서열 데이타마이닝을 통한 단백질 서열 예측기법

조 순 이[†]·이 도 헌[††]·조 광 휘[†††]
원 용 관[††††]·김 병 기[†††††]

요 약

단백질은 아미노산의 선형 중합체(linear polymer)로서 생체의 조직을 구성하고 각종 생화학 반응을 조절하는 역할을 하는 가장 중요한 생체 분자에 속한다. 이러한 단백질의 특성과 기능은 해당 단백질을 구성하는 아미노산의 서열에 의해 결정되기 때문에, 주어진 단백질의 서열을 알아내는 것은 단백질 기능 연구의 출발점이다. 본 논문은 기존의 생화학적 단백질 서열 결정 방법의 단점을 극복할 수 있는 데이터 마이닝 기반 단백질 서열 예측 기법을 제안한다. 복수개의 단백질 절단효소(protease)를 적용함으로써, 서로 중첩된 단백질 조각을 얻어내고, 각 조각의 질량 정보와 단백질 데이타베이스를 이용하여 후보 서열을 식별한다. 얻어진 후보 서열의 조립을 통해 전체 서열을 결정하기 위한, 다중 분할 그래프(multi-partite graph) 구축 및 경로 탐색 기법을 제안한다. 아울러, 대표적인 단백질 서열 데이타베이스인 SWISS-PROT을 이용한 실험을 통해 제안한 방법의 성능을 평가한다.

A Protein Sequence Prediction Method by Mining Sequence Data

Sun-I Cho[†]·Doheon Lee[††]·Kwang-Hwi Cho[†††]
Yong-Gwan Won[††††]·Byoung-Ki Kim[†††††]

ABSTRACT

A protein, which is a linear polymer of amino acids, is one of the most important bio-molecules composing biological structures and regulating bio-chemical reactions. Since the characteristics and functions of proteins are determined by their amino acid sequences in principle, protein sequence determination is the starting point of protein function study. This paper proposes a protein sequence prediction method based on data mining techniques, which can overcome the limitation of previous bio-chemical sequencing methods. After applying multiple proteases to acquire overlapped protein fragments, we can identify candidate fragment sequences by comparing fragment mass values with peptide databases. We propose a method to construct multi-partite graph and search maximal paths to determine the protein sequence by assembling proper candidate sequences. In addition, experimental results based on the SWISS-PROT database showing the validity of the proposed method is presented.

키워드 : 단백질 동정(Protein Identification), 질량 분석(Mass Spectrometry), 단백질 서열 예측(Protein Sequence Prediction), 다중 분할 그래프(Multi-partite Graph)

1. 서 론

인간 유전체 프로젝트(Human Genome Project) 및 다양한 미생물 유전체 프로젝트를 통해 방대한 양의 유전체 관련 정보가 얻어짐에 따라, 바이오정보학(bioinformatics)이 주목을 받게 되었다. 바이오정보학이란 생물학(biology)과 정보학(informatics)의 합성어로서 컴퓨터를 이용하여 바이오 관련 자료를 체계적으로 정리, 분석하고 이용하는 방법론을 제공하는 학문이다.

생체내의 유전정보는, DNA 유전자 정보 중 일부만이 RNA 거쳐 단백질 합성으로 연결되는 이른바 중심 원리(central dogma)라고 불리는 흐름에 따라 전달되는 것으로 알려져 있다. 이러한 중심 원리의 최종 산물인 단백질은 아미노산의 선형 중합체(linear polymer)로서 생체의 조직을 구성하고 각종 생화학 반응을 조절하는 역할을 하는 가장 중요한 생체 분자에 속한다. 이러한 단백질의 특성과 기능은 해당 단백질을 구성하는 아미노산의 서열에 의해 결정되기 때문에, 주어진 단백질의 서열을 알아내는 것은 단백질 기능 연구의 출발점이다.

특정 단백질을 분리하여 여러 가지 실험을 거친 후 서열의 패턴을 비교하여 미지의 단백질(unknown protein)이 어떤 단백질인가 규명하는 것을 단백질 동정(Protein Identification)이라고 한다. 특히 데이터베이스에 서열이 알려져 있지 않은 단백질의 경우는 단백질 예측(Protein Prediction)이라고 한다. 이러한 단백질 서열의 동정과 예측은 단백질 연구 중에서도 가장 핵심적인 연구 분야로 인식되어 많은 실험기법과 도구의 개발 등이 이루어지고 있다.

현재, 단백질 동정에 많이 사용되는 기법 가운데 하나인

※ 본 논문은 보건복지부·정보통신부 2001년 IMT2000 출연금 기술개발사업에 의해 지원 받았음.
† 정 회 원 : 전남대학교 대학원 전산통계학과
†† 정 회 원 : 한국과학기술원 바이오시스템학과 교수
††† 정 회 원 : 숭실대학교 생명정보학과 교수
†††† 종신회원 : 전남대학교 전자컴퓨터정보통신공학부 교수
††††† 종신회원 : 전남대학교 전산학과 교수
논문접수 : 2002년 8월 23일, 심사완료 : 2002년 9월 11일

Protein Expression Profiling은 많은 단백질 연구자들에게 특별한 관심의 대상이 되고 있다[1-3]. 이는 다음과 같은 3단계를 거쳐 이루어진다. 먼저, 2차원 전기영동(2D-PAGE)을 이용하여 세포의 단백질을 분리시킨다. 분리된 단백질 중에서 흥미있는 스팟을 잘라내서 트립신(trypsin)과 같은 효소(protease)를 사용하여 단백질 분자를 작은 조각으로 자르고, MALDI-MS(Matrix Assisted Laser Desorption/Ionization-Mass Spectrometry) 질량 분석을 한다. 마지막으로, 실제 실험을 통해 측정된 질량을 단백질 데이터베이스에 등록되어 있는 질량과 비교한다. 같은 질량을 가진 후보 서열 중에서 매칭되는 조각의 수가 많은 것을 해당 스팟의 단백질로 인식한다. 그렇지 않으면 단백질 서열을 밝히기 위해 에드먼 감성법(Edman degradation)과 같은 복잡한 실험 과정을 거쳐야 한다[4-6].

이러한 과정이 매우 일반적이긴 하나 단백질 동정 과정에서 비슷한 질량을 가진 후보 서열이 너무 많이 생긴다는 문제점이 있다. 측정할 질량/전하량과 샘플의 유형을 명시하여 검색 공간을 제한한다 하여도 후보 서열이 100개 이상이 되면 추적하기가 어렵다. 일부 연구에서는 단백질을 조각내고 아미노산 질량 차이를 측정하는데 직렬 질량 분석법(tandem mass spectrometry)[7, 8]과 같은 보다 정교한 장비를 사용하기도 한다. 그러나 일반 실험실에서 사용하기에는 너무 고가이고 더구나, 단백질 데이터베이스에 아직 등록되지 않은 새로운 단백질은 예측할 수 없다는데 그 한계가 있다.

이에 본 논문은 일반적인 MALDI-MS 장비를 사용하여 단백질의 서열을 동정(identify)하고 나아가 예측(predict)까지 할 수 있는 효과적인 기법을 제안한다. 단백질을 조각내기 위해서 단백질 서열의 특정 위치만을 절단하는 효소를 여러 개 사용하고, 또한 조각들의 후보 서열들을 단백질 서열로 정렬하는데는 다중 분할 그래프를 이용한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 본 연구에서 제안한 단백질 서열의 예측 기법을 제시한다. 4장에서는 실험 및 결과 분석에 대해 기술하고, 5장에서는 결론 및 향후 연구 과제를 기술한다.

2. 관련 연구

2.1 에드먼 감성법

단백질의 아미노산 서열을 알아내는 가장 기본적이고 확실한 기법이다[9]. 실험 단계가 하나의 사이클로 이루어져 있고 사이클을 한 번 돌 때마다 단백질의 N 말단부터 하나씩 잘라낸 아미노산 성분을 식별한다. 이 방법은 아미노산 1개 단위까지 서열을 알아낼 수 있다는 장점이 있으나 화학적 실험 특성상 길어야 아미노산 50개 정도까지 밖에 시퀀싱(sequencing) 할 수 없으며 시퀀싱하는 시간도 오래 걸린다는 한계가 있다. 따라서 정제된 단백질의 N 말단 아미노산 서열과 부분 아미노산 서열을 결정할 때 다른 기법과 병행해서 사용하는 경우가 많다.

2.2 직렬 질량 분석법

두 개의 질량분석기(mass spectrometer)를 직렬로 연결하여 시퀀싱하는 기법이다[10]. 첫 번째 MS에서는 단백질을 조각내어 질량 스펙트럼을 분석하고 그 중에서 관심있는 조각만을 선택하여 두 번째 MS로 보낸다. 두 번째 MS에서는 불활성 기체와 충돌시켜 생긴 단백질 조각을 다시 아미노산 단위로 분해한다. 출력된 스펙트럼을 분석하면 아미노산 성분 뿐만 아니라 순서까지 알 수 있다. 그러나 실험결과 나온 질량 스펙트럼이 매우 복잡하여 이로부터 서열을 직접 유추하는 것은 고도의 기술과 많은 시간을 필요로 하기 때문에, 대부분 데이터베이스에서 알려진 단백질 조각들의 질량 스펙트럼과 비교하여 서열을 찾아낸다.

따라서 이러한 직렬 질량 분석법은 정확한 질량 정보를 제공하여 단백질 동정이 매우 빠르고 효과적이긴 하나 데이터베이스에 없는 새로운 단백질의 서열 예측은 어렵고 장비의 가격도 비싸서 보통의 실험실에서 사용하기에는 한계가 있다.

3. 단백질 서열의 예측 절차

본 논문에서 제안하는 단백질 서열의 예측 기법은 다음과 같은 네 단계로 구성된다.

① Residue Mass Index(RMI)의 구축
② 타겟 단백질의 MALDI-MS 분석
③ 질량 정보에 의한 조각 매칭
④ 다중 분할 그래프를 이용한 조각 정렬

3.1 Residue Mass Index(RMI)의 구축

효소는 그 화학적 성질에 의해 단백질 서열 중에서 특정 위치만을 절단한다. 예를 들면 트립신(trypsin)은 라이신(K)이나 아르기닌(R)이 N 단말 쪽에 나타나는 위치를 절단한다. 따라서 데이터베이스에 있는 모든 알려진 단백질은 절단 효소를 이용하여 조각으로 만들 수 있고, 각 조각의 질량을 계산할 수 있다. 알려진 아미노산이 20개로 개수가 그다지 많지 않고 각 아미노산은 질량 차이에 의해 구분될 수 있으므로 계산은 쉽게 할 수 있다. 주어진 하나의 효소에 의해 생긴 단백질 조각의 질량 정보를 표현한 것을 Residue Mass Vector(RMV)라고 한다. 다음은 RMV의 일례이다.

<Pepsin, P31946, 15-20, VLAAIL, 598.41>

위 RMV는 'P31946이라는 단백질을 Pepsin으로 분해하여 생긴 조각 중에 하나로서 위치는 15에서 20, 아미노산 서열은 VLAAIL, 질량은 598.41이다'를 의미한다. 주어진 단백질 데이터베이스에 일련의 효소들을 이용하여 만든 RMV의 전체집합을 Residue Mass Index(RMI)라고 정의한다. 이 때 사용된 효소 집합을 Protease Set of Fragmentation(PSF)이라고 한다. RMI의 구축은 프로그램만으로 할 수 있기 때문에 매우 간단하며 이미 공개되어 있는 데이터베이스도 있다.

3.2 타겟 단백질의 MALDI-MS 분석

타겟 단백질을 PSF를 이용하여 조각으로 절단한 후 MALDI-

MS 질량 분석을 하면 각 조각의 RMV를 구할 수 있다.

3.3 질량 정보에 의한 조각 매칭

타겟 단백질 조각의 질량과 같거나 비슷한 것을 찾기 위해 RMI를 이용한다. 타겟 조각에 사용한 것과 같은 효소의 RMV만이 매칭 대상이 된다. 예를 들어, <Pepsin, 598.41>과 같은 타겟 조각이 있다면 <Pepsin, P31946, 15-20, VLAAIL, 598.41>, <Pepsin, P40956, 110-115, LVAAIL, 598.41>, <Pepsin, P56231, 19-24, VLIAAL, 598.41>과 같은 매칭 엔트리가 있을 수 있다. 이는 타겟 조각의 서열이 'VLAAIL', 'LVAAIL', 'VLIAAL' 중에 하나가 될 수 있음을 의미한다.

3.4 다중 분할 그래프를 이용한 조각 정렬

3.4.1 다중 분할 그래프를 이용한 조각 정렬 알고리즘

단백질 서열의 조각 정렬의 문제를 해결하기 위해서 다중 분할 그래프를 사용한다. 다중 분할 그래프의 같은 파티션의 노드와는 연결할 수 없고 반드시 다른 파티션의 노드와 연결할 수 있다는 파티션의 개념을 채택하면 각 타겟 조각에 대한 후보 서열들을 파티션으로 매칭할 수 있다. 즉, 타겟 조각의 많은 후보들 중 바른 서열 하나만을 선택하여 정렬해야 하는 문제를 해결할 수 있기 때문이다.

다중 분할 그래프를 이용하여 각 조각들의 후보 서열들을 단백질 서열로 정렬하는 알고리즘은 다중 분할 그래프 구축 <표 1>과 그리디 알고리즘에 의한 가중치 최대 경로 탐색 <표 2>의 두 단계로 되어 있다.

〈표 1〉 알고리즘 1 : Multi-partite Graph의 구축

```
알고리즘 1. Multi-partite Graph의 구축
Input : G = (V, E, P), and ∀i, j  v_i, v_j ∈ V, P(v_i) ≠ P(v_j)
Output : Multi-partite Weighted Graph G = (V, E, P, W)
begin
    for (all connected edges v_i, v_j is not empty) do
        PlusWeight(v_i, v_j)
end
```

알고리즘 1에서 사용하는 함수의 정의는 다음과 같다.

PlusWeight(v_i, v_j) : *HasBridge*(v_i, v_j)가 true이거나 *SameProtein* (v_i, v_j, R)이 true이면 가중치 increment

HasBridge(v_i, v_j, G_b) : v_i의 suffix가 u_i($u_i \in G_b$)의 prefix이고 u_i의 suffix가 v_j의 prefix이면 true

SameProtein(v_i, v_j, R) : 조각 데이터베이스 R에서 v_i, v_j가 같은 단백질로 발견되면 true

〈표 2〉 알고리즘 2 : 최장길이 경로 탐색

```
알고리즘 2. 최장길이 경로 탐색
Input : G = (V, E, P, W)
Output : Max Weighted Path
begin
  MaxWeightedPath := NIL
  for ( v_i ∈ V is not empty) do
  begin
    LocalMaxWeightedPath := GreedyMWP ( v_i ) /* v_i는 start
        vertex */
    if Weight (MaxWeightedPath) < Weight (LocalMaxWeightedPath)
        then
      MaxWeightedPath := LocalMaxWeightedPath
  end
end
procedure GreedyMWP ( v_k )
begin
  if (k == |P|) /* |P|는 파티션의 개수 */
  else
    v_j := MaxWeight(v_k, v_{k+1}) ∀i, 1 ≤ i ≤ k, P(v_{k+1}) ≠ P(v_i)
    MaxWeightPath := v_j Ψ GreedyMWP(v_j) /* Ψ는 ordered
        merge */
  return MaxWeightPath
end
```

알고리즘 2에서 사용하는 함수의 정의는 다음과 같다.

MaxWeight(v_k, v_{k+1}) : v_k와 연결된 파티션 $P(v_{k+1})$ 에지 중 가중치 최대인 노드를 return

사용된 절단 효소 중 기준 효소를 하나 선택하여 다중 분할 그래프를 구축한다. 그래프에서 노드는 매치를, 파티션은 같은 타겟 조각과 매치되는 후보서열들을, 에지의 가중치는 정렬 가능 정도를 의미한다. 모든 에지의 가중치는 0으로 초기화하고 반드시 다른 파티션의 노드와만 연결할 수 있다. 서로 다른 파티션의 두 매치가 같은 데이터베이스의 단백질인 경우와 두 매치간의 브리지 서열이 다른 그룹 중에 있으면 각각 가중치를 1씩 올린다. 기준 노드를 달리하여 위 과정을 반복한다. 가중치의 반영이 끝나면 모든 파티션을 한 번씩만 경유하는 가중치가 최대인 경로를 찾는다.

3.4.2 [예제] 다중 분할 그래프를 이용한 조각 정렬

타겟 단백질을 효소 분해하여 생긴 타겟 조각을 QF라고 하자. RMV 매치에 의해 얻은 후 보서열이 'VLAAIL', 'LVAAIL', 'VLIAAL'일 때, QF : VLAAIL, LVAAIL, VLIAAL라고 표현한다. 효소를 여러 개 사용하면 여러 그룹의 매칭 목록을 구할 수 있다. <표 3>는 세 종류의 효소를 사용하였을 때, 세 그룹의 매칭 목록이 생긴 예를 보여주고 있다.

〈표 3〉 매칭 목록 그룹의 예

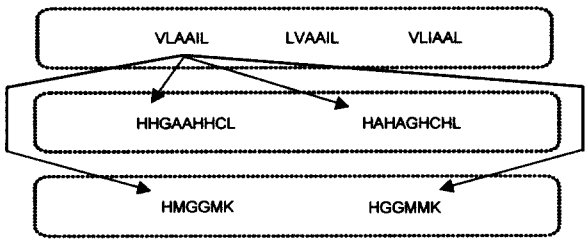
	매칭 목록 그룹
그룹 1 : 효소-Ⅰ	QF 11 : VLAAIL, LVAAIL, VLIAAL QF 12 : HHGAAHHCL, HAHAGHCHL QF 13 : HMGGMK, HGGMMK
그룹 2 : 효소-Ⅱ	QF 21 : VLA, ALV QF 22 : AILHHGA, ALIHHAG, HAHLIGA QF 23 : AHHCLHMGGMK, KCLHHHMGGMA
그룹 3 : 효소-Ⅲ	QF 31 : VLAAILHH, LVAAILHH, AAVLLIHH QF 32 : GAAHHCLHMGGMK, GHAACLHMGGMKH

<표 3>의 매칭 목록 그룹과 <표 4>의 RMI가 있다고 할 때, 먼저 (그림 1)에서처럼 첫 번째 기준 노드인 'VLAAIL'과 다른 파티션의 노드들을 매칭한다. (그림 1)에서 둥근 박스의 서열들은 기준 그룹의 조각들과 매치되는 후보 서열들이다. 다른 그룹의 서열들은 에지의 가중치를 결정할 때 브리지로 사용된다. 다중 분할 그래프를 구축하기 위해 기준이 되는 노드와 다른 파티션의 노드들을 연결하고 각 에지의

〈표 4〉 Residue Mass Index의 예

효 소	단백질	아미노산 서열	위 치	질 량
Pepsin	P 31946	VLAAIL	15~20	598.41
Pepsin	P 31946	HHGAAHHCL	21~29	982.44
Pepsin	P 40956	LVAAIL	110~115	598.41
Pepsin	P 56231	VLIAAL	19~24	598.41
Pepsin	P 12487	HAHAGHCHL	38~46	982.44
Pepsin	P 23654	HGGMMK	39~44	660.30
...	...	...	...	...
Pepsin	P 31946	HMGGMK	30~35	660.30

가중치를 0으로 초기화한다. 에지의 가중치를 결정하기 위해서 3.4.1에서 제시한 규칙인, 브리지 서열이 있는가의 여부와 데이터베이스에 같은 단백질의 조각으로 등록되어 있는지의 여부를 따진다.



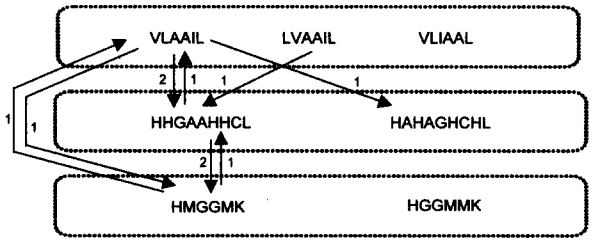
(그림 1) 기준 노드와 매칭하는 다중 분할 그래프

예를 들면 'VLAAIL'과 'HHGAAHHCL' 에지는 QF 22의 'AILHHGA'가 브리지 역할을 하고 있다. 왜냐하면 'AILHHGA'의 접두사가 'VLAAIL'의 접미사로 매치되고 'AILHHGA'의 접미사가 'HHGAAHHCL'의 접두사와 매치되기 때문이다. <표 5>는 그룹 2와 그룹 3 중에서 브리지 서열을 가진 모든 에지를 보여주고 있다. 따라서 'VLAAIL'과 'HHGAAHHCL' 에지에 가중치를 1 올린다. 또한, 'VLAAIL'과 'HHGAAHHCL'은 <표 4>의 RMI에서 'P 31946'이라는 같은 단백질의 조각인 것으로 확인되므로 에지의 가중치를 1 올린다. 'VLAAIL' 기준노드와 다른 파티션의 모든 노드들과의 매칭이 끝나면 다른 기준을 선택하여 위 과정을 반복한다.

〈표 5〉 브리지 서열을 가지고 있는 에지

에 지		브리지 서열
VLAAIL	HHGAAHHCL	AILHHGA
LVAAIL	HHGAAHHCL	AILHHGA
HHGAAHHCL	HMGGMK	GAAHHCLHMGGMK

(그림 2)는 매칭이 끝난 후 가중치가 결정된 다중 분할 그래프이다. 타겟 단백질의 후보 서열 결정의 문제는 다중 분할 그래프에서 각 파티션을 반드시 한번씩 경유하는 경로를 구하는 문제인 해밀토니언 경로(Hamiltonian path) 문제라고 할 수 있다. 따라서 NP-complete한 문제이므로 경험적인 접근 방법인 탐욕(greedy) 알고리즘을 이용한다[11-13]. 각 파티션을 잇는 에지 중에 가중치가 최대인 노드를 연결하여 전체 파티션을 경유하는 경로를 구하면 그 최장 거리 경로가 정렬 가능 정도가 가장 높다고 보고 타겟 단백질의 후보 서열로 결정하는 것이다. 따라서 위 예제에서는 최장 거리 경로(Maximum Weighted Path)를 지나는 노드의 서열을 병합한 'VLAAILHHGAAHHCLHMGGMK'이 타겟 단백질의 후보 서열이 된다.



(그림 2) 완성된 다중 분할 그래프

4. 실험 및 결과 분석

본 논문에서 제안한 다수 개 절단 효소의 사용과 다중 분할 그래프(multi-partite graph)에 의한 단백질 서열 정렬 알고리즘의 타당성을 검증하기 위해 시뮬레이션 실험을 하였다. Pentium Ⅲ 500MHz Dual Processor, RAM 512M의 실험 환경에서, 실험 데이터는 SWISS-PROT Release 40.26의 112892 엔트리 중 서열 길이 25정도 되는 68개의 단백질 서열을 임의로 선택하여 미지의 단백질(unknown protein)로 가정하여 사용하였다. 절단 효소로는 trypsin, chymotrypsin, thermolysin, pepsin, elastase, endoproteinase Glu-C를 사용하였고 기준 그룹은 chymotrypsin에 의해 절단된 조각들을 선택하였다. 다중 분할 그래프를 구축한 후 단백질 후보 서열을 결정하기 위해 경로 탐색을 할 때, 탐욕(greedy) 알고리즘과 모든 경로를 탐색하는(exhausitive) 알고리즘으로 각각 실험을 하여 비교하였다. 실험 결과에 의하면 가중치의 합이 가장 큰 경로의 후보 서열이 미지의 단백질이라고 가정하고 실험 데이터로 입력했던 실제의 서열과 대부분 일치하였다. 따라서 가중치 최대인 경로를 탐색하여 단백질 서열을 결정짓는 탐욕(greedy) 알고리즘의 타당성을 확인할 수 있었다. 실험의 결과는 부록으로 첨부하였다. 다음에서 가중치와 효소의 수에 따라 단백질 서열을 얼마나 정확하게 예측하는지를 분석한다.

4.1 가중치와 단백질 서열 유사도의 관계

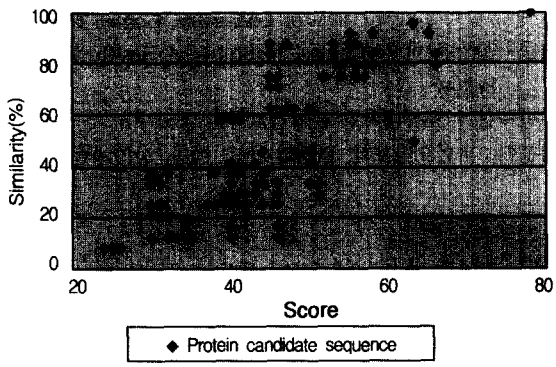
복수 개 효소에 의해 절단된 조각들의 후보 서열들에 다중 분할 그래프(multi-partite graph) 알고리즘을 적용하여 산출된 단백질 후보 서열과 미지의 단백질로 가정하고 입력한 실제 단백질 서열을 비교해 본 결과 (그림 3)에서 보이는 바와 같은 상관 관계를 보였다. 이는 다음과 같이 해석할 수 있다.

가중치의 합(score)이 높을수록 실험으로 예측한 서열과 실제 서열의 유사성(similarity)이 높다는 것을 알 수 있다. 서열의 유사성 측정은 Needleman과 Wunsch의 알고리즘을 이용하였다[14, 15].

다음은 실험 결과 출력된 단백질 후보 서열과 가중치의 합,

유사도를 의미한다. 가중치의 합이 후보 서열 중 가장 높은 (1) 서열은 Needleman & Wunch의 알고리즘에 의해 실제 서열과의 유사성 정도를 계산한 결과, 25/25 즉 1로써 예측한 서열과 실제 서열이 완전히 일치함을 보여준다. 다음에서 실제 서열과 실험에서 예측한 서열을 비교해 본다. 가중치가 가장 높은 77인 서열은 100%의 유사성을 보여 주고 있어 예측한 서열과 실제 서열이 정확하게 일치하고 있음을 알 수 있다.

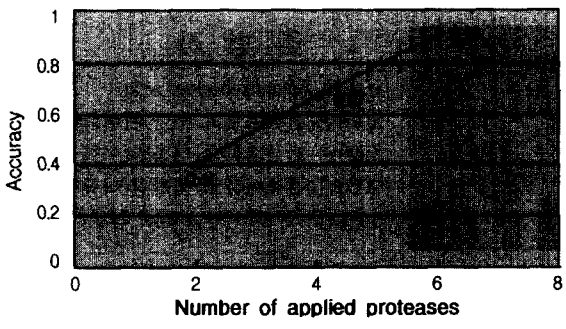
- 실제 단백질 서열
  : GLLSSLSSVAKHVLPHVVPVIAEHL
- 후보 단백질 서열
  GLLSSLSSVAKHVLPHVVPVIAEHL
  //실제 단백질 서열과 일치함
  77, 25/25
  IGLSSLSSVAKHVLPHVVPVIAEHL
  63, 23/25
  GLSSLSSVAKHVLPHVVPVIAEHLI
  58, 20/25
  GLSSLSSVAKHVLIPHVVPVIAEHL
  66, 20/25
  ...



(그림 3) 가중치와 단백질 서열 유사도의 관계

4.2 효소의 수와 단백질 서열 예측의 정확도 관계

단백질 서열을 절단하기 위해 사용된 효소의 수에 따른 단백질 서열 예측의 정확성 정도를 측정해 본 결과, (그림 4)에서 보이는 바와 같은 상관 관계를 보였다. 이는 다음과 같이 해석할 수 있다.



(그림 4) 사용된 효소수와 단백질 서열 예측의 정확성 관계

사용된 효소의 수가 많을수록 단백질 서열을 정확하게 찾는다는 것을 알 수 있다. 효소를 6개 사용했을 때는 91%까지 서열 예측을 정확하게 할 수 있음을 알 수 있다. 이는 절단 부위가 다른 여러 개의 효소를 사용함으로써 생기는 브리지 서열의 수가 증가한 만큼 서열 예측의 정확성을 증가시킴을 확인하는 것이다.

5. 결 론

단백질 동정과 서열 예측은 단백질의 특성과 기능 연구에서 매우 중요한 분야라고 할 수 있다. Protein Expression Profiling을 비롯한 기존의 생화학적 기법에 의한 단백질 동정은 단백질 데이터베이스에 등록되어 있는 계산된 질량과 비교하는 과정에서 비슷한 질량을 가진 후보 서열이 너무 많이 발생하고 더구나, 단백질 데이터베이스에 아직 등록되지 않은 새로운 단백질은 예측할 수 없다는 한계가 있다.

본 논문에서는 이러한 문제를 해결하고자 데이터마이닝 기반 단백질 서열 예측 기법을 제안하였다. 보통의 MALDI-MS 장비를 사용하여 단백질 서열을 동정 및 예측할 수 있는 효과적이고 경제적인 기법이라고 할 수 있다. 복수개의 단백질 절단효소(protease)를 적용함으로써, 서로 중첩된 단백질 조각을 얻어내고, 각 조각의 질량 정보와 단백질 데이터베이스를 이용하여 후보 서열을 식별하였다. 또한, 생긴 후보 서열의 조립을 통해 전체 서열을 결정하기 위해서, 다중 분할 그래프 구축 및 경로 탐색 기법을 제안하였다. 제안한 알고리즘은 타당성을 입증하기 위해서 SWISS-PROT 단백질 데이터베이스에 있는 알려진 단백질 서열로 시뮬레이션 실험을 하였다. 본 논문의 알고리즘을 적용하여 산출된 단백질 후보 서열과 미지의 단백질로 가정하고 입력한 실제 단백질 서열을 비교한 결과 대부분 100% 유사성을 보여 알고리즘의 타당성을 확인할 수 있었다.

본 기법은 현재 실험실에서 가장 많이 사용되고 있는 MALDI-MS 장비로 단백질 동정의 문제를 해결하는 실용적인 접근 방법이다. 예측한 단백질이 새로운 것이면 에드먼 감성법 같은 실험적 방법에 의한 실질적인 단백질 시퀀싱에 유용한 지침이 될 수 있다. 또한, 알려진 서열로부터 각 조각을 예측하기 때문에 타겟 단백질의 부분적 기능에 대한 단서를 제공할 수도 있다. 그러나 단백질 데이터베이스에 비해 조각 데이터베이스(RMI)가 부족하여 서열 예측을 제대로 하지 못하는 경우가 간혹 있었으며, 현재 단백질 동정 과정에서 공통적으로 겪는 후보 서열의 수 때문에 발생하는 시간적, 공간적 복잡성에 대한 한계를 보였다. 향후 후보 서열의 수를 줄여 복잡도를 줄일 수 있는 방향으로 연구를 계속하여야 할 것이다.

참 고 문 헌

[1] M. Mann and M. Wilm, "Error-Tolerant Identification of Peptides in Sequence Data-bases by Peptide Sequence Tags," Anal. Chem, 66, pp.4390-4399, 1994.

[2] A. Shevchenko et al., "Linking Genome and Proteome by Mass Spectrometry : Large-Scale Identification of Yeast Proteins from Two Dimensional Gels," Proc. Nat'l Acad. Sci, 93, pp.14440-14445, 1996.
[3] D. N. Perkins et al., "Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data," Electrophoresis, 20, pp.3551-3567, 1999.
[4] M. Wilm et al., "Femtomole Sequencing of Proteins from Polyacrylamide Gels by Nano-Electrospray Mass Spectrometry," Nature, 379, pp.466-469, 1996.
[5] G. Neubauer et al., "Mass Spectrometry and EST-Database Searching Allows Characterization of the Multi-Protein Spliceosome Complex," Nature Genetics, 20, pp. 46-50, 1998.
[6] John M. Ward, "Identification of Novel Families of Membrane Proteins from the Model Plant Arabidopsis Thaliana," Bioinformatics, 17, pp.560-563, 2001.
[7] Daniel C. Liebler, "Introduction to Proteomics," Humana Press, 2001.
[8] Edmon de Hlffmann, "Tandem Mass Spectrometry : a Primer," Journal of mass spectrometry, Vol.31, pp.129-137, 1996.
[9] Andrew A. et al., "A role for Edman degradation in proteome studies," Electrophoresis, 18, pp.1068-72, 1997.
[10] Ting Chen, "Gene-Finding via Tandem Mass Spectrometry," The ACM-SIGACT Fifth Annual International Conference on Computational Moledular Biology (RECOMB01), pp. 85-92, 2001.
[11] Daniel H. Huson et al., "The Greedy Path-Merging Algorithm for Sequence Assembly," RECOMB, pp.157-163, 2001.
[12] R. M. Idury and M. S. Waterman. "A New Algorithm for DNA sequence assembly," Journal of Computational Biology, 2, pp.291-306, 1995.
[13] Pavel A. Pevzner and Haixu Tang, "Fragment assembly with double-barreled data," Bioinformatics, 17, pp.225S-233S, 2001.
[14] Needleman, S. B. and Wunsch, C. D., "A general method applicable to the search for similarities in the amino acid sequence of two proteins," J. Mol. Bilo, 48, pp.443-453, 1970.
[15] Gusfield, D., "Algorithms on Strings, Trees, and Sequences," Cambridge University Press, 1997.

조 순 이

e-mail : sun2cho@hitel.net
1983년 전남대학교 회계학과 경영학사
1998년 전남대학교 전산통계학과 이학석사
2002년 전남대학교 전산통계학과 박사
관심분야 : 퍼지 데이터베이스, 데이터 마이닝, 바이오정보학

이 도 헌

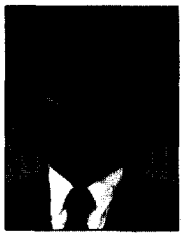
e-mail : dhlee@dbcore.chonnam.ac.kr
1990년 한국과학기술원 전산학과 공학사
1992년 한국과학기술원 전산학과 공학석사
1995년 한국과학기술원 전산학과 공학박사
1996년~2002년 전남대학교 전산학과 및 의학과 조교수
1999년~2000년 미국 Univ. of Texas at Austin, 방문교수
2001년~현재 ACM Transactions on Internet Technology, Associate Editor
2001년~현재 한국데이타마이닝학회 이사
2002년 미국 National Institute of Health, 방문연구원
2002년~현재 한국과학기술원 바이오시스템학과 부교수
관심분야 : 바이오정보학, 데이터마이닝, 데이터베이스

조 광 휘

e-mail : khchol@ssu.ac.kr
1989년 숭실대학교 화학과 이학사
1992년 숭실대학교 화학과 이학석사
2000년 Department of Chemistry, Cornell University, Ph.D.
2000년 Postdoctoral Fellow, Department of Chemistry, Cornell University
2001년~2003년 미국 National Institutes of Health, Research Fellow
2003년~현재 숭실대학교 생명정보학과 조교수
관심분야 : 바이오정보학, 단백질 구조 예측, 신약개발

원 용 관

e-mail : ykwon@chonnam.ac.kr
1986년 한양대학교 전자공학과 학사
1991년 Univ. of Missouri, 컴퓨터공학 석사
1995년 Univ. of Missouri, 컴퓨터공학 박사
1991년~1995년 Univ. of Missouri, Research/Teaching Assistant
1995년~1996년 한국전자통신연구원
1996년~1999년 한국통신 연구개발본부
1999년~현재 전남대학교 정보통신공학부/의학부 부교수
2002년~현재 한국생물정보학회 Proteome Informatics 연구회장
관심분야 : 네트워크관리 및 보안, 컴퓨터비젼, 바이오정보학



김 병 기

e-mail : bgkim@chonnam.ac.kr
1978년 전남대학교 수학과(이학사)
1980년 전남대학교 대학원(이학석사)
2000년 전북대학교 대학원(이학박사)
1981년~현재 전남대학교 컴퓨터정보학부 교수
2002년 한국정보처리학회 부회장, 한국정보처리학회 소프트웨어공학연구회 위원장, 광주정보문화 진흥원 이사
관심분야 : 소프트웨어공학, 객체지향시스템, 지역정보화