

BGGN 213

Foundations of Bioinformatics

Lecture 2

Barry Grant
UC San Diego

<http://thegrantlab.org/bggn213>

Recap From Last Time:

- Bioinformatics is computer aided biology.
 - ▶ Deals with the collection, archiving, organization, and interpretation of a wide range of biological data.
- There are a large number of bioinformatics databases (see [handout!](#)).
- 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](#))
 - 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/>

NCBI Resources How To Sign in to NCBI

All Databases ras Search

NCBI Home Resource List (A-Z) All Resources Chemicals & Bioassays Data & Software DNA & RNA Domains & Structures Genes & Expression Genetics & Medicine Genomes & Maps Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

Welcome to NCBI

The National Center for Biotechnology Information advances health by providing access to biomedical information.

About the NCBI | Mission | Organization | NCBI News

Get Started

- Data: Find data using NCBI software
- Tools: Get NCBI data or software
- How Tos: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Genotypes and Phenotypes

Data from Genome Wide Association studies that link genes and diseases. See study variables, protocols, and analysis.

Resources

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

NCBI Announcements

RefSeq release 69 available on

The full RefSeq release 69 is now available on the FTP site with 74 records describing 52,378,420 ...

Hands on demo (or see following slides)

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 **OMIM**
- 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**

ms - GQuery: Global Cross X

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

NCBI Resources How To Sign in to NCBI

Search NCBI databases

Help

ras

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

Literature		Genes			
Books	1,677	books and reports	EST	3,985	expressed sequence tag sequences
MeSH	402	ontology used for PubMed indexing	Gene	87,165	collected information about gene loci
NLM Catalog	223	books, journals and more in the NLM Collections	GEO DataSets	3,732	functional genomics studies
PubMed	54,672	scientific & medical abstracts/citations	GEO Profiles	1,622,789	gene expression and molecular abundance profiles
PubMed Central	96,114	full-text journal articles	HomoloGene	696	homologous gene sets for selected organisms
Health		PopSet	2,254	sequence sets from phylogenetic and population studies	
ClinVar	759	human variations of clinical significance	UniGene	4,770	clusters of expressed transcripts
dbGaP	120	genotype/phenotype interaction studies	Proteins		
GTR	1,879	genetic testing registry			

8

ms - 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 Hide sidebar >>

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

Filters: Manage Filters

Did you mean ras as a gene symbol?
Search Gene for ras as a symbol.

<< First < Prev Page 1 of 4282 Next > Last >>

Results: 1 to 20 of 85633

i Filters activated: Current only. [Clear all](#) to show 87165 items.

Top Organisms [Tree]

Homo sapiens (1126) **(highlighted)**
Mus musculus (823)
Rattus norvegicus (625)
Oreochromis niloticus (533)
Neolamprologus brichardi (507)
All other taxa (82019)
[More...](#)

Find related data

Database: Select Find items

Search details

ras [All Fields] AND alive [property]

Name/Gene ID	Description	Location	Aliases
<input type="checkbox"/> ras ID: 19412	resistance to audiogenic seizures [Mus musculus (house mouse)]		asr
<input type="checkbox"/> ras ID: 43873	rasberry [Drosophila melanogaster (fruit fly)]	Chromosome X, NC_004354.4 (10744502..10749097)	Dmel_CG1799, CG11485, CG1799, DmelCG1799, EP(X)1093,

(ras) AND "Homo sapiens"[porgn:txid9606]

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 >>

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

① Filters activated: Current only. Clear all to show 1499 items.

Name/Gene ID	Description	Location	Aliases
<input type="checkbox"/> 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
<input type="checkbox"/> 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, K-RAS1, KRAS2, NS, NS2, RASK2

Filters: Manage Filters

Find related data

Database: Select

Find Items

Search details

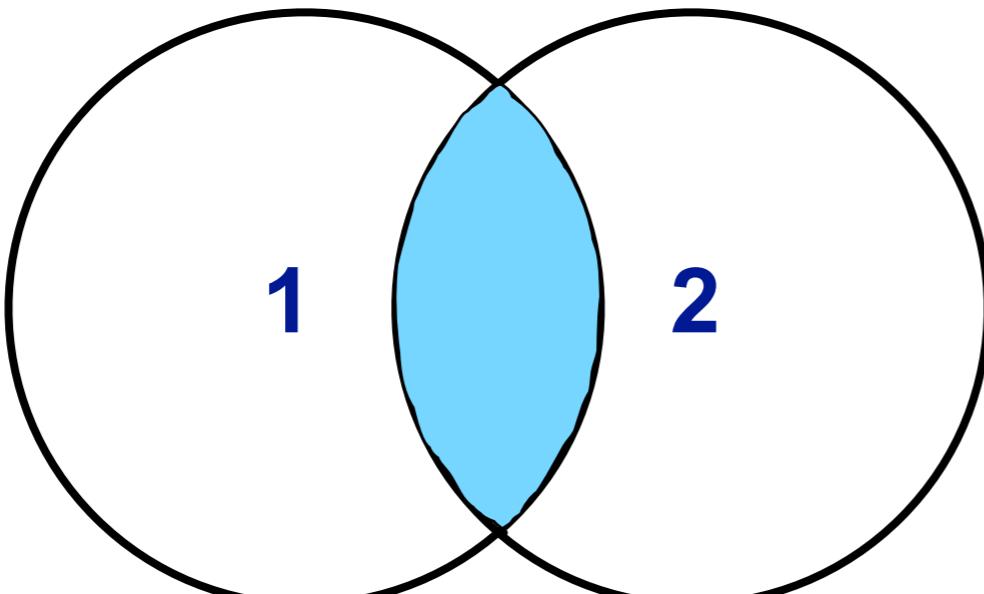
```
ras[All Fields] AND "Homo sapiens"[porgn] AND alive[property]
```

Search See more...

Recent activity Turn Off Clear

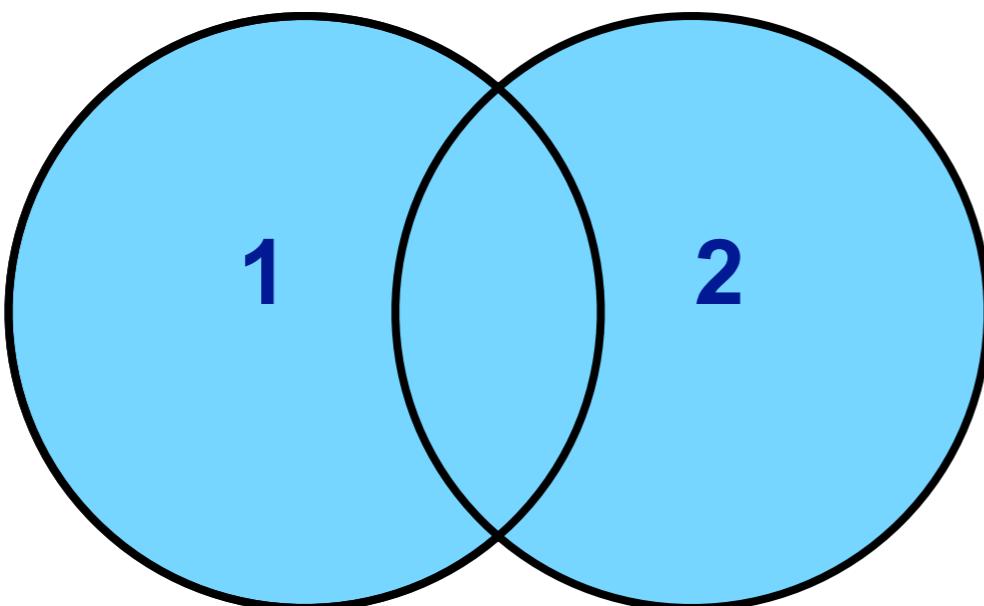
Chromosome locations Select

1 AND 2



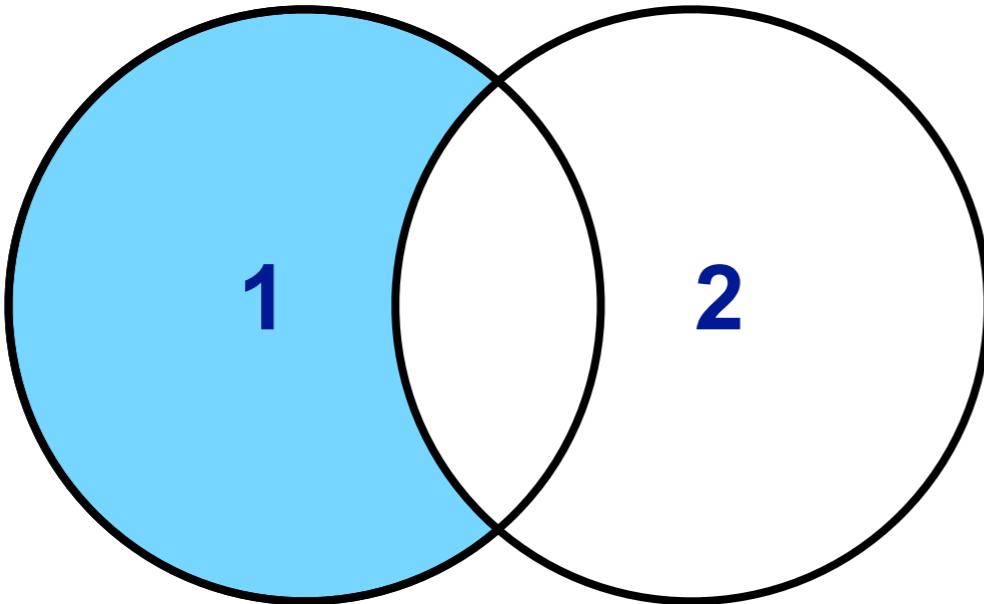
**ras AND disease
(1185 results)**

1 OR 2



**ras OR disease
(134,872 results)**

1 NOT 2



**ras NOT disease
(84,448 results)**

(ras) AND "Homo sapiens"[porgn:txid9606]

Gene Gene (ras) AND "Homo sapiens"[porgn:txid9606] Search

Save search Advanced Help

Show additional filters Hide sidebar >>

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

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

① Filters activated: Current only. Clear all to show 1499 items.

	Name/Gene ID	Description	Location	Aliases
Categories	<input type="checkbox"/> 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
Sequence content	<input type="checkbox"/> 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, K-RAS1, KRAS2, NS, NS2, RAKS2

Gene sources Genomic

Clear all

Gene sources Genomic

Categories

Alternatively spliced

Annotated genes

Non-coding

Protein-coding

Pseudogene

Sequence content

CCDS

Ensembl

RefSeq

Status clear

Current only

Chromosome locations Select

Filters: Manage Filters

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

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)

13

KRAS Kirsten rat sarcoma viral oncogene homolog

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

NCBI Resources How To Sign in to NCBI

Gene Search Help

Display sidebar >> Hide sidebar >>

Example Questions:

What chromosome location and what genes are in the vicinity?

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-RAS2B

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)

Related

14

KRAS KRas gene summary

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)

Chromosome 12 - NC_000012.12

Genomic regions, transcripts, and products

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

Go to nucleotide: Graphics FASTA GenBank

Go to reference sequence details

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

KRAS Kirsten rat sarcoma viral oncogene homolog

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

NCBI Resources How To Sign in to NCBI

Gene Search Help

Display Settings Hide sidebar >>

KRAS **Ki**
(human)]

Gene ID: 3845

Summary

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

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)
Related

16

KRAS KRas protein

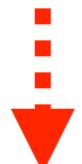
Gene Ontology Provided by GOA

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 signalling 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

The screenshot shows the UniProt-GOA database homepage. At the top, there's a navigation bar with tabs for 'Services', 'Research', 'Training', and 'About us'. Below the navigation bar is a search bar with examples like 'GO:0006915, tropomyosin, P08727' and a 'Search' button. The main content area features a large title 'UniProt-GOA' and a sub-section titled 'Gene Ontology Annotation (UniProt-GOA) Database'. A detailed description of the program follows, mentioning UniProt GO annotation, UniProt Knowledgebase, UniProt biocuration, manual and electronic annotations, and external collaborating groups. At the bottom, it states 'UniProt is a member of the GO Consortium'. On the right side, there's a 'Menu' sidebar with links to 'Downloads', 'Searching UniProt-GOA', 'Annotation Methods', 'Annotation Tutorial', 'Manual Annotation Efforts' (which is currently selected), 'Reference Genome Annotation Initiative', 'Cardiovascular Gene Ontology Annotation Initiative', 'Renal Gene Ontology Annotation Initiative', and 'Exosome Gene'.

KRAS Kiraten rat sarcoma x UniProt-GOA < EMBL-EBI x

www.ebi.ac.uk/GOA

EMBL-EBI

Services Research Training About us

UniProt-GOA

Search

Examples: GO:0006915, tropomyosin, P08727

Overview New to UniProt-GOA FAQ Contact Us

Gene Ontology Annotation (UniProt-GOA) Database

The UniProt GO annotation program aims to provide high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB). The assignment of GO terms to UniProt records is an integral part of UniProt biocuration. UniProt manual and electronic GO annotations are supplemented with manual annotations supplied by external collaborating GO Consortium groups, to ensure a comprehensive GO annotation dataset is supplied to users.

UniProt is a member of the GO Consortium.

Menu

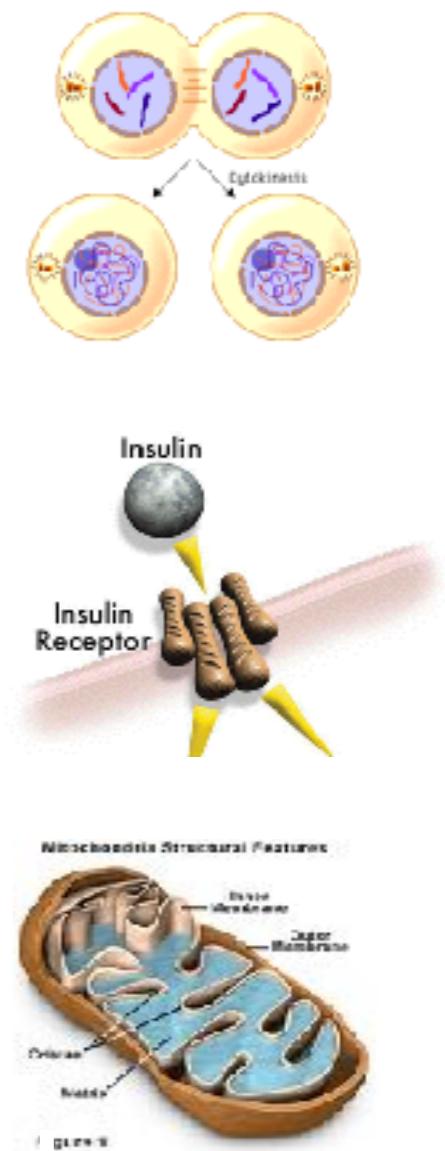
- Downloads
- Searching UniProt-GOA
- Annotation Methods
- Annotation Tutorial
- Manual Annotation Efforts
 - Reference Genome Annotation Initiative
 - Cardiovascular Gene Ontology Annotation Initiative
 - Renal Gene Ontology Annotation Initiative
 - Exosome Gene

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 Process**
A commonly recognized series of events
e.g. cell division, mitosis,
 - ▶ **Molecular Function**
An elemental activity, task or job
e.g. kinase activity, insulin binding
 - ▶ **Cellular Component**
Where a gene product is located
e.g. mitochondrion, mitochondrial membrane



KRAS KRasG12D rat isoform 1

Gene Ontology Provided by GOA

Function Evidence Code Pubs

GDP binding TAS

GMP binding IEA

GTP binding TAS

LRR domain binding TAS

protein binding IEA

protein complex binding TAS

Process Code

Fc-epsilon receptor signalling 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

such as this one

The screenshot shows a gene page for KRAS (X01669.1) with protein accession CAA25828.1. The UniProtKB Link (UniProtKB/Swiss-Prot:P01116) is highlighted with a red box. A red arrow points from the text "Scroll down to very bottom for UniProt link" to this highlighted link.

genomic X01669.1 CAA25828.1

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

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

Additional links

You are here: NCBI > Genes & Expression > Gene Write to the Help Desk

GETTING STARTED	RESOURCES	POPULAR	FEATURED	NCBI INFORMATION
NCBI Education	Chemicals & Bioassays	PubMed	Genetic Testing Registry	About NCBI
NCBI Help Manual	Data & Software	Bookshelf	PubMed Health	Research at NCBI
NCBI Handbook	DNA & RNA	PubMed Central	GenBank	NCBI News
Training & Tutorials	Domains & Structures	PubMed Health	Reference Sequences	NCBI FTP Site
	Genes & Expression	BLAST	Gene Expression Omnibus	NCBI on Facebook
	Genetics & Medicine	Nucleotide	Map Viewer	NCBI on Twitter
	Genomes & Maps	Genome	Human Genome	NCBI on YouTube
	Homology	SNP	Mouse Genome	
	Literature	Gene	Influenza Virus	
	Proteins	Protein	Primer-BLAST	
	Sequence Analysis	PubChem	Sequence Read Archive	
	Taxonomy			

UniProt will detail much more information for protein coding genes

KRAS - GTPase KRas protein

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

Display: None

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

BLAST Align Format Add to basket History Feedback Help video

Function

Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23698361, PubMed:22711838). 2 Publications Curated

Enzyme regulation

Alternates between an inactive form bound to GDP and an active form bound to GTP. Activated by a guanine nucleotide-exchange factor (GEF) and Inactivated by a GTPase-activating protein (GAP). Interaction with SOS1 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

The screenshot shows the UniProtKB interface for protein P01116 (KRAS_HUMAN). Key details include:

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

The "Format" button in the top navigation bar is highlighted with a green box, and a callout bubble says "View FASTA file format".

Function:
Ras proteins bind GDP/GTP and promote cell proliferation (PubMed: 23698361, PDB ID: 1RAS).

Enzyme regulation:
Alternates between an inactive form and an active GTP-bound form. It is a nucleotide-exchange factor (GEF) that 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)	Graphical view	Feature identifier	Actions
Nucleotide binding ⁱ	29 – 35	7	GTP (2 Publications)	Graphical view	Feature identifier	Actions
Nucleotide binding ⁱ	59 – 60	2	GTP (2 Publications)	Graphical view	Feature identifier	Actions

UniProt will detail much more information for protein coding genes

KRAS - GTPase KRas protein

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

Display: None

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

BLAST Align Format Add to basket History Feedback Help video

Function

Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23698361, PubMed:22711838). 2 Publications Curated

Enzyme regulation

Alternates between an inactive form bound to GDP and an active form bound to GTP. Activated by a guanine nucleotide-exchange factor (GEF) and Inactivated by a GTPase-activating protein (GAP). Interaction with SOS1 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			

KRAS - GTPase KRas protein

www.uniprot.org/uniprot/P01116

UniProt

UniProtKB Advanced

BLAST Align Retrieve/ID Mapping Help Contact

P01116 - RASK_HUMAN

Protein: GTPase KRas
Gene: KRAS
Organism: Homo sapiens (Human)
Status: Reviewed - ●●●●●

Display: None

None BLAST Align Format Add to basket History Feedback Help video

FUNCTION

NAMES & TAXONOMY

SUBCELL LOCATION

PATHOL/BIOTECH

PTM / PROCESSING

EXPRESSION

INTERACTION

STRUCTURE

FAMILY & DOMAINS

SEQUENCES (2)

CROSS-REFERENCES

Example Questions:

What positions in the protein are responsible for GTP binding?

Function

Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23698361, PubMed:22711838). 2 Publications Curated

Enzyme regulation

Alternates between an inactive form bound to GDP and an active form bound to GTP. Activated by a guanine nucleotide-exchange factor (GEF) and Inactivated by a GTPase-activating protein (GAP). Interaction with SOS1 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			

Example Questions:

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

KRAS - GTPase KRas protein

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

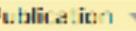
SIMILAR PROTEINS

Top

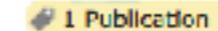
Pathology & Biotech

Involvement in disease¹

LEUKEMIA, ACUTE MYELOGENOUS (AML)

[MIM:601626]: A subtype of acute leukemia, a cancer of the white blood cells. AML is a malignant disease of bone marrow characterized by maturational arrest of hematopoietic precursors at an early stage of development. Clonal expansion of myeloid blasts occurs in bone marrow, blood, and other tissue. Myelogenous leukemias develop from changes in cells that normally produce neutrophils, basophils, eosinophils and monocytes. 

Note: The disease is caused by mutations affecting the gene represented in this entry.

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Natural variant ¹	10 – 10		1. G → GG in one individual with AML; expression in 3T3 cell causes cellular transformation; expression in COS cells activates the Ras-MAPK signaling pathway; lower GTPase activity; faster GDP dissociation rate. 		VAR_034601	

LEUKEMIA, JUVENILE MYELOMONOCYTIC (JMML)

[MIM:607785]: An aggressive pediatric myelodysplastic syndrome/myeloproliferative disorder characterized by malignant transformation in the hematopoietic stem cell compartment with proliferation of differentiated progeny. Patients have splenomegaly, enlarged lymph nodes, rashes, and hemorrhages.

Note: The disease is caused by mutations affecting the gene represented in this entry.

NOONAN SYNDROME 3 (NS3)

[MIM:609942]: A form of Noonan syndrome, a disease characterized by short stature, facial dysmorphic features such as hypertelorism, a downward eyeslant and low-set posteriorly rotated ears, and a high incidence of congenital heart

Example Questions:

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

KRAS - GTPase KRas protein

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

Structure

Secondary structure

Legend: Helix Turn Beta strand

Show more details

3D structure databases

Select the link destinations:

PDBeⁱ RCSB PDBⁱ PDBJⁱ

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	1-154	[>]
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?

The screenshot shows the RCSB PDB homepage. At the top, there is a navigation bar with links for Home, Gmail, Deposit, Search, Visualize, and Analyze. Below this is the main header with the RCSB PDB logo and a search bar. The search bar contains the text "Search by PDB ID, author, macromolecule, sequence, or ligand" and a "Go" button. Below the search bar are links for "Advanced Search" and "Browse by Annotations".

The main content area features the RCSB PDB logo and the text "An Information Portal to 133759 Biological Macromolecular Structures". Below the logo are links for PDB-101, Worldwide PDB, EMDDataBank, NDB, and Worldwide Protein Data Bank Foundation.

A navigation menu at the bottom includes "Structure Summary", "3D View" (which is highlighted with a red box), "Annotations", "Sequence", "Sequence Similarity", "Structure", and "Literature".

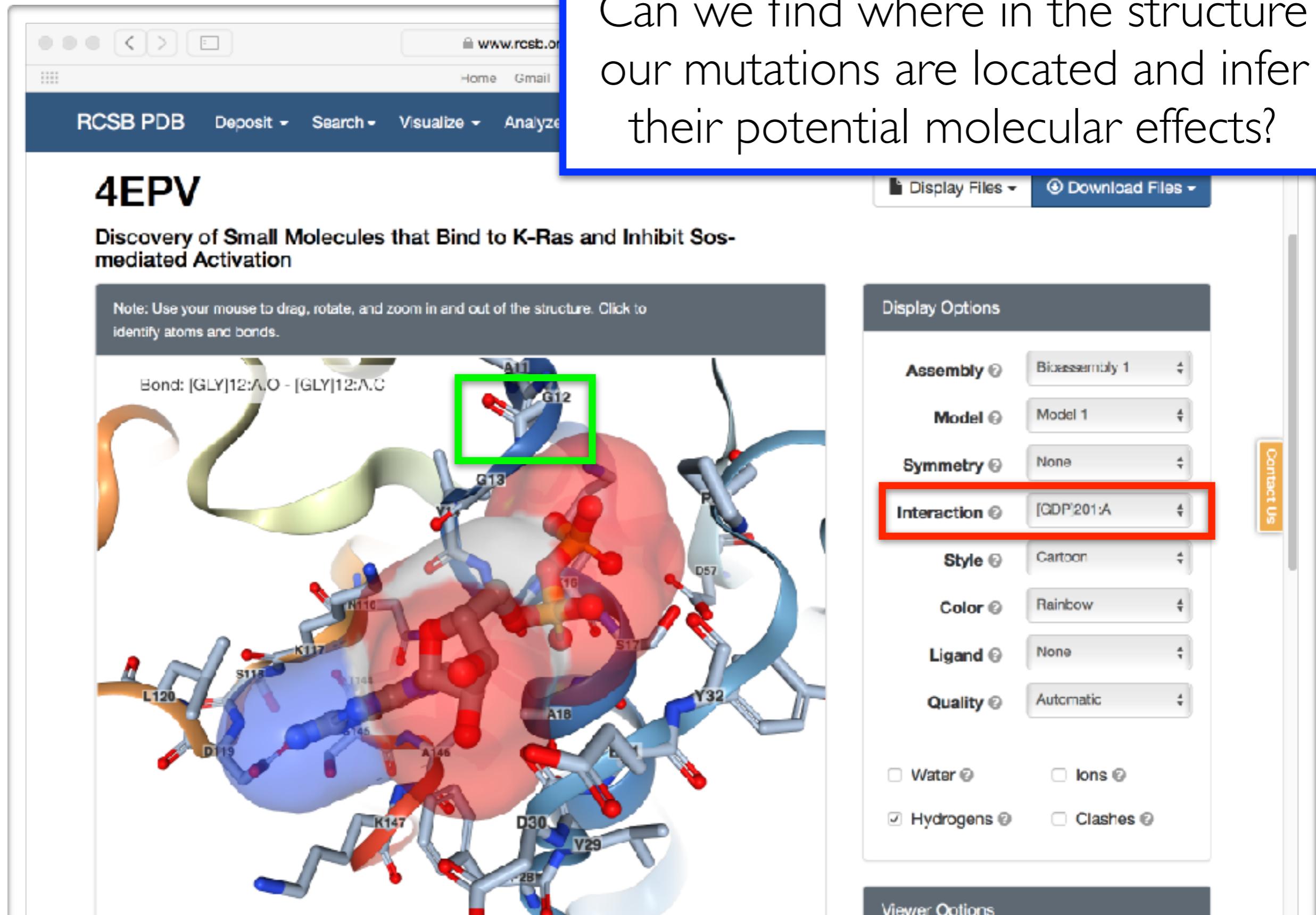
The central part of the page displays a protein structure labeled "4EPV" with the title "Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation". It includes information about the DOI (10.2210/pdb4epv/pdb), classification (HYDROLASE), deposition date (2012-04-17), release date (2012-05-23), deposition authors (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), and mutation(s) (1).

On the right side, there is a "View PDB file format" button with a green speech bubble, and a "Display Files" button with a green box around it, along with a "Download Files" button.

At the bottom, there are links for "View in 3D: NGL or JSmol (in Browser)", "Experimental Data Snapshot", "wwPDB Validation", "3D Report", and "Full Report".

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... ?

The screenshot shows the UniProt protein entry page for P01116 (KRAS - GTPase KRas protein). The 'Display' section has 'None' selected. Under 'Family and domain databases', the Pfam entry is highlighted with a red box. The text 'PFAM is one of the best protein family databases' is overlaid in a red box.

Pfam: PF00071. Ras. 1 hit.
[Graphical view]

PFAM is one of the best protein family databases

Sequences (2)
Sequence statusⁱ: Complete.
Sequence processingⁱ: The displayed sequence is further processed into a mature form.
This entry describes 2 isoformsⁱ 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... ?

KRAS - GTPase KRas protein | Pfam: Family: Ras (PF00071) | pfam.xfam.org/family/PF00071

EMBL-EBI HOME

Family: Ras (PF00071)

Summary **Domain organisation** **Clan** **Alignments** **HMM logo** **Trees** **Curation & model** **Species** **Interactions** **Structures** **Jump to...** 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.

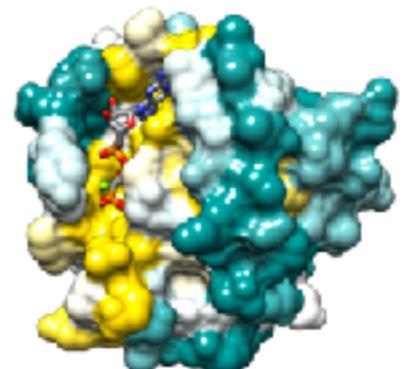
Because these signals result in cell growth and division, overactive Ras signaling can ultimately lead to cancer.^[1] The 3 Ras genes in humans (HRAS, KRAS, and NRAS) are the most common oncogenes in human cancer; mutations that permanently activate Ras are found in 20% to 25% of all human tumors and 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](#) [edit]

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

Identifiers

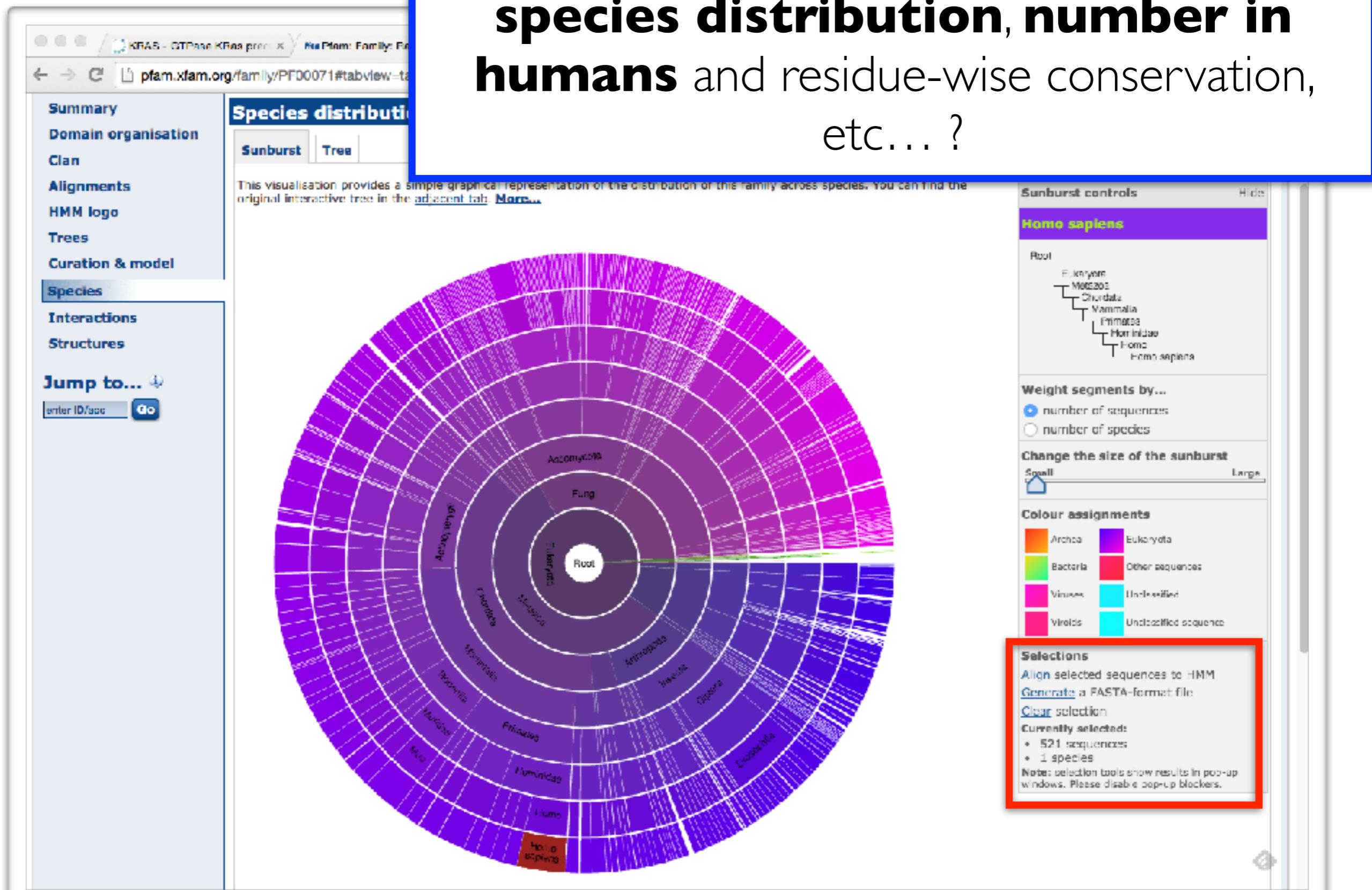
Symbol	Ras
Pfam	PF00071_5
InterPro	IPR013753
PROSITE	PS000017
SCOP	Sp21
SUPERFAMILY	Sp21



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

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... ?

KRAS - GTPase KRas protein
No Pfam: Family: Pro

← → C pfam.xfam.org/family/PF00071#tabview=tab1

Summary Species distribution

Domain organisation

Clan

Alignment

HMM log

Trees

Curation

Species

Interaction

Structure

Jump to another ID/acc

EMBL-EBI

Pfam: Pfam alignment viewer

pfam.xfam.org/family/PF00071/alignment/view?jobId=EDCA403E-9836-11E4-B360-10D3298E2F76

Alignment for selected sequences

Currently showing rows 1 to 30 of 536 rows in this alignment. Show 30 rows of alignment.

P11234/16-178	...KIVVVGCCVGRGADIL...	Q...	FN...	T...	D...	E...	F...	V...	E...	DYDFPK...	-AD...	SYRENWLD...
P01112/5-163	...KLIVVGGCCVGRGADIL...	Q...	LI...	Q...	N...	H...	V...	D...	DYDFPK...	-ED...	SYREQWVID...	
Q14088/38-204	...KIVVVGCCVGRGADIL...	R...	FC...	G...	E...	D...	F...	V...	DYDFPK...	GID...	TYERKVWIDE...	
Q9RN83/7-173	...KIVVVGCCVGRGADIL...	T...	FR...	S...	D...	N...	K...	AVTIT...	GID...	TAUVTVPED...		
P15153/7-178	...KIVVVGCCVGRGADIL...	S...	YT...	E...	N...	A...	P...	TYIFPV...	-ED...	YISAKWVHD...		
Q00194/11-183	...KLLALGCCVGRGADIL...	R...	YT...	D...	N...	E...	P...	XPFICW...	GID...	TYERKVWVSDPN...		
Q11907/13-174	...KIVVVGCCVGRGADIL...	R...	FT...	R...	N...	E...	N...	L...	DSRITI...	GVE...	FATREIQWT...	
P10114/5-165	...KIVVVGCCVGRGADIL...	Q...	FV...	G...	E...	D...	F...	I...	KYMPIT...	-ED...	TYRKEFTRWD...	
P51153/10-171	...KLLLGDDCCVGRGADIL...	R...	FA...	E...	D...	N...	N...	TYISHI...	GID...	TKIRTWDIE...		
P53040/77-241	...RVLVIGCCVGRGADIL...	I...	FA...	Gvhd...	SM...	D...	S...	D-CEVL...	GID...	TYERTLMVD...		
P5042/93-203	...RVLVIGCCVGRGADIL...	I...	FO...	G...	V...	E...	G...	EEAAB...	--H...	TYDREIWWI...		
P01116/5-163	...KIVVVGCCVGRGADIL...	Q...	LI...	Q...	N...	H...	V...	D...	DYDFPK...	-ED...	SYREKWVID...	
Q9J0W7/21-182	...KIVVVGCCVGRGADIL...	R...	YI...	R...	N...	D...	N...	TYIFPV...	GID...	TYTKUVVDE...		
Q9ULC3/11-171	...KIVVVGCCVGRGADIL...	R...	YC...	R...	E...	I...	T...	K...	DYDFPK...	GID...	TYERGVIN...	
Q11807/15-177	...KIVVVGCCVGRGADIL...	C...	FT...	G...	K...	I...	V...	P...	DYDFPK...	-ED...	SYLKHED...	
Q9NKD7/7-202	...KIVVVGCCVGRGADIL...	R...	YN...	E...	R...	R...	F...	D...	T-V3W...	GID...	TYLEQW...	
Q9A062/35-201	...KIVVVGCCVGRGADIL...	R...	YI...	R...	R...	R...	F...	D...	DRTEATI...	GID...	TYERKVWID...	
Q95905/9-174	...KIVVVGCCVGRGADIL...	R...	YI...	H...	D...	H...	V...	D...	TYQATI...	GAA...	TYAKUMVDE...	
P51149/10-175	...KIVVVGCCVGRGADIL...	C...	YN...	H...	K...	S...	N...	OYKMTI...	GID...	TYTKEWVHD...		
Q9ULN5/65-227	...KIVVVGCCVGRGADIL...	R...	FK...	D...	G...	A...	D...	ATFISW...	GID...	TYNEVLDV...		
P57710/14-175	...KIVVVGCCVGRGADIL...	R...	FT...	R...	N...	E...	S...	H...	DSRITI...	GVE...	TYSTRVWHL...	
P51153/11-183	...KIVVVGCCVGRGADIL...	Q...	YT...	D...	G...	K...	N...	APLICW...	GID...	TYERKVWYRas...		
P01111/5-165	...KIVVVGCCVGRGADIL...	Q...	LI...	Q...	N...	H...	V...	D...	DYDFPK...	-ED...	SYREKWVID...	
P11233/16-177	...KIVVVGCCVGRGADIL...	C...	FN...	T...	D...	E...	F...	V...	DYDFPK...	-AD...	SYRENWLD...	
Q9UL25/21-182	...KIVVVGCCVGRGADIL...	R...	YC...	E...	N...	K...	N...	D...	XHITL...	GID...	TYTERKLNIC...	
Q9NP72/10-171	...KIVVVGCCVGRGADIL...	R...	FT...	D...	D...	T...	D...	P...	BLANTI...	GID...	TYVKTISV...	
Q9J0L4/10-171	...KIVVVGCCVGRGADIL...	R...	FA...	D...	D...	T...	V...	E...	DYDFPK...	GID...	TYRIMELD...	
Q9J0W6/7-165	...KIVVVGCCVGRGADIL...	R...	FT...	R...	D...	S...	D...	P...	NINMII...	GID...	TYPTWYOQ...	
Q9UBX7/23-179	...KIVVVGCCVGRGADIL...	R...	FL...	H...	D...	Q...	O...	P...	QQLSTY...	ACT...	TYXHTATV...	
P51157/14-179	...KIVVVGCCVGRGADIL...	C...	FA...	G...	E...	T...	G...	K...	TYKGTI...	GID...	TYLRRITL...	

There are 18 pages in this alignment. Show page 1

Download this alignment.

Close window

can find the

Sunburst controls Hide

Homo sapiens

Root

Eukaryota

M002203

Chordata

Mammalia

Primates

Hominoidea

Homo

Homo sapiens

Weight segments by...

number of sequences

number of species

Change the size of the sunburst

Small Large

Colour assignments

Archaea

Bacteria

Miniviruses

Viruses

Eukaryota

Other sequences

Unclassified

Unclassified sequence

Selections

Align selected sequences to HMM

Generate a FASTA-format file

Clear selection

Currently selected:

- + 521 sequences
- + 1 species

Note: selection tools show results in pop-up windows. Please disable pop-up blockers.

Example Questions:

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

KRAS - GTPase KRas protein | Help | Family: Ras

← → C pfam.xfam.org/family/PF00071#tabview=tab4

EMBL-EBI 

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

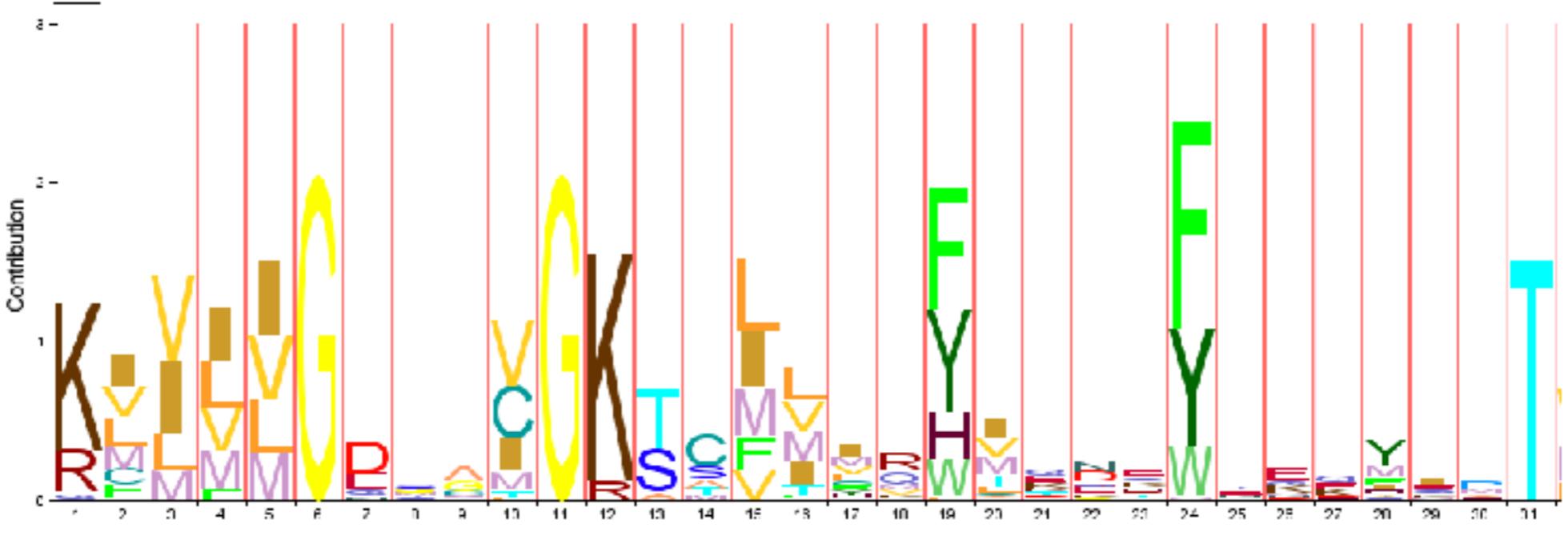
Family: Ras (PF00071)

Summary Domain organisation Clan Alignments **HMM logo** (highlighted with a red box) Trees Curation & model Species Interactions Structures

Jump to... enter ID/acc Go

HMM logo

HMM logos is one way of visualising profile HMMs. Logos provide a quick overview of the properties of an HMM in a graphical form. You can see a more detailed description of HMM logos and find out how you can interpret them [here](#). [More...](#)



Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

Family: Kinesin (PF00225)

 Loading page components (1 remaining)...
[Summary](#)[Domain organisation](#)[Clans](#)[Alignments](#)[HMM logo](#)[Trees](#)[Curation & models](#)[Species](#)[Interactions](#)[Structures](#)[Jump to... !\[\]\(4fa1a734083d1c409ea6d909f0c24706_img.jpg\)](#)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

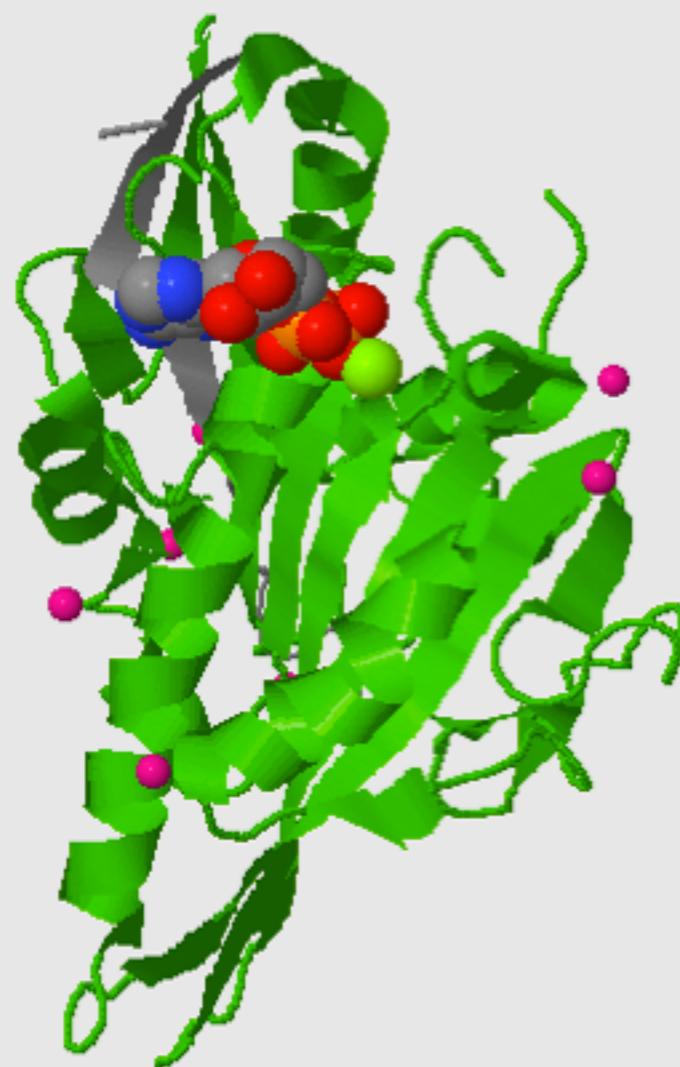
Structures

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
		3kar	A	392 - 723	Jmol AstexViewer SPICE
KI13B_HUMAN	11 - 352	3qbj	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
		1	A	24 - 359	Jmol AstexViewer SPICE



PDB entry 3bfm



Your turn:

What can you find out about “eg5”

Jmol

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

 Close window

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 programming
 - 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 programming
 - Global alignment
 - Local alignment
 - ▶ BLAST heuristic approach

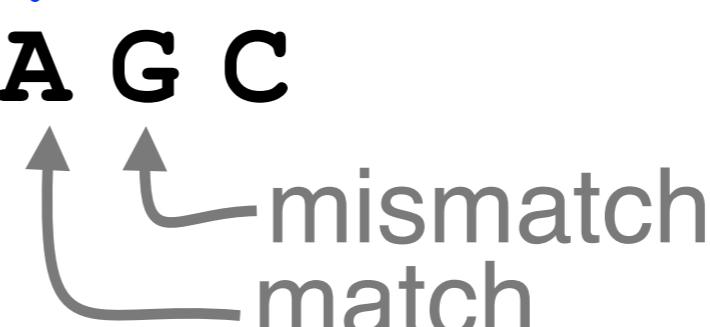
Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.

Seq1 : C A T T C A C

Seq2 : C T C G C A G C

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

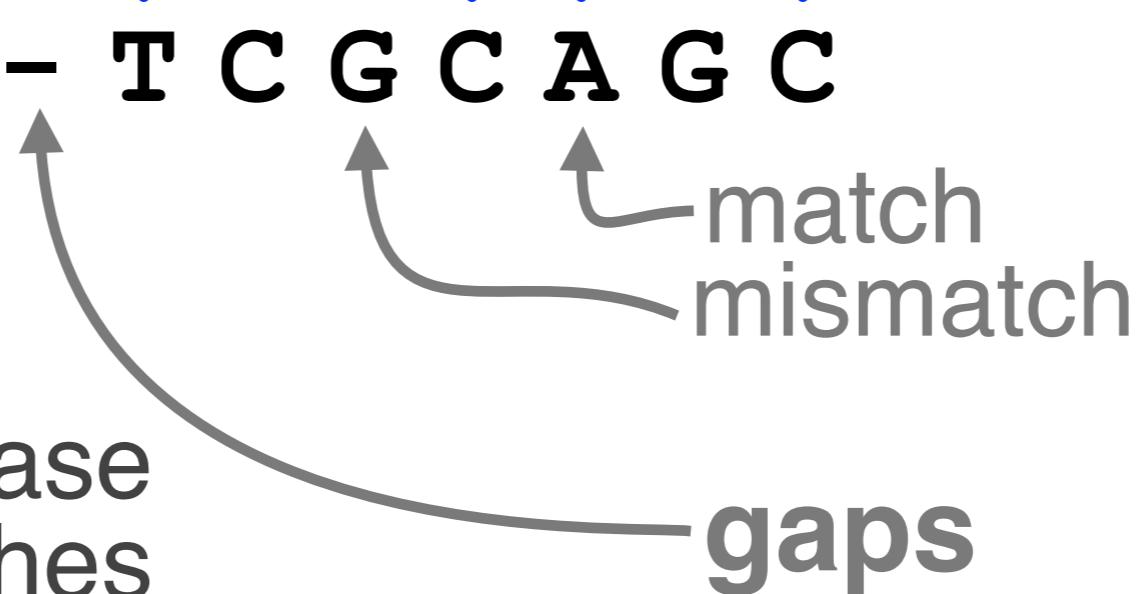
Seq1 : C A T T C A C
 | | |
Seq2 : C T C G C A G C



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.

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



Add gaps to increase number of matches

gaps

match mismatch

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

Gaps represent 'indels'
mismatch represent mutations

match
mismatch } mutation

insertion
deletion } indels

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 sequences**
 - construct such as
- **Mapping of new sequences to a known genome**
 - Looking for differences from reference sequences, indels (insertions or deletions)
 - Mapping transcription factor binding sites via ChIP-Seq (chromatin immuno-precipitation sequencing)
 - Pretty much all next-gen sequencing data analysis

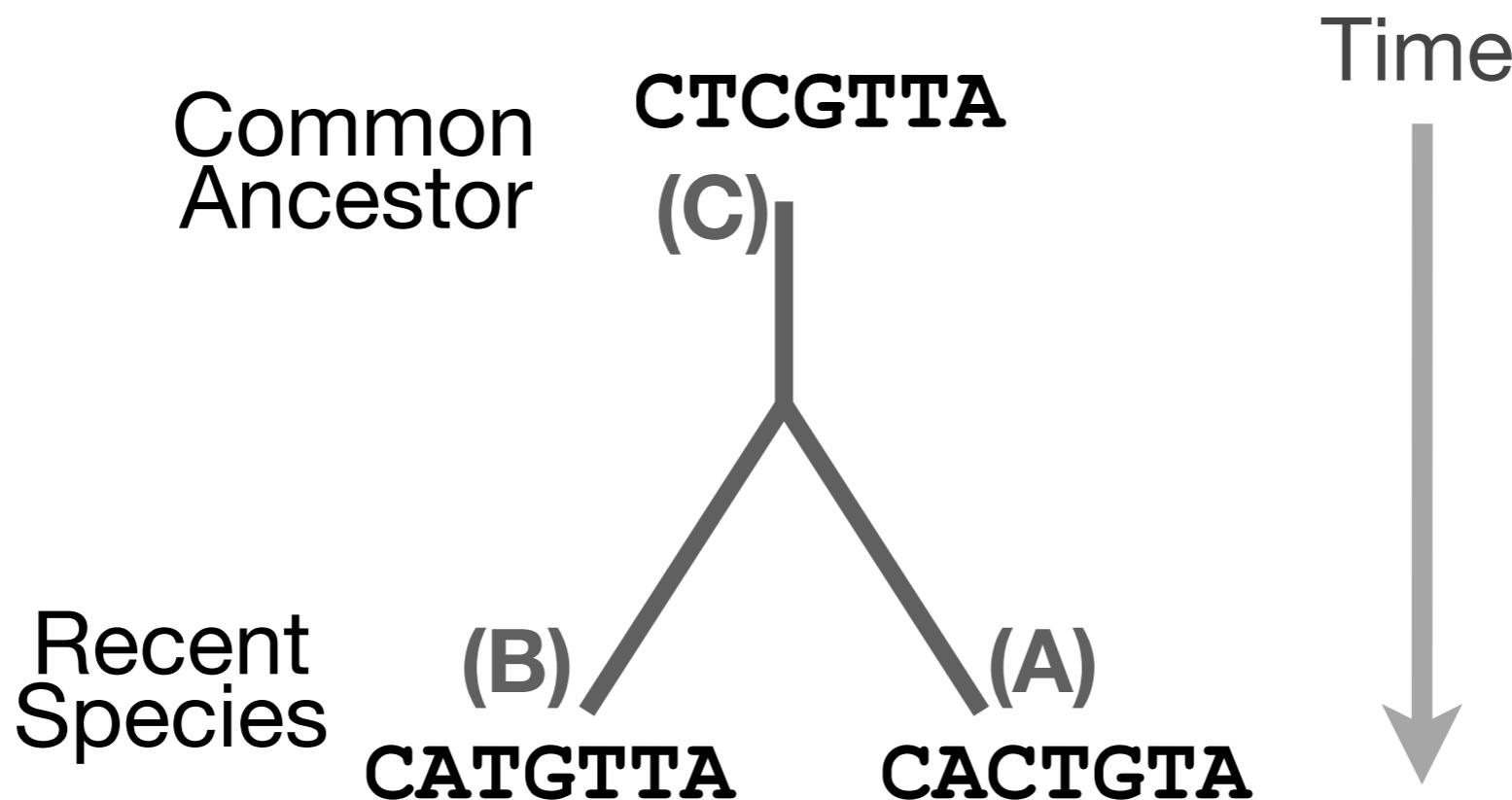
ALIGNMENT FOUNDATIONS

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

Sequence changes during evolution

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

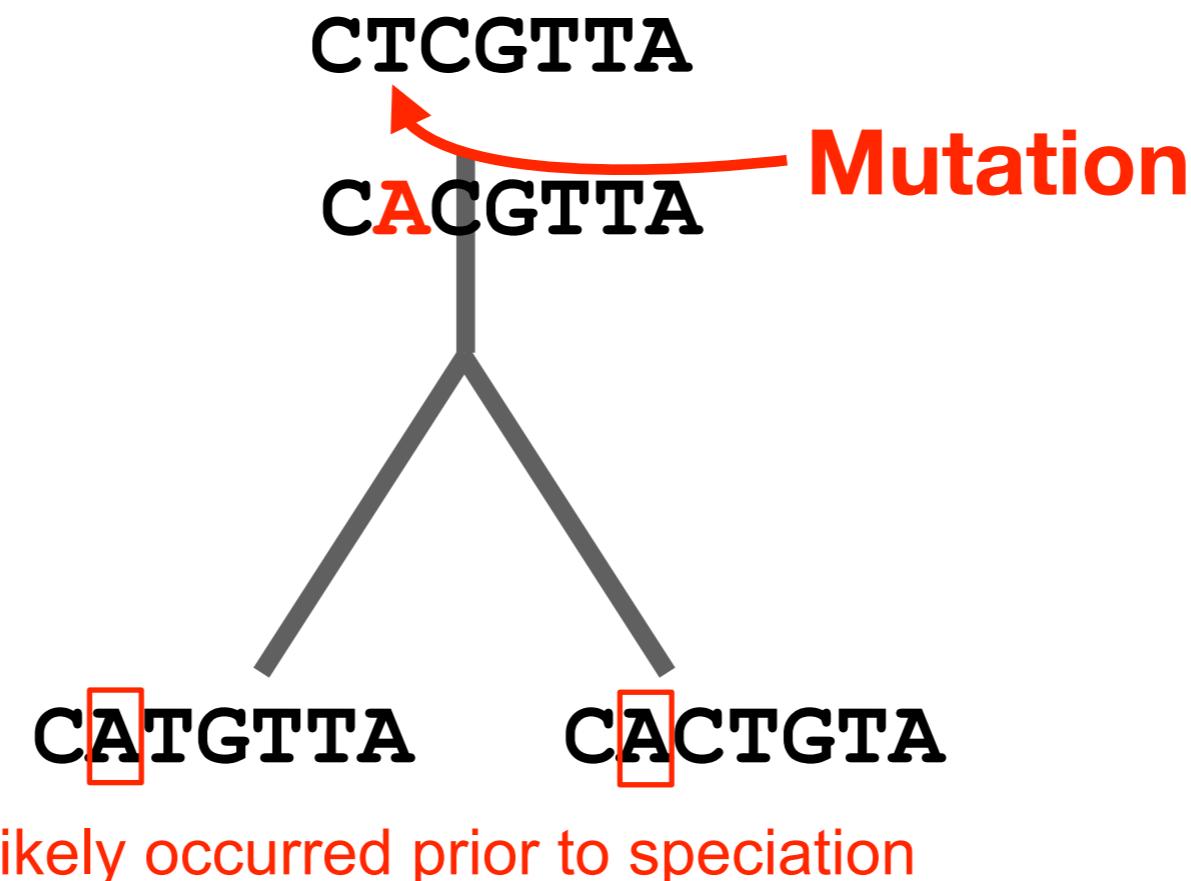
- Mutations/Substitutions
- Deletions
- Insertions



Mutations, deletions and insertions

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

- **Mutations/Substitutions** CTCGTTA → CACGTTA
- Deletions
- Insertions

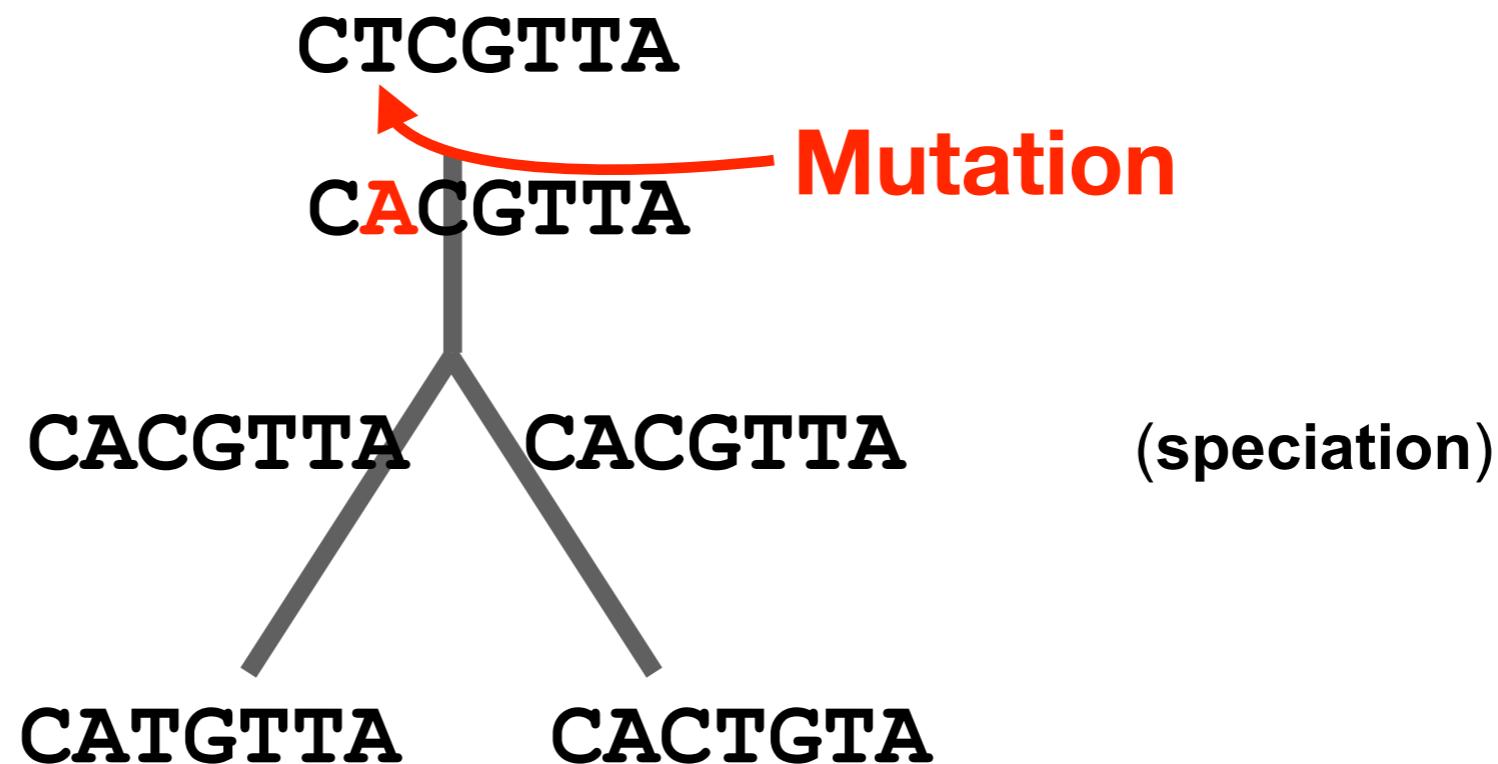


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA

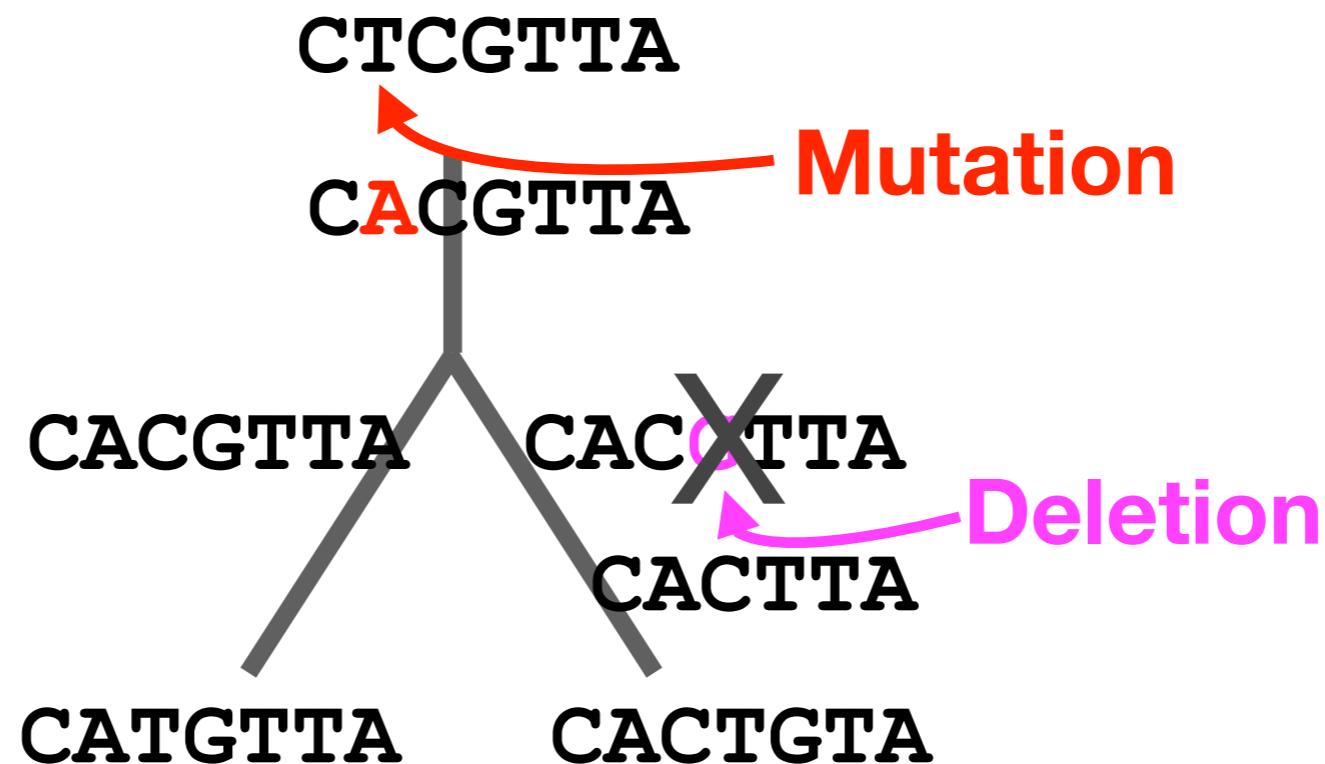


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA
CACGTTA → CACTTA

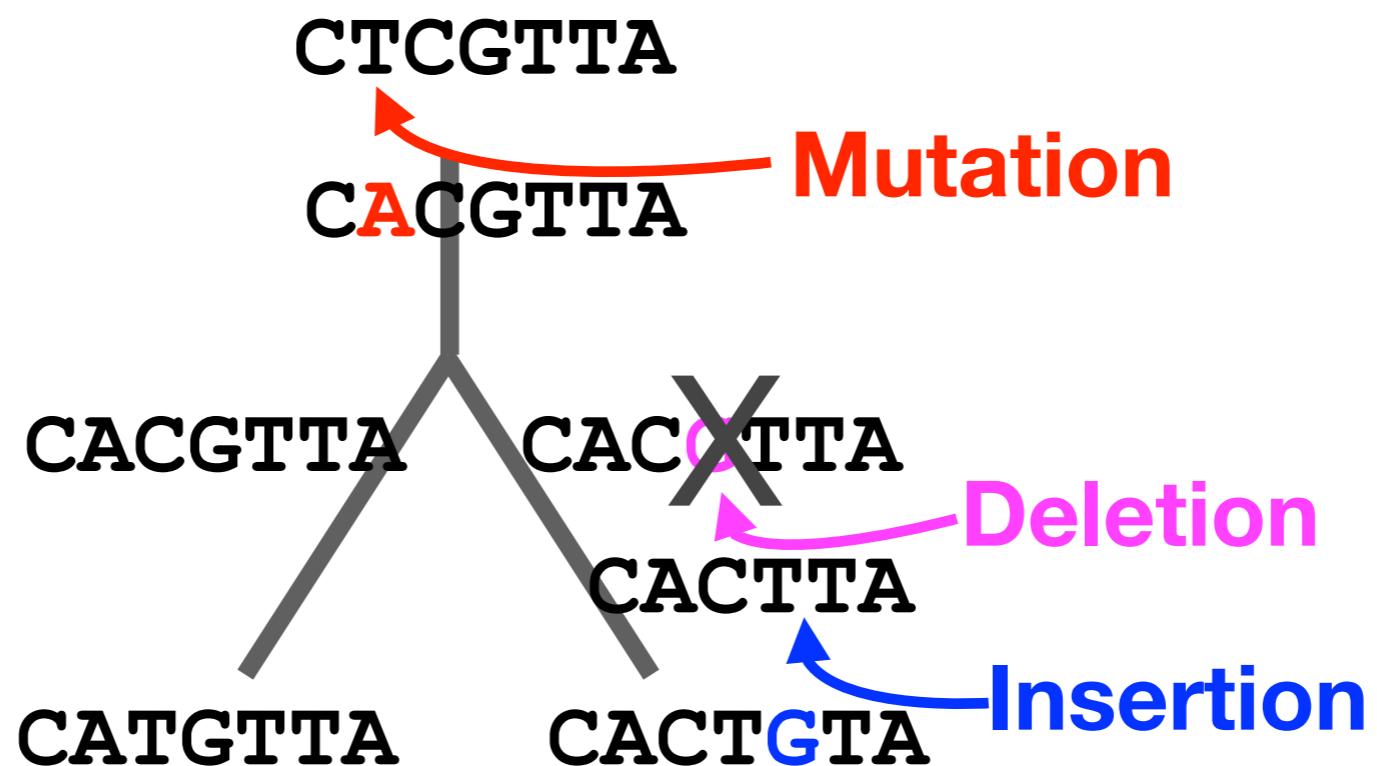


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA
CACGTTA → CACTTA
CACTTA → CACTGTA

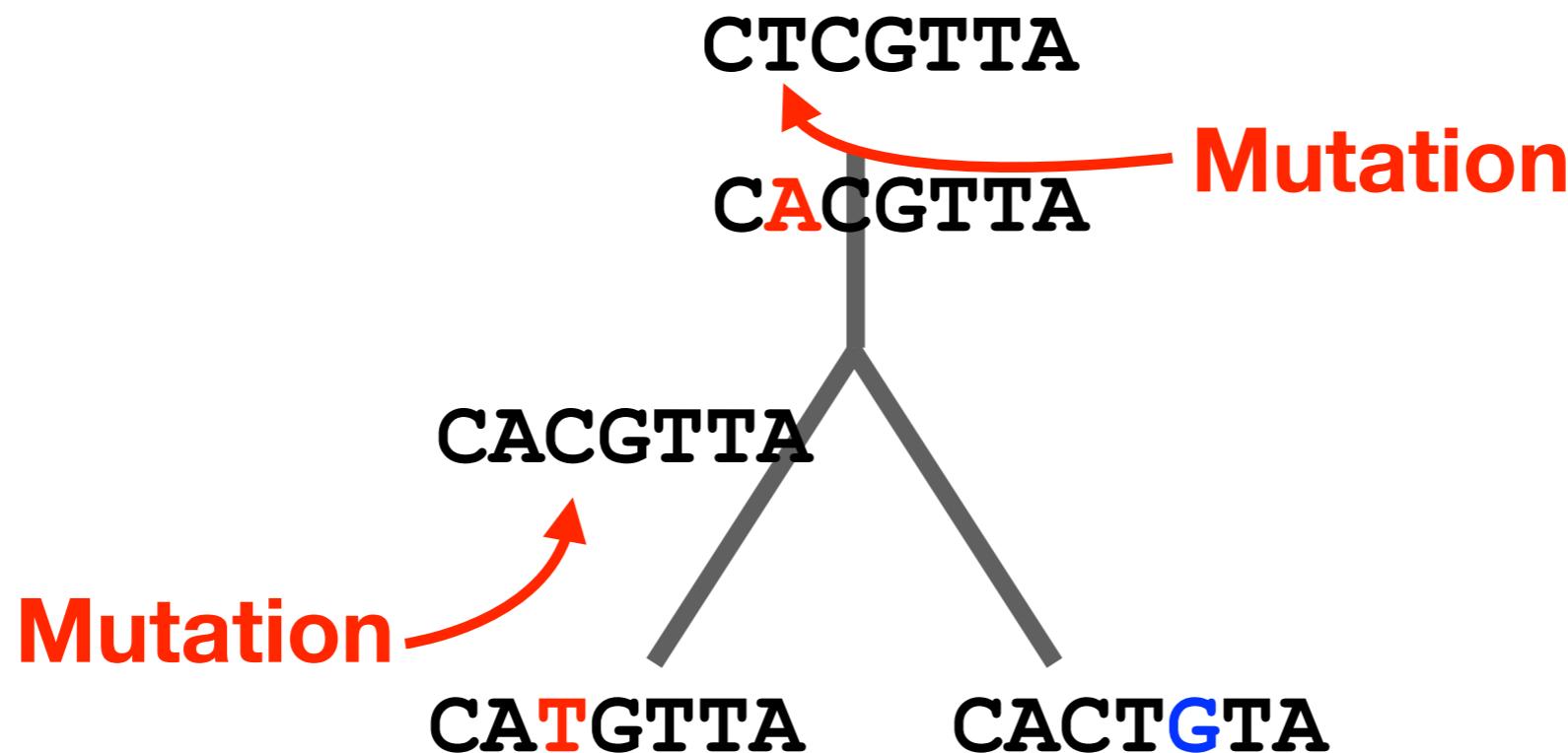


Mutations, deletions and insertions

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

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

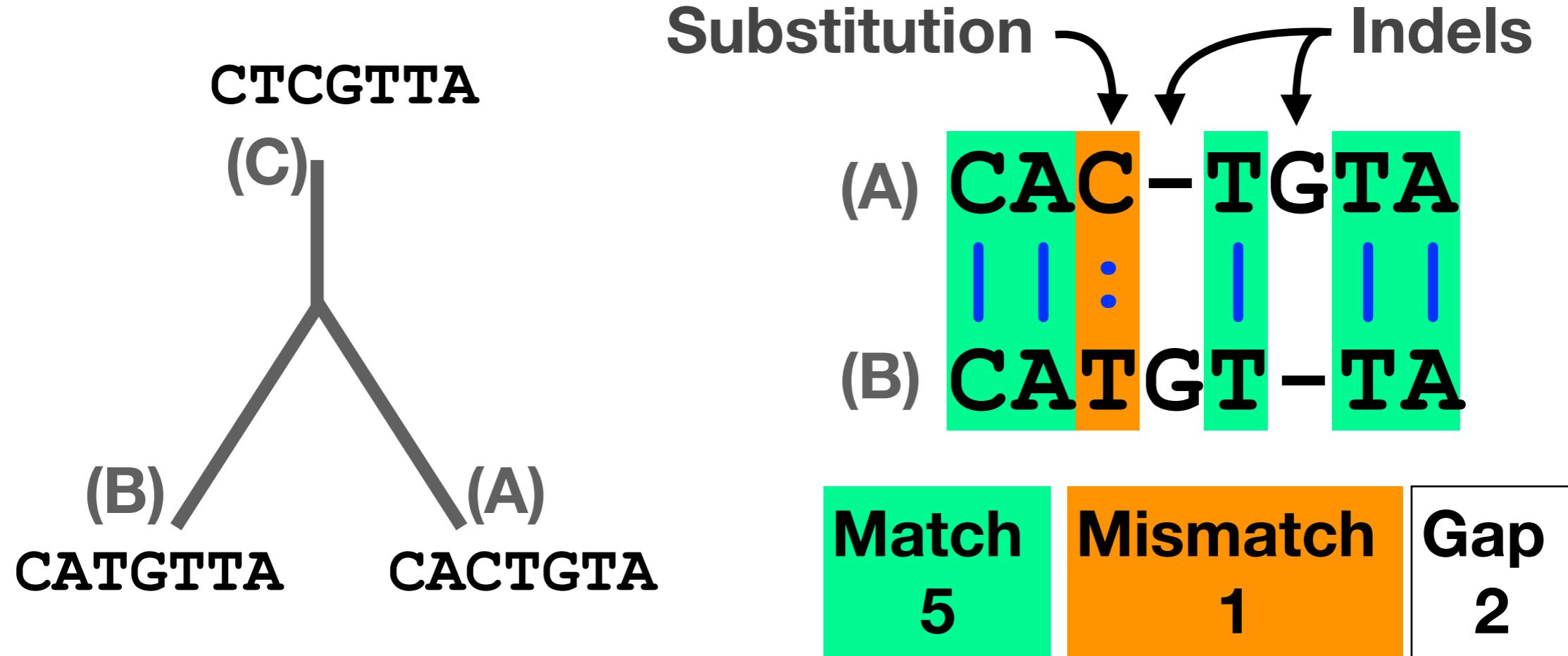
CTCGTTA → CACGTTA
CACGTTA → CATGTTA



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.

CA	CTG	TA
	:	:
CAT	TGT	TA

2.

CA	CTG	T	-A
CA	-T	G	T

3.

CAC	-T	G	TA
	:		
CAT	G	T	-TA

Alternative alignments

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

● 4 matches
● 3 mismatches
○ 0 gaps

● 6 matches
● 0 mismatches
○ 2 gaps

● 5 matches
● 1 mismatch
○ 2 gaps

CACTGTA
|| : : : ||
CATGTAA

This sequence alignment shows a green box for CACTGTA and an orange box for CATGTAA. There are 4 matches (green), 3 mismatches (orange), and 0 gaps (white). A vertical blue bar is positioned between the two sequences.

CACTGT-A
|| | | | |
CA-TGTAA

This sequence alignment shows a green box for CACTGT-A and an orange box for CA-TGTAA. There are 6 matches (green), 0 mismatches (orange), and 2 gaps (white). A vertical blue bar is positioned between the two sequences.

CAC-TGTA
| | : | |
CATGT-TA

This sequence alignment shows a green box for CAC-TGTA and an orange box for CATGT-TA. There are 5 matches (green), 1 mismatch (orange), and 2 gaps (white). A vertical blue bar is positioned between the two sequences.

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

● 6 (+3)
● 0 (+1)
○ 2 (-1) = 16

● 5 (+3)
● 1 (+1)
○ 2 (-1) = 14

CACTGTA
|| : : : ||
CATGTAA

A sequence alignment showing two DNA strands. The top strand is CACTGTA and the bottom strand is CATGTAA. Vertical blue lines indicate matches between A, C, T, G, T, and A. Dots indicate mismatches between the second C and the first A, and between the fourth T and the fifth T. The third T is aligned with a gap in the bottom strand.

CACTGT-A
|| | | | |
CA-TGTAA

A sequence alignment showing two DNA strands. The top strand is CACTGT-A and the bottom strand is CA-TGTAA. Vertical blue lines indicate matches between C, A, C, T, G, T, and A. Dots indicate mismatches between the second C and the first A, and between the fourth T and the fifth T. The third T is aligned with a gap in the top strand.

CAC-TGTA
| | : | |
CATGT-TA

A sequence alignment showing two DNA strands. The top strand is CAC-TGTA and the bottom strand is CATGT-TA. Vertical blue lines indicate matches between C, A, C, T, G, T, and A. Dots indicate mismatches between the second C and the first A, and between the fourth T and the fifth T. The third T is aligned with a gap in the top strand.

Optimal alignments

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

- 4 matches
- 3 mismatches
- 0 gaps

CA	CTG	TA
	:	:
CAT	TGT	TA

- 6 matches
- 0 mismatches
- 2 gaps

CA	CTG	T	-	A
CA	-	TGT	T	A

- 5 matches
- 1 mismatch
- 2 gaps

CA	C	-	T	G	TA
	:				
CAT	G	T	-	TA	

Optimal alignments

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

- 4 matches
- 3 mismatches
- 0 gaps

CA	CTG	TA
	:	:
CAT	TGT	TA

- 6 matches
- 0 mismatches
- 2 gaps

CA	CTG	T	-	A
CA	-	TGT	TA	

- 5 matches
- 1 mismatch
- 2 gaps

CA	C	-	T	G	TA
		:			
CAT	G	T	-	TA	

Optimal alignments

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

- 4 matches
- 3 mismatches
- 0 gaps

CA	CTG	TA
	:	:
CAT	TGT	TA

- 6 matches
- 0 mismatches
- 2 gaps

CA	CTG	-	TA
CA	-TG	TTA	

- 5 matches
- 1 mismatch
- 2 gaps

CAC	-	TG	TA
	:		
CAT	G	T	-TA

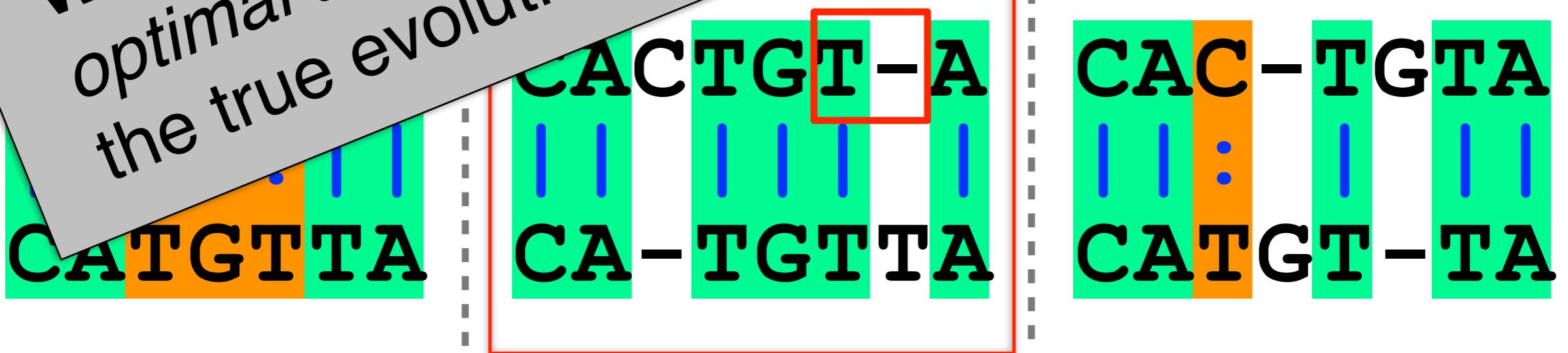
Optimal alignments

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

- 4 matches
- 3 mismatches
- 2 gaps

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

- 3 matches
- 1 mismatch
- 2 gaps



ALIGNMENT FOUNDATIONS

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

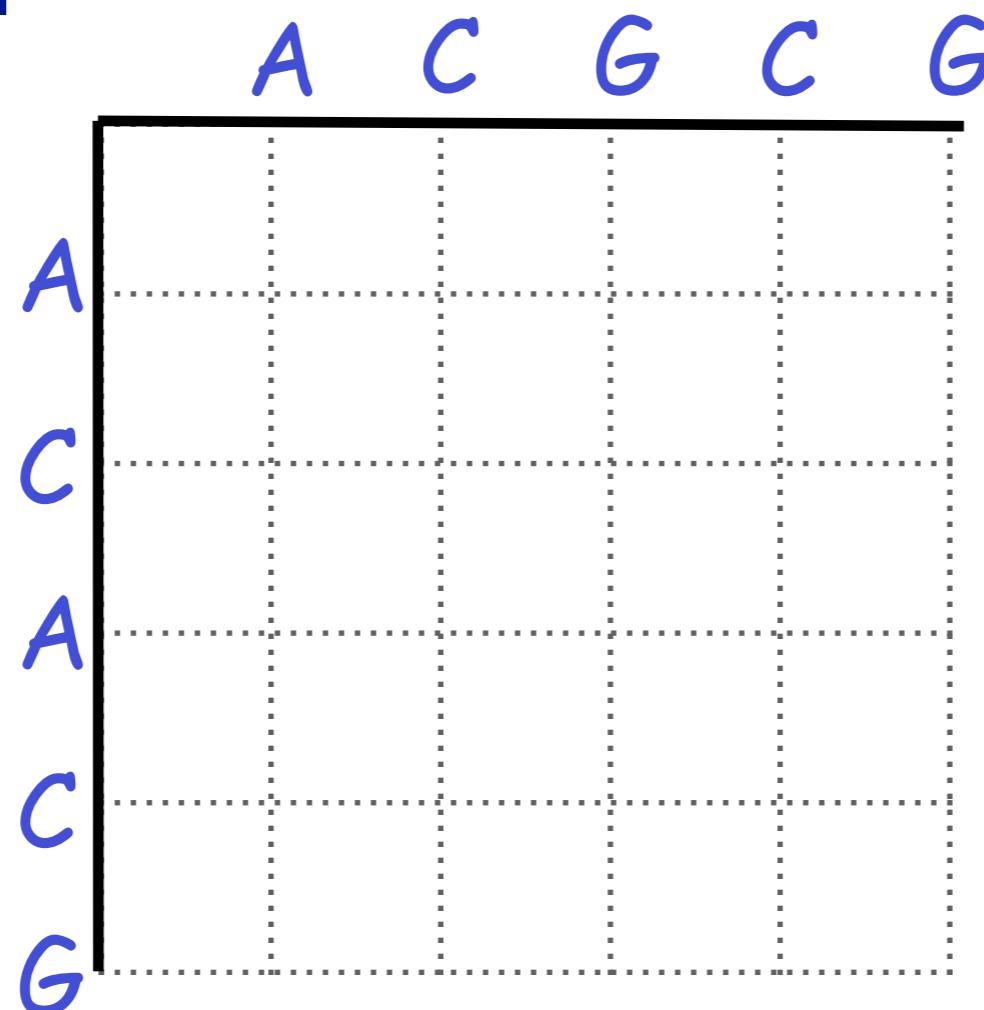
ALIGNMENT FOUNDATIONS

- Why...
 - Why compare biological sequences?
- What...
 - Alignment view of sequence changes during evolution
(matches, mismatches and gaps)
- How...
 - ▶ Dot matrices
 - ▶ D

How do we compute the optimal alignment between two sequences?
 - ▶ BLAST heuristic approach

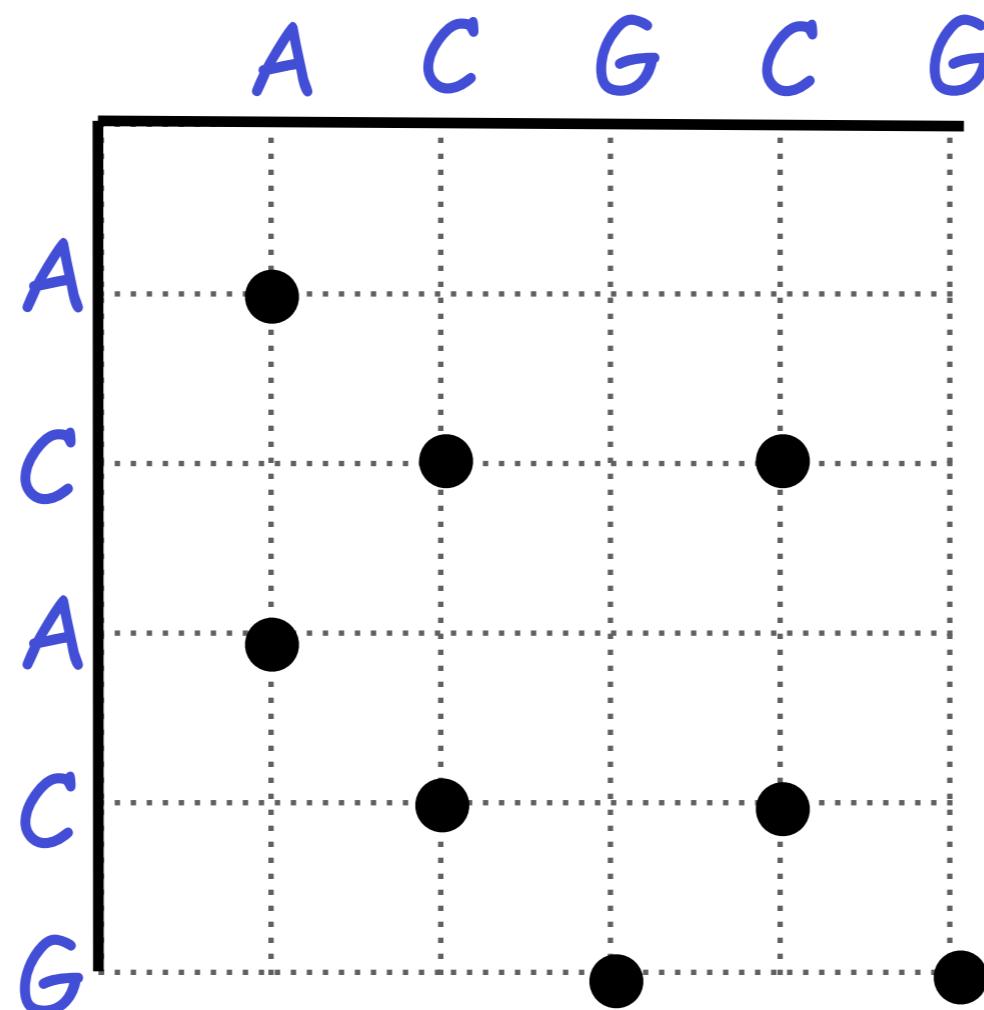
Dot plots: simple graphical approach

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



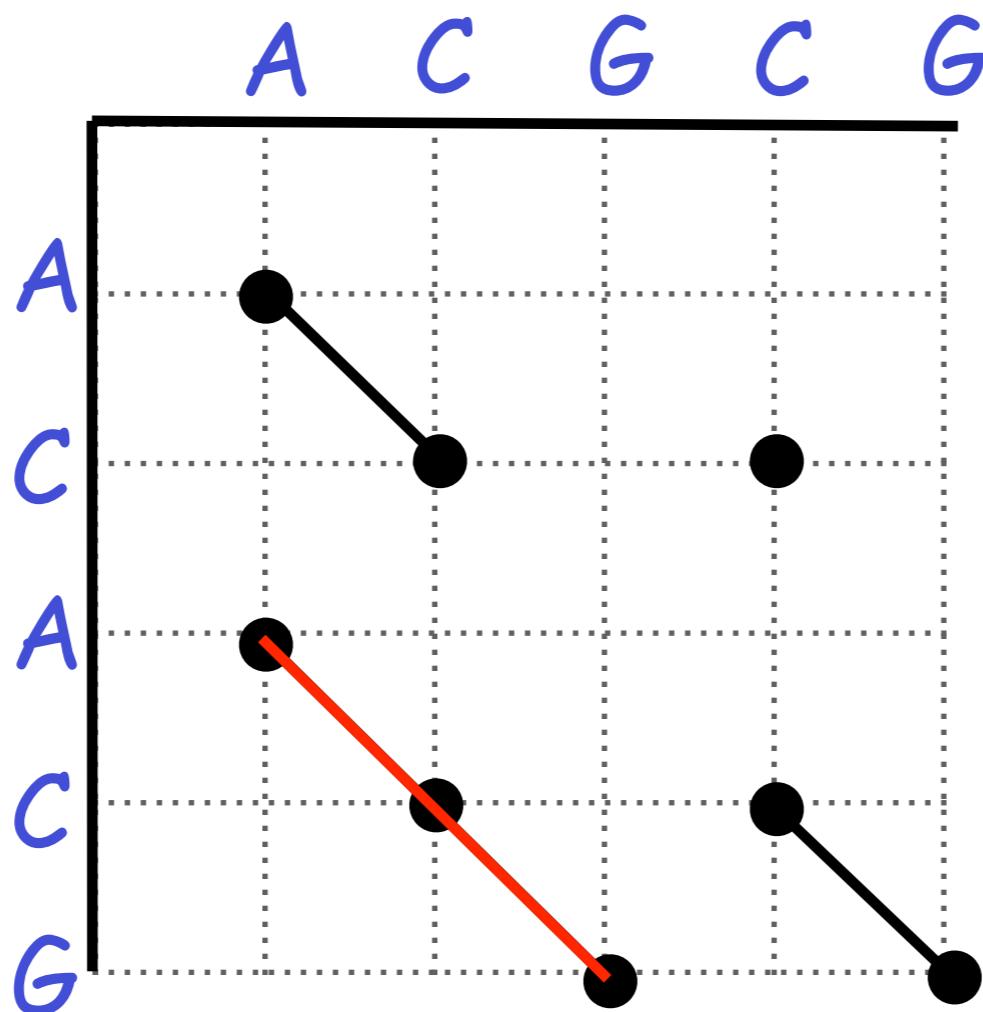
Dot plots: simple graphical approach

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



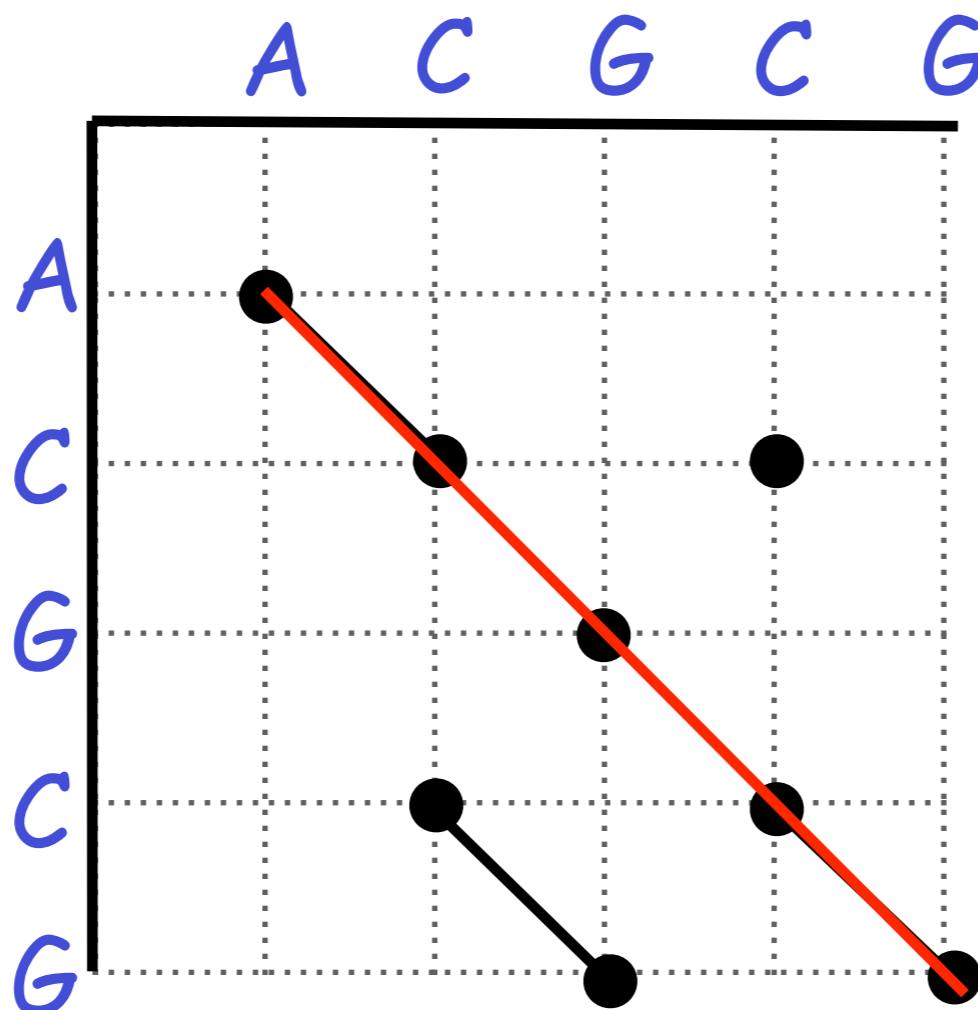
Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



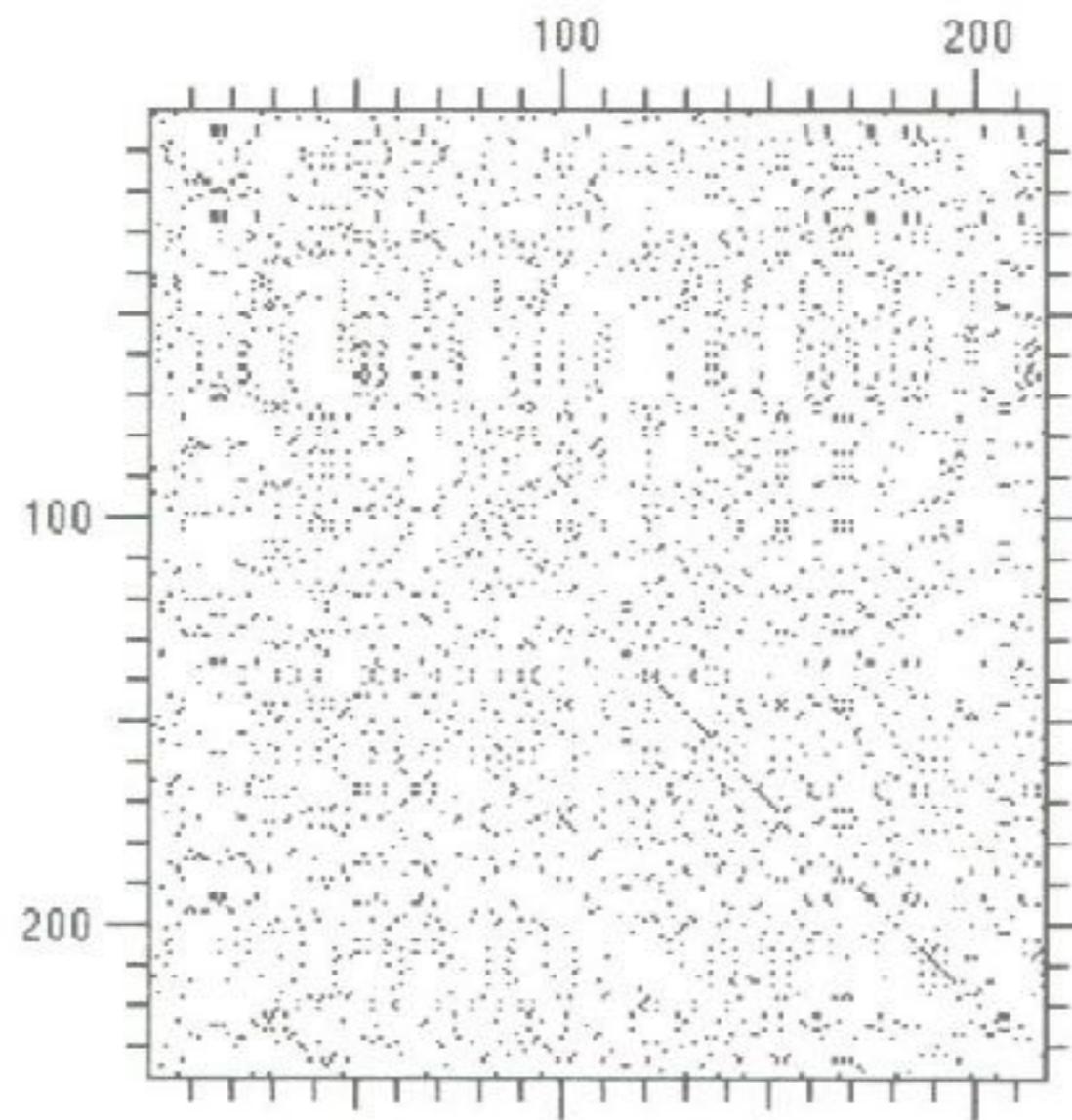
Dot plots: simple graphical approach

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



Dot plots: simple graphical approach

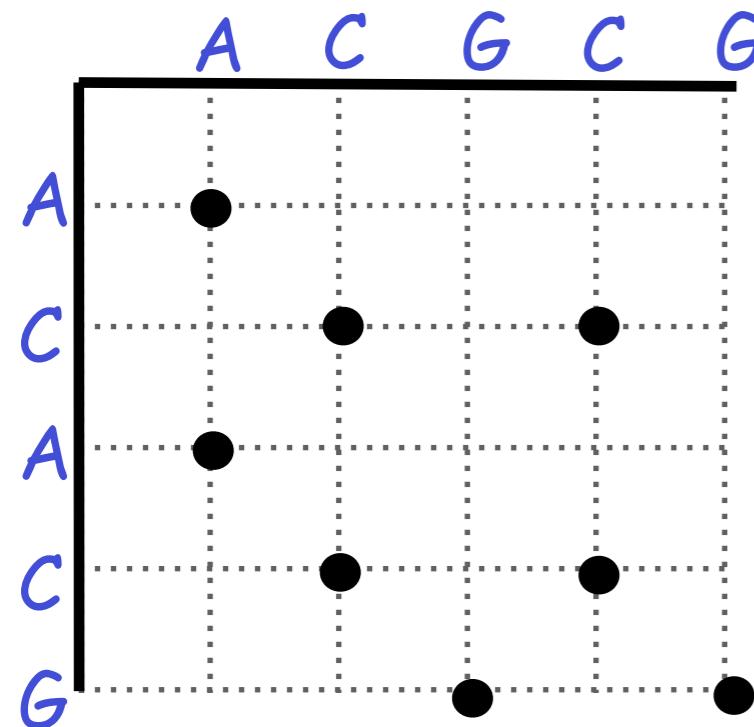
- Dot matrices for long sequences can be noisy



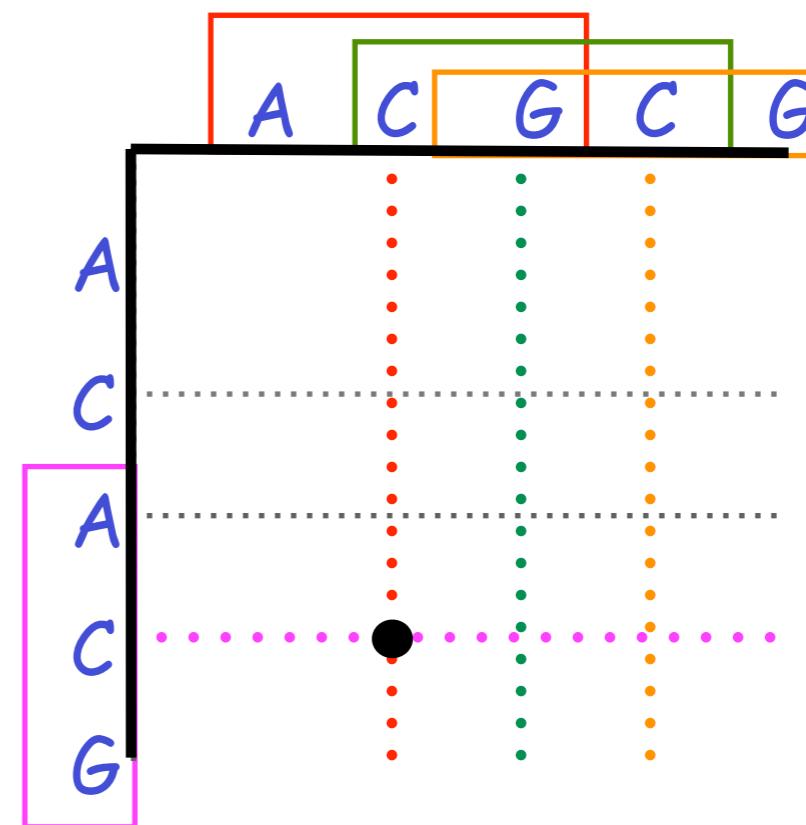
Dot plots: window size and match stringency

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



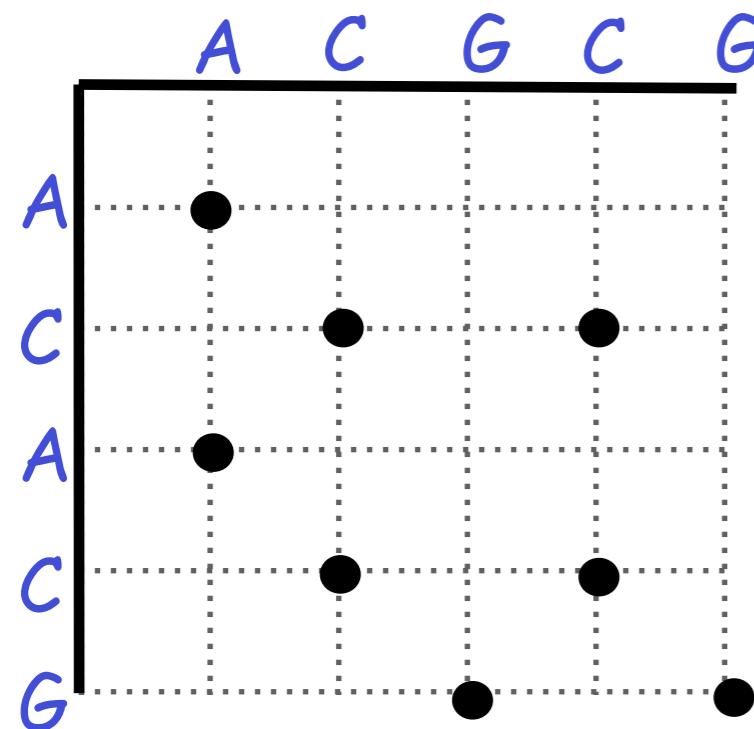
Filter
Window = 3
Stringency = 3



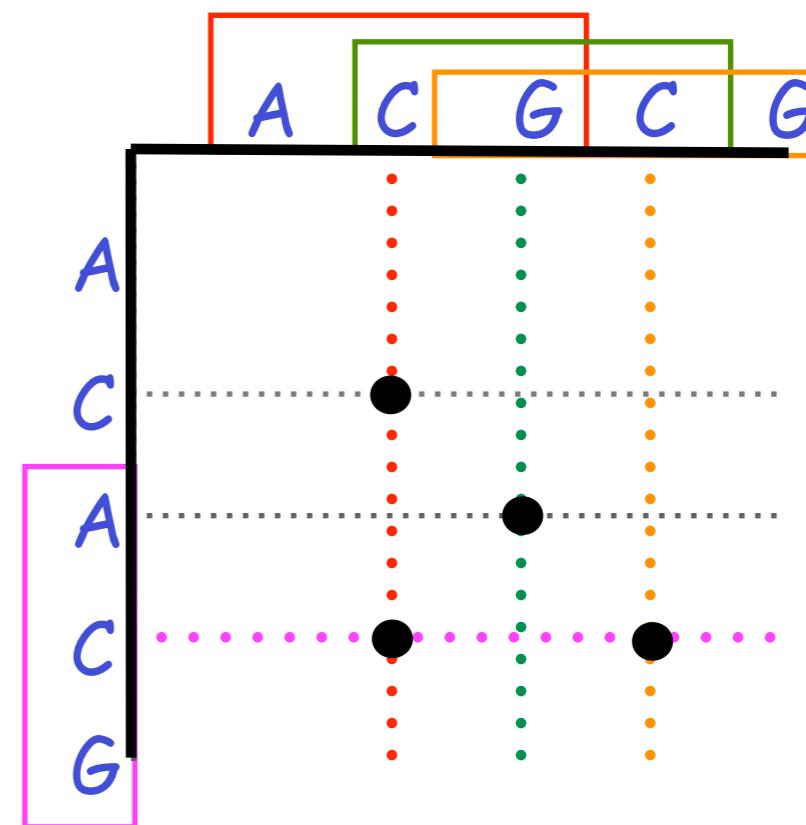
Dot plots: window size and match stringency

Solution: use a window and a threshold

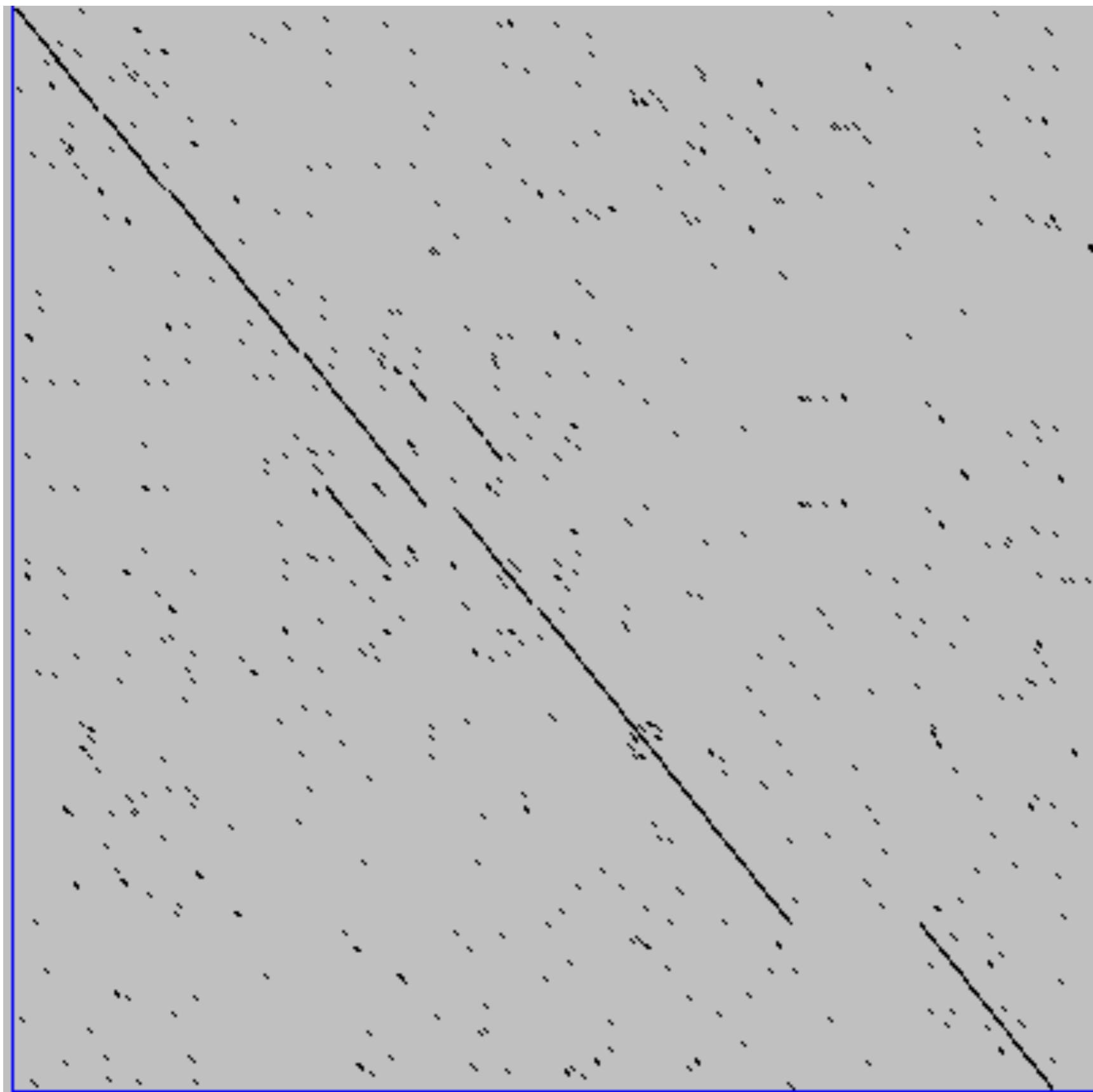
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
 - You have to choose window size and stringency



Filter
Window = 3
Stringency = 2



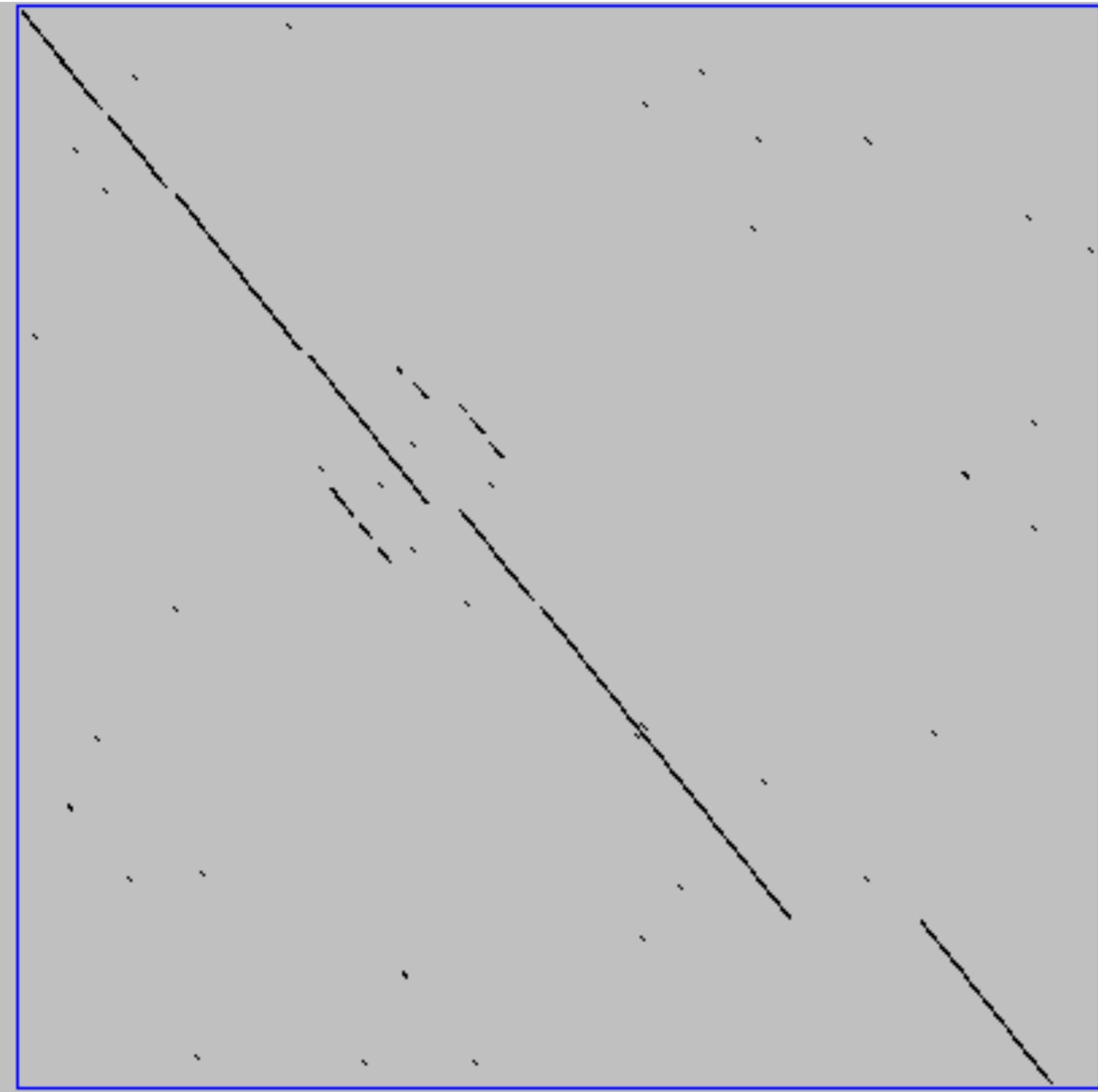
Window size = 5 bases



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

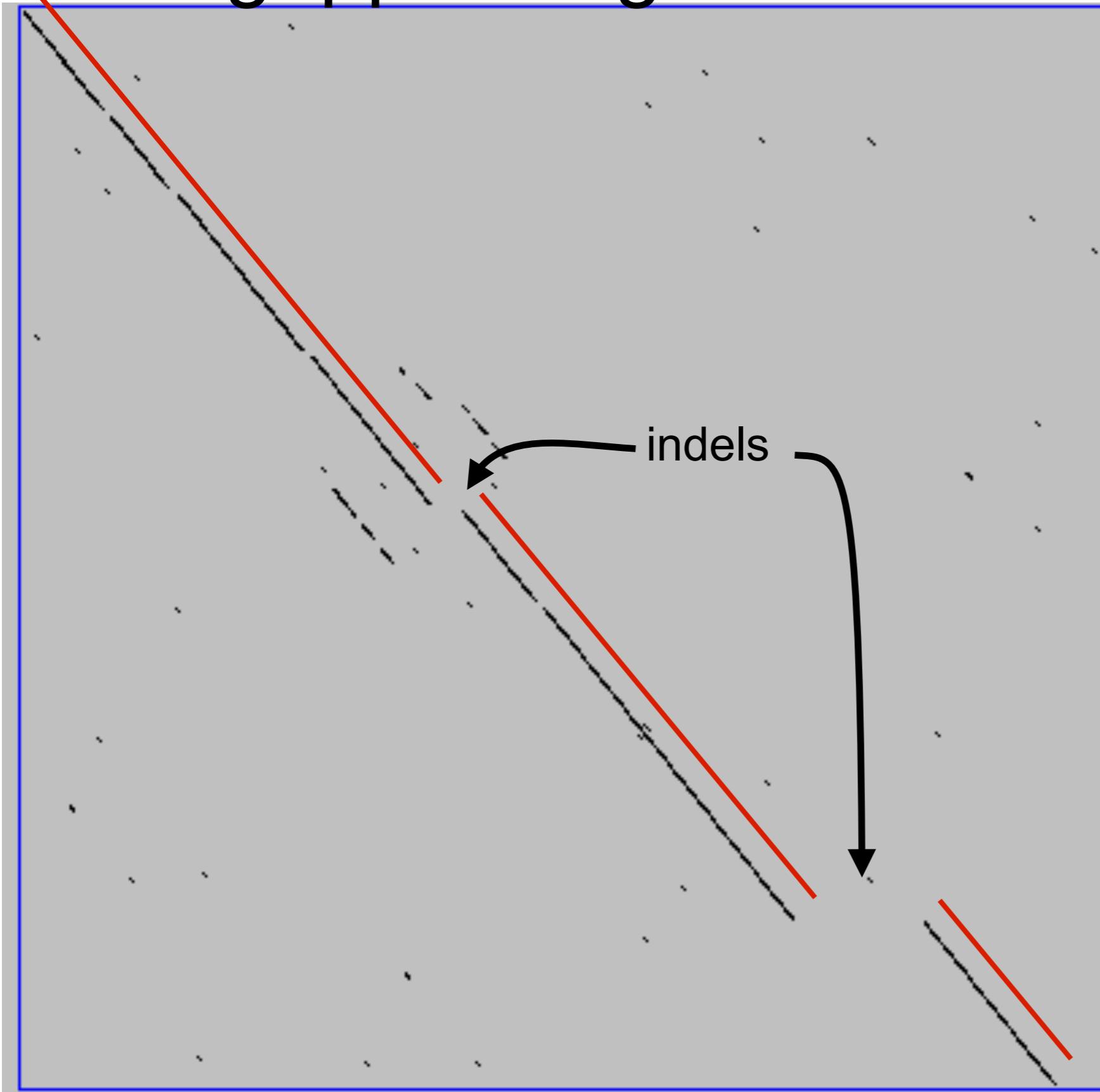


This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

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

Ungapped alignments



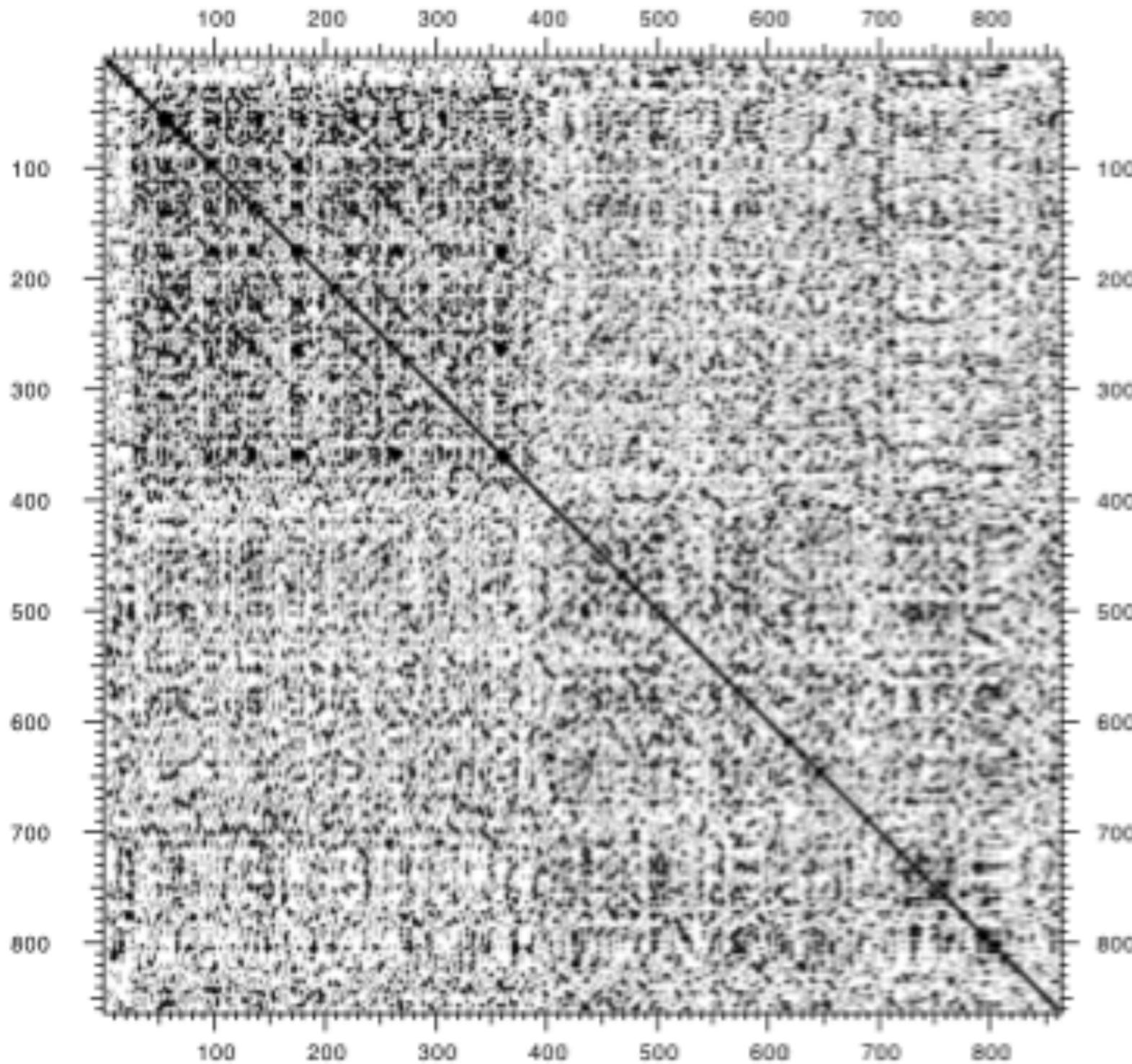
Only **diagonals** can be followed.

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

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

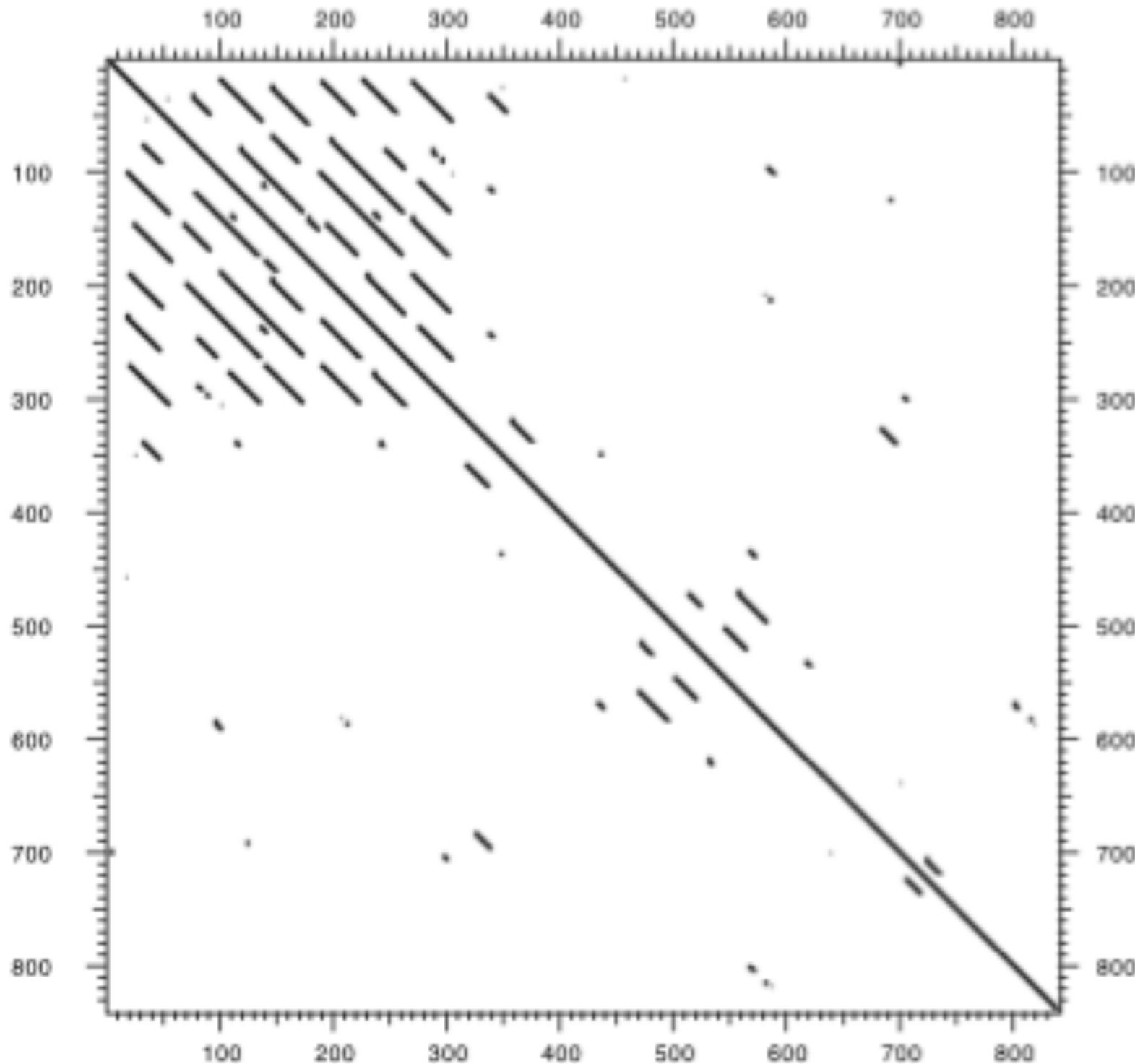


Human LDL receptor
protein sequence
(Genbank P01130)

$$\begin{aligned} W &= 1 \\ S &= 1 \end{aligned}$$

(Figure from Mount, "Bioinformatics sequence and genome analysis")

Repeats



Human LDL receptor
protein sequence
(Genbank P01130)

$$\begin{aligned} W &= 23 \\ S &= 7 \end{aligned}$$

(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/>

Screenshot of the BGGN-213 Dot Plot Comparison of Two Sequences web application.

BGGN-213: Dot Plot Comparison of Two Sequences

Dot plots are a simple graphical approach for the visual comparison of two sequences. They have a long history (see [Meizel and Lenk 1981](#) and references therein) and entail placing one sequence on the vertical axis of a 2D grid (or matrix) and the other on the horizontal. In its simplest form, a dot is placed where the horizontal and vertical sequence values match. That is a dot is produced at position (i,j) if character number i in the first sequence is the same as character number j in the second sequence. More elaborate forms use 'sliding windows' composed of multiple characters and a threshold value, or 'match stringency' for two windows to be considered as matched.

Dot Plot Parameters

Alter the parameters below to change the displayed protein and DNA dot plots. It is important to have a good feel for these parameters when we get to alignment heuristic approaches later.

Window Size: (1 to 10)

Moving window step size: (1 to 10)

Match stringency: (1 to 10)

Protein Dot Plot
wsize = 3 wstep = 3 , nmatch = 2

The Protein Dot Plot shows a diagonal band of dots from (0,0) to (150,150) on a 2D grid. The x-axis is labeled "Sequence 1" and the y-axis is labeled "Sequence 2", both ranging from 0 to 150. The plot title is "Protein Dot Plot" with parameters "wsize = 3 wstep = 3 , nmatch = 2".

DNA Dot Plot
wsize = 3 wstep = 3 , nmatch = 2

The DNA Dot Plot shows a dense cluster of dots forming a diagonal band on a 2D grid. The x-axis is labeled "Sequence 1" and the y-axis is labeled "Sequence 2", both ranging from 0 to 150. The plot title is "DNA Dot Plot" with parameters "wsize = 3 wstep = 3 , nmatch = 2".

Questions for discussion:

- Why does the DNA sequence have more dots than the protein sequence plot?
- How can we increase the signal to noise ratio?
- What does a 'Match stringency' larger than 'Window size' yield, and why?

ALIGNMENT FOUNDATIONS

- **Why...**
 - Why compare biological sequences?
- **What...**
 - Alignment view of sequence changes during evolution
(matches, mismatches and gaps)
- **How...**
 - ▶ Dot matrices
 - ▶ Dynamic programming
 - 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

		j	Sequence 2				
		-	D	P	L	E	
Sequence 1		-	0	-2	-4	-6	-8
—	-	—	0	-2	-4	-6	-8
D	D	-2					
P	P	-4					
M	M	-6					
E	E	-8					

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

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

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

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

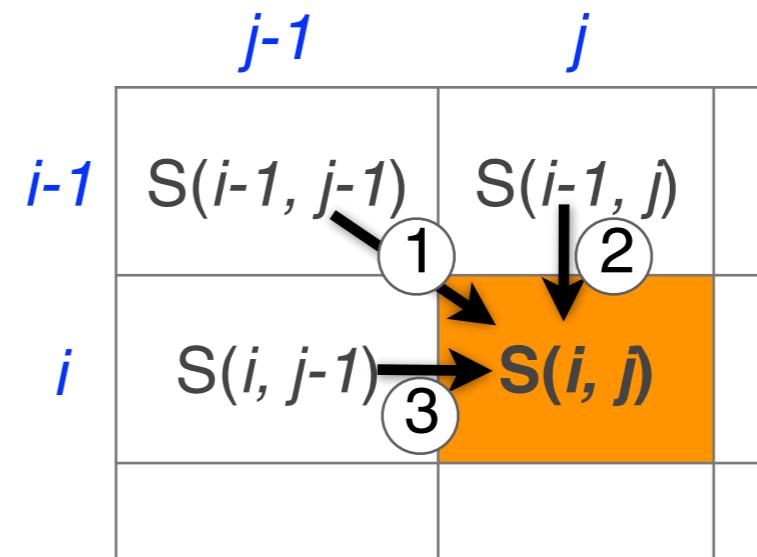
Seq1 : DPME
Seq2 : ----

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

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

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

$$S(i, j) = \text{Max} \begin{cases} S(i-1, j-1) + (\text{mis})\text{match} \\ S(i-1, j) + \text{gap penalty} \\ S(i, j-1) + \text{gap penalty} \end{cases}$$

1
2
3

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

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

- Alignment D
D
- ① $(0) + (+1) = +1 \leq (D-D) \text{ match!}$
- ② $(-2) + (-2) = -4$
- ③ $(-2) + (-2) = -4$

Scoring the alignment matrix

- At each step, the score in the current cell is determined 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)

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

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

① $(-2)+(-1) = -3 \leq (D-P)$ mismatch!

Alignment

② $(-4)+(-2) = -6$

D-
DP

③ $(1)+(-2) = -1$

Scoring the alignment matrix

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

	-	D	P	L	j	E
-	0	-2	-4	-6		-8
D	-2	1	-1	-3		
P	-4					
M	-6					
E	-8					

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

① $(-4)+(-1) = -5 \leq (D-L)$ mismatch

Alignment

② $(-6)+(-2) = -8$

D--
DPL

③ $(-1)+(-2) = -3$

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.

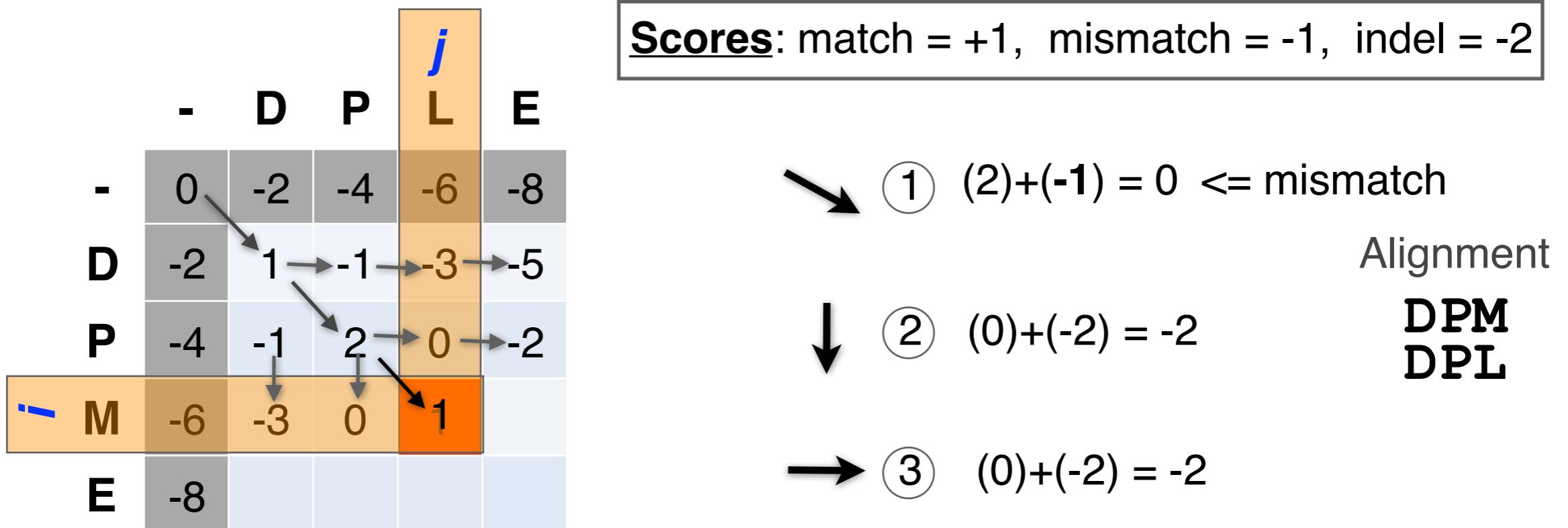
		j				
		-	D	P	L	E
-		0	-2	-4	-6	-8
D	-	-2	1	-1	-3	-5
P	-	-4	-1	2	0	
M	-6					
E	-8					

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

- Alignment
- DP-
DPL
- 1 $(-1)+(-1) = -2$
 - 2 $(-3)+(-2) = -5$
 - 3 $(2)+(-2) = 0$

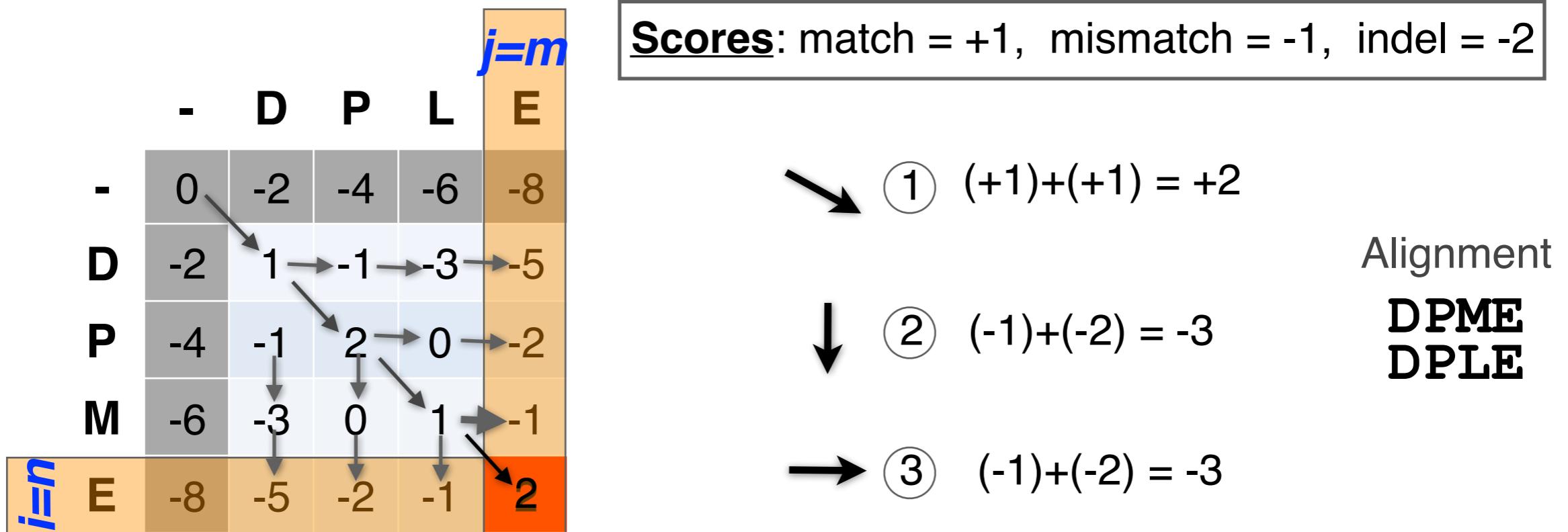
Scoring the alignment matrix

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



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)



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

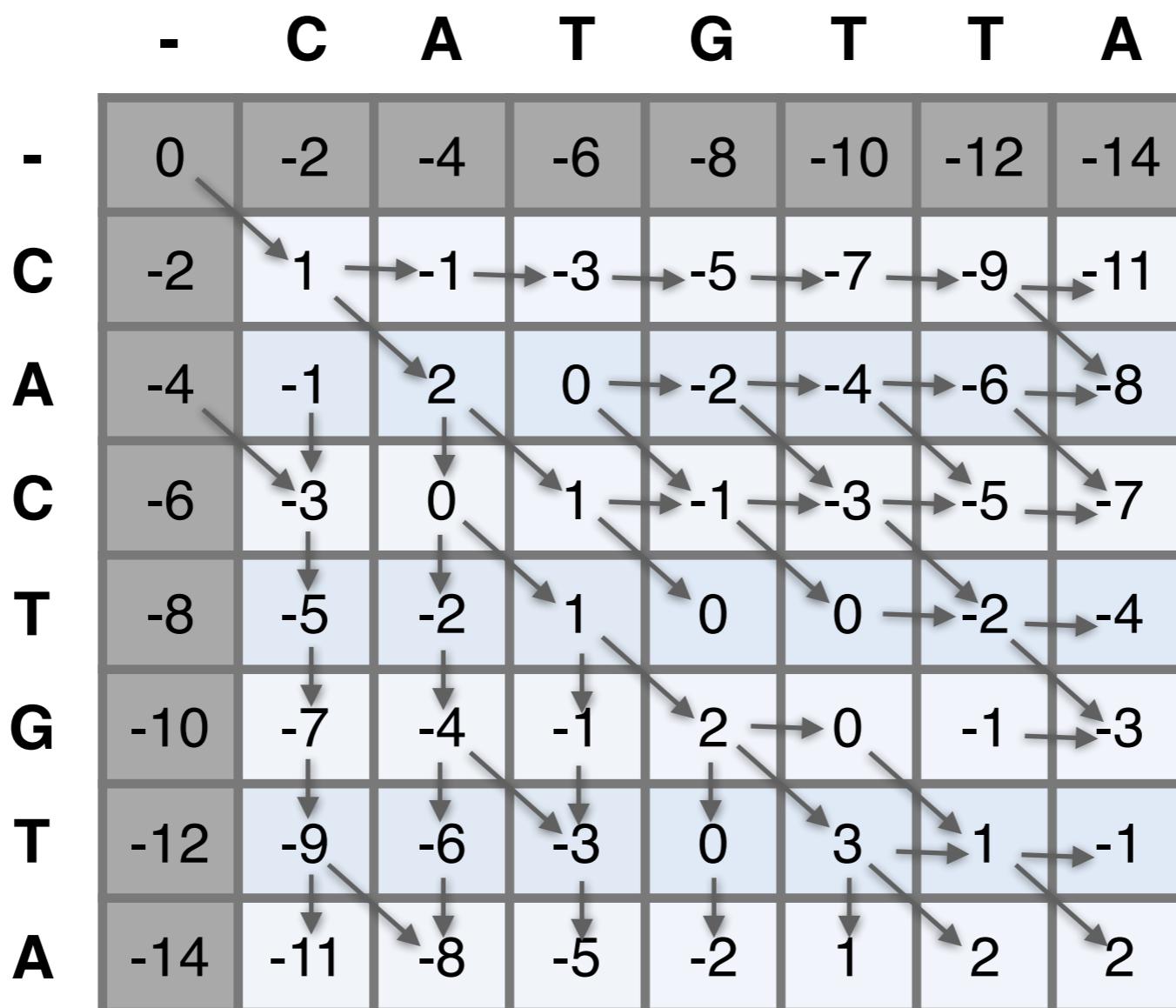
Scores: match = +1, mismatch = -1, indel = -2

	-	D	P	L	E
-	0	-2	-4	-6	-8
D	-2	1	-1	-3	-5
P	-4	-1	2	0	-2
M	-6	-3	0	1	-1
E	-8	-5	-2	-1	2

Alignment
DPME
DPLE

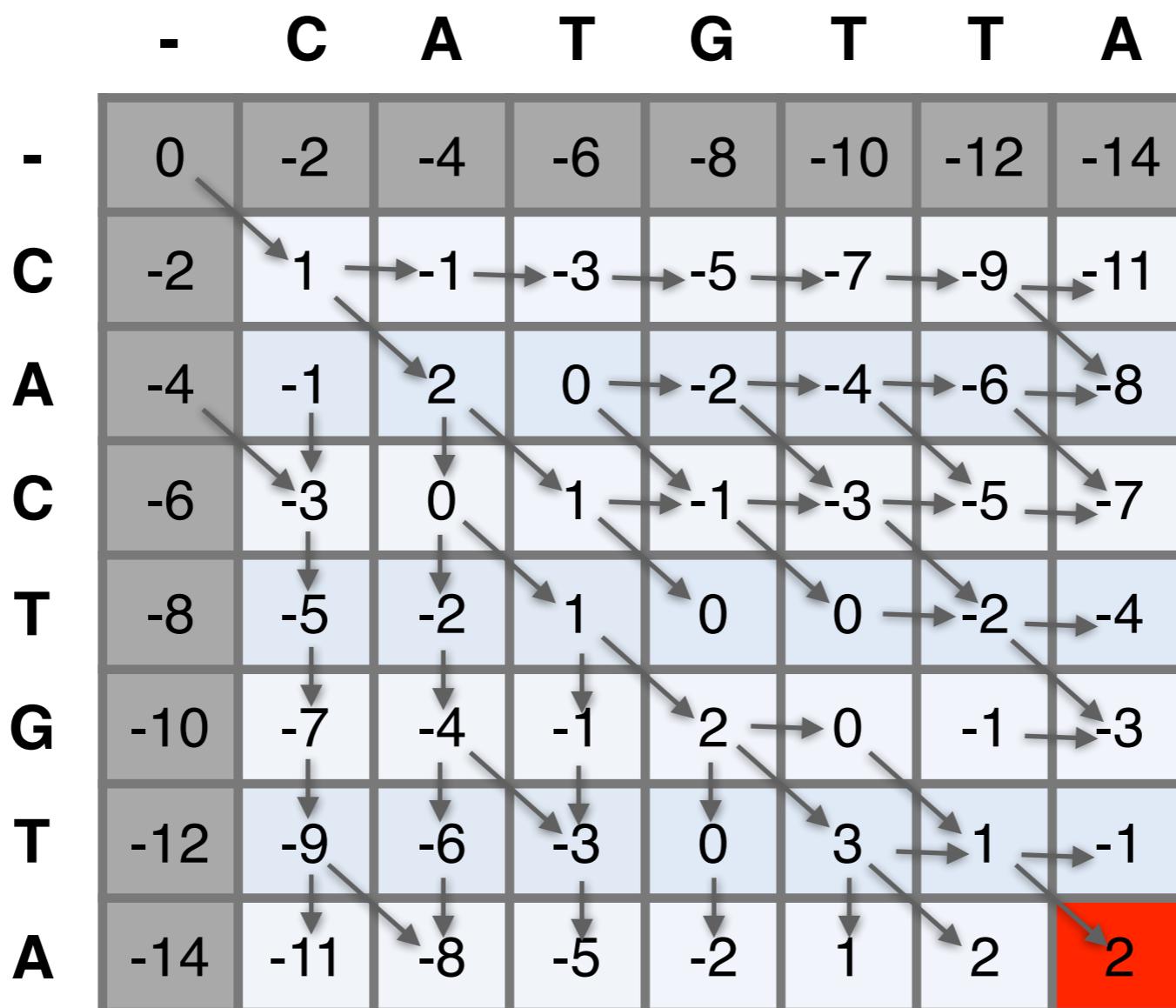
Questions:

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



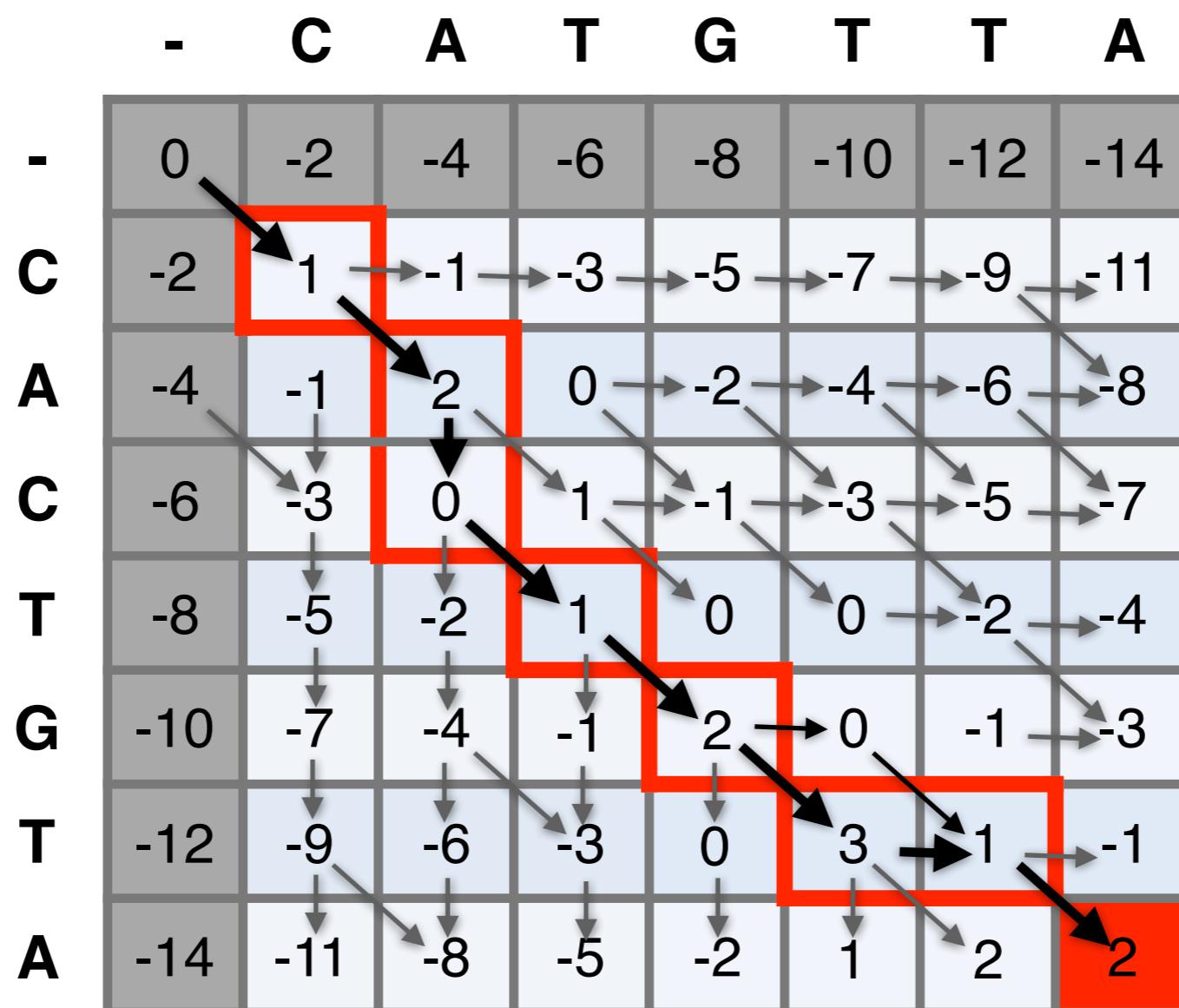
Questions:

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



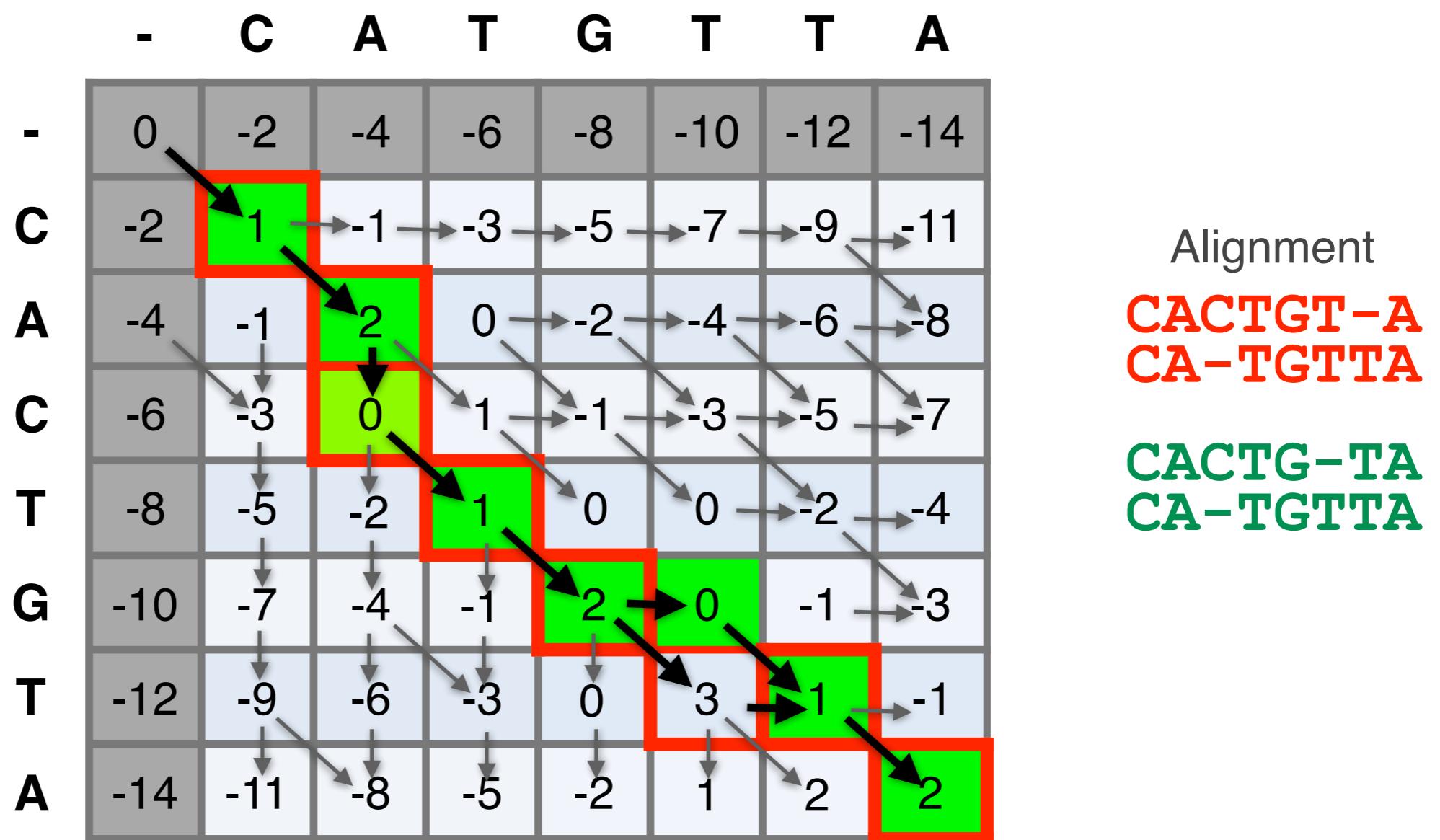
Questions:

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



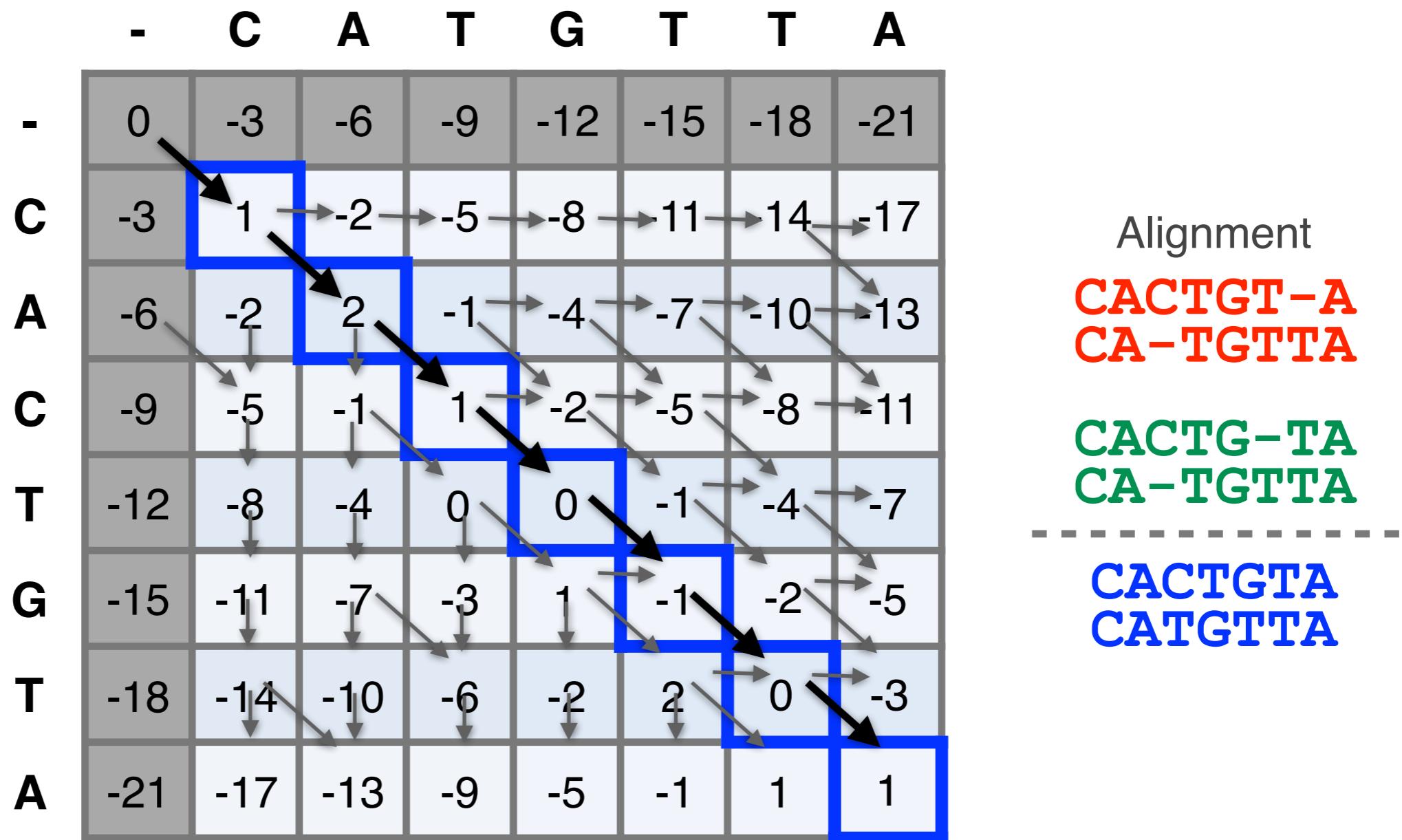
More than one alignment possible

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



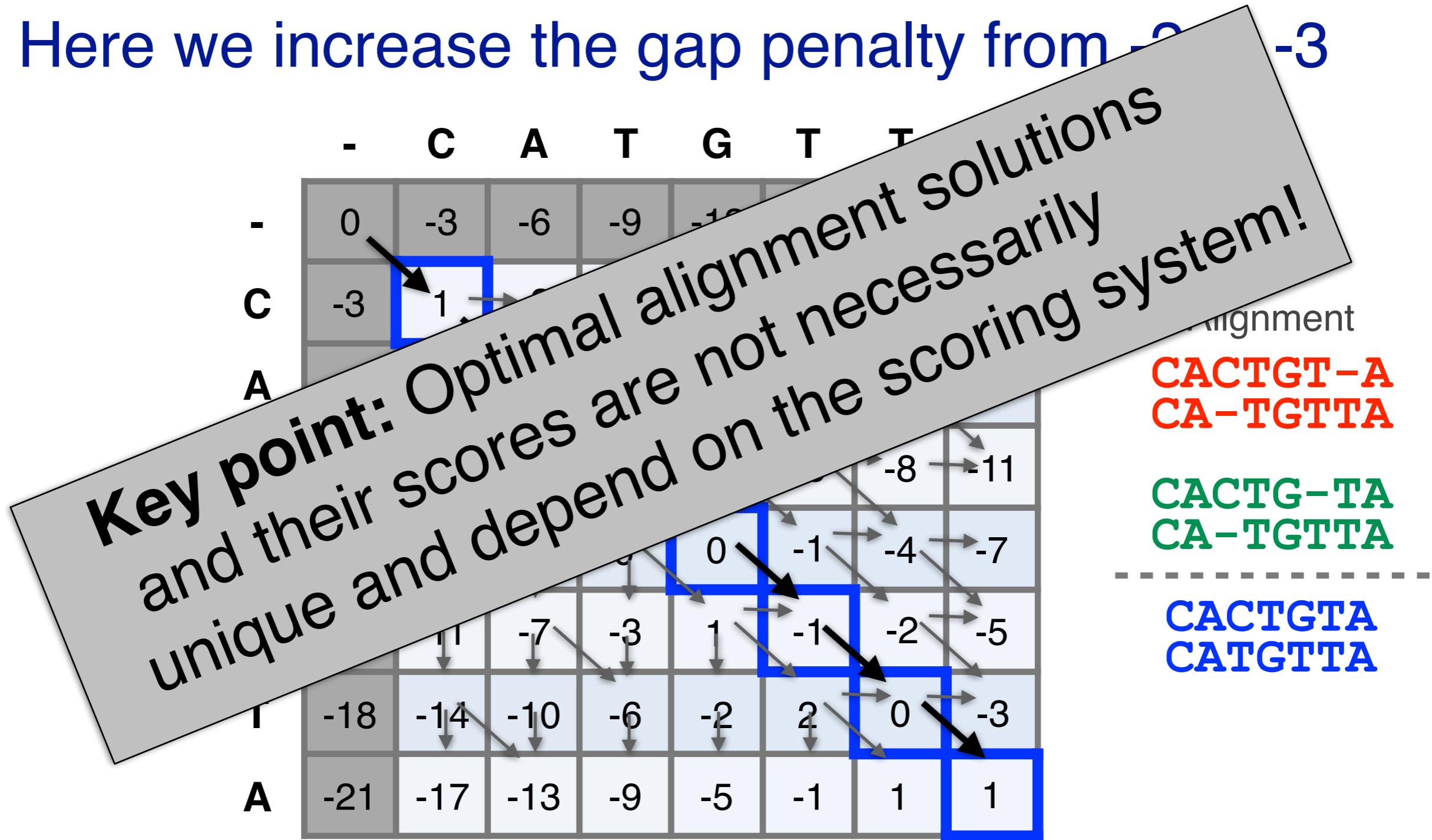
The alignment and score are dependent on the scoring system

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



The alignment and score are dependent on the scoring system

- Here we increase the gap penalty from -9 to -3



Your Turn!

Hands-on worksheet **Sections 2 & 3**

Match: +2

Mismatch: -1

Gap: -2

	A	G	T	T	C
0					
A					
T					
T					
G					
C					

NW DYNAMIC PROGRAMMING

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

	A	G	T	T	C	
A	0	-2	-4	-6	-8	-10
T	-2	+2	0	-2	-4	-6
T	-4	0	+1	+2	0	-2
T	-6	-2	-1	+3	+4	+2
G	-8	-4	0	+1	+2	+3
C	-10	-6	-2	-1	0	+4

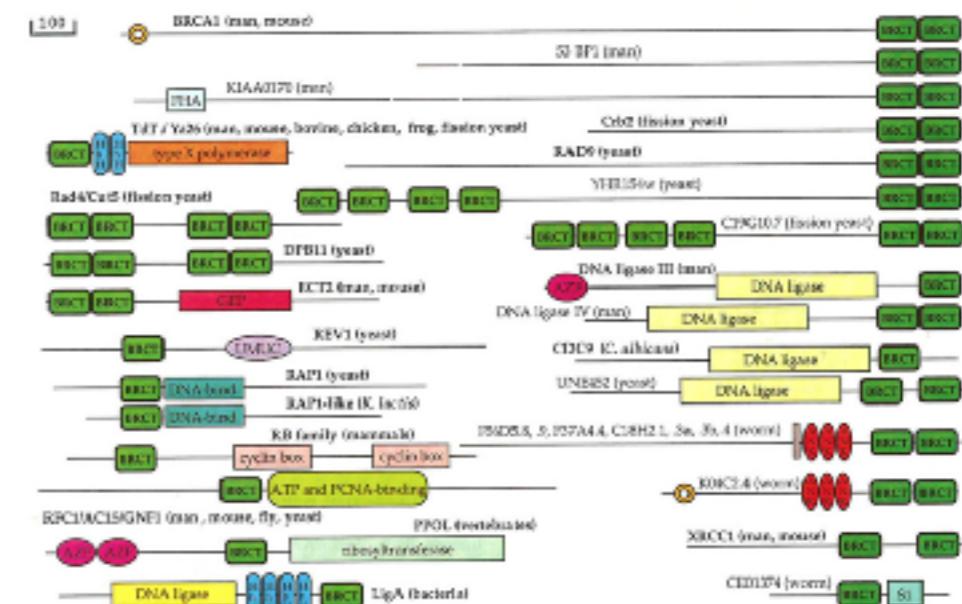
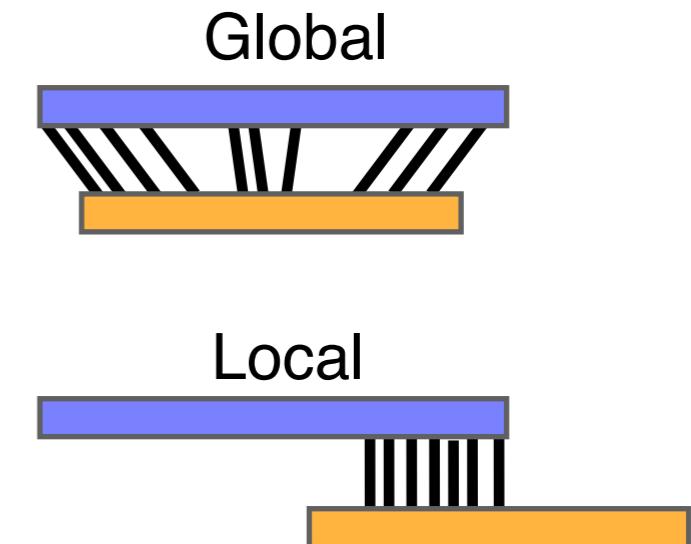
The diagram illustrates the NW Dynamic Programming algorithm for sequence alignment. The grid shows scores for matching (red arrows), mismatching (grey arrows), and gaps (grey arrows). The final score is highlighted in red at the bottom-right corner (+4).

ALIGNMENT FOUNDATIONS

- **Why...**
 - Why compare biological sequences?
- **What...**
 - Alignment view of sequence changes during evolution
(matches, mismatches and gaps)
- **How...**
 - ▶ Dot matrices
 - ▶ Dynamic programming
 - 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

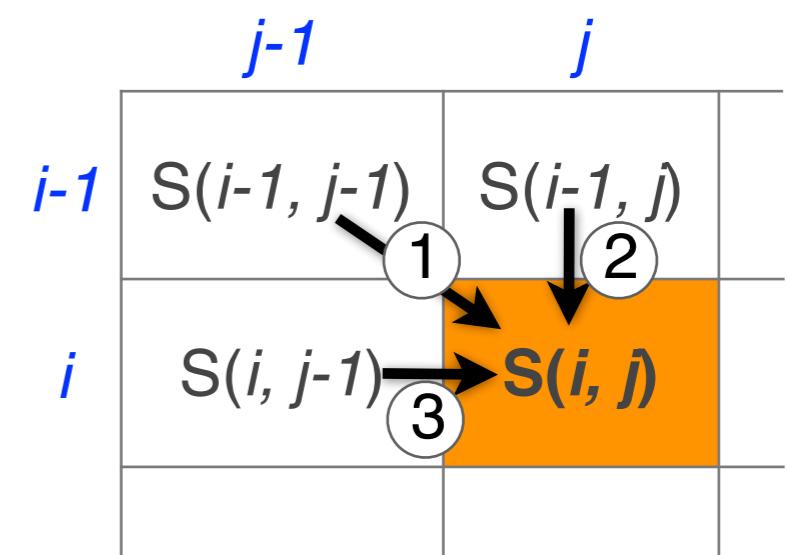
Smith, T.F. & Waterman, M.S. (1981) “Identification of common molecular subsequences.” J. Mol. Biol. 147:195-197.

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

$$S(i, j) = \text{Max} \left\{ \begin{array}{l} S(i-1, j-1) + (\text{mis})\text{match} \\ S(i-1, j) - \text{gap penalty} \\ S(i, j-1) - \text{gap penalty} \\ 0 \end{array} \right\}$$

(1)
(2)
(3)
(4)



Sequence 1

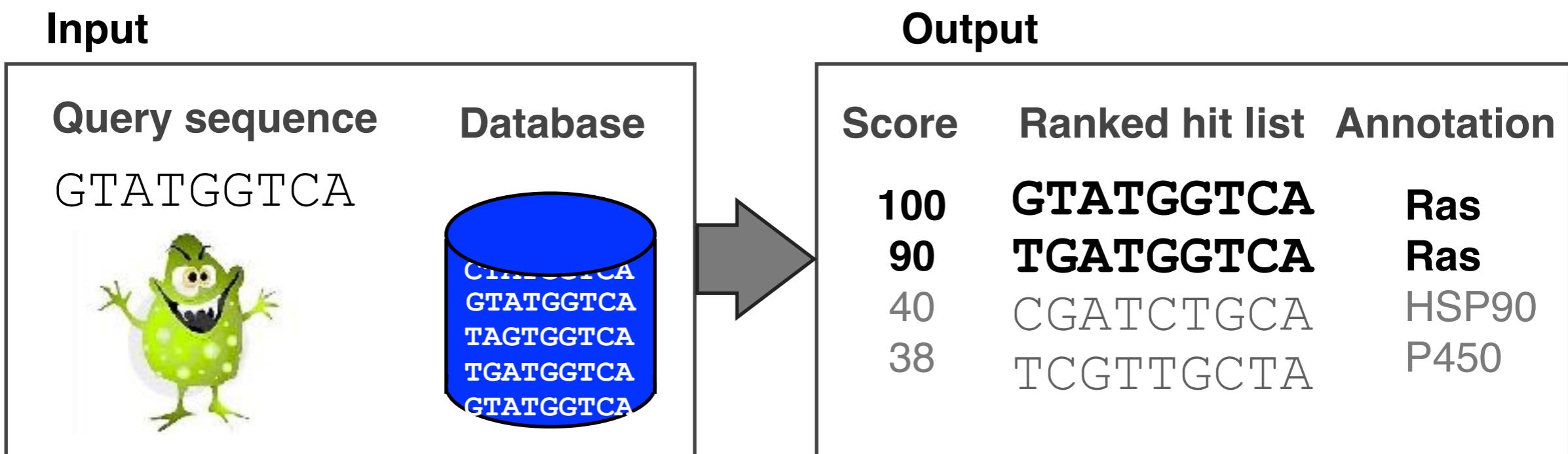
	-	C	A	G	C	C	U	C	G	C	U	U	A	G
-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U	0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	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**

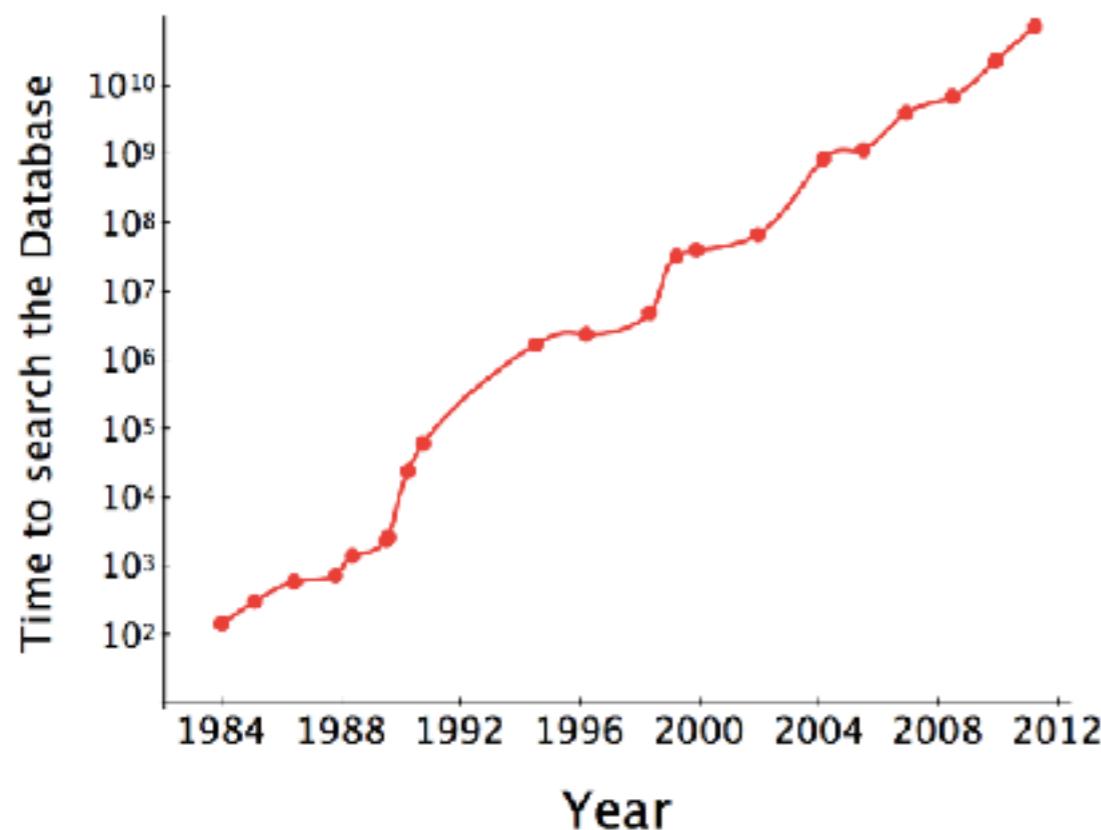
Local alignments can be used for database searching

- **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

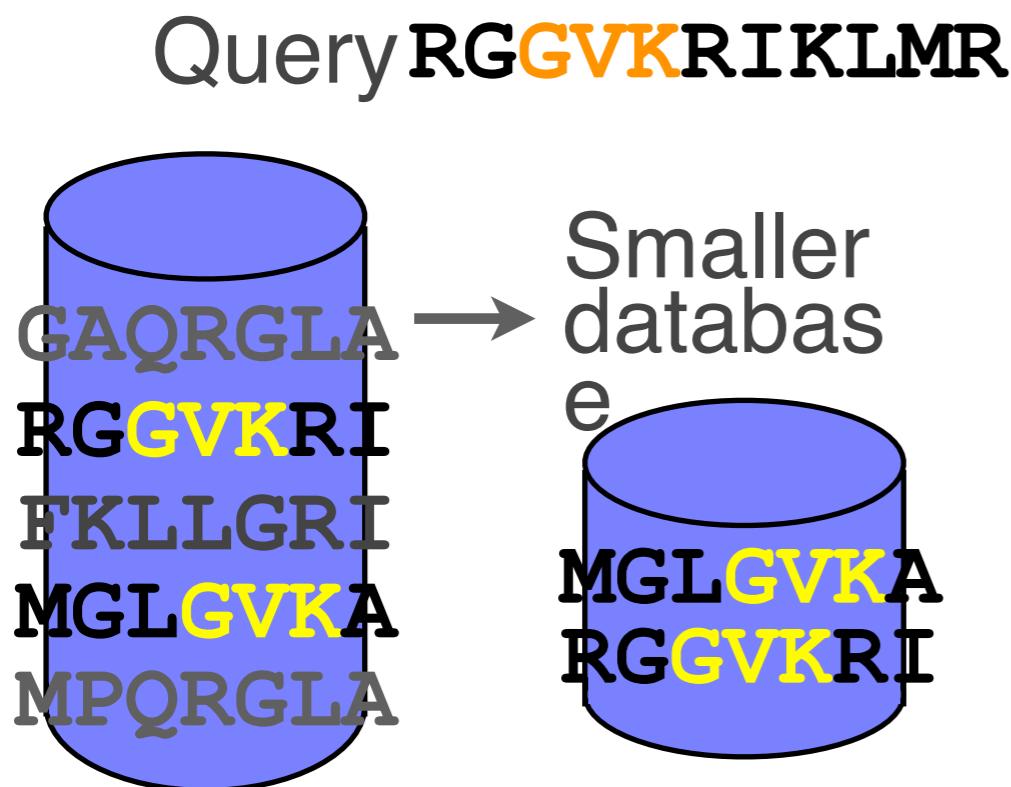
- Due to the rapid growth of sequence databases, search algorithms have to be both efficient and sensitive
 - Time to search with SW is proportional to $m \times n$ (m is length of query, n is length of database), **too slow for large databases!**



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 $m \times n$ (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 programming
 - Global alignment
 - Local alignment
 - ▶ BLAST heuristic approach

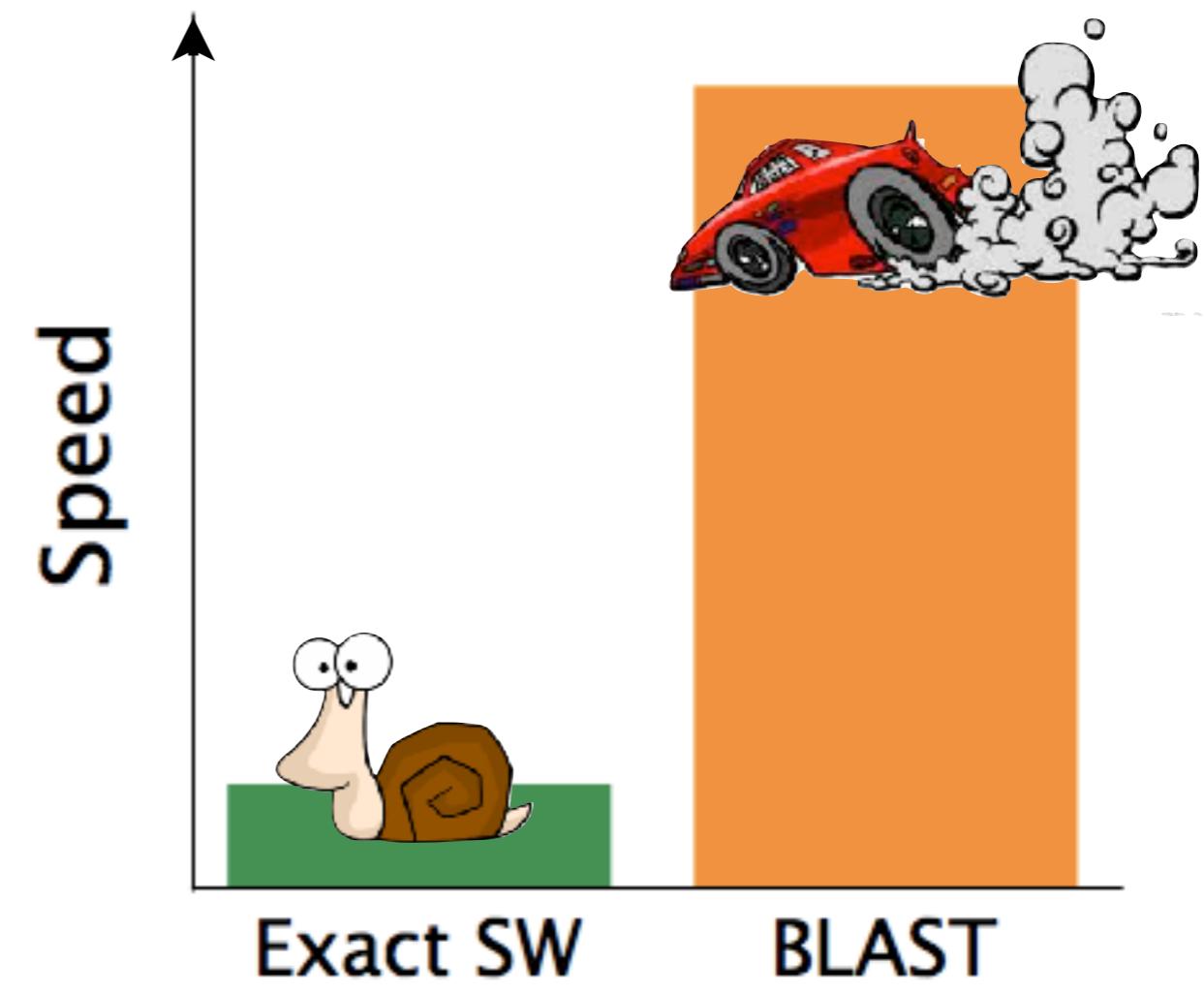
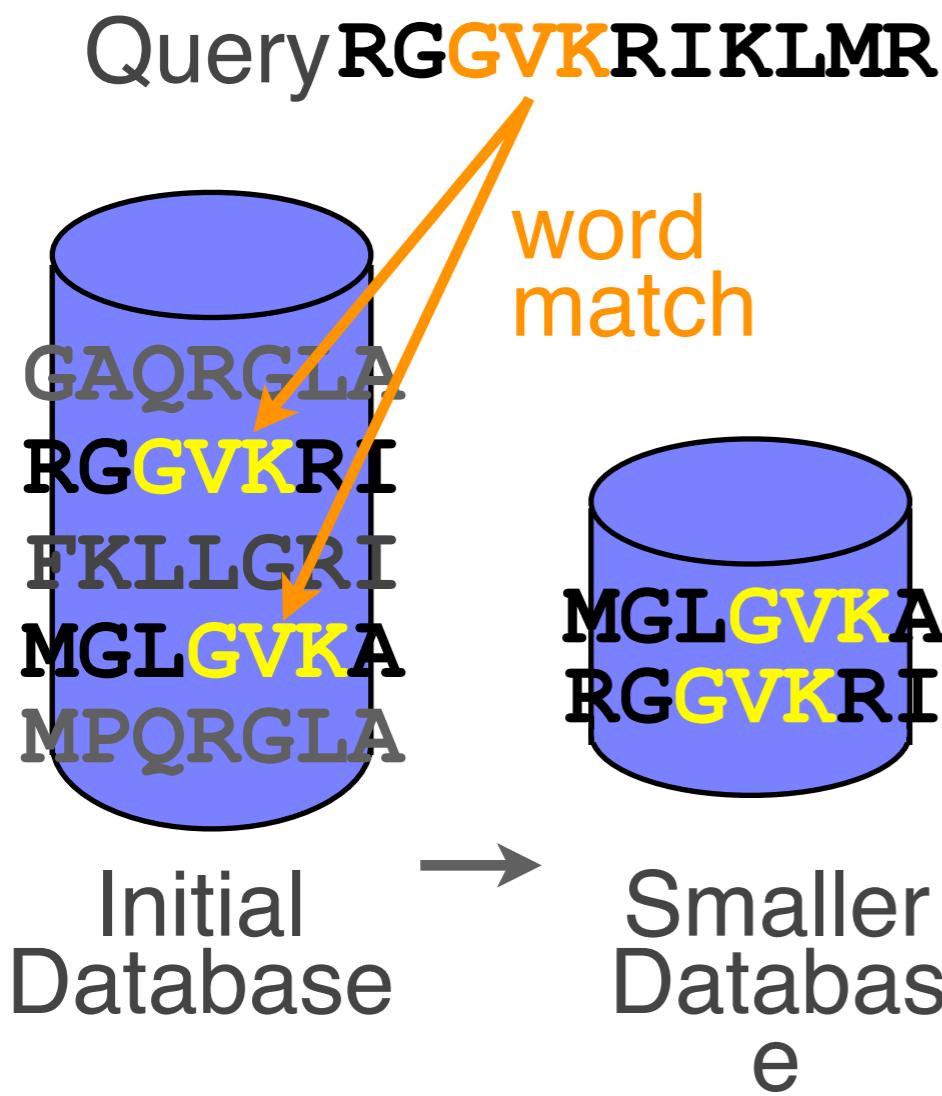
Rapid, heuristic versions of Smith–Waterman: BLAST

- 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) is a simplified form of Smith-Waterman. It is popular because it is **fast**.
 - BLAST finds regions of similarity between sequences
 - BLAST does not find all matches, but it finds many matches by scanning for local 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)
- The sensitivity in exchange for speed
In contrast to SW, BLAST is not guaranteed to find optimal alignments

- BLAST uses this pre-screening heuristic approximation resulting in 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

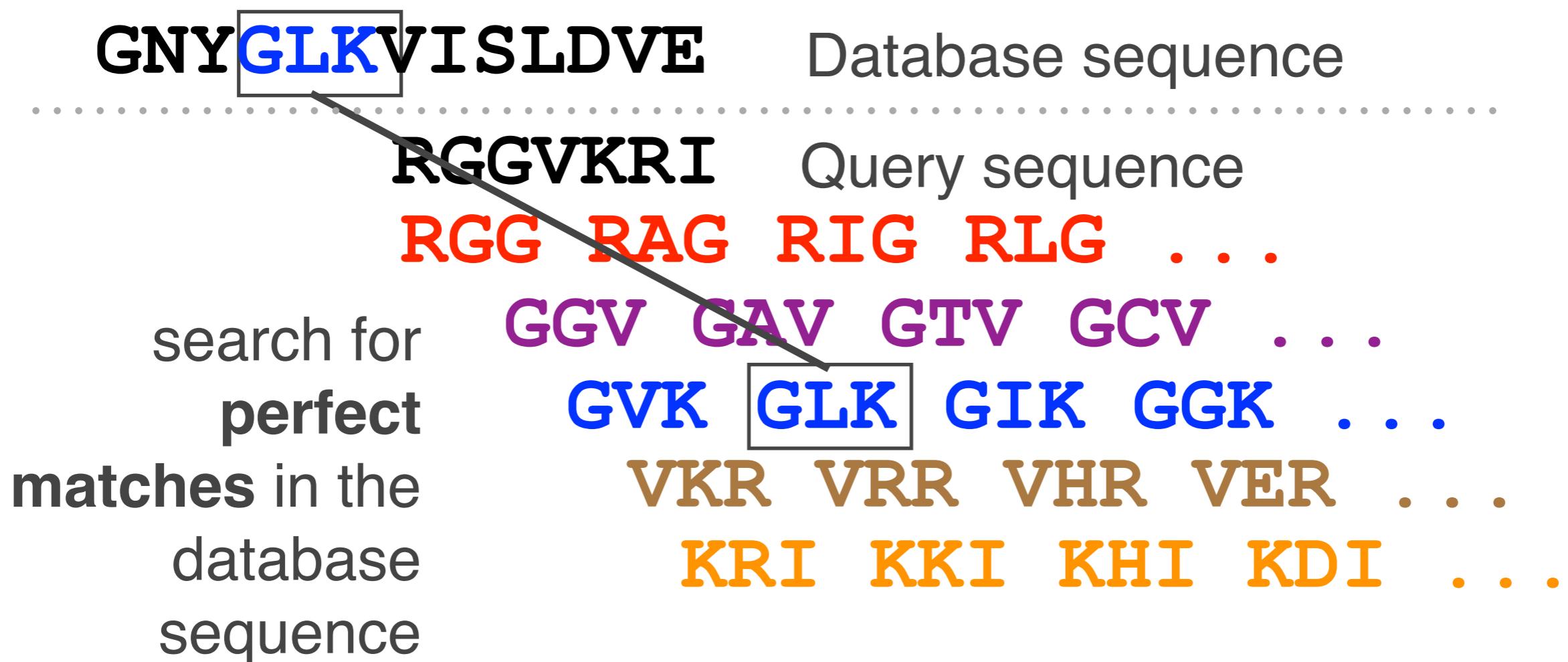
- **Phase 2:** expand word pairs to include those similar to query (defined as those above a similarity threshold to original word, i.e. match scores in substitution matrix)

extend list of
words similar
to query

RGGVKRI Query sequence
RGG RAG RIG RLG . . .
GGV GAV GTV GCV . . .
GVK GAK GIK GGK . . .
VKR VRR VHR VER . . .
KRI KKI KHI KDI . . .

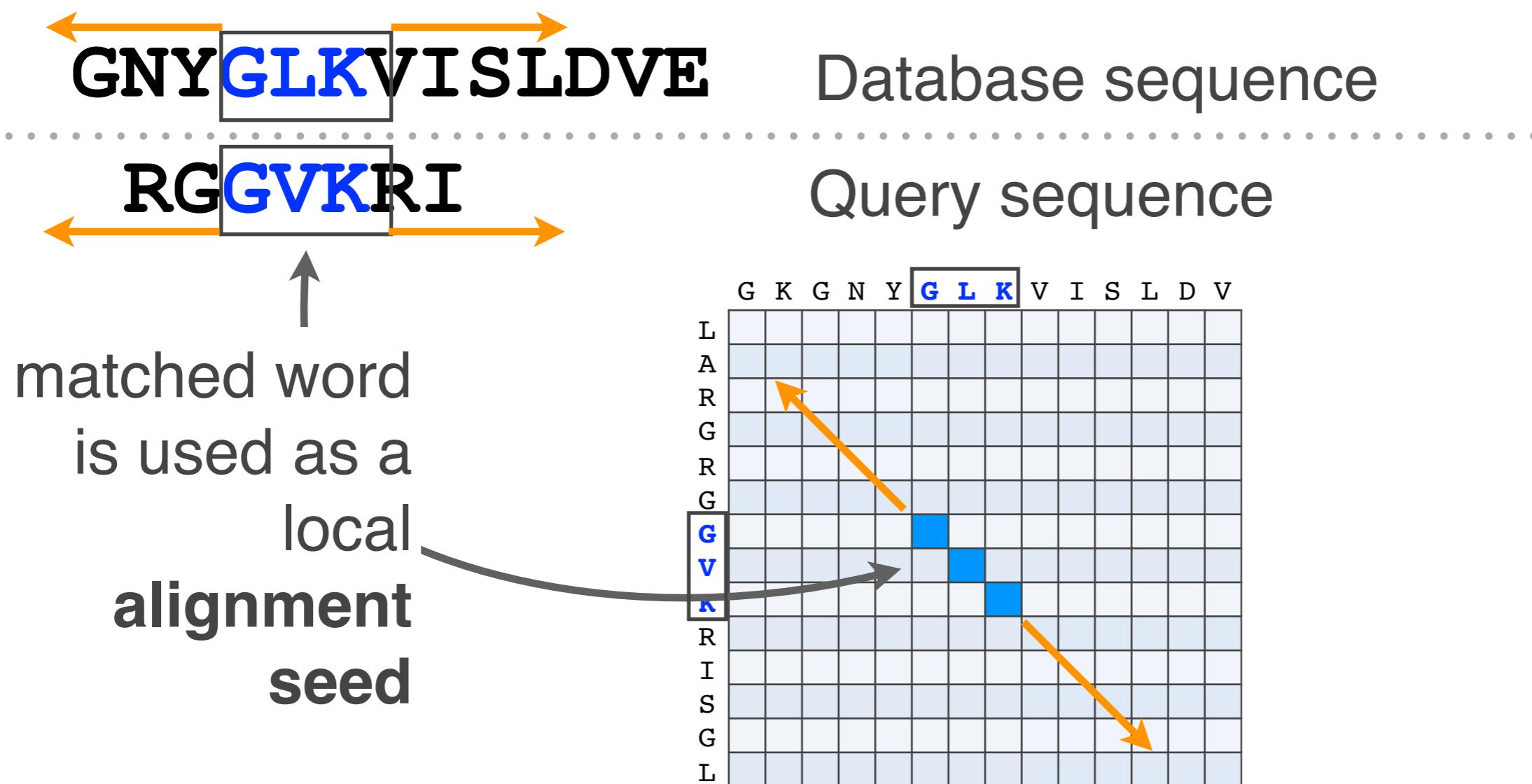
Blast

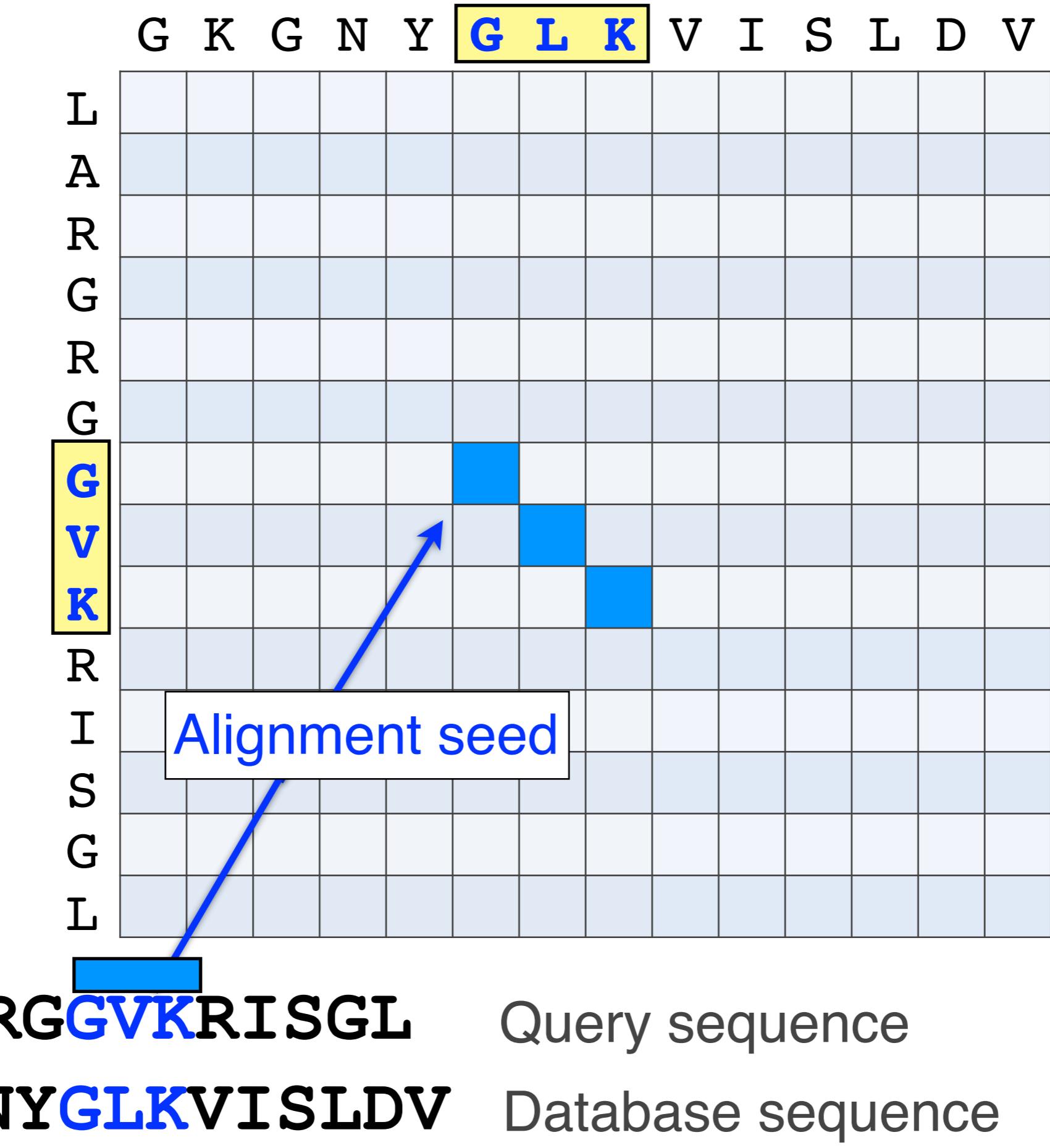
- Phase 3: a database is scanned to find sequence entries that match the compiled word list

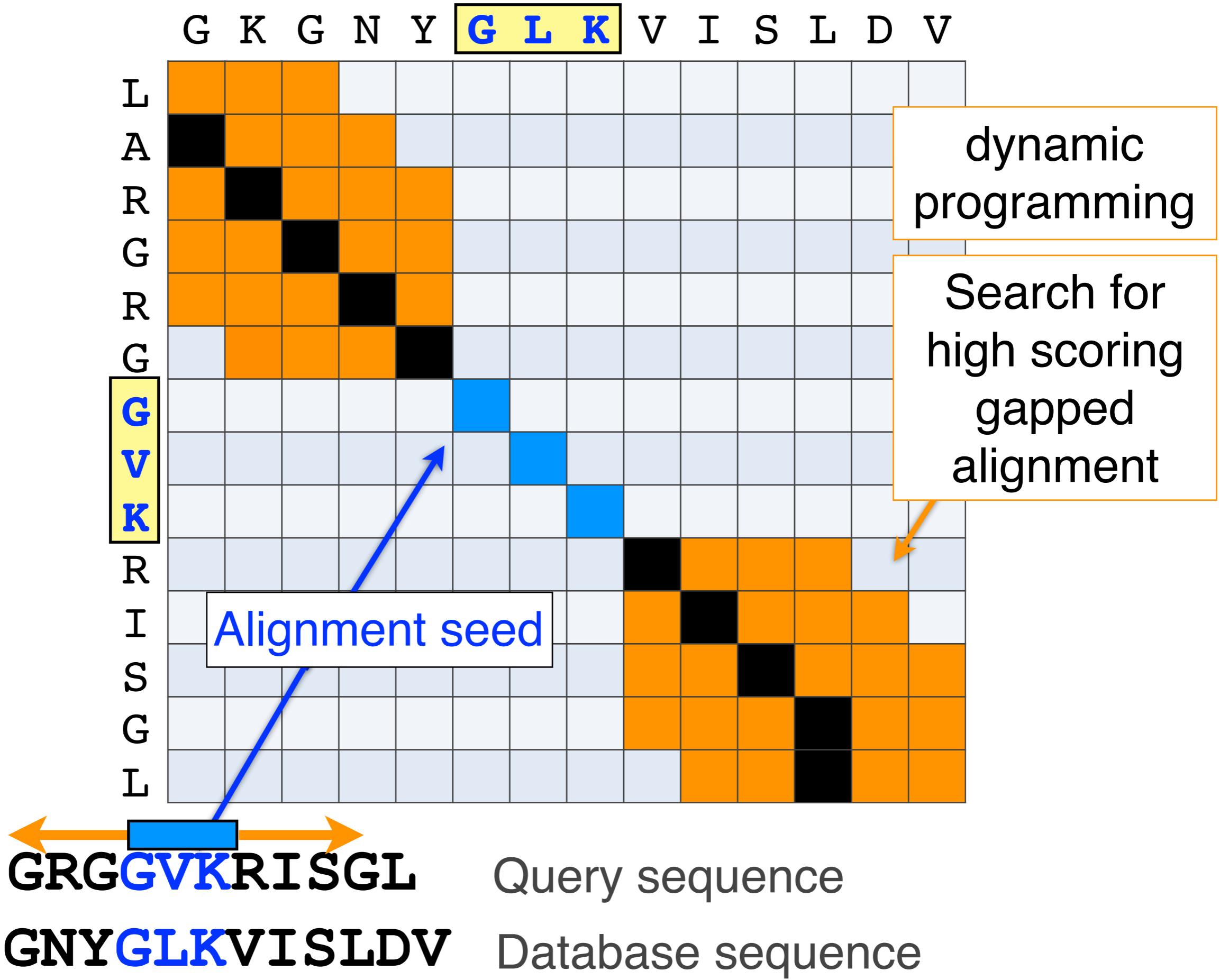


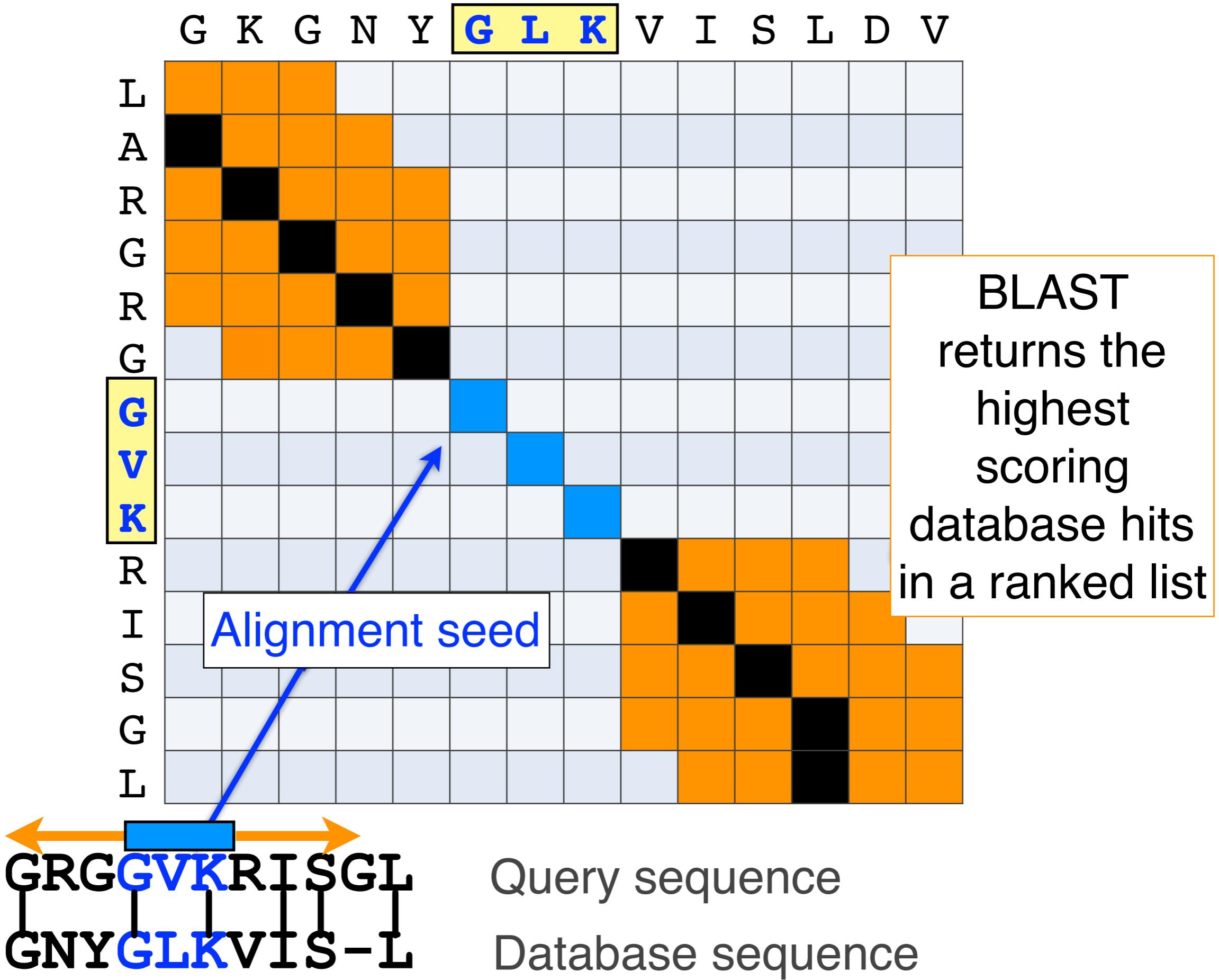
Blast

- Phase 4: the initial database hits are extended in both directions using dynamic programming









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

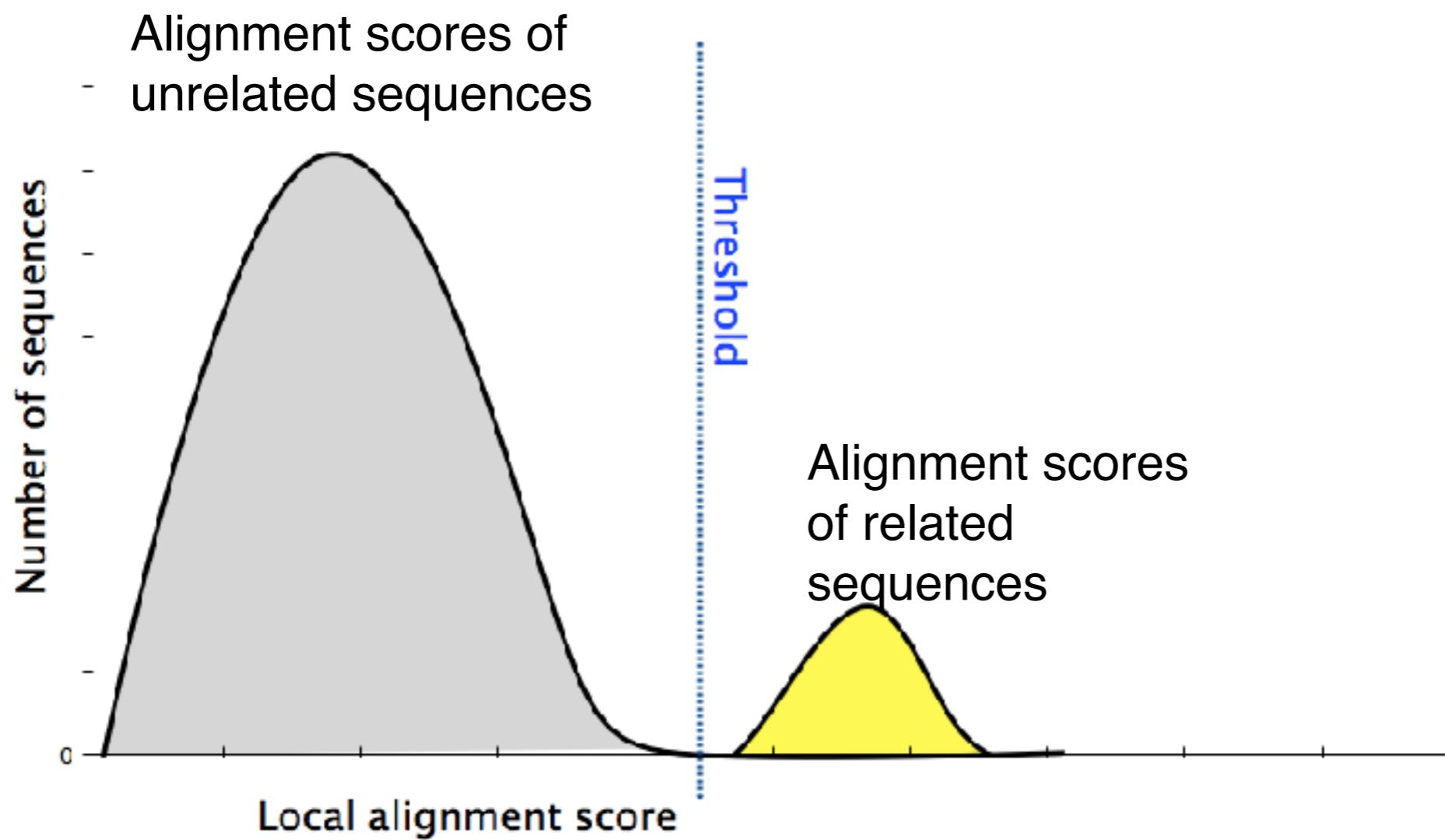
- An important feature of BLAST is the computation of statistical significance for each hit. This is described by the **E value** (expect value)

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

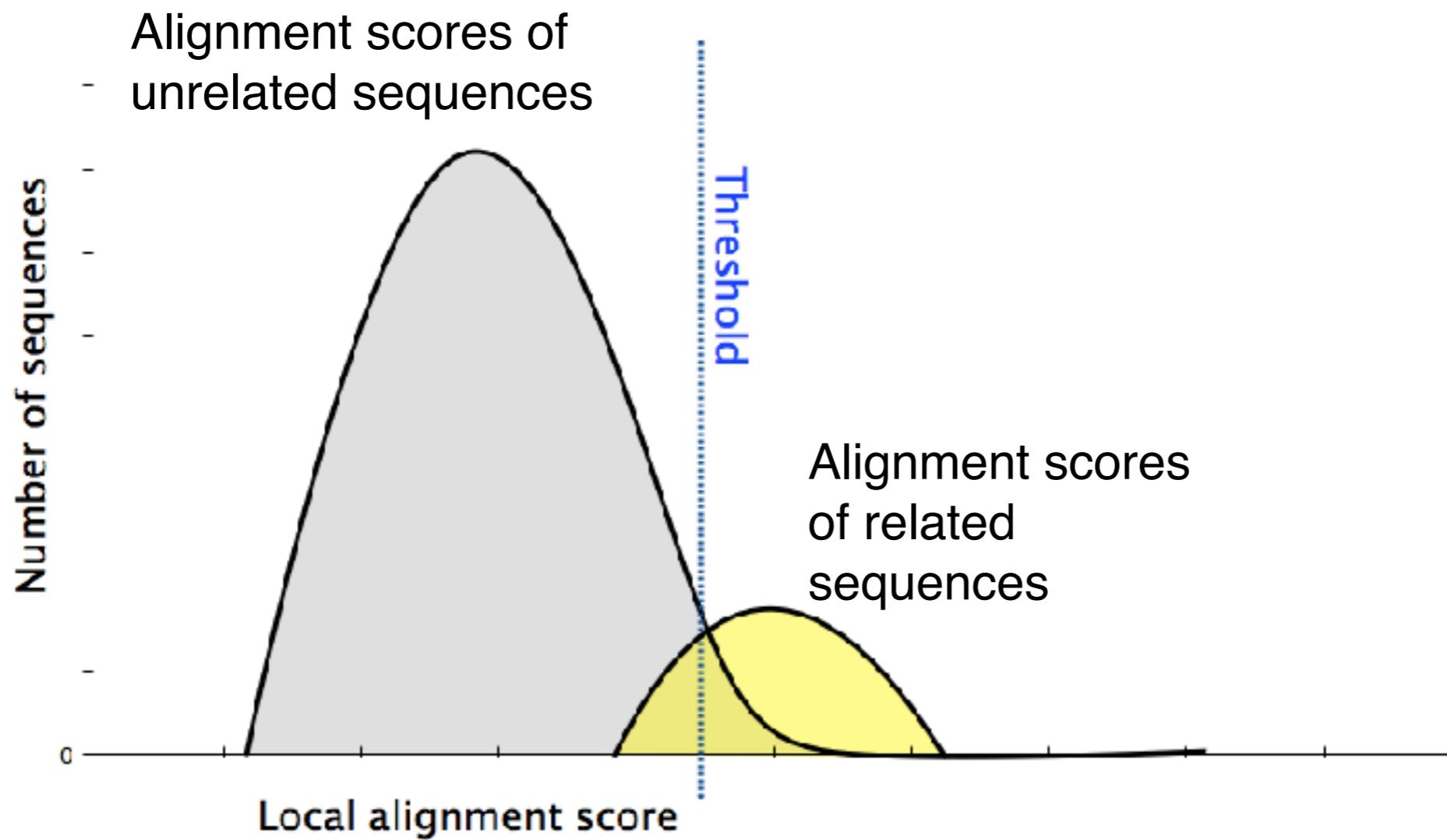
BLAST scores and E-values

- 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

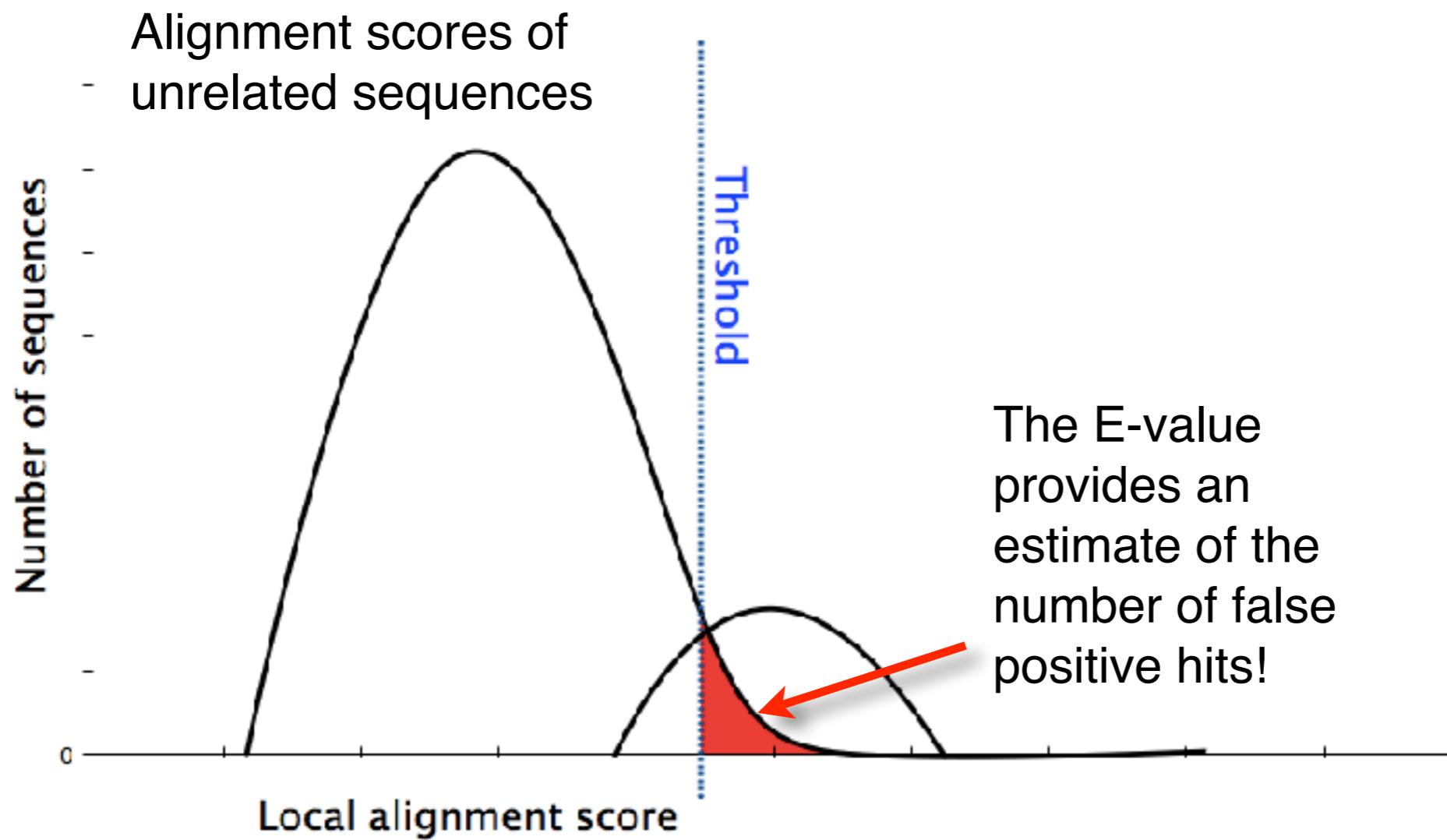
- Ideally, a threshold separates all query related sequences (yellow) from all unrelated sequences (gray)



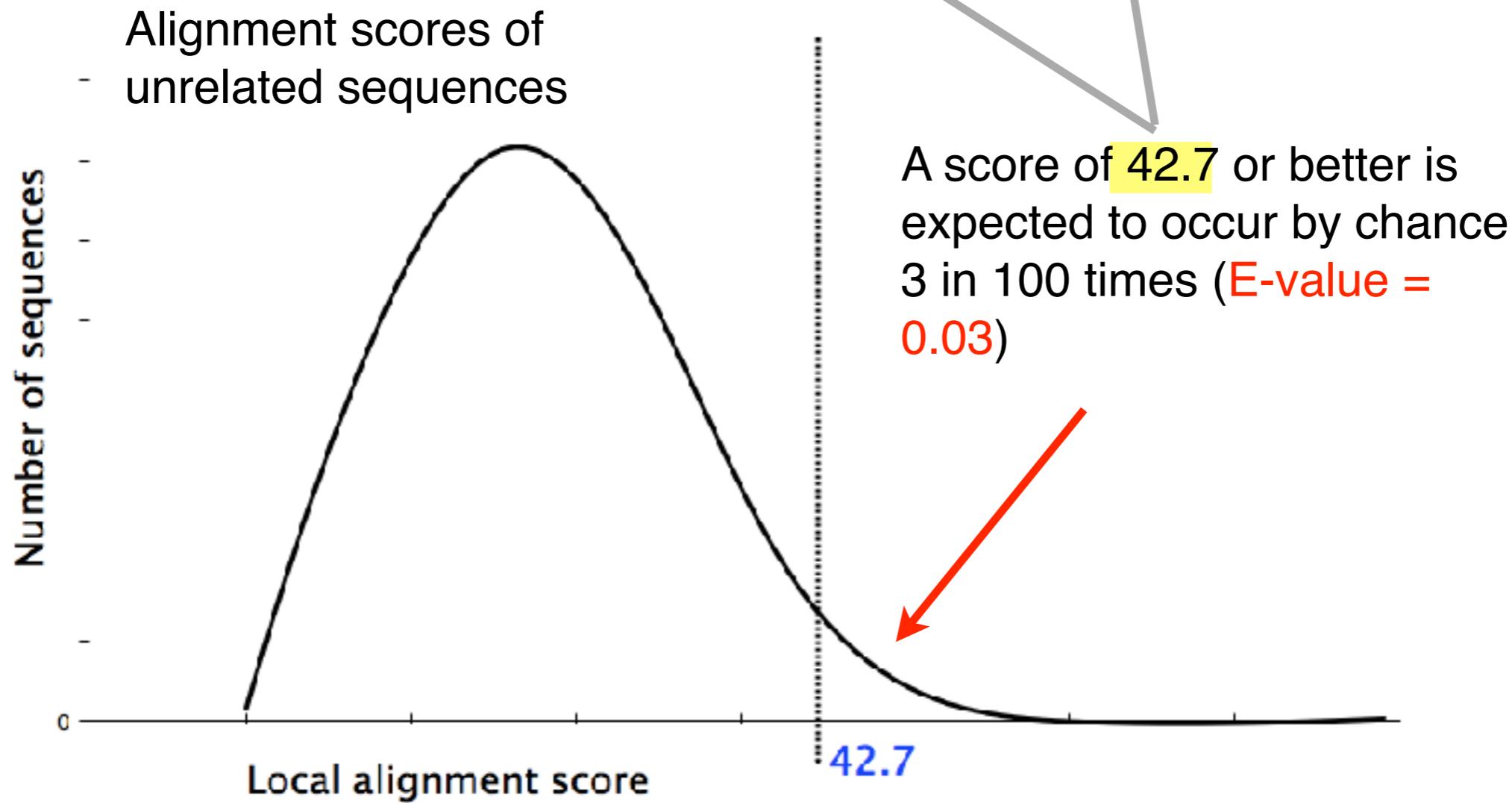
- Unfortunately, often both score distributions overlap
 - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



- 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

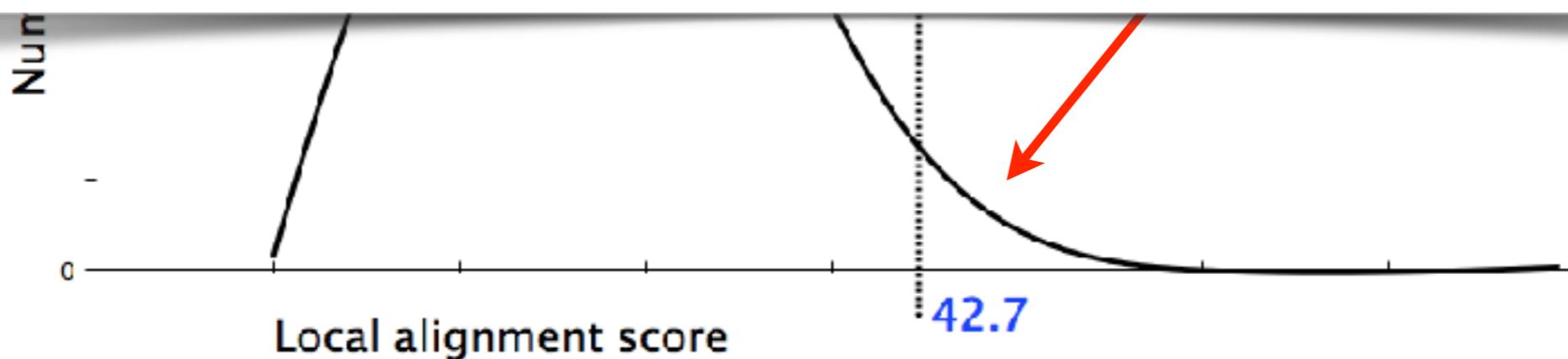


Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo sapiens]	677	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	676	100%	0	98%	AAA20133.1

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>

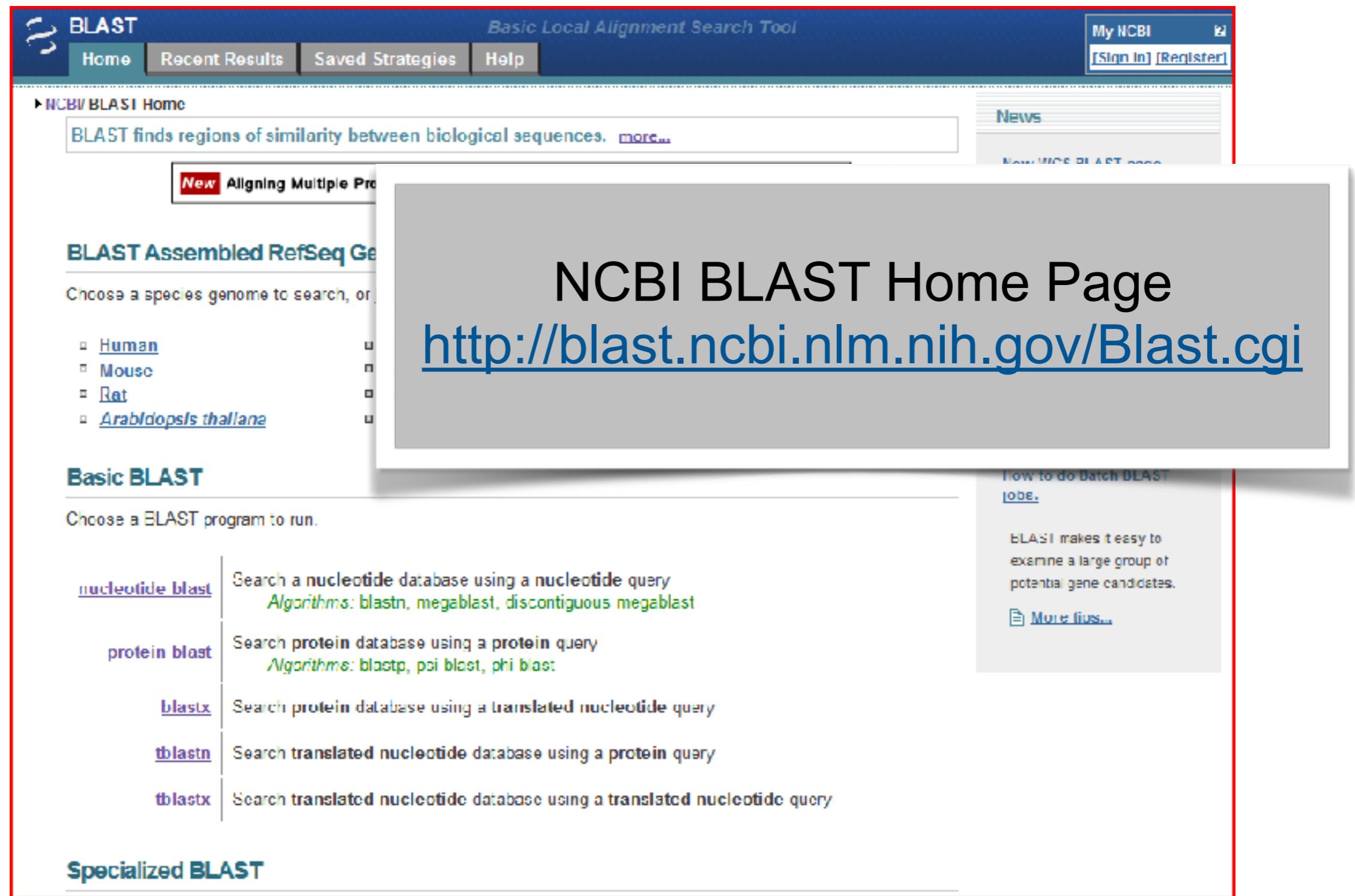


Your Turn!

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



The screenshot shows the NCBI BLAST Home Page. The top navigation bar includes links for Home, Recent Results, Saved Strategies, Help, My NCBI, Sign In, and Register. A banner at the top states "BLAST finds regions of similarity between biological sequences." Below this, a "New Aligning Multiple Pro" button is visible. On the left, there's a section titled "BLAST Assembled RefSeq Genomes" with a list of species: Human, Mouse, Rat, and *Arabidopsis thaliana*. The main content area features a large title "NCBI BLAST Home Page" and the URL "http://blast.ncbi.nlm.nih.gov/Blast.cgi". To the right, a sidebar provides information on "How to do Batch BLAST" and mentions that BLAST makes it easy to examine a large group of potential gene candidates. At the bottom, there are sections for "Basic BLAST" (nucleotide blast, protein blast, blastx, tblastn, tblastx) and "Specialized BLAST".

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

Basic BLAST

Choose a BLAST program to run.

- [nucleotide blast](#) Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontiguous megablast
- [protein blast](#) Search protein database using a protein query
Algorithms: blastp, psi blast, phi blast
- [blastx](#) Search protein database using a translated nucleotide query
- [tblastn](#) Search translated nucleotide database using a protein query
- [tblastx](#) Search translated nucleotide database using a translated nucleotide query

Specialized BLAST

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

NCBI Resources How To My N

Protein Translations of Life

Search: Protein Limits Advanced search Help

Display Settings FASTA Send to:

hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence [NP_000509.1](#)

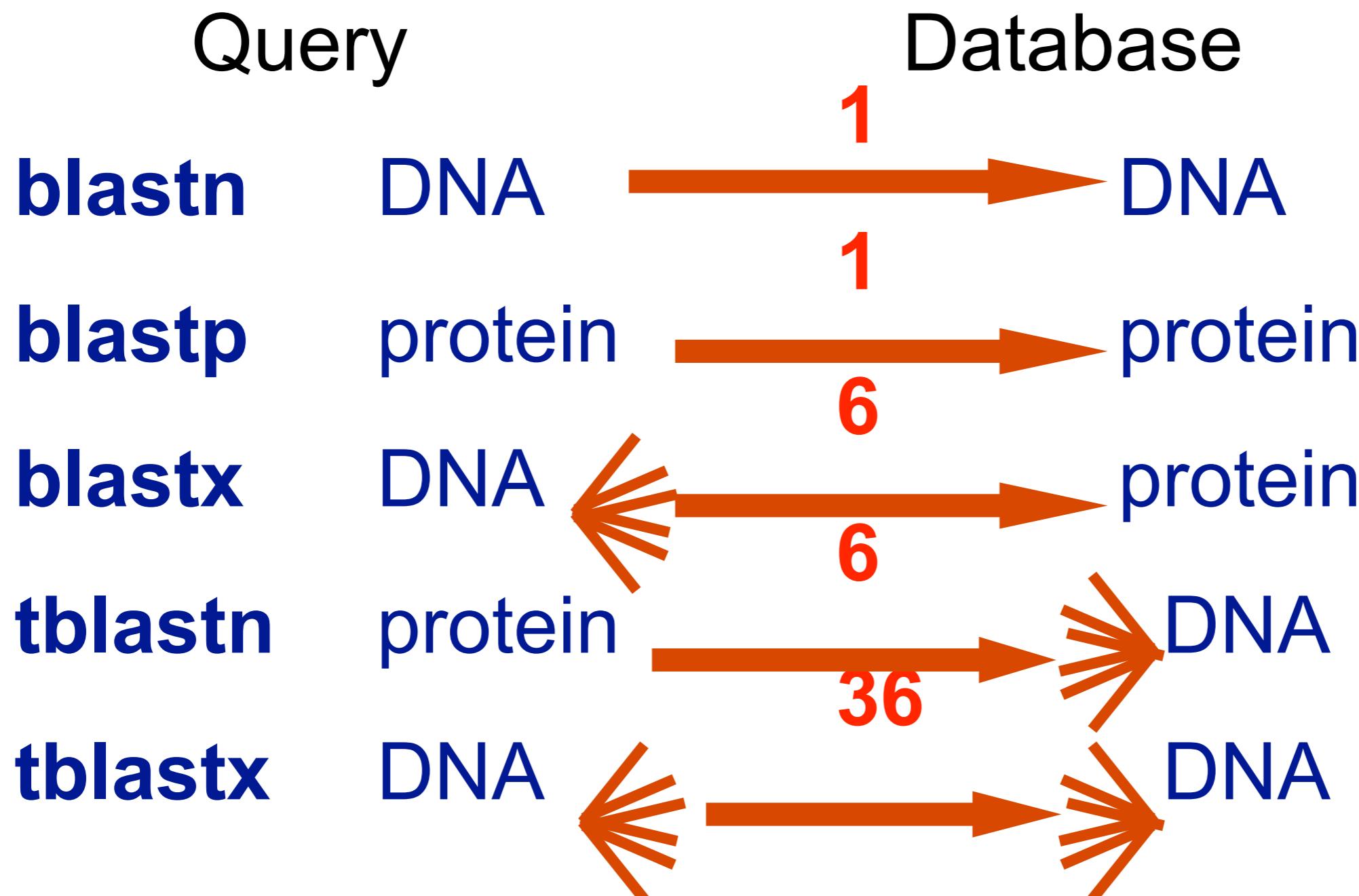
[GenPept](#) [Graphics](#) [Run BLAST](#)

>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHHFGKEFTPPVQAAQKVVAGVAN
ALAHKYH

Identify Conserved Domains

Find in this Sequence

Step 2: Choose the BLAST program

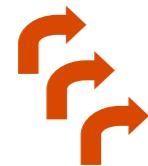


DNA potentially encodes six proteins

5' CAT CAA

5' ATC AAC

5' TCA ACT



5' CATCAACTACAACCTCAAAGACACCCCTTACACATCAACAAACCTACCCAC 3'

3' GTAGTTGATGTTGAGGTTCTGTGGGAATGTGTAGTTGTTGGATGGGTG 5'

5' GTG GGT

5' TGG GTA

5' GGG TAG



Protein BLAST: search protein databases using a protein query

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=Blast+Search

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGVNVDEVGGEALGRLLVYPWTQRFESFGDLSTPDAVMGNPKVKAHCK
KVLGAESDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLGNVLVCVLAHHGKEETPPVQAAYQK
WAGVANALAHKYH
```

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

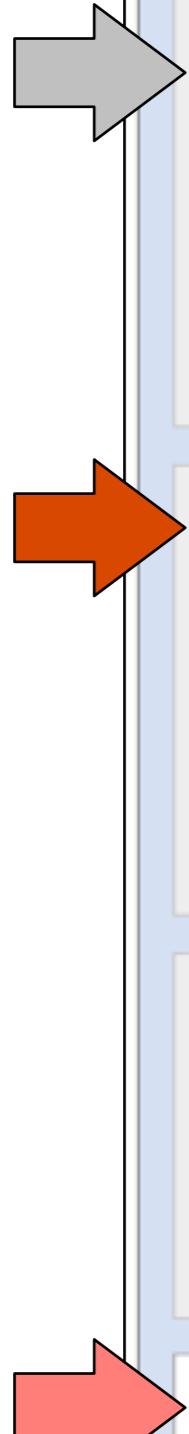
Entrez Query [Optional](#)
Enter an Entrez query to limit search [?](#)

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](#)



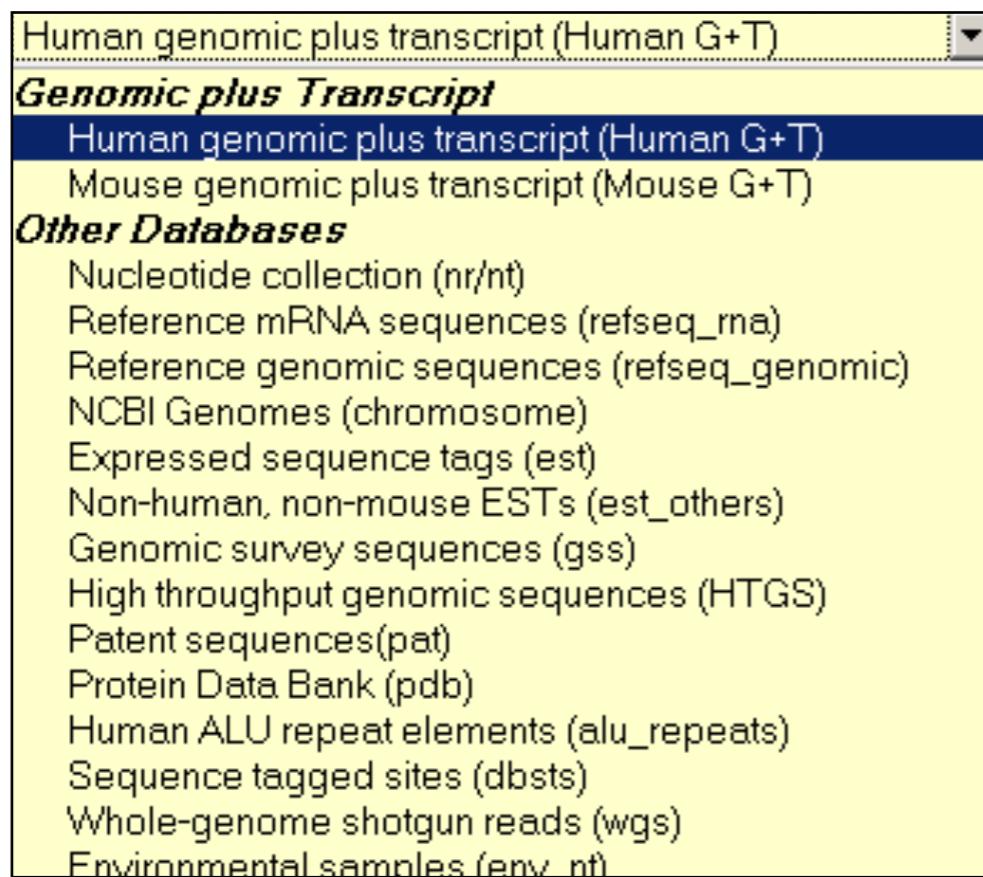
Step 3: Choose the database

nr = non-redundant (most general database)

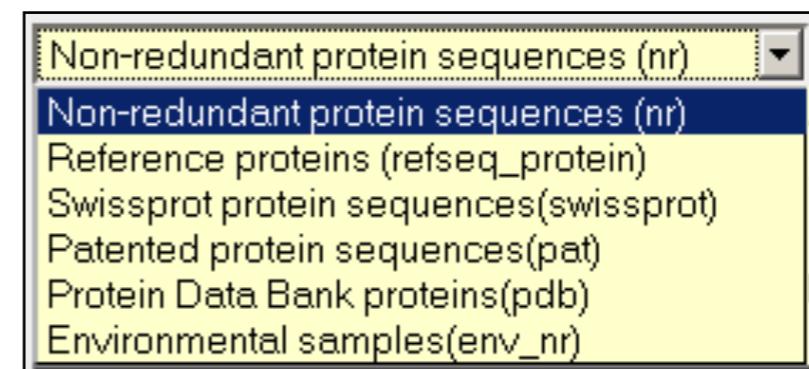
dbest = database of expressed sequence tags

dbsts = database of sequence tag sites

gss = genomic survey sequences



nucleotide databases



protein databases

Protein BLAST: search protein databases using a protein query

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=Blast+Search

Reader

Query subrange

From []

To []

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPKVKAHKG
KVLGAISDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGTVLVCVLAHHGKEITPPVQAAYQK
WAGVANALAHKYH

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 [Exclude](#) [+](#)
Optional

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude Models (XM/XP) Uncultured/environmental sample sequences
Optional

Entrez Query
Optional

Enter an Entrez query to limit search [?](#)

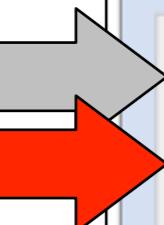
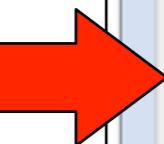
Program Selection

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](#)

Organism 
Entrez 
Settings! 

Step 4a: Select optional search parameters

Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only
 Mask lower case letters

BLAST: Search database Non-redundant protein sequences (nr) using Blastp
 Show results in a new window

Annotations highlight specific parameters:

- A blue double-headed arrow spans from the "Expect threshold" field to the "Word size" field.
- An orange double-headed arrow spans from the "Matrix" dropdown to the "Scoring matrix" label.

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 Reader

BLAST® Basic Local Alignment Search Tool My NCBI [Sign In] [Register]

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp suite/ Formatting Results - FVGUTMRZ013

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

You Tube Learn about the enhanced report Blast report description

gi|4504349|ref|NP_000509.1| hemoglobin

Query ID: Icl|84677 Database Name: nr
Description: gi|4504349|ref|NP_000509.1| hemoglobin subunit Description: All non-redundant GenBank CDS
beta [Homo sapiens] translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type: amino acid Program: BLASTP 2.2.27+ > Citation
Query Length: 147

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.

Query seq. Specific hits Superfamilies

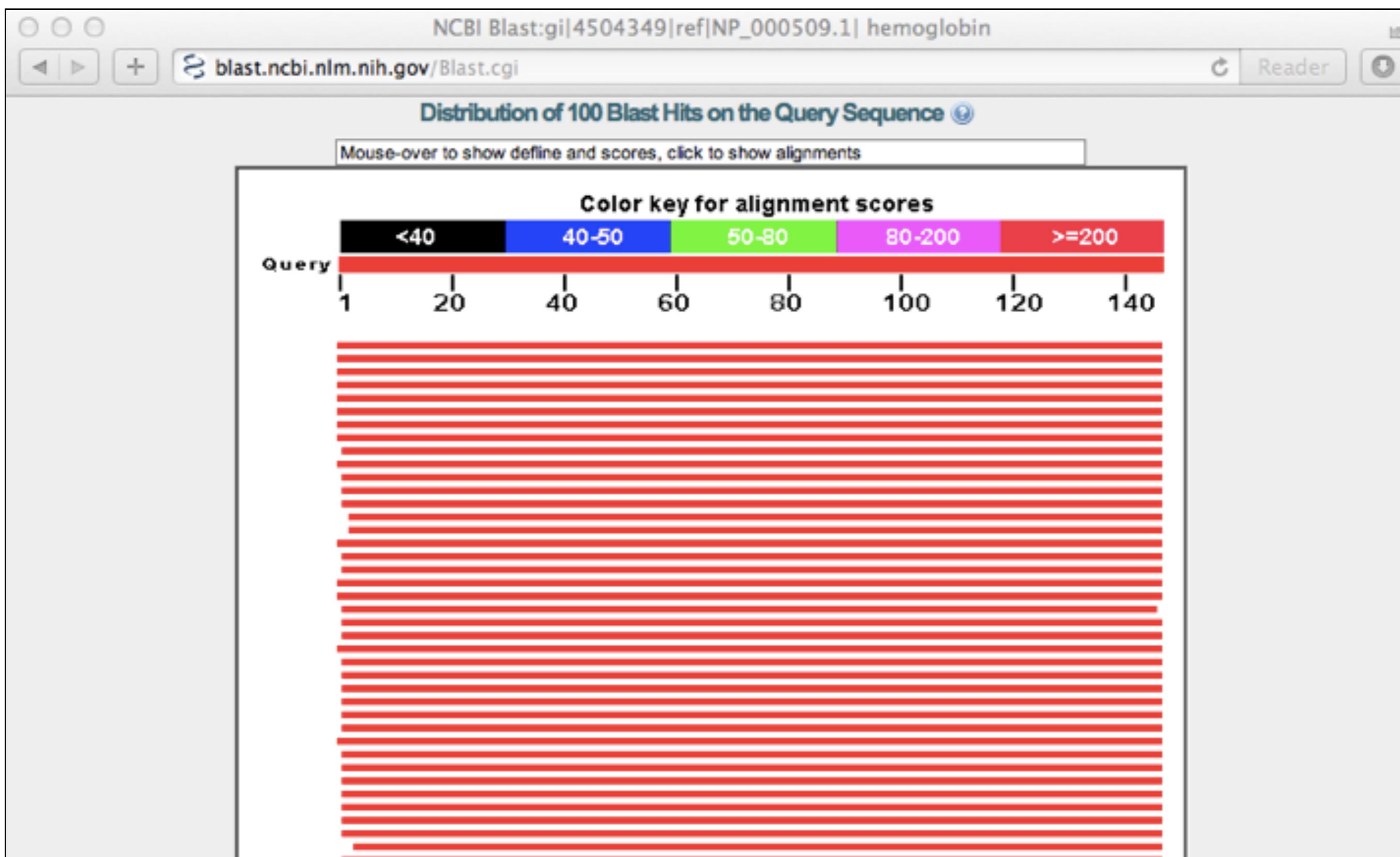
1 25 50 75 100 125 147

heme-binding site globin globin_like superfamily

Distribution of 100 Blast Hits on the Query Sequence ⓘ

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

Further down the results page...



Further down the results page...

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

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

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input type="checkbox"/>	hemoglobin beta [synthetic construct]	301	301	100%	9e-103	100%	AAX37051.1
<input type="checkbox"/>	hemoglobin beta [synthetic construct]	301	301	100%	1e-102	100%	AAX29557.1
<input type="checkbox"/>	hemoglobin subunit beta [Homo sapiens] >ref XP_508242.1 PREDICTED: hemoglobin s	301	301	100%	1e-102	100%	NP_000509.1
<input type="checkbox"/>	RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta	300	300	100%	4e-102	99%	P02024.2
<input type="checkbox"/>	beta globin chain variant [Homo sapiens]	299	299	100%	5e-102	99%	AAN84548.1
<input type="checkbox"/>	beta globin [Homo sapiens] >gb AAZ39781.1 beta globin [Homo sapiens] >gb AAZ39782.1 beta globin [Homo sapiens]	299	299	100%	5e-102	99%	AAZ39780.1
<input type="checkbox"/>	beta-globin [Homo sapiens]	299	299	100%	5e-102	99%	ACU56984.1
<input type="checkbox"/>	hemoglobin beta chain [Homo sapiens]	299	299	100%	6e-102	99%	AAD19696.1
<input type="checkbox"/>	Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound At The Beta Subunit	298	298	99%	9e-102	100%	1COH_B
<input type="checkbox"/>	hemoglobin beta subunit variant [Homo sapiens] >gb AAA88054.1 beta-globin [Homo sapiens]	298	298	100%	1e-101	99%	AAF00489.1
<input type="checkbox"/>	Chain B, Human Hemoglobin D Los Angeles: Crystal Structure >pdb 2YRS D Chain D, H	298	298	99%	2e-101	99%	2YRS_B
<input type="checkbox"/>	Chain B, High-Resolution X-Ray Study Of Deoxy Recombinant Human Hemoglobins Synthesized In Escherichia Coli	297	297	99%	3e-101	99%	1DXU_B
<input type="checkbox"/>	Chain B, Analysis Of The Crystal Structure, Molecular Modeling And Infrared Spectroscopic Characterization Of Human Hemoglobin D Los Angeles	297	297	99%	3e-101	99%	1HDB_B

Further down the results page...

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

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

Download GenPept Graphics ▾ Next ▲ Previous ▲ Descriptions

hemoglobin subunit beta [Homo sapiens]
Sequence ID: ref|NP_000509.1| Length: 147 Number of Matches: 1
► See 84 more title(s)

Range 1: 1 to 147 GenPept Graphics ▾ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
301 bits(770)	1e-102	Compositional matrix adjust.	147/147(100%)	147/147(100%)	0/147(0%)

Query 1 MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60
Sbjct 1 MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60

Query 61 VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 120
Sbjct 61 VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 120

Query 121 KEFTPQAAQKVVAAGVANALAHKYH 147
Sbjct 121 KEFTPQAAQKVVAAGVANALAHKYH 147

Related Information

- Gene - associated gene details
- UniGene - clustered expressed sequence tags
- Map Viewer - aligned genomic context
- Structure - 3D structure displays
- PubChem Bio
- Assay - bioactivity screening

Download GenPept Graphics ▾ Next ▲ Previous ▲ Descriptions

RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta chain
Sequence ID: sp|P02024.2|HBB_GORGO Length: 147 Number of Matches: 1

Range 1: 1 to 147 GenPept Graphics ▾ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
300 bits(767)	4e-102	Compositional matrix adjust.	146/147(99%)	147/147(100%)	0/147(0%)

Related Information

Different output formats are available

NCBI Blast:gi|4504349|ref|NP_000509.1| hemoglobin
blast.ncbi.nlm.nih.gov/Blast.cgi Reader

BLAST® Basic Local Alignment Search Tool My NCBI [Sign In] [Register]

Home Recent Results Saved Strategies Help

NCBI BLAST/blastp suite/ Formatting Results - FVGUTMRZ013

Edit and Resubmit Save Search Strategies ▾ Formatting options Download Change the result display back YouTube Learn about the enhanced report Blast

Formatting options

Show Alignment as HTML Old View Reset form to defaults

Alignment View Query-anchored with letters for identities

Display Graphical Overview Sequence Retrieval NCBI-gi

Masking Character: Lower Case Color: Grey

Limit results Descriptions: 50 Graphical overview: 50 Alignments: 50

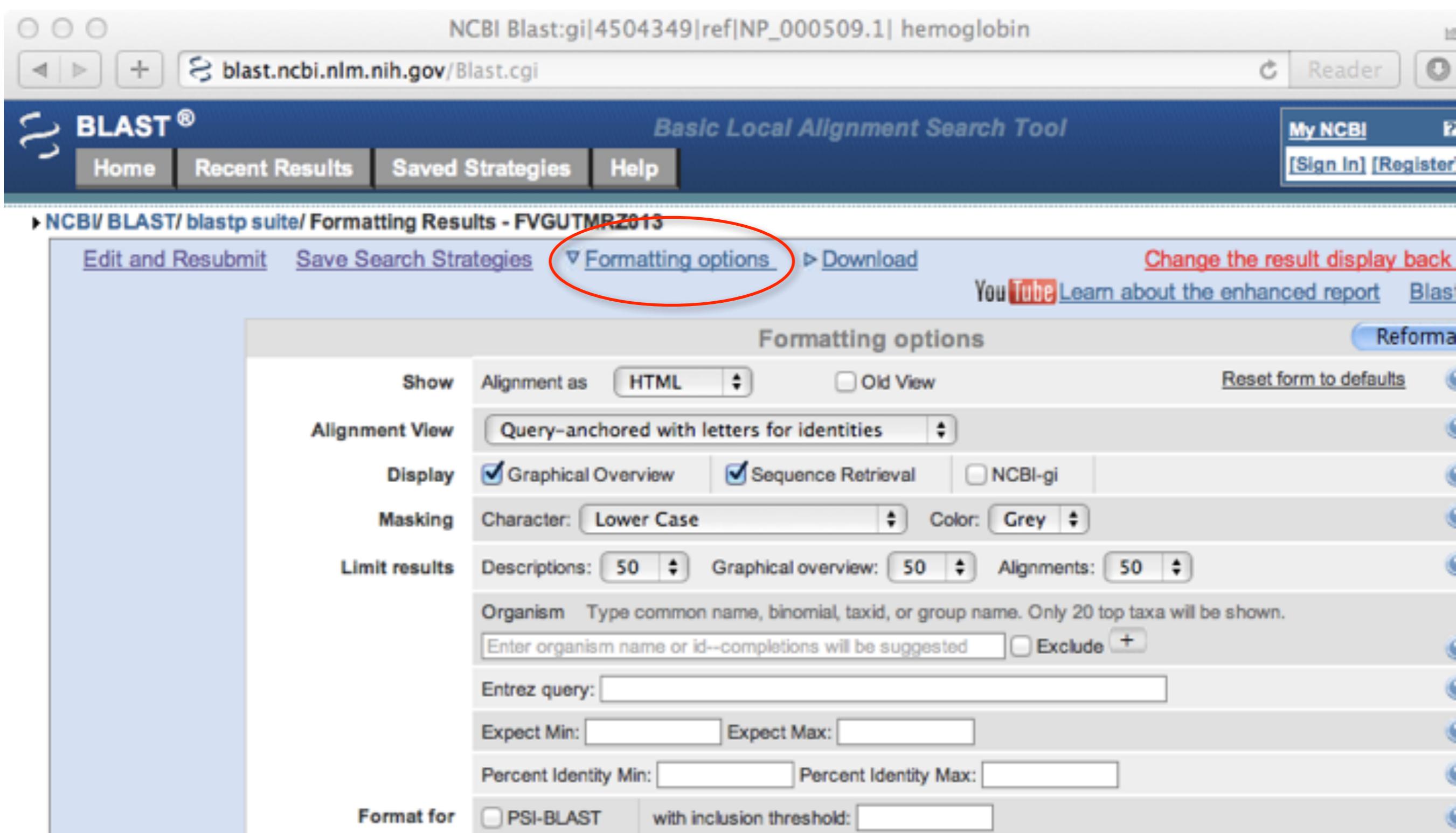
Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.
Enter organism name or id--completions will be suggested Exclude +

Entrez query:

Expect Min: Expect Max:

Percent Identity Min: Percent Identity Max:

Format for PSI-BLAST with inclusion threshold:



gi|4504349|ref|NP_000509.1| hemoglobin

E.g. Query anchored alignments

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

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

Query	Length	Sequence	Score
AAX37051	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAX29557	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
NP_000509	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
P02024	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAN84548	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAZ39780	1	MVHLTPKEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
ACU56984	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFKSFSDLSTPDAVMGNPK	60
AAD19696	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFLESFGDLSTPDAVMGNPK	60
1COH_B	1	VHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
AAF00489	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
2YRS_B	1	VHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1DXU_B	1	MHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1HDB_B	1	VHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1DXV_B	2	HLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
3KMF_C	2	HLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
AAL68978	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
1NQP_B	1	VHLTPEEKSAVTALWGKVNDEVGGKALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1K1K_B	1	VHLTPKEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
AAN11320	1	MVHLTPVEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
XP_002822173	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
1Y85_B	1	VHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1YE0_B	1	MHLTPEEKSAVTALWGKVNDEVGGEALGRLLAVYPWTQRFFESFGDLSTPDAVMGNPK	59
1O1O_B	1	MHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
CAA23759	1	MVHLTPVEEKSAVTAXWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
1YE2_B	1	MHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVFPWTQRFFESFGDLSTPDAVMGNPK	59
1Y5F_B	1	MHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1A00_B	1	MHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPYTQRFFESFGDLSTPDAVMGNPK	59
1HBS_B	1	VHLTPVEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1ABY_B	1	MHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1CMY_B	1	VHLTPKEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59

... and alignments with dots for identities

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

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

Query	Length	Sequence	Length
AAX37051	1	MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAX29557	1	60
NP_000509	1	60
P02024	1	60
AAN84548	1	60
AAZ39780	1K.....	60
ACU56984	1	60
AAD19696	1	60
1COH_B	1	59
AAF00489	1	60
2YRS_B	1	59
1DXU_B	1	M.....	59
1HDB_B	1	59
1DXV_B	2	59
3KMF_C	2	59
AAL68978	1	60
1NQP_B	1K.....	59
1K1K_B	1K.....	59
AAN11320	1V.....	60
XP_002822173	1	60
1Y85_B	1	59
1YE0_B	1	M.....A.....	59
1O1Q_B	1	M.....	59
CAA23759	1V.....X.....	60
1YE2_B	1	M.....F.....	59
1Y5F_B	1	M.....	59
1A00_B	1	M.....Y.....	59

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**.

To Update!

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
Correct optimal score position in matrix used	1
Correct optimal score obtained for given scoring scheme	1
Traceback path(s) clearly highlighted	1
Correct alignment(s) yielding optimal score listed	1
	A+