

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/>

The screenshot shows the NCBI homepage with a search bar containing 'ras'. A red box highlights the search term. The page includes a sidebar with links to various NCBI resources like PubMed, Bookshelf, and BLAST.

<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**

The screenshot shows the search results for 'ras' on the NCBI database search page. A red box highlights the 'Gene' result, which has a count of 87,165 and a description of 'collected information about gene loci'.

<http://www.ncbi.nlm.nih.gov/query?term=ras>

NCBI Resources How To Sign in to NCBI

Gene Gene Search Help

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

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

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)]		rasr
ID: 19412			
ras	rasberry [Drosophila melanogaster (fruit fly)]	Chromosome X, NC_004354.4 (10744502..10749097)	Dmel CG1799, CG11485, CG1799, Dmel CG1799, EP(X)1093,
ID: 43873			

Filters: Manage Filters Top Organisms [Tree] **Homo sapiens (1126)** Mus musculus (623) Rattus norvegicus (625) Ochromis niloticus (533) Neotamias pictus (507) All other taxa (82019) More... Find related data Database: Select Find items Search details ras[All Fields] AND alive[property]

NCBI Resources How To Sign in to NCBI

Gene Gene Search Help

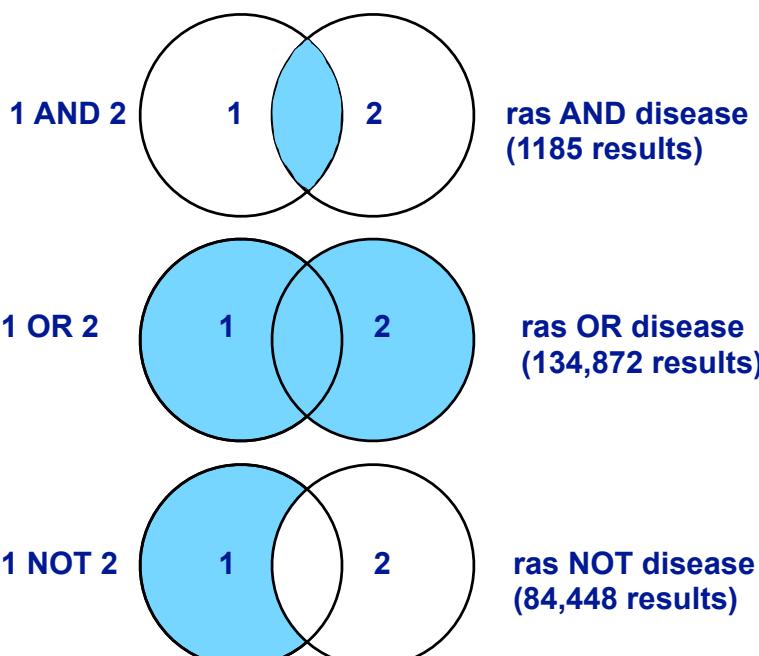
Show additional filters Display Settings: Tabular, 20 per page, Sorted by Relevance Send to: Hide sidebar >

Clear all Gene sources Genomic

Results: 1 to 20 of 1126 << First < Prev Page 1 of 57 Next >> Last >> 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 (114704464..114716894, complement)	RPS-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, DASK2
ID: 3845			

Filters: Manage Filters Find related data Database: Select Find items Search details ras[All Fields] AND "Homo sapiens"[orgn] AND alive[property]



NCBI Resources How To Sign in to NCBI

Gene Gene Search Help

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

Clear all Gene sources Genomic

Results: 1 to 20 of 1126 << First < Prev Page 1 of 57 Next >> Last >> 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 (114704464..114716894, complement)	RPS-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, DASK2
ID: 3845			

Filters: Manage Filters Find related data Database: Select Find items Search details ras[All Fields] AND "Homo sapiens"[orgn] AND alive[property]

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:ENSG0000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
 Gene type protein coding
 RefSeq status REVIEWED
 Organism Homo sapiens
 Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; 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)

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:ENSG0000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
 Gene type protein coding
 RefSeq status REVIEWED
 Organism Homo sapiens
 Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; 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)

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 (25205246..25250923, complement)

Annotation release: 105 Status: previous assembly Assembly: GRCh37.p13 (GCF_000001405_25) Chr: 12 Location: NC_000012.11 (25358100..25403870, complement)

Chromosome 12 - NC_000012.12

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

15

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:ENSG0000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
 Gene type protein coding
 RefSeq status REVIEWED
 Organism Homo sapiens
 Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; 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)

Gene Ontology Provided by GOA

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

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



GO: Gene Ontology

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

UniProt-GOA

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 biocurator . UniProt manual and electronic GO annotations are supplemented with manual annotations supplied by external collaborating GO Consortium groups, to ensure a comprehensive GO annotation dataset is supplied to users .

UniProt is a member of the GO Consortium .

Menu

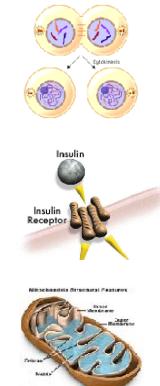
- Downloads
- Searching UniProt-GOA
- Annotation Methods
- Annotation Tutorial
- Manual Annotation Efforts
- Reference Genome Annotation Initiative
- Cardiovascular Gene Ontology Annotation Initiative
- Renal Gene Ontology Annotation Initiative
- Enzyme Gene

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.

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

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

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

Function
Ras protein 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	Graphical view		
Nucleotide binding ¹	29 – 35	7	GTP 2 Publications	Graphical view		
Nucleotide binding ¹	59 – 60	2	GTP 2 Publications	Graphical view		

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

Display: None
 FUNCTION
 NAMES & TAXONOMY
 SUBCELLULAR LOCATION
 PATHOL/BIOTECH
 PTM / PROCESSING
 EXPRESSION
 INTERACTION
 STRUCTURE
 FAMILY & DOMAINS
 SEQUENCES (2)
 CROSS-REFERENCES

Function
Ras protein 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	Graphical view		
Nucleotide binding ¹	29 – 35	7	GTP 2 Publications	Graphical view		
Nucleotide binding ¹	59 – 60	2	GTP 2 Publications	Graphical view		

P01116 - RASK_HUMAN
Pathology & Biotech

Involvement in disease
LEUKEMIA, ACUTE MYELOGENOUS (AML)
[MIM:601625]: 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 changes in cells that normally produce neutrophils, basophils, eosinophils and monocytes. 1 Publication

Note: The disease is caused by mutations affecting the gene represented in this entry.

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Neutral variant ¹	10 – 10	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.	Graphical view	VAR_034601	

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

Note: The disease is caused by mutations affecting the gene represented in this entry.

NONAN 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
Structure

Secondary structure
Legend: Helix Turn Beta strand
Show more details

3D structure databases

Selected link destination	Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
RCSB PDB ¹	1D8D	X-ray	2.00	P	178-188	[>]
RCSB PDB ¹	1D8B	X-ray	3.00	P	178-188	[>]
RCSB PDB ¹	1K2O	X-ray	2.20	C	169-173	[>]
RCSB PDB ¹	1K2P	X-ray	2.10	C	169-173	[>]
RCSB PDB ¹	3GFT	X-ray	2.27	A/B/C/D/E/F	1-164	[>]
RCSB PDB ¹	4DSN	X-ray	2.03	A	2-164	[>]
RCSB PDB ¹	4DSO	X-ray	1.85	A	2-164	[>]
RCSB PDB ¹	4EPR	X-ray	2.00	A	1-164	[>]
RCSB PDB ¹	4EPT	X-ray	2.00	A	1-164	[>]
RCSB PDB ¹	4EPV	X-ray	1.35	A	1-164	[>]
RCSB PDB ¹	4EPW	X-ray	1.70	A	1-164	[>]
RCSB PDB ¹	4EPX	X-ray	1.76	A	1-164	[>]
RCSB PDB ¹	4EPY	X-ray	1.80	A	1-164	[>]
RCSB PDB ¹	4L8G	X-ray	1.52	A	1-164	[>]
RCSB PDB ¹	4LDJ	X-ray	1.15	A	1-164	[>]
RCSB PDB ¹	4LPK	X-ray	1.50	A/B	1-169	[>]

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?

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

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

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

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

[View in 3D: NGL or JSmol \(in Browser\)](#)

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

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

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]125O - [GLY]125C

Back to UniProt:

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

Display: None
FUNCTION
NAMES & TAXONOMY
SUBCELL. LOCATION
PATHOBIOLOGY
PTM / PROCESSING
EXPRESSION
INTERACTION
STRUCTURE
FAMILY & DOMAINS
SEQUENCES (2)
CROSS REFERENCES
PUBLICATIONS
ENTRY INFORMATION
MICROARRAYS
SIMILAR PROTEINS
Top

Family and domain databases
Gene3D: 3,40,50,300, 1 hit
InterPro: IPR027417, P-loop_NTPase, IPR005225, Small_GTP_bdg_dom, IPR01866, Small_GTPase, IPR020849, Small_GTPase_Res, [Graphical view]
PANTHER: PTI024070, PTI024070, 1 hit
PFam: PF00071, Ras, 1 hit, [Graphical view]
PRINTS: PR00449, RASTNSPRMNG, SVAH1: SM00173, RAS, 1 hit, [Graphical view]
SLPfam: SSF52540, SSF52540, 1 hit
TIGRFAMs: TIGR00231, small_GTP, 1 hit
PROSITE: PS51421, RAS, 1 hit, [Graphical view]

PFAM is one of the best protein family databases

Sequences (2)
Sequence status¹: Complete.
Sequence processing¹: The displayed sequence is further processed into a mature form.
This entry describes 2 isoforms¹ 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: Ras family
Domain organization
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures
Jump to...
InterPro: 0

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 3D 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 to refer to a class of genes coding those proteins. When Ras is activated² 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.^{1,2} The 3 Ras genes in human (h-Ras, KRAS, and NRAS) are the most common oncogenes in human cancer, mutations that permanently activate Ras are found in 20% to 25% of all human tumors and up to 90% in certain types of cancer (e.g., pancreatic cancer). For this reason, Ras inhibitors are being studied as a treatment for cancer, and other diseases with Ras overexpression.

Contents [Edit]

1. Summary
2. Domains
3. Families
3.1 Activation and deactivation
3.2 Membrane attachment
4. Members
5. Kinases
5.1 Receptor-like activation
5.2 Constitutively active Ras

Identifiers
Symbol: Ras
Name: PF00071_E
InterPro: IPR037531_U
PROSITE: PS030617_G
SCOP: 1g21_B
SUPERFAMILY: SP01_P

Example Questions:

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

This visual化 provides a simple graphical representation of the distribution of this family across species. You can find the original interactive tree in the [alignment](#).

Sunburst controls

Species distribution

Jump to... enter ID/acc Go

Selections

- Align selected sequences to HMM
- Download a PASTA-format file
- Copy selected sequences
- Currently selected:
 - 501 sequences
 - None selection tool shows results in pop-up windows. Please disable Java script blockers.

Example Questions:

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

Plan: Pfam alignment viewer

Alignment for selected sequences

Currently showing rows 1 to 30 of 501 rows in this alignment. Show 30 more rows of alignment

P51242/1/1-28	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/29-46	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/47-52	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/53-62	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/63-72	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/73-82	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/83-92	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/93-102	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/103-112	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/113-122	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/123-132	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/133-142	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/143-152	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/153-162	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/163-172	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/173-182	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/183-192	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/193-202	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/203-212	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/213-222	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/223-232	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/233-242	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/243-252	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/253-262	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/263-272	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/273-282	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/283-292	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/293-302	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/303-312	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/313-322	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/323-332	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/333-342	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/343-352	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/353-362	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/363-372	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/373-382	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/383-392	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/393-402	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/403-412	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/413-422	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/423-432	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/433-442	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/443-452	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/453-462	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/463-472	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/473-482	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/483-492	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/493-502	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/503-512	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/513-522	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/523-532	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/533-542	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/543-552	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/553-562	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/563-572	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/573-582	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/583-592	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/593-602	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/603-612	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/613-622	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/623-632	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/633-642	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/643-652	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/653-662	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/663-672	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/673-682	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/683-692	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/693-702	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/703-712	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P	S	R	I	L	N	G	...
P51242/1/713-722	R	I	A	M	E	L	C	F	...	D	D	T	P	S	R	E	V	...	P</td							

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

UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View		
A8BKD1_GIALA	11 - 335	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		
				1i6	A	24 - 359	Jmol AstexViewer SPICE
				B	24 - 359	Jmol AstexViewer SPICE	
				1g0b	A	24 - 359	Jmol AstexViewer SPICE
				1x88	B	24 - 359	Jmol AstexViewer SPICE
		A	24 - 359	Jmol AstexViewer SPICE			

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

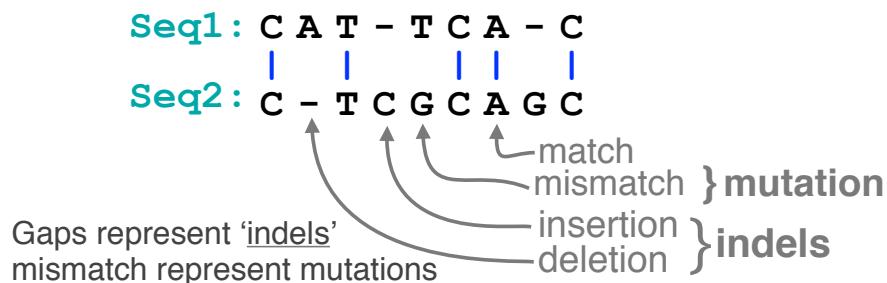
Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.

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!*

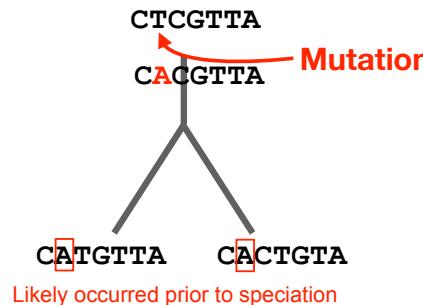
ALIGNMENT FOUNDATIONS

- 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

Mutations, deletions and insertions

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

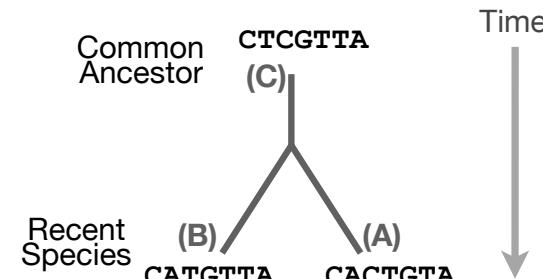
- **Mutations/Substitutions** $\text{CTCGTTA} \rightarrow \text{CAGTTA}$
- Deletions
- Insertions



Sequence changes during evolution

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

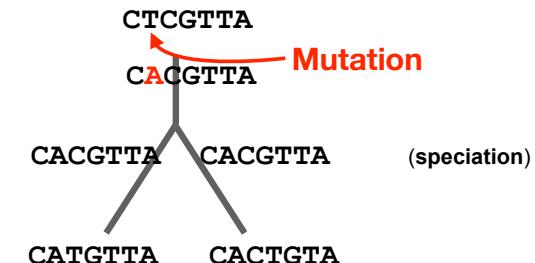
- Mutations/Substitutions
- Deletions
- Insertions



Mutations, deletions and insertions

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

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

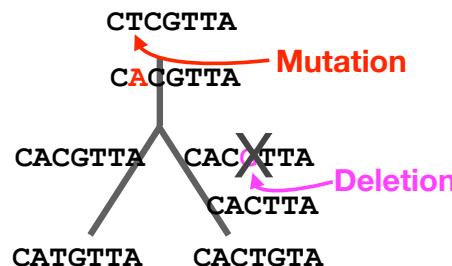


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

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

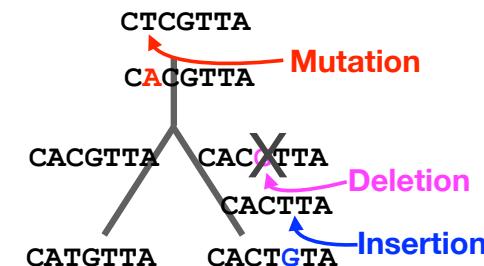


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

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

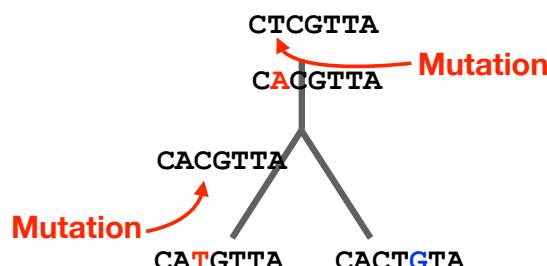


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

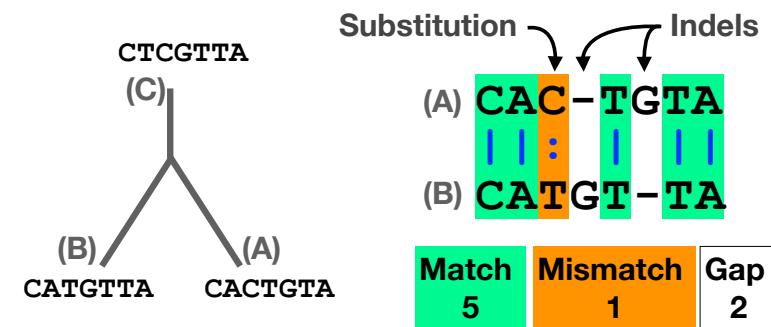
$$\text{CTCGTTA} \rightarrow \text{CACGTTA}$$
$$\text{CACGTTA} \rightarrow \text{CATGTTA}$$



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

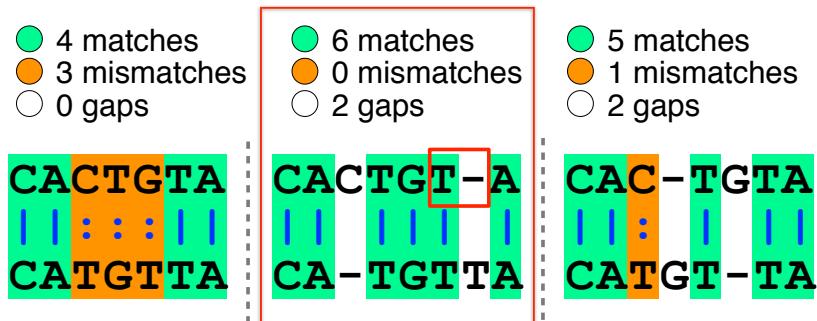
CACTGTA
||:||:
CATGTTA

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

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

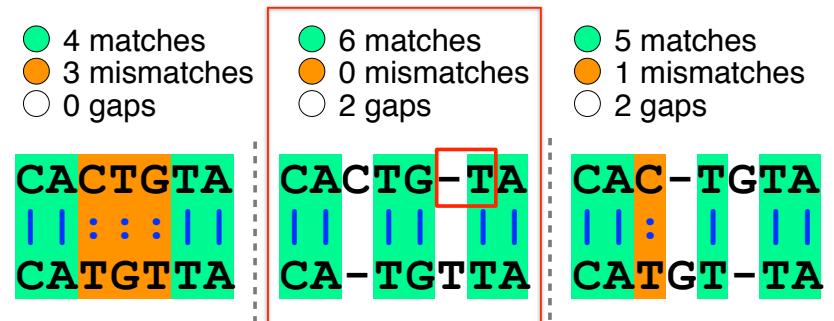
Optimal alignments

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



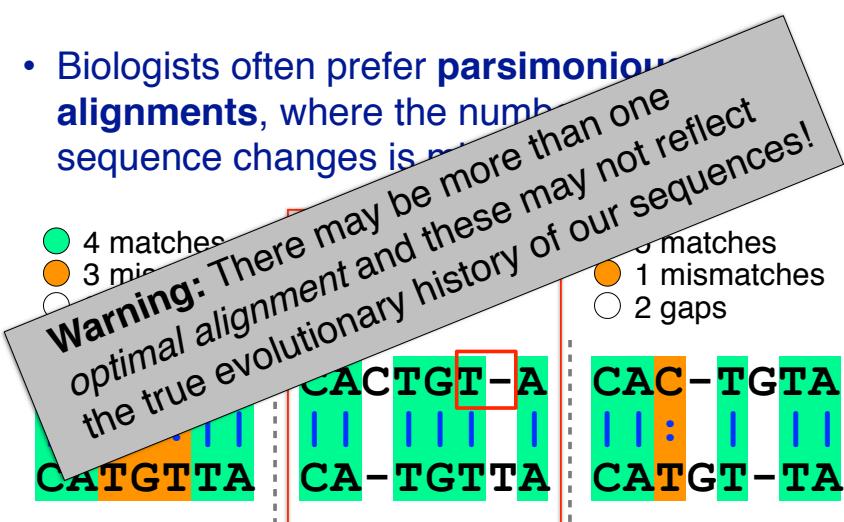
Optimal alignments

- 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

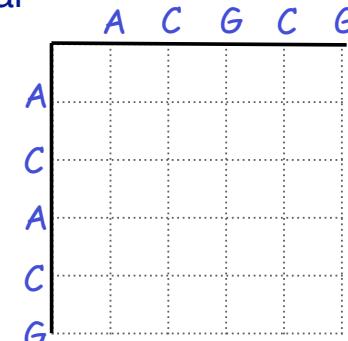
- Why...
 - Why compare biological sequences?
- What...
 - Alignment view of sequence changes during evolution (matches, mismatches and gaps)

• How...

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

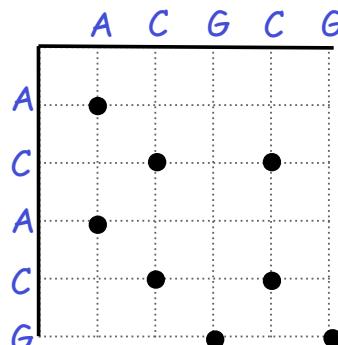
Dot plots: simple graphical approach

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



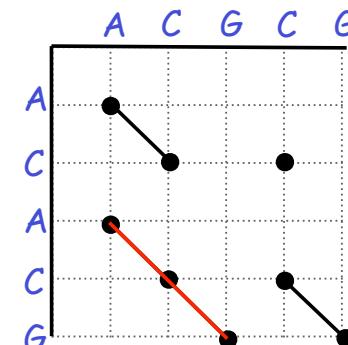
Dot plots: simple graphical approach

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



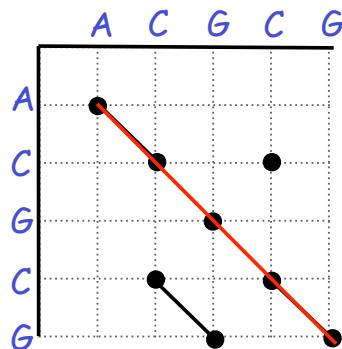
Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



Dot plots: simple graphical approach

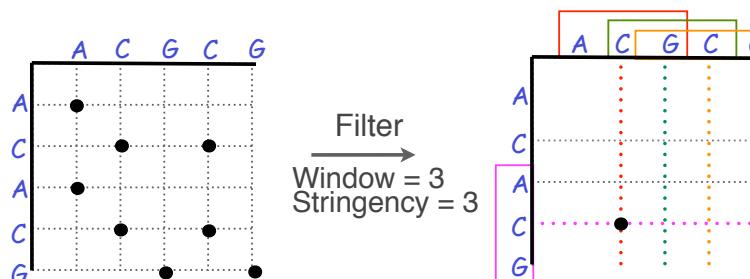
Q. What would the dot matrix of a two identical sequences look like?



Dot plots: window size and match stringency

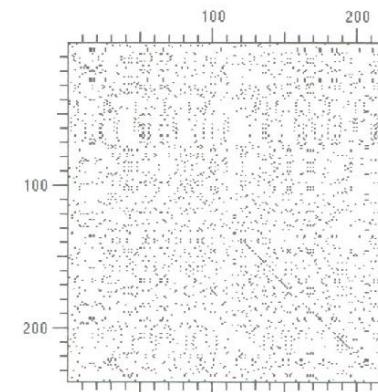
Solution: use a window and a threshold

- 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: simple graphical approach

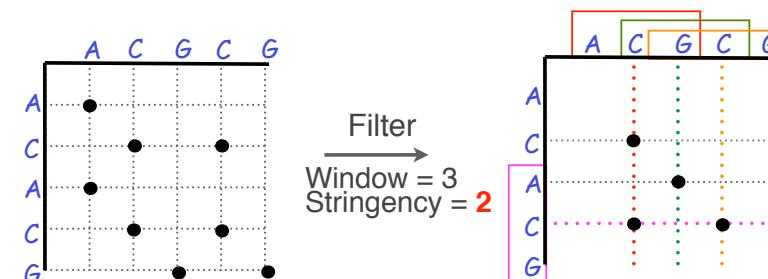
- Dot matrices for long sequences can be noisy



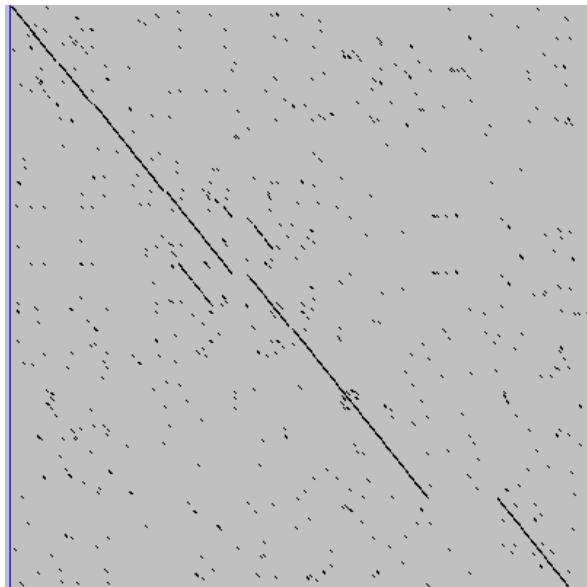
Dot plots: window size and match stringency

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



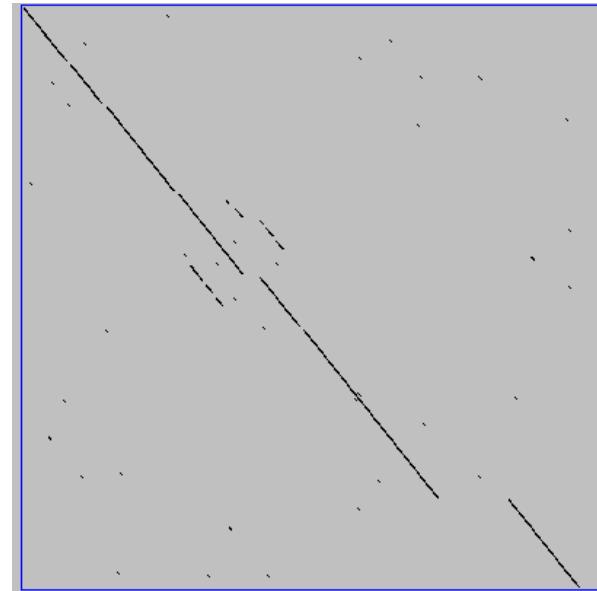
Window size = 5 bases



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

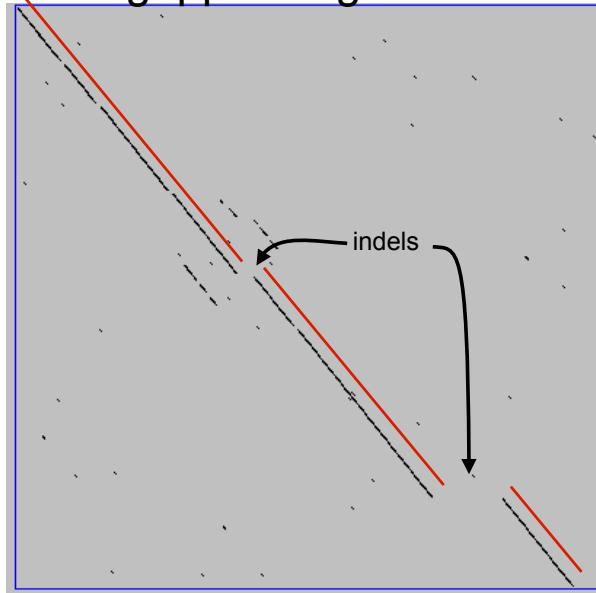


This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

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



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

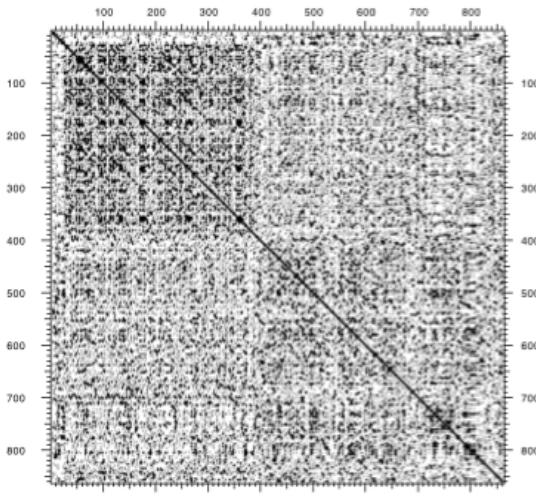
Only **diagonals** can be followed.

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

Uses for dot matrices

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

Repeats

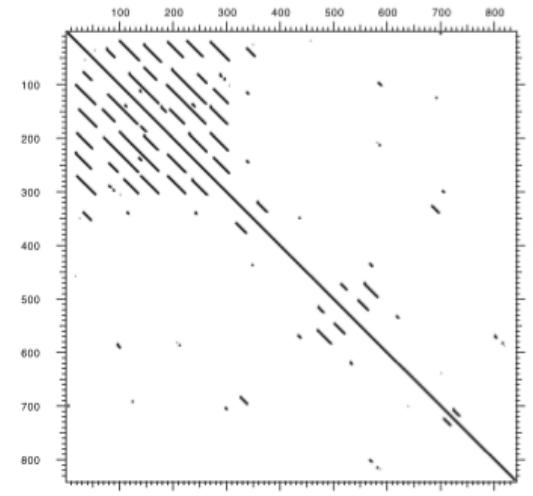


Human LDL receptor
protein sequence
(Genbank P01130)

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

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

Repeats



Human LDL receptor
protein sequence
(Genbank P01130)

$$W = 23 \\ S = 7$$

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

Your Turn!

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

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

The screenshot shows a web-based tool for comparing two sequences. It features two side-by-side dot plots: a 'Protein Dot Plot' and a 'DNA Dot Plot'. Both plots have 'Sequence 1' on the x-axis and 'Sequence 2' on the y-axis, with scales from 0 to 150. The 'Protein Dot Plot' has a 'wslice' value of 3, 'wstep' value of 3, and 'nnmatch' value of 2. The 'DNA Dot Plot' has the same parameters. The 'Dot Plot Parameters' section includes sliders for 'Window Size' (set to 3), 'Moving window step size' (set to 3), and 'Match stringency' (set to 2). Below the plots, there's a 'Questions for discussion:' section with three bullet points:

- Why does the DNA sequence have more dots than the protein sequence plot?
- How can we increase the signal to noise ratio?
- What does a 'Match stringency' value of 3 mean?

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.

81

Scoring the alignment matrix

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

		Sequence 2					
		j	D	P	L	E	
		Scores:	match = +1	mismatch = -1	gap = -2		
		-	0	-2	-4	-6	-8
Sequence 1		D	-2				
		P	-4				
		M	-6				
		E	-8				

Algorithm of Needleman and Wunsch

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



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

Scoring the alignment matrix

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

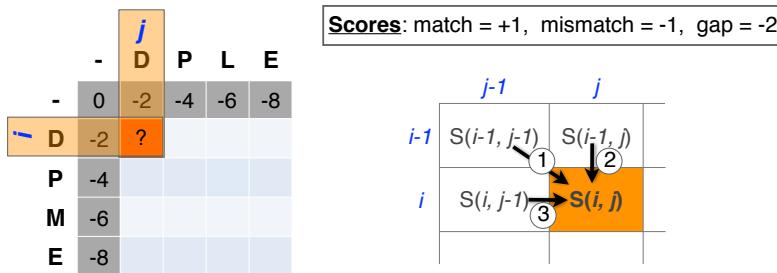
		Sequence 2					
		j	D	P	L	E	
		Scores:	match = +1	mismatch = -1	gap = -2		
		-	0	-2	-4	-6	-8
Sequence 1		D	-2				
		P	-4				
		M	-6				
		E	-8				

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

Seq1: DPME
Seq2: ----

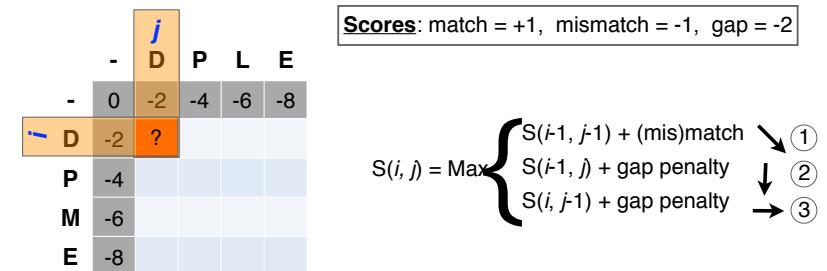
Scoring the alignment matrix

- 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



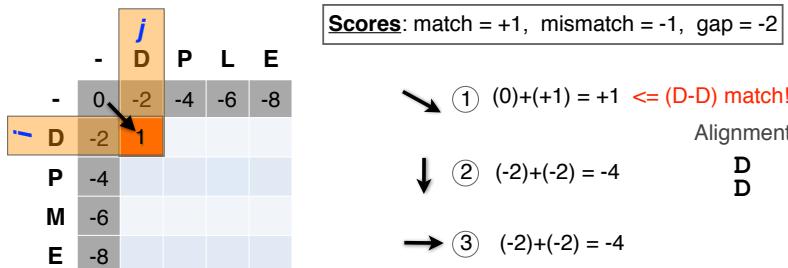
Scoring the alignment matrix

- 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



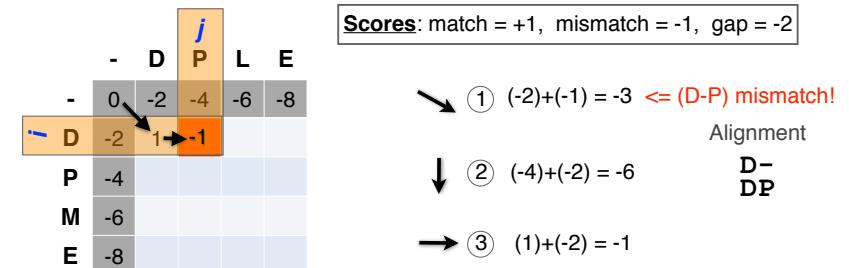
Scoring the alignment matrix

- 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



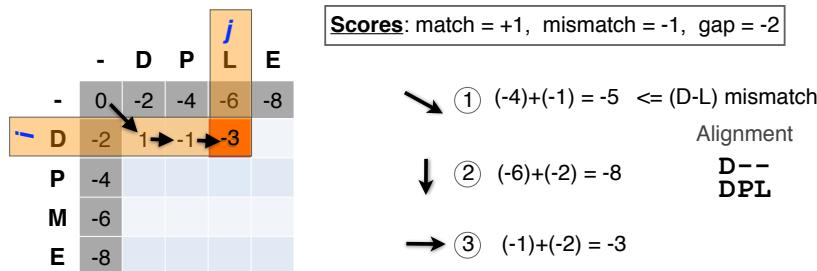
Scoring the alignment matrix

- 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)



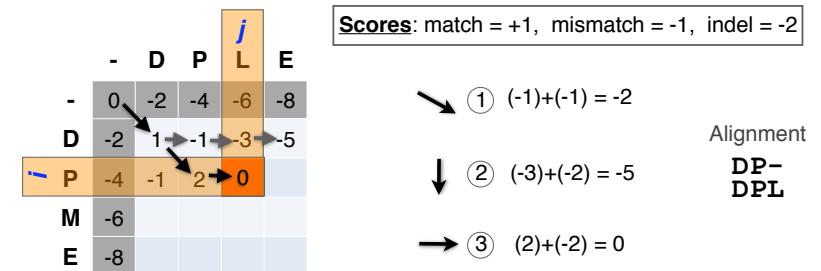
Scoring the alignment matrix

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



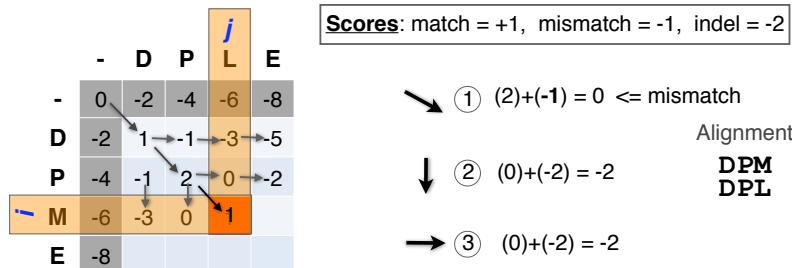
Scoring the alignment matrix

- 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.



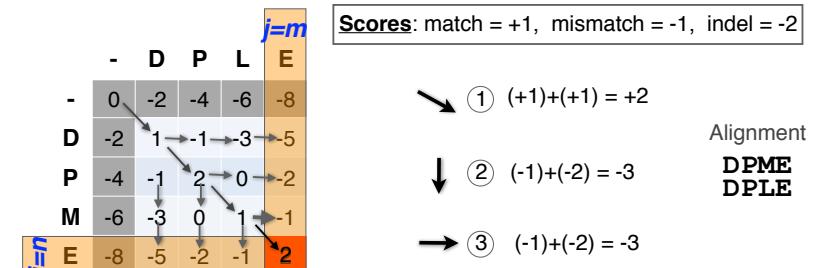
Scoring the alignment matrix

- 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



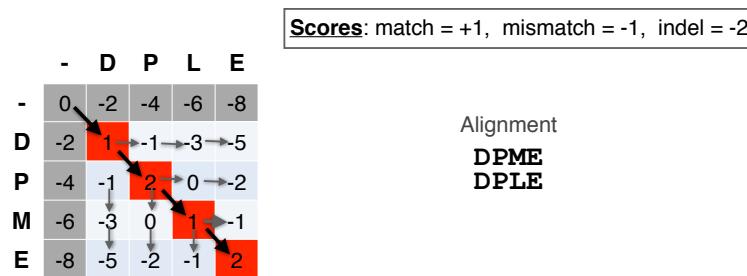
Scoring the alignment matrix

- 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)



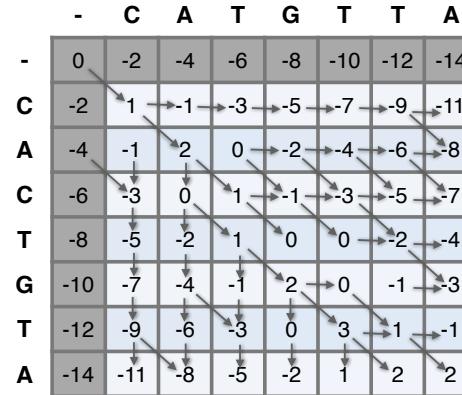
Scoring the alignment matrix

- 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



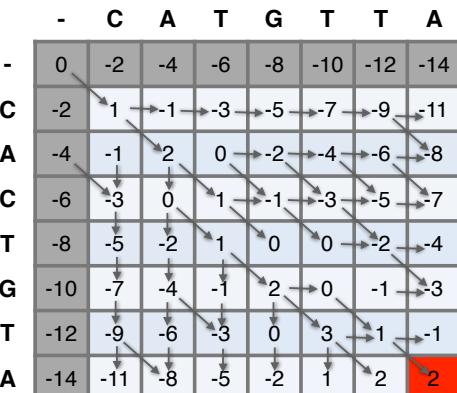
Questions:

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



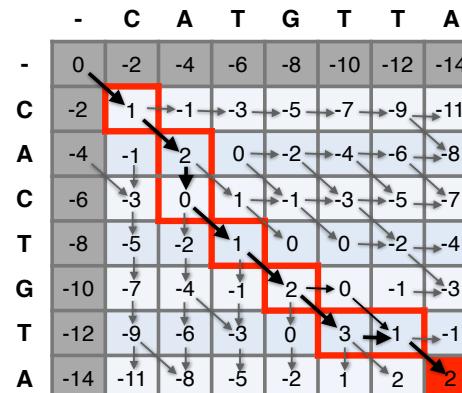
Questions:

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



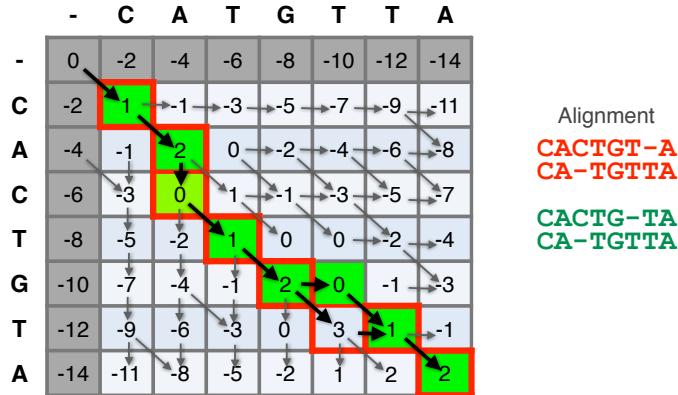
Questions:

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



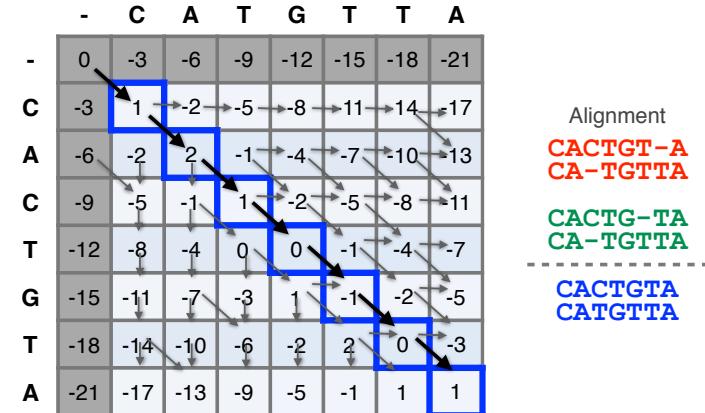
More than one alignment possible

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



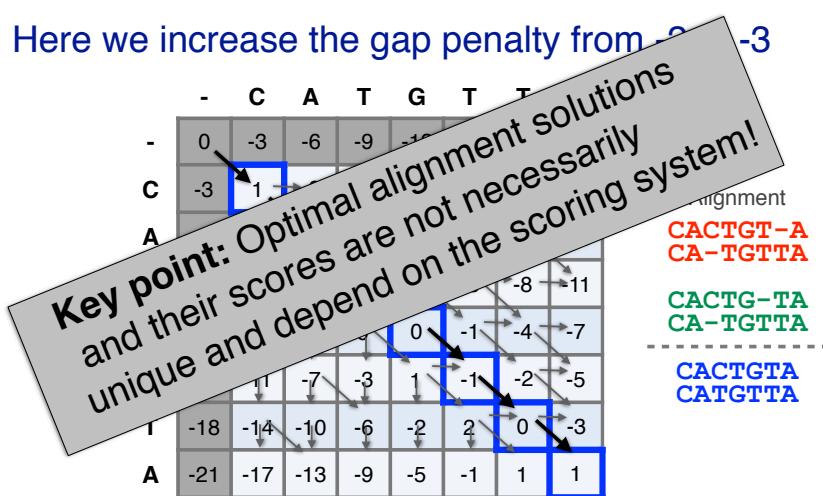
The alignment and score are dependent on the scoring system

- 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



Your Turn!

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

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

	A	G	T	T	C
A	0				
T					
T					
G					
C					

NW DYNAMIC PROGRAMMING

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

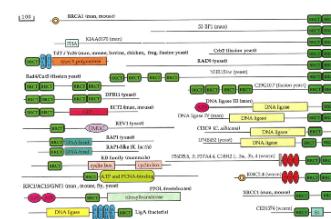
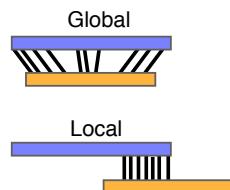
	A	G	T	T	C	
0	-2	-4	-6	-8	-10	
A	-2	+2	0	-2	-4	-6
T	-4	0	+1	+2	0	-2
T	-6	-2	-1	+3	+4	+2
G	-8	-4	0	+1	+2	+3
C	-10	-6	-2	-1	0	+4

ALIGNMENT FOUNDATIONS

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

Global vs local alignments

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



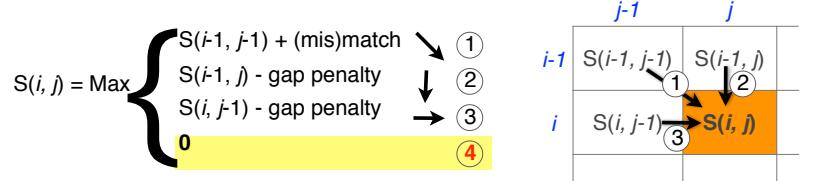
Local alignment: Definition

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

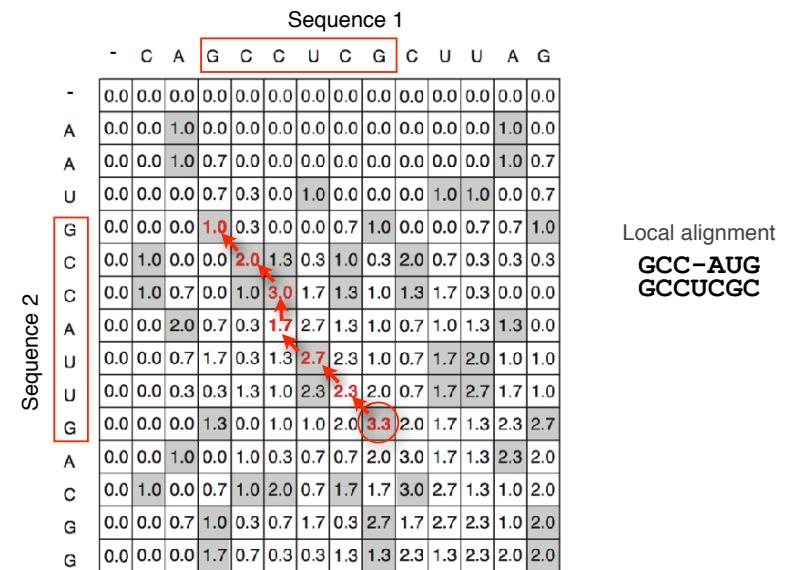
Smith, T.F. & Waterman, M.S. (1981) "Identification of common molecular subsequences." J. Mol. Biol. 147:195-197.

The Smith-Waterman algorithm

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



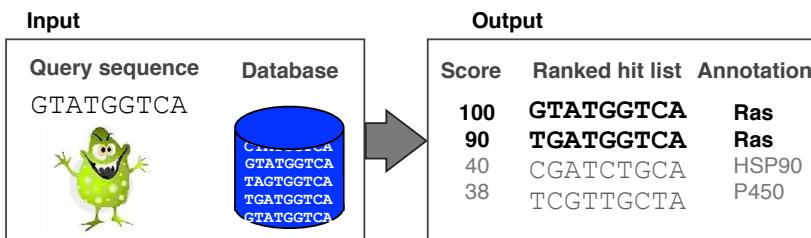
105



106

Local alignments can be used for database searching

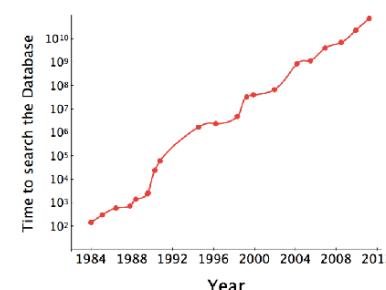
- Goal:** Given a query sequence (Q) and a sequence database (D), find a list of sequences from D that are most similar to Q
 - Input:** Q, D and scoring scheme
 - Output:** Ranked list of hits



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!**

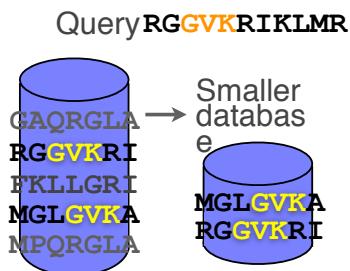


To reduce search time **heuristic algorithms**, such as BLAST, first remove database sequences without a strong local similarity to the query sequence in a quick initial scan.

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

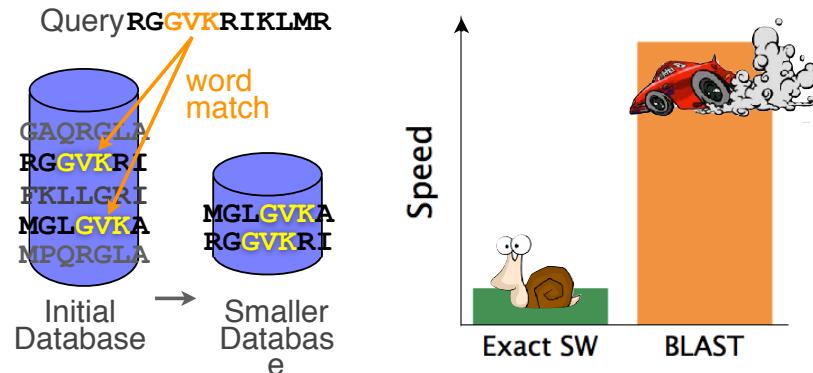
Rapid, heuristic versions of Smith–Waterman: **BLAST**

- 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 two sequences
 - BLAST does not examine every sequence pair in the database, instead it scans the database 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)

111

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 ($w=3$)

RGGVKRI Query sequence
RGG
GGV
GVK
VKR
KRI

generate list
of $w=3$
words for
query

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)

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

115

Blast

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

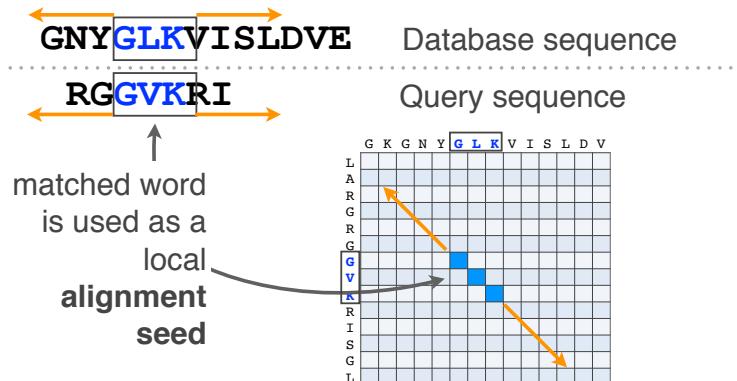
GNYGLKVISLDVE Database sequence
RGGVKRI Query sequence
RGG RAG RIG RLG ...
GGV GAV GTV GCV ...
GVK GLK GIK GGK ...
VKR VRR VHR VER ...
KRI KKI KHI KDI ...

search for perfect matches in the database sequence

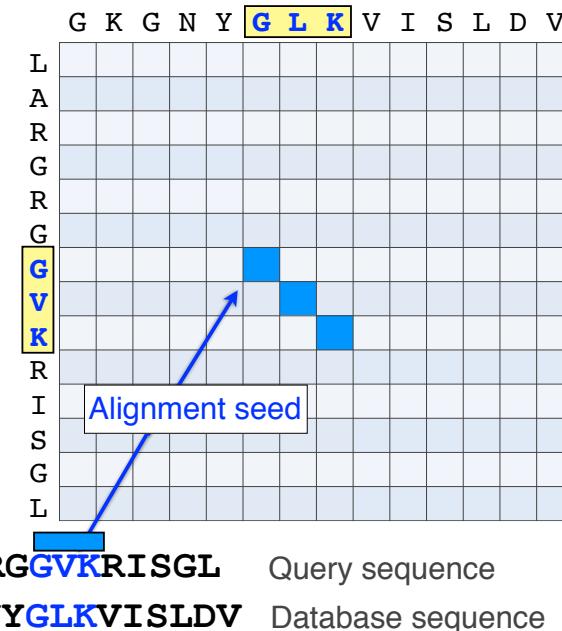
116

Blast

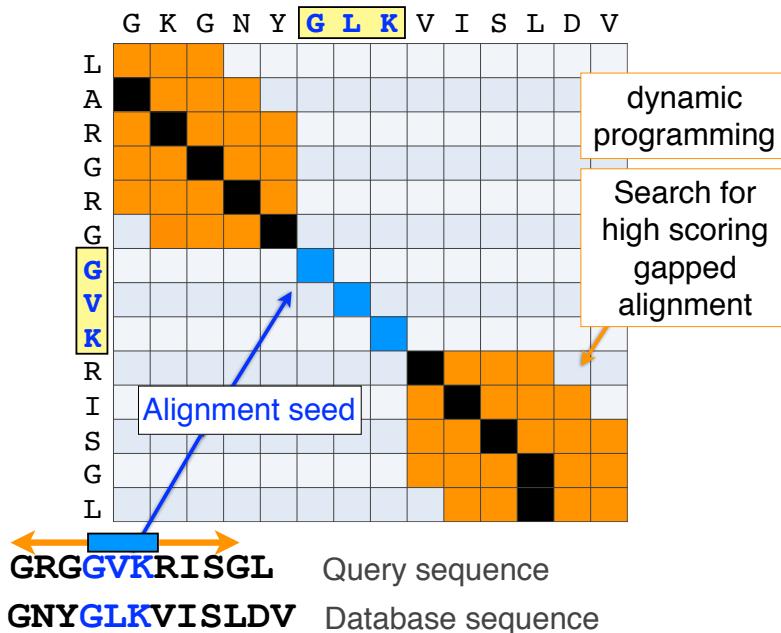
- Phase 4: the initial database hits are extended in both directions using dynamic programming



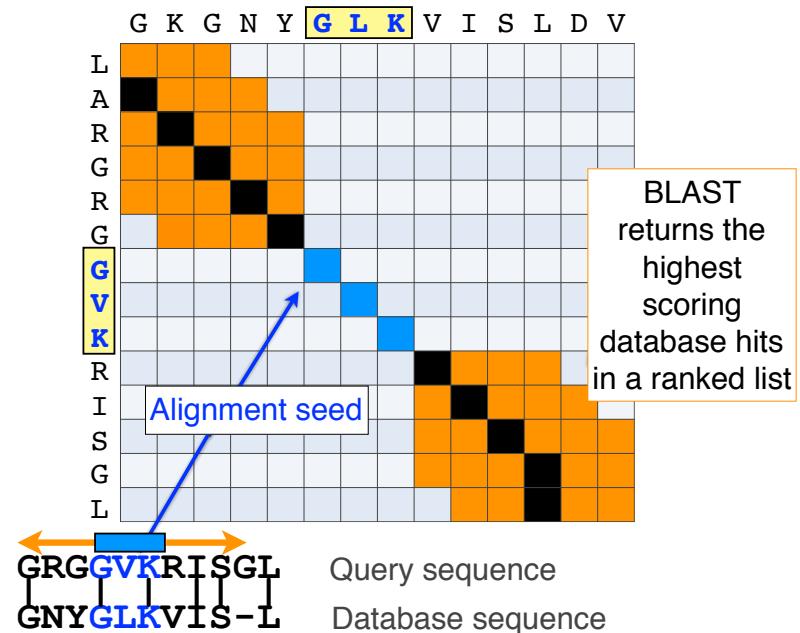
117



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

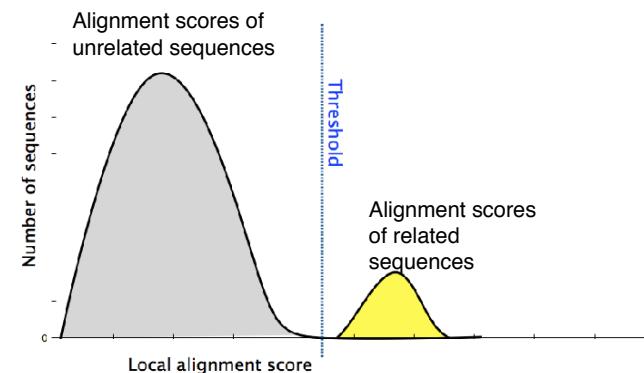
122

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

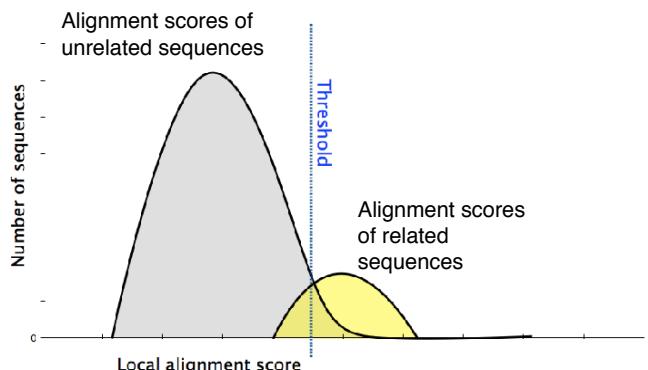
123

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



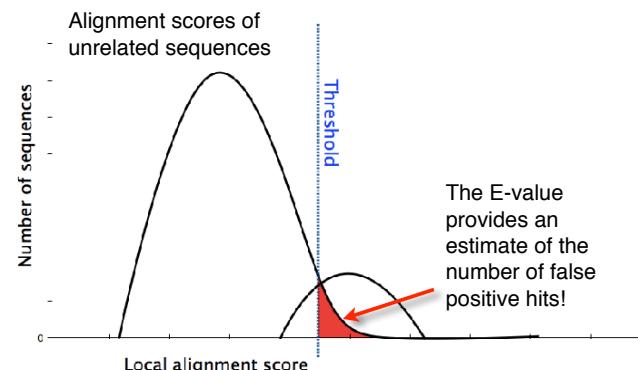
124

- 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



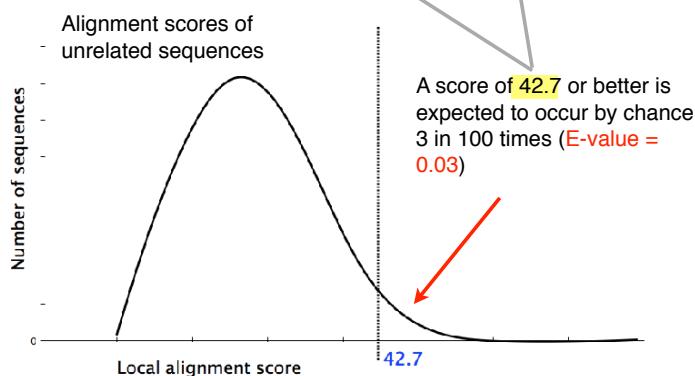
125

- 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



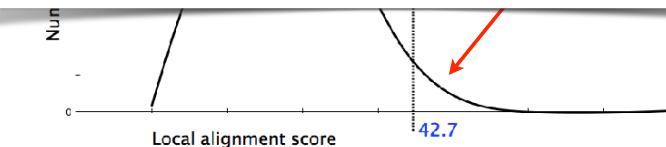
127

Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo 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>



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

The screenshot shows the NCBI BLAST Home Page. At the top, there's a navigation bar with links for Home, Recent Results, Saved Strategies, and Help. A red box highlights the 'Home' link. Below the navigation, there's a search bar with the placeholder 'BLAST finds regions of similarity between biological sequences.' and a 'Search' button. To the right of the search bar, there's a 'My NCBI' section with 'Sign In' and 'Register' buttons. A news feed is visible on the right side. The main content area is titled 'Basic Local Alignment Search Tool' and contains sections for 'NCBI BLAST Home', 'BLAST Assembled RefSeq Genomes', 'Basic BLAST', and 'Specialized BLAST'. Under 'Basic BLAST', there are several 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). A large red box highlights the 'tblastx' option. On the right side of the page, there's a sidebar with a 'How to do BLAST' link and a note about BLAST being useful for examining a large group of potential gene candidates.

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”

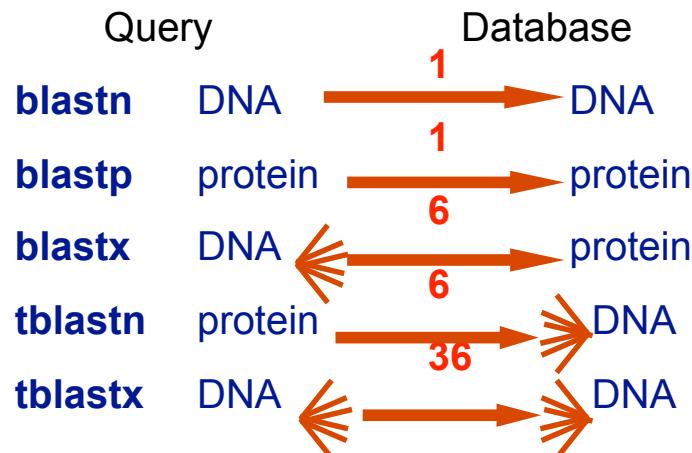
Step 1: Choose your sequence

- 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]'. At the top, there's a search bar with 'Protein' selected and a 'Search' button. Below the search bar, there are 'Display Settings' and a 'FASTA' checkbox, which is checked and highlighted with a red circle. The search results show the sequence information: 'NCBI Reference Sequence: NP_000509.1' and the full sequence: '>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens] MVLHLPPEEKAVTALWGKVNVDEVGGEALGRLLVYVWTQRFESFGDLSTPDAMGNPKVKAHGKVKVLG AFSDGLAHLNDLKGTFATLSELHCKDNLHVDPENFRLLGVNVLVCVLAHHFGKEFTPPVQAYQKVVAGVAN ALAHKYH'. There are also links for 'GenPept' and 'Graphics'.

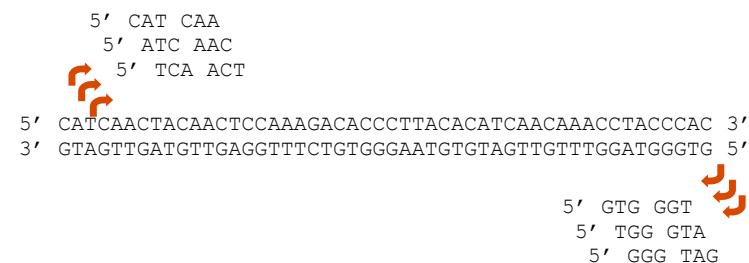
132

Step 2: Choose the BLAST program



133

DNA potentially encodes six proteins



134

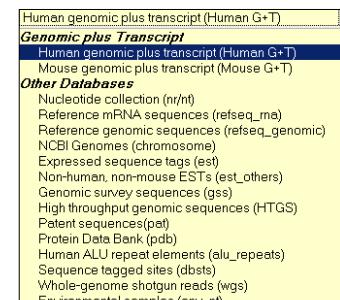
The screenshot shows the Protein BLAST search interface at blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch. Key elements include:

- Enter Query Sequence:** Contains a text input field with the sequence >gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens] MVHLPEEKSAVATLWCKVNDEVGGEALGRLLVPPWTORFESGDSLTPDAVMGNPKVKAHKK KVLAGTSSGALIIDLNLKGTTATLSEIICDKLIIVDPEVNRLGNVLVCVLAAIIIIGKEITPPVQAAQYK WAVCANALAHKYH.
- Choose Search Set:** Includes a dropdown menu set to "Non-redundant protein sequences (nr)". Other options include "Organism" and "Exclude".
- Program Selection:** Includes a dropdown menu set to "blastp (protein-protein BLAST)". Other options include "PSI-BLAST", "PHI-BLAST", and "DELTA-BLAST".
- BLAST:** A large blue button labeled "BLAST".

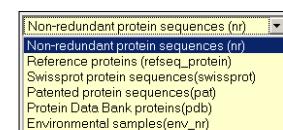
135

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

136

The screenshot shows the Protein BLAST search interface. In the 'Enter Query Sequence' section, a sequence for hemoglobin subunit beta from Homo sapiens is entered. Below it, there are fields for 'Organism' (Non-redundant protein sequences), 'Entrez' (Entrez Query), and 'Settings'. In the 'Algorithm parameters' section, 'blastp (protein-protein BLAST)' is selected. A large red arrow points to the 'Settings' button.

137

Step 4a: Select optional search parameters

This screenshot shows the 'Algorithm parameters' configuration window. Several parameters are highlighted with arrows: 'Expect threshold' (blue arrow), 'Word size' (orange arrow), 'Matrix' (BLOSUM62) (orange arrow), and 'Scoring matrix' (orange arrow). The 'Expect' parameter is labeled 'Expect' in blue, while 'Word size' is labeled 'Word size' in orange. Other visible parameters include 'Max target sequences' (100), 'Short queries' (checkbox), 'Gap Costs' (Existence: 11 Extension: 1), and 'Compositional adjustments'.

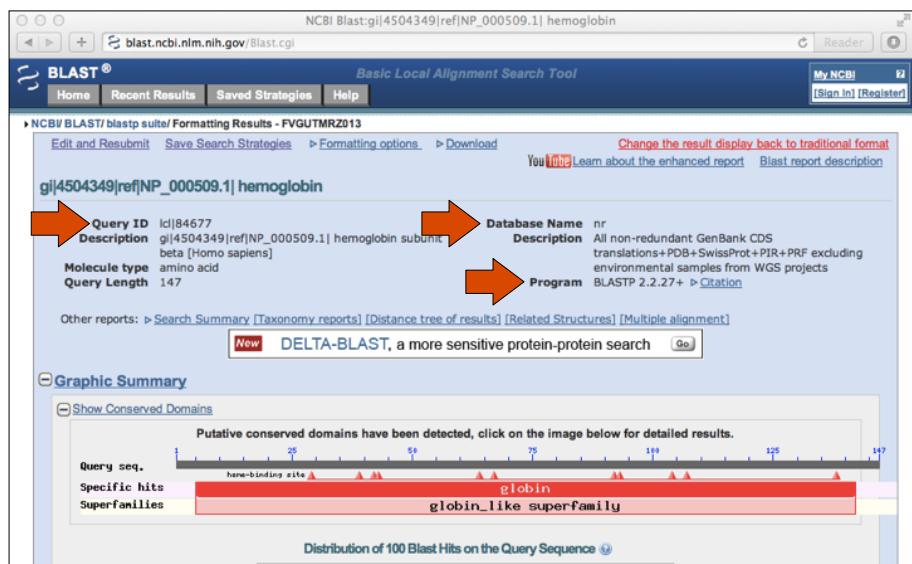
138

Step 4: Optional parameters

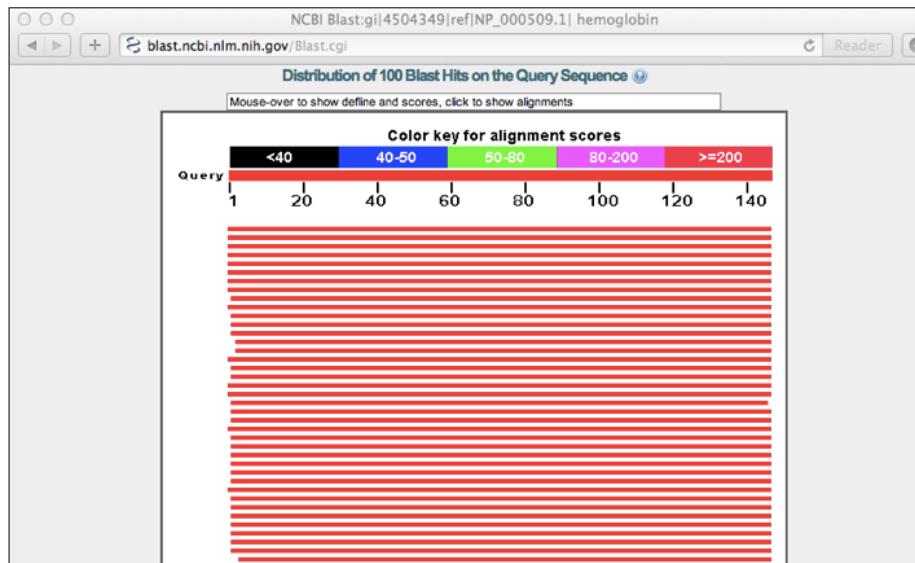
- You can...
 - choose the organism to search
 - change the substitution matrix
 - change the expect (E) value
 - change the word size
 - change the output format

139

Results page



Further down the results page...



Further down the results page...

This screenshot shows a table of sequences producing significant alignments. The table includes columns for Description, Max score, Total score, Query cover, E value, Max ident, and Accession. An orange arrow points to the "E value" column header. The table lists various hemoglobin subunit and chain entries with their respective scores and identities.

Description	Max score	Total score	Query cover	E value	Max ident	Accession
hemoglobin beta [synthetic construct]	301	301	100%	9e-103	100%	AAX37051.1
hemoglobin beta [synthetic construct]	301	301	100%	1e-102	100%	AAX29557.1
hemoglobin subunit beta [Homo sapiens] >ref XP_508242.1 PREDICTED: hemoglobin_s	301	301	100%	1e-102	100%	NP_000509.1
RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hem	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	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 AAA88054.1 beta-globin [Homo sa	298	298	100%	1e-101	99%	AAF00489.1
Chain B. Human Hemoglobin D Los Angeles: Crystal Structure >pdb 2YRS D.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...

This screenshot shows detailed alignment results for hemoglobin subunit beta. It includes a table of matches, a "Related Information" section with links to UniGene, Map Viewer, Structure, PubChem Bio, and Assay databases, and a second table for a different hemoglobin beta chain entry.

Different output formats are available

This screenshot shows the "Formatting options" interface for the NCBI BLAST results. It includes sections for 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), and other settings like Organism, Entrez query, Expect Min, Expect Max, Percent Identity Min, Percent Identity Max, and Format for (PSI-BLAST, inclusion threshold).

E.g. Query anchored alignments

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

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

Reader

Query	Score	Sequence	Start	End	Length
AAX37051	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
AAX29557	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
NP_000509	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
P02024	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
AAN84548	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
AAZ39780	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
ACU56984	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
AAD19696	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
I1COH_B	59	VHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
AAF00489	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
2YRS_B	59	VHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
IDXU_B	59	MHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1HDB_B	59	VHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
IDXV_B	59	HLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	2	59	59
J3KMF_C	59	HLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	2	59	59
AAI68978	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1NQP_B	59	VHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1KIK_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
AAN11320	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
XP_002822173	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
I1Y85_B	59	VHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1Y80_B	59	MHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1O10_B	59	MHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
CAA23759	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
I1Y82_B	59	MHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1Y5F_B	59	MHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1A00_B	59	MHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1HBS_B	59	VHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1ABY_B	59	MHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1CMY_B	59	VHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59

... and alignments with dots for identities

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

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

Reader

Query	Score	Sequence	Start	End	Length
AAX37051	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
AAX29557	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
NP_000509	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
P02024	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
AAN84548	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
AAJ39780	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
ACU56984	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
AAD19696	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1COH_B	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
AAF00489	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
2YRS_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
IDXU_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1HDB_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
IDXV_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	2	59	59
J3KMF_C	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	2	59	59
AAI68978	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
I1NQP_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1KIK_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
AAN11320	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
XP_002822173	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
I1Y85_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1Y80_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1O10_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
CAA23759	60	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	60	60
I1Y82_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1Y5F_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1A00_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1HBS_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1ABY_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59
I1CMY_B	59	MVHLTPEEKSAVTALWGVNVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	1	59	59

Common problems

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

How to handle too many results

- Focus on the question you are trying to answer
 - select “refseq” database to eliminate redundant matches from “nr”
 - Limit hits by organism
 - Use just a portion of the query sequence, when appropriate
 - Adjust the expect value; lowering E will reduce the number of matches returned

How to handle too few results

- Many genes and proteins have no significant database matches
 - remove Entrez limits
 - raise E-value threshold
 - search different databases
 - try scoring matrices with lower BLOSUM values (or higher PAM values)
 - use a search algorithm that is more sensitive than BLAST (e.g. PSI-BLAST or HMMer)

149

FOR NEXT CLASS...

Check out the online:

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

To Update!

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.

Homework Grading

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

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