진화적인 거리 관계를 반영한 계통분석 분류 단위의 3차원 공간 표현

Three-Dimensional Space Embedding of phylogenetic taxonomic units Preserving Evolutionary Distance Relationship

[1]이승희, [1]황경순, [1]이혜리, [1]이건명, [2]이찬희

[1] 충북대학교 전기전자컴퓨터공학부
[2] 충북대학교 미생물학과
E-mail : shlee@aicore.cbnu.ac.kr

요 약

계통발생학적 분석은 바이오인포매틱스에서 분류 단위의 진화적인 관계를 추정하는데 있어서 중요한 문제 중 하나이다. 계통발생학적 분석에서 사용되는 계통수는 분류 단위들 사이의 진화 거리에 대한 정보를 표현하기에는 부족하다. 이 논문은 진화거리 표현을 위해 3차원 공간으로 분류 단위가 표현되기 위한 방법을 제안하고, 2차원 트리 뷰와 함께 3차원 공간에서 분류 단위들을 시각화하기 위한 도구를 개발하였다.

Key Words : 시각화, 다차원 척도법, 계통발생학적 분석, 바이오인포매틱스

1. 서 론

바이오인포매틱스에서 계통발생학적 분석은 유기체, 유전자, 종들과 같은 분류 단위(OTU ; Operational Taxonomic Units)의 진화적인 관계를 추정하는데 있어서 중요한 문제 중 하나이다.[4] 계통수(phylogenetic tree)는 분류 단위 사이의 진화적인 관계를 설명하기 위해 사용된 시각적인 표현이다. 이것은 생물학에서 필수적인 도구 중 하나로서 진화를 이해하고, 유전자 발현을 예측하고, 바이러스 변종의 기원을 결정하는 것과 같이 다양한 방법들로 사용되어왔다.

계통수를 추정하기 위해 다양한 방법들이 개발되어 왔는데, 그것들은 거리 기반 방법들과 특징 기반 방법들로 분류된다. 전형적인 거리 기반 방법으로는 UPGMA(Unweighted Pair Group Method with Arithmetic mean), NJ(neighbor joining), FM(Fitch-Margolish) 등이 있다. [3]

UPGMA 방법은 가장 밀접하게 관련된 두 OTU를 선택하고, 그것들의 공통 조상을 표현하기 위한 노드를 삽입한다. 그런 다음 두 분류 단위 모두 포함하는 집합을 선택된 두 OTU들로 대체하고, 그 쌍으로부터 다른 쌍으로 거리를 수정한다. 앞에서 설명한 과정을 모두 마치면 처음 과정부터 반복한다. NJ 방법은 트리의 다른 가지들 중에서 진화의 다른 정도를 수정하기 위해 디자인된 UPGMA 방법의 변형이다. FM 방법은 OTU들 사이의 경로 길이가 그것들 사이의 거리가 되는 방법으로서 점진적으로 계통수를 구성한다.

일반적으로 사용된 문자 기반 방법들로는, MP(Maximum parsimony) 방법, ML(Maximum likelihood) 방법 그리고 베이지안(Bayesian) 방법 등이 있다. [3] MP 방법은 관측된 데이터 패턴을 설명하기 위해 필요한 가장 적은 문자 변화를 요구하는 트리 위상을 찾는다. ML 방법은 데이터를 설명할 수 있는 가장 적당한 트리 하나를 찾는다. 베이지안 방법은 데이터와 적절한 진화 모델을 설명하기 위한 가장 좋은 트리들의 집합을 찾는다.

2차원 공간에 그려진 계통수에는 그것의 위상과 몇몇 OTU들 사이의 경로 길이가 진화 관계에 대한 정보를 나타낸다. 비록 하나의 OTU가 다른 OTU와 인접해 있는 것이 항상 진화 역사에서도 인접해 있다는 것을 의미하지 않지만, 일부 분석가들은 인접한 OTU들이 밀접하게 관련된 것으로 해석하게 된다. n개의 OTU들을 갖는 이진 계통수와 동일한 의미를 갖는 표현 가능한 트리는 2^n개 이다. 이것은 하나의 OTU가 다른 OTU와 인접해 있지 않더라도 진화 역사에서 인접해 있을 수 있다는 것을 의미한다. 그러므로, 앞에서 설명한 내용은 가까운 OTU들을 계통수 뷰에 어떻게 보여줄 것인가에 대해서 다른 관점을 제공하기 위한 도움이 될 것이다.

이러한 관찰로부터 OTU들 사이의 거리 제약들이 거의 일치하는 방법으로 3차원 공간에 OTU들을 시각화하기 위한 방법을 개발하였다. 제안된 방법은 거리 기반 계통수 추정 방법에 대해서 부가적인 뷰를 제공하기 위해 개발되었다.

이 논문은 다음과 같이 구성된다. 2장에서는 계통수 시각화에 대한 관련 연구를 간단히 살펴본다.

이 논문은 2006년도 교육인적자원부 지방중심연구대학 육성사업의 지원에 의하여 연구되었음.

3장에서는 OTU들을 위해서 제안된 3차원 시각화 방법과 이 방법을 지원하기 위해 개발된 도구를 소개한다. 4장에서는 어떻게 도구가 개발되었는지 설명하고, 5장에서는 결론을 맺는다.

2. 관련 연구

계통수의 시각화는 바이오인포매틱스에서 중요한 문제이다. 계통수를 시각화하는 문제의 다섯 가지 주요 범주에는 배치, 레이블 붙이기과 주석 달기, 탐색, 트리 비교, 조작 그리고 편집 등이 있다. [5] 배치를 연결하는 것은 계통수들을 표현하기 위한 것으로 다양한 트리 토폴리지들이 있다, 그리고 그것들로는 계통도(phylogram), 방사상태의(radial) 구조도, 그리고 기울어진 분기도(slanted cladogram)가 일반적으로 가장 많이 사용된다. 레이블링(labeling)은 각각의 레이블이 읽기 쉽고, 그것의 그래프적인 객체의 관계가 모호하지 않고 어떤 것도 관련 정보를 가리지 않도록 모든 레이블의 위치를 배치하는 문제이다. [5] 탐색은 세 단계의 처리 과정인 전체 보기, 줌과 필터, 그리고 detail-on-demand를 제공함으로서 사용자가 큰 트리를 탐색 가능하게 한다. 트리 비교에 관한 것은 비교된 계통수를 효과적으로 시각화 하는데 매우 도움이 되며 부가적인 문제이다. 또한 계통수의 조작과 편집은 시각화 도구들에 대한 연구 분야로서 [illegible]한 문제가 될 수 있다.

[illegible]수에 대한 시각화를 위해서 다양한 애플리케이션들이 개발되어 왔다. 몇몇 애플리케이션들(PAUL, PHYLIP 등)은 주어진 데이터로부터 계통수를 만들 수 있고 그것들을 그릴 수 있도록 제공한다. Phylodendron는 사용자가 트리와 상호작용이 가능하므로 사용자들이 노드를 선택할 수 있고, 선택된 노드와 관련된 데이터를 볼 수 있다. TreeWiz[12]은 서브트리를 탐색하는 수단으로서 다중 윈도우를 사용하는 인터페이스를 제공한다. TreeJuxtapose[13]은 많은 노드에 대해서 구조적 비교를 지원하기 위한 시각화 기능들을 제공한다.

그 외에 2차원 시각화 애플리케이션은 계통수에 대해서 보다 효과적인 시각 정보를 제공하기 위해 3차원 표현 방법을 사용하기 위한 여러 가지 접근을 시도했다. Arbor 3D[6]은 3차원 공간에 계통수를 표현하고, 가상현실 장비를 사용하여 트리와 사용자가 상호작용이 가능하다. Hughes 등[10]은 매우 큰 계통수를 시각화하기 위해서 3차원 곡선 공간의 사용을 위한 애플리케이션을 개발하였다.

제안된 방법은 계통수 자신을 시각화하기 보다는 계통수에 대한 부가적인 뷰를 제공하는데 중점을 두었다. 제안된 방법으로 만들어진 뷰 덕분에 생물학자들은 계통수를 잘못 이해하지 않도록 약간의 지원을 받게 될 것이다.

3. 제안된 표현 방법

이 장에서는 계통수에 대한 OTU의 3차원 뷰를 생성하기 위한 방법을 소개한다. 먼저, 몇 가지 내용을 정리한다.

S_i = i-th OTUs (i=1, ..., n)

n = 표시되는 OTU의 개수

$P_i = (x_i, y_i, z_i)$: S_i에 대응되는 3D 공간의 좌표

$\delta(S_i, S_j)$ = 서열정렬과 같은 평가 방법으로 얻은 S_i와 S_j사이에 거리

$d(P_i, P_j)$ = 3D 상에 P_i와 P_j의 거리

3.1. 표현 방법(embedding method)

일반적으로 계통수 구성 순서는 다음과 같다. 우선, 다중 서열정렬은 각각의 OTU들에 대응되는 서열에 대하여 처리된다. 다음으로 정렬된 서열 쌍 사이의 거리를 계산한다. 그리고 NJ, UPGMA과 같은 몇몇 계통수 구성 알고리즘이 계통수를 생성하기 위해서 적용된다.

OTU들 사이의 진화 거리에 대한 정보를 시각화하기 위해서 3D 공간에서 그들의 유클리디안 거리에 비례하는 방법으로 3D 공간에 OTU들의 위치를 결정하였다.

문헌상에는 주성분 분석, 요인 분석, 다차원 척도법(MDS;Multi-dimensional Scaling)과 같이 차원을 축소하는 기술들이 알려지고 있다. 다차원 척도법은 OTU들의 거리를 가능한 가깝게 하여 낮은 차원의 공간에 데이터를 표현하기 위한 기술이다. 이것은 계통학적 분석을 위해 3D 공간에 OTU들을 표현하기 위한 매우 적절한 기술이다.

이 관찰로부터 3D 공간에 OTU들을 표현하고 전통적인 2D 계통수 표현과 함께 3D 공간에 그들을 보여주기 위한 방법을 제안한다. 계통수에 대한 3D 보조 보기를 생성하기 위하여 제안한 방법으로 non-metric MDS를 사용하였다. 이 방법은 원래 공간상의 거리 $\delta(S_i, S_j)$ 차가 매핑된 공간상의 거리 $d(P_i, P_j)$와 얼마나 일치하는가를 표현하기 위한 Stress라고 불리는 손실함수를 최소화하기 위해 설계된 MDS이다.

다음은 OTU들의 집합에 대하여 3D 뷰를 생성하기 위한 단계이다.

1단계. 모든 서열 S_i, (i=1, ..., n)에 대하여 다중 서열 정렬 알고리즘을 적용하고 정렬된 서열 S_i', (i=1, ..., n)을 만든다.

2단계. 정렬된 S_i'서열 사이에 한 쌍씩 거리를 계산하고 $\delta(S_i', S_j')$가 S_i'와 S_j'사이의 거리가 되는 거리 행렬 $D = |\delta(S_i', S_j')|_{n \times n}$을 만든다.

3단계. 거리 정보 D로부터 2D 계통수를 구성한다.

4.단계. 정렬된 서열 S_i'에 대응되는 3차원 위치 P_i(i=1, ..., n)를 찾기 위하여 거리 행렬 D에 non-metric MDS를 적용한다.

5단계. 시각화 도구에 거리 행렬, 계통수 T, 그리

고 위치들의 좌표 P_i (i=1, ..., n)를 제공한다.

3.2. 시각화 도구

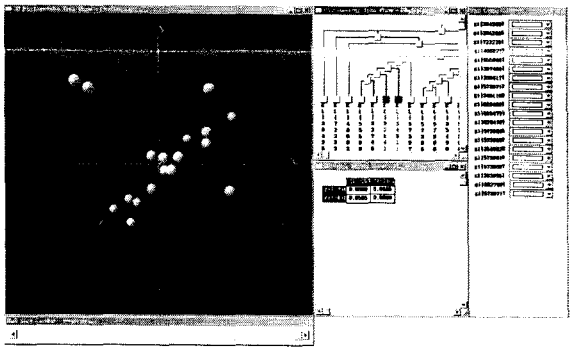
계통발생학적 분석에서 생물학자들에게 도움을 주기 위하여 시각화 도구는 제안된 표현 기술을 이용하여 설계되었다. 이 도구는 2D 계통수와 이것과 대응되는 3D 뷰 모두를 표현한다. 이것은 다음과 같은 특성을 가진다.

- **대화식 제어(interactive control)**: 3D 원소(element) 뷰는 이해하기 쉽도록 이동, 회전, 그리고 크기 조절을 할 수 있다. OTU 원소들은 뷰에서 작은 볼로서 표시되고 그것들의 크기는 제어될 수 있다. 원소들에 대한 레이블은 보기 또는 숨김으로 제어 할 수 있다. 원소의 볼에 대한 색상은 사용자가 임의로 설정할 수 있다.

- **경계 윈도우(the bound windows)**: 2D 계통수 윈도우와 3D 원소 뷰는 경계가 있다. 즉 계통수의 요소는 3D 뷰 윈도우에서 이것에 대응되는 노드로 넘어간다. 한 윈도우 상에서의 연산은 다른 윈도우에 반영된다.

- **특정 원소들의 선택(selection of specific elements)**: 사용자는 특정 원소들을 선택할 수 있고 그들에게 초점을 맞출 수 있다. 단지 선택된 원소들에 대한 거리정보만을 보여줄 수도 있다. 표시된 원소들은 선택할 수 있다.

- **이웃 질의(neighbor query)**: 특정한 원소들에 대해서 이웃한 원소들은 이웃한 범위가 사용자에 의해 제어될 수 있을 때 찾을 수 있다. 원소 사이의 거리는 그 위치들이 non-metric MDS에 의해 얻어졌기 때문에 3D 원소 뷰 공간에서 그것을 계산하는 대신 거리 행렬로부터 얻어진다.



『그림1』 개발된 계통수 분석 도구의 인터페이스

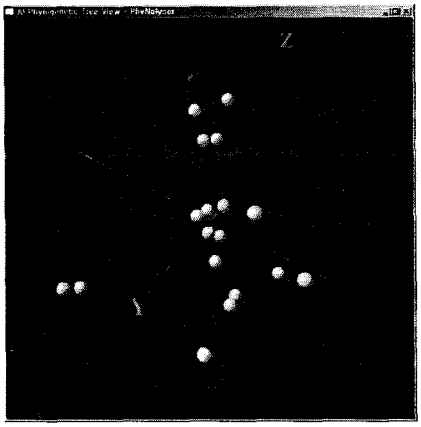
『그림1』 은 PhyNalyser라고 불리는 개발 도구의 인터페이스를 나타낸다. 이 도구는 기본 윈도우로 3D 원소 뷰 윈도우, 계통수 뷰 윈도우, 그리고 거리 행렬 뷰 윈도우를 가지고 있다.

3D 원소 뷰 윈도우는 다양한 보기 연산이 지원되는 3D 공간상에 OTU들을 나타낸다. 계통수 뷰 윈도우는 3D 원소 뷰 윈도우에서 OTU들이 선택되고 선택된 OTU들이 갱신된 계통수를 보여준다. 게다가 계통수에서 OTU들에 대응되는 노드들은 많은 양의 OTU 원소들을 시각화하기 위하여 확장과 축소를 한다.

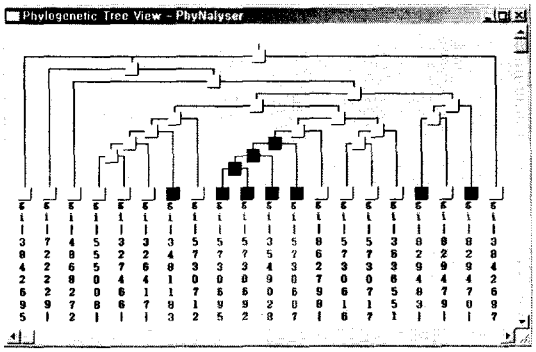
『그림2』 는 OTU를 각 볼 모양 객체로 표시한 3D 원소 뷰 윈도우를 보여준다. 『그림3』 은 서열 집합에 대한 계통수를 나타낸다. 이 그림은 HIV 서열 집합에 대한 계통수의 일부를 보여준다. 『그림4』 는 선택된 원소들에 대한 거리 행렬을 보여주고 있다.

4. 구 현

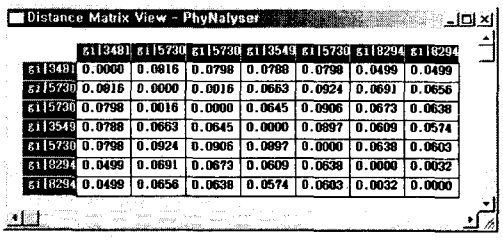
2D 계통수와 이것에 대응되는 3D 뷰에 대한 시각화 도구는 윈도우 환경에서 DirectX API를 사용하여 구현하였다. 이것의 입력으로서 OTU 원소들에 대한 거리 행렬, 계통수 정보, 그리고 OTU 원소에 대한 3D 좌표를 제공한다. 이것들의 입력 요소들은 다른 도구로부터 생성한다.



『그림2』 볼 모양의 객체가 OTU 원소와 대응하는 3차원 원소 뷰 윈도우



『그림3』 계통수를 보여주고 조작하기 위한 계통수 윈도우



「그림4」 선택된 OTU 원소에 대한 거리 쌍을 보여주기 위한 거리 행렬 윈도우

원소에 대한 거리 정보를 얻기 위해서 다중 서열 정렬 도구가 우선 사용되고 정렬된 서열 사이의 쌍 거리를 계산하기 위한 도구가 사용된다. 개발된 시스템에서 ClusterX와 Phylip의 DNADIST 도구는 열(row) 값에 사용한다. 그런 다음 원소에 대해서 계통수 구성 도구를 사용한다. Phylip의 NEIGHBOR 도구는 그 시스템 안에 통합되었다. 3D 좌표의 계산을 위해서 MATLAB 패키지는 원소들의 거리 행렬을 위한 non-metric MDS를 적용하여 구현되었다.

관련된 도구와 패키지의 사용은 개발된 시스템으로서 제시되고, 사용자는 계통발생학적 분석을 위해서 개발된 시각화 시스템을 쉽게 사용할 수 있다.

5. 결 론

몇몇 생물학자들은 2차원 계통수 표현에 대해 만족하지 않았다. 이 논문에서는 3차원 공간에 위치한 OTU들을 표시하기 위한 시각화 방법을 소개하였다. 이 시각화 방법은 2차원 계통수 표현에 대하여 부가적인 보기를 생성하기 위해 사용될 수 있다. 3D 시각화 문제에 대해서 OTU들의 3D 좌표는 OTU들 사이의 거리 제약들을 만족하는 것을 찾는다. 왜냐하면 OTU들 사이의 거리는 유클리안 거리 척도로 유지할 수 없기 때문에 제안한 방법은 non-metric 다차원 척도법 기술을 사용하였다. 제안된 방법은 소프트웨어 도구로 개발되었고 계통발생학적 분석에서 생물학자들에게 도움을 될 것으로 기대한다.

향후 연구과제로서 OTU로부터 3D 좌표를 생성하기 위하여 non-metric MDS의 효과적인 매개변수를 결정하는 것과 계통발생학적 분석에서 유용한 새로운 기능들을 추가하는 것이다.

참 고 문 헌

[1] E. Alpaydin, Introduction to Machine Learning, The MTI Press, 2004.

[2] W. L. Martinez, A. R. Martinez, Exploratory Data Analysis with MATLAB, Chapman & Hall/CRC, 2005.

[3] M. Salemi, A.-M. Vandamme, The Phylogenetic Hanbook: A Practial Approach to DNA and Protein Phylogeny, Cambridge, 2003.

[4] J. Kim, T. Warnow, Tutorial on Phyogenetic Tree Estimation. citeseer.ist.psu.edu/kim99tutorial.html, 1999.

[5] S. F. Carrizo, Phylogenetic Trees: An Information Visualization Perspective. In Proc. the 2nd Asia-Pacific Bioinformatics Conference(APBC2004), Dunedin, New Zealand, 2004.

[6] R. A. Ruths, E. S. Chen, and L. Ellis, Arbor 3D: An interactive environment for examining phylogenetic and taxonomic trees in multiple dimensions, Bioinformatics, 16(11):1003-1009, 2000.

[7] D. A. Keim. Visual Exploration of Large Data Sets, Communications of the ACM, 44(8), 2001.

[8] I. Montealegre and K. St. John, Visualizing Restricted Landscapes of Phylogenetic Trees. In Proc. of the European Conference for Computational Biology(ECCB03), Paris, Sept. 2003.

[9] Z. Yang, Phylogenetic Analysis Using Parsimony and Likelihood Methods, Journal of Molecular Evolution, 42, 1996, pp.294-307.

[10] T. Hughes, Y. Hyun, and D. A. Liberies, Visualising very Large Phylogenetic Trees in Three Dimensional Hyperbolic Space, BMC Bioinformatics, 5(48), 2004.

[11] D. M. Hills, T. A. Heath, K. St. John, Analysis and Visualization of Tree Space, Systematic Biology, 54(3), 2005, pp.471-482.

[12] U. Rost, E. Bornberg-Bauser. TreeWiz: interactive exploration of huge trees. Bioinformatics 18:109-114, 2002.

[13] T. Munzner, F. Guimbretiere, S. Tasiran, L. Zhang, and Y. Zhou. TreeJuxtaposer: Scalable Tree Comparison using Focus+Context with Guaranteed Visibility. In proc. of SIGGRAPH2003, 2003.