Critical analysis of binding avidity hypothesis

Hsiang-Yu Yuan

Date: July 23, 2013

Version 1.1

Section:

Background

System Architecture

Packages Description

Program Flow

Proposed Figures

Software: Projects/BindingAnal/back/BindingAnal-1.1.zip. (Notes: the package doesn’t include the BEAST tree log files)

How to deal with outlier (<http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/>)

Check H1N1 binding score chart.

Background:

This document describes the system design and program flow of the new project based on the previous project entitled *critical analysis of binding avidity hypothesis*. The previous project ends in April, 2013 with main figures generated for thesis work. The files for the old project are mainly located in Z:/Projects/Binding\_Evo, Binding\_Evo\_Tree and Binding\_Phy, etc.

System Architecture

BindingAnal

ParseTree

BuildTree

(from *Binding\_Evo\_Tree*)

BindingScr

(from *BindingChange*)

NetCharge

(same folder name)

*Mparsetree*

*beast\**

*beast\_log*

*mega*

*dscr/BindingChange/v2*

*Mnetcharge\**

*Pnetcharge\**

Ancestral Sequences

Trees

FASTA

Score

Data should store in folder as *keyword\_yyyymmdd*

*e.g.* thesis\_20121002

\*Subversion controlled

System Design

*H3N2 NY*

*BindingScr*

*BuildTree*

*Mnetcharge*

*Pnetcharge*

*copy*

Bindingscore.csv

*dscr/predict bindign*

*copy*

BuildTree\beast

*H3N2 NorAm*

*copy*

out/FASTA

*SampleTrees*

*Sampling*

*SampleYears*

*Annotate Trees*

*Calculate netcharge*

*Extract Anc Seqs*

*merge\_metadata\_nr.m*

merge\_data.mat

merge\_data.mat

generate\_fasta\_for\_beast

*merge\_bding\_2.m*

out/FASTA

*ParseTree*

*Read\_ newick*

beast/

*copy*

*mega: transform*

NEXUS

*Beauty:*

*generate*

*beast*

XML

MCC Tree

*mega: align*

*dscr/predict side change*

*Calculate netcharge*

*dscr/formatyear2.pl*

Packages Description

NetCharge:

* Calculate netcharge distribution
* Integrate binding score with netcharge
* Generate FASTA files for BEAST

BindingScr:

1. Calculate binding score

BuildTree:

1. Generate BEAST Trees

Highlight:

* Annotate net charge on the phylogenetic tree
* Annotate binding score on the phylogenetic tree
* Annotate binding avidity on the simulated phylogenetic tree

Program Flow

Using H3N2\_NY as an example

Creating a phylogenetic tree with netcharge

NetCharge

* 1. Download sequences from Influenza Virus Database and save into Mnetcharge/dat/strain as the following 6 formats:
  + Pnetcharge/seq/hm\_h3n2\_flu\_ny\_any.csv **(METADATA\_ORI)**
  + Pnetcharge/seq/hm\_h3n2\_flu\_ny\_any.fas (not used now)
  + Pnetcharge/seq/hm\_h3n2\_flu\_ny\_any\_del.fas ( \_delimited, not used now)
  + Pnetcharge/seq/hm\_h3n2\_flu\_ny\_any\_gene\_access.txt (allow integration using GenBank ID)
  + Pnetcharge/seq/hm\_h3n2\_flu\_ny\_any\_prot\_access.txt
  + Pnetcharge/seq/hm\_h3n2\_flu\_ny\_any\_simple.fas (**SEQ\_PROT** file, only protein accession)
  1. Download hm\_h3n2\_ny\_ntcds\_simple.fas for BEAST file generating
  2. Extract information from FASTA
  3. Parse age and isolated date information: Pnetcharge/formatyear2.pl(**METADATA\_ORI**) -> **METADATA** (Pnetcharge/seq/hm\_h3n2\_flu\_ny\_ay.csv)

Note: run on DSCR or on Windows.

* 1. Manually create **REDUNDANT\_SET** in ‘hm\_h3n2\_ny\_redundant\_sets.csv’;
  2. Copy **SEQ\_PROT** and **METADATA** into matlab folder **Mnetcharge**.
  3. Calculate netcharge: main\_aa\_dist.m(**SEQ\_PROT**) -> **CHARGE** (hm\_h3n2\_ny\_charge.mat)
  4. Merge metadata: merge\_metadata\_nr (**CHARGE**, **METADATA**, **REDUNDANT\_SET**) -> **MERGE\_DATA**(hm\_h3n2\_ny\_merged\_data.mat)

Note: convert date\_string into numeric date during merging.

{Viruses,binding\_info,gbacc,ages,iso\_date\_num} should become {Viruses,binding\_info,gbacc,ages,iso\_date,iso\_date\_decim,iso\_date\_num}

Original date format: 1993/01/16

* 1. Add gene accession to each protein accession: create\_geneprot\_table(filename, gbacc, merged\_dat) gene accession = filename + '\_gene\_access.txt'.
  2. Merge with predicted binding avidity data from h3n2\_noram\_1968\_2012/scr/bindingscore\_h3n2\_noram\_noh.csv

Note: Check the section in BindingScr to

* 1. Create BEAST FASTA files (hm\_h3n2\_ny\_dna\_beast\_1993\_2006.fas) and save into beast/ for BEAST input
  2. Copy BEAST FASTA file into buildTree/beast/hm\_h3n2\_ny/fasta

PS. Total 686 records from 1993/7-2006/06

BuildTree

* 1. mega:transform Use MEGA to open BEAST FASTA file from BuildTree/beast/hm\_h3n2\_ny/fasta and then convert into NEXUS file into beauti.
  2. Configure BEAST input file using beauty. Note: date format: yyyy-MM-dd.

Length of chain: 10,000,000

Change Log parameters every 10,000

Also,

<logTree id="treeFileLog" logEvery="10000"

<logTree id="substTreeFileLog" logEvery="10000"

Reduce the total number of trees to be 1,000 in the log file. The tree size will be around 100MB.

BindingScr

Input data: hm\_h3n2\_noram\_simple\_1968.fas

* 1. Format sequences (hm\_h3n2\_noram\_simple\_1968.fas; only contains proteinID). First align the H3 template on several sequences (see align\_template.fas). Then copy the H3 aligned result with other sequences (hm\_h3n2\_noram\_simple\_1968\_template.fas).

* 1. Truncate sequences and save as (hm\_h3n2\_noram\_simple\_1968\_truncated.fas; Only retrieve amino acids in certain domain in HA)
  2. Remove gaps (no gaps for H3N2 but some gaps for H1N1). Be sure to remove gaps for H1N1 otherwise might generate wrong predicted structures and corresponding binding scores.

Four files in the folder fasta:

* Original aligned FASTA: hm\_h3n2\_noram\_simple\_1968.fas (for h1n1, one additional file hm\_h1n1\_world\_aligned.fas)
* Aligned Template: align\_template.fas
* Aligned FASTA plus Template: hm\_h3n2\_noram\_simple\_1968\_template.fas
* Truncated and gap removed Aligned FASTA plus Template: hm\_h3n2\_noram\_simple\_1968\_truncated.fas
  1. Upload the FASTA sequences onto DSCR (h3n2\_noram\_1968\_2012/seq\_noh)

1. If need to do multiple alignment (for h1n1), use clustalw2 fast pairwise alignment to generate alignment result on dscr or use bioedit fast multiple alignment on PC (which is preferred).
2. Run readfasta.pl fasta/protein\_h1n1\_ha\_na\_full\_truncated.fas (same length with template) to produce separated sequence files into seq\_noh/
3. Copy template H3.pdb into seq\_noh/.
4. Run job/run\_sidechain.job (make sure –h is included so PDB file won’t include hydrogen atoms) to generate pdb files Predict side Change -> pdb files.
5. Copy pdb files into seq\_noh/pdb/.
6. Run job/run\_binding.job to generate the binding score in scr files.
7. Copy scr files into seq\_noh/scr/.
8. Run parsescore.pl to extract accession and binding score into formatted files. Use excel to parse into a csv file.
9. Copy the score file into mat folder in project 'netcharge'

Note: Use -h option to not adding hydrogen atoms

Store results in seq\_noh (noh means no hydrogen added)

Output: Binding Score, h3n2\_noram\_bding.txt

Jul 26, 2013

* 1. (Option) To build binding score for ancestral nodes, use ancestral\_sequences.prot.fas from Mparsetree. Use MEGA to align and truncate with template. Obtain binding score as ‘h3n2\_ny\_1993\_2005.bsc’. The ID is sequentially assigned from 1 to total nodes. Use excel to extract data and save ‘h3n2\_ny\_binding\_n1371.csv’ into Mparsetree.
  2. Create binding\_txt for tree nodes (main\_binding\_dist('ancestor\_20130713/hm\_h3n2\_ny\_charge\_1993.mat','ancestor\_20130713/h3n2\_ny\_binding\_n1371.csv','ancestor\_20130713/hm\_h3n2\_ny\_binding\_1993.mat')
  3. Annotate binding score on the tree

annotate\_tree\_binding.m

Plot binding score by time

plotBindingByTime.m

pairs = string2pairs('dat/ancestor\_20130713/hm\_h3n2\_ny\_dna\_beast\_1993.topo.nw.mcc.(time).trees',n\_tips);

load('dat/ancestor\_20130713/hm\_h3n2\_ny\_charge\_1993\_2006.mat');

load('dat/ancestor\_20130713/newick\_elements.mat');

For simulated tree, there are three objects, b, d and names.

Extract binding score from names into binding scores array.

How to use b and d to reconstruct tree?

After I have tree use string2pairs to generate pairs.

Use both pairs and binding scores array to create bindingbytime.

Check how to create tree from b and d. <http://www.mathworks.co.uk/help/bioinfo/ref/phytree.html>

(code: Projects\BindingAnal\ParseTree\mat\Mparsetree\plotNetchargeByTime.m)

1. Replace negative values in d to be 0.01, d(find(d<0))=0.01
2. Tree = phytree(b,d)
3. String = getnewickstr(Tree, 'Distances', false, 'BranchNames', false)
4. Copy to nodpad+ and replace all the Leaf smbol.
5. pairs = string2pairs('dat/simphylo/genealogy\_300(2).nw',300)
   1. Running BEAST

Generate ‘hm\_h1n1\_noram\_age\_dna\_beast\_1995\_2006.fas’ in Mnetcharge.

%step1 exclude outliers

metadata = [proj 'hm\_h1n1\_flu\_noram\_ay\_new.csv'];

%step2calculate netcharge

prot\_fasta = [proj 'fasta/hm\_h1n1\_flu\_noram\_any\_simple\_new.fas'];

%step6 generate FASTA file

DNAFile = [proj 'fasta/hm\_h1n1\_noram\_ntcds\_simple1\_new.fas'];

Before running BEAST, guess the date (06/30/yyyy) for any strains that don’t have month and days information.

Do multiple alignment first. Since clustalw sequence identifiers only allow 30 characters. Sequence header I use is {ProteinID\_Date\_NGS\_Netcharge}. Transfer the aln file into FASTA file.

Prerun BEAST with shorter MCMC steps in folder prerun and check whether there are outliers.

Outliers:

ABI19015

AAX56530

ABI20870

AFO66235 (Keep it. Canadian strain, the only strain isolated in 2004)

Remove outliers and run BEAST in 9596-20130730/

95-96 should change to 9506-20130730

Annotate phylogenetic tree with netcharge

Beware:

* + - 1. Need to deal with insertion before alignment. Change --- to CAA.
      2. Use consensus sequence if the ancestral nodes have multiple sequences states. (See taxa2fasta\_new.m). This give the result that the root node has #NGS=7.

To do list:

Aug 1, 2013

Change Beta in individual-based model simulation.

R0 = f x g x n

1. Rho = 1 – 1/R0
2. Rho = 1

Beta = c x rho

Check get\_beta\_list.m

getDeltaV is still same.

Proposed Figures:

Figure 2‑4. The scatter plots of viral netcharge vs host age

Figure 2‑5. The relationship between host age and viral netcharge

Figure 2‑6. The relationship between host age and predicted viral binding score

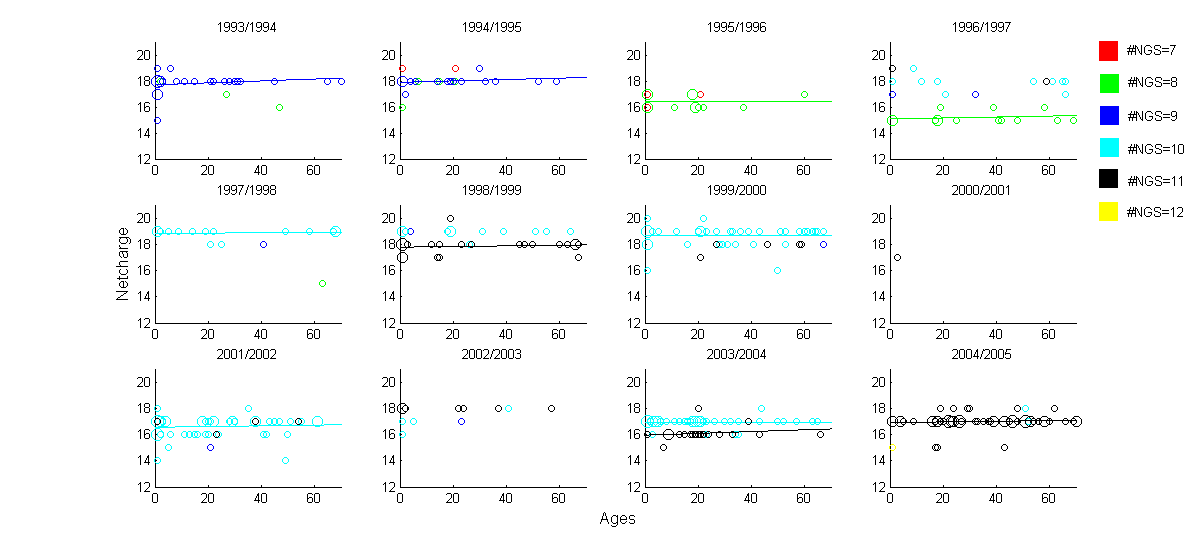
Figure 2‑7. The relationship between host age and predicted viral binding score

Figure 2‑8. Scatterplots of predicted binding scores of viral isolates by their netcharge values

Figure 2‑9. Phylogenetic tree of influenza A/H3N2 viral isolates

Figure 2‑10. A phylogeny of influenza A/H3N2, from viral isolates spanning years 2003 through 2005

Fig1 (Thesis fig2.4):

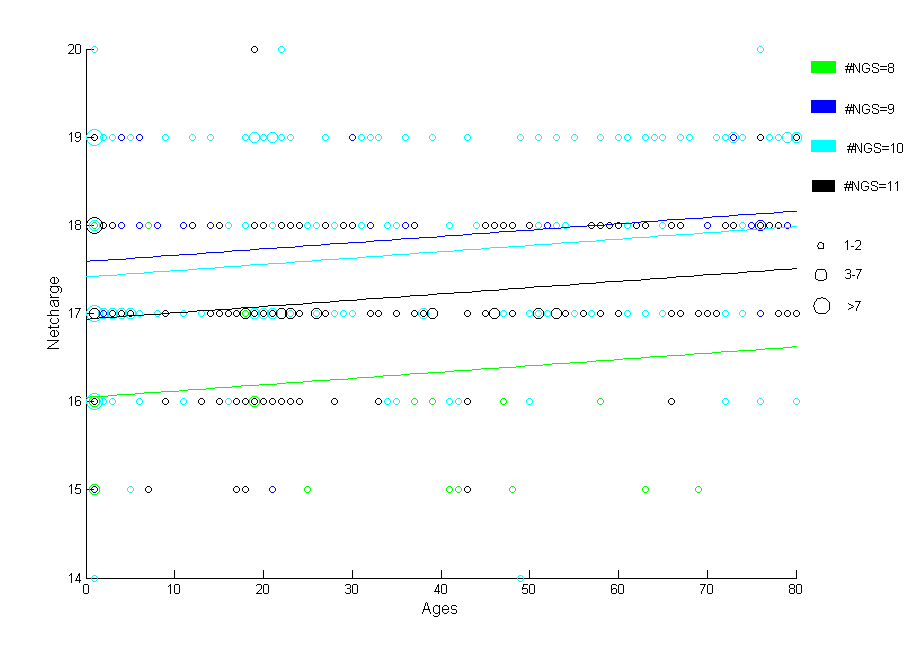


Description: The scatter plots of viral netcharge vs host age. Viral isolates with same number of NGS are grouped by the same color.

Code:

Figure name:

Fig2 (Thesis fig2.5):

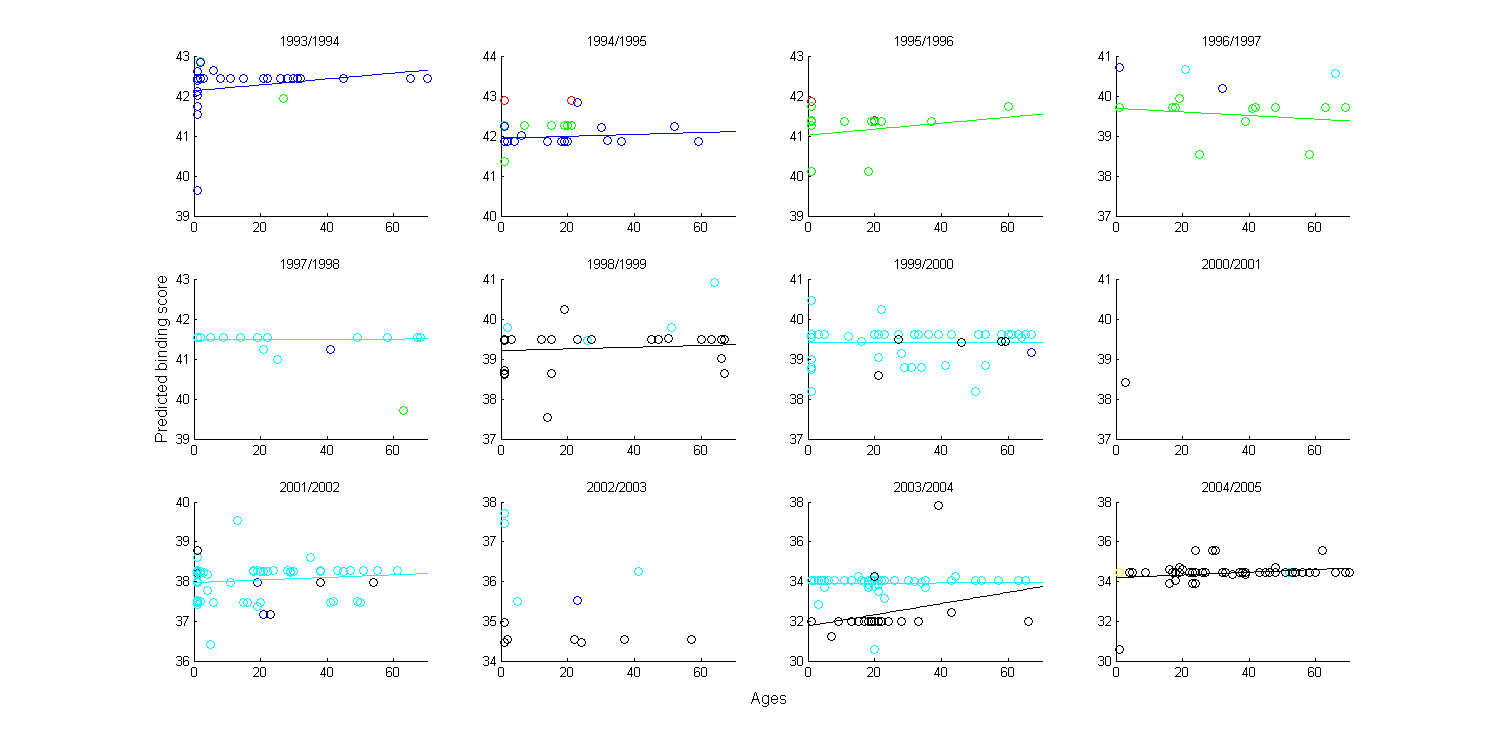


Description: The relationship between host age and viral netcharge. Viral isolates with same number of NGS are grouped by the same color.

Code:

Figure name:

Fig3 (Thesis fig2.6):



Description: The relationship between host age and predicted viral binding score. Viral isolates with the same number of NGSs are grouped by the same color, as in Figure 2‑4.

Code:

Figure name: