The phylogenetic information profile of HIV-1 and the degradation effect of recombination: How to wipe out the Mitos of Ariadne

G.Magiorkinis[1], D.Paraskevis[1], H.Schmidt[2], A.Hatzakis[1]

1. Medical School, National and Kapodistrian University of Athens, Greece
2. von-Neumann-Institut fuer Computing, Forschungsgruppe Bioinformatik, Germany

Corresponding Author: A. Hatzakis, MD, PhD

Professor of Epidemiology and Preventive Medicine

National Retrovirus Reference Center

Department of Hygiene and Epidemiology

Athens University Medical School

Mikras Asias 75

11527, Athens

Greece

Tel.: +302107462090

Fax: +302107462190

e-mail: ahatzak@med.uoa.gr

Introduction

Several arguments have been compiled to second the suggestion that genetic recombination has a pivotal role in the evolutionary pathways. Firstly, recombination is known to be the landslide in the more complex species' proliferation (e.g in the animals as sexual reproduction) (Maynard Smith 1978; de Visser and Elena 2007). Thus, the presence of recombination in the upper levels of the evolution is indisputable and supposed to testify an evolutionary advantage. Apart from this intuitive evolutionary benefit, recombination is confirmed to take place in a great variety of species from viruses up to humans, meaning that it has a universal effect in the diversification of life. Finally, it is known that it provides the recombinant subjects several evolutionary properties, such as the "escape from Muller's ratchet" and "evolutionary broad jumping", both accelerating and/or accommodating the expansion of the progeny (Burke 1997; Chao 1997).

Apart from the evolutionary aspect of recombination, it has been suggested that it has a devastating effect in the reconstruction of the species' phylogeny. Traditionally, phylogenetic trees have been implemented to describe the evolutionary relationships of the species by connecting the taxa with "lines" (Semple and Steel 2004). Strictly geometrically speaking, the phylogenetic trees are one-dimensional linear plots meaning that the line segments do not produce closed shapes forming outlined surfaces. The basic assumption underlying this aspect of evolution is that each strain has a single progeny and the obvious effect is that these graphics cannot be used for describing the phylogeny of recombinants: during recombination each subject has more than one direct ancestor. On the other hand networks have been suggested to overcome this single-progeny assumption (Huson 1998; Huson and Bryant 2006). Networks in contrast to trees may produce closed shapes permitting

branch interconnection and providing the framework for alternative histories: through networks one can have more than one direct ancestor.

Nevertheless, during recombination each progeny usually contributes to the descendant's genome by providing large genomic segments rather than small genomic fragments or single nucleotides. Based on this observation, the assumption of the single progeny may be implemented on separate segments[1] coming from distinct progenitors as long as they can be somehow distinguished. As a logical consequence, an approach that is widely used especially in the phylogenetic analyses of HIV, is the breakdown of the genome to potentially non recombinant segments guided by an "exploratory" analysis (e.g. bootscanning plot) and followed by phylogenetic analysis based on conventional trees (Salminen et al. 1995). The term exploratory methodology of recombination is used to describe a set of graphical methods based on the sliding window concept (similarity plot, bootscanning plot, bayesian scanning plot) (Lole et al. 1999; Paraskevis et al. 2005) which can be used at the beginning of the recombination analysis to suggest the possible recombination pattern.

HIV-1 is known to recombine frequently, since it is documented that at least 20% of the circulating strains are intersubtype recombinants (Peeters and Sharp 2000; Quinones-Mateu et al. 2002) and is one of the most sequenced organisms (http://hiv-web.lanl.gov), providing thus a very good subject to test the theoretical hypotheses about recombination. It was to our intention to analyze the distribution of the phylogenetic information along the HIV-1 genome and confirm the theoretical effect of recombination onto conventional phylogenetic tree reconstruction regarding the phylogenetic informative content by using a well characterized set of recombinant HIV-1 strains (Magiorkinis et al. 2003).

[1] We consider as genomic *segment* a set of consecutive nucleotides that is a subset of the genome, whereas as *area* the genomic segment sized 400 nt.

Materials and Methods

Sequence Data and Alignments

For methodological reasons, explained in the following sections, we formed three distinct nucleotide sequence datasets: the recombinant dataset, the subtype reference dataset and the baseline information dataset.

We composed the recombinant dataset with 33 previously analyzed full-length HIV recombinant sequences (Magiorkinis et al. 2005) (Table 1). We composed the subtype reference dataset with the following full-length sequences (isolate numbers): subtype A (U455, 92UG037), B (LAI, RF), C (C2220, 92BR025), D (ELI, NDK), F (FIN9363, 93BR020, 95CMMP255), G (92NG083, HH8793), H (VI991, 90CF056), J (SE7022, SE7887), K (96CM-MP535C). We composed the baseline information dataset with sequences (accession numbers): subtype A (AF361872), B (AF049495), C (AF110964), D (AF484487), F (AF077336), G (AY772535), H (AF005496), J (AF082394), K (AJ249235).

For each separate dataset we constructed multiple sequence alignments by means of the Clustal W program (Thompson, Higgins, and Gibson 1994) which we subsequently manually corrected.

Puzzle scanning plots

We build puzzle scanning plots in a similar way to the bootscanning plot (Salminen et al. 1995) according to the exploratory methodology: a window of given size (called puzzle scanning window, w_{ps}) is slid stepwise along the sequence alignment and phylogenetic analysis is performed by means of the Tree-Puzzle program (Schmidt et al. 2002). For each reconstructed tree we store and plot the following values: 1) the puzzling support values supporting the closer relationship among the query sequence and one of the rest of the defined strains/groups, 2) the number of the fully resolved and partially resolved quartets

We build the puzzle scanning plots using the following parameters:

1) w_{ps} size: 400 nt

2) step size: 50 nt

3) evolutionary model: Tamura Nei (Tamura and Nei 1993), 4 discrete categories as an approximation of a gamma distributed substitution rate to model the rate heterogeneity among sites (Yang 1994)

We chose the w_{ps} to be sized 400 nt because this is empirically defined to contain enough phylogenetic signal across the HIV-1 genome (Magiorkinis et al, 2003). We choose the step to be sized 50 nt since it seems to provide a fair trade between the analytical detail and computational intensiveness. For example, if we use steps sized 1 nt the analysis needs 50 times more computations, while the gain in the analytical detail is practically nilpotent.

Hardware – Software Implementation

We performed the analysis on an 8-node Linux cluster implementing unix-shell scripting along with the linux parallel versions of the previously mentioned software components (Thompson, Higgins, and Gibson 1994; Schmidt et al. 2002). We developed a simple program (parser) for extracting the previously mentioned values from the Tree-Puzzle output files using macros in the Windows platform.

Sequence similarity across the genome

In order to calculate the sequence similarity across the HIV-1 genome we implement another exploratory algorithm: a window of given size is slid stepwise along the sequence alignment and sequence distances are calculated by means of the DNADIST program of the PHYILIP package implementing the Jukes and Cantor nucleotide substitution model (Jukes and Cantor 1969; Felsenstein 2005). For each sliding window we calculate the average

intersubtype *p* distance for the subtype reference dataset. We calculate the *p*-distance (proportion of nucleotide sites at which two sequences being compared are different) by transformation of the Jukes-Cantor distances (Jukes and Cantor 1969) and the similarity (~~sim~~) as the complementary of *p* up to1 ($\sqrt{sim}=1-p$).

Information Baseline

In order to quantify the effect of recombination as an additive or reductive difference of the phylogenetic informative content in the sequences of the progenitors, we suggest the calculation of a non-recombinant standard of the phylogenetic information (HIV-1 non-recombinant information baseline, H_{ib}). For this we chose a set of non-recombinant HIV-1 strains (one strain for each subtype, distinct from those used in the reference dataset) and for each ~~one~~ we performed the same analysis as for the recombinant strains. Subsequently we averaged the informative content of their sequences for each w_{ps} and set it as a reasonable point estimator of H_{ib}.

The choice of using an additional non-recombinant sequence in order to estimate the H_{ib} instead of using merely the subtype reference dataset is justified in the need ~~of~~ comparing phylogenetic constructions that contain the same number of taxa (19 sequences for the recombinant and non-recombinant calculations). We support this necessity by the following consideration: The statistical power (Cohen 1988) of the estimations-tests is known to be related to the number of the observations and to the number of the inferred parameters in ~~such~~ a deterministic way that we can never infer more parameters than the number of the observations of the dataset. On the other hand, the phylogenetic tree is known to be a complex estimator composed of a stack of distinct estimations such as the branch lengths and nodes. Consequently, the statistical power of the phylogenetic inference of the tree is related to the number of the branches-nodes and the length of the sequence alignment (more specifically the

number of informative sites). Since the trees constructed during the puzzle scanning of the recombinant and the non-recombinant baseline contained the same number of taxa (19) and had the same size of w_{ps} (400), phylogenetic estimations should have potentially and theoretically the same statistical power and, consequently, the informative content will not be degraded due to inference of additional parameters.

Definition of Variables

In order to elucidate the variables used in our analysis a terminology is defined as follows:

Recombinant area: The *true recombination breakpoint* is the point which separates genomic segments having different progenitors. Since we cannot determine the *true recombination breakpoint* we estimate it as the point which separates genomic segments supporting different direct progenitors through phylogenetic analysis. This support should be suggested by the exploratory analysis and confirmed by subsequent reconstruction of the phylogenetic tree. However, since we cannot be deterministic on the exact location of the *true recombination breakpoint*, we define in this analysis as *recombination area* the region sized 400 nt in the middle of which an estimated *recombination breakpoint* has been inferred. This area has the property that the probability to contain the *true recombination breakpoint,* given that recombination is a fact, empirically approximates 1. Consequently, the w_{ps} sized 400 nt are divided into *recombinant areas* and *non-recombinant areas*.

q_u (unresolved quartets): We define the number of not fully resolved quartets (q_u) as the sum of partially unresolved (q_{pu}) and fully unresolved quartets (q_{fu})

$$q_u = q_{pu} + q_{fu}$$

p_u (proportion of unresolved quartets): We define the proportion of unresolved quartets (p_u) to be equal to q_u divided by the overall number of the ~~attempted~~ EXAMINED quartets (n_q)

$$p_u = \frac{q_u}{n_q}$$

d_u (difference of p_u ~~in~~ between the recombinants and non recombinants): We define the ~~difference~~ d_u to be ~~equal to the~~ THE PROPORTION OF p_u ~~in~~ OF a specific genomic segment of the recombinant (p_{ur}) minus the p_u in the regarding genomic segment of the averaged non-recombinant baseline (p_{un})

WHAT IS SEGMENT? WPS?

$$d_u = p_{ur} - p_{un}$$

HOW DO YOU ~~DEFINE~~ COMPUTE P_{ur} ~~(P_{und})~~ FOR THE RECOMBINANT? OR DO YOU MEAN REC. BASELINE?

s_u (standardized d_u): We define the standardized d_u (s) to be equal to the d_u divided be the p_{un}

$$s_u = \frac{d_u}{p_{un}} = \frac{p_{ur} - p_{un}}{p_{un}}$$

HOW DO YOU COMPARE P_{ur} AND P_{un} EXACTLY?

Statistics

CORRELATION BETWEEN WHAT?

We implement parametric statistics for variables approximating the normal distribution: ~~t-test~~ for hypothesis testing and the Pearson's coefficient of determination ~~(R squared)~~ R^2 to assess correlation. We implement non-parametric statistics for variables departing from the normal distribution: Spearman's rank correlations coefficient to assess correlation (*rho*).

STILL REF- MISSING (3RD TIME)

THE READER WAS NO CLUE HOW, WHY THOSE AND WHAT FOR IF 0

GIVE A SENTENCED THAT/WHY IT IS EXPECTED THAT $P_{ur} \geq P_{un}$ (OTHER WISE $d_u \leq 0$), TO HELP THE READER. ACTUALLY I EXPECT THE OPOSITE!

Information measures - Observations

Information is supposed to be the measure which quantifies the update on our prior beliefs as soon as an observation has taken place (Kullback 1997). In phylogeny the universal prior belief is that all species have a common ancestor. This ~~may be~~ WE simply ~~represented~~ IF by a star-like tree among the species (Figure 1). A fully resolved tree is ~~the~~ HOWEVER one that contains only bifurcations. Consequently, phylogenetic information may be considered as the data measure

BUT WHY?

ie INNER NODE HAVE EXACTLY 3 ADJACENT BRANCHES,

HOW CAN A "DATA MEASURE" UPDATE PRIOR BELIEVES

updating our common ancestor prior beliefs towards a fully resolved tree. A way to quantify this amount of information is the percentage of fully resolved quartets during the quartet-puzzling algorithmic reconstruction (Strimmer and von Haeseler 1996). Thus, we firstly define standardized measures for the phylogenetic information based on unresolved, partially and fully resolved quartets as ~~previously~~ described. Consequently, the measurement used in the analysis below is the standardized information *s* which is inversely related to the phylogenetic information: a positive value of *s* means that more unresolved quartets exist for the query strain than expected from the baseline set meaning that less phylogenetic information is contained in this alignment fragment.

Our main goal was to compare the phylogenetic information contained in the recombinant regions in contrast to the non-recombinant. However, definition of the observation of information is not intuitive and care must be taken in order to define a variable that can be used comparatively. Firstly, for each strain the genome ~~was~~ divided into segments called after *recombinant areas* (as previously defined) and ~~into~~ segments defined by the rest of the genome (Figure 2); *s* as previously defined was chosen to describe the phylogenetic informative content of them.

The following possible options were examined in order to define a reliable and comparable statistic for both the recombination area and the remaining genome:

Consideration 1: The statistical observation is the phylogenetic information contained in each one of the w_{ps}, while the null hypothesis to be checked is that the phylogenetic information in-between w_{ps} defined as recombinant areas and w_{ps} defined as non-recombinant areas is equal.

This consideration leads to test the distribution of all the possible w_{ps}; however, the statistical observations under this consideration are not independent since these windows are partially overlapping. We examined the strength of linear correlation among adjacent

overlapping windows and the autocorrelation coefficient (Pearson, R-squared) for adjacent windows (lag=1) was found to be 0.58 ($P<0.001$).

Consideration 2: The statistical observation is the phylogenetic information contained in each genomic segment and the genomic segments are divided into recombinant areas and into the non-recombinant segments (Figure 2). The null hypothesis is that the phylogenetic information contained into the recombinant areas and the non-recombinant segments is equal.

We considered two different approaches for calculating the phylogenetic information in theses genomic segments:

a) Information (s) can be calculated by conducting phylogenetic analysis on the alignment fragments defined as recombinant areas and the non-recombinant segments. The genomic segments defined as recombinant areas have fixed size of 400 nt, while the remaining (non-recombinant) genomic segments have size determined by the limits of the recombinant areas and may vary from 400 up to more than 5000 nt. Consequently, we compare the phylogenetic information from genomic segments having unequal size thus introducing bias into the analysis: larger regions (data) tend to have smaller s because they contain more sites and consequently the inferences have more statistical power than in small regions.

b) Information for each genomic segment can be calculated as the average s of the w_{ps} having their midpoints inside the limits of the corresponding genomic segments. Thereby the measurements are comparable (w_{ps} have potentially the same statistical power since they have the same length) in-between the genomic segments which have varying size and the recombination areas which have the fixed sized of 400 nt. Still there is a theoretical disturbance of the dependency of the observations since the w_{ps} of the remaining genomic segments and the recombination areas partially overlap. We calculated the strength of this correlation and we found that the previously described autocorrelation is now attenuated,

WHAT IS THAT — NOT UNDERSTANDABLE W/O DEF AND REF.

since the Pearson's autocorrelation coefficient for (lag=1) is reduced from 0.58 ($P<0.001$) for the s of the overlapping w_{ps} down to 0.07 ($P=0.13$) for the averaged s of the adjacent genomic segments making the observations virtually independent.

As a logical consequence we use as statistical observation of the information is the average s as described in Consideration 2b WITH the null hypothesis ~~to be checked if~~ that the phylogenetic information contained BETWEEN the recombinant areas and the non-recombinant segments is equal

Implementation

Each full-length sequence was aligned against the profile of the reference dataset and subsequently a puzzle-scanning plot was built. The numbers of resolved, partially resolved and fully resolved quartets for each w_{ps} were collected to form a database. Coordinates of recombination breakpoints from 34 HIV-1 recombinant strains were used from a previous analysis (Magiorkinis et al. 2005), according to which 235 recombinant areas were defined overall, while the remaining genomic segments were 251. The average s (Figure 2) for both recombinant areas and remaining segments, is approximately normally distributed. Hence, the mean of the averaged s is a good estimator of the phylogenetic information for the genomic segments.

DB NOT DESCRIBED YET - HOW SHOULD THE READER KNOW. SOUNDS LIKE THERE IS A DB ON THE WEB.

A LOT OF THE ABOVE SOUNDS HANDWAVING:
→ TOO MANY "THEORETICALLY", "POTENTIALLY" "COULD/SHOULD", "VIRTUALLY", "MIGHT" ~~EAB~~, "COULD COUNT"

YOU SHOULD EXPLAIN EXPLICITLY ~~WHAT~~ HOW POSITIVE AND NEGATIVE S ~~IS~~ HAS TO BE INTERPRETED. IF S IN INFORMATION, WHY AND WHAT IS NEGATIVE INFORMATION?

Results

As shown in Figure 3, the distributions of the *s* in both the recombinant areas and the rest of the genome are strongly shifted towards positive values ~~meaning~~ THIS MEANS that the phylogenetic information is reduced with regard to the H_{ib}. This might have been caused because the recombination events are temporally old in such a way that the recombinants have diverged from the parental strains and significantly evolved towards saturation of the phylogenetic signal, which is now recorded as attenuation of the phylogenetic information.

In Figure 4 we show how the average $\overline{q_i}$ calculated for the non recombinant set fluctuates across the HIV-1 genome as well as the *sm*; the plots are overlaid ~~to graphically check the possibility of~~ TO VISUALIZE co-fluctuation. The overlaid plots ~~are~~ suggestive of a moderate correlation that prompted for testing ~~measurably~~ the strength and significance of this correlation. The Spearman's rank correlation for non-overlapping respective windows of intersubtype *s* and *sm* suggested a moderate though significant correlation (*rho*=0.66, P<0.001) (Figure 5).

In order to test the null hypothesis as proposed by Consideration 2b, we performed a standard t-test to compare the means of the averaged *s* between the recombinant areas and the non-recombinant segments. The mean averaged *s* is higher in the recombinant areas (0.051 vs 0.047), however it was not found to be significant (*P*=0.15, two-sided, unequal variances). Failure of significance may be debited to the perturbation of the assumption underlying the standard t-test which requires that the statistical observations need to be drawn from a single normal distribution. However, each recombinant strain has pursued different evolutionary pathways and the phylogenetic information has been disrupted in different proportions for each strain meaning that *s* from different strains might not be comparable. In terms of statistics this means that *s* is not drawn from a single normal distribution, but from several distinct normal distributions (as many as the recombinant strains).

WHY DO YOU USE S AS ABBREVIATION FOR "INFORMATION"? BETTER SOMETHING WITH "I" (I_S, $I_{STD\ INFO}$, PI, ...)

To deal with this problem we transformed the null hypothesis as proposed by Consideration 2b as follows: if the recombinant areas and the non recombinant segments have the same amount of phylogenetic information then the difference of the phylogenetic information in between the recombinant areas and the non-recombinant segments in each strain should be zero or else randomly distributed around zero in such a way that the proportion of the strains having positive difference should be 0.5 or else distributed according to the binomial distribution. More specifically, for each recombinant we calculate the mean of the averaged *s* of the recombinant and the mean of the averaged *s* of the non-recombinant segments as reasonable estimators of the phylogenetic information; in that way we determine whether the mean phylogenetic information was higher in the recombinant areas or the non-recombinant segments. Finally, we calculate the proportion of the recombinants having less phylogenetic information in the recombinant regions (positive difference) than in the non-recombinant ones (negative difference). This proportion is distributed according to a binomial distribution and does not suffer of parametric assumptions. The number of the recombinant strains having less phylogenetic information in the recombinant regions was found to be 24 (73%) and this corresponds to P=0.01 (two-sided test). Consequently, recombinants tend to have less phylogenetic information in the recombinant areas than in the non-recombinant segments.

TO LONG A SENTENCE → REPHRASE e.g. 3 SENTENCES.

DESCRIBE THIS EARLIER

THIS MEANS

→ 24 RECOMBINANT STRAINS OF ~ 32 RECOMBINANT STRAINS ?

WHAT DO THE OTHER 27% STRAINS SHOW ?

YOU NEED EITHER TO DESCRIBE WHY AND HOW YOUR TESTS ARE APPLIED, OR YOU (STILL) HAVE TO GIVE REFERENCE. YOU CANNOT ASSUME THAT ALL READER KNOW THIS, AND IF NOT, YOU HAVE TO DIRECT THEM TO WHERE TO FIND THIS KNOWLEDGE.

Discussion

The effect of recombination in reconstructing the evolutionary history of species is still a vexed question. Several ways have been proposed in order to cope with this kind of data such as the use of networks instead of trees to summarize the chimeric history of evolution (Huson 1998; Huson and Bryant 2006). The puzzling algorithm adopts this network approach by including the partially resolved quartet trees in the tree building procedure (Strimmer and von Haeseler 1996; Strimmer and von Haeseler 1997). These partially resolved quartet trees are in fact networks in their simplest forms. Moreover the fully unresolved quartet trees as implemented in the puzzling algorithm can be considered as the plethoric network of 4 taxa. Regarding the puzzling algorithm, the decrease of not fully resolved quartets is considered evidence degradation of the phylogenetic informative content of the analyzed data. Consequently, by fusion of the network solution for recombinant evolution and the quartet approach for quantifying the phylogenetic information, we a-priori anticipate that recombination should result to degradation of the phylogenetic informative content. The current analysis attempts to verify and detect this theoretical prediction on a real dataset.

Firstly we calculated the distribution of the phylogenetic information along the HIV-1 genome and established a moderate relationship in between sequence similarity and the amount of phylogenetic information of the same region. We have previously shown that intersubtype recombination in HIV-1 has a significant strong correlation with sequence similarity; this could count as a confounder for the relationship observed in between the phylogenetic information and recombination. On the other hand there is a phase difference in between the distributions of recombination frequency and sequence similarity that is not observed in between phylogenetic information and sequence similarity. Moreover the relationship in between the sequence similarity and phylogenetic information is observed in

BETWEEN = RELATING TWO THINGS
IN BETWEEN = SURROUNDED BY THINGS

the baseline non-recombinant dataset. Consequently, the relationship in between sequence similarity and phylogenetic information is not confounded by the recombination. This relationship between sequence similarity and phylogenetic information should be taken into account when design of a molecular epidemiology study is online; selection of one region for massive sequencing based on high similarity for better laboratory performance might lead to minimal resolution of the phylogenetic relationships among the strains.

Furthermore, our analysis evidenced that recombination is a force that not only influences the ability of the data to provide the true tree, by leading to erroneous estimations, but also degrades the informative content of the data to infer any tree (true or wrong). This relationship in between recombination and phylogenetic information could be confounded by the sequence similarity of the region. Nevertheless, as stated before recombination and sequence similarity have a phase shift across the HIV-1 genome. Additionally, the phylogenetic information *s* has been adjusted and standardized for each region and, hence, for sequence similarity.

This relationship, being proven hereafter, brings up the issue of the inheritance of the recombinant areas: these segments as they are transferred through generations and provide information of the evolution after the recombination event, however, they are in fact evolutionary dead-ends since the phylogenetic information before the recombination event is corrupted. This is usually recorded as segments without a specific progenitor and the result in taxonomic studies is not to be able to classify them e.g. in the phylogeny of HIV-1 many recombinant strains have been isolated carrying segments that cannot be classified (Magiorkinis et al. 2005). Hence recombination is increasing the diversification of life, however it diminishes the traits to the roots of the phylogeny and leaves us inside the evolutionary labyrinth looking for the way out.

Figures

Figure 1:

Transition from an unresolved (A) to a fully resolved (C) through a partially resolved quartet tree (B).

Figure 2:

Classification of the areas and segments into recombinant and non-recombinant.

Figure 3:

Distributions of *s* for the recombinant (B) and the non-recombinant (A) regions along with normal density lines.

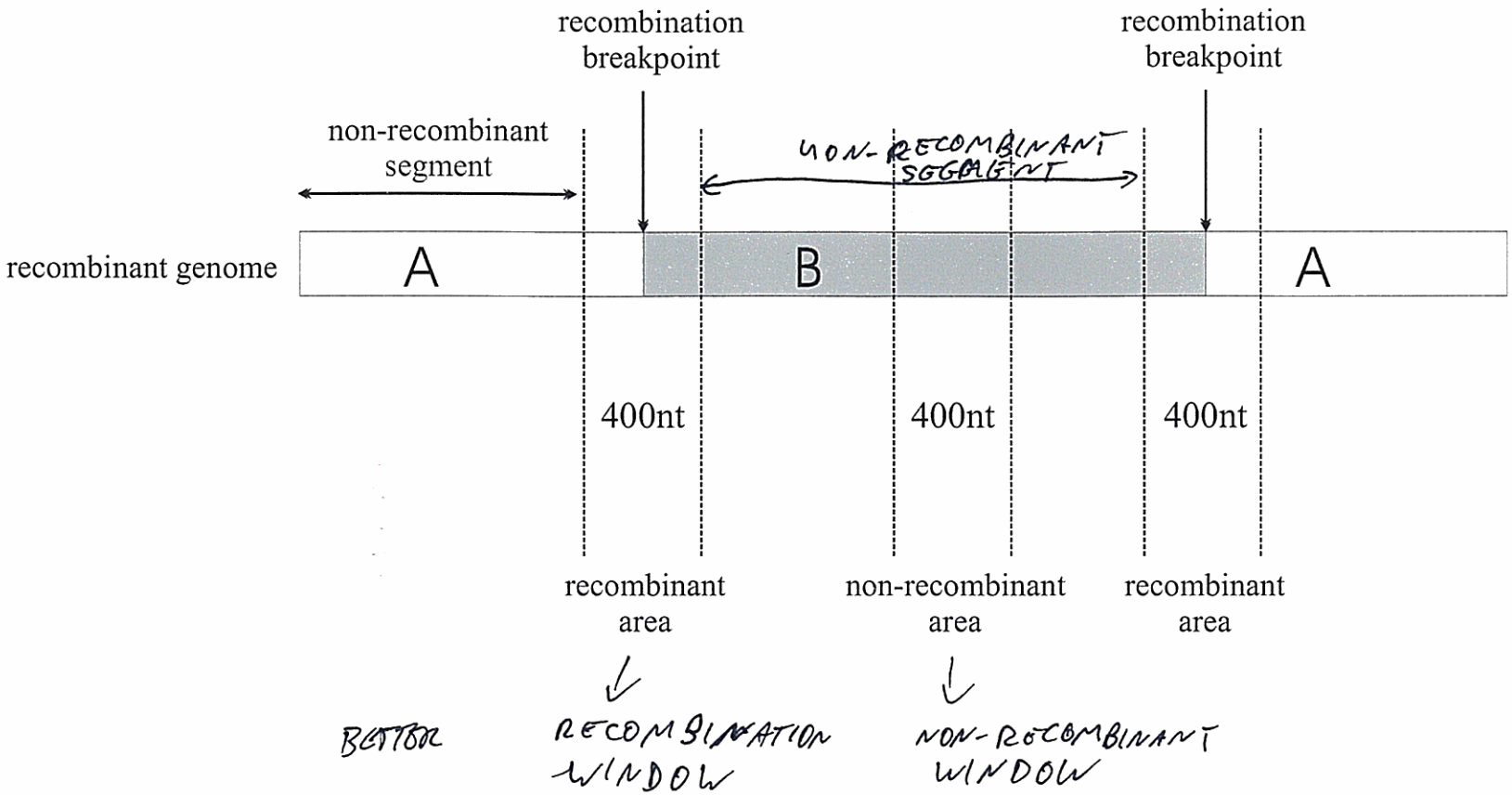
Figure 4:

Average *q* and *sm* calculated for the non recombinant along the HIV-1 genome. Smoothing splines were used to produce the lines from the scatter plots.

Figure 5:

Scatter plot of the *s* and *sm* of non overlapping w_{ps} in the non recombinant dataset.

TO BE HONEST AS A REFEREE I WOULD STILL REJECT OUR PAPER WHILE READING THE METHODS PART! WE HAVE TO TAKE THE READER BY THE HAND AND LEAD HIM/HER THROUGH THE PAPER, WHICH IS NOT THE CASE IN METHODS + RESULTS. IT STILL READS LIKE PATCHWORK, MANY PARTS ARE UNMOTIVATED AT THE LOCATION THEY ARE PLACED!
I LIKE THE DISCUSSION MUCH BETTER THIS TIME, BUT METHODS AND RESULTS SECTION DO NOT CLEARLY LEAD TO IT.

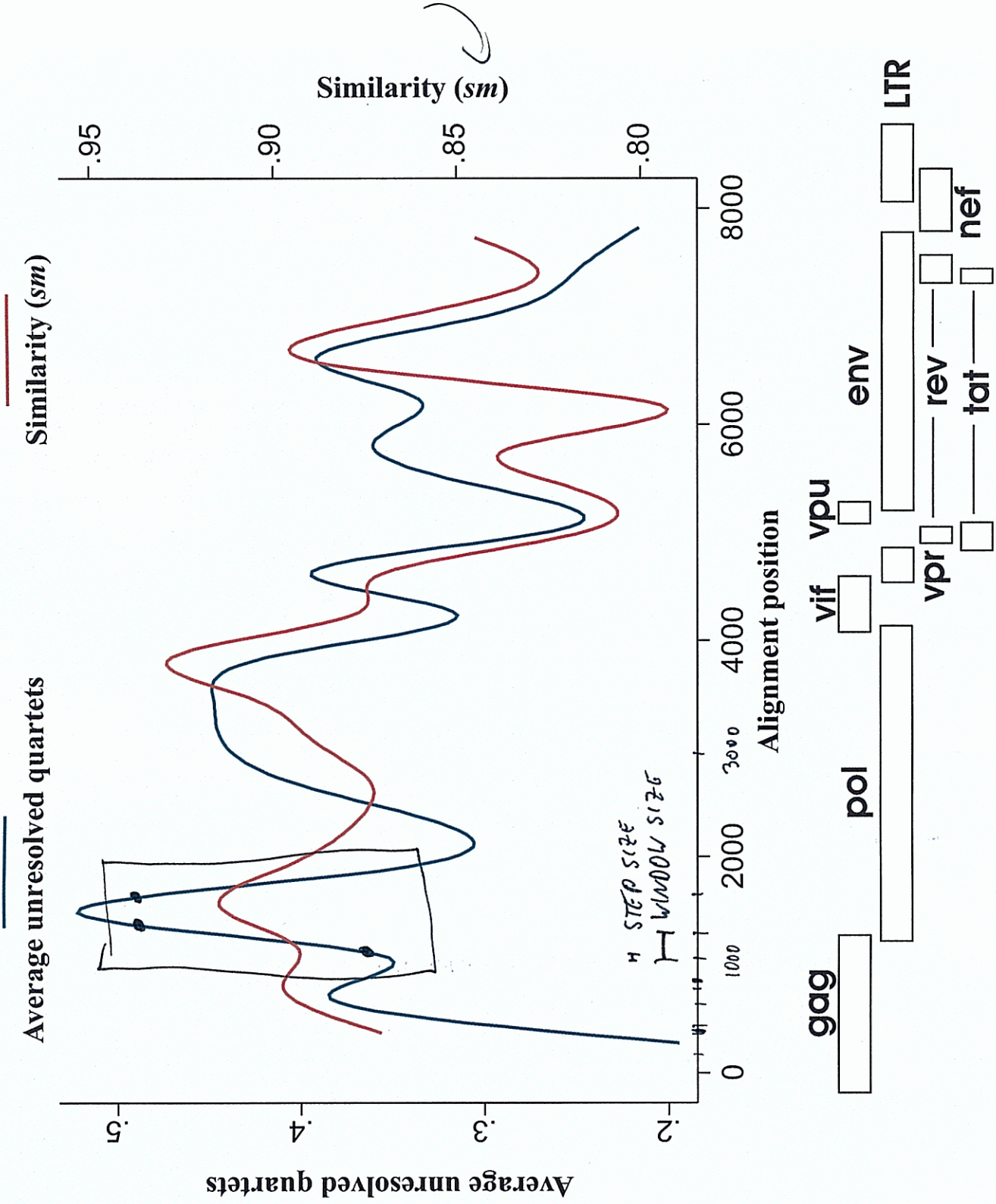


IF WE USE WINDOW INSTEAD OF AREA, IT'S DIRECTLY LINKED TO THE FIXED WINDOW-SIZE OF 400nt WINDOW = WINDOW INSTEAD OF ~~AREA~~ = WINDOW.

FIG 2

ADD THE KNOW BREAKPOINTS HERE AS WELL!

FIG. 4



DESCRIBE WHAT METHOD OF SPLINING YOU USE. SPLINING CAN LEAD TO LARGE, EXCESSES (SEE BOX ABOVE), WHICH SHOULD BE AVOIDED.
(ARTIFICIAL

(THAT, BESIDES ARTIFICIAL CONTINUITY, ARE THE REASON WHY I DON'T LIKE SPLINING!)

MAYBE YOU CAN ADD TWO SCALE BARS FOR WINDOW + STEP SIZE → WOULD CLARIFY THAT SPLINING HAS NOT MUCH (IMPACT.