

BGGN 213
Foundations of Bioinformatics
 Lecture 2
 Barry Grant
 UC San Diego
<http://thegrantlab.org/bggn213>

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 <i>why</i> and <i>how</i> 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 the search bar containing 'ras'. A red box highlights the search bar. A diagonal banner across the page reads 'Hands on demo (or see following slides)'. The page includes a navigation menu on the left, a 'Welcome to NCBI' message, and various resource links.

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 search bar contains 'ras' and the results show 'About 2,978,774 search results for "ras"'. The results are organized into columns for Literature, Genes, and Health. The 'Gene' entry is highlighted with a red box, showing 87,165 results and 'collected information about gene loci'.

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

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

Gene: ras

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

Results: 1 to 20 of 85633

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

Name/Gene ID	Description	Location	Aliases
ras ID: 19412	resistance to audiogenic seizures [<i>Mus musculus</i> (house mouse)]		asr
ras ID: 43873	raspberry [<i>Drosophila melanogaster</i> (fruit fly)]	Chromosome X, NC_004354.4 (10744502..10749097)	Dmel_CG1799, CG11485, CG1799, DmelCG1799, EP(X)1093,

Top Organisms: **Homo sapiens (1126)**, Mus musculus (823), Rattus norvegicus (625), Oreochromis niloticus (533), Neolamprologus brichardi (507), All other taxa (82019)

www.ncbi.nlm.nih.gov/gene

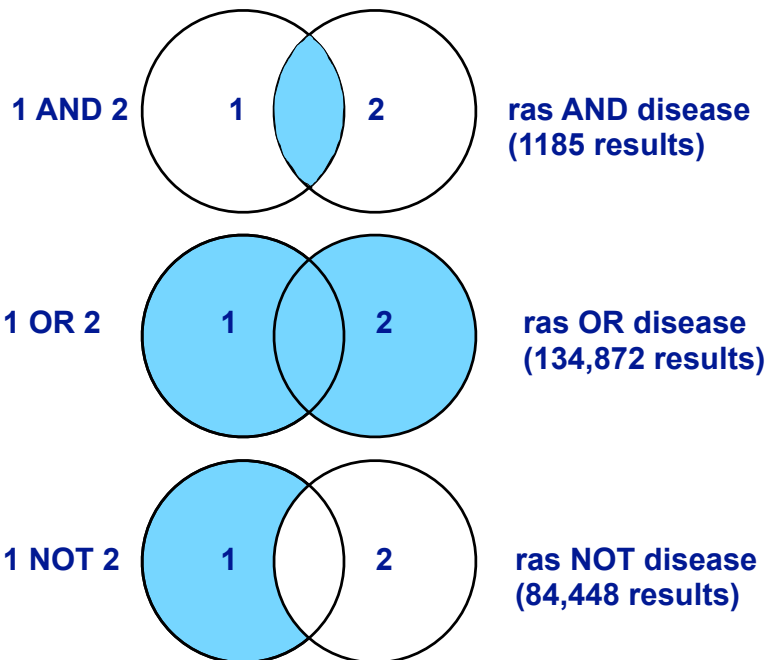
Gene: (ras) AND "Homo sapiens"[porgn:..txid9606]

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

Results: 1 to 20 of 1126

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

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KIRAS1, KRAS2, NS,



www.ncbi.nlm.nih.gov/gene

Gene: (ras) AND "Homo sapiens"[porgn:..txid9606]

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

Results: 1 to 20 of 1126

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

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KIRAS1, KRAS2, NS,

Gene: KRAS Kirsten rat sarcoma viral oncogene homolog [*Homo sapiens* (human)]
Gene ID: 3845, updated on 4-Jan-2015

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)

Summary

Official Symbol KRAS provided by HGNC
Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source HGNC:HGNC:6407
See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
Gene type protein coding
RefSeq status REVIEWED
Organism *Homo sapiens*
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

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

Genomic context

Summary

Official Symbol KRAS provided by HGNC
Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source HGNC:HGNC:6407
See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
Gene type protein coding
RefSeq status REVIEWED
Organism *Homo sapiens*
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Side-Note: Function, like beauty, is in the eye of the beholder...

Genomic context

Location: 12p12.1 See KRAS in [Epigenomics](#), [MapViewer](#)
Exon count: 6

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

Chromosome 12 - NC_000012.12

Map showing KRAS gene location on chromosome 12. Coordinates: 25952141 to 25436297. Neighboring genes: LRRP, CAS1C1, LVRRS, RPL39P27.

Genomic regions, transcripts, and products

Go to reference sequence details
 Genomic Sequence: NC_000012.12 chromosome 12 reference GRCh38 Primary Assembly
 Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)

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

General gene information

Summary

Official Symbol KRAS provided by HGNC
Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source HGNC:HGNC:6407
See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
Gene type protein coding
RefSeq status REVIEWED
Organism *Homo sapiens*
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

GO: Gene Ontology

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

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

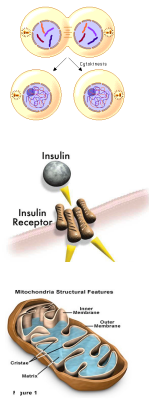
Process	Evidence Code	Pubs
Fc-epsilon receptor signaling pathway	TAS	
GTP catabolic process	IEA	
MAPK cascade	TAS	
Ras protein signal transduction	TAS	
actin cytoskeleton organization	IEA	
activation of MAPKK activity	TAS	
axon guidance	TAS	
blood coagulation	TAS	

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



www.ncbi.nlm.nih.gov/gene/3845#gene-ontology

Gene Ontology Provided by GOA

Function Evidence Code Pubs

GDP binding TAS

GMP binding IEA

GTP binding TAS

LRR domain binding IEA

protein binding TAS

protein complex binding TAS

Process Code Pubs

Fc-epsilon receptor signaling pathway TAS

GTP catabolic process IEA

MAPK cascade TAS

Ras protein signal transduction TAS

actin cytoskeleton organization IEA

activation of MAPKK activity TAS

axon guidance TAS

blood coagulation TAS

The 'Gene Ontology' or **GO** is actually maintained by the EBI so lets switch or link over to **UniProt** also from the EBI.

Scroll down to UniProt link

UniProt will detail much more information for protein coding genes such as this one

genomic X01669.1 CAA25828.1

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

Protein Accession Links

P01116.1 GenPept UniProtKB Link

UniProtKB/Swiss-Prot:P01116

Additional links

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

GETTING STARTED RESOURCES POPULAR FEATURED NCBI INFORMATION

NCBI Education Chemicals & Bioassays PubMed Genetic Testing Registry About NCBI

NCBI Help Manual Data & Software Bookshelf PubMed Health Research at NCBI

NCBI Handbook DNA & RNA PubMed Central GenBank NCBI News

Training & Tutorials Domains & Structures PubMed Health Reference Sequences NCBI FTP Site

Genes & Expression BLAST Gene Expression Omnibus Gene Expression Omnibus

Genetics & Medicine Nucleotide Map Viewer Human Genome NCBI on Facebook

Genomes & Maps Genome SNP Mouse Genome NCBI on Twitter

Literature Literature SNP Gene Influenza Virus Primer-BLAST NCBI on YouTube

Proteins Protein Protein BLAST Sequence Read Archive

Sequence Analysis PubChem

Taxonomy

Scroll down to UniProt link

UniProt will detail much more information for protein coding genes

www.uniprot.org/uniprot/P01116

UniProtKB

P01116 - RASK_HUMAN

Protein GTPase KRas

Gene KRAS

Organism Homo sapiens (Human)

Status Reviewed - Experimental evidence at protein level!

Display None BLAST Align Format Add to basket History Feedback Help video

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

www.uniprot.org/uniprot/P01116

UniProtKB

P01116 - RASK_HUMAN

Protein GTPase KRas

Gene KRAS

Organism Homo sapiens (Human)

Status Reviewed - Experimental evidence at protein level!

Display None BLAST Align Format Add to basket History Feedback Help video

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

```
>sp|P01116|RASK_HUMAN GTPase KRas OS=Homo sapiens GN=KRAS PE=1 SV=1
MEYKIAVVGAGGVGKSALTIQLQNIHFVDEYDFEEDSFRKQVYIDECCLLLEFAG
QEYSANRQDMYRTEGFLCFVAINNFKSFEDLHRYRQIKRVDKSEDPHVLGNKCDL
PSRTVDTKQAQLARSYGFPIETSANKTRQVEDAFYLVREIRQYRKKIKSEKKTFCG
VKIKKCIIM
```

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 level¹

Display None

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

Function¹
 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 regulation¹
 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 -			

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

P01116 - RASK_HUMAN
 Protein: GTPase KRas
 Gene: KRAS
 Organism: Homo sapiens (Human)
 Status: Reviewed - Experimental evidence at protein level¹

Display None

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

Function¹
 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 regulation¹
 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 -			

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

P01116 - RASK_HUMAN
 Protein: GTPase KRas
 Gene: KRAS
 Organism: Homo sapiens (Human)
 Status: Reviewed - Experimental evidence at protein level¹

Display None

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

Pathology & Biotech¹
Involvement in disease¹
 LEUKEMIA, ACUTE MYELOGENOUS (AML)
 [MIM:601626]: A subtype of acute leukemia, a cancer of the white blood cells. AML is a malignant disease of bone marrow characterized by maturational arrest of hematopoietic precursors at an early stage of development. Clonal expansion of myeloid blasts occurs in bone marrow, blood, and other tissue. Myelogenous leukemias develop from changes in cells that normally produce neutrophils, basophils, eosinophils and monocytes. #1 Publication -

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

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Natural variant ¹	10 - 10	1	G → GG in one individual with AML; expression in 3T3 cell causes cellular transformation; expression in CDS cells activates the Ras-MAPK signaling pathway; lower GTPase activity; faster GDP dissociation rate. #1 Publication -		VAR_034601	

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

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

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

P01116 - RASK_HUMAN
 Protein: GTPase KRas
 Gene: KRAS
 Organism: Homo sapiens (Human)
 Status: Reviewed - Experimental evidence at protein level¹

Display None

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

Structure¹
 Secondary structure
 Legend: Helix Turn Beta strand
 Show more details

3D structure databases

Select the link destinations:	Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
<input type="checkbox"/> PDBs ¹	108D	X-ray	2.00	P	178-188	[*]
<input checked="" type="checkbox"/> RCSB PDB ¹	10BE	X-ray	3.00	P	178-188	[*]
<input type="checkbox"/> PDBs ¹	1KZO	X-ray	2.20	C	169-173	[*]
	1KZP	X-ray	2.10	C	169-173	[*]
	3GFT	X-ray	2.27	A/B/C/D/E/F	1-164	[*]
	4DSN	X-ray	2.03	A	2-164	[*]
	4DSO	X-ray	1.85	A	2-164	[*]
	4EPR	X-ray	2.00	A	1-164	[*]
	4EPT	X-ray	2.00	A	1-164	[*]
	4EPV	X-ray	1.35	A	1-164	[*]
	4EPW	X-ray	1.70	A	1-1	
	4EPX	X-ray	1.76	A	1-1	
	4EPY	X-ray	1.80	A	1-1	
	4L8G	X-ray	1.52	A	1-1	
	4LDJ	X-ray	1.15	A	1-164	[*]
	4LPK	X-ray	1.50	A/B	1-169	[*]

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?

Lets view the 3D structure:

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

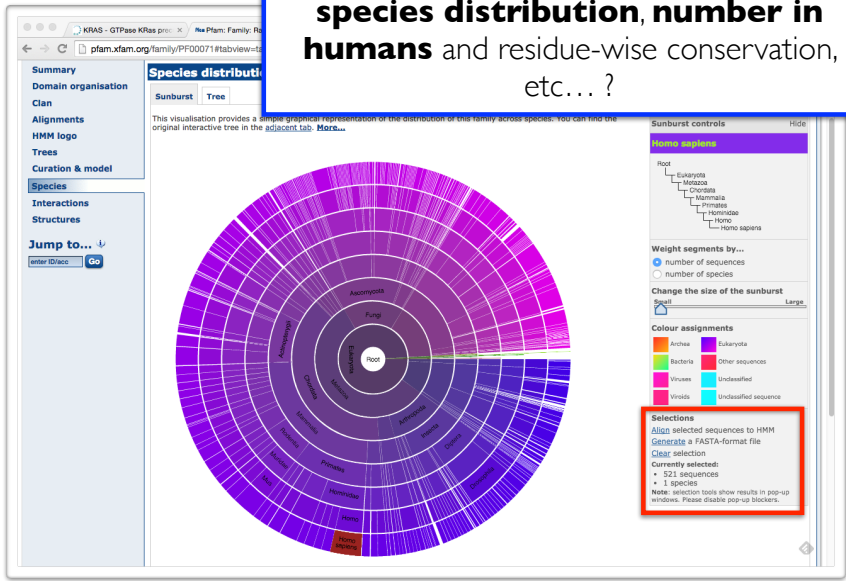
Back to UniProt:

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

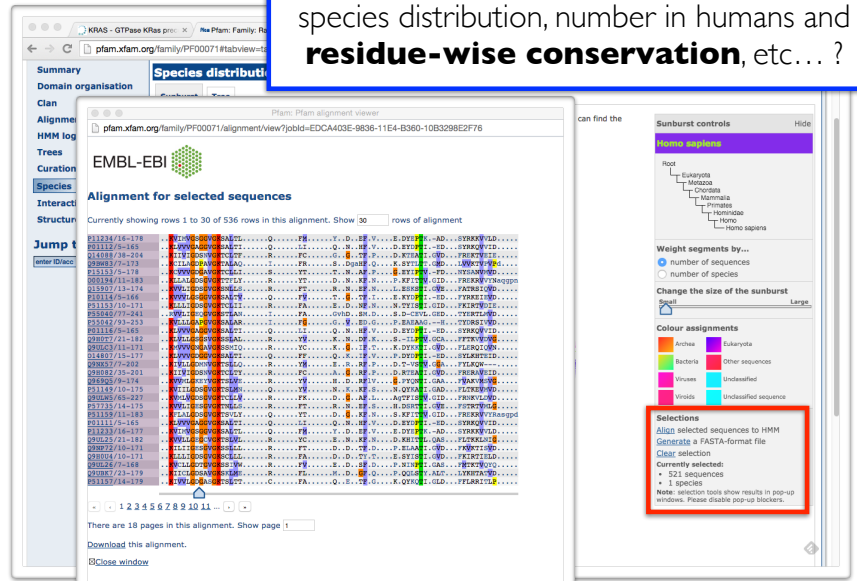
Example Questions:

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

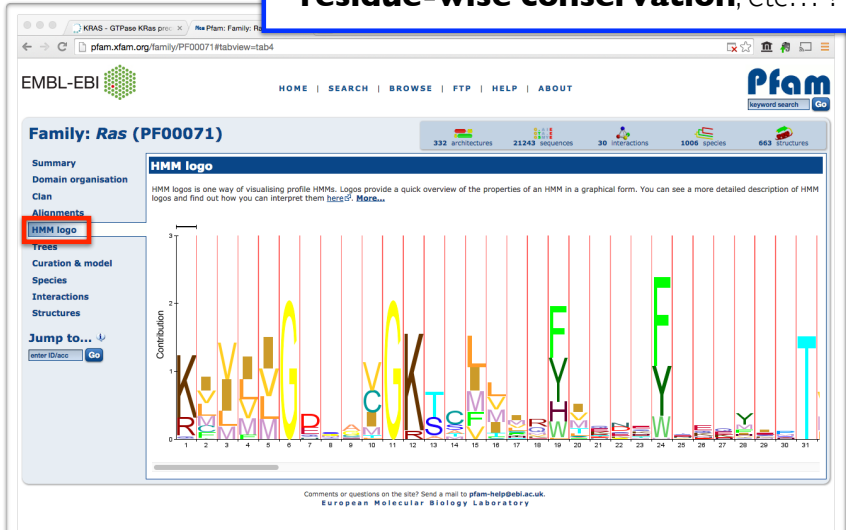
Example Questions:
 What is known about the protein family, its **species distribution, number in humans** and **residue-wise conservation**, etc... ?



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



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



Recap: Major NCBI and EBI databases

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

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 <i>why</i> and <i>how</i> of comparing sequences
Alignment Algorithms	Hands-on exploration of alignment algorithms and applications

ALIGNMENT FOUNDATIONS

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

ALIGNMENT FOUNDATIONS

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

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

Seq1: C A T T C A C

Seq2: C T C G C A G C

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

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

↑ mismatch
 ↑ match

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

↑ match
 ↑ mismatch
 ↑ gaps

Add gaps to increase number of matches

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

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

↑ match
 ↑ mismatch } mutation
 ↑ insertion } indels
 ↑ deletion } indels

Gaps represent 'indels'
 mismatch represent mutations

Why compare biological sequences?

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

Practical applications include...

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

Practical applications include...

- Similarity searching of databases
 - Protein structure prediction
- Assembly of sequence reads into a longer construct such as a bacterial genome
- 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!

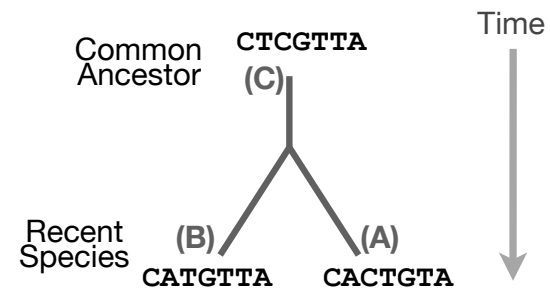
ALIGNMENT FOUNDATIONS

- **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

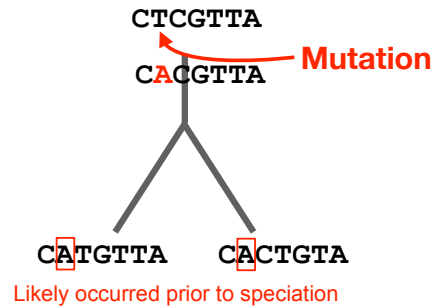


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA

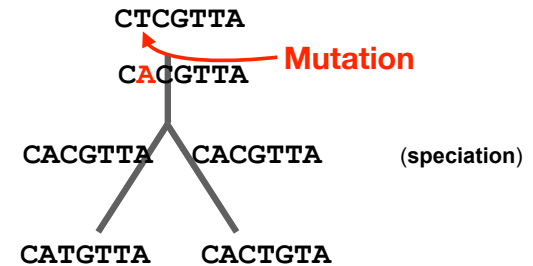


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA



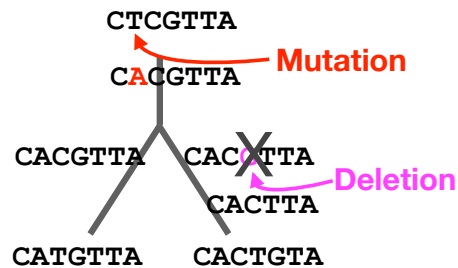
Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA

CACGTTA → CACTTA



Mutations, deletions and insertions

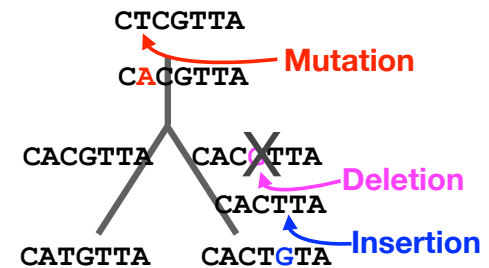
There are three major types of sequence change that can occur during evolution.

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA

CACGTTA → CACTTA

CACTTA → CACTGTA

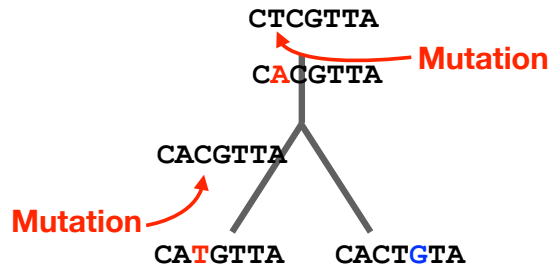


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

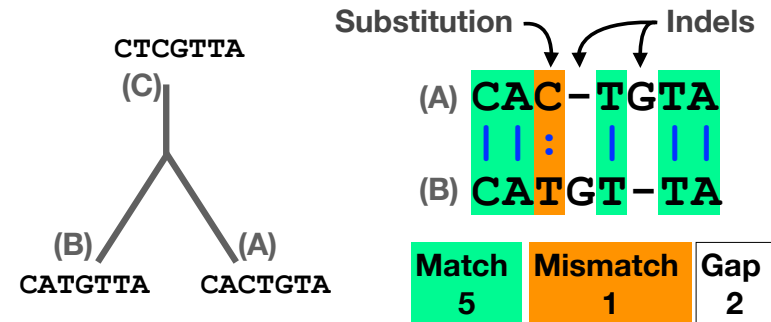
CTCGTTA → CACGTTA
CACGTTA → CATGTTA



Alignment view

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

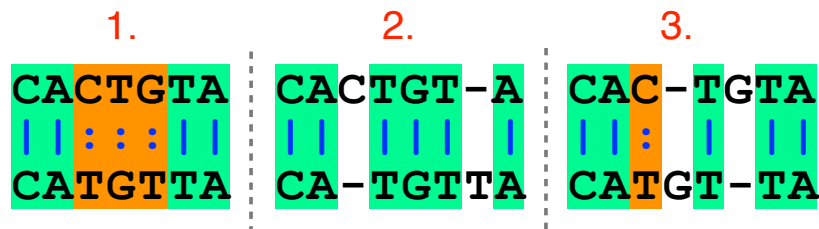
- Mismatches represent mutations/substitutions
- Gaps represent insertions and deletions (indels)



Alternative alignments

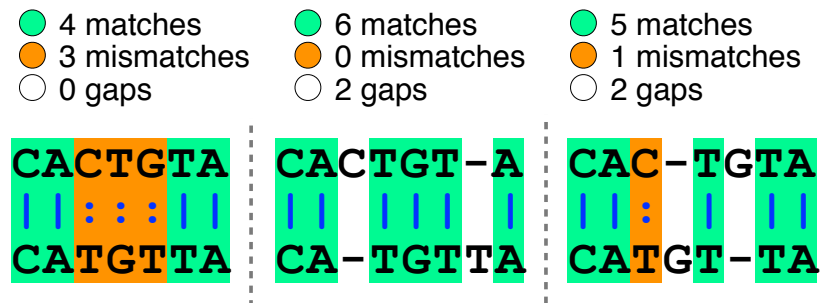
- Unfortunately, finding the correct alignment is difficult if we do not know the evolutionary history of the two sequences

Q. Which of these 3 possible alignments is best?



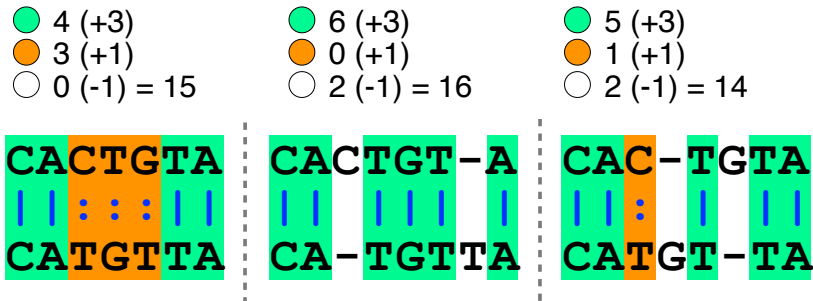
Alternative alignments

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



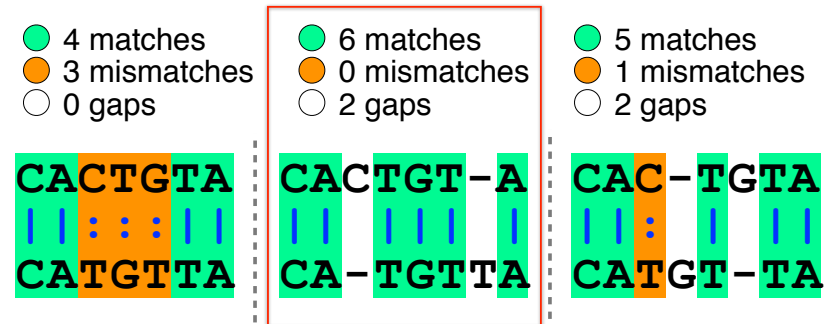
Scoring alignments

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



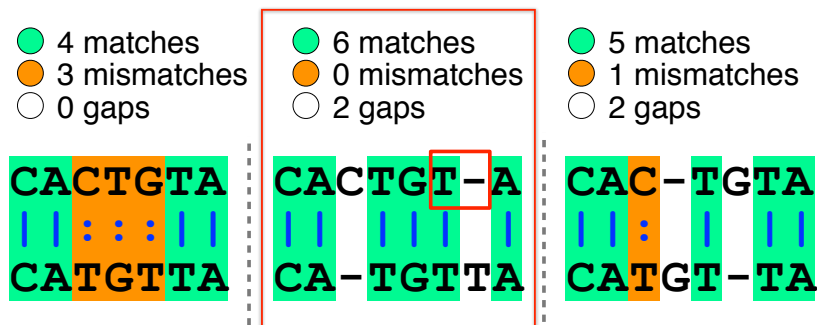
Optimal alignments

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



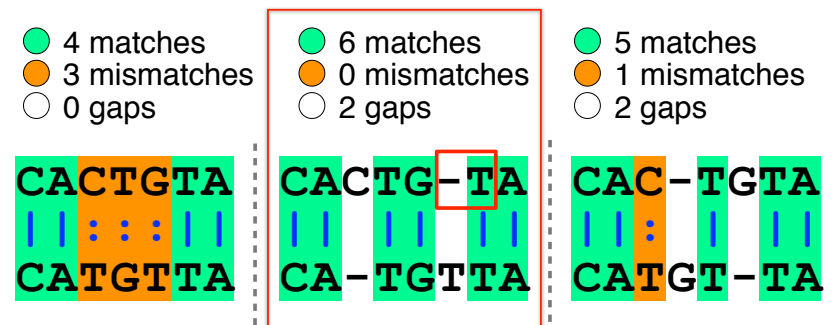
Optimal alignments

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



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 sequence changes is minimized

- 4 matches (green circle)
- 3 mismatches (orange circle)
- 2 gaps (white circle)
- 5 matches (yellow circle)
- 1 mismatch (red circle)

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

CACTGT-A
 CA-TGTTA
 CAC-TGTA
 CATGT-TA

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

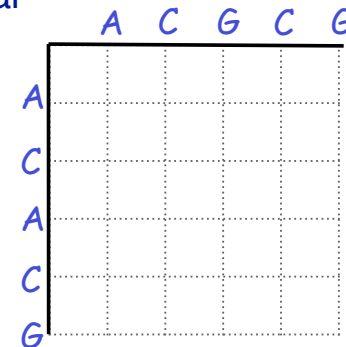
- **Why...**
 - Why compare biological sequences?
- **What...**
 - Alignment view of sequence changes during evolution (matches, mismatches and gaps)

- **How...**
 - ▶ Dot matrices
 - ▶ Dynamic programming
 - Global alignment
 - Local alignment
 - ▶ BLAST heuristic approach

How do we compute the optimal alignment between two sequences?

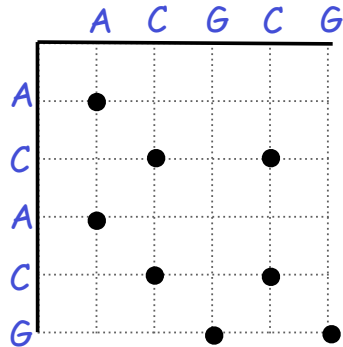
Dot plots: simple graphical approach

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



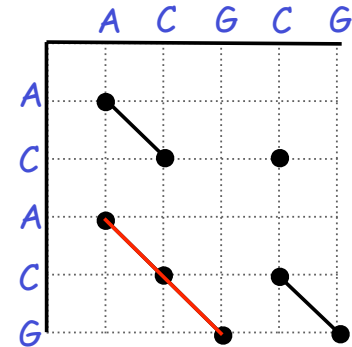
Dot plots: simple graphical approach

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



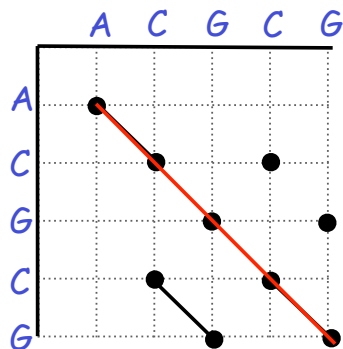
Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



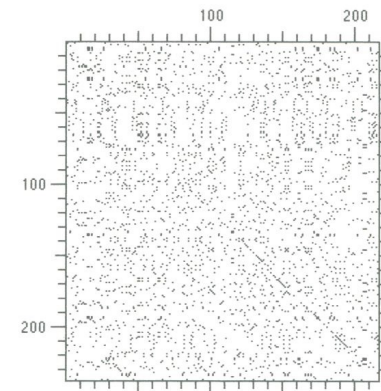
Dot plots: simple graphical approach

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



Dot plots: simple graphical approach

- Dot matrices for long sequences can be noisy

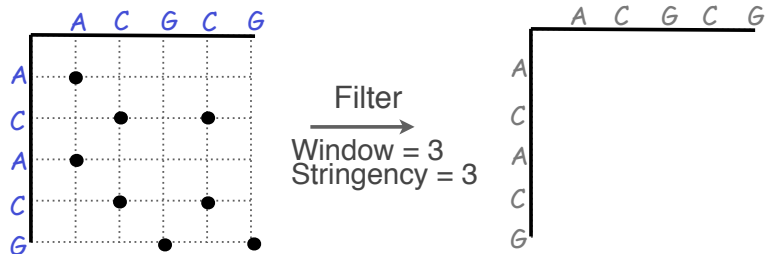


Dot plots: window size and match stringency

Solution: use a window and a threshold

- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.

- You have to choose window size and stringency

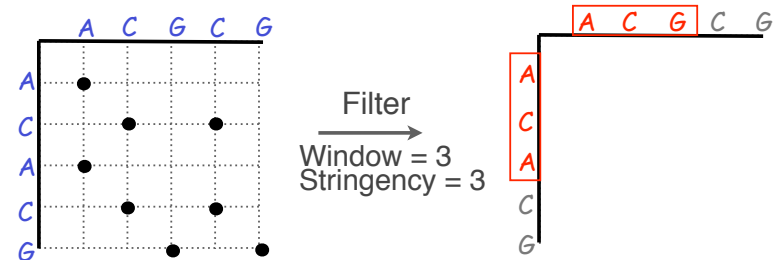


Dot plots: window size and match stringency

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

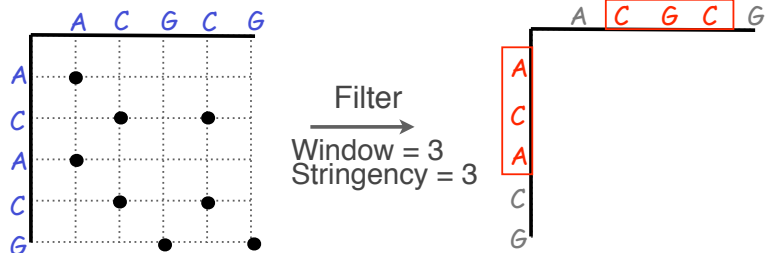


Dot plots: window size and match stringency

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

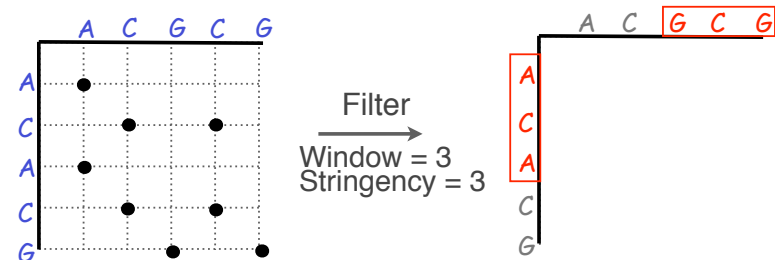


Dot plots: window size and match stringency

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

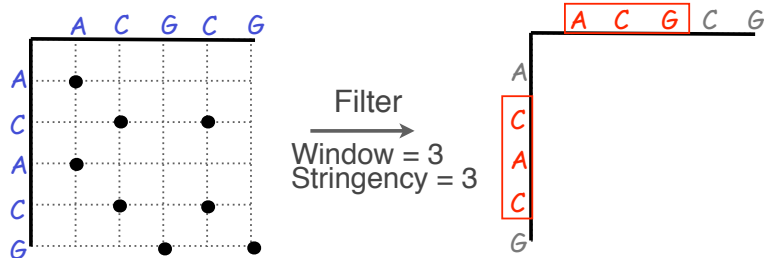


Dot plots: window size and match stringency

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

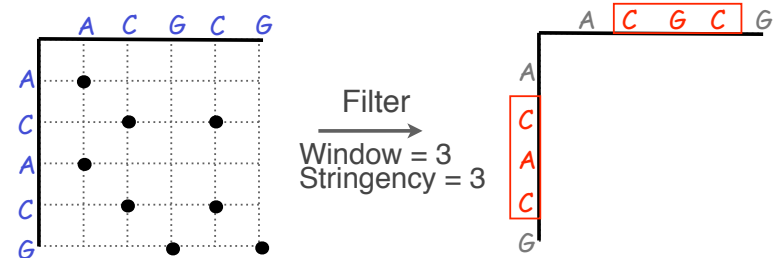


Dot plots: window size and match stringency

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

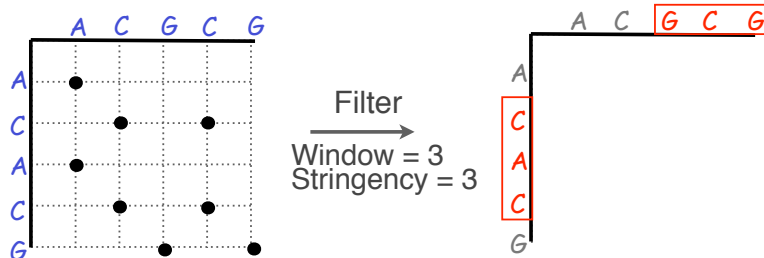


Dot plots: window size and match stringency

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

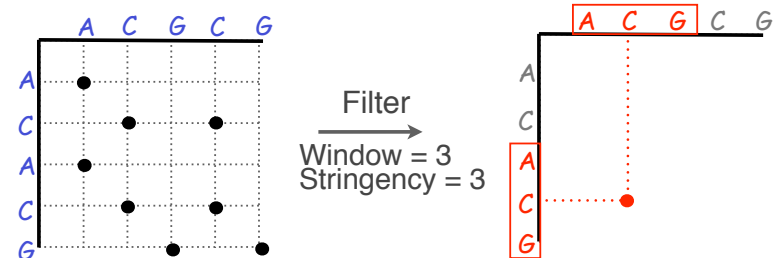


Dot plots: window size and match stringency

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

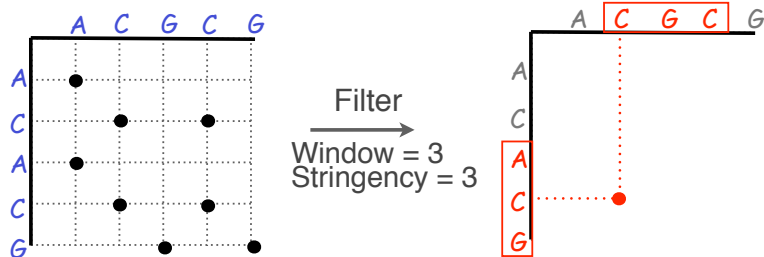


Dot plots: window size and match stringency

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

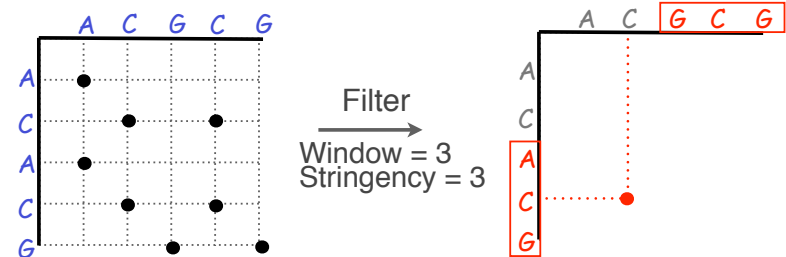


Dot plots: window size and match stringency

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

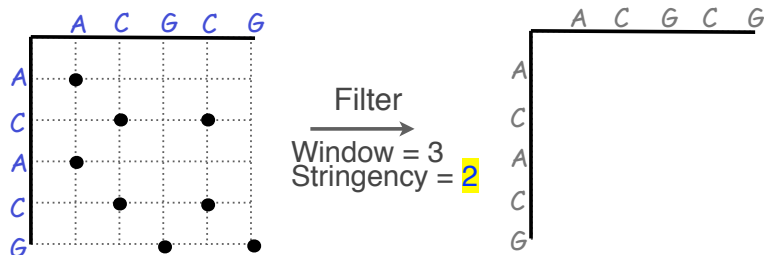


Dot plots: window size and match stringency

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

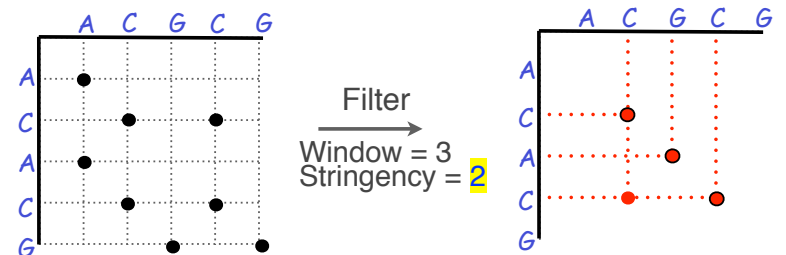


Dot plots: window size and match stringency

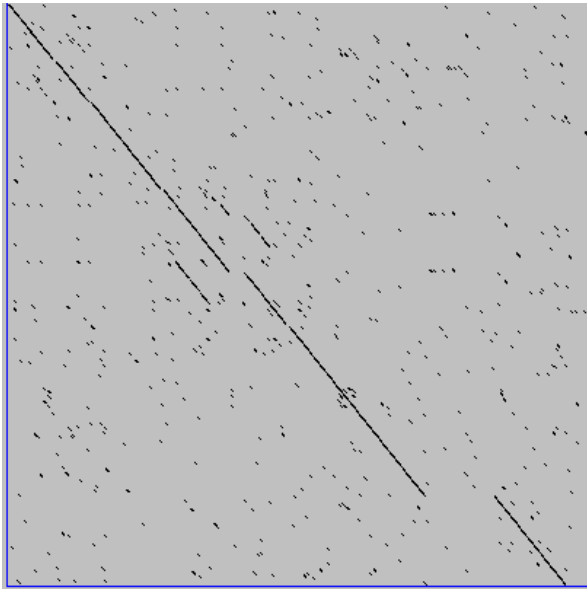
Solution: use a window and a threshold

- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.

- You have to choose window size and stringency



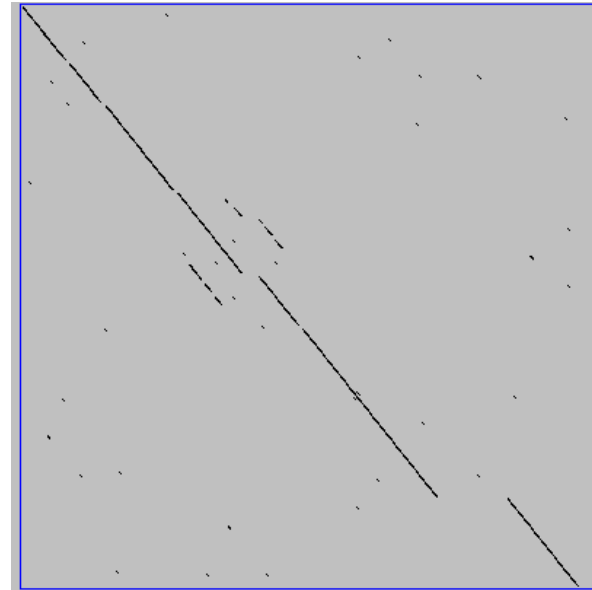
Window size = 5 bases



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

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

Window size = 7 bases



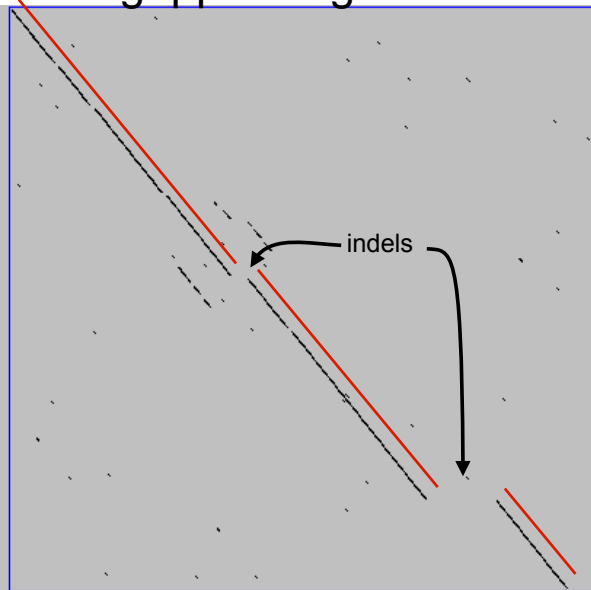
This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

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

Bigger window (kmer) fewer matches to consider

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

Ungapped alignments



Only **diagonals** can be followed.

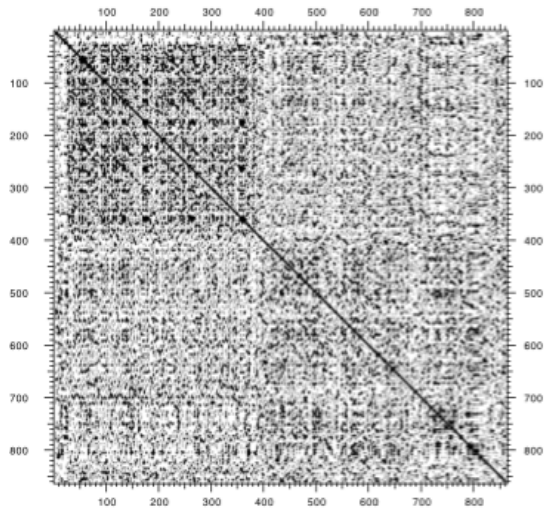
Downward or rightward paths represent **insertion** or **deletions** (gaps in one sequence or the other).

Uses for dot matrices

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

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

Repeats

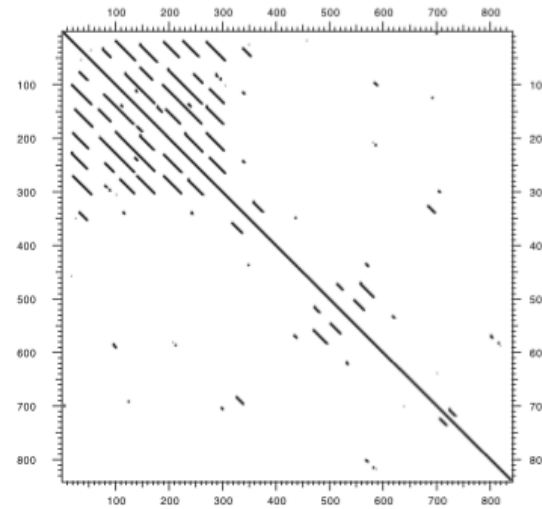


Human LDL receptor
protein sequence
(Genbank P01130)

$W = 1$
 $S = 1$

(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**)

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

BGGN-213: Dot Plot Comparison of Two Sequences

Dot plots are a simple graphical approach for the visual comparison of two sequences. They have a long history (see [Maizel and Lenk 1981](#) and references therein) and entail placing one sequence on the vertical axis of a 2D grid (or matrix) and the other on the horizontal. In its simplest form, a dot is placed where the horizontal and vertical sequence values match. That is a dot is produced at position (i, j) if character number i in the first sequence is the same as character number j in the second sequence. More elaborate forms use "sliding windows" composed of multiple characters and a threshold value, or "match stringency" for two windows to be considered as matched.

Dot Plot Parameters

After 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:

Moving window step size:

Match stringency:

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

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

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

Questions for discussion:

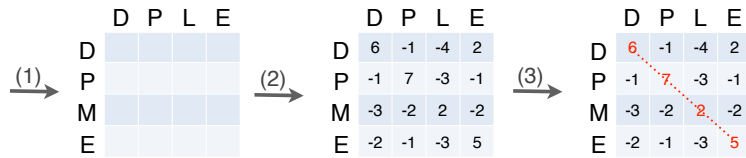
- Why does the DNA sequence have more dots than the protein sequence plot?
- How can we increase the signal to noise ratio?
- What does a "match stringency" parameter mean? "Match stringency" value and why?

ALIGNMENT FOUNDATIONS

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

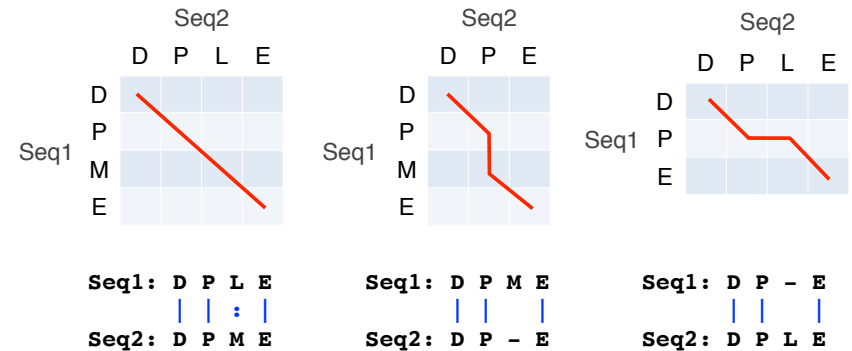
The Dynamic Programming Algorithm

- The dynamic programming algorithm can be thought of an extension to the dot plot approach
 - One sequence is placed down the side of a grid and another across the top
 - Instead of placing a dot in the grid, we compute a score for each position
 - Finding the optimal alignment corresponds to finding the path through the grid with the best possible score



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

Different paths represent different alignments



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

Algorithm of Needleman and Wunsch

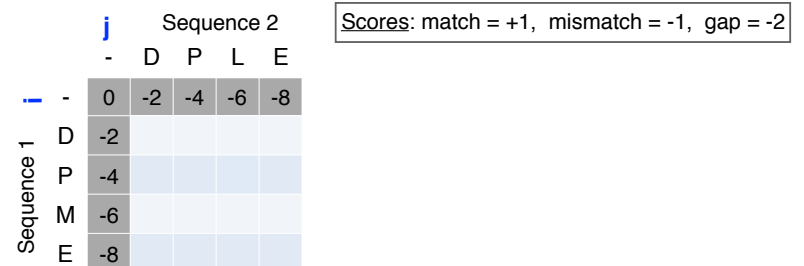
- The Needleman–Wunsch approach to global sequence alignment has three basic steps:
 - setting up a 2D-grid (or alignment matrix),
 - scoring the matrix, and
 - identifying the optimal path through the matrix



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.

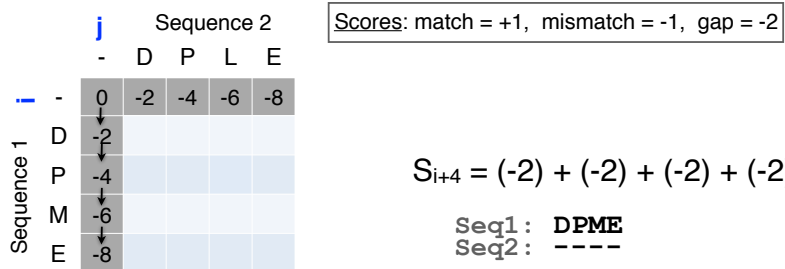
Scoring the alignment matrix

- 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



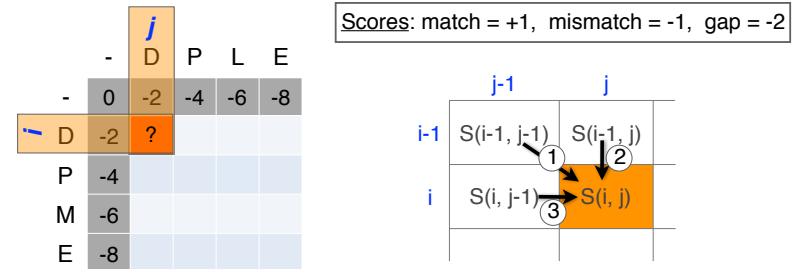
Scoring the alignment matrix

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



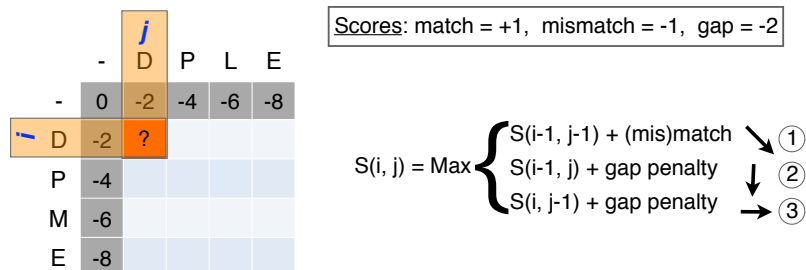
Scoring the alignment matrix

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



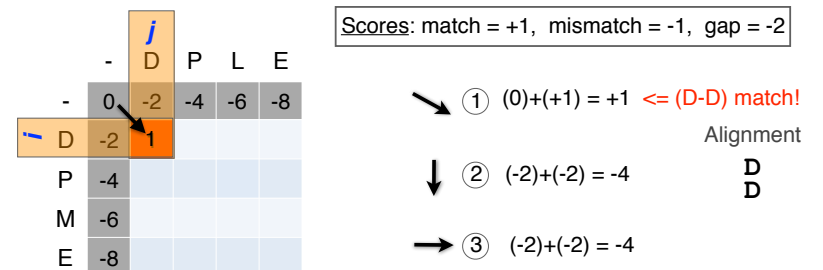
Scoring the alignment matrix

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



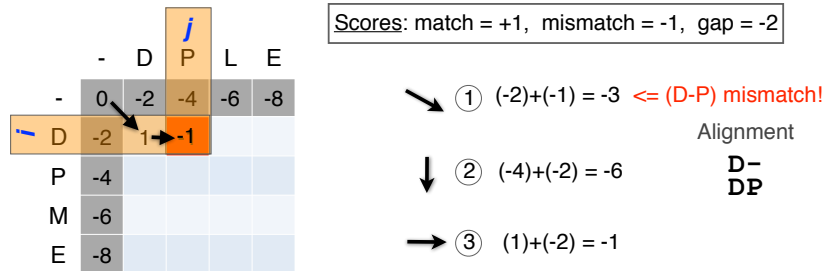
Scoring the alignment matrix

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



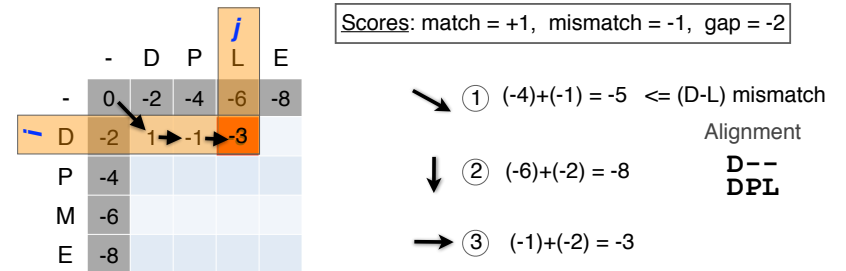
Scoring the alignment matrix

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



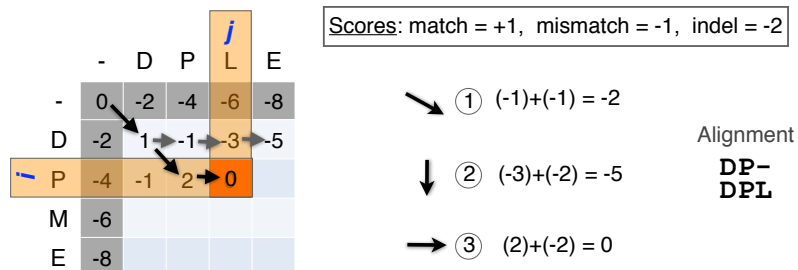
Scoring the alignment matrix

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



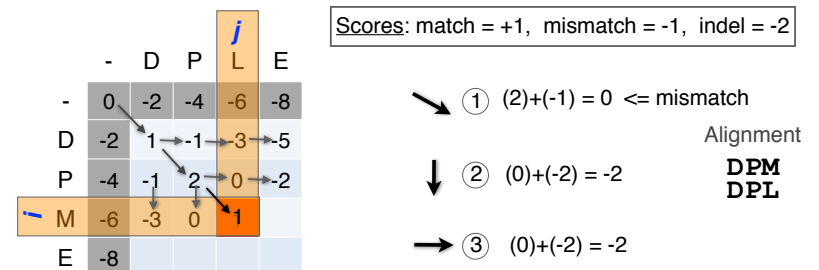
Scoring the alignment matrix

- For the highlighted cell, the corresponding score ($S_{i,j}$) refers to the score of the optimal alignment of the first i characters from sequence 1, and the first j characters from sequence 2.



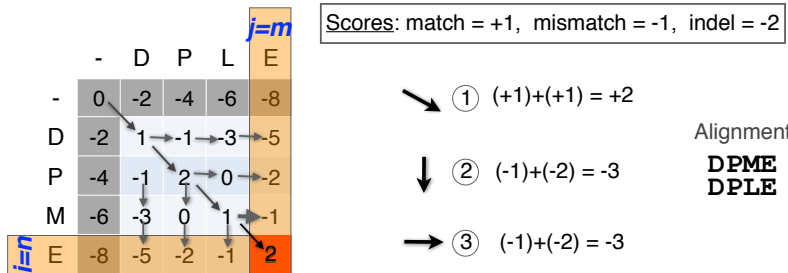
Scoring the alignment matrix

- 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



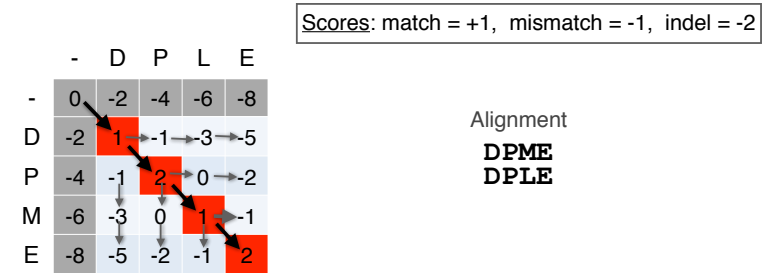
Scoring the alignment matrix

- 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)



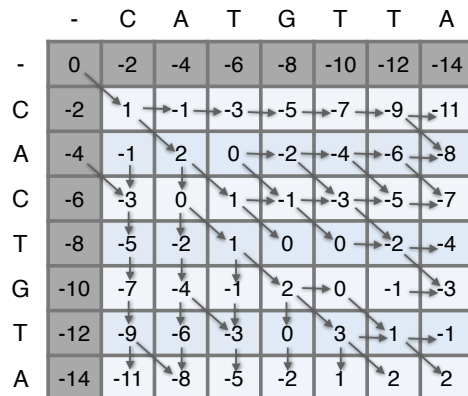
Scoring the alignment matrix

- 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



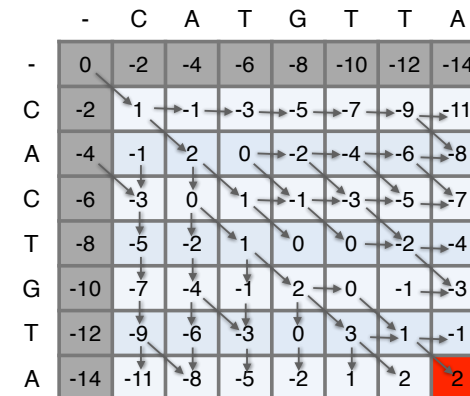
Questions:

- 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

	-	C	A	T	G	T	T	A
-	0	-2	-4	-6	-8	-10	-12	-14
C	-2	1	-1	-3	-5	-7	-9	-11
A	-4	-1	2	0	-2	-4	-6	-8
C	-6	-3	0	1	-1	-3	-5	-7
T	-8	-5	-2	1	0	0	-2	-4
G	-10	-7	-4	-1	2	0	-1	-3
T	-12	-9	-6	-3	0	3	1	-1
A	-14	-11	-8	-5	-2	1	2	2

More than one alignment possible

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

	-	C	A	T	G	T	T	A
-	0	-2	-4	-6	-8	-10	-12	-14
C	-2	1	-1	-3	-5	-7	-9	-11
A	-4	-1	2	0	-2	-4	-6	-8
C	-6	-3	0	1	-1	-3	-5	-7
T	-8	-5	-2	1	0	0	-2	-4
G	-10	-7	-4	-1	2	0	-1	-3
T	-12	-9	-6	-3	0	3	1	-1
A	-14	-11	-8	-5	-2	1	2	2

Alignment
CACTGT-A
CA-TGTTA
CACTG-TA
CA-TGTTA

The alignment and score are dependent on the scoring system

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

	-	C	A	T	G	T	T	A
-	0	-3	-6	-9	-12	-15	-18	-21
C	-3	1	-2	-5	-8	-11	-14	-17
A	-6	-2	2	-1	-4	-7	-10	-13
C	-9	-5	-1	1	-2	-5	-8	-11
T	-12	-8	-4	0	0	-1	-4	-7
G	-15	-11	-7	-3	1	-1	-2	-5
T	-18	-14	-10	-6	-2	2	0	-3
A	-21	-17	-13	-9	-5	-1	1	1

Alignment
CACTGT-A
CA-TGTTA
CACTG-TA
CA-TGTTA
CACTGTA
CATGTTA

The alignment and score are dependent on the scoring system

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

	-	C	A	T	G	T	T	A
-	0	-3	-6	-9	-12	-15	-18	-21
C	-3	1	-2	-5	-8	-11	-14	-17
A	-6	-2	2	-1	-4	-7	-10	-13
C	-9	-5	-1	1	-2	-5	-8	-11
T	-12	-8	-4	0	0	-1	-4	-7
G	-15	-11	-7	-3	1	-1	-2	-5
T	-18	-14	-10	-6	-2	2	0	-3
A	-21	-17	-13	-9	-5	-1	1	1

Alignment
CACTGT-A
CA-TGTTA
CACTG-TA
CA-TGTTA
CACTGTA
CATGTTA

Key point: Optimal alignment solutions and their scores are not necessarily unique and depend on the scoring system!

Your Turn!

Hands-on worksheet Sections 2 & 3

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

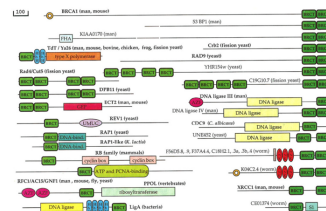
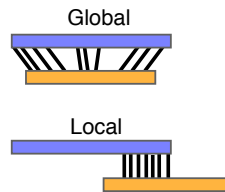
		A	G	T	T	C
	0					
A						
T						
T						
G						
C						

ALIGNMENT FOUNDATIONS

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

Global vs local alignments

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



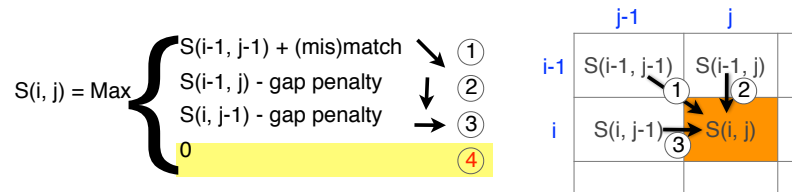
Local alignment: Definition

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

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

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

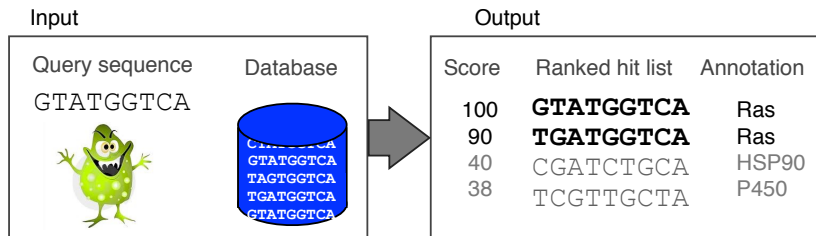


		Sequence 1														
		-	C	A	G	C	C	U	C	G	C	U	U	A	G	
Sequence 2	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
	A	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	A	0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
	U	0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7	1.0
	G	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0	1.0
	C	0.0	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3	0.3
	C	0.0	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0	0.0
	A	0.0	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0	0.0
	U	0.0	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0	1.0
	U	0.0	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0	1.0
	G	0.0	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3	2.7	2.7
	A	0.0	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0	2.0
	C	0.0	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0	2.0
G	0.0	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0	2.0	
G	0.0	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0	2.0	

Local alignment
GCC-UCG
GCCAUUG

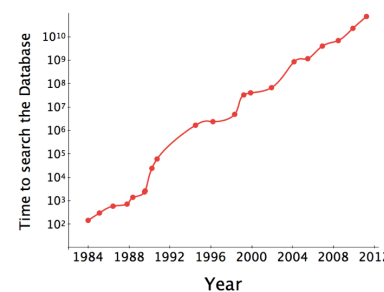
Local alignments can be used for database searching

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



The database search problem

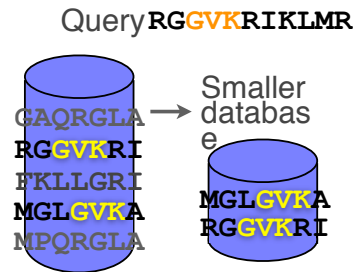
- Due to the rapid growth of sequence databases, search algorithms have to be both efficient and sensitive
 - Time to search with SW is proportional to $m \times n$ (m is length of query, n is length of database), too slow for large databases!



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

The database search problem

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



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

ALIGNMENT FOUNDATIONS

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

Rapid, heuristic versions of Smith–Waterman: BLAST

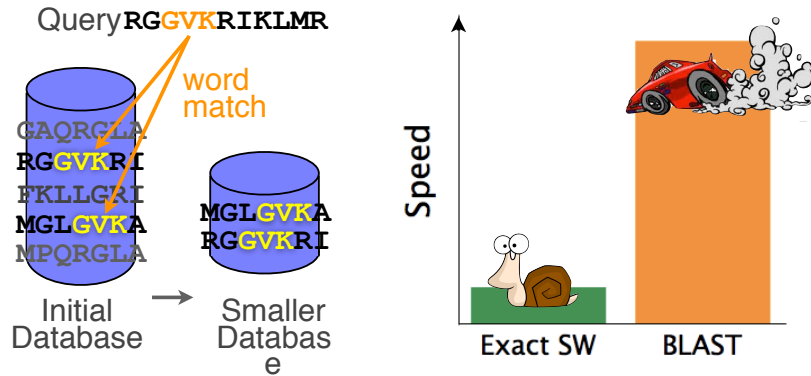
- BLAST (Basic Local Alignment Search Tool) is a simplified form of Smith-Waterman (SW) alignment that is popular because it is fast and easily accessible
 - BLAST is a heuristic approximation to SW - It examines only part of the search space
 - BLAST saves time by restricting the search by scanning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

Rapid, heuristic versions of Smith–Waterman: BLAST

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

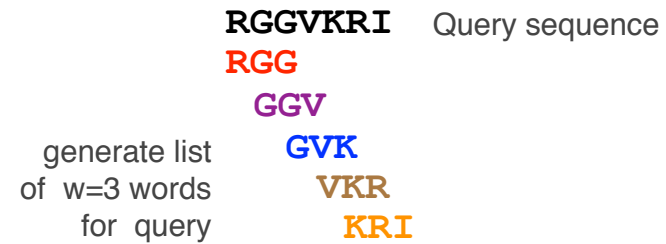
“The central idea of the BLAST algorithm is to confine attention to sequence pairs that contain an initial word pair match”
Altschul et al. (1990)

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



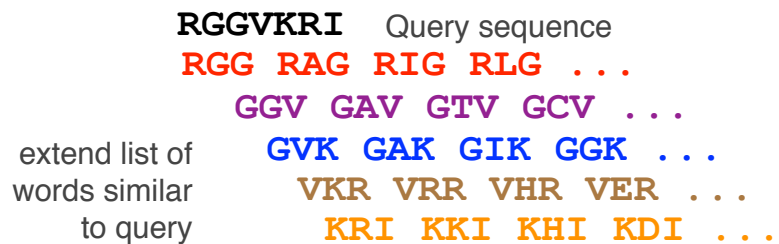
How BLAST works

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



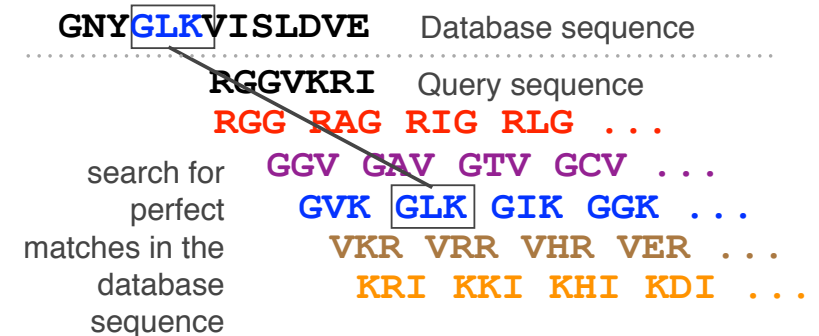
Blast

- 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)



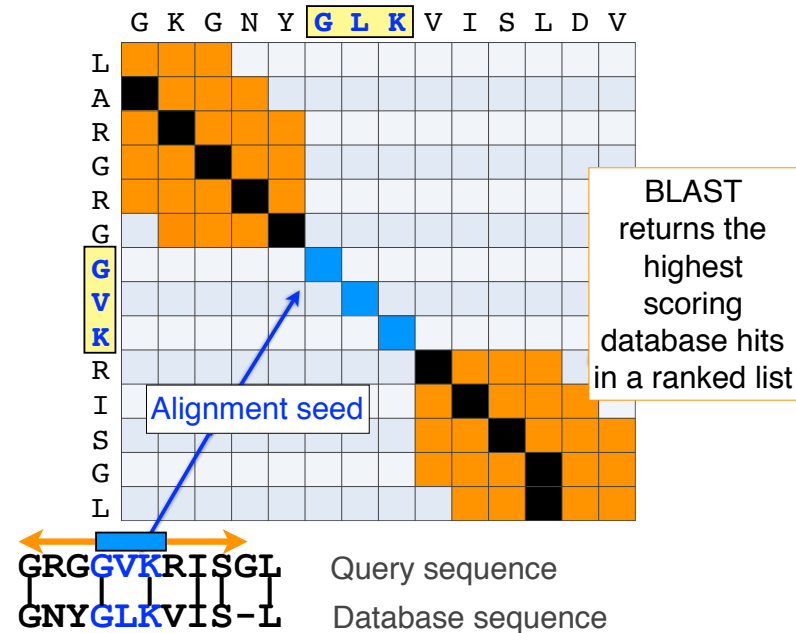
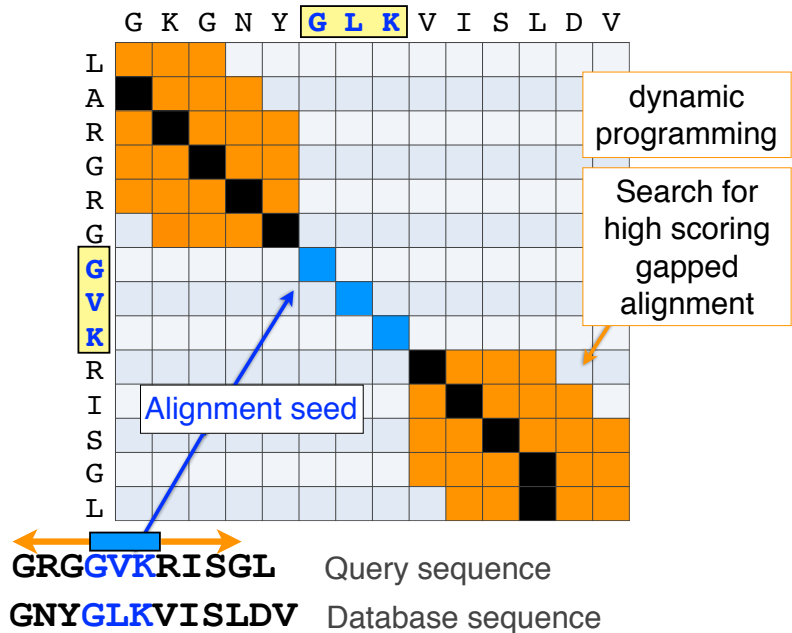
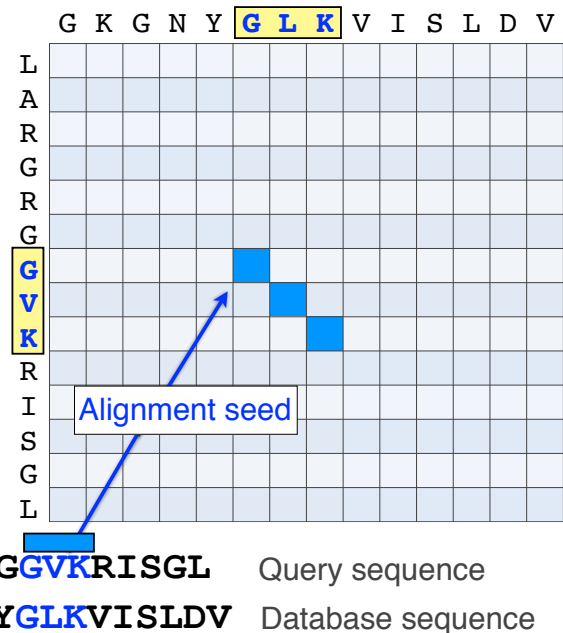
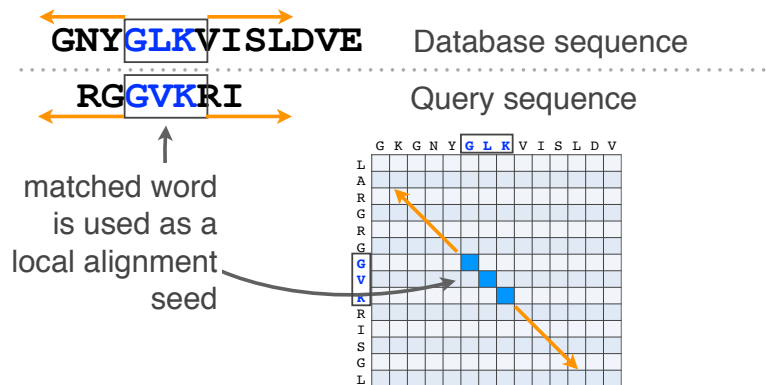
Blast

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



Blast

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



BLAST output

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

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

129

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

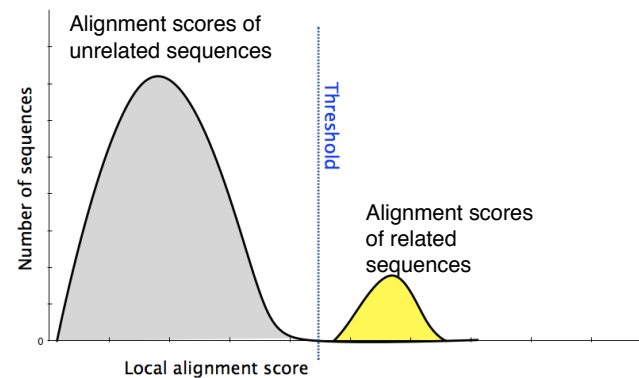
130

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

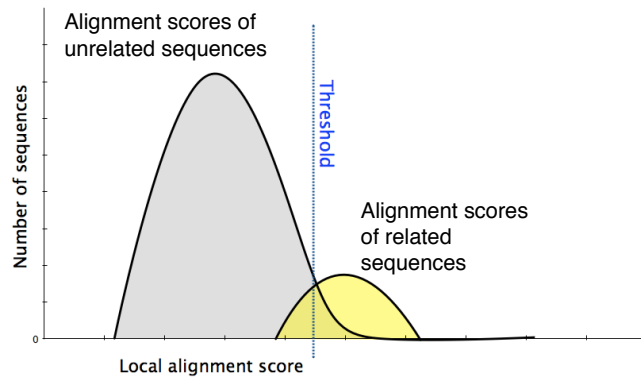
131

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



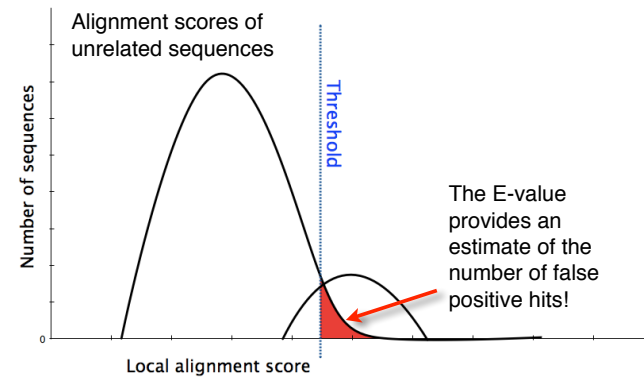
132

- 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



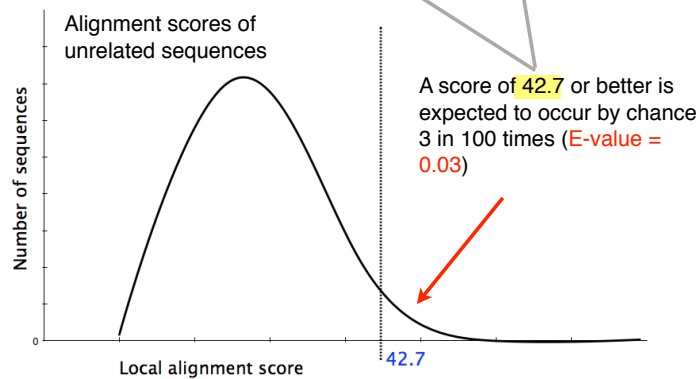
133

- 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



134

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



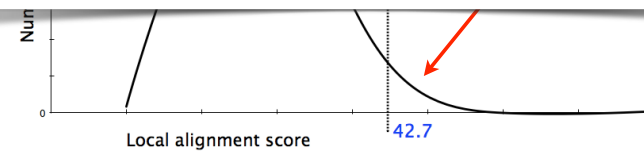
135

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

In general E values < 0.005 are usually significant.

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

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



136

FOR NEXT CLASS...

Check out the online:

- ✓ **Reading:** Sean Eddy's "What is dynamic programming?"
- ✓ **Feedback:** Muddy point assessment link online
- ✓ **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+

REFERENCE SLIDES...

Additional reference slides for the motivated student

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"

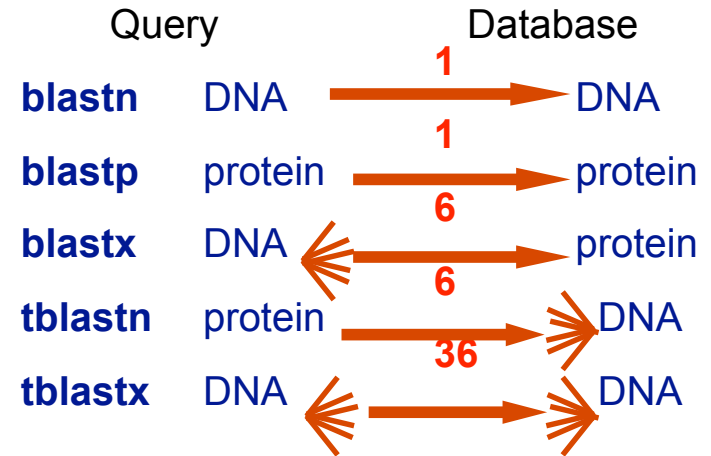
Step 1: Choose your sequence

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



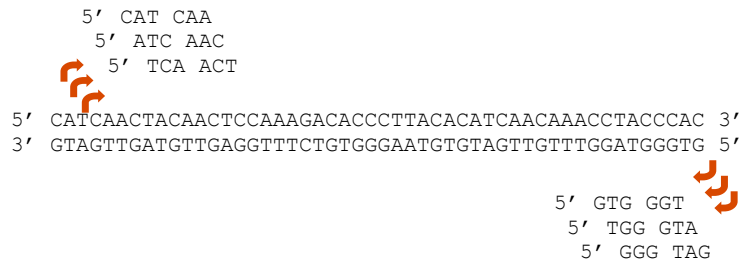
141

Step 2: Choose the BLAST program

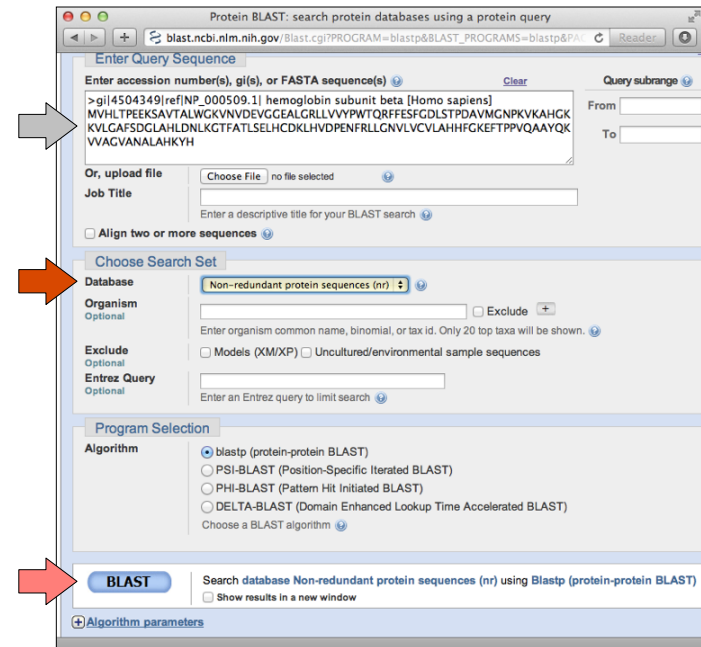


142

DNA potentially encodes six proteins



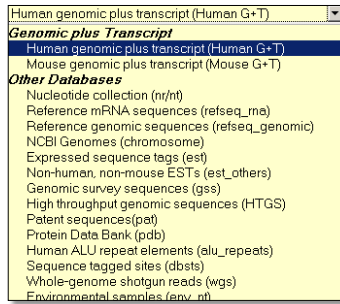
143



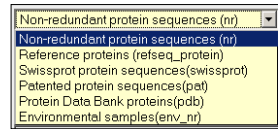
144

Step 3: Choose the database

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



nucleotide databases

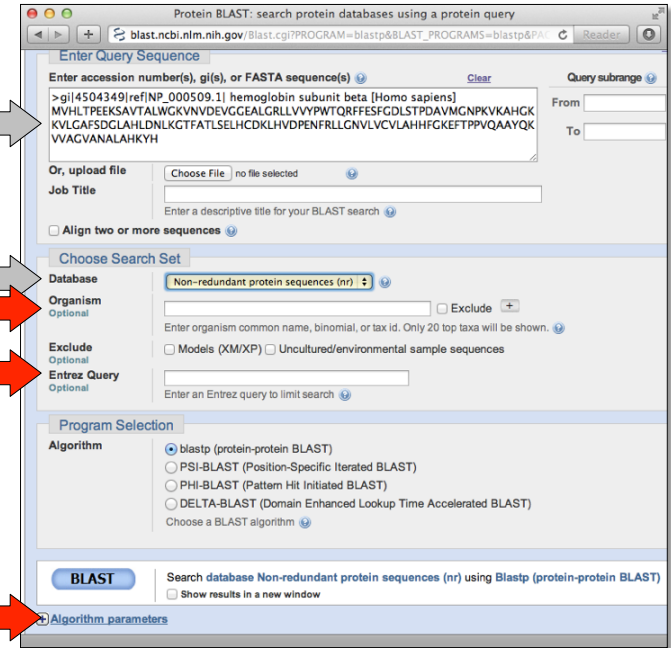


protein databases

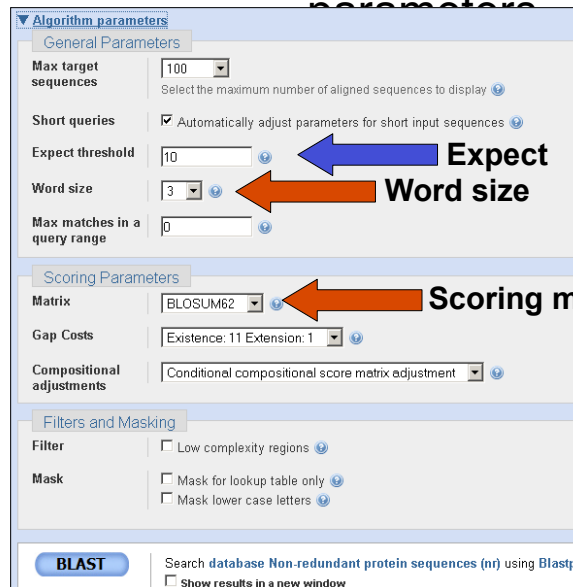
Organism

Entrez

Settings!



Step 4a: Select optional search parameters



Expect

Word size

Scoring matrix

Step 4: Optional parameters

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

Results page

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

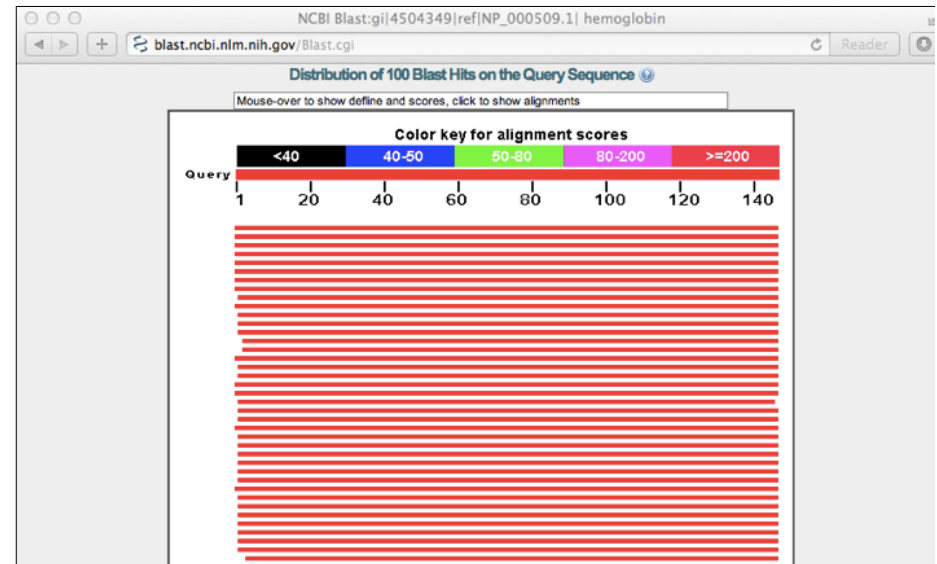
Query ID: gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
 Description: beta [Homo sapiens]
 Molecule type: amino acid
 Query Length: 147

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

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

Specific hits: globin
 Superfamilies: globin_like superfamily

Further down the results page...

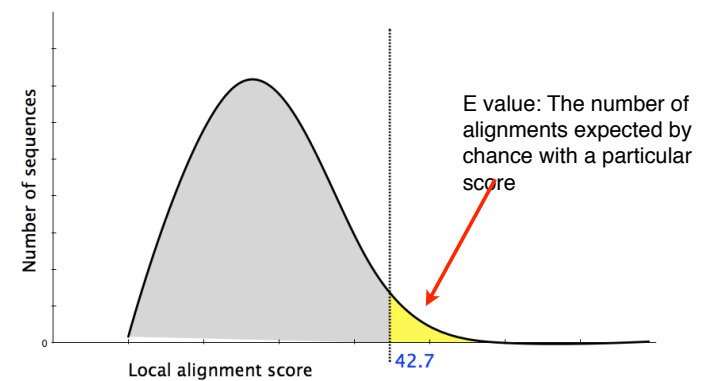


Further down the results page...

Sequences producing significant alignments:

Description	Max score	Total score	Query cover	E value	Max ident	Accession
hemoglobin beta [synthetic construct]	301	301	100%	9e-103	100%	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=Her	300	300	100%	4e-102	99%	P02024.2
beta globin chain variant [Homo sapiens]	299	299	100%	5e-102	99%	AA84548.1
beta globin [Homo sapiens] >gb AAZ39781.1 beta globin [Homo sapiens] >gb AAZ3978	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%	1CQH_B
hemoglobin beta subunit variant [Homo sapiens] >gb AAA88054.1 beta-globin [Homo sa	298	298	100%	1e-101	99%	AAF00469.1
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

Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo	677	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	676	100%	0	98%	AAA20133.1
Kinesin-14 heavy chain [Danio rerio]	595	595	88%	0	78%	XP_00320703
hypothetical protein EGK_18589	48.2	52	40%	0.03	32%	ELK35081.1
mKIAA4102 protein [Mus musculus]	42.7	42.7	38%	3.02	24%	EHH28205.1



E values in BLAST

- Each alignment gets a score determined from the alignment and doesn't take into account the full length of the query, target or database
- The E value is what you want to look at
- E value = Expect
 - How often do I expect an alignment with this score give the length of my query and the size of the database
 - $E = Kmne^{-\lambda S}$
 - K and λ are scaling factors
 - S is the score
 - m – length of query, n – length of database
 - E corrects for multiple comparisons, i.e., query compared to many sequences – proportional to length of database and query for a given S (score)

Further down the results page...

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

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

Download GenPept Graphics

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

Query 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60
MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK
Sbjct 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

Query 61 VKAHGKVLGAFSDGLAHLNLDKGTFTLSELHCCKLHVDPENFRLLGNLVLCVLAHHPG 120
VKAHGKVLGAFSDGLAHLNLDKGTFTLSELHCCKLHVDPENFRLLGNLVLCVLAHHPG
Sbjct 61 VKAHGKVLGAFSDGLAHLNLDKGTFTLSELHCCKLHVDPENFRLLGNLVLCVLAHHPG 120

Query 121 KEFTPPVQAAYQKVVAGVANALAHRYH 147
KEFTPPVQAAYQKVVAGVANALAHRYH
Sbjct 121 KEFTPPVQAAYQKVVAGVANALAHRYH 147

Download GenPept Graphics

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

Range 1: 1 to 147	Expect	Method	Identities	Positives	Gaps
Score	4e-102	Compositional matrix adjust.	146/147(99%)	147/147(100%)	0/147(0%)

Download GenPept Graphics

Related Information

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

Different output formats are available

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

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

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp suite/ Formatting Results - FVGUTMR2013

Edit and Resubmit Save Search Strategies **Formatting options** Download

Change the result display back

You Tube Learn about the enhanced report

Formatting options

Show Alignment as HTML Old View Reset form to defaults

Alignment View Query-anchored with letters for identities

Display Graphical Overview Sequence Retrieval NCBI-gi

Masking Character: Lower Case Color: Grey

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

Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.

Enter organism name or id-completions will be suggested Exclude

Entrez query:

Expect Min: Expect Max:

Percent Identity Min: Percent Identity Max:

Format for PSI-BLAST with inclusion threshold:

gi|4504349|ref|NP_000509.1| hemoglobin

E.g. Query anchored alignments

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

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

Query 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

AAK37051 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

AAK29557 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

NP_000509 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

P02024 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

AAK84548 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

AAZ39780 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

ACU56984 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

AAI19696 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

ICOH_B 1 VHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

AAF00489 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

2YRS_B 1 VHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

1DXU_B 1 MHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

1HDB_B 1 VHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

1DKV_B 2 HLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

3RMC_C 2 HLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

AAI68978 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

1NOP_B 1 VHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

1K1K_B 1 VHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

AAI11320 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

KP_002822173 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

1Y85_B 1 VHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

1YE0_B 1 MHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

1O10_B 1 MHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

CAA23759 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 60

1YE2_B 1 MHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

1Y5F_B 1 MHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

1A00_B 1 MHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

1HBS_B 1 VHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

1ABY_B 1 MHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

1CHY_B 1 VHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDVAMGNPK 59

... and alignments with dots for identities

Accession	Score	Identity
Query	1	MVHLTPEEKSAVTALMGKVNVDEGVEALGRLLVVYPWTQRFFESFGDLSTPDAMVGNPK 60
AAX37051	1 60
AAX29557	1 60
NP_000509	1 60
P02024	1 60
AAN84548	1 60
AAZ39780	1K..... 60
ACU56984	1K..... 60
AAD19696	1L..... 60
ICOH_B	1 59
AAF00489	1 60
ZYRS_B	1 59
IDXU_B	1	M..... 59
IHDB_B	1 59
IDXV_B	2 59
3KMF_C	2 59
AAL68978	1K..... 60
INOP_B	1K..... 59
IKIK_B	1V..... 60
AAN11320	1 60
XP_002822173	1 60
IY85_B	1 59
IYE0_B	1	M.....A..... 59
IOIO_B	1	M..... 59
CAA23759	1V.....X..... 60
IYE2_B	1	M.....F..... 59
IY5F_B	1	M..... 59
IA00_B	1	M.....Y..... 59

Common problems

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

158

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

159

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)

160