Computer Aided Drug Design

——Sequence Analysis

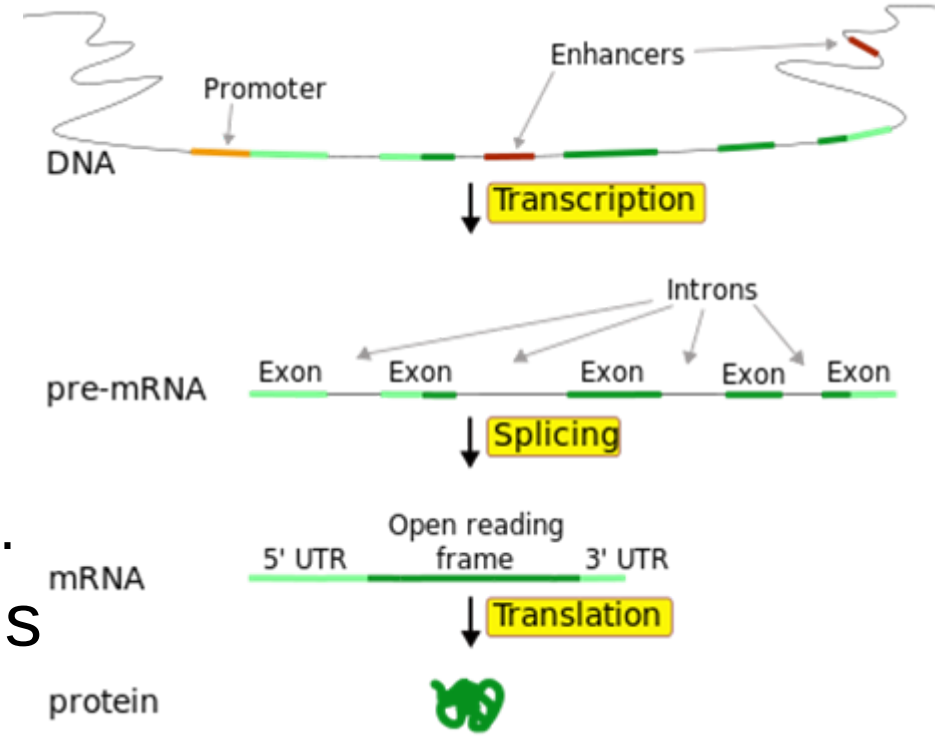
Qin Xu
http://cbb.sjtu.edu.cn/~qinxu/CADD.htm

Course Outline

- Introduction and Case Study
- Drug Targets
  - Sequence analysis
  - Protein structure prediction
  - Molecular simulation
- Drug Design
  - Combinatorial library
  - 3D-QSAR
  - Statistical methods
- Molecular Docking

Basic problems in sequence analysis

- Sequence alignment
- Genome construction
  - genes → proteins
  - the non-coding sequences
- Functional element finding
  - Exons, introns, promoters, …
- Regulatory network analysis
  - Transcription factors



Targets

Sequence Alignment

- DNA/RNA alignment
  - ACGT,U
- Protein alignment
  - 20 amino acids

- Global alignment and local alignment
  - Lower score

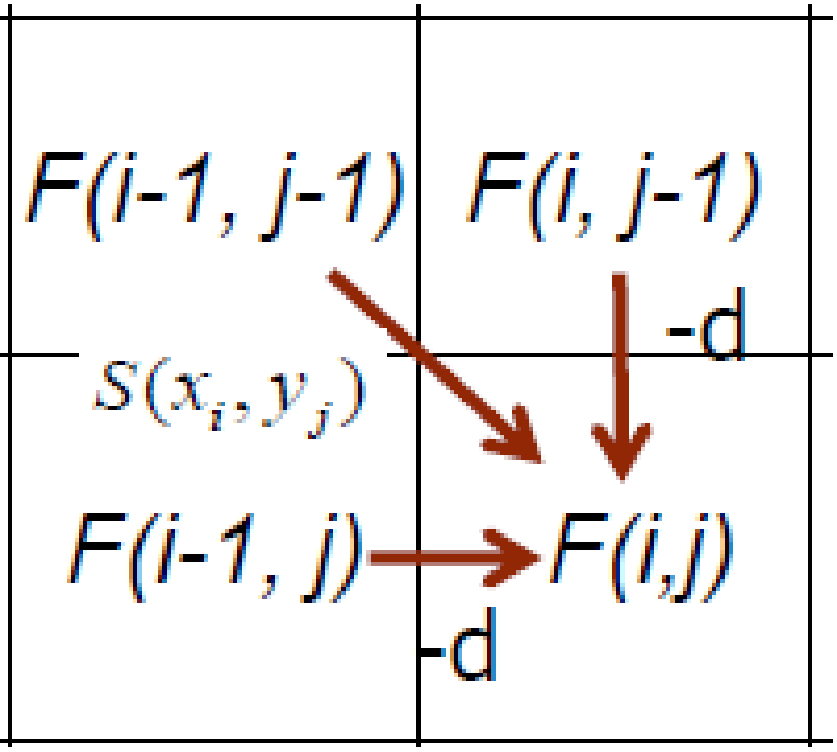
Dynamic programming sequence alignment algorithms

- Needleman/Wunsch global alignment
- Smith/Waterman local alignment
- Linear and affine gap penalties

Needleman/Wunsch global alignment

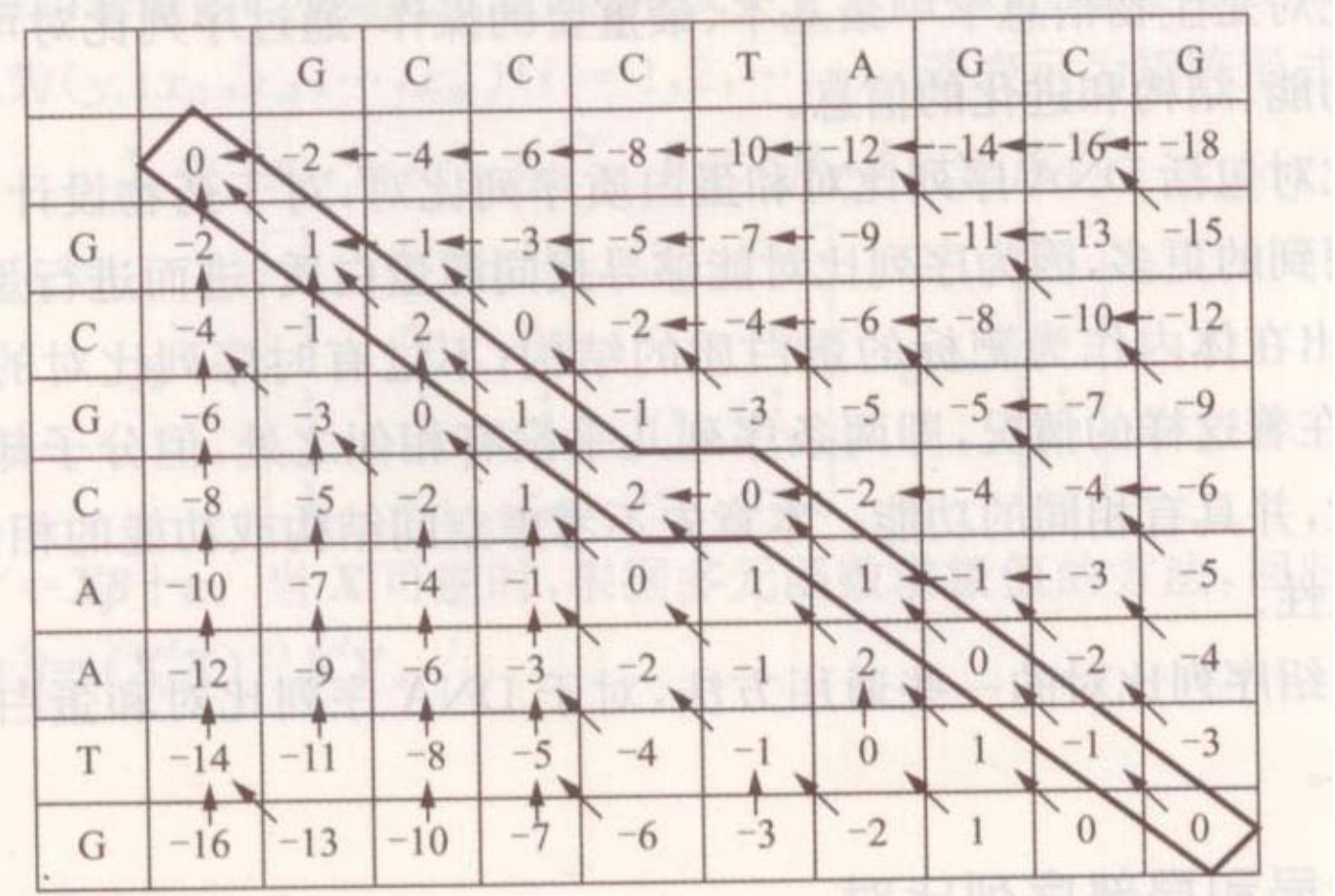
- Needleman, SB and Wunsch, CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins", *J. Mol. Biol.* 48:443-453, 1970.
- Two sequences $X = x_1 \ldots x_n$ and $Y = y_1 \ldots y_m$
- Let $F(i, j)$ be the optimal alignment score *of* $X_{1 \ldots i}$ of X up to x_i and $Y_{1 \ldots j}$ of Y up to Y_j *(*$0 \leqslant i \leqslant n$, $0 \leqslant j \leqslant m$*), then we have*

Needleman/Wunsch global alignment



$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$



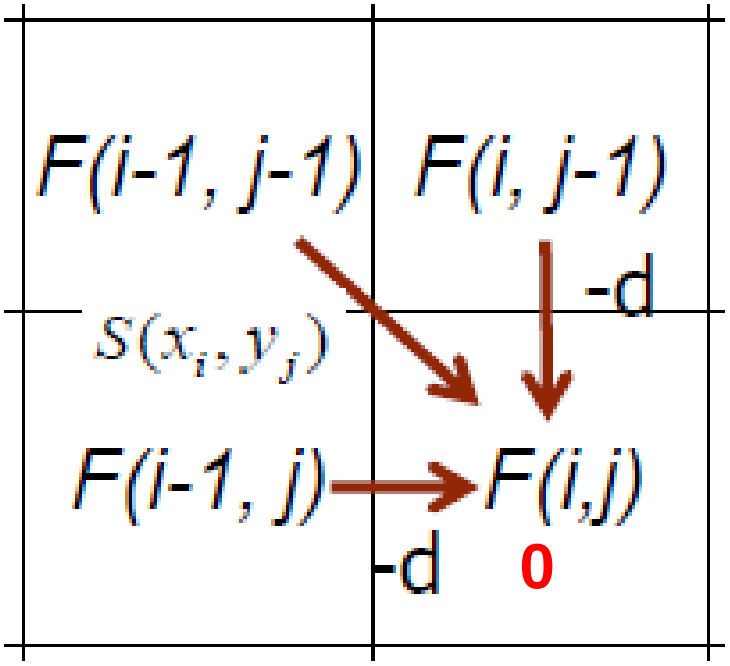
		G	C	C	C	T	A	G	C	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	-2	1	-1	-3	-5	-7	-9	-11	-13	-15
C	-4	-1	2	0	-2	-4	-6	-8	-10	-12
G	-6	-3	0	1	-1	-3	-5	-5	-7	-9
C	-8	-5	-2	1	2	0	-2	-4	-4	-6
A	-10	-7	-4	-1	0	1	1	-1	-3	-5
A	-12	-9	-6	-3	-2	-1	2	0	-2	-4
T	-14	-11	-8	-5	-4	-1	0	1	-1	-3
G	-16	-13	-10	-7	-6	-3	-2	1	0	0

图 2-1　Needleman-Wunsch 算法二维表

d=2; $s(x_i,y_j)$= 1 (same) | -1 (different)

Smith/Waterman local alignment

Smith, TF and Waterman, MS. Identification of Common Molecular Subsequences", *J. Mol. Biol.* 147: 195-197, 1981



$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} 0 \\ F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	0	1
C	0	0	2	1	1	0	0	0	2	0
G	0	1	0	1	0	0	0	1	0	3
C	0	0	2	1	2	0	0	0	2	1
A	0	0	0	1	0	1	1	0	0	1
A	0	0	0	0	0	0	2	0	0	0
T	0	0	0	0	0	1	0	1	0	0
G	0	1	0	0	0	0	0	1	0	1

图 2-2　Smith-Waterman 算法二维表

d=2; $s(x_i,y_j)$= 1 (same) | -1 (different)

Linear and affine gap penalties

- Linear: $w(k) = k\,d$
- Affine: $w(k) = d + (k-1)\,e$
- Let $M(i,j)$, $I_x(i,j)$, $I_y(i,j)$ be the best scores up to (i,j):
  - $M(i,j)$: x_i is aligned to y_j;
  - $I_x(i,j)$: x_i is aligned to a gap;
  - $I_y(i,j)$ y_j is aligned to a gap

then we have

$$M(i,j) = \max\begin{cases} M(i-1,j-1) + s(x_i, y_j), \\ I_x(i-1,j-1) + s(x_i, y_j), \\ I_y(i-1,j-1) + s(x_i, y_j); \end{cases}$$

$$I_x(i,j) = \max\begin{cases} M(i-1,j) - d, \\ I_x(i-1,j) - e; \end{cases}$$

$$I_y(i,j) = \max\begin{cases} M(i,j-1) - d, \\ I_y(i,j-1) - e. \end{cases}$$

Gap penalties

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

PAM

(Point accepted mutation matrix)

The BLOSUM62 matrix (**BLO**cks **SU**bstitution **M**atrix)

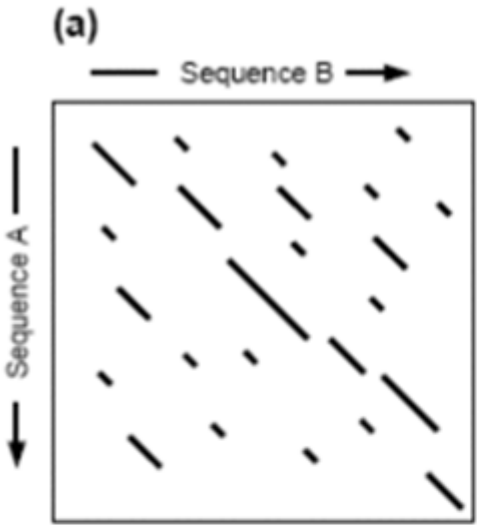
Improved algorithms

- Gotoh, O. An improved algorithm for matching biological sequences, *J. Mol. Biol.* 162:705-708, 1982
  - More efficient Needleman/Wunsch and Smith/Waterman algorithms.
- Myers, E. W. and Miller, W. Optimal alignment in linear space", *CABIOS* 4: 11-17, 1988.
  - A method to reduce the memory cost from O(n^2) to O(n)
- Henikoff, S. and Henikoff, J.G. Amino Acid Substitution Matrices from Protein Blocks. *PNAS* 89 (22): 10915–10919, 1992.
  - BLOSUM, gap penalty matrix for sequence alignment of proteins.

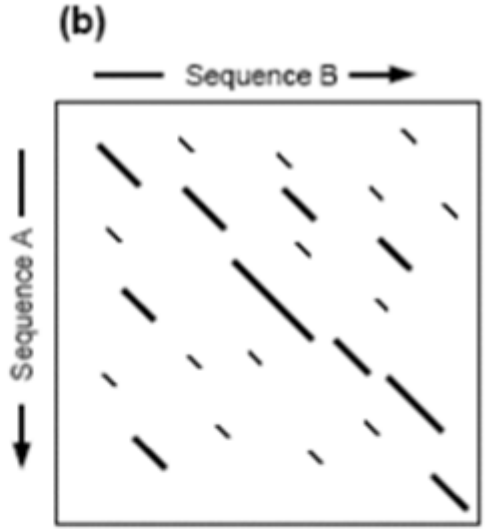
Heuristic algorithms

- Trading optimality, completeness, accuracy, or precision for speed
  - Much faster
  - May completely miss the optimal alignment
- How to save work?
  - Seed: find significant gap-less matches and then extend them
  - Searching in the diagonal band of the matrix
- Two important algorithms for alignment
  - FASTA
  - BLAST

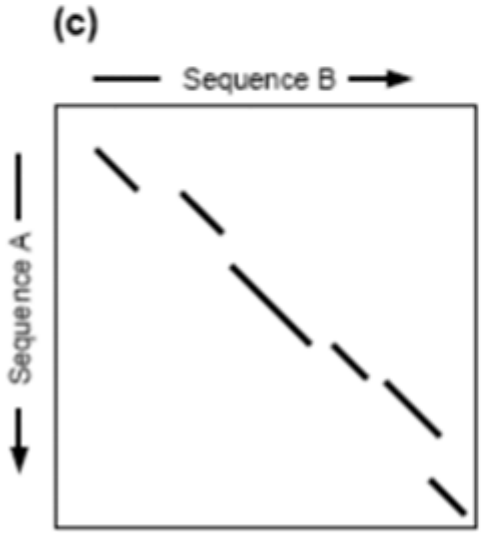
FASTA Algorithm



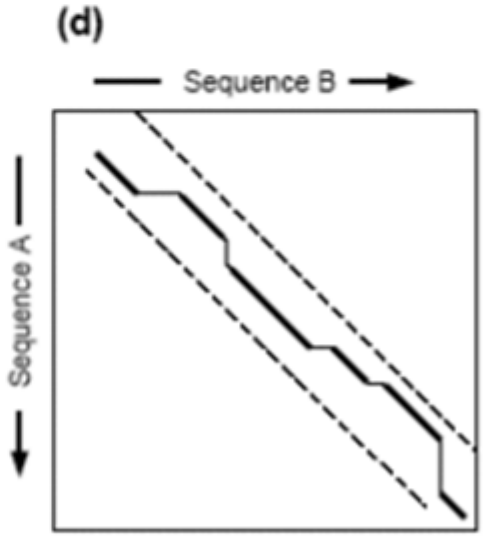
Find runs of identities



Re-score using PAM matrix
Keep top scoring segments.



Apply "joining threshold"
to eliminate segments that
are unlikely to be part of the alignment
that includes highest scoring segment.



Use dynamic programming
to optimise the alignment in a
narrow band that encompasses
the top scoring segments.

FASTA

- FASTA = "FAST-All" = "FAST-P" (protein) + "FAST-N" (nucleotide)
- FASTA file format

  >1SIO:A|PDBID|CHAIN|SEQUENCE

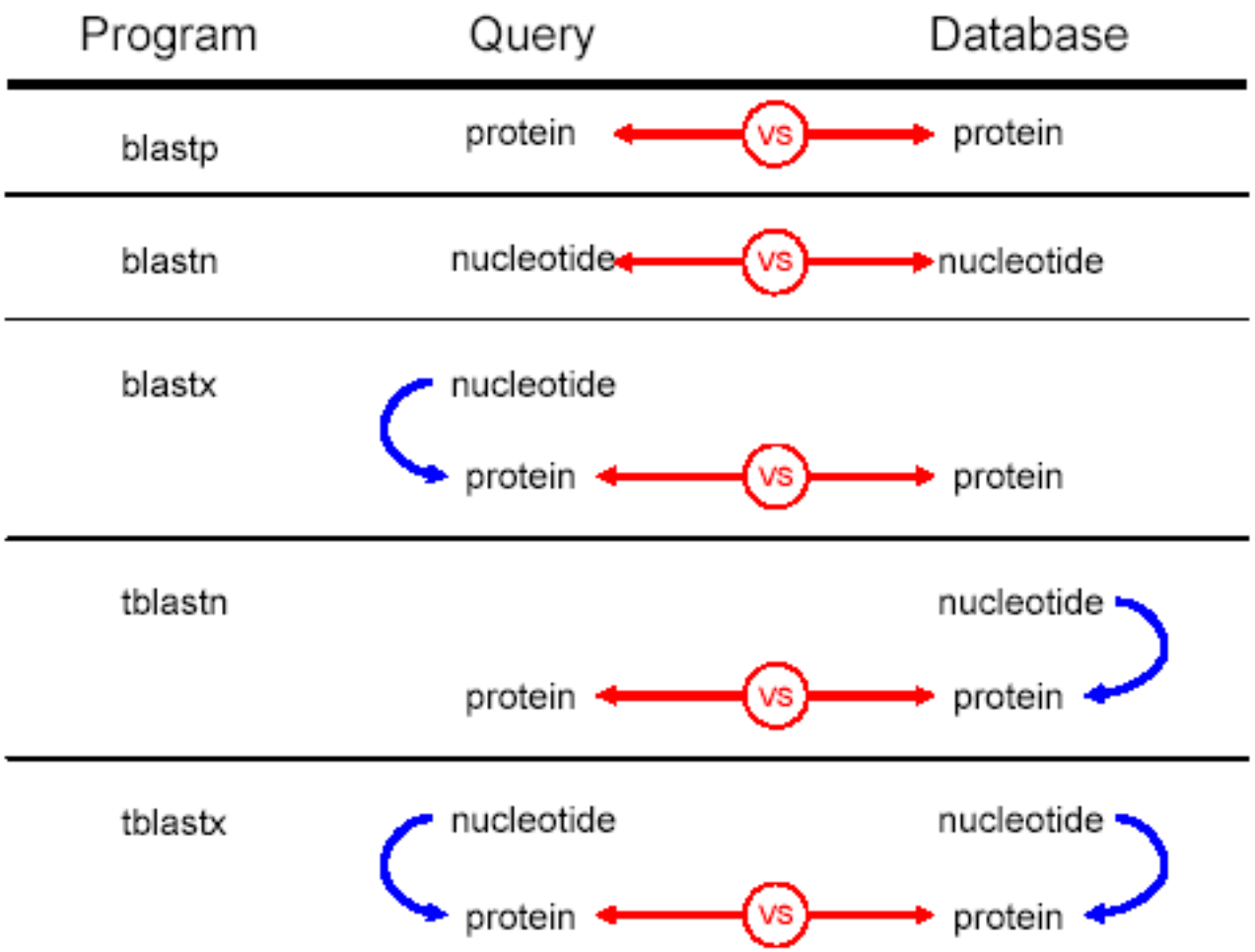
  AAPTAYTPLDVAQAYQFPEGLDGQGQCIAIIELGGGYD
  EASLAQYFASLGVPAPQVVSVSVDGASNQPTGDPSGP
  DGEVELDIEVAGALAPGAKFAVYFAPNTDAGFLDAITTA
  IHDPTLKPSVVSISWGGPEDSWTSAAIAAMNRAFLDAA
  ALGVTVLAAAGDSGSTDGEQDGLYHVDFPAASPYVLA
  CGGTRLVASGGRIAQETVWNDGPDGGATGGGVSRIF
  PLPAWQEHANVPPSANPGASSGRGVPDLAGNADPAT
  GYEVVIDGEATVIGGTSAVAPLFAALVARINQKLGKAVG
  YLNPTLYQLPADVFHDITEGNNDIANRAQIYQAGPGWD
  PCTGLGSPIGVRLLQALLPSASQPQP

Basic Local Alignment Search Tool (BLAST)

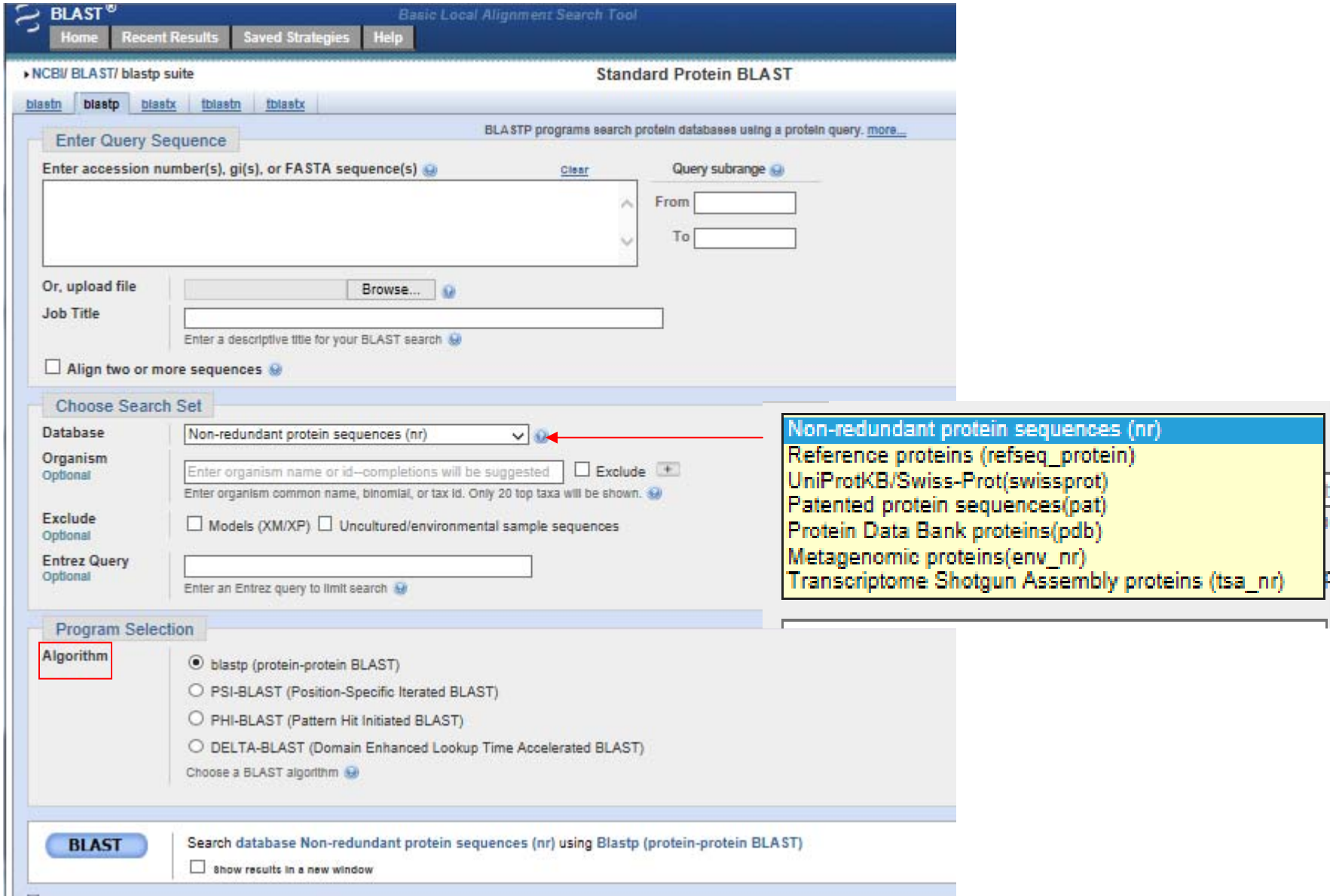
- **Publications:**
  - **Ungapped BLAST** – Altschul et al., 1990
  - **Gapped BLAST, PSI-BLAST** - Altschul et al., 1997
- **Input:**
  - **Query (target) sequence：** DNA, RNA or Protein
  - **Scoring Scheme** – gap penalties, substitution matrix for proteins, identity/mismatch scores for DNA/RNA
  - **Word length W** – typical is **W**=3 for proteins and **W**=11 for DNA/RNA
- **Output:**
  - Statistically significant matches

Sequencing by BLAST

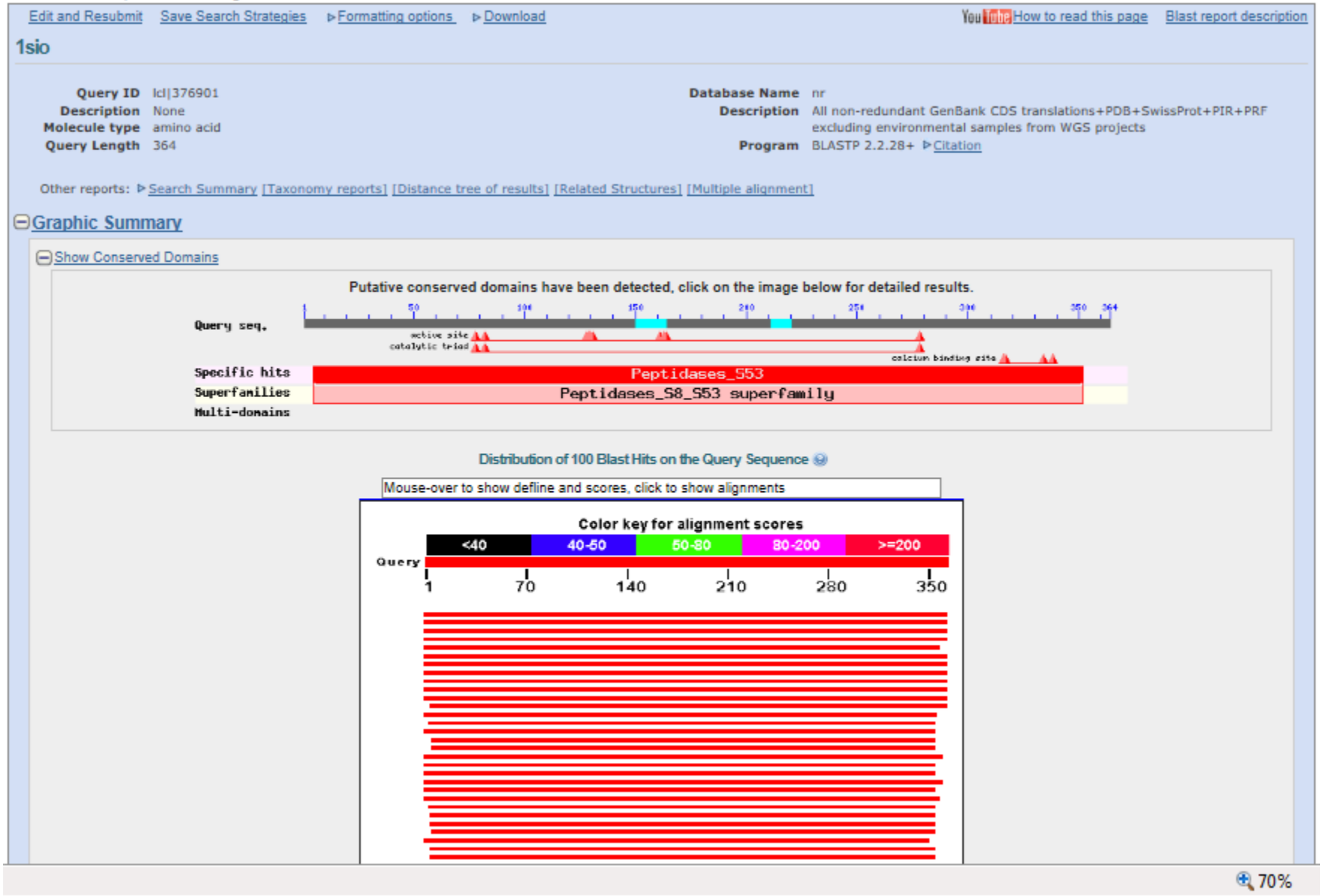
Different programs are available according to the type of query



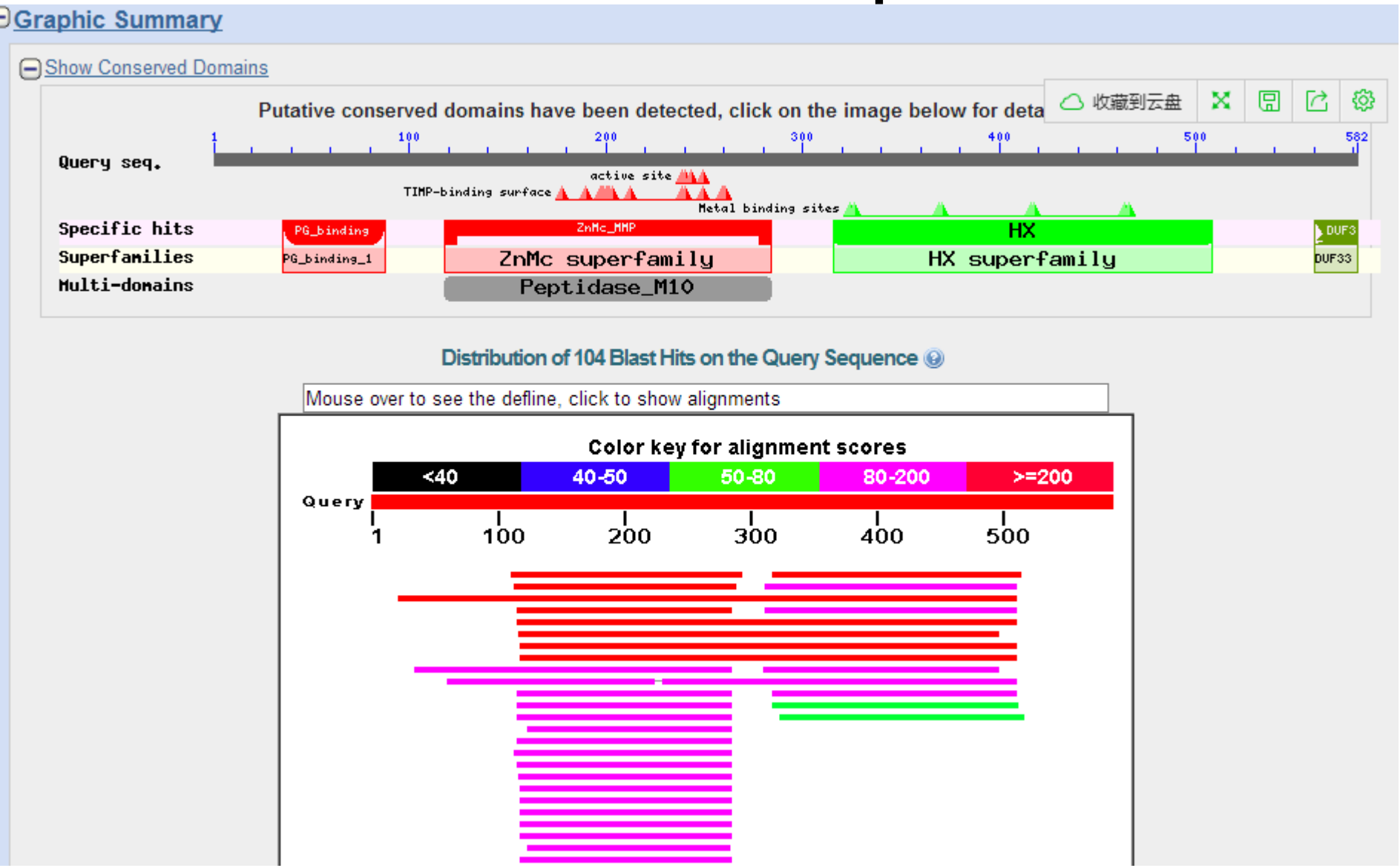
Standard Protein BLAST



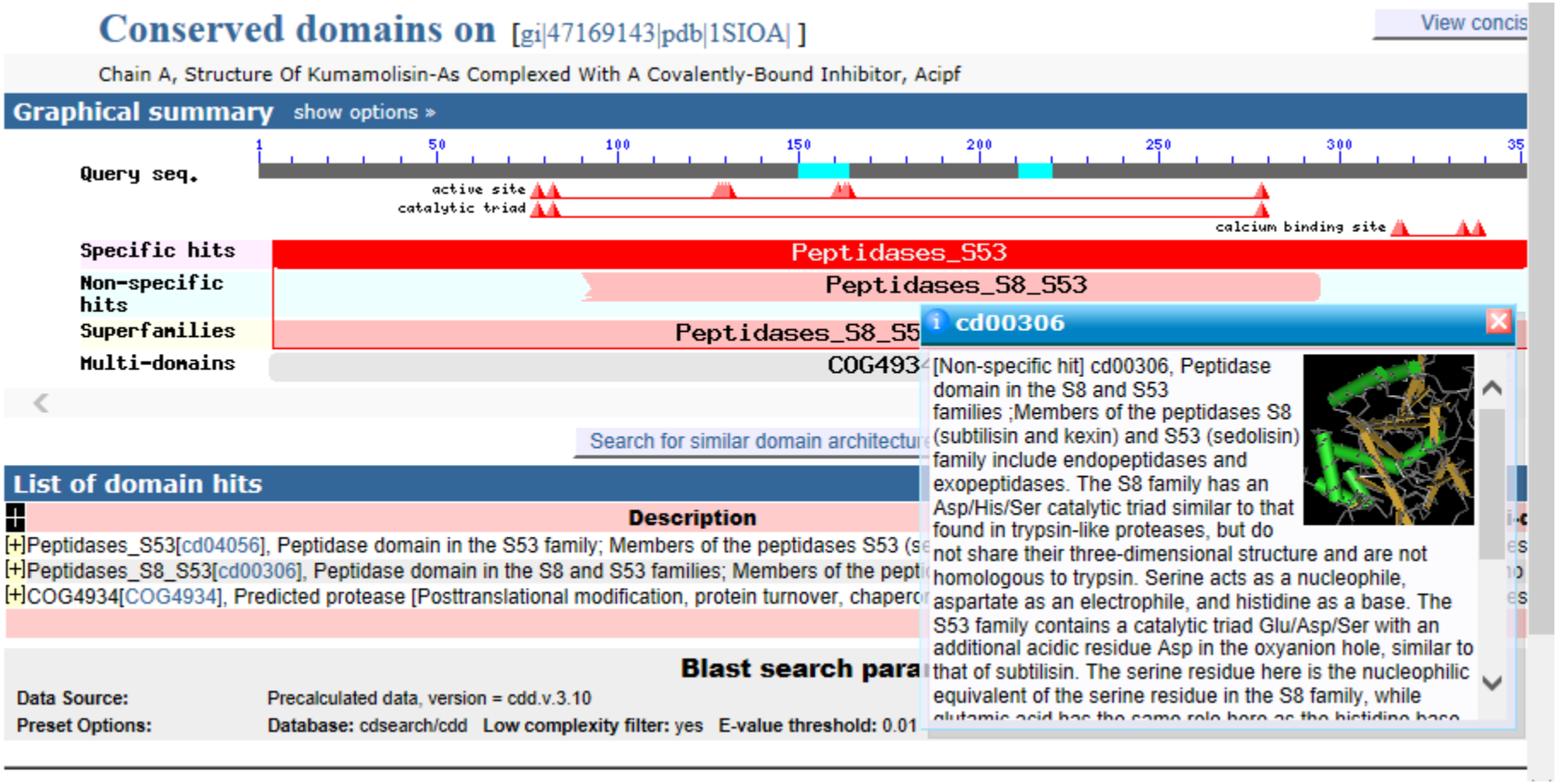
Results of BLAST



Results of BLASTp on PDB



Results about conserved domain



Description of conserved domain

List of domain hits

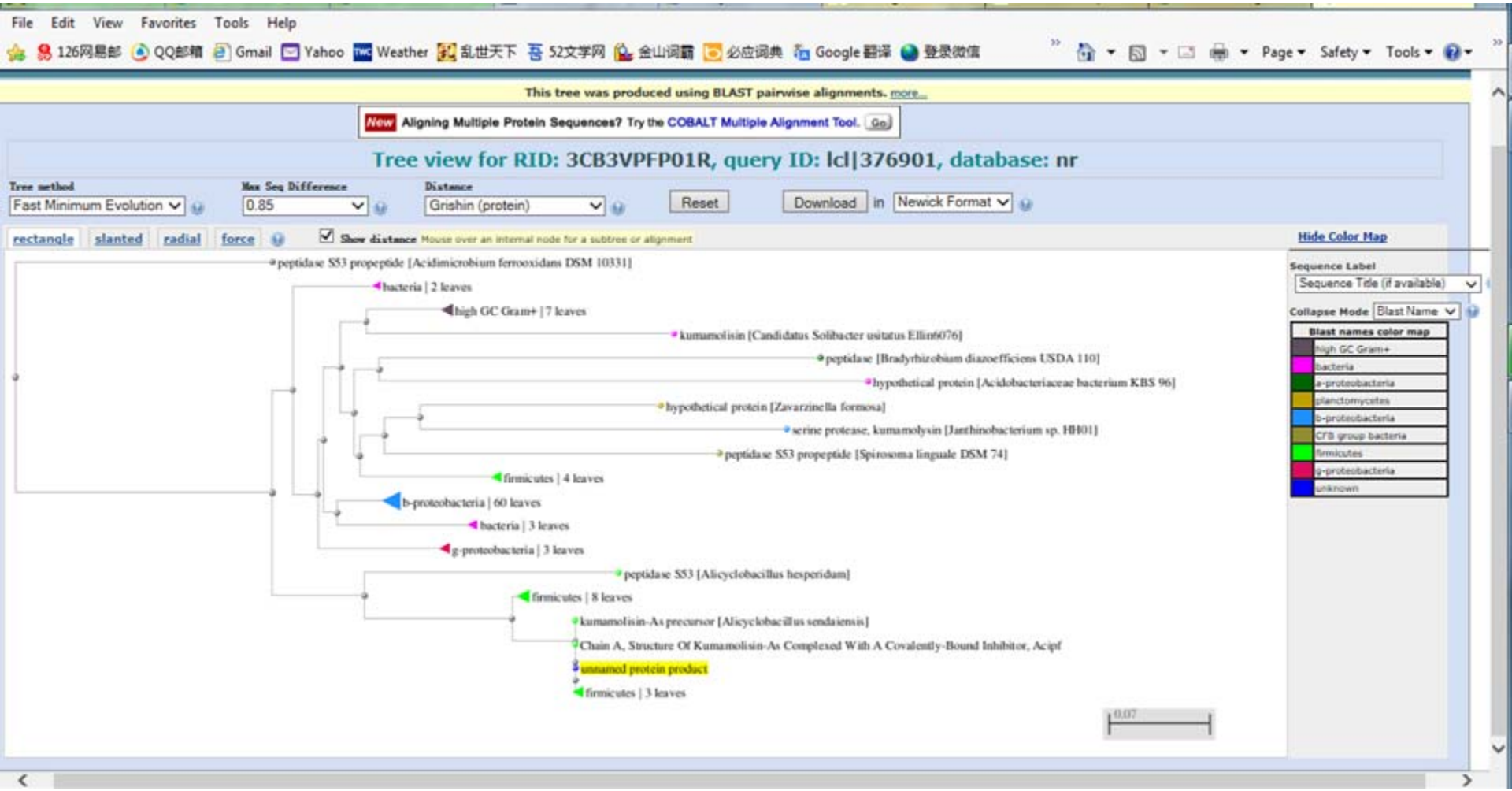
	Description	PssmId	Multi-
[–]	Peptidases_S53[cd04056], Peptidase domain in the S53 family; Members of the peptidases S53 (sedolisin) family include endopeptidases and exo	173788	yes

Peptidase domain in the S53 family; Members of the peptidases S53 (sedolisin) family include endopeptidases and exopeptidases sedolisin, kumamolysin, and (PSCP insensitive Carboxyl Proteinase. The S53 family contains a catalytic triad Glu/Asp/Ser with an additional acidic residue Asp in the oxyanion hole, similar to that of Asn stability of these enzymes may be enhanced by calcium, some members have been shown to bind up to 4 ions via binding sites with different affinity. Some members contain disulfide bonds. These enzymes can be intra- and extracellular, some function at extreme temperatures and pH values. Characterized sedolisins include Kuma extracellular calcium-dependent thermostable endopeptidase from Bacillus. The enzyme is synthesized with a 188 amino acid N-terminal preprotein region which is cl extraction into the extracellular space with low pH. One kumamolysin paralog, kumamolisin-As, is believed to be a collagenase. TPP1 is a serine protease that functio tripeptidyl exopeptidase as well as an endopeptidase. Less is known about PSCP from Pseudomonas which is thought to be an aspartic proteinase.

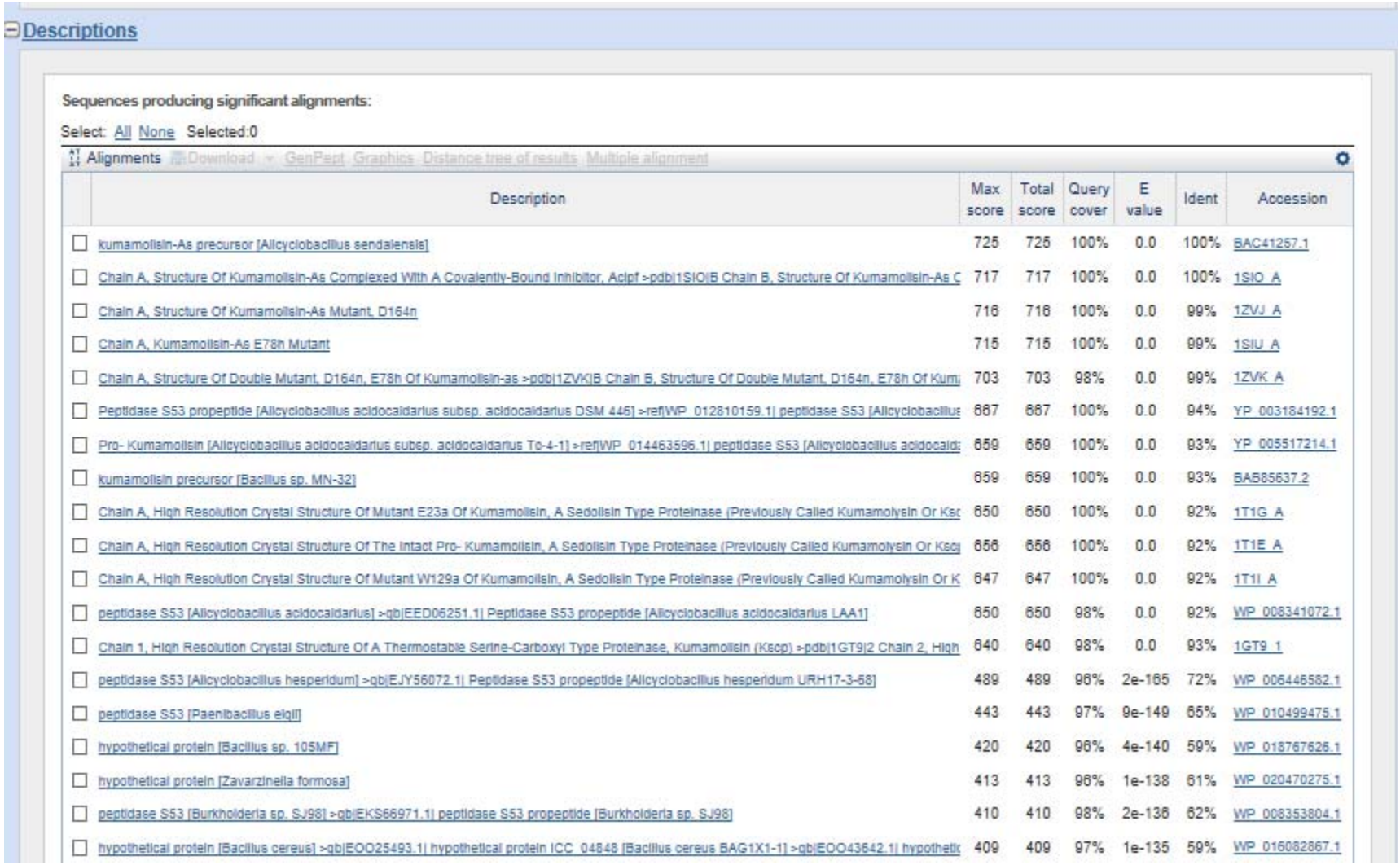
Cd Length: 361 **Bit Score:** 321.19 **E-value:** 4.01e-107

```
                    10        20        30        40        50        60        70        80
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1SIOA         5 AYTPLDVAQAYQFPE-GLDGQGQCIAIIELGGGYDEASLAQYFASL-GVPAPQVVSVSVDGASNqPTGDPSGPDGEVELD 82
Cd:cd04056    1 GYTPADLAALYNIPPlGYTGSGQTIGIIEFGGGYYNPSDLQTFFQLfGLPAPTVFIVVVIGGGN-APGTSSGWGGEASLD 79

                    90       100       110       120       130       140       150       160
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1SIOA        83 IEVAGALAPGAKFAVYFAPNT-DAGFLDAITTAIHDPTLKPSVVSISWGGPEDSWTSAAIAAMNRAFLDAAALGVTVLAA 161
Cd:cd04056   80 VEYAGAIAPGANITLYFAPGTvTNGPLLAFLAAVLDNPNLPSVISISYGEPEQSLPPAYAQRVCNLFAQAAAQGITVLAA 159

                   170       180       190       200       210       220       230       240
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1SIOA       162 AGDSGSTDGEQDG---LYHVDFPAASPYVLACGGTRLVA-SGGRIAQETVWNDGPDGGATGGGVSRIFPLPAWQEHA-NV 236
Cd:cd04056  160 SGDSGAGGCGGDGsgtGFSVSFPASSPYVTAVGGTTLYTgGTGSSAESTVWSSEGGWGGSGGGFSNYFPRPSYQSGAvLG 239

                   250       260       270       280       290       300       310       320
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1SIOA       237 PPSANPGASSGRGVPDLAGNADPATGYEVVIDGEATVIGGTSAVAPLFAALVARINQKLGKA----VGYLNPTLYQLPAD 312
Cd:cd04056  240 LPPSGLYNGSGRGVPDVAANADPGTGYLVVVNGQWYLVGGTSAAAPLFAGLIALINQARLAAgkppLGFLNPLLYQLAAT 319

                   330       340       350       360
                    ....*....|....*....|....*....|....*....|...
1SIOA       313 ---VFHDITEGNNDIANrAQIYQAGPGWDPCTGLGSPIGVRLL 352
Cd:cd04056  320 apsAFNDITSGNNGGCG-GAGYPAGPGWDPVTGLGTPNFAKLL 361
```

Distance tree



Descriptions of searched results



Descriptions

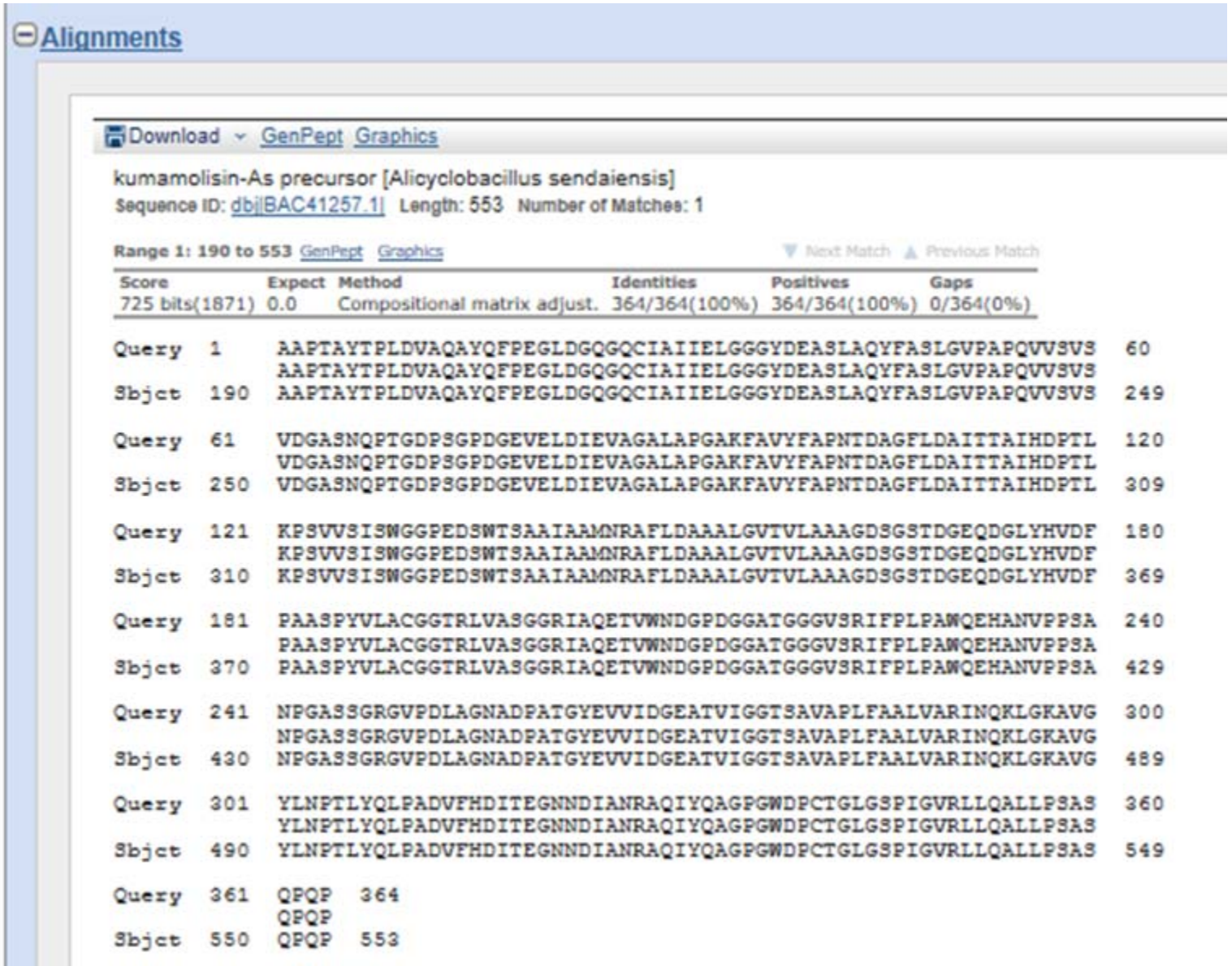
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
kumamolisin-As precursor [Alicyclobacillus sendaiensis]	725	725	100%	0.0	100%	BAC41257.1
Chain A, Structure Of Kumamolisin-As Complexed With A Covalently-Bound Inhibitor, Aclpf >pdb|1SIO|B Chain B, Structure Of Kumamolisin-As C	717	717	100%	0.0	100%	1SIO_A
Chain A, Structure Of Kumamolisin-As Mutant, D164n	716	716	100%	0.0	99%	1ZVJ_A
Chain A, Kumamolisin-As E78h Mutant	715	715	100%	0.0	99%	1SIU_A
Chain A, Structure Of Double Mutant, D164n, E78h Of Kumamolisin-as >pdb|1ZVK|B Chain B, Structure Of Double Mutant, D164n, E78h Of Kumi	703	703	98%	0.0	99%	1ZVK_A
Peptidase S53 propeptide [Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446] >ref|WP_012810159.1| peptidase S53 [Alicyclobacillus	667	667	100%	0.0	94%	YP_003184192.1
Pro- Kumamolisin [Alicyclobacillus acidocaldarius subsp. acidocaldarius Tc-4-1] >ref|WP_014463596.1| peptidase S53 [Alicyclobacillus acidocald	659	659	100%	0.0	93%	YP_005517214.1
kumamolisin precursor [Bacillus sp. MN-32]	659	659	100%	0.0	93%	BAB85637.2
Chain A, High Resolution Crystal Structure Of Mutant E23a Of Kumamolisin, A Sedolisin Type Proteinase (Previously Called Kumamolysin Or Ksc	650	650	100%	0.0	92%	1T1G_A
Chain A, High Resolution Crystal Structure Of The Intact Pro- Kumamolisin, A Sedolisin Type Proteinase (Previously Called Kumamolysin Or Kscp	656	656	100%	0.0	92%	1T1E_A
Chain A, High Resolution Crystal Structure Of Mutant W129a Of Kumamolisin, A Sedolisin Type Proteinase (Previously Called Kumamolysin Or K	647	647	100%	0.0	92%	1T1I_A
peptidase S53 [Alicyclobacillus acidocaldarius] >gb|EED06251.1| Peptidase S53 propeptide [Alicyclobacillus acidocaldarius LAA1]	650	650	98%	0.0	92%	WP_008341072.1
Chain 1, High Resolution Crystal Structure Of A Thermostable Serine-Carboxyl Type Proteinase, Kumamolisin (Kscp) >pdb|1GT9|2 Chain 2, High	640	640	98%	0.0	93%	1GT9_1
peptidase S53 [Alicyclobacillus hesperidum] >gb|EJY56072.1| Peptidase S53 propeptide [Alicyclobacillus hesperidum URH17-3-68]	489	489	96%	2e-165	72%	WP_006446582.1
peptidase S53 [Paenibacillus elgii]	443	443	97%	9e-149	65%	WP_010499475.1
hypothetical protein [Bacillus sp. 10SMF]	420	420	96%	4e-140	59%	WP_018767626.1
hypothetical protein [Zavarzinella formosa]	413	413	96%	1e-138	61%	WP_020470275.1
peptidase S53 [Burkholderia sp. SJ98] >gb|EKS66971.1| peptidase S53 propeptide [Burkholderia sp. SJ98]	410	410	98%	2e-136	62%	WP_006353804.1
hypothetical protein [Bacillus cereus] >gb|EOO25493.1| hypothetical protein ICC_04848 [Bacillus cereus BAG1X1-1] >gb|EOO43642.1| hypothetic	409	409	97%	1e-135	59%	WP_016082867.1

Alignment



Alignments

Download GenPept Graphics

kumamolisin-As precursor [Alicyclobacillus sendaiensis]

Sequence ID: dbj|BAC41257.1| Length: 553 Number of Matches: 1

Range 1: 190 to 553 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
725 bits(1871)	0.0	Compositional matrix adjust.	364/364(100%)	364/364(100%)	0/364(0%)

```
Query  1    AAPTAYTPLDVAQAYQFPEGLDGQGQCIAIIELGGGYDEASLAQYFASLGVPAPQVVSVS  60
            AAPTAYTPLDVAQAYQFPEGLDGQGQCIAIIELGGGYDEASLAQYFASLGVPAPQVVSVS
Sbjct  190  AAPTAYTPLDVAQAYQFPEGLDGQGQCIAIIELGGGYDEASLAQYFASLGVPAPQVVSVS  249

Query  61   VDGASNQPTGDPSGPDGEVELDIEVAGALAPGAKFAVYFAPNTDAGFLDAITTAIHDPTL  120
            VDGASNQPTGDPSGPDGEVELDIEVAGALAPGAKFAVYFAPNTDAGFLDAITTAIHDPTL
Sbjct  250  VDGASNQPTGDPSGPDGEVELDIEVAGALAPGAKFAVYFAPNTDAGFLDAITTAIHDPTL  309

Query  121  KPSVVSISWGGPEDSWTSAAIAAMNRAFLDAAALGVTVLAAAGDSGSTDGEQDGLYHVDF  180
            KPSVVSISWGGPEDSWTSAAIAAMNRAFLDAAALGVTVLAAAGDSGSTDGEQDGLYHVDF
Sbjct  310  KPSVVSISWGGPEDSWTSAAIAAMNRAFLDAAALGVTVLAAAGDSGSTDGEQDGLYHVDF  369

Query  181  PAASPYVLACGGTRLVASGGRIAQETVWNDGPDGGATGGGVSRIFPLPAWQEHANVPPSA  240
            PAASPYVLACGGTRLVASGGRIAQETVWNDGPDGGATGGGVSRIFPLPAWQEHANVPPSA
Sbjct  370  PAASPYVLACGGTRLVASGGRIAQETVWNDGPDGGATGGGVSRIFPLPAWQEHANVPPSA  429

Query  241  NPGASSGRGVPDLAGNADPATGYEVVIDGEATVIGGTSAVAPLFAALVARINQKLGKAVG  300
            NPGASSGRGVPDLAGNADPATGYEVVIDGEATVIGGTSAVAPLFAALVARINQKLGKAVG
Sbjct  430  NPGASSGRGVPDLAGNADPATGYEVVIDGEATVIGGTSAVAPLFAALVARINQKLGKAVG  489

Query  301  YLNPTLYQLPADVFHDITEGNNDIANRAQIYQAGPGWDPCTGLGSPIGVRLLQALLPSAS  360
            YLNPTLYQLPADVFHDITEGNNDIANRAQIYQAGPGWDPCTGLGSPIGVRLLQALLPSAS
Sbjct  490  YLNPTLYQLPADVFHDITEGNNDIANRAQIYQAGPGWDPCTGLGSPIGVRLLQALLPSAS  549

Query  361  QPQP  364
            QPQP
Sbjct  550  QPQP  553
```

CLUSTAL

- Multiple sequence alignment
  - Do a pairwise alignment
  - Create a (or use a user-defined tree)
  - Use the guide tree to carry out a multiple alignment
- Three main version
  - ClustalW: command line interface
  - ClustalX: graphical user interface
  - Clustal Omega: the latest online version

http://www.clustal.org/

CLUSTAL

Clustal: Multiple Sequence Alignment

Multiple alignment of nucleic acid and protein sequences

Clustal Omega

- Latest version of Clustal - fast and scalable (can align hundreds of thousands of sequences in hours), greater accuracy due to new HMM alignment engine
- Command line/web server only (GUI public beta available soon)

ClustalW/ClustalX

- "Classic Clustal"
- GUI (ClustalX), command line (ClustalW), web server versions available

http://www.clustal.org/

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments.

STEP 1 - Enter your input sequences

Enter or paste a set of PROTEIN sequences in any supported format:

```
AAPTAYTPLDVAQAYQFPEGLDGQGQCIAIIELGGGYDEASLAQYFASLGVPAPQVVSVSVDGASNQPTGDPSGPDGEVE
LDIEVAGALAPGAKFAVYFAPNTDAGFLDAITTAIHDPTLKPSVVSISWGGPEDSWTSAAIAAMNRAFLDAAALGVTVLA
AAGDSGSTDGEQDGLYHVDFPAASPYVLACGGTRLVASGGRIAQETVWNDGPDGGATGGGVSRIFPLPAWQEHANVPPSA
NPGASSGRGVPDLAGNADPATGYEVVIDGEATVIGGTSAVAPLFAALVARINQKLGKAVGYLNPTLYQLPADVFHDITEG
NNDIANRAQIYQAGPGWDPCTGLGSPIGVRLLQALLPSASQPQP
```

Or, upload a file: Browse...

STEP 2 - Set your parameters

OUTPUT FORMAT Clustal w/o numbers

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

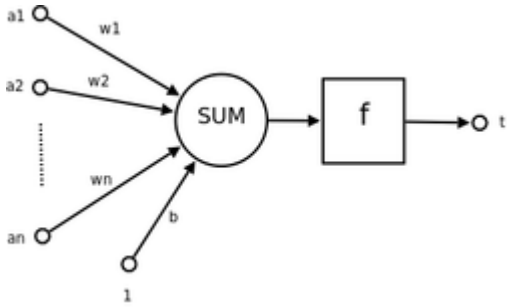
Submit

Course Outline

- Introduction and Case Study
- Drug Targets
  - Sequence analysis
  - Protein structure prediction
  - Molecular simulation
- Drug Design
  - Combinatorial library
  - 3D-QSAR
  - Statistical methods
- Molecular Docking

Classes of Secondary Structural Prediction

- Chou-Fasman
  - P_α, P_β, P_{turn}
  - Overlap
- Amino Acid Hydrophobicity
- k-Nearest Neighbor algorithm (kNN，最邻近法，相似片段法)
- Artificial Neural Networks（ANNs，人工神经网络）



Amino Acid Hydrophobicity

AA	Abrr.	Hydrophobicity	AA	Abrr.	Hydrophobicity
Ala	A	1.8	Leu	L	3.8
Arg	R	-4.5	Lys	K	-3.9
Asn	N	-3.5	Met	M	1.9
Asp	D	-3.5	Phe	F	2.8
Cys	C	2.5	Pro	P	-1.6
Gln	Q	-3.5	Ser	S	-0.8
Glu	E	-3.5	Thr	T	-0.7
Gly	G	-0.4	Trp	W	-0.9
His	H	-3.2	Tyr	Y	-1.3
Ile	I	4.5	Val	V	4.2

Methods of Secondary Structural Prediction

- DPM（双重预测方法）
- DSC
- PHDsec
- SOMPA
- MLRC
- Jpred
- Predict Protein

Jpred

http://www.compbio.dundee.ac.uk/~www-jpred/

Jpred 3
Incorporating Jnet

A Secondary Structure Prediction Server

Update (19 Aug 2014): JNet v.2.3.1. Please see news page for details.

Sequence:

Help
Advanced

Make Prediction Clear

The Barton Group - The University of Dundee

Citation: Cole C, Barber JD & Barton GJ. *Nucleic Acids Res.* **2008**. 35 (suppl. 2) W197-W201 [link]
More citations

Jpred results

```
                  : 1---------11--------21--------31--------41--------51--------61--------71--------81--------91--------101-------111-
OrigSeq           : MSPAPRPPRCLLLPLLTLGTALASLGSAQSSSFSPEAWLQQYGYLPPGDLRTHTQRSPQSLSAAIAAMQKFYGLQVTGKADADTMKAMRRPRCGVPDKFGAEIKANVRRKRYAI

Jnet              : ------HHHHHHHHHHHHHH----------HHHHHHHHHHH---------------HHHHHHHHHHHHHH----------HHHHHHHHH-----------------EE------
jhmm              : ---HHHHHHHHHHHHHHHH----------HHHHHHHHHHH---------------HHHHHHHHHHHHHH----------HHHHHHHHH-----------------EE------
jpssm             : -------HHHHHHHHHHHHH----------HHHHHHHHHH---------------HHHHHHHHHHHHHH----------HHHHHHHHH-----------------E-------

Lupas 14          : ------------------------------------------------------------------------------------------------------------------------
Lupas 21          : ------------------------------------------------------------------------------------------------------------------------
Lupas 28          : ------------------------------------------------------------------------------------------------------------------------

Jnet_25           : -------BBBBBBBBBBBBBBBBB-------B--B--BB--BB-B---------------B--BB--BB-BB-B-BBB-B---BB-BB--B-BBB-------B---B----BBB
Jnet_5            : ---------BBBBBBBBBBBB------------B--BB--------------------B---B--B--------------B-----------------------------------
Jnet_0            : ---------B---B--------------------B---B-------------------------B-----------------------------------------------
Jnet Rel          : 999312378999999999860367887777527899999862367777777777764289999999999851678887662789999862367777777777765101243577
```

Notes

Key:

```
Colour code for alignment:
Blue              - Complete identity at a position
Shades of red     - The more red a position is, the higher the level of
                    conservation of chemical properties of the amino acids
Jnet              - Final secondary structure prediction for query
jalign            - Jnet alignment prediction
jhmm              - Jnet hmm profile prediction
jpssm             - Jnet PSIBLAST pssm profile prediction

Lupas             - Lupas Coil prediction (window size of 14, 21 and 28)

Note on coiled coil predictions  - = less than 50% probability
                                 c = between 50% and 90% probability
                                 C = greater than 90% probability

Jnet_25           - Jnet prediction of burial, less than 25% solvent accesibility
Jnet_5            - Jnet prediction of burial, less than 5% exposure
Jnet_0            - Jnet prediction of burial, 0% exposure
Jnet Rel          - Jnet reliability of prediction accuracy, ranges from 0 to 9, bigger is better.
```

SOMPA

Pôle BioInformatique Lyonnais
Network Protein Sequence Analysis

NPS@ is the IBCP contribution to PBIL in Lyon, France

[HOME] [NPS@] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D] [SuMo] [Positions] [PBIL]

Monday, August 26th 2013: added support of BCL2DB database (see news)

SOPMA SECONDARY STRUCTURE PREDICTION METHOD

[Abstract] [NPS@ help] [Original server]

Sequence name (optional) :

Paste a protein sequence below : help

Output width : 70

SUBMIT CLEAR

Parameters

Number of conformational states : 4 (Helix, Sheet, Turn, Coil)

Similarity threshold : 8

Window width : 17

SOMPA results

https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html

SOPMA result for : UNK_374690

Abstract Geourjon, C. & Deléage, G., SOPMA: Significant improvement in protein secondary structure prediction by consensus prediction from multiple alignments., Cabios (1995) 11, 681-684

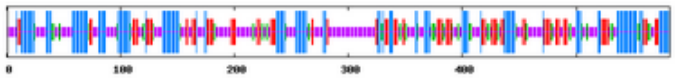
View SOPMA in: [AnTheProt (PC) , Download...] [HELP]

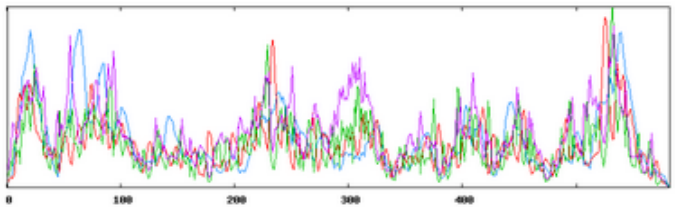
```
        10        20        30        40        50        60        70        80        90       100
         |         |         |         |         |         |         |         |         |         |
MSPAPRPPRCLLLPLLTLGTALASLGSAQSSSFSPEAWLQQYGYLPPGDLRTHTQRSPQSLSAAIAAMQKFYGLQVTGKADADTMKAMRRPRCGVPDKFG
ccccccccchheeehhhhhhhhhhhtccccccccchhhhhtttcccttccccccccccchhhhhhhhhhhhhhteeecccccchhhhhhhccccccccccthh
AEIKANVRRKRYAIQGLKWQHNEITFCIQNYTPKVGEYATYEAIRKAFRVWESATPLRFREVPYAYIREGHEKQADIMIFFAEGFHGDSTPFDGEGGFLA
hhhhhhhhhtceeeeeccccttheeeeeetcccccchhhhhhhhhhhhhhhhtccccceeccceeeehccccchheeeeeettcccccccccccceee
HAYFPGPNIGGDTHFDSAEPWTVRNEDLNGNDIFLVAVHELGHALGLEHSSDPSAIMAPFYQWMDTENFVLPDDDRRGIQQLYGGESGFPTKMPPQPRTT
eeccccccccccccctteeeeecttcettcceeeeehhhhhhhhtcccccccchhhhhhhhhhcttceeccccccccchhheeccccccccccccccccc
SRPSVPDKPKNPTYGPNICDGNFDTVAMLRGEMFVFKERWFWRVRNNQVMDGYPMPIGQFWRGLPASINTAYERKDGKFVFFKGDKHWVFDEASLEPGYP
ccccccccccccccccccccccccceehhhttheeeehhhheeeecttceettcccccchhtcccchhhhhctttteeeeeettceeeeetcccctteo
KHIKELGRGLPTDKIDAALFWMPNGKTYFFRGNKYYRFNEELRAVDSEYPKNIKVWEGIPESPRGSFMGSDEVFTYFYKGNKYWKFNNQKLKVEPGYPKS
hhhhhhttteccccchheeeecttteeeeettceeehhhhhhhhhhtccccceeettcccccccccccchheeeeeettcceeeeccceeeeettcccc
ALRDWMGCPSGGRPDEGTEEETEVIIIEVDEEGGGAVSAAAVVLPVLLLLLVLAVGLAVFFFRRHGTPRRLLYCQRSLLDKV
hhhhhtcccttcccccccchhheeeeeeectttccchhhhhhhhhhhhhhhhhhhhttthhheeeettttccceeeehhhhhhhhh
```

Sequence length : 582

SOPMA :

Alpha helix	(Hh) :	161 is	27.66%
3_{10} helix	(Gg) :	0 is	0.00%
Pi helix	(Ii) :	0 is	0.00%
Beta bridge	(Bb) :	0 is	0.00%
Extended strand	(Ee) :	112 is	19.24%
Beta turn	(Tt) :	66 is	11.34%
Bend region	(Ss) :	0 is	0.00%
Random coil	(Cc) :	243 is	41.75%
Ambigous states	(?) :	0 is	0.00%
Other states	:	0 is	0.00%





Parameters :

Window width	:	17
Similarity threshold	:	8
Number of states	:	4

Prediction result file (text): [SOPMA]
Intermediate result file (text): [BLASTP on NRPROT]

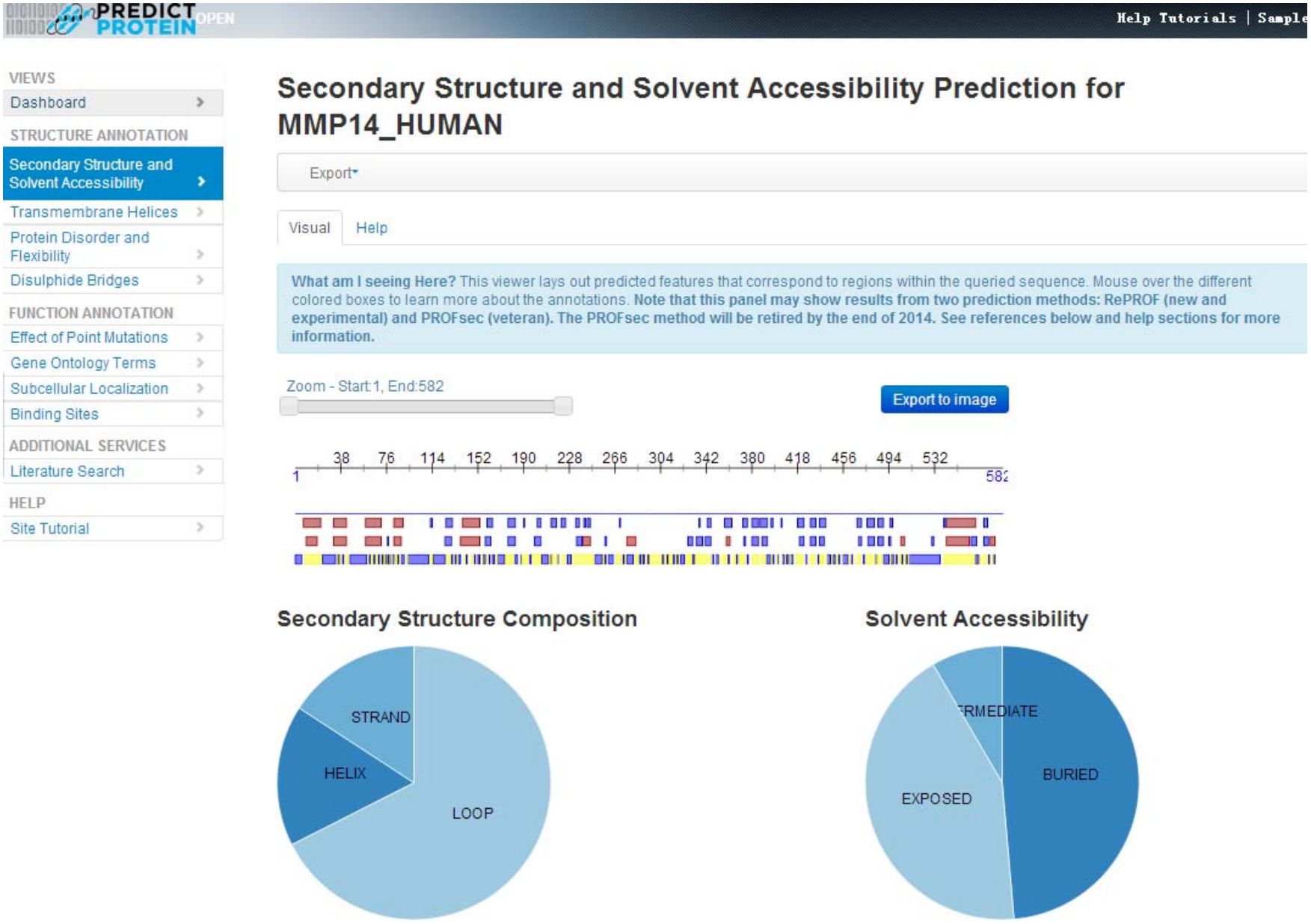
PredictProtein

https://www.predictprotein.org/

MSPAPRPPRCLLLPLLTLGTALASLGSAQSSSFSPEAWLQQYGYLPPGDLRTHTQRSPQSLSAAIAAMQK
FYGLQVTGKADADTMKAMRRPRCGVPDKFGAEIKANVRRKRYAIQGLKWQHNEITFCIQNYTPKVGEYAT
YEAIRKAFRVWESATPLRFREVPYAYIREGHEKQADIMIFFAEGFHGDSTPFDGEGGFLAHAYFPGPNIG
GDTHFDSAEPWTVRNEDLNGNDIFLVAVHELGHALGLEHSSDPSAIMAPFYQWMDTENFVLPDDDRRGIQ
QLYGGESGFPTKMPPQPRTTSRPSVPDKPKNPTYGPNICDGNFDTVAMLRGEMFVFKERWFWRVRNNQVM
DGYPMPIGQFWRGLPASINTAYERKDGKFVFFKGDKHWVFDEASLEPGYPKHIKELGRGLPTDKIDAALF
WMPNGKTYFFRGNKYYRFNEELRAVDSEYPKNIKVWEGIPESPRGSFMGSDEVFTYFYKGNKYWKFNNQK
LKVEPGYPKSALRDWMGCPSGGRPDEGTEEETEVIIIEVDEEGGGAVSAAAVVLPVLLLLLVLAVGLAVF
FFRRHGTPRRLLYCQRSLLDKV

Clear PredictProtein [Example Input 1
Example Input 2]

PredictProtein results



Course Outline

- Introduction and Case Study
- Drug Targets
  - Sequence analysis
  - Protein structure prediction
  - Molecular simulation
- Drug Design
  - Combinatorial library
  - 3D-QSAR
  - Statistical methods
- Molecular Docking