2003 정보 및 제어 학술회의 (2003.11.21-22)

HABIT : 질병 진단 시스템[1)]

HABIT : Cancer Diagnosis System

김 기 성*, 온 승 엽**, 강 경 남***

*항공대학교 컴퓨터 공학과: (E-mail : expkks@mail.hau.ac.kr)
** 항공대학교 컴퓨터 공학과 :(전화:(02)300-0146(교환 86), E-mail : syohn@mail.hau.ac.kr)
*** 항공대학교 컴퓨터 공학과: (E-mail : knkang@mial.hau.ac.kr)

Abstract : In this paper we proposes a new technique for identification of breast cancer by classification of proteome pattern generated from 2-D polyacrylamide gel electrophoresis (2-D PAGE) and development of cancer diagnosis system : HABIT.

Proteome patterns reflect the underlying pathological state of a human organ and it is believed that the anomalies or diseases of human organs are identified by the analysis or classification of the patterns. Proteome patterns consist of quantitative information of the spots such as their size, position, and density in the proteome image produced from 2-D PAGE, for the image mining of proteome pattern, SVM(support vector machine) and GA(genetic algorithm) are used to generate a decision model for the identification of breast cancer. The decision model was then used to classify an independent set of test proteome patterns into the affecter and unaffecter classes. The proposed technique was tested by actual clinical test samples and showed a good performance of a hit ratio of 90%

Keywords : Bioinformatics, Proteome, Cancer, SVM, GA,

Ⅰ. 서론

본 논문에서는 프로테옴(Proteome) 패턴을 이용한 질병 조기 진단 시스템을 제안하고 이 시스템이 질병 조기 진단에 효율적으로 응용될 수 있음을 보이고자 한다. 바이오 인포매틱스(Bioinformatics) 는 바이오 기술(BT)와 정보 기술(IT)분야의 학제간 연구를 통해 방대한 양의 데이터를 빠르고 효율적으로 가공하고 처리하는 기술의 하나로 생명체가 지니고 있는 광대한 양의 정보를 수집, 저장, 분석하고 그 결과를 제약, 식품, 농업, 환경 등의 분야에 이용하여 부가가치를 창출하는 기술이라고 할 수 있다.

단백체학(Proteomics)은 생명체의 전체 유전자, 즉 유전체(Genome)에 의해 발현되는 모든 단백질들의 총합을 일컫는 프로테옴을 다루는 학문으로, 어떤 단백질이, 얼마의 양으로 어떤 환경에서 발현되는가를 파악하는 것을 목적으로 한다. 생명체의 유전체는 모든 세포에서 동일한 형태로 존재하며, 생명체가 수행하는 기능의 이론적인 면만을 제시할 수 있음에 반해, 프로테옴은 세포가 처해 있는 환경, 조직별로 유동적으로 존재하며, 세포의 실질적인 기능을 표현해 준다. 이러한 이유로 급속한 속도로 밝혀지고 있는 미지의 유전자들의 기능을 밝혀내고자 하는 Functional genomics의 한 부분으로 새롭게 부각되고 있다.

Functional genomics는 세포 내에서 일어나는 실제적인 현상들을 전체 단백질 단계에서 통합적으로 파악하는 수단을 제공하기에 특정 질병의 발병 기작과 질병의 종류에 따른 증상의 상호 연관성을 단백질 단계에서 밝혀낼 수 있다. 이는 질병의 조기 진단, 발현 예측에 응용될 수 있는데 특히 암과 같은 질병의 경우 조기에 발견할 수 있다면 높은 치료율을 기대할 수 있다.

현재 질병을 조기에 발견하기 위한 시료로써 눈물, 타액, 혈청등이 사용되고 있는데 이중에서 혈청 단백질(Serum Proteome)을 이용한 연구가 가장 활발하다.[1][2] 혈청은 확보가 용이하고 일상적인 임상검사에서 분석되는 시료임으로 폐암과 같은 조기진단이나 예후 판정이 어려운 질환에서 우선적으로 고려되는 최적의 시료라 할 수 있다.

혈청 단백질의 성분분석 방법은 HC(Hierarchical clustering), K-means, SOM(Self-Organizing Map), PCA(Principal component analysis)등이 응용되어 왔으며, 최근에 들어서는 SVM(Support Vector Machine)과 같은 새로운 기계학습 기법이 사용되고 있다. 현재 이러한 분석방법을 지원하는 범용 분석 소프트웨어들은 조기 질병 진단이라는 목적에 특화된 것이 아니기에 사용의 편의성이 낮고 결과분석이 쉽지 않다.

1) 본 논문은 과학 기술부, 한국과학 재단 지정 경기도 지역협력연구센터(RRC)인 한국항공대학교 인터넷정보검색연구센터의 지원에 의한 것임

HABIT의 시스템의 개발목적은 프로테옴 분석을 통한 질병 조기 진단 시스템의 개발에 있다. 위 시스템의 개발에서 얻는 장점은 조기 진단 및 질병 발현 예측, 진단 소요시간, 비용, 인력의 절약, 유전자와 유전자 외적 요인에 의한 현상추적의 용이화, 정상 조직과 질병 조직간 단백질 발현의 차이정보를 획득할 수 있다는 점이다. 본 논문에서는 앞서 소개한 여러 가지 기계 학습 기법을 통합한 암 조기 진단 시스템의 혈청 단백질 데이터 분석 프로그램을 개발한 과정과 알고리즘, 프로그램을 통해서 수행한 결과를 제시하고, 결론적으로 이 프로그램이 혈청 데이터 분석에 효율적으로 응용될 수 있음을 보이고자 한다.

II. 프로테옴을 이용한 질병 진단 과정

1. 프로테옴

프로테옴이란 프로테인(Protein) 과 게놈(Genome)의 합성어로 특정 조건과 특정 시점에 한 시료(조직, 기관, 세포, Body Fluid) 가 갖는 단백질의 총체를 말하며 이를 연구하는 학문을 단백체학(Proteomics) 이라 한다. 일반적으로 게놈에 의해 mRNA로 전사가 되면 그 전사가 된 것을 리보솜 안에서 단백질로 번역해 낸다. 하지만 게놈의 저오는 전사과정에서 intron 이라는 비 정보영역과 exon이라는 정보영역을 선택적으로 취하게 되는데 이것을 다시 번역과정에서 변형시켜 최종적으로 단백질로 발현된다. 따라서 정확한 유전자 배열을 안다고 해도 실제 세포내에서 기능을 담당하는 단백하고는 차이가 존재하게 된다. 또한 유전자 발현의 최종 산물인 프로테옴은 게놈에 비해서 훨씬 다양하고 복잡성을 띄므로 전체의 프로테옴을 관찰하고 연구한다는 것은 매우 어려운 일로 여겨졌다. 이런 연유로 많은 바이오 연구 프로젝트들이 프로테옴 보다는 유전자 염기서열 분석에 집중되어 왔는데 염기서열 분석만으로는 단백질의 발현을 예측할 수 없고, 단백질이 변형된 것은 밝혀낼 방법이 없으므로 기존의 연구로는 질병의 원인, 진행에 대한 연구가 한계에 도달하게 되어 프로테옴 연구가 활성화되게 되었다.

프로테옴 분석은 다음과 같은 특성을 가지고 있다.

(1) 정제 과정 없이 조직, 개체등 시료에 존재하는 모든 단백질을 펼쳐 분석가능

(2) 유전자의 발현 정보를 한 번에 확인할 수 있다.

(3) 유전자에 의한 현상과 유전자 외적 요인에 의한 현상추적이 용이하다.

(4) 정상조직과 질병조직, 그리고 좋은 품종과 나쁜 품종간 단백질 발현의 차이를 알 수 있다.

현재 많은 게놈연구 프로젝트들이 프로테옴 연구 프로젝트로 전환하고 있으며 [3][4] 표 1은 게놈과 프로테옴을 비교한 것이다.

	게놈	프로테옴
정 의	유전정보의 총체	단백질의 총체
특 성	개체마다 유일	한 개체에 수없이 많음
목 적	생명체 설계도의 해독	기능단위 총체적 해독
용 도	연구, 산업의 기본 Data	산업과 직결된 Data

표 1 게놈과 프로테옴의 비교

2. 프로테옴 이미지 생성

프로테옴 분석과정의 워크 플로우는 다음과 같다.

(1) Sample preparation
(2) 2-D electrophoresis
(3) Staining
(4) Proteome Image Acquisition
(5) Proteome Image Analysis
(6) Protein excision
(7) Proteion Identification

프로테옴 이미지 생성은 워크 플로우의 (1)~(4) 과정을 통해 생성된다. 프로테옴 이미지의 일반적 제작 과정은 먼저 길이 18cm broad range IPG strip (pH3-10, linear, nonlinear, pH 4~7, pH6~9)를 사용하거나 혹은 1 unit ranges (pH4.5-5.5 / pH5-6등등) IPG strip을 사용하여 시료 내에 존재하는 단백질을 각각의 등정점(PI)에 따라 2차원 전기 영동법에 의해 분리한다.. 그 다음으로 단백질 크기에 따른 SDS-PAGE를 이용한 분리과정인데 등정점에 의한 단백질의 분리 후 사용자가 요구하는 폴리아크릴 아미드의 농도에 따라 구배 젤(통상 8- 16%)을 제조하여 크기에 따라 각각의 단백질을 분리한다. 마지막으로 젤을 발색하는데 전기 영동법으로 분리된 단백질의 image는 Colloidal Coomassie blue (CBB G250, R250), Sypro-Ruby, MALDI compatible silver staining 법등을 이용해 발색을 한다. 다음 그림 1은 위와 같은 과정을 통해 얻은 프로테옴 이미지의 샘플이다.

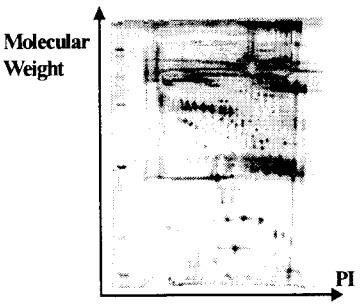

그림 1 2-D 프로테옴 이미지
Fig 1 Proteom Image

획득된 프로테옴 이미지는 GS-710 Calibrated Imaging Densitometer, Molecular Imager FX등과 같

은 스캐너로 디지털화 하여 컴퓨터에 저장한다.

3. PDQuest에 의한 Image Analysis

저장된 프로테옴 이미지는 프로테옴 이미지 분석 소프트웨어인 PDQuest[5] 이용하여 전 처리 작업, 스팟 검출(Spot Detection), 피쳐 추출(Feature Extraction) 작업을 수행한다. 이미지의 전 처리 작업은 Image Sizing 과 Orientation, 필터링등을 통해 조정한다. 그림 2는 스팟 검출과정의 그림이다. 최초 획득 이미지의 스팟은 사용자가 직접 선택하거나 PDQuest에서 제공되는 자동 모드로 검출 한다. 사용자는 검출된 스팟을 라벨링 하고 마스터로 저장한다.

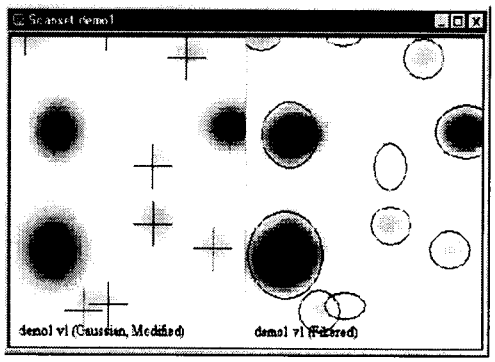

그림 2 스팟 검출

Fig 2 Spot Detecting

그림 3은 PDQuest내에서 마스터와 각 프로테옴 이미지들 간의 실험구성을 보여주는 것으로 마스터에서 검출되고 라벨링된 스팟들은 매치셋(Matchset)이라고 하며 스팟을 검출하는 기준이 된다. 새로 입력된 이미지는 매치셋에서 가장 근접한 위치의 스팟에 같은 번호로 라벨링 하고 각 스팟들은 Center X, Center Y, Size X, Size Y, Peak Val, Quantity, Norm_Qty, Quality 필드로 엑셀파일로 저장한다.

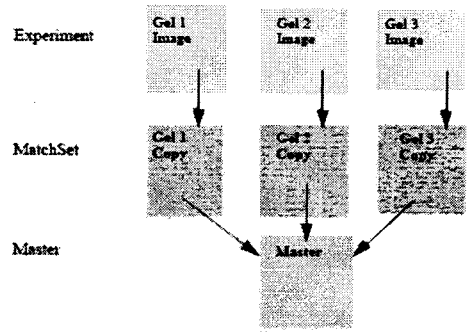

그림 3 매치셋과 실험 이미지의 실험구성

Fig 3 Matchset, Image Experiment

각 필드의 상세내용은 다음과 같다. Center X, Y는 스팟의 절대위치 좌표이다. Size X, Y는 원점에서 X, Y 축 방향으로의 스팟 크기이다. Peak Val는 스팟내에서 Intensity값의 최대값을 말하며, Qunatity는 검출 된 스팟의 전체 intensity값으로 젤 이미지의 특정 스팟의 단백질 양을 말한다. Norm_Qty는 Quantity의 값을 노말라이즈 한 값으로 산출 공식은 Raw Spot Quantity * Scaling Factor로 계산되어 지며 Scaling Factor는 상수값으로써 사용자가 보통 직접 지정하여 사용한다. Quality는 0~100까지의 값을 가지며 갓 스팟의 다음의 속성을 기반으로 계산되어 진다.

(1) 가우시안 : 스팟이 가우시안 모델과 얼마나 일치하는지에 따라 결정되어진다.

(2) X, Y Streaking : X, Y축 방향으로 젤 스트리킹이 얼마나 영향을 미쳤는지 결정한다.

(3) 오버랩(Overlap) : 스팟끼리 서로 겹친 정도

(4) Linear range of Scanner : 스팟의 최대 인텐시티 값이 스캐너의 선형범위 내에 위치하는 정도

만약 스팟이 가우시안 모델에 완벽하게 맞고 스트리킹, 오버랩이 없고 스캐너의 선형 범위 내에 존재하면 그 퀄리티는 100으로 계산된다.

III. 클래시피케이션 및 키 피쳐 추출

1. SVM

SVM은 기본적으로 2개의 클래스를 갖는 객체들을 분리해내는 방법이다. 그림 4를 보기 되면 클래스를 구분하는 하이퍼 플레인(Hyper Plane)을 볼 수가 있는데 이런 하이퍼 플레인은 무수히 많개 존재할 수 있다. 이중에서 양 측의 트레이닝 샘플을 최대한 가까이 위치시키면서 마진(Margin)을 최대화 하는 최적 분류면(Optimal Plane)을 찾는 문제이다.

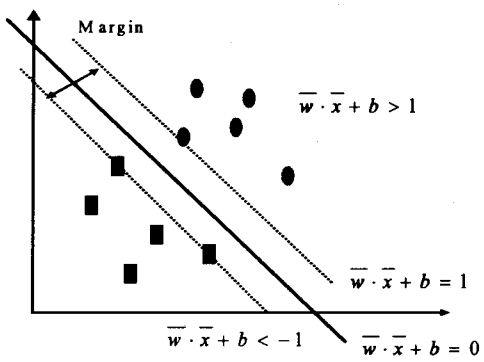

그림 4 최적 분류면

Fig 4 Optimal Plane

Objective Functions

$$\text{Min}\ \frac{w'w}{2} \qquad (1)$$

Subject to

$$y_i\left(w'x_i+b\right)\geq 1 \qquad (2)$$

라그랑지 함수로 계산하면 원 문제를 얻고 여기에

KKT(Karush-Kuhn-Tucker)조건을 적용시키고 쌍대문제(Dual Problem)을 해결하면 다음의 관계를 얻는다.

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \qquad (3)$$

Decision Functions은

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i x_i \cdot x + b = 0 \qquad (4)$$

본 실험에서는 $f(x) < 0$ 일 경우 암 환자로 $f(x) > 0$ 일 경우 정상인으로 분류하였다. [6][7]

2. GA(Genetic Algorithm)

유전자 알고리즘(GA)은 자연계 있어서 생물이 갖는 환경에서의 적응능력을 취급하는 것으로 생물의 유전과 진화의 메커니즘을 공학적으로 모델화한 것이다. 본 연구에서는 GA Search 기법(Probabilistic searching)을 이용하여 최적의 피처 스팟 셋을 찾기 위해 사용되었다. 그림 5는 본 실험에서 사용된 GA Cycle이다. 유전자 알고리즘은 트레이닝 데이터내의 많은 프로테옴 패턴들 중에 랜덤하게 선택된 스팟 셋으로 SVM을 통해 클래시피케이션 하고 나온 결과에 대해서 적합도(fitness)테스트를 한 후 GA Operator를 적용시켜 새로운 세대를 생성해낸다 이 과정이 1세대이며 100세대 정도 반복하면서 최고의 적중률을 보인 패턴의 스팟을 추출해 내어 키 피처 스팟 셋으로 선정하고 이를 알려지지 않은 샘플에 적용하여 테스트한다.

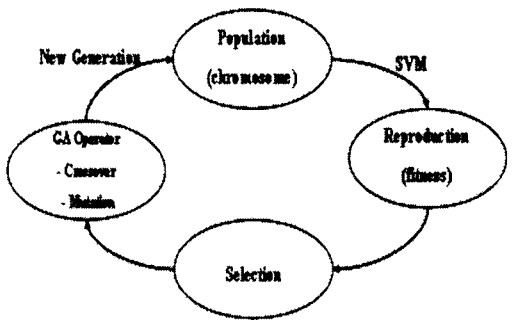

그림 5 GA Cycle
Fig 5 GA Cycle

IV HABIT 시스템 개념 및 이용

1. HABIT 시스템 개념

본 시스템의 주요 개발 목적은 프로테옴 패턴을 이용한 질병 진단 과정에서 데이터 셋의 선택, 분류, 키 피처 추출과정을 자동하는데 있다. 그림 6은 조기 질병 진단 시스템의 전체 개념도이다.

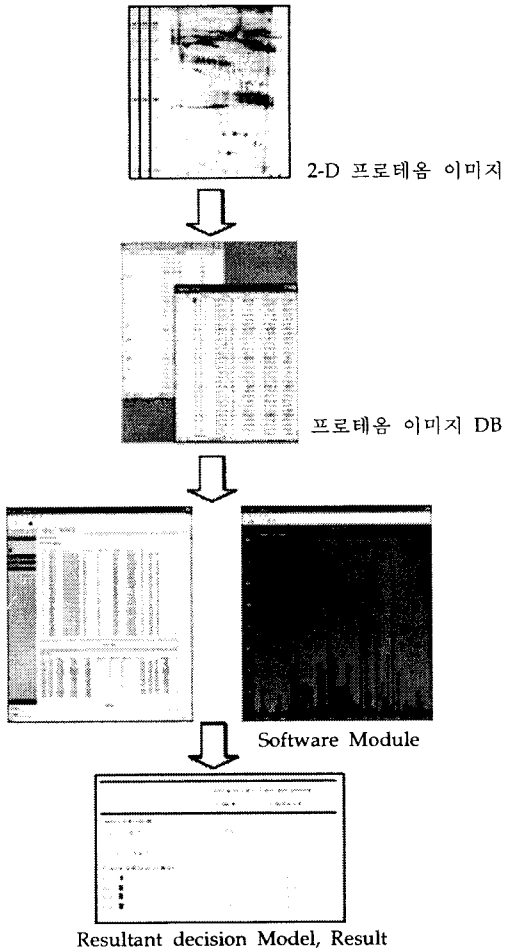

그림 6 조기 질병 진단 시스템 개념도
Fig 6 Cancer Diagnosis System

HABIT 시스템의 전체 구성은 프로테옴 이미지 DB와 Software 모듈로 크게 나눌 수가 있다. 그림 7은 HABIT 시스템의 소프트웨어 아키텍처이다.

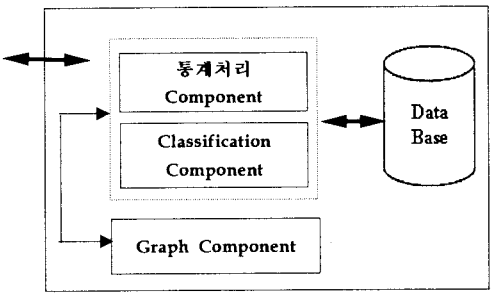

그림 7 HABIT 시스템 소프트웨어 아키텍처
Fig 7 HABIT System Software Architecture

각 컴포넌트의 기능은 다음과 같다.

(1) 데이터 셀렉션 Component

데이터를 DB로부터 셀렉션 하는 기능을 가진 컴포넌트로써 암의 종류, 진행도, 환자의 성별, 나이대, 스팟별 조건으로 데이터 셋을 셀렉션 하거나 랜덤하게 DB에서 셀렉션 하는 기능을 수행한다.

(2) 통계처리 Component

통계처리 기능을 수행하는 컴포넌트로서 사용자에 의해 선택된 데이터의 평균, 분산, 표준편차, Correlation, Covariance, 표준편차의 상관관계를 계산하여 수치로서 사용자에게 제공한다.

(3) Classification Component

Classification 기능을 수행하며 현재 Classifier로 SVM을 지원하고 있다.

(4) Graph Component

통계처리, Classification의 결과를 2-D 그래프로 시각화 하거나 특정 환자의 젤 이미지를 3차원 그래프로 표현, 다른 환자, 정상인 데이터와의 비교, 도시할 수 있다.

(5) HABIT DB

암 환자와 정상인의 혈청 이미지 데이터와 병력기록을 저장한 DB로서 Spot data table, 병력 테이블로 구성되어 있으며 현재 시스템에서는 유방암 환자와 정상인 전체 345명의 병력 데이터와 스팟 67개의 데이터를 가지고 DB를 구성했으며 DBMS는 MS사의 Access를 사용하였다.

2. 유방암의 시험 결과

HABIT을 사용하여 실험한 결과를 표 2에 도시했다. 유방암 환자 63명, 정상인 65명의 데이터를 입력으로 트레이닝 셋에 환자 30, 정상인 30, 전체 스팟을 테스트셋에는 환자 33, 정상인 35명을 스팟은 GA를 통해 추출한 키 피처 스팟 셋 28개를 사용하였다.

실험 1			
범 주	환자	정상인	스팟
Total	63명	65명	67개
Training	30명	30명	
Test	33명	35명	
결 과			
전체	Sensitivity	Specificity	스팟
90.0%	100%	88.57%	28개

표 2 유방암을 이용한 실험결과

V 결 론

개발된 질병 조기 진단 시스템의 분석 도구 : HABIT은 현재 질병진단에 특화된 통계적 분석, 클래시피케이션 분석법을 제공하며 그 결과를 수치나 그래프로써 보여준다.

클래시피케이션의 결과는 실험 데이터의 선택 조건을 변경시킴으로서 그 결과가 크게 달라진다. 기존의 이러한 분석방법을 제공하는 프로그램들은 일반적인 연구목적에 활용되도록 만들어짐으로써 실험 데이터의 선택이 제한적이다. SPSS나 Matlab과 같은 툴들은 실험 데이터의 Selection 작업이 수동적이어서 HABIT처럼 질병진단에 특화된 기능 제공이 없거나 미약하다. 향후 과제는 현재 개발된 질병 진단 툴 이 초기모델로 개발된 것이기에 사용자 요구사항에 대한 분석과 개념이 명확하지 않았다는 점이다. 차후 보완을 통해서 문제점을 개선하고 추가적인 클래시피케이션 기법을 제공할 예정이다. 본 시스템이 성공적으로 개발된다면 질병 진단 시스템으로써의 역할을 충분히 하리라 기대된다.

참고문헌

[1] E.F Petricoin, et, al,, "Use of Proteomic Patterns in Serum to Identify Ovarian Cancer", The Lancet, Vol. 359, pp.572~577, Feb, 2002

[2] I.Pucci-Minafia, et, al,, "Proteomic Patterns of Cultured Breast Cancer Cells and Epithelial Mammary Cells", New York Academy of Science, Vol., 963, pp. 122-139, 2002

[3] APAF Chiron Project

[4] Swiss institute of Bioinformatics 2-D polyacrylamide gel electrophoresis image DB

[5] *PDQuest User Guide for Version 7.0*, Bio-Rad Laboratories, 2001

[6] V. N. Vapnik, "The Nature of Statistical Learning Theory", Springer, 2000

[7]N.Cristianini and J.Shawe-Taylor, "An Introduction to Support Vector Machine and other kernel-based learning methods". Cambridge University Press, 2000