

Recap From Last Time:

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

Today's Menu

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

Primary, secondary & composite databases

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

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

DATABASE VIGNETTE

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

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

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

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

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

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

The screenshot shows the NCBI homepage with a search bar containing 'ras'. The search results are displayed on the right, including sections for Genotypes and Phenotypes and NCBI Announcements. The 'Genotypes and Phenotypes' section features a diagram of a pedigree chart.

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 website. The results are categorized into Literature, Genes, Health, and Proteins. The 'Genes' category is highlighted with a red box, and a red arrow points from the 'Gene' result in the search results to the 'Gene' section in the summary. The summary text reads: 'About 2,978,774 search results for "ras"'. The 'Gene' section shows 87,165 results.

NCBI Resources How To Sign in to NCBI

Gene Gene Search Help

Show additional filters Save search Advanced

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)]		ras
ID: 19412			
ras	rasberry [Drosophila melanogaster (fruit fly)]	Chromosome X, NC_004354.4	Dmel_CG1799, CG11485, CG1799, DmelCG1799, EP(X)1093
ID: 43873			

Find related data Database: Select Find items

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

Top Organisms [Tree] Homo sapiens (1126) Mus musculus (623) Rattus norvegicus (625) Oryctolobus niloticus (533) Neotamias leucurus (507) All other taxa (82019) More...

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

NCBI Resources How To Sign in to NCBI

Gene Gene Search Help

Show additional filters Save search Advanced

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

Filters: Manage Filters

Results: 1 to 20 of 1126

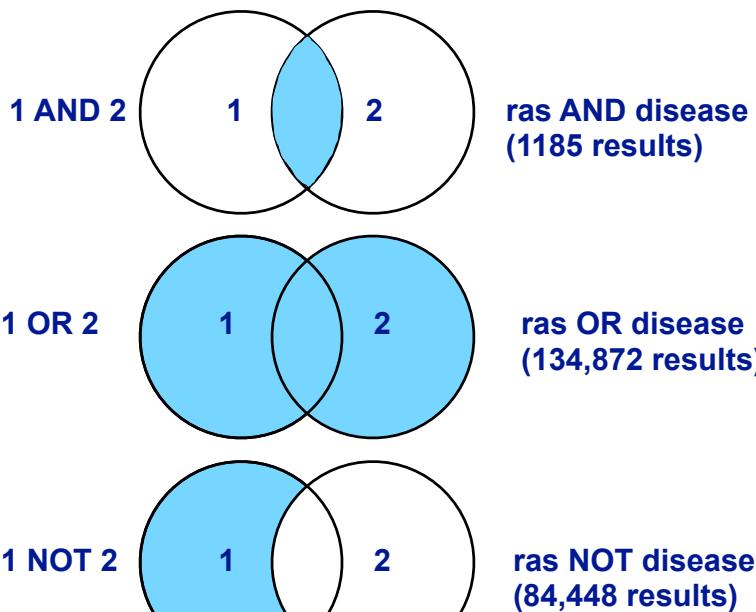
Filters activated: Current only. Clear all to show 1499 items.

Name/Gene ID	Description	Location	Aliases
NRAS	neuroblastoma oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (114704454..114716894, complement)	RP5-1000E10.2, ALPSA, CMNS, N-ras, NCMS1, NS6, NRAS
ID: 4893			
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, KRAS1, KRAS2, NS, NS2, K-RAS2
ID: 3645			

Find related data Database: Select Find items

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

Recent activity Turn Off Clear



11

NCBI Resources How To Sign in to NCBI

Gene Gene Search Help

Show additional filters Save search Advanced

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 oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (114704454..114716894, complement)	RP5-1000E10.2, ALPSA, CMNS, N-ras, NCMS1, NS6, NRAS
ID: 4893			
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, K-RAS2
ID: 3645			

Find related data Database: Select Find items

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

Recent activity Turn Off Clear

12

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

NCBI Resources How To Sign in to NCBI

Gene Gene Advanced Search Help

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:6407
 See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
 Gene type protein coding
 RefSeq status REVIEWED
 Organism Homo sapiens
 Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrini; Hominidae; Homo
 Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-
 13

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)

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

NCBI Resources How To Sign in to NCBI

Gene Gene Advanced Search Help

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:6407
 See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
 Gene type protein coding
 RefSeq status REVIEWED
 Organism Homo sapiens
 Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrini; Hominidae; Homo
 Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-
 14

Table of contents

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

- 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)

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

NCBI Resources How To Sign in to NCBI

Gene Gene Advanced Search Help

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

Genomic context

Location: 12p12.1 Exon count: 6

Annotation release: 106 Status: current Assembly: GRCh38 (GCF_000001405_26) Chr: 12 Location: NC_000012.12 (2505246..25250923, complement)

Annotation release: 105 Status: previous assembly Assembly: GRCh37.p13 (GCF_000001405_25) Chr: 12 Location: NC_000012.11 (2536180..25403870, complement)

Genomic regions, transcripts, and products

Genomic Sequence: NC_000012.12 chromosome 12 reference GRCh38 Primary Assembly Go to reference sequence details Go to nucleotide: Graphics Fasta GenBank Nucleotide 15

Chromosome 12 - NC_000012.12

LRMP LYRMP5 LOC10421617 RPL36P27

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)

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

NCBI Resources How To Sign in to NCBI

Gene Gene Advanced Search Help

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:6407
 See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
 Gene type protein coding
 RefSeq status REVIEWED
 Organism Homo sapiens
 Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrini; Hominidae; Homo
 Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-
 16

Table of contents

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

- 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)

The screenshot shows the NCBI Gene Ontology page. At the top, there is a red box around the header "Gene Ontology Provided by GOA". Below this, there are two tables. The first table is for "Function" and lists categories like GDP binding, GMP binding, GTP binding, LRR domain binding, protein binding, and protein complex binding, each with evidence codes (IEA, IPI, IDA) and PubMed links. The second table is for "Process" and lists categories like Fc-epsilon receptor signaling pathway, GTP catabolic process, MAPK cascade, Ras protein signal transduction, actin cytoskeleton organization, activation of MAPKK activity, axon guidance, and blood coagulation, also with evidence codes (TAS, IEA) and PubMed links. A red arrow points downwards from the bottom of the "Function" table towards the UniProt-GOA page.

GO: Gene Ontology

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

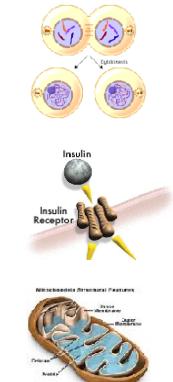
The screenshot shows the UniProt-GOA database homepage. At the top, it features the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. The main title is "UniProt-GOA" with a subtitle "Gene Ontology Annotation (UniProt-GOA) Database". Below the title, there is a brief description of the UniProt GO annotation program's goal: to provide high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB). It mentions the assignment of GO terms to UniProt records as an integral part of UniProt biocuration. The UniProt manual and electronic GO annotations are supplemented with manual annotations supplied by external collaborating GO Consortium groups. A note states that UniProt is a member of the GO Consortium. On the right side, there is a "Menu" section with links for Downloads, Searching UniProt-GOA, Annotation Methods, Annotation Tutorial, Manual Annotation Efforts, Reference Genome Annotation Initiative, Cardiovascular Gene Ontology Annotation Initiative, Renal Gene Ontology Annotation Initiative, and 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

P01116 - RASK_HUMAN

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

Display: None

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 ¹	10 - 18	9	GTP	2 Publications		
Nucleotide binding ¹	29 - 35	7	GTP	2 Publications		
Nucleotide binding ¹	59 - 60	2	GTP	2 Publications		

UniProt will detail much more information for protein coding genes

P01116 - RASK_HUMAN

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

Display: None

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 ¹	10 - 18	9	GTP	2 Publications		
Nucleotide binding ¹	29 - 35	7	GTP	2 Publications		
Nucleotide binding ¹	59 - 60	2	GTP	2 Publications		

View FASTA file format

UniProt will detail much more information for protein coding genes

P01116 - RASK_HUMAN
Protein: GTPase KRas
Gene: KRAS
Organism: Homo sapiens (Human)
Status: Reviewed - Experimental evidence at protein level

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	Identifier	Actions
Nucleotide binding ¹	10 - 18	9	GTP 2 Publications	Graphical view	VAR_034601	
Nucleotide binding ¹	29 - 35	7	GTP 2 Publications	Graphical view	VAR_034601	
Nucleotide binding ¹	59 - 60	2	GTP 2 Publications	Graphical view	VAR_034601	

P01116 - RASK_HUMAN
Protein: GTPase KRas
Gene: KRAS
Organism: Homo sapiens (Human)
Status: Reviewed - Experimental evidence at protein level

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	Identifier	Actions
Nucleotide binding ¹	10 - 18	9	GTP 2 Publications	Graphical view	VAR_034601	
Nucleotide binding ¹	29 - 35	7	GTP 2 Publications	Graphical view	VAR_034601	
Nucleotide binding ¹	59 - 60	2	GTP 2 Publications	Graphical view	VAR_034601	

P01116 - RASK_HUMAN
Protein: GTPase KRas
Gene: KRAS
Organism: Homo sapiens (Human)
Status: Reviewed - Experimental evidence at protein level

Pathology & Biotech:

Involvement in disease:

- [MIM:601626]: A subtype of acute leukemia, a cancer of the white blood cells. AML is a malignant disease of bone marrow characterized by maturation arrest of hematopoietic precursors at an early stage of development. Clonal expansion of myeloid blasts occurs in bone marrow, blood, and other tissue. Myelogenous leukemias develop from cells that normally produce neutrophils, basophils, eosinophils and monocytes. 1 Publication
- Note: The disease is caused by mutations affecting the gene represented in this entry.

Regions:

Feature key	Position(s)	Length	Description	Graphical view	Identifier	Actions
Natural variant ¹	10 - 18	9	1 G → GG 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	Graphical view	VAR_034601	

Leukemia, Acute Myelogenous (AML)
[MIM:601626]: A subtype of acute leukemia, a cancer of the white blood cells. AML is a malignant disease of bone marrow characterized by maturation arrest of hematopoietic precursors at an early stage of development. Clonal expansion of myeloid blasts occurs in bone marrow, blood, and other tissue. Myelogenous leukemias develop from cells that normally produce neutrophils, basophils, eosinophils and monocytes. 1 Publication

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

Noonan Syndrome 3 (NS3)
[MIM:609942]: A form of Noonan syndrome, a disease characterized by short stature, facial dysmorphic features such as hypertelorism, a downward eyelid and low-set posteriorly rotated ears, and a high incidence of congenital heart

P01116 - RASK_HUMAN
Protein: GTPase KRas
Gene: KRAS
Organism: Homo sapiens (Human)
Status: Reviewed - Experimental evidence at protein level

Structure:

Secondary structure: Legend: Helix Turn Beta strand

3D structure databases:

Select the link destination:	Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
<input checked="" type="checkbox"/> PDB	1D8D	X-ray	2.00	P	178-188	[x]
<input checked="" type="checkbox"/> RCSB PDB ²	1DR	X-ray	3.00	P	178-188	[x]
<input type="checkbox"/> PDB	1K2O	X-ray	2.20	C	169-173	[x]
<input type="checkbox"/> PDB	1K2P	X-ray	2.10	C	169-173	[x]
<input type="checkbox"/> PDB	3GFT	X-ray	2.27	A/B/C/D/E/F	1-164	[x]
<input type="checkbox"/> PDB	4DSN	X-ray	2.03	A	2-164	[x]
<input type="checkbox"/> PDB	4DSQ	X-ray	1.85	A	2-164	[x]
<input type="checkbox"/> PDB	4EPR	X-ray	2.00	A	1-164	[x]
<input type="checkbox"/> PDB	4EPT	X-ray	2.00	A	1-164	[x]
<input type="checkbox"/> PDB	4EPV	X-ray	1.35	A	1-164	[x]
<input type="checkbox"/> PDB	4EPW	X-ray	1.70	A	1-164	[x]
<input type="checkbox"/> PDB	4EPX	X-ray	1.76	A	1-164	[x]
<input type="checkbox"/> PDB	4EPY	X-ray	1.80	A	1-164	[x]
<input type="checkbox"/> PDB	4L8G	X-ray	1.52	A	1-164	[x]
<input type="checkbox"/> PDB	4LDJ	X-ray	1.15	A	1-164	[x]
<input type="checkbox"/> PDB	4LPK	X-ray	1.50	A/B	1-160	[x]

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?

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.R., Phan, J., Burns, M.C., Olejniczak, E.T., Watsonson, A.G., Lee, T., Rosanesse, O.W., Fealk, 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

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]121-O-[GLY]121-C

Assembly: Biocomplex 1 Model: Model 1 Symmetry: None
Interaction: IQDP201A Style: Carbon Color: Rainbow Ligand: None Quality: Automatic
Water Ions Hydrogens Clashes
Viewers Options

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

FAMILY & DOMAINS

- Phylogenetic tree
- Pfam
- Treponem
- TFB

PFAM is one of the best protein family databases

PFAM is one of the best protein family databases

Sequences (2)
Sequence status: Complete.
Sequence processing: The displayed sequence is further processed into a mature form.
This entry describes 2 isoforms produced by alternative splicing. [Align](#)

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

Family: Ras (PF00071)

Summary
Domain organization
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures
Jump to... [InterPro](#) [Go](#)

Summary: Ras family
Other inclusion annotations and additional family information from a range of different sources. These sources can be selected via the tabs below.

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

This is the Wikipedia entry entitled "Ras subfamily". More...

Ras is the name given to a family of related proteins which is ubiquitously expressed in all cell membranes and organelles. All Ras proteins 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 archetypal member of the Ras superfamily of proteins, which are all related in 1D structure and regulate diverse cell behaviors.

The name "Ras" is an abbreviation of "Rat Sarcoma", reflecting the way the first members of the protein family were discovered. The term "Ras" is also used in the field of genomics to denote a class of genes encoding these proteins.

When Ras is activated it by interacting with GTP. 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 uncontrolled and overactive signaling inside the cell, even in the absence of incoming signals.

Because these signals result in cell growth and division, overactive Ras signaling can ultimately lead to cancer.^[1] The 3 Ras genes in human (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 tumors and up to 90% in certain types of cancer (e.g., pancreatic cancer).^[2] For this reason, Ras inhibitors are being studied as a treatment for cancer, and other diseases with Ras overexpression.

Contents

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

Identifiers

Symbol	Ras
Name	PF00071; R
InterPro	IPR027531; R
PROSITE	PS0000017; R
SCOP	sc2149
SUPERFAMILY	spf0071

species distribution, number in humans and residue-wise conservation, etc... ?

Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures
Jump to... ↻
enter ID/acc ↗

Species distribution

Sunburst Tree

This visualization provides a graphical representation of the distribution of the human genome species. You can find the original interactive tree in the adjacent tab [Mammals](#).

Sunburst controls

Home genome

Root

- non-coding
- exons
- introns
- other exons
- other introns
- unclassified sequences

Weight segments by...

- number of segments
- number of species

Change the size of the sunburst

Small Large

Color assignments

orange	blue/yellow
green	other exons
magenta	other introns
pink	unclassified
yellow	unclassified sequences

Selections

Align selected sequences to HMM

Current selection: **non-coding**

Currently selected:

- 501 requirements
- 2,487 genes

Note: Some requirements may show results in pop-up windows. Please disable pop-up blockers.

species distribution, number in humans and residue-wise conservation, etc... ?

EMBL-EBI

Alignment for selected sequences

Currently showing rows 1 to 30 of 358 rows in this alignment. Show all rows of alignment

Sequence ID	Accession	Organism	Length	Start	End	Score	Align.
PF00071_1	PF00071	C. elegans	100	1	100	100	100.000
PF00071_2	PF00071	Yersinia enterocolitica	100	1	100	100	100.000
PF00071_3	PF00071	Escherichia coli	100	1	100	100	100.000
PF00071_4	PF00071	Streptomyces coelicolor	100	1	100	100	100.000
PF00071_5	PF00071	Neurospora crassa	100	1	100	100	100.000
PF00071_6	PF00071	Arabidopsis thaliana	100	1	100	100	100.000
PF00071_7	PF00071	Artemia franciscana	100	1	100	100	100.000
PF00071_8	PF00071	Aspergillus nidulans	100	1	100	100	100.000
PF00071_9	PF00071	Aspergillus oryzae	100	1	100	100	100.000
PF00071_10	PF00071	Aspergillus fumigatus	100	1	100	100	100.000
PF00071_11	PF00071	Aspergillus terreus	100	1	100	100	100.000
PF00071_12	PF00071	Aspergillus niger	100	1	100	100	100.000
PF00071_13	PF00071	Aspergillus flavus	100	1	100	100	100.000
PF00071_14	PF00071	Aspergillus terreus	100	1	100	100	100.000
PF00071_15	PF00071	Aspergillus oryzae	100	1	100	100	100.000
PF00071_16	PF00071	Aspergillus fumigatus	100	1	100	100	100.000
PF00071_17	PF00071	Aspergillus flavus	100	1	100	100	100.000
PF00071_18	PF00071	Aspergillus niger	100	1	100	100	100.000
PF00071_19	PF00071	Aspergillus terreus	100	1	100	100	100.000
PF00071_20	PF00071	Aspergillus oryzae	100	1	100	100	100.000
PF00071_21	PF00071	Aspergillus flavus	100	1	100	100	100.000
PF00071_22	PF00071	Aspergillus niger	100	1	100	100	100.000
PF00071_23	PF00071	Aspergillus terreus	100	1	100	100	100.000
PF00071_24	PF00071	Aspergillus oryzae	100	1	100	100	100.000
PF00071_25	PF00071	Aspergillus flavus	100	1	100	100	100.000
PF00071_26	PF00071	Aspergillus niger	100	1	100	100	100.000
PF00071_27	PF00071	Aspergillus terreus	100	1	100	100	100.000
PF00071_28	PF00071	Aspergillus oryzae	100	1	100	100	100.000
PF00071_29	PF00071	Aspergillus flavus	100	1	100	100	100.000
PF00071_30	PF00071	Aspergillus niger	100	1	100	100	100.000

Show all 358 rows | Show page 1 | 1 2 3 4 5 6 7 8 9 10 11 | Show page 1

can find the sunburst controls Help

Home **options**

Root

- + **viruses**
- + **Mycobacterium**
 - + **Yersinia**
 - + **Escherichia**
 - + **Salmonella**
 - + **Shigella**
 - + **Proteus**
 - + **Enterobacter**
 - + **Pseudomonas**
 - + **Acinetobacter**
 - + **Neisseria**
 - + **Haemophilus**
 - + **Leptospira**

Weight segments by ...

- number of sequences
- number of species

Change the size of the sunburst Large

Colour assignments

- Human
- Bacteriophage
- Bacteria
- Virus
- Unknown
- Unassigned sequence

Selections

Align selected sequences to HMM

Open in current browser window

Current selection:

- 1 selected sequences
- 1 species

Note: long lists may result in pop-up windows. Please disable pop-up blockers.

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

○ ○ ○

Pfam: Family: Kinesin (PF00225)

http://pfam.janelia.org/family/kinesin#tabview=tab8 RSS Google

HMM
janelia farm research campus

Pfam
keyword search Go

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Family: Kinesin (PF00225)

>Loading page components (1 remaining)...

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

Interactions

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

Tubulin	Tubulin_C	Kinesin	Tubulin	Kinesin
Tubulin	Tubulin_C	Kinesin	Tubulin	Kinesin

126 architectures 4150 sequences 6 Interactions 248 species 114 structures

HMM logo

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

Contribution

Jump to... ↻

InterPro ID: [GO](#)

EMBL-EBI HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Family: Ras (PF00071)

332 architectures 21243 sequences 36 interactions 1096 species 563 structures

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
KI13B_HUMAN	11 - 352	1f9w	A	392 - 723	Jmol AstexViewer SPICE
			B	392 - 723	Jmol AstexViewer SPICE
		3kar	A	392 - 723	Jmol AstexViewer SPICE
		1g0b	A	11 - 352	Jmol AstexViewer SPICE
		3gbj	B	11 - 352	Jmol AstexViewer SPICE
			C	11 - 352	Jmol AstexViewer SPICE
		1i16	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
		A	24 - 359	Jmol AstexViewer SPICE	

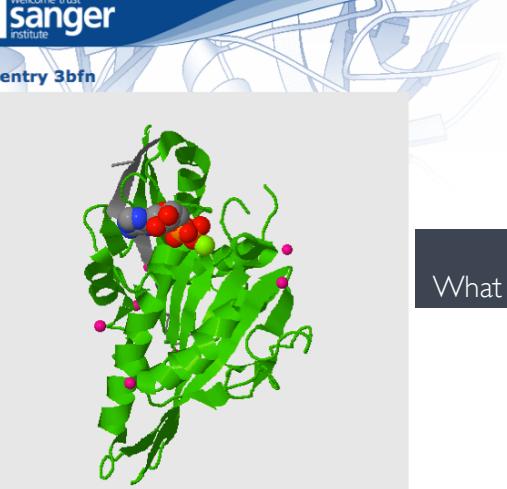
Summary
Domain organisation
Clans
Alignments
HMM logo
Trees
Curation & models
Species
Interactions
Structures
Jump to... ↴
enter ID/acc Go

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"

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.

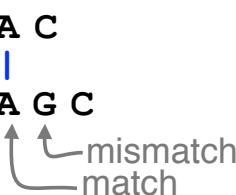
Seq1: C A T T C A C

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.

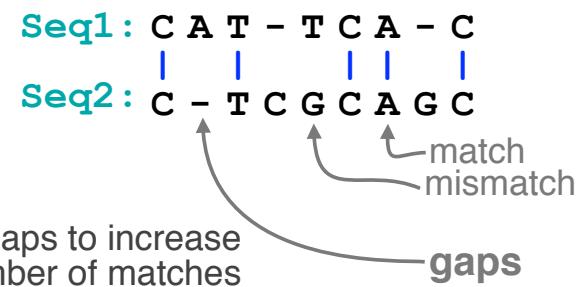
Seq1: C A T T C A C
Seq2: C T C G C A G C



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.

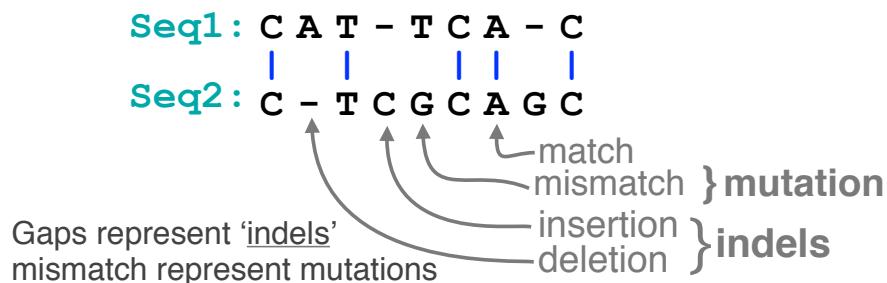
Seq1: C A T - T C A - C
Seq2: C - T C G C A G C



Add gaps to increase number of matches

gaps

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!*

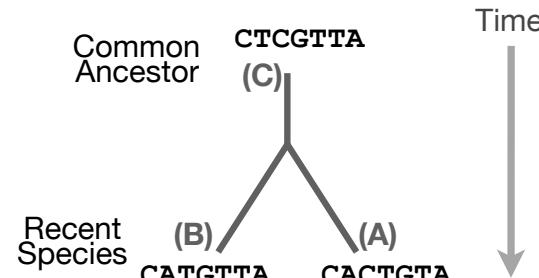
ALIGNMENT FOUNDATIONS

- 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.

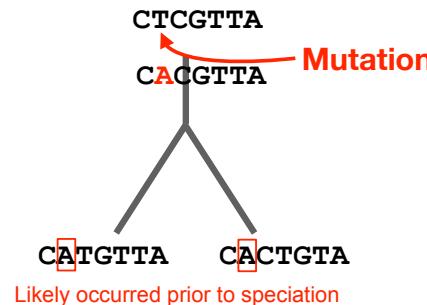
- Mutations/Substitutions
- Deletions
- Insertions



Mutations, deletions and insertions

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

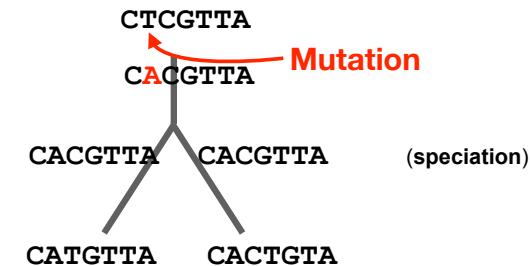
- **Mutations/Substitutions** $\text{CTCGTTA} \rightarrow \text{CACGTTA}$
- Deletions
- Insertions



Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

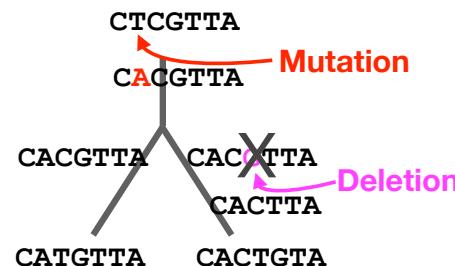


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

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

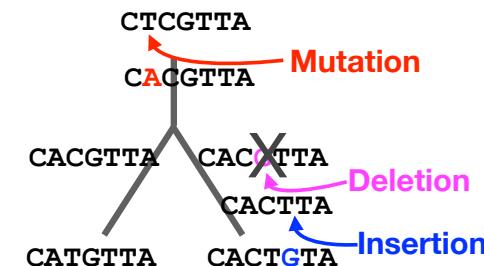


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

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

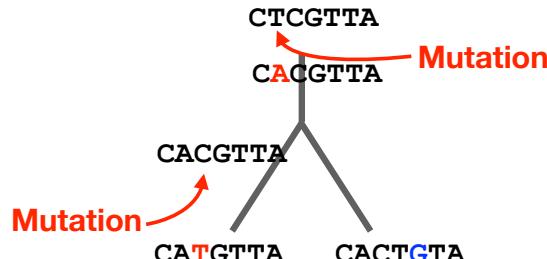


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

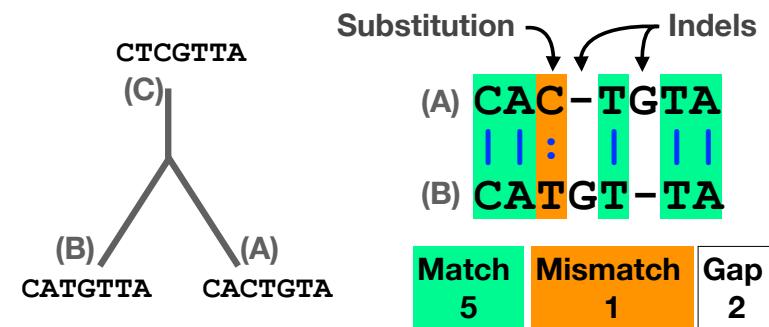
$$\text{CTCGTTA} \rightarrow \text{CACGTTA}$$
$$\text{CACGTTA} \rightarrow \text{CATGTTA}$$



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

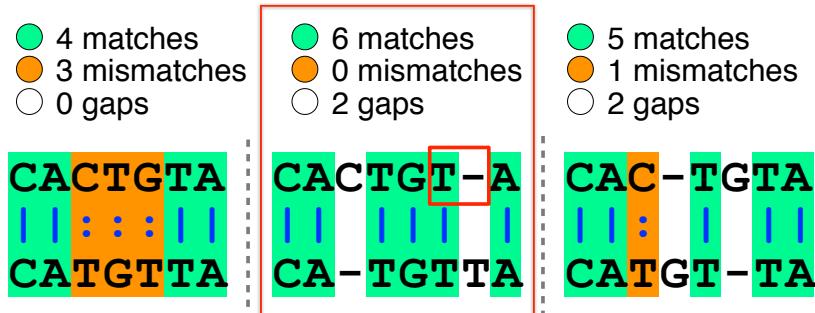
CACTGTA
||:||:
CATGTTA

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

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

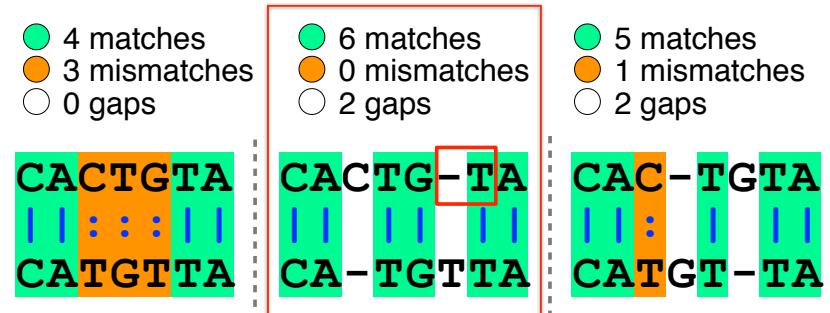
Optimal alignments

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



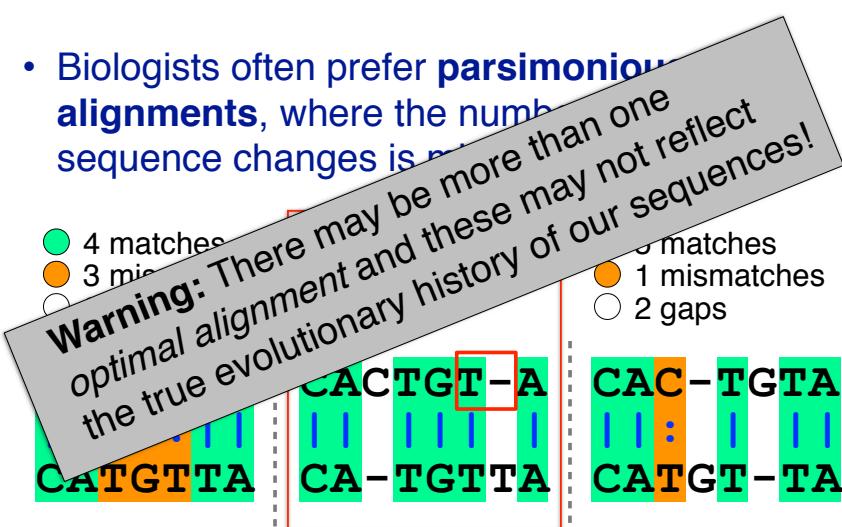
Optimal alignments

- 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

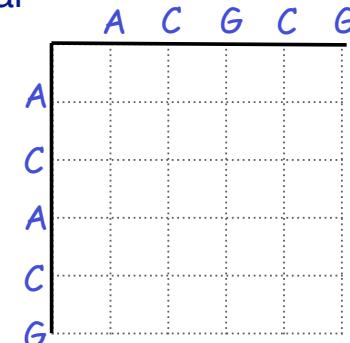
- Why...
 - Why compare biological sequences?
- What...
 - Alignment view of sequence changes during evolution (matches, mismatches and gaps)

• How...

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

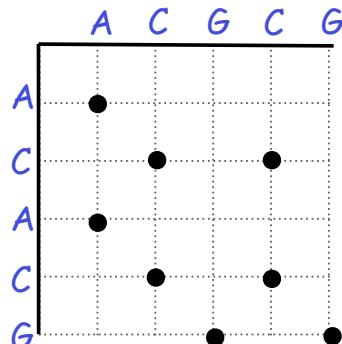
Dot plots: simple graphical approach

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



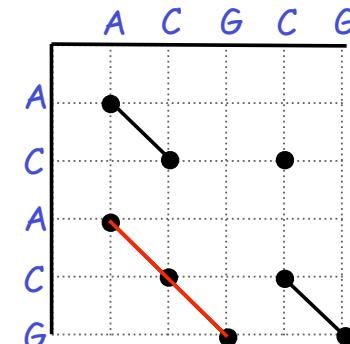
Dot plots: simple graphical approach

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



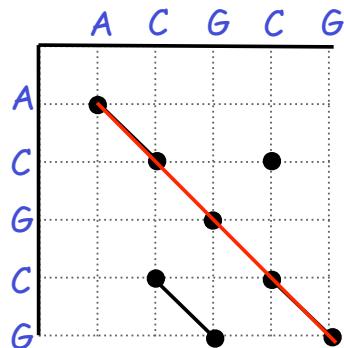
Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



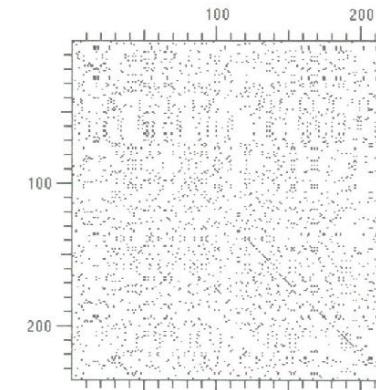
Dot plots: simple graphical approach

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



Dot plots: simple graphical approach

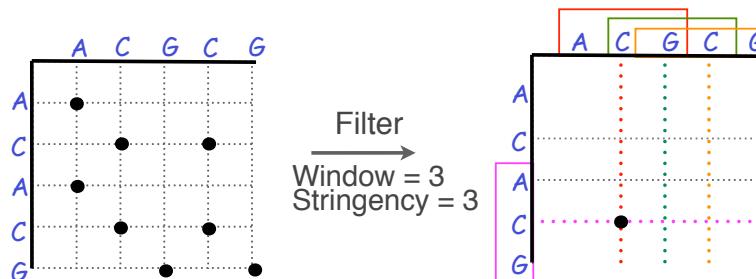
- Dot matrices for long sequences can be noisy



Dot plots: window size and match stringency

Solution: use a window and a threshold

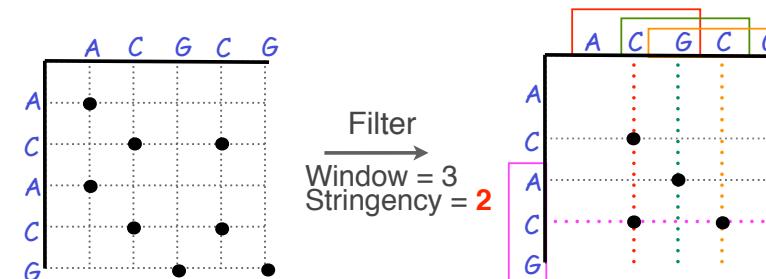
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
 - You have to choose window size and stringency



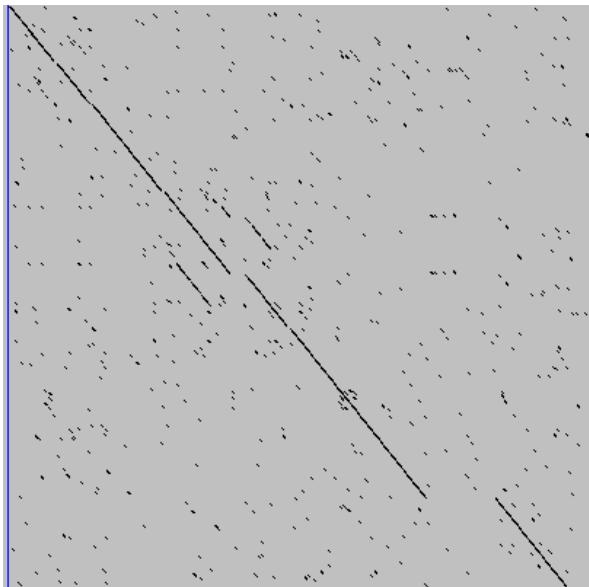
Dot plots: window size and match stringency

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



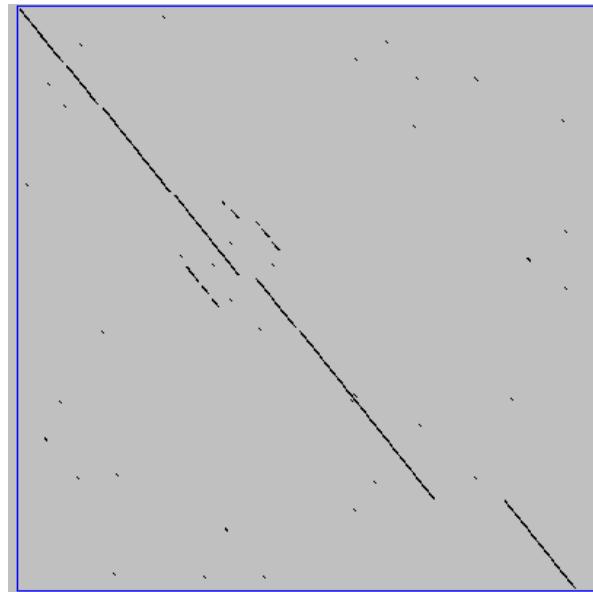
Window size = 5 bases



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

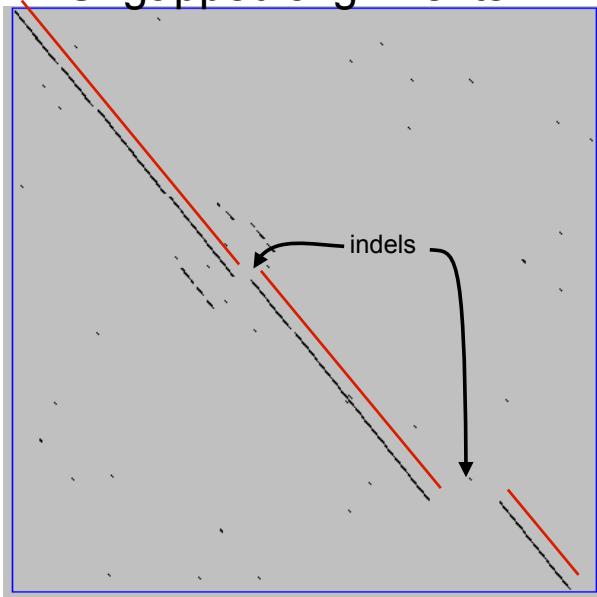


This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

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.

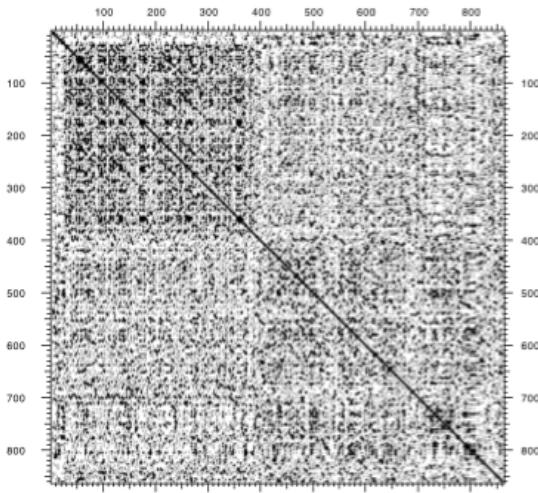
Downward or rightward paths represent **insertion** or **deletions** (gaps in one sequence or the other).

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

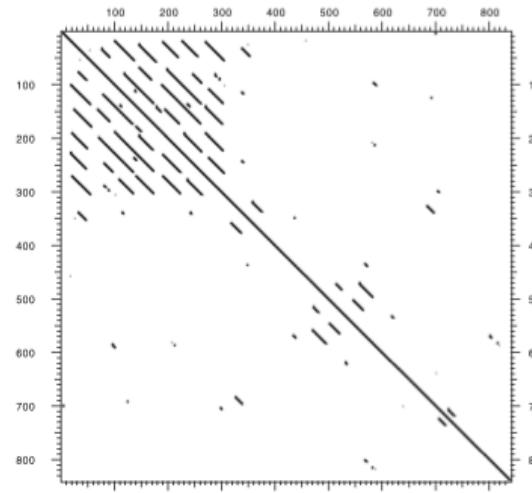


Human LDL receptor
protein sequence
(Genbank P01130)

$$W = 1 \\ S = 1$$

(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/>

The screenshot shows a web-based application 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" sliders for "Window Size" (set to 5), "Moving window step size" (set to 2), and "Match stringency" (set to 8). To the right, there are two dot plots: "Protein Dot Plot" (wslice = 5, wstep = 2, nmismatch = 2) and "DNA Dot Plot" (wslice = 3, wstep = 3, nmismatch = 2). Both plots show a diagonal line of dots with some scattered points. At the bottom, there is a URL "https://bioboot.shinyapps.io/dotplot2/" and 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: 5

Moving window step size: 2

Match stringency: 8

Protein Dot Plot: wslice = 5, wstep = 2, nmismatch = 2

DNA Dot Plot: wslice = 3, wstep = 3, nmismatch = 2

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

Questions for discussion:

- Why does the DNA sequence have more dots than the protein sequence plot?
- How can we increase the signal to noise ratio?
- What does it mean when one dot has more than 100 red/orange dots next to it?

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**

	D	P	L	E
D	6	-1	-4	2
P	-1	7	-3	-1
M	-3	-2	2	-2
E	-2	-1	-3	5

(1) → (2) → (3) →

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.

81

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

	D	P	L	E
D	6	-1	-4	2
P	-1	7	-3	-1
M	-3	-2	2	-2
E	-2	-1	-3	5

(1) → (2) → (3) →

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

Scoring the alignment matrix

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

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

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

Scoring the alignment matrix

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

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

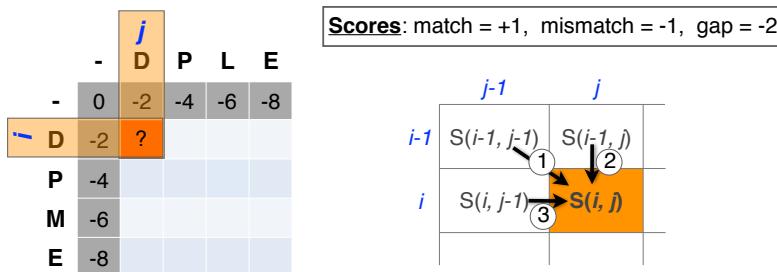
Scores: match = +1, mismatch = -1, gap = -2

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

Seq1: DPME
Seq2: ----

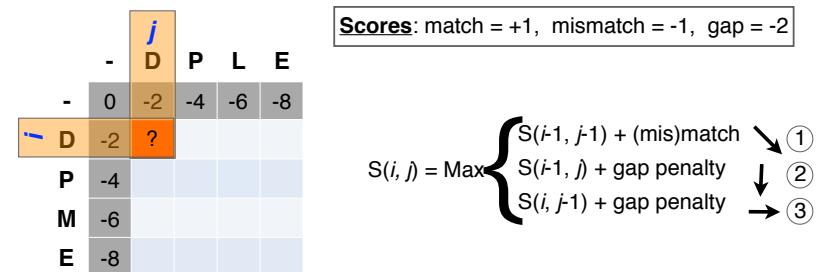
Scoring the alignment matrix

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



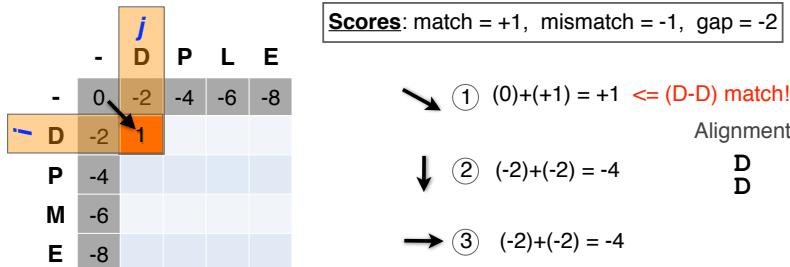
Scoring the alignment matrix

- 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



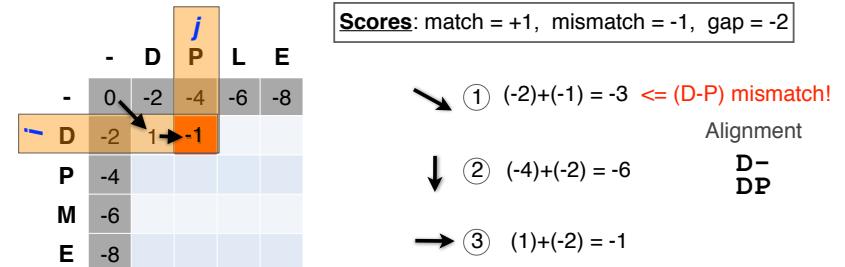
Scoring the alignment matrix

- 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



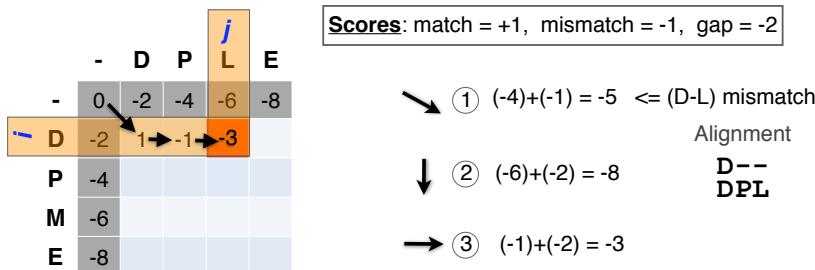
Scoring the alignment matrix

- 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)



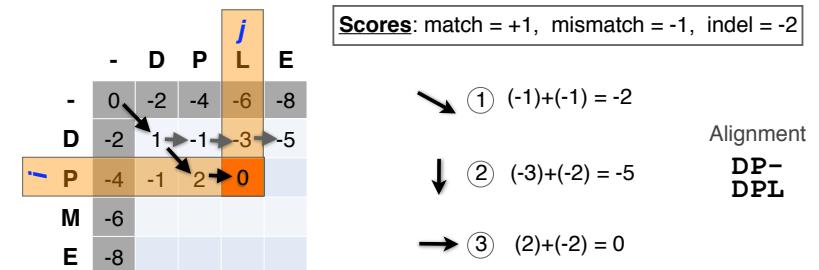
Scoring the alignment matrix

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



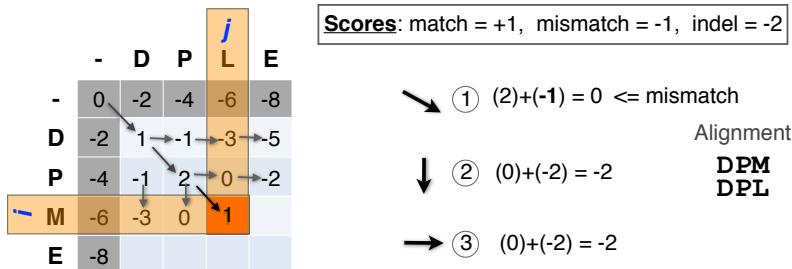
Scoring the alignment matrix

- 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.



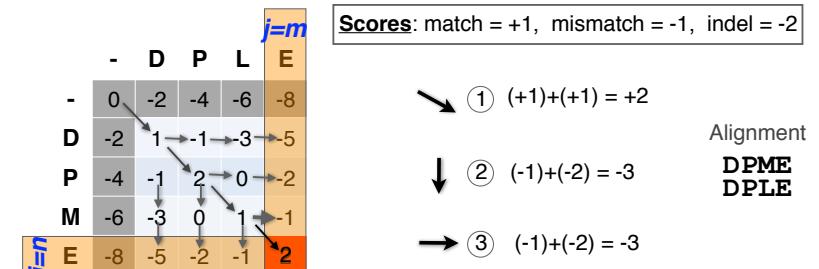
Scoring the alignment matrix

- 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



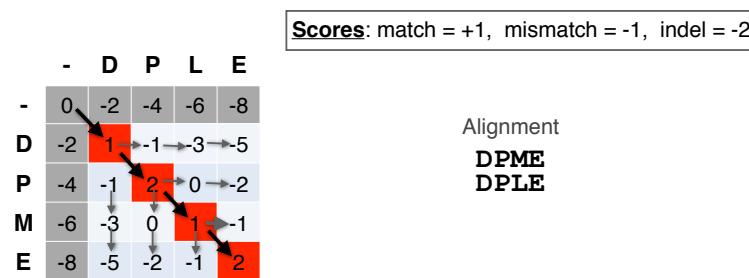
Scoring the alignment matrix

- 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)



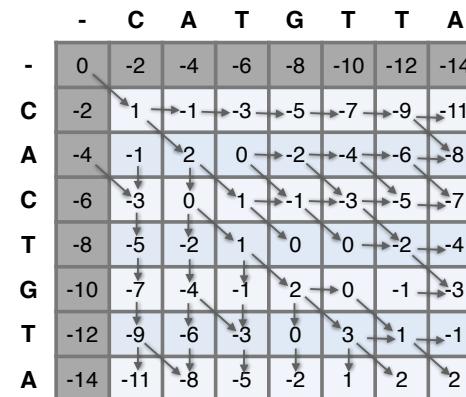
Scoring the alignment matrix

- 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



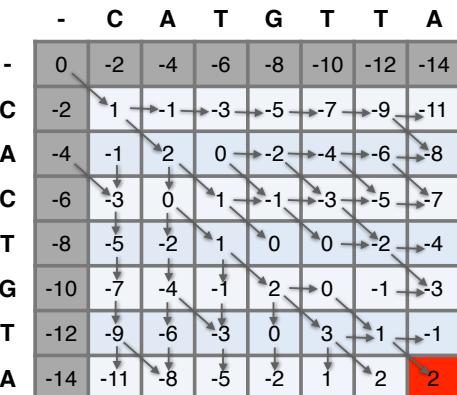
Questions:

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



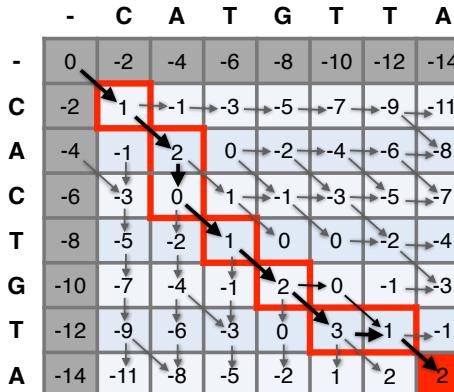
Questions:

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



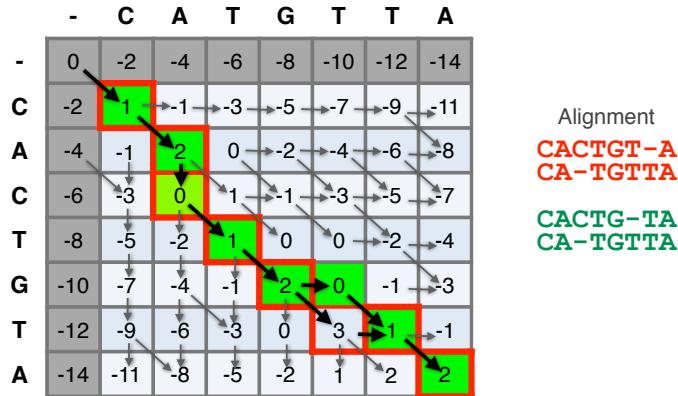
Questions:

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



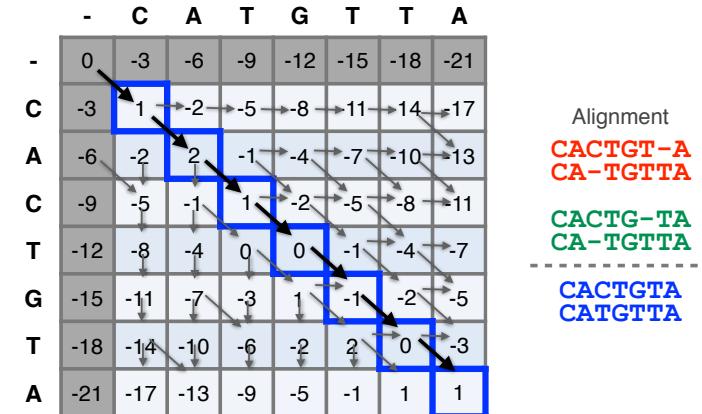
More than one alignment possible

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



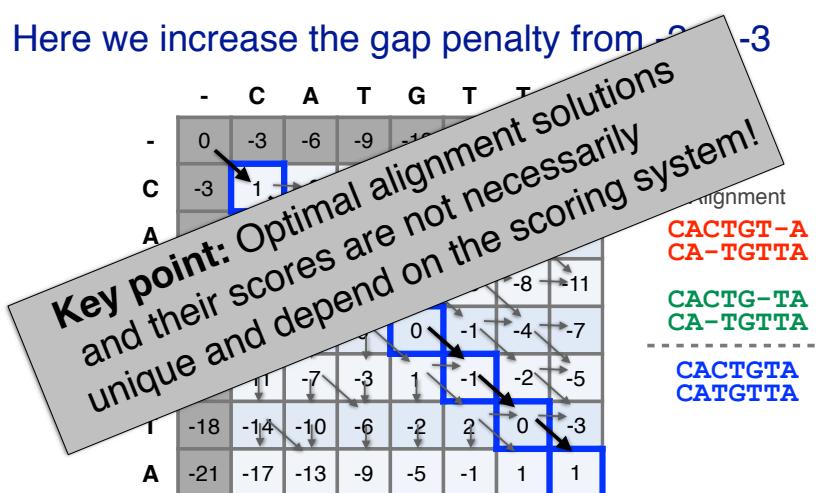
The alignment and score are dependent on the scoring system

- 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



NW DYNAMIC PROGRAMMING

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

	A	G	T	T	C
A	0				
T					
T					
G					
C					

ALIGNMENT FOUNDATIONS

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

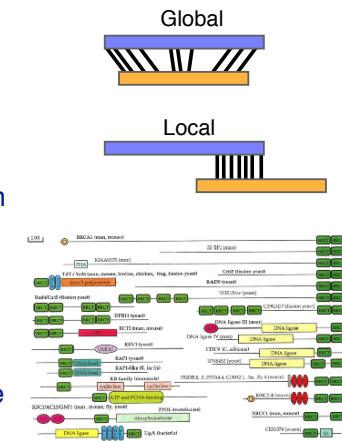
Local alignment: Definition

- 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.

Global vs local alignments

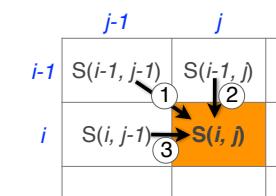
- Needleman-Wunsch is a **global alignment** algorithm
 - Resulting alignment spans the complete sequences end to end
 - This is appropriate for closely related sequences that are similar in length
- For many practical applications we require **local alignments**
 - Local alignments highlight sub-regions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

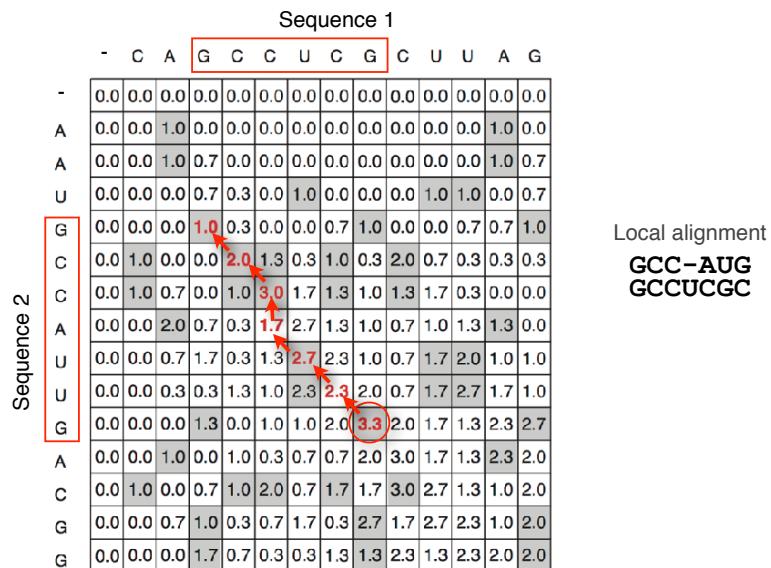


The Smith-Waterman algorithm

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

$$S(i, j) = \max \begin{cases} S(i-1, j-1) + (\text{mis})\text{match} \\ S(i-1, j) - \text{gap penalty} \\ S(i, j-1) - \text{gap penalty} \\ 0 \end{cases}$$

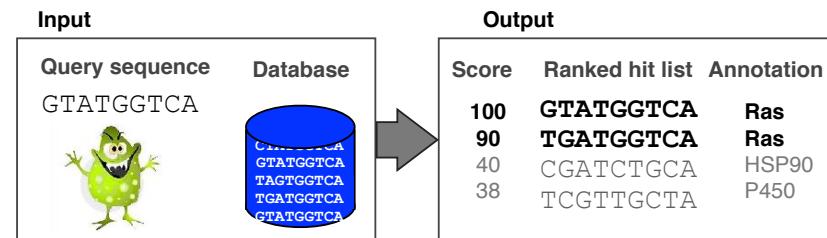




105

Local alignments can be used for database searching

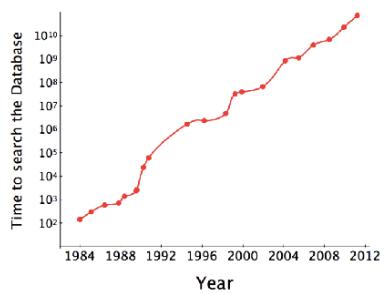
- **Goal:** Given a query sequence (Q) and a sequence database (D), find a list of sequences from D that are most similar to Q
 - **Input:** Q, D and scoring scheme
 - **Output:** Ranked list of hits



106

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!**

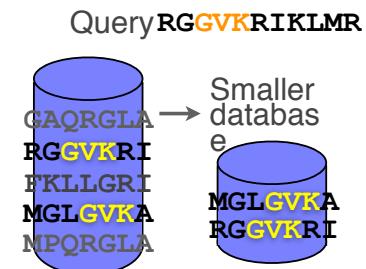


To reduce search time **heuristic algorithms**, such as BLAST, first remove database sequences without a strong local similarity to the query sequence in a quick initial scan.

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

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**
 - BLAST finds regions of local similarity between two sequences
 - BLAST does not examine the entire search space by scanning database sequences for likely matches before performing more rigorous alignments
 - “The central idea of the BLAST algorithm is to confine attention to sequence pairs that contain an initial **word pair match**” Altschul et al. (1990)
 - Sensitivity in exchange for speed
 - In contrast to SW, BLAST is not guaranteed to find optimal alignments

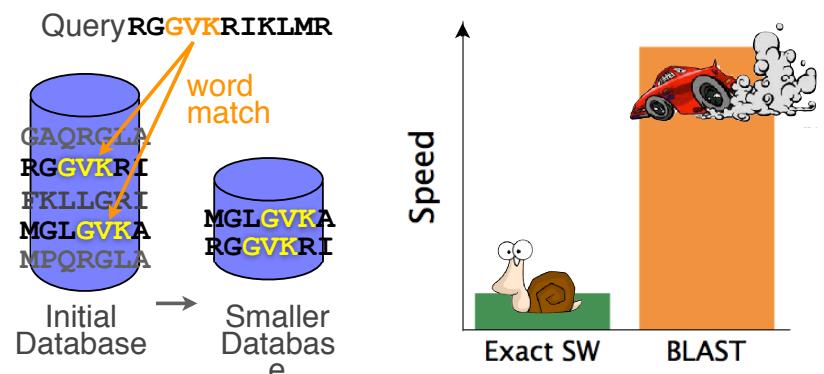
111

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 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

110

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



112

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

113

- 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

114

Blast

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

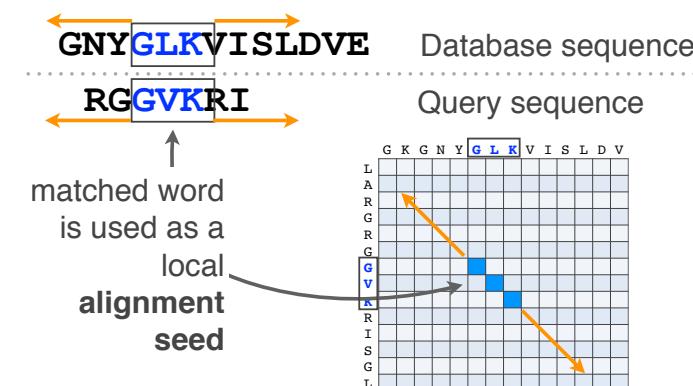
GNYGLKVVISLDVE Database sequence
RGGVKRI Query sequence
RGG RAG RIG RLG ...
GGV GAV GTV GCV ...
GVK GLK GIK GGK ...
VKR VRR VHR VER ...
KRI KKI KHI KDI ...

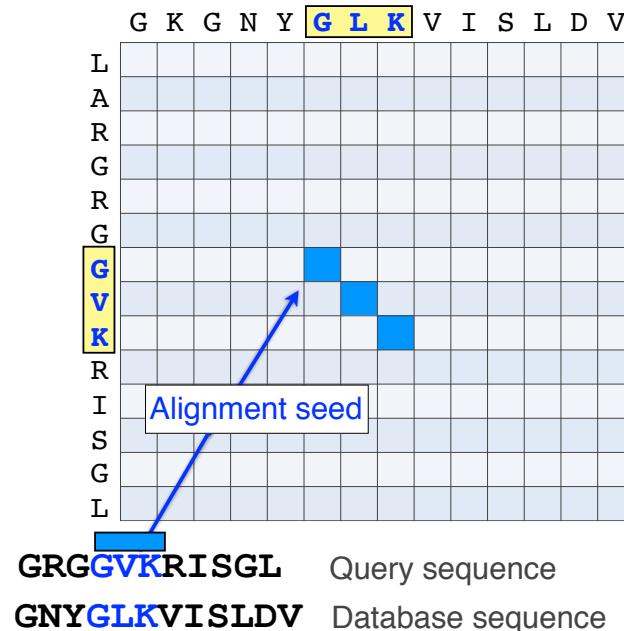
search for
perfect
matches in the
database
sequence

115

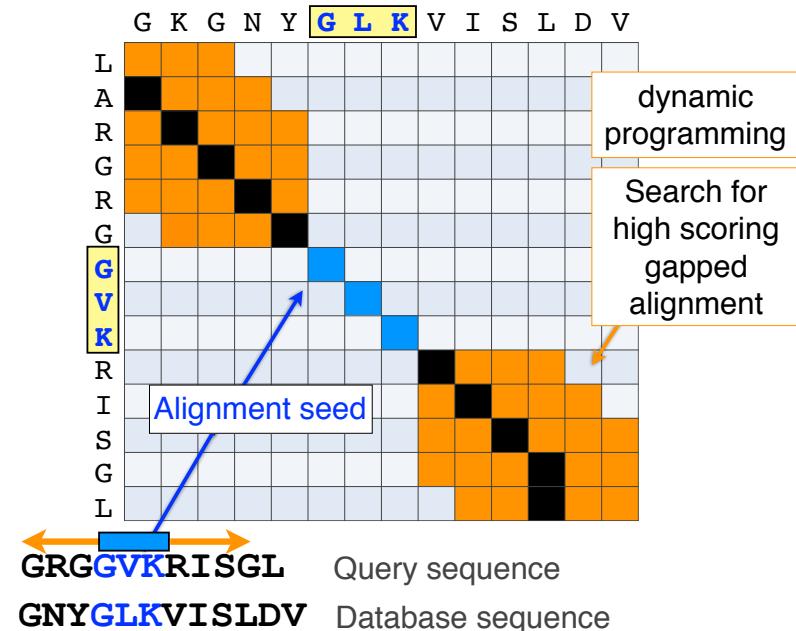
Blast

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

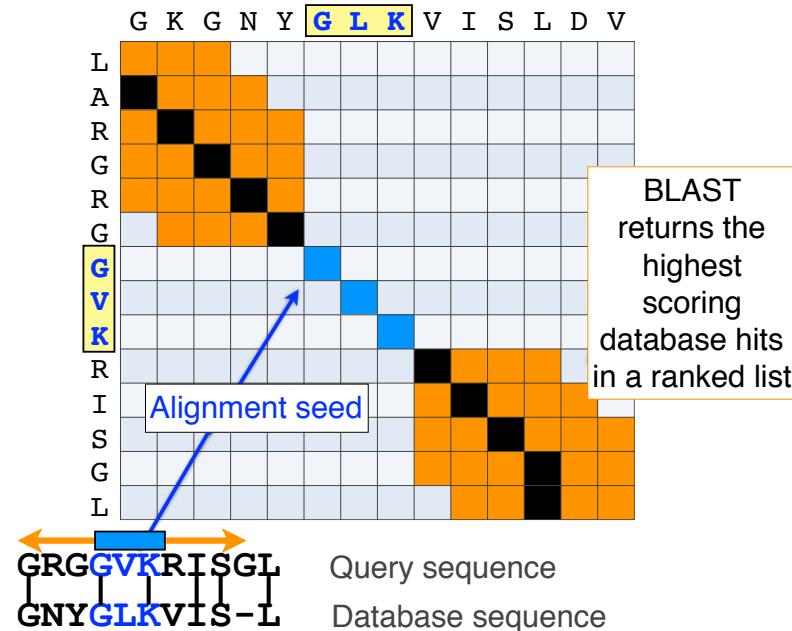




117



118



119

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

120

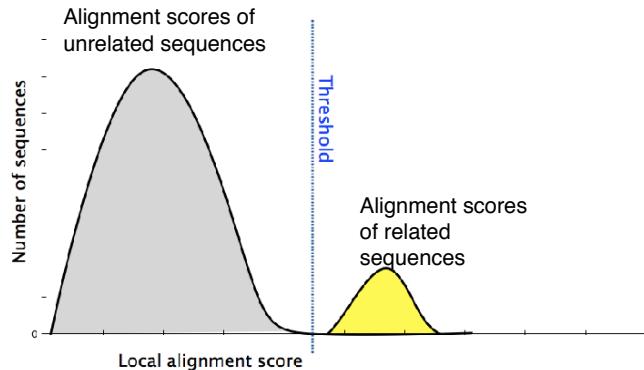
Statistical significance of results

- 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

121

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



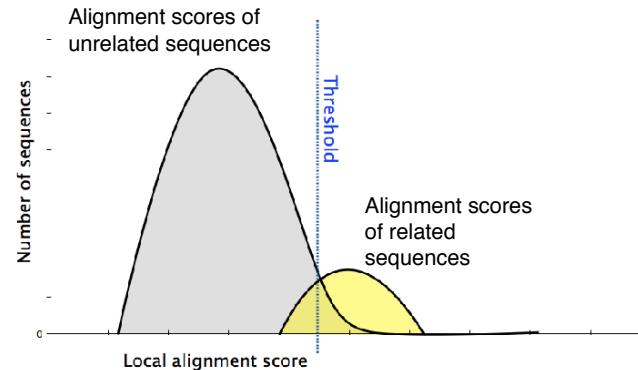
123

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

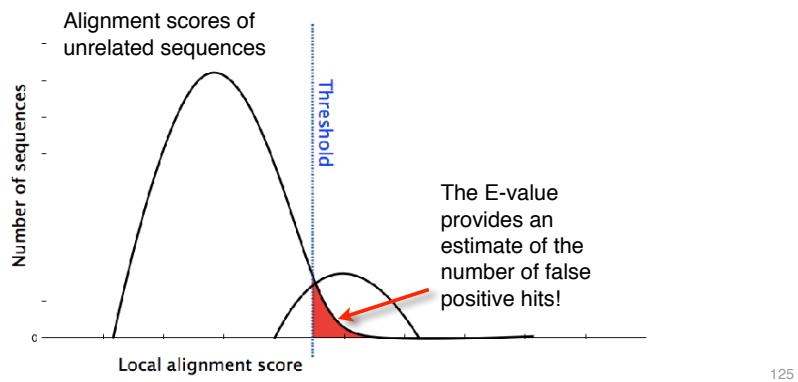
122

- 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

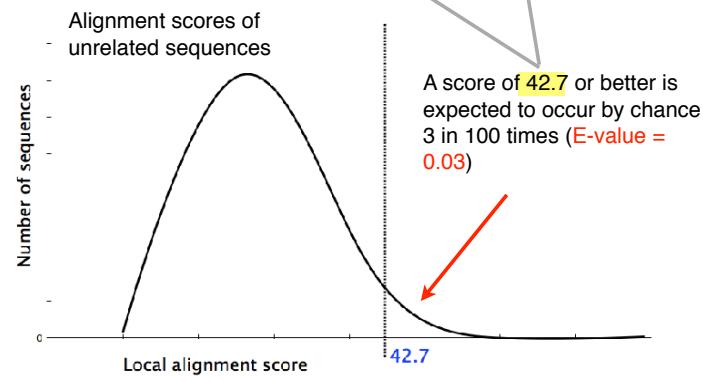


124

- 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



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

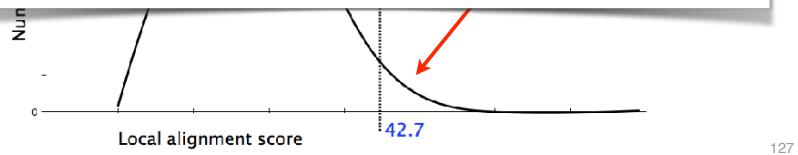


Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo sapiens]	677	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	676	100%	0	98%	AAA20133.1

In general E values < 0.005 are usually significant.

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

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



Your Turn!

Hands-on worksheet Sections 4 & 5

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

Practical database searching with BLAST

The screenshot shows the NCBI BLAST Home Page. At the top, there's a navigation bar with links for Home, Recent Results, Saved Strategies, Help, My NCBI (Sign In/Register), and News. Below the navigation is a search bar with the placeholder "BLAST finds regions of similarity between biological sequences." A red box highlights the "Basic BLAST" section. This section contains several search options: "nucleotide blast" (Search a nucleotide database using a nucleotide query), "protein blast" (Search protein database using a protein query), "tblastx" (Search protein database using a translated nucleotide query), "tblastn" (Search translated nucleotide database using a protein query), and "tblastx" (Search translated nucleotide database using a translated nucleotide query). To the right of these options is a note: "BLAST makes it easy to examine a large group of potential gene candidates." Below the search options is a "Specialized BLAST" section.

129

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”

130

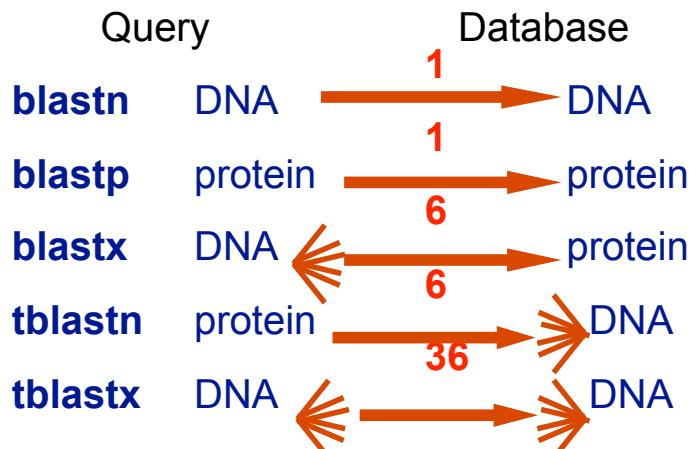
Step 1: Choose your sequence

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

The screenshot shows the NCBI Protein search results for the sequence "hemoglobin subunit beta [Homo sapiens]". The search interface includes a "Display Settings" dropdown set to "FASTA" (circled in red), a "Search" button, and a "Clear" button. Below the search bar, the sequence name "hemoglobin subunit beta [Homo sapiens]" is displayed, along with its NCBI Reference Sequence (NP_000509.1) and GenPept link. The sequence itself is shown in FASTA format: >gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens] MVLHLPEEKSAVTALWGVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPKVKAHGKKVLG AFSDGLAHLNDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCLAHFFGKEFTPPVQAYQKVVAGVAN ALAHKYH. The bottom of the page has links for "Analyze this sequence", "Run BLAST", "Identify Conserved Domains", and "Find in this Sequence".

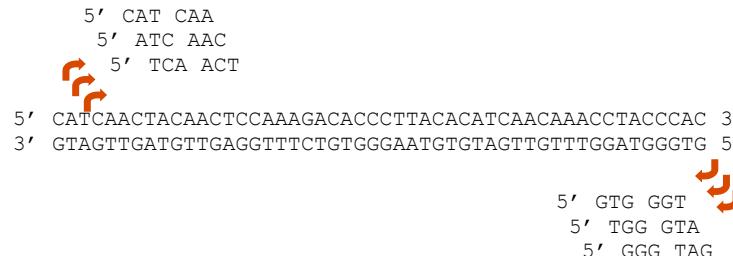
131

Step 2: Choose the BLAST program



132

DNA potentially encodes six proteins



133

Protein BLAST: search protein databases using a protein query
 Enter accession number(s), gi(s), or FASTA sequence(s) >gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
 MVHLTPEEKSAVALWGKVNVDEVGGEALGRLLVYPWTQRFESFCDLSTPDAVMGNPKVKAHGK
 KVLAGFSDGLAIIDNLKGTTATLSELICDKLIVDPMENTRLLGNVLVCVLAIIIFGKEITPPVQAAYQK
 VAGVANALAHKYH
 Or, upload file Choose File no file selected
 Job Title
 Align two or more sequences
 Choose Search Set
 Database Non-redundant protein sequences (nr)
 Organism Optional
 Exclude
 Enter a descriptive title for your BLAST search
 Align two or more sequences
 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

134

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

Human genomic plus transcript (Human G+T)
Genomic plus Transcript
 Human genomic plus transcript (Human G+T)
 Mouse genomic plus transcript (Mouse G+T)
Other Databases
 Nucleotide collection (nr/nnt)
 Reference mRNA sequences (refseq_mrna)
 Reference genomic sequences (refseq_genomic)
 NCBI Genomes (chromosome)
 Expressed sequence tags (est)
 Non-human, non-mouse ESTs (est_others)
 Genomic survey sequences (gss)
 High throughput genomic sequences (HTGS)
 Patent sequences (pat)
 Protein Data Bank (pdb)
 Environmental samples (env_nr)

nucleotide databases

Non-redundant protein sequences (nr)
Non-redundant protein sequences (nr)
 Non-redundant protein sequences (nr)
 Reference proteins (refseq_protein)
 Swissprot protein sequences (swissprot)
 Patented protein sequences (pat)
 Protein Data Bank proteins (pdb)
 Environmental samples (env_nr)

protein databases

135

Protein BLAST: search protein databases using a protein query
 Enter accession number(s), gi(s), or FASTA sequence(s) >gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
 MVHLTPEEKSAVALWGKVNVDEVGGEALGRLLVYPWTQRFESFCDLSTPDAVMGNPKVKAHGK
 KVLAGFSDGLAIIDNLKGTTATLSELICDKLIVDPMENTRLLGNVLVCVLAIIIFGKEITPPVQAAYQK
 VAGVANALAHKYH
 Or, upload file Choose File no file selected
 Job Title
 Align two or more sequences
 Choose Search Set
 Database Non-redundant protein sequences (nr)
Organism Optional
 Exclude
 Enter a descriptive title for your BLAST search
 Align two or more sequences
 Program Selection
 Algorithm blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
 Choose a BLAST algorithm
 BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
 Show results in a new window
 + Algorithm parameters

Organism
 Entrez
 Settings

136

Step 4a: Select optional search parameters

The screenshot shows the 'Algorithm parameters' section of the BLAST search interface. It includes:

- General Parameters:** Max target sequences (100), Short queries (Automatically adjust parameters for short input sequences checked), Expect threshold (10), Word size (3), Max matches in a query range (0).
- Scoring Parameters:** Matrix (BLOSUM62 selected), Gap Costs (Existence: 11 Extension: 1), Compositional adjustments (Conditional compositional score matrix adjustment checked).
- Filters and Masking:** Filter (Low complexity regions unchecked), Mask (Mask for lookup table only unchecked, Mask lower case letters unchecked).
- BLAST:** Search database Non-redundant protein sequences (nr) using Blastp, Show results in a new window checked.

Arrows point to the 'Expect threshold' field (labeled 'Expect'), the 'Word size' field (labeled 'Word size'), and the 'Matrix' dropdown (labeled 'Scoring matrix').

Step 4: Optional parameters

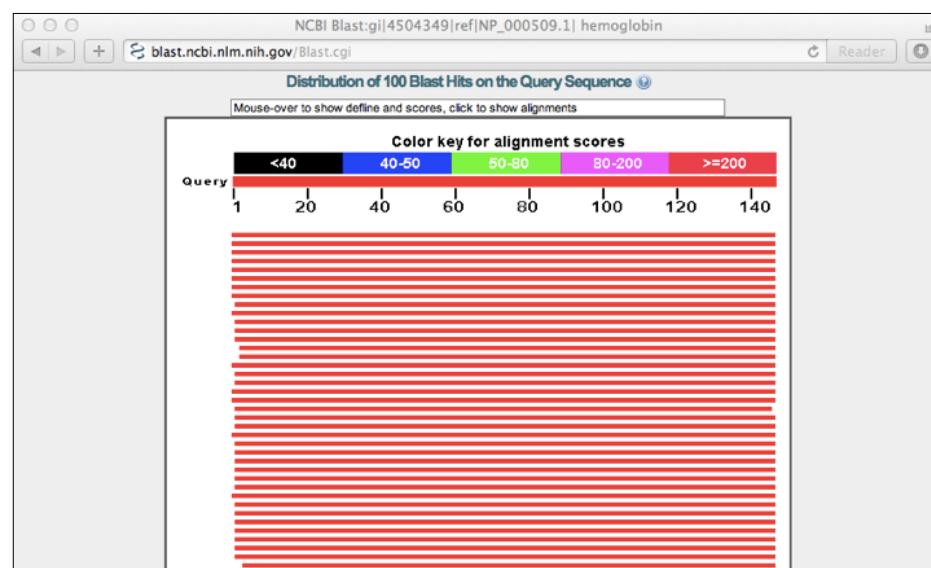
- You can...
 - choose the organism to search
 - change the substitution matrix
 - change the expect (E) value
 - change the word size
 - change the output format

Results page

The screenshot shows the NCBI BLAST results page for the query sequence gi|4504349|ref|NP_000509.1| hemoglobin. Key information displayed includes:

- Query ID:** gi|4504349|ref|NP_000509.1| hemoglobin
- Description:** Query Description: gi|4504349|ref|NP_000509.1| hemoglobin subunit beta (Homo sapiens); Molecule type: amino acid; Query Length: 147.
- Database Name:** nr (All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects)
- Program:** BLASTP 2.2.27+
- Graphic Summary:** Shows putative conserved domains detected along the query sequence, including a 'globin' domain and a 'globin_like superfamily' domain.

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%	AAZ37051.1
hemoglobin beta [synthetic construct]	301	301	100%	1e-102	100%	AAZ29557.1
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	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 AAZ39782.1	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 A	298	298	99%	9e-102	100%	1COH_B
hemoglobin beta subunit variant [Homo sapiens] >gb AAE88054.1 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 Spectroscop	297	297	99%	3e-101	99%	1HDB_B

Further down the results page...

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

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

Download GenPept Graphics ▾ Next ▲ Previous ▲ Descriptions

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

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

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

Query 1 MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60
Sbjct 1 MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60

Query 61 VKAHGGKVLGAFSDGLAHLNLKGTFATLSELHCDKLHVDPENFRLLGNVILVCVLAHHIFG 120
Sbjct 61 VKAHGGKVLGAFSDGLAHLNLKGTFATLSELHCDKLHVDPENFRLLGNVILVCVLAHHIFG 120

Query 121 KEFTPPVQAAAYKVAVANALAHKYH 147
Sbjct 121 KEFTPPVQAAYKVAVANALAHKYH 147

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

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

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

Different output formats are available

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

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

BLAST® Basic Local Alignment Search Tool My NCBI

[Sign In] [Register]

Home Recent Results Saved Strategies Help

► NCBI/ BLAST/ blastp suite/ Formatting Results - FVGUTMP2013

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

Formatting options Reform

Show Alignment as HTML Old View Reset form to defaults

Alignment View Query-anchored with letters for identities

Display Graphical Overview Sequence Retrieval NCBI-gi

Masking Character: Lower Case Color: Grey

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

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

Entrez query:

Expect Min: Expect Max:

Percent Identity Min: Percent Identity Max:

Format for PSI-BLAST with inclusion threshold:

gi|4504349|ref|NP_000509.1| hemoglobin

E.g. Query anchored alignments

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

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

Query	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAZ37051	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAZ29557	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
NP_000509	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
P02024	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAN84548	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAZ39780	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
ACU56984	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAD19696	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
1COH_B	1	VHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
AAF00489	1	VHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
2YRS_B	1	VHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1DXU_B	1	MHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1HDB_B	1	MHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1DKY_B	2	HLTPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
3KMF_C	2	HLTPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
AAL68978	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
1NQD_B	1	VHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1KIK_B	1	VHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
AAN11320	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
XP_02822173	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
1YB5_B	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1YE0_B	1	MHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1O10_B	1	MHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
CAA23759	1	MVHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
1YE2_B	1	MHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1YSF_B	1	MHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1AO0_B	1	MHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1HBS_B	1	MHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1ABY_B	1	MHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59
1CMC_B	1	MHLTPPEEKSAVTALNGKVNNDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	59

... and alignments with dots for identities

The screenshot shows a BLAST search results page from NCBI. The query sequence is hemoglobin (NP_000509.1). The results table includes columns for Query, ID, Sequence, and Score. The sequence alignment is shown below the table, where identical bases are represented by dots and different bases by letters. The alignment length is 60.

Query	ID	Sequence	Score
AAK37051	1	MVHLTPPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK	60
AAK29557	1	60
NP_000509	1	60
P02024	1	60
AAH84548	1	60
AAZ39780	1X.....	60
ACU56984	1	60
AAD19696	1L.....	60
IC0H_B	1	59
AAF00489	1	60
ZYR8_B	1	59
IDXU_B	1M.....	59
IHDB_B	1	59
IDXV_B	2	59
JXMF_C	2	59
AAI68978	1K.....	60
INOP_B	1	59
IKIK_B	1K.....	59
AAAN11320	1V.....	60
XP_002822173	1	60
LY85_B	1	59
LY80_B	1M.....A.....	59
IO10_B	1M.....	59
CAA23759	1V.....X.....	60
LYE2_B	1M.....F.....	59
LY5F_B	1M.....	59
IA00_B	1M.....Y.....	59

Common problems

- Selecting the wrong version of BLAST
- Selecting the wrong database
- Too many hits returned
- Too few hits returned
- Unclear about the significance of a particular result - are these sequences homologous?

146

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

147

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)

148

Summary of key points

- Sequence alignment is a fundamental operation underlying much of bioinformatics.
- Even when optimal solutions can be obtained they are not necessarily unique or reflective of the biologically correct alignment.
- Dynamic programming is a classic approach for solving the pairwise alignment problem.
- Global and local alignment, and their major application areas.
- Heuristic approaches are necessary for large database searches and many genomic applications.

FOR NEXT CLASS...

Check out the online:

- [Reading](#): Sean Eddy's "What is dynamic programming?"
- [Homework](#): (1) [Quiz](#), (2) [Alignment Exercise](#).

Homework Grading

Both (1) quiz questions and (2) alignment exercise carry equal weights (*i.e.* 50% each).

(Homework 2) Assessment Criteria	Points	
Setup labeled alignment matrix	1	
Include initial column and row for GAPs	1	
All alignment matrix elements scored (<i>i.e.</i> filled in)	1	
Evidence for correct use of scoring scheme	1	
Direction arrows drawn between all cells	1	
Evidence of multiple arrows to a given cell if appropriate	1	D
Correct optimal score position in matrix used	1	C
Correct optimal score obtained for given scoring scheme	1	B
Traceback path(s) clearly highlighted	1	A
Correct alignment(s) yielding optimal score listed	1	A+