Antibody Characterization with Next Generation Sequencing (NGS) Using GroupMyAbs Shiny Application

Volha Tryputsen,[1] Jocelyn Sendecki,[2] Mark Tornetta[2]

[1]Janssen Research & Development, Raritan, NJ; [2] Janssen Research & Development, Spring House, PA

ABSTRACT

Background:

Next-generation sequencing (NGS), phage display technology and high throughput capacities enables scientists to screen antibody therapy candidates at a level never possible before. NGS information makes possible to characterize antibodies based on their HCDR3 sequence and further group them into families before moving to hit-to-lead stage of drug discovery and development. However, there was no method or software available in-house, tailored for antibody discovery with capabilities to apply biophysical rules to classify the volume of sequences generated.

Methods:

A web based Shiny application **GroupMyAbs** was developed as a collaboration between statisticians and scientists to allow apply biophysical properties for further antibody characterization to the NGS data. Several Multiple Sequence Alignment (MSA) algorithms implemented in the app enable sequence comparability. A method was developed to both: evaluate pair-wise differences between sequences and objectively classify them further into families.

Results:

The app provides custom-made and interactive data visualization, enables refined antibody classification in a mathematically-driven manner, considerably increases efficiency and reduces resources, and insures reproducibility and traceability. This all decreases bias in decision making during the hit-to-lead stage in biologics drug discovery. The app further enables NGS to effectively replace primary screening in an antibody discovery project.

OBJECTIVE

The objective of this project was to develop:

- a method to classify antibodies into families using the NGS data of an antibody variable region as an input;
- a tool which enables antibody classification analyses in an objective, automated, tracible, user friendly and efficient manner

METHODS

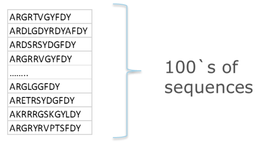
Three steps in antibody classification

1. Make antibody amino acid sequences **comparable**
2. Evaluate pair-wise **differences** between the sequences
3. Classify sequences into **families**

1. Make sequences comparable

Two random sequences X and Y (**Figure 1**) must be aligned before they can be compared. In reality, scientists deal with hundreds or even thousands of sequences which have to be arranged by certain regions within the sequences (HCDR3) that may be a consequence of functional or structural relationship.

Figure 1. Sequences are not comparable



Three multiple sequence alignment (MSA) algorithms have been implemented in GroupMyAbs shiny application:

- ClustalW
- ClustalOmega
- Muscle

There are upwards of many MSA algorithms in use; these 3 were selected on the basis of popularity, efficiency, implementation in R environment, and discussion with the scientist.
At the end of this step all sequences are aligned and therefore are comparable.

2. Evaluate differences between the sequences

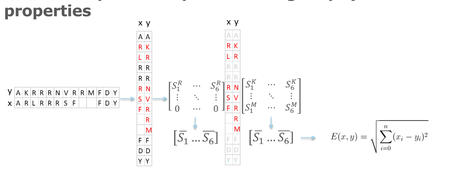
After the sequences are aligned, biophysical properties of amino acids are used to evaluate the differences between a pair of sequences, following the process reflected in **Figure 2:**

- Mismatches in amino acids at each position of aligned sequences are flagged
- For a vector of mismatches respective biophysical scores are used for both sequences: X and Y
- A vector of mean scores is calculated for sequence X and Y separately
- Two mean vectors are taking as an input into Euclidean distance formula E(x, y)
  - where x and y are two sequences,
  - *i* corresponds to a biophysical score.

The same procedure is applied to all pairs of sequences.

During this step, GroupMyAbs enables users to go from aligned comparable sequences to the distance matrix which reflects on sequence dissimilarities.

Figure 2. The process of evaluating the difference between a pair of sequences using biophysical properties



Amino acid	Hydrophobicity at pH 7 [1]	Hydrophilicity scale [2]	Avg side chain orientation angle scale [3]	Polarity scale [4]	H bond donors [2]	pI [5]
A	41	0.28	0.51	0.00	0.17	6.00
R	-14	1.00	0.37	1.00	0.73	10.76
N	-28	0.66	0.12	0.07	0.39	5.41
D	-55	0.79	0.14	0.96	0.30	2.77
C	49	0.07	0.91	0.03	0.31	5.07
Q	-10	0.65	0.02	0.07	0.53	5.65

Note: amino acid mismatches between two sequences X and Y are highlighted in red in Figure 2

3. Classify sequences into families

Hierarchical **cluster analysis** is applied to the set of dissimilarity measures and a dendrogram is used to visualize a classification tree.

It is important to objectively determine the best number of clusters. Five different **indices** (methods) were used to decide on the number of clusters:

- Frey
- McClain
- Cindex
- Silhouette
- Dunn

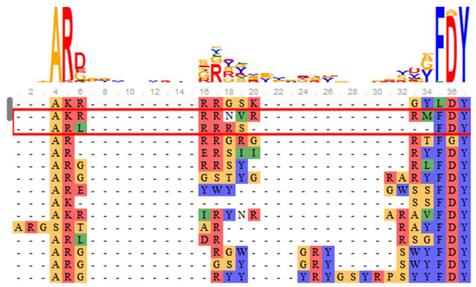
The five indices were chosen on the basis of being able to process dissimilarity measure as an input. The user is given flexibility to evaluate the number of clusters proposed by each method and consider the most frequently proposed number as the optimal number of clusters.

Multidimensional scaling (MDS) is leveraged to visualize the level of similarities between sequences in two dimensional space as an additional unsupervised visualization tool. That is to say, the closer the points to each other on MDS plot, the more likely they are to belong to the same family.

RESULTS & DISCUSSION

The results of MSA are displayed in an interactive manner (partial view is given in **Figure 3**) with the size of an amino acid on the top of the array, corresponding to the frequency of that amino acid in that position. A user of GroupMyAbs has an option of exporting, sorting and filtering of sequences and a choice of a color scheme for MSA results.

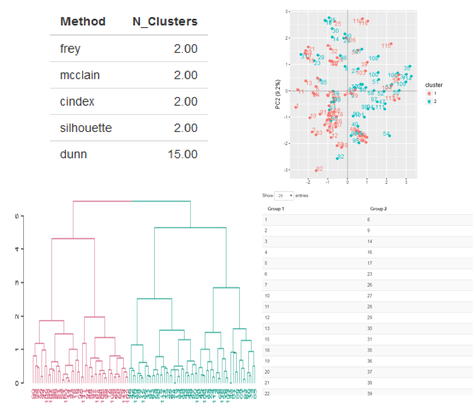
Figure 3. Dynamic Multiple Sequence Alignment



Note: ClustalW method of dynamic multiple sequence alignment was used to produce Figure 3

The results of hierarchical clustering is represented by the dendrogram (**Figure 5b**). The color coding of the branches in the dendrogram is a function of the number of clusters, suggested by different indices (**Figure 5a**). Cluster membership (**Figure 5d**) is also displayed in interactively and has a search option.
The MDS plot (**Figure 5c**) aids in visualizing similarities and dissimilarities between the sequences. Color coding for MSD is equivalent to the one for dendrogram and is a function of an optimal number of clusters.

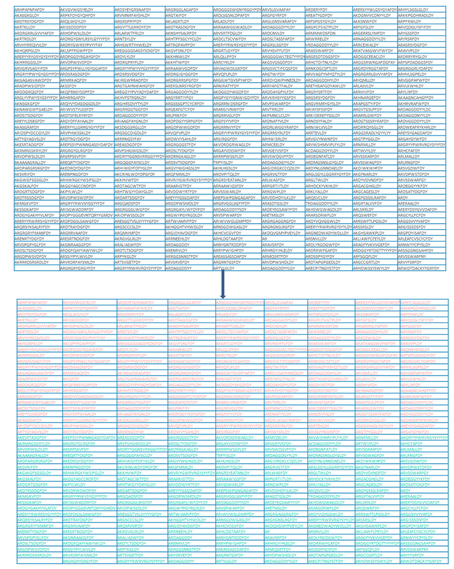
Figure 5.
Suggested number of clusters (top-left, 5a)
Visualization of cluster analysis (bottom-left, 5b)
MDS plot (top-right, 5c)
Cluster membership (bottom-right, 5d)



Method	N_Clusters
frey	2.00
mcclain	2.00
cindex	2.00
silhouette	2.00
dunn	15.00

After classification is complete, the user has an option of downloading MSA results as a report. The results of classification are included as a part of the downloadable report as well.
As depicted in **Figure 6**, **GroupMyAbs** application enable users to go from amino sequences of hundreds of antibodies to classified families.

Figure 6. Antibody Amino Acid Sequences prior (top) and after classification (bottom)



CONCLUSION

GroupMyAbs shiny application:

- Provides customized, objective, mathematically-driven analysis and visualization of NGS data;
- Reduces resources, insures traceability and decreases bias in decision making during the hit-to-lead stage in biologics drug discovery;
- Enables NGS to effectively replace primary screening in an antibody discovery project.

REFERENCES

1. Bodenhofer U, Bonatesta E, Horejs-Kainrath C and Hochreiter S (2015). "msa: an R package for multiple sequence alignment." *Bioinformatics*, **31**(24), pp. 3997–3999.
2. Charrad M et al. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1-36, 2014

PHARMACEUTICAL COMPANIES OF Johnson&Johnson