FASTA 실습

Bioinformatics Lecture

By Prof. Keun Woo Lee

경상대학교 생화학과 생물정보학연구실
Bioinformatics Lab http://bio.gsnu.ac.kr

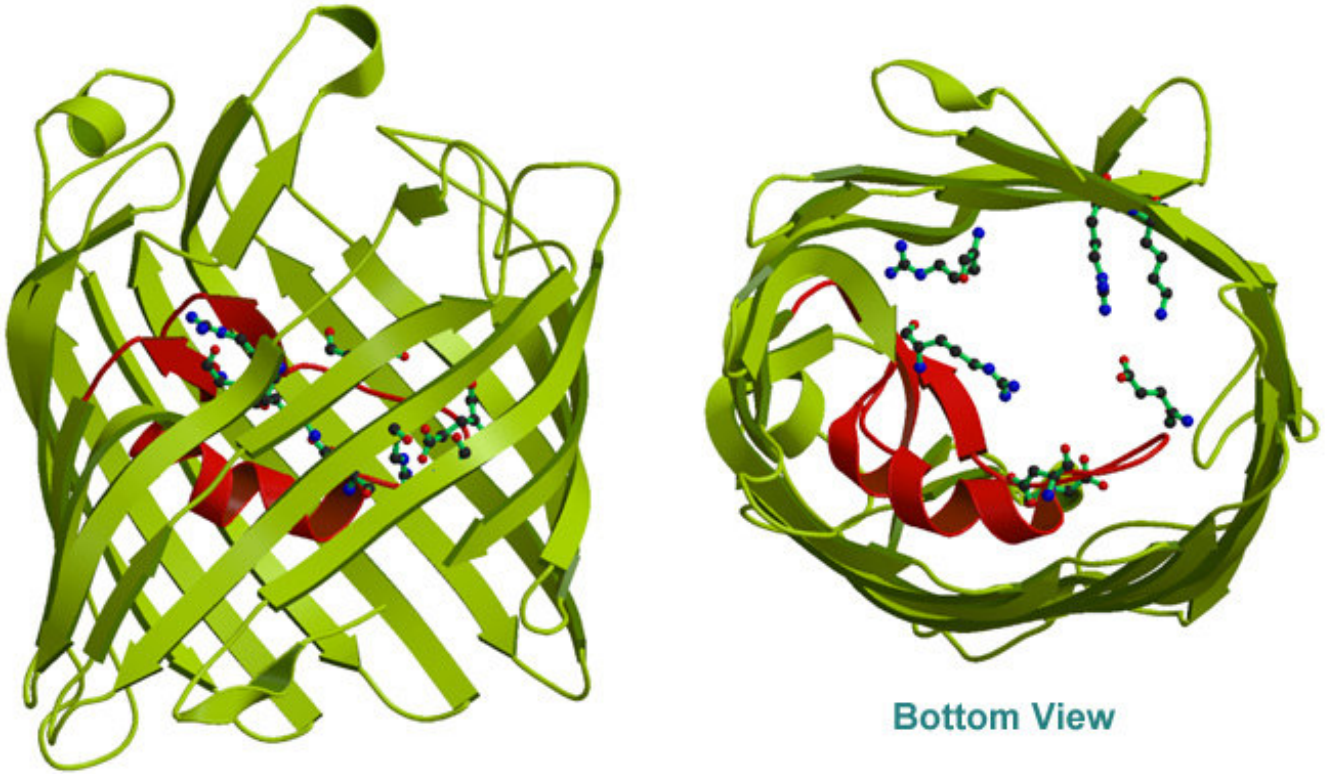
강의 목표

1. 세계적인 서열검색 프로그램인 FASTA를 사용해 본다. 수강생의 직접적인 체험을 통해 이 프로그램의 기능에 대한 이해도를 높인다.

2. E. coli OmpF Porin 단백질의 아미노산 서열을 이용하여 비슷한 서열을 갖는 단백질을 검색 하여 본다.

E. coli OmpF Porin 단백질의 구조

일반적으로 Porin 은 박테리아의 세포벽 바깥쪽 막에 존재하며 영양분 및 이온들의 통로로 알려져 있다. 이번 시간 실습에 사용될 구조는 E. coli 에 존재하는 OmpF 라 불리우는 단백질이다. 총 340개의 아미노산으로 이루어져 있으며 16개의 베파병풍구조들로 배럴을 구성하고 배럴 중간에는 통로의 개폐를 조절하는 L3 루프라고 불리는 알파나선구조가 존재한다.



Porin 단백질의 서열 준비

E. Coli 의 OmpF 단백질의 서열은 340개의 아미노산이다. FASTA를 사용하고 싶으면 아미노산 서열들을 모두 1문자코드로 작성하여야 한다. 즉 알라닌은 A, 글라이신은 G 처럼 340개의 1문자코드로 늘어뜨려 놓아야 FASTA의 입력란에 집어 넣을 수 있다.

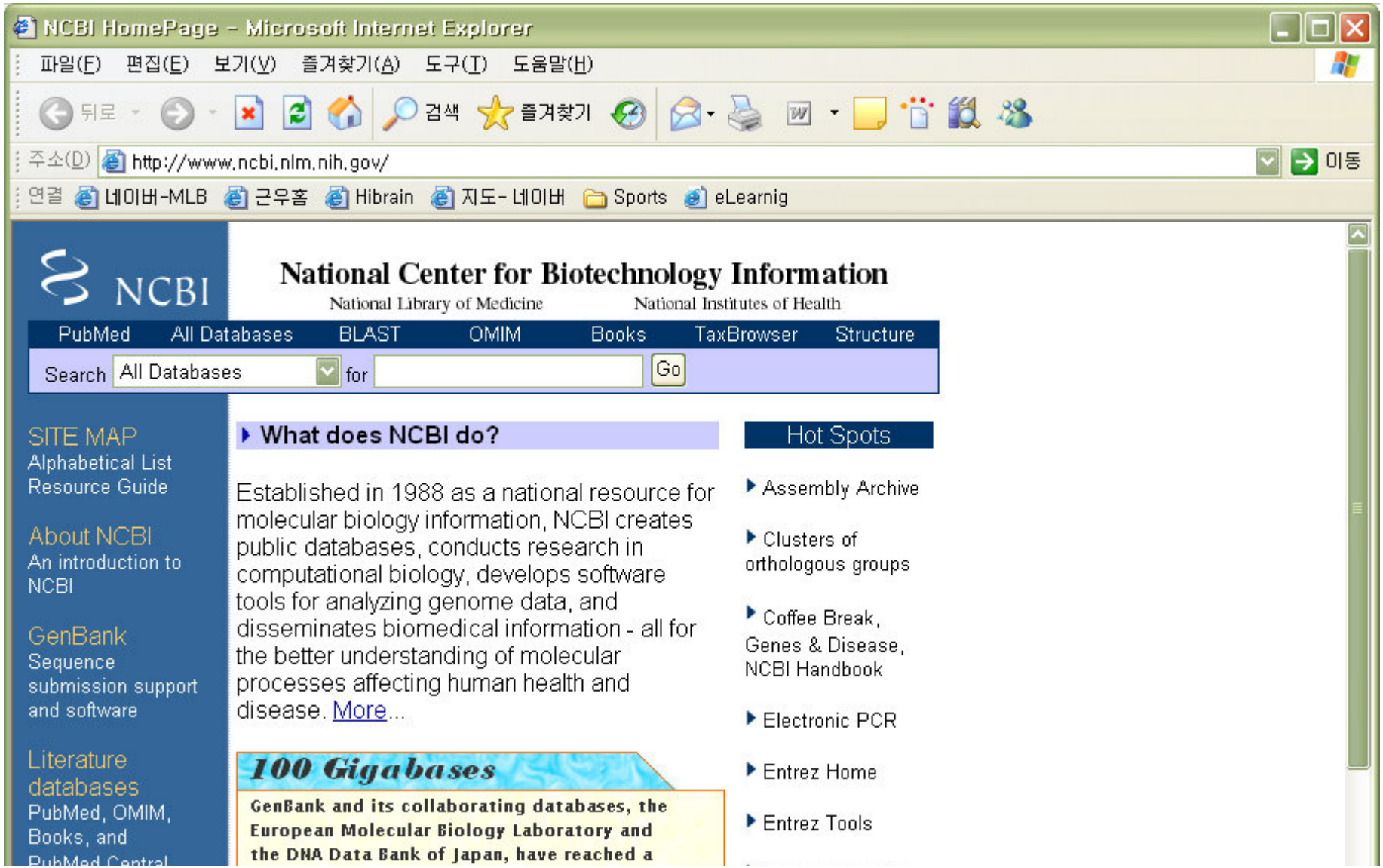
이런 서열들은 유럽의 최대 생물정보서버 홈인 EBI (European Bioinformatics Institute)나 미국 최대의 생물정보서버 홈인 NCBI (National Center for Biotechnology Information) 홈페이지를 방문하여 구할 수 있다.

준비된 340개의 1문자코드로 작성된 아미노산 서열:

aeiynkdgnk vdlygkavgl hyfskgngen syggngdmty arlgfkgetq insdltgygq
weynfqgnns egadaqtgnk trlafaglky advgsfdygr nygvvydalg ytdmlpefgg
dtaysddffv grvggvatyr nsnffglvdg lnfavqylgk nerdtarrsn gdgvggsisy
eyegfgivga ygaadrtnlq eaqplgngkk aeqwatglky danniylaan ygetrnatpi
tnkftntsgf anktqdvllv aqyqfdfglr psiaytkska kdvegigdvd lvnyfevgat
yyfnknmsty vdyiinqids dnklgvgsdd tvavgivyqf

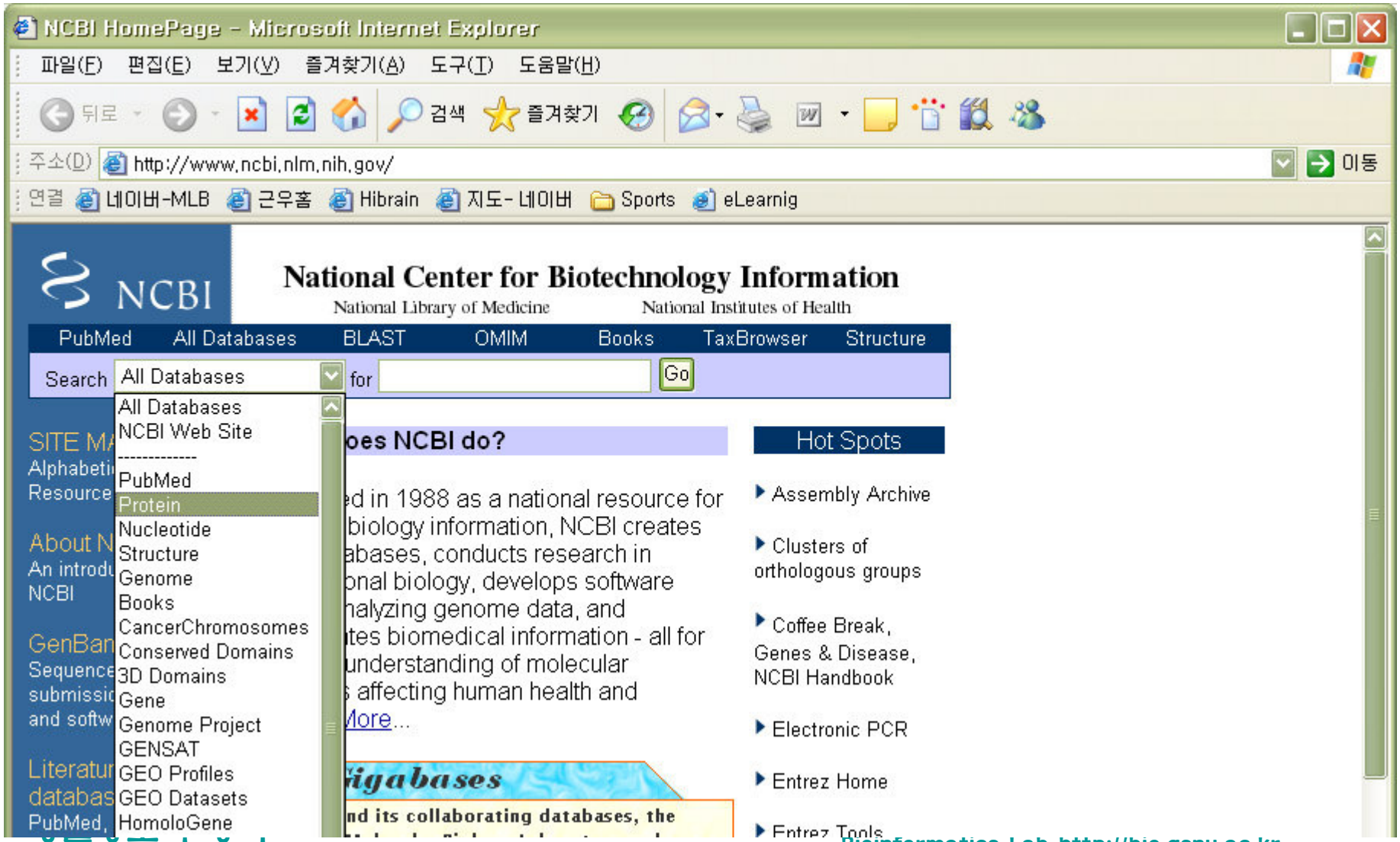
NCBI에서 단백질 서열구하기

웹브라우저를 열고 주소에 http://www.ncbi.nlm.nih.gov/ 라고 친다.



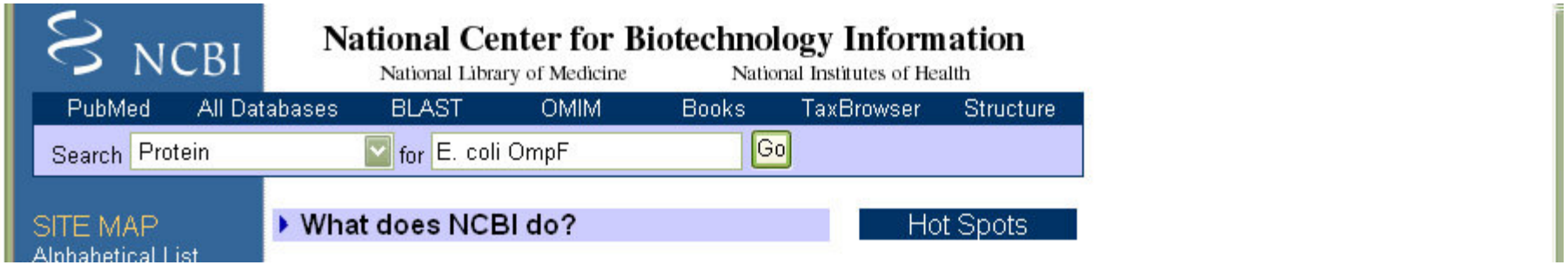
NCBI에서 단백질 서열구하기

왼쪽 조그만 창에서 DB 종류를 단백질(protein)로 선택한다.



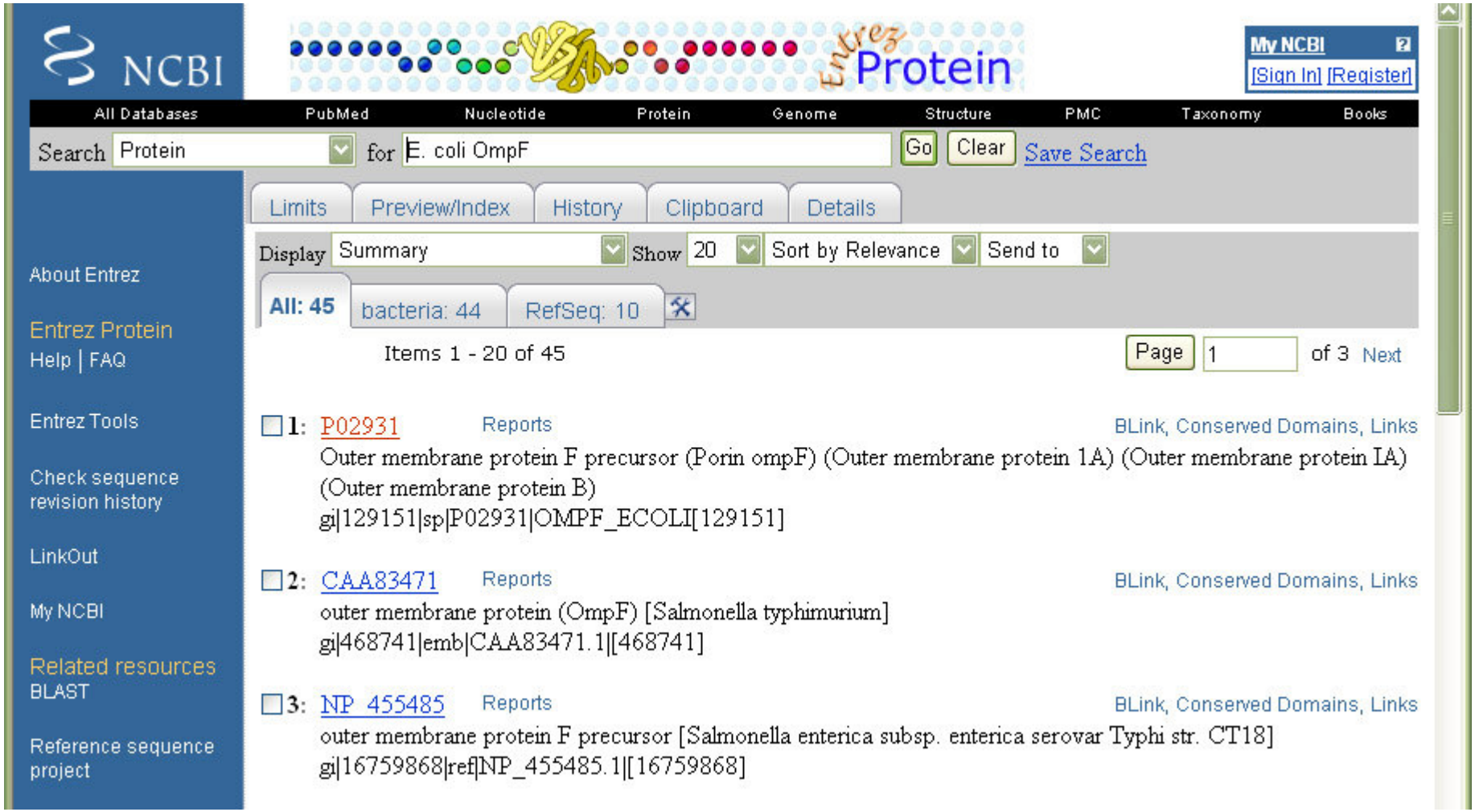
NCBI에서 단백질 서열구하기

그런다음 검색창에 “E. coli OmpF” 란 키워드를 입력하고 Go 버튼을 누른다.



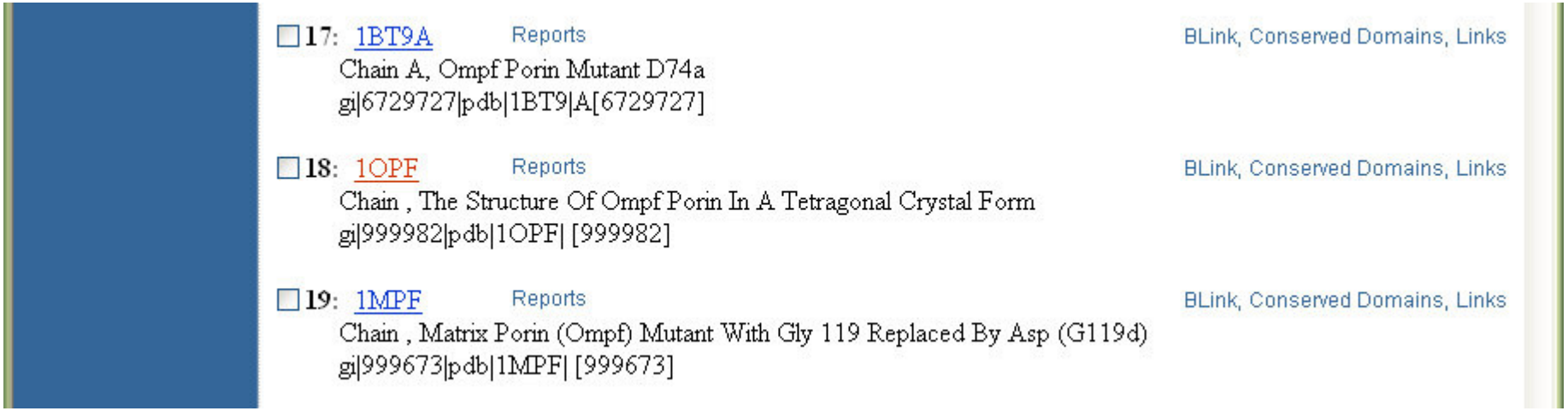
NCBI에서 단백질 서열구하기

그러면 다음과 같이 관련된 45개의 자료가 검색된다.



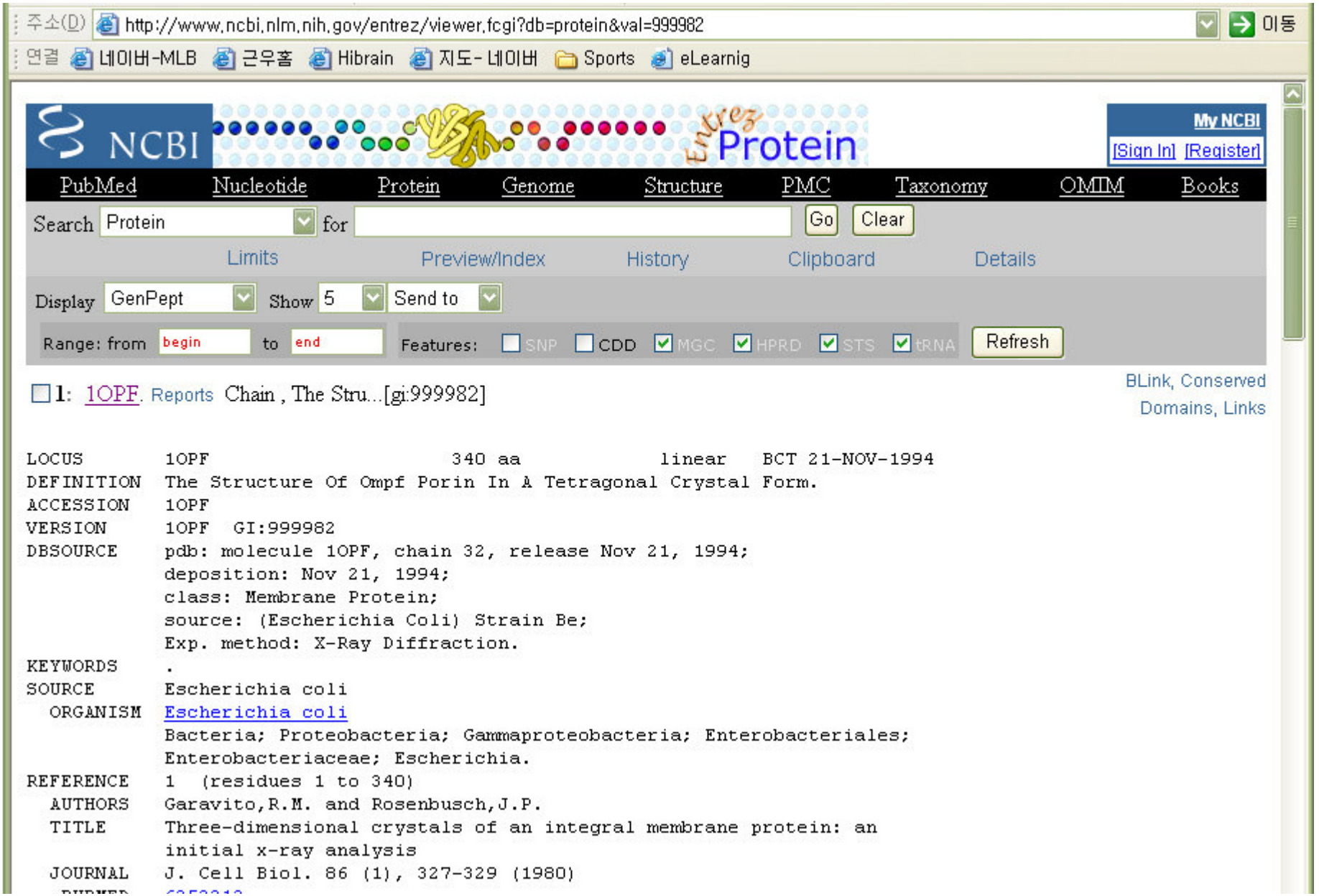
NCBI에서 단백질 서열구하기

조금 아래로 내려가면 18번에 원하는 자료를 만날 수 있다. 앞에 있는 많은 자료는 OmpF 가 아니고 Precursor들에 관한 자료가 많으니 주의하기 바란다. 그리고 이근처 자료들의 아이디이름이 앞의 주어진 자료들과는 달리 4자리 코드명인데, 이는 Protein Data Bank (PDB)에 이미 그 단백질의 3차원(입체)구조가 이미 올라가 있음을 의미이다.



1OPF를 클릭하면 다음처럼, 이 단백질이 속한 세균의 이름, 이 데이터를 저장한 시간, 관련된 논문 소개등, 그 서열에 대한 자세한 설명이 다음처럼 주어진다.

NCBI에서 단백질 서열구하기



NCBI에서 단백질 서열구하기

2차구조 설명등 다양한 자료내용을 지나 그 페이지의 제일 밑까지 내려가면 그 단백질의 서열에 관한 정보가 주어진다. 10개 서열마다 공백이 하나 들어가고 한 행마다 60개 서열이 나열된다.

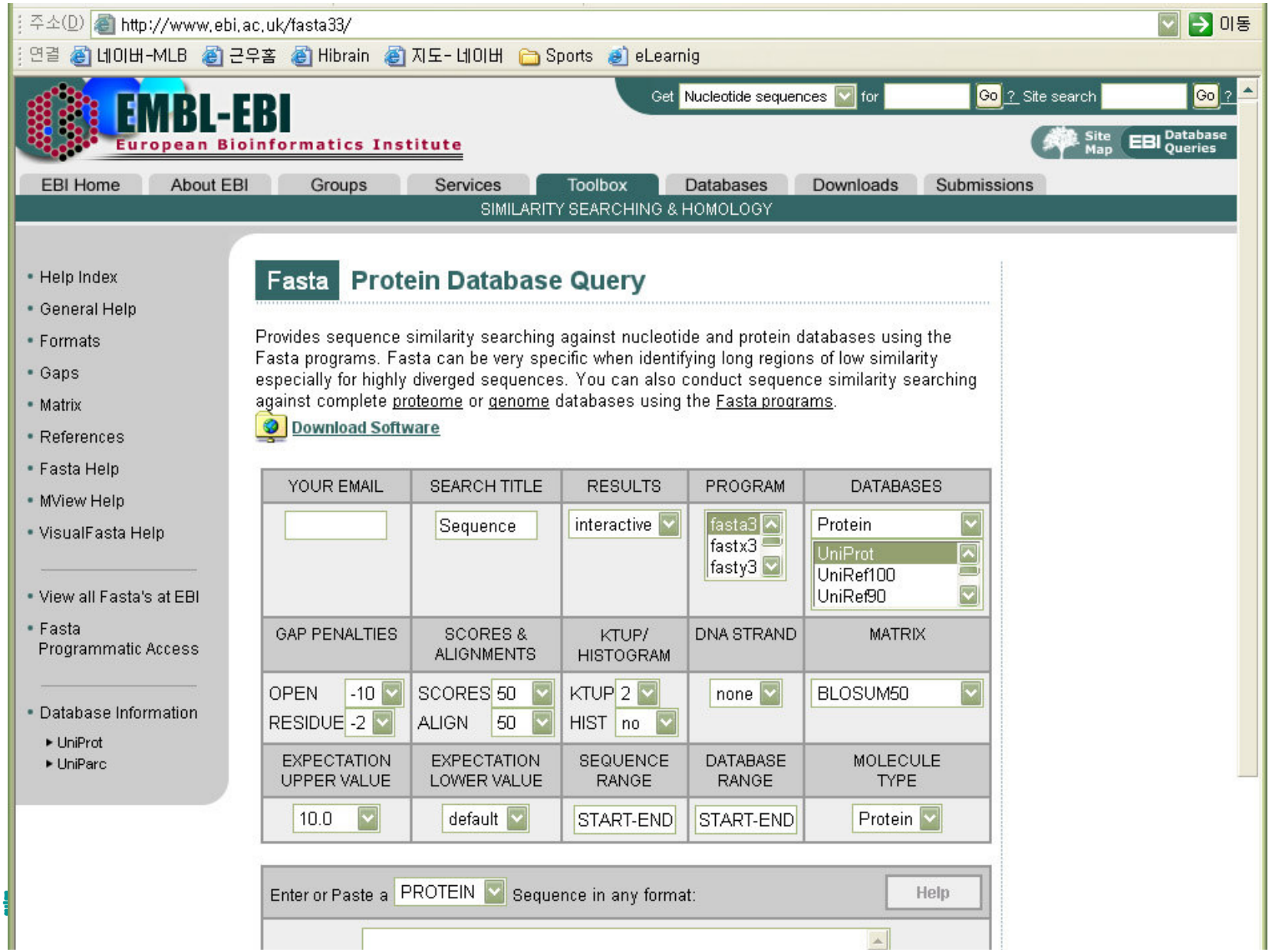
이서열을 노트패드 등에 복사 한 뒤 앞에 있는 숫자를 지워서 저장한뒤, 나중에 FASTA나 BLAST 검색때 사용하여라.

```
     SecStr          328..334
                     /sec_str_type="sheet"
                     /note="strand 30"
     SecStr          335..340
                     /sec_str_type="sheet"
                     /note="strand 31"
ORIGIN
        1 aeiynkdgnk vdlygkavgl hyfskgngen syggngdmty arlgfkgetq insdltgygq
       61 weynfqgnns egadaqtgnk trlafaglky advgsfdygr nygvvydalg ytdmlpefgg
      121 dtaysddffv grvggvatyr nsnffglvdg lnfavqylgk nerdtarrsn gdgvggsisy
      181 eyegfgivga ygaadrtnlq eaqplgngkk aeqwatglky danniylaan ygetrnatpi
      241 tnkftntsgf anktqdvllv aqyqfdfglr psiaytkska kdvegigdvd lvnyfevgat
      301 yyfnknmsty vdyiinqids dnklgvgsdd tvavgivyqf
//
```

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

FASTA 실습: 홈페이지 열기

웹브라우저를 열고 주소 검색창에 http://www.ebi.ac.uk/fasta33/ 라고 친다.



FASTA 실습: 선택사항 소개

앞에서 본 바와 같이 다양한 선택사항이 존재 하고 각각의 설명은 왼쪽 메뉴중 Fasta Help에 자세히 설명 되어져 있다.

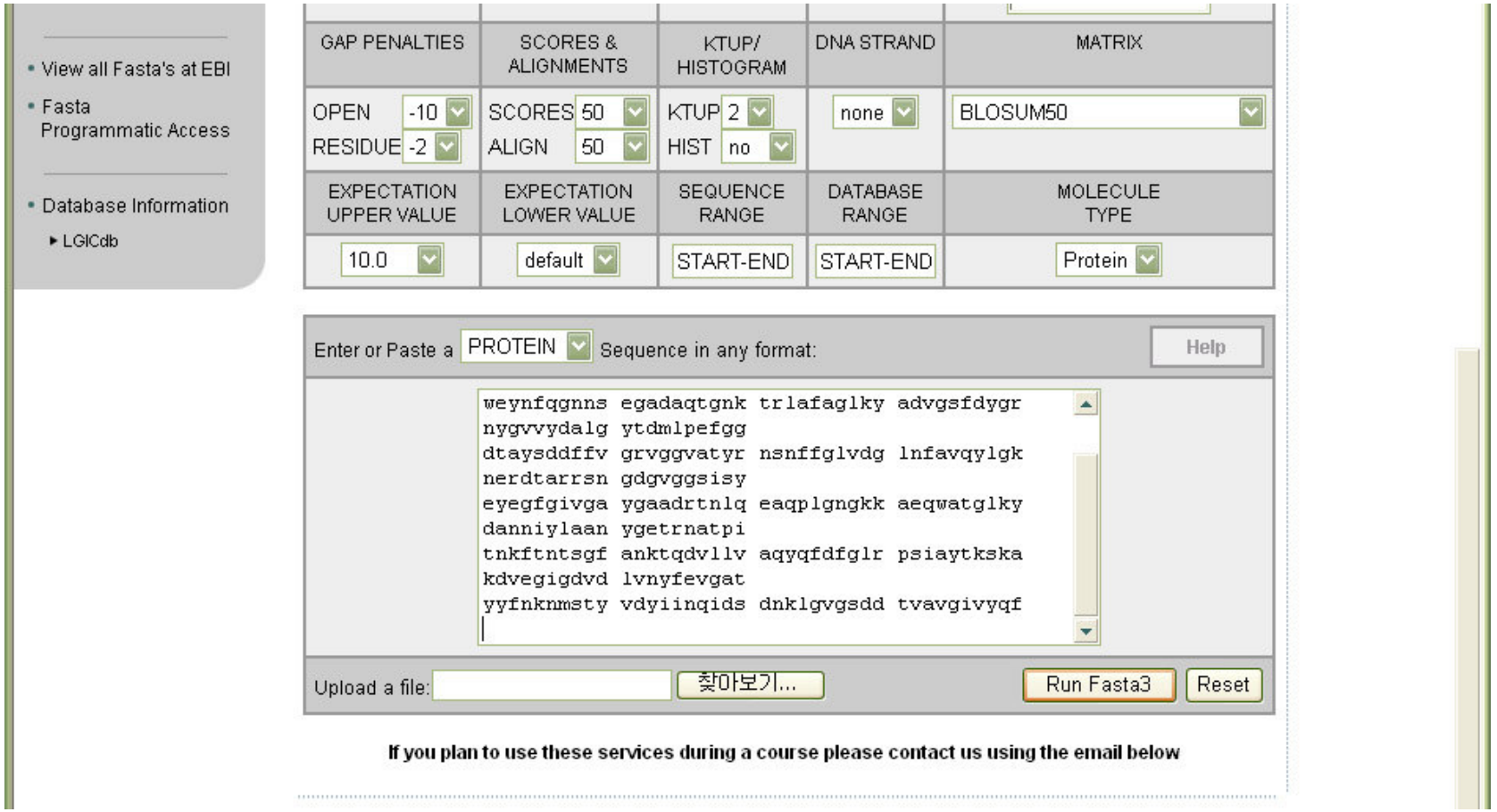
Program, Databases, Matrix등의 선택에 주의 하고 나머지 사항은 기본값을 사용할 것. 아래 프로그램의 종류에 대하여 설명을 실었다.

•PROGRAM 종류

Program	Function (기능)
Fasta3	단백질 혹은 DNA서열을 같은 유형의 DB서열과 비교
fastx/y3	DNA서열을 단백질서열 DB와 비교 (DNA서열을 번역한 후)
tfastx/y3	단백질서열 번역한 DNA서열 DB와 비교
fasts3	연결된 펩타이드서열을 PDB와 비교
fastf3	혼합된 펩타이드서열을 PDB와 비교

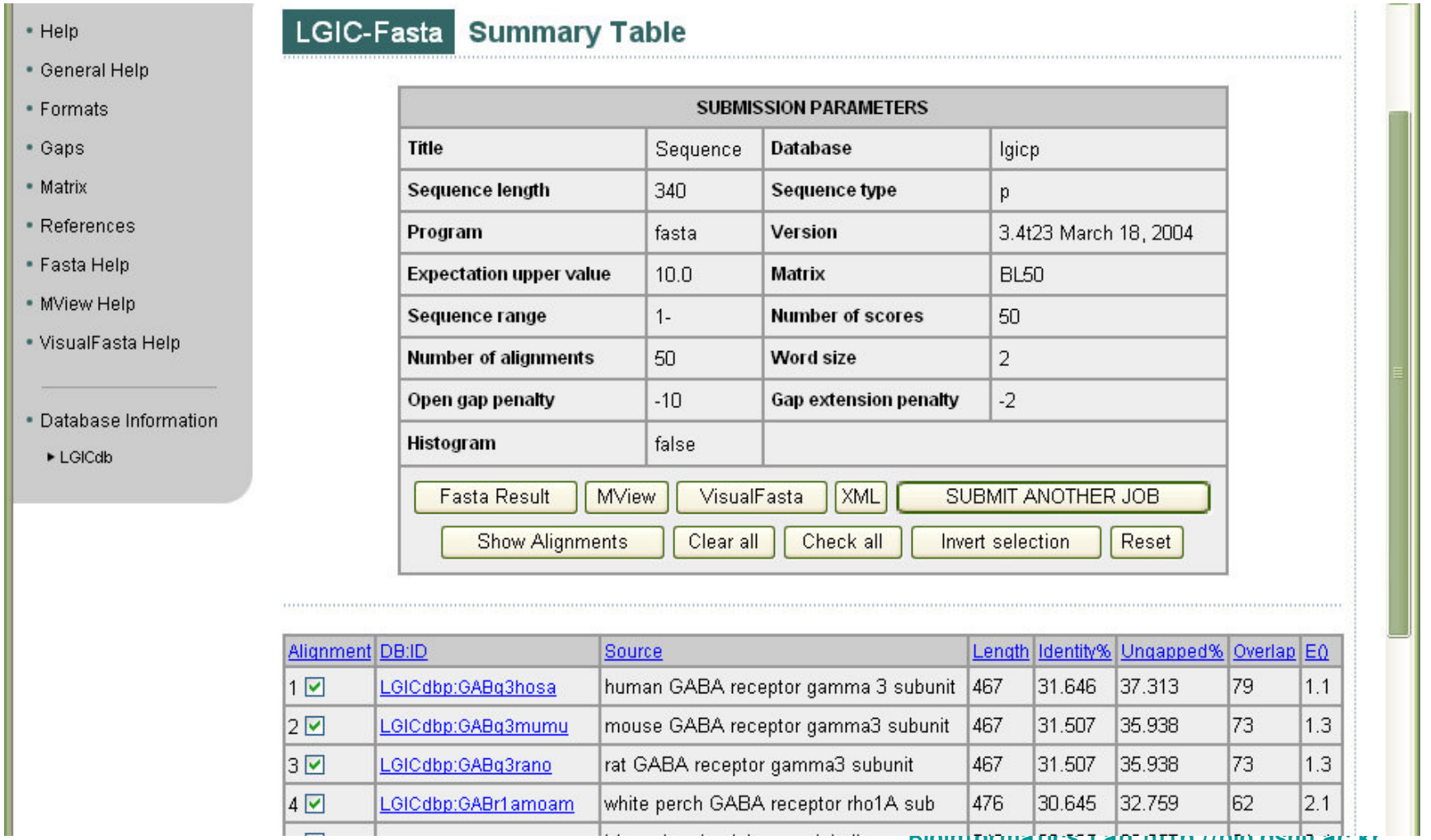
FASTA 실습: 작업수행 시키기

앞에서 준비한 340개의 서열을 복사 와 붙여넣기를 이용하여 데이터 입력창에 넣어주고 “Run Fasta3” 버튼을 누르고 작업이 끝날때 까지 기다린다 (이 작업은 몇 분 이내에 끝남).



FASTA 실습: 결과보기

작업이 끝나면 총11개의 검색결과가 다음과 같은 결과출력화면이 나타난다. 제일 앞에 Summary Table이 주어지고 결과를 살펴보는 여러가지 옵션이 주어진다. 요약표 아래에 주어진 각 리스트를 누르면 이름을 비롯한 그 단백질에 대한 다양한 정보가 주어져 있다.



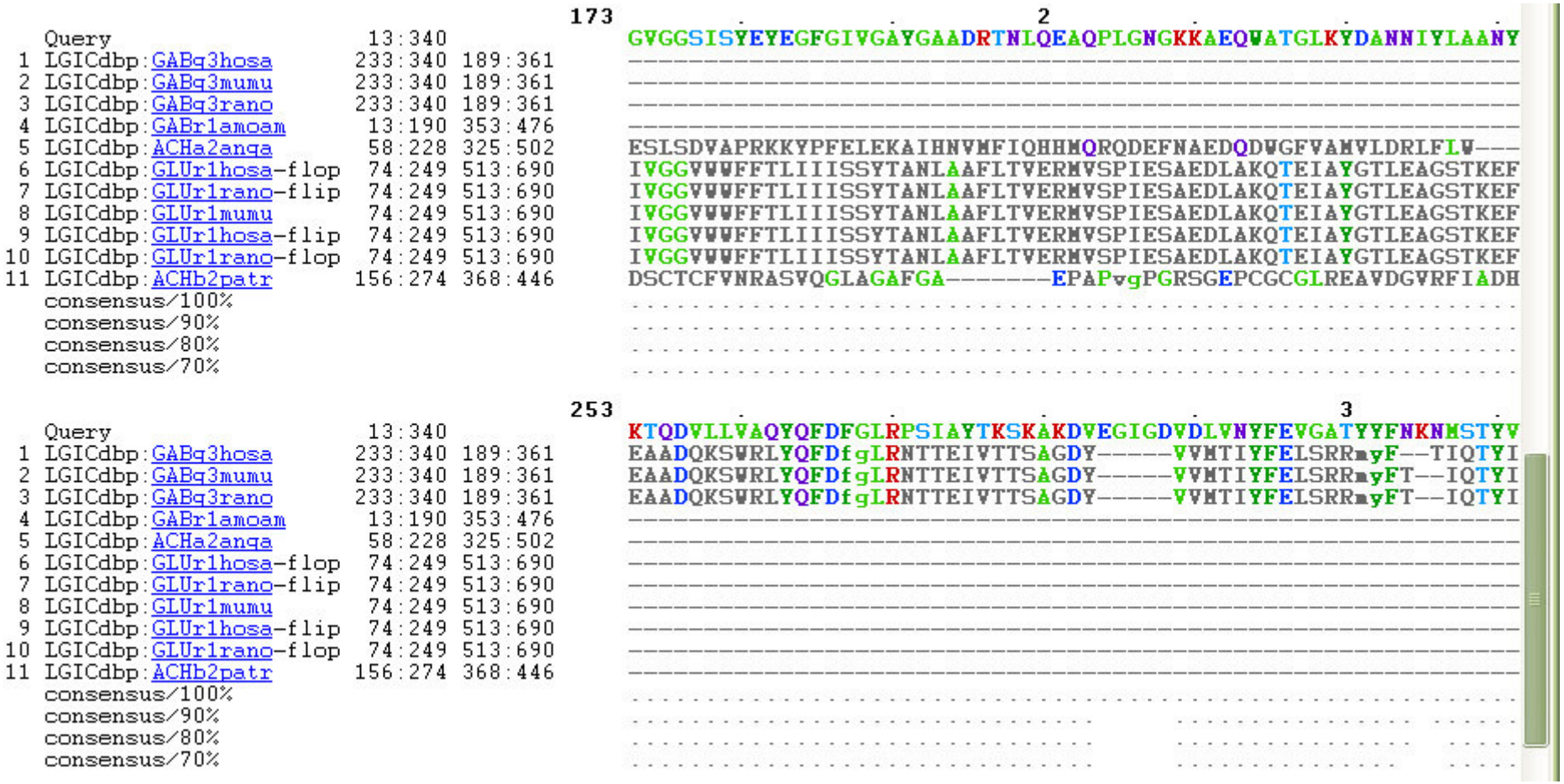
FASTA 실습: 결과보기/기본보기

결과를 살펴보는 여러가지 옵션중에서 “Fasta Result” 나 “Show Alignment”를 누르면 다음처럼 Text모드로 주어진 서열정렬 결과들을 보여 준다.

```
The best scores are:                                          opt bits E(606)
LGICdbp:GABg3hosa human GABA receptor gamma 3 sub  ( 467)   76   26.3     1.1
LGICdbp:GABg3mumu mouse GABA receptor gamma3 subu  ( 467)   75   26.1     1.3
LGICdbp:GABg3rano rat GABA receptor gamma3 subuni  ( 467)   75   26.1     1.3
LGICdbp:GABr1amoam white perch GABA receptor rho1  ( 476)   72   25.4     2.1
LGICdbp:ACHa2anga Mosquito nicotinic acetylcholin  ( 546)   72   25.5     2.4
LGICdbp:GLUr1hosa-flop human glutamate receptor G  ( 906)   72   25.7     3.4
LGICdbp:GLUr1rano-flip rat glutamate receptor GLU  ( 907)   72   25.7     3.4
LGICdbp:GLUr1mumu mouse glutamate receptor GLUR1   ( 907)   72   25.7     3.4
LGICdbp:GLUr1hosa-flip human glutamate receptor G  ( 907)   72   25.7     3.4
LGICdbp:GLUr1rano-flop rat glutamate receptor GLU  ( 907)   72   25.7     3.4
LGICdbp:ACHb2patr Chimpanzee nicotinic acetylchol  ( 446)   62   23.2     9.6

>>LGICdbp:GABg3hosa human GABA receptor gamma 3 subunit   (467 aa)
 initn:  39 init1:  39 opt:  76  Z-score: 94.5  bits: 26.3 E():  1.1
Smith-Waterman score: 76;  31.646% identity (37.313% ungapped) in 79 aa overlap (263-336:219

            240       250       260       270       280       290
Sequen ETRNATPITNKFTNTSGFANKTQDVLLVAQYQFDF-GLRPSIAYTKSKAKDVEGIGDVDL
                                   ::::: :::  .   . .: .:       : .
LGICdb FSSYGYPKEEMIYRWRKNSVEAADQKSWRLYQFDFMGLRNTTEIVTTSAGDY-----VVM
      190       200       210       220       230       240

             300       310       320       330        340
Sequen VNYFEVGATY-YFNKNMSTYVDYIINQIDSDNKLGVGSDDT---VAVGIVYQF
        . :::.  . ::   ..::.:  :.:  .  .   .. . .: :   .:.::
LGICdb TIYFELSRRMGYF--TIQTYIPCILTVVLSWVSFWIKKDATPARTALGITTVLTMTTLST
           250         260       270       280       290       300

LGICdb IARKSLPRVSYVTAMDLFVTVCFLFVFAALMEYATLNYYSSCRKPTTTKKTTSLLHPDSS
             310       320       330       340       350       360

>>LGICdbp:GABg3mumu mouse GABA receptor gamma3 subunit    (467 aa)
 initn:  39 init1:  39 opt:  75  Z-score: 93.3  bits: 26.1 E():  1.3
Smith-Waterman score: 75;  31.507% identity (35.938% ungapped) in 73 aa overlap (263-333:219

            240       250       260       270       280       290
Sequen ETRNATPITNKFTNTSGFANKTQDVLLVAQYQFDF-GLRPSIAYTKSKAKDVEGIGDVDL
                                   ::::: :::  .   . .: .:       : .
LGICdb FSSYGYPKEEMIYRWRKNSVEAADQKSWRLYQFDFMGLRNTTEIVTTSAGDY-----VVM
      190       200       210       220       230       240
```

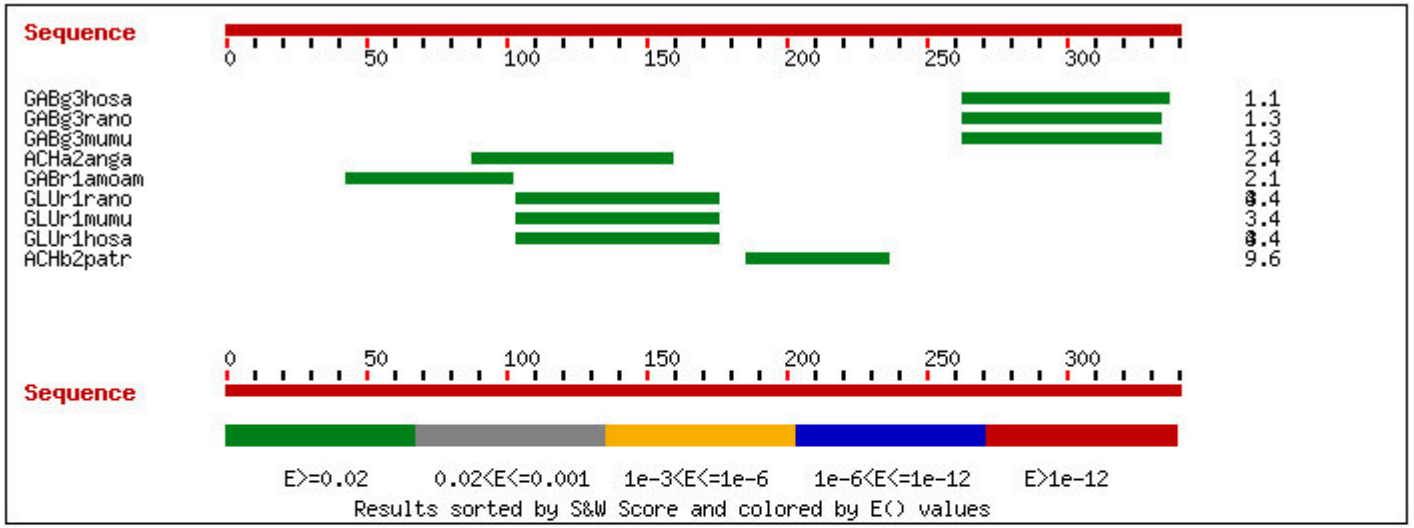
FASTA 실습: 결과보기/MView

옵션중에서 “MView” 누르면 다음처럼 모든서열들이 다중정렬된 결과물을 컬러로 보여 준다.



FASTA 실습: 결과보기/VisualFasta

옵션중에서 “VisualFasta” 누르면 다음처럼 모든서열들이 다중정렬된 결과물을 컬러로 유전자도표 처럼 보여 준다.




```
FASTA searches a protein or DNA sequence data bank
 version 3.4t23 March 18, 2004
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

Query library @ vs +lgicp library
searching /ebi/services/idata/v1331/fastadb/lgicp library

  1>>>Sequence - 340 aa
 vs  LGIC Protein library

 349836 residues in   606 sequences
  Expectation_n fit: rho(ln(x))= 4.6875+/-0.00924; mu= 10.5248+/- 0.578
 mean_var=67.9846+/-27.886, 0's: 0 Z-trim: 0  B-trim: 0 in 0/25
 Lambda= 0.155550
```

FASTA 실습: 1문자 코드표

FASTA에서 사용하는 핵산및 아미노산들의 1문자 코드표.

핵산 코드 테이블		아미노산 코드 테이블	
A 아데닌	U 유리딘	A 알라닌	P 프롤린
G 구아닌	R (G, A) 퓨린	C 시스테인	Q 글루타민
T 티민	Y (T, C) 피리미딘	D 아스파르트산	R 아르기닌
C 시토신	K (G, T) 케토	E 글루탐산	S 세린
	M (A, C) 아미노	F 페닐알라닌	T 트레오닌
	S (G, C) 강한 수소 결합 염기	G 글리신	V 발린
	W (T, A) 약한 수소 결합 염기	H 히스티틴	W 트립토판
	B G, T, C 중 하나를 의미	I 이소루신	Y 티로신
	D G, A, T 중 하나를 의미	K 라이신	U 셀레노 시스테인
	H A, C, T 중 하나를 의미	L 루신	B 아스파르트산 혹은 아스파라긴
	V G, C, A 중 하나를 의미	M 메티오닌	Z 글루탐산 혹은 글루타민
	N A, G, C, T 중 하나를 의미	N 아스파라긴	X 불특정 아미노산에 해당
	– 서열 간 갭을 의미		* 번역 정지 코돈
			– 서열 간 갭을 의미

강의 요약

1. 세계적인 서열검색 프로그램인 FASTA를 이용하여 E. coli OmpF Porin 단백질의 아미노산 서열과 연관된 서열을 갖는 단백질을 검색 하여 보았다.

2. 검색 결과 10개의 비슷한 서열을 발견할 수 있었다.

3. 1문자코드로 된 서열을 마련하는 방법도 소개하였다.