JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Development of an Informatics Platform for Therapeutic Protein and Peptide Analytics

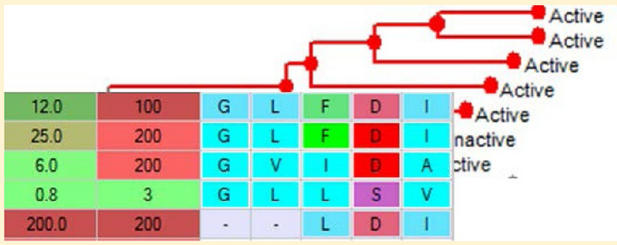
Mark R. Hansen and Hugo O. Villar*

Altoris, Inc., 7770 Regents Rd #557, San Diego, California 92122, United States

Eric Feyfant

Aileron Therapeutics, 281 Albany Street, Cambridge, Massachusetts 02139, United States

ABSTRACT: The momentum gained by research on biologics has not been met yet with equal thrust on the informatics side. There is a noticeable lack of software for data management that empowers the bench scientists working on the development of biologic therapeutics. SARvision|Biologics is a tool to analyze data associated with biopolymers, including peptides, antibodies, and protein therapeutics programs. The program brings under a single user interface tools to filter, mine, and visualize data as well as those algorithms needed to organize sequences. As part of the data-analysis tools, we introduce two new concepts: mutation cliffs and invariant maps. Invariant maps show the variability of properties when a monomer is maintained constant in a position of the biopolymer. Mutation cliff maps draw attention to pairs of sequences where a single or limited number of point mutations elicit a large change in a property of interest. We illustrate the program and its applications using a peptide data set collected from the literature.



INTRODUCTION

Biologics research has grown extremely fast over the past decade.[1] Breakthroughs in protein therapeutics have been provided for the treatment of major inflammatory diseases,[2] cancer,[3] and even neurological disorders[4] as well as their more established role in the management of diabetes[5] or side effects of chemotherapy.[6] Their successes have resulted in major shifts in pharmaceutical research.[7]

Large pharmaceutical companies have been acquiring biological-molecule platforms to be integrated into their research organizations. At the same time, the number of a smaller biopharmaceutical companies solely devoted to the development of antibodies, peptides,[8] or RNA-based[9] therapeutics continues to grow. In most of these companies, large-scale discovery efforts are underway for biologics that parallel in intensity the work done for small-molecule therapeutics. Research informatics groups are developing tools to cope with the new reality of significant collections of biopolymers with large associated data sets.[10] As companies tackle the registration process for biologics and other infrastructure obstacles, bench scientists have a critical need for tools to analyze these large data collections.

In the same way that chemoinformatics filled in the need for tools to relate structure to activity in drug discovery when large chemical libraries became the norm, a parallel need exists today for research on biopolymers.[10] When optimizing small molecules, researchers seek to identify relationships among the molecules being pursued and their physicochemical and structural similarity. Likewise, for biologics, libraries of related compounds are generated as sequence variations derived from an endogenous protein or a promising lead. The goal can be the identification of variants of the biopolymers with improved efficacy, specificity, pharmacokinetics, immunogenicity, or toxicity.

Until recently, researchers were able to keep their data on peptides or antibodies with simple spreadsheets because the data sets were small. However, those means to analyze data are quickly being overtaken by the amount of data to be processed and the multivariate nature of the process of therapeutic discovery. The use of spreadsheets is due to their ubiquitous nature, but they are ill-suited to deal with sequence information. At the other end of the research informatics spectrum, sophisticated analysis can typically be done with the excellent tools available for structure-based protein design.[11] Nevertheless, structure-based analysis can be impractical when projects may involve hundreds of molecules evaluated in multiple assays. Tools for organizing and analyzing biopolymer data sets that provide some of the basic features of spreadsheets with a smart understanding of sequence data are scarcely found[12] but are becoming a critical bottleneck to increase the productivity of the bench scientist.

SARvision|Biologics is an application developed to allow bench scientists to relate a biopolymer sequence to other associated data, such as bioactivity or physical properties, regardless of the size of the data to be analyzed.[13] This tool

Received: June 7, 2013
Published: September 19, 2013

dx.doi.org/10.1021/ci400333x | J. Chem. Inf. Model. 2013, 53, 2774−2779

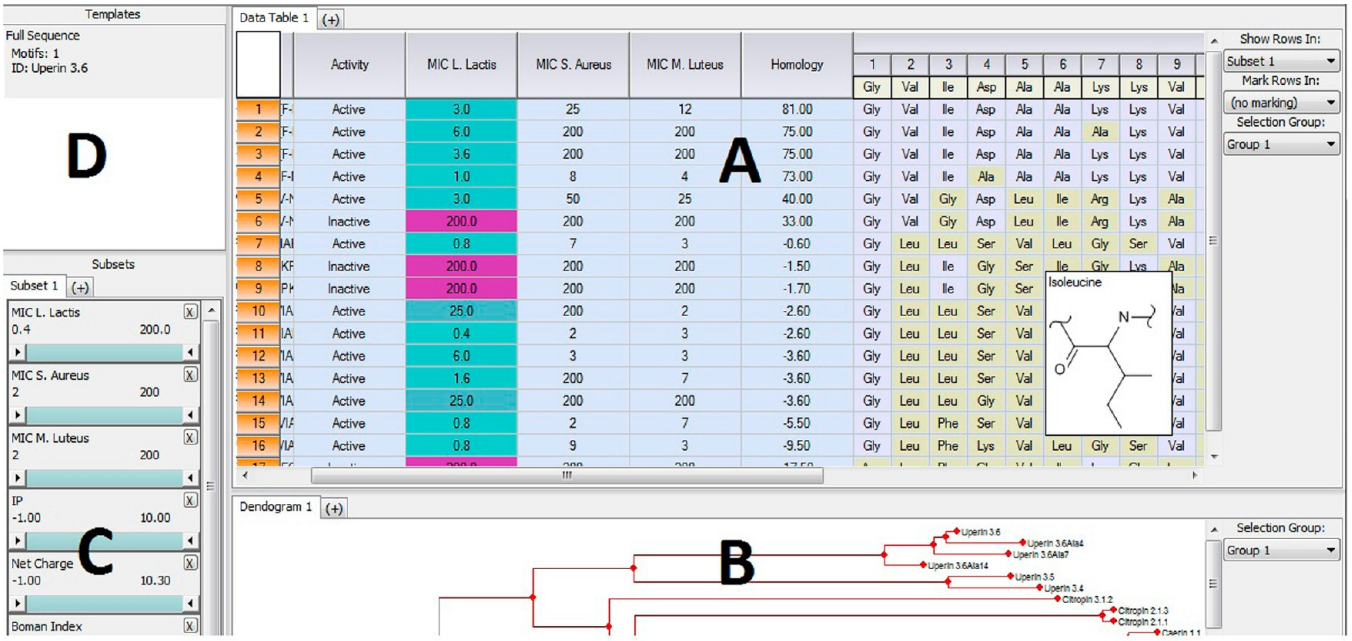


Figure 1. User interface for SARvision|Biologics. The interface contains four subpanels: (A) the main panel where sequence and other data are displayed; (B) a panel for filtering with visual aids such as dendrograms, plots, and invariant or mutation cliff maps; (C) a control panel for filtering according to numeric properties; and (D) a template definition and control panel.

brings together basic data manipulation, analysis, and visualization tools under a platform that is capable of handling primary-structure information.

At a first glance, the program can be regarded as a smart spreadsheet with bioinformatics tools embedded to deal with sequence information, while in the same application we provide tools to analyze the numerical and text data. Once these basic requirements for preliminary data analysis are satisfied, the stage is set for enhanced algorithms to facilitate the identification of relations between sequence and biological or physicochemical parameters.

In the next sections, the implementation of the program is described, and some new concepts for research informatics applicable to biologics research are introduced. The use of the program is illustrated using some public domain information on peptides. Peptides were chosen because in some ways they present the widest, most complex data sets, but they also make clear the potential use of this software in other areas such as antibody research, epitope mapping, or phage display.

IMPLEMENTATION

SARvision|Biologics (Altoris, Inc., San Diego, CA; www.chemapps.com) is a desktop application written in C++ and designed to run under MS Windows. The most straightforward use of the program allows the user to read-in sequence and other associated data from a comma-delimited file. Alternatively, it can be embedded into workflow applications. In the back-end, there are some challenges faced during the implementation that are worth discussing and that required the use of simple knowledge-based algorithms to address them.

Data Input. An initial challenge faced in the development of this tool was the multiplicity of nomenclatures that exist for biopolymers. Specifically, the annotations found in the literature and in data repositories when dealing with covalently modified biopolymers is quite large. Some of the variability in nomenclature is being standardized with the deployment of registration systems, but there is still a lack of centralized systems to deal with information in this area. In addition to the HELM approach,[10] other attempts to obtain a systematic nomenclature for biopolymers were reported.[14−16] Some research groups have adopted ad-hoc nomenclatures, which although they are easy to understand for a human are difficult for the development of algorithms to deal with information on biopolymers because of their variability. We are primarily interested in the data analysis aspects of the biologics problem, but we had to confront the variability of nomenclatures used in the emerging registration systems to bring the data into the analysis environment. Users may have already deployed an approach, so we wanted to provide great flexibility regarding the input files. We addressed the challenge of multiplicity in the nomenclatures by a combination of simple rules to be followed by the users when building the input files combined jointly with the use of knowledge-based techniques. Two knowledge bases that can be generated and edited by the users at will are at the center of the solution that we used.

The first knowledge base contains a list of monomers of interest. A platform with broad applicability in the area of biologics research requires addressing the issue of the potentially large number of monomers that can be used. In addition to natural monomers (such as the 20 natural amino acids or the basic nucleotides), there are large numbers of unnatural monomers that can be used for research purposes. The monomer knowledge base includes properties that can be associated with it, such as molecular weight, structure as SMILES strings, and different physicochemical properties that can be the basis of additive calculations or coloring schema. In addition, for each residue, a list of synonyms can be added, which alleviates the lack of uniformity in the nomenclature.

The second knowledge base is a list of transformations that the monomers may undergo. Examples of transformations include stereoisomery, methylation, and pegylation. The modifiers knowledge base shows how the modification is to

dx.doi.org/10.1021/ci400333x | J. Chem. Inf. Model. 2013, 53, 2774−2779

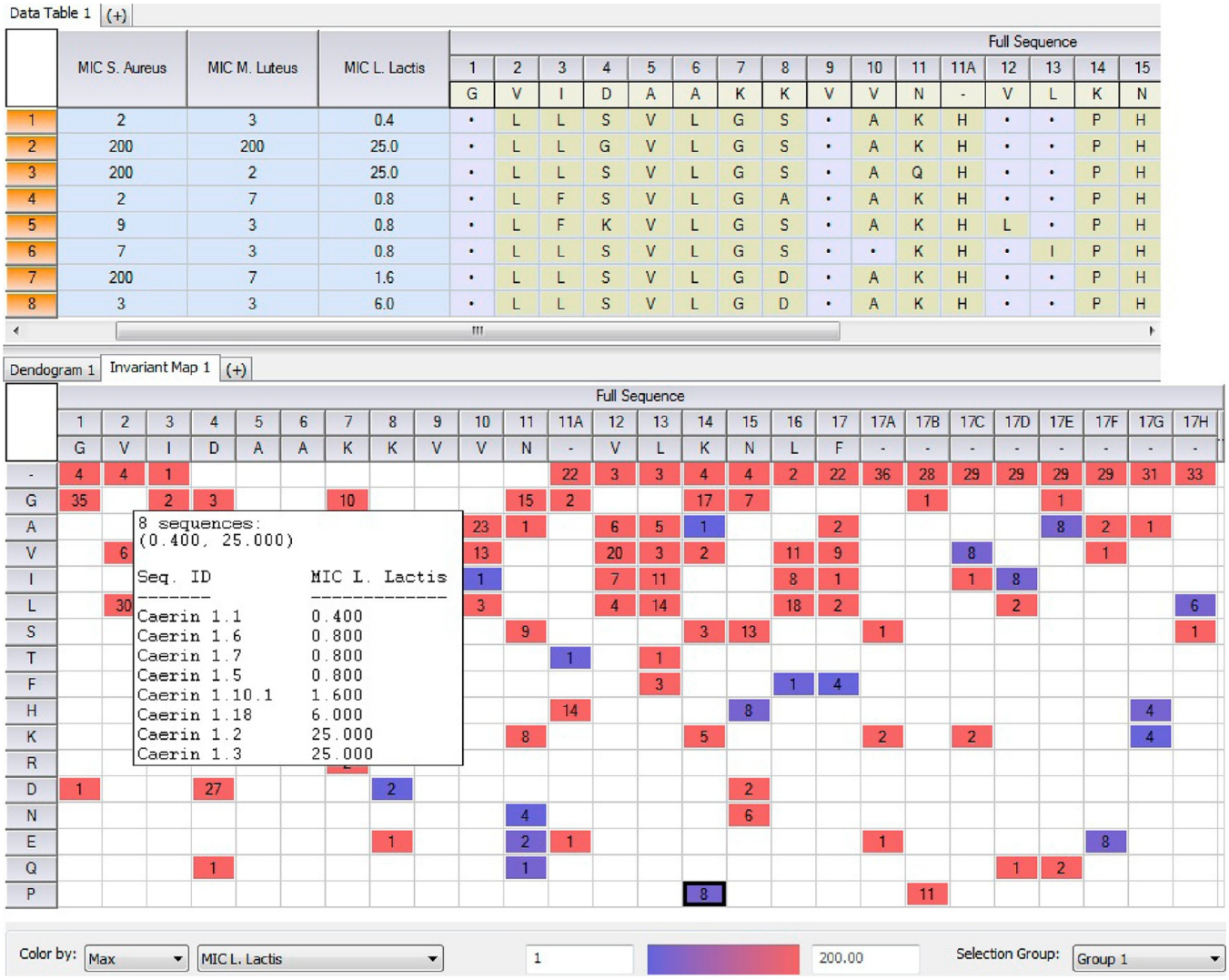


Figure 2. Invariant maps. This is a two-entry heat-map table that records the residues tested at each position of the biopolymer. Each cell is colored depending on the values of the variable selected when a given residue remains constant at a position in the sequence. When more than one value is available for a given position (multiple biopolymers have the selected residue in the given position), the user can select if the cells are to be colored according to the maximum, minimum, or average value.

be found in the input file and the changes it introduces to the monomer properties, for example, in molecular weight. Other issues, such as cyclization, are addressed by the user in the input file. A numeric character following a residue indicates that the residue is linked to another with the same number. Processing the input through both knowledge bases enables the program to read a very large number of monomers, including those that have been modified, while providing the researcher with some flexibility as to the nomenclature used. The user can quickly adapt any nomenclature used in existing data repositories by simply editing the knowledge base, whereas in others some minor scripting may be needed.

Sequence information is processed by Clustal W[17] in the background as it is read-in. The user has the option to read in prealigned files as well. Associated data can be simultaneously read or it can be read in from a separate set of files in comma-delimited format. The alignments can be edited by the user using the program interface by inserting or deleting gaps.

Interface. The interface has multiple working areas, as shown in Figure 1. For illustration purposes, we compiled a list of 50 anti-infectives peptides together with their minimum inhibitory concentration (MIC) values for three different bacterial strains (*Lactococcus lactis*, *Staphylococcus aureus*, and *Micrococcus luteus*) from the literature. The main window, labeled A in Figure 1, displays the sequence and a spreadsheet with any other data that was read in, which, in this case, is the MIC for the three bacterial strains. Cells in the spreadsheet containing data can be sorted in increasing or decreasing order. Any position in the sequence can be sorted as well. The order for the sorting of the monomers is defined by the user in the monomer knowledge base, which includes a priority column that is used to rank the residues' priority. As with all other data in the knowledge base, the column and therefore the sorting order can be edited by the user.

Data columns can be colored according to a gradient that depends on their values, which is typically referred as a heat map. New columns can be generated that are the result of Boolean or arithmetic operations on existing data.

Monomers in the table are colored distinctively if they are modified (for example, it is the D-amino acid). The color of the residue name depends on what the modification is, as defined in the Modifiers Table. The background color of a monomer

dx.doi.org/10.1021/ci400333x | J. Chem. Inf. Model. 2013, 53, 2774−2779

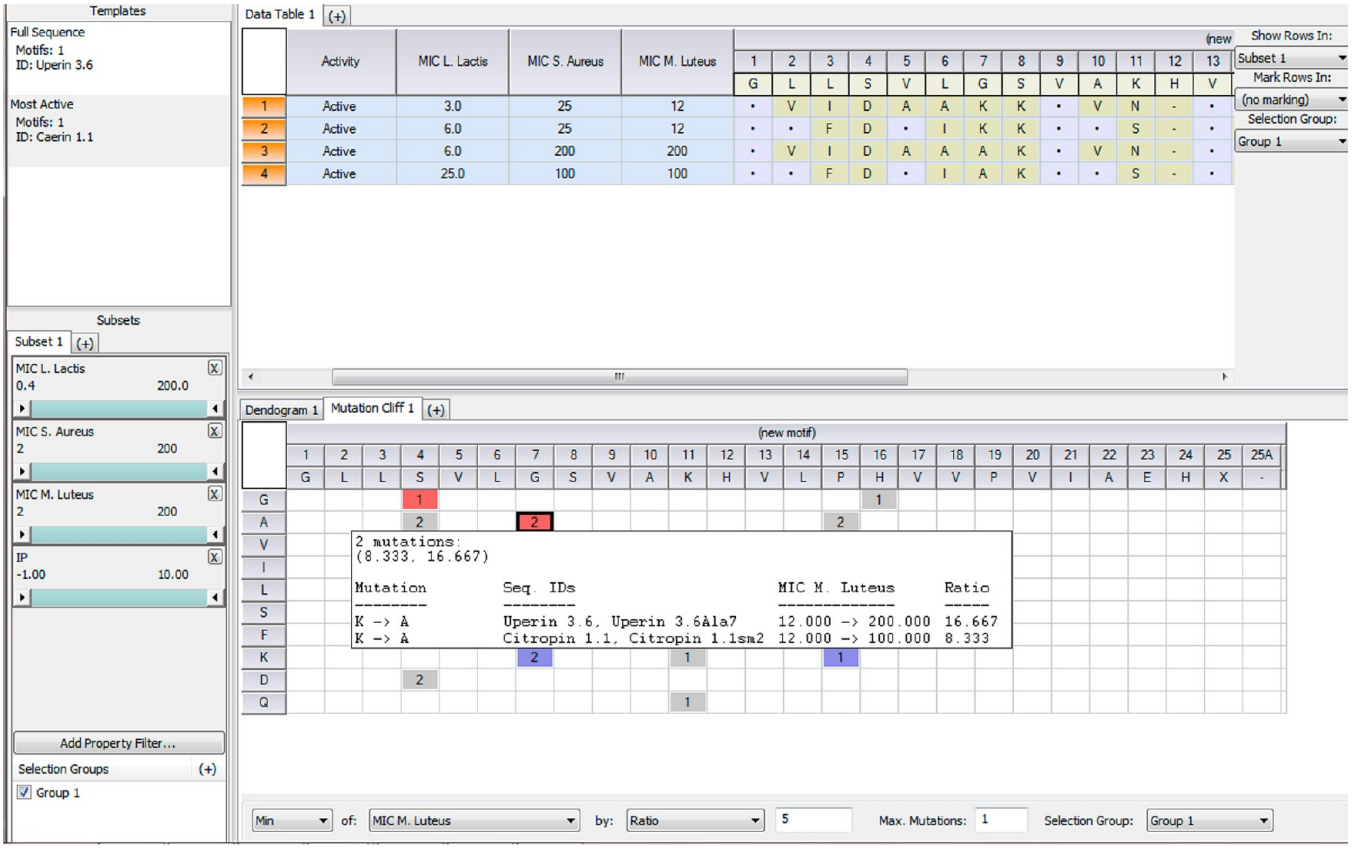


Figure 3. Mutation cliffs. This is a two-way entry heat-map that provides a tool to quickly identify the positions in the sequence that are the most affected by residue replacements in beneficial or detrimental directions. The cell colors indicate the direction of the change. The user can select the number of mutations that can occur simultaneously to meet the criteria as well as the expected change expressed either as a ratio or as a difference. When the cursor hovers over a cell, the mutation and its associated values are shown.

cell can be chosen on the basis of the monomer properties. For example, the background could be colored based on charge or hydrophobicity if the parameters are provided in the monomer knowledge base. In the figure, a blue background shows that the residue is conserved relative to the template, but a gray background shows that it is not.

When the cursor is placed on a monomer cell, it displays the information available on it, which may include its structure, weight, and list of modifications. In the example, the structure of isoleucine is shown. This feature is particularly valuable when dealing with unnatural monomers because the codes used may not be sufficiently descriptive. The depiction is based on the SMILE string provided by the user for the monomer in the knowledge base.

Filtering. The program provides different ways to filter information. The user can select ranges for the numeric properties to be displayed and can limit the number of compounds under analysis. The area labeled C in Figure 1 shows how the data filtering has been implemented with sliding rulers.

Another type of filtering is by sequence similarity. As the data was read in and the sequences aligned, the program generates a dendrogram. The user can select branches of the tree, which results in displaying in the spreadsheet only the subset of sequences in the leaves, as shown in Figure 1 (the area labeled B). The dendrogram enables the user to analyze subsets of related sequences and thus to filter into structurally related sequences.

Template Definition. The template is a reference sequence used to organize the data set. The first sequence read in from the input file becomes the template sequence by default. In the current version of the program, where a multiple sequence alignment algorithm is used, templates are mostly meant to orient the user with respect to the numbering system. In addition, the program adds a homology and identity parameter relative to the template that may be useful to sort the data according to similarity to the template.

Figure 1D shows where different templates can be generated and selected. The user can define more than one template. The templates can be different reference sequences in the data sets but also different regions in a sequence, for example, users working on protein engineering may be interested in only particular areas of the protein, such as the complementarity determining regions in antibodies. The sequences are aligned on the basis of the complete protein sequence. Even if the template focuses in on a region of the protein, in this version of the software the alignment is based on the entire sequence. This area of the interface allows the user to select among different templates that can be defined and to toggle among them promptly. Multiple templates give the user access to different organizations of the data depending on the template selected.

■ SEQUENCE−ACTIVITY RELATIONS

As data sets for biologics become more abundant and complex, scientists need new ways to visualize and identify patterns. To

dx.doi.org/10.1021/ci400333x | J. Chem. Inf. Model. 2013, 53, 2774−2779

this end, we are putting forward two new tools to identify trends in the data that have been found to be useful in the analysis that has been implemented in SARvision|Biologics.

Invariant Maps. In some instances, analysis of the data when a certain residue remains invariant can provide valuable insights regarding the importance of that position in the protein or peptide for the activity being analyzed.

An invariant map is a two-way entry table where the columns are the position of the residues numbered according to the template and the rows are a list of all of the amino acids found in the collected data. Each cell is then colored according to the maximum, minimum, or average value of a chosen property. In Figure 2, we show an example of an invariant map for the antibacterial peptides.

The advantages of this representation are many. First, it allows the user to see what residues were tested for each position. Cells that remain blank correspond to residues that were not tested at that position. For example, in the sample data set, only G and D were tested in position 1. The number in each cell reports the number of molecules with that residue at the position. Second, the color in the cell indicates the trend for that residue in that position. In Figure 2, the color of the cell depends on the maximum value for the antimicrobial activity, expressed as minimum inhibitory concentration (MIC), against *L. lactis*, as selected by the user using the menus at the bottom of the user interface. The cell is blue if the maximum MIC for the subset still retains significant antibacterial activity, and red is for those sets where at least one peptide is essentially inactive. The user could choose averages or minima as well.

The interface is set up so that if the user clicks on one of the cells, the associated biopolymers are shown. Figure 2 shows that the user has selected P at position 14, which is slightly off-blue. When the mouse hovers over the cell, the user can see the list of peptides in the set. At the same time, the data table shows the peptides in question, which can be seen to have P as an invariant residue in position 14 of the template.

Mutation Cliffs. Researchers are typically interested in small structural changes that elicit a large response in the activity being analyzed. In the case of small molecules, those changes are described as activity cliffs and can be analyzed by matching pairs.[18,19]

For biopolymers, we aimed to recreate a similar notion with the concept of mutation cliff maps. The maps are a visualization of changes in the primary sequence of the biopolymer that elicit large changes in activity. The biopolymer changes can be the result of single point mutations or multiple point mutations. The program compares all pairs of sequences that differ in up to the number of point mutations selected, and it highlights those that show the desired change in activity. The results are displayed in a mutation cliff map.

A mutation cliff map is a two-way entry table and can be seen in Figure 3. As for the invariant maps, the residues used at any position are listed as rows, whereas the columns are the aligned sequences position-numbered according to the template. Here again, the sequence of the template is provided under the position number solely for the purpose of orienting the user, as it plays no role in defining the mutation cliff. Each of the cells is colored depending on the change in a selected property. A change of x-fold in activity at a given position will result in the coloring of the cell, where x is selected by the user. The change can be a ratio for properties such as binding constants or a difference for properties such as thermal stability. Blank cells indicate that there are no pairs of sequences that differ in less than the number of point mutations selected.

The figure shows the mutation cliff maps for pairs of peptide antibiotics. The user selects, at the bottom of the program, the property, for example, MIC for *M. luteus*, and the ratio, for example, as a 5-fold difference, for single point mutations. There are quite a few blank cells, indicating that there are not many sequences that differ in a single point mutation. Blue-colored cells indicate that one or more pairs of sequences have been found that lower the MIC when looking at a single mutation by at least 5-fold. Red cells show the converse, and gray is for pairs that do not satisfy the x-fold condition. The mutation cliff maps have built in symmetry because a point mutation that improves the activity appears in reverse results to demonstrate a loss of activity. Therefore, for every red cell, there will be a blue cell. Gray cells indicate that there is a point mutation and that it does not satisfy the criteria for a cliff. The number in the cell indicates how many pairs of sequences show mutations to the residue corresponding to the row.

When the user hovers over the cell with the mouse, the mutations and property values are shown. If the user selects the cell, then a pop-up table shows the peptides and associated information. The example shows two mutations in position 7 where any residue in that position has been mutated to Ala. Because the cell is red, we know that the mutation to Ala in position 7 resulted in a loss of activity. The pop-up table shows the two pairs of sequences where the residue being mutated was Lys. A point mutation of Lys to Ala in position 7 results in loss of activity in both cases. Single point mutations are the simplest case, but the analysis could be carried out allowing for multiple mutations. The figure also illustrates the antisymmetry in the table. The cell that shows that mutations to Lys in position 7 is colored blue, indicating the mutations to Lys increase activity.

Mutation cliffs maps provide a quick assessment of the positions in the peptide (or any other biopolymer) that are the most sensitive to variability. Among other uses, it can greatly speed up the analysis of residue-scan results. The concept that we illustrate here with peptides can be easily applied to other types of biopolymers for any properties of interest.

FINAL REMARKS

The importance gained by biologics in research has revealed a significant deficiency in our tool chest in research informatics. Although we have large collections of tools for small-molecule discovery, the tools available for biologics research do not provide the same level of sophistication in the analysis. Until recently, protein and peptide engineering involved a limited number of compounds, and the design effort has been focused on analysis of structural information. The large volumes of data are creating a need for some filtering prior to the use of molecular modeling and other advanced computational techniques.

Here, we have introduced our SARvision|Biologics platform to identify relations between sequence and function. We attempted to create a flexible platform that can handle the variability in nomenclature found and is able to handle the various types of biopolymers. By an adequate definition of the knowledge bases associated with the program, the user could potentially work with peptides, proteins, polynucleotides, and other biopolymers. Once a platform for data management and visualization is established, we can work toward developing new

algorithms specifically geared to the analysis of therapeutic biopolymers.

Once the platform is established, new concepts to summarize knowledge are likely to emerge. Here, we introduced a concept parallel to the activity cliffs used in small-molecule cheminformatics. Mutation cliff and invariant maps provide new avenues to look at data related to biopolymers, but the field is likely to witness a significant expansion in the number of tools available for informatics in biologics research. We believe that SARvision|Biologics provides a solid but flexible platform for bench scientists who are trying to cope with the data being collected.

AUTHOR INFORMATION

Corresponding Author

*E-mail: hugo@altoris.com.

Notes

The authors declare the following competing financial interest(s): SARvision Biologics is a commercial product sold by one of our organizations.

REFERENCES

(1) Leader, B.; Baca, Q. J.; Golan, D. E. Protein therapeutics: A summary and pharmacological classification. *Nat. Rev. Drug Discovery* **2008**, *7*, 21−39.

(2) Chan, A. C.; Carter, P. J. Therapeutic antibodies for autoimmunity and inflammation. *Nat. Rev. Immunol.* **2010**, *10*, 301−316.

(3) Weiner, L. M.; Surana, R.; Wang, S. Monoclonal antibodies: Versatile platforms for cancer immunotherapy. *Nat. Rev. Immunol.* **2010**, *10*, 317−327.

(4) Sharma, A.; Sharma, H. S. Monoclonal antibodies as novel neurotherapeutic agents in CNS injury and repair. *Int. Rev. Neurobiol.* **2012**, *102*, 23−45.

(5) Borgoño, C. A.; Zinman, B. Insulins: Past, present, and future. *Endocrinol. Metab. Clin. North Am.* **2012**, *41*, 1−24.

(6) Cooper, K. L.; Madan, J.; Whyte, S.; Stevenson, M. D.; Akehurst, R. L. Granulocyte colony-stimulating factors for febrile neutropenia prophylaxis following chemotherapy: Systematic review and meta-analysis. *BMC Cancer* **2011**, *11*, 404-1−404-11.

(7) Nagle, P. C.; Nicita, C. A.; Gerdes, L. A.; Schmeichel, C. J. Characteristics of and trends in the late-stage biopharmaceutical pipeline. *Am. J. Managed Care* **2008**, *14*, 226−229.

(8) Herrero, E. P.; Alonso, M. J.; Csaba, N. Polymer-based oral peptide nanomedicines. *Ther. Delivery* **2012**, *3*, 657−368.

(9) Burnett, J. C.; Rossi, J. J. RNA-based therapeutics: Current progress and future prospects. *Chem. Biol.* **2012**, *19*, 60−71.

(10) Zhang, T.; Li, H.; Xi, H.; Stanton, R. V.; Rotstein, S. H. HELM: A hierarchical notation language for complex biomolecule structure representation. *J. Chem. Inf. Model.* **2012**, *52*, 2796−2806.

(11) Pantazes, R. J.; Grisewood, M. J.; Maranas, C. D. Recent advances in computational protein design. *Curr. Opin. Struct. Biol.* **2011**, *21*, 467−472.

(12) Vielmetter, J.; Tishler, J.; Ary, M. L.; Cheung, P.; Bishop, R. Data management solutions for protein therapeutic research and development. *Drug Discovery Today* **2005**, *10*, 1065−1071.

(13) *SARvision|Biologics*, version 1.1.; Altoris, Inc.: San Diego, CA.

(14) Siani, M. A.; Weininger, D.; Blaney, J. M. CHUCKLES: A method for representing and searching peptide and peptoid sequences on both monomer and atomic levels. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 588−593.

(15) Jensen, J. H.; Hoeg-Jensen, T.; Padkjær, S. B. Building a BioChemformatics database. *J. Chem. Inf. Model.* **2008**, *48*, 2404−2413.

(16) Chen, W. L.; Leland, B. A.; Durant, J. L.; Grier, D. L.; Christie, B. D.; Nourse, J. G.; Taylor, K. T. Self-contained sequence representation (SCSR): Bridging the gap between bioinformatics and cheminformatics. *J. Chem. Inf. Model.* **2011**, *51*, 2186−2208.

(17) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673−4680.

(18) Maggiora, G. M. On outliers and activity cliffs–why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

(19) Stumpfe, D.; Bajorath, J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.