Petascale Distributed Data Analysis

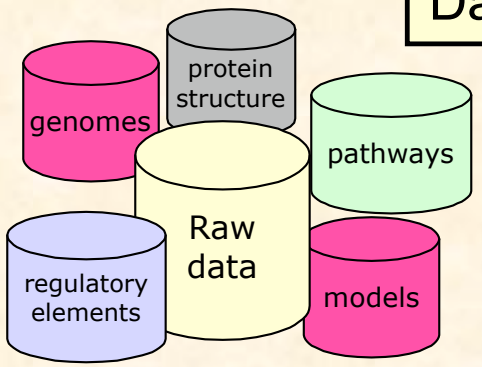
Important issues for mining massive biological data sets

Scalable

Popular methods do not scale in terms of time and storage

Distributed

Existing methods work on single centralized dataset. Data transfer is prohibitive



High-dimensional

Need new methods that scale up with the number of dimensions

Dynamic

Most methods work with static data - Changes lead to complete re-computation

Computational Feasibility on a Teraflop Computer

Biological Data Growth Trend:

Genome Assembly	300TB/genome
Protein Structure Prediction	PetaByte
Simulations of Bionetworks	1000s of PBs

Algorithmic Complexity:

Calculate means	O(n)
Calculate FFT	O(n log(n)
Clustering algorithms	$O(n^2)$

Data size, n	Algorithm Complexity			
	n	n*log*(n)	n^2	n^3
1MB	10^{-6} sec.	10^{-5} sec.	1 sec.	11 days
100MB	10^{-4} sec.	10^{-3} sec.	3 hrs	31 millenia
10GB	10^{-2} sec.	0.1 sec.	3 yrs.	10^{11}x age of the Universe

Bottom line: Bigger Computers aren't going to solve our problems
We need breakthroughs in modeling and simulation algorithms