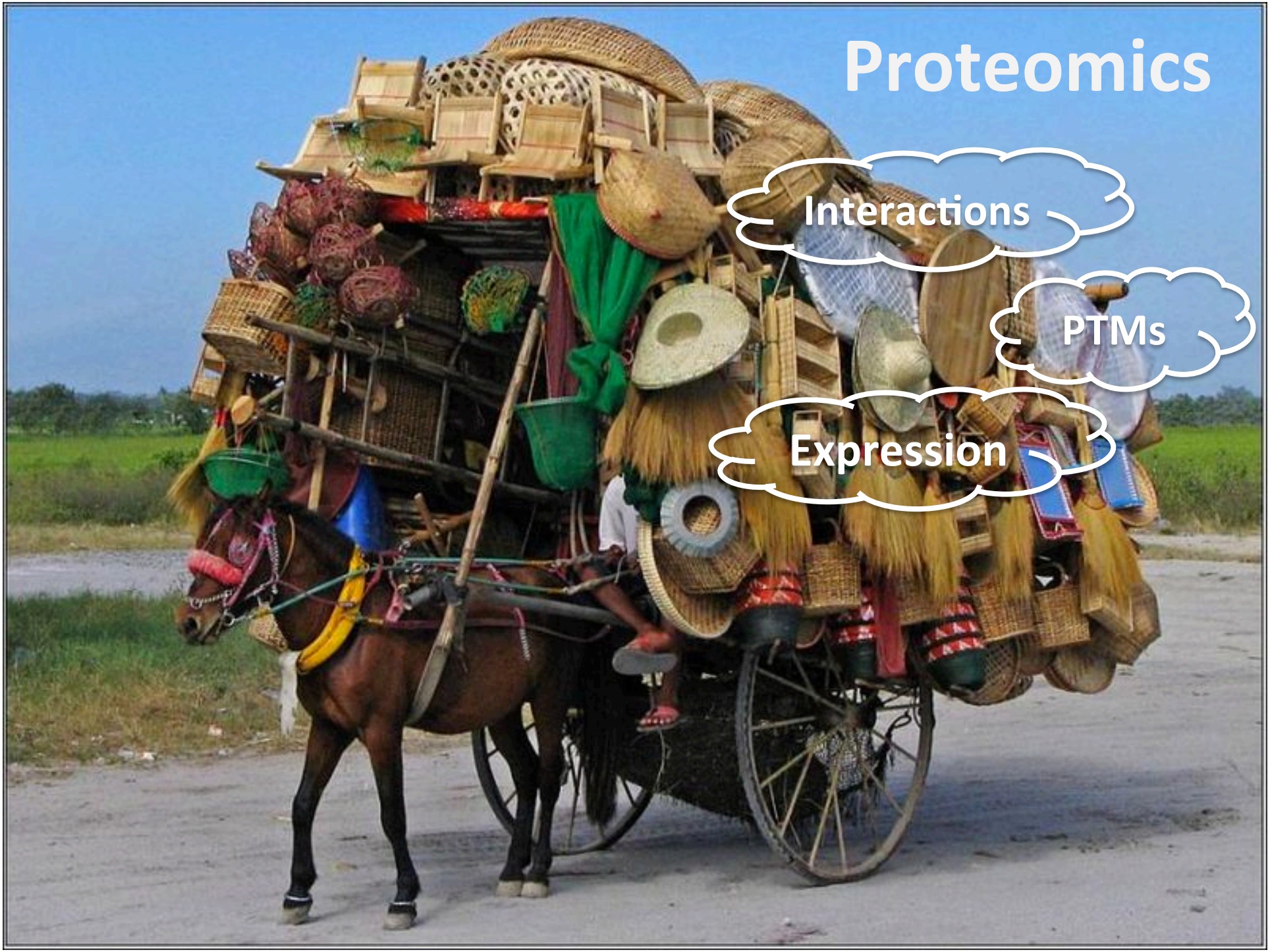
Efficient Processing of Models for Large-scale Shotgun Proteomics Data

Himanshu Grover, Ph.D.
Vanathi Gopalakrishnan, Ph.D.
University of Pittsburgh

C-Big 2012, Pittsburgh, USA
14th October, 2012

Outline

- Background on Proteins and Shotgun Proteomics
- Computational modeling framework:
  - Context-sensitive Peptide Identification (**CSPI**)
- Problem Statement
- Methods for efficient handling
- Challenges and Future Work


Proteomics
Interactions
PTMs
Expression

Proteomics

Mass Spectrometry

Analytical tool to identify unknown compounds

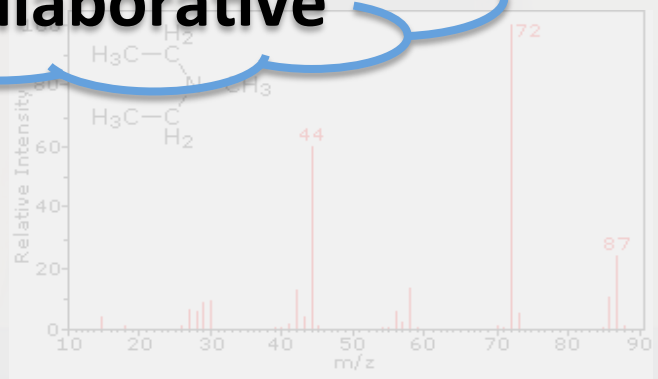
Complex

Sample

↓

Ionization → Mass Analyzer → Detector

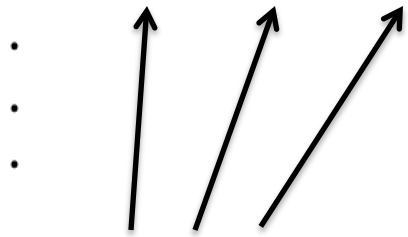
Collaborative



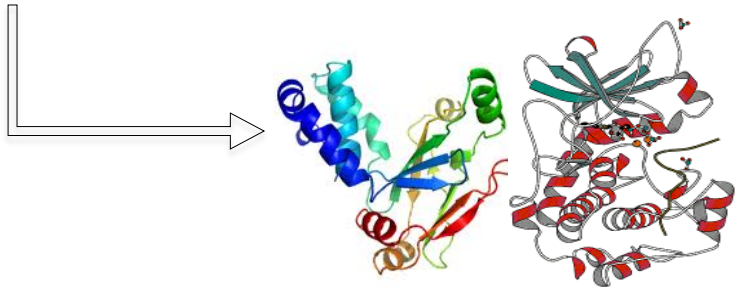
Amino Acids and Proteins

>IPI:IPI00000005.1 Tax_Id=9606 Gene_Symbol=NRAS GTPase NRas
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
QEEYSAMRDQYMRTGEGFLCVFAINNSKSFADINLYREQIKRVKDSDDVPMVLVGNKCDL
PTRTVDTKQAHELAKSYGIPFIETSAKTRQGVEDAFYTLVREIRQYRMKKLNSSDDGTQG
CMGLPCVVM
>IPI:IPI00000115.1 Tax_Id=9606 Gene_Symbol=CNIH4 Isoform 1 of Protein cornichon homolog 4
MEAVVFVFSLLDCCALIFLSVYFIITLSDLECDYINARSCCSKLNKWVIPELIGHTIVTV
LLLMSLHWFIFLLNLPVATWNIYRYIMVPSGNMGVFDPTEIHNRGQLKSHMKEAMIKLGF
HLLCFFMYLYSMILALIND

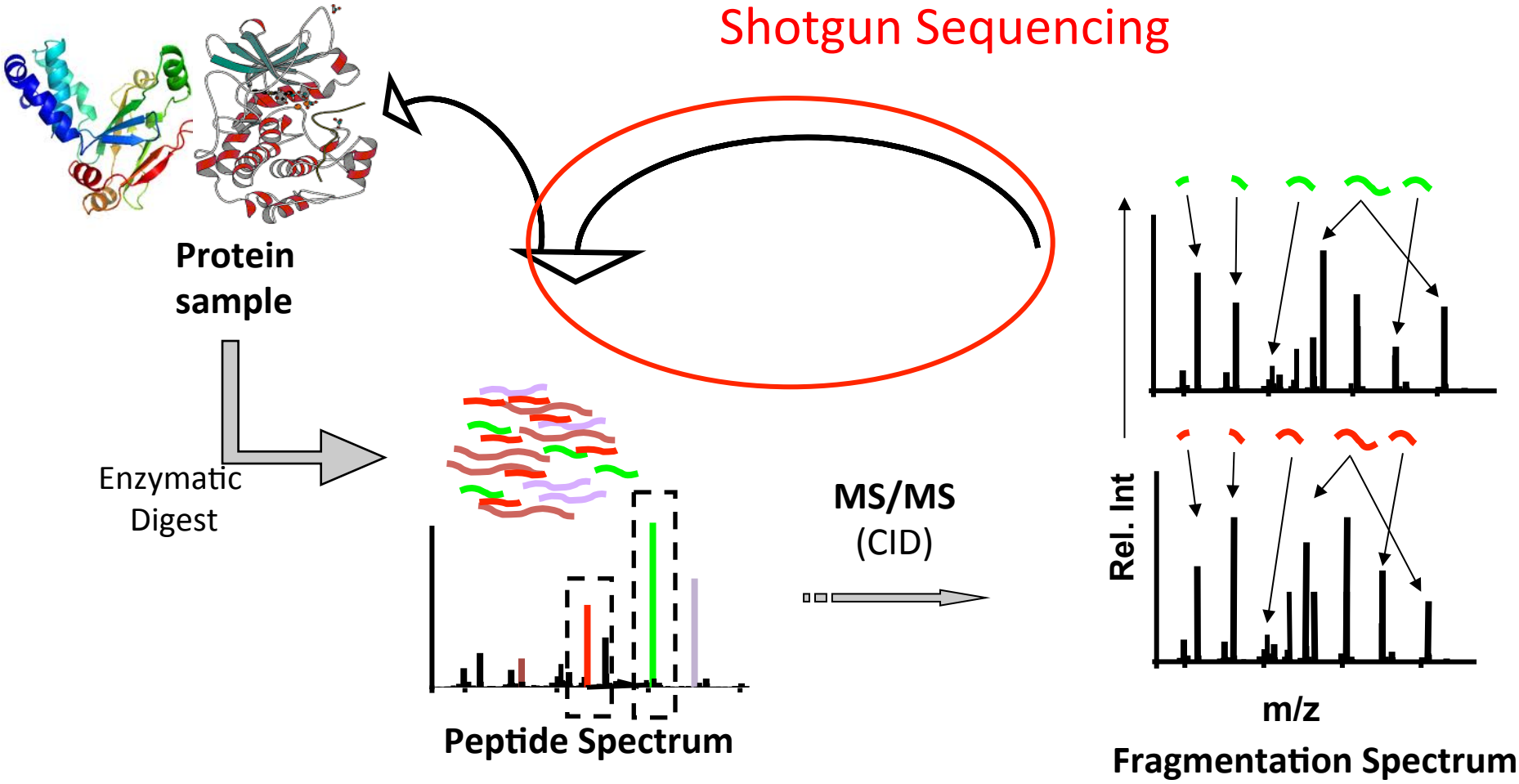
.
.
.



Amino Acids

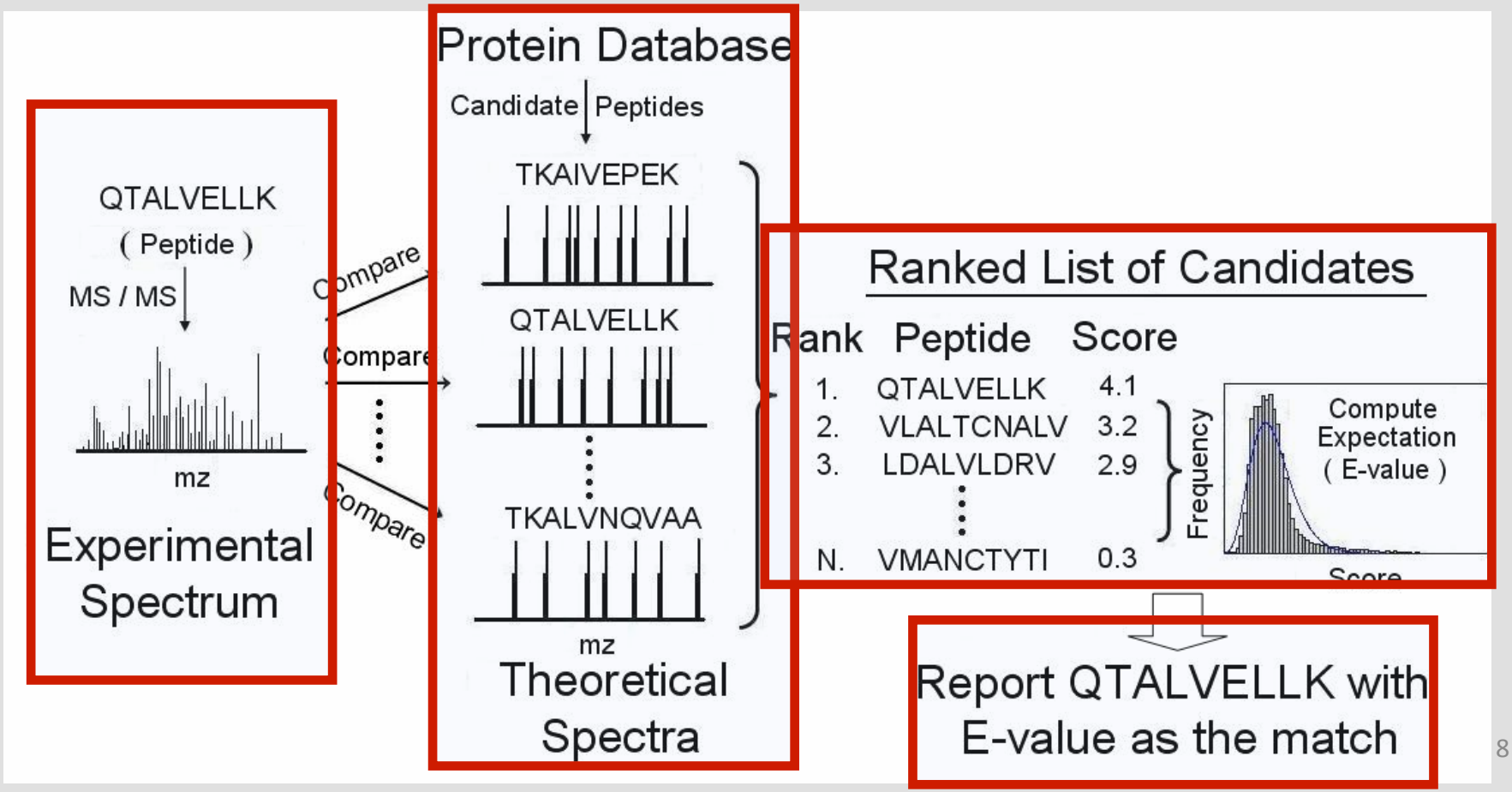


Shotgun Proteomics: Protein/Peptide Identification



Database Searching

Predominant methodology for peptide ID from MS/MS



Fact !!

< 30% of spectra are confidently assigned with peptides

- Noise
- Variability
- Inadequate scoring systems

Computational Bottlenecks

- **High volume and rate of data generation**
  - 24*7
  - 200 – 400 ^ 3 spectra per day from moderate sized labs

- **Large protein databases:** ~90 K protein sequences for Humans
  - Constrained searches:
    - ~5-10 ^ 6 unique peptides in database
    - ~10-20 ^ 3 peptides per spectrum
  - Unconstrained searches
    - Over billion peptides

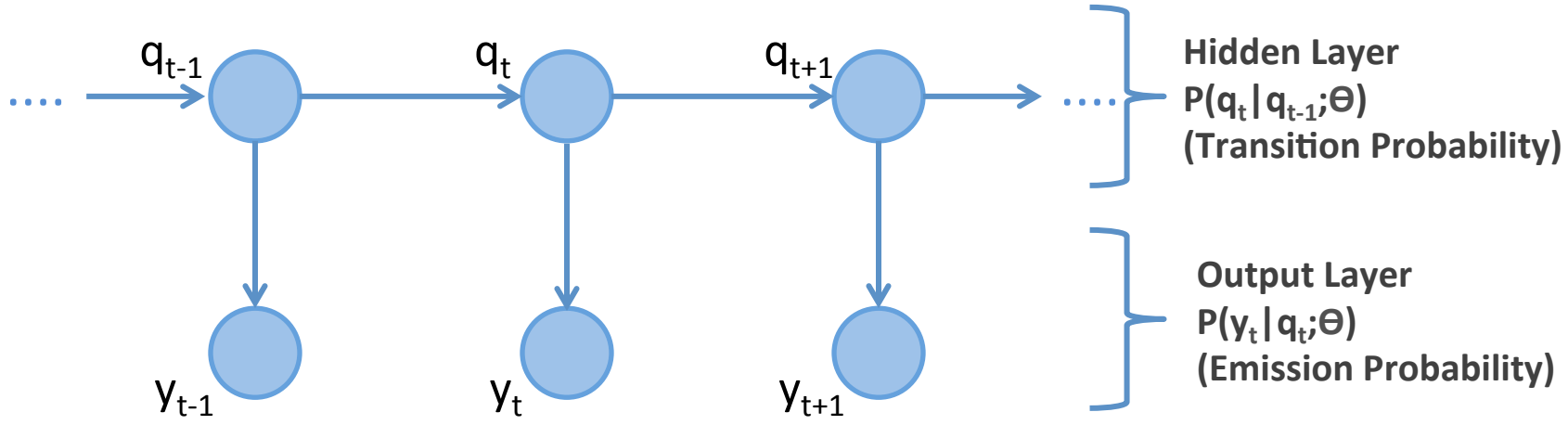
Context-**S**ensitive **P**eptide **I**dentification (**CSPI**) Framework Demystified

Grover et. al. (2012), OMICS (*submitted for publication*)

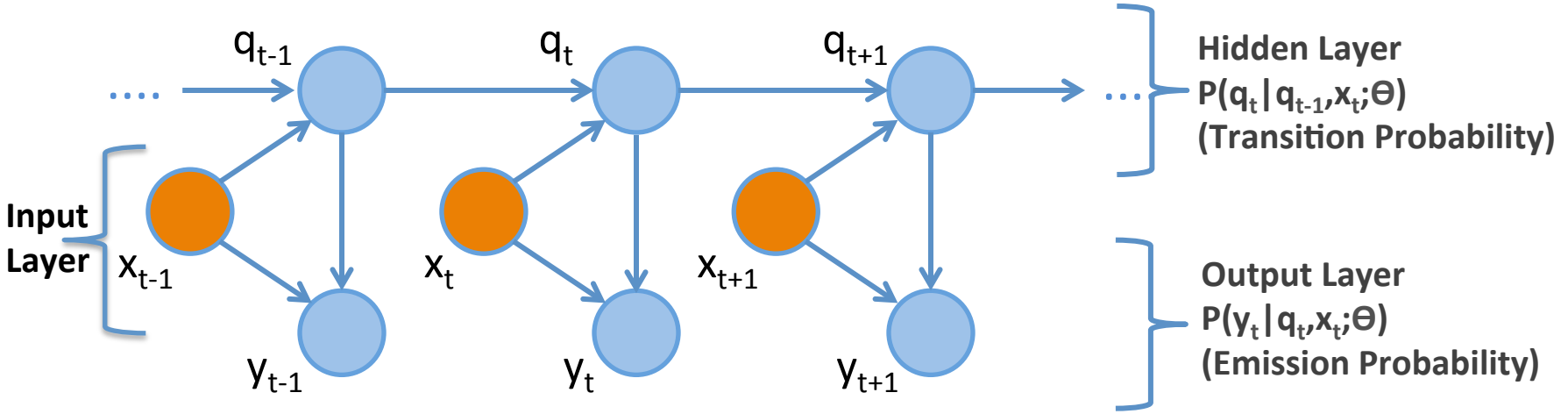
- Novel probabilistic framework
  - Scalable and flexible

- **Specific Goal:** Model influence of peptide physicochemical **context** on the observed peak heights (intensities) in fragmentation spectra

Input-Output Hidden Markov Models (IO-HMM)

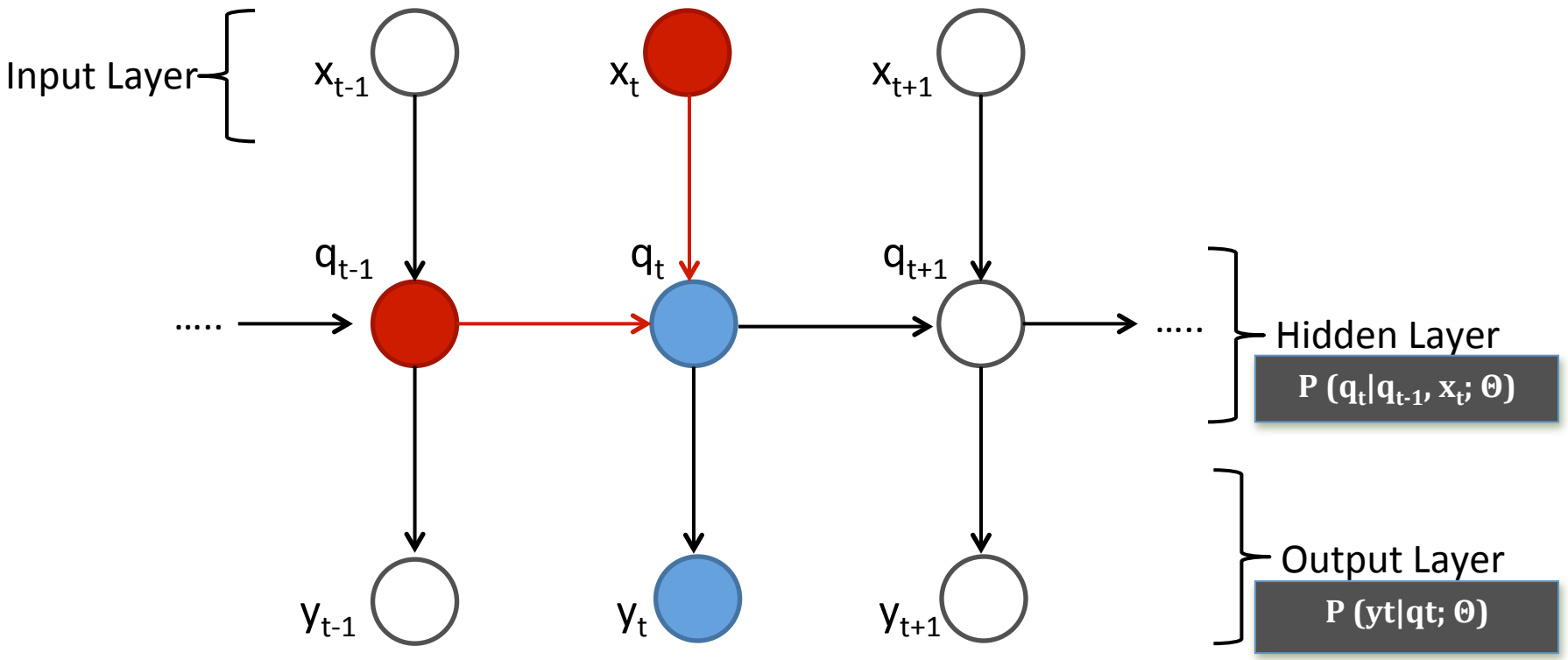


Classical Hidden Markov Model

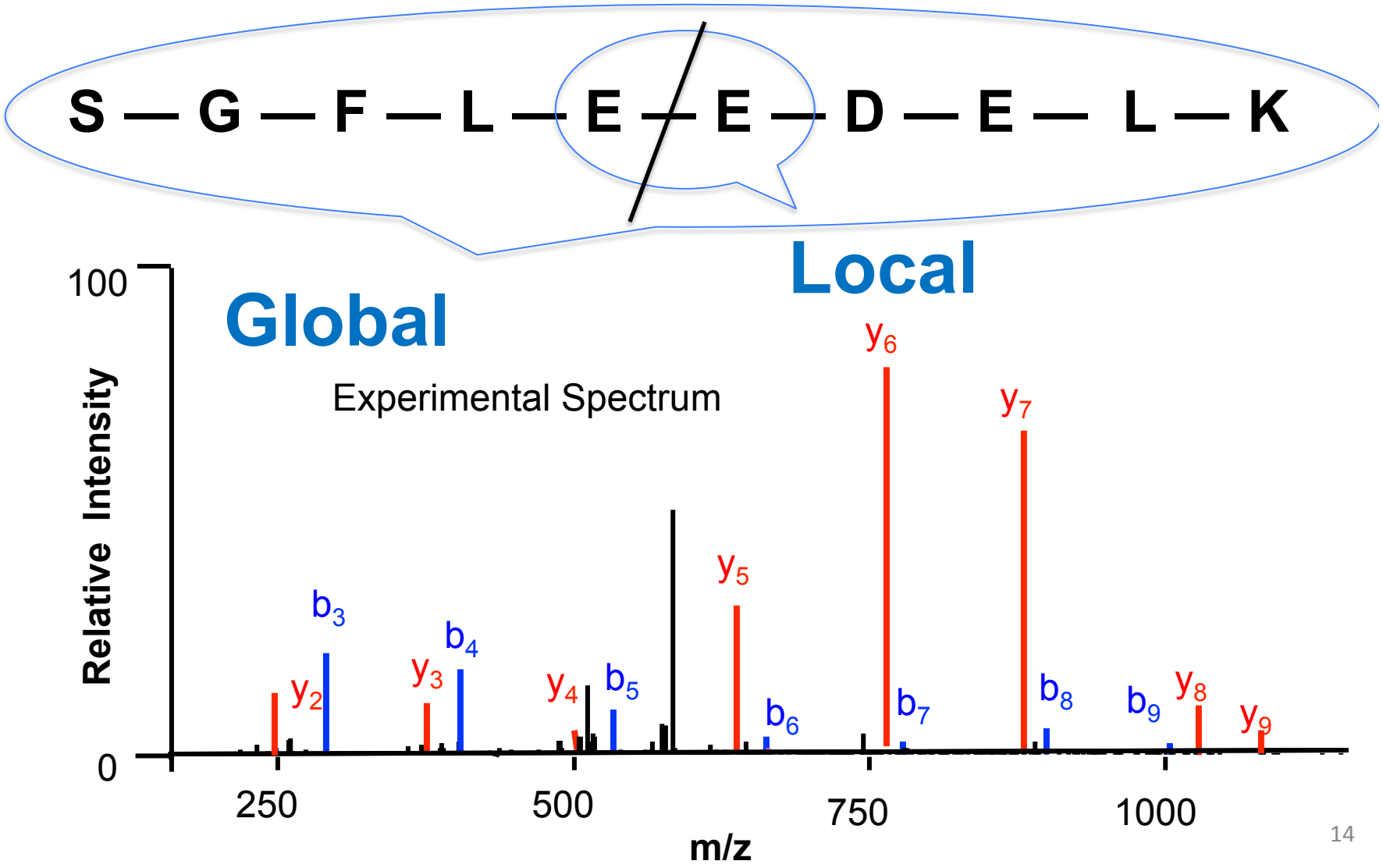


Input-output Hidden Markov Model

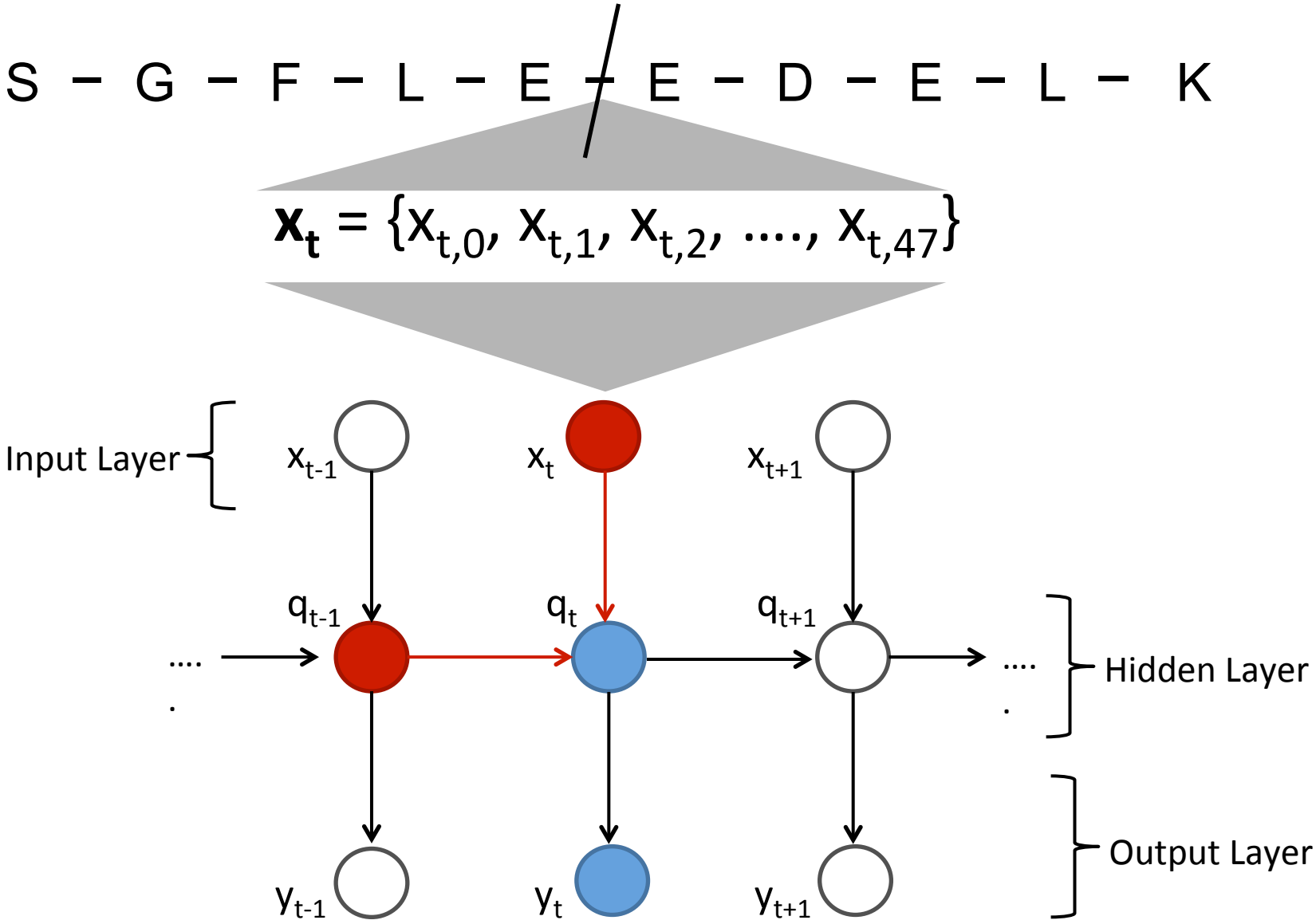
CSPI Model Structure



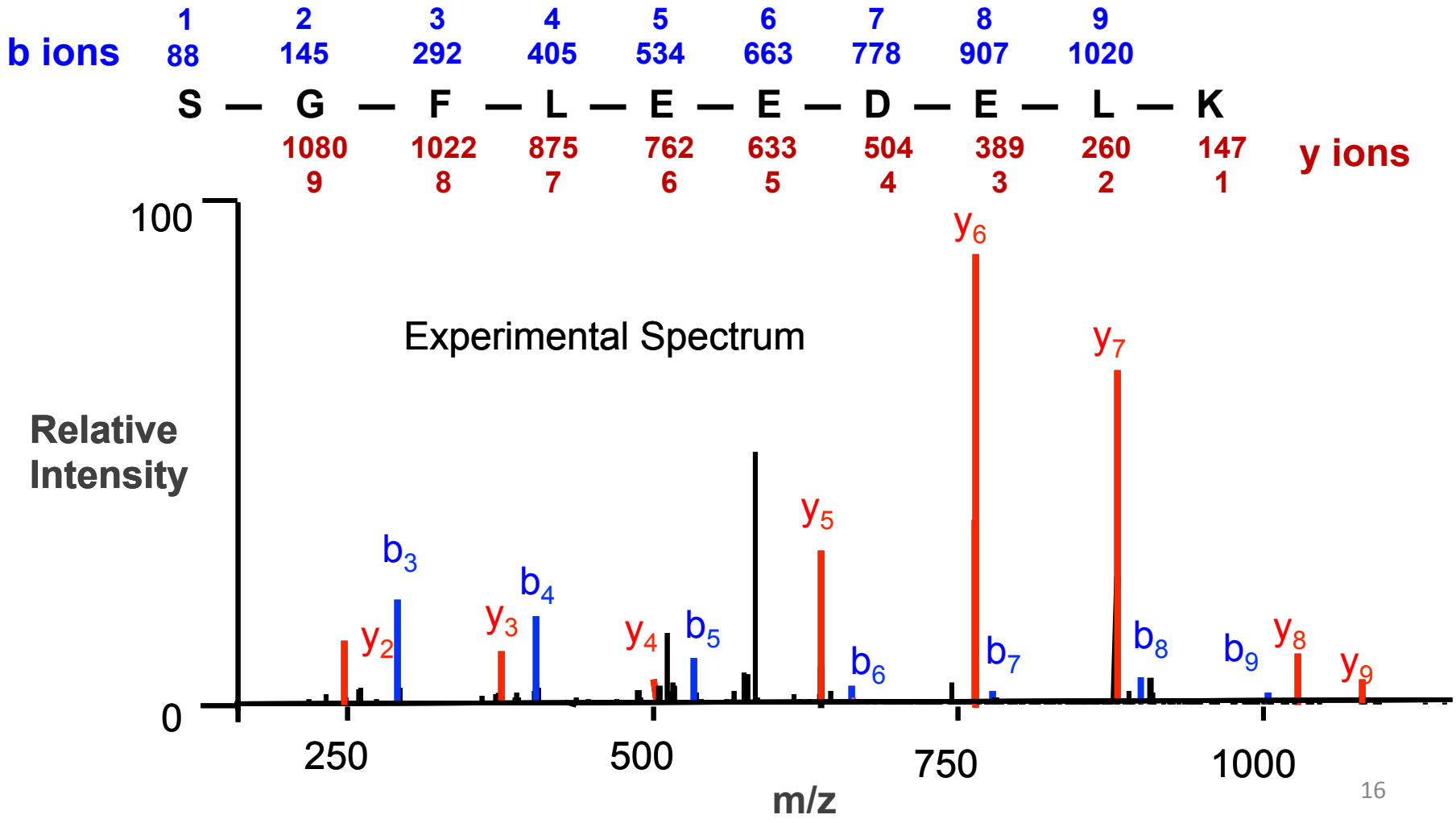
Input Layer: Peptide Physicochemical Context



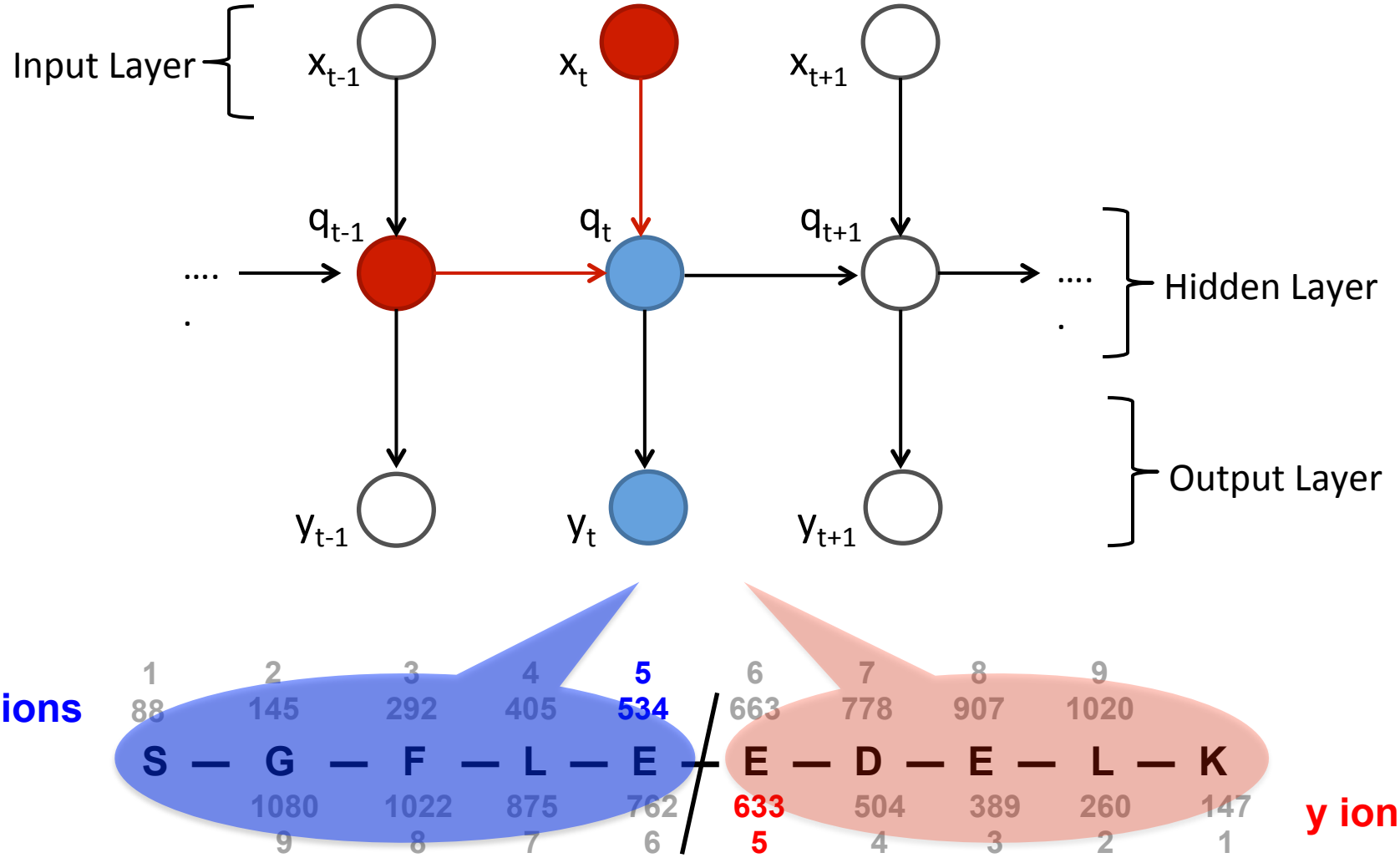
‘Context’ in the context of CSPI



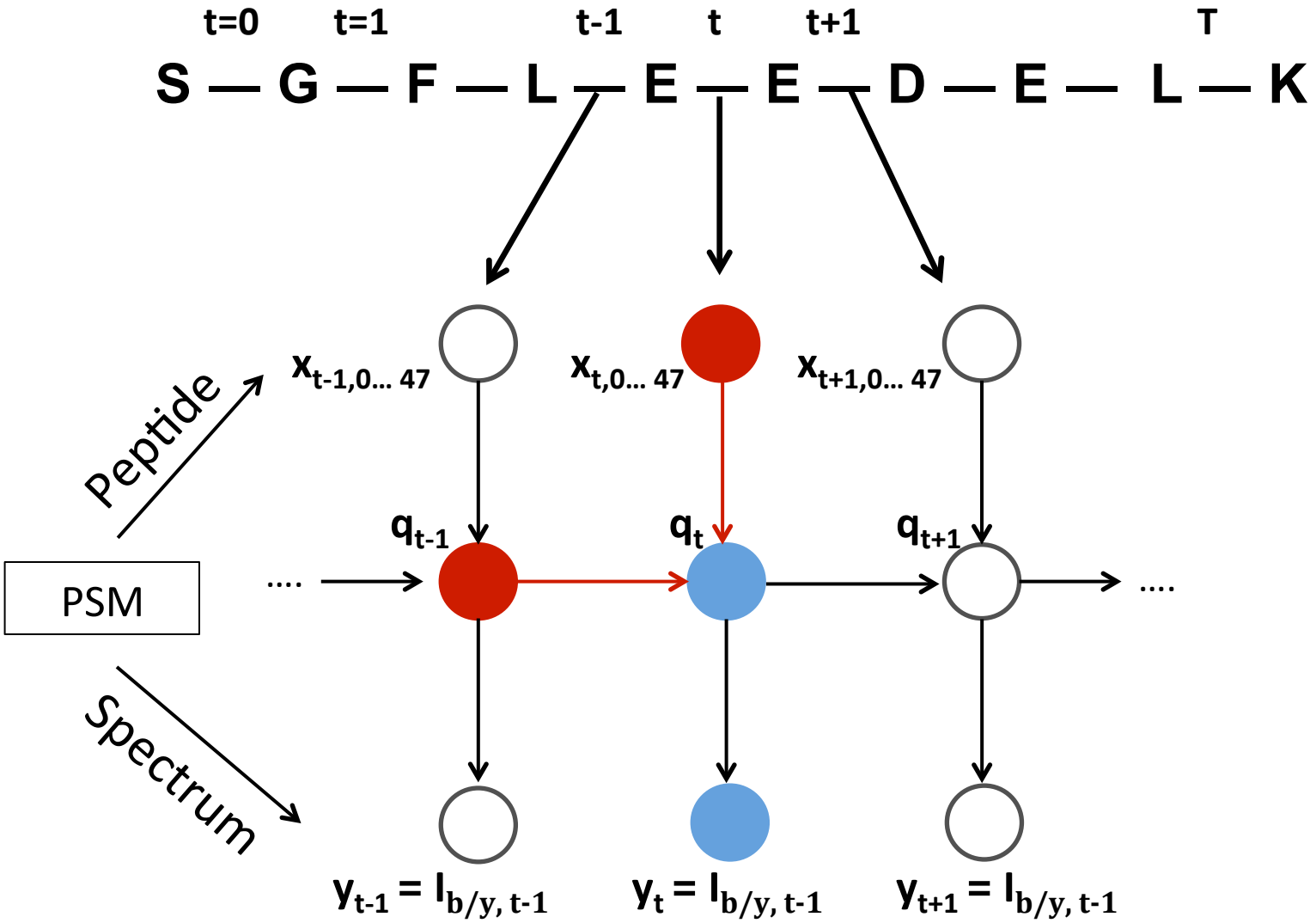
Matching A Peptide with Experimental Spectra



Normalized Intensities in context of CSPI



Summary

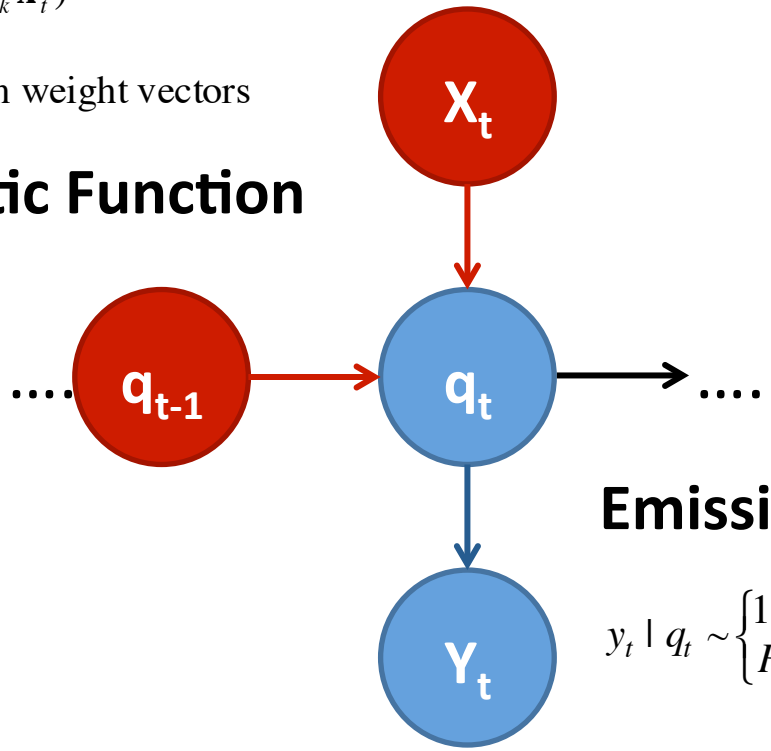

t=0
t=1
t-1
t
t+1
T
S — G — F — L — E — E — D — E — L — K
Peptide
PSM
Spectrum
$x_{t-1,0\ldots 47}$
$x_{t,0\ldots 47}$
$x_{t+1,0\ldots 47}$
q_{t-1}
q_t
q_{t+1}
....
....
$y_{t-1} = I_{b/y,\, t-1}$
$y_t = I_{b/y,\, t-1}$
$y_{t+1} = I_{b/y,\, t-1}$

Parameterization: Transition/Emission Functions

$$P(q_t \mid q_{t-1} = j, x_t; \Theta_{q_t}) = \begin{cases} \dfrac{1}{1 + \sum_{k=1}^{S} \exp(\mathbf{w}_k^T \mathbf{x}_t)} & \text{if } y_t = \text{"NA"} \\ \dfrac{\exp((\mathbf{w}_i^T \mathbf{x}_t))}{1 + \sum_{k=1}^{S} \exp(\mathbf{w}_k^T \mathbf{x}_t)} \quad ; i = 1,2,\ldots,s-1 & \text{if } y_t \text{!=" NA"} \end{cases}$$

where w_i^T are *the* Logistic Regression weight vectors

Logistic Function



Emission Distr^ns

$$y_t \mid q_t \sim \begin{cases} 1.0 & \text{if } y_t = 0 \\ P(\Theta) & \text{if } y_t > 0 \end{cases}$$

where $P = \{Exp(\lambda),\ Be(\alpha,\beta),\ N(\mu,\sigma^2)\}$

Parameter Estimation

- Parameters to estimate per CSPI model (4 hidden states):
  - Over 700 (Logistic function weights, Emission distribution parameters)
- Maximum Likelihood
  - Generalized Expectation Maximization algorithm (GEM)

Inference: Log-likelihood Ratio

- Score: Log Likelihood Ratio

$$CSPI\_Score = \log\left(\frac{P(\text{Spectrum intensities} \mid \text{PeptideSeq};\ \Theta_{True})}{P(\text{Spectrum intensities} \mid \text{PeptideSeq};\ \Theta_{Null})}\right)$$

- Computed using Forward Procedure

Computational bottleneck

- **Database searching**
  - Extract candidate peptides (sub-strings) for each spectrum
- **Candidate Peptides' scoring**
  - **200-400 ^ 3 spectra * ~10-20 ^ 3 peptides**
  - CSPI:
    - Increases performance but…
    - takes ~5-8 seconds per spectrum to evaluate candidates (under constrained searches)

Database Searching

- **Mass-range query**
  - Amino acids (characters) have masses
- **Goal:**
  - Search for sub-strings with a (roughly) specific mass
- **Naïve Approach:**
  - Scan the protein database for each query

Indexed Database Searching

- **Berkeley DB:** key-value store
  - Pre-compute
  - Key: Mass of peptide
  - Value: Location and length of peptide
- Multiple index files
- Time (per query): < 1 sec

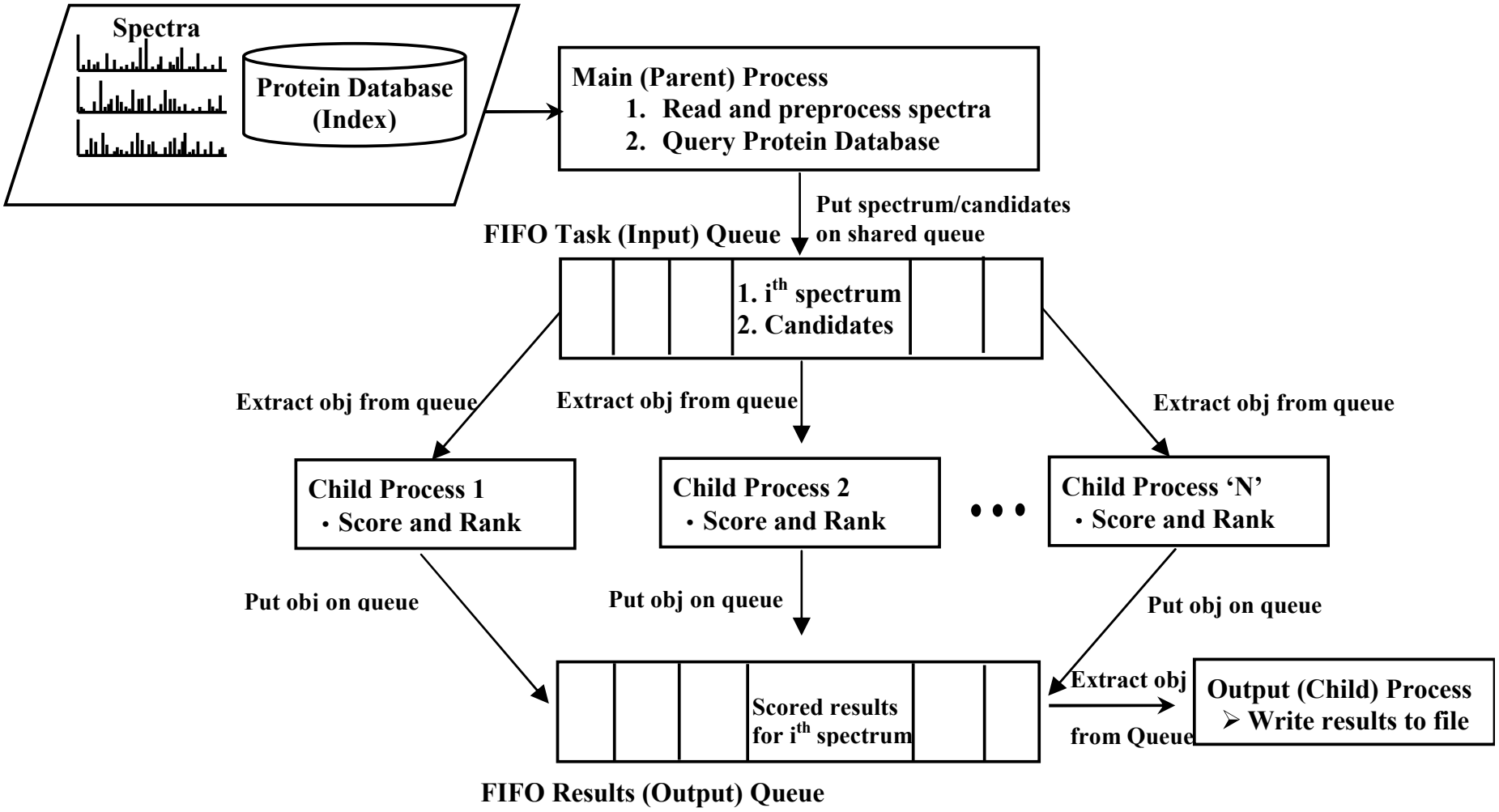
Challenge

- Works well for constrained database searches:
  - Time to generate
  - Size
- Issues with unconstrained searches
- Potential solution:
  - Parallel generation and query
  - Simple synchronization primitives and multiple index files facilitates

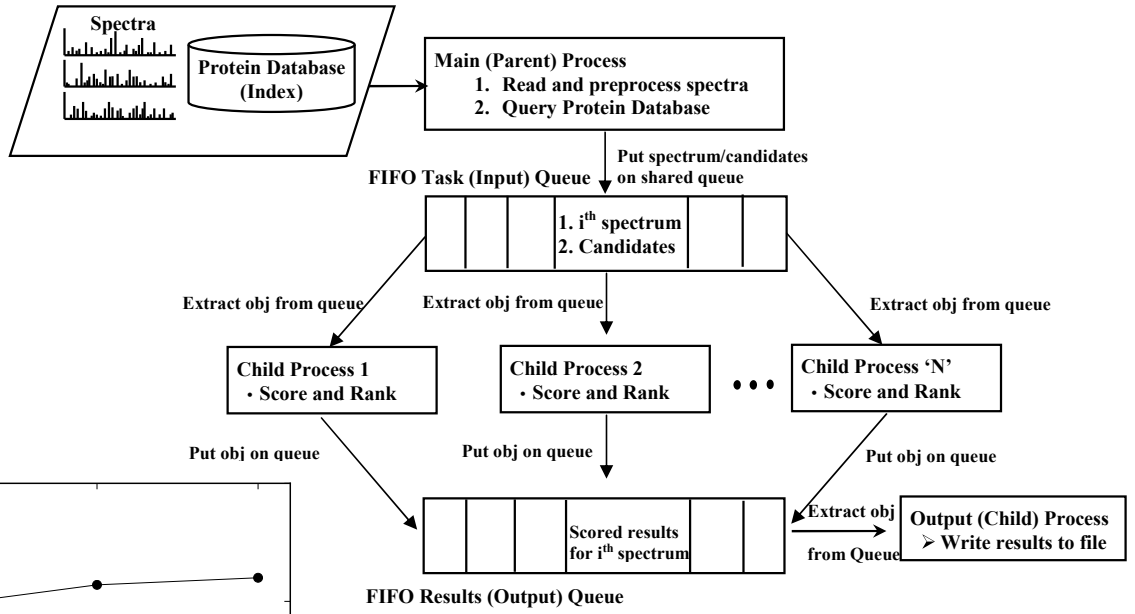
Candidate Peptide Scoring

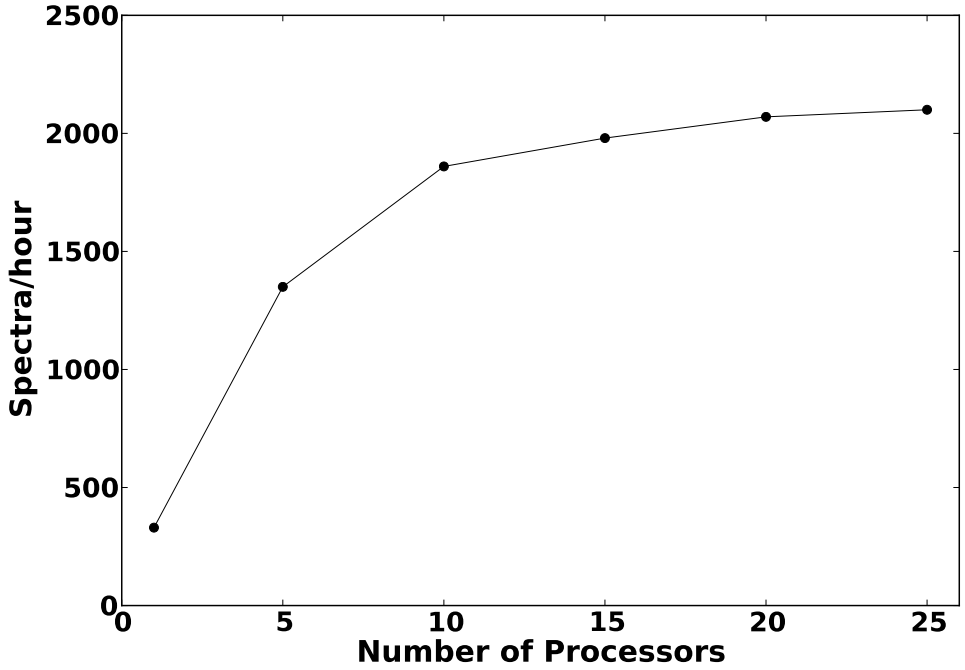
- Embarrassingly parallel
  - For each spectrum, searching and scoring/ ranking is independent of others
- Utilize multiprocessing

Parallel Implementation



Parallel Implementation





Challenges and Potential Solutions

- Spectrum-level parallelization
- Candidate-level optimization can provide further gains:
  - Non-trivial:
    - Careful profiling of individual steps
    - IPC overhead vs. performance gain
      - Protein Database Size
      - Search Constraints

Conclusions and Future Work

- Complex and computationally intensive algorithms
- Collaborative efforts are required for robust analyses (evidence combination)
  - requires efficient processing
  - better parameter estimates
- Further efficiency improvements
- Other applications:
  - Time-series
    - Gene-Expression + Protein-expression
    - MicroRNA expression + Gene Expression
    - Stimulus/Response

Acknowledgements

- Funding Agencies:
  - This work was supported in part by the following grants: NIGMS Award Number K25GM071951, NIH Award Number P41RR006009 and NLM Award Number R01LM010950 to Dr. Vanathi Gopalakrishnan.

Thanks

Questions?