N-그램 기법을 이용한 단백질데이터베이스 색인
N-gram Indexing for Protein Sequence Database

김진숙[1], 황미녕[1], 안설아[1], 손현석[1,2]
[1]한국과학기술정보연구원 바이오인포매틱스센터
대전 유성구 어은동 52번지
{jinsuk, mnhwang, snowy, ssh}@kisti.re.kr

[2]서울대학교 보건대학원
서울특별시 종로구 연건동 28번지

요 약

단백질과 유전자 서열 데이터베이스의 크기가 급격히 증가하면서 생물학자들이 서열 검색 시스템을 사용하는데 많은 시간이 소모되고 있다. 그러나 정보 기술의 발달에 맞추어 대규모 텍스트 데이터베이스에서 문자열 검색을 위한 정보 검색 알고리즘이 개발되고 있고, 그 결과 또한 상당히 성공적이다. 본 논문에서는 이런 알고리즘을 생물학적 데이터에 적용시키는 방법을 제시한다. 4가지의 n-gram 색인 기법(tri-gram, tetra-gram, penta-gram, hexa-gram)으로 PIR-NREF 데이터베이스의 단백질 서열을 색인화하고, 검색의 정확성과 속도를 측정하였다. Penta-gram 방법이 BLASTP와 비교하여 검색의 정확도가 비교적 양호하였고, 검색 속도는 38배 정도 더 빠르게 나타났다.

Abstract

Though the sequence databases of proteins and DNAs are increasing in size exponentially, still exhaustive sequence search systems are commonly used in conducting biological researches. However, due to the advancement of information technologies, many information retrieval algorithms have been developed to search strings in large-scale text databases and are proved to be successful. We propose that these algorithms could also be applied to the biological data. Four n-gram indexing methods (tri-gram, tetra-gram, penta-gram, and hexa-gram) were applied to extract indexes from protein sequences of the PIR-NREF database, and their retrieval effectiveness and speed were measured. Penta-gram method showed the best results that its retrieval effectiveness matches for BLASTP and its retrieval speed was about 38 times faster than BLASTP program. Our protein sequence search service is accessible at http://proses.kisti.re.kr.[1]

[1] 본 연구는 한국IBM의 일부 지원 하에 수행되었음

1. 서론

단백질과 DNA의 서열 해독이 늘어남에 따라, 새로이 밝혀진 서열과 유사하거나 상동성이 높은 서열을 기존의 데이터베이스에서 검색해 내는 일의 중요성이 부각되고 있다. 서열 사이의 상동성을 찾아내는 일은 단백질이나 DNA의 기능을 밝히는 중요한 단서가 되기도 한다. 현재 많이 사용되는 유전자 및 단백질 서열에 대한 검색시스템은 질의 서열과 데이터베이스에 있는 각각의 서열 간의 부분 정렬 점수(local alignment score)를 계산하여 이 점수를 기준으로 결과를 보여준다. 이런 소모적인 검색 방법의 시간 복잡도는 질의어와 검색 대상이 되는 데이터베이스의 서열의 총 개수뿐 만 아니라 각각의 서열의 길이에도 큰 영향을 받기 때문에 계산량이 매우 커지게 된다. 따라서 사용자가 인내할 수 있는 시간 내에 검색 결과를 보여줄 수 있는 새로운 방법이 필요하다.

생물학적 데이터베이스의 크기가 급속하게 증가함에 따라 검색 시간의 효율성 문제는 심각하게 대두되고 있다. 기존의 부분 정렬 알고리즘이 내포하는 계산량 부하를 줄이기 위해 경험적인(heuristic) 검색 알고리즘에 대한 관심이 높아지게 되었다. 이런 방법 중 하나가 후보 서열을 미리 뽑아놓고, 이들 후보군과 질의어간에 부분 정렬을 자세하게 계산하여 최종 결과를 찾는 방법이다. 이 방법 중 대표적인 것으로 FASTA는 *k-tuple* 방법론을 사용하는데, 여기서 *k*는 서열에서 추출한 고정된 작은 정수로서 *k-tuple*은 부분문자열 단위가 된다. FASTA의 *k-tuple* 방법론은 2단계를 거치게 된다 (Lipman *et al.*, 1985). 첫번째 단계로 입력 질의어와 데이터베이스의 서열에서 동일한 k-tuple 문자열을 찾아낸다. 그리고 2단계로 각각의 부분 서열들 사이에서 가능한 부분 정렬을 계산하게 된다. 이 *k-tuple* 방법의 보다 발전된 형태가 BLAST(Basic Local Alignment Search Tool)(Altschul, 1991)로 현재 가장 많이 사용되고 있는 생물학적 서열의 검색 방법이다. 이와는 다르게, 정보 검색론의 접근 시각에서 FASTA의 1단계 방법을 n-gram 색인의 역파일 사용으로 확장시킨 연구 방법도 있다(Williams *et al.*, 2002). 이 연구의 결과물이 CAFE라는 원형 시스템으로, 이 시스템 또한 2단계 검색 방법을 채택하고 있다. 첫 번째 단계는 거친 검색(coarse search) 단계로 입력 서열의 유사성을 보여주기 위해 서열의 부분을 선택하는 역색인(inverted index) 방법을 사용한다. 두 번째 단계는 고운 검색(fine search) 단계로 1단계에서 나온 결과 서열들의 순위를 매긴다. 이 CAFE 시스템은 BLAST와 비교했을 때 속도와 검색 효율성에서 뛰어나다고 한다.

이와는 독립적으로, 전산학 분야에서, “정보 검색” 시스템은 대용량의 문서 데이터베이스에서 효율적이고 적합한 검색 방법론으로 인정 받아 왔다(Salton *et al.*, Witten *et al.*, 1999). 이 연구는 원문(full-text) 색인 시스템과 인터넷 정보 검색 분야에 성공적으로 적용되었다. 반면 생물학적 데이터베이스 검색 도구들은 생물 서열 데이터베이스에서 질의 서열과 가장 유사한 서열들을 찾아내는 것이 목표이다. 서열 데이터베이스에서 유사 서열을 찾는 것과 문서 데이터베이스에서 유사한 문서를 찾아내는 것은 그 개념과 방법론에서 보면 매우 유사한 작업이 될 수 있고, 실제로 CAFE 시스템의 예에서 그 가능성을 엿볼수 있다. 생물학적

서열을 DNA 염기나 아미노산 부호로 쓰여진 문서로 간주하고, 알맞은 색인 방법을 사용하면, 정보검색 방법론을 생물학적 데이터베이스의 검색 방법에 적용할 수 있다. 본 연구에서는 n-gram 방법론을 기반으로 하는 단백질 서열의 색인 기법을 제시하고, 이 방법론을 이용하여 PIR-NREF 데이터베이스의 검색을 수행함으로써 n-gram 색인 방법의 검색 속도와 그 효용성에 대해서 논의하고자 한다.

2. 실험방법

2.1 시스템 환경

본 연구에서는 단백질 서열 데이터베이스의 정보 검색 시스템을 Berkeley DB라는 임베디드 데이터베이스(Loverso *et al.*, 2002)와 C++ STL(Standard Template Libraries)로 구현하였다. 시스템의 이름은 ProSeS(Protein Sequence Search)로 명명되었으며, FASTA 형식의 단백질 서열 데이터로부터 n-gram 색인을 추출하여 각 색인들을 Berkeley DB 시스템에서 제공하는 B+ 트리로 저장하게 된다. N-gram 색인 방법에 대해서는 2.4절에서 상세히 다루도록 한다.

모든 실험은 펜티엄-IV 2.4GHz의 듀얼 CPU, 3Gb의 시스템 메모리와 Ultra-160 SCSI 하드 디스크의 데스크탑 리눅스 시스템 환경에서 이루어졌다. 테스트 시에 큰 부하가 걸리는 다른 작업은 없도록 하였다.

2.2. 테스트 데이터

생물학적 서열 데이터베이스에서 검색의 효율성을 측정하기 위한 합리적인 테스트 컬렉션을 만드는 것은 쉬운 일이 아니다. Williams *et al.*(2002)는 PIR Super-family 데이터베이스를 테스트 데이터베이스로 선택하였다. PIR-데이터베이스는 전체 정렬(global alignment)과 동일한 의미를 가지는 다중 정렬(multiple alignment) 방법을 사용하여 각각의 서열들을 슈퍼-패밀리로 분류해 놓고 있다(Wu *et al.*, 2003). 그러나 최근의 서열 검색의 연구 경향은 전체 정렬 방법론보다 부분 정렬 방법론으로 초점을 맞추고 있기 때문에(Altschul *et al.*, 1997), PIR 슈퍼-패밀리 데이터베이스를 검색의 유효성을 측정하는데 사용하는 것은 적합하지 못하다. 따라서 본 실험에서는 PIR-NREF 데이터베이스(Wu *et al.*, 2002)를 테스트 데이터베이스로 선정하고, ProSeS와 BLASTP의 성능을 비교하고자 한다. BLAST의 여러 도구 모음 중 하나인 BLASTP는 단백질 서열 검색 프로그램으로 질의 단백질 서열과 가장 부분 정렬이 높은 서열을 찾아준다(Altschul *et al.*, 1997). 실험 당시의 PIR-NREF 데이터베이스는 1.26 공개 버전으로 127만 건의 단백질 서열을 포함하고 있으며, 평균 서열 길이는 317개이고 총 4억 5백만 개의 아미노산으로 구성되어 있었다.

2.3 N-gram 색인

서양어와 같은 문서는 단어나 어구로 구성되는데, 일반적으로 공백(띄어쓰기)에 의해 구

분이 가능하다. 이런 특성을 이용하여, 정보 검색 시스템에서는 일반적으로 문서에서 각 단어 또는 어구를 추출하여 역 파일(inverted file)에 저장한다(Salton *et al.*, 1983, Witten *et al.*, 1999). 그러나 DNA나 단백질 서열과 같은 생물학적인 서열은 공백이 없는 문자열(string)이며, 이를 서양어의 단어나 어구에 해당하는 의미 있는 단위로 구분하는 것은 거의 불가능한 일이다. 이런 상황에 맞는 경험적인(heuristic) 색인 방법이 몇 가지 있는데, 이 중의 하나가 n-gram 토큰 방법론이다. N-gram을 이용한 색인 방법론은 단어의 경계가 없거나 모호한 중국어 문서(Wilkinson, 1998)나 OCR 문서(Harding *et al.*, 1997)에 잘 적용된다. 즉, 생물학적인 서열은 어구의 경계가 없는 문서로 볼 수 있고, 이런 문서의 색인에 잘 적용되는 n-gram 색인 기법을 여기에 적용하는 것이 가능하다.

N-gram은 FASTA 연구 방법론(Lipman *et al.*, 1985)에 언급되는 "*k-tuple*"의 또 다른 명칭으로 BLAST 연구론(Altschul *et al.*, 1997)에서는 "*w-mer*"로 사용되기도 한다. 즉 n-gram은 서열 안의 일정한 간격으로, 이 간격은 길이가 *n*으로 고정된 부분 문자열로 겹쳐지게 된다. 예를 들어 ACEPITCH의 단백질 서열에서 *n*이 4라면, 최종적으로 나오는 n-gram은 ACEP, CEPI, EPIT, PITC, ITCH가 된다.

여기서 표 1에서 보듯이 *n*은 3, 4, 5, 6으로 '길이'는 부분문자열의 고정된 길이를 의미한다. '알파벳 개수'는 구분할 수 있는 아미노산 코드의 개수를 의미하고 '색인어 개수' 단백질 서열 데이터베이스에서 나타날 수 있는 고유한 색인어의 최대값을 의미한다.

표 1. 네 가지 N-gram의 특징

기호	길이	알파벳 개수	색인어 개수
N3-A20	3	20	8,000 (20^3)
N4-A20	4	20	160,000 (20^4)
N5-A20	5	20	3,200,000 (20^5)
N6-A18	6	18	34,012,224 (18^6)

단백질 서열은 20개의 아미노산 코드가 조합된 스트링이다[2]. 트리그램(tri-gram)(N3-A20), 테트라그램(tetra-gram)(N4-A20), 펜타그램(penta-gram)(N5-A20)에서는 20개의 알파벳이 서열을 색인어로 나누기 위해서 사용된다. 그러나 헥사그램(hexa-gram)(N6-18)에서는 18개의 알파벳 문자만 사용한다. 이것은 20개의 알파벳 문자를 헥사그램에 적용하게 되면, 고유한 색인어가 6천 4백만 개나 되어, 현재의 시스템 메모리에 수용할 수 있는 한계를 넘기 때문이다. 이런 문제점을 해결하기 위해서 20개의 아미노산 중 2개를 그와 유사한 아미노산으로 치환시키도록 하였다. BLOSUM62 점수 행렬(Altschul, 1991)에서 가장 높은 점

[2] 실제로 아미노산의 개수는 21이다. 그러나 PIR-NREF 데이터베이스에서 selenocysteine 아미노산은 출현 빈도가 현저히 낮아서 무시할만한 수준이다. B, Z, X의 와일드카드 아미노산 코드들도 낮은 출현 빈도로 인해 이번 실험에서는 빼기로 한다.

수를 보이는 (V, I)와 (F, Y)를 각각 I와 Y로 치환 시킨다. 그 결과 18개의 아미노산으로 만들 수 있는 고유 색인어 수는 약 3천 4백만 개가 되어, 메모리 상에서 작업이 가능하다.

N-gram들은 각각 Berkeley DB 시스템에 역색인으로 저장되게 된다. 역 파일은 검색 키(searchable key)와 저장 위치 리스트(posting list)가 기록된다. 예를 들면 다음과 같다.

"ACEP" 36 (2), 127(3), 1074(1), ...

여기에서 테트라그램 ACEP는 36번째 서열에서 두 번 나타나고, 127번째 서열에서 세 번, 1074번째 서열에서 한 번 나타남을 의미한다. 검색을 위한 색인 사용에서 디스크 작업을 줄이기 위해서 저장 위치 리스트(posting list)는 범용 압축 알고리즘인 gzip으로 압축된다.

2.4 서열 유사성 측정

정보 검색 분야에서 질의어와 대상 문서의 유사성을 측정하기 위한 모델은 여러 가지가 있다. 이들 중에서 가장 많이 연구되고, 널리 사용되는 모델이 벡터 공간 모델이다(Witten *et al.*, 1999). 이 모델에서는 질의어와 검색 대상 문서를 각각 고차원 공간의 벡터로 나타낸다. 그 후에 이들 벡터의 내적을 문서 길이와 관계된 정규화 계수로 나누어 줌으로써 질의어와 문서의 유사성을 측정할 수 있다. 질의 서열 q와 대상 서열 d의 유사성 $(Sim(q, d))$는 다음과 같이 정의한다.

$$Sim(q,d) = \frac{1}{W_d} \cdot \sum_{t \in q \wedge d} (w_{q,t} \cdot w_{d,t})$$

$$w_{q,t} = \log(f_{q,t} + 1) \cdot \log(\frac{N}{f_t} + 1)$$

$$w_{d,t} = \log(f_{d,t} + 1) \cdot \log(\frac{N}{f_t} + 1)$$

$$W_d = \log(1 + \sum_{t \in d} f_{d,t})$$

여기에서 $f_{s,t}$는 서열 s안의 n-gram 토큰의 빈도를, N은 데이터베이스 안의 총 서열의 개수를, f_t는 n-gram 토큰 t가 한 번 이상 나타나는 서열들의 개수를, $w_{s,t}$는 질의 서열 또는 대상 서열 s안의 토큰 t의 가중치를 의미하고, W_d는 대상 서열 d의 길이를 뜻한다.

2.5 성능평가

검색성능을 평가하기 위해 본 연구에서는 PIR-NREF 데이터베이스로부터 100개의 단백질 서열을 임의로 뽑았다. 이들 서열 데이터는 길이가 최소 51에서 1609 아미노산 사이였고 평균 길이는 316이었다. 이들 테스트 질의어와 PIR-NREF 데이터베이스를 상대로 한 BLAST 결과는 본 연구의 웹 사이트에서 자세히 확인할 수 있다. 이들 서열들을 각각 BLAST를 수행하여, 상동성이 높은 결과 서열들을 최대 1000개까지 얻었다. 다른 설정은

그대로 두었으나, E-value는 0.0001로 설정하였다. 이 결과를 n-gram 색인 방법을 이용한 검색의 유효성을 측정하기 위한 참조 테스트 집합(reference test set)으로 삼기로 한다. 그 다음으로, 같은 질의어 테스트 그룹을 ProSeS에서 PIR-NREF 데이터베이스를 대상으로 실험하여 최대 10,000개의 상동성이 높은 서열들을 검색하였다. N-gram 색인 방법의 상대적인 성능 평가를 측정하기 위해서 BLAST 결과에 대한 재현률(recall)과 정확도(precision)를 계산하였다. 이 재현률과 정확도는 정보 검색의 결과의 유효성 평가가 사용되는 측정 방법(Salton *et al.*, 1983)이다. 본 연구에서 정확도(p)는 특정 검색 지점에서의 관계 있는 BLAST 결과의 비율을 의미한다. 즉,

$$p = \frac{\text{BLAST sequences retrieved}}{\text{BLAST and non - BLAST sequences retrieved}}$$

대조적으로 재현률(r)은 특정 검색 지점에서의 전체 BLAST 결과에 대한 관계 있는 BLAST 결과의 비율이다. 즉,

$$r = \frac{\text{BLAST sequences retrieved}}{\text{Total number of BLAST sequences}}$$

4가지 n-gram 색인 방법에 대한 상대적인 결과의 유효성을 11점 평균 재현률-정확도 그래프로 나타낸다. 모든 입력 서열들에 대해서 재현률이 0.0, 0.1, 0.2, ..., 1.0인 11개의 지점에서의 평균 정확도를 그림 1과 같이 볼 수 있다. 이 방법은 각각의 재현률 지점에서 상대적인 유효성을 비교하기에도 적당한 방법이다.

3. 결과 및 토론

3.1 검색 정확성

그림 1은 4가지 n-gram 색인 방법에 따른 평균 재현률-정확도를 그래프로 나타낸 것이다. 그래프에서 알 수 있듯이 헥사그램(N6-A18)과 펜타그램(N5-A20)의 정확도 수치가 다른 두 가지 색인 기법보다 높게 나왔다. 트리그램(N3-A20)의 검색 효율성은 매우 저조하고, 테트라그램(N4-A20)에서는 그 보다는 높게 나왔다. N3-A20, N4-A20, N5-A20, N6-A18의 n-gram 색인 방법에서의 11점 평균 재현률-정확율 값은 0.1376, 0.5038, 0.6342와 0.6337였다. 일반적인 문서의 검색에서의 11점 평균 재현률-정확도의 값이 0.25에서 0.40 범위인 것과 비교한다면(Voorhees, 2002, Oard *et al.*, 2002), N5-A20과 N6-N18의 약 0.63의 수치는 상당히 높은 수치임을 알 수 있다. 이 결과를 토대로 펜타그램과 헥사그램의 색인 기법의 검색 결과가 BLAST의 검색 결과와 상당히 일치한다는 것을 알 수 있다.

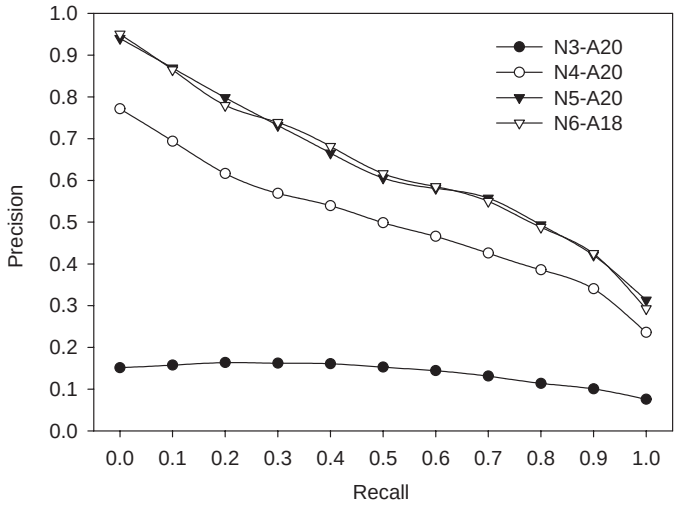

그림 1. 네 가지 n-gram 색인방법에 대한 평균 11점 재현율-정확도 그래프

그림 1에서는 n-gram의 간격의 길어지면, 헥사그램(N6-A18)을 제외하고는 검색의 유효성이 증가한다는 것도 알 수 있다. 앞 절에서 헥사그램은 다른 3가지 색인 방법과 달리 아미노산 코드를 18개만 사용해야 한다는 것을 설명했었다. 만약 헥사그램이 20개의 알파벳 코드를 모두 사용한다면, 검색의 유효성이 조금 더 증가할 것이라고 볼 수 있다. 그러나, 그 증가 폭은 그림 1과 표 2에서 테트라그램과 펜타그램의 평균 재현률-정확율의 차이가 2.8배나 되는 것에 비해서 상대적으로 적을 것으로 추측된다. 이는 n-gram의 간격이 커질수록, 그에 따른 정보도 증가하게 될 것이고 좁은 영역에서의 정렬의 정확도가 감소하기 때문이다. 따라서, 이번 실험으로 n-gram의 길이는 6을 넘어서면 검색의 유효성이 낮아진다는 것을 알 수 있다.

펜타그램에서 재현률이 0.1일 때의 평균 정확도는 0.869이다. 이것은 ProSeS의 검색결과가 재현률이 낮은 범위에서는 BLAST의 검색 결과와 거의 일치한다는 것을 의미한다. 100개의 질의 서열들에 대한 BLASTP 검색 결과의 평균 테스트 집합이 335개가 나왔다. 즉, 재현률이 0.1이면 34개의 서열이 검색되고, 이는 색인 기법이 펜타그램일 때의 ProSeS의 검색 결과에서 30개와(재현률 0.1에서의 34개의 서열의 86.9%) 일치한다는 것이다.

더욱이 ProSeS와 BLASTP의 검색 결과에서 재현률이 0.1, 0.4, 0.7과 1.0일 때 검색결과가 완전히 일치하는 것이 73, 51, 38, 17개가 나왔다. 이것은 ProSeS와 BLASTP가 재현률 0.1에서는 73개의 서열이 동일한 결과로, 0.4일 때는 51개가 동일하게 나타났다는 것이다. 그러나 9개의 질의 서열들은 재현률이 0.1일 때 평균 정확율이 0.1 보다 적게 나오기도 했다. 평균 정확율이 낮게 나온 질의 서열 중의 하나가 97번째 질의어인데, BLASTP로 PIR-NREF 데이터베이스를 검색하면 20개의 서열이 검색되는데 ProSeS에서는 6개만 검색되었다.

```
>NF01265541 Similar to 5'-nucleotidase, cytosolic III (Fragment) [Xenopus
          laevis]
         Length = 294

 Score =  127 bits (320), Expect = 1e-28
 Identities = 96/288 (33%), Positives = 127/288 (44%), Gaps = 79/288 (27%)

Query: 15  PRALTDKMTLIRDAGPSKFQVF-----------------PTP---------ISEQGDAYY 48
           P  L DK+T I+   G  K Q+                  PT          IS++G
Sbjct: 20  PEGLQDKITRIQRGGQEKLQIISDFDMTLSRFSRNGERCPTCYNIIDNSNIISDEGRK-- 77

Query: 49  DAKRQALYDHYHPLEISPVIPIDEKTKLMEEWWGKTHELLIEGGLTYDAIKKSVANSSIA 108
             K + L+D Y+PLEI P    I+EK  LM  EWW K H+L  E  +   D + + V  S
Sbjct: 78  --KLKCLFDIYYPLEIDPKKSIEEKYPLMVEWWSKAHDLFYEQRIQKDRLAQVVKESQAT 135

Query: 109 FREGVSELFEFLEKKEIPVLIFSAGLADVIEEVTLKSISLLELLSYFCCLYNEYAFVAYS 168
            R+G    F   L ++EIP+ IFSAG+ DV+EE
Sbjct: 136 LRDGYDLFFNSLYQREIPLFIFSAGIGDVLEE---------------------------- 167

Query: 169 HSYQVLRQNLDRTFKNVKIVSNRMVFNDDGQLVSFKGKLIHVLNKNEHALDMAAPLHDRL 228
               ++RQ        N K+VSN M F+D+G L  FKG LIH   NKN   L
Sbjct: 168 ----IIRQ-AGVFHPNTKVVSNYMDFDDNGILTGFKGDLIHTYNKNSSVL---------- 212

Query: 229 GVDIGEEDEENVNMKERRNVLLMGDHLGDLRMSDGLD-YETRISIGFL 275
                ++ E    +   R N+LL+GD LGDL M+DG+   E   I IGFL
Sbjct: 213 -----KDTEYFKEISHRTNILLLGDTLGDLTMADGVSTVENIIKIGFL 255
```

그림 2. 97번째 질의 서열에 대한 BLAST 결과의 두 번째 항목. “NF00667350 hypothetical protein At2g38680 [Arabidopsis thaliana]”. 동일한 질의에 대해 ProSeS는 이 서열을 검색하지 못했다.

그림 2는 97번째 질의 서열의 BLASTP 검색 시 두 번째로 나오는 결과 서열과의 일치 정도를 보여주고 있다. 이 결과 서열은 펜타그램의 ProSeS에서는 검색되지 않는 서열이다. 그림 2의 정렬 형태를 주의 깊게 보면, 질의어의 129번째 시작하는 IFSAG만이 유일하게 질의어와 결과 서열간에 동시에 존재하는 펜타그램이다. 이는 ProSeS의 유사성 측정 방법인 벡터 공간 모델에서는 질의서열과 대상 서열간의 일치하는 색인어가 하나 밖에 없기 때문에 유사성의 측정 수치가 낮게 나올 수 밖에 없다. 따라서 이런 서열은 ProSeS,의 검색 결과에서는 낮은 순위로 밀려나게 되는 것이다.

이와는 다르게, 그림 3은 97번째 질의 서열의 BLASTP의 세 번째 결과의 일치 정도를 보여주고 있다. 이 결과 서열은 ProSeS에서도 동일하게 나타난다. 여기에서는 질의 서열의 129번째 위치에서 시작하는 IFSAG와 256번째 위치에서 시작하는 GDLRM, 이 두 개의 일치하는 펜타그램을 찾을 수 있다. 따라서, ProSeS에서는 이 서열이 그림 2의 서열보다 더 높은 유사성 측정 값을 얻게 되는 것이다. 이런 점 때문에 그림 2의 서열보다 그림 3의 서열이 ProSeS의 결과에서는 보다 높은 순위에 있게 되는 것이다.

그림 2와 그림 3의 서열이 BLASTP에서는 두 번째와 세 번째에 나타나지만, ProSeS에서는 그림 3의 서열은 검색되지만 그림 2의 서열은 검색되지 않는다. 이런 사실을 통해서 ProSeS의 유사성 측정 방법론이 BLASTP의 부분 정렬 방법과 차이점을 보인다는 것을 알 수 있다. 따라서, ProSeS에서 사용되는 벡터 공간 모델의 유사성 측정 방법에 대한 개선과 보완이 필요하다.

```
>NF01178199 10 days embryo whole body cDNA, RIKEN full-length enriched
          library, clone:2610024B13 product:HSPC233 (PYRIMIDINE
          5'-NUCLEOTIDASE) (EC 3.1.3.5) (URIDINE 5' MONOPHOSPHATE
          HYDROLASE 1) (SIMILAR TO HYPOTHETICAL PROTEIN) homolog
          [Mus musculus]
         Length = 331

 Score =  124 bits (310), Expect = 2e-27
 Identities = 78/222 (35%), Positives = 118/222 (53%), Gaps = 49/222 (22%)

Query: 55  LYDHYHPLEISPVIPIDEKTKLMEEWWGKTHELLIEGGLTYDAIKKSVANSSIAFREGVS 114
           L + Y+ +E+ PV+ ++EK    M EW+ K+H LLIE G+    +K+ VA+S +   +EG
Sbjct: 122 LKEQYYAIEVDPVLTVEEKFPYMVEWYTKSHGLLIEQGIPKAKLKEIVADSDVMLKEGYE 181

Query: 115 ELFEFLEKKEIPVLIFSAGLADVIEEVTLKSISLLELLSYFCCLYNEYAFVAYSHSYQVL 174
            LF   L++  IPV IFSAG+ DV+EEV  ++                         HS
Sbjct: 182 NLFGKLQQHGIPVFIFSAGIGDVLEEVIRQA--------------------GVYHS---- 217

Query: 175 RQNLDRTFKNVKIVSNRMVFNDDGQLVSFKGKLIHVLNKNEHALDMAAPLHDRLGVDIGE 234
                    NVK+VSN M F+++G L   FKG+LIHV NK++ AL              +
Sbjct: 218 ---------NVKVVSNFMDFDENGVLKGFKGELIHVFNKHDGALK--------------- 253

Query: 235 EDEENVNMKERRNVLLMGDHLGDLRMSDGL-DYETRISIGFL 275
               +   +K+  N++L+GD  GDLRM+DG+  + E  + IG+L
Sbjct: 254 NTDYFSQLKDNSNIILLGDSQGDLRMADGVANVEHILKIGYL 295
```

그림 3. 97번째 질의 서열에 대한 BLAST 결과의 세 번째 항목. “NF01178199 10 days embryo whole body cDNA ... [Mus musculus]”. 이 서열은 ProSeS 또한 검색하였다.

검색의 유효성 측면에서는 펜타그램이 단백질 서열의 검색에 가장 적당한 색인 기법임을 알 수 있었다. 다음 절에서는 각각의 방법에서의 검색 시간을 비교해보도록 하겠다.

3.2 공간과 속도

표 2는 PIR-NREF 데이터베이스를 색인 하는데 걸리는 시간, 색인 파일의 크기, 검색 시간과 11점 평균 정확도가 나타나 있다. PIR-NREF 데이터베이스의 1.26 release 버전에는 약 404,532,594개의 아미노산이 있음을 감안하여 볼 때 색인 작업을 거치고 난 후에는 데이터베이스가 3.7배에서 14.0배까지 확장됨을 알 수 있다. 펜타그램(N5-A20)일 때에는 초기 서열 데이터베이스의 크기보다 7.6배가 색인 공간이 커지지만, 검색 속도는 가장 빨랐다. 비록 현재의 디스크 저장 기술이 급속히 발달한다 할지라도, 단백질 서열의 데이터베이스 크기는 보다 더 급속하게 커지고 있기 때문에 상대적으로 저장 공간에 대한 부하가 존재하게 된다. 뉴클레오타이드 데이터베이스의 색인 압축 기술(Williams *et al.*, 1997)이나 매우 빈번한 n-gram은 색인을 하지 않는 *index stopping* 방법론(Williams *et al.*, 1996) 등을 ProSeS에도 잘 적용하면 좀 더 효율적인 결과를 보일 것으로 생각된다.

표 2. 네 가지 색인 방법과 BLAST의 색인크기, 색인시간, 검색속도, 11점 평균 재현율-정확도.

N-gram	역파일(MB)	색인 시간(hr)	검색시간(초)	11점 정확도
N3-A20	1,514	1.91	16.70	0.1376
N4-A20	3,125	2.40	3.88	0.5038
N5-A20	3,075	3.97	1.17	0.6342
N6-A18	5,655	8.94	7.61	0.6337
BLAST	583	0.05	44.10	-

그림 4에서 보듯이 색인 시간은 n-gram의 고유 색인어 수에 크게 영향을 받음을 알 수 있다. 즉 n-gram의 길이가 길수록 색인화 시간은 훨씬 더 소요된다. 색인화 시간이 몇 시간씩 소요됨에도 불구하고, 색인화 작업은 시스템 관리자 권한이기 때문에 다른 서열 검색 작업을 하는 다중사용자에 비해 작업 관리 측면에서 우선 순위를 부여 받을 수 있다. 또한, 역 파일 리스트를 관리하는 메모리 구조를 C++ STL 라이브러리를 사용하지 않고, 보다 효율적인 방법으로 구현하면 시간 절약이 가능할 것으로 보여진다.

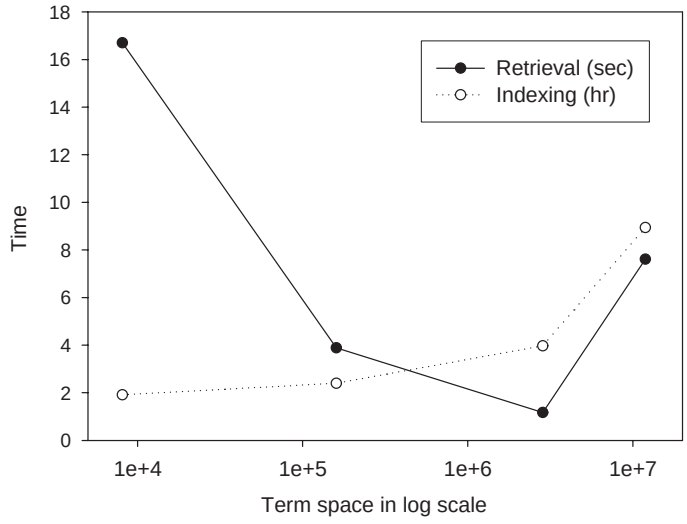

그림 4. 네 가지 n-gram 색인방법의 고유색인어 수에 대한 색인 시간 및 검색 속도 그래프. 고유색인어의 수(Term Space)는 로그 척도로 표시하였다. 그래프 상의 각 원은 왼쪽부터 오른쪽으로 N3-A20, N4-A20, N5-A20, N6-A18를 나타내며 각각의 고유 색인어 수는 8.0×10^3, 1.6×10^4, 2.8×10^6, 1.2×10^7이다.

다른 한편으로, 검색 속도는 재미있는 경향을 보여준다. N-gram의 간격 길이가 3에서 5로 증가하면, 검색 속도는 매우 빨라짐을 알 수 있다(그림 4). 그러나 n이 6이 되면, 검색 속도는 느려져 버린다. 이는 색인 구조의 크기가 펜타그램에 비해 10배나 크기 때문에 많은 부하가 걸리기 때문이다. 펜타그램으로 색인 했을 때 하나의 질의 서열을 검색하는데 걸리는 시간은 평균 1.17초가 걸린다(표 2). 이것은 BLASTP에 비해서 38배 빠르다. 이 결과를 통해서 펜타그램으로 색인 했을 때가 속도 면에서 가장 효율을 보장할 수 있다는 것을 알 수 있다.

4. 결론

본 논문에서는 n-gram 색인 기법을 이용해서 단백질 서열을 검색하는 방법에 대해서 논하였다. 특히, 펜타그램일 때가 속도와 검색의 유효성 면에서 가장 효율적임을 알 수 있었다. 전통적인 검색 방법론인 n-gram 색인 방법론의 성능이 좋은 편이지만, 부분 정렬 사이의 공백(gap)을 처리하는 방법과 단백질 서열 사이의 유사성을 측정하는 벡터 공간 모델에 대한 연구가 더 필요하다. N-gram 색인 기반의 검색시스템은 다른 응용 프로그램의 확대가 쉽다는 장점을 가지고 있다. ProSeS,의 경우에는 간단한 데이터 마이닝 기반의 *keyword suggestion* 기능과 텍스트 분류 알고리즘에 기반한 *subcellular localization* 기능, *super-family classification* 기능을 추가로 제공한다.

현재의 시간 소모적인 비교 방법(exhaustive comparison system)을 사용하는 시스템은 검색 시간 측면에서 한계에 직면하고 있다. 생물학적 데이터베이스가 날이 갈수록 기하급수적으로 커지는 미래에는 이런 시스템을 보기 힘들게 될 것이다. 본 연구에서 제안하는 색인 기반의 서열 검색 시스템인 ProSeS가 기존의 생물학적 데이터베이스 검색 도구를 대체 또는 보완하는 실질적인 하나의 대안이 될 수 있을 것이다.

참고문헌

Altschul, S. F. (1991) Amino Acid Substitutions Matrices from an Information Theoretic Perspective. *J. Mol. Biol.*, **219**, 555-665

Altschul, S. F., Madden, T. L, Schaffer, A. A., Zhang, J., Zhang, Z., Muller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research*, **25**, 3389-3402

Harding, S. M., Croft, W. B. and Weir, C. (1997) Probabilistic Retrieval of OCR Degraded Text Using N-grams. *European Conference on Digital Libraries*, 1997, 345-359

Lipman, D. J. and Pearson. W. R. (1985) Rapid and Sensitive Protein Similarity Searched. *Science*, **227**, 1435-1441

Loverson, S. and Seltzer, M. (2002) Tree Houses and Real Houses: Research and Commercial Software, *Proceedings of the 2nd Workshop on Industrial Experiences with Systems* Software (WIESS'02*),* Berkeyley, CA: USENIX Association, December 2002, pp.55-66

Oard, D. W and Gey, F. C. (2002) The TREC 2002 Arabic/English CLIR Track. *NIST Special Publication 500-251: The Eleventh Text REtrieval Conference (TREC 2002)*, 2002

Salton, G. and McGill, M. J. (1983) *Introduction to Modern Information Retrieval,* McGraw-Hill, New York, 1983

Voorhees, E. M. (2002) Overview of TREC 2002. *NIST Special Publication 500-251: The Eleventh Text Retrieval Conference (TREC 2002)*, 2002

Wilkinson, R. (1998) Chinese Document Retrieval at TREC-6. *NIST Special Publiction 500-240: The Sixth Text Retrieval Conference(TREC 6)*, 25-29

Williams, H. E. and Zobel, J. (1996) Indexing Nucleotide Databases for Fast Query Evaluation. *In Peter M. G. Apers, Mokrane Bouzeghoub, and Georges Gardarin, Proc. International Conference on Advances in Database Technology (EDBT)*, Avignon, France, 275-288

Williams, H. E. and Zobel, J. (1997) Compression of Nucleotide Databases for Fast Searching. *Computer Applications in the Biosciences*, **13**, 549-554

Williams, H. E. and Zobel, J. (2002) Indexing and Retrieval for Genomic Databases. *IEEE Transactions on Knowledge and Data Engineering*, **14**, 63-78

Witten, I. H., Moffat, A., and Bell, T. C. (1999) *Managing Gigabytes: Compressing and Indexing Documents and Images*, The Second Edition, Morgan Kaufmann Publishing, San Francisco, 1999

Wu, C., Huang, H., Arminski, L., CastroAlvear, J., Chen, Y., Hu, Z., Ledley, R. S., Lewis, K. C., Mewes, H.W., Orcutt, B. C., Suzek, B. E., Tsugita, A., Vinayaka, C. R., Yeh, L. S., Zhang, J. and Barker, W. C. (2002) The Protein Information Resource: An Integrated Public Resource of Functional Annotation of Proteins. *Nucleic Acids Research,* **30**, 35-37

Wu, C. H., Huang, H., Yeh, L.-S. L., and Barker, W. C. (2003) Protein Family Classification and Functional Annotation. *Computational biology and Chemistry*, **27**, 37-47