Alignment Tools

Dr. Fayyaz ul Amir Afsar Minhas
PIEAS Biomedical Informatics Research Lab
Department of Computer and Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad, Pakistan
http://faculty.pieas.edu.pk/fayyaz/

Tools

- UGENE
  - A free bioinformatics tool for analysis of biological data
- Download
  - http://ugene.unipro.ru/download.html

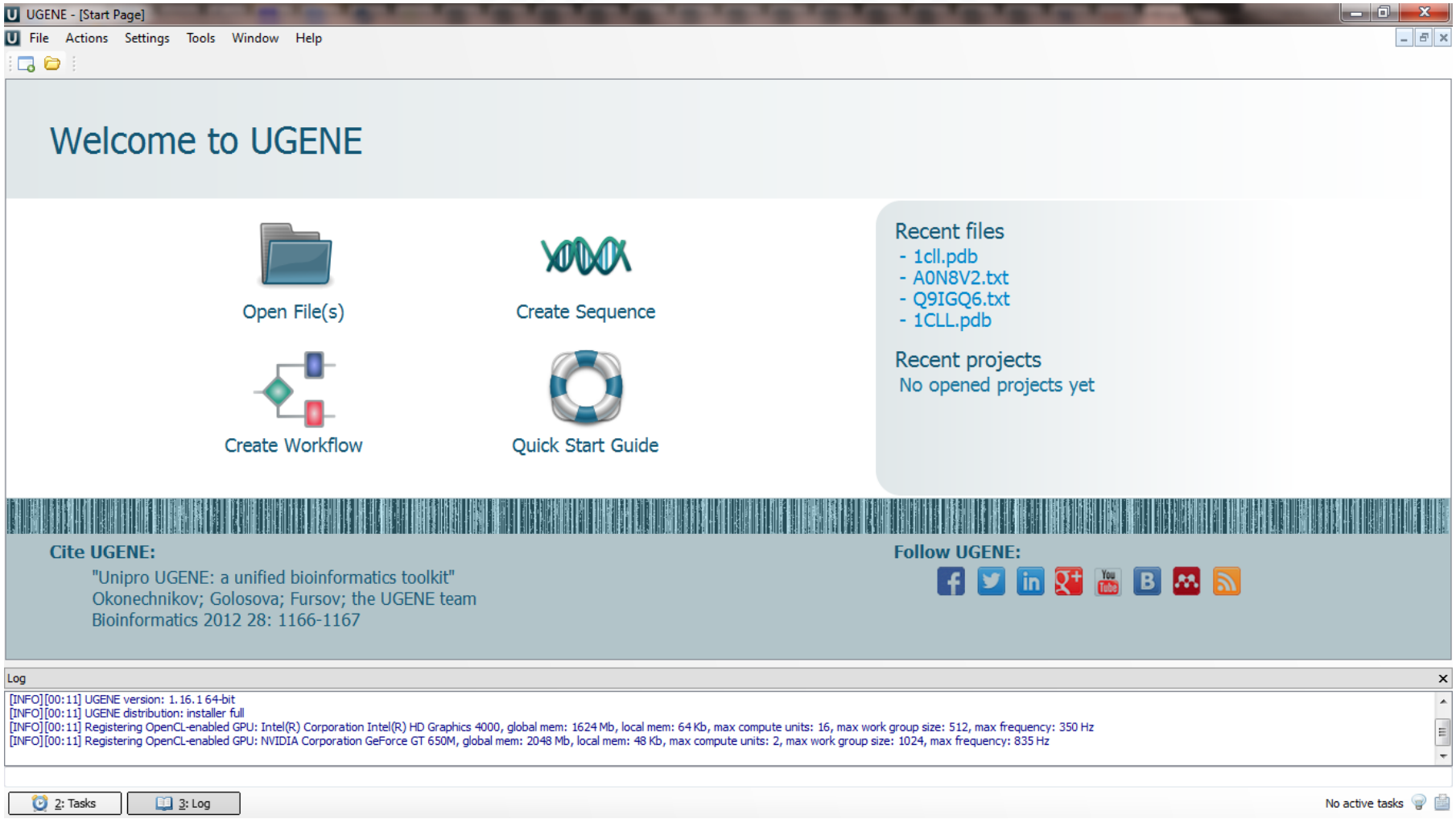
What can you do with UGENE for proteins?

- Create, edit and annotate sequences
- Searching in sequences
- MSA
- Online database search
- BLAST
- Alignments
- 3D Visualization of structures
- Secondary Structure Prediction
- Phylogenetic trees
- …

Tasks covered today

- Creating a project
- Read a protein sequence
- Searching (Exact and inexact)
- Find basic properties of the protein sequence
- Introduction to workflow design: Converting the file format
- BLASTING
- Connecting to remote database for downloading sequence
- Visualization of protein structure
- Introduction to annotations
- Using DAS for sequence annotation

The UGENE Interface

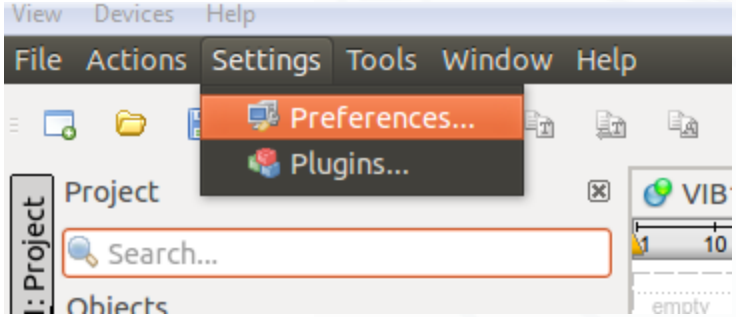


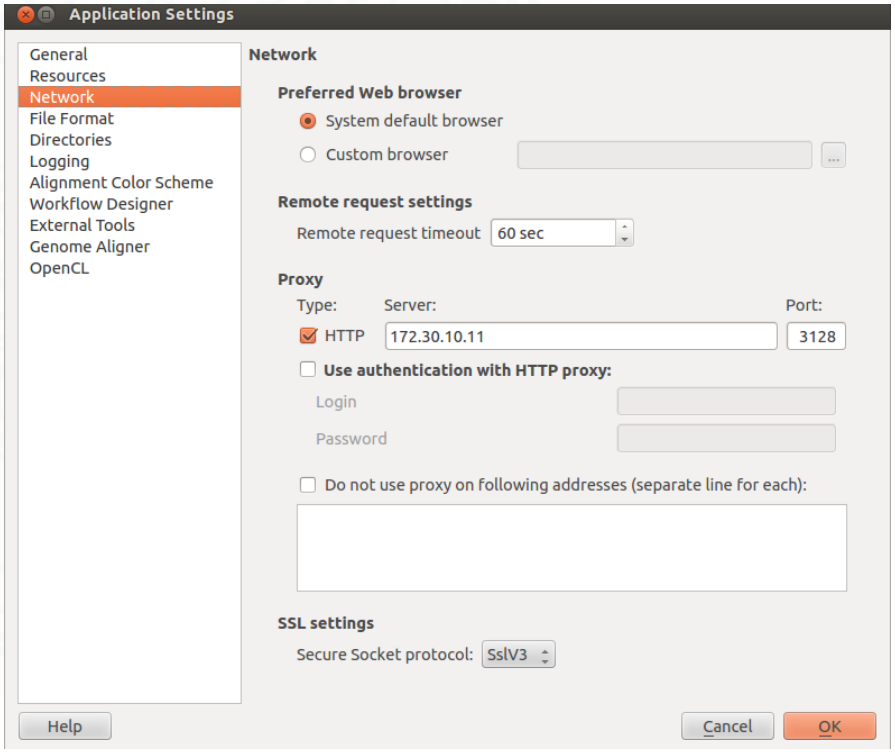
UGENE Projects

- UGENE projects allow you save your Bioinformatics files, workflows and objects
  - A project can have multiple files
    - FASTA
  - A file can have multiple objects
    - Sequences
    - Structures
  - Workflows (covered in the next lectures) manipulate the objects to produce others

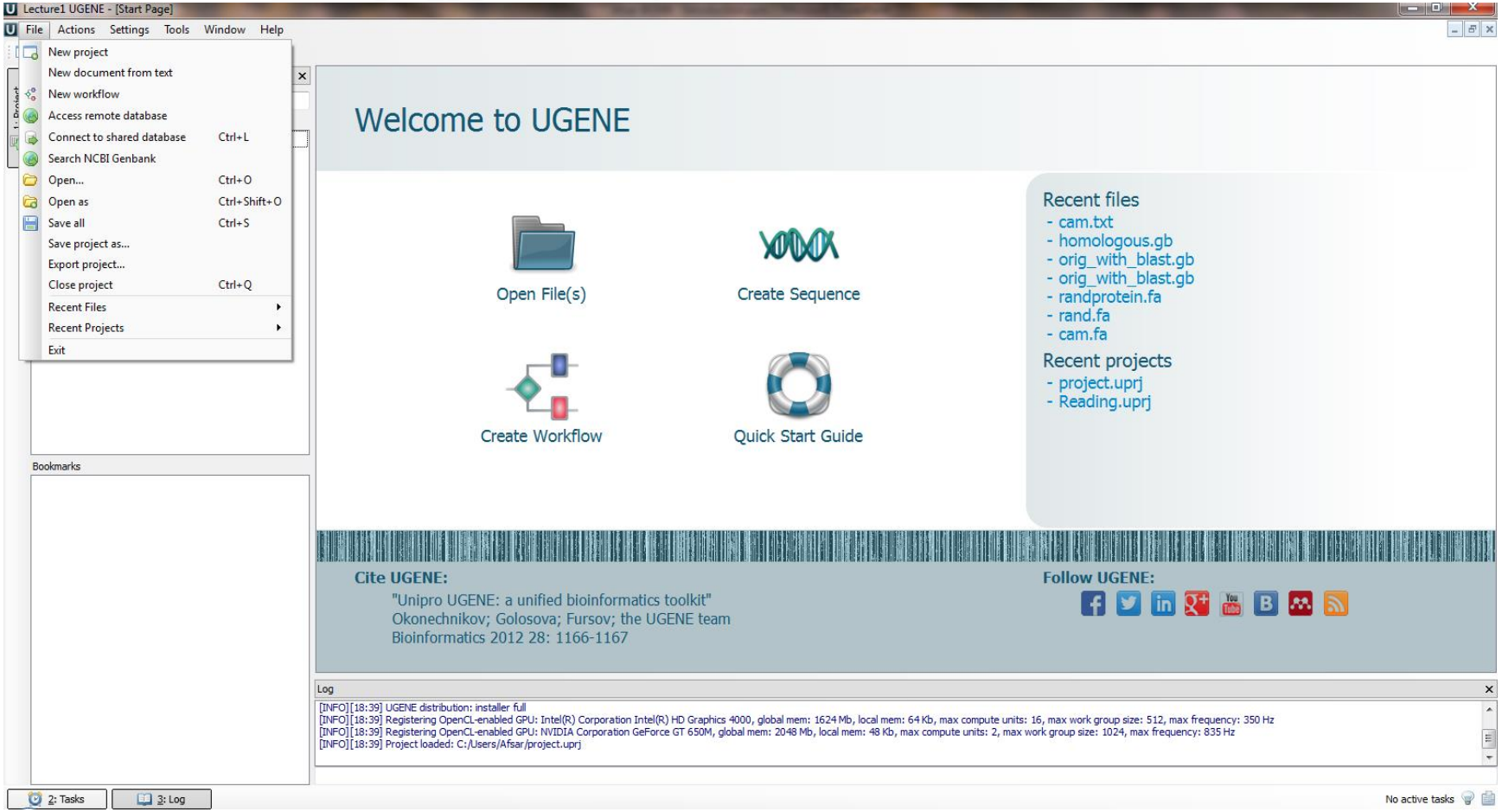
Setting up proxy

- If you are behind a proxy server you will need to specify it in UGENE for it to be able to access the Internet

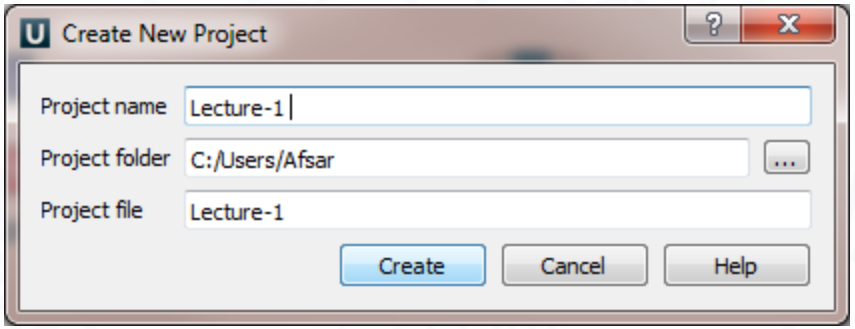




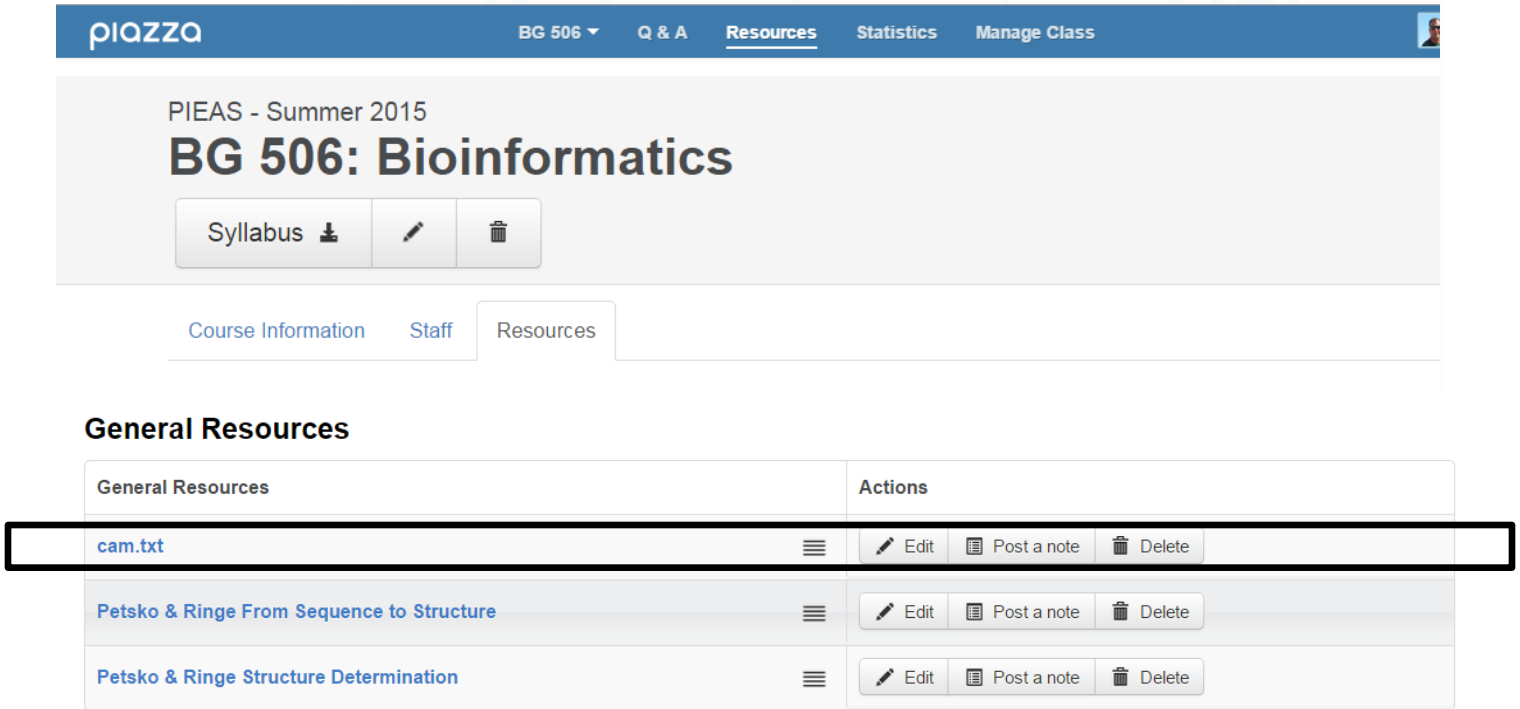
Create a new project



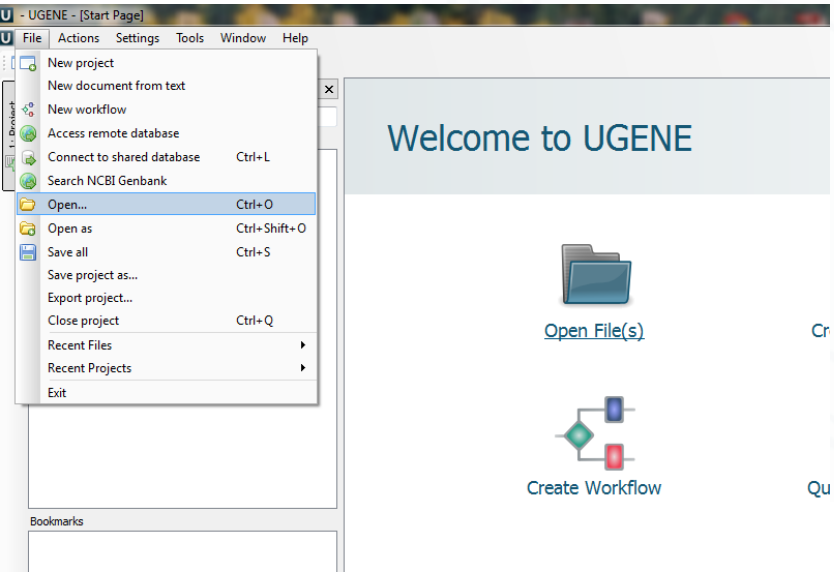
Create a new project

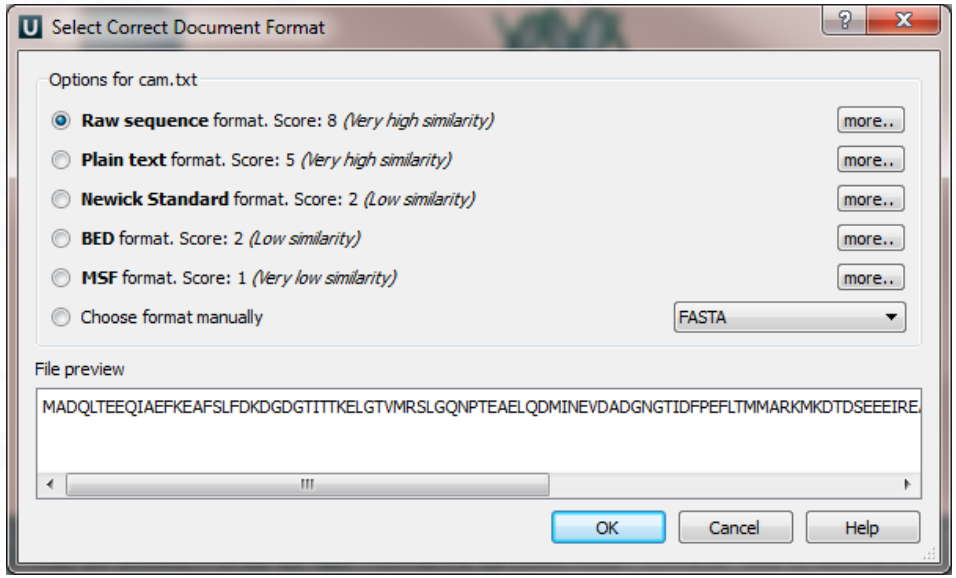


- Download the cam.txt file from Piazza and save it to your Desktop

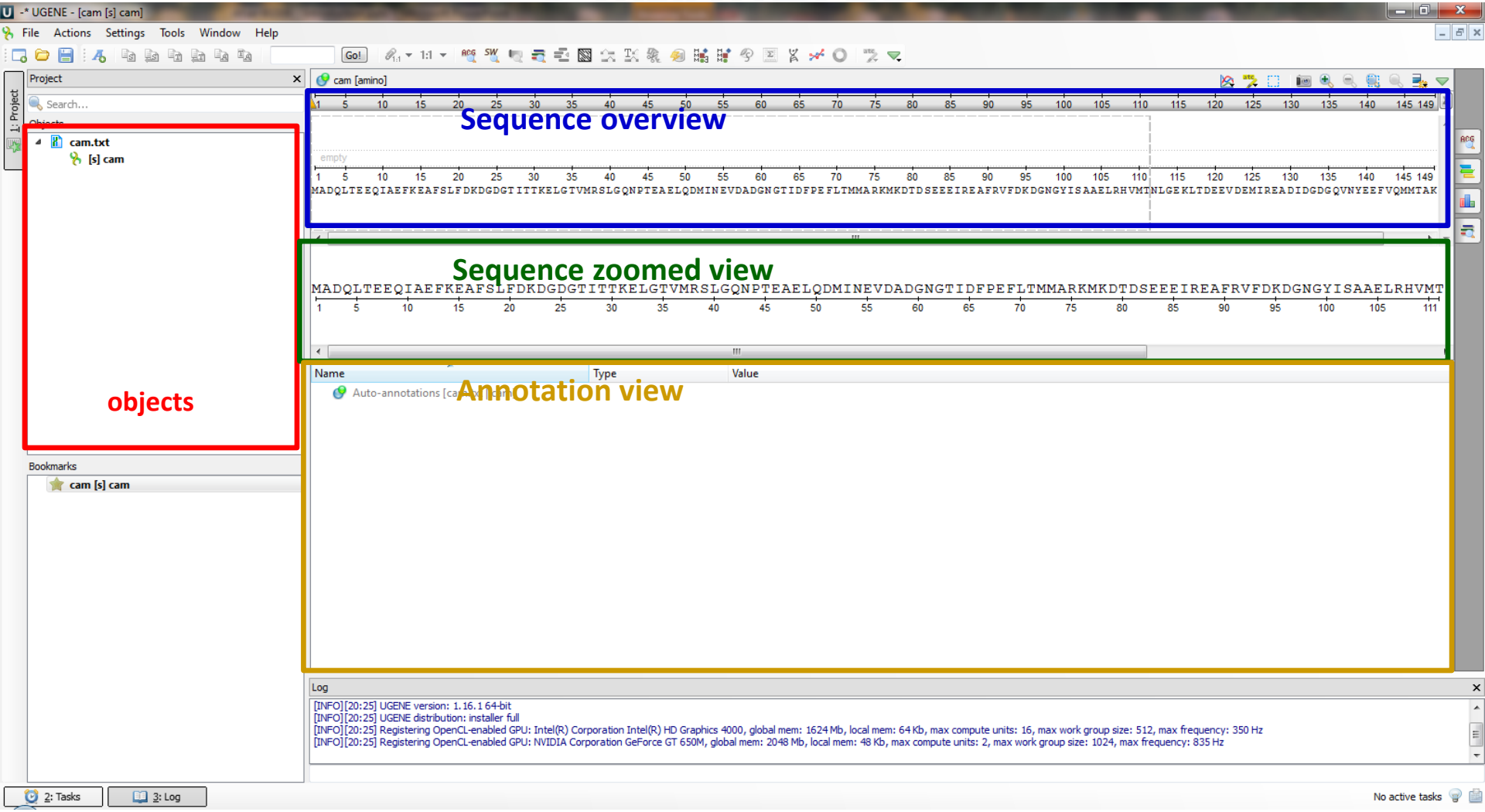


- File > Open > (File Dialog) and open ‘cam.txt’
- A “Select correct document format” dialog opens
- Use it to select “Raw Sequence” and click ok

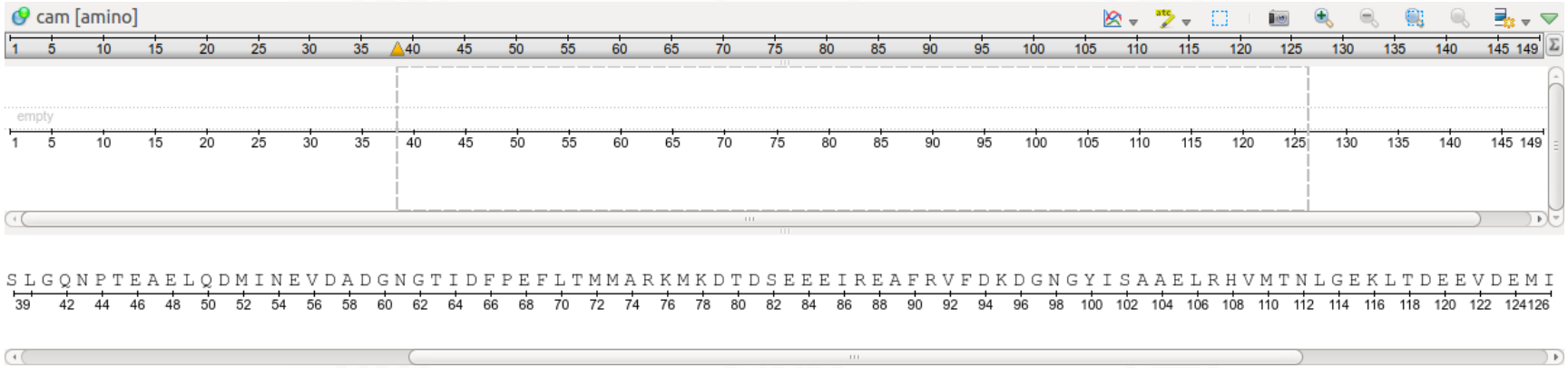


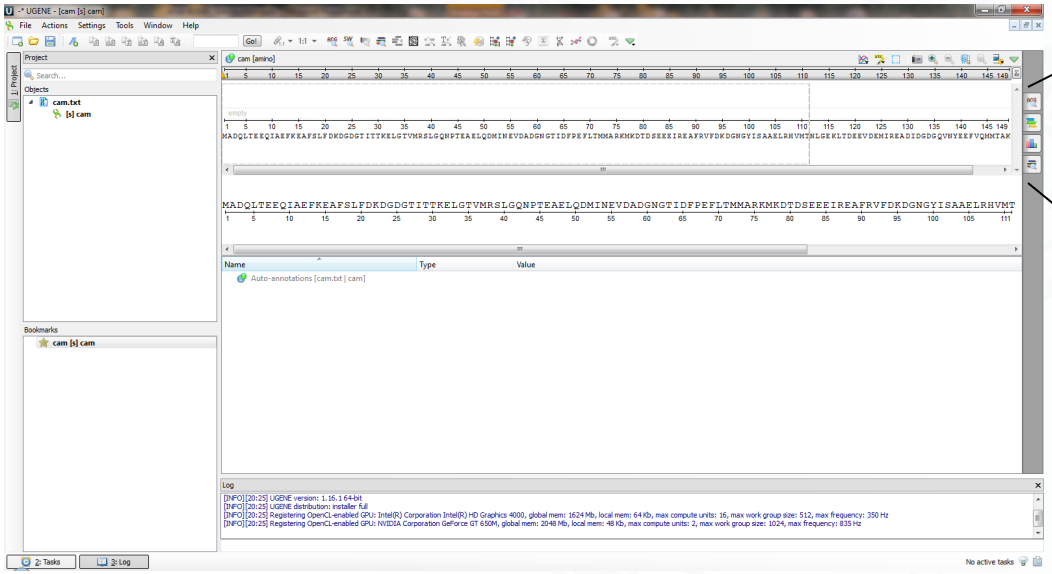


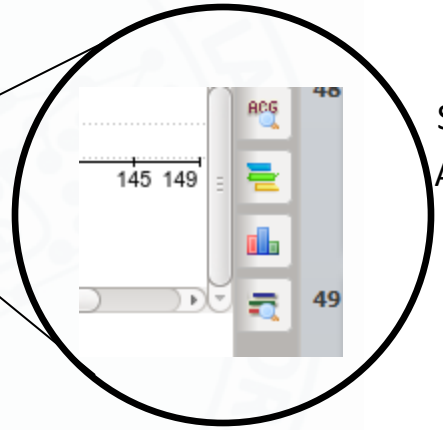
UGENE Sequence View



Looking at sequences







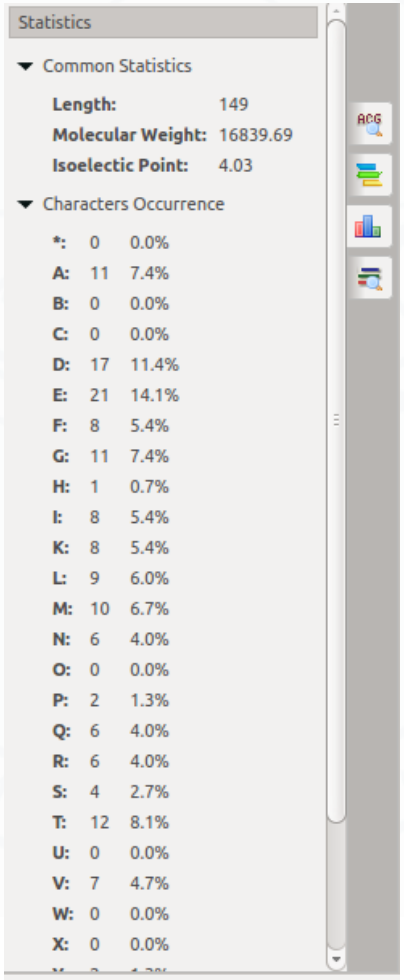
Search

Annotations highlighti

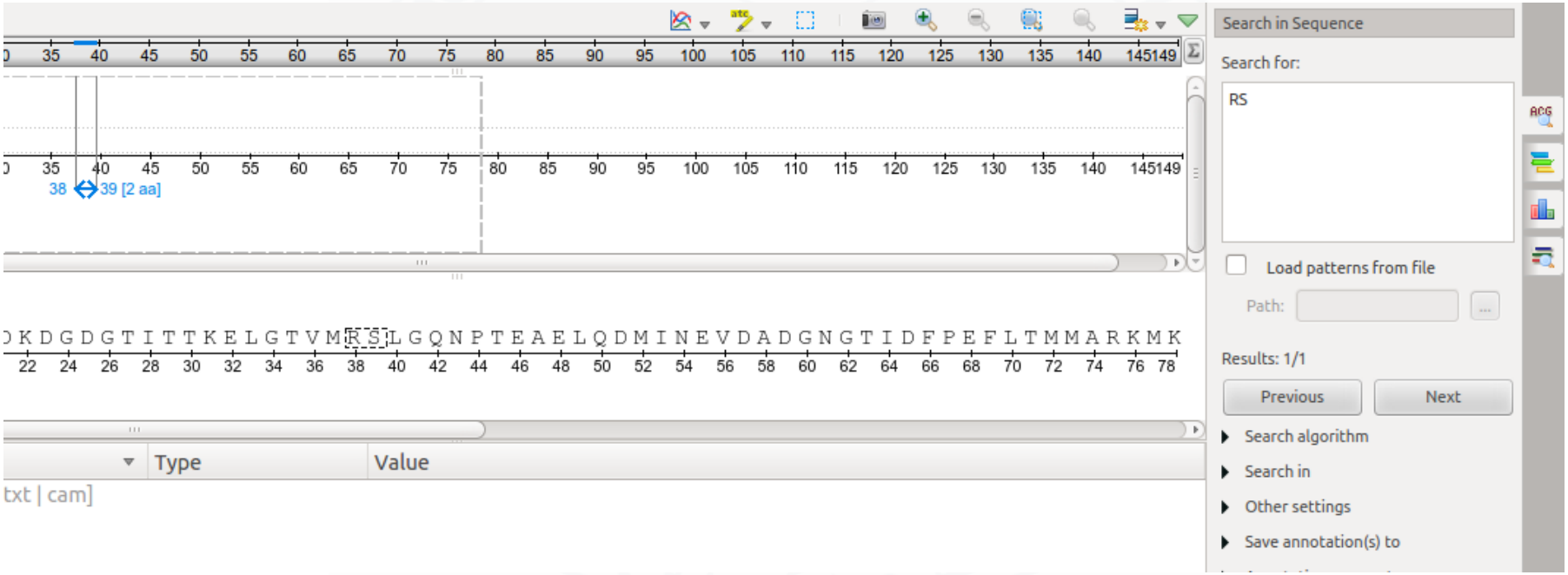
Statistics

DAS Annotations

Looking at sequence statistics



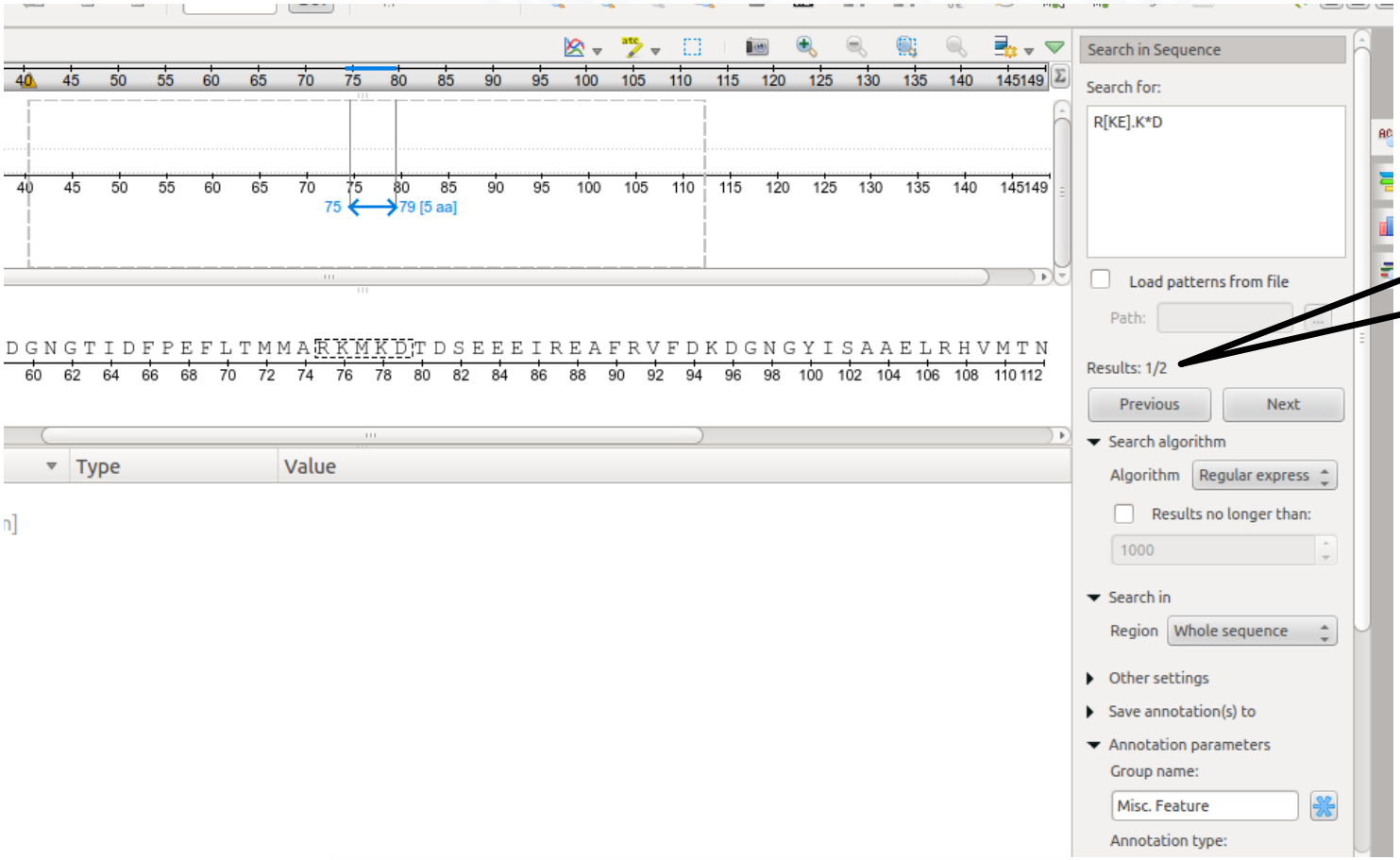
Searching in sequences



Inexact Searching

- Using Regular Expressions
- Select “Search Algorithm” > Regular Expressions
- We can find a pattern such as
  - R[KE].K*D
  - Starts with R followed by a K or an E followed by any other character (.) followed by zero or more occurences of K (*) followed by a D

Inexact Searching



There are two occurrences of this pattern in the sequence

Regular Expressions (Exercise 1-1)

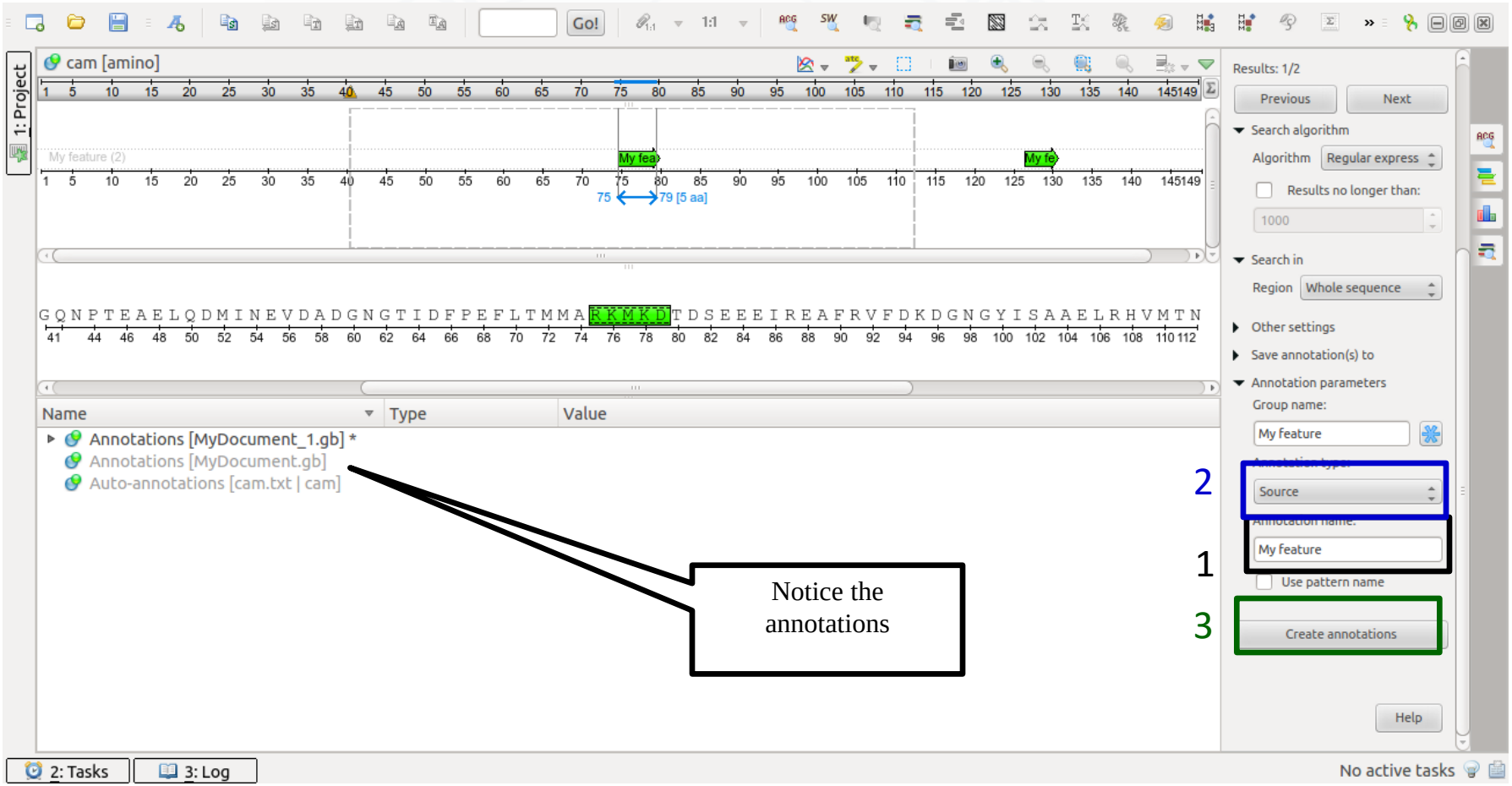
- To find any patterns you can use regular expressions
- More details on how to build regular expressions is available online
- Build a regular expression that, within the cam sequence, searches for
  - three consecutive E's followed by two other characters followed by another E
  - Two consecutive E's followed by 3 other characters followed by another E
- Once you are done, post your answers to piazza
- Please include the question or exercise number

More on searching

- You can also play with other options within the search to locate a certain pattern (exact or inexact) within the whole or a part of the sequence using “search in”
- You can also restrict the results you get in “other settings”

Creating annotations

- Such searches can be used for creating annotations
  - Annotations are saved in .gb files
  - Here in “Annotation parameters” give
    - Group name of your choice
    - Annotation type and annotation name of your choice and then click “create annotations”
  - You can then click on the annotation you have created to view where they occur in the sequence


cam [amino]
My feature (2)
My fea
My fe
75
79 [5 aa]
Name
Type
Value
Annotations [MyDocument_1.gb] *
Annotations [MyDocument.gb]
Auto-annotations [cam.txt | cam]
Notice the annotations
Results: 1/2
Previous
Next
Search algorithm
Algorithm
Regular express
Results no longer than:
1000
Search in
Region
Whole sequence
Other settings
Save annotation(s) to
Annotation parameters
Group name:
My feature
2
Source
1
My feature
Use pattern name
3
Create annotations
Help
2: Tasks
3: Log
No active tasks

Where is this used?

- Let's say you know that the binding site in a certain protein has the general pattern R..K and you want to find out where in the protein does this occur, i.e., where is the binding site
- You can also annotate it to keep a record

Exercise 1-2

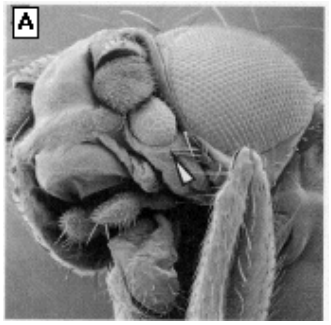
- You are given a file on piazza called cam_binders3.fasta
- Download that file
- It contains 3 sequences of proteins that bind Calmodulin
- R..K is a known Calmodulin binding motif
- Let's search for R..K within these three proteins
- And create annotations with
  - Annotation type: Misc. Binding Site
  - Annotation Name: cam binding R..K
- Once you are done report the positions of R..K within each protein on Piazza
- This way you can find the correct binding site of Calmodulin within these proteins for 2 out of the 3 given proteins
- For background reading:
  - Yap, Kyoko L., Justin Kim, Kevin Truong, Marc Sherman, Tao Yuan, and Mitsuhiko Ikura. "Calmodulin Target Database." *Journal of Structural and Functional Genomics* 1, no. 1 (March 1, 2000): 8–14. doi:10.1023/A:1011320027914.

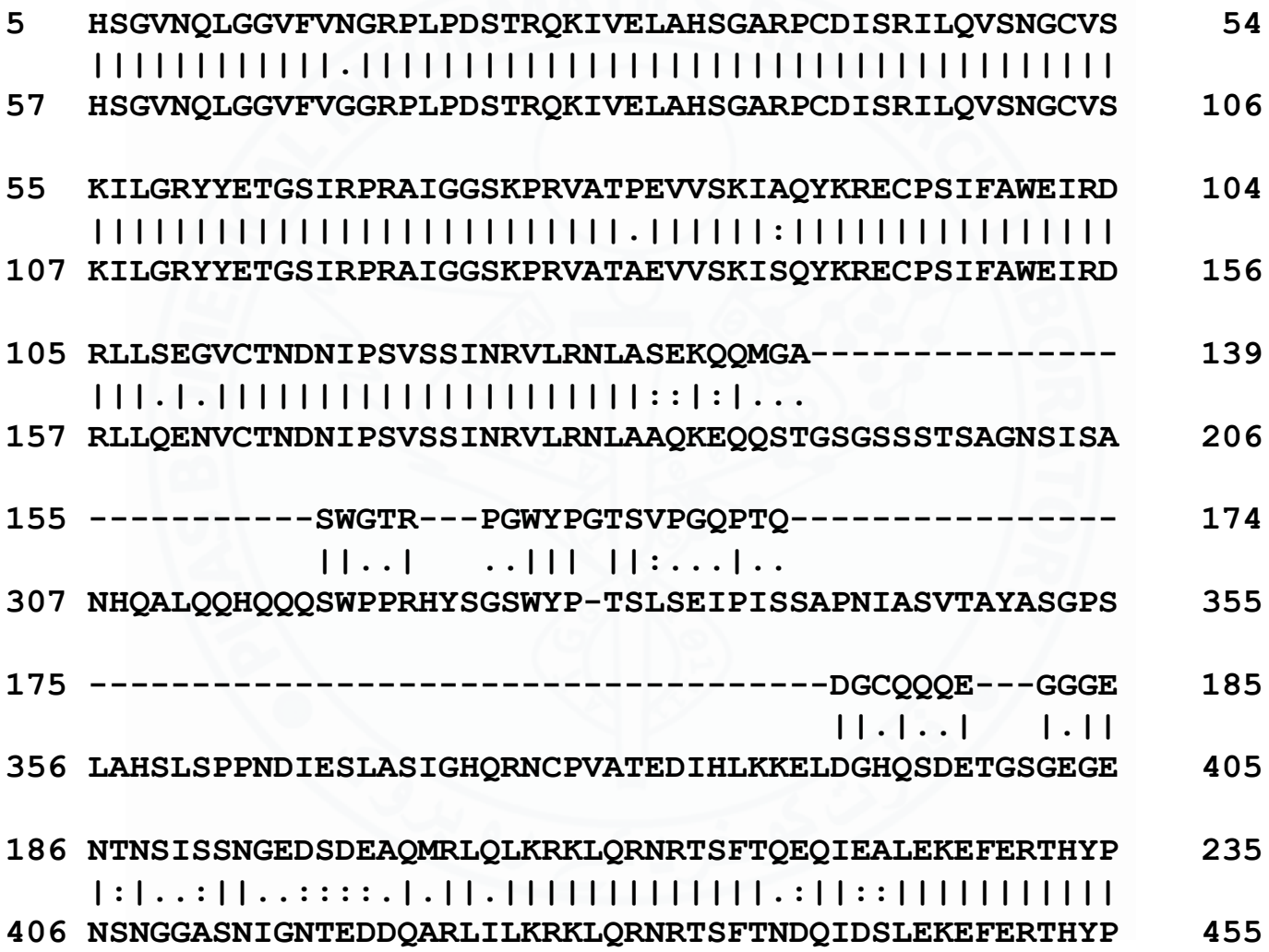
Practical alignment

- Definition
  - procedure of comparing two or more sequences by looking for a series of characteristics (residue identity, similarity, and so on) that match up in both and maximize conservation, in order to assess overall similarity.

Practical Application: Global Alignment

- Inferring the function of a human protein
- Given the human aniridia protein
  - HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIR PRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNL ASEKQQMGASWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKR KLQRNRTSFTQEQIEALEKEFERTHYP
  - Uniprot Accession: P26367
- This sequence shows good alignment with the "eyeless" protein from *D. Melanogaster*
  - HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIR PRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNL AAQKEQQSTGSGSSSTSAGNSISANHQALQQHQQQSWPPRHYSGSWYPTSLSEIPISSAPNIA SVTAYASGPSLAHSLSPPNDIESLASIGHQRNCPVATEDIHLKKELDGHQSDETGSGEGENSN GGASNIGNTEDDQARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYP
  - Uniprot Accession: O18381




5 HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVS 54
57 HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVS 106
55 KILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRD 104
107 KILGRYYETGSIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRD 156
105 RLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGA 139
157 RLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSGSSSTSAGNSISA 206
155 SWGTR---PGWYPGTSVPGQPTQ 174
307 NHQALQQHQQQSWPPRHYSGSWYP-TSLSEIPISSAPNIASVTAYASGPS 355
175 DGCQQQE---GGGE 185
356 LAHSLSPPNDIESLASIGHQRNCPVATEDIHLKKELDGHQSDETGSGEGE 405
186 NTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYP 235
406 NSNGGASNIGNTEDDQARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYP 455

Practical Application: Global Alignment

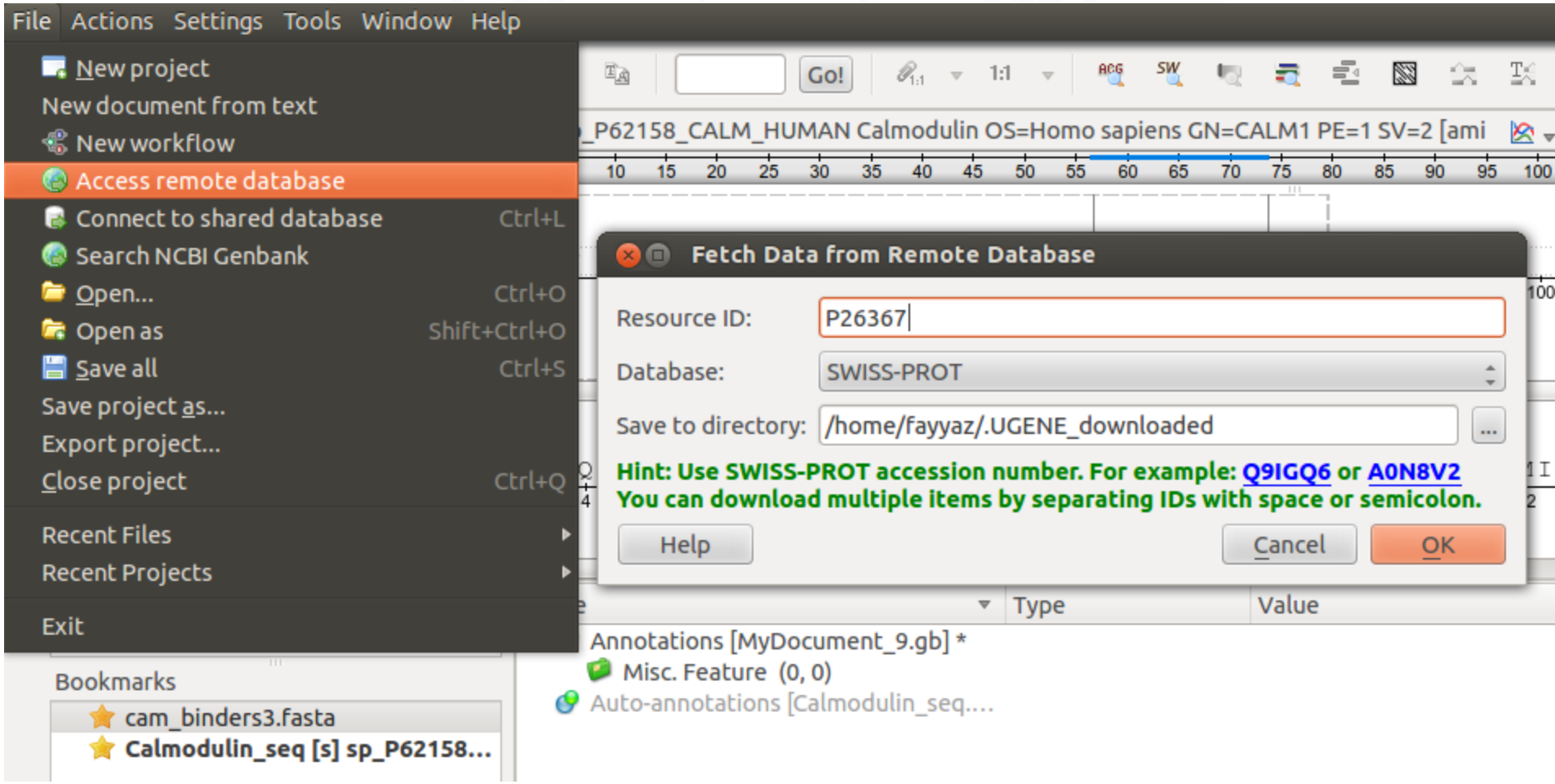
- Repeat this experiment in UGENE
- Based on this alignment Quiring et al. inferred the following

- *"The finding that ey of Drosophila, Small eye of the mouse, and human Aniridia are encoded by homologous genes suggests that eye morphogenesis is under similar genetic control in both vertebrates and insects, in spite of the large differences in eye morphology and mode of development."*

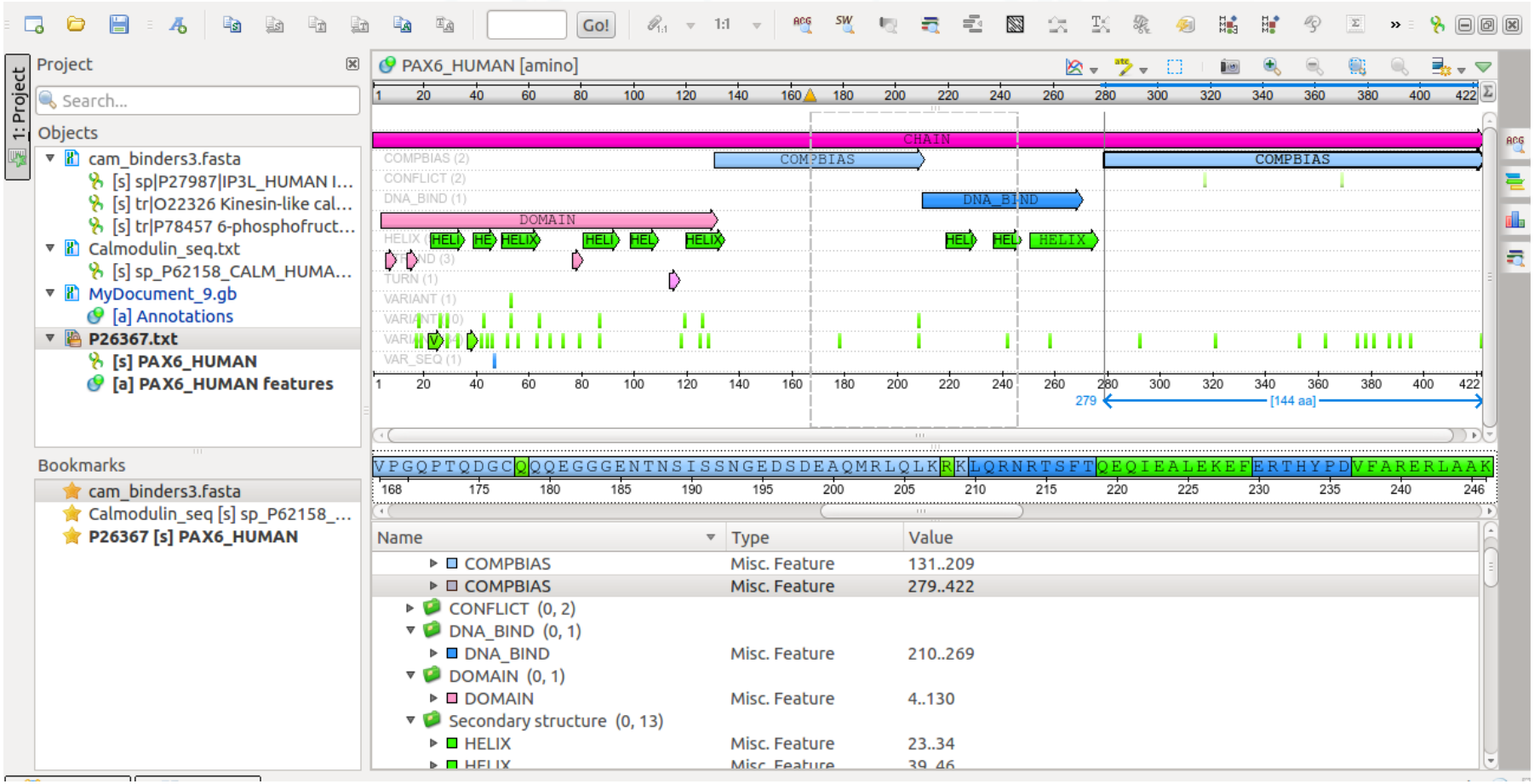
- Quiring, R., U. Walldorf, U. Kloter, and W. J. Gehring. "Homology of the Eyeless Gene of Drosophila to the Small Eye Gene in Mice and Aniridia in Humans." *Science (New York, N.Y.)* 265, no. 5173 (August 5, 1994): 785–89.

http://en.wikipedia.org/wiki/PAX6

In UGENE: Accessing Remote Databases

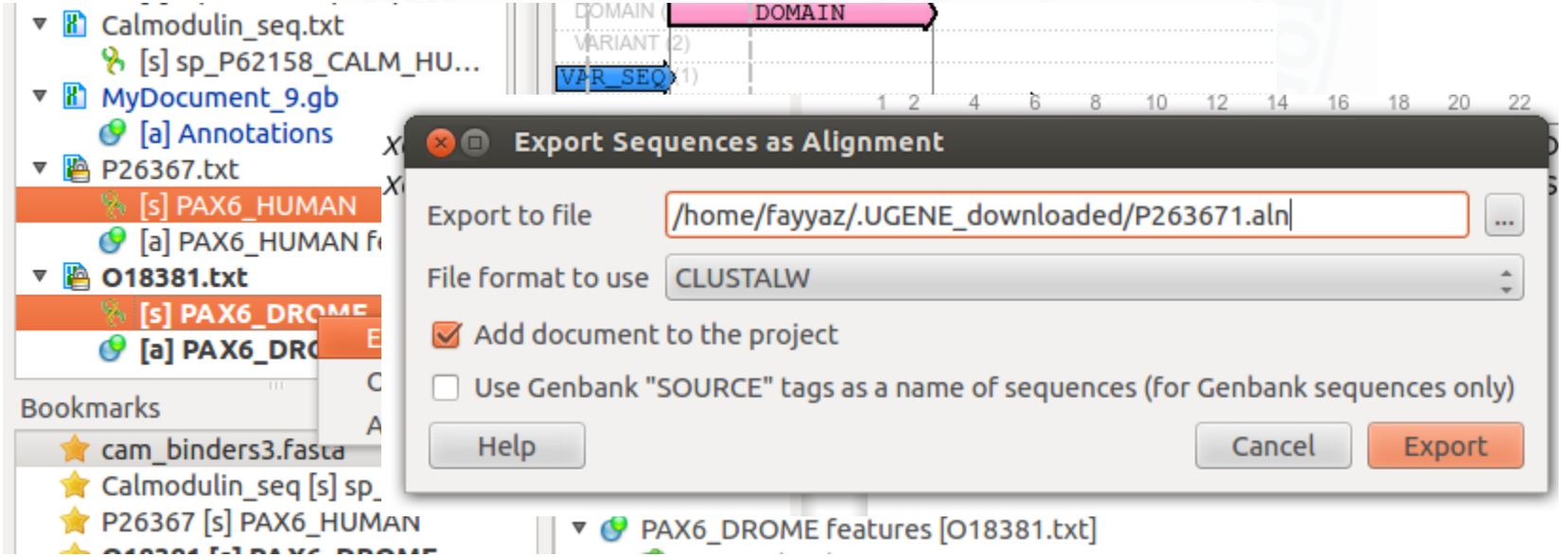


After Loading



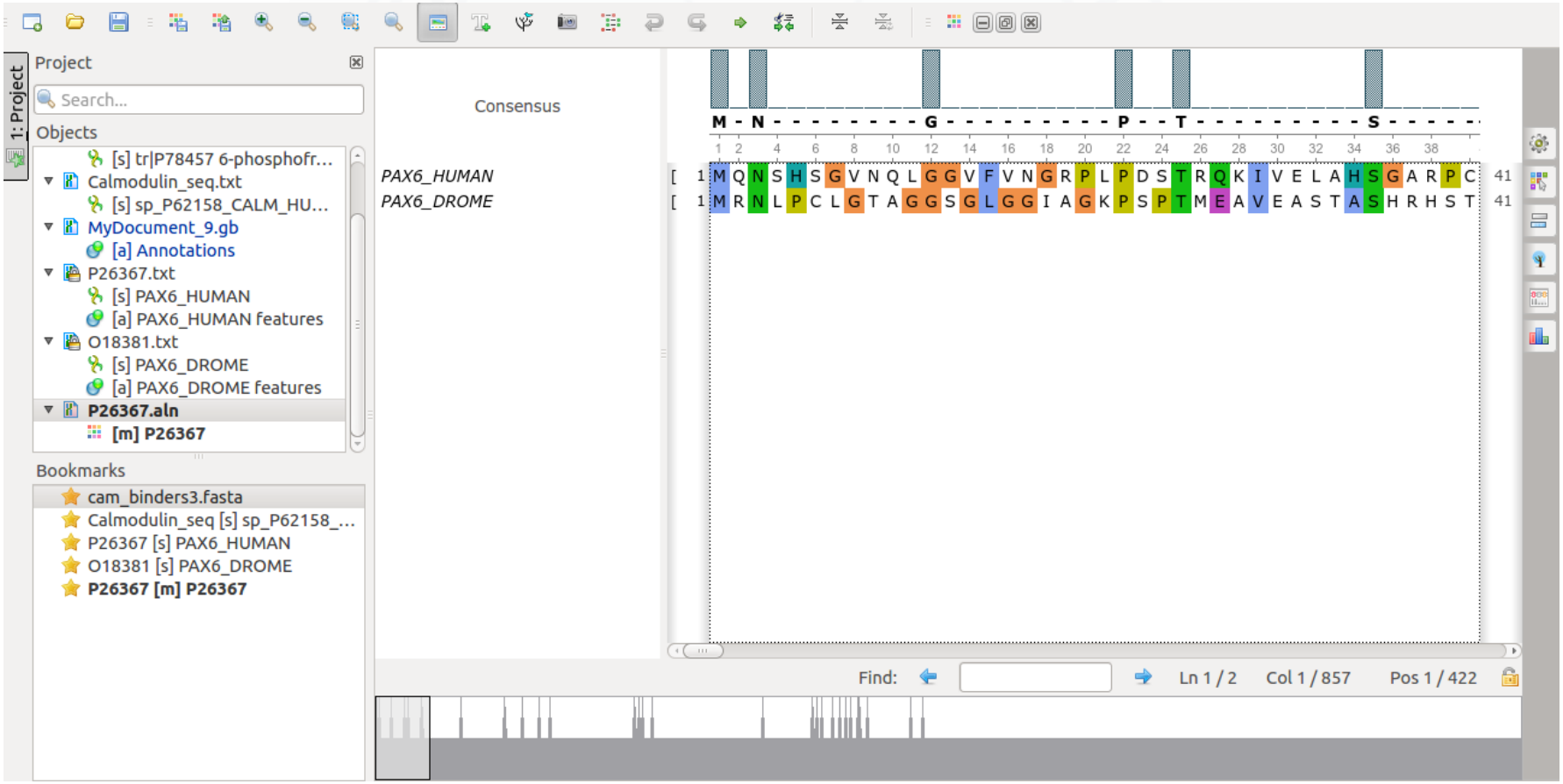
Aligning

- Similarly load O18381
- To align them select both the sequences (ctrl+left click) and export the sequences as alignment as shown below



Aligning

- This will open the alignment view however the sequences aren't aligned yet



Aligning

- To do the alignment follow these steps

Exercise: Simian Sarcoma Virus & Cancer

- R. F. Doolittle, M. W. Hunkapiller, L. E. Hood, S. G. Devare, K. C. Robbins, S. A. Aaronson, and H. N. Antoniades. Simian sarcoma virus oncogene, ν -sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. Science, 221:275–277, 1983.
- Human Platelet Derived Growth Factor: http://www.uniprot.org/uniprot/A0A024R1R5.fasta
- Woolly Monkey Sarcoma Virus V-SIS
- http://www.uniprot.org/uniprot/Q7LZ02.fasta
- Download sequences and obtain the sequence alignment using Online Needleman Wunsch: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=blastn&BLAST_SPEC=GlobalAln&LINK_LOC=BlastHomeLink
- Report the sequence identity

FASTA, BLAST, BLAT & HHBlits

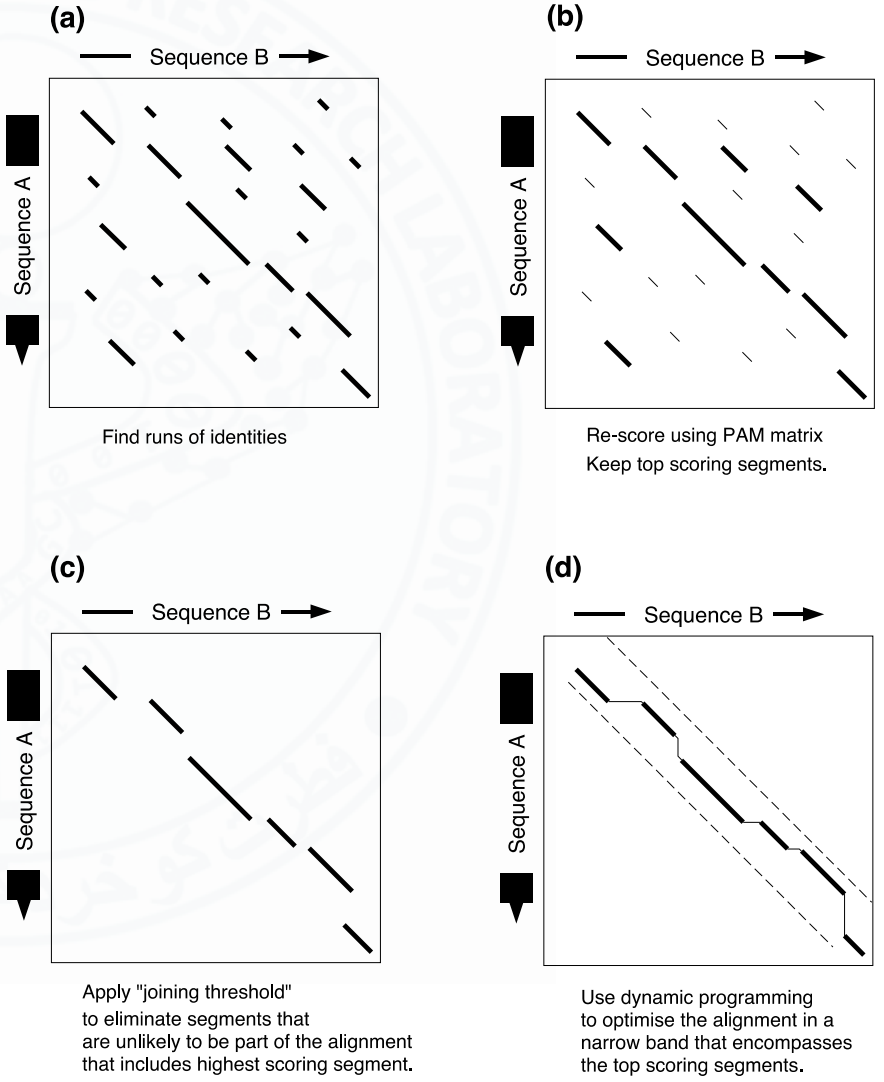
- Are heuristic tools for alignemnts

Motivation

- Smith-Waterman algorithm too slow for searching large sequence databases
- Most sequences are not homologous to the query so there is no need for the alignment
- Use heuristic methods:
  - FASTA
  - BLAST

FASTA

- Idea: in order for two sequences to be similar, need a run of identical letters
  - This is called seeding
- Only sequences that have high scoring segments need to be aligned

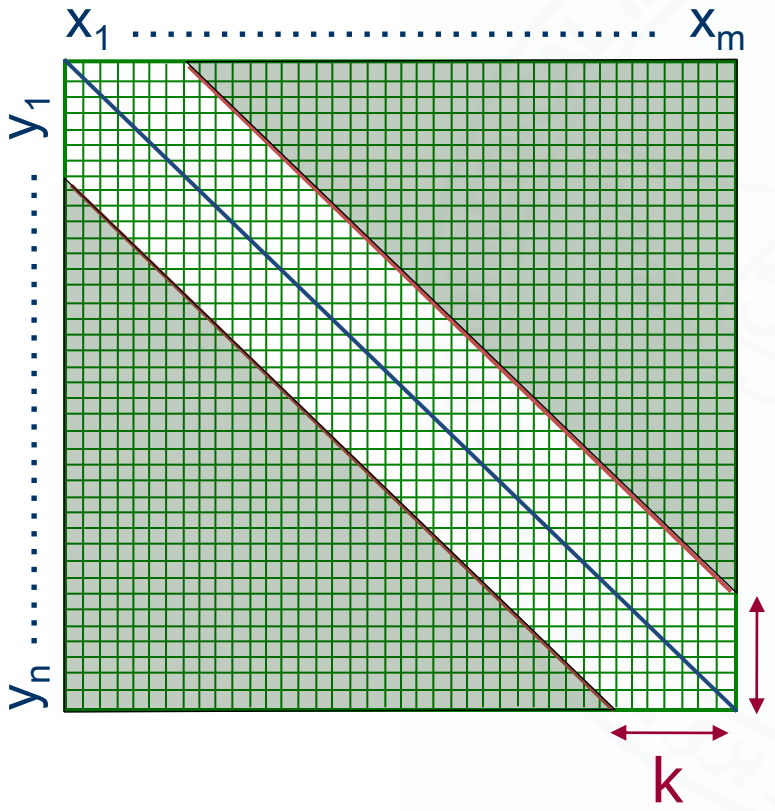


Lipman, D.J. and Pearson, W.R., Rapid and sensitive protein similarity searches, Science, 227:1435--1441, 1985.

Finding all matching k-mers

- Store the positions of all k-mers in the query sequence
- Scan for matches with those k-mers
- k=2 for protein sequences.
- Need an efficient method for retrieving k-mer positions: hash table!

Bounded Dynamic Programming



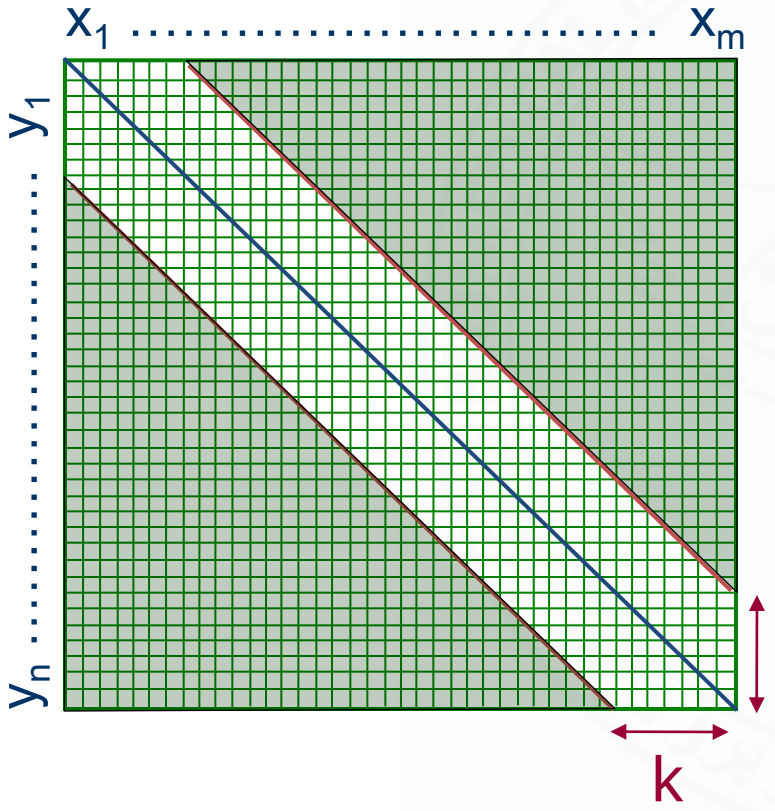
Assume that x and y are very similar

Assumption: #gaps(x, y) < k

We can align x and y more efficiently:

Time, Space: O(n × k)

Bounded Dynamic Programming



Initialization:

S(i,0), S(0,j) undefined for i, j > k

Iteration:

For i = 1,…,n

For j = max(1, i – k),…,min(n, i+k)

$$S(i, j) = \max \begin{cases} S(i-1, j-1) + s(x_i, y_j) \\ S(i, j-1) - d, \text{ if } j > i - k \\ S(i-1, j) - d, \text{ if } j < i + k \end{cases}$$

Can extend to the affine gap case

BLAST: Basic Local Alignment and Search Tool[1]

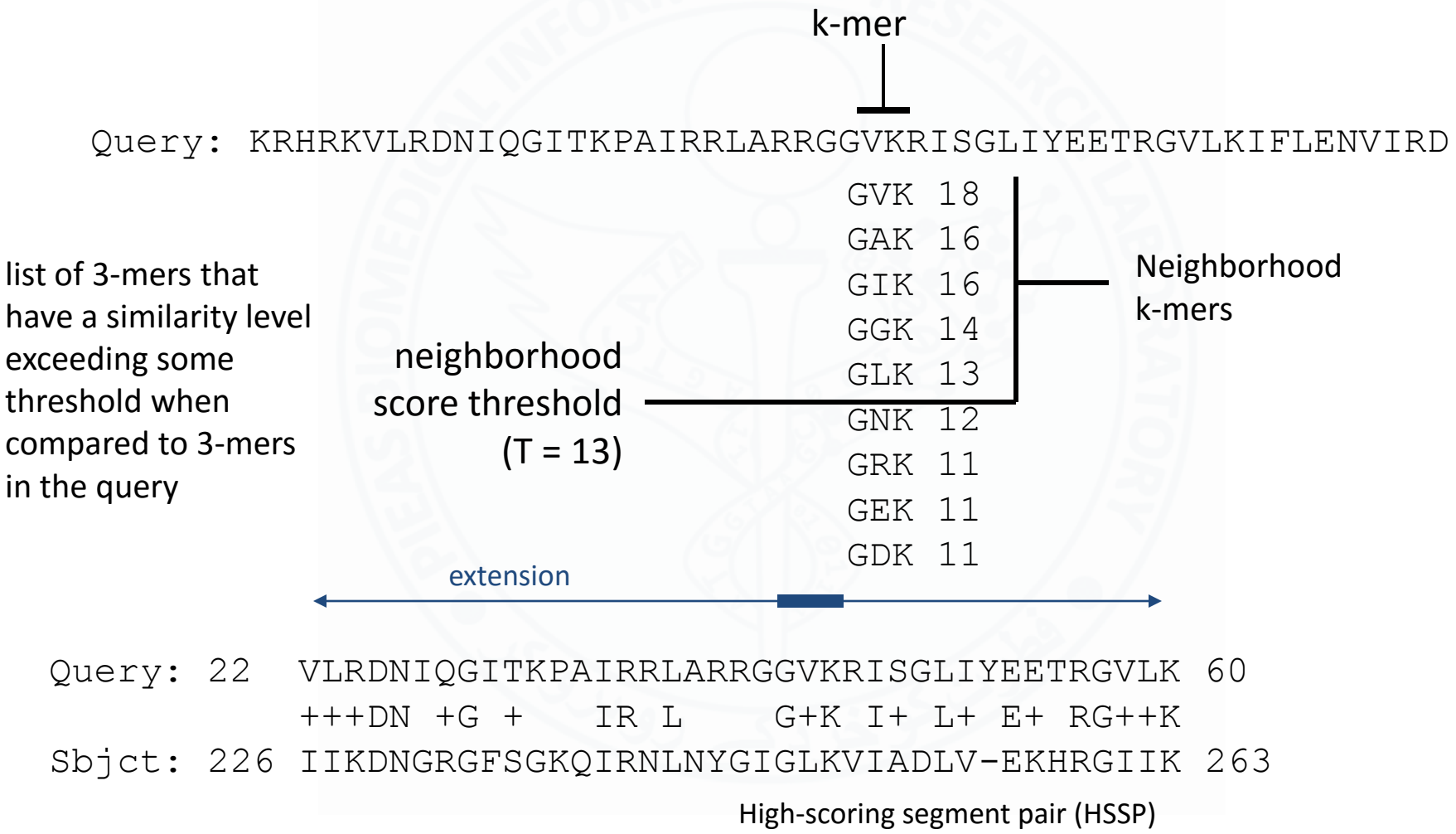
- Similar idea to FASTA:
  - Does not compute alignments that do not look promising
  - Fast alignment: does not consider complete edit graph
- Difference from FASTA: seed matches do not need to match exactly
- Great improvement in speed, with a modest decrease in sensitivity with respect to SW.

[1]Altschul, SF, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. J Mol Biol 215(3):403-10, 1990.

Outline of BLAST

- Eliminate low complexity regions
- Create an “index” of 3-mers from the query sequence
- Search for hits in the database
- Extend hits into HSSPs: High Scoring Segment Pairs (ungapped)
- Extend HSSPs into a local alignment
- Compute “E-values”

BLAST algorithm (cont'd)




```
Query: 22  VLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLK 60
           +++DN +G +   IR L    G+K I+ L+ E+ RG++K
Sbjct: 226 IIKDNGRGFSGKQIRNLNYGIGLKVIADLV-EKHRGIIK 263
```

High-scoring segment pair (HSSP)

Indexing the query

- Compile a list of high-scoring 3-mers (have a similarity level exceeding some threshold when compared to 3-mers in the query)
- Typical size of list: 50 times the length of the sequence

Original BLAST

- **Extending a hit**
  - *Ungapped* extensions until score (sum of subst. matrix elements) falls below some threshold
- **Output**
  - All local ungapped alignments with score > threshold

Original ungapped BLAST

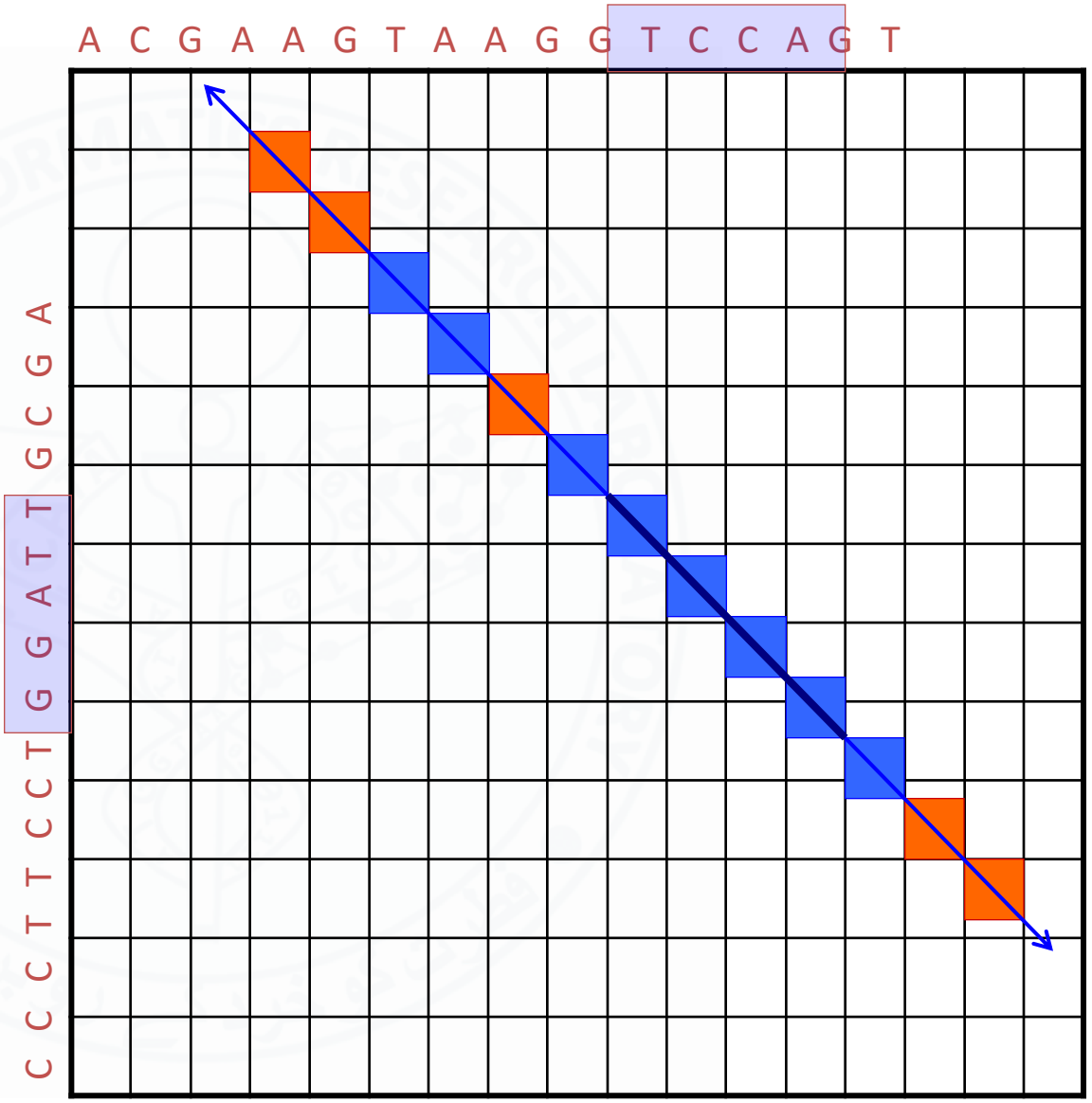
The seed **GGTC** initiates an alignment

Extension with no gaps

Output:

GTAAGGTCC

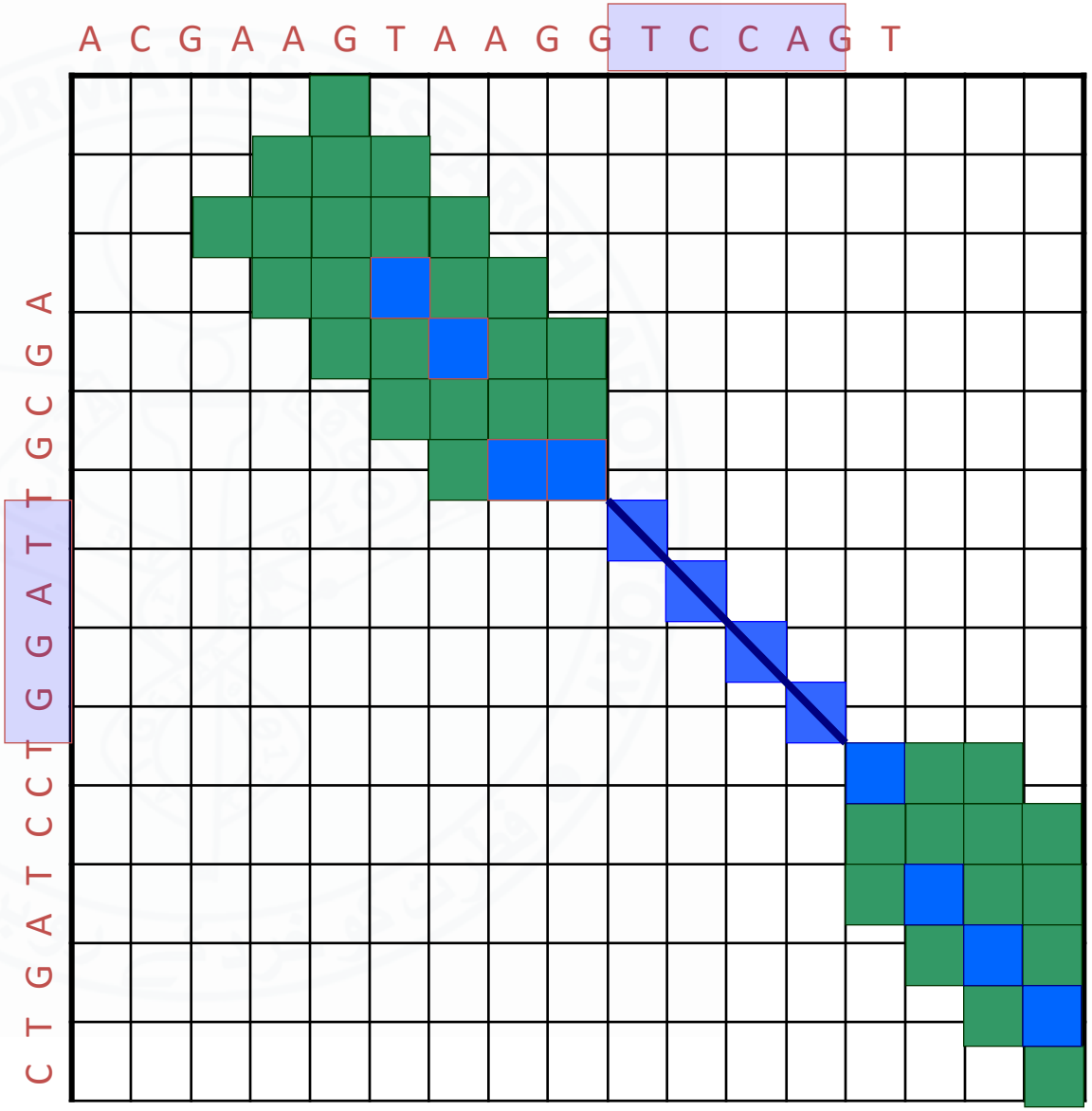
GTTAGGTCC



Gapped BLAST

- Search for seed. THEN:
- Extend with gaps in a band around seed (anchor) until score < *threshold*
- Result:

GTAAGGTCCAG
T
GTTAGGTC-AGT



BLAST: Locally Maximal Segment Pairs

- A segment pair is locally maximal if its score can't be improved by extending or shortening
- BLAST finds all locally maximal segment pairs with scores above some threshold

- Output: Statistically significant locally maximal segment pairs (more likely to be of biological interest)

BLAST statistics

- We want to assign probabilities to BLAST scores.

  $p = P(\text{two random bases are equal})$
  - The event of a mismatch followed by t matches has probability $(1 - p)p^t$
  - There are nm places to begin the event.
  - The expected # of such events: $mn\,(1 - p)p^t$
  - Rare event: use a Poisson distribution with mean $mn\,(1 - p)p^t$

  $P(\text{alignment of length t or longer}) = 1 - P(\text{no such event})$

  $= 1 - \exp(-mn\,(1 - p)p^t)$

  This is known as the *extreme value distribution*

$$f_\lambda(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

BLAST statistics

- # matches with score greater than θ has mean $E(\theta) = Kmne^{-\lambda\theta}$; K is a constant, m,n are the lengths of the two compared sequences
  - Parameter λ is positive root of:

    $\Sigma_{x,y}\ p_x p_y\, e^{\delta(x,y)} = 1,$

    where p_x, p_y are aa frequencies

Sample BLAST output

- Blast of human beta globin protein against zebra fish

```
                                                                    Score
Sequences producing significant alignments:                          (bits)   E

gi|18858329|ref|NP_571095.1| ba1 globin [Danio rerio] >gi|147757...   171   3e-44
gi|18858331|ref|NP_571096.1| ba2 globin; SI:dZ118J2.3 [Danio rer...   170   7e-44
gi|37606100|emb|CAE48992.1| SI:bY187G17.6 (novel beta globin) [D...   170   7e-44
gi|31419195|gb|AAH53176.1| Ba1 protein [Danio rerio]                  168   3e-43

ALIGNMENTS
>gi|18858329|ref|NP_571095.1| ba1 globin [Danio rerio]
Length = 148

 Score =  171 bits (434), Expect = 3e-44
 Identities = 76/148 (51%), Positives = 106/148 (71%), Gaps = 1/148 (0%)

Query: 1   MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60
           MV  T  E++A+  LWGK+N+DE+G +AL R L+VYPWTQR+F +FG+LS+P A+MGNPK
Sbjct: 1   MVEWTDAERTAILGLWGKLNIDEIGPQALSRCLIVYPWTQRYFATFGNLSSPAAIMGNPK 60

Query: 61  VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 120
           V AHG+ V+G     + ++DN+K T+A LS +H +KLHVDP+NFRLL + +    A  FG
Sbjct: 61  VAAHGRTVMGGLERAIKNMDNVKNTYAALSVMHSEKLHVDPDNFRLLADCITVCAAMKFG 120

Query: 121 KE-FTPPVQAAYQKVVAGVANALAHKYH 147
           +  F   VQ A+QK +A V +AL  +YH
Sbjct: 121 QAGFNADVQEAWQKFLAVVVSALCRQYH 148
```

Sample BLAST output

- Blast of human beta globin DNA against human DNA

```
                                                                    Score     E
Sequences producing significant alignments:                        (bits) Value

gi|19849266|gb|AF487523.1| Homo sapiens gamma A hemoglobin (HBG1...   289   1e-75
gi|183868|gb|M11427.1|HUMHBG3E Human gamma-globin mRNA, 3' end        289   1e-75
gi|44887617|gb|AY534688.1| Homo sapiens A-gamma globin (HBG1) ge...   280   1e-72
gi|31726|emb|V00512.1|HSGGL1 Human messenger RNA for gamma-globin     260   1e-66
gi|38683401|ref|NR_001589.1| Homo sapiens hemoglobin, beta pseud...   151   7e-34
gi|18462073|gb|AF339400.1| Homo sapiens haplotype PB26 beta-glob...   149   3e-33

ALIGNMENTS
>gi|28380636|ref|NG_000007.3| Homo sapiens beta globin region (HBB@) on chromosome 11
          Length = 81706
 Score =  149 bits (75), Expect = 3e-33
 Identities = 183/219 (83%)
 Strand = Plus / Plus

Query: 267   ttgggagatgccacaaagcacctggatgatctcaagggcacctttgcccagctgagtgaa 326
             || ||| | ||    | || |  |||||| ||||| ||||||||||    ||||||||
Sbjct: 54409 ttcggaaaagctgttatgctcacggatgacctcaaaggcacctttgctacactgagtgac 54468


Query: 327   ctgcactgtgacaagctgcatgtggatcctgagaacttc 365
             ||||||||| |||||||||| ||||| ||||||||||||
Sbjct: 54469 ctgcactgtaacaagctgcacgtggaccctgagaacttc 54507
```

Flavors of BLAST

- blastn: Nucleotide-nucleotide
- blastp: Protein-protein
- blastx: Translated query vs. protein database (all six frames)
- tblastn: Protein query vs. translated database (finding homologous protein coding regions in unannotated nucleotide sequences – ESTs and draft genomes)
- tblastx: Translated query vs. translated database (6 frames each - expensive)

Versions of BLAST

- PSI-BLAST
  - Find members of a protein family or build a custom position-specific score matrix
- Megablast:
  - Search longer sequences with fewer differences
  - When the number of query sequences is large

Timeline

- 1970: Needleman-Wunsch global alignment algorithm
- 1981: Smith-Waterman local alignment algorithm
- 1985: FASTA
- 1990: BLAST (basic local alignment search tool)
- 2000s: BLAST has become too slow in - faster algorithms are developed
  - Pattern Hunter
  - BLAT
- 2009: BLAST+: a complete re-write of BLAST
- 2010: Next generation sequencing: need super fast algorithms!

BLAT: BLAST-Like Alignment Tool

- BLAT builds an index of the database and scans through the query sequence, whereas BLAST builds an index of the query and then scans through the database
- What is the advantage of this approach?
- Potential disadvantage?

Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002.

BLAT: BLAST-Like Alignment Tool

- Builds an index of all 11-mers that are not heavily involved in repeat regions (what are the tradeoffs in the choice of k-mer size?)

BLAT: BLAST-Like Alignment Tool

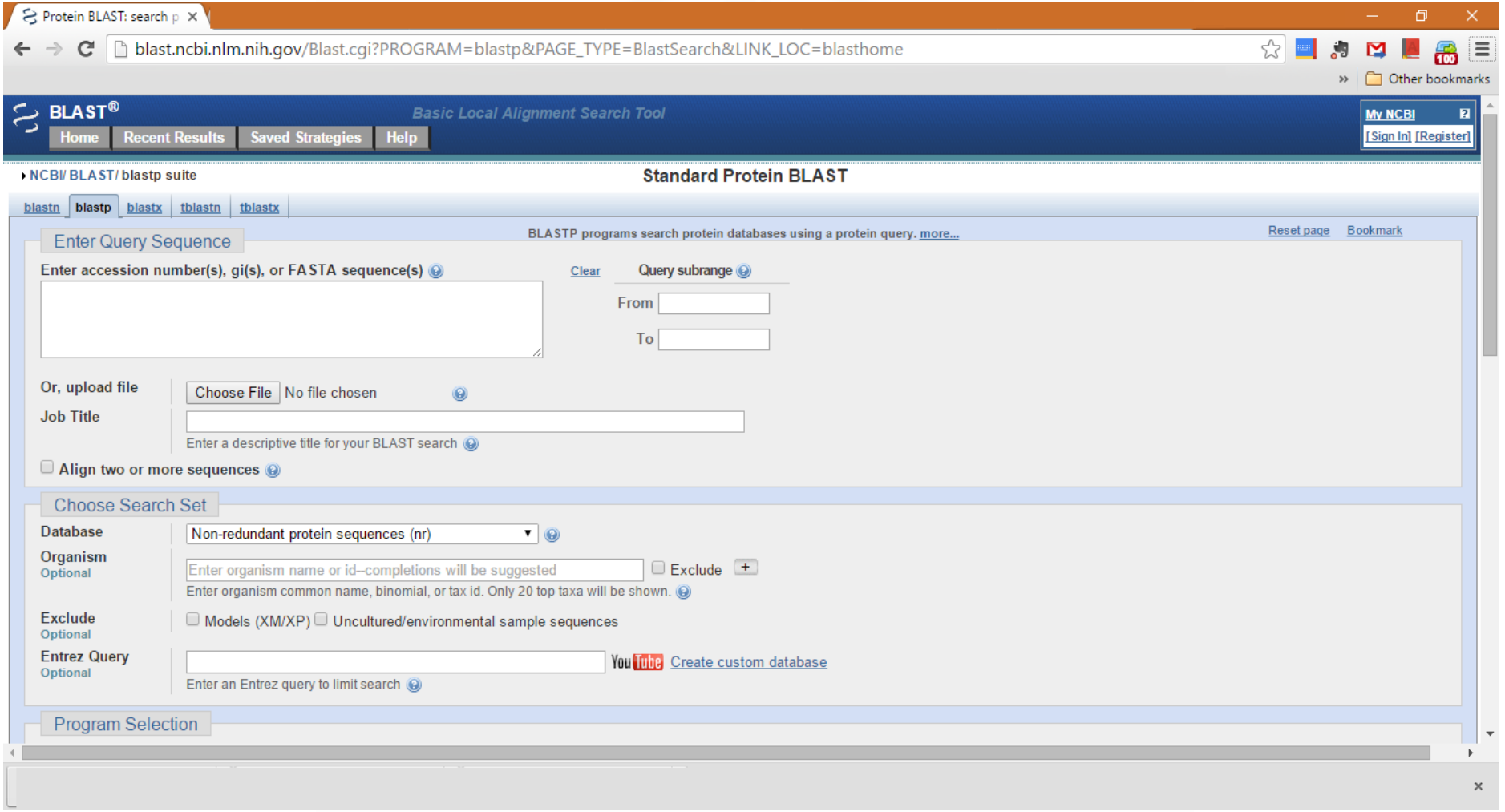
- Designed to find matches for sequences that are closely related (originally built to help annotate vertebrate genomes)
- Models splice junction dimers and performs "spliced alignment" (BLAST will generate matches to individual exons).
- Much faster than BLAST

BLAT

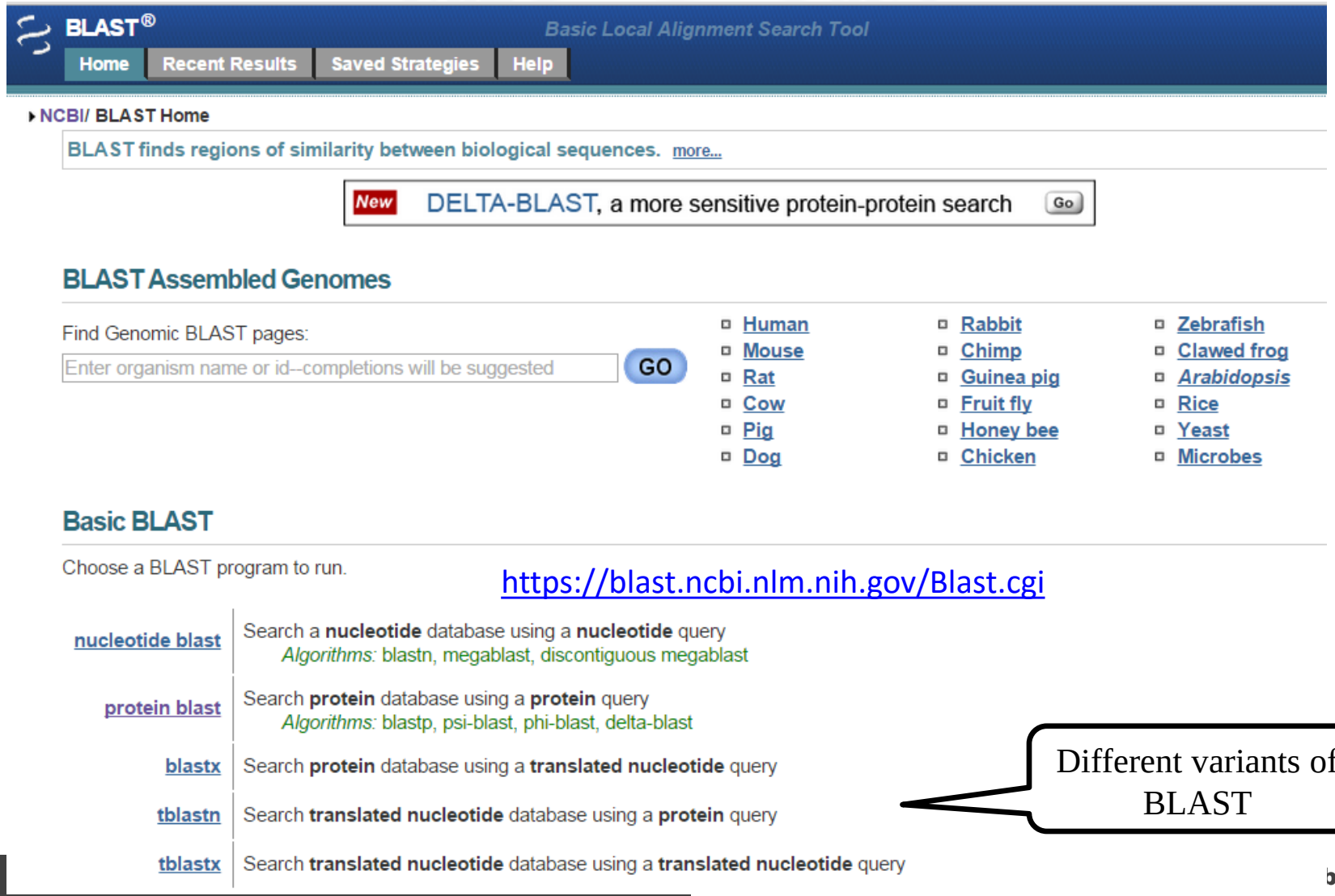
- Opening sentence in the paper:

  Some might wonder why in the year 2002 the world needs another sequence alignment tool.

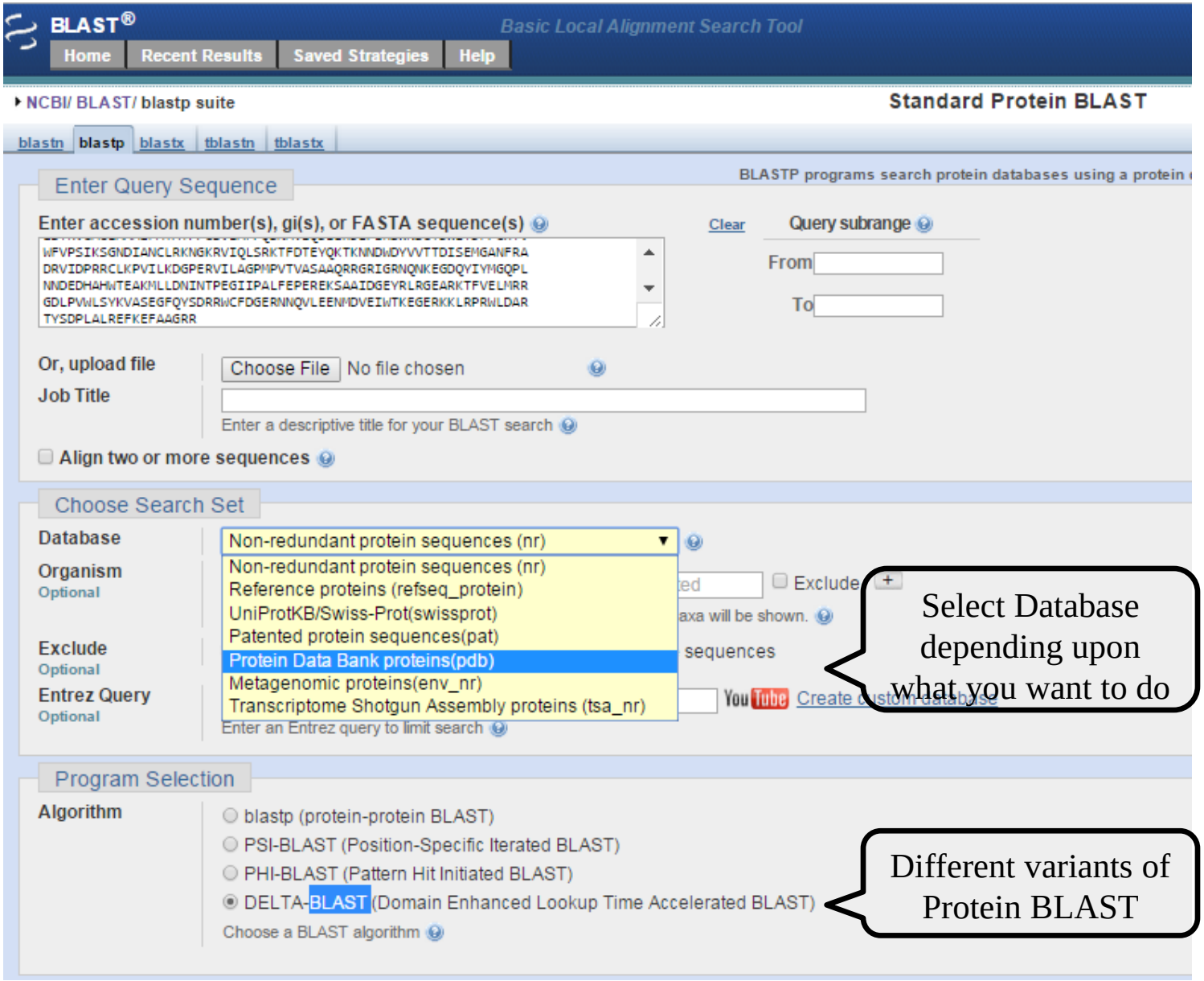
A BLAST demo…


Protein BLAST: search p
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
Other bookmarks
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein query. more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Choose File
No file chosen
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query
Optional
Create custom database
Enter an Entrez query to limit search
Program Selection

Finding Homologs: BLAST



Different variants of BLAST


BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastp suite
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Choose File
No file chosen
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Non-redundant protein sequences (nr)
Reference proteins (refseq_protein)
UniProtKB/Swiss-Prot(swissprot)
Patented protein sequences(pat)
Protein Data Bank proteins(pdb)
Metagenomic proteins(env_nr)
Transcriptome Shotgun Assembly proteins (tsa_nr)
Organism
Optional
Exclude
Exclude
Optional
sequences
Entrez Query
Optional
Create custom database
Enter an Entrez query to limit search
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
Select Database depending upon what you want to do
Different variants of Protein BLAST

Protein BLAST Variants

- BlastP: protein-protein BLAST
  - simply compares a protein query to a protein database.
- PSI-BLAST: Position-Specific Iterated BLAST
  - Allows the user to build a PSSM (position-specific scoring matrix) using the results of the first BlastP run.)
- PHI-BLAST: Pattern Hit Initiated BLAST
  - performs the search but limits alignments to those that match a pattern in the query.
- DELTA-BLAST: Domain Enhanced Lookup Time Accelerated BLAST
  - constructs a PSSM using the results of a Conserved Domain Database search and searches a sequence database.

Selecting Database

- Non-redundant Protein Sequences (nr)
  - All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
  - 67,897,003 Sequences
  - To build profiles or simply find related sequences
- Reference Proteins (refseq_protein)
  - NCBI Protein Reference Sequences
  - 49,158,634 Sequences
  - To build profiles or simply find related sequences
- Non-redundant UniProtKB/SwissProt sequences.
  - 461146 Sequences
  - Known Functions
- PDB
  - 77, 4658 Sequences
  - Known Structures

Assume you are given a protein

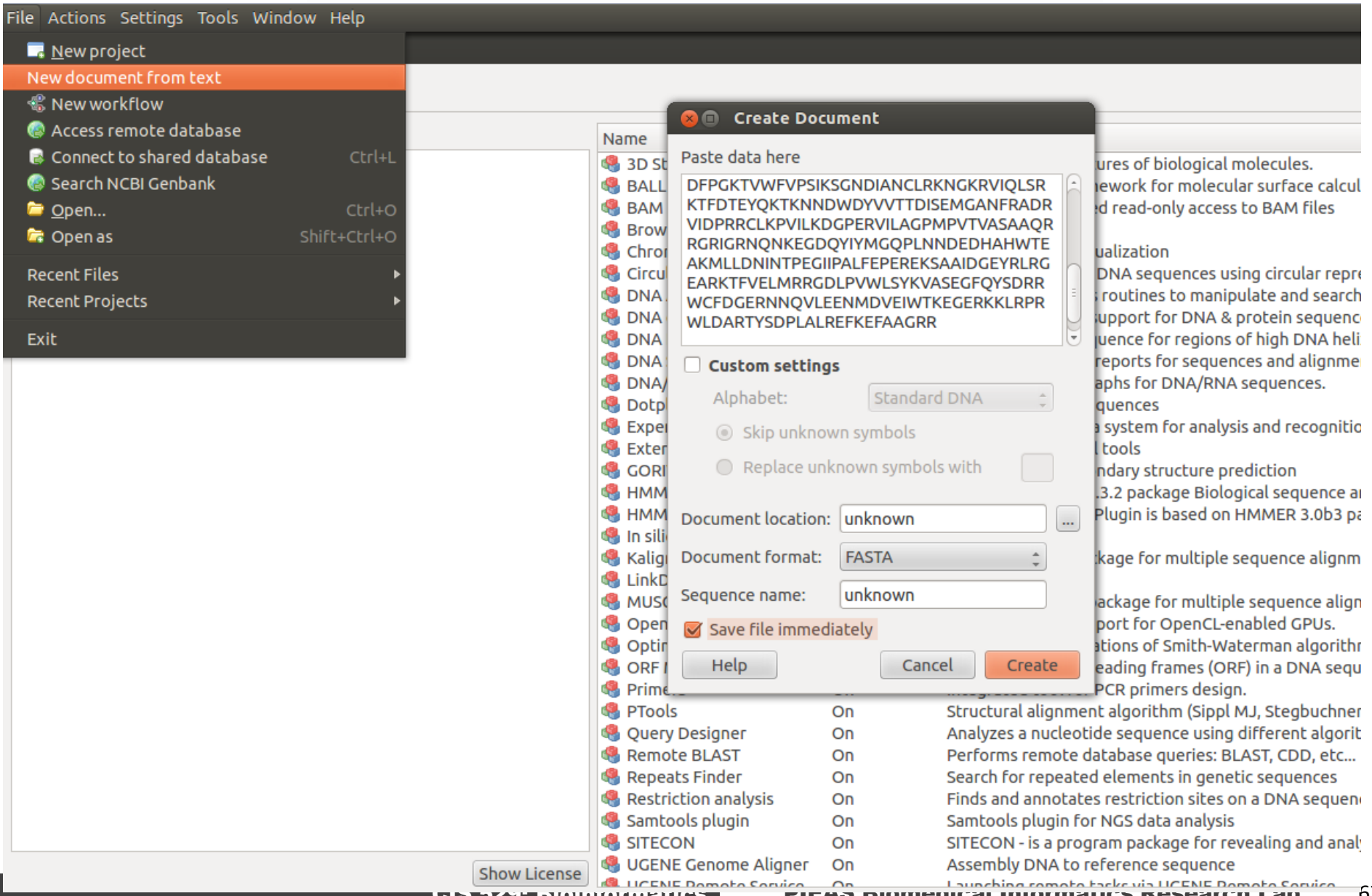
SGVLWDTPSPPEVEKAVLDDGIYRILQRGLLGRSQVGVGVFQEGVFHTMWHVTRGAVLMYQGKRLE
PSWASVKKDLISYGGGWRFQGSWNTGEEVQVIAVEPGKNPKNVQTAPGTFKTSEGEVGAIALDFKPG
TSGSPIVNREGKIVGLYGNGVVTTSGTYVSAIAQAKASQEGPLPEIEDEVFRKRNLTIMDLHPGSGKTRR
YLPAIVREAIKRKLRTLVLAPTRVVASEMAEALKGMPIRYQTTAVKSEHTGKEIVDLMCHATFTMRLLSP
VRVPNYNMIIMDEAHFTDPASIAARGYISTRVGMGEAAAIFMTATPPGSVEAFPQSNAVIQDEEKDIP
ERSWNSGYDWITDFPGKTVWFVPSIKSGNDIANCLRKNGKRVIQLSRKTFDTEYQKTKNNDWDYVVT
TDISEMGANFRADRVIDPRRCLKPVILKDGPERVILAGPMPVTVASAAQRRGRIGRNQNKEGDQYIY
MGQPLNNDEDHAHWTEAKMLLDNINTPEGIIPALFEPEREKSAAIDGEYRLRGEARKTFVELMRRGD
LPVWLSYKVASEGFQYSDRRWCFDGERNNQVLEENMDVEIWTKEGERKKLRPRWLDARTYSDPLAL
REFKEFAAGRR

- What can you find out about the function of the protein?
- This protein is taken from Dengue Virus

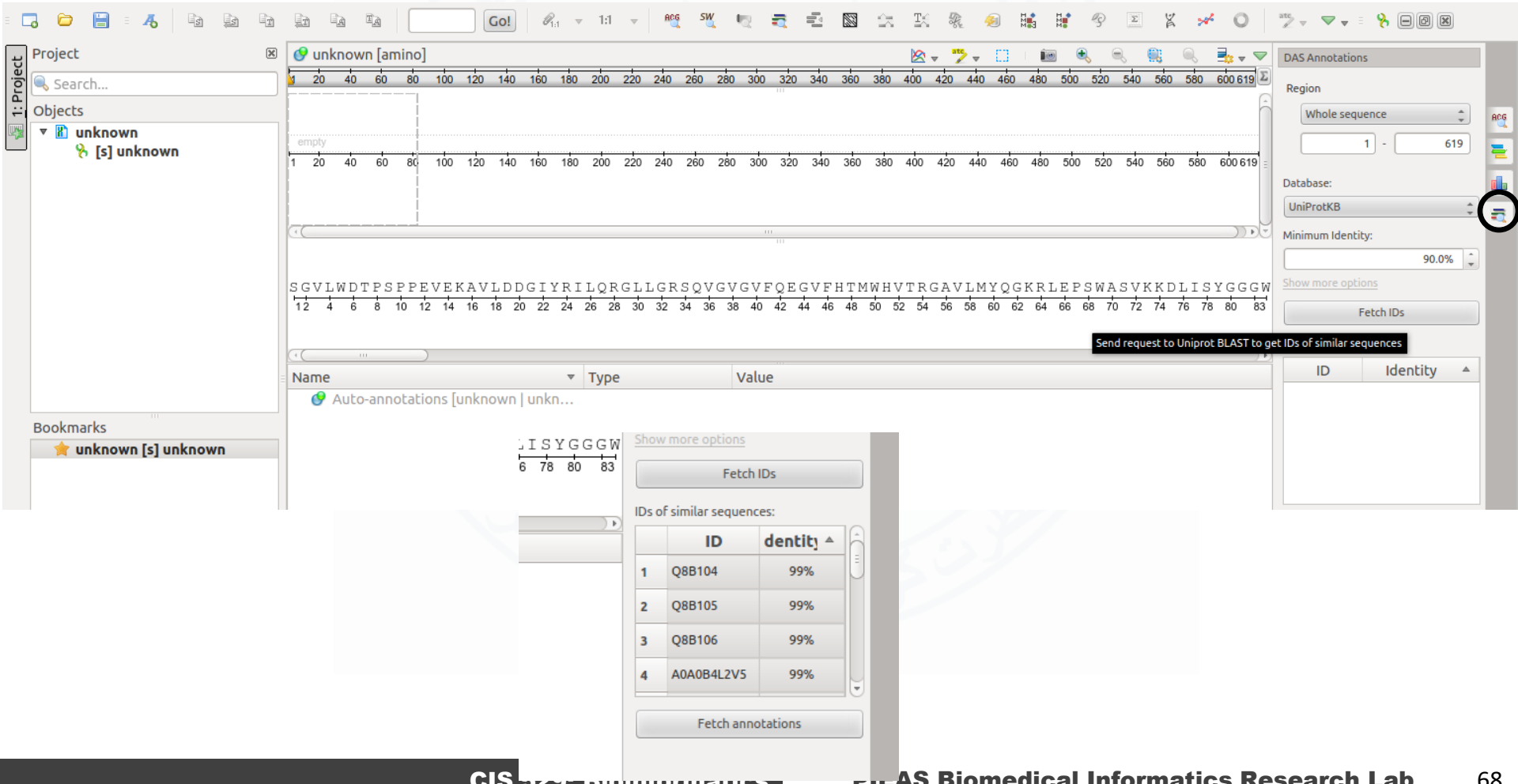
Inferring Function using DAS

- Uniprot Protein Distributed Annotation System (DAS) Server
  - Provides access to sequences and annotations from UniprotKB
  - Access to Gene Ontology Annotations
- Can search given a sequence for all annotations on related proteins or parts of the sequences
- Accessible using DASty (http://www.ebi.ac.uk/dasty/index.html) or UGENE

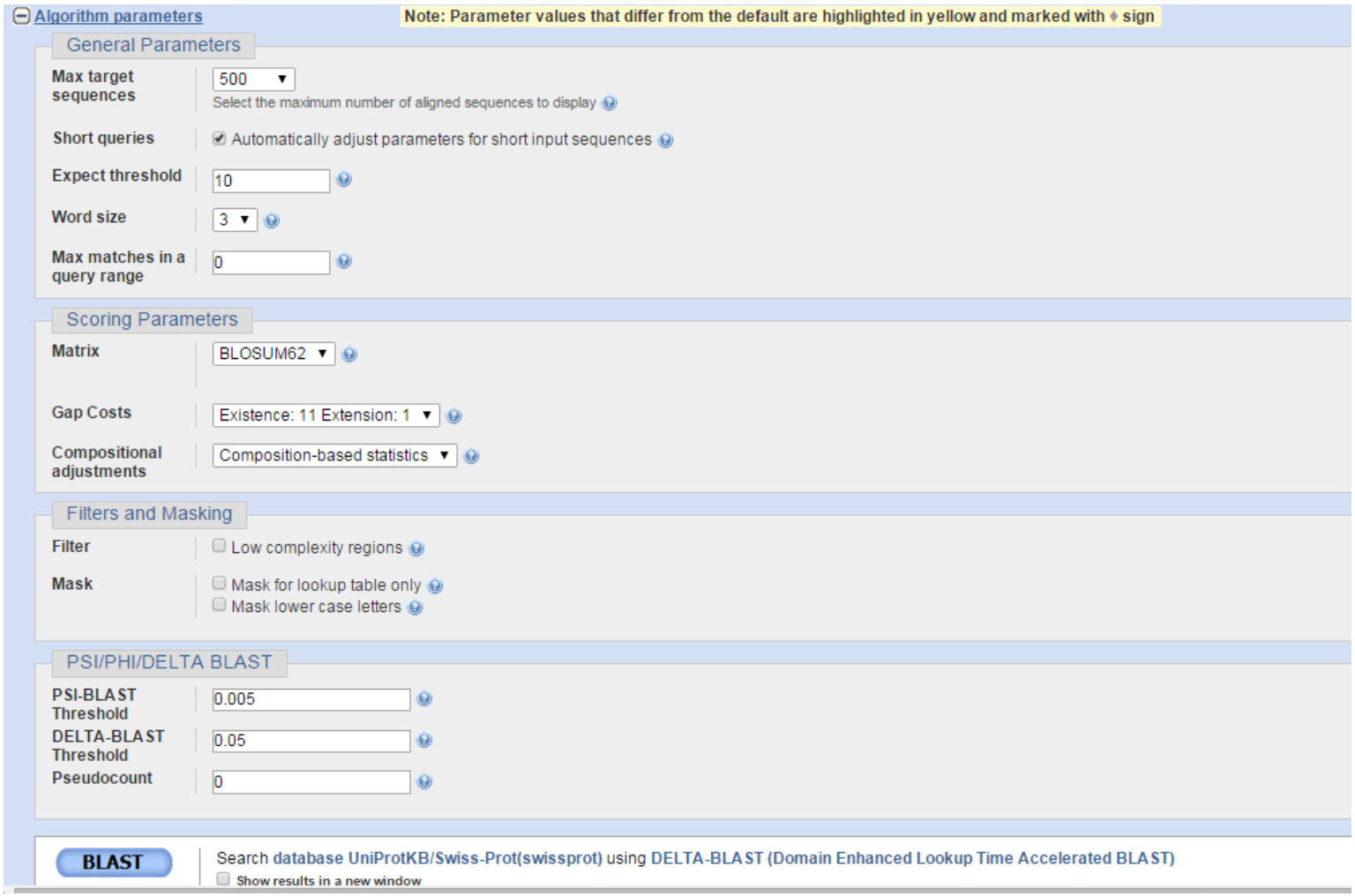
http://www.ebi.ac.uk/uniprot-das


File Actions Settings Tools Window Help
New project
New document from text
New workflow
Access remote database
Connect to shared database Ctrl+L
Search NCBI Genbank
Open... Ctrl+O
Open as Shift+Ctrl+O
Recent Files
Recent Projects
Exit
Create Document
Paste data here
DFPGKTVWFVPSIKSGNDIANCLRKNGKRVIQLSR
KTFDTEYQKTKNNDWDYVVTTDISEMGANFRADR
VIDPRRCLKPVILKDGPERVILAGPMPVTVASAAQR
RGRIGRNQNKEGDQYIYMGQPLNNDEDHAHWTE
AKMLLDNINTPEGIIPALFEPEREKSAAIDGEYRLRG
EARKTFVELMRRGDLPVWLSYKVASEGFQYSDRR
WCFDGERNNQVLEENMDVEIWTKEGERKKLRPR
WLDARTYSDPLALREFKEFAAGRR
Custom settings
Alphabet: Standard DNA
Skip unknown symbols
Replace unknown symbols with
Document location: unknown
Document format: FASTA
Sequence name: unknown
Save file immediately
Help
Cancel
Create
Name
PTools On Structural alignment algorithm (Sippl MJ, Stegbuchner
Query Designer On Analyzes a nucleotide sequence using different algorit
Remote BLAST On Performs remote database queries: BLAST, CDD, etc...
Repeats Finder On Search for repeated elements in genetic sequences
Restriction analysis On Finds and annotates restriction sites on a DNA sequen
Samtools plugin On Samtools plugin for NGS data analysis
SITECON On SITECON - is a program package for revealing and anal
UGENE Genome Aligner On Assembly DNA to reference sequence
Show License

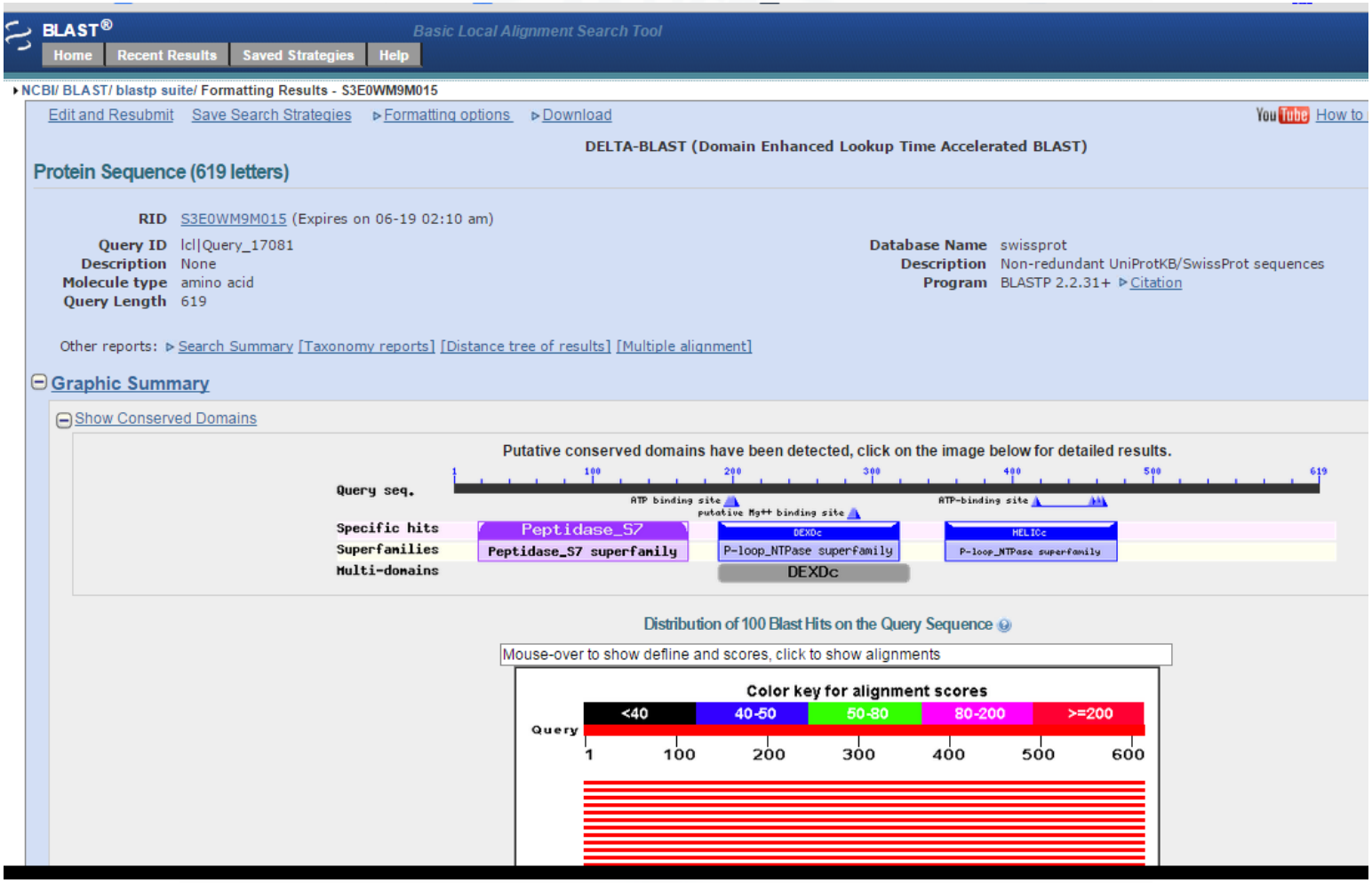
Using DAS


Project
Search...
Objects
unknown
[s] unknown
Bookmarks
unknown [s] unknown
unknown [amino]
SGVLWDTPSPPEVEKAVLDDGIYRILQRGLLGRSQVGVGVFQEGVFHTMWHVTRGAVLMYQGKRLEPSWASVKKDLISYGGGW
Name
Type
Value
Auto-annotations [unknown | unkn...
DAS Annotations
Region
Whole sequence
1
619
Database:
UniProtKB
Minimum Identity:
90.0%
Show more options
Fetch IDs
Send request to Uniprot BLAST to get IDs of similar sequences
ID
Identity
Show more options
Fetch IDs
IDs of similar sequences:
ID
dentity
1 Q8B104 99%
2 Q8B105 99%
3 Q8B106 99%
4 A0A0B4L2V5 99%
Fetch annotations

BLAST Parameters



Result


BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastp suite/ Formatting Results - S3E0WM9M015
Edit and Resubmit
Save Search Strategies
Formatting options
Download
How to
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Protein Sequence (619 letters)
RID S3E0WM9M015 (Expires on 06-19 02:10 am)
Query ID lcl|Query_17081
Description None
Molecule type amino acid
Query Length 619
Database Name swissprot
Description Non-redundant UniProtKB/SwissProt sequences
Program BLASTP 2.2.31+ Citation
Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]
Graphic Summary
Show Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
Query seq.
ATP binding site
putative Mg++ binding site
ATP-binding site
Specific hits
Superfamilies
Multi-domains
Peptidase_S7
Peptidase_S7 superfamily
DEXDc
P-loop_NTPase superfamily
DEXDc
HELICc
P-loop_NTPase superfamily
Distribution of 100 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
100
200
300
400
500
600

class: standard.
created: Feb 1, 1994.
sequence updated: Feb 1, 1994.
annotation updated: Jan 7, 2015.
xrefs: M87512.1, A42551
xrefs (non-sequence databases): ProteinModelPortal:P33478,

GO:0044167, GO:0042025, GO:0016021, GO:0019028, GO:0019031,
GO:0055036, GO:0005524, GO:0008026, GO:0003725, GO:0046872,
GO:0004482, GO:0004483, GO:0003724, GO:0003968, GO:0004252,
GO:0070008, GO:0005198, GO:0075512, GO:0039654, GO:0039520,
GO:0006355, GO:0039564, GO:0039574, GO:0039502, GO:0006351,
GO:0039694, GO:0019062, Gene3D:2.60.40.350, Gene3D:2.60.98.10,

Gene Ontology Information

InterPro:IPR011492, InterPro:IPR000069, InterPro:IPR013755,
InterPro:IPR001122, InterPro:IPR026470, InterPro:IPR001157,
InterPro:IPR000752, InterPro:IPR000487, InterPro:IPR000404,
InterPro:IPR001528, InterPro:IPR002535, InterPro:IPR000336,
InterPro:IPR001850, InterPro:IPR027287, InterPro:IPR014412,
InterPro:IPR011998, InterPro:IPR013754, InterPro:IPR014001,
InterPro:IPR001650, InterPro:IPR014756, InterPro:IPR026490,
InterPro:IPR027417, InterPro:IPR000208, InterPro:IPR007094,
InterPro:IPR002877, InterPro:IPR029063, InterPro:IPR009003,

InterPro:
Prediction of families, domains and sites

Pfam:PF01003, Pfam:PF07652, Pfam:PF02832, Pfam:PF00869,
Pfam:PF01004, Pfam:PF00948, Pfam:PF01005, Pfam:PF01002,
Pfam:PF01350, Pfam:PF01349, Pfam:PF00972, Pfam:PF01570,
Pfam:PF01728, Pfam:PF00271, Pfam:PF00949, PIRSF:PIRSF003817,
SMART:SM00487, SMART:SM00490, SUPFAM:SSF50494, SUPFAM:SSF52540,
SUPFAM:SSF53335, SUPFAM:SSF56983, SUPFAM:SSF81296,
TIGRFAMs:TIGR04240, PROSITE:PS51527, PROSITE:PS51528,
PROSITE:PS51192, PROSITE:PS51194, PROSITE:PS50507, PROSITE:PS51591

KEYWORDS Activation of host autophagy by virus; ATP-binding; Capsid protein; Clathrin-mediated endocytosis of virus by host; Cleavage on pair of basic residues; Complete proteome; Disulfide bond; Fusion of virus membrane with host endosomal membrane; Fusion of virus membrane with host membrane; Glycoprotein; Helicase; Host endoplasmic reticulum; Host membrane; Host nucleus; Host-virus interaction; Hydrolase; Inhibition of host innate immune response by virus; Inhibition of host interferon signaling pathway by virus; Inhibition of host STAT2 by virus; Inhibition of host TYK2 by virus; Membrane; Metal-binding; Methyltransferase; mRNA capping; mRNA processing; Multifunctional enzyme; Nucleotide-binding; Nucleotidyltransferase; Phosphoprotein; Protease; RNA-binding; RNA-directed RNA polymerase; S-adenosyl-L-methionine; Secreted; Serine protease; Transcription; Transcription regulation; Transferase; Transmembrane; Transmembrane helix; Viral attachment to host cell; Viral envelope protein; Viral immunoevasion; Viral penetration into host cytoplasm; Viral RNA replication; Virion; Virus endocytosis by host; Virus entry into host cell.

SOURCE Dengue virus 1 Singapore/S275/1990
ORGANISM Dengue virus 1 Singapore/S275/1990
Viruses; ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Flaviviridae; Flavivirus; Dengue virus group.

REFERENCE 1 (residues 1 to 3396)
AUTHORS Fu,J., Tan,B.H., Yap,E.H., Chan,Y.C. and Tan,Y.H.
TITLE Full-length cDNA sequence of dengue type 1 virus (Singapore strain S275/90)

FEATURES	Location/Qualifiers
source	1..3396 /organism="Dengue virus 1 Singapore/S275/1990" /host="Aedes aegypti (Yellowfever mosquito) (Culex aegypti)" /host="Aedes albopictus (Asian tiger mosquito) (Stegomyia albopicta)" /host="Homo sapiens (Human)" /db_xref="taxon:33741"
Protein	1..3396 /product="Genome polyprotein" /EC_number="3.4.21.91" /EC_number="3.6.1.15" /EC_number="3.6.4.13" /EC_number="2.1.1.56" /EC_number="2.1.1.57" /EC_number="2.7.7.48" /UniProtKB_evidence="Inferred from homology"
Region	1..3396 /region_name="Mature chain" /experiment="experimental evidence, no additional details recorded" /note="Genome polyprotein. /FTId=PRO_0000405206."
Region	1..101 /region_name="Topological domain" /experiment="experimental evidence, no additional details recorded" /note="Cytoplasmic. {ECO:0000255}."
Region	1..100 /region_name="Mature chain" /experiment="experimental evidence, no additional details recorded" /note="Capsid protein C. {ECO:0000250}. /FTId=PRO_0000037894."
Region	5..114 /region_name="Flavi_capsid" /note="Flavivirus capsid protein C; pfam01003" /db_xref="CDD:250285"
Region	33..74 /region_name="Region of interest in the sequence" /experiment="experimental evidence, no additional details recorded" /note="Hydrophobic; homodimerization of capsid protein C. {ECO:0000250}."
Site	100..101 /site_type="cleavage" /experiment="experimental evidence, no additional details recorded" /note="Cleavage; by viral protease NS3. {ECO:0000255}."
Region	101..114 /region_name="Propeptide" /experiment="experimental evidence, no additional details recorded" /note="ER anchor for the protein C, removed in mature form by serine protease NS3. {FTId=PRO_0000037895."

Sequence Databases

- https://en.wikipedia.org/wiki/Biological_database
- https://en.wikipedia.org/wiki/List_of_biological_databases
- GENBANK

Protein databases

- Genbank: All DNA sequences
- Uniprot

http://en.wikipedia.org/wiki/UniProt

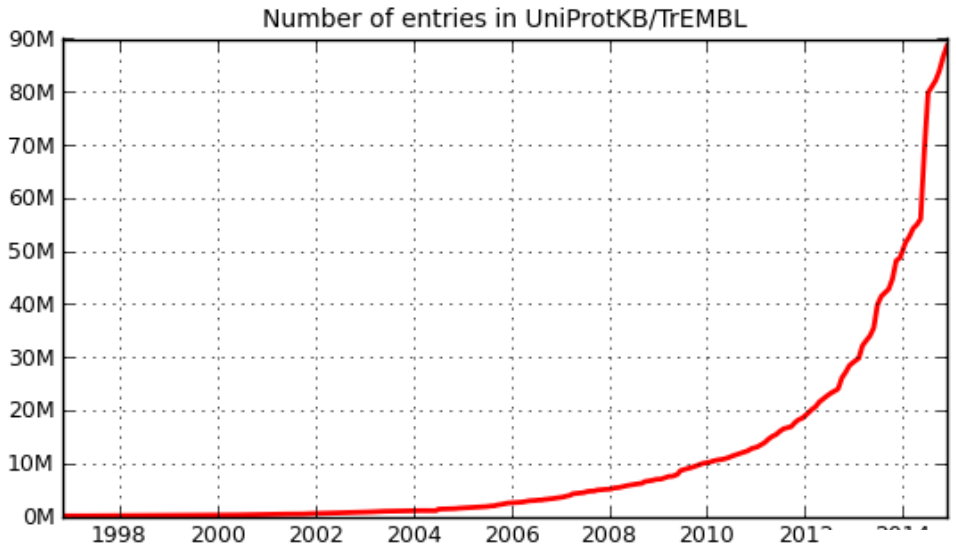
- SCOP: Structural classification of proteins
  - manual classification of protein structural domains based on similarities of their structures and amino acid sequences

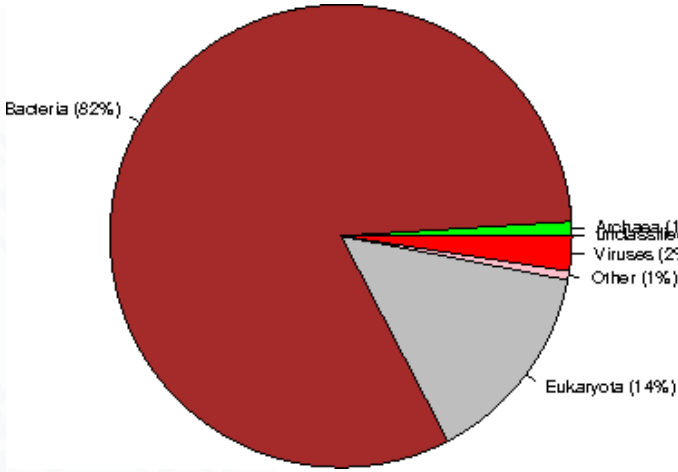
http://en.wikipedia.org/wiki/Structural_Classification_of_Proteins_database

- Pfam
  - Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models.

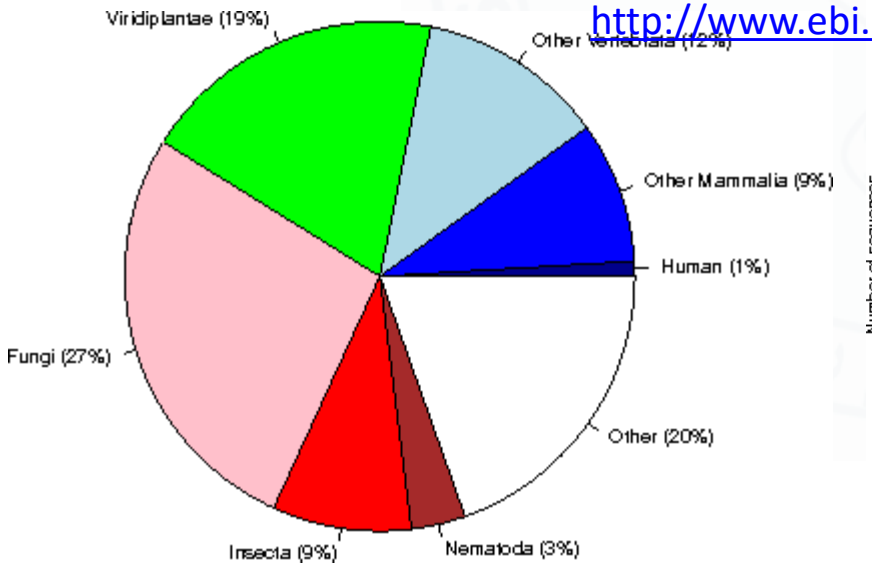
http://en.wikipedia.org/wiki/Pfam

Uniprot Statistics

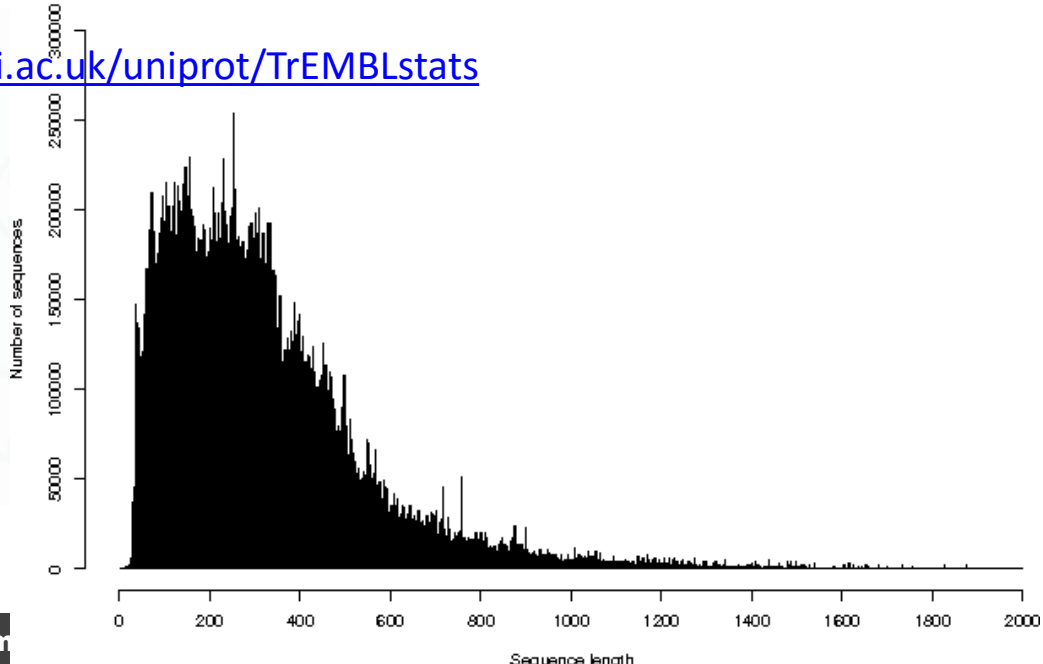




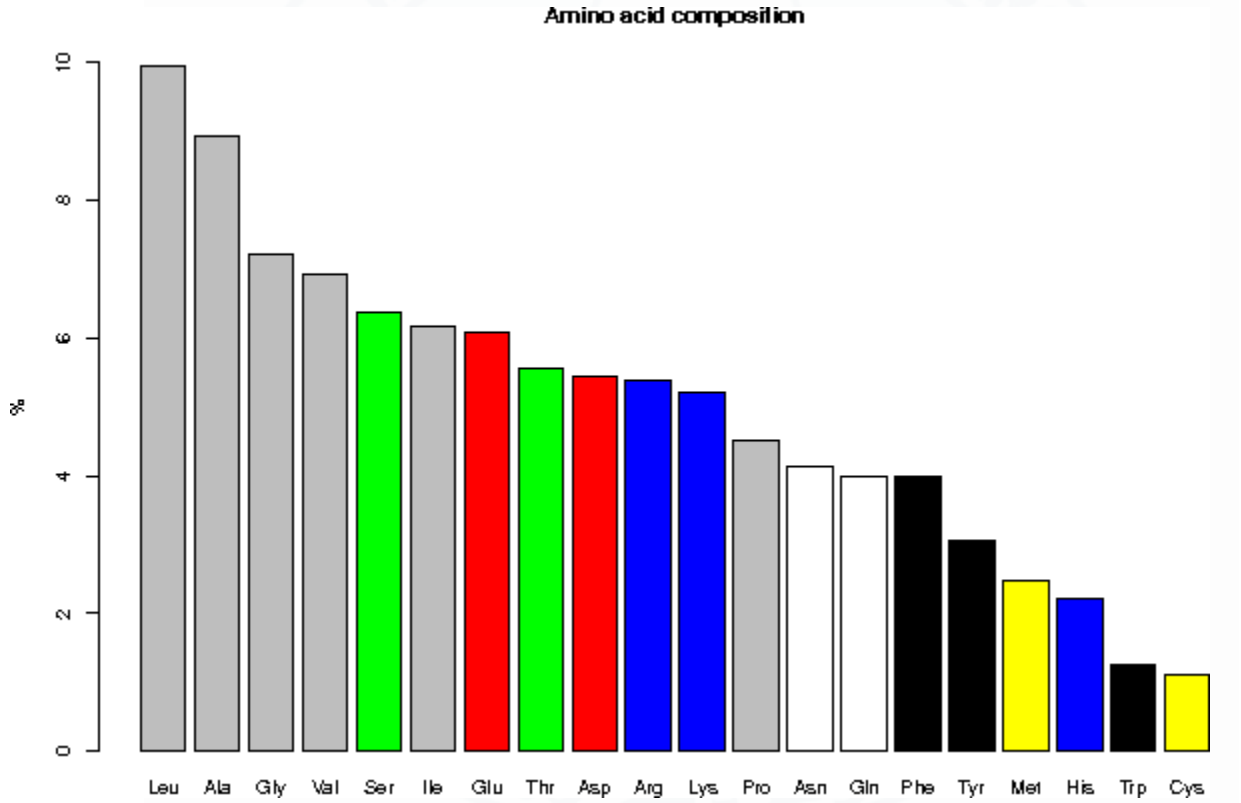
Length distribution of the sequences



http://www.ebi.ac.uk/uniprot/TrEMBLstats



Usage of amino acids



gray = aliphatic, red = acidic, green = small hydroxy, blue = basic, black = aromatic, white = amide, yellow = sulfur

Uniprot structure

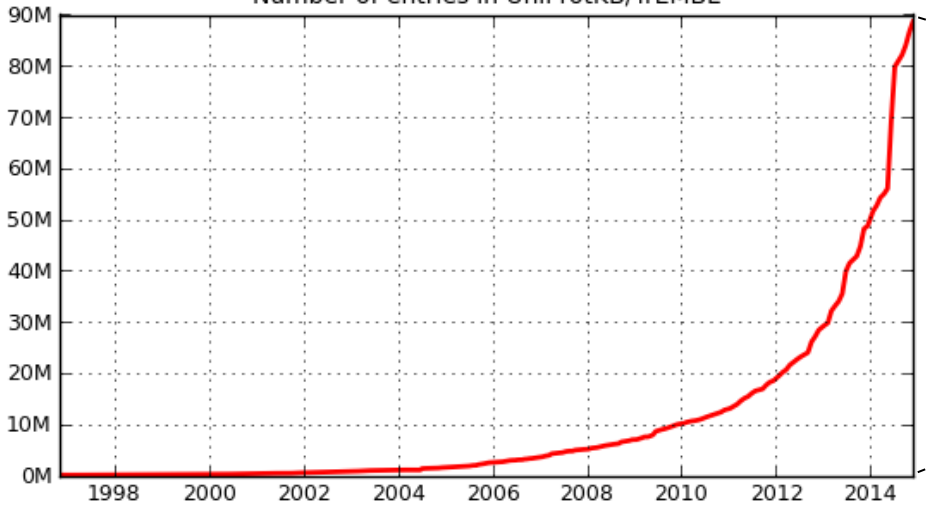
- "The primary mission of Universal Protein Resource (UniProt) is to support biological research by maintaining a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community"

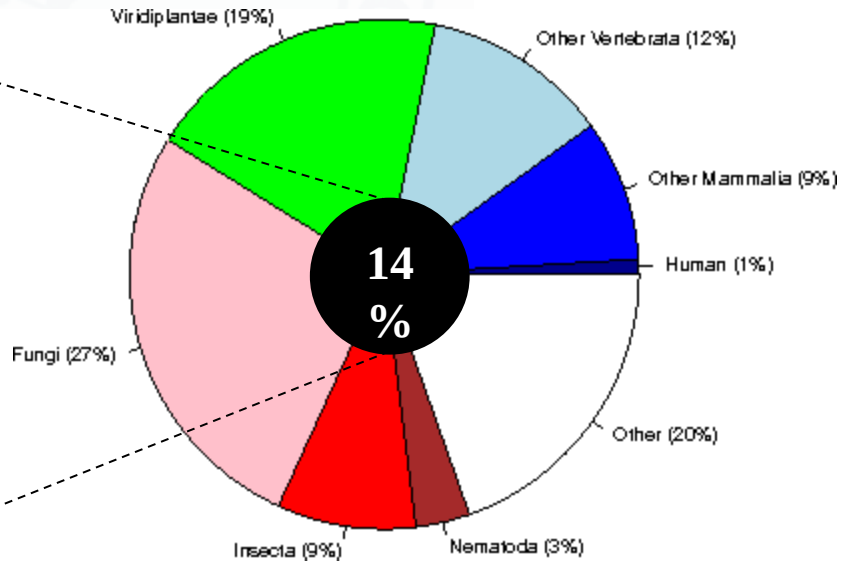
- Each protein has a uniprot accession number

"Ongoing and future developments at the Universal Protein Resource," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D214–D219, Jan. 2011.

Protein databases: Sequence

- Sequence
  - UniProt
    - UniProt Knowledgebase (UniProtKB): Expertly curated annotations
    - The UniProt Archive (UniParc): Reflects history of all proteins
    - UniProt Reference Clusters (UniRef): Closely related proteins are merged
    - UniProt Metagenomic and Environmental Sequences Database (UniMES)





Uniprot structure

- UniProt Knowledgebase (UniProtKB)
  - Expertly curated database,
  - Central access point for integrated protein information
  - Cross-references to multiple sources.
- The UniProt Archive (UniParc)
  - is a comprehensive sequence repository
  - Reflects the history of all protein sequences
- UniProt Reference Clusters (UniRef)
  - Merge closely related sequences based on sequence identity to facilitate sequence similarity searches
  - Using CD-HIT
- UniProt Metagenomic and Environmental Sequences Database (UniMES)
  - Caters for the expanding area of metagenomic data.

Uniprot structure

- UniProtKB/Swiss-Prot
  - 542,782 sequence entries in 2014_03 release
  - Reviewed, manually annotated entries
  - Annotations include
    - Protein and gene names
    - Function
    - Enzyme-specific information such as catalytic activity, cofactors and catalytic residues
    - Subcellular location
    - Protein-protein interactions
    - Pattern of expression
    - Locations and roles of significant domains and sites
    - Ion-, substrate- and cofactor-binding sites
    - Protein variant forms produced by natural genetic variation, RNA editing, alternative splicing, proteolytic processing, and post-translational modification
- UniProtKB/TrEMBL
  - Automatic annotation

Example

- http://www.uniprot.org/uniprot/P62158

Protein Databank

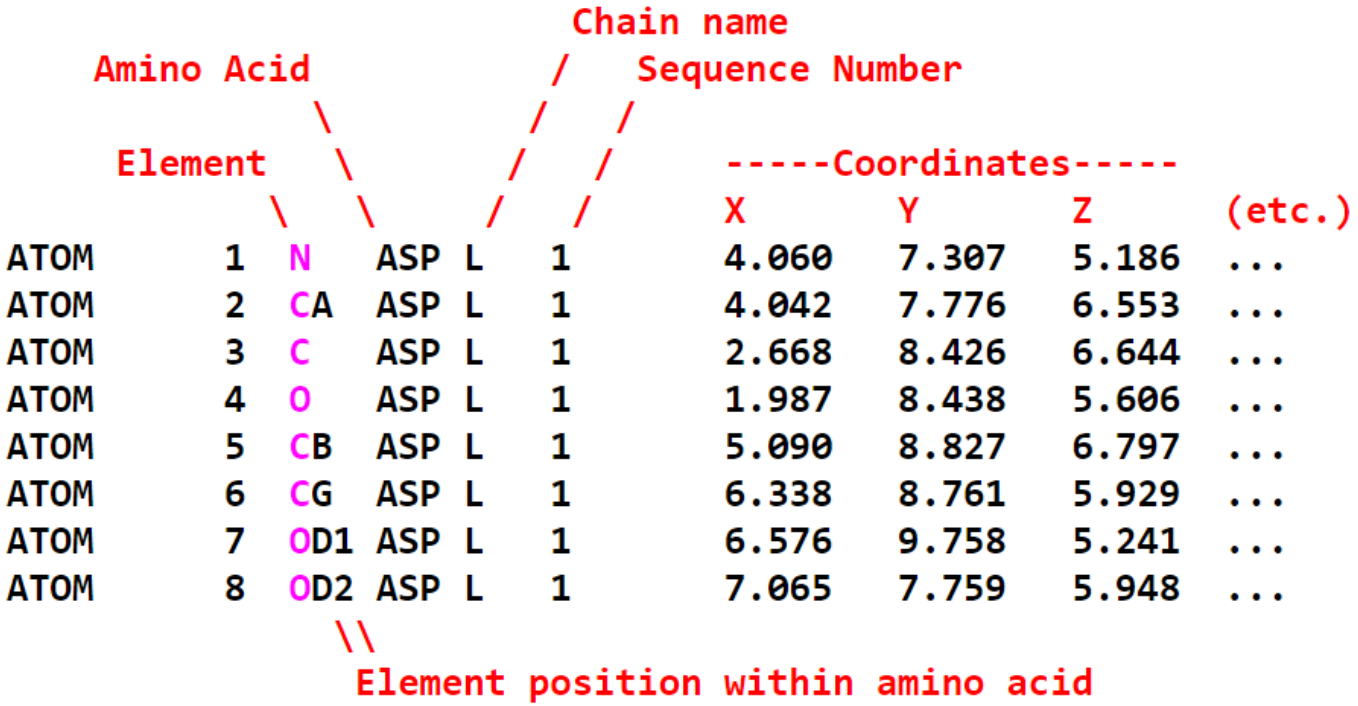
- http://proteopedia.org/wiki/index.php/Believe_It_or_Not
- http://en.wikipedia.org/wiki/Protein_Data_Bank

Experimental Method	Proteins	Nucleic Acids	Protein/Nucleic Acid complexes	Other	Total
X-ray diffraction	85406	1558	4547	5	91516
NMR	9281	1092	218	7	10598
Electron microscopy	579	67	190	0	836
Hybrid	63	3	2	1	69
Other	157	4	6	13	180
Total:	95486	2724	4963	26	**103199**

The PDB File

- Also contains header and other information

Atomic Coordinates: PDB Format



http://en.wikipedia.org/wiki/Protein_Data_Bank_(file_format)

http://proteopedia.org/wiki/index.php/Atomic_coordinate_file

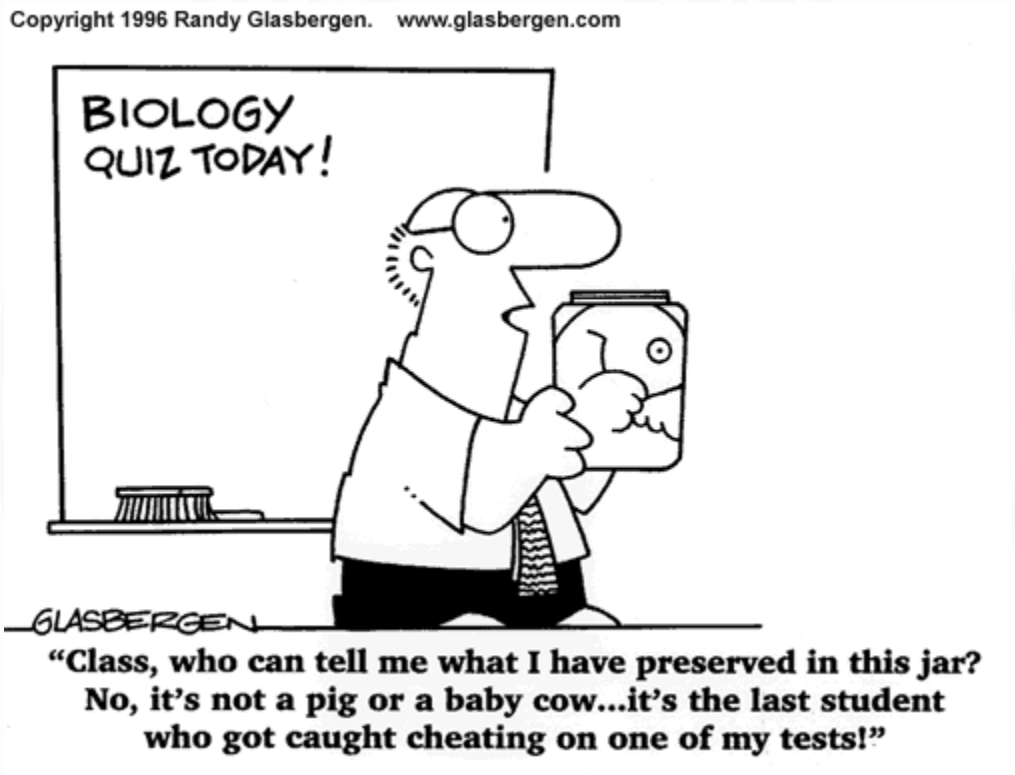
Other tools

- OCA
- FirstGlance
- PDBSum
- ccPDB

Tools for PDB

Database name	Database description
DSSP	Secondary structure of proteins http://swift.cmbi.ru.nl/gv/dssp/
HSSP	Multiple sequence alignments of UniProtKB against PDB files http://swift.cmbi.ru.nl/gv/hssp/
PDBREPORT	Reports about errors and anomalies in macromolecules http://swift.cmbi.ru.nl/gv/pdbreport/
PDB_REDO	Re-refined PDB files solved by X-ray crystallography http://www.cmbi.ru.nl/pdb_redo/
PDBFINDER	Searchable summaries of PDB file information ftp://ftp.cmbi.ru.nl/pub/molbio/data/pdbfinder/
PDBFINDER2	As PDBFINDER, but with much extra information added ftp://ftp.cmbi.ru.nl/pub/molbio/data/pdbfinder2/
PDB_SELECT	Quality-sorted culled lists of protein chains in the PDB http://swift.cmbi.ru.nl/gv/select/
WHY_NOT	Explanation why entries in other databases cannot exist http://www.cmbi.ru.nl/WHY_NOT/

R. P. Joosten, T. A. H. te Beek, E. Krieger, M. L. Hekkelman, R. W. W. Hooft, R. Schneider, C. Sander, and G. Vriend, "A series of PDB related databases for everyday needs," *Nucleic Acids Res.*, vol. 39, no. suppl 1, pp. D411–D419, Jan. 2011.



End of Lecture