Sequence Alignment Practical

Partners

KTH VETENSKAP OCH KONST
THE UNIVERSITY OF EDINBURGH
BSC
IRB
MAX-PLANCK-GESELLSCHAFT
MANCHESTER 1824 The University of Manchester
Ian Harrow Consulting
archer

Funding

Horizon 2020
European Union Funding
for Research & Innovation

Reusing this material

Aims (i)

- Familiarise yourself with running parallel programs

- Run a real parallel code
    - on different numbers of cores
    - measure the time taken (elapsed wallclock time, cpu time)
    - observe increase in performance (Amdahl's law? – see later)

- Gain experience using an HPC system
    - transferring files
    - compiling
    - using a batch system

Aims (ii)

- Get you running on ARCHER
- Sort out practical details
    - usernames, passwords
    - idiosyncrasies of your Windows / Mac / Linux laptop
    - …
- Please ask for assistance if you need it!
    - demonstrators are here to help

Sequence Alignment

- Use HMMER (*phmmer)* to search a protein database for matches to a given protein sequence:

```
Alignments for each domain:
== domain 1  score: 2966.8 bits;  conditional E-value: 0
sp|P01024|CO3_HUMAN    1 mgptsgpsllllllthlplalgspmysiitpnilrleseetmvleahdaqgdvpvtvtvhdfpgkklvlssektvltpatnhmgnvtf 88
                         m ptsgpsllllll  lp+alg+pmys+itpnilrleseet+vleah  qg + v+vtvhdfp+kk vls+e t l     +++ vt+
          CO3_BOVIN    1 MKPTSGPSLLLLLLASLPMALGNPMYSMITPNILRLESEETVVLEAHGGQGTIQVSVTVHDFPAKKQVLSNENTQLNSNNGYLSTVTI 88
                         67************************************************************************************** PP

Alignments for each domain:
== domain 1  score: 2956.0 bits;  conditional E-value: 0
sp|P01024|CO3_HUMAN    1 mgptsgpsllllllthlplalgspmysiitpnilrleseetmvleahdaqgdvpvtvtvhdfpgkklvlssektvltpatnhmgnvtf 88
                         mg tsgp llllllt lplalg p+y+iitpn+lrlesee +vleah+ qgd+ v+vtvhdfp+k+ vlsse t l  a n+++ v +
            CO3_PIG    1 MGSTSGPRLLLLLLTSLPLALGDPIYTIITPNVLRLESEEMVVLEAHEGQGDIRVSVTVHDFPAKRQVLSSETTTLNNANNYLSTVNI 88
                         89************************************************************************************** PP

Alignments for each domain:
== domain 1  score: 1343.7 bits;  conditional E-value: 0
sp|P01024|CO3_HUMAN  938 mnktvavrtldperlgregvqkedippadlsdqvpdtesetrillqgtpvaqmtedavdaerlkhlivtpsgcgeqnmigmtptviav 1025
                         mnktvavrtldpe+lg+ gvqke+ip ad+sdqvp teset+illqgtpvaqmteda+d erlkhlivt sgcgeqnmi+mt tviav
          CO3_RABIT    1 MNKTVAVRTLDPENLGQGGVQKEEIPSADISDQVPGTESETKILLQGTPVAQMTEDAIDGERLKHLIVTGSGCGEQNMIAMTHTVIAV 88
                         8*************************************************************************************** PP
```

phmmer

- Alternative to BLAST-style sequence alignment
- Generates Hidden Markov Model profile for query sequence
- Three-stage filter for probabilistic alignment scoring against target database
- Returns top scoring matches, subject to chosen thresholds

phmmer – parallel execution

- Two parallel execution modes (mutually exclusive):
  - Threads (“pthreads”) - restricted to single node
  - MPI - can span many nodes
- Both use task farm / work queue model
- These concepts will be explained later in the course..

Practical

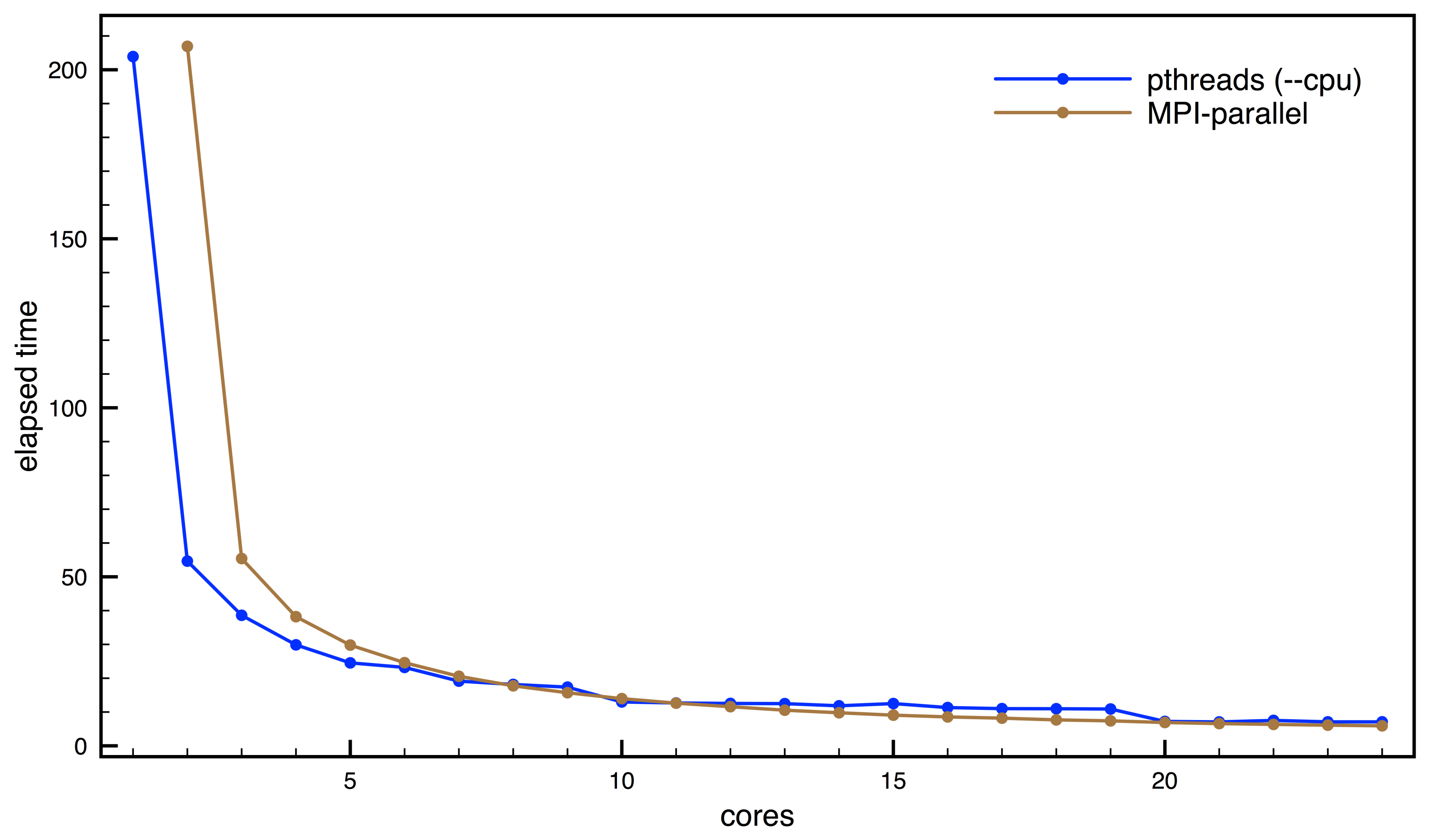
- Log on to ARCHER
- Download HMMER, unpack, and compile
- Run on one core on login node
- Run in parallel on compute nodes using batch system
    - Threaded
    - MPI
- Record times for different numbers of cores
- Plot performance figures

- See the instruction sheet..

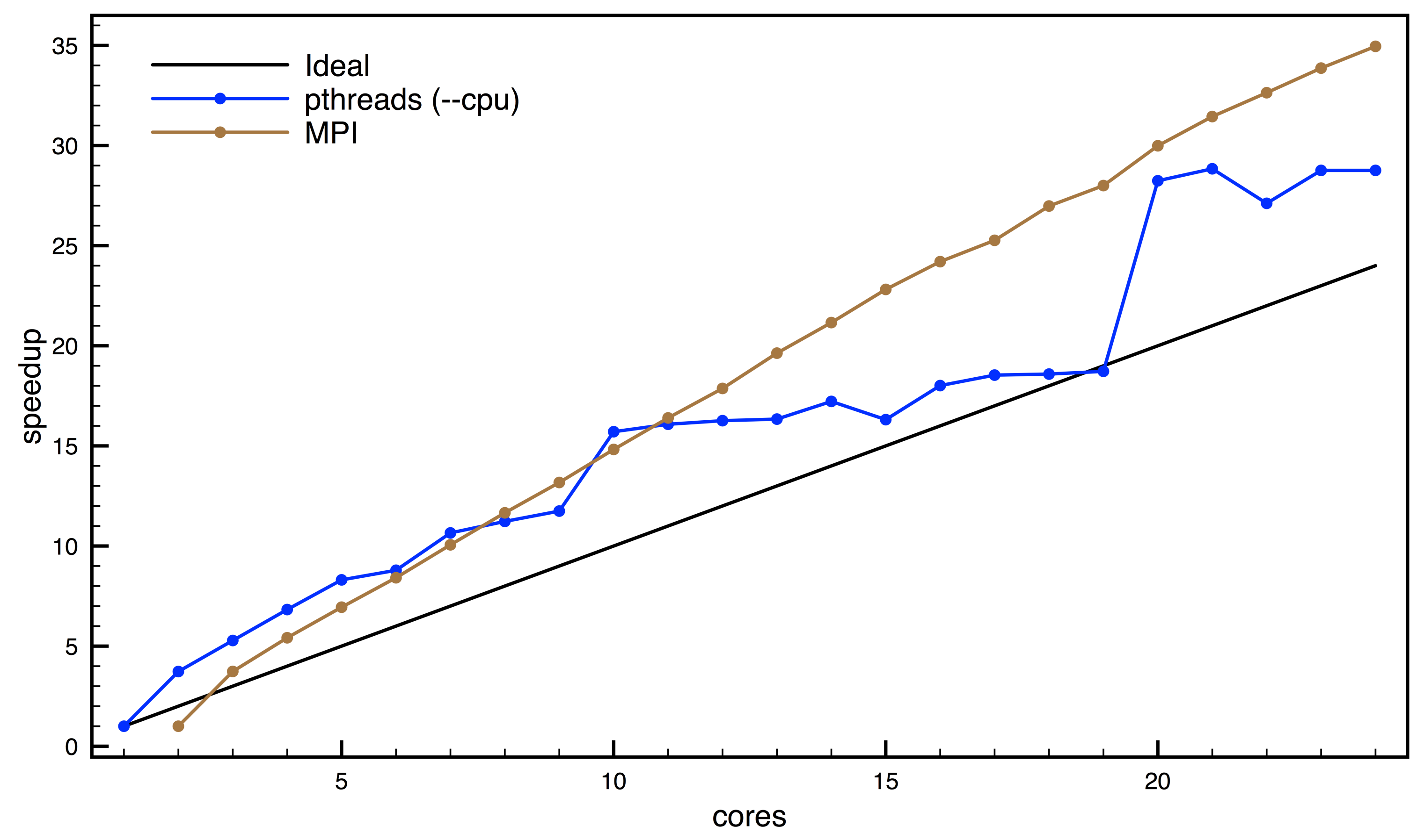
Sequence Alignment Practical

Outcomes

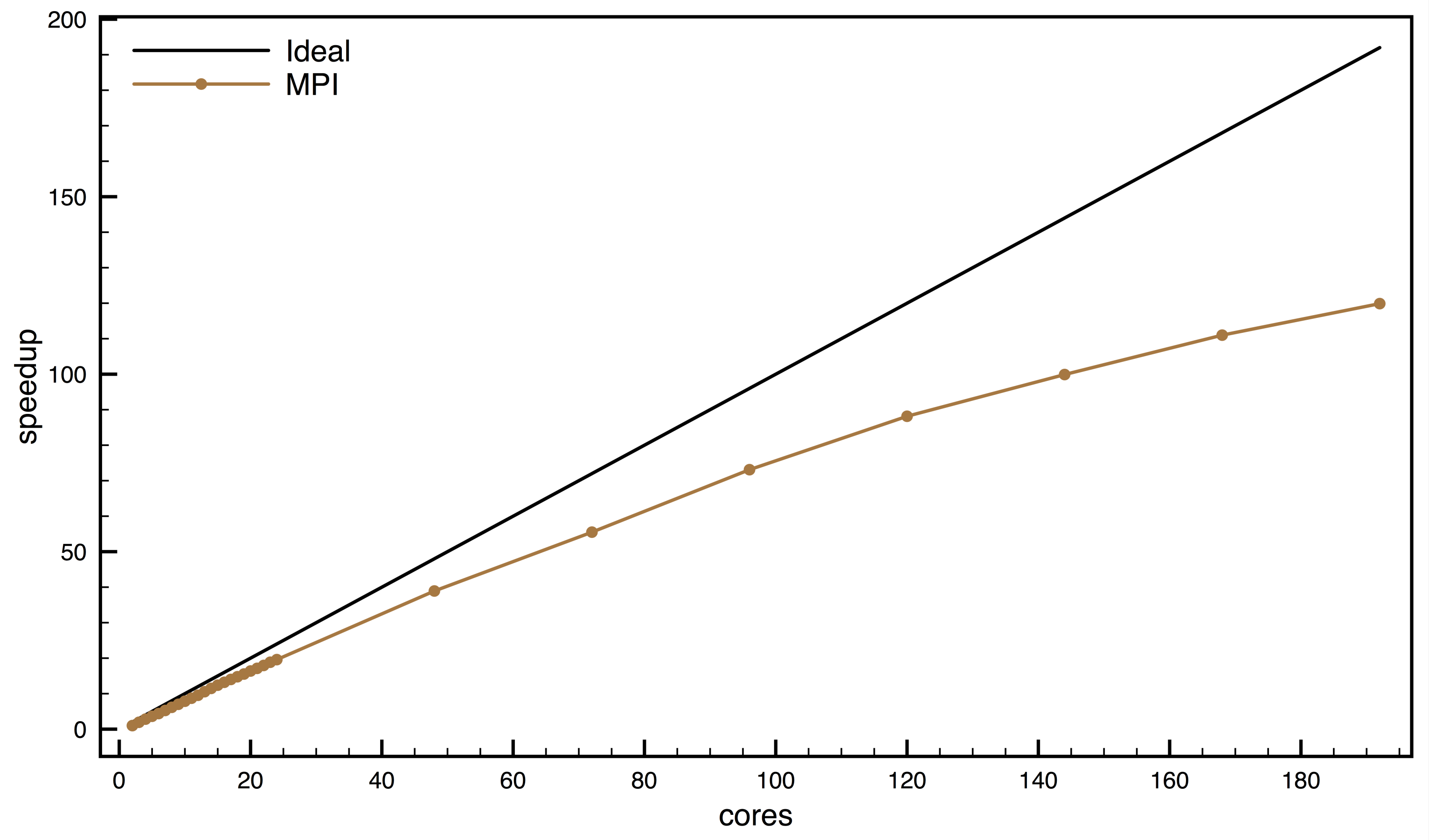
C3 protein & Swiss-Prot mammals



C3 protein & Swiss-Prot mammals



C3 protein & Swiss-Prot all



C3 protein & Swiss-Prot all