Support vector machine evaluation of peptide identification via mass spectrometry

William Stafford Noble
Department of Genome Sciences
Department of Computer Science and Engineering
University of Washington
SVMs in computational biology

- Splice site recognition
- Protein sequence similarity detection
- Protein functional classification
- Regulatory module search

- Protein-protein interaction prediction
- Gene functional classification from microarray data
- Cancer classification from microarray data
• Dave Anderson
• Don Payan
• Weiqun Li
Rigel, Inc.
Outline

- **The task**: Recognizing correct peptide matches.
- **The data**: Thirteen informative features.
- **The algorithm**: The support vector machine.
- **The results**: Multiple data sets, and comparison to other approaches.
Protein mass spectrometry

(Siuzdak 1996)
Mass spectrum

(Siuzdak 1996)
Tandem mass spectrometry

Protein sample

Digestion by trypsin

Peptide pool

Mass spectrometer

Ionization

Single peptide

Mass spectrometer

Peptide spectrum
Database search

Protein sample

Tandem mass spectrometer

Observed spectra

SEQUEST

Sequence database

Predicted peptides
Trypsin

SEQUEST

Predicted peptides
- GDIFYPGYCPDVK
- PVNDFDLSAFAGAWHEIAK
- LPLENENQGK
- CTIAEYK
- YDGK
- ASVYNSFVSNGVK
- EYMEGDLEIAPDAK
- YTK
- QGK
- YVMTFK
- FGQK
- VVNR

Theoretical spectra

Observed spectrum
The learning task

- We are given SEQUEST output: paired observed and theoretical spectra.
- Question: Is the pairing correct?
We need to choose
• the feature set and
• the learning algorithm.
Properties of the observed spectrum

- Total peptide mass. Too small yields little information; too large (>25 amino acids) yields uneven fragmentation.
- Charge (+1, +2 or +3). Provides some evidence about amino acid composition.
- Total ion current. Proportional to the amount of peptide present.
- Peak count. Small indicates poor fragmentation; large indicates noise.
Observed vs. theoretical spectra

- Mass difference.
- Percent of ions matched.
- Percent of peaks matched.
- Percent of peptide fragment ion current matched.
- Preliminary SEQUEST score.
- Cross-correlation.
- Cross-correlation rank.
Percent of ions matched

\[
\frac{7 \text{ matched ions}}{11 \text{ possible ions}} = 63.6\% 
\]
Percent of peaks matched

Number of matched peaks
Total number of peaks

Observed peaks
Percent of peptide fragment ion current matched.

Total intensity of matched peaks
Total intensity of all peaks

This metric weights for matching large peaks

Observed peaks
Preliminary SEQUEST score

The score $S_p$

- is only computed for pairs within a defined mass tolerance,
- accounts for percent of ion matches, continuity, and other factors, and
- can be computed efficiently.
Cross-correlation

\[ R_t (x, y) = \sum_{i=0}^{n-1} x[i]y[i+t] \]

- Theoretical and observed spectra are \( x \) and \( y \), and \( t \) is the offset between them.
- The correlation is computed via FFT.
- \( C_n \) (a.k.a. Xcorr) is the maximal \( R_t \) divided by the mean \( R_t \) for \(-75 < t < 75\), normalized to 1.0.
- Cross-correlation is only computed for the top-scoring 500 peptides.
- Correlation rank is the location of the theoretical spectrum in a list ranked by cross-correlation.
Top-ranked vs. second-ranked peptides

• Change in cross-correlation. Compute the difference in $C_n$ for the top-ranked and second-ranked peptide. 0.1 or greater indicates a significant difference between the first two choices.

• Percent sequence identity. Usually anti-correlated with change in cross-correlation.
### Positive examples

- LNNHAGKWGFHVDGLYK
- LNNHAGKWGFHVDGLYK
- LNNHAGKWGFHVDGLYK
- VHSHLNECTR
- ENVSDHNPSTFLSER
- VHSHLNECTR
- WEIRYKWLDEKLR
- EFPDRHQAOCNQGR
- VHSHLNECTR
- VHSHLNECTR
- ENVSDHNPSTFLSER
- TVHISECDTQSGRMHAR

### Negative examples

- FIGPSMNWVK
- ENVSDHNPSTFLSER
- NHEAVHFLQKPR
- GRVQAQTVYNPDVVK
- VDGIALFWMVWGLRQSR
- AVREAVNHGQLIK
- DHGQSSLTR
- ITSEIPDGEKPSSGTVGELNFR
- ESRFYRQVEGER
- ITSEIPDGEKPSSGTVGELNFR
- NAQYLVLRK
- DMYGDLIK
- ITOIIGEEGIDPENLR
- ITQVYKGHLTITGALNKR
- ITOIIGEEGIDPENLR
- IGSFQPDGLLLFLFR
- LGSIPRAVKARTR
- SVHSSTPLNLK
- TVQAIKDIFR
- LFPGGNEIGMVWNWKR
- DVSQGLORDEEYGPNMIVK
- DVSQGLORDEEYGPNMIVK
- DVSQGLORDEEYGPNMIVK
- VNNRQDNLFR
- RELGPAVELYN
- NQGETLHTGSLPR
- FWKXVMTPEOLK
- ITEEAPRNTIPVRFENWEGDCRHK
- VETQIGGWOK
Fisher criterion score

\[
\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}
\]

Low score

High score
### Feature ranking

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>delta $C_n$</td>
<td>2.861</td>
</tr>
<tr>
<td>% match total ion current</td>
<td>2.804</td>
</tr>
<tr>
<td>$C_n$</td>
<td>2.444</td>
</tr>
<tr>
<td>% match peaks</td>
<td>2.314</td>
</tr>
<tr>
<td>$S_p$</td>
<td>1.158</td>
</tr>
<tr>
<td>mass</td>
<td>0.704</td>
</tr>
<tr>
<td>charge</td>
<td>0.488</td>
</tr>
<tr>
<td>rank $S_p$</td>
<td>0.313</td>
</tr>
<tr>
<td>peak count</td>
<td>0.209</td>
</tr>
<tr>
<td>sequence similarity</td>
<td>0.115</td>
</tr>
<tr>
<td>% ion match</td>
<td>0.079</td>
</tr>
<tr>
<td>total ion current</td>
<td>0.026</td>
</tr>
<tr>
<td>delta mass</td>
<td>0.024</td>
</tr>
</tbody>
</table>
## Pairwise feature ranking

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>% match TIC-delta Cn</td>
<td>4.741</td>
</tr>
<tr>
<td>% match peaks-delta Cn</td>
<td>4.233</td>
</tr>
<tr>
<td>% match TIC-Cn</td>
<td>3.819</td>
</tr>
<tr>
<td>delta Cn-Cn</td>
<td>3.597</td>
</tr>
<tr>
<td>delta Cn-charge</td>
<td>3.563</td>
</tr>
<tr>
<td>% match peaks-Cn</td>
<td>3.377</td>
</tr>
<tr>
<td>delta Cn-mass</td>
<td>3.119</td>
</tr>
<tr>
<td>% match TIC-% match peaks</td>
<td>2.823</td>
</tr>
<tr>
<td>% ion match-delta Cn</td>
<td>2.812</td>
</tr>
<tr>
<td>Sp-delta Cn</td>
<td>2.799</td>
</tr>
<tr>
<td>% match TIC-Sp</td>
<td>2.579</td>
</tr>
<tr>
<td>% match peaks-Sp</td>
<td>2.383</td>
</tr>
<tr>
<td>% match TIC-% ion match</td>
<td>2.097</td>
</tr>
<tr>
<td>% ion match-mass</td>
<td>2.091</td>
</tr>
<tr>
<td>Cn-charge</td>
<td>1.943</td>
</tr>
<tr>
<td>Sp-mass</td>
<td>1.922</td>
</tr>
<tr>
<td>% match TIC-charge</td>
<td>1.898</td>
</tr>
<tr>
<td>Cn-mass</td>
<td>1.884</td>
</tr>
<tr>
<td>Sp-Cn</td>
<td>1.881</td>
</tr>
<tr>
<td>% ion match-Cn</td>
<td>1.827</td>
</tr>
<tr>
<td>Sp-charge</td>
<td>1.770</td>
</tr>
<tr>
<td>% match peaks-mass</td>
<td>1.668</td>
</tr>
<tr>
<td>% match peaks-charge</td>
<td>1.528</td>
</tr>
<tr>
<td>% match TIC-% ion match</td>
<td>1.473</td>
</tr>
</tbody>
</table>
Support vector machine
Support vector machine

Locate a plane that separates positive from negative examples.

Focus on the examples closest to the boundary.
Support vector machine learning

- The SVM learning algorithm finds a linear decision boundary.
- The hyperplane maximizes the margin; i.e., the distance from any training example.
- The optimization is convex; the solution is sparse.
- A soft margin allows for noise in the training set.
- A complex decision surface can be learned by using a non-linear kernel function.
Kernel matrix representation
Kernel matrix representation
\[ K(X, Y) = ((X \cdot Y) + 1)^3 \]
\[ K(X, Y) = \exp\left( -\frac{\|X - Y\|^2}{2\sigma^2} \right) \]
Kernel function

- Let \( p(x,y) \) be the function that computes a 13-element vector of parameters for a pair of spectra, \( x \) and \( y \).
- The kernel function \( K \) operates on pairs of observed and theoretical spectra:

\[
K(S_o^A : S_t^A, S_o^B : S_t^B) = K(p(S_o^A, S_t^A), p(S_o^B, S_t^B)) \\
= (p(S_o^A, S_t^A) \cdot p(S_o^B, S_t^B) + 1)^2
\]
Experimental design

- Data consists of one 13-element vector per predicted peptide.
- Each feature is normalized to sum to 1.0 across all examples.
- The SVM is tested using leave-one-out cross-validation.
- The SVM uses a second-degree polynomial, normalized kernel with a 2-norm asymmetric soft margin.
Three data sets

• **Set 1:** Ion trap mass spectrometer. Sequest search on the full non-redundant database.

• **Set 2:** Ion trap mass spectrometer. Sequest search on human NRDB.

• **Set 3:** Quadrupole time-of-flight mass spectrometer. Sequence search on human NRDB.
## Data set sizes

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ion-trap NRDB</td>
<td>497</td>
<td>479</td>
<td>976</td>
</tr>
<tr>
<td>Ion-trap HNRDB</td>
<td>696</td>
<td>465</td>
<td>1161</td>
</tr>
<tr>
<td>QTOF HNRDB</td>
<td>1017</td>
<td>523</td>
<td>1540</td>
</tr>
</tbody>
</table>
Tuning SVM parameters

- The polynomial degree controls the dimensionality of the feature space.
- The “diagonal factor” controls the softness of the margin.
Varying the polynomial

**Graph:**
- **Y-axis:** ROC score
- **X-axis:** Polynomial degree
- Data set: Ion trap NRDB
- Lines indicated:
  - Black diamond: Polynomial
  - Open square: Radial basis

**Legend:**
- Gaussian curve
- Polynomial curve
Adjusting the soft margin

Ion trap NRDB data set.
Conversion to probabilities

- Hold out a subset of the training examples.
- Use the hold-out set to fit a sigmoid.
  \[
  \Pr(y = 1|f) = \frac{1}{1 + e^{Af+B}}
  \]
  
  - This is equivalent to assuming that the SVM output is proportional to the log-odds of a positive example.

\( y = \text{label} \)
\( f = \text{discriminant} \)
\[ \Pr(x) = \frac{1}{1 + e^{-1.15x + 0.114}} \]
Ubiquitin is a 76 amino acid protein that, when covalently attached to other cell proteins, targets them to the proteasome for degradation.

Ubiquitin attachment can also be used to regulate cellular processes by mechanisms other than degradation.

Proteins are labeled with ubiquitin in lysed cells using cellular enzyme systems specialized for attaching ubiquitin to proteins.

Peptides of proteins affinity extracted with an anti-epitope tag-ubiquitin antibody are analyzed by mass spectrometry.

Proteins identified by only 1 or 2 peptides are analyzed by the SVM and the appropriate training set to calculate the probability that the peptide sequence match is correct.
## Results

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Prob</th>
<th>Protein</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTIAQGGVLPNIQAVLLPK</td>
<td>0.971</td>
<td>Histone H2A</td>
<td>known to be ubiquitinated (positive control)</td>
</tr>
<tr>
<td>AENYDIPSADR</td>
<td>0.929</td>
<td>E1 ubiquitin activating enzyme</td>
<td>important component of the ubiquitin-proteasome system</td>
</tr>
<tr>
<td>NKLDFLRPYTVPNK</td>
<td>0.858</td>
<td>26S proteasome beta 7 subunit</td>
<td>combined with other data, indicates affinity extraction of the proteasome</td>
</tr>
<tr>
<td>VLVALYEEPEKPNSALDFLK</td>
<td>0.922</td>
<td>c-myc binding protein</td>
<td>binds c-myc, which when disregulated is oncogenic in a variety of cancers</td>
</tr>
</tbody>
</table>
Future work

• Test the SVM’s generalization to other data sets.
• Develop a more complete feature set.
• Design algorithms for other mass spec instruments.
• Combine peptide-level predictions into protein-level predictions.
• Anderson, DC, W Li, DG Payan and WS Noble. “A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores.” *Journal of Proteome Research.*

• [http://www.gs.washington.edu/~noble](http://www.gs.washington.edu/~noble)