

# BGGN 213

## Foundations of Bioinformatics Lecture 2

Barry Grant  
UC San Diego

<http://thegrantlab.org/bggn213>

## Today's Menu

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

## Recap From Last Time:

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

## Primary, secondary & composite databases

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

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

# DATABASE VIGNETTE

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

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

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

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

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

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

The screenshot shows the NCBI homepage with a search bar containing 'ras'. The search results are displayed on the right, under the heading 'Welcome to NCBI'. The results include various links such as 'NCBI Home', 'Resource List (A-Z)', and 'All Resources'. A section titled 'Genotypes and Phenotypes' is also visible.

## Example Vignette Questions:

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

The screenshot shows the search results for 'ras' on the NCBI Global Cross Query page. The results are organized into sections: Literature, Genes, Health, and Proteins. The 'Genes' section is highlighted with a red box around the 'Gene' entry, which shows 87,165 results. Other entries in the 'Genes' section include EST, GEO DataSets, GEO Profiles, HomoloGene, PopSet, UniGene, and Proteins.

[ras - Gene - NCBI](http://www.ncbi.nlm.nih.gov/gene/?term=ras)

NCBI Resources How To Sign in to NCBI

Gene Gene Search Save search Advanced Help

Show additional filters Display Settings: Tabular, 20 per page, Sorted by Relevance Send to: Hide sidebar >

Filters: Manage Filters

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

Results: 1 to 20 of 85633

Filters activated: Current only. Clear all to show 87165 items.

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

Find related data Database: Select Find items

Search details ras[All Fields] AND alive[property]

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

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

9

[ras AND "Homo sapiens"\[porgn:\\_txid9606\] - Gene - NCBI](http://www.ncbi.nlm.nih.gov/gene/?term=(ras)&term=(Homo+sapiens)[porgn:_txid9606])

NCBI Resources How To Sign in to NCBI

Gene Gene Search Save search Advanced Help

Show additional filters Display Settings: Tabular, 20 per page, Sorted by Relevance Send to: Hide sidebar >

Filters: Manage Filters

Results: 1 to 20 of 1126

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

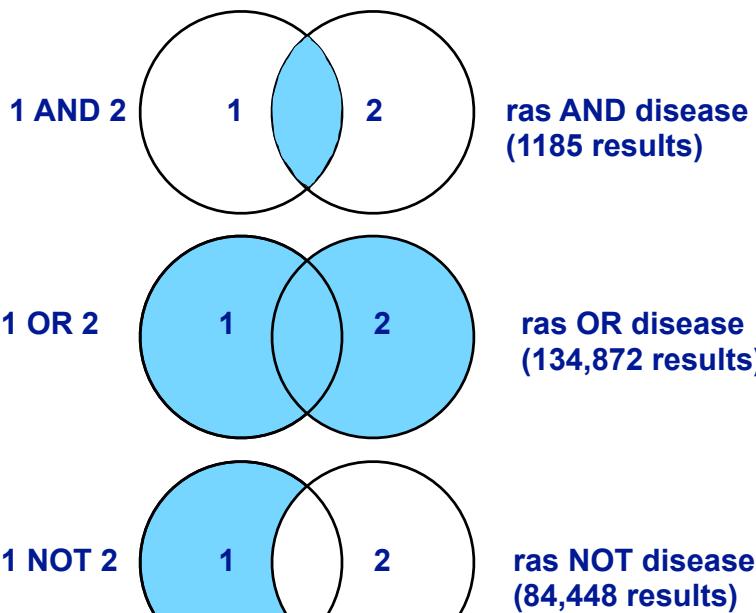
Name/Gene ID	Description	Location	Aliases
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (11470446..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS	Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)]	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, RAS2

Find related data Database: Select Find items

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

Recent activity Turn Off Clear

10



11

[ras AND "Homo sapiens"\[porgn:\\_txid9606\] - Gene - NCBI](http://www.ncbi.nlm.nih.gov/gene/?term=(ras)&term=(Homo+sapiens)[porgn:_txid9606])

NCBI Resources How To Sign in to NCBI

Gene Gene Search Save search Advanced Help

Show additional filters Display Settings: Tabular, 20 per page, Sorted by Relevance Send to: Hide sidebar >

Filters: Manage Filters

Results: 1 to 20 of 1126

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

Name/Gene ID	Description	Location	Aliases
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (11470446..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS	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, RAS2

Find related data Database: Select Find items

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

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:ENSG0000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193

Gene type protein coding  
RefSeq status REVIEWED  
Organism Homo sapiens  
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo  
Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Table of contents

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

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

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:ENSG0000133703; 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)

Location: 12p12.1  
Exon count: 6

Annotation release: 106 Status: current Assembly: GRCh38 (GCF\_000001405.26)  
Annotation release: 105 Status: previous assembly Assembly: GRCh37.p13 (GCF\_000001405.25)

Chromosome 12 - NC\_000012.12

LRMP ← LYRMS ← LOC100421617 ← KRAS ← RPL39P27 →

Genomic regions, transcripts, and products

Genomic Sequence: NC\_000012.12 chromosome 12 reference GRCh38 Primary Assembly  
Go to nucleotide: Graphics Fasta GenBank

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

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

Gene ID: 3845

**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:ENSG0000133703; 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)

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

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

## GO: Gene Ontology

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

The 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.

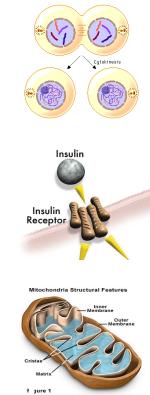
UniProt is a member of the GO Consortium.

## 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



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

UniProtKB/Swiss-Prot:P01116

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

UniProt will detail much more information for protein coding genes

P01116 - RASK\_HUMAN

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

Function

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

Enzyme regulation Alternates between an inactive form bound to GDP and an active form bound to GTP. Activated by a guanine nucleotide-exchange factor (GEF) and inactivated by a GTPase-activating protein (GAP). Interaction with SOS1 promotes exchange of bound GDP by GTP. 3 Publications

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding <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

P01116 - RASK\_HUMAN

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

Function

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

Enzyme regulation Alternates between an inactive form bound to GDP and an active form bound to GTP. Activated by a guanine nucleotide-exchange factor (GEF) and inactivated by a GTPase-activating protein (GAP). Interaction with SOS1 promotes exchange of bound GDP by GTP. 3 Publications

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding <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			

View FASTA file format

```
>sp|P01116|RASK_HUMAN GTPase KRas OS=Homo sapiens GN=KRAS PE=1 SV=1 MTEYKLVVVGGAGGVGKSAITIQLIQNHFDEYDPTEIDSYRKVQVILGETCLLDILPTAG QEEYSAMRDQYMTREGFGLCVFAINNTKSFEDIHRYRQEIKRKVQDSVEDPVMVLGNKCDL PSFVVDTRQADLARSYGIPPIETTSAKTRQRQVEDAFYTTLVREIRQYRLKKISKEEKTFCG VKIKRCIM
```

UniProt will detail much more information for protein coding genes

**Example Questions:**  
What variants of this enzyme are involved in gastric cancer and other human diseases?

**Example Questions:**  
Are high resolution protein structures available to examine the details of these mutations?

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

Structure Summary    **3D View**    Annotations    Sequence    Sequence Similarity    Structure    Literature

**4EPV**  
Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation

DOI: 10.2210/pdb4epv/pdb  
Classification: HYDROLASE  
Deposited: 2012-04-17 Released: 2012-05-23  
Deposition author(s): Sun, Q., Burke, J.P., Phan, J., Burns, M.C., Olejniczak, E.T., Waterson, A.G., Lee, T., Rossanese, O.W., Fesik, S.W.  
Organism: Homo sapiens  
Expression System: Escherichia coli  
Mutation(s): 1

Experimental Data Snapshot    wwPDB Validation    3D Report    Full Report  
Method: X-RAY DIFFRACTION    Metric    Percentile Ranks    Value

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

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

**4EPV**  
Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation

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

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

**Display Options**

- Assembly: Biossembly 1
- Model: Model 1
- Symmetry: None
- Interaction: **IGDP201-A**
- Style: Cartoon
- Color: Rainbow
- Ligand: None
- Quality: Automatic
- Water:
- Ions:
- Hydrogens:
- Clashes:

**Viewer Options**

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

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

**FAMILY AND DOMAIN DATABASES**

- Gene3D<sup>1</sup>: 3,40,50,300. 1 hit.
- InterPro<sup>1</sup>: IPR027417. P-loop\_NTPase.  
IPR005225. Small\_GTP-bd\_dom.  
IPR001806. Small\_GTPase.  
IPR020849. Small\_GTPase\_Ras.  
[Graphical view]
- PANTHER<sup>1</sup>: PTHR24070. 1 hit.
- PFam<sup>1</sup>: PF00071. Ras. 1 hit.  
[Graphical view]
- PRINTS<sup>1</sup>: PR00449. RASTRNSFRMNG.
- SMART<sup>1</sup>: SM00173. RAS. 1 hit.  
[Graphical view]
- SUPFAM<sup>1</sup>: SSF52540. SSF52540. 1 hit.
- TIGRFAMS<sup>1</sup>: TIGR00231. small\_GTP. 1 hit.
- PROSITE<sup>1</sup>: PS51421. RAS. 1 hit.  
[Graphical view]

**Sequences (2)**  
Sequence status<sup>1</sup>: Complete.  
Sequence processing<sup>1</sup>: The displayed sequence is further processed into a mature form.  
This entry describes 2 isoforms<sup>1</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... ?

**Family: Ras (PF00071)**

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

**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](#) | [Wikidata: 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/Wnt 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 GTPases, 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 these 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 signalling 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).

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

**Contents** | [Edit](#)

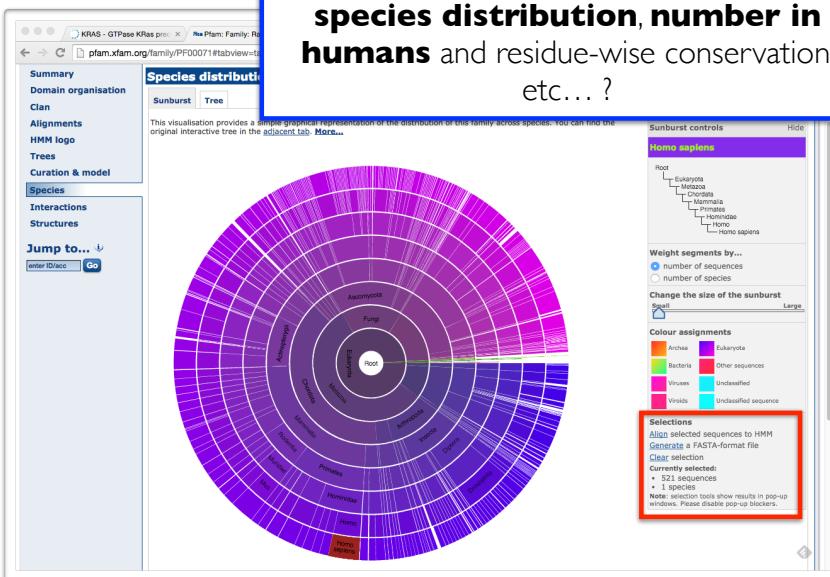
**Identifiers**

Symbol	Ras
Pfam	PF00071_0
InterPro	IPR013753_0
PROSITE	POCO0017_0
SCOP	Sp21_0
SUPERFAMILY	Sp21_0

**3D Structure**

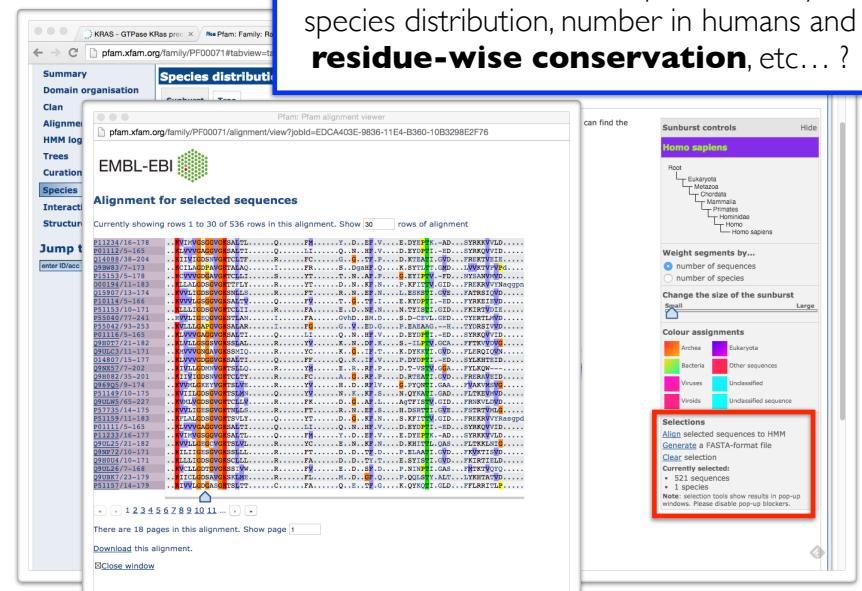
## Example Questions:

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



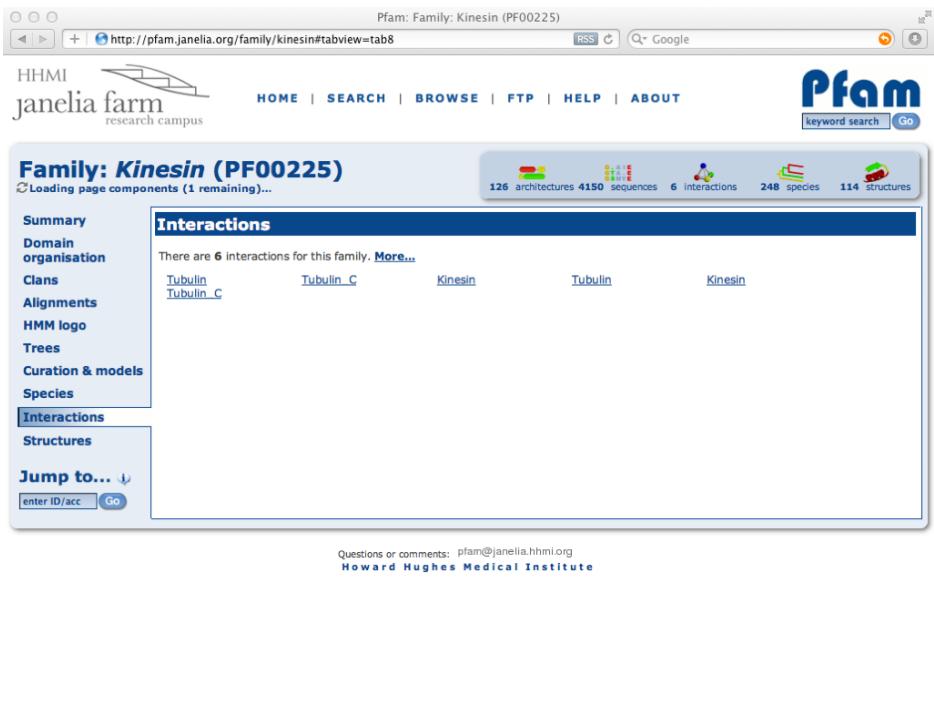
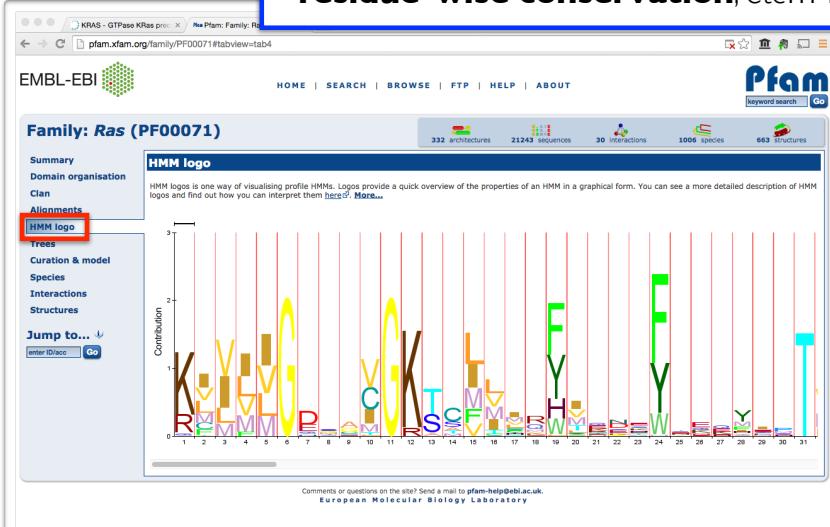
## Example Questions:

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



## Example Questions:

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



Pfam: Family: Kinesin (PF00225) <http://pfam.janelia.org/family/kinesin#tabview=tab9>

HHMI  
janelia farm  
research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

**Pfam**  
keyword search | Go

**Family: Kinesin (PF00225)**

**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 PDBer group, to allow us to map Pfam domains onto UniProt sequences and three-dimensional protein structures. The table below shows the structures on which the **Kinesin** domain has been found.

UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View
A8BKD1_GIALA	11 - 335	2vvg	A	11 - 335	Jmol AstexViewer SPICE
			B	11 - 335	Jmol AstexViewer SPICE
CENPE_HUMAN	12 - 329	1t5c	A	12 - 329	Jmol AstexViewer SPICE
			B	12 - 329	Jmol AstexViewer SPICE
KAR3_YEAST	392 - 723	1f9t	A	392 - 723	Jmol AstexViewer SPICE
		1f9u	A	392 - 723	Jmol AstexViewer SPICE
		1f9v	A	392 - 723	Jmol AstexViewer SPICE
		1f9w	A	392 - 723	Jmol AstexViewer SPICE
			B	392 - 723	Jmol AstexViewer SPICE
KI13B_HUMAN	11 - 352	3kar	A	11 - 352	Jmol AstexViewer SPICE
			B	11 - 352	Jmol AstexViewer SPICE
			C	11 - 352	Jmol AstexViewer SPICE
		1li6	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		1g0b	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		1x88	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
			A	24 - 359	Jmol AstexViewer SPICE

**Summary**  
**Domain organisation**  
**Clans**  
**Alignments**  
**HMM logo**  
**Trees**  
**Curation & models**  
**Species**  
**Interactions**  
**Structures**  
**Jump to... ↴**  
enter ID/acc Go

Pfam: Jmol <http://pfam.janelia.org/structure/viewer?viewer=jmol&id=3bfm>

Pfam: Family: Kinesin (PF00225) Pfam: Jmol

welcome trust sanger institute

**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.

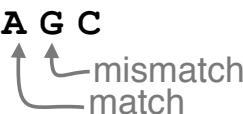
**Seq1:** C A T T C A C

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

[Screencast Material]

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

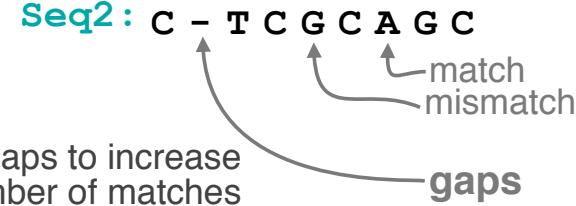
**Seq1:** C A T T C A C  
**Seq2:** C T C G C A G C



Two types of character correspondence

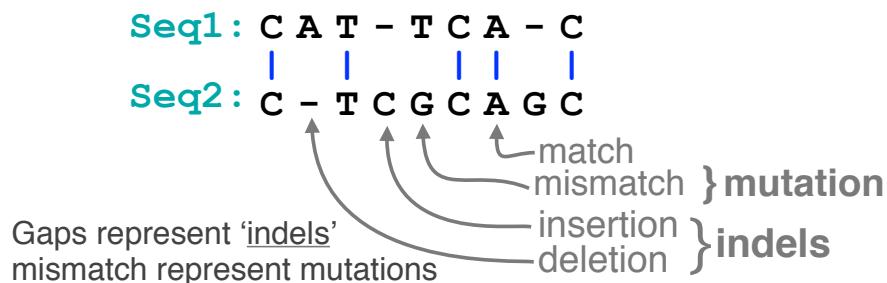
**Basic Idea:** Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.

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



Add gaps to increase number of matches

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



## Why compare biological sequences?

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

Practical applications include...

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

Practical applications include...

- **Similarity searching of databases**
    - Protein structure prediction
  - **Assembly of sequence reads** into a longer construct such
  - **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
- N.B. Pairwise sequence alignment is arguably the most fundamental operation of bioinformatics!*

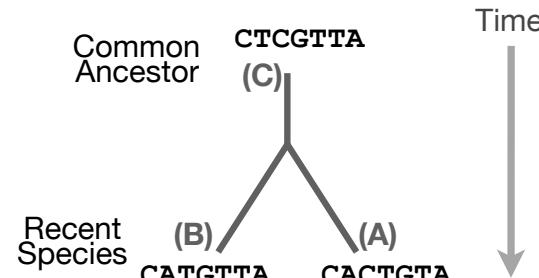
## ALIGNMENT FOUNDATIONS

- 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.

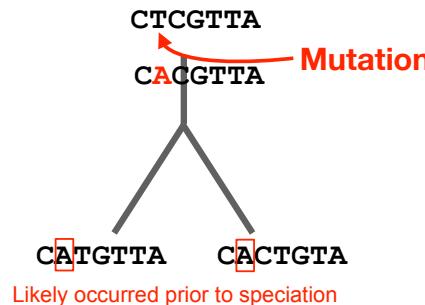
- Mutations/Substitutions
- Deletions
- Insertions



## Mutations, deletions and insertions

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

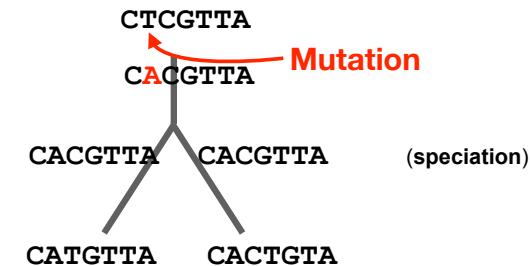
- **Mutations/Substitutions**       $\text{CTCGTTA} \rightarrow \text{CACGTTA}$
- Deletions
- Insertions



## Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

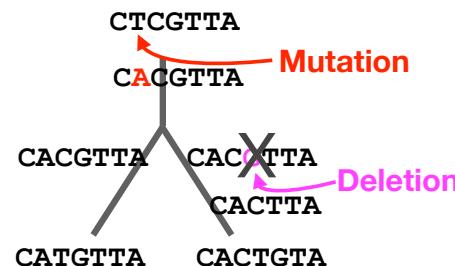


## Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

$$\text{CTCGTTA} \rightarrow \text{CACGTTA}$$
$$\text{CACGTTA} \rightarrow \text{CACTTA}$$

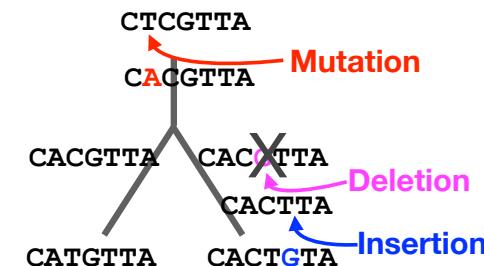


## Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

$$\text{CTCGTTA} \rightarrow \text{CACGTTA}$$
$$\text{CACGTTA} \rightarrow \text{CACTTA}$$
$$\text{CACTTA} \rightarrow \text{CACTGTA}$$

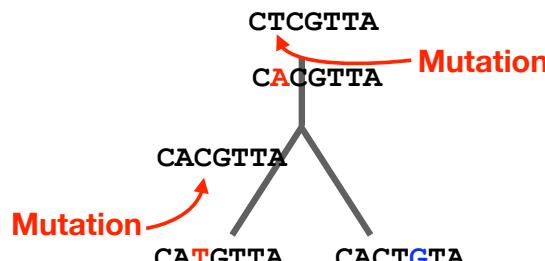


## Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

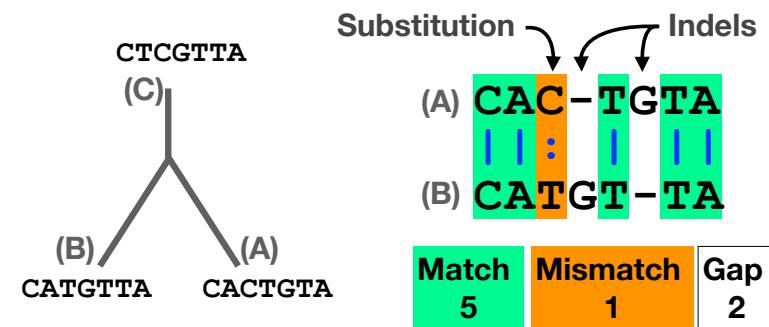
$$\text{CTCGTTA} \rightarrow \text{CACGTTA}$$
$$\text{CACGTTA} \rightarrow \text{CATGTTA}$$



## Alignment view

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

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



## Alternative alignments

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

Q. Which of these 3 possible alignments is best?

1.

CACTGTA  
||:||:  
CATGTTA

2.

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

3.

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

## Alternative alignments

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

● 4 matches  
● 3 mismatches  
○ 0 gaps

● 6 matches  
● 0 mismatches  
○ 2 gaps

● 5 matches  
● 1 mismatch  
○ 2 gaps

CACTGTA  
||:||:  
CATGTTA

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

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

## Scoring alignments

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

● 4 (+3)  
● 3 (+1)  
○ 0 (-1) = 15

● 6 (+3)  
● 0 (+1)  
○ 2 (-1) = 16

● 5 (+3)  
● 1 (+1)  
○ 2 (-1) = 14

CACTGTA  
||:||:  
CATGTTA

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

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

## Optimal alignments

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

● 4 matches  
● 3 mismatches  
○ 0 gaps

● 6 matches  
● 0 mismatches  
○ 2 gaps

● 5 matches  
● 1 mismatch  
○ 2 gaps

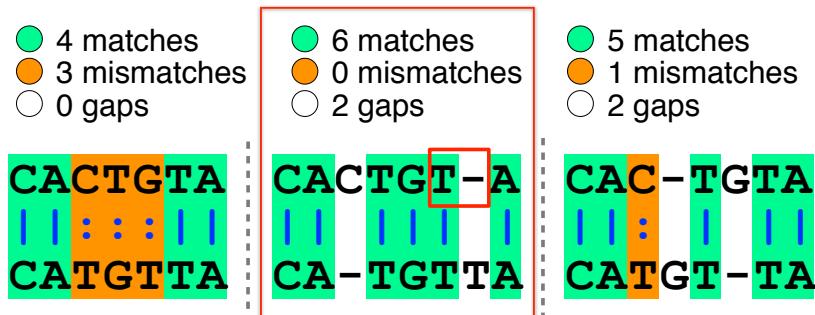
CACTGTA  
||:||:  
CATGTTA

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

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

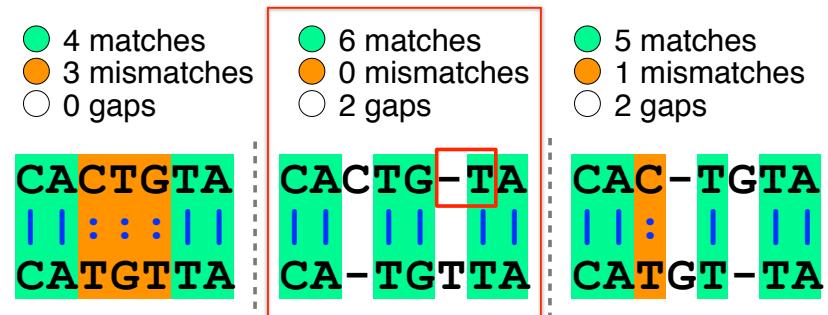
## Optimal alignments

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



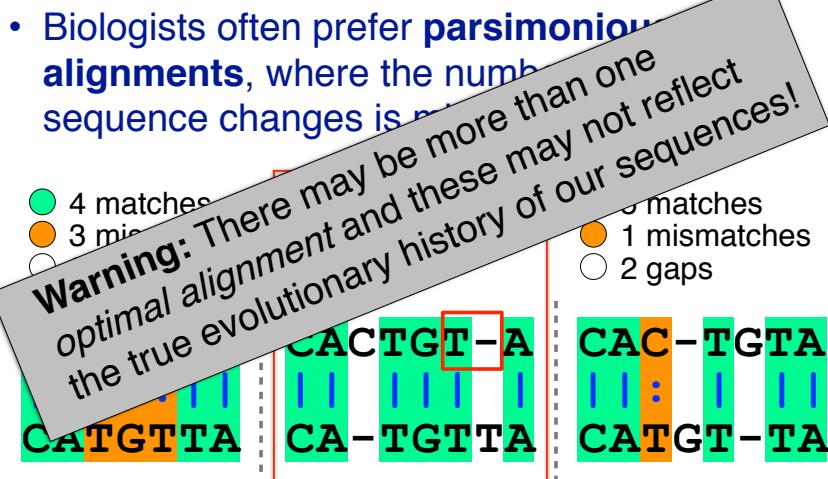
## Optimal alignments

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



## Optimal alignments

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



## ALIGNMENT FOUNDATIONS

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

### How...

- Dot matrices
- Dynamic programming
  - Global alignment
  - Local alignment
- BLAST heuristic approach

## ALIGNMENT FOUNDATIONS

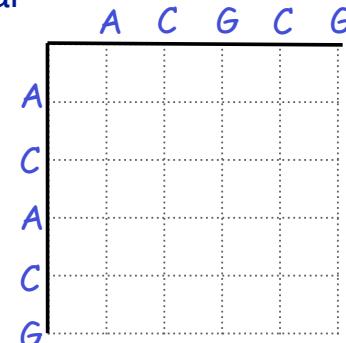
- Why...
  - Why compare biological sequences?
- What...
  - Alignment view of sequence changes during evolution (matches, mismatches and gaps)

### • How...

- Dot matrices
- D
- How do we compute the optimal alignment between two sequences?
- BLAST heuristic approach

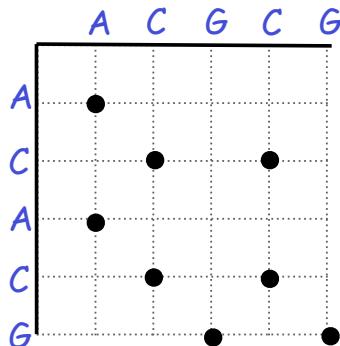
## Dot plots: simple graphical approach

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



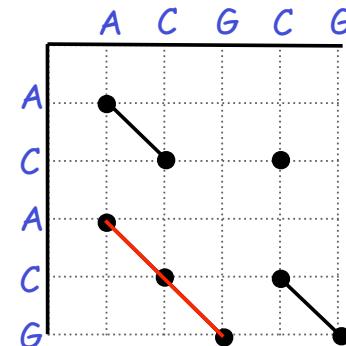
## Dot plots: simple graphical approach

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



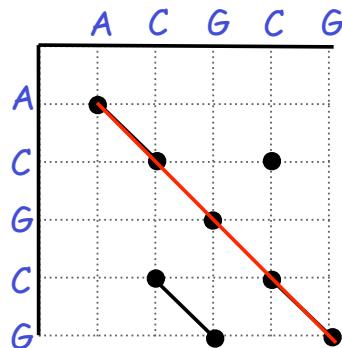
## Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



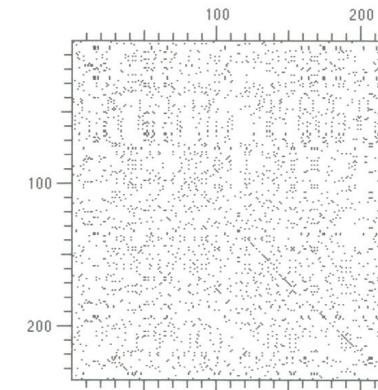
## Dot plots: simple graphical approach

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



## Dot plots: simple graphical approach

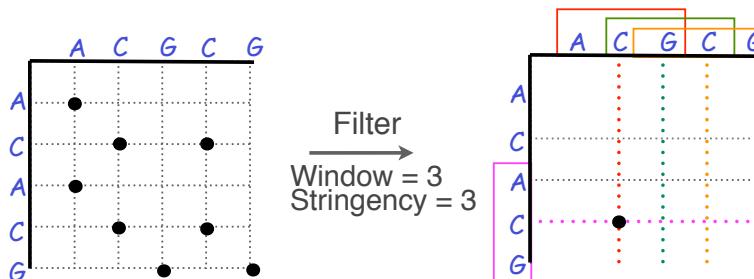
- Dot matrices for long sequences can be noisy



## Dot plots: window size and match stringency

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

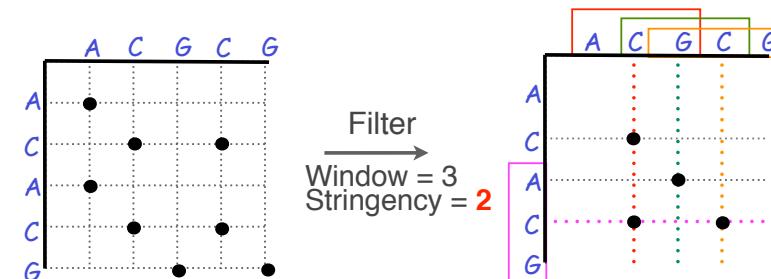
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
  - You have to choose window size and stringency



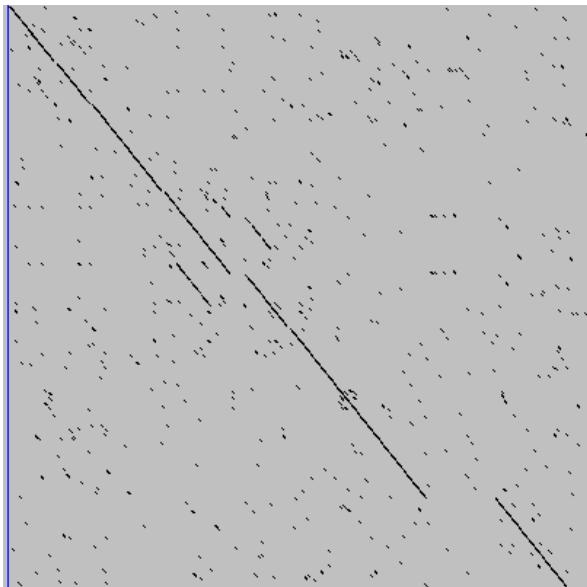
## Dot plots: window size and match stringency

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

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



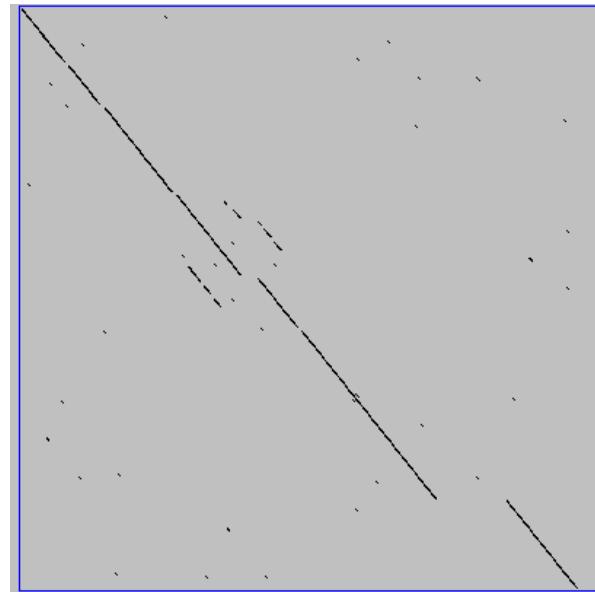
## Window size = 5 bases



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

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

## Window size = 7 bases

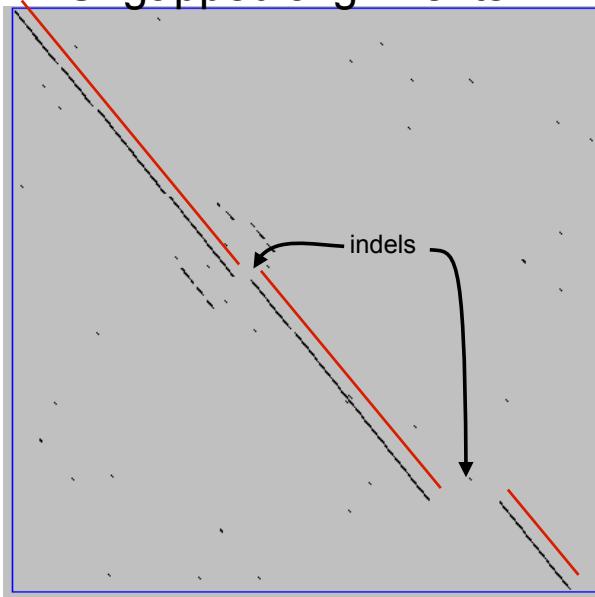


This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

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

Bigger window (kmer)  
fewer matches to consider

## Ungapped alignments



Only **diagonals** can be followed.

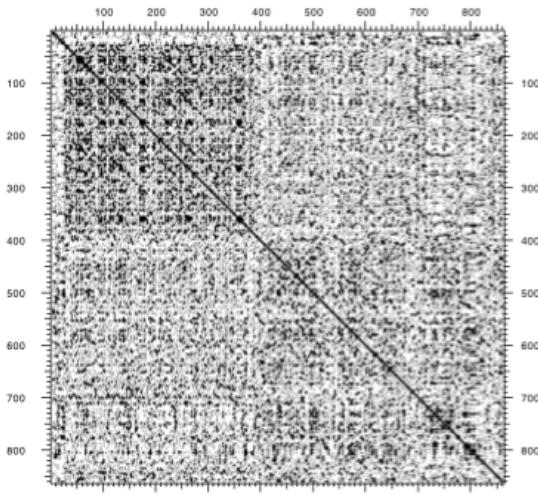
Downward or rightward paths represent **insertion** or **deletions** (gaps in one sequence or the other).

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

## Uses for dot matrices

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

## Repeats

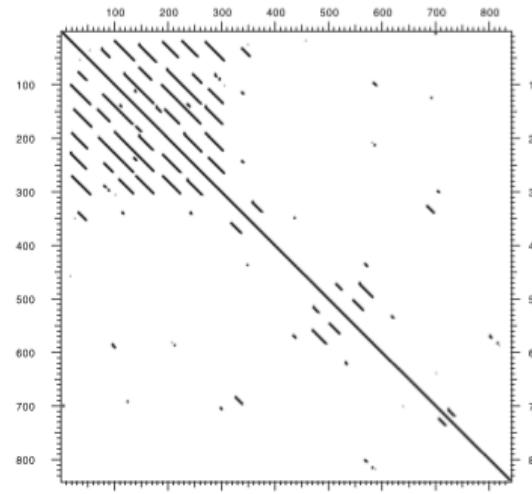


Human LDL receptor  
protein sequence  
(Genbank P01130)

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

(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/>

The screenshot shows a web-based tool for comparing two sequences. At the top, it says "BGGN-213: Dot Plot Comparison of Two Sequences". Below this is a detailed description of what dot plots are. On the left, there are "Dot Plot Parameters" with sliders for "Window Size" (set to 3), "Moving window step size" (set to 2), and "Match stringency" (set to 2). To the right are two dot plots: "Protein Dot Plot" and "DNA Dot Plot", both showing diagonal patterns of dots. At the bottom, there's a section titled "Questions for discussion" with three bullet points.

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: 3

Moving window step size: 2

Match stringency: 2

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

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

Sequence 2

Sequence 1

Sequence 2

Sequence 1

Questions for discussion:

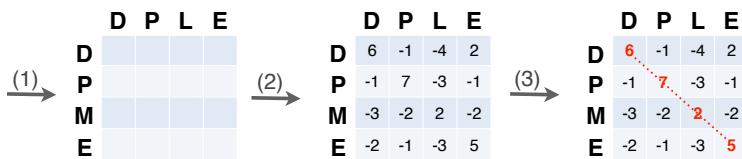
- Why does the DNA sequence have more dots than the protein sequence plot?
- How can we increase the signal to noise ratio?
- What does it mean when a window size is larger than the match stringency?

## 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

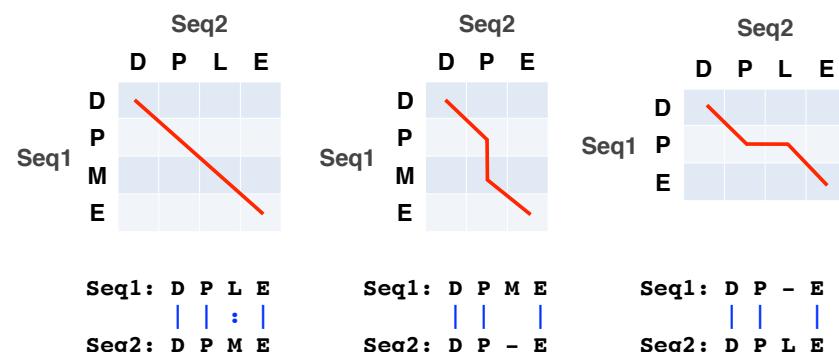
# The Dynamic Programming Algorithm

- 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.

Different paths represent different alignments



Matches are represented by diagonal paths & indels with horizontal or vertical path segments

## Algorithm of Needleman and Wunsch

- The Needleman–Wunsch approach to global sequence alignment has three basic steps:
  - (1) setting up a 2D-grid (or **alignment matrix**),
  - (2) **scoring the matrix**, and
  - (3) identifying the **optimal path** through the matrix



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

## Scoring the alignment matrix

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

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

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

## Scoring the alignment matrix

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

Sequence 2

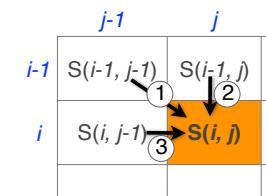
	j	-	D	P	L	E	
Sequence 1	I	-	0	-2	-4	-6	-8
	D	-2					
	P	-4					
	M	-6					
	E	-8					

**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

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



## 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

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

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

## 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					

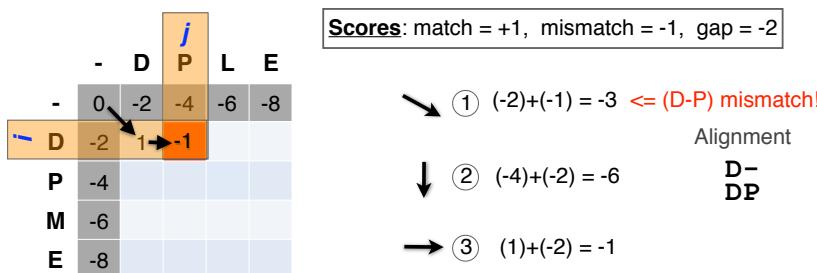
→ ①  $(0) + (+1) = +1$  <= (D-D) match!

$$\downarrow \quad (2) \quad (-2) + (-2) = -4$$

## Alignment

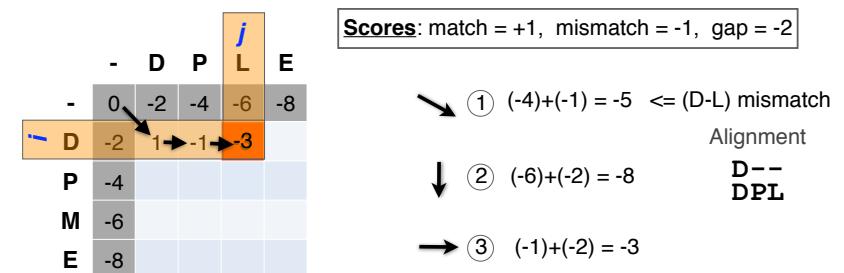
## Scoring the alignment matrix

- 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)



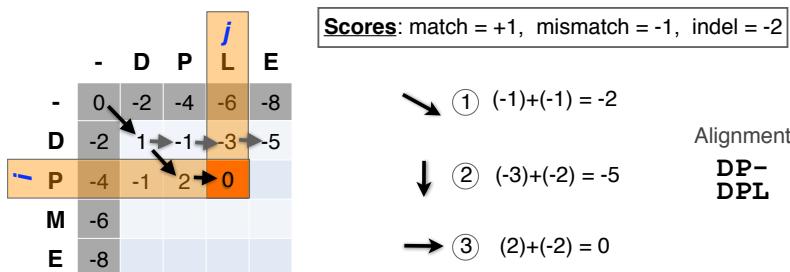
## Scoring the alignment matrix

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



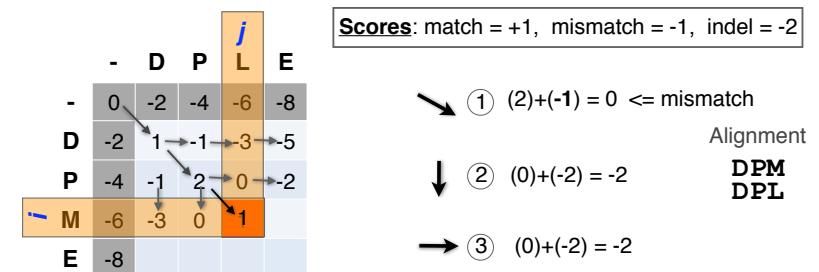
## Scoring the alignment matrix

- 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.



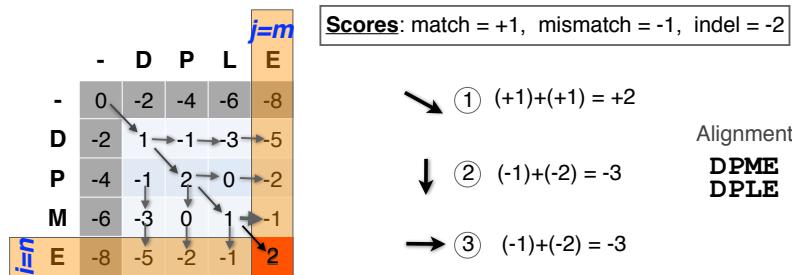
## Scoring the alignment matrix

- 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



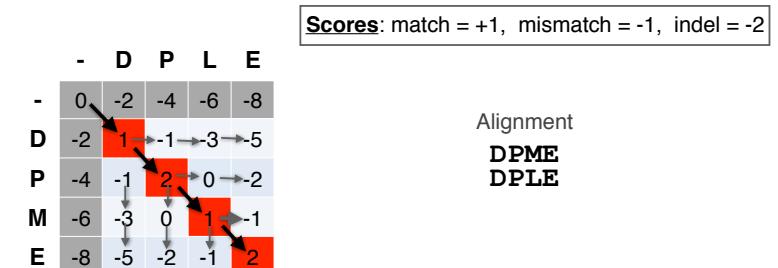
## Scoring the alignment matrix

- 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)



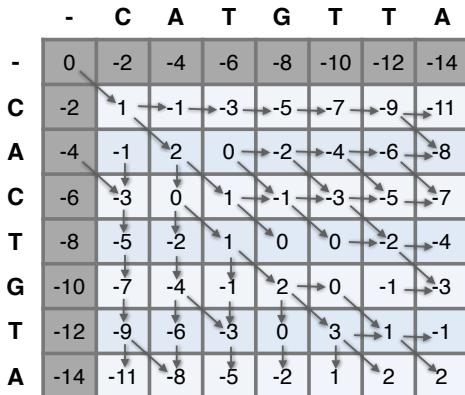
## Scoring the alignment matrix

- 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



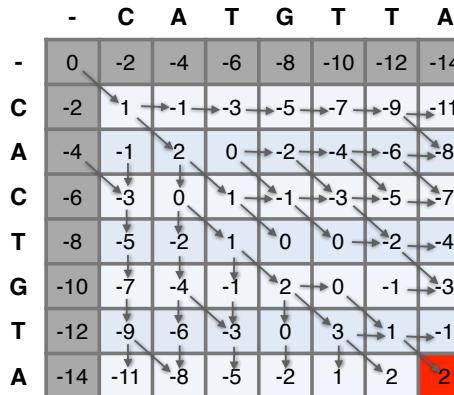
## Questions:

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



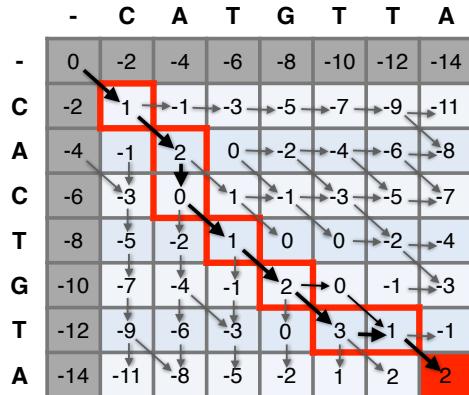
## Questions:

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



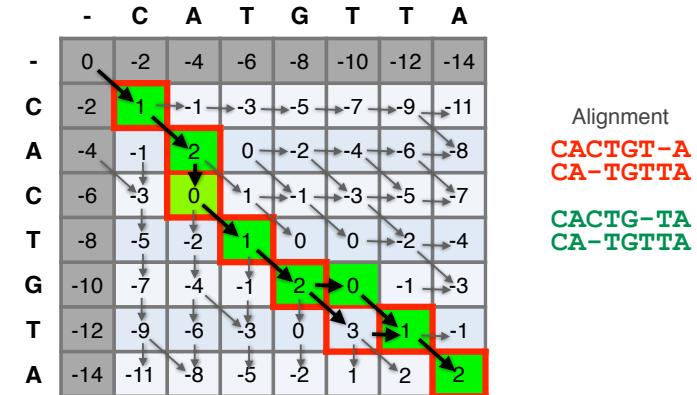
## Questions:

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



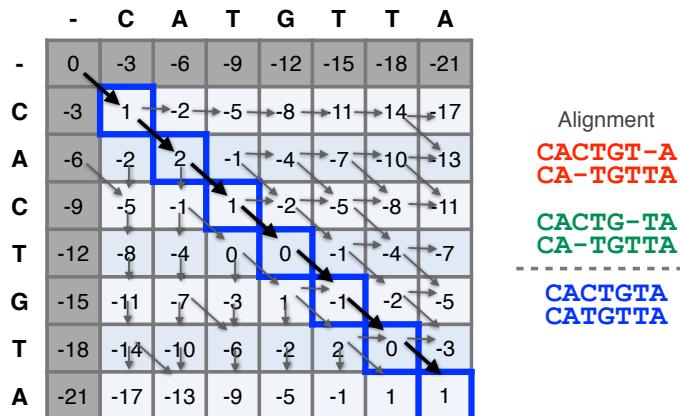
## More than one alignment possible

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



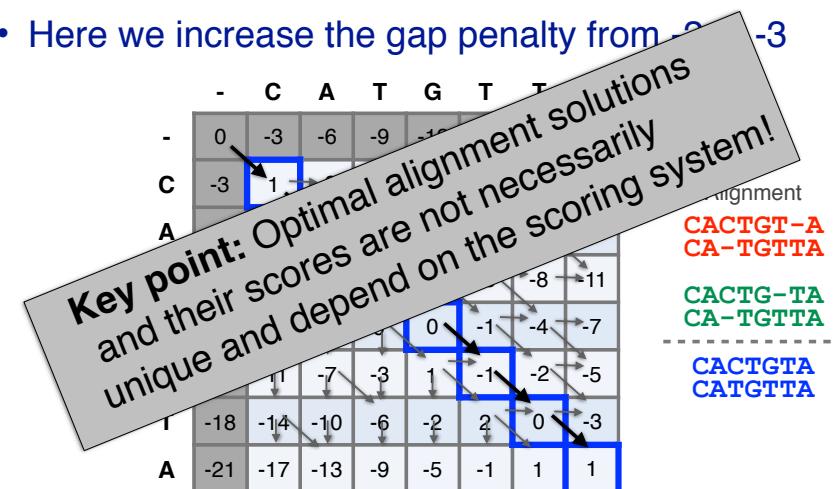
The alignment and score are dependent on the scoring system

- 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 -3 to -2



# Your Turn!

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

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

	A	G	T	T	C
A	0				
T					
T					
G					
C					

## NW DYNAMIC PROGRAMMING

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

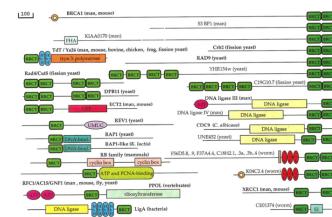
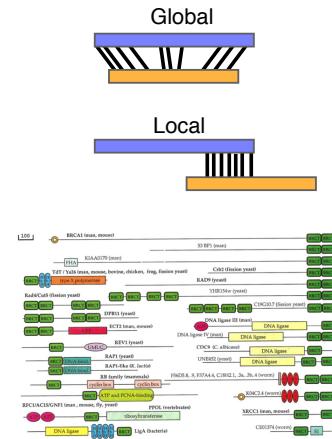
	A	G	T	T	C	
A	0	-2	-4	-6	-8	-10
T	-2	+2	0	-2	-4	-6
T	-6	-2	-1	+3	+4	+2
G	-8	-4	0	+1	+2	+3
C	-10	-6	-2	-1	0	+4

## ALIGNMENT FOUNDATIONS

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

## Global vs local alignments

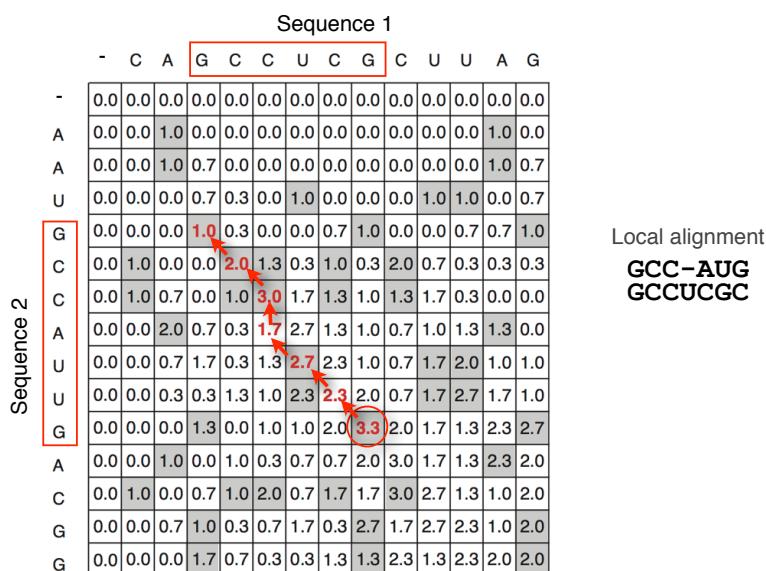
- Needleman-Wunsch is a **global alignment** algorithm
  - Resulting alignment spans the complete sequences end to end
  - This is appropriate for closely related sequences that are similar in length
- For many practical applications we require **local alignments**
  - Local alignments highlight sub-regions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



## Local alignment: Definition

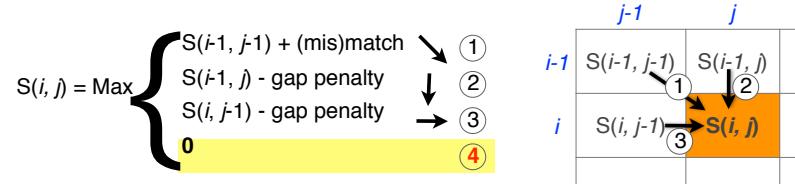
- Smith & Waterman proposed simply that a local alignment of two sequences allow arbitrary-length segments of each sequence to be aligned, with no penalty for the unaligned portions of the sequences. Otherwise, the score for a local alignment is calculated the same way as that for a global alignment

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



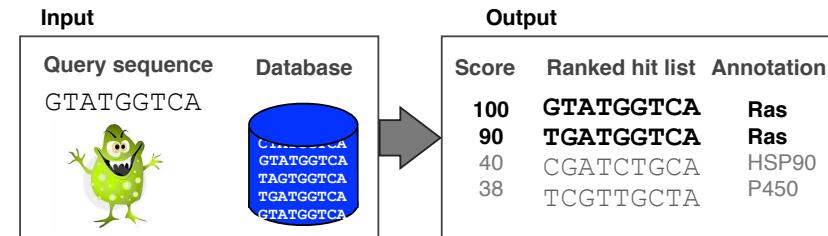
## The Smith-Waterman algorithm

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



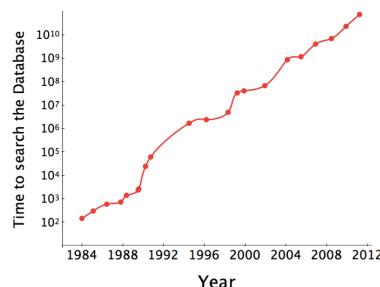
## Local alignments can be used for database searching

- 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

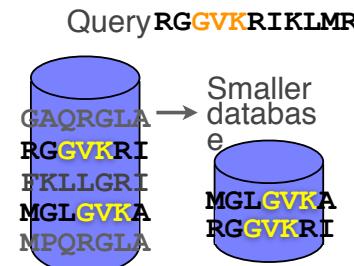
- Due to the rapid growth of sequence databases, search algorithms have to be both efficient and sensitive
  - Time to search with SW is proportional to  $m \times n$  ( $m$  is length of query,  $n$  is length of database), **too slow for large databases!**



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**

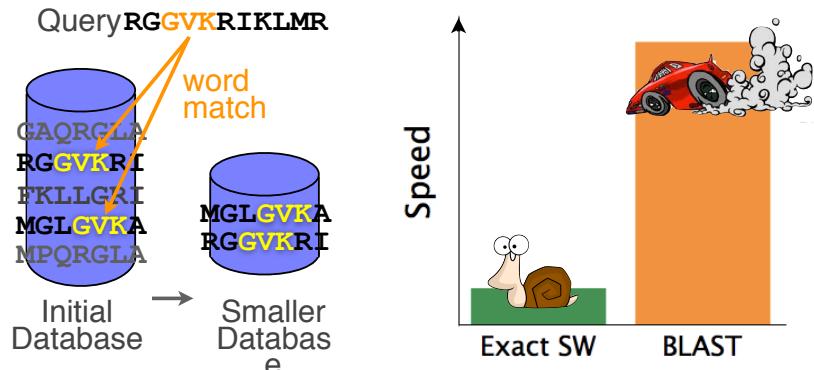
- BLAST (Basic Local Alignment Search Tool) is a simplified form of Smith-Waterman (SW) alignment that is popular because it is **fast** and **easily accessible**
  - BLAST is a heuristic approximation to SW - It examines only part of the search space
  - BLAST saves time by restricting the search by 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 (SW) that is popular because it is **fast**
  - BLAST finds regions of local sequence similarity
  - BLAST does not search by scanning all matches before performing a search
  - Altschul et al. (1990) note that sensitivity in exchange for speed is contrast to SW, BLAST is not guaranteed to find optimal alignments

“The central idea of the BLAST algorithm is to confine attention to sequence pairs that contain an initial **word pair match**”

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



## How BLAST works

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

**RGGVKRI** Query sequence  
**RGG**  
**GGV**  
**GVK**  
**VKR**  
**KRI**

generate list of  $w=3$  words for query

## Blast

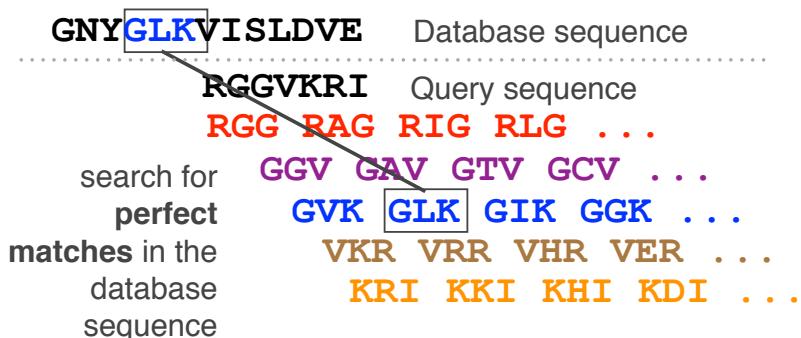
- Phase 2: expand word pairs to include those similar to query (defined as those above a similarity threshold to original word, i.e. match scores in substitution matrix)

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

extend list of words similar to query

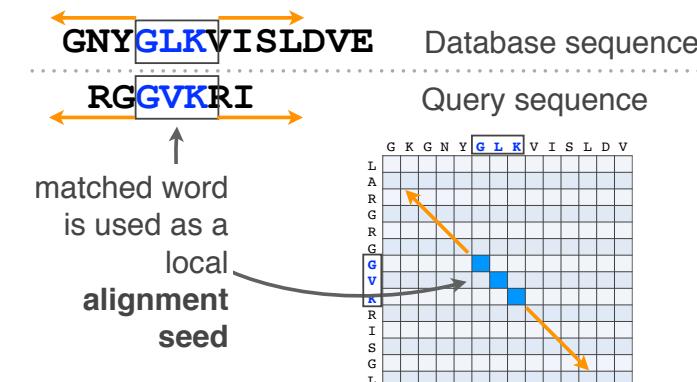
## Blast

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

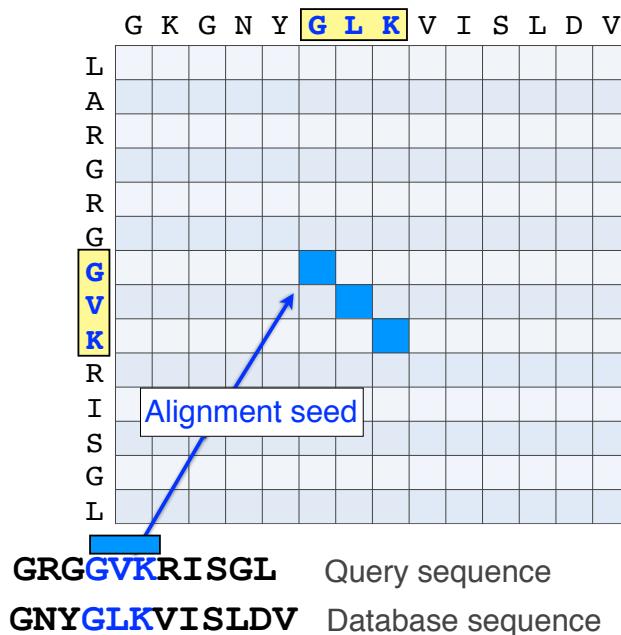


## Blast

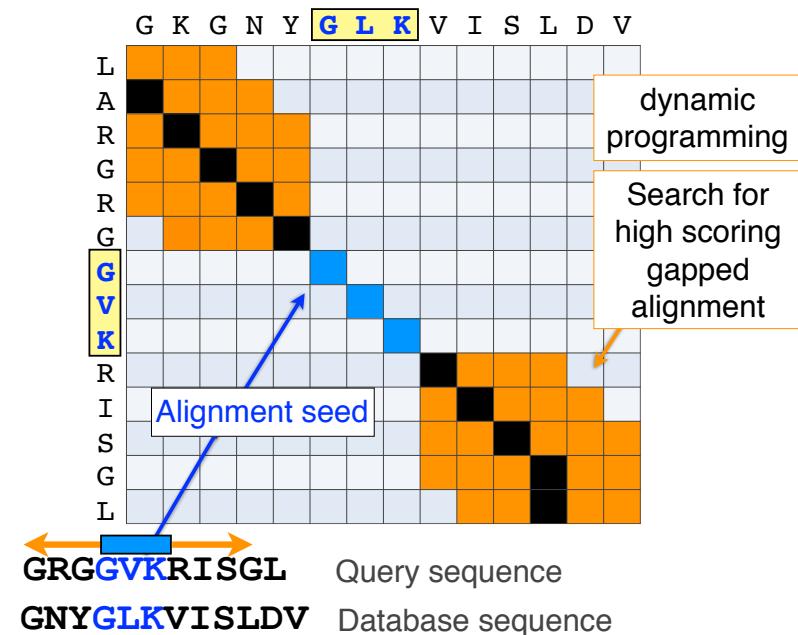
- **Phase 4:** the initial database hits are extended in both directions using dynamic programming



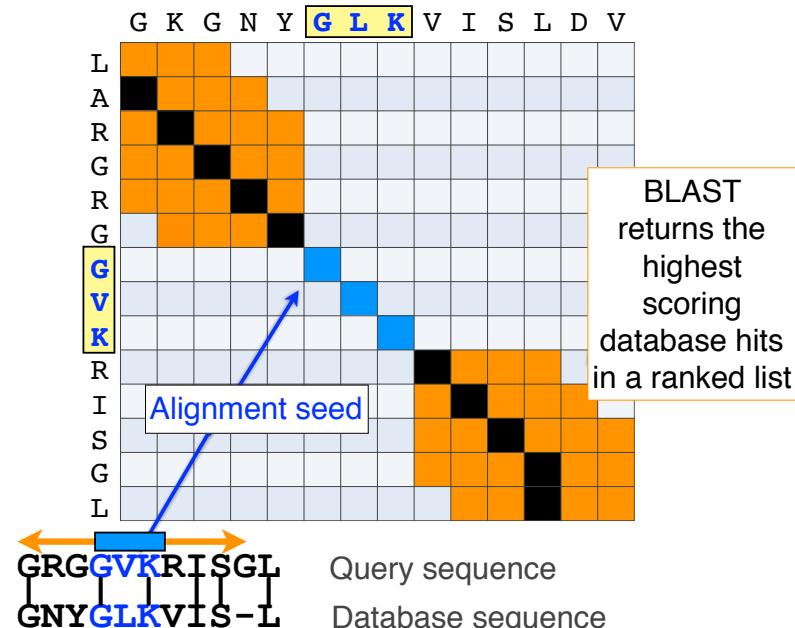
118



119



120



121

## 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

122

## 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

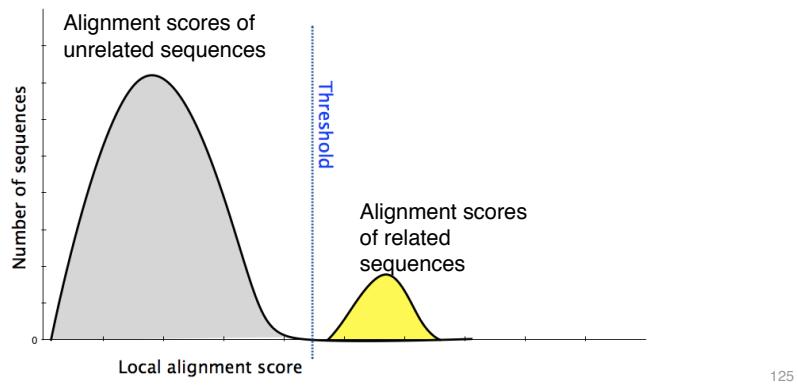
123

## BLAST scores and E-values

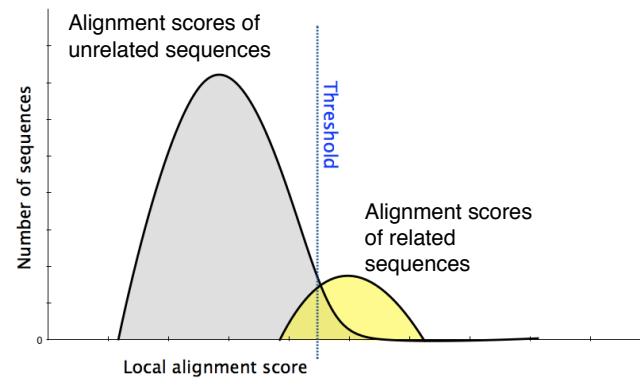
- The **E value** is the **expected** number of hits that are as good or better than the observed local alignment score (with this score or better) if the query and database are **random** with respect to each other
  - i.e. the number of alignments expected to occur by chance with equivalent or better 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

124

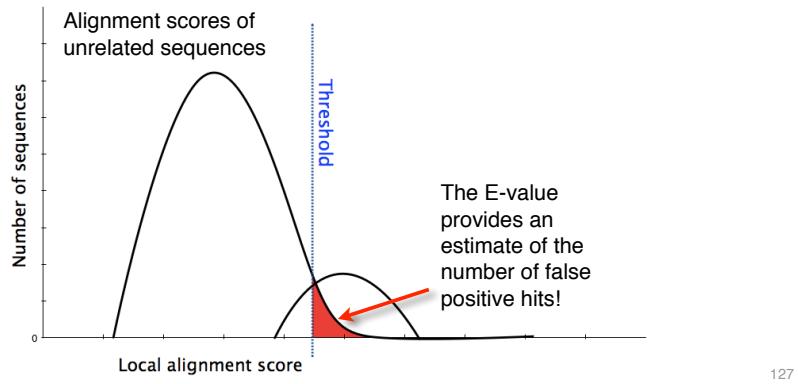
- Ideally, a threshold separates all query related sequences (yellow) from all unrelated sequences (gray)



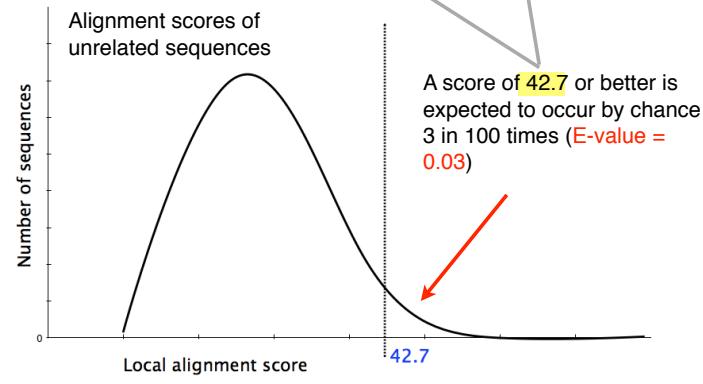
- Unfortunately, often both score distributions overlap
  - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



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



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

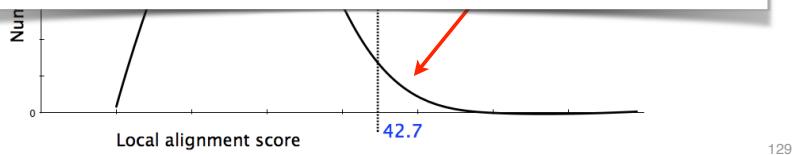


Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo]	677	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	676	100%	0	98%	AAA20133.1

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>



## Practical database searching with BLAST

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

131

## 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

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

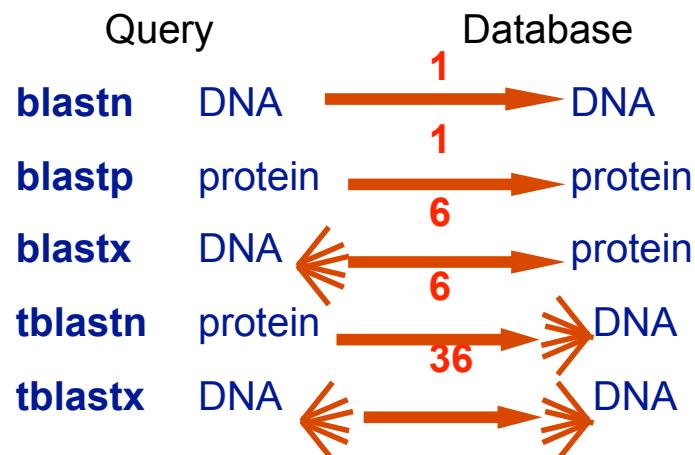
## Step 1: Choose your sequence

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

NCBI Resources How To My N  
Protein Translations of Life Search: Protein Limits Advanced search Help  
Display Settings  FASTA Send to: Change region shown  
hemoglobin subunit beta [Homo sapiens]  
NCBI Reference Sequence NP\_000509.1  
GenPept Graphics  
>gi|4504349|ref|NP\_000509.1| hemoglobin subunit beta [Homo sapiens]  
MVHLTPEEKSATVALWGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPKVKAHGKKVLG  
AFSDGLAHLNLKGTFATLSELHCDKLHVDFENFRLGLNVLCVLAHHFGEFTPPVQAAYQKVVAGVAN  
ALAHKYH

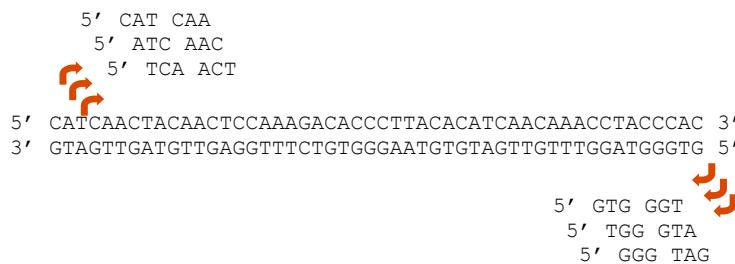
133

## Step 2: Choose the BLAST program



134

DNA potentially encodes six proteins



135

Protein BLAST: search protein databases using a protein query  
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST\_PROGRAMS=blastp&PAGE\_TYPE=BlastSearch  
Reader  
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]  
MVHLTPEEKSATVALWGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPKVKAHGKKVLG  
KVLGAFSDGLAHLNLKGTFATLSELHCDKLHVDFENFRLGLNVLCVLAHHFGEFTPPVQAAYQK  
VVACVANALAHKYH  
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 Optional  
Exclude From To  
Enter Query Entrez Query  
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

136

## Step 3: Choose the database

nr = non-redundant (most general database)  
 dbest = database of expressed sequence tags  
 dbsts = database of sequence tag sites  
 gss = genomic survey sequences

nucleotide databases

protein databases

Organism

Entrez

Settings!

## Step 4a: Select optional search parameters

Expect

Word size

Scoring matrix

## Step 4: Optional parameters

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

## Results page

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

**BLAST® Basic Local Alignment Search Tool**

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/BLAST/blasp suite/Formatting Results - FVGUTMRZ013

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

You can Learn about the enhanced report Blast report description

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

Query ID Id|84677  
Description gi|4504349|ref|NP\_000509.1| hemoglobin subunit beta [Homo sapiens]  
Molecule type amino acid  
Query Length 147

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

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

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

**Graphic Summary**

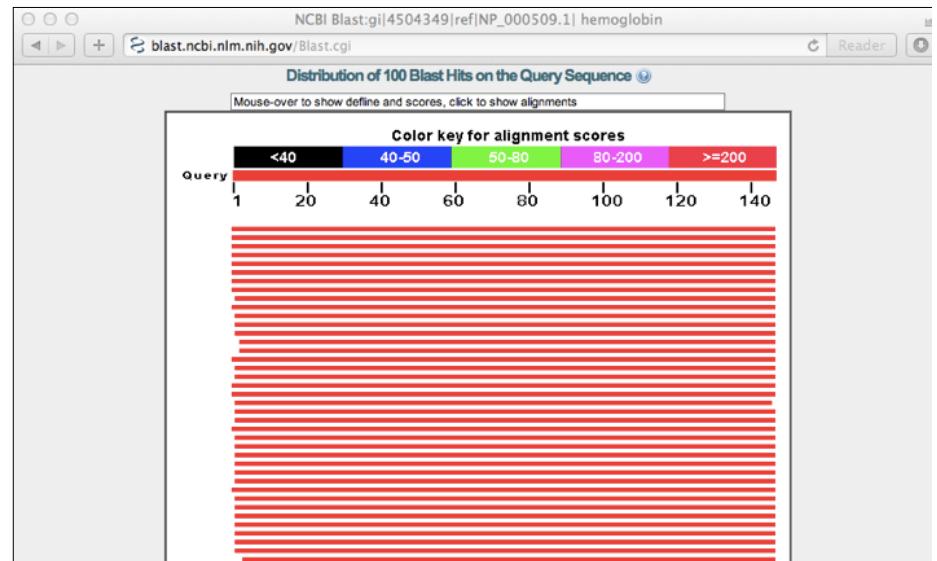
Show Conserved Domains

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

Query seq. 25 54 75 111 125 147  
Specific hits hem-binding site globin  
Superfamilies globin\_like superfamily

Distribution of 100 Blast Hits on the Query Sequence

## Further down the results page...



## Further down the results page...

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

blast.ncbi.nlm.nih.gov/Blast.cgi 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
hemoglobin beta [synthetic construct]	301	301	100%	9e-103	100%	AAX37051.1
hemoglobin beta [synthetic construct]	301	301	100%	1e-102	100%	AAX29557.1
hemoglobin subunit beta [Homo sapiens] >ref XP_508242.1  PREDICTED: hemoglobin s	301	301	100%	1e-102	100%	NP_000509.1
RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta chain	300	300	100%	4e-102	99%	P02024.2
beta-globin chain variant [Homo sapiens]	299	299	100%	5e-102	99%	AAN84548.1
beta-globin [Homo sapiens] >gb AAZ39781.1  beta-globin [Homo sapiens] >gb AAZ39781.1  beta-globin [Homo sapiens]	299	299	100%	5e-102	99%	AAZ39780.1
beta-globin [Homo sapiens]	299	299	100%	5e-102	99%	ACU56984.1
hemoglobin beta chain [Homo sapiens]	299	299	100%	6e-102	99%	AAD19696.1
Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound At The Beta Subunit	298	298	99%	9e-102	100%	1COH_B
hemoglobin beta subunit variant [Homo sapiens] >gb AAA88054.1  beta-globin [Homo sapiens]	298	298	100%	1e-101	99%	AAF00489.1
Chain B, Human Hemoglobin D Los Angeles: Crystal Structure >pdb 2YRSID Chain D, H	298	298	99%	2e-101	99%	2YRS_B
Chain B, High-Resolution X-Ray Study Of Deoxy Recombinant Human Hemoglobins Syn	297	297	99%	3e-101	99%	1DXU_B
Chain B, Analysis Of The Crystal Structure, Molecular Modeling And Infrared Spectroscopic	297	297	99%	3e-101	99%	1HDB_B

## Further down the results page...

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

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

Download GenPept Graphics Next Match Previous Match

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

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 KVHLITPEEKSAVTALNGKVNVDENVGEALGRLRLLVVPNTQRFFESFGDLSTPDAVMGNPK  
Sbjct 1 KVHLITPEEKSAVTALNGKVNDEVENVGEALGRLRLLVVPNTQRFFESFGDLSTPDAVMGNPK

Query 61 VKAHGKVKVLGAFSDGLAHLNDLNLKGTFATLSSELHCDKLHVDPENFRLLGNVLVCVLAHHIFG  
Sbjct 61 VKAHGKVKVLGAFSDGLAHLNDLNLKGTFATLSSELHCDKLHVDPENFRLLGNVLVCVLAHHIFG

Query 121 KEFTPPVQAAAYKVVAAGVANALAHKYI  
Sbjct 121 KEFTPPVQAAAYKVVAAGVANALAHKYI

Range 1: 1 to 147 GenPept Graphics

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

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

Related Information

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

Range 1: 1 to 147 GenPept Graphics

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

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

The screenshot shows the NCBI BLAST search interface. At the top, it displays 'NCBI Blast:gi|4504349|ref|NP\_000509.1| hemoglobin'. Below this is the 'Basic Local Alignment Search Tool' header. A red oval highlights the 'Formatting options' section, which contains various settings for displaying search results. The 'Alignment View' dropdown is set to 'Query-anchored with letters for identities'. The 'Display' section includes 'Graphical Overview' (checked), 'Sequence Retrieval' (checked), and 'NCBI-gi' (unchecked). The 'Limit results' section shows 'Descriptions: 50', 'Graphical overview: 50', and 'Alignments: 50'. There are also fields for 'Organism', 'Entrez query', 'Expect Min.', 'Expect Max.', and 'Percent Identity Min.' and 'Max.'. At the bottom, there's a 'Format for' section with 'PSI-BLAST' checked and an 'inclusion threshold' field.

## E.g. Query anchored alignments

The screenshot shows the NCBI BLAST search results for the query 'NP\_000509'. The results are displayed in a table where each row represents a hit. The columns show the query sequence, the hit ID, the hit sequence, and the percentage identity (ranging from 59 to 60%). The hits are mostly identical, with minor variations indicated by dots and letters. The first few rows are:

Query	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
AAX37051	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
AAX29557	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
NP_000509	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
P02024	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
AAN84548	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
AAZ39780	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
ACU56984	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
AAD19696	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
1COH_B	1	VHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
AAF00489	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
2YRS_B	1	VHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
IDXU_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1HDB_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
IDKV_B	2	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
3KMF_C	2	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
AAE68978	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
1NQP_B	1	VHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1K1K_B	1	VHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
AAN11320	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
XP_002822173	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
1Y85_B	1	VHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1YE0_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1O1O_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
CAA23759	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
1YE2_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1Y5F_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1A00_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1HBS_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1ABY_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1CMY_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59

## ... and alignments with dots for identities

The screenshot shows the NCBI BLAST search results for the query 'NP\_000509'. The results are displayed in a table where each row represents a hit. The columns show the query sequence, the hit ID, the hit sequence, and the percentage identity (ranging from 59 to 60%). The hits are mostly identical, with minor variations indicated by dots and letters. The first few rows are:

Query	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
AAX37051	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
AAX29557	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
NP_000509	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
P02024	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
AAN84548	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
AAZ39780	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
ACU56984	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
AAD19696	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
1COH_B	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
AAF00489	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
2YRS_B	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
IDXU_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1HDB_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
IDKV_B	2	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
3KMF_C	2	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
AAE68978	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
1NQP_B	1	VHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1K1K_B	1	VHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
AAN11320	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
XP_002822173	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
1Y85_B	1	VHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1YE0_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1O1O_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
CAA23759	1	MVHLTPPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	60
1YE2_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1Y5F_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1A00_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1HBS_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1ABY_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	59
1CMY_B	1	MHLTPEEKS...GALGRLLVVYWPWTQRFFESFGDLSTPDAVMGNPK	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

149

## 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)

150

## 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