

BioPerf: A Benchmark Suite to Evaluate High-Performance Computer Architecture on Bioinformatics Applications

David A. Bader,
Georgia Tech.

Yue Li Tao Li
University of Florida

Vipin Sachdeva
UNM

Oct. 7, 2005



Motivation

- Bioinformatics is becoming an increasingly important domain
- What is Bioinformatics?
- Computational challenge of bioinformatics applications

Previous Work

- General benchmark suites-SPEC
- Domain-specific benchmarks, e.g. TPC, EEMBC, SPLASH, SPLASH-2
- Few special benchmark for bioinformatics

Contributions of this Work

- Propose a benchmark suite-BioPerf which spans a wide variety of bioinformatics application
- Performance study on PowerPC G5 and the Mambo simulator from IBM.

Outline

- Background
- BioPerf Benchmark Suite
- Performance Study of BioPerf benchmarks
- Conclusions

The Area of Bioinformatics

- Sequence Analysis
- Sequence Homology and Gene Finding
- Phylogeny Analysis
- Protein Structure Analysis

Selected Benchmarks

Sequence Analysis	Blast,Fasta,ClustalW, T-Coffee, Hmmer
Gene Finding	Glimmer
Phylogeny Analysis	Phylip,GRAPPA
Protein Structure Analysis	CE,Predator

Outline

- Background
- BioPerf Benchmark Suite
- Performance Study of BioPerf benchmarks
- Conclusions

Blast

- Basic Local Alignment Search Tool
- Developed by NCBI
- The most important bioinformatics application for its popularity

Input dataset for Blast

Blast	blastp blastn	The homo sapiens hereditary haemochromatosis protein Non-redundant protein sequence nr developed by NCBI
-------	----------------------	---

FASTA

- Also do a pairwise sequence alignment

FASTA	Fasta34 ssearch	The human LDL receptor precursor nr
-------	--------------------	---

ClustalW

- Multiple sequence alignment(MSA) program

ClustalW	Clustalw Clustalw_smp	317 Ureaplasma's gene sequences from NCBI Bacteria genomes database
----------	--------------------------	--

T-Coffee

- A sequential MSA similar to ClustalW with higher accuracy and complexity

T-coffee	Tcoffee	50 sequences of average length 850 extracted from the Prefab database
----------	---------	---

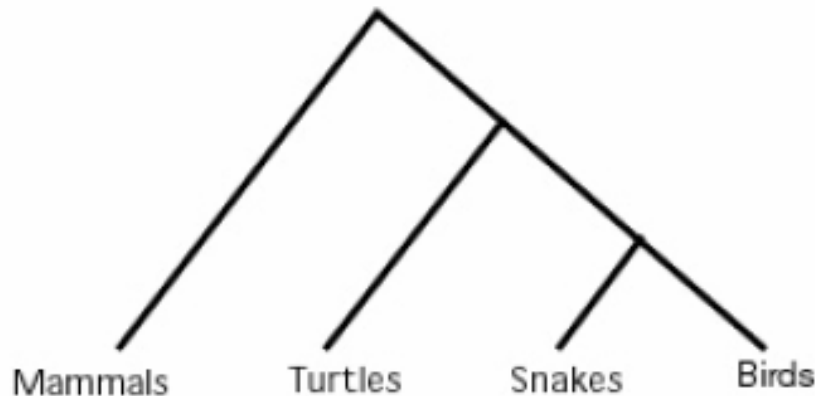
Hmmer

- Align multiple sequences by using hidden Markov models

Hmmer	hmmsearch hmmpfam	Brine shrimp globin HMM of 50 aligned globin sequences
-------	----------------------	--

Phylogenetic Reconstruction

- Study the evolution of all sequences and all species



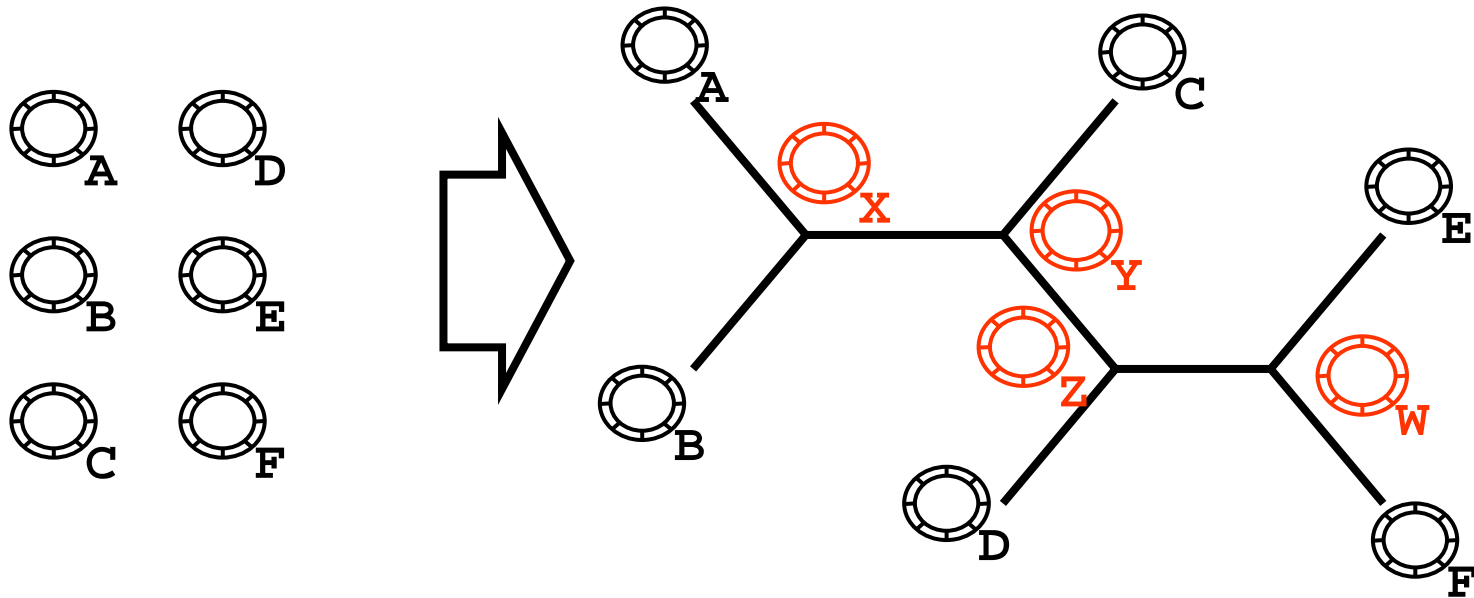
- Find the *best* among all possible trees.
- Given n taxa, number of possible trees $(2n-3)!!$
 - 10 taxa 2 million trees
- Approaches like maximum parsimony, maximum likelihood among others

Phylogeny: Phylip

- Collection of programs for inferring phylogenies
- Methods include
 - Maximum parsimony
 - Maximum likelihood
 - Distance based methods.
- **Input:** Aligned dataset of 92 cyclophilins proteins of eukaryotes each of length 220

Phylogeny: GRAPPA

- Gene order based phylogeny



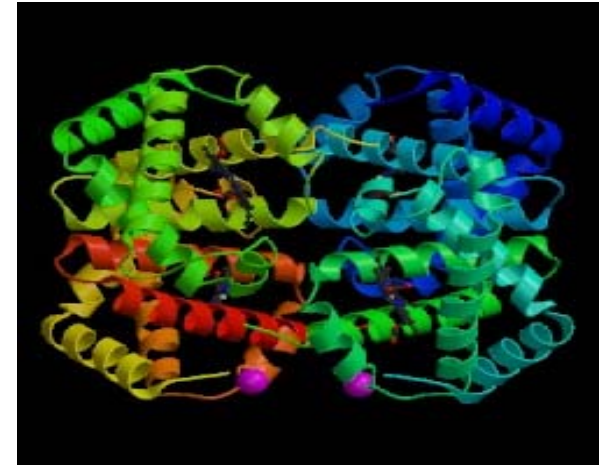
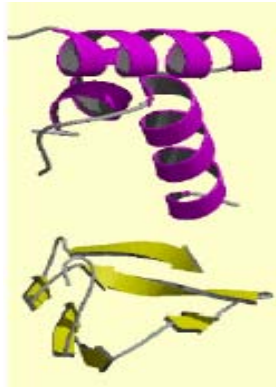
Input: 12 bluebell flower species of 105 genes

Protein Structure Prediction

5' atgcccaagctgaat ... 3'

atg ccc aag ctg aat ...

M P K L N ...



- Find the sequences, three dimensional structures and functions of all proteins and vice-versa
 - Why computationally?
 - Experimental Techniques slow and expensive
 - Problems with computational approach
 - Little understanding of how structure develops
 - Does function really follow structure ? Well

Protein Structure : Predator

- Tool for finding protein structures.
- Relies on local alignments from BLAST, FASTA.
- **Input:** 20 sequences from Swissprot each of length about 7000 residues.

CE(Combinatorial Extension)

- Find structural similarities between the primary structures of pairs of proteins.

CE	CE	Two different types of hemoglobin which is used to transport oxygen
----	----	---

Gene-Finding: Glimmer

- **Gene-Finding:** Find regions of genome which code for proteins.
- Widely used gene finding tool for microbial DNA.
- **Input:** Bacteria genome consisting of 9.2 million base pairs

Why BioPerf ?

- Previous attempts have been incomplete
 - Analysis on old architectures (*Biobench*)
- Description of input sets is incomplete
- Previous suites not available for download

BioPerf characteristics

- Freely redistributable Source codes.
- Pre-compiled binaries (PowerPC, x86, Alpha).
- Scalable Input datasets with each code for fair comparisons.
- Scripts for installation, running and collecting outputs
- Documentation for compiling and using the suite
- Parallel codes where available
- Available for download from www.bioperf.org

BioPerf: Applications Summary

Area	Package	Executables
Sequence homology Word-based Profile-based	BLAST HMMER	<i>blastp, blastn</i> <i>hmmpfam, hmmsearch</i>
Sequence Alignment Pairwise Multiple Multiple	FASTA CLUSTALW TCOFFEE	<i>ssearch, fasta</i> <i>clustalw, clustalw_smp</i> <i>tcoffee</i>
Phylogeny Parsimony/Likelihood Gene Rearrangement	PHYLIP GRAPPA	<i>dnapenny, promlk</i> <i>grappa</i>
Protein Structure Prediction Gene Finding Molecular Dynamics	PREDATOR GLIMMER CE	<i>predator</i> <i>glimmer, glimmer-package</i> <i>ce</i>

Alpha binaries & Simpoint (Li, Li)

- We have pre-compiled Alpha binaries for the majority of benchmarks for simulation.
- In order to reduce the simulation time, we collect the simulation points for those benchmarks by using SimPoint.

BioPerf performance (Bader, Sachdeva)

- Analysis at the instruction and memory level on PowerPC
- Livegraph data helps to visualize performance as it varies during a run
- Identify bottlenecks of current processors and make inputs for better performance on future processors
- Ongoing work using Mambo simulator (IBM PERCS)

Conclusions

- Bioinformatics is a rapidly evolving field of increasing importance
- BioPerf is a complete bioinformatics workload
- Allows people to analyze performance without dealing with complexities of bioinformatics

Thanks for attending the talk

- Questions ?