碩士學位論文

단백질 시퀀스를 이용한 PIS 시스템과 단백질체학 시스템 설계

濟州大學校 大學院

電算統計學科

梁　根　倬

2006年　12月

碩士學位論文

단백질 시퀀스를 이용한 PIS 시스템과 단백질체학 시스템 설계

濟州大學校 大學院

電算統計學科

梁 根 倬

2006年 12月

단백질 시퀀스를 이용한 PIS 시스템과 단백질체학 시스템 설계

指導敎授 李 鳳 奎

梁 根 倬

이 論文을 理學 碩士學位 論文으로 提出함

2006年 12月

梁根倬의 理學 碩士學位 論文을 認准함

審査委員長 _____ ㊞

委 員 _____ ㊞

委 員 _____ ㊞

濟州大學校 大學院

2006年 12月

목 차

그림 목차

표 목차

초록

단백질 정보분석은 데이터베이스 관련분야와 분석도구(analysis tools: analysis programs) 개발관련 등으로 나눌 수 있다. 하나의 단백질 정보를 컴퓨터의 저장 단위인 한 바이트(byte)에 저장하고 이를 컴퓨터에서 관리하고 분석하기 위해서는 데이터베이스 시스템의 개발과 관리가 기본적이다. 분석도구에 대한 연구는 단백질 서열을 대상으로 하여, 유전자의 기능을 유추하거나 서열간의 관계에 대한 정보를 추출하는 등, 서열로 표현되어 있는 생명체의 암호를 판독하는 과정에 관련된 전반적 분석 방법을 연구하고 이 방법을 표현하는 프로그램을 개발하는 것이다. 이 분야 이론에서는 생물학 서열의 특성을 반영하여 생물학 서열의 성질과 관계성을 수학적으로 정의하여 중진 정의에 의한 최적의 해답을 구하는 방법을 다루게 된다. 그 외에도 실험을 통해 얻은 DNA의 서열에서 단백질을 발현하는 유전자가 있는가를 예측 하는 것(gene prediction이라고 한다), 미지의 DNA나 단백질 서열과 비교하여 (homology search라고 한다) 그 미지 서열의 기능을 밝히는 것, 한 생물의 유전체를 다른 생물체의 유전체와 총체적으로 비교하는 것(comparative genomics)등과 같이 유전체 프로젝트의 종료와 함께 새롭게 개발되는 방법들도 많이 등장하고 있다.

본 논문에서는 미지의 단백질 서열을 데이터베이스 내의 서열과 비교하여 그 미지 서열의 기능을 밝히는 분야에 해당되는 부분서열을 이용하는 새로운 단백질 동정 시스템을 제안한다. 구현된 시스템은 입력되는 단백질의 부분서열을 기존의 데이터베이스내의 서열들과 비교하는 새로운 방법을 사용한다. 또한 기존의 질량에 의한 분석도 할 수 있도록 함으로써, 단백질 분석 기능을 향상시킨다.

I. 서론

생명체의 전체 유전자인 지놈(genome)에 의해 발현되는 모든 단백질의 총합인 프로테옴(Proteome)을 다루는 프로테오믹스(Proteomics)는 단백질을 대량으로 분석하고 이들의 상호기능관계 지도를 작성하며 구조분석을 통해 궁극적으로 특정 단백질과 이를 만드는 유전자의 기능을 동시에 밝혀내는 것을 목적으로 한다.

2001년 HGP(Human Genome Project)에 의해서 인간 유전체 염기서열이 다 밝혀져 있지만 그것만 가지고는 유전자 산물의 기능을 알 수가 없고, 이것이 전사(transcription)되어 단백질 생성 수준에서 조절된다고 하더라도 최종적으로 세포 내에서의 기능 여부는 얼마나 정교하고 적절하게 단백질 합성 후 변형되는가에 달려 있기 때문이다. 즉 최종적으로 완벽한 모양이 갖추어진 단백질을 분석하지 않고는 그 유전자의 세포 내 기능을 알 수 없는 것이다. 이런 의미에서 21세기 기술문명을 이끌어나갈 고부가가치, 고성장 산업으로 기대되며 다른 산업에 대한 파급 효과가 클 것으로 예상된다.

하나의 생물은 무수히 많은 수의 단백질로 구성되어 있다. 유전체는 하나의 세포 안에서 변화가 없이 일정하게 유지되는 반면 단백체는 조직의 특성, 상태, 건강과 같은 환경요인에 따라서 시시각각으로 변화한다. 따라서, DNA에서 단백질로 번역된 후에도 단백질 자체에서 다양화가 이뤄지며, 번역 후 단백질 변형은 현재 약 300여 가지 이상으로 알려져 있다. 이렇듯 Proteomics는 유전체 및 유전자에 관련된 방대한 양의 정보를 이용한 분석 및 실험이 필수적이기 때문에 생물정보학의 활용이 필수적이다. 즉 Proteomics에 관련된 정보를 컴퓨터를 이용하여 신속하게 저장/가공/처리하는 방법이 개발되어야 하며, 이런 소프트웨어는 현재 Proteomics 연구자들의 단백질 분석에 필수적으로 사용되고 있다.

Proteomics를 위한 생물정보학 연구 중 하나가 PIS(Protein Identification System)이다. 이것은 실험을 통해서 얻어진 새로운 단백질을 범주를 기존 단백질 정보를 이용하여 예측하는 방법으로 데이터베이스와 비교 시스템으로 구성된다. 즉 기존에 기능이 알려진 단백질들을 데이터베이스에 저장하고, 입력되는 새로운 단백질 정보와 데이터베이스내의 정보를 비교하여 가장 유사한 단백질 군으로 분류하는 방법이으로 Proteomics 관련 연구자들이 가장 많이 사용하는 생물정보학 시스템이다. 현재까지 이 분야에 대해서 많은 연구가 진행되고 있으며 대표적인 알고리즘으로는 MOWSE 등이 알려져 있다. 그러나 기존의 알고리즘의 경우 단백질의 분자량에만 의존한 분석 방법이기 때문에 부분적인 서열정보를 활용할 수 없다는 단점이 있다. 또한 분자량의 비교 방법에 있어서도 빈도수만을 이용하기 때문에 상이한 단백질과 매칭되는 단점을 가지고 있다.

본 논문에서는 부분서열을 이용하는 새로운 단백질 동정 시스템을 제안한다. 구현된 시스템은 입력되는 단백질의 부분서열을 기존의 데이터베이스내의 서열들과 비교하는 새로운 방법을 사용한다. 또한 기존의 질량에 의한 분석도 할 수 있도록 함으로써, 단백질 분석 기능을 향상시킨다. 시스템의 구현을 위해 SWISSPROT에 등록된 약 210,000개의 기능이 규명된 단편 단백질 정보를 자체 ORACLE DB로 구현하였다. 실험에 사용되는 테스트 데이터는 NCBI의 nr 데이터베이스내에 있는 2900여개의 단백질 정보를 이용하였다. 실험 방법은 먼저 nr 데이터베이스내의 모든 단백질 정보를 입력으로 하여 blast search를 수행하였다. 이때 index로는 SWISSPROT DB를 이용하였다. 이 실험에서 index에 가장 잘 매칭되는 nr 데이터베이스내의 단백질 2900개를 선택한 후, 이 선택된 서열을 제안된 시스템에 입력하여 결과를 얻었다. 얻어진 결과는 서로 비교하여 일치 여부를 확인하는 방법을 통하여 성능을 판단하였다.

본 논문의 구성은 다음과 같다. II 장에서는 Proteomics에 대한 기본적인 개념과 이에 관련된 생물정보학 여러 연구결과의 개용 및 장/단점을 알아본다. III장에서는 본 논문에서 제안하는 전체 시스템과 PIS관련 알고리즘에 대해서 다룬다. IV장에서는 실제 구현된 시스템에 서 얻어진 실험결과를 바탕으로 제안한 방법의 타당성을 검토한다. 마지막으로 V장에서는 본 논문에 대한 결론과 요약을 다룬다.

II. 배경 및 관련 연구 동향

1. Proteomics를 위한 중심이론 (Central Dogma)

DNA(DeoxyriboNucleic Acid)는 아데닌(Adenine), 시토신(Cytosine), 구아닌(Guanine), 티민(Thymine)의 네 가지 뉴클레오티드(Nucleotides)로 구성된 중합체이다. DNA는 보통 두 가닥의 형태로 나타나며 두 가닥은 이중나선 형태로 서로를 감싼다. 각 염기들은 서로 쌍을 이루고 있으며, 한 가닥 위에서 하나의 A는 항상 다른 가닥 위의 T와, G는 C와 쌍을 이루고 있다. 하나의 긴 서열로서의 유전체는 각각의 기능을 가진 유전자로 나뉘어진 후 작용하게 되는데, 각 유전자는 단백질 코딩, RNA 규정, 비전사의 기능을 담당하게 된다.

DNA로부터 전사되어 만들어지는 세 가지 주요 RNA(RiboNucleic Acid)는 구조적으로 DNA와 유사하지만, A, C, G, U(우라실)로 이루어진 단일 가닥으로 존재한다. 세 개의 주요 RNA는 유전체로부터 단백질 합성 장소인 리보솜으로 정보를 전달하는 mRNA (messenger RNA), 단백질을 이루는 아미노산을 리보솜으로 이동시키는 tRNA(transfer RNA), 리보솜 구조 형성의 주요 요소가 되는 rRNA(ribosomal RNA)이다. 이 중에서 rRNA는 mRNA를 잡아서 단백질로의 번역 과정을 촉매하는 일에 참여한다.

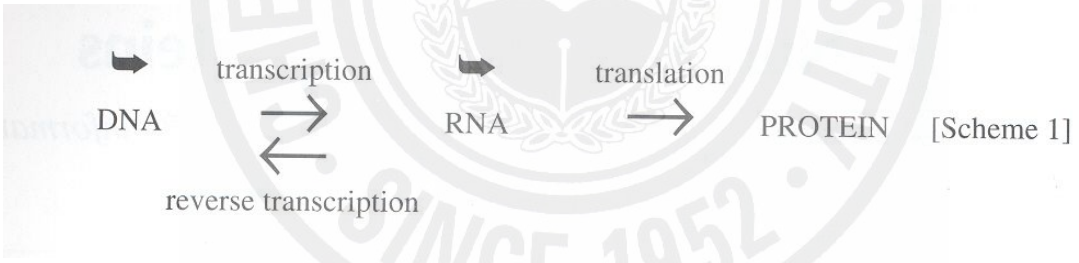

그림 (1) 분자 생물학의 Central Dogma

mRNA가 단백질로 번역되는 과정은 유전체의 정보가 세포 안에서 기능을 수행하기 위해 거치는 마지막 과정이다. 세 개의 RNA 염기 단위인 코돈(Codon)은 표 1에서 나타나는 특성을 가지는 20개의 아미노산(Amino acid)들에 대응한다. 4개의 뉴클리오타이드 A,U,G,C로 이룰 수 있는 모든 조합은 64개이지만 실제로는 그중 20개의 아미노산이 구성되는 것이다. 의 조합을 가지고 있지만 실제로는 20개의 아미노산으로 번역된다. (그림 2 참조)

첫 번째 핵산	두 번째 핵산				세 번째 핵산
	U	C	A	G	
U	UUU 페닐알라닌 (Phe)	UCU 세린 (Ser)	UAU 티로신 (Tyr)	UGU 시스테인 (Cys)	U
	UUC 페닐알라닌 (Phe)	UCC 세린 (Ser)	UAC 티로신 (Tyr)	UGC 시스테인 (Cys)	C
	UUA 루신 (Leu)	UCA 세린 (Ser)	UAA 정지, 글루타민[1]	UGA 정지, 트립토판[2, 3] 시스테인[4] 셀레노시스테인[5]	A
	UUG 루신 (Leu)	UCG 세린 (Ser)	UAG 정지, 글루타민[1]	UGG 트립토판 (Trp)	G
C	CUU 루신 (Leu)	CCU 프롤린 (Pro)	CAU 히스티딘 (His)	CGU 아르기닌 (Arg)	U
	CUC 루신 (Leu)	CCC 프롤린 (Pro)	CAC 히스티딘 (His)	CGC 아르기닌 (Arg)	C
	CUA 루신 (Leu)	CCA 프롤린 (Pro)	CAA 글루타민 (Gln)	CGA 아르기닌 (Arg)	A
	CUG 루신 (Leu), 세린 (Ser)[6]	CCG 프롤린 (Pro)	CAG 글루타민 (Gln)	CGG 아르기닌 (Arg)	G
A	AUU 이소루신 (Ile)	ACU 트레오닌 (Thr)	AAU 아스파라긴 (Asn)	AGU 세린 (Ser)	U
	AUC 이소루신 (Ile)	ACC 트레오닌 (Thr)	AAC 아스파라긴 (Asn)	AGC 세린 (Ser)	C
	AUA 이소루신 (Ile)	ACA 트레오닌 (Thr)	AAA 라이신 (Lys)	AGA 아르기닌 (Arg)	A
	AUG 메티오닌 (Met) 혹은 시작	ACG 트레오닌 (Thr)	AAG 라이신 (Lys)	AGG 아르기닌 (Arg)	G
G	GUU 발린 (Val)	GCU 알라닌 (Ala)	GAU 아스파르트산 (Asp)	GGU 글리신 (Gly)	U
	GUC 발린 (Val)	GCC 알리닌 (Ala)	GAC 아스파르트산 (Asp)	GGC 글리신 (Gly)	C
	GUA 발린 (Val)	GCC 알라닌 (Ala)	GAA 글루탐산 (Glu)	GGA 글리신 (Gly)	A
	GUG 발린 (Val)	GCG 알라닌 (Ala)	GAG 글루탐산 (Glu)	GGG 글리신 (Gly)	G

그림 (2) 61개의 코돈을 나타내는 유전자 코드

표 1은 아미노산 기호와 그 특성들을 나타내고 있다. 또한 각 단위는 질량[dalton], 표면적[Å2], 부피[Å3], 등전위점 값 [25℃] 이다. 각각의 아미노산은 아미노(Amino)그룹과 카르복실(Carboxyl)그룹으로 구성되어 있고, 아미노산들은 펩티드 결합(Peptide bond)이라는 화학적 사슬을 인접한 아미노산의 아미노산그룹과 카르복실 그룹 사이에 형성한다.

단백질은 DNA의 이중 나선(double helix)구조와 달리 펩티드 결합으로 이루어진 아미노산의 서열인 1차 구조(the primary structure)와 알파 나선구조(α-helix), 베타 평판 구조

(β-sheet), 회전 구조(turn)로 나눌 수 있는 2차 구조, 두세 개의 인접한 2차 구조들이 서로 합쳐져서 만들어지는 초월 2차 구조, 이들 구조 블록들이 합쳐진 3차 구조 등을 가지고 있다.

표 (1) 아미노산특성과 아미노산 기호.

아미노산 (Amino Acid)	3문자 코 드	1문자 코 드	질량	표면적	부피	등전위점 값
알라닌 (Alanine)	ALA	A	71.09	115	88.6	10.76
아르기닌 (Arginine)	ARG	R	156.19	225	173.4	2.98
아스파르트산 (Aspartic acid)	ASP	D	114.11	150	111.1	-
아스파라긴 (Asparagine)	ASN	N	115.09	160	114.1	5.02
시스테인 (Cysteine)	CYS	C	103.15	135	108.5	3.08
글루탐산 (Glutamic acid)	GLU	E	129.12	190	138.4	-
글루타민 (Glutamine)	GLN	Q	128.14	180	143.8	6.604
글리신 (Glycine)	GLY	G	57.05	75	60.1	7.64
히스티딘 (Histidine)	HIS	H	137.14	195	153.2	6.038
이소루신 (Isoleucine)	ILE	I	113.16	175	166.7	6.036
루신 (Leucine)	LEU	L	113.16	170	166.7	9347
라이신 (Lysine)	LYS	K	128.17	200	168.6	5.74
메티오닌 (Methionine)	MET	M	131.19	185	162.9	5.91
페닐알라닌 (Phenylalanine)	PHE	F	147.18	210	189.8	6.3
프롤린 (Proline)	PRO	P	97.12	145	112.7	5.68
세린 (Serine)	SER	S	87.08	115	89	-
트레오닌 (Threonine)	THR	T	101.11	140	116.1	5.88
트립토판 (Tryptophan)	TRP	W	186.12	255	227.8	5.63
티로신 (Tyrosine)	TYR	Y	163.18	230	193.6	6.002
발린 (Valine)	VAL	V	99.14	155	140	

하나의 생물은 무수히 많은 수의 단백질로 구성되어 있다. 유전체는 하나의 세포 안에서 변화가 없이 일정하게 유지되는 반면 단백체는 조직의 특성, 상태, 건강과 같은 환경요인에 따라서 시시각각으로 변화한다. 따라서, DNA에서 단백질로 번역된 후에도 단백질 자체에서 다양화가 이뤄지며, 번역 후 단백질 변형은 현재 약 300여 가지 이상으로 알려져 있다. 이것은 하나의 단백질은 하나의 유전자로부터 만들어지는 것이 아니라 동일 유전자상에서 단백질들은 다른 상이한 활성 및 기능을 가지는 단백질들로 만들어 진다는 것과, 생체 내 정보를 얻고자 할 때 DNA나 RNA가 지니는 유전정보만으로는 그 기능 규명에 한계를 가진다는 것을 말한다. 즉, 유전자의 염기서열에 의한 산물인 단백질에 대하여 분석이 되지 않는다면 유전자 기능 규명에 대한 의미가 없어진다.

2. Proteomics 관련 생물정보학 방법

1) 단백질 분자량 측정 및 질량분석

단백질과 같은 분자량이 큰 생체 고분자의 정확한 질량을 알 수 있는 방법은 화학적 구조와 아미노산 서열에 근거를 두고 분자량을 계산하는 것이다. 단백질 동정에 사용 가능한 실험 데이터로는 단백질의 등전위점값(pI) 측정, 분자량, 펩티드 질량 지문추적법, 아미노산 조성, 서열 분석 등이다. 이 중 등전위점값 측정을 제외하고 나머지 방법들은 모두 질량분석에 의존한다.

단백질의 질량에 대한 정보는 대상 단백질의 탐색 및 동정을 위한 작업에 필수적이지만, 등전위점값(pI)의 추가로 인해 보다 자세한 대상 단백질의 특성에 대한 해석이 가능하다.

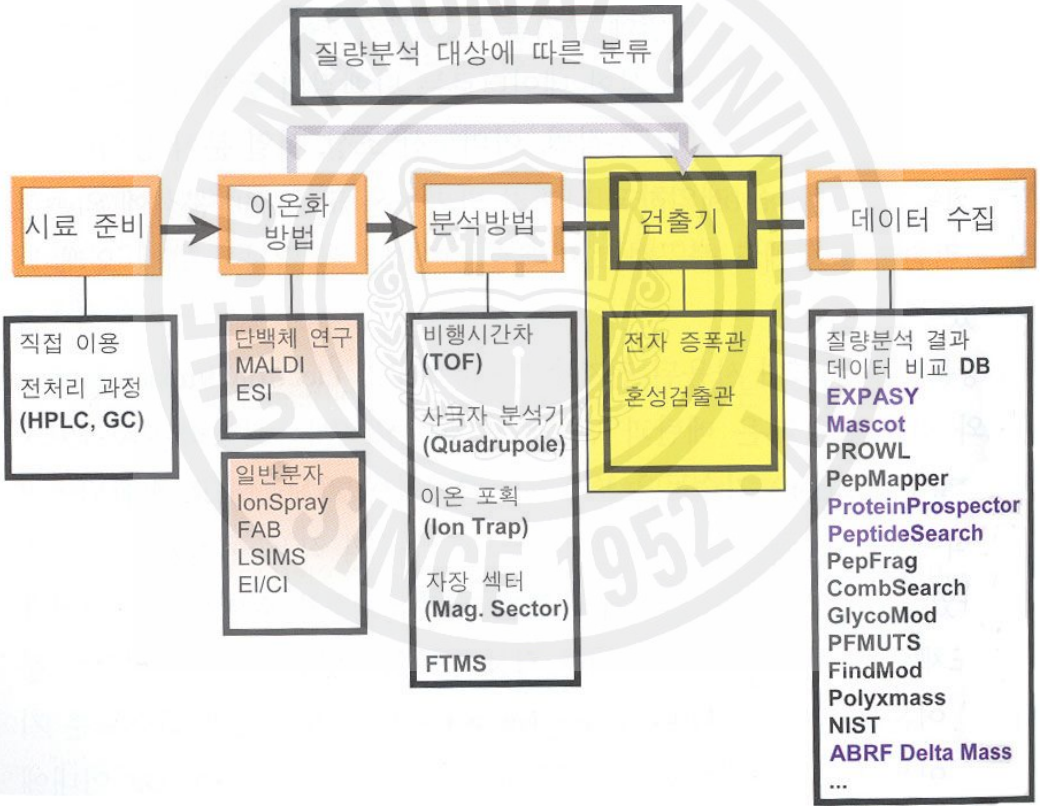

그림 (3) 질량분석 대상에 따른 분류

펩티드 질량지문 추적법(Peptid mass fingerprinting)은 효소를 이용하여 대상 단백질을 특이 위치별로 절단한 후 발생되는 절편들의 질량분석을 통하여 펩티드나 단백질을 동정하기 위한 방법으로서, 기존에 구축된 펩티드나 단백질 질량 데이터베이스에서 유사 질량에 대한 비교 검색을 통하여 실질적인 동정이 이루어진다.

질량분석을 통해서 펩티드의 구조 정보를 얻어내기 위해서는 구조체를 분해

(fragmentation)하고 재조립해야 한다. 분자량이 큰 단백질은 직접적인 서열 결정이 어렵기 때문에 트립신과 같은 단백질 내부 절단 효소를 이용하여 펩티드 사슬들로 절단한 후 단편화를 진행한다.

표 (2) 단백질 절단에 사용되는 분해 효소

분해 효소	R1 특이성
트립신 (trypsin)	Lys, Arg
V8 단백질분해효소 (glutamyl endopeptidase)	Glu, Asp
키모트립신 (chymotrypsin)	Tyr, Tro, Phe>Met, Leu>Gln, Asn>His
펩신 (pepsin)	Phe, Leu, Tyr, Trp>Cys>Glu>Gln
파파인 (papain)	Lys, Arg>His>Gly>Leu>Gln>Glu
엘리스테이즈 (elastase)	대다수의 지방족 혹은 방향족
섭틸리신 (subtilisin)	파파인과 유사
스트렙토 그리세우스 단백질 분해효소 (Strep. griseus protease)	매우 비특이적 절단

2) MOWSE

데이터베이스로부터의 단백질 혹은 펩티드의 질량을 검색하는 알고리즘으로서, 측정된 단백질의 질량 값과 단백질 혹은 펩티드 서열간에 계산된 질량 일치가 우연에 의하여 발생할 확률을 배제하기 위하여 OWL 데이터베이스를 이용한다.

각 질량들을 단편화하기 위해 OWL 데이터베이스의 시퀀스 엔트리 그룹화 단위인 10kDa 단위로 그룹화 한 후, 이 그룹들을 다시 100Da 간격으로 단편화 하고, 각 질량 값들의 빈도수를 측정하여, 제일 큰 빈도수를 가지는 단편의 값을 1로 만든 후에 나머지 단편 값들의 정규화 기준으로 사용한다.

단편화된 질량 값과 단백질 데이터베이스에서 얻은 질량 값을 비교하여 각 값들의 확률점수를 수식 (1)에 의해 계산한다.

$$Score = \frac{50000}{Pn \times H}$$
$$Pn = Product\,of\,distribution\,frequency\,Scores \qquad (1)$$
$$5000 = 'Average'\ Protein\,MW,\ H : 'Hit'$$

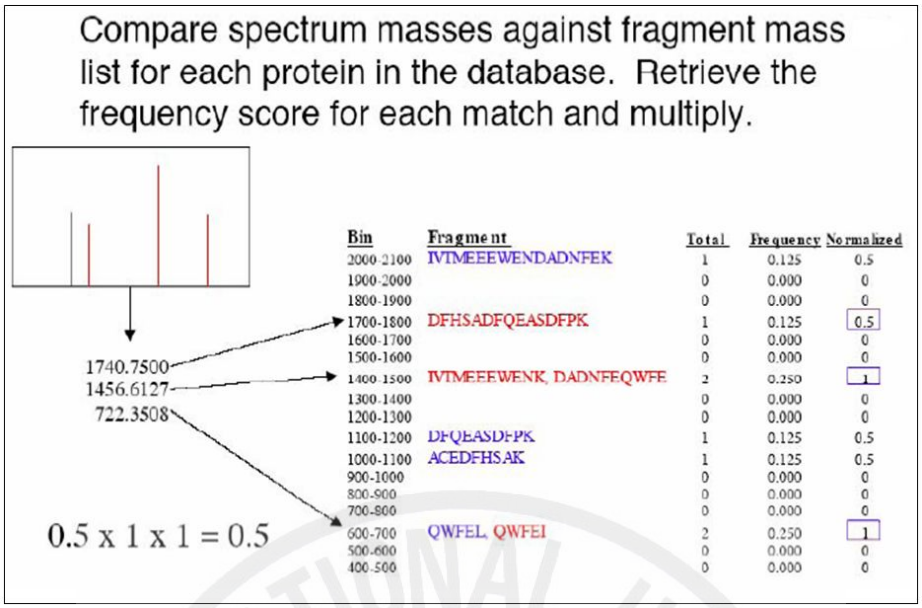

그림 (4) MOWSE 과정

MOWSE 알고리즘은 펩티드 질량분석의 일종으로 단백질 동정, 단백질간의 구조적 차이, 서열 확인과 동질성 분석, 돌연변이 단백질 검출등에 사용할 수 있으나, 동일한 질량값을 가지는 상이한 단백질 혹은 펩티드 서열에 대해서는 사용할 수 없다는 점과, 여러 개의 단백질을 검색 결과로 얻어오기 때문에 검색결과의 신뢰도에 문제가 발생할 수 있다. 신뢰도에 문제가 발생할 경우 해당 검색결과를 재처리해 주는 과정이 필요하기 때문에, 경우에 따라서는 동일한 검색 작업을 반복해야 하는 문제가 발생한다.

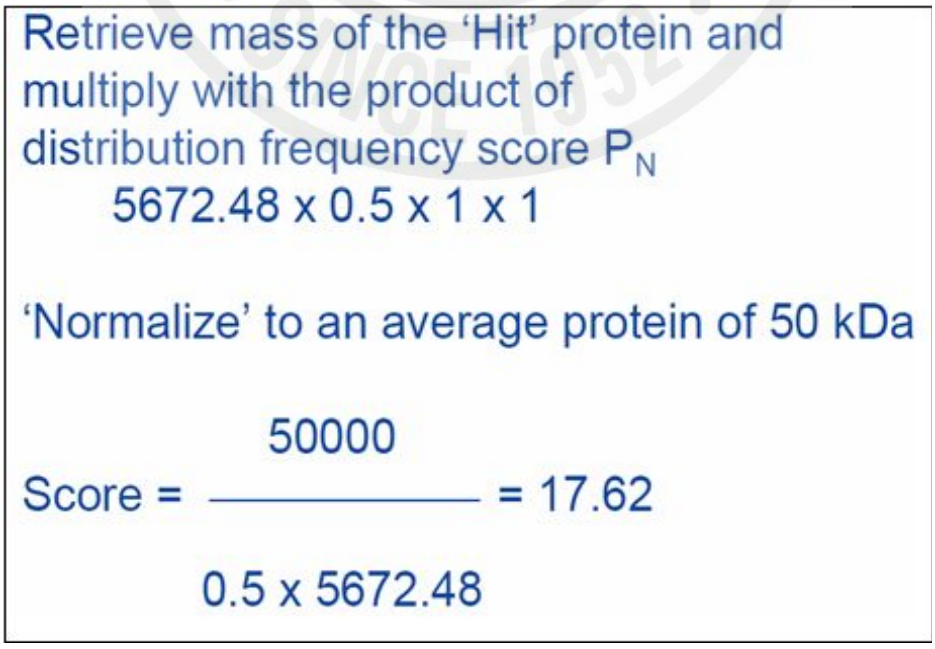

그림 (5) MOWSE 점수 계산

3) Mascot Search

Mascot 은 MOWSE 의 단점을 보완하기 위해 개발되었으며, 확률 기반 스코어링(Probability-based scoring)이 추가되었고, MS/MS 데이터를 지원한다.

단백질 내부 절단 효소를 이용하여 펩티드 사슬로 절단 하여 단편화 한 후 이온화를 진행한다. 각 펩티드 사슬들이 가진 이온들의 질량 차이는 연속적으로 연결된 아미노산들의 서열을 결정하므로 펩티드 서열을 유추할 수 있다.

표 (3) Mascot에서 쓰이는 MS/MS 단편화 시리즈 이온

Ion Type	Ion mass	Low energy CID	High energy CID	PSD	Custom weighting factor
a	[N]+[M]-CO		1	1	☐
a^*	a-NH_3			1	☐
a^0	a-H_2O				
a^{++}	(a+H)/2		1		☐
b	[N]+[M]	1	1	1	☑
b^*	b-NH_3			1	☐
b^0	b-H_2O				
b^{++}	(b+H)/2	1	1		☐
c	[N]+[M] + NH_3				
d	a-partial side chain				
v	y-complete side chain				
w	z-partial side chain				
x	[C]+[M]+CO				
y	[C]+[M]+H_2	1	1	1	☑
y^*	y-NH_3				☐
y^0	y-H_2O				
y^{++}	(y+H)/2	1	1		☐
z	[C]+[M]-NH				

[N], mass of N-term group
[C], mass of C-term group
[M], mass of the sum of the neutral amino acid residue masses

4) 스미스-워터만 알고리즘

스미스-워터만 알고리즘은 최대 공통문자열(LCS, Longest Common Substring)해독법을 동적으로 확장시킨 알고리즘이다. LCS는 문자열 X와 문자열 Y가 주어졌을 때, 이 두 문자열을 만족하는 가장 긴 공통문자열 S를 구하는 것으로, LCS는 최종적으로 수식 (2) 의 상관관계를 얻을 수 있다.

$$L(i:j) = max\{L(i-1, j-10+A[i] == B[j], L(i, j-1), L(i-1, j)\}$$
$$L(i,j) : \cap_{\max} = (A[1:i], B[1:j]) \quad (2)$$

스미스-워터만 알고리즘은 유사 단백질 서열을 2차원 매트릭스상의 어느 지점에서나 구할 수 있도록 음(-)이 아닌 양수를 이용하여 출발점을 독립화 하고, 일치되는 두 서열 간(X, Y)에는 친화도 A_f(X, Y)의 개념을 넣어 일정량의 수를 가감한다. 일반적으로 일치되는 서열 간에는 양(+)의 값을 가산하고, 불일치하는 서열에 대해서는 음(-)의 값을 넣어 감산하는데, 서열간 갭(Gap)에 대한 감산 정도를 일차 비례함수 $d_r = c_0 + kc_1$ 으로 제안하였다.

총감산치 d_r은 음(-)의 값을 가진다.

$$S(i,j) = MAX \begin{cases} 0, S(i-1, j-1) + A_f(X, Y), \\ \max\limits_{k=1\to j} \{S(i, j-k)\} + c_0 + kc_1, \\ \max\limits_{k=1\to i} \{S(i-k, j)\} + c_0 + kc_1, \\ \max\limits_{k=1\to i} \max\limits_{k=1\to i} \{S(i-k, j-m)\} + c_0 + (k+m)c1 \end{cases} \quad (3)$$

$S(i,j)$에는 몇가지 규칙에 의해 전 단계에 존재하는 행렬로부터 3개의 정보가 전해진다. 3개의 정보는 대각선으로 존재하는 $S(i-1, j-1)$의 행렬값에 친화도를 더한 후 $S(i,j)$로 갱신하는 것과, $S(i,j)$의 좌측과 상위에 존재하는 행렬의 값들 중 점수가 높은 값을 탐색 후 각각의 갭 감산치를 $S(i,j)$에 적용하는 것이다.

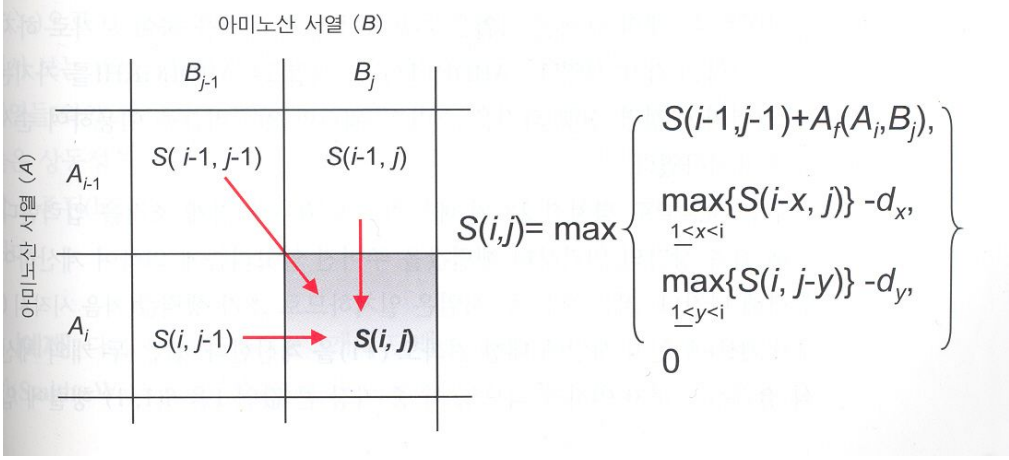

그림 (6) 스미스-워터만 알고리즘 동작원리

5) ClustalW 알고리즘

주어진 각 서열들에 서로 다른 가중치(Weight)를 부여하는 방식으로서, 다중 정렬을 수행하기위해 검색하고자 하는 서열들을 두개씩 짝지어 쌍정렬을 수행하고 이들 중에 가장 연관성이 높은 서열들을 뽑아낸다. 이후 점차적으로 덜 연관된 서열그룹들을 처음 구한 서열에 하나씩 붙여나가는 알고리즘이다.

모든 서열쌍을 정렬한 후 각 쌍들간의 유사도를 결정하는 데 있어서 BLOSUM 이나 PAM 매트릭스를 사용하고, N-J법을 이용하여 유사도 계통수를 작성한 다음 쌍정렬 결과들을 가장 밀접하게 연관된 그룹으로부터 가장 멀리 떨어진 그룹순으로 갭을 부여하면서 연결한다. 갭 벌점은 계통수에 존재하는 각 서열들 간의 가중치 계산으로 얻어진 값을 사용한다.

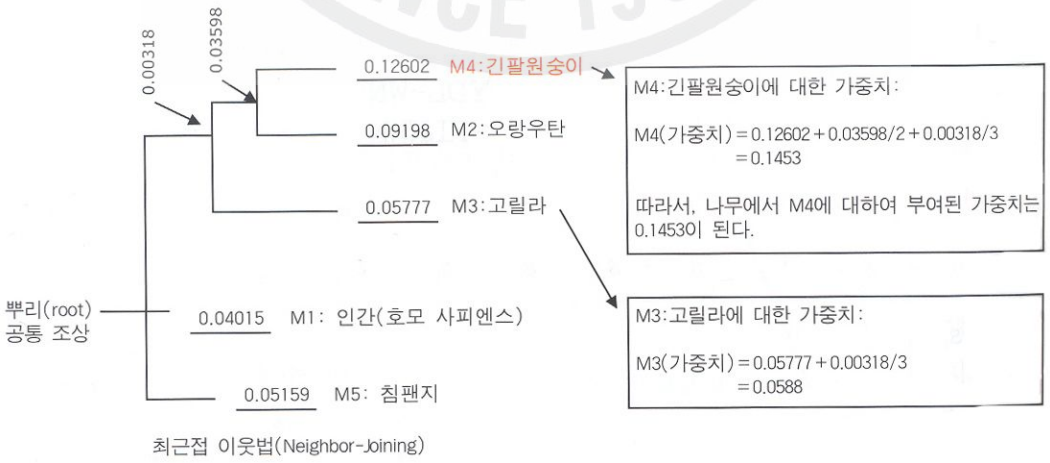

그림 (7) ClusatalW에서 가중치를 구하는 방법

ClustalW 알고리즘은 서열들의 다중정렬 속도가 빠르고 계통수 제작에 가장 많이 이용되고, 단백질 상관분석을 통한 도메인 추출과 아미노산의 가변성에 대한 해석이 가능하다는 장점을 가지고 있지만, 민감도 문제로 인하여 30% 이하의 아미노산 서열 유사도를 지니는 단백질들에 대하여 만족할 만한 결과를 제공하지 못하고 정렬된 결과가 항상 최적의 결과를 지니고 있다고 확신하기 어렵다는 단점이 있다. 이것은 초기 쌍정렬 과정 중 비교적 높은 서열 유사도를 얻은 정렬쌍들에 대해서는 재정렬 과정중에 재조정이 서열간 재조정이 요구되더라도 수정이 어렵기 때문에 세세한 부분 서열간에는 최적이 결과를 제공한다고 보기 어렵다.

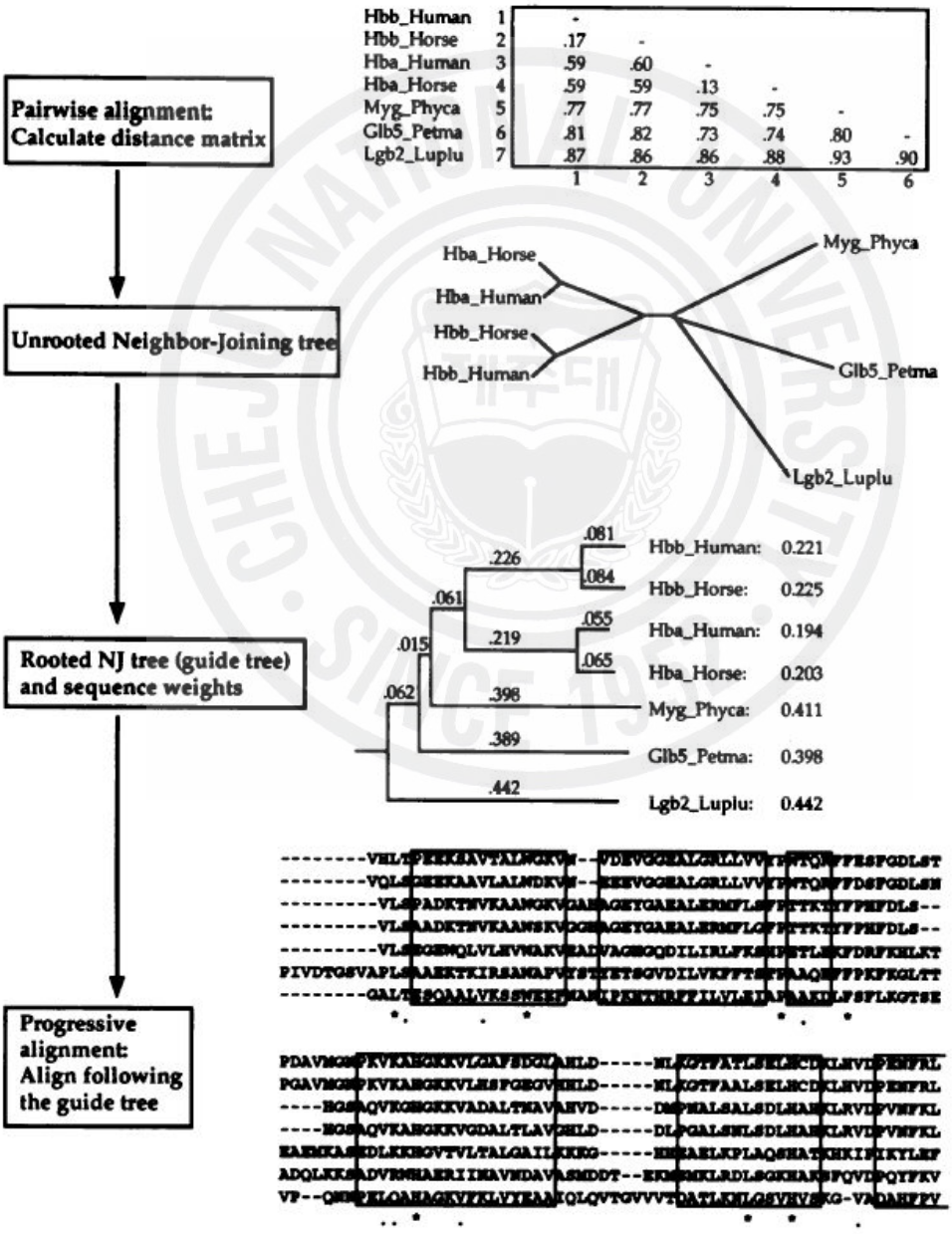

그림 (8) ClustalW 알고리즘 진행도

6) BLAST

BLAST(Basci Local Alignment Search Tool)는 서열 데이터베이스 탐색 도구로서 가장 광범위하게 이용되는 방법으로 1990년에 개발되었고, 1995년 이후로 웹기반 서비스를 제공하고 있다. 초기 최적점 탐색 단계에서 치환 매트릭스를 이용하여 작은 서열이지만 높은 유사성 점수를 지니는 최적점들을 빠른 속도로 검색할 수 있으며, 초기 탐색 단계완 확장 탐색 단계로 나누어 수행 된다.

초기 탐색에 사용되는 부분문자열 길이(w)는 3~5개 정도가 제안되고, 주어진 문자열과 데이터베이스 안의 단백질 서열들과의 부분정렬 점수가 임계치(T) 이상이 되는 서열들만을 선별한다.

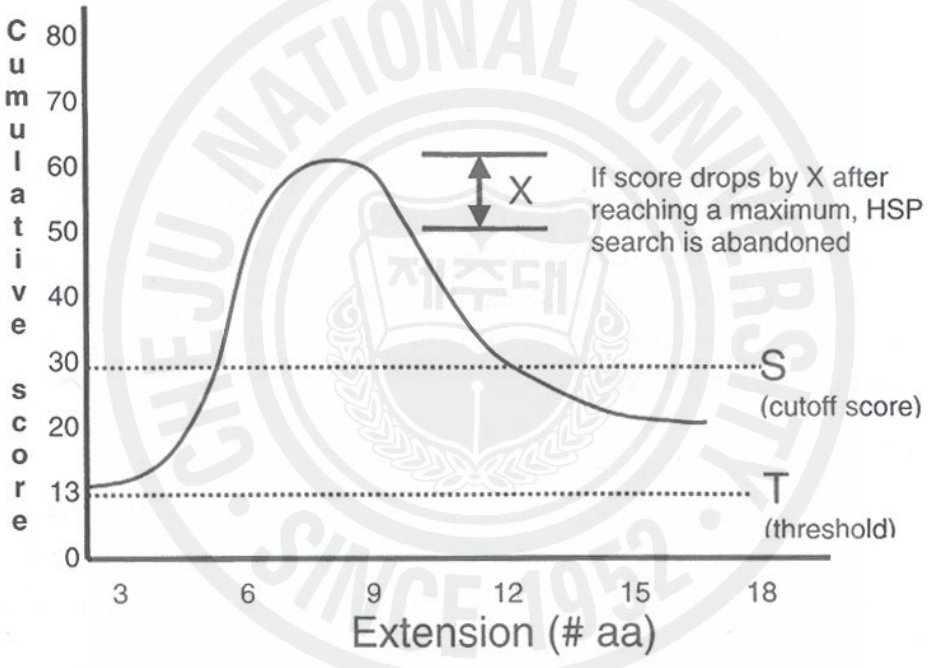

그림 (9) S 와 T 결정

확장 탐색 단계에서는 문자열과 일치하는 단백질 서열들의 일치 지점을 출발점으로 하여 양방향으로 유사성 측정 점수가 음(-) 혹은 주어진 값(S_c) 이하로 떨어질 때 까지 확장 검색을 실시한다.

BLAST는 검색 속도 증가에 대한 보상으로 일치 서열에 대한 민감도를 떨어뜨리고, 짧은 서열이나 특정 반복 단백질 서열이 존재할 때 예측도가 급격히 저하된다. 단백질을 포함한 생체 내에는 반복 유전자 서열들이 작게는 300bp ~ 1.5kbp까지 산재되어 발견된다.

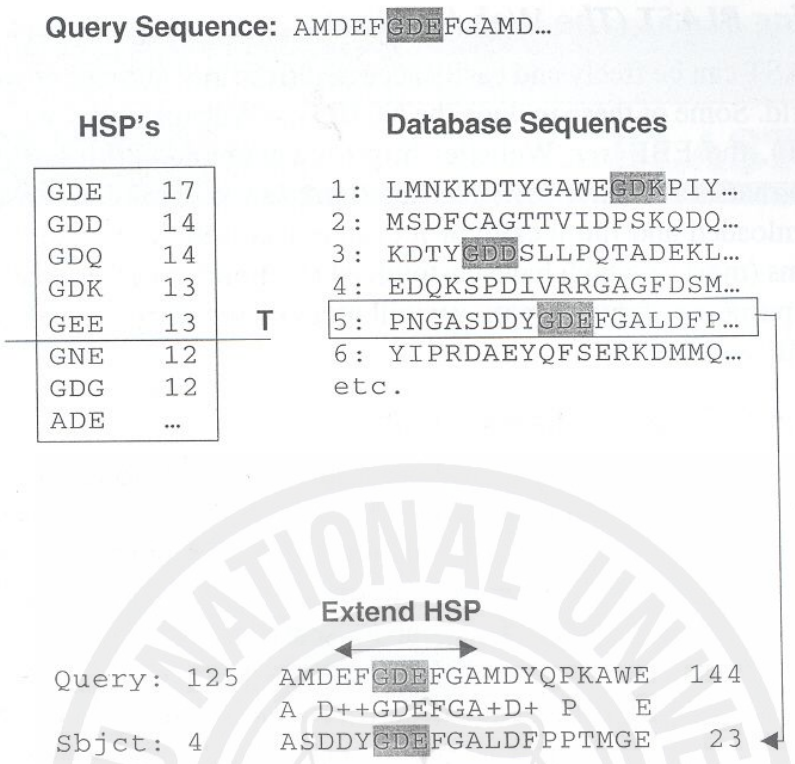

그림 (10) Blast 검색 과정

3. 기존 단백질 동정 시스템의 문제점

EMOWSE, Mascot Search 알고리즘은 단백질 질량값과 펩티드 사슬 단편 이온값을 필요로한다. 하지만 단백질을 구성하고 있는 20가지 아미노산들에는 루신(Leu), 이소루신(Ile)의 질량값과 등전위값(pl)이 동일하며, 라이신(Lys)과 글리신(Gly)의 분자량 차이는 0.04Da 으로 서열 결정에서 혼돈을 야기하는 아미노산이라고 할 수 있다. 또한, 동일한 질량값을 가지는 상이한 펩티드 사슬을 구분에 어려움을 가지고 있고, 이로 인해 단백질 질량 분석법을 기반으로 하는 알고리즘에 얻어진 결과에 대한 신뢰도에 문제가 생긴다.

스미스-워터만 알고리즘과 BLAST, ClustalW 알고리즘의 경우는 치환 매트릭스(PAM, BLOSUM)를 사용하여 서로 다른 아미노산 서열이 얼마나 정렬될 수 있는가를 구분하여 측정점수를 제공하는 것이다. 따라서 치환 매트릭스는 서열 결정 단계에서 매우 중요한 역할을 제공하지만, 두 단백질 서열 사이의 최적 정렬값을 탐색할 때 적용하는 치환행렬 값들의 차이에 따라 전혀 다른 정렬 결과를 얻을 수 있다.

PAM(Point Accepted Mutation)매트릭스는 20가지 아미노산을 20×20의 배열로 구성하고 각각 일대일로 치환될 확률을 계산하여 기록한 것이다. 평균적으로 100개의 아미노산 중 치환 돌연변이가 1개 존재할 때 두 서열은 진화적으로 하나의 PAM 단위를 가지고 있는 것으로 정의한다. 서열간의 유사도 R_{ij}는 수식 (4) 에 의해 계산된다.

$$R_{ij} = \frac{q_{ij}}{p_i p_j} \qquad (4)$$

q_{ij} 는 질의서열 내의 특정 아미노산 i와j가 대체되는 빈도수 이고, $p_i p_j$는 완료서열 내의 특정 아미노산 i와 j가 나타나는 빈도수이다.

실제로 사용되는 PAM 매트릭스는 PAM1 매트릭스뿐만 아니라, PAM67, PAM120, PAM250 등이 사용되는데, PAM 단위가 n 번 반복될 경우 나타나는 아미노산들의 치환 확률을 수치화하여 표시한 도표이다.

BLOSUM(BLOcks SUbstitution Matrix)매트릭스는 PAM 매트릭스의 단점을 보완하기 위하여 블록 대 블록 비교를 통한 치환 행렬값을 결정하였다. 가능한 모든 단백질들의 다중 정렬을 통하여 서로 서열이 일치하는 블록을 검색 후, 이들의 일치 정도가 각각 45%, 50%, 62%를 가지는 개별 블록들을 선별하고, 각각의 블록 안에서 서열 일치도를 이루는 서열들만 집단화하여 통계적으로 PAM 과 유사하게 구한 매트릭스이다.

아미노산 i와 j가 서로 쌍을 이룰 빈도를 f_{ij} 라고 정의할 때, 이 것의 분율은 행 혹은 지정된 블록 내 전체 쌍에 대한 특정 쌍의 비로 나타낼 수 있다.
별도로 우연히 일치된 상황이라면 주어진 쌍들은 어떤 기댓값 e_{ij}의 빈도로 발생된다.

p_i는 관측된 하나의 아미노산이 특정 아미노산 i일 확률이다.

$$q_{ij} = \frac{f_{ij}}{\sum f_{ij}}$$

$$e_{ij} = \begin{cases} 2p_i p_j & i \neq j \\ p_i^2 & i = j \end{cases} \qquad (5)$$

$$p_i = p_{ii} + \frac{1}{2}\sum_{i \neq j} q_{ij}$$

결론적으로, 우연에 의하여 두 아미노산이 일치된 결과와 실제 관측된 통계값 사이를 보정하여 이의 상대적인 비(ratio)를 표시한 오즈 매트릭스(odds matrix)는 $\frac{q_{ij}}{e_{ij}}$가 되며, $\log_2$ 스케일로 변형시킨 형태를 로그-오즈비율(log odds ratio)라고 한다. BLOSUM 치환 행렬의 *i, j*행렬값들은 이 비율을 통하여 결정되며, 각 매트릭스의 행렬값들로 된다.

$$\text{로그-오즈 비율 } (i,j) = 2log_2\frac{q_{ij}}{e_{ij}} = 2log_2\frac{q_{ij}}{p_i p_j} \qquad (6)$$

BLOSUM (*x*) 매트릭스는 일정 수준 이하의 유사도 일치를 보이는 블록들로부터 만들어진 BLOSUM 매트릭스를 말하며, 62% 이내의 서열 유사도를 보이는 블록들을 이용하여 만들어진 매트릭스를 BLOSUM62라 부른다.

PAM 과 BLOSUM 매트릭스 모두 아미노산 서열을 이용하고 있다. PAM은 계통수에 기준을 두어 분지가 발생한지 얼마 안 된 시점의 서열을 이용하여 치환율을 구하고 행렬로 표시하여 이를 PAM1 이라 정의하고, 이 치환율 자체를 여러번 곱하여 진화적 거리를 늘렸다. 반면, BLOSUM은 미리 정해진 서열 유사도를 가지는 집단을 선별하여 이 집단 안에서만 치환율을 구하였다. 따라서 BLOSUM 에서는 이미 정해진 유사도 안에서 사용하였을 때만 의미를 갖는다.

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
A  2
R -2  6
N  0  0  2
D  0 -1  2  4
C -2 -4 -4 -5  4
Q  0  1  1  2 -5  4
E  0 -1  1  3 -5  2  4
G  1 -3  0  1 -3 -1  0  5
H -1  2  2  1 -3  3  1 -2  6
I -1 -2 -2 -2 -2 -2 -2 -3 -2  5
L -2 -3 -3 -4 -6 -2 -3 -4 -2  2  6
K -1  3  1  0 -5  1  0 -2  0 -2 -3  5
M -1  0 -2 -3 -5 -1 -2 -3 -2  2  4  0  6
F -4 -4 -4 -6 -4 -5 -5 -5 -2  1  2 -5  0  9
P  1  0 -1 -1 -3  0 -1 -1  0 -2 -3 -1 -2 -5  6
S  1  0  1  0  0 -1  0  1 -1 -1 -3  0 -2 -3  1  3
T  1 -1  0  0 -2 -1  0  0 -1  0 -2  0 -1 -2  0  1  3
W -6  2 -4 -7 -8 -5 -7 -7 -3 -5 -2 -3 -4  0 -6 -2 -5 17
Y -3 -4 -2 -4  0 -4 -4 -5  0 -1 -1 -4 -2  7 -5 -3 -3  0 10
V  0 -2 -2 -2 -2 -2 -2 -1 -2  4  2 -2  2 -1 -1 -1  0 -6 -2  4
```

그림 (11) PAM250 매트릭스

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
A  4
R -1  5
N -2  0  6
D -2 -2  1  6
C  0 -3 -3 -3  9
Q -1  1  0  0 -3  5
E -1  0  0  2 -4  2  5
G  0 -2  0 -1 -3 -2 -2  6
H -2  0  1 -1 -3  0  0 -2  8
I -1 -1 -3 -3 -1 -3 -3 -4 -3  4
L -1 -1 -3 -4 -1 -2 -3 -4 -3  2  4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4
```

그림 (12) BLOSUM62 매트릭스

III. 제안 시스템

제안하는 단백질 동정 시스템은 크게 3개의 큰 부분으로 구성된다. 1) 핵심적인 부분에 해당되는 새로운 부분서열에 기반한 단백질 분류 기법과 그에 대한 구현, 2) 단백질 동정을 수행하는데 필요한 DB, 3) Proteomics 데이터에 대한 annotation등을 위한 Gene DB이다. 특히 기존의 시스템과는 달리 제안 시스템에서는 기존의 파일 형식으로 된 Gene DB와 단백질 정보를 직접 로컬 데이터베이스로 구현하였기 때문에 독립적인 운용이 가능하며, 이식 및 확장이 용이하다는 부가적인 장점이 있다.

1. 지원시스템 구축

포괄적인 Bioinformatics 시스템을 구축하기 위해서는 Proteomics 시스템과 Genomics 시스템 구축이 필요하다. 전 세계적으로 유전체를 포함한 Bioinformatics 정보량은 해마다 2배 이상씩 성장하고 있으며, 이에 따라 파일시스템 데이터베이스에 적합하게 설계되어 있는 현재의 유전체 데이터베이스를 새롭게 설계해야 한다.

현재 미국국립생물공학정보센터(National Center for Biotechnology Information)에 로스 알라모스 국립연구소(Los Alamos National Laboratory)로부터 이관된 GeneBank 는 세계에서 가장 많은 유전체 염기서열 정보를 소장하고 있으며, 각 염기서열에 대한 정보를 이용하여 데이터베이스를 구축하고 있다.

하지만, GeneBank 의 데이터베이스는 플랫파일 시스템으로서 기하급수적으로 늘어나는 유전체정보량의 처리와, 유전체 검색 서비스의 제공에 유연하게 대처할 수 없다. GeneBank 에서 현재 배포하고 있는 파일은 LOCUS, DEFINITION, ACCESSION, KEYWORDS, SOURCE, REFERENCE, FEATURE, ORIGIN 등으로 구성되어 있다. (표4 참고)

각 줄은 80열로 되어 있고, 각 줄은 1-10열에 기재되는 키워드 부분과 13-80열까지 기록되는 내용으로 구성된다. 키워드와 서브 키워드는 모두 대문자로 표시하고, 서브 키워드는 바로 위에 나타나는 키워드에 종속된 정보임을 표시한다. 1-10열이 공백으로 표시되어 있는 줄은 윗줄의 내용과 연결되는 것을 의미하고, 아래에서 첫 번째 줄에 있는 이중사선(//)은 각 항목의 끝을 나타낸다. 즉, LOCUS 키워드는 항목의 시작을 알리고, 이중사선은 항목을 끝을 알린다. (표5 참고)

GeneBank에서 제공하는 파일에는 수많은 변이가 생성되어 있으며, 단백질 또는 RNA를 암호화하는 부분에 대한 정보나 실험으로 증명된 생물학적으로 중요한 부분에 대한 정보 등을 수록하는 FEATURES 테이블은 그 자체만으로도 내용이 방대하고, 수많은 하부코드(source, gene, CDS, rRNA 등)를 가지고 있다.

Genomics 시스템의 설계를 위해서는 파일내부에 작성되어 있는 변이의 처리와 중복되어 나타날 수 있는 키워드와 서브 키워드 간의 관계의 유동적인 처리가 중요하다. 각 파일들의

변이를 일일이 찾아내는 것은 불가능하다고 판단되어, 변이를 고려하여 유전체 데이터베이스 설계를 하였다.

Genomics 데이터베이스 전체 주키로는 ACCESSION과 VERSION 키워드를 기준으로 하여 일괄적으로 넘버링 하였으며, 데이터베이스 내부에서 관련내용을 검색하고자 할 경우는 해당 주키를 이용하여 검색한다. 각 테이블들은 서로 관련성이 있는 키워드와 부 키워드 내용으로 이루어져 있으며, GeneBank에서 배포하는 파일 형태를 참조하여 각 레코드 길이를 정하였다. (부록1. Genomics 데이터베이스 구성도 참고)

전체 시스템을 구축하기 위해서는 NCBI GeneBank 와 같은 유전체 데이터베이스 뿐만 아니라, Enzyme, Gene 과 같은 작지만 꼭 필요한 데이터베이스 구축이 병행되어야 한다. 이는, 유전체 검색 이후에 제공할 수 있는 분해, 결합, 변이 등을 시뮬레이션 하는 과정에서 필요로 한다. 병행하여 구축한 데이터베이스 구성도를 부록에 첨부하였다.

표 (4) GeneBank field definitions

Field	Description
LOCUS	A short mnemonic name for the entry, chosen to suggest the sequence's definition. Mandatory keyword/exactly one record.
DEFINITION	A concise description of the sequence. Mandatory keyword/one or more records.
ACCESSION	The primary accession number is a unique, unchanging code assigned to each entry. Mandatory keyword/one ore more records.
VERSION	A compound identifire consisting of the primary accession number and a numeric version number associated with the current version of the sequence datain the record. This is followed by an integer key (a "GI") assigned to the sequence by NCBI. Mandatory keyword/exactly one record.
SOURCE	Common name of the organism or the name most frequently used in the literature. Mandatory keyword in all annotated entries/one or more records/includes one subkeyword.
ORGANISM	Fomal scientific name of the organism(first line) and taxonomic classification levels(second and subsequent lines). Mandatory subkeyword in all annotated entries/two or more records.
REFERENCE	Citations for all articles containing data reported in this entry. Includes four subkeywords and may repeat. Mandatory keyword/one or more records.
COMMENT	Cross-references to other sequence entries, comparisons to other collections, notes of changes in LOCUS names, and other remarks. Optional keyword/one or more records/may include blank records.
FEATURES	Table containing information on portions of the sequence that code for proteins and RNA molecules and information on experimentally determined sites of biological significance. Optional keyword/one or more records.
ORIGIN	Specification of how the first base of the reported sequence is operationally located within the genome. Where possible, this includes its localtion within a larger genetic map. Mandatory keyword/exactly one record.
//	Entry termination symbol. Mandatory at the end of ad entry/exactly one record.

표 (5) GeneBank에서 배포하는 유전체 정보파일 형식

```
LOCUS       AAB2MCG1              289 bp    DNA     linear   PRI 23-AUG-2002
DEFINITION  Aotus azarai beta-2-microglobulin precursor exon 1.
ACCESSION   AF032092
VERSION     AF032092.1  GI:3265027
KEYWORDS    .
SEGMENT     1 of 2
SOURCE      Aotus azarai (Azara's night monkey)
  ORGANISM  Aotus azarai
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Platyrrhini; Cebidae; Aotinae; Aotus.
REFERENCE   1  (bases 1 to 289)
  AUTHORS   Canavez,F.C., Ladasky,J.J., Muniz,J.A., Seuanez,H.N., Parham,P. and
            Cavanez,C.
  TITLE     beta2-Microglobulin in neotropical primates (Platyrrhini)
  JOURNAL   Immunogenetics 48 (2), 133-140 (1998)
  MEDLINE   98298008
   PUBMED   9634477
REFERENCE   2  (bases 1 to 289)
  AUTHORS   Canavez,F.C., Ladasky,J.J., Seuanez,H.N. and Parham,P.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-OCT-1997) Structural Biology, Stanford University,
            Fairchild Building Campus West Dr. Room D-100, Stanford, CA
            94305-5126, USA
FEATURES             Location/Qualifiers
     source          1..289
                     /organism="Aotus azarai"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:30591"
     sig_peptide     134..193
     exon            <134..200
                     /number=1
     intron          201..>289
                     /number=1
ORIGIN
        1 gtccccgcgg gccttgtcct gattggctgt ccctgcgggc cttgtcctga ttggctgtgc
       61 ccgactccgt ataacataaa tagaggcgtc gagtcgcgcg ggcattactg cagcggacta
      121 cacttgggtc gagatggctc gcttcgtggt ggtggccctg ctcgtgctac tctctctgtc
      181 tggcctggag gctatccagc gtaagtctct cctcccgtcc ggcgctggtc cttcccctcc
      241 cgctcccacc ctctgtagcc gtctctgtgc tctctggttt cgttacctc
//
```

2. 단백질 DB의 구축

Bioinformatics 시스템에서 중요한 부분을 담당하는 분야가 단백체학(Proteomics)분야이다. 유전자의 최종산물인 단백질의 기능을 규명하기 위해서는, 단백질을 구성하고 있는 아미노산 서열과 함께 단백질의 3차원 구조가 필요하다.

단백체학에서 사용되고 있는 대부분의 구조분석 프로그램은 Brookhaven PDB(Protein Data Bank)형식의 파일을 사용한다. PDB 데이터에는 단백질 원자들의 3차원 공간상의 좌표, 서열정보, 실험정보 및 참조 정보 등이 포함되어 있으며, 이 형식은 80줄 정도의 규정된 형식을 사용하고 있다. 구조에 대한 기본적인 정보를 가진 헤더 부분과 참고 문헌, 분석능(resolution), 결정분석용 그래픽 환경변수(Crystallographic parameter)와 서열 등을 알 수 있으며, 2차구조에 관한 정보와 원자에 대한 정보를 담고 있다. 파일은 HEADER, ORIGXn, SCALEn, COMPND, EXPDTA, SOURCE, ATOM, DBREF, SEQRES, TITLE 등의 레코드로 구성 되어 있다. (표6 참고)

대체로 PDB 파일 각 줄의 첫 6열은 레코드 타입을 나타내며, 11열 이후부터는 해당 레코드의 내용을 나타내고 있다. (표7 참고)

PDB 파일을 데이터베이스화하기 위해서 각 레코드 타입들 중에서 필수항목들과 함께 원자 정보가 들어 있는 ATOM 레코드와 보조 인자(Cofactor), 기질(substrate), 이온, 그 외 단백질 체인과 공유결합을 하지는 않지만 단백질 체인과 연결될 수 있는 잔기에 대한 정보를 포함하는 HETATM 레코드를 포함하였다.

단백체학에서 구축하는 단백질 데이터베이스는 구조분석 프로그램에서만 쓰이기 위한 것이 아니기 때문에 각 단백질을 구성하고 있는 아미노산 서열들에 대한 정보도 필요하다. 각 서열들에 대한 정보는 공용 단백질 데이터베이스인 SWISS-PROT에서 제공하는 파일을 이용하였다. SWISS-PROT 데이터베이스는 높은 수준의 annotation과 자료들간의 중복이 거의 없는 단백질 정보를 제공한다.

SWISS-PROT 파일 역시 80줄 정도의 규정된 형식을 사용한다. 발현 유전자, 단백질의 특징, 단백질을 구성하는 아미노산 서열 등의 정보가 포함되어 있고, ID, AC, DT, DE, GN 등의 레코드로 구성되어 있다. 표8 에서 SWISS-PROT 파일 레코드를 알 수 있다.

데이터베이스화를 위한 단백질 데이터들은 ID를 주키로 사용하여 분류하였으며, 파일 레코드 중 DE, GN, OX, FT, CC, KW, SQ 등으로 선별하여 테이블을 분류하였다. 그리고, SWISS-PROT이 공용 단백질 데이터베이스인 것에 착안 하여, 파일의 DR 줄에 나올 수 있는 모든 참조 데이터베이스를 ID 와 매칭 시키는 테이블을 따로 작성해 주었다. 이것은 단백질 구조분석과 아미노산 서열중심으로 구성된 단백질 데이터베이스 시스템을 언제든지 확장할 수 있도록 하기 위함이다. 단백체학에서 사용 가능하도록 만들어둔 단백질 데이터베이스 구성도를 부록에 첨부 하였다.

표 (6) PDB Record Format

Record	Description
HEADER	First line of the entry, contains PDB ID code, classification, and date of deposition. (Mandatory)
ORIGXn	Transformation from orthogonal coordinates to the submitted coordinates. (n = 1, 2, or 3), (Mandatory)
SCALEn	Transformation from orthogonal coordinates to fractional crystallographic coordinates (n = 1, 2, or 3), (Mandatory)
COMPND	Description of macromolecular contents of the entry. (Mandatory)
KEYWDS	List of keywords describing the macromolecule. (Mandatory)
SOURCE	Biological source of macromolecules in the entry. (Mandatory)
TITLE	Description of the experiment represented in the entry. (Mandatory)
ATOM	Atomic coordinate records for standard groups. (Optional, Madatory if standard residues exist)
DBREF	Reference to the entry in the sequence database(s). (Optional, Mandatory for each peptide chain with a length greater than ten(10) residues, and for nucleic acid entries that exist in the Nucleic Acid Databae(NDB)).
HELIX	Identification of helical substructures. (Optional)
HET	Identification of non-standard groups or residues (heterogens), (Optional, Mandatory if non-standard group other than water appears in the entry)
MTRIXn	Transformations expressing non-crystallographic symmetry(n = 1, 2, or 3). There may be multiple sets of these records. (Optional, Mandatory if the complete asymmetric unit must be generated from the fiven coordinates using non-crystallographic symmetry)
SEQRES	Primary sequence of backbone residues. (Optional, Mandatory if ATOM records exist)
HETATM	Atomic coordinate records for heterogens. (Optional, Mandatory if non-standard grouop appears)
TER	Chain terminator. (Optional, Mandatory if ATOM records exist)
REMARK	General remarks, some are structured and some are free form. (Mandatory)

표 (7) PDB 파일 형식

```
HEADER     DNA                                                  28-MAY-98   XXXX
TITLE    THE INTRINSIC STRUCTURE AND STABILITY OF OUT-OF-ALTERNATION
TITLE     2 BASE PAIRS IN Z-DNA
COMPND     5'-D(*(CH3)CP*GP*GP*CP*(CH3)CP*G)-3
KEYWDS     Z-DNA DOUBLE HELIX
EXPDTA     X-RAY DIFFRACTION
AUTHOR     P.S.HO B.F.EICHMAN, B.BASHAM, G.P.SCHROTH
JRNL          AUTH    B.F.EICHMAN, G.P.SCHROTH, B.E.BASHAM, P.S.HO
REMARK   1
SEQRES   1 A     6   +C    G    G    C  +C    G
HETNAM     CH3 METHYL GROUP
FORMUL   3  CH3     4(C1 H3)
FORMUL   7  HOH    *35(H2 O1)
LINK         C    CH3 A   1                   C5   +C A   1
LINK         C    CH3 A   5                   C5   +C A   5
CRYST1   17.790   30.900   44.360  90.00  90.00  90.00 P 21 21 21     8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.056200  0.000000  0.000000        0.00000
SCALE2      0.000000  0.032360  0.000000        0.00000
SCALE3      0.000000  0.000000  0.022540        0.00000
ATOM      1  O5*  +C A   1      19.586  17.809  18.122  1.00 12.09       O
ATOM      7  C2*  +C A   1      16.834  16.280  17.538  1.00 11.38       C
TER     121        G A   6
ATOM    122  O5*  +C B   7      19.554  18.333  37.567  1.00 11.61       O
ATOM    126  C3*  +C B   7      16.583  18.498  37.564  1.00  9.94       C
TER     242        G B  12
HETATM  243  C    CH3 A   1      15.830  21.653  18.777  1.00 10.00       C
HETATM  250  O    HOH     16      15.486   7.233  35.513  1.00 36.11       O
HETATM  251  O    HOH     17      19.159   9.022  37.958  1.00 10.93       O
CONECT   15  243
CONECT  246  218
MASTER        0    0    0    0    0    0    0    6  279    2    8    2
END
```

표 (8) SWISS-PROT field definitions

Line Code	Contents	Occurence in an entry
ID	Identificaton	Once; starts the entry
AC	Accession number(s)	once or more
DT	Date	Three times
DE	Description	Once or more
GN	Gene name(s)	Optional
OS	Organism species	once or more
OG	Organelle	Optional
OC	Organism classification	Once or more
OX	Taxonomy cross-reference(s)	Once or more
RN	Reference number	Once or more
RP	Reference position	Once or more
RC	Reference comment(s)	Optional
RX	Reference cross-reference(s)	Optional
RA	Reference authors	Once or more (Optional if RG line)
RG	Reference group	Once or more (Optional if RA line)
RT	Reference title	Optional
RL	Reference location	Once or more
CC	Comments or notes	Optional
DR	Database cross-references	Optional
KW	Keywords	Optional
FT	Feature table data	Optional
SQ	Sequence header (blanks) sequence data	Once Once or more
//	Termination line	Once; ends the entry

3. 부분서열을 이용한 새로운 PIS (Protein Identification Search) 알고리즘

질량분석을 통한 단백질 서열을 확인하는 MOWSE는 동일한 질량값을 가지는 서열의 구분이 불가능 하다는 단점이 있다는 것을 앞서 확인하였다. 그리고, 현재 광범위 하게 사용되는 스미스-워터만, ClustalW, BLAST 서열 정렬 알고리즘은 고정된 형태의 PAM, BLOSUM 매트릭스를 사용하고 있다.

본 논문에서 제안하는 알고리즘은 단백질 데이터베이스 전체의 서열 전체를 이용하여 동적으로 단백질 서열 블록 매트릭스를 구성하고, 동적으로 구성되는 매트릭스를 이용하여 질의 하고자 하는 단백질 아미노산 서열 블록과 데이터베이스에 존재하는 각 단백질 서열 블록을 매칭시키는 것을 기본 아이디어로 한다.

1) PFBM (Protein Fragment Block Matrix)

PFBM(Protein Fragment Block Matrix)을 구성하기 위해서 단백질 데이터베이스에 있는 모든 단백질 서열을 대상으로 기본 아미노산 1문자 코드 20개와 그 외의 코드 1개를 하나로 묶어 21×21 행렬을 구성한다.

21×21 행렬에는 전체 단백질 서열에서 무작위로 추출한 아미노산 n개로 구성된 단편서열 n개에 대한 서열 위치별 아미노산의 빈도수($f_{k_{ij}}$)를 측정하여 정규화한 값 ($M_{k_{ij}}$)가 들어간다. 정규화 하기 전에 측정된 각 아미노산들의 빈도수 비율을 동일화하기 위하여 단편 서열의 첫 자리를 기준으로 하여 각 X개의 단편서열을 추출한다.

$$M_{kij} = \frac{f_{kij}}{\sum_{j=1}^{21} f_{kij}} \qquad (7)$$

$$i \leq k \leq n-1,\ i,j = Amino\ acid\ Index$$

주어진 단편서열에 따라서 PFBM의 각 셀에는 0의 값이 들어갈 수 있다. 정렬후 계산 과정에서의 상황을 고려하여 실제 빈도수에는 영향을 주지 않는 범위의 소수(0.01)를 부여한다. 아미노산 n개로 이루어진 단편 아미노산들의 위치별 서열 간 결합도를 측정해야 하기 때문에 $n-1$개의 PFBM을 구성하고, 각 행렬들의 값을 정규화 한다. 이 때 정규화 되어진 값은 단편서열 내에서 i행에 대응하는 $n-1$번째 위치의 아미노산과 j열에 대응하는 n번째 위치의 아미노산의 결합도를 의미한다.

각 단편서열을 기준으로 정해진 아미노산과 아미노산의 결합도를 표시한 PFBM은 $n-1$개를 생성한다. 표9 는 구성된 PFBM 의 예를 보여준다. PFBM은 PAM 이나 BLOSUM과는 달리 항상 동적으로 구성되므로, 표9 는 한 예일 뿐이다.

표 (9) Protein Fragment Block Matrix 구성모습

1 \ 2	A	R	D	N	C	E	Q
A	0.0407	0.0288	0.0229	0.0123	0.0060	0.0300	0.0159
R	0.0339	0.0332	0.0217	0.0169	0.0062	0.0202	0.0176
D	0.0241	0.0195	0.0227	0.0183	0.0055	0.0263	0.0200
N	0.0305	0.0161	0.0190	0.0222	0.0060	0.0249	0.0164
C	0.0266	0.0232	0.0237	0.0161	0.0089	0.0237	0.0164
E	0.0346	0.0205	0.0156	0.0193	0.0067	0.0305	0.0134
Q	0.0288	0.0198	0.0181	0.0173	0.0050	0.0279	0.0293

2 \ 3	A	R	D	N	C	E	Q
A	0.040732	0.02878	0.022927	0.012268	0.006	0.03	0.015878
R	0.033902	0.033171	0.021707	0.016878	0.00622	0.020244	0.017561
D	0.024146	0.019537	0.022683	0.018317	0.005488	0.026341	0.020024
N	0.030488	0.016122	0.019024	0.02222	0.006	0.024878	0.016366
C	0.026585	0.023171	0.023659	0.016146	0.008878	0.023683	0.016415
E	0.034634	0.020512	0.01561	0.019293	0.006732	0.030488	0.013415
Q	0.02878	0.01978	0.018098	0.017341	0.005024	0.027854	0.029268

3 \ 4	A	R	D	N	C	E	Q
A	0.043171	0.018098	0.023902	0.020244	0.004537	0.023415	0.016585
R	0.029512	0.026341	0.020756	0.013488	0.006317	0.037073	0.018805
D	0.039024	0.022707	0.020024	0.019049	0.008	0.024146	0.016829
N	0.024634	0.017829	0.021	0.020732	0.009341	0.027317	0.016854
C	0.029024	0.016366	0.01978	0.016829	0.017341	0.019537	0.01439
E	0.028049	0.021976	0.026341	0.017585	0.003585	0.031732	0.017585
Q	0.028073	0.021488	0.029268	0.021	0.007707	0.026366	0.023415

4 \ 5	A	R	D	N	C	E	Q
A	0.042439	0.017829	0.02	0.017585	0.006683	0.021732	0.011512
R	0.030756	0.037805	0.023415	0.01561	0.006317	0.023927	0.026829
D	0.030244	0.022927	0.023171	0.01978	0.00839	0.019268	0.013902
N	0.02561	0.013439	0.020732	0.027073	0.005805	0.02439	0.020268
C	0.027317	0.02	0.021244	0.018537	0.01278	0.025366	0.013659
E	0.029268	0.029049	0.016098	0.016146	0.005829	0.032927	0.023415
Q	0.026341	0.01878	0.031707	0.021	0.008902	0.036829	0.022439

단편서열을 구성하는 아미노산의 개수보다 하나 적은 $n-1$개의 매트릭스로 구성되는 PFBM은 기존 서열정렬 알고리즘에 이용되기 어렵다. 기존 단백질 서열정렬 알고리즘들은 단백질을 구성하는 아미노산 서열자체를 사용하기 때문에, 아미노산 서열을 블록화 하여 사용하는 별도의 단백질 서열정렬 알고리즘이 필요하다.

이에 본 논문에서는 PFBM을 이용한 알고리즘인 PIBS를 제안하고자 한다.

2) PIBS(Protein Identification Block Search)

PIBS(Protein Identification Block Search)는 PFBM의 구성을 위해 필요한 단편화작업을 수행하고, 단편서열을 구성하는 아미노산의 개수는 n개로 동일해야 한다는 전제조건을 가진다. 길이 l인 질의 단백질의 각 블록 값(Q_l)은 단편서열 각 위치별 아미노산에 PFBM의 값(M_{kij})을 적용한 후 모든 값을 더한다.

PIBS에서 질의 단백질의 블록 개수 l은 단백질 전체서열 개수를 단편서열의 아미노산 개수 n으로 나눈 값 중 소수점 이하를 버림으로서 얻어진다. 질의 단백질과 비교 대상이 되는 데이터베이스에 저장된 각 비교 단백질들의 각 블록 값(C_l) 역시 질의 단백질의 블록 값을 구하는 절차와 동일하게 이뤄지며, PFBM의 값을 사용한다.

$$\begin{aligned} Q_l &= M_{1a1a2} + M_{2a2a3} + \cdots + M_{(n-1)a(n-1)an} \\ C_l &= M_{1a1a2} + M_{2a2a3} + \cdots + M_{(n-1)a(n-1)an} \end{aligned} \tag{8}$$

단백질 서열의 비교를 위해 질의 단백질 블록을 기준으로 비교 단백질들의 블록 개수를 결정한다. 질의 단백질의 블록보다 비교 단백질의 블록이 많을 경우에는 질의 단백질의 블록 숫자만큼만 사용하며, 비교 단백질의 블록이 적을 경우에는 비교 단백질의 마지막 블록인 $l - Ce$블록부터 l블록까지의 값을 -1로 변환한다.

PFBM을 이용하여 질의 단백질과 비교 단백질들의 블록 값을 정한 뒤에 각 블록값을 비교하여 두 서열의 각 블록 매칭 값(B_l)을 검색한다. 이 때, 서열 검색의 범위를 조정할 수 있는 범위값(b)를 지정하게 되는데, 범위값(b)의 결정은 PFBM 구성시 0을 대신한 값(0.001)보다 크고, 1보다 작은 값으로 결정한다.

$$B_l \begin{cases} Q_l = C_l \pm b & 1 \\ Q_l \neq C_l \pm b & 0 \end{cases} \quad (1 \leqq l \leqq m\ ,\ L = 5 \times m\ )$$

$$S = \frac{\sum_{l=1}^{m} B_l}{m} \tag{9}$$

최종 유사도(S)는 비교 단백질의 총 블록 매칭 값에 따라 0에서 1까지의 값을 가지며(수식 (9)), 1에 가까울수록 두 단백질 서열의 유사도는 높은 것으로 판명된다.

표 (10) Portein Identification Block Search

질의 단백질	gi|63540667|ref|XP_145161.4|		
비교 단백질 이름	Score	Matching	Non Matching
RPOC_BRUME	0.3303	73	148
RPOH_THECE	0.3212	71	150
RS17_CAEEL	0.3212	71	150
RRF_LEGPL	0.9076	68	153
RRPL_BUNYW	0.3031	67	154
RS17_CRIGR	0.3031	67	154
RS4E_METMA	0.3	66	154
RS12_COXBU	0.2954	65	155
질의 단백질	gi|38564771|gb|AAR23813.1|		
비교 단백질 이름	Score	Matching	Non Matching
CORO7_MOUSE	0.314286	33	72
SYK_PYRKO	0.291262	30	73
TIM16_RAT	0.291262	30	73
TAL_RALSO	0.291262	30	73
SYM_METMP	0.288462	30	74
T23O_CAEEL	0.278846	29	75
TATB_ECOL6	0.277228	28	73
TERF1_HUMAN	0.27619	29	76

IV. 실험 및 결과

PFBM을 구성하고, PIBS 알고리즘을 실행하기 위해서 단백체학 시스템에 구축한 SWISS-PROT 데이터베이스를 사용하였다. SWISS-PROT 데이터베이스에는 총 206,585개의 단백질 정보가 들어 있다.

PFBM을 구성하기 위해서 무작위로 추출하는 단편서열의 길이를 5개로 정하였으며, 단편화 서열의 개수는 21×100으로 하였다. 단편화 서열의 길이에 의해서 PFBM을 구성하는 매트릭스의 개수는 4개가 되며, 단백질 서열 5,000개당 하나의 PFBM을 구성하였다. 대상이 되는 단백질 서열은 총 206,585개이기 때문에 41개의 PFBM 묶음이 만들어 졌고, PIBS에 적용하기 위해 전체 PFBM의 평균값을 구하여 데이터베이스 서열 전체에 적용 할 수 있는 평균 PFBM을 구성하였다.

PIBS 알고리즘에 적용할 질의 서열은 미 국립 생명공학 정보 센터(National Center for Biotechnology Informaticn)에서 제공하는 서열검색 프로그램인 BLAST에서 사용하는 nr(non-redundant databases) 단백질 데이터베이스에서 3000개의 서열을 추출하여 활용하였다. 3000개의 단백질 서열은 실험환경을 고려하여 100개의 단백질 서열 30개로 나눴다.

PIBS 알고리즘에 200개의 질의 서열을 입력하고 구해진 PFBM에 따라 각 질의 서열들을 블록화 한 뒤, SWISS-PROT에 있는 총 단백질 서열(206,858개)을 대상으로 검색하였다. 검색의 대상이 되는 단백질 서열의 수는 PFBM 구성할 때와 동일하게 5,000개로 제한하였고, 42회 반복하였다.

표 (11)은 PIBS 알고리즘으로 검색한 질의 서열과 결과 서열을 NCBI에서 다시 검색하여 동일한 생물 종에 속하는 단백질들만 표시한 것이다. 질의 서열과 동일한 생물 종에 속하는 결과 서열은 총 65개로서 33%의 서열검색 성능을 보여주었다.

표 (11) PIBS에 의한 검색 결과와 NCBI 검색 결과 매칭

Query Name	Result Name	Score	Matching	Non-Matching
gi|18676606				
gi|15887008				
gi|225874	DCNL2_HUMAN	0.337838	25	49
gi|23125950	COFH_SYNEL	0.877049	321	45
gi|223657	CN102_MOUSE	0.333333	109	218
	CPSM_MOUSE	0.333333	109	218
gi|23127872				
gi|23127809				
gi|23128304				

gi|23128099				
gi|224719	DNS2B_RAT	0.421053	16	22
gi|27734646	DNS2B_RAT	0.369048	31	53
gi|21542396	CRUM1_HUMAN	0.315789	30	65
gi|30316381	CN102_MOUSE	0.363636	12	21
	CPSM_MOUSE	0.363636	12	21
gi|23023431				
gi|226183				
gi|157979				
gi|17986277	CSK22_HUMAN	0.292398	100	242
	CN105_HUMAN	0.291176	99	241
gi|17921995	CSK22_HUMAN	0.308271	41	92
	CN105_HUMAN	0.303704	41	94
gi|15807619				
gi|12834988	MLN64_MOUSE	0.473684	18	20
gi|456277				
gi|53727547				
gi|56270269				
gi|57791178				
gi|6031200	PRGB_HUMAN	0.318182	21	45
	RAB38_HUMAN	0.307692	20	45
gi|32034805	DNAJ_ACTAC	0.394737	30	46
gi|446631	PRGB_HUMAN	0.390244	48	75
gi|447094	PAQR3_MOUSE	0.327731	39	80
	JAK1_MOUSE	0.305085	36	82
gi|448295				
gi|448355	VGFR2_RAT	0.441176	15	19
	T53I2_RAT	0.441176	15	19
gi|739455	PRLR_RAT	0.886364	78	10
	UNC5A_RAT	0.363636	32	56
gi|46140409				
gi|46141134				
gi|15146344				
gi|53761486				
gi|14165442				
gi|14165446				
gi|14165448				
gi|14165450				
gi|45517137				

gi|1094717				
gi|46134347				
gi|45505651				
gi|18201915				
gi|18201915	GLPK3_HUMAN	0.283077	92	233
gi|18201917	GLPK3_HUMAN	0.304348	105	240
	UPK2_HUMAN	0.300578	104	242
gi|18201919				
gi|460329	XYLF_PSEPU	0.344262	21	40
gi|460331				
gi|46134806				
gi|46134869				
gi|12839920				
gi|21700195				
gi|46135337				
gi|223760				
gi|225990				
gi|156055				
gi|157472				
gi|157473	PAX6_DROME	0.325000	39	81
gi|157474	CP6D5_DROME	0.327586	38	78
gi|46131327				
gi|46132262				
gi|46132613				
gi|45520016				
gi|50403743	T53I2_RAT	0.315217	29	63
gi|160966				
gi|46311254				
gi|46315246				
gi|46316235				
gi|47573347				
gi|47574034				
gi|47575086				
gi|166345				
gi|115326				
gi|13432104	GLPK3_HUMAN	0.304348	105	240
	UPK2_HUMAN	0.300578	104	242
gi|53796537				
gi|117580				
gi|309818				

gi|309820				
gi|53800138				
gi|53800150				
gi|53800401				
gi|171550				
gi|134192				
gi|12852157	ATRAP_RAT	0.315315	35	76
	T53I2_RAT	0.303571	34	78
gi|26380307	STX6_RAT	0.410256	16	23
	T53I2_RAT	0.358974	14	25
gi|12852168	WD51B_MOUSE	0.409091	27	39
gi|14017957				
gi|3077806				
gi|3970860	LGI4_HUMAN	0.413333	31	44
	CJ049_HUMAN	0.388889	28	44
	CD9_HUMAN	0.373333	28	47
gi|387050				
gi|191151				
gi|4240137	THAP8_HUMAN	0.343558	112	214
	THAP2_HUMAN	0.340491	111	215
	UPK2_HUMAN	0.338415	111	217
gi|192262				
gi|192287	OLF63_MOUSE	0.472727	26	29
gi|1504038	HXC10_HUMAN	0.292373	69	167
	ICAM2_HUMAN	0.290598	68	166
gi|12856672				
gi|1708299				
gi|15021422				
gi|15021511				
gi|7024451				

V. 결론

본 논문에서는 index 데이터베이스내에 있는 단백질 서열에서 임의로 추출한 부분서열로부터 서열내의 각 아미노산이 결합하는 빈도를 추출하고 이를 이용하여 score matrix를 이용하여 질의되는 단백질 서열을 동정하는 새로운 알고리즘을 제안하였다. 시스템의 구현을 위해 SWISSPROT에 등록된 210,000개의 기능이 규명된 단편 단백질 정보를 자체 ORACLE DB로 구현하였다. 실험에 사용되는 테스트 데이터는 NCBI의 nr 데이터베이스내에 있는 2900여개의 단백질 정보를 이용하였다. 실험 방법은 먼저 nr 데이터베이스내의 모든 단백질 정보를 입력으로하여 blast search를 수행하였다. 이때 index로는 SWISSPROT DB를 이용하였다. 이 실험에서 index에 가장 잘 매칭되는 nr 데이터베이스내의 단백질 2900개를 선택한 후, 이 선택된 서열을 제안된 시스템에 입력하여 결과를 얻었다. 결과를 통하여 제안한 알고리즘이 효과적으로 단백질을 분류할 수 있음을 보였다.

그러나 아직 제한적인 데이터에서의 실험 환경이기 때문에 기존의 방법에 비하여 높은 성능을 가진다고 할 수는 없는 것으로 판단된다. 일반적으로 단백질 동정 시스템은 사용하는 DB 즉 index들에 의해서 성능이 좌우되는 것이 일반적인 현상이다. 따라서 현재 우리가 사용한 index인 swissprot이 아닌 다른 DB에 대해서도 충분한 실험을 해야 할 것이다.

현재 구축된 PFBM과 PIBS는 완성된 알고리즘이 아니기 때문에, 개선의 여지가 많은 것이 사실이다. PFBM은 항상 동적으로 구성되고 있어서, PAM 이나 BLOSUM 과 같은 고정형태의 스코어링 매트릭스와 비교하여 더 유연한 대응이 가능한 반면, 동적인 구성에 걸리는 시간을 단축해야 하는 문제점을 갖고 있다. 앞으로 하드웨어 성능이 발전함에 따라 차차 해결되어질 문제로 생각되지만, 매트릭스를 더 빠르게 구성할 수 있는 프로그램 작성 방법의 개선이 있어야 한다. PFBM을 이용한 PIBS는 표 (12)에서 보여지는 것처럼 30개 이하의 질의 블록을 갖는 단백질 시퀀스를 검색할 때에는 부정확한 결과를 제공한다. 단백질을 검색하기 전에 질의 단백질 블록 개수와 비교 단백질 블록 개수를 비교하여 비슷한 블록 개수를 가지고 있는 비교 단백질 군을 생성하는 부분의 개선이 필요하다. 그리고, BLAST 알고리즘과 스미스-워터만 알고리즘에서 사용되는 갭 감점에 대한 부분이 알고리즘상에서 구현되어 있지 않기 때문에, 질의 단백질 블록과 비교 단백질 블록간에 일치하지 않는 갭에 대한 처리가 이루어 지지 않고 있다. 향후 갭 처리와 관련하여 기존 알고리즘에 사용되는 로그-오즈 비율을 PIBS에 맞게 수정하여 적용 하는 방법을 생각해 볼 수 있다.

표 (12) PIBS 결과중 30개 이하의 블록을 갖는 단백질에 대한 검색 결과

질의 단백질	gi|41616114|tpg|DAA03151.1		
비교 단백질 이름	Score	Matching	Non Matching
LIMC_RHOER	0.395833	19	29
LEP_GORGO	0.395833	19	29
SIX1_MOUSE	0.354167	17	31
DPOE_CRYNE	0.354167	17	31
질의 단백질	gi|1122499|gb|AAB00416.1		
비교 단백질 이름	Score	Matching	Non Matching
SFXN3_MOUSE	0.454545	10	12
SELD_CAEEL	0.454545	10	12
RECR_TREDE	0.454545	10	12
ACDG_METJA	0.428571	9	12

본 논문에서 제안한 방법은 이러한 제약점에도 불구하고, 기존의 방법이 가지는 문제점의 일부분을 개선하려는 노력의 일환으로 기존 시스템의 문제점을 분석하고 해당 문제에 대한 해결책을 제시하고 있기 때문에 향후의 계속적인 연구에 많은 도움을 줄 수 있을 것으로 생각되어지며, 아울러 현재 기초연구에 머물고 있는 생물정보학 관련 툴의 개발에 이바지 할 것으로 생각한다.

Ⅵ. 참고문헌

Anderson, N.L., Anderson, N.G. *Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis* 19 (1998) pp.1853-1861

Bairoch, A., Apweiler, R. *The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res.* 28. (2000) pp.45-48

David N. Perkins, Darryl J. C. Pappin, David M. Creasy, John S. Cottrell. *Probability-bases protein identification by searching seuence databases usgin mass spectrometry data.* Electrophoresis 20. (1999) pp.3551-3567

Henikoff, S., Henikoff, J. G. *Amino acid substitution matrices from protein blocks.* Proc. Natl. Acad. Sci. USA 89. (1992) pp.10915-10919

I.Hsuan Yang., Chien-Ppin Huang, Kun-Mao Cho. *A fast algorithm for computing a longest common increasing subsequence.* Information Processing Letters 93 (2005). pp.249-253

Julie D.Thompson., Desmond G.Higgins. and Toby J.Gibson. *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Research. Vol. 22, No. 22. (1994) pp.4673-4680

Jungblut, P.m Wittmann-Liebold, B. *Protein analysis on a genomic scale.* J. Biotechnology 41 (1995) pp.111-120

L.Bergroth., H.Hakonen., T.Raita. *A Survey of Longest Common Subsequence Algorithms.* Department of Computer Science University of Turku 20520 Turku Finland. pp.39-47

Stephen F.Altschul, Warren Gish, Webb Miller Eugene, W.Myers and David J. Lipman. *Basic Local Alignment Search Tool.* J. MOL. Biol 215. (1990) pp.406-410

Stephen F.Altschul. *Protein alignment scoring system sensitive at all evolutionary distances.* J. MOL. Biol 36. (1993) pp.290-300

T.F.Smith., M.S.WaterMan. *Identification of Common Molecular Subsequences.* Reprinted from J. Mol. Biol. 147. (1981) pp.195-197

Roepstorff, P., Fohlman, J. *Proposal for a common nomenclature for sequence ions in mass spectra of peptides.* Biomed. Mass Spectrom. 11 (1984) p.601

Warren J. Ewens and Gregory R. Grant. *Statistical methods in bioinformatics:an introduction.* Springer-Verlag New York, (2001)

Ⅶ. 부록

1. Genome Database 구성도

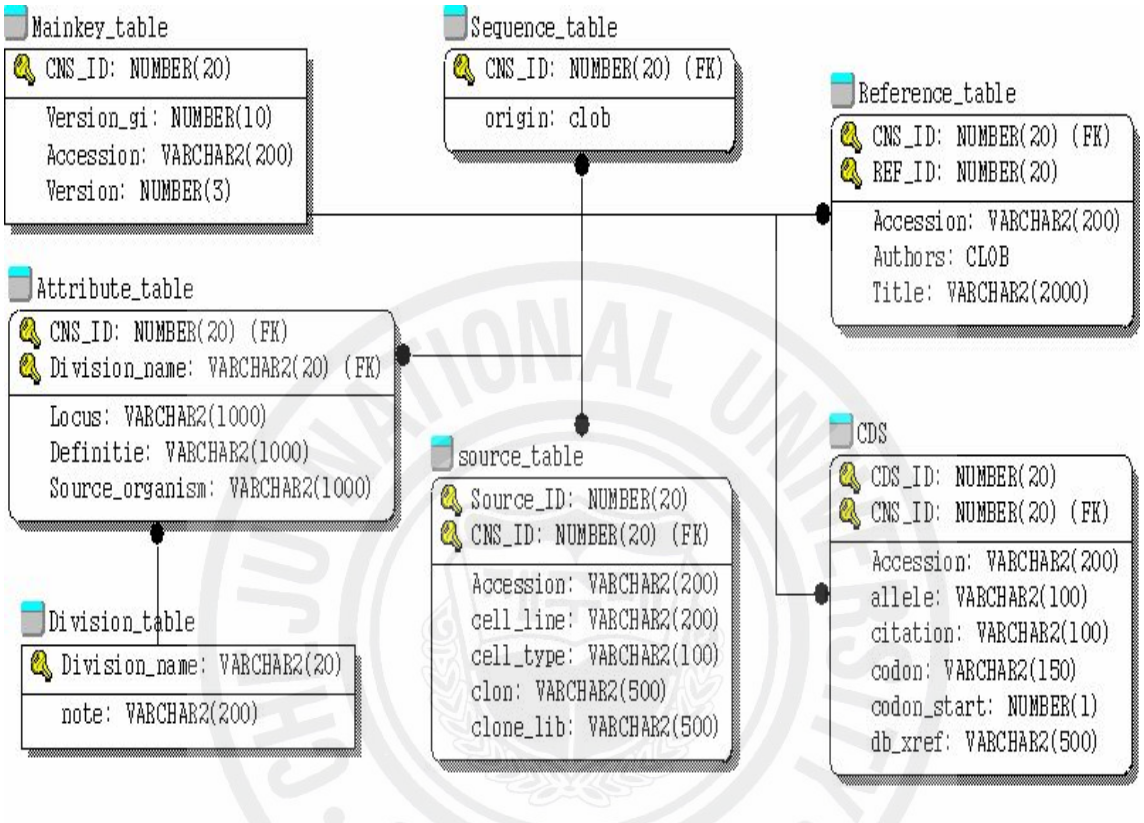

그림 (13) Genome Database 구성

2. 단백질 Database 구성도

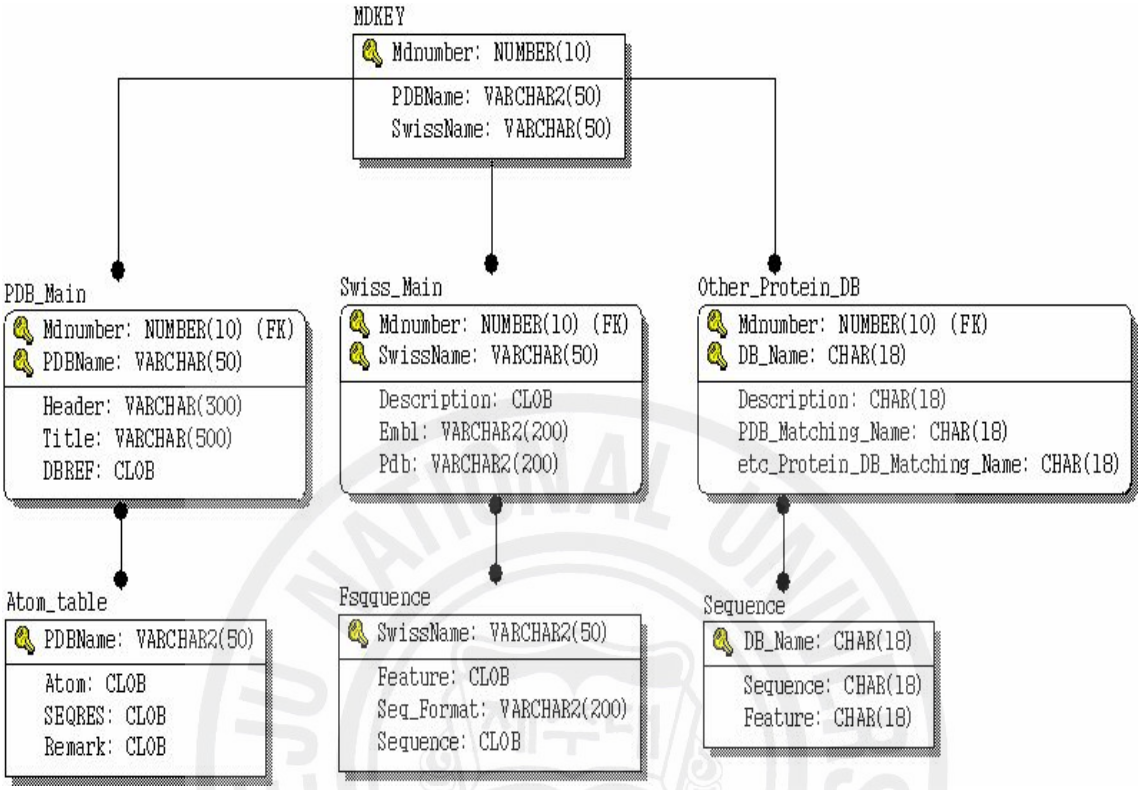

그림 (14) Proteome Database 구성

3. 지원시스템

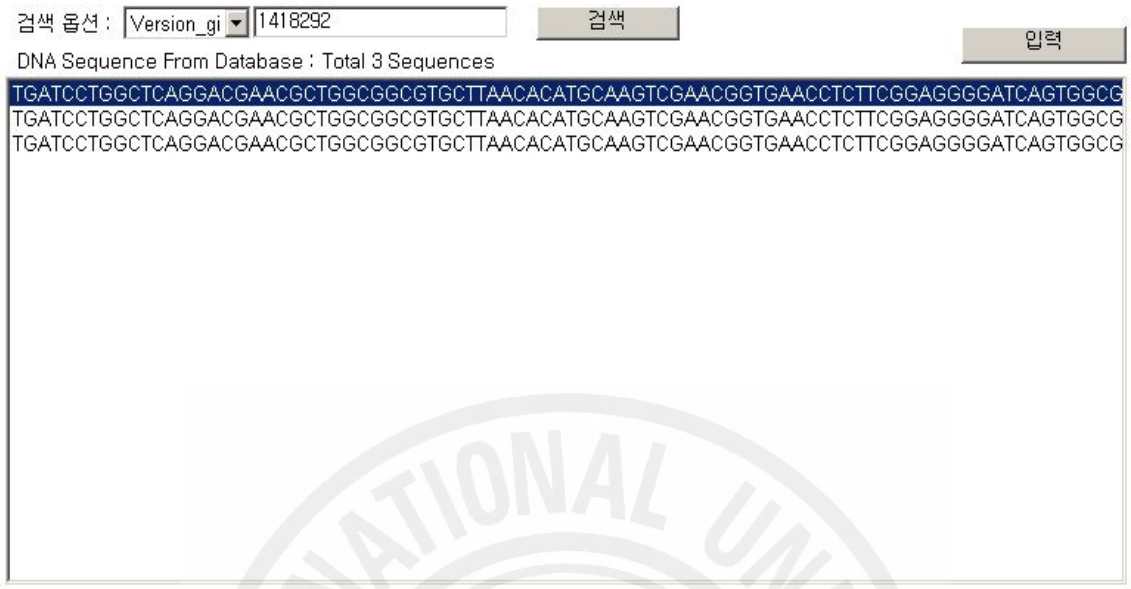

그림 (15) Genome 검색

그림 (16) Genome 변형

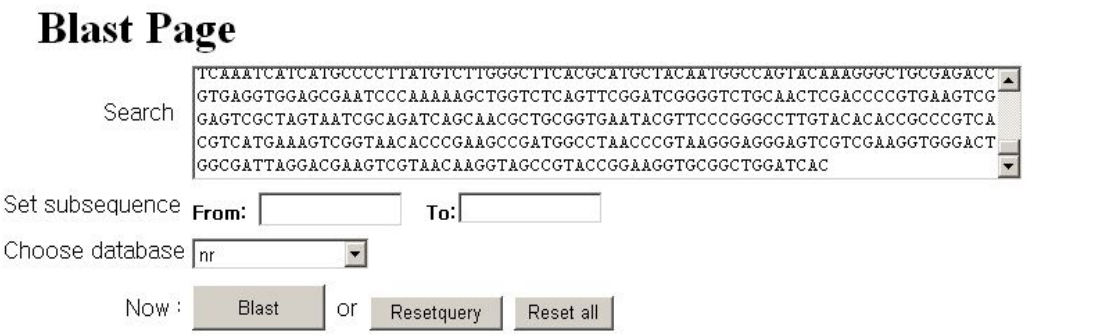

그림 (17) BLAST 검색화면

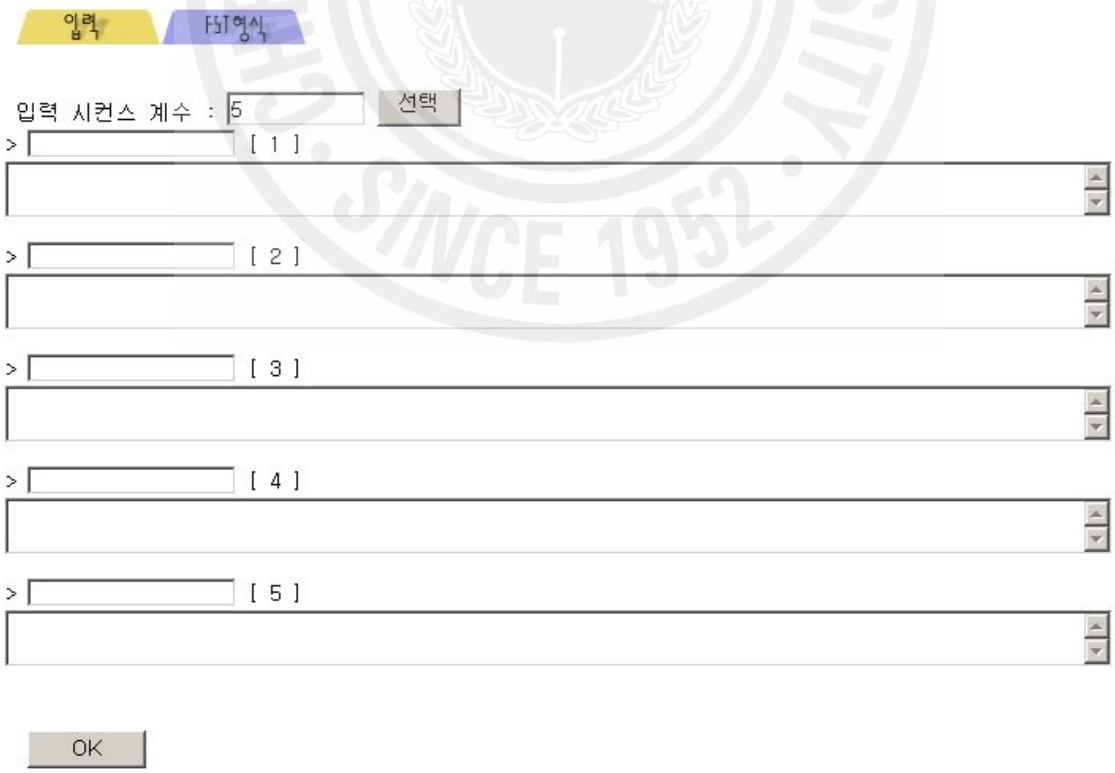

그림 (18) Contig Assembly 화면

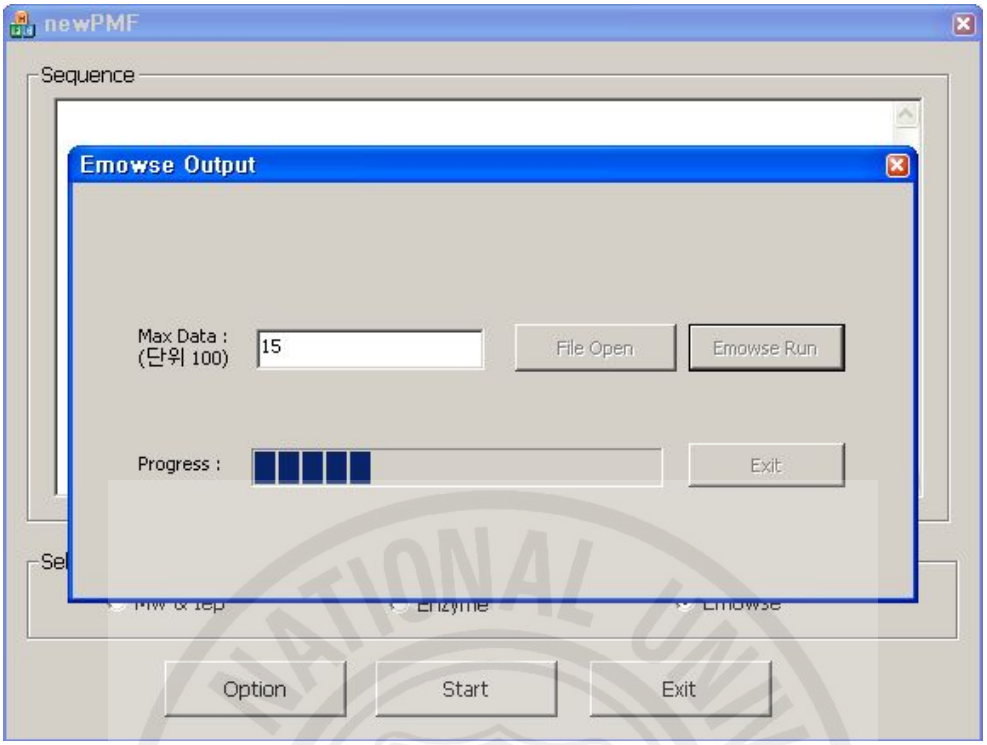

그림 (19) EMOSE 검색 화면