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
• 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
Mass Spectrometry

Analytical tool to identify unknown compounds

Sample → Ionization → Mass Analyzer → Detector

Complex

Collaborative
Amino Acids and Proteins

> IPI:IPI0000005.1 Tax_Id=9606 Gene_Symbol=NRAS GTPase NRas
MTEYKLVVVGAGGGVGSALTIQLiQNHFVDEYDPTIEDSYRKQVVIDGECLLLDILDTAG
QEEYSAAMDQYMRTGEGLCVFAINNSKSFADINLYREQIKRVKDSDDVPMLVGNKCDL
PTRTVDTKQAHELAKSYGIPFIETSAKTRQGVEDAFTLYLREIRQYRMKKNSSDDGTDQG
CMGLPCVVM

> IPI:IPI00000115.1 Tax_Id=9606 Gene_Symbol=CNIH4 Isoform 1 of Protein cornichon homolog 4
MEAVVFVFSLLDCALIIFSYYFIITLSDLEYINARCCSKLNKWKVIPELIGHTIVTV
LLLLMLHWFIFLNLNPVATWNIYRYIMVPSGNGMVFDPTEIHNRGQLKSHMKEAMIKLGF
HLLCFFMYLYSIMILALIND

Amino Acids
Shotgun Proteomics: Protein/Peptide Identification
Database Searching

Predominant methodology for peptide ID from MS/MS
< 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
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 Layer**
  - $x_{t-1}$
  - $x_t$
  - $x_{t+1}$
- **Hidden Layer**
  - $q_{t-1}$
  - $q_t$
  - $q_{t+1}$
- **Output Layer**
  - $P(y_t | q_t; \Theta)$ (Emission Probability)
  - $P(y_t | q_t, x_t; \Theta)$ (Emission Probability)

**Input-output Hidden Markov Model**

- **Input Layer**
  - $x_{t-1}$
  - $x_t$
  - $x_{t+1}$
- **Hidden Layer**
  - $q_{t-1}$
  - $q_t$
  - $q_{t+1}$
- **Output Layer**
  - $P(y_t | q_t, x_t; \Theta)$ (Emission Probability)
CSPI Model Structure

Input Layer:
- \( x_{t-1} \)
- \( q_{t-1} \)
- \( y_{t-1} \)

Hidden Layer:
- \( x_t \)
- \( q_t \)
- \( y_t \)

Output Layer:
- \( x_{t+1} \)
- \( q_{t+1} \)
- \( y_{t+1} \)

\[
P(q_t | q_{t-1}, x_t; \Theta) \]
\[
P(y_t | q_t; \Theta) \]
Input Layer: Peptide Physicochemical Context

S — G — F — L — E — E — D — E — L — K

Global

Experimental Spectrum

Local

Relative Intensity

m/z

0 250 500 750 1000

y2 y3 y4 b3 b4 b5 y5 b6 b7 b8 y6 y7 y8 y9
‘Context’ in the context of CSPI

\[ x_t = \{x_{t,0}, x_{t,1}, x_{t,2}, \ldots, x_{t,47}\} \]
Matching A Peptide with Experimental Spectra

b ions

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>145</td>
<td>292</td>
<td>405</td>
<td>534</td>
<td>663</td>
<td>778</td>
<td>907</td>
<td>1020</td>
</tr>
</tbody>
</table>

S — G — F — L — E — E — D — E — L — K

y ions

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>147</td>
<td>260</td>
<td>389</td>
<td>504</td>
<td>389</td>
<td>260</td>
<td>147</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Experimental Spectrum

Relative Intensity

m/z

0 250 500 750 1000

b3 b4 y2 y3 y4 b5 y5 b6 b7 b8 b9 y8 y9
Normalized Intensities in context of CSPI
Summary

- $x_{t-1,0...47}$
- $x_{t,0...47}$
- $x_{t+1,0...47}$

- $y_{t-1} = l_{b/y, t-1}$
- $y_t = l_{b/y, t}$
- $y_{t+1} = l_{b/y, t-1}$

PSM

Peptide

Specturm
Parameterization: Transition/Emission Functions

\[ P(q_t | q_{t-1} = j, x_t; \Theta_q) = \begin{cases} 
\frac{1}{1 + \sum_{k=1}^{S} \exp(w_k^T x_t)} & \text{if } y_t = "NA" \\
\frac{\exp((w_i^T x_t))}{1 + \sum_{k=1}^{S} \exp(w_k^T x_t)} & \text{if } y_t \neq "NA" 
\end{cases} \]

; \ i = 1, 2, ..., s - 1

where \( w_i^T \) are the Logistic Regression weight vectors

where \( P(q_t | q_{t-1}) \) is the transition probability distribution.

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

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

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

- Computed using Forward Procedure
Computational bottleneck

- **Database searching**
  - Extract candidate peptides (sub-strings) for each spectrum

- **Candidate Peptides’ scoring**
  - $200-400 \times 3$ spectra $\times \sim 10-20 \times 3$ peptides
  - **CSPI:**
    - Increases performance but...
    - Takes $\sim 5-8$ seconds per spectrum to evaluate candidates (under constrained searches)
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

Main (Parent) Process
1. Read and preprocess spectra
2. Query Protein Database

FIFO Task (Input) Queue
- Put spectrum/candidates on shared queue

Child Process 1
- Score and Rank

Child Process 2
- Score and Rank

- Scored results for i\textsuperscript{th} spectrum

Child Process 'N'
- Score and Rank

Output (Child) Process
- Write results to file

FIFO Results (Output) Queue
Parallel Implementation

- Main (Parent) Process
  1. Read and preprocess spectra
  2. Query Protein Database

- FIFO Task (Input) Queue
  - Put spectrum/candidates on shared queue

- Child Process 1
  - Score and Rank

- Child Process 2
  - Score and Rank

- Child Process ‘N’
  - Score and Rank

- FIFO Results (Output) Queue
  - Extract obj from queue
  - Put obj on queue

Graph:

- X-axis: Number of Processors
- Y-axis: Spectra/hour

Graph shows an increasing trend in Spectra/hour as the number of processors increases.
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
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?