Learning from Diversity
Epitope Prediction with Sequence and Structure Features
using an Ensemble of Support Vector Machines

Rob Patro and Carl Kingsford
Center for Bioinformatics and Computational Biology
University of Maryland
Nov. 16, 2010
Overview

Challenge: epitope-antibody recognition

Solution: ensemble of support vector machines

- Trained with probabilistic extension
- Variety of feature classes: physicochemical properties, string kernels, structure
- Performance of individual methods and ensemble
Problem Overview

The Challenge

Binding with linear epitopes
“Simpler” sequence $\rightarrow$ affinity relation

The Details

Measure binding affinity $\text{aff}(p_i) \in [0, 65536]$

$C^+ = \{p_i \mid \text{aff}(p_i) \in [10000, 65536]\}$
6,841 binders

$C^- = \{p_i \mid \text{aff}(p_i) \in [0, 1000]\}$
20,437 non-binders

Learn a function to predict binding

$f : \mathcal{P} \rightarrow [0, 1]$

$f(p_i) \geq 0.5 \implies p \in C^+$

$f(p_i) < 0.5 \implies p \in C^-$
System Overview

Individual classifiers trained on various features

Decision Trees, Boosted / Bagged / Random Forests, Naive Bayes, Logistic Regression, Maximum Entropy Classification, (Balanced) Winnow Classifiers, etc.

Support Vector Machines (SVM)

Aggregate scores of classifiers

Produces prediction for binding class

Unlikely Binder

Likely Binder
Probabilistic SVMs

Ideally we want a confidence in each prediction \textbf{(Platt:1999)}.

For each prediction, we obtain a posterior probability.

Allows ranking of predictions by posterior.

Aids in classifier combination.
Combining Predictions

Probabilistic SVMs trained on various features using various kernels, with various parameters

Combined by weighted voting:

\[ p_{\text{ensemble}}^+ (x^i) = \frac{1}{A} \sum_{j=0}^{M} a^j p^j_+ (x^i) \]

- Weight of the \( j^{\text{th}} \) classifier based on cross-validation performance
- Normalizing constant
- Posterior of positive class label for \( j^{\text{th}} \) classifier
Choosing Features

To train SVMs we translate each peptide $p^i$ into a feature vector $x^i$.

Good features are essential.

- Good features should be discriminative.
- Lead to class separability.
- Be efficient to compute.

- Real features capture partial information.
- Separate data subsets.
- Are often complementary.

Consider many useful features → predictive power.
Which Features?

**Sequence Features**
- k-spectrum kernel
- mismatch kernel
- substitution kernel
- string subsequence kernel
- sparse spatial sample kernel

**Physico-Biochemical Features**
- BLOSUM encoding
- AAIndex encoding
- Local composition

**Structure Features**
- Peptide/Structure shape complementarity

ILAMRSHYPF

[Chemical structure of an amino acid]

[Diagram of a protein structure]
Sequence Features (String Kernels)

String kernels assign a similarity to a pair of strings

K-spectrum kernel
Consider all $(K)$ k-mers that occur in the training set
Encode each peptide as a vector $\mathbf{v} \in \mathbb{R}^K$

$$v_j = \begin{cases} 
1 & \text{if } p \text{ contains the } j^{th} \text{ k-mer} \\
0 & \text{otherwise}
\end{cases}$$

or $v_j$ can encode the frequency of the $j^{th}$ k-mer in the peptide.

Other string kernels: Mismatch kernel, Substitution kernel, Restricted gappy kernel, String subsequence kernel, Sparse Spatial Sample (SSS) kernel
Compositional Features

Consider physicochemical properties of each peptide sequence

Hydropathy, Antigenicity, Structure preference etc.

Average property over entire peptide

Map each peptide to a scalar \( v \in \mathbb{R} \)
Local Compositional Features

Physicochemical features can be useful but are global.

Epitope is only a subset of the peptide.

Consider a sliding window of a given length $w$.
Move window along the peptide from left to right.
Average values over window.
Concatenate output to represent the peptide.
Orthogonal Encoding

Orthonormal representation proposed by Qian1988

Map each amino acid $a_j \in p^i$ to a 20 long bit-vector $v_j$

$x^i = v_0 v_1 \ldots v_{k-1}$ for an amino acid of length $k$
Orthogonality is not actually important in our application

Replace the indicator vector by something more informative

e.g. a row from a BLOSUM or PAM matrix
AAIndex Encoding

The Amino Acid Index (AAIndex) ([Kawashima2008](#)) compiles a growing list of different phyisicochemical and biochemical properties of amino acids . . . 544 to date!

Is it possible to make use of all this information? Use non-linear factor matrix of AAIndex ([Nanni2010](#))
Structural Features

Consider how well IgG and a peptide “fit” together.

**Poor Shape**

**Good Shape**

Complementarity

Experimentally measured IgG conformation

Choose most common sidechain positions

Relax energy

Approximate native peptide conformation

Compute 2000 "best" dockings using ZDock (Chen, Li, and Wang 2003)

Feature vector given by histogram of docking scores
## Results Table

<table>
<thead>
<tr>
<th>Features</th>
<th>AUROC</th>
<th>AUPR</th>
<th>∆AUROC</th>
<th>∆AUPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-spectrum</td>
<td>0.85</td>
<td>0.70</td>
<td>-0.043</td>
<td>-0.072</td>
</tr>
<tr>
<td>Sparse Spatial Sample</td>
<td>0.87</td>
<td>0.73</td>
<td>-0.023</td>
<td>-0.042</td>
</tr>
<tr>
<td>Nonlinear Fisher Mat.</td>
<td>0.86</td>
<td>0.69</td>
<td>-0.024</td>
<td>-0.082</td>
</tr>
<tr>
<td>Statistical Analysis Mat.</td>
<td>0.85</td>
<td>0.67</td>
<td>-0.025</td>
<td>-0.102</td>
</tr>
<tr>
<td>BLOSUM Encoding</td>
<td>0.86</td>
<td>0.70</td>
<td>-0.024</td>
<td>-0.072</td>
</tr>
<tr>
<td>Local Composition*</td>
<td>0.88</td>
<td>0.74</td>
<td>-0.013</td>
<td>-0.032</td>
</tr>
<tr>
<td>Structure</td>
<td>0.74</td>
<td>0.53</td>
<td>-0.153</td>
<td>-0.242</td>
</tr>
<tr>
<td><strong>ensemble</strong></td>
<td><strong>0.893</strong></td>
<td><strong>0.772</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2nd Place</strong></td>
<td><strong>0.892</strong></td>
<td><strong>0.766</strong></td>
<td>-0.001</td>
<td>-0.006</td>
</tr>
<tr>
<td><strong>3rd Place</strong></td>
<td><strong>0.864</strong></td>
<td><strong>0.691</strong></td>
<td>-0.029</td>
<td>-0.081</td>
</tr>
<tr>
<td><strong>4th Place</strong></td>
<td><strong>0.855</strong></td>
<td><strong>0.689</strong></td>
<td>-0.038</td>
<td>-0.083</td>
</tr>
</tbody>
</table>

*using various physicochemical features*
Performance Curves ROC

Receiver Operating Characteristic Curve

- ensemble
- kspectrum
- SSS
- PE: NLF
- PE: SA
- PE: SBLO
- LPE: SWIN
- struct

True Positive Rate vs. False Positive Rate
Performance Curves P/R

Precision/Recall Curve

- ensemble
- kspectrum
- GSS
- PE: NLF
- PE: SA
- PE: SBLO
- LPE: SWIN
- struct

Precision

Recall
Conclusions

- Many good features exist
- They capture *some* non-overlapping information
- Ensemble solutions, used properly, are effective
- Structure features are hard to compute
- Much room for improvement here
- Simple features should not be discounted
- The local composition feature was the best *single* classifier
- We didn’t encounter anyone using this in the literature!
Thanks

Funding:

NIH grant 1R21AI085376 and NSF grant 0849899 to C.K.

For many interesting and useful conversations:

Geet Duggal, Darya Filippova, Justin Malin, Guillaume Marçais, Saket Navlakha, Emre Sefer