Introduction to Linux and BLAST

Keith Jolley

Department of Biology

UNIVERSITY OF OXFORD
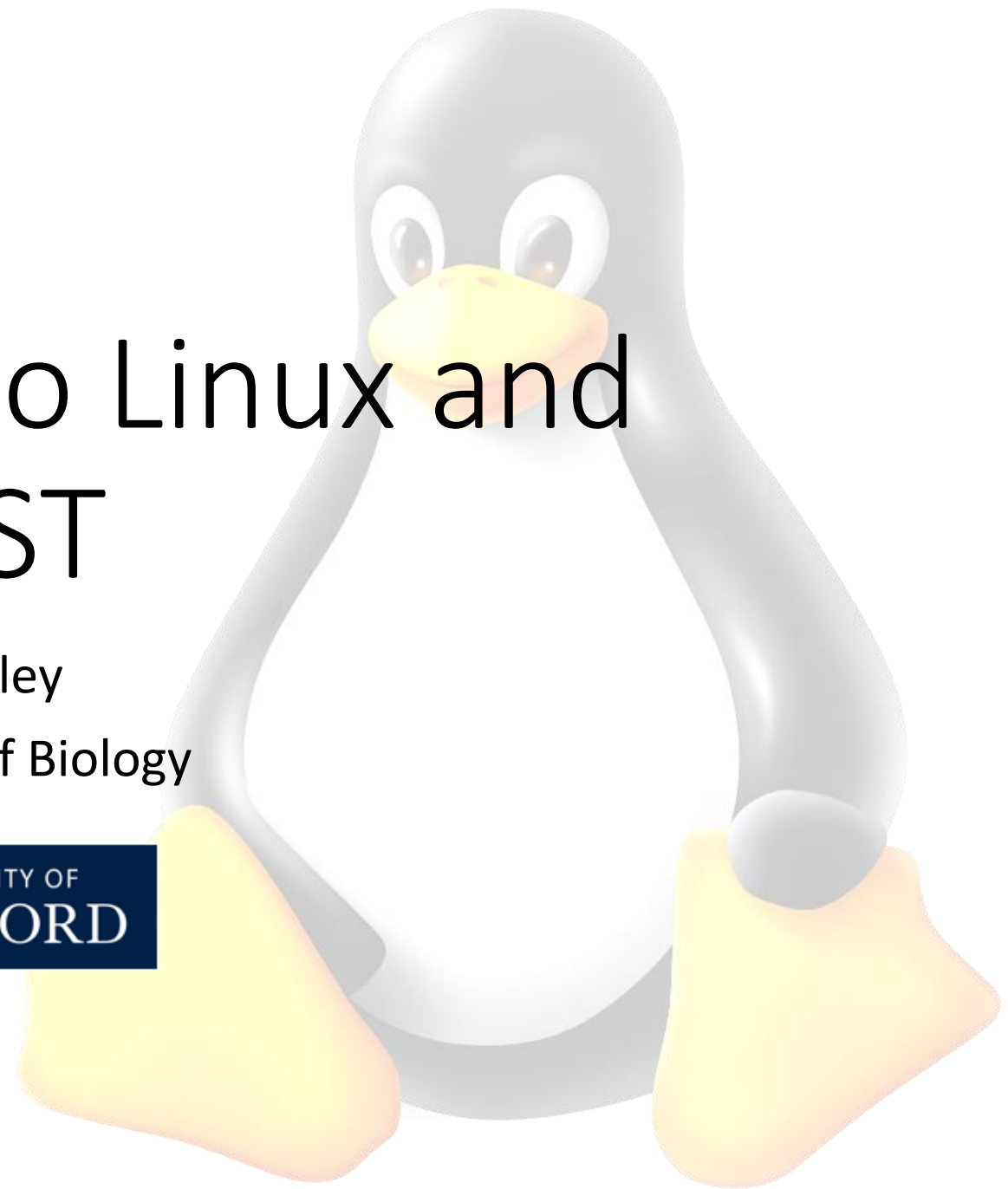
UNIX

- Development dates back to 1960s.

- Philosophy is to use a large range of simple, dependable tools that each do one simple task.

- Combining tools facilitates complex analysis.

- Popular for high-performance computing.

Ken Thompson (sitting) and Dennis Ritchie at PDP-11 (photo: Peter Hamer) circa 1970

UNIX/Linux

- UNIX is the operating system of choice for engineering and scientific computing.

- Linux is a free Unix clone.

- Linux runs on most hardware.
  - Embedded systems
  - Mobile phone
  - Desktop computers
  - Super-computing clusters

The command line

keith@zoo-jura[~]

File Edit View Search Terminal Help

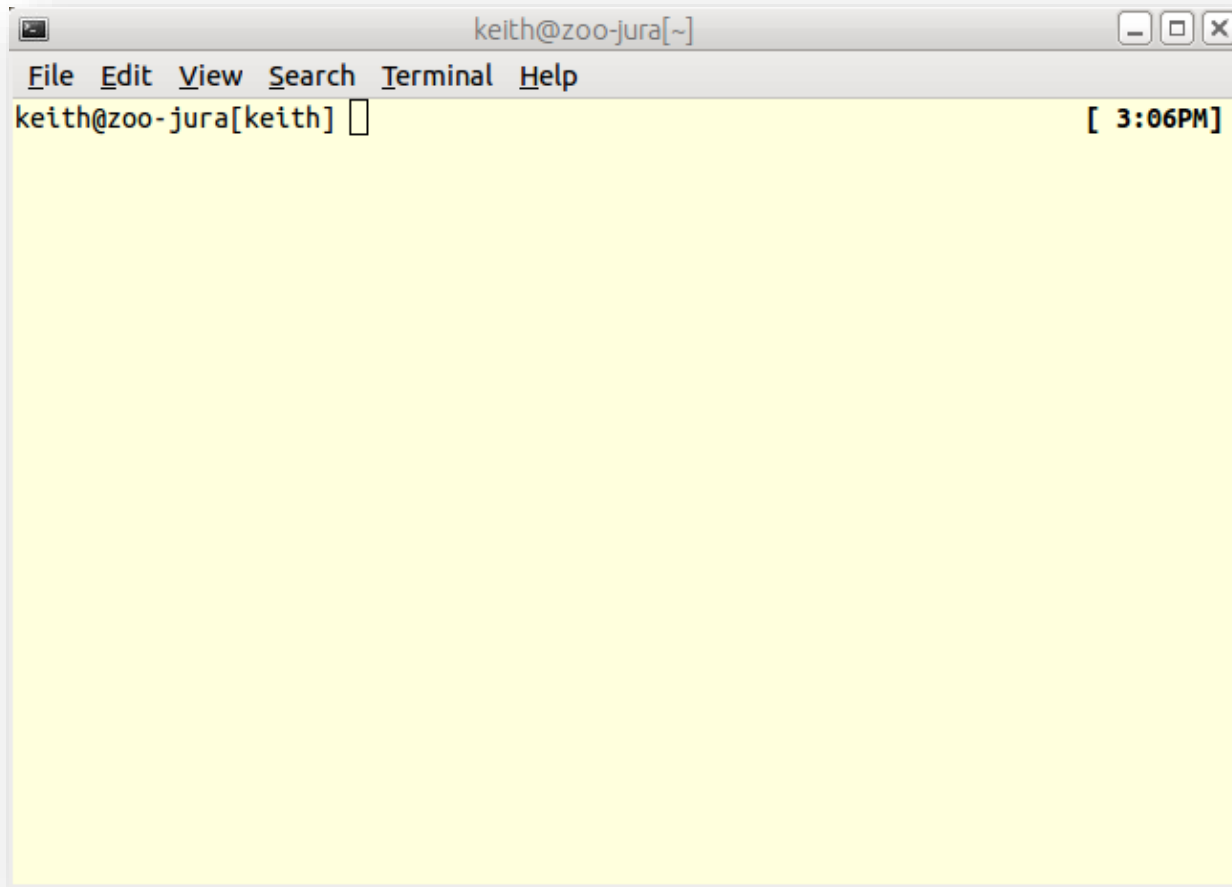keith@zoo-jura[keith] ▯   [3:06PM]
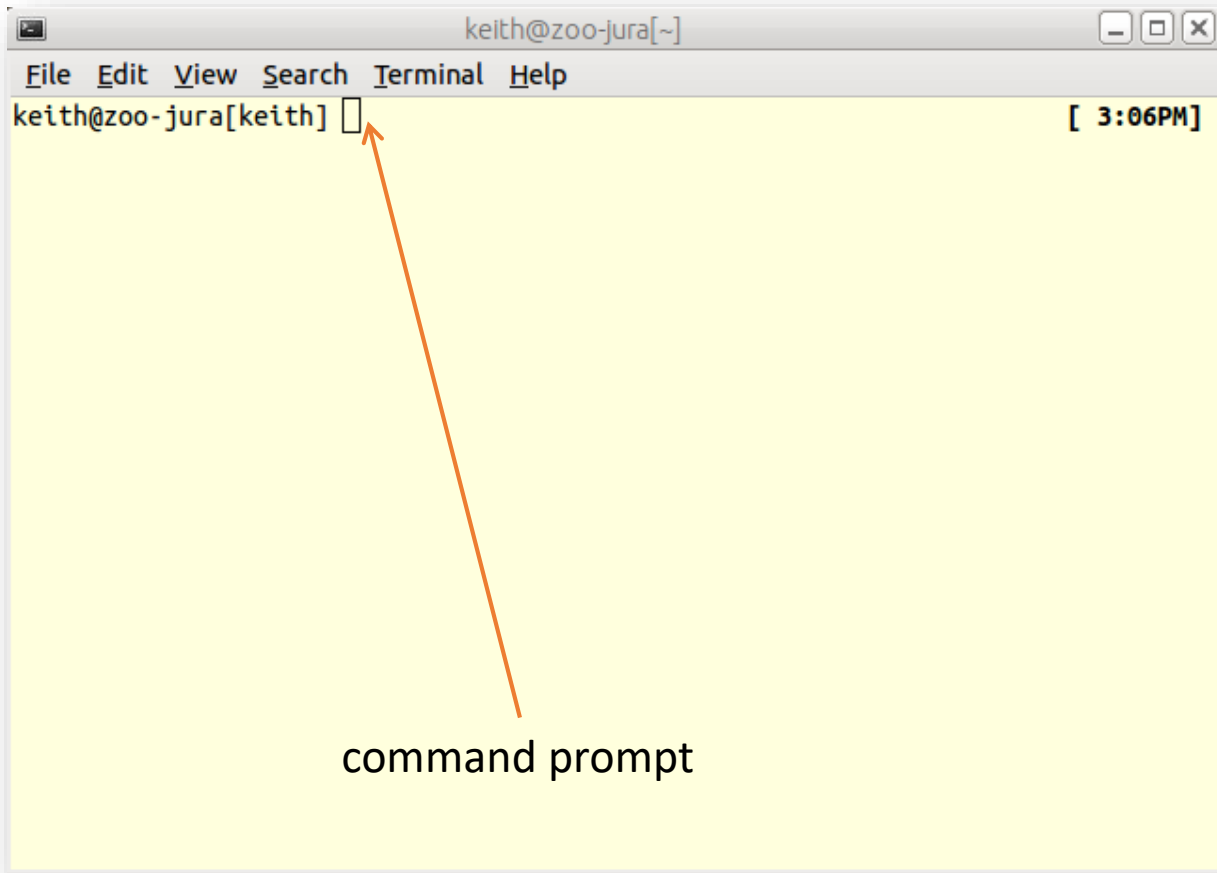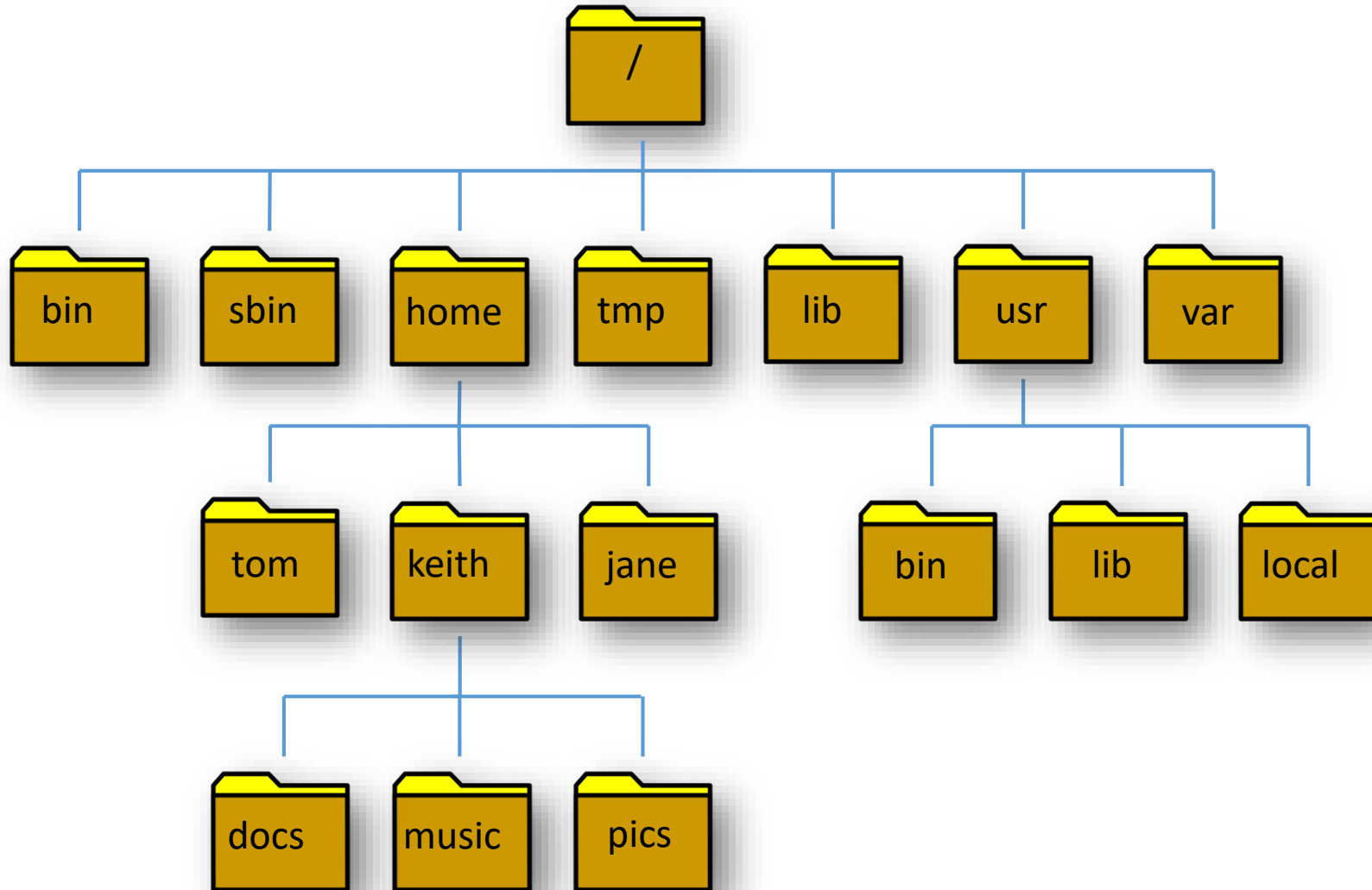
command prompt

- A command line, or terminal, is a text-based interface to the system.

- You enter commands by typing them on the keyboard and feedback is given to you similarly as text.

- The command line usually presents you with a prompt.

The command line

- Command syntax must be **exact!**
- Commands are **case-sensitive**.

The Linux file system is a hierarchy

Commands to navigate Linux filesystem

- ls (list)
  - Lists contents of current directory
- cd (change directory)
  - Change current focus to a new directory
  - Examples:
    - **cd documents** (enters directory called 'documents' in current directory)
    - **cd ..** (moves up a directory in the filesystem hierarchy)
    - **cd** (takes you to your home directory)
- pwd (print working directory)
  - Shows full path to current directory
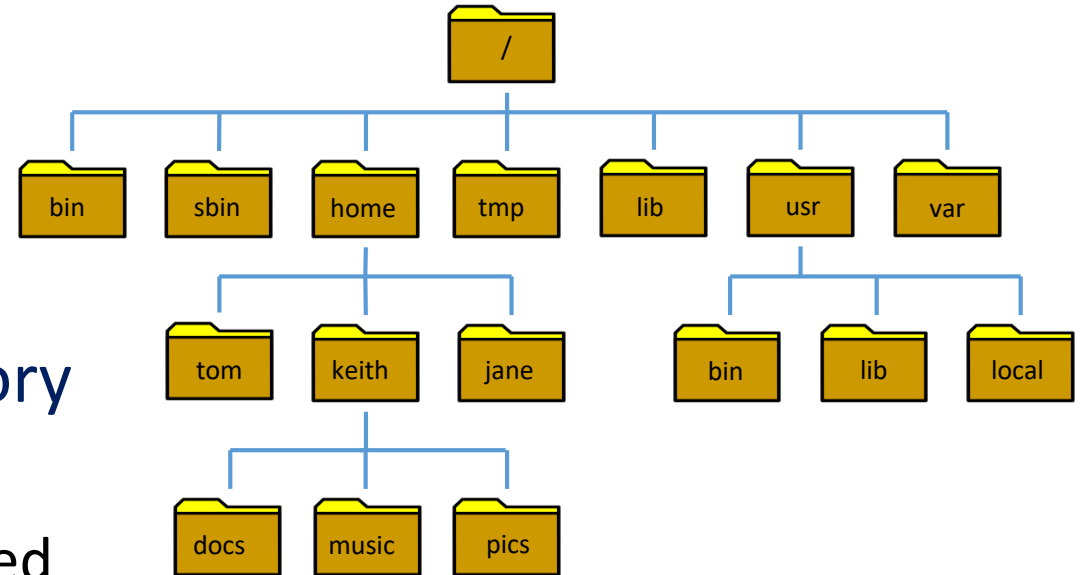
Commands that operate of files

- mv (move file or directory)
  - Examples
    - `mv file1 dir1` (move 'file1' in to existing sub-directory 'dir1')
  - If destination does not exist, then file is renamed
    - `mv file1 file2` (renames 'file1' to 'file2')
- rm (remove file)
  - Examples
    - `rm file1` (remove 'file1')
    - `rm -fr dir1` (remove dir1 and all its contents)

BLAST

Basic Local Alignment Search Tool

What is BLAST?

- Basic BLAST search
  - What is BLAST?
  - Different BLAST programs
  - BLAST databases you can search
  - Where can I run BLAST on the web?
- The BLAST algorithm
  - Stand-alone program is used behind-the-scenes in most sequence comparison applications

What is BLAST?

- BLAST stands for
  Basic Local Alignment Search Tool

- Why BLAST is popular?
  - Good balance of sensitivity and speed
  - Reliable
  - Flexible

- Local alignments; search query divided into short sequences ('words') and exact matches identified. Once found, these matches are extended

BLAST extends local alignments

DNA 1 on horizontal axis = 244 bases
DNA 2 on vertical axis = 248 bases

Click on plot to get positional data
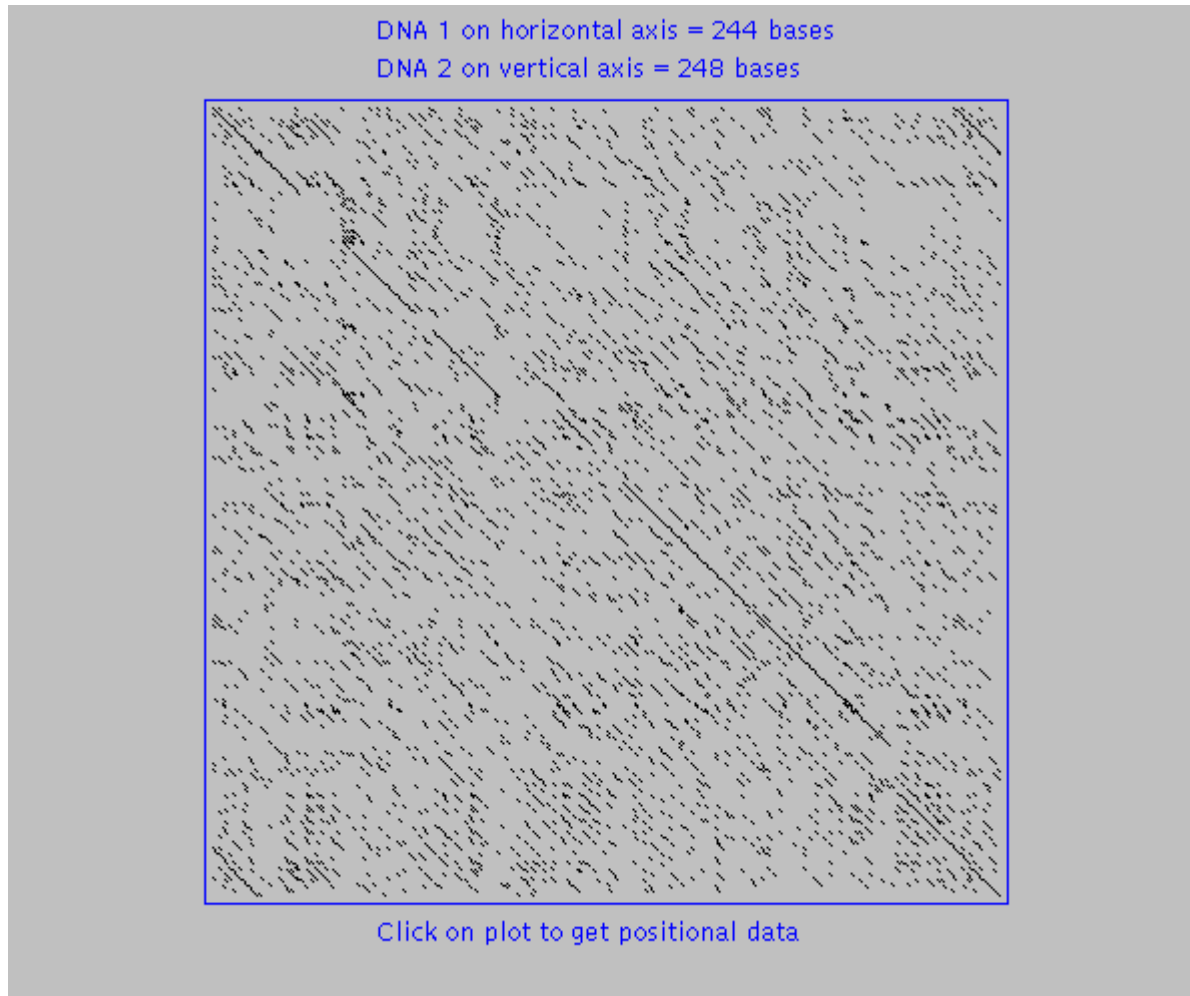
- Many nucleotide matches purely by chance (1:4)
- BLAST only considers significant 'seed' matches (default word size: 11)
- Scoring system takes account of matches, mismatches, gap formation and gap extension

BLAST output

1. List of sequences with scores
   - Raw score
     - Higher is better
     - Depends on aligned length
   - Expect Value (E-value)
     - Smaller is better
     - Independent of length and database size
2. List of alignments

```
Score =  248 bits (129), Expect = 1e-63
Identities = 213/263 (80%), Gaps = 34/263 (12%)
Strand = Plus / Plus


Query: 161  atatcaccacgtcaaaggtgactccaactcca---ccactccattttgttcagataatgc 217
            ||||||||||||||||||||||||||||||||    |     | |   || |||||||||||||
Sbjct: 481  atatcaccacgtcaaaggtgactccaact-tattgatagtgttttatgttcagataatgc 539


Query: 218  ccgatgatcatgtcatgcagctccaccgattgtgagaacgacagcgacttccgtcccagc 277
            |||||||    |||||||||||||||||||||| || |              |||||||||||
Sbjct: 540  ccgatgactttgtcatgcagctccaccgattttg-g------------ttccgtcccagc 586


Query: 278  c-gtgcc--aggtgctgcctcagattcaggttatgccgctcaattcgctgcgtatatcgc 334
            |    || |   |  |||||||||||||||||||||||||||||||||||||| |||||||||
Sbjct: 587  caatgacgta-gtgctgcctcagattcaggttatgccgctcaattcgctgggtatatcgc 645


Query: 335  ttgctgattacgtgcagctttcccttcaggcggga-------------ccagccatccgtc 382
            ||||||||||||||||||||||||||||||||||||          |||||||||||||||
Sbjct: 646  ttgctgattacgtgcagctttcccttcaggcgggattcatacagcggccagccatccgtc 705


Query: 383  ctccatatc-accacgtcaaagg 404
            ||||||||| |||||||||||||
Sbjct: 706  atccatatcaaccacgtcaaagg 728
```

BLAST Programs

Program	Database (Subject)	Query
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Protein	Nt. ➤ Protein
TBLASTN	Nt. ➤ Protein	Protein
TBLASTX	Nt. ➤ Protein	Nt. ➤ Protein

The BLOSUM62 matrix is used by BLAST to align amino acids and determine if two non-identical amino acids are 'positives'

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =   -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

BLAST ®

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.   Learn more
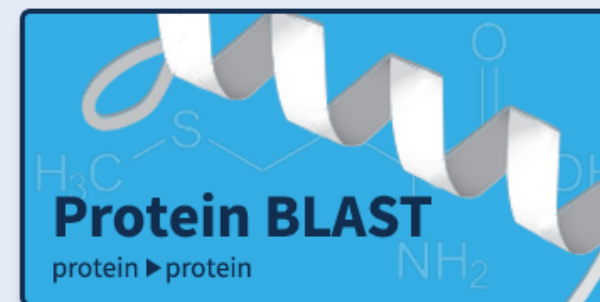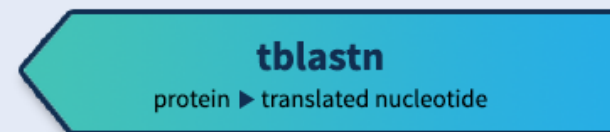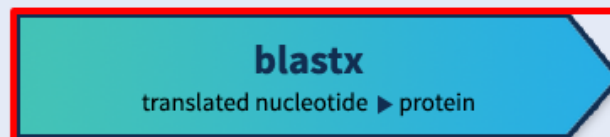
Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

BLAST Genomes

| Enter organism common name, scientific name, or tax id | Search |

Human   Mouse   Rat   Microbes

Home | Recent Results | Saved Strategies | Help

My NCBI
[Sign In] [Register]

NCBI/ BLAST/ blastx

Translated BLAST: blastx

blastn | blastp | **blastx** | tblastn | tblastx

BLASTX search protein databases using a translated nucleotide query. more...

Reset page   Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)   Clear

```
GGCGGCAAAA AACTTCAACA TACAATATGT CGCACCGCAC CCCCGCCAAA
CCTACGGGCT TTCCGGCGTA ACCACATTAA ATTCGCCCTA TGTCATCGAA
GACTATATTT TGCGCGAGAT TAAGAAAAAC CCGCATACGA GGTATGAAAT
TTATACCTTT TTCAGCGGCG CGGCGTTGAC GATGAAGGAT TTTCCCAATG
TGCACGTTTA CGCATTGAAA CCGGCTTCCC TTCCGGAACA TTATGGCTC
AAGCCGGTGT ATGCCCTGTT TACCCAATCC GGCATCCCGA TTTTGACATT T
```

Query subrange

From [   ]

To [   ]

Or, upload file   [Browse...] No file selected.

Genetic code   [Standard (1) ▾]

Job Title   [   ]
Enter a descriptive title for your BLAST search

☐ **Align two or more sequences**

Choose Search Set

Database   [Non-redundant protein sequences (nr) ▾]

Organism
Optional
[Enter organism name or id--completions will be suggested]   ☐ Exclude   [+]
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional
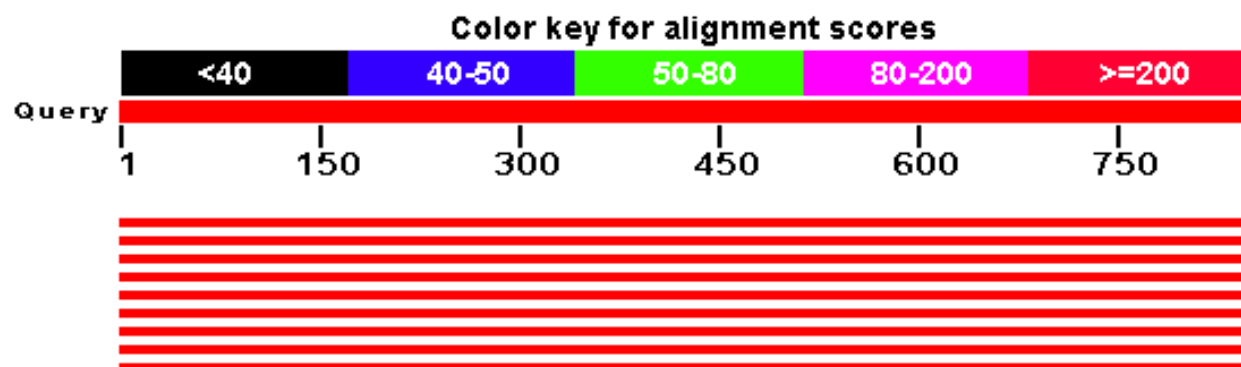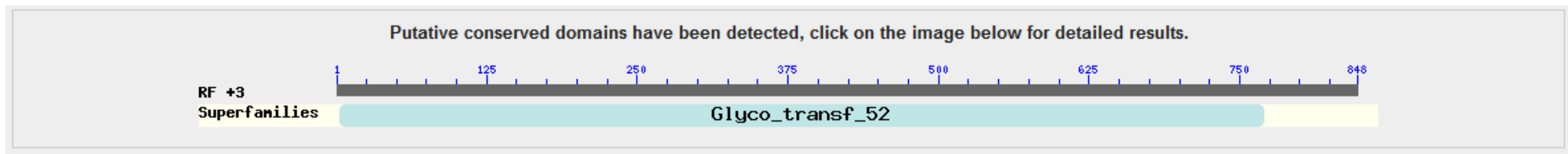☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query
Optional
[   ]   YouTube Create custom database
Enter an Entrez query to limit search

[**BLAST**]   Search database Non-redundant protein sequences (nr) using Blastx (search protein databases using a translated nucleotide query)
☐ Show results in a new window

[+] Algorithm parameters

NCBI BLAST output

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

RF +3

Superfamilies: Glyco_transf_52

Color key for alignment scores

| <40 | 40-50 | 50-80 | 80-200 | >=200 |

Query

1   150   300   450   600   750

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
alpha-2,3-sialyltransferase [Neisseria meningitidis]	576	576	99%	0.0	99%	AAC44543.1
CMP-N-acetylneuraminate-beta-galactosamide-alpha-2 3-sialyltransferase [Neisseria meningitidis]	576	576	99%	0.0	99%	WP_002234614.1
CMP-N-acetylneuraminate-beta-galactosamide-alpha-2 3-sialyltransferase [Neisseria meningitidis]	575	575	99%	0.0	99%	WP_014580539.1
CMP-N-acetylneuraminate-beta-galactosamide-alpha-2 3-sialyltransferase [Neisseria meningitidis]	574	574	99%	0.0	99%	WP_002236898.1
CMP-N-acetylneuraminate-beta-galactosamide-alpha-2 3-sialyltransferase [Neisseria meningitidis]	573	573	99%	0.0	98%	WP_002230437.1
CMP-N-acetylneuraminate-beta-galactosamide-alpha-2 3-sialyltransferase [Neisseria meningitidis]	573	573	99%	0.0	98%	WP_002239594.1

Sequence alignment output

CMP-N-acetylneuraminate-beta-galactosamide-alpha-2 3-sialyltransferase [Neisseria meningitidis]

Sequence ID: ref|WP_002234614.1| Length: 371   Number of Matches: 1

▷ See 29 more title(s)

Range 1: 86 to 367 GenPept   Graphics   ▼ Next Match   ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
576 bits(1485)	0.0	Compositional matrix adjust.	279/282(99%)	279/282(98%)	0/282(0%)	+3

```
Query  3    NEKYDYYFKQIKDKAERAYFFHLPYDLNKSFNVIPTMAELKVKSMLLPKVKRTYLASLEK  182
            NEKYDYYFKQIKDKAERAYFFHLPY LNKSFN IPTMAELKVKSMLLPKVKR YLASLEK
Sbjct  86   NEKYDYYFKQIKDKAERAYFFHLPYGLNKSFNFIPTMAELKVKSMLLPKVKRIYLASLEK  145

Query  183  VSIAAFLSTYPDAEIKTFDDGTGNLIQSSSYLGDEFSVNGTIKRNFARMMIGDWSIAKTR  362
            VSIAAFLSTYPDAEIKTFDDGTGNLIQSSSYLGDEFSVNGTIKRNFARMMIGDWSIAKTR
Sbjct  146  VSIAAFLSTYPDAEIKTFDDGTGNLIQSSSYLGDEFSVNGTIKRNFARMMIGDWSIAKTR  205

Query  363  NASDEHYTIFKGLKNIMDDGRRKMTYLPLFDASELKAGDETGGTVRILLGSPDKEMKEIS  542
            NASDEHYTIFKGLKNIMDDGRRKMTYLPLFDASELKAGDETGGTVRILLGSPDKEMKEIS
Sbjct  206  NASDEHYTIFKGLKNIMDDGRRKMTYLPLFDASELKAGDETGGTVRILLGSPDKEMKEIS  265

Query  543  EKAAKNFNIQYVAPHPRQTYGLSGVTTLNSPYVIEDYILREIKKNPHTRYEIYTFFSGAA  722
            EKAAKNFNIQYVAPHPRQTYGLSGVTTLNSPYVIEDYILREIKKNPHTRYEIYTFFSGAA
Sbjct  266  EKAAKNFNIQYVAPHPRQTYGLSGVTTLNSPYVIEDYILREIKKNPHTRYEIYTFFSGAA  325

Query  723  LTMKDFPNVHVYALKPASLPEDYWLKPVYALFTQSGIPILTF  848
            LTMKDFPNVHVYALKPASLPEDYWLKPVYALFTQSGIPILTF
Sbjct  326  LTMKDFPNVHVYALKPASLPEDYWLKPVYALFTQSGIPILTF  367
```

Gotchas

- BLAST will usually return a result!
  - The top match may not be significant
  - E-values up to 10 returned by default
- Domains are often shared by different proteins
  - Significant match to parts of the sequence, but not to others

Bioinformatics practical 1: Introduction to the Linux command line and BLAST

- This practical will be run using a Linux virtual machine that is installed on each of the laptops.

- Bioinformatics practicals and data are in the course GitHub repository:

- https://github.com/WCSCourses/Molecular_Approaches_Clinical_Microbiology_2024

- Navigate to 'modules/bioinformatics/Bioinformatics.md'