

# 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/>

NCBI National Center for Biotech x www.ncbi.nlm.nih.gov Resources How To Sign in to NCBI All Databases ras Search

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

Welcome to NCBI

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

About the NCBI | Mission | Organization | NCBI News

Get Started

- [Data](#): Search and analyze data using NCBI software
- [Tools](#): Get NCBI data or software
- [How Tos](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

Genotypes and Phenotypes

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

Resources

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

NCBI Announcements

RefSeq release 69 available on

The full RefSeq release 69 is now available on the FTP site with 74 records describing 50,276,469 ..

Hands on demo (or see following slides)

# Example Vignette Questions:

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

ras - GQuery: Global Cross X

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

NCBI Resources How To Sign in to NCBI

## Search NCBI databases

Help

ras

Search

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

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

8

ras - Gene - NCBI

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

NCBI Resources How To Sign in to NCBI

Gene Gene ras Search Save search Advanced Help

Show additional filters Hide sidebar >>

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

Filters: Manage Filters

Did you mean ras as a gene symbol?  
Search Gene for [ras](#) as a symbol.

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

**Results: 1 to 20 of 85633**

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

**Top Organisms [Tree]**

- Homo sapiens (1126)
- Mus musculus (823)
- Rattus norvegicus (625)
- Oreochromis niloticus (533)
- Neolamprologus brichardi (507)
- All other taxa (82019)

More...

**Find related data**

Database: Select

Find items

**Search details**

ras [All Fields] AND alive [property]

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

(ras) AND "Homo sapiens" | [X](#)

www.ncbi.nlm.nih.gov/gene

NCBI Resources How To Sign in to NCBI

Gene Gene (ras) AND "Homo sapiens"[porgn:\_txid9606] Search Help

Show additional filters Hide sidebar >

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

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

i Filters activated: Current only. [Clear all](#) to show 1499 items.

Filters: [Manage Filters](#)

Find related data

Database: Select

Find items

Search details

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

Search See more...

Recent activity

Turn Off Clear

Name/Gene ID Description Location Aliases

<input type="checkbox"/> <a href="#">NRAS</a> ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [ <i>Homo sapiens</i> (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
<input type="checkbox"/> <a href="#">KRAS</a> ID: 3845	Kirsten rat sarcoma viral oncogene homolog [ <i>Homo sapiens</i> (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, NS2, RAKS2

Categories

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

Sequence content

- CCDS
- Ensembl
- RefSeq

Status

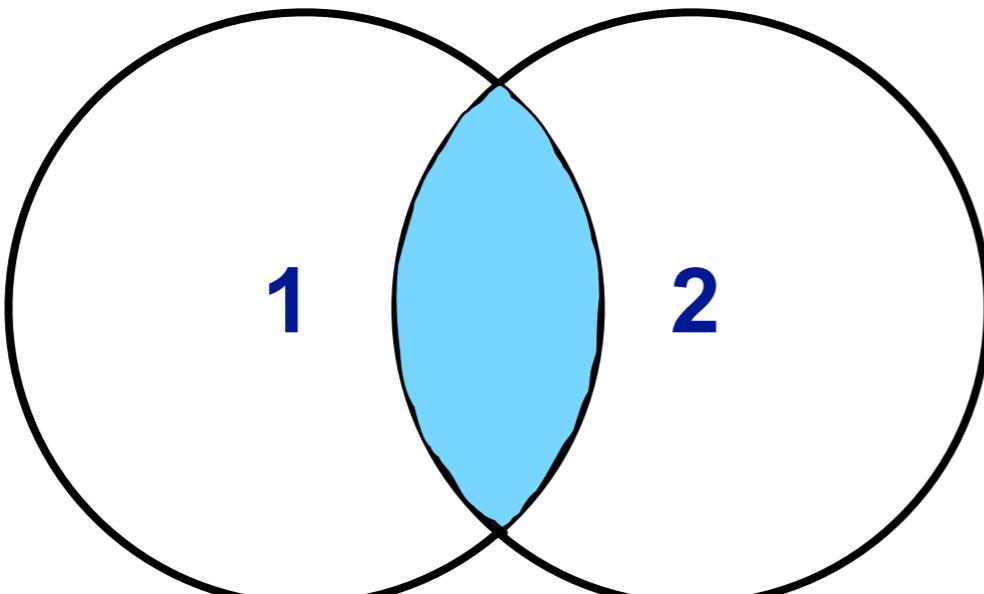
- clear
- Current only

Chromosome locations

Select

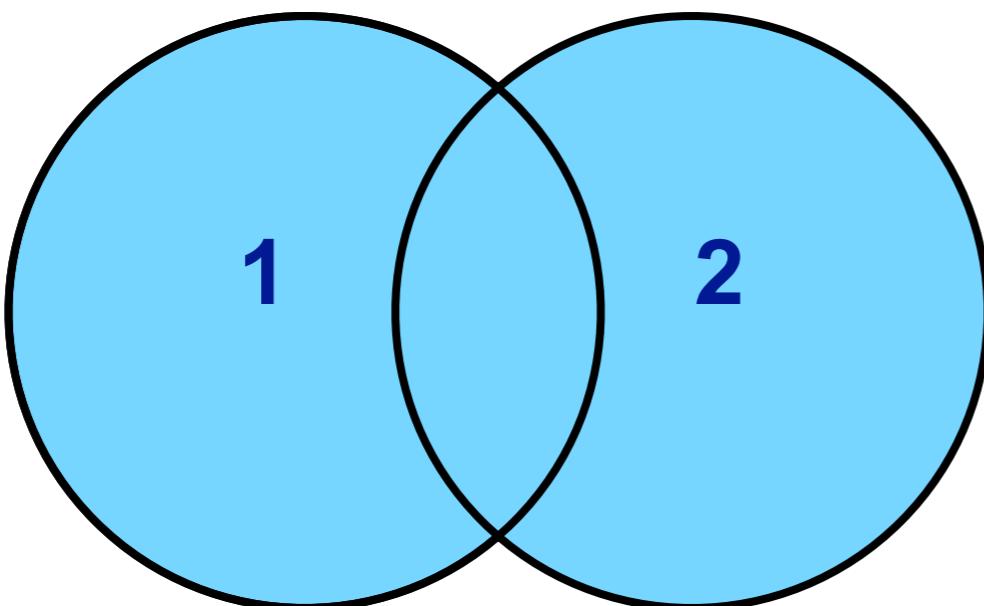
10

**1 AND 2**



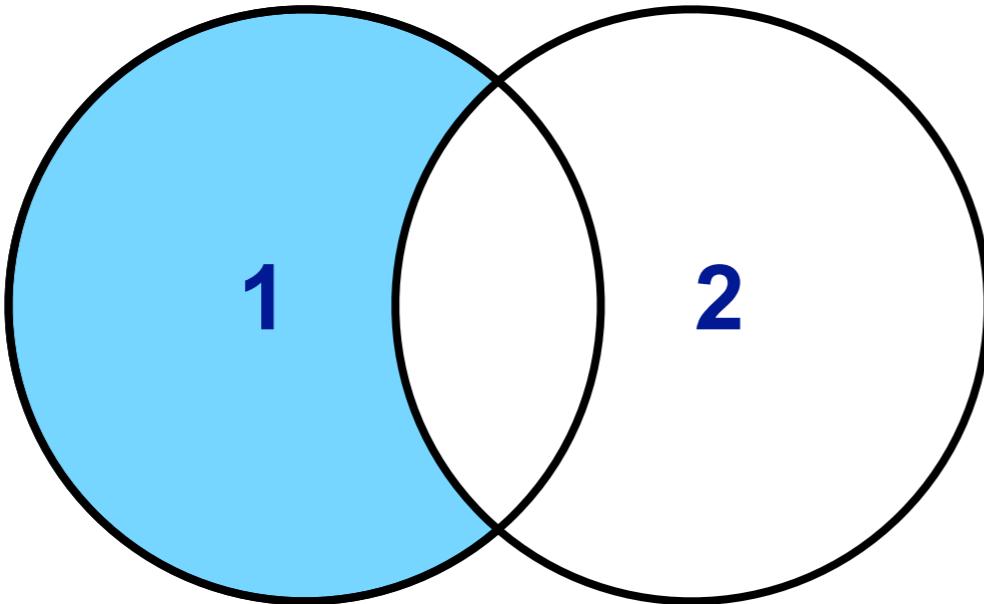
**ras AND disease  
(1185 results)**

**1 OR 2**



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

**1 NOT 2**



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

(ras) AND "Homo sapiens" | X

www.ncbi.nlm.nih.gov/gene

NCBI Resources How To Sign in to NCBI

Gene Gene (ras) AND "Homo sapiens"[porgn:\_txid9606] Search Save search Advanced Help

Show additional filters Hide sidebar >>

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

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

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

Gene sources Genomic Categories

Alternatively spliced Annotated genes Non-coding Protein-coding Pseudogene

Sequence content CCDS Ensembl RefSeq Status clear

✓ Current only Chromosome locations Select

Name/Gene ID Description Location Aliases

<input type="checkbox"/> NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
<input type="checkbox"/> KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, NS2, RAKS2

Filters: Manage Filters

Find related data

Database: Select

Find items

Search details

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

Search See more...

Recent activity Turn Off Clear

12

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

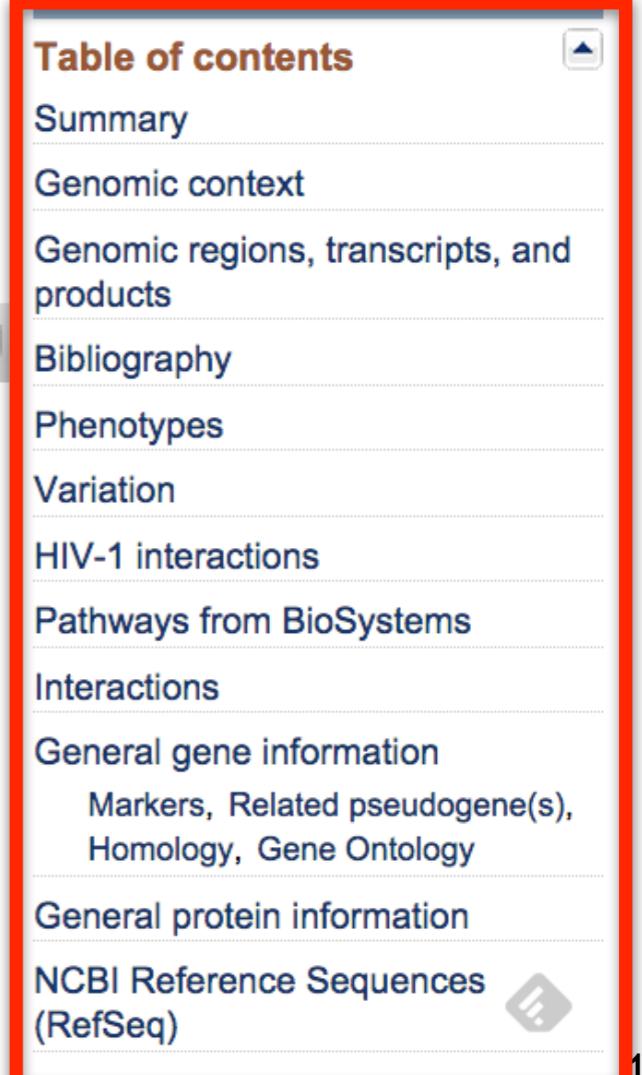
**Organism** [Homo sapiens](#)

**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 \(RefSeq\)](#)



KRAS Kirsten rat sarcoma

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

NCBI Resources How To Sign in to NCBI

Gene Search Help Hide sidebar >>

**Example Questions:**

What chromosome location and what genes are in the vicinity?

Display S KRAS (human) Gene ID: 3845, updated on 4-Jan-2015

**Summary**

**Official Symbol** KRAS provided by HGNC

**Official Full Name** Kirsten rat sarcoma viral oncogene homolog provided by HGNC

**Primary source** HGNC:HGNC:6407

**See related** Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193

**Gene type** protein coding

**RefSeq status** REVIEWED

**Organism** Homo sapiens

**Lineage** Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

**Also known as** NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Table of contents

Summary

**Genomic context**

Genomic regions, transcripts, and products

Bibliography

Phenotypes

Variation

HIV-1 interactions

Pathways from BioSystems

Interactions

General gene information

Markers, Related pseudogene(s), Homology, Gene Ontology

General protein information

NCBI Reference Sequences (RefSeq)

Related documents

14

KRAS Kirsten rat sarcoma

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

**Genomic context**

**Location:** 12p12.1      **Exon count:** 6

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

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

**Chromosome 12 - NC\_000012.12**

[ 25052101 ► ] [ 25436297 ► ]

LRMP → LYRM5 ← CASC1 ← KRAS → LOC100421617 ← RPL39P27

**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

KRAS Kirsten rat sarcoma

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

NCBI Resources How To Sign in to NCBI

Gene

Display Settings

KRAS Ki (human) ] Gene ID: 3845

Summary

**Example Questions:**

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

**Official Symbol** KRAS provided by HGNC

**Official Full Name** Kirsten rat sarcoma viral oncogene homolog provided by HGNC

**Primary source** HGNC:HGNC:6407

**See related** Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193

**Gene type** protein coding

**RefSeq status** REVIEWED

**Organism** Homo sapiens

**Lineage** Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

**Also known as** NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Search Help Hide sidebar >>

Table of contents

Summary

Genomic context

Genomic regions, transcripts, and products

Bibliography

Phenotypes

Variation

HIV-1 interactions

Pathways from BioSystems

Interactions

General gene information

Markers, Related pseudogene(s), Homology, Gene Ontology

General protein information

NCBI Reference Sequences (RefSeq)

Related documents

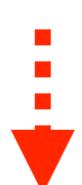
KRAS Kirsten rat sarcoma

Gene Ontology [Provided by GOA](#)

Function	Evidence Code	Pubs
<a href="#">GDP binding</a>	<a href="#">IEA</a>	
<a href="#">GMP binding</a>	<a href="#">IEA</a>	
<a href="#">GTP binding</a>	<a href="#">IEA</a>	
<a href="#">LRR domain binding</a>	<a href="#">IEA</a>	
<a href="#">protein binding</a>	<a href="#">IPI</a>	<a href="#">PubMed</a>
<a href="#">protein complex binding</a>	<a href="#">IDA</a>	<a href="#">PubMed</a>

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

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



# GO: Gene Ontology

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

The screenshot shows a web browser window with the UniProt-GOA database. The address bar shows 'www.ebi.ac.uk/GOA'. The page has a dark header with 'EMBL-EBI' and a green circular logo. Below the header is a teal banner with the text 'UniProt-GOA'. A search bar contains 'Examples: GO:0006915, tropomyosin, P06727' and a 'Search' button. The main content area features a large title 'Gene Ontology Annotation (UniProt-GOA) Database'. Below it is a paragraph about the UniProt GO annotation program. To the right is a sidebar titled 'Menu' with a list of links.

KRAS Kirsten rat sarcoma × UniProt-GOA < EMBL-EBI ×

www.ebi.ac.uk/GOA

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

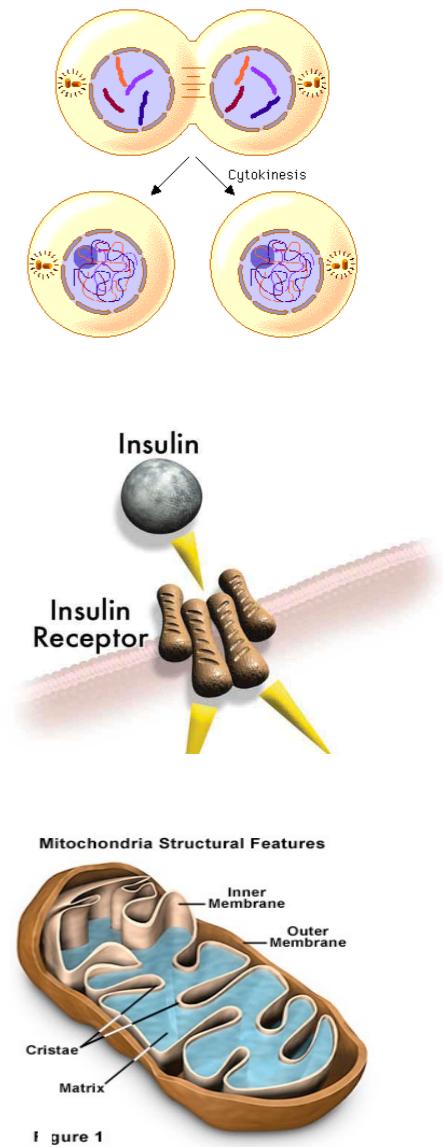
- [Downloads](#)
- [Searching UniProt-GOA](#)
- [Annotation Methods](#)
- [Annotation Tutorial](#)
- [Manual Annotation Efforts](#)
  - [Reference Genome Annotation Initiative](#)
  - [Cardiovascular Gene Ontology Annotation Initiative](#)
  - [Renal Gene Ontology Annotation Initiative](#)
  - [Exosome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

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

KRAS Kirsten rat sarcoma X www.ncbi.nlm.nih.gov/gene/3845#gene-ontology

genomic X01669.1 CAA25828.1

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

Protein Accession	Links
P01116.1	<a href="#">GenPept Link</a> <a href="#">UniProtKB Link</a> <a href="#">GenPept</a> <a href="#">UniProtKB/Swiss-Prot:P01116</a>

Additional links

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

**GETTING STARTED**

- NCBI Education
- NCBI Help Manual
- NCBI Handbook
- Training & Tutorials

**RESOURCES**

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

**POPULAR**

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

**FEATURED**

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

**NCBI INFORMATION**

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

UniProtKB/Swiss-Prot:P01116

UniProtKB/Swiss-Prot link

UniProt link

UniProt will detail much more information for protein coding genes

KRAS - GTPase KRas prec

www.uniprot.org/uniprot/P01116

UniProtKB Advanced

BLAST Align Retrieve/ID Mapping Help Contact

Basket

## P01116 - RASK\_HUMAN

Protein: GTPase KRas  
Gene: KRAS  
Organism: Homo sapiens (Human)  
Status: Reviewed - Experimental evidence at protein level<sup>i</sup>

Display: None

BLAST Align Format Add to basket History Feedback Help video

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

### Function<sup>i</sup>

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

### Enzyme regulation<sup>i</sup>

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 <sup>i</sup>	10 – 18	9	GTP 2 Publications			
Nucleotide binding <sup>i</sup>	29 – 35	7	GTP 2 Publications			
Nucleotide binding <sup>i</sup>	59 – 60	2	GTP 2 Publications			

UniProt will detail much more information for protein coding genes

KRAS - GTPase KRas prec

www.uniprot.org/uniprot/P01116

UniProtKB Advanced

BLAST Align Retrieve/ID Mapping Help Contact Basket

# P01116 - RASK\_HUMAN

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

Display: None

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

**Format**

**Function**  
Ras proteins bind GDP/GTP and promote cell proliferation (PubMed: 23698361, 23698362)

**Enzyme regulation**  
Alternates between an inactive form and an active form as a nucleotide-exchange factor (GEF). Promotes exchange of bound GDP by GTP. 3 Publications

**Regions**

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

>sp|P01116|RASK\_HUMAN GTPase KRas OS=Homo sapiens GN=KRAS PE=1 SV=1 MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDL PSRTVDTKQAQDLARSYGYIPFIETSAKTRQRVEDAFYTLVREIRQYRLKKISKEEKTPGC VKIKKCIIM

UniProt will detail much more information for protein coding genes

KRAS - GTPase KRas prec

www.uniprot.org/uniprot/P01116

UniProtKB Advanced

BLAST Align Retrieve/ID Mapping Help Contact

Basket

# P01116 - RASK\_HUMAN

Protein: GTPase KRas  
Gene: KRAS  
Organism: Homo sapiens (Human)  
Status: Reviewed - Experimental evidence at protein level<sup>i</sup>

Display: None

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

BLAST Align Format Add to basket History Feedback Help video

## Function<sup>i</sup>

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

### Enzyme regulation<sup>i</sup>

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 <sup>i</sup>	10 – 18	9	GTP 2 Publications			
Nucleotide binding <sup>i</sup>	29 – 35	7	GTP 2 Publications			
Nucleotide binding <sup>i</sup>	59 – 60	2	GTP 2 Publications			

KRAS - GTPase KRas prec x

www.uniprot.org/uniprot/P01116

UniProtKB Advanced

BLAST Align Retrieve/ID Mapping Help Contact

# P01116 - RASK\_HUMAN

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

Display None

BLAST Align Format Add to basket History Feedback Help video

FUNCTION

NAMES & TAXONOMY

SUBCELL. LOCATION

PATHOL/BIOTECH

PTM / PROCESSING

EXPRESSION

INTERACTION

STRUCTURE

FAMILY & DOMAINS

SEQUENCES (2)

CROSS-REFERENCES

## Example Questions:

What positions in the protein are responsible for GTP binding?

### Function<sup>i</sup>

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

### Enzyme regulation<sup>i</sup>

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 <sup>i</sup>	10 – 18	9	GTP 2 Publications			
Nucleotide binding <sup>i</sup>	29 – 35	7	GTP 2 Publications			
Nucleotide binding <sup>i</sup>	59 – 60	2	GTP 2 Publications			

# Example Questions:

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

KRAS - GTPase KRas prec x

www.uniprot.org/uniprot/P01116

Display None

FUNCTION

NAMES & TAXONOMY

SUBCELL. LOCATION

PATHOL/BIOTECH

PTM / PROCESSING

EXPRESSION

INTERACTION

STRUCTURE

FAMILY & DOMAINS

SEQUENCES (2)

CROSS-REFERENCES

PUBLICATIONS

ENTRY INFORMATION

MISCELLANEOUS

SIMILAR PROTEINS

▲ Top

**Pathology & Biotech<sup>1</sup>**

**Involvement in disease<sup>i</sup>**

**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 <sup>i</sup>	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?

The screenshot shows the Uniprot protein entry page for KRAS (P01116). The 'STRUCTURE' section is highlighted with a red box. A blue box highlights the secondary structure visualization, showing a sequence of green, blue, and pink bars representing beta strands, helices, and turns respectively. A red arrow points to the '4EPV' link in the 3D structure database table, which is also highlighted with a green box. A red box with the text 'Open link in a new tab!' surrounds the '4EPV' row.

**Display** None

**Structure<sup>1</sup>**

**Secondary structure**

1 Legend: Helix Turn Beta strand 189

Show more details

**3D structure databases**

Select the link destinations:

PDB<sup>i</sup>

RCSB PDB<sup>i</sup>

PDBj<sup>1</sup>

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	[»]
<b>4EPV</b>	X-ray	1.35	A	1-164	[»]
4EPW	X-ray	1.70	A	1-1	
4EPX	X-ray	1.76	A	1-1	
4EPY	X-ray	1.80	A	1-1	
4L8G	X-ray	1.52	A	1-1	
4LDJ	X-ray	1.15	A	1-164	[»]
4LPK	X-ray	1.50	A/B	1-169	[»]

▲ 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 links for Home, Gmail, Deposit, Search, Visualize, and Analyze. Below this is the main header with the RCSB PDB logo and a search bar. The search bar contains the text "Search by PDB ID, author, macromolecule, sequence, or ligand" and a "Go" button. There are also links for Advanced Search and Browse by Annotations. To the right of the search bar is a map of the world with various protein structures overlaid.

Below the header, there are several logos: PDB-101, Worldwide Protein Data Bank, EMDDataBank, Nucleic Acid Database, and Worldwide Protein Data Bank Foundation. A social media sharing bar with icons for Facebook, Twitter, and YouTube is also present.

The main content area features a navigation menu with tabs: Structure Summary, 3D View (which is highlighted with a red box), Annotations, Sequence, Sequence Similarity, Structure, and Literature. A green speech bubble points to the "View PDB file format" link in the Literature tab.

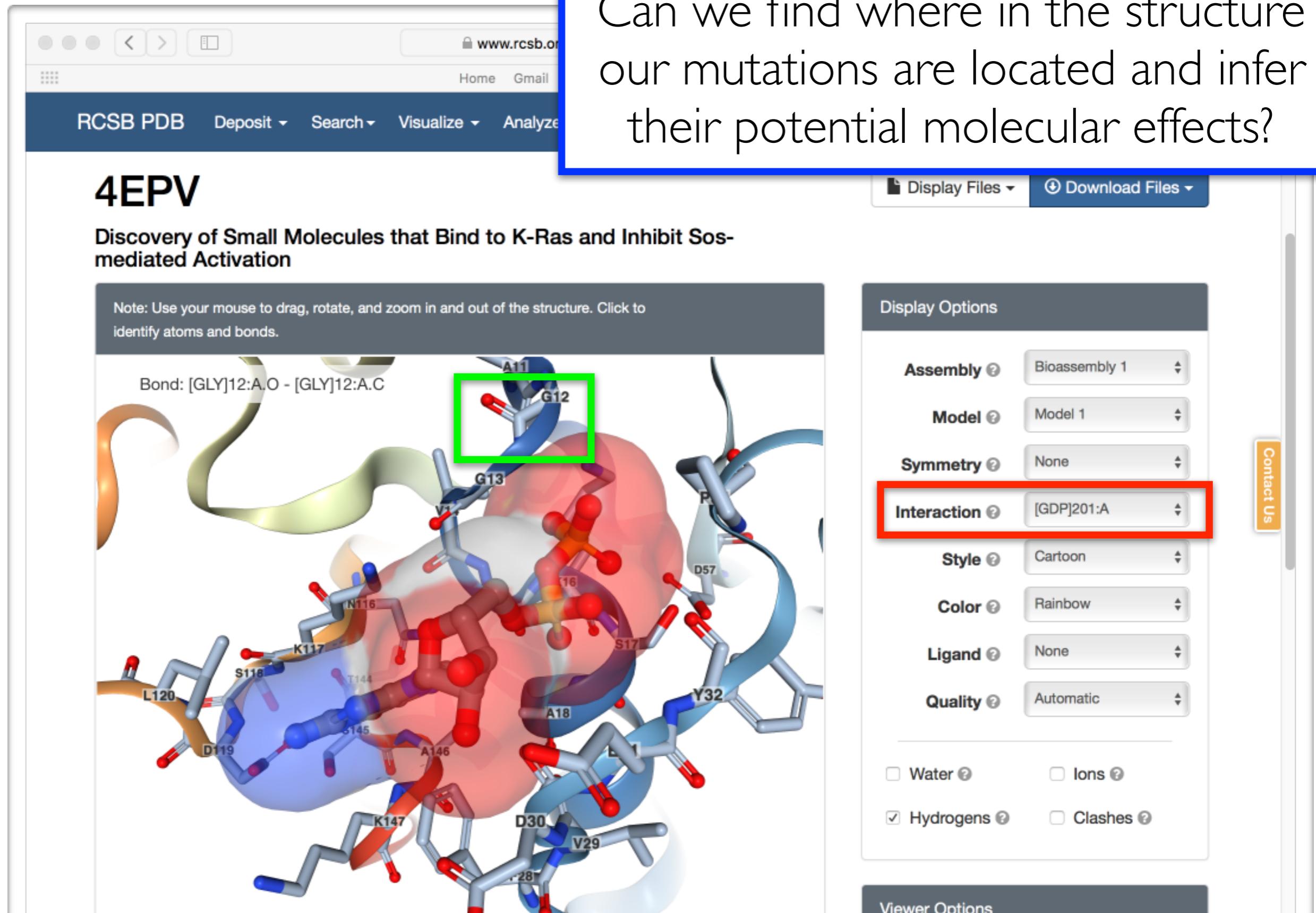
On the left side, there is a ribbon diagram of a protein structure labeled "Biological Assembly 1". On the right side, the structure code "4EPV" is displayed in large letters, followed by the title "Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation". Below the title, the DOI is listed as "DOI: 10.22110/pdb4epv/pdb". Further down, the classification is listed as "HYDROLASE", and the deposition details are "Deposited: 2012-04-17 Released: 2012-05-23". The deposition authors are "Sun, Q., Burke, J.P., Phan, J., Burns, M.C., Olejniczak, E.T., Waterson, A.G., Lee, T., Rossanese, O.W., Fesik, S.W.". The organism is "Homo sapiens", the expression system is "Escherichia coli", and there is 1 mutation.

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

A green box highlights the "Display Files" dropdown menu in the "Download Files" section of the right sidebar.

# Lets view the 3D structure:

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



## Back to UniProt:

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

KRAS - GTPase KRas prec X 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

OrthoDB

PhylomeDB<sup>i</sup> P01116

TreeFam<sup>i</sup> TF3

Family and domain databases

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

Sequences (2)<sup>i</sup>

Sequence status<sup>i</sup>: Complete.

Sequence processing<sup>i</sup>: The displayed sequence is further processed into a mature form.

This entry describes 2 isoforms<sup>i</sup> produced by alternative splicing. Align

PFAM is one of the best  
protein family databases

# Example Questions:

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

The screenshot shows a web browser window with the URL [pfam.xfam.org/family/PF00071](http://pfam.xfam.org/family/PF00071). The page header includes the EMBL-EBI logo and a navigation bar with links like HOME, SEARCH, and HELP. The main content area displays information for the protein family PF00071, specifically KRAS - GTPase KRas precursor. It includes a summary table with columns for architectures, sequences, interactions, species, and structures, and a detailed description of the Ras family.

## Family: Ras (PF00071)

**Summary**

**Domain organisation**

**Clan**

**Alignments**

**HMM logo**

**Trees**

**Curation & model**

**Species** (highlighted with a red box)

**Interactions**

**Structures**

**Jump to...** ⓘ

enter ID/acc Go

### Summary: Ras family

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

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

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

#### Ras subfamily [Edit Wikipedia article](#)

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

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

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

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

Because these signals result in cell growth and division, overactive Ras signaling can ultimately lead to [cancer](#).<sup>[1]</sup> 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](#)).<sup>[2]</sup> For this reason, Ras inhibitors are being studied as a treatment for cancer, and other diseases with Ras overexpression.

[Contents](#) [hide]

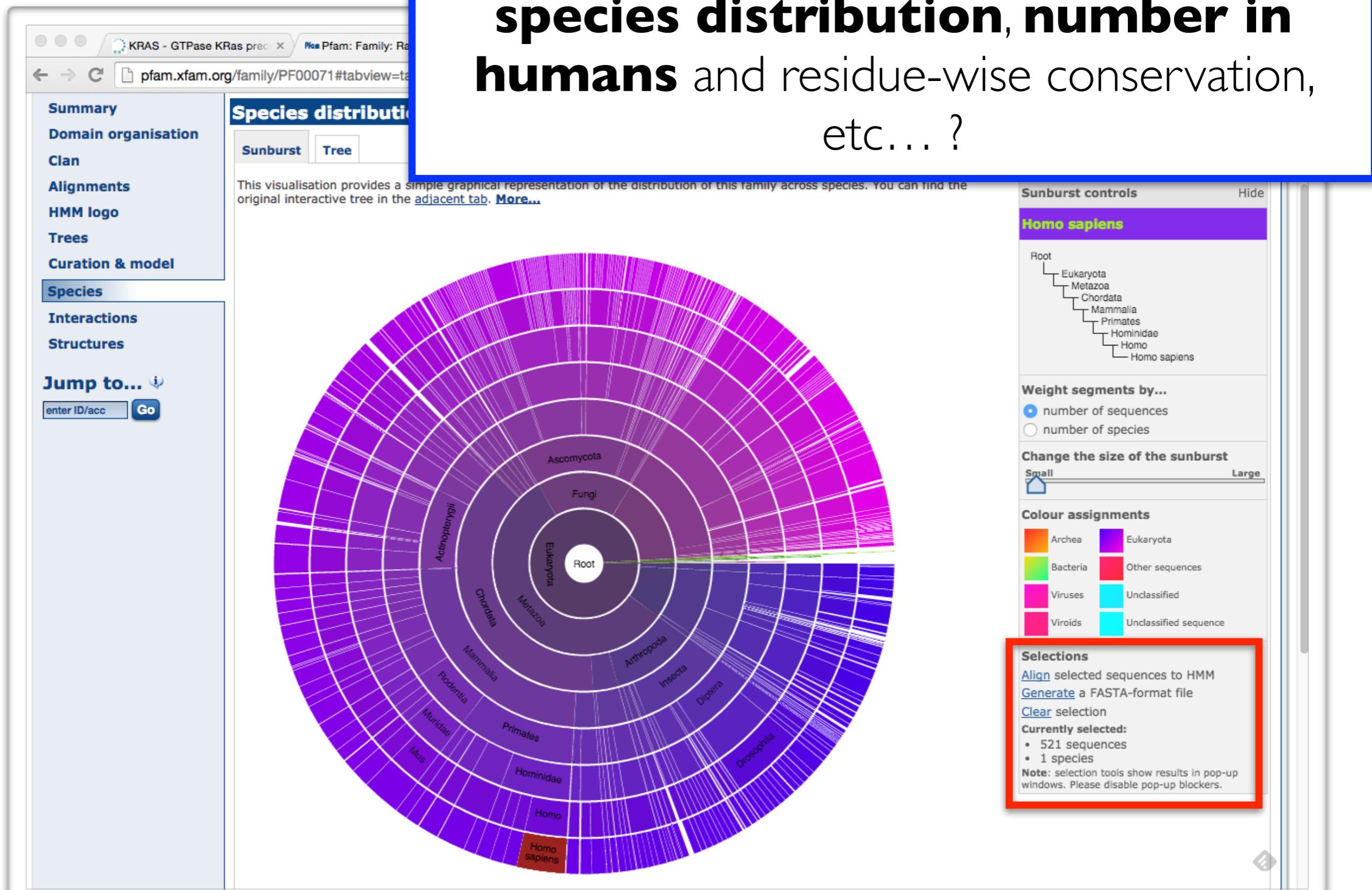
1 History  
2 Structure  
3 Function  
  3.1 Activation and deactivation  
  3.2 Membrane attachment  
4 Members  
5 Ras in cancer  
  5.1 Inappropriate activation  
  5.2 Constitutively active Ras

**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... ?

KRAS - GTPase KRas prec X More Pfam: Family: Re...

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

Summary Species distribution

Domain organisation Sunburst Tree

Clan Alignment HMM log Trees Curation Species Interactions Structures

Jump to enter ID/acc

Pfam: Pfam alignment viewer pfam.xfam.org/family/PF00071/alignment/view?jobId=EDCA403E-9836-11E4-B360-10B3298E2F76

EMBL-EBI

Alignment for selected sequences

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

P11234/16-178	KVIMVGSCGGVGKSAITL	Q	FM	Y	D	EF	V	E	DYEPIK	-AD	SYRKVVLD
P01112/5-165	KLVVVGAGGVGKSAITI	Q	LI	Q	N	HF	V	D	EYDPTI	-ED	SYRKQVVID
Q14088/38-204	KIIIVGDSNVGKTCILTF	R	FC	G	G	TF	P	D	KTEAII	GVD	FREKTVEIE
Q9BW83/7-173	KCILAGDPAVGKTAIIAQ	I	FR	S	DgaHF	Q	K	SYTLT	GMD	IVVKTVPVPd	
P15153/7-178	KCVVVGDDGAVGKTCILLI	S	YT	T	N	AF	P	C	EYIPTV	-FD	NYSANVMVD
D00194/11-183	KLLALGDSCGVGKTCFLY	R	YT	D	N	KF	N	P	KFIFTIV	GID	FREKRVVYNaaqgn
Q15907/13-174	KVVLIGDSCGVGKSNLIS	R	FT	R	N	EF	N	L	ESKSII	GVE	FATRSIQVD
P10114/5-166	KVVVLGSCGGVGKSAITV	Q	FV	T	G	TF	I	E	KYDPTI	-ED	FYRKEIEVD
P51153/10-171	KLLLIGDSCGVGKTCII	R	FA	E	D	NF	N	N	TYISII	GID	FKIRTVDIE
P55040/77-241	RVVLLIGEQGVGKSTLAN	I	FA	Gvhd	SM	D	S	D	CEVL	GED	TYERTLMVD
P55042/93-253	KVLLLGAPGVGKSAALAR	I	FG	G	V	ED	G	P	EEAAAG	--H	TYDRSIVVD
P01116/5-165	KLVVVGAGGVGKSAITI	Q	LI	Q	N	HF	V	D	EYDPTI	-ED	SYRKQVVID
Q9H07/21-182	KLVLLGSGSGVGKSSLAL	R	YV	K	N	DF	K	S	ILPTIV	GCA	FFTAKVVDVG
Q9ULC3/11-171	KMVVVGNGAVVGKSSMIQ	R	YC	K	G	IF	T	K	DYKII	GVD	FLERQIQVN
Q14807/15-177	KLVVVGDGGVGKSAITI	Q	FF	Q	K	IF	V	P	DYDPTI	-ED	SYLKHTIED
Q9NX57/7-202	KIVLLGDMNVGKTSILQ	R	YM	E	R	RF	P	D	T-VSIV	GGA	FYLKQW
Q9H082/35-201	KIIIVGDSNVGKTCILY	R	FC	A	G	RF	P	D	RTEAII	GVD	FRERAVEID
Q969Q5/9-174	KVVMLGKEYVGKTSILVE	R	YV	H	D	RFIV	V	C	PYQNTI	GAA	FVAKVMSVC
P51149/10-175	KVIILGDSGVGKTSILMN	Q	YV	N	K	KF	S	N	QYKAI	GAD	FLTKEVMVD
Q9ULW5/65-227	KVMLVGDSCGVGKTCILV	R	FK	D	G	AF	L	Ag	TFISIV	GID	FRNKVLDVD
P57735/14-175	KVVLIGECSGVGKTNLLS	R	FT	R	N	EF	S	H	DSRTII	GVE	FSTRTVMLG
P51159/11-183	KFLALGDSCGVGKTSVLY	Q	YT	D	G	KF	N	S	KFIFTIV	GID	FREKRVVYRasgpd
P01111/5-165	KLVVVVGAGGVGKSAITI	Q	LI	Q	N	HF	V	D	EYDPTI	-ED	SYRKQVVID
P11233/16-177	KVIMVGSCGGVGKSAITL	Q	FM	Y	D	EF	V	E	DYEPIK	-AD	SYRKVVLD
Q9UL25/21-182	KVVLLGEGCGVGKTSIVL	R	YC	E	N	KF	N	D	KHITIL	QAS	FLTKKLNIC
Q9NP72/10-171	KILIIGECSGVGKSSILL	R	FT	D	D	TF	D	P	ELAATI	GVD	FKVKTISVD
Q9HOU4/10-171	KLLLIGDSCGVGKSCILL	R	FA	D	D	TY	T	E	SYISII	GVD	FKIRTIELD
Q9UL26/7-168	KVCLLGDGTGVGKSSIVW	R	FV	E	D	SF	D	P	NINPI	GAS	FMTKTVQYQ
Q9UBK7/23-179	KIICLGDSAVGKSKLME	R	FL	M	D	GF	Q	P	QQLSY	ALT	LYKHTATVD
P51157/14-179	KIVVLGDCASGKTSIT	C	FA	Q	E	TF	G	K	QYKQII	GLD	FFLRRITLP

1 2 3 4 5 6 7 8 9 10 11 ...

There are 18 pages in this alignment. Show page 1

Download this alignment.

Close window

can find the

Sunburst controls Hide

**Homo sapiens**

Root  
Eukaryota  
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... ?

KRAS - GTPase KRas prec < More Pfam: Family: Ras

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

EMBL-EBI 

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

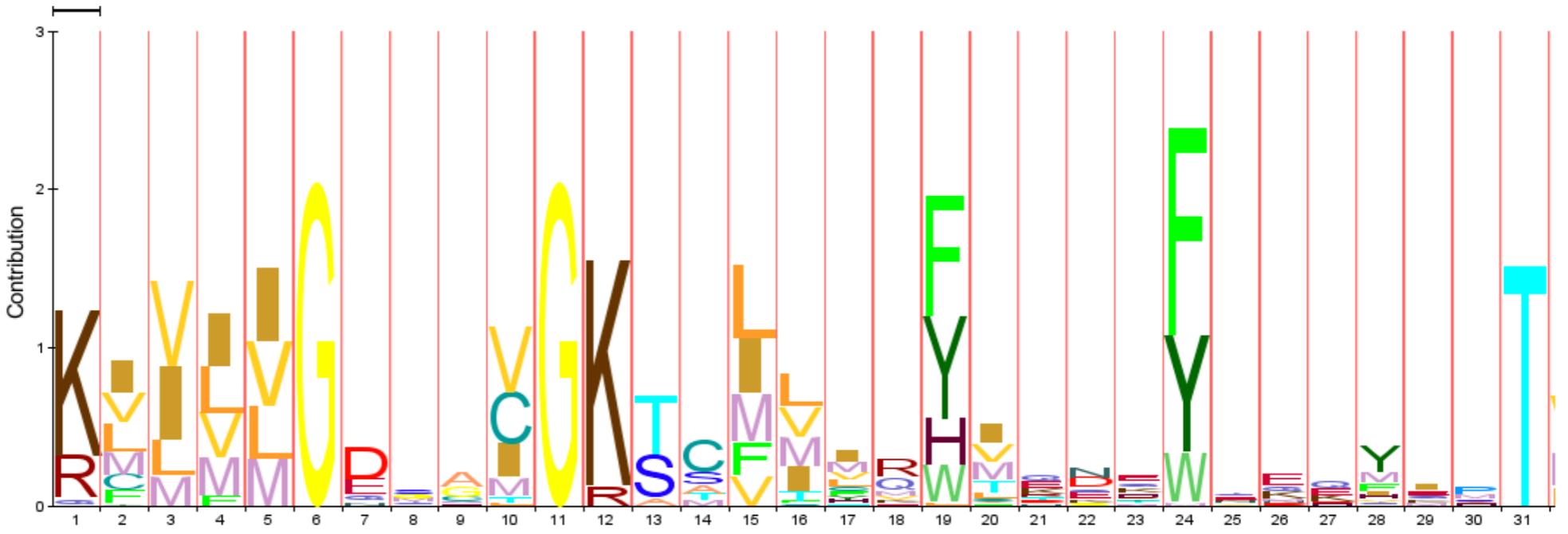
## Family: Ras (PF00071)

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

Jump to... enter ID/acc Go

### HMM logo

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



Comments or questions on the site? Send a mail to [pfam-help@ebi.ac.uk](mailto:pfam-help@ebi.ac.uk).  
European Molecular Biology Laboratory

# 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... !\[\]\(4fa1a734083d1c409ea6d909f0c24706\_img.jpg\)](#)
 [Go](#)

## Interactions

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

[Tubulin](#)
[Tubulin\\_C](#)
[Tubulin\\_C](#)
[Kinesin](#)
[Tubulin](#)
[Kinesin](#)

# Family: Kinesin (PF00225)

  
**126** architectures   
**4150** sequences   
**6** Interactions   
**248** species   
**114** structures
**Summary****Domain organisation****Clans****Alignments****HMM logo****Trees****Curation & models****Species****Interactions****Structures****Jump to...**

enter ID/acc

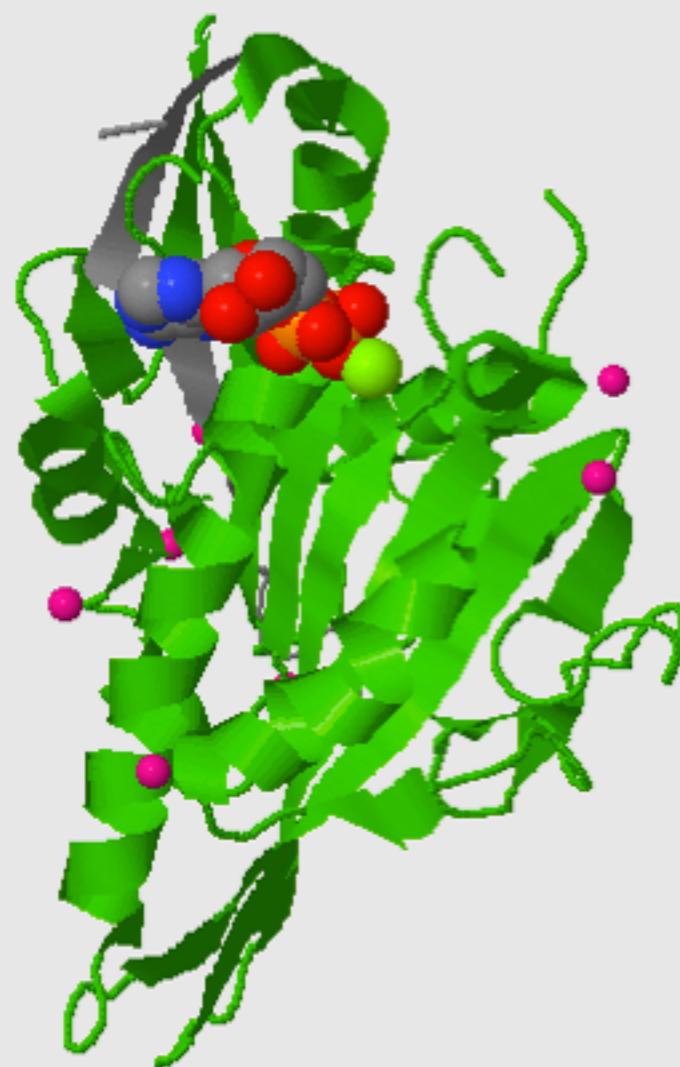
## Structures

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

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



PDB entry 3bfm



## Your turn:

What can you find out about “eg5”

Jmol

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

 Close window

# Today's Menu

## Classifying Databases

Primary, secondary and composite Bioinformatics databases

## Using Databases

**Vignette** demonstrating how major Bioinformatics databases intersect

## Major Biomolecular Formats

How nucleotide and protein sequence and structure data are represented

## Alignment Foundations

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

## Alignment Algorithms

**Hands-on** exploration of alignment algorithms and applications

# ALIGNMENT FOUNDATIONS

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

# ALIGNMENT FOUNDATIONS

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

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

**Seq1 :** C A T T C A C

**Seq2 :** C T C G C A G C

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

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



Two types of character correspondence

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

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

Add gaps to increase number of matches

match mismatch

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

Gaps represent 'indels'  
mismatch represent mutations

match  
mismatch } mutation

insertion  
deletion } indels

# Why compare biological sequences?

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

# Practical applications include...

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

# Practical applications include...

- **Similarity searching of databases**
    - Protein structure prediction
  - **Assembly of sequences**
    - construct such as
  - **Mapping of new sequences to a known genome**
    - Looking for differences from reference sequences, indels (insertions or deletions)
    - Mapping transcription factor binding sites via ChIP-Seq (chromatin immuno-precipitation sequencing)
    - Pretty much all next-gen sequencing data analysis
- N.B. Pairwise sequence alignment is arguably the most fundamental operation of bioinformatics!

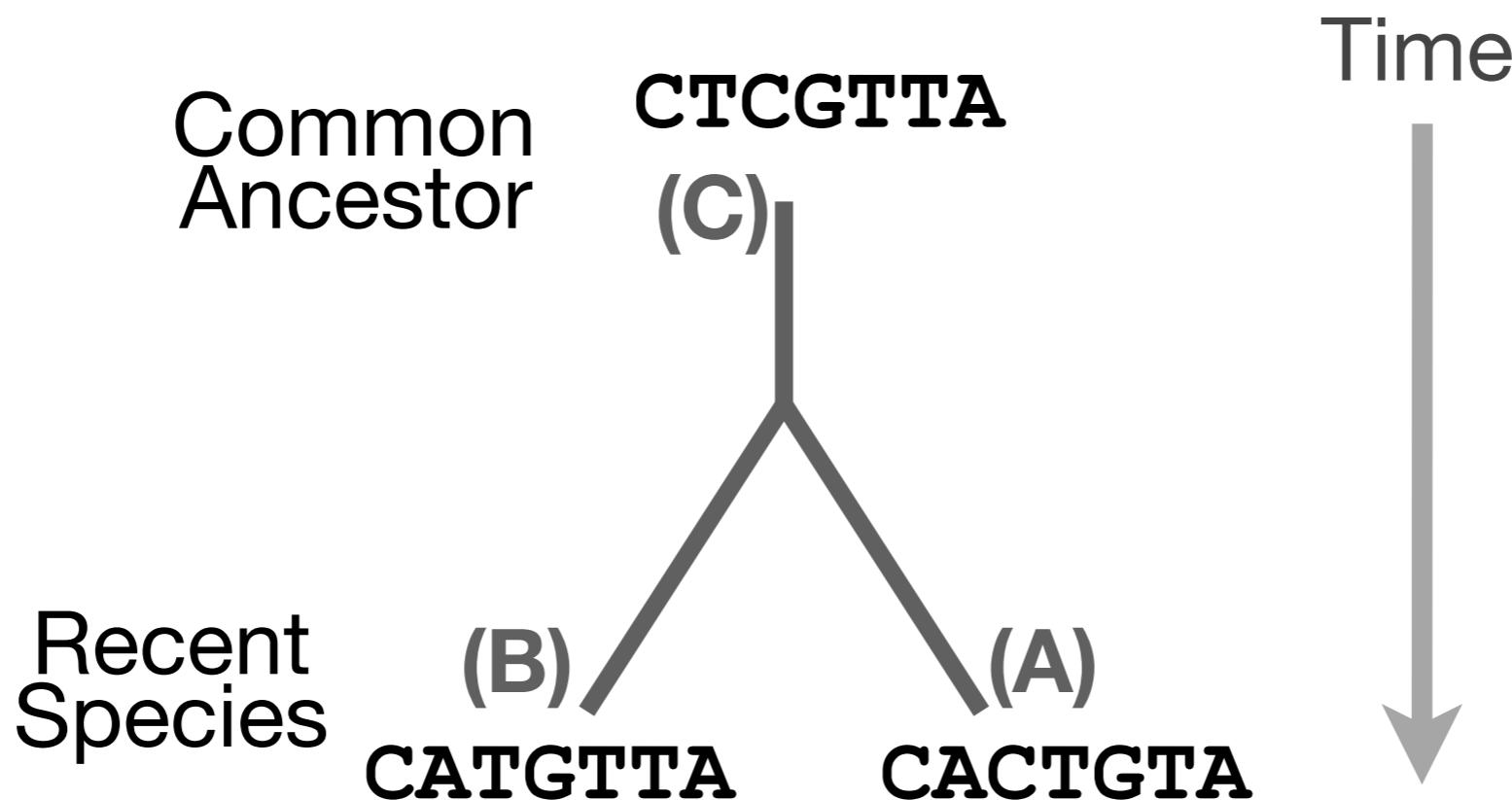
# ALIGNMENT FOUNDATIONS

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

# Sequence changes during evolution

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

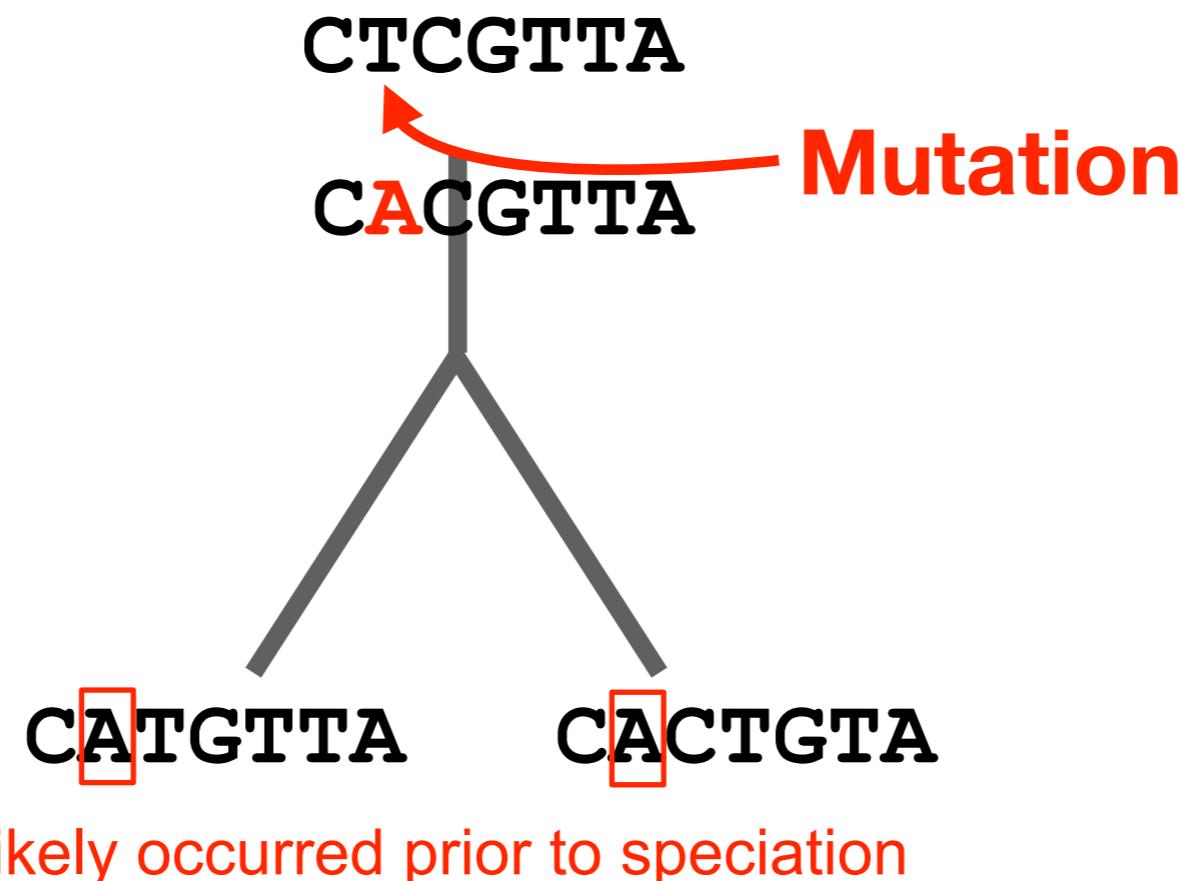
- Mutations/Substitutions
- Deletions
- Insertions



# Mutations, deletions and insertions

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

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

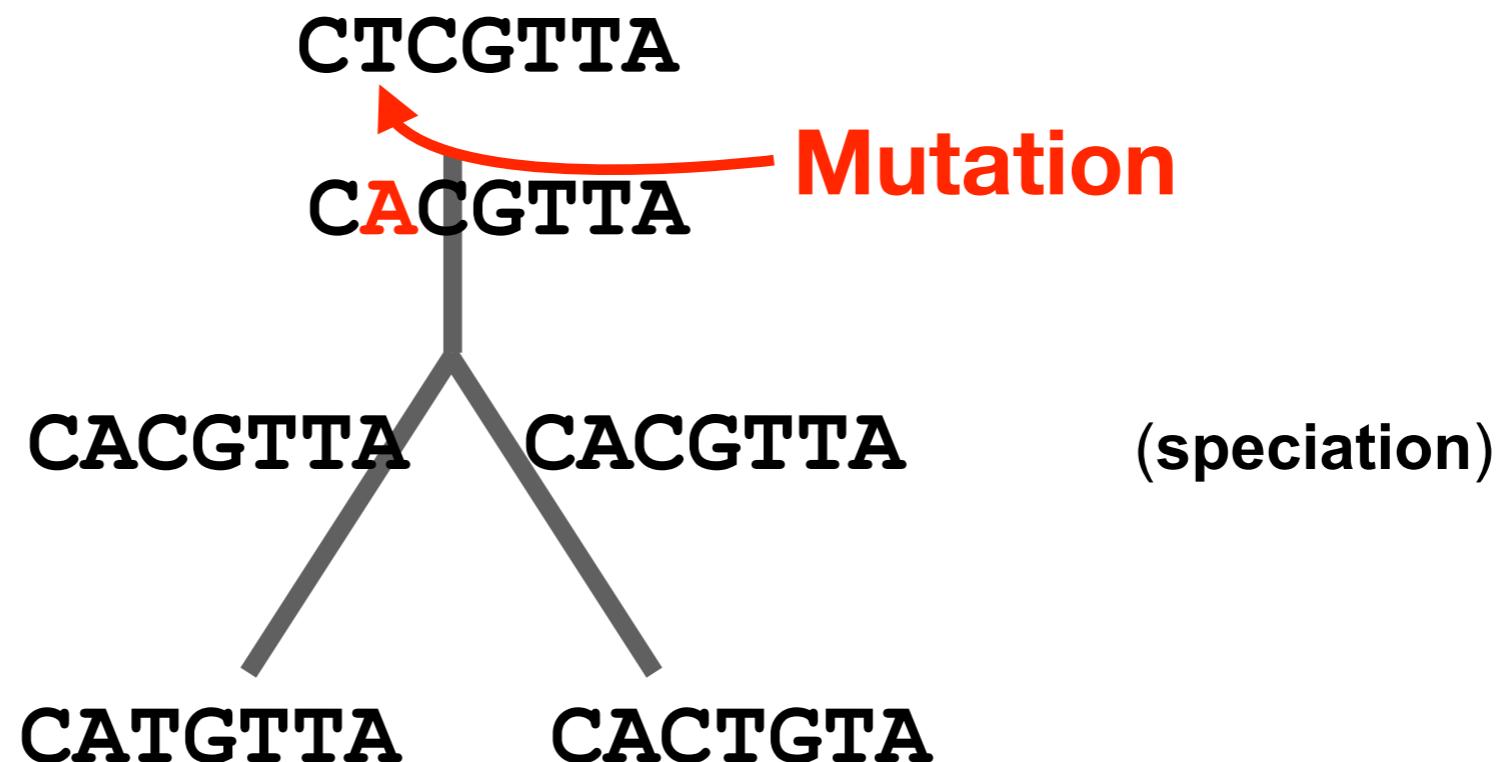


# Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA

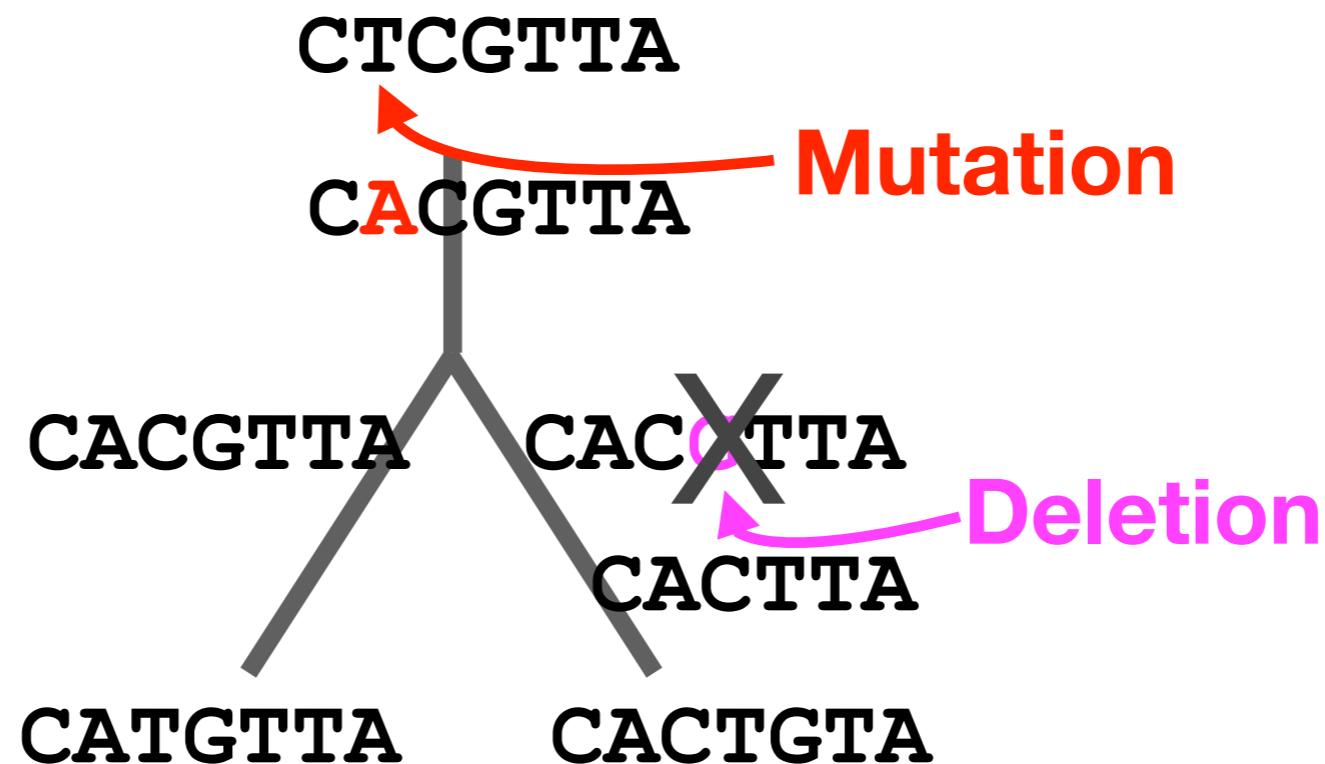


# Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA  
CACGTTA → CACTTA

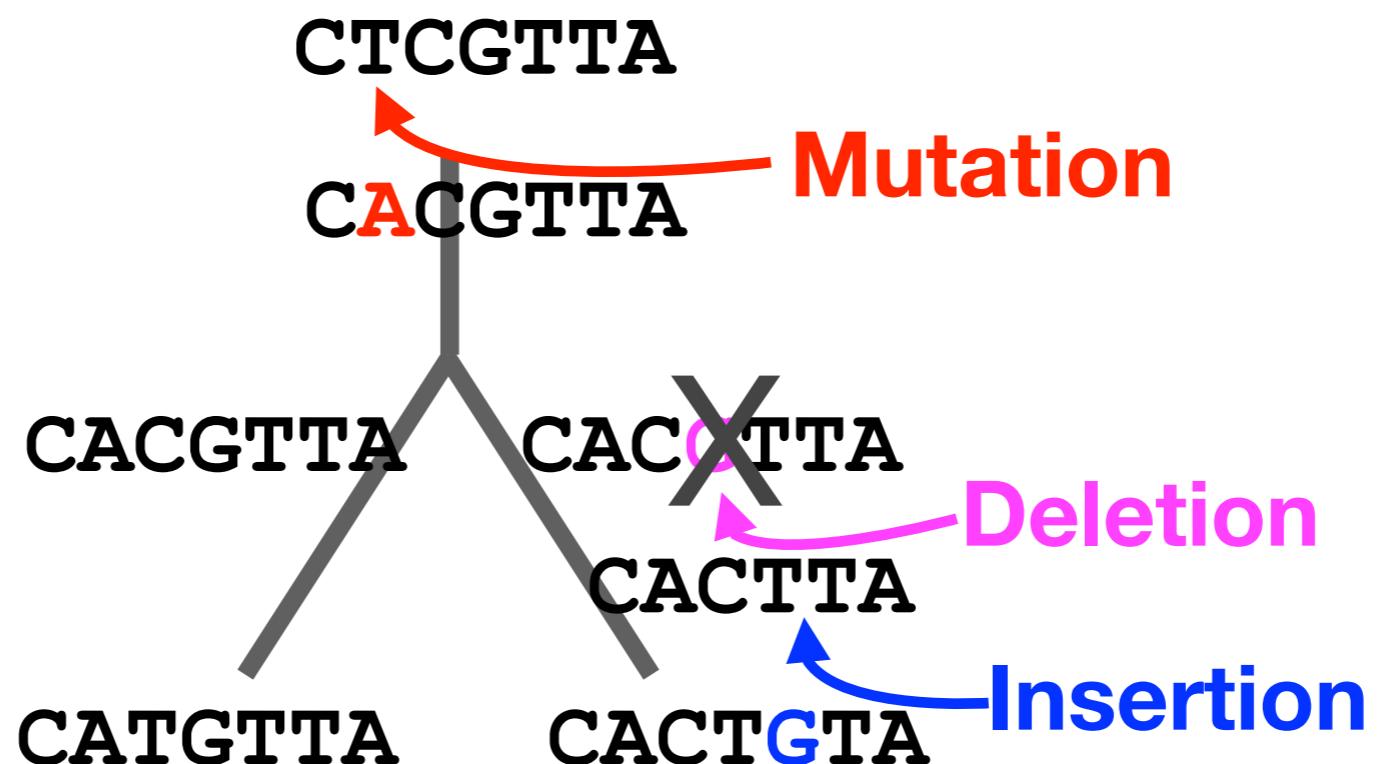


# Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA  
CACGTTA → CACTTA  
CACTTA → CACTGTA

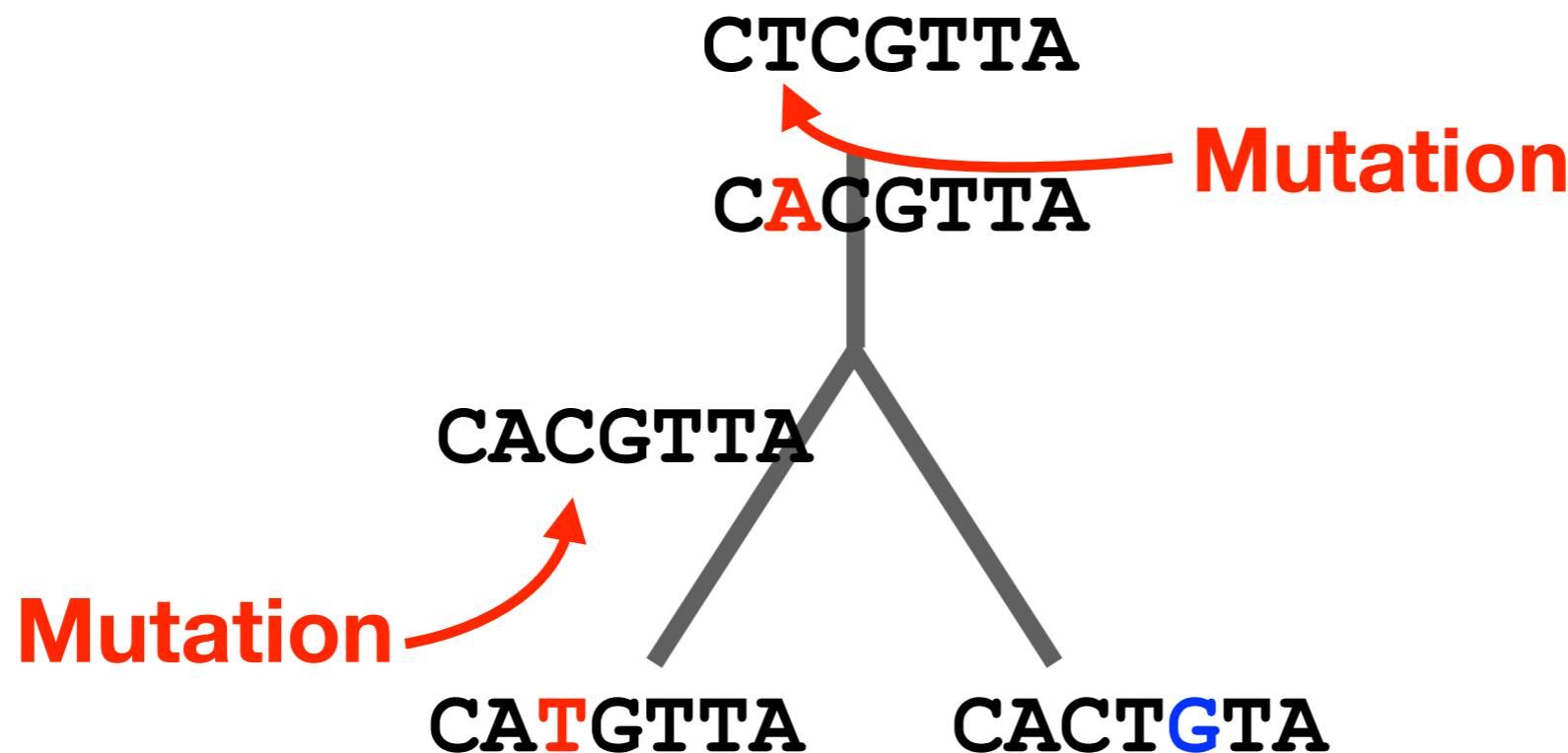


# Mutations, deletions and insertions

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

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

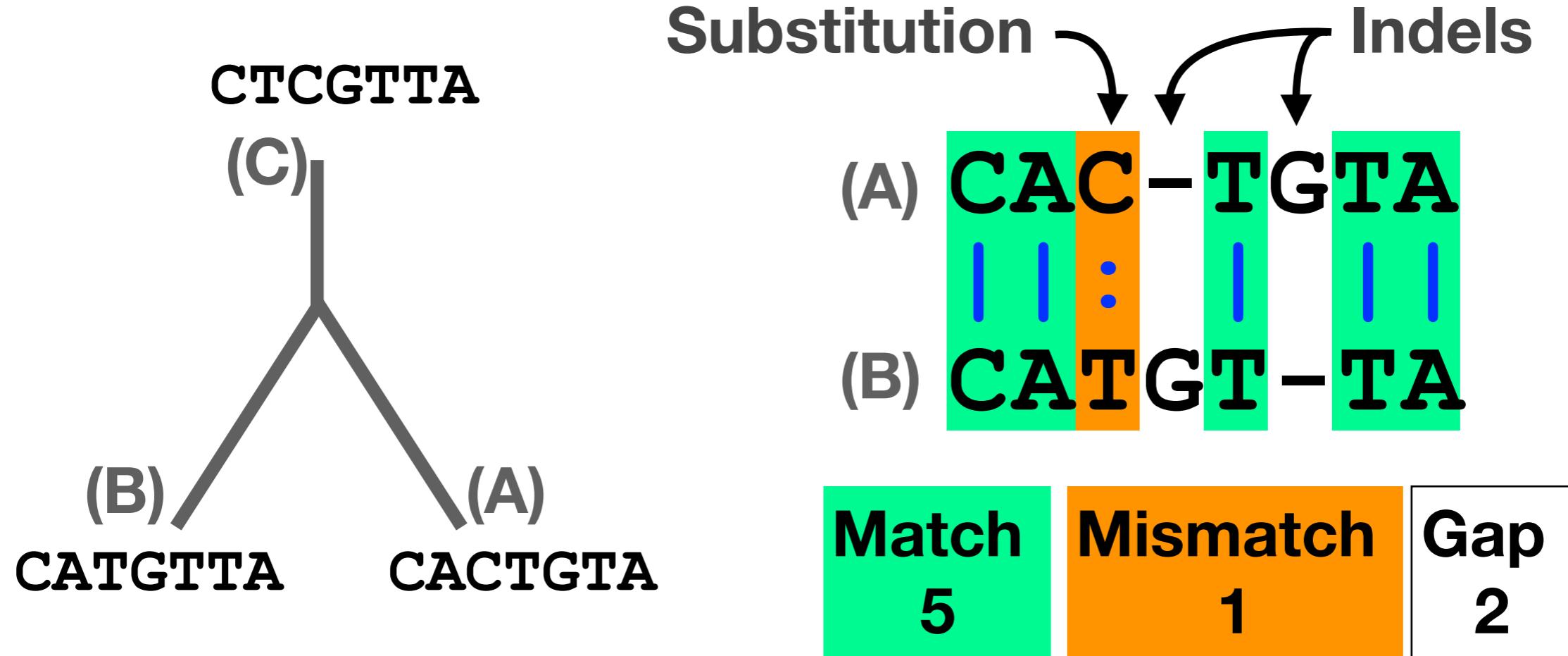
CTCGTTA → CACGTTA  
CACGTTA → CATGTTA



# Alignment view

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

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



# Alternative alignments

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

**Q.** Which of these 3 possible alignments is best?

1.

CA	CTG	TA
	:	:
CAT	TGT	TA

2.

CA	CTG	T	-A
CA	-T	G	T

3.

CAC	-T	G	TA
	:		
CAT	G	T	-TA

# Alternative alignments

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

● 4 matches  
● 3 mismatches  
○ 0 gaps

● 6 matches  
● 0 mismatches  
○ 2 gaps

● 5 matches  
● 1 mismatch  
○ 2 gaps

CACTGTA  
|| : : : ||  
CATGTAA

A sequence alignment between "CACTGTA" and "CATGTAA". The first four bases (CACT) are highlighted in green, indicating 4 matches. The fifth base (T) is orange, indicating a mismatch. The last two bases (GA) are also orange, indicating they are not aligned with any bases in the other sequence. There are no gaps.

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

A sequence alignment between "CACTGT-A" and "CA-TGTAA". The first five bases (CACTG) are green, indicating 5 matches. The next base (T) is orange, indicating a mismatch. The last three bases (-A, T, A) are orange, indicating they are not aligned with any bases in the other sequence. There are no gaps.

CAC-TGTA  
| | : | |  
CATGT-TA

A sequence alignment between "CAC-TGTA" and "CATGT-TA". The first three bases (CAC) are green, indicating 3 matches. The next base (T) is orange, indicating a mismatch. The last four bases (-TGTA) are orange, indicating they are not aligned with any bases in the other sequence. There are two gaps (indicated by blue vertical bars).

# 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

Sequence alignment diagram showing two DNA sequences:

CA	CTG	TA
	:	:
CAT	TGT	TA

The first sequence is colored green (methylated) and the second orange (unmethylated). Blue vertical bars indicate matches between C, T, G, and T.

Sequence alignment diagram showing two DNA sequences:

CA	CTG	T	-A	
CA	-T	G	T	A

The first sequence is green and the second orange. Blue vertical bars indicate matches between C, T, G, and T. A gap is shown with a dash and a score of -1.

Sequence alignment diagram showing two DNA sequences:

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

The first sequence is green and the second orange. Blue vertical bars indicate matches between C, G, T, and T. A gap is shown with a dash and a score of -1.

# Optimal alignments

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

- 4 matches
- 3 mismatches
- 0 gaps

CA	CTG	TA
	:	:
CAT	TGT	TA

- 6 matches
- 0 mismatches
- 2 gaps

CA	CTG	T	-	A
CA	-	TGT	T	A

- 5 matches
- 1 mismatch
- 2 gaps

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

# Optimal alignments

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

- 4 matches
- 3 mismatches
- 0 gaps

CA	CTG	TA
	:	:
CAT	TGT	TA

- 6 matches
- 0 mismatches
- 2 gaps

CA	CTG	T	-	A
CA	-	TGT	TA	

- 5 matches
- 1 mismatch
- 2 gaps

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

# Optimal alignments

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

- 4 matches
- 3 mismatches
- 0 gaps

CA	CTG	TA
	:	:
CAT	TGT	TA

- 6 matches
- 0 mismatches
- 2 gaps

CA	CTG	-	TA
CA	-TGT	TA	

- 5 matches
- 1 mismatch
- 2 gaps

CAC	-	TG	TA
	:		
CAT	G	T	-TA

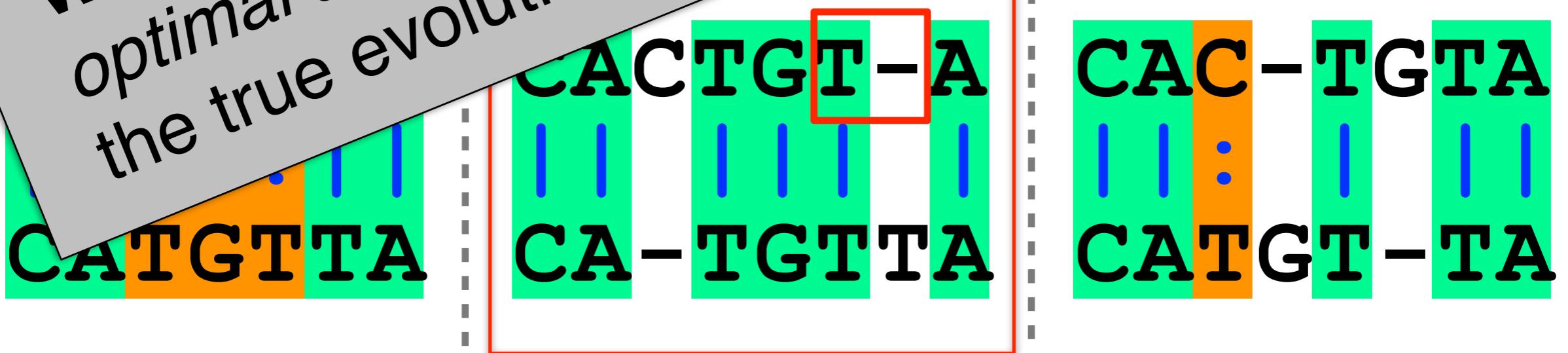
# Optimal alignments

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

- 4 matches
- 3 mismatches
- 2 gaps

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

- 3 matches
- 1 mismatch
- 2 gaps



# ALIGNMENT FOUNDATIONS

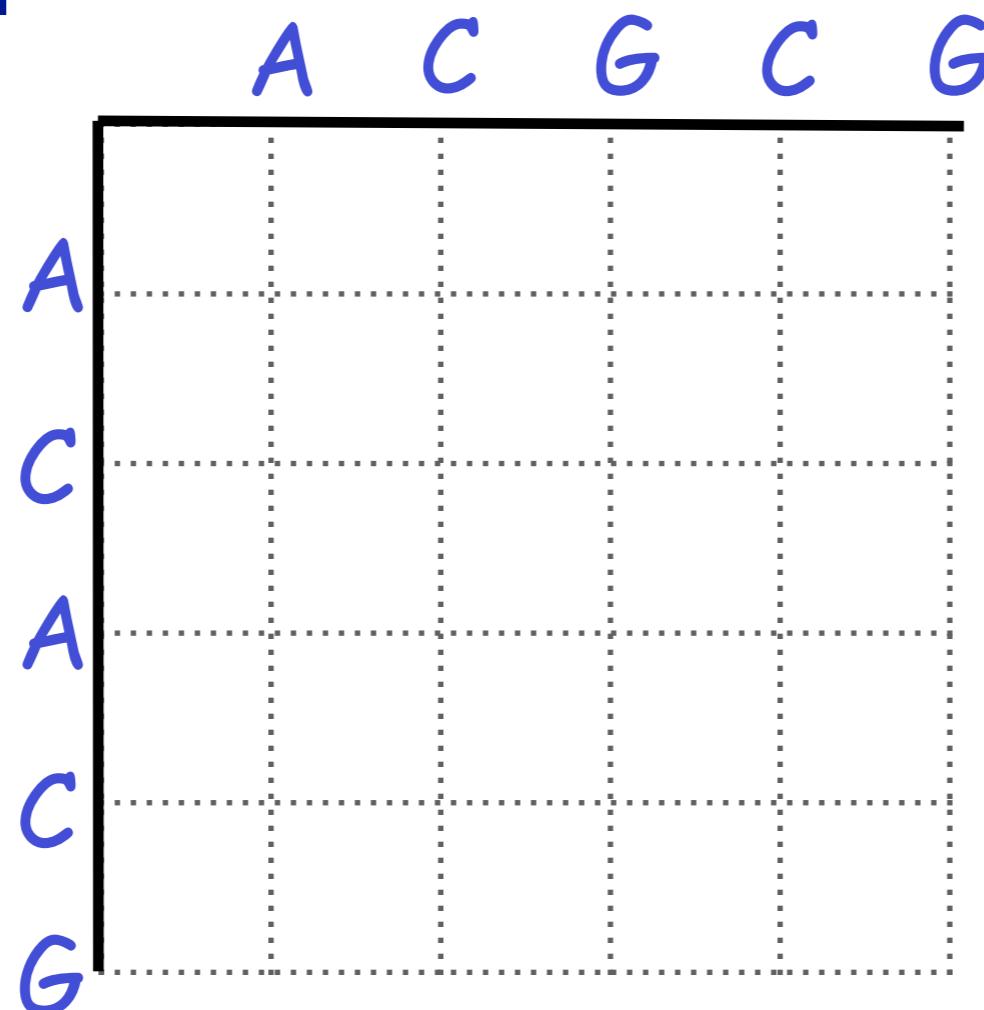
- **Why...**
  - Why compare biological sequences?
- **What...**
  - Alignment view of sequence changes during evolution  
(matches, mismatches and gaps)
- **How...**
  - ▶ Dot matrices
  - ▶ Dynamic programming
    - Global alignment
    - Local alignment
  - ▶ BLAST heuristic approach

# ALIGNMENT FOUNDATIONS

- Why...
    - Why compare biological sequences?
  - What...
    - Alignment view of sequence changes during evolution  
(matches, mismatches and gaps)
  - How...
    - ▶ Dot matrices
    - ▶ D
    - ▶ B
- How do we compute the optimal alignment between two sequences?

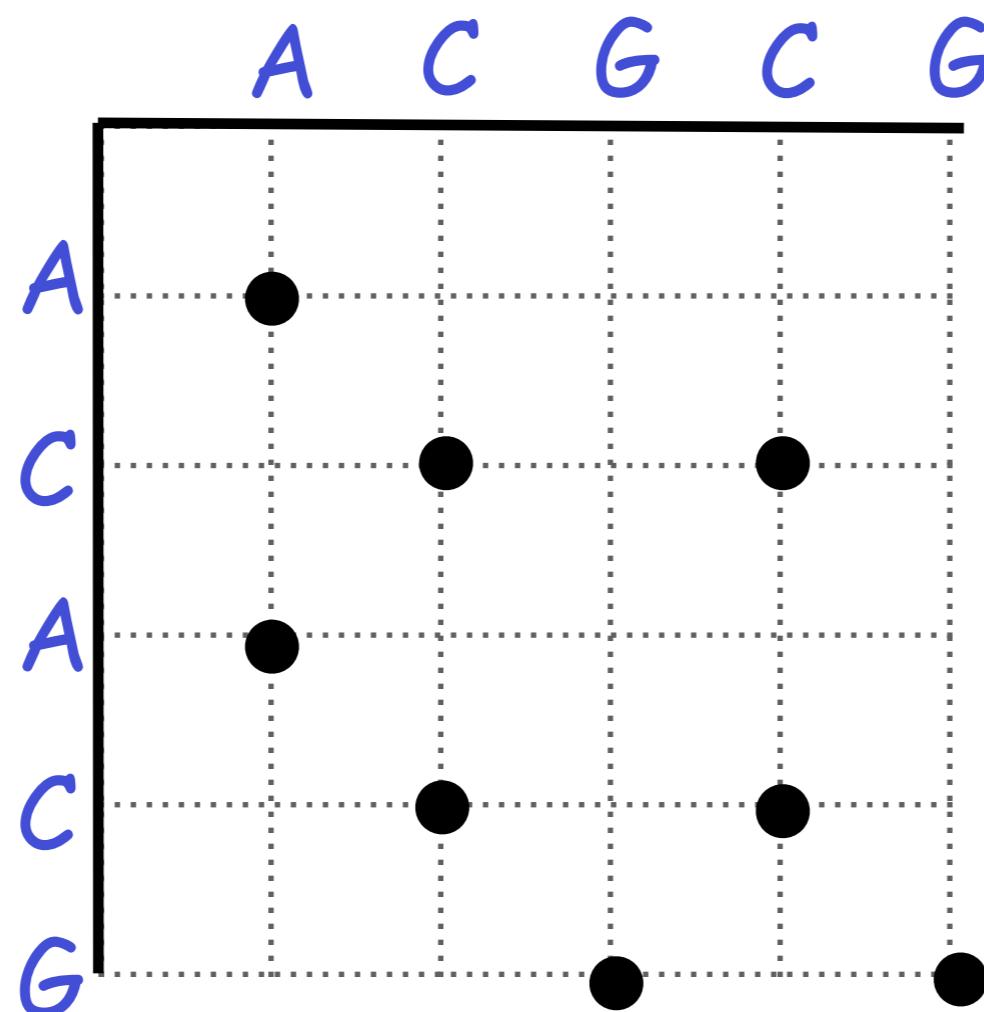
# Dot plots: simple graphical approach

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



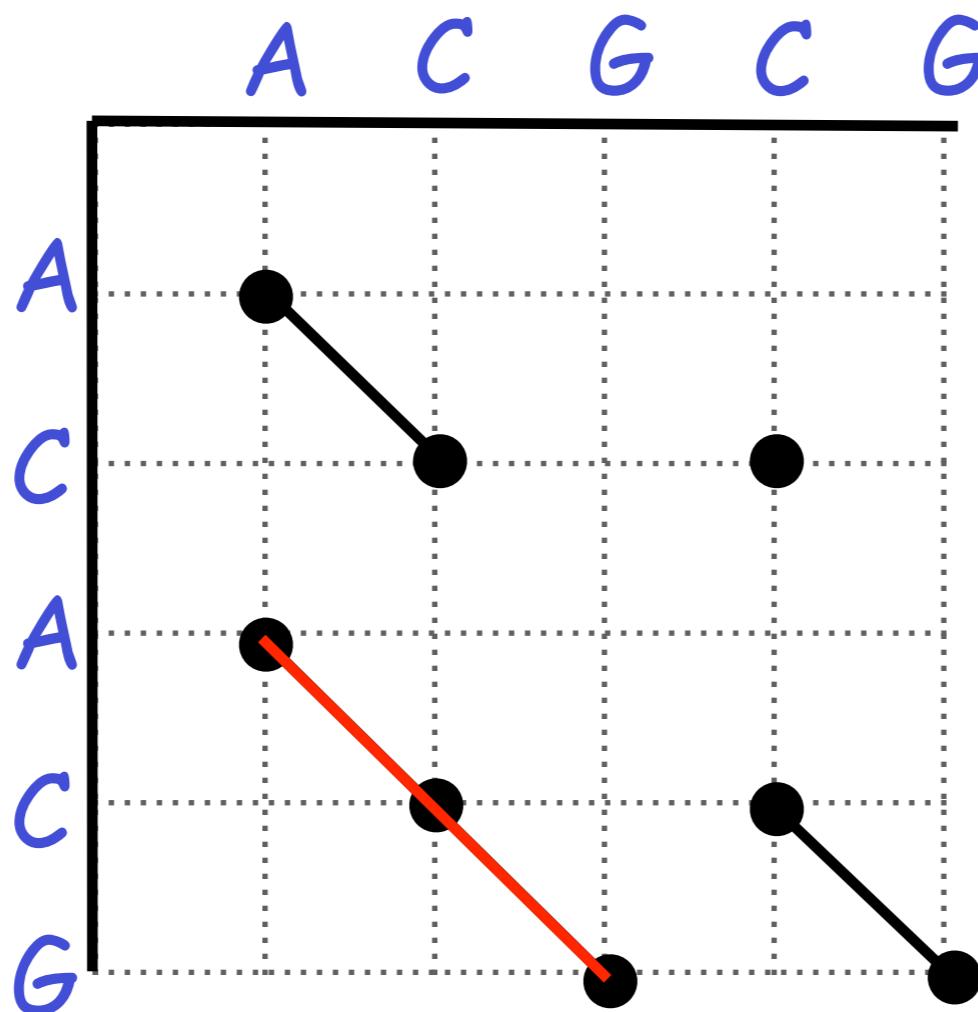
# Dot plots: simple graphical approach

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



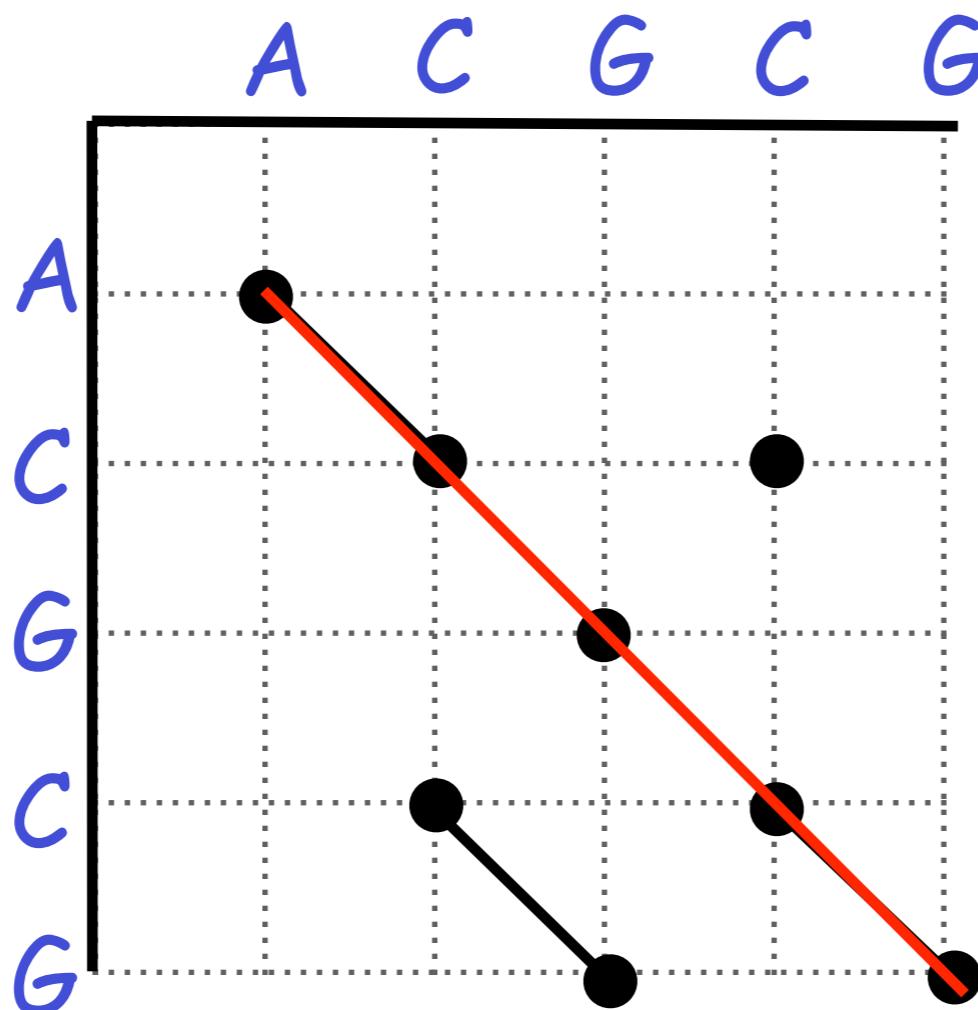
# Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



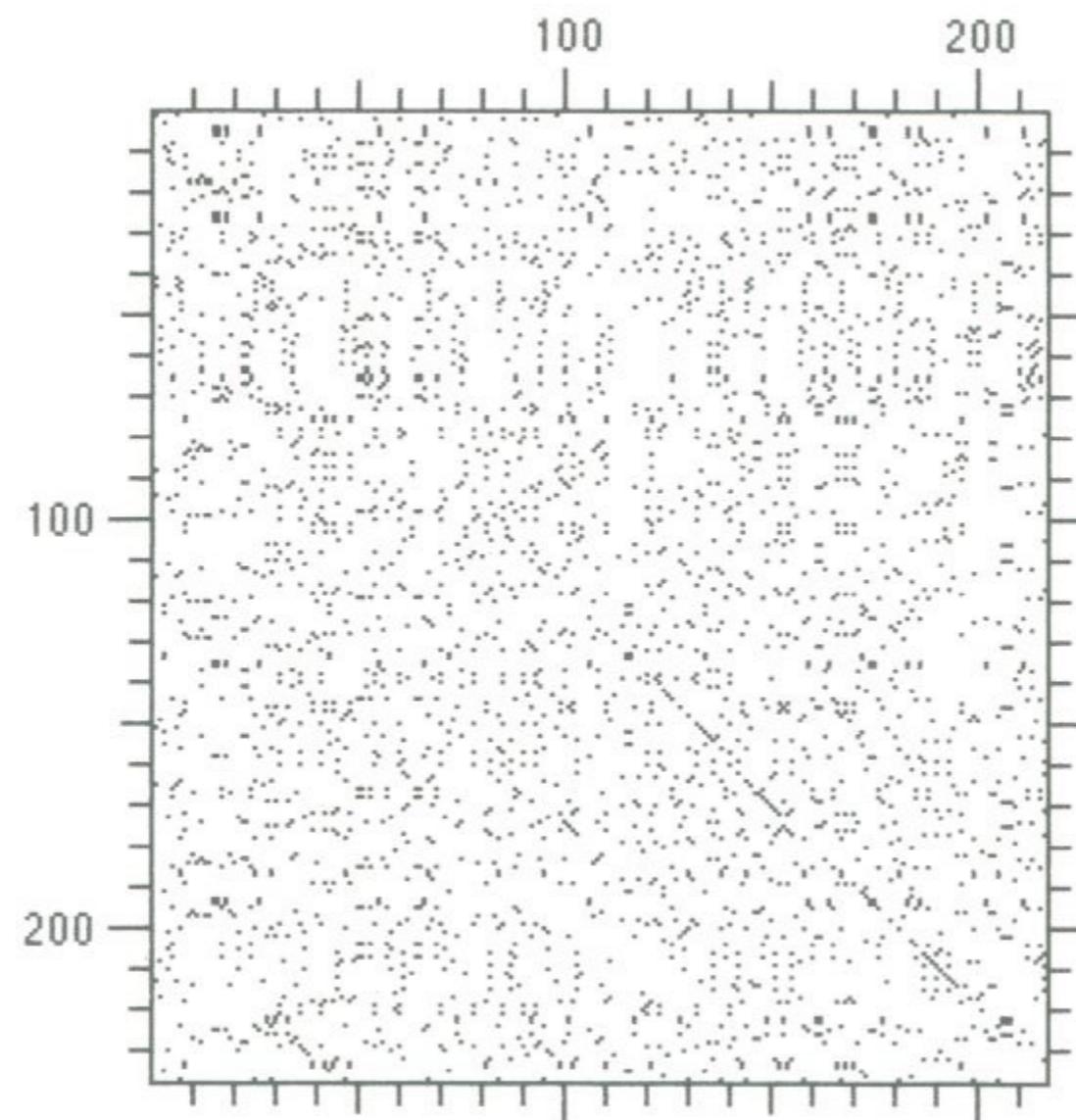
# Dot plots: simple graphical approach

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



# Dot plots: simple graphical approach

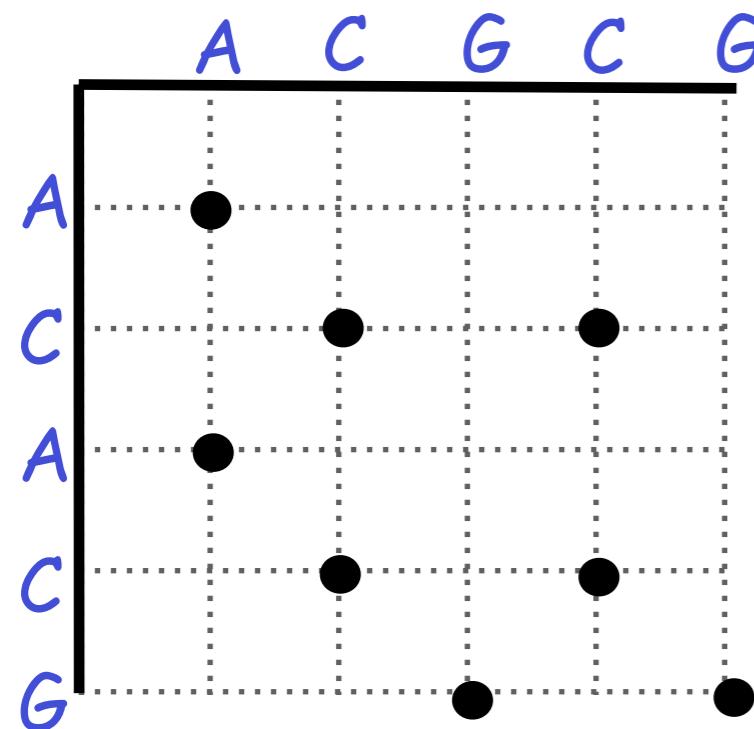
- Dot matrices for long sequences can be noisy



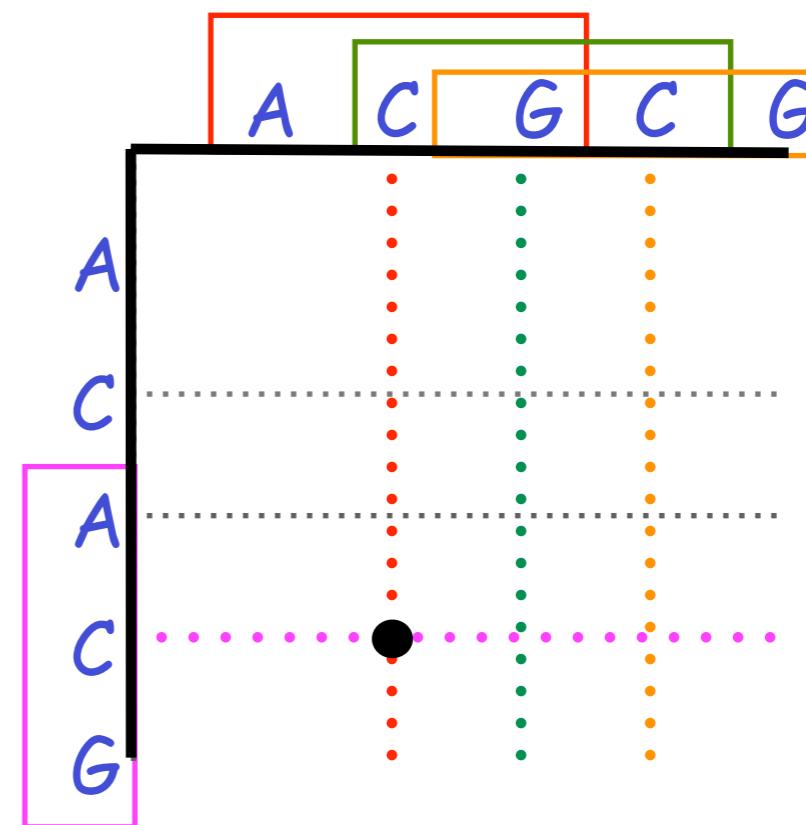
# Dot plots: window size and match stringency

**Solution:** use a window and a threshold

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



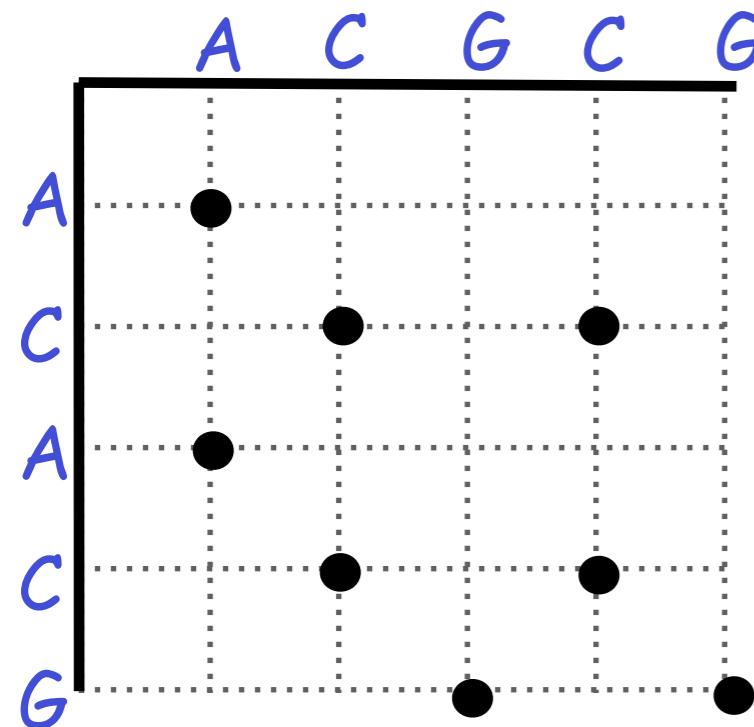
Filter  
Window = 3  
Stringency = 3



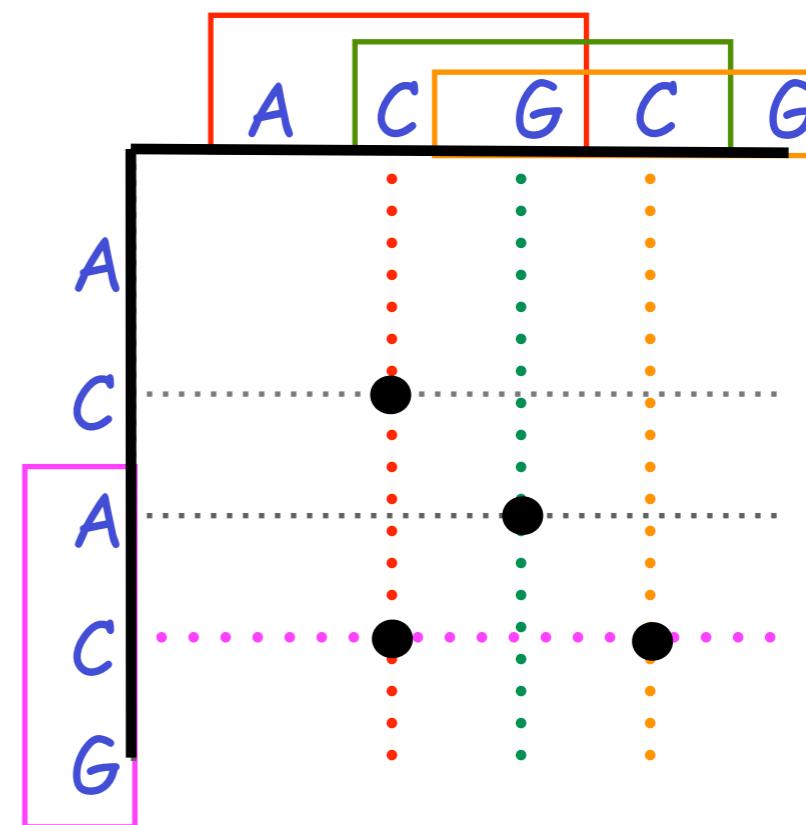
# Dot plots: window size and match stringency

**Solution:** use a window and a threshold

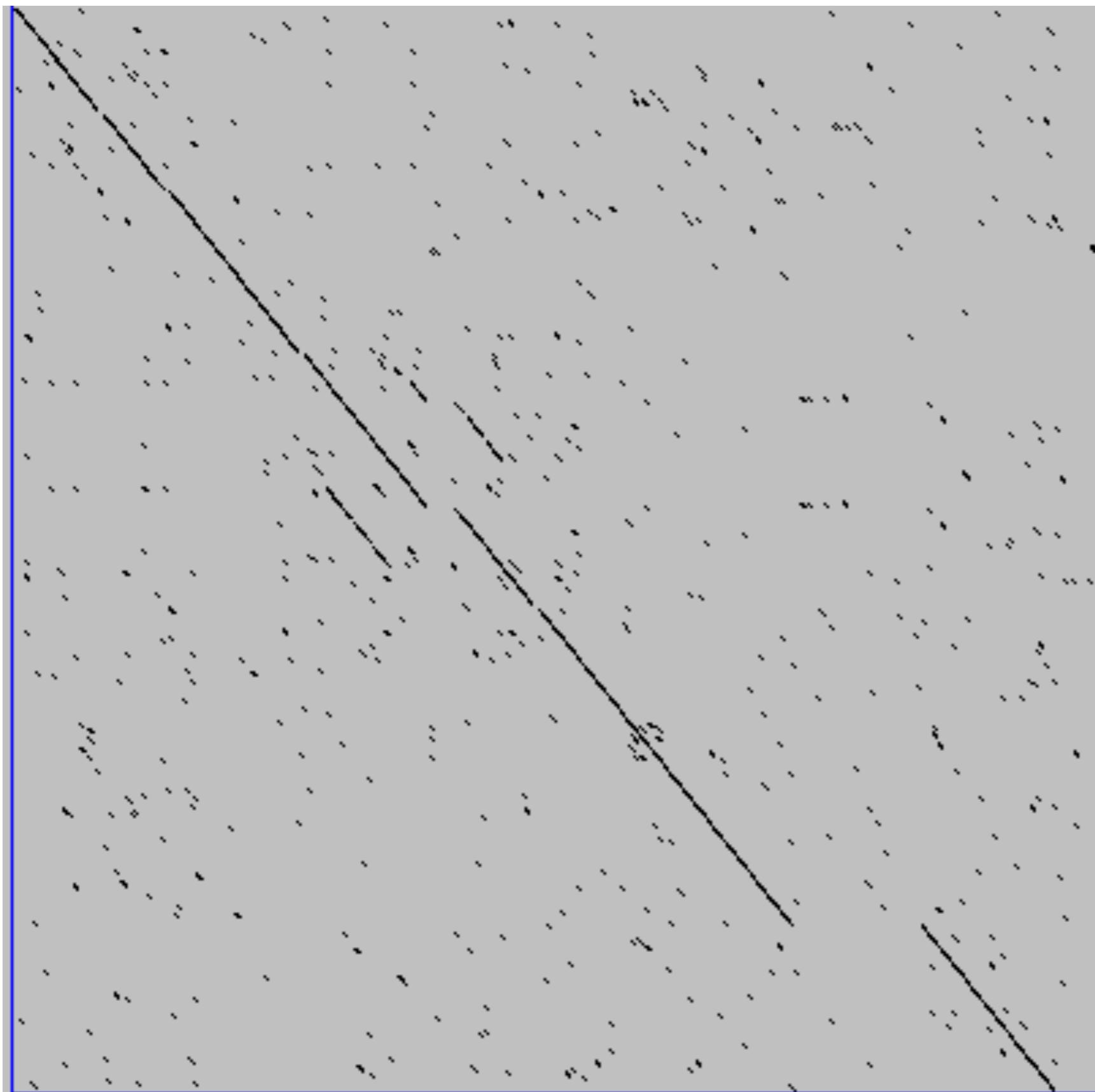
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
  - You have to choose window size and stringency



Filter  
Window = 3  
Stringency = 2



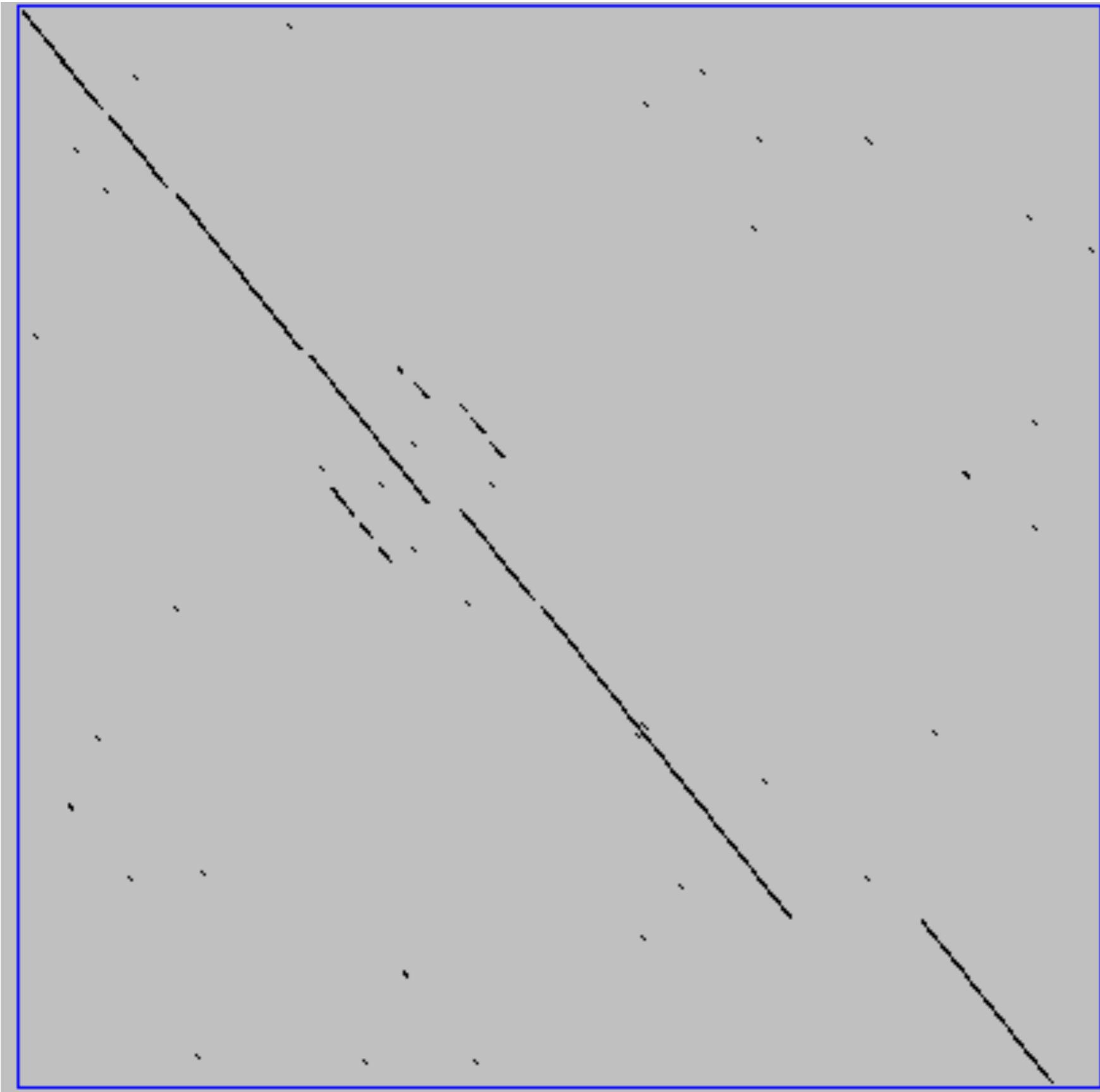
# Window size = 5 bases



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

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

# Window size = 7 bases

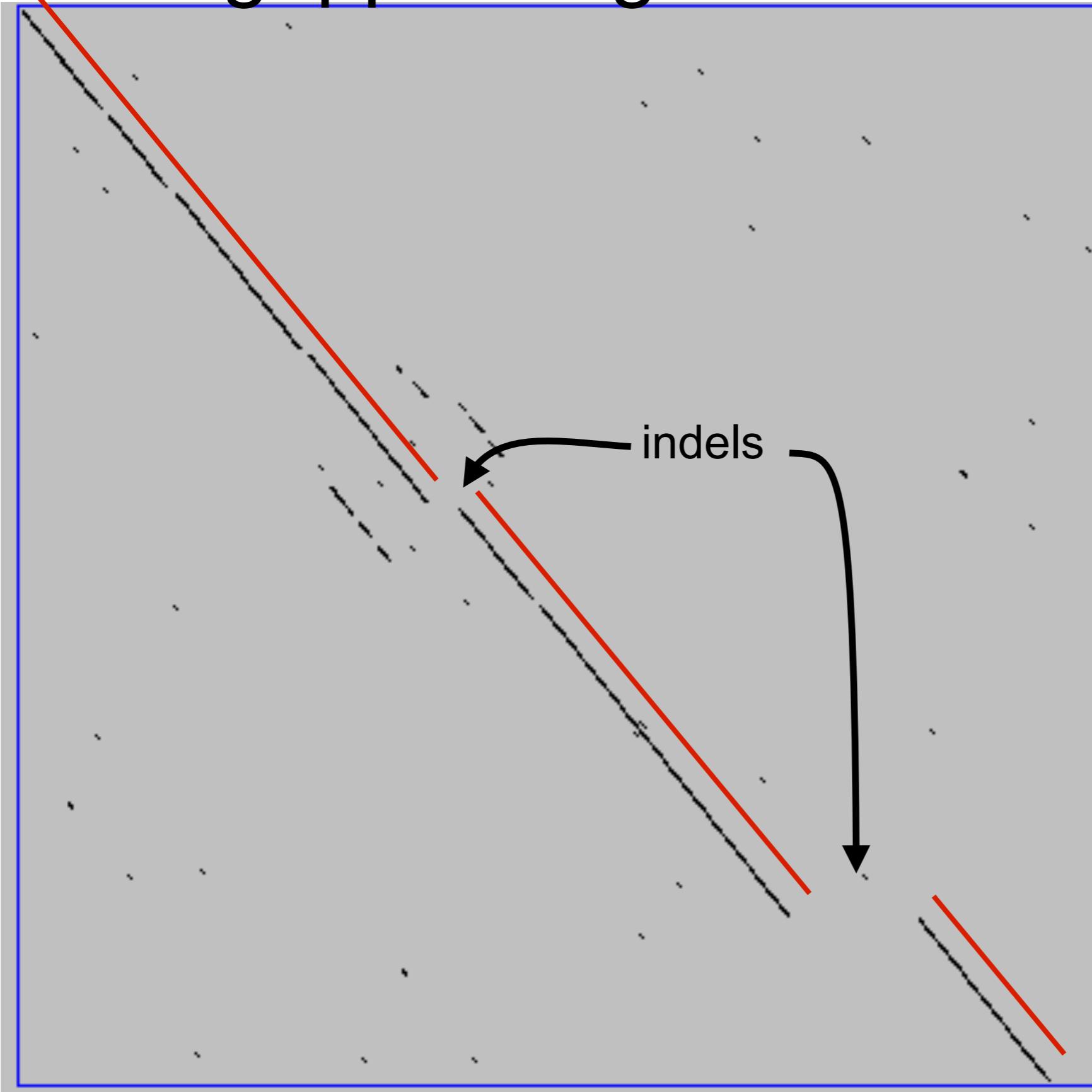


This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

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

Bigger window (kmer)  
fewer matches to consider

# Ungapped alignments



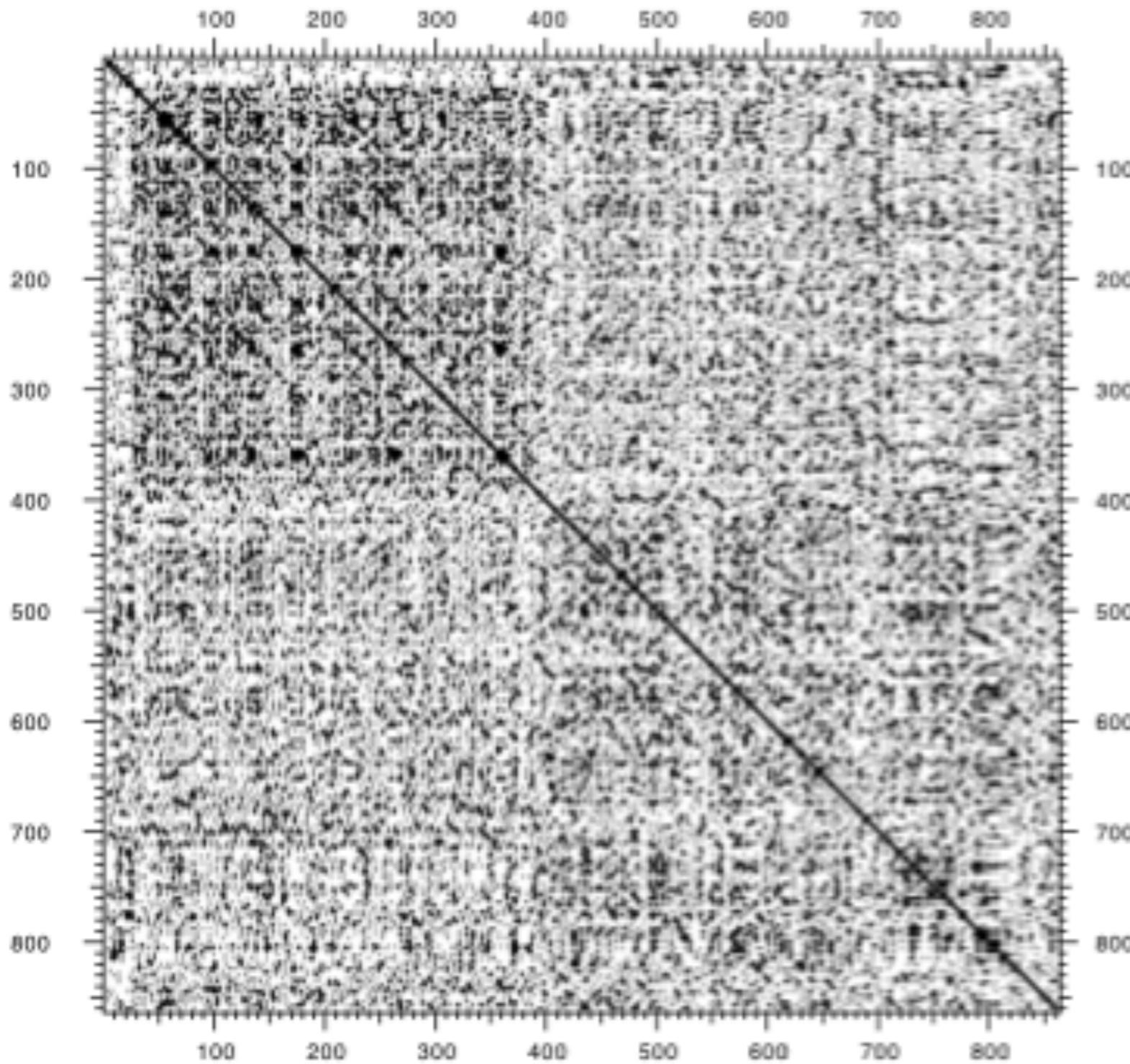
Only **diagonals** can be followed.

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

# Uses for dot matrices

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

# Repeats

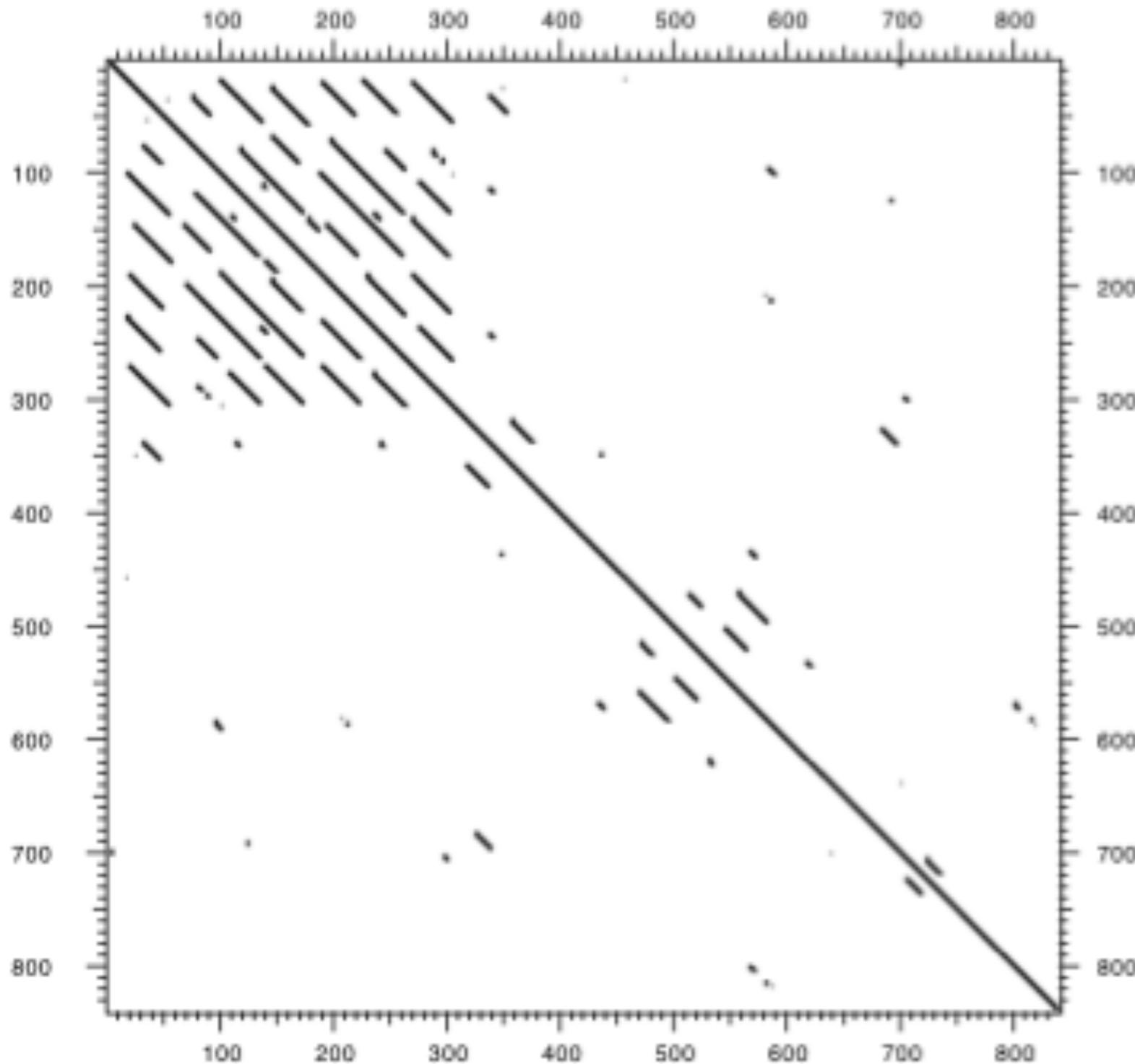


Human LDL receptor  
protein sequence  
(Genbank P01130)

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

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

# Repeats



Human LDL receptor  
protein sequence  
(Genbank P01130)

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

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

# Your Turn!

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

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

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

**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 2 3 4 5 6 7 8 9 10

**Moving window step size:**  1 2 3 4 5 6 7 8 9 10

**Match stringency:**  1 2 3 4 5 6 7 8 9 10

**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

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

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

# Scoring the alignment matrix

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

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

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

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

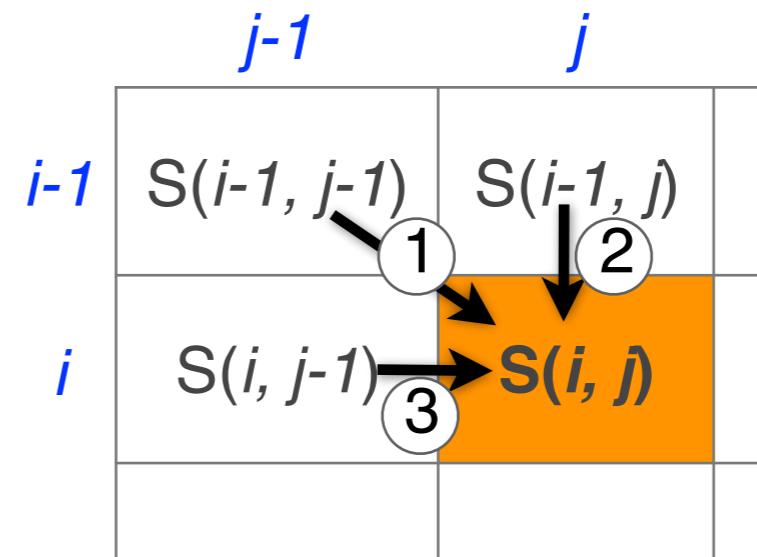
Seq1 : DPME  
Seq2 : ----

# Scoring the alignment matrix

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

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

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



# Scoring the alignment matrix

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

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

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

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

1  
2  
3

# Scoring the alignment matrix

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

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

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

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

# Scoring the alignment matrix

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

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

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

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

②  $(-4)+(-2) = -6$  Alignment  $D-DP$

③  $(1)+(-2) = -1$

# Scoring the alignment matrix

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

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

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

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

Alignment

②  $(-6)+(-2) = -8$

D--  
DPL

③  $(-1)+(-2) = -3$

# Scoring the alignment matrix

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

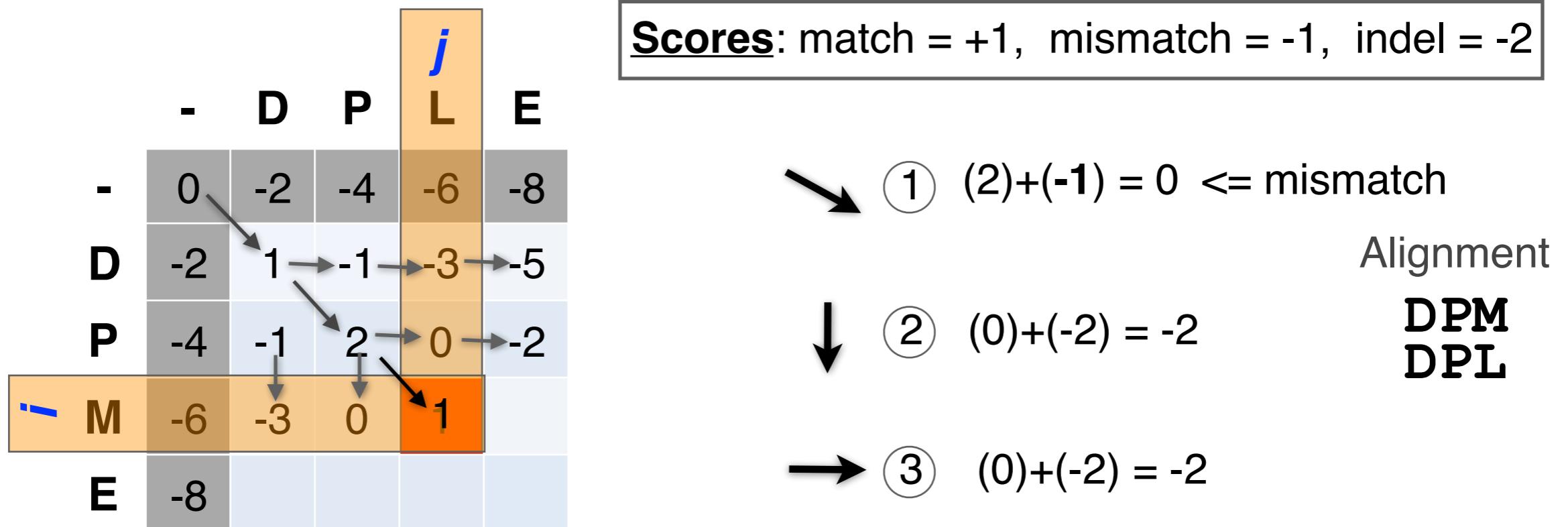
		j				
		-	D	P	L	E
-		0	-2	-4	-6	-8
D	-2	1	-1	-3	-5	
P	-4	-1	2	0		
M	-6					
E	-8					

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

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

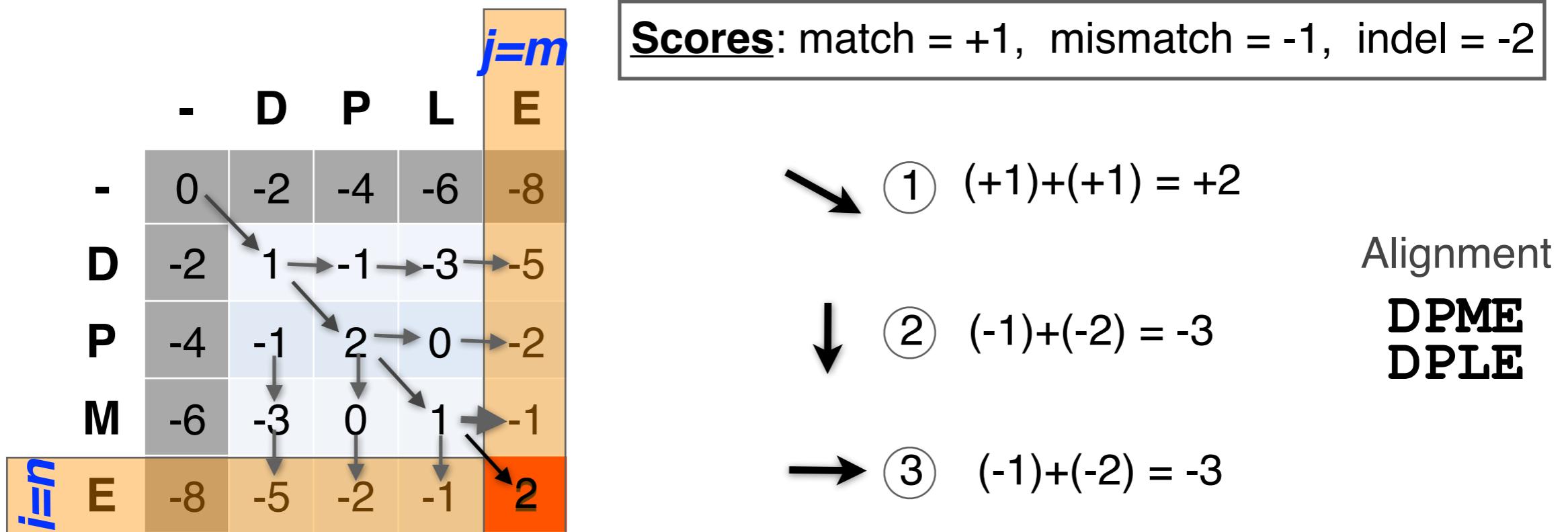
# Scoring the alignment matrix

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



# Scoring the alignment matrix

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



# Scoring the alignment matrix

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

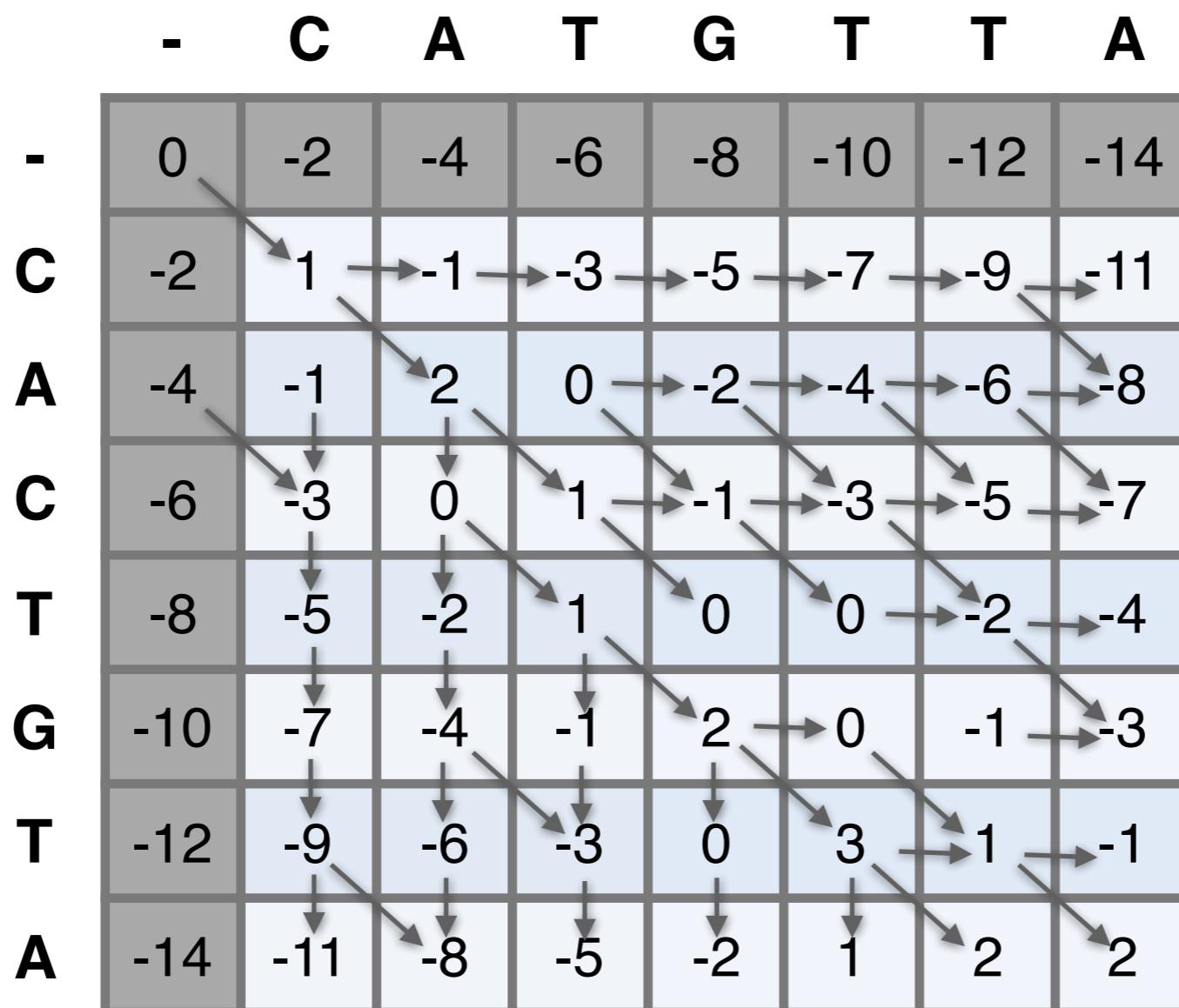
**Scores:** match = +1, mismatch = -1, indel = -2

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

Alignment  
**DPME**  
**DPLE**

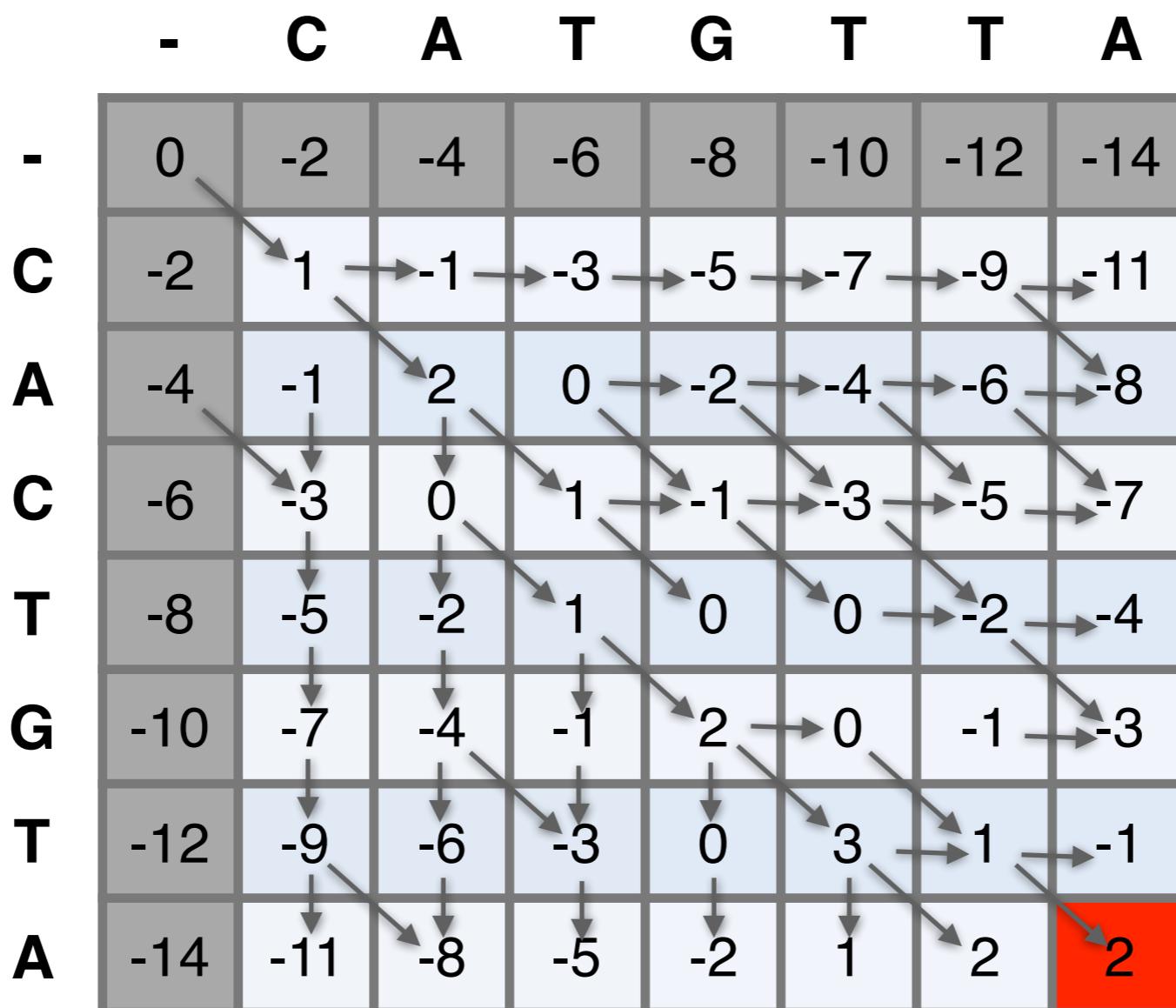
# Questions:

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



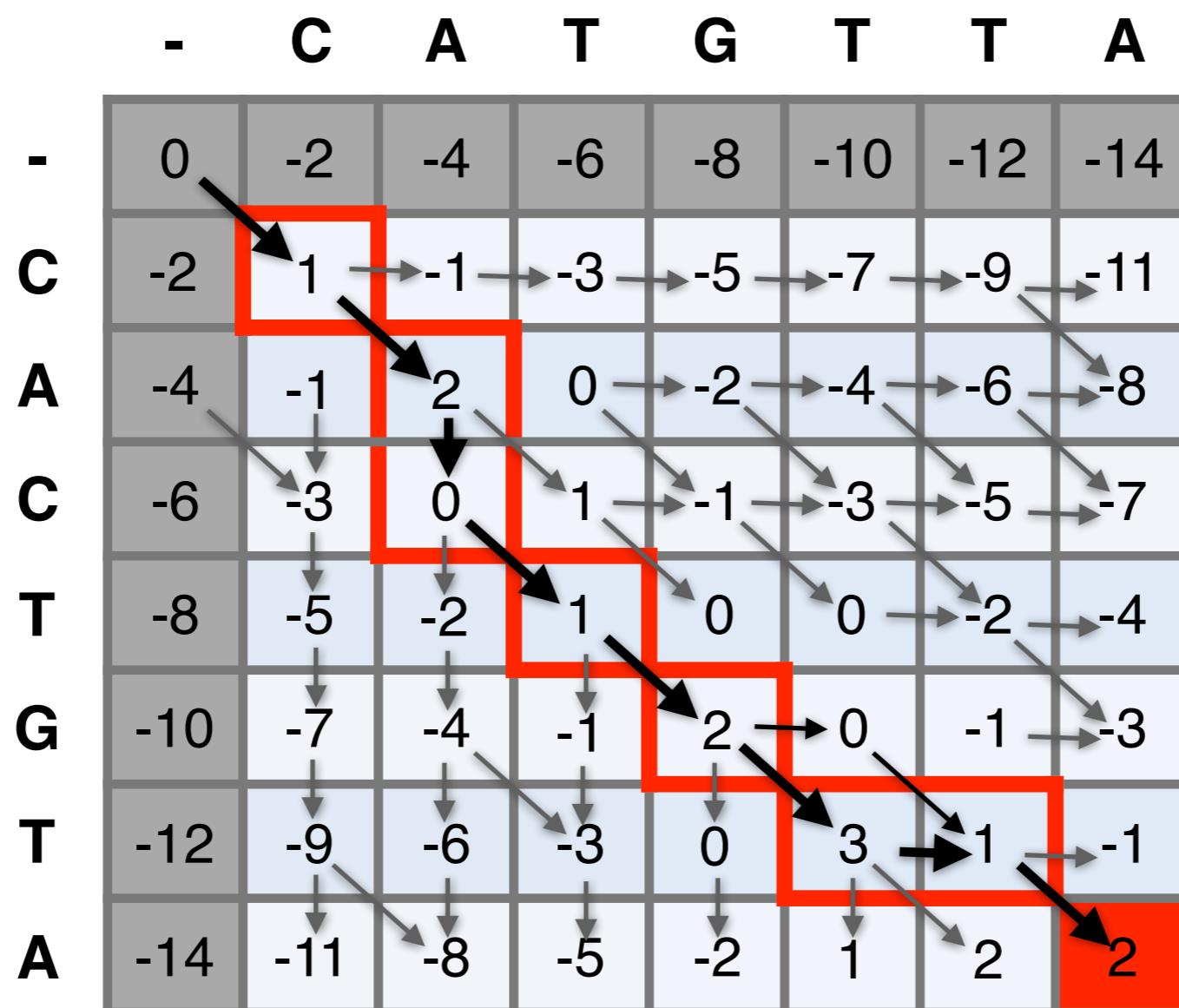
# Questions:

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



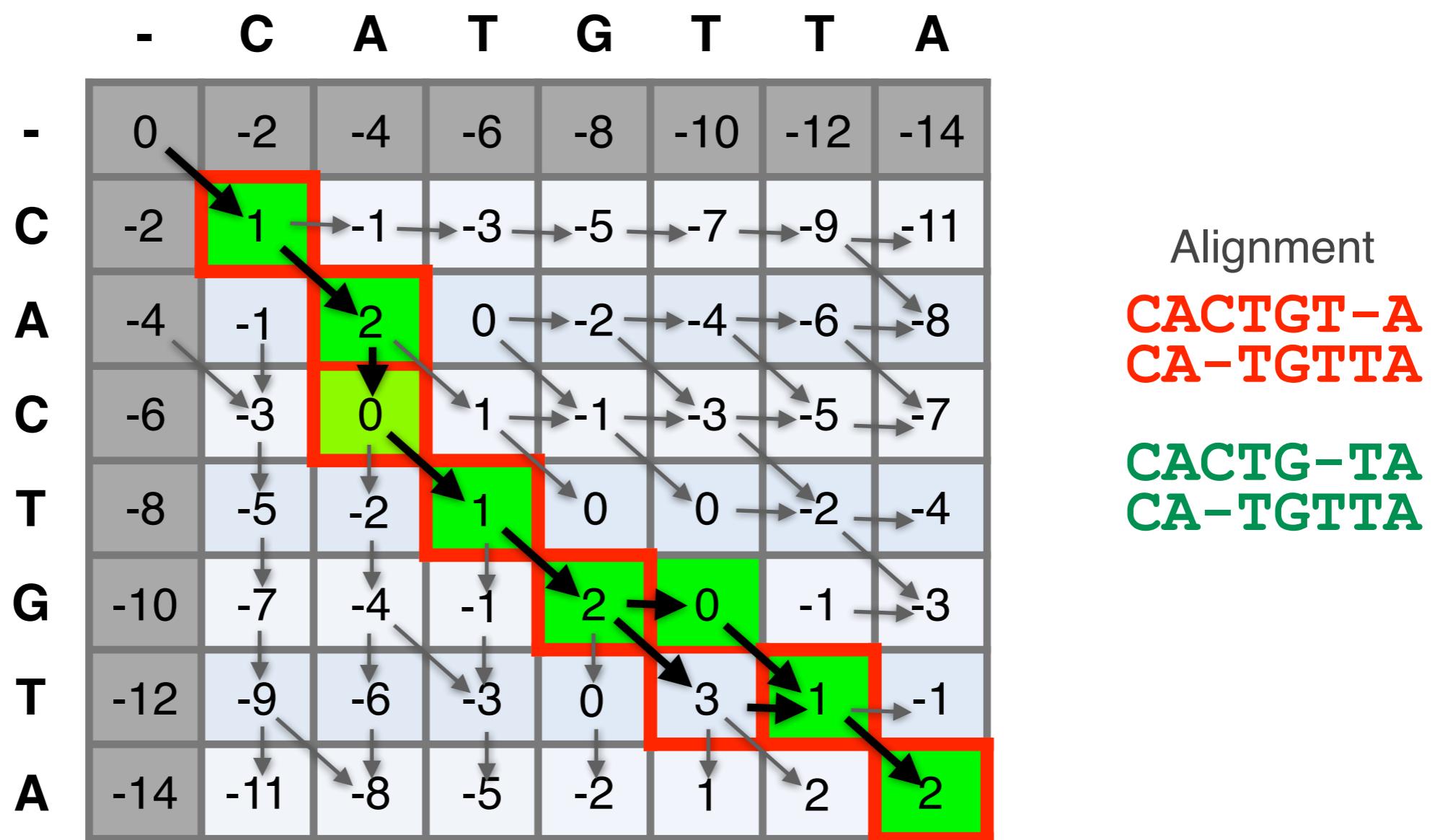
# Questions:

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



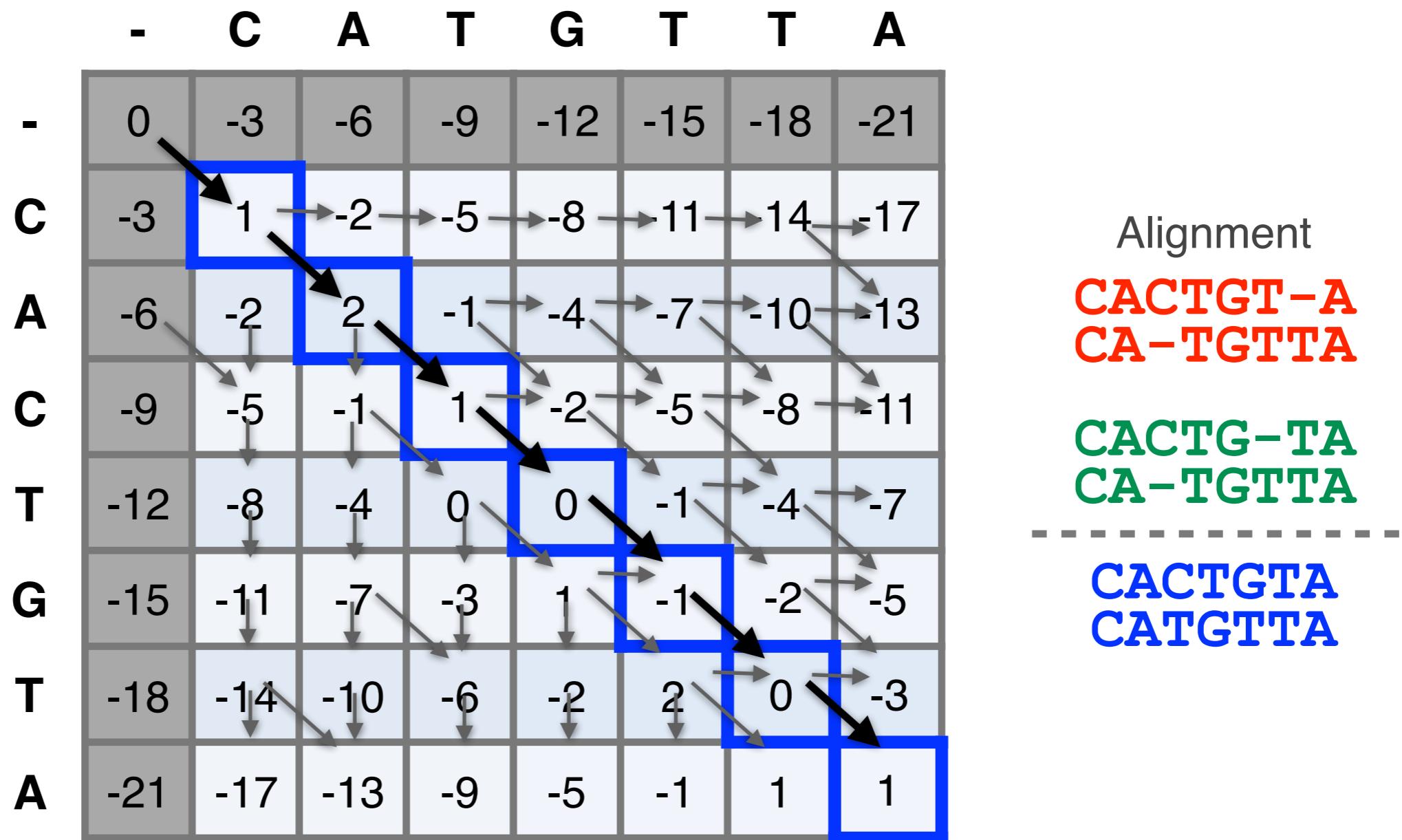
# More than one alignment possible

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



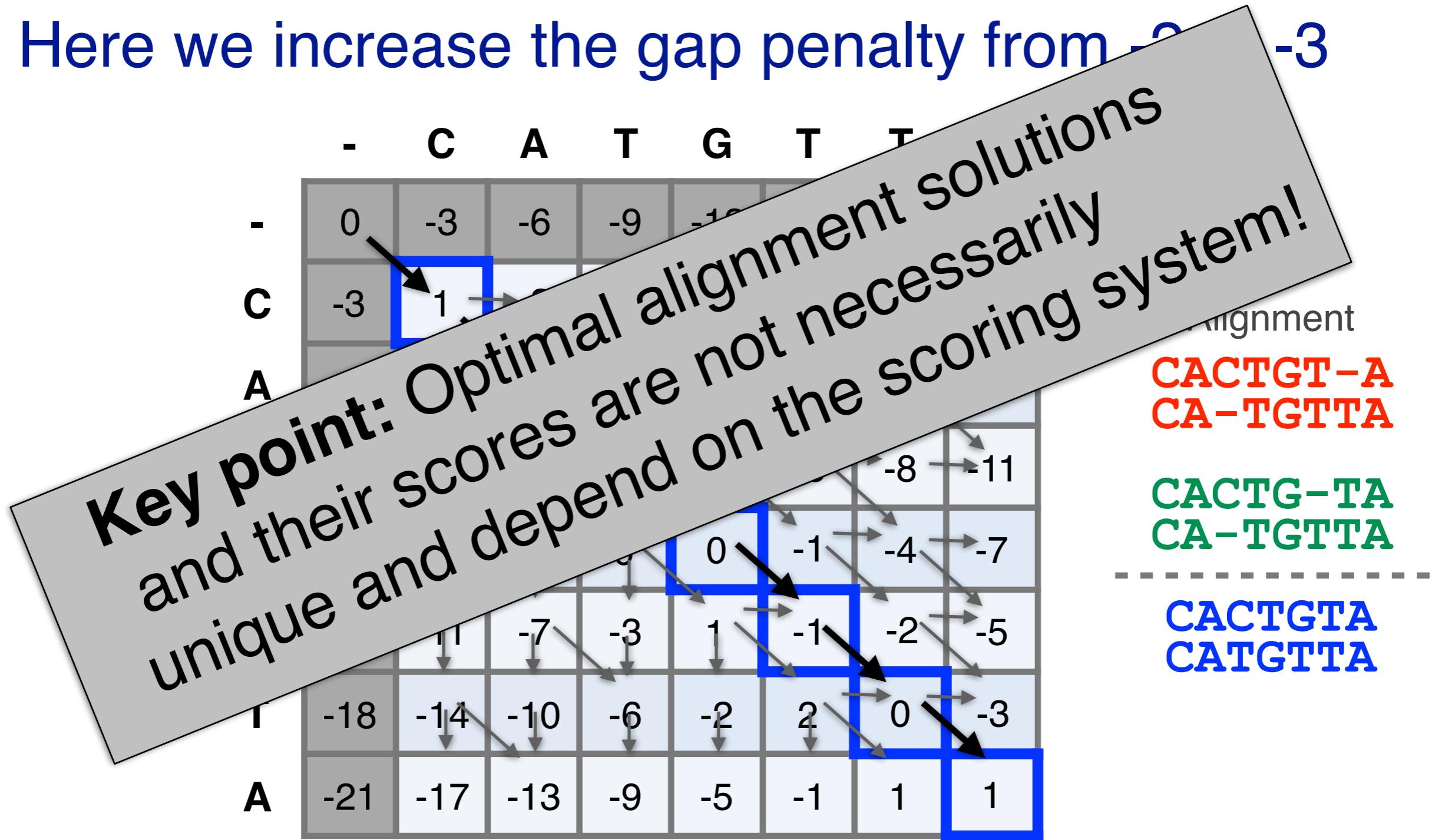
# The alignment and score are dependent on the scoring system

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



# The alignment and score are dependent on the scoring system

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



# NW DYNAMIC PROGRAMMING

Match: +2

Mismatch: -1

Gap: -2

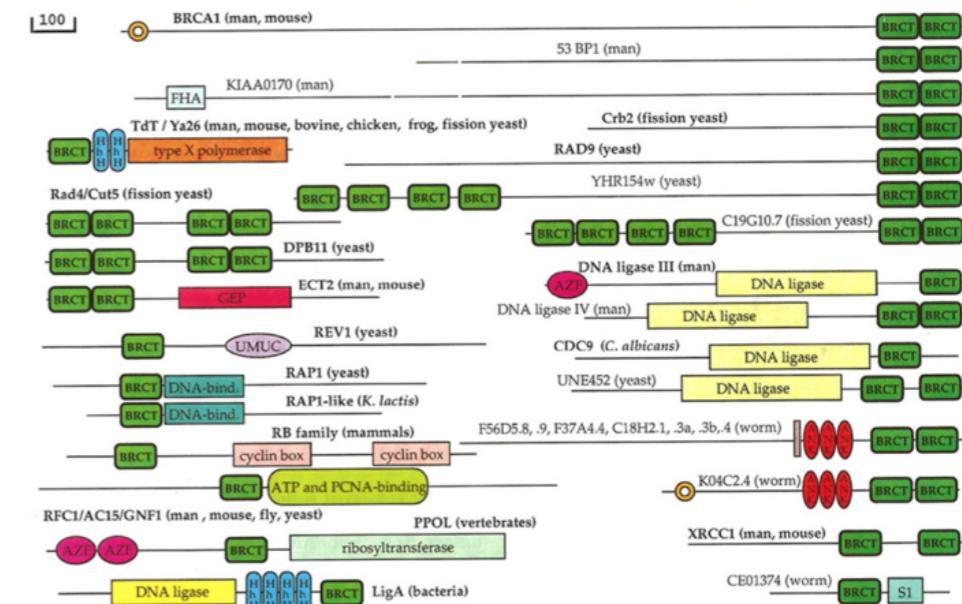
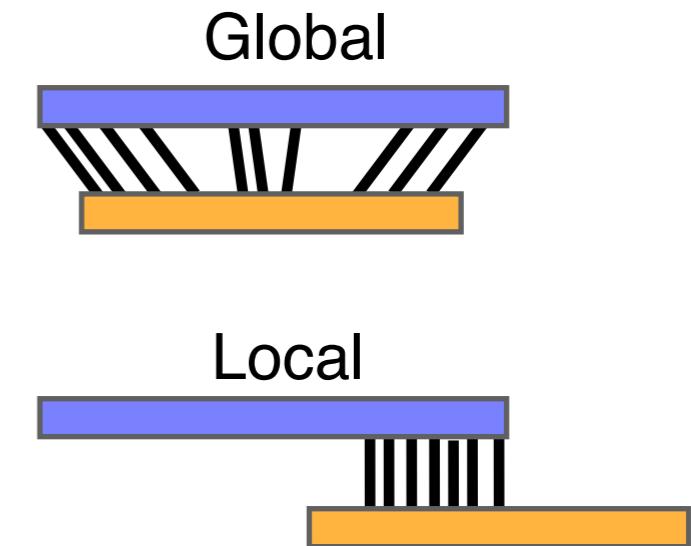
	A	G	T	T	C
A	0				
T					
T					
G					
C					

# 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

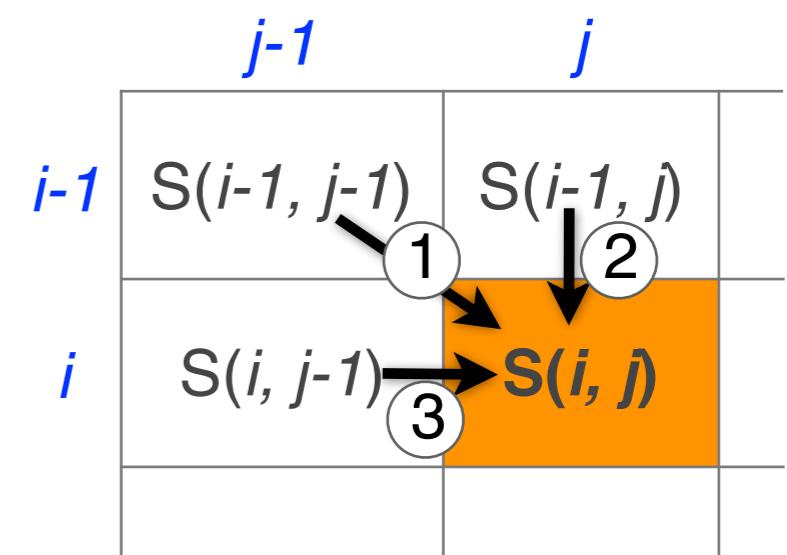
Smith, T.F. & Waterman, M.S. (1981) “Identification of common molecular subsequences.” J. Mol. Biol. 147:195-197.

# 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

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

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



Sequence 1

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

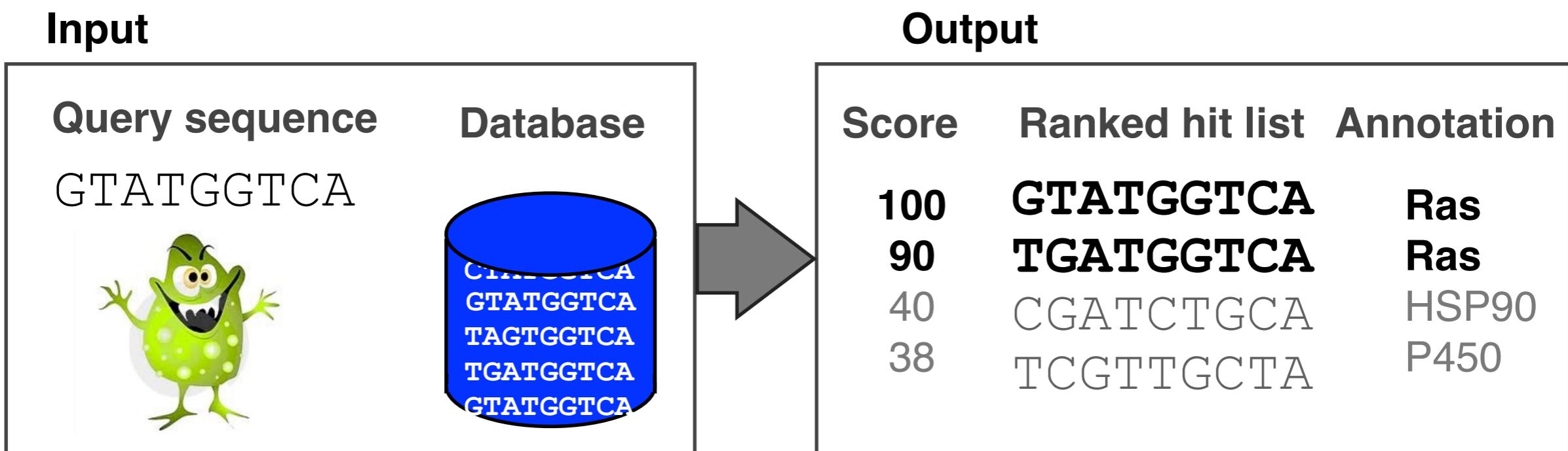
Local alignment

**GCC-AUG**

**GCCUCGC**

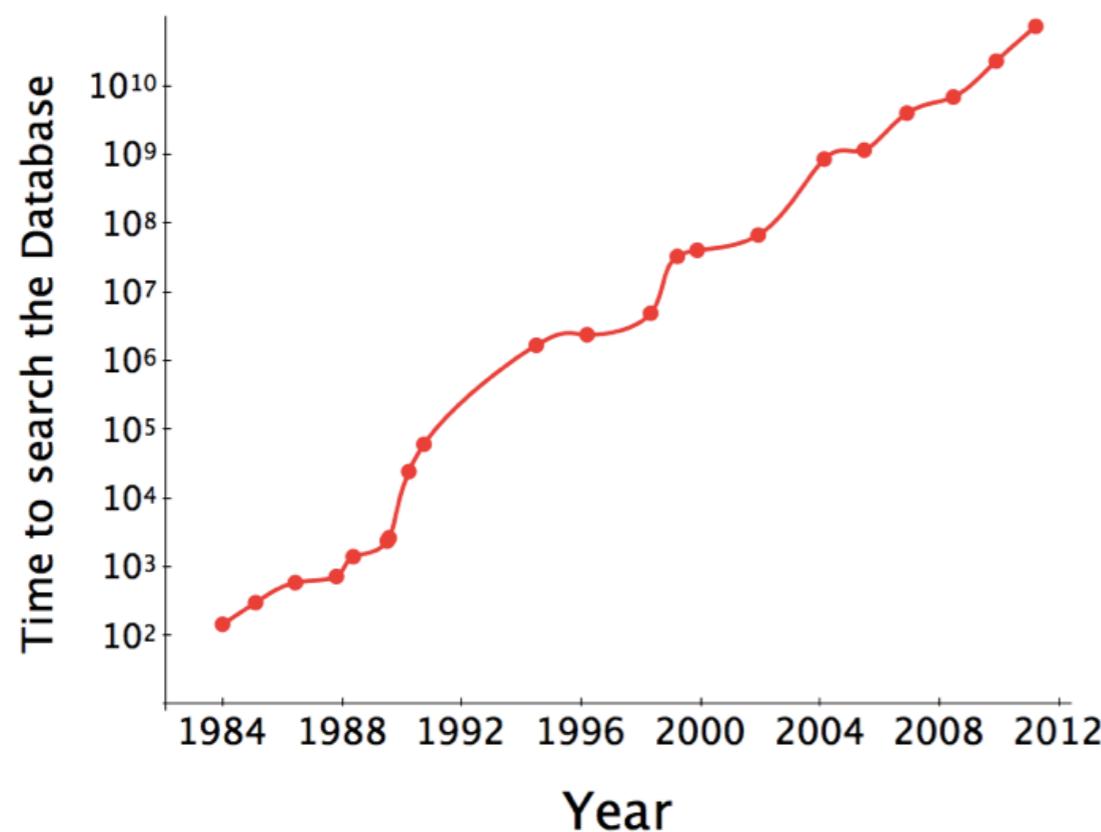
# Local alignments can be used for database searching

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



# The database search problem

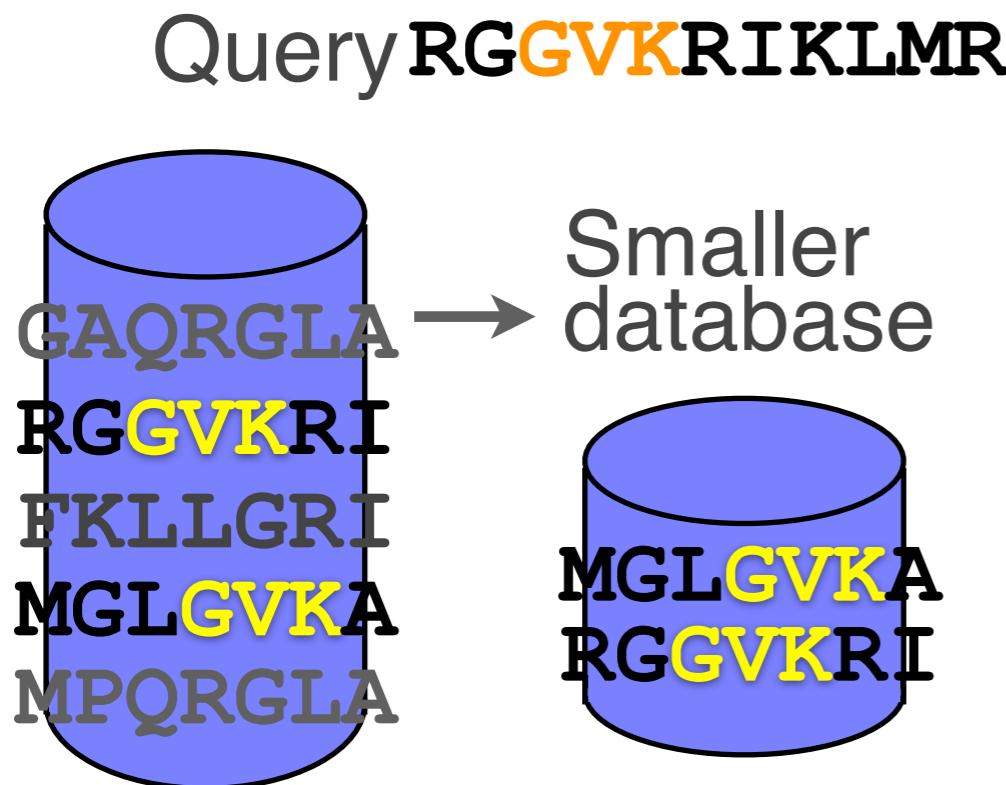
- Due to the rapid growth of sequence databases, search algorithms have to be both efficient and sensitive
  - Time to search with SW is proportional to  $m \times n$  ( $m$  is length of query,  $n$  is length of database), **too slow for large databases!**



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

# The database search problem

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



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

# ALIGNMENT FOUNDATIONS

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

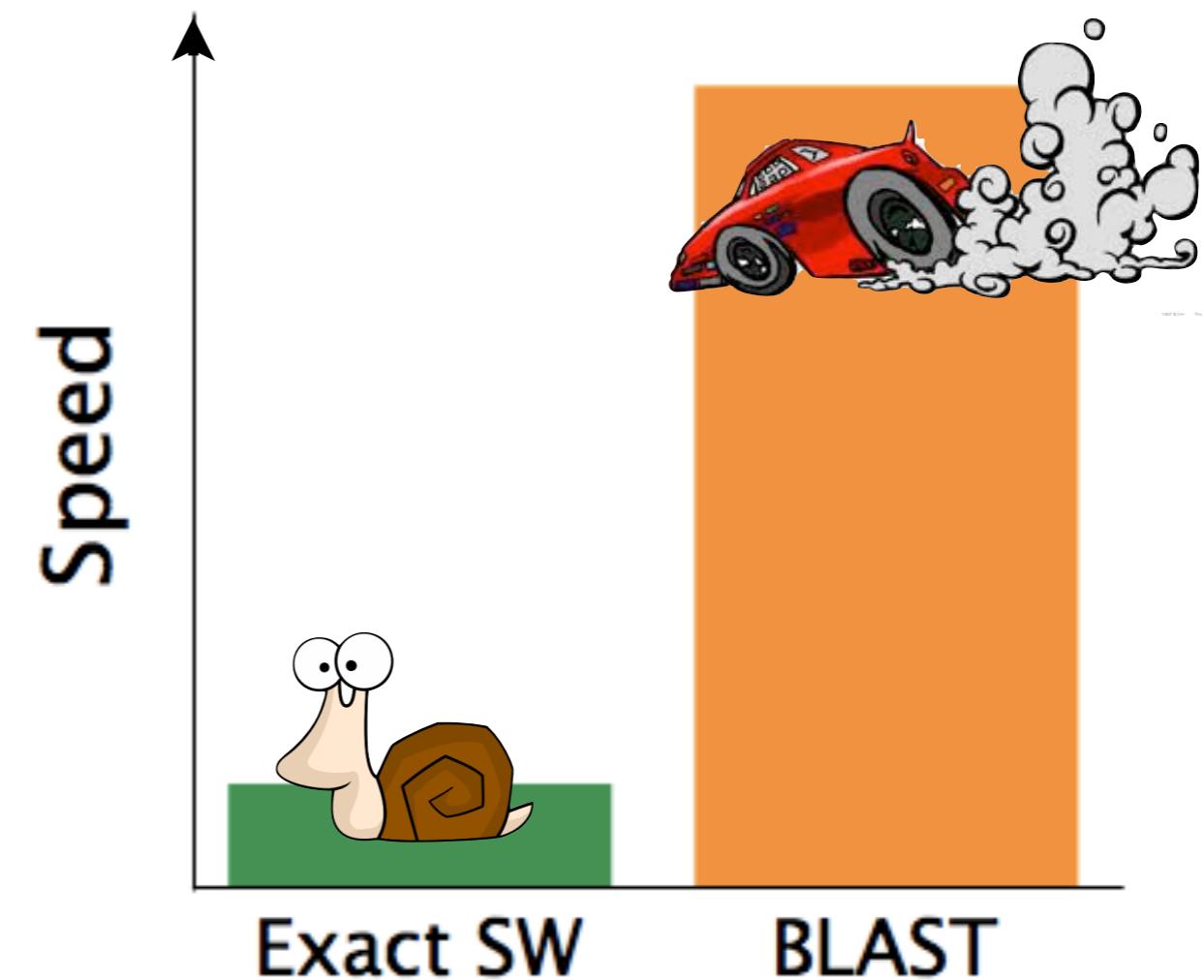
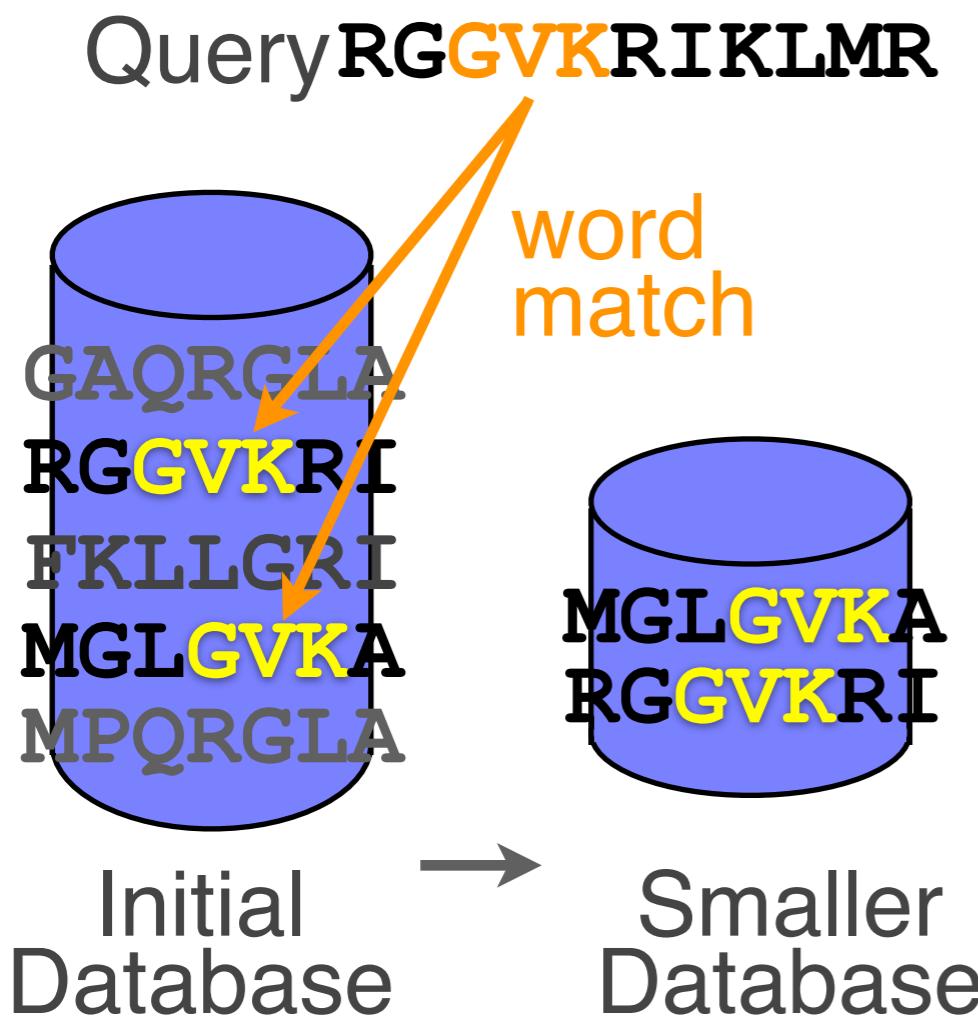
# Rapid, heuristic versions of Smith–Waterman: BLAST

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

# Rapid, heuristic versions of Smith–Waterman: BLAST

- BLAST (Basic Local Alignment Search Tool) is a simplified form of Smith-Waterman. It is popular because it is **fast**.
    - BLAST finds regions of similarity between sequences
    - BLAST does not find all matches, but it finds many matches by scanning for local alignments
- “The central idea of the BLAST algorithm is to confine attention to sequence pairs that contain an initial **word pair match**”  
Altschul et al. (1990)
- The sensitivity in exchange for speed  
In contrast to SW, BLAST is not guaranteed to find optimal alignments

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



# How BLAST works

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

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

**RGGVKRI**      Query sequence

**RGG**

**GGV**

**GVK**

**VKR**

**KRI**

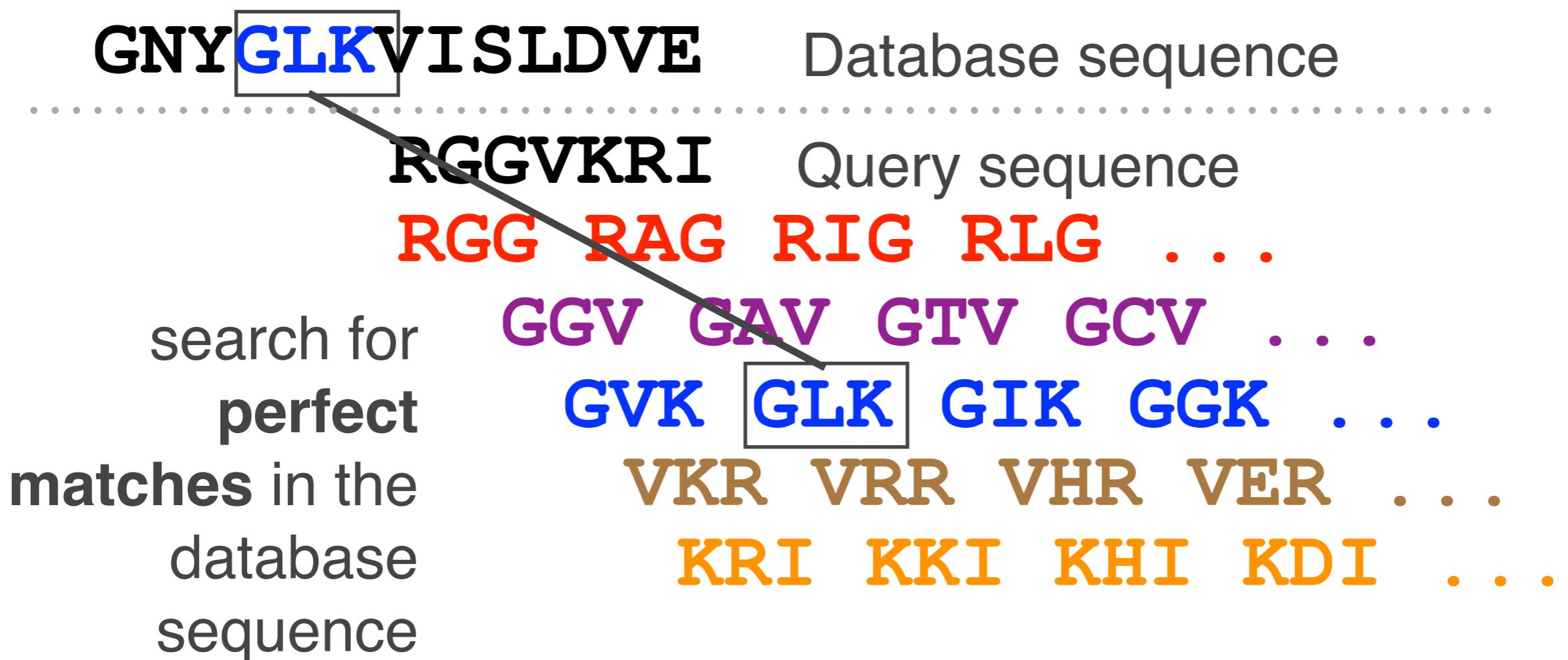
# Blast

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

	<b>RGGVKRI</b>	Query sequence
extend list of words similar to query	<b>RGG RAG RIG RLG . . .</b>	
	<b>GGV GAV GTV GCV . . .</b>	
	<b>GVK GAK GIK GGK . . .</b>	
	<b>VKR VRR VHR VER . . .</b>	
	<b>KRI KKI KHI KDI . . .</b>	

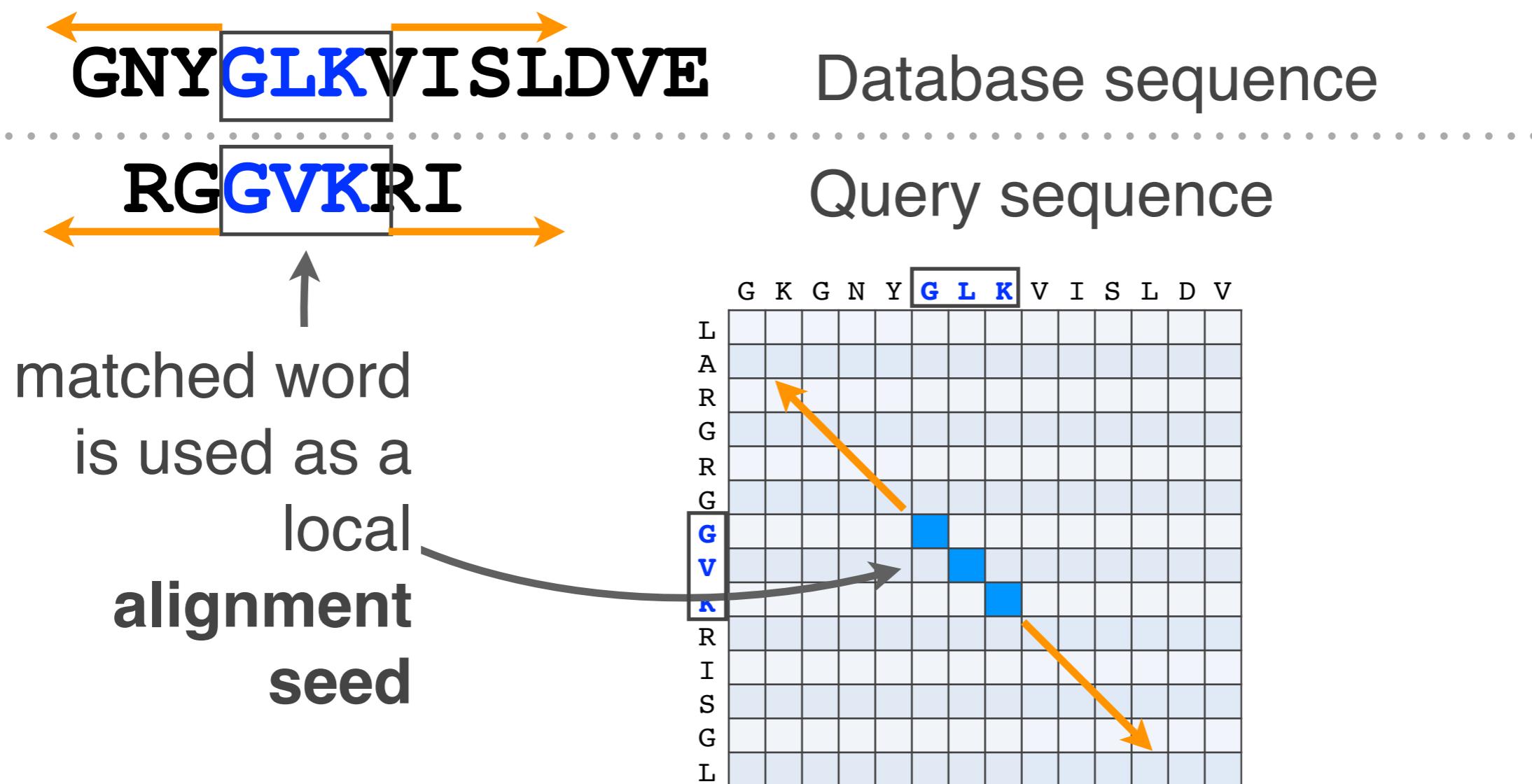
# Blast

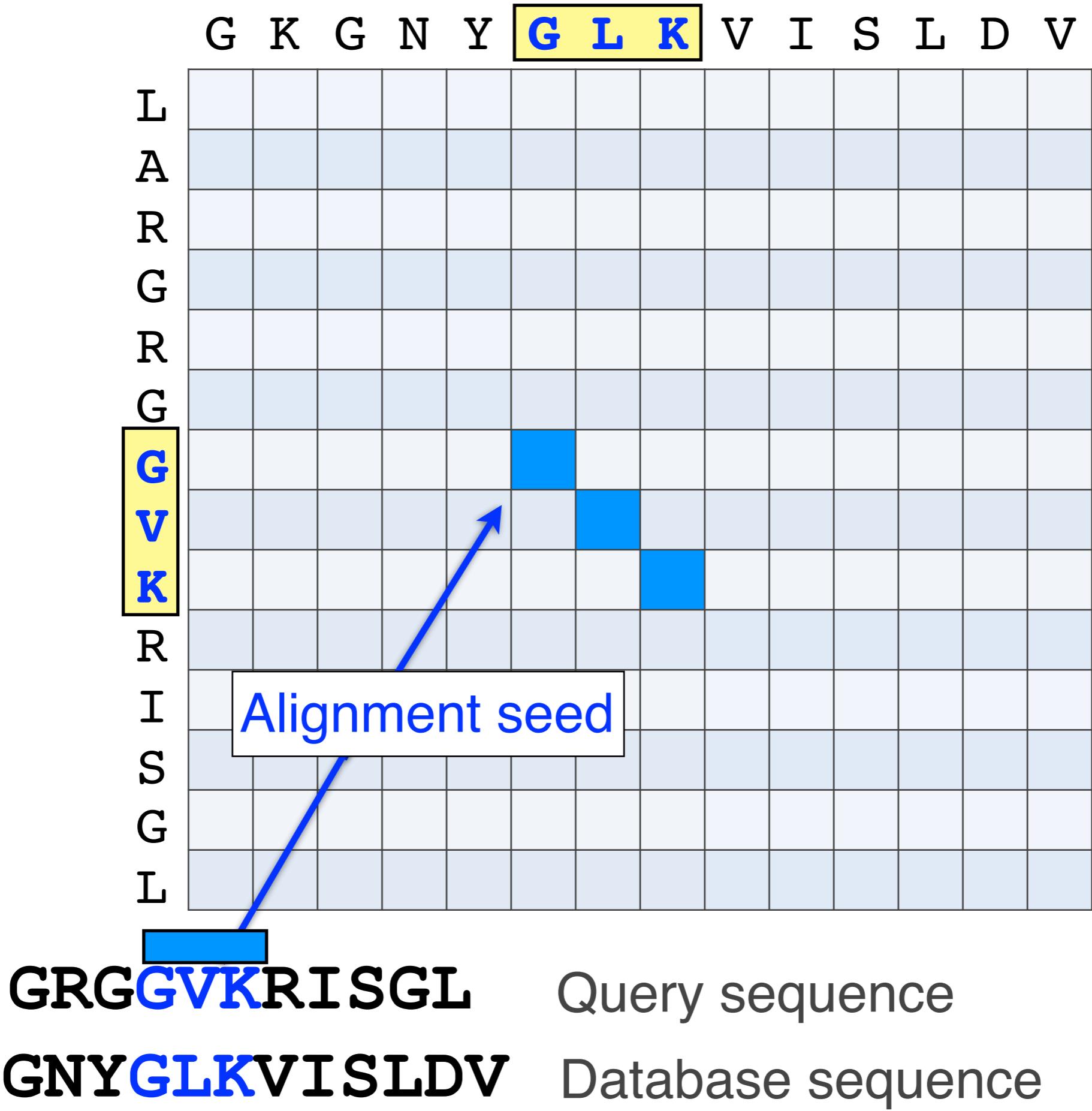
- Phase 3: a database is scanned to find sequence entries that match the compiled word list

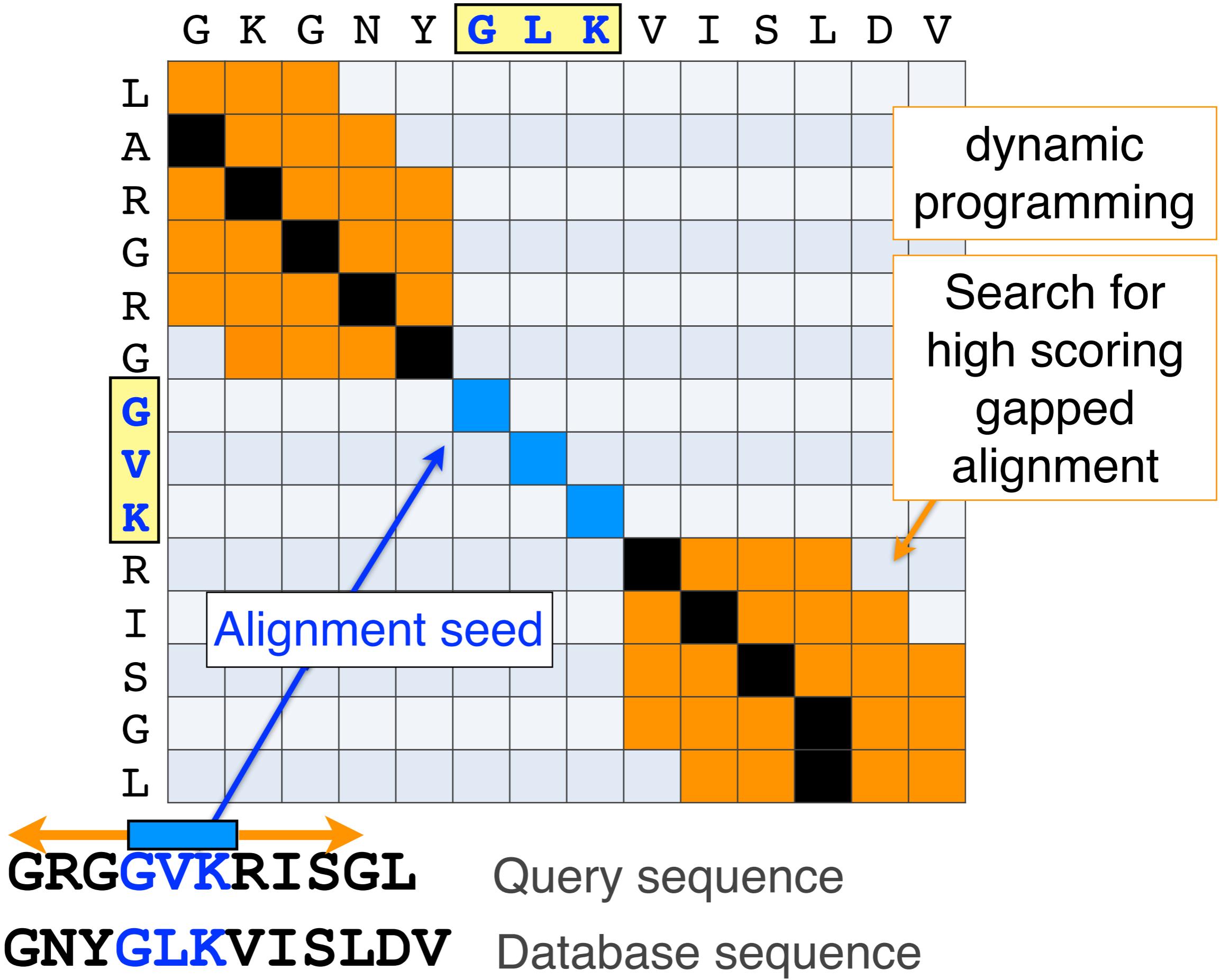


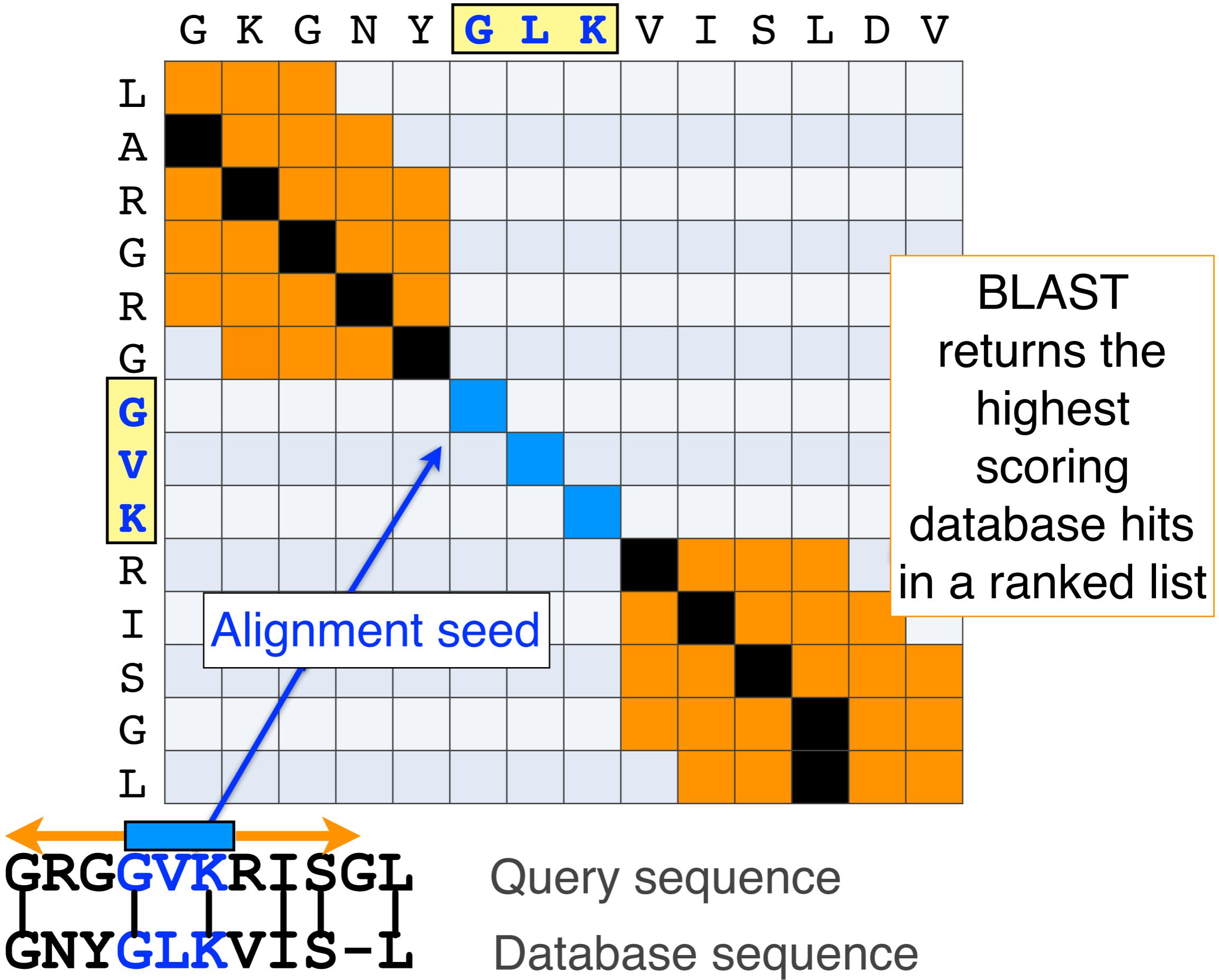
# Blast

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









# BLAST output

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

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

# Statistical significance of results

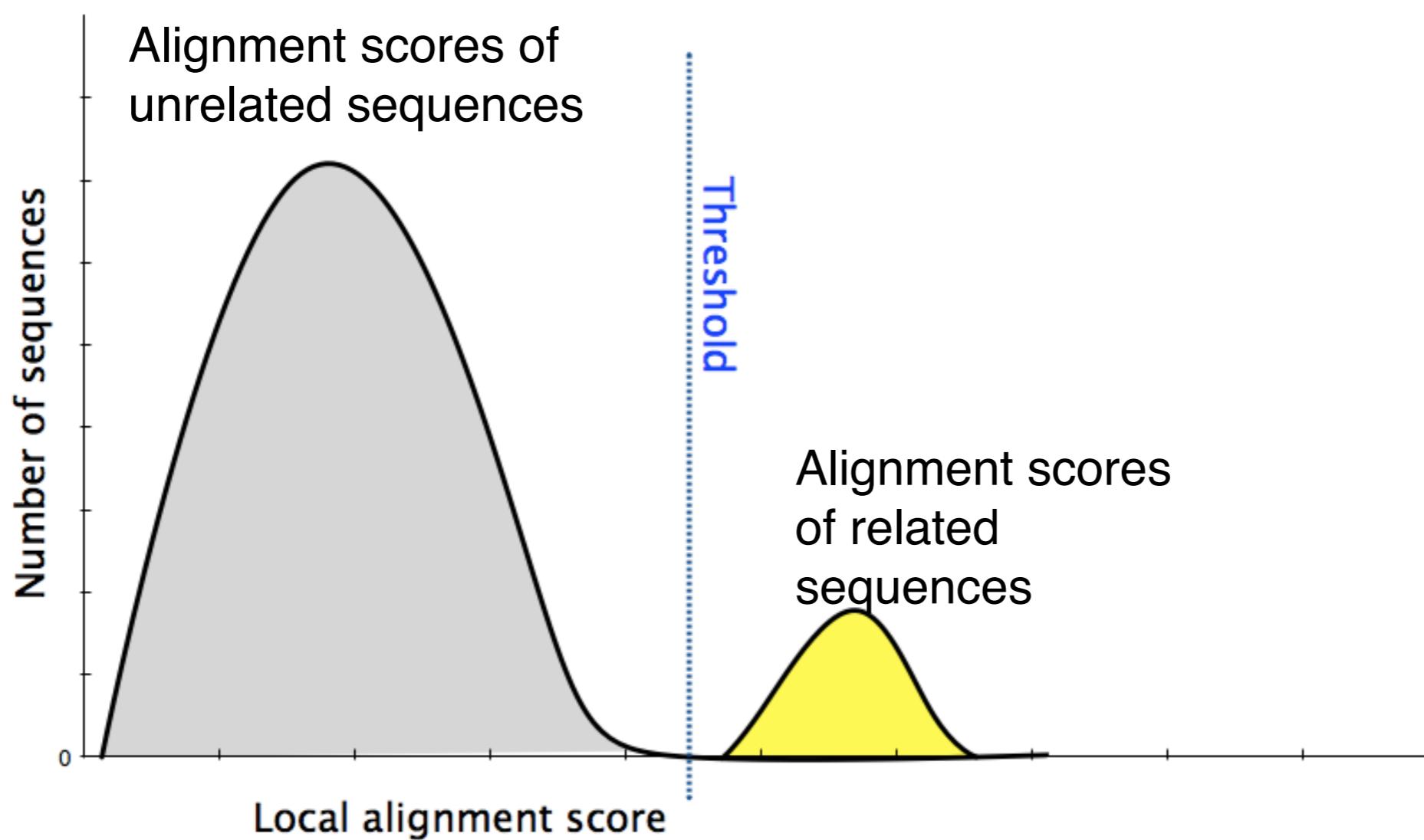
- An important feature of BLAST is the computation of statistical significance for each hit. This is described by the **E value** (expect value)

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

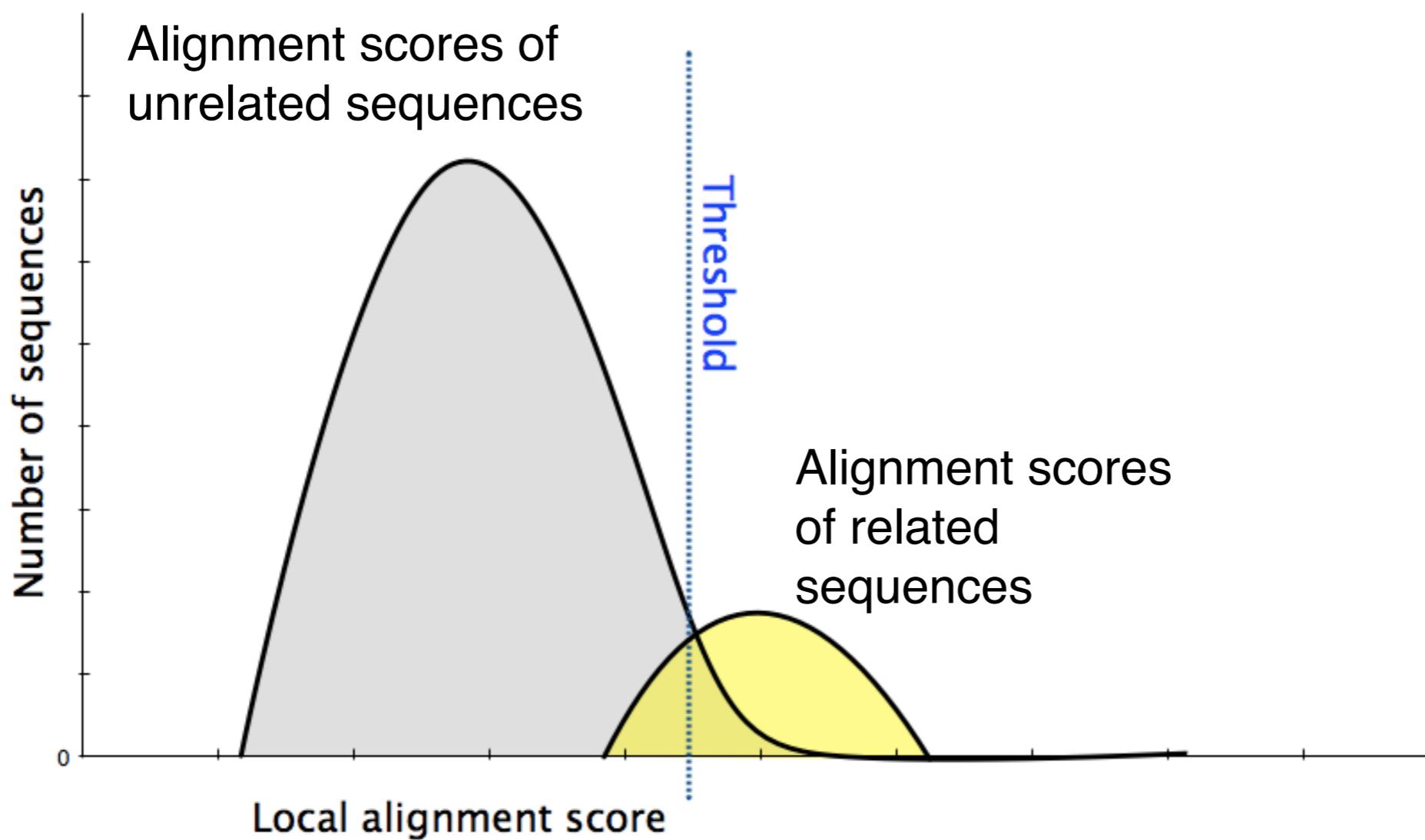
# BLAST scores and E-values

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

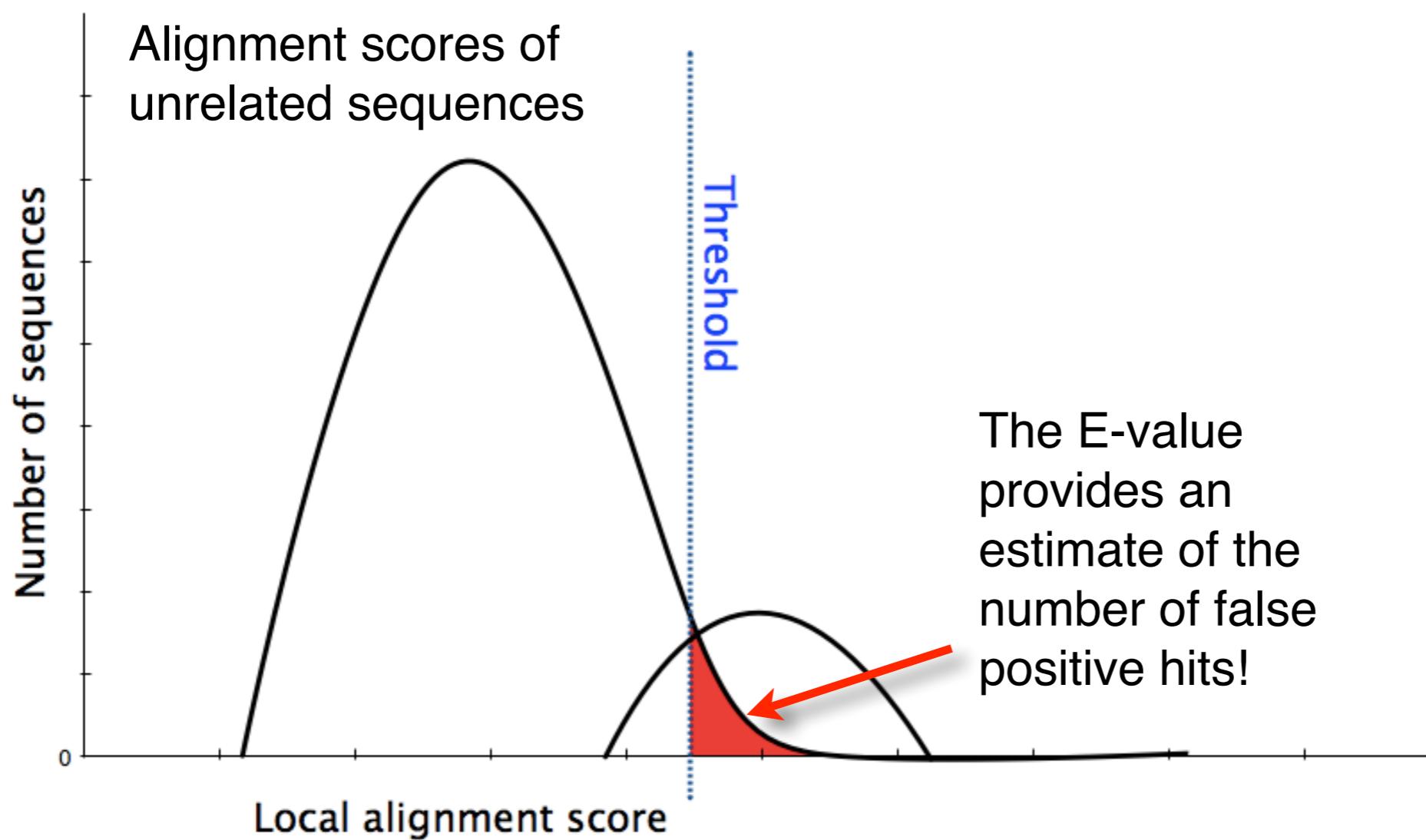
- Ideally, a threshold separates all query related sequences (yellow) from all unrelated sequences (gray)



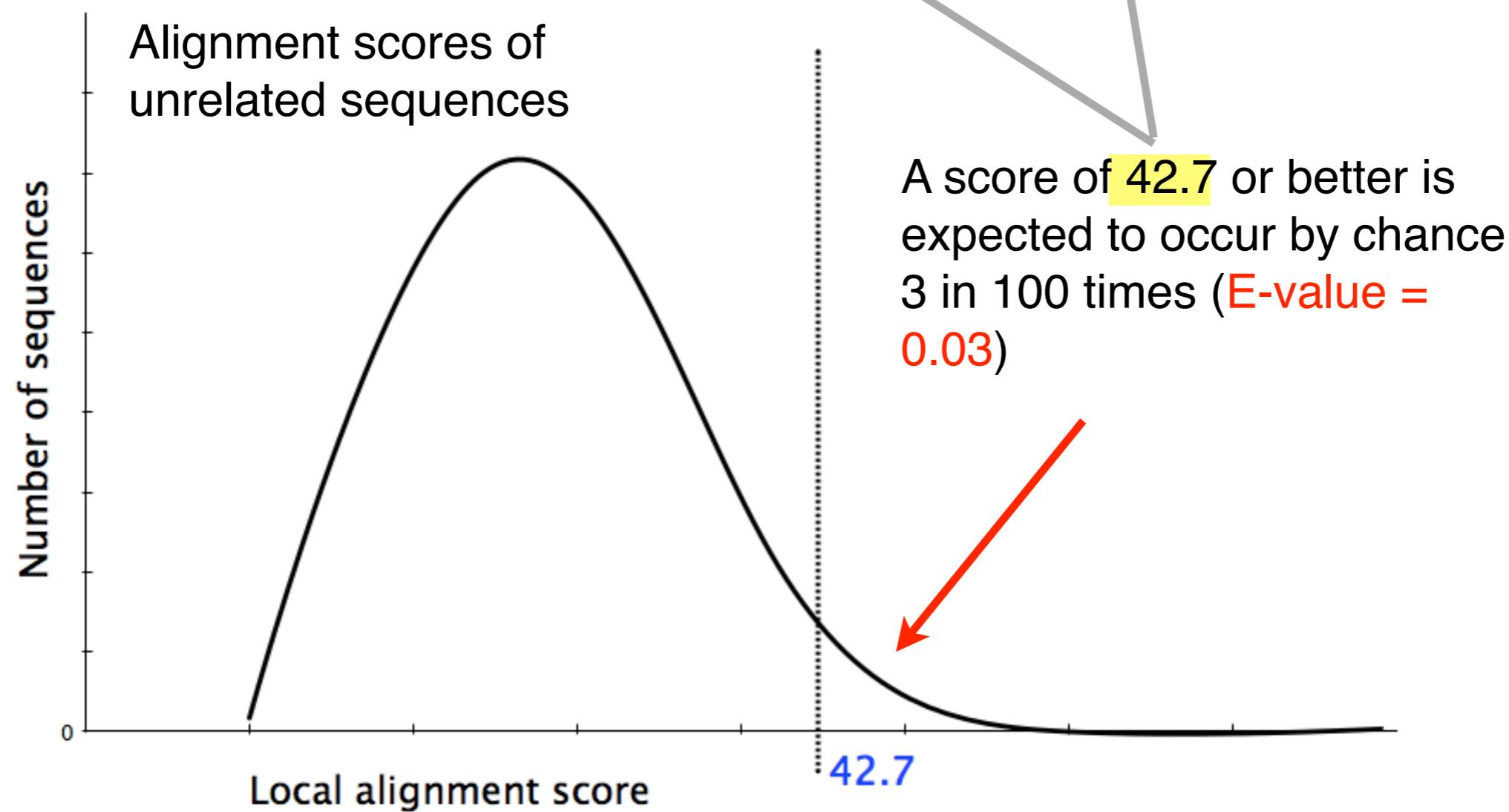
- Unfortunately, often both score distributions overlap
  - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



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



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

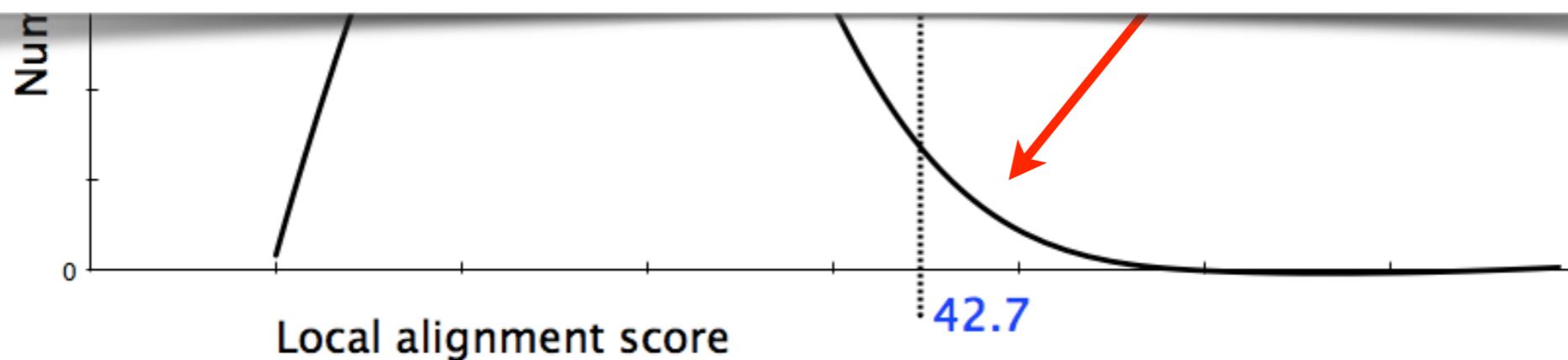


Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo sapiens]	677	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	676	100%	0	98%	AAA20133.1

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

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

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

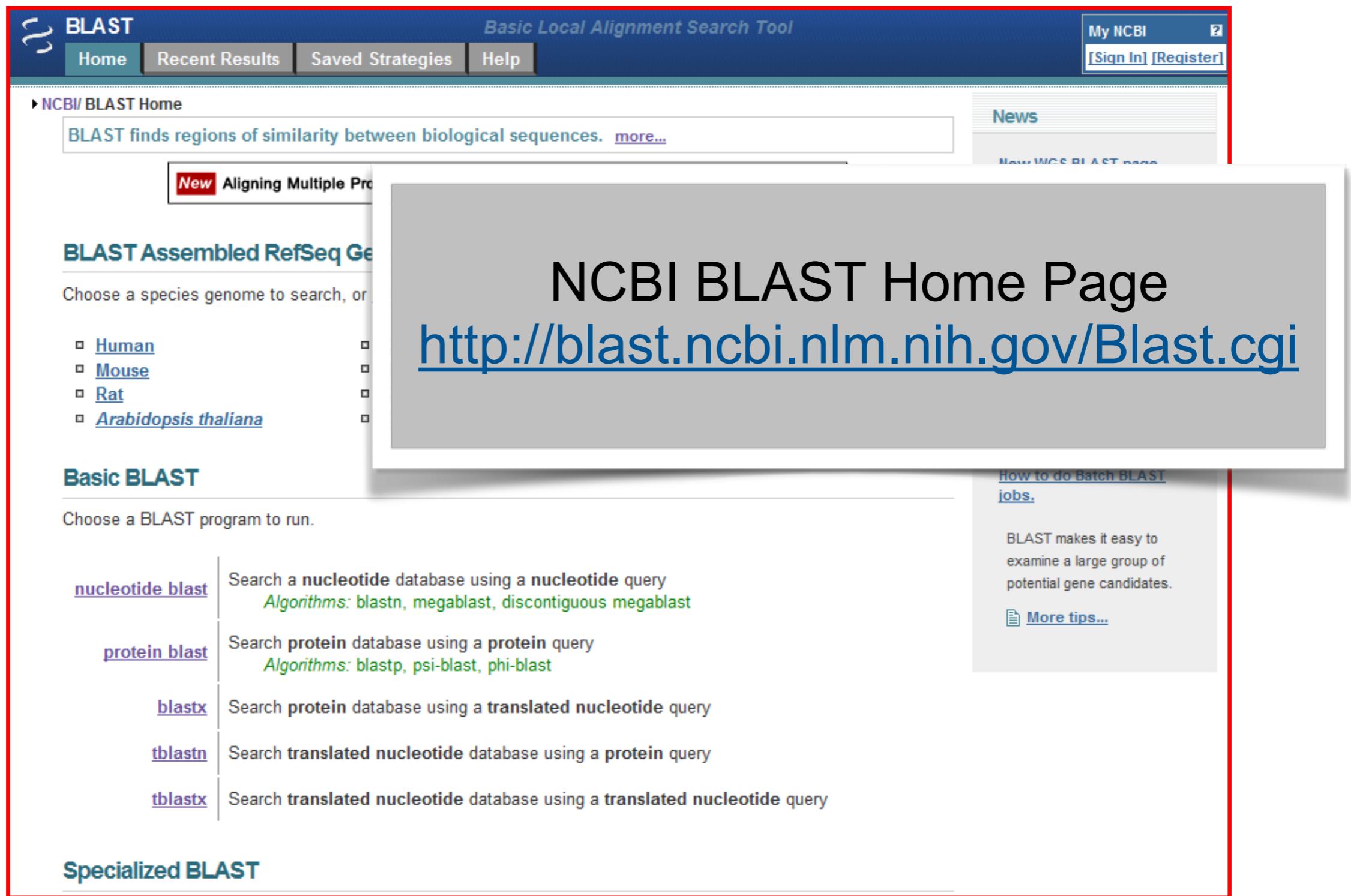


# Your Turn!

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

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

# Practical database searching with BLAST



The screenshot shows the NCBI BLAST Home Page. The top navigation bar includes links for Home, Recent Results, Saved Strategies, Help, My NCBI, Sign In, and Register. A banner at the top states "BLAST finds regions of similarity between biological sequences." Below this, a "Basic BLAST" section allows users to choose a species genome (Human, Mouse, Rat, Arabidopsis thaliana) and a BLAST program (nucleotide blast, protein blast, blastx, tblastn, tblastx). To the right, a large callout box highlights the NCBI BLAST Home Page URL: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. A sidebar on the right provides tips for batch BLAST jobs.

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

**Basic BLAST**

Choose a BLAST program to run.

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

**Specialized BLAST**

[How to do Batch BLAST jobs.](#)  
BLAST makes it easy to examine a large group of potential gene candidates.  
[More tips...](#)

# 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”

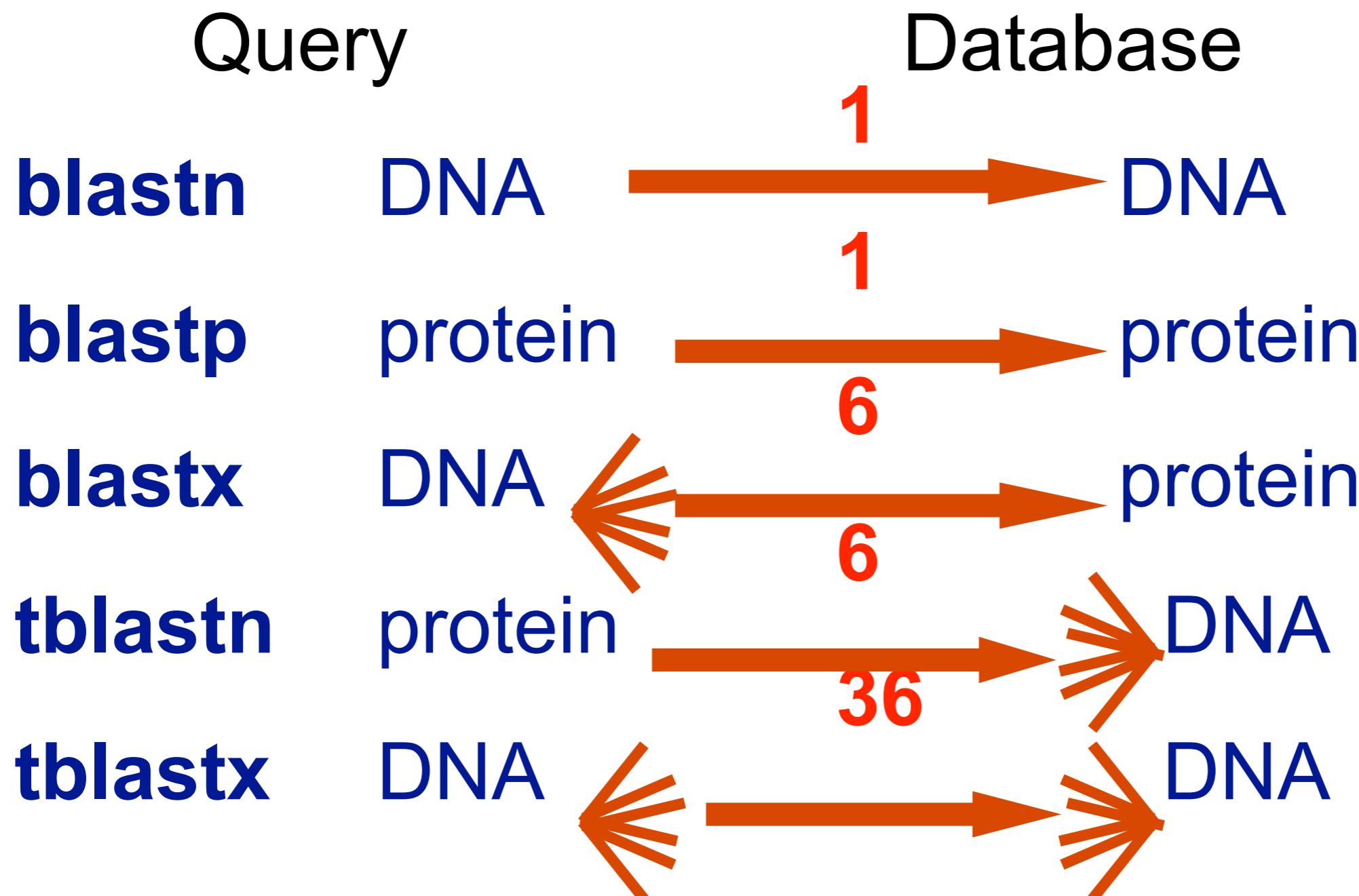
# Step 1: Choose your sequence

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

The screenshot shows the NCBI Protein search results for "hemoglobin subunit beta [Homo sapiens]". At the top, there's a navigation bar with links for NCBI, Resources, How To, Limits, Advanced search, Help, and a My N link. Below the navigation is a search bar with "Search: Protein" and a dropdown menu. The main content area displays the protein information with the following details:

- Display Settings:** A checkbox labeled "FASTA" is checked and circled in red.
- NCBI Reference Sequence:** A link labeled "NP\_000509.1" is circled in red.
- Protein Name:** hemoglobin subunit beta [Homo sapiens]
- NCBI Reference Sequence:** NP\_000509.1
- GenPept:** GenPept
- Graphics:** Graphics
- Sequence:** >gi|4504349|ref|NP\_000509.1| hemoglobin subunit beta [Homo sapiens]  
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKVKAHGKKVLG  
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHHFGKEFTPPVQAAQKVVAGVAN  
ALAHKYH
- Buttons and Links on the right:**
  - Send to: (dropdown menu)
  - Change region shown
  - Analyze this sequence
  - Run BLAST
  - Identify Conserved Domains
  - Find in this Sequence

## Step 2: Choose the BLAST program



# DNA potentially encodes six proteins

5' CAT CAA

5' ATC AAC

5' TCA ACT



5' CATCAACTACAACCTCAAAGACACCCCTTACACATCAACAAACCTACCCAC 3'

3' GTAGTTGATGTTGAGGTTCTGTGGGAATGTGTAGTTGTTGGATGGGTG 5'

5' GTG GGT

5' TGG GTA

5' GGG TAG



Protein BLAST: search protein databases using a protein query

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST\_PROGRAMS=blastp&PAGE\_TYPE=Blast+Search

**Enter Query Sequence**

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

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

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

Job Title

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

Align two or more sequences [?](#)

**Choose Search Set**

Database [Non-redundant protein sequences \(nr\)](#) [?](#)

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

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

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

**Program Selection**

Algorithm  blastp (protein-protein BLAST)  
 PSI-BLAST (Position-Specific Iterated BLAST)  
 PHI-BLAST (Pattern Hit Initiated BLAST)  
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)  
Choose a BLAST algorithm [?](#)

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

[Algorithm parameters](#)



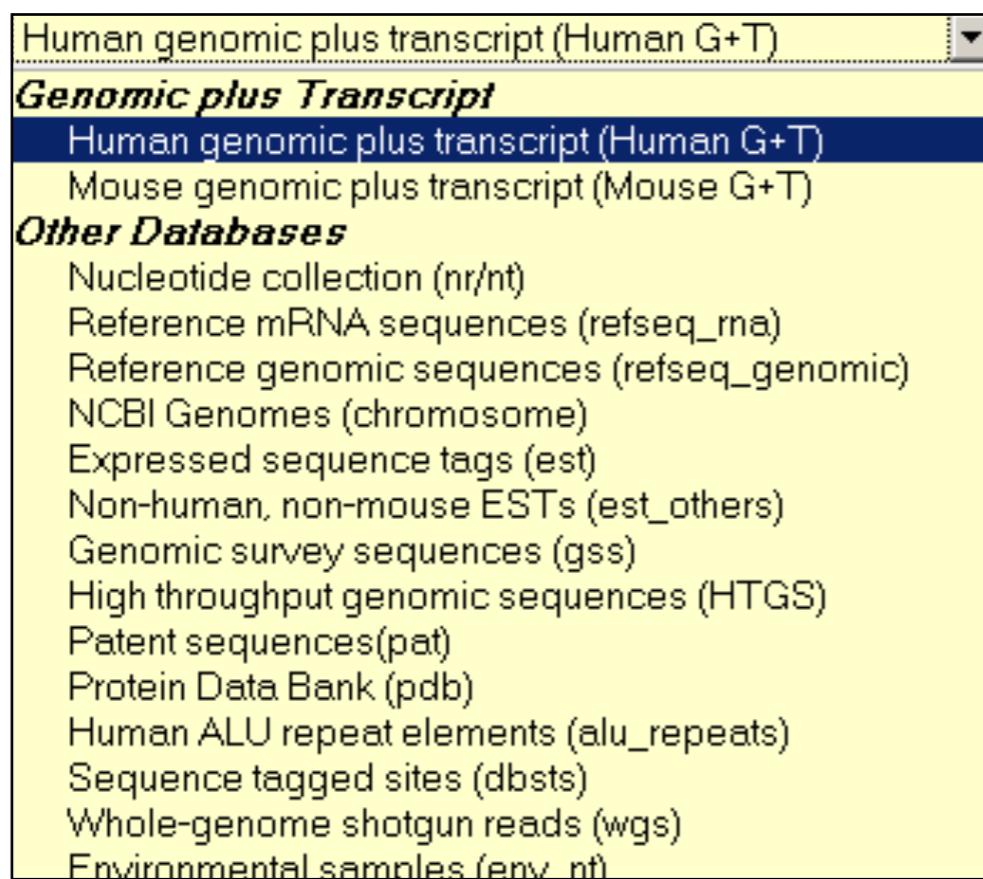
# Step 3: Choose the database

nr = non-redundant (most general database)

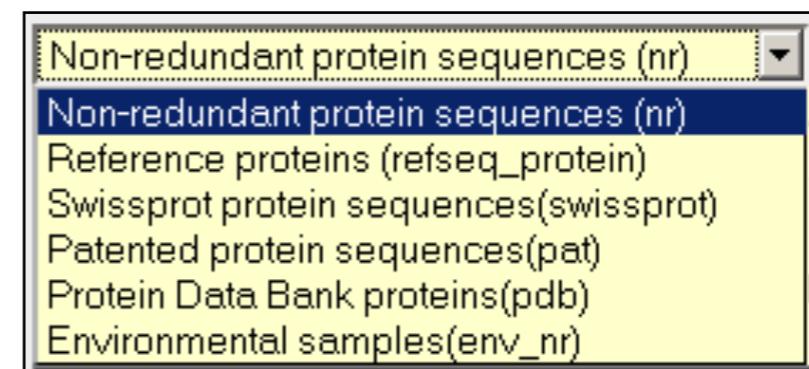
dbest = database of expressed sequence tags

dbsts = database of sequence tag sites

gss = genomic survey sequences



nucleotide databases



protein databases

Protein BLAST: search protein databases using a protein query

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST\_PROGRAMS=blastp&PAGE\_TYPE=Blast+Search

Reader

Query subrange

From [ ] To [ ]

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s)

>gi|4504349|ref|NP\_000509.1| hemoglobin subunit beta [Homo sapiens]  
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGK  
KVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLGNVLVCVLAHHFGKEFTPPVQAAYQK  
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: Non-redundant protein sequences (nr)

Organism:   Exclude  Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

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

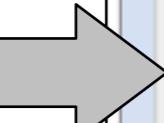
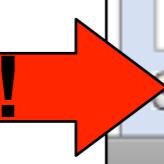
Entrez Query:  Enter an Entrez query to limit search

**Program Selection**

Algorithm:  blastp (protein-protein BLAST)  
 PSI-BLAST (Position-Specific Iterated BLAST)  
 PHI-BLAST (Pattern Hit Initiated BLAST)  
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)  
Choose a BLAST algorithm

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

**Algorithm parameters**

**Organism**   
**Entrez**   
**Settings!** 

# Step 4a: Select optional search parameters

**Algorithm parameters**

General Parameters

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

**Short queries**:  Automatically adjust parameters for short input sequences

**Expect threshold**: 10

**Word size**: 3

**Max matches in a query range**: 0

**Scoring Parameters**

**Matrix**: BLOSUM62

**Gap Costs**: Existence: 11 Extension: 1

**Compositional adjustments**: Conditional compositional score matrix adjustment

**Filters and Masking**

**Filter**:  Low complexity regions

**Mask**:  Mask for lookup table only  
 Mask lower case letters

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

Annotations for Step 4a:

- A blue double-headed arrow points between the "Expect threshold" field and the text "Expect".
- An orange double-headed arrow points between the "Word size" field and the text "Word size".
- An orange double-headed arrow points between the "Matrix" dropdown and the text "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 Reader

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

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp suite/ Formatting Results - FVGUTMRZ013

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

You Tube Learn about the enhanced report Blast report description

gi|4504349|ref|NP\_000509.1| hemoglobin

Query ID: Icl|84677 Database Name: nr  
Description: gi|4504349|ref|NP\_000509.1| hemoglobin subunit Description: All non-redundant GenBank CDS  
beta [Homo sapiens] translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects  
Molecule type: amino acid Program: BLASTP 2.2.27+ ► Citation  
Query Length: 147

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

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

Graphic Summary

Show Conserved Domains

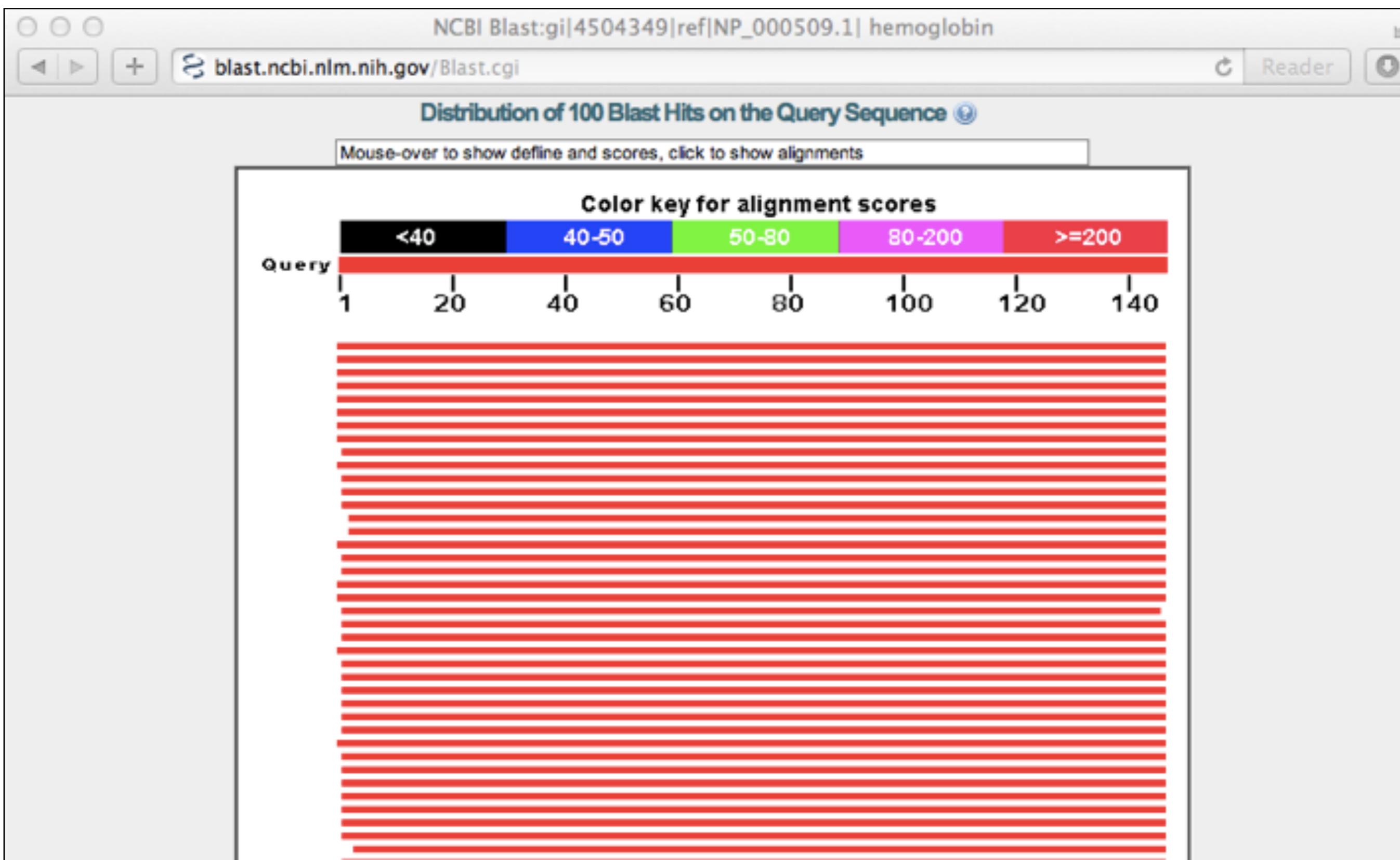
Putative conserved domains have been detected, click on the image below for detailed results.

Query seq. 1 25 50 75 100 125 147  
Specific hits: hem-binding site globin  
Superfamilies: globin\_like superfamily

Distribution of 100 Blast Hits on the Query Sequence ⓘ

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

# Further down the results page...



# Further down the results page...

NCBI Blast:gi|4504349|ref|NP\_000509.1| hemoglobin

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

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

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

# Further down the results page...

NCBI Blast:gi|4504349|ref|NP\_000509.1| hemoglobin

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

Download GenPept Graphics ▾ Next ▲ Previous ▲ Descriptions

hemoglobin subunit beta [Homo sapiens]  
Sequence ID: ref|NP\_000509.1| Length: 147 Number of Matches: 1  
► See 84 more title(s)

---

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

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

Query 1 MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60  
Sbjct 1 MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60

Query 61 VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 120  
Sbjct 61 VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 120

Query 121 KEFTPQAAVYQKVVAGVANALAHKYH 147  
Sbjct 121 KEFTPQAAVYQKVVAGVANALAHKYH 147

---

Related Information

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

Download GenPept Graphics ▾ Next ▲ Previous ▲ Descriptions

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

---

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

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

---

Related Information

# Different output formats are available

NCBI Blast:gi|4504349|ref|NP\_000509.1| hemoglobin  
blast.ncbi.nlm.nih.gov/Blast.cgi Reader

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

Home Recent Results Saved Strategies Help

NCBI BLAST/blastp suite/ Formatting Results - FVGUTMRZ013

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

Formatting options

Show Alignment as HTML  Old View Reset form to defaults

Alignment View Query-anchored with letters for identities

Display  Graphical Overview  Sequence Retrieval  NCBI-gi

Masking Character: Lower Case Color: Grey

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

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

Entrez query:

Expect Min: Expect Max:

Percent Identity Min: Percent Identity Max:

Format for  PSI-BLAST with inclusion threshold:

gi|4504349|ref|NP\_000509.1| hemoglobin

# E.g. Query anchored alignments

NCBI Blast:gi|4504349|ref|NP\_000509.1| hemoglobin

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

Query	Score	Sequence	Length
AAX37051	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAX29557	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
NP_000509	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
P02024	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAN84548	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAZ39780	1	MVHLTPKEKS A VTALWGKVN DEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	60
ACU56984	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAD19696	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFLESFGDLSTPDAVMGNPK	60
1COH_B	1	VHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
AAF00489	1	MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
2YRS_B	1	VHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1DXU_B	1	MHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1HDB_B	1	VHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1DXV_B	2	HLTPEEKSAVTALWGKVNDEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	59
3KMF_C	2	HLTPEEKSAVTALWGKVNDEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	59
AAL68978	1	MVHLTPEEKSAVTALWGKVNDEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	60
1NQP_B	1	VHLTPEEKSAVTALWGKVNDEVGG KALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	59
1K1K_B	1	VHLTPKEKS A VTALWGKVN DEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	59
AAN11320	1	MVHLTPVEKS A VTALWGKVN DEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	60
XP_002822173	1	MVHLTPEEKSAVTALWGKVNDEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	60
1Y85_B	1	VHLTPEEKSAVTALWGKVNDEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	59
1YE0_B	1	MHLTPEEKSAVTALWGKVNDEVGG E ALGR LLAV Y PWTQRFFESFGDLSTPDAVMGNPK	59
1O1O_B	1	MHLTPEEKSAVTALWGKVNDEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	59
CAA23759	1	MVHLTPVEKS A VTAXWGKVN DEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	60
1YE2_B	1	MHLTPEEKSAVTALWGKVNDEVGG E ALGR LLVVFPWTQRFFESFGDLSTPDAVMGNPK	59
1Y5F_B	1	MHLTPEEKSAVTALWGKVNDEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	59
1A00_B	1	MHLTPEEKSAVTALWGKVNDEVGG E ALGR LLVV Y PYTQRFFESFGDLSTPDAVMGNPK	59
1HBS_B	1	VHLTPVEKS A VTALWGKVN DEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	59
1ABY_B	1	MHLTPEEKSAVTALWGKVNDEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	59
1CMY_B	1	VHLTPKEKS A VTALWGKVN DEVGG E ALGR LLVV Y PWTQRFFESFGDLSTPDAVMGNPK	59

# ... and alignments with dots for identities

NCBI Blast:gi|4504349|ref|NP\_000509.1| hemoglobin

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

Reader

Query	Length	Sequence	Score
<a href="#">AAX37051</a>	1	MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
<a href="#">AAX29557</a>	1	.....	60
<a href="#">NP_000509</a>	1	.....	60
<a href="#">P02024</a>	1	.....	60
<a href="#">AAN84548</a>	1	.....	60
<a href="#">AAZ39780</a>	1	.....K.....	60
<a href="#">ACU56984</a>	1	.....	60
<a href="#">AAD19696</a>	1	.....	60
<a href="#">1COH_B</a>	1	.....	59
<a href="#">AAF00489</a>	1	.....	60
<a href="#">2YRS_B</a>	1	.....	59
<a href="#">1DXU_B</a>	1	M.....	59
<a href="#">1HDB_B</a>	1	.....	59
<a href="#">1DXV_B</a>	2	.....	59
<a href="#">3KMF_C</a>	2	.....	59
<a href="#">AAL68978</a>	1	.....	60
<a href="#">1NQP_B</a>	1	.....K.....	59
<a href="#">1K1K_B</a>	1	.....K.....	59
<a href="#">AAN11320</a>	1	.....V.....	60
<a href="#">XP_002822173</a>	1	.....	60
<a href="#">1Y85_B</a>	1	.....	59
<a href="#">1YE0_B</a>	1	M.....A.....	59
<a href="#">1O1Q_B</a>	1	M.....	59
<a href="#">CAA23759</a>	1	.....V.....X.....	60
<a href="#">1YE2_B</a>	1	M.....F.....	59
<a href="#">1Y5F_B</a>	1	M.....	59
<a href="#">1A00_B</a>	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.e. filled in)	1
Evidence for correct use of scoring scheme	1
Direction arrows drawn between all cells	1
Evidence of multiple arrows to a given cell if appropriate	1
Correct optimal score position in matrix used	1
Correct optimal score obtained for given scoring scheme	1
Traceback path(s) clearly highlighted	1
Correct alignment(s) yielding optimal score listed	1
	A+