BIMM 143

Introduction to Bioinformatics
Lecture 2

Barry Grant

UC San Diego

http://thegrantlab.org/bimm143

Recap From Last Time:

- Bioinformatics is computer aided biology.
  - ‣ Deals with the collection, archiving, organization, and interpretation of a wide range of biological data.

- The NCBI and EBI are major online bioinformatics service providers.

- Introduced via **hands-on session** the BLAST, Entrez, GENE, OMIM, UniProt, Muscle and PDB bioinformatics tools and databases.

    - Muddy point assessment (see results)

- There are a large number of bioinformatics databases (see handout!).

- Also covered: Course structure; Supporting course website, Ethics code, and Introductions…

Today's Menu

Classifying Databases	Primary, secondary and composite Bioinformatics databases
Using Databases	**Vignette** demonstrating how major Bioinformatics databases intersect
Major Biomolecular Formats	How nucleotide and protein sequence and structure data are represented
Alignment Foundations	**Introducing the *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/

Hands on demo (or see following slides)

Example Vignette Questions:

- What chromosome location and what genes are in the vicinity of a given query gene? NCBI **GENE**

- What can you find out about molecular functions, biological processes, and prominent cellular locations? EBI **GO**

- What amino acid positions in the protein are responsible for ligand binding? EBI **UniProt**

- What variants of this gene are associated with gastric cancer and other human diseases? NCBI **OMIN**

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

- Are high resolution protein structures available to examine the details of these mutations? How might we explain their potential molecular effects? RCSB **PDB**

10

1 AND 2

**ras AND disease
(1185 results)**

1 OR 2
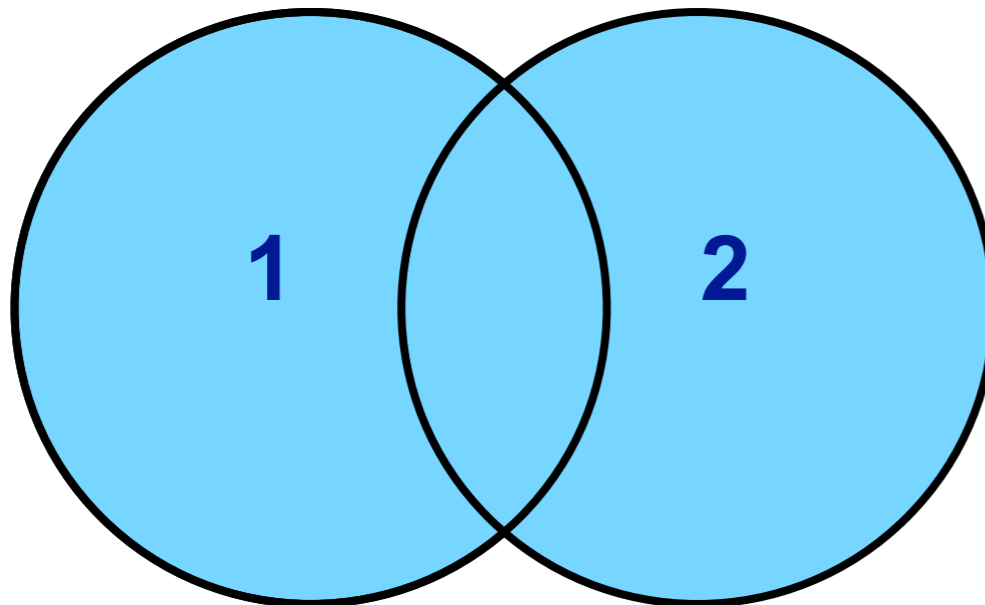
**ras OR disease
(134,872 results)**

1 NOT 2
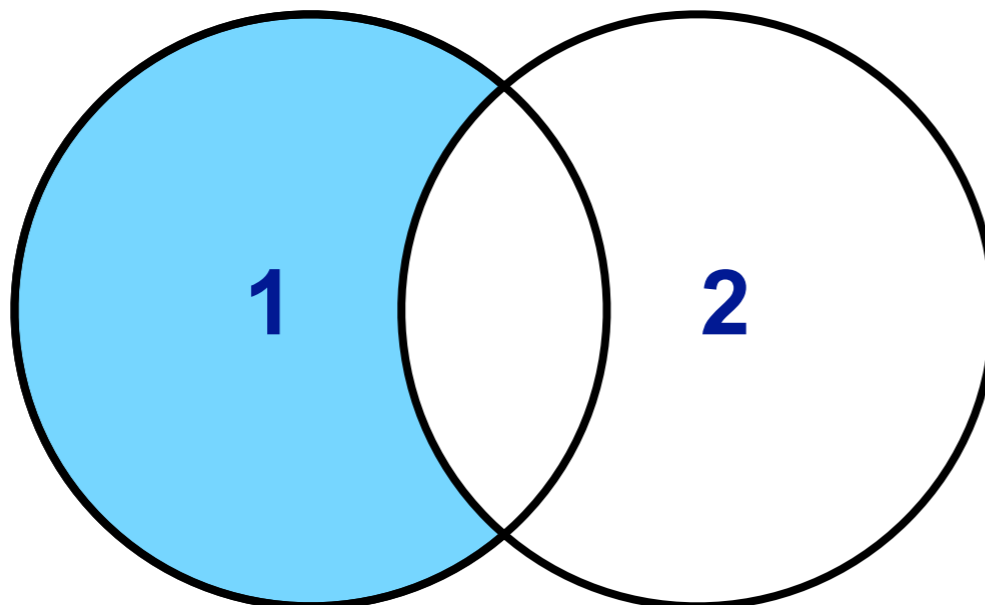
**ras NOT disease
(84,448 results)**

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

15

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

Gene Ontology   Provided by GOA

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

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

Process	Evidence Code	Pubs
Fc-epsilon receptor 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

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

UniProt will detail much more information for protein coding genes such as this one
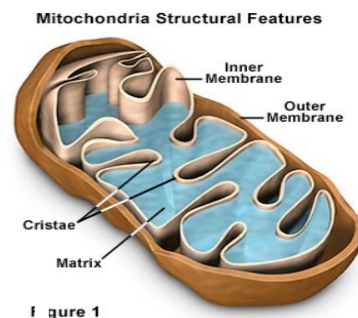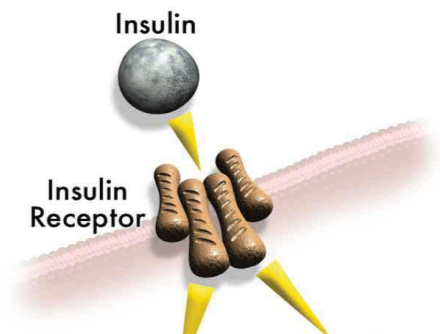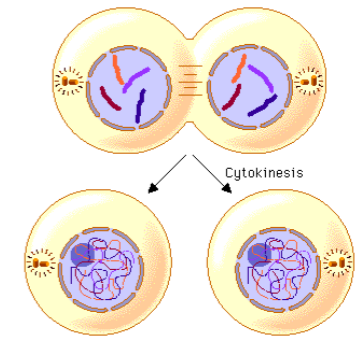
KRAS Kirsten rat sarcoma   ×

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

| genomic | X01669.1 | | CAA25828.1 |

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

Protein Accession	Links	
	GenPept Link	UniProtKB Link
P01116.1	GenPept	UniProtKB/Swiss-Prot:P01116

Scroll down to Very bottom for **UniProt** link

▼ **Additional links**   ⚠ ?

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

GETTING STARTED
NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials

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

POPULAR
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

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

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

UniProt will detail much more information for protein coding genes

UniProt will detail much more information for protein coding genes

KRAS - GTPase KRas prec ✕

www.uniprot.org/uniprot/P01116

UniProt

UniProtKB ▾   Advanced ▾ 🔍

BLAST   Align   Retrieve/ID Mapping   Help Contact

🛒 Basket ▾

P01116 - RASK_HUMAN

Protein	**GTPase KRas**
Gene	**KRAS**
Organism	*Homo sapiens (Human)*
Status	⭐ Reviewed - ⦿⦿⦿⦿⦿ - Experimental evidence at protein level[i]

View FASTA file format

Display   None   🔍 BLAST   ≡ Align   🗋 Format   🛒 Add to basket   ▶ Help video

☑ FUNCTION

☑ NAMES & TAXONOMY

☑ SUBCELL. LOCATION

☑ PATHOL./BIOTECH

☑ PTM / PROCESSING

☑ EXPRESSION

☑ INTERACTION

☑ STRUCTURE

☑ FAMILY & DOMAINS

☑ SEQUENCES (2)

☑ CROSS-REFERENCES

Function[i]

Ras proteins bind GDP/GTP and po
proliferation (PubMed:23698361, F

Enzyme regulation[i]

Alternates between an inactive for
nucleotide-exchange factor (GEF)
promotes exchange of bound GDP by GTP. 🏷 3 Publications ▾

```
>sp|P01116|RASK_HUMAN GTPase KRas OS=Homo sapiens GN=KRAS PE=1 SV=1
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDL
PSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYTLVREIRQYRLKKISKEEKTPGC
VKIKKCIIM
```

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

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

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

View PDB file format

RCSB PDB   Deposit ▾   Search ▾   Visualize ▾   Analyze ▾

RCSB PDB

An Information Portal to **133759** Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligand   Go

Advanced Search | Browse by Annotations

PDB-101   WORLDWIDE PDB PROTEIN DATA BANK   EMDataBank Unified Data Resource for 3DEM   NUCLEIC ACID DATABASE   Worldwide Protein Data Bank Foundation

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure... | ...ment

Literature

📄 Display Files ▾   ⊕ Download Files ▾

Contact Us

Biological Assembly 1

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

🔗 View in 3D: NGL or JSmol (in Browser)

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

wwPDB Validation   🔳 3D Report | Full Report

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?

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… ?

HHMI
janelia farm
research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam
keyword search Go

Family: *Kinesin* (PF00225)

126 architectures **4150** sequences **6** interactions **248** species **114** structures

- **Summary**
- **Domain organisation**
- **Clans**
- **Alignments**
- **HMM logo**
- **Trees**
- **Curation & models**
- **Species**
- **Interactions**
- **Structures**

Jump to... ⓘ

enter ID/acc   Go

Structures

For those sequences which have a structure in the Protein DataBank⧉, we use the mapping between UniProt⧉, PDB and Pfam coordinate systems from the 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
A8BKD1_GIALA	11 - 335	2vvg	A	11 - 335	Jmol AstexViewer SPICE⧉
			B	11 - 335	Jmol AstexViewer SPICE⧉
CENPE_HUMAN	12 - 329	1t5c	A	12 - 329	Jmol AstexViewer SPICE⧉
			B	12 - 329	Jmol AstexViewer SPICE⧉
KAR3_YEAST	392 - 723	1f9t	A	392 - 723	Jmol AstexViewer SPICE⧉
		1f9u	A	392 - 723	Jmol AstexViewer SPICE⧉
		1f9v	A	392 - 723	Jmol AstexViewer SPICE⧉
		1f9w	A	392 - 723	Jmol AstexViewer SPICE⧉
			B	392 - 723	Jmol AstexViewer SPICE⧉
		3kar	A	392 - 723	Jmol AstexViewer SPICE⧉
KI13B_HUMAN	11 - 352	3gbj	A	11 - 352	Jmol AstexViewer SPICE⧉
			B	11 - 352	Jmol AstexViewer SPICE⧉
			C	11 - 352	Jmol AstexViewer SPICE⧉
		1ii6	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⧉

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 **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**

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 programing
    - 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 programing
     - Global alignment
     - Local alignment
  ‣ BLAST heuristic approach

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

Seq1: C A T T C A C

Seq2: C T C G C A G C

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

```
Seq1: C A T T C A C
          |     | |
Seq2: C T C G C A G C
```

mismatch
match

Two types of character correspondence

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

```
Seq1: C A T - T C A - C
      |   |     | |   |
Seq2: C - T C G C A G C
```
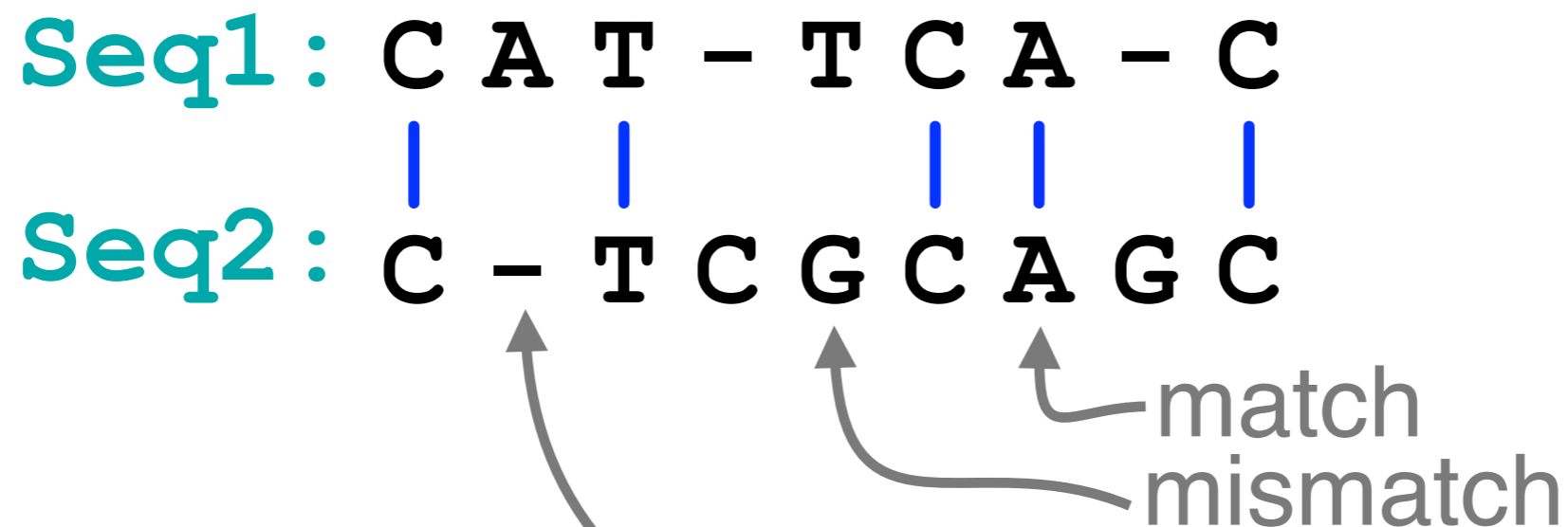
match
mismatch

Add gaps to increase number of matches

gaps

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

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

match
mismatch } mutation
insertion } indels
deletion

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~~s~~ construct
  such as a bact~~e~~

- Mappin~~g~~ to a known genome
  ~~look~~ing for differences from reference
  ~~~~s, indels (insertions or deletions)
  ~~~~ping 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 programing
     - 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

Time

Common Ancestor **CTCGTTA (C)**

Recent Species **(B)** **(A)**

CATGTTA **CACTGTA**

Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → CACGTTA

CTCGTTA

CACGTTA ← **Mutation**

CATGTTA   CACTGTA

Likely occurred prior to speciation

Mutations, deletions and insertions

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

– Mutations/Substitutions

– Deletions

– Insertions

CTCGTTA ⟶ C**A**CGTTA

CTCGTTA

C**A**CGTTA ⟵ **Mutation**

CACGTTA   CACGTTA   (**speciation**)

CATGTTA   CACTGTA

Mutations, deletions and insertions

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

- Mutations/Substitutions
- Deletions
- Insertions

CTCGTTA → C**A**CGTTA

CAC**G**TTA → CACTTA

CACTTA → CACT**G**TA

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)

CTCGTTA
(C)

(B)
CATGTTA

(A)
CACTGTA

Substitution   Indels

(A) CAC-TGTA
    || : | ||
(B) CATGT-TA

Match 5   Mismatch 1   Gap 2

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

🟢 4 (+3)
🟠 3 (+1)
⚪ 0 (-1) = 15

🟢 6 (+3)
🟠 0 (+1)
⚪ 2 (-1) = 16

🟢 5 (+3)
🟠 1 (+1)
⚪ 2 (-1) = 14

Optimal alignments

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

Optimal alignments

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

Optimal alignments

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

Optimal alignments

- Biologists often prefer parsimonious alignments, where the number of sequence changes is ~~~~~~~~~~~~

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

🟢 4 matches

🟠 3 mismatches

⚪

🟠 1 mismatches

⚪ 2 gaps

```
CACTGT-A       CAC-TGTA
| | | | | | |  | | : | | |
CA-TGTTA       CATGT-TA
```

CATGTTA

ALIGNMENT FOUNDATIONS

- **Why…**

  - Why compare biological sequences?

- **What…**

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

- **How…**

  ‣ Dot matrices
  ‣ Dynamic programing
      - Global alignment
      - Local alignment
  ‣ BLAST heuristic approach

ALIGNMENT FOUNDATIONS

- **Why…**
  - Why compare biological sequences?

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

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

Dot plots: simple graphical approach

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

Dot plots: simple graphical approach

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

Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence

Dot plots: simple graphical approach

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

Dot plots: 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

Filter

Window = 3
Stringency = 3

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

Filter

Window = 3
Stringency = 3

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

Filter

Window = 3
Stringency = 3

Dot plots: window size and match stringency

Solution: use a <u>window</u> and a <u>threshold</u>
– 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

Filter

Window = 3
Stringency = 3

Dot plots: window size and match stringency

Solution: use a <u>window</u> and a <u>threshold</u>
– 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

Filter

Window = 3
Stringency = 3

Dot plots: window size and match stringency

Solution: use a <u>window</u> and a <u>threshold</u>
– 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

Filter

Window = 3
Stringency = 3

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

Filter

Window = 3
Stringency = 3

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

Filter

Window = 3
Stringency = 3

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

Filter

Window = 3
Stringency = 3

Dot plots: window size and match stringency

Solution: use a <u>window</u> and a <u>threshold</u>
– 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

Filter

Window = 3
Stringency = 3

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

Filter

Window = 3
Stringency = 2

Dot plots: window size and match stringency

Solution: use a <u>window</u> and a <u>threshold</u>

– 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

Filter

Window = 3
Stringency = 2

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 place. 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).

indels

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/

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

ALIGNMENT FOUNDATIONS

- **Why…**

  - Why compare biological sequences?

- **What…**

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

- **How…**

  ‣ Dot matrices

  ‣ Dynamic programing
      - 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

(1) →

	D	P	L	E
D				
P				
M				
E				

(2) →

	D	P	L	E
D	6	-1	-4	2
P	-1	7	-3	-1
M	-3	-2	2	-2
E	-2	-1	-3	5

(3) →

	D	P	L	E
D	6	-1	-4	2
P	-1	7	-3	-1
M	-3	-2	2	-2
E	-2	-1	-3	5

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 <u>diagonal paths</u> & indels with <u>horizontal or vertical path</u> 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

j   Sequence 2

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

		-	D	P	L	E
i	-	0	-2	-4	-6	-8
	D	-2				
	P	-4				
	M	-6				
	E	-8				

Sequence 1

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

j   Sequence 2

	-	D	P	L	E
-	0	-2	-4	-6	-8
D	-2				
P	-4				
M	-6				
E	-8				

Sequence 1

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

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

```
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

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

	-	D	P	L	E
-	0	-2	-4	-6	-8
D	-2	?			
P	-4				
M	-6				
E	-8				

j

	j-1	j
i-1	S(i-1, j-1)	S(i-1, j)
i	S(i, j-1)	S(i, j)

1 2 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 of the three directions gives the highest score?
  - keep track of this score and direction

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

	-	D	P	L	E
-	0	-2	-4	-6	-8
D	-2	?			
P	-4				
M	-6				
E	-8				

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

Scoring the alignment matrix

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

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

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

① (0)+(+1) = +1 <= (D-D) match!

Alignment

② (-2)+(-2) = -4

D
D

③ (-2)+(-2) = -4

Scoring the alignment matrix

- At each step, the score in the current cell is determine 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)

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

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

j

① (-2)+(-1) = -3 <= (D-P) mismatch!

② (-4)+(-2) = -6

③ (1)+(-2) = -1

Alignment

D−
DP

Scoring the alignment matrix

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

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

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

j

① (-4)+(-1) = -5 <= (D-L) mismatch

Alignment

② (-6)+(-2) = -8

```
D--
DPL
```

③ (-1)+(-2) = -3

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.

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

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

① (-1)+(-1) = -2

② (-3)+(-2) = -5

③ (2)+(-2) = 0

Alignment

```
DP-
DPL
```

Scoring the alignment matrix

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

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

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

j

① (2)+(-1) = 0 <= mismatch

② (0)+(-2) = -2

③ (0)+(-2) = -2

Alignment

```
DPM
DPL
```

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)

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

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

① (+1)+(+1) = +2

② (-1)+(-2) = -3

③ (-1)+(-2) = -3

Alignment

DPME
DPLE

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

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

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

Alignment

```
DPME
DPLE
```

Questions:

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

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

Questions:

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

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

Questions:

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

More than one alignment possible

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

Alignment

CACTGT-A
CA-TGTTA

CACTG-TA
CA-TGTTA

The alignment and score are dependent on the scoring system

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

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

Alignment

CACTGT-A
CA-TGTTA

CACTG-TA
CA-TGTTA

CACTGTA
CATGTTA

The alignment and score are dependent on the scoring system

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

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

Alignment

CACTGT–A
CA–TGTTA

CACTG–TA
CA–TGTTA

CACTGTA
CATGTTA

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

NW DYNAMIC PROGRAMMING

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

		A	G	T	T	C
	0					
A						
T						
T						
G						
C						

ALIGNMENT FOUNDATIONS

- **Why…**
  - Why compare biological sequences?

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

- **How…**
  - ‣ Dot matrices
  - ‣ Dynamic programing
    - 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

Global

Local

Local alignment: Definition

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

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

The Smith-Waterman algorithm

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

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

	j-1	j
i-1	S(i-1, j-1)	S(i-1, j)
i	S(i, j-1)	S(i, j)

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

Input

Query sequence

GTATGGTCA

Database

CTATGGTCA
GTATGGTCA
TAGTGGTCA
TGATGGTCA
GTATGGTCA

Output

Score	Ranked hit list	Annotation
100	**GTATGGTCA**	Ras
90	**TGATGGTCA**	Ras
40	CGATCTGCA	HSP90
38	TCGTTGCTA	P450

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 x 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 x n (m is length of query, n is length of database), too slow for large databases!

Query **RGGVKRIKLMR**

Smaller database

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.

GAQRGLA
RGGVKRI
FKLLGRI
MGLGVKA
MPQRGLA

MGLGVKA
RGGVKRI

ALIGNMENT FOUNDATIONS

- **Why…**
  - Why compare biological sequences?

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

- **How…**
  - ‣ Dot matrices
  - ‣ Dynamic programing
    - - 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 To... simplified form of Smith-Waterman... that is popular because it is f... ...ble
  - BLAST finds regions ...gical sequences
  - BLAST ...the search by scanning ...ly matches before performing ...ents
  - ...me sensitivity in exchange for speed
  - ...ntrast to SW, BLAST is not guaranteed to find optimal alignments

"The central idea of the BLAST algorithm is to confine attention to sequence pairs that contain an initial word pair match"

Altschul et al. (1990)

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

Query **RGGVKRIKLMR**

word match

GAQRGLA
RGGVKRI
FKLLGRI
MGLGVKA
MPQRGLA

Initial Database → Smaller Database

MGLGVKA
RGGVKRI

Speed

Exact SW   BLAST

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
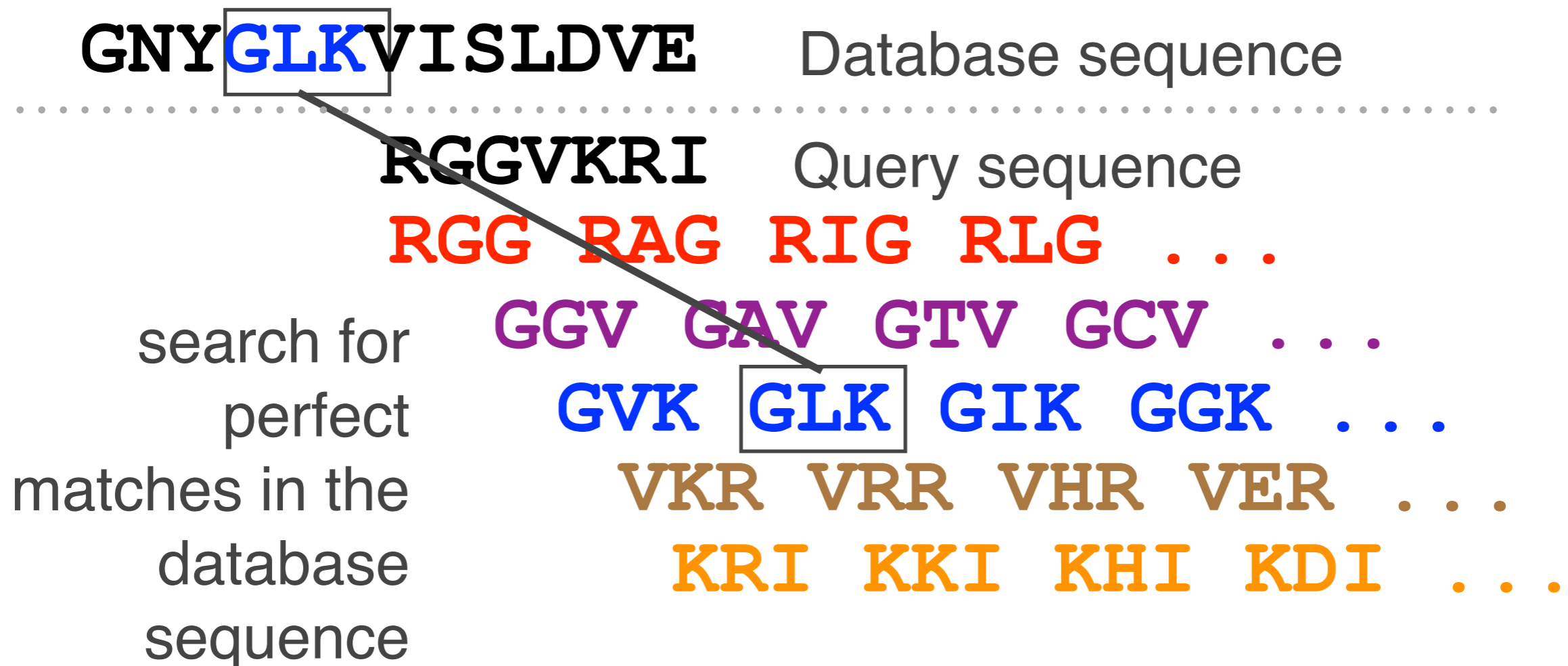
RGG RAG RIG RLG ...

GGV GAV GTV GCV ...

extend list of words similar to query

GVK GAK GIK GGK ...

VKR VRR VHR VER ...

KRI KKI KHI KDI ...

Blast

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

GNY⟨GLK⟩VISLDVE Database sequence

RGGVKRI Query sequence

RGG RAG RIG RLG ...

GGV GAV GTV GCV ...

search for
perfect
matches in the
database
sequence

GVK ⟨GLK⟩ GIK GGK ...

VKR VRR VHR VER ...

KRI KKI KHI KDI ...

Blast

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

GNYGLKVISLDVE   Database sequence

RGGVKRI   Query sequence

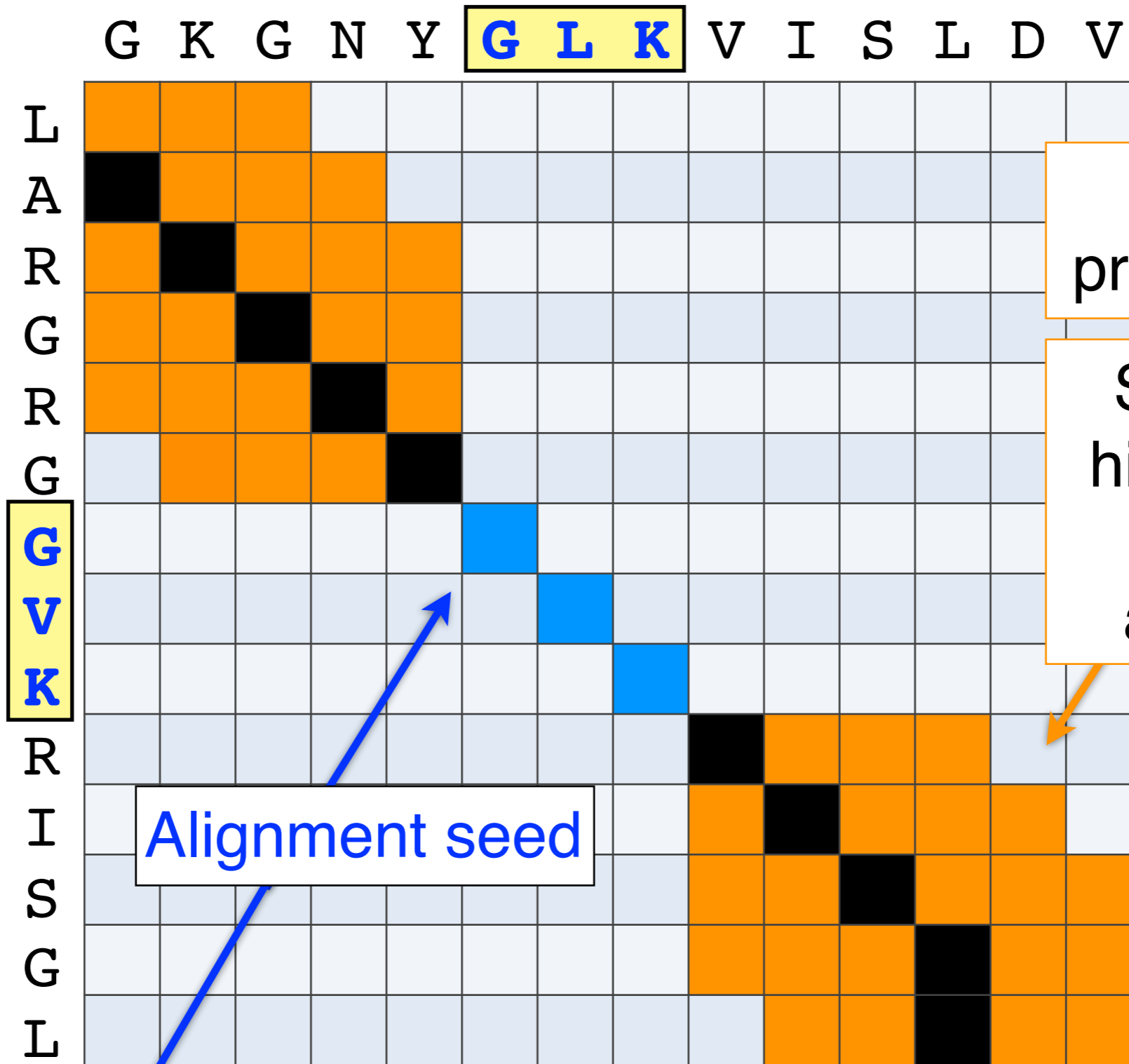matched word is used as a local alignment seed

Alignment seed

GRG**GVK**RISGL   Query sequence

GNY**GLK**VISLDV   Database sequence

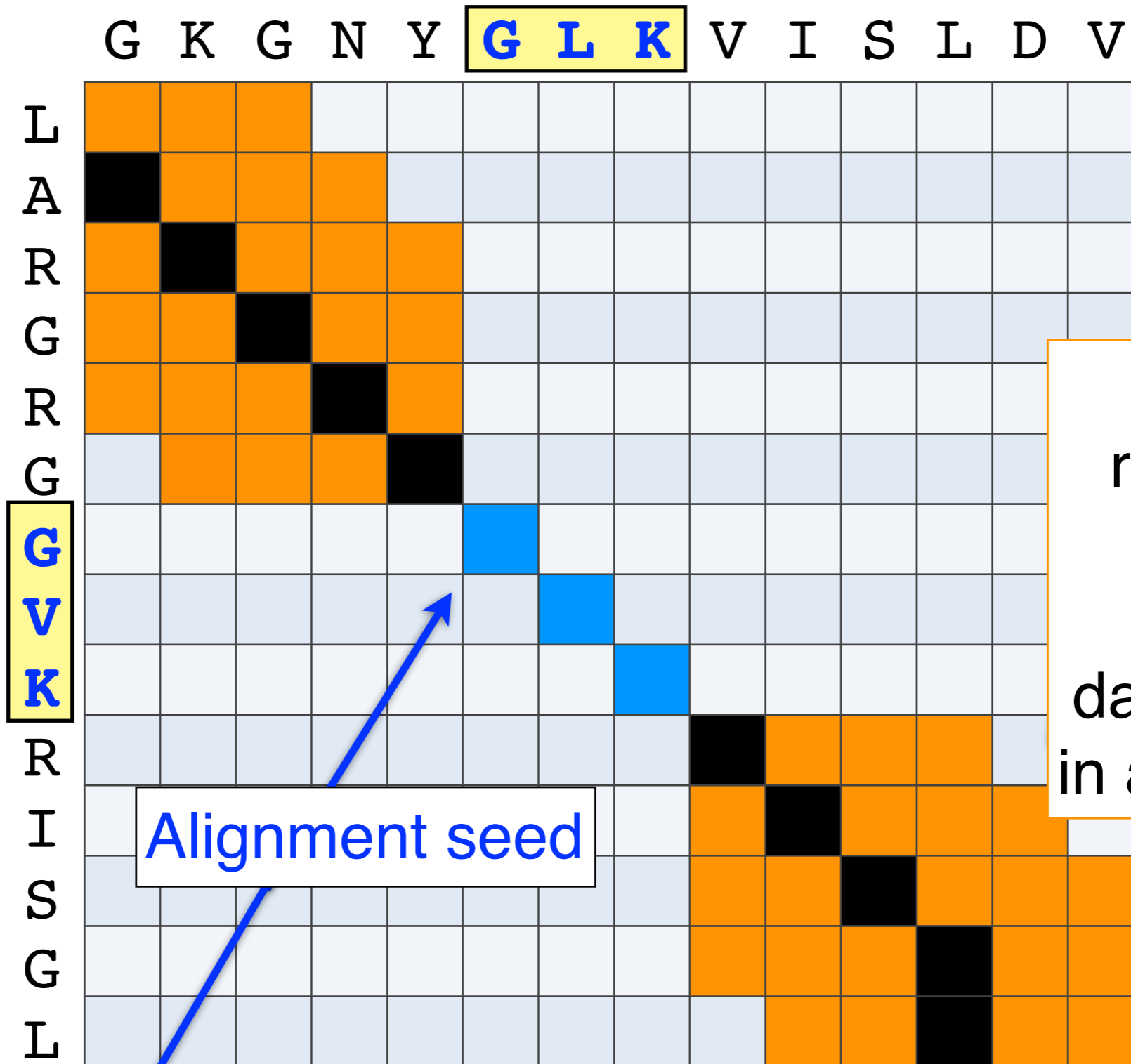dynamic programming

Search for high scoring gapped alignment

Alignment seed

GRG**GVK**RISGL — Query sequence

GNY**GLK**VISLDV — Database sequence

BLAST returns the highest scoring database hits in a ranked list

Alignment seed

GRG**GVK**RISGL   Query sequence

GNY**GLK**VIS-L   Database sequence

BLAST output

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

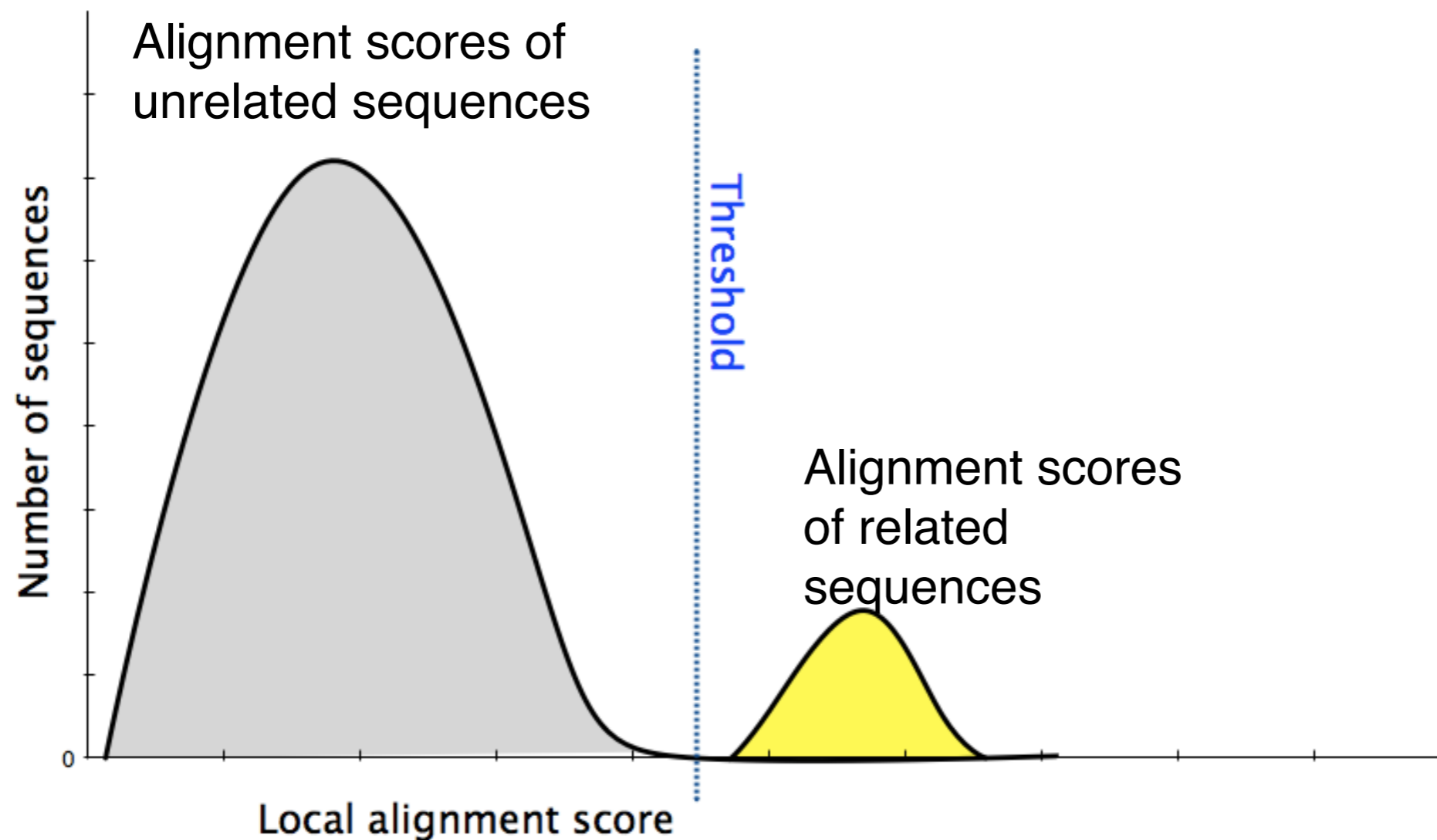| Description	Max score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo sapiens]	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	100%	0	98%	AAA20133.1
Kinesin-14 heavy chain [Danio rerio]	595	88%	0	78%	XP_00320703
hypothetical protein EGK_18589	48.2	40%	0.03	32%	ELK35081.1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

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)

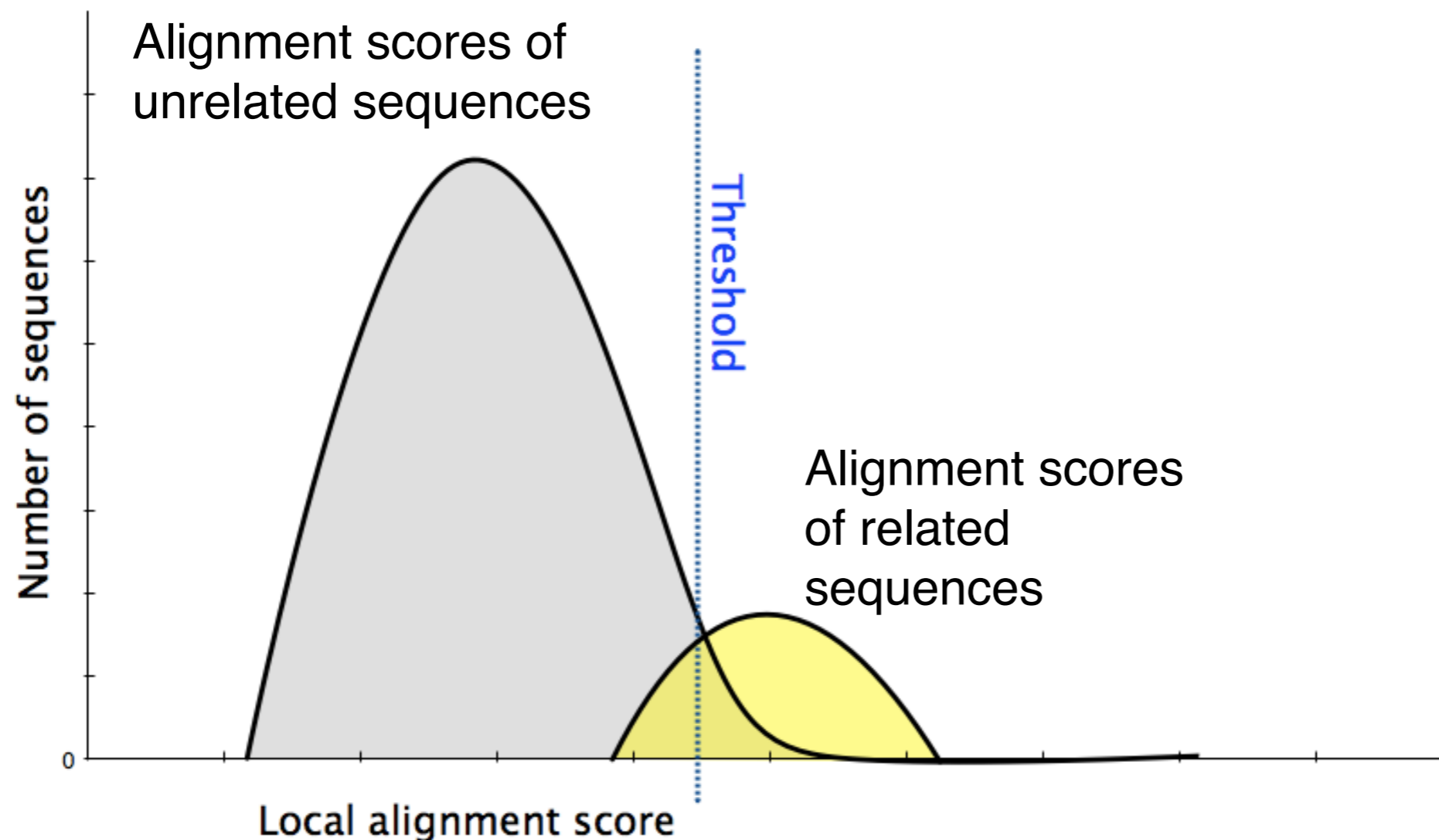| Description	Max score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo sapiens]	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	100%	0	98%	AAA20133.1
Kinesin-14 heavy chain [Danio rerio]	595	88%	0	78%	XP_00320703
hypothetical protein EGK_18589	48.2	40%	0.03	32%	ELK35081.1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

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
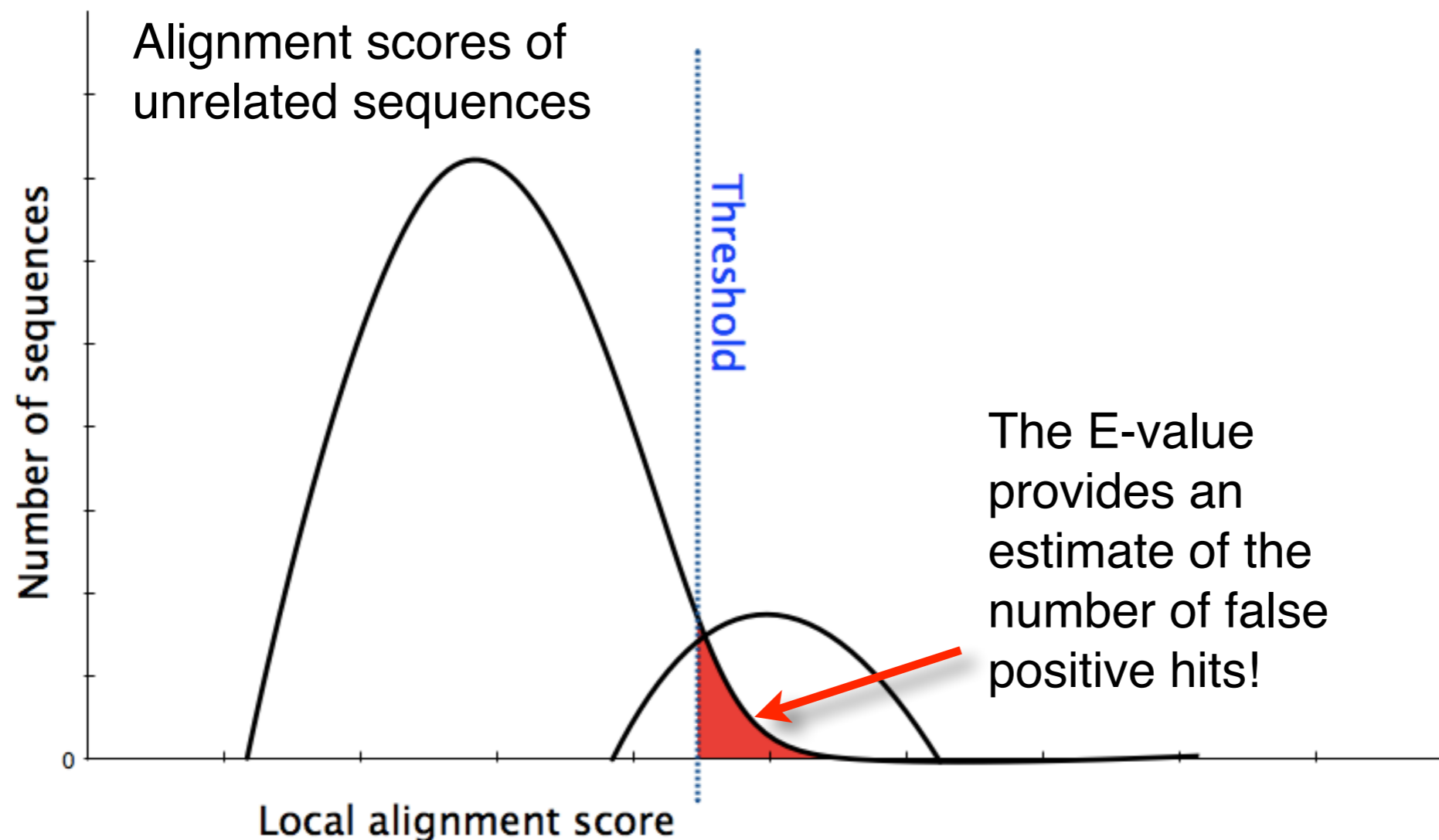 - This is equivalent to selecting alignments with score above a certain score threshold

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

Alignment scores of unrelated sequences

Threshold

Alignment scores of related sequences

Number of sequences

Local alignment score

- 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

Alignment scores of unrelated sequences

Threshold

Alignment scores of related sequences

Number of sequences

0

Local alignment score

- 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

Alignment scores of
unrelated sequences

Threshold

Number of sequences

The E-value
provides an
estimate of the
number of false
positive hits!

0

Local alignment score

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

Alignment scores of unrelated sequences

A score of 42.7 or better is expected to occur by chance 3 in 100 times (E-value = 0.03)

Number of sequences

0

42.7

Local alignment score

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

In general *E* values < 0.005 are usually significant.

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

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

Num

0

Local alignment score

42.7

Your Turn!

Hands-on worksheet **Sections 6**

‣ Please do answer the last lab review question (**Q19**).

‣ We encourage discussion and exploration!

Summary

- 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.

- Reviewed global and local alignment approaches and their major application areas.

- Heuristic approaches are often necessary for large database searches and many genomic applications.

FOR NEXT CLASS…

Check out the online:

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

☑ **Homework**: (1) **Quiz**, (2) **Alignment Exercise**.

Homework Grading

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

(Homework 2) Assessment Criteria	Points	
Setup labeled alignment matrix	1	
Include initial column and row for GAPs	1	
All alignment matrix elements scored (i.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	D
Correct optimal score position in matrix used	1	C
Correct optimal score obtained for given scoring scheme	1	B
Traceback path(s) clearly highlighted	1	A
Correct *alignment(s)* yielding optimal score listed	1	A+

REFERENCE SLIDES...

Additional reference slides for the motivated student

Practical database searching with BLAST

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

Step 1: Choose your sequence

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

Step 2: Choose the BLAST program

Query   Database

blastn   DNA   **1** →   DNA

blastp   protein   **1** →   protein

blastx   DNA   **6** →   protein

tblastn   protein   **6** →   DNA

tblastx   DNA   **36** →   DNA

DNA potentially encodes six proteins

```
5' CAT CAA
  5' ATC AAC
    5' TCA ACT

5' CATCAACTACAACTCCAAAGACACCCTTACACATCAACAAACCTACCCAC 3'
3' GTAGTTGATGTTGAGGTTTCTGTGGGAATGTGTAGTTGTTTGGATGGGTG 5'

                                   5' GTG GGT
                                     5' TGG GTA
                                       5' GGG TAG
```

Step 3: Choose the database

nr = non-redundant (most general database)

dbest = database of expressed sequence tags

dbsts = database of sequence tag sites

gss = genomic survey sequences

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/nt)
Reference mRNA sequences (refseq_rna)
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_nt)

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

nucleotide databases

149

Step 4a: Select optional search parameters

Expect

Word size

Scoring matrix

Step 4: Optional parameters

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

Results page

Further down the results page...

Further down the results page...

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

blast.ncbi.nlm.nih.gov/Blast.cgi
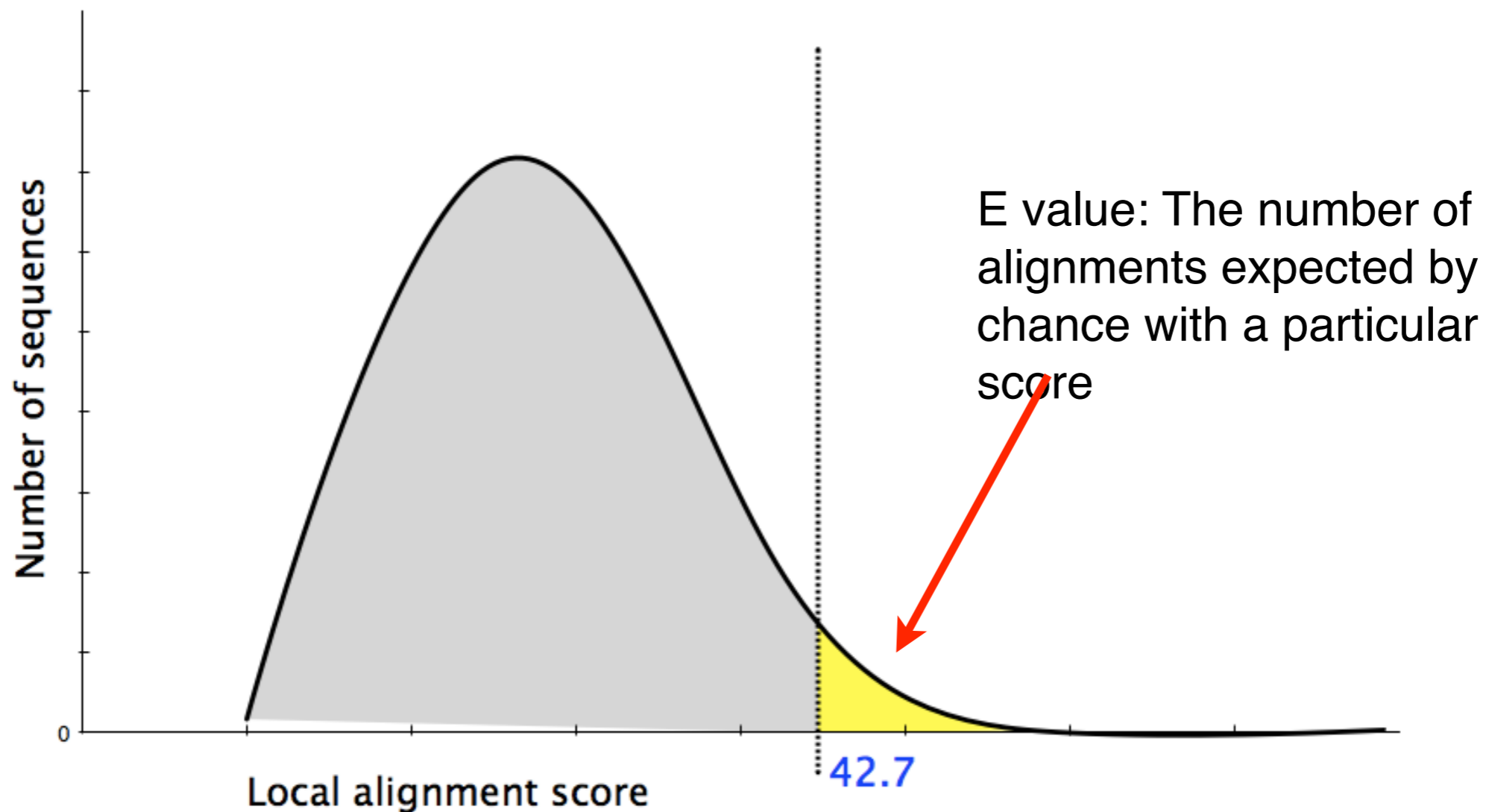
Sequences producing significant alignments:

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=Hen	300	300	100%	4e-102	99%	P02024.2
☐ beta globin chain variant [Homo sapiens]	299	299	100%	5e-102	99%	AAN84548.1
☐ beta globin [Homo sapiens] >gb|AAZ39781.1| beta globin [Homo sapiens] >gb|AAZ3978;	299	299	100%	5e-102	99%	AAZ39780.1
☐ beta-globin [Homo sapiens]	299	299	100%	5e-102	99%	ACU56984.1
☐ hemoglobin beta chain [Homo sapiens]	299	299	100%	6e-102	99%	AAD19696.1
☐ Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound At	298	298	99%	9e-102	100%	1COH_B
☐ hemoglobin beta subunit variant [Homo sapiens] >gb|AAA88054.1| beta-globin [Homo sa	298	298	100%	1e-101	99%	AAF00489.1
☐ Chain B, Human Hemoglobin D Los Angeles: Crystal Structure >pdb|2YRS|D Chain D, H	298	298	99%	2e-101	99%	2YRS_B
☐ Chain B, High-Resolution X-Ray Study Of Deoxy Recombinant Human Hemoglobins Syn	297	297	99%	3e-101	99%	1DXU_B
☐ Chain B, Analysis Of The Crystal Structure, Molecular Modeling And Infrared Spectroscop	297	297	99%	3e-101	99%	1HDB_B

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

E value: The number of alignments expected by chance with a particular score

42.7

E values in BLAST

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

Further down the results page...

Different output formats are available

E.g. Query anchored alignments

... and alignments with dots for identities

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

```
Query        1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK   60
AAX37051     1    ............................................................   60
AAX29557     1    ............................................................   60
NP_000509    1    ............................................................   60
P02024       1    ............................................................   60
AAN84548     1    ............................................................   60
AAZ39780     1    .......K....................................................   60
ACU56984     1    ...............................................K............   60
AAD19696     1    .........................................L..................   60
1COH_B       1    .............................................................   59
AAF00489     1    ............................................................   60
2YRS_B       1     ............................................................   59
1DXU_B       1    M............................................................   59
1HDB_B       1    .............................................................   59
1DXV_B       2    .............................................................   59
3KMF_C       2    .............................................................   59
AAL68978     1    ............................................................   60
1NQP_B       1    ...............................K.............................   59
1K1K_B       1    ......K......................................................   59
AAN11320     1    .......V....................................................   60
XP_002822173 1    ............................................................   60
1Y85_B       1     ............................................................   59
1YE0_B       1    M........................................A...................   59
1Q1Q_B       1    M............................................................   59
CAA23759     1    .......V.......X............................................   60
1YE2_B       1    M....................................F.......................   59
1Y5F_B       1    M............................................................   59
1A00_B       1    M..............................................Y............   59
```

Common problems

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

How to handle too many results

- Focus on the question you are trying to answer

  – select "refseq" database to eliminate redundant matches from "nr"

  – Limit hits by organism

  – Use just a portion of the query sequence, when appropriate

  – Adjust the expect value; lowering E will reduce the number of matches returned

How to handle too few results

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