단백질 동정을 위한 Mowse 스코어링 방법의 성능 개선

정민아* · 김치연**

Performance Improvement of Mowse Scoring Method for Protein Identification

Min-A Jung* · Chi-Yeon Kim**

이논문은 2006년도 목포대학교 학술연구비 지원에 의하여 연구되었음

요 약

본 논문은 단백질 동정에 이용하는 펩타이드-매스 핑거프린팅 툴 중 하나인 Mowse의 성능을 개선하는 방법을 제안한다. Mowse에서 빈발 요소 행렬은 단백질과 펩타이드 질량에 대하여 일정한 간격으로 생성되어 행렬의 각 원소의 값은 펩타이드의 빈발횟수에 따라 계산된다. 현재 이러한 행렬을 생성하는데 있어서 정해진 간격으로 생성되는데 이러한 간격의 값이 작아질수록 스코어링 값은 정확해진다. 그러나 이러한 간격의 값이 작아질수록 행렬의 크기는 증가하게 되며 이에 따라 스코어링 계산의 복잡도도 증가하게 된다. 본 논문에서는 행렬의 크기를 현재와 같이 유지하면서 스코어링 값을 정확하게 계산하기 위한 새로운 방법을 제안한다. 현재 Mowse에서 검색대상이 되는 단백질 데이터베이스의 분포를 고려하여 비선형적으로 행렬의 간격의 값을 정하는 방법 즉, 임의의 단백질 질량 값이 많은 곳에서는 행렬의 간격을 작게 결정하는 반면 단백질 질량 값이 적은 곳에서는 행렬의 간격을 크게 결정하는 방법을 새롭게 제안하였다. 또한, 성능평가는 Mowse 스코어링 방법과 본 논문에서 제안한 새로운 스코어링 방법에 관하여 수행하고 분석결과를 제시하였다.

ABSTRACT

In this paper, we propose the method that improve the performance of the Mowse. Mowse is the tool of the peptide mass fingerprinting that is used the identification of protein. In Mowse, frequency factor matrix is generated to regular interval for protein and peptide mass and the value of each elements is calculated to frequency of peptide. We propose new method for calculation of exact scoring value maintaining same size of matrix. The proposed method is that decide interval of matrix considering distribution of protein database. That is, interval of matrix is decided to small in many value of protein mass and is decided to large in few value of protein mass. We present the performance result both Mowse scoring method and the proposed scoring method.

키워드
단백질 동정, 펩타이드-매스 핑거프린팅, Mowse, 빈발요소행렬

* 목포대학교 컴퓨터교육과
** 목포해양대학교 해양전자통신공학부

접수일자 : 2007. 6. 5

Ⅰ. 서 론

프로테오믹스(Proteomics) 분야의 발전은 질량분석기술(Mass spectrometry)이 생체분자(단백질, 펩타이드)까지 가능해졌기 때문이라고 말하는 것은 과언이 아니다. 단백질 동정은 젤상에서 실제 관찰되는 단백질과 유전체 데이터상의 단백질을 연결시키는 작업이라 할 수 있다. 최근 질량분석기술에 의한 신속한 단백질 동정은 대량으로 생명현상을 분석할 수 있는 가능성을 제시하고 있으며, 그 분석대상 물질이 고분자인 단백질에까지 가능하게 되었다[1, 2].

질량분석기술을 이용한 단백질 동정방법중의 하나인 펩타이드-매스 핑거프린팅(PMF:Peptide mass fingerprinting)은 단백질 동정에 아미노산 서열을 이용하지 않고 단백질 조각 즉 펩타이드의 질량을 이용한다. 즉 펩타이드 질량은 펩타이드를 구성하는 아미노산 조성에 따른 고유한 속성이라는 점에 착안한 방법이다. 단백질은 특정 제한효소처리하게 되면 펩타이드들로 잘려지게 되는데 이때 잘려지는 부위는 단백질이 가지고 있는 특정 서열에 의존한다. 따라서 단백질의 서열을 알고 있으면 이론적으로 펩타이드의 질량을 예측할 수 있다. MALDI-TOF 라는 질량 분석기를 이용하면 이들 펩타이드들의 정확한 질량을 잴 수 있다. 한편 컴퓨터는 데이터베이스에 있는 모든 단백질에 트립신으로 처리할 경우 얻어지는 조각들의 이론상 질량을 계산하여 새로운 데이터베이스를 만든다. 제한효소로부터 얻은 펩타이드 질량 값을 이 데이터베이스와 비교하여 서로 일치하는 단백질을 찾아낸다. 이러한 일치정도를 표현하기 위한 중요한 요소는 스코어링 방법이다. 결과적으로 PMF방법을 통해서, 신속하게, 실험자의 펩타이드 질량 리스트를 통해서, 데이터베이스내의 가장 유사한 단백질을 검색해 낼 수 있다[3, 4].

위에서 제시한 검색방법은 전산학적 방법을 이용하게 되며, 해당 질량 리스트로부터, 데이터베이스내의 단백질을 찾는 알고리즘 및 웹 검색 서비스 등이 몇 가지 존재한다. 국외에서는 PMF를 위한 알고리즘들로는 PeptideSearch, PeptIdent, Mowse, Mascot, MS-Fit, ProFound 등이 개발되어 왔고 이러한 알고리즘들을 통합하여 사용할 수 있게 하는 연구가 국내에서도 진행되어 왔다. 이러한 알고리즘들을 이용하는데 있어서 스코어링 방법은 중요한 요소 중의 하나로 볼 수 있다. 그러나 기존의 알고리즘들은 사실상 오차를 많이 내제하고 있는 질량분석데이터에 대해서 정확하지 못한 검색결과를 출력하게 되는 경우가 많다. 알고리즘마다 검색결과가 상이하게 나타나며, 실험자로 하여금 해당 검색결과를 신뢰할 수 있는지에 대한 추가적은 정보 또한 제공하지 못하고 있다. 그 외에도 자신만의 서열DB에 대한 PMF 검색 등의 요구사항 등을 만족시켜주지 못한다[5, 6, 7, 8, 9, 10, 11].

이러한 알고리즘들 중의 하나인 Mowse 스코어링 방법은 다른 PMF알고리즘에서도 적용하여 많이 사용하고 있다. 따라서 Mowse 스코어링 방법을 개선하고자 하는 연구는 매우 중요하다고 할 수 있다. 본 논문에서는 Mowse 스코어링 방법을 개선하기 위한 새로운 방법을 제시한다. 이로 인하여 단백질의 동정을 보다 빠른 시간내에 수행할 수 있으며, 보다 정확하게 수행할 수 있다. 또한 Mowse 스코어링 방법을 적용하여 사용하고 있는 다른 PMF 소프트웨어 툴들의 성능도 개선되리라 기대하며, 이러한 연구를 토대로 단백질 신규성과 기능을 결정하는 과정을 간단하고 정교하게 수행할 수 있다[6].

Ⅱ. 관련연구

2.1 기존의 PMF 소프트웨어 툴의 특성 조사[6, 7, 8, 9, 10, 11]

PMF를 위한 각 소프트웨어 툴들의 특성을 조사하고 각 툴에서 스코어링을 계산할 때 영향을 미치는 매개변수들과 스코어링 방법을 분석한다.

가) PMF를 위한 툴들은 다음과 같다.

- PeptideSearch, - PeptIdent
- Mowse, - Mascot
- MS-Fit, - ProFound

나) 일반적으로 위에서 제시한 PMF 툴들에서 사용하는 매개변수들은 다음과 같다.

- Protein pI range
  : 단백질의 Isoelectric point
- Protein mass range
  : 단백질의 질량(kDa)
- Methionine
  : Methionine 의 산화여부
- Cysteine

: 샘플준비동안 Cysteine에 의한 단백질 변형 여부

- Database

: 검색을 위해 사용되는 데이터베이스

- Taxonomy

: 검색될 종

- Missed cleavage

: 허용되는 Missed cleavage의 최대 수

- Required number of peptides

: 단백질 동정을 위해 매칭되어야 하는 펩타이드의 최대 수

- Peptide charge state

: 질량분석동안 펩타이드의 양성화또는 중성화 여부

- Cleavage agent

: 펩타이드를 생성하는데 사용되는 효소나 시약

- Peptide mass tolerance

: 펩타이드 질량을 비교하는데 있어서 허용되는 질량의 오류범위

- List of peptide masses

: 알려지지 않은 단백질로부터 생성된 펩타이드들에 대하여 실험적으로 측정된 질량

- Report hits

: 검색결과로 보여지는 단백질의 리스트 수

다) PMF 툴들에서 검색하는 데이터베이스는 다음과 같다.

- NCBInr, - Swiss-Prot, - OWL

다음 표 2.1은 위에서 제시한 각 PMF 툴들에 대하여 각 툴에서 사용한 스코어링 방법, 데이터베이스, 제공되는 검색필드에 관한 비교분석을 보인다.

표 1. PMF툴 비교

Table. 1 Comparison of PMF tools

PMF툴	스코어링 방법	데이터 베이스	검색필드	비교
PeptideSearch	Count the number of matching masses	Non-redundant protein db(nrdb	Prot_MR, No_Pept, CA, MT, List_PM, No_hit	
PeptIdent	Count the number of matching masses	Swiss Prot, TrEMBL	PI, Prot_MR, Taxo, No_Pept, CA, MT, List_PM, No_hit	Taking into account known protein modification
Mowse	Frequency of each peptide mass	OWL	Prot_MR, CA, MT, List_PM	
Mascot	Probability-based scoring(mowse scoring)	OWL, NCBInr, dbEST, Random	Prot_MR, DB, Taxo, CA, List_PM, No_hit	
MS-Fit	mowse scoring	NCBInr, Swiss Prot, OWL, Genpept, Ludwignr, pdbEST.human	PI, Prot_MR, DB, Taxo, No_Pept, CA, List_PM, No_hit	
ProFound	Use a Bayesian probability-based scoring	NCBInr	Prot_MR, Taxo, No_Pept, CA, List_PM,	Use Tandem mass specctrometry (MS/MS) data

Ⅲ. Mowse 분석[6]

Mowse에서 단백질 동정을 하기 위해 사용되는 데이터베이스와 데이터베이스 구조를 조사하고 스코어링 하기 위해 사용되는 매개변수들과 에서 스코어링 하기 위해 사용되는 빈발 요소 행렬(Frequency factor matrix), Mowse 요소 행렬(Mowse factor matrix), 최종 스코어링 하는 방식을 분석한다.

3.1 Mowse의 source 데이터베이스

Mowse의 source 데이터베이스는 OWL 비중복 복합 서열 데이터베이스(non-redundant composite sequence database)로 이러한 OWL 데이터베이스는 4개의 공적으로 접근가능한 데이터베이스인 SWISS-PROT, PIR(1-3), GenBank(translation)과 NRD-3D로부터 유지된다. Mowse 의 source 데이터베이스의 첫 번째 인터넷 버전은 61,000 단백질이 저장되어 있으며, 이러한 단백질로부터 대략 15,000,000 펩타이드들이 생성되어 있다. 이러한 Mowse 펩타이드 데이터베이스는 OWL 데이터베이스의 새로운 버전에 따라 2달 만에 갱신되고 있다. SWISS-PROT은 Swiss Institute of Bioinformatics에서 제공하는 데이터베이스로 단백질의 기능, 구조 정보 등 단백질별로 상세한 정보를 가지고 있는 단백질 서열 데이터베이스이다. 또한 TrEMBL을 사용하여 중복체크, 자동화된 주석 처리, 전문가 분석을 통한 신뢰성있는 정보 제공이 특징이다.

3.2 데이터베이스의 구조

Mowse 데이터베이스는 다음과 같이 세 개의 이진파

일로 구성되어 있다.

- 펩타이드 질량 파일(MW 파일)

: 이 파일에는 각 펩타이드들에 대한 질량값이 저장되어 있으며, 저장되는 순서는 OWL 데이터베이스에 저장되어 있는 단백질의 순서와 일치한다.

- OWL 엔트리 인덱스 (MDX 파일)

: MW 파일에 저장되어 있는 질량값에 대하여 관련되어 있는 OWL 식별자 코드들에 대한 인덱스들이 저장되어 있다.

- 전체 서열 질량 파일(SMW 파일)

: 원래의 OWL서열에 대하여 계산된 질량이 저장되어 있으며, 이러한 질량 값들은 MW 파일에 있는 각 단백질들에 대한 엔트리와 같은 순서로 저장되어 있다.

3.3 PMF를 위해 사용되는 Mowse 매개변수

현재 사용되고 있는 매개변수 외 다른 매개변수의 사용이 스코어링의 정확성을 높일 수 있는지 분석한다.

- Protein mass range
- Cleavage agent
- Peptide mass tolerance
- List of peptide masses

3.4 스코어링에 사용되는 빈발 요소 행렬

빈발 요소 행렬 F는 다음과 같으며 각 열의 간격은 단백질 질량 10kDa 기준으로, 각 행의 간격은 펩타이드 질량 100Da기준으로 나누어져서 구성된다.

$$F = \begin{pmatrix} f_{1,1} \\ fi,j \end{pmatrix} \quad (1)$$

행렬의 각 원소는 각 단백질 서열 엔트리가 처리됨에 따라 펩타이드 질량의 빈발횟수에 따른 값이 증가한다. 다음 단계로 각 원소는 각 열에 대하여 각 열의 원소 개수 중 가장 큰 값으로 나누어짐으로써 정규화한다. 결과적으로 다음과 같이 Mowse 요소 행렬의 각 원소 $m_{i,j}$가 정해진다.

$$m_{i,j} = \frac{f_{i,j}}{\left| f_{i,j} \right|_{\max \ in \ j}} \quad (2)$$

그리고, 각 엔트리에 대한 스코어 값은 다음과 같이 계산된다.

$$Score = \frac{50000}{M_{Prot} \times \prod_{n} m_{i,j}} \quad (3)$$

이때 M_{prot} 은 단백질 엔트리의 질량을 의미하며, n은 실험에 의한 펩타이드와 데이터베이스의 이론적 질량이 일치하는 펩타이드들에 대응되는 Mowse 요소 행렬의 값들을 의미한다. 질량이 큰 단백질(>200kDa)의 스코어가 증가되는 것을 방지하게 위해 단백질 전체의 질량의 평균값인 50kDa을 의미한다.

Ⅳ. 스코어링 방법 개선 및 성능평가

4.1 Mowse 스코어링 방법 개선

가) 빈발 요소 행렬 생성에 대한 새로운 방법 제안

Mowse에서 빈반 요소 행렬은 단백질과 펩타이드 질량에 대하여 일정한 간격으로 생성되어 행렬의 각 원소의 값은 펩타이드의 빈발횟수에 따라 계산된다. 현재 이러한 행렬을 생성하는데 있어서 정해진 간격으로 생성되는데 이러한 간격의 값이 작아질수록 스코어링 값은 정확해진다. 그러나 이러한 간격의 값이 작아질수록 행렬의 크기는 증가하게 되며 이에 따라 스코어링 계산의 복잡도도 증가하게 된다. 그러므로 행렬의 크기를 현재와 같이 유지하면서 스코어링 값을 정확하게 계산하기 위한 새로운 방법이 필요하다. 그러므로 현재 Mowse에서 검색대상이 되는 단백질 데이터베이스의 분포를 고려하여 비선형적으로 행렬의 간격의 값을 정하는 방법 즉, 임의의 단백질 질량 값이 많은 곳에서는 행렬의 간격을 작게 결정하는 반면 단백질 질량 값이 적은 곳에서는 행렬의 간격을 크게 결정하였다.

나) 스코어링 계산시 매개변수의 단계적 적용

Mowse의 빈발 요소 행렬은 단백질과 펩타이드의 각 질량에 의존한다. 스코어링을 계산하는데 있어서 다른 매개변수 즉, 예를 들어 pI 값 등을 통해서 먼저 검색의 범위를 줄일 수 있도록 단계적으로 매개변수들을 적용하였다.

다) 단백질과 펩타이드 데이터베이스에 대하여 질량값에 따른 분포 제공

빈발 요소 행렬을 비선형적인 간격으로 생성하기 위해 검색 대상이 되는 OWL 데이터베이스의 특성을 파악하고 데이터베이스의 분포를 고려하는데, 이러한 분포

를 검사하는 방법과 행렬 생성시 간격을 결정하는 방법을 형식화하였다.

4.2. 개선된 Mowse 스코어링 방법의 성능 비교

개선된 Mowse 스코어링 결과는 다음과 같이 입력으로 펩타이드 질량이 주어졌을때 단백질 동정의 결과는 아래와 같다.

```
Using data fragments of:
    1086.2, 1399.6, 2030.2, 2424.7, 2930.3
    3086.3, 5423.0, 6082.8

1 100K_RAT 100 KDA PROTEIN (EC 6.3.2.-).
2 POLG_MCFA GENOME POLYPROTEIN [CONTAINS: CAPSID
PROTEIN C (CORE PROTEIN);
3 PGCV_HUMAN VERSICAN CORE PROTEIN PRECURSOR
(LARGE FIBROBLAST PROTEOGLYCAN
4 POL1_BAYMJ GENOME POLYPROTEIN 1 [CONTAINS:
CYTOPLASMIC INCLUSION PROTEIN
5 RRPB_CVMJH RNA-DIRECTED RNA POLYMERASE (EC 2.7.7.48)
(ORF1B).
```

또한, 실험결과 예측되는 5가지 단백질 중 1과 2에 에 대한 스코어링 결과는 다음과 같다.

```
1 : 100K_RAT 1.277e+06 100368.6 0.750
    100 KDA PROTEIN (EC 6.3.2.-).
    Mw Start End Seq
    1086.3 358 367 CATTPMAVHR
    1399.6 6 17 GDFLNYALSLMR
    2424.7 290 312 VFMEDVGAEPGSILTELGGFEVK
    2930.3 671 698 QLILASQSSDADAVFSAMDLAFAVDLCK
    3086.3 458 485 QLSIDTRPFRPASEGNPSDDPDPLPAHR
    *6082.8 817 870
  QDLVYFWTSSPSLPASEEGFQPMPSITIRPPDDQHLPTANTCISR...
    No Match 2030.2 5423.0
```

```
2 : POLG_MCFA 2.612e+05 373265.6 0.500
    GENOME POLYPROTEIN [CONTAINS: CAPSID PROTEIN C (CORE
PROTEIN); MATRIX PROTEIN (ENVELOPE PROTEIN M); MAJOR ENVELOPE
PROTEIN E; NONSTRUCTURAL PROTEINS NS1, NS2A, NS2B, NS4A AND NS4B;
HELICASE (NS3); RNA-DIRECTED RNA POLYMERASE (EC 2.7.7.48) (NS5)].
    Mw Start End Seq
    2423.7 3233 3251 TSWSVHQYHEWMTTDDMLR
    2931.2 951 977 EYTPDTLSDPSDQALFIPPAWGGPISR
    *3088.7 2174 2200 SYMDSDLVKWVILGSCLICGVLAWEMR
    6084.9 1327 1382
  AHQPTVAAVLAFTMVVLFLYMEQTNVSMELEFISAGETPEGVSTE.
    No Match 1086.2 1399.6 2030.2 5423.0
```

개선된 Mowse 스코어링 방법에 관하여 성능평가를 하기 위하여 기존의 Mowse 툴과 개선된 Mowse 툴을 비교 분석한 결과는 그림 1과 같다.

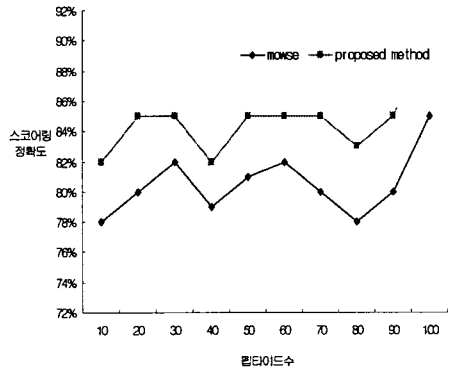


그림 1. 성능평가
Fig. 1 Performance Results

단백질 질량 값이 많은 곳에서는 행렬의 간격을 작게 결정하는 반면 단백질 질량 값이 적은 곳에서는 행렬의 간격을 크게 결정하였다. 그 결과 기존의 Mowse 방법에 비해 스코어링 계산속도가 향상되어 정확하고 빠르게 스코어링 결과를 볼 수 있었다. 고려할 사항은 행렬의 생성은 데이터베이스의 갱신이 어느정도 이루어질 경우 주기적으로 이루어져야 한다. 그러므로 데이터베이스 관리자는 데이터베이스 갱신 상황을 파악하고 주기적으로 행렬이 생성될 수 있도록 해야 한다.

V. 결 론

본 논문에서는 단백질 동정을 위한 알고리즘 중의 하나인 Mowse 스코어링 방법을 개선하고, 기존의 알고리즘과의 비교분석을 실시하였다. 결과적으로 제안한 알고리즘을 적용한 결과 기존의 스코어링 방법에 비해 성능이 향상됨을 보였다. 단백질 동정을 위해 사용하는 Mowse의 스코어링 방법은 다른 PMF알고리즘에서도 적용하여 사용하고 있는 만큼 Mowse 스코어링 방법을 개선하고자 하는 연구는 매우 중요하다고 할 수 있다. Mowse 스코어링 방법의 성능 개선을 통하여 단백질의 동정을 보다 빠른 시간 내에 수행할 수 있으며, 보다 정확하게 수행할 수 있고, Mowse 스코어링 방법을 적용하

여 사용하고 있는 다른 PMF 소프트웨어 툴들의 성능 개선에 활용할 수 있다. 결과적으로, 이러한 연구를 토대로 단백질 신규성과 기능을 결정하는 과정을 간단하고 정교하게 수행할 수 있으며, 프로테오믹스를 이용한 신약개발 등 관련 기술의 산업적 유용성에 있어서, 약물의 직접 타깃이 되는 단백질을 연구함으로써 신약개발 가능성을 한층 높이고, 이렇게 개발된 신약이 고부가가치를 창출할 수 있을 것이다.

향후 연구로는 Mowse 스코어링 방법을 적용하여 사용하고 있는 다른 PMF 소프트웨어 툴에 본 논문에서 제안한 알고리즘을 적용할 것이다.

참고문헌

[1] S. R. Penningron and M. J. Dunn, Proteomics From Protein Sequence To Function, 2001. Springer-Verlag.

[2] M. R. Wilkins, K. L. Williams, R. D. Appel, and D. F. Hochstrasser, Proteom Research: New Frontiers in Fuctional Geomics, Springer-Verlag.

[3] G. Kris and Joel Vandekerckhove, "Protein identification methods in proteomics," Electrophoresis, 21, 2000, pp.1145-1154.

[4] C. L. Daniel, Introduction to Proteomics, Humana Press.

[5] Tang, C., Zhang, W., Fenyo, D., and Chait, B. T., "Assessing the performance of Different Protein Identification Algorithms," 48th ASMS Conference, June 11-15, 2000.

[6] Pappin, D. J. C., Hojrup, P., and Bleasby., "Rapid Identification of Proteins by Pepide-Mass Fingerprinting," Current Biology, 3, 1993, PP.327-332.

[7] Perkins, D. N., Pappin, D. J. C., Creasy, D. M. and Cottrell, J. S., "Probability-based protein identification by searching sequence databases using mass spectrometry data," Electrophoresis, 20, 1999, pp.3551-3567.

[8] Zhang, W. and Chait, B. T., "ProFound - An expert system for protein identification using mass spectrometric peptide mapping information," Anal. Chem., 72(11), 2000, pp.2482-2489.

[9] P. R. Baker and K. R. Clauser, 1995. http://prospector.ucsf.edu.

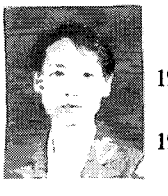
[10] M. R. Wilkins, E. Gasteiger, A. Bairoch, J. C. Sanchez, K. L. Williams, R. D. Appel, and D. F. Hochstrasser, Protein Identification and Analysis Tools in the ExPASy Server in: 2-D Proteom Analysis Protocols, 1998, Humana Press.

[11] M. R. Wilkins, E. Gasteiger, C. Wheeler, I. Lindskog, J. C. Sanchez, A. Bairoch, R. D. Appel, M. D. Dunn, and D. F. Hochstrasser, "Multiple parameter cross-species protein identification using MultiIdent, Electrophoresis, 19(18), 1998, p.3199-3206.

저자소개

정 민 아(Min-A Jung)

1992년 전남대학교 전산통계학과(학사)
1994년 전남대학교 대학원 전산통계학과(이학석사)
2002년 전남대학교 대학원 전산통계학과(이학박사)
2002년 ~ 2003년 광주과학기술원 정보통신학과 Post-Doc.
2005년 ~ 현재 목포대학교 컴퓨터 교육과 교수
※관심분야 : 데이터마이닝, 생물정보학, 데이터베이스, 정보보호



김 치 연(Chi-yeon Kim)

1992년 2월 전남대학교 전산통계학과 (이학사)
1994년 2월 전남대학교 대학원 전산통계학과 (이학석사)
1999년 8월 전남대학교 대학원 전산통계학과 (이학박사)
2002년 3월 ~ 현재 목포해양대학교 해양전자통신공학부 교수
※ 관심분야 : 이동 컴퓨팅, 전자상거래, 데이터마이닝