

Recap From Last Time:

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

Today's Menu

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

Primary, secondary & composite databases

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

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

DATABASE VIGNETTE

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

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

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

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

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

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

The screenshot shows the NCBI homepage with the search term 'ras' entered in the search bar. The page includes links to various databases like GenBank, RefSeq, and PubMed, as well as sections for 'Welcome to NCBI', 'Get Started', and 'Genotypes and Phenotypes'.

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 **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**

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

Databases	Count	Description
Literature	Books: 1,677, MeSH: 402, NLM Catalog: 228, PubMed: 54,672, PubMed Central: 96,114	books and reports, ontology used for PubMed indexing, books, journals and more in the NLM Collections, scientific & medical abstracts/ citations, full-text journal articles
Genes	EST: 3,985, GEO DataSets: 3,732, GEO Profiles: 1,822,789, HomoloGene: 696, PopSet: 2,254, UniGene: 4,770	expressed sequence tag sequences, functional genomics studies, gene expression and molecular abundance profiles, homologous gene sets for selected organisms, sequence sets from phylogenetic and population studies, clusters of expressed transcripts
Health	ClinVar: 759, dbGaP: 120, GTR: 1,879	human variations of clinical significance, genotype/phenotype interaction studies, genetic testing registry
Proteins	None	None

Results: 1 to 20 of 85633

Name/Gene ID	Description	Location	Aliases
ras (ID: 10412)	resistance to audogenic seizures [Mus musculus (house mouse)]	asr	
ras (ID: 43873)	rasberry [Drosophila melanogaster (fruit fly)]	Chromosome X, NC_004354, CG11485, CG1799, Dmel:CG1799, EPI(X)1063,	

Results: 1 to 20 of 1126

Name/Gene ID	Description	Location	Aliases
NRAS (ID: 4683)	neuroblastoma RAS viral (v-ras) oncogene [Homo sapiens (human)]	Chromosome 1, NC_000012.11 (114704464..114716864), ALPS4, CMNS, N-ras, NCMS1, NS8, NRAS	RPS-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS8, NRAS
KRAS (ID: 3645)	Kirsten rat sarcoma viral oncogene [Homo sapiens (human)]	Chromosome 12, NC_000012.12 (2520824..25250923), RAS2, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS	C-K-RAS, CFC2, K-RAS2, K-RAS4B, K-RAS1, KRAS2, NS

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)**

Results: 1 to 20 of 1126

Name/Gene ID	Description	Location	Aliases
NRAS (ID: 4683)	neuroblastoma RAS viral (v-ras) oncogene [Homo sapiens (human)]	Chromosome 1, NC_000012.11 (114704464..114716864), ALPS4, CMNS, N-ras, NCMS1, NS8, NRAS	RPS-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS8, NRAS
KRAS (ID: 3645)	Kirsten rat sarcoma viral oncogene [Homo sapiens (human)]	Chromosome 12, NC_000012.12 (2520824..25250923), RAS2, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS	C-K-RAS, CFC2, K-RAS2, K-RAS4B, K-RAS1, KRAS2, NS

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-8407
See related: Ensembl:ENSG00000133703; HPRD_01817; MIM:190070;
Gene type: protein coding
RefSeq status: REVIEWED
Organism: Homo sapiens
Lineage: Eukaryota; Metazoa; Chordata; Craniota; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euchortoglires; Primates; Haplorhini; Catarrini; Hominoidea; Homo
Also known as: NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-
13

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

KRAS (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-8407
See related: Ensembl:ENSG00000133703; HPRD_01817; MIM:190070;
Gene type: protein coding
RefSeq status: REVIEWED
Organism: Homo sapiens
Lineage: Eukaryota; Metazoa; Chordata; Craniota; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euchortoglires; Primates; Haplorhini; Catarrini; Hominoidea; Homo
Also known as: NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-
14

Genomic context

Location: 12p12.1
Exon count: 6

Annotation release	Status	Assembly	Chr	Location
106	current	GRC:38 (GCF_000001405_26)	12	NC_000012.12 (25205246..25250923, complement)
105	previous assembly	GRC:37.p13 (GCF_000001405_25)	12	NC_000012.11 (2536190..25403670, complement)

Chromosome 12 - NC_000012.12

Genomic regions, transcripts, and products

Genomic Sequence: NC_000012.12 chromosome 12 reference GRC:38 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank Nucleotide

15

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

KRAS (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-8407
See related: Ensembl:ENSG00000133703; HPRD_01817; MIM:190070;
Gene type: protein coding
RefSeq status: REVIEWED
Organism: Homo sapiens
Lineage: Eukaryota; Metazoa; Chordata; Craniota; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euchortoglires; Primates; Haplorhini; Catarrini; Hominoidea; Homo
Also known as: NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-
16

Gene Ontology Provided by GOA

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

Process	Evidence Code	Pubs
Fc epsilon receptor 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	

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

17

UniProt-GOA

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

UniProt is a member of the GO Consortium .

Menu

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

18

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

19

GO Ontologies

- There are three ontologies in GO:

► Biological Process

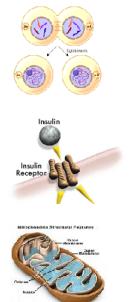
A commonly recognized series of events
e.g. cell division, mitosis,

► 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



30

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

↓ Scroll down to
↓ **UniProt** link

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

The screenshot shows a browser window with the URL www.ncbi.nlm.nih.gov/gene/30469/gene-ontology. The page displays gene information for P01116.1, including its UniProtKB Link ([UniProtKB/Swiss-Prot:P01116](#)). A red box highlights this link. Below the main content, there's a section titled "Additional links" which includes the UniProtKB link again. A red arrow points from the text "Scroll down to Very bottom for UniProt link" to the UniProtKB link in the "Additional links" section.

 Scroll down to
Very bottom for
UniProt link

UniProt will detail much more information for protein coding genes

P01116 - RASK_HUMAN

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

Display None

FUNCTION

NAMES & TAXONOMY

STRUCTURE & LOCATION

PATHOLOGY

PTM / PROCESSING

EXPRESSION

INTERACTION

STRUCTURE

FAMILY & DOMAINS

SEQUENCES

CROSS REFERENCES

Function

Responsible for GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23698361, PubMed:22711838), 2 Publications, 2 Curates

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 2 Publications			
Nucleotide binding ²	29 - 35	7	GTP ² 2 Publications			
Nucleotide binding ³	59 - 60	2	GTP ² 2 Publications			

UniProt will detail much more information for protein coding genes

UniProt will detail much more information for protein coding genes

P01116 - RASK_HUMAN

Protein | GTPase KRas
Gene | KRAS
Organism | Homo sapiens (Human);
Status | Reviewed - Experimental evidence at protein level¹

Display | Home

FUNCTION | Name & taxonomy | Search location | Pathobiology | PTM / Processing | Expression | Interaction | Structure | Family & Domains | Sequences (2) | Cross-references

[View FASTA file format](#) [Help video](#)

Function

Ras protein bind GDP/GTP and promote proliferation (PubMed:23698361, 8)

Enzyme regulation

Alternates between an inactive form nucleotide-exchange factor (GEF) promotes exchange of bound GDP by GTP. [Full publication](#)

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding ¹	10 - 18	9	GTP			
Nucleotide binding ²	29 - 35	7	GTP			
Nucleotide binding ³	59 - 60	2	GTP			

UniProt will detail much more information for protein coding genes

[View FASTA](#)

UniProt will detail much more information for protein coding genes

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... ?

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... ?

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... ?

Pfam: Family: Kinesin (PF00225) <http://pfam.janelia.org/family/kinesin#tabview=tab9>

HHMI
janelia farm
research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam
keyword search | Go

Family: Kinesin (PF00225)

Structures

For those sequences which have a structure in the Protein DataBank[®], we use the mapping between UniProt[®], PDB and Pfam coordinate systems from the PDBe[®] 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
ABBK01_GIALA	11 - 335	2vga	A	11 - 335	Jmol AstexViewer SPICE
			B	11 - 335	Jmol AstexViewer SPICE
CENPE_HUMAN	12 - 329	115c	A	12 - 329	Jmol AstexViewer SPICE
			B	12 - 329	Jmol AstexViewer SPICE
KAR3_YEAST	392 - 723	1pt	A	392 - 723	Jmol AstexViewer SPICE
		1pu	A	392 - 723	Jmol AstexViewer SPICE
		1px	A	392 - 723	Jmol AstexViewer SPICE
		1pw	A	392 - 723	Jmol AstexViewer SPICE
		3kar	A	392 - 723	Jmol AstexViewer SPICE
KIF13B_HUMAN	11 - 352	3qbj	A	11 - 352	Jmol AstexViewer SPICE
			B	11 - 352	Jmol AstexViewer SPICE
			C	11 - 352	Jmol AstexViewer SPICE
		1t6	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		1q0b	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		1x88	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
			A	24 - 359	Jmol AstexViewer SPICE

Jump to... ↴
enter ID/acc ↴

Pfam: Jmlol <http://pfam.janelia.org/structure/viewer?viewer=jmlol&id=3bfm>

Pfam Family: Kinesin (PF00225) Pfam: Jmlol

welcome trust sanger institute

PDB entry 3bfm

Your turn:
What can you find out about "eg5"

PDB	Chain	Start	End	UniProt	ID	Start	End	Pfam family	Colour
	A	49	368	KIF22_HUMAN		49	368	Kinesin (PF00225)	

[Close window](#)

Today's Menu

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

ALIGNMENT FOUNDATIONS

- **Why...**

- Why compare biological sequences?

- **What...**

- Alignment view of sequence changes during evolution (matches, mismatches and gaps)

- **How...**

- Dot matrices
- Dynamic programming
 - Global alignment
 - Local alignment
- BLAST heuristic approach

ALIGNMENT FOUNDATIONS

- **Why...**

- Why compare biological sequences?

- **What...**

- Alignment view of sequence changes during evolution (matches, mismatches and gaps)

- **How...**

- Dot matrices
- Dynamic programming
 - Global alignment
 - Local alignment
- BLAST heuristic approach

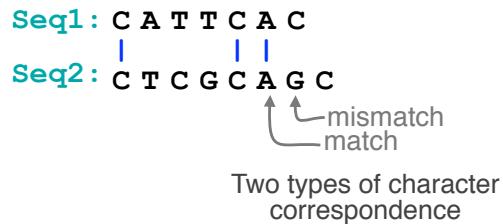
Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.

Seq1: C A T T C A C

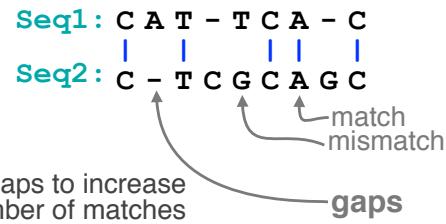
Seq2: C T C G C A G C

[Screencast Material]

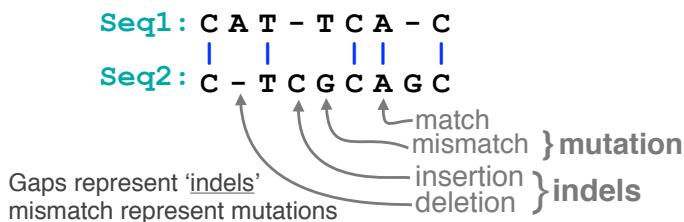
Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.



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



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



Why compare biological sequences?

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

Practical applications include...

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

Practical applications include...

- **Similarity searching of databases**
 - Protein structure prediction
- **Assembly of sequence reads** into a longer construct such 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)

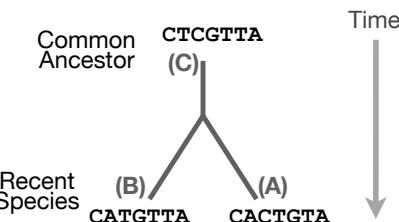
• How...

- 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.

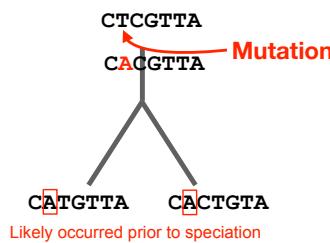
- Mutations/Substitutions
- Deletions
- Insertions



Mutations, deletions and insertions

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

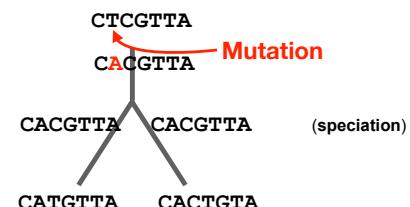
- **Mutations/Substitutions** $\text{CTCGTTA} \rightarrow \text{CACGTTA}$
- Deletions
- Insertions



Mutations, deletions and insertions

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

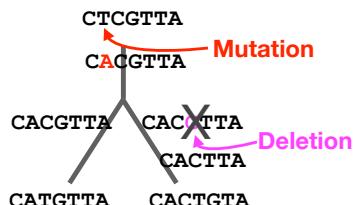
- Mutations/Substitutions $\text{CTCGTTA} \rightarrow \text{CACGTTA}$
- Deletions
- Insertions



Mutations, deletions and insertions

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

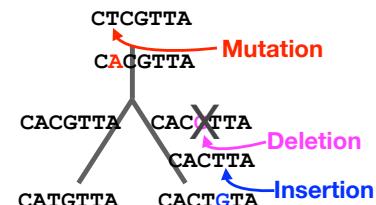
- Mutations/Substitutions
- **Deletions**
- Insertions



Mutations, deletions and insertions

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

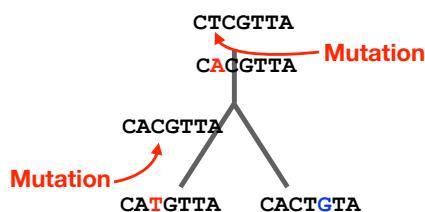
- Mutations/Substitutions
- Deletions
- **Insertions**



Mutations, deletions and insertions

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

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

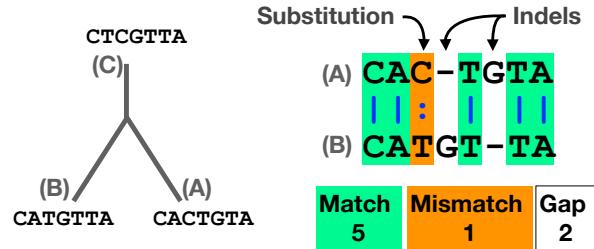


$\text{CTCGTTA} \rightarrow \text{CACGTAA}$
 $\text{CACGTAA} \rightarrow \text{CATGTAA}$

Alignment view

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

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



Alternative alignments

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

Q. Which of these 3 possible alignments is best?

1.

CACTGTA
|| : : : ||
CATGTAA

2.

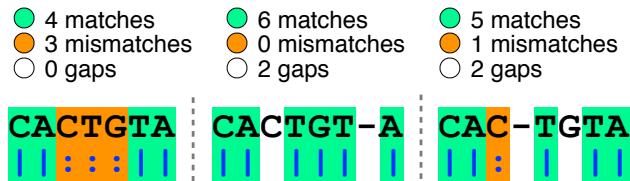
CACTGT-A
|| : : : ||
CA-TGTTA

3.

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

Alternative alignments

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



Scoring alignments

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

● 4 (+3) ● 3 (+1) ○ 0 (-1) = 15	● 6 (+3) ● 0 (+1) ○ 2 (-1) = 16	● 5 (+3) ● 1 (+1) ○ 2 (-1) = 14
---	---	---

CACTGTA
|| : : : ||
CATGTAA

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

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

Optimal alignments

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

● 4 matches ● 3 mismatches ○ 0 gaps	● 6 matches ● 0 mismatches ○ 2 gaps	● 5 matches ● 1 mismatch ○ 2 gaps
---	---	---

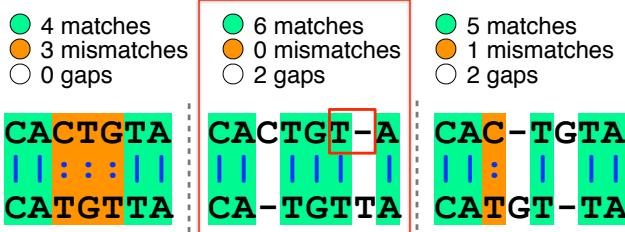
CACTGTA
|| : : : ||
CATGTAA

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

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

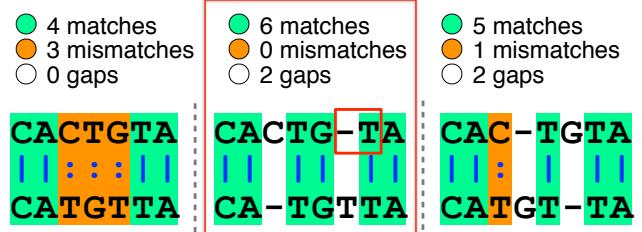
Optimal alignments

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



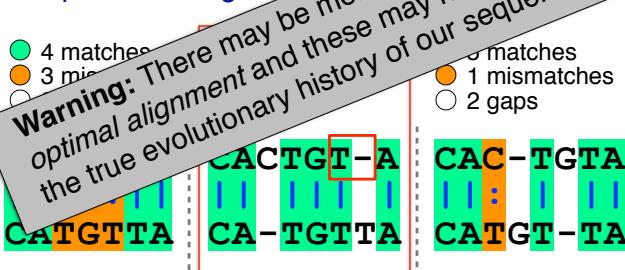
Optimal alignments

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



Optimal alignments

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



ALIGNMENT FOUNDATIONS

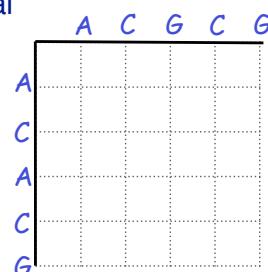
- Why...
 - Why compare biological sequences?
- What...
 - Alignment view of sequence changes during evolution (matches, mismatches and gaps)
- How...
 - Dot matrices
 - Dynamic programming
 - Global alignment
 - Local alignment
 - BLAST heuristic approach

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

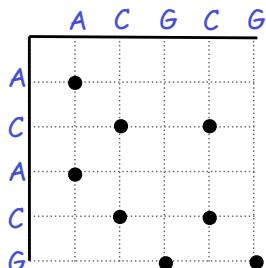
Dot plots: simple graphical approach

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



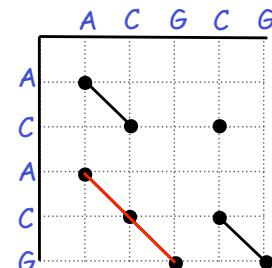
Dot plots: simple graphical approach

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



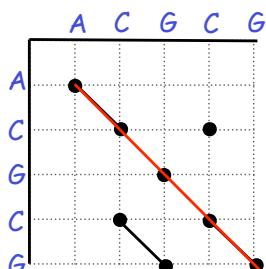
Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



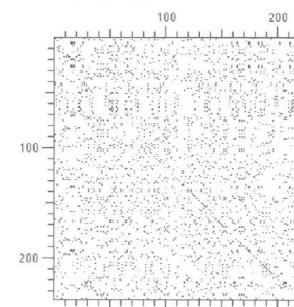
Dot plots: simple graphical approach

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



Dot plots: simple graphical approach

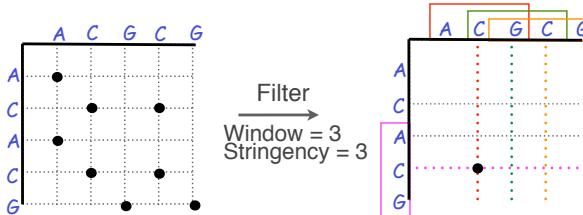
- Dot matrices for long sequences can be noisy



Dot plots: window size and match stringency

Solution: use a window and a threshold

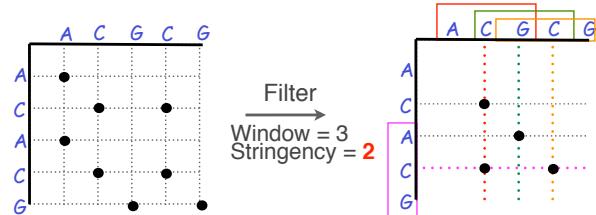
- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.
- You have to choose window size and stringency



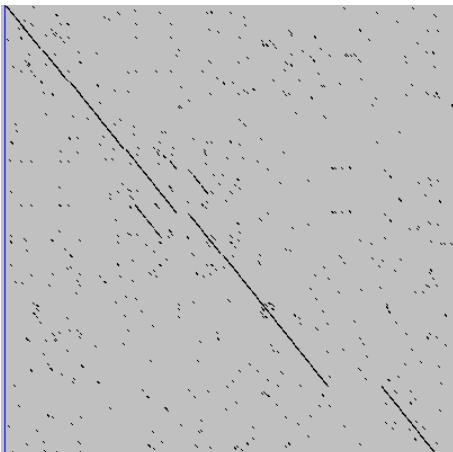
Dot plots: window size and match stringency

Solution: use a window and a threshold

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



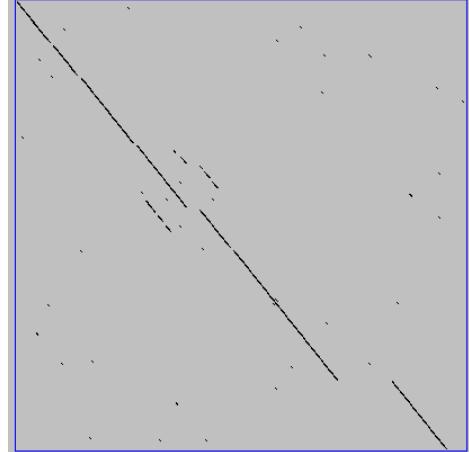
Window size = 5 bases



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

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

Window size = 7 bases

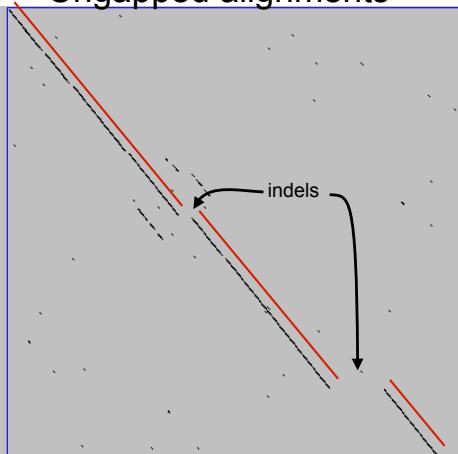


This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

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

Bigger window (kmer) fewer matches to consider

Ungapped alignments



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

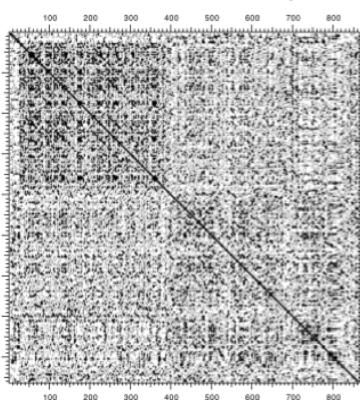
Only **diagonals** can be followed.

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

Uses for dot matrices

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

Repeats

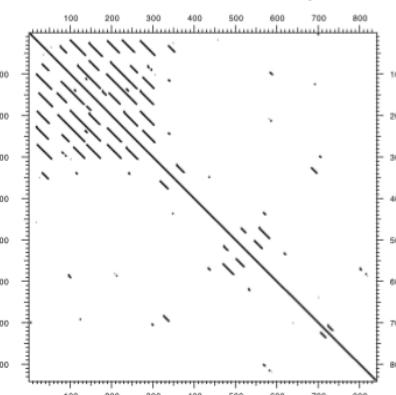


Human LDL receptor protein sequence (Genbank P01130)

$W = 1$
 $S = 1$

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

Repeats



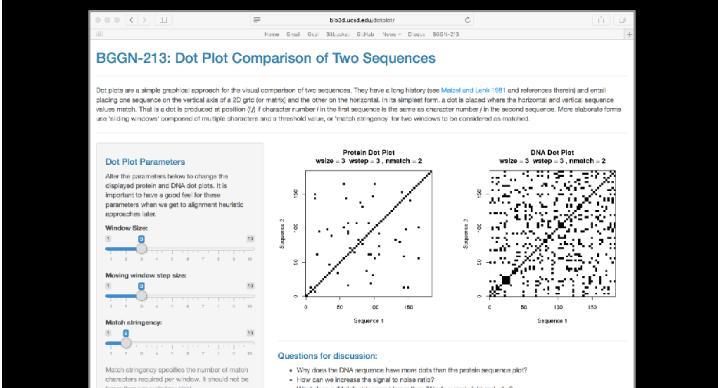
Human LDL receptor protein sequence (Genbank P01130)

$W = 23$
 $S = 7$

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

Your Turn!

Exploration of dot plot parameters (hands-on worksheet **Section 1**)
<http://bio3d.ucsd.edu/dotplot/> <https://bioboot.shinyapps.io/dotplot/>



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.

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	
-		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

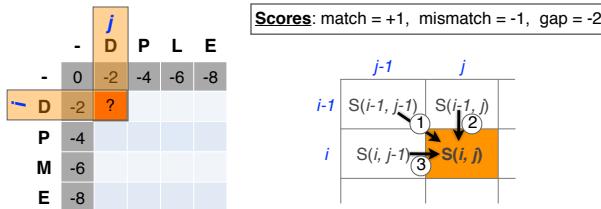
		Sequence 2				
		Scores: match = +1, mismatch = -1, gap = -2				
		D	P	L	E	
-		0	-2	-4	-6	-8
-		D	-2			
-		P	-4			
-		M	-6			
-		E	-8			

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

Seq1: DPME
Seq2: ----

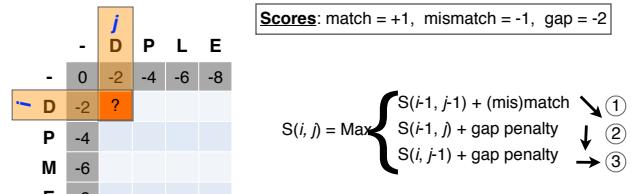
Scoring the alignment matrix

- 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



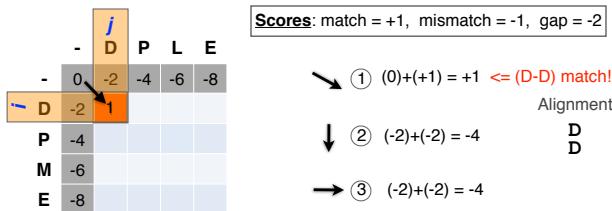
Scoring the alignment matrix

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



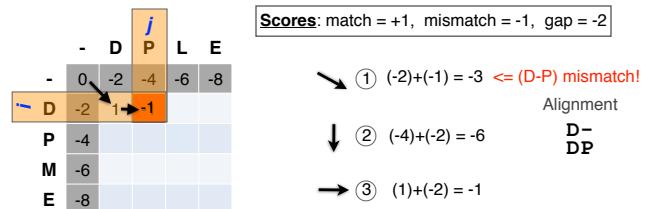
Scoring the alignment matrix

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



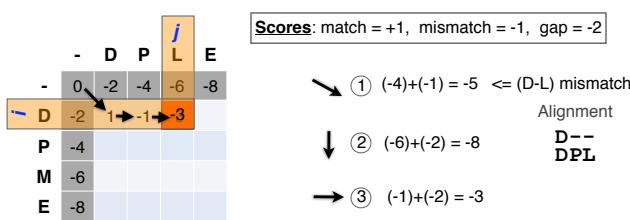
Scoring the alignment matrix

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



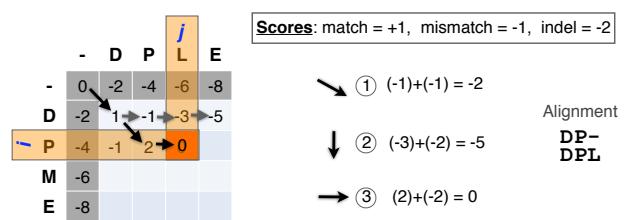
Scoring the alignment matrix

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



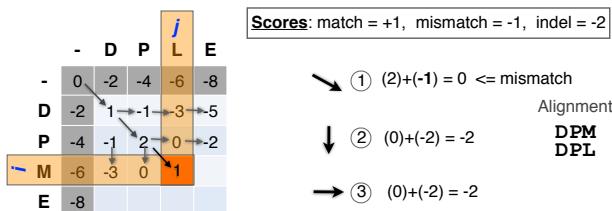
Scoring the alignment matrix

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



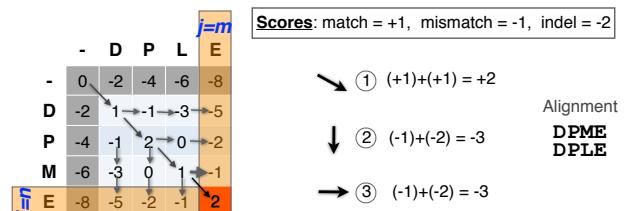
Scoring the alignment matrix

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



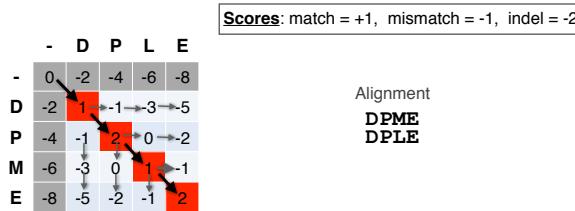
Scoring the alignment matrix

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



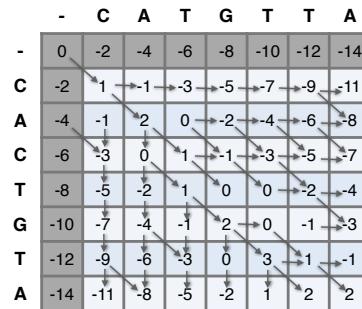
Scoring the alignment matrix

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



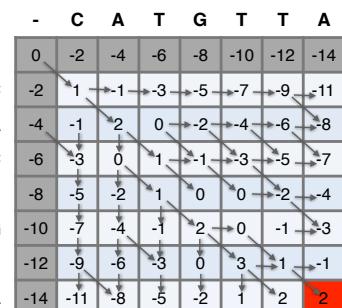
Questions:

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



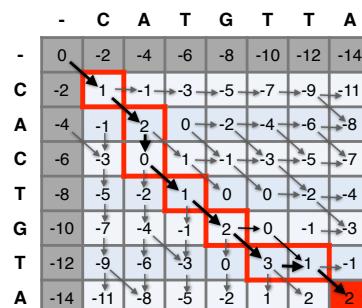
Questions:

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



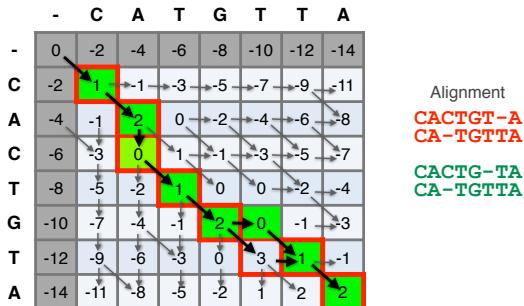
Questions:

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



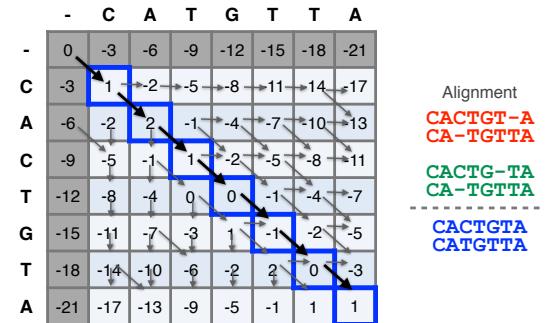
More than one alignment possible

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



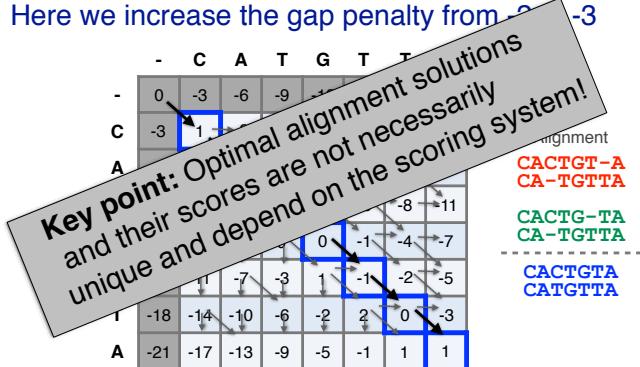
The alignment and score are dependent on the scoring system

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



The alignment and score are dependent on the scoring system

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



Your Turn!

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

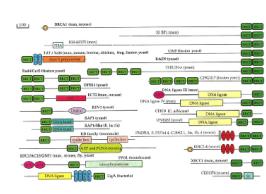
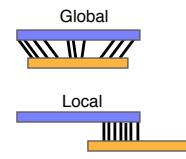
	A	G	T	T	C
A	0				
T					
T					
G					
C					

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

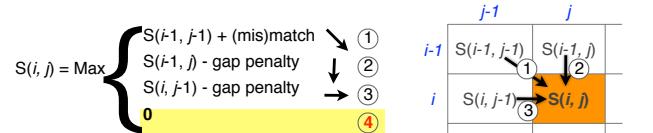
- Smith & Waterman proposed simply that a local alignment of two sequences allow arbitrary-length segments of each sequence to be aligned, with no penalty for the unaligned portions of the sequences.
- Otherwise, the score for a local alignment is calculated the same way as that for a global alignment

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

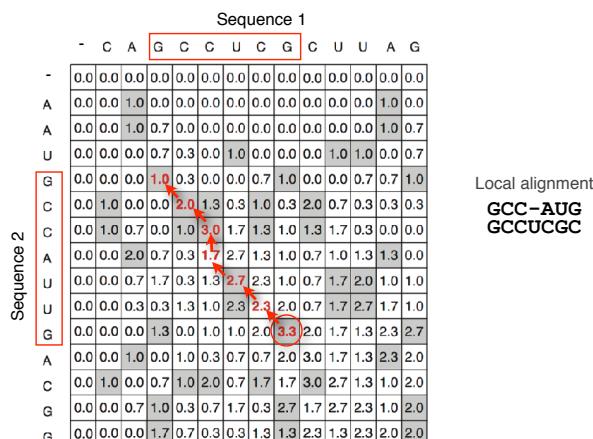
104

The Smith-Waterman algorithm

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



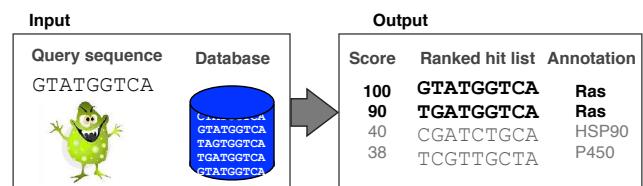
105



106

Local alignments can be used for database searching

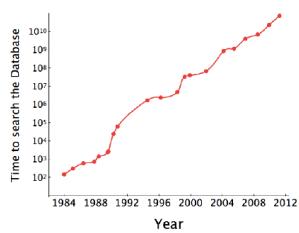
- Goal:** Given a query sequence (Q) and a sequence database (D), find a list of sequences from D that are most similar to Q
- Input:** Q, D and scoring scheme
- Output:** Ranked list of hits



107

The database search problem

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

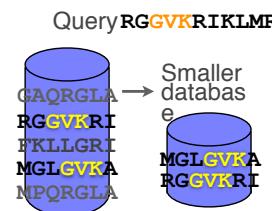


To reduce search time **heuristic algorithms**, such as BLAST, first remove database sequences without a strong local similarity to the query sequence in a quick initial scan.

108

The database search problem

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



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

109

ALIGNMENT FOUNDATIONS

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

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

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

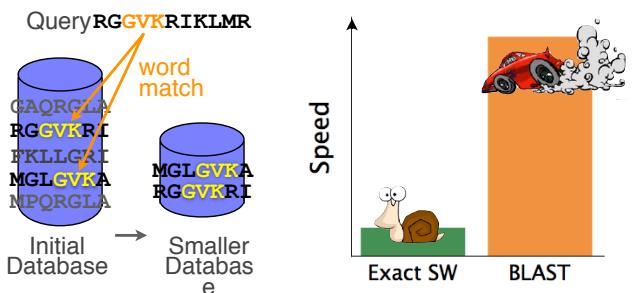
111

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

- BLAST (Basic Local Alignment Search Tool) is a simplified form of Smith-Waterman (SW). It is popular because it is **fast**.
– BLAST finds regions of similarity between two sequences
- BLAST does not examine every possible pair of sequences, but instead performs a search by scanning database sequences for likely matches before performing more rigorous alignments
- Altschul et al. (1990) “The central idea of the BLAST algorithm is to confine attention to sequence pairs that contain an initial **word pair match**”
- BLAST sacrifices some sensitivity in exchange for speed
- In contrast to SW, BLAST is not guaranteed to find optimal alignments

112

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



113

How BLAST works

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

RGGVKRI Query sequence
RGG
GGV
GVK
VKR
KRI

generate list of $w=3$ words for query

114

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

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

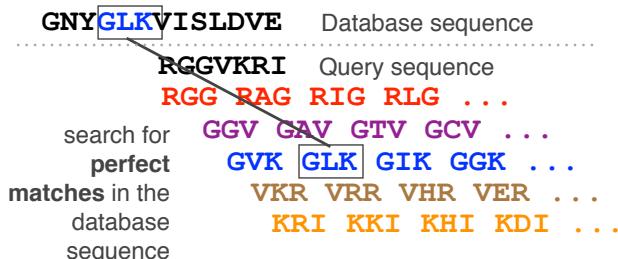
extend list of words similar to query

115

Blast

Blast

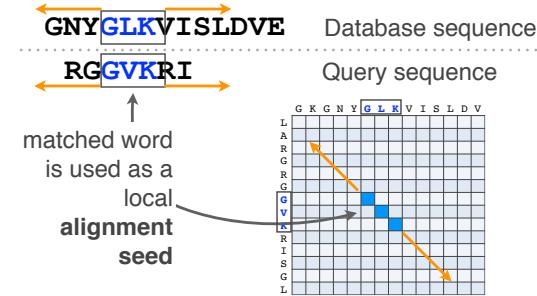
- Phase 3: a database is scanned to find sequence entries that match the compiled word list



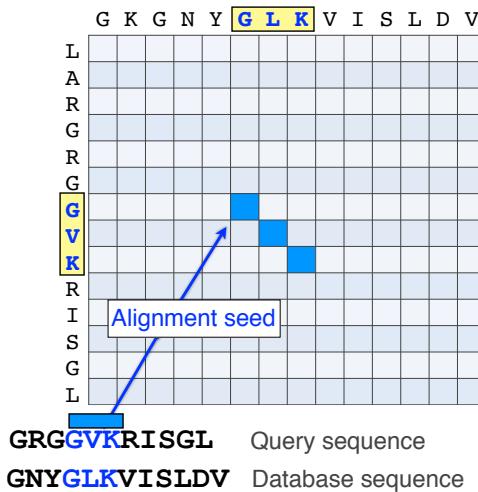
116

Blast

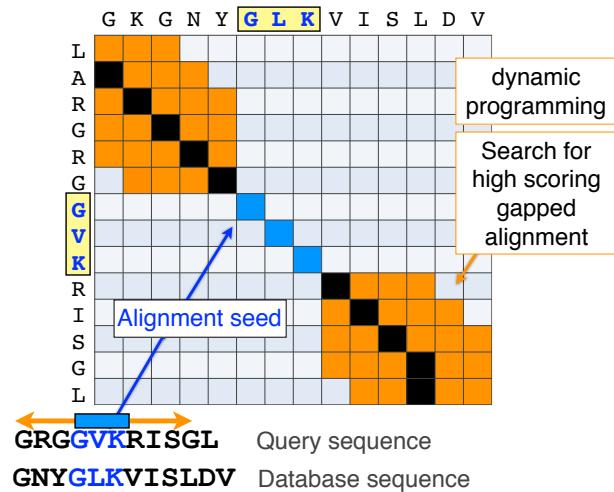
- Phase 4: the initial database hits are extended in both directions using dynamic programming



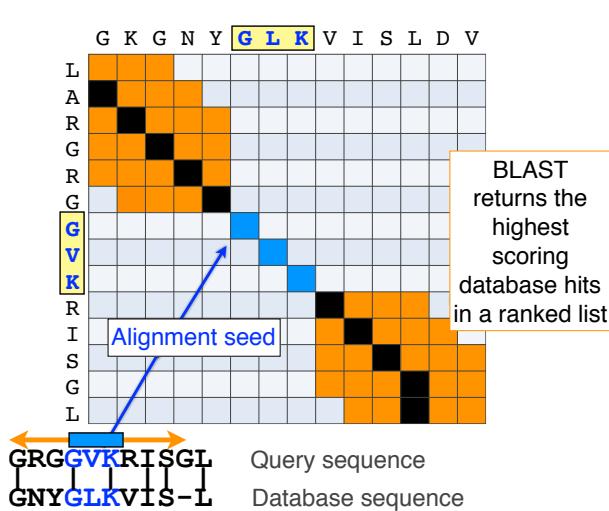
117



118



119



120

BLAST output

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

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

121

Statistical significance of results

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

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

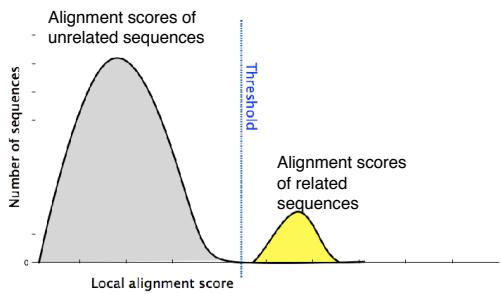
122

BLAST scores and E-values

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

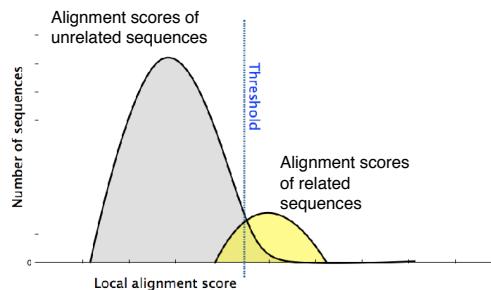
123

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



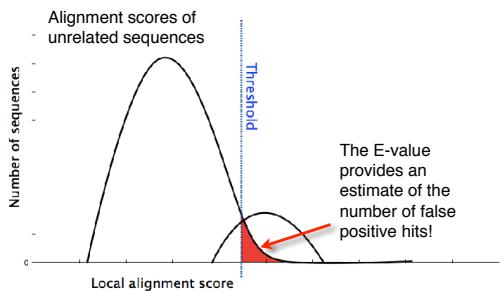
124

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



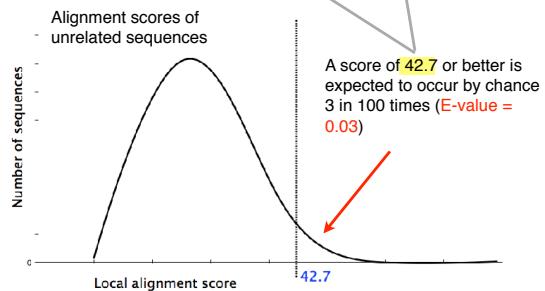
125

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



126

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



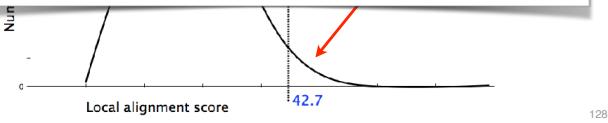
127

Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo sapiens]	677	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	676	100%	0	98%	AA20133.1

In general E values < 0.005 are usually significant.

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

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



Your Turn!

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

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

Practical database searching with BLAST

The screenshot shows the NCBI BLAST Home Page with the URL <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. The search term 'Arabidopsis thaliana' has been entered into the search field. Below the search bar, a list of results is displayed under the heading 'BLAST Assembled RefSeq Genomes'. The first result is for 'Arabidopsis thaliana' with a score of 1000. A link to 'View Aligning Multiple Pairs' is also present.

Score	Organism
1000	Arabidopsis thaliana

Below the main search area, there are sections for 'Basic BLAST' and 'Specialized BLAST'. The 'Basic BLAST' section includes links for 'nucleotide blast', 'protein blast', 'tblastx', 'tblastn', and 'tBlastx'. The 'Specialized BLAST' section includes links for 'nucleotide blast', 'protein blast', 'tblastx', 'tblastn', and 'tBlastx'.

Practical database searching with BLAST

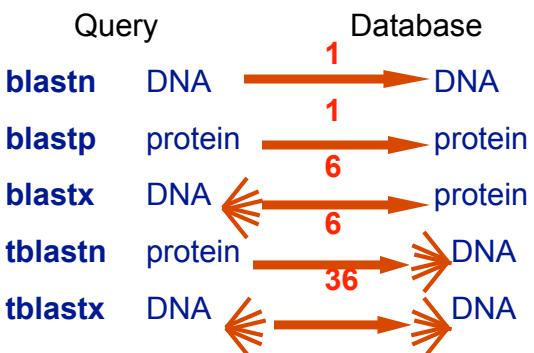
- There are four basic components to a traditional BLAST search
 - (1) Choose the sequence (query)
 - (2) Select the BLAST program
 - (3) Choose the database to search
 - (4) Choose optional parameters
 - Then click “BLAST”

Step 1: Choose your sequence

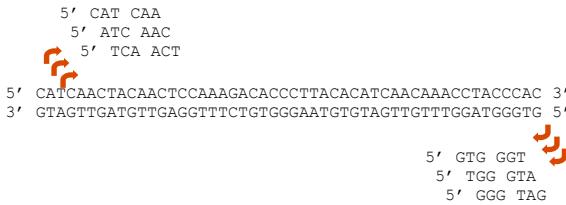
- Sequence can be input in FASTA format or as accession number

The screenshot shows the NCBI Protein search interface. The search term 'Protein' is entered in the search bar, and the results are for 'Translations of Life'. The main result is 'hemoglobin subunit beta [Homo sapiens]'. Below the title, the accession number NP_000509.1 is shown. The sequence is displayed with several regions highlighted by red circles: 'Display Settings' (FASTA), the protein name, the accession number, and two segments of the sequence starting with 'NP...' and 'AL...'. On the right side, there are links for 'Send to', 'Change region shown', 'Analyze this sequence', 'Run BLAST', 'Identify Conserved Domains', and 'Find in this Sequence'.

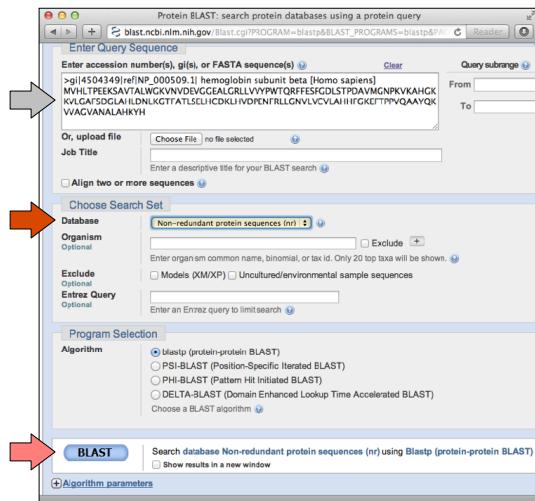
Step 2: Choose the BLAST program



DNA potentially encodes six proteins



134



135

Step 3: Choose the database

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

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

Microrna collection (miRNA)

Reference mRNA sequences (refseq_mrna)

NCBI Genomes (chromosome)

Non-human, non-mouse EST's (est_others)

Genomic survey sequences (gss)

High throughput genomic sequences (HTGS)

Patient sequences (pat)

Protein Data Bank (pdb)

Human ALU repeat elements (alu_repeats)

Sequence tagged sites (dbsts)

Whole-genome shotgun reads (wgs)

Environmental samples (env_nt)

nucleotide databases

Non-redundant protein sequences (nr)

Reference proteins (refseq_protein)

Swissprot protein sequences (swissprot)

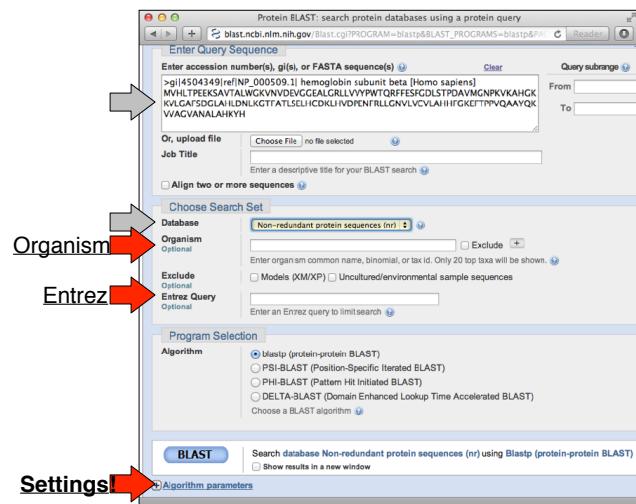
Potential protein sequences (pft)

Protein Data Bank proteins (pdb)

Environmental samples (env_nt)

protein databases

136



137

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

138

Step 4: Optional parameters

You can...

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

139

Results page

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

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI BLAST blast suite! Formating Results - FVGUTMR2013

Query ID gi|4504349|ref|NP_000509.1| hemoglobin
Description beta (Homo sapiens)
Database All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type amino acid
Query Length 147

Program BLAST 2.2.2+ > Citation

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

New DELTA-BLAST, a more sensitive protein-protein search

Graphic Summary

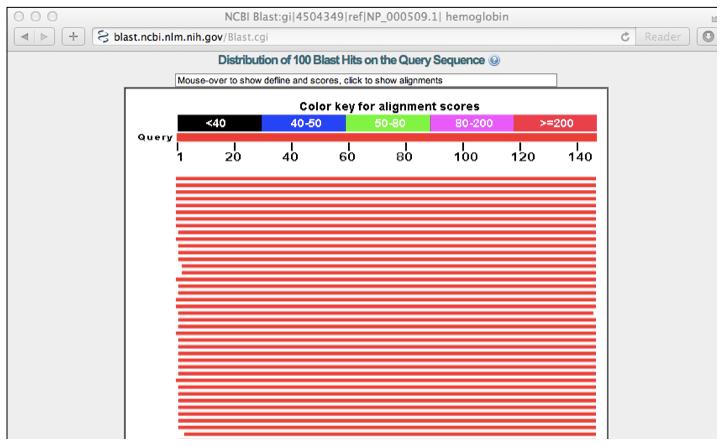
Show Conserved Domains

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

Query seq. hem-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 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=Hemoglobin subunit beta [Homo sapiens]	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 beta globin [Homo sapiens]	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 At The Beta Chain	298	298	99%	9e-102	100% 1COH_B
hemoglobin beta subunit variant [Homo sapiens] >gb AAA88054.1 beta-globin [Homo sapiens]	298	298	100%	1e-101	99% AAF0489.1
Chain B, Human Hemoglobin D Los Angeles: Crystal Structure >pdb 2YR8D 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 Spectroscopic	297	297	99%	3e-101	99% 1HDB_B

Further down the results page...

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

Download GenPept Graphics

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

Range 1 to 147 Seq. 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	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	60			
Subjct 1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	60			
Query 61	VKAAGCKVLGNSFGCLAHLDNLKGTFATSLHCLCDLJVDPENPFLRGVLVCLAHIFHG	120			
Subjct 61	VKAAGCKVLGNSFGCLAHLDNLKGTFATSLHCLCDLJVDPENPFLRGVLVCLAHIFHG	120			
Query 121	KETFPVQAAQKPVQAVGANALAKHYH	147			
Subjct 121	KETFPVQAAQKPVQAVGANALAKHYH	147			

Range 1 to 147 Seq. Graphics

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

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

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI BLAST blast suite! Formating Results - FVGUTMR2013

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

Change the result display back to traditional format

You can learn about the enhanced report Blast report description

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 top 20 taxon will be shown.
Enter organism name or id—completions will be suggested Exclude

Entrez query:

Expect Min: Expect Max: Percent Identity Min: Percent Identity Max:

Format for: PSI-BLAST with inclusion threshold:

gi|4504349|ref|NP_000509.1| hemoglobin

E.g. Query anchored alignments

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

Query	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	60
AAZ37051	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	60
NP_000509	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	60
P02024	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	60
AAN84548	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	60
AAZ39780	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	60
ACU56984	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	60
AAZ39782	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	60
2YR8D	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	60
IDXU_B	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IHD8_B	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IHD8_C	2	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IHD8_D	2	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IHD8_E	2	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IOP9_B	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IOP9_D	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IOP9_E	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IYB2_B	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IYB2_D	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IYB2_E	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IYB8_B	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IYB8_D	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IYB8_E	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IABY_B	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IABY_D	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
IABY_E	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59
ICMY_B	1	MVHLTPEEKSAVALGKVNVDVEVGGEALGRLLVVPPWQRFFESGDLSTPDAVGMNPX	59

... and alignments with dots for identities

The screenshot shows a BLAST search results page from NCBI. The query sequence is a short peptide: MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPK. The results list various proteins from different organisms, each with its ID, name, and a sequence alignment. The alignments use dots to represent identities and dashes for mismatches. The first few results are: P02024 (AAH84548), AAH31380 (AAH31380), AAC016984 (AAC016984), AAD19696 (AAD19696), and ICG1_B (ICG1_B). The alignments show significant homology between the query and the target sequences.

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?

147

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

148

How to handle too few results

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

149

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](#).

To Update!

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
Correct optimal score position in matrix used	C
Correct optimal score obtained for given scoring scheme	B
Traceback path(s) clearly highlighted	A
Correct alignment(s) yielding optimal score listed	A+