HIV Database Workshop

[Website: www.hiv.lanl.gov]
[Contact: seq-info@lanl.gov]

Presenters: Will Fischer & Bette Korber

Database PIs: Bette Korber, Thomas Leitner, Karina Yusim

Additional database staff: Werner Abfalterer, Will Fischer, Brian Foley, Peter Hraber, Elisabeth Sharon Fung, Robert Funkhouser, Kumkum Ganguly, Jenni Macke, James Szinger, and Hyejin Yoon

Project Officer: Stuart Shapiro, NIAID, NIH


Theoretical Biology and Biophysics, T-6
Los Alamos National Laboratory
Los Alamos HIV Sequence Database Overview

Will Fischer

Summer School on Quantitative Systems Immunology,
Boston University
June 10-15, 2013

slides available as PDF documents:
http://www.hiv.lanl.gov/content/sequence/HIV/HIVWORKSHOP/index.html
Workshop Topics

HIV Sequence Database and Immunology Database

General introduction
Sequence search interface – alignments and basic trees
Geography search interface
Database Alignments

Tools:
- Genecutter – processing nucleotide sequences
- Treemaker – phylogenetic trees via neighbor-joining
- HIV/SIV sequence locator tool
- Hypermut – detection of APOBEC-mediated hypermutation
- Highlighter – visualization of mutations in related sequences
- Protein Feature Accent
Workshop Goals

- Understanding the database content, how information was obtained, and what is available
- Database searching
- Examples of using tools for analyses
The HIV Databases

- **HIV Sequence database** – founded 1986, G. Myers
  - Relational database, data from GenBank with added fields from the literature
  - Alignments – align indels and reduce multiple sequences per person
  - Annual hard copy and reviews
  - Web search interfaces: subtype, phenotype, geographic, sampling year…
  - Analysis tools

- **HIV Immunology database** – founded 1995, B. Korber
  - Comprehensive HIV epitope database, > 300-400 new papers per year
  - Integrate HIV immunological and sequence data
  - Annual hard copy and reviews
  - Web search interfaces: epitope, protein, HLA type, immunogen, keywords
  - Analysis tools for immunologists

- **HIV Vaccine database** – founded 2003, J. Mokili
  - A searchable relational database of published primate vaccine trials
Database usage over time

HIV database usage 2008-2011

2012 statistics

Hits: ~12 million  downloads: 537 Gb  visits: 364,000  visitors: 165,000

http://www.hiv.lanl.gov/awstats/awstats.pl
Histogram output

This histogram shows the distribution of sequences from your query across the entire HIV-1 genome. At each position across the genome, the number of sequences overlapping with that position is plotted. The colors represent different subtypes.
Primate Lentiviruses

Alignments: http://www.hiv.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html

Van Heuverswyn, Nature 2006
Keele, Science 2006
Corbet, J. Virol 2000
Foley, HIV database

Positive Chimps
HIV-1 M, N, O

P. t. troglodytes
P. t. schweinfurthii

Los Alamos National Laboratory
Entry page at http://www.hiv.lanl.gov/

The HIV databases contain data on HIV genetic sequences, immunological epitopes, drug resistance-associated mutations, and vaccine trials. The website also gives access to a large number of tools that can be used to analyze these data. This project is funded by the Division of AIDS of the National Institute of Allergy and Infectious Diseases (NIAID), a part of the National Institutes of Health (NIH). Click on any of the links below to access a database. Editorial Board

SEQUENCE DATABASE ➤ VACCINE DATABASE ➤
IMMUNOLOGY DATABASE ➤ OTHER VIRUSES ➤

News

New Features for Epitope Location Finder (ELF)
ELF displays known and predicted epitopes found within a protein sequence query. ELF results now include both Class I (CTL) and Class II (helper) epitopes. In addition to predicting epitopes based on anchor residues, ELF now includes predictions from the Class I and Class II Binding Predictions tools at the Immune Epitope Database (IEDB). 13 March 2012

New Features for HIV BLAST
HIV BLAST has new features. It now allows the user to find best matches among only subtyped sequences, or sequences of a specific subtype. It allows the resulting sequences to be downloaded fully aligned. 01 March 2012

New Option for N-GlycoSite
The N-GlycoSite tool predicts N-linked glycosylation sites in amino acid sequences. A new option allows the user to exclude sites with a second-position proline, which is disfavored for N-linked glycosylation. 29 February 2012

HIV Antibody Search Results More Specific
The antibody search interface in the HIV Immunology database is now more specific. Searches from the Author, Keyword, and Note fields now display only those notes and references that relate directly to the search. The user may still opt to display all, if desired 09 February 2012

New Options for Quickalign
The Quickalign tool aligns any short protein or nucleotide sequence with database sequences. New options provide additional ways to calculate and display frequency by position, and allow the user to include the surrounding region in the alignment. 08 February 2012

Questions or comments? Contact us at hivinfo@lanl.gov
HIV sequence database

Programs and Tools
- Search Interface retrieves HIV and SIV sequences, which can then be aligned and used to build trees
- Geography Search Interface retrieves HIV sequences based on geographical distribution
- Tools for working with sequences lists all our online tools, organized by function

Alignments
- HIV Primate Alignments includes Consensus and Ancestral Sequences, Subtype Reference Alignments, and Complete Alignments

Information
- HIV Sequence Compendium print or order our annual publication
- Tutorials and other information unpublished web-based content
- Links to other HIV/AIDS tools and information

About this website
- FAQ general Information about this website
- Site Statistics usage information for www.hiv.lanl.gov

How to Cite this Database
- Editorial Board

News:

New Features for HIV BLAST
HIV BLAST has new features. It now allows the user to find best matches among only subtyped sequences, or sequences of a specific subtype. It allows the resulting sequences to be downloaded fully aligned. 01 March 2012

New Option for N-GlycoSite
The N-GlycoSite tool predicts N-linked glycosylation sites in amino acid sequences. A new option allows the user to exclude sites with a second-position proline, which is disfavored for N-linked glycosylation. 29 February 2012

HIV Antibody Search Results More Specific
The antibody search interface in the HIV Immunology database is now more specific. Searches from the Author, Keyword, and Note fields now display only those notes and references that relate directly to the search. The user may still opt to display all, if desired. 09 February 2012

New Options for Quickalign
The Quickalign tool aligns any short protein or nucleotide sequence with database sequences. New options provide additional ways to calculate and display frequency by position, and allow the user to include the surrounding region in the alignment. 08 February 2012

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

Copyright © 2001-2016 LANL All rights reserved | Disclaimer/Privacy
Search Interface

- **Help**
  - Tips at the top of the page are often overlooked
    - Ranges, operators, wildcards, logical groupings
  - Mouse-over provides brief descriptions; click field names for details in Help file

- **Searches**
  - Searches are case-insensitive
  - Records are searchable through sequence, patient, genomic region, or publication information and can be matched to the genomic region of a user-provided alignment
  - First seven fields will appear in search results page by default
  - A “*” in a textbox will cause that field to be included in the results page
  - Patient information (Infection year, Infection country) is different than sequence information (Sampling year and Sampling country)
  - Problematic sequence filters (hypermutation, frequent ambiguities, potential contamination)

- **Analysis**
  - Build a tree with user alignment, search results and subtype reference sequences combined

- **Results**
  - Can download aligned or unaligned sequences
  - Alignments are based on multiple pairwise alignments – alignments are good, but need hand editing for an optimal alignment
  - Select all or a subset of sequences for download
  - Sequences can be re-ordered by clicking on fields at the top of the page
We will search for country = Brazil (BR)

We will search for complete genomes.
Results for HIV-1 complete genomes from Brazil

Choose “One sequence/patient” to remove very similar sequences (only available if a region is selected).
Select a few sequences and make tree, allows us to add a reference set to our data and align them.
Choice of outgroup influences the tree. In general, choose next closest sequences to the “ingroup”. In this case our Brazilian sequences are all HIV-1 M group.

These settings minimally influence relative branch lengths, but rarely alter the tree topology.

Optional mailback, and tree title
ATV java-based view for quick look, cannot save/print

Obtaining your sequences of interest and having them aligned to a good reference set was the whole point of this. The tree was just a first check on data and alignment quality.

Save alignment, run GeneCutter or use BioEdit or SeAl to view/adjust.
Save alignment, use BioEdit or SeAl to view/adjust.

Send alignment to GeneCutter or HIV-Align first, is usually best.

http://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html

Brazil Genomes Plus Subtype Reference Set, as downloaded
New search: all complete genomes; then look at geographic and subtype distribution of the sequences
Each continent’s pie chart is clickable to “zoom in” on that continent.

Likewise for each country once you are zoomed in to the continent level.

Most complete genomes in the HIV database are subtype B. But subtype C is more prevalent in human infections. Beware of this type of sampling bias.
Click through pie to get (e.g.,) all sequences from Brazil.
Tools

- **Analysis and Quality Control**
  - HIV BLAST finds sequences similar to yours in the HIV database.
  - N-Glycosite finds potential N-linked glycosylation sites.
  - RIP 3.0 (Recombinant Identification Program) detects HIV-1 subtypes and recombination.

- **Alignment and sequence manipulation**
  - HIValign uses our HMM alignment models to align your sequences.
  - Gapstreeze removes columns with more than a given % of gaps.
  - ElimDupes Given an alignment or set of unaligned nucleotide or protein sequences, this tool compares the sequences and eliminates any duplicates.

- **Phylogenetics**
  - TreeMaker generates a neighbor-joining phylogenetic tree.
  - PhyML generates a maximum likelihood phylogenetic tree.
  - TreeRate finds the phylogenetic root of a tree and calculates evolutionary rate.

- **Format and display**
  - Protein Feature Accent provides an interactive 3-D graphic of HIV proteins; the user can map a sequence feature (a short functional domain, epitope, or amino acid) and see where it occurs spatially in the 3D structure.
  - Highlighter highlights mismatches, matches, transition and transversion mutations, and silent and non-silent mutations in an alignment of nucleotide sequences.
  - SeqPublish makes alignment publication-ready.
  - Recombinant HIV drawing tool highlights regions of the genome on a graphically representation.
The HIV database sequence analysis tool set

![Image of the tool set](image_url)

**Click top level to link to full page of tools**

<table>
<thead>
<tr>
<th>DATABASES</th>
<th>SEARCH</th>
<th>ALIGNMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tools</td>
<td>Publications</td>
<td>Guides</td>
</tr>
<tr>
<td>Index of all tools</td>
<td>HIV-BLAST</td>
<td>Quality Control</td>
</tr>
<tr>
<td>ADRA</td>
<td>HIVAlign</td>
<td>Consensus Align</td>
</tr>
<tr>
<td>Branchlength</td>
<td>Hypermut</td>
<td>Rainbow Tree</td>
</tr>
<tr>
<td>Codon Alignment</td>
<td>JpHIV at GOBICS</td>
<td>Recombinant HIV-1 Drawing Tool</td>
</tr>
<tr>
<td>Consensus Maker</td>
<td>Mosaic Vaccine Tool Suite</td>
<td>RIP</td>
</tr>
<tr>
<td>ELF</td>
<td>Motif Scan</td>
<td>Sequblish</td>
</tr>
<tr>
<td>ElimDupes</td>
<td>N-Glycosite</td>
<td>Sequence Locator</td>
</tr>
<tr>
<td>Entropy</td>
<td>PCORD</td>
<td>SNAP</td>
</tr>
<tr>
<td>FindModel</td>
<td>PepMap</td>
<td>SUDI Subtyping</td>
</tr>
<tr>
<td>Format Converter</td>
<td>PeptGen</td>
<td>SynchAlign</td>
</tr>
<tr>
<td>Gap Strip/Squeeze</td>
<td>PhyloPlace</td>
<td>Translate</td>
</tr>
<tr>
<td>GenBank Entry Generation</td>
<td>PhyML</td>
<td>TreeMaker</td>
</tr>
<tr>
<td>Gene Cutter</td>
<td>PhyloMAC</td>
<td>TreeRate</td>
</tr>
<tr>
<td>Heatmap</td>
<td>Poisson-Fitter</td>
<td>VESPA</td>
</tr>
<tr>
<td>Hpitopos</td>
<td>Protein Feature Accent</td>
<td>External Tools</td>
</tr>
<tr>
<td>Highlighter</td>
<td>Protein Structure</td>
<td></td>
</tr>
</tbody>
</table>

**News:**

Sequence Locator Improved output for multiple queries

For input of multiple sequences, Sequence Locator now provides links to download the summary information as tab-delimited text files of coordinates. 05 December 2012

Last modified: Tue Jan 26 10:10 2010
Tools are organized in groups by function/purpose.

Most tools have explanation pages, and sample data sets.

Many tools were inspired by user comments, please ask for more.
We tend to list only tools of great use in HIV research. Many of these tools are essential, such as either BioEdit or SeAl for alignment viewing and correction.

http://www.hiv.lanl.gov/content/sequence/HIV/HIVTools.html
Pre-Built Sequence alignments

- Originally based on iterations of manual and HMM alignments
- Yearly updates using HMM and manual corrections
- Alignments are in reading frame (codon aligned)
- Contain non-redundant data (one sequence per patient)
- Compendium alignments show fewer sequences than web version
- Reference alignments contain up to four representatives of each subtype. One of each CRF.
- Protein alignments may contain frameshift compensations
- Subtype consensus with ties resolved, as well as maximum likelihood ancestors, are available for reagent production
- Special interest alignments are being added
  - Sequence sets of particular research interest
  - Suggestions welcome to tkl@lanl.gov
All (complete) = one per patient, all sequences for which we have a complete genome.

Subtype Reference = 4 representatives of each subtype, plus one of each circulating intersubtype recombinant form (CRF) of the M group, plus 4 O group, N group, P group and SIV-CPZ

Consensus/Ancestral computed from master alignment periodically.

HIV-2/SIV-SMM and primate lentivirus alignments also available here.
Gene Cutter

- Unconventional Alignment/Homology program specific for HIV/SIV
- "Cuts out" specified genes and proteins from sets of DNA sequences
  - Aligns to HXB2 via HMMer (or to SIV-Mac239 for HIV-2 and SIV-SMM)
  - Splits input sequences into genes, if desired
  - Aligns DNA sequences by codon, and translates them with interpretation of IUPAC ambiguity codes (e.g. R/Y for purine/pyrimidine)
- Useful for processing new sequence data
  - annotating full length genomes
  - pulling out regions of interest from raw sequence data
- For each gene/region, maintains a list of anomalies
  - stop codons
  - codons containing multi-state characters
  - codons containing indels
- Input sequences may be aligned or unaligned
- Results may be better if the HXB2 sequence is included as a reference in your input file
GeneCutter Result
Alignments viewed with Pixel
http://www.hiv.lanl.gov/content/sequence/pixel/pixel.html

Our data aligned to reference set by search tool:
(output of search and tree build was input to GeneCutter)

Can also be viewed with BioEdit, Se-Al or other multiple sequence alignment editors.
Treemake

Check for phylogenetic relatives:
- TreeMaker produces a Neighbor Joining tree for a quick comparison
- TreeMaker uses PAUP* for its calculations; a few model options are available
- Reference sequences can be included, and are aligned to the input automatically
- Trees are displayed using PHYLIP and ATV
- The alignment used for the tree can also be downloaded
- A Phyml interface is also available

http://www.hiv.lanl.gov/content/sequence/PHYML/interface.html
Paste or type a DNA alignment here.

OR upload an alignment file here.

http://www.hiv.lanl.gov/components/sequence/HIV/treemaker/treemaker.html
http://tree.bio.ed.ac.uk/software/figtree/

- hypermutated
- contaminant
- recombinant
HIV/SIV Sequence Locator Tool

- Instantly computes position numbers of DNA or protein fragments relative to a reference strain (HXB2r for HIV-1, SMM239 for SIV)
  - Such numbers, often included in the literature, are frequently incorrect
- Shows the location of the sequence on an HIV map
- Presents protein translations of DNA sequences
- Can be used for input into the search interface, to align a new sequence you have generated with the database set
- Can also retrieve reference sequences
  - by coordinates (range of base or amino-acid positions)
  - by single position (retrieves flanking sequences)
HIV Sequence Locator Tool

Purpose: This tool has several purposes. It can find the start and end coordinates (relative to the reference strain HXB2) of your input sequence(s) and show which genes or proteins it covers, along with a graphical view of the location of your sequence(s) relative to the reference sequence. The tool will display both the nucleotide sequence and protein translation of your input as it aligns to HXB2. It will also check the reverse complement of your input sequence, and report the orientation with the best match. Another use is to retrieve a section of the HXB2 reference sequence based on its coordinates.

How to use: To find the coordinates for your sequence, either upload or paste your sequence (any format) in the box below, or (for database sequences only) enter GenBank accession numbers. To retrieve the HXB2 sequence for a set of coordinates (see HIV coordinate map), enter the coordinates and choose the region. To retrieve the entire gene or protein, enter coordinate values of "1" and "end". To retrieve a single nucleotide or range with its surrounding 42-nucleotide sequence, enter the single coordinate in the "from" field and check the box. For more details, see Sequence Locator Explanation.

Useful Links:
- HXB2 numbering
- SIVmm239 numbering (review articles)
- HXB2 spreadsheet
- SIVmm239 spreadsheet (spreadsheets with base-by-base annotation)

Find the location of a sequence

- Sequence type: Let program decide, HIV, SIV
- Paste your input here
- or upload your file

Retrieve a region by its coordinates

- Enter coordinates: from to
- Region: Complete
- Retrieve: Nucleotide or protein output, Include surrounding region

Paste or type a DNA or protein sequence here.

OR enter numeric coordinates here.
### Sequence Locator:

**Table of genomic regions touched by query sequence. Query protein translation in blue.**

<table>
<thead>
<tr>
<th>CDS</th>
<th>NA position relative to CDS start in HXB2</th>
<th>NA position relative to query sequence start</th>
<th>NA position relative to HXB2 genome start</th>
<th>AA position relative to protein start in HXB2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gag</td>
<td>352 -&gt; 483</td>
<td>1 -&gt; 132</td>
<td>1141 -&gt; 1272</td>
<td>118 -&gt; 161</td>
</tr>
<tr>
<td></td>
<td><strong>AAADTGHSNQVSQNPIVQNIQGQVMHQAI.SPRTLNAWVKVVEE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p17</td>
<td>352 -&gt; 396</td>
<td>1 -&gt; 45</td>
<td>1141 -&gt; 1185</td>
<td>118 -&gt; 132</td>
</tr>
<tr>
<td></td>
<td><strong>AAADTGHSNQVSQNY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p24</td>
<td>1 -&gt; 87</td>
<td>46 -&gt; 132</td>
<td>1186 -&gt; 1272</td>
<td>1 -&gt; 29</td>
</tr>
<tr>
<td></td>
<td><strong>PIVQNIQGQVMHQAI.SPRTLNAWVKVVEE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sequence below includes up to 42 bases of context surrounding query sequence.**

<table>
<thead>
<tr>
<th>Reference Strain</th>
<th>Type</th>
<th>Region</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>HXB2</td>
<td>nuc</td>
<td>complete</td>
<td>1141</td>
<td>1272</td>
</tr>
</tbody>
</table>

**Retrieved Sequence:**

AAGCTGACACAGGACACAGCAATCGGTTCAGCCAAAATATTACCCTATAGTGCAAGACATCCAGGGCAATGGTACA
TCAGGCCCATATCACCCTAGAACCTTTAATATGCTTGGTAAAGTAGTAGAAGAG

**Organism:** HIV
Hypermutation

- Detects APOBEC related A->G hypermutation as default
- Can be adapted to detect any fuzzy motif in relation to a control pattern
Cumulative mutation graph is useful.

High ratio of G -> A vs. A -> G indicates hypermutation.
Highlighter

- Highlights mutations relative to a reference strain, particularly useful for intra-patient analyses.
- Highlights:
  - syn/non-syn
  - transition/transversion
  - Apobec motifs
- Sorts on similarity
- Visualize recombination of closely related sequences
Nonrandom distribution of mutations evident.

Sample Set is from a possible dual Infection, with intra-subtype recombinants evident.
Protein Feature Accent

- Highlights region of interest in an HIV structure
- You can upload a PDB structure, or use one of our annotated Env structures
- You can upload your own alignment and get an entropy map
http://www.hiv.lanl.gov/content/sequence/PROTVIS/html/protvis.html

List of “recommended” PDB entries

Only a gp120 alignment is provided so far. We hope to add others. You can paste in your own.
Selected region gets highlighted in structure

Many display options in Jmol are “built in” to this web tool. Use the Jmol command script box below for other commands.

One of the color schemes is “color by entropy” based on diversity in the alignment added below.

Selected region gets highlighted in structure
Please let us know if you have questions, comments or suggestions

seq-info@lanl.gov

Bette Korber: btk@lanl.gov
Will Fischer: wfischer@lanl.gov