전 유전체 가시화를 위한 Component Software 개발

For whole genomes which is readablility component software development

정우근
Chung Woo-Keun
부산대학교 컴퓨터공학과
wkchung@pusan.ac.kr

ABSTRACT

정보화 시대에 있어서, 많은 양의 데이터들을 접하게 됨에 따라 대용량 데이터를 사용자에게 가독성 있게 시각화 하기위한 문제는 최근 중요한 이수가 되오 있다.일반적으로 대용량 데이터를 시각화하기 위해서는 브라우저의 수평바, 또는 수직바와 같은 인터페이스를 사용한다. 이와 같은 방법으로는 방대한 양을 가지는 여러 가지의 정보를 한 눈에 비교하기는 아주 힘든 일이다. 본 보고서에서는 에서는 방대한 양의 데이터, 또는 정보들을 가독성있게 보여주는 Component Software를 제안한다. 이 방법은 다수의 대용량들을 동시에 가독성있게 시각화 하기 위한 시스템이다. 이를 증명하기위해 대용량 데이터에 적합한 실험데이터로써, 'HERV(Human Endogenous retrovirus)를 사용하였다. 본 보고서에서는 가독성 있는 Component Software를 제안하고, 'HERV' 데이터를 적용하여 성능을 측정하여 본다.

KEYWORDS Visualization, Component Software, Bioinformatics

1 Motivation

우리는 일상 생활에 있어서 많은 것을 보고, 접하게 된다. 그 중에서도 스크린, 모니터 등 시각화를 제공하는 것을 통하여 정보, 또는 데이터들을 보고 접하게 된다. 하지만 우리는 이 것들을 가독성있게 또는 많은 데이터 양을 비교, 또는분석을 위하여 와이드 모니터 또는 대형 스크린 화면에 해당 하는 정보, 또는 데이터를 시각화하여 접할 수 있다. 모니터에서 이러한 데이터 및 정보들을 접할 경우, 기본적인 시각화 툴에서 제공하는 수직바, 또는 수평바를 통하여 볼수 있다. 하지만 시각화 툴에서 제공하는 이러한 인터페이스로도 충분히 가독성 있는 시각화를 제공하나, 큰 용량을 가지고 있는 데이터 또는 수많은 데이터들을 동시에 비교하기에는 다소 어려움이 존재한다. 본 논문에서는 이러한 어려움을 해결하고자 가독성있는 제공해주는 Componenet Software를 제시한다. 이 다음장에서는 위와 같은 데이터들 즉, 큰 용량과 많은 정보를 가지고 있는 데이터들에 대하여 가독성있는 가시화를 제공해주는 Software에 대하여 조사해보았다.

2 Component Software의 전체 구조 및 주요 클래스

본단락에서는 본 논문에서 제시한 Component Software의 주요 클래스 및 함수들에 대하여 설명하겠다. 본 논문에서 제시한 Component Software는 Java로 되어 있으며, 직접적인 데이타 출력 및 가시화를 제공하는 것은 Java Applet이다. Java 언어의 플랫폼 독립적 특성 때문에 다양한 시스템 환경을 고려하여 Java를 선택하였다. 이러한 환경 및 사용자의 고려가 Java가 좋은 언어가 된다. 데이타 저장 및 Load는 MySQL 이 담당하고 있다. 이제 각 중요 클래스와 함수에 대하여 알아보도록 하자.

그림 1. 본 그림은 DNA Sequence의 정보를 보여주고 있는 그림이다. 하지만 우측에서부터 정보가 출력되지 않아 가독성이 떨어지는 것을 알수가 있다.

2.1 메인 클래스

본 단락에서는 사용자에게 직접적인 출력과 Interface을 제공하는 클래스를 설명하고자 한다. 메인 클래스는 Java 언어로 이루어져 있으며, Java 언어의 Applet으로 이루어져 있다. Java의 플랫폼 독립적 특성 때문에 본 논문에서 제시한 Component Software는 사용자의 다양한 OS 환경에도 무리없이 동작 할수 있다.

2.2 입력 처리 클래스

본 단락에서는 사용자의 입력 담당 및 입력에 대한 쿼리 저장 및 쿼리 결과등을 담당하는 클래스를 설명하고자 한다. InputInterfacePanel는 사용자에게 GUI를 제공하며 GUI를 통하여 입력을 받는다. 이 입력받은 사항을 QueryInputManager에게 전달하여 사용자 선택에 따른 사항들을 쿼리로 바꾸고, 현재 쿼리를 저장한다. 이 쿼리 값을 QueryResultManager에 전송한다. QueryResultManager는 이 쿼리 값을 저장시키고, 이 쿼리 명령에 따라 서버의 ServerAgent 클래스와 연동하여 검색을 시작한다. 검색이 끝나면 검색 결과 값을 넘겨 받아. QueryResultManager에게 이 값을 전달한다. ServerAgent 클래스의 주요 기능은 서버상태 확인과 서버와의 접속이며, 접속 방법은 TCP/IP이다. 본 클래스는 사용자의 선택에 따라 용량이 큰 데이터에 나타나는 정보들을 검색하고 가져오는 역할을 담당한다.

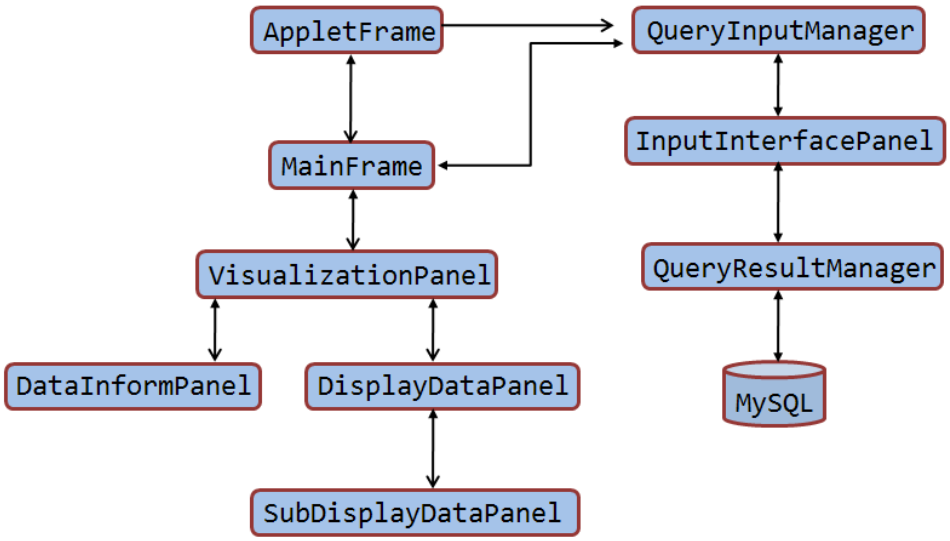

그림 2. 본 논문에서 제시한 Componenet Software의 구조

2.3 시각화 클래스

본 단락에서는 본 논문에서 제시한 가독성 높은 시각화를 제공하는 Component Software 에서 직접적인 시각화를 담당하는 클래스를 설명하고자 한다. VisualizationPanel 클래스는 전 다락에서 앞서 말한 QueryInputManager 클래스에 저장된 사용자의 선택에 따라 용량이 큰 데이터에 나타나는 정보들을 검색하고 가져오는 역할을 담당한다. 한 예로 그림 3에서 보이는 것이 QuryInputManager 클래스에 사용자의 검색에 따른 정보값을 쿼리로 저장 시켜 그 결과 값을 본 논문에서 제공하는 Component Software에 적용 시킨 사례이다. 빨간 사각형 안에 들어가 있는 것이 바로 사용자의 선택에 따른 쿼리 값에 존재하는 것들 가독성 높은 시각화를 제공한 것이다. 가로의 막대가 전체의 구간을 의미하는 것이고, 검은 실선들이 이 전체의 구간에 나타나는 정보 값들이다. 녹색 박스의 값이 바로 이 특정 데이터의 크기를 나타내고 있다. DataInformPanel은 사용자가 InputInterfacePanel에서 검색하거나 입력한 정보값 또는 쿼리 결과에서 가독성 높은 데이터를 출력하는 것을 제외 하고 정보에 해당하는 것을 출력해주는 것을 나타내고 있다. 데이터의 크기, 또는 데이터이름등을 가지고 있다. DisplayDataPanel, SubDisplayDataPanel은 VisualzationPanel 클래스의 보조 역할을 맡고 있다. 이 클래스들은 방대한 양의 데이터들을 가독성높은 가시화를 제공하기 위하여, 직접적인 계산작업과 출력하는 작업을 담당하고 있는 클래스이다.

그림 3. Component Software의 초기 단계의 그림이다. 빨간색 테두리안의 정보가 가독성높은 시각화를 제공하는 것이고, 녹색 테두리안의 정보가 현재 데이터의 크기를 보여주고 있다.

3 Data 준비

본 단락에서는 본 논문에서 제시한 Component Software의 실험을 적용한 결과를 설명하고자 한다. 실험에 임하기 앞서 실험 데이터에 대하여 살펴보자.

3.1 HERV

ERV(endogenous retrovirus)는 RNA 바이러스의 한 유형이다. 유전자로써 RNA와 이 RNA를 숙주의 DNA에 끼워 넣기 위한 역전사효소를 가지고 있으며, 단백질 캡시드는 숙주 세포에서 빠져나올 때 다음 숙주 세포로 침입하기 위한 막단백질을 붙여놓은 외막으로 둘러싸여 있다. 대개의 레트로 바이러스는 숙주 세포를 죽이지 않고 이용만 하기 대문에, 면역 체계가 바이러스만 인식해서 공격하는 것만으로는 레트로바이러스에 이미 감염된 세포를 없애지 못하므로 대부분 만성 감염으로 진행하게 된다. 따라서 혈액 중의 바이러스 숫자는 극히 적지만 감염자체는 계속 진행되는 형상이 나타난다. 따라서 치료를 위해서는 레트로바이러스의 감염 자체를 막는 방법 또는 레트로바이러스에 감염된 세포를 파괴하는 방법이 필요하다. 인간에 감염되는 대표적인 레트포 바이러스는 C형 간염 바이러스가 있다. 인간의 genome 상에는 다수의 ERV가 존재하고 있으며, 약 8%를 점유하고 있다. 인간 내생레트로바이러스(HERV)는 인간의 전 genome 상에 넓게 분포하고 있다.
이와 같이 본 논문에서 제시한 Component Software에 실험데이터로 HERV를 사용하기에는 부족함이 없는 듯하다. 또한 본 논문에서는 인간뿐만이 아닌, 다른 종들에 대해서도 ERV 데이터를 적용시켜 보았다. 다음 단락에서는 HERV 데이터의 저장한 방법과 실험을 적용할 결과를 살펴보겠다.

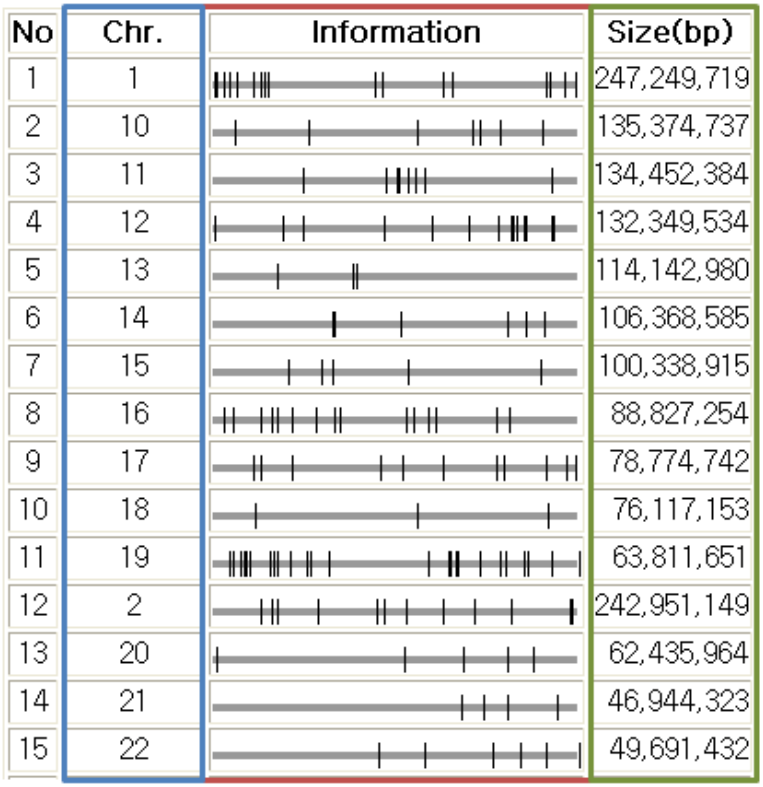

그림 4. 본 논문에서 제시한 가독성 높은 시각화를 제공하는 툴에 HERV 데이터를 적용 시킨 모습.

3.2 HERV 적용

본 단락에서는 앞 단락에서 선보였던 HERV 데이터로서 본 논문에서 제시한 Component Software의 실험에 적용할 것이다. 실험에 들어가기 앞서 하나의 HERV 데이터로써 실험한 결과를 살펴보자. 그림 **??** 은 하나의 HERV 데이터로써 실험한 결과이다. 테이블 안에 빨간색 테두리 안의 정보 값은 가독성 높은 시각화를 제공하고 있는 것이고, 녹색 테두리 안의 정보 값은 앞서 말한 가독성 높은 시각화를 제공하려 하는 데이터의 크기를 나타내고 있다. 본 논문에서는 가독성 높은 시각화를 제공하면서 많은 데이터를 동시에 시각화하며, 비교함으로써 효율을 높이는 것을 목적으로 하고 있다. 구체적인 결과 값을 나타내기 위하여, HERV 데이터를 데이터베이스로 구성하기로 하였다. 표 **??** 에 보이는 것이 현재 데이터베이스에 저장되어 있는 HERV 테이블의 구조이다. DBMS로써는 MySQL를 사용하였다.

표 1. HERV 데이터가 저장되어 있는 구조를 보여주고 있다.

Species	hName	SCORE	START	END	QSIZE	IDEN	CHRO	rand	hStart	hEnd	Span
HU	HERV-FC1	4338	1	4619	4629	97.2%	7	-	6393061	63937625	4565
CH	HERV4_I	2516	936	5797	6339	86.2%	5	+	34789017	34795601	6585
OR	HERV49I	468	4077	4820	6331	88.5%	8	+	177856002	177856921	920
RH	HERV19I	868	1	1171	5586	90.7%	X	+	112233067	112234651	1585

MySQL에 저장된 HERV 정보를 이용하여 본 논문에서 제시한 Component Software에 적용 시켜 보았다.

그림 ?? 에 보이는 빨간색 테두리는 각 데이터들의 크기를 나타내며, 빨간색 테두리는 순차적으로 해당 데이터들의 정보를 시각화하고 있으며, 파란색 테두리는 가독성 높은 시각화를 제외 하고 나타날 수 있는 정보 값이다.

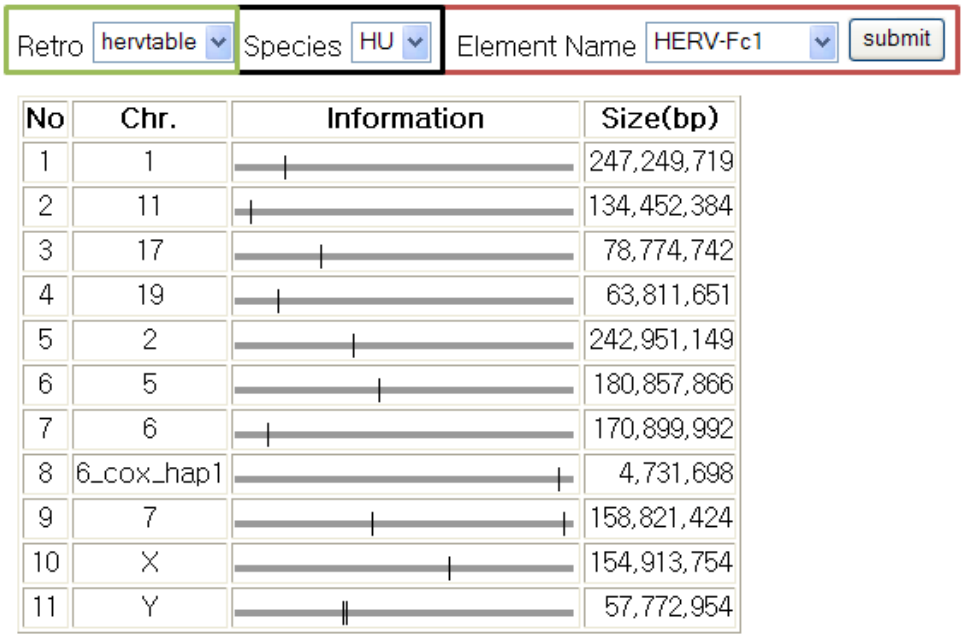

그림 5. 많은 데이터를 한 눈에 보여주는 결과물. 녹색, 검정, 빨간 색 테두리 안의 Option 버튼은 각 종 또는 사용자의 선택적인 선택을 제공해주고 있으며, 빨간색 테두리안의 Submit 버튼이 시각화를 제공한다.

4 결과 및 추후 과제.

우리는 방대한 양의 데이터, 또는 가독성이 떨어지는 정보들을 접하게 된다. 방대한 양의 데이터, 또는 가독성이 떨어지는 정보들을 가독성 있게 보기위하여, 다양한 시각화를 제공하는 툴을 제공한다. 하지만 현재 제공되는 시각화 툴을 사용하더라고 좀 더 다양한 정보와 비교 분석을 필요할 때가 존재한다. 그리하여 우리는 이러한 어려움을 해결하고자 본 논문에서는 방대한 양의 데이터, 또는 가독성이 떨어지는 정보들에 대하여 가독성이 높은 시각화를 제공하는 Component Software를 제시하였다. 본 논문에서 제시한 Component Software는 Java 언어로 이루어져 있다. Java 언어의 플랫폼 독립적 특성 때문에 다양한 시스템 환경을 고려하는 특성이 사용자의 환경을 고려하지 않아도 된다. 실험 데이터로써는 ERV중 하나인 HERV를 사용하였다. ERV는 RNA의 바이러스의 한 유형이며, 인간의 genome 상에는 다수의 ERV가 넓게 분포 하고 있다. 이러한 HERV 특성이 본 논문에서 제시한 가독성 높은 시각화를 제공하는 Component Software의 좋은 예가 되었다. 본 시스템에 HERV 데이터를 적용한 결과 가독성 높은 시각화를 제공함과 동시에 각 데이터 간에 비교 분석을 할 수 있는 기능 까지 제공하였다. 본 시스템의 구현으로써 넓게 분포 하고 있는 정보 데이터를 가독성이 높은 시각화를 제공 할 수 있게 되었다.

추후 과제로써는 현재 본 논문에서 제시한 시스템에서 가독성 높은 가시화와 가시화를 제공하는 것을 제외한 정보를 표시 해주고 있으나, 가독성 높은 가시화의 정보중에 즉, 그림 ??에 존재하는 수직 막대에 대한

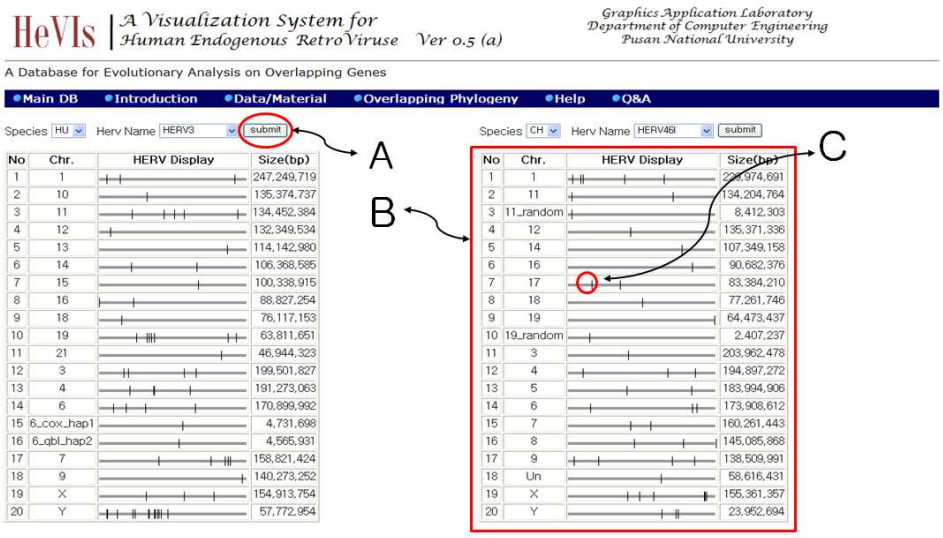

그림 6. 본 논문에서 제안한 컴포넌트 웨어의 결과물이다. 가독성있는 가시화를 보고싶은 데이터 값을 선택하여 (A) Submit을 통하여 데이터를 전송하면 (B) 와 같은 결과물이 제공된다. (C) 각 종에 존재하는 HERV 데이터 들이다.

정보 값들이 존재하지 않고 있다. 막대한 양의 데이터를 함축 시켜 시각화를 제공하였기 때문에 작은 수직 막대 하나에도 정보들이 존재하고 있다. 이와 같이 작은 데이터 하나하나에 대한 정보값을 제공, 사용자에게 조금 더 다양한 이벤트를 제공한다는 것이다. 본 시스템에서 제공한 가독성 높은 가시화를 제공하는 한편, 사용자에게 다양한 이벤트를 제공하여, 본 시스템에서 제공한 시각화 중 작은 정보까지 좀 더 세밀하게 제공할 수 있다면, 본 시스템을 사용하는 사용자들의 연구 활동에 좀 더 효과적인 툴로 자리잡게 될 것이다.

참고 문헌

1. Joanna Jakubowska, Ela Hunt, Matthew Chalmers, Martin McBride, and Anna F. Dominiczak, "Visgenome," *Bioinformatics*, vol. 23, no. 19, pp. 2641–2642, 2007.
2. Kim Rutherford, Julian Parkhill, James Crook, Terry Horsnell, Peter Rice, Marie-Adele Rajandream, and Bart Barrell, "Artemis: sequence visualization and annotation," *Bioinformatics*, vol. 16, no. 10, pp. 944–945, 2000.
3. Ela Hunt and Neil Hanlon, "Syntenyvista," in *NordiCHI '04: Proceedings of the third Nordic conference on Human-computer interaction*, New York, NY, USA, 2004, pp. 455–456, ACM.
4. David Nix and Michael Eisen, "Gata: a graphic alignment tool for comparative sequence analysis," *BMC Bioinformatics*, vol. 6, no. 1, pp. 9, 2005.
5. SE Lewis, SMJ Searle, N Harris, M Gibson, V Iyer, J Richter, C Wiel, L Bayraktaroglu, E Birney, MA Crosby, JS Kaminker, BB Matthews, SE Prochnik, CD Smith, JL Tupy, GM Rubin, S Misra, CJ Mungall, and ME Clamp, "Apollo: a sequence annotation editor," *Genome Biology*, vol. 3, no. 12, pp. research0082.1–0082.14, 2002, This article is part of a series of refereed research articles from Berkeley Drosophila Genome Project, FlyBase and colleagues, describing Release 3 of the Drosophila genome, which are freely available at http://genomebiology.com/drosophila/.
6. JW Kent, "Blat: the blast-like alignment tool.," *Genome Res*, vol. 12, pp. 656–664, 2002.
7. Kushal Chakrabarti and Lior Pachter, "Visualization of Multiple Genome Annotations and Alignments With the K-BROWSER," *Genome Research*, vol. 14, no. 4, pp. 716–720, 2004.

8. Nameeta Shah, Olivier Couronne, Len A. Pennacchio, Michael Brudno, Serafim Batzoglou, E. Wes Bethel, Edward M. Rubin, Bernd Hamann, and Inna Dubchak, "Phylo-VISTA: interactive visualization of multiple DNA sequence alignments," *Bioinformatics*, vol. 20, no. 5, pp. 636–643, 2004.
9. Jeffrey Heer, Stuart K. Card, and James A. Landay, "prefuse: a toolkit for interactive information visualization," in *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, New York, NY, USA, 2005, pp. 421–430, ACM.
10. Gregg A. Helt, Suzanna Lewis, Ann E. Loraine, and Gerald M. Rubin, "BioViews: Java-Based Tools for Genomic Data visualization," *Genome Research*, vol. 8, no. 3, pp. 291–305, 1998.
11. Lincoln D. Stein, Christopher Mungall, ShengQiang Shu, Michael Caudy, Marco Mangone, Allen Day, Elizabeth Nickerson, Jason E. Stajich, Todd W. Harris, Adrian Arva, and Suzanna Lewis, "The Generic Genome Browser: A Building Block for a Model Organism System Database," *Genome Research*, vol. 12, no. 10, pp. 1599–1610, 2002.
12. Feng Lu, Ji Zhang, and Yanhong Zhou, "A computational framework and browser for supporting automatic genome annotation," in *GCCW '06: Proceedings of the Fifth International Conference on Grid and Cooperative Computing Workshops*, Washington, DC, USA, 2006, pp. 389–396, IEEE Computer Society.
13. Robert D. Finn, James W. Stalker, David K. Jackson, Eugene Kulesha, Jody Clements, and Roger Pettett, "Proserver," *Bioinformatics*, vol. 23, no. 12, pp. 1568–1570, 2007.