



BIMM 143

Introduction to Bioinformatics

Lecture 2

Barry Grant

UC San Diego

<http://thegrantlab.org/bimm143>

Recap From Last Time:

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

Today's Menu

Classifying Databases

Primary, secondary and composite Bioinformatics databases

Using Databases

Vignette demonstrating how major Bioinformatics databases intersect

Major Biomolecular Formats

How nucleotide and protein sequence and structure data are represented

Alignment Foundations

Introducing the *why* and *how* of comparing sequences

Alignment Algorithms

Hands-on exploration of alignment algorithms and applications

Primary, secondary & composite databases

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

- **Primary databases** (or archival databases) consist of data derived experimentally.
 - ▶ **GenBank**: NCBI's primary nucleotide sequence database.
 - ▶ **PDB**: Protein X-ray crystal and NMR structures.
- **Secondary databases** (or derived databases) contain information derived from a primary database.
 - **RefSeq**: non redundant set of curated reference sequences primarily from GenBank
 - **PFAM**: protein sequence families primarily from UniProt and PDB
- **Composite databases** (or metadatabases) join a variety of different primary and secondary database sources.
 - **OMIM**: catalog of human genes, genetic disorders and related literature
 - **GENE**: molecular data and literature related to genes with extensive links to other databases.

DATABASE VIGNETTE

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

“What do you know about that ‘Kras’ gene the speaker kept taking about?”

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

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

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

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

The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. The browser's address bar displays www.ncbi.nlm.nih.gov/. The NCBI logo and navigation menu are visible at the top. A search bar is present, with the text "All Databases" on the left and "ras" entered in the search field, which is highlighted with a red square. A "Search" button is located to the right of the search bar. A large, diagonal watermark in red text reads "Hands on demo (or see following slides)".

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 and the National Library of Medicine provide access to biomedical and health information and health by providing access to biomedical and health information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#)

Get Started

- [Data](#): Get NCBI data or software
- [How To](#): 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.

NCBI Announcements

RefSeq release 69 available on...

The full RefSeq release 69 is not available on the FTP site with 74 records describing 52,276,469...

Example Vignette Questions:

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

Search NCBI databases

[Help](#)

ras

Search

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

Literature

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

Health

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

Genes

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

Proteins

Gene
[Save search](#) [Advanced](#) [Help](#)

[Show additional filters](#)

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

Send to:

[Hide sidebar >>](#)

Filters: [Manage Filters](#)

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

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.

- [Clear all](#)
- Gene sources**
- Genomic
- Mitochondria
- Organelles
- Plasmids
- Plastids
- Categories**
- Alternatively spliced
- Annotated genes
- Non-coding
- Protein-coding
- Pseudogene
- Sequence content**
- CCDS
- Ensembl
- RefSeq

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

Find related data

Database:

Search details

ras[All Fields] AND alive[property]

Gene

Gene

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

Search

Help

Show additional filters

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

Send to:

Hide sidebar >>

Filters: Manage Filters

Clear all

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

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

Find related data

Database:

Select

Find items

Search details

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

Search

See more...

Recent activity

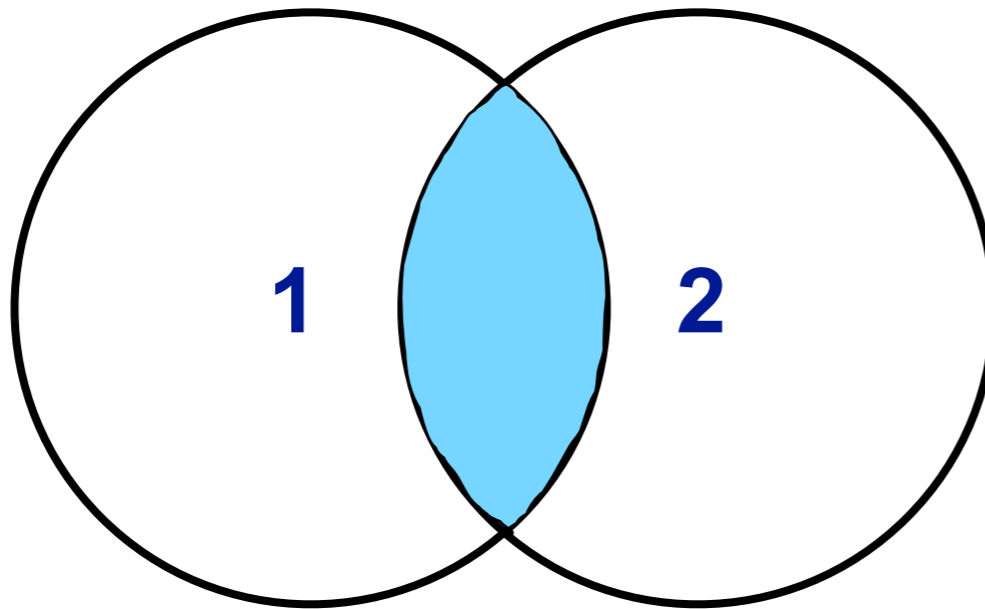
Turn Off Clear

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, KI-RAS1, KRAS2, NS,

Status clear
Current only

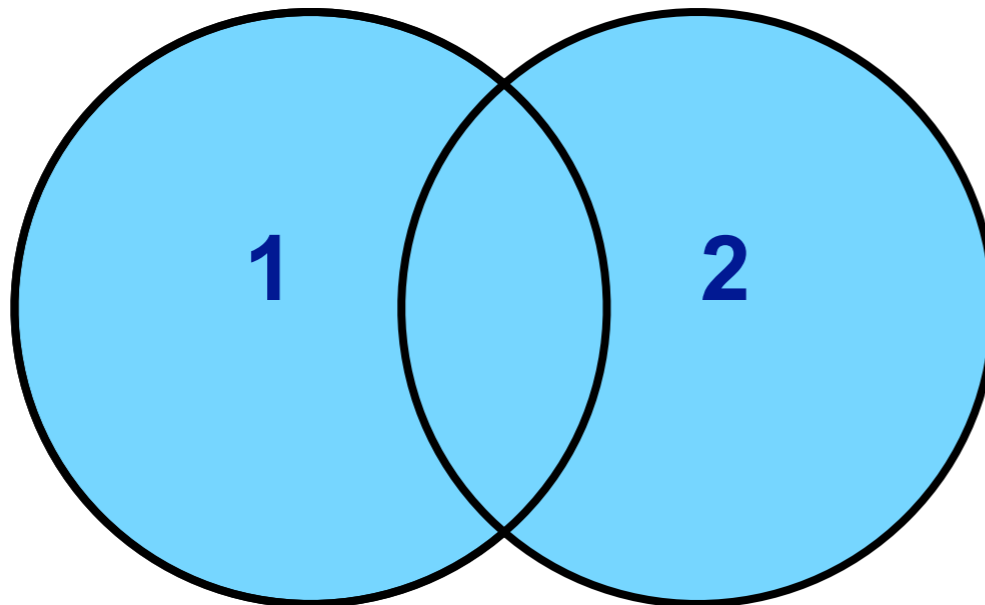
Chromosome locations
Select

1 AND 2



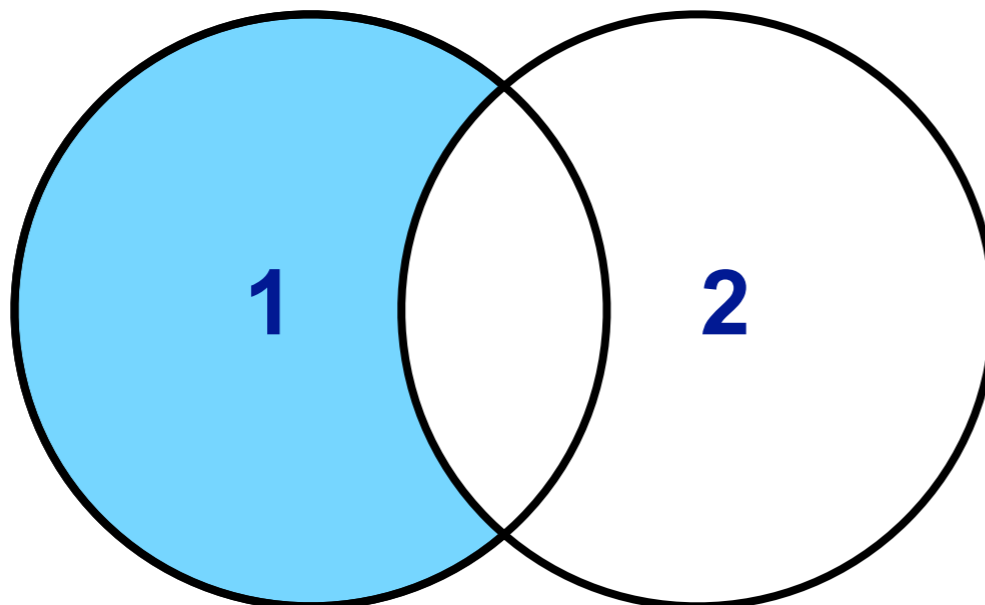
**ras AND disease
(1185 results)**

1 OR 2



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

1 NOT 2



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

Show additional filters

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

Hide sidebar >>

Filters: Manage Filters

Clear all

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

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

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

- Gene sources
- Genomic
- Categories
- Alternatively spliced
- Annotated genes
- Non-coding
- Protein-coding
- Pseudogene
- Sequence content
- CCDS
- Ensembl
- RefSeq
- Status clear
- Current only
- Chromosome locations

Name/Gene ID	Description	Location	Aliases
<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, KI-RAS1, KRAS2, NS, NS3, RASK2

Gene
[Advanced](#) [Help](#)

[Display Settings:](#) Full Report [Send to:](#)

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)



Gene

Search

Help

Display S

Hide sidebar >>

KRAS
(human)

Example Questions:
What chromosome location and what genes are in the vicinity?

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)

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-

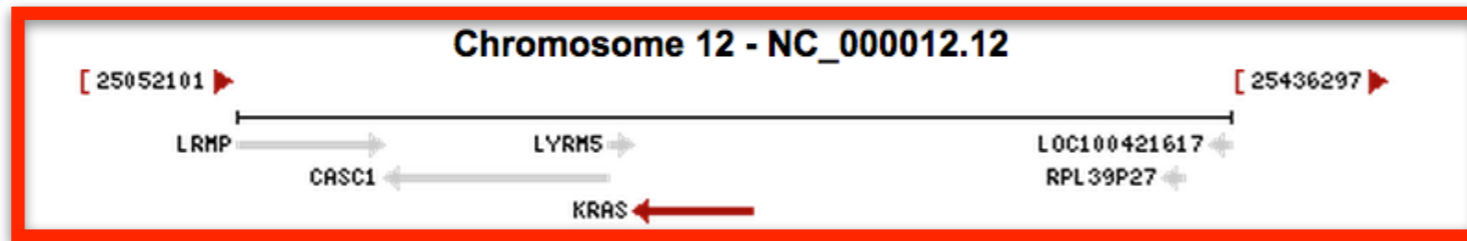
Genomic context

Location: 12p12.1

Exon count: 6

See KRAS in [Epigenomics](#), [MapViewer](#)

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



Genomic regions, transcripts, and products

Go to [reference sequence details](#)

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

Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)

- [BioAssay by Target \(List\)](#)
- [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](#)

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?

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 sequences

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-

Gene Ontology [Provided by GOA](#)

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

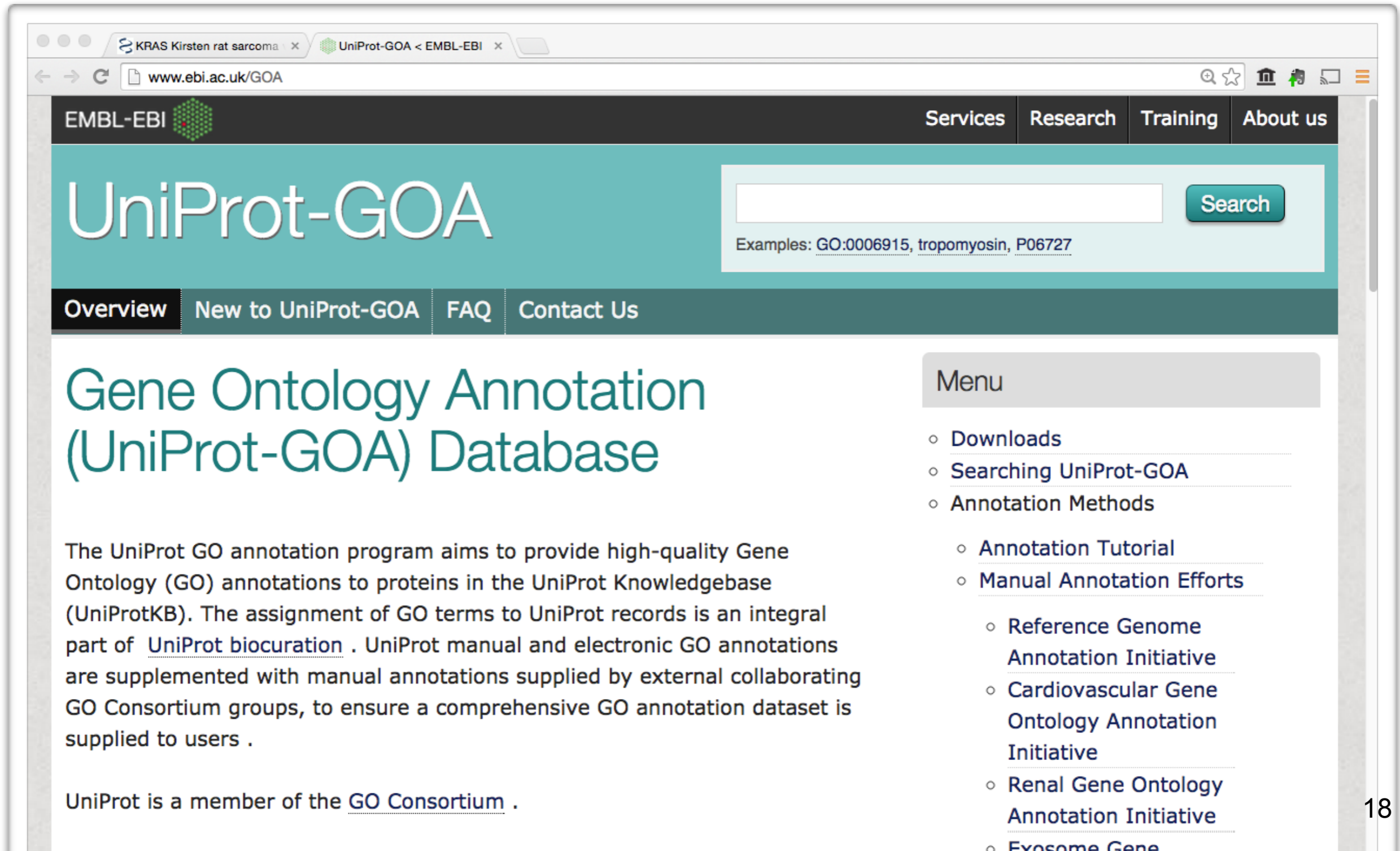
Items 1 - 25 of 33 < Prev Page 1 of 2 Next >

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



GO: Gene Ontology

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



The screenshot shows the UniProt-GOA website interface. At the top, there are browser tabs for 'KRAS Kirsten rat sarcoma' and 'UniProt-GOA < EMBL-EBI'. The address bar shows 'www.ebi.ac.uk/GOA'. The main header features the EMBL-EBI logo and navigation links for 'Services', 'Research', 'Training', and 'About us'. Below this is a large teal banner with the 'UniProt-GOA' logo and a search bar. The search bar contains a text input field and a 'Search' button. Below the search bar, there are example search terms: 'Examples: [GO:0006915](#), [tropomyosin](#), [P06727](#)'. A secondary navigation bar includes links for 'Overview', 'New to UniProt-GOA', 'FAQ', and 'Contact Us'. The main content area features the title 'Gene Ontology Annotation (UniProt-GOA) Database' in a large teal font. Below the title is a paragraph of text describing the UniProt GO annotation program. To the right of the main text is a 'Menu' section with a list of links: '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', and 'Exosome Gene'.

EMBL-EBI [Services](#) [Research](#) [Training](#) [About us](#)

UniProt-GOA

[Search](#)

Examples: [GO:0006915](#), [tropomyosin](#), [P06727](#)

[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

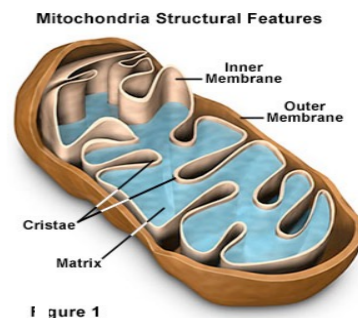
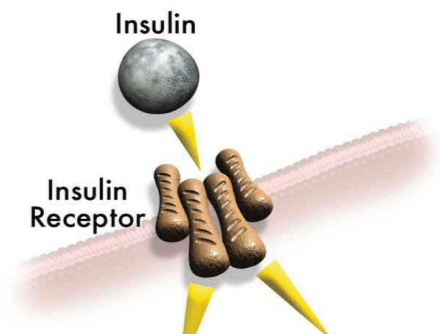
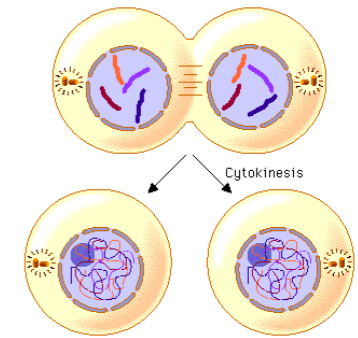
- [Downloads](#)
- [Searching UniProt-GOA](#)
- [Annotation Methods](#)
- [Annotation Tutorial](#)
- [Manual Annotation Efforts](#)
- [Reference Genome Annotation Initiative](#)
- [Cardiovascular Gene Ontology Annotation Initiative](#)
- [Renal Gene Ontology Annotation Initiative](#)
- [Exosome Gene](#)

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 Kirsten rat sarcoma

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

Gene Ontology [Provided by GOA](#)

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

Process

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

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

⋮ Scroll down to
▼ **UniProt** link

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

The screenshot shows the NCBI Gene page for KRAS (3845). The browser address bar shows the URL: www.ncbi.nlm.nih.gov/gene/3845#gene-ontology. The page displays genomic coordinates (X01669.1 to CAA25828.1) and a list of protein accessions. The UniProtKB link for P01116.1 is highlighted with a red box. A red arrow points to the UniProt link with the text: "Scroll down to Very bottom for UniProt link".

Protein Accession	Links
P01116.1	GenPept 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

UniProtKB Advanced

BLAST Align Retrieve/ID Mapping Help Contact

P01116 - RASK_HUMAN

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

Display

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

Enzyme regulationⁱ
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.

Regions

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

UniProt will detail much more information for protein coding genes

P01116 - RASK_HUMAN

Protein | **GTPase KRas**
Gene | **KRAS**
Organism | *Homo sapiens (Human)*
Status | Reviewed - ●●●●● - Experimental evidence at protein levelⁱ

Display: None

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

Functionⁱ
Ras proteins bind GDP/GTP and promote cell proliferation (PubMed:23698361, PMID:10550000).

Enzyme regulationⁱ
Alternates between an inactive form and an active form. The active form is a GTP-bound protein 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			
Nucleotide binding ⁱ	29 – 35	7	GTP 2 Publications			
Nucleotide binding ⁱ	59 – 60	2	GTP 2 Publications			

```
>sp|P01116|RASK_HUMAN GTPase KRas OS=Homo sapiens GN=KRAS PE=1 SV=1
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDITAG
QEEYSAMRDQYMRGTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSQVPMVLVGNKCDL
PSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYTLVREIRQYRLKKISKEEKTGPGC
VKIKKCIIM
```


UniProt will detail much more information for protein coding genes

P01116 - RASK_HUMAN

Protein | **GTPase KRas**
Gene | **KRAS**
Organism | *Homo sapiens (Human)*
Status | Reviewed - ●●●●● - Experimental evidence at protein levelⁱ

Display None

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

Functionⁱ

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 regulationⁱ

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			

P01116 - RASK_HUMAN

Protein | **GTPase KRas**
 Gene | **KRAS**
 Organism | *Homo sapiens (Human)*
 Status | Reviewed -

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](#)

Functionⁱ

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

Enzyme regulationⁱ

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.

Regions

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

Example Questions:
 What positions in the protein are responsible for GTP binding?



Example Questions:

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

Display None **Pathology & Biotech¹**

- 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

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. [1 Publication](#)

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 ⁱ	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. 1 Publication		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?

Display **None** **Structure¹**

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

Secondary structure
1
Legend: Helix Turn Beta strand

3D structure databases

Select the link destinations:
 PDBeⁱ
 RCSB PDBⁱ
 PDBj¹

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	1-164	[>]
4EPX	X-ray	1.76	A	1-164	[>]
4EPY	X-ray	1.80	A	1-164	[>]
4L8G	X-ray	1.52	A	1-164	[>]
4LDJ	X-ray	1.15	A	1-164	[>]
4LPK	X-ray	1.50	A/B	1-169	[>]

▲ Top

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 website interface. At the top, there is a navigation bar with 'RCSB PDB', 'Deposit', 'Search', 'Visualize', and 'Analyze'. Below this is the PDB logo and a search bar. The main content area features a horizontal menu with tabs: 'Structure Summary', '3D View' (highlighted with a red box), 'Annotations', 'Sequence', 'Sequence Similarity', and 'Structure Alignment'. Below the menu, there are buttons for 'Display Files' (highlighted with a green box) and 'Download Files'. A green callout box with the text 'View PDB file format' points to the 'Display Files' button. The protein entry '4EPV' is displayed, including its title 'Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation', DOI, classification as 'HYDROLASE', deposition and release dates, authors, organism, and expression system. A 3D ribbon diagram of the protein structure is shown on the left. At the bottom, there are sections for 'Experimental Data Snapshot' and 'wwPDB Validation'.

Lets view the 3D structure:

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

www.rcsb.org

Home Gmail

RCSB PDB Deposit Search Visualize Analyze

4EPV

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

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Click to identify atoms and bonds.

Bond: [GLY]12:A.O - [GLY]12:A.C

Display Files Download Files

Assembly Bioassembly 1

Model Model 1

Symmetry None

Interaction [GDP]201:A

Style Cartoon

Color Rainbow

Ligand None

Quality Automatic

Water Ions

Hydrogens Clashes

Viewer Options

Contact Us

Back to UniProt:

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

www.uniprot.org/uniprot/P01116

Display **None**

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

▲ Top

OrthoDBⁱ E0V
PhylomeDBⁱ P01
TreeFamⁱ TF3

Family and domain databases

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


PFAM is one of the best protein family databases

Sequences (2)ⁱ

Sequence 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... ?

EMBL-EBI  HOME

Family: Ras (PF00071)

332 architectures 21243 sequences 30 interactions 1006 species 663 structures

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

Jump to...

Summary: Ras family

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

[Wikipedia: Ras subfamily](#) [Wikipedia: Ras superfamily](#) [Pfam](#) [InterPro](#)

This is the Wikipedia entry entitled "[Ras subfamily](#)". [More...](#)

Ras subfamily [Edit Wikipedia article](#)

This article is about p21/Ras protein. For the p21/waf1 protein, see [p21](#).

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

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

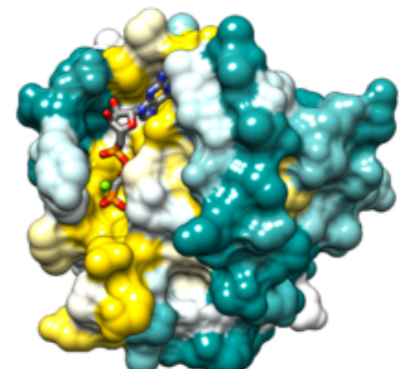
When Ras is 'switched on' by incoming signals, it subsequently switches on other proteins, which ultimately turn on genes involved in **cell growth**, **differentiation** and **survival**. As a result, mutations in ras genes can lead to the production of permanently activated Ras proteins. This can cause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

Because these signals result in cell growth and division, overactive Ras signaling can ultimately lead to **cancer**.^[1] The 3 Ras genes in humans (**HRAS**, **KRAS**, and **NRAS**) are the most common **oncogenes** in human **cancer**; mutations that permanently activate Ras are found in 20% to 25% of all human tumors and up to 90% in certain types of cancer (e.g., **pancreatic cancer**).

^[2] For this reason, Ras inhibitors are being studied as a treatment for cancer, and other diseases with Ras overexpression.

Contents [\[hide\]](#)

- History
- Structure
- Function
 - 3.1 Activation and deactivation
 - 3.2 Membrane attachment
- Members
- Ras in cancer
 - 5.1 Inappropriate activation
 - 5.2 Constitutively active Ras

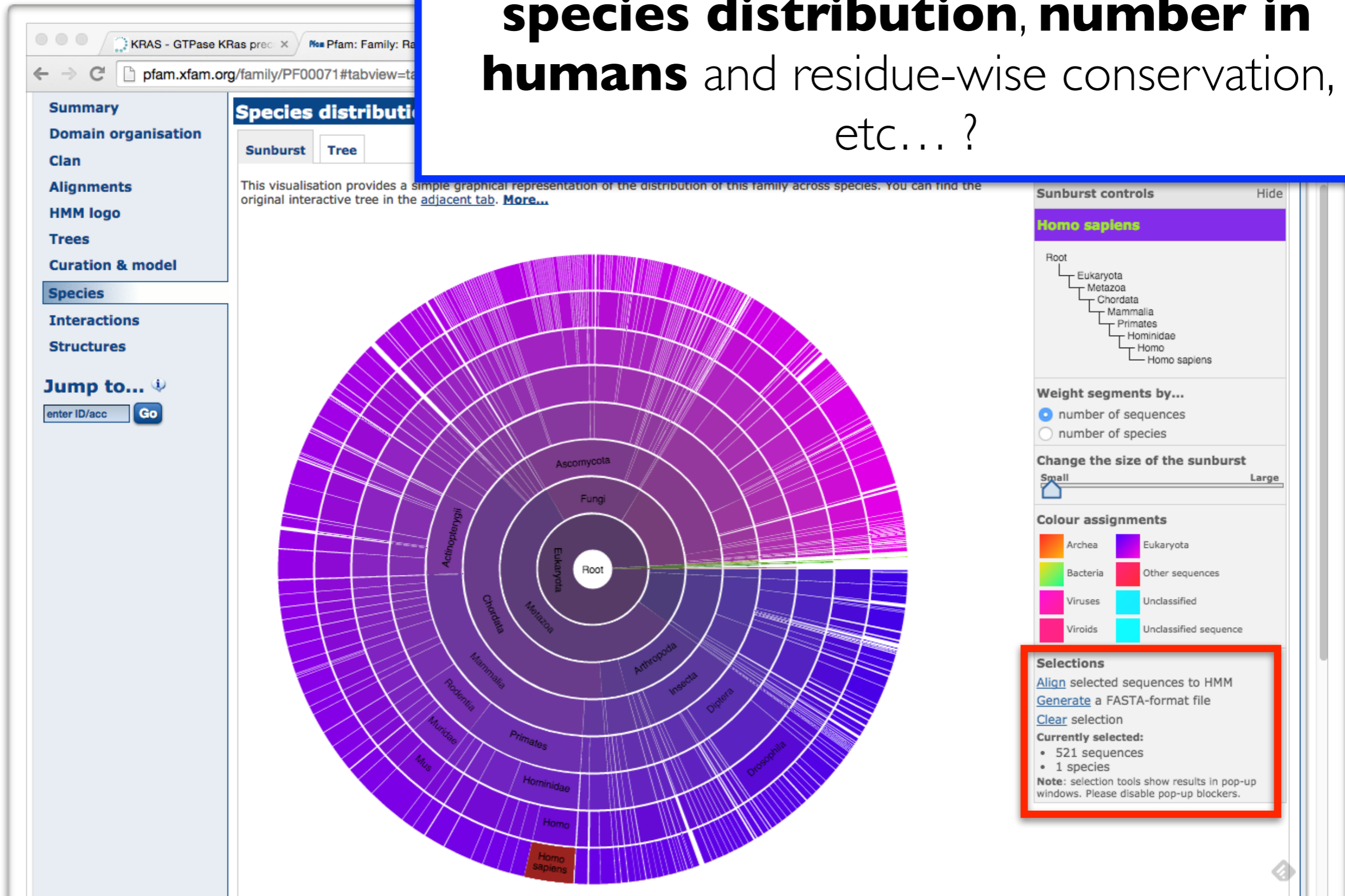


H-Ras structure PDB 121p, surface colored by conservation in Pfam seed alignment: gold, most conserved; dark cyan, least conserved.

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

Example Questions:

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



Example Questions:

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

The image displays a multi-panel view of the Pfam website. The top panel shows the 'Species distribution' section with a tree diagram highlighting 'Homo sapiens'. The middle panel shows an 'Alignment for selected sequences' with a table of protein sequences and their residue alignments. The bottom panel shows 'Sunburst controls' for 'Homo sapiens', including a tree diagram, 'Weight segments by...' options, 'Change the size of the sunburst' slider, and 'Colour assignments' for different taxonomic groups. A red box highlights the 'Selections' section, which includes options to 'Align selected sequences to HMM', 'Generate a FASTA-format file', and 'Clear selection'. The 'Currently selected:' section shows 521 sequences and 1 species.

Species distribution

Alignment for selected sequences

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

Accession	Sequence
P11234/16-178	..KVIMVSGGCVKNSALTL.....Q.....FM.....Y..D..EF.V....E.DYEPTK.-AD...SYRKKVLD....
P01112/5-165	..KLVVVGAGGCVKNSALTI.....Q.....LI.....Q..N..HF.V....D.EYDPTI.-ED...SYRKQVVID....
Q14088/38-204	..KIIIVIGDSNVGKNTCLTF.....R.....FC.....G...TF.P....D.KTEATI.GVD...FREKTVEIE....
Q9BW83/7-173	..KCILAGDPAVGKTALAQ.....I.....FR.....S..DgaHF.Q...K.SYTLT.GMD...LVVKTVPVEd....
P15153/5-178	..KCVVVGDGAVGKTCLLI.....S.....YT.....T..N..AF.P....G.EYIPTV.-FD...NYSANVMVD....
O00194/11-183	..KLLALGDSGCVKNTFLY.....R.....YT.....D..N..KF.N...P.KFITTV.GID...FREKRNVYNagppn
Q15907/13-174	..KVVLI GDSGCVKNSLLS.....R.....FT.....R..N..EF.N...L.ESKSTI.GVE...FATRSIOVD....
P10114/5-166	..KVVVLGSGGCVKNSALTV.....Q.....FV.....T..G..TF.I...E.KYDPTI.-ED...FYRKEIEVD....
P51153/10-171	..KLLLI GDSGCVKNTCLII.....R.....FA.....E..D..NF.N...N.TYISTI.GID...FKIRTVDIE....
P55040/77-241	..RVVLI GEGCVKNSSTLAN.....I.....FA.....GvhD..SM.D...S.D-CEVL.GED...TYERTLMVD....
P55042/93-253	..KVVLLGAPGCVKNSALAR.....I.....FG.....G..V..ED.G...P.EAEAAG.--H...TYDRSIVVD....
P01116/5-165	..KLVVVGAGGCVKNSALTI.....Q.....LI.....Q..N..HF.V....D.EYDPTI.-ED...SYRKQVVID....
Q9H077/21-182	..KLVLLGSGSVKNSLLAL.....R.....YV.....K..N..DF.K...S.-ILPTV.GCA...FFTKVVVDV....
Q9ULC3/11-171	..KVVVVGNGAVGKSSMIQ.....R.....YC.....K...IF.T...K.DYKKEI.GVD...FLERQIQVN....
Q14807/15-177	..KLVVVGDGCVKNSALTI.....Q.....FF.....Q..K..IF.V...P.DYDPTI.-ED...SYLKHTTEID....
Q9NX57/7-202	..KIVLLGDMNVGKNTSLLO.....R.....YM.....E..R..RF.P...D.T-VSTV.GCA...FYLKQW---....
Q9H082/35-201	..KIIIVIGDSNVGKNTCLTY.....R.....FC.....A...G..RF.P...D.RTEATI.GVD...FRERAVEID....
Q969Q5/9-174	..KVVMLGKEYVGNSTLVE.....R.....YV.....H..D..RFIV..G.PYQNEI.GAA...FVAKVMSVG....
P51149/10-175	..KVVILGDSGCVKNTSLMN.....Q.....YV.....N..K..KF.S...N.QYKATI.GAD...FLTKEVMVD....
Q9ULW5/65-227	..KVVMLV GDSGCVKNTCLLV.....R.....FK.....D...AF.L...AgTFISTV.GID...FRNKVLDVD....
P57735/14-175	..KVVLI GESGCVKNTNLS.....R.....FT.....R..N..EF.S...H.DSRTI.GVE...FSTRTVMLG....
P51159/11-183	..KFLALGDSGCVKNTSVLY.....Q.....YT.....D...G..KF.N...S.KFITTV.GID...FREKRNVYRasgpd
P01111/5-165	..KLVVVGAGGCVKNSALTI.....Q.....LI.....Q..N..HF.V....D.EYDPTI.-ED...SYRKQVVID....
P11233/16-177	..KVIMVSGGCVKNSALTL.....Q.....FM.....Y..D..EF.V....E.DYEPTK.-AD...SYRKKVLD....
Q9UL25/21-182	..KVVLLGEGCVKNTSLVL.....R.....YC.....E..N..KF.N...D.KHITL.QAS...FLTKKLNI....
Q9NP72/10-171	..KILII GESGCVKNSLLL.....R.....FT.....D..D..TF.D...P.ELAATI.GVD...FKVKTISVD....
Q9H0U4/10-171	..KLLLI GDSGCVKNSCLL.....R.....FA.....D..D..TY.T...E.SYISTI.GVD...FKIRTIELD....
Q9UL26/7-168	..KVCLLGDTGCVKNSIIVW.....R.....FV.....E..D..SF.D...P.NINPTI.GAS...FMTKTVOYQ....
Q9UBK7/23-179	..KIICLGDSAVGKSKLME.....R.....FL.....M..D..GF.Q...P.QQLSTY.ALT...LYKHTATVD....
P51157/14-179	..KIVVLGDGASGKNTSLTT.....C.....FA.....Q..E..TF.G...K.QYKQTI.GLD...FFLRRITL....

Sunburst controls Hide

Homo sapiens

Root
Eukaryota
Metazoa
Chordata
Mammalia
Primates
Hominidae
Homo
Homo sapiens

Weight segments by...

number of sequences
 number of species

Change the size of the sunburst

Small Large

Colour assignments

	Archea		Eukaryota
	Bacteria		Other sequences
	Viruses		Unclassified
	Viroids		Unclassified sequence

Selections

[Align](#) selected sequences to HMM
[Generate](#) a FASTA-format file
[Clear](#) selection

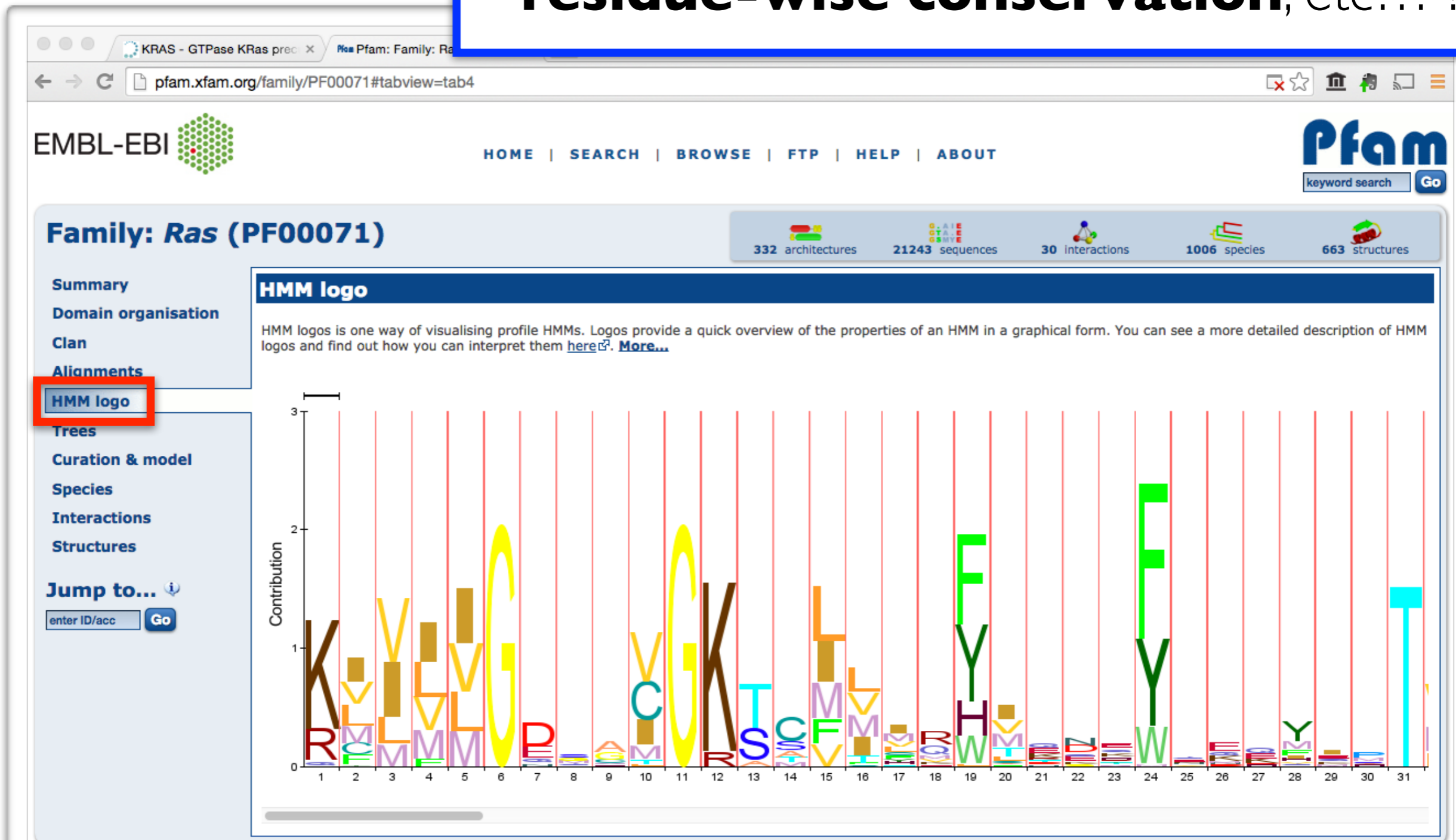
Currently selected:

- 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... ?



Family: *Kinesin* (PF00225)

⌂ Loading page components (1 remaining)...

 126 architectures
  4150 sequences
  6 interactions
  248 species
  114 structures

Summary

Domain organisation

Clans

Alignments

HMM logo

Trees

Curation & models

Species

Interactions

Structures

Jump to...

Interactions

There are **6** interactions for this family. [More...](#)

[Tubulin](#)
[Tubulin_C](#)

[Tubulin_C](#)

[Kinesin](#)

[Tubulin](#)

[Kinesin](#)

Questions or comments: pfam@janelia.hhmi.org

Howard Hughes Medical Institute

Family: *Kinesin* (PF00225)

 126 architectures
  4150 sequences
  6 interactions
  248 species
  114 structures

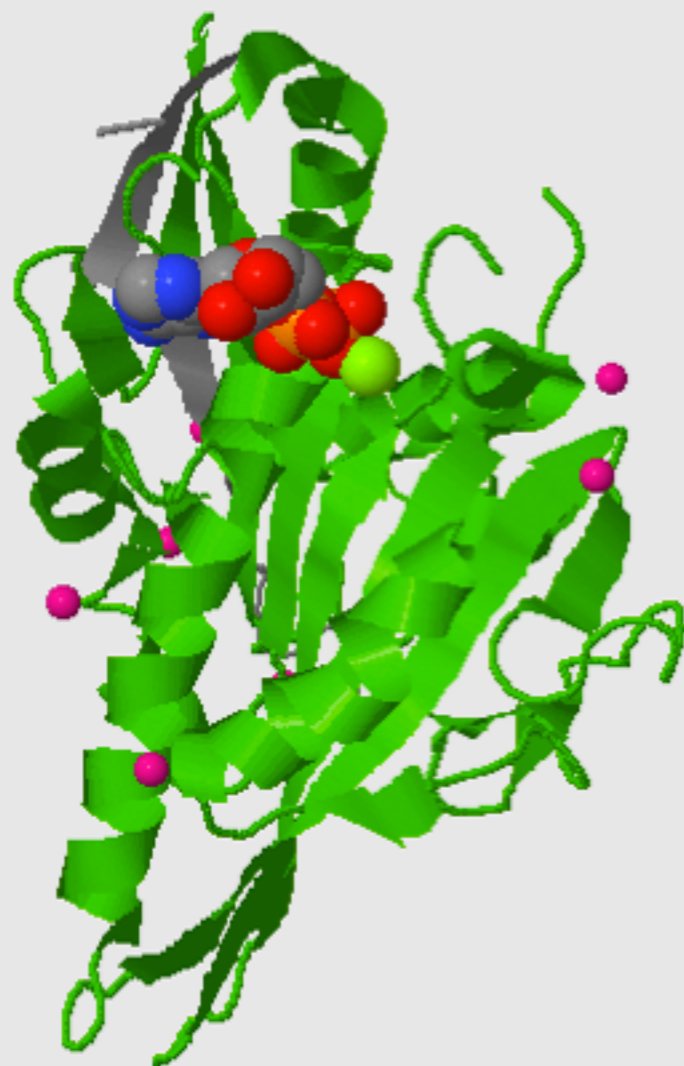
[Summary](#)
[Domain organisation](#)
[Clans](#)
[Alignments](#)
[HMM logo](#)
[Trees](#)
[Curation & models](#)
[Species](#)
[Interactions](#)
[Structures](#)
[Jump to...](#)

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 [PDB](#) 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	2vvq	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
			1f9w	B	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
			1ii6	B	24 - 359	Jmol AstexViewer SPICE
			1q0b	A	24 - 359	Jmol AstexViewer SPICE
				B	24 - 359	Jmol AstexViewer SPICE
			1x88	A	24 - 359	Jmol AstexViewer SPICE
				B	24 - 359	Jmol AstexViewer SPICE
			A	24 - 359	Jmol AstexViewer SPICE	

PDB entry 3bfm



Jmol

Your turn:
 What can you find out about “eg5”

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

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 <i>why</i> and <i>how</i> 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

mismatch
match

Two types of character
correspondence

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

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

match
mismatch

Add gaps to increase number of matches

gaps

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

match
mismatch } **mutation**
insertion } **indels**
deletion }
Gaps represent 'indels'
mismatch represent mutations

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

- **Mapping**

- **Pairwise sequence alignment is arguably the most fundamental operation of bioinformatics!**

- **Looking for differences from reference sequences, indels (insertions or deletions)**

- **Identifying 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)

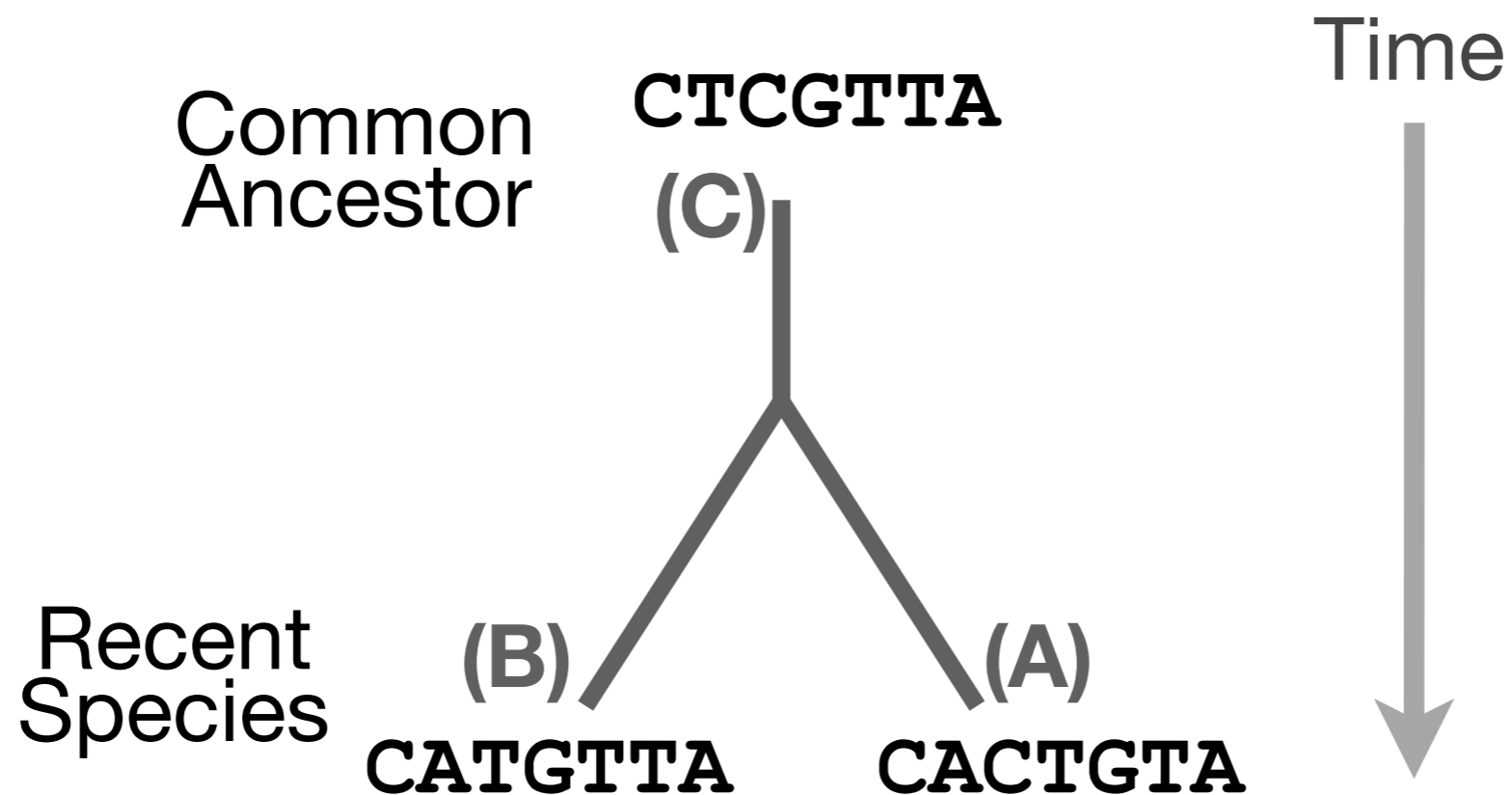
- **How...**

- ▶ 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.

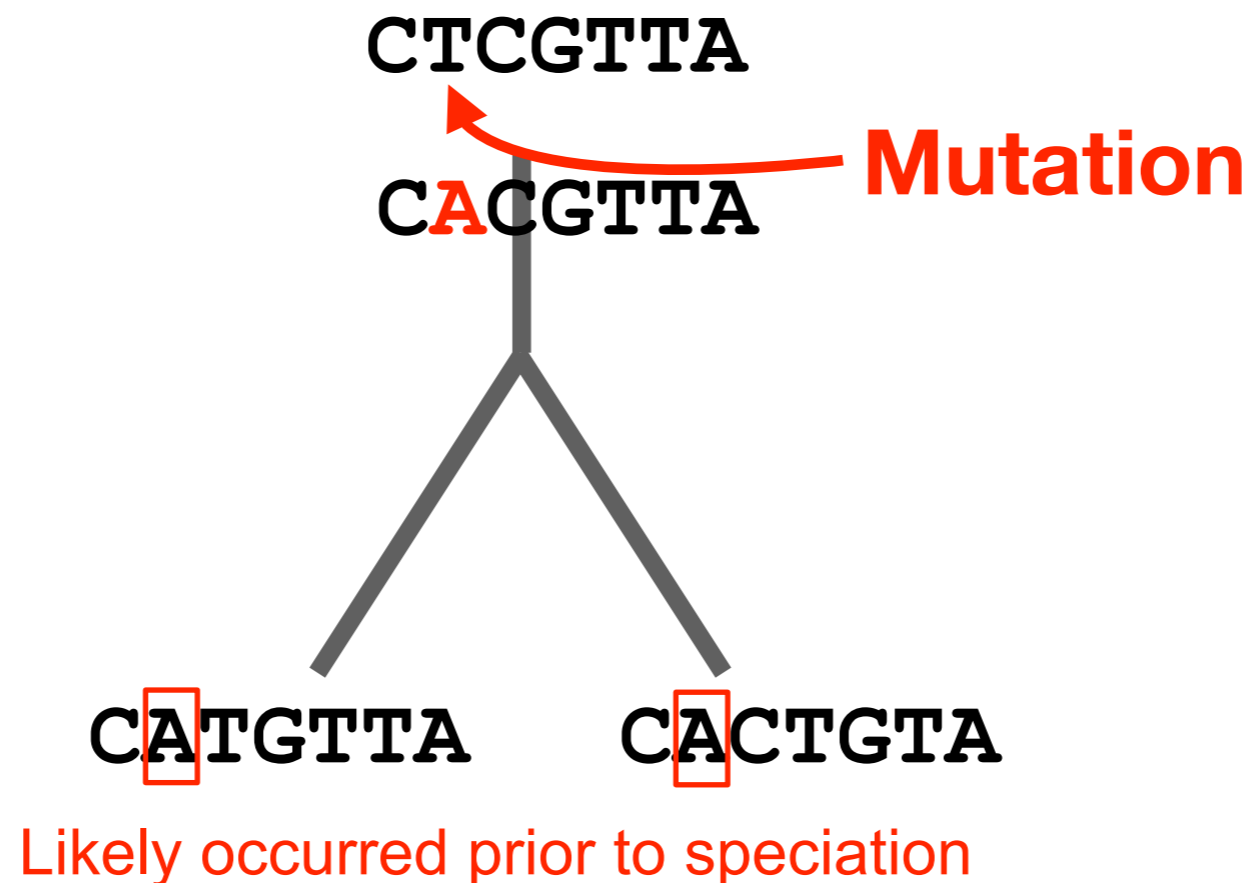
- Mutations/Substitutions
- Deletions
- Insertions



Mutations, deletions and insertions

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

- **Mutations/Substitutions** CTCGTTA → C**A**CGTTA
- Deletions
- Insertions

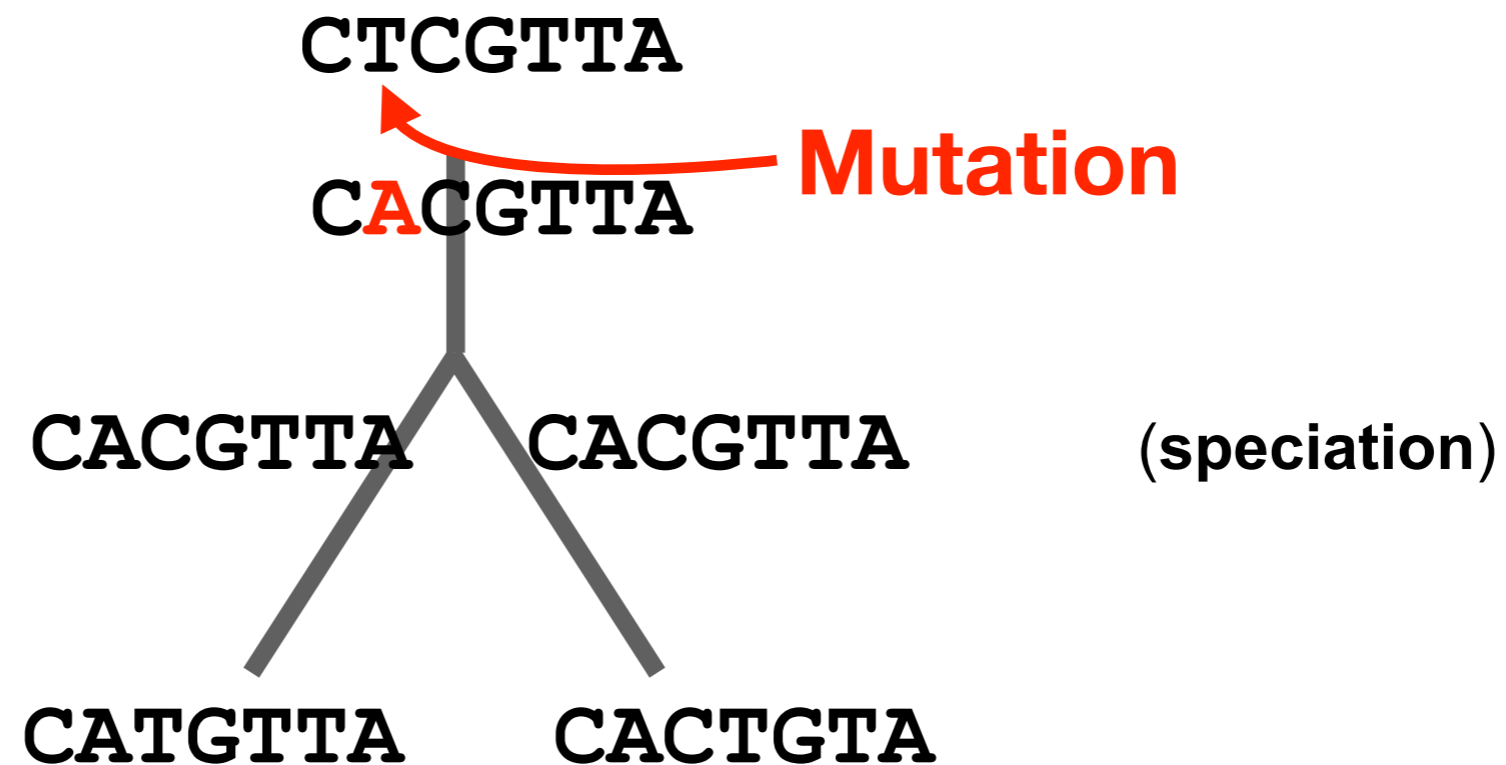


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → C**A**CGTTA



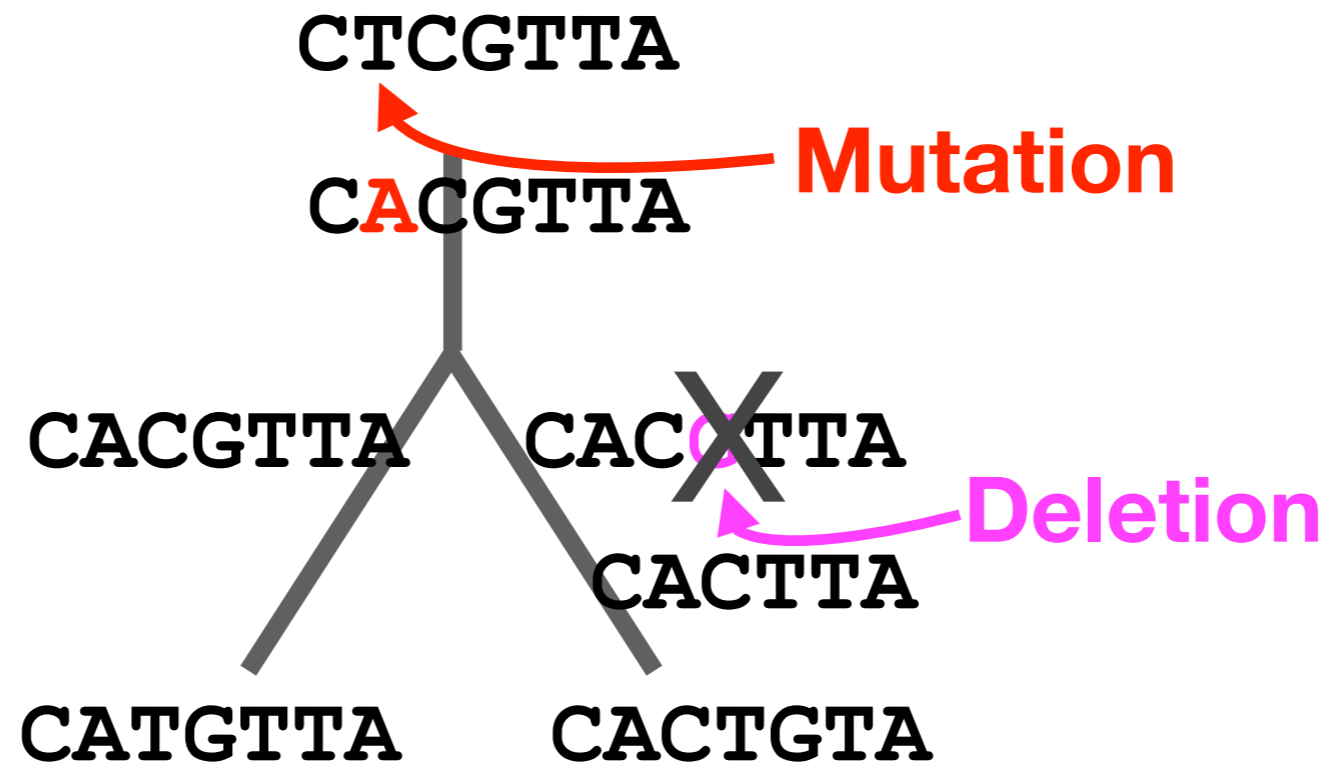
Mutations, deletions and insertions

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

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

CTCGTTA → C**A**CGTTA

CAC**G**TTA → CACTTA



Mutations, deletions and insertions

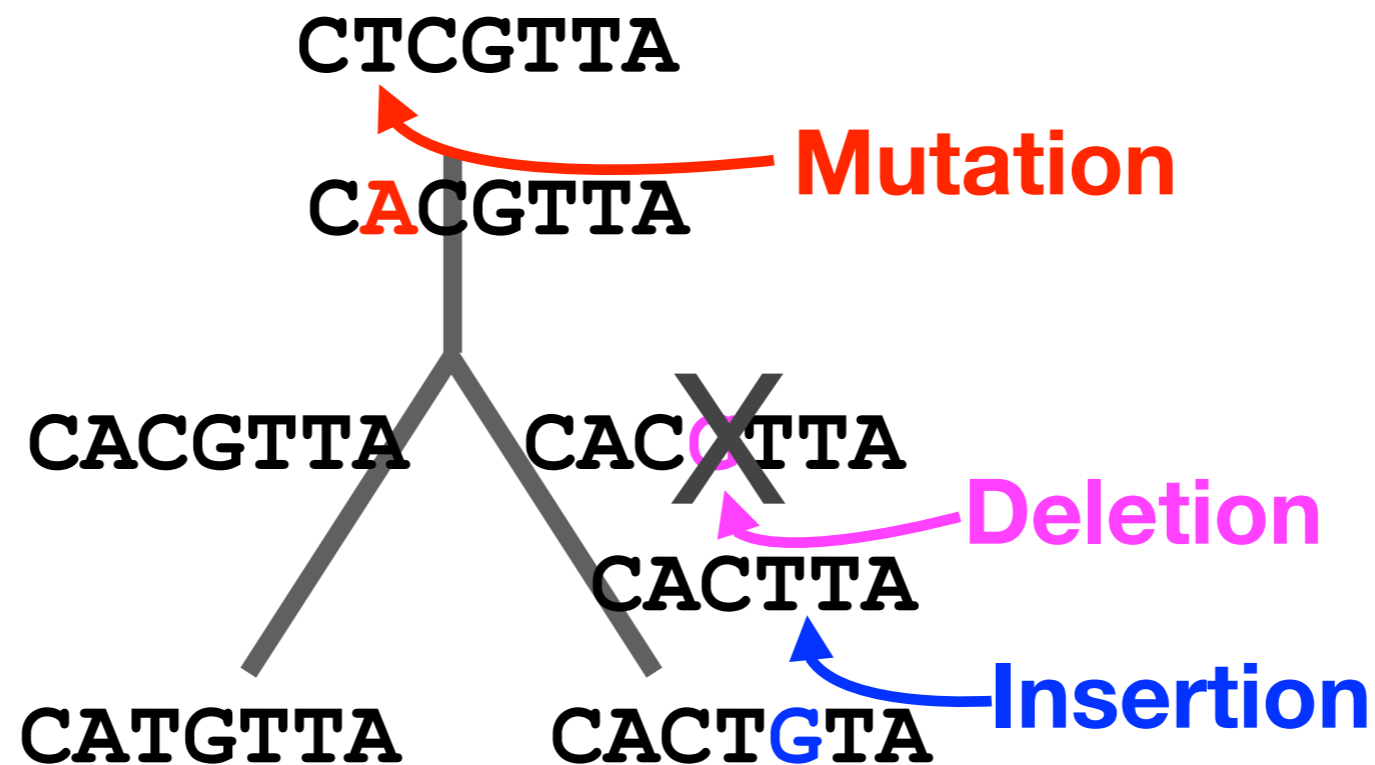
There are three major types of sequence change that can occur during evolution.

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

CTCGTTA → C**A**CGTTA

CAC**G**TTA → CACTTA

CACTTA → CACT**G**TA



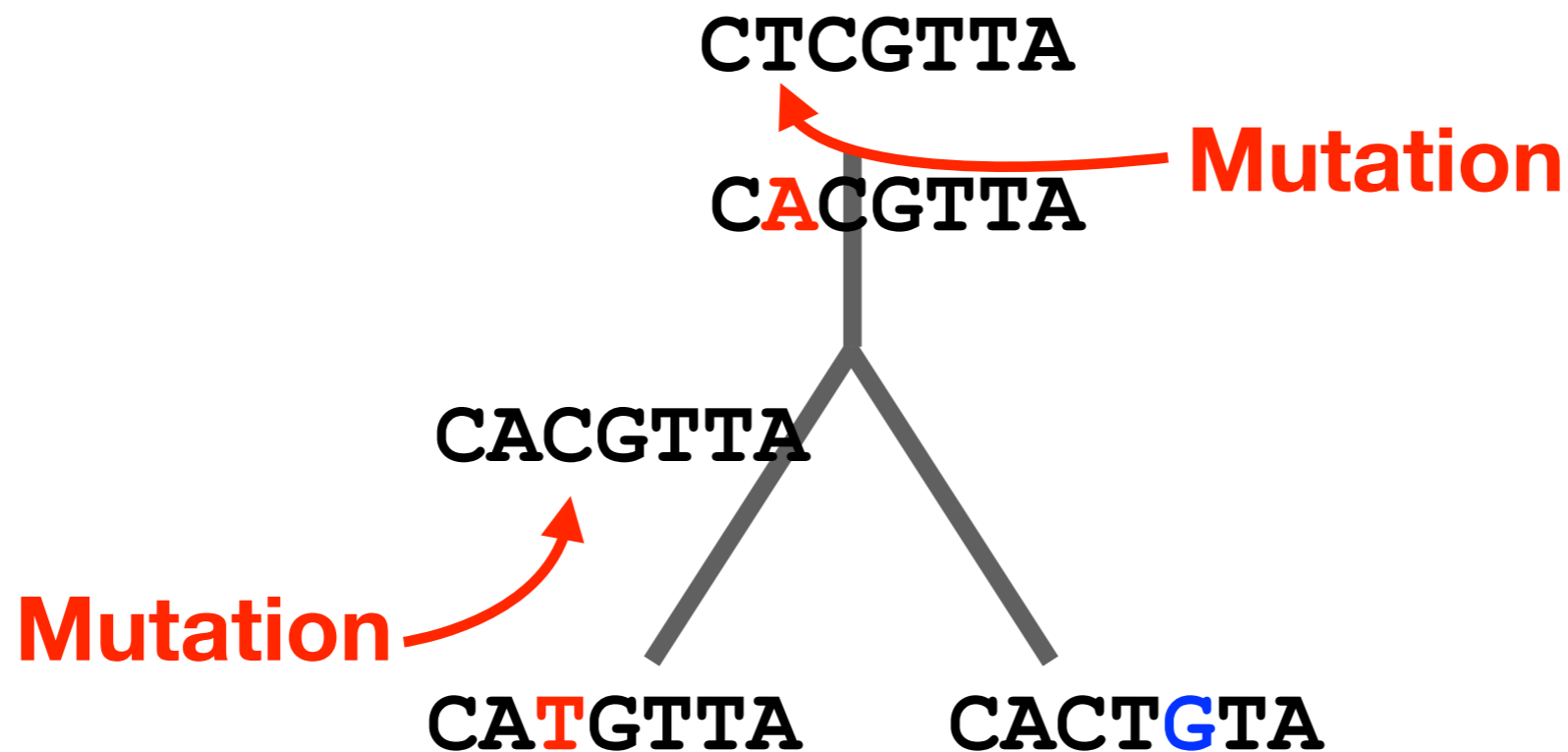
Mutations, deletions and insertions

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

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

CTCGTTA → C**A**CGTTA

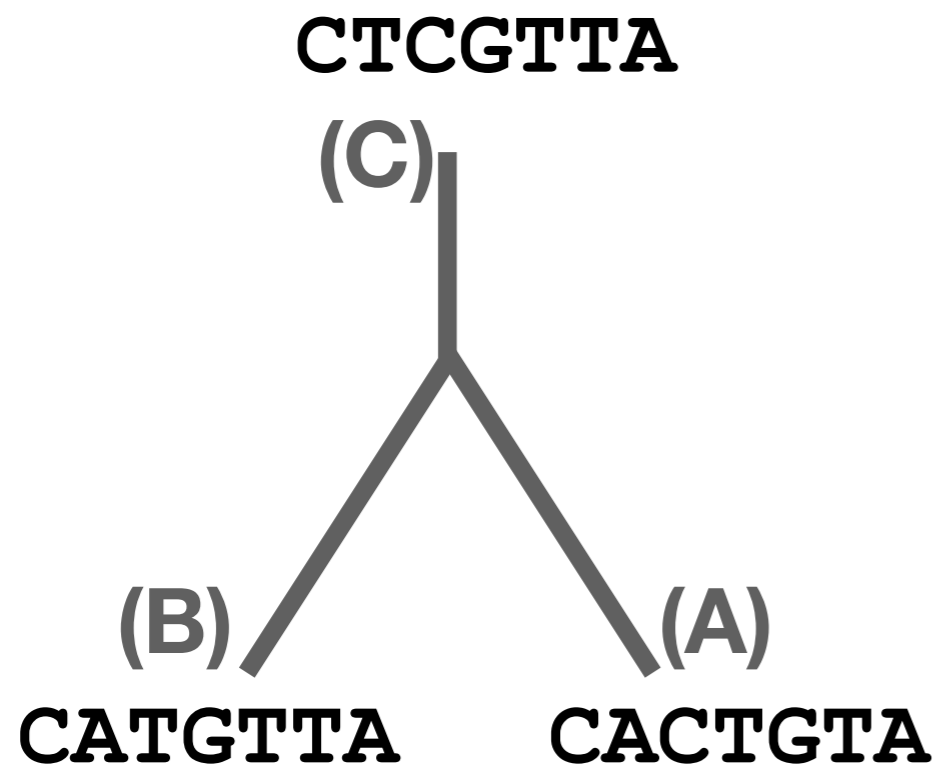
CACGTTA → CA**T**GTTA



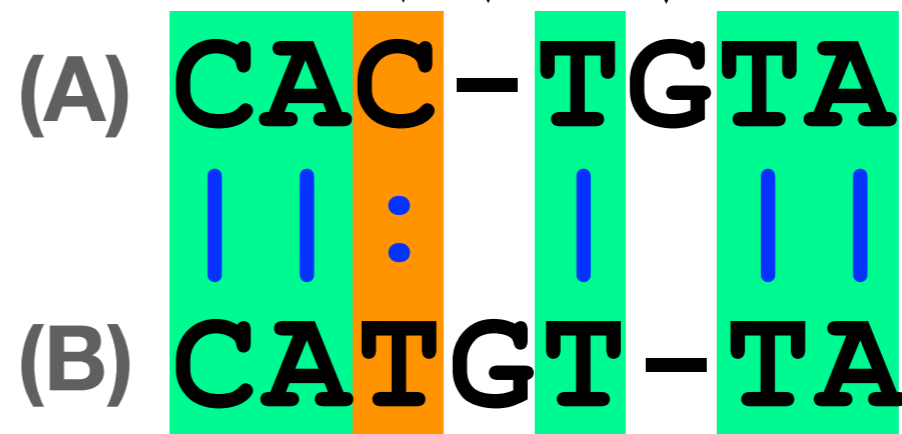
Alignment view

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

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



Substitution Indels

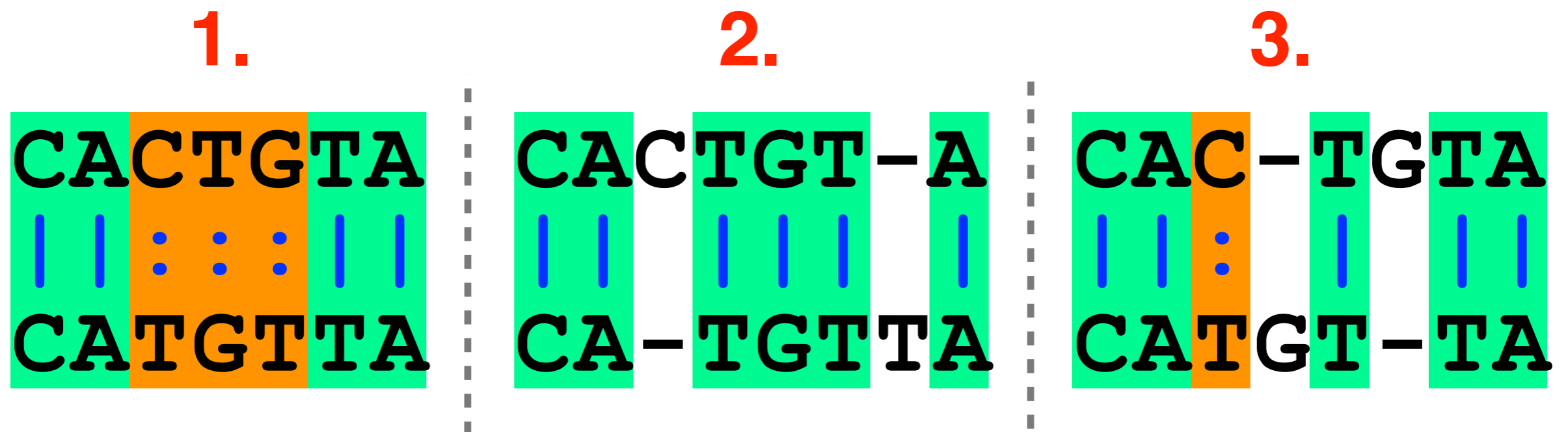


Match	Mismatch	Gap
5	1	2

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?



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 mismatches
○ 2 gaps



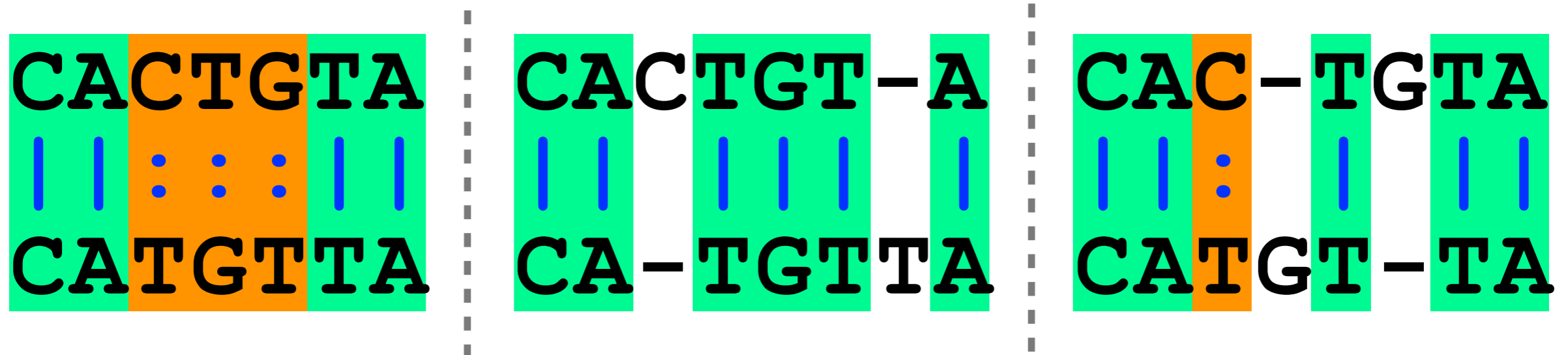
Scoring alignments

- 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



Optimal alignments

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

● 4 matches
● 3 mismatches
○ 0 gaps

CACTGTA
|| : : : ||
CATGTTA

● 6 matches
● 0 mismatches
○ 2 gaps

CAC TGT - A
|| | | | |
CA - TGT TA

● 5 matches
● 1 mismatches
○ 2 gaps

CAC - TGT A
|| : | | |
CATGT - TA

Optimal alignments

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

● 4 matches
● 3 mismatches
○ 0 gaps

CACTGTA
|| : : : ||
CATGTTA

● 6 matches
● 0 mismatches
○ 2 gaps

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

● 5 matches
● 1 mismatches
○ 2 gaps

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

Optimal alignments

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

● 4 matches
● 3 mismatches
○ 0 gaps

CACTGTA
|| : : : ||
CATGTTA

● 6 matches
● 0 mismatches
○ 2 gaps

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

● 5 matches
● 1 mismatches
○ 2 gaps

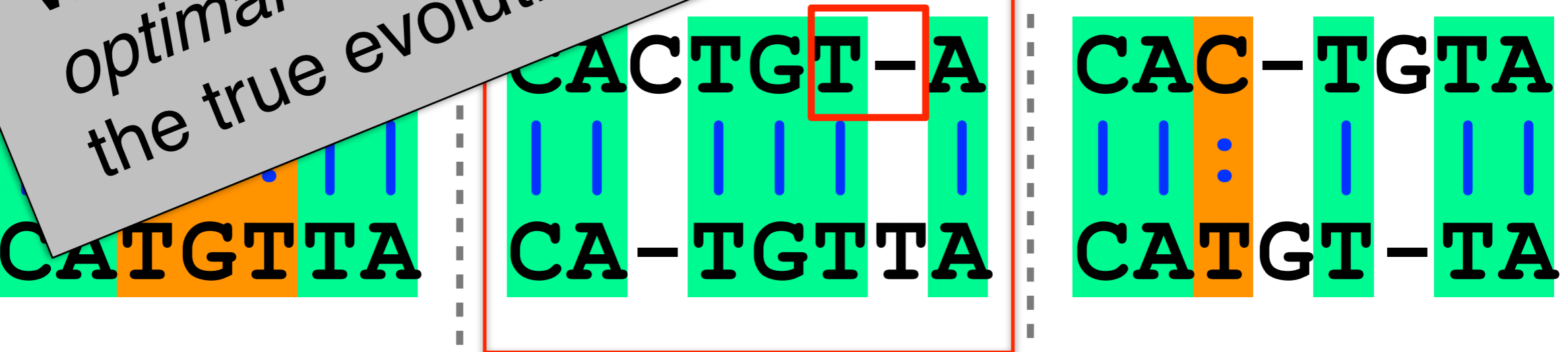
CAC-TGTA
|| : | ||
CATGT-TA

Optimal alignments

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

● 4 matches
● 3 mismatches
○ 2 gaps

● 5 matches
● 1 mismatch
○ 2 gaps



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

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

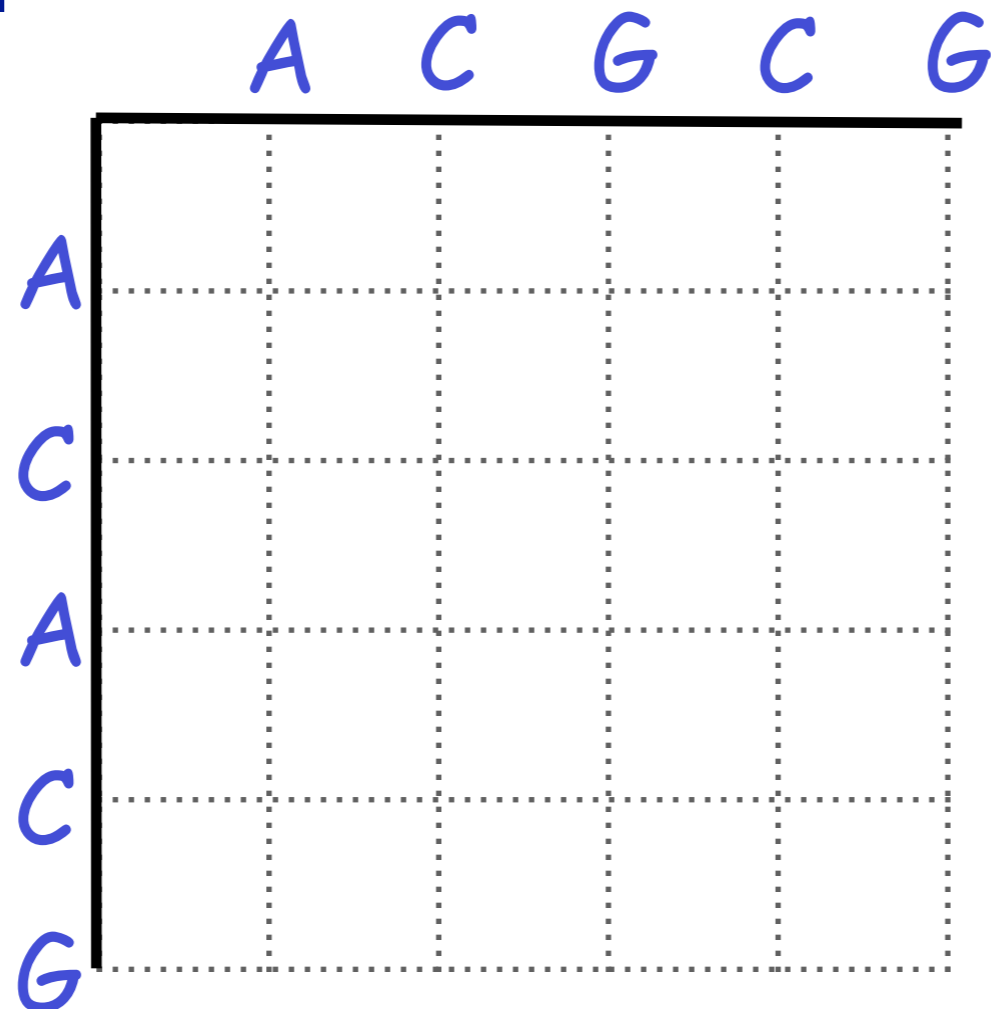
- ▶ D

How do we compute the optimal alignment between two sequences?

- ▶ BLAST heuristic approach

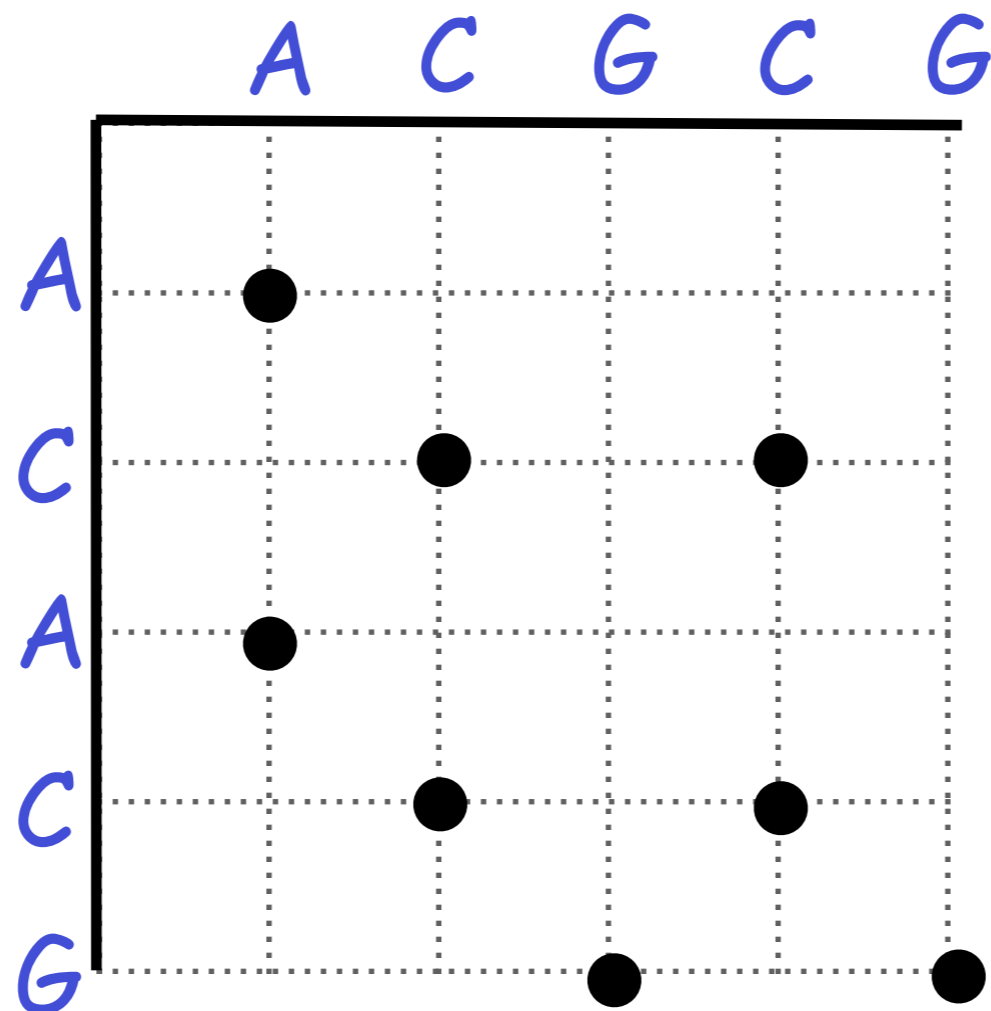
Dot plots: simple graphical approach

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



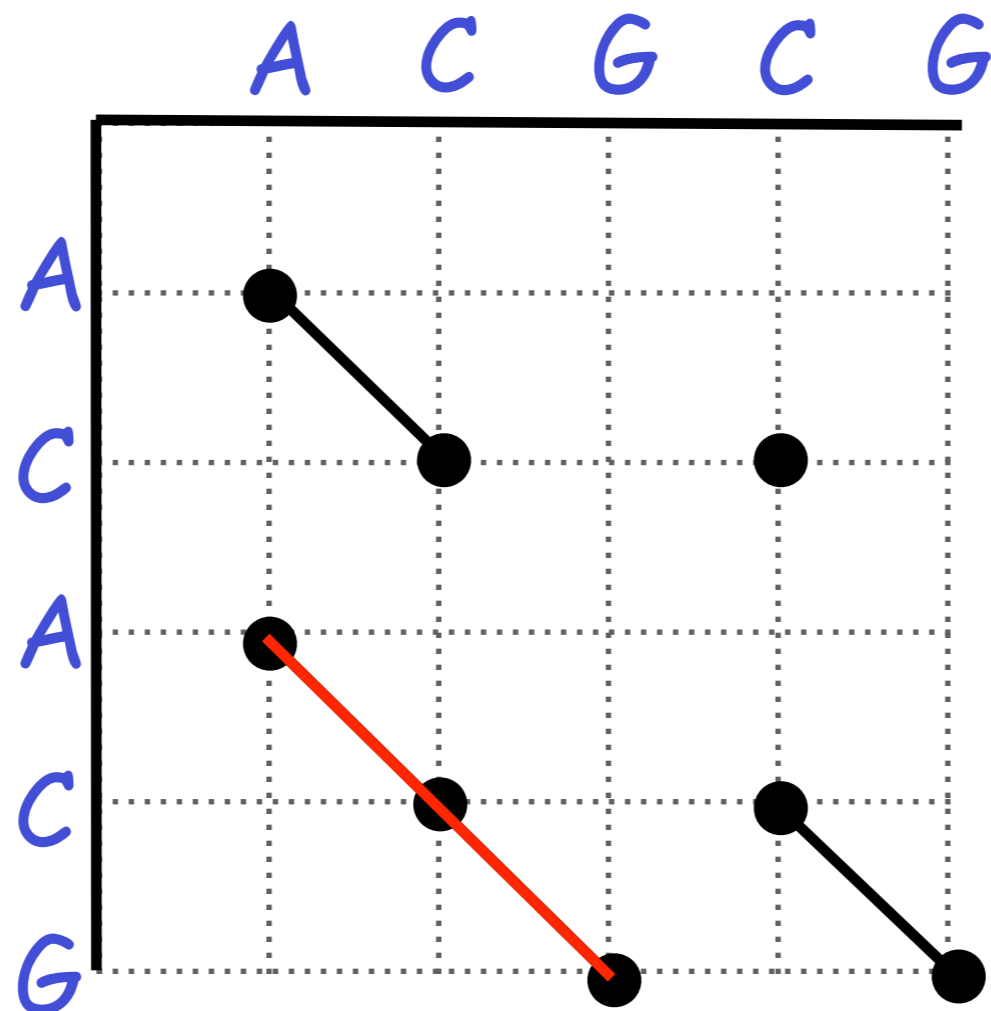
Dot plots: simple graphical approach

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



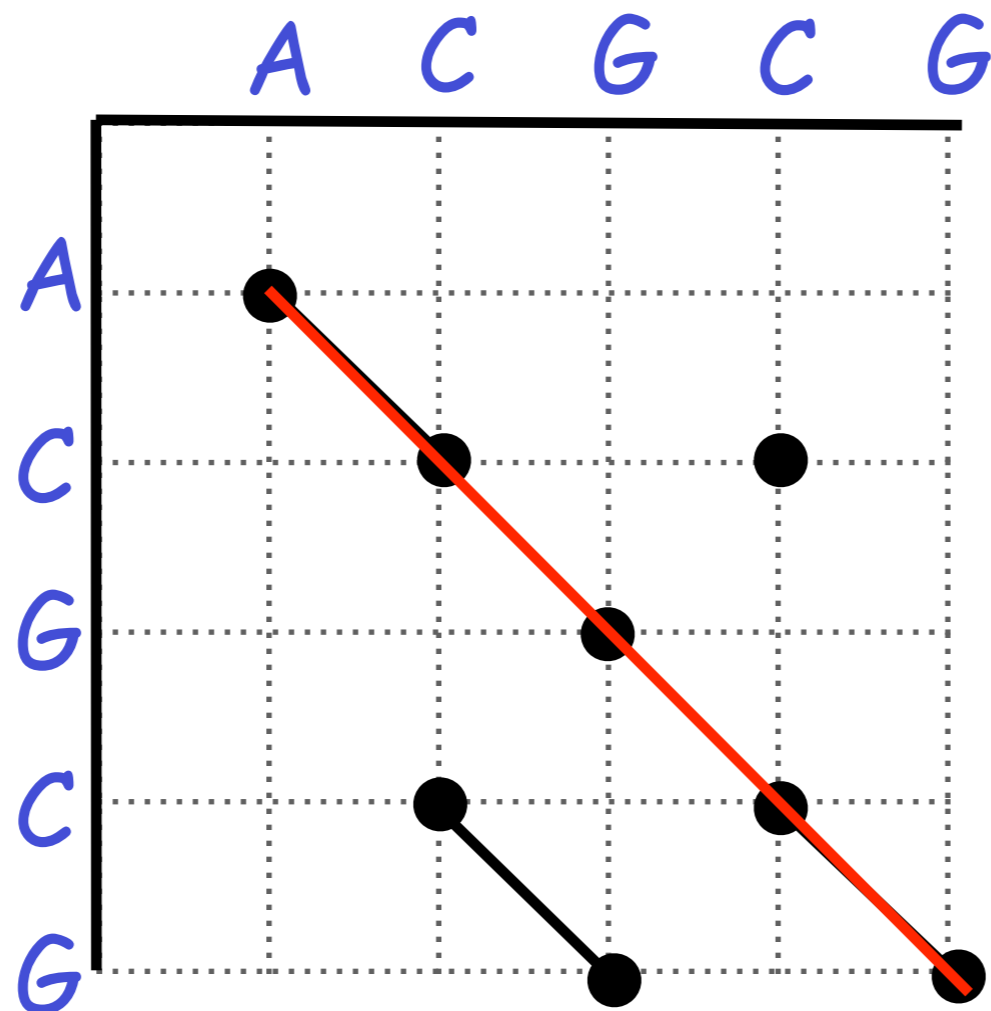
Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



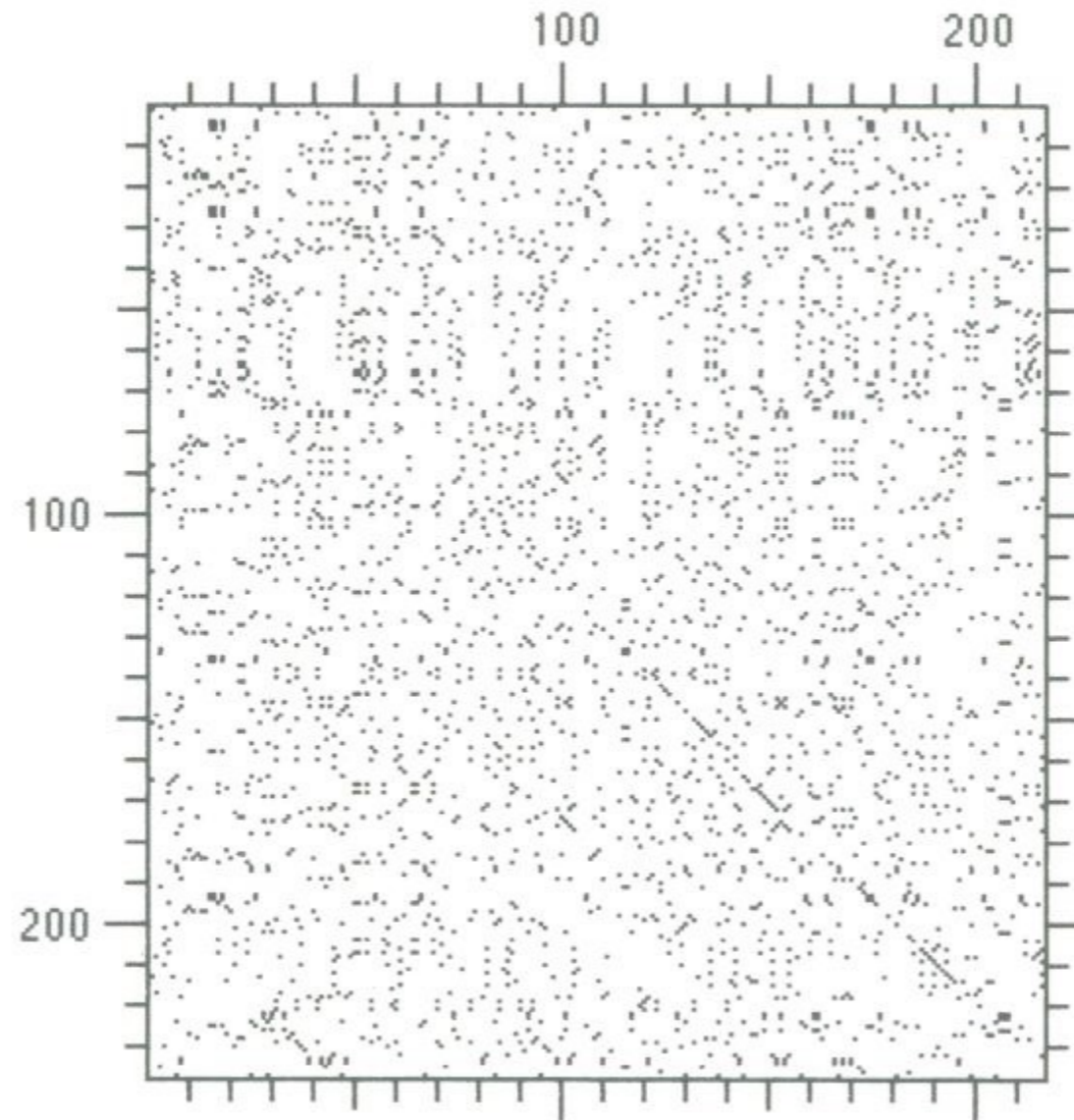
Dot plots: simple graphical approach

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



Dot plots: simple graphical approach

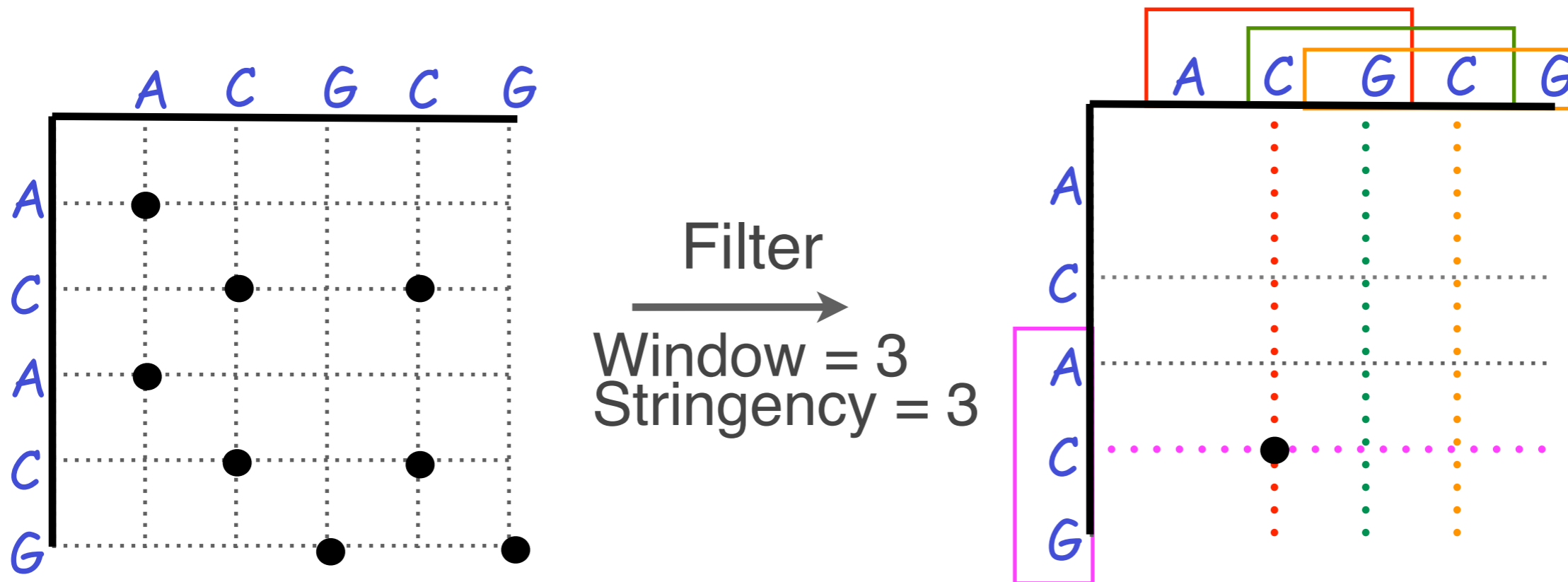
- Dot matrices for long sequences can be noisy



Dot plots: window size and match stringency

Solution: use a window and a threshold

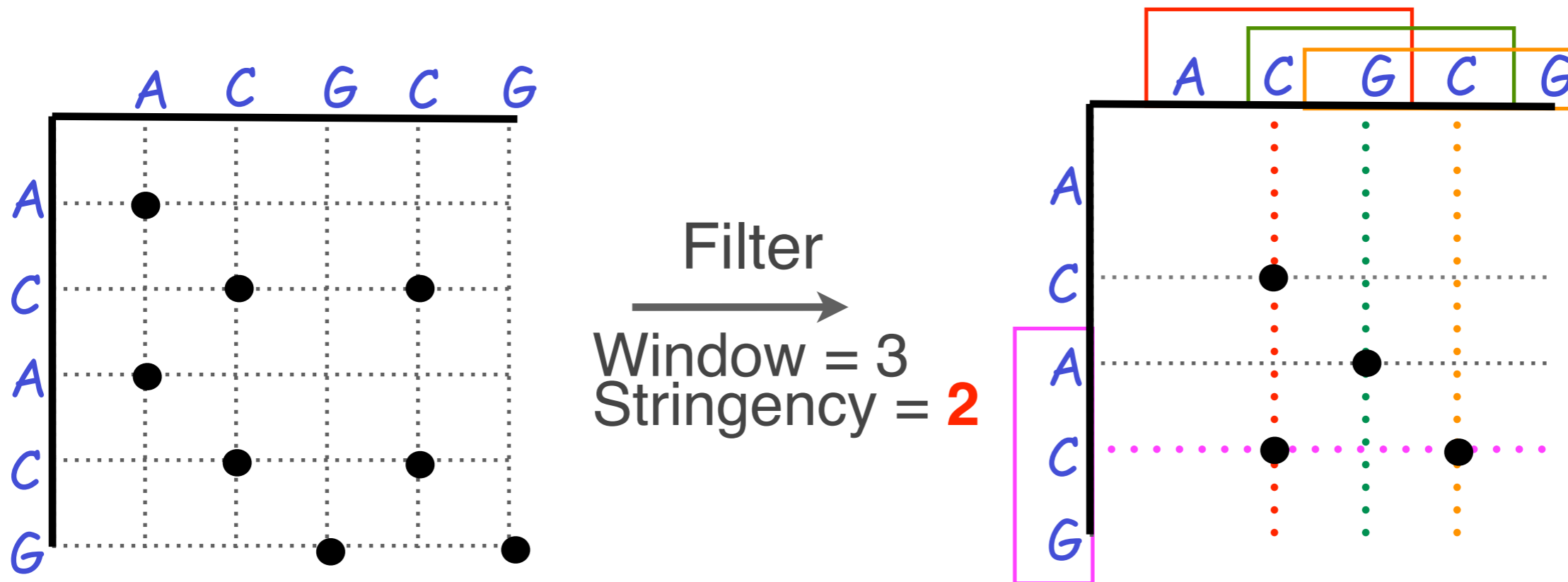
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
- You have to choose window size and stringency



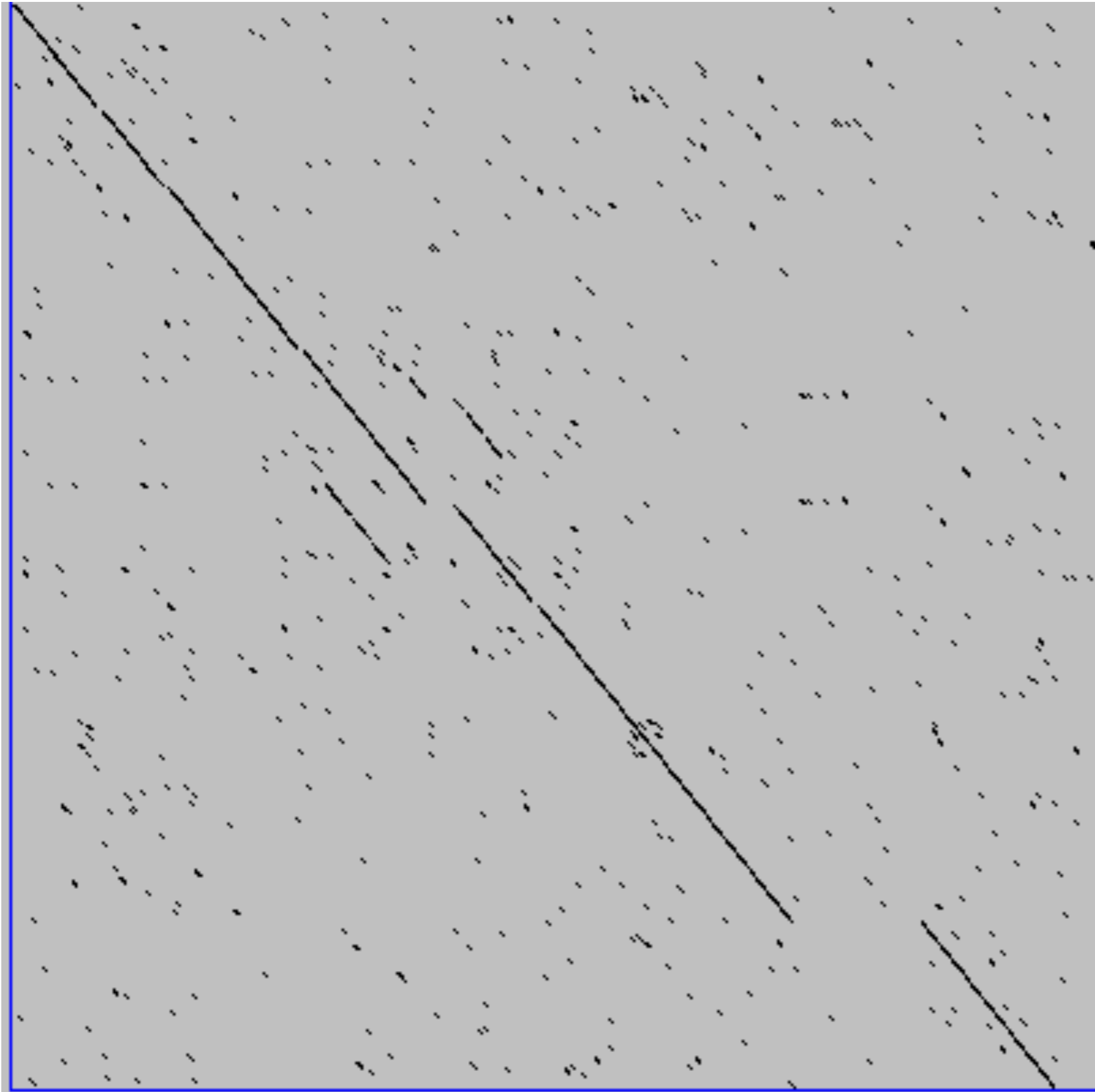
Dot plots: window size and match stringency

Solution: use a window and a threshold

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



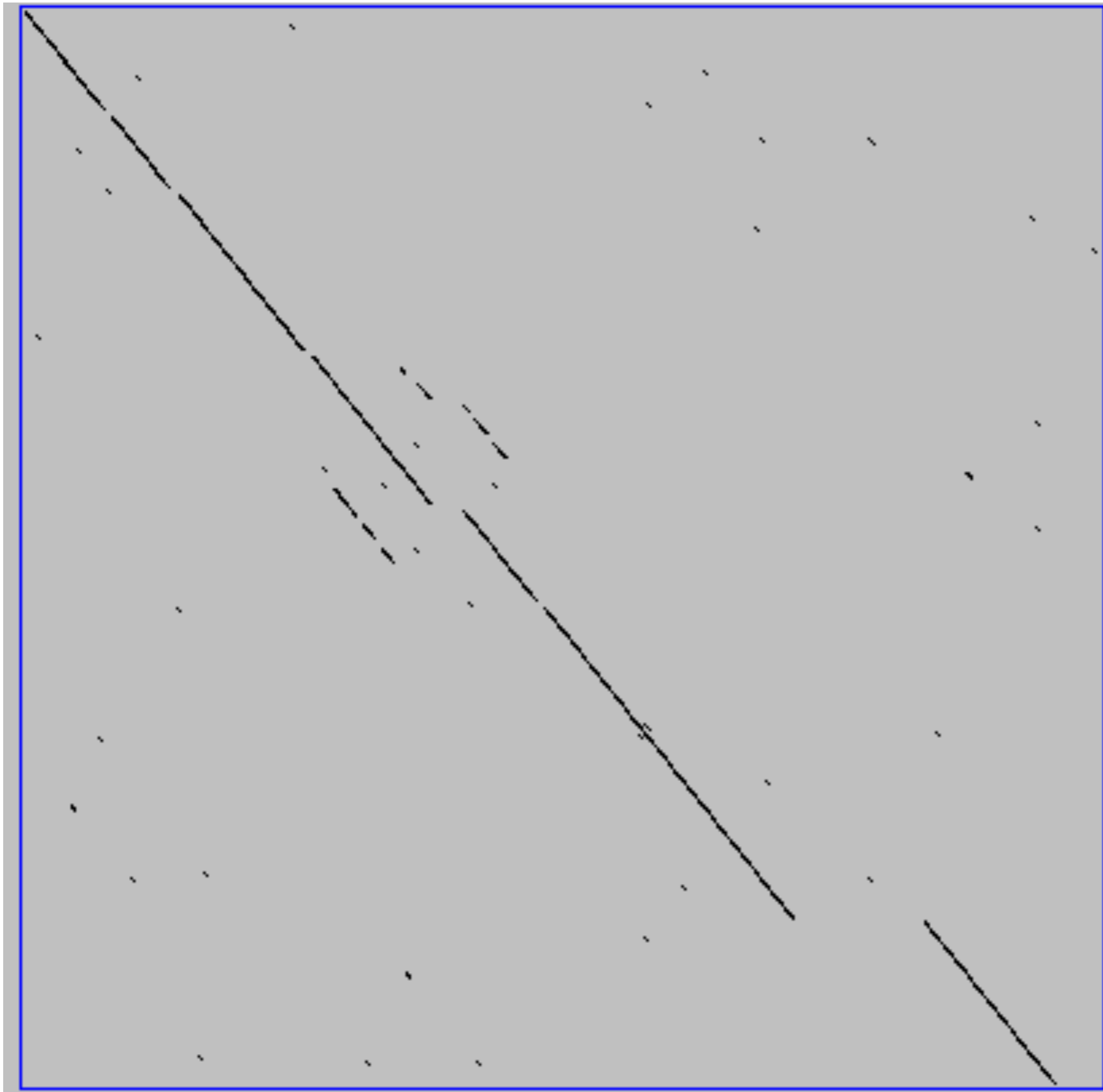
Window size = 5 bases



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

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

Window size = 7 bases

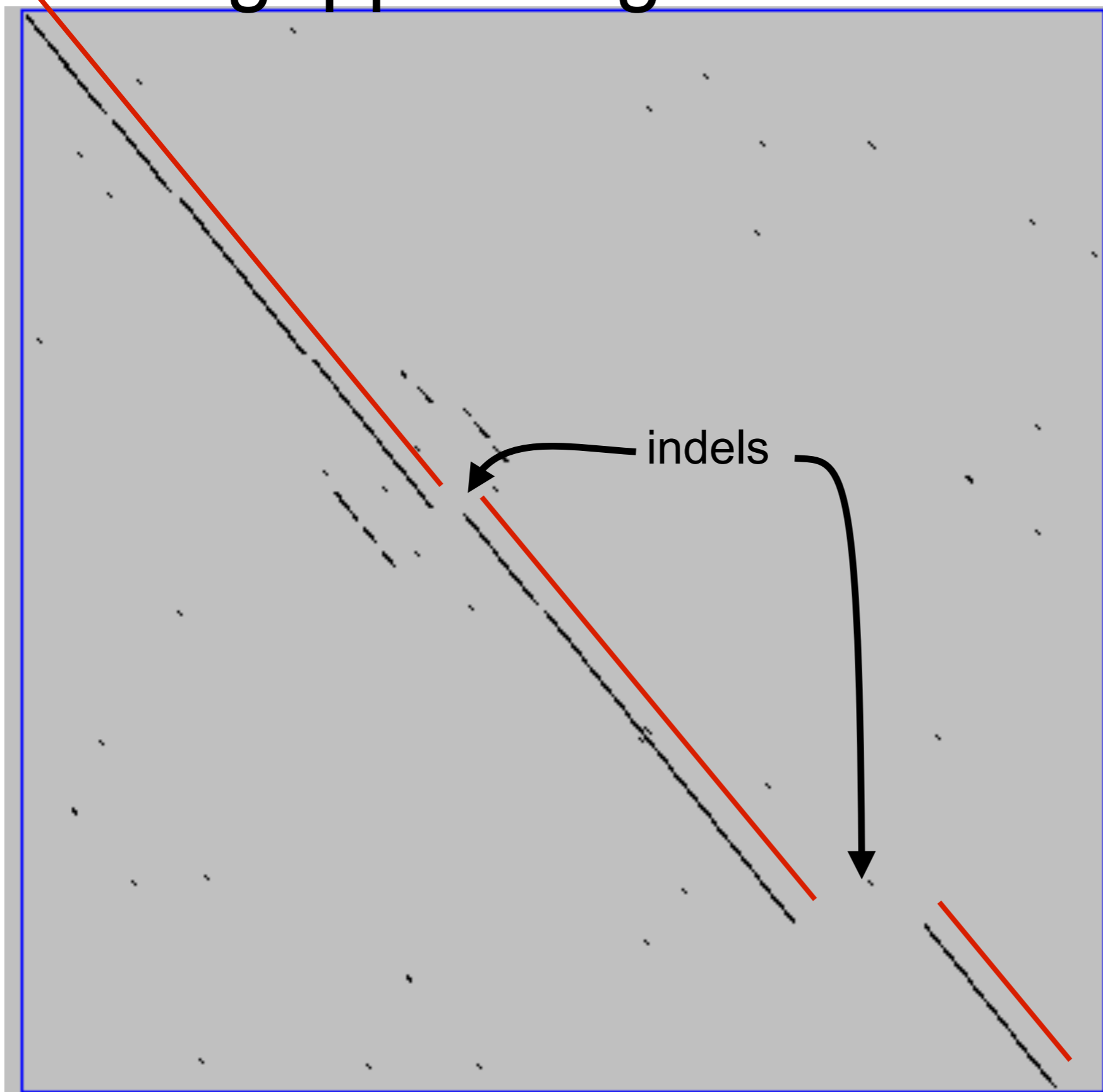


This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

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



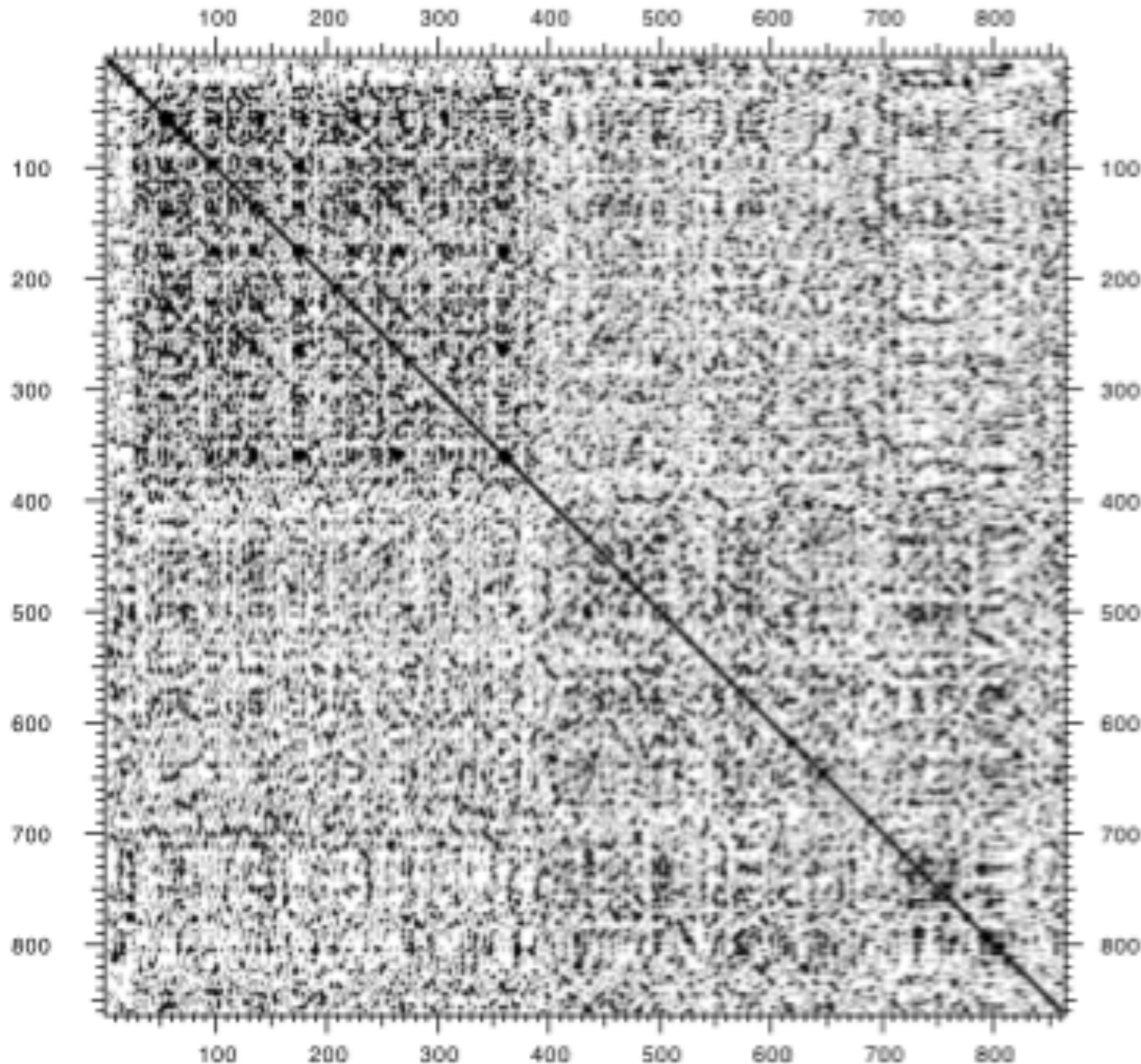
Only **diagonals** can be followed.

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



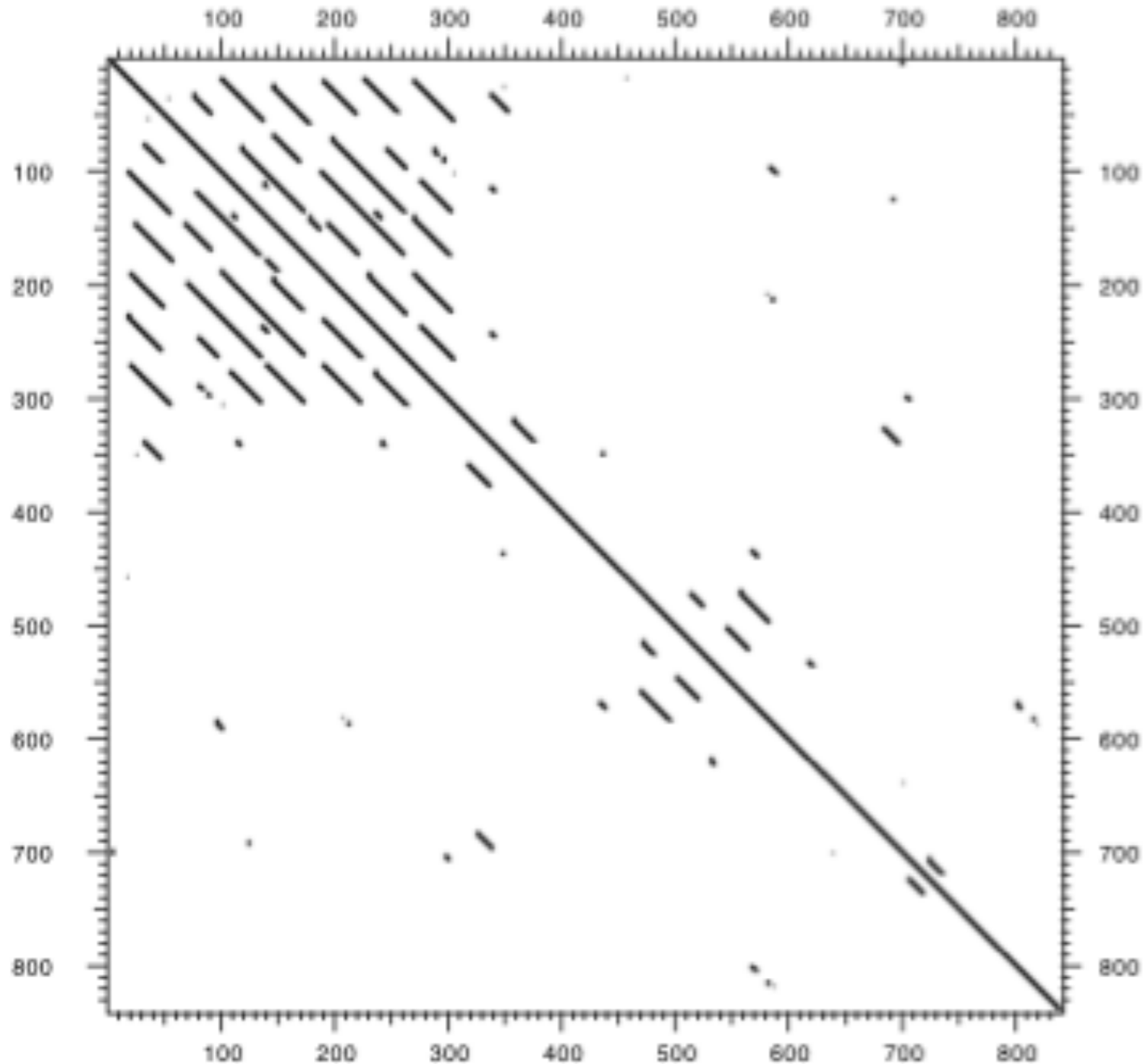
Human LDL receptor
protein sequence
(Genbank P01130)

$$W = 1$$

$$S = 1$$

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

Repeats



Human LDL receptor
protein sequence
(Genbank P01130)

$$W = 23$$

$$S = 7$$

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

Your Turn!

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

<http://bio3d.ucsd.edu/dotplot/>

<https://bioboot.shinyapps.io/dotplot/>

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 [Maizel 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 3 10

Moving window step size: 1 3 10

Match stringency: 1 2 10

Match stringency specifies the number of match characters required per window. It should not be larger than your window size!

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

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

<https://bioboot.shinyapps.io/dotplot2/>

Questions for discussion:

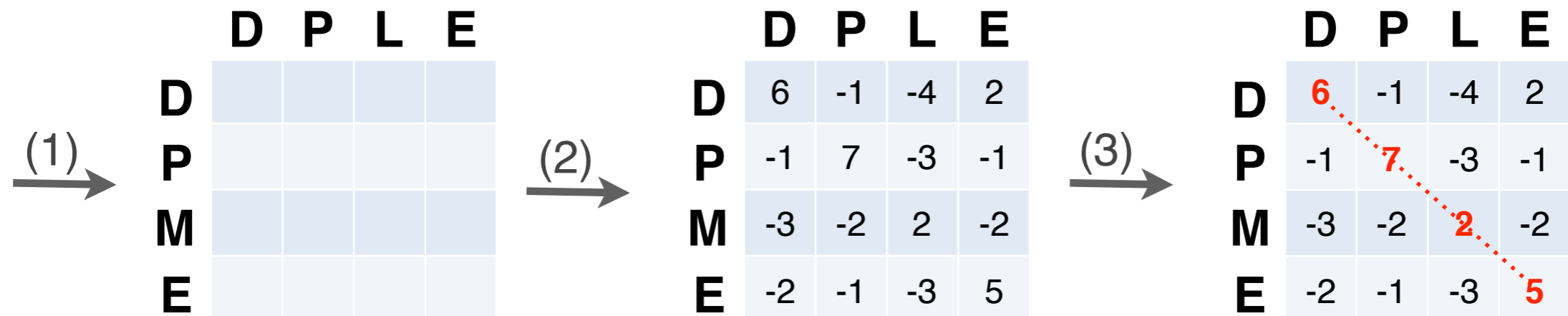
- Why does the DNA sequence have more dots than the protein sequence plot?
- How can we increase the signal to noise ratio?
- What does a 'Match stringency' larger than 'Window size' yield and why?

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

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

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

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

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

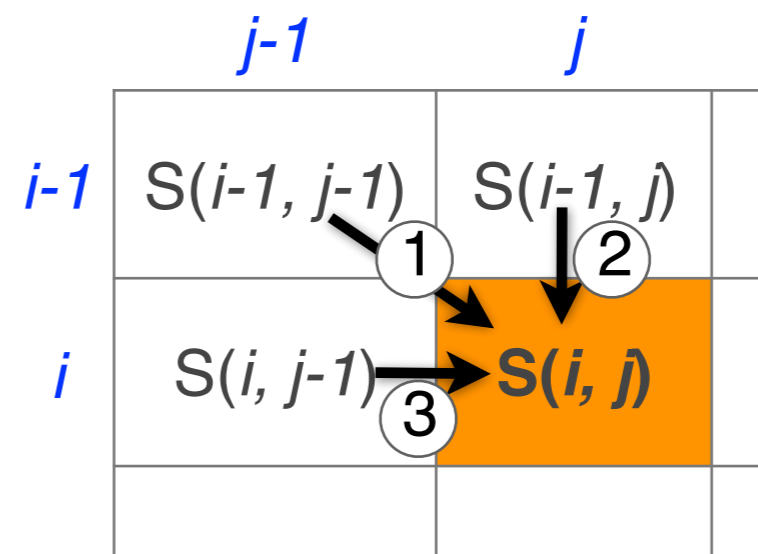
Seq1 : **DPME**
Seq2 : **----**

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

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

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



Scoring the alignment matrix

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

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

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

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

Scoring the alignment matrix

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

		<i>j</i> 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

↙ ① $(0) + (+1) = +1$ **<= (D-D) match!**

↓ ② $(-2) + (-2) = -4$

→ ③ $(-2) + (-2) = -4$

Alignment

D
D

Scoring the alignment matrix

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

	-	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!

↓ ② $(-4) + (-2) = -6$

→ ③ $(1) + (-2) = -1$

Alignment

D-
DP

Scoring the alignment matrix

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

	-	D	P	L	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 - -
D P L

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.

	-	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

↘ ① $(-1) + (-1) = -2$

↓ ② $(-3) + (-2) = -5$

→ ③ $(2) + (-2) = 0$

Alignment

DP-
DPL

Scoring the alignment matrix

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

	-	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	
E	-8				

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

↘ ① $(2) + (-1) = 0 \leq \text{mismatch}$

↓ ② $(0) + (-2) = -2$

→ ③ $(0) + (-2) = -2$

Alignment

DPM
DPL

Scoring the alignment matrix

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

		-	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	

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

→ ① $(+1)+(+1) = +2$

↓ ② $(-1)+(-2) = -3$

→ ③ $(-1)+(-2) = -3$

Alignment

DPME
DPLE

Scoring the alignment matrix

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

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

	-	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?

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

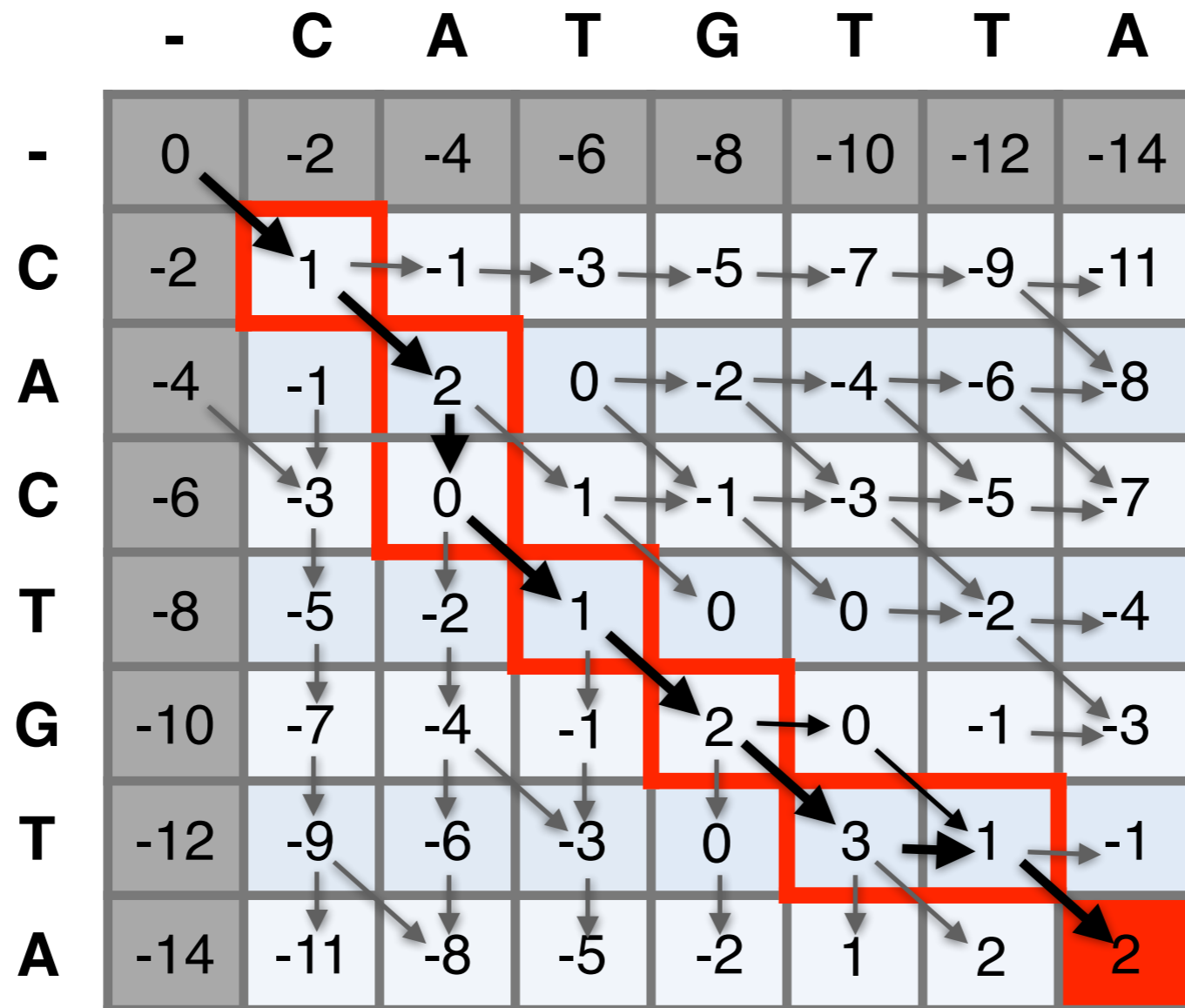
Questions:

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

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

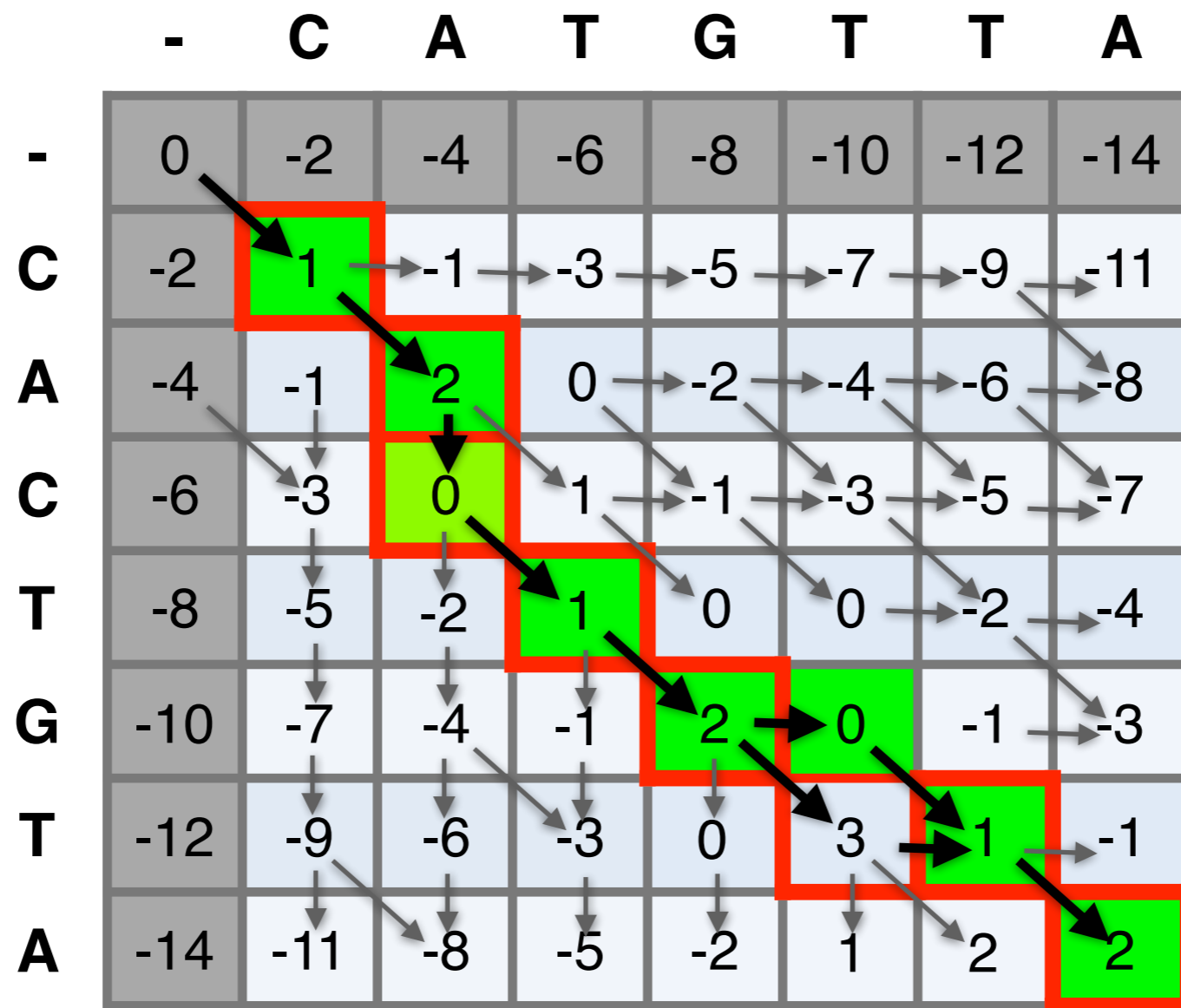
Questions:

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



More than one alignment possible

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



Alignment
CACTGT-A
CA-TGTTA

CACTG-TA
CA-TGTTA

The alignment and score are dependent on the scoring system

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

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

Alignment
CACTGT-A
CA-TGTTA

CACTG-TA
CA-TGTTA

CACTGTA
CATGTTA

The alignment and score are dependent on the scoring system

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

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

Key point: Optimal alignment solutions and their scores are not necessarily unique and depend on the scoring system!

Alignment
CACTGT-A
CA-TGTTA

CACTG-TA
CA-TGTTA

CACTGTA
CATGTTA

NW DYNAMIC PROGRAMMING

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

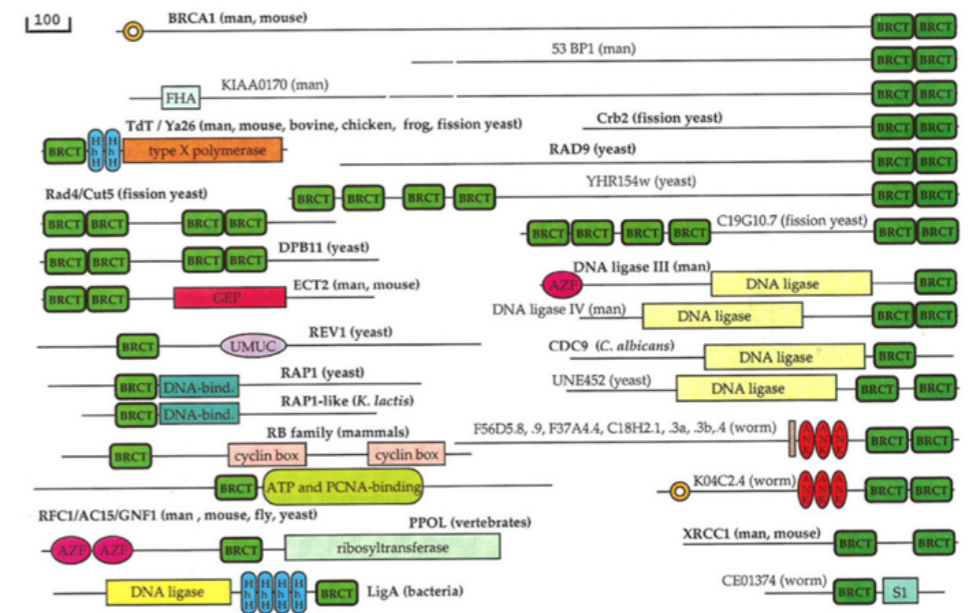
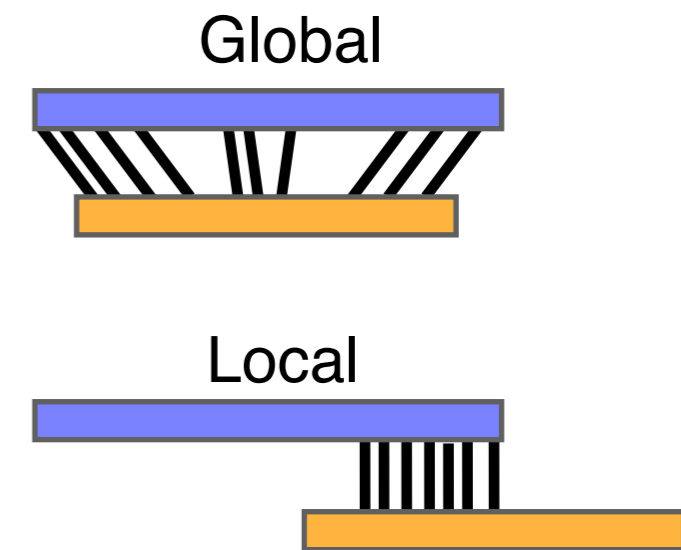
		A	G	T	T	C
	0					
A						
T						
T						
G						
C						

ALIGNMENT FOUNDATIONS

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

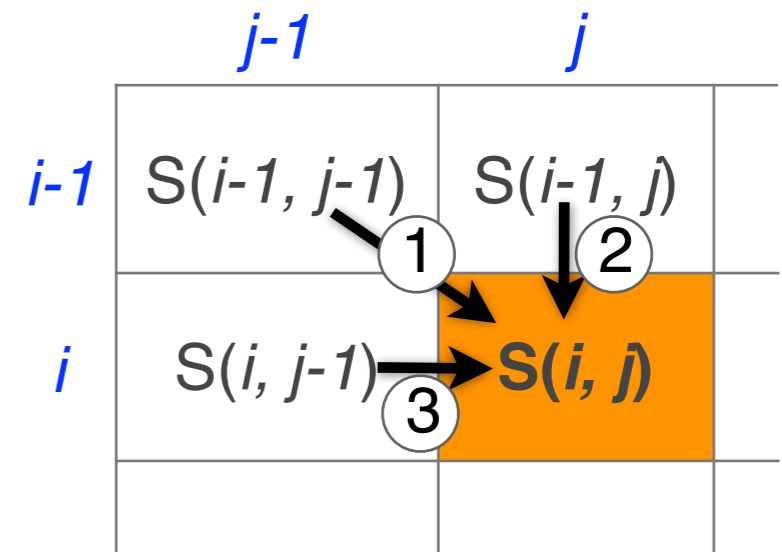
- Smith & Waterman proposed simply that a local alignment of two sequences allow arbitrary-length segments of each sequence to be aligned, with no penalty for the unaligned portions of the sequences. Otherwise, the score for a local alignment is calculated the same way as that for a global alignment

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

$$S(i, j) = \text{Max} \left\{ \begin{array}{l} S(i-1, j-1) + (\text{mis})\text{match} \quad \searrow \textcircled{1} \\ S(i-1, j) - \text{gap penalty} \quad \downarrow \textcircled{2} \\ S(i, j-1) - \text{gap penalty} \quad \rightarrow \textcircled{3} \\ 0 \quad \textcircled{4} \end{array} \right.$$



Sequence 1

- C A **G C C U C G** C U U A G

Sequence 2

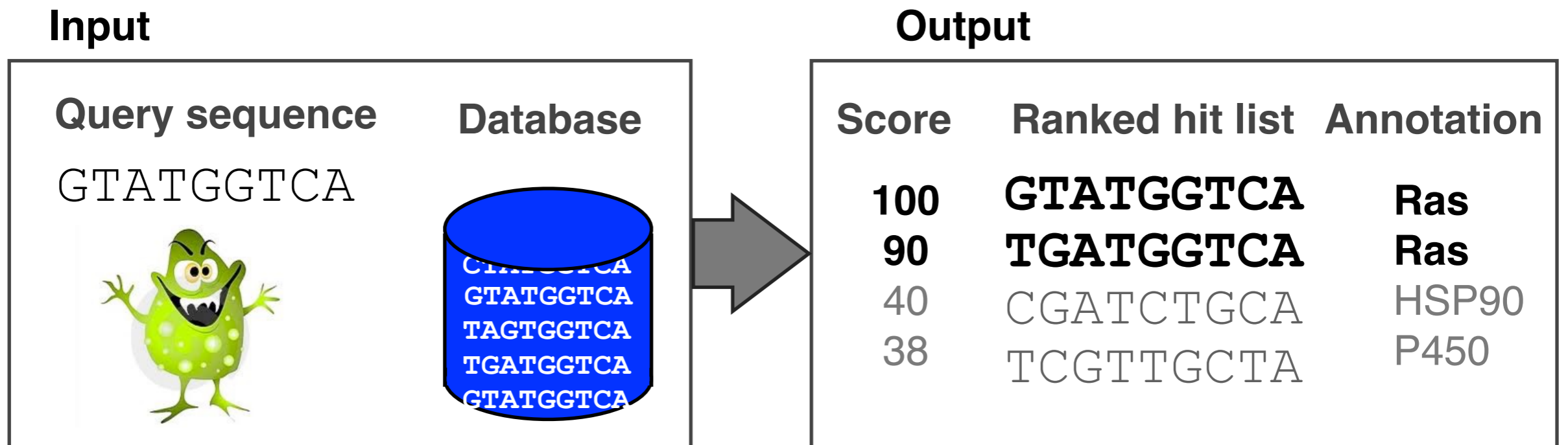
-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U	0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	0.0	1.0	0.0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

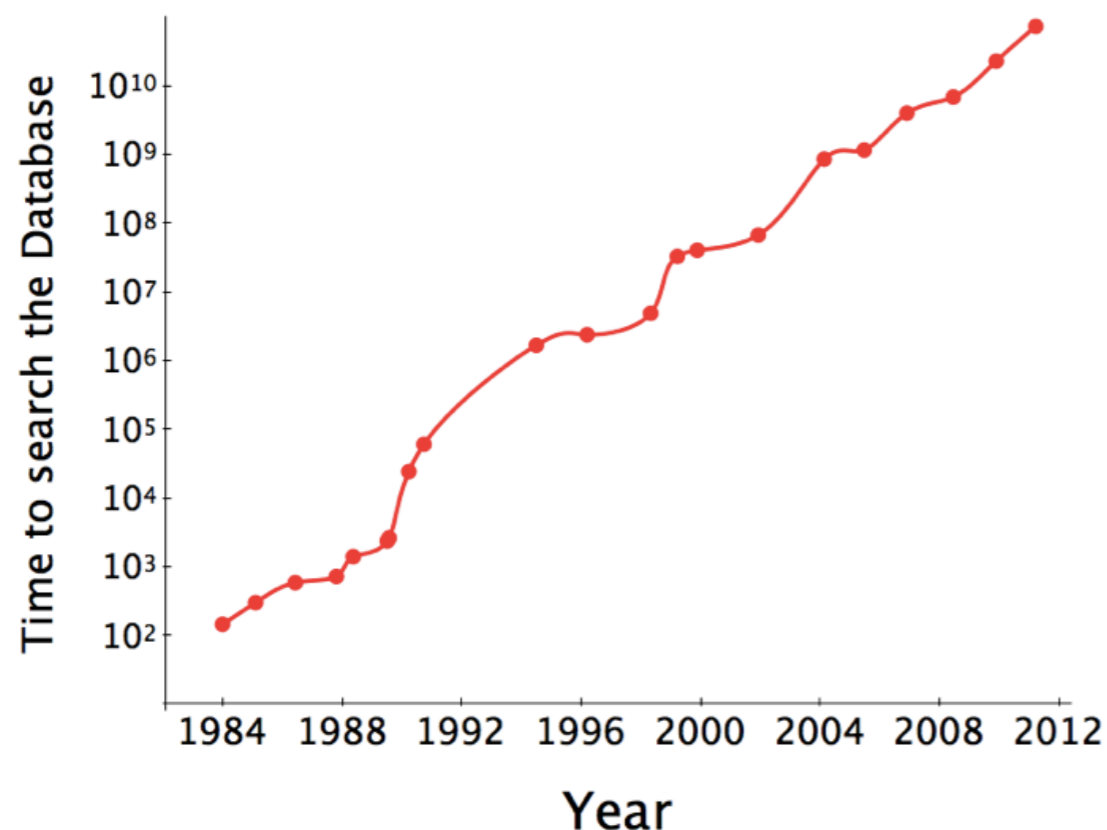
Local alignments can be used for database searching

- **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!**

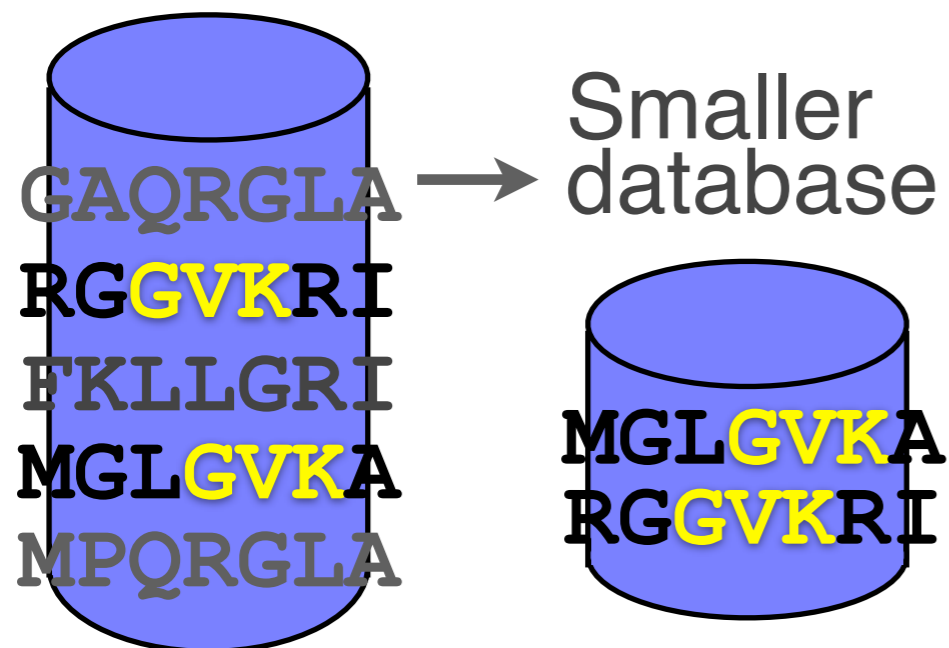


To reduce search time **heuristic algorithms**, such as BLAST, first remove database sequences without a strong local similarity to the query sequence in a quick initial scan.

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!**

Query **RGGVKRIKLMR**



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**

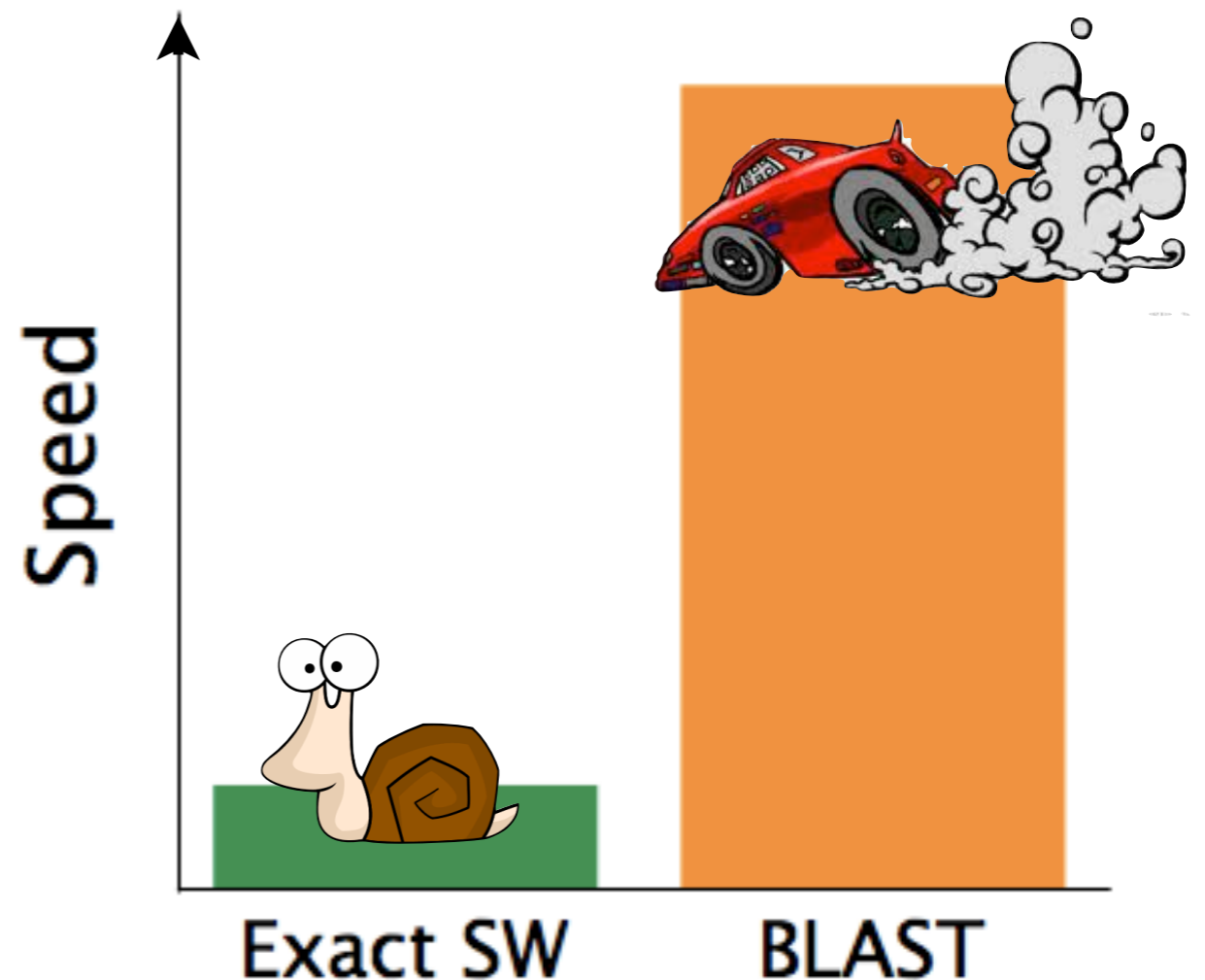
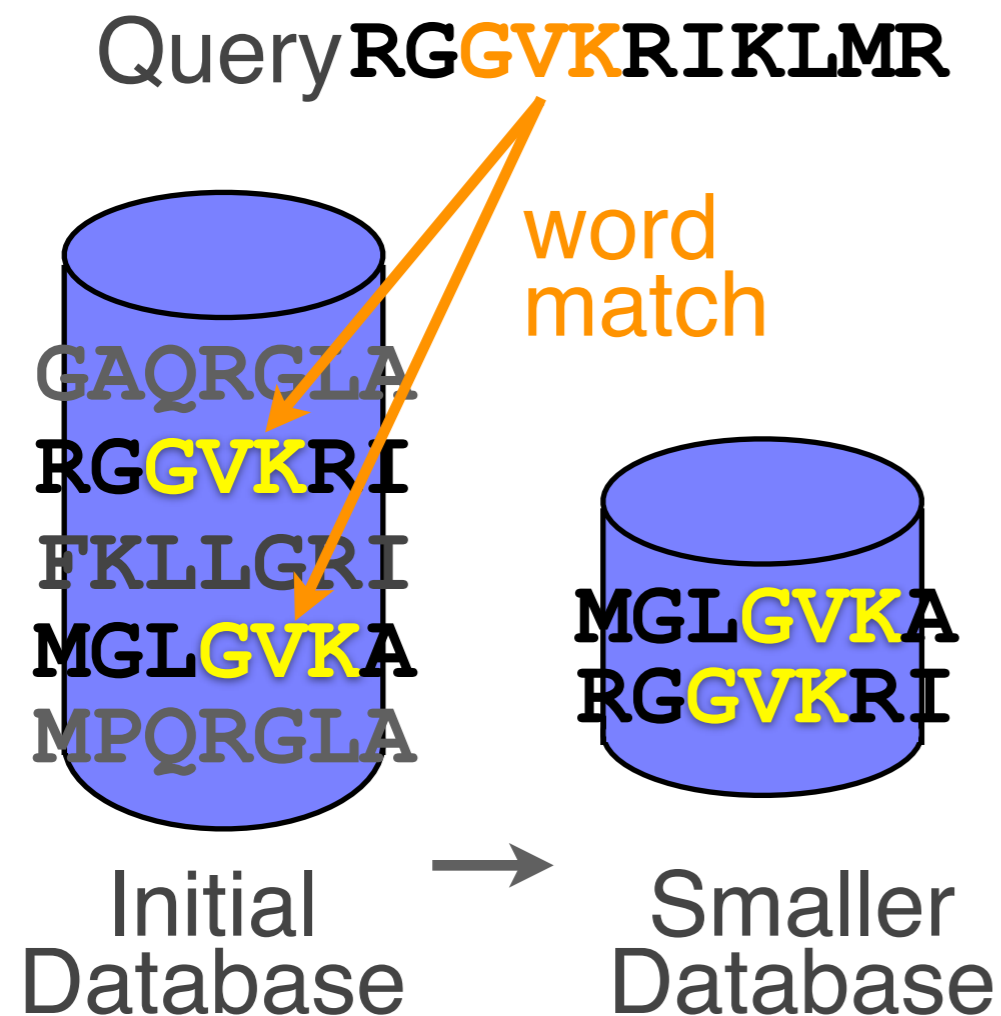
Rapid, heuristic versions of Smith–Waterman: **BLAST**

- 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 (SW) that is popular because it is **fast**
 - BLAST finds regions of high similarity between sequences
 - BLAST uses a heuristic search by scanning for initial **word pair** matches before performing 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)
- ...sensitivity in exchange for speed
- ...ast 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 ($w=3$)

RGGVKRI Query sequence

RGG

GGV

GVK

VKR

KRI

generate list
of $w=3$
words for
query

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)

RGGVKRI Query sequence

RRG RAG RIG RLG . . .

GGV GAV GTV GCV . . .

GVK GAK GIK GGK . . .

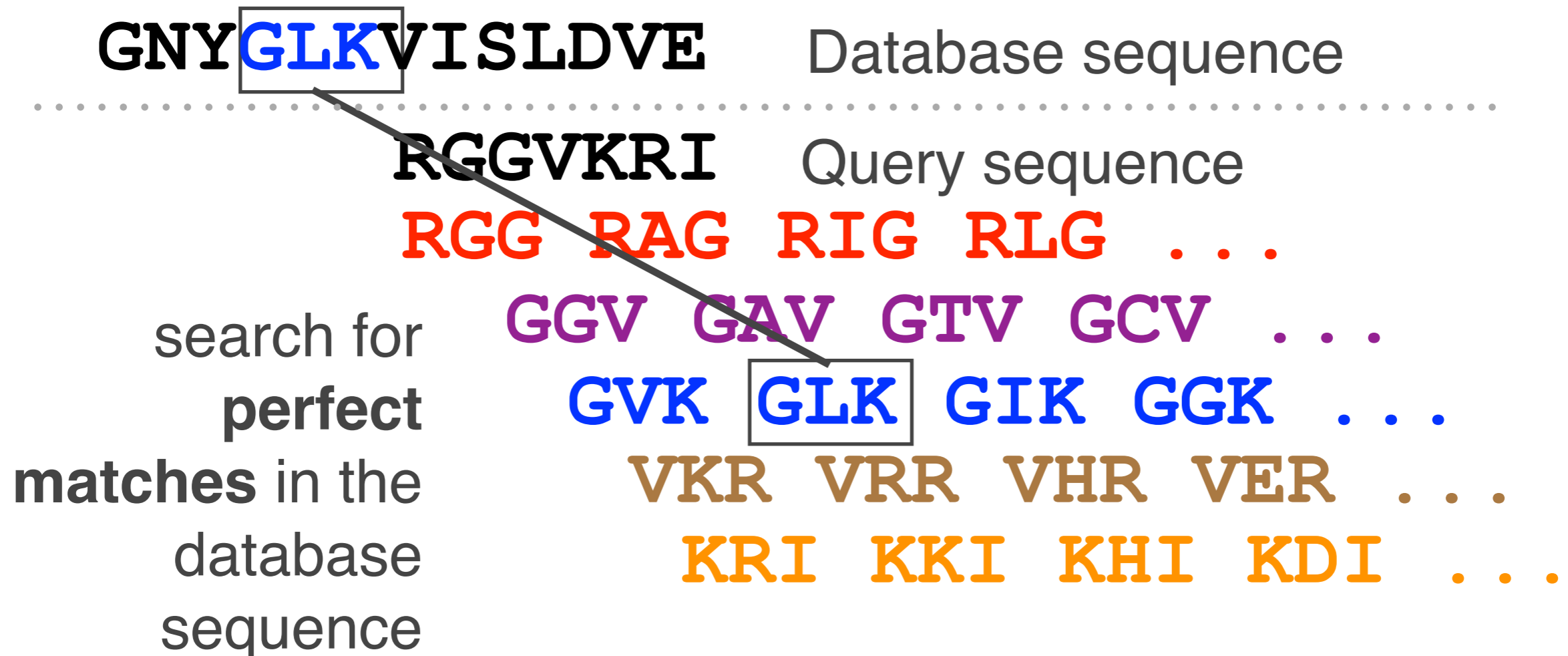
VKR VRR VHR VER . . .

KRI KKI KHI KDI . . .

extend list of
words similar
to query

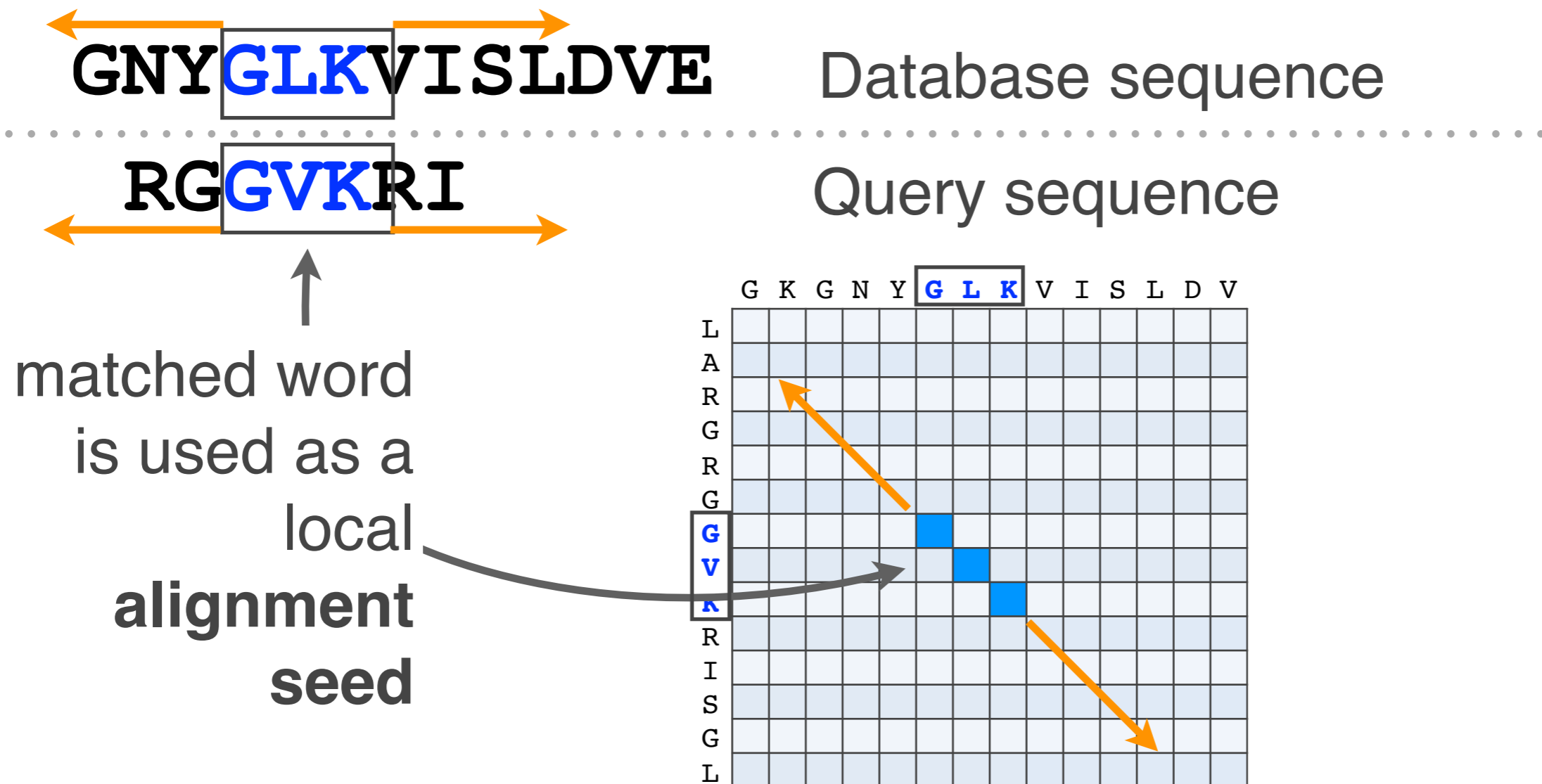
Blast

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

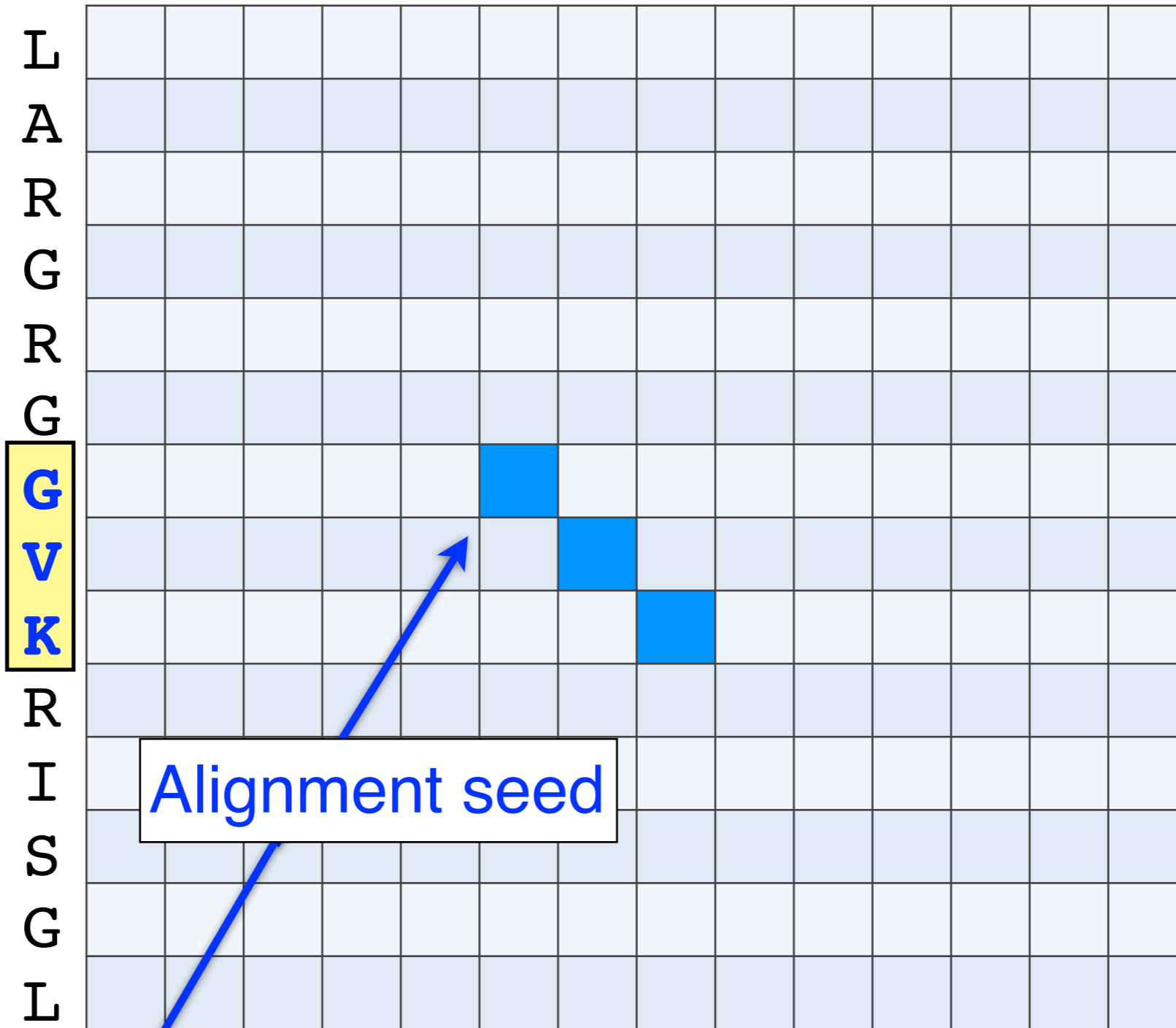


Blast

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



G K G N Y **G L K** V I S L D V



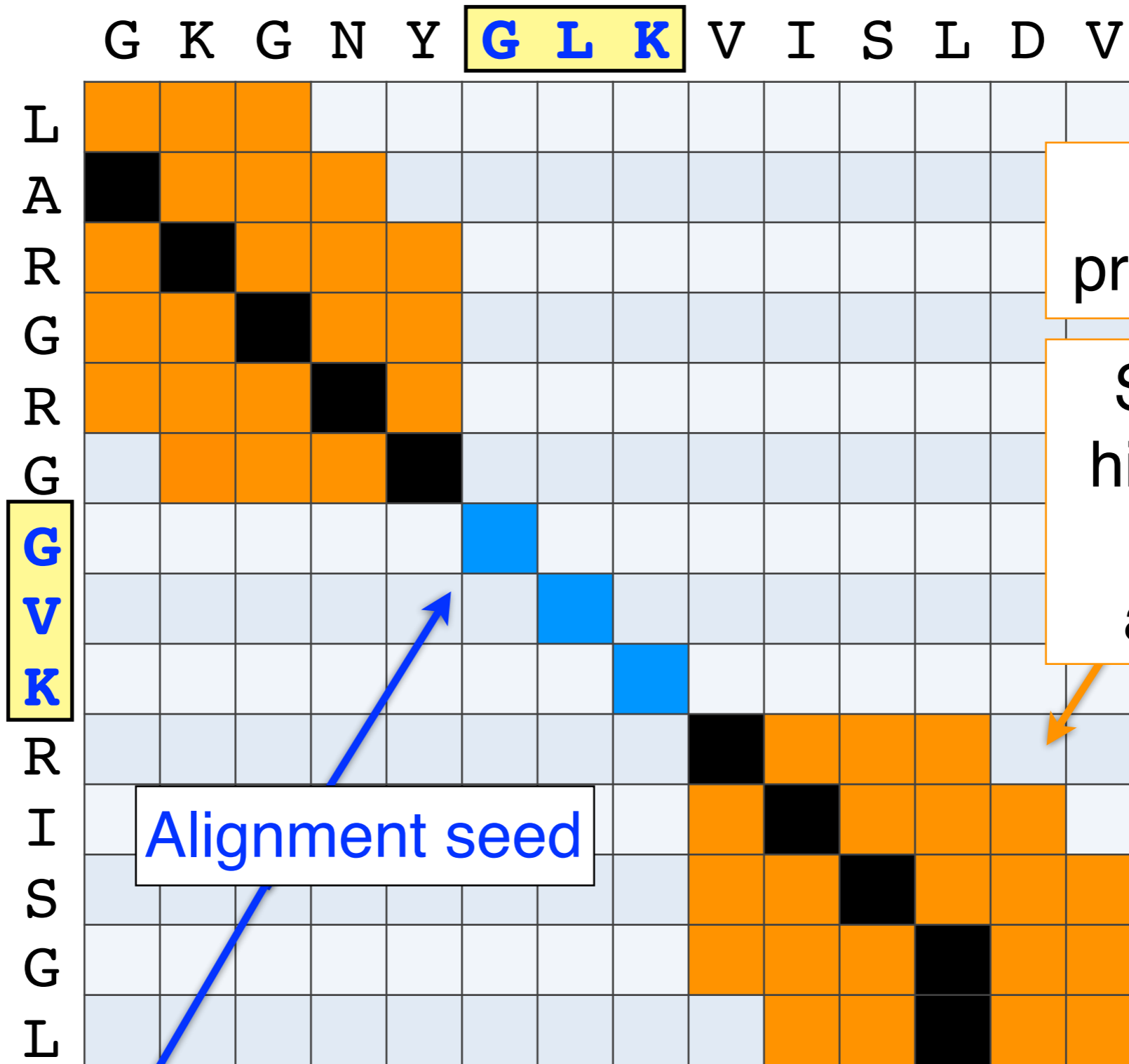
Alignment seed

GRGGVKRISGL

Query sequence

GNY**GLK**VISLDV

Database sequence



dynamic programming

Search for high scoring gapped alignment

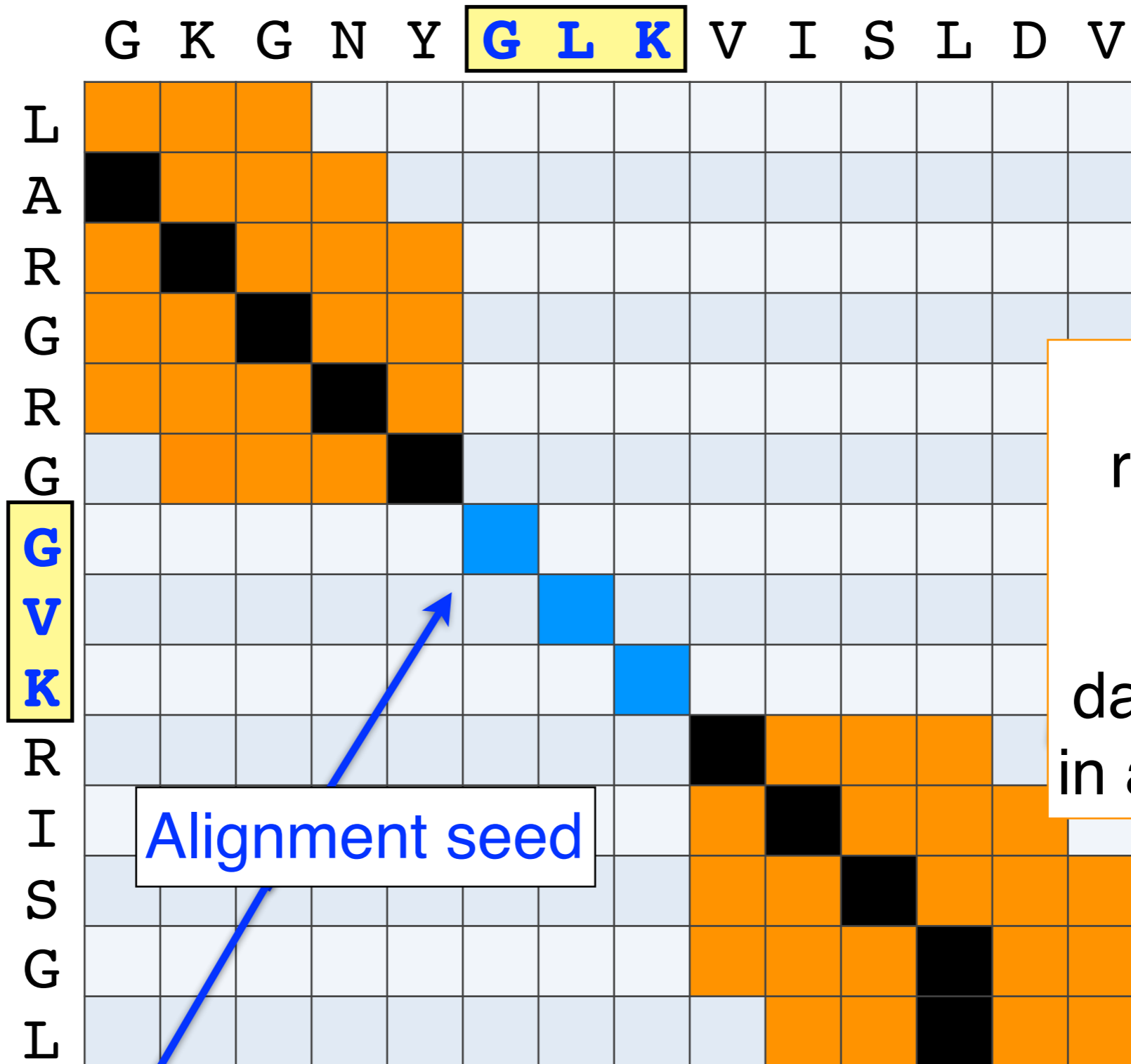
Alignment seed

← **GRGGV** KRISGL →

Query sequence

GNYGLK VISLDV

Database sequence



BLAST returns the highest scoring database hits in a ranked list

Alignment seed



Query sequence

Database sequence

BLAST output

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

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

Statistical significance of results

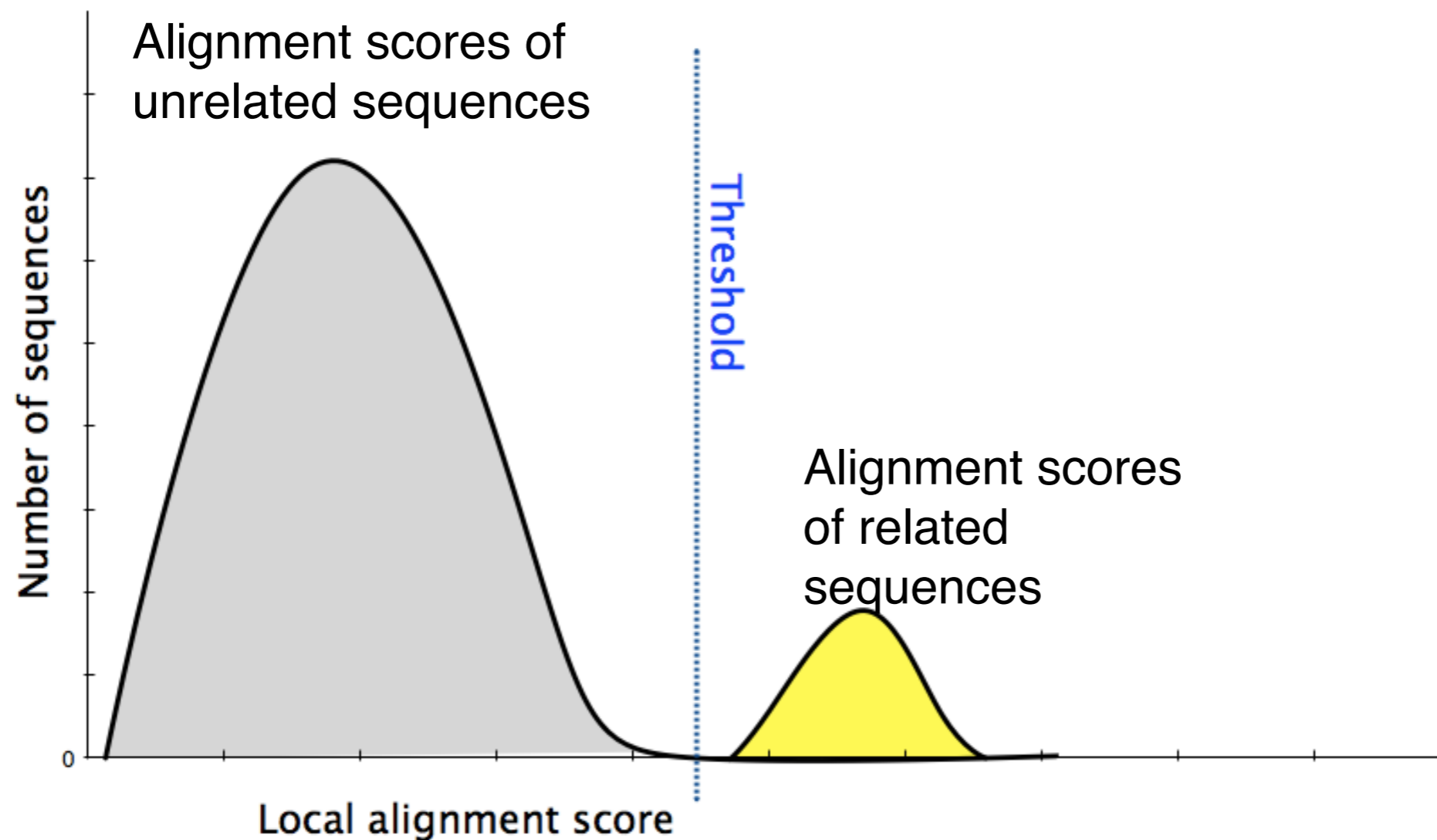
- An important feature of BLAST is the computation of statistical significance for each hit. This is described by the **E value** (expect value)

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

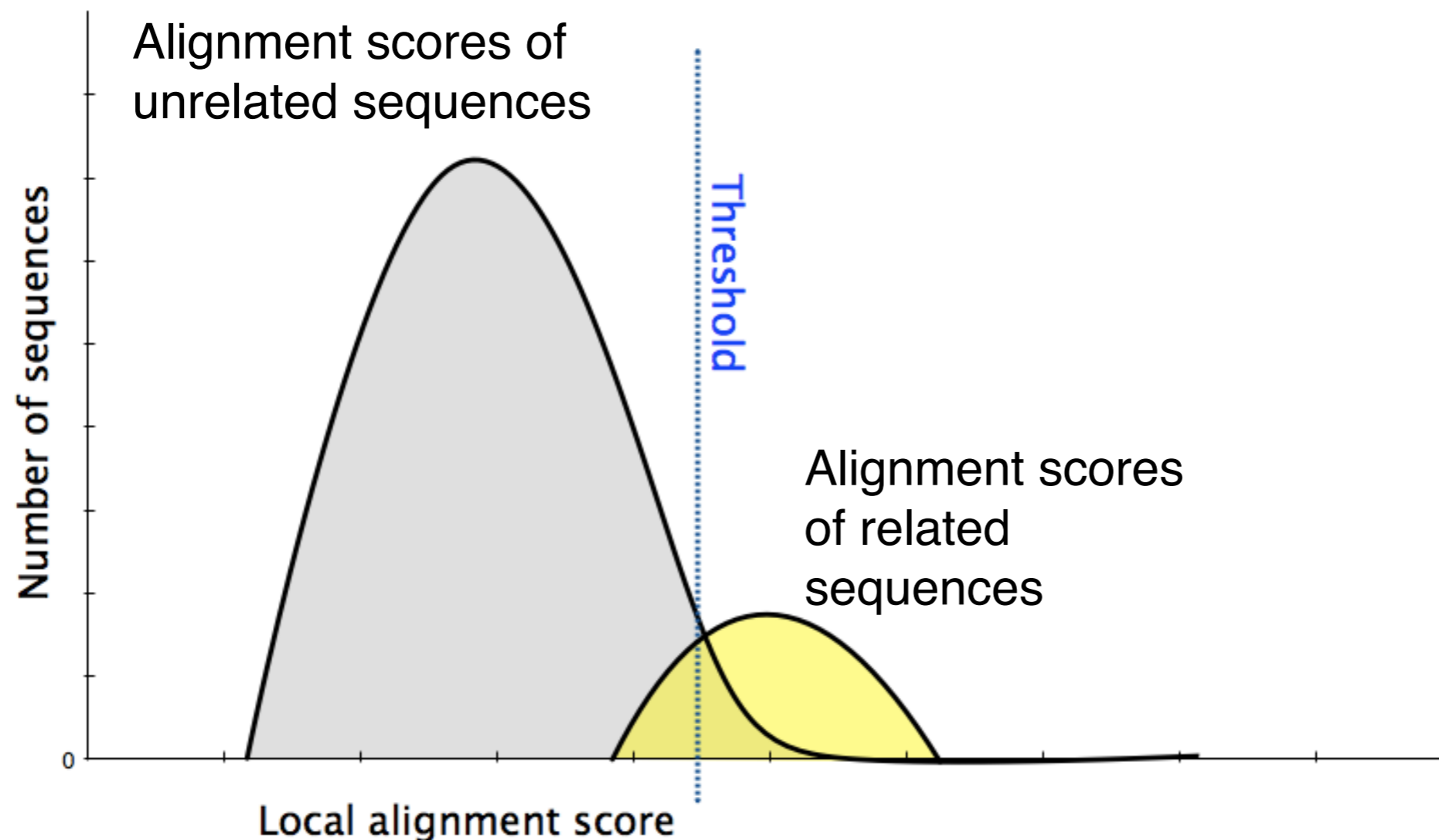
BLAST scores and E-values

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

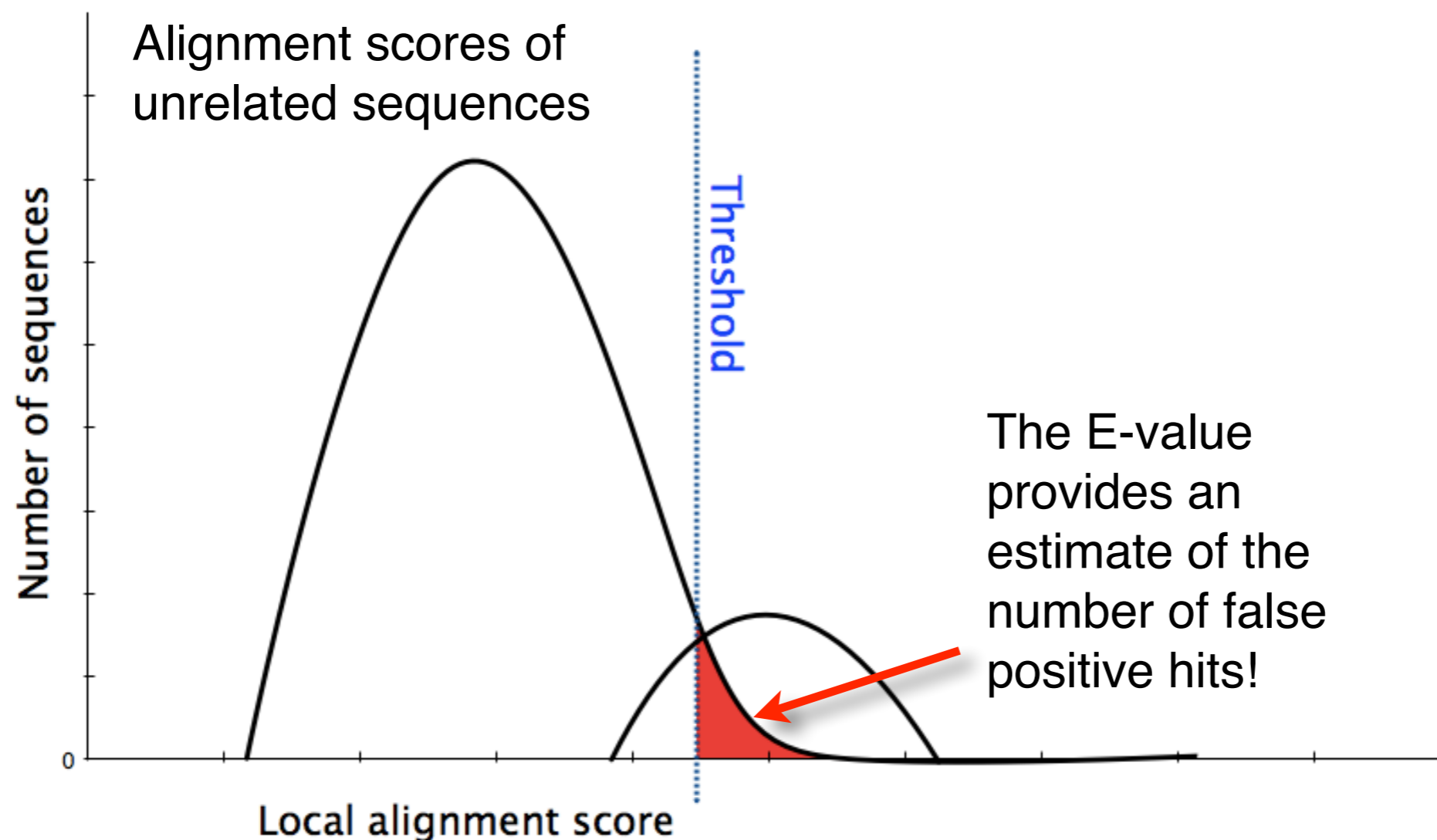
- Ideally, a threshold separates all query related sequences (yellow) from all unrelated sequences (gray)



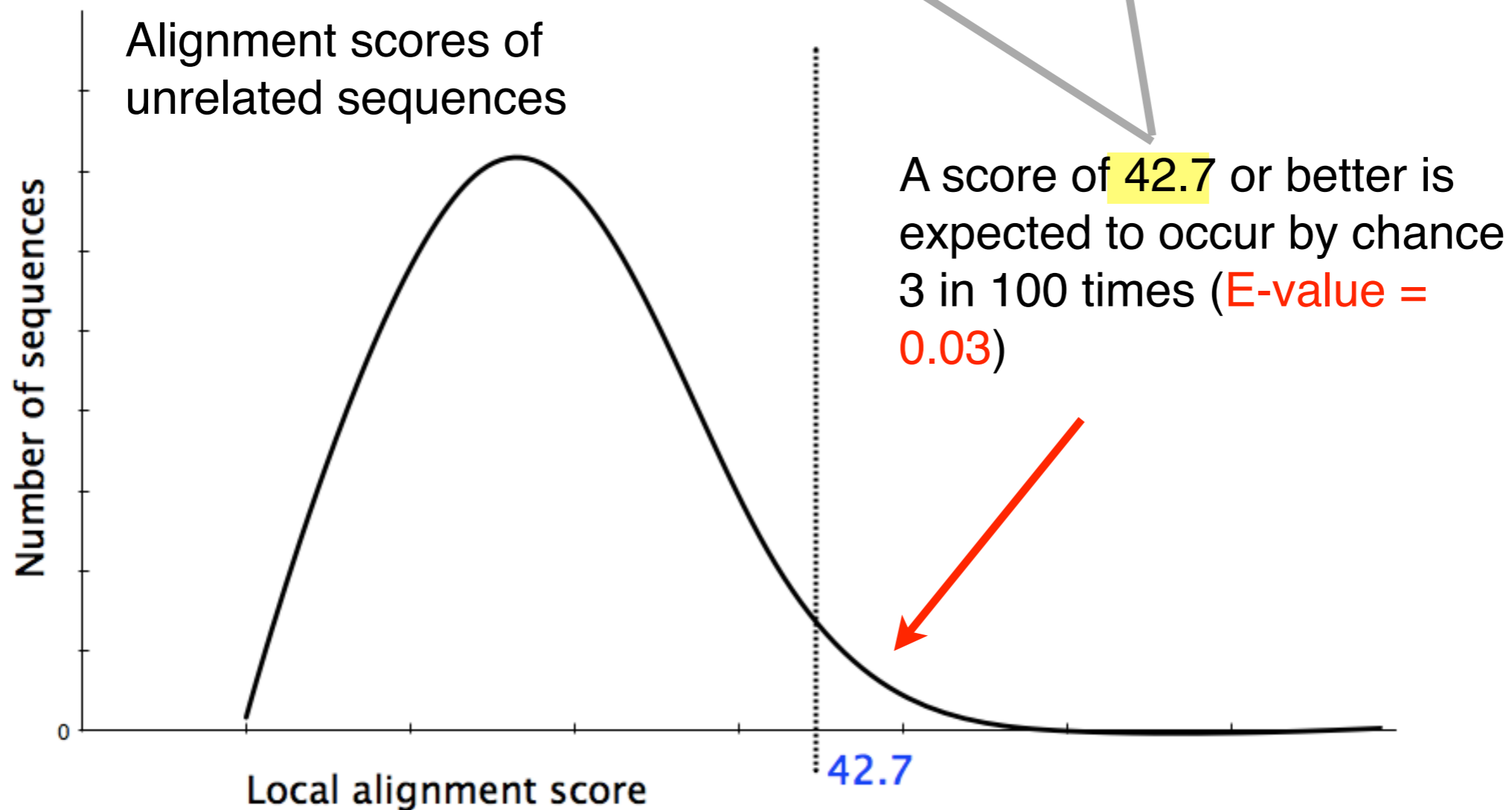
- Unfortunately, often both score distributions overlap
 - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



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



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

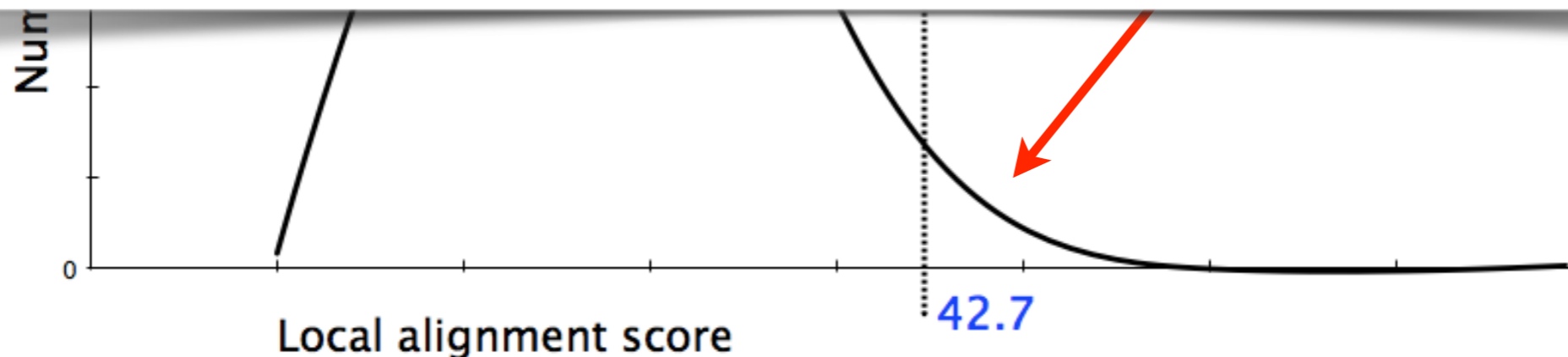


Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo	677	677	100%	0	100%	NP_004512.1
Kif5h protein [Mus musculus]	676	676	100%	0	98%	AAA20133.1

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 image shows a screenshot of the NCBI BLAST Home Page. The page has a blue header with the BLAST logo and the text "Basic Local Alignment Search Tool". Navigation tabs include "Home", "Recent Results", "Saved Strategies", and "Help". In the top right corner, there are links for "My NCBI", "[Sign In]", and "[Register]".

The main content area includes a breadcrumb "NCBI/ BLAST Home" and a descriptive sentence: "BLAST finds regions of similarity between biological sequences. [more...](#)". Below this is a "News" section with a link to "New WGS BLAST page".

A central overlay box with a grey background contains the text: "NCBI BLAST Home Page" and the URL "<http://blast.ncbi.nlm.nih.gov/Blast.cgi>".

Below the overlay, the page is divided into sections:

- BLAST Assembled RefSeq Ge**: "Choose a species genome to search, or" followed by a list of species: [Human](#), [Mouse](#), [Rat](#), and [Arabidopsis thaliana](#).
- Basic BLAST**: "Choose a BLAST program to run." followed by a list of programs:
 - [nucleotide blast](#): Search a **nucleotide** database using a **nucleotide** query. Algorithms: *blastn, megablast, discontinuous 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**

On the right side, there is a "How to do Batch BLAST jobs." section with a link to "More tips..." and a brief description: "BLAST makes it easy to examine a large group of potential gene candidates."

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

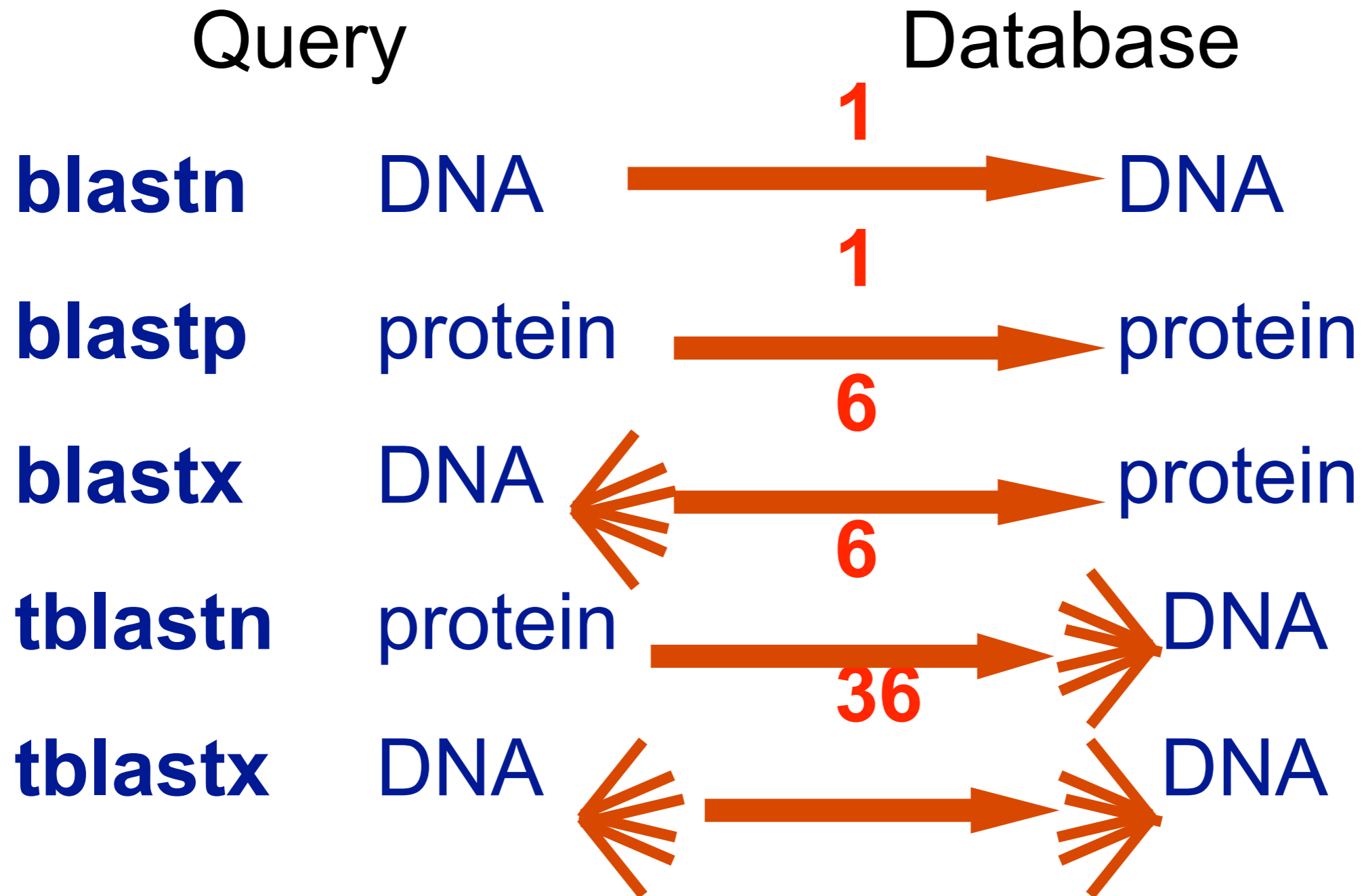
- Sequence can be input in FASTA format or as accession number

The screenshot shows the NCBI Protein search interface. At the top, there is a navigation bar with "NCBI", "Resources", and "How To". Below this, the "Protein" section is visible with the tagline "Translations of Life". A search bar contains the word "Protein" and a dropdown menu. To the right of the search bar are links for "Limits", "Advanced search", and "Help". Below the search bar is a "Search" button and a "Clear" button. On the left side, there is a "Display Settings" link and a dropdown menu showing "FASTA" selected. On the right side, there is a "Send to:" link and a "Change region shown" button. The main content area displays the search results for "hemoglobin subunit beta [Homo sapiens]". Below the title, it shows the "NCBI Reference Sequence" as "NP_000509.1". There are links for "GenPept" and "Graphics". The FASTA format sequence is displayed as follows:

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVEVGGGEALGRLLEVYPWTQRFFESFGDLSTPDVAVMGNPKVKAHGKKVLG
AFSDGLAHLAHLNKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHFGKEFTPPVQAAAYQKVVAVAN
ALAHKYH
```

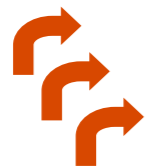
On the right side, there is a section titled "Analyze this sequence" with links for "Run BLAST", "Identify Conserved Domains", and "Find in this Sequence".

Step 2: Choose the BLAST program



DNA potentially encodes six proteins

5' CAT CAA
5' ATC AAC
5' TCA ACT



5' CATCAACTACAACCTCCAAAGACACCCTTACACATCAACAAACCTACCCAC 3'
3' GTAGTTGATGTTGAGGTTTCTGTGGGAATGTGTAGTTGTTTGGATGGGTG 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&PAC

Enter Query Sequence

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

Query subrange [?](#)

From

To

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGK
KVLGAFSDGLAHL DNLKGT FATLSELHCDKLHVDPENFRLGNVLCVLAH HFGKEFTPPVQAA YQK
VVAGVANALAHKYH
```

Or, upload file no file selected [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database [?](#)

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)

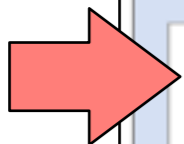
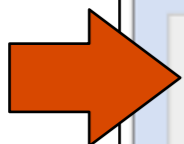
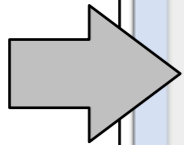
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

Search database **Non-redundant protein sequences (nr)** using **Blastp (protein-protein BLAST)**

Show results in a new window

[+ Algorithm parameters](#)



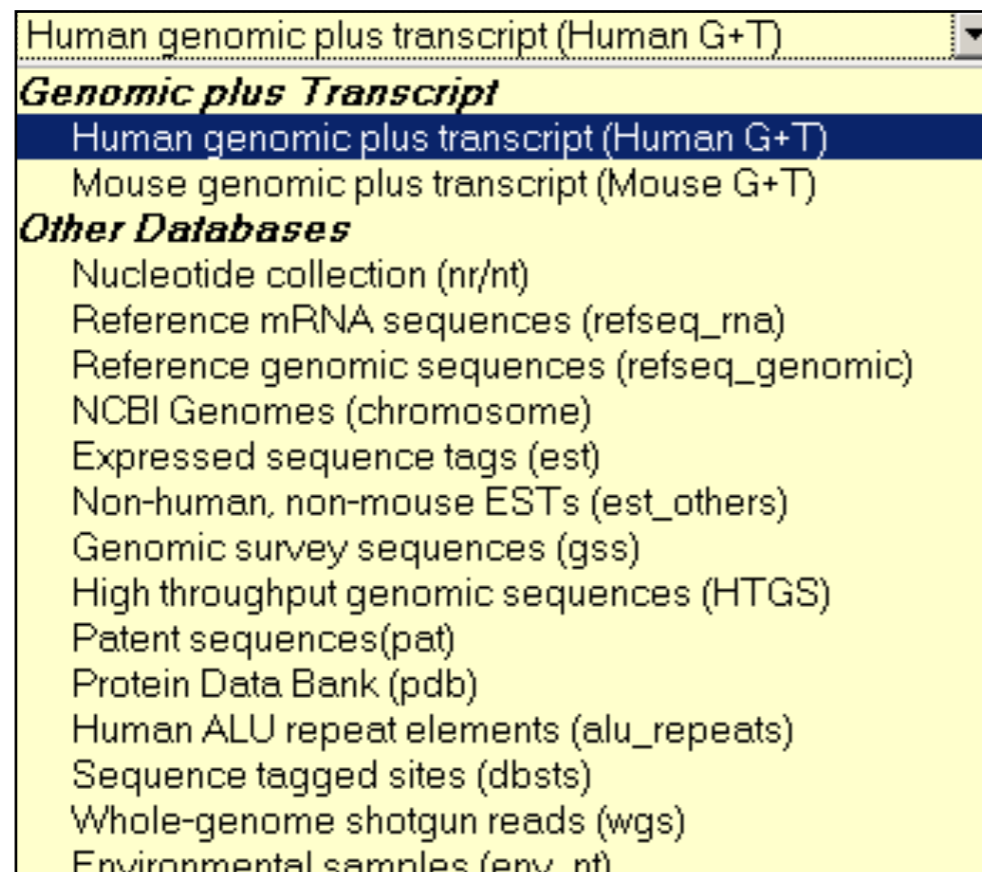
Step 3: Choose the database

nr = non-redundant (most general database)

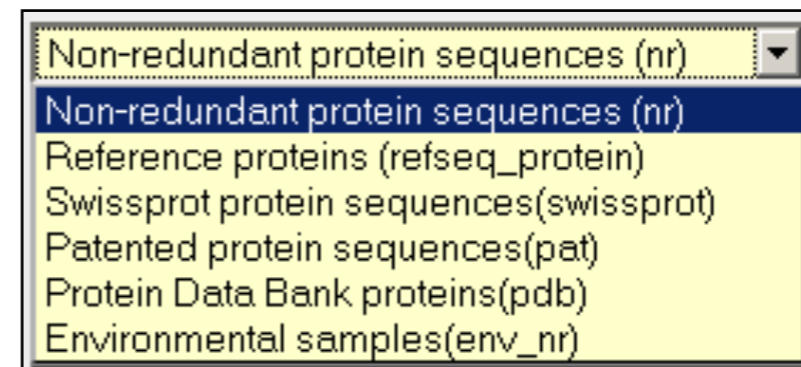
dbest = database of expressed sequence tags

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&PAC

Enter Query Sequence

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

Query subrange [Query subrange](#)

From

To

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVPWTQRFESFGDLSTPDAVMGNPKVKAHGK
KVLGAFSDGLAHLNLDLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQK
VVAGVANALAHKYH
```

Or, upload file no file selected

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

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

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	?
Short queries	<input checked="" type="checkbox"/> 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	<input type="checkbox"/> Low complexity regions	?
Mask	<input type="checkbox"/> Mask for lookup table only	?
	<input type="checkbox"/> Mask lower case letters	?

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

parameters

Expect
Word size

Scoring matrix

Step 4: Optional parameters

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

Results page

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

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

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/BLAST/blastp suite/ Formatting Results - FVGUTMRZ013

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#) [Change the result display back to traditional format](#)

[YouTube](#) [Learn about the enhanced report](#) [Blast report description](#)

gi|4504349|ref|NP_000509.1| hemoglobin

Query ID	lcl 84677	Database Name	nr
Description	gi 4504349 ref NP_000509.1 hemoglobin subunit beta [Homo sapiens]	Description	All non-redundant GenBank CDS 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

Graphic Summary

Show Conserved Domains

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

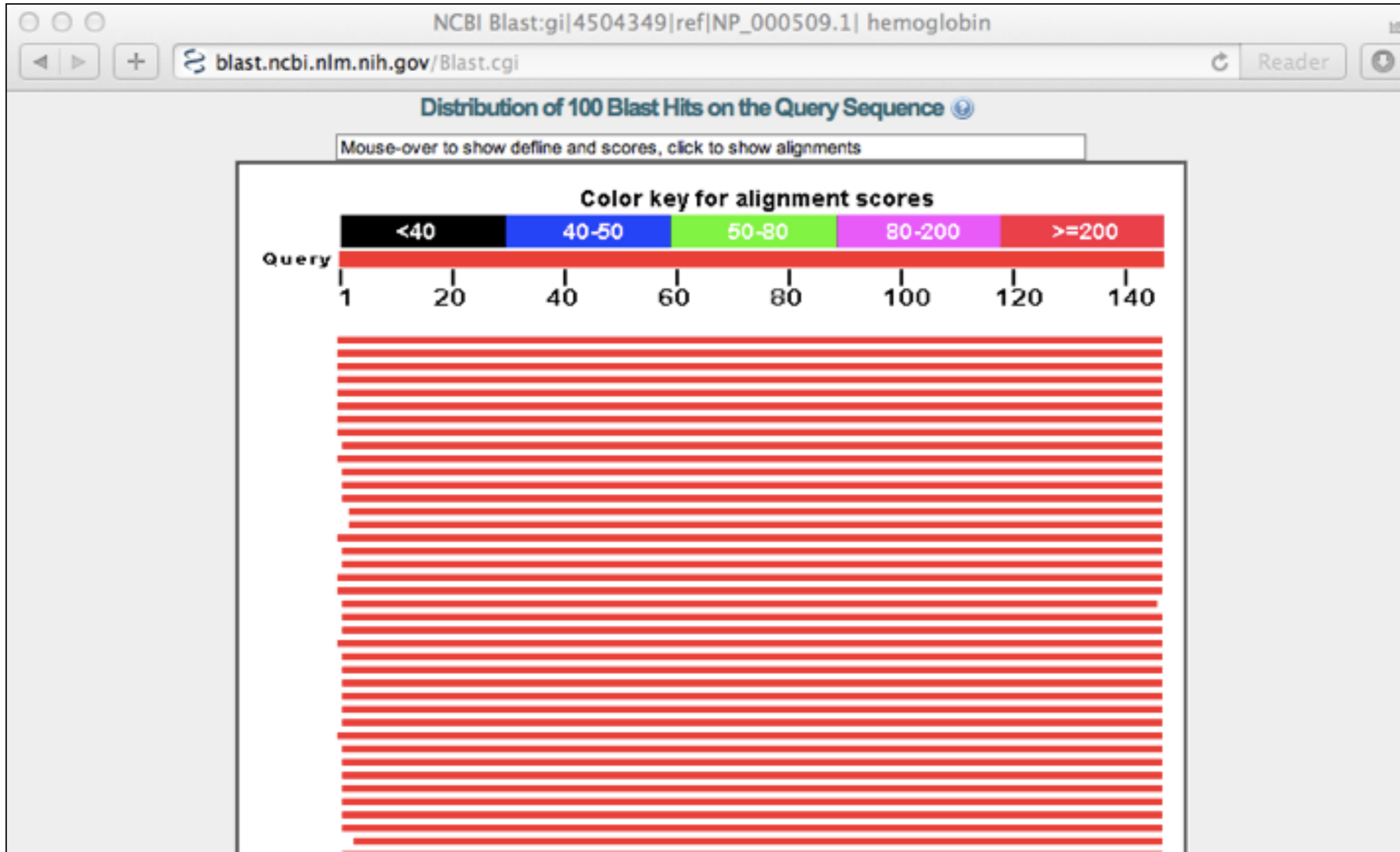
Query seq. 1 25 50 75 100 125 147

Specific hits: heme-binding site, globin

Superfamilies: globin_like superfamily

Distribution of 100 Blast Hits on the Query Sequence

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

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=Hen	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 AAZ3978	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	298	298	99%	9e-102	100%	1COH_B
<input type="checkbox"/>	hemoglobin beta subunit variant [Homo sapiens] >gb AAA88054.1 beta-globin [Homo sa	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 Syn	297	297	99%	3e-101	99%	1DXU_B
<input type="checkbox"/>	Chain B, Analysis Of The Crystal Structure, Molecular Modeling And Infrared Spectroscop	297	297	99%	3e-101	99%	1HDB_B

Further down the results page...

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

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

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		MVHLTPEEKSAVTALWGKVNVDDEVGGEALGRLLEVYPWTQRFFESFGDLSTPDAVMGNPK		60	
Sbjct 1		MVHLTPEEKSAVTALWGKVNVDDEVGGEALGRLLEVYPWTQRFFESFGDLSTPDAVMGNPK		60	
Query 61		VKAHGKKVLGAFSDGLAHLAHDNLKGTFFATLSELHCDKLHVDPENFRLLGNVLCVLAHHFG		120	
Sbjct 61		VKAHGKKVLGAFSDGLAHLAHDNLKGTFFATLSELHCDKLHVDPENFRLLGNVLCVLAHHFG		120	
Query 121		KEFTPPVQAAYQKVVAGVANALAHKYH	147		
Sbjct 121		KEFTPPVQAAYQKVVAGVANALAHKYH	147		

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

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

Different output formats are available

The screenshot shows the NCBI BLAST web interface. The browser address bar displays `blast.ncbi.nlm.nih.gov/Blast.cgi`. The page title is "NCBI Blast:gi|4504349|ref|NP_000509.1| hemoglobin". The BLAST logo and "Basic Local Alignment Search Tool" are visible. Navigation links include "Home", "Recent Results", "Saved Strategies", and "Help". A "My NCBI" section contains "[Sign In]" and "[Register]".

The main content area shows the "Formatting options" menu circled in red. Below the menu, the "Formatting options" panel is expanded, showing various settings:

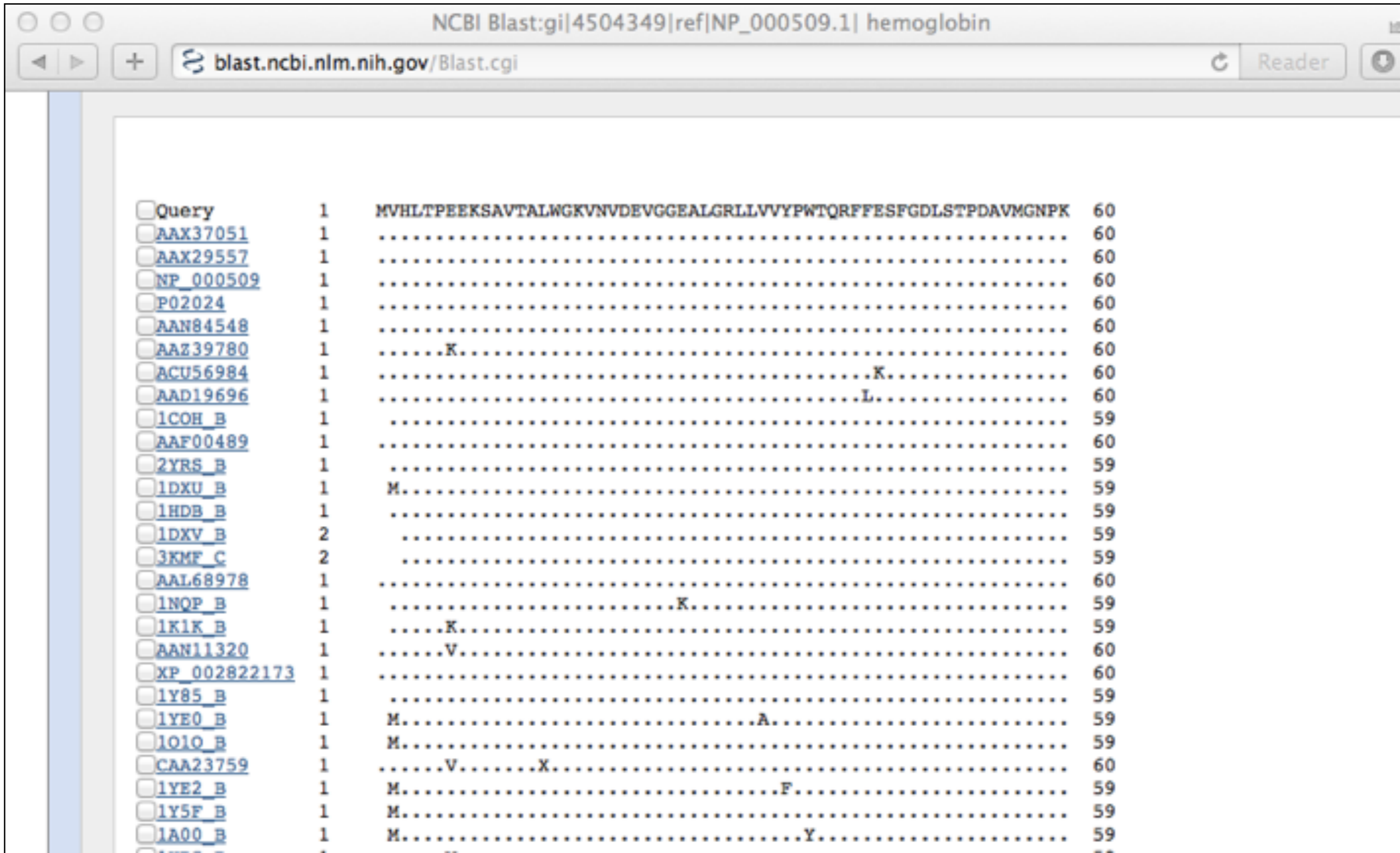
- Show**: Alignment as **HTML** (dropdown), Old View, [Reset form to defaults](#)
- Alignment View**: **Query-anchored with letters for identities** (dropdown)
- Display**: Graphical Overview, Sequence Retrieval, NCBI-gi
- Masking**: Character: **Lower Case** (dropdown), Color: **Grey** (dropdown)
- Limit results**: Descriptions: **50** (dropdown), Graphical overview: **50** (dropdown), Alignments: **50** (dropdown)
- 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:

At the bottom left, the query sequence is shown: `gi|4504349|ref|NP_000509.1| hemoglobin`.

E.g. Query anchored alignments

Sequence ID	Position	Sequence	Length
<input type="checkbox"/> Query	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> AAX37051	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> AAX29557	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> NP_000509	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> P02024	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> AAN84548	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> AAZ39780	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> ACU56984	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> AAD19696	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> 1COH_B	1	VHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> AAF00489	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> 2YRS_B	1	VHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 1DXU_B	1	MHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 1HDB_B	1	VHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 1DXV_B	2	HLPTEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 3KMF_C	2	HLPTEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> AAL68978	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> 1NQP_B	1	VHLTPEEKSAVTALWGKVNVDVGGKALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 1K1K_B	1	VHLTPPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> AAN11320	1	MVHLTPVEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> XP_002822173	1	MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> 1Y85_B	1	VHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 1YE0_B	1	MHLTPEEKSAVTALWGKVNVDVGGGEALGRLLAVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 1O1O_B	1	MHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> CAA23759	1	MVHLTPVEEKSAVTAXWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<input type="checkbox"/> 1YE2_B	1	MHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVFPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 1Y5F_B	1	MHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 1A00_B	1	MHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPYTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 1HBS_B	1	VHLTPVEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 1ABY_B	1	MHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
<input type="checkbox"/> 1CMY_B	1	VHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59

... and alignments with dots for identities



Accession	Length	Sequence	Score
Query	1	MVHLTPEEKSAVTALWGKVNVDVEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAX37051	1	60
AAX29557	1	60
NP_000509	1	60
P02024	1	60
AAN84548	1	60
AAZ39780	1K.....	60
ACU56984	1K.....	60
AAD19696	1L.....	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
1YEQ_B	1	M.....A.....	59
1O1O_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**.

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>i.e.</i> filled in)	1	
Evidence for correct use of scoring scheme	1	
Direction arrows drawn between all cells	1	
Evidence of multiple arrows to a given cell if appropriate	1	D
Correct optimal score position in matrix used	1	C
Correct optimal score obtained for given scoring scheme	1	B
Traceback path(s) clearly highlighted	1	A
Correct alignment(s) yielding optimal score listed	1	A+