

BGGN 213

Foundations of Bioinformatics Lecture 2

Barry Grant
UC San Diego

<http://thegrantlab.org/bggn213>

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

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...

Primary, secondary & composite databases

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

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

DATABASE VIGNETTE

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

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

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

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

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

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

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

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 results of a global cross query for 'ras' on the NCBI website. The results are categorized into Literature, Genes, Health, and Proteins. The 'Genes' section is highlighted with a red box around the 'Gene' entry, which shows 87,165 results. Other categories include Books (1,677), MeSH (402), NLM Catalog (223), PubMed (54,672), PubMed Central (96,114), GEO DataSets (3,732), GEO Profiles (1,622,789), HomoloGene (696), ClinVar (759), dbGaP (120), GTR (1,879), and PopSet (2,254).

[ras - Gene - NCBI](http://www.ncbi.nlm.nih.gov/gene/?term=ras)

NCBI Resources How To Sign in to NCBI

Gene Gene Search Save search Advanced Help

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

Filters: Manage Filters

Did you mean ras as a gene symbol? Search Gene for ras as a symbol.

Results: 1 to 20 of 85633

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

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

Find related data Database: Select Find items

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

Top Organisms [Tree] Homo sapiens (1126) Mus musculus (823) Rattus norvegicus (625) Oryctolophus niloticus (533) Neolamprologus brichardi (507) All other taxa (82019) More...

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

9

[ras AND "Homo sapiens"\[porgn:_txid9606\] - Gene - NCBI](http://www.ncbi.nlm.nih.gov/gene/?term=(ras)&term=(Homo+sapiens)[porgn:_txid9606])

NCBI Resources How To Sign in to NCBI

Gene Gene Search Save search Advanced Help

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

Filters: Manage Filters

Results: 1 to 20 of 1126

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

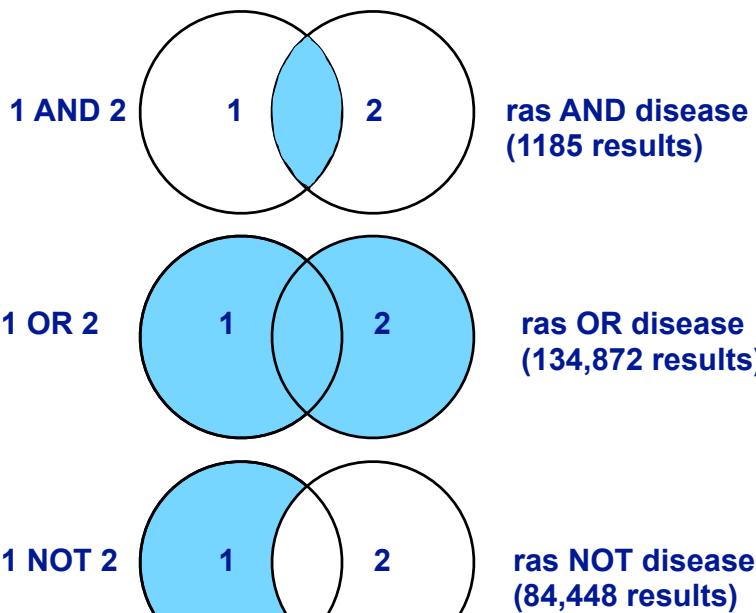
Name/Gene ID	Description	Location	Aliases
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (11470446..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS	Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, NS2, RAS2

Find related data Database: Select Find items

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

Recent activity Turn Off Clear

10



11

[ras AND "Homo sapiens"\[porgn:_txid9606\] - Gene - NCBI](http://www.ncbi.nlm.nih.gov/gene/?term=(ras)&term=(Homo+sapiens)[porgn:_txid9606])

NCBI Resources How To Sign in to NCBI

Gene Gene Search Save search Advanced Help

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

Filters: Manage Filters

Results: 1 to 20 of 1126

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

Name/Gene ID	Description	Location	Aliases
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (11470446..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS	Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, NS2, RAS2

Find related data Database: Select Find items

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

Recent activity Turn Off Clear

12

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

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

Summary

Official Symbol KRAS provided by HGNC
Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source HGNC:HGNC:6407
See related Ensembl:ENSG0000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193

Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; 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)

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

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

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

Summary

Official Symbol KRAS provided by HGNC
Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source HGNC:HGNC:6407
See related Ensembl:ENSG0000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193

Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; 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)

Location: 12p12.1
Exon count: 6

Annotation release: 106 Status: current Assembly: GRCh38 (GCF_000001405.26) Chr: 12 Location: NC_000012.12 (2505246..25250923, complement)

Annotation release: 105 Status: previous assembly Assembly: GRCh37.p13 (GCF_000001405.25) Chr: 12 Location: NC_000012.11 (25358180..25403870, complement)

Genomic regions, transcripts, and products

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

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
Map Viewer
MedGen
Nucleotide

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

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

Gene ID: 3845

Summary

Official Symbol KRAS provided by HGNC
Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source HGNC:HGNC:6407
See related Ensembl:ENSG0000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193

Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; 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)

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

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

GO: Gene Ontology

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

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.

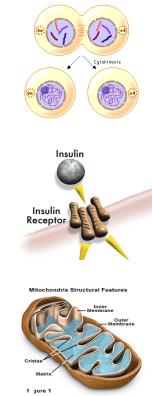
UniProt is a member of the GO Consortium.

Why do we need Ontologies?

- Annotation is essential for capturing the understanding and knowledge associated with a sequence or other molecular entity
- Annotation is traditionally recorded as “free text”, which is easy to read by humans, but has a number of disadvantages, including:
 - Difficult for computers to parse
 - Quality varies from database to database
 - Terminology used varies from annotator to annotator
- Ontologies are annotations using standard vocabularies that try to address these issues
- GO is integrated with UniProt and many other databases including a number at NCBI

GO Ontologies

- There are three ontologies in GO:
 - Biological Process**
A commonly recognized series of events e.g. cell division, mitosis,
 - Molecular Function**
An elemental activity, task or job e.g. kinase activity, insulin binding
 - Cellular Component**
Where a gene product is located e.g. mitochondrion, mitochondrial membrane



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

Scroll down to
UniProt link

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

UniProtKB/Swiss-Prot:P01116

Scroll down to
Very bottom for
UniProt link

UniProt will detail much more information for protein coding genes

P01116 - RASK_HUMAN

Reviewed - Experimental evidence at protein level

Display

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

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

UniProt will detail much more information for protein coding genes

P01116 - RASK_HUMAN

Reviewed - Experimental evidence at protein level

Display

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

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

View FASTA file format

UniProt will detail much more information for protein coding genes

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

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

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

4EPV
Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation
DOI: 10.2210/pdb4epv/pdb
Classification: HYDROLASE
Deposited: 2012-04-17 Released: 2012-05-23
Deposition author(s): Sun, Q., Burke, J.P., Phan, J., Burns, M.C., Olejniczak, E.T., Waterson, A.G., Lee, T., Rossanese, O.W., Fesik, S.W.
Organism: Homo sapiens
Expression System: Escherichia coli
Mutation(s): 1
Experimental Data Snapshot wwPDB Validation 3D Report Full Report
Method: X-RAY DIFFRACTION Metric Percentile Ranks Value

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

4EPV
Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation
Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Click to identify atoms and bonds.
Bond: [GLY]12:A-O - [GLY]12:A-C

Display Options
Assembly: Biossembly 1
Model: Model 1
Symmetry: None
Interaction: IGDPJ201-A
Style: Cartoon
Color: Rainbow
Ligand: None
Quality: Automatic
Water: Ions:
Hydrogens: Clashes:

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

FAMILY & DOMAINS
PFAM: PF00071: Ras. 1 hit. [Graphical view]

PFAM is one of the best protein family databases

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

Family: Ras (PF00071)
Summary Domain organisation Clan Alignments HMM logo Trees Curation & model **Species** Interactions Structures
Jump to... enter ID/acc Go

Summary: Ras family
Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.
Wikipedia: Ras subfamily | Wikipedia: Ras superfamily | Pfam | InterPro
This is the Wikipedia entry entitled "Ras subfamily". More...

Ras subfamily Edit Wikipedia article
This article is about p21/Ras protein. For the p21/Waf1 protein, see p21.
Ras is the name given to a family of related proteins which is ubiquitously expressed in all cell lineages and organs. All Ras protein family members belong to a class of protein called small GTPases, and are involved in transmitting signals within cells (cellular signal transduction). Ras is the prototypical member of the Ras superfamily of proteins, which are all related in 3D structure and regulate diverse cell behaviours.
The name "Ras" is an abbreviation of "Rat Sarcoma", reflecting the way the first members of the protein family were discovered. The name "Ras" is also used to refer to the family of genes encoding these proteins.
When Ras is "switched on" by incoming signals, it subsequently switches on other proteins, which ultimately turn on genes involved in cell growth, differentiation and survival. As a result, mutations in ras genes can lead to the production of permanently activated Ras proteins. This can cause unintended and overactive signalling inside the cell, even in the absence of incoming signals.
Because these signals result in cell growth and division, overactive Ras signalling can ultimately lead to cancer.^[1] The 3 Ras genes in humans (HRAS, KRAS, and NRAS) are the most common oncogenes in human cancer; mutations that permanently activate Ras are found in 20% to 25% of all human tumors and up to 90% in certain types of cancer (e.g., pancreatic cancer).^[2] For this reason, Ras inhibitors are being studied as a treatment for cancer, and other diseases with Ras overexpression.

Contents [Edit]
1 History
2 Structure
3 Function
3.1 Activation and deactivation
3.2 Membrane attachment
4 Members
5 Ras in cancer
5.1 Inappropriate activation
5.2 Constitutively active Ras

Identifiers
Symbol: Ras
Pfam: PF00071_0
InterPro: IPR013753_0
PROSITE: POC00017_0
SCOP: Sp21_0
SUPERFAMILY: Sp21_0

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

Summary
Domain organisation
Clin
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures

Jump to... ↻
 Go

Species distribution
 Sunburst Tree

This visualisation provides a simple graphical representation of the distribution of this family across species. You can find the original interactive tree in the adjacent tab. [More...](#)

Sunburst controls Hide

Home sapiens

- Root
 - Eukaryota
 - Metazoa
 - Bilateria
 - Mammalia
 - Primates
 - Hominoidea
 - Homo
 - Homo sapiens

Weight segments by...

- number of sequences
- number of species

Change the size of the sunburst **Large**

Colour assignments

Orange	Purple	Eukaryota
Yellow	Red	Other sequences
Magenta	Cyan	Unclassified
Green	Blue	Fungi
Dark Blue	Light Blue	Unclassified sequence

Selections

Alien selected sequences to HMM
 Generated a-format file
 Date selection
 Currently selected:

- 521 sequences
- 1 species

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

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

Example Questions:

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

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

Family: Ras (PF00071)

HMM logo

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

Contribution

322 architectures 21243 sequences 30 interactions 1006 species 663 structures

Summary Domain organisation Clan Alignments **HMM logo** Trees Curation & model Species Interactions Structures

Jump to ... ↴
enter ID/acc Go

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.

European Molecular Biology Laboratory

Pfam: Family: Kinesin (PF00225)

<http://pfam.janelia.org/family/kinesin#tabview=tab9>

RSS Google

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

Structures

For those sequences which have a structure in the Protein DataBank², we use the mapping between UniProt³, PDB and Pfam coordinate systems from the PDBeC group, to allow us to map Pfam domains onto UniProt sequences and three-dimensional protein structures. The table below shows the structures on which the **Kinesin** domain has been found.

UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View	
A8BKD1_GIALA	11 - 335	2vgv	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	
		1f9t	A	392 - 723	Jmol AstexViewer SPICE	
		1f9u	A	392 - 723	Jmol AstexViewer SPICE	
		1f9v	A	392 - 723	Jmol AstexViewer SPICE	
		1f9w	A	392 - 723	Jmol AstexViewer SPICE	
		1f9w	B	392 - 723	Jmol AstexViewer SPICE	
		3kar	A	392 - 723	Jmol AstexViewer SPICE	
			A	11 - 352	Jmol AstexViewer SPICE	
			B	11 - 352	Jmol AstexViewer SPICE	
			C	11 - 352	Jmol AstexViewer SPICE	
KI13B_HUMAN	11 - 352	3gbi				
			1i16	A	24 - 359	Jmol AstexViewer SPICE
			1i16	B	24 - 359	Jmol AstexViewer SPICE
			1q0b	A	24 - 359	Jmol AstexViewer SPICE
			1q0b	B	24 - 359	Jmol AstexViewer SPICE
			1x88	A	24 - 359	Jmol AstexViewer SPICE
			1x88	B	24 - 359	Jmol AstexViewer SPICE
			1	A	24 - 359	Jmol AstexViewer SPICE

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 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

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
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!

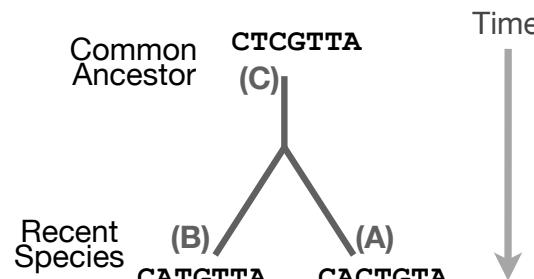
ALIGNMENT FOUNDATIONS

- 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.

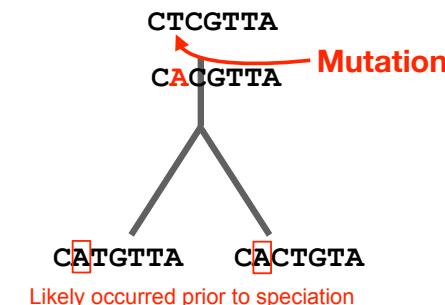
- Mutations/Substitutions
- Deletions
- Insertions



Mutations, deletions and insertions

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

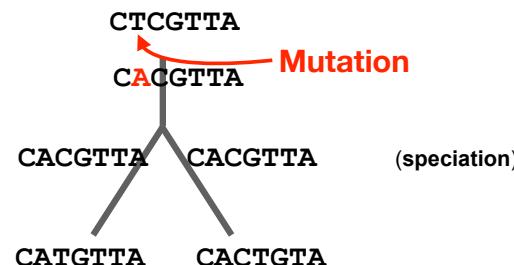
- Mutations/Substitutions
- Deletions
- Insertions



Mutations, deletions and insertions

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

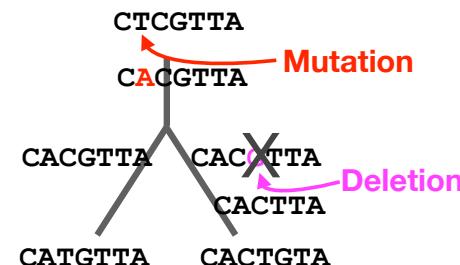
- Mutations/Substitutions
- Deletions
- Insertions



Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

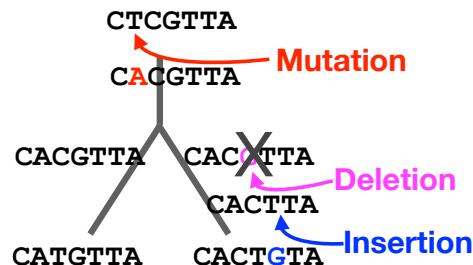


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA
CACGTTA → CACTTA
CACTTA → CACTGTA

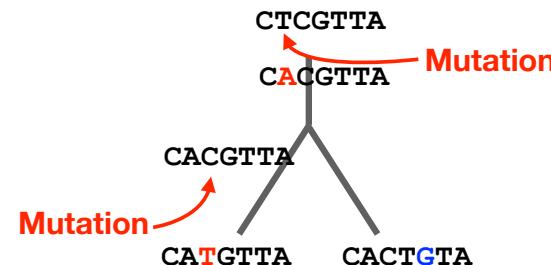


Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

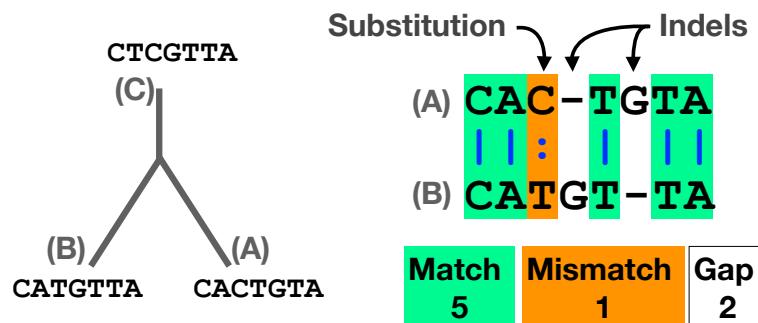
CTCGTTA → CACGTTA
CACGTTA → CATGTTA



Alignment view

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

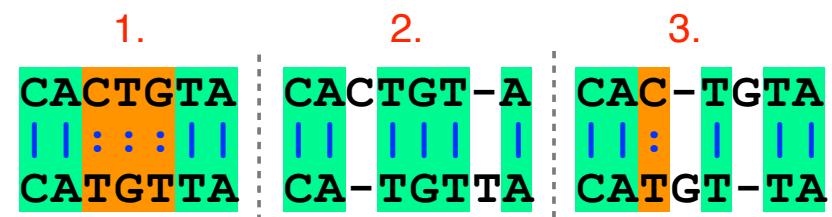
- Mismatches represent mutations/substitutions
- Gaps represent insertions and deletions (indels)



Alternative alignments

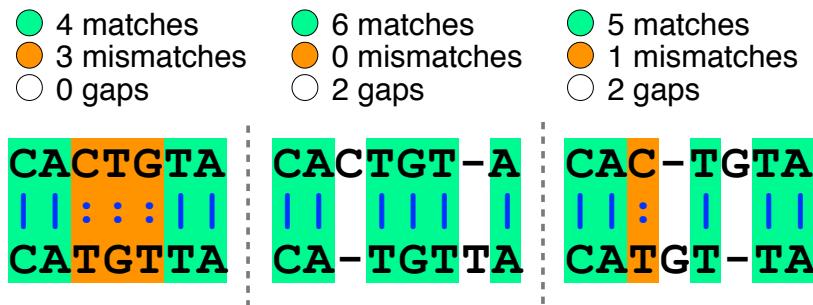
- Unfortunately, finding the correct alignment is difficult if we do not know the evolutionary history of the two sequences

Q. Which of these 3 possible alignments is best?



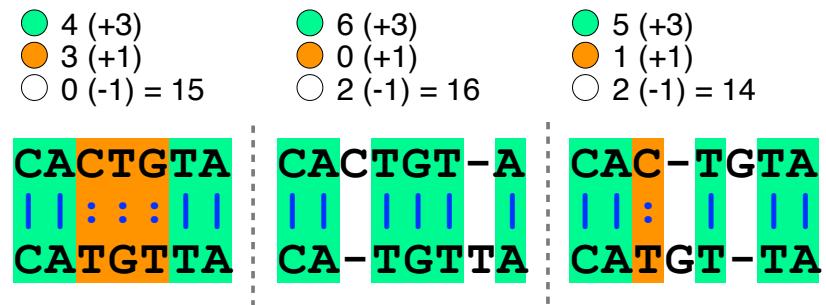
Alternative alignments

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



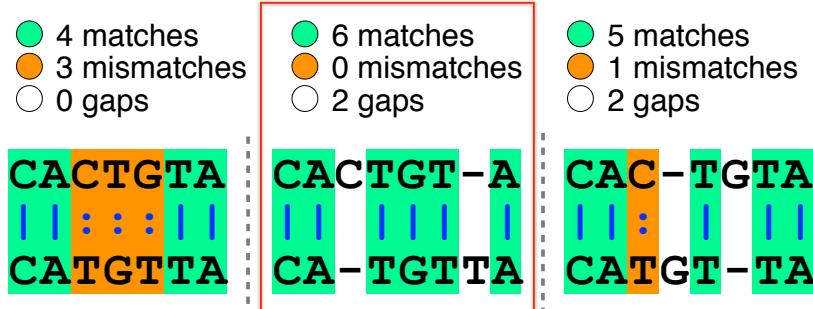
Scoring alignments

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



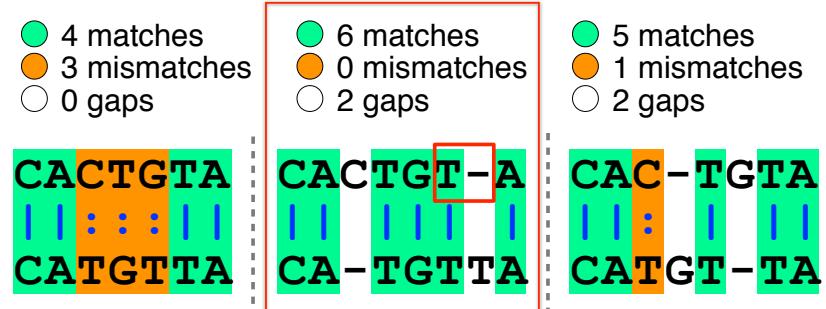
Optimal alignments

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



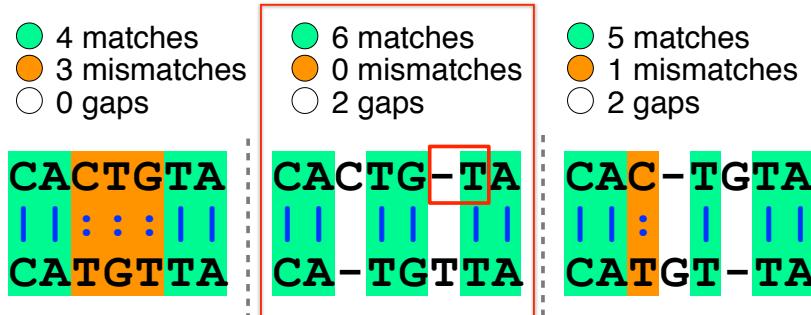
Optimal alignments

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



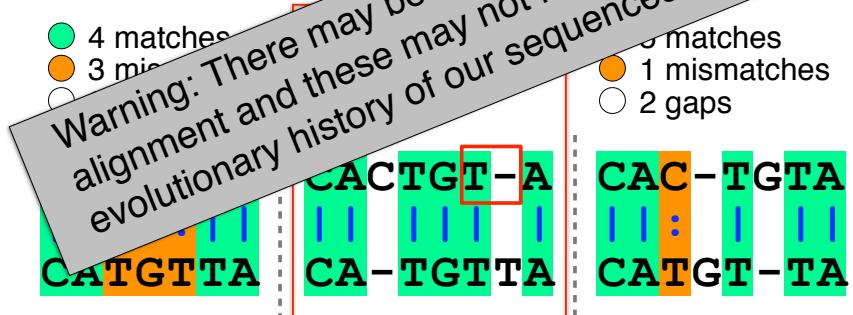
Optimal alignments

- 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.



ALIGNMENT FOUNDATIONS

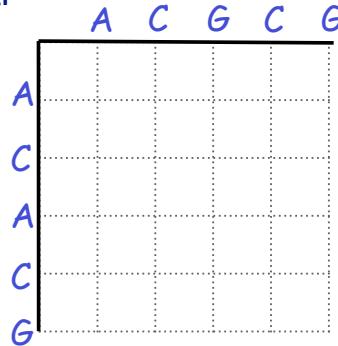
- Why...
 - Why compare biological sequences?
- What...
 - Alignment view of sequence changes during evolution (matches, mismatches and gaps)
- How...
 - Dot matrices
 - Dynamic programming
 - Global alignment
 - Local alignment
 - BLAST heuristic approach

ALIGNMENT FOUNDATIONS

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

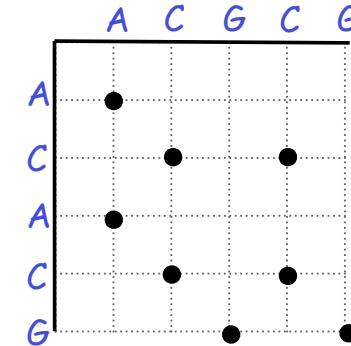
Dot plots: simple graphical approach

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



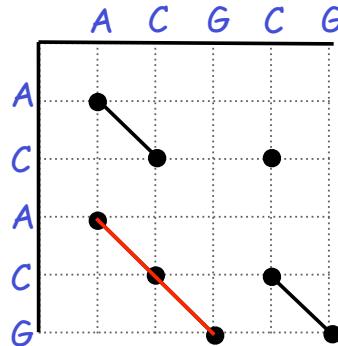
Dot plots: simple graphical approach

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



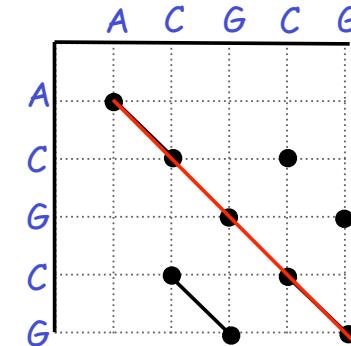
Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



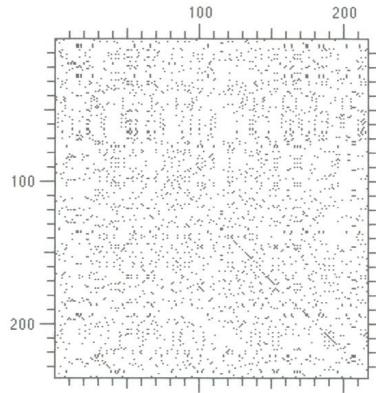
Dot plots: simple graphical approach

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



Dot plots: simple graphical approach

- Dot matrices for long sequences can be noisy

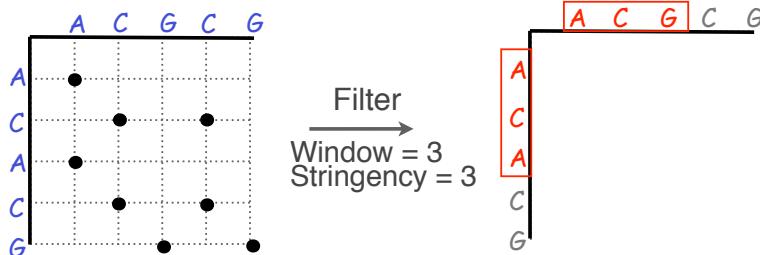


Dot plots: window size and match stringency

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

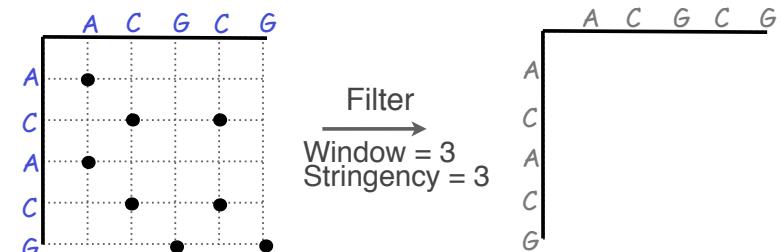


Dot plots: window size and match stringency

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

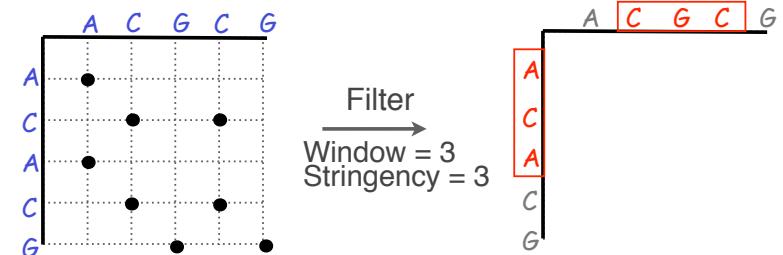


Dot plots: window size and match stringency

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.

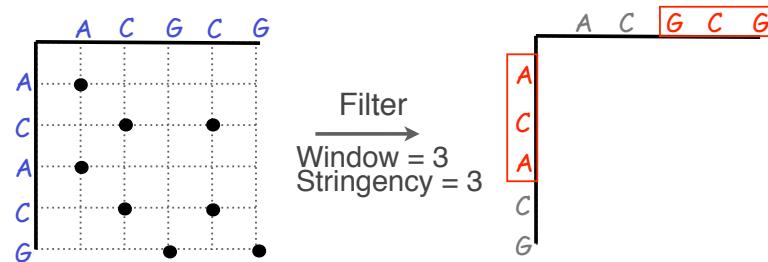
- You have to choose window size and stringency



Dot plots: window size and match stringency

Solution: use a window and a threshold

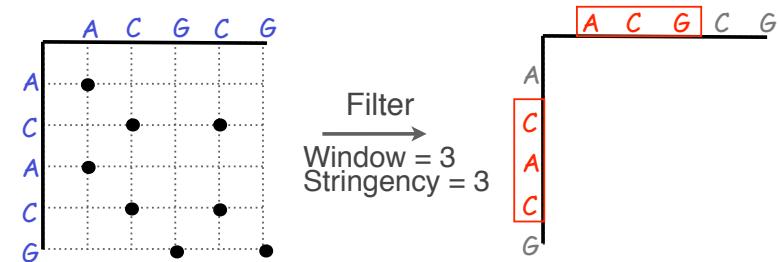
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
 - You have to choose window size and stringency



Dot plots: window size and match stringency

Solution: use a window and a threshold

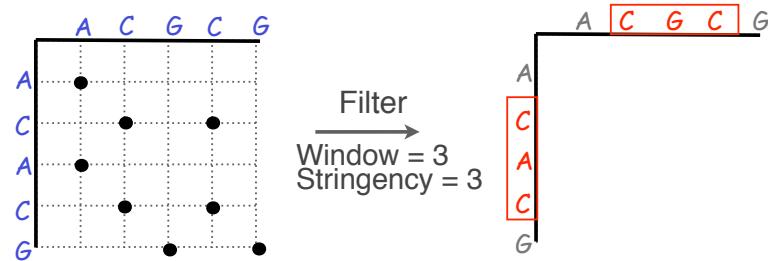
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
 - You have to choose window size and stringency



Dot plots: window size and match stringency

Solution: use a window and a threshold

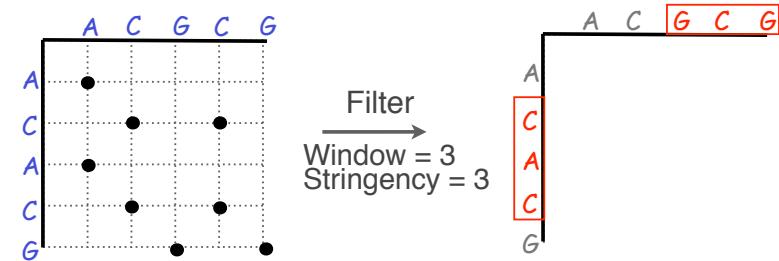
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
 - You have to choose window size and stringency



Dot plots: window size and match stringency

Solution: use a window and a threshold

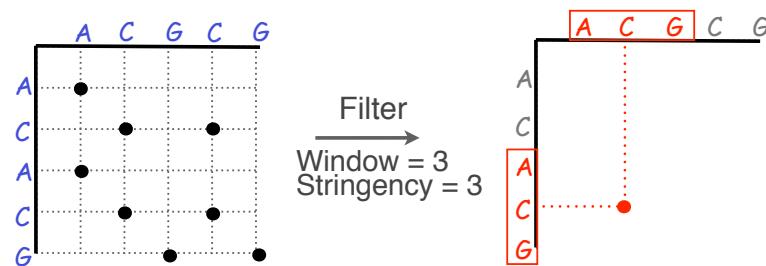
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
 - You have to choose window size and stringency



Dot plots: window size and match stringency

Solution: use a window and a threshold

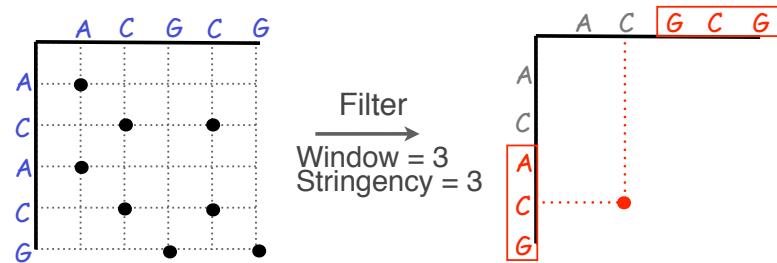
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
 - You have to choose window size and stringency



Dot plots: window size and match stringency

Solution: use a window and a threshold

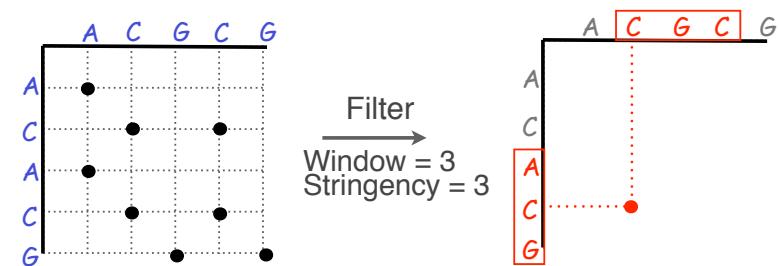
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
 - You have to choose window size and stringency



Dot plots: window size and match stringency

Solution: use a window and a threshold

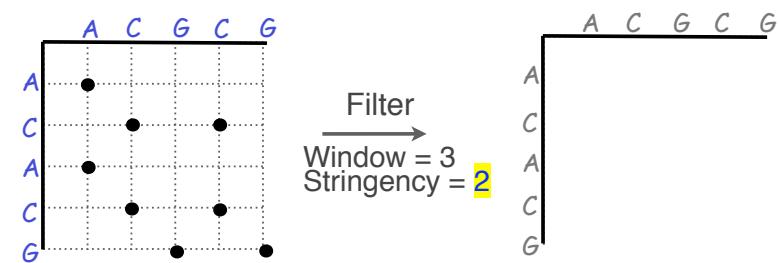
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
 - You have to choose window size and stringency



Dot plots: window size and match stringency

Solution: use a window and a threshold

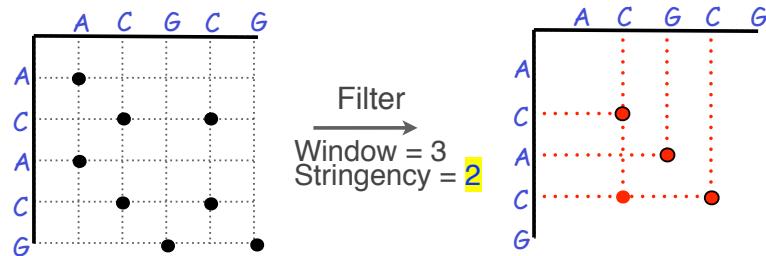
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
 - You have to choose window size and stringency



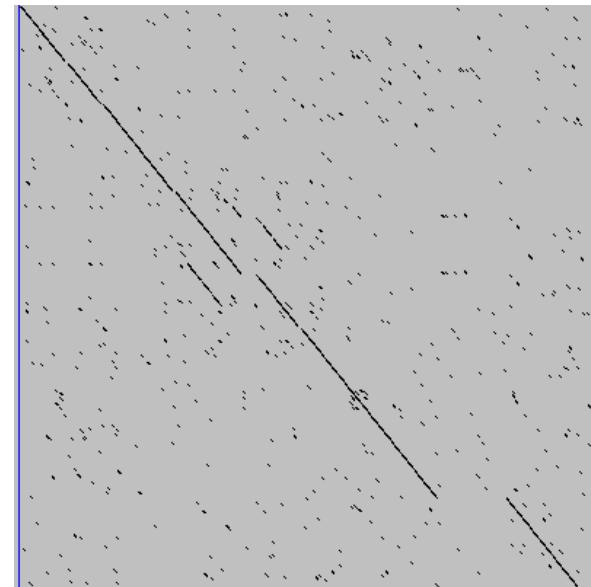
Dot plots: window size and match stringency

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



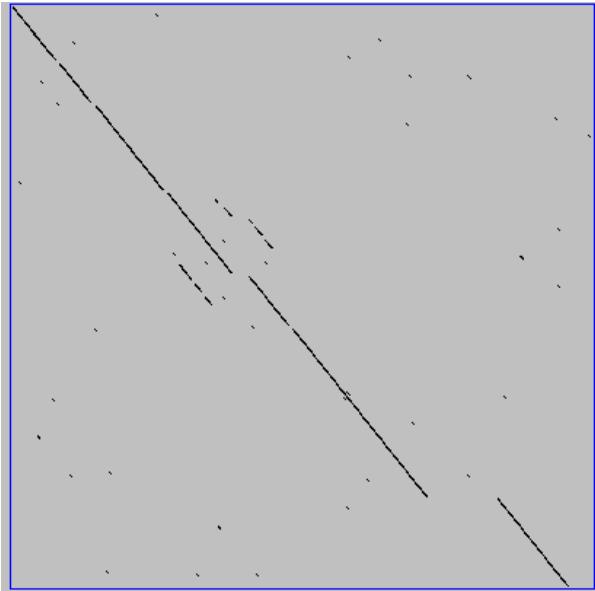
Window size = 5 bases



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



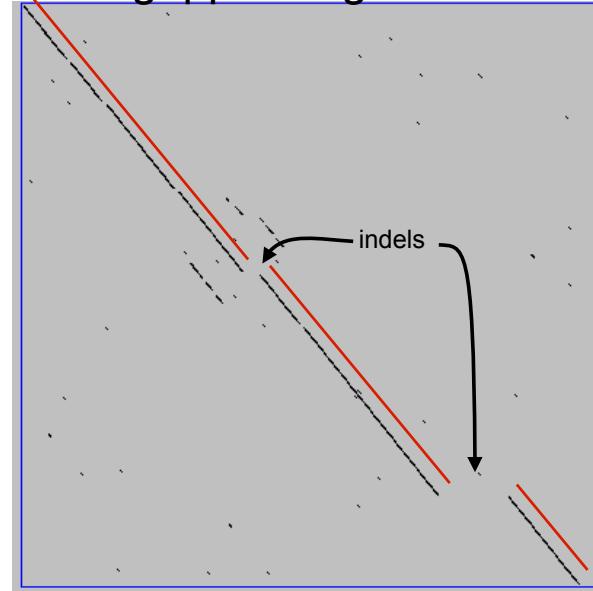
This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

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.

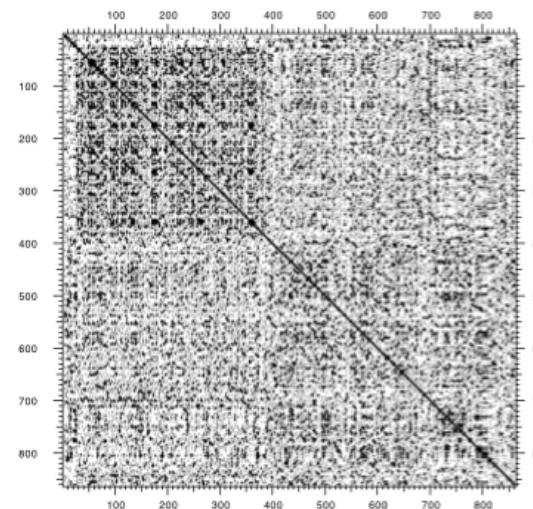
Downward or rightward paths represent **insertion** or **deletions** (gaps in one sequence or the other).

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

Uses for dot matrices

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

Repeats

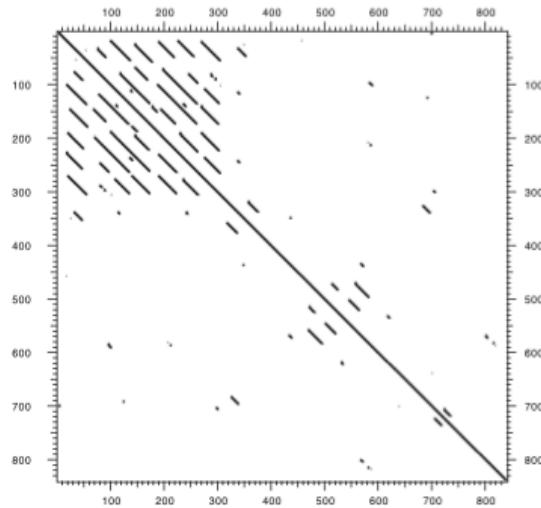


Human LDL receptor
protein sequence
(Genbank P01130)

W = 1
S = 1

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

Repeats



Human LDL receptor
protein sequence
(Genbank P01130)

W = 23
S = 7

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

Your Turn!

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

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

A screenshot of a web application titled "BGGN-213: Dot Plot Comparison of Two Sequences". The interface includes a "Dot Plot Parameters" section with sliders for "Window Size" (set to 3), "Moving window step size" (set to 3), and "Match stringency" (set to 2). To the right are two dot plots: "Protein Dot Plot" and "DNA Dot Plot", both comparing "Sequence 2" against "Sequence 1". The DNA plot shows significantly more dots than the protein plot, indicating higher stringency or a larger window size. Below the plots is a "Questions for discussion:" section with three bullet points.

Dot Plot Parameters

- Window Size: 3
- Moving window step size: 3
- Match stringency: 2

Protein Dot Plot
wszie = 3 wstep = 3, rmatch = 2

DNA Dot Plot
wszie = 3 wstep = 3, rmatch = 2

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" lower than "Window size" yield and who?

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

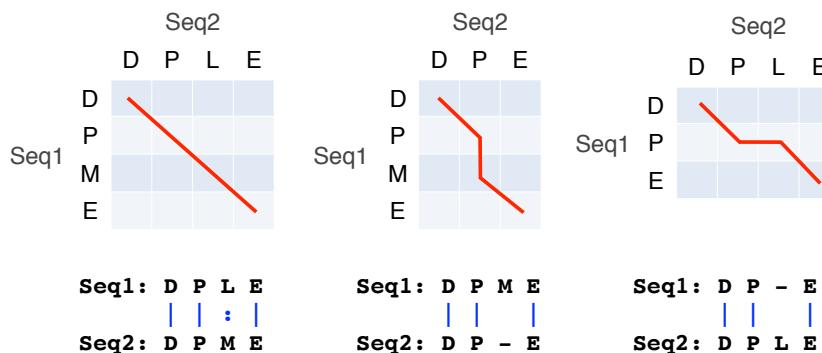
The Dynamic Programming Algorithm

- 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:
 - (1) setting up a 2D-grid (or alignment matrix),
 - (2) scoring the matrix, and
 - (3) identifying the optimal path through the matrix



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

Scoring the alignment matrix

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

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

Scoring the alignment matrix

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

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

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

Seq1: DPME
Seq2: -----

Scoring the alignment matrix

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

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

The diagram illustrates the three possible paths to the cell $S(i, j)$ from the surrounding cells $S(i-1, j-1)$, $S(i-1, j)$, and $S(i, j-1)$. Path 1 is indicated by a top-left arrow from $S(i-1, j-1)$ to $S(i, j)$. Path 2 is indicated by a top-right arrow from $S(i-1, j)$ to $S(i, j)$. Path 3 is indicated by a bottom-left arrow from $S(i, j-1)$ to $S(i, j)$.

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

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

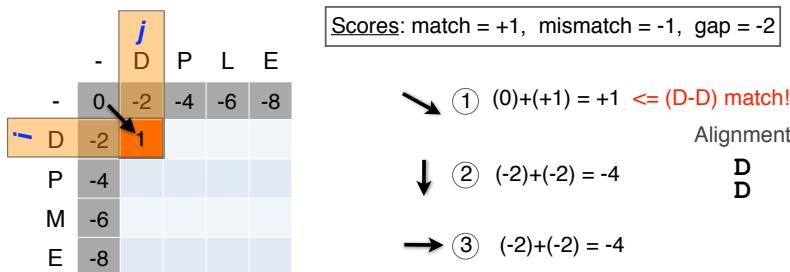
$S(i, j) = \text{Max} \left\{ \begin{array}{l} S(i-1, j-1) + (\text{mis})\text{match} \\ S(i-1, j) + \text{gap penalty} \\ S(i, j-1) + \text{gap penalty} \end{array} \right\}$

Diagram illustrating the calculation of the score $S(i, j)$ based on the three possible paths:

- Path 1: $S(i-1, j-1) + (\text{mis})\text{match}$ (indicated by arrow 1)
- Path 2: $S(i-1, j) + \text{gap penalty}$ (indicated by arrow 2)
- Path 3: $S(i, j-1) + \text{gap penalty}$ (indicated by arrow 3)

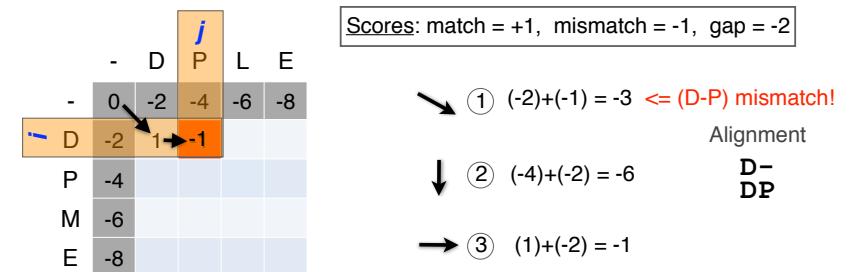
Scoring the alignment matrix

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



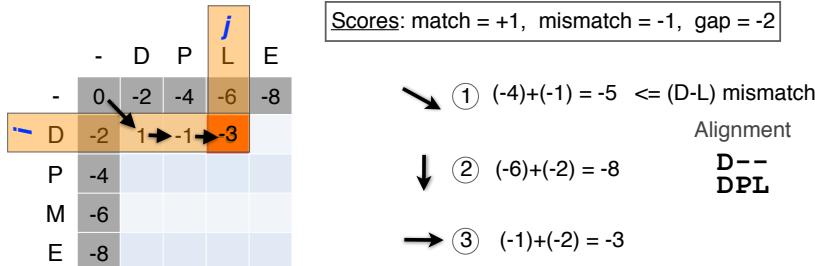
Scoring the alignment matrix

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



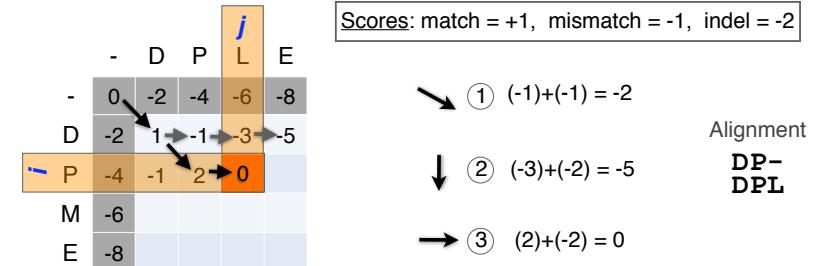
Scoring the alignment matrix

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



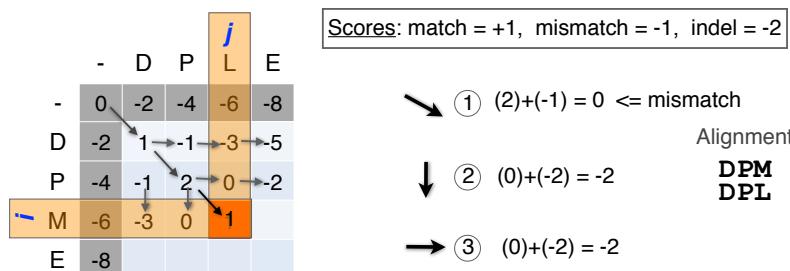
Scoring the alignment matrix

- For the highlighted cell, the corresponding score ($S_{i,j}$) refers to the score of the optimal alignment of the first i characters from sequence1, and the first j characters from sequence2.



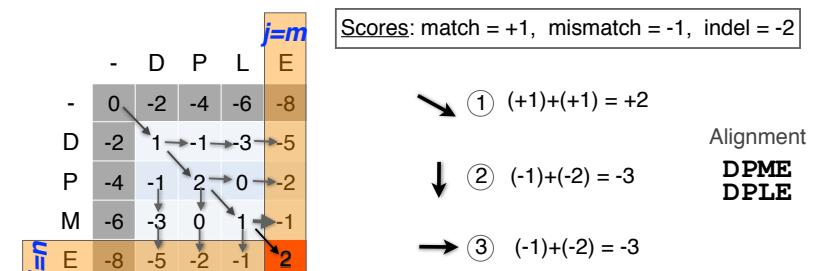
Scoring the alignment matrix

- At each step, the score in the current cell is determined by the scores in the neighboring cells
 - The maximal score and the direction that gave that score is stored



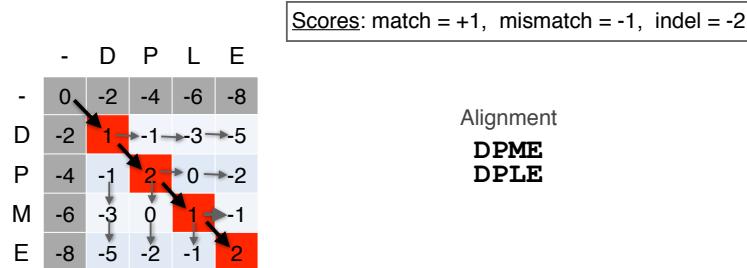
Scoring the alignment matrix

- The score of the best alignment of the entire sequences corresponds to $S_{n,m}$
 - (where n and m are the length of the sequences)



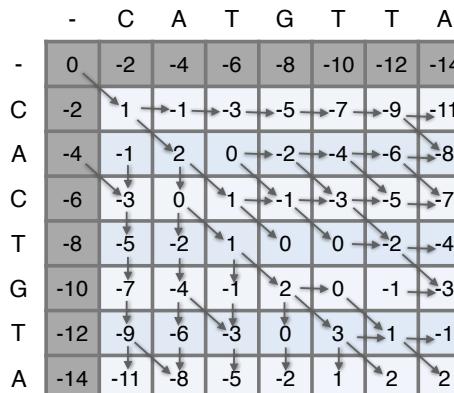
Scoring the alignment matrix

- To find the best alignment, we retrace the arrows starting from the bottom right cell
 - N.B. The optimal alignment score and alignment are dependent on the chosen scoring system



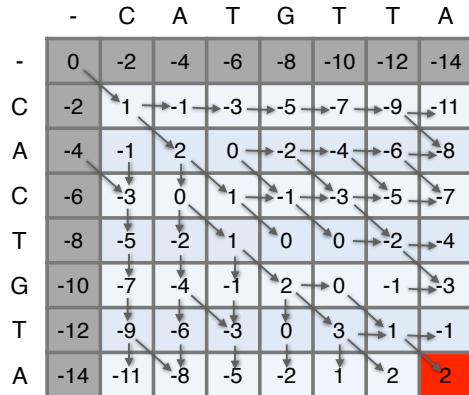
Questions:

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



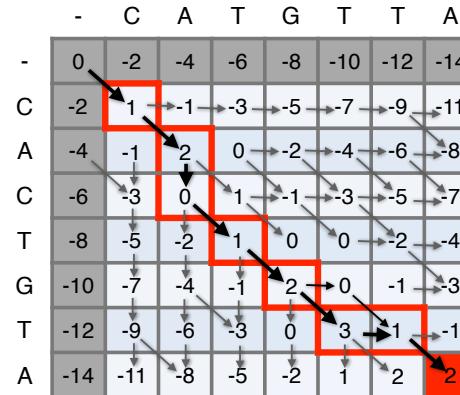
Questions:

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



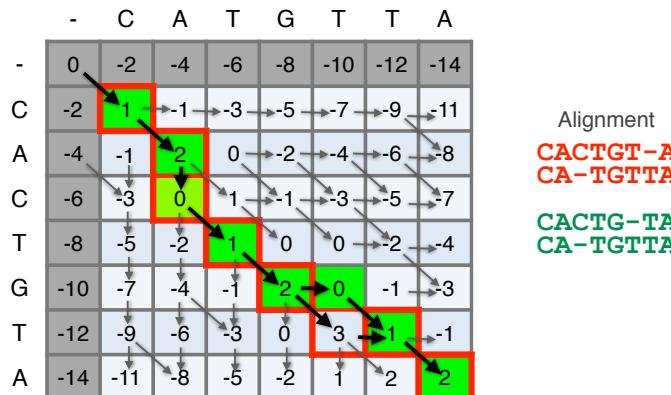
Questions:

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



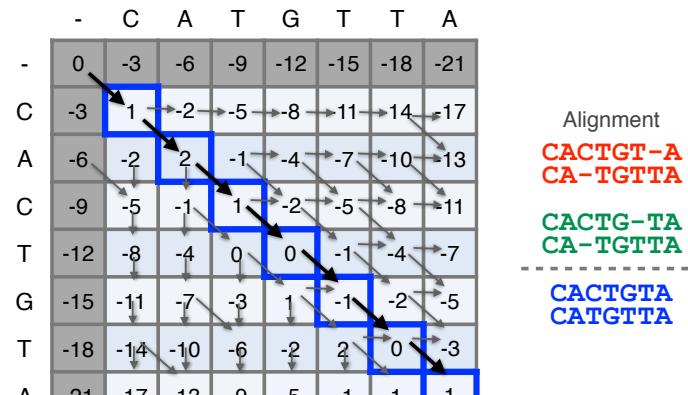
More than one alignment possible

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



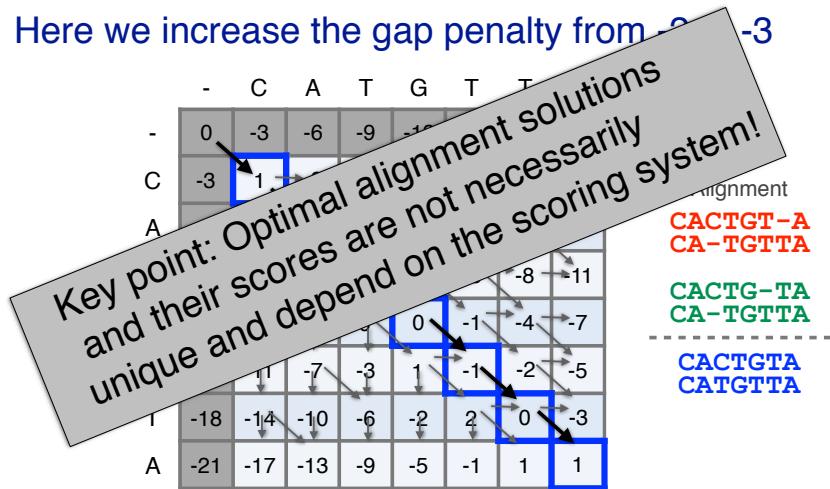
The alignment and score are dependent on the scoring system

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



The alignment and score are dependent on the scoring system

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



Your Turn!

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

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

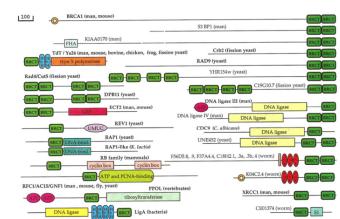
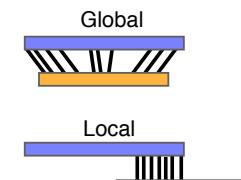
	A	G	T	T	C
A	0				
T					
T					
G					
C					

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

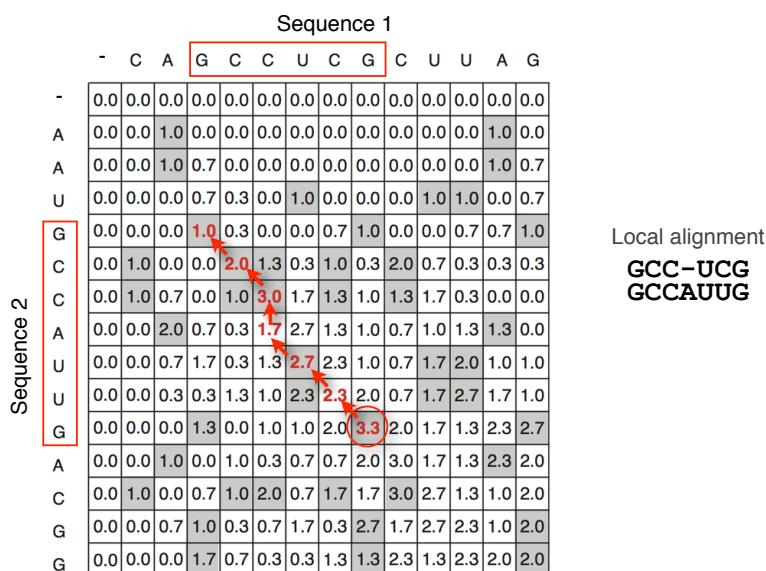
- Needleman-Wunsch is a global alignment algorithm
 - Resulting alignment spans the complete sequences end to end
 - This is appropriate for closely related sequences that are similar in length
- For many practical applications we require local alignments
 - Local alignments highlight sub-regions (e.g.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



Local alignment: Definition

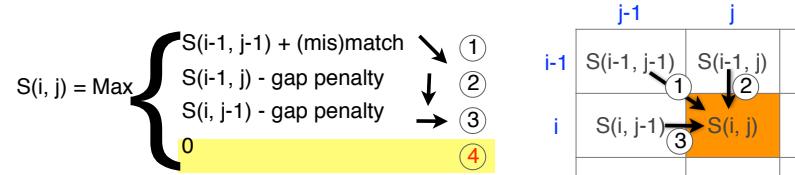
- Smith & Waterman proposed simply that a local alignment of two sequences allow arbitrary-length segments of each sequence to be aligned, with no penalty for the unaligned portions of the sequences. Otherwise, the score for a local alignment is calculated the same way as that for a global alignment

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



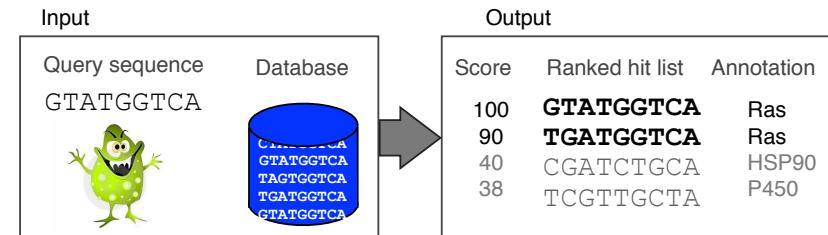
The Smith-Waterman algorithm

- 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



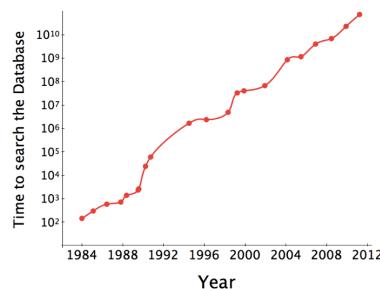
Local alignments can be used for database searching

- 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

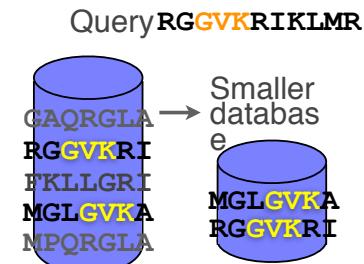
- Due to the rapid growth of sequence databases, search algorithms have to be both efficient and sensitive
 - Time to search with SW is proportional to $m \times n$ (m is length of query, n is length of database), too slow for large databases!



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

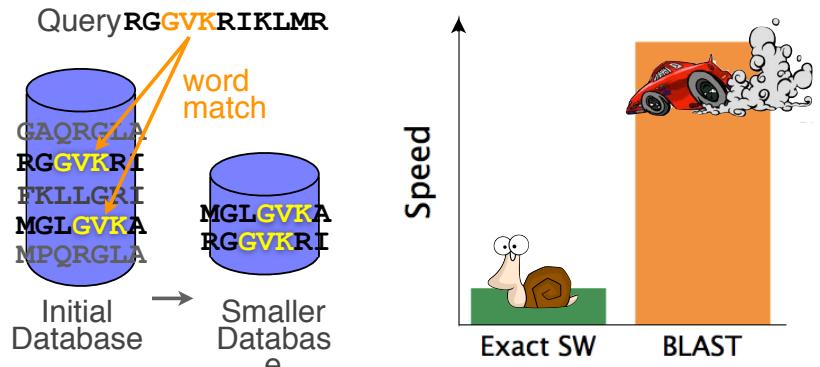
Rapid, heuristic versions of Smith–Waterman: BLAST

- 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 that is popular because it is faster
 - BLAST finds regions of local sequence similarity
 - BLAST uses a “word pair” search by scanning sequence pairs that contain an initial word pair match”
- “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)
- Trade off some sensitivity in exchange for speed
- In contrast to SW, BLAST is not guaranteed to find optimal alignments

- 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 ($w=3$)

RGGVKRI Query sequence
RGG
GGV
GVK
VKR
KRI

generate list of $w=3$ words for query

Blast

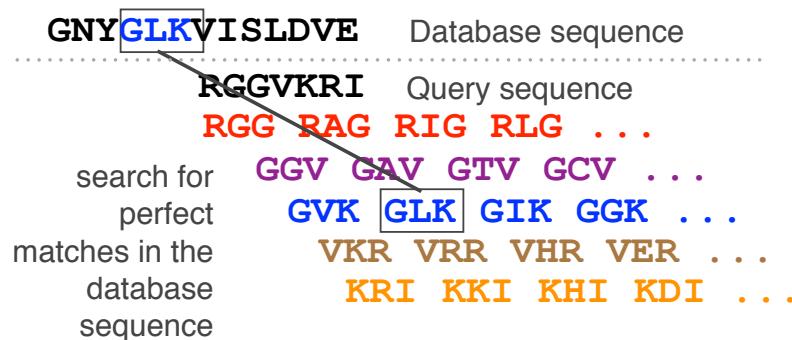
- Phase 2: expand word pairs to include those similar to query (defined as those above a similarity threshold to original word, i.e. match scores in substitution matrix)

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

extend list of words similar to query

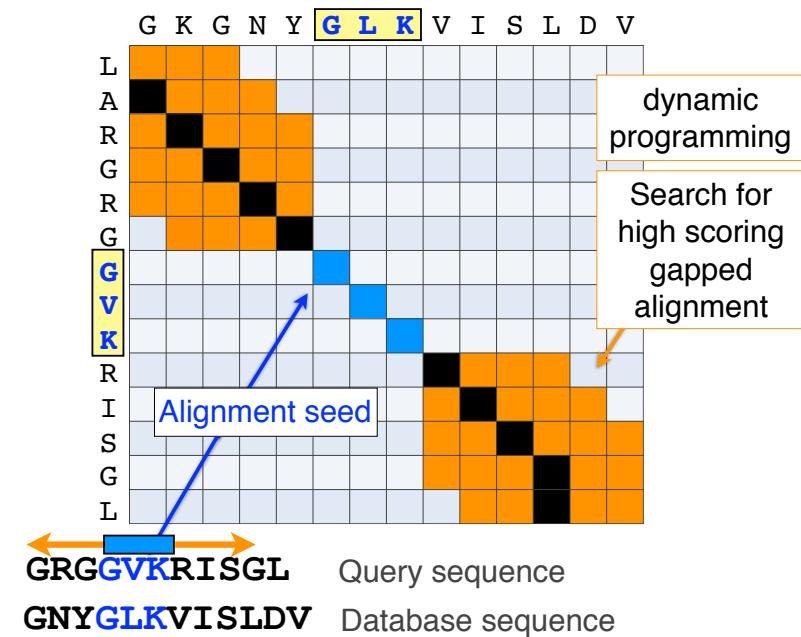
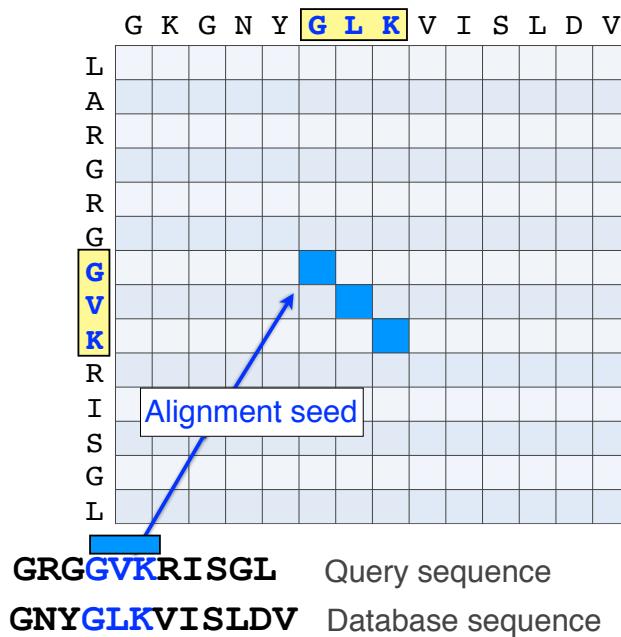
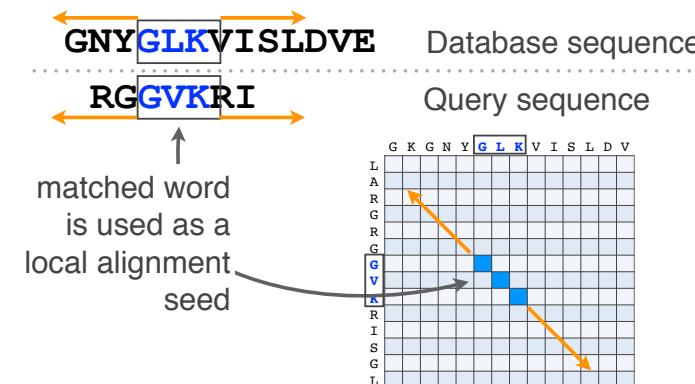
Blast

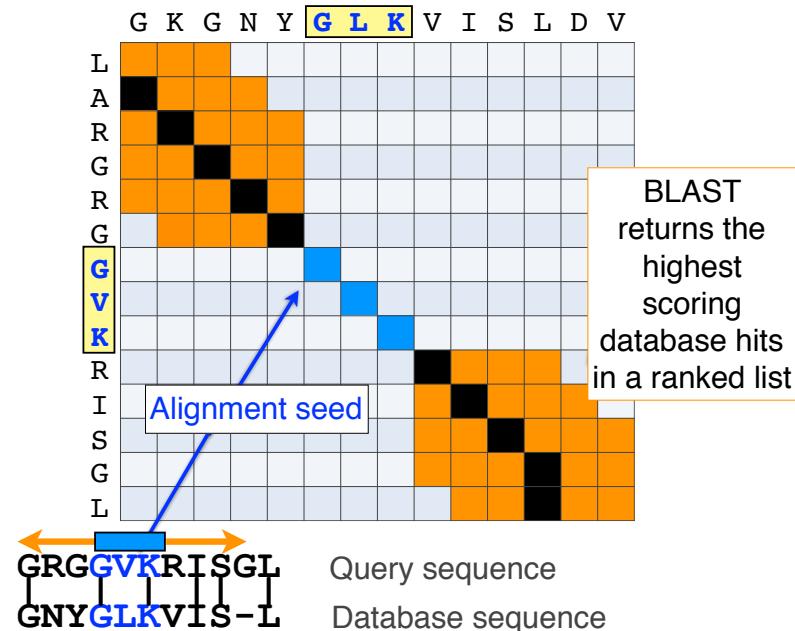
- Phase 3: a database is scanned to find sequence entries that match the compiled word list



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

130

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

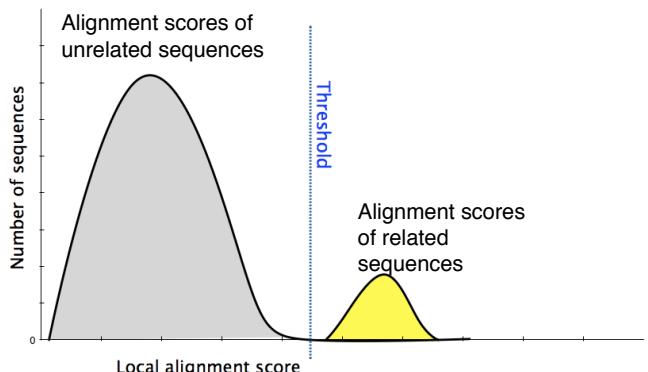
131

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

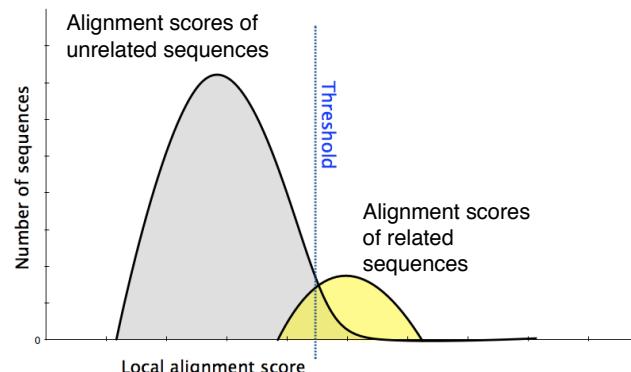
132

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



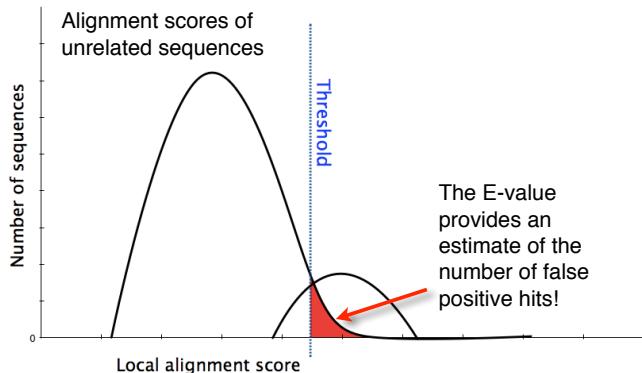
133

- 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



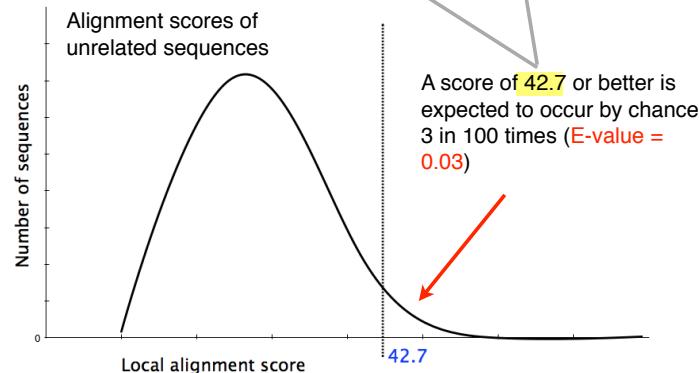
134

- 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



135

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



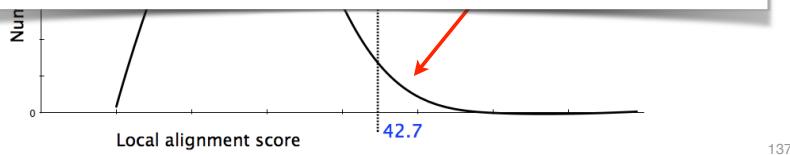
136

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
Kit5b 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>



137

FOR NEXT CLASS...

Check out the online:

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

Your Turn!

Hands-on worksheet **Sections 4 (& 5)**

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

Homework Grading

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

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

REFERENCE SLIDES...

Additional reference slides for the motivated student

142

Step 1: Choose your sequence

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

The screenshot shows the NCBI Protein search page. At the top, there's a search bar labeled "Search: Protein". Below it, under "Display Settings", there's a checked checkbox for "FASTA" which is circled in red. The main content area displays a protein sequence for "hemoglobin subunit beta [Homo sapiens]" with its accession number NP_000509.1. The sequence itself is also circled in red. At the bottom of the page, there's some detailed sequence information and a link to "GenPept Graphics".

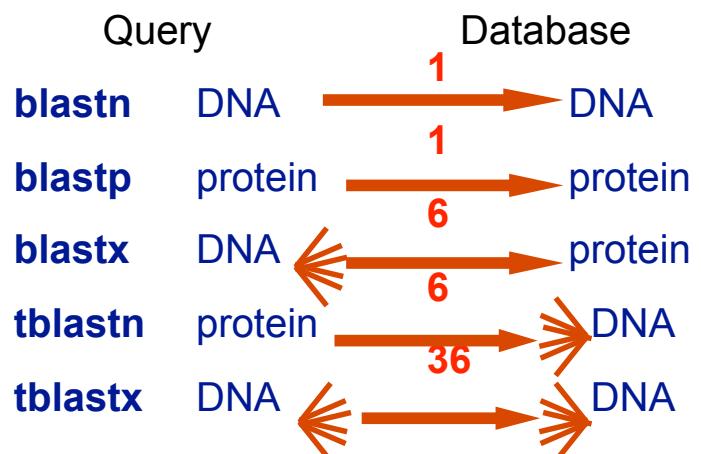
143

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”

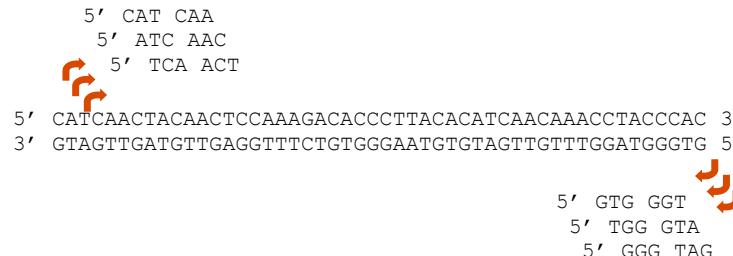
142

Step 2: Choose the BLAST program



144

DNA potentially encodes six proteins



145

Protein BLAST: search protein databases using a protein query
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) >gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSATAVLWGVNVNDEVGGEALGRLLVVYPWTQRFESFGDLSTPAVMGNPKVKAHGK
KVLCGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLCNVLVCVLAHHFCKEFTPVQAAYQK
VVAGVANALAHKYH

Or, upload file no file selected
Job Title
Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)
Organism:
Optional: Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
 Models (XMP) Uncultured/environmental sample sequences

Exclude: Optional:
Entrez Query:
Optional: Enter an Entrez query to limit search

Program Selection

Algorithm: blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm

BLAST
 Show results in a new window

[Algorithm parameters](#)

146

Step 3: Choose the database

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

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

nucleotide databases

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

protein databases

147

Protein BLAST: search protein databases using a protein query
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) >gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSATAVLWGVNVNDEVGGEALGRLLVVYPWTQRFESFGDLSTPAVMGNPKVKAHGK
KVLCGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLCNVLVCVLAHHFCKEFTPVQAAYQK
VVAGVANALAHKYH

Or, upload file no file selected
Job Title
Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)
Organism:
Optional: Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
 Models (XMP) Uncultured/environmental sample sequences

Exclude: Optional:
Entrez Query:
Optional: Enter an Entrez query to limit search

Program Selection

Algorithm: blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm

BLAST
 Show results in a new window

[Algorithm parameters](#)

148

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

149

Step 4: Optional parameters

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

Results page

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

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/BLAST blastp suite/Formatting Results - FVGUTMRZ013

Edit and Resubmit Save Search Strategies > Formatting options > Download Change the result display back to traditional format

YouTube Learn about the enhanced report Blast report description

gi|4504349|ref|NP_000509.1| hemoglobin

Query ID: gi|4504349|ref|NP_000509.1| hemoglobin subunit beta (Homo sapiens)

Database Name: nr

Description: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects

Molecule type: amino acid

Query Length: 147

Program: BLASTP 2.2.27+ > Citation

Other reports: > Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment]

Now DELTA-BLAST, a more sensitive protein-protein search Go

Graphic Summary

Show Conserved Domains

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

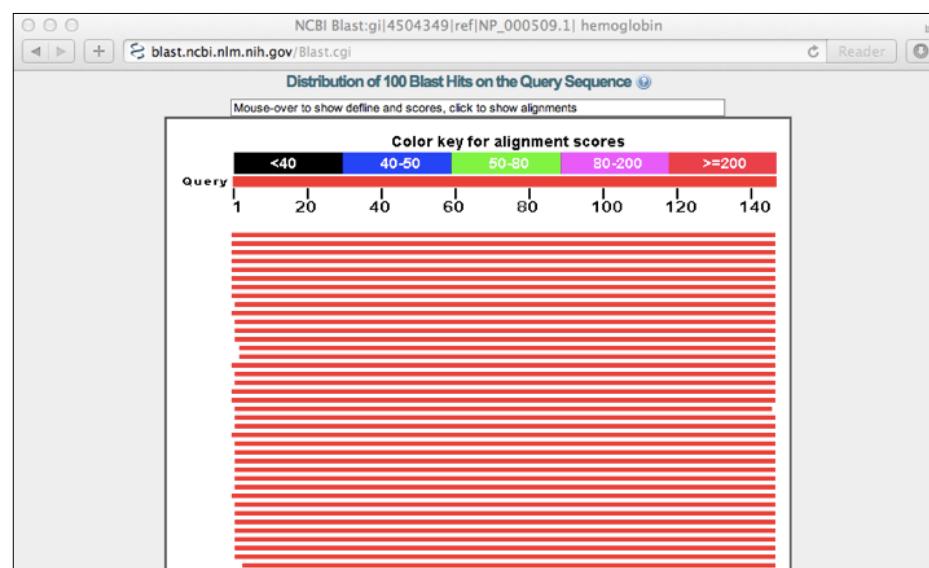
Query seq. Specific hits Superfamilies

25 50 75 100 125 147

heme-binding site globin globin_like superfamily

Distribution of 100 Blast Hits on the Query Sequence

Further down the results page...



Further down the results page...

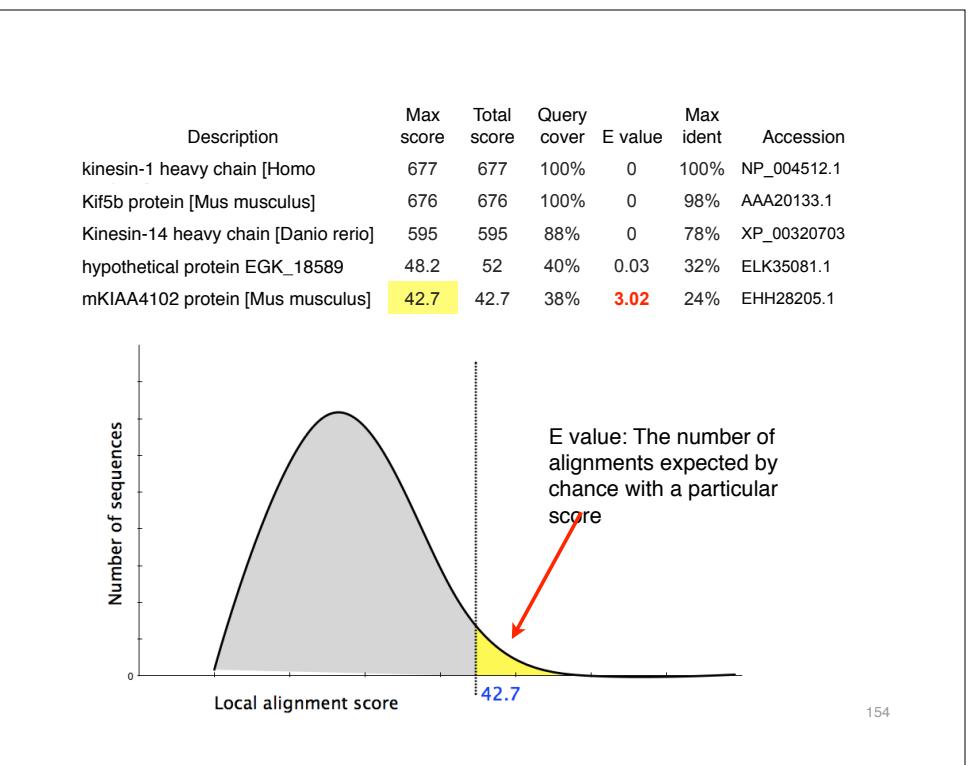
NCBI Blast:gi|4504349|ref|NP_000509.1| hemoglobin

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Max ident	Accession
hemoglobin beta [synthetic construct]	301	301	100%	9e-103	100%	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=Hb	300	300	100%	4e-102	99%	P02024.2
beta globin chain variant [Homo sapiens]	299	299	100%	5e-102	99%	AAN84548.1
beta globin [Homo sapiens] >gb AAZ39781.1 beta globin [Homo sapiens] >gb AAZ39782.1	299	299	100%	5e-102	99%	AAZ39780.1
beta-globin [Homo sapiens]	299	299	100%	5e-102	99%	ACU56984.1
hemoglobin beta chain [Homo sapiens]	299	299	100%	6e-102	99%	AAD19696.1
Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound A	298	298	99%	9e-102	100%	1COH_B
hemoglobin beta subunit variant [Homo sapiens] >gb AA88054.1 beta-globin [Homo sa	298	298	100%	1e-101	99%	AAF00489.1
Chain B, Human Hemoglobin D Los Angeles: Crystal Structure >pdb 2YRSID Chain D, H	298	298	99%	2e-101	99%	2YRS_B
Chain B, High-Resolution X-Ray Study Of Deoxy Recombinant Human Hemoglobins Syn	297	297	99%	3e-101	99%	1DXU_B
Chain B, Analysis Of The Crystal Structure, Molecular Modeling And Infrared Spectroscop	297	297	99%	3e-101	99%	1HDB_B



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 given the length of my query and the size of the database
 - $E = Kmne^{-S}$
 - 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

hemoglobin subunit beta [Homo sapiens]

Sequence ID: ref|NP_000509.1| Length: 147 Number of Matches: 1

► See 84 more title(s)

Range 1: 1 to 147 GenPept Graphics

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 KVHLITPEEKSAVTALNGKVNVDDEVGGAEALGRLRLLVYPWTQRFESFGDLSTPDAVMGNPK 60
Sbjct 1 KVHLITPEEKSAVTALNGKVNDEVGGAEALGRLRLLVYPWTQRFESFGDLSTPDAVMGNPK 60

Query 61 VKAHGKVKVLGAFSDGLAHLNDLNLKGTFATLSSELHCDKLHVDPENFRLLGNVLVCVLAHHIFG 120
Sbjct 61 VKAHGKVKVLGAFSDGLAHLNDLNLKGTFATLSSELHCDKLHVDPENFRLLGNVLVCVLAHHIFG 120

Query 121 KEFTPFPVQAAYKVVAAGVANALAHKYI 147
Sbjct 121 KEFTPFPVQAAYKVVAAGVANALAHKYI 147

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

RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta chain

Sequence ID: sp|P02024.2|HB_GORG Length: 147 Number of Matches: 1

Range 1: 1 to 147 GenPept Graphics

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

Related Information

Different output formats are available

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

Basic Local Alignment Search Tool

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.

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

Query	Score	Sequence
AAX37051	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
AAX29557	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
NP_000509	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
P02024	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
AAN84548	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
AAZ39780	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
ACU56984	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
AAD19696	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
1COH_B	1	VHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
AAF00489	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
2YRS_B	1	VHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
IDXU_B	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
1HDB_B	1	VHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
IDXV_B	2	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
3KMF_C	2	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
AAE68978	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
1NQP_B	1	VHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
1K1K_B	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
AAN11320	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
XP_002822173	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
IYB5_B	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
IYE0_B	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
1O1O_B	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
CAA23759	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
IYE2_B	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
IY5F_B	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
IA00_B	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
IHBS_B	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
IABY_B	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
1C9Y_B	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK

... and alignments with dots for identities

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

Query	Score	Sequence
AAX37051	1	MVHLTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK
AAX29557	1
NP_000509	1
P02024	1
AAN84548	1
AAZ39780	1K.....
ACU56984	1
AAD19696	1
1COH_B	1L.....
AAF00489	1
2YRS_B	1
IDXU_B	1
1HDB_B	1
IDXV_B	2
3KMF_C	2
AAE68978	1
1NQP_B	1K.....
1K1K_B	1V.....
AAN11320	1
XP_002822173	1
IYB5_B	1
IYE0_B	1A.....
1O1O_B	1V.....X.....
CAA23759	1F.....
IYE2_B	1
IY5F_B	1
IA00_B	1Y.....

Common problems

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

How to handle too many results

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

161

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)

162