An Investigation on the Efficiency of Sequence Alignment Tools

Justin Chao, Chase Meyer, Bria Lacour

Background/Research Significance

- Sequence Alignment (SA) - the alignment of biological sequences (DNA, RNA, or protein), which are assumed to have an evolutionary relationship, meaning that they share a common ancestor
  - Global (Needleman-Wunsch) and local (Smith-Waterman)
  - Requires computation of very large matrices
- As of now, many SA tools available don't take full advantage of the computational power of the TACC supercomputers.
- SA can give insight into the function of proteins and specific genes, prove homology (the sharing of a common ancestor) based on the calculated similarity

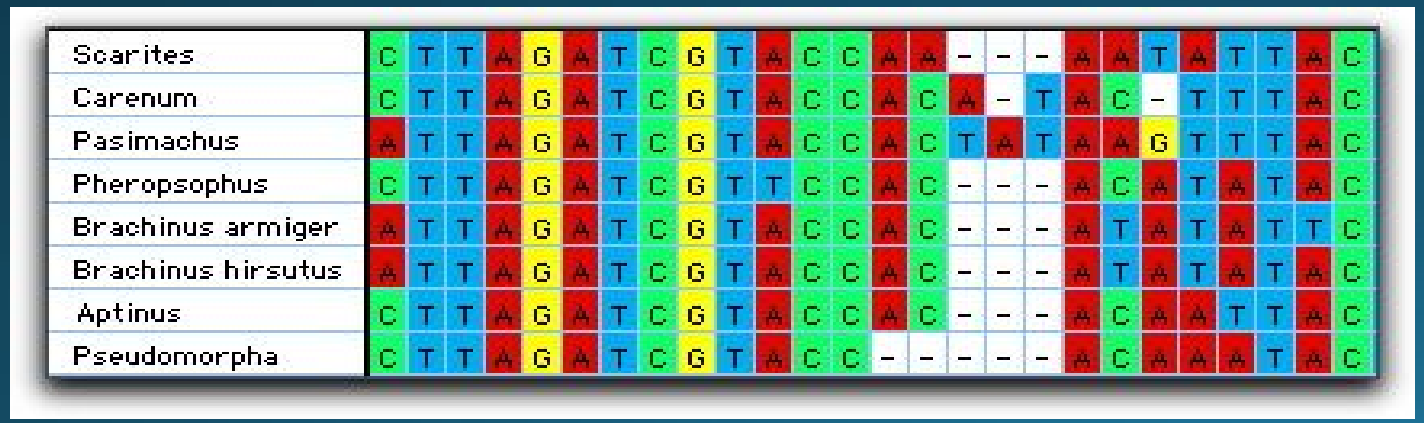

Figure 1: Example of BLAST Multiple Sequence Alignment Results

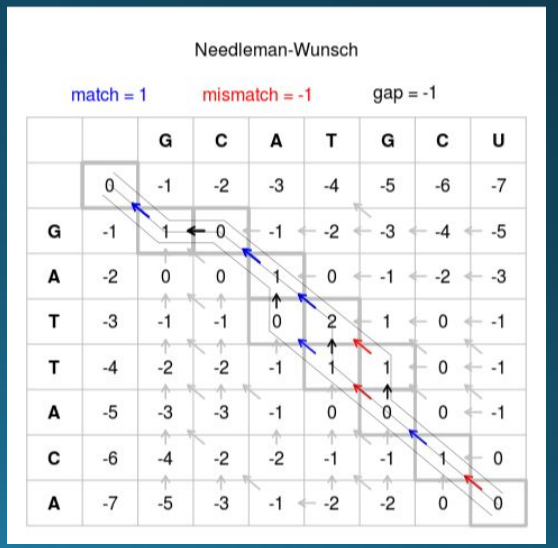

Figure 2: Needleman-Wunsch pairwise sequence alignment

Goals

- Create a global sequence alignment tool using Needleman-Wunsch Algorithm
- Optimize for supercomputer and compare efficiency to industry standard MSA tools
  - Ended up writing our code dynamically
  - Profiled code to reduce bottlenecks in performance

Sequence Alignment Output Example

```
meepqsdpsvepplsqetfsdlwkllpennvlsplpsqamddlmlspddieqwftedpgdeaprmpeaappvapapaaptpaapapapswp
medsqsdmsielplsqetfsclwkllppddilpttatgspnsmedlflpqdvaellegpeealqvsapaaqepgteapapvapasatpwp
91
90
reverse seq 1: MEEPQSDPSVEPPLSQETFSDLWKLLPENNVL------SPLPSQAMDDLMLSP-DDIEQWFTEDPGDEAPRMPEAAPPV-APAPA-AP-T--PA--APAPA-PSWP
reverse seq 2: MEDSQSDMSIELPLSQETFSCLWKLLPPDDILPTTATGSP-NS--MEDLFL-PQDVAE--LLE--G------PEEALQVSAPA-AQEPGTEAPAPVAPASATP-WP
Score = 265
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVL------SPLPSQAMDDLMLSP-DDIEQWFTEDPGDEAPRMPEAAPPV-APAPA-AP-T--PA--APAPA-PSWP
ME  QSD S E PLSQETFS LWKLLP    L      SP  S  M DL L P D  E    E  G      PE A  V APA A  P T  PA  APA A P WP
MEDSQSDMSIELPLSQETFSCLWKLLPPDDILPTTATGSP-NS--MEDLFL-PQDVAE--LLE--G------PEEALQVSAPA-AQEPGTEAPAPVAPASATP-WP
```

Alignment of TP53 in Mouse and Humans

Demonstration of our code…

Profiling Results

```
                     Call graph (explanation follows)


granularity: each sample hit covers 2 byte(s) no time propagated

index % time    self  children    called     name
                0.00    0.00     393/154073      finalize [4]
                0.00    0.00  153680/154073      main [12]
[1]      0.0    0.00    0.00  154073          match_score [1]
-----------------------------------------------
                0.00    0.00  153272/153272      main [12]
[2]      0.0    0.00    0.00  153272          max_array [2]
-----------------------------------------------
                0.00    0.00       2/2           finalize [4]
[3]      0.0    0.00    0.00       2          reverseSeq [3]
-----------------------------------------------
                0.00    0.00       1/1           main [12]
[4]      0.0    0.00    0.00       1          finalize [4]
                0.00    0.00     393/154073      match_score [1]
                0.00    0.00       2/2           reverseSeq [3]
-----------------------------------------------
                0.00    0.00       1/1           main [12]
[5]      0.0    0.00    0.00       1          printMatrix [5]
-----------------------------------------------
```

Comparison to Other SA Tools

Each alignment method was measured for efficiency by running the alignment of the sequences with the time command

	Open Source Code - NW	Ours- NW	MAFFT	MUSCLE	Clustal Omega
Real Time	0.026s	0.024s	0.004s	0.014 s	0.005s
User Time	0.016s	0.016s	0.000s	0.004 s	0.000s
System Time	0.006s	0.004s	0.000s	0.004 s	0.000s

Figure 3: Table of Efficiency of Selected MSA Tools