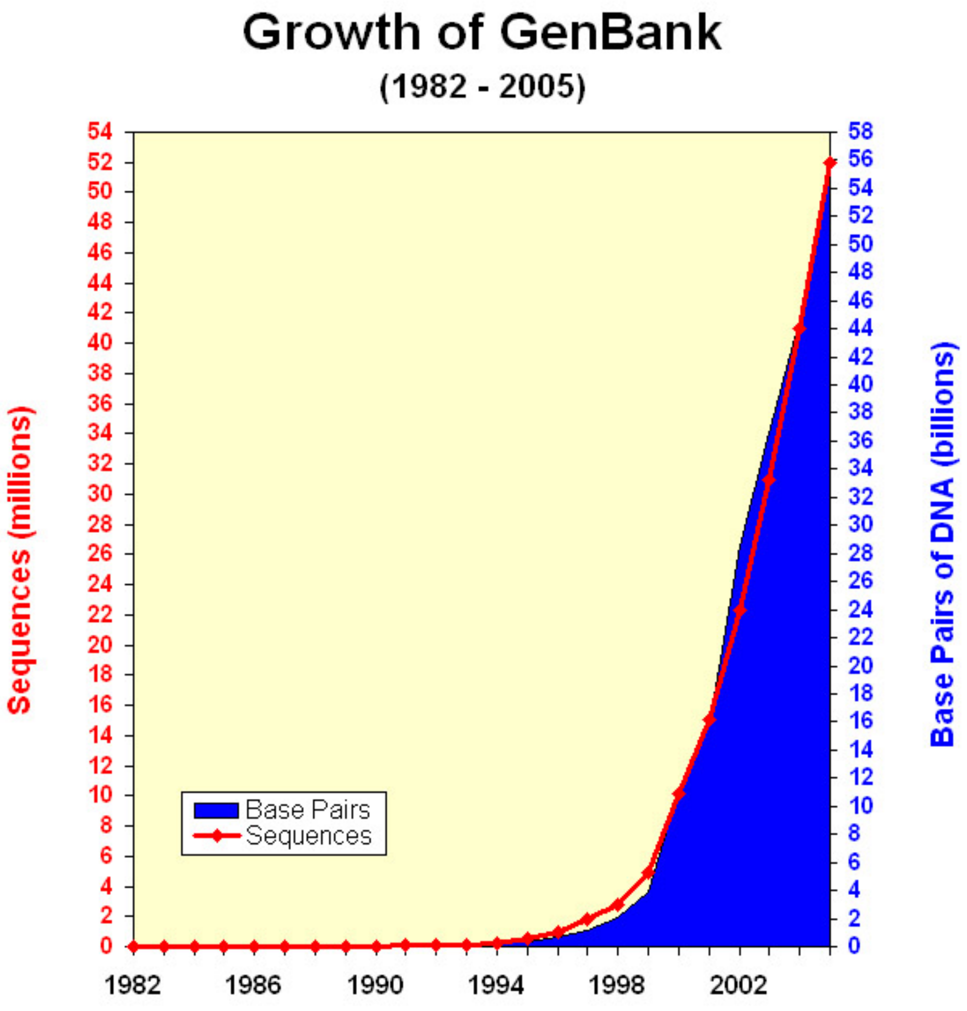
Fast-growing datasets meet slowly improving memory bandwidth and latency



Doubling time for sequence databases is currently ~18 months

According to Moore's Law, doubling time for processor speed is ~18 months.

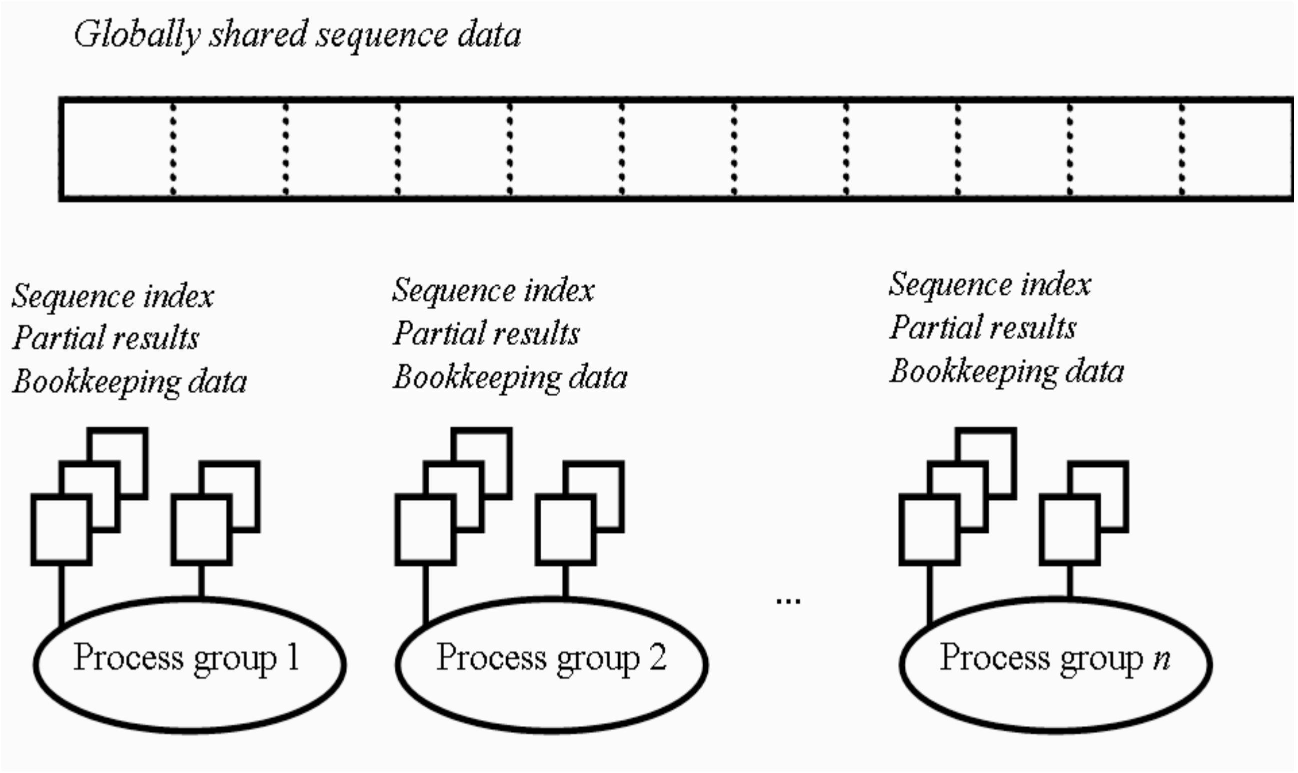
Time for doubling of bandwidth to memory and to disk = 2.7 years*

During this same time, memory *latency* only improves by 20%, and disk *latency* only improves by 30%*

*source: Patterson DA, "Latency Lags Bandwidth: Recognizing the chronic imbalance between bandwidth and latency, and how to cope with it", *Comm. ACM*. 47(10): **2004**, 71-75

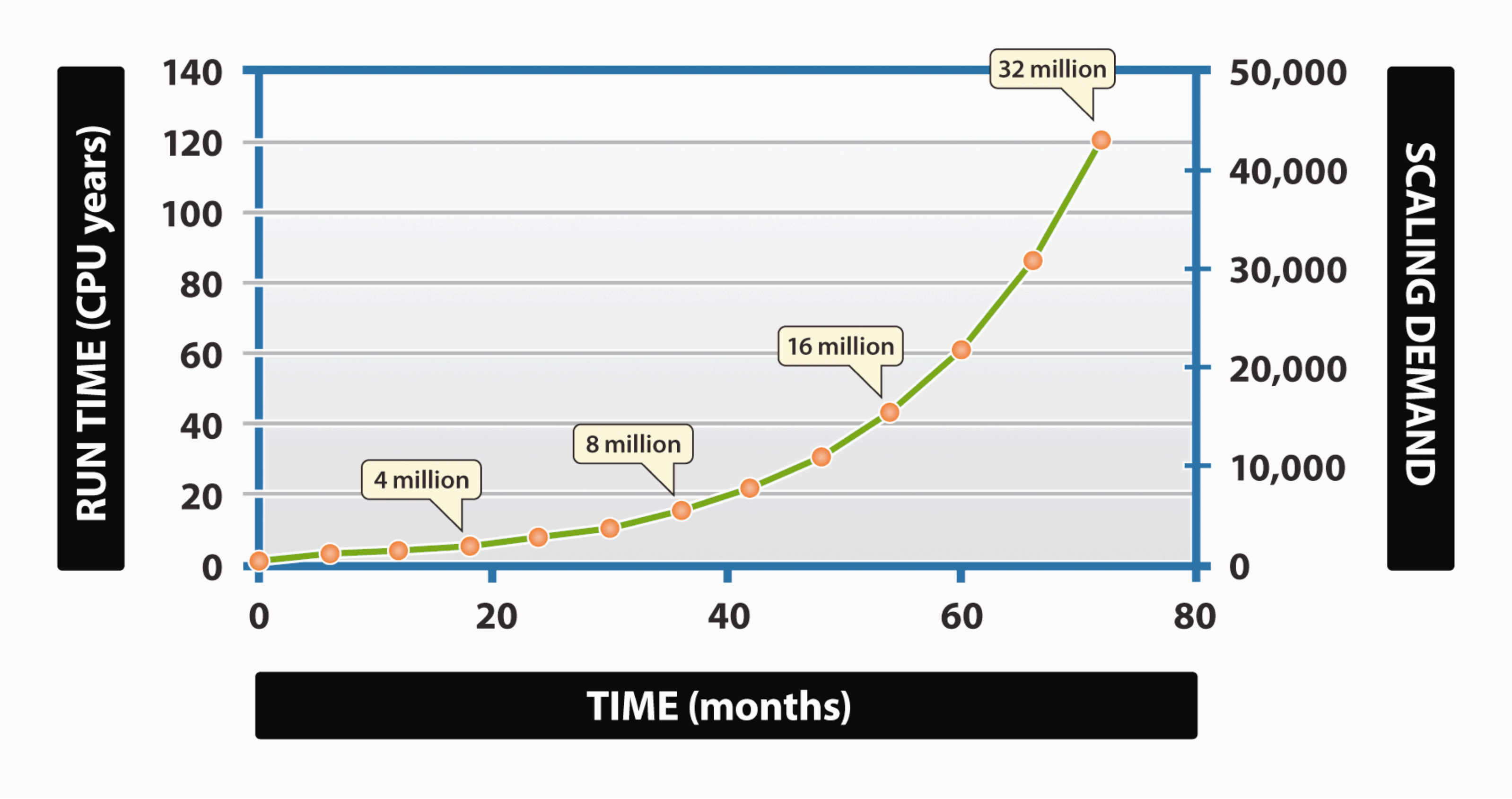
ScalaBLAST
Key: Memory Management

- Use large aggregate memory to share a single copy of the target database
- Hide latency by prefetching sequences in blocks.
- Each process group operates on independent query sets

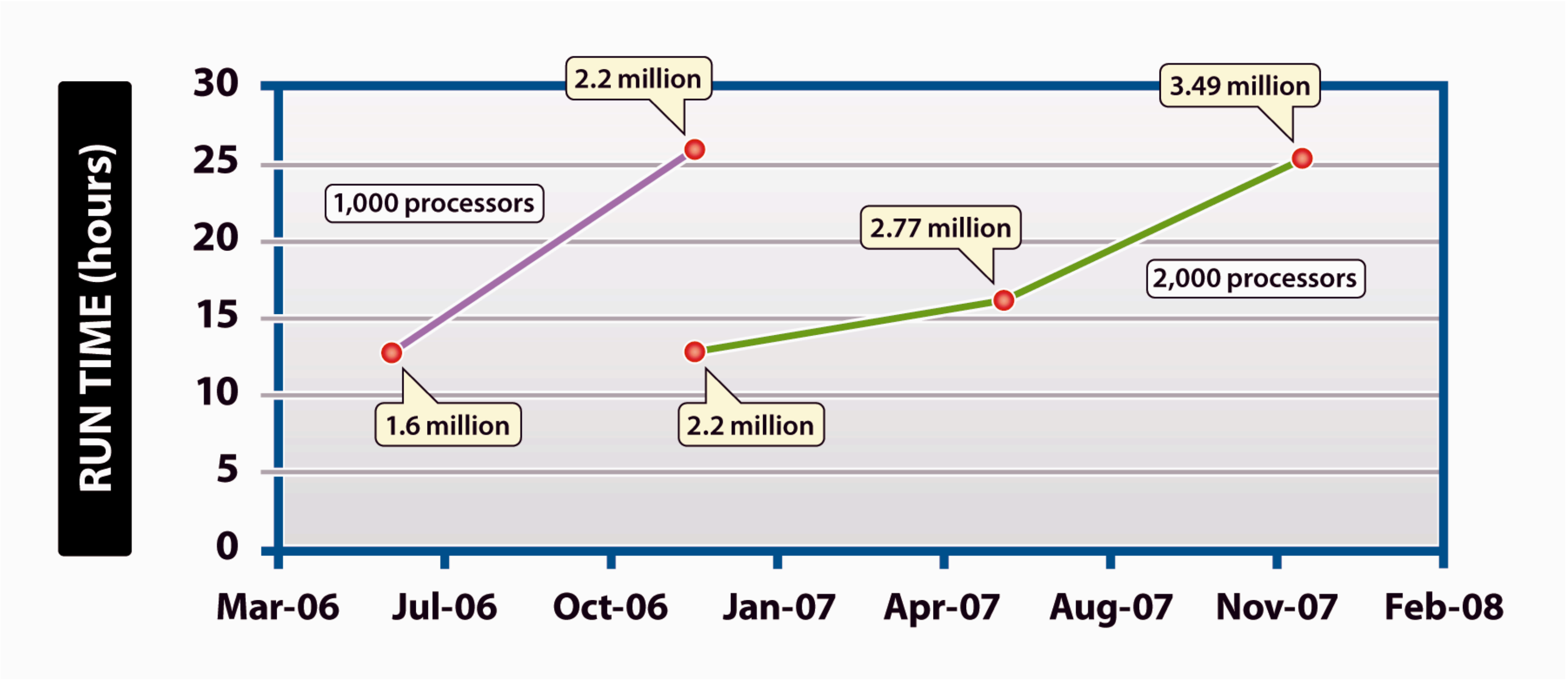


Demand for parallel BLAST

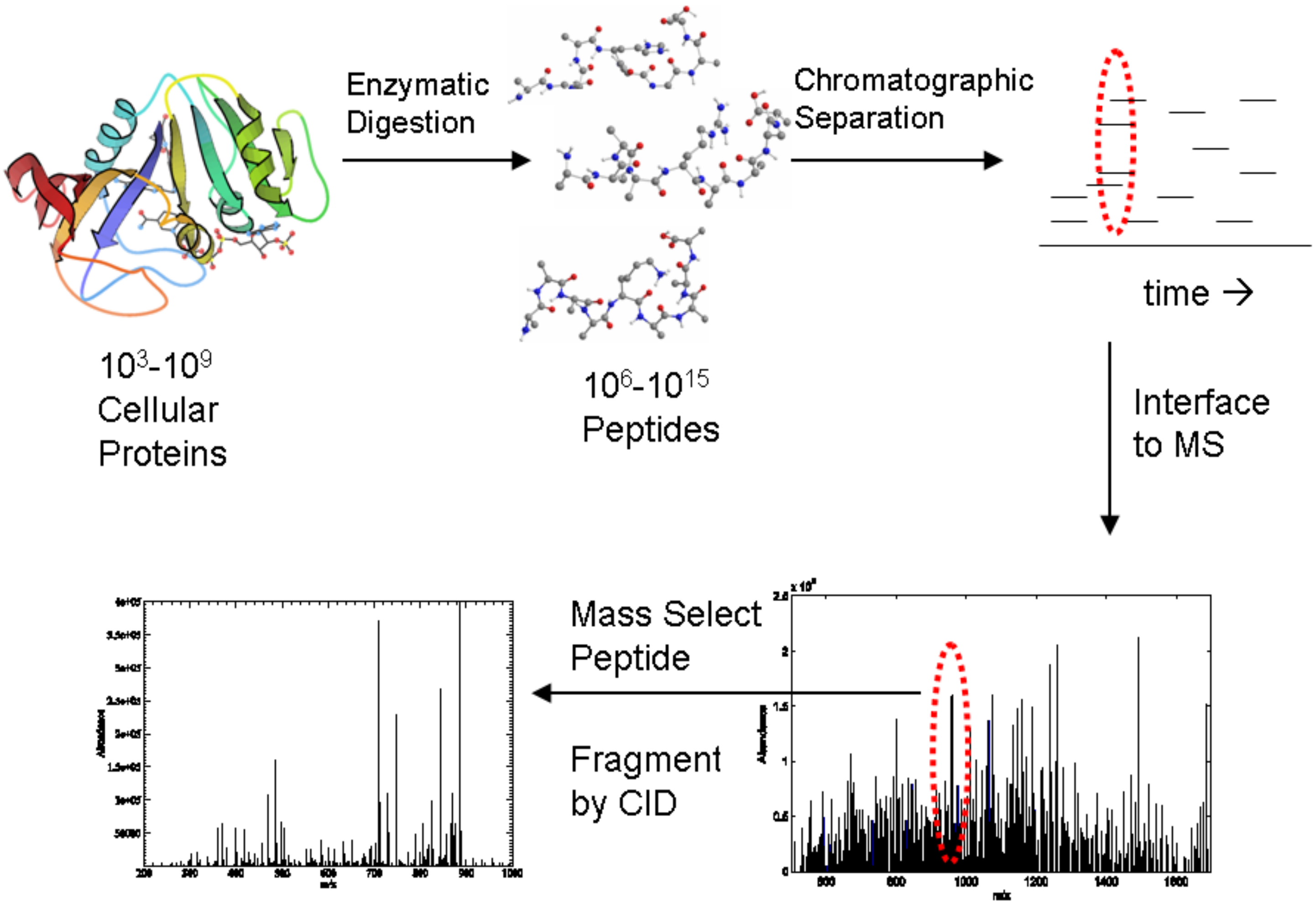
Computing time needed to perform an all vs. all calculation grows exponentially even though compute power increases with time. Scaling demand is calculated as the number of processors required to perform an all vs. all BLAST run within 24 hours at the expected memory bandwidth capacity available at the time of the run. ScalaBLAST scales to thousands of processors, but increased scaling demand will require running on tens of thousands of processors within 3 years. Callouts indicate anticipated database size over time.



Keeping pace with sequence data



Mass Spectrometry-based Proteomics



10^3-10^9 Cellular Proteins

Enzymatic Digestion

10^6-10^{15} Peptides

Chromatographic Separation

time →

Interface to MS

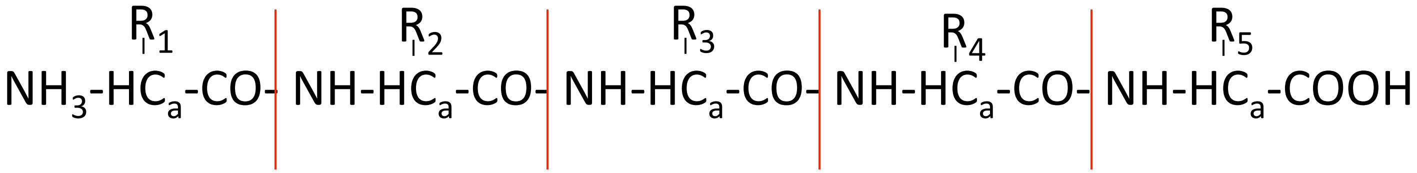
Mass Select Peptide

Fragment by CID

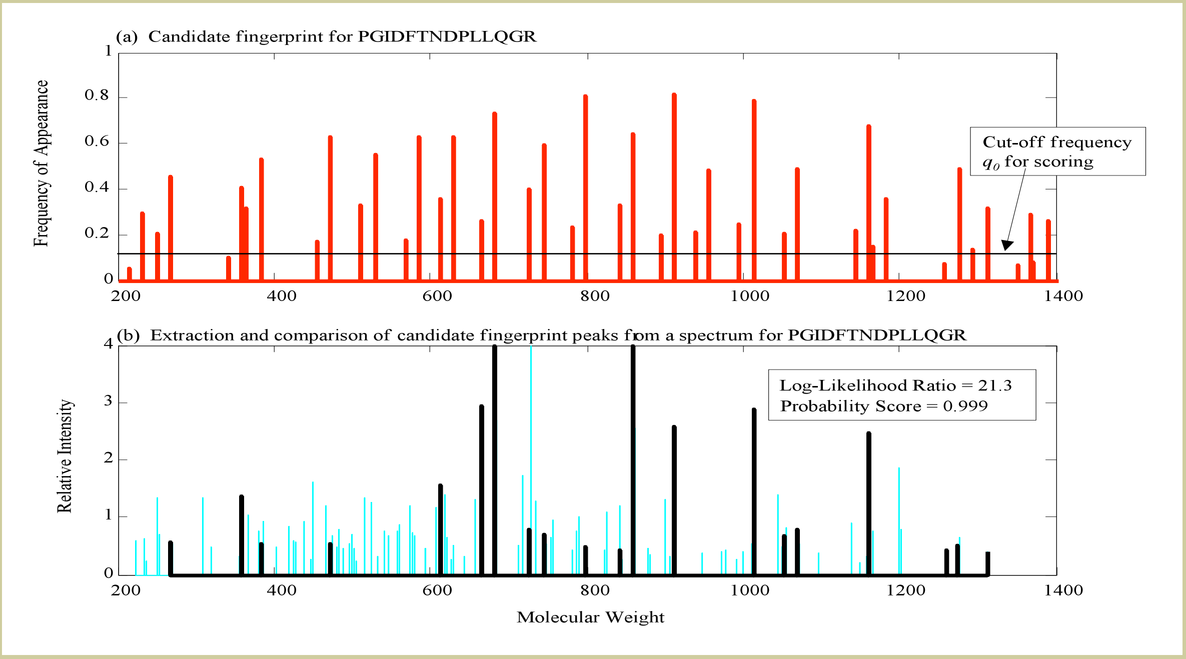
Terabytes of MS-proteomics data at EMSL

Comparing Models to Data

Generic model spectra don’t reflect the diversity of the data



Model spectrum


(a) Candidate fingerprint for PGIDFTNDPLLQGR

Cut-off frequency q_0 for scoring

(b) Extraction and comparison of candidate fingerprint peaks from a spectrum for PGIDFTNDPLLQGR

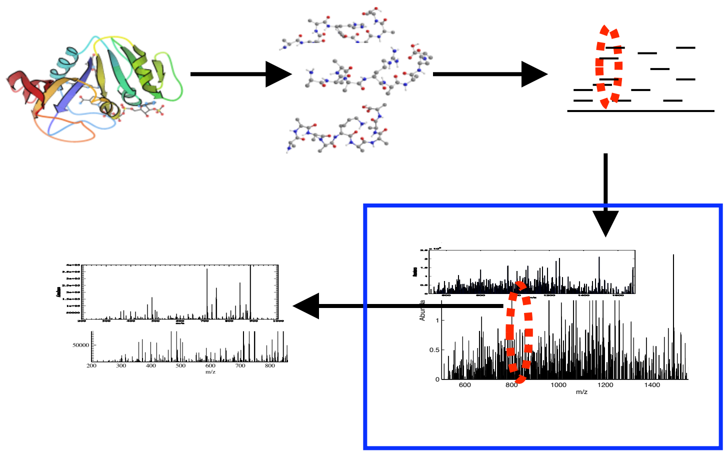
Log-Likelihood Ratio = 21.3
Probability Score = 0.999

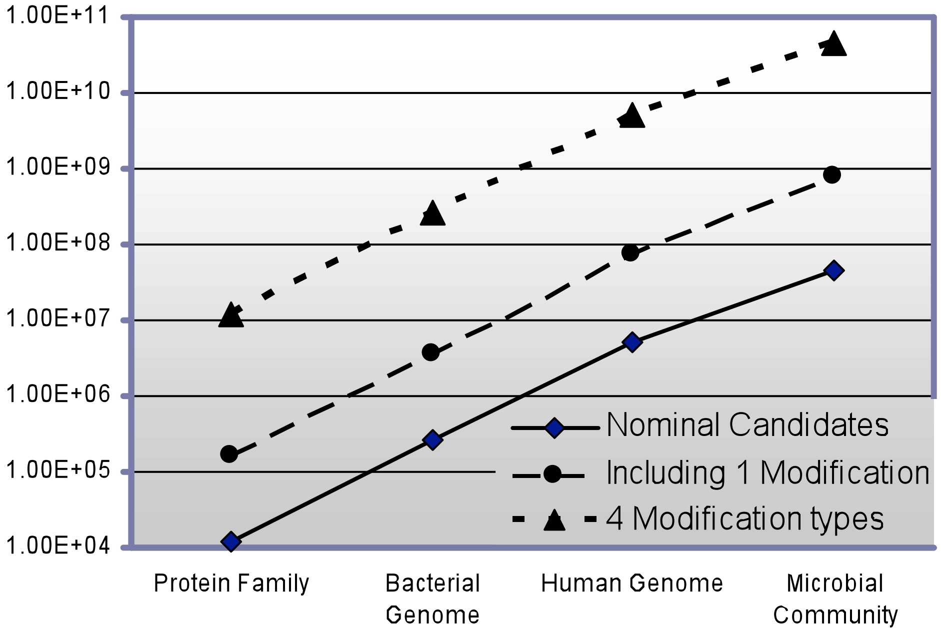
Actual spectrum- Highly variable

Models:

- **Statistical avg.**
- **Physical**
- **Experimental**

Peptide Candidates Per Spectrum





1. Not all peptides are candidate matches for each spectrum
2. Mass & chg selection

1 out of 10^5-10^{11} must be selected as the correct peptide