

Prediction of Peptide Activity using Bayesian Learning

Ton van Daelen, Robert Brown,

David Rogers

SciTegic, Inc., San Diego, CA

Outline

- Goals of this project
- Bayesian learning
 - Descriptors for peptides
- MHC peptide data set
 - Validation of the model
 - Scoring virtual peptide libraries
- Conclusions

Goals

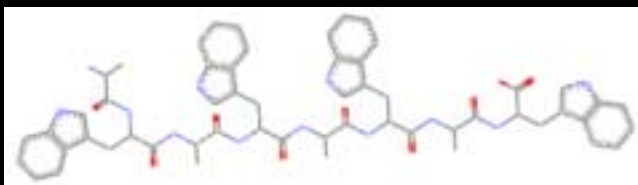
- Relate peptide immunogenic potency to peptide sequence, using a frequency based approach
 - Investigate applicability of Bayesian statistics
 - Generate suitable descriptor set
- “Propose” novel peptides with high probability of exhibiting immunogenic potency

Bayesian Classification

- Build a model which estimates the likelihood that a given data sample is from a "good" subset of a larger set of samples
- Characteristics
 - Relative predictor
 - Good with high-dimensional data
 - Scales linearly; efficient for large data sets
 - Works for a *few* as well as *many* 'good' examples
 - Can model broad classes of compounds (i.e. multiple modes of action)

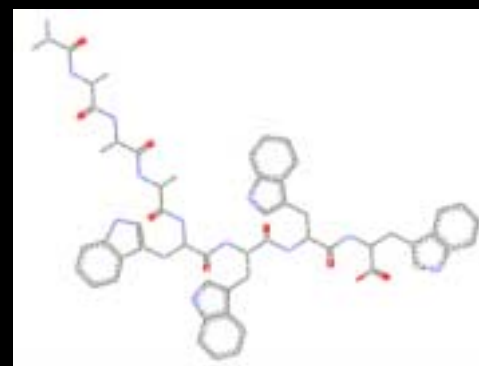
Bayesian Modeling Applied to Chemical Structure

- Choose appropriate descriptors
 - Continuous (e.g. molweight, LogP, #donors, #Acceptors)
 - Structural fingerprints
- This doesn't work well for peptides
 - Molecular fingerprints describe presence and absence of substructural features, not their arrangement with respect to one another



AAAAWWWW

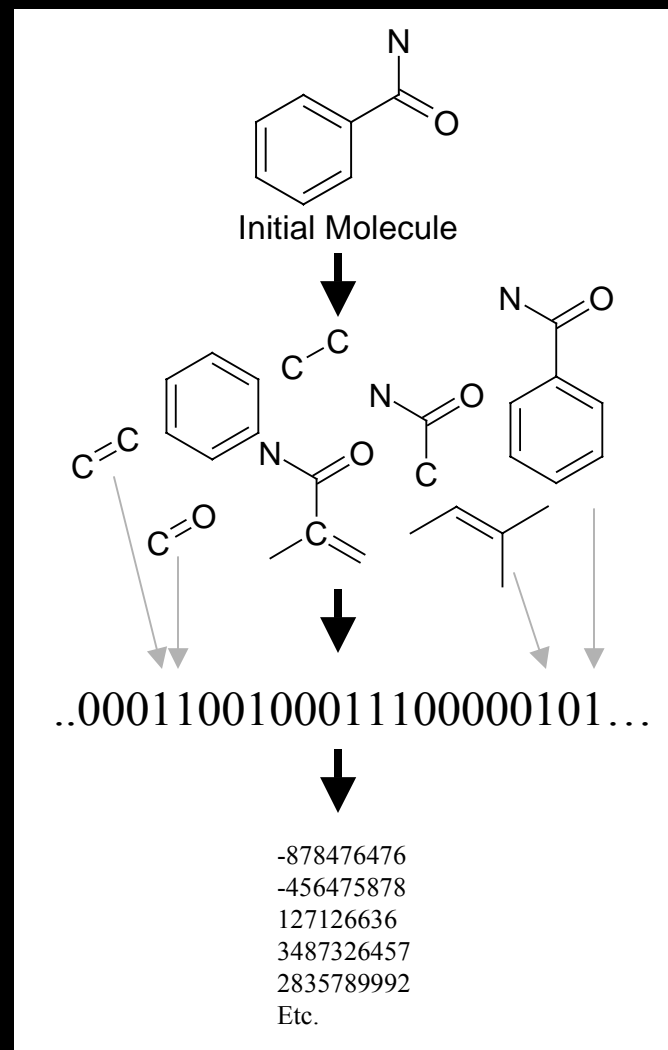
≠



AWAWAWAW

Generating Molecular Fingerprints

- Structure is fragmented according to a particular set of rules
- Each fragment generates a code that represents an 'on' bit in the final bit string
- Fingerprint bits indicate presence and absence of certain structural features



Bayesian Statistics Applied to Non-Chemistry Data

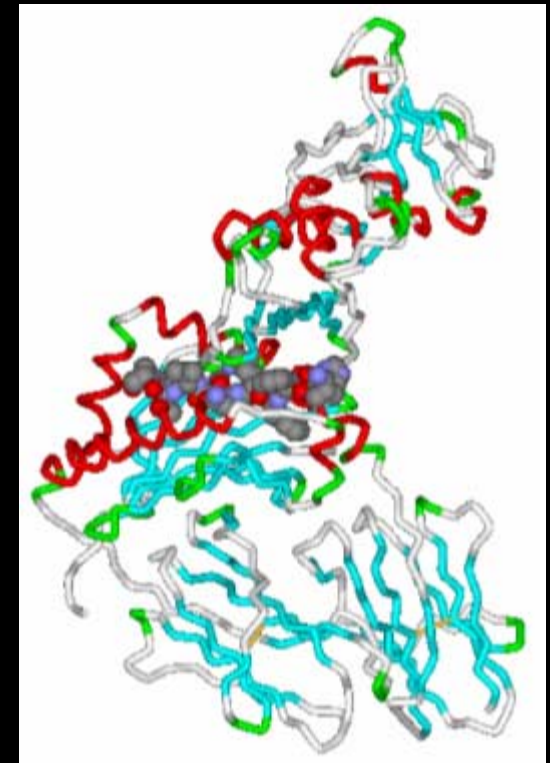
- Sequence based descriptors
- Procedure
 - For a given sequence, select all consecutive amino-acid codes of size 1 .. N
 - N is typically between 3 and 6
 - Keep only unique examples
- Example
 - Simple AA sequence
 - VDAELEN
 - Text descriptor (fingerprint) of maximum length 4 has 21 ‘bits’
 - V;VD;VDA;VDAE;D;DA;DAE;DAEL;A;AE;AEL;AELE;E;EL;ELE;ELEN;L;LE;LEN;EN;N

Derived Fingerprints

- Goal
 - Reduced size of fingerprint space supporting smaller data sets
 - For interpreting fundamental properties of the sequence
- Hydrophobicity
 - Translate each AA into Polar and A-Polar
 - Example (max length 4):
 - Raw sequence: **VDAELEN**
 - Pol. sequence: **APAPAPP**
 - Pol. fingerprint: **A;AP;APA;APAP;P;PA;PAP;PAPA;PAPP;APP;PP**
- Chemical class
 - Translate each AA into **A**liphatic, **C**yclic, Aromatic (**E**), Hydroxyl or **S**ulphur groups, **b**asic, acidic/amides (**O**)
 - Example (max length 4):
 - Raw sequence: **VDAELEN**
 - Chem. sequence: **AOAOAOO**
 - Chem. fingerprint: **A;AO;AOA;AOAO;O;OA;OAO;OAOA;OAOO;AOO;OO**

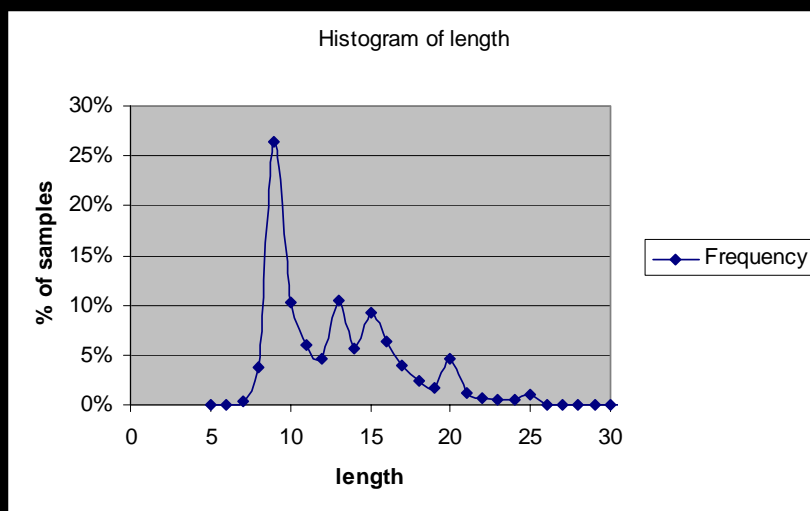
MHC-Binding Peptides

- Major Histocompatibility Complex (MHC) is a large complex of tightly linked genes that encodes molecules involved in many aspects of the immune response
- MHC is the most polymorphic gene complex known (coding to Class I and Class II MHC proteins)
- The proteins expressed by the MHC genes are the receptors for large virus and bacteria proteins. The peptides in our study represent fragments of the virus/bacteria that bind to the receptor



MHC Peptide Data Set

- 13,362 peptides binding to human and mouse MHC proteins
- Most peptides between 8 and 16 amino acids in length



source	mhc	Frequency	# duplicates
human	HLA-A2	962	334
human	HLA-DR4	924	263
human	HLA-DR1	715	57
mouse	I-Ad	418	
human	HLA-B27	348	90
mouse	I-Ek	335	
human	HLA-DR2	324	1
mouse	I-Ed	300	
human	HLA-DR11	295	184
mouse	H-2Kd	293	1
human	HLA-DR7	283	77
mouse	I-Ak	276	1
human	HLA-A3	260	16
human	HLA-B35	226	9
human	HLA-A11	211	24
human	HLA-DR5	200	
mouse	H-2Kb	196	2
human	HLA-DR3	189	3
human	HLA-B8	185	
mouse	H-2Kk	183	
human	HLA-B7	176	8
mouse	I-Ag7	171	
mouse	H-2Db	166	
human	HLA-D?	165	
human	HLA-DR17	154	7
human	HLA-A24	149	2
human	HLA-?	143	6

Immunogenic Potency Data

- Immunogenic potency (activity) determined by T-cell based assays
- Activity is reported as PD50 (peptide dose 50)
 - Dose giving 50% lysis (cell destruction)

PD50	Activity	# samples
< 1nm	High	556
10 μm-1μm	Moderate	1100
> 10 μm	Inactive	379
	Unknown	11327

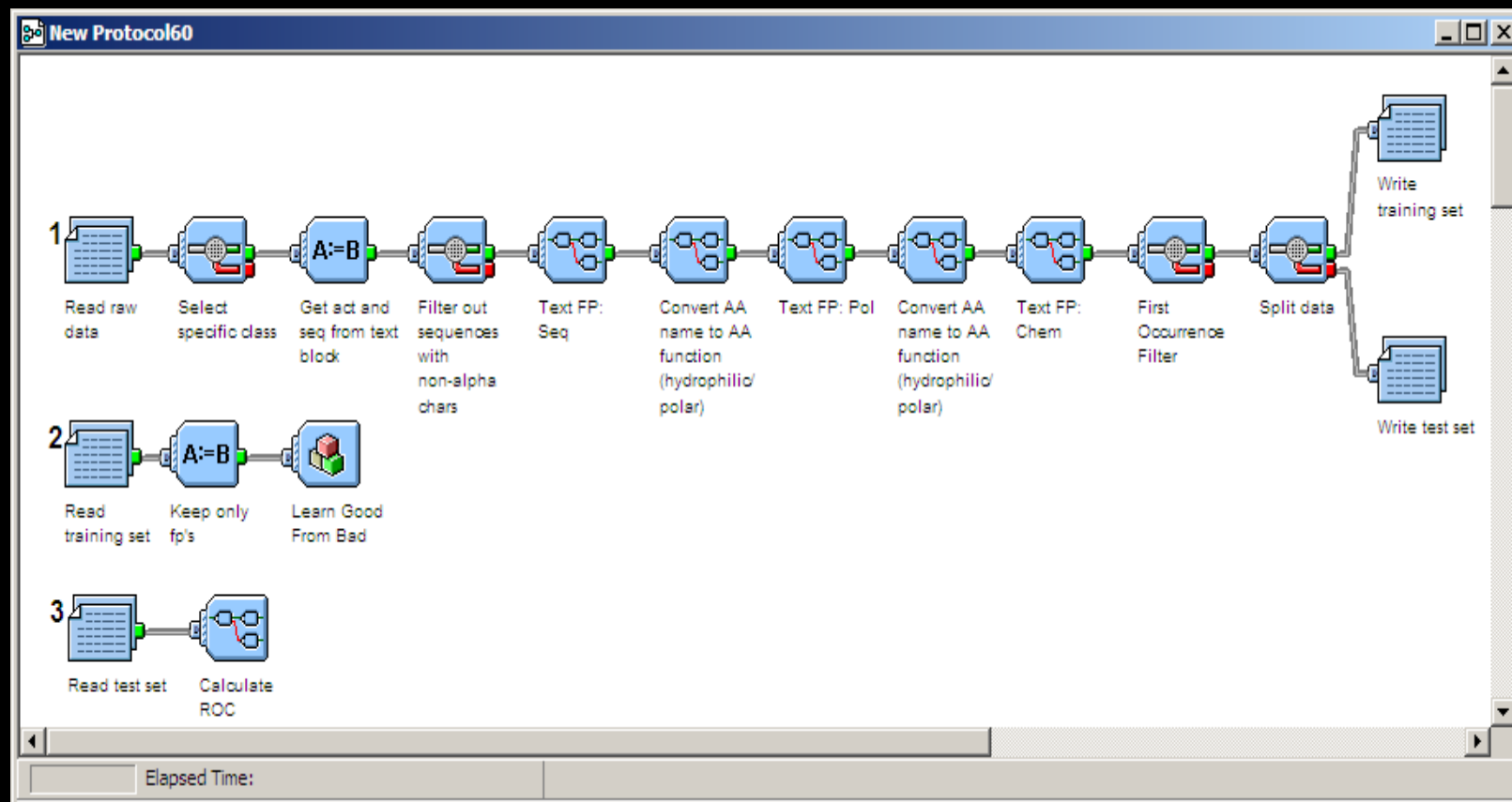
Source	Sequence	MHC	Activity
human	KGILGIVFTLTV	HLA-A2	Unknown
human	FSHDYRGSTSHRL	HLA-DR7	Unknown
human	EREEALTTNVWIEMQWCDYR	HLA-DR2	Unknown
human	EELAKQAEELAKL	HLA-DP9	Unknown
human	HDYNFVKAINAIQKSW	HLA-DR11	Unknown
human	MGLLECCARCLVGAPFASLV	HLA-DR51	Unknown
human	FLHSGTAKSV	HLA-A2	Unknown
mouse	AEGASYTVANKAKGIT	I-As	High
human	WWAGVGRAR	HLA-DR11	Unknown
human	LXRGSMXGL	HLA-B7	Unknown
human	LPCRIKQII	HLA-B51	Unknown
human	QYIKAQSKFIGITE	HLA-DR7	Unknown
human	IPLPKQYQPY	HLA-B35	Unknown
human	EYKLVVVGADGVGKSALT	HLA-DR15	Unknown
human	CFLAMLSLFICGTAGIFLMA	HLA-DR4	Unknown
mouse	YAATASTMDHARHGFLPRHRD	I-Ag7	Unknown
human	RLVTLKDIV	HLA-A2	Unknown
human	VHFFKNIVAPRTP	HLA-DR15	Unknown
human	XLDSDXERL	HLA-A2	Unknown
human	LIGFRKEIGRMLNI	HLA-DR1	Unknown
human	ASCIGLIY	HLA-B35	High
mouse	FEDTGNLI	H-2Kk	Unknown
human	RRYQKSTVL	HLA-B27	Inactive
mouse	FESNFATQATNR	I-Ak	Unknown
mouse	QADHAAHAEIN	I-Ad	Unknown
human	APRSNGMVX	HLA-B7	Unknown
mouse	DLIAYLKQQTK	I-Ek	Inactive
human	SRYWAIRTRSGGI	HLA-DR7	Unknown
human	MLDSVPLLLG	HLA-A2	Unknown

Raw Data Example

```
>HUM1014A#
MHC MOLECULE: HLA-A2, CLASS-1, (HUMAN)#
METHOD: cytotoxicity as.#
ACTIVITY: yes, little#
BINDING: yes, ?#
SOURCE: Influenza matrix protein (53-68)#
DB REFERENCE: SWISS: (VMT1_I ACKB, VMT1_I AANN, VMT1_I AFOW, VMT1_I AWIL, VMT1_I AMAN, #
& VMT1_I ABAN, VMT1_I ALE1, VMT1_I AUSS, VMT1_I AUDO, VMT1_I APUE, #
& VMT1_I ALE2, VMT1_I AZI1)#
& PIR1: (JN0392, MFI V61, MFI V, MFI V1M, MFI VC, MFI V1K, MFI VWS)#
& PIR2: (S04052, S04056, S14616, S04054, S04050, S07429, S04058)#
REFERENCES: bodmer89a#
COMMENT: #
SUMMARY: HLA-A2, actyesl, bi ndyesu, SPLTKGI LGFVFTLTV*#
SEQUENCE: SPLTKGI LGFVFTLTV*
```

- Data to extract
 - Source (human, mouse, ...)
 - Peptide sequences, source, activity data

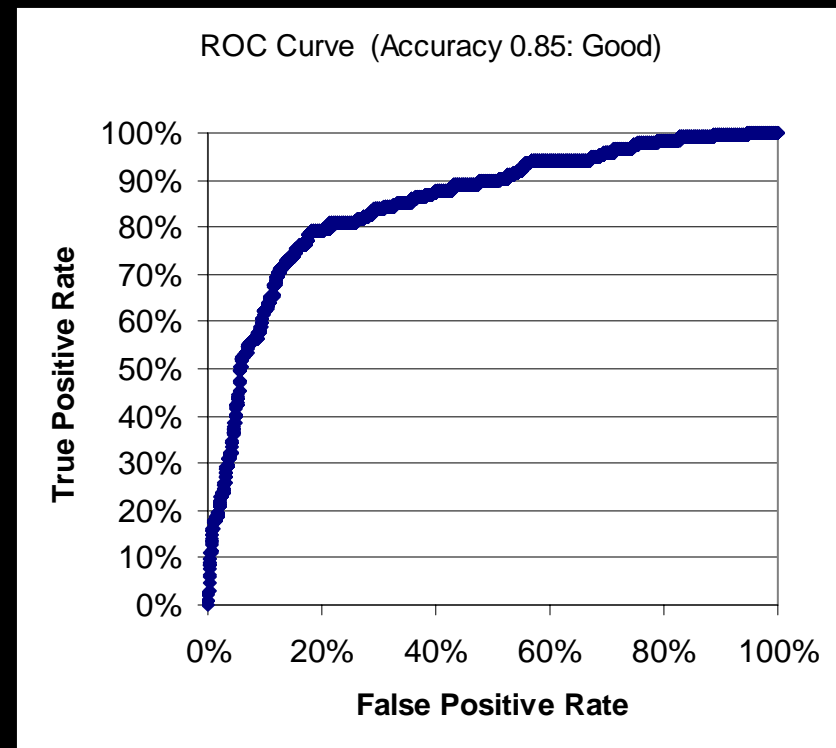
Work Flow



Data pipelining – Pipeline Pilot

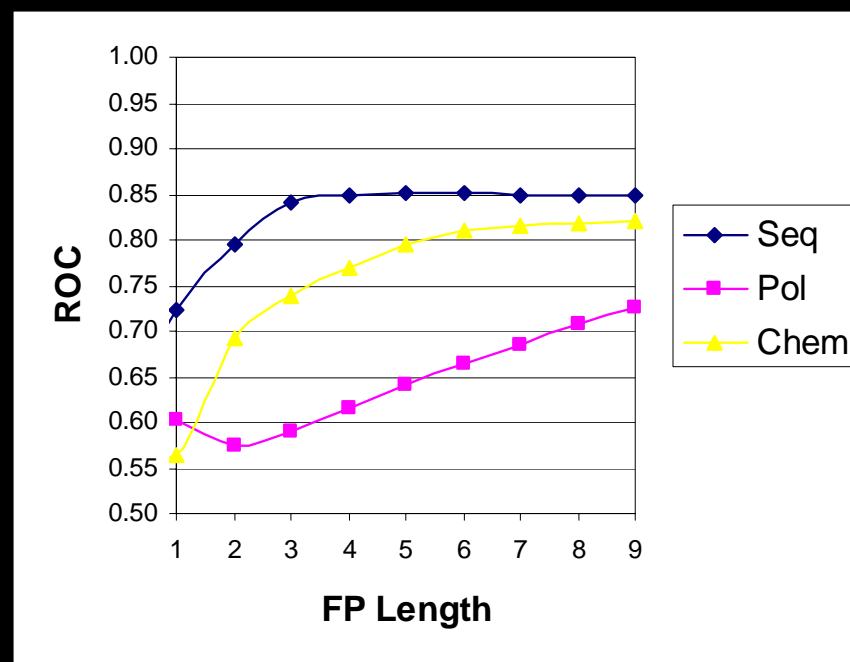
Validating the MHCPEP Model

- Receiver operating characteristics (ROC) curve
 - Method for validating classification models
 - Shows sensitivity versus specificity
- Area under the curve is a measure of test accuracy (1=perfect, 0.5=random)
- MHC activity model
 - Sequence fingerprint
 - Length = 5
 - ROC = 0.85



ROC as a Function of Fingerprint Length

- ROC analyses were performed for different fingerprint types and fingerprint lengths
- Sequence fingerprint is most powerful descriptor
- Sequence fingerprint has converged at length 5
- Polarization and chemical fingerprints converge more slowly and result in poorer model



The MHCPEP Bayesian Model

- Best features
 - AA sequence

Bin ID	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16
Bin Value	PGNYP	PGNY	APGN	APGNY	NYP	PGN	FAP	GNYP	APG	FAPG	GNY	YPAL	NYPA	NYPAL	FAPGN	YPA
Feature Count	22	22	22	22	34	25	29	26	45	21	31	34	22	22	12	38
Subset Count	22	22	22	22	28	22	24	22	30	18	22	23	17	17	12	23

- Polar/a-polar

Bin ID	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Bin Value	AAAAP	AAAPA	AAPAP	AAAAA	AAAP	PPAPP	AAAA	AAA	PAPPA	APAPP
Feature Count	1038	1212	1204	738	2107	1043	1313	2450	1035	1073
Subset Count	73	82	79	48	132	65	81	148	59	60

- Chemical class

Bin ID	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Bin Value	AEAES	BOASB	EAESA	CAOEC	AAEAE	ACAOE	EAES	AEABO	OECAA	ABOAS
Feature Count	47	57	46	19	61	23	61	72	27	76
Subset Count	24	27	23	13	25	13	24	26	13	27

The MHCPEP Bayesian Model

- ‘Worst’ features
 - AA sequence

Bin ID	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20
Bin Value	WV	SG	VR	IS	TD	FI	RS	KF	YN	YD	AEI	LN	NK	YE	QAV	QQ	WV	GG	DQ	RW
Feature Count	164	161	154	146	144	115	107	99	92	88	83	174	168	73	71	70	67	151	66	65
Subset Count	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0

- Polar/a-polar

Bin ID	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
Bin Value	PPPPP	PAPAP	PPPAA	AAPAA	APAAP	APAA	PAAPP	PPAAP	PAPAA	APAAA
Feature Count	397	699	1012	1116	1051	2041	1003	1076	992	982
Subset Count	7	13	26	29	29	65	33	37	35	35

- Chemical class

Bin ID	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
Bin Value	SSO	ASOAA	AOAS	AAAAO	OEAAA	BBB	ABAB	EAAAO	OSSA	SBB
Feature Count	195	118	106	101	99	96	93	92	92	90
Subset Count	0	0	0	0	0	0	0	0	0	0
Normalized Probability	-2.31	-1.87	-1.78	-1.74	-1.72	-1.70	-1.67	-1.66	-1.66	-1.64

Proposing Peptides with High Activity

- Can we generate novel peptides with high likelihood of being active?
- A combinatorial peptide library built from 20 amino acids:
 - Peptide of length 10: $20^{10} = 1.0E13$
 - The haystack is too big!
- Need a more focused search
 - Select all peptide fragments of size 2 among best 200 features
 - Combine with 20 amino acids
 - Randomly pick a 'bit' to build peptides of size 8, 9, 10

Building blocks

A F G I L M P V W C D E H K N Q R S T Y
 KQ VF GF CD AP PA TK FA LK NY FV YP YL FT

Peptides generated

VF;AP;G;NY;PA;L VFAPGNYPAL
 I;L;G;FV;FT;LK;L ILGFVFTLKL

Scoring the Virtual Library

- Generated 100M ‘virtual’ peptides (5h30 min)
- Scored and sorted virtual and known peptides using Bayesian model
- High scoring virtual compounds suggest novel peptide chemistries to explore or motives to search for in protein data bank

Source	MHC	Sequence	Bits	MHCPEP	Activity
human	HLA-B51	ILGFVFTLTV		36.19	Unknown
human	HLA-A2	ILGFVFTLTV		36.19	Unknown
human	HLA-A69	ILGFVFTLTV		36.19	Moderate
		VFAPGNYPAL	VF;AP;G;NY;PA;L	35.76	
human	HLA-A2	GILGFVFTLT		35.26	Moderate
		ILGFVFTLKL	I;L;G;FV;FT;LK;L	34.36	
mouse	H-2Db	FAPGNYPAL		34.02	Unknown
mouse	H-2Kb	FAPGNYPAL		34.02	High
mouse	H-2Kd	FAPGNYPAL		34.02	Unknown
		LKQAPGNYPAL	L;KQ;AP;G;NY;PA	33.88	
		YLKQAPGNYP	YL;KQ;AP;G;N;YP	33.52	
		PALGFVFTLK	PA;L;GF;VF;T;LK	32.95	
		YLGfVFTLKL	YL;GF;V;FT;LK;L	32.71	
human	HLA-A2	ILGFVFTLT		32.57	Moderate
		FAPGFVFTLK	FA;P;GF;V;FT;LK	32.52	
		AYLKQAPGNYP	A;YL;KQ;AP;G;NY	32.49	
human	HLA-B51	LGFVFTLTV		32.34	Unknown
		FAPGNYPATL	F;AP;G;NY;PA;T;L	32.32	
		VKQAPGNYPAL	V;KQ;AP;G;NY;PA	32.18	
		PGNYPAYLKQ	P;G;NY;PA;YL;KQ	32.14	
		YLGfVFTLKQ	YL;GF;V;FT;L;KQ	32.09	
		MGFAPGNYPAL	M;GF;AP;G;NY;PA	32.08	
		TVFAPGNYPAL	T;VF;AP;G;NY;PA	31.85	
		VFAYLKQATK	V;FA;YL;KQ;A;TK	31.69	
		TKQAPGNYPAL	TK;Q;AP;G;NY;PA	31.67	
mouse	I-Es	IAYLKQATK		31.58	Unknown
mouse	I-Ek	IAYLKQATK		31.58	Unknown
mouse	I-Eb	IAYLKQATK		31.58	Unknown
		QVFAPGNYPAL	Q;VF;AP;G;NY;P;A	31.55	
		VFAPGNYPAL	VF;AP;G;NY;PA	31.54	
		LGFVFTLTK	L;GF;V;FT;L;TK	31.46	

Scoring the Virtual Library

- Generated 100M ‘virtual’ peptides (5h30 min)
- Scored and sorted virtual and known peptides using Bayesian model
- High scoring virtual compounds suggest novel peptide chemistries to explore or motives to search for in protein data bank

Source	MHC	Sequence	Bits	MHCPEP	Activity
		PAPGNYPALK	PA;P;G;NY;PA;LK	31.32	
		FAPGNYPAFT	F;AP;G;N;Y;PA;FT	31.31	
human	HLA-A2	GILGFVFTL		31.26	Moderate
mouse	H-2Db	GILGFVFTL		31.26	Unknown
human	HLA-A69	GILGFVFTL		31.26	Unknown
		GFVFTLKQAP	GF;VF;T;LK;Q;AP	31.25	
		VFAPGNYPAG	VF;AP;G;NY;PA;G	31.1	
human	HLA-A2	AILGFVFTL		31.05	High
		VFAPGNYPAN	VF;AP;G;N;YP;A;N	31.02	
		APGNYPALTK	AP;G;N;Y;PA;L;TK	31.01	
		LFAYLKQATK	L;FA;YL;KQ;A;TK	30.94	
		GNYPAFVFTL	G;NY;PA;FV;FT;L	30.6	
		FAYLKQATKQ	FA;YL;KQ;A;T;KQ	30.6	
		FAPYLG FVFT	FA;P;YL;GF;V;FT	30.59	
		GFAPGFVFTL	GF;AP;GF;V;FT;L	30.56	
		LGFVFTLKL	L;G;F;VF;T;LK;L	30.51	
		PALGFVFTL	PA;L;GF;V;FT;L	30.5	
		IAYLGFVFTK	I;A;YL;GF;VF;TK	30.49	
		QYLG FVFTLK	Q;YL;GF;V;FT;LK	30.48	
mouse	H-2Kb	QAPGNYPAL		30.47	High
		GNYLGFVFTL	G;NY;L;GF;VF;T;L	30.47	
		APGNYPFVFT	AP;G;NY;P;FV;FT	30.46	
		GAPGNYPALK	G;AP;G;NY;PA;LK	30.31	
		YNYPALGFVF	Y;NY;PA;L;GF;VF	30.3	
		APGFVFTLKQ	AP;GF;VF;T;L;K;Q	30.28	
		APGNYPALFA	A;P;G;NY;PA;L;FA	30.26	
		APGNYPALK	AP;G;NY;PA;LK	30.15	
human	HLA-A2	KILGFVFTL		30.14	High
		FAYLKQATKI	FA;YL;KQ;A;TK;I	30.11	
		YLG FVFTLTG	YL;GF;V;FT;L;T;G	30.1	
		GFVFAPGNY	GF;VF;AP;G;NY	30.07	
mouse	H-2Kb	RAPGNYPAL		30.06	High

Conclusions

- This initial exploration showed
 - Fast approximate modeling works for peptide data
 - Fingerprint based descriptors of sequence fragments gives good results over broad range of data
 - Chemical based abstractions gave poorer results for MHC binding (can be more useful for other types or binding)
- Pipeline Pilot is a good prototyping environment

Acknowledgements

- Carol Gorst (SciTegic)
- Andrei Caracoti (SciTegic)
- Visit us at the exhibition at the MDL booth