데이터 베이스를 이용한 웹 기반 계통수 추론 시스템 설계†

김신석*, 황부현*
*전남대학교 전산학과
e-mail:{sskim,bhhwang}@sunny.chonnam.ac.kr

Design of Web-based Phylogentic Tree Inference System Using DataBase

Shin-Suck Kim*, Bu-Hyun Hwang*
*Dept of Computer Science, Chonnam National University

요 약

계통수는 특정 객체의 분류 즉 특정 객체로부터 추출한 염기서열을 이용하여 그 객체의 소속 분류 집단을 결정하기 위해서 사용될 수 있다. 만약 특정지역에서 획득한 토끼의 종을 구분하기 위해서 이미 분류된 토끼의 염기서열들을 가지고 염기서열들과의 관계를 표현하는 계통수를 제작함으로써, 객체를 분류 할 수 있다. 계통수 제작은 기존의 계통수 제작 도구들(MEGA등)이 사용되지만, 이러한 계통수 제작 도구는 객체의 어떤 특성에 의해서 종이 나뉘어지는 가는 예측 할 수 없다. 계통수 제작에 이용되는 염기서열 데이터는 기존의 염기서열 데이터 베이스들(EMBL,GenBank,DDBJ)에서 인터넷을 이용하여 찾을 수 있지만, 계통생물학을 위해 누적된 데이터가 아니므로, 계통수 제작을 위해서는 사용이 제한적이다. 또 계통수 제작 도구을 사용하기 위해서는 자신이 관련 염기서열 데이터를 수집하여야 한다. 본 논문은 웹기반 계통수 추론 시스템을 제시한다. 본 시스템은 염기서열 데이터를 검색하여, 계통분류 즉 계통수 제작을 위한 데이터로 저장하고, 이를 이용하여 계통수를 그릴 수 있다. 또한 이렇게 저장된 데이터는 데이터 마이닝 분류 기법을 사용하여, 각 객체 분류 집단을 모델링하며, 분류 속성을 예측할 수 있다.

1. 서론

최근 생물학의 문제를 전산학의 기법으로 해결하려는 생물정보학이 활발히 연구되고 있다. 특히 DNA 염기서열 분석이 자동화가 되어가고 있어, 새로운 염기서열 정보가 세계적인 자료은행들(EMBL, GenBank, DDBJ등)에 의해서 누적되고 있다. 이러한 자료를 바탕으로 서열의 특성 및 진화적 관계를 파악하고 데이터 마이닝 기법을 통하여 새로운 지식을 발견하고자 하는 연구가 진행되고 있다[1][2].

염기서열 데이터는 생물의 진화관계를 추적하는 계통수 제작에서의 이용 또한 활발하다. 염기서열은 DNA → RNA → Protein →→→ phenotype의 흐름을 갖는다. 이러한 일련의 유전자의 흐름에서 최종산물인 표현형을 비교 관찰하기보다는 세포 내·외적 환경의 영향을 적게 받는 DNA 즉 서열데이터를 직접 비교 분석하는 일이 생물의 진화관계를 추적하는데 객관적인 중요한 단서를 제공한다[2].

현재 가장 널리 이용되는 DNA염기서열을 이용한 계통수 제작 방법은 염기서열간의 유사도를 측정하여 각 표현형(객체)간의 차이를 나타내는 방법이다. 이 방법은 염기서열의 차이가 생물 객체군간의 차이의 근본이라는 점을 토대로 그 차이가 적을수록 가까운 관계임을 의미한다[3]. 이를 바탕으로 표현된 계통수는 분류된 집단간의 상대적인 차이는 알 수 있지만, 분류된 집단의 특성을 정의 할 수 없다. 또한 기존의 자료은행들의 자료는 계통수 제작을 목적으로 누적된 자료가 아니기 때문에 그 이용이 제한적이다.

본 논문에서는 계통생물학에서 필수적인 단계인 계통수를 제작하는 시스템을 제안한다. 본 시스템에서 각 서열데이터는 계통수를 제작하기 위한 형태로 데이터 베이스에 저장된다. 이러한 데이터는 웹 상에서 공유되어 사용되고, 계통수 제작시 필요한 염기서열 데이터를 검색할 수 있으며, 저장된 염기서열 데이터를 이용하여, 각 객체 분류군의 특성을 모

† 이 논문은 1999년도 한국과학재단 특정기초 연구비 지원에 의한 결과임(과제번호:1999-2-303-006-3)

델링하는 방법을 제공한다. 이 방법은 데이터 마이닝의 분류기법에 의해 이루어진다.

2. 관련 연구

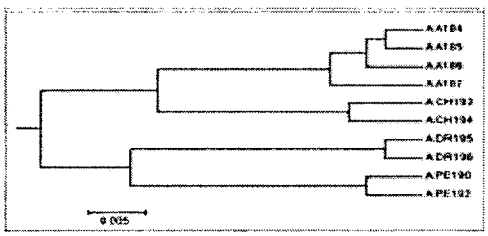
2.1 계통수 추론

이 절에서는 현재 생물학에서 유전체 염기서열을 이용한 계통수 추론 도구의 하나인 MEGA[4]를 통해서 기존 유전자 계통수에 대하여 살펴본다.

DNA 염기서열을 이용하여 유전자 계통수를 제작하는 대표적인 방법은 UPGMA (Unweighted Pair Group Methods using arithmetic Averages)방법과 절약분석을 이용한 방법 등이 있다. 이 중 UPGMA는 유사도를 바탕으로 하여, 유사도가 가장 높은 객체들을 우선적으로 하나의 군으로 모으는 방법이다. 2개 이상의 객체들로 이루어진 군간의 유사도는 각 객체의 유사도의 평균값을 취한다[2].

두 염기서열간의 유사도는 각 위치에서 염기가 같은가 다른가를 가지고 계산한다. 이러한 유사도 측정은 각 염기간의 변이 특성을 고려하여 계산해야 한다. 실제 생체 내에서는 A⇔G, C⇔G로 변환되는, 즉 같은 퓨린 염기가 퓨린 염기로, 피리미딘 염기가 피리미딘 염기로 치환되는 확률과 퓨린 염기가 피리미딘 염기로 치환될 확률이 다르기 때문이다.

MEGA에서는 이러한 염기 변이의 특성 등을 고려하여 계통수를 추론한다. 다음 [그림 1]은 MEGA에서 유사도에 바탕을 둔 UPGMA방법으로 그려진 계통수의 예이다.



[그림 1] MEGA에 의한 계통수(UPGMA방법사용)

2.2 데이터 마이닝 분류기법

데이터 미아닝은 통상 대용량의 실제 데이터로부터, 미리 알려지지 않았지만, 잠재적으로 유용한, 암시적인 정보를 발굴하는 작업이다[5]

데이터 마이닝은 탐사하고자하는 대상지식에 따라 데이터분류, 데이터 클러스터링, 연관규칙, 요약, 유사 시계열 패턴 등의 방법을 적용할 수 있다. 이 중 데이터 분류란 이미 분류된 객체 집단군 즉, 학습 데이터에 대한 분석을 바탕으로 아직 분류되지 않는 객체의 소속집단을 결정하는 작업이다[6].

이러한 결정 트리 분류기법의 대부분은 다음 2단계로 이루어진다[7].

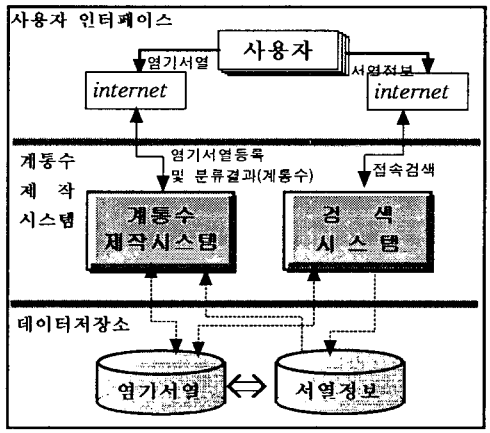
- 트리 성장 단계(Tree Building)
  각 속성을 평가하여 최적 분할 속성(best split attribute)을 선택하고, 이를 이용하여 학습데이터 집합을 최적 분할 속성값에 따라 부분집합으로 분할 된다. 이러한 선택-분할 과정은 부분집합에 순환적으로 계속된다. 데이터 집합이 동일한 클래스이면 멈춘다.
- 트리 전지 단계(Tree Pruning)
  트리 성장 단계에서 완성된 트리에서 오류유발 데이터와 통계적 변동을 가지는 가지들을 제거하는 과정이다.

최근 염기서열 데이터로부터 데이터 마이닝 기법을 이용하여 새로운 지식들을 발견하고자 하는 연구가 진행되고 있다[8].

3. 웹기반 계통수 추론 시스템

3.1 시스템 구성

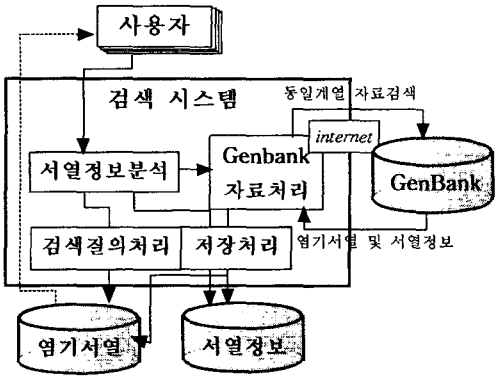
본 시스템은 [그림 2]와 같이, 이미 분류된 염기서열을 검색하는 검색시스템과, 검색결과 나온 염기서열과 사용자가 추출한 서열을 비교하여 진화 관계를 표현하는 계통수 제작시스템으로 구성된다.



[그림2] 시스템 전체 구조

3.2 검색시스템

사용자로부터 서열정보를 받아, 서열정보에 대한 검색을 Genbank와 데이터베이스을 통해서 이루어진다. 검색시스템은 기존의 염기서열자료가 누적된 GenBank로부터 동일계열의 자료를 검색하고 저장한다. 또한 이렇게 저장된 자료는 다시 계통수을 제작하는 단계에서 사용된다.



[그림3] 검색시스템 구조

3.2.1 서열정보분석

계통수 추론을 위한 염기서열은 특정 위치의 유전자를 가지고 이루어지고 있다. 또한 생물의 계통은 종 - 속 - 과 - 목 - 강 - 문 - 계의 단계적인 단위로 표현되며, 계통수는 특정 단위에 대하여 만들어진다. 그러므로 사용자의 서열정보로부터 그 계통 단위를 받아 동일 계열의 정보를 검색할 수 있다. 이러한 검색은 인터넷을 통한 검색과 본 시스템의 데이터 저장소를 통해서 이루어진다.

3.2.2 GenBank 자료처리

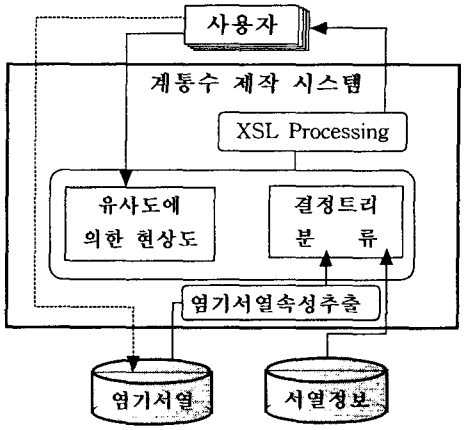
이 모듈에서는 GenBank에서 얻어진 결과를 처리하는 모듈이다. GenBank의 자료는 계통수 제작에 필요한 자료가 아니다. 각 객체간의 비교를 위한 자료는 특정 부위에서(ex 미토콘드리아 cytochrome b) 산출된 동일한 길이를 가지고 있어야 한다. 이 모듈에서는 필요한 염기서열만을 추출하여 사용자에게 제공되고, 데이터 베이스에 저장된다.

3.2.3 검색질의 처리 및 저장처리

검색 또는 저장처리 모듈은 데이터베이스의 자료를 사용하기 위한 모듈이다. GenBank 자료처리의 결과로 얻어지는 자료를 데이터베이스에 저장하는 저장처리 모듈과, 사용자의 질의 요청을 데이터베이스에 요구하고 그 결과를 사용자에 제공한다.

3.3 계통수 제작 시스템

계통수 제작 시스템은 [그림 4]와 같이, 기존 계통수 제작 방법에 의한 유사도에 의한 현상도 모듈과 데이터 베이스의 정보를 이용하는 결정트리 분류 모듈로 구성되었다. 이 두 모듈은 통합되어 결과를 하나의 XML로 표현한다.



[그림 4] 계통수 제작 시스템

3.3.1 유사도에 의한 현상도

계통수를 제작하는 방법들은 다양하고, 그 형태 또한 다양하다[3]. 본 시스템에서는 우선적으로 현재 가장 널리 이용되고 있는 유사도에 바탕을 둔 UPGMA 방법에 의한 현상도로써 계통관계를 표현한다. 또한 유사도의 측정은 염기서열의 배열에 따른 상동, 비상동문제, 아미노산 치환등을 고려하는 Tamura-Nei distance 계산 방법[4]을 사용한다.

3.3.2 염기서열 속성 추출

염기서열은 AGTC의 네 개의 문자로 이루어진 문자열이다. 이러한 문자열을 이용해서 결정트리를 분류하기는 힘들다. 그래서 본 시스템은 원시 염기서열로부터 서열간의 차이점을 염기량, 패턴, 위치에 따른 변화량을 이용해서 다시 표현한다.

3.3.3 결정트리 분류

생물학에서 사용되는 계통수는 종간 분류 속성을 예측 할 수 없다. 그러나 결정트리 분류를 이용한 방법은 종간 분류 속성의 예측이 가능하고, 종 분류 모델을 형성할 수 있다.

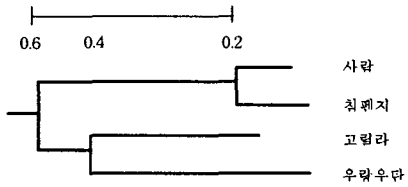
이러한 결정트리 분류방법은 이미 분류된 염기서열이 염기서열속성 추출단계를 거쳐 추출된 속성과 서열 정보를 사용한다. 클래스로 사용되는 서열 정보는 생물학적 계층 단위(아종-종-속-과 등)로 이루어져 있었기 때문에 본 시스템에서는 각 단위별로 그 결정트리를 생성한다. 결정트리 분류 모듈은 사용자로부터 아직 분류되지 않은 염기서열을 받아, 기존 데이터로부터 얻어진 종 분류 모델을 사용하여 종을 분류한다.

3.3.4 XML을 통한 결과 제공

유사도에 의한 결과와 결정트리 분류에 의한 결과는 XML로 표현되어 사용자에게 제공된다. 이렇게 표현된 결과는 웹 브라우저에서의 표현을 위해

XSL로 처리한다.

만약 사람, 침팬지, 고릴라, 오랑우탄의 현상도(계통수)를 [그림 5]와 같이 가정하면, [그림 6]과 같이 XML로 표현될 수 있다. 아래 그림은 사람과 침팬지, 고릴라와 우랑우탄이 유사한 집단임을 나타낸다.



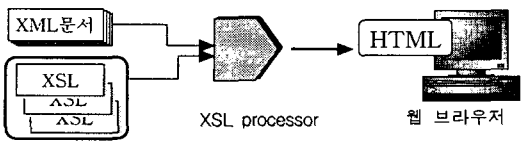
[그림 5] 비유사도에 의한 현상도

```
<?xml version="1.0" encoding="EUC-KR"?>

<PhylogenticTree>
- <split attribute="G+C" attvalue="10" distence="0.6" splitunit
  ="famliy" unitname="영장류" nodeposition="root">

  - <split distence="0.2" nodeposition="Yes" Splitunit=
      "genus" unitname="영장류">

    - <leafnode splitunit="species" unitname="사람">
        사람
      </leafnode>

    - <leafnode splitunit="species" unitname="침팬지">
        침팬지
      </leafnode>
    </split>
  - <split distence="0.4" nodeposition="No" Splitunit=
      "genus" unitname="영장류">

    - <leafnode splitunit="species" unitname="원숭이">
        고릴라
      </leafnode>

    - <leafnode splitunit="species" unitname="우랑우탄">
        오랑우탄
      </leafnode>
    </split>
  </split>
</PhylogenticTree>
```

[그림 6] [그림 5]에 대한 XML

[그림 6]의 <PhylogenticTree>는 2개의 <split>를 가지며, split에는 여러 개의 속성이 있다. 그리고 최종 객체는 <leafnode>로 표현되었다.

3.3.5 XSL Processing



[그림 7] XML문서처리 모델

[그림6]과 같은 XML문서는 [그림7]과 같은 XSL Processing 과정을 통하여 웹 브라우저에서 계통수로 표현된다. 이러한 과정은 동일 결과에 대하여 다양한 XSL을 사용하여 사용자에게 여러 가지 계통수 형태를 보여줄 수 있다.

4. 결론

본 논문에서 제시한 웹기반의 계통수 제작 시스템은 별도의 프로그램 설치과정 없이 인터넷 웹 브라우저를 통해 계통수를 그릴 수 있다. 또한 데이터베이스와 연동되어 사용되므로 자료를 공유하고 검색할 수 있으며, 본 데이터 베이스에는 기존의 유전자 데이터베이스를 검색하여 계통 분류에 적합한 데이터로 변환되어진 데이터들이 저장된다. 데이터 베이스의 이용은 기존의 계통수에서 예측하지 못하였던 분류 정보를 데이터 마이닝 기법을 이용하여 예측한다. 그리고 이러한 결과들은 XML로 표현되고, XSL을 이용하여 웹 브라우저를 통하여 다양한 형태로 사용자에게 제공된다.

참고문헌

[1] 장병탁, "바이오정보기술(BIT)와 바이오지능(Biointelligence)" 2001 춘계학술발표회 초청강연, p46 -95.

[2] 김기중, "분자생물학적 자료와 계통수 제작," 한국유전학회 2000, p259-272.

[3] Sokal R.R., P.H.A. Sneath, "Principles of numerical taxonomy," W.H. Freeman and Co., SanFrancisco.

[4] MEGA : M. Nei(Pennsylvania State Univ.) group, IBM PC용 program.

[5] 이도헌 "데이터 마이닝 : 개념 및 연구 동향," 데이터베이스 연구회지 13권 4호, 1996.

[6] 정민아, 이도헌, "데이터의 다중 추상화 수준을 위한 결정트리," 한국정보과학회 학술발표 논문집(B), p82-84.

[7] Manish Mehta, Rakesh Agrawal and Jorma Rissanen, "SLIQ:A Fast Scalable Classifier for Data Mining," EDBT 96, Avignon, France, March 1996.

[8] Setubal J, Meidanis J., "Introduction to Computational Molecular Biology," Boston, MA:PWS Publishing Company, 1997.