A New Distributed System for Large-Scale Sequence Analysis

www.cs.virginia.edu **Douglas Blair and Gabriel Robins**

Department of **Computer Science**
School of Engineering and Applied Science
University of Virginia

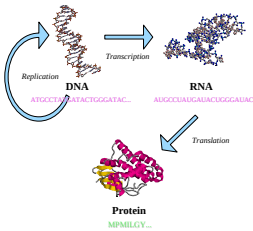
(434) 982-2207 {dmb4x, robins}@cs.virginia.edu

Presented at **Intelligent Systems for Molecular Biology**, August 19-23, 2000, San Diego, CA

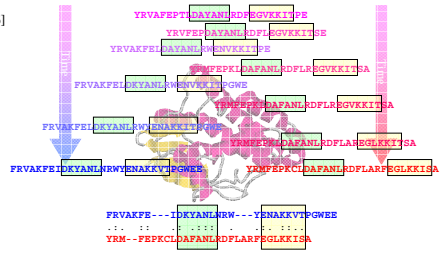
Bioinformatics

"Central Dogma" of Molecular Biology

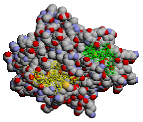
Basic mechanisms in all living organisms [Crick, ~1956]

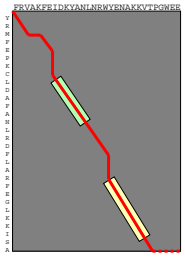


Proteins and Evolution



Sequence Alignment





Algorithms and Statistics

Sequence Comparison Dynamic Programming Algorithms:

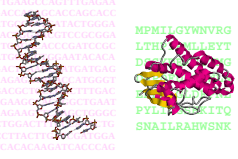
- Needleman-Wunsch [Needleman & Wunsch, 1970]
- Smith-Waterman [Smith & Waterman, 1981]
- Smith-Waterman with gaps [Gotoh, 1982]
- FASTA [Pearson & Lipman, 1988]
- BLAST [Altschul et al., 1990]

Statistical Significance:

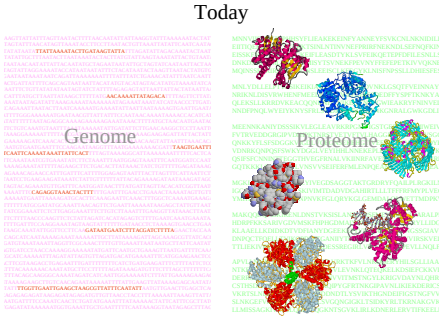
- Distribution of Smith-Waterman scores [Karlin & Altschul 1990]
- Distribution of n SW scores [Karlin & Altschul 1993]
- Empirical distribution for gapped scores [Altschul & Gish 1996]

Data Avalanche

Yesterday



Today



Genomes and Proteomes

Organism	Year Completed	Genome Size (Base Pairs)	Proteome Size (# of Proteins)
Mycoplasma Genitalium	1995	~588,000	480
Escherichia Coli	1997	~4,600,000	4,289
Saccharomyces Cerevisiae	1997	~11,000,000	~6,600
Caenorhabditis Elegans	1998	~86,000,000	~14,300
Drosophila Melanogaster	2000	~137,000,000	~13,500
Homo Sapiens	~2000	~3,100,000,000	~30,000-60,000

~~≥ 35~~

31 complete microbial genomes (87 in progress)

Many new microbial genomes every year

Many other higher organisms' genomes being sequenced

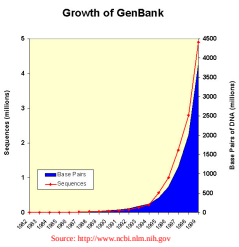
Runaway Growth

Advances in sequencing technology

Exponentially increasing data volume

GenBank: 8.6 billion nucleotides (Jun 2000)
9.5 billion nucleotides (Aug 2000)

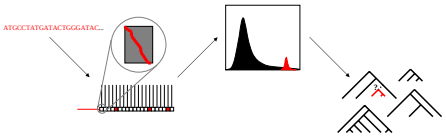
Data growing faster than computer speeds:
- Data volume doubles every 12 months
- Moore's Law: 18-month doubling time



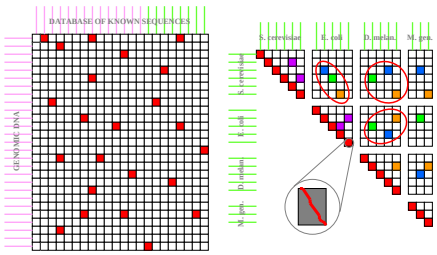
Paradigm Shift

Old Sequence Analysis Paradigm

Record new experimentally derived sequence
Compare to known sequences in database
Determine statistical significance of comparison scores
Deduce biological and evolutionary relationships

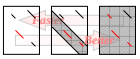


New Sequence Analysis Paradigm: Genomics and Comparative Genomics



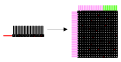
Challenges

- Computing power growing less quickly than data volume



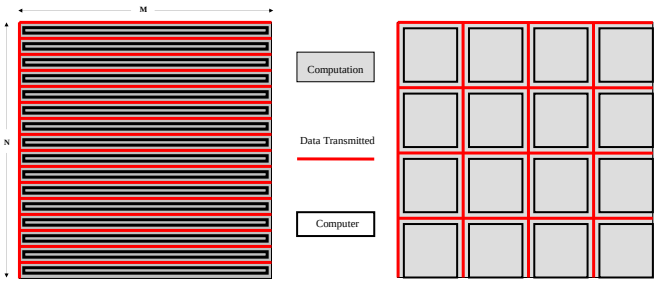
- Heuristic methods are faster but less sensitive

- Computation grows *quadratically* with data volume



- Current parallel implementations scale poorly

Solution: Break the Data Bottleneck

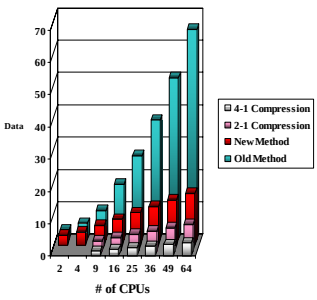


M+(N/*k*) Data/CPU
(*k*×M)+N Total Data

k Computers
(M×N)/*k* Work/CPU
M×N Total Work

(M+N)/√*k* Data/CPU
(M+N)×√*k* Total Data

Data Transmitted



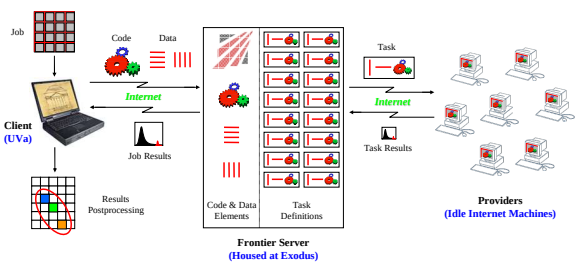
Old vs. New

Data ≡ Work	Data << Work -- "Square" tasks minimize data/computation
Data expansion takes as long as or longer than comparison operations	Data expansion relatively inexpensive - compression becomes worthwhile
Entire library required everywhere simultaneously (Poor NFS server…)	Tasks self-contained, compact, independent - exquisitely parallel
Parallelism constrained to number of machines not starved for data	Parallelism constrained only by the available number of machines

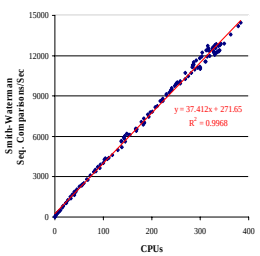
Paves the way for Massively Parallel Computation
Ability to take advantage of more processors encourages use of more sensitive, computationally demanding techniques

Implementation & Results

Test Platform: Parabon Frontier



Scalability



Future Directions

- Further Smith-Waterman optimizations
- Investigation of novel methods for estimating statistical significance
- Other methods (BLAST, FASTA, HMMs, GeneWise, etc.)
- Data compression
- Implementation of DNA-protein and DNA-DNA comparisons
- Large-scale structure-structure comparison
- Large-scale sequence-structure threading/comparison
- Human Genome vs. GenBank scale searches
- Java 1.3 JVM for Provider Compute Engine (Faster than C!)
- Other projects (e.g. Maximum Likelihood Tree Searches)

"Determine never to be idle. No person will have occasion to complain of the want of time, who never loses any. It is wonderful how much may be done, if we are always doing."

-- **Thomas Jefferson, May 5, 1787**