Multiple Sequence Alignment

Presenter: Brian Foley

btf@lanl.gov

HIV Databases

Theoretical Biology and Biophysics,

Los Alamos National Laboratory

www.hiv.lanl.gov
seq-info@lanl.gov

Presenter: Brian Foley

btf@lanl.gov

HIV Databases

Los Alamos NATIONAL LABORATORY

Topics

Multiple Sequence Alignment, Primarily HIV-1

General introduction:
HIV, presents special challenges for alignment due to frequent insertions and deletions (in/dels) relative to most other organisms.

Multiple genes, and overlapping reading frames make it difficult to align the genome "in frame" for convenient translation to amino acid sequence.

Tools for aligning HIV genomes and genes.

Special problems and tips for dealing with them.

We will not cover alignment of short reads to genome; Bowtie etc.

Los Alamos
NATIONAL LABORATORY

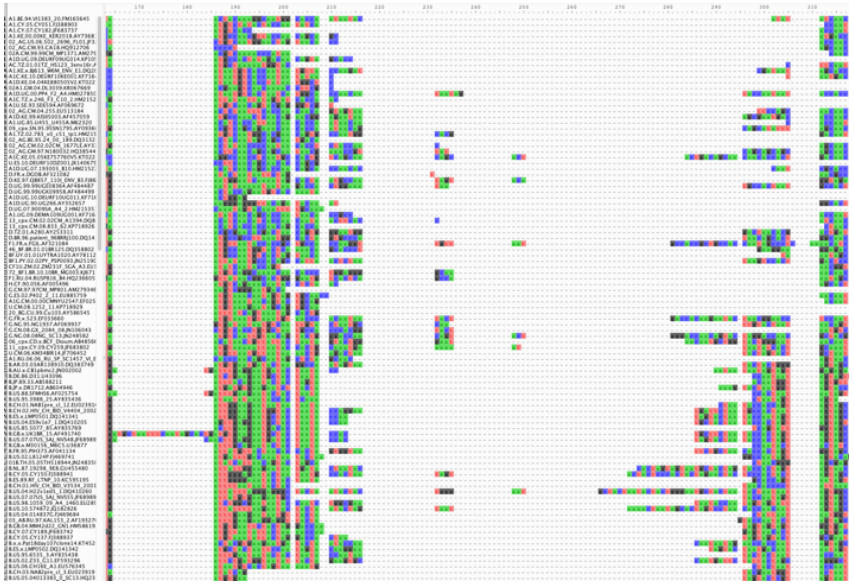
Pairwise vs multiple alignment

Pairwise alignment can be local or global.
Simple, quick and fairly unambiguous.
“Optimal” alignment is well defined.

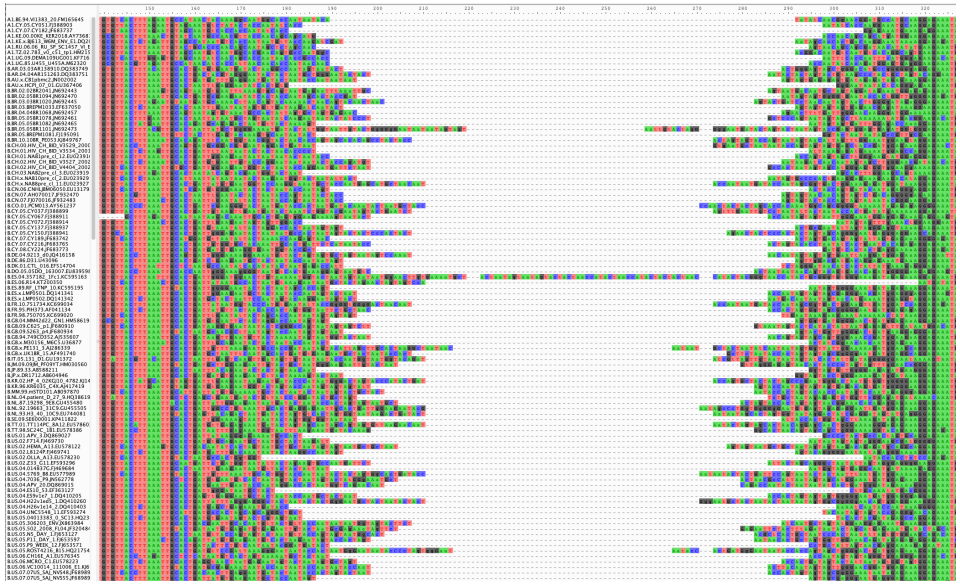
BLAST
Needle (EMBOSS)
Water (EMBOSS)
Align0

Multiple alignment is much more difficult.
“Optimal” alignment is debatable.

ClustalW ClustalOmega
MUSCLE
MAFFT
GeneCutter (HIV specific)



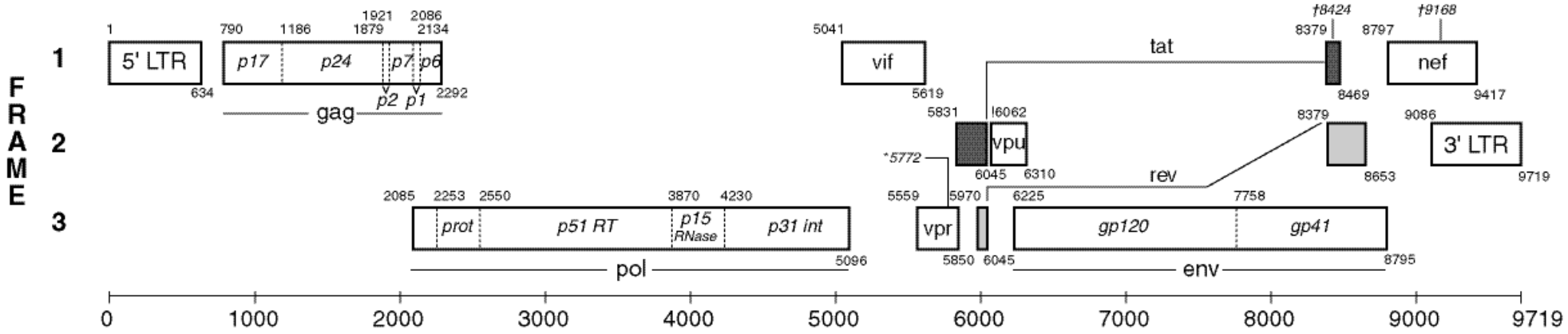
Typical alignment of HIV-1 env by MAFFT or similar tools. Cannot be translated to amino acid sequences.



GeneCutter preferentially puts gaps in between codons, splits regions of uncertain alignment in the center.

Los Alamos
NATIONAL LABORATORY

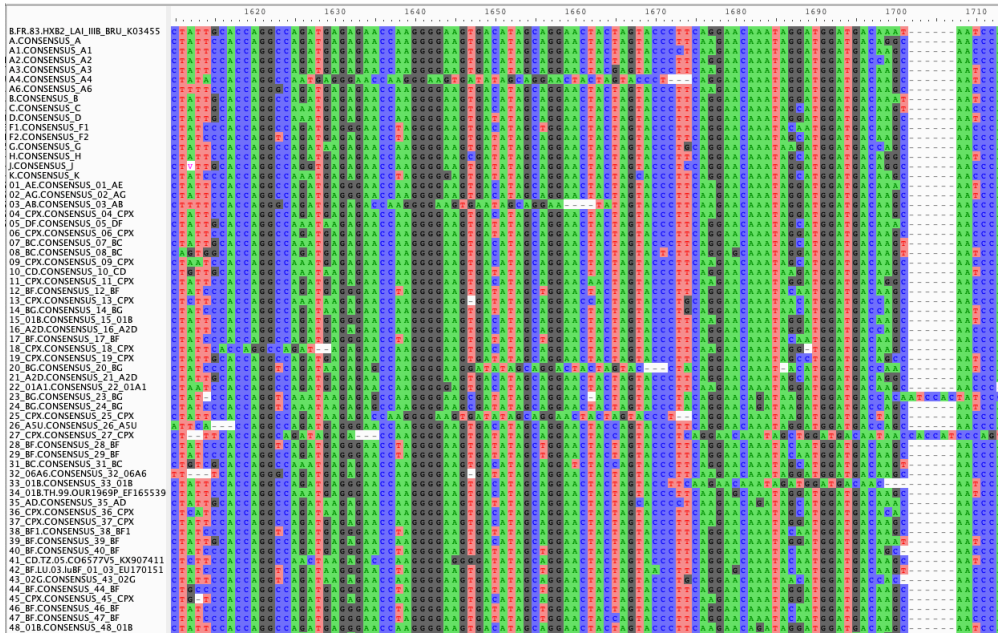
HIV-1 Genome

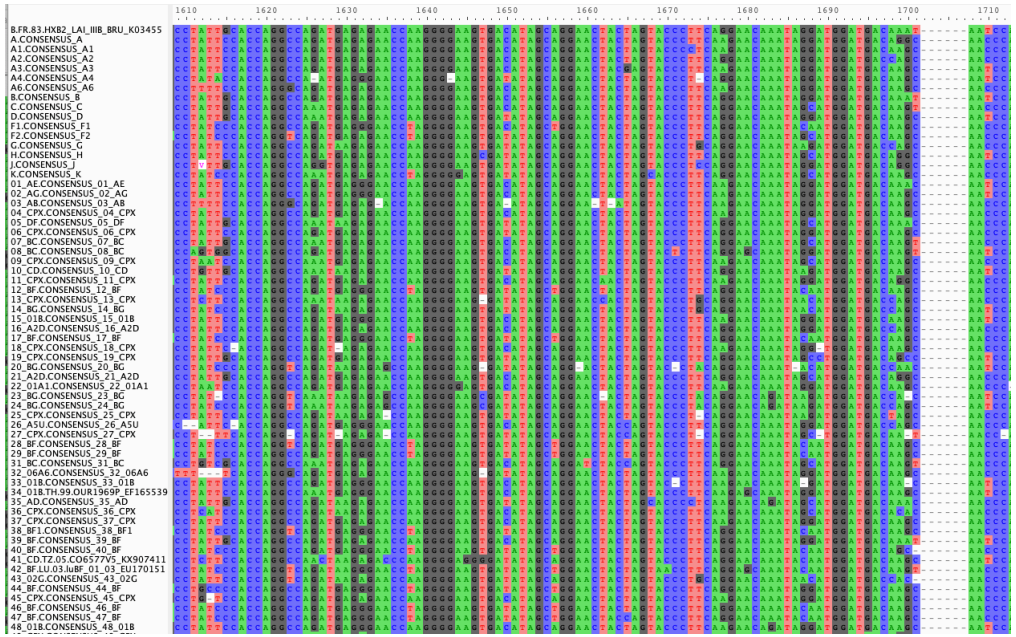


https://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html

Nine proteins/polyproteins produced, 7 regions with overlapping reading frame, one with triple overlap (tat/rev/gp41). Thus, no single genome alignment can be easily translated to the amino acid sequences for all proteins.

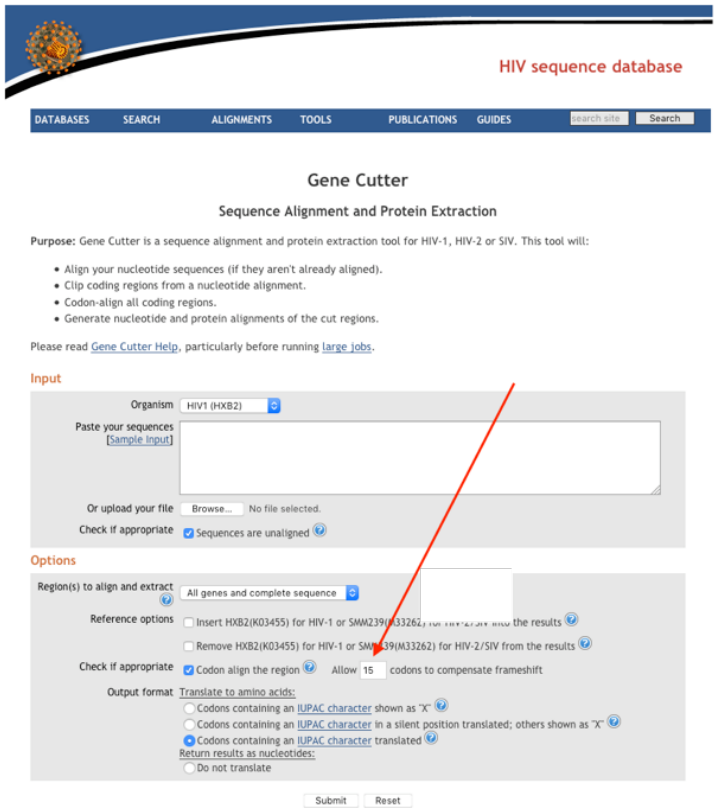
Los Alamos
NATIONAL LABORATORY





GeneCutter has an option for how many codons between compensating in/dels are allowed to be misaligned.

Top example set to 25 codons, bottom zero codons.



Los Alamos
NATIONAL LABORATORY

For many purposes, MAFFT or other multiple alignment tools are the best option, Factors to consider are speed, ability to add the alignment process to a “pipeline” or automated script or program, ease of use, etc. as well as overall “quality” or usefulness of the resulting alignment.

Often, a first pass “quick and dirty” alignment can be highly useful in checking the data, for quality, length variations, etc. before more time-consuming methods are applied.

Always use some multiple alignment viewing tool to at least spot-check the alignment before proceding to further steps. Aliview (Mac) BioEdit (Windows), Pixel, etc..

https://ormbunkar.se/aliview/
http://www.mbio.ncsu.edu/BioEdit/bioedit.html
https://www.hiv.lanl.gov/content/sequence/pixel/pixel.html

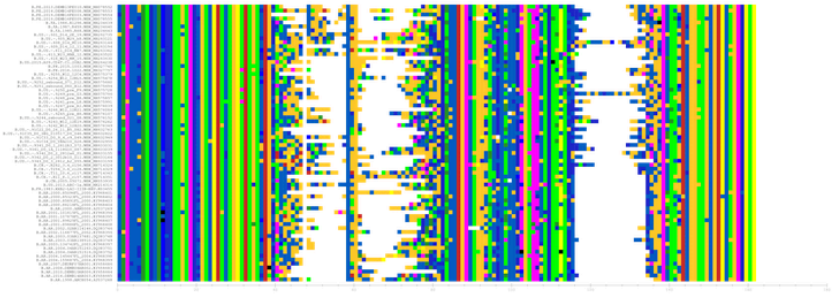
Pixel

Input & Options

Alignment: GlycoExampleV12adj.FASTA
Sequence type: AA
Color scheme: Default
Show differences: No
Plot size: Automatically compute.
Residue scale: 3 x 3 pixels
Show sequence names: Yes
Font size: 3
Show scale bar: Yes
Scale bar size: 3
Margins: 0.5 inches

Results

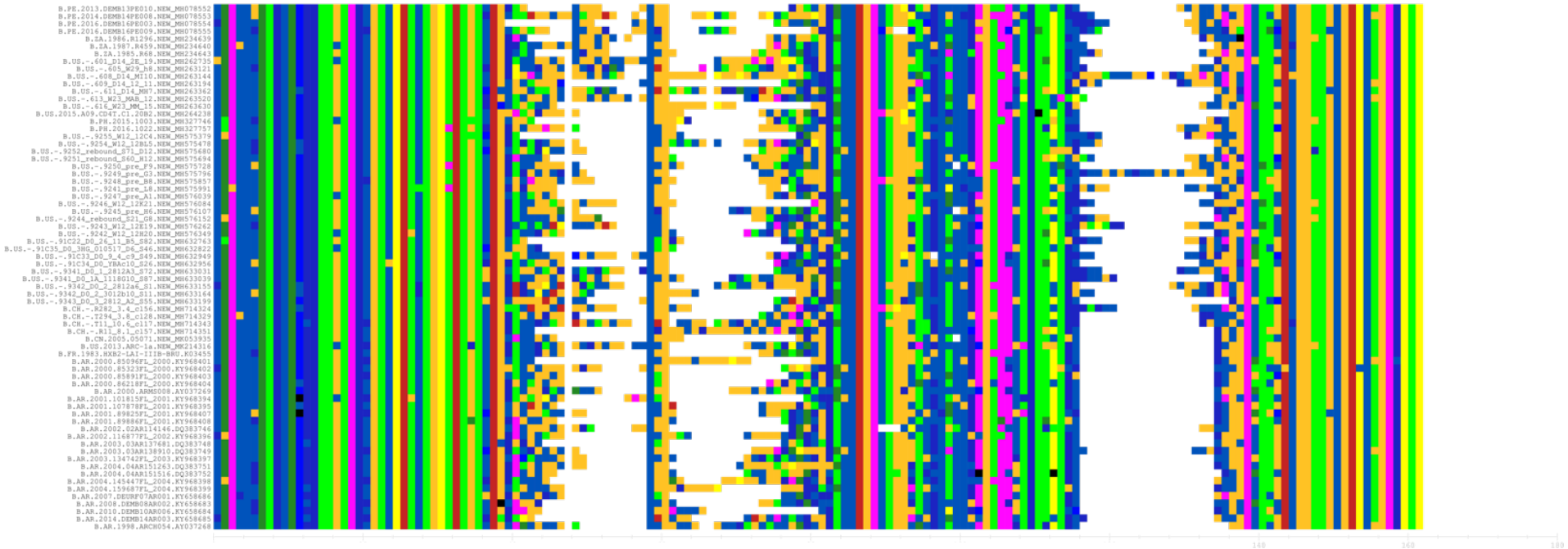
[View large] Download [EPS] [PDF] [PNG] [Legend] [All]

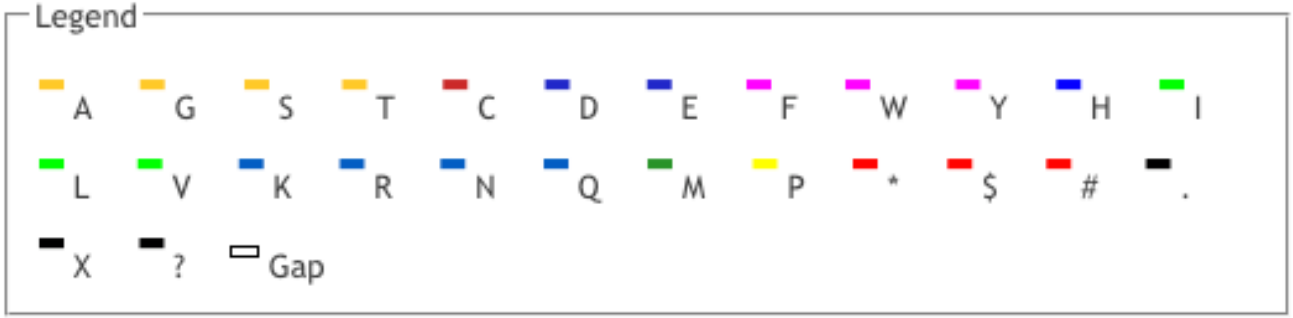


Questions or comments? Contact us at seq-info@lanl.gov.

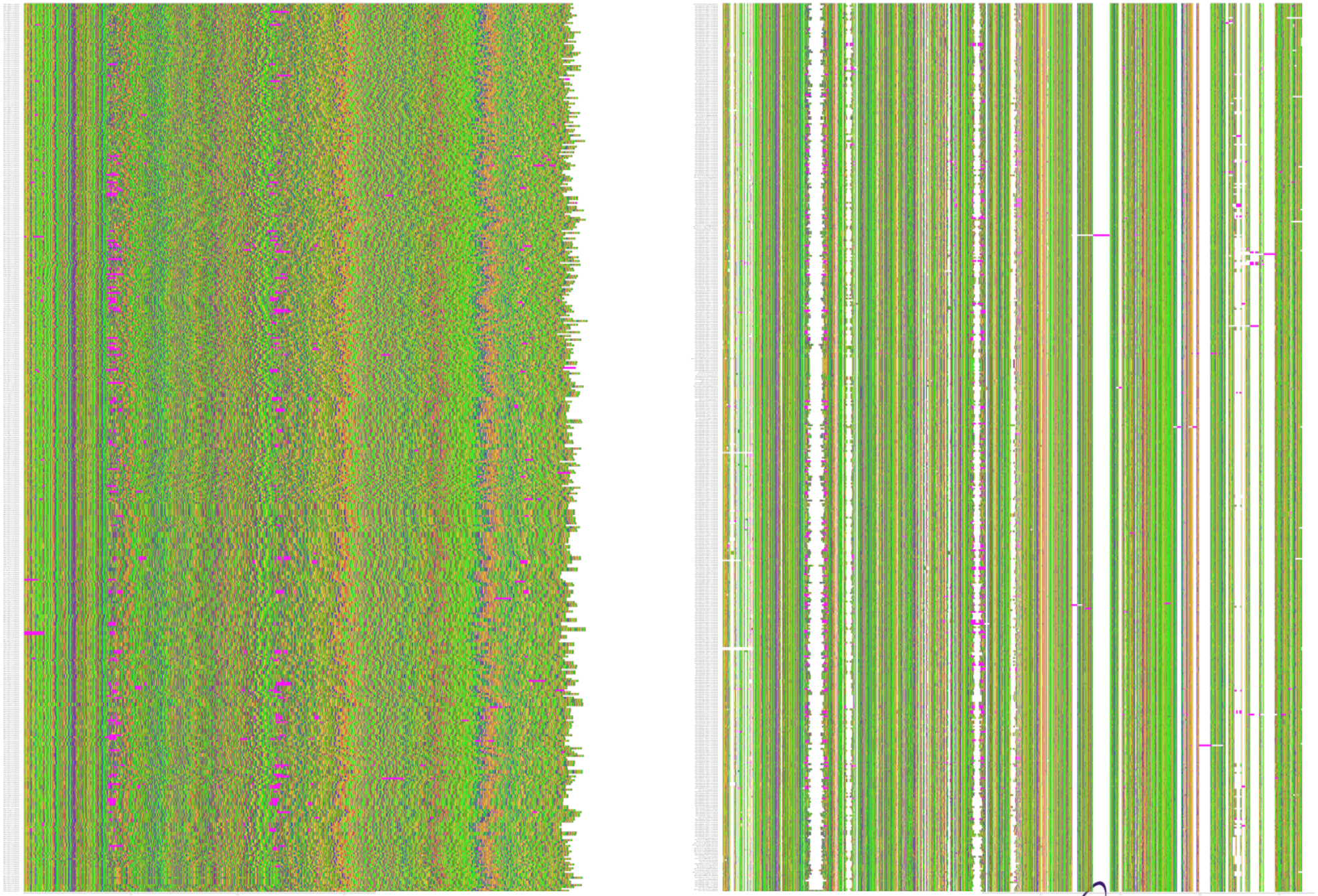
Los Alamos NATIONAL LABORATORY

PIXEL view of HIV-1 Env V1-V2 region amino acid alignment

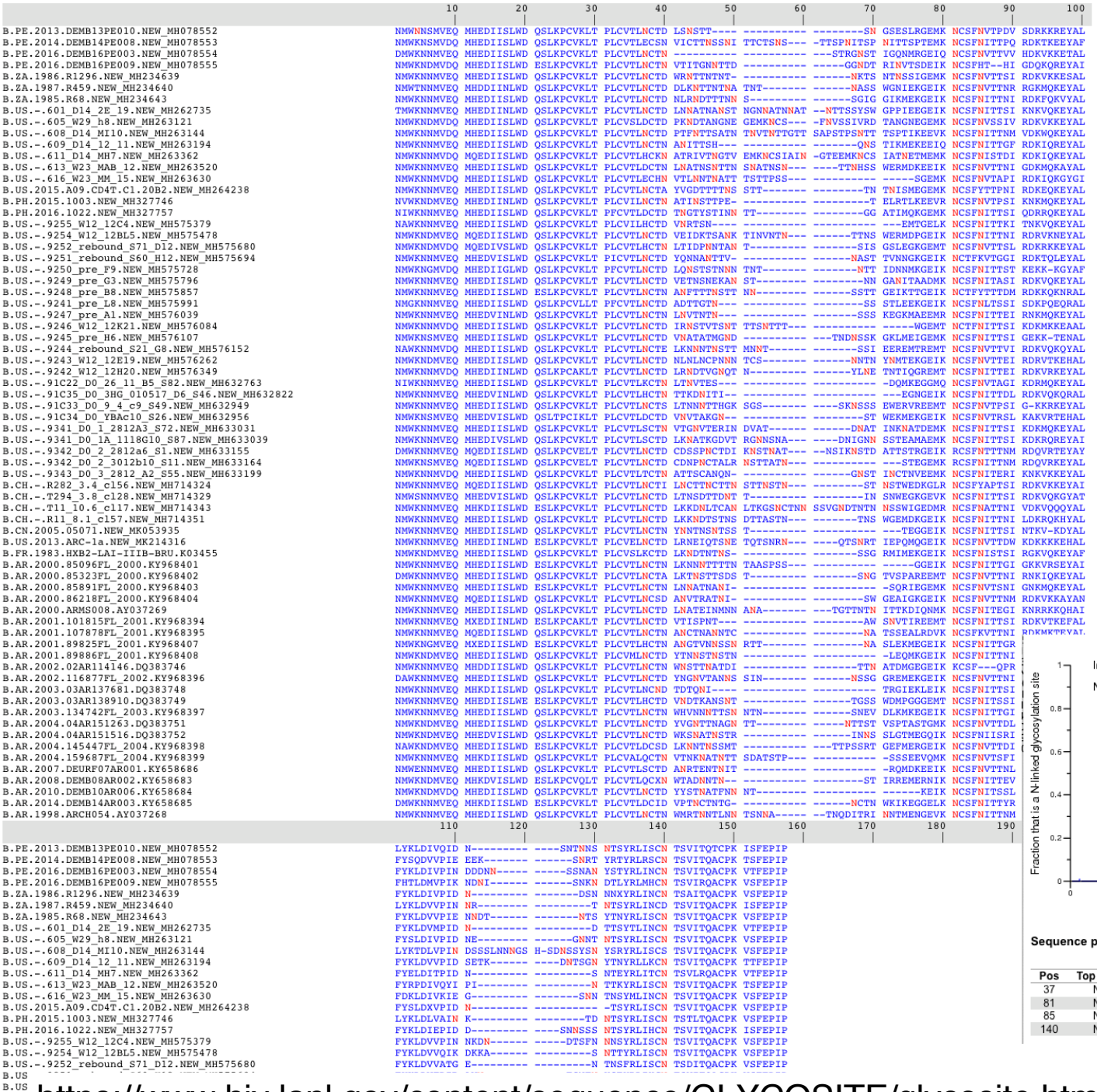




Los Alamos
NATIONAL LABORATORY

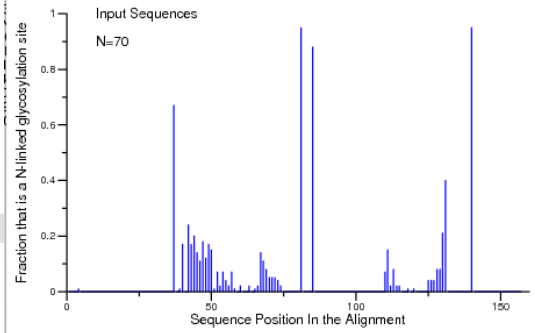


Los Alamos
NATIONAL LABORATORY



Envelope V1-V2 (and other regions) has highly variable N-linked glycosylation sites. The N-glycosite tool highlights N-x-(S/T) sites and computes statistics on them.

"Optimal" alignment may or may not involve adjusting these sites specifically, depending on what the purpose of the alignment is.



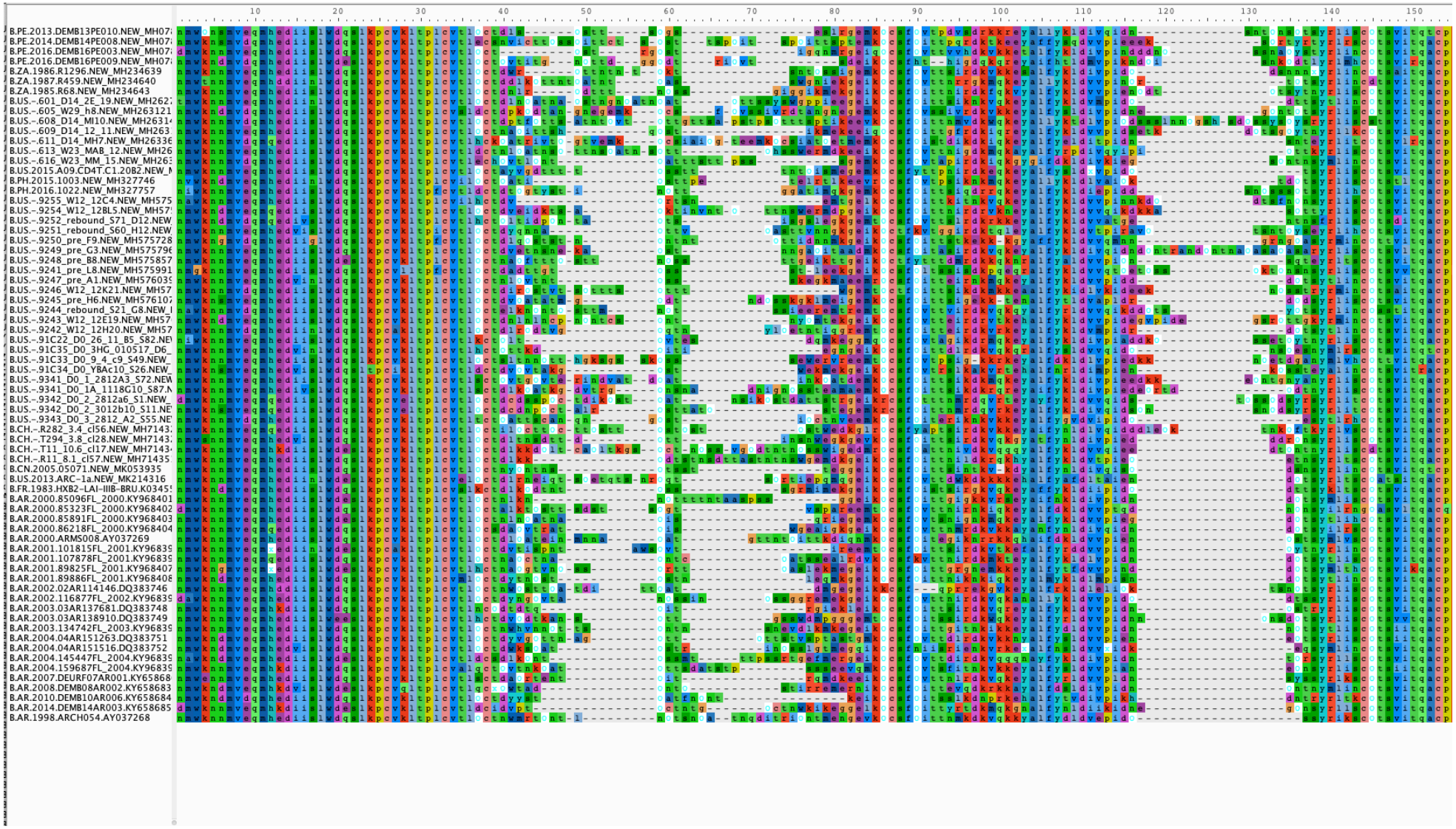
Sequence position(s) of High N-glycosylation Site (>60%)

Pos	Top seq	Num of N-glycosylation	Fraction
37	N	47	0.671
81	N	67	0.957
85	N	62	0.886
140	N	67	0.957

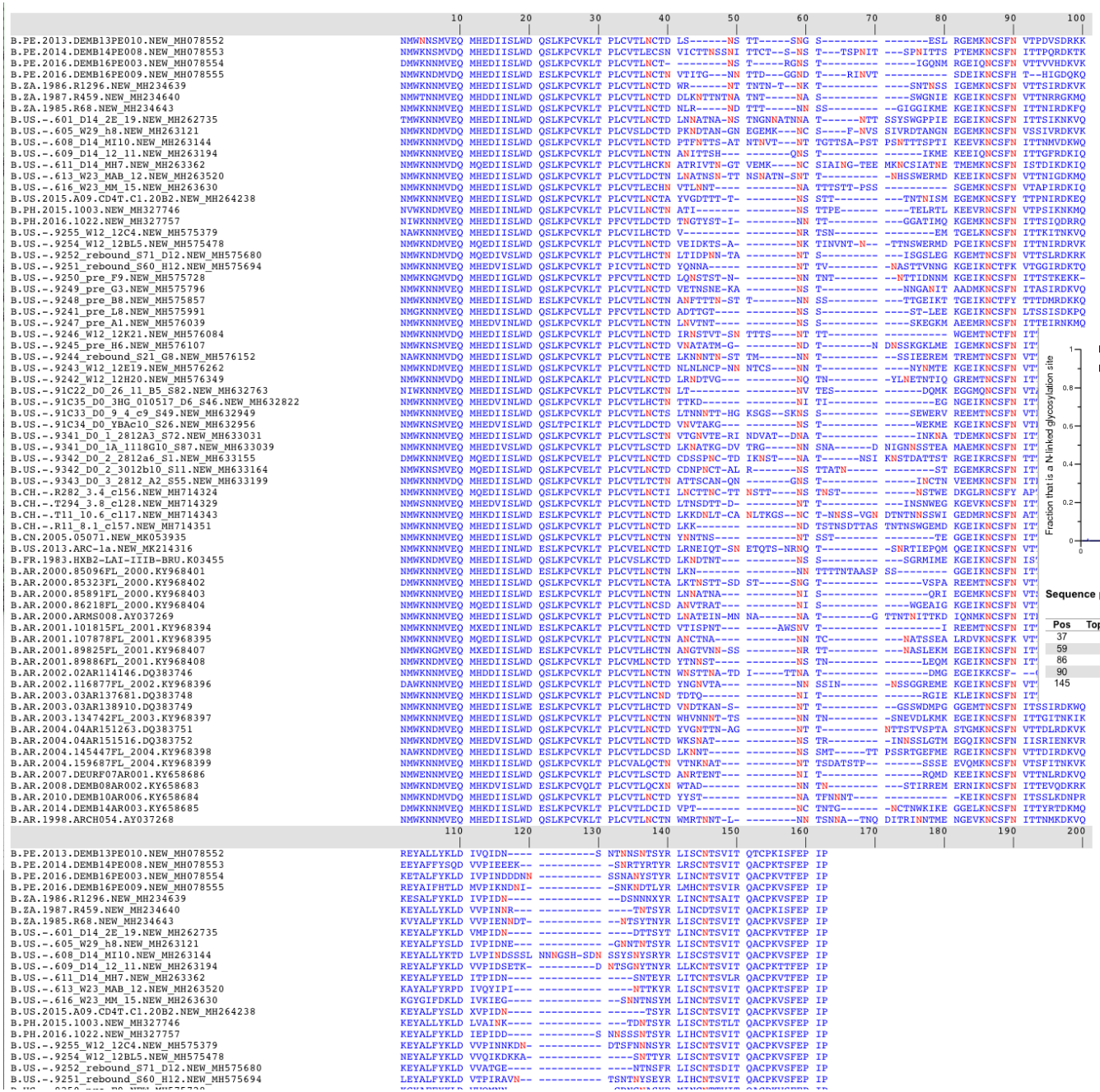
https://www.hiv.lanl.gov/content/sequence/GLYCOSITE/glycosite.html

Los Alamos NATIONAL LABORATORY

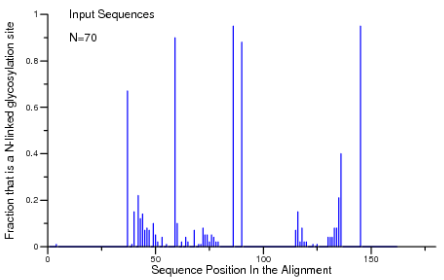
Glycosite can output FASTA file with the N of N-x-(ST) recoded as “O”



Los Alamos NATIONAL LABORATORY



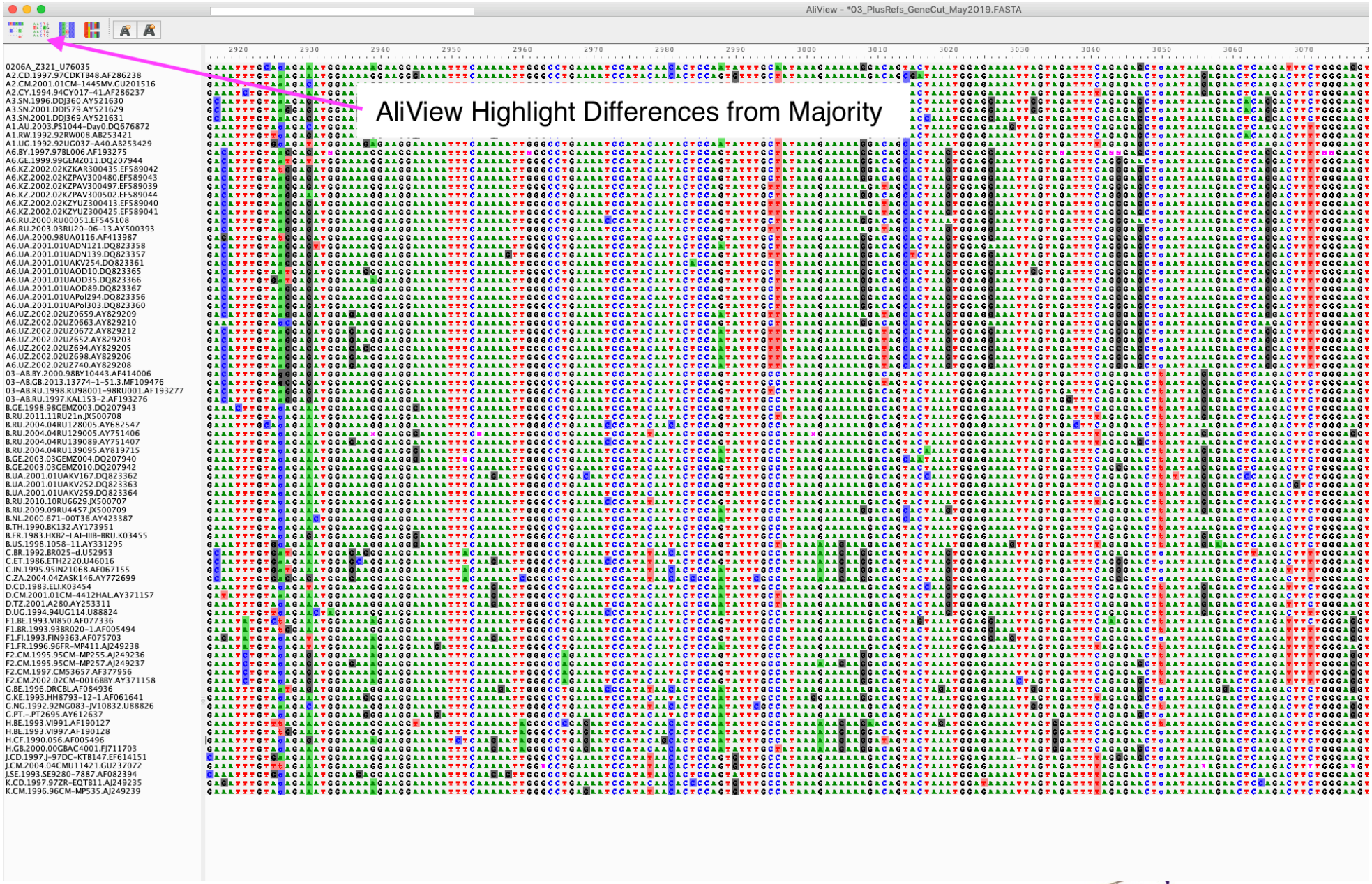
V1 region adjusted to move one N-x-S/T site nearer to center of loop.



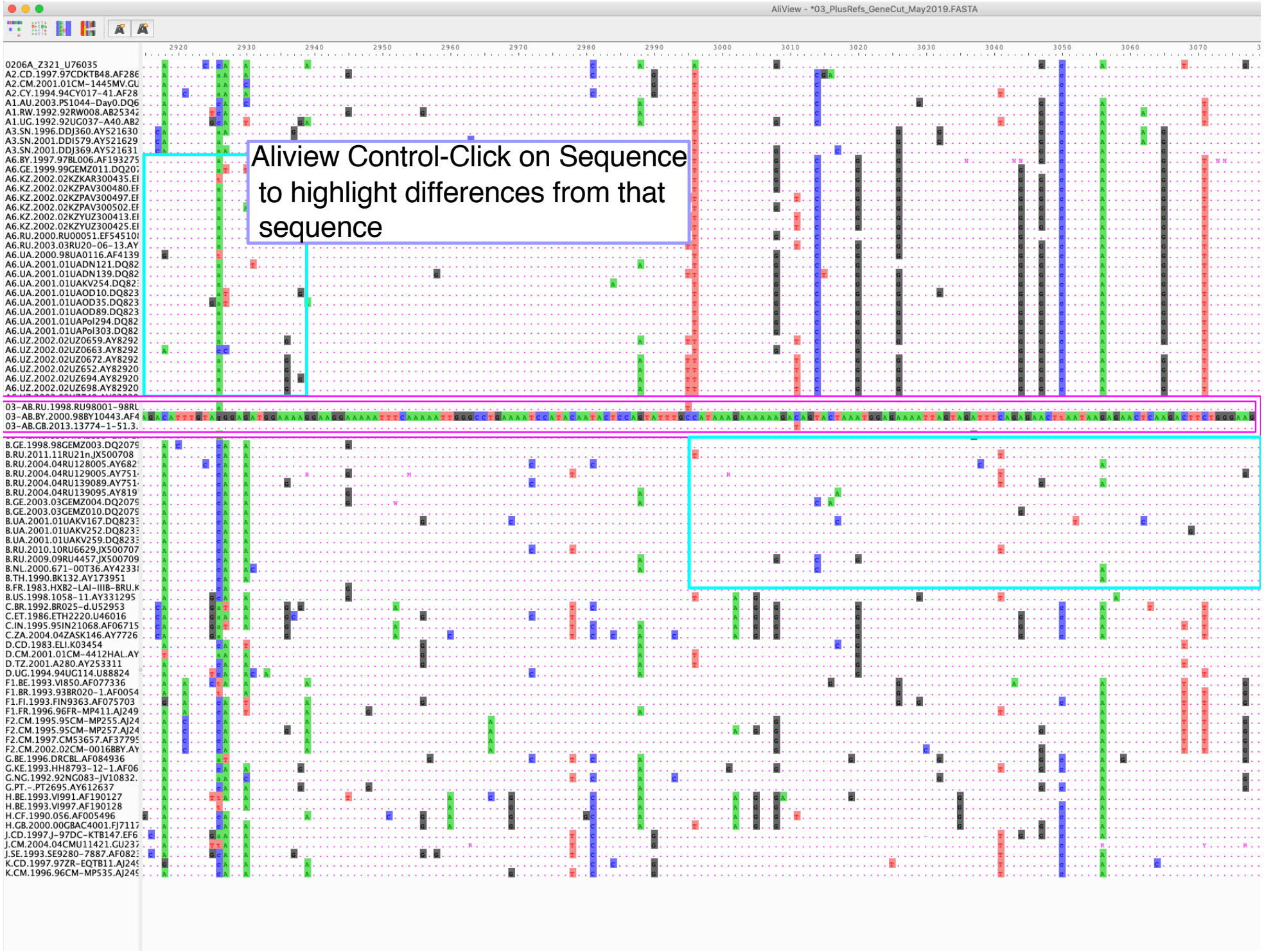
Sequence position(s) of High N-glycosylation Site (>60%)

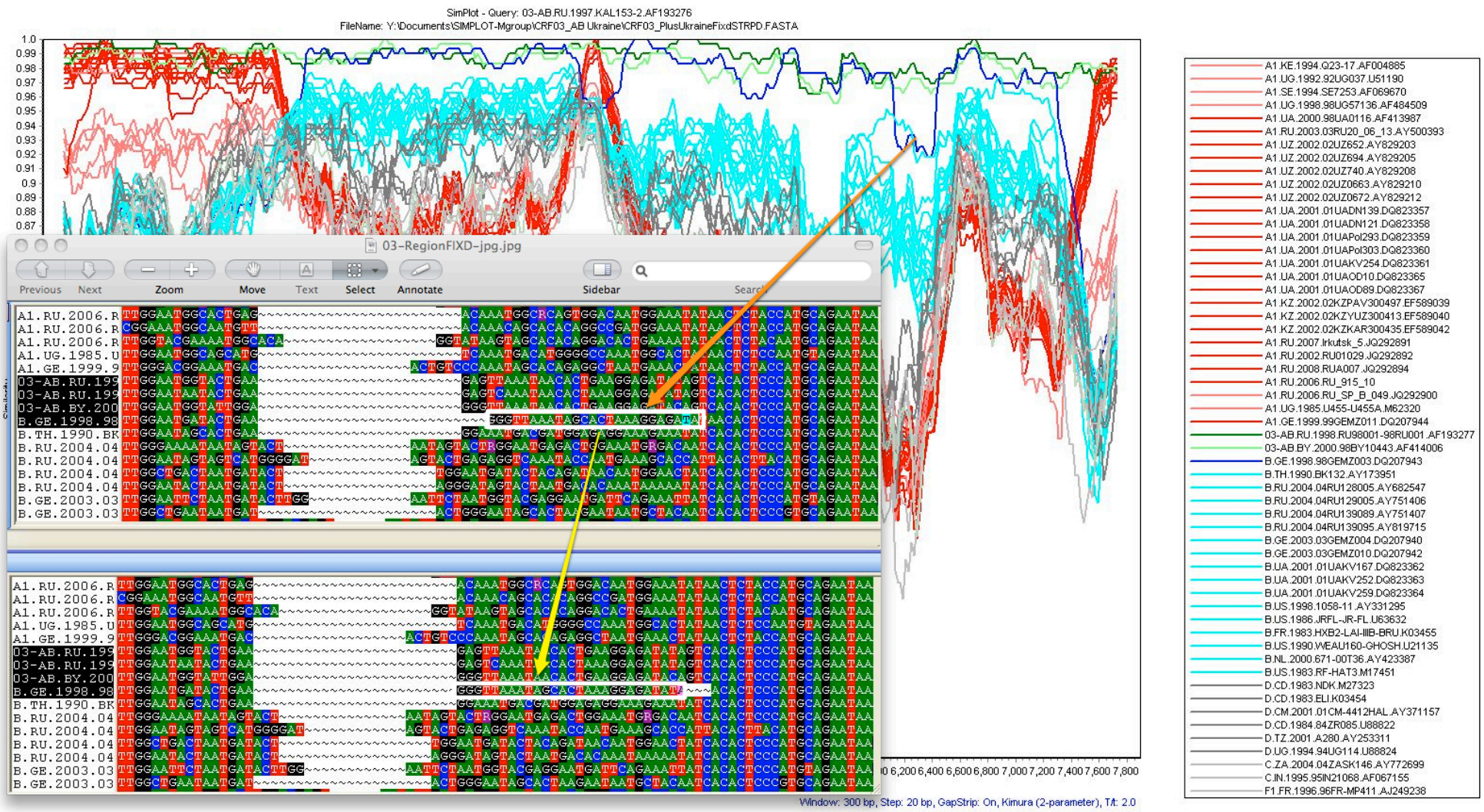
Pos	Top seq	Num of N-glycosylation	Fraction
37	N	47	0.671
59	N	63	0.900
86	N	67	0.957
90	N	62	0.886
145	N	67	0.957

Los Alamos NATIONAL LABORATORY



Los Alamos
NATIONAL LABORATORY

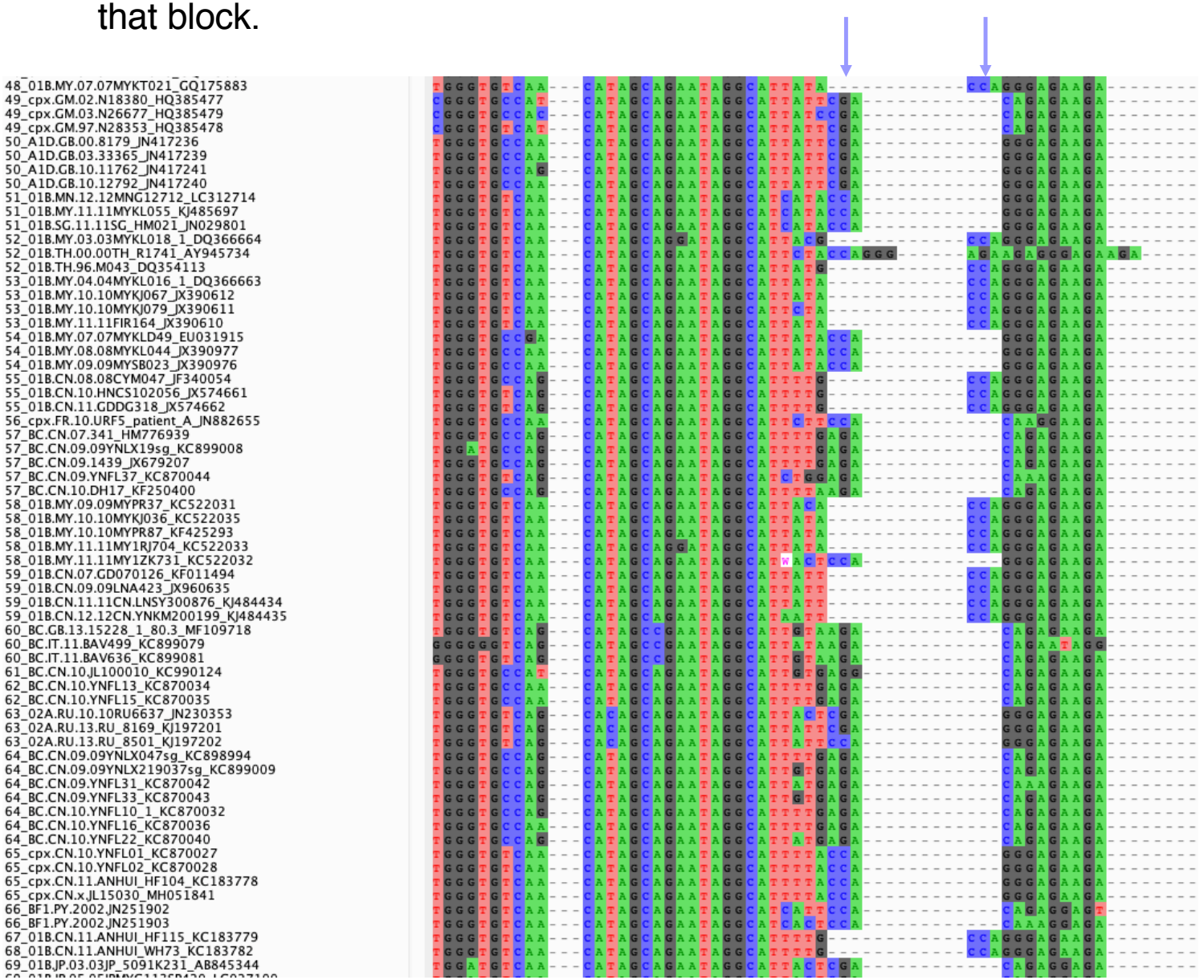

AliView - *03_PlusRefs_GeneCut_May2019.FASTA
Aliview Control-Click on Sequence to highlight differences from that sequence



Even a very small region of misalignment, hypermutation, or poor sequence quality can have a large impact on similarity plots, phylogenetic trees, and other analyses. Similarity plots can be quite useful for identifying sites in a multiple sequence alignment that should be scrutinized, and corrected if in error, as this example shows.
https://sray.med.som.jhmi.edu/SCRoftware/simplot/

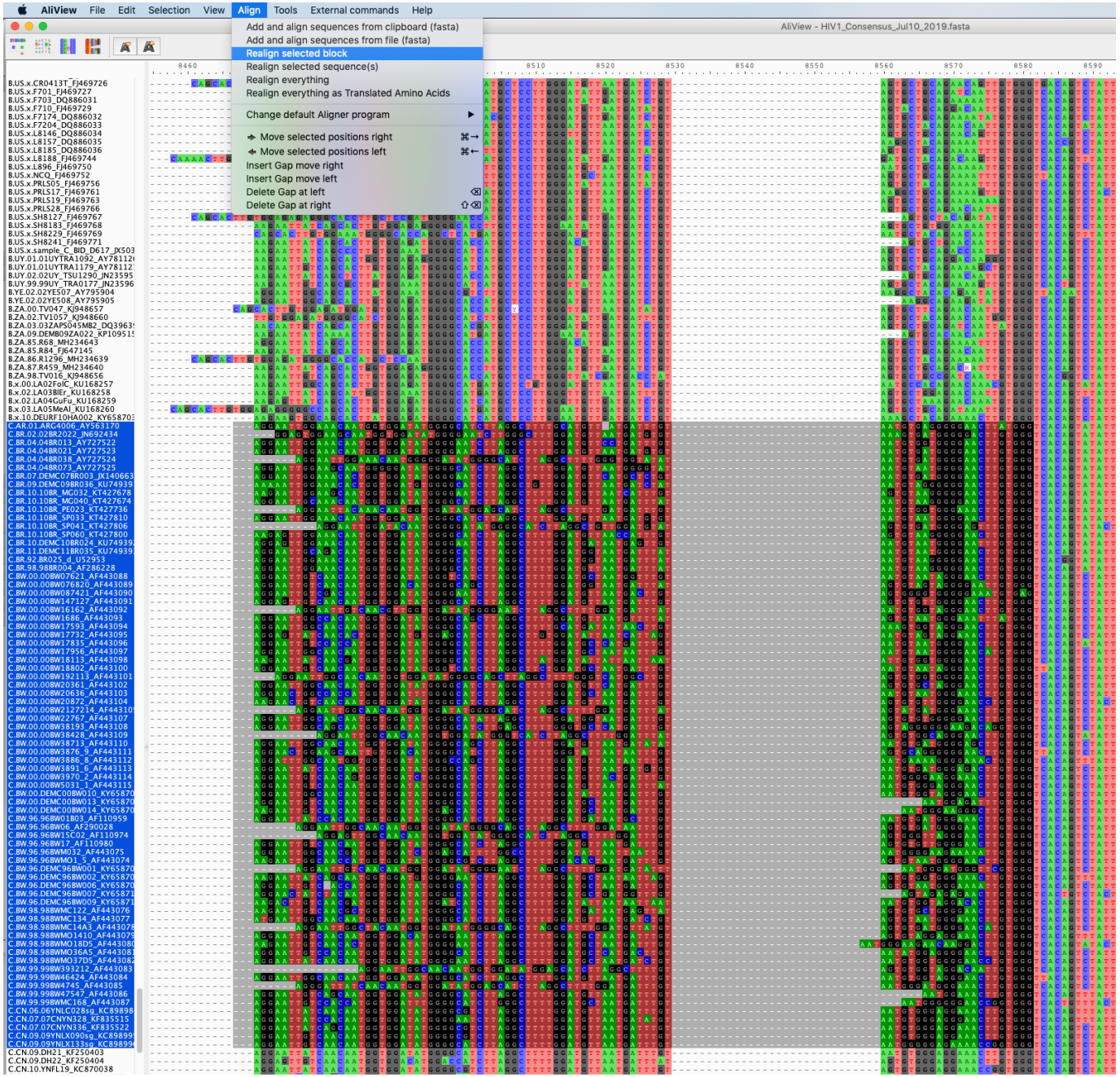
Los Alamos
NATIONAL LABORATORY

One codon being alternatively aligned in many sequences is a very common issue. Easiest to fix by sorting on one of these columns, to get all CCA-right or CCA-left together in one block, then adjusting that block.



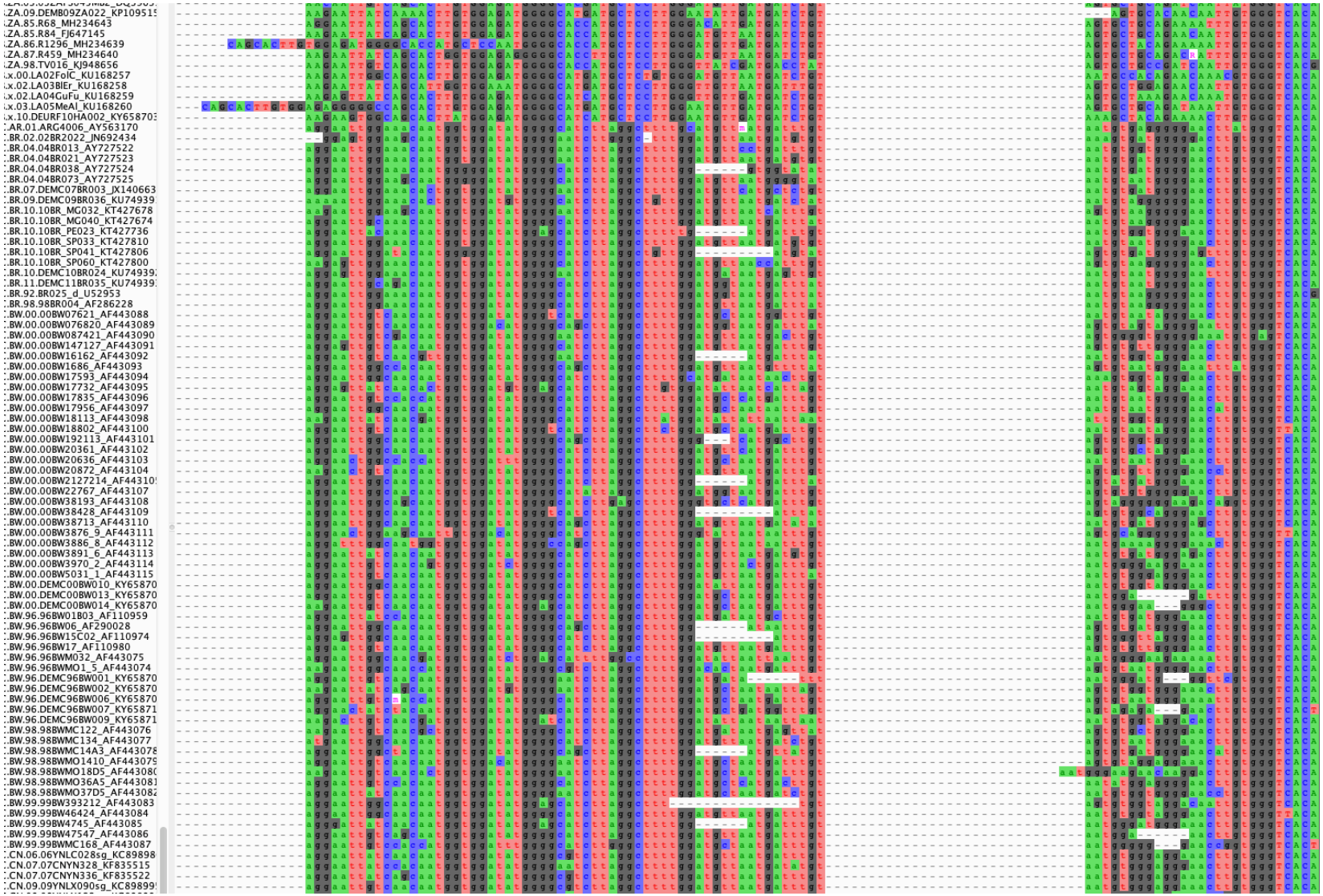
Los Alamos NATIONAL LABORATORY

AliView has built-in alignment tools. Select a block and re-align only that block, is a very useful tool.



Los Alamos NATIONAL LABORATORY

AliView has built-in alignment tools. Select a block and re-align only that block, is a very useful tool.



Note some codons now “broken”; gap in codon rather than in between codons.

Los Alamos
NATIONAL LABORATORY

AliView has built-in alignment tools. Select a block and re-align only that block, is a very useful tool.

Aliview does not currently support re-aligning a selected block as amino acids.

We try to work with software developers to ask for additions of such features, when reasonable.

Los Alamos
NATIONAL LABORATORY

“Perfecting” a large alignment can be a never-ending task. It is important to know when to call it “good enough”.

Also consider methods such as gap-stripping which will often automatically remove regions of uncertain alignment such as the Env V1 and V2 hypervariable loop regions.

Feel free to write to me btf@lanl.gov or seq-info@lanl.gov for advice or help.

Los Alamos
NATIONAL LABORATORY