BIMM 143

Introduction to Bioinformatics

Lecture 2

Barry Grant

UC San Diego

http://thegrantlab.org/bimm143

Recap From Last Time:

- Bioinformatics is computer aided biology.
  - Deals with the collection, archiving, organization, and interpretation of a wide range of biological data.
- The NCBI and EBI are major online bioinformatics service providers.
- Introduced via **hands-on session** the BLAST, Entrez, GENE, OMIM, UniProt, Muscle and PDB bioinformatics tools and databases.
    - Muddy point assessment (see results)
- There are a large number of bioinformatics databases (see handout!).
- Also covered: Course structure; Supporting course website, Ethics code, and Introductions…

Today's Menu

Classifying Databases	Primary, secondary and composite Bioinformatics databases
Using Databases	**Vignette** demonstrating how major Bioinformatics databases intersect
Major Biomolecular Formats	How nucleotide and protein sequence and structure data are represented
Alignment Foundations	**Introducing the *why* and *how* of comparing sequences**
Alignment Algorithms	**Hands-on** exploration of alignment algorithms and applications

Primary, secondary & composite databases

Bioinformatics databases can be usefully classified into *primary*, *secondary* and *composite* according to their data source.

- **Primary databases** (or *archival databases*) consist of data derived experimentally.
  - **GenBank**: NCBI's primary nucleotide sequence database.
  - **PDB:** Protein X-ray crystal and NMR structures.

- **Secondary databases** (or *derived databases*) contain information derived from a primary database.
  - **RefSeq**: non redundant set of curated reference sequences primarily from GenBank
  - **PFAM**: protein sequence families primarily from UniProt and PDB

- **Composite databases** (or *metadatabases*) join a variety of different primary and secondary database sources.
  - **OMIM**: catalog of human genes, genetic disorders and related literature
  - **GENE**: molecular data and literature related to genes with extensive links to other databases.

DATABASE VIGNETTE

You have just come out a seminar about gastric cancer and one of your co-workers asks:

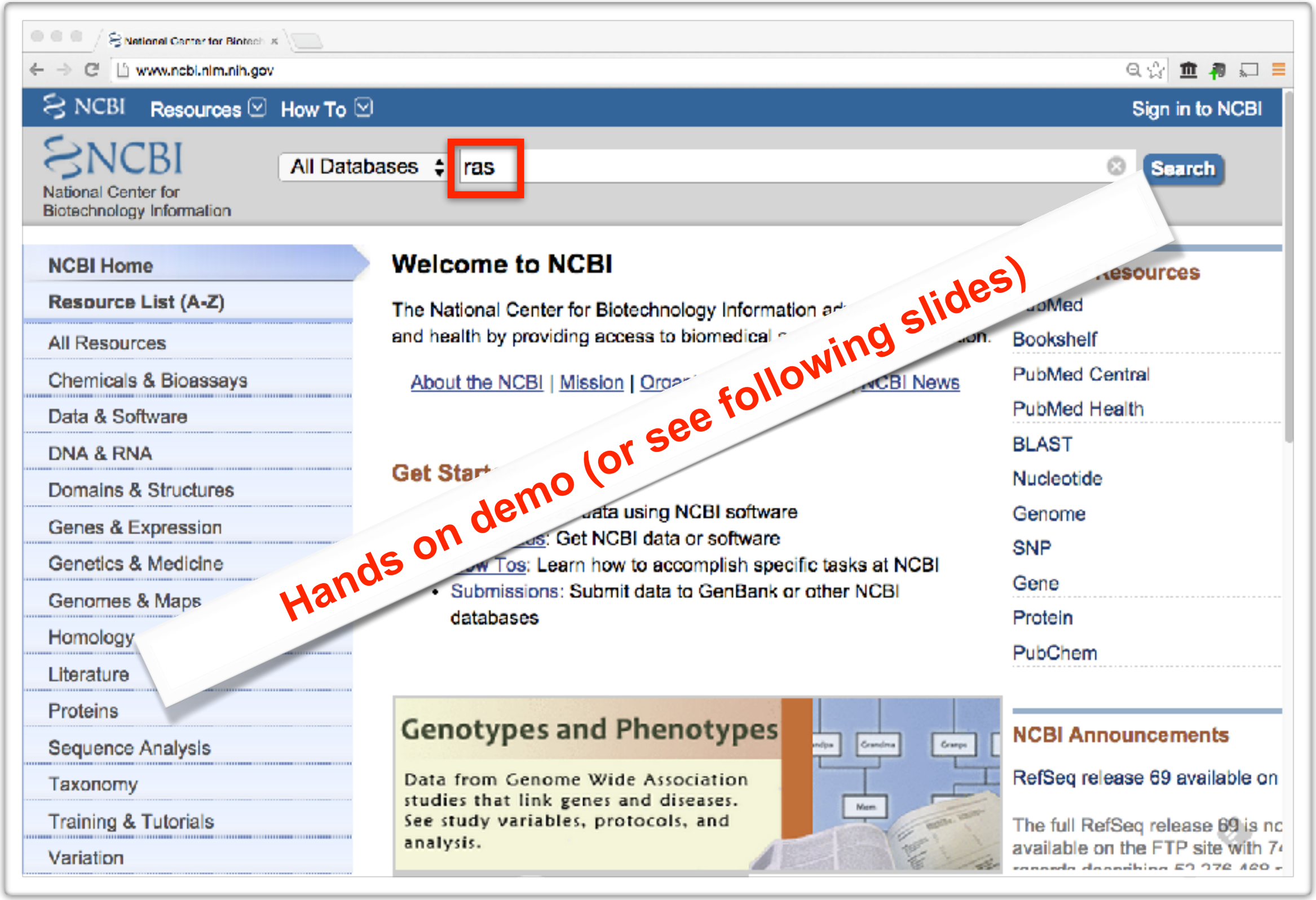
> *"What do you know about that 'Kras' gene the speaker kept taking about?*"

You have some recollection about hearing of 'Ras' before. How would you find out more?

- Google?
- Library?
- **Bioinformatics databases at NCBI and EBI!**

http://www.ncbi.nlm.nih.gov/

http://www.ncbi.nlm.nih.gov/



Example Vignette Questions:

- What chromosome location and what genes are in the vicinity of a given query gene? NCBI **GENE**
- What can you find out about molecular functions, biological processes, and prominent cellular locations? EBI **GO**
- What amino acid positions in the protein are responsible for ligand binding? EBI **UniProt**
- What variants of this gene are associated with gastric cancer and other human diseases? NCBI **OMIN**
- What is known about the protein family, its species distribution, number in humans and residue-wise conservation? EBI **PFAM**
- Are high resolution protein structures available to examine the details of these mutations? How might we explain their potential molecular effects? RCSB **PDB**

www.ncbi.nlm.nih.gov/gquery/?term=ras

NCBI Resources How To Sign in to NCBI

Search NCBI databases

Help

ras Search

About 2,978,774 search results for "ras"

Literature

Books	1,677	books and reports
MeSH	402	ontology used for PubMed indexing
NLM Catalog	223	books, journals and more in the NLM Collections
PubMed	54,672	scientific & medical abstracts/citations
PubMed Central	96,114	full-text journal articles

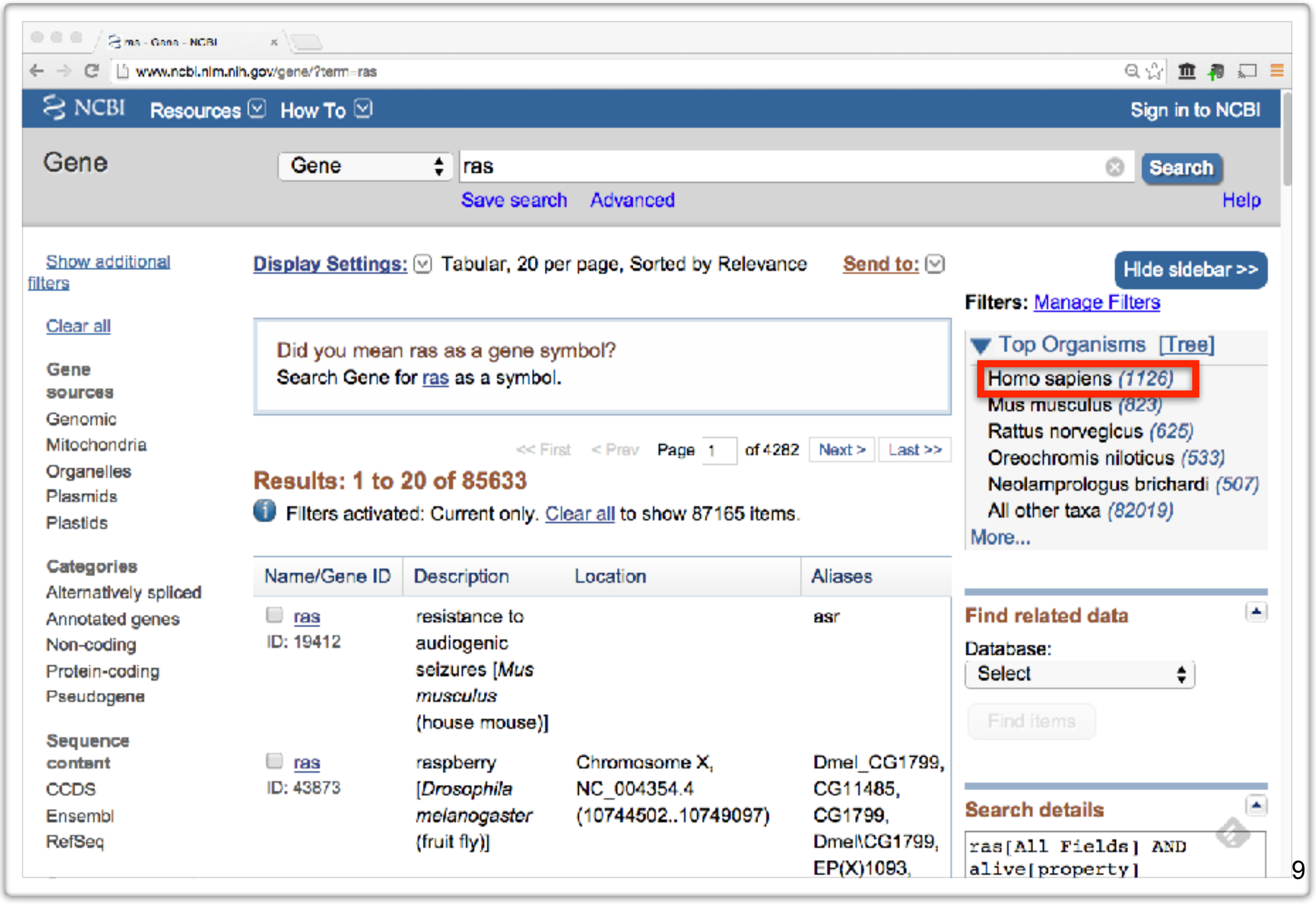
Health

ClinVar	759	human variations of clinical significance
dbGaP	120	genotype/phenotype interaction studies
GTR	1,879	genetic testing registry

Genes

EST	3,985	expressed sequence tag sequences
Gene	87,165	collected information about gene loci
GEO DataSets	3,732	functional genomics studies
GEO Profiles	1,622,789	gene expression and molecular abundance profiles
HomoloGene	696	homologous gene sets for selected organisms
PopSet	2,254	sequence sets from phylogenetic and population studies
UniGene	4,770	clusters of expressed transcripts

Proteins


ras - Gene - NCBI
www.ncbi.nlm.nih.gov/gene/?term=ras
NCBI
Resources
How To
Sign in to NCBI
Gene
Gene
ras
Search
Save search
Advanced
Help
Show additional filters
Display Settings: Tabular, 20 per page, Sorted by Relevance
Send to:
Hide sidebar >>
Filters: Manage Filters
Clear all
Did you mean ras as a gene symbol?
Search Gene for ras as a symbol.
Top Organisms [Tree]
Homo sapiens (1126)
Mus musculus (823)
Rattus norvegicus (625)
Oreochromis niloticus (533)
Neolamprologus brichardi (507)
All other taxa (82019)
More...
Gene sources
Genomic
Mitochondria
Organelles
Plasmids
Plastids
Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene
Sequence content
CCDS
Ensembl
RefSeq
<< First
< Prev
Page 1
of 4282
Next >
Last >>
Results: 1 to 20 of 85633
Filters activated: Current only. Clear all to show 87165 items.
Name/Gene ID
Description
Location
Aliases
ras
ID: 19412
resistance to audiogenic seizures [Mus musculus (house mouse)]
asr
ras
ID: 43873
raspberry [Drosophila melanogaster (fruit fly)]
Chromosome X, NC_004354.4 (10744502..10749097)
Dmel_CG1799, CG11485, CG1799, Dmel\CG1799, EP(X)1093,
Find related data
Database:
Select
Find items
Search details
ras[All Fields] AND alive[property]

www.ncbi.nlm.nih.gov/gene

NCBI Resources How To Sign in to NCBI

Gene

Gene (ras) AND "Homo sapiens"[porgn:__txid9606] Search

Help

Show additional filters

Display Settings: Tabular, 20 per page, Sorted by Relevance Send to:

Hide sidebar >>

Filters: Manage Filters

Clear all

Results: 1 to 20 of 1126 << First < Prev Page 1 of 57 Next > Last >>

Filters activated: Current only. Clear all to show 1499 items.

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene

Sequence content
CCDS
Ensembl
RefSeq

Status clear
✓ Current only

Chromosome locations

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [*Homo sapiens* (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [*Homo sapiens* (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS1, KRAS2, NS,

Find related data

Database:
Select

Find items

Search details

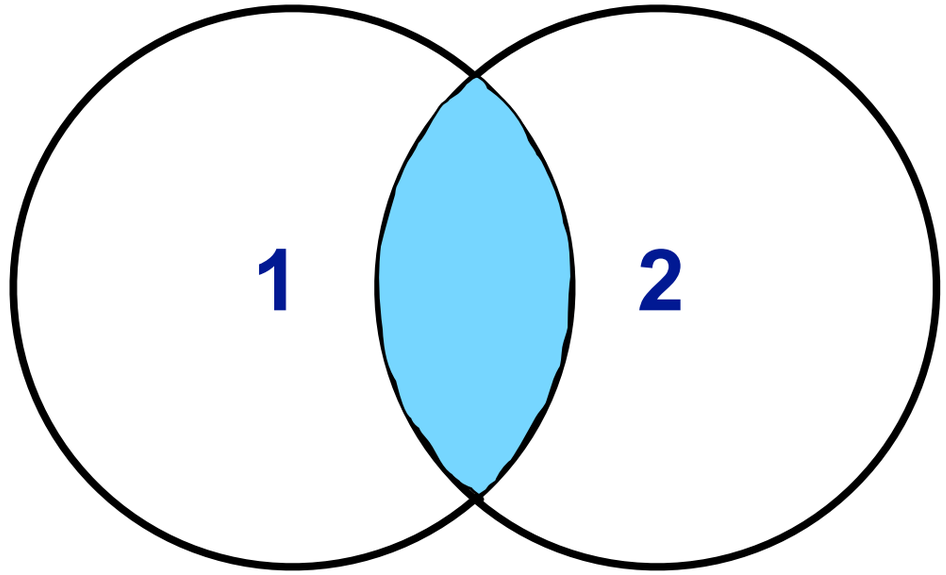
```
ras[All Fields] AND "Homo
sapiens"[porgn] AND
alive[property]
```

Search See more...

Recent activity

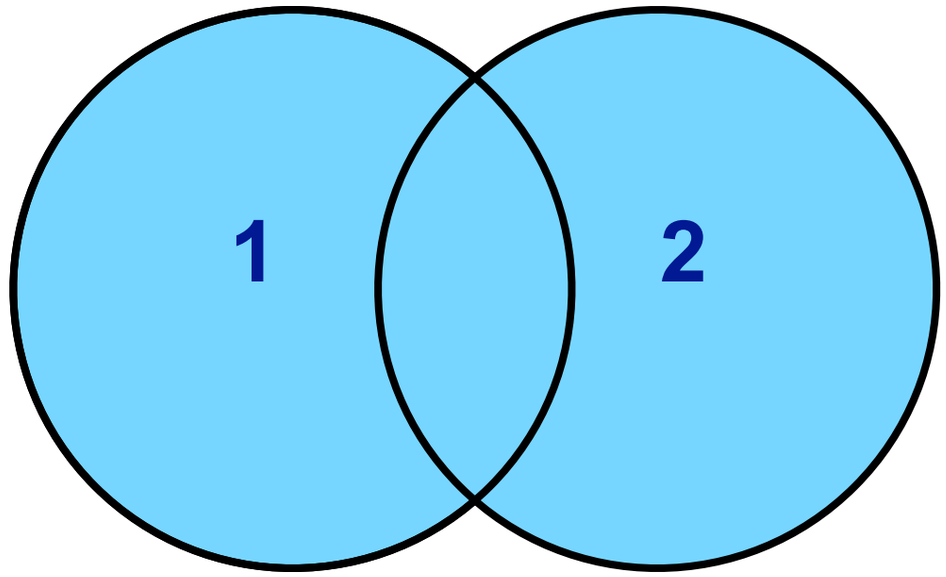
Turn Off Clear

1 AND 2



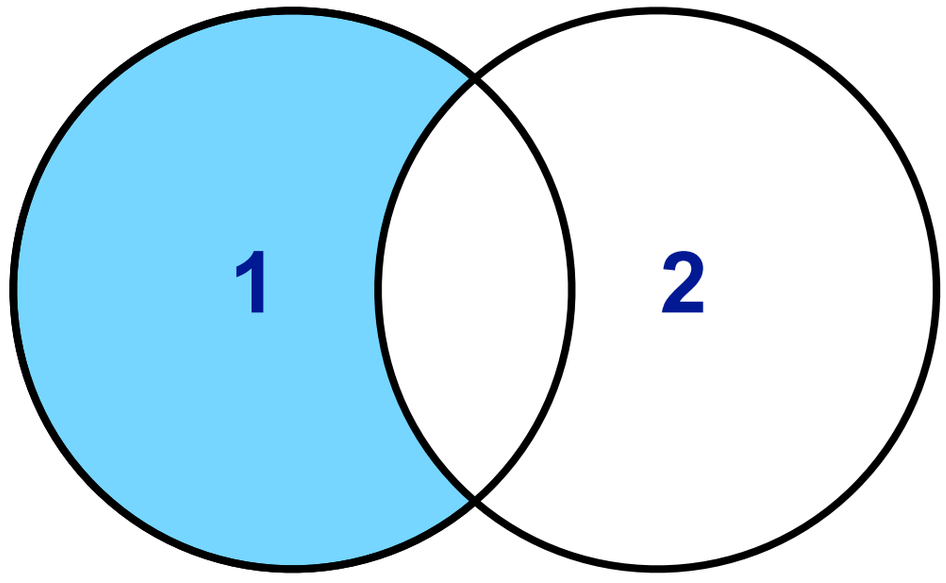
ras AND disease (1185 results)

1 OR 2



ras OR disease (134,872 results)

1 NOT 2



ras NOT disease (84,448 results)

www.ncbi.nlm.nih.gov/gene

NCBI Resources How To Sign in to NCBI

Gene

Gene (ras) AND "Homo sapiens"[porgn:__txid9606] Search

Save search Advanced Help

Show additional filters

Display Settings: Tabular, 20 per page, Sorted by Relevance Send to:

Hide sidebar >>

Clear all

Results: 1 to 20 of 1126 << First < Prev Page 1 of 57 Next > Last >>

Filters activated: Current only. Clear all to show 1499 items.

Filters: Manage Filters

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene

Sequence content
CCDS
Ensembl
RefSeq

Status clear
✓ Current only

Chromosome locations

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [*Homo sapiens* (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [*Homo sapiens* (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS1, KRAS2, NS,

Find related data

Database:
Select

Find items

Search details

```
ras[All Fields] AND "Homo
sapiens"[porgn] AND
alive[property]
```

Search See more...

Recent activity

Turn Off Clear

www.ncbi.nlm.nih.gov/gene/3845

NCBI Resources How To Sign in to NCBI

Gene Gene Search

Advanced Help

Display Settings: Full Report Send to: Hide sidebar >>

KRAS Kirsten rat sarcoma viral oncogene homolog [*Homo sapiens* (human)]

Gene ID: 3845, updated on 4-Jan-2015

Summary

Official Symbol	KRAS provided by HGNC
Official Full Name	Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source	HGNC:HGNC:6407
See related	Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Variation
- HIV-1 Interactions
- Pathways from BioSystems
- Interactions
- General gene information
  - Markers, Related pseudogene(s), Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)

www.ncbi.nlm.nih.gov/gene/3845

NCBI Resources How To Sign in to NCBI

Gene Search Help

Display S Hide sidebar >>

KRAS

(human)

Gene ID: 3845, updated on 4-Jan-2015

Summary

Official Symbol	KRAS provided by HGNC
Official Full Name	Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source	HGNC:HGNC:6407
See related	Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Variation
- HIV-1 Interactions
- Pathways from BioSystems
- Interactions
- General gene information
  - Markers, Related pseudogene(s), Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)

Example Questions:

What chromosome location and what genes are in the vicinity?

www.ncbi.nlm.nih.gov/gene/3845#genomic-context

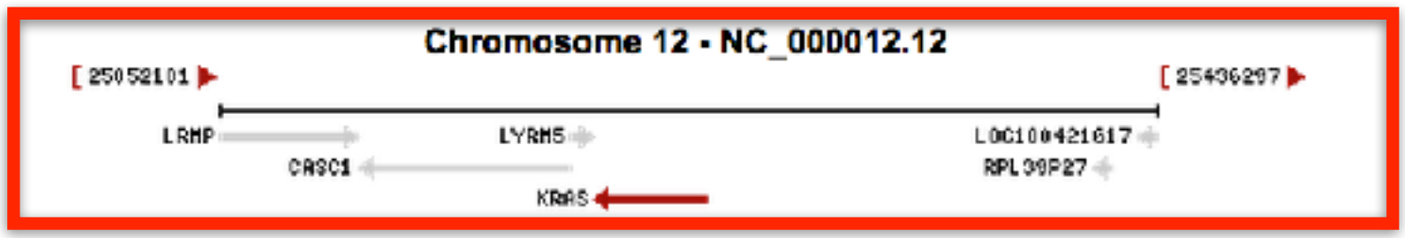
Genomic context

Location: 12p12.1

Exon count: 6

See KRAS in Epigenomics, MapViewer

Annotation release	Status	Assembly	Chr	Location
106	current	GRCh38 (GCF_000001405.26)	12	NC_000012.12 (25205246..25250923, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	12	NC_000012.11 (25358180..25403870, complement)



Genomic regions, transcripts, and products

Go to reference sequence details

Genomic Sequence: NC_000012.12 chromosome 12 reference GRCh38 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank

BioAssay by Target (Summary)
BioAssay, by Gene target
BioAssays, RNAi Target, Active
BioAssays, RNAi Target, Tested
BioProjects
BioSystems
Books
CCDS
ClinVar
Conserved Domains
dbVar
EST
Full text in PMC
Full text in PMC_nucleotide
Gene neighbors
Genome
GEO Profiles
GTR
HomoloGene
Map Viewer
MedGen
Nucleotide

www.ncbi.nlm.nih.gov/gene/3845

NCBI Resources How To Sign in to NCBI

Gene Search Help

Display Settin

KRAS Ki
(human)]

Gene ID: 3845

Summa

Example Questions:

What 'molecular functions', 'biological processes', and 'cellular component' information is available?

Hide sidebar >>

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Variation
- HIV-1 interactions
- Pathways from BioSystems
- Interactions
- General gene information
  - Markers, Related pseudogene(s), Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)

Official Symbol	KRAS provided by HGNC
Official Full Name	Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source	HGNC:HGNC:6407
See related	Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

www.ncbi.nlm.nih.gov/gene/3845#gene-ontology

Gene Ontology Provided by GOA

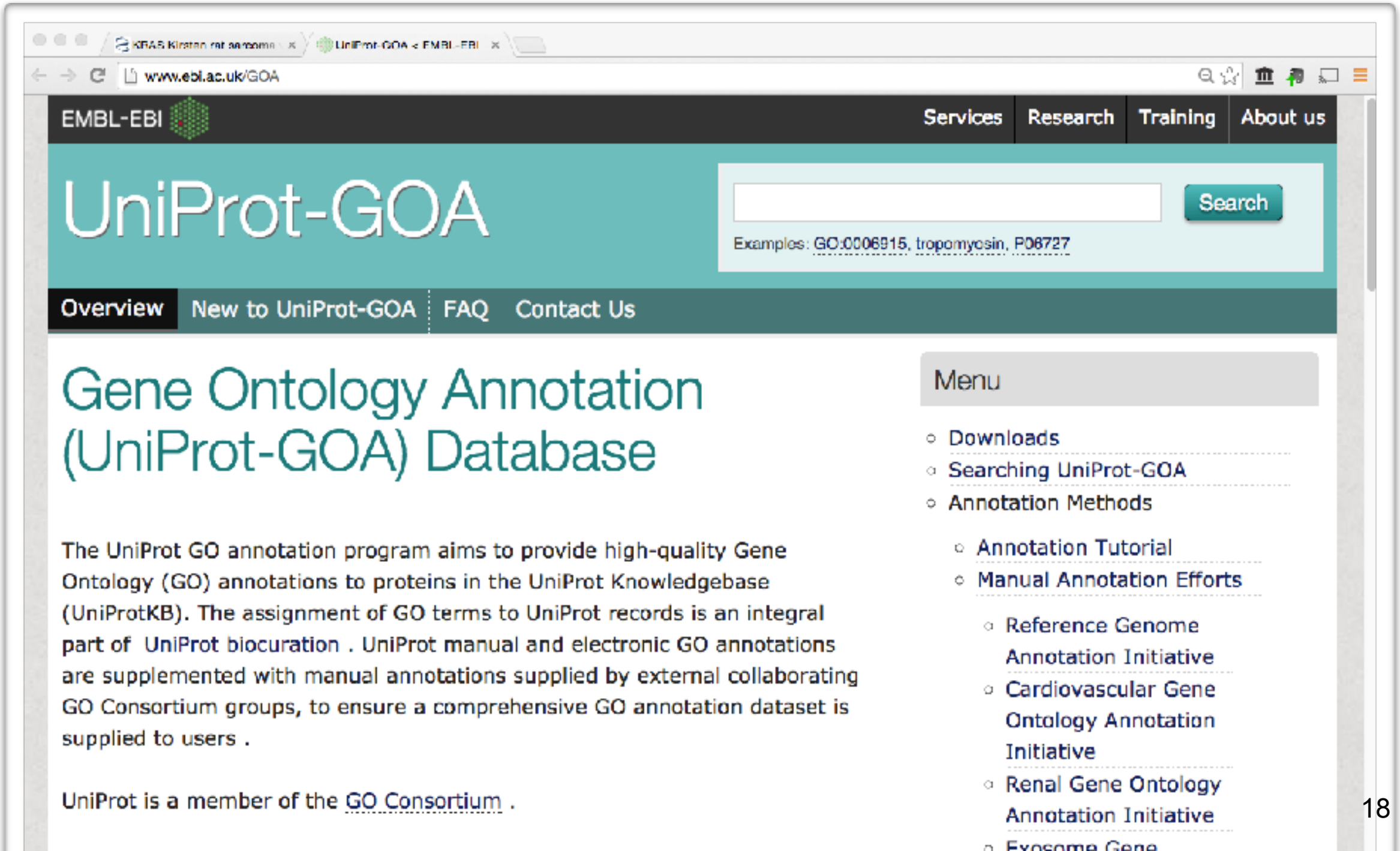
Function	Evidence Code	Pubs
GDP binding	IEA	
GMP binding	IEA	
GTP binding	IEA	
LRR domain binding	IEA	
protein binding	IPI	PubMed
protein complex binding	IDA	PubMed

Items 1 - 25 of 33 < Prev Page 1 of 2 Next >

Process	Evidence Code	Pubs
Fc-epsilon receptor signaling pathway	TAS	
GTP catabolic process	IEA	
MAPK cascade	TAS	
Ras protein signal transduction	TAS	
actin cytoskeleton organization	IEA	
activation of MAPKK activity	TAS	
axon guidance	TAS	
blood coagulation	TAS	

GO: Gene Ontology

GO provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data

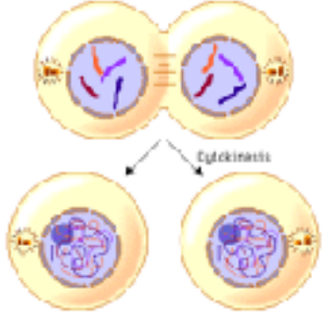


Why do we need Ontologies?

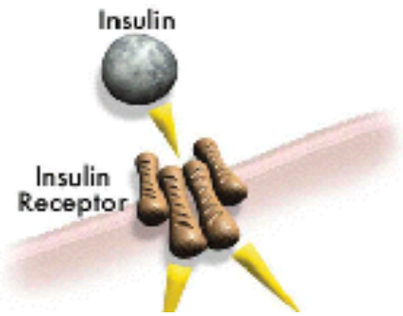
- Annotation is essential for capturing the understanding and knowledge associated with a sequence or other molecular entity
- Annotation is traditionally recorded as “free text”, which is easy to read by humans, but has a number of disadvantages, including:
  - Difficult for computers to parse
  - Quality varies from database to database
  - Terminology used varies from annotator to annotator
- Ontologies are annotations using standard vocabularies that try to address these issues
- GO is integrated with UniProt and many other databases including a number at NCBI

GO Ontologies

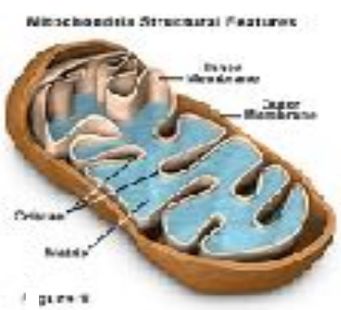
- There are three ontologies in GO:
  - **Biological <u>Process</u>**
    A commonly recognized series of events
    e.g. cell division, mitosis,



  - **Molecular <u>Function</u>**
    An elemental activity, task or job
    e.g. kinase activity, insulin binding



  - **Cellular <u>Component</u>**
    Where a gene product is located
    e.g. mitochondrion, mitochondrial membrane



www.ncbi.nlm.nih.gov/gene/3845#gene-ontology

Gene Ontology Provided by GOA

Function	Evidence Code	Pubs
GDP binding		
GMP binding		
GTP binding		
LRR domain binding		
protein binding		
protein complex binding		

Process	Evidence Code	Pubs
Fc-epsilon receptor signaling pathway	TAS	
GTP catabolic process	IEA	
MAPK cascade	TAS	
Ras protein signal transduction	TAS	
actin cytoskeleton organization	IEA	
activation of MAPKK activity	TAS	
axon guidance	TAS	
blood coagulation	TAS	

The 'Gene Ontology' or GO is actually maintained by the EBI so lets switch or link over to UniProt also from the EBI.

Scroll down to **UniProt** link

UniProt will detail much more information for protein coding genes such as this one

www.ncbi.nlm.nih.gov/gene/3845#gene-ontology

genomic X01669.1 CAA25828.1

Items 1 - 25 of 43 < Prev Page 1 of 2 Next >

Protein Accession	Links	
	GenPept Link	UniProtKB Link
P01116.1	GenPept	UniProtKB/Swiss-Prot:P01116

Additional links

Scroll down to Very bottom for **UniProt** link

You are here: NCBI > Genes & Expression > Gene

Write to the Help Desk

GETTING STARTED
NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials

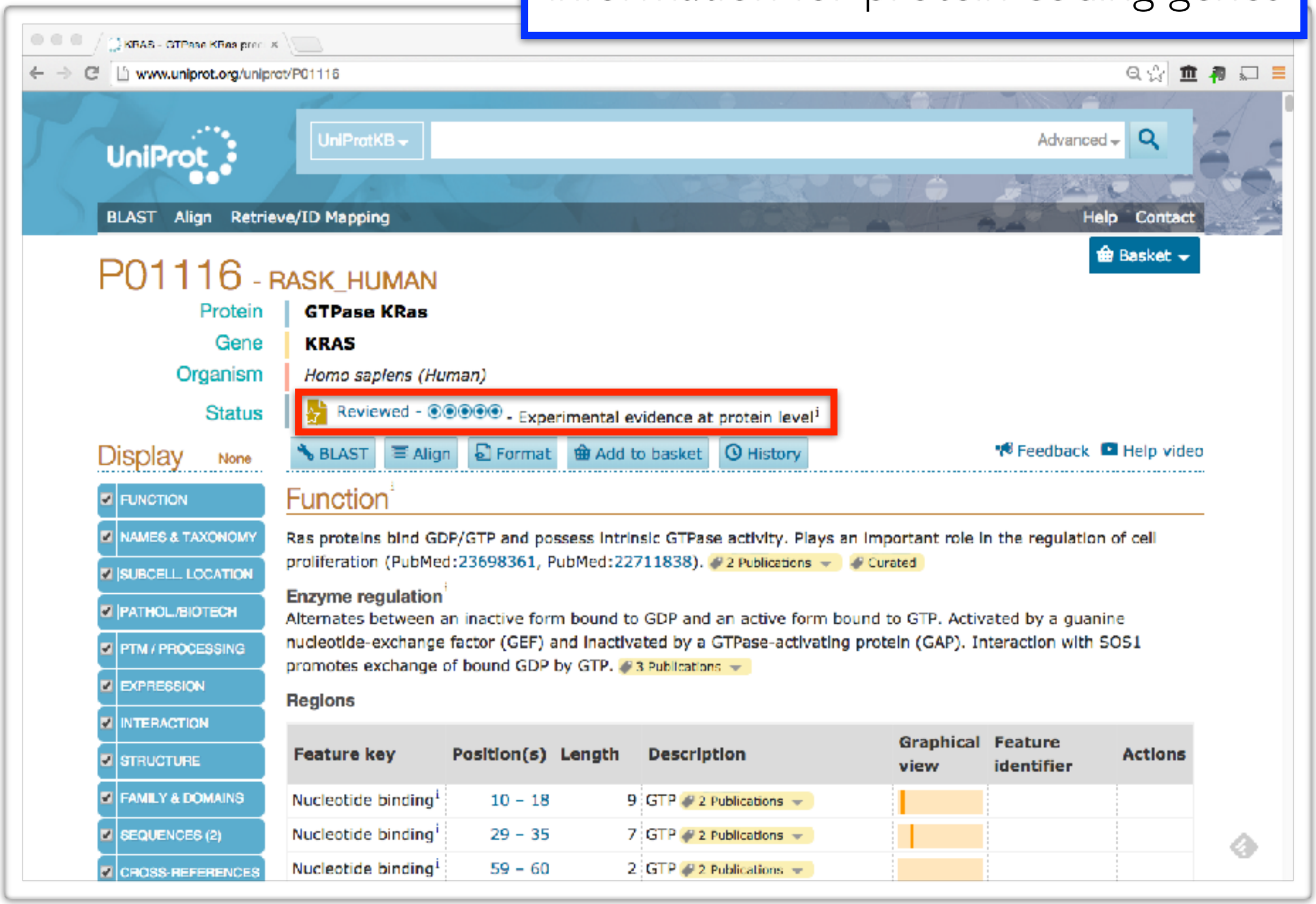
RESOURCES
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy

POPULAR
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

FEATURED
Genetic Testing Registry
PubMed Health
GenBank
Reference Sequences
Gene Expression Omnibus
Map Viewer
Human Genome
Mouse Genome
Influenza Virus
Primer-BLAST
Sequence Read Archive

NCBI INFORMATION
About NCBI
Research at NCBI
NCBI News
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter
NCBI on YouTube

UniProt will detail much more information for protein coding genes



UniProt will detail much more information for protein coding genes

www.uniprot.org/uniprot/P01116

UniProtKB | Advanced | BLAST | Align | Retrieve/ID Mapping | Help | Contact | Basket

P01116 - RASK_HUMAN

Protein	**GTPase KRas**
Gene	**KRAS**
Organism	*Homo sapiens (Human)*
Status	Reviewed - ◉◉◉◉◉ - Experimental evidence at protein level[i]

BLAST | Align | Format | Help video

View FASTA file format

```
>sp|P01116|RASK_HUMAN GTPase KRas OS=Homo sapiens GN=KRAS PE=1 SV=1
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDL
PSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYTLVREIRQYRLKKISKEEKTPGC
VKIKKCIIM
```

Display None

- FUNCTION
- NAMES & TAXONOMY
- SUBCELL. LOCATION
- PATHOL./BIOTECH
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (2)
- CROSS-REFERENCES

Function

Ras proteins bind GDP/GTP and po
proliferation (PubMed:23698361,

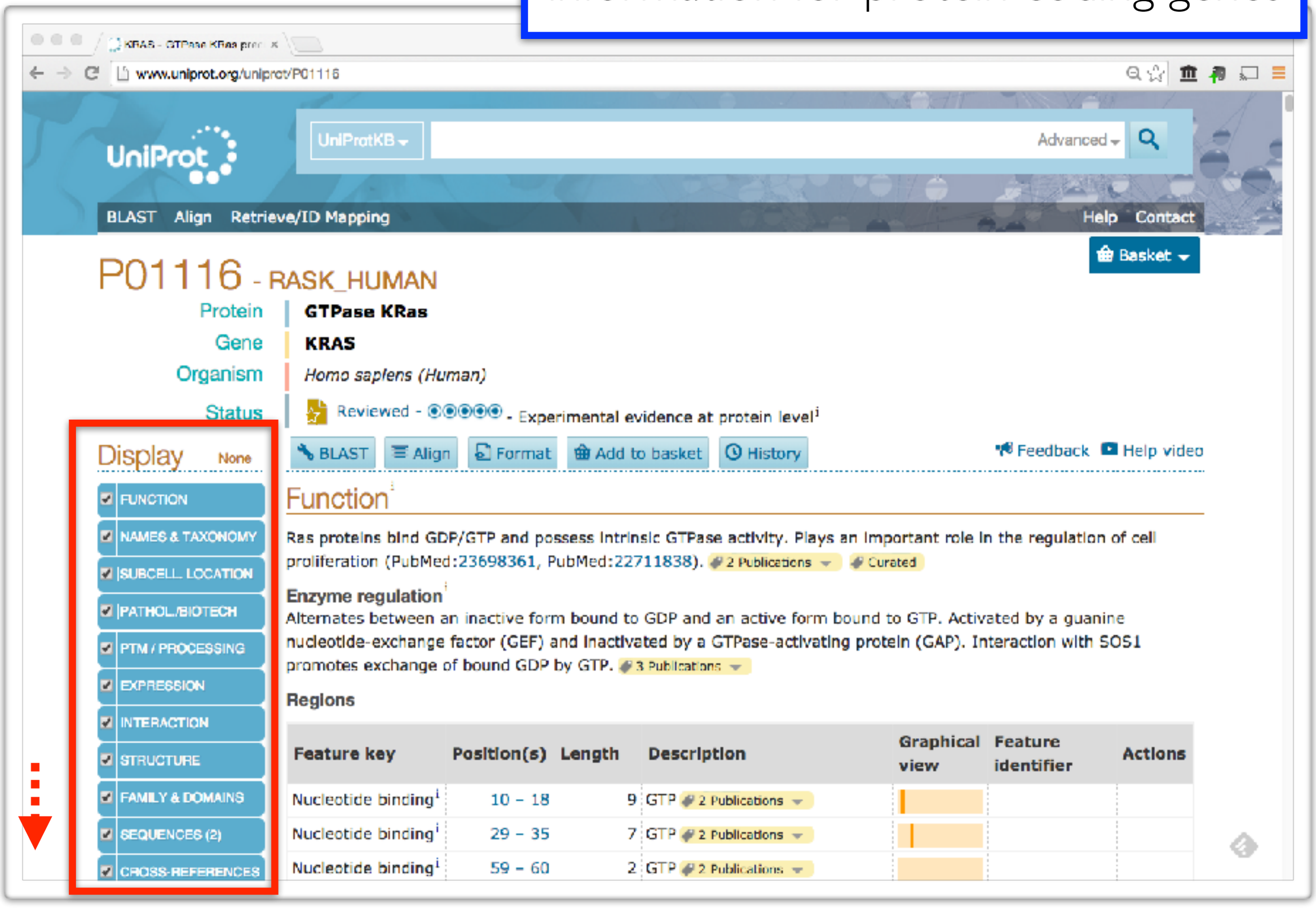
Enzyme regulation

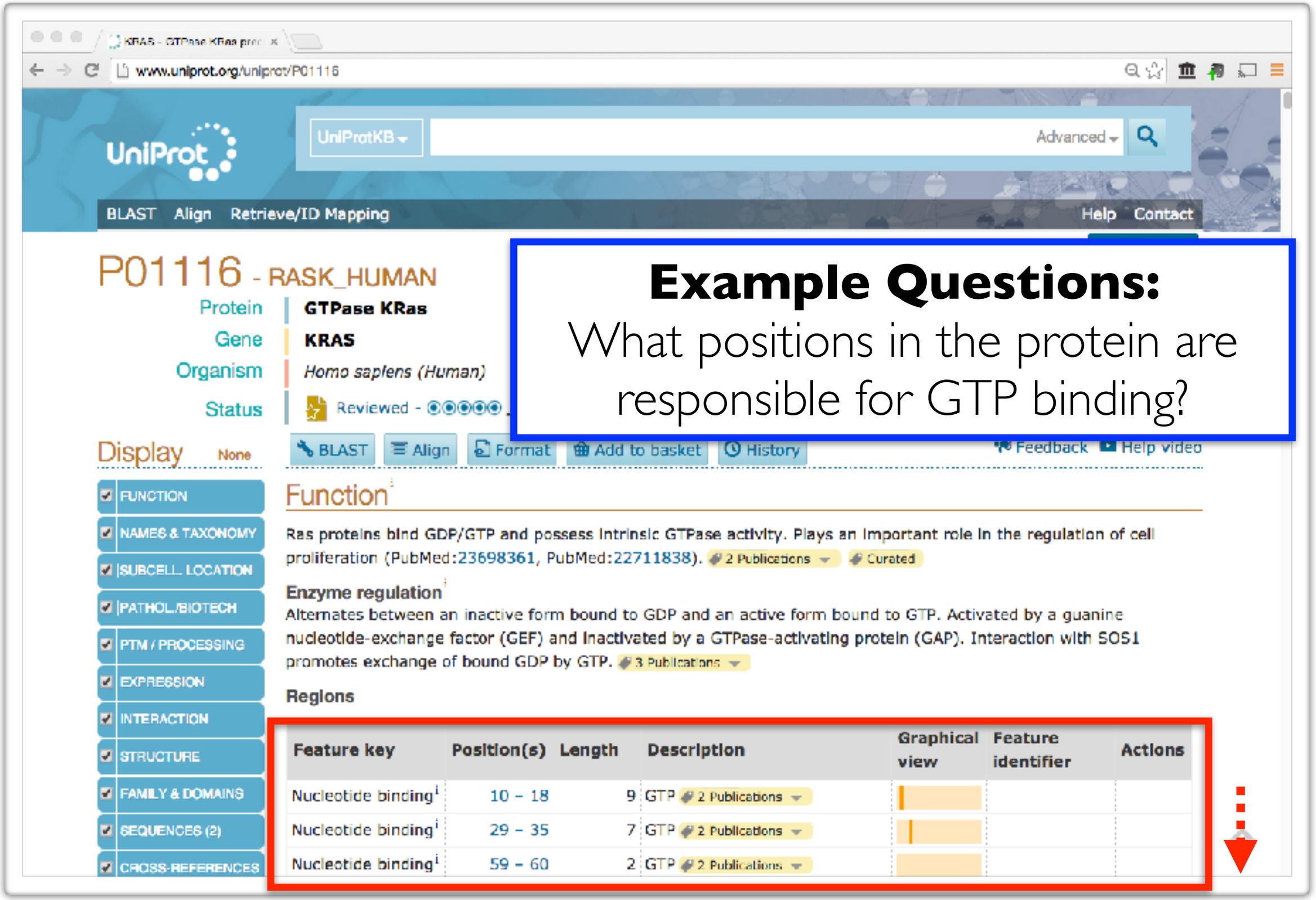
Alternates between an inactive for
nucleotide-exchange factor (GEF)
promotes exchange of bound GDP by GTP. 3 Publications

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding[i]	10 – 18	9	GTP 2 Publications			
Nucleotide binding[i]	29 – 35	7	GTP 2 Publications			
Nucleotide binding[i]	59 – 60	2	GTP 2 Publications			

UniProt will detail much more information for protein coding genes



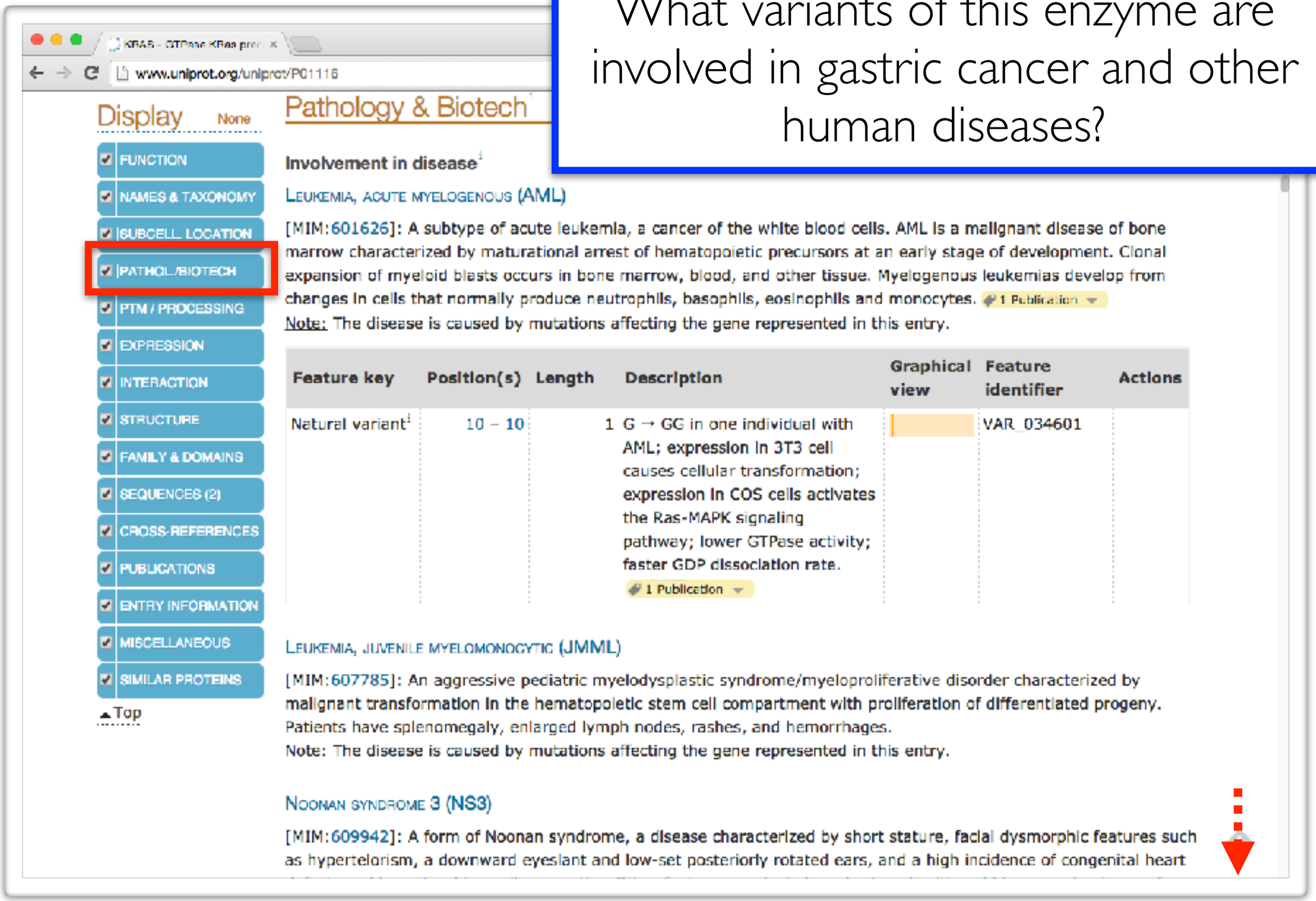


Example Questions:

What positions in the protein are responsible for GTP binding?

Example Questions:

What variants of this enzyme are involved in gastric cancer and other human diseases?



Example Questions:

Are high resolution protein structures available to examine the details of these mutations?

www.uniprot.org/uniprot/P01116

Display None

- FUNCTION
- NAMES & TAXONOMY
- SUBCELL. LOCATION
- PATHOL./BIOTECH
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (2)
- CROSS-REFERENCES
- PUBLICATIONS
- ENTRY INFORMATION
- MISCELLANEOUS
- SIMILAR PROTEINS

Top

Structure

Secondary structure

1 ... 189

Legend: Helix Turn Beta strand

Show more details

3D structure databases

Select the link destinations:

- PDBe
- RCSB PDB
- PDBj

Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
1D8D	X-ray	2.00	P	178-188	[»]
1D8E	X-ray	3.00	P	178-188	[»]
1KZO	X-ray	2.20	C	169-173	[»]
1KZP	X-ray	2.10	C	169-173	[»]
3GFT	X-ray	2.27	A/B/C/D/E/F	1-164	[»]
4DSN	X-ray	2.03	A	2-164	[»]
4DSO	X-ray	1.85	A	2-164	[»]
4EPR	X-ray	2.00	A	1-164	[»]
4EPT	X-ray	2.00	A	1-164	[»]
4EPV	X-ray	1.35	A	1-164	[»]
4EPW	X-ray	1.70	A	1-1	
4EPX	X-ray	1.76	A	1-1	
4EPY	X-ray	1.80	A	1-1	
4L8G	X-ray	1.52	A	1-1	
4LDJ	X-ray	1.15	A	1-164	[»]
4LPK	X-ray	1.50	A/B	1-169	[»]

Open link in a new tab!

Lets view the 3D structure:
Can we find where in the structure our mutations are located and infer their potential molecular effects?

RCSB PDB Deposit Search Visualize Analyze

RCSB PDB PROTEIN DATA BANK An Information Portal to 133759 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligai Go

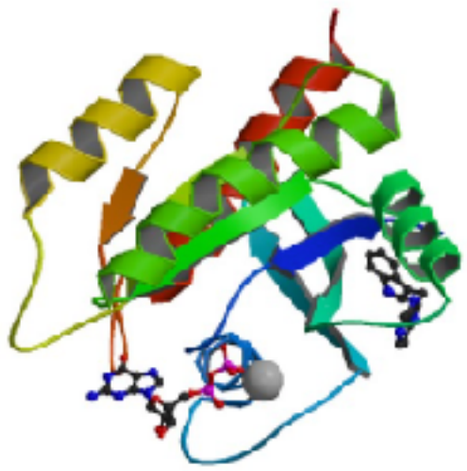
Advanced Search | Browse by Annotations

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure | ...ment | Literature

View PDB file format

Display Files | Download Files

Biological Assembly 1



View in 3D: NGL or JSmol (in Browser)

4EPV

Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation

DOI: 10.2210/pdb4epv/pdb

Classification: HYDROLASE
Deposited: 2012-04-17 **Released:** 2012-05-23
Deposition author(s): Sun, Q., Burke, J.P., Phan, J., Burns, M.C., Olejniczak, E.T., Waterson, A.G., Lee, T., Rossanese, O.W., Fesik, S.W.
Organism: Homo sapiens
Expression System: Escherichia coli
Mutation(s): 1

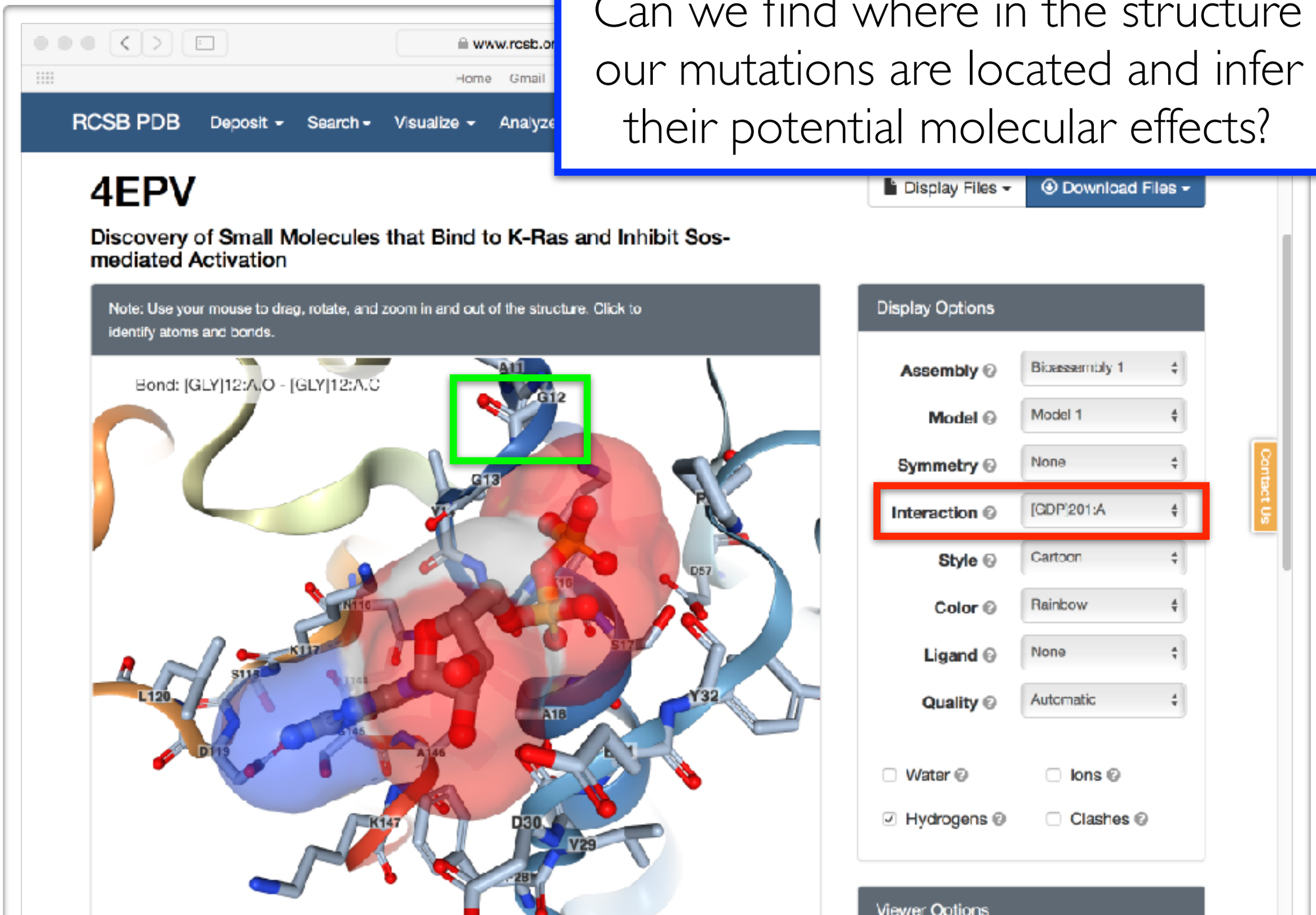
Experimental Data Snapshot **wwPDB Validation** 3D Report | Full Report

Method: X-RAY DIFFRACTION

Metric | Percentile Ranks | Value

Lets view the 3D structure:

Can we find where in the structure our mutations are located and infer their potential molecular effects?



Back to UniProt:

What is known about the protein family, its species distribution, number in humans and residue-wise conservation, etc… ?

www.uniprot.org/uniprot/P01116

Display None

- FUNCTION
- NAMES & TAXONOMY
- SUBCELL. LOCATION
- PATHOL./BIOTECH
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (2)
- CROSS-REFERENCES
- PUBLICATIONS
- ENTRY INFORMATION
- MISCELLANEOUS
- SIMILAR PROTEINS

▲Top

PhylomeDB[i] P01

TreeFam[i] TF3

Family and domain databases

Gene3D[i]	3.40.50.300. 1 hit.
InterPro[i]	IPR027417. P-loop_NTPase. IPR005225. Small_GTP-bd_dom. IPR001806. Small_GTPase. IPR020849. Small_GTPase_Ras. [Graphical view]
PANTHER[i]	PTHR24070. PTHR24070. 1 hit.
Pfam[i]	PF00071. Ras. 1 hit. [Graphical view]
PRINTS[i]	PR00449. RASTRNSFRMNG.
SMART[i]	SM00173. RAS. 1 hit. [Graphical view]
SUPFAM[i]	SSF52540. SSF52540. 1 hit.
TIGRFAMs[i]	TIGR00231. small_GTP. 1 hit.
PROSITE[i]	PS51421. RAS. 1 hit. [Graphical view]

PFAM is one of the best protein family databases

Sequences (2)[i]

Sequence status[i]: Complete.

Sequence processing[i]: The displayed sequence is further processed into a mature form.

This entry describes **2** isoforms[i] produced by **alternative splicing**. Align

Example Questions:

What is known about the protein family, its **species distribution**, number in humans and residue-wise conservation, etc… ?

pfam.xfam.org/family/PF00071

EMBL-EBI HOME

Family: *Ras* (PF00071)

332 architectures | 21243 sequences | 30 interactions | 1006 species | 663 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to... enter ID/acc Go

Summary: Ras family

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: Ras subfamily | Wikipedia: Ras superfamily | Pfam | InterPro

This is the Wikipedia entry entitled "Ras subfamily". More...

Ras subfamily Edit Wikipedia article

This article is about p21/Ras protein. For the p21/waf1 protein, see p21.

Ras is the name given to a family of related proteins which is ubiquitously expressed in all cell lineages and organs. All Ras protein family members belong to a class of protein called small GTPase, and are involved in transmitting signals within cells (cellular signal transduction). Ras is the prototypical member of the Ras superfamily of proteins, which are all related in 3D structure and regulate diverse cell behaviours.

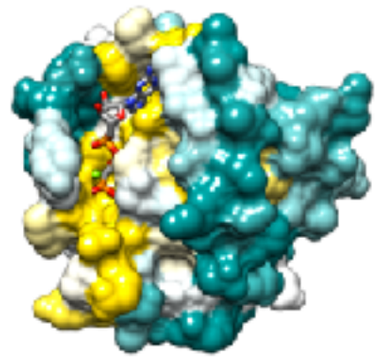
The name 'Ras' is an abbreviation of 'Rat sarcoma', reflecting the way the first members of the protein family were discovered. The name ras is also used to refer to the family of genes encoding those proteins.

When Ras is 'switched on' by incoming signals, it subsequently switches on other proteins, which ultimately turn on genes involved in cell growth, differentiation and survival. As a result, mutations in ras genes can lead to the production of permanently activated Ras proteins. This can cause unintended and overactive signalling inside the cell, even in the absence of incoming signals.

Because these signals result in cell growth and division, overactive Ras signaling can ultimately lead to cancer.[1] The 3 Ras genes in humans (HRAS, KRAS, and NRAS) are the most common oncogenes in human cancer; mutations that permanently activate Ras are found in 20% to 25% of all human tumors and up to 90% in certain types of cancer (e.g., pancreatic cancer).[2] For this reason, Ras inhibitors are being studied as a treatment for cancer, and other diseases with Ras overexpression.

Contents [hide]

1 History
2 Structure
3 Function
  3.1 Activation and deactivation
  3.2 Membrane attachment
4 Members
5 Ras in cancer
  5.1 Inappropriate activation
  5.2 Constitutively active Ras

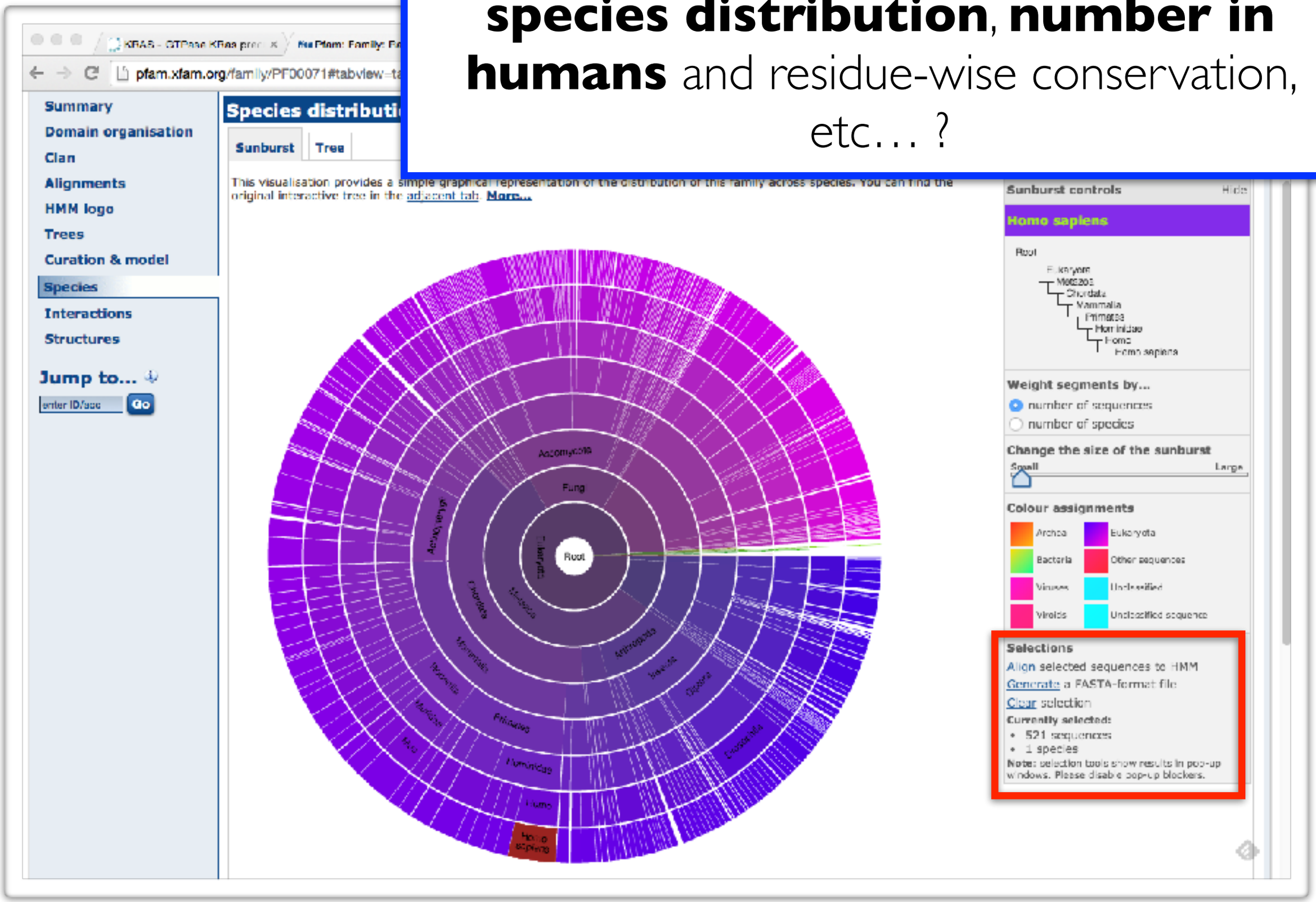


H-Ras structure PDB 121p, surface colored by conservation in Pfam seed alignment: gold, most conserved; dark cyan, least conserved.

Identifiers	
Symbol	Ras
Pfam	PF00071
InterPro	IPR013753
PROSITE	PDOC00017
SCOP	5p21
SUPERFAMILY	5p21

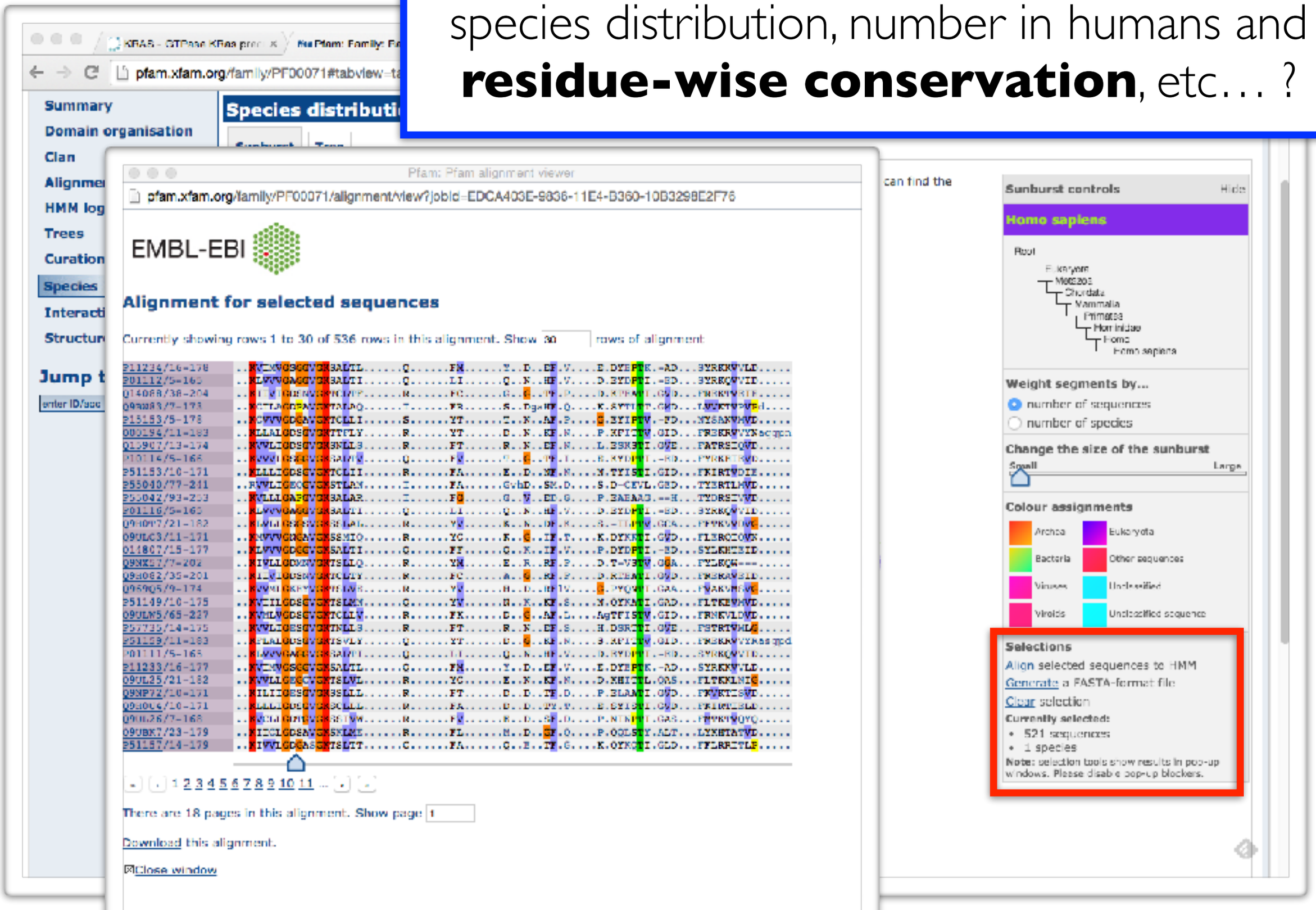
Example Questions:

What is known about the protein family, its **species distribution**, **number in humans** and residue-wise conservation, etc… ?



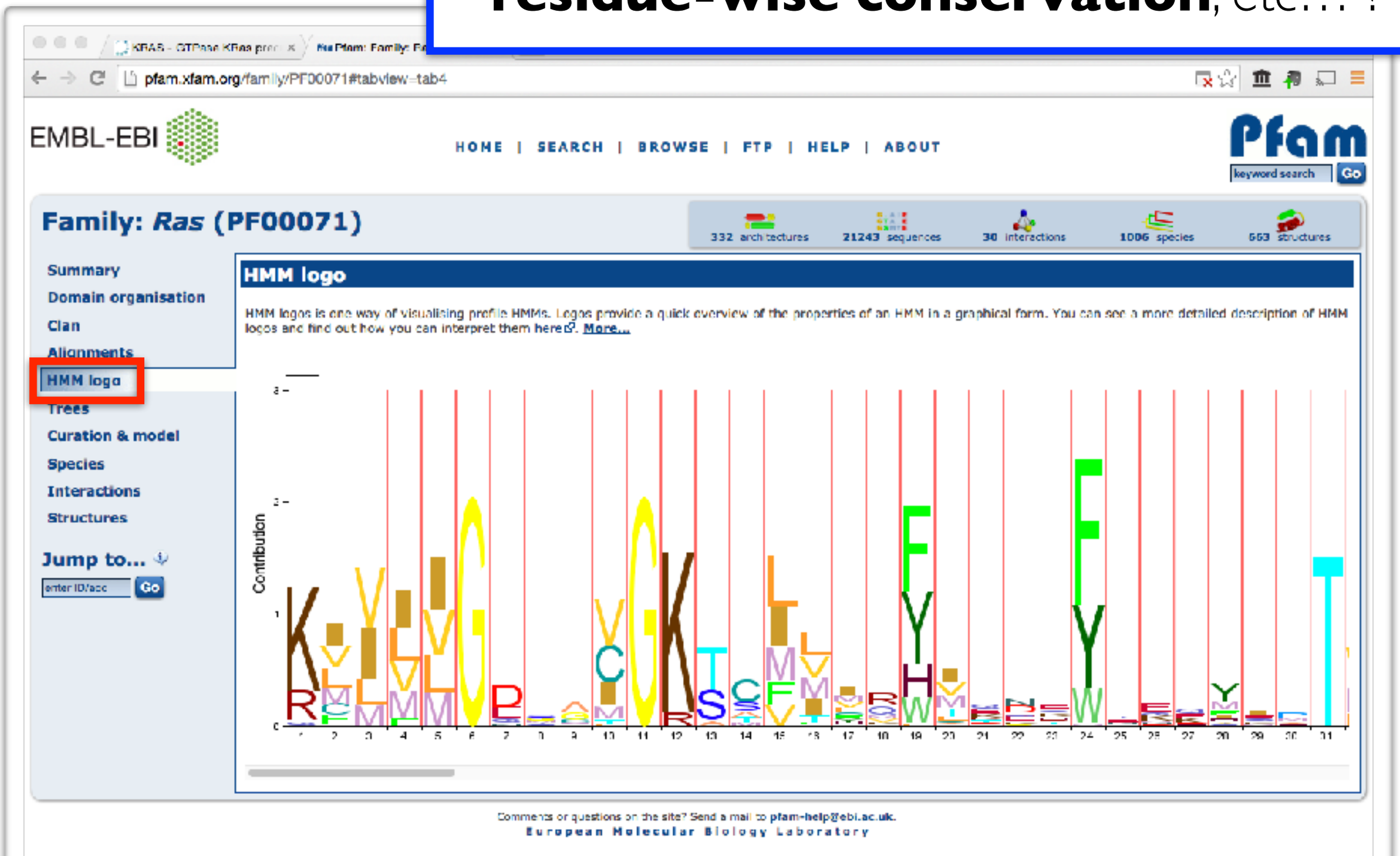
Example Questions:

What is known about the protein family, its species distribution, number in humans and **residue-wise conservation**, etc… ?



Example Questions:

What is known about the protein family, its species distribution, number in humans and **residue-wise conservation**, etc… ?



Family: *Kinesin* (PF00225)

Loading page components (1 remaining)...

126 architectures 4150 sequences 6 interactions 248 species 114 structures

Summary
Domain organisation
Clans
Alignments
HMM logo
Trees
Curation & models
Species
Interactions
Structures

Jump to...
enter ID/acc Go

Interactions

There are **6** interactions for this family. **More...**

Tubulin
Tubulin_C
Tubulin_C
Kinesin
Tubulin
Kinesin

Family: *Kinesin* (PF00225)

126 architectures 4150 sequences 6 interactions 248 species 114 structures

Summary
Domain organisation
Clans
Alignments
HMM logo
Trees
Curation & models
Species
Interactions
Structures

Jump to... enter ID/acc Go

Structures

For those sequences which have a structure in the Protein DataBank, we use the mapping between UniProt, PDB and Pfam coordinate systems from the PDBe group, to allow us to map Pfam domains onto UniProt sequences and three-dimensional protein structures. The table below shows the structures on which the **Kinesin** domain has been found.

UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View
A8BKD1_GIALA	11 - 335	2vvg	A	11 - 335	Jmol AstexViewer SPICE
			B	11 - 335	Jmol AstexViewer SPICE
CENPE_HUMAN	12 - 329	1t5c	A	12 - 329	Jmol AstexViewer SPICE
			B	12 - 329	Jmol AstexViewer SPICE
KAR3_YEAST	392 - 723	1f9t	A	392 - 723	Jmol AstexViewer SPICE
		1f9u	A	392 - 723	Jmol AstexViewer SPICE
		1f9v	A	392 - 723	Jmol AstexViewer SPICE
		1f9w	A	392 - 723	Jmol AstexViewer SPICE
			B	392 - 723	Jmol AstexViewer SPICE
		3kar	A	392 - 723	Jmol AstexViewer SPICE
KI13B_HUMAN	11 - 352	3gbj	A	11 - 352	Jmol AstexViewer SPICE
			B	11 - 352	Jmol AstexViewer SPICE
			C	11 - 352	Jmol AstexViewer SPICE
		1ii6	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		1q0b	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		1x88	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
			A	24 - 359	Jmol AstexViewer SPICE

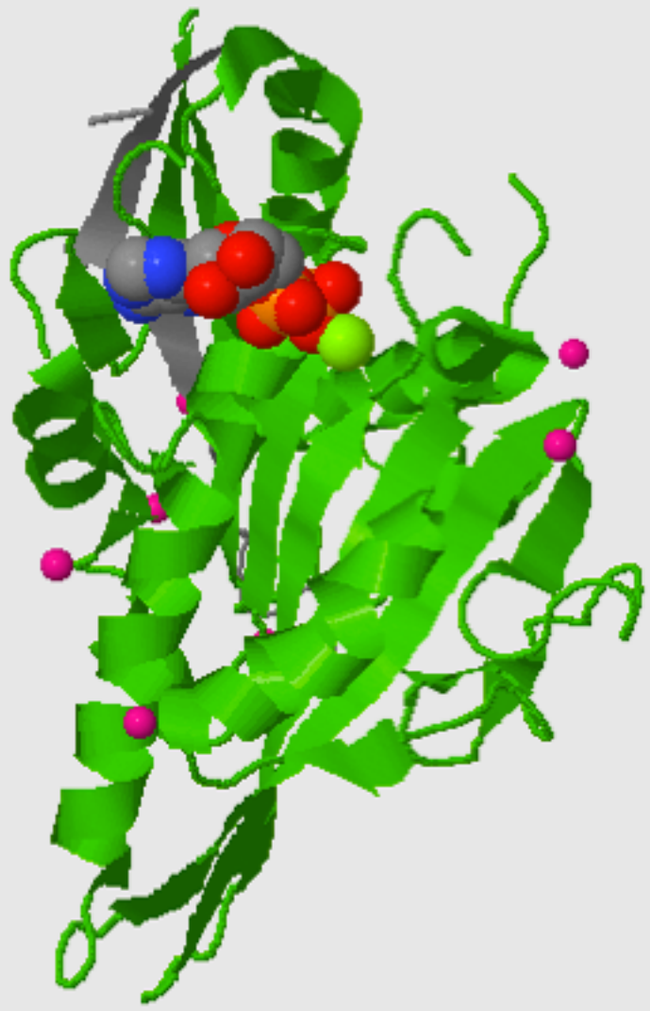
Pfam: Jmol

http://pfam.janelia.org/structure/viewer?viewer=jmol&id=3bfn

Pfam: Family: Kinesin (PF00225) | Pfam: Jmol

wellcome trust sanger institute

PDB entry 3bfn



Jmol

PDB			UniProt			Pfam family	Colour
Chain	Start	End	ID	Start	End		
A	49	368	KIF22_HUMAN	49	368	Kinesin (PF00225)	

☒Close window

Your turn:

What can you find out about "eg5"

Today's Menu

Classifying Databases	Primary, secondary and composite Bioinformatics databases
Using Databases	**Vignette** demonstrating how major Bioinformatics databases intersect
Major Biomolecular Formats	How nucleotide and protein sequence and structure data are represented
Alignment Foundations	**Introducing the *why* and *how* of comparing sequences**
Alignment Algorithms	**Hands-on** exploration of alignment algorithms and applications

ALIGNMENT FOUNDATIONS

- **Why…**
  - Why compare biological sequences?
- **What…**
  - Alignment view of sequence changes during evolution (matches, mismatches and gaps)
- **How…**
  - Dot matrices
  - Dynamic programing
    - Global alignment
    - Local alignment
  - BLAST heuristic approach

ALIGNMENT FOUNDATIONS

- **Why…**
  - Why compare biological sequences?
- **What…**
  - Alignment view of sequence changes during evolution (matches, mismatches and gaps)
- **How…**
  - Dot matrices
  - Dynamic programing
    - Global alignment
    - Local alignment
  - BLAST heuristic approach

Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.

```
Seq1: C A T T C A C

Seq2: C T C G C A G C
```

[Screencast Material]

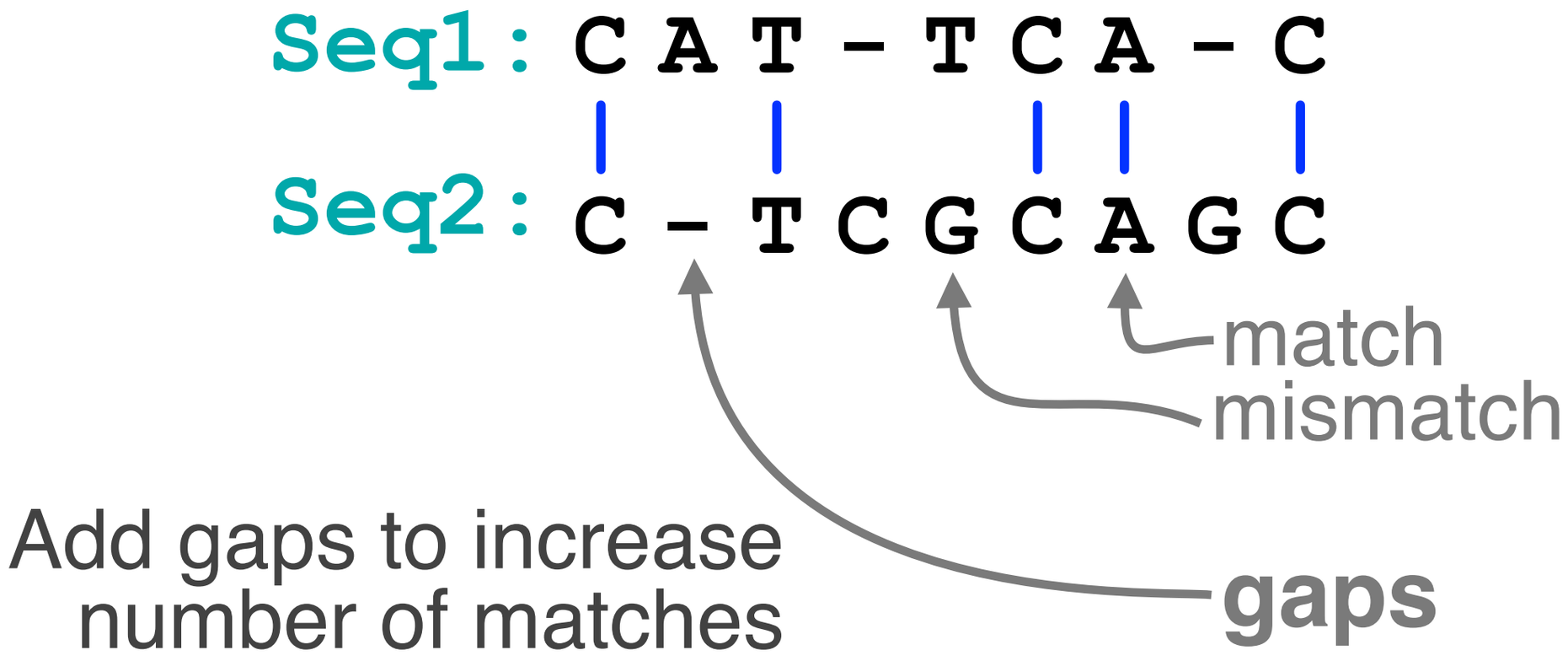
Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.

```
Seq1: C A T T C A C
        |       | |
Seq2: C T C G C A G C
```

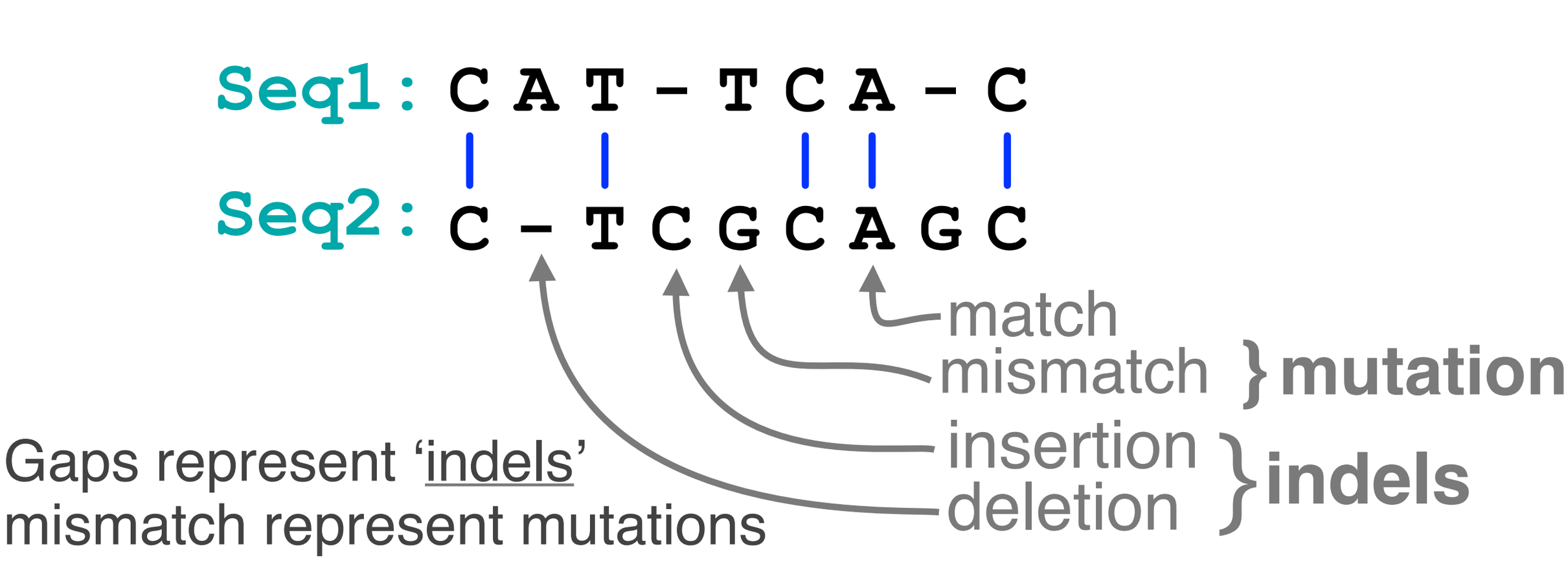
mismatch
match

Two types of character correspondence

Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.



Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.



Why compare biological sequences?

- To obtain **functional or mechanistic insight** about a sequence by inference from another potentially better characterized sequence
- To find whether two (or more) genes or proteins are **evolutionarily related**
- To find **structurally or functionally similar regions** within sequences (e.g. catalytic sites, binding sites for other molecules, etc.)
- Many practical bioinformatics applications…

Practical applications include...

- **Similarity searching of databases**
  – Protein structure prediction, annotation, etc...
- **Assembly of sequence reads** into a longer construct such as a genomic sequence
- **Mapping sequencing reads to a known genome**
  – "Resequencing", looking for differences from reference genome - SNPs, indels (insertions or deletions)
  – Mapping transcription factor binding sites via ChIP-Seq (chromatin immuno-precipitation sequencing)
  – Pretty much all next-gen sequencing data analysis

Practical applications include...

- **Similarity searching of databases**
  - Protein structure prediction
- **Assembly of sequen** ger construct such
- **Mappi** **ds to a known genome**
  - king for differences from reference
    - s, indels (insertions or deletions)
  - ping transcription factor binding sites via ChIP-Seq (chromatin immuno-precipitation sequencing)
  - Pretty much all next-gen sequencing data analysis

N.B. Pairwise sequence alignment is arguably the most fundamental operation of bioinformatics!

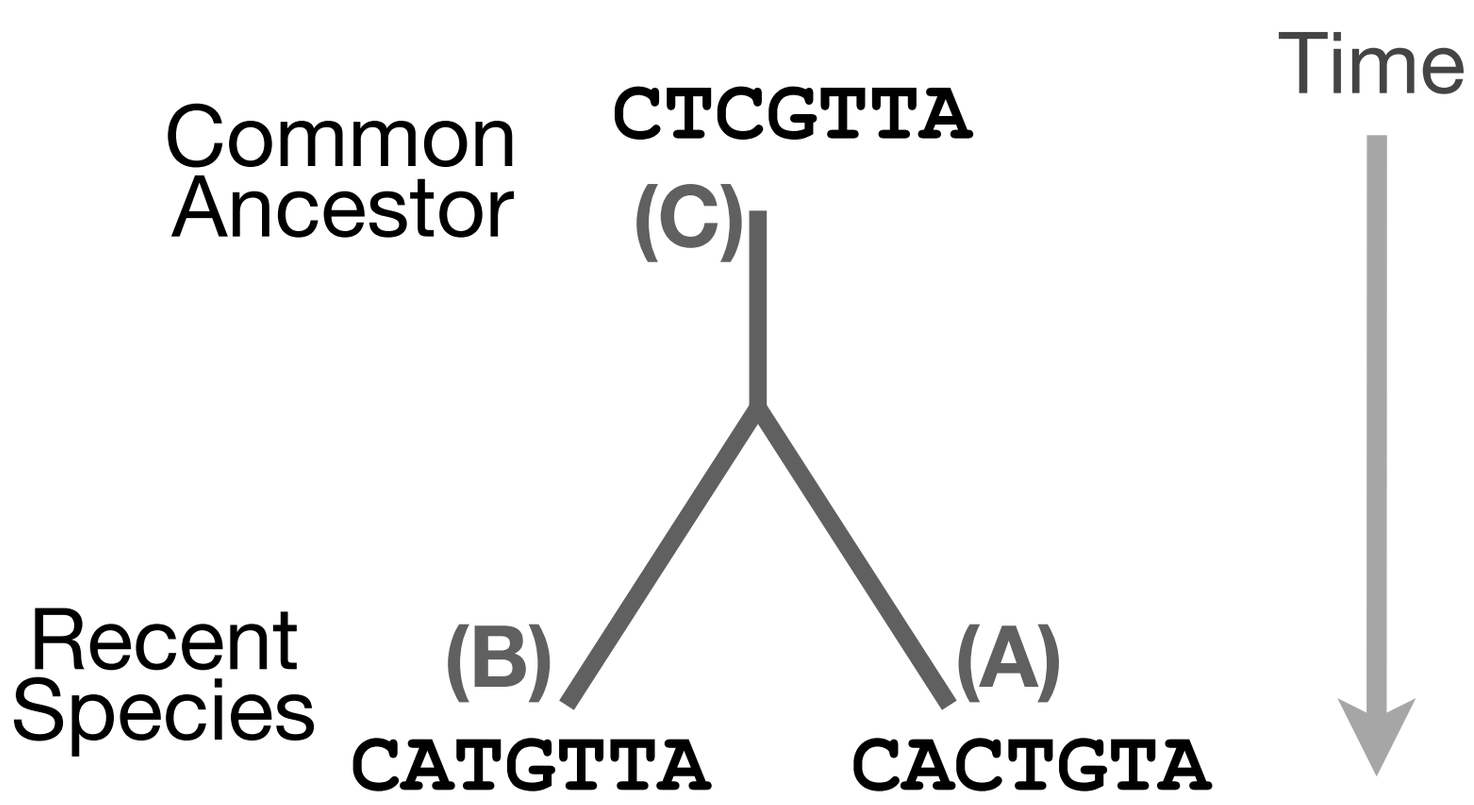
ALIGNMENT FOUNDATIONS

- **Why…**
  - Why compare biological sequences?
- **What…**
  - Alignment view of sequence changes during evolution (matches, mismatches and gaps)
- **How…**
  - Dot matrices
  - Dynamic programing
    - Global alignment
    - Local alignment
  - BLAST heuristic approach

Sequence changes during evolution

There are three major types of sequence change that can occur during evolution.

- Mutations/Substitutions
- Deletions
- Insertions

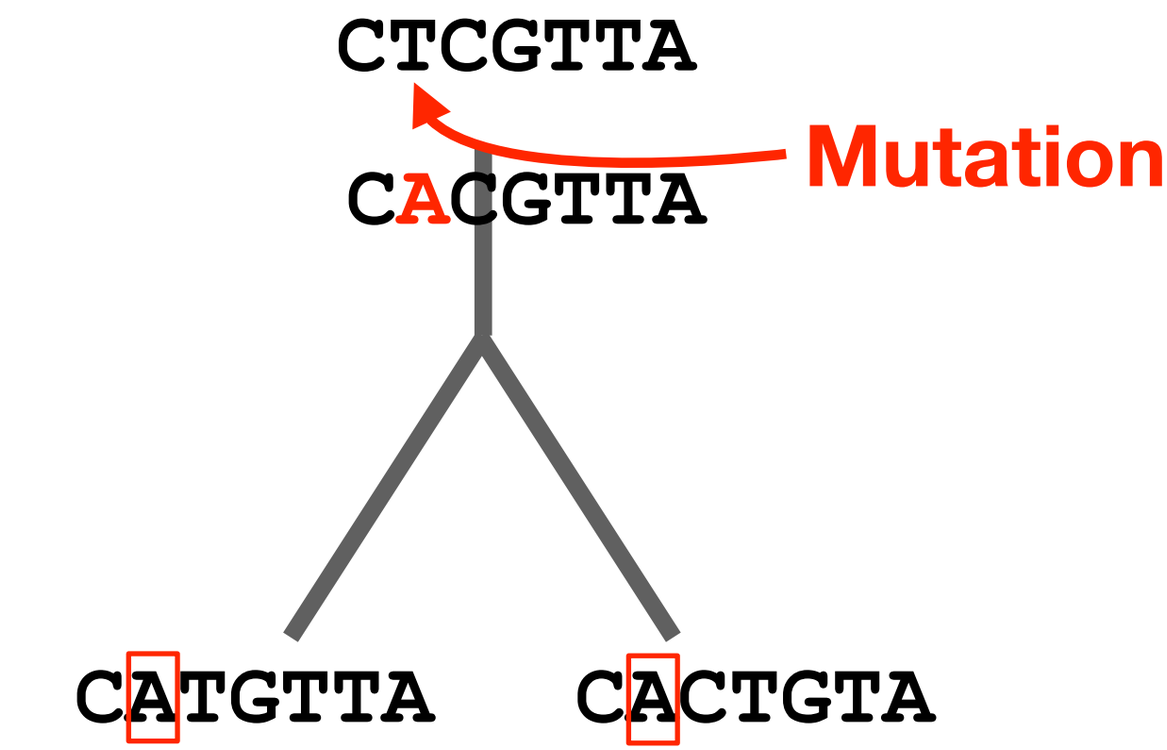


Mutations, deletions and insertions

There are three major types of sequence change that can occur during evolution.

- **Mutations/Substitutions**
- Deletions
- Insertions

CTCGTTA → CACGTTA



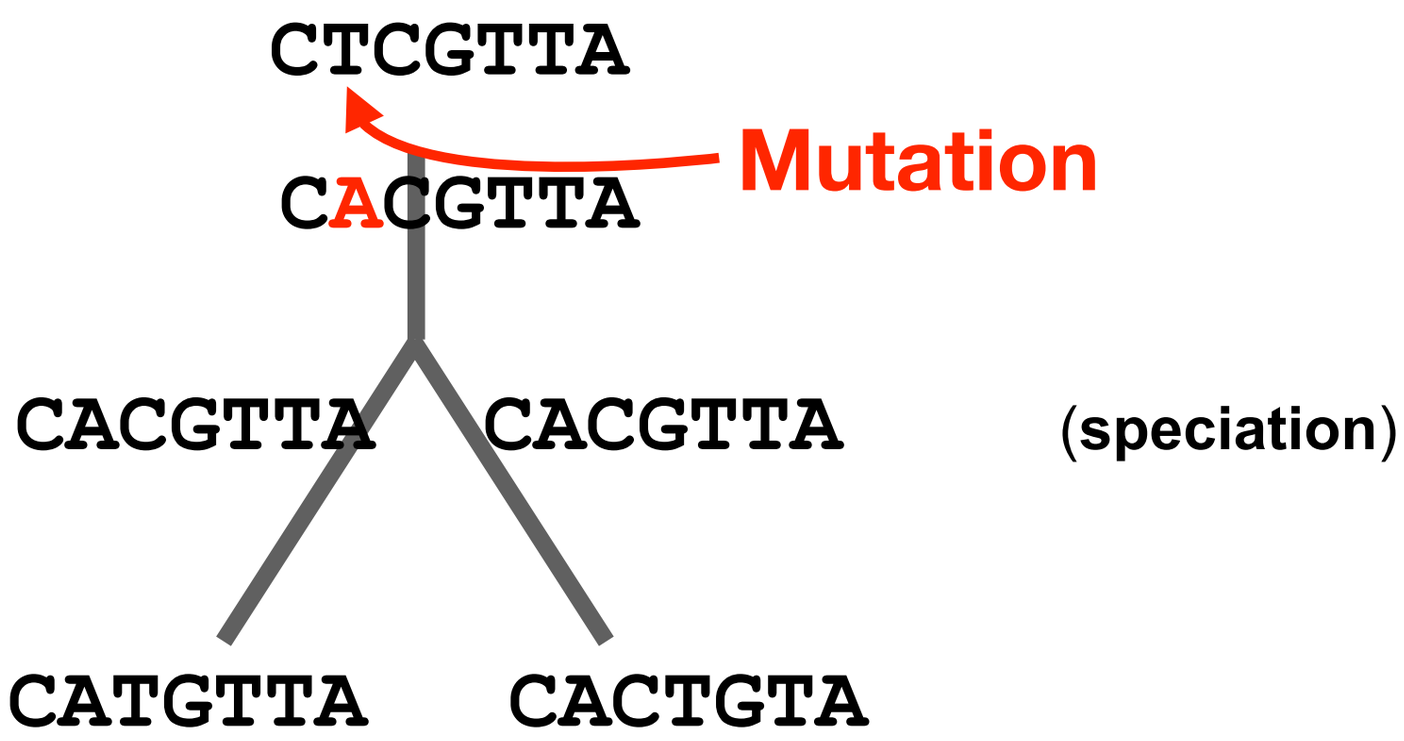
Likely occurred prior to speciation

Mutations, deletions and insertions

There are three major types of sequence change that can occur during evolution.

- Mutations/Substitutions
- Deletions
- Insertions

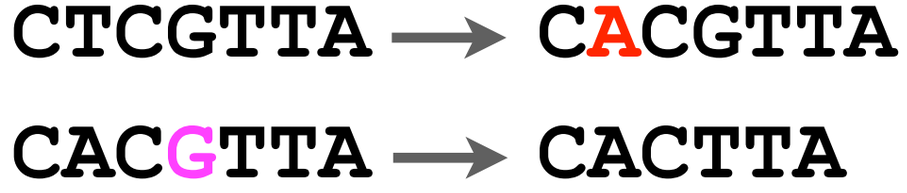
CTCGTTA → CACGTTA

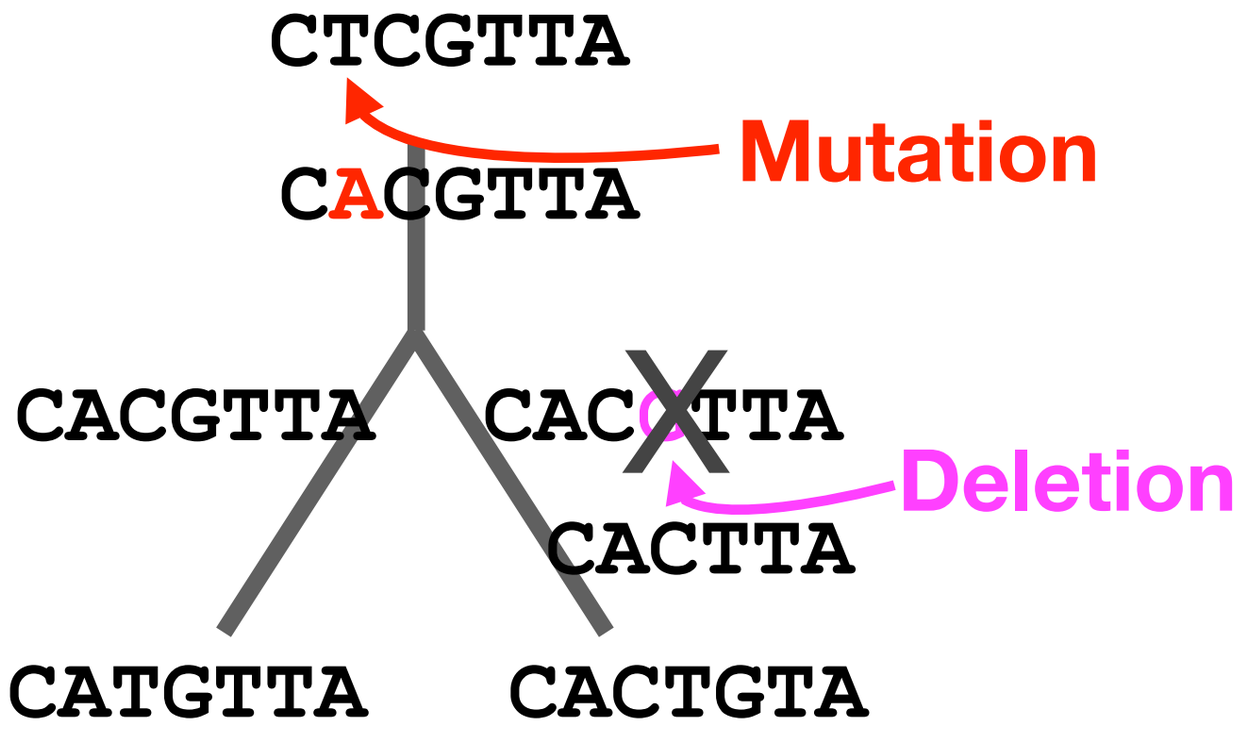


Mutations, deletions and insertions

There are three major types of sequence change that can occur during evolution.

- Mutations/Substitutions
- **Deletions**
- Insertions





Mutations, deletions and insertions

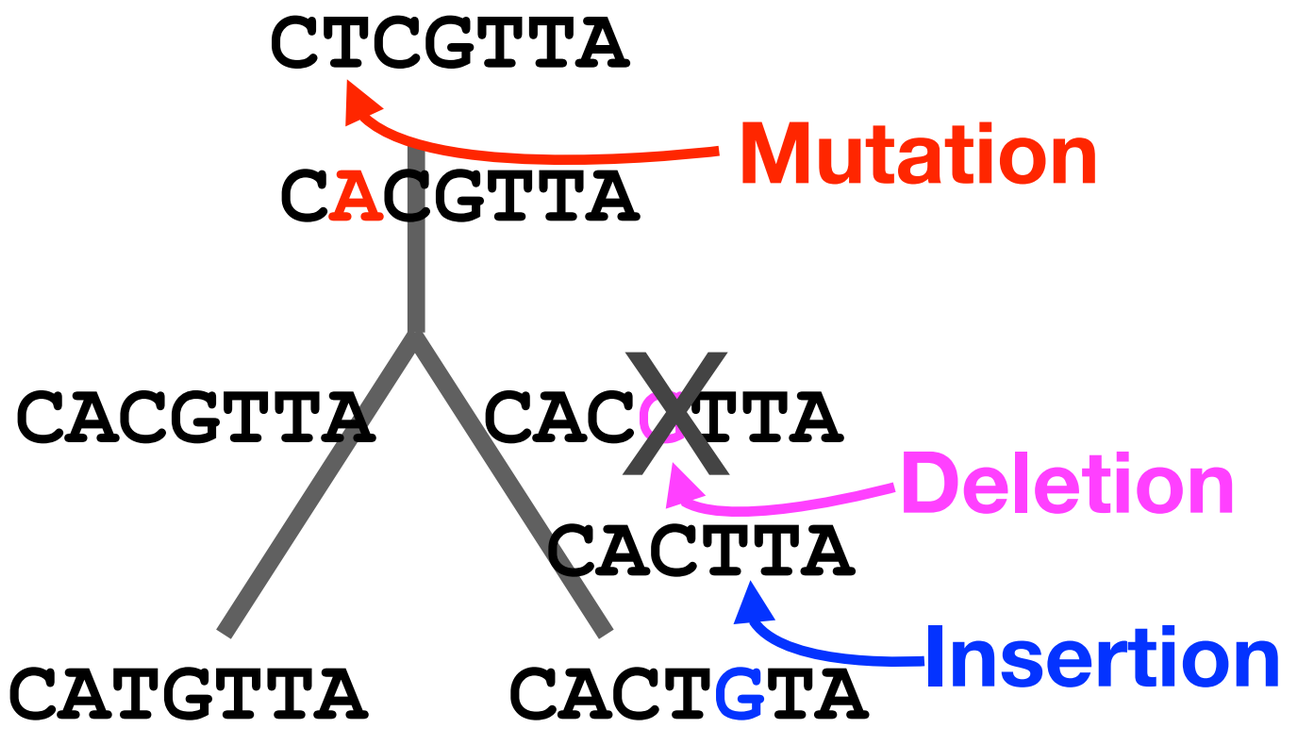
There are three major types of sequence change that can occur during evolution.

- Mutations/Substitutions
- Deletions
- **Insertions**

CTCGTTA → CACGTTA

CACGTTA → CACTTA

CACTTA → CACTGTA



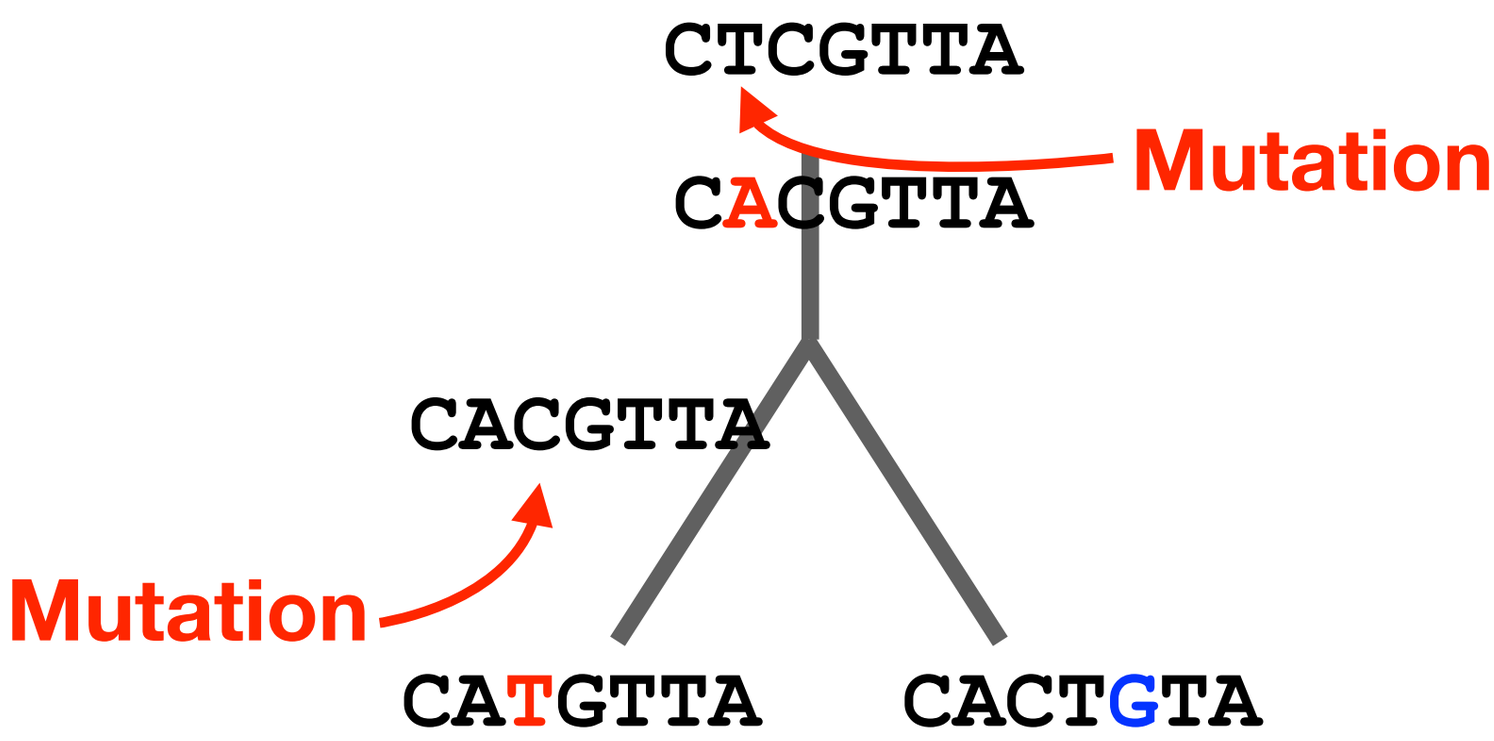
Mutations, deletions and insertions

There are three major types of sequence change that can occur during evolution.

- **Mutations/Substitutions**
- Deletions
- Insertions

CTCGTTA → CACGTTA

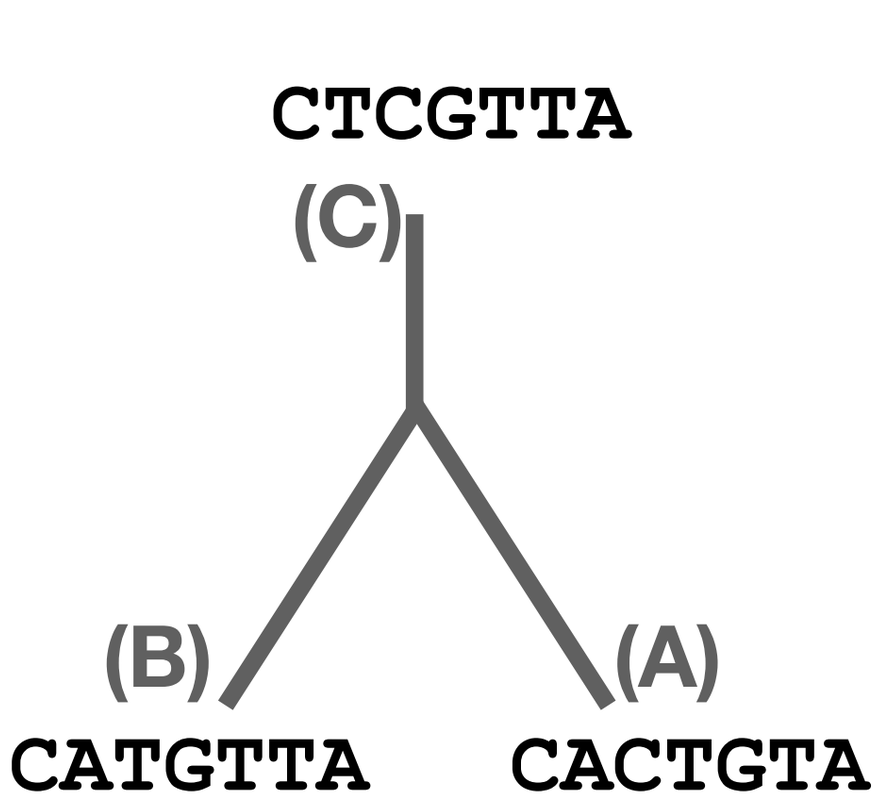
CACGTTA → CATGTTA

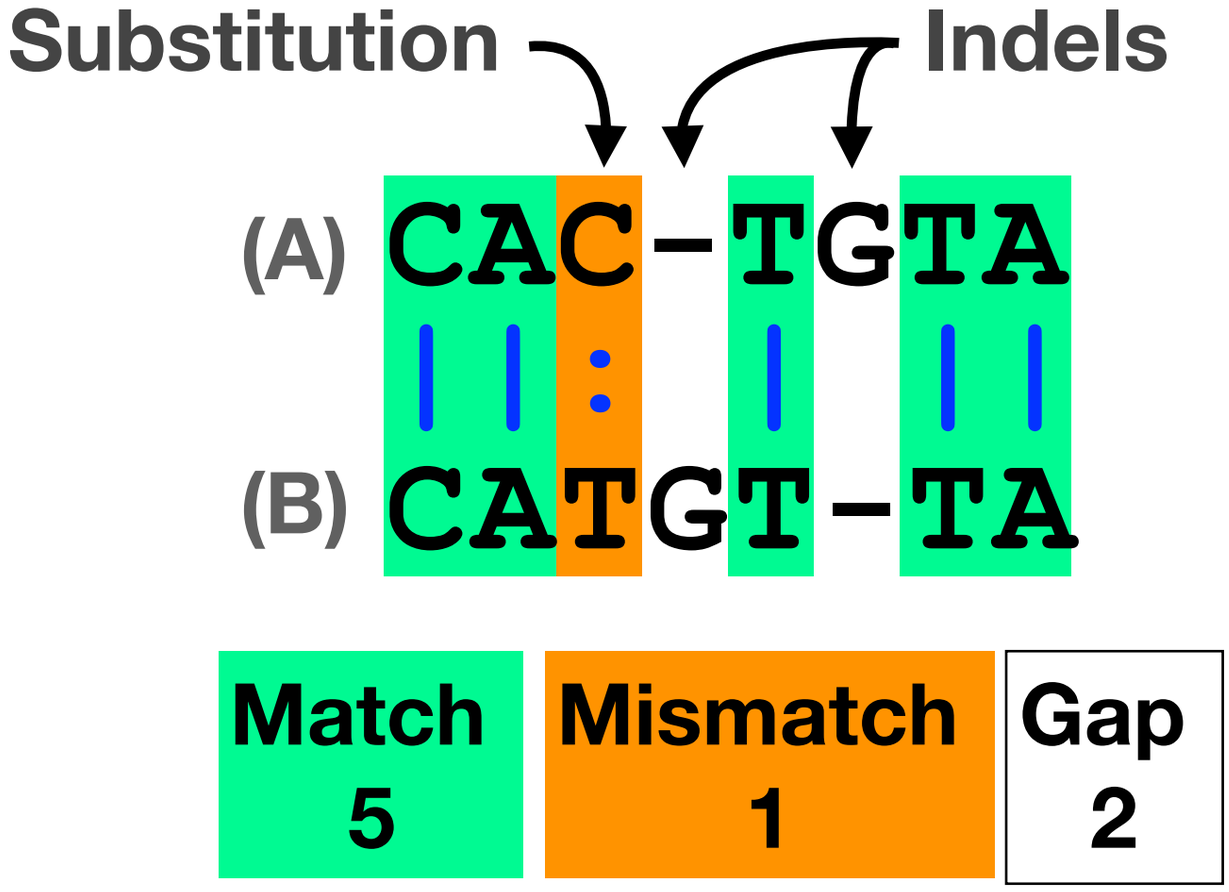


Alignment view

Alignments are great tools to visualize sequence similarity and evolutionary changes in homologous sequences.

- **Mismatches** represent mutations/substitutions
- **Gaps** represent insertions and deletions (indels)





Alternative alignments

- Unfortunately, finding the correct alignment is difficult if we do not know the evolutionary history of the two sequences

Q. Which of these 3 possible alignments is best?

1.

```
CACTGTA
| | : : : | |
CATGTTA
```

2.

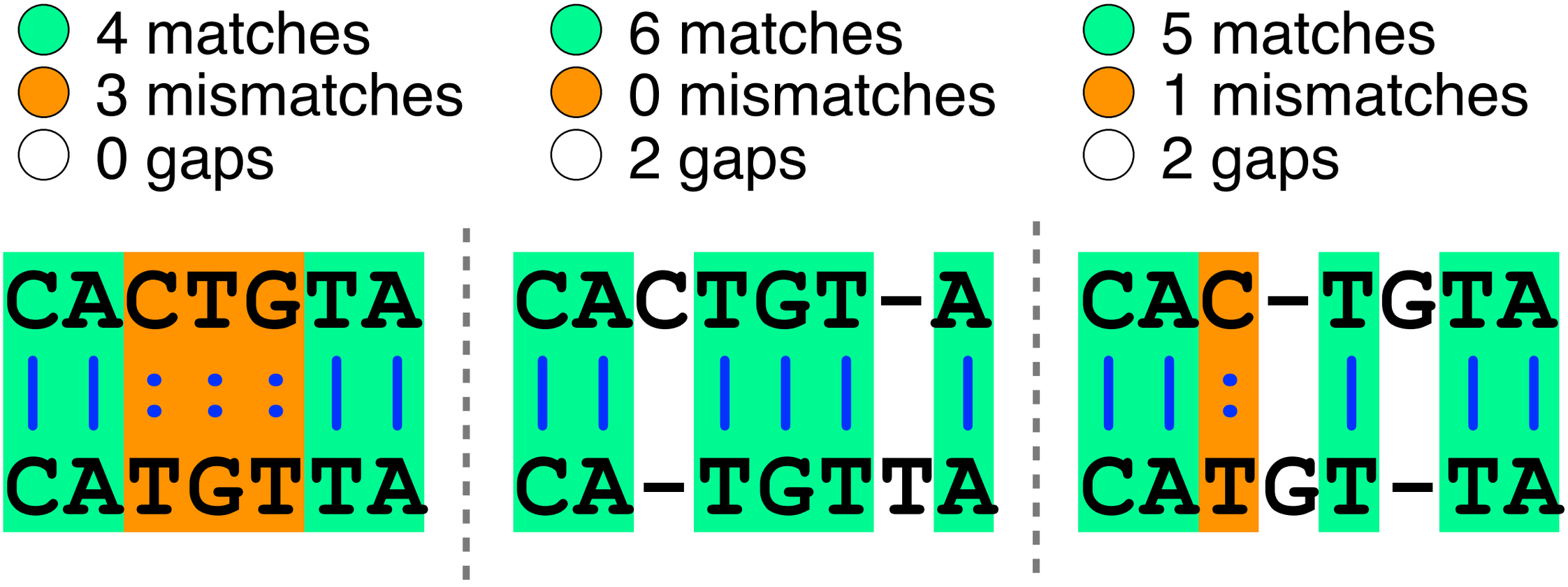
```
CACTGT-A
| |  | | | |
CA-TGTTA
```

3.

```
CAC-TGTA
| | :  |  | |
CATGT-TA
```

Alternative alignments

- One way to judge alignments is to compare their number of matches, insertions, deletions and mutations



Scoring alignments

- We can assign a score for each match (+3), mismatch (+1) and indel (-1) to identify the **optimal alignment** *for this scoring scheme*

4 (+3)
3 (+1)
0 (-1) = 15

```
CACTGTA
| | : : | |
CATGTTA
```

6 (+3)
0 (+1)
2 (-1) = 16

```
CACTGT-A
| |  | | |
CA-TGTTA
```

5 (+3)
1 (+1)
2 (-1) = 14

```
CAC-TGTA
| | :  | |
CATGT-TA
```

Optimal alignments

- Biologists often prefer **parsimonious alignments**, where the number of postulated sequence changes is minimized.

4 matches
3 mismatches
0 gaps

```
CACTGTA
| | : : : | |
CATGTTA
```

6 matches
0 mismatches
2 gaps

```
CACTGT-A
| |  | | |  |
CA-TGTTA
```

5 matches
1 mismatches
2 gaps

```
CAC-TGTA
| | :  |  | |
CATGT-TA
```

Optimal alignments

- Biologists often prefer **parsimonious alignments**, where the number of postulated sequence changes is minimized.

4 matches
3 mismatches
0 gaps

```
CACTGTA
| | : : : | |
CATGTTA
```

6 matches
0 mismatches
2 gaps

```
CACTGT-A
| |  | | |  |
CA-TGTTA
```

5 matches
1 mismatches
2 gaps

```
CAC-TGTA
| | :  |  | |
CATGT-TA
```

Optimal alignments

- Biologists often prefer **parsimonious alignments**, where the number of postulated sequence changes is minimized.

4 matches
3 mismatches
0 gaps

```
CACTGTA
| | : : : | |
CATGTTA
```

6 matches
0 mismatches
2 gaps

```
CACTG-TA
| |  | |  | |
CA-TGTTA
```

5 matches
1 mismatches
2 gaps

```
CAC-TGTA
| | : |  | |
CATGT-TA
```

Optimal alignments

- Biologists often prefer **parsimonious alignments**, where the numb… sequence changes is m…

4 matches
3 mis…
…

Warning: There may be more than one optimal alignment and these may not reflect the true evolutionary history of our sequences!

```
CATGTTA
```

```
CACTGT-A
| |  | | |
CA-TGTTA
```

4 matches
1 mismatches
2 gaps

```
CAC-TGTA
| | : |  |
CATGT-TA
```

ALIGNMENT FOUNDATIONS

- **Why…**
  - Why compare biological sequences?
- **What…**
  - Alignment view of sequence changes during evolution (matches, mismatches and gaps)

- **How…**
  - Dot matrices
  - Dynamic programing
    - Global alignment
    - Local alignment
  - BLAST heuristic approach

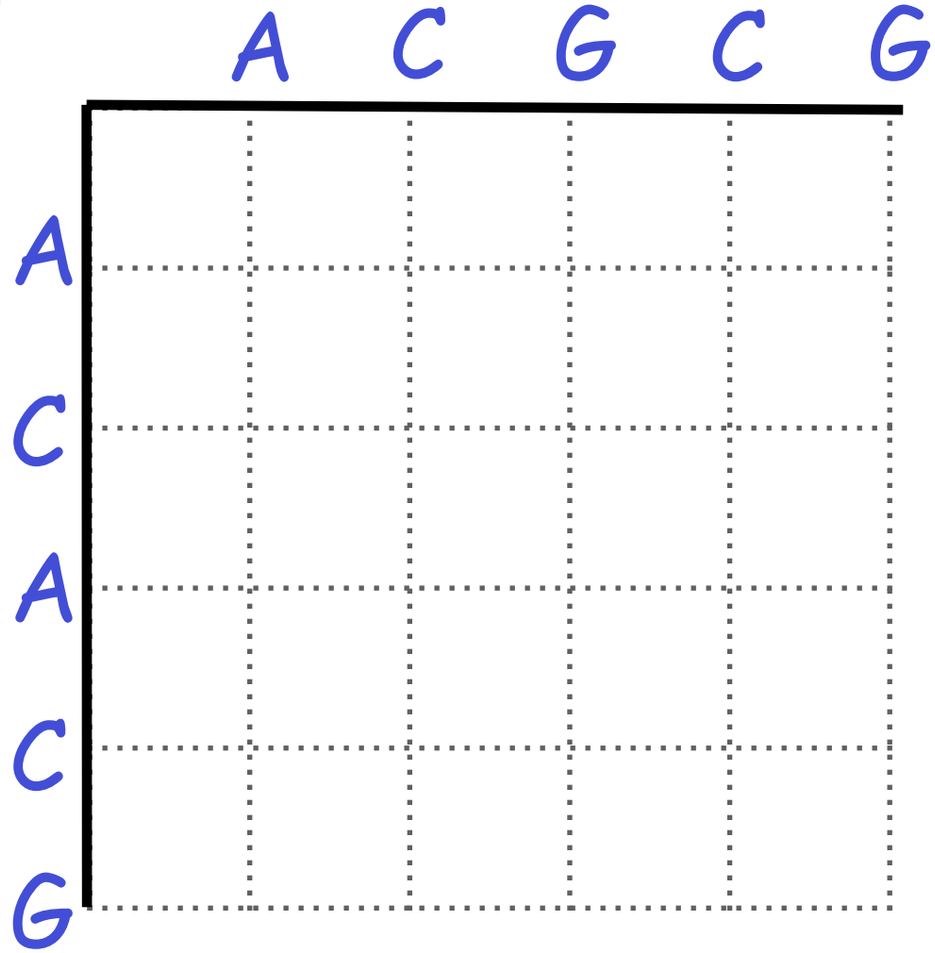
ALIGNMENT FOUNDATIONS

- **Why…**
  - Why compare biological sequences?
- **What…**
  - Alignment view of sequence changes during evolution (matches, mismatches and gaps)
- **How…**
  - Dot matrices
  - D
    -
    -
  - BLAST heuristic approach

How do we compute the optimal alignment between two sequences?

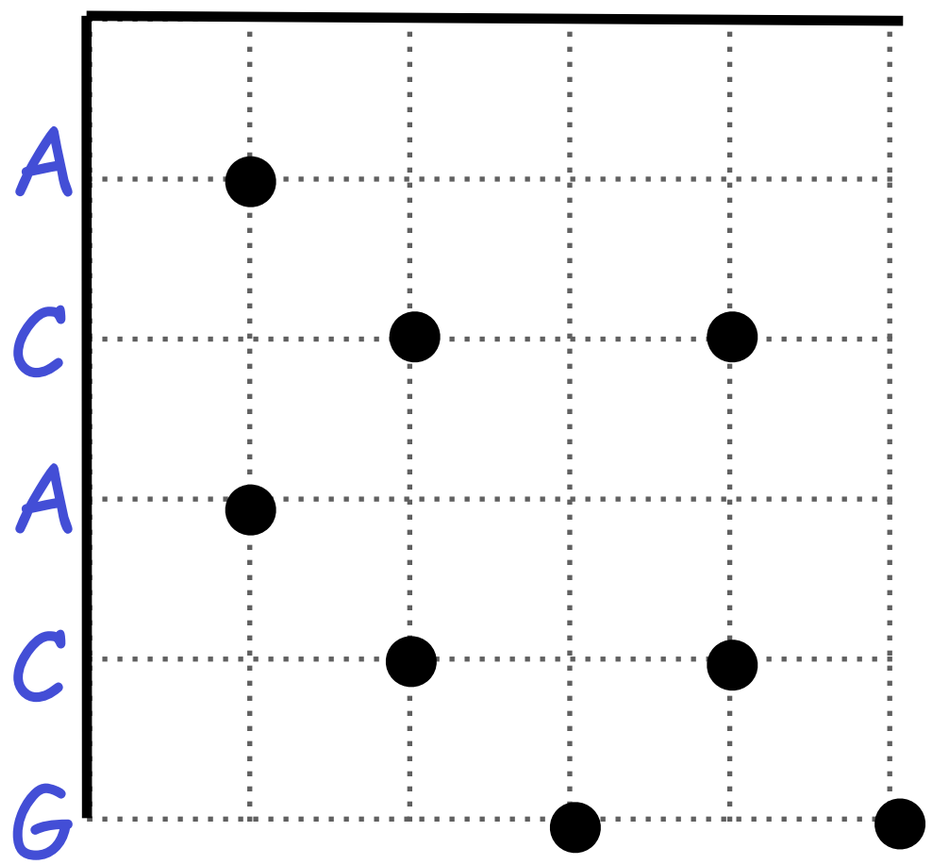
Dot plots: simple graphical approach

- Place one sequence on the vertical axis of a 2D grid (or matrix) and the other on the horizontal



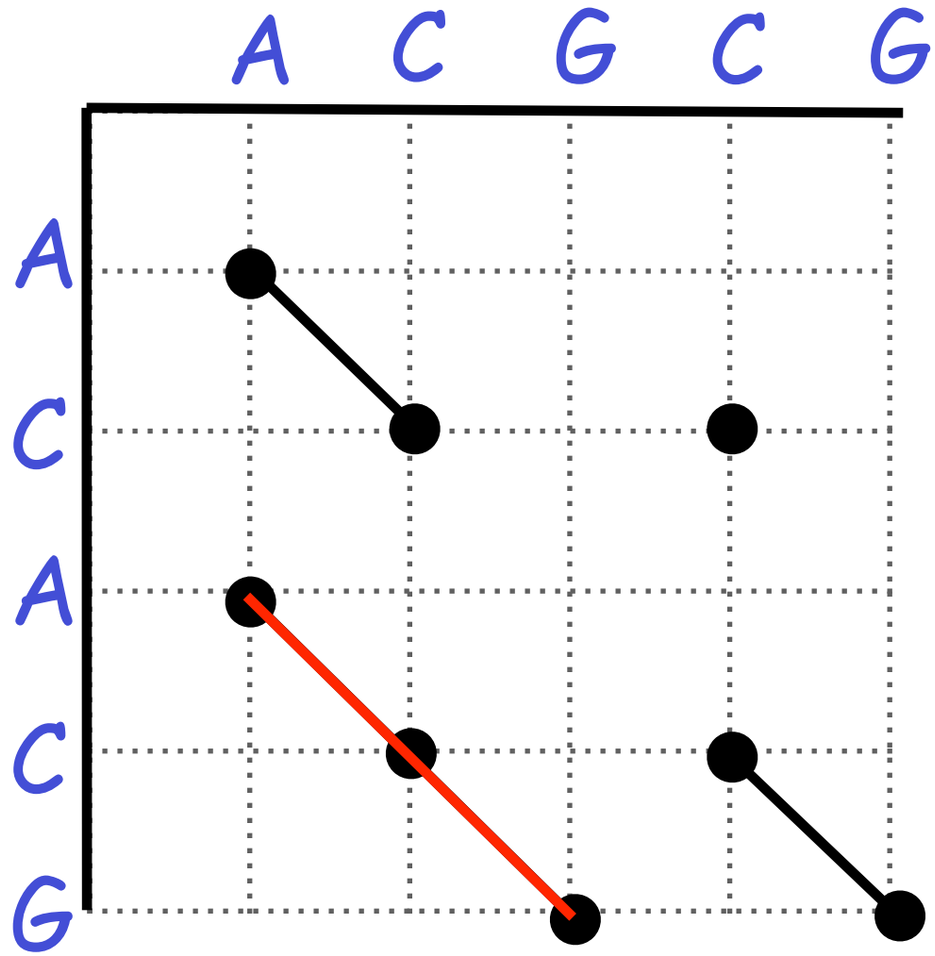
Dot plots: simple graphical approach

- Now simply put dots where the horizontal and vertical sequence values match



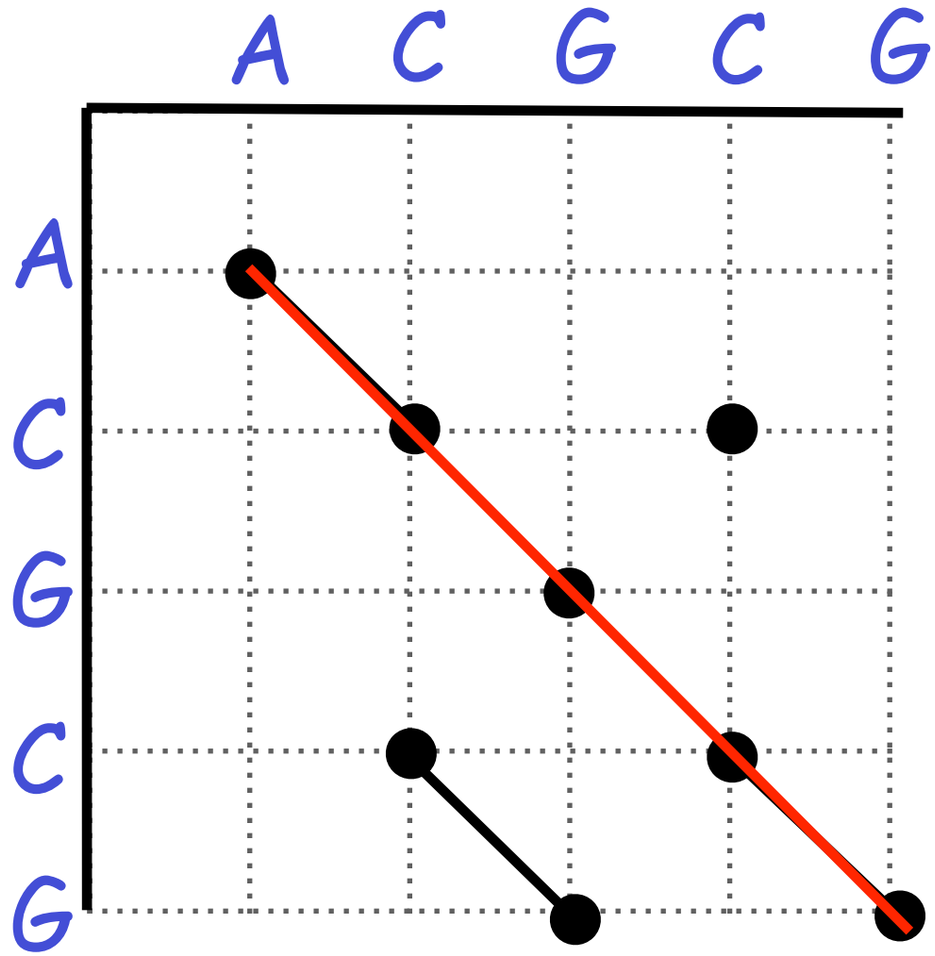
Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



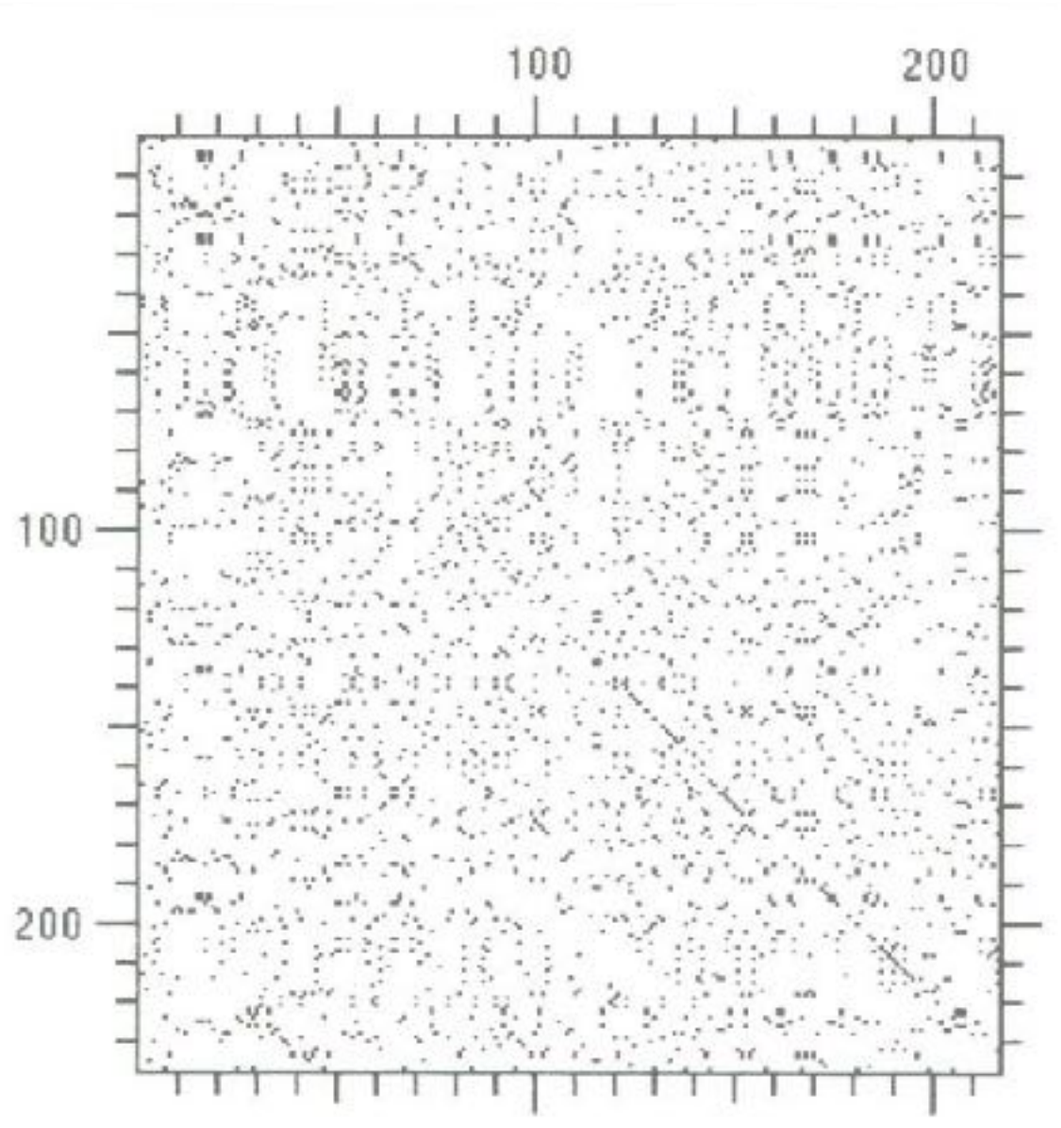
Dot plots: simple graphical approach

Q. What would the dot matrix of a two identical sequences look like?



Dot plots: simple graphical approach

- Dot matrices for long sequences can be noisy



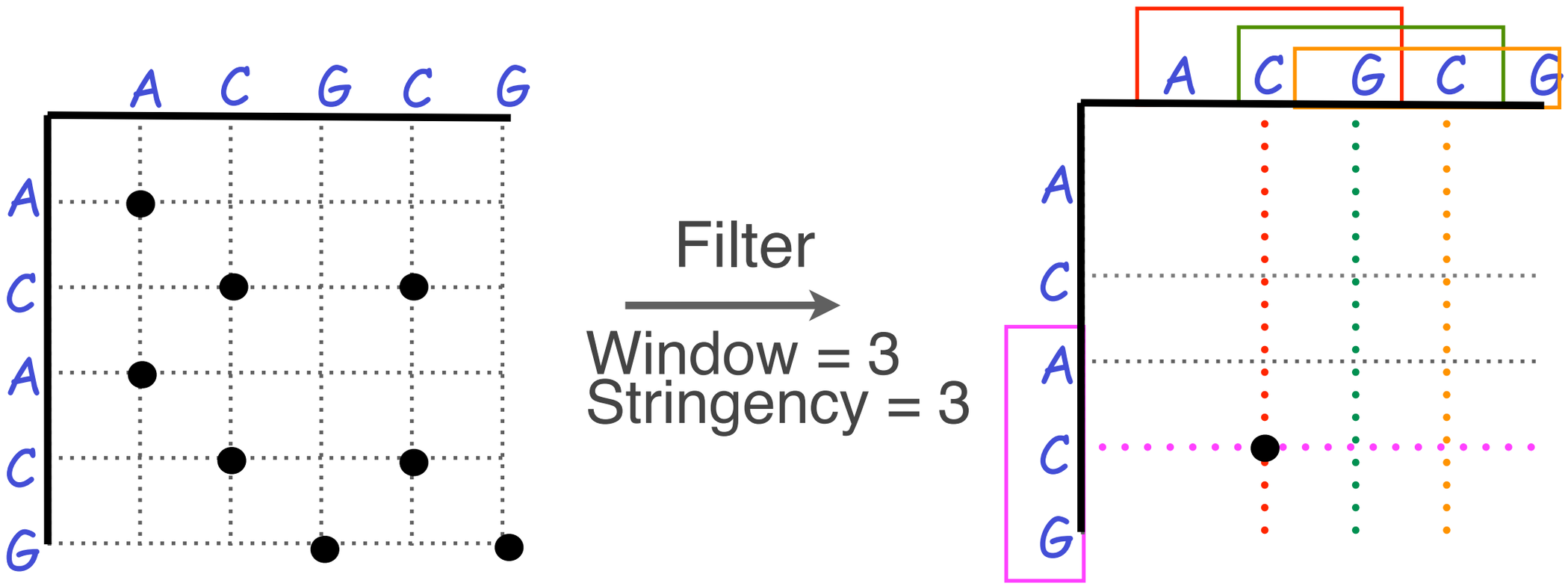
Dot plots: window size and match stringency

Solution: use a window and a threshold

– compare character by character within a window

– require certain fraction of matches within window in order to display it with a dot.

- You have to choose window size and stringency



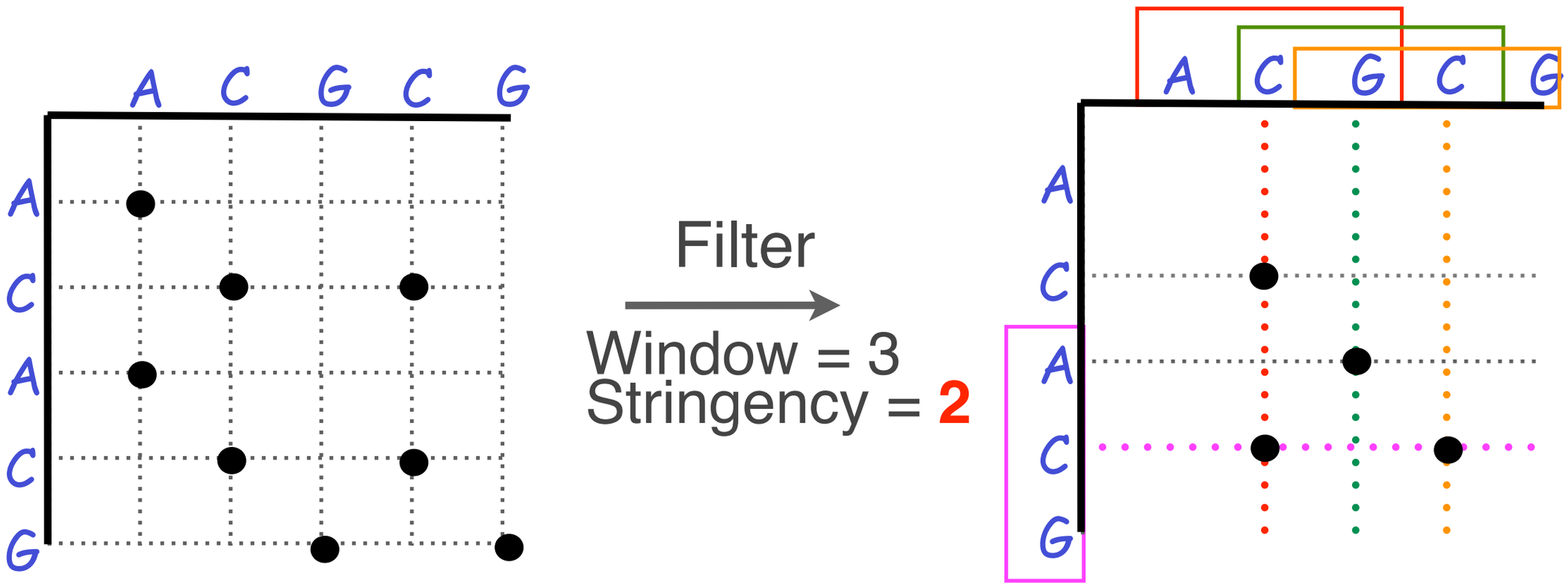
Dot plots: window size and match stringency

Solution: use a window and a threshold

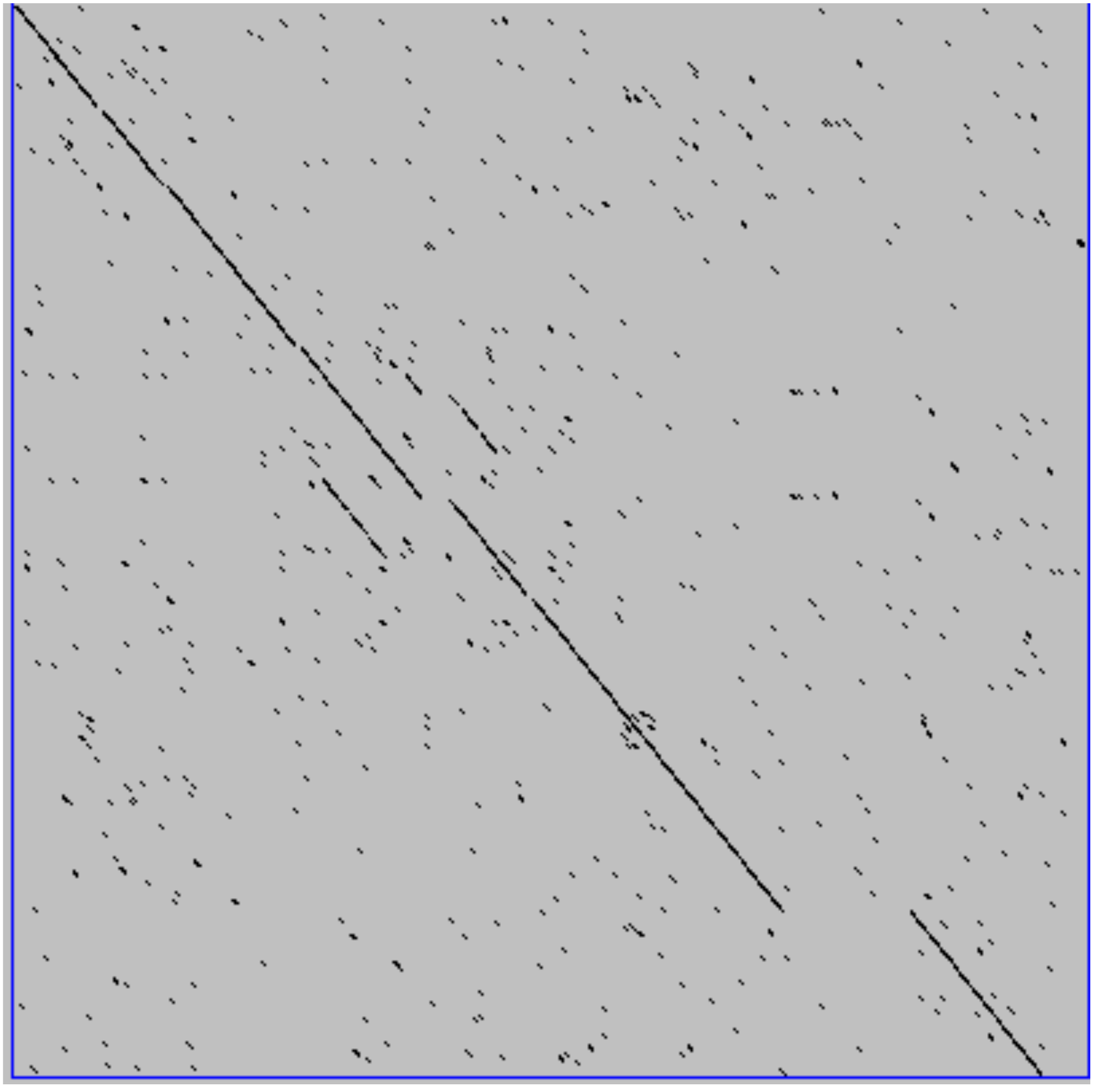
– compare character by character within a window

– require certain fraction of matches within window in order to display it with a dot.

- You have to choose window size and stringency



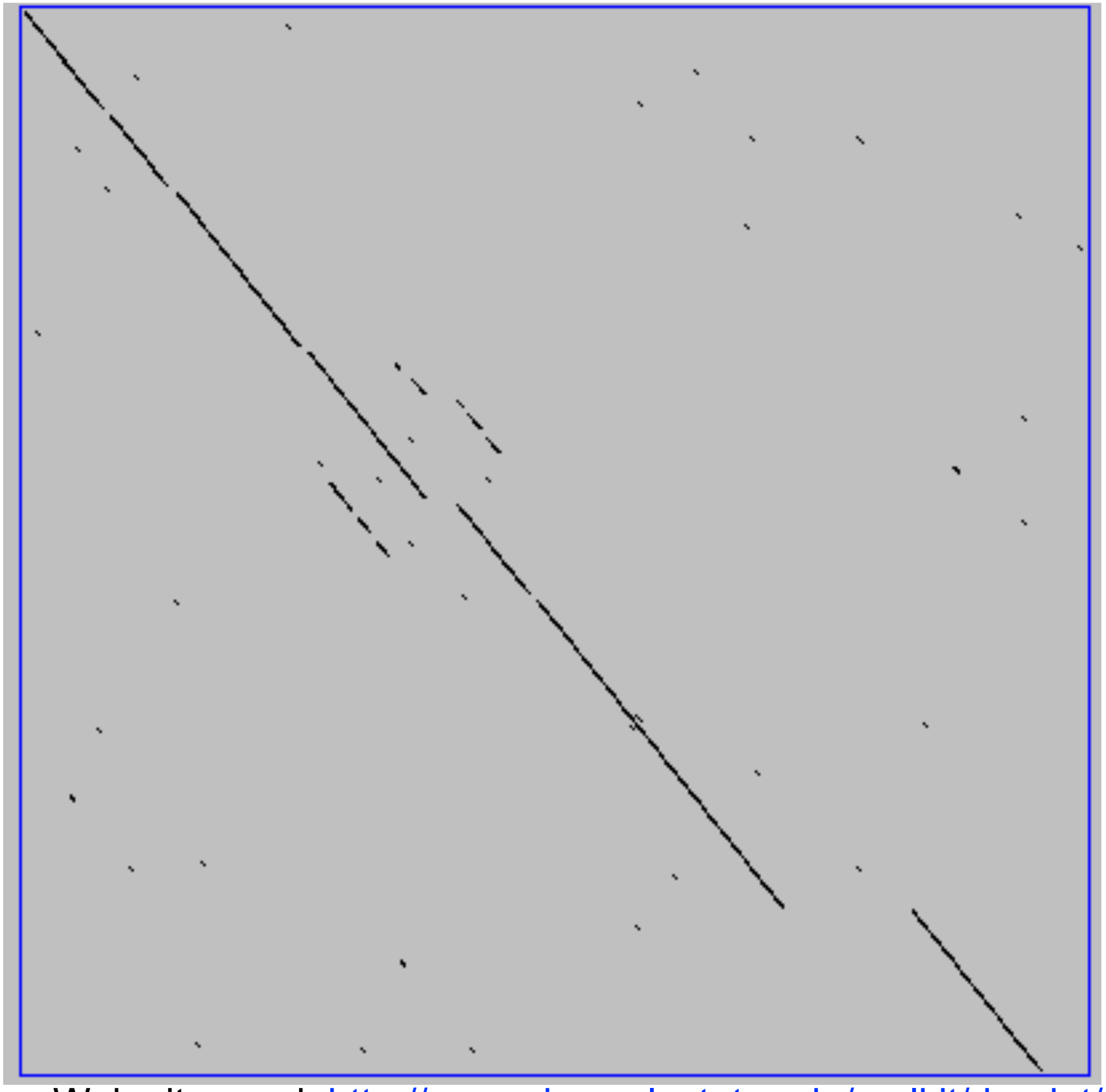
Window size = 5 bases



A dot plot simply puts a dot where two sequences match. In this example, dots are placed in the plot if 5 bases in a row match perfectly. Requiring a 5 base perfect match is a **heuristic** – only look at regions that have a certain degree of identity.

Do you expect evolutionarily related sequences to have more word matches (matches in a row over a certain length) than random or unrelated sequences?

Window size = 7 bases



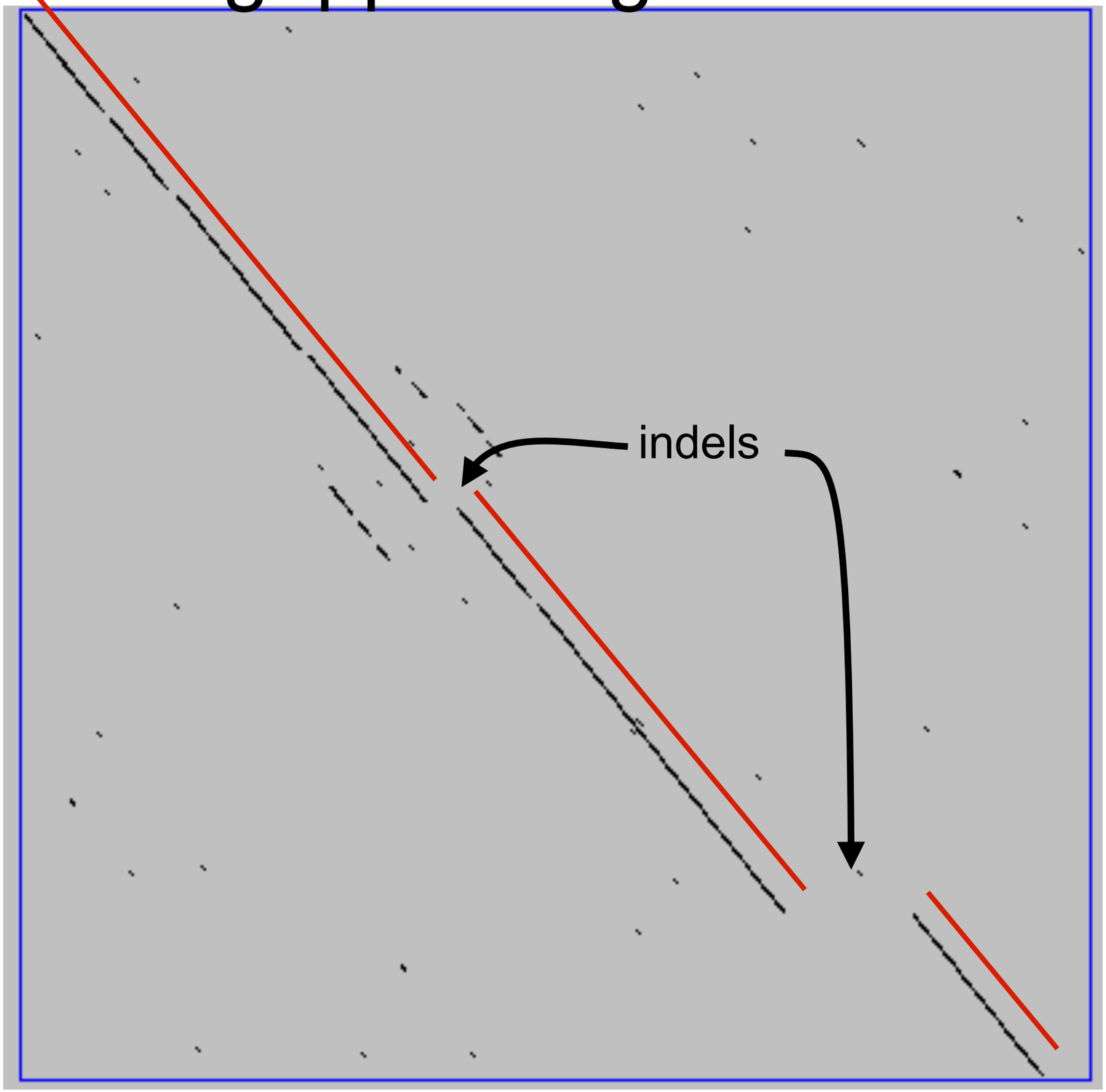
This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be place. Noise is reduced.

Using windows of a certain length is very similar to using words (kmers) of N characters in the heuristic alignment search tools

Bigger window (kmer) fewer matches to consider

Web site used: http://www.vivo.colostate.edu/molkit/dnadot/

Ungapped alignments



Only **diagonals** can be followed.

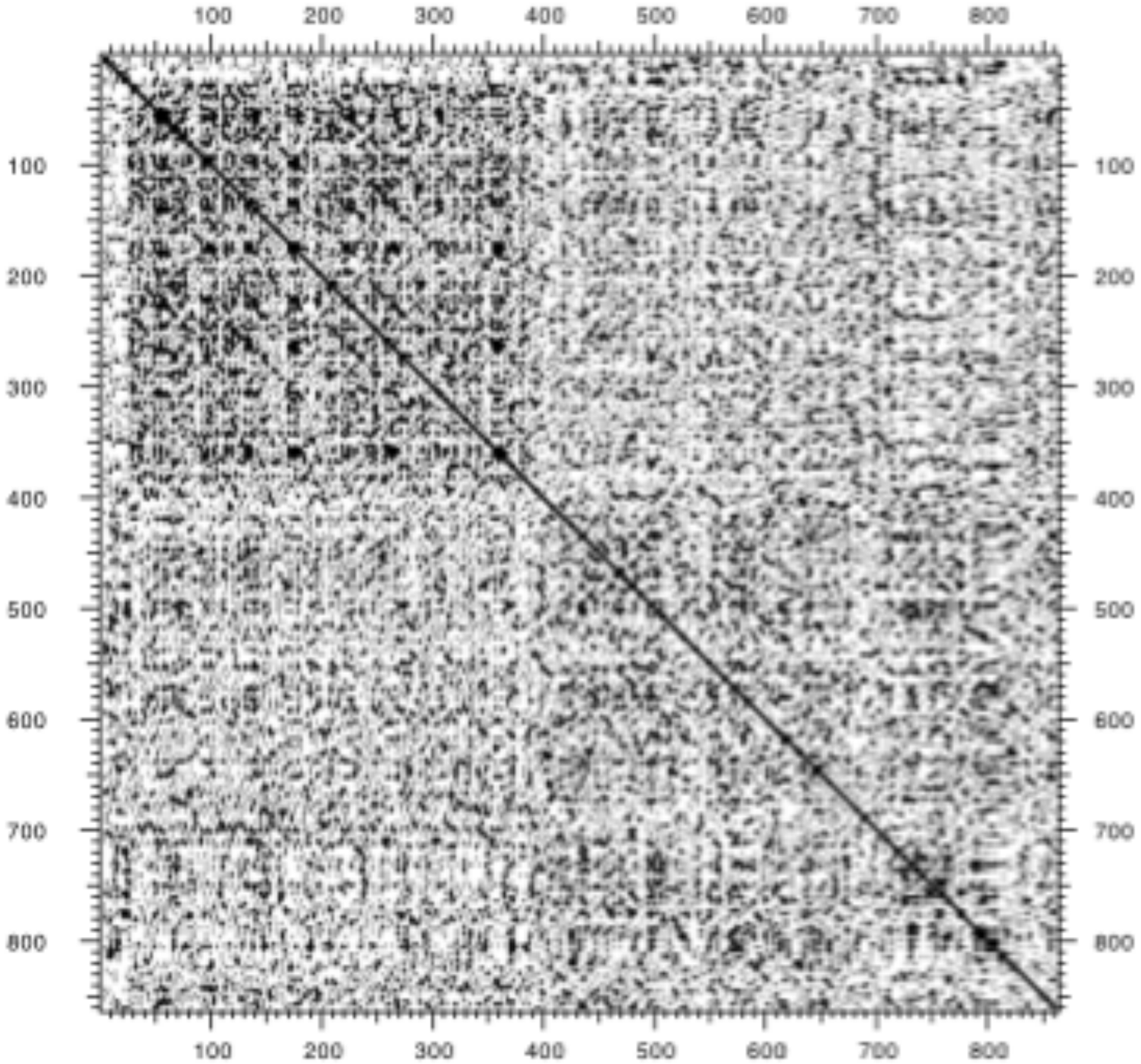
Downward or rightward paths represent **insertion** or **deletions** (gaps in one sequence or the other).

Web site used: http://www.vivo.colostate.edu/molkit/dnadot/

Uses for dot matrices

- Visually assessing the similarity of two protein or two nucleic acid sequences
- Finding local repeat sequences within a larger sequence by comparing a sequence to itself
  - Repeats appear as a set of diagonal runs stacked vertically and/or horizontally

Repeats



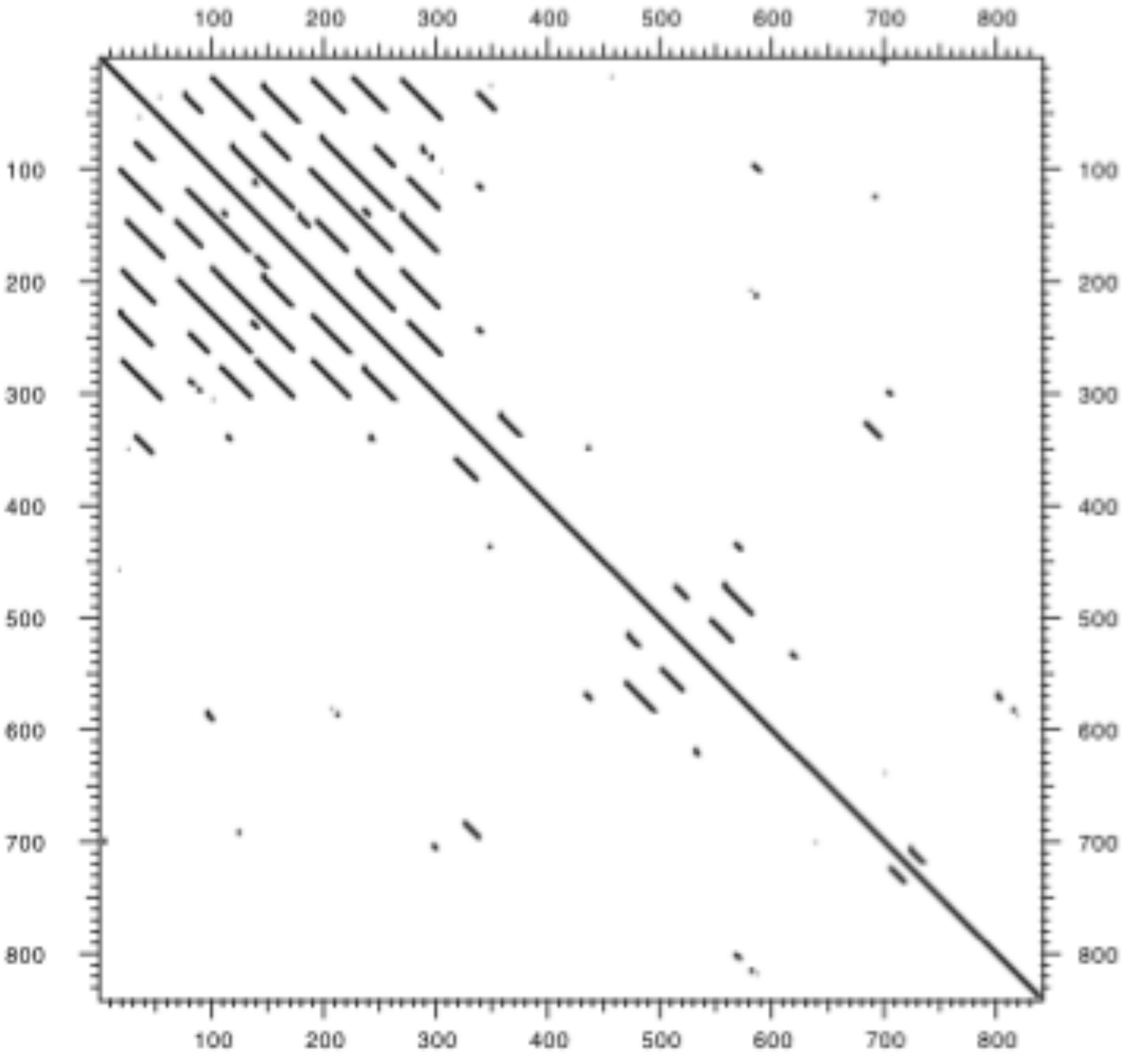
Human LDL receptor protein sequence (Genbank P01130)

W = 1

S = 1

(Figure from Mount, “Bioinformatics sequence and genome analysis”)

Repeats



Human LDL receptor protein sequence (Genbank P01130)

W = 23

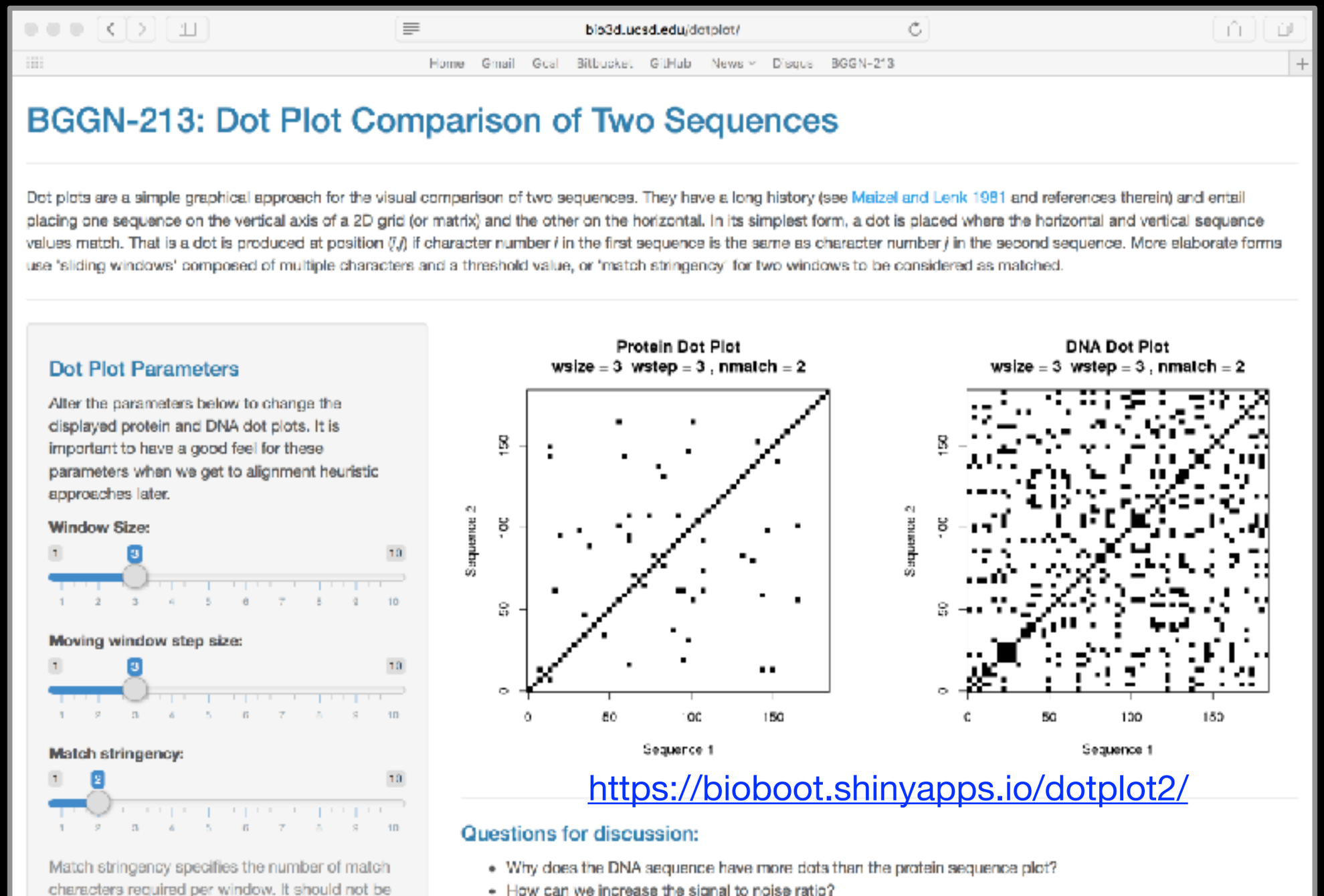
S = 7

(Figure from Mount, “Bioinformatics sequence and genome analysis”)

Your Turn!

Exploration of dot plot parameters (hands-on worksheet **Section 1**)

http://bio3d.ucsd.edu/dotplot/ https://bioboot.shinyapps.io/dotplot/



ALIGNMENT FOUNDATIONS

- **Why…**
  - Why compare biological sequences?
- **What…**
  - Alignment view of sequence changes during evolution (matches, mismatches and gaps)
- **How…**
  - Dot matrices
  - Dynamic programing
    - Global alignment
    - Local alignment
  - BLAST heuristic approach

The Dynamic Programming Algorithm

- The dynamic programming algorithm can be thought of an extension to the dot plot approach
  - One sequence is placed down the side of a grid and another across the top
  - Instead of placing a dot in the grid, we **compute a score** for each position
  - Finding the optimal alignment corresponds to finding the path through the grid with the **best possible score**

Needleman, S.B. & Wunsch, C.D. (1970) "A general method applicable to the search for similarities in the amino acid sequences of two proteins." J. Mol. Biol. 48:443-453.

Algorithm of **Needleman and Wunsch**

- The Needleman–Wunsch approach to global sequence alignment has three basic steps:

  (1) setting up a 2D-grid (or **alignment matrix**),

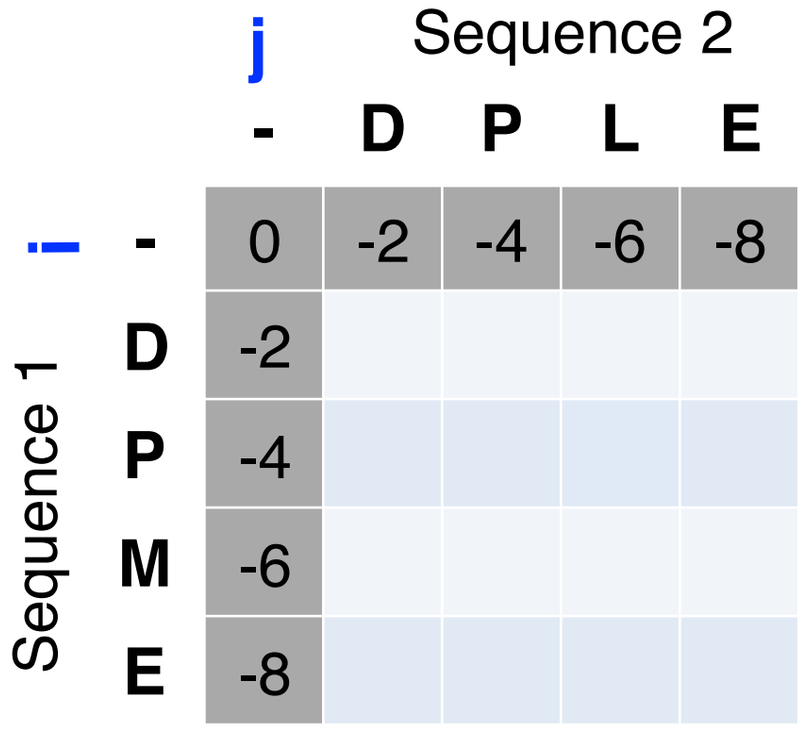
  (2) **scoring the matrix**, and

  (3) identifying the **optimal path** through the matrix

Needleman, S.B. & Wunsch, C.D. (1970) "A general method applicable to the search for similarities in the amino acid sequences of two proteins." J. Mol. Biol. 48:443-453.

Scoring the alignment matrix

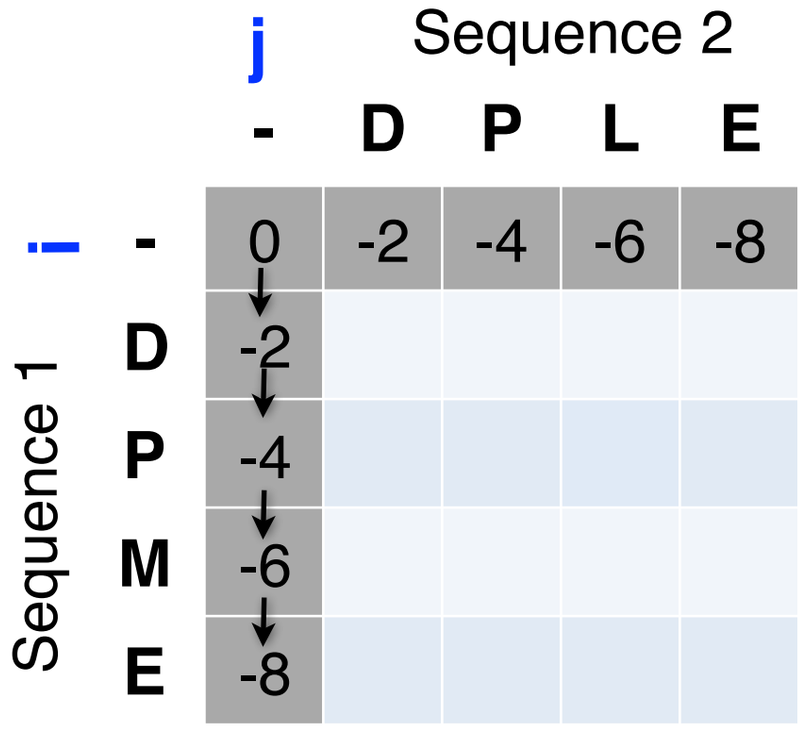
- Start by filling in the first row and column – these are all indels (gaps).
  - Each step you take you will add the **gap penalty** to the score ($S_{i,j}$) accumulated in the previous cell



Scores: match = +1, mismatch = -1, gap = -2

Scoring the alignment matrix

- Start by filling in the first row and column – these are all indels (gaps).
  - Each step you take you will add the **gap penalty** to the score ($S_{i,j}$) accumulated in the previous cell



Scores: match = +1, mismatch = -1, gap = -2

$S_{i+4} = (-2) + (-2) + (-2) + (-2)$

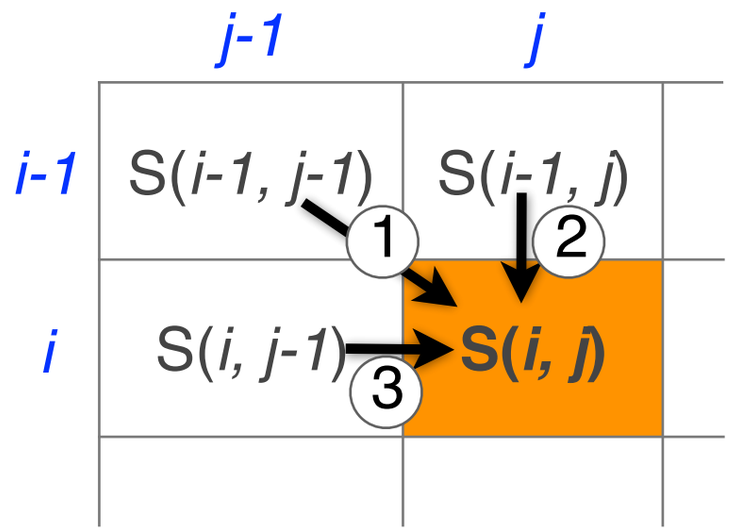
```
Seq1: DPME
Seq2: ----
```

Scoring the alignment matrix

- Then go to the empty corner cell (upper left). It has filled in values in up, left and diagonal directions
  - Now can ask which of the three directions gives the highest score?
  - keep track of this score and direction

		j			
	-	D	P	L	E
-	0	-2	-4	-6	-8
i D	-2	?			
P	-4				
M	-6				
E	-8				

Scores: match = +1, mismatch = -1, gap = -2



Scoring the alignment matrix

- Then go to the empty corner cell (upper left). It has filled in values in up, left and diagonal directions
  - Now can ask which of the three directions gives the highest score?
  - keep track of this score and direction

		j			
	-	D	P	L	E
-	0	-2	-4	-6	-8
i D	-2	?			
P	-4				
M	-6				
E	-8				

Scores: match = +1, mismatch = -1, gap = -2

$$S(i, j) = \text{Max}\begin{cases} S(i\text{-}1, j\text{-}1) + \text{(mis)match} & \searrow\ 1 \\ S(i\text{-}1, j) + \text{gap penalty} & \downarrow\ 2 \\ S(i, j\text{-}1) + \text{gap penalty} & \rightarrow\ 3 \end{cases}$$

Scoring the alignment matrix

- Then go to the empty corner cell (upper left). It has filled in values in up, left and diagonal directions
  - Now can ask which direction gives the highest score
  - keep track of direction and score

		j -	*j* D	*j* P	*j* L	*j* E
	-	0	-2	-4	-6	-8
i -	D	-2	1			
	P	-4				
	M	-6				
	E	-8				

Scores: match = +1, mismatch = -1, gap = -2

↘ ① (0)+(+1) = +1 <= (D-D) match!

↓ ② (-2)+(-2) = -4

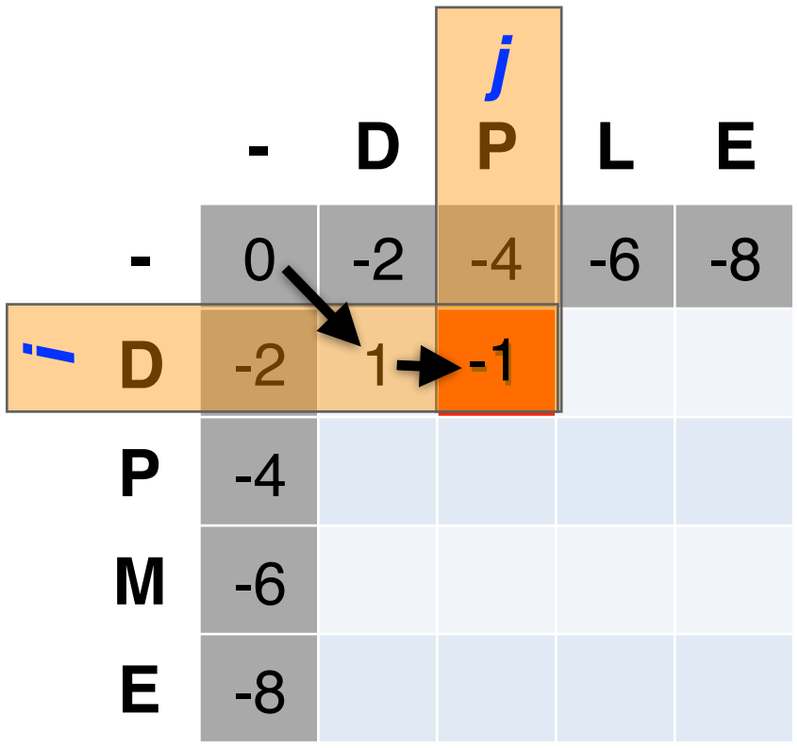
→ ③ (-2)+(-2) = -4

Alignment

D
D

Scoring the alignment matrix

- At each step, the score in the current cell is determine by the scores in the neighboring cells
  - The maximal score and the direction that gave that score is stored (we will use these later to determine the optimal alignment)



Scores: match = +1, mismatch = -1, gap = -2

① (-2)+(-1) = -3 <= (D-P) mismatch!

② (-4)+(-2) = -6

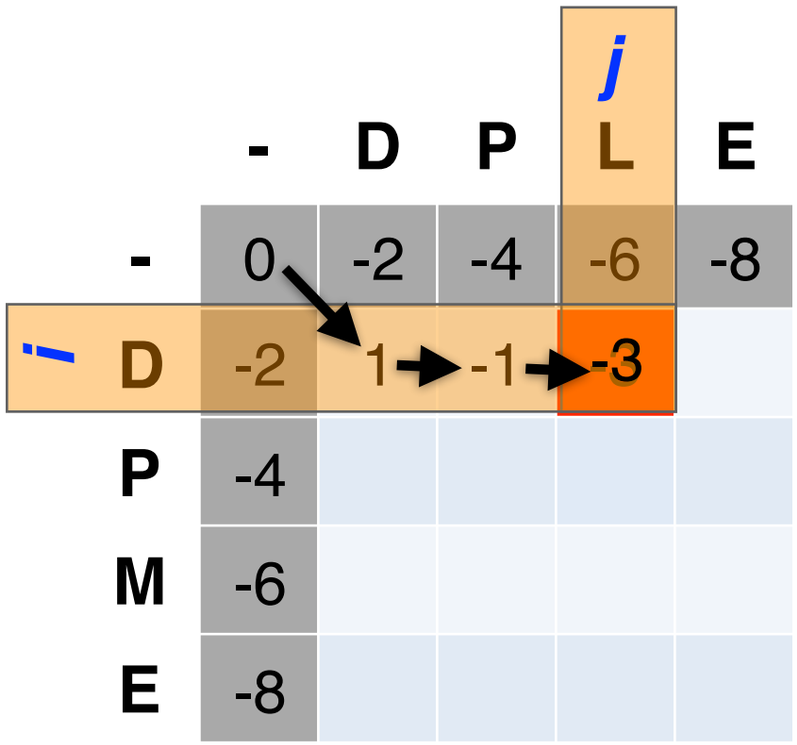
③ (1)+(-2) = -1

Alignment

D-
DP

Scoring the alignment matrix

- We will continue to store the alignment score ($S_{i,j}$) for all possible alignments in the alignment matrix.



Scores: match = +1, mismatch = -1, gap = -2

① (-4)+(-1) = -5 <= (D-L) mismatch

② (-6)+(-2) = -8

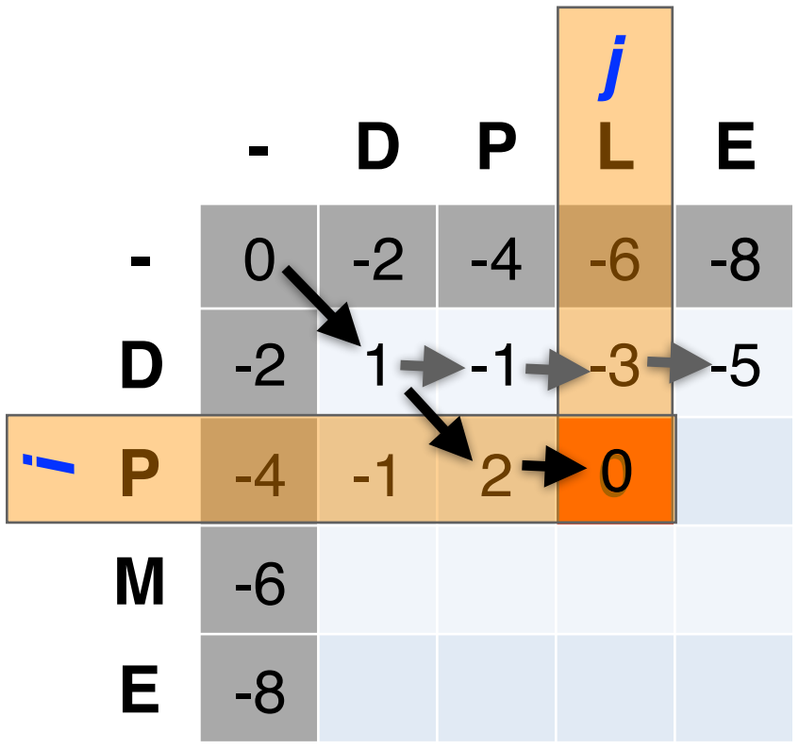
③ (-1)+(-2) = -3

Alignment

```
D--
DPL
```

Scoring the alignment matrix

- For the highlighted cell, the corresponding score ($S_{i,j}$) refers to the score of the optimal alignment of the first *i* characters from sequence1, and the first *j* characters from sequence2.



Scores: match = +1, mismatch = -1, indel = -2

① (-1)+(-1) = -2

② (-3)+(-2) = -5

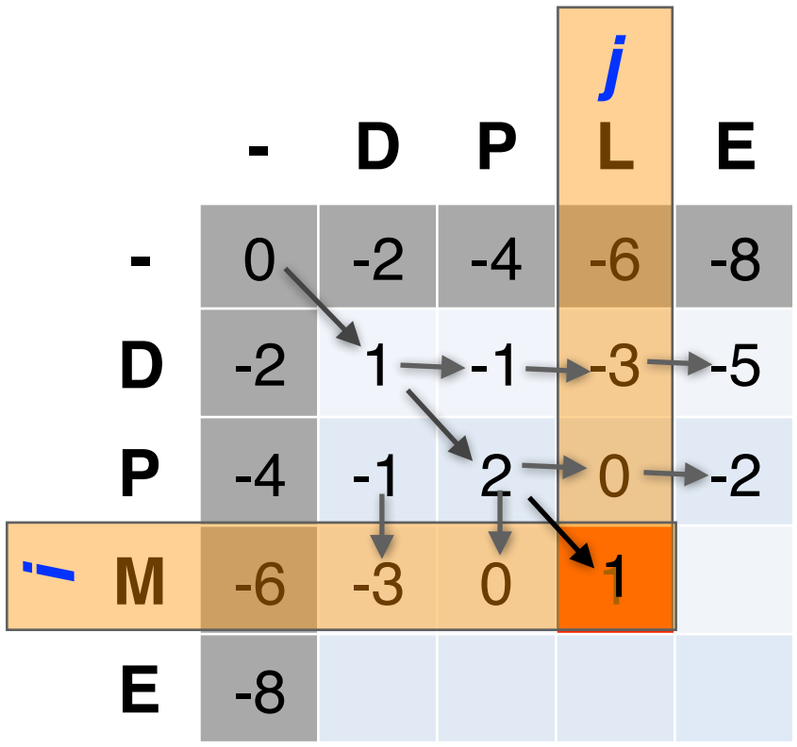
③ (2)+(-2) = 0

Alignment

DP-
DPL

Scoring the alignment matrix

- At each step, the score in the current cell is determine by the scores in the neighboring cells
  - The maximal score and the direction that gave that score is stored



Scores: match = +1, mismatch = -1, indel = -2

① (2)+(**-1**) = 0 <= mismatch

② (0)+(-2) = -2

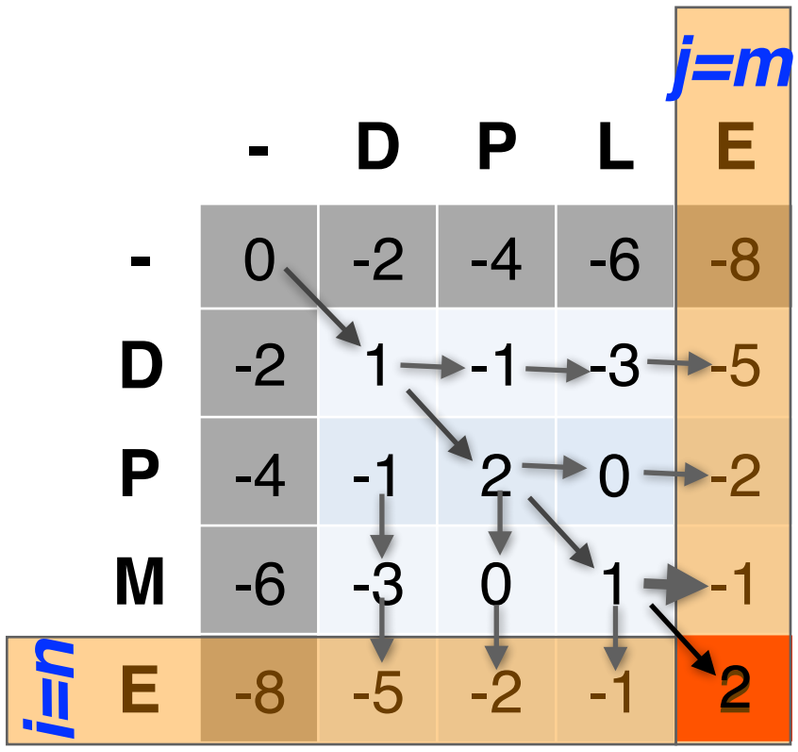
③ (0)+(-2) = -2

Alignment

DPM
DPL

Scoring the alignment matrix

- The score of the best alignment of the entire sequences corresponds to $S_{n,m}$
  - (where *n* and *m* are the length of the sequences)



Scores: match = +1, mismatch = -1, indel = -2

① (+1)+(+1) = +2

② (-1)+(-2) = -3

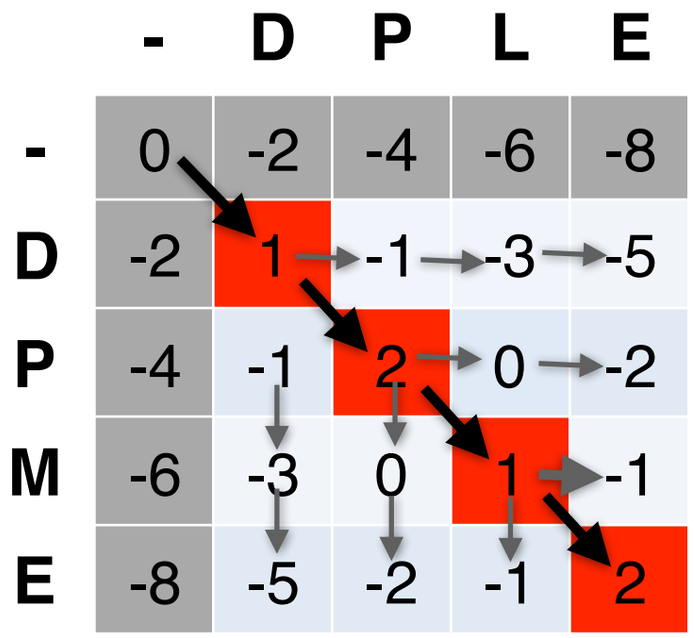
③ (-1)+(-2) = -3

Alignment

DPME
DPLE

Scoring the alignment matrix

- To find the best alignment, we retrace the arrows starting from the bottom right cell
  - N.B. The optimal alignment score and alignment are dependent on the chosen scoring system



Scores: match = +1, mismatch = -1, indel = -2

Alignment

DPME
DPLE

Questions:

- What is the optimal score for the alignment of these sequences and how do we find the optimal alignment?

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

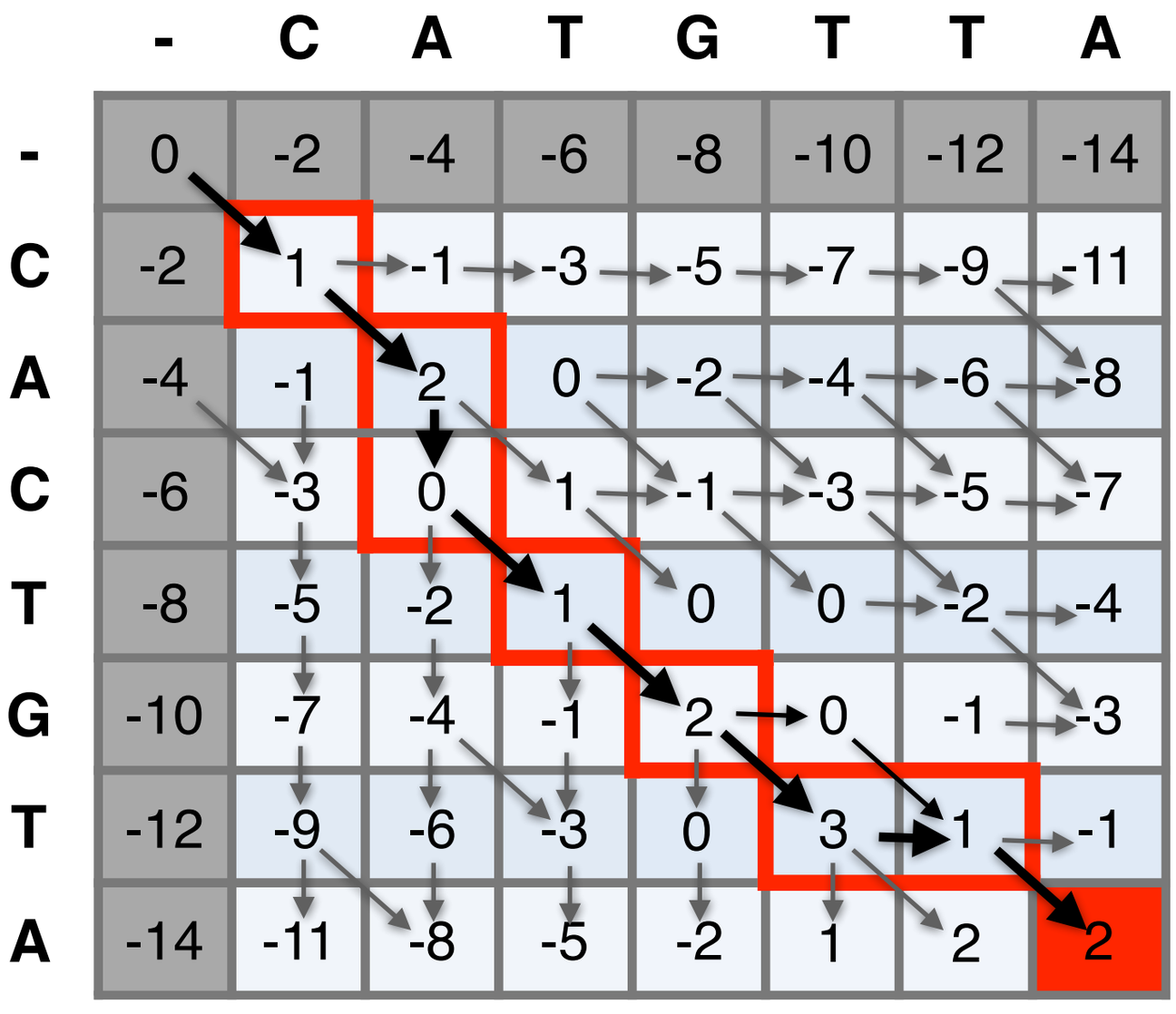
Questions:

- What is the optimal score for the alignment of these sequences and how do we find the optimal alignment?

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

Questions:

- To find the best alignment we retrace the arrows starting from the bottom right cell



More than one alignment possible

- Sometimes more than one alignment can result in the same optimal score

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

Alignment

CACTGT-A
CA-TGTTA

CACTG-TA
CA-TGTTA

The alignment and score are dependent on the scoring system

- Here we increase the gap penalty from -2 to -3

	-	C	A	T	G	T	T	A
-	0	-3	-6	-9	-12	-15	-18	-21
C	-3	1	-2	-5	-8	-11	14	-17
A	-6	-2	2	-1	-4	-7	-10	-13
C	-9	-5	-1	1	-2	-5	-8	-11
T	-12	-8	-4	0	0	-1	-4	-7
G	-15	-11	-7	-3	1	-1	-2	-5
T	-18	-14	-10	-6	-2	2	0	-3
A	-21	-17	-13	-9	-5	-1	1	1

Alignment

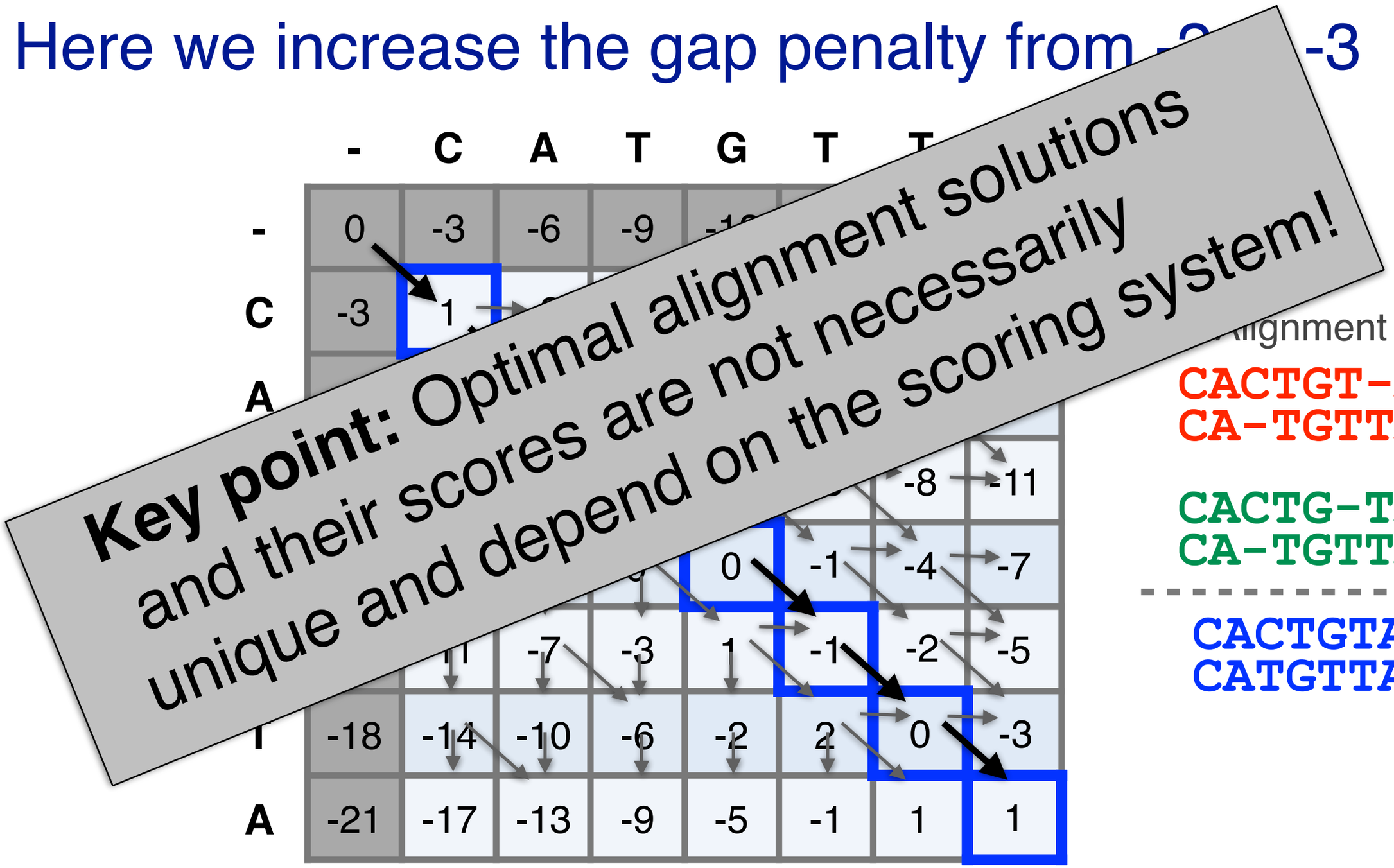
CACTGT-A
CA-TGTTA

CACTG-TA
CA-TGTTA

CACTGTA
CATGTTA

The alignment and score are dependent on the scoring system

- Here we increase the gap penalty from -2 to -3



NW DYNAMIC PROGRAMMING

Match: +2
Mismatch: -1
Gap: -2

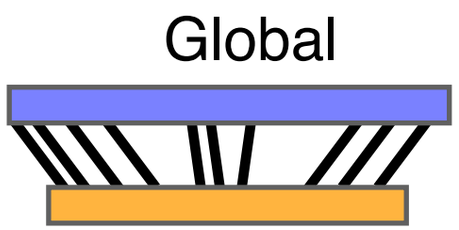
		A	G	T	T	C
	0					
A						
T						
T						
G						
C						

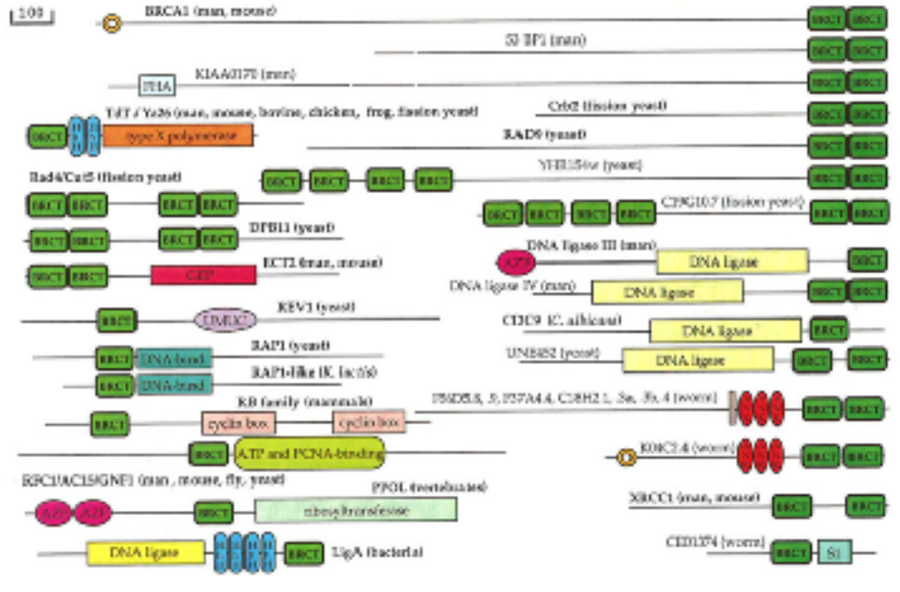
ALIGNMENT FOUNDATIONS

- **Why…**
  - Why compare biological sequences?
- **What…**
  - Alignment view of sequence changes during evolution (matches, mismatches and gaps)
- **How…**
  - Dot matrices
  - Dynamic programing
    - Global alignment
    - Local alignment
  - BLAST heuristic approach

Global *vs* local alignments

- Needleman-Wunsch is a **global alignment** algorithm
  - Resulting alignment spans the complete sequences end to end
  - This is appropriate for closely related sequences that are similar in length
- For many practical applications we require **local alignments**
  - Local alignments highlight sub-regions (*e.g.* protein domains) in the two sequences that align well





Local alignment: Definition

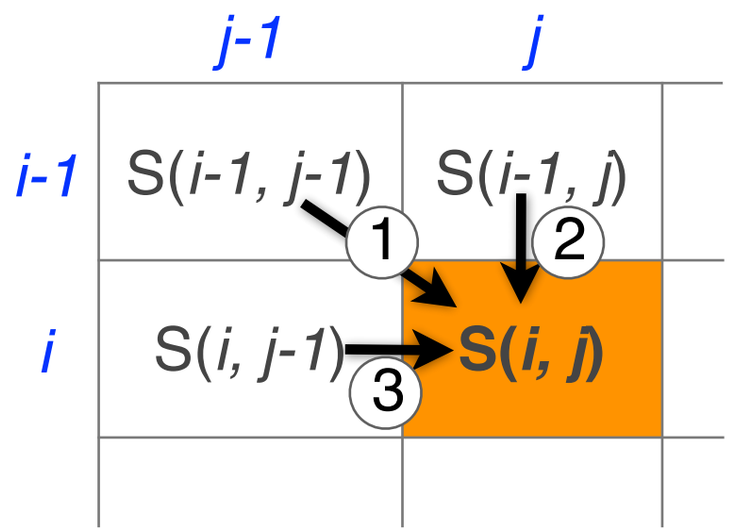
- Smith & Waterman proposed simply that a local alignment of two sequences allow arbitrary-length segments of each sequence to be aligned, with no penalty for the unaligned portions of the sequences. Otherwise, the score for a local alignment is calculated the same way as that for a global alignment

Smith, T.F. & Waterman, M.S. (1981) “Identification of common molecular subsequences.” J. Mol. Biol. 147:195-197.

The Smith-Waterman algorithm

- Three main modifications to Needleman-Wunsch:
  - Allow a node to start at 0
  - The score for a particular cell cannot be negative
    - if all other score options produce a negative value, then a zero must be inserted in the cell
  - Record the highest- scoring node, and trace back from there

$$S(i, j) = \text{Max} \begin{cases} S(i\text{-}1, j\text{-}1) + \text{(mis)match} & \searrow \ (1) \\ S(i\text{-}1, j) - \text{gap penalty} & \downarrow \ (2) \\ S(i, j\text{-}1) - \text{gap penalty} & \rightarrow \ (3) \\ \mathbf{0} & (4) \end{cases}$$



Sequence 1

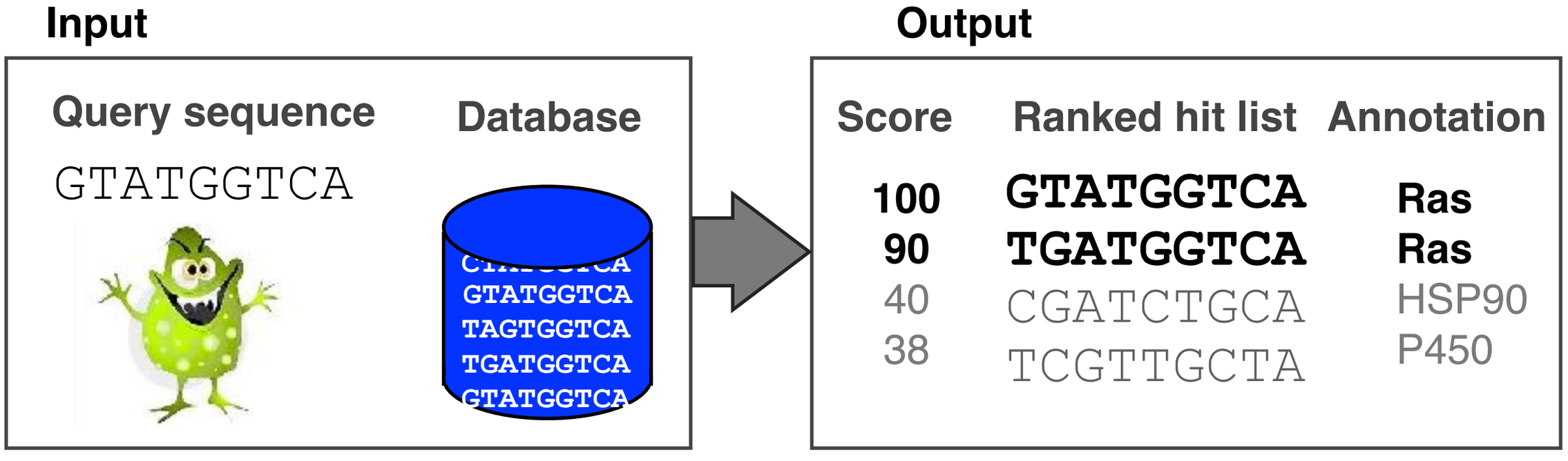
Sequence 2	-	C	A	G	C	C	U	C	G	C	U	U	A	G
-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U	0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	0.0	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C	0.0	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
A	0.0	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
U	0.0	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0
U	0.0	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0
G	0.0	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3	2.7
A	0.0	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
C	0.0	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
G	0.0	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
G	0.0	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

Local alignment

GCC-AUG
GCCUCGC

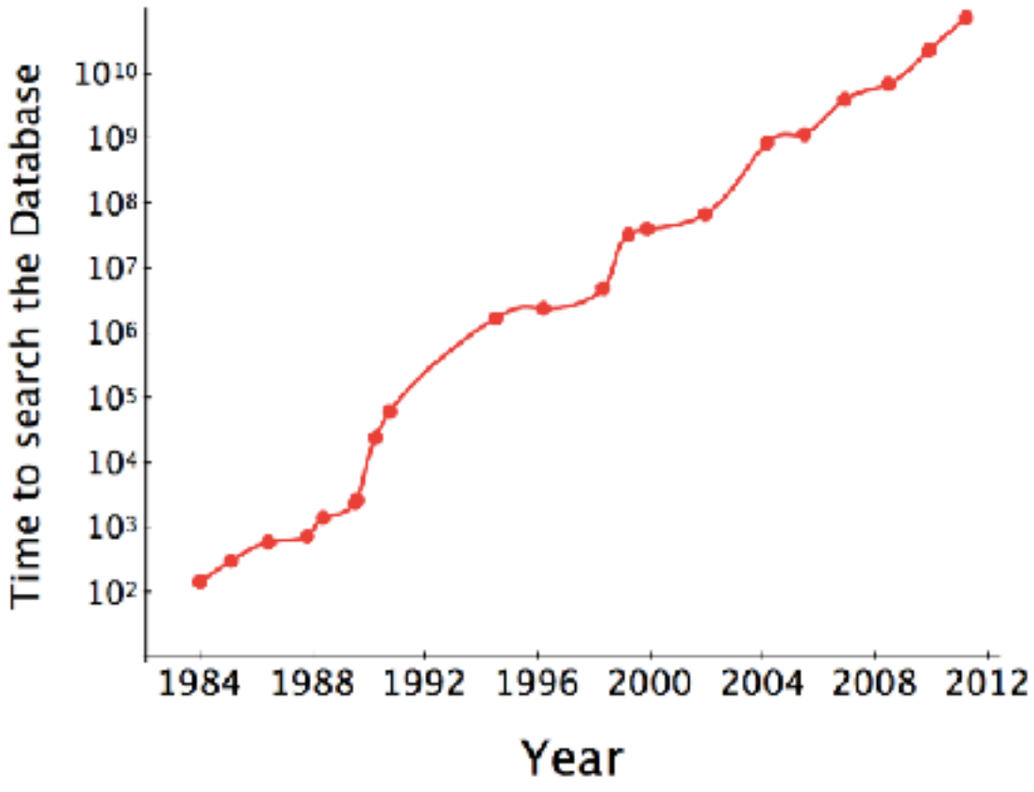
Local alignments can be used for database searching

- **Goal**: Given a query sequence (Q) and a sequence database (D), find a list of sequences from D that are most similar to Q
  - **Input**: Q, D and scoring scheme
  - **Output**: Ranked list of hits



The database search problem

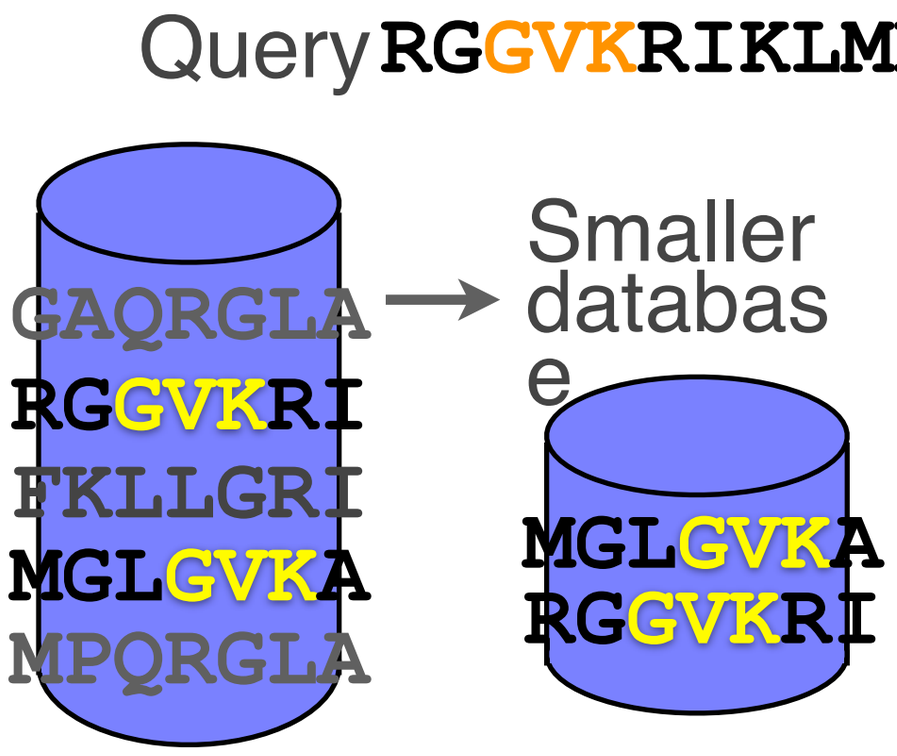
- Due to the rapid growth of sequence databases, search algorithms have to be both efficient and sensitive
  - Time to search with SW is proportional to *m* x *n* (*m* is length of query, n is length of database), **too slow for large databases**!



To reduce search time **heuristic algorithms**, such as BLAST, first remove database sequences without a strong local similarity to the query sequence in a quick initial scan.

The database search problem

- Due to the rapid growth of sequence databases, search algorithms have to be both efficient and sensitive
  - Time to search with SW is proportional to *m* x *n* (*m* is length of query, n is length of database), **too slow for large databases**!



To reduce search time **heuristic algorithms**, such as BLAST, first remove database sequences without a strong local similarity to the query sequence in a quick initial scan.

ALIGNMENT FOUNDATIONS

- **Why…**
  - Why compare biological sequences?
- **What…**
  - Alignment view of sequence changes during evolution (matches, mismatches and gaps)
- **How…**
  - Dot matrices
  - Dynamic programing
    - Global alignment
    - Local alignment
  - BLAST heuristic approach

Rapid, heuristic versions of Smith–Waterman: **BLAST**

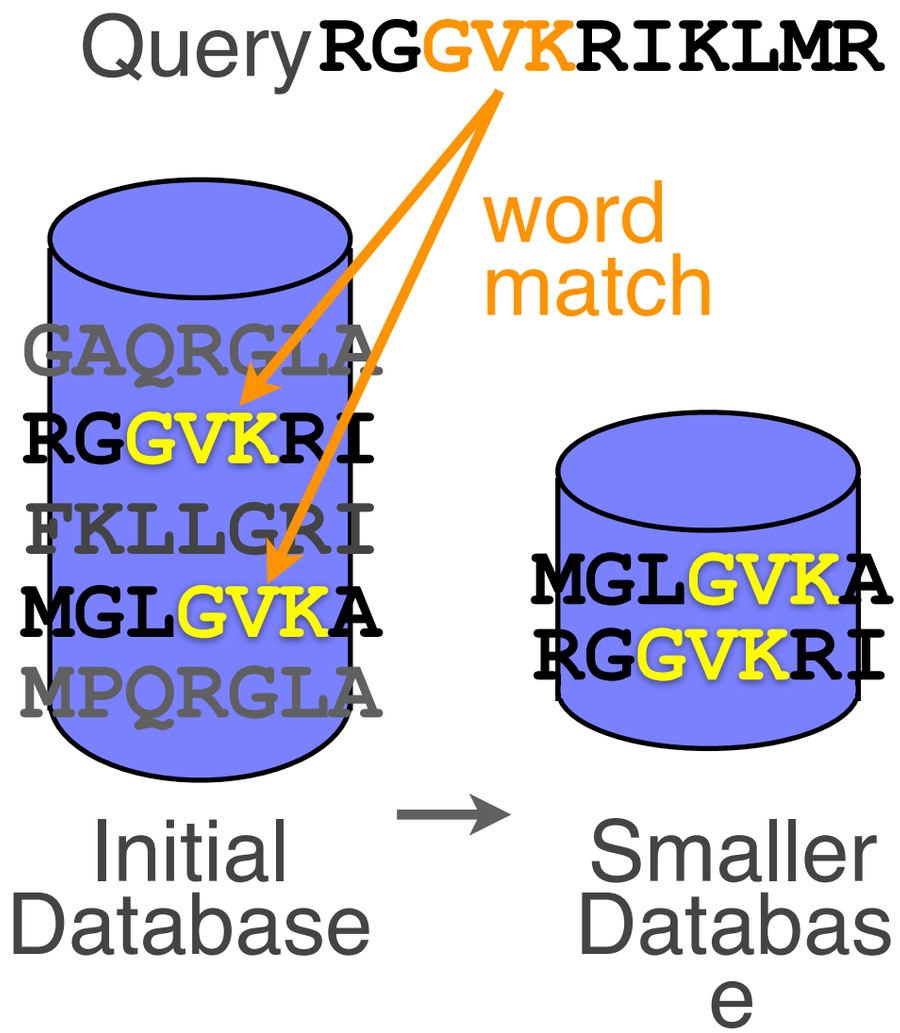
- BLAST (Basic Local Alignment Search Tool) is a simplified form of Smith-Waterman (SW) alignment that is popular because it is **fast** and **easily accessible**
  - BLAST is a heuristic approximation to SW - It examines only part of the search space
  - BLAST saves time by restricting the search by scanning database sequences for likely matches before performing more rigorous alignments
  - Sacrifices some sensitivity in exchange for speed
  - In contrast to SW, BLAST is not guaranteed to find optimal alignments

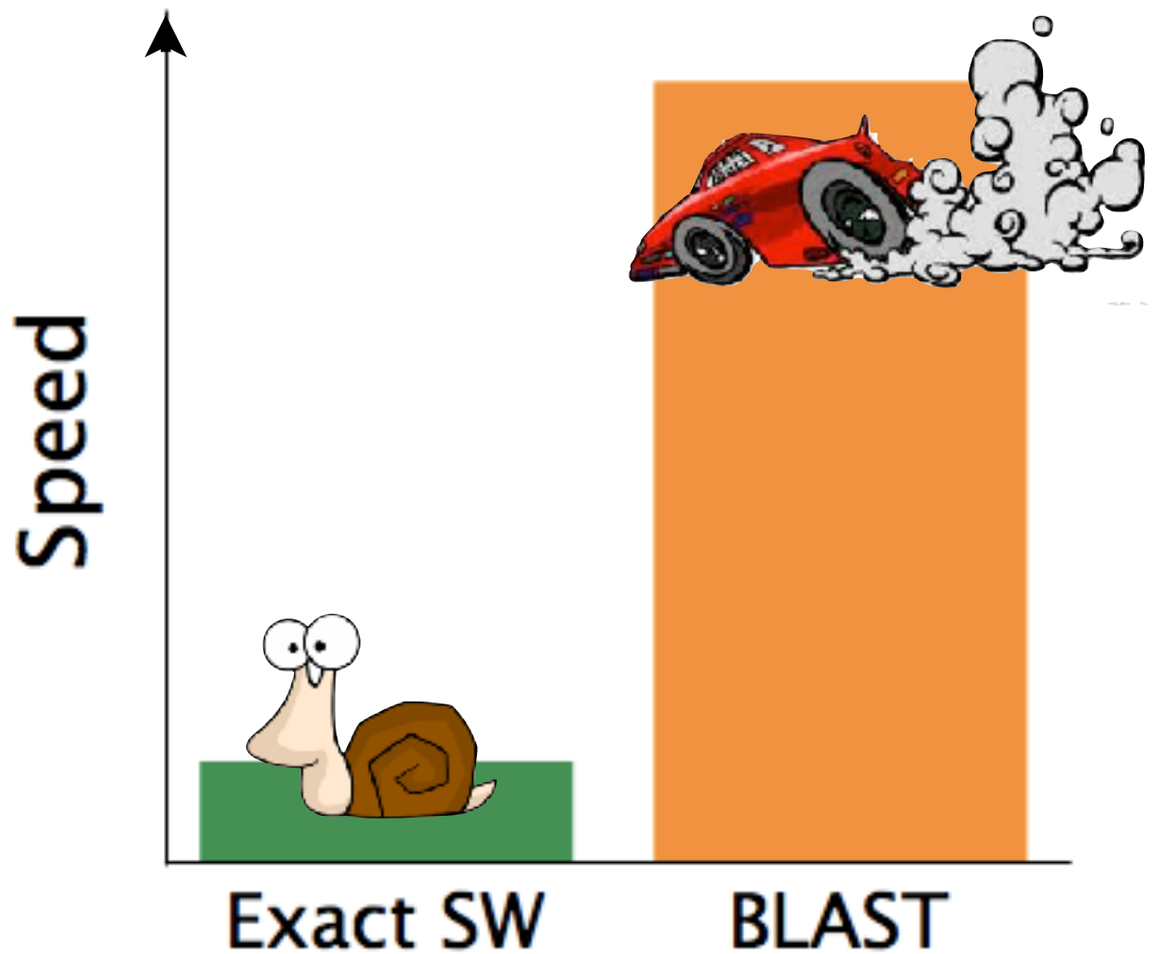
Rapid, heuristic versions of Smith–Waterman: **BLAST**

- BLAST (Basic Local Alignment Search Tool)
  simplified form of Smith-Waterman (
  is popular because it is **fast**
  – BLAST finds regions of
  sequences
  – BLAST
  d
  – Trades sensitivity in exchange for speed
  – In contrast to SW, BLAST is not guaranteed to find optimal alignments

"The central idea of the BLAST algorithm is to confine attention to sequence pairs that contain an initial **word pair** match"
Altschul et al. (1990)

- BLAST uses this pre-screening heuristic approximation resulting in an an approach that is about 50 times faster than the Smith-Waterman





How BLAST works

- Four basic phases
  - **Phase 1**: compile a list of query word pairs (w=3)

generate list of **w=3 words** for query

RGGVKRI Query sequence

RGG

GGV

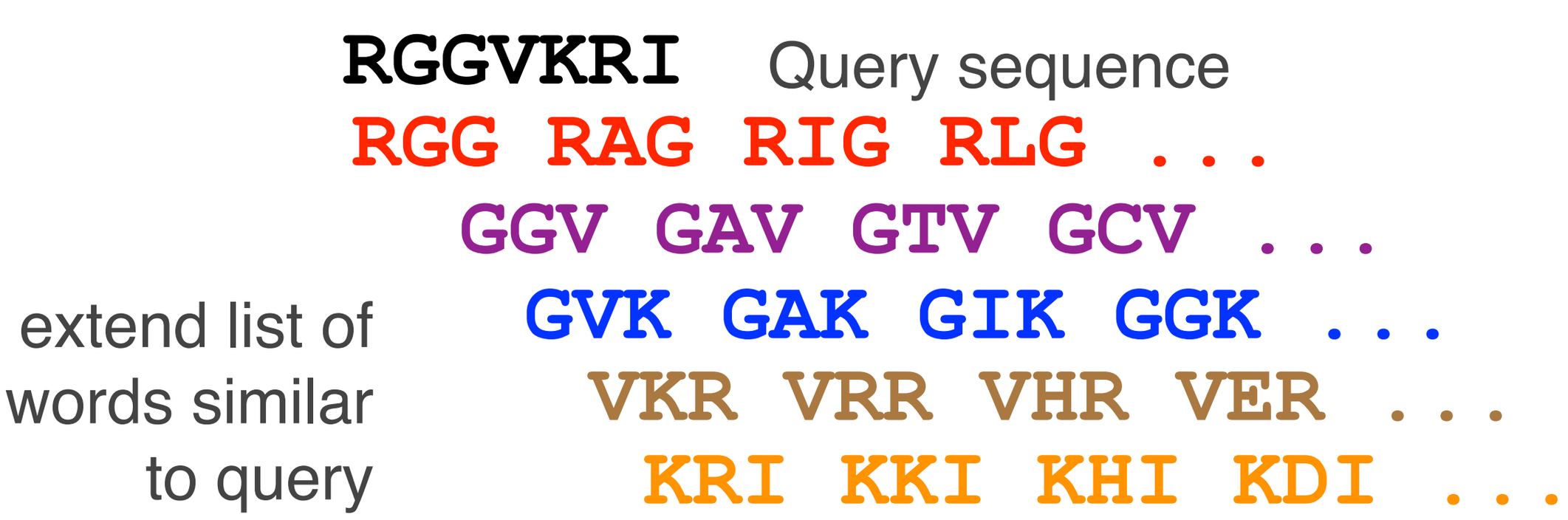
GVK

VKR

KRI

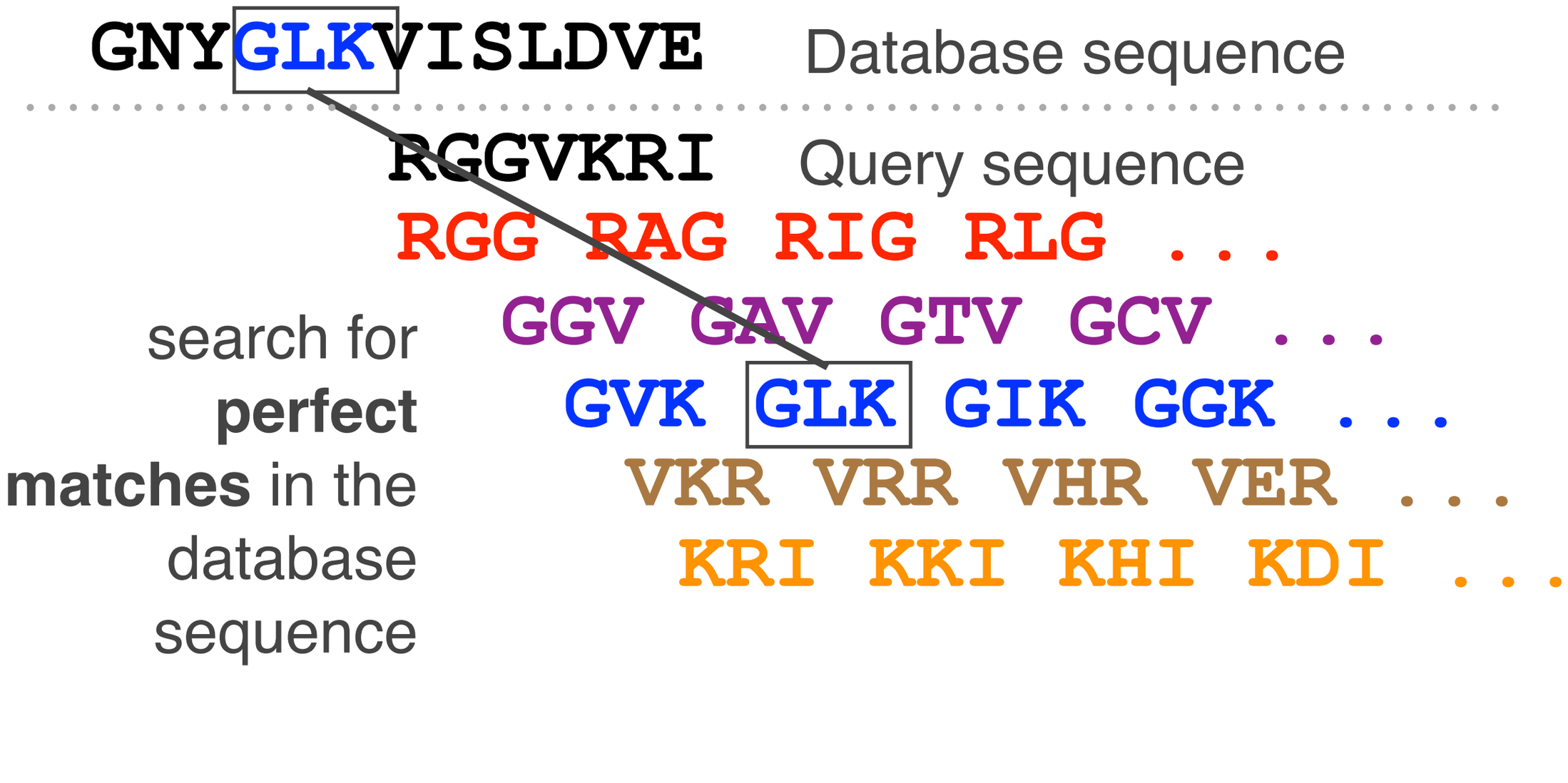
Blast

– **Phase 2**: expand word pairs to include those similar to query (defined as those above a similarity threshold to original word, i.e. match scores in substitution matrix)



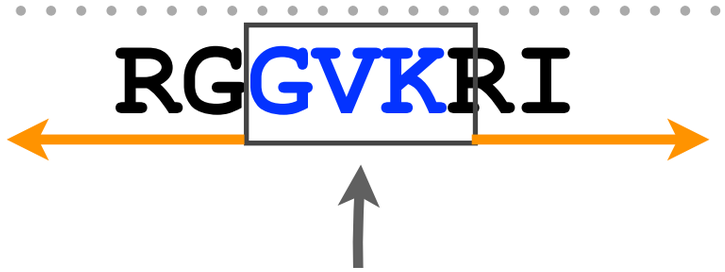
Blast

– **Phase 3**: a database is scanned to find sequence entries that match the compiled word list



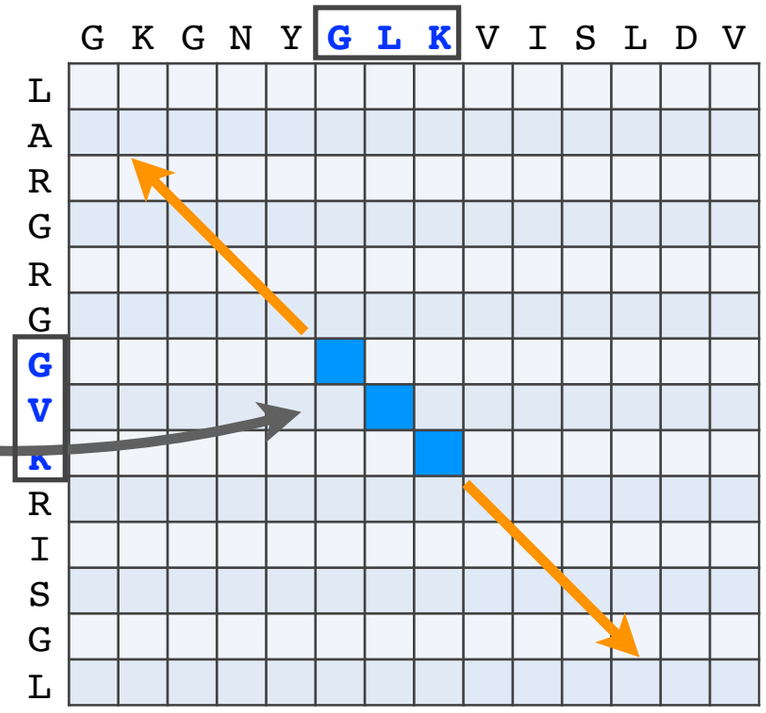
Blast

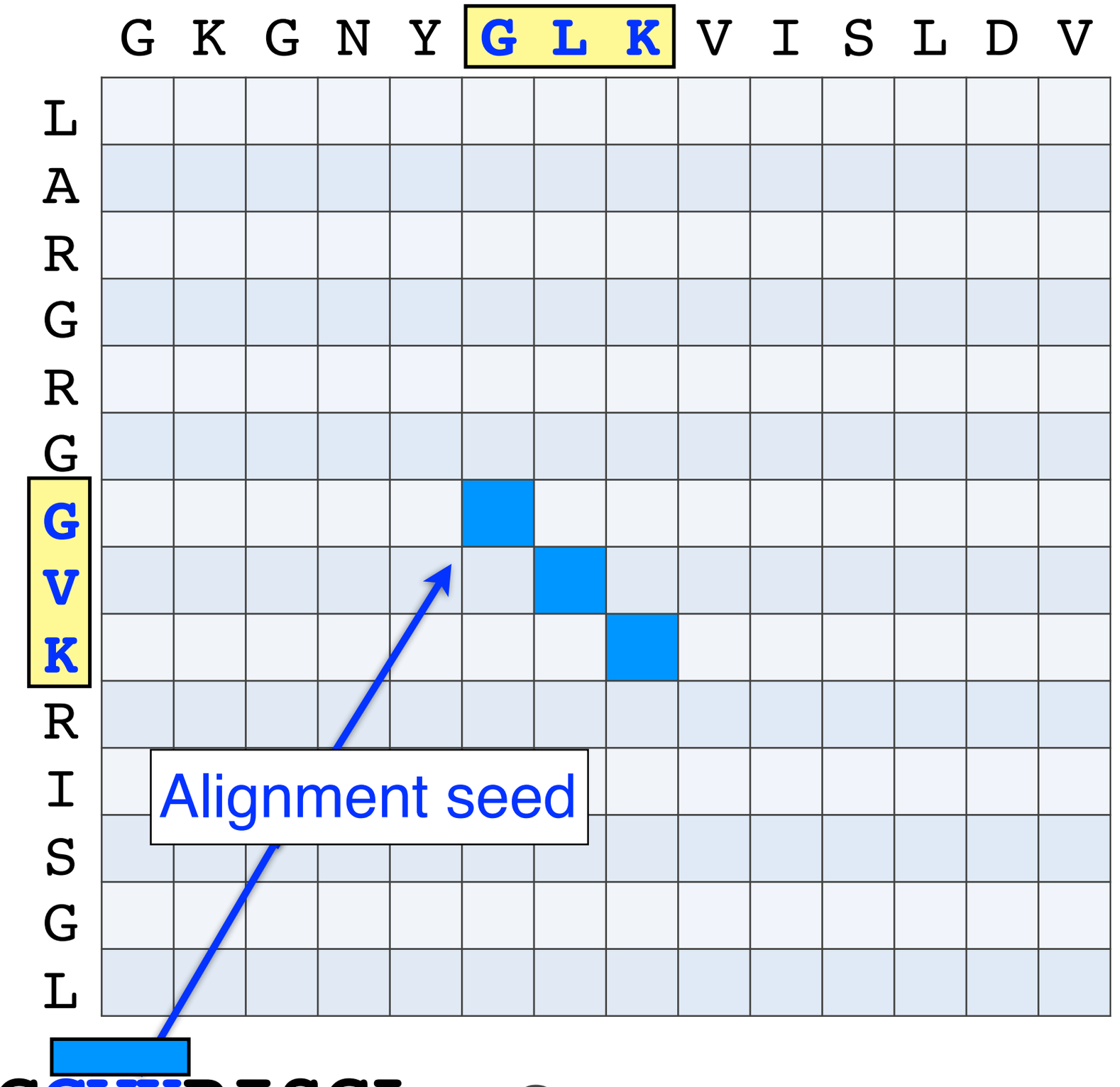
– **Phase 4**: the initial database hits are extended in both directions using dynamic programing



matched word is used as a local **alignment seed**

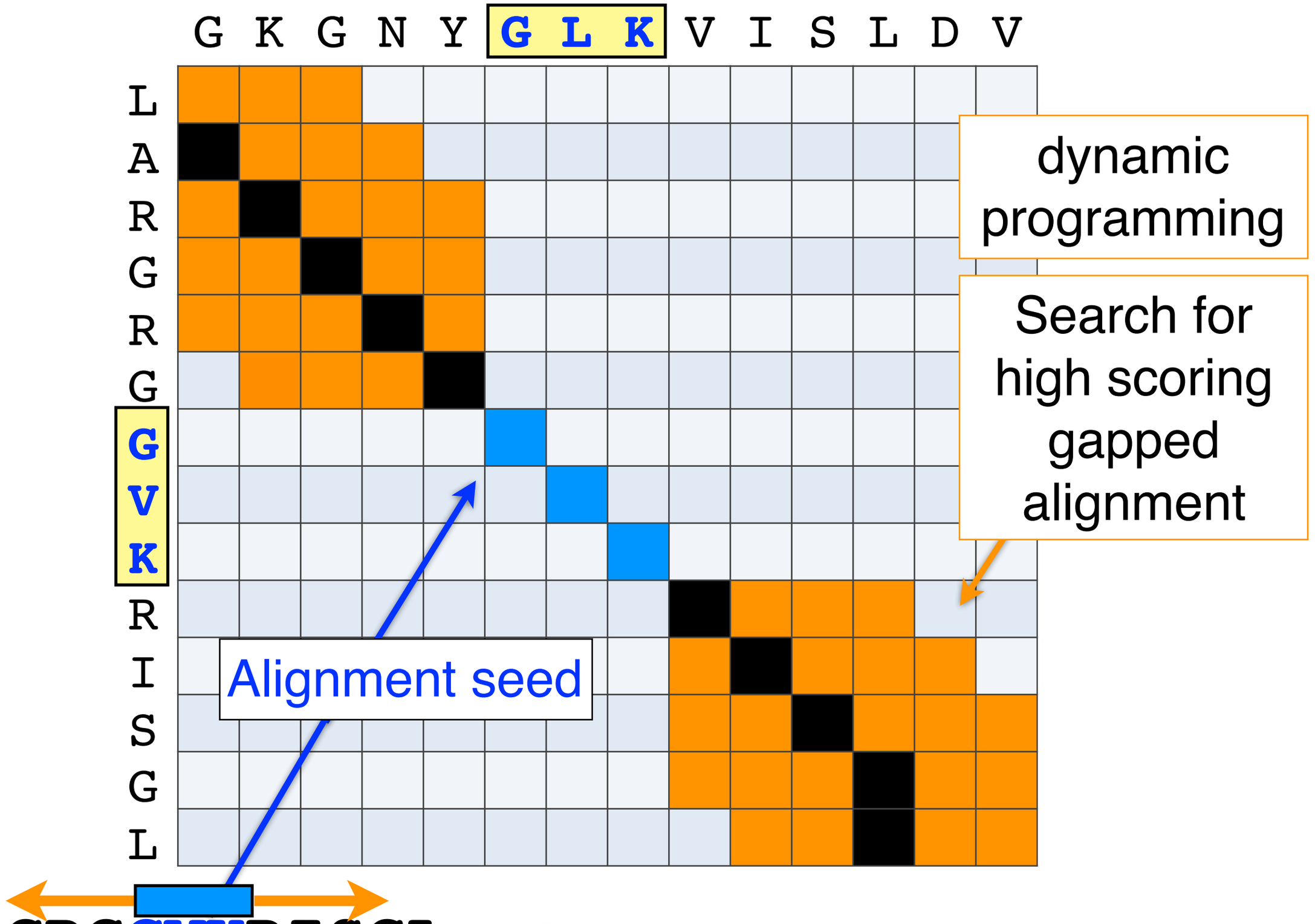
Query sequence





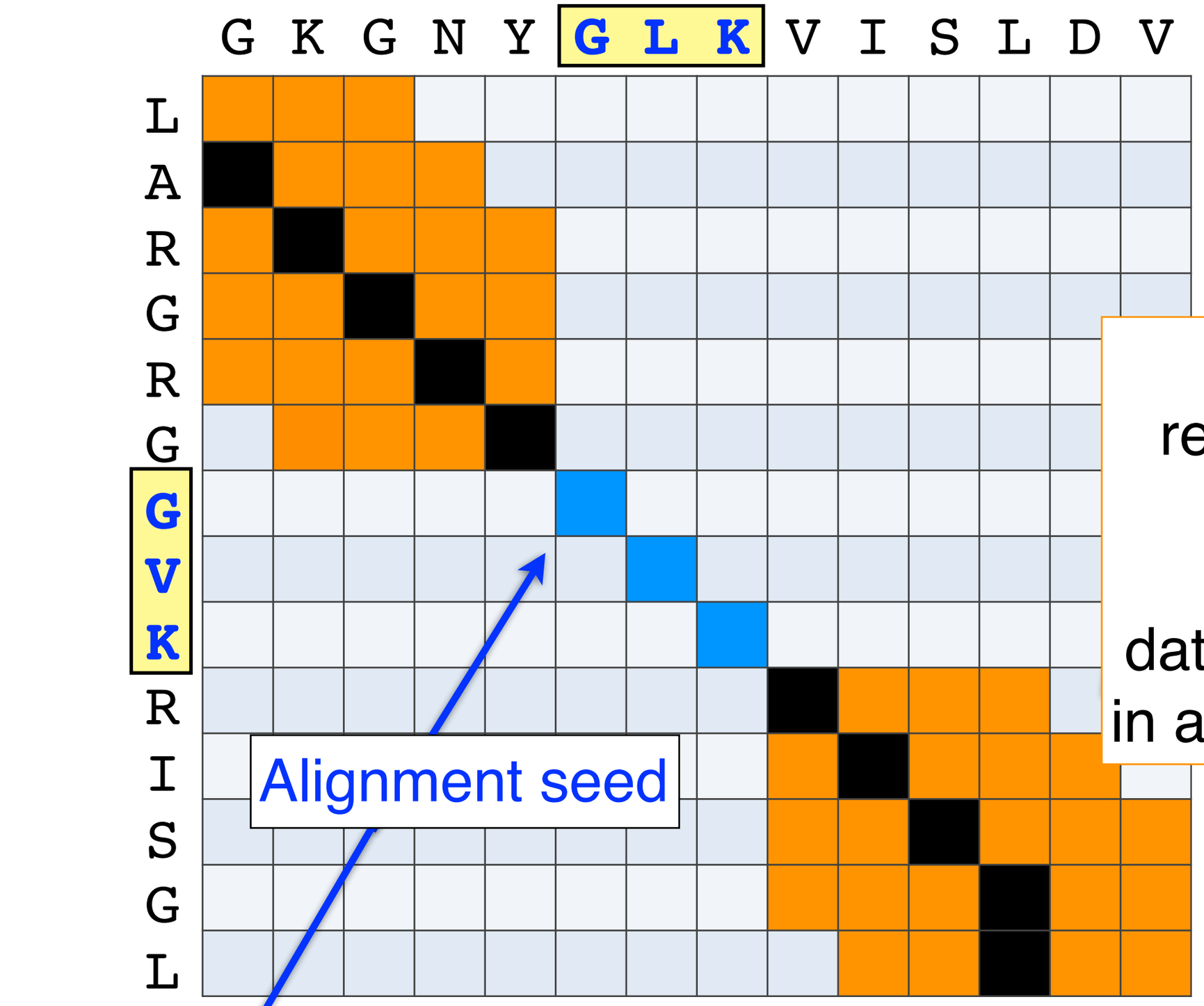
GRGGVKRISGL Query sequence

GNYGLKVISLDV Database sequence



GRGGVKRISGL Query sequence

GNYGLKVISLDV Database sequence



BLAST returns the highest scoring database hits in a ranked list

GRGGVKRISGL
GNYGLKVIS-L

Query sequence

Database sequence

BLAST output

- BLAST returns the highest scoring database hits in a ranked list along with details about the target sequence and alignment statistics

Description	Max score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo sapiens]	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	100%	0	98%	AAA20133.1
Kinesin-14 heavy chain [Danio rerio]	595	88%	0	78%	XP_00320703
hypothetical protein EGK_18589	48.2	40%	0.03	32%	ELK35081.1
mKIAA4102 protein [Mus musculus]	42.7	38%	3.02	24%	EHH28205.1

Statistical significance of results

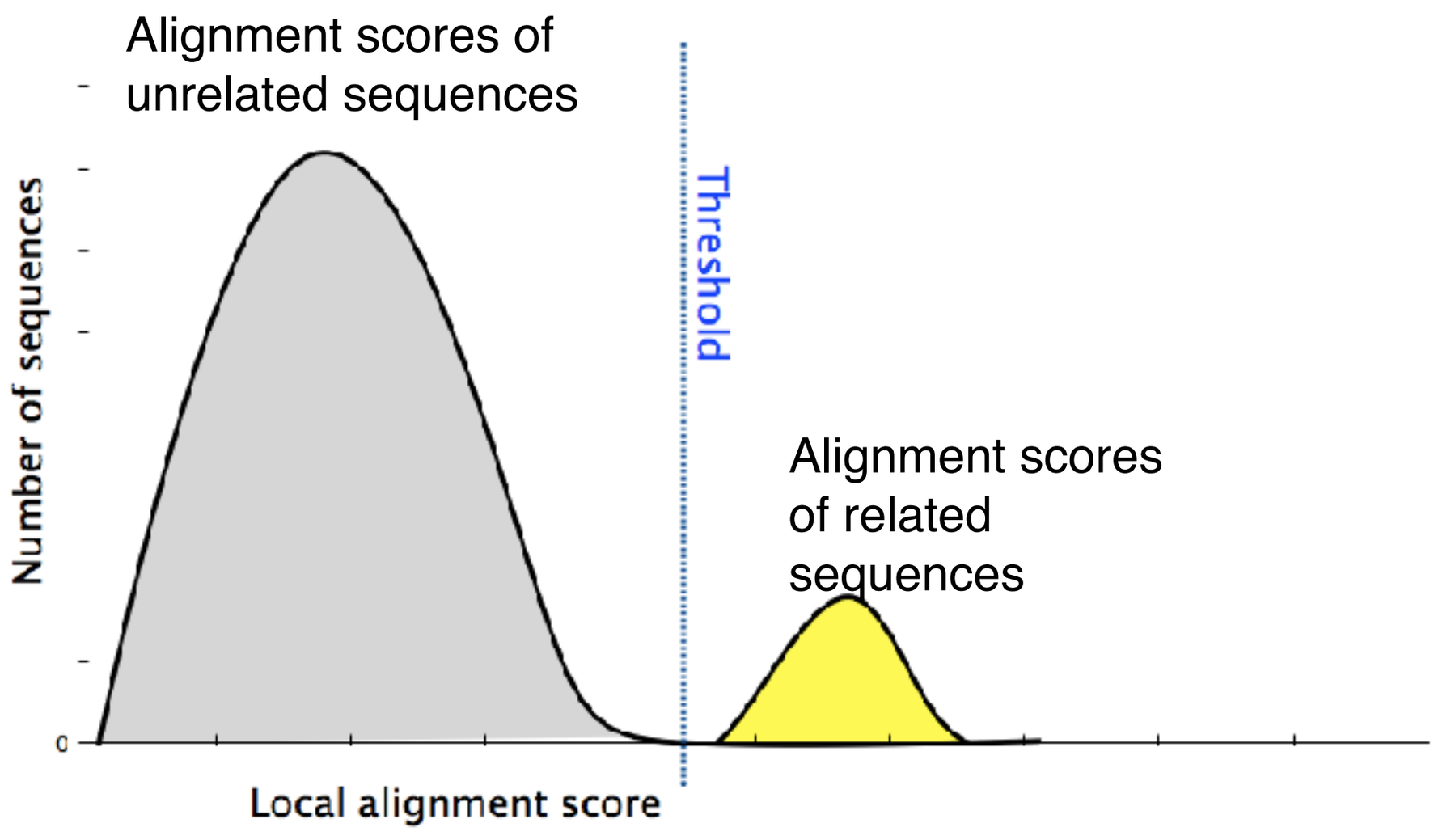
- An important feature of BLAST is the computation of statistical significance for each hit. This is described by the **E value** (expect value)

Description	Max score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo sapiens]	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	100%	0	98%	AAA20133.1
Kinesin-14 heavy chain [Danio rerio]	595	88%	0	78%	XP_00320703
hypothetical protein EGK_18589	48.2	40%	0.03	32%	ELK35081.1
mKIAA4102 protein [Mus musculus]	42.7	38%	3.02	24%	EHH28205.1

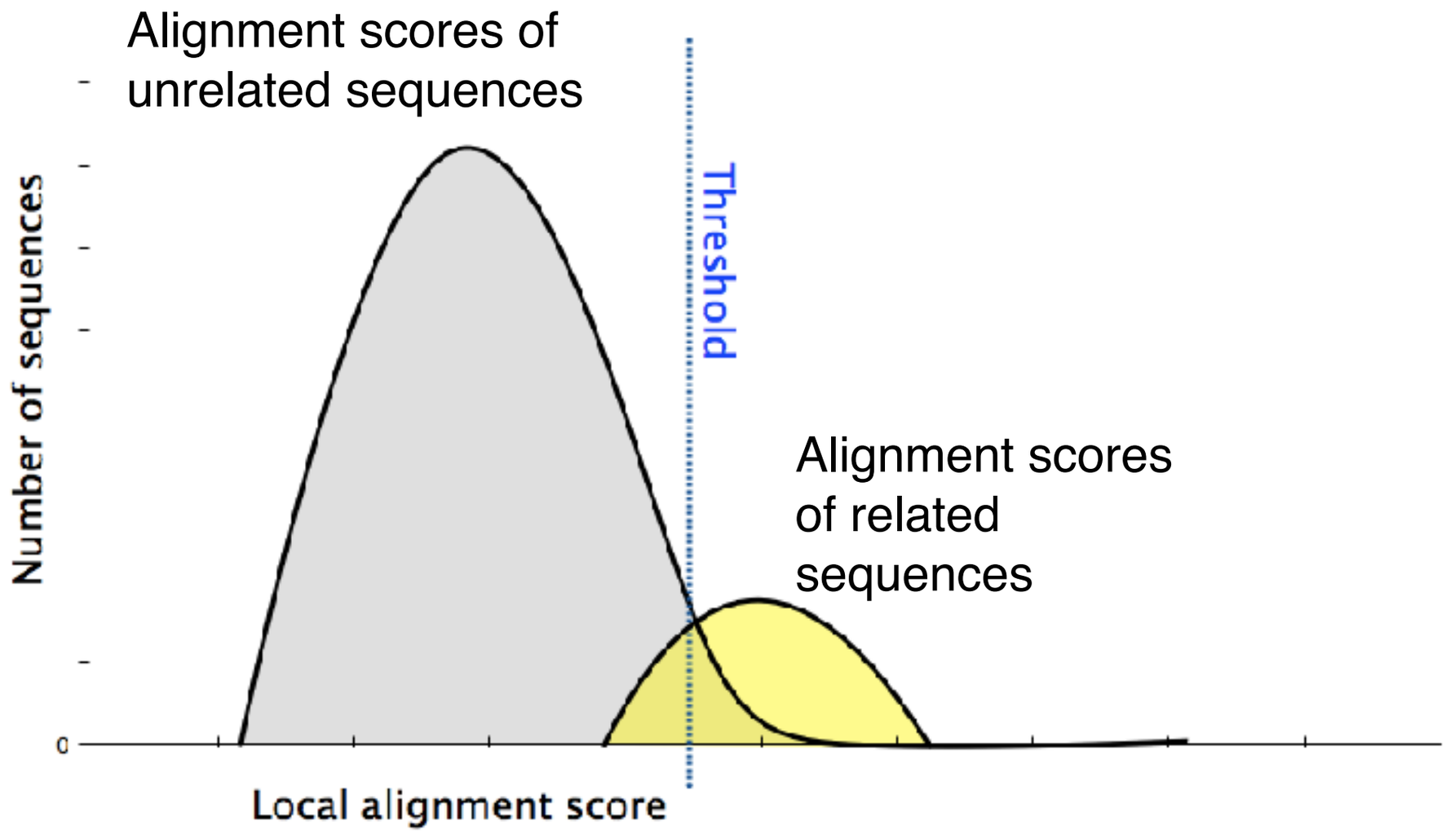
BLAST scores and E-values

- The **E value** is the **expected** number of hits that are as good or better than the observed local alignment score (with this score or better) if the query and database are **random** with respect to each other
  - *i.e.* the number of alignments expected to occur by chance with equivalent or better scores
- Typically, only hits with E value **below** a significance threshold are reported
  - This is equivalent to selecting alignments with score above a certain score threshold

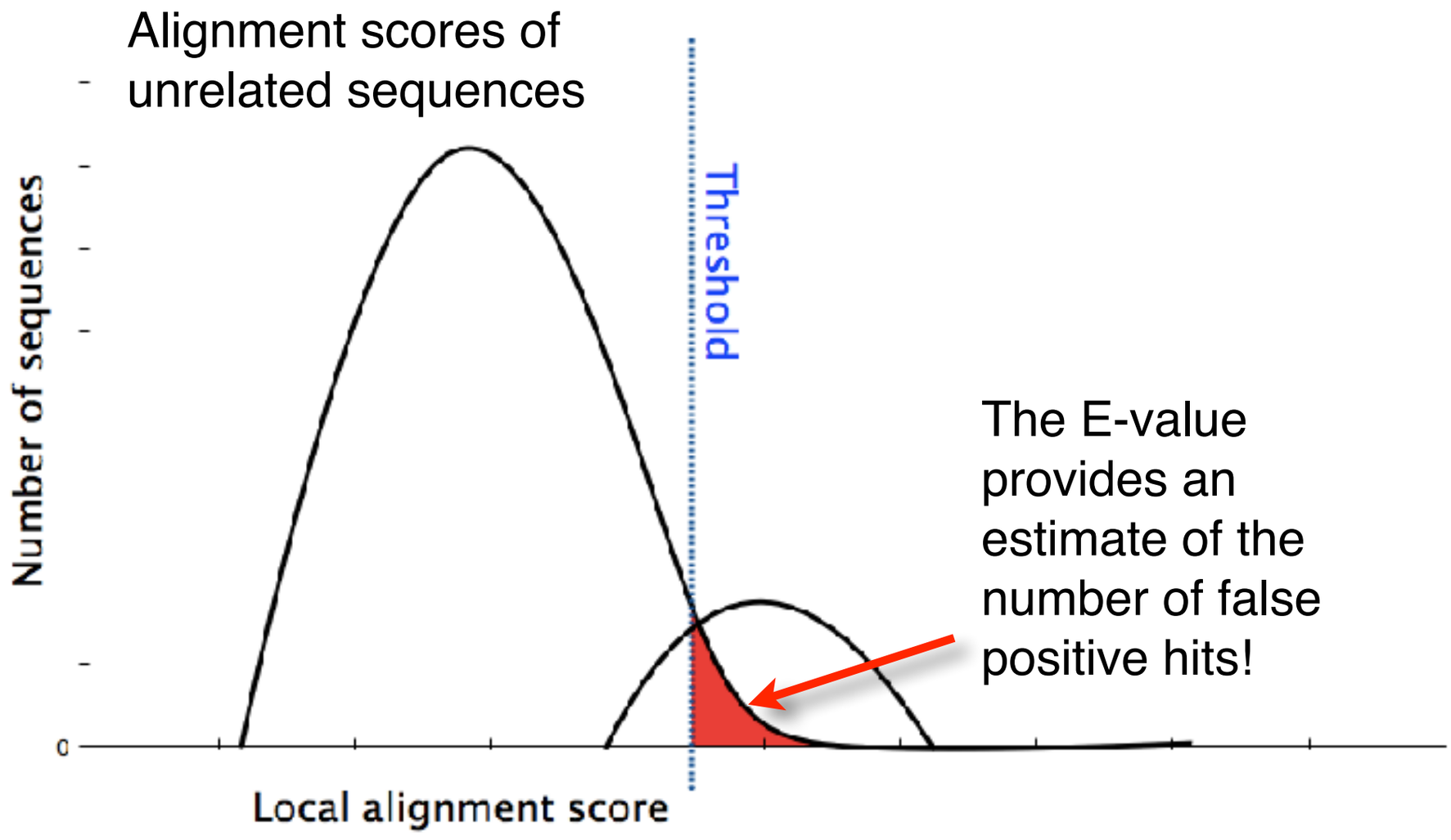
- Ideally, a threshold separates all query related sequences (yellow) from all unrelated sequences (gray)



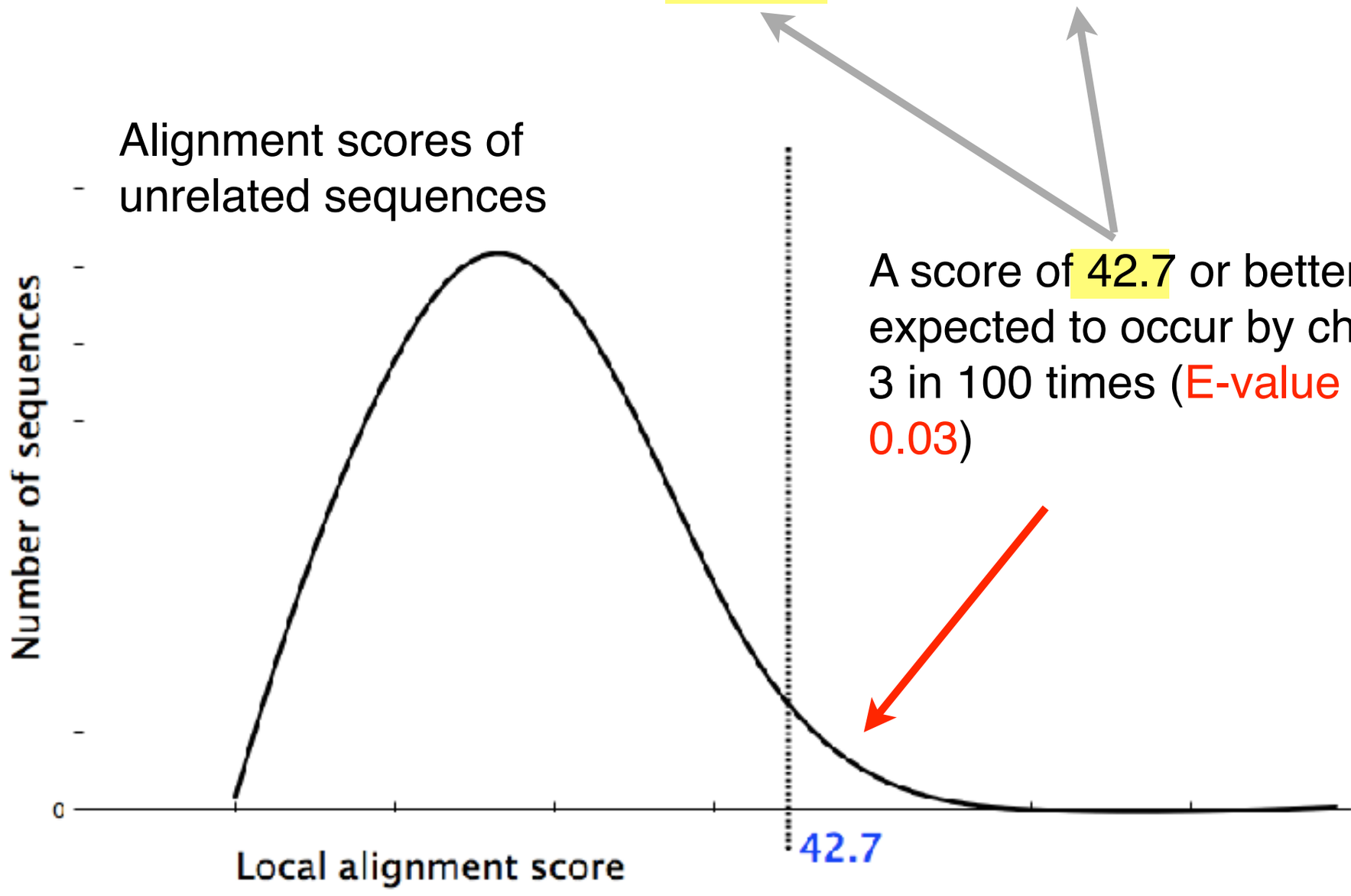
- Unfortunately, often both score distributions overlap
  - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



- Unfortunately, often both score distributions overlap
  - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



Description	Max score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo sapiens]	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	100%	0	98%	AAA20133.1
Kinesin-14 heavy chain [Danio rerio]	595	88%	0	78%	XP_00320703
hypothetical protein EGK_18589	42.7	40%	0.03	32%	ELK35081.1



Alignment scores of unrelated sequences

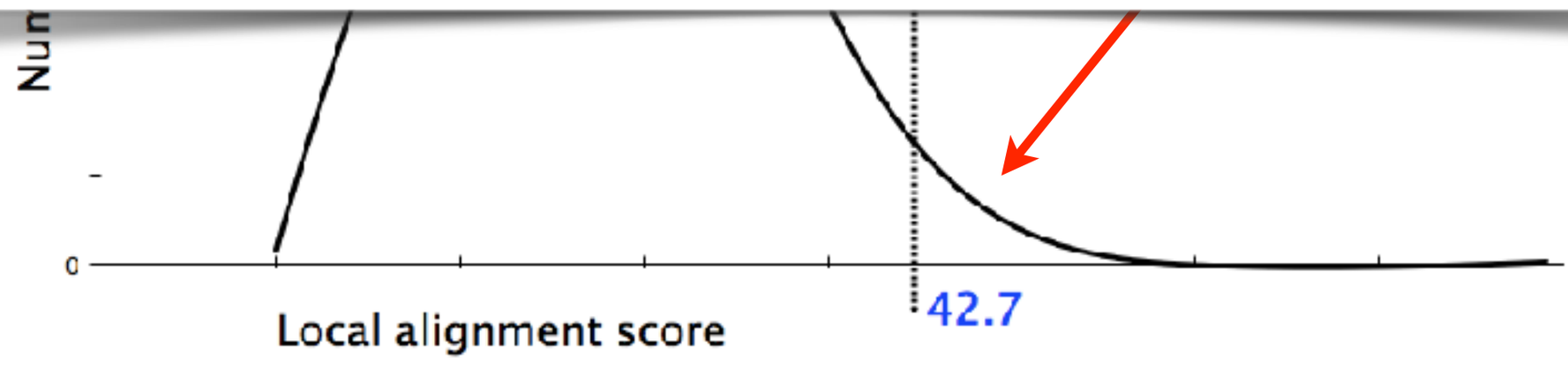
A score of 42.7 or better is expected to occur by chance 3 in 100 times (E-value = 0.03)

Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo	677	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	676	100%	0	98%	AAA20133.1

In general *E* values < 0.005 are usually significant.

To find out more about *E* values see: "*The Statistics of Sequence Similarity Scores*" available in the help section of the NCBI BLAST site:

http://www.ncbi.nlm.nih.gov/blast/tutorial/Altschul-1.html

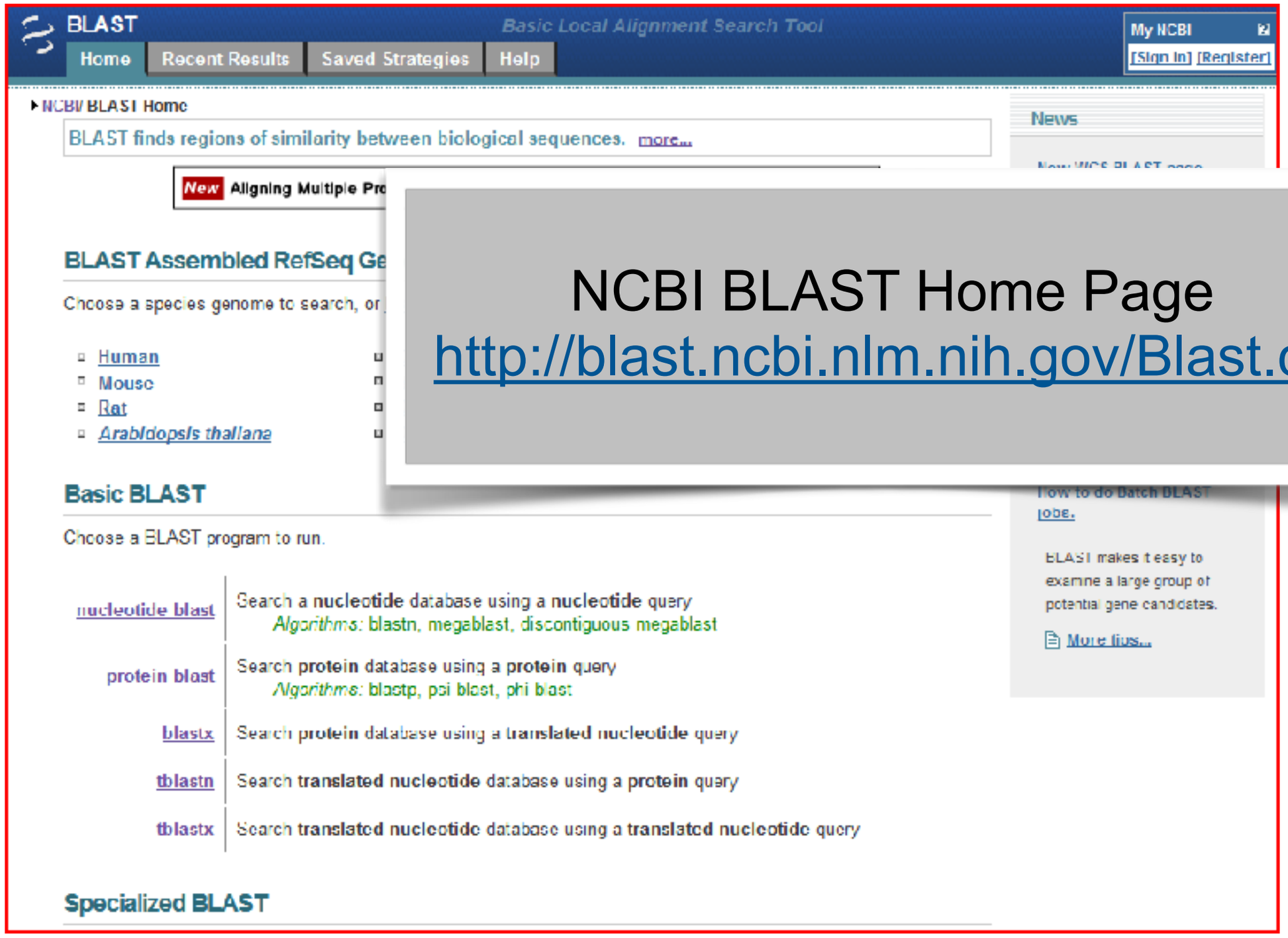


Your Turn!

Hands-on worksheet **Sections 4** & **5**

- Please do answer the last lab review question (**Q19**).
- We encourage discussion and exploration!

Practical database searching with BLAST



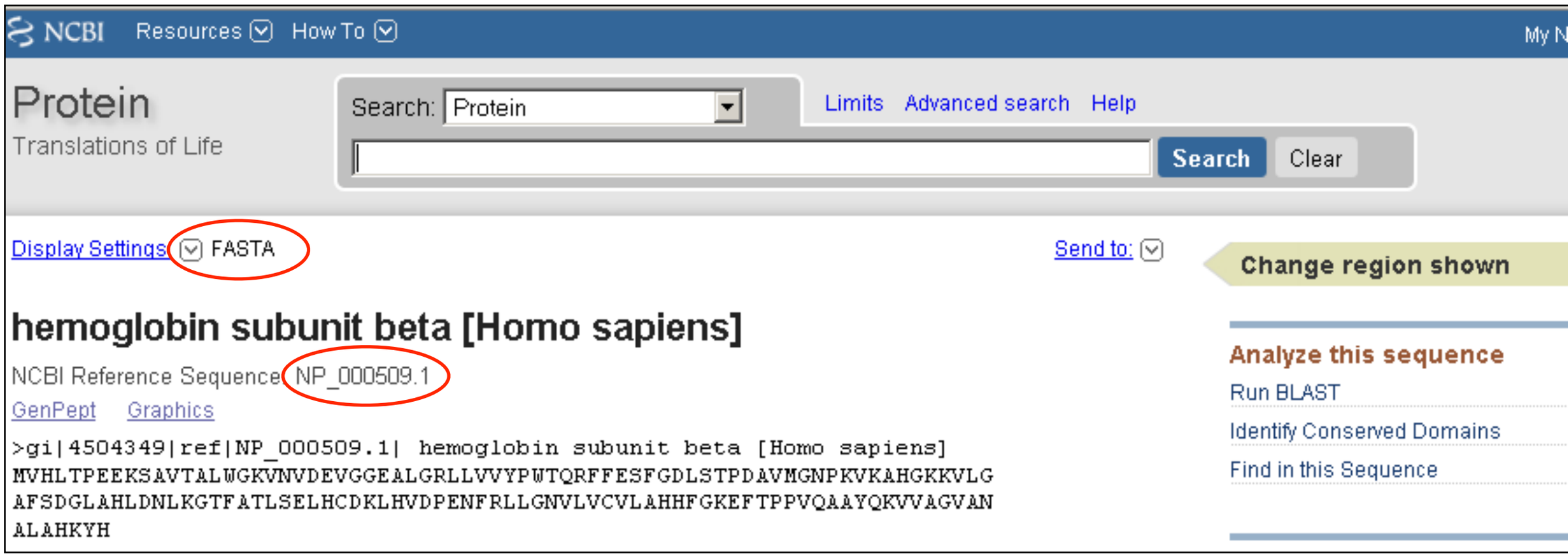
NCBI BLAST Home Page
http://blast.ncbi.nlm.nih.gov/Blast.cgi

Practical database searching with BLAST

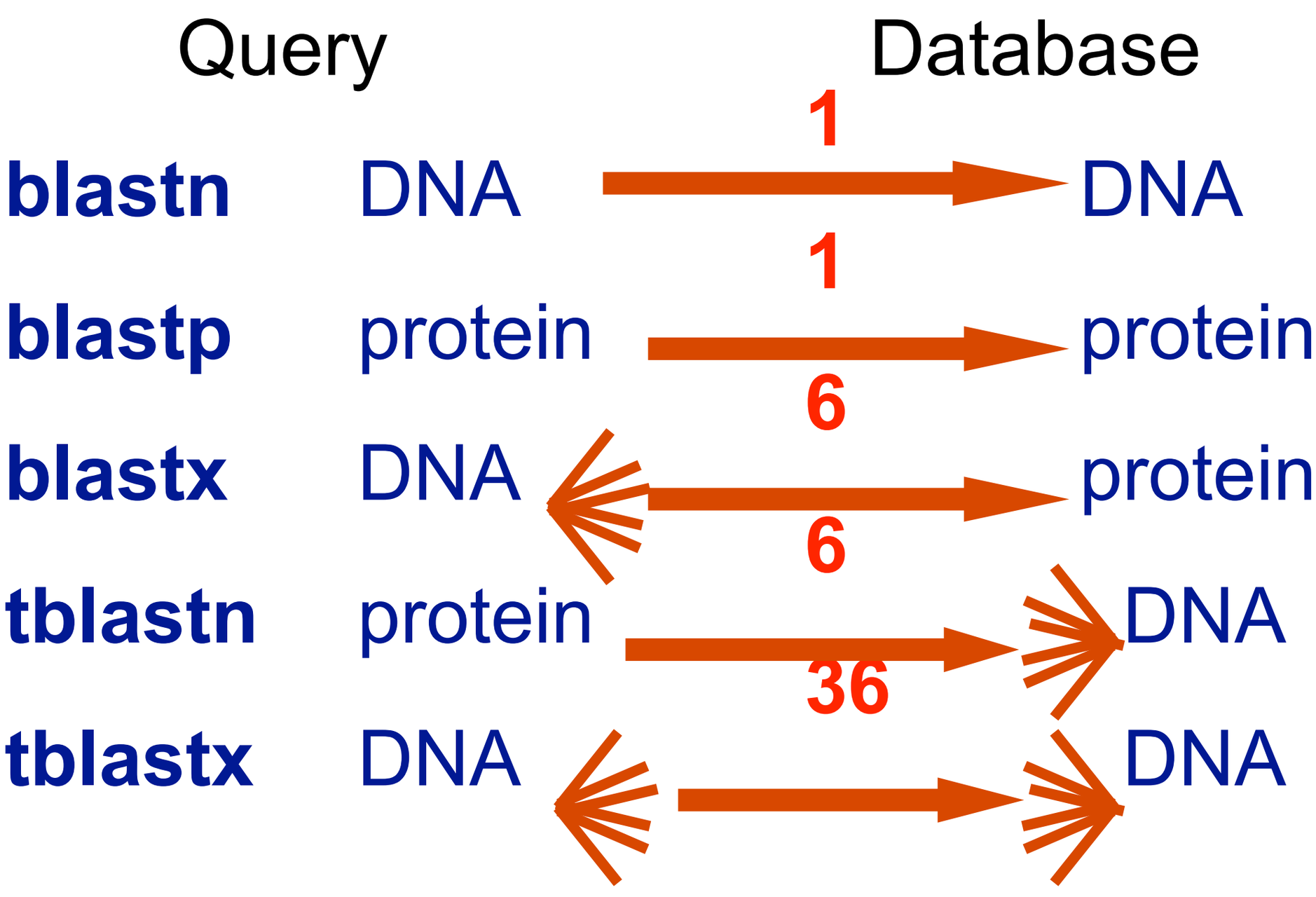
- There are four basic components to a traditional BLAST search
  - (1) Choose the sequence (query)
  - (2) Select the BLAST program
  - (3) Choose the database to search
  - (4) Choose optional parameters
- Then click “BLAST”

Step 1: Choose your sequence

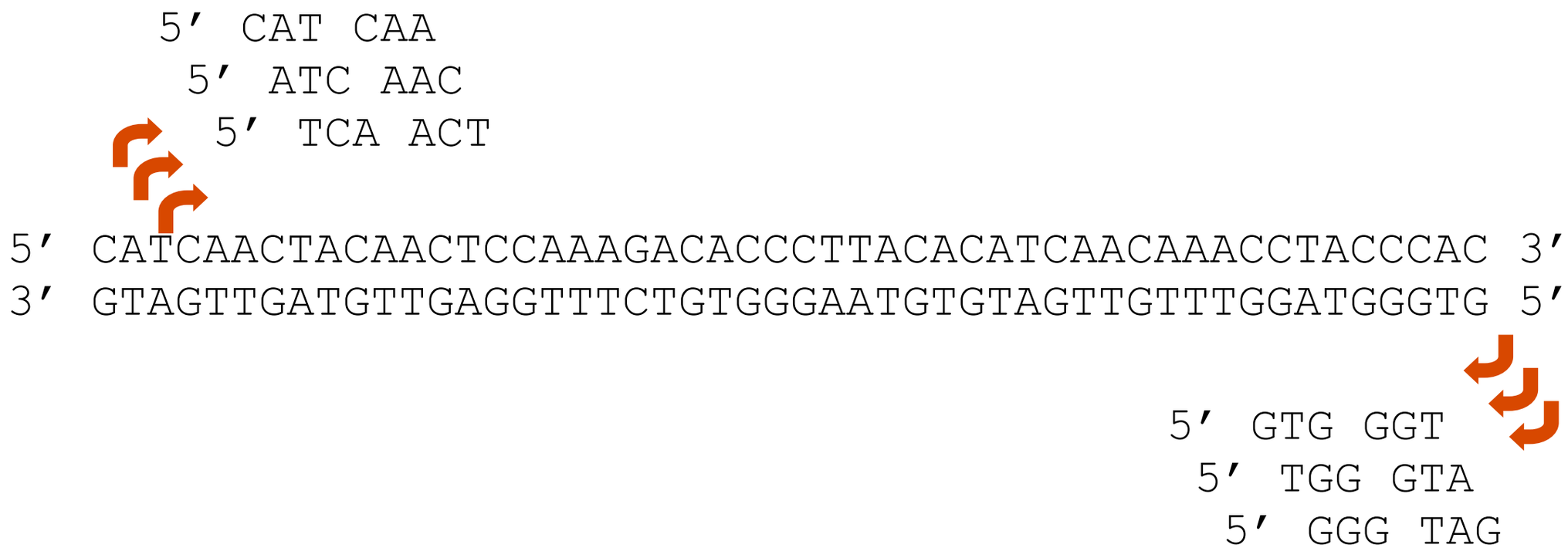
- Sequence can be input in FASTA format or as accession number

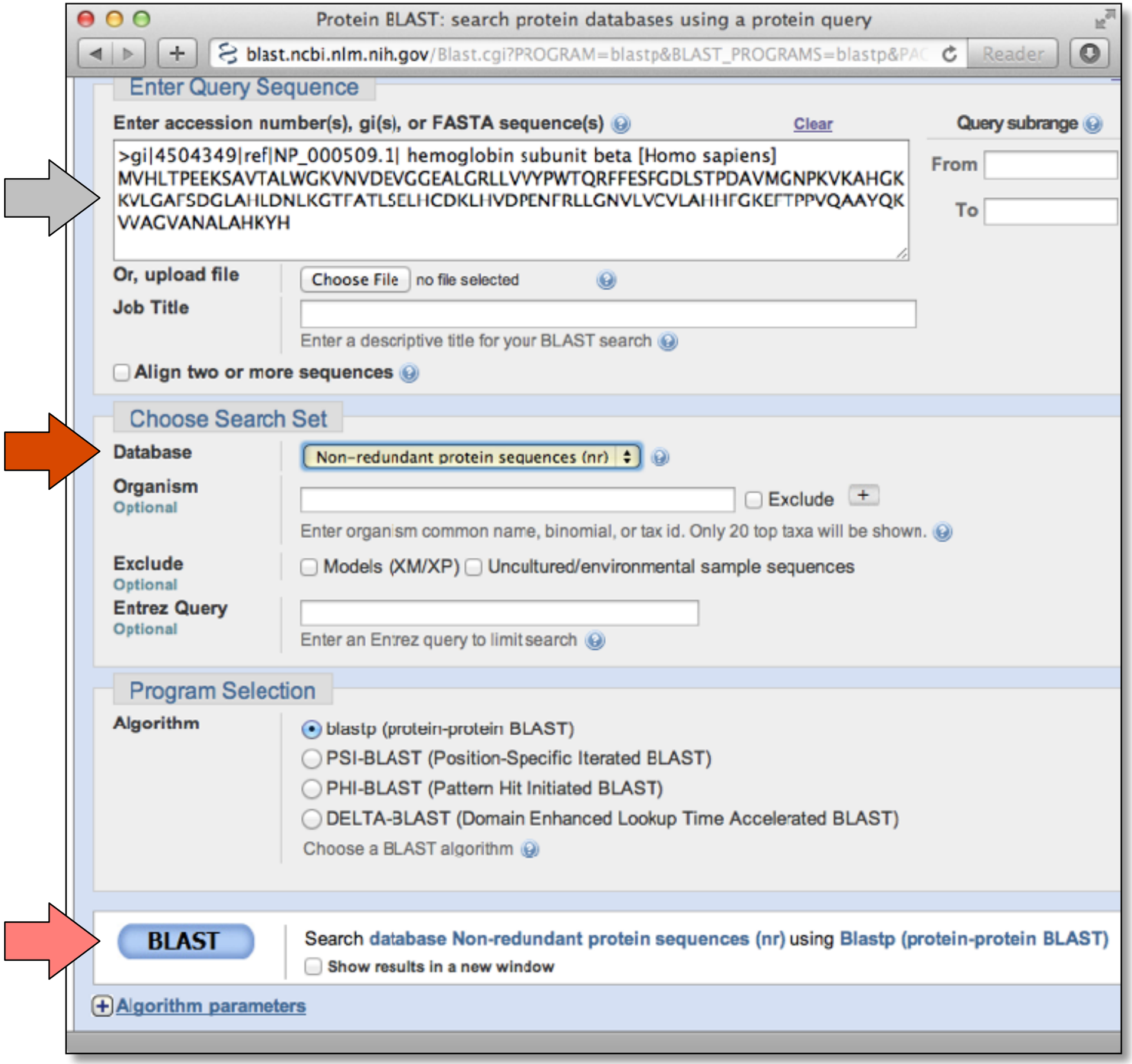


Step 2: Choose the BLAST program



DNA potentially encodes six proteins




Protein BLAST: search protein databases using a protein query
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PA
Reader
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGK
KVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQK
VVAGVANALAHKYH
Or, upload file
Choose File
no file selected
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
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
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
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters

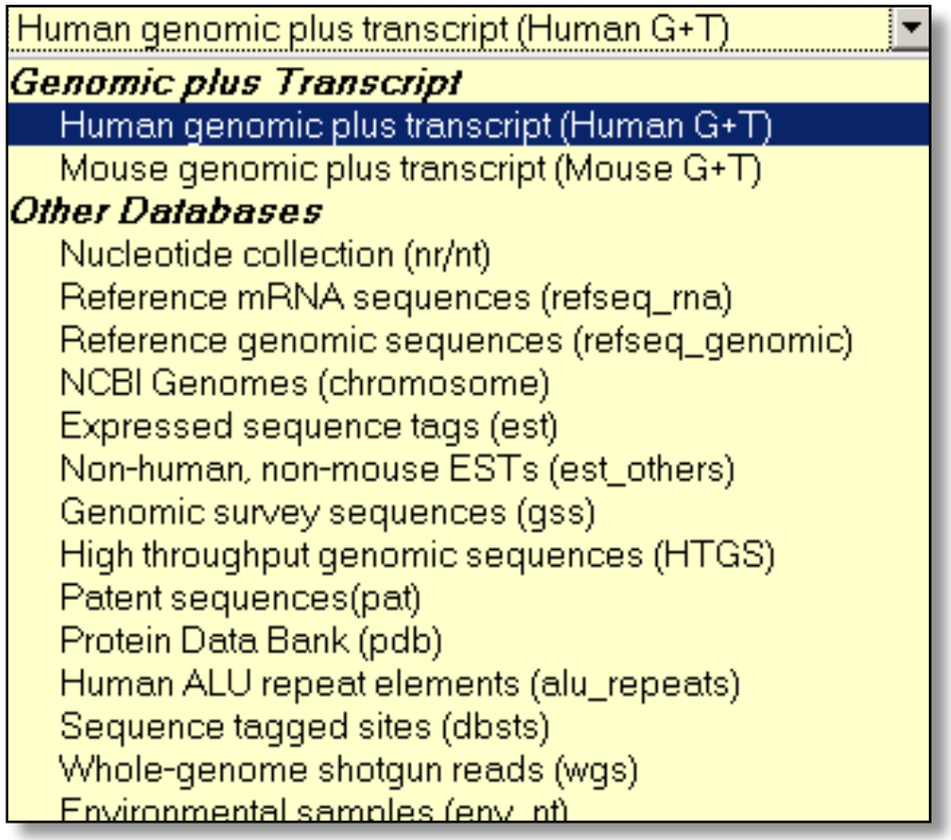
Step 3: Choose the database

nr = non-redundant (most general database)

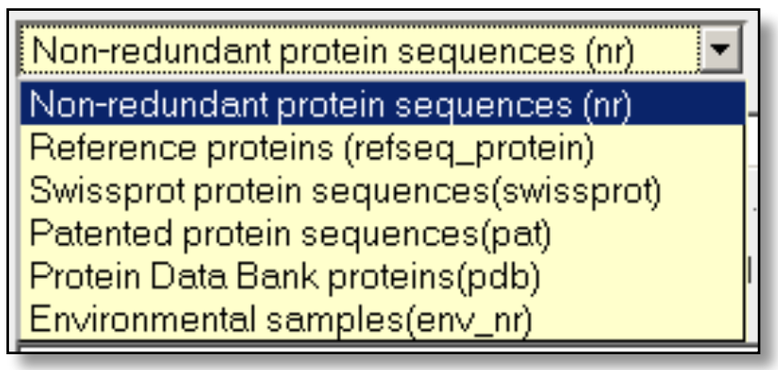
dbest = database of expressed sequence tags

dbsts = database of sequence tag sites

gss = genomic survey sequences



nucleotide databases

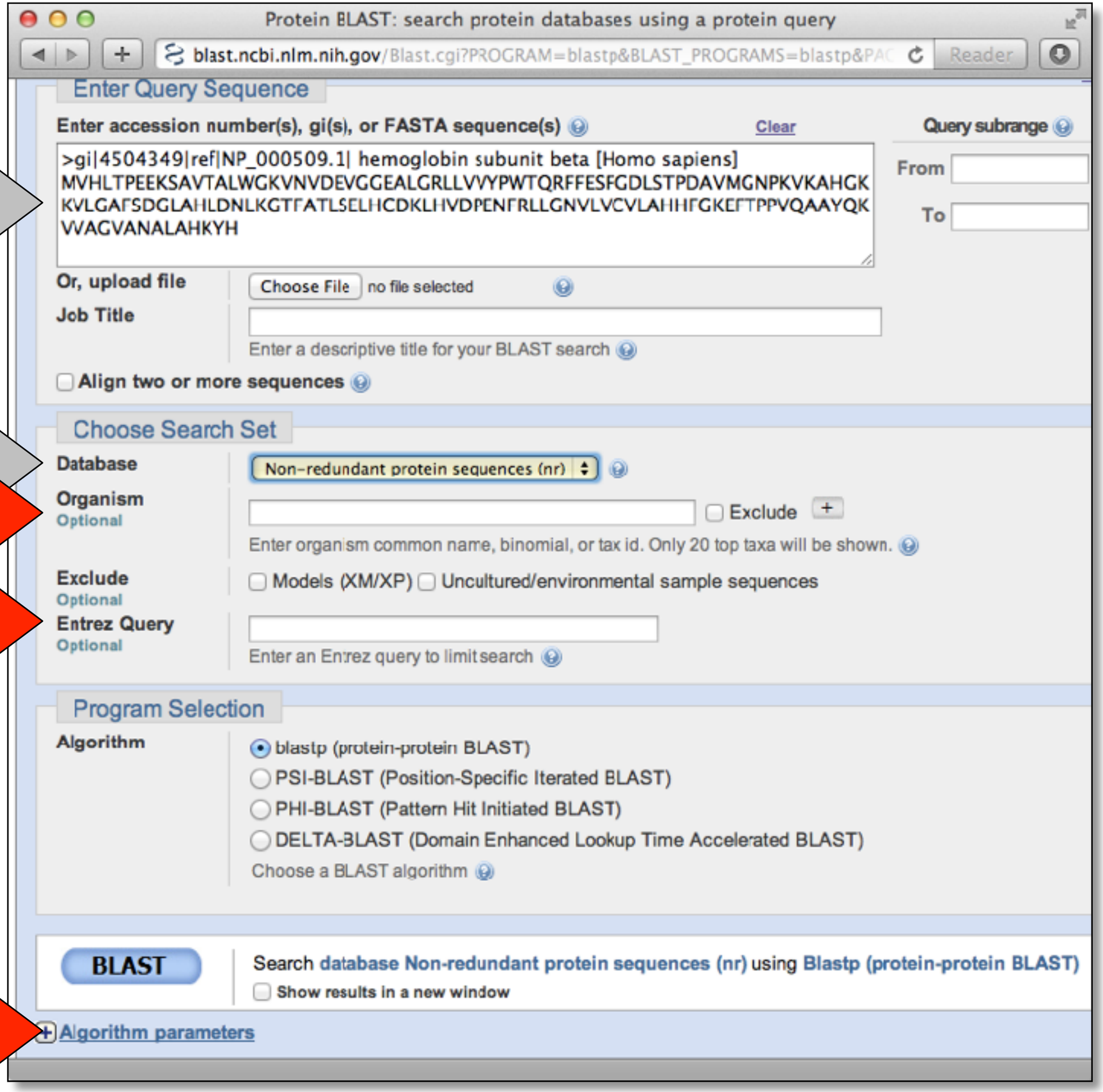


protein databases

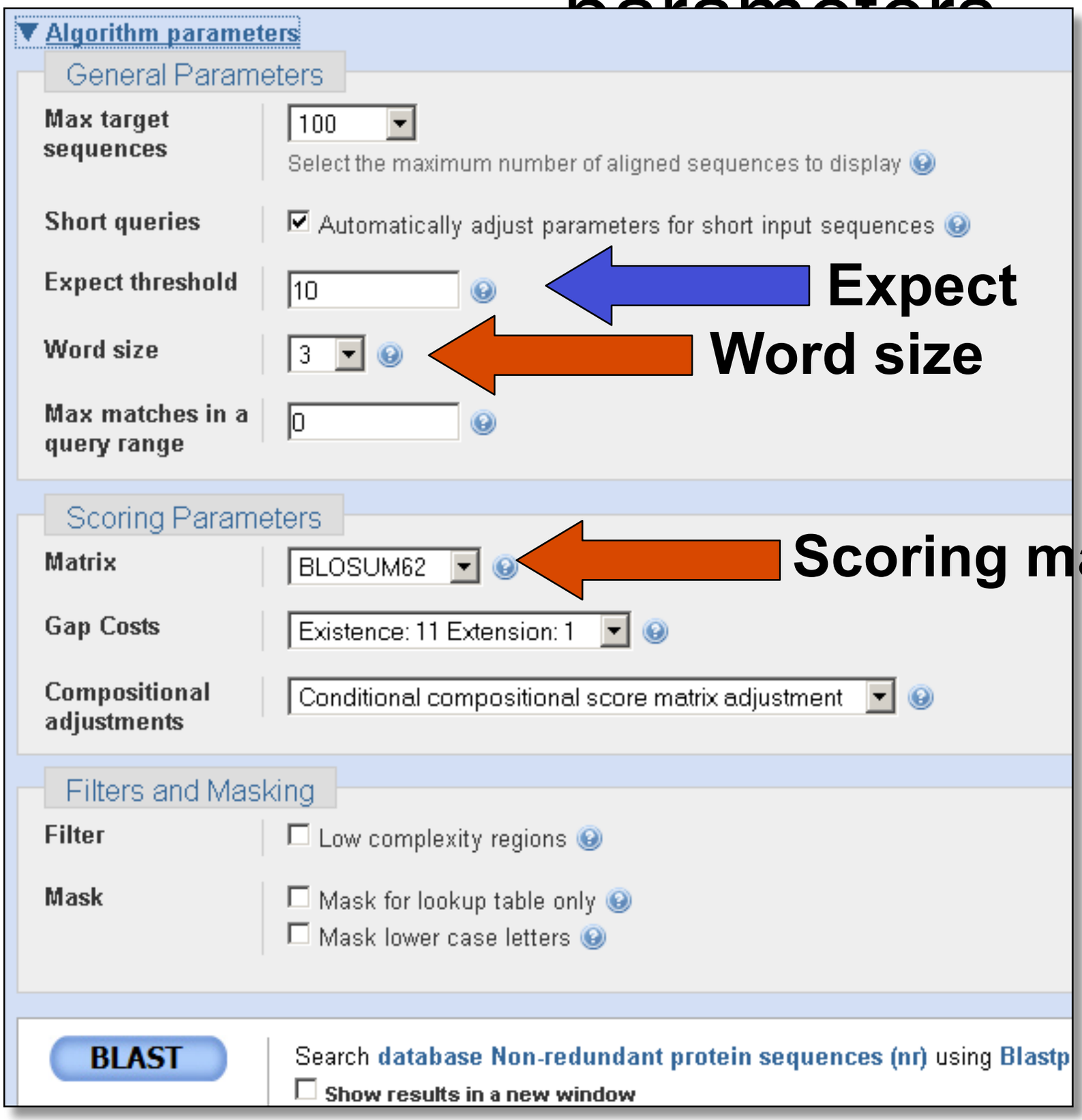
Organism

Entrez

Settings!



Step 4a: Select optional search parameters



Step 4: Optional parameters

- You can...
  - choose the organism to search
  - change the substitution matrix
  - change the expect (E) value
  - change the word size
  - change the output format

Results page

NCBI Blast:gi|4504349|ref|NP_000509.1| hemoglobin

blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - FVGUTMRZ013

Edit and Resubmit | Save Search Strategies | ▷Formatting options | ▷Download | Change the result display back to traditional format

YouTube Learn about the enhanced report | Blast report description

gi|4504349|ref|NP_000509.1| hemoglobin

Query ID	lcl|84677
Description	gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
Molecule type	amino acid
Query Length	147

Database Name	nr
Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program	BLASTP 2.2.27+ ▷Citation

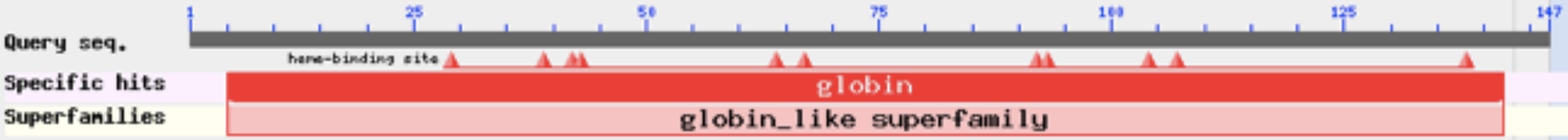
Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment]

New DELTA-BLAST, a more sensitive protein-protein search Go

Graphic Summary

Show Conserved Domains

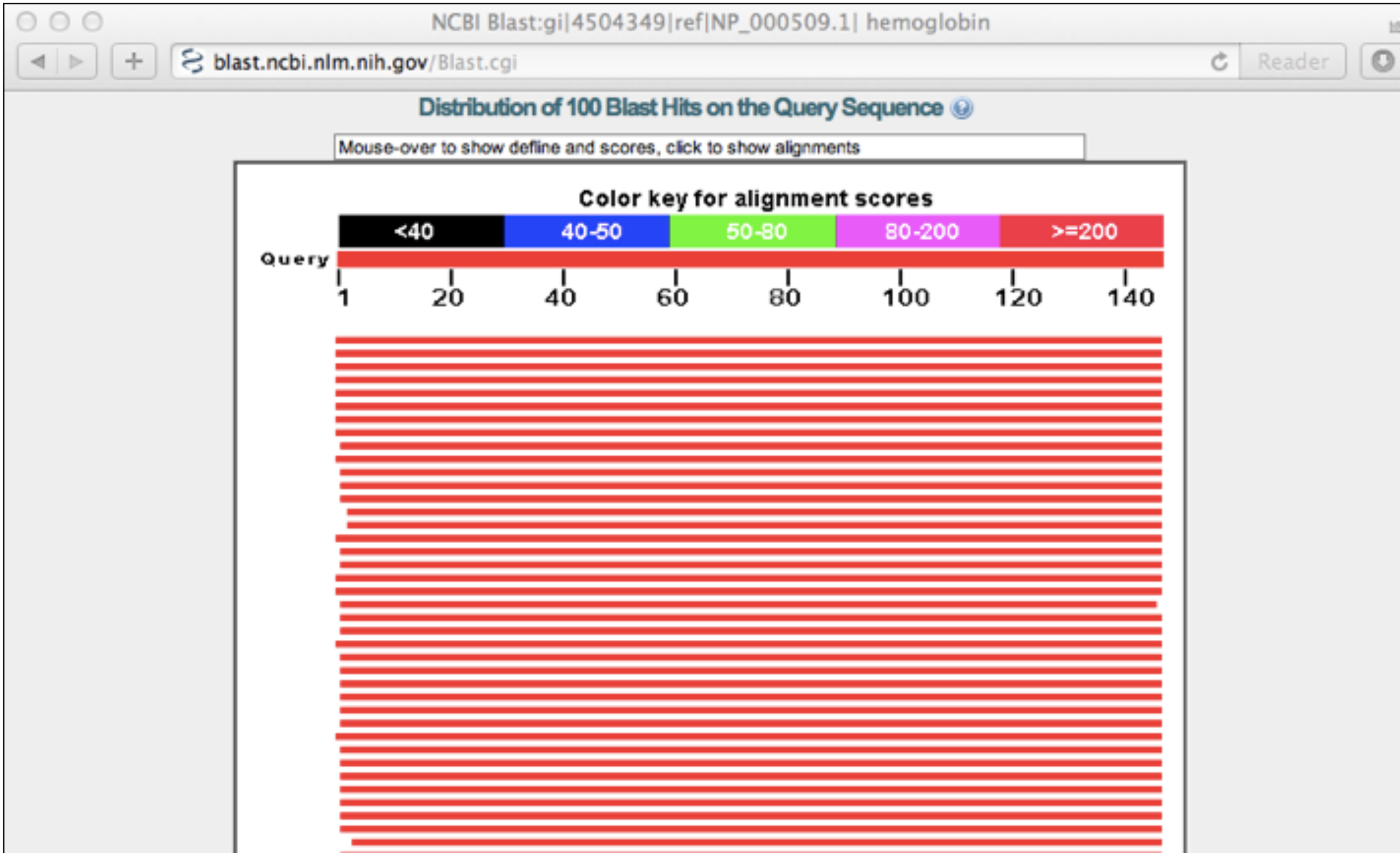
Putative conserved domains have been detected, click on the image below for detailed results.



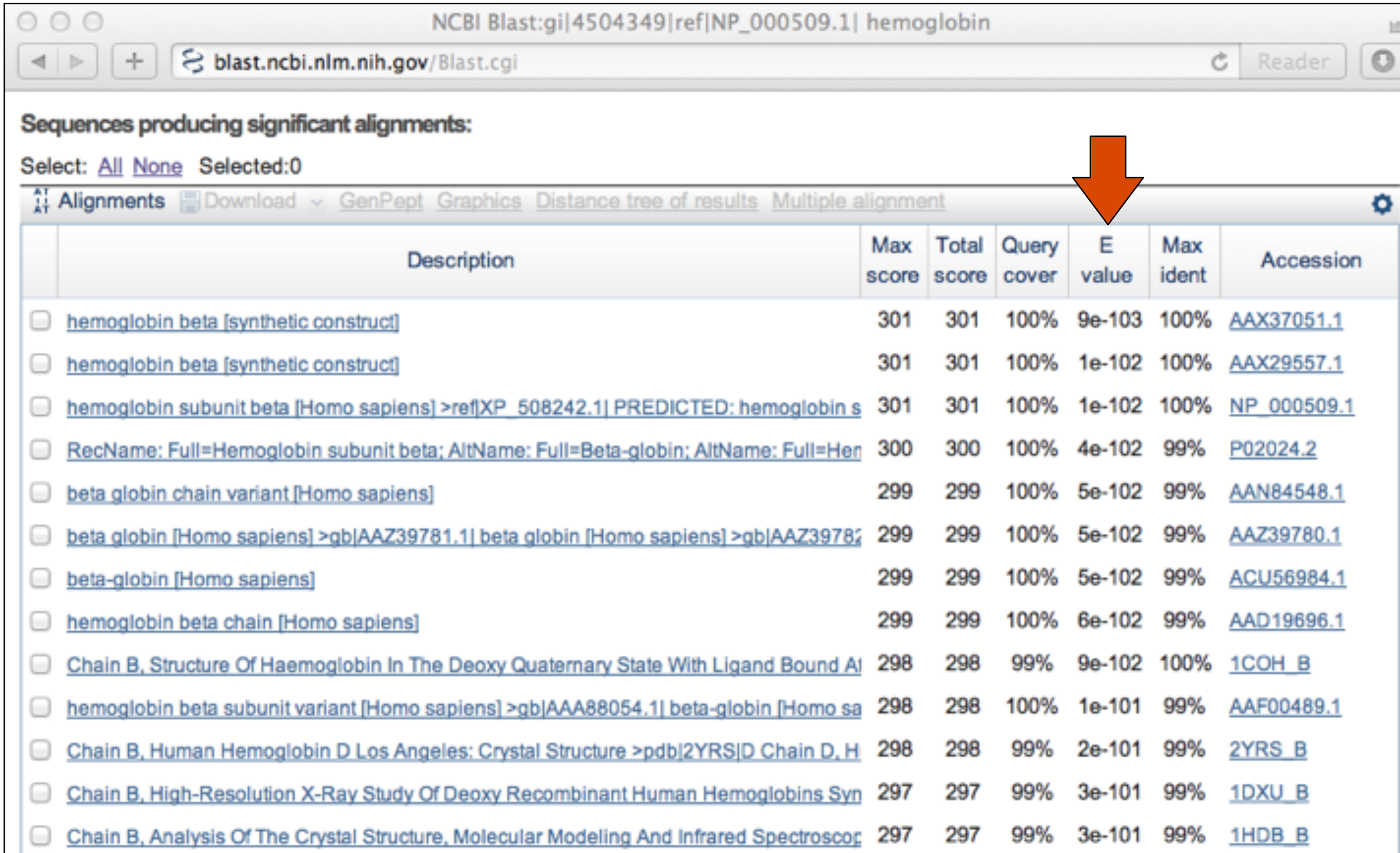
Distribution of 100 Blast Hits on the Query Sequence

Mouse over to show defline and scores, click to show alignments

Further down the results page...

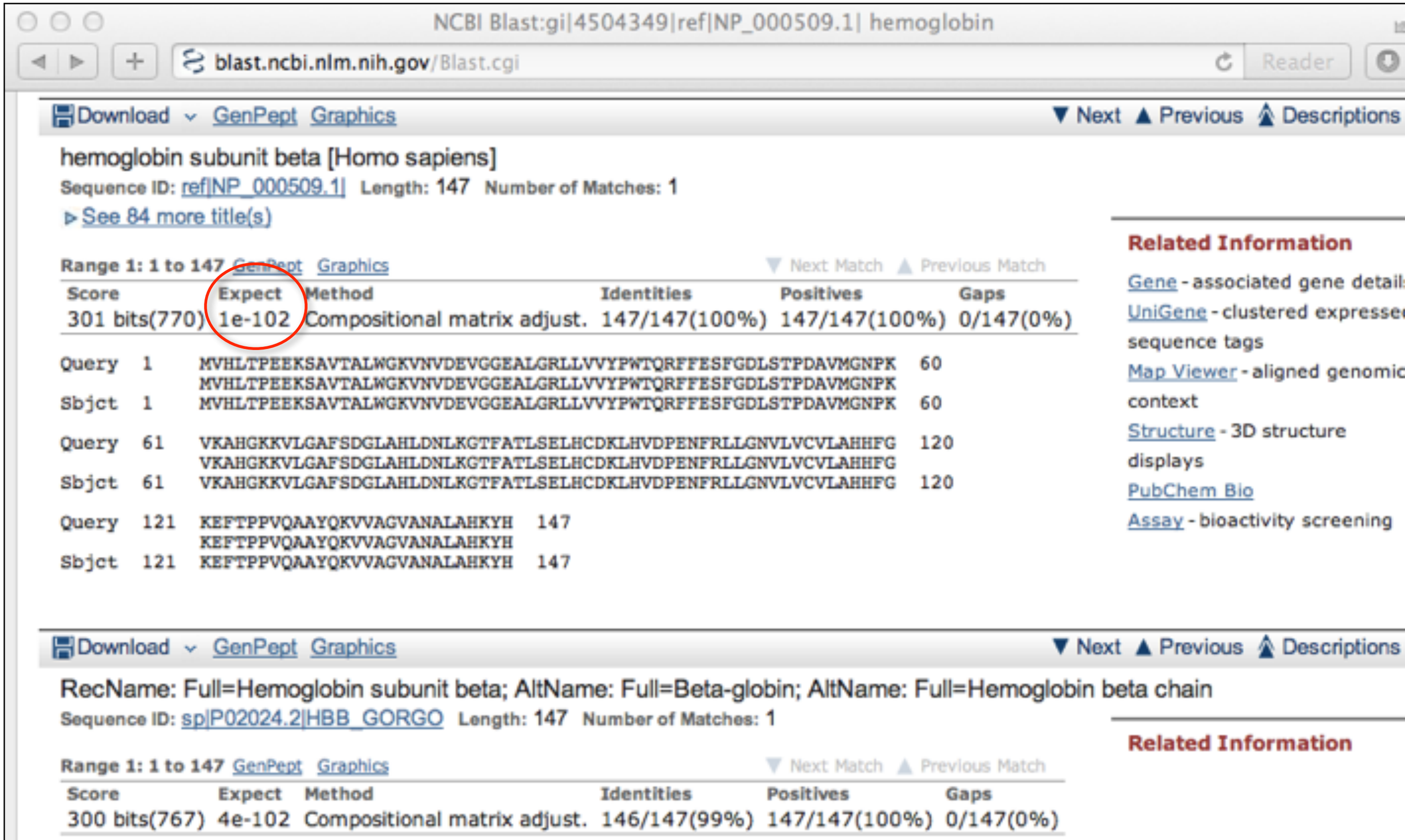


Further down the results page...

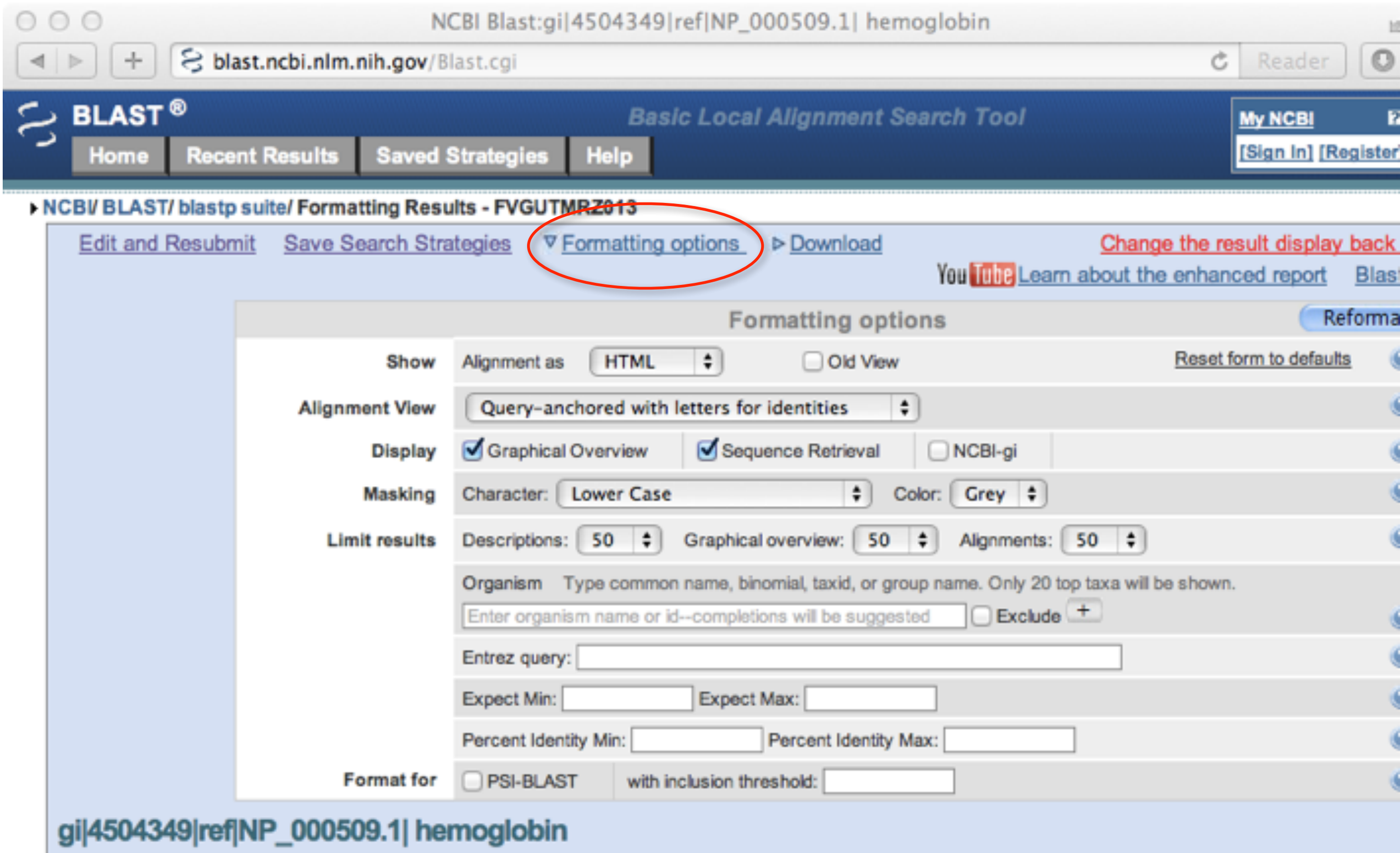


Description	Max score	Total score	Query cover	E value	Max ident	Accession
hemoglobin beta [synthetic construct]	301	301	100%	9e-103	100%	AAX37051.1
hemoglobin beta [synthetic construct]	301	301	100%	1e-102	100%	AAX29557.1
hemoglobin subunit beta [Homo sapiens] >ref|XP_508242.1| PREDICTED: hemoglobin s	301	301	100%	1e-102	100%	NP_000509.1
RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Her	300	300	100%	4e-102	99%	P02024.2
beta globin chain variant [Homo sapiens]	299	299	100%	5e-102	99%	AAN84548.1
beta globin [Homo sapiens] >gb|AAZ39781.1| beta globin [Homo sapiens] >gb|AAZ3978	299	299	100%	5e-102	99%	AAZ39780.1
beta-globin [Homo sapiens]	299	299	100%	5e-102	99%	ACU56984.1
hemoglobin beta chain [Homo sapiens]	299	299	100%	6e-102	99%	AAD19696.1
Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound At	298	298	99%	9e-102	100%	1COH_B
hemoglobin beta subunit variant [Homo sapiens] >gb|AAA88054.1| beta-globin [Homo sa	298	298	100%	1e-101	99%	AAF00489.1
Chain B, Human Hemoglobin D Los Angeles: Crystal Structure >pdb|2YRS|D Chain D, H	298	298	99%	2e-101	99%	2YRS_B
Chain B, High-Resolution X-Ray Study Of Deoxy Recombinant Human Hemoglobins Syn	297	297	99%	3e-101	99%	1DXU_B
Chain B, Analysis Of The Crystal Structure, Molecular Modeling And Infrared Spectroscop	297	297	99%	3e-101	99%	1HDB_B

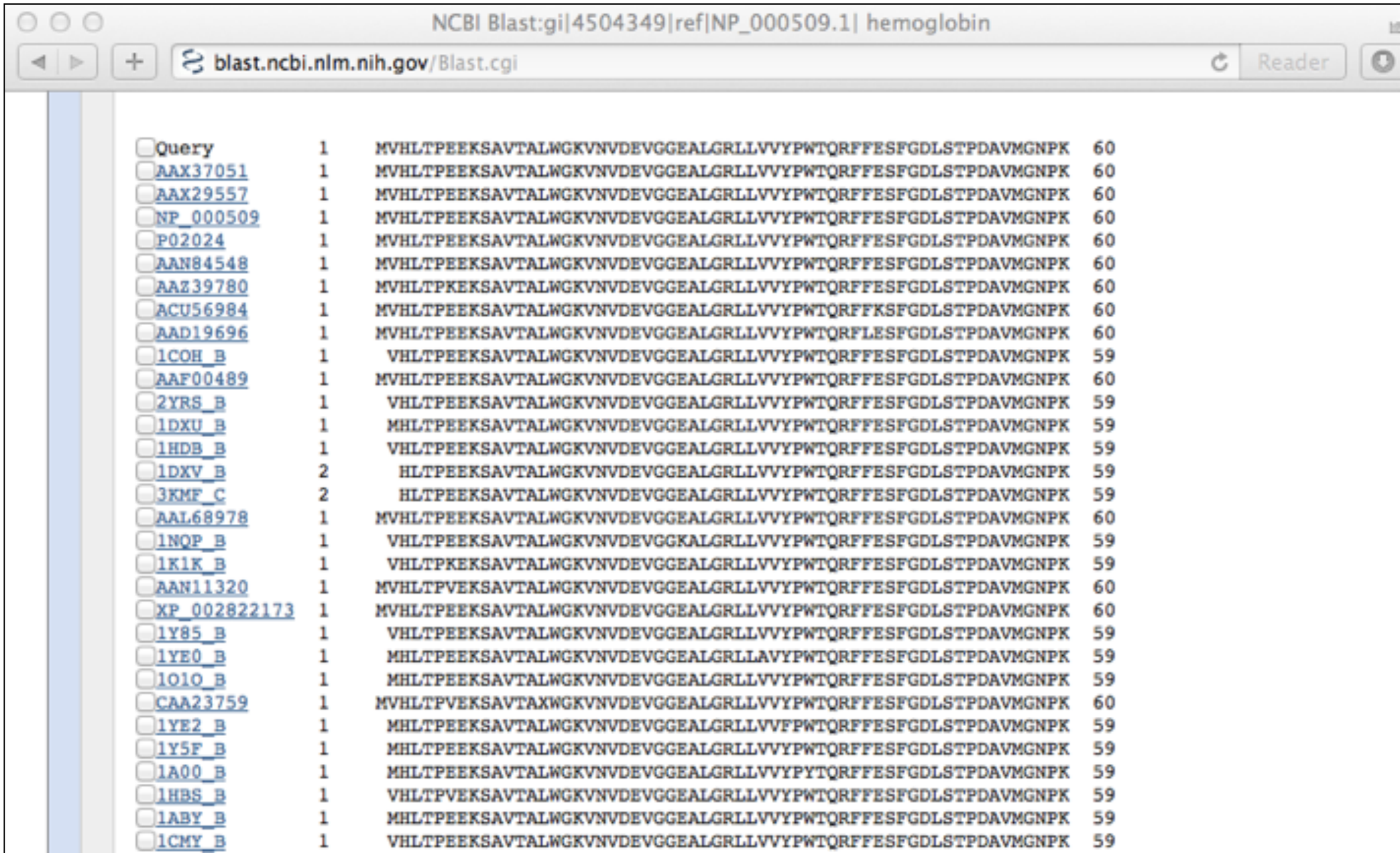
Further down the results page...



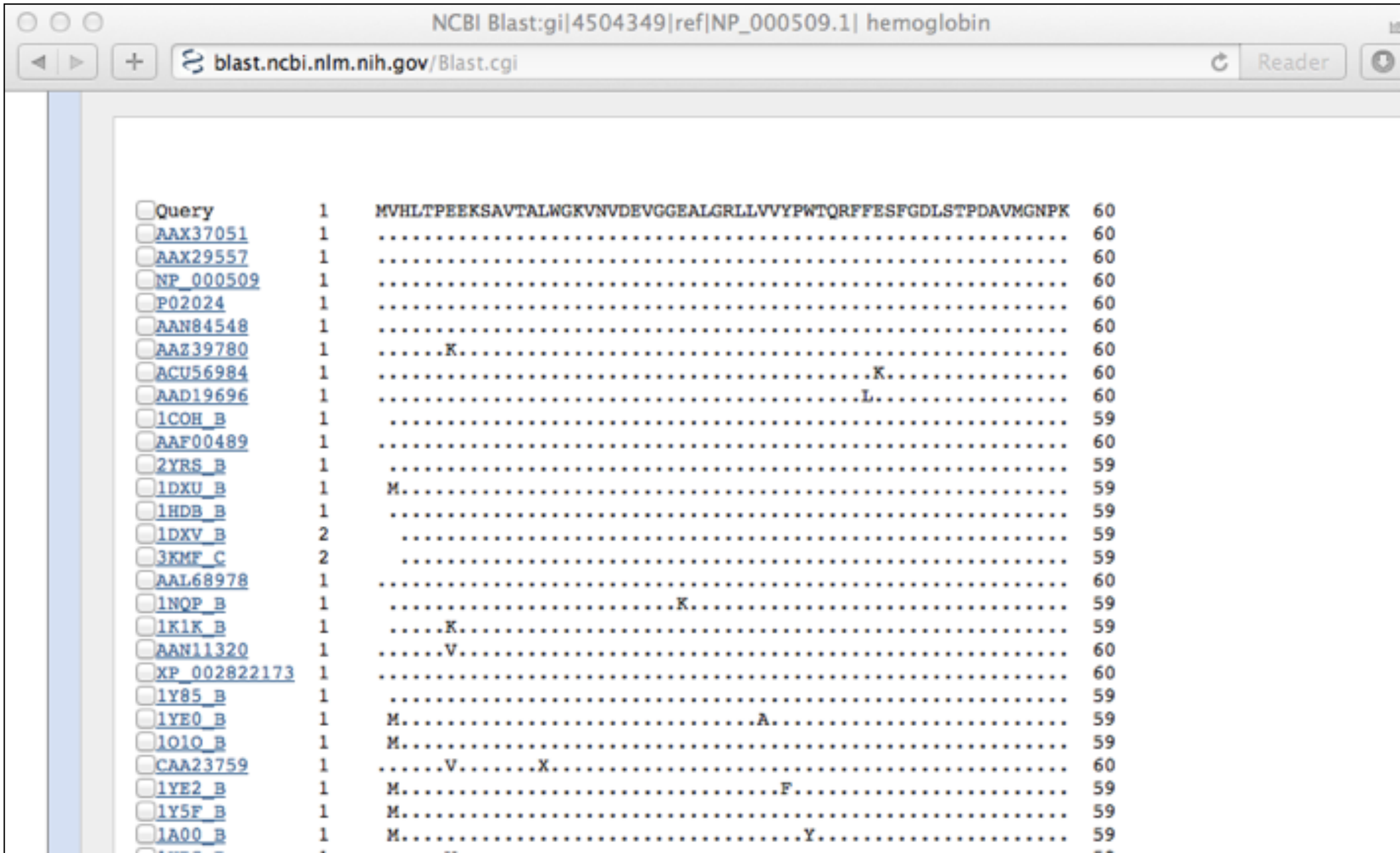
Different output formats are available



E.g. Query anchored alignments



... and alignments with dots for identities



Common problems

- Selecting the wrong version of BLAST
- Selecting the wrong database
- Too many hits returned
- Too few hits returned
- Unclear about the significance of a particular result - are these sequences homologous?

How to handle too many results

- Focus on the question you are trying to answer
  - select “refseq” database to eliminate redundant matches from “nr”
  - Limit hits by organism
  - Use just a portion of the query sequence, when appropriate
  - Adjust the expect value; lowering *E* will reduce the number of matches returned

How to handle too few results

- Many genes and proteins have no significant database matches
  - remove Entrez limits
  - raise E-value threshold
  - search different databases
  - try scoring matrices with lower BLOSUM values (or higher PAM values)
  - use a search algorithm that is more sensitive than BLAST (*e.g.* PSI-BLAST or HMMer)

Summary of key points

- Sequence alignment is a fundamental operation underlying much of bioinformatics.
- Even when optimal solutions can be obtained they are not necessarily unique or reflective of the biologically correct alignment.
- Dynamic programming is a classic approach for solving the pairwise alignment problem.
- Global and local alignment, and their major application areas.
- Heuristic approaches are necessary for large database searches and many genomic applications.

FOR NEXT CLASS…

Check out the online:

- **Reading**: Sean Eddy's "What is dynamic programming?"
- **Homework**: (1) **Quiz**, (2) **Alignment Exercise**.

Homework Grading

Both (1) quiz questions and (2) alignment exercise carry equal weights (*i.e.* 50% each).

(Homework 2) Assessment Criteria	Points	
Setup labeled alignment matrix	1	
Include initial column and row for GAPs	1	
All alignment matrix elements scored (i.e. filled in)	1	
Evidence for correct use of scoring scheme	1	
Direction arrows drawn between all cells	1	
Evidence of multiple arrows to a given cell if appropriate	1	D
Correct optimal score position in matrix used	1	C
Correct optimal score obtained for given scoring scheme	1	B
Traceback path(s) clearly highlighted	1	A
Correct *alignment(s)* yielding optimal score listed	1	A+