韓國 컴퓨터情報學會 論文誌
第13卷 第2號, 2008. 3.

2008-13-2-1-12

대용량 자료 실시간 시각화를 위한 레벨 수준 표현 인터페이스 설계

이 도 훈*

Level Scale Interface Design for Real-Time Visualizing Large-Scale Data

DoHoon Lee*

요 약

자료를 시각적으로 표현하는 방법은 입력자료나 출력자료의 형태에 따라 많은 방법들이 제시되었다. 복잡하거나 방대한 자료 또는 정보를 시각적으로 표현하기 위해서 LOD와 같은 방법을 사용하고 특정부분을 지정하여 확대하는 방법을 주로 사용하고 있다. 본 논문에서는 생물정보와 같은 대용량 자료의 동적이고 실시간으로 배율을 표현할 수 있는 레벨수준 표현을 위한 인터페이스 설계 방법을 제안한다. 이는 기존의 LOD나 특정지역의 단순한 확대만을 위한 것이 아니라 동적으로 특정 영역을 축소 또는 확대해야 할 경우 실시간으로 표현할 수 있는 방법이다. 축소 또는 확대영역의 폭을 크게 했다가 어느 시점에서 매우 정교하게 조절할 수 있다. 제안된 방법으로 방대한 유전체 자료를 표현하는데 접목하여 구현하였고 매우 편리함을 보여주었다.

Abstract

Various visualizing methods have been proposed according to the input and output types. To show complex and large-scale raw data and information, LOD and special region scale method have been used for them. In this paper, I propose level scale interface for dynamic and interactive controlling large scale data such as bio-data. The method has not only advantage of LOD and special region scale but also dynamic and real-time processing. In addition, the method supports elaborate control from large scale to small one for visualization on a region in detail. Proposed method was adopted for genome relationship visualization tool and showed reasonable control method.

▸ Keyword : 인터페이스 설계(Interface Design), LOD, 대용량 자료(Large Scale data), Level-scale Interface

• 제1저자 : 이도훈
• 접수일 : 2008. 2. 25, 심사일 : 2008. 3. 3, 심사완료일 : 2008. 3. 14.
* 부산대학교 정보컴퓨터공학부 교수
※이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음

I. 서 론

컴퓨터가 일반화되면서 키보드, 스크린, 마우스 등은 인간과 컴퓨터 처리기관과의 훌륭한 인터페이스를 제공하고 있다. 나아가 소프트웨어도 단순한 처리 프로그램에서 벗어나 이제 인간공학적 측면이나 효율적 처리를 위해서 설계되고 이를 위한 전문적인 형태의 연구가 계속되고 있다. 이와 관련하여 인간이 보다 편안하게 일을 수행할 수 있을까 하는 연구는 일찍이 심리적 인식론에서 인터페이스 설계를 편리성방향을 제시하였고[1] 오늘날 그 형태도 다양하게 제공하고 있다[14]. 컴퓨터를 통해 표현해야할 자료의 종류에 따라 보다 다양한 형태의 표현 방법들이 제시되었다[3]. 전통적인 전산학 기반의 데이터의 크기는 주어진 해상도를 크게 넘어서지 않거나 상수(10배 이하의 상수)배 이하의 크기로 조절이 가능한 수준의 자료를 다루어 왔다. 전통적인 문제는 대부분 상용화 될 만큼 해결되었거나 학문적으로 잘 증명되고 구현되었다는 것이 보편적인 결론이다. 이들의 특징은 그 처리되는 입력의 형태가 잘 정리된 것, 혹은 특정 양식을 띄고 있는 것이 보편적인 형태이다. 이런 전통적인 문제가 해결되고 난 후의 문제는 그 동안 관심의 대상이 아닌 분야의 가공되지 않은 자료(row data)를 처리하고 그 자료로부터 정보를 얻어야 하는 문제로 옮겨가고 있다. 예를 들면, 위성에서 수시로 쏟아져 오는 기후와 해수 온도에 대한 정보나 기타 여태까지 고려하지 않았던 자료가 무한정으로 쌓이고 있다. 이런 자료들의 생산은 위성과 같은 자연적인 자료에만 국한된 것이 아니다. 기존에 존재하는 시스템에 의해 재생산되는 자료의 양은 일반적으로 측정할 수 있는 측도를 넘어서고 있다[3,15].

재생산된 자료의 양이 대규모로 쏟아져 나오는 분야 중에 하나가 생물정보학 분야이다[4,5]. 이는 공개된 여러 가지 프로그램에 의해 생산되는 자료는 박테리아 정도만 하더라도 유전체일 경우 기본적으로 백만 단위(bp)를 초과하게 된다. 백만단위의 자료를 표현하려면 약 1000 픽셀을 표현할 수 있는 일반적인 해상도에서 1000배 이상의 크기를 시각화해야하는 문제가 발생한다. 1000배 이상의 축소/확대 영역을 가지는 문제는 기존의 전산학 범위의 입력 자료 형태에서 볼 수 없던 자료형태이다. 해상도의 경우 일반적으로 사용되고 있는 간단한 문서의 확대가 크게 몇 백배 되는 경우가 있다. 이런 보편적인 자료는 그 형태가 매우 단순하여 LOD[2] 개념이 필요하지 않는 문서가 대부분이다. 본 논문에서 다루고자하는 자료의 형태는 단순한 문서 정보가 아니라 생물정보에서 다루는 다양한 형태의 정보를 제공하는 동적인 LOD 형태의 인터페이스를 요구하고 있다.

본 논문에서 제안하고자 하는 동적인 인터페이스는 유전체 염기서열처럼 10Mbp 정도의 길이를 가지는 자료를 다루는데 적합하다. 기존의 표현 가능한 2048 픽셀에 몇 배정도 큰 자료를 표현하는 것이 아니라 적어도 주어진 화면의 몇 백배에서 몇 천배 큰 자료를 표현해야 할 때, 이를 보다 효율적으로 조절할 수 있는 인터페이스 설계가 필요하다.

II. 대용량 정보처리 문제점

요즘 생물정보학 분야에서 가장 활발하게 연구되고 있는 분야가 정보의 시각화이다. 이를 위한 많은 연구결과가 발표되었고 시스템이 공개되어 있다. 이 프로그램들의 대부분 목적기반에 시각화 수행한다. 가장 쉽고 널리 알려진 NCBI Entrez[13]에서 다양한 프로그램들을 볼 수 있다. AceDB[6], Synteny- Vista[7], Apollo[8], Artemis[9], Ensemble[10], K-Browser[11], NCBI Mapviewer[12]는 대부분 LOD을 기반으로 하고 있다. 그리고 필요에 따라 새로운 창을 제공하는 방법으로 정보를 시각화하고 있다.

생물정보라고 모두 대용량은 아니다. 구조정보, 단백질 서열 정보 등은 상대적으로 유전체 전체를 다루는 정보가 아니기 때문에 상대적으로 적은 용량이다. 이에 반해 염기서열은 미생물이나 다양한 바이러스 등 비교적 적은 양이라고 하는 서열이라 할지라도 백만 bp 단위를 가진 크기가 많다. 이런 유전체 서열은 그 기능이나 역할이 잘 알려지고 상호작용도 잘 정의되어 알려진 정보도 있으나 대부분 그 정보가 알려져 있지 않다. 따라서 각 서열의 모든 위치의 정보를 하나하나 점검하고 이를 보관하여 처리해야 한다.

본 논문에서 언급하는 대용량 자료는 이와 같은 박테리아 DNA와 그에 관련된 정보이다. 이와 같이 모든 정보를 보관하여 처리한다는 의미는 보편적으로 백만단위(Mbp)를 다루면서 이를 시각화하도록 처리하는 데는 많은 시간이 소요됨은 물론 실시간 상호작용을 요하는 인터페이스 디자인에 있어서는 치명적인 장애요인이 된다. LOD방법에 의한 시각화도 처리량이 절대적으로 많아지면 처리속도에 심각한 장애를 초래한다. 또한 시각화하고자 하는 정보의 종류에 따라 동적인 변화를 실시간으로 처리하기 위해서는 새로운 LOD 방법이 필요하다. 따라서 한 픽셀에 한 정보를 표현하는 방법을 탈피하는 능동적이고 동적인 LOD(Level of Detail) 방법에 의한 인터페이스 디자인이 필요하다.

앞선 설명을 일반적으로 사용되는 잘 정의된 소프트웨어 입력 자료와 본 논문에서 언급하고자 하는 생물정보 자료와의

차이를 비교 설명하면 다음과 같다. 표1에서 언급했듯이 미리 정의된 형식에 의한 입력 자료들은 그 획일성이나 구조가 소프트웨어 인터페이스 설계에 많은 편의성을 제공한다. 반면에 생물학자나 의학자들의 연구에 도움을 주기위해 제공되는 생물정보학 관련 자료들은 소프트웨어 개발을 위한 인터페이스 설계를 위한 부가적인 작업을 요구한다. 또한 그 크기가 일반적인 소프트웨어 입력 자료의 크기와 비교가 힘들 만큼 차이가 난다. 이를 처리하여 시각적으로 정보를 제공하는 일은 기존의 시각화를 위해 처리하던 방식으로는 시간적 제약으로 인해 효율적으로 대응하지 못한다. 이를 위해 새로운 처리 방법이 필요하다.

표 1. 일반 입력자료와 생물정보 자료의 비교
Table 1. Comparison of bio-data and general software input data

항 목	일반입력자료	생물정보의 서열자료
형태	잘정의된 형식을 갖춘 형태 (Well defined input data)	연구자들이 읽기 쉽게 만든 문서 형태 (Natural format oriented)
중심	프로그램	사람(연구자)
크기	K - 100 K	100 K - 10 M
입력자료 의미	잘 정의됨	부분적으로 정의됨
입력자료 형식	프로그램에 맞게 설계됨	사용자가 읽기 쉽게 표현
입력자료 해석	대부분 아는 정보	대부분 의미 모르는 정보
목적	프로그램 수행	생물정보 해석

III. 레벨수준 배율 표현

2절에서 언급한 생물정보는 다음과 같은 성질을 가지고 있다. 처리하고자 하는 입력이 대용량이며 종류에 따라 그 양이 매우 다양하다. 또한 부과적인 정보를 가지고 있기 때문에 단순히 주어진 입력만으로 처리가 완료되는 경우가 드물다. 이를 모니터에 표현하고자 하나 그 제한된 크기 때문에 방대한 정보를 표현하기가 힘들다. 기존의 LOD와 같은 방법으로 처리하려면 그 처리 속도 때문에 효율적이지 못하다. 기존의 LOD 방식은 점진적 배율 조정으로 인해 많은 단계적 시간 소모가 발생하게 된다. 따라서 보다 동적인 배율 조정이 가능한 인터페이스 설계가 필요하다. 이와 같은 동적인 배율을 위한 방법중의 하나가 특정 영역을 지정하여 즉시 확대하는 방법이다. 이미지를 확대할 때 많이 사용하는 방법이다. 이 방법과 앞서 언급한 LOD방법 모두 장단점을 가지고 있다. 표2에서 정리한 것과 같이 LOD는 배율의 폭이 급격히 크거나 자료의 크기가 클때, 특정영역 확대기법은 축소를 위한 방법에서 한계를 가지고 있다. 표현되어야 할 시각적 정보의 동적인 표현을 위한 인터페이스는 다음과 같은 요소를 만족해야 한다. 빠른 실시간 상호 작용 기능, 전체 자료 보기 기능, 빠른 정보 표현 그리고 LOD에서 중요한 성질인 동적인 레벨 조정이 기능해야 한다.

표 2. LOD와 특정영역 확대기법의 비교
Table 2. Comparison of LOD and area scale method

특징	LOD	특정영역확대기법
좋은 점	점진적 축소/확대	동적이고 특정영역 확대기능 우수
한 계	Scale 폭이 크거나 동적인 확대/축소	축소에 한계
시 간	시간 소요	확대-빠름
응용분야	점진적인 확대/축소 자료 활용	필요한 부분 확대를 즉시 알고 싶은 자료

위와 같은 요소를 만족하는 인터페이스 설계를 위해 본 논문에서는 절차적인 LOD표현이 아니라 동적인 레벨 조절이 가능한 레벨수준 표현 인터페이스(Level-Scale Interface)를 제안한다. 축소비율 조정을 동적으로 표현하는 레벨 수준 인터페이스 설계를 만드는 방법은 다음과 같다. 최저 축소 비율 단위를 지정하여 십진수 자리수 형태로 표현한다. 1, 1/2, 1/3, ... , 1/9와 같이 분모가 일의 자리로 구성되어 바로 조절할 수 있다. 그 다음 단계로 최저 비율에 일정한 배율(10)을 곱하여 만들 수 있다. 예를 들어 1의 자리에 10를 곱하여 10의 자리를 그 배율로 삼는다. 따라서 1/10, 1/20, 1/30, ... , 1/90의 배율을 조정할 수 있는 조절바를 만든다. 이와 같은 방법을 사용하면 다양한 축소비율을 만들 수 있다.

이와 같이 자료의 크기에 따른 동적인 레벨 수준 인터페이스 방법의 특징은 다음과 같이 요약할 수 있다. 절차적 ZOOM in/out 탈피와 동적인 ZOOM In/out 를 용이하게 할 수 있다. 각 레벨은 다음과 같이 표현할 수 있다. 각 레벨별로 1/(n*1000), 1/(n*100), 1/(n*10), 1/n 이고 레벨별 단계는 $1 \leq n \leq 9$ 이다.

이를 보다 단계적으로 표시한 것이 그림1이다. 그림1의 왼쪽은 1/9100 배율을 표현하기 위한 트리의 모양을 표현하

고 있다. 1/1000 배율을 9에 설정하고 1/100 배율을 1에 설정하면 원하는 배율을 얻을 수 있다. 이와같은 조절 인터페이스는 동적인 배율 조정을 지원하게 되고 LOD와 같은 점진적 배율 조정에 의해 낭비되는 시간을 없앨 수 있다.

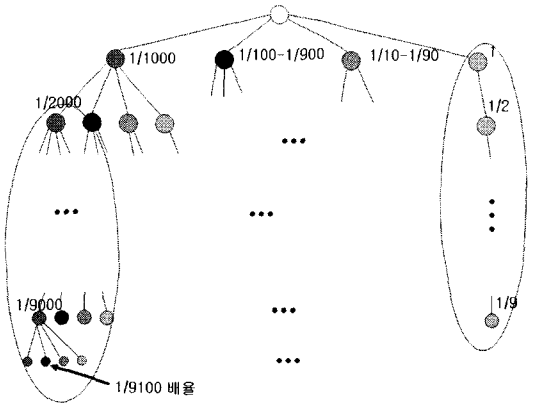


그림 1. 레벨수준 배율 표현을 위한 트리구조. 1/9100배율을 위해서는 1/1000레벨과 1/100레벨 배율을 한번 사용하면 된다(왼쪽). 오른쪽 타원내 레벨은 1/9배까지의 배율을 표현하기 위한 단계이다.
Fig 1. Tree for representation of level-scale.

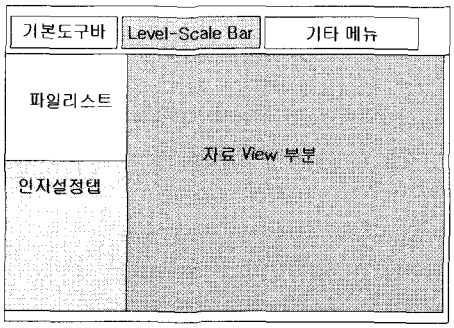
표현할 수 있는 윈도우의 크기를 W라 하고 입력크기를 N이라 하자. 1픽셀에 표현되어야 할 정보의 크기를 W_s라 하면 $W_s = \lceil \frac{N}{W} \rceil$가 된다. 입력된 정보를 한 윈도우에 전부 표현하기 위한 축소배율은 $1/W_s$ 이고 조절바를 설계하기 위해서 몇개의 바가 필요한지를 계산해야 한다. 그림1 트리의 루트노드의 차수를 의미하고 b라 표현한다. 1픽셀에 저장할 크기 W_s는 다음과 같이 표현할 수 있다.

$$W_x = c * 10^k$$

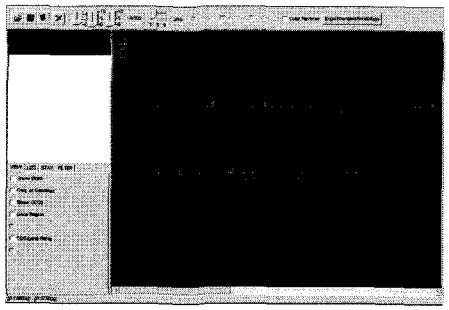
조절바의 개수 b는 다음과 같다.

$$b = \begin{cases} k+2, c \geq 5 \\ k+1, c < 5 \end{cases}$$

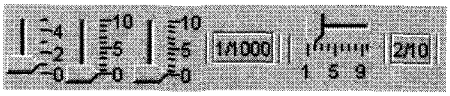
예를 들어 1M 정보를 1000 픽셀의 윈도에 표현하려고 한다고 하자. 먼저 주어진 1,000,000 /1,000 = 1,000이다. 즉 W_s가 1,000 이다. 이는 1×10^3 이므로 조절바를 4개를 두어야 한다.



(a) 전체시스템의 인터페이스 설계



(b) 구현된 인터페이스

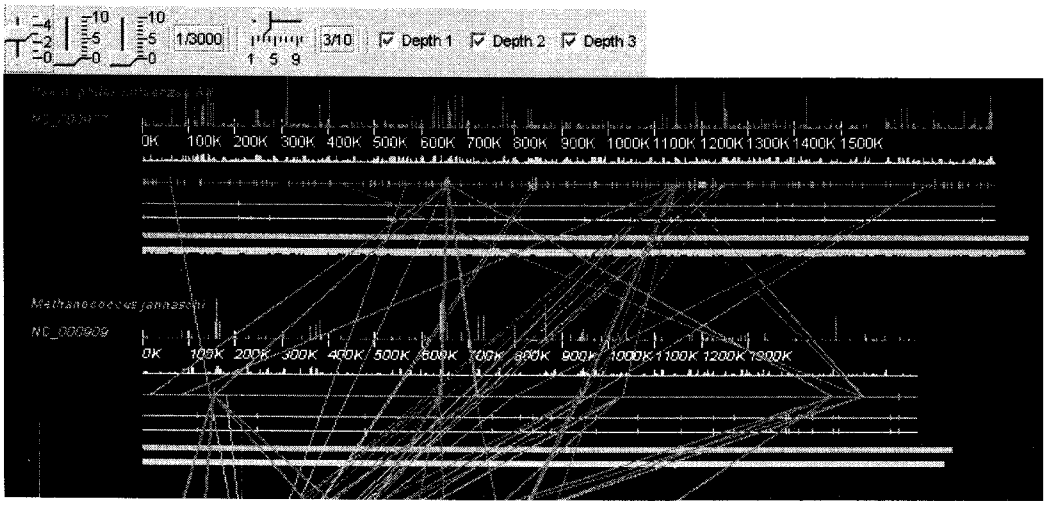


(c) x축 정보를 위한 레벨 수준 표현을 위한 인터페이스(확대)

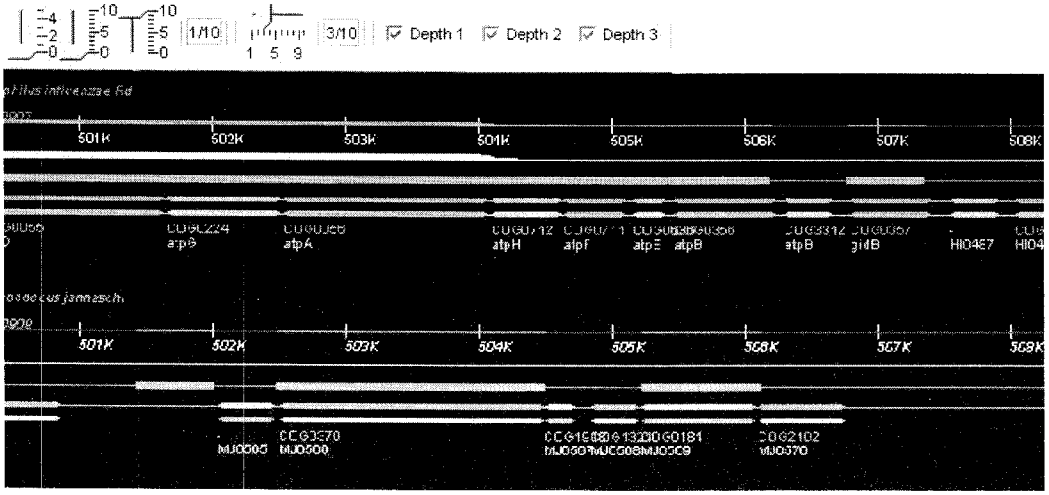
그림 2 시스템 인터페이스
Fig 2. System interface

IV. 실험 및 평가

본 논문에서 제안한 동적인 레벨 수준 인터페이스를 그 크기가 매우 다양한 유전체 데이터를 다루는 문제에 적용하였다. 기본적으로 그 크기가 10K에서 수백M 정도의 길이를 가진다. 따라서 수백M되는 길이를 1000 픽셀에 표현해야 한다. 생물관련 자료중 서열 정보는 선형적으로 표현이 가능하기 때문에 고배율을 적용할 때 1000 픽셀에 정보가 있으면 한 픽셀에 표현하는 방식으로 중간 처리 속도를 개선하였다. 그럼에도 불구하고 이 정보를 실시간으로 표현하는데는 궁극적으로 얼마나 많은 유전체 정보를 처리하느냐에 달려있었다. 따라서 점진적 확대/축소 방법인 LOD로는 효율적인 실시간 상호작용할 수 있는 기능이 매우 불편하였다.



(a) 전체를 보기 위한 고배율 축소(1/3000)



(b) 보다 상세한 정보를 보기 위한 저배율축소(1/10)

그림 3. 배율에 따른 조정. (a) 전체를 볼수 있는 배율과 (b) 내용을 보다 상세하게 볼수 있는 배율 보기

Fig 3. Scale control.

제안한 방법을 구현하기 위해 사용된 환경은 다음과 같다. Pentium 4 CPU 3.00 GHz에 Windows XP 운영체제상에서 Java언어로 구현하였다. 사용된 자료는 작게는 500K bp에서 최대 5M bp 크기를 가진 박테리아 유전체이고 박테리아 유전체간의 관계를 보여주는 프로그램을 구현하였다. 실험에 사용한 구체적인 생물자료는 다음과 같다. *Haemophilus influenzae Rd*(19.M), *Mycoplasma genitalium*(580K), *Methanococcus jannaschii*(1.7M), *Mycoplasma pneumoniae*(800K), *Escherichia coli K12*(4.6M), *Helicobacter pylori*(1.6M), *Mycobacterium tuberculosis H37Rv*(4.4M), *Bacillus subtilis*(4.2M), *Bacillus halodurans*(4.2M), *Staphylococcus aureus subsp. aureus N315*(2.8)의 유전자 서열을 이용하여 유전체간 관계를 표현하는 데 제안한 확대/축소 인터페이스를 접목하였다.

본 논문에서 제안한 인터페이스는 그림 1(a)에서 언급한 레벨수준 조절을 할 수 있는 기능이다. 이 기능은 고배율 혹은 저배율 확대/축소를 매우 동적으로 지원함으로써 LOD나

특정부분 확대 기능의 약점을 보완한 인터페이스를 구현하였다. 그림2(a)는 전체 시스템의 설계와 이를 바탕으로 한 구현된 화면을 (b)에서 보이고 있다. (c)는 실제로 본 논문에서 제안한 레벨수준 조절 인터페이스를 확대하여 보이고 있다. 확대영역을 직접 지정하여 제어하는 특정영역 확대기법의 제어력은 미비하나 다른 기능들은 현재 많이 사용하고 있는 방법에 비해 매우 편리함을 보이고 있다. 다만 이를 위한 객관적인 비교 방법이 알려진 바가 없는 관계로 점성적 평가로 인해 객관성이 다소 결여된다고 볼 수 있다.

표 3. 제안된 방법의 기능별 비교표
Table 3. Functional comparison of proposed method

기능	LOD	특정영역 확대기법	Level Scale (제안한 방법)
축소기능	○	○	○
확대기능	○	X	○
동적인 확대	X	○	○
동적인 축소	X	X	○
실시간 적응력	X	○(확대) X(축소)	○
확대영역 제어력	X	○	X

그림2 (c)와 같이 레벨별 배율 표현 인터페이스를 적용함으로써 보다 적은 횟수의 조정으로 원하는 정보를 보거나 추론하는데 효과적이였다. 그림3 (a) 에서는 전체적인 특징을 보기 위해 고배율 축소를 통해 유전체간 관계를 알 수 있도록 보여주는 것으로 레벨수준 인터페이스부분과 그 결과에 따른 결과를 편집하여 보여주고 있다. 이때 보여주는 관계는 전체적인 윤곽만 필요하므로 세세한 정보보다는 관계된 영역만을 보여주도록 한다. (b)는 (a)를 통해 필요한 유전자를 탐색하고자 할 때나 보존정도가 높은 특징부분을 찾고자 할 때 사용한다. 이는 보다 세밀한 정보를 알기 위해 저배율 축소를 통해 탐색하는 장면을 보이고 있다. 그림3(b)에서와 같이 매우 구체적인 정보를 볼 수 있도록 지원한다. 표3은 확대/축소를 할수 있는 기법과 그에 대응되는 요소에 대한 기능을 비교한 결과이다. 제안한 인터페이스가 방대한 자료의 정보를 실시간 상화작용 기능을 다른 기법에 비해 상대적으로 용이하게 지원함을 볼 수 있다.

V. 결론

정보를 시각적으로 보여주는 문제는 복잡하거나 원시자료를 다루는 분야일수록 그 중요도는 점점 높아지고 있다. 자료가 정형화되어 있거나 특정한 형태로 통일되어 있을 때는 굳이 시각화할 필요성이 없었기 때문이다. 본 논문에서는 대용량 정보를 시각화하는데 필요한 실시간 축소/확대를 위한 레벨수준 표현 인터페이스 방법을 제시하였다. 제시한 방법으로 실제로 유전체 정보를 다루는 문제에 적용하여 기존의 LOD나 특정영역 확대기법 보다 편리하게 접목됨을 구현을 통해 보였다.

참고문헌

[1] L.J. Bannon, "A Pilgrim's process: From cognitive science to cooperative design", AI and Society, Vol. 4, no. 4, Fall Issue, pp. 259-275, 1990.

[2] H. Clark, "Hierarchical Geometric Models for Visible Surface Algorithms," Communication of the ACM, Vol. 19, no. 10, pp. 547-554, 1976.

[3] F. Murta호, T. Taskaya, P. Contreras, J. Mothe, and K. Englmeier, "Interative Visaul User Interfaces: A Survey," AI Review Vol. 19, pp. 263-283, 2003.

[4] Y. Lue, P. Guo, S. Hasegawa, and M. Sato, "An Interactive Molecular Visaulizatio System for Education in Immersive Multi-projection Virtual Environment," Proc. of the 3rd Int. Conf. on Image and Graphics, 2004.

[5] P. McDermott, J. Sinnott, D. Thorne, and S. Pettifer, "An Architecture for Visualization and Interactive Analysis of Proteins," Proc. of the 4th Int. Conf. on Coordinated & Multiple Views in Exploratory Visualization, 2006.

[6] R. Durbin, and Mieg, J.T.A.C. elegans Database, Doc. code and data available from anonymous FTP servers at lirmm.lirmm.fr, (1991-)

[7] E. Hunt, et. al., "The Visual Language of synteny," OMICS, 8(4), pp. 289-305, 2004.

[8] S.E. Lewis, "Apollo: a sequence annotation editor," Genome Biology, 2002.
[9] K. Rutherford, et. al., "Artemis: sequence visualization and annotation," Bioinformatics, Vol. 16, No.10, 2000.
[10] Ensemble database, http://www.ensembl.org.
[11] K. Chakrabarti and L. Pachter, "Visualization of multiple genome annotations and alignments with the K-BROWSER," Genome Research, 2004.
[12] Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/
[13] NCBI Entrez, http://www.ncbi.nih.gov/Entrez/
[14] 강선경, 정성태, 이상설, "EOG와 마커인식을 이용한 착용형 사용자 인터페이스," 한국컴퓨터정보학회논문지, Vol. 11, No. 6, 2006.
[15] 정규장, "웹 기반 공간데이터 공통 컴포넌트 설계 기법," 한국컴퓨터정보학회논문지, Vol. 9, No. 1, 2004.

저 자 소 개

이 도 훈

1986년 부산대학교 계산통계학과
1992년 부산대학교 전자계산학과(석사)
1997년 부산대학교 전자계산학과(박사)
1995년~2005년 밀양대학교 컴퓨터공학부 교수
2004년~2006년 미국Indiana Univ. 객원교수
2006년~현재 부산대학교 정보컴퓨터공학부 교수
관심분야: 물리기반 모델링, 비주얼컴퓨팅, 생물정보학