
Organization and analysis of information for biotherapeutics research

Mark R. Hansen, Hugo O. Villar

Altoris, Inc. , La Jolla, CA

Eric Feyfant

Global Biotherapeutic Technologies Dept., Pfizer Inc., Cambridge, MA

Analytics for Biologics

❖ Current state of the art

- 3D Structure to activity
- Possible when small datasets are available

❖ Lack of tools to correlate sequence to activity

- Macros in MS Excel

❖ Narrow applicability

- Peptides, Antibodies, Protein Engineering, Polynucleotides, need to be served

SARvision|Biologics

❖ Solution for research informatics

- SAR analysis, not production or inventory
- Primary Sequence to Activity
 - For 3D and bioinformatics solutions exist

❖ Three broad development considerations

- Data handling
- Data organization
- Primary structure analytics

❖ Integration with other applications possible

Spreadsheet format

❖ Basic Visualization

- Sorting
 - Sequence or alphanumerical data
- Heat maps
- Properties: hydrophobicity secondary structure

❖ Basic Calculations

- Molecular weight isoelectric point, etc.
- Boolean or arithmetic operations on columns

SARvision | Biologics

SARvision | Biologics

FileViewDataTools

Full Sequence

Motifs: 1

ID: GLP1-7137

Full Sequence

Motifs: 1

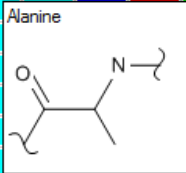
ID: GLP1-7137

Data	(+)		ID	IC50	EC50	MW	pl	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
								H	A	E	G	T	F	T	S	D	V	S	S	Y	L	E	G	Q
1		16	0.15	0.22	3167.43	4.54		*	*	D	*	*	*	*	*	*	*	*	*	*	*	*	*	*
2		Exendin-4	0.22	0.30	4169.55	4.39		*	G	*	*	*	*	*	*	*	L	*	K	Q	M	*	E	E
3		15	0.09	0.40	3323.62	5.36		*	*	D	*	*	*	*	*	*	*	*	*	*	*	*	*	*
4		18	0.24	0.50	3165.50	5.31		*	*	L	*	*	*	*	*	*	*	*	*	*	*	*	*	*
5		1	0.18	0.50	3181.45	4.54		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
6		19	0.09	0.70	3339.73	7.12		*	*	M	*	*	*	*	*	*	*	*	*	*	*	*	*	*
7		20	0.19	0.80	3183.54	5.31		*	*	M	*	*	*	*	*	*	*	*	*	*	*	*	*	*
8		17	0.40	0.80	3321.69	7.12		*	*	L	*	*	*	*	*	*	*	*	*	*	*	*	*	*
9		7	0.15	0.80	3337.64	5.41		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
10		2	0.32	0.90	3347.67	4.60		F	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
11		GLP1-7137	0.31	0.90	3337.64	5.41		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
12		8	0.33	1.40	3181.45	4.54		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
13		11	0.47	2.50	3365.69	5.41		*	V	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
14		21	1.50	3.50	3295.61	7.12		*	*	S	*	*	*	*	*	*	*	*	*	*	*	*	*	*
15		12	1.40	3.50	3209.50	4.54		*	V	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
16		6	2.70	5.40	3363.67	4.60		Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
17		22	4.90	6.40	3139.42	5.31		*	*	S	*	*	*	*	*	*	*	*	*	*	*	*	*	*
18		3	1.60	7.00	3191.48	4.14		F	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
19		9	2.90	15.00	3353.64	5.41		*	S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
20		10	3.80	17.00	3197.45	4.54		*	S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
21		28	19.50	98.00	3181.45	4.54		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
22		24	19.50	98.00	3151.47	5.31		*	*	V	*	*	*	*	*	*	*	*	*	*	*	*	*	*
23		27	9.30	103.00	3337.64	5.41		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
24		23	9.30	103.00	3307.66	7.12		*	*	V	*	*	*	*	*	*	*	*	*	*	*	*	*	*
25		25	216.00	126.00	3336.70	9.05		*	*	K	*	*	*	*	*	*	*	*	*	*	*	*	*	*
26		4	3.30	127.00	3386.71	4.60		W	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
27		5	4.60	152.00	3230.52	4.14		W	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
28		13	5.70	180.00	3379.72	5.41		*	L	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
29		14	16.80	218.00	3223.53	4.54		*	L	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
30		26	231.00	354.00	3180.51	7.12		*	*	K	*	*	*	*	*	*	*	*	*	*	*	*	*	*

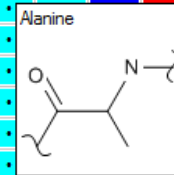
Color by Property

Data																													
	EC50	pI	Full Sequence																										
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	
			H	A	E	G	T	F	T	S	D	V	S	S	Y	L	E	G	Q	A	A	K	E	F	I	A	W	L	
1	0.90	5.41	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
2	0.50	4.54	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		
3	0.90	4.60	F	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		
4	7.00	4.14	F	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		
5	127.00	4.60	W	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		
6	152.00	4.14	W	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		
7	5.40	4.60	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		
8	0.80	5.41	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		
9	1.40	4.54	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		

Alanine



Data																										
	IC50	EC50	pI	Full Sequence																						
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
				H	A	E	G	T	F	T	S	D	V	S	S	Y	L	E	G	Q	A	A	K	E	F	I
1	0.31	0.90	5.41	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
2	0.18	0.50	4.54	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
3	0.32	0.90	4.60	F	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
4	1.60	7.00	4.14	F	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
5	3.30	127.00	4.60	W	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
6	4.60	152.00	4.14	W	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
7	2.70	5.40	4.60	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
8	0.15	0.80	5.41	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
9	0.33	1.40	4.54	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
10	2.90	15.00	5.41	*	S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
11	3.80	17.00	4.54	*	S	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
12	0.47	2.50	5.41	*	V	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
13	1.40	3.50	4.54	*	V	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
14	5.70	180.00	5.41	*	L	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
15	16.80	218.00	4.54	*	L	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
16	0.09	0.40	5.36	*	*	D	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
17	0.15	0.22	4.54	*	*	D	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*



Data Handling Challenges

- ❖ Chemoinformatics tools inadequate
 - Mostly based on 2D
- ❖ Information can be dispersed
- ❖ Biopolymers are built in blocks
 - Extensive repertoire
 - e.g. unnatural amino acids, isomers, etc.
 - Chemically modified
 - cyclization, glycosylation, etc.

Data Handling

❖ Long lists of monomeric units

- Stereoisomers, chemical modifications, etc.

❖ Combinatorial problem

- Monomeric units times modifications

❖ Process information from two files:

- Monomer Units: nucleotides, amino acids, etc.
- Chemical Modifications: operations on the monomers

Monomer Datafile

❖ Structural information

- SMILES strings

❖ Color schemes

- Clustal, Hydrophobicity, Secondary structure, etc.

❖ Synonyms

- G == Gly == GLY == Glycine, etc.

❖ Monomer properties

- MW, pKa, etc.

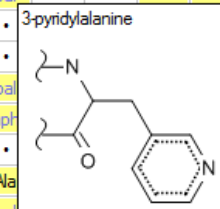
❖ Large collections of unnatural monomers

Modifier Datafile

- ❖ Name and synonyms for modifiers
 - stereochemistry, glycosylation, N-methylation, etc.
- ❖ Color Scheme
 - highlight modified residues
- ❖ Property Modification
 - i.e. deviation introduced by the modification (+MW)

Sequence characteristics

IC50					ShortSeq																									
SST1	SST2	SST3	SST4	SST5	15	16	17	18	19	20	21	22	23	24	25	26	27	28	28A											
					Ala	Gly	Cys(1)	Lys	Asn	Phe	Phe	Trp	Lys	Thr	Phe	Thr	Ser	Cys(1)	-											
10000.0	1.8	43.0	66.00	0.62	-	amp	•(1)	-	-	-	Tyr	•	•	-	-	-	Val	•(1)	Thr											
197.0	1.9	52.0	1.00	43.00	-	-	•(1)	Phe	-	-	•	•	•	•	•	•	-	•(1)	-											
10000.0	2.1	4.4	10000.00	5.60	-	Phe	•(1)	-	-	-	•	•	•	-	-	-	Thr	•(1)	Thr											
3.9	2.2	7.1	3.80	3.90	•	•	•(1)	•	•	•	•	•	•	•	•	•	•	•(1)	-											
3.2	2.3	3.5	2.50	2.40	•	•	•(1)	•	•	•	•	•	•	•	•	•	•	•(1)	-											
112.0	5.0	11.0	1.40	9.50	-	-	•(1)	Asn	-	•	•	•	•	•	•	•	-	•(1)	-											
10000.0	5.4	3.1	10000.00	0.70	-	Phe	•(1)	-	-	-	Tyr	•	•	-	-	-	Val	•(1)	Trp											
5.3	15.0	39.0	0.60	13.00	-	-	•(1)	•	-	•	•	•	•	•	•	-	-	•(1)	-											
10.3	15.5	8.2	0.51	4.80	-	-	•(1)	-	-	•	•	•	•	•	Tyr	-	-	•(1)	-											
10.0	16.0	8.0	0.50	4.80	-	-	•(1)	•	-	•	•	•	•	•	Tyr	•	•	•(1)	-											
117.0	26.0	36.0	1.80	20.30	-	-	•(1)	-	-	•	Tyr	•	•	•	•	-	-	•(1)	-											
79.0	28.0	222.0	3.30	8.60	-	-	•(1)	-	-	•	•	•	•	•	•	•	•	•(1)	-											
27.0	41.0	13.0	1.80	46.00	-	-	•(1)	-	-	•	•	•	•	•	•	-	-	•(1)	-											
6.5	43.0	10.0	1.30	24.00	-	-	•(1)	-	-	•	•	•	•	•	Tyr	-	-	•(1)	-											
27.0	54.0	22.0	1.30	63.00	-	-	•(1)	-	-	•	Tyr	•	•	•	•	-	-	•(1)	-											
517.0	56.0	263.0	1.20	34.00	-	Tyr	•(1)	-	-	•	aph	•	•	•	•	-	-	•(1)	-											
330.0	57.0	347.0	1.10	51.00	-	Tyr	•(1)	-	-	•	Ala	•	•	•	•	-	-	•(1)	-											
450.0	71.0	271.0	0.88	30.00	-	-	•(1)	-	-	•	aph	•	•	•	•	-	-	•(1)	-											
348.0	81.0	171.0	10000.00	524.00	-	nal	•(1)	-	-	-	pal	•	•	-	-	-	Val	•(1)	nal											
59.0	95.0	189.0	1.20	31.00	-	-	•(1)	Asn	-	•	•	•	•	•	•	-	-	•(1)	-											
5.3	130.0	13.0	0.70	14.00	-	-	•(1)	-	-	•	•	•	•	•	•	-	-	•(1)	-											
413.0	163.0	192.0	1570.00	382.00	-	nal	•(1)	-	-	-	pal	•	•	•	•	-	-	•(1)	nal											
327.0	170.0	247.0	1.10	240.00	-	-	•(1)	-	-	•	aph	•	•	•	•	-	-	•(1)	-											
13.0	179.0	57.0	1.60	19.00	-	Tyr	•(1)	-	-	•	•	•	•	•	•	-	-	•(1)	-											
10000.0	183.0	897.0	0.98	199.00	-	-	•(1)	-	-	•	Ala	•	•	•	•	-	-	•(1)	-											
1200.0	203.0	379.0	10000.00	1860.00	-	nal	•(1)	-	-	-	pal	•	•	•	•	-	Ala	•(1)	nal											
309.0	213.0	273.0	267.00	190.00	-	-	•(1)	•	-	•	•	•	•	•	•	-	•	•(1)	-											
270.0	260.0	135.0	1.90	663.00	-	Tyr	•(1)	-	-	•	aph	•	•	•	•	-	-	•(1)	-											
213.0	347.0	10000.0	1.20	10000.00	-	-	•(1)	-	-	•	aph	•	•	•	•	-	-	•(1)	-											
10000.0	531.0	10000.0	229.00	10000.00	-	-	•(1)	-	-	-	•	•	•	•	•	-	-	•(1)	-											
415.0	543.0	243.0	728.00	968.00	-	nal	•(1)	-	-	-	pal	•	•	•	•	-	-	Leu	•(1)	nal										
10000.0	598.0	10000.0	10000.00	10000.00	-	-	•(1)	-	-	•	•	•	•	•	•	-	-	•(1)	-											
1000.0	622.0	624.0	2.00	692.00	-	Tyr	•(1)	-	-	•	Ala	•	•	•	•	-	-	•(1)	-											



Input standardization examples

❖ Cyclic

- YCFFWKTF(C-C Cyclized)
- YC1KFWZTFTC1
- YC1KE2FWZTFK2SC1

❖ Stereochemistry

- C1NFFd-WKTFTC1

❖ Chemical Modifications

- [Nme]C1KFFWKTFSC1

❖ Unnatural amino acids

- [nal]C1[pal]WKVC1[nal]

Data Handling

❖ Some level of standardization required

- Lack of systematic nomenclature
 - Specially for peptides
- e.g. modifications shown in brackets

❖ Retain as much flexibility as possible

- Use of synonyms for amino acids and modifications

User can fully edit the information and enter additional data as needed

Data Organization

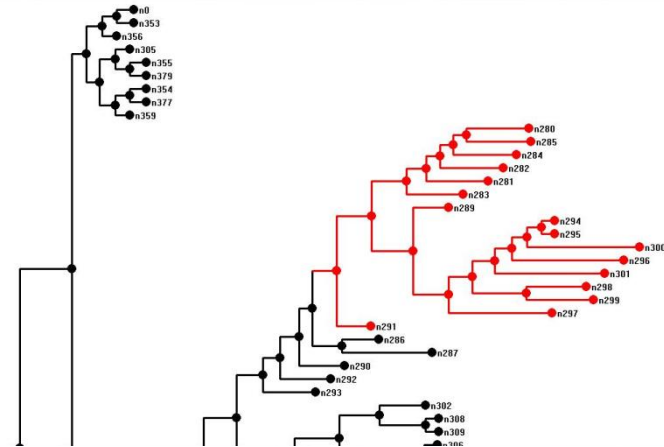
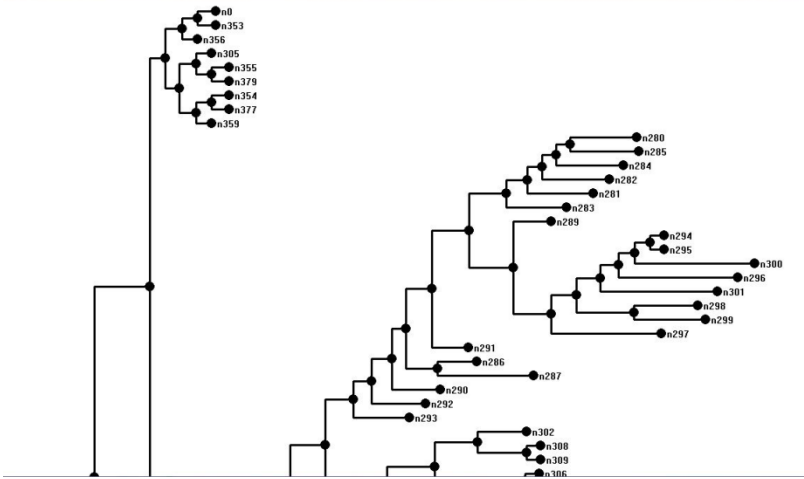
❖ Align sequences as read-in

- Clustal and other algorithms
- Read in aligned sequences
- User able to edit alignments

❖ Potential use of different substitution matrices

- Currently BLOSUM62
- Unnatural monomers present an issue

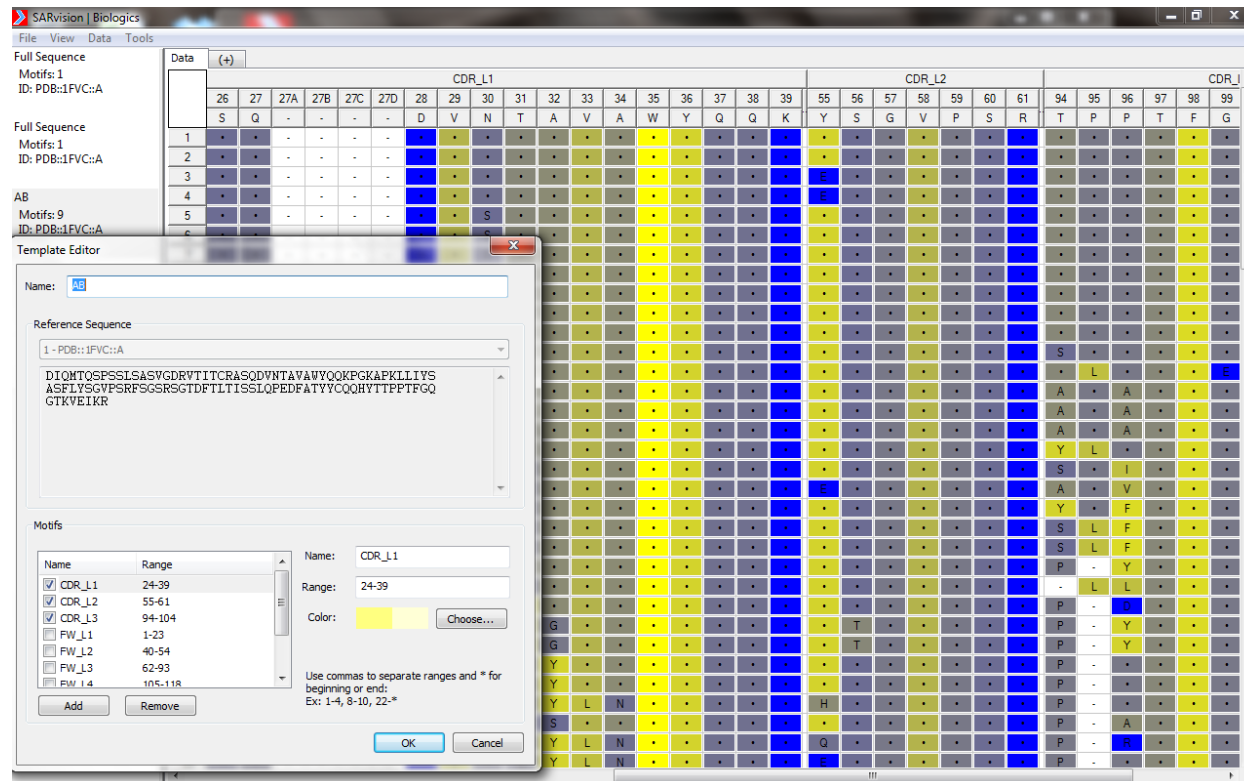
From Alignments to SAR



	Identity	Homology								CDR_L2							CDR_L3											
			33	34	35	36	37	38	39	55	56	57	58	59	60	61	94	95	96	97	98	99	100	101	102	103	104	
			V	A	W	Y	Q	Q	K	Y	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	
1	1.00	562.00	V	A	W	Y	Q	Q	K	Y	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	
2	1.00	562.00	V	A	W	Y	Q	Q	K	Y	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	
3	0.99	553.00	V	A	W	Y	Q	Q	K	E	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	
4	0.99	553.00	V	A	W	Y	Q	Q	K	E	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	
5	0.97	541.00	V	A	W	Y	Q	Q	K	Y	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	
6	0.97	541.00	V	A	W	Y	Q	Q	K	Y	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	
7	0.97	541.00	V	A	W	Y	Q	Q	K	Y	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	
8	0.97	541.00	V	A	W	Y	Q	Q	K	Y	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	
9	0.97	541.00	V	A	W	Y	Q	Q	K	Y	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	
10	0.97	541.00	V	A	W	Y	Q	Q	K	Y	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	
11	0.95	526.00	V	A	W	Y	Q	Q	K	Y	S	G	V	P	S	R	T	P	P	T	F	G	Q	G	T	K	V	

Motifs based analysis

❖ Antibody Engineering, CDRs



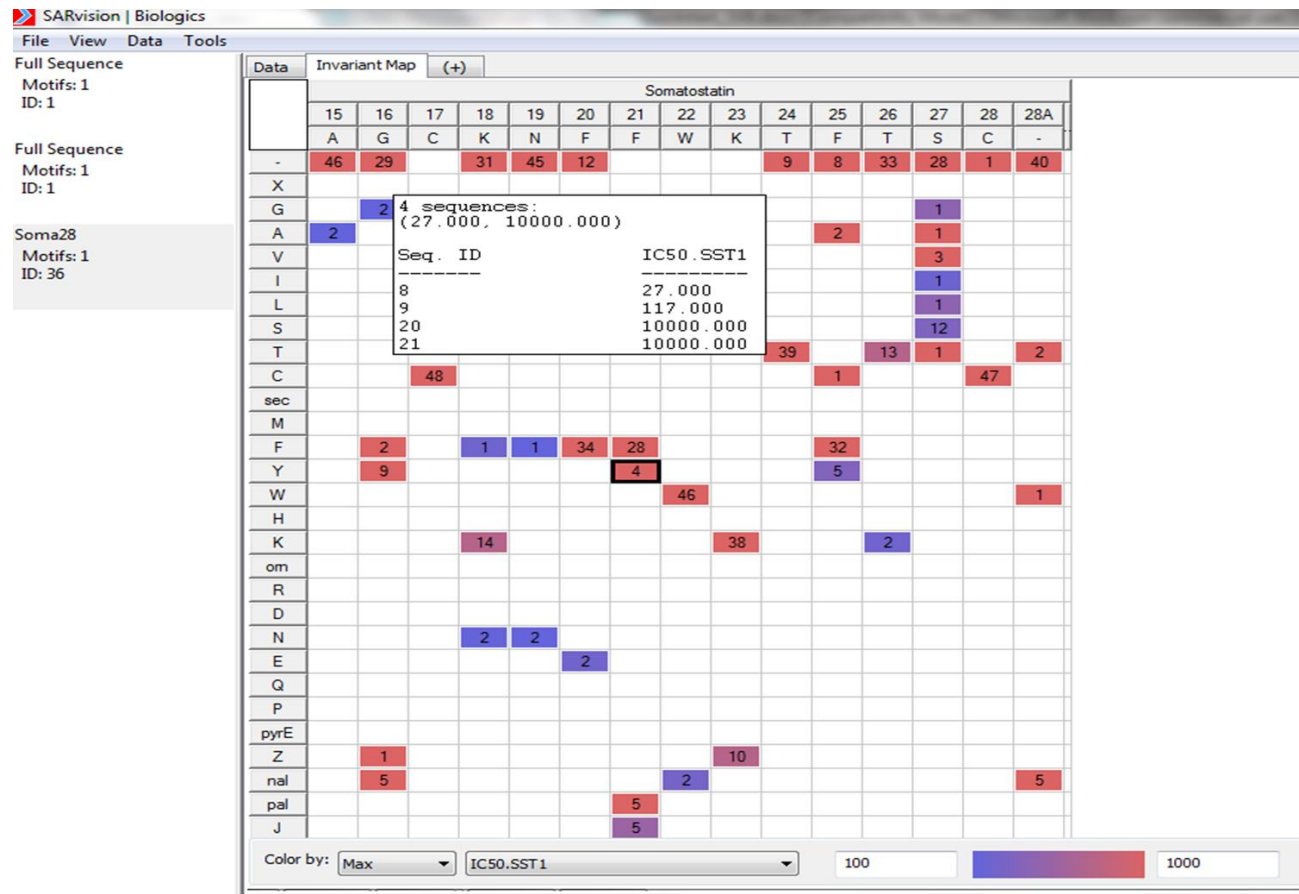
Antibodies from the Protein DB, CDRs shown colored by Hopps Wood scores

SAR for Biotherapeutics

- ❖ Tools that exploit unique nature are needed
- ❖ State of the art involves use of 3D structure
 - Need supplementary tools to analyze larger datasets
- ❖ Identify patterns in aligned sequences
- ❖ New tools are required
 - Invariant maps
 - Activity Cliffs

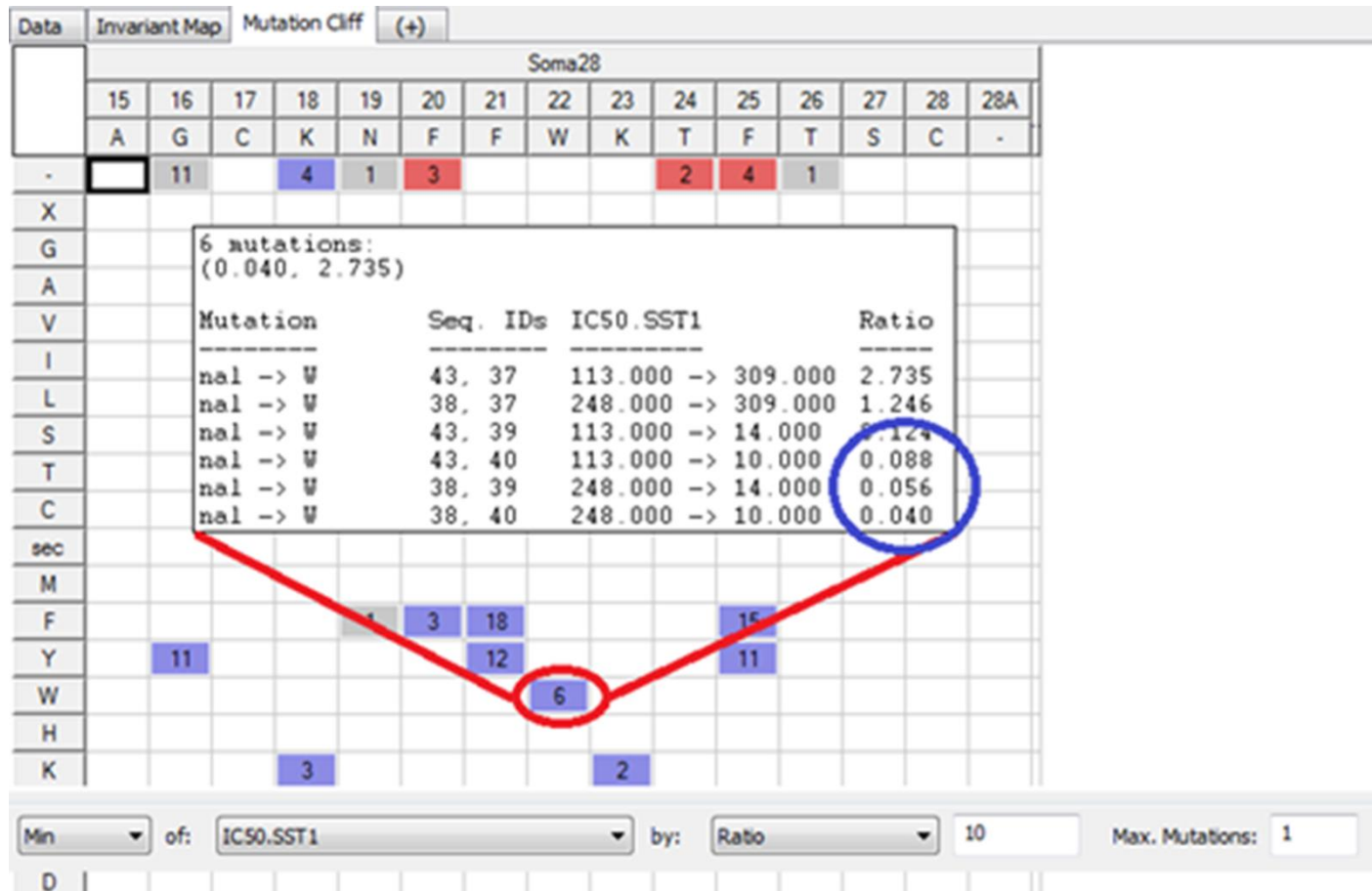
Invariant Maps

- ❖ Activity trends as a position in the sequence is kept invariant

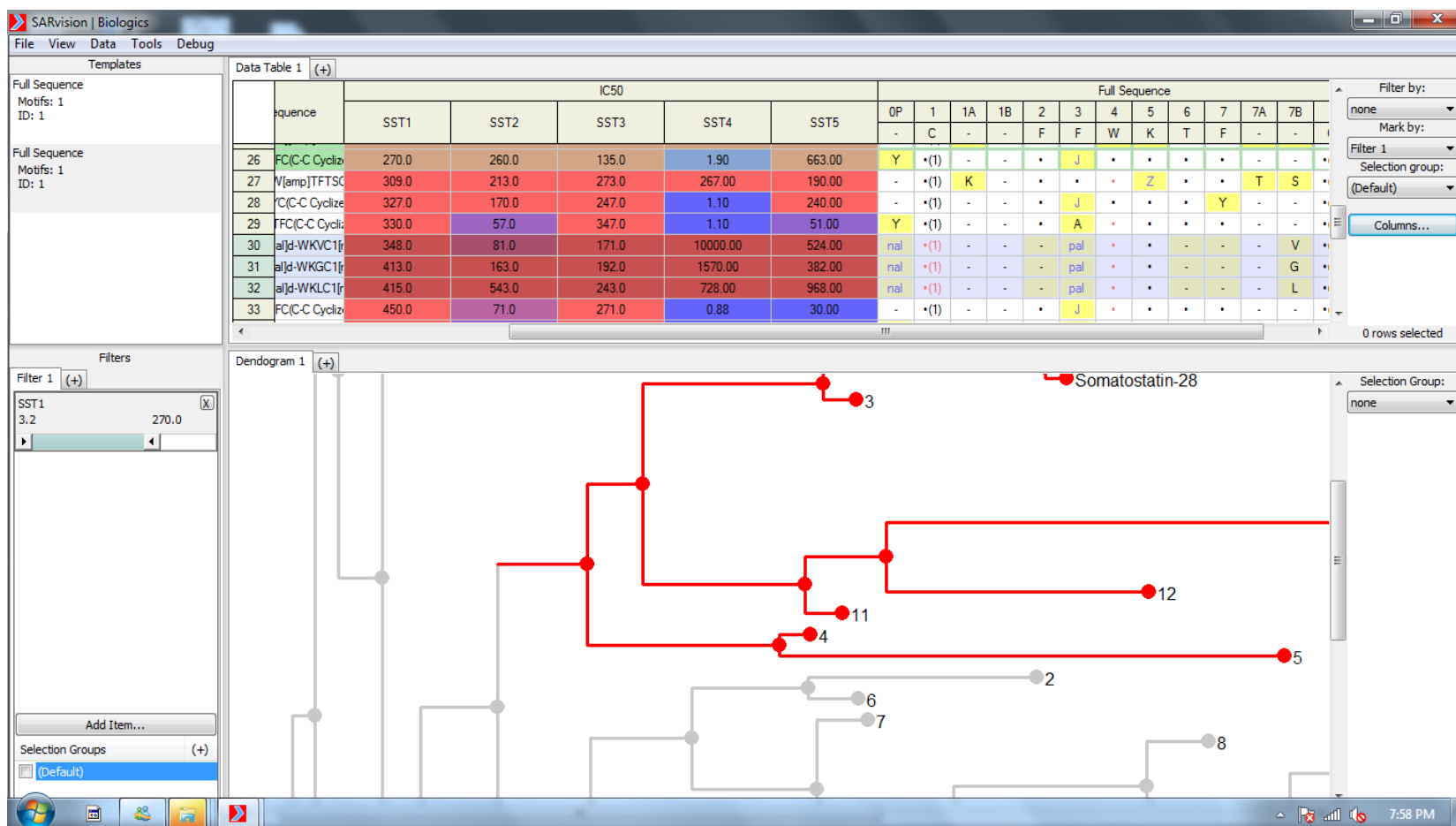


Mutation Cliffs

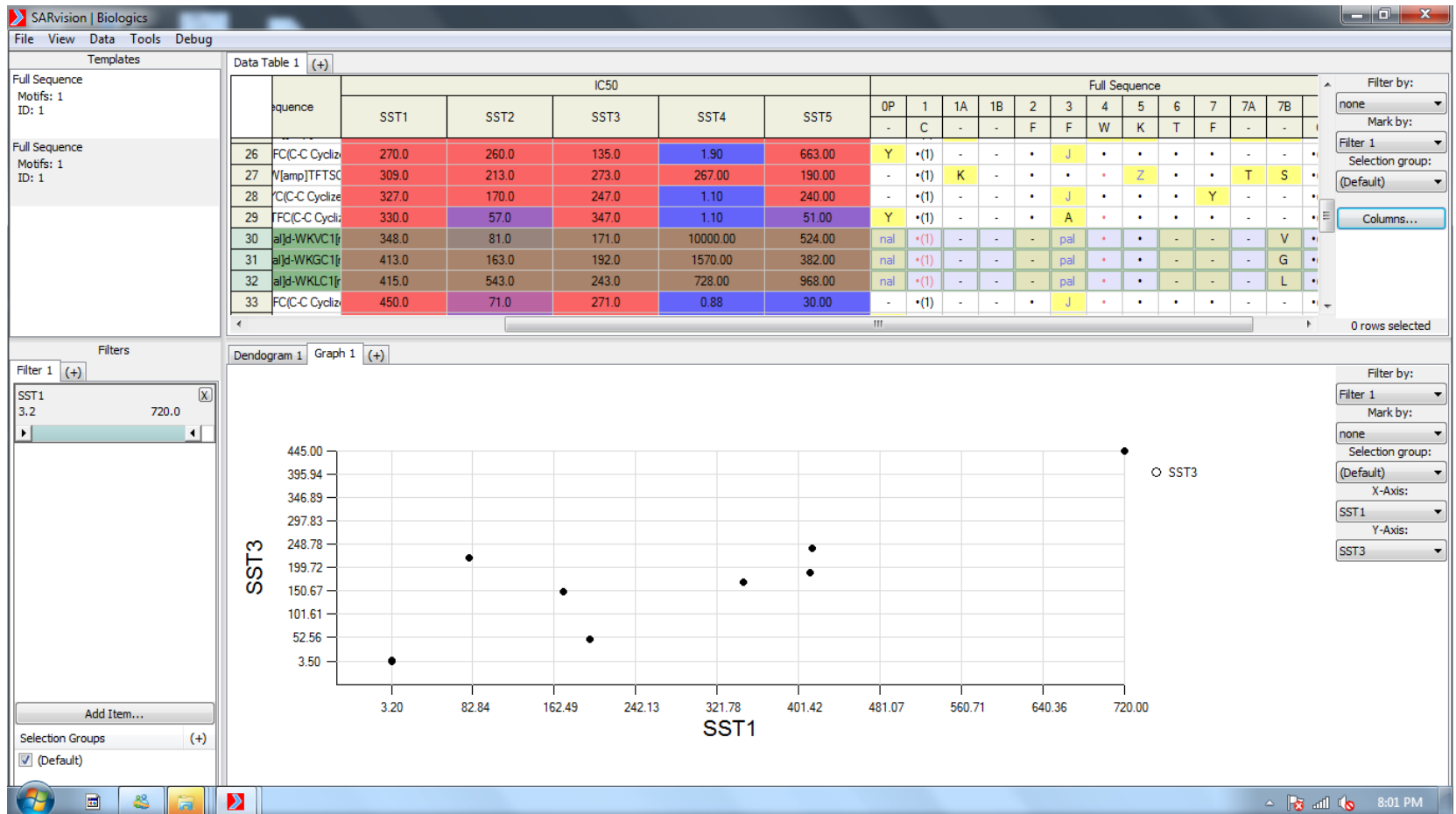
- ❖ Sequence mutations that elicit a significant change in activity



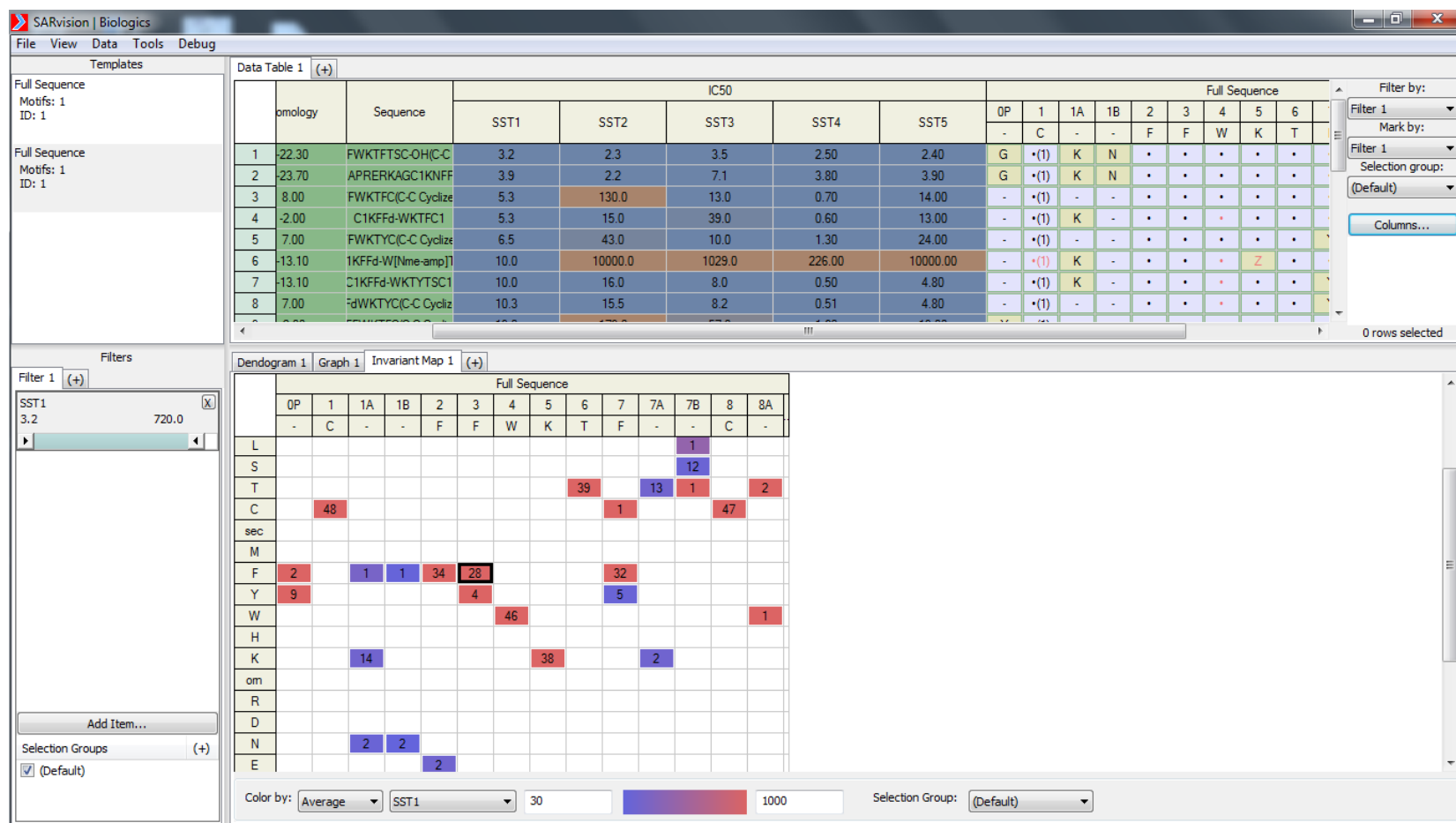
Current Version



Current Version



Current Version



Summary

- ❖ **SARvision | Biologics** was developed to fill in a gap in research informatics
- ❖ Representation, Organization and Analysis need to be reexamined to serve a broad range of compounds
- ❖ Tools to supplement 3D analysis that will not always be possible with large datasets
- ❖ Tools for analysis, Invariant Maps and Mutation Cliffs aid in defining SAR

ALTORIS, INC.

7660-H Fay Ave #347
La Jolla, CA 92037

www.altoris.com
www.chemapps.com