

**BIMM 143**  
**Introduction to Bioinformatics**  
 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.
- There are a large number of bioinformatics databases (see [handout!](#)).
- The **NCBI** and **EBI** are major online bioinformatics service providers.
- Introduced via **hands-on session** the **BLAST**, **Entrez**, **GENE**, **OMIM**, **UniProt**, **Muscle** and **PDB** bioinformatics tools and databases.
  - Muddy point assessment (see [results](#))
- Also covered: Course structure; Supporting course website, Ethics code, and Introductions...

**Today's Menu**

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

**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 OMIN**
- What is known about the protein family, its species distribution, number in humans and residue-wise conservation? **EBI PFAM**
- Are high resolution protein structures available to examine the details of these mutations? How might we explain their potential molecular effects? **RCSB PDB**

Search NCBI databases

ras

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

Category	Count	Description
<b>Literature</b>		
Books	1,677	books and reports
MeSH	402	ontology used for PubMed indexing
NLM Catalog	223	books, journals and more in the NLM Collections
PubMed	54,672	scientific & medical abstracts/citations
PubMed Central	96,114	full-text journal articles
<b>Health</b>		
ClinVar	759	human variations of clinical significance
dbGaP	120	genotype/phenotype interaction studies
GTR	1,879	genetic testing registry
<b>Genes</b>		
EST	3,985	expressed sequence tag sequences
Gene	87,165	collected information about gene loci
GEO DataSets	3,732	functional genomics studies
GEO Profiles	1,822,789	gene expression and molecular abundance profiles
HomoloGene	696	homologous gene sets for selected organisms
PopSet	2,254	sequence sets from phylogenetic and population studies
UniGene	4,770	clusters of expressed transcripts
<b>Proteins</b>		

Gene

ras

Save search Advanced Search Help

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

Filters: Manage Filters

Top Organisms [Tree]

- Homo sapiens (1126)
- Mus musculus (829)
- Rattus norvegicus (625)
- Oryzomys latipes (533)
- Neolamprologus brichardi (507)
- All other taxa (82019)

Results: 1 to 20 of 85633

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

Name/Gene ID	Description	Location	Aliases
ras ID: 19412	resistance to autologous selizuro (Mus musculus [house mouse])		asr
ras ID: 43873	raspberry [Drosophila melanogaster [fruit fly]]	Chromosome X, NC_004354.4 (10744502..10749007)	Dmel_CG1799, CG11485, CG1799, DmelCG1799, EPX1093

Gene

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

Save search Advanced Search Help

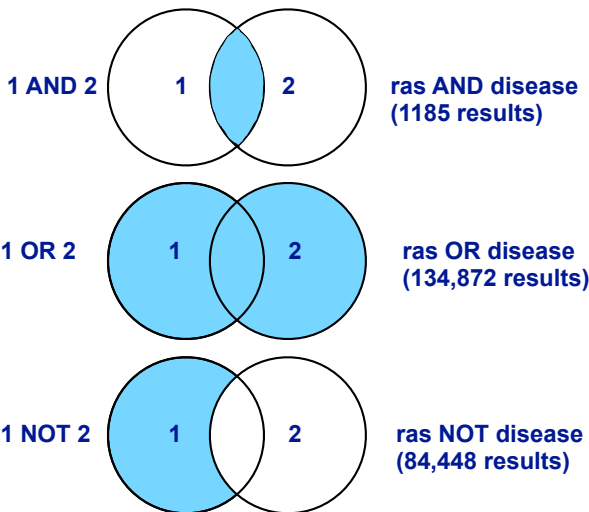
Display Settings: Tabular, 20 per page, Sorted by Relevance

Filters: Manage Filters

Results: 1 to 20 of 1126

Filters activated: Current only. Clear all to show 1439 items.

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4953	neuroblastoma RAS viral (v-ras) oncogene homolog (Homo sapiens [human])	Chromosome 1, NC_000011.11 (114704464..114716884, complement)	RPS100E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3648	Kirsten rat sarcoma viral oncogene homolog (Homo sapiens [human])	Chromosome 12, NC_000012.12 (25205248..2520923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS6



Gene

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

Save search Advanced Search Help

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

Filters: Manage Filters

Results: 1 to 20 of 1126

Filters activated: Current only. Clear all to show 1439 items.

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4953	neuroblastoma RAS viral (v-ras) oncogene homolog (Homo sapiens [human])	Chromosome 1, NC_000011.11 (114704464..114716884, complement)	RPS100E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3648	Kirsten rat sarcoma viral oncogene homolog (Homo sapiens [human])	Chromosome 12, NC_000012.12 (25205248..2520923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS6

NCBI Resources How To Sign in to NCBI

Gene KRAS

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

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

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

**Summary**

Official Symbol: KRAS provided by HGNC  
 Official Full Name: Kirsten rat sarcoma viral oncogene homolog provided by HGNC  
 Primary source: HGNC:HGNC:8407  
 See related: Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTFLUMG00000171193

Genetic type: protein coding  
 RefSeq status: REVIEWED  
 Organism: *Homo sapiens*  
 Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Homiidae; 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 pseudogenes, Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)

NCBI Resources How To Sign in to NCBI

Gene KRAS

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

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

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

**Summary**

Official Symbol: KRAS provided by HGNC  
 Official Full Name: Kirsten rat sarcoma viral oncogene homolog provided by HGNC  
 Primary source: HGNC:HGNC:8407  
 See related: Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTFLUMG00000171193

Genetic type: protein coding  
 RefSeq status: REVIEWED  
 Organism: *Homo sapiens*  
 Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Homiidae; 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 pseudogenes, Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)

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

Genomic context

Location: 12p12.1 See KRAS in Epigenomics, MapViewer

Exon count: 6

Annotation release	Status	Assembly	Chr.	Location
108	current	GRCh38 (GCF_000001405.25)	12	NC_000012.12 (25205246..25250823, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	12	NC_000012.11 (25358190..25403870, complement)

Chromosome 12 - NC\_000012.12

Genomic regions, transcripts, and products

Genomic Sequence: NC\_000012.12 chromosome 12 reference GRCh38 Primary Assembly

NCBI Resources How To Sign in to NCBI

Gene KRAS

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

**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:8407  
 See related: Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTFLUMG00000171193

Genetic type: protein coding  
 RefSeq status: REVIEWED  
 Organism: *Homo sapiens*  
 Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Homiidae; 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 pseudogenes, Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)

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

Gene Ontology Provided by GOA

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

Items 1 - 25 of 33

Process	Evidence Code	Pubmed
Fc-gamma 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

EMBL-EBI UniProt-GOA

Search

Example: GO:0006115, isoprenylation, P08727

Overview | New to UniProt-GOA | FAQ | Contact Us

### Gene Ontology Annotation (UniProt-GOA) Database

The UniProt GO annotation program aims to provide high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB). The assignment of GO terms to UniProt records is an integral part of UniProt bioinformatics. 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.

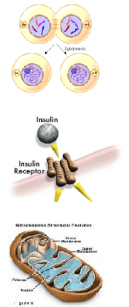
- Downloads
- Searching UniProt-GOA
- Annotation Methods
  - Annotation Tutorial
  - Manual Annotation Efforts
- Reference Genome Annotation Initiative
- Cardiovascular Gene Ontology Annotation Initiative
- Renal Gene Ontology Annotation Initiative
- Enzyme Gene

# 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



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

Scroll down to Very bottom for UniProt link

UniProt will detail much more information for protein coding genes

Reviewed - Experimental evidence at protein level!

View FASTA file format

UniProt will detail much more information for protein coding genes

View FASTA file format

UniProt will detail much more information for protein coding genes

**Function**  
Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23695361, PubMed:22711838). @2 Publications @ Citations

**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

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding <sup>1</sup>	10 - 18	9	GTP @2 Publications			
Nucleotide binding <sup>1</sup>	29 - 35	7	GTP @2 Publications			
Nucleotide binding <sup>1</sup>	59 - 60	2	GTP @3 Publications			

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

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding <sup>1</sup>	10 - 18	9	GTP @2 Publications			
Nucleotide binding <sup>1</sup>	29 - 35	7	GTP @2 Publications			
Nucleotide binding <sup>1</sup>	59 - 60	2	GTP @3 Publications			

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

**Pathology & Biotech**  
**Involvement in disease**  
LEUKEMIA, ACUTE MYELOIDOUS (AML) [MIM:601635]: 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 tissues. Myelogenous leukemias develop from changes in cells that normally produce neutrophils, basophils, eosinophils and monocytes. @1 Publications

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Natural variant <sup>1</sup>	10 - 10	1	G - GC in one individual with AML; expression in JTC3 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	

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

**Structure**  
Secondary structure  
Legend: Helix Turn Beta strand  
Show more details

3D structure databases

Select the 3D structure database	Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
<input type="checkbox"/> RCSB PDB	1D8D	X-ray	2.00	P	178-188	[+]
<input checked="" type="checkbox"/> RCSB PDB	1D8E	X-ray	3.00	P	178-188	[+]
<input type="checkbox"/> PDB	1K2O	X-ray	2.20	C	169-173	[+]
<input type="checkbox"/> PDB	1K2P	X-ray	2.10	C	169-173	[+]
<input type="checkbox"/> PDB	3GFT	X-ray	2.27	A/B/C/D/E/F	1-154	[+]
<input type="checkbox"/> PDB	4D5N	X-ray	2.03	A	2-154	[+]
<input type="checkbox"/> PDB	4D5O	X-ray	1.85	A	2-154	[+]
<input type="checkbox"/> PDB	4E9K	X-ray	2.00	A	1-154	[+]
<input type="checkbox"/> PDB	4EPT	X-ray	2.00	A	1-154	[+]
<input checked="" type="checkbox"/> PDB	4EPV	X-ray	1.35	A	1-154	[+]
<input type="checkbox"/> PDB	4EPW	X-ray	1.70	A	1-154	[+]
<input type="checkbox"/> PDB	4EPX	X-ray	1.76	A	1-154	[+]
<input type="checkbox"/> PDB	4EPY	X-ray	1.80	A	1-154	[+]
<input type="checkbox"/> PDB	4L8G	X-ray	1.52	A	1-154	[+]
<input type="checkbox"/> PDB	4LDJ	X-ray	1.15	A	1-154	[+]
<input type="checkbox"/> PDB	4L7K	X-ray	1.50	A/B	1-160	[+]

**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?

**3D View**

**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., Burns, J.D., Phan, J., Burns, M.C., Olejniczak, E.T., Waters, A.G., Lee, J., Rossman, G.W., Fesik, S.W.  
Organism: Homo sapiens  
Expression System: Escherichia coli  
Mutation(s): 1

**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

Display Options

- Assembly: Blower-ruy 1
- Model: Model 1
- Symmetry: None
- Interaction: GDP(201A)**
- Style: Cartoon
- Color: Rainbow
- Ligand: None
- Quality: Automatic



Family: *Kinesin* (PF00225)

HHMI janelia farm research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

126 architectures 4150 sequences 6 interactions 248 species 114 structures

**Summary**

**Domain organisation**

**Alignments**

**HMM logo**

**Trees**

**Curation & models**

**Species**

**Interactions**

**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 PDBsum 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
ARBK01_GIALA	11 - 335	2vva	A	11 - 335	<a href="#">Jmol AstexViewer SPICE</a>
			B	11 - 335	<a href="#">Jmol AstexViewer SPICE</a>
CENPL_HUMAN	12 - 329	115c	A	12 - 329	<a href="#">Jmol AstexViewer SPICE</a>
			B	12 - 329	<a href="#">Jmol AstexViewer SPICE</a>
KAR3_YEAST	392 - 723	1f9t	A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
K113B_HUMAN	11 - 352	3qbi	A	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
			B	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
			C	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
		1i6	A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
		1q0b	A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
		1x88	A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>

Jump to...  
enter ID:acc. Go

Family: *Kinesin* (PF00225)

HHMI janelia farm research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

126 architectures 4150 sequences 6 interactions 248 species 114 structures

**Summary**

**Domain organisation**

**Alignments**

**HMM logo**

**Trees**

**Curation & models**

**Species**

**Interactions**

**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 PDBsum 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
ARBK01_GIALA	11 - 335	2vva	A	11 - 335	<a href="#">Jmol AstexViewer SPICE</a>
			B	11 - 335	<a href="#">Jmol AstexViewer SPICE</a>
CENPL_HUMAN	12 - 329	115c	A	12 - 329	<a href="#">Jmol AstexViewer SPICE</a>
			B	12 - 329	<a href="#">Jmol AstexViewer SPICE</a>
KAR3_YEAST	392 - 723	1f9t	A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
K113B_HUMAN	11 - 352	3qbi	A	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
			B	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
			C	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
		1i6	A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
		1q0b	A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
		1x88	A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>

Jump to...  
enter ID:acc. Go

**Your turn:**  
What can you find out about "eg5"?

Jmol

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

Close window

# Today's Menu

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

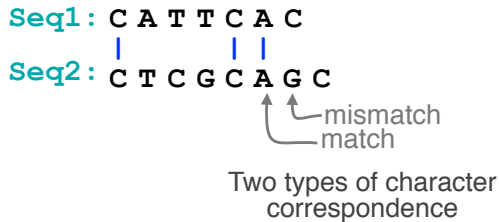
**Basic Idea:** Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.

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

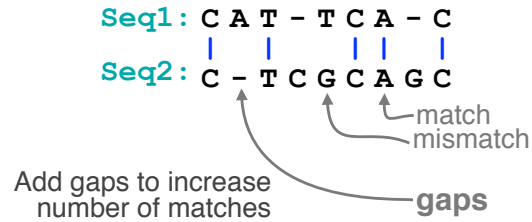
**Seq2:** C T C G C A G C

[Screencast Material]

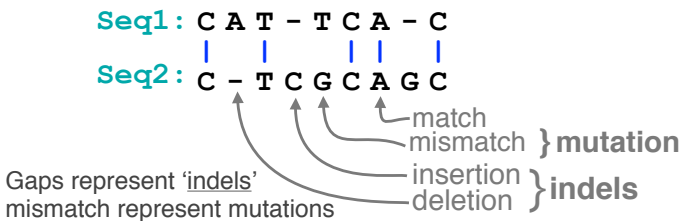
**Basic Idea:** Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.



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



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



### Why compare biological sequences?

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

### Practical applications include...

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

### Practical applications include...

- **Similarity searching of databases**
  - Protein structure prediction
- **Assembly of 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

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



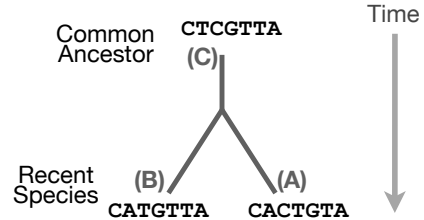
# ALIGNMENT FOUNDATIONS

- 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.

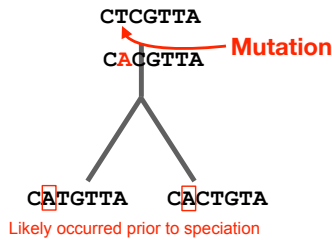
- Mutations/Substitutions
- Deletions
- Insertions



# Mutations, deletions and insertions

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

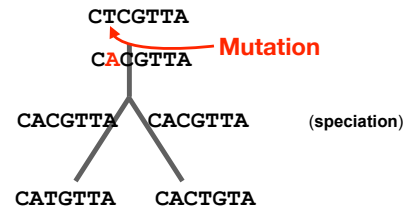
- **Mutations/Substitutions**    CTCGTTA → CACGTTA
- Deletions
- Insertions



# Mutations, deletions and insertions

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

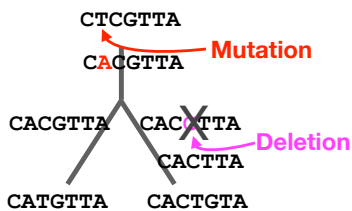
- **Mutations/Substitutions**    CTCGTTA → CACGTTA
- Deletions
- Insertions



# Mutations, deletions and insertions

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

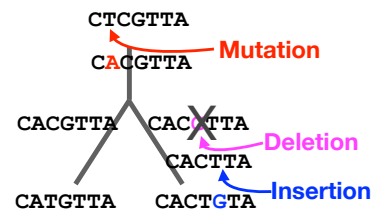
- **Mutations/Substitutions**    CTCGTTA → CACGTTA
- **Deletions**                    CACGTTA → CACTTA
- Insertions



# Mutations, deletions and insertions

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

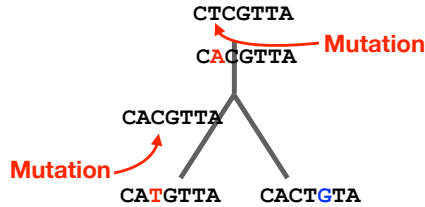
- **Mutations/Substitutions**    CTCGTTA → CACGTTA
- **Deletions**                    CACGTTA → CACTTA
- **Insertions**                    CACTTA → CACTGTA



## Mutations, deletions and insertions

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

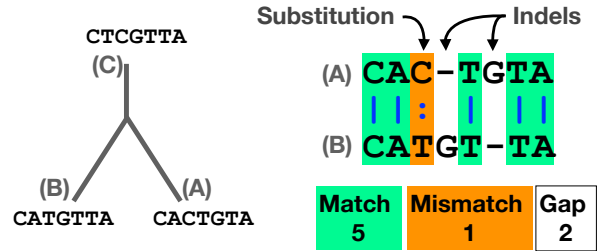
- **Mutations/Substitutions** CTCGTTA → CACGTTA
- Deletions CACGTTA → CATGTTA
- Insertions



## Alignment view

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

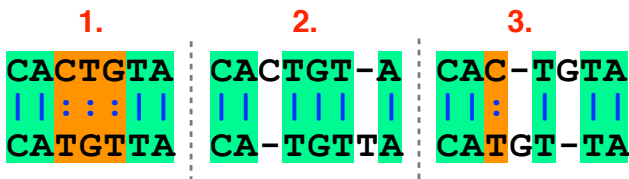
- **Mismatches** represent mutations/substitutions
- **Gaps** represent insertions and deletions (indels)



## Alternative alignments

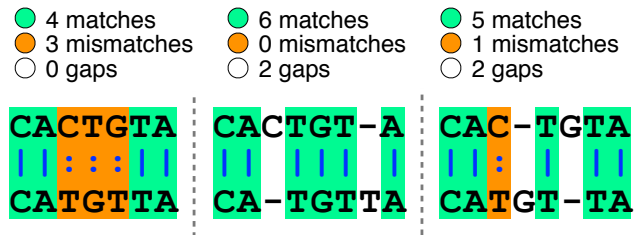
- Unfortunately, finding the correct alignment is difficult if we do not know the evolutionary history of the two sequences

Q. Which of these 3 possible alignments is best?



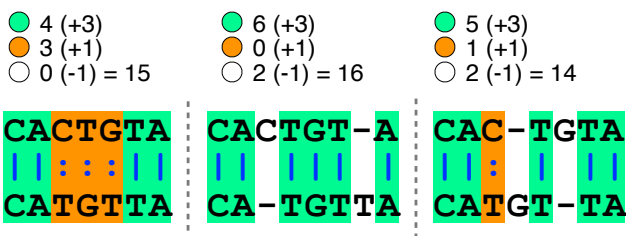
## Alternative alignments

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



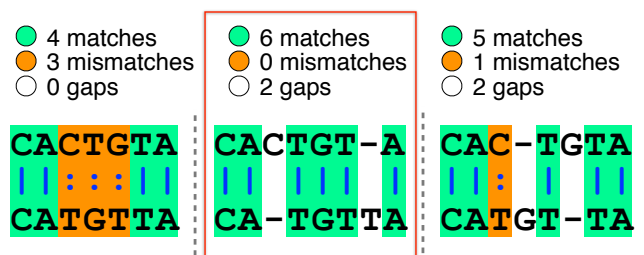
## Scoring alignments

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



## 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.

● 4 matches  
● 3 mismatches  
○ 0 gaps

CACTGTA  
| | | | |  
CATGTTA

● 6 matches  
● 0 mismatches  
○ 2 gaps

CACTGT-A  
| | | | |  
CA-TGTTA

● 5 matches  
● 1 mismatches  
○ 2 gaps

CAC-TGTA  
| | | | |  
CATGT-TA

## Optimal alignments

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

● 4 matches  
● 3 mismatches  
○ 0 gaps

CACTGTA  
| | | | |  
CATGTTA

● 6 matches  
● 0 mismatches  
○ 2 gaps

CACTG-TA  
| | | | |  
CA-TGTTA

● 5 matches  
● 1 mismatches  
○ 2 gaps

CAC-TGTA  
| | | | |  
CATGT-TA

## Optimal alignments

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

● 4 matches  
● 3 mismatches  
○ 0 gaps

CACTGTA  
| | | | |  
CATGTTA

● 6 matches  
● 0 mismatches  
○ 2 gaps

CACTGT-A  
| | | | |  
CA-TGTTA

CAC-TGTA  
| | | | |  
CATGT-TA

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

## ALIGNMENT FOUNDATIONS

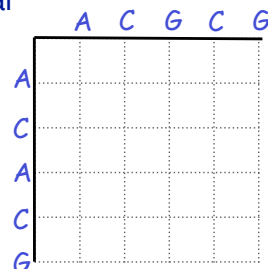
- Why...**
  - Why compare biological sequences?
- What...**
  - Alignment view of sequence changes during evolution (matches, mismatches and gaps)
- How...**
  - Dot matrices
  - Dynamic programming
    - Global alignment
    - Local alignment
  - BLAST heuristic approach

## ALIGNMENT FOUNDATIONS

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

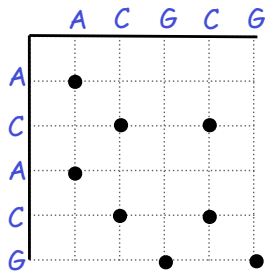
## Dot plots: simple graphical approach

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



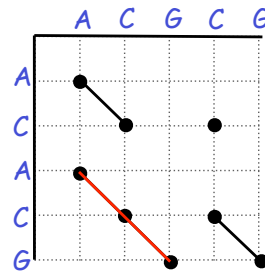
## Dot plots: simple graphical approach

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



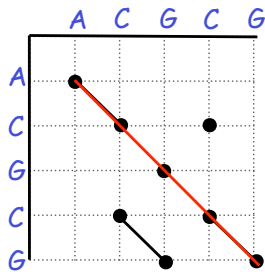
## Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



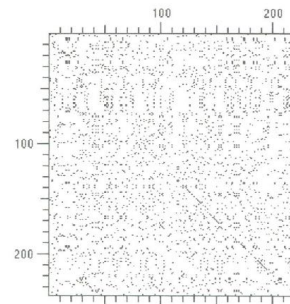
## Dot plots: simple graphical approach

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



## Dot plots: simple graphical approach

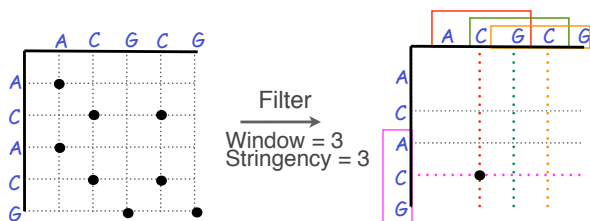
- Dot matrices for long sequences can be noisy



## Dot plots: window size and match stringency

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

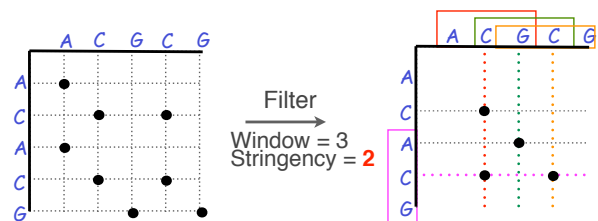
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
- You have to choose window size and stringency



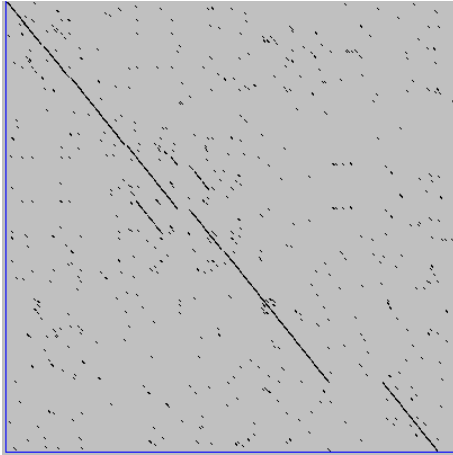
## Dot plots: window size and match stringency

**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



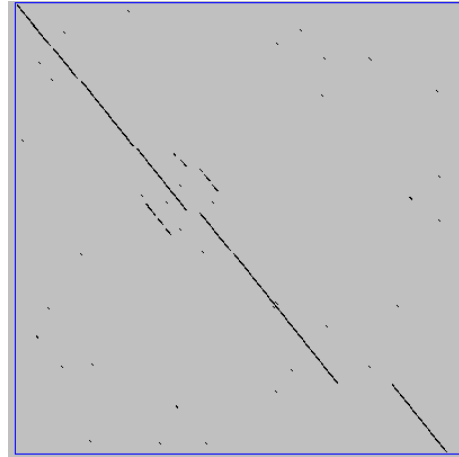
## Window size = 5 bases



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



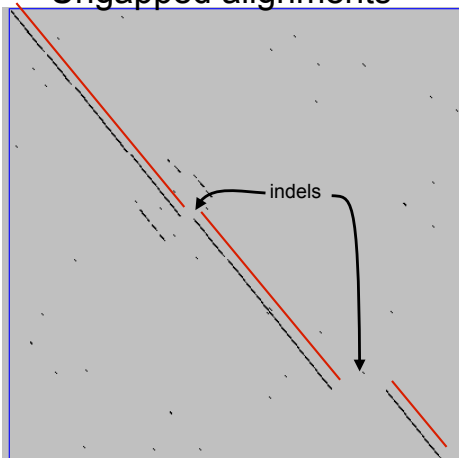
This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

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

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

## Ungapped alignments



Only **diagonals** can be followed.

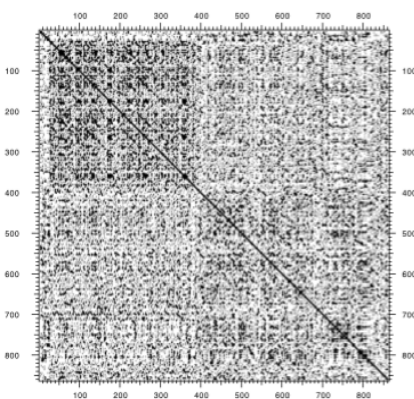
Downward or rightward paths represent **insertion** or **deletions** (gaps in one sequence or the other).

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

## Uses for dot matrices

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

## Repeats

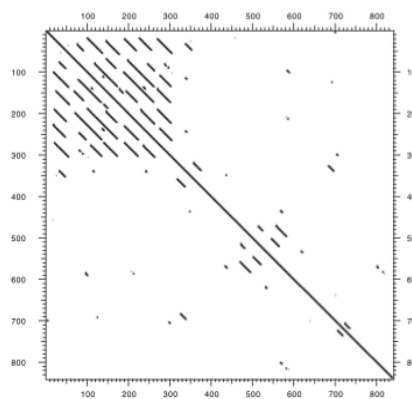


Human LDL receptor protein sequence (Genbank P01130)

$W = 1$   
 $S = 1$

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

## Repeats



Human LDL receptor protein sequence (Genbank P01130)

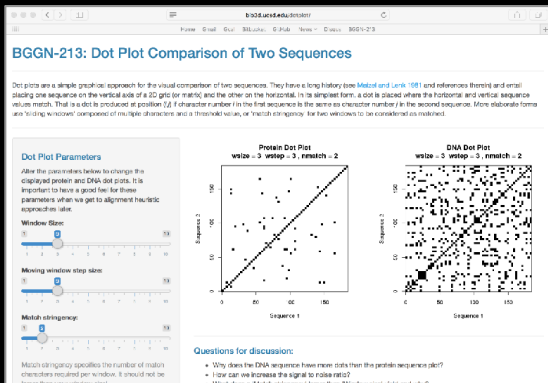
$W = 23$   
 $S = 7$

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

# Your Turn!

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

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



## ALIGNMENT FOUNDATIONS

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

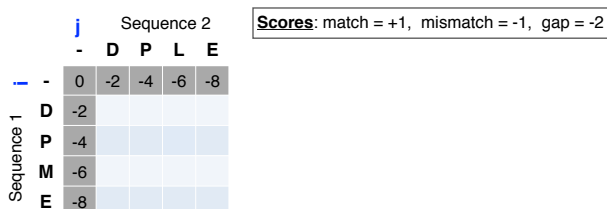
- The Needleman–Wunsch approach to global sequence alignment has three basic steps:
  - (1) setting up a 2D-grid (or **alignment matrix**),
  - (2) **scoring the matrix**, and
  - (3) identifying the **optimal path** through the matrix



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.

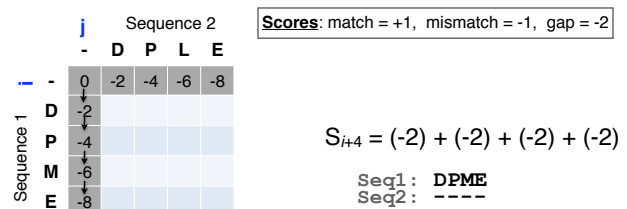
## Scoring the alignment matrix

- 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



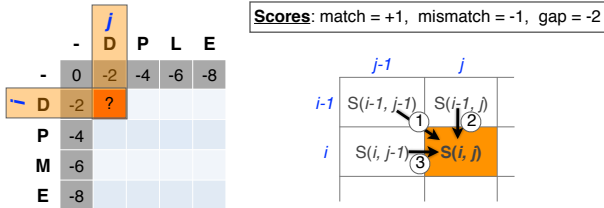
## Scoring the alignment matrix

- 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



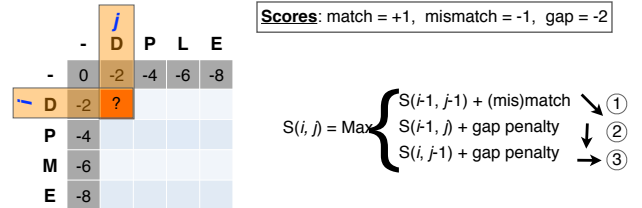
## Scoring the alignment matrix

- 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



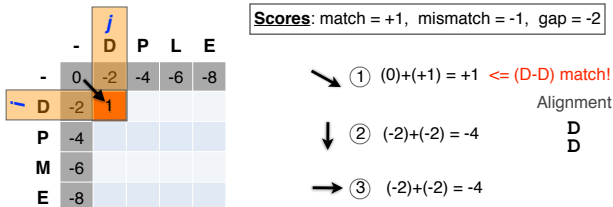
## Scoring the alignment matrix

- 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



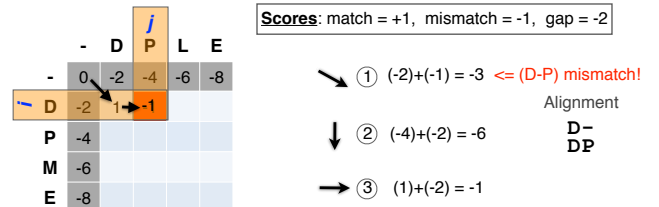
## Scoring the alignment matrix

- 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



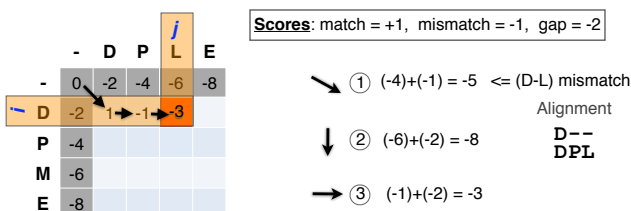
## Scoring the alignment matrix

- 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)



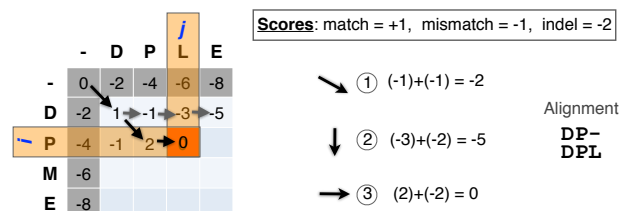
## Scoring the alignment matrix

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



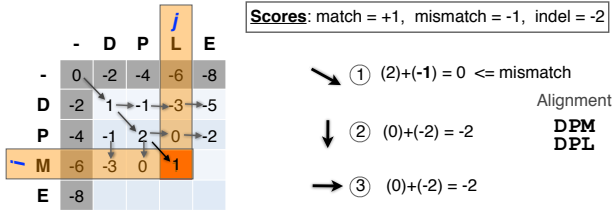
## Scoring the alignment matrix

- 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.



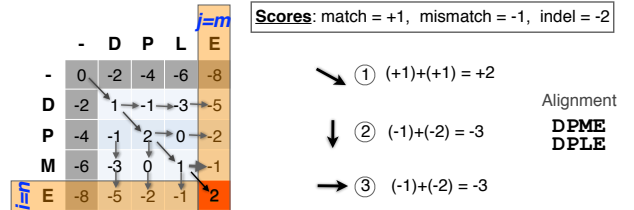
## Scoring the alignment matrix

- 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



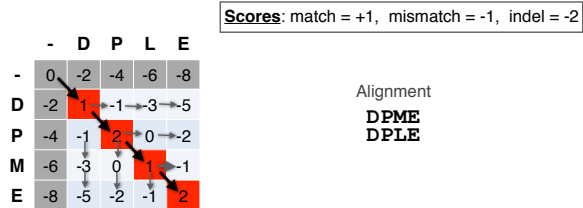
## Scoring the alignment matrix

- 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)



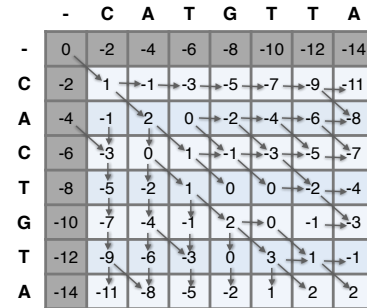
## Scoring the alignment matrix

- 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



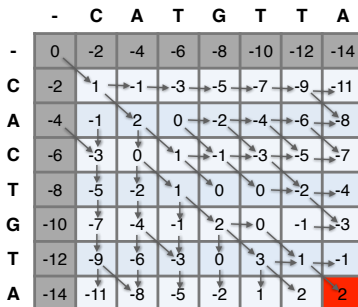
## Questions:

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



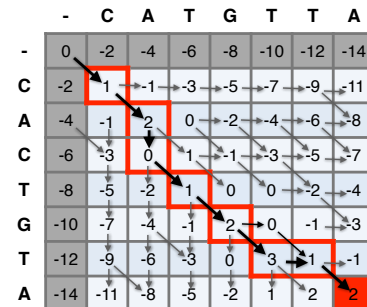
## Questions:

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



## Questions:

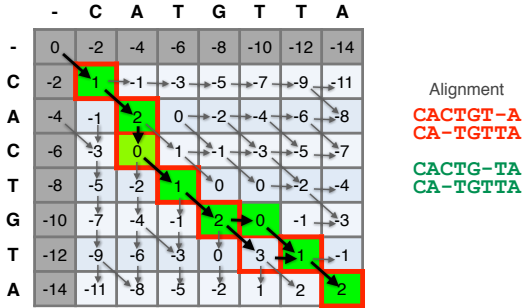
- To find the best alignment we retrace the arrows starting from the bottom right cell





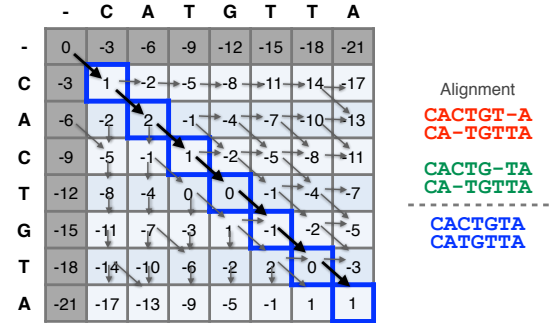
## More than one alignment possible

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



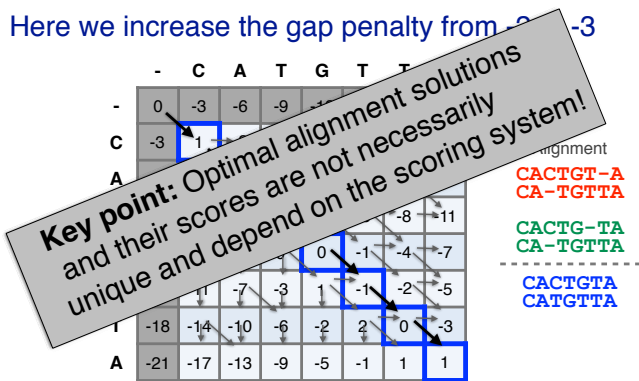
## The alignment and score are dependent on the scoring system

- 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 -2 to -3



## Your Turn!

### Hands-on worksheet Sections 2 & 3

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

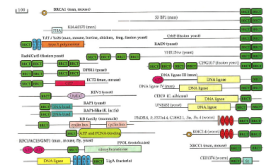
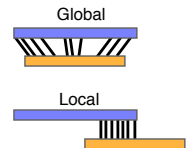
		A	G	T	T	C
	0					
A						
T						
T						
G						
C						

## 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

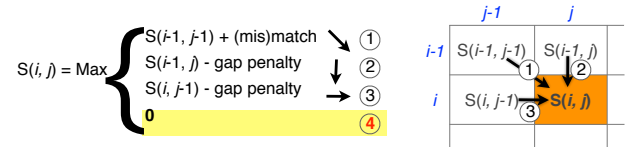
- Smith & Waterman proposed simply that a local alignment of two sequences allow arbitrary-length segments of each sequence to be aligned, with no penalty for the unaligned portions of the sequences. Otherwise, the score for a local alignment is calculated the same way as that for a global alignment

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

104

## 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



105

		Sequence 1															
		-	C	A	G	C	C	U	C	G	C	U	U	A	G		
Sequence 2	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	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	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	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	0.0	1.0	1.0	0.0	0.7	0.0
	G	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.0	0.7	0.7	1.0	0.0
	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	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	1.0	0.0	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	0.0	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	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	1.0	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	2.0	2.0
	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	2.0	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	2.0	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	2.0	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	2.0	2.0	

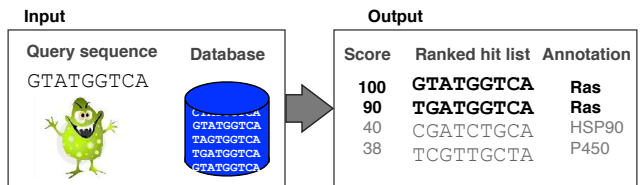
Local alignment  
GCC-AUG  
GCCUCGC

106

## 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

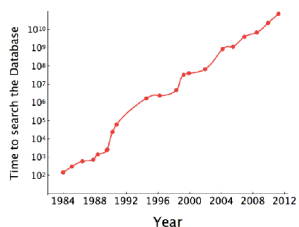
- Input:** Q, D and scoring scheme
- Output:** Ranked list of hits



107

## 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!**

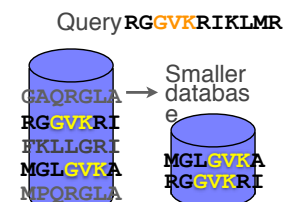


To reduce search time **heuristic algorithms**, such as BLAST, first remove database sequences without a strong local similarity to the query sequence in a quick initial scan.

108

## 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.

109

## 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

111

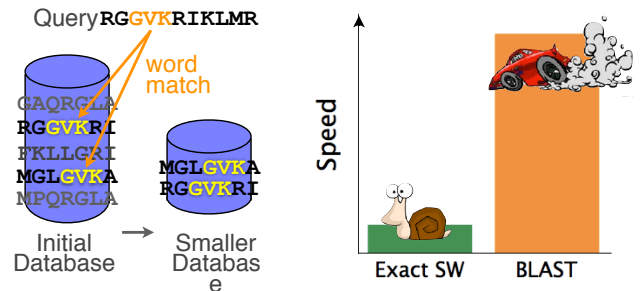
## Rapid, heuristic versions of Smith–Waterman: **BLAST**

- 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 finds regions of local similarity between query sequences
  - 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

“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)

112

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



113

## 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

114

## Blast

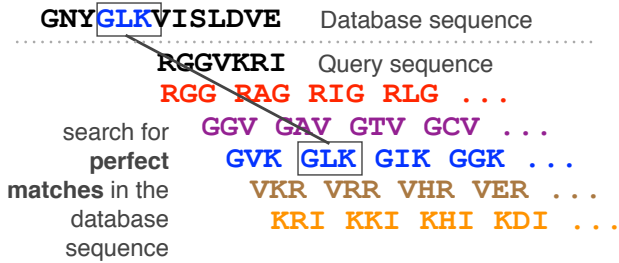
- **Phase 2:** expand word pairs to include those similar to query (defined as those above a similarity threshold to original word, i.e. match scores in substitution matrix)

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

115

## Blast

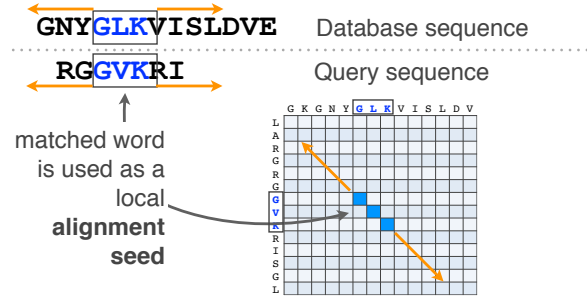
- **Phase 3:** a database is scanned to find sequence entries that match the compiled word list



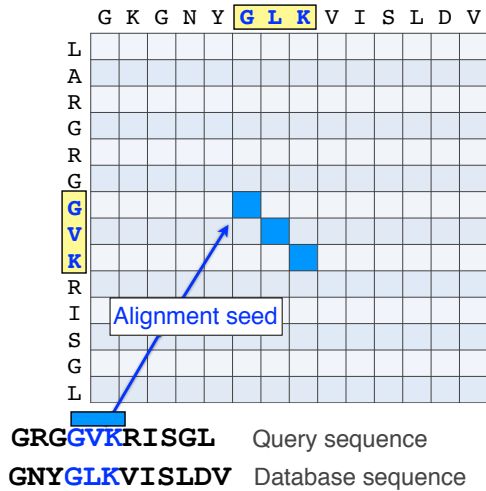
116

## Blast

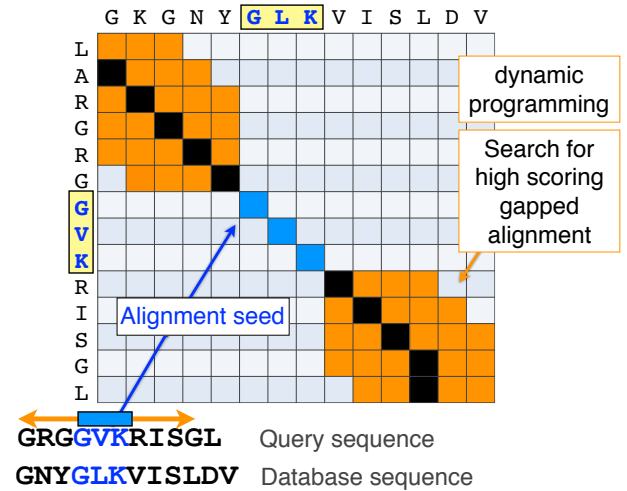
- **Phase 4:** the initial database hits are extended in both directions using dynamic programming



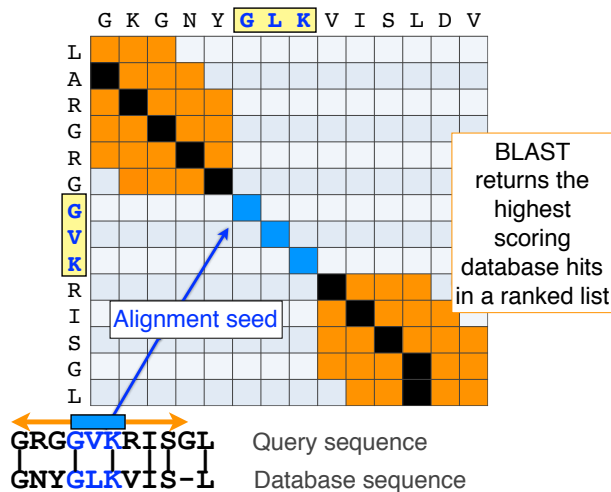
117



118



119



120

## 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

121

## 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

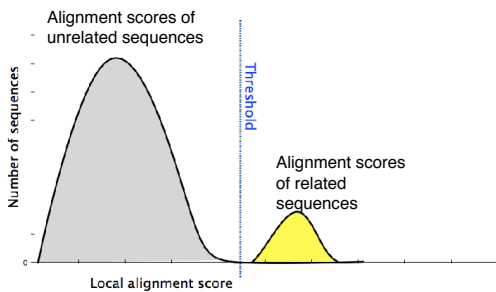
122

## 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

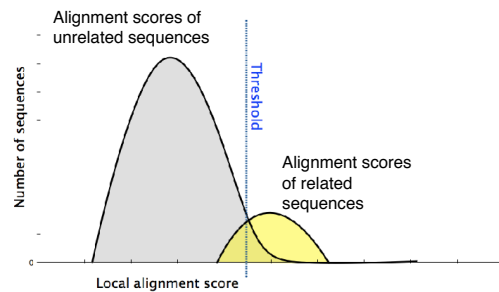
123

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



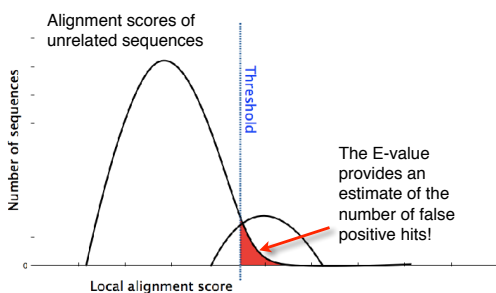
124

- 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



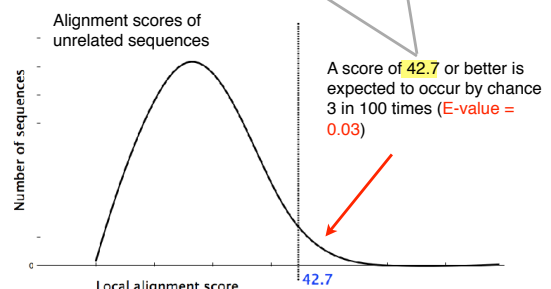
125

- 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



126

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



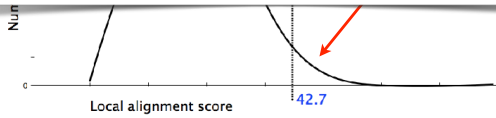
127

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

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

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

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



128

## 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

130

## 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”

131

## Step 1: Choose your sequence

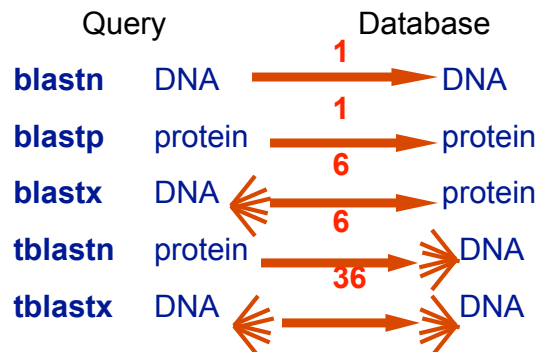
- Sequence can be input in FASTA format or as accession number

```

>gi|45043491|ref|NP_000509.1| hemoglobin subunit beta (Homo sapiens)
MVHLTPEEKSAVTALWGKLVNDEVGSEALGRLLVVPTQRFESFGDLSTPDAVGNPKVKAHGKEVVG
AFSDGLAHLNLRKQTFATLSLCLRDLRVDPEFRLGNVLCVLAHFFGKEFTPPVQAAYQKRVAGVAV
ALANKVY
  
```

132

## Step 2: Choose the BLAST program



133

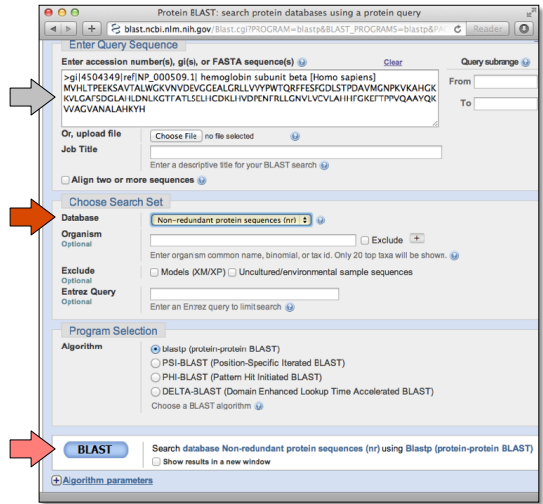
# DNA potentially encodes six proteins

```

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

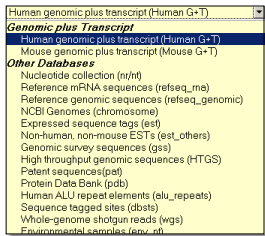
5' CATCAACTACAACCTCCAAAGACACCCCTTACACATCAACAAACCTACCCAC 3'
3' GTAGTTGATGTTGAGGTTTCTGTGGGAATGTGTAGTTGTTGGATGGGTG 5'

5' GTG GGT
5' TGG GTA
5' GGG TAG
    
```

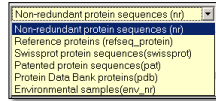


## Step 3: Choose the database

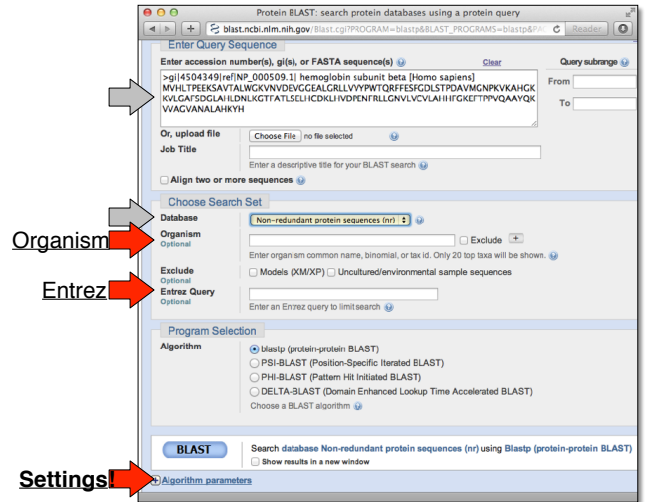
- nr = non-redundant (most general database)
- dbest = database of expressed sequence tags
- dbsts = database of sequence tag sites
- gss = genomic survey sequences



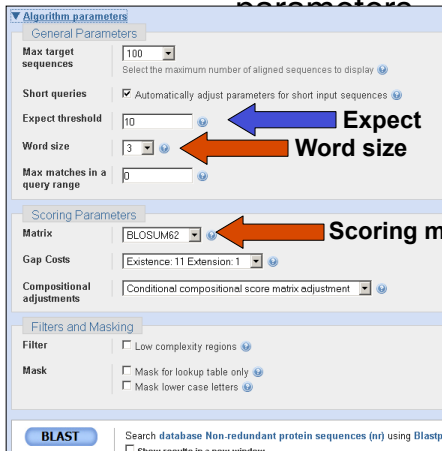
nucleotide databases



protein databases



## Step 4a: Select optional search parameters



## Step 4: Optional parameters

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

## Results page

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

BLAST® Basic Local Alignment Search Tool

Query ID: gi|4504349|ref|NP\_000509.1| hemoglobin

Database Name: nr

Program: BLASTP

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

## Further down the results page...

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

Distribution of 100 Blast Hits on the Query Sequence

Color key to show alignments scores

Query: 1 20 40 60 80 100 120 140

## Further down the results page...

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

Sequences producing significant alignments:

Description	Max score	Total score	Query cover	E value	Max ident	Accession
hemoglobin beta [synthetic construct]	301	301	100%	9e-103	100%	AAK3705.1
hemoglobin beta [synthetic construct]	301	301	100%	1e-102	100%	AAK29557.1
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 chain variant [Homo sapiens]	299	299	100%	5e-102	99%	P02024.2
beta globin [Homo sapiens] >gb AAZ39781.1  beta globin [Homo sapiens] >gb AAZ39781	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%	AAI19896.1
Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound A	298	298	99%	9e-102	100%	1CQH_B
hemoglobin beta subunit variant [Homo sapiens] >gb AAA8054.1  beta-globin [Homo s	298	298	100%	1e-101	99%	AAF00489.1
Chain B, Human Hemoglobin D, Los Angeles, Crystal Structure >pdb 2YRSJD  Chain D, H	298	298	99%	2e-101	99%	2YRS_B
Chain B, High-Resolution X-Ray Study Of Deoxy Recombinant Human Hemoglobins Syn	297	297	99%	3e-101	99%	1DXU_B
Chain B, Analysis Of The Crystal Structure, Molecular Modeling And Infrared Spectroscop	297	297	99%	3e-101	99%	1HDB_B

## Further down the results page...

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

hemoglobin subunit beta [Homo sapiens]

Sequence ID: ref|NP\_000509.1| Length: 147 Number of Matches: 1

Expect: 1e-102

Related Information: UniGene, Map Viewer, PubChem Bio

## Different output formats are available

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

BLAST® Basic Local Alignment Search Tool

Formatting options

Show: Alignment as HTML

Alignment View: Query-anchored with letters for identities

Masking: Character Lower Case

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

## E.g. Query anchored alignments

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

Query-anchored alignments

Query: AAK3705.1, AAK29557, NP\_000509, P02024, AAN84548, AAN39780, ACU56984, AAI19896, 1CQH\_B, AAF00489, 2YRS\_B, 1DXU\_B, 1HDB\_B, AAL69278, AAL002, AAN10489, AAN11320, XP\_002822173, 1YB5\_B, 1YB0\_B, 1010\_B, CAK37259, 1YB2\_B, 1YSP\_B, 1A00\_B, 1HDB\_S, 1AN1\_B, 1CHY\_B



## ... and alignments with dots for identities

Accession	Identity
Query	1
AAK37031	1
AAK29557	1
NP_000509	1
P02024	1
AAN84548	1
AAN39790	1
ACU56984	1
AAJ19696	1
LSOJL	1
AAN90489	1
ZY88_B	1
LDXU_B	1
LH99_B	1
LDXV_B	2
3KHP_C	2
AAI68978	1
LH9P_B	1
LK1K_B	1
AAN11320	1
XP_007822173	1
LY85_B	1
LY80_B	1
LO10_B	1
CAK3759	1
LY82_B	1
LY5P_B	1
LA00_B	1

## 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?

147

## 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

148

## 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)

149

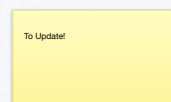
## Summary of key points

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