾는데 더 큰 의미를 갖는다. GenoVA는 이를 위하여 블록노드를 점수와 블록노드의 길이를 제한함으로써 누락되었거나 아직 밝혀지지 않은 영역을 찾을 가능성이 높은 블록을 추출하는 기능을 제공한다. Fig. 7은 BLAST 점수를 3,000점에서 6,000점

Table 1. Sample genomes and statistical reports used in GenoVA

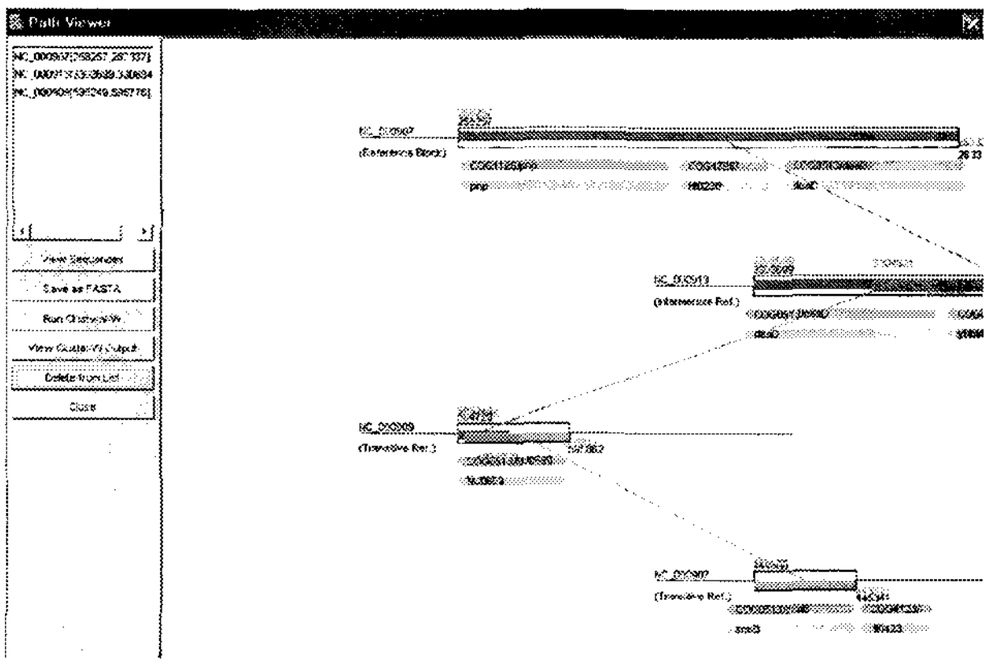
Acc_No	Genome Name	Size	Ratio of Cover	Ratio of BNCover	#CoG	#Gene
NC_000907	Haemophilus influenzae Rd	1830138	70.67516	70.67516	1657	1788
NC_000908	Mycoplasma genitalium	580074	96.75869	96.75869	484	523
NC_000909	Methanococcus jannaschi	1664970	18.773851	18.773851	1729	1773
NC_000912	Mycoplasma pneumoniae	816394	81.44205	81.44205	689	696
NC_000913	Escherichia coli K12	4639221	42.36166	42.36166	4279	4395
NC_000921	Helicobacter pylori, strain J9	1643831	29.61527	29.61527	1491	1495
NC_000962	Mycobacterium tuberculosis H37R	4411529	18.372633	18.372633	3927	3926
NC_000964	Bacillus subtilis	4214814	59.34563	59.34563	4112	4234
NC_002570	Bacillus halodurans	4202353	56.75249	56.75249	4066	4167
NC_002745	Staphylococcus aureus subsp. aureus N315	2814816	51.948795	51.948795	2593	2664



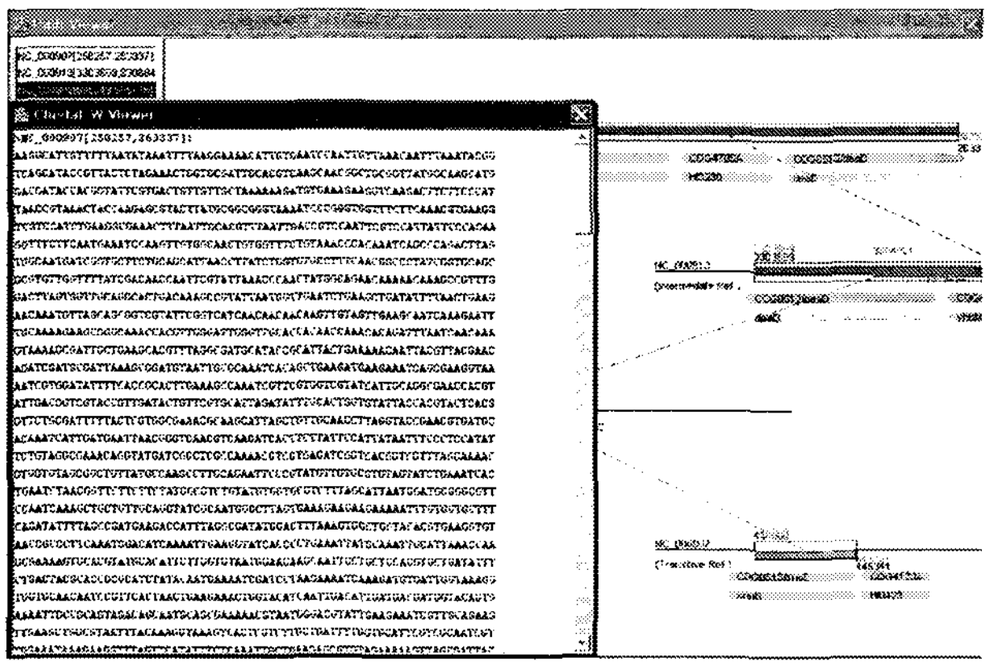
(a)



(b)



(c)



(d)

Fig. 6. Various functions of GenoVA. It has many functions for visualizing: block node, COG and gene location, clustering in multi-graph by clicking special block node. (a) Block node and gene/COG location. In this figure we can see that all gene are always not conserved. (b) Average frequency and no. of conserved genomes. (c) Example of the clustering in multi-graph hierarchically. This interface shows to handle multiple alignment given block node area using ClustalW and to display FASTA format sequence. A block node of *NC_000907* is connected a block node of *NC_000912*, *NC_000909* respectively. This transitive connection show that common family COG0513 is appeared. So these blocks contains same family. (d) Sequence of given clustering in multi-graph.

Fig. 7. Interface of filtering and its modified block node. Gray nodes indicate filter out when BLAST score range is 30,000~60,000 and length of block node is 100~5,000.

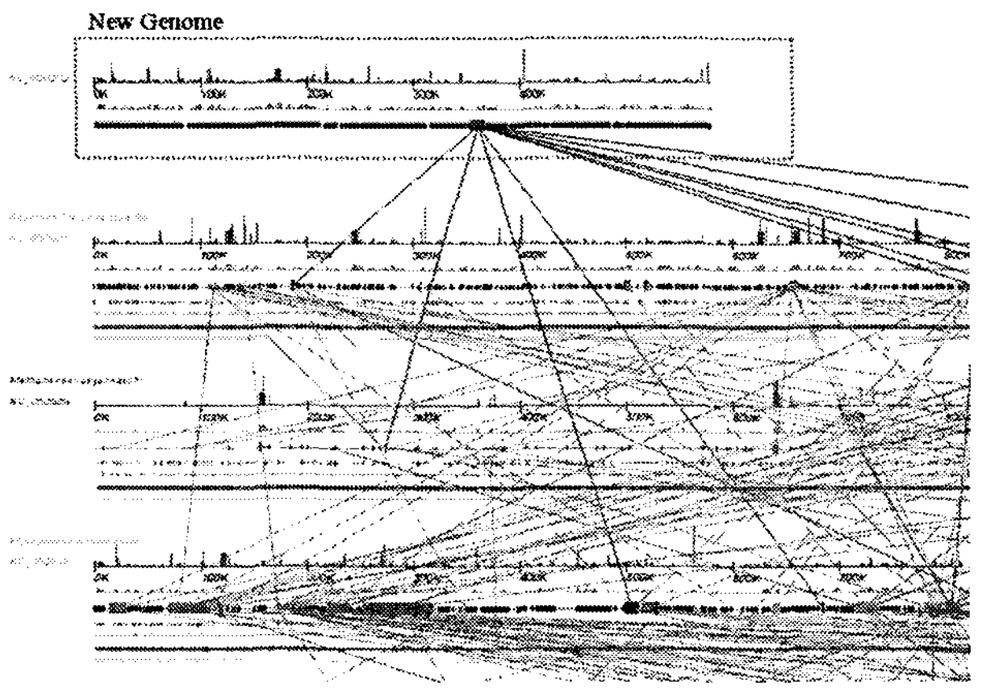


Fig. 8. Example fo rough computational annotation. New genome, called it NC_000000, is annotated easily using COG, GenBank and multi-graph. Reference block connected many other regions in difference genomes.

사이의 블록노드와 노드의 길이가 100 bp에서 5,000 bp가 되는 노드만을 추출할 수 있도록 했다. 그 결과 이 조건에 맞지 않은 블록노드는 모두 회색노드로 변한다.

제안한 시스템을 이용한다면 새로운 유전체 서열을 매우 간단하고 빠르게 annotation 할 수 있다. 참조 유전체가 새로운 유전체라고 한다면 비교되는 유전체가 많을수록 더 많은 annotation 정보를 얻을 수 있을 것이다. Fig. 8은 새로운 유전체내의 한 블록노드와 그 클러스터링 노드들을 보이고 있다. GenoVA는 이와 같이 annotation된 결과를 제공한다.

Table 2와 Table 3이 제공되는 annotation 정보의 일부이다. Table 2는 새로운 유전체에 대한 annotation 정보인데 각 블록노드에 연관된 비교된 유전체의 위치와 COG 정보 및 유전자 정보를 제공한다. Table 3은 잘 알려져 있는 유전체이지만 참조 유전체를 기준으로 비교되는 유전체간의 annotation

Table 2. Computational annotation result

```
Reference Genome : NC_000907, Haemophilus influenzae Rd,
                                        Genome length : 1830138
-------------------------------------------------------------------------
Block_No  | Block..Region || Related COGs(Name : PID : Gene)
            Compared Genome | Block..Region
          .... Related COGs of compared Genoem(Name:PID:Gene)
=========================================================
1 | 1..1028 || COG0057 : 16271977 : gapdH
        NC_002570 | 3661958..3662948
                  ....COG1508 : 15616125 : sigL
        NC_000913 | 1860794..1861789
                  ....COG0676 : 16129734 : yeaD
        NC_000964 | 2966149..2967128
                  ....COG1733 : 16079955 : ytcD
        NC_002745 | 832685..833638
                  ....COG1314 : 15926455 : secG
                  ....COG1647 : 15926456 : -
                  ....COG0557 : 15926457 : rnr
                  ....COG0691 : 15926458 : ssrP
        NC_002570 | 3263806..3264825
                  ....COG0330 : 15615716 : hflC
        NC_002745 | 1719433..1720400
                  ....COG0642 : 15927270 : phoR
                  ....COG0745 : 15927271 : phoP
        NC_000964 | 3480779..3481768
                  ....COG0477 : 16080449 : araE
        NC_000962 | 1613306..1614282
                  ....COG0243 : 15608580 : bisC
        NC_000908 | 371163..371941
                  ....- : 12045168 : hmw1
                  ....- : 12045169 : -
                  ....- : 12045170 : -
                  ---------------------------
2 | 2394..2655 || COG1022 : 16271978 : -
        NC_000962 | 1754765..1755024
                  ....COG1053 : 15608690 : frdA
                  ....COG0479 : 15608691 : frdB
                  ....COG3029 : 15608692 : frdC
                  ....COG3080 : 15608693 : frdD
        NC_002745 | 272498..272721
                  ....COG4670 : 15925938 : -
        NC_002570 | 2103338..2103512
                  ....COG1012 : 15614573 : BH2010
        NC_002570 | 3219822..3219949
                  ....COG0072 : 15615672 : pheT
                  ....COG0016 : 15615673 : pheS
                  ....COG0566 : 15615674 : BH3112
                  ....- : 15615675 : BH3113
3 | 3036..3311 || COG0561 : 16271979 : -
        NC_002745 | 603939..604211
                  ....COG0438 : 15926243 : -
        NC_002570 | 1248136..1248380
                  ....COG0745 : 15613716 : BH1153
                  ....COG0642 : 15613717 : BH1154
        NC_000913 | 3873771..3874006
                  ....COG0187 : 16131567 : gyrB
                  ....COG1195 : 16131568 : recF
                  ....COG0592 : 16131569 : dnaN
                  ....COG0593 : 16131570 : dnaA
        NC_002745 | 942857..943027
```

Reference genome is *Haemophilus* influenzae Rd (NC_000907) and location of conserved region, COG and Gene of reference genome, compared genome respectively.

Table 3. Computational annotation of new genome

```
Reference Genome : NC_000000, null, Genome length : 580074
-------------------------------------------------------------------------
Block_No  | Block..Region || Related COGs(Name : PID : Gene)
            Compared Genome | Block..Region
          .... Related COGs of compared Genoem(Name:PID:Gene)
=========================================================
1 | 51..28488
        NC_000912 | 131..12059
                  ....COG0358 : 13507753 : dnaE
                  ....COG0189 : 13507754 : -
                  ....COG0189 : 13507755 : rimK
                  ....COG0190 : 13507756 : mtd1
                  ....COG1132 : 13507757 : pmd1
                  ---------------------------
13 | 214028..214636
        NC_000912 | 190406..191000
                  ....- : 13507885 : -
                  ....- : 13507886 : -
        NC_000912 | 170191..170782
                  ....COG1525 : 13507872 : -
                  ....COG3839 : 13507873 : ugpC
                  ....COG1175 : 13507874 : ugpA
                  ....COG0395 : 13507875 : ugpE
        NC_000912 | 571178..571404
                  ....- : 13508213 : -
        NC_000912 | 342026..342366
                  ....- : 13508027 : -
        NC_000912 | 246094..246422
                  ....- : 13507944 : -
        NC_000912 | 441818..442032
                  ....- : 13508109 : -
                  ....- : 13508110 : -
        NC_000912 | 606607..606912
                  ....- : 13508241 : -
                  ---------------------------------
14 | 215022..215399
        NC_000912 | 566532..566865
                  ....- : 13508206 : -
        NC_000912 | 186285..186672
                  ....- : 13507881 : orf6
                  ....- : 13507882 : -
        NC_000912 | 132218..132542
                  ....COG0016 : 13507844 : pheS
                  ....COG0072 : 13507845 : pheT
        NC_000912 | 497963..498284
                  ....- : 13508154 : P37
                  ....COG1120 : 13508155 : P29
                  ....COG3639 : 13508156 : P69
                  ....COG0816 : 13508157 : -
                  ....COG0013 : 13508158 : alaS
                  ....COG0584 : 13508159 : glpQ
                  ....COG0477 : 13508160 : -
                  ....COG0482 : 13508161 : -
                  ....- : 13508162 : -
                  ....COG2739 : 13508163 : ylxM
                  ....COG0552 : 13508164 : ftsY
                  ....COG1196 : 13508165 : -
        NC_000912 | 437231..437606
                  ....- : 13508106 : -
        NC_000912 | 197258..197580
                  ....- : 13507891 : -
        NC_000912 | 611094..611284
                  ....- : 13508245 : -
```

All block node can be annotated in detail via simple tool such as BLAST and ClustalW.

된 정보를 보여주고 있다. COG나 유전자가 잘 알려진 정보는 그 family나 이름이 표시되지만 아직 알려지지 않은 부분을 아무것도 표시되지 않는다. 이 표에서 보여주는 빈공란이 항상 생물학적인 의미를 갖는 것은 아니다. 그 이유는 블록노드가 생물학적 의미를 갖는 영역이 아니기 때문이다.

본 논문은 박테리아 유전체들의 DNA 서열을 기본으로 하여 다중 정렬을 서열쌍 정렬 프로그램인 BLAST를 모든 유전체간에 수행하여 그 결과에 따라 블록노드를 만들고 다중 관계그래프로 클러스터링하였다. 이 다중 관계 그래프를 활용하여 annotation된 결과를 시각화하고 특정 블록노드의 경로를 계층적으로 보여줌으로써 누락된 정보를 시각적으로 인지할 수 있도록 했다. 또한 새로운 유전체 annotation을 위하여 생물학적인 의미보다는 클러스터링된 정보를 빠르게 적용함으로써 개략적인 전사 annotation을 수행할 수 있는 도구이다. 다만 이런 결과는 그대로 접목하기 보다는 생물학적 의미를 부여하기 위해 보다 부가적인 작업을 수행해야 하는 단점을 가지고 있다. 향후 이런 생물학적 annotation을 위해 부가적인 기능을 보완해야 한다.

요 약

생물정보학의 발전으로 다양한 형태의 생물정보가 컴퓨터 프로그램에 의해 양산되고 있다. 단순한 서열간의 비교나 작은 규모의 자료를 처리하기 보다는 다각화된 정보와 대규모의 생물정보를 취급하고 있다. 그 중에서 시각화와 annotation를 위한 도구개발은 지난 10년간 많은 연구가 되고 있는 분야이다. 그럼에도 일반화된 도구 개발은 생물정보의 다양성과 사용자 요구의 다양화로 인해 매우 어렵다. 본 논문에서는 유전체간 알려진 정보와 다중 관계 그래프를 이용하여 이를 annotation하고 시각화하는 GenoVA 시스템을 제안한다. 다중 정렬을 위한 몇 개의 프로그램이 존재하지만 그 방법들이 서열내의 복잡성 때문에 많은 정보가 누락된다. 따라서 제안된 방법에서는 pairwise alignment를 확장하여 모든 유전체간 비교를 통해 연관성 도출한다. 유전체간 보존되는 영역의 빈도수와 BLAST 점수가 높은 것을 블록노드라 하고 이들 간의 연관관계를 다중 관계 그래프로 표현하였다. 또한 GenoVA는 알려진 정보, COG, 유전자를 시각화하고 다중 관계 그래프의 한 영역을 중심으로 클러스터링된 경로를 계층적으로 보여주었다. 이때 누락되거나 알려지지 않은 유전자나 다른 annotation 정보 추출할 수 있다. 본 논문의 실험을 위해 열 개의 박테리아 유전체가 사용되었고 시각화와 annotation을 위한 자료로 활용하였다. GenoVA는 새로운 유전체에 대한 개략적이고 전산적 annotation을 직관적이고 편리하게 제공한다.

감사의 글

이 논문은 부산대학교 자유과제 학술연구비에 의해 지원되었음.

References

1. Carver, T. J., K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell and J. Parkhill. 2005. ACT: the artemis comparison tool. *Bioinformatics* **21**, 3422-3423.
2. Chakrabarti, K. and L. Pachter, 2004. Visualization of multiple genome annotations and alignments With the K-BROWSER. *Genome Res.* **14,** 716-720.
3. Choudhuri, J. V., C. Schleiermacher, S. Kurtz and R. Giegerich. 2004. GenAlyzer: interactive visualization of sequence similarities between entire genomes. *Bioinformatics* **20,** 1964-1965.
4. Darling, A. C., B. Mau, F. R. Blattner and N. T. Perna. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14,** 1394-1403.
5. Enault, F., K. Suhre, C. Abergel, O. Poirot and J. M. Claverie. 2003. Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics* **19**. i105-i107.
6. Lynn, A. M., C. K. Jain, K. Kosalai, P. Barman, N. Thakur, H. Batra and A. Bhattacharya. 2001. An automated annotation tool for genomic DNA sequences using GeneScan and BLAST. *J. of Genetics* **80**, 9-16.
7. Majoros, W. H., M. Pertea, and S. L. Salzberg. 2004. TigrScan and GlimmerHMM: two open-source ab initio eukaryotic gene-finders. *Bioinformatics* **20** 2878-2879.
8. McCauley, S., S. de Groot, T. Mailund and J. Hein. 2007. Annotation of Selection strengths in viral genomes. Bioinformatics **23**, 2978-2986.
9. Rasko, D. A. 2005. Visualization of comparative genomic analyses by BLAST score ratio. MC *Bioinformatics* **6**, 1471-2105.
10. Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream and B. Barrell. 2000. Artemis:sequence visualization and annotation. *Bioinformatics* **16**, 944-945.
11. Shah, N., O. Couronne, L. A. Pennacchio, M. Brudno, S. Batzoglou, E. W. Bethel, E. M. Rubin, B. Hamann and I. Dubchak. 2004. Phylo-VISTA: interactive visualization of multiple DNA sequence alignments. *Bioinformatics* **20**, 636-643.
12. Stothard, P. and D. S. Wishart. 2006. Automated bacterial genome analysis and annotation. *Current Opinion in Microbilogy* **9,** 505-510.
13. Zhao, J., D. Che and L. Cai. 2006. Comparative pathway annotation with protein-DNA interaction and operon information via graph tree decomposition. *Proc. of Pacific Symposium on Biocomputing* **12,** 496-507.