Bioinformatics I -- 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 ellicit the cellular immune response, that is, activated T-cell clones.

B-cell epitopes

- B-cell epitopes are peptides or other bio-molecules that bind specifically to antibodies.

B and T cells recognize different epitopes of the same protein

T-cell epitope

Denatured antigen

Linear (often) peptide 8-37 aa

Internal (often)

<u>Binding to T cell receptor:</u>

K_d 10^{-5} – 10^{-7} M (low affinity)

Slow on-rate, slow off-rate (once bound, peptide may stay associated for hours to many days)

B-cell epitope

Native or denatured (rare) antigen

Sequential (continuous) or conformational (discontinuous)

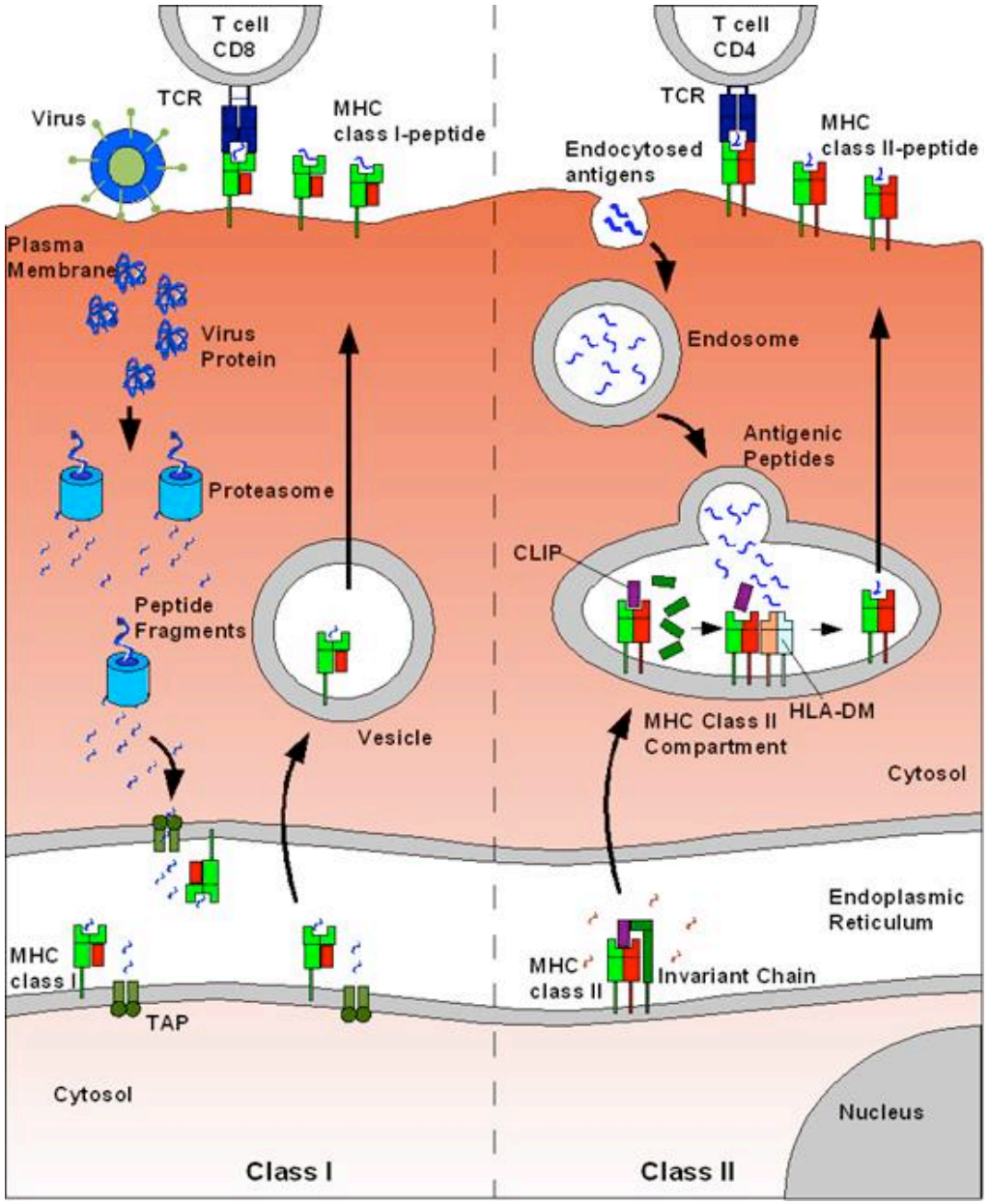
Accessible, hydrophilic, mobile, usually on the surface or could be exposed as a result of physicochemical change

<u>Binding to antibody:</u>

K_d 10^{-7} – 10^{-11} M (high affinity)

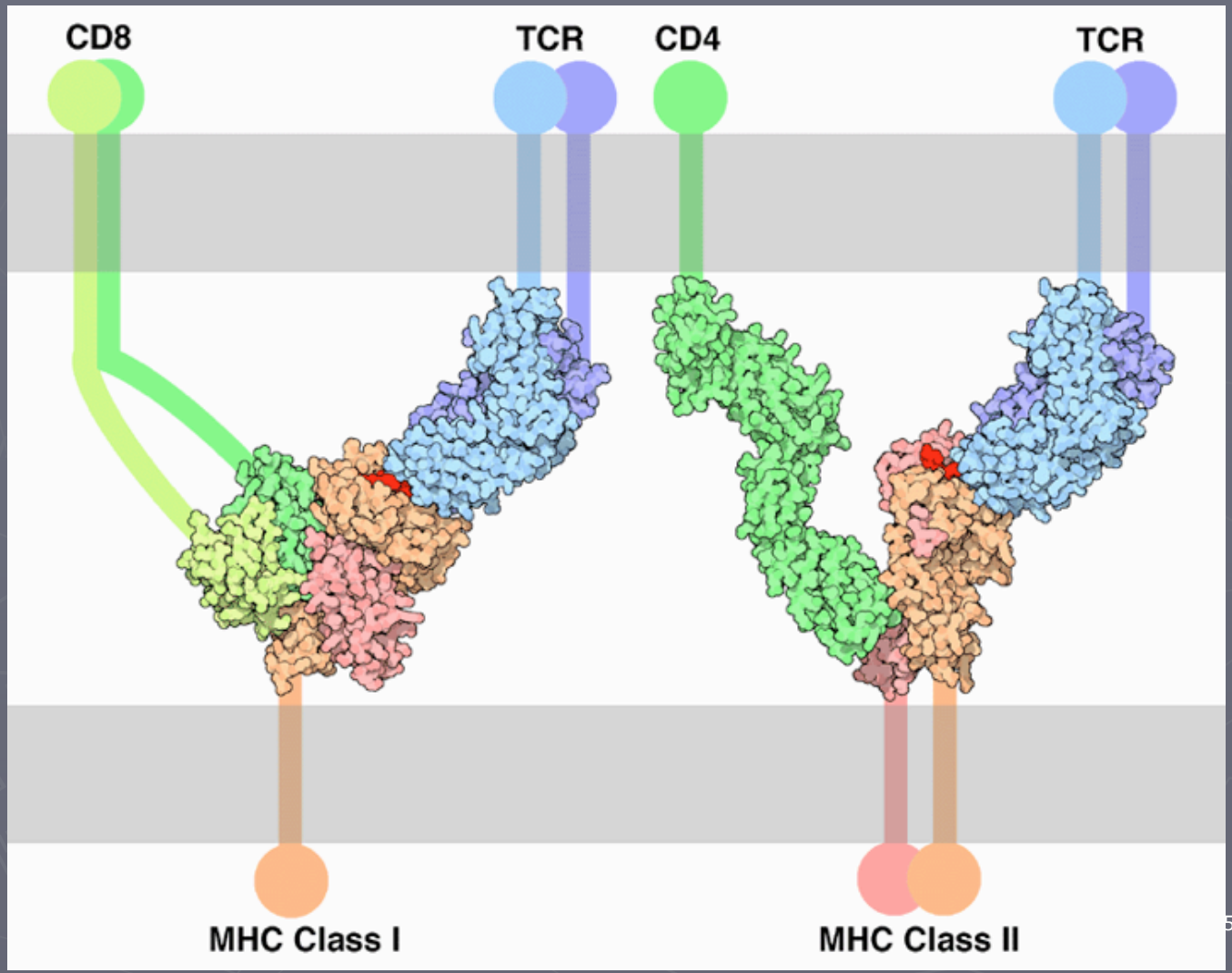
Rapid on-rate, variable off-rate

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.

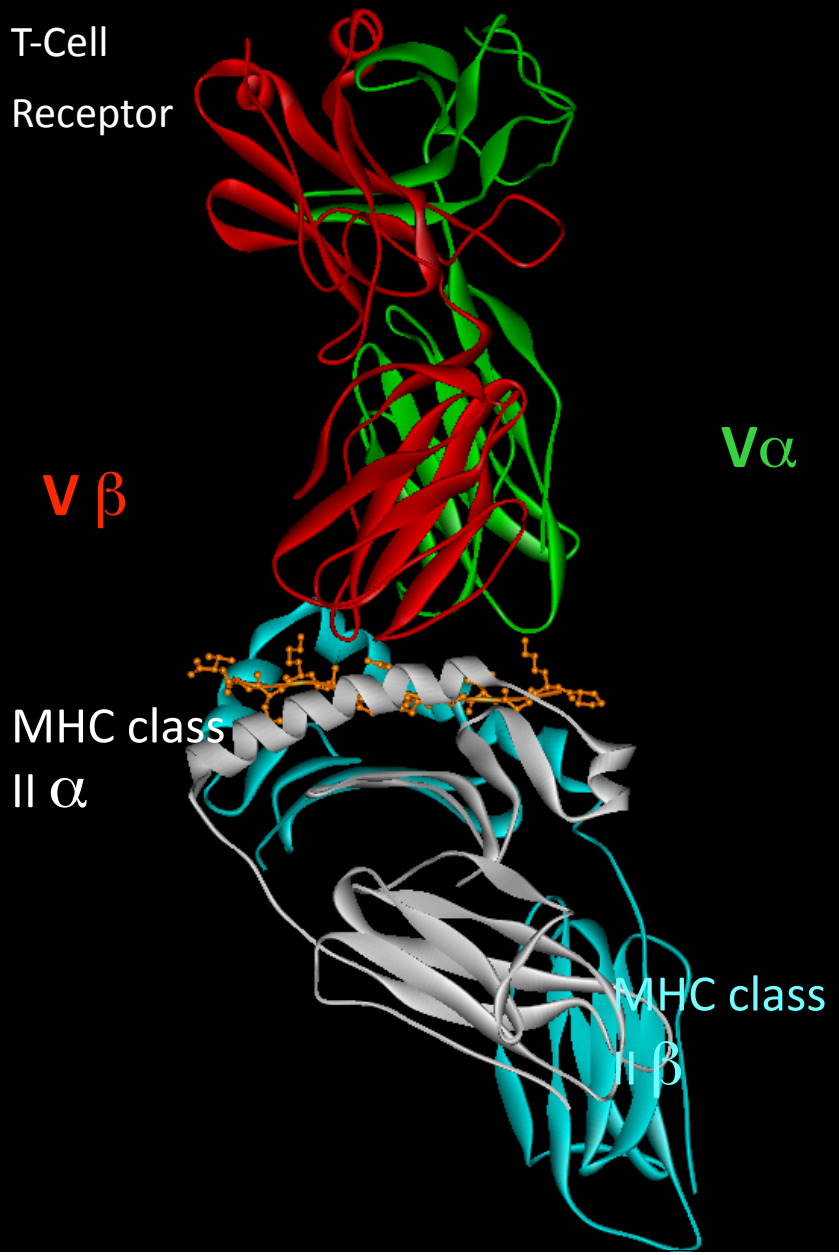


MHC class II gets peptides from **outside the cell** through phagocytosis. These are **exogenous**, usually. Blocked from peptides in the ER. Exposed to peptides from **lysosomes**, in a vesicle. recognized by **CD4** T-cells.

Modes of interaction with different T-cells.

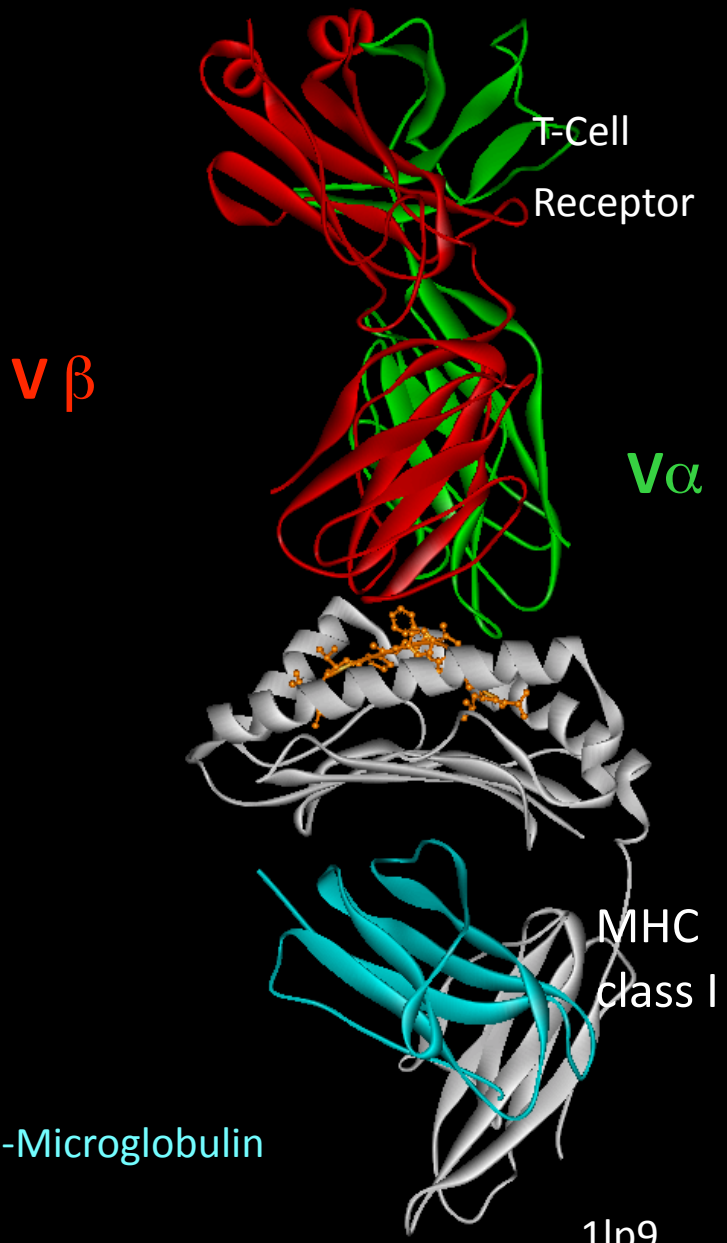


Complex Of A Human TCR, Influenza HA Antigen Peptide (PKYVKQNTLKLAT) and MHC Class II



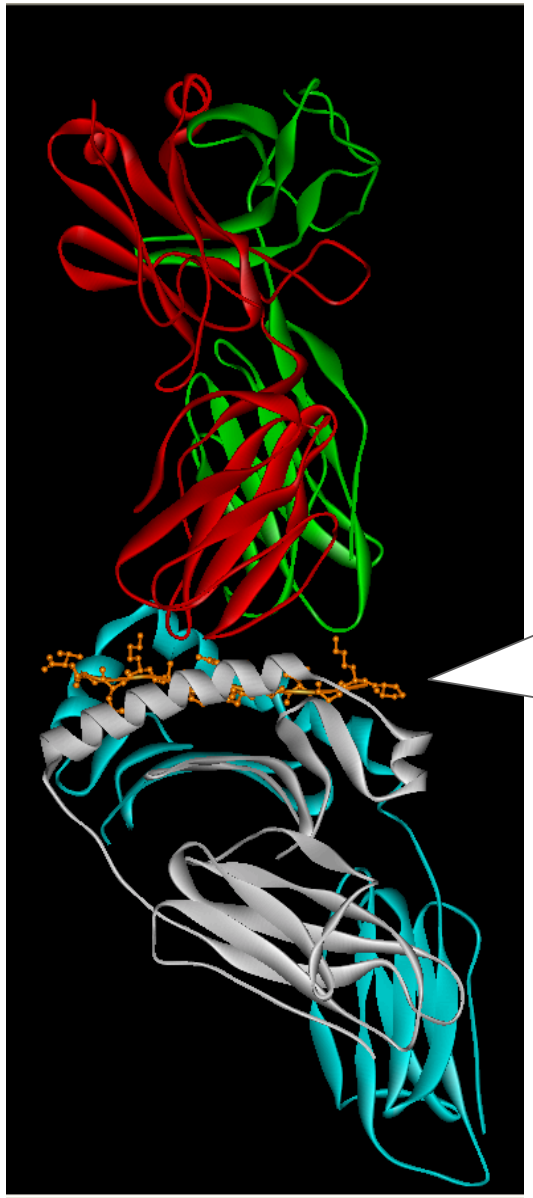
1fyt

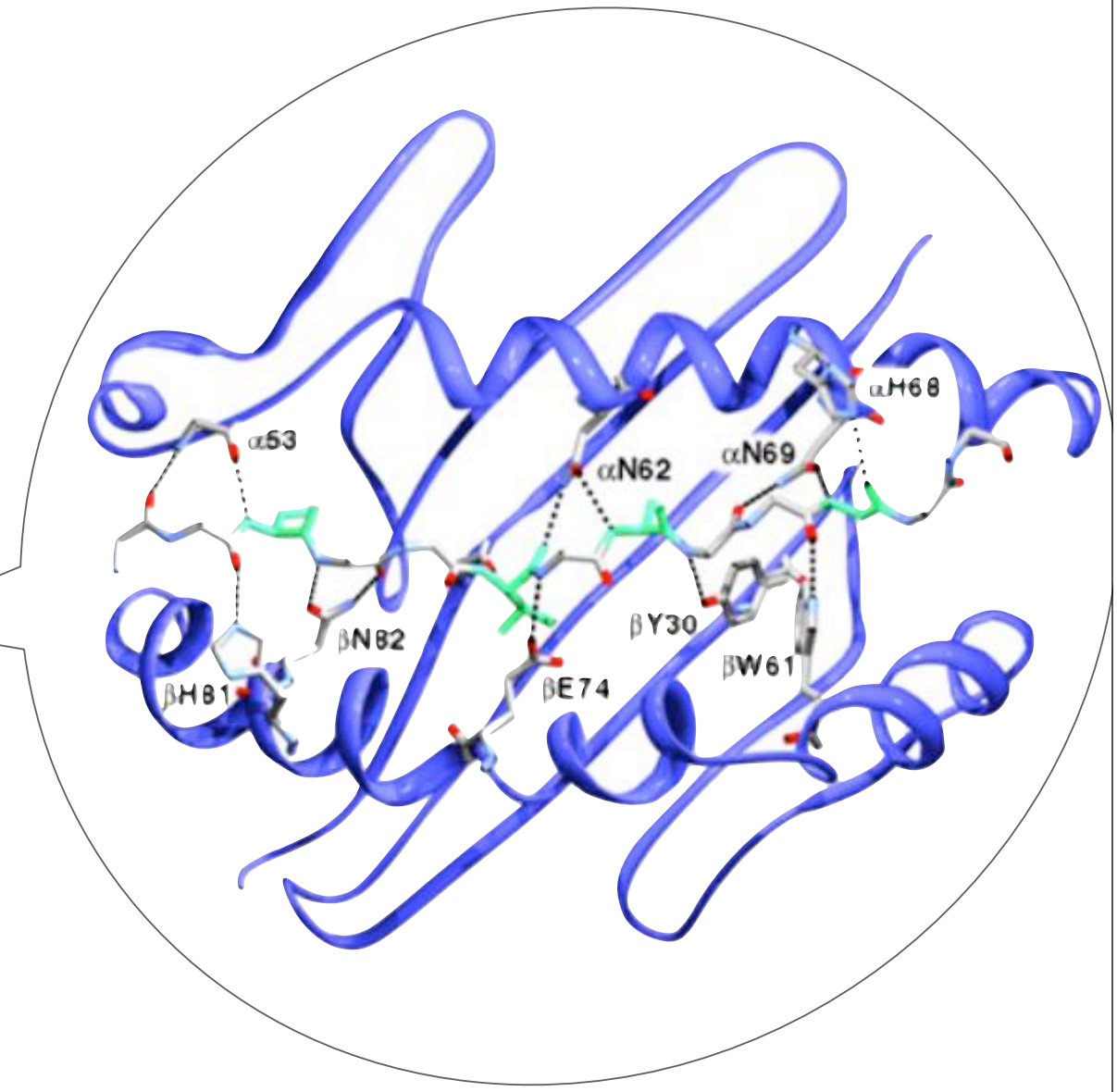
Xenoreactive Complex AHIII 12.2 TCR bound to P1049 (ALWGFFPVLS) /HLA-A2.1



1lp9

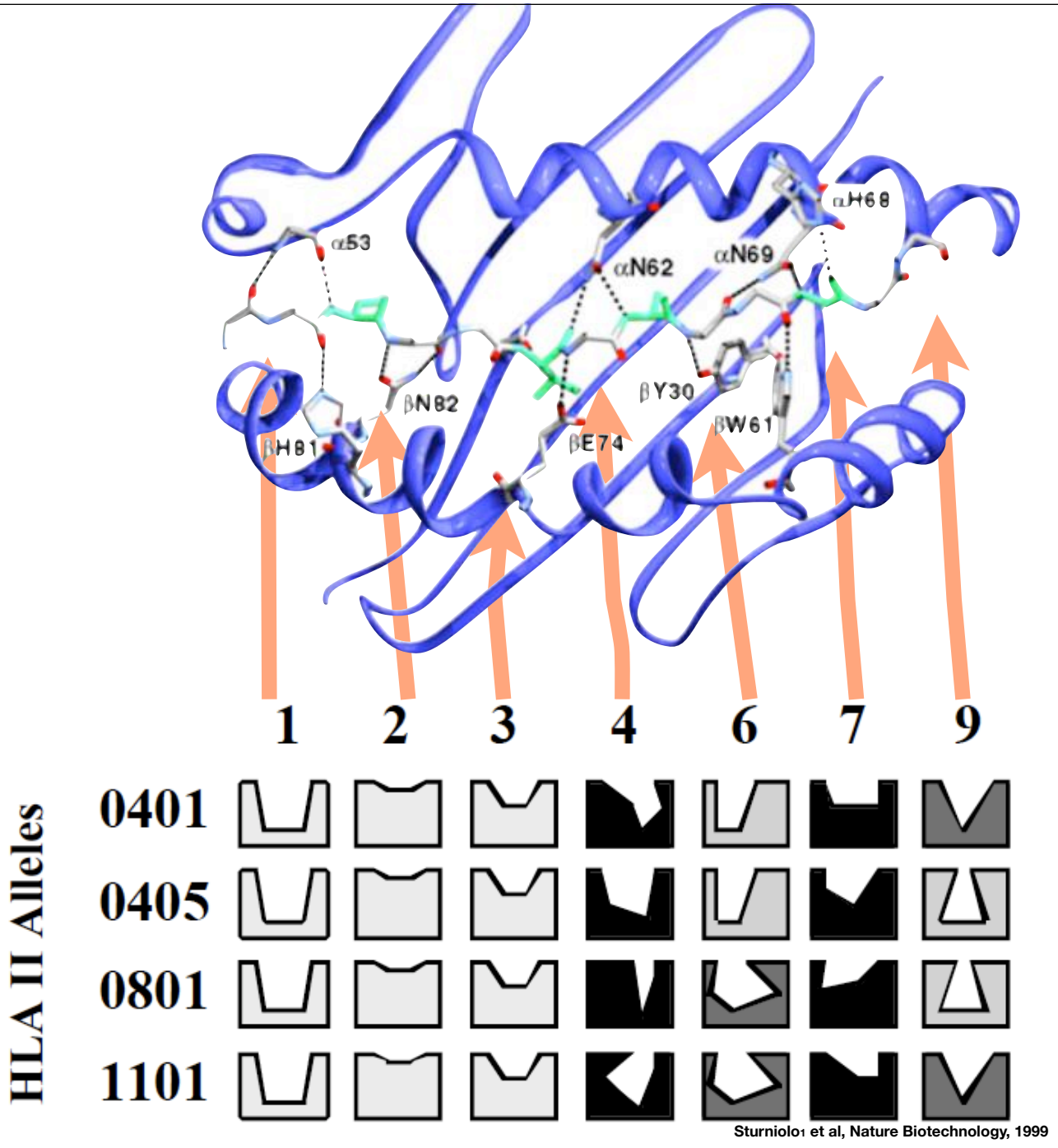
The strength of interaction with TCR depends on the strength of the peptide binding to MHC class II.





Long, open-ended pocket sandwiched between two alpha helices.

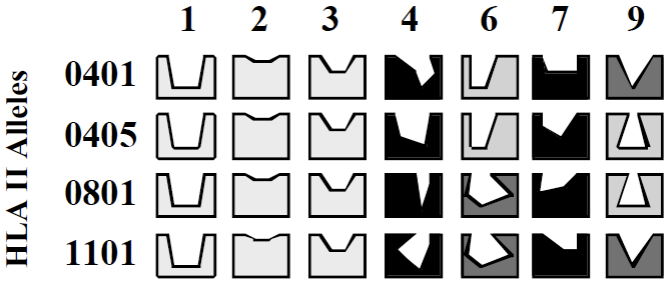
Different alleles of MHC II will have <u>different AAs in the binding pocket</u>.
The pocket can be divided up into 9 regions, each having one of several **shapes**, depending on the **allele**. Alleles are shuffled so that all combinations exist. In this view pocket shapes and specificities map directly to alleles.

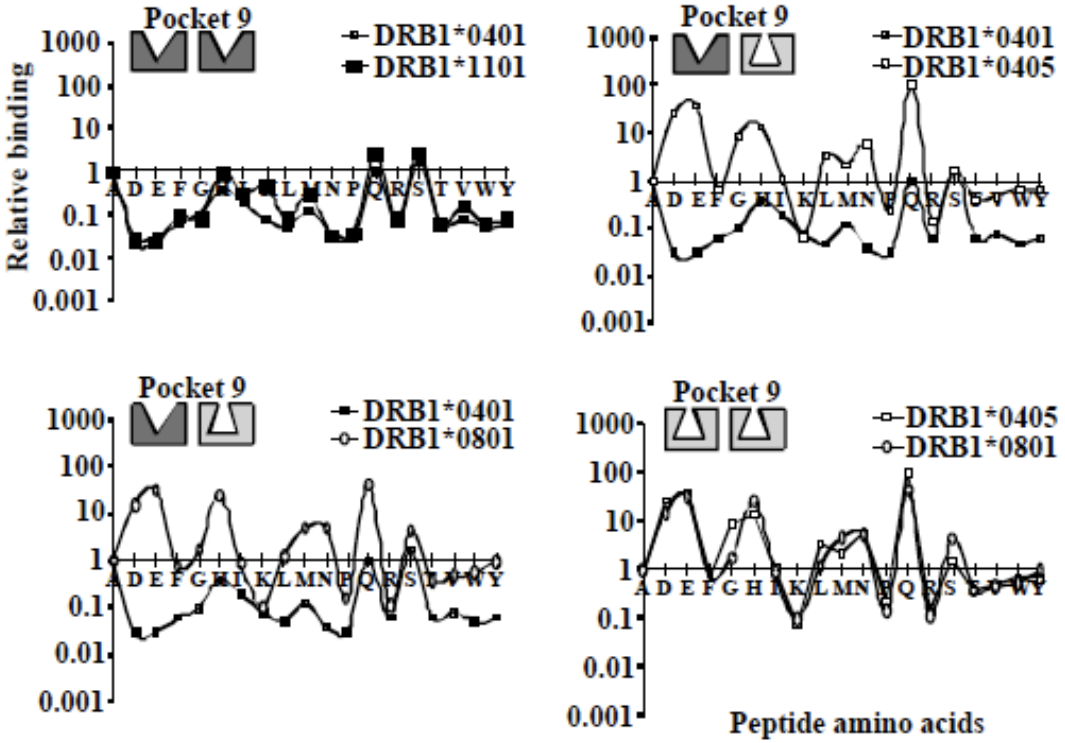


Sturniolo et al, Nature Biotechnology, 1999

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.

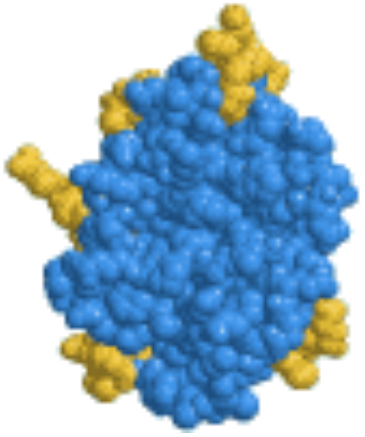




Sturniolo1 et al, Nature Biotechnology, 1999

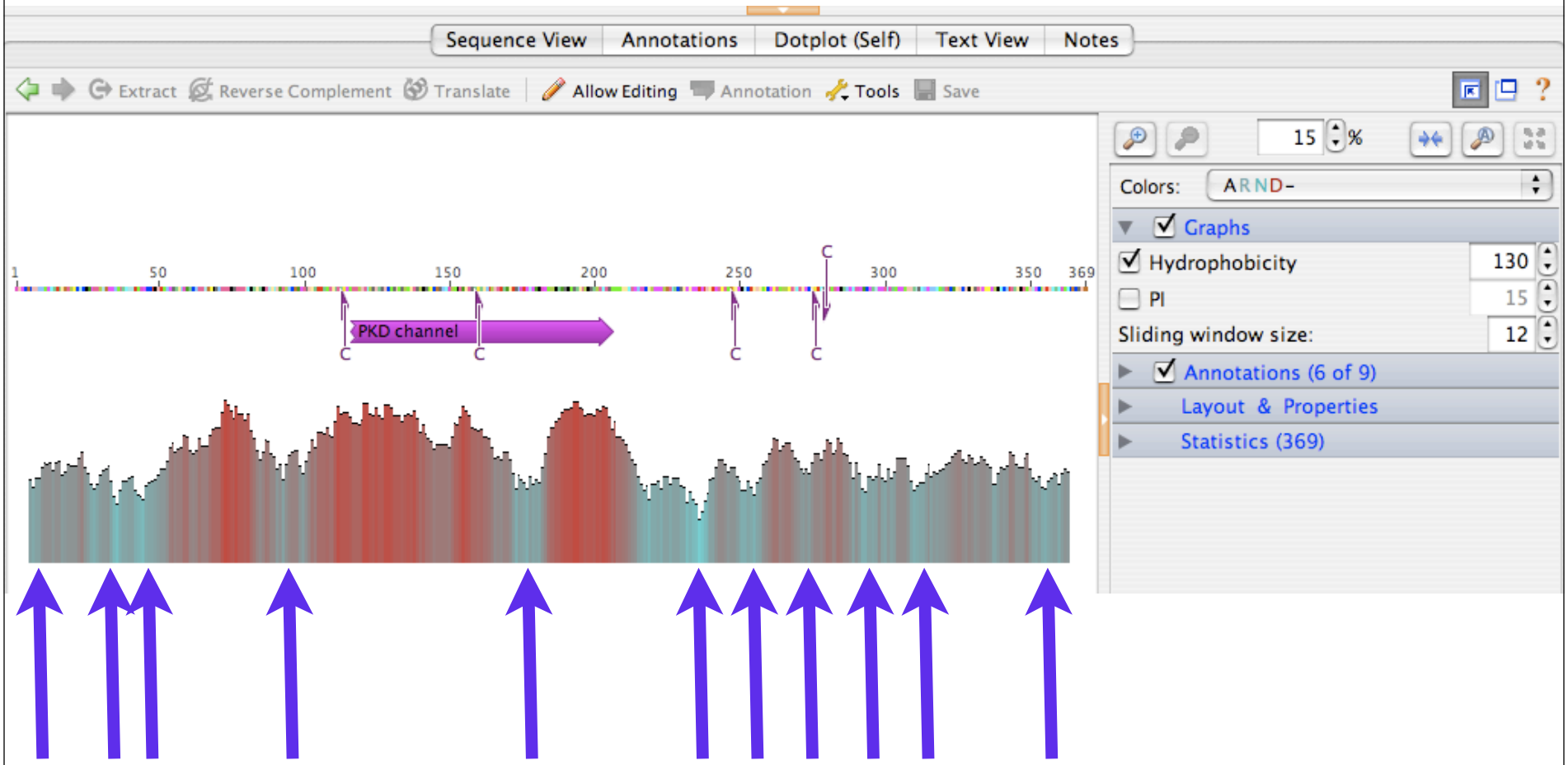
Properties of B-cell epitopes

- 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

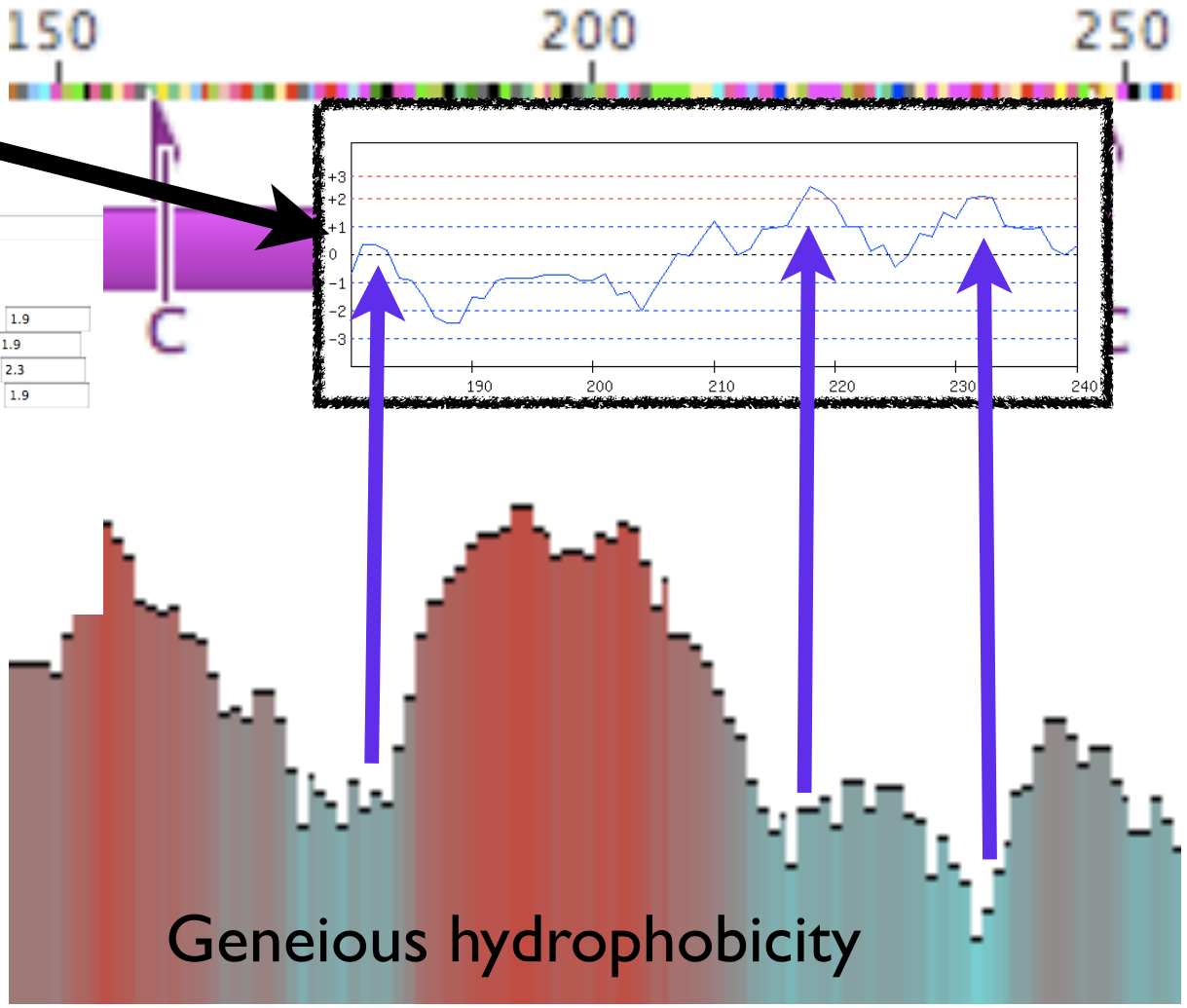
In Geneious: Graph hydrophobicity. Look for low hydrophobicity (hydrophilicity)



Potential B-cell epitopes

BcePRED compared to Geneious hydrophobicity plot

Part of BcePRED output



Or Submit sequences from file :

Threshold [-3 to 3] :

Hydrophilicity:	2	Flexibility:	1.9
Accessibility:	2	Turns:	1.9
Exposed Surface:	2.4	Polarity:	2.3
Antegenic Propensity:	1.8	Combined:	1.9

Select physico-chemical properties to use:

For multiple selection use Ctrl Key

- Hydrophilicity (Parker et al., Biochemistry, 25, 5425 (1986))
- Flexibility (Karplus et al., Naturwissenschaften, 72, 212 (1985))
- Accessibility (Emini et al., J.virol., 55, 836 (1985))
- Turns (Pellequer et al., Immunol.Lett., 36, 8 3(1993))

Clear fields | Submit sequence

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

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*.
  - PAProC, 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.
  - CTLPred, *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 (miHAgs), 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**
  - EPIPREDICT, *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 (miHAgs), at the Department of Immunology, University of Tuebingen, Germany.

Go here to see this list of servers with links:
http://imgt.cines.fr/textes/Immunoinformatics.html

SNPs = single nucleotide polymorphisms

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

NCBI : SNP database

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.

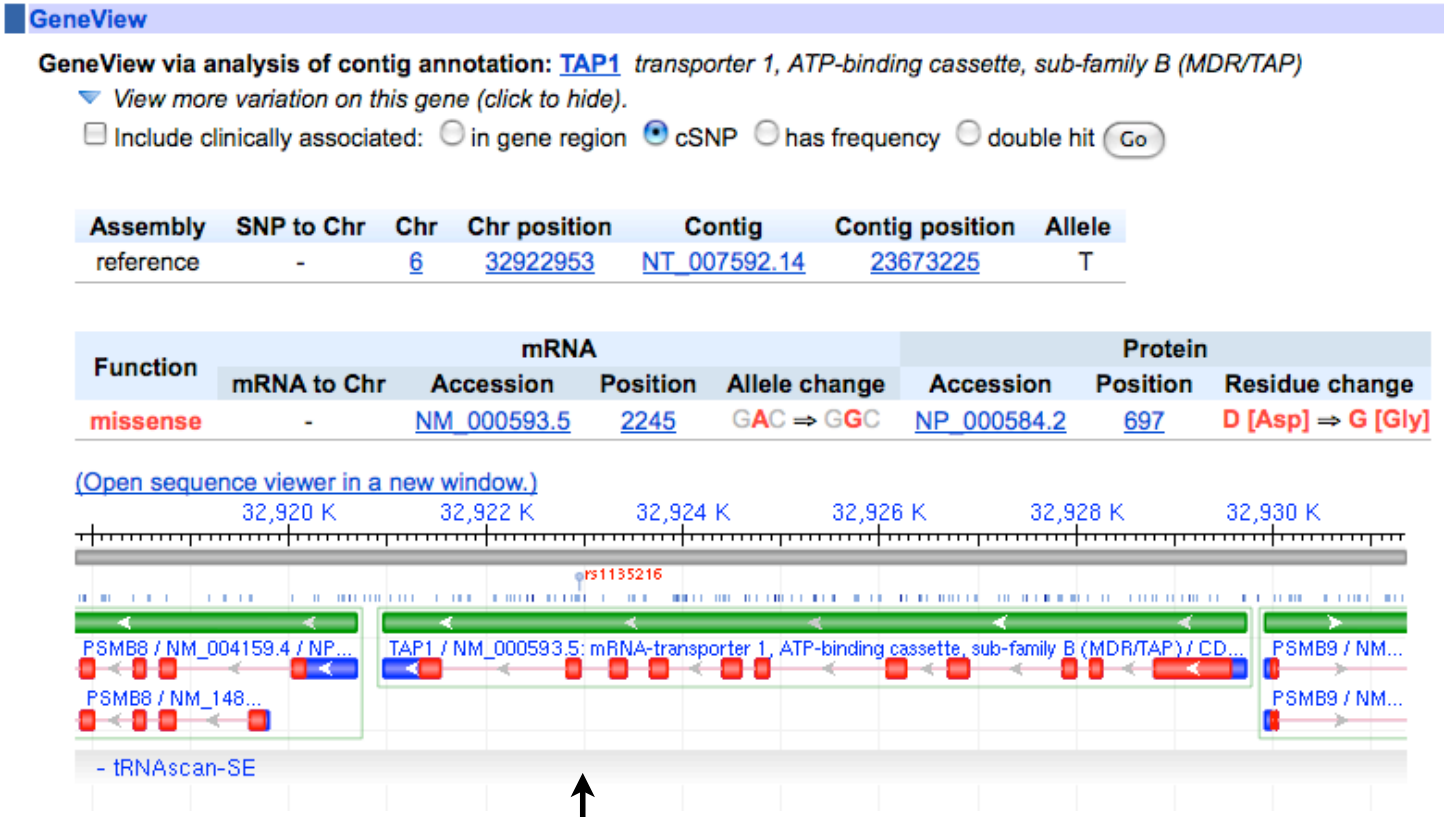
a protein sequence

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....

1st position	2nd position U	C	A	G	3rd position
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

GeneView page from dbSNP link

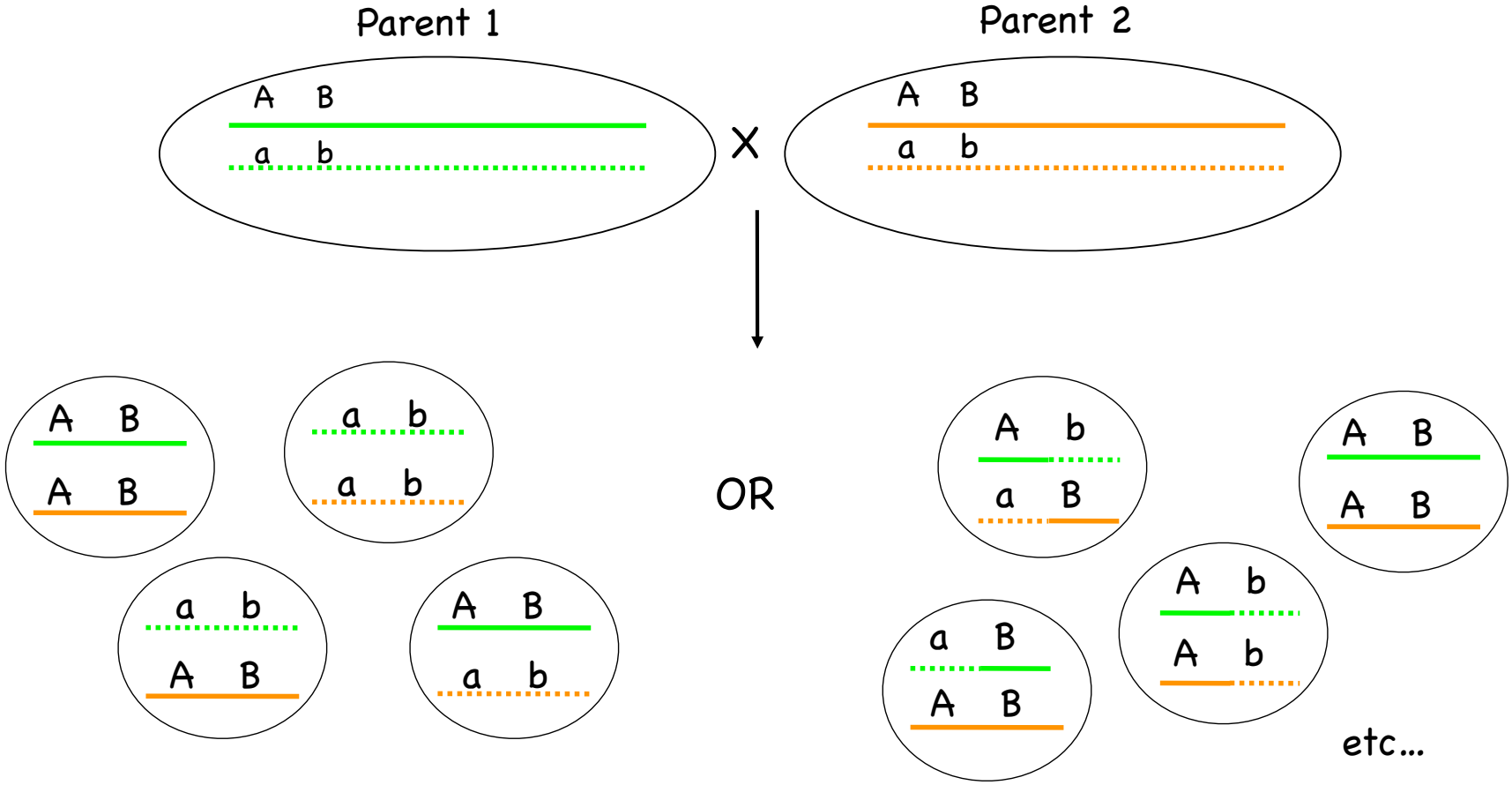


Goes to chromosome 6 navigator window.

The International HapMap Project

http://hapmap.ncbi.nlm.nih.gov/

Basic Concepts



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

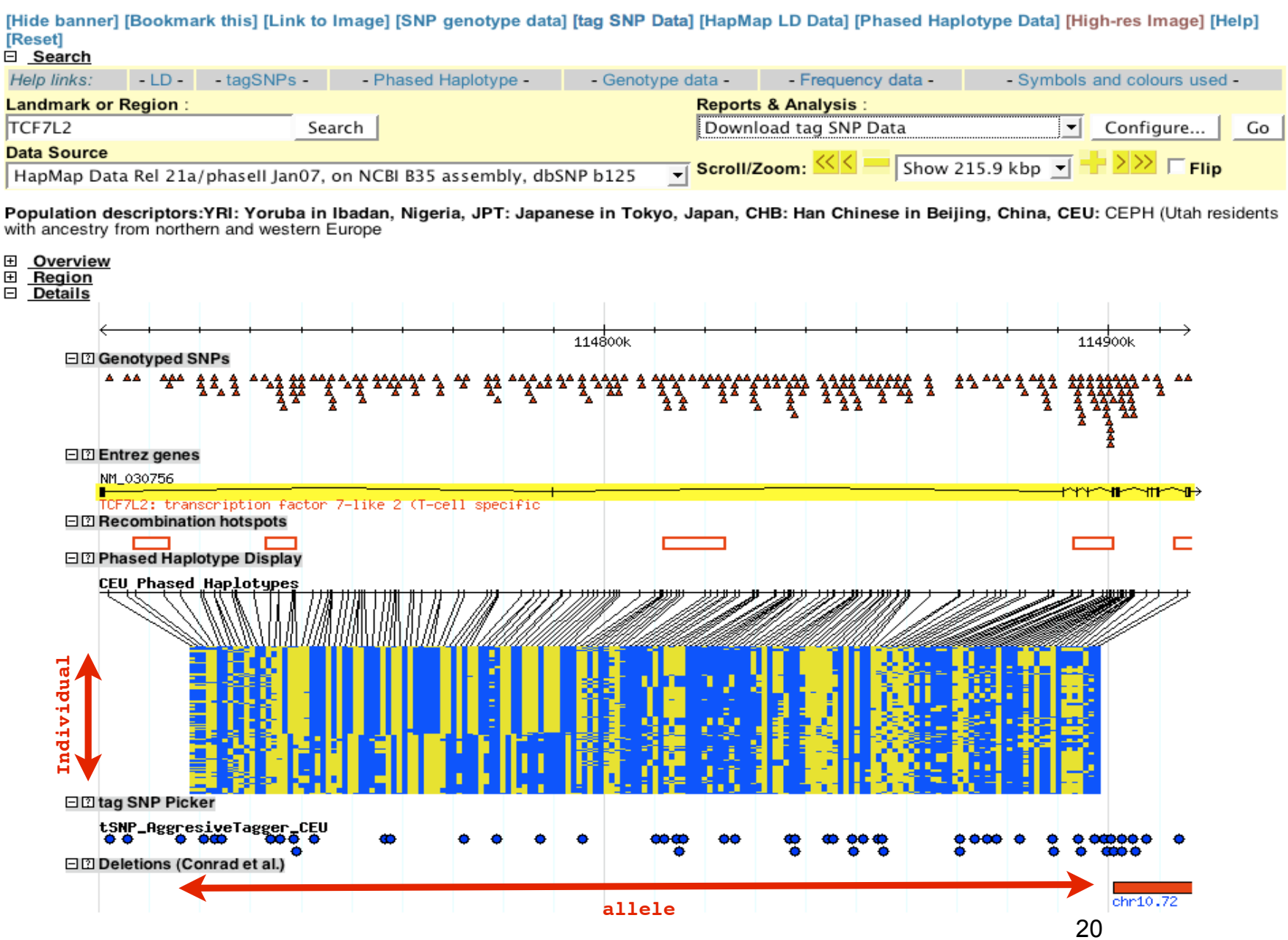
Low LD -> Recombination
Many possibilities

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 hapltype "blocks" first be infered. In the majority of cases when you are investigating associtionwithin a candidate gene you are likely to start of 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,)


[Hide banner] [Bookmark this] [Link to Image] [SNP genotype data] [tag SNP Data] [HapMap LD Data] [Phased Haplotype Data] [High-res Image] [Help] [Reset]
Search
Help links: - LD - - tagSNPs - - Phased Haplotype - - Genotype data - - Frequency data - - Symbols and colours used -
Landmark or Region :
TCF7L2
Search
Reports & Analysis :
Download tag SNP Data
Configure...
Go
Data Source
HapMap Data Rel 21a/phaseII Jan07, on NCBI B35 assembly, dbSNP b125
Scroll/Zoom:
Show 215.9 kbp
Flip
Population descriptors:YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe
Overview
Region
Details
114800k
114900k
Genotyped SNPs
Entrez genes
NM_030756
TCF7L2: transcription factor 7-like 2 (T-cell specific
Recombination hotspots
Phased Haplotype Display
CEU Phased Haplotypes
Individual
tag SNP Picker
tSNP_AggresiveTagger_CEU
Deletions (Conrad et al.)
allele
chr10.72

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?

Does it have a homolog of known structure? **blastp**

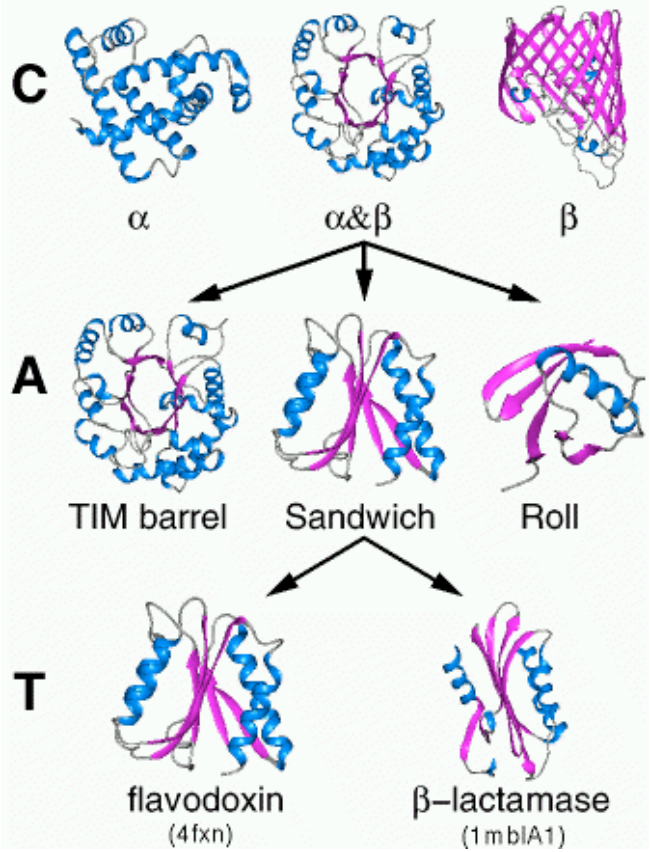
If so, align it to the homolog and model it (take Bioinformatics 2 to learn how)

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

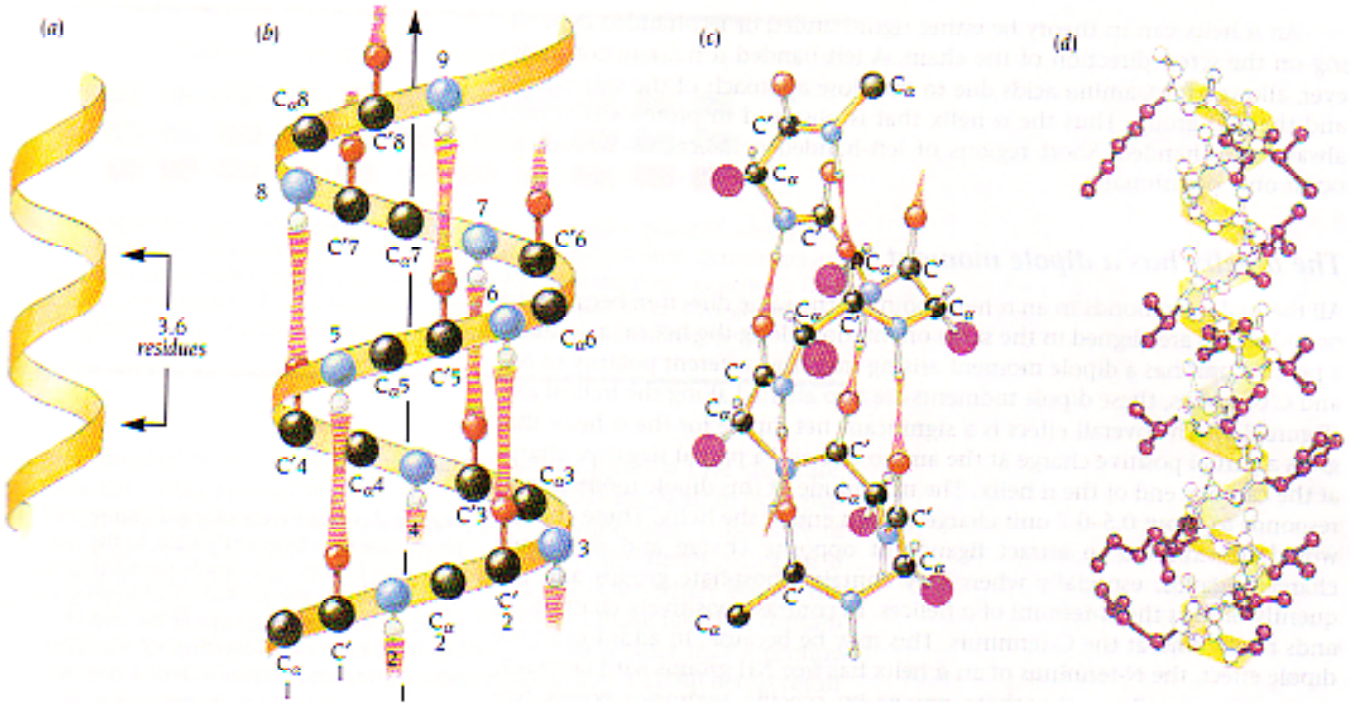


http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html

Structural heirarchy 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

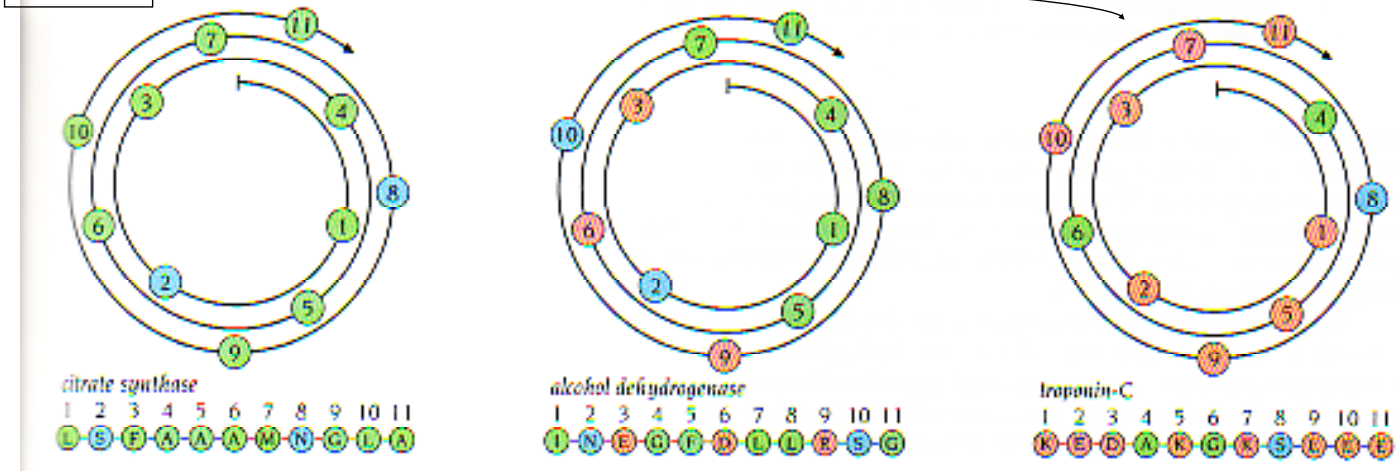
non-polar

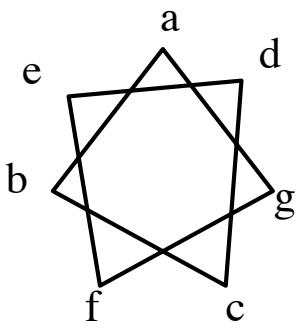
amphipathic

polar

Table 2.1 Amino acid sequences of three α helices

1. Leu - Ser - Phe - Ala - Ala - Ala - Met - Asn - Gly - Leu - Ala -
2. Ile - Asn - Glu - Gly - Phe - Asp - Leu - Leu - Arg - Ser - Gly -
3. Lys - Glu - Asp - Ala - Lys - Gly - Lys - Ser - Glu - Glu - Gln -



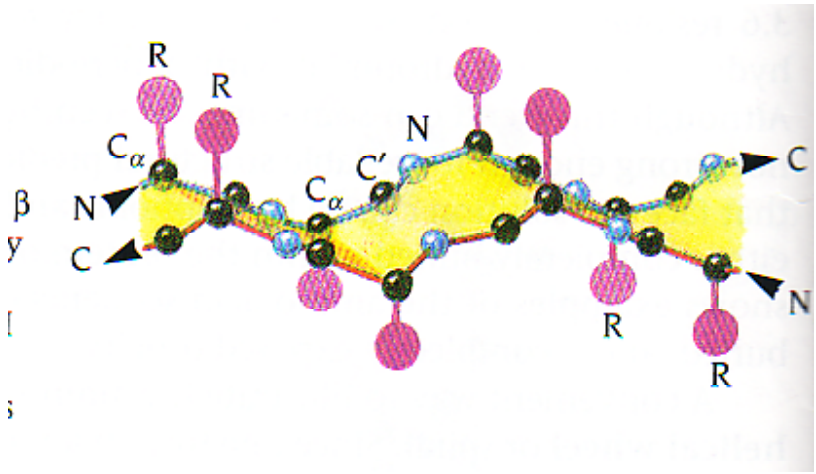


Two ways to display position of sidechain on a helix.

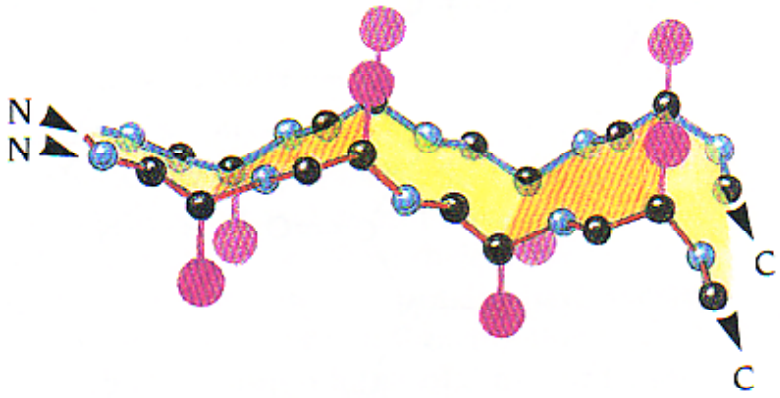
For amphipathic and non-polar, sidechains line up in a favorable way.

beta-strand

Antiparallel:



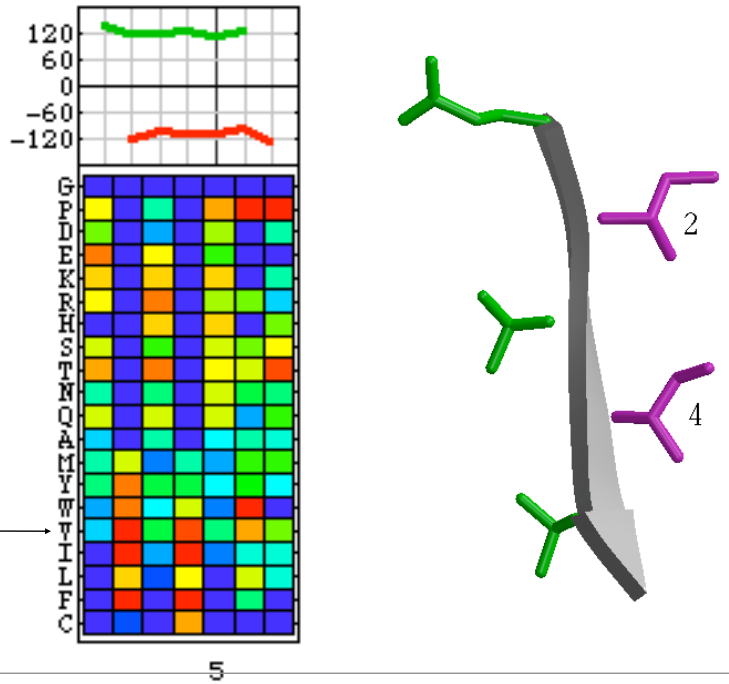
Parallel:



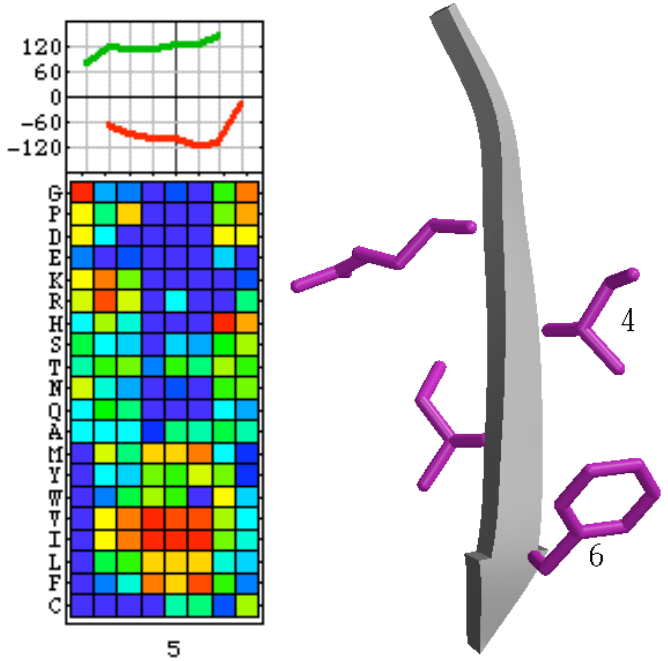
Two sequence motifs for beta strand

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

Amphipathic



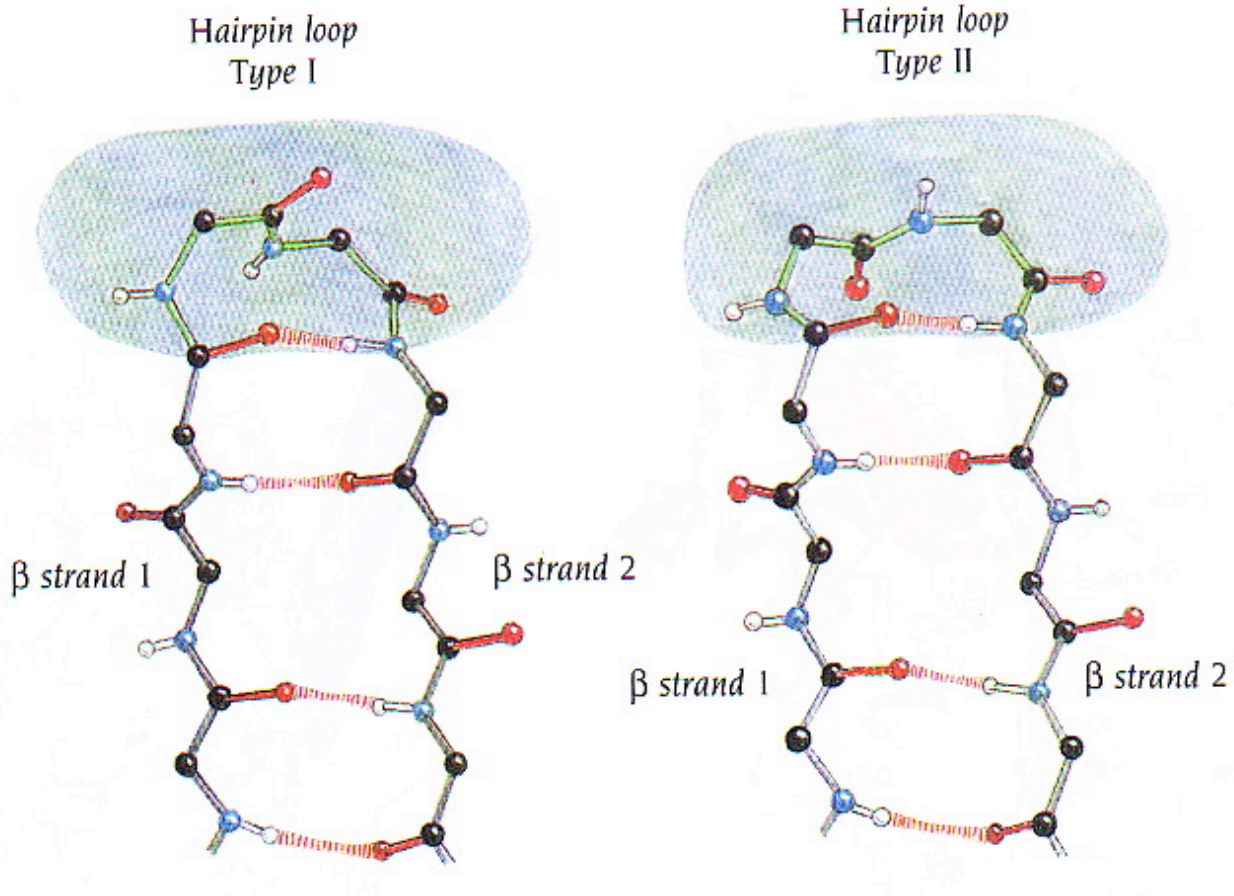
Hydrophobic



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

Found in the buried middle strands of sheets in 3-layer proteins.

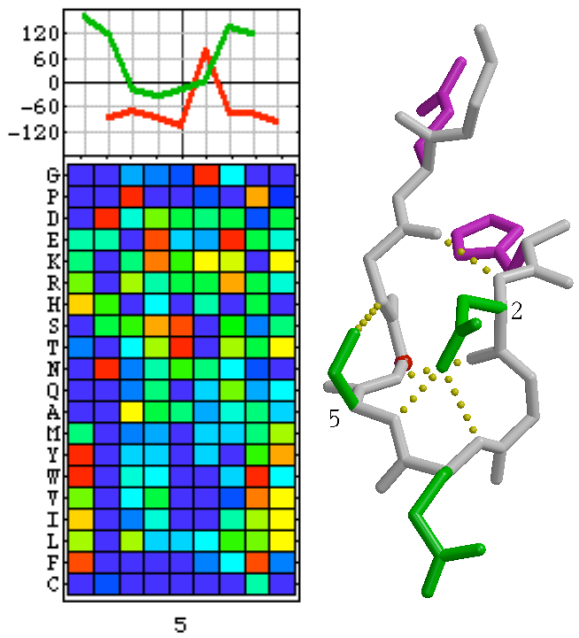
Local structure: beta hairpins



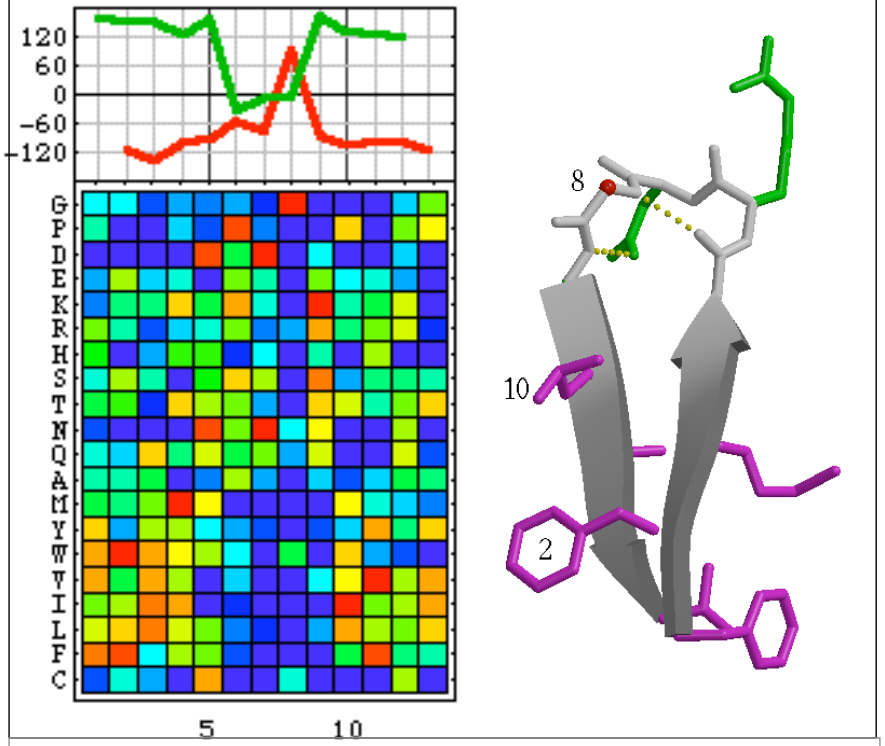
Two adjacent antiparallel beta strands = a beta hairpin

Shown are “tight turns”, 2 residues in the loop region (shaded).
Hairpins can have as many as 20 residues in the loop region.

hairpin sequence motifs

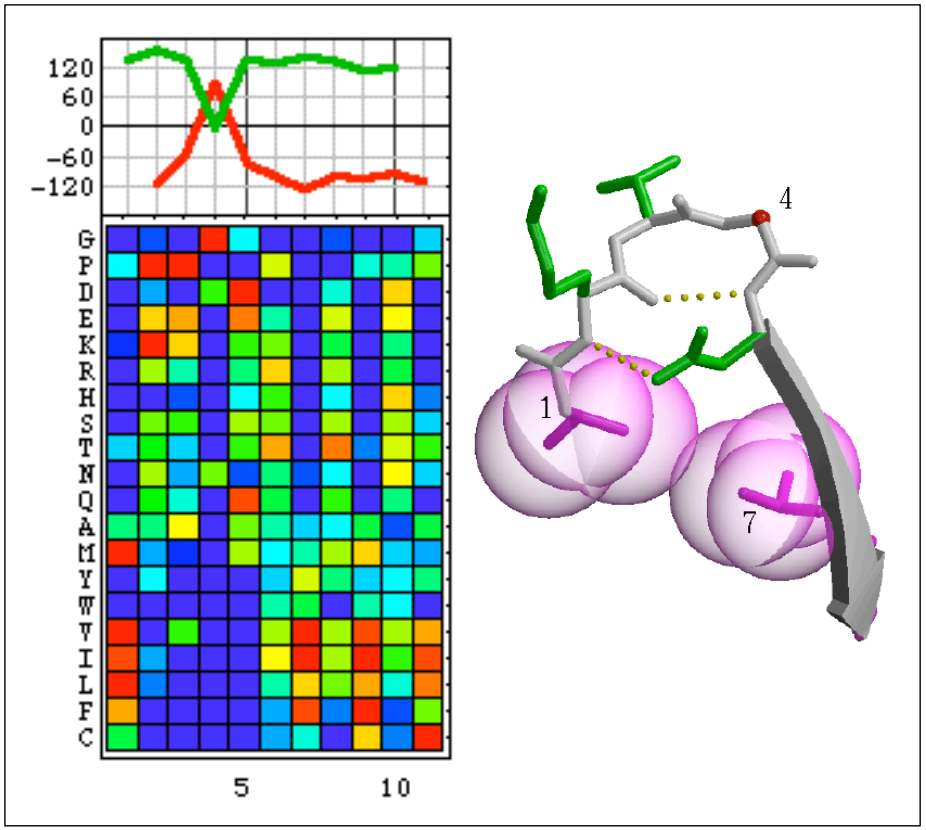


"Serine beta-hairpin" (also called an "alpha turn"). A specific pattern (DPESG) forms an alpha-helical turn 4-residues long.



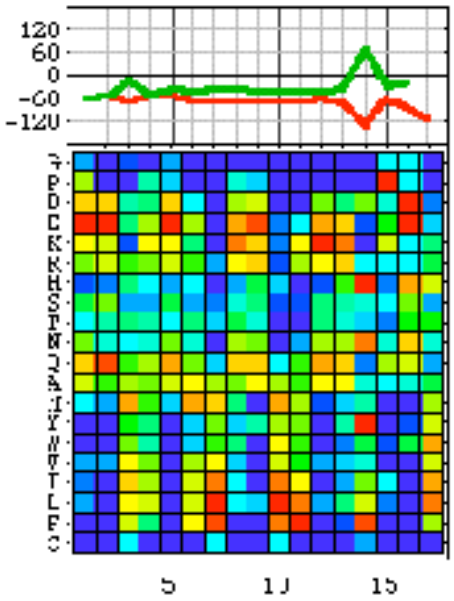
"Extended Type-1 hairpin". A type-1 "tight turn" has only 2 residues in the turn. This motif, more common than the tight turn, has an additional Pro or polar sidechain. Pattern: PDG.

diverging turn motif

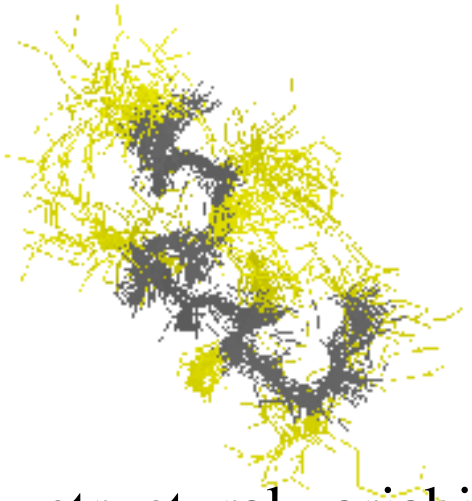


"Diverging turns" have a Type-2 beta turn and two strands that do not pair. The consensus sequence pattern is PDG. The residue before G can be anything polar, but not a D or an N.

Proline helix C-cap motif

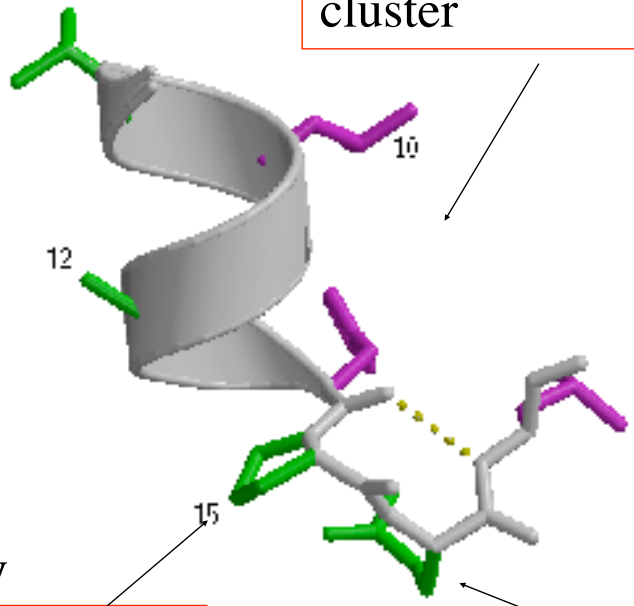


Sequence pattern=
...nppnnpp[HNYF]P[DE]n



structural variability

note:
hydrophobic
cluster



Pro blocks helix

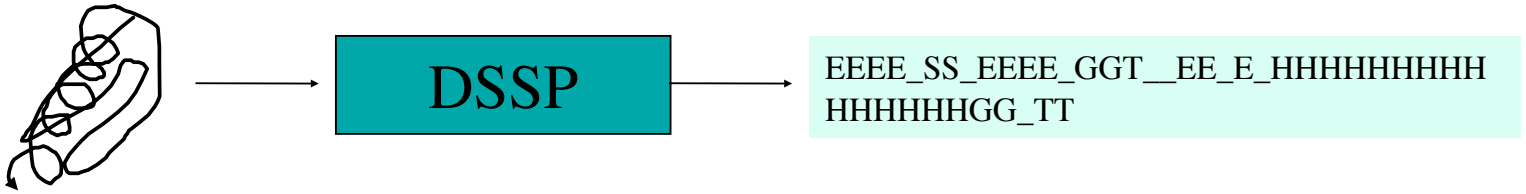
D or E stabilizes
tight turn

"n"=non-polar
"p"=polar
[...]=alternative aa's

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

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!)

_= unclassified. None-of-the-above. Generic loop, or beta-strand with no regular H-bonding.

collectively called

L

for Loop

Accuracy of 3-state predictions

```
True SS:    EEEE_SS_EEEE_GGT__EE_E_HHHHHHHHHHHHHHHHGG_TT
Prediction: EEEELLLLHHHHHHLLLLEEEEEHHHHHHHHHHHHHHHHHHLL
```

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:

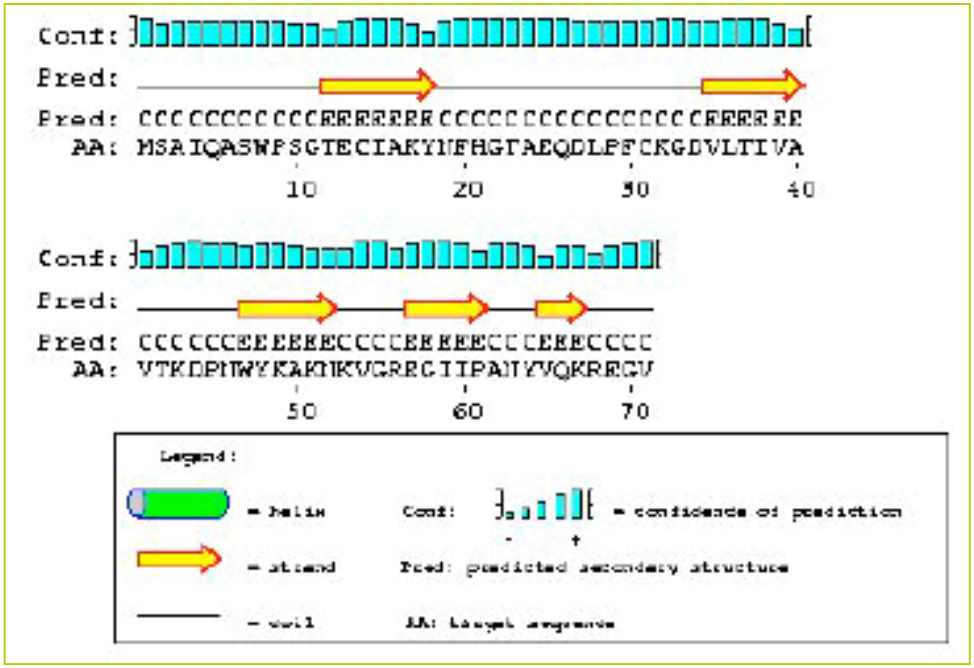
PHD (Burkhard Rost) -- 72-74% Q3

HMMSTR (Bystroff) -- 74-75% Q3

Psi-pred (David T. Jones) -- 76-78% Q3

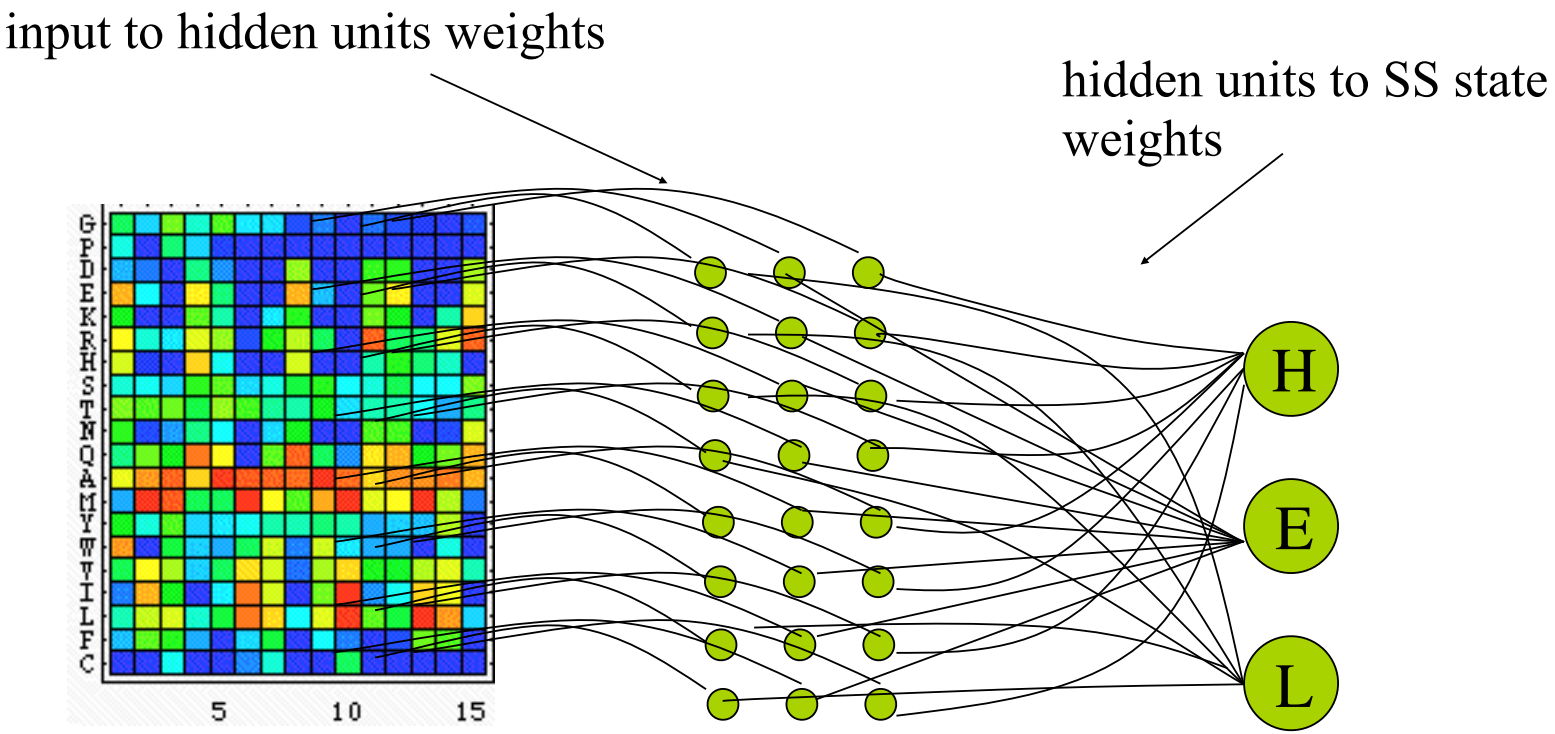
PSI-pred-- a secondary structure predictor

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.

Reminder

Making a sequence profile

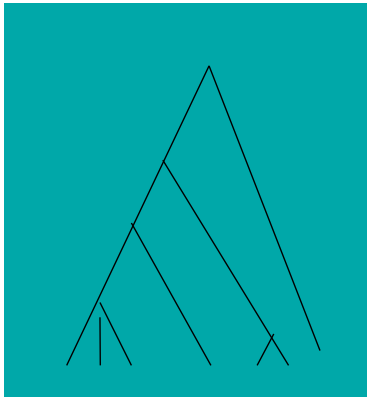
1. Multiple sequence alignment

VIASAVRTA
VIVDAGRSA
VIASGVRTA
VIVAAKRTA
VIVSAVRTP
VIVSAARTA
VIVSAVRTP
VIVDAGRTA
VIVDAGRTA
VIVSGARTP
VIVDFGRTP
VIVSATRTP
VIVSATRTP
VIVGALRTP
VIVSATRTP
VIVSATRTP
VIASAARTA
VIVDAIRTP
VIVAAYRTA
VIVSAARTP
VIVDAIRTP
VIVSAVRTA
VIVAAHRTA

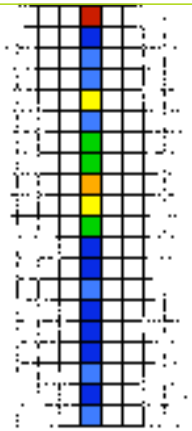
3. sum weights of each amino acid.

$$P_{ij} = \frac{\sum_{k=seqs} w_k \delta(s_{kj} = aa_i)}{\sum_{k=seqs} w_k}$$

2. sequence weights from phylo.tree



4. Sequence profile, probabilities of 20 amino acids

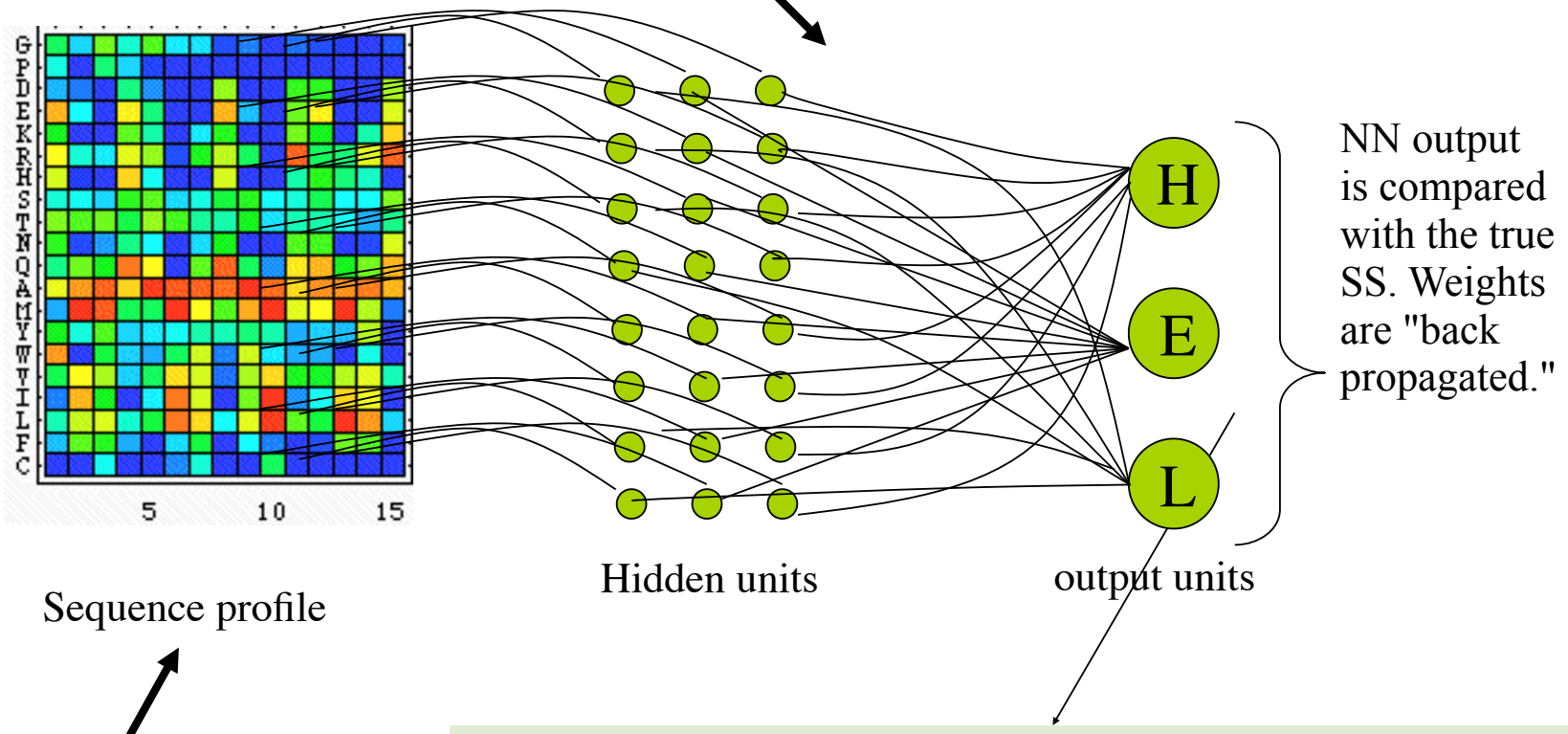


aa

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

Psi-Pred: Training the neural network (NN)

weights are found that *minimize errors*



Protein database provides both input and output

```
True SS:    EEEE_SS_EEEE_GGT__EE_E_HHHHHHHHHHHHHHHHGG_TT
Prediction: EEEELLLLHHHHHHHLLLLEEEEEHHHHHHHHHHHHHHHHHHLL

Errors:     00000000111111000000101000000000000000000100
```

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

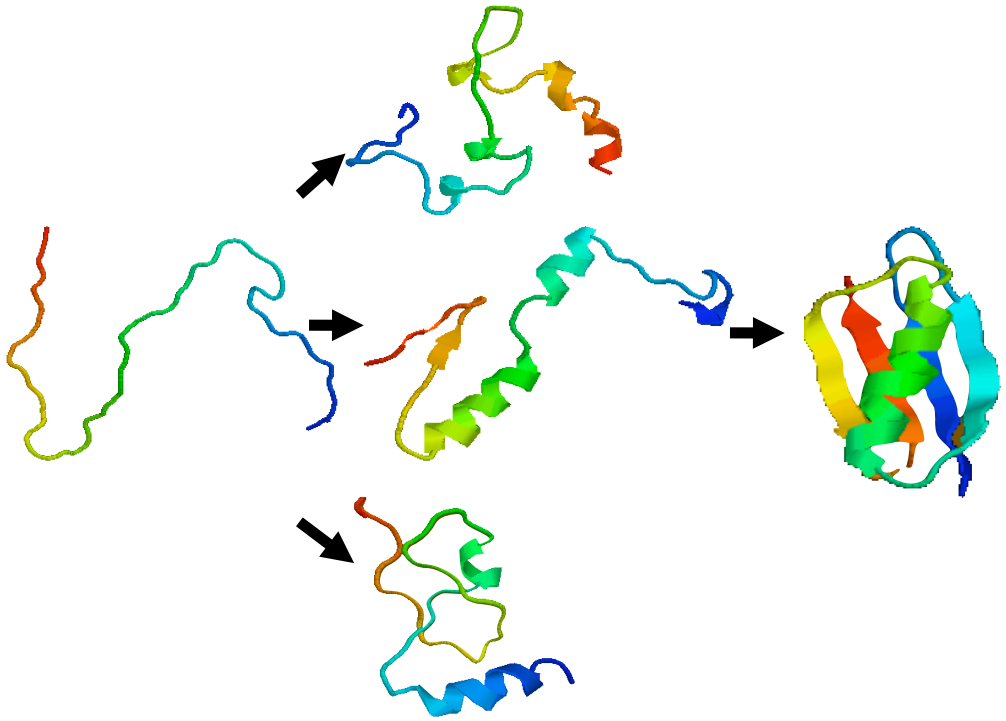
- GOR
- Chou Fasman
- PHD/PROF
- ZPRED
- PREDATOR

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! really???

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.