BLAST – DNA Search

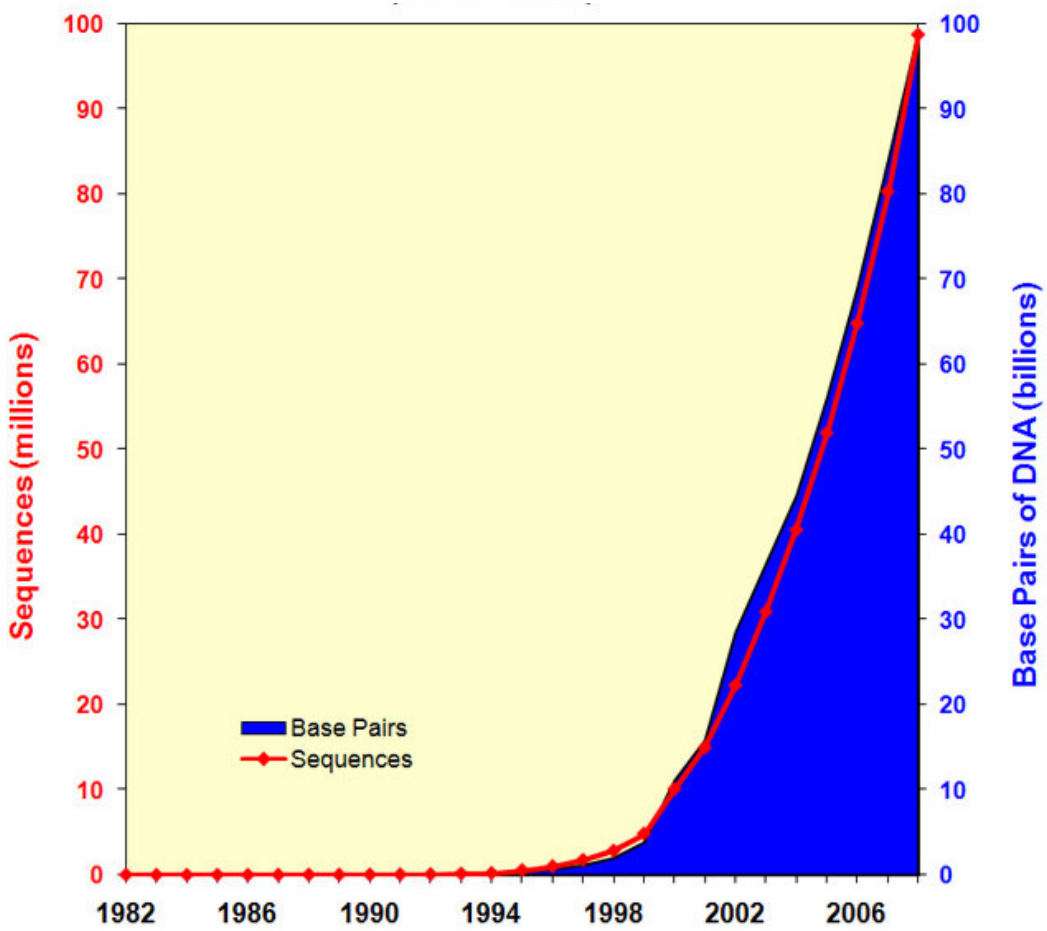
Helen Durrant

(Excuse the lack of a pretty background, had to remove it because the file was too big..)

DNA Sequencing

- DNA and proteins can be sequenced.
- Being able to compare sequences is important for biologists.
- A sequence database contains known sequences of any or all organisms.

GenBank – Growth Statistics (1982 – 2008)



Basic Logical Alignment Search Tool

- The need for speed in sequence comparison is obvious..
- The BLAST algorithm involves comparing a given sequence against a sequence database, to find similar sequences.

A Quick Note on Sequence Comparison

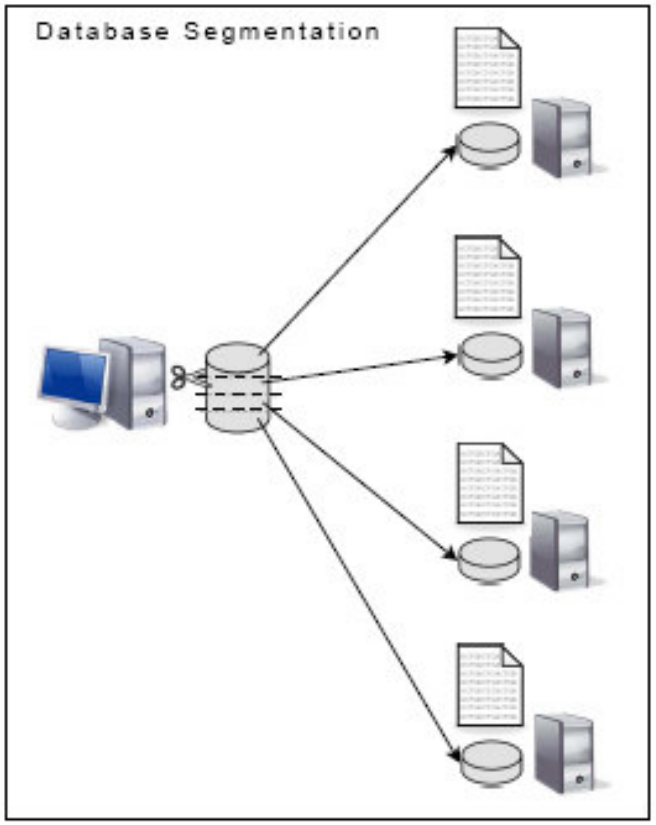
- Evolution causes amino acid sequences of an organism's proteins to gradually alter.

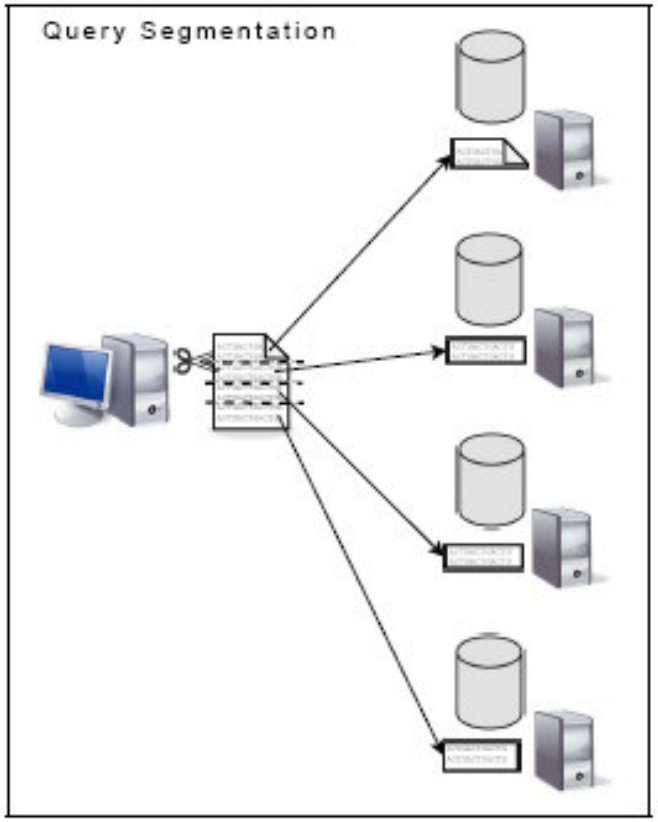
- Each amino acid is more or less likely to mutate into various other amino acids.

Speeding Up BLAST

- Heuristics increase the speed of the comparison algorithm.
- Our problem is embarrasingly parallel..
- Fine grained parallelism can be acheived at hardware level.

Database vs. Query Segmentation





Database Segmentation

Advantages:

- Less demanding memory requirements
- Can utilise a large number of machines regardless of number of queries

Disadvantage:

- Higher parallel search overhead, local results need to be merged globally

Query Segmentation

Advantages:

- Low parallelisation overhead
- Can achieve linear speedup
- Can use optimisation

Disadvantages:

- Amount of memory required
- Paging can occur
- Load imbalance

mpiBLAST

- Open source, parallel BLAST
- Master-slave model
- Searching done in workers:
  - Search all queries against a subset of DB
  - Generate partial results
- Output processing done in master:
  - Merge partial results from workers
  - Fetch needed sequence data, output results

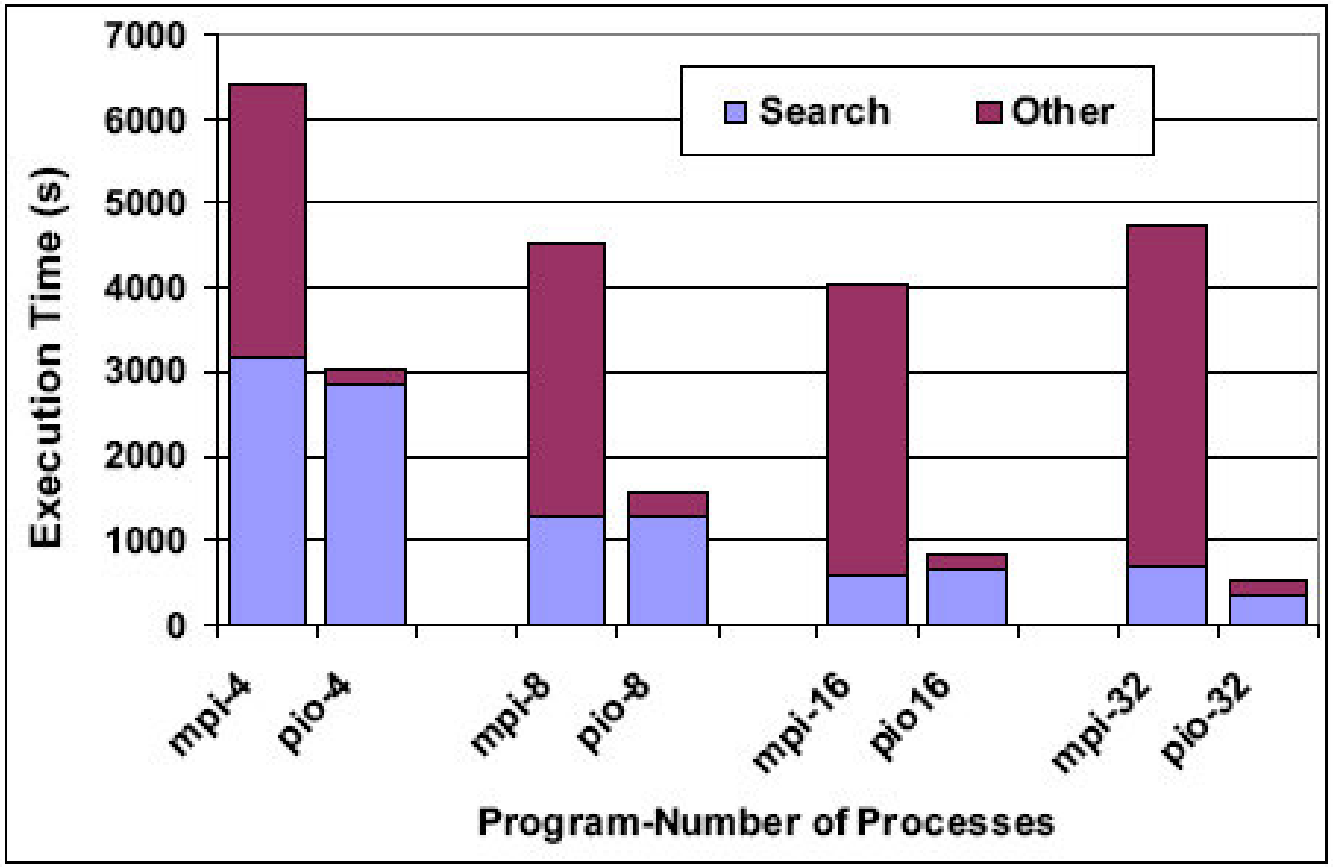
mpiBLAST Development

- Several important issues needed to be addressed:
  - Static database partitioning (inflexible)
  - Data handling overhead (limits scalability)
  - Output handling by the master (inefficient)

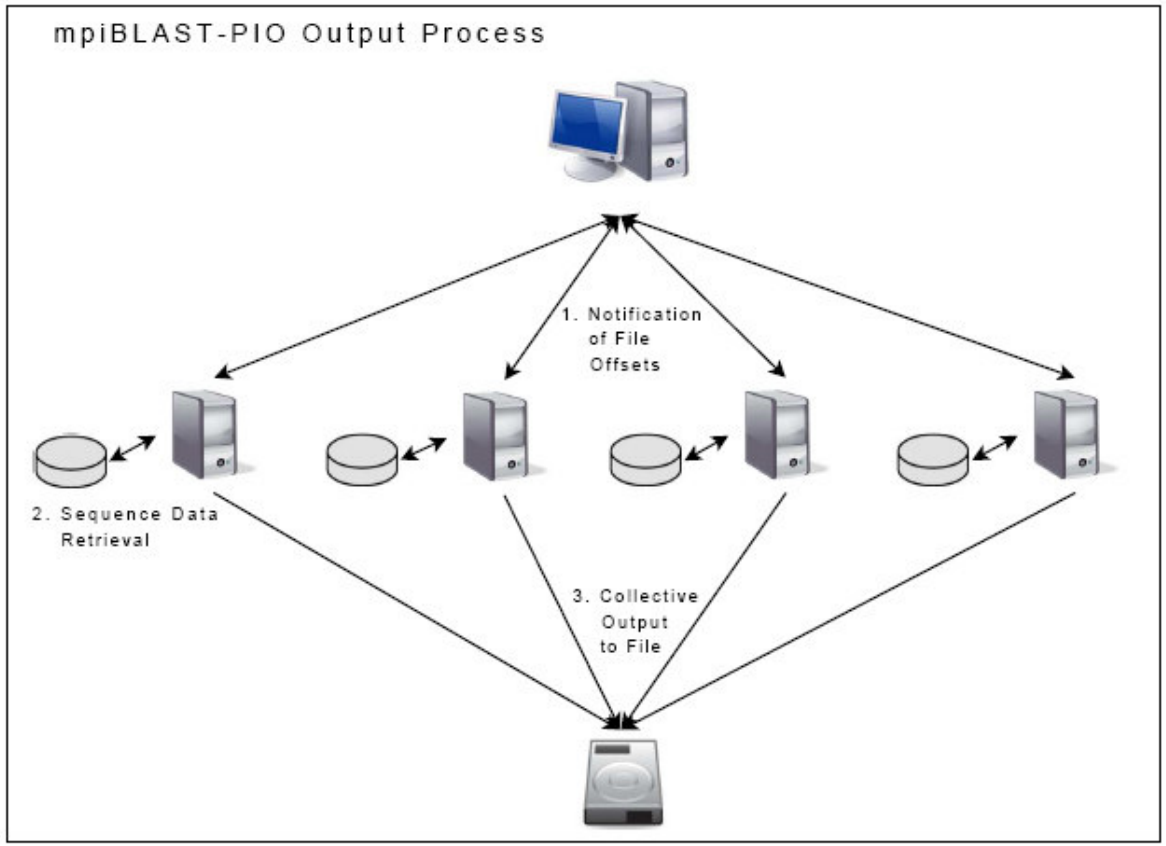
pioBLAST

- Research prototype of an efficient parallel BLAST.
- Implements:
  - Dynamic database partitioning
  - Caching of important data
  - Parallel I/O on shared files
  - More efficient processing of results

mpiBLAST-1.2 vs. pioBLAST



mpiBLAST-1.4 to mpiBLAST-PIO



Future Directions for mpiBLAST

- Fault tolerance
- Database updates
- Task threading at the master

Conclusion

- DNA sequence comparisons are both computationally expensive and embarrassingly parallel.
- mpiBLAST has achieved super linear speedup of sequential sequence comparison algorithms.
- There are many other, less widely used techniques that I haven't covered.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipmanl, D.J.: Basic local alignment search tool. Journal of Molecular Biology 215 (1990) 403410

[2] Archuleta J.S., Tilevich E., F.W.: A maintainable software architecture for fast and modular bioinformatics sequence search. In: 23rd IEEE International Conference on Software Maintenance. (2007)

[3] Archuleta J.S., Feng W., T.E.: A pluggable framework for parallel pairwise sequence search. In: International Conference of the IEEE Engineering in Medicine and Biology Society. (2007)

[4] Baxevanis, A.D., Ouellette, B.F.F., eds.: Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Wiley-Interscience (2005)

[5] Braun, R.C., Pedretti, K.T., Casavant, T.L., Scheetz, T.E., Birkett, C.L., Roberts, C.A.: Parallelization of local blast service on workstation clusters. Future Generation Comp. Syst. 17 (2001) 745754

[6] Chang, C.: Blast implementation on BEE2. Technical report, University of California, Berkeley (2004)

References

[7] Darling, A.E., Carey, L., Feng, W.C.: The design, implementation, and evaluation of mpiBLAST. In: In Proceedings of ClusterWorld 2003. (2003)

[8] Goddard, C.J.: Analysis and Abstraction of Parallel Sequence Search. PhD thesis, Faculty of the Virginia Polytechnic Institute and State University, Blacksburg, Virginia (2007)

[9] Grant, J.D., Jr., R.L.D., Manion, F.J., Ochs, M.F.: BeoBLAST: distributed blast and psi-BLAST on a beowulf cluster. Bioinformatics 18 (2002) 765766

[10] Kim, H.S., Jang, W.H., Han, D.S.: Communication protocols and message formats for BLAST parallelization on cluster systems. In: AINA Workshops, IEEE Computer Society (2008) 820825

[11] Rangwala, H., Lantz, E., Musselman, R., Pinnow, K., Smith, B., Wallenfelt, B.: Massively parallel BLAST for the blue gene/l. In: In High Availability and Performance Computing Workshop. (2005)

[12] Lin, H., Ma, X., Chandramohan, P., Geist, A., Samatova, N.F.: Efficient data access for parallel BLAST. In: IPDPS, IEEE Computer Society (2005)