CHEMAPPS
Scientific Software and Discovery Consulting

Application Note

Facilitating the identification of Sequence Activity Relationships in antibodies with SARvision|Biologics

The increased interest in peptides, antibodies and other biopolymers as therapeutic agents has made obvious that a significant gap exists in research informatics. Most of the efforts in the past and even today aim to deploy molecular modeling techniques to interpret data for biologics, which can be suitable when biopolymer data sets are small. However, as larger datasets are generated in biotherapeutics research, that approach becomes unmanageable, due to a lack of tools to organize and analyze such data. In collaboration with scientists researching biotherapeutic agents, we developed SARvision|Biologics to aid in the analysis of structure activity relationships, for peptides epitopes and paratopes, as well as RNA or other biopolymers. The program includes some unique analysis tools for determining relations between biological responses and biopolymers at the sequence level.

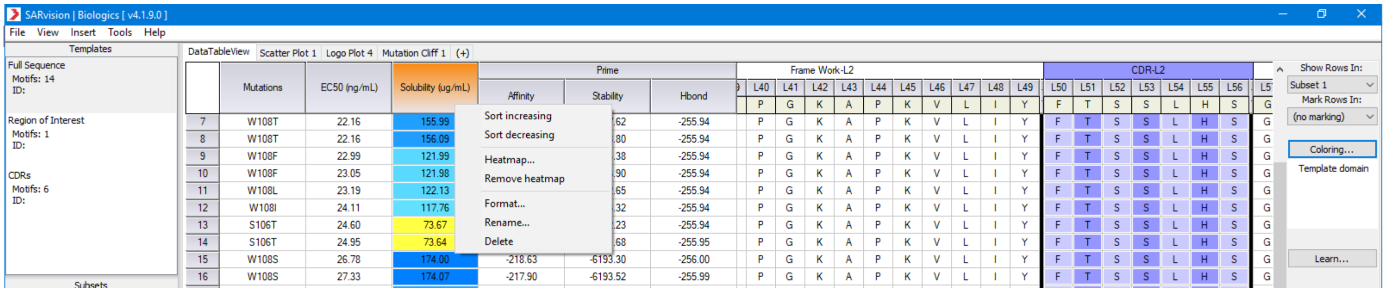
The Problem

Tools in research informatics have been mostly designed to deal with large datasets involving small molecules. In recent years the approval of several biotherapeutics (peptides, antibodies, polynucleotides) for clinical research or to market, has spurred significant research with biopolymers. Nowadays, it is not uncommon to deal with hundreds of biopolymers and their responses in batteries of tests aimed to evaluate their pharmacological properties. Cheminformatics has dealt effectively with the challenge of relating small molecule properties to biological function, but those tools are not equally suitable to deal with biotherapeutics. The optimization of the pharmacological properties of biopolymers is carried out by replacing (mutating), deleting or modifying residues or nucleotides. The use of a relatively more limited ensemble of building blocks which are put together in various arrangements, and sometimes with limited modifications, makes the analysis of their activity patterns different from the more significant changes that can be seen in small molecules. There is a lack of tools to perform the analysis of data in the context of macromolecules. Research projects in biologics involve much larger compound collections than in the past. What was possible to undertake with a simple spreadsheet, it is becoming unmanageable without new tools.
SARvision|Biologics is a *desktop application* aimed to fill in the gap that exists in the research informatics arena to deal with data on biologics. **SARvision|Biologics** provides a platform to read-in and organize data on biological polymers. The program has tools to aid in the identification of sequence activity relationships, for peptides, proteins, RNA and other biopolymers.

For more information contact us at info@altoris.com

Organizing Data on VEGF Antibodies

The use of monoclonal antibodies as therapeutics requires optimizing several key attributes, such as binding affinity and specificity, stability, solubility, pharmacokinetics among others. In some cases an engineered antibody should be compatible with the attachment of additional antibody domains (bispecific antibodies) and cytotoxic drugs (antibody-drug conjugates). Addressing these multiple conditions requires the management of significant amounts of data and the development of a solid understanding of the relations between sequence and the different parameters to be optimized. In some cases, experimental data can be supplemented with the results of computational methods expanding the list of parameters to be scrutinized. **SARvision|Biologics** offers a means to relate computational and experimental data to sequence, thus complementing molecular modeling and computational biology tools. A *smart spreadsheet able to understand sequences* with all the tools needed to identify their relationship to data. With **SARvision|Biologics** we can quickly read in and organize data with a tool that understands how to organize biopolymers.

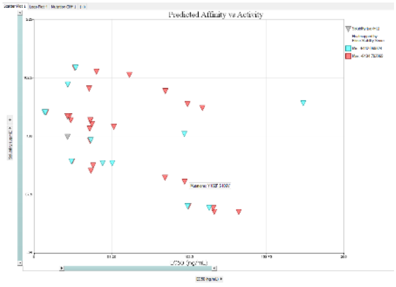


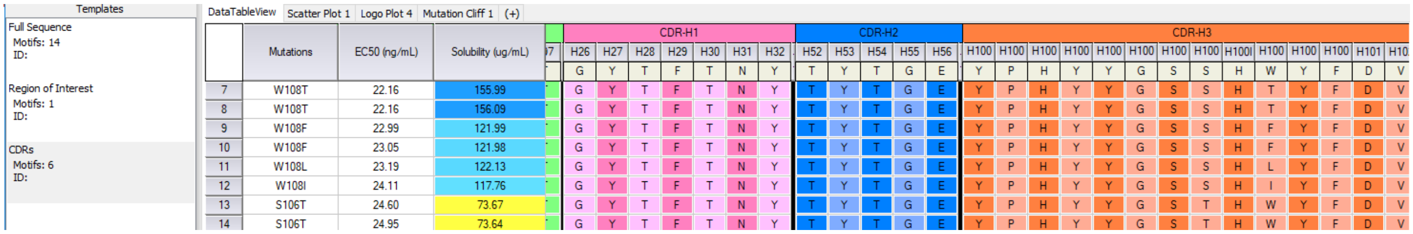
Show Rows In: Subset 1
Mark Rows In: (no marking)

				Prime			Frame Work-L2										CDR-L2							
	Mutations	EC50 (ng/mL)	Solubility (ug/mL)	Affinity	Stability	Hbond	L40	L41	L42	L43	L44	L45	L46	L47	L48	L49	L50	L51	L52	L53	L54	L55	L56	L5
							P	G	K	A	P	K	V	L	I	Y	F	T	S	S	L	H	S	G
7	W108T	22.16	155.99		.62	-255.94	P	G	K	A	P	K	V	L	I	Y	F	T	S	S	L	H	S	G
8	W108T	22.16	156.09		.80	-255.94	P	G	K	A	P	K	V	L	I	Y	F	T	S	S	L	H	S	G
9	W108F	22.99	121.99		.38	-255.94	P	G	K	A	P	K	V	L	I	Y	F	T	S	S	L	H	S	G
10	W108F	23.05	121.98		.90	-255.94	P	G	K	A	P	K	V	L	I	Y	F	T	S	S	L	H	S	G
11	W108L	23.19	122.13		.65	-255.94	P	G	K	A	P	K	V	L	I	Y	F	T	S	S	L	H	S	G
12	W108I	24.11	117.76		.32	-255.94	P	G	K	A	P	K	V	L	I	Y	F	T	S	S	L	H	S	G
13	S106T	24.60	73.67		.23	-255.94	P	G	K	A	P	K	V	L	I	Y	F	T	S	S	L	H	S	G
14	S106T	24.95	73.64		.68	-255.95	P	G	K	A	P	K	V	L	I	Y	F	T	S	S	L	H	S	G
15	W108S	26.78	174.00	-218.63	-6193.30	-256.00	P	G	K	A	P	K	V	L	I	Y	F	T	S	S	L	H	S	G
16	W108S	27.33	174.07	-217.90	-6193.52	-255.99	P	G	K	A	P	K	V	L	I	Y	F	T	S	S	L	H	S	G

In **SARvision|Biologics** sequence data can be loaded from comma delimited files (*.csv) or from FastA files (*.fasta), together with any additional information placed into a spreadsheet. Whenever users have preferred tools to carry out the alignment, those *alignments can easily be read-in*. Otherwise as the program *reads-in the sequences and aligns them using different tools* such as a Clustal or Needleman-Wunsch, with many different options for substitution matrices or the ability to read-in your own. Users can manually override any of the automated alignments and insert or delete gaps. The result is a spreadsheet that contains the aligned sequences and all other data uploaded, as shown above. In this case we are showing a mix of experimental and computed data that was read-in by the user.

Once the data is loaded into the program, the user can perform the tasks typical of any spreadsheet, such as *sort data*, including sorting residue columns, *edit numerical data*, or perform simple operations on the columns. Heat maps for numerical data can be a powerful means to identify trends in the data, as that shown for the solubility column in the figure above. New columns can be defined from the result of *operations* on existing ones. For example rather than looking at the affinity constants we can quickly generate a column that is the logarithm of the existing column, when looking at selectivity, we can define ratios. There is a large number of operations that are possible, including AND/OR type operations on the columns. Note that you can also sort the residues at any one position, or use residue counts as part of your equations. The program includes some *graphic capabilities* that are fully interactive, as you highlight some elements of the graph they are highlighted in the tables and viceversa.





Templates

Full Sequence
Motifs: 14
ID:

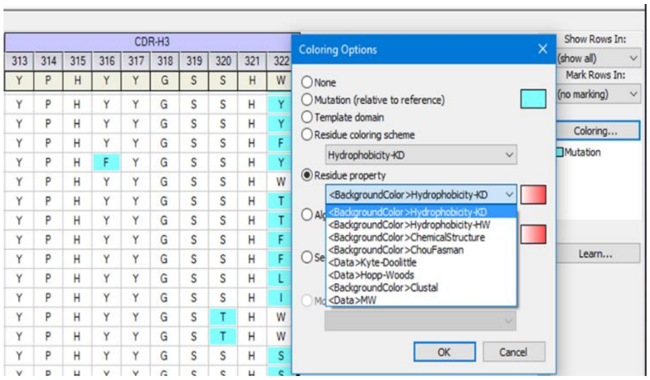
Region of Interest
Motifs: 1
ID:

CDRs
Motifs: 6
ID:

DataTableView | Scatter Plot 1 | Logo Plot 4 | Mutation Cliff 1 | (+)

					CDR-H1							CDR-H2					CDR-H3													
	Mutations	EC50 (ng/mL)	Solubility (ug/mL)	7	H26	H27	H28	H29	H30	H31	H32	H52	H53	H54	H55	H56	H100	H100	H100	H100	H100	H100	H100	H100	H100I	H100	H100	H100	H101	H10
					G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	W	Y	F	D	V
7	W108T	22.16	155.99		G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	T	Y	F	D	V
8	W108T	22.16	156.09		G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	T	Y	F	D	V
9	W108F	22.99	121.99		G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	F	Y	F	D	V
10	W108F	23.05	121.98		G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	F	Y	F	D	V
11	W108L	23.19	122.13		G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	L	Y	F	D	V
12	W108I	24.11	117.76		G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	I	Y	F	D	V
13	S106T	24.60	73.67		G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	T	H	W	Y	F	D	V
14	S106T	24.95	73.64		G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	T	H	W	Y	F	D	V

SARvision|Biologics allows thus the user to keep *different templates* to *go back and forth between different ways to organize the sequences*. For example, the user can select certain regions of the biopolymer to be displayed, including non-contiguous regions of the protein. This is particularly important when doing protein or antibody engineering, where large portions of the sequence may remain unchanged. As different templates are defined, the identity and sequence are recomputed automatically. Users can store and *share sequence* profiles that can be quickly deployed, for example zero in on only those regions of interest, or labeling residues according to well established numbering conventions such as Chotia or Kabsch-Sanders, or any other definition found useful by the team.

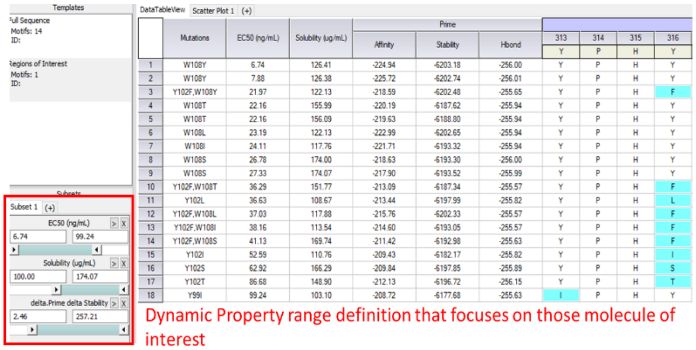


CDR-H3									
313	314	315	316	317	318	319	320	321	322
Y	P	H	Y	Y	G	S	S	H	W
Y	P	H	Y	Y	G	S	S	H	Y
Y	P	H	Y	Y	G	S	S	H	Y
Y	P	H	Y	Y	G	S	S	H	F
Y	P	H	F	Y	G	S	S	H	Y
Y	P	H	Y	Y	G	S	S	H	W
Y	P	H	Y	Y	G	S	S	H	T
Y	P	H	Y	Y	G	S	S	H	T
Y	P	H	Y	Y	G	S	S	H	F
Y	P	H	Y	Y	G	S	S	H	F
Y	P	H	Y	Y	G	S	S	H	L
Y	P	H	Y	Y	G	S	S	H	I
Y	P	H	Y	Y	G	S	T	H	W
Y	P	H	Y	Y	G	S	T	H	W
Y	P	H	Y	Y	G	S	S	H	S

Coloring Options

○ None
○ Mutation (relative to reference)
○ Template domain
○ Residue coloring scheme
Hydrophobicity-KD
● Residue property
<BackgroundColor>Hydrophobicity-KD

Different *coloring schemes* for the sequences and the data can be deployed. For sequences coloring can simply show residues that are not conserved, or colored according to properties, such as hydrophobicity or charge. Particularly useful is the use of filters that could be used to display or highlight only those sequences that fall within selected property ranges. The selection is done using sliders that provide a dynamic view of the data as the ranges of values are changed.



Templates

Full Sequence
Motifs: 14
ID:

Regions of Interest
Motifs: 1
ID:

Subsets

Subset 1 | (+)

EC50 (ng/mL): 6.74 – 99.24
Solubility (ug/mL): 100.00 – 174.07
delta.Prime delta Stability: 2.46 – 257.21

DataTableView | Scatter Plot 1 | (+)

				Prime						
	Mutations	EC50 (ng/mL)	Solubility (ug/mL)	Affinity	Stability	Hbond	313	314	315	316
							Y	P	H	Y
1	W108Y	6.74	126.41	-224.94	-6203.18	-256.00	Y	P	H	Y
2	W108Y	7.88	126.38	-225.72	-6202.74	-256.01	Y	P	H	Y
3	Y102F,W108Y	21.97	122.13	-218.59	-6202.48	-255.65	Y	P	H	F
4	W108T	22.16	155.99	-220.19	-6187.62	-255.94	Y	P	H	Y
5	W108T	22.16	156.09	-219.63	-6188.80	-255.94	Y	P	H	Y
6	W108L	23.19	122.13	-222.99	-6202.65	-255.94	Y	P	H	Y
7	W108I	24.11	117.76	-221.71	-6193.32	-255.94	Y	P	H	Y
8	W108S	26.78	174.00	-218.63	-6193.30	-256.00	Y	P	H	Y
9	W108S	27.33	174.07	-217.90	-6193.52	-255.99	Y	P	H	Y
10	Y102F,W108T	36.29	151.77	-213.09	-6187.34	-255.57	Y	P	H	F
11	Y102L	36.63	108.67	-213.44	-6197.99	-255.82	Y	P	H	L
12	Y102F,W108L	37.03	117.88	-215.76	-6202.33	-255.57	Y	P	H	F
13	Y102F,W108I	38.16	113.54	-214.60	-6193.05	-255.57	Y	P	H	F
14	Y102F,W108S	41.13	169.74	-211.42	-6192.98	-255.63	Y	P	H	F
15	Y102I	52.59	110.76	-209.43	-6182.17	-255.82	Y	P	H	I
16	Y102S	62.92	166.29	-209.84	-6197.85	-255.89	Y	P	H	S
17	Y102T	86.68	148.90	-212.13	-6196.72	-256.15	Y	P	H	T
18	Y99I	99.24	103.10	-208.72	-6177.68	-255.63	I	P	H	Y

Dynamic Property range definition that focuses on those molecule of interest

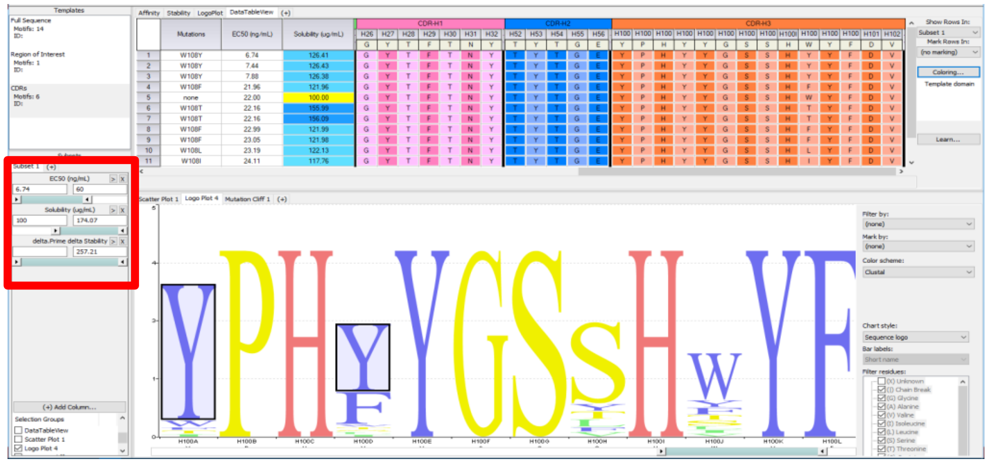
So far, we have loaded the data efficiently and quickly have been able to group compounds, sort data in different ways, create heat maps and even generate new data from the existing to look for selectivity. **SARvision|Biologics** will with these simple operations *save significant time* to the researchers and *eliminate many tedious, repetitious tasks* that were preventing them from looking at the data from different sides. Alignments and all the analysis can easily be shared among different members of the team.

Now that an efficient platform exists that permits the quick organization and reorganization of data, we can start to develop a sense of the features in the antibody that confer affinity or provide a desired solubility.

Looking for Sequence Activity Relationships

There are numerous tools in **SARvision|Biologics** that can be used to identify trends in the data, from *dendrograms* that provide a graphic measure of relation among the sequences, to *invariant maps* that allow us to examine what happens to the data whenever a given residue is at a certain position. However, to identify the patterns that increase affinity, two tools are of particular value: Logo Plots and Mutation Cliffs. In this dataset, we know that mutations have been carried out in the CDR3 of the heavy chain. Using our templates we can reduce all our analysis to that region.

A *LogoPlot*, sometimes called sequence logos, provide a graphical representation of the sequence conservation of amino acids in a series of proteins. We combine the LogoPlots with the use of filters to provide a dynamic view of the frequency with which residues are found for a certain range of property values. In the figure we look at the sequence logo for those antibodies that have good EC50 values and reasonable solubility.



Full Sequence
Motifs: 14
ID:

Region of Interest
Motifs: 1
ID:

CDRs
Motifs: 6
ID:

	Mutations	EC50 (ng/mL)	Solubility (ug/mL)	CDR-H1							CDR-H2					CDR-H3													
				H26	H27	H28	H29	H30	H31	H32	H52	H53	H54	H55	H56	H100	H100	H100	H100	H100	H100	H100	H100	H100I	H100	H100	H100	H101	H102
				G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	W	Y	F	D	V
1	W108Y	6.74	126.41	G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	Y	Y	F	D	V
2	W108Y	7.44	126.43	G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	Y	Y	F	D	V
3	W108Y	7.88	126.38	G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	Y	Y	F	D	V
4	W108F	21.96	121.96	G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	F	Y	F	D	V
5	none	22.00	100.00	G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	W	Y	F	D	V
6	W108T	22.16	155.99	G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	T	Y	F	D	V
7	W108T	22.16	156.09	G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	T	Y	F	D	V
8	W108F	22.99	121.99	G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	F	Y	F	D	V
9	W108F	23.05	121.98	G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	F	Y	F	D	V
10	W108L	23.19	122.13	G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	L	Y	F	D	V
11	W108I	24.11	117.76	G	Y	T	F	T	N	Y	T	Y	T	G	E	Y	P	H	Y	Y	G	S	S	H	I	Y	F	D	V

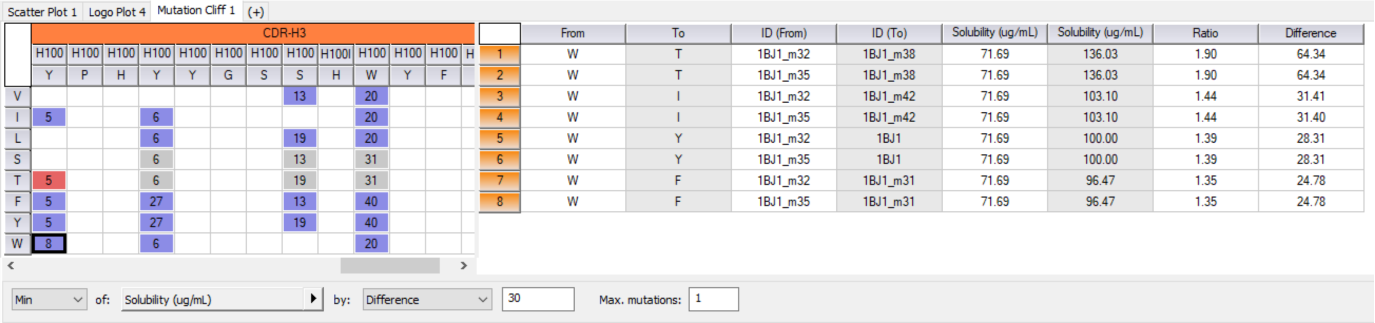
Subset 1
EC50 (ng/mL): 6.74 – 60
Solubility (ug/mL): 100 – 174.07
delta.Prime delta Stability: 257.21

Filter by: (none)
Mark by: (none)
Color scheme: Clustal
Chart style: Sequence logo
Bar labels: Short name
Filter residues:
□ (X) Unknown
☑ (|) Chain Break
☑ (G) Glycine
☑ (A) Alanine
☑ (V) Valine
☑ (I) Isoleucine
☑ (L) Leucine
☑ (S) Serine
☑ (T) Threonine

Selection Groups
□ DataTableView
□ Scatter Plot 1
☑ Logo Plot 4

It is important to note that all graphs are interconnected, that is, selecting a portion of the graph can be used to highlight any other graphic or table open in the program. For example selecting Tyr in the logoplot for positions H100A and H100D results in filtering the table with only those molecules that have those residues in that positions.

Mutation Cliffs are also very powerful tool. The program provides a two way entry table, as columns each position of the protein is shown and each row shows all the different residues that have been tried for that project. In the example below, in position H100, Ile has been tried 5 times and Trp 8 times in different cases. Below the user chooses what type of property they are interested in and what threshold level in that property is of interest. In this case, we select solubility, and we want to see pairs of molecules that differ in solubility by 30 units. We can also choose ratios. Each cell is then colored red gray or blue depending on whether there is at least one pair of molecules that meet the criteria in a positive or negative direction. Gray is used for those pairs that don’t have any pair meeting the criteria. When a cell is chosen (in this case W in H100, we can see the cases that have been identified, for example going from W-> T resulted in a difference in solubility of 64 units.



	CDR-H3												
	H100	H100	H100	H100	H100	H100	H100	H100	H100I	H100	H100	H100	H
	Y	P	H	Y	Y	G	S	S	H	W	Y	F	
V								13		20			
I	5			6						20			
L				6				19		20			
S				6				13		31			
T	5			6				19		31			
F	5			27				13		40			
Y	5			27				19		40			
W	8			6						20			

	From	To	ID (From)	ID (To)	Solubility (ug/mL)	Solubility (ug/mL)	Ratio	Difference
1	W	T	1BJ1_m32	1BJ1_m38	71.69	136.03	1.90	64.34
2	W	T	1BJ1_m35	1BJ1_m38	71.69	136.03	1.90	64.34
3	W	I	1BJ1_m32	1BJ1_m42	71.69	103.10	1.44	31.41
4	W	I	1BJ1_m35	1BJ1_m42	71.69	103.10	1.44	31.40
5	W	Y	1BJ1_m32	1BJ1	71.69	100.00	1.39	28.31
6	W	Y	1BJ1_m35	1BJ1	71.69	100.00	1.39	28.31
7	W	F	1BJ1_m32	1BJ1_m31	71.69	96.47	1.35	24.78
8	W	F	1BJ1_m35	1BJ1_m31	71.69	96.47	1.35	24.78

Min of: Solubility (ug/mL) by: Difference 30 Max. mutations: 1

Biopolymers require new tools for the definition of SARs. **SARvision|Biologics** provides some new types of analysis that is specifically geared to identify trends in activity. We plan to continue our research to identify new tools that can be valuable to study the properties of biopolymers and provide guidance as to what steps to take next to accelerate your discovery engine.