AnnGen: Learning the Relationship between the Primary Structure of HIV Envelope Glycoproteins and Neutralization Activity of Particular Antibodies

Ghiţă I. Cristian

Abstract

Finding the efficiency of different antibodies against various strains of HIV-1 viruses requires expensive procedures and a preliminary screening could reduce the cost by highlighting the antibodies most probable to be effective. This paper presents an open-source code called AnnGen with an UI to develop artificial neural networks to predict the outcome of future experiments in immunology by using already existing data. The code uses either self-organizing maps to map the patterns of amino acids in HIV-1 envelope glycoproteins or feedforward networks to learn the relationship between the above amino acids and the half maximal inhibitory concentration (I50) of antibodies.

Keywords:

HIV-1; glycoproteins; antibodies; neutralization data; artificial neural network

Code metadata

Current Code version	V1.0
Permanent link to code / repository used of this code version	https://github.com/icghita/ann-relationship-between-hiv-and-antibodies
Legal Code License	Matlab Trial License Id 4479911
Code Versioning system used	git
Software Code Language used	Matlab R2017b
Compilation requirements, Operating environments & dependencies	Windows, Linux, Mac
If available Link to developer documentation / manual	https://github.com/icghita/ann-relationship-between-hiv-and-antibodies/wiki
Support email for questions	ic.ghita@gmail.com

1 Introduction

The HIV-1 virus enters the target cells by using envelope glycoprotein (ENV) gp160, with its subunits gp41 and gp120, the latter which interacts with CD4 receptors on the target T-cells[1]. ENV gp160 is a highly glycosylated glycoprotein organized into five conserved regions, called C1 to C5, and five variable regions, V1 to V5. Its variability, described as sequence variations in the variable regions V1 to V5 and in gp120, leads to resistance to certain antibodies[2] and discovering the relationship between the glycoprotein structure and which antibodies are effective at binding with it constitute a path to developing broadly neutralizing antibodies (bNAbs)[3][4][5].

2 Problems and Background

Multiple bNAbs have been discovered, each with variable neutralization potency against different HIV-1 strains[6][7][8][9]. The variation of neutralization data with respect to the HIV strains is complex[10][11][12] and can be mapped using machine learning[13]. For this, antibodies neutralization data is expressed as IC_{50}, the concentration at which infectivity is reduced by 50%[14].

Two different kinds of networks were used: feedforward neural networks and self-organizing maps. The task of feedforward network, also called supervised training, is to learn from an input dataset and a target dataset, the efficacy of the learning being described by a performance factor, and then use that knowledge to predict new sets of data[15]. That of self-organizing maps, known as unsupervised learning, is to form, using only an input dataset, detectors of different signal patterns[16]. The performance factor used for feedforward networks is the mean squared error:

$$mse = \frac{1}{N}\sum_{i=1}^{N}(t_i - a_i)$$

Where *N* is the number of input-target pairs, *t* is the target and *a* is the output of the network for the input corresponding to the target.

The amino acids in a protein, which are used as the input in these networks, are typically codified in .FASTA files as strings of letters, which must be converted to values between -1 and 1 to be fed to the network. The network is sent as its input the array containing all the converted amino acids of a protein, and all the proteins in a .FASTA file make a dataset.

The way in which the input data of a neural network is formatted can influence the result, thus different codifications for the amino acids were implemented as in Figure 1 from [17].

For codification A-20 , called “A (Numerical)” in the UI, amino acids are converted to integers, instead of boolean, then normalized to [0, 1] before being fed to the network. The network thus produced will be said to have a single layer with its input dimension being the length of the protein (Figure 2). For the other codifications named in the UI: “A-6 (Properties codification)”, “A-9 (Properties codification)” and “B (Raw Properties)”, each amino acid is converted into 6 (for A-6 and B) or 9 (for A-9) values representing physical properties or representations of these properties, which again are normalized to [0,1]. The resulting network will be said to have 6 or 9 layers, each with an input size the length of the protein (Figure 3).

The antibody data can also be formatted by using classes. A class contains all the virus strain entries with antibody neutralization values between two limits. Three classes are created separated by two limits, the classes being codified as “0”,” 0.5” and “1”.

3 Software Framework

The software has been implemented in Matlab, using functions from the Neural Network Toolbox and the user interface has been made using GUIDE. The training algorithm for feedforward networks can be chosen from a list of all available in the Toolbox, some being more efficient for smaller datasets, like Levenberg-Marquardt, while other being more suited for larger datasets, like Scaled Conjugate Gradient. In the case of self-organizing maps, four distance functions are available for calculating the distances between neurons when training the network: Euclidean, box, link and Manhattan distance.

Hardware resources: Performance can be improved for feedforward networks using parallelism. If this option selected, the software will attempt to create a thread for each of the CPU's cores and divide the problem among them. If a graphical processing unit (GPU) is available, it can be used to train a feedforward network, but using only the Scaled Conjugate Gradient algorithm.

4 Practical Usage

The software can be downloaded from the repository at https://github.com/icghita/ann-relationship-between-hiv-and-antibodies. Documentation with details about its usage can be found at https://github.com/icghita/ann-relationship-between-hiv-and-antibodies/wiki.

The software can be used by running the AnnGen.m file with Matlab, preferably 2016 or above, and with the Matlab path set to the one containing all the source and figure files.

5 Illustrative Example

Using an Intel i7-6700HQ processor, we have made a feedforward neural network with the following parameters:

- Protein Codification: A (numerical)
- No. of ANN iterations: 100
- No. of Hidden Neurons: 10
- Training Function: Levenberg-Marquardt
- Data Division: Training from 0 to 70%, Validation from 70% to 85% and Test from 85% to 100%
- No Parallel Computing, GPU or Classes

The input files can be found in the repository and are taken from http://www.hiv.lanl.gov [18] and compiled from [19][20][21][22] [7][6][9][14][23]. The storage file is a path to a .mat file.

The correlation coefficients are: 0.78152 for the training set, 9.2171 for the target set and 0.94829 for the test set.

Figure 1

Amino Acid	Representation (A)			Representation (B)					
	20	6	9	Volume	Bulkiness	Flexibility	Polarity	Aromaticity	Charge
A	10000000000000000000	53500a	100001000	0.1677	0.4433	0.2490	0.3951	0.0	0.5
C	01000000000000000000	45500a	000001001	0.3114	0.5506	0.2048	0.7441	0.0	0.5
D	00100000000000000000	a5001a	001001000	0.3054	0.4532	0.8675	1.0000	0.0	0.5
E	00010000000000000000	a60010	001000100	0.4970	0.5567	0.8112	0.9136	0.0	0.5
F	00001000000000000000	185a0a	100100101	0.7725	0.8976	0.0763	0.0370	0.6667	0.5
G	00000100000000000000	525000	000001000	0.0	0.0	1.0	0.5062	0.0	0.5
H	00000010000000000000	565010	010100111	0.5569	0.5632	0.1124	0.6790	0.5556	0.5
I	00000001000000000000	37500a	100010011	0.6467	0.9852	0.6707	0.0370	0.0	0.5
K	00000000100000000000	a7a02a	010000111	0.6946	0.6738	0.6867	0.7901	0.0	1.0
L	00000000010000000000	375000	100010001	0.6467	0.9852	0.2811	0.0	0.0	0.5
M	00000000001000000000	375005	100000100	0.6108	0.7033	0.0	0.0988	0.0	0.5
N	00000000000100000000	655020	000001000	0.3174	0.5156	0.6747	0.8272	0.0	0.5
P	00000000000010000000	555000	100000010	0.1766	0.7679	0.8594	0.3827	0.0	0.5
Q	00000000000001000000	66502a	000000100	0.4910	0.6048	0.7952	0.6914	0.0	0.5
R	00000000000000100000	a8a046	010000110	0.7246	0.5955	0.9398	0.6914	0.0	1.0
S	00000000000000010000	645010	000001011	0.1737	0.3222	0.8514	0.5309	0.0	0.5
T	00000000000000001000	55501a	000001010	0.3473	0.6771	0.5984	0.4568	0.0	0.5
V	00000000000000000100	365007	100011010	0.4850	0.9945	0.3655	0.1235	0.0	0.5
W	00000000000000000010	0a5a10	100100110	1.0	1.0	0.0402	0.0617	1.0	0.5
Y	00000000000000000001	285a18	000100111	0.7964	0.8008	0.5020	0.1605	0.6667	0.5
X	11111111111111111111	555555	111111111						

Note: (A) binary representations (20-, 6-, and 9-bin), used by Brusic et al. (1995), where rep "6" assigns a 6-place string where each place is a scalar value for a feature (hydrophobicity, volume, charge, aromatic side chain, hydrogen bonds) or a correction bit. "a" stands for 10; "Rep 9" is an intermediate representation using a feature-based grouping of amino acids. (B) used by Milik et al. (1998), where every amino acid is represented by six properties (volume, bulkiness, flexibility, polarity, aromaticity, and charge).

Figure 2

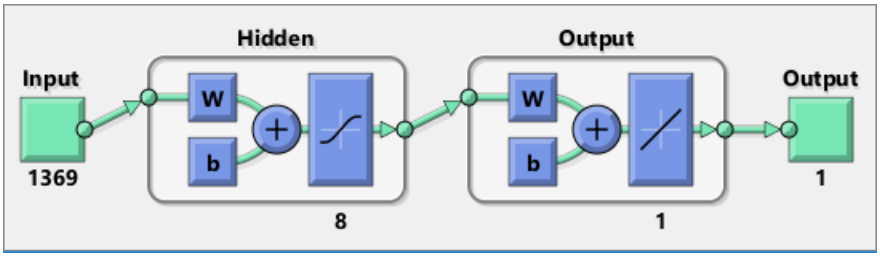


Figure 3

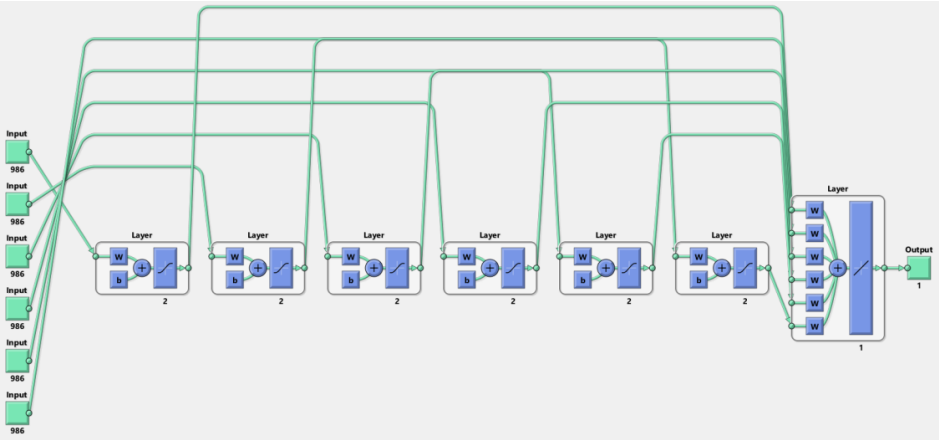
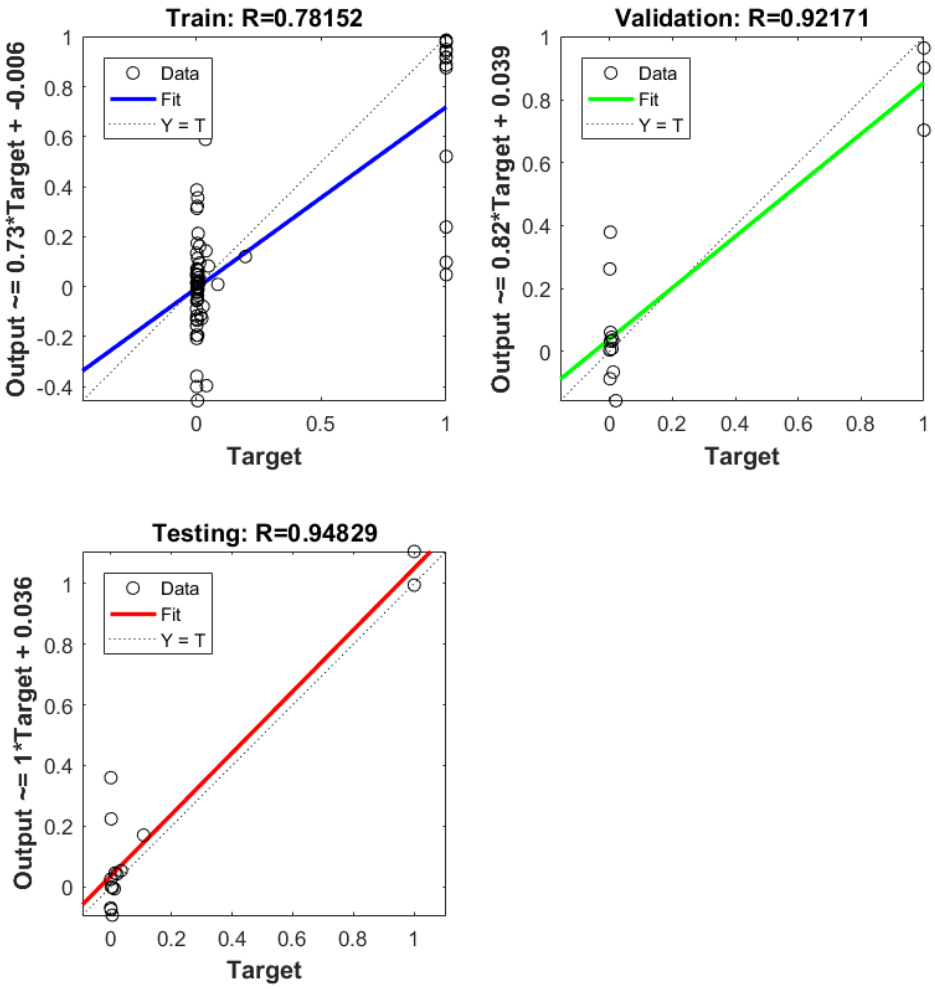


Figure 4



6 Conclusions

We have presented AnnGen, an open source tool for taking advantage of the capabilities of the Neural Network Toolbox of Matlab in the context of identifying patterns in proteins. Further updates will be intended to optimize the codification of proteins and the algorithms used.

Acknowledgements

This work is based on the paper [13] and has been made thanks to the supervision of Prof. Cătălin Buiu.

References

[1] J. de Wit and S. Seager, "Crystal Structure of a soluble cleaved HIV-1 envelope trimer," *Science (80-.).*, vol. 342, no. 6165, pp. 1473–1477, Jan. 2014.

[2] A. Merk and S. Subramaniam, "HIV-1 envelope glycoprotein structure.," *Curr. Opin. Struct. Biol.*, vol. 23, no. 2, pp. 268–76, Apr. 2013.

[3] J. Chen *et al.*, "Mechanism of HIV-1 Neutralization by Antibodies Targeting a Membrane-Proximal Region of gp41," *J. Virol.*, 2014.

[4] A. J. Hessell *et al.*, "Broadly neutralizing human anti-HIV antibody 2G12 is effective in protection against mucosal SHIV challenge even at low serum neutralizing titers.," *PLoS Pathog.*, vol. 5, no. 5, p. e1000433, May 2009.

[5] L. Yang and P. Wang, "Passive immunization against HIV/AIDS by antibody gene transfer.," *Viruses*, vol. 6, no. 2, pp. 428–47, Mar. 2014.

[6] J. F. Scheid *et al.*, "Sequence and Structural Convergence of Broad and Potent HIV Antibodies That Mimic CD4 Binding," *Science (80-.).*, 2011.

[7] X. Wu *et al.*, "Rational Design of Envelope Identifies Broadly Neutralizing Human Monoclonal Antibodies to HIV-1," *Science (80-.).*, vol. 329, no. 5993, p. 856 LP-861, Aug. 2010.

[8] X. Wu *et al.*, "Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing," *Science (80-.).*, 2011.

[9] L. M. Walker *et al.*, "Broad neutralization coverage of HIV by multiple highly potent antibodies," *Nature*, 2011.

[10] G. Frey, H. Peng, S. Rits-Volloch, M. Morelli, Y. Cheng, and B. Chen, "A fusion-intermediate state of HIV-1 gp41 targeted by broadly neutralizing antibodies," *Proc. Natl. Acad. Sci.*, 2008.

[11] B. K. Chakrabarti *et al.*, "Direct Antibody Access to the HIV-1 Membrane-Proximal External Region Positively Correlates with Neutralization Sensitivity," *J. Virol.*, 2011.

[12] A. P. West, L. Scharf, J. Horwitz, F. Klein, M. C. Nussenzweig, and P. J. Bjorkman, "Computational analysis of anti-HIV-1 antibody neutralization panel data to identify potential functional epitope residues," *Proc. Natl. Acad. Sci.*, vol. 110, no. 26, pp. 10598–10603, Jun. 2013.

[13] C. Buiu, M. Putz, and S. Avram, "Learning the Relationship between the Primary Structure of HIV Envelope Glycoproteins and Neutralization Activity of Particular Antibodies by Using Artificial Neural Networks," *Int. J. Mol. Sci.*, vol. 17, no. 10, p. 1710, Oct. 2016.

[14] D. C. Montefiori, *Evaluating neutralizing antibodies against HIV, SIV, and SHIV in luciferase reporter gene*

assays. Curr. Protoc. Immunol, 2005.

[15] T. M. Mitchell, *Machine Learning*. McGraw-Hill, Inc., 1997.

[16] T. Kohonen, "The Self-Organizing Map," *Proceedings of the IEEE*, vol. 78, no. 9. pp. 1464–1480, 1990.

[17] Y. E. Khudyakov, *Medicinal Protein Engineering*. 2008.

[18] C. Kuiken, B. Korber, and R. W. Shafer, "HIV Sequence Databases," *AIDS Rev.*, vol. 5, no. 1, pp. 52–61, 2003.

[19] R. Diskin *et al.*, "Restricting HIV-1 pathways for escape using rationally designed anti–HIV-1 antibodies," *J. Exp. Med.*, vol. 210, no. 6, p. 1235 LP-1249, Jun. 2013.

[20] H. Mouquet *et al.*, "Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies," *Proc. Natl. Acad. Sci.*, 2012.

[21] J. Huang *et al.*, "Broad and potent neutralization of HIV-1 by a gp41-specific human antibody," *Nature*, 2012.

[22] N. A. Doria-Rose *et al.*, "HIV-1 Neutralization Coverage Is Improved by Combining Monoclonal Antibodies That Target Independent Epitopes," *J. Virol.*, 2012.

[23] R. Diskin *et al.*, "Increasing the Potency and Breadth of an HIV Antibody by Using Structure-Based Rational Design," *Science (80-.).*, vol. 334, no. 6060, p. 1289 LP-1293, Dec. 2011.