NCBI의 소개 (II)
BLAST

Moo-Sang Kim
Department of Aquatic Life Medicine, Pukyong National University, Busan 608-737, Korea

I. BLAST

BLAST (Basic Local Alignment Search Tool)는 뉴클레오타이드 데이터베이스 (nucleotide database)와 단백질 데이터베이스 (protein database)에 있는 자료들로부터 query와 유사한 서열을 찾기 위한 신속한 검색 방법을 제공하는 NCBI에서 이용할 수 있는 하나의 tool이다. 서열비교 혹은 서열정렬 (sequence alignments) 분석의 목적은 관심 있는 서열의 유사성과 차이점을 분석하여 염기와 아미노산 수준에서 서열간의 구조적, 기능적 및 진화론적 관련성을 추론하는 것에 있다. 즉, 잘 제작된 query와의 정렬들로부터 유전자의 구조적, 기능적 정보뿐만 아니라 진화적인 정보도 추출할 수 있다는 이야기이다. BLAST에서 사용하는 알고리듬은 global alignment뿐만 아니라 local alignment도 탐지하기 때문에, 관련이 없는 단백질 내에 파묻혀 있는 유사성 (similarity)의 영역도 탐지될 수 있고, 이들 모든 타입의 유사성은 미지의 단백질의 기능에 대한 중요한 단서를 제공할 수도 있다. 이런 특성으로 인하여, 생물학자들은 BLAST와 같은 서열 정렬 프로그램을 사용하여 짧은 시간 안에 거대한 서열 데이터베이스로부터 query 서열과 일치하는 수십 개의 유사 서열을 찾아낼 수 있게 되었으며, 또한 요즈음은 이 BLAST가 생물학자들이 생물학 데이터베이스를 사용할 때 가장 먼저 접하게 되는 일반적인 도구가 되었다.

아래는 이런 BLAST의 사용법과 그 사용으로부터 얻은 결과를 해석하는데 도움을 주기 위한 정보를 제공하려 한다. 그리고 아래의 내용 대부분은 “NCBI Education의 BLAST information” (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html)의 원문을 참고로 하였음을 공지한다.

1. Algorithm

BLAST에서는 gap을 가지고 있지 않은 정렬의 부분을 High Scoring-Segment Pairs (HSP)라고 한다. BLAST 에 의해 검색되기 위해서는 정렬부분에 cutoff 점수 S보다 큰 점수를 가지는 HSP를 포함하여야 한다. 이 cutoff 점수는 사용자에 의해 변경될 수 있으나, 문제는 어떠한 cutoff 점수를 사용하여야 하는가를 결정하기 쉽기가 않다는 점이다. 이는 Karlin-Altschul의 통계방법에 근거한 것이기 때문인데, cutoff 점수 S대신에 기대 cutoff E (expect ation cutoff E)를 대신 사용하기도 한다. 만일 기대 cutoff 점수를 사용하면 BLAST는 주어진 검색 조건에 근거하여 cutoff S점수를 계산한 후 이를 검색에 이용한다. 그리고 BLAST에 소개된 혁신적인 것은 neighborhood words (확장 단어)의 개념이다. 정확하게 일치하는 단어를 찾는 대신에, 비교되는 데이터베이스 서열의 단어(길이 W)들 중에서 질의서열의 단어(길이 W)들과 비교하여 T 이상의 값을 가지는 단어가 있는가를 점검하는 것이다. 이때 비교하는 단어의 길이 W를 증가시키면, sensitivity를 희생하지 않고 검색 속도를 증가시킬 수 있다. 그러므로 T가 속도와 sensitivity를 결정짓는 가장 결정적인 요소가 된다. 만일 T의 값이 증가되면, 배경 단어 일치 (background word hit , 혹은 noise)는 감소하게 되며 검색속도는 보다 단축되게 된다. T의 값을 감소시키면, 보다 유사하지 않은 서열들이 발견되게 된다. word hit이 발견되면 이로부터 정렬의 좌우로 확장하며 확장되는 단어 (neighborhood word)가 적어도 S인가를 계산하게 된다. BLAST 2.0에서는 gap을 고려하여 좌우로 확장하게 된다. 이때 통계적인 추정방법에 의해 더 확장하여도 원하는 HSP를 발견할 가능성이 적을 때는 계산을 중지하게 된다. 이 방법을 요약하면 다음과 같다 (그림 1과 2).

(1) 질의 서열로부터 3개의 단백질 혹은 11개의 염기로 이루어진 단어들을 구성한다. 이 단어들과 비교하여 T 값보다 큰 값을 가지는 단어들의 list를 만든다. 만들어진 list들의 단어들을 비교하려는 데이터베이스내의 서열과 비교한다.
(2) 만일 list들의 단어들과 비교도중 hit이 발생하면, 단어를 확장하여 확장된 단어 (neighborhood word)의 값이 S보다 크거나 같은가를 계산한다 (HSP 여부계산).
(3) HSP를 가지는 서열들을 추출하고 서열 내에 복수개의 HSP가 존재하면 통계적인 방법을 사용하여 이들 HSP를 연결한다.
아주 미약한 유사도를 가지지만, 그러나 중요한 서열을 데이터베이스 내에서 검색하기 위해서 사용하는 방법 중에 profile 검색방법이 있다. 이 방법을 BLAST에 간단하게 적용시킨 방법이 있는데, 이를 PSI-BLAST라 한다. profile이란 conserved protein domain내의 각 position에서의 20개의 아미노산의 빈도수를 가지는 테이블이다. PSI-BLAST에서 profile은 공백상태에서 만들어지면, 그 횟수를 반복할수록 정제된 profile이 만들어지게 된다. 최초에 하나의 질의 서열을 사용하여 일반적인 데이터베이스 검색을 시도한다. 이 초기의 검색 결과에서 아주 현저한 복수 서열 정렬로부터 profile이 만들어지게 된다. 필요하다면 이 과정이 다시 반복되고 새로운 서열이 발견된 정렬로부터 profile을 다시 정제하게 된다.

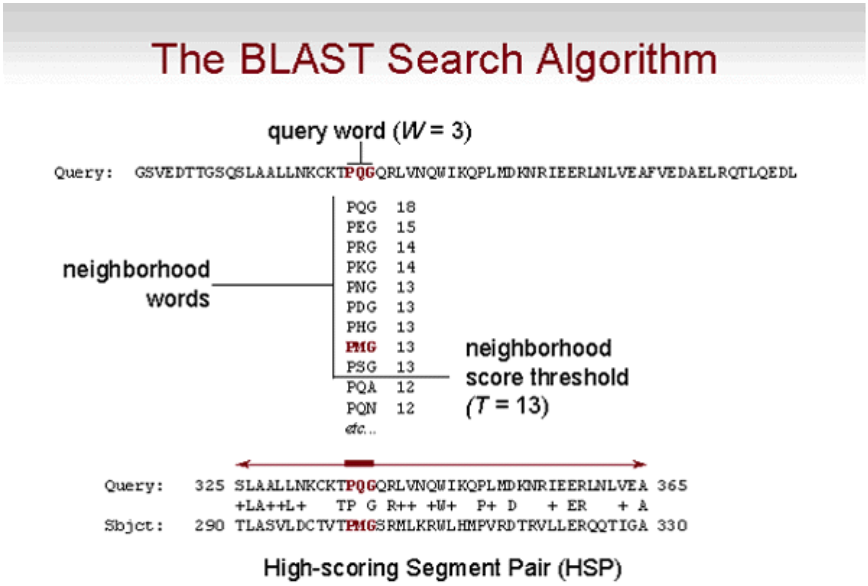

그림 1. BLAST Search Algorithm.

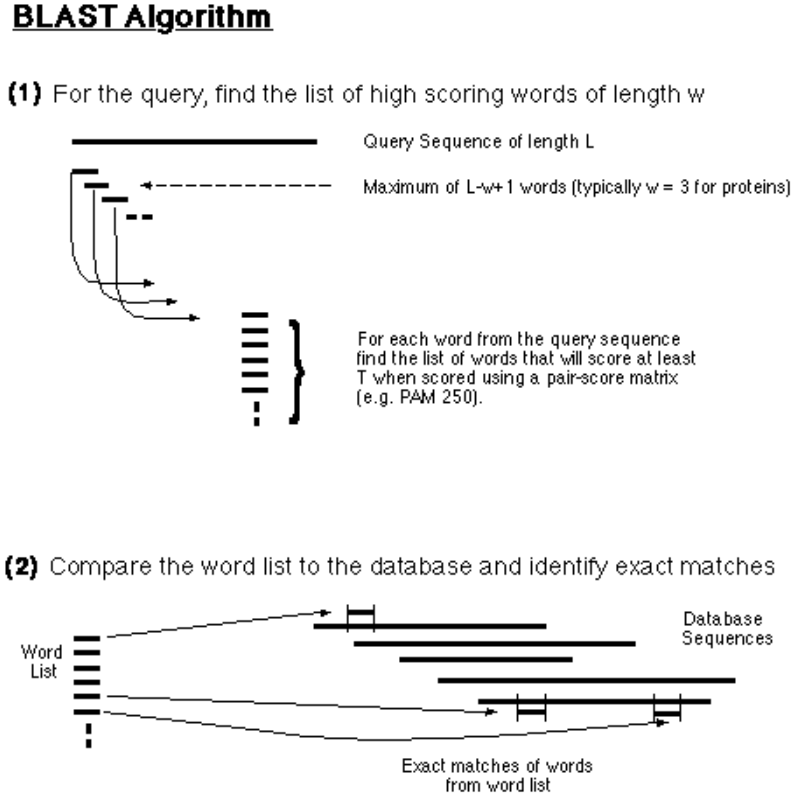

그림 2. BLAST 알고리듬을 이용한 검색 방법.

Figure from Barton, G.J. Protein Sequence Alignment and Database Scanning (University of Oxford, Laboratory of Molecular Biophysics)

BLAST는 quary 서열과 gap 없이 일정 값 이상의 HSP를 기록하지 못하는 서열들을 미리 제거한다. 그래서 FASTA에 비해 훨씬 비교속도가 빠르다. 하지만 두 서열이 특정 부분이 높은 일치성을 가지고 있지는 않지만 대부분의 서열에서 유사성을 가지고 있는 경우에 BLAST는 검색을 해 낼 수 가 없다. Short repeat sequence 나 특정한 residue들이 많이 존재하는 서열들을 query 서열로 이용하였을 경우, 별로 연관성이 없는 서열들이 결과로 나오는 경우도 있다. 이런 결과들을 피하기 위하여 BLAST는 filtering하는 기능을 기본적으로 가지고 있다. 결국 repeat sequence같은 것들은 검색하기 이전에 제거된다는 사실을 기억해야 한다. FASTA와 마찬가지로 BLAST도 단백질 서열을 위해 개발된 프로그램이다. 염기 서열의 검색이 가능하지만 sensitivity가 떨어지므로 염기 서열로 염기 데이터베이스를 검색해 진화적으로 떨어져 있는 서열을 찾고자 한다면 FASTA를 사용하는 더욱 좋은 결과를 얻을 가능성이 높다.

2. BLAST Homepage

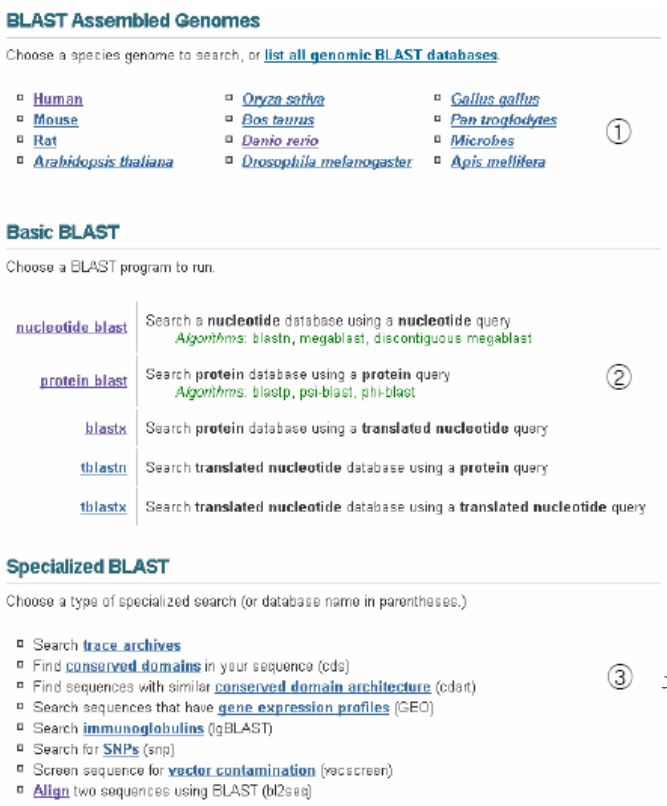

그림 3. BLAST Homepage.

BLAST 홈페이지에 들어가면 위의 그림과 같이 크게 3부분으로 나누어지며 (그림 3), 그것에 대하여 간단하게 아래에 설명하고자 한다.

① BLAST Assembled Genomes: 검색되는 해당 database를 제한 할 수 있다. 즉, 사람, 마우스, 등의 검색하고자하는 genomic BLAST database의 종 list를 볼 수 있으며, 그 외 추가되는 genomic BLAST database를 보기 위에서는 “list all genomic BLAST database"를 클릭하면 볼 수 있으며, 또한 해당 종을 선택하여 실행 할 수 있다.

② Basic BLAST: 기본적으로 BLAST 검색을 실행하기 위한 “BLAST program"을 선택할 수 있다. 아래는 그 선택 가능한 프로그램과 해당 프로그램에 대한 설명이다. 그리고 굵은 글씨체로 표시한 프로그램은 일반적으로 생물과학자들이 자주 사용하는 프로그램을 표시하였다.

Nucleotide blast (BLASTN): query로 입력되는 DNA 서열을 가지고 뉴클레오타이드 sequence 데이터베이스 (nucleotide sequence database)에 대하여 비교하는 프로그램.- 염기 서열간의 비교

Protein blast (BLASTP): query로 입력되는 아미노산 sequence (amino acid query sequence)을 가지고 단백질 sequence 데이터베이스 (protein sequence database)에 대하여 비교하는 프로그램.- 단백질 서열간의 비교

BLASTX: query로 입력되는 DNA 서열이 reading frames으로 변환되어 생긴 단백질 서열을 가지고 단백질 sequence 데이터베이스 (protein sequence database)에 대하여 비교하는 프로그램으로, 미지의 DNA서열로부터 만들어질 수 있는 단백질을 검색하는 경우에 사용한다.- 입력한 염기서열을 6개의 frame으로 변환 후 단백질 서열 database와 비교

TBLASTN: 동적으로 reading frame으로 번역된 뉴클레오타이드 sequence 데이터베이스(nucleotide sequence database)를 대상으로 단백질sequence(protein query sequence)을 비교하는 프로그램.-염기서열 database를 6 frame으로 변환 후 입력한 단백질 서열과 비교

TBLASTX: query로 입력되는 DNA 서열을 6 frame으로 번역하여 6 frame으로 번역된 뉴클레오타이드 sequence 데이터베이스를 대상으로 비교하는 프로그램이며, BLAST 웹페이지에서 balstx프로그램은 nr 데이터베이스와 함께 사용할 경우 계산량이 너무 많으므로 사용 할 수 없다.- 입력한 염기 서열과 염기서열 database를 모두 6 frame으로 변환후 비교

③ Specialized BLAST: 보다 전문화된 BLAST 검색을 위한 링크가 되어 있으며, 염기서열 결정시에 혼입된 vector 서열의 검출이나, Immunoglobulin 서열의 검색, 매우 길고 유사성이 낮은 서열간의 비교가 가능한 MegaBLAST 등과 같은 tool에 대한 링크를 제공한다. 특히, Bl2seq 프로그램은 유사한 서열 두 개를 입력하였을 때 두 서열의 alignment를 수행하는 program 이다.

2.1 검색용 BLAST 프로그램 선택 방법

아래는 생물학자들이 미지의 서열에 대한 검색을 할 때 가장 기본적으로 행하는 방법을 순서적으로 설명하겠다.

①. BLAST 검색 페이지를 연다. (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi)

②. Basic BLAST란의 "Program" 메뉴들 중에서 원하는 프로그램을 선택하여 클릭한다.
(예를 들면, 위 그림 3에서 ②에서 nucleotide blast를 클릭을 한다.)

③ 아래와 같은 BLAST query 입력 페이지가 열린다. 이 페이지에는 크게 5 부분 (A-E)으로 나누어지며, 그 각 부분에 대한 보다 자세한 설명은 아래에 하겠다.

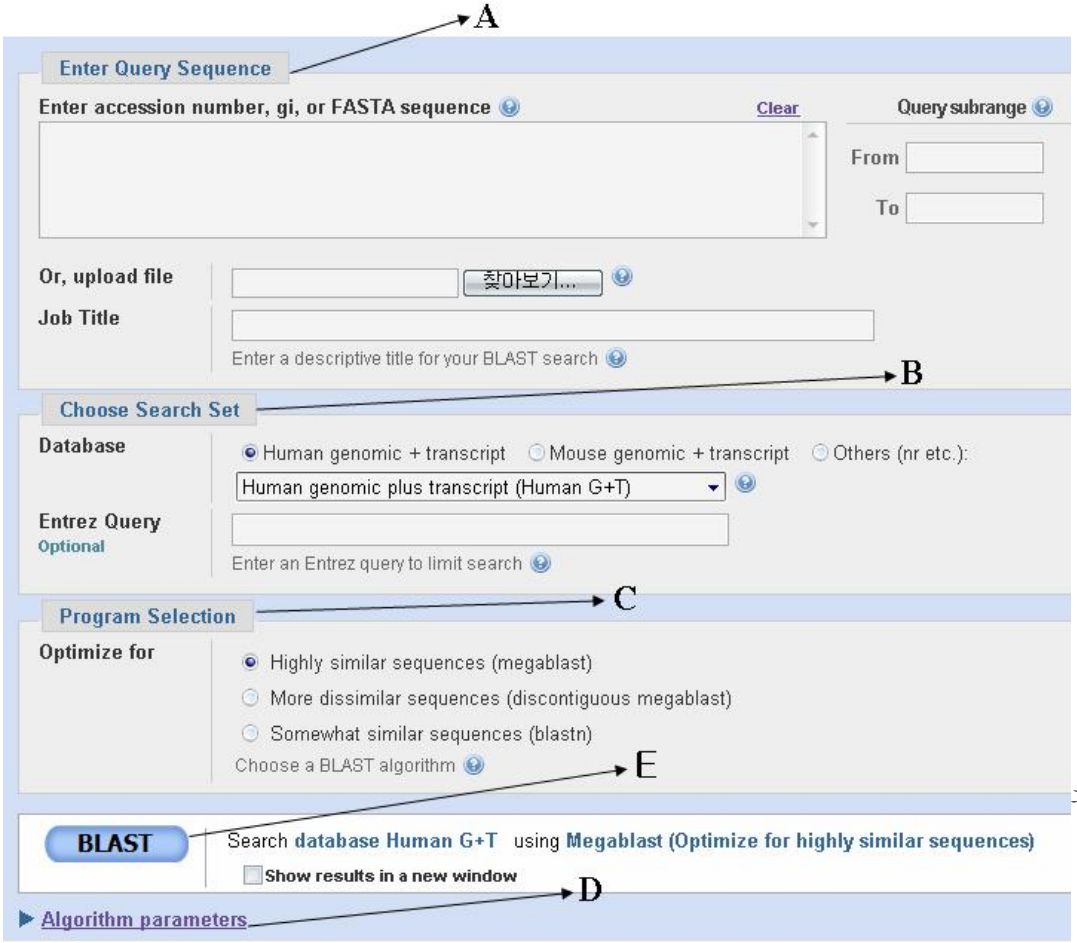

그림 4. Nucleotide BLAST.

A. Enter Query Sequence: 이란에는 내가 질의하고자하는 query 서열을 입력할 수 있다. 그 입력방법으로는 3가지 방법이 있다. 그 첫 번째 방법은 GenBank내의 Accession number나 Genbank Identifier (GI) number를 알고 있다면, 이를 사용하여 BLAST 검색을 할 수도 있다. 두 번째 방법은 직접적으로 그 공란에 query 서열을 입력 혹은 copy/past로 입력하는 방법이다. 이때 그 query 서열의 형식은 FASTA 포맷 혹은 서열만으로 입력할 수 있다. 마지막 세 번째로는 따로 query서열을 메모장 혹은 워드패드를 이용하여 text 파일로 작성해야 하며, 그 파일을 찾아보기 버튼을 눌러 찾아 선택하는 방법이다.

B. Choose Search Set: 검색하고자 하는 database를 선택할 수 있다. 만약 human genomic과 transcript와 관련되는 DNA 서열과 검색하고자 한다면 “Human genomic + transcript”를 선택하고, 그것이 만약 mouse와 관련된 database와 검색하고자 한다면 “Mouse genomic + transcript”를 선택한다. 마지막으로 human과 mouse를 제외한 다른 database와 비교하고자 한다면 “Others (nr etc.)"를 선택하면 된다.

C. Program Selection:

- Highly similar sequences (megablast)는 뉴클레오타이드 서열을 동정하기 위한 선택 툴이다. 알려지지 않은 서열을 동정하기 위한 가장 좋은 방법은 그 서열이 이미 공개 데이터베이스에 존재하는지를 알아보는 것이다. 만일 데이터베이스 서열이 특색이 잘 규정 지어진 서열이라면, 많은 생물학적 정보를 얻을 수 있을 것이다. MEGABLAST, discontiguous-MegaBLAST와 blastn 모두는 이러한 목표를 이루기 위해 사용되어 질 수 있다. 그러나 MEGABLAST는 매우 유사한 서열들 사이에 긴 정렬을 효과적으로 찾기 위해 특별히 고안되었으므로, 질의 서열과 동일한 매치를 찾기 위한 방법으로 가장 좋은 툴이 될 것이다. expect value significance cut-off에 덧붙여, MEGABLAST는 정렬에 대한 adjustable percent identify cut-off도 제공하고 있다.

- More dissimilar sequences (discontiguous megablast)는 질의서열과 유사하지만 동일하지는 않은 뉴클레오타이드를 찾을 경우에는 더 나은 방법이다. BLAST 뉴클레오타이드 알고리즘은 질의 서열을 words라 불리는 작은 서열로 나눔으로써 유사 서열을 찾는다. 프로그램은 첫번째 질의 words에 대한 정확한 매치를 찾는다(words hits). 그리고 나서 BLAST 프로그램은 이 word hits를 다중 step으로 확장 시켜 마지막 gapped 정렬을 생성한다. BLAST 검색의 민감도를 지배하는 중요한 파라미터 중 하나는 바로 최초 words 길이 혹은 크기이다(흔히 word size). blastn이 MEGABLAST보다 훨씬 민감한 가장 중요한 이유가 바로 blastn이 더 짧은 기본 word size(11)를 사용하기 때문이다. 따라서 blastn은 다른 종으로부터 관계 있는 뉴클레오타이드 서열에 대한 정렬을 찾는데 있어서 MEGABLAST보다 더 좋다. blastn의 word size는 적당하며, 검색민감도를 높이기 위해 기본값에서 최소 7 까지 낮출 수 있다. 검색 민감도는 새로이 도입된 discontiguous MegaBLAST 페이지를 사용함으로써 더 향상될 수 있다. 이 페이지는 같은 이름의 알고리즘을 사용하며, Ma et.al.에 의해 보고된 것과 유사하다. 정렬 확장의 seeds로써 정확한 word 매치를 요구하기 보다, discontiguous MegaBLAST는 주형(template)의 더 긴 window내 non-contiguous

word를 사용한다. 코딩모드에서 세 번째 염기 흔들림은 첫번째와 두 번째 코돈 위치에서 매치를 찾는데 초점을 맞춤으로써 고려된 반면, 세 번째 위치에서의 미스매치는 무시되었다. 같은 word size를 사용하여 discontiguous MEGABLAST에서 검색하는 것은 같은 word size를 사용한 표준 blastn보다 훨씬 더 민감하고 효율적이다. 이러한 이유에서 이런 형식의 검색에서는 discontiguous MegaBLAST를 추천한다. 필요하다면 alternative non-coding 패턴도 지정할 수 있다. 그 자세한 방법은 아래의 site를 참고로 하기를 바란다.

http://www.ncbi.nlm.nih.gov/blast/discontiguous.html

http://www.ncbi.nlm.nih.gov/Web/Newsltr/FallWinter02/blastlab.html

discontiguous MegaBLAST를 위한 고유 파라미터는 다음과 같다.

word size : 11 이나 12 두 가지 옵션으로 제한된다.

template :16, 18, 21 세가지 옵션만이 있다.

template type : coding(0), non-coding(1) 또는 양쪽모두(2)

뉴클레오타이드-뉴클레오타이드 검색이 다른 종내 homologous 단백질코딩영역을 찾는데 있어 가장 좋은 방법이 아니라는 점이 중요하다. 이런 작업은 직접 단백질-단백질 BLAST 검색이나 translated BLAST 검색으로 단백질 레벨에서 수행하는 것이 더 좋다. 이는 코돈 degeneracy가 있고, 아미노산 서열에 더 많은 정보가 있으며, 단백질-단백질 BLAST에 사용되는 더 정교한 알고리즘과 스코어링 매트릭스가 있기 때문이다.

- Somewhat similar sequences (blastn)는 짧은 뉴클레오타이드 motif 검색에 유용하다. 20 base 미만의 짧은 서열은 표준은 뉴클레오타이드-뉴클레오타이드 BLAST 세팅 하에서 데이터베이스 엔트리에 대한 어떤 중요 매치로 찾지 못할 때가 종종 있다. 대개 이 경우는 expect value에 의해 지배되는 significance threshold가 너무 엄격하거나, 기본 word size가 너무 높기 때문이다. 표준 BLAST 페이지상의 두 expect value와 word size를 적당히 맞추어서 짧은 서열에 대해 적용할 수 있다. 그러나 이미 짧은 서열에 대해 최적화된 결과를 주기 위해 미리 정해진 이런 값을 가진 페이지를 제공하고 있다. 이 페이지의 일반적인 사용은 PCR이나 hybridization의 특이성(specificity)을 체크하기 위해서 이다. PCR primer 쌍을 체크하기 위한 유용한 방법은 맨 처음 20 혹은 그 이상의 N을 두 primer 사이에 넣어서 두개를 잇고, 이어진 쌍을 하나의 서열로 검색하는 것이다. BLAST는 local 정렬만을 찾고, 자동적으로 양쪽 strands를 찾기 때문에, 잇거나 검색을 하기 전에 primers의 다른 한쪽을 뒤집을 필요는 없다. 질의서열은 애매모호한 base를 가지고 있으면 안 된다. AACNNNNNNRTAYG (StySQI recognition site)나 TGGNNNNNNGCCAA (NF-1 binding site)처럼 degenerate bases를 가진 일치하는(consensus) motifs는 이런 타입의 검색에 적용될 수 없을 것이다.

D. Algorithm parameters: 기타 검색을 위한 첨가적인 option parameter를 결정할 수 있다. 특별한 경우가 아닐 때는 이 parameter들을 default의 경우 그대로 놓아 주는 것이 좋다. 그리고 아래에 각 parameter들에 대한 간단한 설명이다.

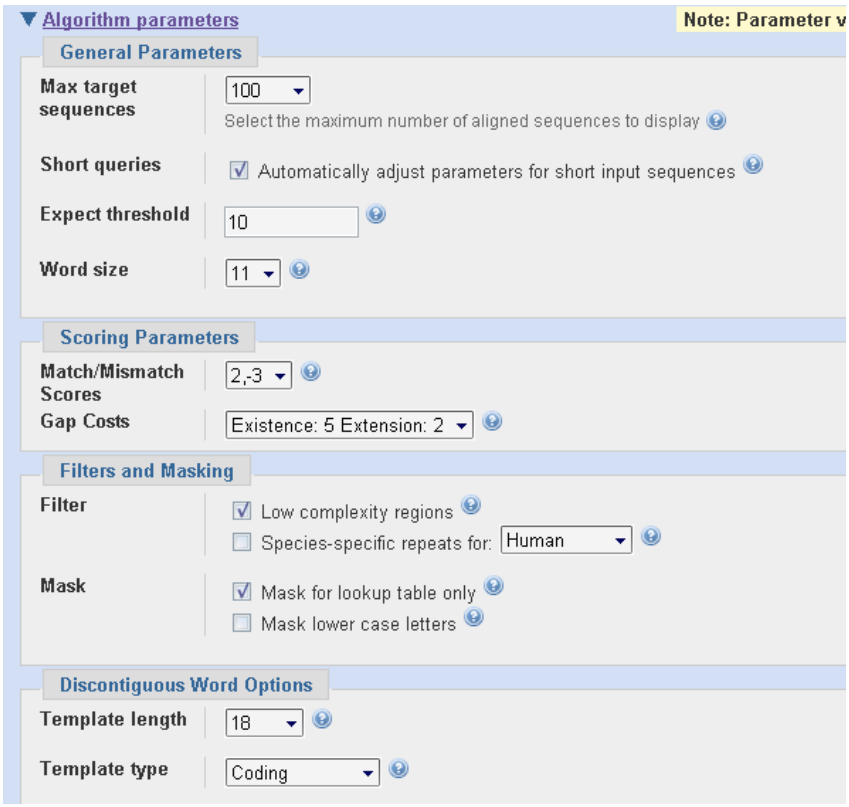

그림 5. Algorithm parameter를 지정하기 위한 창.

Max target sequences

보고될 matching sequence의 갯수를 지정된 숫자로 제한하여 보여준다. 별도의 설정을 하지 않는 경우에는 기본적으로 100을 사용한다.

EXPECT Threshold

database sequences에 대하여 보고할 matches를 한정하는 statistical significance threshold값. 별도의 지정이 없는 경우, 기본값은 10이다. 10의 의미는 Karlin and Altschul (1990)의 stochastic model에 따라, 순전히 우연에 의해 10개의 matches가 발견될 수 있다는 것이다. match의 statistical significance가 EXPECT threshold보다 큰 경우, 해당 matches는 보고되지 않는다. EXPECT threshold 값이 작을수록 좀더 엄격한 검색을 하는 것으로 검색 결과로 보고될 matches가 줄어든다. 분수값도 받아들인다.

FILTER (Low-complexity)

Query sequence에서 low compositional complexity를 가진 segments를 mask off한다. Wootton &Fedrhen(Computers and Chemistry, 1993)의 SEG 프로그램에 의해 필터링이 결정된다. BLASTN의 경우에는 Tatusov and Lipman 의 DUST에 의해 필터링이 결정된다. 필터링을 통해 database sequences에 대해 특정 matching용으로 이용가능한 query sequence의 biologically interesting regions는 그대로 두면서,
statistically significant하지만 biologically uninteresting reports(예를 들어, 공통적인 acidic-, basic- 또는 proline-rich regions에 대한 hits)를 BLAST 출력에서 제거할 수 있다.
필터링은 query sequence또는 query sequence의 translation에만 적용되고 database sequence에는 필터링을 하지 않는다.
별도의 필터링 지정이 없는 경우, 기본적으로 BLASTN에는 DUST필터링을, 그외의 BLAST 프로그램들에는 SEG 필터링이 사용된다.
SWISS-PROT내에 있는 sequence에 SEG 필터링을 적용하는 경우에는 필터링을 통해 아무것도 마스크되지 않는 것은 이상한 일이 아니다. 따라서 필터링이 항상 효과가 있을 것으로 기대할 수는 없다. 또한 어떤 경우에는 sequence전체가 마스크될 수도 있는데 이 때는 필터링하지 않은 query sequence를 사용한 matches의 statistical significance를 의심해볼 필요가 있다.

Low Complexity Filtering

필터링에 관련된 별도의 지정을 하지 않아도 BLAST 2.0 서비스에서는 사용자가 입력한 query sequence에서 low compositional complexity regions(LCR)을 필터링한다. Low complexity regions는 일반적으로 compositional bias하기 때문에 significant position-by-position alignment보다 높게 scoring되기 쉬우므로 BLAST 검색 결과 보고에서 잠재적으로 high score를 받을 수 있는 matches(예를 들어 proline-rich region이나 poly-A tails에 대한 hits로 표시될 수 있음)들을 필터링을 통해 제거하고 BLAST 통계자료에서 pairwise alignment의 specificity를 반영할 수 있는 영역들은 보존한다. blastn 프로그램을 사용한 query에서는 필터링 프로그램으로 DUST를 사용하고, 다른 BLAST 프로그램에서는 SEG를 사용한다. 필터링 프로그램에서 탐지된 low complexity sequence는 nucleotide sequence에서는 'N'으로 대치하고(예를 들어 "NNNNNNNNNNN"), protein sequence인 경우에는 'X'로 대치한다(예를 들어 "XXXXXXXXX"). 필터링을 사용하지 않고 싶다면 BLAST 서비스의 "Advanced options"에 있는 "Filter"옵션을 사용하여 필터링 기능을 해제하면 된다.

FILTER (Human repeats)

이 옵션을 사용하면 Human repeates(LINE's 와 SINE's)가 마스크되므로 이들 repeats를 가지고 있을 수도 있는 human sequence에 특히 유용한 옵션이다.

FILTER (Mask for lookup table only)

이 옵션은 BLAST가 사용하는 참조 테이블(lookup table)을 만들 목적으로만 마스크한다. BLAST extension은 마스킹없이 수행된다.

E. BLAST 버튼: 이상과 같이 blast 검색을 위한 query 서열과 모든 parameter들의 입력이 완료 되었을때 이 버튼을 클릭하면 그 검색의 결과를 볼 수 있다.

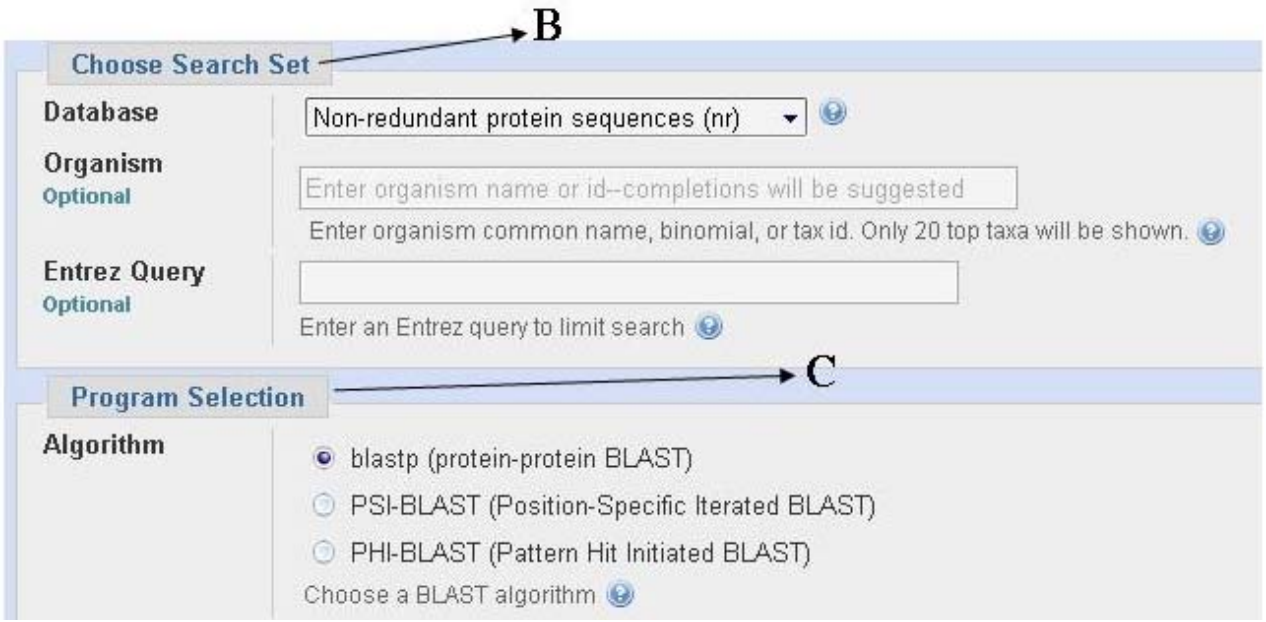

그림 6. BLAST 첫페이지에서 protein blast를 선택한 후에 나타난 창. protein에 관련하는 검색을 할 수 있다.

2.2 pBLAST를 선택한 후

그림 3에서 Basic BLAST란의 "Program" 메뉴들 중에서 protein blast 프로그램을 선택하여 클릭하면 다른 전반적인 것은 nucleotide blast를 선택한 후에 열리는 창과 거의 동일 한데, 단지 위의 그림 6에서처럼 차이가 나는 부분이 있는데 그것만을 설명하겠다.

B. Choose Search Set은 검색하고자 하는 database를 선택하는 것이며, 그 database에 대한 내용은 아래 표에 나타내었다.

표 2.1 단백질 서열 데이터베이스 내용

데이터베이스	내용설명
nr	env_nr에 있는 것을 제외한 non-redundant GENBANK translations + PDB + SwissProt + PIR + PRF
refseq	NCBI reference sequence 프로젝트에서 나온 단백질 서열
swissprot	SWISS-PROT 단백질 서열 데이터베이스의 최근 주요 릴리즈 (no incremental updates)
pat	GENBANK의 특허부(patent division)에서 온 단백질 서열들.
pdb	Protein Data Bank의 3차원 구조 레코드로부터 유도된 서열들
env_nr	토양이나 해양 샘플로부터 나온 배양되지 않은 bacterial 샘플과 같은 환경적 샘플로부터 나온 서열로 부터의 CDS translation된 서열

C. Program selection

Protein blast에서의 program은 blastp (protein-protein BLAST), PSI-BLAST (Position-Specific Iterated BLAST) 그리고 PHI-BLAST (Pattern Hit Initiated BLAST)가 있으며, 검색 목적에 따라 적당한 것을 선택하면 된다.

blastp는 일반적으로 단백질 검색을 위하여 고안되었다. blastp는 질의 아미노산 서열을 동정하고, 단백질 데이터베이스 내 유사한 서열을 동정하는 두 가지 목적으로 사용된다. 다른 BLAST 프로그램들처럼, blastp는 유사한 local 영역을 찾기 위해 고안되었다. 서열 유사성이 전체 서열로 확장될 때, blastp는 global 정렬도 기록할 것이며, 이것이 단백질 동정 목적에 오히려 더 좋은 결과이다. 동정 검색의 명확한 결과를 위하여 "low-complexity filter" 모두를 꺼 본다. 뉴클레오타이드 BLAST와는 달리 단백질 검색을 위한 MEGABLAST만한 것이 없으므로, 웹을 통한 배치검색이 불가능하다. 배치 단백질 BLAST를 하기 위해서는 netblast(blastcl3)를 살펴보도록 한다. 이 툴을 설명하는 문서는 netblast.txt 이다.

PSI-BLAST는 좀더 민감한 단백질-단백질 유사성 검색을 위해 고안되었다. 두 개의 아미노산이 바뀔 확률이 특정 단백질 서열에서의 아미노산 위치에 의존할 (position specific) 가능성을 고려해야 한다. Position-Specific Iterated(PSI)-BLAST는 이와 같이 특정 위치에 있는 아미노산 서열의 바뀔 가능성에 대해서 다른 점수로 분리하여 주어진다면 보다 감도가 높게 단백질 서열간의 정렬이 실행될 것이며, 이는 매우 멀리 떨어진 관계된 단백질을 찾아내는데 있어 유용하게 할 것이다. 표준 단백질-단백질 BLAST 검색에서 중요한 hit을 찾는데 실패하거나, "hypothetical protein" 이나 "similar to .."라는 메시지와 함께 hit이 출력될 때 이용할 수 있는 방법이 PSI-BLAST이다. PSI-BLAST는 우선 1차적으로 표준 단백질-단백질 BLAST 검색프로그램이다. 이 프로그램은 포함된 threshold (기본값=0.005) 보다 낮은 Expect value를 가지고 나온 서열의 정렬로부터 position-specific scoring matrix (PSSM)를 만든다. PSSM은 검색의 다음 반복에서 정렬을 평가하는데 사용될 것이다. 내부 threshold보다 낮은 새로운 데이터베이스의 어떤 hit이라도 새로운 PSSM을 만드는데 포함된다. PSI-BLAST 검색은 부차적 반복 (subsequent iteration)에서 더 이상 새로운 데이터베이스 서열을 추가하지 않을 때 수렴한다고 할 수 있다. hit 옆의 체크박스를 체킹하여 내부 threshold 내에 속하지 않는 데이터베이스 hit도 다음번의 PSSM에 추가할 수 있다. 이미 선택된 hit 역시 체크박스의 체크를 없애서 선택에서 제외시킬 수 있다. PSSM은 질의서열 특이적(query specific) 이다. 한 데이터베이스에 대한 PSI-BLAST 검색동안 만들어진 PSSM을 저장하여 같은 질의에 대한 다른 데이터베이스의 검색에서 이를 사용할 수도 있다. 이렇게 하기 위해서는 "Formatting BLAST" 페이지(첫번째 반복 이후의 과정에서)의 "Format" 섹션의 풀다운 메뉴에서 "Alignment"를 "PSSM"으로 바꾸어야 한다. 그리고 나서 검색을 포맷하고, 결과로 나오는 ASCII로 인코딩된 PSSM을 복사하여 새 PSI-BLAST 페이지의 PSSM 창에 이를 붙이면 된다.

PHI-BLAST는 제한된 단백질 패턴 검색을 할 수 있다. Pattern-Hit Initiated (PHI)-BLAST는 사용자에 의해 명시된 패턴을 포함하고 있고, 패턴 부근의 질의서열과 유사한 단백질을 검색하기 위하여 고안되었다. 이 이중요건은 패턴을 포함하고 있지만 질의서열에 대한 true homology를 가지고 있지 않은 많은 데이터베이스의 hits를 줄이기 위한 의도이다. PHI-BLAST를 실행하기 위해서는 하나 이상의 패턴형식을 가진 질의서열을 "Search" 박스에 넣고, 페이지의 "Option" 섹션에 있는 "PHI pattern"박스에 패턴을 넣는다. 패턴은 반드시 PROSITE의 문법협의(syntax convention)에 따라야 한다. 주어진 검색에서 오직 하나의 패턴만이 사용될 수 있다. PROSITE 형식의 샘플질의 서열과 샘플 패턴이 PHI-BLAST의 테스트 수행을 위해 그 예를 아래에 나타내었다.

* PROSITE 형식의 샘플질의 서열 *

>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase
MSHIQIPPGLTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAATPRQSLGHPPPEPGPDR
VADAKGDSESEEDEDLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQEACKDILLF
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGS
FGELALMYNTPRAATIVATSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSERMKIVDVIGEK
IYKDGERIITQGEKADSFYIIESGEVSILIRSRTKSNKDGGNQEVEIARCHKGQYFGELALVTNKPRAAS
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNISHYEEQLVKMFGSSVDLGNLGQ

* 샘플 패턴 *

[LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV]

3. BLAST의 결과 해석

여기부터는 위의 설명대로 어떤 분석하고자 하는 query 서열과 여러 가지 parameter들의 입력이 끝나고 나서 검색을 한 결과물에 대한 분석을 이장부터 하고자 한다. 우선, 설명의 용이성을 위하여 Methanococcus jannaschii에서 나온 미지의 archaebacterial protein MJ0577을 sample query sequence로 하여 non-redudant database를 대상으로 BLASTP로 검색할 때의 BLAST 출력을 예로 사용할 것이며, 이후 설명은 BLAST 출력에서 디스플레이되는 순서를 따라 각각의 내용을 기술할 것이다.

BLAST 검색에 대한 결과의 창을 보면, 크게 5부분으로 나누어진다. 그 첫 번째 부분은 대괄적인 query와 Database에 대한 내용이 나오는데, 즉, 수행한 분석의 Job title, BLAST의 참고문헌, 그 검색에 대한 고유한 분석번호, 검색에 사용된 database에 대한 설명 그리고 어떤 질의한 Query에 대한 설명이 나타난다 (그림 7). 여기서 아래의 결과의 그림 일부를 보면, 검색에 사용된 Database는 "All non-redundant GenBank CDS"가 사용되었고 그 Database에는 6,340,496개의 서열이 보관되어 있다는 것을 알 수 있다. 그리고 질의 서열에 대한 정보도 간단하게 나오는 데, 즉, 그 질의 서열의 길이는 162개의 아미노산으로 이루어 져있다는 간단한 정보가 나타난다.

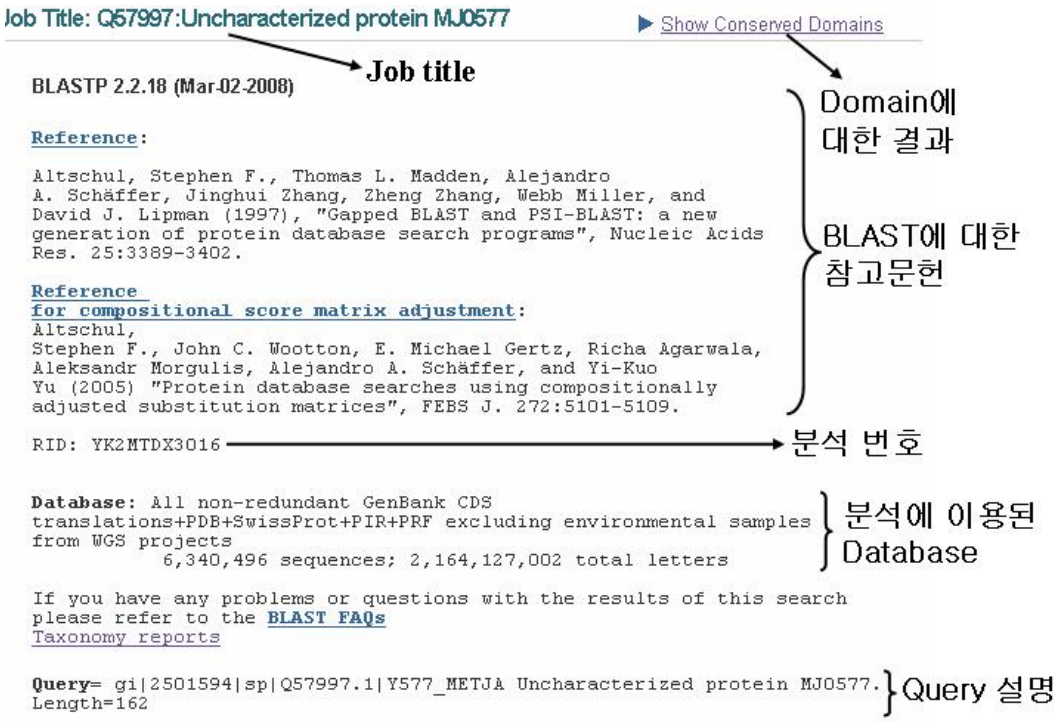

그림 7. 검색 결과창의 도입부.

여기서 입력한 질의 서열인 query 단백질 서열의 Domain에 대한 결과를 보기를 원한다면 "Show Conserved Domains"을 클릭하면 그 결과를 새로운 창으로 나타난 그림과 함께 볼 수 있으며, 그 새로운 창의 그림을 클릭하면 그 domain에 대하여 보다 자세하게 확인할 수 있다. 그리고 일반적으로 BLAST에 분석하고자 하는 질의 서열 (query)를 입력한 후 실행을 하면 대부분 빠른 시간에 그 결과를 볼 수 있다. 하지만, 분석하고자 하는 이의 사정상 그 결과물을 바로 볼 수 없을 때, 혹은 후에 이 분석 결과를 다시 보고자 할 때, 실행 버튼을 클릭한 후 창이 바뀌면서 완전한 결과가 나타나기 전에 중간 과정으로 분석번호 (RID no)를 표시한 창이 나타나는데 이 분석 번호를 잘 메모하여 추후에 각 검색 창의 위에 "Recent Results" (그림 8의 ①) 버튼을 클릭하면 아래와 같은 최근 결과보기 창에 들어가 그 메모한 "RID no" (그림 8의 ②)를 입력한 후 "Go" (그림 8의 ③)를 클릭하면 똑같은 그 결과물을 얻을 수 있다.

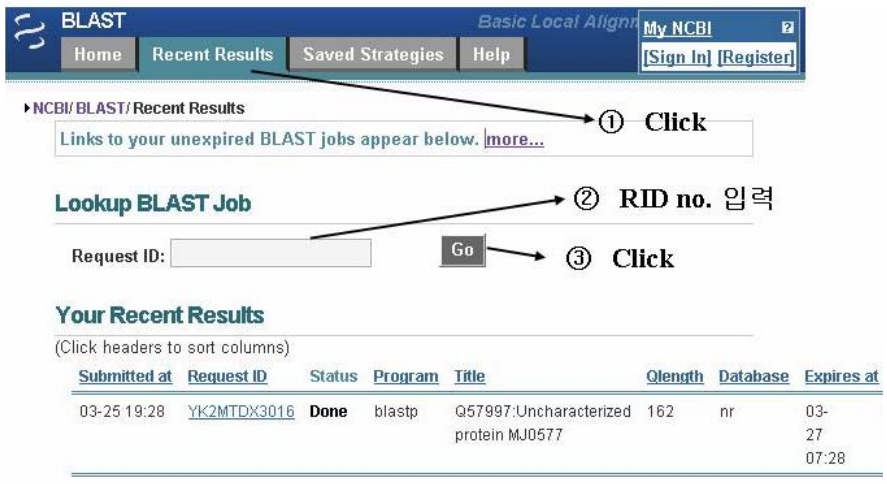

그림 8. BLAST는 검색한 결과를 History하는 기능을 가지고 있으며, 그 기능을 이용하여 이전 검색한 결과들을 다시 볼 수 있다.

이제는 검색 결과물의 5부분 중에서 2번째인 Graphical overview에 대하여 설명하겠다 (그림 9). 이 부분은 질의 서열에 대한 결과를 사용자의 이해를 돕기 위하여 쉽게 그림으로 나타낸 것이다 (그림 9).

① 질의 서열의 검색에 대한 유사서열의 갯수와 함께 graphical overview에 대한 설명의 시작을 알려준다.

② 아래 여러 색깔의 bar들은 질의 서열과 유사도 점수를 색깔로 표시할 뿐만 아니라, 어느 정도의 길이만큼이 질의 서열과 유사한지를 나타내는 것이다.

③ 질의 서열과의 대상 서열간의 유사도에 대한 점수를 색깔화 할 것이다. 예를 들면, 유사도 점수가 200 이상일 때 빨간색으로 나타내며, 그것이 질의 서열과 가장 유사할 가능성이 높다는 것을 말한다.

④ 질의 서열을 도식화한 것이다. 아래 길이에 대한 표시가 나타나며 이번 검색에서 입력된 질의 서열의 길이와 그 아래의 유사서열간의 유사 부위가 어디인지를 대충 알 수 있게 한다.

⑤ 이런 bar의 각각에 마우스 cursor를 올리면 그 bar (그 서열)에 대한 Genbank accession no., 그 서열의 간단한 설명, 유사도 점수 (S) 그리고 E값이 창에 나타난다.

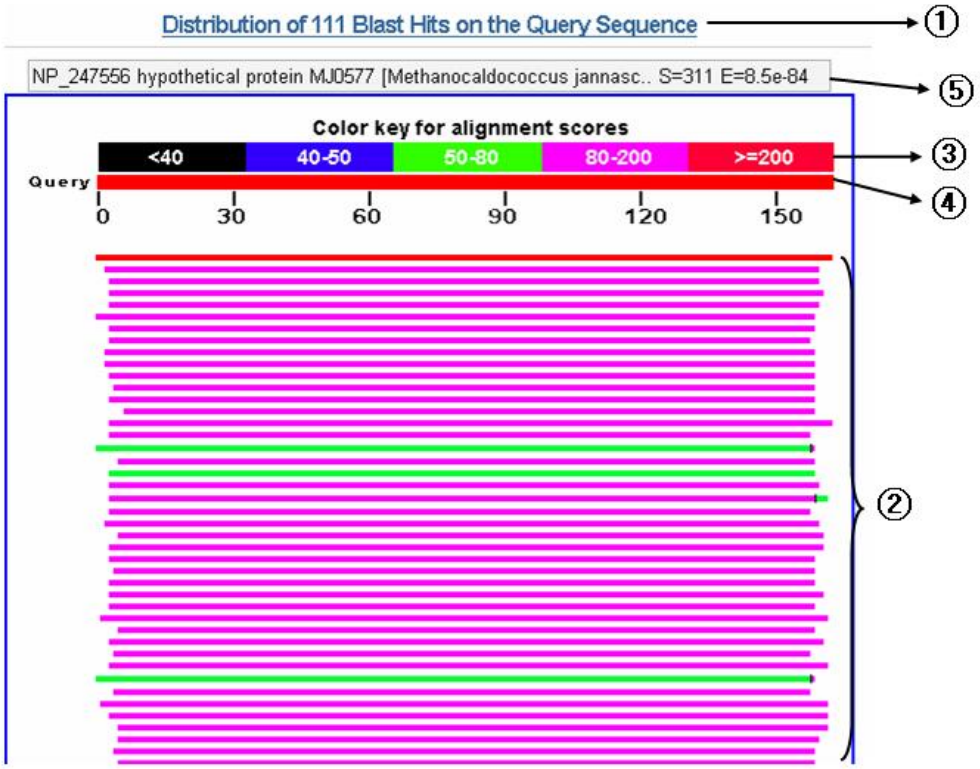

그림 9. 검색한 후 결과창의 graphical overview 부분.

검색 결과물의 5부분 중에서 3번째 부분에는 이번 검색에서 얻은 질의 서열과 유사한 서열에 대한 내용을 간단하게 설명한다 (그림 10). 여기에 나타나는 정보는 위의 ⑤과 유사한 정보들을 볼 수 있는데 한 가지 차이가 있다면 각 서열 정보의 자세한 내용을 보기 위한 link가 연결되어 있다. 이들 Genbank에 보관된 서열의 정보에 대해서는 앞의 NCBI-introduction 장에서 Genbank에 대한 내용을 설명했으니, 여기서는 그 자세한 설명은 생략하겠다. 위의 ⑤와 마찬가지로 질의 서열과 대상 서열간의 유사도 S값과 E값을 보여주며, 그 들중에 유사도 S값이 아래 그림과 같이 link가 연결되어 있음을 알 수 있다. 이 link를 클릭하면 질의 서열과 유사서열간의 정렬에 대한 결과를 볼 수 있다. 또한 위에 표시된 “Distance tree of results”를 클릭하면 질의 서열을 중심으로 다른 종 혹은 유사유전자와 관계를 암시할 수 있는 간단한 phylogenetic tree에 대한 정보를 볼 수 있다.

그림 10. 검색 결과에서 질의 서열과 유사한 서열에 대한 내용을 간단히 설명한 부분.

검색 결과물의 5부분 중에서 4번째 부분에는 질의 서열과 관계하는 유사서열 간의 서열정렬 (sequence alignment)의 결과를 볼 수 있다 (그림 11). 이 정보에서는 각 유사서열에 대한 간단한 설명과 그리고 보다 자세한 유사서열의 정보를 보기 위한 link가 있다. 그 link를 클릭하면 유사서열에 대한 보다 자세한 서열정보를 볼 수 있다. 아래 그림을 보면 질의 서열의 3번째 서열로부터 159번째 서열까지가 유사서열의 6번째 서열부터 150번째 서열까지의 서열정렬을 볼 수 있고, 그 사이에 "-"은 보다 나은 정렬을 위하여 삽입된 gap을 표시한 것이다. 그리고 두 서열간에 일치한 서열은 두 서열 정렬 가운데 일치한 각 아미노산 표기로 표시했으며, 또한 "+"는 그 두 서열간 아미노산의 유사성이 높다는 것을 표시하는 것이다. 그리고 그 정렬위에는 질의서열과 유사서열간의 유사도의 S값과 E값이 나오고, 그리고 유사성 (Positives)와 일치성 (Identities)을 %로 표시되어 있다. 서열 데이타베이스에 대한 BLAST 검색은 수십 혹은 수백 개의 후보 정렬을 내놓는다. 이들 정렬 결과로부터 어떤 것이 정말 중요한 상보성을 가진 것이고 어떤 것이 관련 없는 것인지는 raw score, bit score, e-value로 판단할 수 있다.

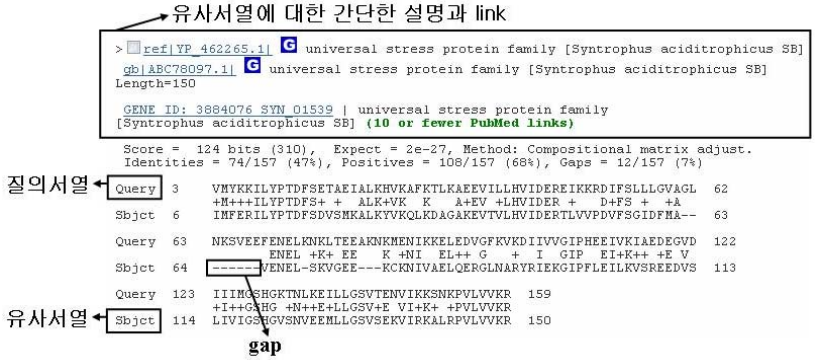

그림 11. 검색 결과에서 질의 서열과 관계하는 유사서열 간의 서열의 정렬의 결과를 나타내는 부분.

마지막으로, 검색 결과물의 5부분 중에서 5번째 부분은 검색에 대한 통계자료를 볼 수 있다. 여기서 나오는 자료는 S값, E값고 같은 통계치를 계산하는데 도움을 주는 수치들이 있다. 이런 수치들이 어떤 공식으로 어떻게 계산하여 S 혹은 E 값과 같은 통계치를 계산하는지는 여기서 생략하도록 하겠다. 예를 들면, 만약 BLAST의 E-value가 1e-12이면 0에 가까운 수이다. 따라서 쿼리서열이 우연히 조합되어 만들졌을 가능성이 매우 작다(검색 데이터베이스를 Random Data라고 가정했으므로). 즉, 랜덤서열중에 비슷한게 거의 없다는 의미이므로, 어떤 정보를 담고 있을 확률이 높다는 뜻이 되며, 그 값이 더욱 작을수록 그 값은 유의미하다고 이해하도록 하자.

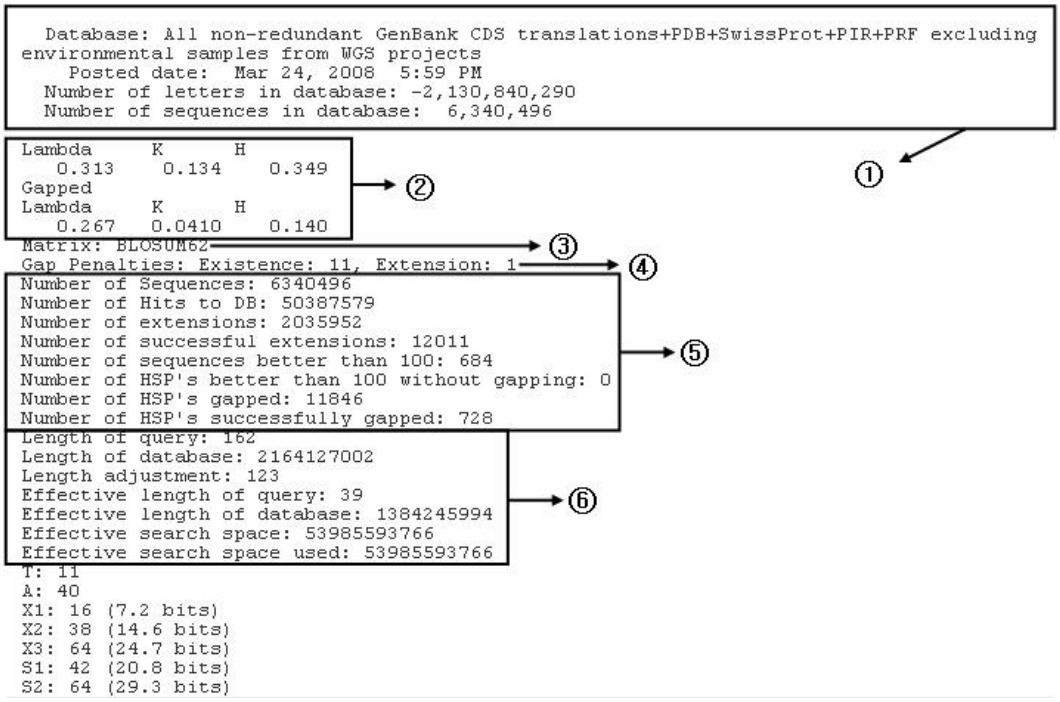

그림 12. 검색 결과에 대한 통계자료.

① 검색에 사용된 database, 그 database의 최신 갱신 날짜와 최신 갱신 시점에서의 database의 크기
② 검색의 결과로부터 계산된 lambda, K 및 H 의 값
③ 사용된 scoring matrix 의 종류
④ Gap creation cost 및 gap extension cost
⑤ HSPs 확인까지 이르는 현재 BLAST search 과정의 여러 단계에서의 값
⑥ 현재 BLAST 검색에서 사용된 database size, query size용으로 계산된 값