Sequence Alignment Practical

Horizon 2020
European Union Funding
for Research & Innovation

Reusing this material

Aims

- Gain experience using an HPC system
    - Transferring files
    - Compiling source code
    - Using a batch system

- Run a real bioinformatics software package in parallel
    - Run on different numbers of cores
    - Measure time taken (elapsed **wallclock time**, **cpu time**)
    - Observe how performance scales
    - Relate performance to how software solves problem in parallel

Sequence Alignment

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

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


Alignments for each domain:
== domain 1  score: 2956.0 bits;  conditional E-value: 0
sp|P01024|CO3_HUMAN     1 mgptsgpslllllllthlplalgspmysiitpnilrleseetmvleahdaqgdvpvtvtvhdfpgkklvlssektvltpatnhmgnvtf 88
                          mg tsgp lllllll lplalg p+y+iitpn+lrlesee +vleah+ qgd+ v+vtvhdfp+k+ vlsse t l  a n+++ v +
             CO3_PIG     1 MGSTSGPRLLLLLLLTSLPLALGDPIYTIITPNVLRLESEEMVVLEAHEGQGDIRVSVTVHDFPAKRQVLSSETTTLNNANNYLSTVNI 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 ARCHER2

- Download HMMER source code, unpack, and build (compile)

- Run HMMER on one core on login node

- Run HMMER in parallel on compute nodes using batch system

  - Threaded

  - MPI

- Record runtimes for different numbers of cores

- Plot performance figures

- See the instruction sheet..