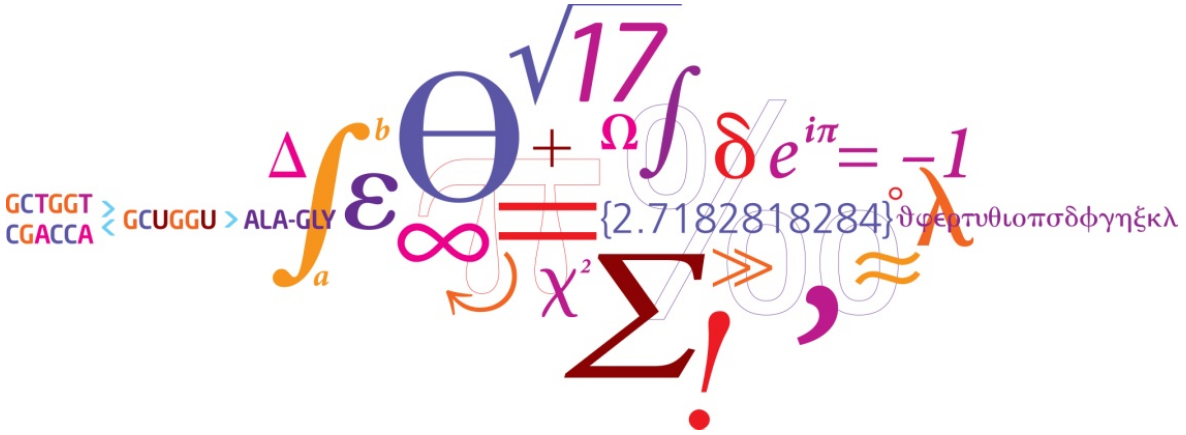
Conformational B-cell epitope Prediction

Paolo Marcatili



DTU Biosys
Department of Systems Biology

Prediction of linear epitopes

- Pro
- easily predicted computationally
- easily identified experimentally
- immunodominant epitopes in many cases
- do not need 3D structural information
- easy to produce and check binding activity experimentally

Con

- only ~10% of epitopes can be classified as “linear”
- weakly immunogenic in most cases
- most epitope peptides do not provide antigen-neutralizing immunity
- in many cases represent hypervariable regions

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

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

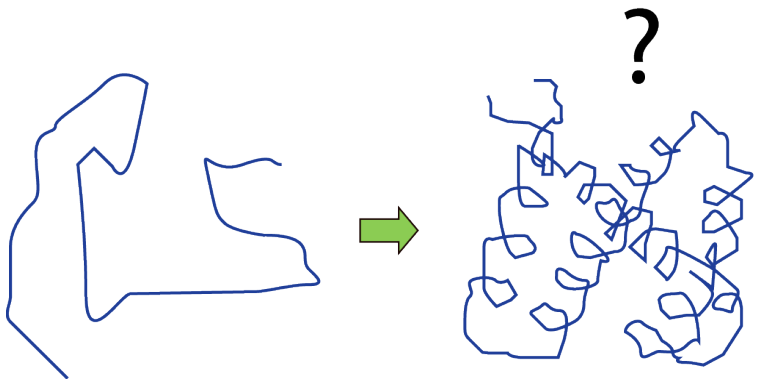
- Both methods are time consuming and not easily done in a larger scale

- Structure prediction
  - Homology modeling
  - Fold recognition

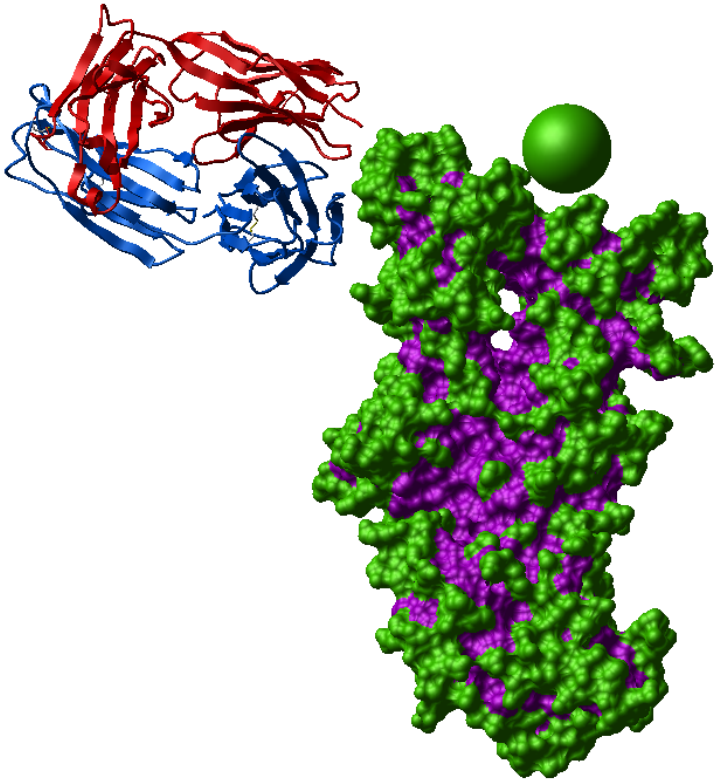
- Less time consuming, but there is a possibility of incorrect predictions, specially in loop regions

Protein structure prediction

- 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



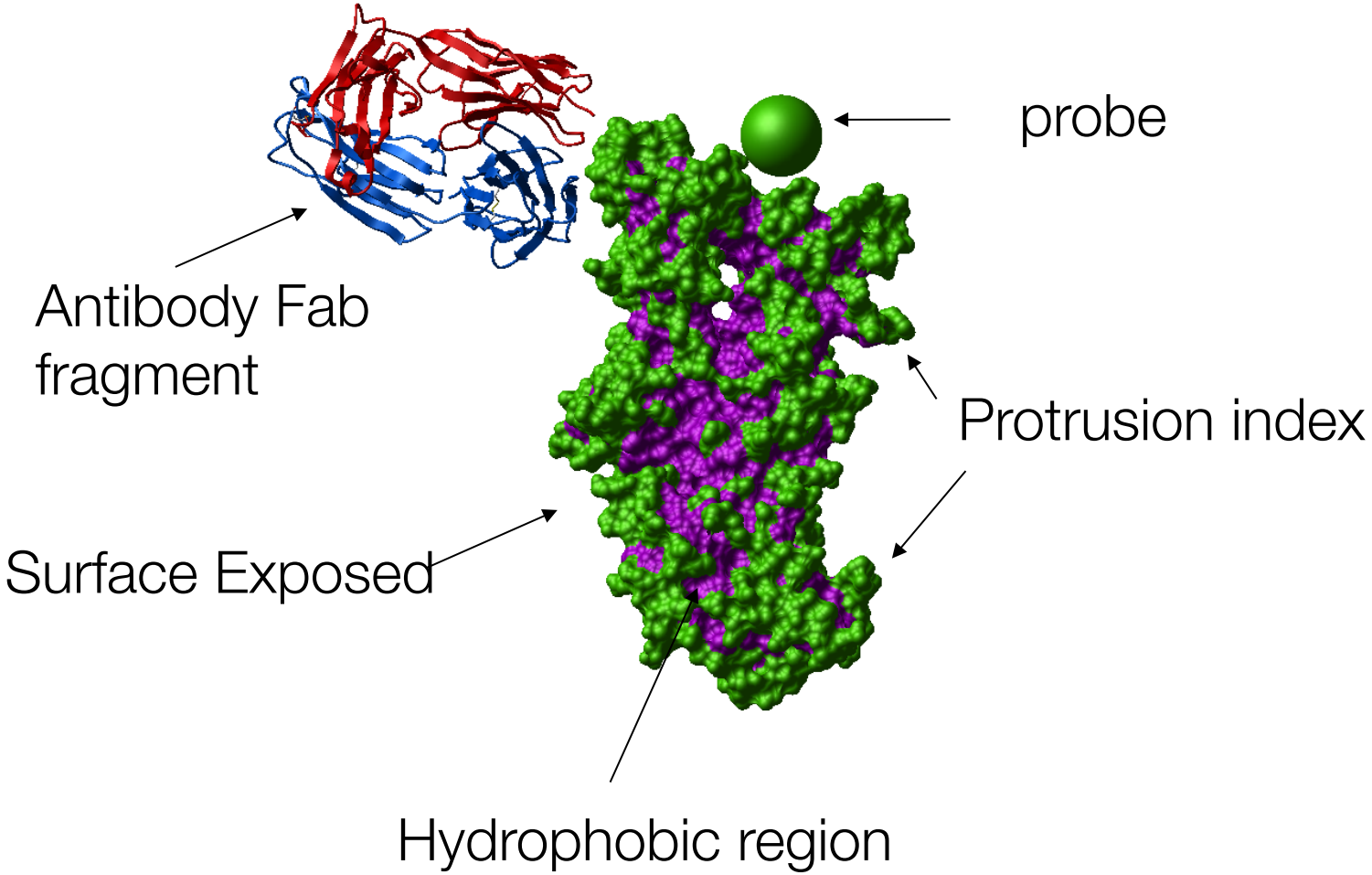
What does antibodies recognize in a protein?



The binding interaction

	Epitope characteristics	Reference
Size	• 10-25 residues is involved in binding • 15±4 residues is involved in binding • 22±8 residues is involved in binding • 600-1000 $Å^2$ is buried upon binding • 847±279 $Å^2$ accesible surface area • The epitope plane (see results): 401±133$Å^2$ when approximated by an ellipse • Thickness (see results): 8.2±2.0Å	• Van Regenmortal (2009) • Present study • Sun et al., (2011) • Rubinstein et al. (2008) • Sun et al., (2011) • Present study • Present study
Shape	• Flat rugged area • Flat oblong (ellipse) shaped area	• Rubinstein et al. (2008) • Present study
Segmentation	• Above 60% epitope residues exists in linear stretches of 3 or more residues • 85% of epitopes has a linear stretch of 5 or more residues	• Rubinstein et al. (2008) and present study • Sun et al., (2011) and present study
Secondary structure	• Enriched by loops • Depleted of strands and helixes	• Rubinstein et al. (2008) and Ofran et al. (2008)
Epitope position on the antigen	• Epitopes are more surface exposed than the remaining antigen • Epitopes protrude from the antigen surface	• Andersen et al. (2006) and Rubinstein et al. (2008) • Thornton et al. (1986)
Orientation relative to the antibody	• Epitopes bind predominantly in a -30 to 60 degrees angle relative to the light to heavy antibody chain direction	• Present study
Amino acid composition	• Enriched by polar and charged amino acids and depleted of hydrophobic amino acids compared to non-epitope antigen residues, surface exposed antigen residues or general protein composition • No significant deviation from the non-epitope antigen surface, however a tendency for depletion of small hydrophobic amino acids is observed	• Andersen et al. (2006); Ofran et al. (2008); Rubinstein et al. (2008); Zhao and Li (2010) and Sun et al., (2011) • Present study
Amino acid coorporativeness	• Pairs of Tyr:Tyr, Cys:Pro, Asn:Tyr, Gly:Tyr, Asp:Pro, Thr:Tyr and Arg:Tyr are more frequently observed in epitopes compared to the remaining antigen surface • Pairs of Asn:Tyr, His:Tyr and His:Met are more frequently observed in epitopes	• Rubinstein et al. (2008) • Sun et al., (2011)
Spatial amino acid composition	• Hydrophobic core flanked by charged amino acids • Preferable; hydrophobic amino acids closes to the antibody, then hydrophilic and furthest away positive charged amino acids	• Present study* • Present study

What does antibodies recognize in a protein?



Structure base prediction methods

Name	Input	Implemented epitope feature	LA	Performance	Ref
SVM based prediction of linear B-cell epitopes (Linear)	Sequence	- Amino acid composition. (Position specific amino acid composition using a 20 amino acid window of the primary sequence)	SVM	75% accuracy AUC = 0.84	Wee et al., 2010 [36]
Identification of conformational B-cell epitopes in an antigen from its primary sequence (conformational)*	Sequence	- Amino acid composition. (Pattern composition, which is the percentage of a given amino acid in a 15 amino acid stretch. See text)	SVM	86% accuracy MCC = 0.73	Ansari and Raghava 2010 [37]
EPSVR (Conformational)	Structure	Propensity scores based on: - Epitope amino acid composition - Amino acid conservation - Side chain energy score - Contact number - Surface planarity score - Secondary structure composition.	SVR	AUC = 0.597	Liang et al., 2010 [34]
Mining for the antibody-antigen interacting associations that predict the B cell epitopes (Conformational)	Antigen-Antibody Sequence	- Paratope-epitope co-occurring patterns of interacting residue pairs. - Primary sequence cooperative of both paratope and epitope	None	AUC = 0.593	Zhao and Li 2009 [21]
Epitopia: a web-server for predicting B-cell epitopes	Sequence or structure	- Patch on surface of the protein equal to average epitope size - Amino acid composition - Frequency of helices. - RSA, - RSA (probe = 9Å), - Average curvature of atoms in epitopes, - Proportion of patch atoms that reside within 4Å from a convex hull of the antigen.	Naïve Bayes Classifier	AUC = 0.62. The authors furthermore report 89.4% success (self invented criteria)	Rubinstein et al., 2009 [38] Rubinstein et al., 2008 [28]
COBEpro: a novel system for predicting continuous B-cell epitopes (Linear)	Sequence	Measures of similarities between query peptides and a set of known epitope fragments: - Number of amino acids present in both sequence - Number of dimers , trimers etc present in both sequences.	SVM	AUC = 0.628	Sweredoski and Baldi 2009 [39]
ElliPro: a new structure-based tool for the prediction of antibody epitopes (Conformational)	Structure	- Protrusion Index (PI) Implements the Thornton method (see text) and clusters amino acid based on their PI value.	None	AUC = 0.732	Ponomarenko et al., (2008) [29]
Pepito: improved discontinuous B-cell epitope prediction using multiple threshold and half sphere exposure (Conformational)	Structure	- Amino acid composition - Half-sphere exposure (Hamelryck 2005). Number of C-alpha atoms in 9Å lower and upper half-sphere individually.	None	AUC = 0.683	Sweredoski and Baldi 2009 [40]
Identification of discontinuous antigenic determinants on proteins based on shape complementarity (Conformational)	Structure	- Paratope – epitope shape complementarity - Surface accessibility	None	AUC = 0.634	Rapberger et al., 2007 [41]
Predicting B cell epitope residues with network topology based amino acid indices (Linear)	Sequence	- Relative connectivity propensity scores, based on network topology (connection between C-alpha atoms within 8Å), but similar to contact number to some extent half-sphere exposure	None	AUC = 0.796 (limited dataset)	Huang et al., 2007 [42]
Prediction of linear B-cell epitopes using amino acid pair antigenicity scale (Linear)	Sequence	- Amino acid cooperativeness in epitopes in relation to non-epitopes	SVM	MCC = 0.37	Chen et al., 2007 [33]
DiscoTope: Prediction of residues in discontinuous B-cell epitopes using protein 3D structure (Conformational)	Structure	- Amino acid composition - Residue contact number (number of C-alpha in a 10Å sphere)	None	AUC = 0.62	Andersen et al., 2006 [25]

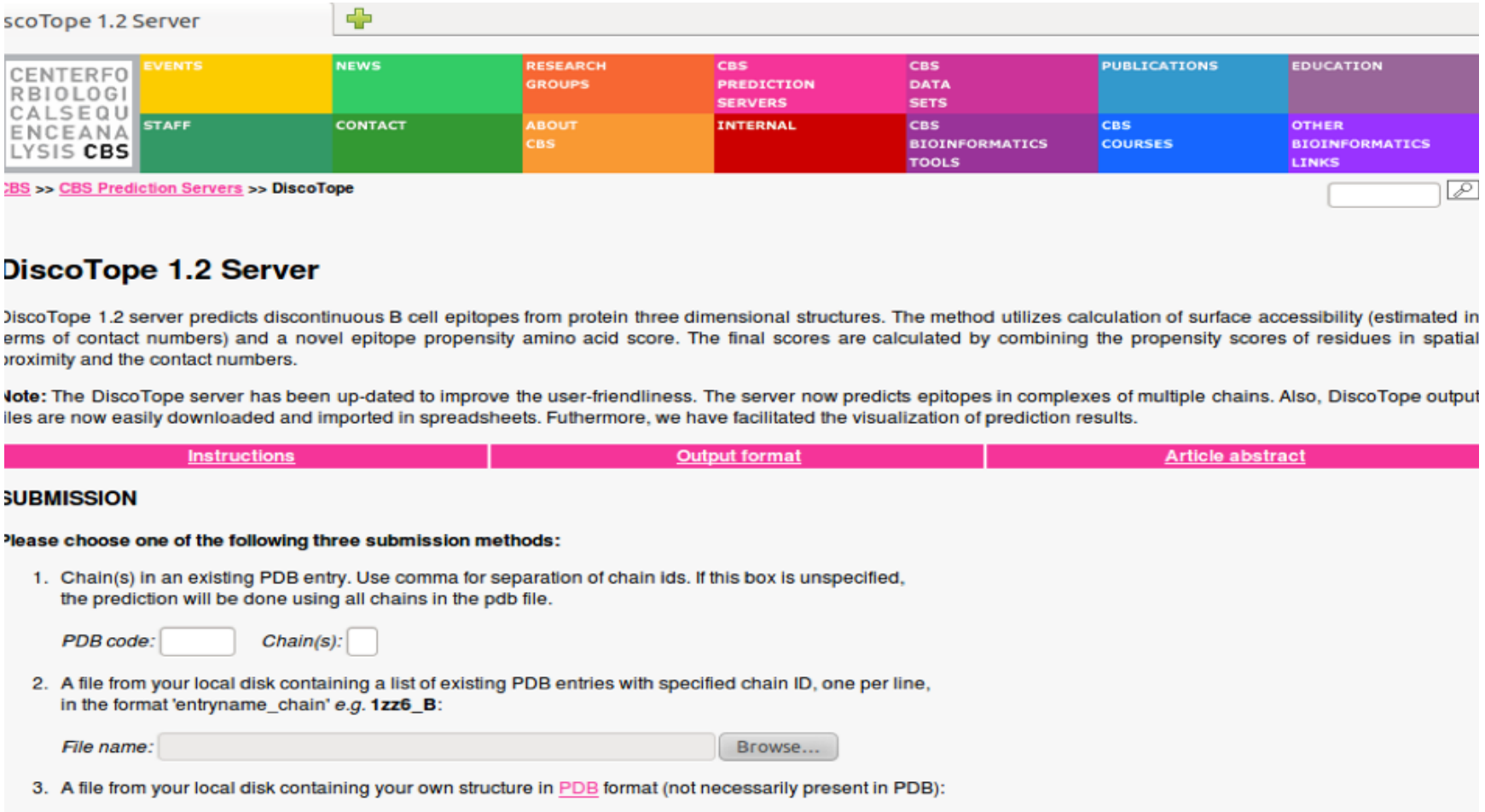
* The author of this report has his doubts about this work. See text.

DiscoTope

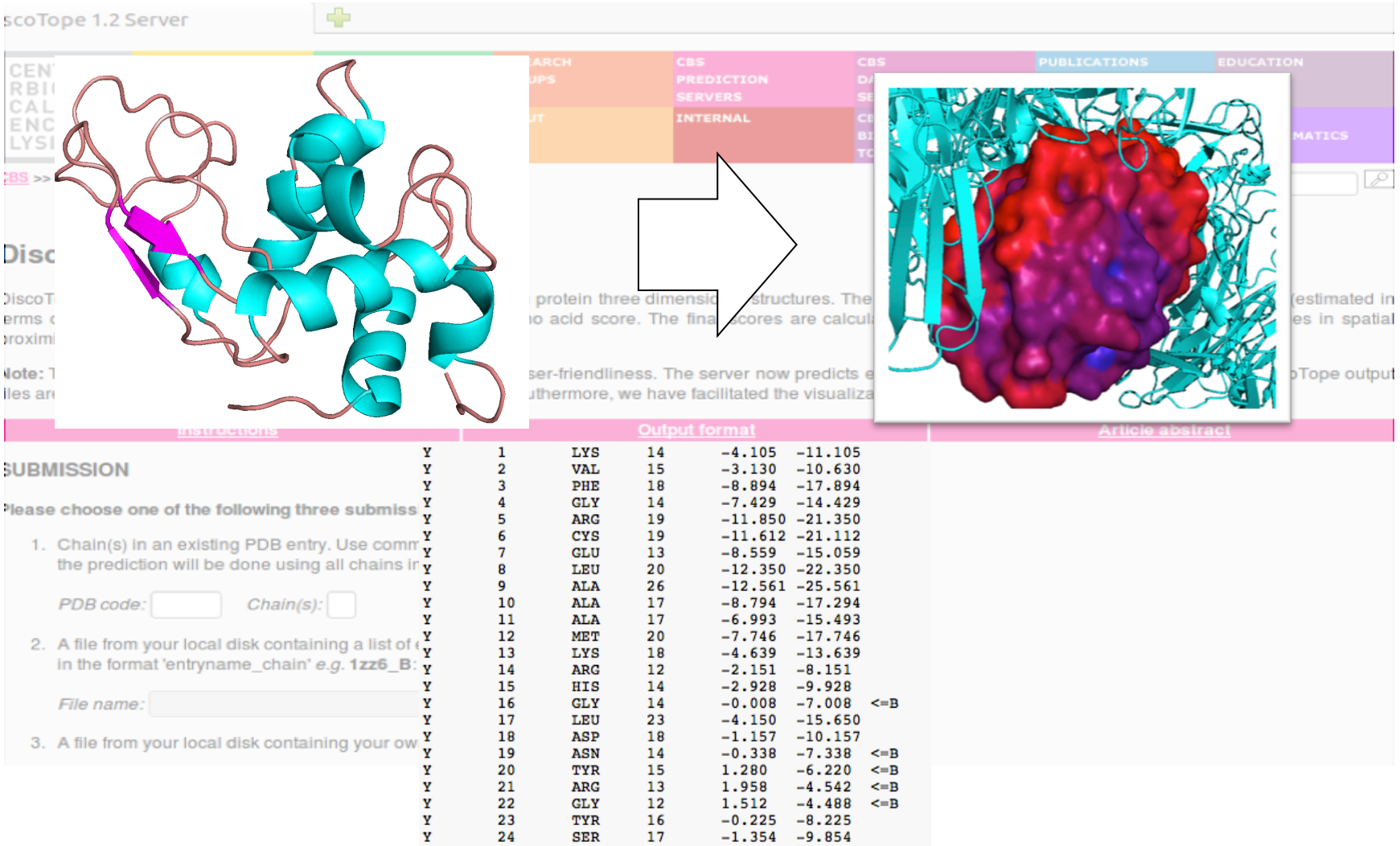
- Prediction of residues in discontinuous B cell epitopes using protein 3D structures

Pernille Haste Andersen, Morten Nielsen and Ole Lund, Protein Science 2006

Predicting B-cell epitopes

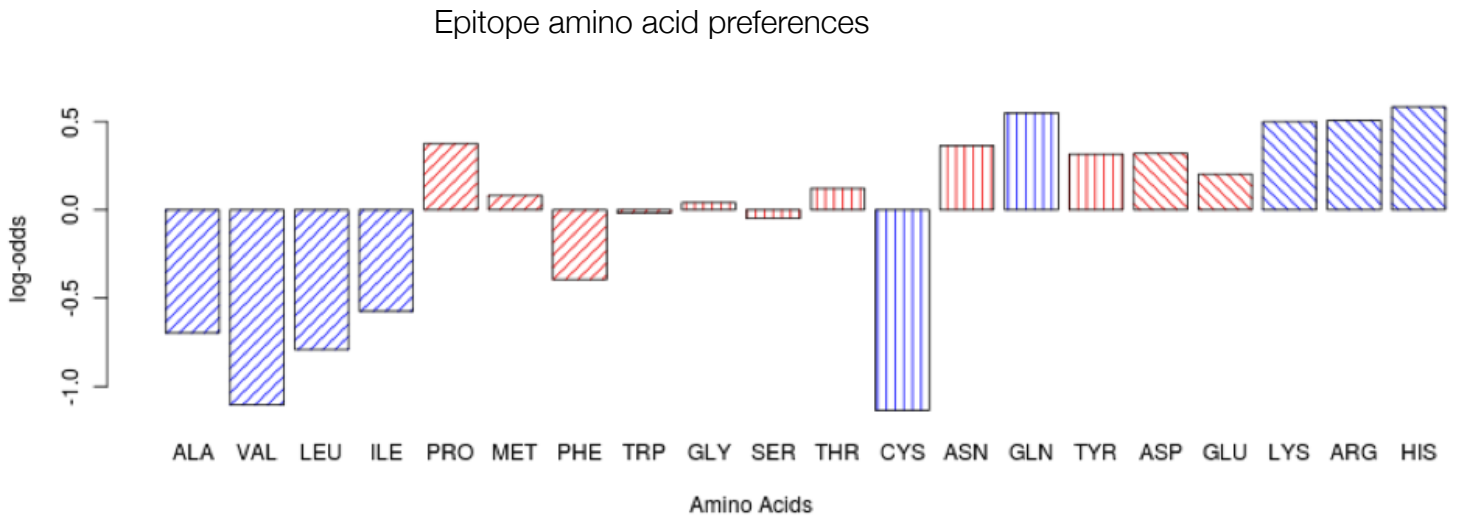


The DiscoTope method



The DiscoTope method

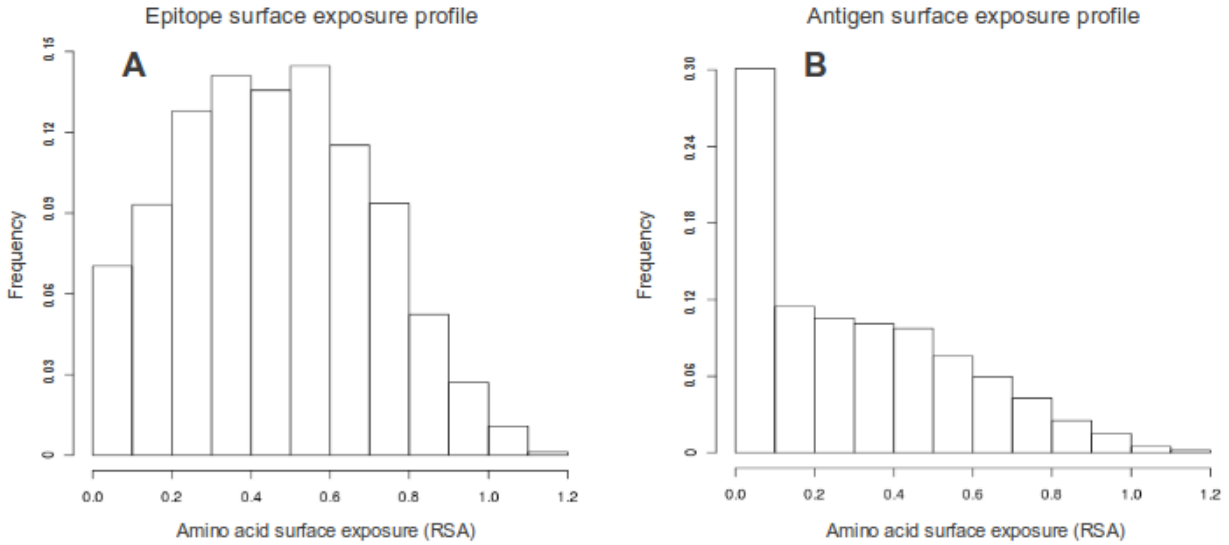
- Some amino acids are preferred and disliked in the epitope



Kringelum et al, 2011, manuscript in preparation

The DiscoTope method

- Some amino acids are preferred and disliked in the epitope
- Epitopes reside on the surface of the protein



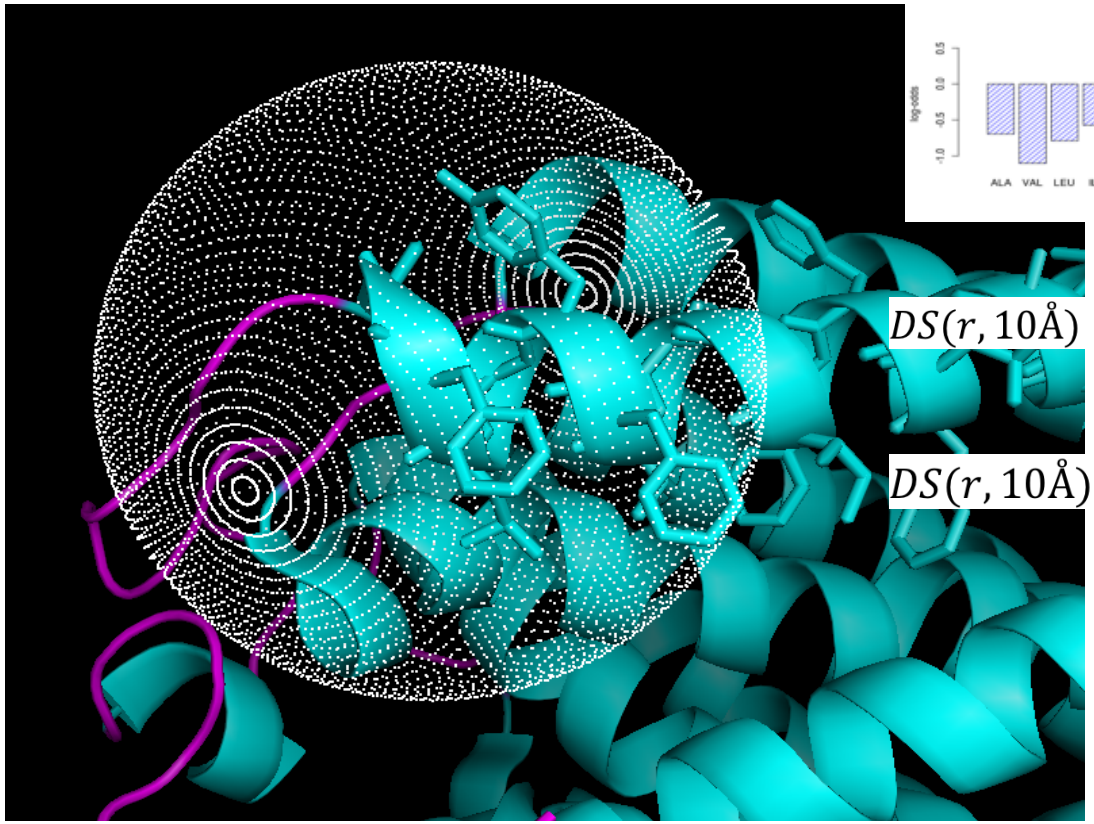
Kringelum et al, 2011

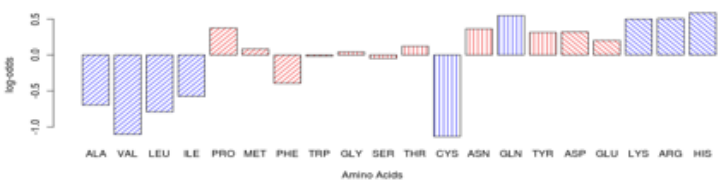
The DiscoTope method

•Predictions score for each residue are calculated by summing the epitope likelihood (propensity) of surrounding residues and subtracting the neighbor count

$$DS(r) = ps(r, 10\text{Å}) - 0.5 * N_{neighbors}$$

$$ps(r,k) = \sum_i P_{epi}(r_i)\ if\ |r - r_i| < k$$





$$DS(r, 10\text{Å}) = 2 * P(Tyr) + P(Phe) + P(Ale) - 0.5 * 4$$

$$DS(r, 10\text{Å}) = 2 * 0.25 - 0.3 - 0.7 - 0.5 * 4 = -2.5$$

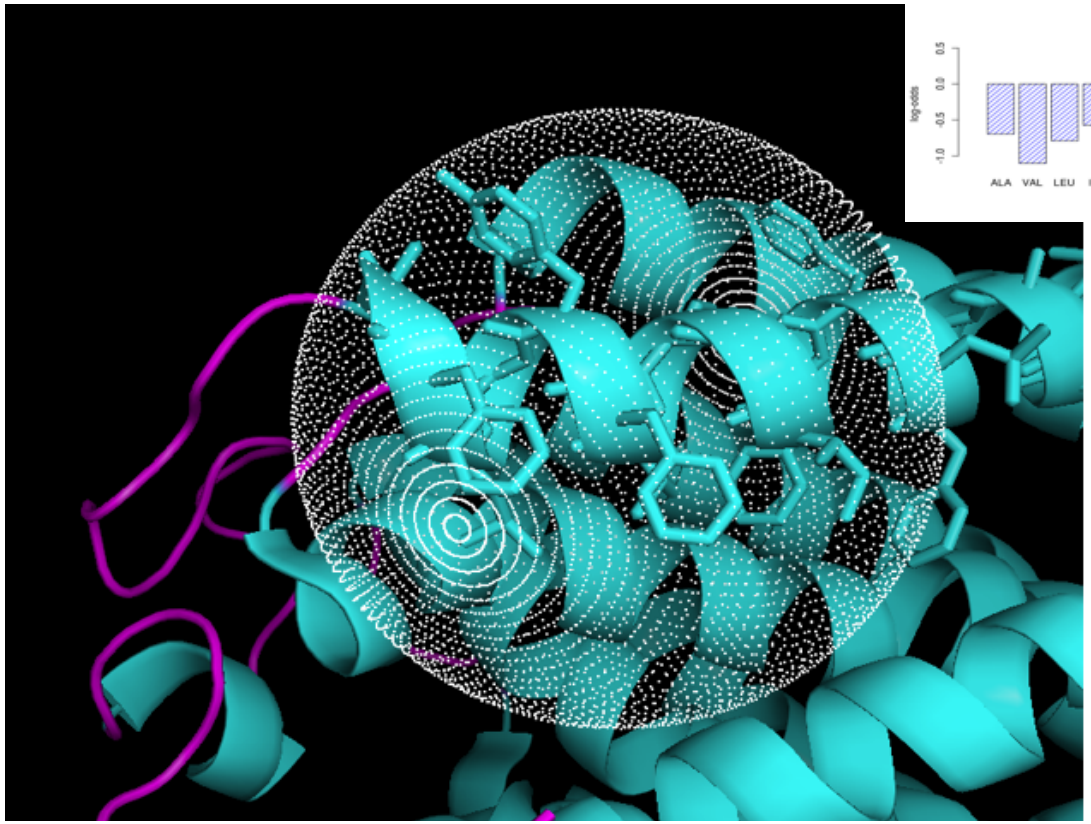
Andersen et al., 2006

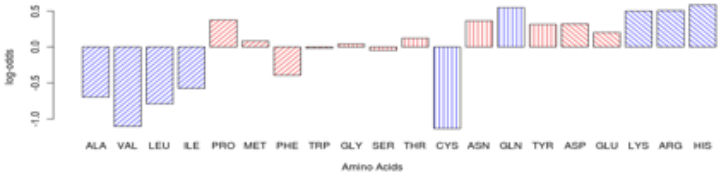
The DiscoTope method

•Predictions score for each residue are calculated by summing the epitope likelihood (propensity) of surrounding residues and subtracting the neighbor count

$$DS(r) = ps(r, 10\text{Å}) - 0.5 * N_{neighbors}$$

$$ps(r,k) = \sum_i P_{epi}(r_i)\ if\ |r - r_i| < k$$





Performance: A_{roc} = 0.700

On a dataset of 75 antigen-antibody complexes divided in 25 proteins

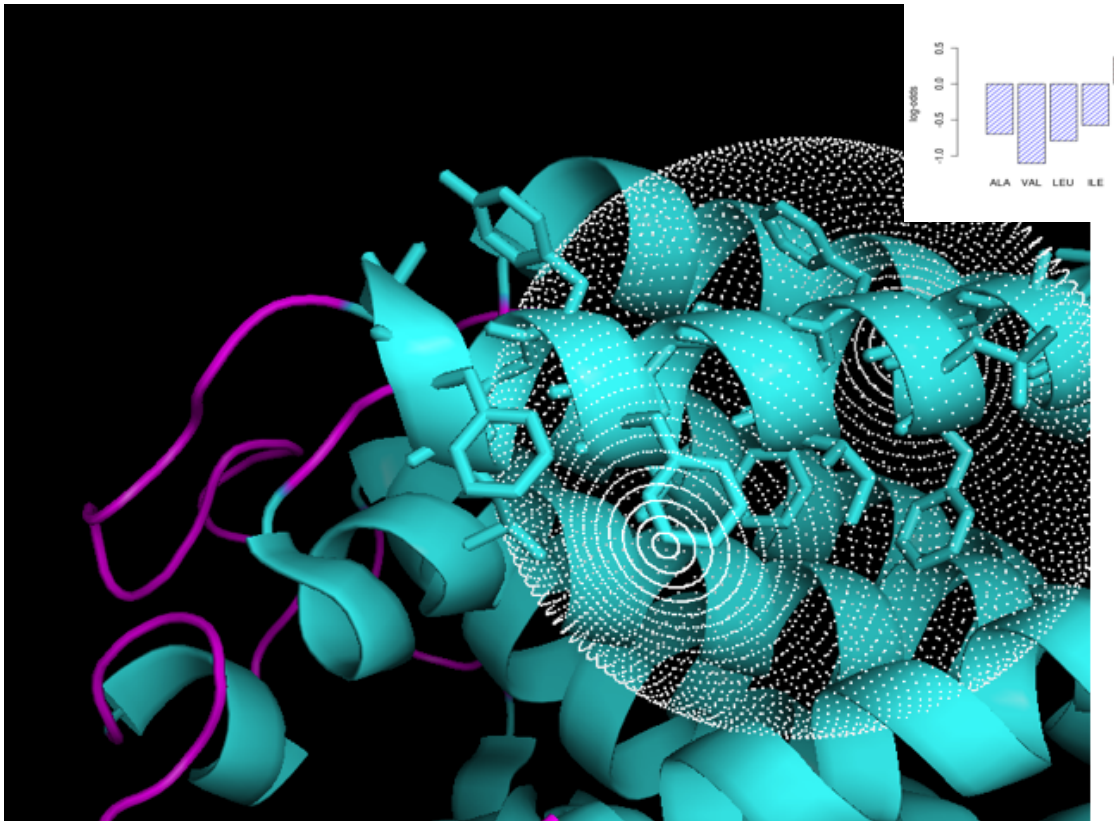
Andersen et al., 2006

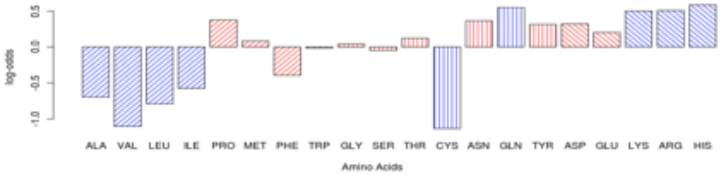
The DiscoTope method

•Predictions score for each residue are calculated by summing the epitope likelihood (propensity) of surrounding residues and subtracting the neighbor count

$$DS(r) = ps(r, 10\text{Å}) - 0.5 * N_{neighbors}$$

$$ps(r,k) = \sum_i P_{epi}(r_i)\ if\ |r - r_i| < k$$





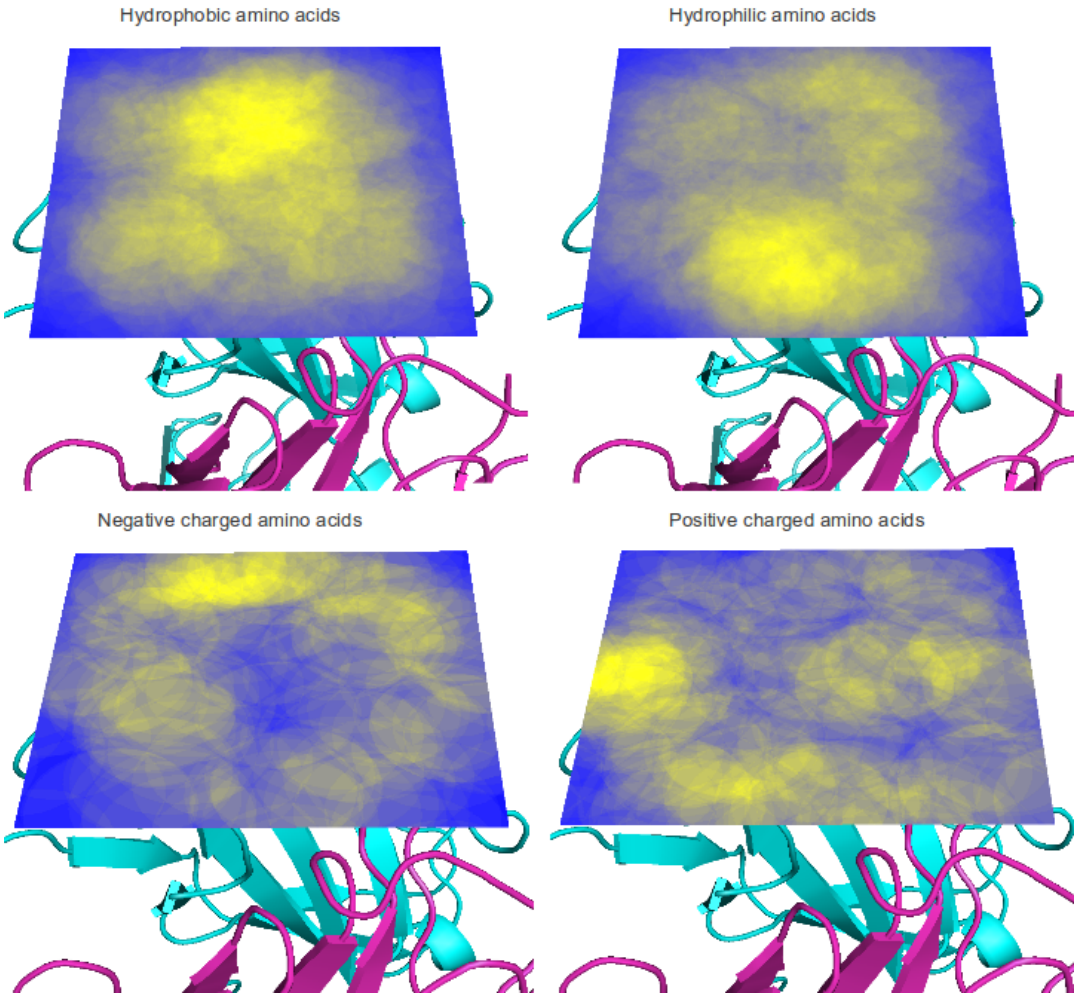
Performance: A_{roc} = 0.700

On a dataset of 75 antigen-antibody complexes divided in 25 proteins

Andersen et al., 2006

However position matters

•Uneven spatial distribution of amino acid in epitopes



Kringelum et al, 2012,

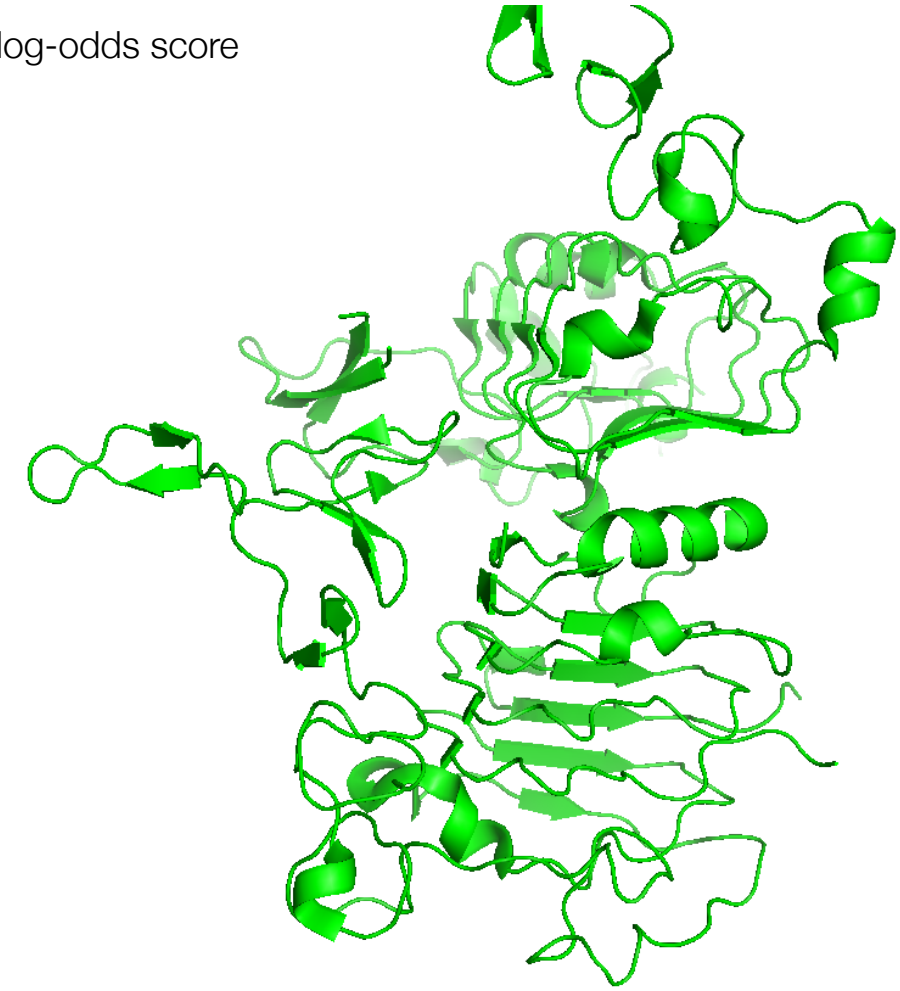
Propensity score function

Weight factor

Amino acid log-odds score

$$PS(r) = \sum_i \beta_i \cdot ls(r_i)$$

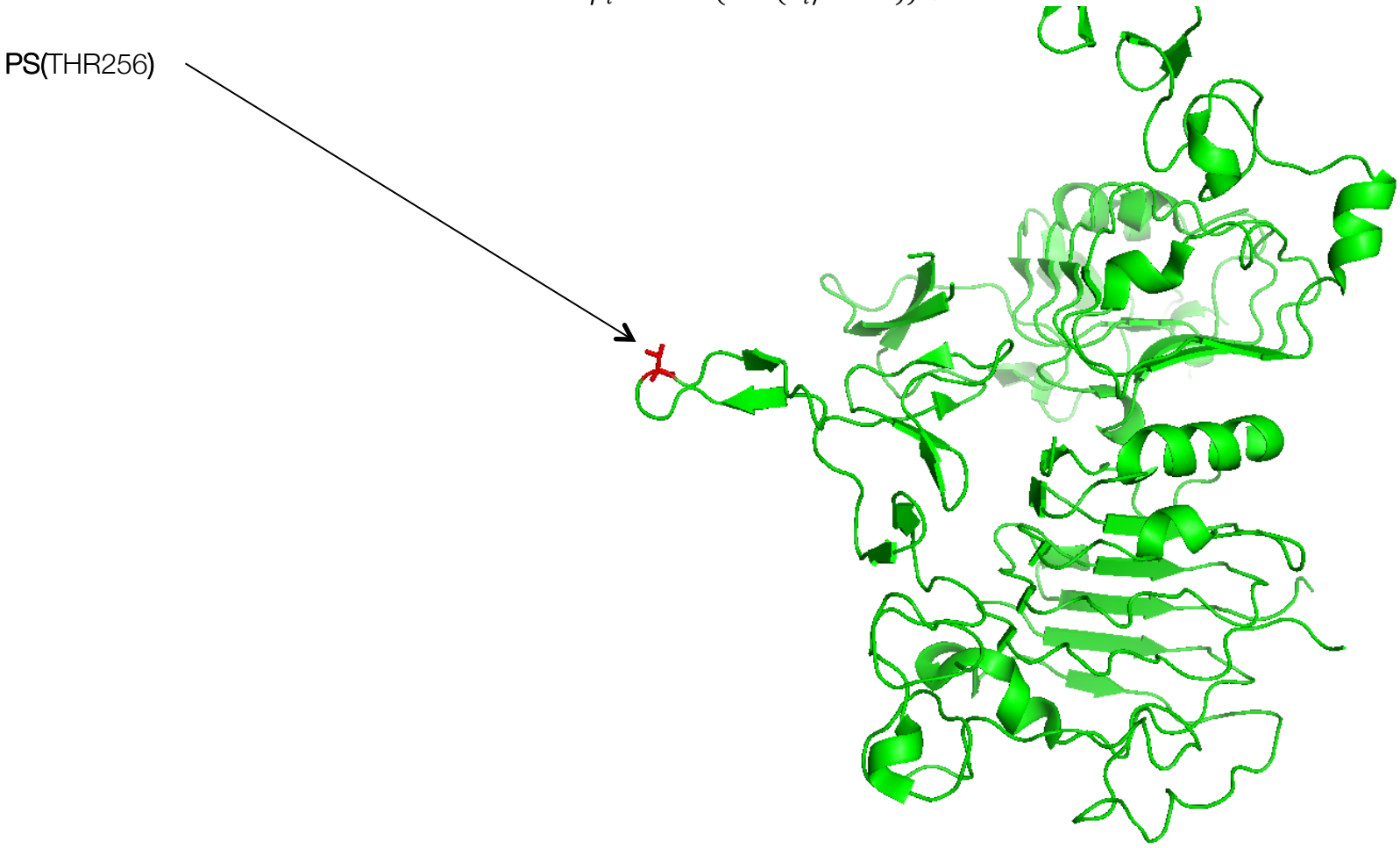
$$\beta_i = 0.8 * (1 - (d_i/21.6\text{Å})) + 0.2$$



Propensity score function

$$PS(r) = \sum_i \beta_i \cdot ls(r_i)$$

$$\beta_i = 0.8 * (1 - (d_i/21.6\text{Å})) + 0.2$$



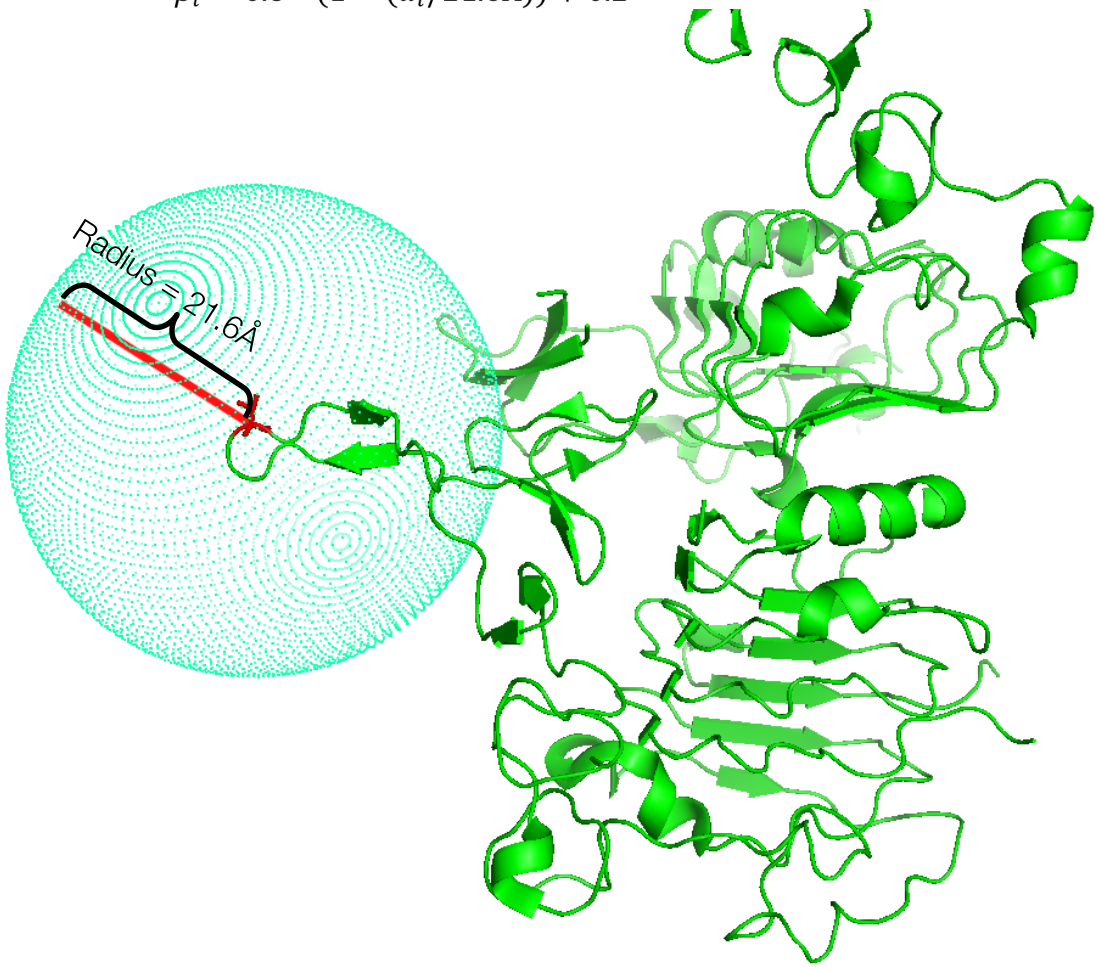
Propensity score function

PS(THR256)

1) Identify Neighbor residues within 21.6Å (C_α-C_α)

$$PS(r) = \sum_i \beta_i \cdot ls(r_i)$$

$$\beta_i = 0.8 * (1 - (d_i/21.6\text{Å})) + 0.2$$



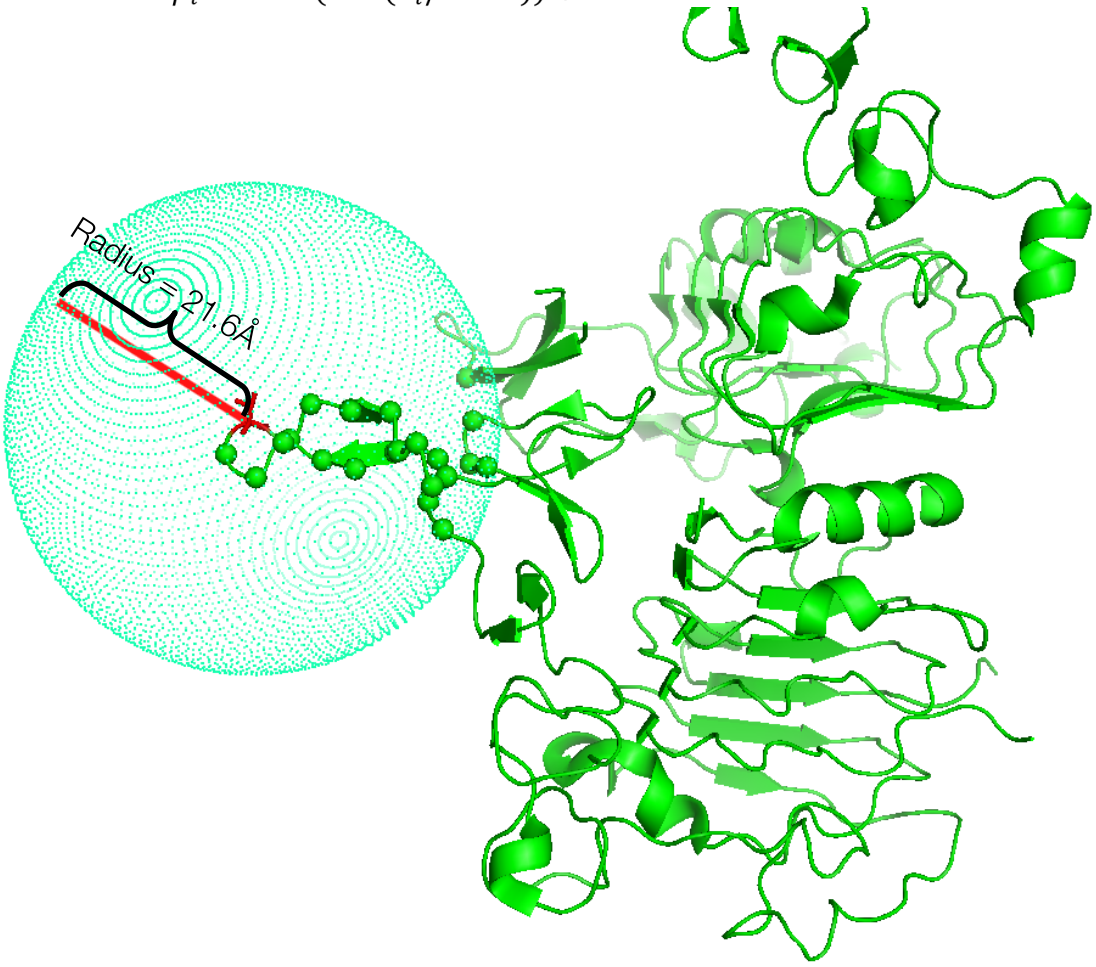
Propensity score function

PS(THR256)

1) Identify Neighbor residues within 21.6Å (C_α-C_α)

$$PS(r) = \sum_i \beta_i \cdot ls(r_i)$$

$$\beta_i = 0.8 * (1 - (d_i/21.6\text{Å})) + 0.2$$



Propensity score function

PS(THR256)

1) Identify Neighbor residues within 21.6Å (C_α-C_α)
2) Calculate summed propensity score:

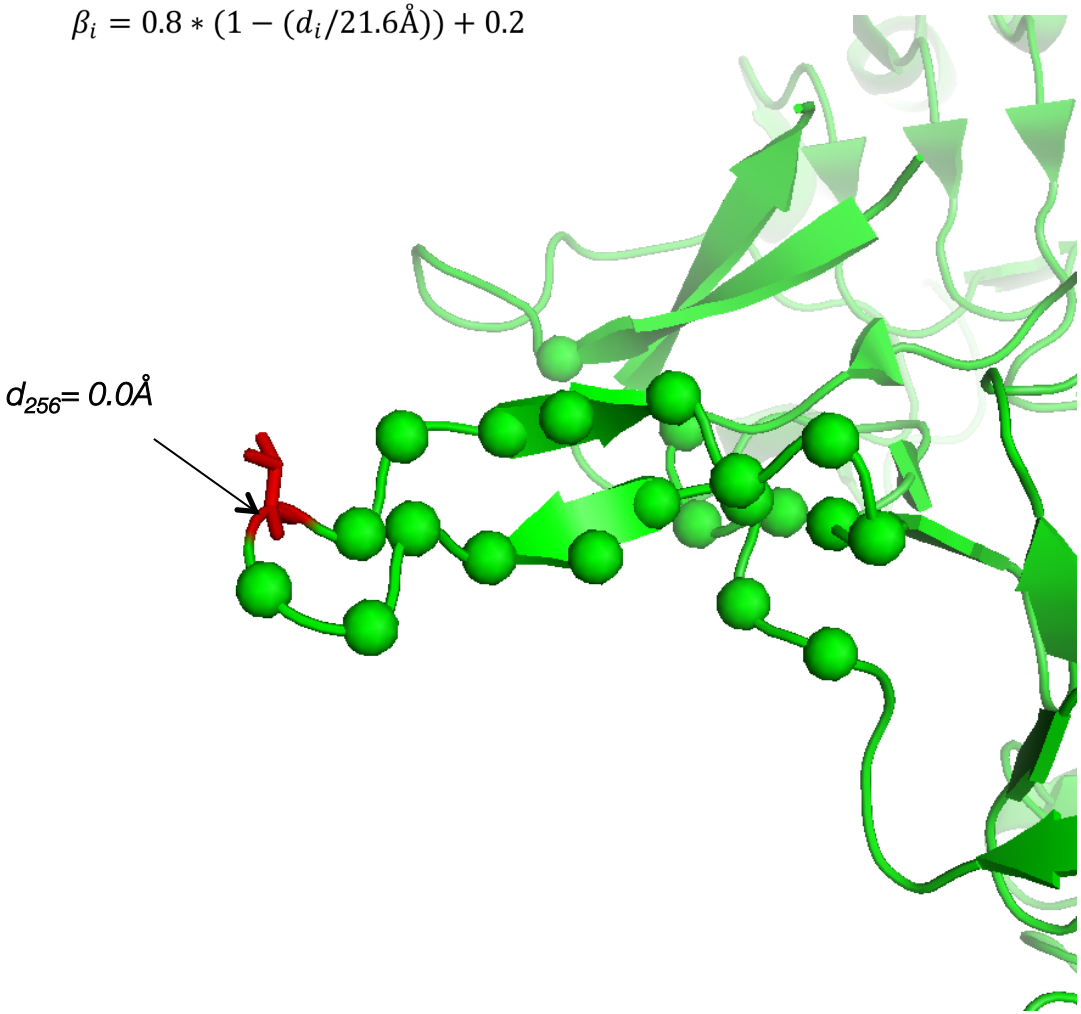
ls(THR) = -0.23

β_{256}= 0.8*(1-**0.0**/21.6)+0.2 = 1.0

ls(THR)*β_{256} = -0.23*1.0 = **-0.23**

$$PS(r) = \sum_i \beta_i \cdot ls(r_i)$$

$$\beta_i = 0.8 * (1 - (d_i/21.6\text{Å})) + 0.2$$



Propensity score function

PS(THR256)

1) Identify Neighbor residues within 21.6Å (C_α-C_α)
2) Calculate summed propensity score:

ls(THR) = -0.23

β_{256}= 0.8*(1-**0.0**/21.6)+0.2 = 1.0

ls(THR)*β_{256} = -0.23*1.0 = **-0.23**

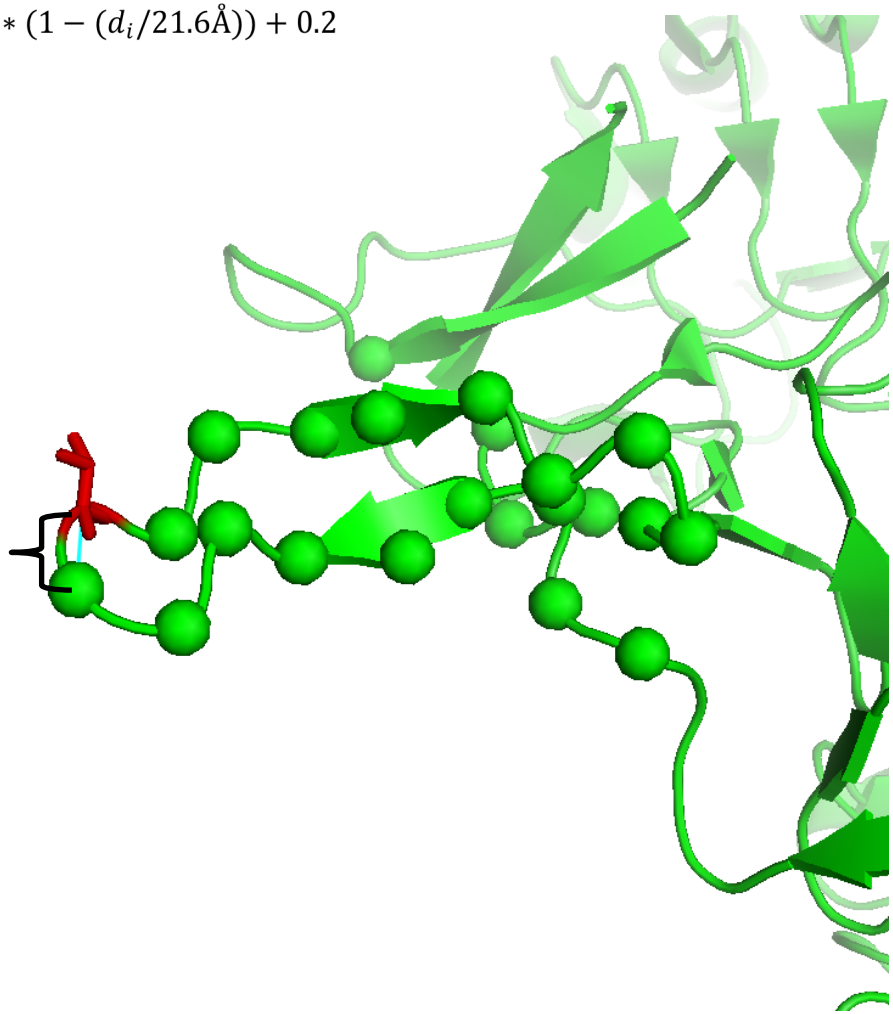
ls(ASP) = 2.5

β_{255}=0.8*(1-**3.8**/21.6)+0.2 = 0.86

ls(THR)*β_{255} = 2.5*0.86 = **2.15**

$$PS(r) = \sum_i \beta_i \cdot ls(r_i)$$

$$\beta_i = 0.8 * (1 - (d_i/21.6\text{Å})) + 0.2$$



Propensity score function

PS(THR256)

1) Identify Neighbor residues within 21.6Å (C_α-C_α)
2) Calculate summed propensity score:

ls(THR) = -0.23

β_{256}= 0.8*(1-**0.0**/21.6)+0.2 = 1.0

ls(THR)*β_{256} = -0.23*1.0 = **-0.23**

---

ls(ASP) = 2.5

β_{255}=0.8*(1-**3.8**/21.6)+0.2 = 0.86

ls(ASP)*β_{255} = 2.5*0.86 = **2.15**

---

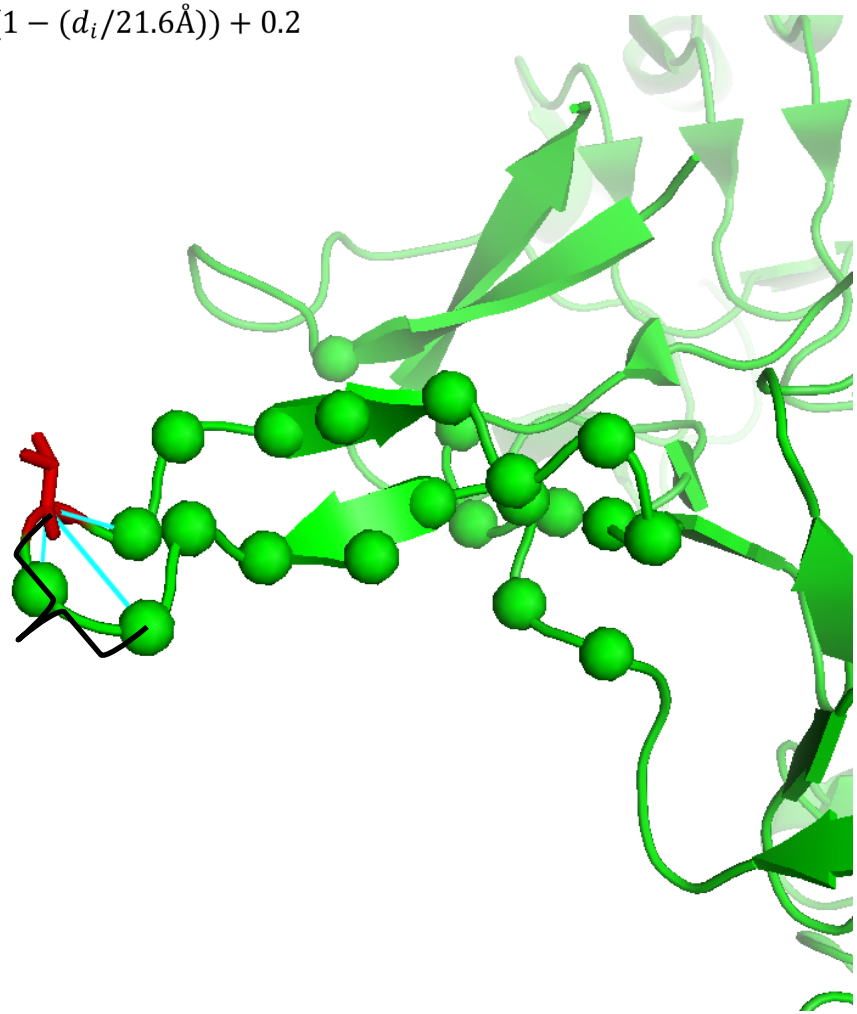
ls(THR) = -0.23

β_{254}=0.8*(1-**6.1**/21.6)+0.2 = 0.77

ls(THR)*β_{254} = -0.23*0.77 = **-0.18**

$$PS(r) = \sum_i \beta_i \cdot ls(r_i)$$

$$\beta_i = 0.8 * (1 - (d_i/21.6\text{Å})) + 0.2$$



d_{254}= 6.1Å

Propensity score function

PS(THR256)

1) Identify Neighbor residues within 21.6Å (C_α-C_α)
2) Calculate summed propensity score:

ls(THR) = -0.23

β_{256}= 0.8*(1-**0.0**/21.6)+0.2 = 1.0

ls(THR)*β_{256} = -0.23*1.0 = **-0.23**

---

ls(ASP) = 2.5

β_{255}=0.8*(1-**3.8**/21.6)+0.2 = 0.86

ls(ASP)*β_{255} = 2.5*0.86 = **2.15**

---

ls(THR) = -0.23

β_{254}=0.8*(1-**6.1**/21.6)+0.2 = 0.77

ls(THR)*β_{254} = -0.23*0.77 = **-0.18**

.....

.....

.....

---

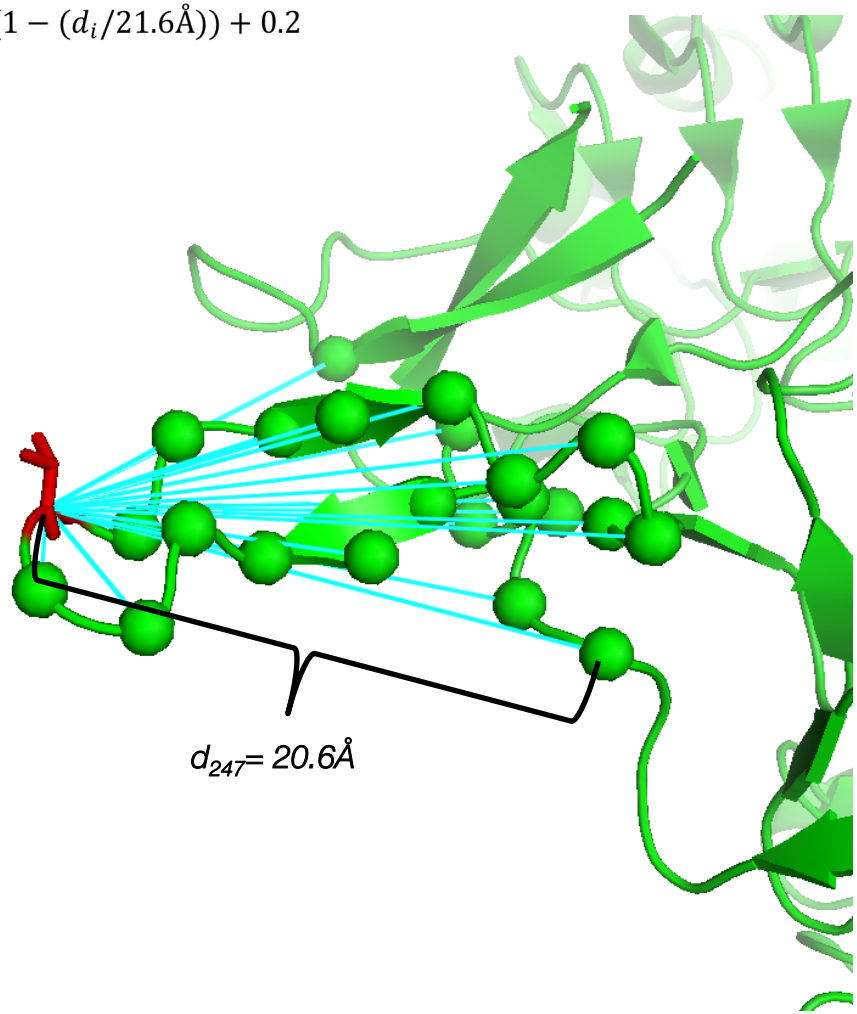
ls(PRO) = 1.2

β_{254}=0.8*(1-**20.6**/21.6)+0.2 = 0.24

ls(PRO)*β_{254} = 1.2*0.24 = **0.29**

$$PS(r) = \sum_i \beta_i \cdot ls(r_i)$$

$$\beta_i = 0.8 * (1 - (d_i/21.6\text{Å})) + 0.2$$



Propensity score function

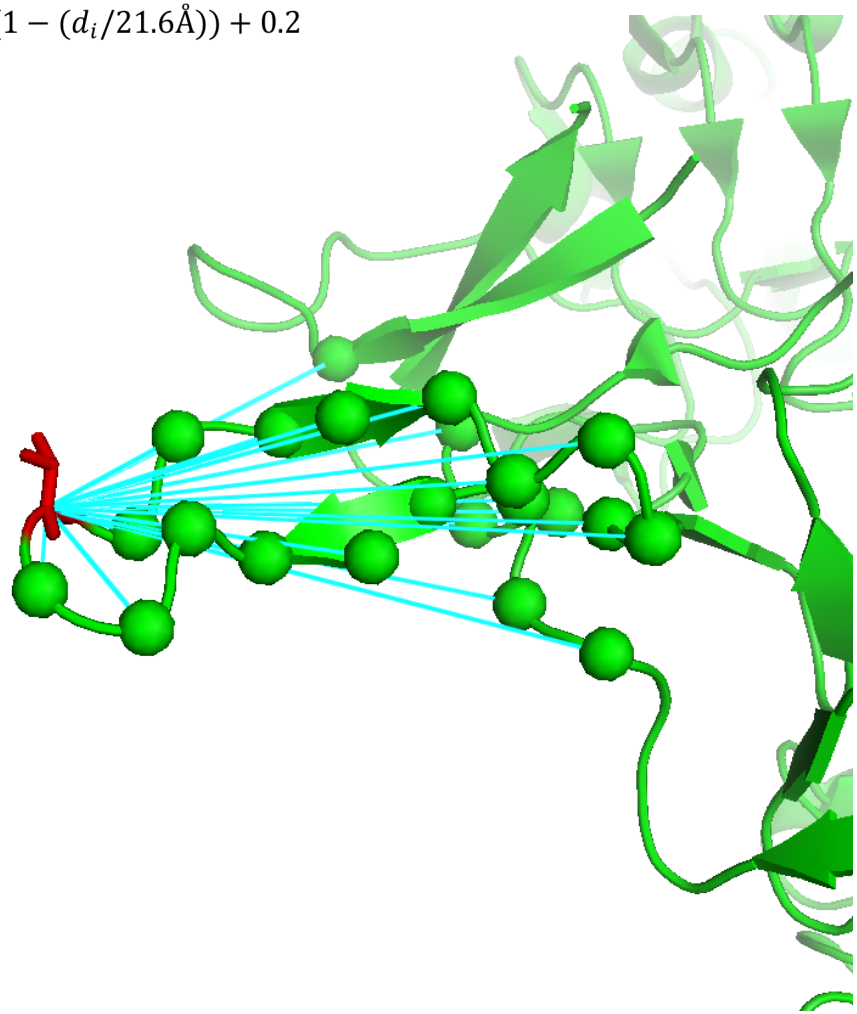
PS(THR256)

1) Identify Neighbor residues within 21.6Å (C_α-C_α)
2) Calculate summed propensity score:

ls(THR) = -0.23	
β_{256}= 0.8*(1-**0.0**/21.6)+0.2 = 1.0	
ls(THR)*β_{256} = -0.23*1.0	= **-0.23**
ls(ASP) = 2.5	
β_{255}=0.8*(1-**3.8**/21.6)+0.2 = 0.86	
ls(ASP)*β_{255} = 2.5*0.86	= **2.15**
ls(THR) = -0.23	
β_{254}=0.8*(1-**6.1**/21.6)+0.2 = 0.77	
ls(THR)*β_{254} = -0.23*0.77	= **-0.18**
.....	
.....	
.....	
ls(PRO) = 1.2	
β_{254}=0.8*(1-**20.6**/21.6)+0.2 = 0.24	
ls(PRO)*β_{254} = 1.2*0.24	= **0.29**
Summation of scores	
PS(THR256)	= **0.39**

$$PS(r) = \sum_i \beta_i \cdot ls(r_i)$$

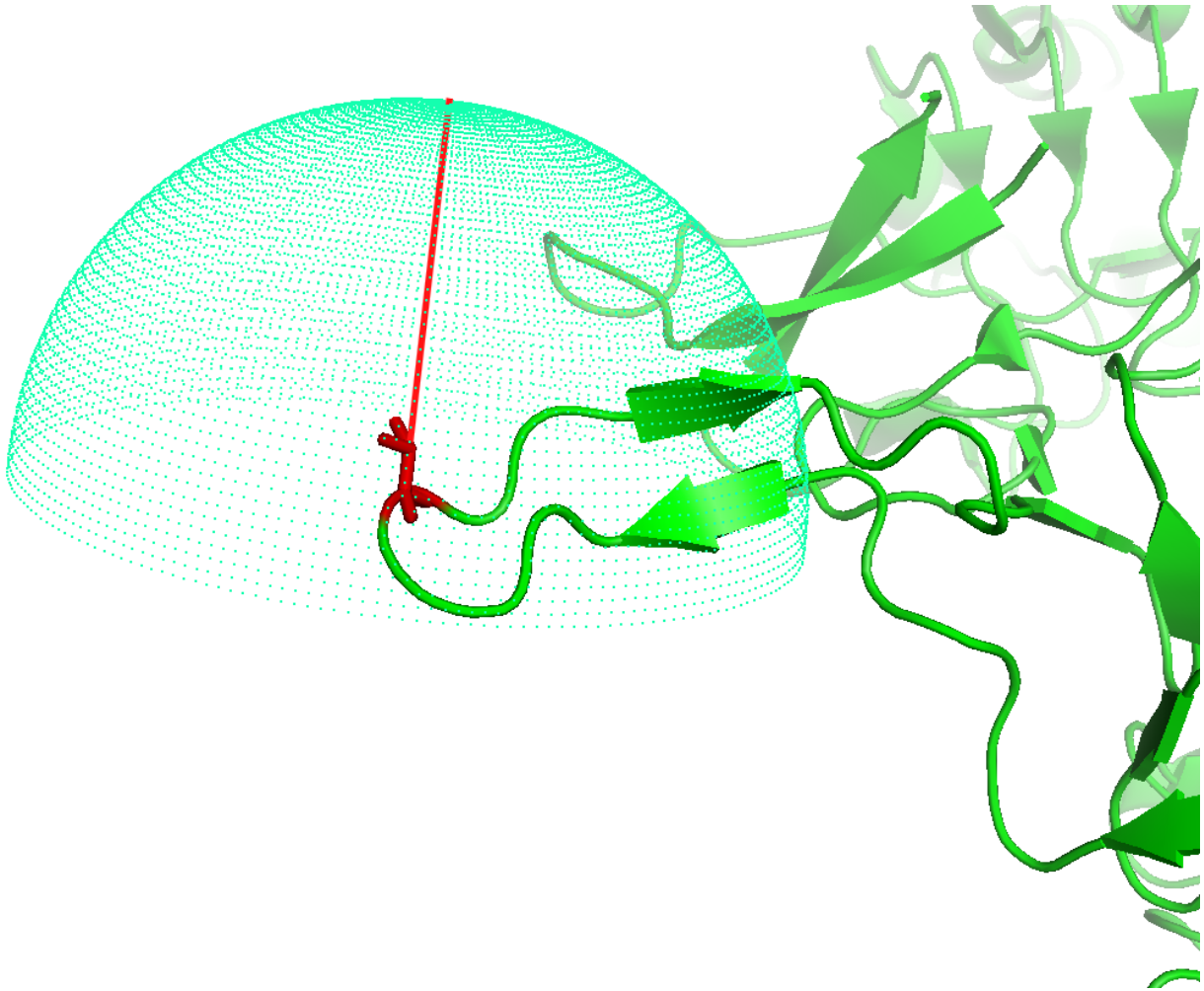
$$\beta_i = 0.8 * (1 - (d_i/21.6\text{Å})) + 0.2$$



Half-sphere exposure

HS(THR256)

1) Create the upper half-sphere, which is the half-sphere where the residue side-chain is pointing

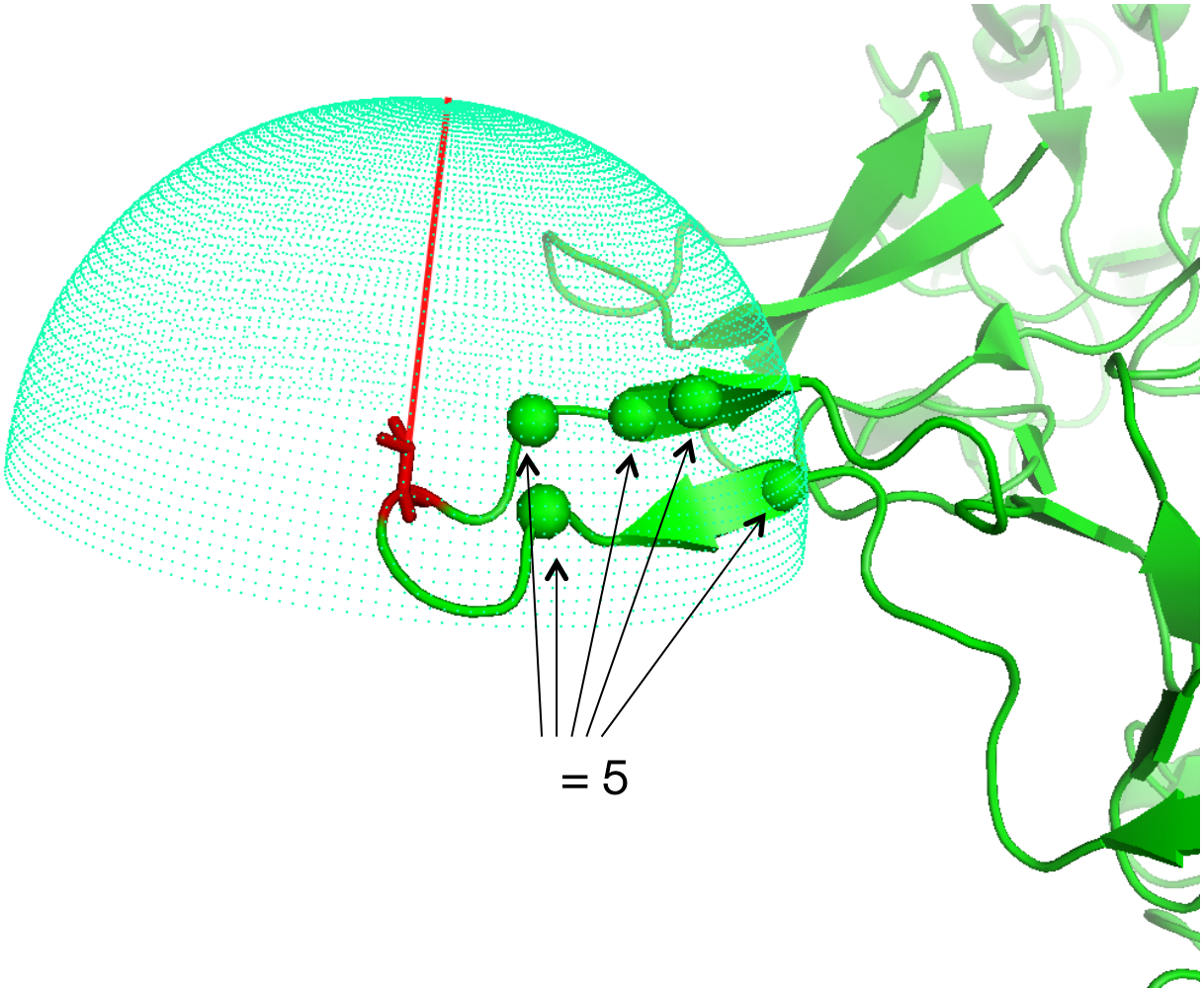


Half-sphere exposure

HS(THR256)

1) Create the upper half-sphere, which is the half-sphere where the residue side-chain is pointing
2) Count neighbor residues within the half-sphere (nr of C_α-atoms)
3) As high counts means highly buried the counts are multiplied by -1

HS(THR256) = -5



The DiscoTope 2.0 Score

DS(THR256)

1) Calculate Propensity score
2) Calculate half-sphere exposure
3) The final score is a weighted sum of the Propensity score and half-sphere exposure

PS(THR256) = **0.39**

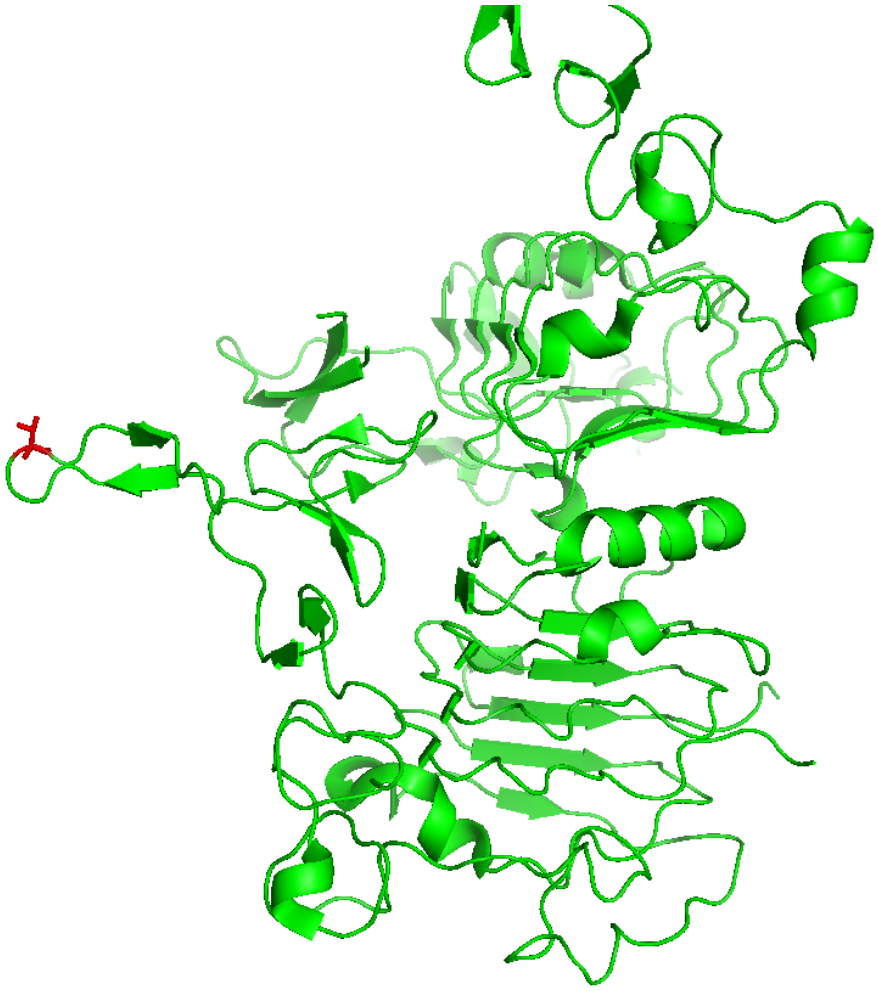
HS(THR256) = **-5**

α = 0.115

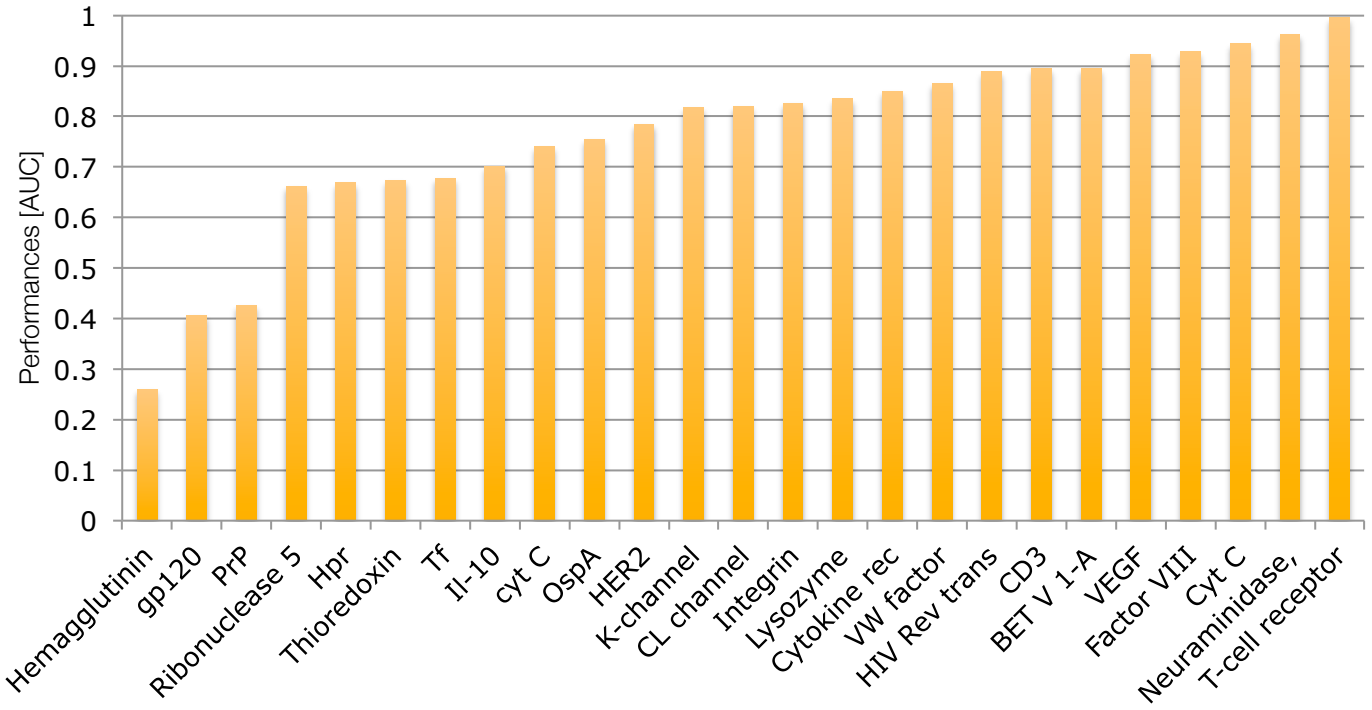
DS = (1-α)***PS** + α***HS**

DS(THR256) = 0.885*0.39 + 0.115*(-5)

DS(THR256) = **-0.23**

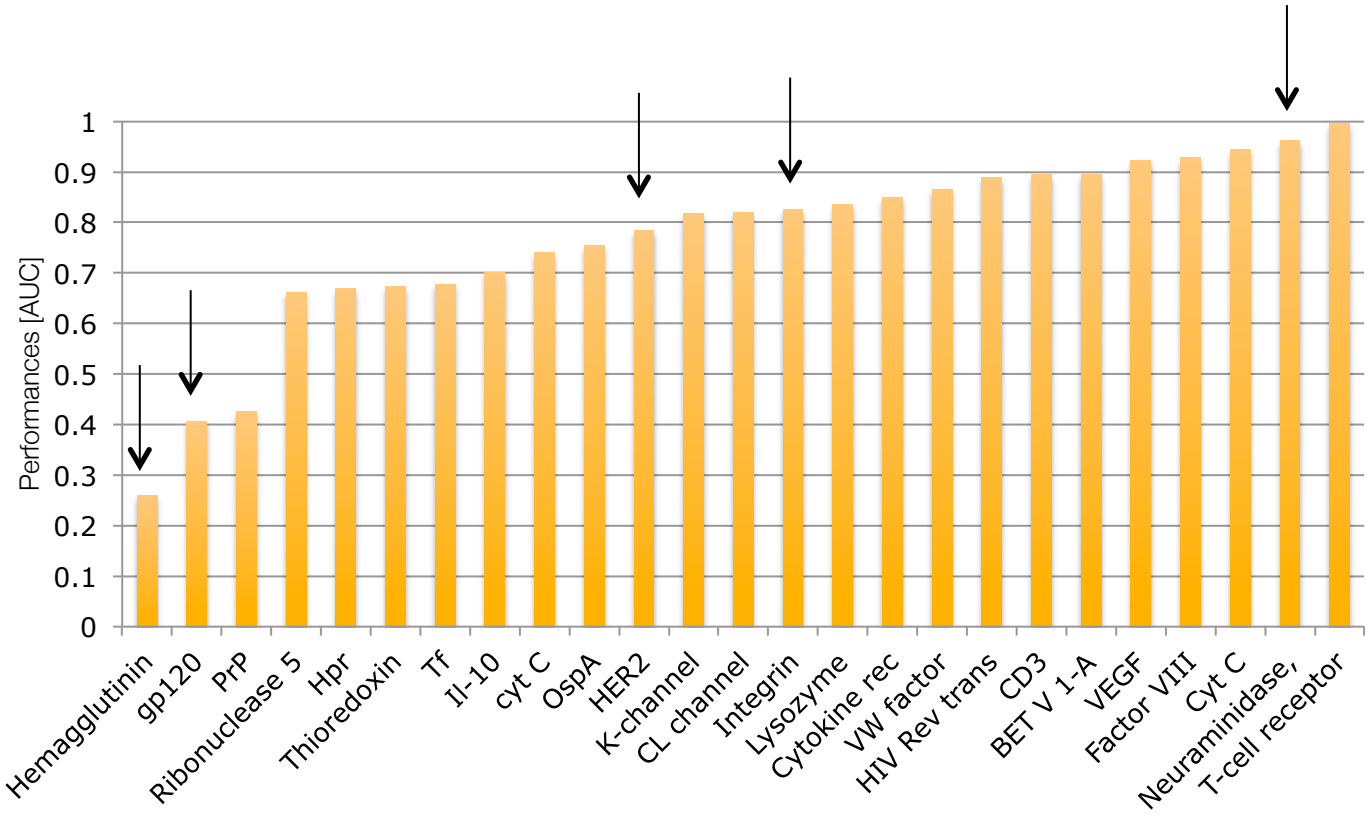


Performance and limitations



Average AUC = 0.741

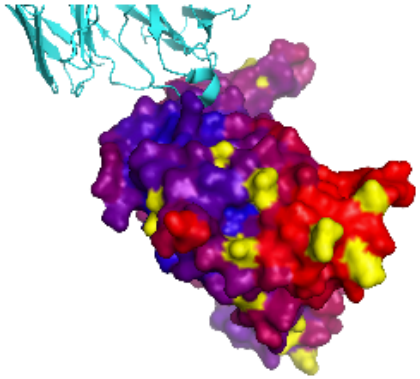
Performance and limitations



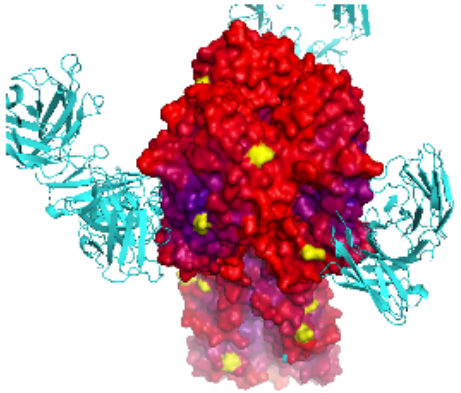
Glycosylated proteins

Performance and limitations

- Glycosylation effects predictions

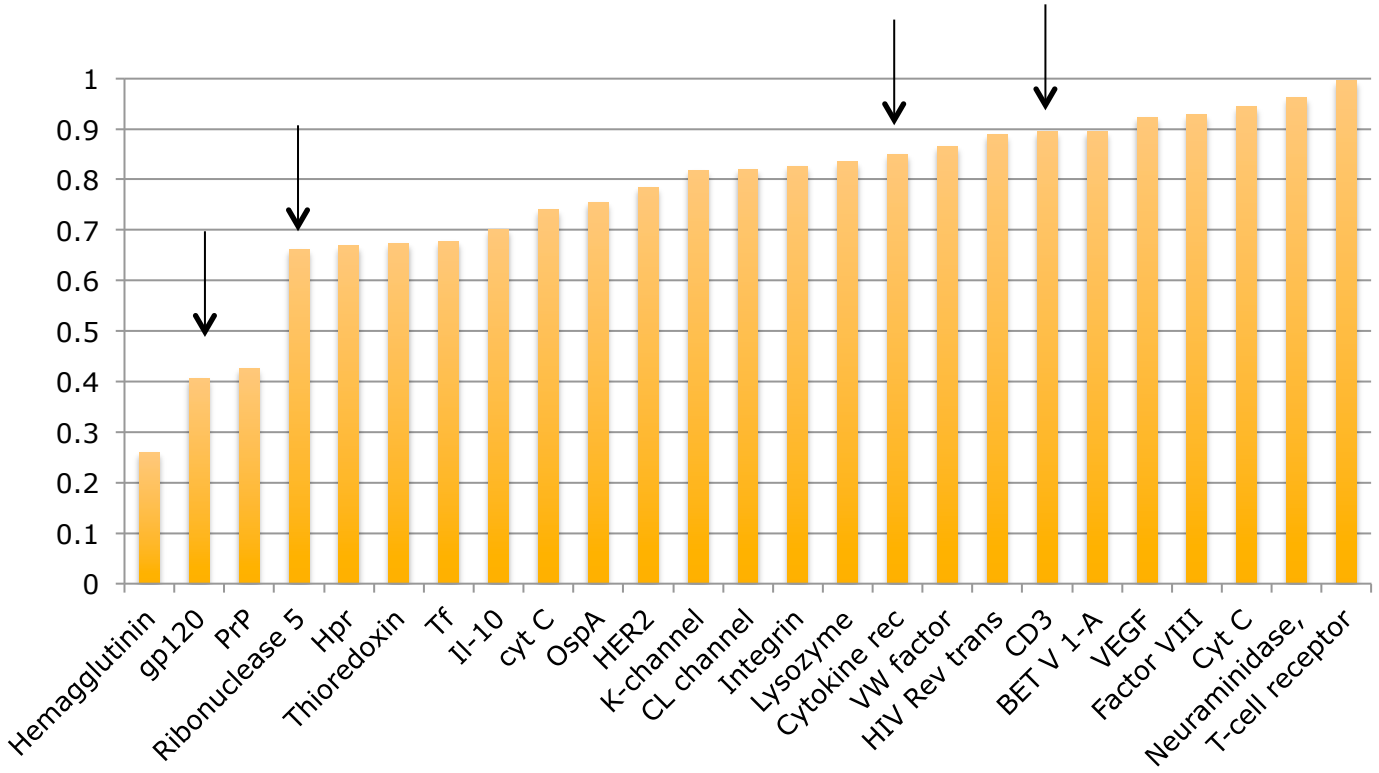


Gp120



Hemagglutinin

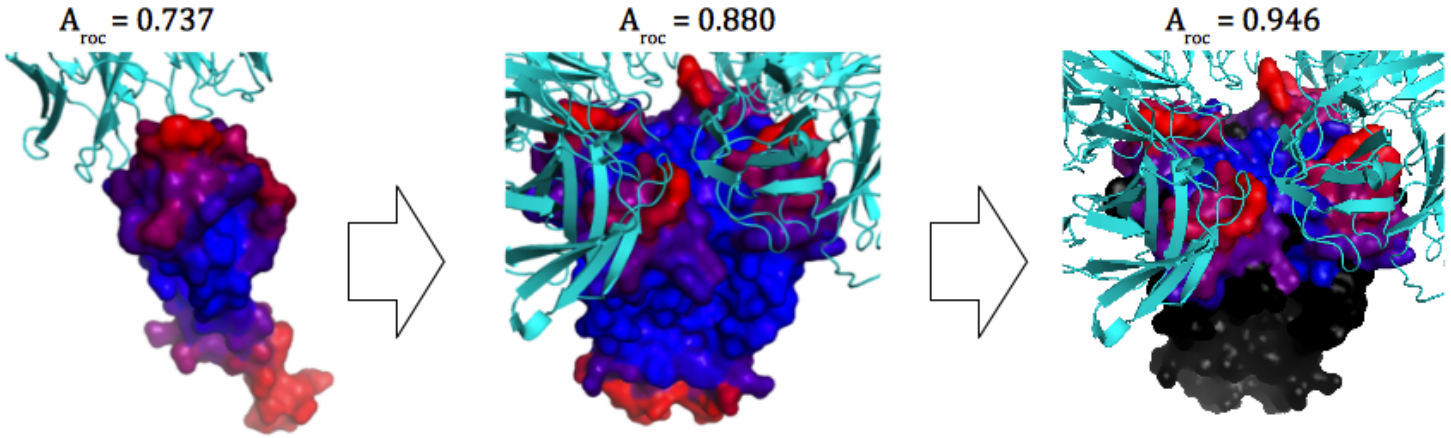
Performance and limitations



Small fragments (<120 residues) of larger biological units

Performance and limitations

- Inclusion of biological units enhance performance



Potassium Channel

Performance and limitations

- External Benchmark Dataset
    - 52 antigen:antibody structures
    - 33 homology groups
- Performance: 0.731 AUC

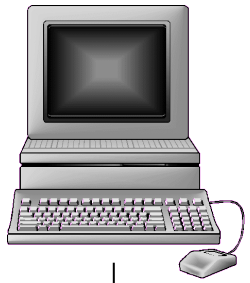
# Residues		DiscoTope-2.0	DiscoTope-1.2	PEPITO	ElliPro	SEPPA
15	PPV	0.178	0.176	0.169	0.134	0.142
	Sens	0.168	0.150	0.147	0.134	0.135
30	PPV	0.141	0.133	0.138	0.120	0.138
	Sens	0.262	0.220	0.237	0.234	0.257

Conclusions

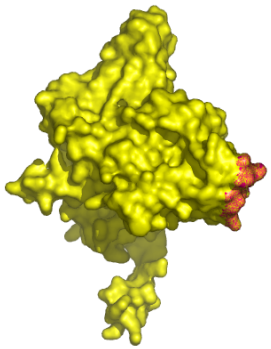
- DiscoTope V2.0 outperforms similar methods
  - High performance on 15/25
  - Medium performance on 7/25
  - Fail on 3/25
- Inclusion of surface measures does only slightly enhance predictions
- Use the entire biological unit, when possible
  - Small fragments (< 120 residues) have lower performance
- Glycosylation might course the prediction to fail
  - Check for clash between predicted epitopes and glycosylation sites

Rational vaccine design

>PATHOGEN PROTEIN
KVFGRCELAAAMKRHGLDNYRGYS
LGNWVCAAKFESNF

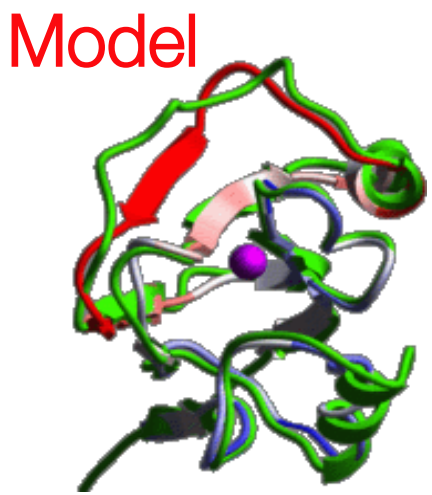


Rational Vaccine
Design



Rational B-cell epitope design

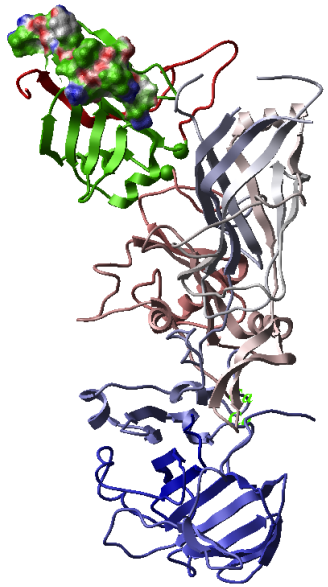
- Protein target choice
- Structural analysis of antigen



- 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

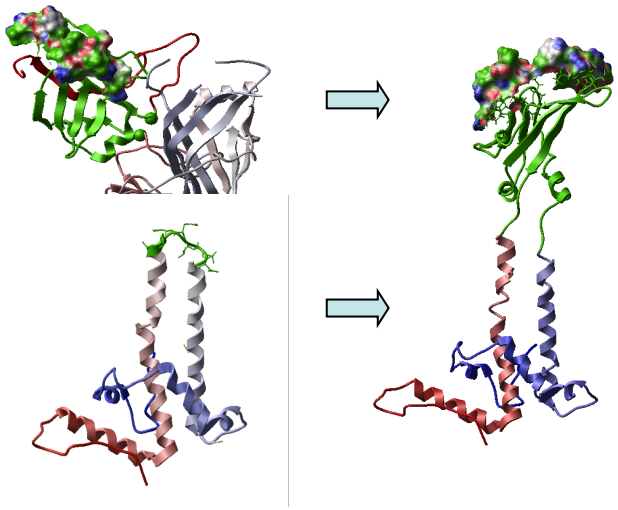
- Protein target choice
- Structural annotation
- Epitope prediction and ranking



- 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

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