트레이닝 데이터 생성과 의사 결정 트리를 이용한 계통수 생성 방법

채 덕 진[†] · 신 예 호[††] · 천 태 영[†††]
고 홍 선[†††] · 류 근 호[††††] · 황 부 현[†††††]

요 약

전통적인 동물 계통수(系統樹)는 초기발생 형질에 기초하여 몸 구조가 단순한 것에서 복잡한 것으로 동물문(animal phylum)들을 배열하는 것이다. 현재 활발하게 연구 진행되는 분자수준에서의 분자계통 분류학(Molecular Systematics) 연구들이 이런 경향을 재평가하고 새로운 계통과 진화의 의미를 제시하고 있다. 본 논문에서는 한 염기서열로부터 획득할 수 있는 특성 값들을 추출하여 트레이닝 데이터를 생성하고, 생성된 데이터를 기반으로 데이터마이닝 기법중의 하나인 분류기법(classification)을 사용하여 계통수를 생성하였다. 실험용 데이터는 미토콘드리아 염기서열을 사용하였으며 생물학분야에서 사용하는 분석 프로그램인 MEGA 프로그램을 사용하여 이를 증명하였다. 비록 마이닝을 수행한 결과는 생물학적 실험을 거쳐 정확성을 검증 받아야 하지만 인터넷상에 떠다니는 무수한 유전체들에 대한 유효한 분류기준을 제시할 수 있고 계통수 제작을 위한 실험에 소요되는 많은 시간과 노력들을 줄일 수 있다.

The Training Data Generation and a Technique of Phylogenetic Tree Generation using Decision Tree

Duck Jin Chai[†] · Ye Ho Shin[††] · Chun Tae Young[†††]
Koh Hung Sun[†††] · Keun Ho Ryu[††††] · Buhyun Hwang[†††††]

ABSTRACT

The traditional animal phylogenetic tree is to align the body structure of the animal phylums from simple to complex based on the initial development character. Currently, molecular systematics research based on the molecular, it is on the fly, is again estimating prior trend and show the new genealogy and interest of the evolution. In this paper, we generate the training set which is obtained from a DNA sequence and apply to the classification. We made use of the mitochondrial DNA for the experiment, and then proved the accuracy using the MEGA program which is analysis program, it is used in the biology field. Although the result of the mining has to be proved through biological experiment, it can provide the methodology for the efficient classify and can reduce the time and effort to the experiment.

키워드 : 데이터마이닝(Data Mining), 분류규칙(Classification), 계통수(Phylogenetic Tree), 계통분류학(Molecular Systematics)

1. 서 론

계통분류학이란 각 분류군에서 계통유연관계(phylogenetic relationship)를 나타내는 형질들을 비교, 분석하여 각 분류군의 진화계통 유연관계를 밝히는 계통학적 연구 방법론으로 주로 형태적 형질과 DNA 염기서열 등의 분자 생물학적 형질을 이용하여 계통도(cladogram)를 작성한다.

지난 20여년 동안의 분자 생물학의 급진적 발전은 현대 계통분류학에 커다란 변혁을 가져왔다. 특히 유전의 근원 물질인 DNA나 RNA를 분리·조작·분석하는 기술의 발전은 현대 계통 분류학의 가장 중요한 실험방법으로 자리 잡고 있다.

DNA → mRNA → protein ----→ phenotype(표현형)으로 표현되는 일련의 유전자의 흐름에서 최종산물인 phenotype을 비교 관찰하기보다는 세포 내·외적 환경의 영향을 적게 받는 DNA를 직접 비교 분석하는 것이 생물의 진화역사를 추적하는데 객관적인 중요한 단서를 제공할 수 있다[15].

많은 분야에서 데이터베이스에 저장되는 데이터의 양이 증가하고, 데이터베이스의 응용 범위가 확대됨에 따라 대용

※ 이 연구는 한국과학재단(1999-2-303-006-3)의 연구비 지원으로 연구되었음.
† 준 회 원 : 전남대학교 대학원 전산학과
†† 정 회 원 : 극동대학교 정보통신학부 교수
††† 정 회 원 : 충북대학교 생물학과 교수
†††† 종신회원 : 충북대학교 전기전자 및 컴퓨터공학부 교수
††††† 정 회 원 : 전남대학교 전산학과 교수
논문접수 : 2003년 5월 6일, 심사완료 : 2003년 7월 3일

량 데이터베이스로부터 유용한 지식을 발견하고자 하는 데이터 마이닝(data mining) 기술에 대한 연구가 활발히 진행되고 있다[1, 5, 8]. 데이터 마이닝 기술은 특성화(characterization), 군집화(clustering), 분류(classification), 연관 규칙(association rule), 경향 분석(trend analysis), 패턴 분석(pattern analysis) 등으로 나눌 수 있다. 특히 통계학과 관련하여 분류나 군집화기법 등은 생물정보학에서도 이미 응용되고 있다[2, 5, 9-13].

분류기법이란 이미 분류된 객체 집단군 즉, 학습 데이터(training data)에 대한 분석을 바탕으로 아직 분류되지 않는 객체의 소속 집단을 결정하는 작업이다[7]. 현재까지 제안된 여러 가지 분류 모델(classification model) 중 의사 결정 트리는 인간이 이해하기 쉬운 형태를 가지고 있기 때문에 매우 유용하다. 본 논문에서는 유전체 염기서열로부터 각 염기의 위치와 양이 종을 분류하는데 큰 의미가 있다는 가정하에 먼저 각 염기에 가중치를 주어 몇 가지 속성들을 추출하여 분류기법을 적용하기 위한 속성 테이블을 생성한다. 그리고 이를 기반으로 데이터마이닝 툴인 See5를 사용하여 의사 결정 트리를 생성한다. 생성된 트리는 계통분류학 분야에서 사용될 수 있는 의미 있는 정보인지를 확인하기 위해 MEGA 프로그램을 사용하여 증명한다. 분류기법을 사용하는 이유는 동물 계통 분류는 유전적 유연 관계를 통해 동물들의 계통수를 분류하는 것이기 때문에 마이닝 기법 중 분류기법과 매우 잘 결합될 수 있기 때문이다.

계통수는 각 객체간의 관계를 염기서열의 유사도로 표현하고 이러한 유사도는 염기서열간의 차이로써 표현된다. 그러나 유사도로 표현되는 염기서열간의 차이는 비교 대상 염기서열들 사이의 상대적 차이를 나타내는 것일 뿐 개별 염기서열들의 자체 특성을 반영하지 못한다. 본 논문에서 사용되는 의사 결정 트리는 유사도로 표현되는 염기서열간의 차이가 갖는 문제를 해결하기 위해 객체의 속성을 이용하여 분류 집합의 특성을 설명할 수 있으며, 분류 집합간의 차이를 속성으로 표현하고 규칙화 할 수 있다. 물론 실제적인 생물학적 실험을 거쳐 증명하여야 한다 하더라도 실험에 소요되는 많은 시간이나 노력들을 해소할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 본 논문과 관련된 계통분류학과 데이터마이닝 기법인 분류 기법에 대해서 알아보고, 3장에서는 유전체 염기서열 데이터로부터 의사 결정 트리를 생성하기위한 속성을 추출하는 방법에 대해서 정의한다. 그리고 4장에서 속성들을 사용하여 전산학 분야에서 사용하는 의사 결정 트리 프로그램과 생물학 분야에서 사용하는 계통수 제작 프로그램의 결과를 비교 평가한다. 마지막으로 5장에서 결론을 내리고 문제점과 추후 연구 방향을 논한다.

2. 관련 연구

2.1 동물 분자계통 분류학(Animal Molecular Systematics)

분자계통 분류학의 초기 연구들은 대부분 ribosomal DNA에 근거하여 이루어졌다. 이 연구 결과들은 기존의 형태 형질을 통하여 의문시 되었던 여러 문제들을 해결해 주었으나 새로운 문제점들을 제시하기도 했다. 이로 인해 많은 분자계통분류학자들은 한 종류의 유전자만으로 어떤 결론을 유도하는 것은 어렵다고 보고, 가급적이면 다양하고 많은 유전자들의 염기서열을 결정하여 이를 계통분류에 이용하려는 시도의 필요성을 인식하기 시작하였다.

미토콘드리아 유전자들을 동물의 계통분류 연구에 이용하는 연구들은 지난 10여년 동안 매우 활발하게 이루어져 왔으며 대부분의 경우가 몇몇 특정 유전자들 즉, ribosomal RNA, cytochrome b, CO1, ND1 and ND2 등의 전체 또는 부분적인 염기서열만을 가지고서 수행되었다. 최근에는 유전자 염기서열을 결정하는 기법들이 급속도로 발전하면서 미토콘드리아 전체의 염기서열이나 유전자 배열 순서를 가지고 계통 분류에 적용하려는 시도들이 행하여지고 있다.

2.2 DNA 염기서열을 이용한 동물계통분류

분자생물학적 자료들을 이용한 동물계통 분류 방법에는 단백질 분석 방법, DNA/DNA 잡종 방법, Microsatellite와 Minisatellite, 제한효소 절단 위치 비교, RAPDs와 AFLP, DNA와 RNA의 염기서열 비교 방법 등이 있다. <표 1>은 앞에서 언급한 계통 생물학에 이용되는 분자 생물학적 방법들의 활용범위를 나타내고 있다[15].

현재 유전자 크로닝(gene cloning), 중합효소 연쇄반응(PCR)에 의한 증폭, 염기서열 결정(Sequencing), 게놈지도 작성(genome mapping) 기술 등과 같은 여러 가지 분자생물학적 기술의 보편화 및 기계화로 짧은 시간에 다량의 유전학적 정보를 얻을 수 있고, DNA 염기서열 분석은 점점 자동화가 되어가고 있는 추세이다. 더욱이 EMBL, Genbank, DDBJ와 같은 세계적인 염기서열 자료은행들은 하루가 다르게 새로운 염기서열 정보가 누적되어지고 있다. DNA 염기서열의 차이가 생물종간 차이의 근본이고, 염기서열의 비교분석이 용이하며 누구나 웹을 통해 데이터를 공유할 수 있다는 장점으로 인해 계통분류학에 염기서열을 이용하는 것이 효과적이다[15].

2.3 의사 결정 트리 분류(Decision-Tree Classification) 기법

과거 컴퓨터 데이터 관리의 성공으로 대용량의 데이터 축적이 이루어졌다. 이러한 데이터는 수동적인 데이터로 이는 데이터 마이닝이나 KDD에 의해 능동적인 정보로 바뀌어 질 수 있다. 데이터 분류는 능동적인 정보를 생성하는

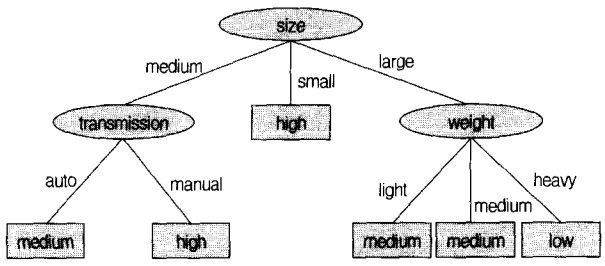
데이터 마이닝의 한 부분이다. 데이터 분류의 목적은 입력 데이터를 분석하여 각 클래스에 대한 정확한 표현(description)이나 모델을 개발하는 것이다.

〈표 1〉 계통생물학의 문제점을 풀기 위하여 이용되는 분자생물학적 방법의 적용범위

(- : 부적당, L : 제한적, + : 적당, E : 적당하나 경비과다)

문제점	DNA 잡종	제한효소 분석	RAPDs/ AFLP	DNA/RNA 염기서열
유전자진화	-	L	-/-	+
집단생물학	-	+	L/-	+
생식집단구조	-	L	L/-	E
크론의 범위	-	+	+/+	E
대립유전자빈도	-	+	-/L	L
어버이검증	-	L	L/+	E
개체의 연관관계	-	L	L/L	E
지리적 변이	-	+	L/L	+
잡종현상	-	+	+/-	E
종의범위	-	+	+/-	+
계통(0~5백만년)	L	+	-/-	+
계통(5백만년~5천만년)	+	+	-	+
계통(5천만년~5억년)	L	L	-	+
계통(5억년~35억년)	-	-	-	+

입력 데이터는 속성이나 특성에 대한 레코드로 구성된다[7]. 또한, 의사 결정 트리에서 각 non-leaf 노드는 하나의 속성을 나타내고, non-leaf 노드로부터 뻗어 나오는 각 가지는 속성 값을 나타낸다. 그리고 각 leaf 노드는 클래스를 나타낸다. (그림 1)은 의사 결정 트리를 보여주는 하나의 예이다. 이 트리는 세 가지 속성(size, transmission, weight)과 클래스(high, medium, low)를 가진 트레이닝 데이터에 대한 의사 결정 트리이다.



(그림 1) 의사 결정 트리의 예

학습 데이터로부터 데이터 분류집합의 특성을 추출하는 의사 결정 트리 분류기법은 CART, ID3, C4.5, SLIQ 등의 방법이 연구되었다. 의사 결정 트리 분류기법의 대부분은 다음 두 단계로 이루어진다[7].

• 트리 성장 단계(Tree Building Phase)

이 단계에서는 학습데이터 집합의 반복적인 분할(partitioning)로 초기 의사 결정 트리가 만들어진다. 학습데이터 집합은 하나의 속성을 사용하여 2개 또는 그 이상의 부분으로 나누어진다. 이 과정은 각 부분들이 하나의 클래스에 속할 때까지 반복적으로 순환된다. 이러한 트리 성장 알고리즘은 (그림 2)와 같다.

```
MakeTree (trainingdata T)
    Partition (T) ;
Partition (Data S)
    if (S의 모든 레코드가 동일한 class) then return ;
    각 속성 A에 대하여 분할 평가(evaluate splits)
    찾아진 최적의 분할 속성 S에 따라 S1과 S2로 분할

    Partitioin(S1) ;
    Partition(S2) ;
```

(그림 2) 의사 결정 트리 성장 알고리즘

• 트리 전지 단계(Tree Pruning Phase)

트리 성장 단계에서 완성된 트리에서 오류 유발 데이터를 포함하는 가지들 또한 생성된다. 이 단계의 목적은 오류를 포함하는 가지를 제거함으로써 예측 오류를 최소화하는 것이다. (그림 3)은 MDL 기반 의사 결정 트리 전지 알고리즘이다[17].

1. Initialization. At the leaves s of T gather the counts ni(s) i = 0, 1, ···, m-1, and compute S(s) = - lnPT(0) + S0(c(s))
2. Recursively in bottom-up order, put ni(s) = $\sum_j$ ni(sj), i = 0, 1, ···, m-1, the sum over all children sj of s, and set

$$S(s) = \min \begin{Bmatrix} -\ln PT(0) + s0(c(s)), \\ -\ln PT(1) + L(thrs) + \sum_j S(sj) \end{Bmatrix}$$

   If the first element is smaller than or equal to the second, delete all children.
3. Continue until the root λ is reached.

(그림 3) MDL-based Decision Tree Pruning Algorithm

3. 유전체 데이터의 수집 및 특성

본 논문에서 사용된 유전자 염기서열 데이터는 충북대학교 동물분류학 연구실에서 수집한 데이터와 NCBI의 GenBank에 등록된 유전자 염기서열 데이터를 검색하여 수집한 데이터이다. GenBank 유전자 데이터베이스에서는 기본적으로 미토콘드리아 cytochrome b 유전자 염기서열을 검색하여, 대표적으로 양서류(개구리), 파충류(뱀, 유혈목이, 구렁이), 조류(도요새, 매), 어류(붕어, 쉬리) 그리고 무척추동물(산호, 흡충류)이 수집되었다. 데이터를 선별할 때는 미토콘드리아 cytochrome b 유전자 전체 염기서열이 등재된 것을 우선적으로 수집하였다.

3.1 유전자 염기서열 데이터의 분류군별 통계치

수집된 각 분류군별 유전자 nucleotide에 대해서 산술평

균값을 계산하였더니 <표 2>와 같은 결과가 나타났다.

〈표 2〉 각 분류군별 유전자 nucleotide 구성 평균값

(단위 : base pair)

Base / 분류군	A (Adenine)	T (Thymine)	G (Guanine)	C (Cytosine)
양 서 류	262	332	161	353
파 충 류	348	319	129	312
조 류	322	270	136	380
어 류	292	318	166	332
무척추동물	210	452	280	167

<표 2>의 결과를 살펴보면 척추동물(양서류, 파충류, 조류, 어류)의 Cytochrome b 유전자 염기구성의 특징은 Guanine(G)의 함량이 다른 Adnine(A), Cytosine(C) 그리고 Thymine(T) 보다 현저히 적게 나타남을 알 수 있다. 이와는 대조적으로 무척추동물에서는 Cytosine(C)의 함량이 다른 염기들보다 적었고, 특히 Thymine(T)의 함량이 월등히 많은 452 base pair를 보였다.

3.2 각 분류군별 Cytochrome b 유전자 염기서열들의 평균 크기

각 분류군마다 수집된 Cytochrome b 유전자 염기서열들은 각각 다른 크기를 나타내지만 척추동물과 무척추동물의 경우 모두 미토콘드리아 유전체 전체(약 16000개의 염기서열) 내에서 차지하는 비율은 양쪽 모두 각각 약 7.0%(1120개의 염기서열) 정도가 되었고 유전자의 평균 크기는 <표 3>과 같다.

〈표 3〉 수집된 유전자의 분류군별 평균 크기

(단위 : base pair)

양서류	파충류	조류	어류	무척추동물
1143.0	1121.0	1143.0	1141.0	1144.8

3.3 각 분류군별 수집 유전자

각 분류군에 포함된 유전자는 전체 25종으로서 이들은 모두 미토콘드리아 DNA의 Cytochrom b에 대해 완전한 염기서열을 갖고 있는 유전자이다. 이들을 각 분류군에 따라 정리하면 <표 4>와 같다. <표 4>는 양서류, 파충류, 조류, 어류, 무척추동물 강(綱)에서 수집한 유전체 데이터들의 리스트이다.

<표 4>에 나타난 자료들은 모두 Cytochrome b 유전자의 완전한 서열 정보를 가지고 있는 데이터들로서 그 크기는 평균 1140bp 정도가 된다.

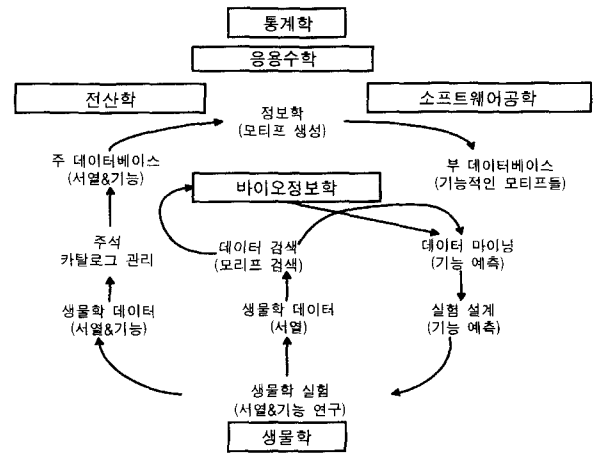
4. 계통수 추론을 위한 분류 규칙 생성 방법

이 절에서는 유전체 정보 분석을 수행할 수 있는 유전체 마이닝 모델을 정의한다. 정의하고자 하는 마이닝 모델은 마이닝 기법 중에서 분류기법(Classification)을 기반으로 한다. 왜냐하면 동물 계통 분류는 유전적 유연 관계를 통해 동물들의 계통수를 분류하는 것이기 때문에 마이닝 기법 중 분류기법과 매우 잘 결합될 수 있기 때문이다.

〈표 4〉 분류군별 유전자 리스트(양서류, 조류, 어류, 무척추동물 綱)

분 류 군	유전자 리스트
양 서 류	개구리(AF205091, AF205090, AF205089, AF205088, AF205087)
파 충 류	뱀(AY099996), 유혈목이(AF471051), 구렁이(AF337173, AF283644, AF283643)
조 류	도요새(AF417924), 매(NC_000878, AF090338, U8337,U83306)
어 류	붕어(NC_002079, AB006953), 쉬리(AP002923, NC_003164, AB085739)
무척추동물	산호(NC_003522, NC_000933, AF338425), 흡충류(NC_002354, NC_002546)

유전체 마이닝은 (그림 4)에서 볼 수 있는 바와 같이 생명공학 분야에서 유전체 서열화나 기능성 유전자 발견을 위한 기초 지식을 찾아내는데 적용됨으로서 매우 중요한 역할을 담당하고 있다[14].



(그림 4) 생물정보학과 생물학의 관계

본 논문에서는 동물의 미토콘드리아 유전자(mitochondrial DNA : mtDNA) 염기서열의 몇 가지 특징을 이용하여 진화적 관계의 분석 문제에 적용하였다. 기존 생물학에서의 진화적 관계의 분석은 진화적 관계의 거리(distance)를 측정하여 계통수를 제작하는 형태로 진행되었다. 이에 반해 본 논문에서는 각 분류군 들로부터 추출한 mtDNA 염기서열에서 획득할 수 있는 특성 값들을 이용하여 데이터 마이닝 기법들 중 하나인 분류기법에 적용하였다.

4.1 분류 모델 정립의 의미

생물의 유전 정보를 담고 있는 염기서열이 환경이나 돌

연변이 등의 요인에 의해 염기 치환이 일어나게 된다면 유전정보는 변화되어진다. 예를 들어, 만약 염기서열 10번째와 11번째에 위치한 염기가 AG이였던 것이 다른 염기로 치환 됐다고 가정하면 이 염기서열에 담겨있는 유전정보는 변화될 수 있다. 즉, 병이나 돌연변이 등의 현상이 발생할 수 있다는 것이다. 그러므로 본 논문에서는 DNA 염기서열이 단순히 랜덤하게 배열되어 있지 않고 각 염기서열의 위치마다 제각기 의미를 담고 있고 각 위치에서 어떤 염기가 배열되어 있는지 또한 중요한 근거로 가정한다. 즉, 종 분류에 있어서 염기서열의 위치, 염기서열의 위치에서 A, G, T, C가 가지는 의미 그리고 전체 염기서열에서 A, G, T, C가 어떤 비율을 이루고 있으며 이것이 또한 어떤 의미를 갖는지를 파악한다.

계통분류학의 일반적인 결과는 트리로 표현되어진다. 그리고 이러한 형태의 트리는 데이터 마이닝 기법 중 분류의 결과와 유사하다. 따라서 전체 염기서열에서 속성들을 추출하고 이러한 속성들을 이용하여 각 종들에 대한 의사 결정 트리(decision tree)를 구성한다.

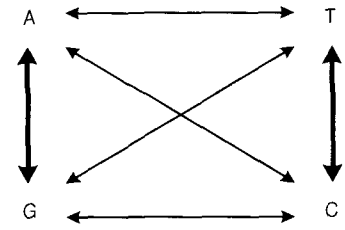
4.2 염기서열에 대한 속성 추출

기존의 계통분류학에서는 각각의 염기서열 사이의 유사성을 비교함으로써, 염기서열간의 관계만을 나타내었다. 이는 DNA의 근본적인 특성을 이용하지 못하고 단지 염기치환에 의존하여 각 분류군간의 거리(아미노산 치환, nucleotide 치환 등)를 표현하며, 다른 부분은 형태학적인 특징을 이용하였다.

유전 정보가 변환되는 것은 염기서열의 치환에 의해서 이루어진다. 그러므로 A, G, T, C의 DNA 염기서열이 단순히 랜덤 하게 배열되지 않고 A, G, T, C가 차지하는 비율, 위치 등에 따라 유전적 의미를 가지고 있다는 가정 하에 본 논문에서는 염기서열에 가중치를 두어 크게 3가지의 속성(횡적가중치, 위치가중치, 각 염기들의 양)으로 표현하였다. 가중치를 둔 이유는 염기서열들의 위치와 양에 대해 차등을 주어 의미를 파악하기 위함이다. 예를 들면, 첫 번째에 A라는 염기가 위치하고 5번째에는 T라는 염기가 위치하고 있을 때, 각 염기 A와 T에 가중치를 주고 다시 그 염기의 위치에 대한 가중치를 주면 어떤 위치에 어떤 염기가 존재하는 가에 따라 전체 염기서열의 가중치들의 합은 다르게 나타난다. 그러므로, 가중치의 변화에 따라 어떤 결과들이 도출되는 지를 확인할 수 있다.

4.2.1 A, G, T, C의 횡적 가중치

원시 데이터(염기 서열)를 이용하여 의사 결정 나무를 작성하기 위해서는 각 서열의 속성 추출 단계를 거쳐야한다. 본 논문에서는 염기서열의 속성 추출을 위하여 각 염기에 가중치를 두었다. 가중치는 (그림 5)와 같이 AGTC 모두가 자유로운 치환이 가능한 염기서열의 치환 모델에서 퓨린(A, G)에는 "1"의 가중치를 피리미딘(T, C)에는 "2"의 가중치를 두었다. 1과 2의 가중치는 고정된 값이 아닌 유동적인 값이다. 즉, 퓨린계와 피리미딘계를 구별하기 위한 값이므로 이들의 값은 모델을 어떻게 세우는가에 따라 변화될 수 있는 값이다. 또, 퓨린과 피리미딘계에 속하는 A, G, T, C 각각에 대해서도 서로 다른 가중치를 부여할 수도 있다.



(그림 5) 염기서열 치환모델

본 논문에서의 퓨린과 피리미딘계 가중치 차등은 상대적으로 각 염기서열 사이의 치환 빈도가 적은 피리미딘계가 염기 자체의 고유한 특성을 보다 잘 반영하기 때문이다. (그림 6)는 퓨린과 피리미딘계의 가중치를 염기서열 AGTTCC에 반영한 것이다.

1	2	3	4	5	6
A	G	T	T	C	C

A, G : "1"
T, C : "2"

횡적가중치 : 1 + 1 + 2 + 2 + 2 + 2 = 10

(그림 6) AGTTCC에 대한 횡적가중치의 예

4.2.2 A, G, T, C의 위치가중치

염기서열의 치환은 유전정보를 변화시킬 수 있다. 그러므로 염기서열의 위치가 아무런 의미 없이 단순하게 배열되어 있지 않다는 가정 하에 각 염기의 위치가 속성으로 잘 표현될 수 있도록 염기 위치에 따라 가중치를 둔다. 분자생물학에서 염기의 위치에 따라 그 특성은 달라질 수 있다. 염기서열의 시작을 기점으로 하여 차례대로 순서를 정하면 1st, 2nd, 3rd 처럼 각 위치에 존재하는 위치 값과 (1)에서 구한 횡적가중치 값을 곱하여 나타내었다. 가중치들을 곱하는 이유는 염기의 위치뿐만 아니라 각 염기들의 수평적인 관계를 알아보기 위함이다. 예를 들면, (1)의 예에서의 위치가중치는 (그림 7)과 같다.

1	2	3	4	5	6
A	G	T	T	C	C

A, G : "1"
T, C : "2"

횡적가중치 : 1×1 + 2×1 + 3×2 + 4×2 + 5×2 + 6×2 = 39

(그림 7) AGTTCC에 대한 위치가중치의 예

〈표 5〉 염기서열의 원시 데이터 및 속성테이블의 예

위치 / OTU	1 - 10 SITE	횡적 가중치	위치 가중치	A	G	T	C	class
AF205091	A T G G C A C C T A	15	86	3	2	2	3	양서류
AY099996	A T G A C C C C A C	16	93	3	1	1	5	파충류
AF417924	A T G G C C C C A A	15	83	3	2	1	4	조 류
NC_002079	A T G G C A A G C C	14	81	3	3	1	3	어 류
NC_003522	A T G C C A T T G C	16	91	2	2	3	3	무척추

4.2.3 A, G, T, C, A+T content 그리고 G+C content

분자생물학에서 G+C content등 각각 염기비율에 따른 변화가 많이 연구되고 있어 본 논문에서도 역시 트레이닝 데이터를 이루는 속성으로 선택한다. 실제 원시 데이터의 경우에는 매우 긴 염기서열로 구성되어 있지만, <표 5>에서는 설명의 편의를 위해 각 1종에 대한 10개 site만을 비교한다.

4.3 의사 결정 트리 생성 및 기존 유전자 계통수와의 비교

이 절에서는 본 논문에서 제안한 속성테이블(training set)로부터 의사 결정 트리를 생성하고 동물분류학 분야에서 적용하던 유전적 거리를 이용한 계통 분류 방법에 의한 실험결과를 비교하고 평가한다. 의사 결정 트리 생성은 데이터마이닝 툴인 See5[16]를 이용하여 생성하고 기존 동물분류학 분야에서의 분류 방법은 계통 분류에 사용하는 MEGA 프로그램을 이용한다. See5는 기본적으로 C4.5 알고리즘을 구현한 프로그램으로써 [16]사이트에서 다운로드 받을 수 있다.

4.3.1 See5를 이용한 의사 결정 트리 생성

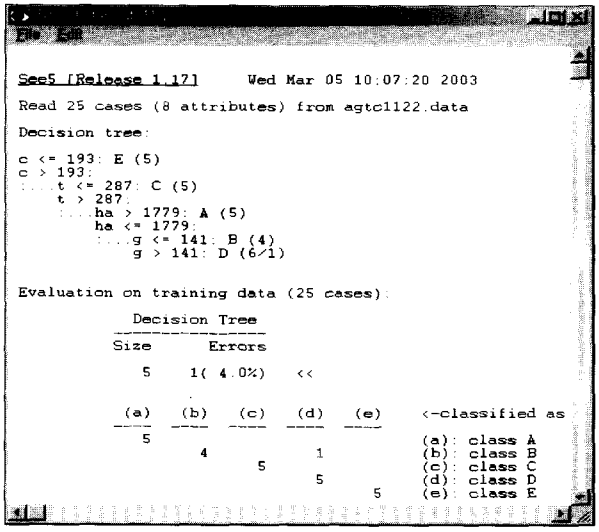
See5에 사용될 속성테이블은 <표 6>과 같다. 데이터는 앞에서 설명한 양서류, 파충류, 조류, 어류, 그리고 무척추동물들에 대한 각 5개종의 서열 데이터에서 추출한 값들이다.

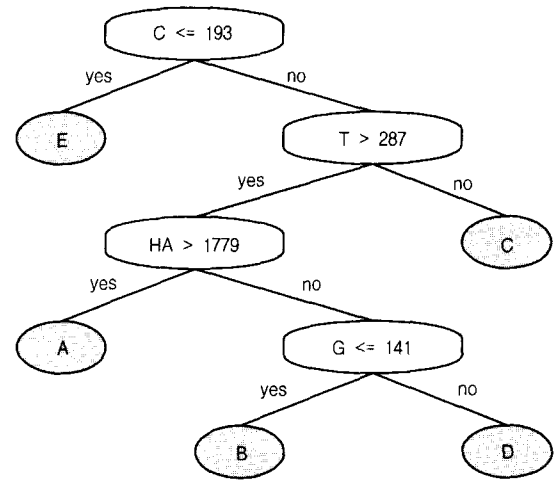
위와 같은 속성테이블이 주어졌을 때 See5 마이닝 툴을 사용하기 위해서는 기본적으로 두 개의 파일이 필요하다. 하나는 위의 속성테이블의 데이터 값을 가지고 있는 파일이고 다른 하나는 속성테이블에 대한 정보를 가지고 있는 파일이다. 이들은 각각 .data와 .names라는 확장자를 갖는다. 속성테이블에 대한 정보 중에서 가장 중요한 것은 각 속성들이 수치적(연속적) 속성인지 아니면 범주형 속성인지에 대한 것과 클래스들의 분포이다. 본 논문에서 사용하고자 하는 <표 6>의 속성 테이블은 클래스 속성을 제외한 8개의 속성으로 이루어져있다. 이 8개의 속성은 모두 수치적 속성이고 클래스들의 분포는 각 클래스가 모두 5개의 레코드들을 가지고 있다.

실험은 파충류를 포함한 것과 포함하지 않은 두 가지의 경우로 진행하였다. 왜냐하면, 결과에서 알 수 있듯이 파충류를 포함 시켰을 때에는 에러가 발생하였기 때문이다. (그림 8)과 (그림 9)는 파충류를 포함한 경우와 그렇지 않은 경우에 See5에서 수행한 결과와 그 결과를 기준으로 의사 결정 트리를 생성한 것이다. 각 속성들의 이름은 편의상 a(염기 A), g(염기 G), t(염기 T), c(염기 C), at(A와 T의 개수의 합), gc(G와 C의 개수의 합), ha(횡적가중치), pa(위치가중치)와 같이 정의하였다. 알파벳 A, B, C, D, E는 각 클래스를 나타낸다.

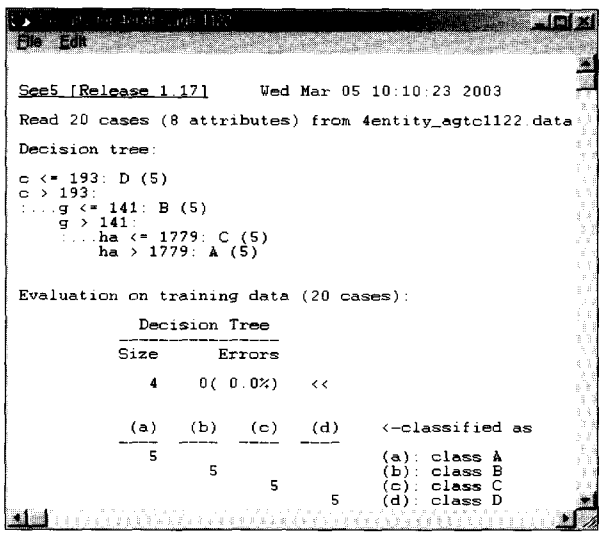
〈표 6〉 의사 결정 트리 생성을 위한 속성테이블(Training set)

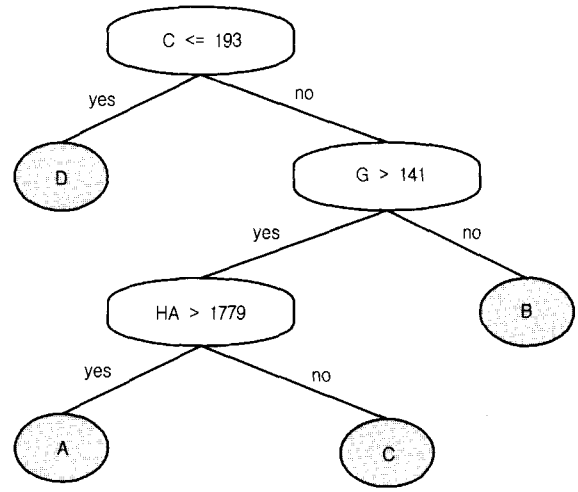
A	G	T	C	A+T	G+C	횡적가중치 (HA)	위치가중치 (PA)	클래스
268	171	336	335	604	506	1781	996402	양서류
263	157	326	364	589	521	1800	1002965	양서류
258	156	322	374	580	530	1806	1007733	양서류
263	163	340	344	603	507	1794	1000081	양서류
259	161	340	350	599	511	1800	1004256	양서류
303	155	300	352	603	507	1762	982186	파충류
333	135	320	322	653	457	1752	974557	파충류
367	117	324	302	691	419	1736	963756	파충류
373	119	325	293	698	412	1728	959370	파충류
367	123	326	294	693	417	1730	961397	파충류
313	138	287	372	600	510	1769	987799	조 류
325	137	274	374	599	511	1758	987487	조 류
325	137	274	374	599	511	1758	987487	조 류
334	131	249	396	583	527	1755	986566	조 류
316	141	269	384	585	525	1763	992884	조 류
313	159	324	314	637	473	1748	975067	어 류
313	159	324	314	637	473	1748	975067	어 류
258	183	299	370	557	553	1779	992569	어 류
258	183	299	370	557	553	1779	992569	어 류
319	150	348	293	667	443	1751	976036	어 류
236	258	446	170	682	428	1726	963276	무척추
255	237	425	193	680	430	1728	969485	무척추
236	258	446	170	682	428	1726	963276	무척추
162	340	417	191	579	531	1718	957537	무척추
162	309	527	112	689	421	1749	973625	무척추





(A : 양서류, B : 파충류, C : 조류, D : 어류, E : 무척추)
(그림 8) 파충류를 포함한 경우의 See5 결과와 의사 결정 트리





(A : 양서류, B : 조류, C : 어류, D : 무척추)
(그림 9) 파충류를 포함하지 않은 경우의 See5 결과와 의사 결정 트리

(그림 8)과 (그림 9)를 보면 파충류를 포함하지 않는 경우는 에러가 발생하지 않고 4개의 클래스로 정확히 분류가 되었다. 그러나 파충류를 포함한 경우에는 다른 4개의 클래스에서는 에러가 발생하지 않았지만 파충류중의 하나가 어류로 잘못 분류되는 것을 발견하였다. <표 7>은 위의 결과에서 나타날 수 있는 분류 규칙들을 정리한 것이다.

4.3.2 MEGA의 결과

MEGA에서는 염기서열들간의 distance를 측정하여 그 관계를 계통수로 표현한다.

MEGA 프로그램의 이용시에는 distance 측정 요소의 선택과 Transition과 Transversion의 이용유무 그리고 계통수 추론을 위한 방법 선택이 필요하게된다. 본 논문에서는 입력 데이터를 아래의 요소들로 선택하여 결과를 산출하였다.

<표 7> 의사 결정 트리에 의한 분류 규칙

파충류 유, 무	분류 규칙(c, g, t : 염기)
유	규칙 1 : IF c <= 193 THEN 무척추 규칙 2 : IF c > 193 AND t <= 287 THEN 조류 규칙 3 : IF c > 193 AND t > 287 AND ha <= 1779 THEN 양서류 규칙 4 : IF c > 193 AND t > 287 AND ha <= 1779 AND g <= 141 THEN 파충류 규칙 5 : IF c > 193 AND t > 287 AND ha <= 1779 AND g > 141 THEN 어류
무	규칙 1 : IF c <= 193 THEN 무척추 규칙 2 : IF c > 193 AND g <= 141 THEN 조류 규칙 3 : IF c > 193 AND g > 141 AND HA <= 1779 THEN 어류 규칙 4 : IF c > 193 AND g > 141 AND HA > 1779 THEN 양서류

• distance 측정 요소(Tajima-Nei distance)

1984년 제안된 방법으로 Tajima와 Nei에 의해서 제안된 방법이다. 염기의 치환수 예측에 관한 방법이며, 거리 측정 방법은 다음과 같다.

$$d(\text{distance}) = -b \times \log e(1 - p/b)$$

p : 서로 다른 염기가 나타날 확률

b : 각 염기에 따른 치환에 대한 가중치의 표현

• 계통수 추론 방법(Transitions + Transversions)

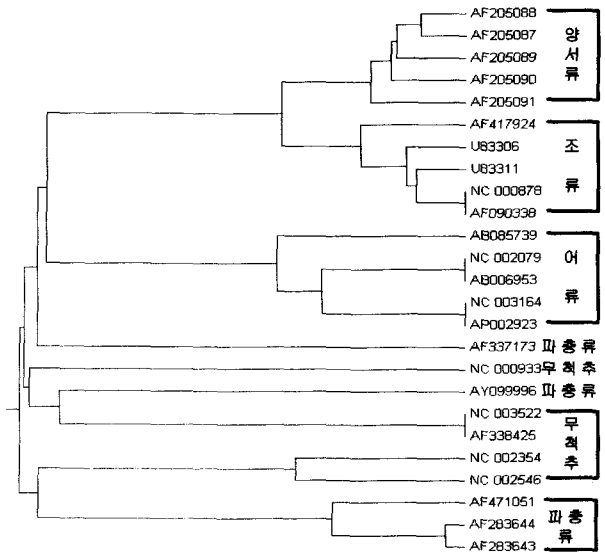
Transition과 Transversion은 치환에서 Transition 또는 Transversion만을 고려할 수도 있다. 이는 Transition 치환과 Transversion 치환은 생물학적으로 발생확률의 차이가 있기 때문이다.

• UPGMA

UPGMA(Unweighted Pair Group Methods using Arithmetic averages)에 근거한 비유사도 지수를 측정한 다음 이 비유사도 지수가 낮은 군들을 하나의 그룹으로 묶는 방법을 통해 계통수를 제작하는 방법이다. (그림 10)은 MEGA에서 UPGMA를 이용하여 제작한 계통수이다.

4.3.3 See5의 결과와 Mega 프로그램의 결과 비교

본 논문에서는 파충류를 포함하지 않은 경우의 Mega 프로그램의 결과를 삽입하지 않았지만 결과는 See5 프로그램의 결과와 같이 나온 것을 알 수 있었다. 그러나 파충류를 포함한 경우는 두 프로그램 모두에서 에러가 검출되었다. 파충류를 포함한 경우 See5에서는 파충류중의 하나가 어류로 잘못 분류되는 것을 발견하였지만 (그림 10)에서 나타나듯이 Mega 프로그램의 결과 또한 무척추와 파충류를 명확히 분류하지 못하는 것을 볼 수 있다. 이는 생물학적 관점



(그림 10) MEGA에서 UPGMA를 이용한 계통수

에서 검증이 된 Mega 프로그램에서도 비슷한 문제가 발생하는 것으로 보아 본 논문에서 제안한 속성테이블의 생성 방법이나 분류 기법상의 문제라기보다는 파충류 DNA 데이터 자체가 갖는 모호성의 문제로 해석하는 것이 타당하리라 판단된다. 아울러 분류대상 유전체의 상대적 차이를 토대로 하는 Mega 프로그램 이용결과와 개별적 유전체의 특성을 기반으로 분류를 수행하는 의사 결정 트리 방법이 오류 데이터의 간섭을 제외한 나머지 영역에서 동일한 결과를 도출함도 확인할 수 있다. 더욱이 See5를 이용한 결과는 어떤 염기의 특성 때문에 각 분류군들이 분류가 되었는지 알 수 있지만 Mega 프로그램은 단순히 염기서열간의 거리만을 가지고 분류하기 때문에 분류가 어떻게 이루어지는 것을 명확히 알 수 없다는 것을 확인할 수 있다.

5. 결론 및 추후 연구방향

본 논문에서는 염기서열 데이터에서 의사 결정 트리 생성을 위한 트레이닝 데이터를 생성하는 방법을 제안하고 분류 기법을 이용하여 동물 분자계통 분류학의 계통수 제작에 응용해 보았다. 생물학 데이터는 양서류, 파충류, 조류, 어류, 무척추 5개 분류군 25개 종을 대상으로 실험하였다. 의사 결정 트리를 생성하기 위한 트레이닝 데이터 집합은 각 염기서열 데이터에서 A, G, T, C에 대한 횡적, 위치 가중치와 A + T content, G + C content 그리고 AGTC 각각의 content를 사용하여 생성하였다. 실험 방법은 데이터 마이닝 분야에서 사용하는 툴 중의 하나인 See5와 생물학 분야에서 사용하는 툴인 MEGA를 사용하여 같은 데이터를 대상으로 결과를 비교 분석하였다. 위의 결과에서 확인되었듯이 본 논문에서 제안한 방법에 의해 생성한 트레이닝 데이터를 가지고 의사 결정 트리를 이용하여 계통수를 제작하는 방법이 분류학 분야에서 사용하는 MEGA라는 프로그램의 결과와 일치함을 확인할 수 있었다. 또한, 제안한 방법은 MEGA에서 알 수 없는 계통수의 특성들을 발견할 수 있다. 즉, 염기서열 데이터의 어떤 속성들 때문에 각 분류군들이 자신들이 속한 분류군으로 분류가 되는지를 (그림 8)이나 (그림 9)와 같이 생성된 의사 결정 트리로써 확인이 가능하다. 비록 마이닝의 결과를 바로 수용하기는 정확성 검증을 거쳐야 하겠지만 생물학 분야에서 소요되는 많은 시간과 노력을 줄이는 데에는 큰 역할을 수행하리라 본다. 향후 연구방향으로는 트레이닝 집합을 생성시 염기서열의 특성을 정확히 반영할 수 있는 속성 추출이 필요할 것이다.

참 고 문 헌

[1] P. Adriaans and D. Zantinge, "Data Mining," Addison-Wesley, 1996.

[2] Alvis Brazma, Inge Jonassen, Ingvar Eidhammer and David

Gilbert, "Approaches to the automatic discovery of pattern biosequences," The Journal of Computational Biology, November, 1997.

[3] J. L. Boore, T. M. Collins, D. Stanton, L. L. Daehler and W. M. Brown, "Deducing the patterns of arthropod phylogeny from mitochondrial DNA rearrangements," Nature 376, pp.163-165, 1995.

[4] T. D. Kocher, W. K. Thomas, A. Meyer, S. V. Edwards, S. Paabo, F. X. Villablanca and A. C. Wilson, "Dynamics of mitochondrial DNA evolution in animals : amplification and sequencing with conserved primers," Proc. Natl. Acad. Sci. USA. 16, pp.6196-6200, 1989.

[5] Usama Fayadd, Gregory Piatetsky-Shapiro and Padhraic Smyth, Chapter 1 From Data Mining to Knowledge Discovery : An Overview, Advances in Knowledge Discovery and Data Mining, AAAI Press, pp.1-34, 1996.

[6] J. L. Boore, D. V. Lavrov and W. M. Brown, "Gene translocation links insects and crustaceans," Nature 392, pp. 667-668, 1998.

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

[8] Peiter Adriaans and Dolf Zantinge, "Data Mining," Addition Wesley, 1996.

[9] P. S. Bradley, U. M. Fayyad and O. L. Mangasarian, "Data Mining : Overview and Optimization Opportunities," http://elib.stanford.edu, Technical Report MP-TR-98-01, 1998.

[10] M. Rebhan, V. Chalifa-Caspi, J. Prilusky and D. Lancet, "GeneCards : a novel functional genomics compendium with automated data mining and query reformulation support," Bioinformatics, Vol.14, No.8, pp.656-664, 1998.

[11] R. J. Hilderman, H. J. Hamilton and N. Cercone, "Data Mining in Large Databases Using Domain Generalization Graphs," Dept. of CS, Univ. of Regina, submitted for publication, 1998.

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

[13] T. Zhang, "Data Clustering for Very Large Datasets Plus Applications," A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Dept. of CS, Univ. of Wisconsin, 1997.

[14] 정재훈 "생물정보학과 인터넷 자원", 한국유전학회, 유전 제3권, pp.176-200, 2000.

[15] 김기중, 분자생물학적 자료와 계통수 제작, 한국유전학회, 유전 제3권, pp.259-271, 2000.

[16] RuleQuest Research Data Mining Tools, See5, http://www.relequest.com/index.html.

[17] M. Mehta, J. Rissanen and R. Agrawal, "MDL-based Decision Tree Pruning," Proc. of the 1st Int'l Conference on Knowledge Discovery in Databases and Data Mining, Montreal, Canada, August, 1995.

채 덕 진

e-mail : djchai@sunny.chonnam.ac.kr

1999년 동신대학교 컴퓨터학과(이학사)

2001년 전남대학교 대학원 전산통계학과 (이학석사)

2001년~현재 전남대학교 대학원 전산학과 박사과정

관심분야 : 데이터 마이닝, Bioinformatics 등

신 예 호

e-mail : snowman@kdu.ac.kr

1996년 군산대학교 컴퓨터과학과(학사)

1998년 충북대학교 대학원 전자계산학과 (석사)

2002년 충북대학교 대학원 전자계산학과 (박사)

2002년~현재 극동대학교 정보통신학부 전임강사

관심분야 : 능동 데이터베이스, 시간, 공간, 시공간 데이터베이스, 데이터 마이닝, 이동객체

천 태 영

e-mail : tommy_chun@yahoo.com

1993년 충북대학교 자연과학대학 생물학과 (학사)

1995년 충북대학교 생물학과 대학원(이학석사)

2002년 충북대학교 생물학과 대학원(이학박사)

1997년~1998년 호주 University of Western Sydney 연구원

2001년 우즈베키스탄 교육부 과학기술자문(생물학분야)

2000년~현재 충북대학교 생물학과 강사

2002년~현재 청주교육대학교 과학교육학과 강사

관심분야 : Molecular systematics, Molecular ecology, Bioinformatics

고 흥 선

e-mail : syskoss@cbucc.chungbuk.ac.kr

1973년 서울대학교 사범대학 생물교육학 (학사)

1976년 서울대학교대학원 동물학 전공 (이학석사)

1980년 캐나다 University of Tronto (이학박사)

1980년~현재 충북대학교 생물학과 교수

1997년~현재 (사)생물자원연구회 대표이사

2001년~현재 중국 산동대학교 객좌교수

2002년~현재 충북대학교 자연과학대학 학장

관심분야 : Molecular systematics, Conservation biology, Molecular ecology

류 근 호

e-mail : khryu@dblab.chungbuk.ac.kr

1976년 숭실대학교 전산학과(이학사)

1980년 연세대학교 산업대학원 전산전공(공학석사)

1988년 연세대학교 대학원 전산전공(공학박사)

1976~1986년 육군군수 지원사 전산실(ROTC 장교), 한국전자통신 연구원(연구원), 한국방송통신대 전산학과(조교수) 근무

1989년~1991년 Univ. of Arizona Research Staff(TempIS 연구원, Temporal DB)

1986년~현재 충북대학교 전기전자 및 컴퓨터공학부 교수

관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 객체 및 지식베이스 시스템, 지식기반 정보검색 시스템, 데이터 마이닝, 데이터베이스 보안 및 Bio-Informatics 등

황 부 현

e-mail : bhhwang@chonnam.chonnam.ac.kr

1978년 숭실대학교 전산학과(학사)

1980년 한국과학기술원 전산학과(공학석사)

1994년 한국과학기술원 전산학과(공학박사)

1980년~현재 전남대학교 전산학과 교수

관심분야 : 분산시스템, 분산 데이터베이스 보안, 객체지향 시스템, 전자상거래