

BIMM 143
Introduction to Bioinformatics
 Lecture 2
 Barry Grant
 UC San Diego
<http://thegrantlab.org/bimm143>

Recap From Last Time:

- Bioinformatics is computer aided biology.
 - Deals with the collection, archiving, organization, and interpretation of a wide range of biological data.
- 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](#))
- There are a large number of bioinformatics databases (see [handout!](#)).
- 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/>

Hands on demo (or see following slides)

Example Vignette Questions:

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

Search NCBI databases

ras

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

Literature		Genes	
Books	1,677	EST	3,985
MeSH	402	Gene	57,165
NLM Catalog	223	GEO DataSets	3,732
PubMed	54,672	GEO Profiles	1,622,789
PubMed Central	96,114	HomoloGene	696
Health		PopSet	2,254
ClinVar	759	UniGene	4,770
dbGaP	120	Proteins	
GTR	1,879		

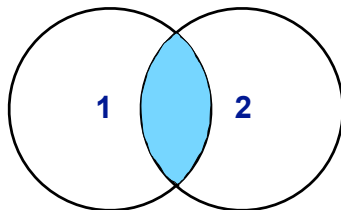
Gene search results for 'ras'. The search bar contains 'ras'. The results show 1 to 20 of 85633 items. The top organisms list includes Homo sapiens (1126), Mus musculus (823), Rattus norvegicus (625), Oreochromis niloticus (533), and Neolamprologus brichardi (507). The table below shows the first two results:

Name/Gen ID	Description	Location	Aliases
ras ID: 15412	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,

Gene search results for '(ras) AND "Homo sapiens"[porgn:..txid9606]'. The search bar contains the query. The results show 1 to 20 of 1126 items. The table below shows the first two results:

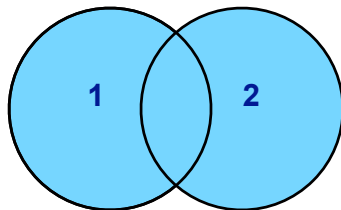
Name/Gen ID	Description	Location	Aliases
NRAS ID: 4693	neuroblastoma RAS viral (v-ras) oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3645	Kirsten rat sarcoma viral oncogene homolog [<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, NS3, RAS2

1 AND 2



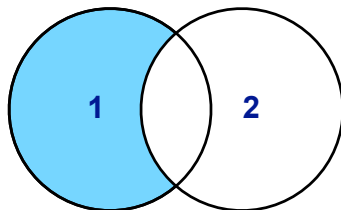
ras AND disease (1185 results)

1 OR 2



ras OR disease (134,872 results)

1 NOT 2



ras NOT disease (84,448 results)

Gene search results for '(ras) AND "Homo sapiens"[porgn:..txid9606]'. The search bar contains the query. The results show 1 to 20 of 1126 items. The table below shows the first two results, with KRAS highlighted:

Name/Gen ID	Description	Location	Aliases
NRAS ID: 4693	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: 3645	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, NS3, RAS2

NCBI Resources How To Sign in to NCBI

Gene Search Help

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

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

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

Summary

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

Table of contents

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

13

NCBI Resources How To Sign in to NCBI

Gene Search Help

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

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

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

Summary

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

Table of contents

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

14

NCBI Resources How To Sign in to NCBI

Gene 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 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

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](#)

15

NCBI Resources How To Sign in to NCBI

Gene Search Help

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

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

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

Summary

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

Table of contents

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

16

www.ebi.ac.uk/ontology/gene-ontology

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

Items 1 - 25 of 33 < Prev Page 1 of 2 Next >

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

GO: Gene Ontology

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

EMBL-EBI UniProt-GOA

Services Research Training About us

Search

Examples: GO:006915, topotecan, P08727

Overview New to UniProt-GOA FAQ Contact Us

Gene Ontology Annotation (UniProt-GOA) Database

The UniProt GO annotation program aims to provide high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB). The assignment of GO terms to UniProt records is an integral part of UniProt biocuration. UniProt manual and electronic GO annotations are supplemented with manual annotations supplied by external collaborating GO Consortium groups, to ensure a comprehensive GO annotation dataset is supplied to users.

UniProt is a member of the [GO Consortium](#).

Menu

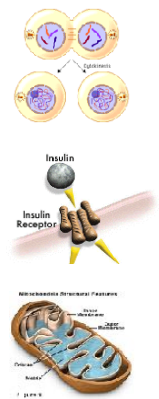
- Downloads
- Searching UniProt-GOA
- Annotation Methods
 - Annotation Tutorial
 - Manual Annotation Efforts
 - Reference Genome Annotation Initiative
 - Cardiovascular Gene Ontology Annotation Initiative
 - Renal Gene Ontology Annotation Initiative
 - Flybase Gene

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



Gene Ontology Provided by GOA

Function

- GDP binding
- GMP binding
- GTP binding
- LRR domain binding
- protein binding
- protein complex binding

Process

- Fe-epsilon receptor signaling pathway
- GTP catabolic process
- MAPK cascade
- Ras protein signal transduction
- actin cytoskeleton organization
- activation of MAPKK activity
- axon guidance
- blood coagulation

Evidence Code

- TAS
- IEA
- TAS
- TAS
- IEA
- TAS
- TAS
- TAS

Pubs

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

CAA25828.1

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

Protein Accession

P01116.1

Links

- GenPept Link
- UniProtKB Link
- GenPept
- UniProtKB/Swiss-Prot:P01116

Additional links

You are here: NCBI > Genes & Expression > Gene

GETTING STARTED

- NCBI Education
- NCBI Help Manual
- NCBI Handbook
- Training & Tutorials

RESOURCES

- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Hematology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy

POPULAR

- PubMed
- Books&f
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

FEATURED

- Genetic Testing Registry
- PubMed Health
- GenBank
- Reference Sequences
- Gene Expression Omnibus
- Map Viewer
- Human Genome
- Mouse Genome
- Influenza Virus
- Primer-BLAST
- Sequence Read Archive

NCBI INFORMATION

- About NCBI
- Research at NCBI
- NCBI News
- NCBI FTP Site
- NCBI on Facebook
- NCBI on Twitter
- NCBI on YouTube

Scroll down to UniProt link

UniProt will detail much more information for protein coding genes

UniProtKB

P01116 - RASK_HUMAN

Protein: GTPase KRas

Gene: KRAS

Organism: Homo sapiens (Human)

Status: Reviewed - Experimental evidence at protein level!

Display: None

Function

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

Enzyme regulation

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

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding ¹	10 - 18	9	GTP #2 Publications -			
Nucleotide binding ¹	29 - 35	7	GTP #2 Publications -			
Nucleotide binding ¹	59 - 60	2	GTP #2 Publications -			

UniProt will detail much more information for protein coding genes

UniProtKB

P01116 - RASK_HUMAN

Protein: GTPase KRas

Gene: KRAS

Organism: Homo sapiens (Human)

Status: Reviewed - Experimental evidence at protein level!

Display: None

View FASTA file format

```
>sp|P01116|RASK_HUMAN|GTPase_KRas|OS=Homo sapiens|GN=KRAS|PE=1|SV=1|
MFEKRIYFVGGDVGKSAITLGLIQHFFVDEYDFPEIDSRFRKVVYDSDCLDLELFDG
QREYRANRDQVMATPCDFLGVFAINNKRFSDIHHVVRQLKRVKDESDVNNVLQVKKDL
PSRTVDTVTKQADLARSYGFPIETSATKTRQVVEDAFYTLVREIRGYRLKIKISKEETFGC
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 level!

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

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

Regions

Feature key	Position(s)	Length	Description	Graphical view	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 level!

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

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

Regions

Feature key	Position(s)	Length	Description	Graphical view	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?

Pathology & Biotech
Involvement in disease¹
 LEUKEMIA, ACUTE MYELOID (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 -

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 COS cells activates the Ras-RAPK 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 canted 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?

Structure
 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"/> PDBa ¹	1D8D	X-ray	2.00	P	178-188	[*]
<input checked="" type="checkbox"/> RCSB PDB ²	1D8E	X-ray	3.00	P	178-188	[*]
<input type="checkbox"/> PDBj ³	1K2O	X-ray	2.20	C	169-173	[*]
	1K2P	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	[*]
	4DSQ	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	1-164	[*]
	4EPX	X-ray	1.76	A	1-164	[*]
	4EPY	X-ray	1.80	A	1-164	[*]
	4L8G	X-ray	1.52	A	1-164	[*]
	4LDJ	X-ray	1.15	A	1-164	[*]
	4LPK	X-ray	1.50	A/B	1-169	[*]

Open link in a new tab!

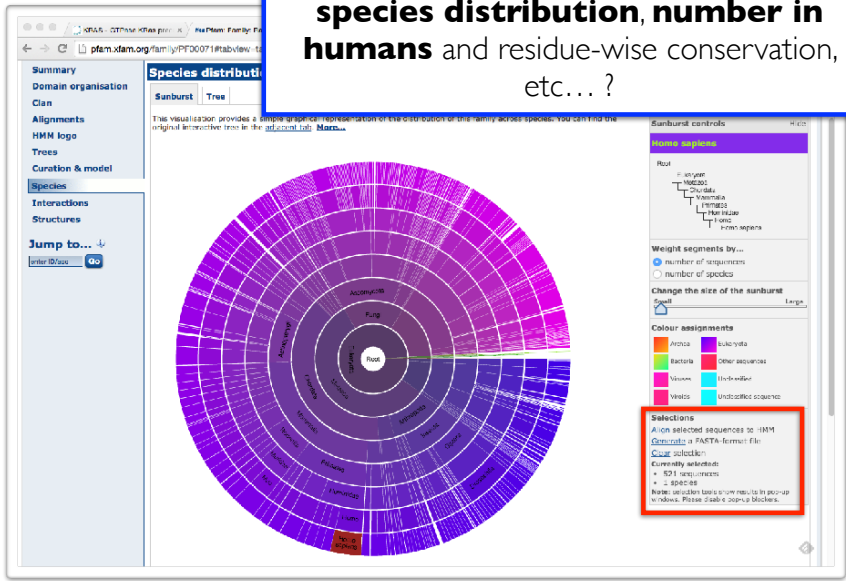
Lets view the 3D structure:
Can we find where in the structure our mutations are located and infer their potential molecular effects?

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

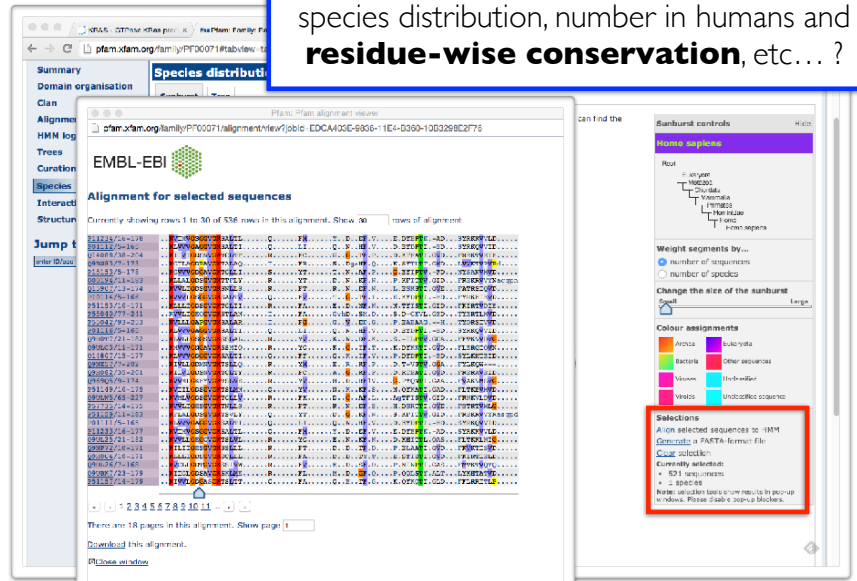
Back to UniProt:
What is known about the protein family, its species distribution, number in humans and residue-wise conservation, etc... ?

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

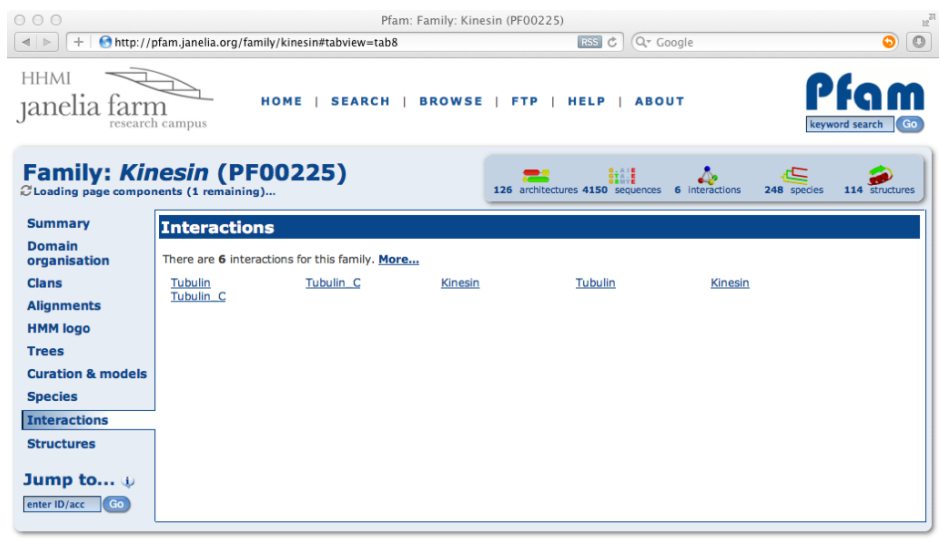
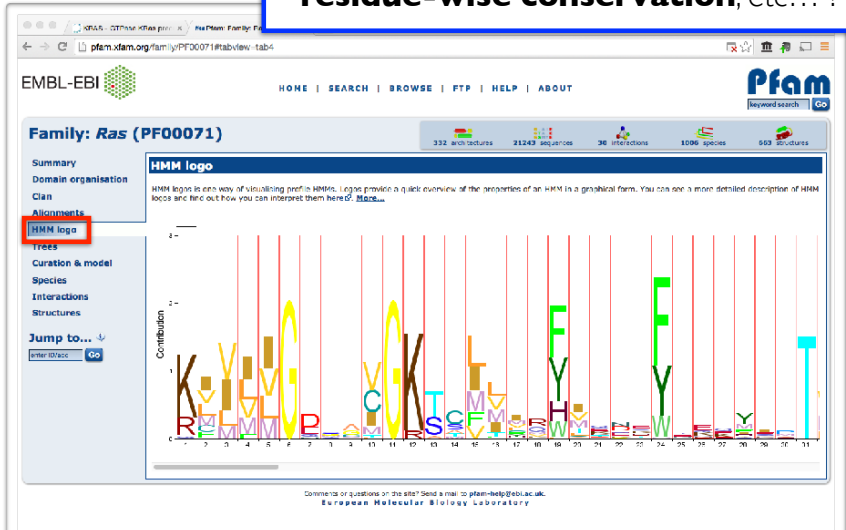
Example Questions:
 What is known about the protein family, its **species distribution, number in humans** and **residue-wise conservation**, etc... ?



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... ?



Questions or comments: pfam@janelia.hmi.org
 Howard Hughes Medical Institute

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)

126 architectures 4150 sequences 6 interactions 248 species 114 structures

Summary

Domain organisation

Clans

Alignments

HMM logo

Trees

Curation & models

Species

Interactions

Structures

Jump to...

enter ID/acc Go

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
ABBKD1_GIALA	11 - 335	2vva	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
			A	392 - 723	Jmol AstexViewer SPICE
			A	392 - 723	Jmol AstexViewer SPICE
			A	392 - 723	Jmol AstexViewer SPICE
			B	392 - 723	Jmol AstexViewer SPICE
			A	392 - 723	Jmol AstexViewer SPICE
KI13B_HUMAN	11 - 352	3qbj	A	11 - 352	Jmol AstexViewer SPICE
			B	11 - 352	Jmol AstexViewer SPICE
			C	11 - 352	Jmol AstexViewer SPICE
			A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
			A	24 - 359	Jmol AstexViewer SPICE
1g0b	24 - 359	1g0b	B	24 - 359	Jmol AstexViewer SPICE
			A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
			A	24 - 359	Jmol AstexViewer SPICE
1x88	24 - 359	1x88	B	24 - 359	Jmol AstexViewer SPICE
			A	24 - 359	Jmol AstexViewer SPICE

PFam: Family: Kinesin (PF00225)

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

wellcome 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	Kinesin (.PF00225)		
A	49	368	KIF22_HUMAN	49	368			

[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

↑ 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

↑ mismatch
↑ match

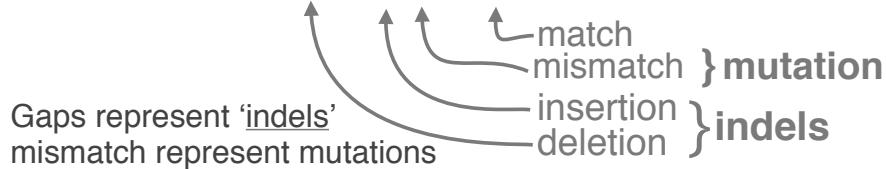
Add gaps to increase number of matches

gaps

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

Seq1: C A T - T C A - C

Seq2: C - T C G C A G C



Why compare biological sequences?

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

Practical applications include...

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

Practical applications include...

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

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

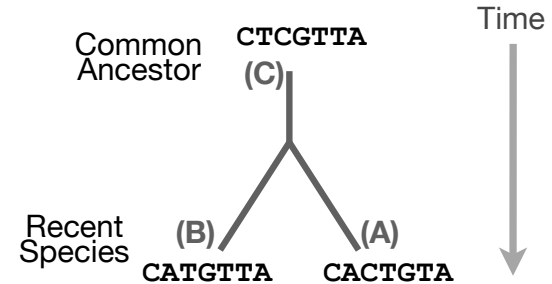
ALIGNMENT FOUNDATIONS

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

Sequence changes during evolution

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

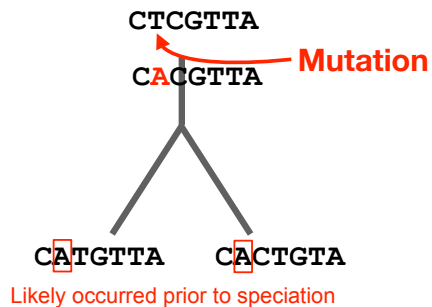
- Mutations/Substitutions
- Deletions
- Insertions



Mutations, deletions and insertions

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

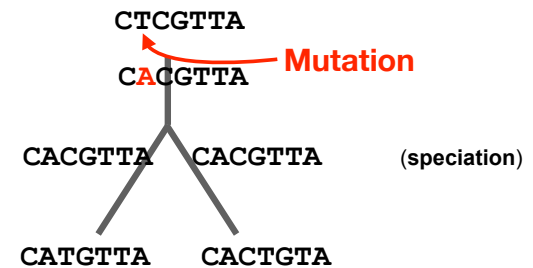
- **Mutations/Substitutions** CTCGTTA → CACGTTA
- Deletions
- Insertions



Mutations, deletions and insertions

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

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

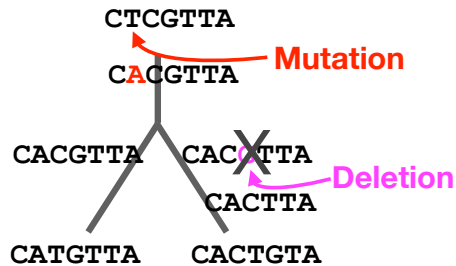


Mutations, deletions and insertions

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

- Mutations/Substitutions
- **Deletions**
- Insertions

CTCGTTA → CACGTTA
CACGTTA → CACTTA

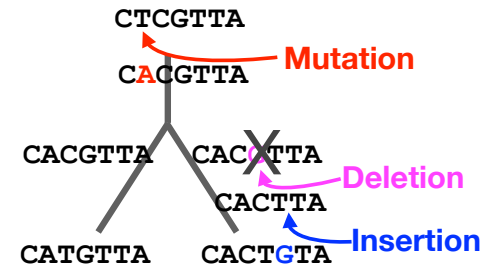


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- **Insertions**

CTCGTTA → CACGTTA
CACGTTA → CACTTA
CACTTA → CACTGTA

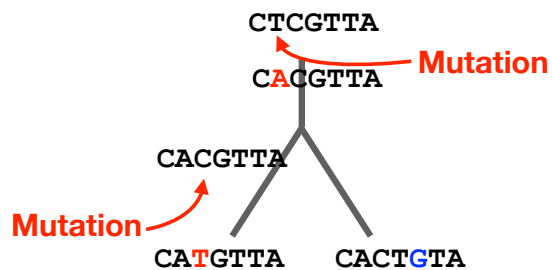


Mutations, deletions and insertions

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

- **Mutations/Substitutions**
- Deletions
- Insertions

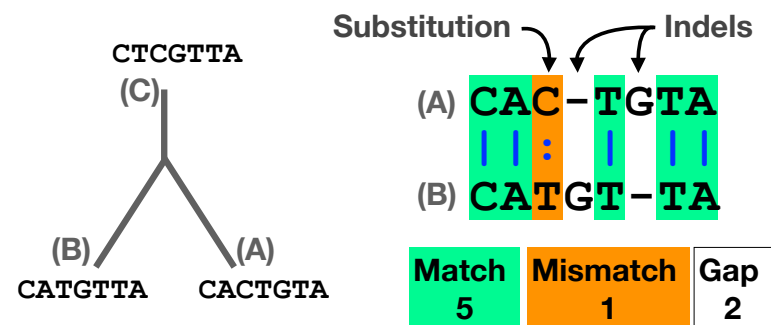
CTCGTTA → CACGTTA
CACGTTA → CATGTTA



Alignment view

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

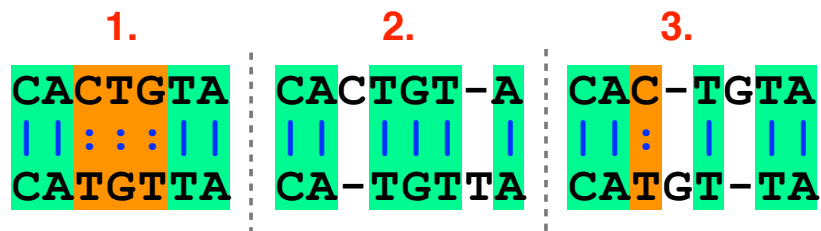
- **Mismatches** represent mutations/substitutions
- **Gaps** represent insertions and deletions (indels)



Alternative alignments

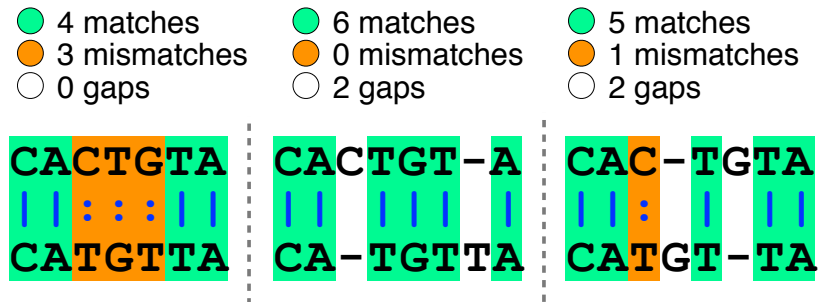
- Unfortunately, finding the correct alignment is difficult if we do not know the evolutionary history of the two sequences

Q. Which of these 3 possible alignments is best?



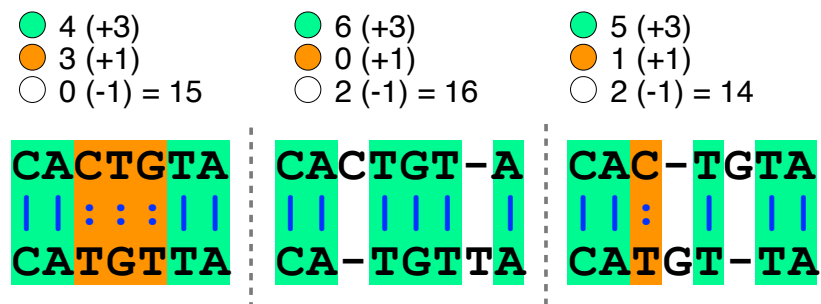
Alternative alignments

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



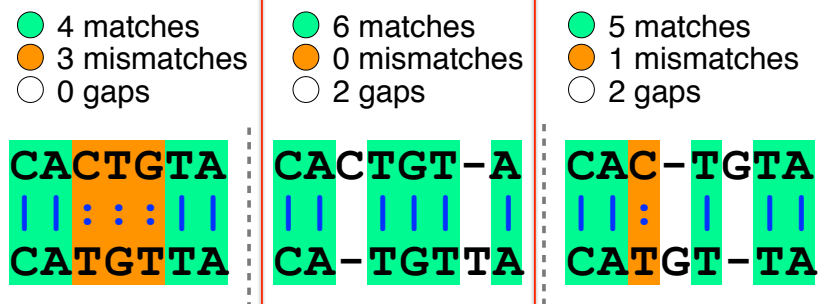
Scoring alignments

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



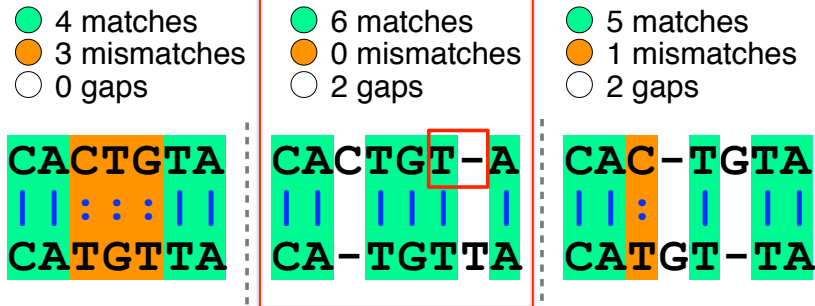
Optimal alignments

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



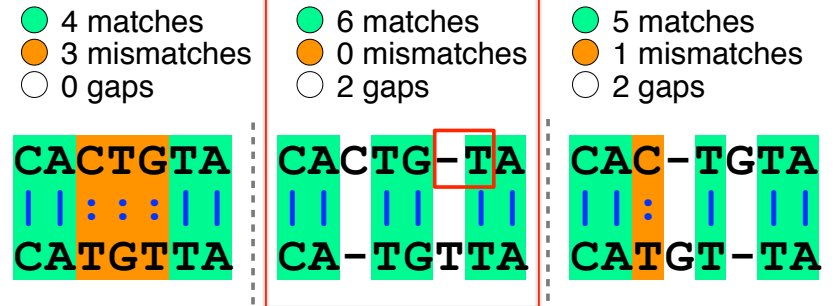
Optimal alignments

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



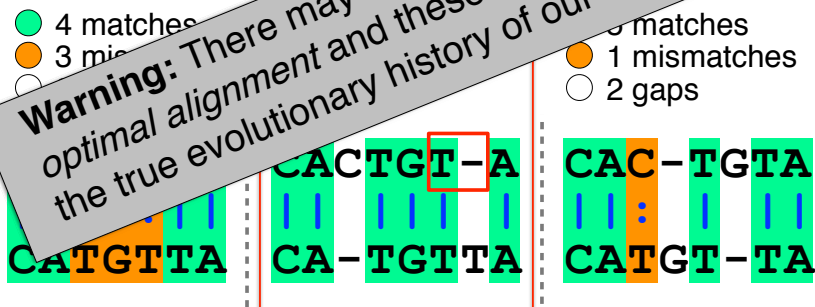
Optimal alignments

- 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 postulated sequence changes is minimized.



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

ALIGNMENT FOUNDATIONS

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

ALIGNMENT FOUNDATIONS

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

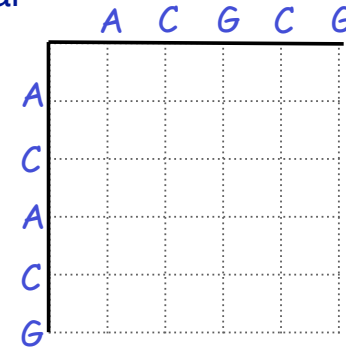
- **How...**

- Dot matrices
- D
- BLAST heuristic approach

How do we compute the optimal alignment between two sequences?

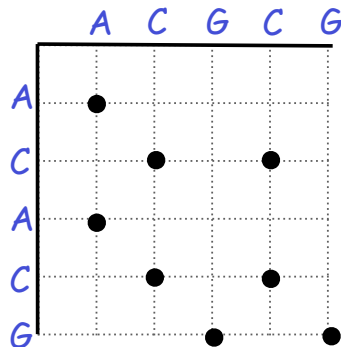
Dot plots: simple graphical approach

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



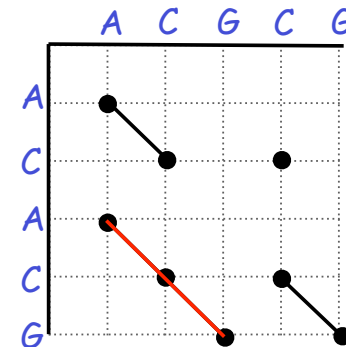
Dot plots: simple graphical approach

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



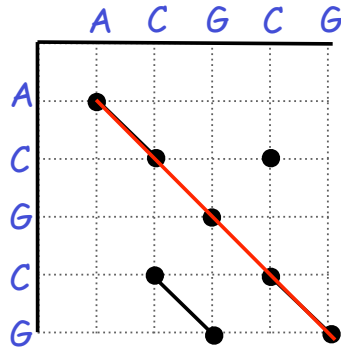
Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



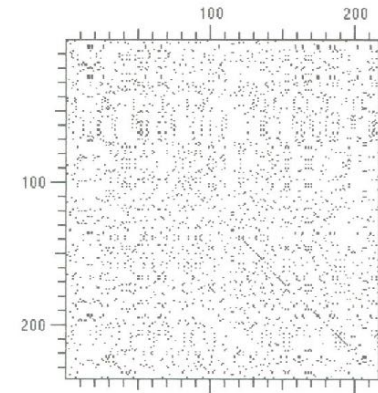
Dot plots: simple graphical approach

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



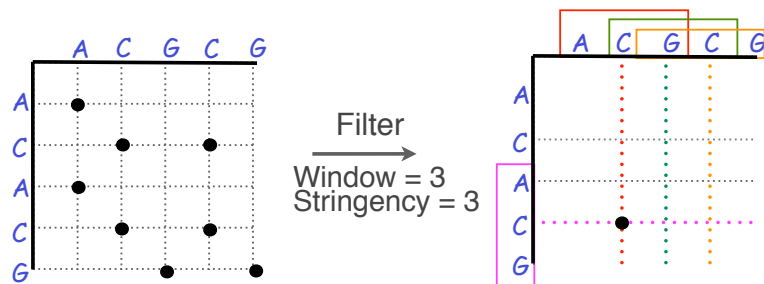
Dot plots: simple graphical approach

- Dot matrices for long sequences can be noisy



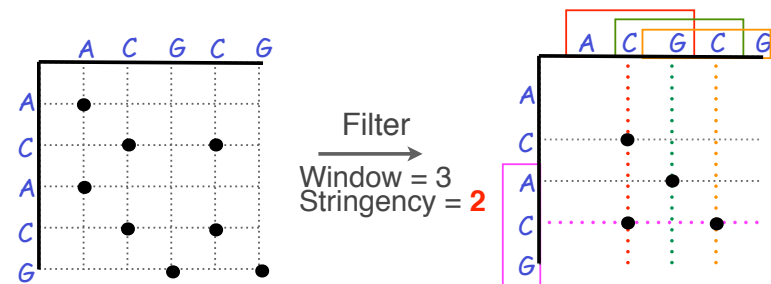
Dot plots: window size and match stringency

- 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

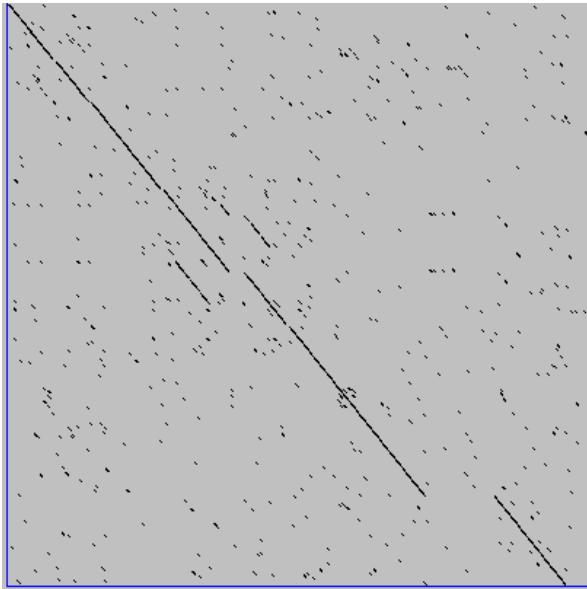


Dot plots: window size and match stringency

- 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



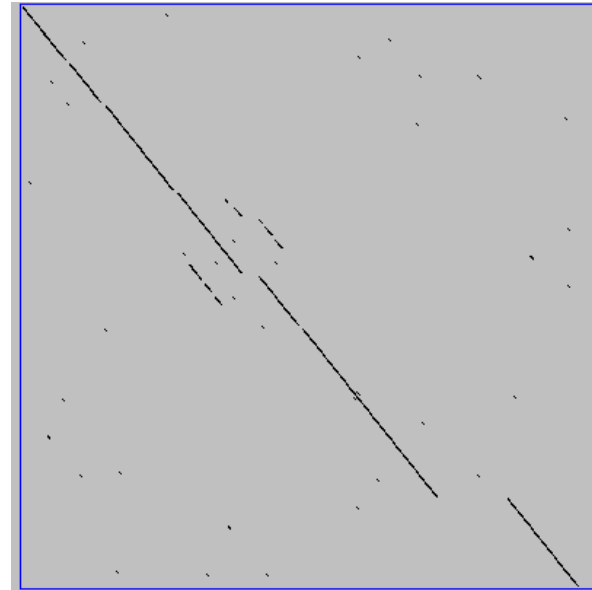
Window size = 5 bases



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



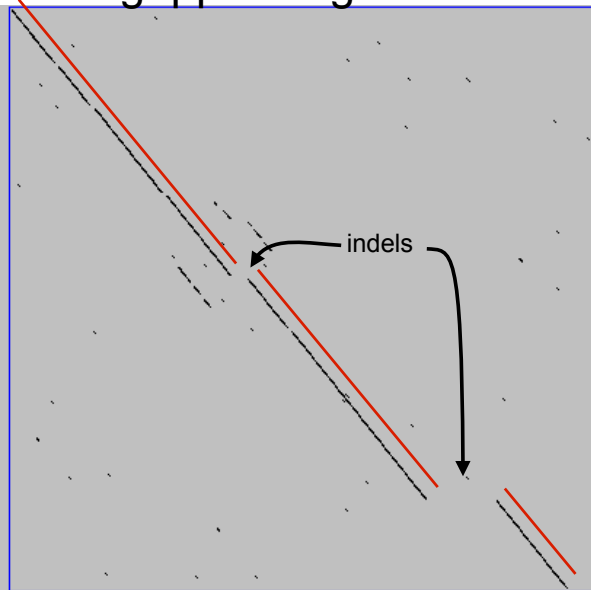
This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

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.

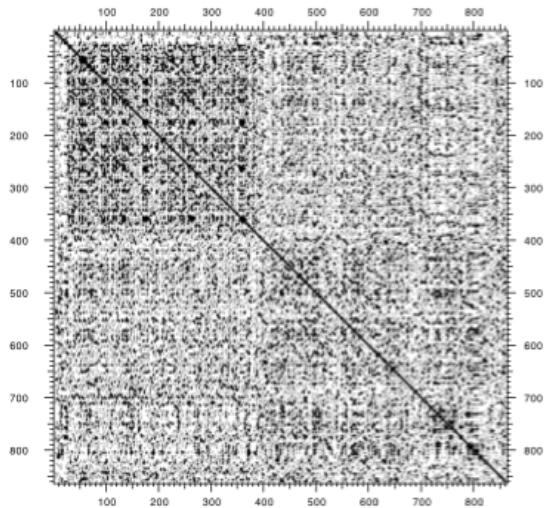
Downward or rightward paths represent **insertion** or **deletions** (gaps in one sequence or the other).

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

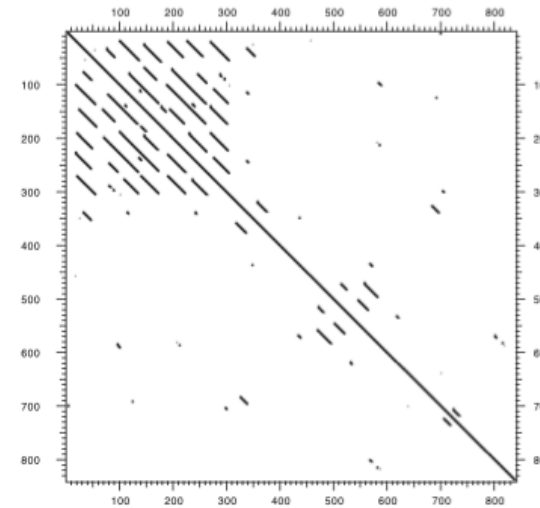


Human LDL receptor
protein sequence
(Genbank P01130)

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**)

<http://bio3d.ucsd.edu/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 [Meitzel and Lenik 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 (j, i) if character number i in the first sequence is the same as character number j in the second sequence. More elaborate forms are "sliding windows" composed of multiple characters and a threshold value, or "match stringency": the two sequences to be considered as matched.

Dot Plot Parameters

Alter the parameters below to change the displayed protein and DNA dot plots. It is important to have a good feel for these parameters when we get to alignment heuristic approaches later.

Window Size: [Slider: 1 to 15, set to 3]

Moving window step size: [Slider: 1 to 15, set to 3]

Match stringency: [Slider: 1 to 15, set to 3]

Match stringency specifies the number of match characters required per window. It should not be lower than the window size.

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" value mean? What value?

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

Algorithm of Needleman and Wunsch

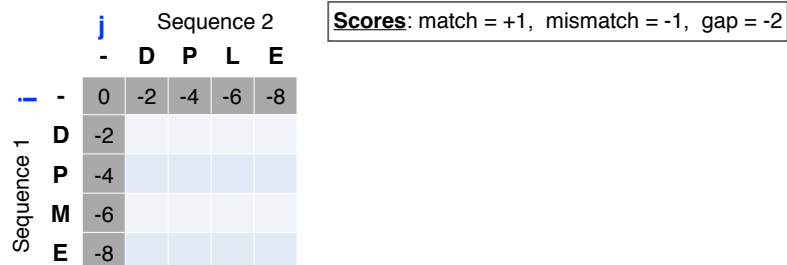
- The Needleman–Wunsch approach to global sequence alignment has three basic steps:
 - setting up a 2D-grid (or **alignment matrix**),
 - scoring the matrix**, and
 - identifying the **optimal path** through the matrix



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.

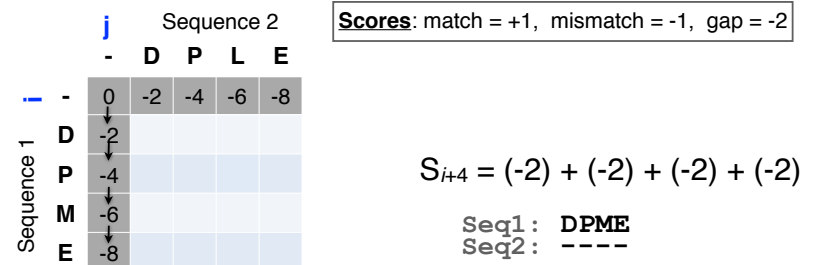
Scoring the alignment matrix

- 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

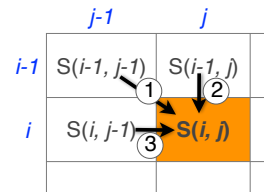


Scoring the alignment matrix

- 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

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

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



Scoring the alignment matrix

- 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

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

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

$$S(i, j) = \text{Max} \begin{cases} S(i-1, j-1) + (\text{mis})\text{match} & \rightarrow \textcircled{1} \\ S(i-1, j) + \text{gap penalty} & \rightarrow \textcircled{2} \\ S(i, j-1) + \text{gap penalty} & \rightarrow \textcircled{3} \end{cases}$$

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

		<i>j</i>			
	-	D	P	L	E
-	0	-2	-4	-6	-8
D	-2	1			
P	-4				
M	-6				
E	-8				

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

- $\rightarrow \textcircled{1} (0) + (+1) = +1 \leq (D-D) \text{ match!}$
 Alignment: $\begin{matrix} D \\ D \end{matrix}$
- $\downarrow \textcircled{2} (-2) + (-2) = -4$
 Alignment: $\begin{matrix} D \\ DP \end{matrix}$
- $\rightarrow \textcircled{3} (-2) + (-2) = -4$

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)

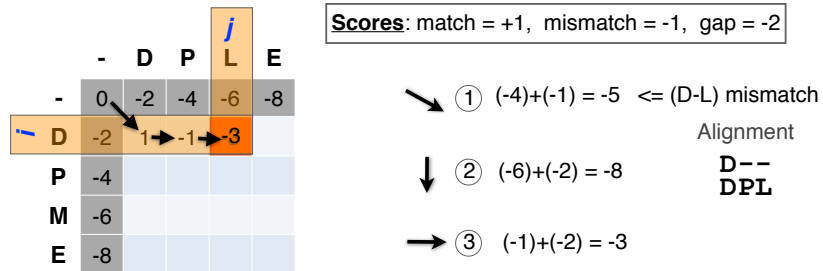
		<i>j</i>			
	-	D	P	L	E
-	0	-2	-4	-6	-8
D	-2	1	-1		
P	-4				
M	-6				
E	-8				

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

- $\rightarrow \textcircled{1} (-2) + (-1) = -3 \leq (D-P) \text{ mismatch!}$
 Alignment: $\begin{matrix} D \\ DP \end{matrix}$
- $\downarrow \textcircled{2} (-4) + (-2) = -6$
 Alignment: $\begin{matrix} D \\ DP \end{matrix}$
- $\rightarrow \textcircled{3} (1) + (-2) = -1$

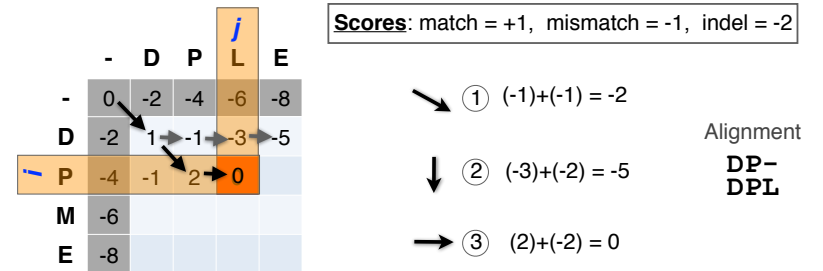
Scoring the alignment matrix

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



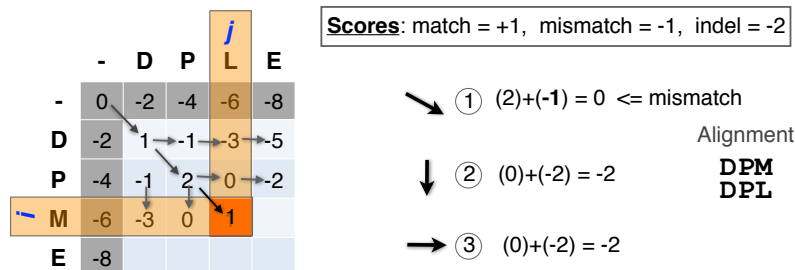
Scoring the alignment matrix

- 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.



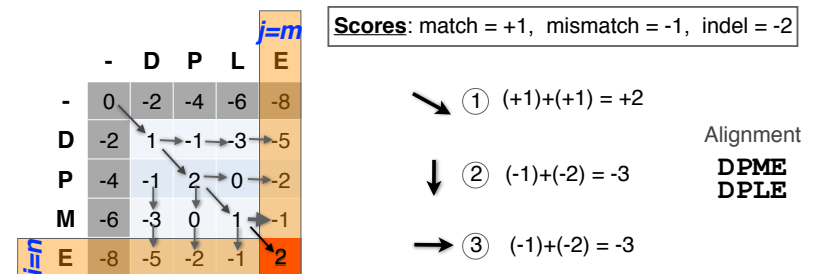
Scoring the alignment matrix

- 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

	-	D	P	L	E
-	0	-2	-4	-6	-8
D	-2	1	-1	-3	-5
P	-4	-1	2	0	-2
M	-6	-3	0	1	-1
E	-8	-5	-2	-1	2

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

Alignment

DPME
DPLE

Questions:

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

	-	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

Questions:

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

	-	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

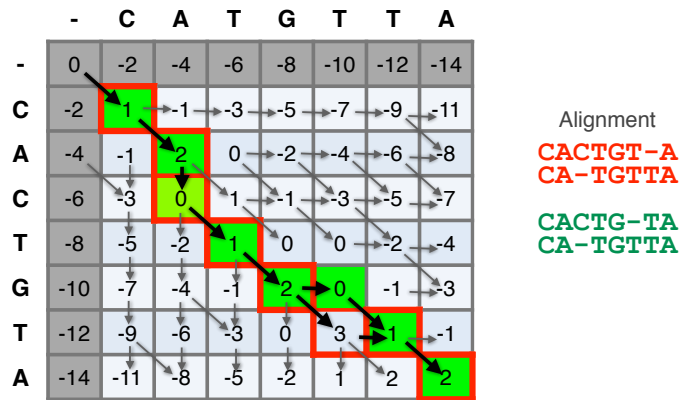
Questions:

- 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

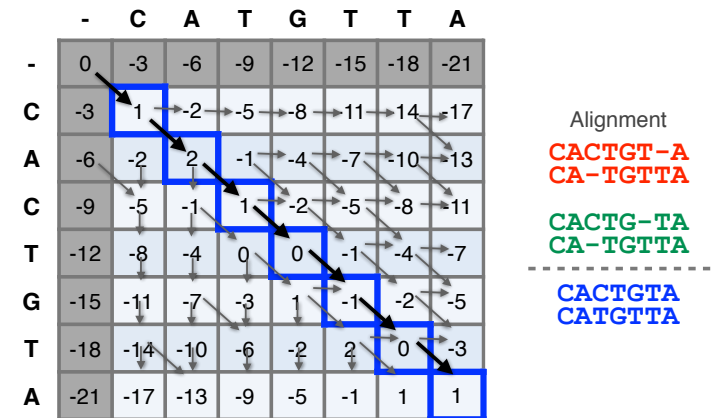
More than one alignment possible

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



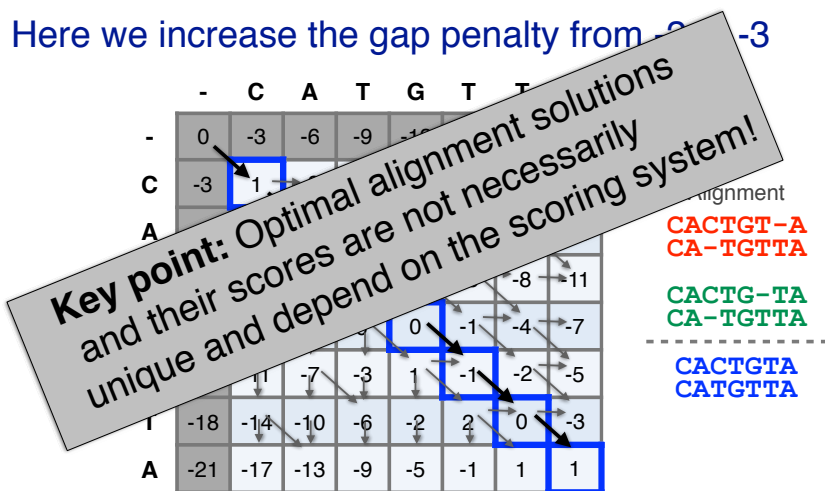
The alignment and score are dependent on the scoring system

- 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



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

NW DYNAMIC PROGRAMMING

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

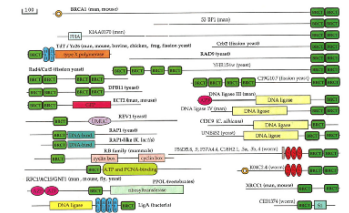
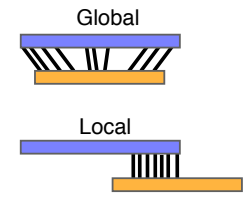
		A	G	T	T	C
	0					
A						
T						
T						
G						
C						

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

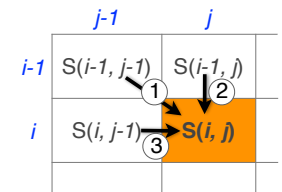
- Smith & Waterman proposed simply that a local alignment of two sequences allow arbitrary-length segments of each sequence to be aligned, with no penalty for the unaligned portions of the sequences. Otherwise, the score for a local alignment is calculated the same way as that for a global alignment

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

$$S(i, j) = \text{Max} \begin{cases} S(i-1, j-1) + (\text{mis})\text{match} & \rightarrow \textcircled{1} \\ S(i-1, j) - \text{gap penalty} & \rightarrow \textcircled{2} \\ S(i, j-1) - \text{gap penalty} & \rightarrow \textcircled{3} \\ 0 & \rightarrow \textcircled{4} \end{cases}$$

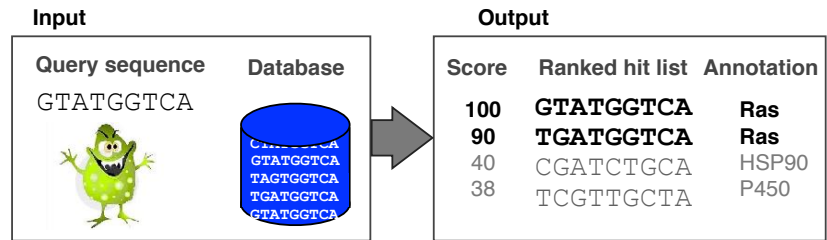


		Sequence 1															
		-	C	A	G	C	C	U	C	G	C	U	U	A	G		
Sequence 2	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
	A	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	A	0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
	U	0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0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

Local alignment
GCC-AUG
GCCUCGC

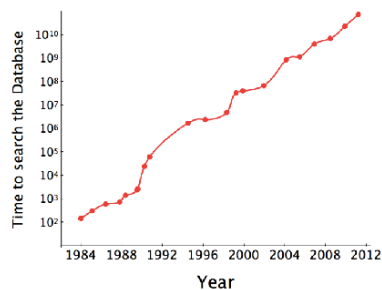
Local alignments can be used for database searching

- **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

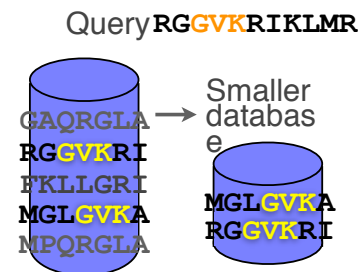
- Due to the rapid growth of sequence databases, search algorithms have to be both efficient and sensitive
 - Time to search with SW is proportional to $m \times n$ (m is length of query, n is length of database), **too slow for large databases!**



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**

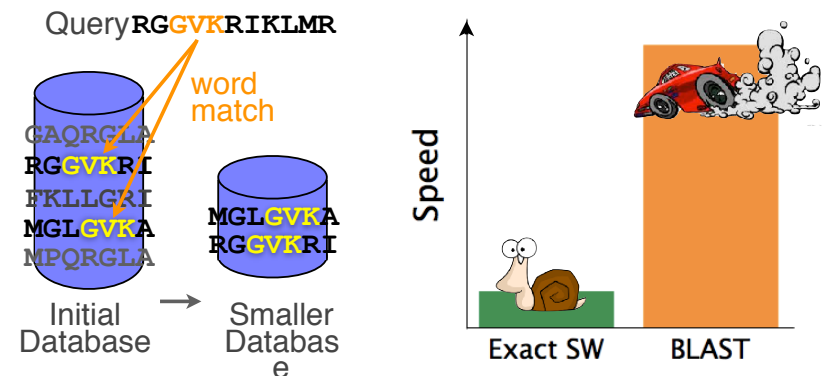
- **BLAST (Basic Local Alignment Search Tool)** is a simplified form of Smith-Waterman (SW) alignment that is popular because it is **fast** and **easily accessible**
 - BLAST is a heuristic approximation to SW - It examines only part of the search space
 - BLAST saves time by restricting the search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

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

- 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 ($w=3$)

generate list of $w=3$ words for query

RGGVKRI Query sequence
GGV
GVK
VKR
KRI

113

Blast

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

extend list of words similar to query

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

114

Blast

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

search for perfect matches in the database sequence

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

115

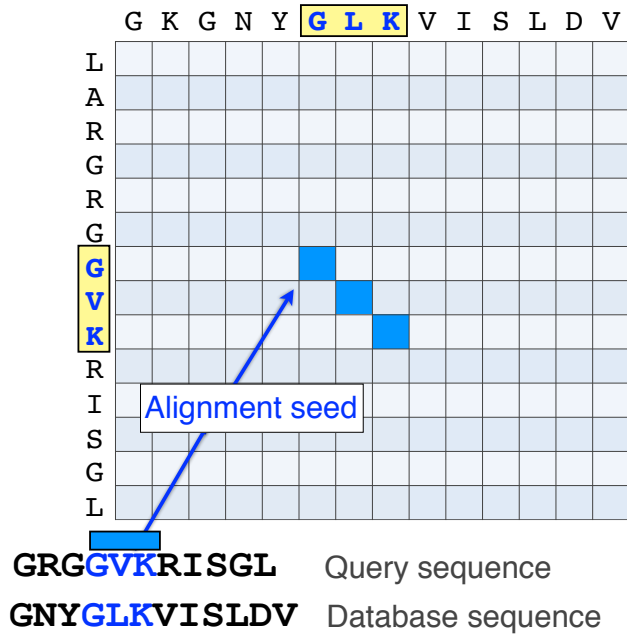
Blast

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

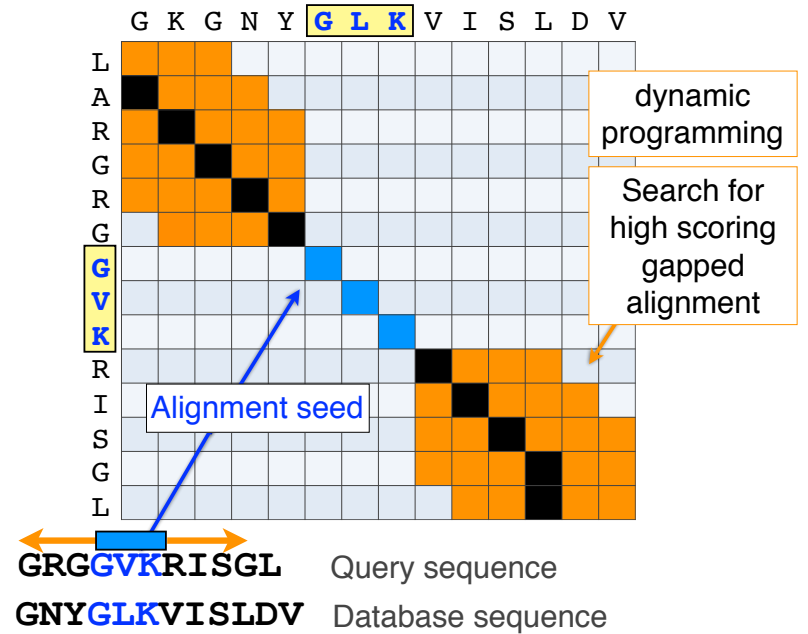
matched word is used as a local alignment seed

GNVGLKVISLDVE Database sequence
RGGVKRI Query sequence

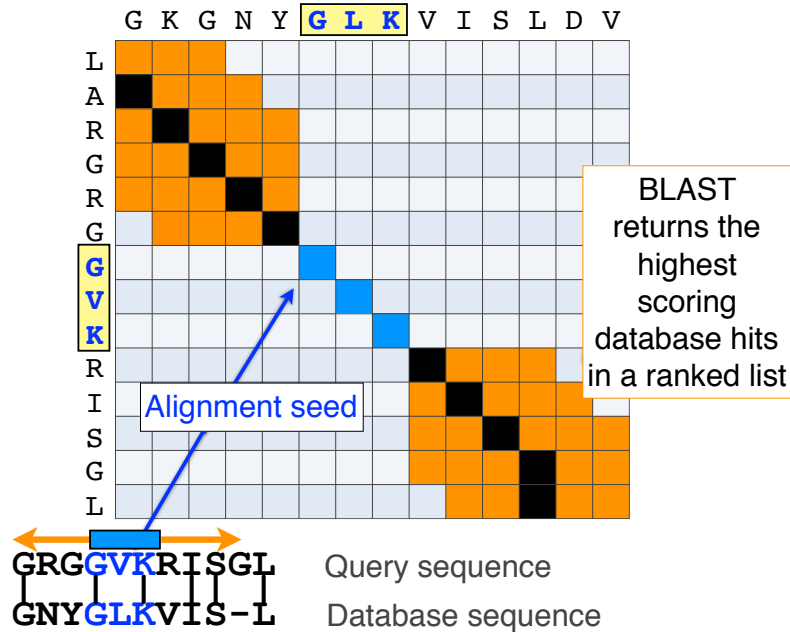
116



117



118



119

BLAST output

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

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

120

Statistical significance of results

- An important feature of BLAST is the computation of statistical significance for each hit. This is described by the **E value** (expect value)

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

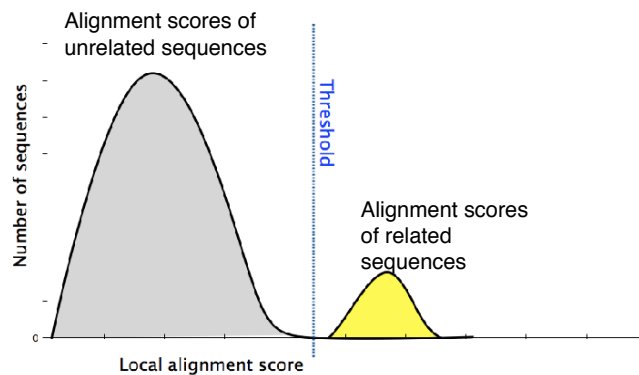
121

BLAST scores and E-values

- The **E value** is the **expected** number of hits that are as good or better than the observed local alignment score (with this score or better) if the query and database are **random** with respect to each other
 - i.e.* the number of alignments expected to occur by chance with equivalent or better scores
- Typically, only hits with E value **below** a significance threshold are reported
 - This is equivalent to selecting alignments with score above a certain score threshold

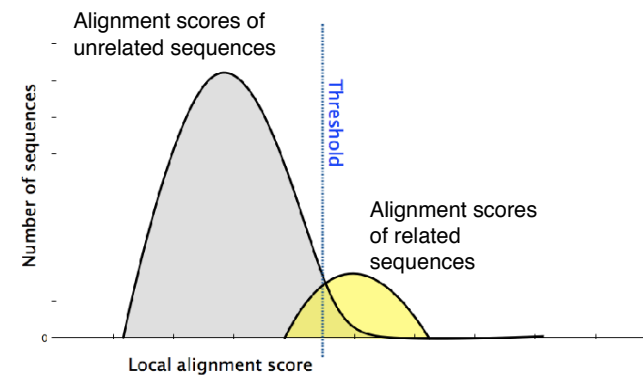
122

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



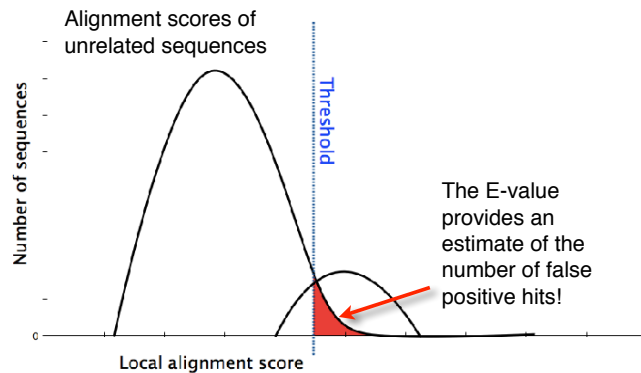
123

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



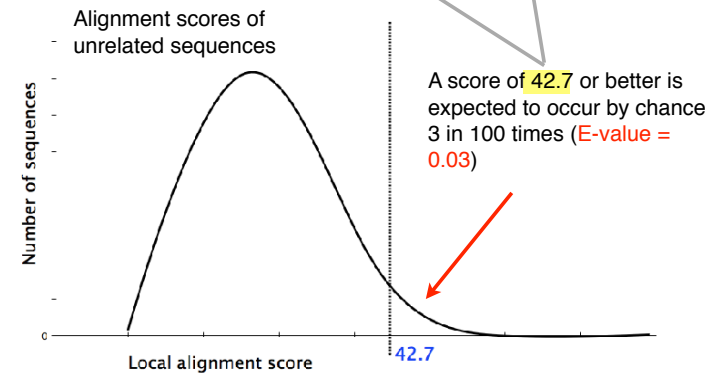
124

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



125

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



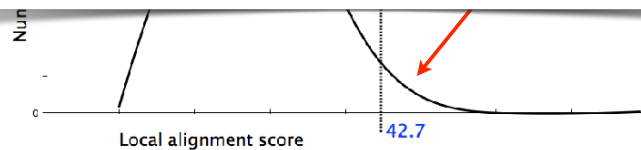
126

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>



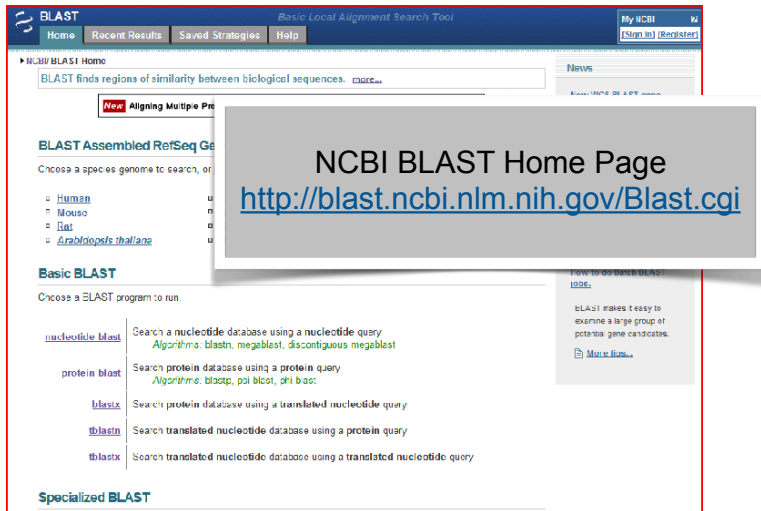
127

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



Practical database searching with BLAST

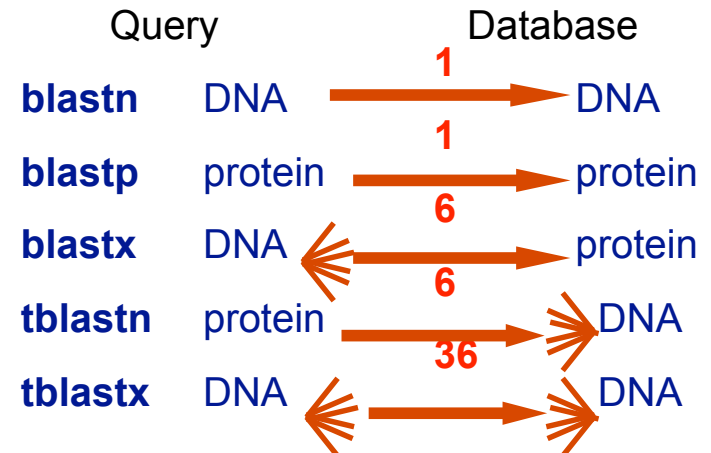
- There are four basic components to a traditional BLAST search
 - (1) Choose the sequence (query)
 - (2) Select the BLAST program
 - (3) Choose the database to search
 - (4) Choose optional parameters
- Then click “BLAST”

Step 1: Choose your sequence

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



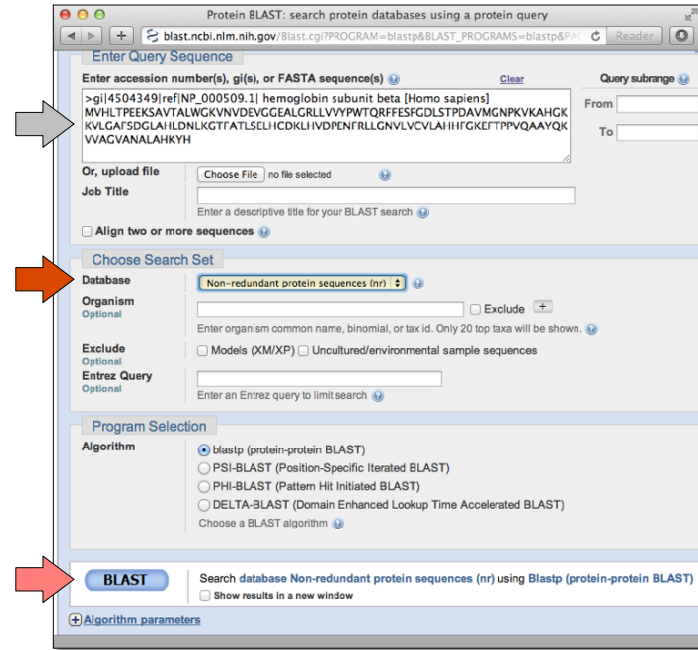
Step 2: Choose the BLAST program



DNA potentially encodes six proteins

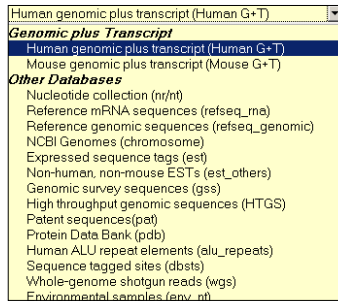
```

5' CAT CAA
5' ATC AAC
5' TCA ACT
5' CATCAACTACAACCTCCAAAGACACCCTTACACATCAACAACCTACCCAC 3'
3' GTAGTTGATGTTGAGGTTTCTGTGGGAATGTGTAGTTGTTGGATGGGTG 5'
5' GTG GGT
5' TGG GTA
5' GGG TAG
    
```

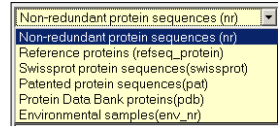


Step 3: Choose the database

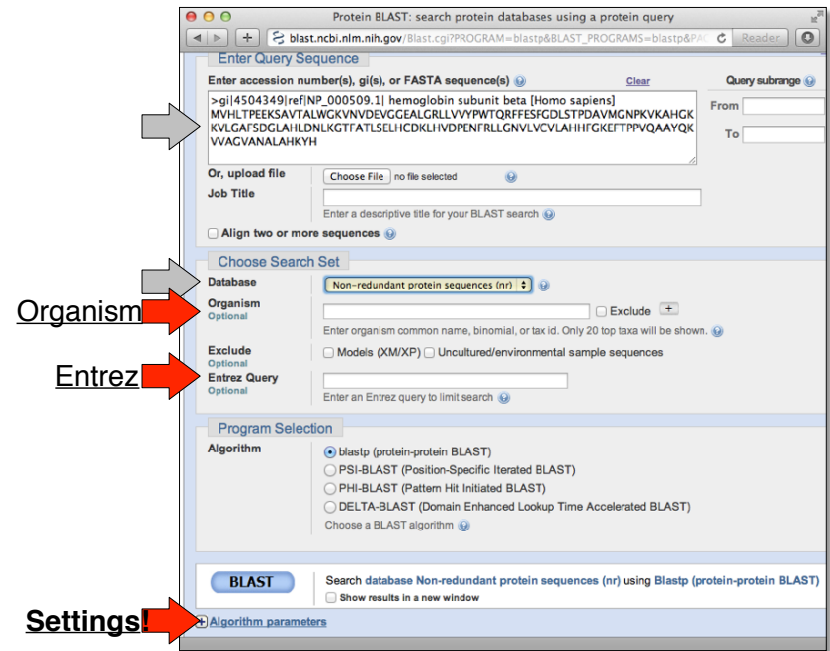
- nr = non-redundant (most general database)
- dbest = database of expressed sequence tags
- dbsts = database of sequence tag sites
- gss = genomic survey sequences



nucleotide databases



protein databases



Step 4a: Select optional search parameters

Algorithm parameters

General Parameters

Max target sequences: 100
 Short queries: Automatically adjust parameters for short input sequences
 Expect threshold: 10
 Word size: 3
 Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62
 Gap Costs: Existence: 11 Extension: 1
 Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: Low complexity regions
 Mask: Mask for lookup table only
 Mask lower case letters

BLAST Search database Non-redundant protein sequences (nr) using Blastp
 Show results in a new window

137

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

138

Results page

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

BLAST Basic Local Alignment Search Tool

NCBI/BLAST/blastp suite/ Formatting Results - FVGUTMRZ013

gi|4504349|ref|NP_000509.1| hemoglobin

Query ID: Id|84677
 Description: gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
 Molecule type: amino acid
 Query Length: 147

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

Graphic Summary

Show Conserved Domains

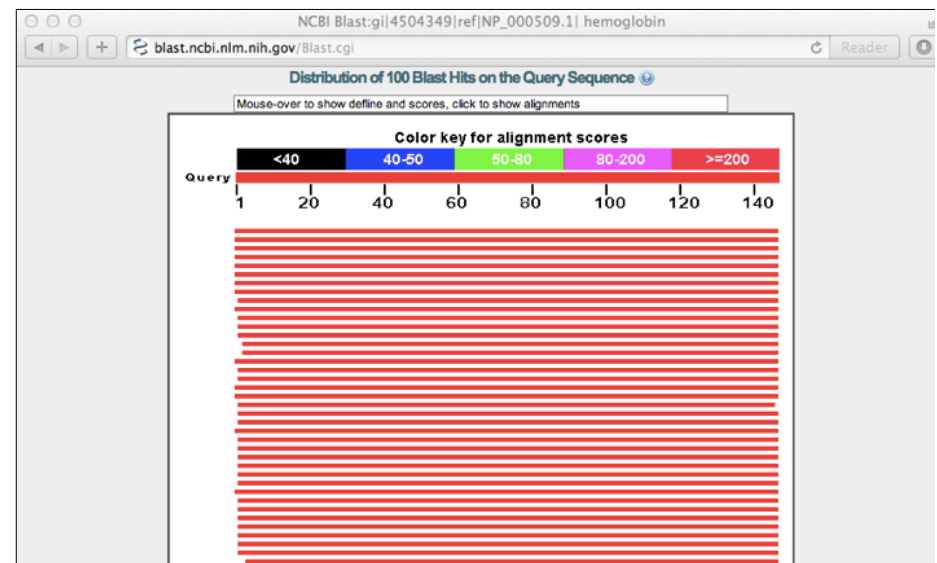
Putative conserved domains have been detected, click on the image below for detailed results.

Query seq.

Specific hits: globin
 Superfamilies: globin_like superfamily

Distribution of 100 Blast Hits on the Query Sequence

Further down the results page...



Further down the results page...

Sequences producing significant alignments:

Select: All None Selected:0

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%	AAN84548.1
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%	1COH_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

Further down the results page...

hemoglobin subunit beta [Homo sapiens]

Sequence ID: [ref|NP_000509.1|](#) Length: 147 Number of Matches: 1

[See 84 more title\(s\)](#)

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

Query 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYVPTQRFESFGDLSTPDVAVGNPK 60
 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYVPTQRFESFGDLSTPDVAVGNPK 60
 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYVPTQRFESFGDLSTPDVAVGNPK 60

Sbjct 1 MVHLTPEEKSAVTALMGKVNDEVGGEALGRLLVYVPTQRFESFGDLSTPDVAVGNPK 60

Query 61 VKAHGKVLGAFSDGLAHLNHLKGTFTLSELHCDKLHVDPENFRLLGNLVLCVLAHHPG 120
 VKAHGKVLGAFSDGLAHLNHLKGTFTLSELHCDKLHVDPENFRLLGNLVLCVLAHHPG 120

Sbjct 61 VKAHGKVLGAFSDGLAHLNHLKGTFTLSELHCDKLHVDPENFRLLGNLVLCVLAHHPG 120

Query 121 KEFTFPVQAAYQKVVAGVANALAHRYH 147
 KEFTFPVQAAYQKVVAGVANALAHRYH 147

Sbjct 121 KEFTFPVQAAYQKVVAGVANALAHRYH 147

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

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

Different output formats are available

NCBI 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](#)

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:

[g|4504349|ref|NP_000509.1| hemoglobin](#)

E.g. Query anchored alignments

NCBI BLAST® Basic Local Alignment Search Tool

NCBI/BLAST/blastp suite/ Formatting Results - FVGUTMR2013

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#) [Change the result display back](#)

Query anchored alignments

Query	Score	Expect	Method	Identities	Positives	Gaps
AAX37051	1					
AAX29557	1					
NP_000509	1					
P02024	1					
AAN84548	1					
AAZ39780	1					
ACU56984	1					
AAD19696	1					
1COH_B	1					
AAF00469	1					
2YRS_B	1					
1HDB_B	1					
1DXU_B	1					
1DKV_B	2					
3RMP_C	2					
AAL68978	1					
1NQP_B	1					
1K1K_B	1					
AAN11320	1					
XP_002822173	1					
1Y85_B	1					
1YE0_B	1					
1O10_B	1					
CAA23759	1					
1YE2_B	1					
1Y5F_B	1					
1A00_B	1					
1HBS_B	1					
1ABY_B	1					
1CHY_B	1					

... and alignments with dots for identities

Accession	Score	Alignment
Query	1	MVHLTPEEKSAVTALMGKVNVDVEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPK 60
AAK37051	1 60
AAK29557	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..... 59
AAN11320	1V..... 60
XP_002822173	1V..... 60
IY85_B	1	M.....A..... 59
IY50_B	1	M.....A..... 59
IOIO_B	1	M.....V.....X..... 59
CAA23759	1V.....X..... 60
IY52_B	1	M.....F..... 59
IY5F_B	1	M.....F..... 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?

146

How to handle too many results

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

147

How to handle too few results

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

148

Summary of key points

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

FOR NEXT CLASS...

Check out the online:

Reading: Sean Eddy's "What is dynamic programming?"

Homework: (1) [Quiz](#), (2) [Alignment Exercise](#).

Homework Grading

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

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