



Parallel Sequence Search with Accelerators

Professor Wu FENG, feng@cs.vt.edu, 540-231-1192

Depts. of Computer Science and Electrical & Computer Engineering





What is Sequence Search?

- Search for a particular sequence in a database of known sequences.

Why is Sequence Search Important?

Nucleic Acids Research

[Journal List](#) > [Nucleic Acids Res](#) > v.34(Database issue); Jan 1, 2006

Nucleic Acids Res. 2006 January 1; 34(Database Issue): D668–D672.
Published online 2005 December 28. doi: 10.1093/nar/gkj067.

[Copyright](#) © The Author 2006. Published by Oxford University Press. All rights reserved

DrugBank: a comprehensive resource for *in silico* drug discovery and exploration

Science



Magazine

News

STKE

Careers

Multimedia

Collections

The Search for Unrecognized Pathogens

David A. Relman

The distribution and diversity of microorganisms in the world are far greater than have been previously appreciated. Molecular, cultivation-independent methods have played a key role in this insight. To what extent do humans remain ignorant of microbial diversity within the human body and the settings in which microorganisms cause human disease? In addition to implicating microbial agents in nontraditional infectious diseases, the use of methods such as broad-range polymerase chain reaction, representational difference analysis, expression library screening, and host gene expression profiling may force a reassessment of the concepts of microbial disease causation.

Linux Tackles Deadly SARS Virus

By [Dan Orzech](#)

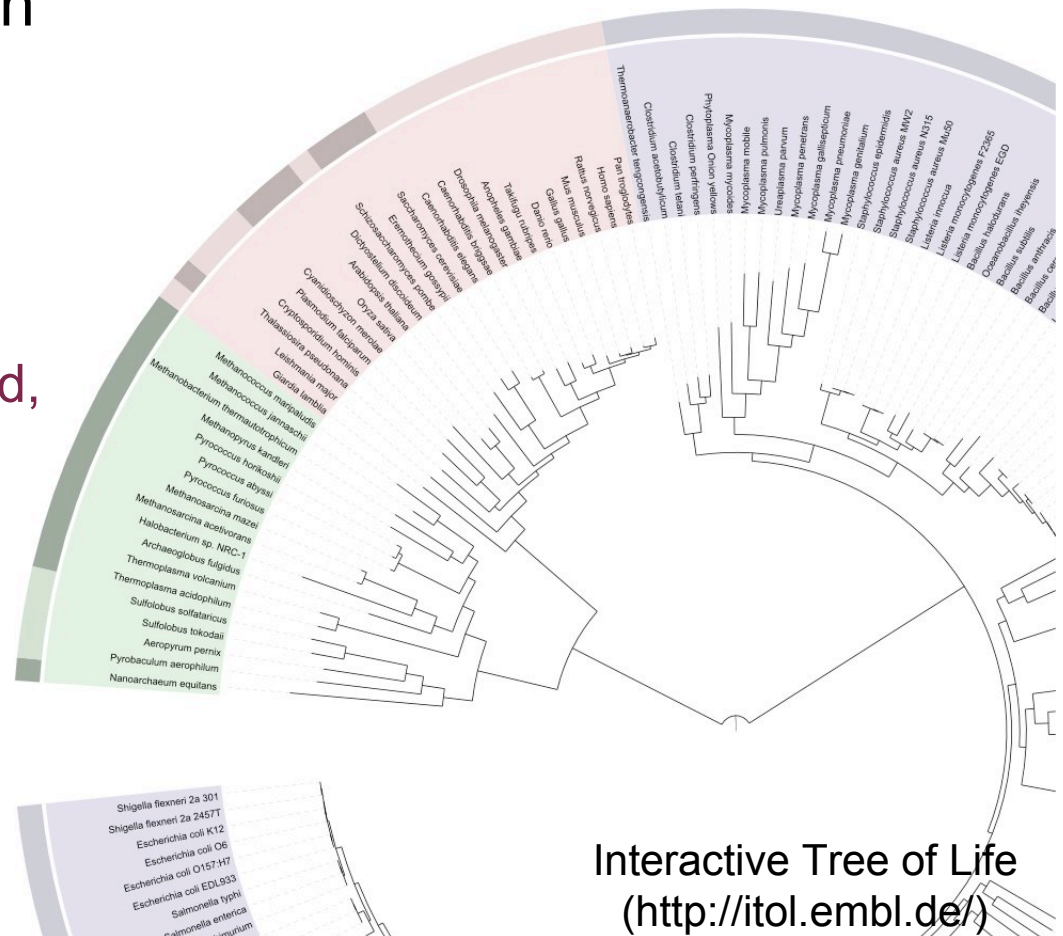
April 16, 2003: With the help of a cluster of PCs running Linux, Canadian researchers have made a major breakthrough in tackling the deadly SARS virus.

With the help of a cluster of PCs running Linux, Canadian researchers have made a major breakthrough in tackling the deadly SARS virus.



What is “Wrong” with Sequence Search?

- Sequence search is *the* bottleneck for most bioinformatics research
- Tree of Life
 - Search for ancestry to improve quality of life
 - Only 0.1M sequenced, and 100M undiscovered,
- Pathogen Detection
 - Search for uniqueness to identify threats in real time



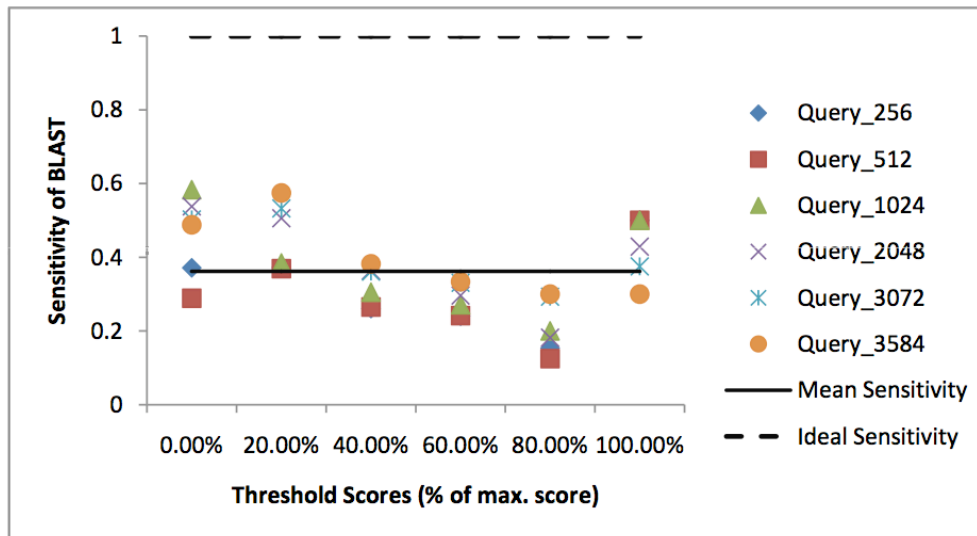


Solutions

mpiBLAST v1.5

- Accelerates discovery and innovation in pairwise sequence search via a cluster.
- Scales with 93% efficiency onto IBM Blue Gene/P (to appear at SC|08).

Problem: Compromised sensitivity

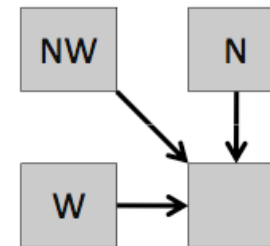
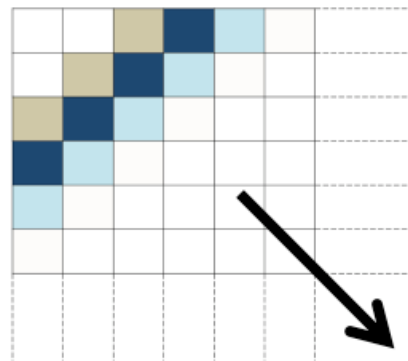


Nodes	Time	Speedup
1	176,880	1.00
2	68,640	2.58
4	39,109	4.52
8	7,730	22.88
16	3,683	48.03
32	2,321	76.21
64	1,021	173.24
128	579	305.49



Smith-Waterman at a Glance

- Features
 - Sequential: Dynamic Programming
 - One of the 13 dwarves of high-performance computing (HPC)
 - Robust Parallel Transformation: Wavefront Algorithm

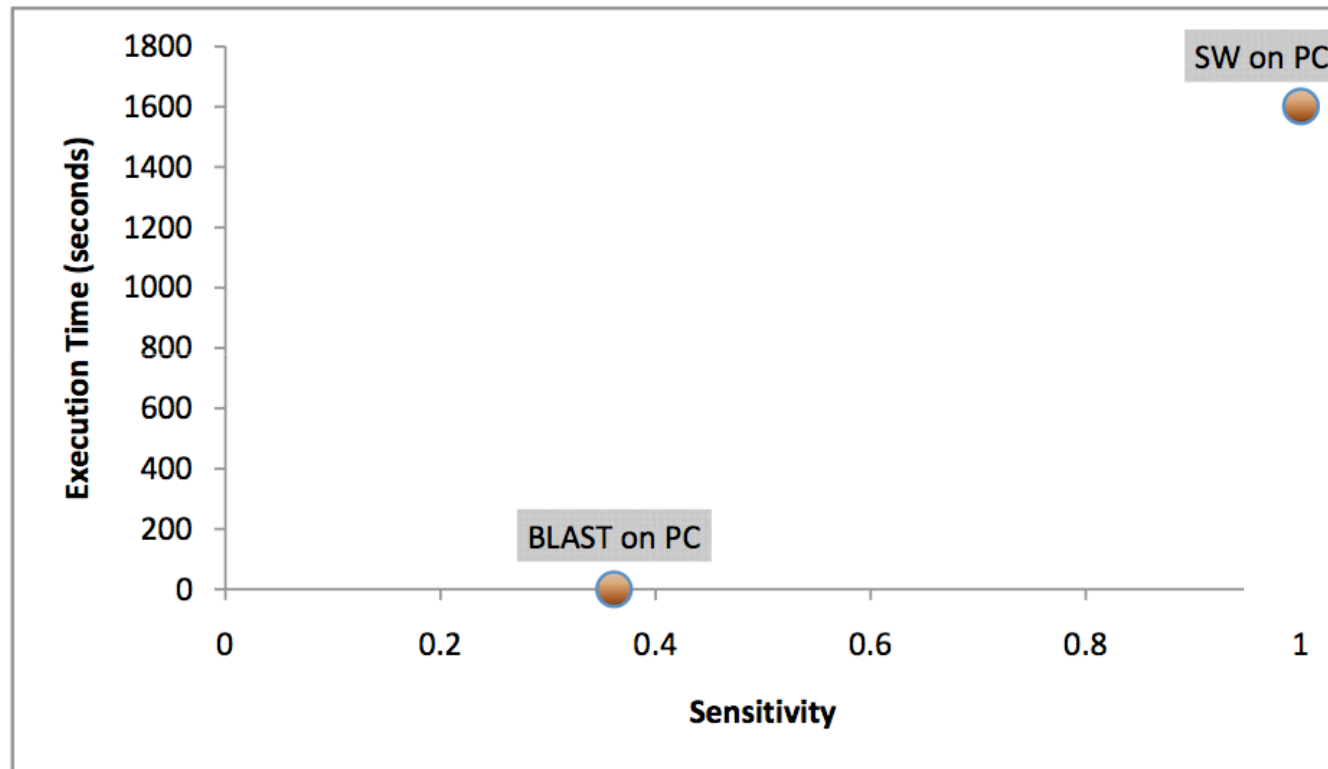


- Mapping
 - ‘Tile’ computation per accelerator unit (i.e., Cell SPE, GPGPU multiprocessor/SIMD unit).



Smith-Waterman (SW) Algorithm

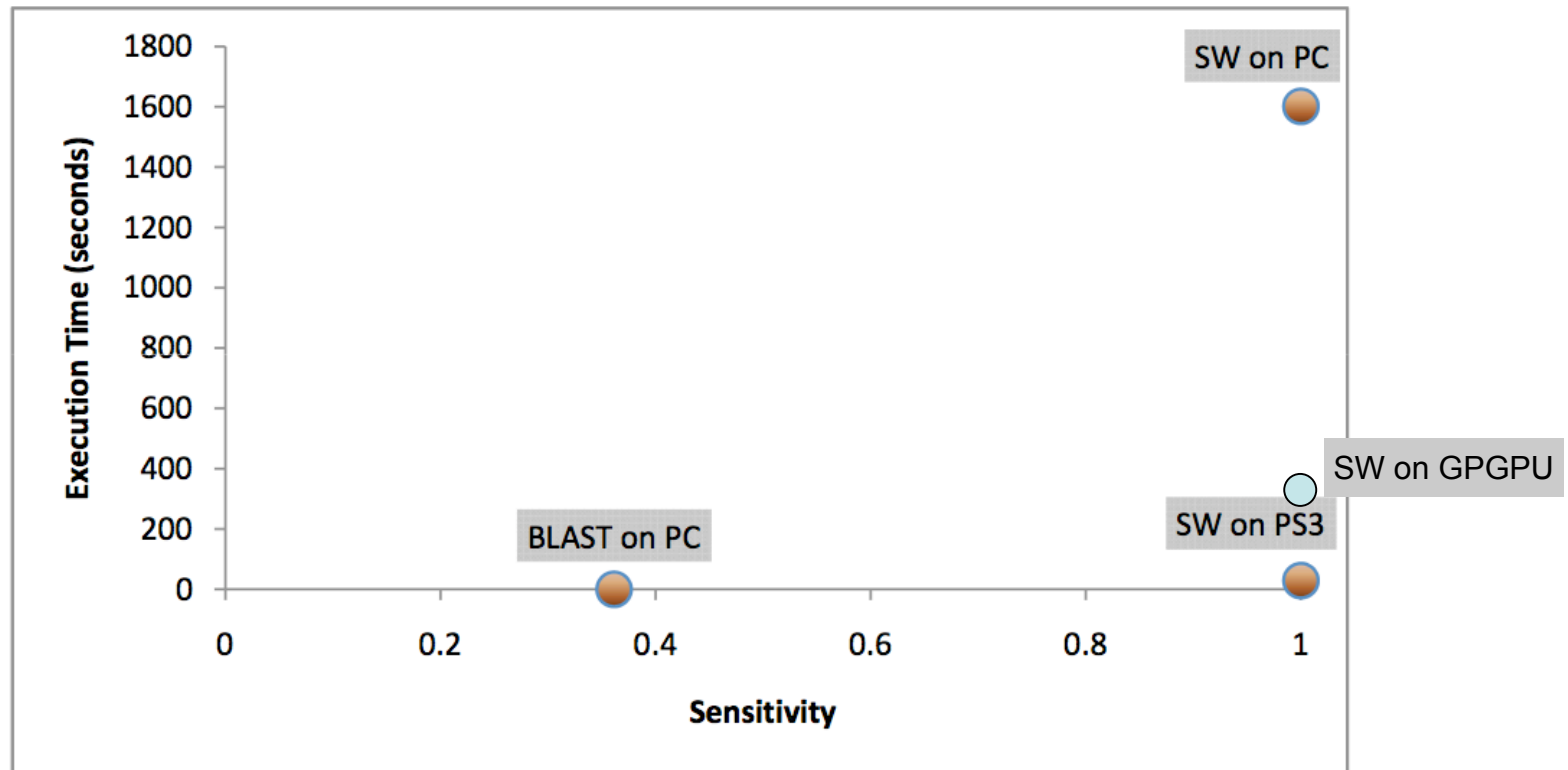
- *Optimal*, but *SLOW*, local sequence alignment algorithm





Smith-Waterman (SW) Algorithm

- *Optimal*, but *SLOW*, local sequence alignment algorithm





Smith-Waterman on the Cell

Sequence IDs	Host Processor	Host + 16 Accelerator Cores
Seq_1	10.1876	0.119327
Seq_2	24.1743	0.224449
Seq_3	31.186	0.287019
Seq_4	40.9207	0.382252
Seq_5	51.7262	0.463263
Seq_6	61.6872	0.553428
Seq_7	71.9487	0.621144
Seq_8	91.4295	0.843931

82x to 118x speedup