Phylogenetic studies on the origin of HIV-1

Elizabeth Villa Rodriguez
University of Illinois
Physics Department
PHYS498BIN: Statistical Physics Applied to
Biological Information and Complexity

In 1999, Edward Hooper proposed that the human inmunodeficiency virus type 1 (HIV-1) could have originated in the bathes of live oral polio vaccine (OPV). He claimed that this vaccine was produced in chimpanzee kidney epithelial cell cultures rather than in monkey kidney cell cultures, as stated by the producers of OPV at the Winstar Institute in Philadelphia. The vaccine was applied to about one million people in Congo in the late fifties. This is claimed to have enabled the transfer to humans of chimpanzee simian immunodeficiency virus, widely accepted to be the origin of HIV-1.
On the other hand, phylogenetic analysis of HIV-1 indicates that the virus originated before the campaign giving rise to a theory suggesting natural transfer between human and chimpanzee.
The HIV virus has been somewhat complicated in its evolution. The tree of the M-group, the mail group of HIV, has a special structure where one can find simultaneous appearance of viral subtypes. This is interpreted by the authors as many events of transfer or "multiple viral lineages" from chimpanzees to human.
A group from Oxford led by Andrew Rambaut analyzed 197 HIV-1 sequences sampled in 1997 in Congo, a likely location for the origin of HIV-1 group M for both hypothesis[1].
The group used the ClustalW program to align the predicted Congo amino acid sequences, plus 223 sequences of HIV-1 from all over they world. fixing it by hand in necessary places. The resulting alignment was used as a guide for the placement of indels (insertions/deletions) in the nucleotide sequences. The result of the phylogenetic tree shows that the Congo stains have a high

genetic diversity, like the global strains, and many Cono lineages were found basal to the origin of each subtype.
In this work, they compare if the structure of the phylogeny differed from that of HIV-1 M by comparing the subtype diversity ratio (SDR) of the two. (Figure 1) the SDR is the ratio of the mean within-subtype pairwise distance to the mean between-subtype pairwise distance. For calculating the SDR, they we used a heuristic algorithm to assign subtypes such that the subtype diversity ratio was minimized.
The Congo and global phylogenies differ significantly in the SDR statistic, with the former showing no more subtype structure than phylogenetic trees simulated under a model of exponential population.
In figure 1 we can see how subtypes can be clearly identified in the distribution of pairwise distances for the global sequences, and distinction between intra- and intersubtype comparisons for the Congo sequences is less clear. This means that for two chosen Congo sequences, it is difficult to determine unambiguously whether they belong to the same or to different subtypes. The authors even say that the Congo and global phylogenies probably result from different epidemiological histories. Many Congo strains appear to be basal, and they think it means that each global subtype is the result of the chance exportation of some Congo strains to other geographical regions, thus producing an apparent starburst in the phylogeny, and says it also suggests that previous phylogenetic analysis has underestimated the number of lineages that pre-date 1957-60, and hence underestimated the minimum number of cross-species transmissions necessary to reconcile the OPV hypothesis with phylogenetic data.
The results of this group are very exciting. The accusation against the Wismar Institute was very serious, and people even talked about the contamination of the vaccine by the virus was deliberately covered up. The authors make then a clear point in assuring that the oldest common ancestor of HIV-1 in froup M which gave rise to the pandemic strains, was prestent in a human host long before the first OPV field trails were conducted in the Congo during the late 1950s using phylogenetic tools: crystal clear results.
If this was not enough, other groups published papers in the same issue of Nature [2, 3] and in Science [4] . The later paper reported on DNA chimpanzee search in the OPV pools, and none was found, but monkey DNA was indeed found. The other two papers also deal with tests on early OPV stocks prepared at the Wistar Institute in the 1950s, and also show that they were propagated in the kidney cells of rhesus monkesys, and that they lack

nucleic-acid sequences related to either HIV or chimpanzees.
The studies above, phylogenetic and molecular, represent an important result that let us eliminate the most accepted hyphotesis about HIV origin until today. It shows how bioinformatic tools can be applied to several different areas of interest in research today.

References

[1] A. Rambaut, D. L Robertson, O. G. Pybus, M. Peeters, and E. C. Holmes. Phylogeny and the origin of hiv-1. *Nature*, 410:1047–1048, April 2001.

[2] P. Blancou et al. '. *Nature*, 410:1046–1047, 2001.

[3] N. Berry et al. '. *Nature*, 410(1046-1047), 2001.

[4] Poinar H, Kuch M, and Paabo S. Molecular analysis of oral polio vaccine samplesj. *Science*, 292:743–744, April 2001.

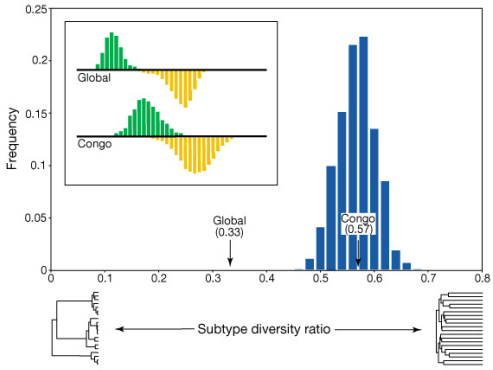

Figure 1: Maximum-likelihood phylogenies were estimated for three V3-V5 data sets: HIV-1 sequences from the Democratic Republic of Congo (423 base pairs), global isolates (426 base pairs), and Congo and global isolates combined (396 base pairs). Given a phylogeny with tips labelled according to subtype, the subtype diversity ratio (SDR) was calculated as the mean path length between tips of the same subtype divided by the mean path length between tips of different subtypes. For the phylogeny of the global isolates, 11 subtypes were allocated according to standard HIV-1 nomenclature7. For the Congo phylogeny, 11 subtypes were allocated so as to minimize the SDR score, using a heuristic optimization algorithm. This assignment is that which gives the maximum possible subtype structure for the Congo phylogeny. The global phylogeny gave an SDR of 0.33 and the Congo a value of 0.57. The analysis was repeated after removal of the Congo and global sequences previously identified as intersubtype recombinants4, 5. The analysis will only be affected if recombination breakpoints fall within the V3-V5 region, so excluding recombinants changes the SDR only marginally (0.35 for the global phylogeny; 0.58 for the Congo). SDR values were similar when Congo isolates were assigned to different numbers of subtypes (for example, 0.59 and 0.55 in the case of 8 and 14 subtypes, respectively). To assess the significance of the difference between the global and Congo SDRs, a null distribution was obtained by simulating phylogenies under an exponential growth coalescent process inferred from env gene sequences of subtype A (ref. 6), which is common in Africa. The frequency distribution of minimum SDR values for these simulated phylogenies is shown in blue. Inset: normalized frequency distributions of intrasubtype path lengths (above the line) and intersubtype path lengths (below the line), plotted on the same horizontal scale (0.0-0.8 substitutions per site), for the global and Congo phylogenies.