

Database Protein ID

- Sequest identifications
 - Uses the m/z ratio of the peptide before fragmentation (first MS step)
 - Uses MS/MS spectrum.
 - Protein database is digested *in silico*
 - Model MS/MS protein fragment spectra created (based on how peptides theoretically would fragment in the collision induced dissociation process)
 - Actual MS/MS spectrum is compared to model spectrum and a cross correlation score (XCorr) used as scoring metric. Matches from at least three to six peptides derived from the same protein are typically required to positively identify a protein
- Database search: very time consuming



Cross-Correlation

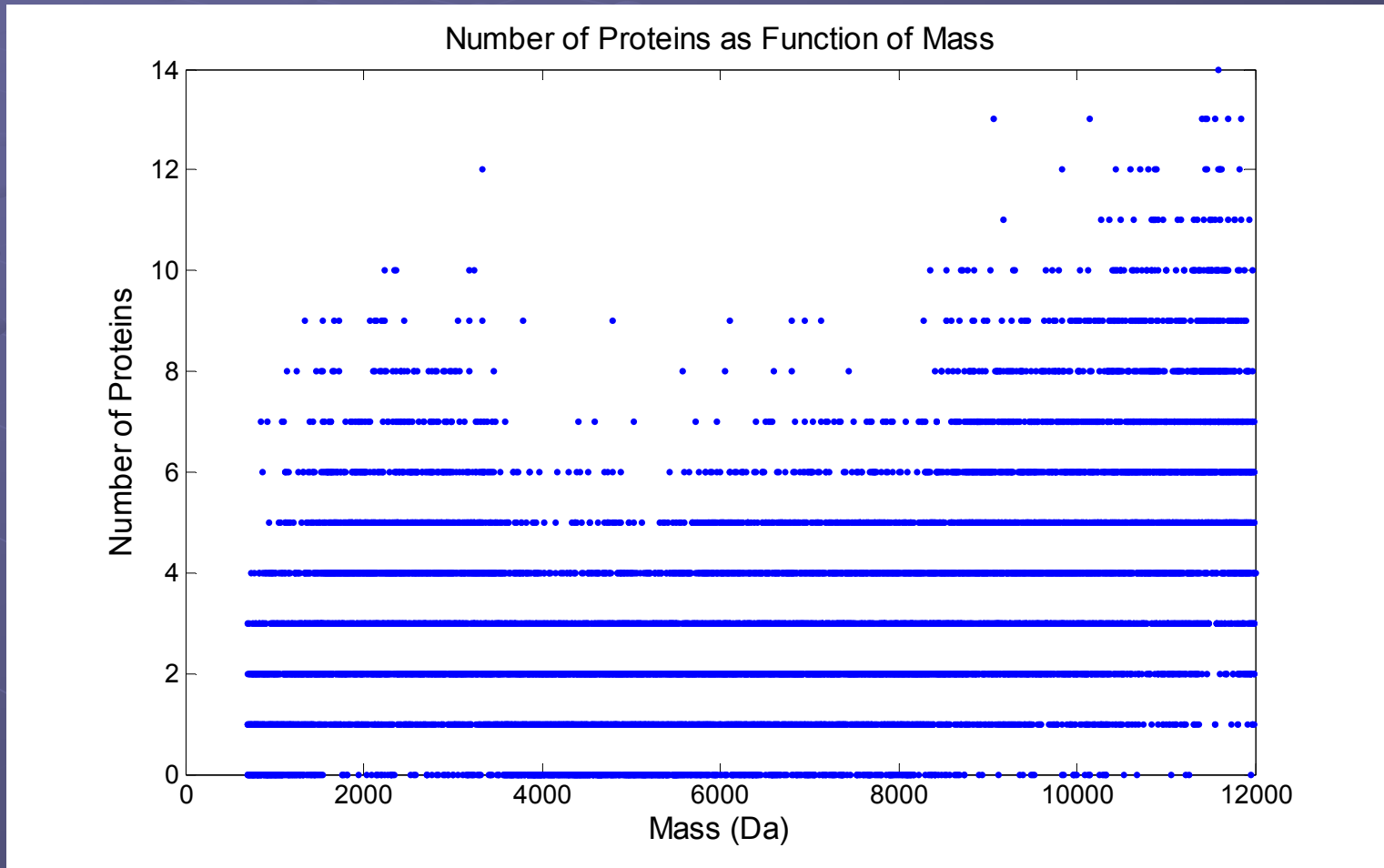
$$R_{xy}(m) = E\{x_{n+m}y_n^*\} = E\{x_n y_{n-m}^*\}$$

where x_n and y_n are jointly stationary random processes, $-\infty < n < \infty$

$$\hat{R}_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m}y_n^* & m \geq 0 \\ \hat{R}_{yx}^*(-m) & m < 0 \end{cases}$$



The number of human proteins per mass unit for various masses is rarely unique



Alterovitz, G., Afkhami, E. & Ramoni, M. in *Focus on Robotics and Intelligent Systems Research*, ed. Columbus, F. Nova Science Publishers, Inc., New York, 2005 (In press).

Poisson

- Often, the Poisson distribution is used to model situations involving counts or arrivals during an interval of time.
- In this case, at each mass unit, an average number of proteins are expected to 'arrive' during this interval.
- However, the Poisson distribution has only one parameter (λ) which is equal to the mean and variance. Yet, since the variance is 4.82 (compared to 3.19 for mean), the Poisson model is not a good fit in this case

$$f(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

Negative Binomial

- The negative binomial distribution can be used (parameters p and r)
- Modeling daily road accidents at certain highway locations.
 - As there can be high variances in this scenario (daily accidents dependent on the day's weather conditions, etc), negative binomial models have been used instead of Poisson in such cases.

The negative binomial distribution is commonly defined as:

$$f(x|r, p) = \binom{r+x-1}{x} p^r (1-p)^x$$

Negative binomial distribution for $r \in \mathbb{Z}^+$

However, when parameter r is not restricted in integer values, the more general expression becomes:

$$f(x|r, p) = \frac{\Gamma(r+x)}{\Gamma(r)\Gamma(x+1)} p^r (1-p)^x$$

The gamma function from above is defined as:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$



Gamma Distribution

- The gamma distribution (not to be interchanged with the gamma function) is useful in measuring failure times
- Superset of the exponential distribution (it allows for an additional dependence on the 'age' of the item).
- In this case, the failure times would be the count of proteins that are confined to a certain mass before the next mass window starts.



Gamma Distribution

$$f(x) = \frac{\left(\frac{x-\mu}{\beta}\right)^{\gamma-1} \exp\left(-\frac{x-\mu}{\beta}\right)}{\beta\Gamma(\gamma)} \quad x \geq \mu; \gamma, \beta > 0$$

where γ is the shape parameter, μ is the location parameter,
 β is the scale parameter

Where the gamma function from above is defined as:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

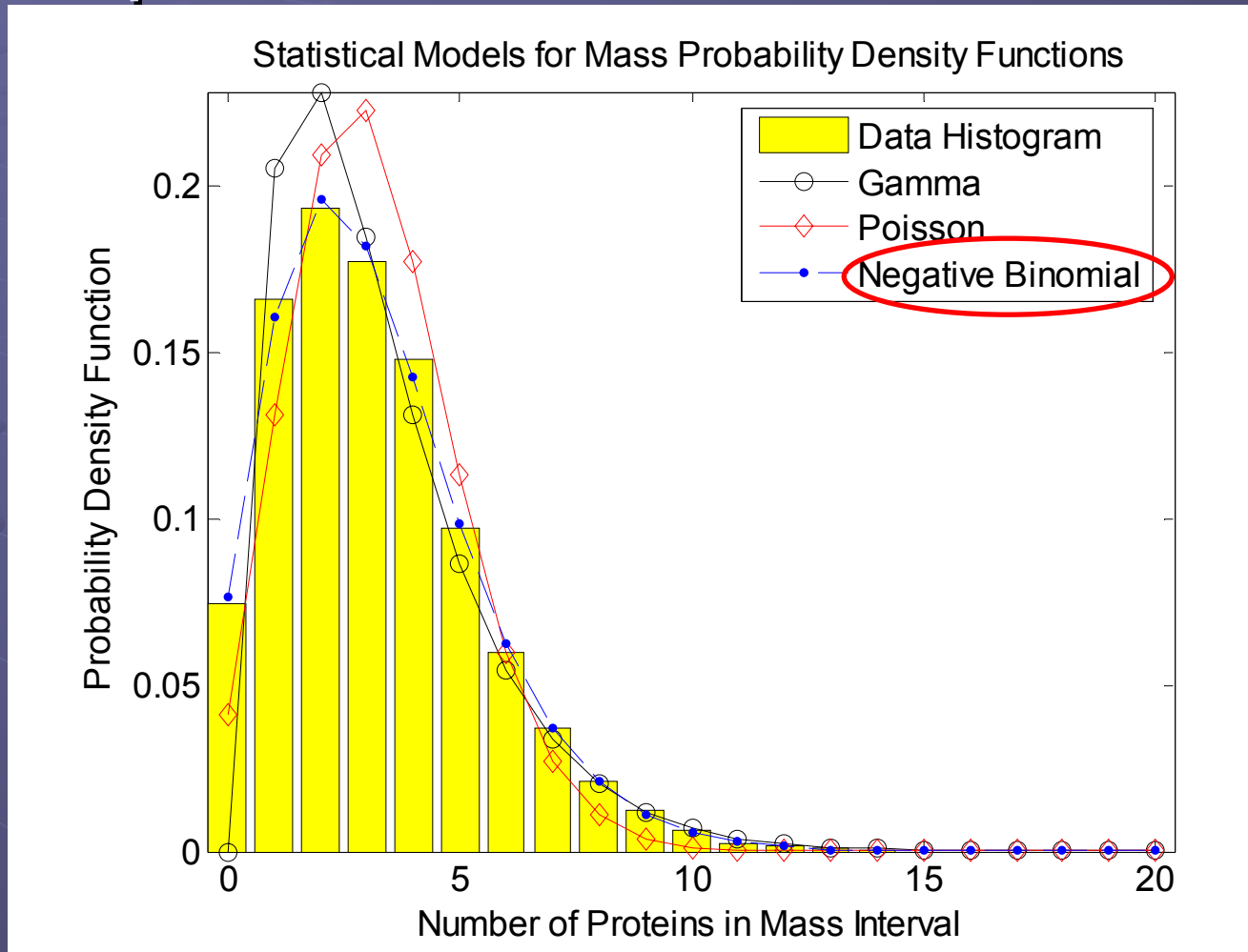


Best Fit

- The negative binomial model (with MLE estimates of $r=6.19$ and $p=0.660$) was the best fit by several measures.
- The negative binomial distribution had the best log likelihood score at: -2.39×10^4 .



The probability density function of protein counts/mass unit



HST's CHIP

chip
Children's Hospital Informatics Program

■ Home ■ Contact ■ Sitemap ■ Intranet

Research People Funding Training Resources

Research The research at the Children's Hospital Informatics Program spans a wide range of problems in bioinformatics and clinical informatics. Our goal is to make significant contributions to biomedical research and patient care by understanding and utilizing various types of genomic and proteomic data and by developing innovative hardware and software technologies.

[Research Area Listing by Members](#)

Bioinformatics

Our research interests include development of statistical and computational techniques for analyzing gene expression data from microarrays under various experimental settings and for analyzing SNP (Single Nucleotide Polymorphism) data in performing large-scale association studies. We also develop tools to integrate various databases effectively, and we combine information from various genomic and proteomic data to gain insights into biological pathways.

[Bioinformatics Projects](#)

Courtesy of Children's Hospital Informatics Program. Used with permission.



Harvard-MIT
Division of Health
Science & Technology

Proteomic Engineering

Automation, Engineering, and Science for Clinical Applications

Home

Research

Publications

Downloads

Contact

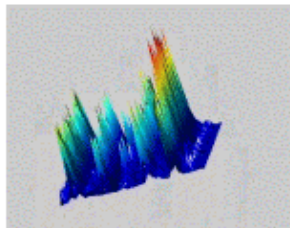
Topic Menu

Home

Computation

Automation/Robot

Mission

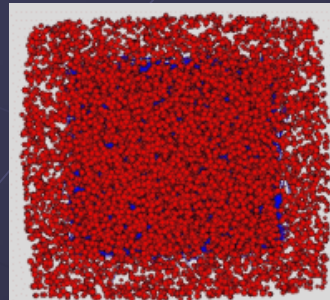


Proteomic Engineering seeks to combine engineering principles with the medical sciences to devise new technologies for exploration of clinically relevant proteins. Through use of automation, computer analysis, and interdisciplinary collaborations, new approaches are being developed.

News Links

Spring 2004 - Analysis and robot pipelined automation [protocol](#) now online.

Copyright (c) 2004. All Rights Reserved.



Bioinformatics and Proteomics Databases

Gil Alterovitz




























Sequence Databases

- Entrez: MEDLINE (articles), genetic, protein sequences
 - Matlab: [getgenbank](#) Retrieve sequence information from GenBank database
 - Matlab: [getgenpept](#) Retrieve sequence information from GenPept database
- SwissProt- Proteins
- EMBL- Nucleotides, other
 - Matlab: [getembl](#) Retrieve sequence information from EMBL database
- TrEMBL- Proteins translated from RNA



- Reference
- Sequence
- Meta-Info
- Classifications/Vocabularies

Entrez

 PubMed: biomedical literature citations and abstracts ?	 Books: online books ?
 PubMed Central: free, full text journal articles ?	 OMIM: online Mendelian Inheritance in Man ?
	 Site Search: NCBI web and FTP sites ?
 Nucleotide: sequence database (GenBank) ?	 UniGene: gene-oriented clusters of transcript sequences ?
 Protein: sequence database ?	 CDD: conserved protein domain database ?
 Genome: whole genome sequences ?	 3D Domains: domains from Entrez Structure ?
 Structure: three-dimensional macromolecular structures ?	 UniSTS: markers and mapping data ?
 Taxonomy: organisms in GenBank ?	 PopSet: population study data sets ?
 SNP: single nucleotide polymorphism ?	 GEO Profiles: expression and molecular abundance profiles ?
 Gene: gene-centered information ?	 GEO DataSets: experimental sets of GEO data ?
 HomoloGene: eukaryotic homology groups ?	 Cancer Chromosomes: cytogenetic databases ?
 PubChem Compound: small molecule chemical structures ?	 PubChem BioAssay: bioactivity screens of chemical substances ?
 PubChem Substance: chemical substances screened for bioactivity ?	 GENSAT: gene expression atlas of mouse central nervous system ?
 Journals: detailed information <i>about</i> the journals indexed in PubMed and other Entrez databases ?	 MeSH: detailed information about NLM's controlled vocabulary ?
 NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections ?	

BLAST

- BLAST- Basic Local Alignment Search Tool.
 - It allows for searching nucleotide and protein databases for similar sequences. It uses alignment algorithm and mapping of similarity regions to find matches.
 - Matlab: [getblast](#) Get BLAST report from NCBI web site



Image courtesy of Ben Fry. Used with permission



Harvard-MIT
Division of Health
Science & Technology

Protein Structure

- Protein Data Bank- a database of biological 3-D structure information.
 - Matlab: [getpdb](#) Retrieve protein structure data from PDB database
 - Matlab: [getpir](#) Retrieve sequence data from PIR-PSD database



Microarray Database

- **Gene Expression Omnibus**- a database of gene expression (e.g. from microarray experiments) and molecular abundance.
 - Matlab: [getgeodata](#) Get Gene Expression Omnibus (GEO) data
 - Matlab: [gprread](#) Read microarray data from a GenePix Results (GPR) file
 - Matlab: [imageneread](#) Read microarray data from an ImaGene Results file



Protein Interactions

- [BIND - Biomolecular Interaction Network Database](#) (Chris Hogue/Gary Bader)- free.
- [DIP - Database of Interacting Proteins](#) – free.
- [cPath](#) (Gary Bader)- cancer related, free.
- Millenium's PARIS- proprietary.
- [Ingenuity Systems](#)- subscription-based



Mass Spec Databases

- Open Proteomics Database- MS/MS
- Peptide Atlas Repository- MS/MS
- NCI Center for Cancer Research- SELDI



Integration Databases

- GO (Gene Ontology)- Categorizes genes by function and cell compartment location. AmiGO can be used view this data.
- SeqHound- to translate between different accession numbers (like social security numbers of a genetic or protein sequence)



Other Matlab functions

- [Gethmmalignment](#) Retrieve multiple aligned sequences from the PFAM database
- [gethmmprof](#) Retrieve profile hidden Markov models from the PFAM database
- [gethmmtree](#) Get phylogenetic tree data from PFAM database
- [fastawrite](#) Write to file with FASTA format

- Plus read versions of Matlab functions to load local files.

