Machine Learning Techniques for peptide optimization from sequence information

Mark R. Hansen, Jane Razumovskaya, Jason Hodges, Hugo O Villar
Altoris, Inc., San Diego, CA
www.altoris.com info@altoris.com

Background

The increased interest in peptides, antibodies and other biopolymers as therapeutic agents has made obvious that a significant gap exists in research informatics for this class of drugs. Most of the tools available for peptide design aim to deploy molecular modeling techniques that can be powerful but as datasets become larger and more complex, the tools available to visualize, manage and organize the data are deficient. The problem is particularly acute for peptides where the common use of unnatural amino acids or chemical modifications precludes some of the techniques that are used for protein therapeutics. There is a clear need for tools that facilitate the discovery of relations between sequence and data. Two general types of methods are needed. First, techniques for **exploratory data analysis**, where we aim to discover relations in available data. Second **predictive analytics** techniques that aim to develop models with predictive power. Exploratory data analysis precedes predictive analytics, since a clear understanding of the available data is needed prior to embarking in the development of predictive models.

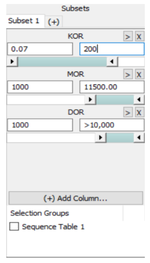
Exploratory Data Analysis for Biopolymers

In the past we have developed a tool to discover relations between data and sequences, a sequence activity relationship tool, named SARvision|Biologics.

Accurate Sequence Alignments are a critical aspect to the analysis and the predictive capabilities. *If the starting alignments are incorrect, then all subsequent studies will be inaccurate*.

Pairwise template based or Multisequence Alignment
Substitution matrices (PAM, Blosum, Identity, **Custom**)
Manual adjustments

A spreadsheet that relates aligned sequences to data provides the first layer of data analysis:

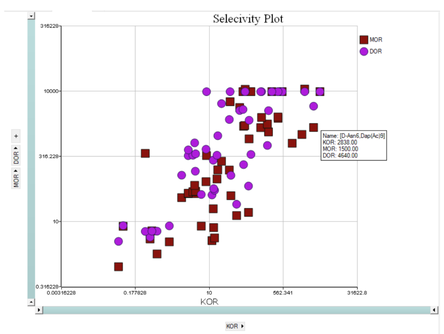


Sequence Table 1

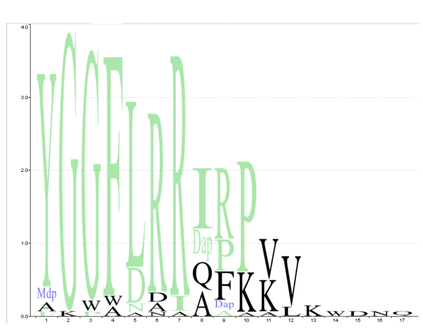
	Sequence	KOR	MOR	DOR
1	[NBzl-Y]GGF[d-D]1RR[dap]1P[K-NH2]	30.30	5880.00	>10000.00
2	[Ac-Y]GGFLRI[R-NH2]	51.00	4315.00	3700.00
3	YGPFLRIRP[K-NH2]	61.00	>10000.00	3900.00
4	[NBzl-Y]GGF[d-N]RR[Ac-dap]P[K-NH2]	66.90	1660.00	>10000.00
5	[NBzl-F]GGF[d-D]1RR[dap]1P[K-NH2]	68.70	11300.00	>10000.00
6	[Nme-Y]GGF[d-D]1RR[dap]1P[K-NH2]	160.00	1490.00	>10000.00
7	[NBzl-Y]GGF[d-D]1RR[dap]1A[K-NH2]	168.00	2550.00	>10000.00

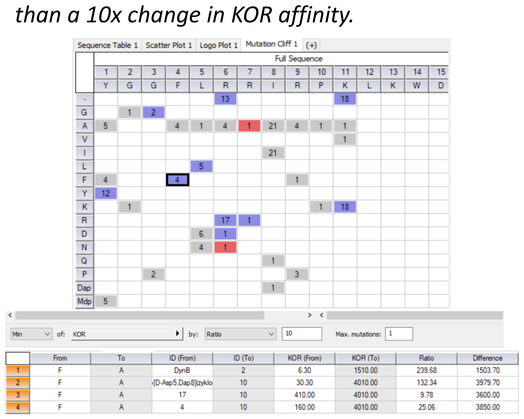
Use tools to subset the data and quickly identify trends in the data. In this case, we carried out an alignment of a series of dynorphin opioid peptides. The tools were available to easily identify the most selective Kappa peptides.

Tools to Identify Critical Residues are important to develop a, understanding of residues that are critical for activity. Models used for predictive techniques should reflect the observations made in the exploratory analysis of data.



Different tools have to be used simultaneously to identify trends in the SAR data. Simple scatter plots, LogoPlots and Mutation Cliffs are some of the tools that can be used to that end. The LogoPlot shows in green the frequency of residues at each position of the KOR selective ligands. In position 5 there is some variability, while positions 2 and 4 have no variability, for the set selected. Mutation cliffs show where mutations in a position happen to have a significant effect in activity. For example in 4 compounds F->A mutations resulted in more than a 10x change in KOR affinity.

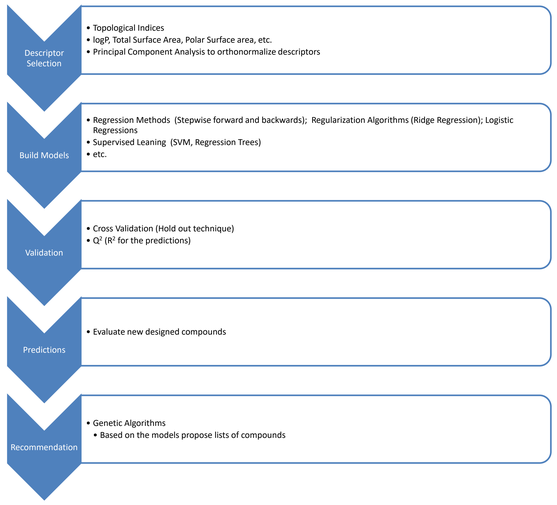


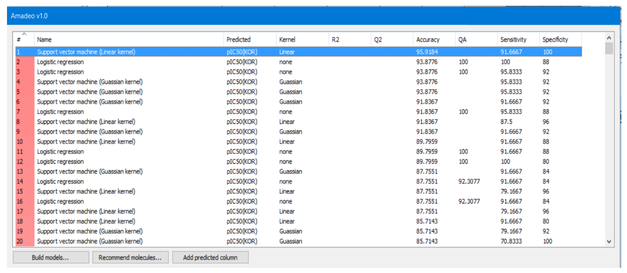


From	To	ID (From)	ID (To)	KOR (From)	KOR (To)	Ratio	Difference
1	F	A	DynB	2	6.30	1510.00	239.68	1503.70
2	F	A	{D-Asp5,Dap8}tyklo	10	30.30	4010.00	132.34	3979.70
3	F	A	17	10	410.00	4010.00	9.78	3600.00
4	F	A	4	10	160.00	4010.00	25.06	3850.00

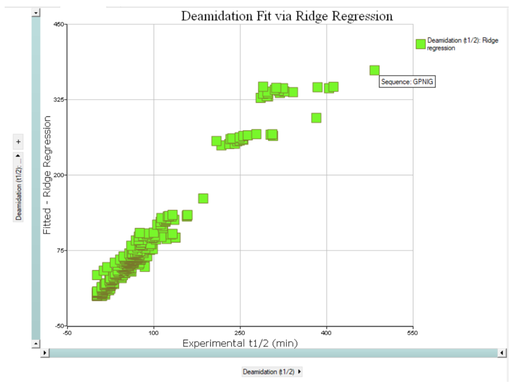
Machine Learning for Peptides

Machine Learning methods can be powerful when applied to the optimization of a compound. Unfortunately, the most appropriate methods for each property to be optimized can be different and require some expertise in their use. A computer system that automatically examines the utility of different machine learning techniques for a dataset and selects the most predictive methods for different properties of clinical interest for peptides was created. The program automatically will **LEARN** about your data, **BUILD** and **SELECT** the most predictive models, and make **RECOMMENDATIONS** as to what compounds to make next.





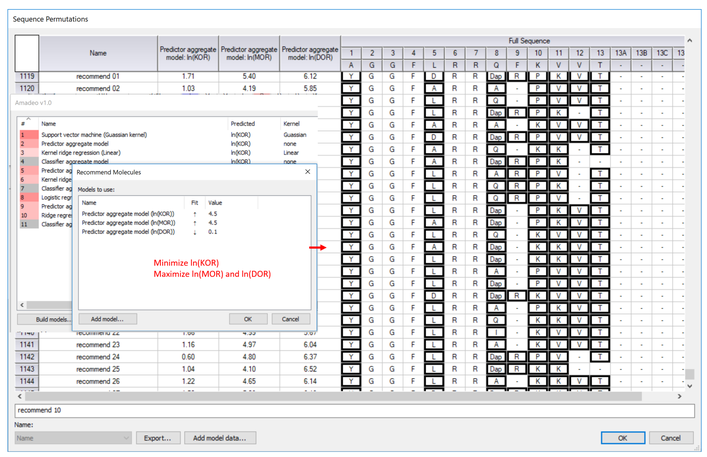
Over 100 models were tried for each Opiate receptor. The program automatically ranks them according to reliability, and in the case shown, SVM comes out as the technique with the most reliability. For other properties, such as deamidation half-life Ridge regression provides a good method (R^2= 0.96; Q^2=0.95)



The exploration of multiple methods and monomer descriptors is now fully automated, which makes these methods available to the bench scientist who may not be familiar with machine learning and artificial intelligence algorithms.

In some cases, the most predictive models are those that result from combining the responses of several models. The program we describe aims to optimize both. The program categorizes the predictive power of the different techniques so that it can recommend molecules that satisfy several parameters simultaneously.

Note that the program is able to handle numerical and categorical data, such as is the case for aggregation.



The models created can be combined to generate lists of molecules that optimize the different parameters of interest. The program combines the different models and proposes new molecules using Genetic Algorithms based on the list of available monomers. The list can be open to any combination, or permutations of residues in selected peptides can be optimized.

In this example, the program will recommend molecules that maximize the pIC50 for KOR and minimize it for MOR and DOR.

Future Directions

The combination of these best in class algorithms provides an avenue to solve the multifactorial problem of peptide optimization. Our ultimate goal is to develop a decision support system that guides the optimization of peptides towards the definition of a clinical candidate minimizing the number of peptides that need to be evaluated and useful to non-experts. Still some open questions remain:

- What is the best way to combine the different models to ensure they are predictive?
- Are the properties we selected optimal ? How to incorporate 3D information?
- Can other sources of data be incorporated as filters that enhance the accuracy of the predictions?

We are currently seeking projects that may benefit from the use of these tools, and that enable us to improve on them.

References

1. SARvision|Biologics, Altoris Inc., San Diego, CA www.chemapps.com
2. Hansen, Villar & Feyfant, J. Chem. Inf. Model. 2013, 53, 2774–2779
3. Joshi, Murray & Aldrich Biopolymers. 2017 Sep;108(5) ; Ramos-Colon et al. J Med Chem. 2016 59:10291; Joshi, Murray & Aldrich J Med Chem. 2015 58:8783; Patkar, Murray & Aldrich J Med Chem. 2009 52:6814