전 유전체 단계적 가시화를 위한 새로운 컴포넌트 소프트웨어

정우근*, 조치영+, 조환규*
*부산대학교 컴퓨터공학과, +부산외국어대학교 비즈니스 IT학부
e-mail: cycho@pufs.ac.kr, {wkchung,hgcho}@pusan.ac.kr

A New Component Software for Hierarchical Visualization for Whole Genomes

Woo-Keun Chung*, Chi-Young Cho+, Hwan-Gue Cho*
*Dept of Computer Science, Pusan University
+Dept of Business IT, Pusan University Foreign School

요 약

게놈 데이터의 지속적인 증가로 인해 생물정보학에서 유전체 정보를 체계적으로 저장하고 열람하는 효과적이고 효율적인 시스템을 확립하는 것은 중요한 일이다. 잘 알려진 게놈 정보의 계층적 구조처럼 우리는 게놈의 내부 구조를 연구하기 위한 우수한 툴도 필요하다. 게놈 연구에 있어서 한 가지 문제는 유전체 정보는 너무 거대해서 표준적인 정보 처리를 이용하는 간단한 툴로는 작업하기 어려운 점이다. 예를 들어 특정 게놈 데이터 크기는 100메가 바이트를 넘는다. 추가적으로 유전자, promoters, retro-elements(HERV), operons, exon-introns와 같은 많은 게놈 요소들이 있다. 전통적으로 생물학자들은 게놈 연구를 위해 툴을 아무거나 사용하지 않고, 보통 그들의 연구에 좋은 툴을 채택하려 노력했다. 게놈 연구에서 기본적이고 주시할만한 단계는 다른 종과 유전체 요소를 비교하기 위해 위치를 인식할 수 있도록 하나의 화면에 모든 게놈 데이터를 시각화하는 것이다. 생물학자에게 툴의 개발은 많은 시간이 걸리고 시행착오를 겪기 쉬운 일이다. 이 논문에서 우리는 전체 게놈 중 어떤 게놈 요소를 시각화하는 컴포넌트 웨어의 셋을 제안한다. 그리하여 실험을 목적으로 생물학자를 만나서 우리의 셋을 이용하여 컴포넌트를 조합하여 소프트웨어를 만드는 것은 비교적 간단한 작업이다. 이 실험에서 우리는 HERV와 연동되는 게놈 요소를 보여주는 툴을 어떻게 우리의 컴포넌트 웨어를 간단히 조합하여 구축하는지를 보여주겠다.

1. 서론

인간 게놈 프로젝트의 성공은 지속적인 성장된 게놈 정보를 입증한다. 그것들은 두 가지로 나눌 수가 있다. 첫 번째 단계는 유전자과 같은 유전체 요소의 기본적인 물리 정보이다. 게놈 정보의 또 다른 단계는 유전자 또는 증가(감소)와 같은 게놈 요소의 사이에서 서로 연결되는 정보, 유전자들 끼리 연결되어 있는 구조를 표현한다. 본 논문에서는 첫 번째 단계의 문제점에 대하여 초점을 둔다. 좋은 시각화는 지놈에 숨겨진 흥미로운 특성을 밝혀내기 위해서 중요하다. 잘 알려진 도구 그리고 방법들은 일반 적인 데이터를 처리하기 위해 개발되었으나 생물학적 연구의 유전자 특성을 가시화 하기에는 다소 어려움이 있다. 그것들은 일반적인 데이터 그리고 유전체학 데이터의 가시화하기 위한 두 가지 다른 것들이 있다. 첫 번째, 한 모니터

표 1 : Genome Browser의 예

Browser	특징	참고
K-Browser	비교 분석, 세부 정보 시각화	[1]
Prefuse	정보에 대한 동적 처리 가능	[2]
VisGenome	비교 분석, 사용자 편의 GUI	[3]
Appolo	annotation 가시화	[4]
GATA	플랫폼 독립적, 비교분석 가능	[5]
ECRBrowser	웹 기반, genome alingment	[6]
Artemis	annotation 가시화, 정보 분석	[7]
Phylo	human, mouse alignment 제공	[8]
GGB	DB 와 web 기반 브라우저 제공	[9]

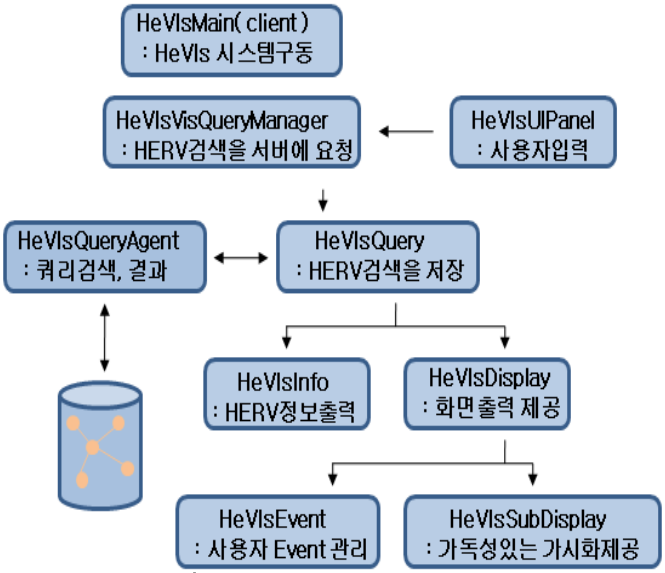
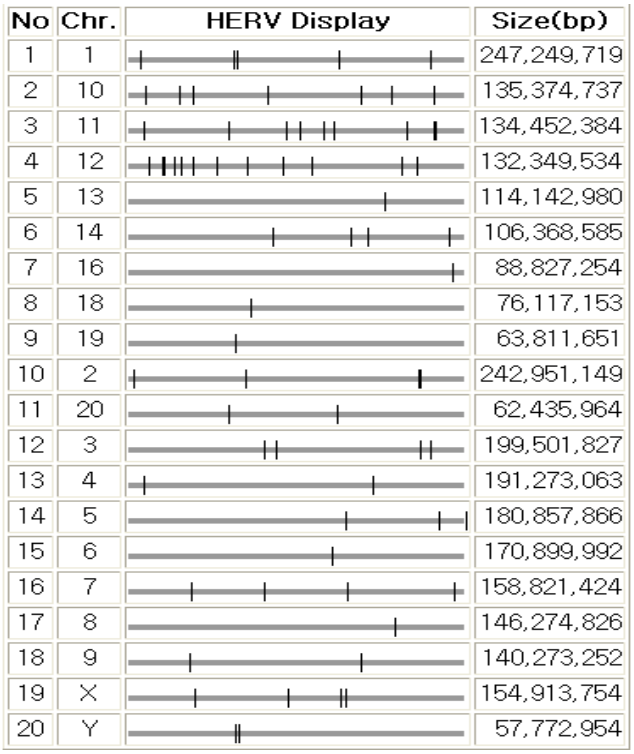


그림 1 : 본 시스템의 전체 구조도

에 유전체학 데이터를 표현하기에는 너무나 방대한 데이터이다. 그러므로 우리는 이 문제에 해결하기 위한 도구를 개발하고자 한다. 표 1 은 현재 제공 되고 있는 브라우저의 예이다. 이 논문에서는 우리는 효율적이고 효과적이며 신뢰성 있는 유전체학 시각화 시스템의 프로토 타입을 구축하기 위한 컴포넌트 웨어의 형태를 갖춘 시스템을 제안한다.

2.Component Ware 시스템 구성



No	Chr.	HERV Display	Size(bp)
1	1		247,249,719
2	10		135,374,737
3	11		134,452,384
4	12		132,349,534
5	13		114,142,980
6	14		106,368,585
7	16		88,827,254
8	18		76,117,153
9	19		63,811,651
10	2		242,951,149
11	20		62,435,964
12	3		199,501,827
13	4		191,273,063
14	5		180,857,866
15	6		170,899,992
16	7		158,821,424
17	8		146,274,826
18	9		140,273,252
19	X		154,913,754
20	Y		57,772,954

그림 2 : 본 논문에서 제안한 Component Ware의 클래스 구조도 및 데이터/제어 흐름을 나타내고 있다.

본 단락에서는 Component Ware의 주요 클래스에 대해서 설명하겠다. 그림 2는 본 시스템에서 제안한 컴포넌트 웨어의 초기 구성도. HERV Display가 바로 본 논문에서 제안한 가독성 있는 시각화를 제공한 부분이다. HERV Display에 존재하는 검은 실선들이 큰 데이터에 존재하는 특정 값들이다. 그림 2 에 보이는 Chr은 크로모좀 값를 나타내고 있으며 HERV Display는 각 크로모좀에 존재하는 HERV 들의 존재를 나타내고 있다. Size(bp)는 해당하는 크로모좀의 사이즈를 나타내고 있다. 현재 HERV 값은 Human에 HERV-Fc2값을 나타내고 있다. 본 시스템을 이용하여 가독성 있는 가시화 도구를 개발할 수 있게 되었다. Component Ware 시스템의 주요 클래스는 크게 클라이언트 측과 서버 측으로 구분된다. 클라이언트 측은 사용자에게 가독성 있는 가시화를 제공하기 위하여 사용자의 편의를 도모하는 인터페이스 환경을 제공한다. 클라이언트측은 사용자에게 우리가 한 눈에 볼수 없는 즉, 가독성이 떨어지는 방대한 양의 데이터를 가독성 있게 가시화하는 기능을 제공한다. 또한 사용자의 선택에 의한 선택적인 쿼리 또한 가능하다. 서버측은 크게 두 부분을 나뉜다. 쿼리 결과값 처리 및 DBMS와의 통신을 담당하는 부분으로 나뉜다. 쿼리 결과값 처리는 사용자의 선택에 따른 쿼리문 저장 및 쿼리문 저장 그리고 결과값 저장 및 결과값 전송하는 부분으로 나뉜다. DBMS와 통신을 담당하는 부분은 해당 서버측 과의 연동 및 접속으로 나뉜다.

3. Component Ware의 주요 클래스

3-1. Query Manager 클래스

사용자의 선택에 따른 쿼리 명령에 따라 서버의 Query Agent 클래스와 연동하여 검색결과를 전송한 후, 결과 값을 넘겨 받아 Query Result에게 이 결과를 전송하는 역할을 수행한다. 이 클래스의 주요기능은 서버상태 확인과 서버와 TCP/IP 접속이다. 사용자의 선택에 따른 쿼리는 해당 영역에 존재하는 특정한 구간들이 가지는 위치 정보 값들의 갯수 및 시작, 끝 정보들을 가져올 수 있다.

3-2. Query Result 클래스

Query Manager에 실행에 의해 서버의 Query Agent에서 보내오는 결과는 Query Result에 저장된다. 이전 단락에서 보였던 Query Manager의 실행에 의하여 사용자의 선택에 따른 쿼리 결과를 Query Agent로 전송하고 Query Agent에서는 그에 따른 결과 값은 Query Result로 저장된다. Query Agent에서 전송되어 지는 결과 값으로

표 2 : 본 시스템에서 제안한 컴포넌트 웨어에 적용될 데이터를 가지고있는 테이블의 구조도이다. 본 논문에서 실험 데이터 로 적용된 HERV 테이블의 구조를 나타내고 있다.

species	hervname	score	start	end	qsize	identity	chr	starand	hstart	hend	span
HU	HERV-Fc1	4338	1	4619	4629	97.2	7	-	63933061	63937625	4565
HU	HERV-Fc1	4306	9	4629	4629	96.7	2	-	83955044	83959645	4602
HU	HERV-Fc1	4171	1	4629	4629	95.5	Y	-	18912636	18917211	4576
HU	HERV-Fc1	4177	1	4487	4629	96.8	7	-	152737382	152741825	4444
HU	HERV-Fc1	4177	1	4629	4629	95.5	Y	+	18231346	18235920	4575
HU	HERV-Fc1	3773	1	4629	4629	96	11	+	8906579	5915735	9157
HU	HERV-Fc1	154	1364	1769	4629	73.8	X	+	96985190	96985517	328
HU	HERV-Fc1	35	572	698	4629	63.8	5	-	76216587	76216713	127
HU	HERV-Fc1	32	4551	4587	4629	94.6	6_cox	-	4466651	4466689	39
HU	HERV-Fc1	23	4195	4246	4629	80.8	1	+	36574320	36574371	52

는 가독성있는 가시화를 위한 위치 값들 그리고 해당 요소들의 특정 정보 값들이 함께 쿼리의 결과로 전송된다.

3-3. HeVIs Display 클래스

HeVIS Diaplay 클래스는 전 단락에서 선보였던 Query Result 클래스에 저장된 특정한 정보 값들을 출력하는 기능을 수행한다. 한 예로 그림 1에는 Quert Result에 저장되어 있는 특정한 정보 값들을 보여 주고 있는 결과물이다. 그림 1에 보이는 테이블 값들 중 HERV Display가 바로 가독성 있게 가시화를 제공한 것이다. 가로로 제공되는 바가 한 눈에 보기 힘든 값을 가지고 있는 전체의 구간이며 그 간에 검은 색 실선들이 바로 Query Result에 가지고 있는 위치 정보 값들이다.

3-4. HeVIsSubDisplay 클래스

전 단락에서 설명하였던 HeVIS Diaplay에서 선보이는 Query Result에 저장된 특정 정보 값들을 출력 시 가독성 있는 시각화를 적용 시켜줄 수 있도록 해주는 클래스이다. 본 단락에서 제시하는 클래스는 Query Result에 저장되어 있는 저장되어 있는 특정한 정보 값들 보여주고, 해당 영역에 존재하는 값들의 정보 값들을 보여주는 역할을 맡고 있다.

3-5. HeVIs Client 클래스

본 논문에서 제시한 시스템은 Java로 구성되어 있다. 클라이언트 부분은 Java Applet 시스템으로 구성되어 있다. Java의 플랫폼 독립적인 특성 때문에 다양한 시스템 환경을 위한 도구 제작으로써는 Java가 좋은 도구이다.

4. HERV 시각화를 위한 예

본 논문에서 제시한 컴포넌트 웨어의 실험을 위하여 HERV를 적용시켰다. HERV(Human Endogenous Retro Virus)는 인간 내생 레트로바이러스라는 것이다. 이러한 ERV는 provirus DNA의 형태로서 닭, 쥐 ,소, 말, 원숭이 등의 거의 모든 동물에 존재하고 있다. 인간의 게놈상에는 다수의 ERV가 존재하고 있으며 약 8%를 점유하고 있다. 인간 내생 레트로바이러스 (HERV)는 인간의 전 게놈상에 넓게 분포하고 있다. 인간의 유전체 데이터는 약 100메가를 넘는다. 이렇게 넓게 분포하는 인간의 전 유전체 데이터에서 특정 부분적으로 존재하는 내생 레트로 바이러스를 한 눈에 보는 것은 쉬운 일이 아니다. 본 논문에서 제시한 컴포넌트 웨어에 실험에 HERV를 적용시키기에 가장 적합하다. 앞에서 말한 바와 같이 HERV 데이터는 방대한 양을 가지고 있으며, 사용자가 가독성 있는 시각화를 유해서는 도구가 필요하다. 그림1 은 HERV 존재하는 하나의 데이터를 가지고 가독성 있는 시각화를 제공한 것이다.

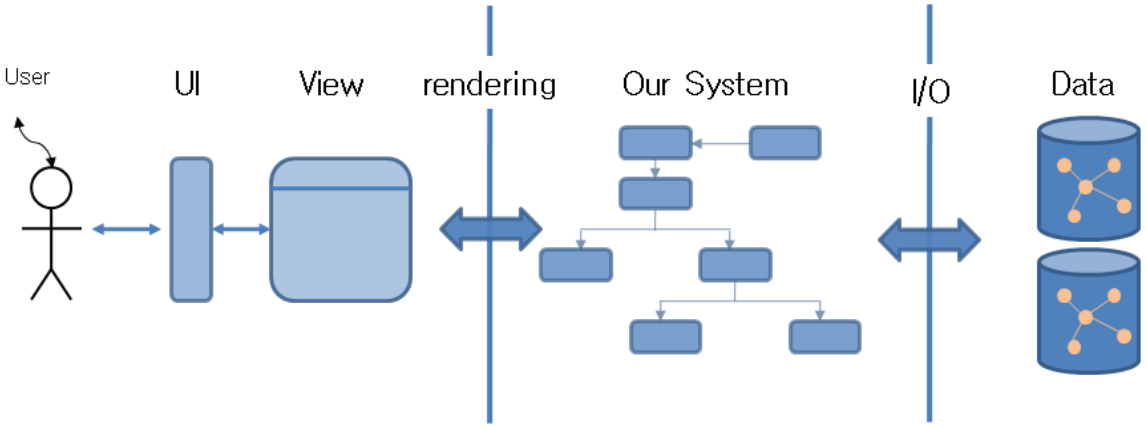


그림 3 : 사용자는 오직 컴포넌트 웨어의 유저 인터페이스 사용법만 알고 있으면, 우리의 시스템이 데이터베이스와 연동 사용자에게는 결과 값을 제공한다.

사용자의 이벤트를 담당해주는 클래스는 HeVIsEvent 클래스로서 사용자가 이행하는 이벤트를 효과적으로 처리하기 위하여, 특정 이벤트에 따르는 정보 값들을 저장하고 해당 하는 이벤트에 효과적으로 대응하기 위하여 이 정보 값 들을 해당 사용자가 원하는 것들로 처리하기 위하여 해당하는 클래스들에게 전송해준다. 또한 단계적 시각화를 이행도중 사용자가 특정 부분에 대하여 가독성 있는 시각화를 원하는 부분에 대한 정보 값들을 가지고 있는 클래스는 HeVIsInfo가 담당하고 있다.

표1 에 보이는 값들은 HERV 데이터들이다. 표1 에 보이는 여러 값들중 **hstart, hend** 값이 특정 지역의 시작, 끝 값이다. **species**는 종을 나타내고 있으며, **hervname**은 종에 존재하는 HERV 들의 이름을 나타내고 있다.

5.결론

본 논문에서는 HERV와 연동되는 유전체 요소를 보여주는 툴을 제작하였다. 본 논문에서 제시한 툴은 넓게 분포하고 있는 데이터 또는 가독성이 떨어지는 즉, 너무 방대한 자료를 담고 있는 자료에 대한 가시화를 제공하였다. 본 논문에서 제시하였던 컴포넌트 웨어에 실험 데이터로

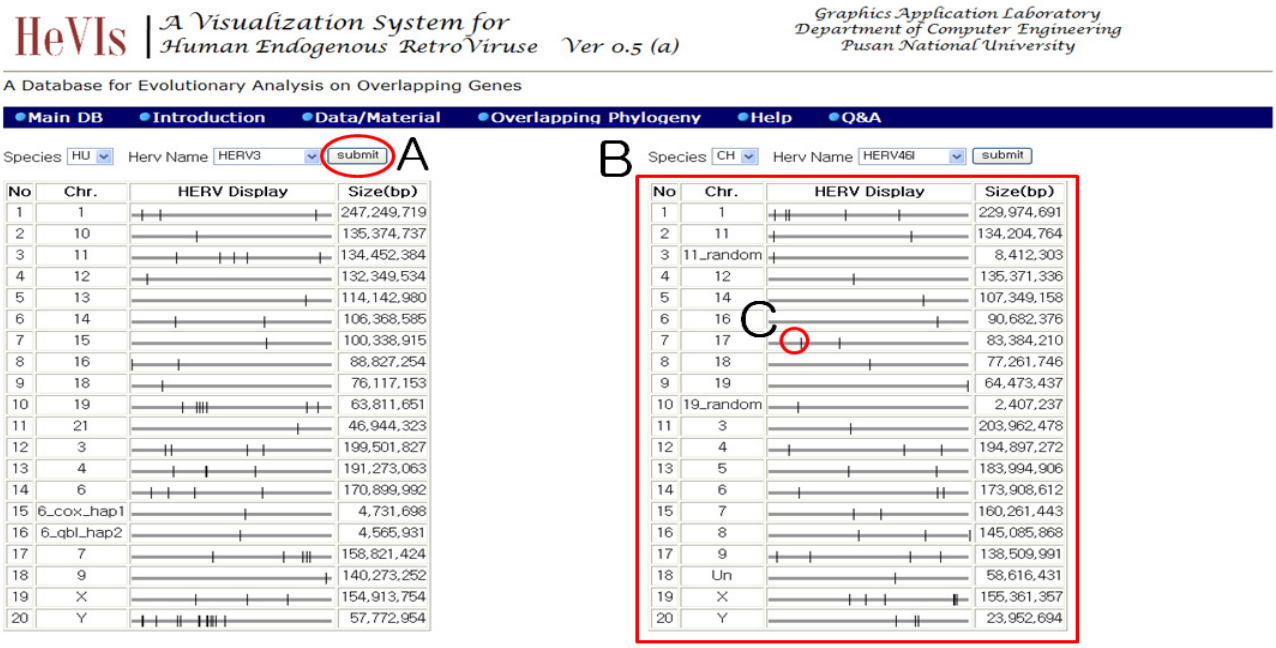


그림 4 본 논문에서 제안한 컴포넌트 웨어의 결과물이다. 가독성있는 가시화를 보고싶은 데이터 값을 선택하여(A) Submit을 통하여 데이터를 전송하면 (B) 와 같은 결과물이 제공된다. (C) 각 종에 존재하는 HERV 데이터 들이다.

HERV를 사용하였다. 실험 데이터로 사용된 HERV는 방대한 양의 자료를 가지고 있음에도 불구하고, 컴포넌트 웨어에서는 가독성 있는 가시화를 제공하였다. 또한 본 시스템은 Java 플랫폼 기반의 언어로 구성되어 있어서 다양한 시스템 환경에 좋은 환경을 제공하였다. 이 컴포넌트 웨어는 본 논문에서 실험 데이터를 HERV로 사용하였다. 하지만 앞서말한 실험데이터를 제외하고도 방대한 자료나 가독성이 떨어지는 가시화를 충분히 사용자의 목적에 맞게 쉽게 제작하고 좋은 도구로써 활용이 가능할 것이다.

참고문헌

[1] K. Chakrabarti and L. Pachter. Visualization of multiple genome annotations and alignments with the k-browser. *Genome Research*, 14:716 - 720, 2004.

[2] J. Heer, S. K. Card, and J. A. Landay. prefuse: a toolkit for interactive information visualization. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421 -430, New York, NY, USA, 2005. ACM.

[3] H. Jakubowka, E. Hunt, M. Chalmers, M. McBride, and A. F.Dominiczak. Visgenome:visualisation of single comparative genome representations. *Bioinformatics*, 23(19):2641-2642, 2007.

[4] S. Lewis, S. Searle, N. Harris, M. Gibson, V. Iyer, J. Richter, C.Wiel, L. Bayraktaroglu, E. Birney, M. Crosby, J. Kaminker, B. Matthews, S. Prochnik, C. Smith, J. Tupy, G. Rubin, S. Misra, C. Mungall, and M. Clamp. Apollo: a sequence annotation editor. *Genome Biology*, 3:0082.1-0082.14, 2002.

[5] D. A. Nix and M. B. Eisen. Gata : A graphic alignment tool for comparative sequence analysis. *Bioinformatics*, 6, 2005.

[6] I. Ovchrenko, M. A. Nobrega, G. G. Loots, and L. Stubbs. Ecr browser: a tool for visualizing and accessing data from comparisons of multiople vertebrate genomes. *Nucleic Acids Research*, 32, 2004.

[7] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. Rajandream, and B. Barrell. Artemis :sequence visualization and annotation .*Bioinfomatics*, 16(10):944-945, 2000.

[8] N. Shah, O. Couronne, L. A.Pennacchio, M. Brudno, S. Batzoglou, E. Bethel, E. M.Rubin, B. Hamann, , and I. Dubchak.Phylo-vista: interactive visualization of multiple dna sequence alignments. *Bio Informatics*, 20(5):636-643, 2004.

[9] L. D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, A. Arva, and S. Lewis. The generic genome browser: A building block for a model organism system database. *Genome Biology*, 12(1599-1610), 2002