B CELL EPITOPES AND PREDICTIONS
OUTLINE

• What is a B-cell epitope?

• How can you predict B-cell epitopes?
WHAT IS A B-CELL EPITOPE?

- B-cell epitopes:
  - Accessible structural feature of a pathogen molecule.
  - Antibodies are developed to bind the epitope specifically using the complementary determining regions (CDRs).

Antibody Fab fragment
THE BINDING INTERACTIONS

- Salt bridges
- Hydrogen bonds
- Hydrophobic interactions
- Van der Waals forces

Binding strength
B-CELL EPITOPE CLASSIFICATION

B-cell epitope: **structural feature of a molecule or pathogen, accessible and recognizable by B-cell receptors and antibodies**

Linear epitopes
One segment of the amino acid chain

Discontinuous epitope (with linear determinant)

Discontinuous epitope
Several small segments brought into proximity by the protein fold
BINDING OF A DISCONTINUOUS EPITOPE

Antibody FAB fragment complexed with Guinea Fowl Lysozyme (1FBI).

Black: Light chain, Blue: Heavy chain, Yellow: Residues with atoms distanced < 5Å from FAB antibody fragments.

Guinea Fowl Lysozyme

KVFGRCELAAAMKRHLGDNRYRGYSLGNWVCAAKFESNFNSQNRNDG
DYGVLNSRWYCDGRTPGSRNLCCNIPCSALQSSDTATANCACKIVSDG
GMNAWVAVWRCCKGTDVRVWIKGCR
B-CELL EPITOPE ANNOTATION

• Linear epitopes:
  • Chop sequence into small pieces and measure binding to antibody

• Discontinuous epitopes:
  • Measure binding of whole protein to antibody

• The best annotation method: X-ray crystal structure of the antibody-epitope complex
B-CELL EPITOPE DATA BASES

• Databases:
  • IEDB, Los Alamos HIV database, Protein Data Bank, Antijen, BciPep

• Large amount of data available for linear epitopes

• Few data available for discontinuous
SEQUENCE-BASED METHODS FOR PREDICTION OF LINEAR EPITOPES

- **Protein hydrophobicity** – hydrophilicity algorithms
  - Parker, Fauchere, Janin, Kyte and Doolittle, Manavalan
  - Sweet and Eisenberg, Goldman, Engelman and Steitz (GES), von Heijne

- **Protein flexibility prediction algorithm**
  - Karplus and Schulz

- **Protein secondary structure prediction algorithms**
  - PsiPred (D. Jones)

- **Protein “antigenicity” prediction**:
  - Hopp and Woods, Welling

TSQDLSVPPLASCCKDNIASTSVTLGCLVTGYLP
MSTTVTWDGSLNKVNTTFPTTFHETYGLHISISV
QVTASGKWAQRFTCVAHAESTAIKTSACALNFIPPTVKLFHSSCNVPVGDTTTIQLLCLISGYVPGDMEVIWLVDGQKAATNIFPYTAPGTKEGNVTSTHSLENITQGEWVSQKTYTCQVTQFGFTFKEARKCSES

Friday, 11 June 2010
PROPENSITY SCALES: THE PRINCIPLE

• The Parker hydrophilicity scale

• Derived from experimental data

<table>
<thead>
<tr>
<th>Residue</th>
<th>Propensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>2.46</td>
</tr>
<tr>
<td>E</td>
<td>1.86</td>
</tr>
<tr>
<td>N</td>
<td>1.64</td>
</tr>
<tr>
<td>S</td>
<td>1.50</td>
</tr>
<tr>
<td>Q</td>
<td>1.37</td>
</tr>
<tr>
<td>G</td>
<td>1.28</td>
</tr>
<tr>
<td>K</td>
<td>1.26</td>
</tr>
<tr>
<td>T</td>
<td>1.15</td>
</tr>
<tr>
<td>R</td>
<td>0.87</td>
</tr>
<tr>
<td>P</td>
<td>0.30</td>
</tr>
<tr>
<td>H</td>
<td>0.30</td>
</tr>
<tr>
<td>C</td>
<td>0.11</td>
</tr>
<tr>
<td>A</td>
<td>0.03</td>
</tr>
<tr>
<td>Y</td>
<td>-0.78</td>
</tr>
<tr>
<td>V</td>
<td>-1.27</td>
</tr>
<tr>
<td>M</td>
<td>-1.41</td>
</tr>
<tr>
<td>I</td>
<td>-2.45</td>
</tr>
<tr>
<td>F</td>
<td>-2.78</td>
</tr>
<tr>
<td>L</td>
<td>-2.87</td>
</tr>
<tr>
<td>W</td>
<td>-3.00</td>
</tr>
</tbody>
</table>

Hydrophilicity
PROPENSITY SCALES: THE PRINCIPLE

\[\frac{-2.78 + -1.27 + 2.46 + 1.86 + 1.26 + 0.87 + 0.3}{7} = 0.39\]

Prediction scores:

0.39  0.1  0.6  0.9  1.0  1.2  2.6  1.0  0.9  0.5  -0.5

Epitope
TURN PREDICTION AND B-CELL EPITOPES

• Pellequer found that 50% of the epitopes in a data set of 11 proteins were located in turns.

Turn propensity scales for each position in the turn were used for epitope prediction.

Pellequer et al., Immunology letters, 1993
• Extensive evaluation of propensity scales for epitope prediction

• Conclusion:
  – Basically all the classical scales perform close to random!
  – Other methods must be used for epitope prediction
BEPIPRED

• Parker hydrophilicity scale

• PSSM

• PSSM based on linear epitopes extracted from the Antijen database

• Combination of the Parker prediction scores and PSSM leads to prediction score

• Tested on the Pellequer dataset and epitopes in the HIV Los Alamos database
## PSSM

<table>
<thead>
<tr>
<th>Position</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.28</td>
</tr>
<tr>
<td>2</td>
<td>9.39</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>4</td>
<td>5.2</td>
</tr>
<tr>
<td>5</td>
<td>7.9</td>
</tr>
</tbody>
</table>

### Sequence:

```
....LISTFVDEKRGSDIVEDLILKDENKTTVI....
```

### Prediction Value:

$$2.46 + 1.86 + 1.26 + 0.87 + 0.3 = 6.75$$

*Prediction value*
ROC EVALUATION

Evaluation on HIV Los Alamos data set
BEPIPRED PERFORMANCE

• Pellequer data set:
  - Levitt  AROC = 0.66
  - Parker  AROC = 0.65
  - BepiPred AROC = 0.68

• HIV Los Alamos data set
  - Levitt  AROC = 0.57
  - Parker  AROC = 0.59
  - BepiPred AROC = 0.60
BEPIPRED

• BepiPred conclusion:
  • On both of the evaluation data sets, Bepipred was shown to perform better
  • Still the AROC value is low compared to T-cell epitope prediction tools!
  • Bepipred is available as a webserver:
    • www.cbs.dtu.dk/services/BepiPred
## Prediction of Linear Epitopes

<table>
<thead>
<tr>
<th><strong>Pro</strong></th>
<th><strong>Con</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>easily predicted computationally</td>
<td>only ~10% of epitopes can be classified as “linear”</td>
</tr>
<tr>
<td>easily identified experimentally</td>
<td>weakly immunogenic in most cases</td>
</tr>
<tr>
<td>immunodominant epitopes in many cases</td>
<td>most epitope peptides do not provide antigen-neutralizing immunity</td>
</tr>
<tr>
<td>do not need 3D structural information</td>
<td>in many cases represent hypervariable regions</td>
</tr>
<tr>
<td>easy to produce and check binding activity experimentally</td>
<td></td>
</tr>
</tbody>
</table>
SEQUENCE BASED PREDICTION METHODS

• Linear methods for prediction of B cell epitopes have low performances
• The problem is analogous to the problems of representing the surface of the earth on a two-dimensional map
• Reduction of the dimensions leads to distortions of scales, directions, distances
• The world of B-cell epitopes is 3 dimensional and therefore more sophisticated methods must be developed

Regenmortel 1996, Meth. of Enzym. 9.
SO WHAT IS MORE SOPHISTICATED?

• Use of the three dimensional structure of the pathogen protein

• Analyze the structure to find surface exposed regions

• Additional use of information about conformational changes, glycosylation and trans-membrane helices
SOURCES OF THREE-DIMENSIONAL STRUCTURES

- Experimental determination
  - X-ray crystallography
  - NMR spectroscopy
- Structure prediction
  - Homology modeling
  - Fold recognition
- Both methods are time consuming and not easily done in a larger scale
- Less time consuming, but there is a possibility of incorrect predictions, specially in loop regions
PROTEIN STRUCTURE PREDICTION METHODS

• Homology/comparative modeling
  >25% sequence identity (seq 2 seq alignment)

• Fold-recognition
  <25% sequence identity (Psi-blast search/ PSSM 2 seq alignment)

• Ab initio structure prediction
  0% sequence identity
WHAT DOES ANTIBODIES RECOGNIZE IN A PROTEIN?

A: Everything accessible to a 10 Å probe on a protein surface

THE CEP SERVER

• Conformational epitope server
  http://202.41.70.74:8080/cgi-bin/cep.pl

• Uses protein structure as input

• Finds stretches in sequences which are surface exposed
THE DISCOTOPE SERVER

• CBS server for prediction of discontinuous epitopes

• Uses protein structure as input

• Combines propensity scale values of amino acids in discontinuous epitopes with surface exposure

• [http://www.cbs.dtu.dk/services/DiscoTope](http://www.cbs.dtu.dk/services/DiscoTope)
DISCOTOPE

• Prediction of residues in discontinuous B cell epitopes using protein 3D structures

Pernille Haste Andersen, Morten Nielsen and Ole Lund, Protein Science 2006
A DATA SET OF DISCONTINUOUS B CELL EPITOPES

- Structures of antibodies/antigen protein complexes in the Protein DataBank

- Dr. Andrew Martin’s SACS database (available at http://www.bioinf.org.uk/abs/sacs) was used to get an overview of PDB entries

- Epitopes in the data set were identified by finding residues within 4Å from heavy or light chains in the Abs

- We used homology grouping and cross-validation for the training and testing of the method to avoid biasing towards specific antigens

- The 5 sets used for cross-validated training/testing are available at: http://www.cbs.dtu.dk/suppl/immunology/DiscoTope.php

An example: The epitope of the outer surface protein A from Borrelia Burgdorferi (1OSP)
## AMINO ACIDS IN EPITOPES

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>G</th>
<th>A</th>
<th>V</th>
<th>L</th>
<th>I</th>
<th>M</th>
<th>P</th>
<th>F</th>
<th>W</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>e/E</td>
<td>0.09</td>
<td>0.07</td>
<td>0.05</td>
<td>0.08</td>
<td>0.04</td>
<td>0.02</td>
<td>0.06</td>
<td>0.03</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.08</td>
<td>0.07</td>
<td>0.10</td>
<td>0.06</td>
<td>0.03</td>
<td>0.05</td>
<td>0.05</td>
<td>0.02</td>
<td>0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>C</th>
<th>T</th>
<th>Q</th>
<th>N</th>
<th>H</th>
<th>Y</th>
<th>E</th>
<th>D</th>
<th>K</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>e/E</td>
<td>0.03</td>
<td>0.08</td>
<td>0.04</td>
<td>0.04</td>
<td>0.02</td>
<td>0.04</td>
<td>0.06</td>
<td>0.07</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.06</td>
<td>0.04</td>
<td>0.05</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
</tr>
</tbody>
</table>
LOG-ODDS RATIOS OF AMINO ACIDS IN DISCONTINUOUS EPITOPEs

Frequencies of amino acids in epitope residues compared to frequencies of non-epitope residues

Several discrepancies compared to the Parker hydrophilicity scale

Predictive performance (AUC) of B cell epitopes:
- Parker: 0.614
- Epitope log–odds: 0.634

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Parker</th>
<th>Log-odds Ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>2.460</td>
<td>0.691</td>
</tr>
<tr>
<td>E</td>
<td>1.860</td>
<td>0.346</td>
</tr>
<tr>
<td>N</td>
<td>1.640</td>
<td>1.242</td>
</tr>
<tr>
<td>S</td>
<td>1.500</td>
<td>-0.145</td>
</tr>
<tr>
<td>Q</td>
<td>1.370</td>
<td>1.082</td>
</tr>
<tr>
<td>G</td>
<td>1.280</td>
<td>0.189</td>
</tr>
<tr>
<td>K</td>
<td>1.260</td>
<td>1.136</td>
</tr>
<tr>
<td>T</td>
<td>1.150</td>
<td>-0.233</td>
</tr>
<tr>
<td>R</td>
<td>0.870</td>
<td>1.180</td>
</tr>
<tr>
<td>P</td>
<td>0.300</td>
<td>1.164</td>
</tr>
<tr>
<td>H</td>
<td>0.300</td>
<td>1.098</td>
</tr>
<tr>
<td>C</td>
<td>0.110</td>
<td>-3.519</td>
</tr>
<tr>
<td>A</td>
<td>0.030</td>
<td>-1.522</td>
</tr>
<tr>
<td>Y</td>
<td>-0.780</td>
<td>0.030</td>
</tr>
<tr>
<td>V</td>
<td>-1.270</td>
<td>-1.474</td>
</tr>
<tr>
<td>M</td>
<td>-1.410</td>
<td>0.273</td>
</tr>
<tr>
<td>I</td>
<td>-2.450</td>
<td>-0.713</td>
</tr>
<tr>
<td>F</td>
<td>-2.780</td>
<td>-1.147</td>
</tr>
<tr>
<td>L</td>
<td>-2.870</td>
<td>-1.836</td>
</tr>
<tr>
<td>W</td>
<td>-3.000</td>
<td>-0.064</td>
</tr>
</tbody>
</table>

*Amino acids are listed with descending hydrophilicity using the values of the Parker scale.*
DISCOTOPE: A PREDICTION METHOD USING 3D STRUCTURES

A combination method:

- Addition of epitope log-odds values of residues in spatial proximity
- Contact numbers

```
LIST..FVDEKRPGSDIVED......ALILKDENKTTVI.
```

Contact number: 7

```
-0.145
+0.691+0.346+1.136+1.180+1.164
+0.346
+1.136
```

Sum of log-odds values

```
W
```

DiscoTope prediction value
Receiver Operator Characteristics (ROC) curves were used for performance measures.

The reported performance is an average of the AUC values of the non-homologous groups of antigens:

- Parker: 0.614  Seq.-based
- Epitope log–odds: 0.634  Seq.-based
- Contact numbers: 0.647  Str.-based
- Naccess: 0.673  Str.-based
- DiscoTope: 0.711  Seq./Str.-based
EVALUATION EXAMPLE AMA I

• Apical membrane antigen 1 from *Plasmodium falciparum* (not used for training/testing)

• Two epitopes were identified using phage-display, sequence variance analysis and point-mutation

  (green backbone)

• Most residues identified as epitopes were successfully predicted by DiscoTope

  (black side chains)

DiscoTope is available as webservice: [http://www.cbs.dtu.dk/services/DiscoTope/](http://www.cbs.dtu.dk/services/DiscoTope/)
**RECENT DEVELOPMENTS**

**APPLICATIONS NOTE**

*Structural bioinformatics*

**PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure**

Michael J. Sweredoski\(^1,2\) and Pierre Baldi\(^1,2,\ast\)

\(^1\)Department of Computer Science and \(^2\)Institute for Genomics and Bioinformatics, University of California, Irvine, 92697-3435, California, USA

**BMC Bioinformatics**

Software

**ElliPro: a new structure-based tool for the prediction of antibody epitopes**

Julia Ponomarenko\(^1,2\), Huynh-Hoa Bui\(^3\), Wei Li, Nicholas Fusseder, Philip E Bourne\(^1,2\), Alessandro Sette\(^4\) and Bjoern Peters\(^4\)

Address: \(^1\)San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA, \(^2\)Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA, \(^3\)Isis Pharmaceuticals, Inc., 1896 Rutherford Road, Carlsbad, California 92008, USA and \(^4\)La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, California 92037, USA
# Secondary Structure in Epitopes

<table>
<thead>
<tr>
<th>Sec struct:</th>
<th>H</th>
<th>T</th>
<th>B</th>
<th>E</th>
<th>S</th>
<th>G</th>
<th>I</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log odds ratio</td>
<td>-0.19</td>
<td>0.30</td>
<td>0.21</td>
<td>-0.27</td>
<td>0.24</td>
<td>-0.04</td>
<td>0.00</td>
<td>0.17</td>
</tr>
</tbody>
</table>

- **H**: Alpha-helix (hydrogen bond from residue i to residue i+4)
- **G**: 310-helix (hydrogen bond from residue i to residue i+3)
- **I**: Pi helix (hydrogen bond from residue i to residue i+5)
- **E**: Extended strand
- **B**: Beta bridge (one residue short strand)
- **S**: Bend (five-residue bend centered at residue i)
- **T**: H-bonded turn (3-turn, 4-turn or 5-turn)
- **C**: Coil
RATIONAL VACCINE DESIGN

PATHOGEN PROTEIN
KVFGRCSELAAAMKRHLGNYRGY
SLGNWVCAAKFESNF

Rational Vaccine Design
RATIONAL B-CELL EPITOPE DESIGN

• Protein target choice

• Structural analysis of antigen

- Known structure or homology model
- Precise domain structure
- Physical annotation (flexibility, electrostatics, hydrophobicity)
- Functional annotation (sequence variations, active sites, binding sites, glycosylation sites, etc.)
RATIONAL B-CELL EPITOPE DESIGN

• Protein target choice

• Structural annotation

• Epitope prediction and ranking

- Surface accessibility
- Protrusion index
- Conserved sequence
- Glycosylation status
RATIONAL B-CELL EPITOPE DESIGN

• Protein target choice
• Structural annotation
• Epitope prediction and ranking

• Optimal Epitope presentation

- Fold minimization, or
- Design of structural mimics
- Choice of carrier (conjugates, DNA plasmids, virus like particles)
- Multiple chain protein engineering
MULTI-EPITOPE PROTEIN DESIGN

Rational optimization of epitope-VLP chimeric proteins:

- Design a library of possible linkers (<10 aa)
- Perform global energy optimization in VLP (virus-like particle) context
- Rank according to estimated energy strain
CONCLUSIONS

• Rational vaccines can be designed to induce strong and epitope-specific B-cell responses

• Selection of protective B-cell epitopes involves structural, functional and immunogenic analysis of the pathogenic proteins

• When you can: Use protein structure for prediction

• Structural modeling tools are helpful in prediction of epitopes, design of epitope mimics and optimal epitope presentation