Thesis Defense

Intercell develops *vaccines*  for the *prevention and treatment*  of *infectious diseases*  and cancer. Here we present you with some facts about our  *technologies*,  *research*, *ethics*,  *partnerships* and *investor relations*  .

Parametrization and Classification of Peptides for the Detection of Linear B-cell Epitopes

Johannes Söllner

Scientific question(s)

Can Machine Learning approaches improve state of the art B-cell epitope prediction for proteins/peptides?

Which features are most relevant to make a peptide antigenic?

Basic definitions

Antigen: “any substance (as a toxin or enzyme) that stimulates the production of antibodies

Epitope: “a molecule or portion of a molecule capable of binding to the combining site of an antibody. ...“ (B-cell epitopes)

Structural epitopes: Binding surfaces formed by amino acids non-adjacent on the sequence level. (3D-conformation).

Linear (sequential) epitopes: formed by adjacent amino acids

We only investigated predictions neglecting 3D-structure i.e. **only linear epitopes**

State of the Art

1. Published mainly in 1980s an 1990s

2. Almost exclusively monoparameteric propensity scales (hydrophilicity, secondary structure prevalence, ...)

3. No classification systems except for the “gold standard“ by Kolaskar and Tongaonkar (1990)

Machine learning - principle

Knowledge Based

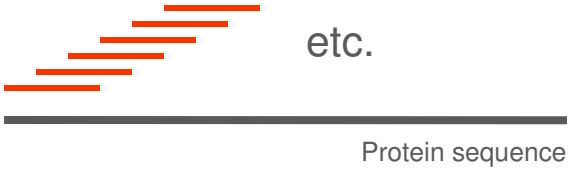
And it works like this:

2. See, this is an antigen
3. See, this is NOT an antigen

And see, here I have another item:
Is this an antigen?

Translation to practice

1. Split protein to be investigated in overlapping peptides



2. Calculate descriptors (values) for each peptide
3. Compare to descriptors of known epitopes an non-epitopes by means of machine learning
4. Assign the class (i.e. epitopic/non-epitopic) of the most similar examples.

What are peptide descriptors/attributes/parameters?

Peptide Parameters

1. Published propensity scales (hydrophilicity, flexibility,...) → **Protscale**
2. MOE based (small molecule attributes for amino acids)
3. Neighbourhood parameters
4. Amino acid frequencies (likelihood to be member of an epitope)
5. Complexity parameters

⇒ Calculated mean values for each peptide

Focus: Neighbourhoods

Definition of Neighbourhood

Word m neighbours a central character *i* in a distance of *x* characters if it can be found in the same string x characters upstram AND/OR downstream of character *i*. (non exclusive OR)

e.g. For *m* = AF, *i* = G, *x* = 3
AAGGIAFVAHGCKLAFAIAIAGGATGHA

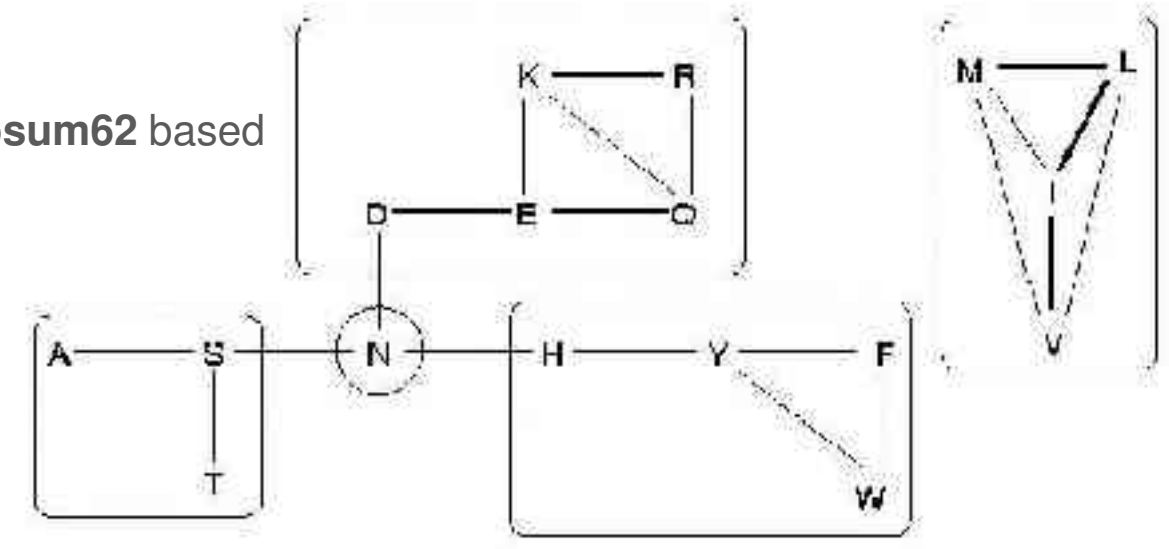
Calculated Probabilites

- Upstream
- Downstream
- Combined
- Mean between Upstream and Downstream

Reduced alphabets (for neighbourhoods)

Treat several amino acids as one – based e.g. on chemical similarity, hydrophilicity, flexibility, secondary structure prevalence, ...

Blosum62 based



- Functional abstractions are possible
- Statistical basis of neighbourhoods improves

Total number of parameters

In total **18920** attributes

⇒ Selection necessary

The dataset

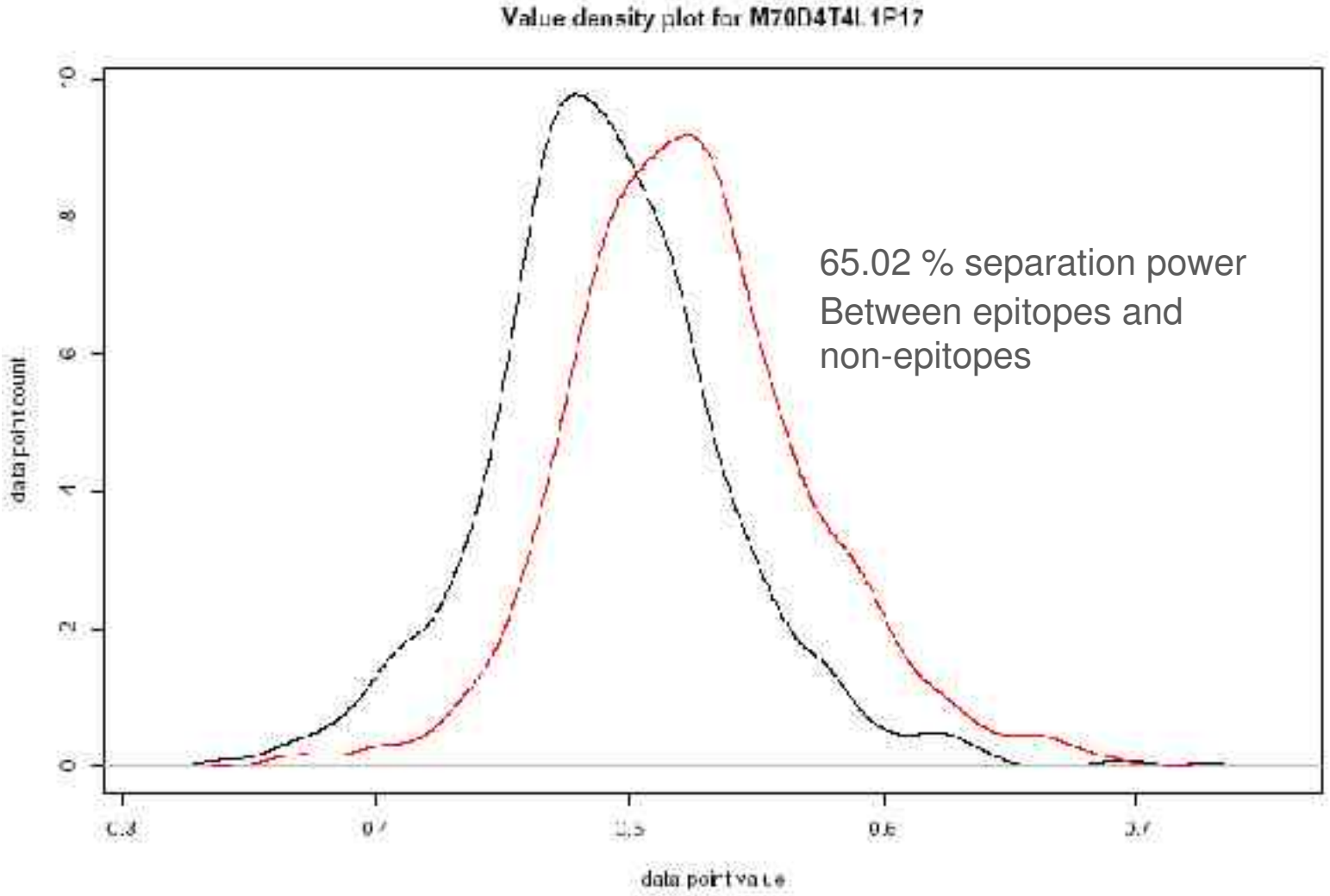
Machine learning is empirical i.e. examples have to be provided

1. BciPep: 211 epitopes (multi organism)
2. FIMM: 293
3. Intercell Antigen Identification Program (AIP): 1333 (13 bacterial pathogens)

Reference (non immunogenic) peptides:

- Excised from AIP unselected proteins but screened organisms
- Randomly excised from proteins
- Excised peptides had the same length as epitopes
- Balanced, i.e. The same number as epitopes and references

Separation Power of Parameters



Parameter Ranking

alphabet	type	name	mc	sd_mc	mcv	sd_mcv
unreduced	D	M70D4T4L1P22	65.29	1.29	65.07	2.44
unreduced	D	M66D3T3L1P22	65.03	1.30	64.79	2.38
unreduced	D	M70D4T4L1P17	65.02	1.25	64.94	2.50
unreduced	D	M70D4T4L1P19	64.97	1.23	64.84	2.46
unreduced	D	M66D3T3L1P17	64.71	1.18	64.71	2.48
-	E	AAfrq3.1	64.66	1.31	64.44	2.40
unreduced	D	M66D3T3L1P19	64.66	1.25	64.44	2.50
unreduced	D	M63D2T2L1P22	64.62	1.35	64.47	2.44
alphabet 3	D	M250D4T4L1P17	64.61	1.35	64.79	2.38
unreduced	D	M69D4T3L1P22	64.57	1.33	64.56	2.66
alphabet 4	D	M340D4T4L1P17	64.56	1.34	64.74	2.51
unreduced	D	M66D3T3L1P21	64.52	1.26	64.35	2.32
alphabet 3	D	M246D3T3L1P17	64.43	1.38	64.63	2.43
unreduced	D	M70D4T4L1P21	64.38	1.31	64.17	2.54
alphabet 4	D	M336D3T3L1P17	64.36	1.36	64.40	2.57
alphabet 3	D	M249D4T3L1P17	64.20	1.39	64.23	2.22
unreduced	D	M63D2T2L1P17	64.20	1.22	64.46	2.66
alphabet 3	D	M243D2T2L1P17	64.18	1.40	64.39	2.56
alphabet 4	D	M339D4T3L1P17	64.17	1.37	64.27	2.42
unreduced	D	M69D4T3L1P19	64.16	1.32	63.93	2.50
alphabet 4	D	M333D2T2L1P17	64.15	1.43	64.30	2.49
unreduced	D	M63D2T2L1P19	64.14	1.23	64.30	2.62
unreduced	D	M69D4T3L1P17	64.13	1.28	64.06	2.64
unreduced	D	M63D2T2L1P21	64.06	1.31	64.20	2.70

mc mean correct
sd standard deviation
mcv mean correct (validation)

Physico-Chemical Parameters

Neighbourhood parameters are at least equally prominent

The best physico-chemical parameters

REF	NAME	PARAMETER TYPE
	AAfrq3.1	Difference frequency
[41]	M&J.1	Contact energy
[36]	M&P.1	Hydrophobicity
[10]	Delage&Roux2.1	beta-turn
[19]	Guy.1	Hydrophobicity
MOE	SlogP_VSA4.1	Subdivided Surface Area
[18]	Rose&Zehfus.1	Mean fractional area loss
[37]	Lifson&Sander2.1	Total beta-strand
[10]	Delage&Roux3.1	Beta-sheet
MOE	MACCS(115).1	I,L,M = 1
MOE	SlogP_VSA1.3	Subdivided Surface Area
[37]	Lifson&Sander3.1	Parallel beta-strand
MOE	MACCS(149).1	I,L,V = 1
[41]	M&J.3	Contact energy
[36]	M&P.3	Hydrophobicity

MC – Mean value of correct classifications

Parameter Selection

- Too many parameters for practical application
- Selection of complementary and powerful parameters
- The Waikato Environment for Knowledge Analysis (WEKA)
- Based selection on multiple samples

Parameter Selection

GenericSearch	BestFirst
M63D2T2L1P3 (81/109)	vsa_don2 (91/109)
Delsage&Roux3.1 (79/109)	M63D2T2L1P4 (78/109)
M78D4T3L1P3 (78/109)	M63D2T2L1P3 (76/109)
M341D1T1L2P27 (78/109)	SMR_VSA6.1 (68/109)
M63D2T2L1P1 (78/109)	M78D4T3L1P1 (62/109)
M66D3T3L1P2 (78/109)	kS_ssNH 1 (58/109)
M66D3T3L1P3 (77/109)	M151D1T1L1P17 (57/109)
M78D4T3L1P17 (77/109)	AAfrq3.3 (55/109)
SlogP_VSA3.1 (75/109)	AAfrc3.3 (55/109)
M69D4T3L1P17 (75/109)	Levit3.1 (55/109)
Levitt3.1 (75/109)	M166D3T3L2P28 (54/109)
a_base.2 (74/109)	M165D3T2L2P1 (54/109)
M63D2T2L1P17 (74/109)	M68D4T2L1P4 (52/109)
Chou&Cosman.1 (73/109)	M61D1T1L1P18 (50/109)
M78D4T3L1P1 (73/109)	M66D3T3L1P2 (50/109)
M66D3T3L1P17 (73/109)	Levit2.3 (48/109)
M68D4T2L1P22 (73/109)	PEOE_VSA_PPOS.2 (47/109)
vsa_don.2 (73/109)	M427D4T1L1P17 (47/109)
SlogP_VSA3.3 (72/109)	M242D2T1L1P1 (46/109)
M69D4T3L1P3 (72/109)	Delsage&Roux2.1 (45/109)
M78D4T3L1P2 (72/109)	M78D4T3L1P2 (44/109)
PEOE_VSA_PPOS.2 (72/109)	Delsage&Roux2.3 (44/109)
M66D3T3L1P1 (71/109)	M78D4T3L1P2 (44/109)
Jonin2.1 (71/109)	Delsage&Roux2.3 (44/109)
SMR_VSA6.1 (71/109)	M163D2T2L2P3 (44/109)

Creation of Classifiers

What is a Classifier?

A Black box telling (predicting) the class of a data item

- 14 different parameter selections, 4 – 80 members per set
- Dimension reduction through PCA (Principal Component Analysis)

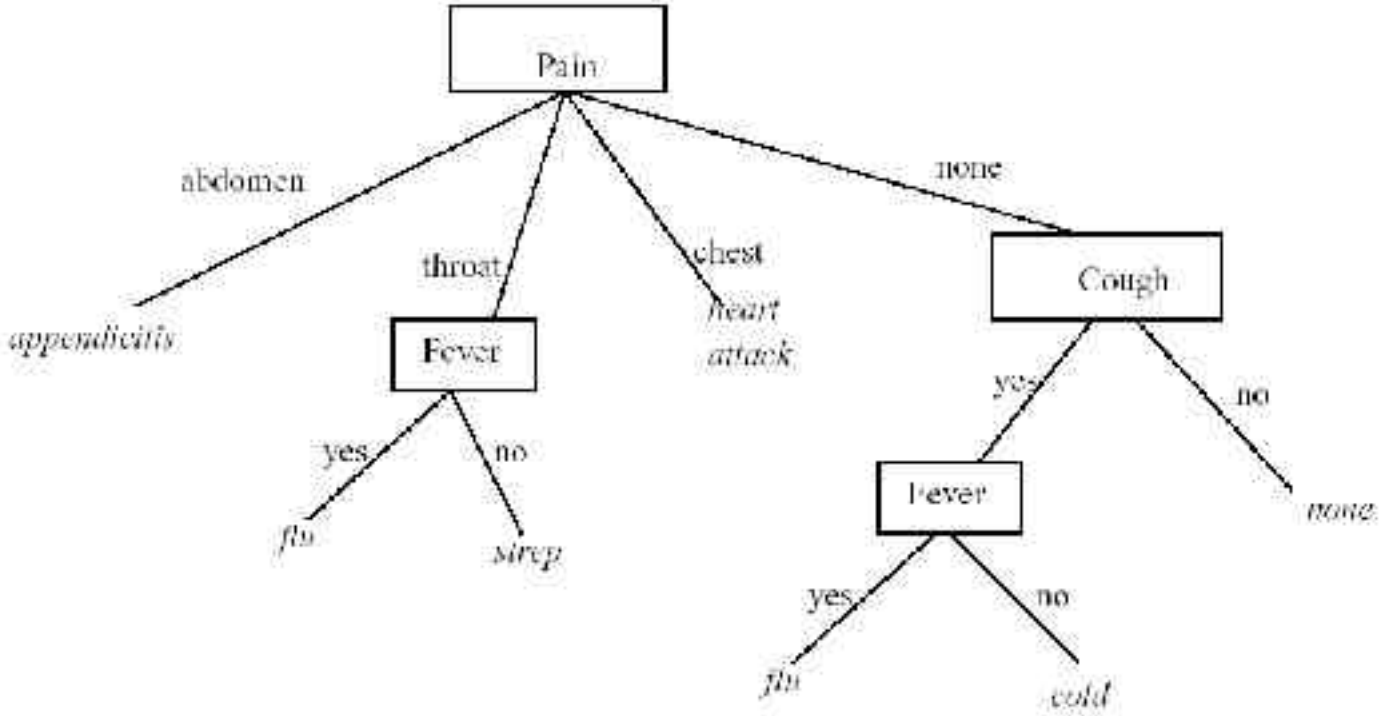
Used classification algorithms

1. J48/C4.5 (decision trees)
2. IBk (nearest neighbours classifiers)

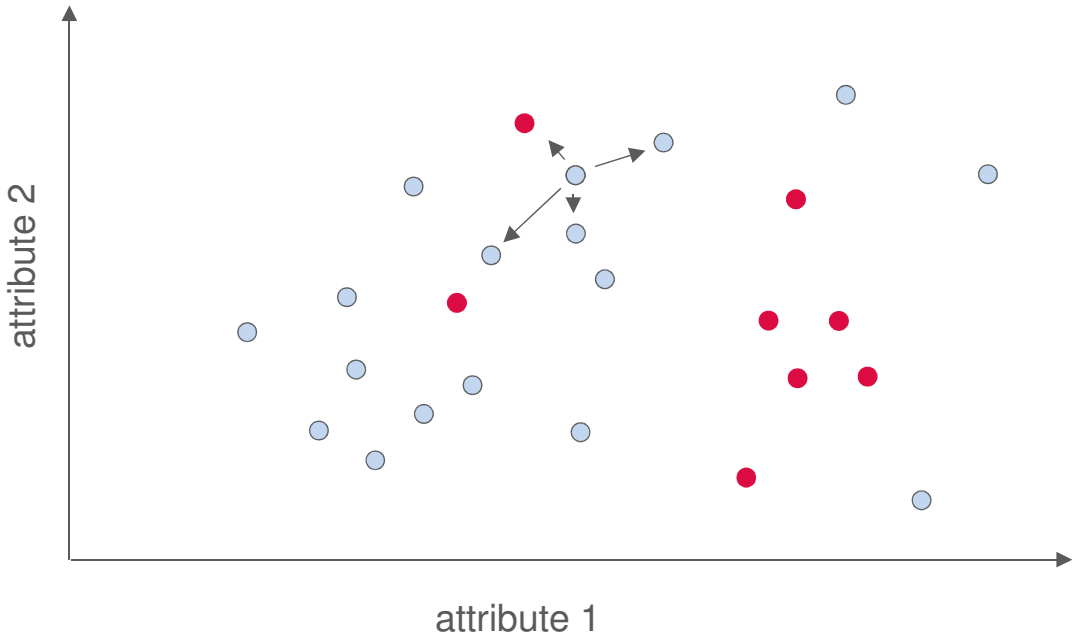
More powerful methods exist (e.g. Neural networks),
but lead to less understandable models.

For each of the two algorithms, different setups (e.g. algorithm options) were possible,
resulting in more than 10^6 setups.

Decision Trees



IBk (k-nearest neighbour or lazy learning)



The Confusion Matrix

The rates of True and False classifications have to be taken into consideration. One way is the representation in a confusion matrix.

	predictions **P**	predictions **N**	
Known classes **P**	TP	FN	**P**
Known classes **N**	FP	TN	**N**

P postives (e.g. epitopes)
N negatives (e.g. non-epitopes)

TP True Positive
TN True Negatives

FP False Positives
FN False Negatives

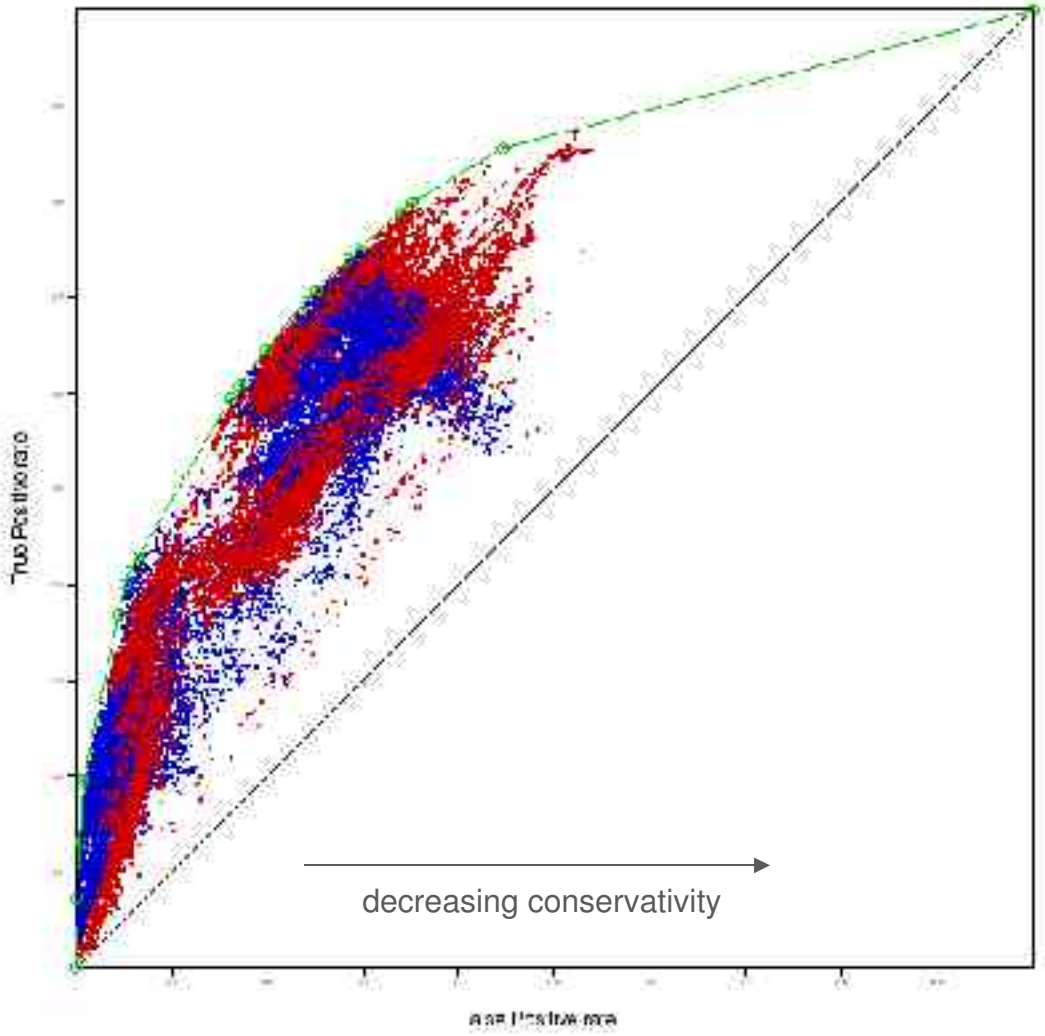
Finding Optimal Classifiers

$$\mathrm{TPrate} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{FPrate} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$

- Can be used to obtain an ROC (Receiver Operator Characteristic) graph
- Can be used to find best setups by means of the ROC convex hull

ROC Convex Hull



All used setups in an ROC graph. The green curve forms the ROC convex hull.

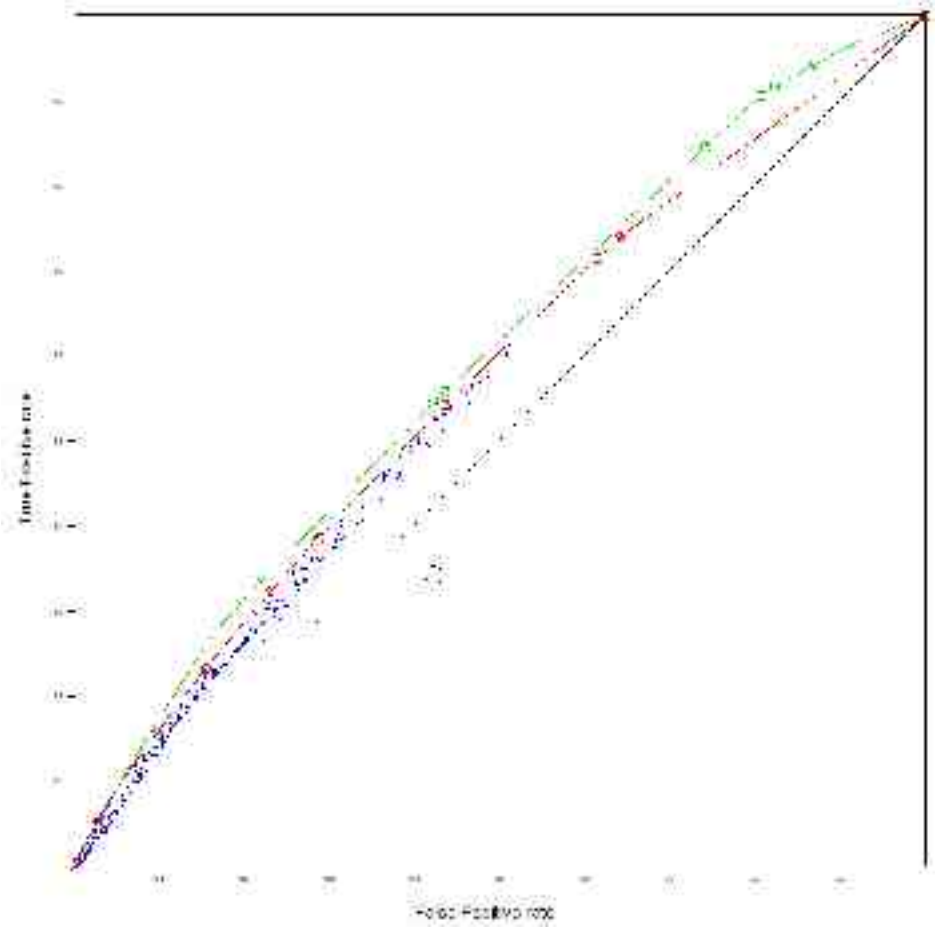
Classifiers forming the hull are by definition optimal

Classifiers

- 17 Classifiers selected from 1164041
- Were combined by voting (ensemble learning)
- Validation on an independent dataset
- Comparison to the gold standard

Validation

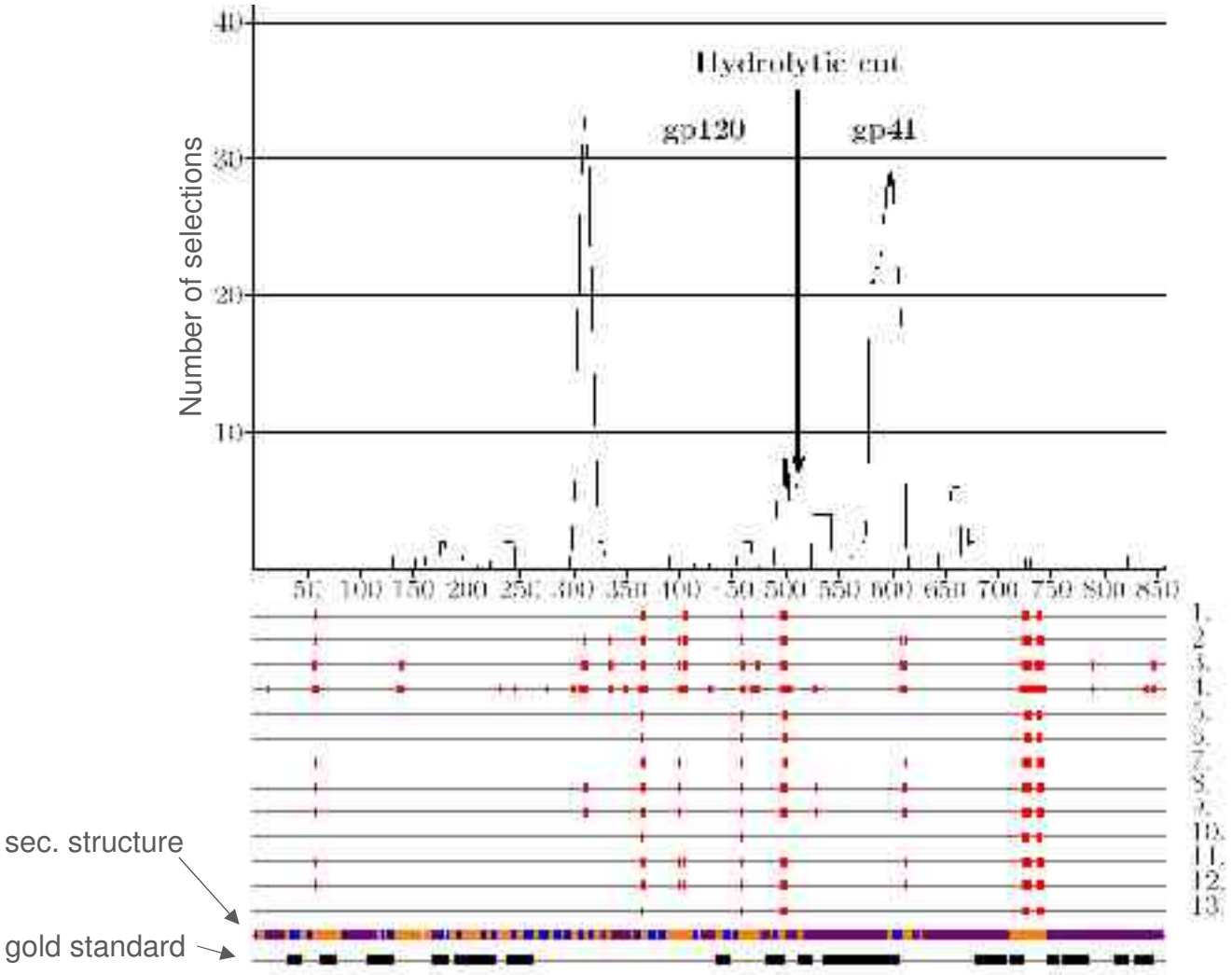
For method validation, a dataset of overlapping peptides from 7 published proteins[1] was assembled ⇒ 4923 peptides of determined antigenicity



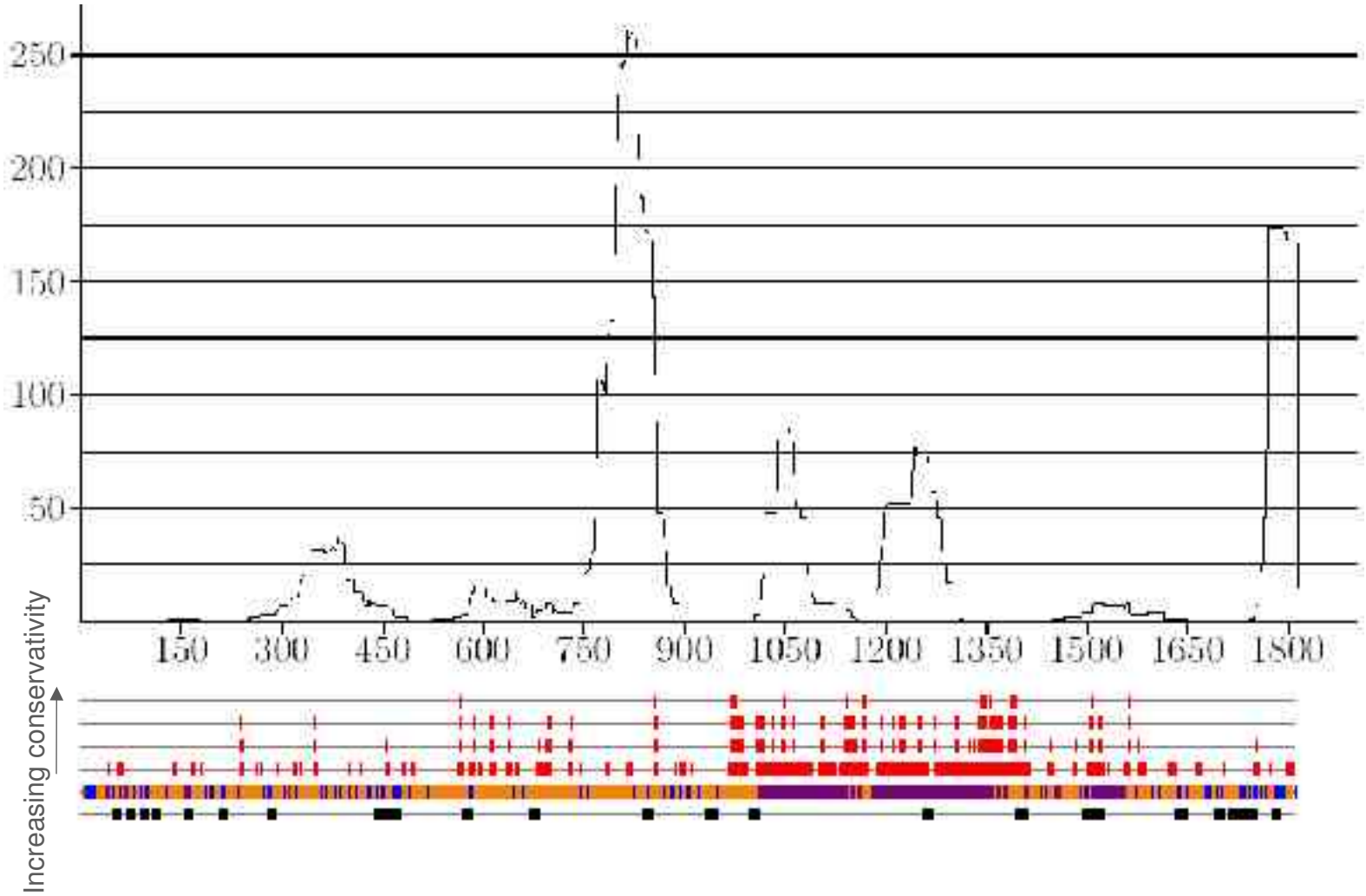
- All our classifiers are above the first meridian
- The gold standard by Kolaskar and Tongaonkar is not above the first meridian

([1]HIV gp160, S. Mutans SpaP, HCV E1, HCV E2, N. meningitidis p64K, SARS E2, HHV-5 UL32 aa490-660)

HIV gp160 – ensemble classifiers



AIP data


250
200
150
100
50
150
300
450
600
750
900
1050
1200
1350
1500
1650
1800
Increasing conservativity

Summary

- Compared new and published parameters
- Determined optimal single parameters
- Selected parameter sets for machine learning
- Evaluated two machine learning strategies

- Neighbourhood parameters can be useful for antigenicity analysis
- BLOSUM62 based alphabets are similarly powerful as ureduced alphabets
- Contact Energy, Hydrophobicity, beta-turn, beta-sheet, accessible surface area and van der Waals potential
- Machine learning classifiers are consistenly better than the gold standard

Acknowledgements

Intercell
Martin Hafner
Jörg Fritz
Alexander von Gabain

Biovertis
Uwe von Ahsen

Emergentec
Bernd Mayer
Martin Willensdorfer

University of Leipzig
Peter F. Stadler