Proceedings of the 35th European Peptide Symposium
Patrick B. Timmons, Chandralal M. Hewage, Michal Lebl (Editors)
European Peptide Society & PSP, 2018

Sequence-activity relationship analysis for peptide optimization using machine learning techniques

Hugo Villar[1], Jane Razumovskaya[2], Jason Hodges[2], Mark Hansen[2]

[1]University of California San Diego, United States
[2]Altoris, United States

https://doi.org/10.17952/35EPS.2018.281

A significant growth in the number of biologic drugs has led to noteworthy investments in research on bioactive peptides by the pharmaceutical industry. The number and diversity of peptides that are generated and evaluated in a project aimed to identify a clinical candidate has been rising, that created a need to manage ever larger data sets. The lack of adequate informatics tools to manage, analyze or visualize the data and sequences created has become evident to the point where tools used in the past, such as Excel result grossly inadequate for most projects and existing tools do not provide the analytical tools or an integrated platform to correlate sequences with trends in the data.

A few years back [1], SARvision|Biologics was developed to allow bench scientists relate a biopolymer sequence to other associated data, such as bioactivity or physical properties, regardless of the size of the data to be analyzed [1,2]. While developing the tool we were able to dissect the elements that are needed to maximize the information that can be extracted in retrospective studies of peptide series and to develop predictive models. We describe here several aspects that should be attended to create informative structure sequence relationships.

Monomer and Peptide Registration: A system to register monomers and the biopolymers is a precondition for any subsequent analysis. Chemical databases can be used to maintain a list of the monomers. In our case, monomer registration is handled *via* a flat file that contains each of the monomers, including their structure as SMILES strings, and physicochemical properties used to determine whole peptide properties, or to color the sequence according to different properties such as for example hydrophobicity. A database of over 700 amino acids that are used for research purposes is our default. The list could be significantly longer if modifications frequently made on the monomers are considered. Simply allowing for the different enantiomers would more than double the list. A second file that contains a list modifications could be used to avoid very large flat files, and operate on the residue list to generate the needed residue [1,2]. HELM notation [3] is used to register and read–in the biopolymers, but in many cases simple line notations suffice [2].

Sequence Alignments. Once sequences are read in, residues that serve equivalent functions in the sequence have to be identified, this is required to show what changes in the sequence result in changes in the peptide properties. The identification of equivalent residues in a sequence is usually done by analysis of the structure of the peptide or using sequence alignments. Clustal V [4] allows for multiple sequence alignments and is used as a default in our analysis but this approach is not suitable for unnatural amino acids. In those cases, the challenge is to create substitution matrices to appropriately score the alignments. After exploring a variety of possibilities substitution matrices based on similarity in computed physicochemical properties provide a good alternative. For short sequences global sequence alignments such as Needleman-Wunsch should be preferred to local sequence alignments such as FASTA [5].

Exploratory Data Analysis should then be carried out to inform the development of more complex statistical models. That is, patterns, trends, outliers, unexpected results in any existing data, using visual and quantitative methods to get a sense of clues that suggest logical next steps. This retrospective analysis of the data, requires the use of a variety of visualization techniques combined with means to identify trends in the data. SARvision|Biologics includes a number of techniques including Logo Plots, heat maps, and scatter plots, among others that can be correlated with the properties of the residues. Mutation cliffs and invariant maps [1] are part of the standard battery of tools when exploring the SAR of peptides. The exploratory data analysis is not a scripted process but one that is follow to build a construct of what is known so far about the peptides under scrutiny. This analysis is very important to develop a sense of models that can be built and their range of validity. Many times, exploring the existing data set with different visualization tools can provide insights that in themselves can suggest the right direction for the optimization of the peptides.

Predictive models [2] can then forecast different properties of interest. In many cases, molecular modeling techniques with analysis of three dimensional structures have been used. These methods confront the challenge

of identifying the bioactive form of the peptide, i.e. the structure that actually is responsible for the biological activity, given the conformational flexibility of peptides, as well as the potential for different protonation states or tautomeric forms. Machine learning techniques offer an alternative, despite challenges of their own, that include the choice of molecular descriptors, to the selection of the most appropriate technique given the myriad of options available. One option is not to choose and build a very large number of models, using a variety of techniques, from linear regression to support vector machine. Hundreds of models can be generated in a few seconds on a regular personal computer that can be tested for their ability to fit the data and make predictions on data sets that were not part of the model training. The most predictive among those models can be combined and used to make recommendations as to what changes to make in the peptide sequence to optimize the property of interest. Indeed, models for multiple properties can be generated, such as potency, cell penetration, deamidation or half-life, among others. The recommendations can then be made on how to optimize multiple properties in parallel, by applying the different models. This is a significant departure from the typical SAR analysis where properties are optimized one at a time in an iterative cycle.

Our view is that these predictive techniques based on machine learning, when properly applied can greatly reduce the number of steps required to optimize peptides.

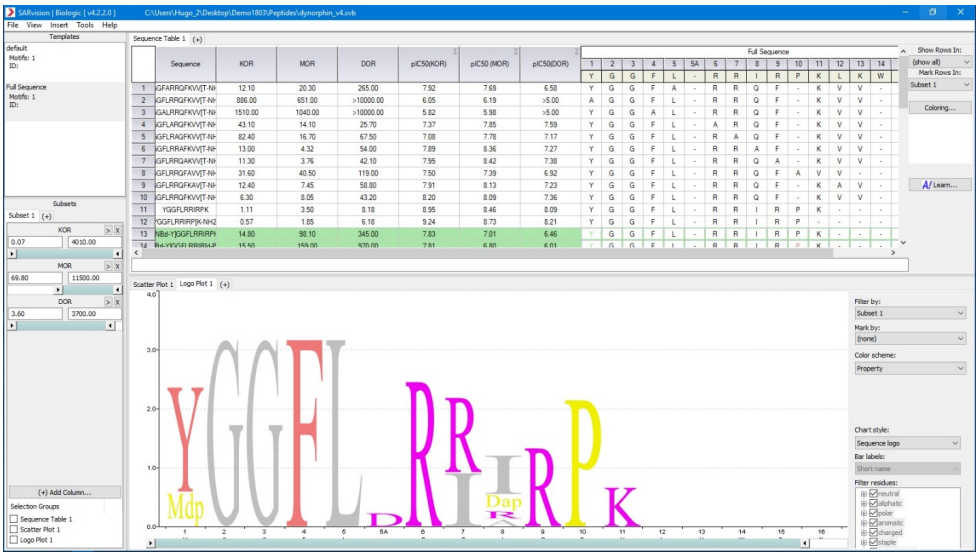

Figure 1: An example of Exploratory Data Analysis for a series of dynorphin analogs. Sequences are aligned with the data in the top right quadrant, while the bottom right shows a Logo Plot for the data. The panels on the right offer the possibility of subsetting the data. Other graphs are available including scatter plots, dendrograms, mutation cliff and invariant maps. All panels are responsive, since changes or selections in one affect the others.

References

[1] Hansen MR, Villar HO, Feyfant E., Development of an informatics platform for therapeutic protein and peptide analytics. J Chem Inf Model. 2013; 53:2774-2779.

[2] ChemApps Resources page, http://www.chemapps.com/resources (accessed: October 2018).

[3] Zhang T, Li H, Xi H, Stanton RV, Rotstein SH., HELM: a hierarchical notation language for complex biomolecule structure representation. J Chem Inf Model. 2012;52:2796-2806.

[4] Higgins DG, Bleasby AJ, Fuchs R. CLUSTAL V: improved software for multiple sequence alignment. Comput Appl Biosci. 1992; 8:189-191.

[5] Morgenstern, B. "Local versus Global Alignments." Sequence Alignment: Methods, Models, Concepts, and Strategies, edited by Michael S. Rosenberg, 1st ed., University of California Press, 2009, pp. 39–54.