Bioinformatics 1 -- Lecture 23

- Immunoinformatics, cont’d...
- T-cell epitope prediction
  - MHC I/II binding prediction
- B-cell epitope prediction
- SNPs
  - t-SNPs
T-cell epitopes

- T-cell epitopes are short peptide sequences that elicit the cellular immune response, that is, activated T-cell clones.

B-cell epitopes

- B-cell epitopes are peptides or other biomolecules that bind specifically to antibodies.
B and T cells recognize different epitopes of the same protein

<table>
<thead>
<tr>
<th>T-cell epitope</th>
<th>B-cell epitope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denatured antigen</td>
<td>Native or denatured (rare) antigen</td>
</tr>
<tr>
<td>Linear (often) peptide 8-37 aa</td>
<td>Sequential (continuous) or conformational (discontinuous)</td>
</tr>
<tr>
<td>Internal (often)</td>
<td>Accessible, hydrophilic, mobile, usually on the surface or could be exposed as a result of physicochemical change</td>
</tr>
</tbody>
</table>

**Binding to T cell receptor:**

- $K_d \ 10^{-5} - 10^{-7} \text{ M (low affinity)}$
- Slow on-rate, slow off-rate (once bound, peptide may stay associated for hours to many days)

**Binding to antibody:**

- $K_d \ 10^{-7} - 10^{-11} \text{ M (high affinity)}$
- Rapid on-rate, variable off-rate
MHC class I gets peptides from the cytoplasm. These are endogenous, usually. Exposed to peptides from the proteosome, in the ER. Recognized by CD8 T-cells.

MHC class II gets peptides from outside the cell through phagocytosis. These are exogenous, usually. Blocked from peptides in the ER. Exposed to peptides from lysosomes, in a vesicle. Recognized by CD4 T-cells.
Modes of interaction with different T-cells.
Complex Of A Human TCR, Influenza HA Antigen Peptide (PKYVKQNTLKLAT) and MHC Class II

Xenoreactive Complex AHIII 12.2 TCR bound to P1049 (ALWGFPVLS) /HLA-A2.1

T-Cell Receptor

Vα

Vβ

MHC class II α

β-2-Microglobulin

1fyt

1lp9
The strength of interaction with TCR depends on the strength of the peptide binding to MHC class II.

Long, open-ended pocket sandwiched between two alpha helices.
Different alleles of MHC II will have different AAs in the binding pocket. The pocket can be divided up into 9 regions, each having one of several shapes, depending on the allele. Alleles are shuffled so that all combinations exist. In this view pocket shapes and specificities map directly to alleles.

Sturniolo et al, Nature Biotechnology, 1999
Pocket profiles for MHC class II

The program TEPITOPE calculates the binding energy and allotype simultaneously, assuming the binding energy is the sum of “pocket profiles”. Each pocket profile is the ratio of binding affinity before/after one position (position 9 in this example) is switched to alanine (A). Resulting profiles at pocket 9 are similar for different peptide sequences. The pocket profiles depend on the allotype, and the pocket.

Sturniolo et al, Nature Biotechnology, 1999
Properties of B-cell epitopes

- Must be on the protein surface
- Must be exposed to solvent
- May be linear in sequence, of “conformational”
  - Most algorithms predict linear epitopes
Geneious hydrophobicity plot

In Geneious: Graph hydrophobicity. Look for low hydrophobicity (hydrophilicity)

Potential B-cell epitopes
BcePRED compared to Geneious hydrophobicity plot

BcePRED allows user to choose from several published sets of physico-chemical properties. These are applied in a sliding window scan of the sequence.

Geneious hydrophobicity
 Immunoinformatics Servers

• **Prediction of proteasome cleavages**
  - **MAPPP**, MHC-I Antigenic Peptide Processing Prediction, *combined proteasome cleavage and MHC ligand prediction*.
  - **NetChop Prediction Server**, *produces neural network predictions for cleavage sites of the human proteasome*.
  - **PAPoC**, Prediction Algorithm for Proteasomal Cleavages

• **Prediction of MHC I binding peptides**
  - **CombiPRED**, a matrix-based MHC Class I prediction tool that combines MHC allele matrices from three MHC prediction programs - nHLAPred, BIMAS and SYFPEITHI, part of a pipeline of tools for vaccine design applied to bacteria.
  - **CTL.Pred**, a SVM and ANN based CTL epitope prediction.
  - **HLA Peptide Binding Predictions**, Bioinformatics and Molecular Analysis Section (BIMAS), *a method based on profiles and predicted half-time of dissociation of a given MHC class I - peptide complex*.
  - **MHCPred**, quantitative prediction of peptide-MHC binding.
  - **NetMHC**, *prediction of peptide binding to HLA alleles using artificial neural networks (ANNs) and hidden Markov models (HMMs)*.
  - **nHLAPred**, a neural network based MHC Class-I Binding Peptide Prediction Server.
  - **PREDEP**, MHC Class I epitope prediction (see Resources).
  - **ProPred-I**, the Promiscuous MHC Class-I Binding Peptide Prediction Server.
  - **RANKPEP**, *prediction of binding peptides to MHC (class I and class II) molecules*.
  - **SMM**, *prediction of high affinity HLA-A2-binding peptides, based on an matrix-based algorithm*.
  - **SNEP**, single nucleotide polymorphism (SNP)-derived Epitope Prediction program for minor histocompatibility antigens (miHAs), at the Department of Immunology, University of Tuebingen, Germany.
  - **SVMHC**, *a machine learning method based on the support vector machine package SVM-light*.
  - **SYFPEITHI T cell epitope prediction**, *a method based on profiles*.

• **Prediction of MHC II binding peptides**
  - **EPIBPREDICT**, *prediction of HLA-class II restricted T cell epitopes and ligands*.
  - **ProPred**, MHC Class-II Binding Peptide Prediction Server, *uses quantitative matrices*.
  - **RANKPEP**, *prediction of binding peptides to MHC (class I and class II) molecules*.
  - **SNEP**, single nucleotide polymorphism (SNP)-derived Epitope Prediction program for minor histocompatibility antigens (miHAs), at the Department of Immunology, University of Tuebingen, Germany.

Go here to see this list of servers with links:
What are polymorphisms?

- Genetic differences between individuals in a population.
- Changes related to alleles
  - **Single nucleotide polymorphisms** (one base substitution)
    - Noncoding
    - Coding
      - synonymous -- same amino acid, different codon
      - non-synonymous
        - missense -- change in amino acid
        - nonsense -- stop codon
  - Frame-shifts
    - One or more base insertion/deletion

SNPs = single nucleotide polymorphisms
How To: View all SNPs associated with a gene

Starting with...

a gene name

1. Search the Gene database with the gene name. If you know the gene symbol and species, enter them as follows: tpo[sym] AND human[orgn]
2. Click on the desired gene.
3. In the list of Links on the right, click "SNP:GeneView". If the link is not present, no SNPs are currently linked to this gene.

a nucleotide or protein accession number (e.g. NM_001126)

1. Search the Nucleotide or Protein database with the accession number.
2. In the Links menu in the upper right, click on "GeneView in dbSNP". If the link is not present, click on the "Gene" link in the same menu and continue at step 3 above under "a gene name".

a nucleotide sequence

1. Go to the BLAST home page and click "nucleotide blast" under Basic BLAST.
2. Paste the sequence in the query box.
3. Enter the name of the organism of interest in the "Organism" box. Click the BLAST button.
4. Click on the desired sequence from the results.
5. Continue at step 2 under "a nucleotide or protein accession number" above.

a protein sequence

1. Go to the BLAST home page and click "protein blast" under Basic BLAST.
2. Paste the sequence in the query box.
3. Enter the name of the organism of interest in the "Organism" box. Click the BLAST button.
4. Click on the desired sequence from the results.
5. Continue at step 2 under "a nucleotide or protein accession number" above.
You’re going to need this....

<table>
<thead>
<tr>
<th>1st position</th>
<th>2nd position</th>
<th>3rd position</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>U</td>
<td>U</td>
</tr>
<tr>
<td>Phe Ser Tyr Cys</td>
<td>Phe Ser Tyr Cys</td>
<td>U C A G</td>
</tr>
<tr>
<td>Leu Ser STOP STOP</td>
<td>Leu Ser STOP Trp</td>
<td>U C A G</td>
</tr>
<tr>
<td>Leu Pro His Arg</td>
<td>Leu Pro His Arg</td>
<td>U C A G</td>
</tr>
<tr>
<td>Leu Pro Gin Arg</td>
<td>Leu Pro Gin Arg</td>
<td>U C A G</td>
</tr>
<tr>
<td>Ile Thr Asn Ser</td>
<td>Ile Thr Asn Ser</td>
<td>U C A G</td>
</tr>
<tr>
<td>Ile Thr Lys Arg</td>
<td>Ile Thr Lys Arg</td>
<td>U C A G</td>
</tr>
<tr>
<td>Val Ala Asp Gly</td>
<td>Val Ala Asp Gly</td>
<td>U C A G</td>
</tr>
<tr>
<td>Val Ala Glu Gly</td>
<td>Val Ala Glu Gly</td>
<td>U C A G</td>
</tr>
<tr>
<td>Val Ala Glu Gly</td>
<td>Val Ala Glu Gly</td>
<td>U C A G</td>
</tr>
</tbody>
</table>
GeneView page from dbSNP link

GeneView via analysis of contig annotation: **TAP1** transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)

- View more variation on this gene (click to hide).
- Include clinically associated: ☐ in gene region ☐ cSNP ☐ has frequency ☐ double hit ☑

<table>
<thead>
<tr>
<th>Assembly reference</th>
<th>SNP to Chr</th>
<th>Chr</th>
<th>Chr position</th>
<th>Contig</th>
<th>Contig position</th>
<th>Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>6</td>
<td>32922953</td>
<td>NT_007592.14</td>
<td>23673225</td>
<td>T</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Function</th>
<th>mRNA to Chr</th>
<th>mRNA Accession</th>
<th>mRNA Position</th>
<th>mRNA Allele change</th>
<th>Protein Accession</th>
<th>Protein Position</th>
<th>Protein Residue change</th>
</tr>
</thead>
<tbody>
<tr>
<td>missense</td>
<td>-</td>
<td>NM_000593.5</td>
<td>2245</td>
<td>CAC -&gt; GCG</td>
<td>NP_000594.2</td>
<td>697</td>
<td>D [Asp] -&gt; G [Gly]</td>
</tr>
</tbody>
</table>

(Open sequence viewer in a new window.)

Goes to chromosome 6 navigator window.
The International HapMap Project

Basic Concepts

Parent 1

Parent 2

A B
a b
A B
a b
A B
a b
A B
a b
A B
a b
A B
a b
A B
a b

Parent 1 and Parent 2 cross to form offspring with the following genotypes:

A B
a b
A B
a b
A B
a b
A B
a b
A B
a b
A B
a b
A B
a b

High LD -> No Recombination
(r^2 = 1) SNP1 “tags” SNP2

Low LD -> Recombination
Many possibilities
Tagging SNPs, tSNPs

• SNPs that are highly correlated are redundant information

• tSNPs are selected as the minimal non-redundant set of SNPs in a population, such that the genotypes can be reconstructed from the tSNPs.

• tSNPs allow genotyping with fewer steps
  – PCR amplification experiments determine which base is present.

• Block based tagging

Block based tagging requires that haplotype "blocks" first be inferred. In the majority of cases when you are investigating association within a candidate gene you are likely to start off with a large number of potential SNPs to choose from, and using various measures of linkage disequilibrium and inferred haplotypes it is possible to define 'haplotype blocks' of markers that are in strong LD with each other, but not with those in other blocks. The exact definition of a haplotype block is open to interpretation, and there are a number of different methods for choosing your haplotype blocks (Gabriel et al 2002, )
Uses of SNPs

• Personalized medicine
  – Sensitivity to
    – diseases
    – drugs
    – chemicals
    – pathogens
    – vaccines

• Livestock breeding

• Human migrations
What can we find out about a protein structure given its sequence?

<table>
<thead>
<tr>
<th>Does it have a homolog of known structure?</th>
<th>blastp</th>
</tr>
</thead>
</table>
If so, align it to the homolog and model it (take Bioinformatics 2 to learn how)
If not, you can still get:
  • secondary structure  psi-pred
  • transmembrane regions  TMHMM
  • coiled-coil regions  COILS
  • disordered regions  disopred
  • local structure  HMMSTR
Protein classification : CATH

- Class
- Architecture
- Topology
- Homology

http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html
Structural hierarchy of proteins

- Primary structure
- Secondary structure
- Local structure
- super-secondary structure
- domains, folds
- Global, multi-domain (tertiary structure)
- Quaternary structure
Secondary structure

Alpha helix  Right-handed
            3.6 residues/turn
            i->i+4 H-bonds

Overall dipole N+->C-
3 types of Alpha helix

- non-polar
- amphipathic
- polar

Two ways to display position of sidechain on a helix.

For amphipathic and non-polar, sidechains line up in a favorable way.
beta-strand

Antiparallel:

Parallel:
Two sequence motifs for beta strand

Amphipathic

Found at the edges of a sheet, or when one side of the sheet is exposed to solvent (i.e. 2-layer proteins).

Note preference for beta-branched aa’s: I,V,T

Hydrophobic

Found in the buried middle strands of sheets in 3-layer proteins.
Local structure: beta hairpins

Two adjacent antiparallel beta strands = a beta hairpin

Shown are “tight turns”, 2 residues in the loop region (shaded). Hairpins can have as many as 20 residues in the loop region.
“Serine beta-hairpin” (also called an “alpha turn”). A specific pattern (DPESG) forms an alpha-helical turn 4-residues long.

“Extended Type-1 hairpin”. A type-1 “tight turn” has only 2 residues in the turn. This motif, more common than the tight turn, has an additional Pro or polar sidechain. Pattern: PDG.
“Diverging turns” have a Type-2 beta turn and two strands that do not pair. The consensus sequence pattern is PDG. The residue before G can be anything polar, but not a D or an N.
Proline helix C-cap motif

Sequence pattern=
...nppnnppp[HNYF]P[DE]n

“n”=non-polar
“p”=polar
[...] = alternative aa’s

Pro blocks helix
D or E stabilizes tight turn

Locations of non-polar (magenta) and polar (green) sidechains

note: hydrophobic cluster

structural variability
secondary structure alphabet

3D protein coordinates may be converted to a 1D secondary structure representation using DSSP or STRIDE

DSSP= Database of Secondary Structure in Proteins

Both programs use hydrogen bonding patterns (see next slide)
DSSP symbols

\( H \) = helix backbone angles (-50, -60) and H-bonding pattern (i-> i+4)

\( E \) = extended strand backbone angles (-120, +120) with beta-sheet H-bonds (parallel/anti-parallel are not distinguished)

\( S \) = beta-bridge (isolated backbone H-bonds)

\( T \) = beta-turn (specific sets of angles and 1 i->i+3 H-bond)

\( G \) = 3-10 helix or turn (i,i+3 H-bonds)

\( I \) = Pi-helix (i,i+5 Hbonds) (rare!)

Accuracy of 3-state predictions

Q3-score = % of 3-state symbols that are correct

Measured on a "test set"

Test set == An independent set of cases (protein) that were not used to train, or in any way derive, the method being tested.

Best methods:

PHD (Burkhard Rost) -- 72-74% Q3

HMMSTR (Bystroff) -- 74-75% Q3

Psi-pred (David T. Jones) -- 76-78% Q3
PSI-pred-- a secondary structure predictor

http://bioinf.cs.ucl.ac.uk/psipred/psiform.html

PSI-PRED (Jones et al.) is currently the best server for secondary structure prediction, according to CASP results.
Psi-Pred: A neural network

Prediction (each position) is the state with the greatest sum of weights.
Psi-pred: a neural net

(Step 1) Run PSI-Blast --> output sequence profile

(Step 2) 15-residue sliding window = 315 weights, multiplied by hidden weights in 1st neural net. Output is 3 weights (1 weight for each state H, E or L) per position.

(Step 3) 45 input weights, multiplied by weights in 2nd neural network, summed. Output is final 3-state prediction.
1. Multiple sequence alignment

2. Sequence weights from phylo.tree

3. Sum weights of each amino acid.

4. Sequence profile, probabilities of 20 amino acids

\[ P_{ij} = \frac{\sum_{k=\text{seqs}} w_k \delta(s_{kj} = \text{aa}_i)}{\sum_{k=\text{seqs}} w_k} \]

- Red = high prob ratio (LLR > 1)
- Green = background prob ratio (LLR ≈ 0)
- Blue = low prob ratio (LLR < -1)

Reminder:
1. Multiple sequence alignment
2. Sequence weights from phylo.tree
3. Sum weights of each amino acid.
4. Sequence profile, probabilities of 20 amino acids
Psi-Pred: Training the neural network (NN)

weights are found that minimize errors

Protein database provides both input and output

NN output is compared with the true SS. Weights are "back propagated."

True SS: EEEE_SS_EEEE_GGT__EE_E______________________TT
Prediction: EEEE_LLLL_________________LLLLEEEE__LL
Errors: 00000000111110000010100000000000000000100
What can you do with a secondary structure prediction?

(1) Find out if a homolog of unknown structure is **missing** any of the SS (secondary structure) units, i.e. a helix or a strand.

(2) Find out whether a helix or strand is **extended/shortened** in the homolog.

(3) Model a large insertion or terminal domain (possibly).

(4) Test remote homology (compare 3-state pred to known SS when sequence homology is very low, i.e. < 20%).

Secondary structure-based alignment doesn't work!
Other methods for secondary structure prediction

- GOR
- Chou Fasman
- PHD/PROF
- ZPRED
- PREDATOR
Why does it work?

Proteins fold via a “2-state” model: folded <=> unfolded.

No intermediates are observed. It’s all-or-none. Structure depends on the entire sequence!

If secondary structure depends on the entire sequence, then why is a 15-residue window enough to predict SS in ≈75-80% of cases?

Fast folding. Early folding events eliminate alternative pathways.