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.
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 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 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 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 vice-versa.
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. 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. 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.

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.

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.