THE FIRST THING YOU SHOULD DO IS
- ADD REFERENCES !!!
- RESTRUCTURE THE METHOD'S DESCRIPTION.

The phylogenetic information profile of HIV-1 and the devastating effect of recombination.

G.Magiorkinis[1], P.Gouvas[2], D.Paraskevis[1], H.Schmidt[3], G.Mentzas[2], A.Hatzakis[1]

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

CBIV

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

NOT A SINGLE REFERENCE YET !?!

Introduction

Genetic recombination has been suggested to have a pivotal role in the evolutionary pathways. Since it is observed in the more complex species (e.g in the animals as sexual reproduction), it is thought to be historically more recent in the evolution of proliferation; consequently its presence in the upper levels of the evolution is supposed to testify a self-evidential 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. It provides the recombinant subjects several evolutionary properties, such as the "escape from Muller's ratchet" and "broad evolutionary jumps", both accelerating and/or accommodating the expansion of the progeny.

REF ?

VERY HARD TO UNDERSTAND → REPHRASE

WITH?

However, it has been suggested that recombination has a devastating effect in the reconstruction of the species' ~~phylogeny~~ HISTORY. The obvious effect is that one-dimensional linear plots such as the tree-like graphics cannot be used for describing the phylogeny: during recombination each subject SEQUENCE CAN HAVE has more than one ancestor. On the other hand, network like plots have been proposed to provide more realistic phylogenetic reconstructions. Another approach that is widely used especially in the virology of HIV is the breakdown of the genome IN to potentially non recombinant segments guided by an exploratory analysis of the genome and followed by phylogenetic analysis based on conventional (one dimensional-linear) trees.

OTHER WORD (NOT A PROPER CONTRADICTION)

NEEDS DEFINITION WHAT IS THAT

OTHER WORD THAN SUBJECT, SEQUENCE? OR: DIFFERENT PARTS OF ~~THE~~ SEQUENCES HAVE DIFFERENT HISTORIES, I.E. DIFFEREN ANCESTRY.

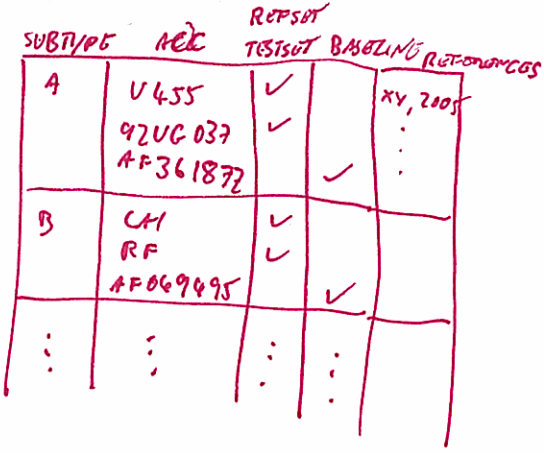
WHAT IS THAT

WHAT IS 1D-LIN IN A TREE-CONTE

It was to our intention to analyze the distribution of the phylogenetic information along the HIV-1 genome and resolve the effect of recombination onto conventional phylogenetic tree reconstruction regarding the phylogenetic informative content by using a well characterized set of recombinant HIV-1 strains.

fully ??

NEEDS A LOT OF RE-WRITING TO MAKE IN UNDERSTANDABLE FOR THE NORMAL USER. I ONLY UNDERSTAND IT ON THE BASIS OF PRIOR DISCUSSIONS, BUT NOT FROM THE TEXT!



Materials and Methods

Data

The recombinant dataset was composed of 33 previously analyzed full-length HIV recombinant sequences (Magiorkinis et al.) (Table 1). The ~~sequence~~ reference dataset was composed of subtypes (isolate numbers): 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). The baseline sequence dataset was composed of subtypes (accession numbers): A (AF361872), B (AF049495), C (AF110964), D (AF484487), F (AF077336), G (AY772535), H (AF005496), J (AF082394), K (AJ249235).

Multiple sequence alignments were achieved by means of the Clustal W program (Thompson et al., 1994).

Puzzle scanning plots

Puzzle scanning plots were built in a similar way to the bootscanning plot according to the exploratory methodology: a window of given size (called puzzle scanning window, *psw*) is slid stepwise along the sequence alignment and ~~phylogenetic analysis~~ is ~~performed~~ by means of the Tree-Puzzle program. For each built tree the following ~~figures~~ are stored and plotted: 1) the puzzling ~~step~~ values supporting the possible monophyli among the query sequence and the rest of the defined strains/groups, 2) the number of the fully resolved, ~~and~~ partially resolved quartets.

The puzzle scanning plots were built using the following parameters:

1) *psw* size: 400 nt
2) step size: 50 nt

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

The analysis was performed on an 8-node Linux cluster implementing unix-shell scripting along with the linux parallel versions of the previously mentioned software components (REF Clustalw-mpi, Parallel tree-puzzle). A simple parser for extracting the previously mentioned figures was developed using macros in the Windows platform.

Sequence similarity across the genome

In order to calculate the sequence similarity across the HIV-1 genome another exploratory algorithm was implemented in a similar way to the bootscanning plot: a window of given size is slid stepwise along the sequence alignment and sequence distances were calculated by means of the DNADIST program of the PHYILIP package. For each slid window the average intersubtype *p* distance was calculated for the subtype reference dataset. The *p* distance was calculated by transformation of the Jukes-Candor distances and similarity (*sm*) is calculated as the complementary of *p* up to1 (*sm*=1-*p*).

information Baseline

In order to quantify the difference (additive or reductive) in the phylogenetic informative content caused by recombination, a non-recombinant standard of the phylogenetic information had to be calculated (HIV-1 non-recombinant information baseline, *Hib*). For this reason we chose a set of non-recombinant HIV-1 strains (one 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 *psw* and set it as a reasonable point estimator of *Hib*.

The choice of using an additional non-recombinant sequence in order to estimate the *Hib* instead of using merely the subtype reference dataset is justified ~~in~~ By the need ~~of~~ to compar~~ing~~ e phylogenetic constructions that contain the same number of taxa (19 sequences for the recombinant and non-recombinant calculations). Since the trees during the puzzle scanning of the recombinant and the non-recombinant baseline ~~were~~ are built containing the same number of taxa (19), phylogenetic inferences should have potentially and theoretically the same statistical power and, consequently, the informative content should not be degraded due to inference of additional parameters.

WHY?

POWER OF WHAT?

Definition of Variables

In order to elucidate the variables used in our analysis a terminology ~~was~~ will be defined as follows:

Recombination breakpoint: ~~As~~ We define a recombination breakpoint ~~was defined~~ as that ~~the~~ region in the middle of which a switch of the monophyletic clustering of the query sequence ~~was~~ is observed.

WHAT MIDDLE? DEFINE ON THE BASIS OF THE SLIDING WINDOWS!

→ WHAT IS MONOPHYL. CLUSTERING?

u (unresolved quartets): number of not fully resolved quartets = sum of partially unresolved and fully unresolved quartets

q (proportion of unresolved quartets): *u* / overall number of the attempted quartets

d (difference of *p*): *q* in a specific region (window?) of the recombinant – *q* in the regarding region of the averaged non-recombinant baseline

WHAT IS P? GIVE IT A NAME.

WHAT'S THE DIFFERENCE?

s (standardized *d*): *d*/(*q* in the regarding region of the non recombinant baseline)

AS SUGGESTED I THE CORRECTIONS OF THE LAST MANUSCRIPT:

VARIABLES
u, q, d, p, s, P
BETTER (SEE LAST TIME)
q_{unres}, q_{res}
q_{PART}, q_{nonres}

REGION = WINDOW? (I ASKED THAT BEFORE!) OR REGION ~~SEM~~ COLLECTION OF CONSECUTIVE WINDOWS WITH SAME TREE ?!?

$q_{nonres} = q_{unres} + q_{PART}$

$q = q_{res} + q_{nonres} = \binom{|ALL\ SEQ|}{4}$

USE MORE SPEAKING VARIABLE

(CHECK WITH MY OLD CORRECTIONS)

RESULTS ?

Information measures - Observations

Our main goal was to compare the phylogenetic information contained in the recombinant and the non-recombinant regions. However, definition of the observation of information is not intuitive and care has been taken in order to define a variable that can be used as comparative parameter. Firstly, for each strain the genome was divided into regions called after "recombination breakpoints" (as previously defined) and into the remaining genomic regions; *s* was chosen to describe the phylogenetic informative content of them.

The following considerations were made in order to define a reliable and comparable estimator for both the reombination brekpoint and the remaining regions:

Consideration 1: Compare all the respective *psw* defining whether a *psw* contains or not a recombination breakpoint.

If we chose to compare all the possible *psw* then the observations would not be independent since these windows are partially overlapping (Pearson's autocorrelation coefficient for adjacent windows (lag=1) is 0.58 (P<0.001))

Consideration 2: Compare the phylogenetic information content in the recombinant and non-recombinant regions. Two different approaches for calculating the contained phylogenetic information were considered.

a) Information (*s*) could be calculated by conducting phylogenetic analysis on the areas (alignment fragments) defined as "recombination breakpoints" and the remaining genome. However, in this case we would have compared the phylogenetic information of regions of unequal size biasing thus the analysis: larger regions (data) tend to have smaller *s* because they contain more sites and consequently the inferences have more power than in the small regions.

WHAT IS A REC. AND NON-REC. REGION? YOU HAVE REC + NON-REC. SEQUENCES! (SIMILAR QUESTION TO LAST TIME!)

NOT PROPERLY DEFINED (SEE LAST TIME!)

REF (SEE LAST TIME)

WHAT ABOUT NON OVERLAPPING WINDOWS OF SAME SIZE? HOWEVER, I HAVE NO GOOD IDEA HOW TO SPLIT THAT BETWEEN BREAKPOINTS

b) Information could be calculated as the average *s* of the *psw* that have their midpoints inside the limits of the corresponding regions. Thereby the measurements are comparable (*psw* have potentially the same statistical power since they have the same length), while the autocorrelation is attenuated (Pearson's autocorrelation coefficient (lag=1) is reduced from 0.58 (P<0.001) for the overlapping *psw* down to 0.07 (P=0.13) for the adjacent regions) making the observations virtually independent.

STILL, WHAT IS A REGION? GIVE A CLEAR DEFINITION

AND WHAT ABOUT THE OVERLAP CRITICIZED BEFORE.

→ YOU HAVE TO EXPLAIN TO THE READER WHAT THE VALUES OF 0,58 OR 0,07 MEAN. I STILL MISS ALL REFERENCES!

YOU SHOULD RESTRUCTURE THE WHOLE METHODS PART! IN THE CURRENT FOR ONE CANNOT UNDERSTAND IT!

I SUGGEST A STRUCTURE LIKE

① WE HAVE AN ALIGNED DATA SET, WE SPLIT THE DATA SET INTO 2 SUBSETS, (a) BASELINE SET NOT CONTAINING ANY RECOMBINATED STRAINS, BUT ONLY PRESUMABLY "CLEAN" REFERENCE STRAINS AND (b) THE TEST/ANALYSIS ... SET.

② ~~FOR~~ EACH SUBSET WE BREAK INTO OVERLAPPING WINDOWS OF SIZE W_{SIZE} AND ~~STEP~~ OVERLAP ($W_{SIZE} - W_{STEP}$). FOR EACH WINDOW WE COMPUTE A TREE WITH QP AND STORE: THE TREE, THE SPLITS WITH SUPPORT, THE FRACTION OF UNRESOLVED (q_{UNRES}), PARTLY (q_{PART}) AND RESOLVED (q_{RES}) QUARTETS. THE ~~LATTER TWO~~ LATTER TWO ARE COMBINED TO $q_{nonres} = q_{unres} + q_{PART}$ CONTAINING THE SUM OF NON-RESOLVED TREES. FOR EACH WINDOW i WE GET 2 VALUES $q_{nonres}(i, TEST\ SET)$ AND $q_{nonres}(i, BASE)$.

③ WE COMPUTE AN STANDARDIZED INFORMATION VALUE FOR EACH WINDOW i DEFINED AS $I(i) = \frac{q_{nonres}(i, TEST) - q_{nonres}(i, BASE)}{q_{nonres}(i, BASE)}$.

(-) YOU HAVE OF COURSE TO EXPLAIN HOW TO GET FROM TREES, QUERIES (ALSO MISSING IN MY PROPOSAL - IS EACH ONE USED OR ONLY ONE? SEE LAST CORRECTION!), AND WINDOWS TO REGIONS. A BREAKPOINT CANNOT BE A REGION. THE FORMER IS A POINT ---- THE LATTER IS AN AREA. ---

(+) THE FOLLOWING OF ABOVE'S VALUES ARE THEN PLOTTED INTO DIAGRAMS TO CHARACTERIZE AND OBSERVE THE STANDARDIZED PHYLO_GENETIC INFORMATION I

- IN THE DESCRIPTION YOU SHOULD USE CLEAR ~~DESCRIPTION~~ DEFINITION AN MAKE CLEAR - WHEN AND WHY DO WE USE WINDOWS OR REGIONS
- HOW DO WE GET THEM
- WHY DO WE USE THE BASELINE SET, INCLUDING ALSO A CLEAR DESCRIPTION OF THE AUTOCORRELATION STUFF AND - WHY REGIONS/WINDOWS CAN, THUS, BE USED AS IF INDEPENDENT.

THE I WOULD END THE MATHEMATH SECTION BY SHOWING A TABLE (SEE BEFORE) ~~AND~~ WITH OUR DATA, THAT THEY ARE ALIGNED WITH CCLUSTALW-MPI (WHY NOT T-COFFEE? SHOULD GIVE BETTER RESULTS) AND CHECKED MANUALLY.

AND A "HARDWARE/IMPLEMENTATION" SECTION STATING THAT THE ANALYSIS IS IMPLENTED AS SCRIPTS (WHAT DO YOU USE? BASH, PERL, PYTHON, CAML, JAVA, ...) AND WHAT HARD-WARE IS USED IN OUR CASE.

THIS SHOULD BRING THE DESCRIPTION OF THE METHOD INTO ONE LINE TO BE READ THROUGH - AT THE MOMENT THE PARTS ARE NOT PROPERLY ORDERED.

Results

TO INTRODUCTION (ITS ALL EXPLAINATION)

Information is supposed to be the measure which quantifies the update on our prior beliefs as soon as an observation has taken place. In phylogeny the universal prior belief is that all species have a common ancestor. This may be simply represented by a star-like tree among the species (Figure 1). A fully resolved tree is the one that contains only bifurcations. Consequently, phylogenetic information may be considered as the data measure that updates our prior beliefs towards a fully resolved tree. An indirect way to quantify this amount of information is the percentage of fully resolved quartets during the quartet-puzzling algorithmic reconstruction. We ~~firstly~~ defined standardized measures for the phylogenetic information based on unresolved, partially and fully resolved quartets as previously described. Consequently, the measurement used in the ~~current~~ analysis is ~~the~~ *s* which is inversely related to the phylogenetic information: a positive value of *s* means ~~that~~ more unresolved quartets ~~were calculated in~~ the query strain than ~~in the~~ baseline ~~meaning~~ that less phylogenetic information is contained in this ~~phylogenetic inference~~.

Each full-length sequence was profile aligned to the reference dataset and subsequently a puzzle-scanning plot was built. The numbers of resolved, partially resolved and fully resolved quartets for each ~~psw~~ were collected to form a database. Coordinates of the mosaic patterns were used from a previous analysis (Magiorkinis et al. REF), according to which 235 breakpoints were defined ~~overall~~ while the remaining (not containing breakpoints) regions were 251. The average *s* (Figure 2) for both breakpoints and remaining regions approximately ~~follow~~ normal distributions.

→ UPDATE OF WHAT? OR DOES IT QUANTIFY DIRECTLY?

→ YOU CAN'T REFER TO AN UNROOTED TREE FOR THAT.

ALSO NORMALLY DEFINED ON ROOTED TREES.

I DON'T UNDERSTAND THAT. (SEE ABOVE AS WELL)

WHY "INDIRECT"?

WILL

THAT'S CLUSTLHOOD AND QUARTET MAPPING (REF: STRIMMER + V. HAESELER + NIESELT + V. HAESELER)

BELOW THE STD. INFORMATION S EXPECTED FROM THE SET.

EXISTS FOR

HOW IS THAT DEFINED

THAT MEANS, AUGMENT WINDOW.

METHODS PART

DEFINE

→ MIGHT CHANGE THE WINDOW FRAME BY INSERTING BARS. HOW IS THAT ACCOUNTED FOR?

NOT TO UNDERSTAND WITH BREAKPOINT + REGIONS NOT PROPERLY DEFINED.

IS

LY

ED

Consequently, the mean of the averaged *s* is a good estimator of the phylogenetic information for the recombinant or non-recombinant regions.

As shown in Figure 2, it is obvious that the distributions of the *s* in both the recombination breakpoints and the rest of the genome are strongly shifted towards positive values meaning that the phylogenetic information is reduced with regard to the *Hib*. This may be due to the fact that the recombination events are temporally old in such a way that the recombinants have diverged from the parental strains and significantly evolved towards saturation, which is recorded as noise of the phylogenetic information.

In Figure 3 we show how the average *q* calculated for the non recombinant set fluctuates across the HIV-1 genome as well as the *sm*. The plots were produced by using smoothing splines and were overlaid to graphically check the possibility of co-fluctuation. The overlaid plots were suggestive of a moderate correlation that prompted for testing measurably the strength and significance of this correlation. The Spearman's rank correlation was implemented for non-overlapping respective windows of intersubtype *s* and *sm* that suggested a moderate though significant relationship (*rho*=0.66, P<0.001) (Figure 4).

Moreover, we performed a standard t-test to compare the means of the averaged *s*. The mean averaged *s* is higher in the recombinant regions (0.051 vs 0.047) but it was not found to be significant (*P*=0.15, two-sided, unequal variances). Nevertheless there seems to be perturbation of the underlying assumptions that might have biased the specific test: since each recombinant strain has suffered different evolutionary pathways the degradation of the phylogenetic information has occurred in distinct proportions. Consequently, *s* would not be drawn from a single normal distribution, but from several distinct normal distributions disturbing, thus, the basic

assumption of a standard t-test. To cope with this problem we also tested indirectly for the increase of the phylogenetic information by performing a non-parametric test on the averaged *s* in the recombinant and non-recombinant regions.

For each recombinant the difference in-between the means of the averaged *s* of the recombinant and non-recombinant regions was calculated to define whether the mean phylogenetic information was higher in the recombinant (positive value) or the non-recombinant regions (negative value). Finally, we calculated the proportion of the recombinants that had less phylogenetic information in the recombinant regions than in the non-recombinant ones. 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 regions than in the non recombinant ones.

ALSO THE TESTS MUST BE EXPLAINED IN THE METHODS WORKFLOW IN THE MAT + METH SECTION:

OBSERVING (X) IN THE DIAGRAMS WOULD INDICATE (Y). TO CERTIFY THE APPLY TEST (Z) BECAUSE ... (MAKE CLEAR WHY TESTS CAN BE APPLIED OR WHEN ASSUMPTIONS ARE VIOLATED)

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 (Reference). The puzzling algorithm partially adopts this network idea by including the partially resolved quartet trees in the tree building procedure. 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 tree-puzzling algorithm, the decrease of not fully resolved quartets is considered to evidence degradation of the phylogenetic informative content of the analyzed data. This is especially evidenced when the alignment (data matrix) is short, containing less informative sites and consequently suffering of minimal statistical power (less observations infer the same number of parameters).

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 phyogenetic information of the same region. Finally, 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 statistical power of data to infer any tree (true or wrong).

TO SUMMARIZE CLIMATIC HISTORY OF EVOLUTION.

(NOT REALLY)

QUARTET

CAN BE DRAWN AS

QP USES ALL 2-3 SUPPORTED TREES RANDOMLY CHOSING 1 IN THE CURRENT PUZZLING STEP.

QUARTET

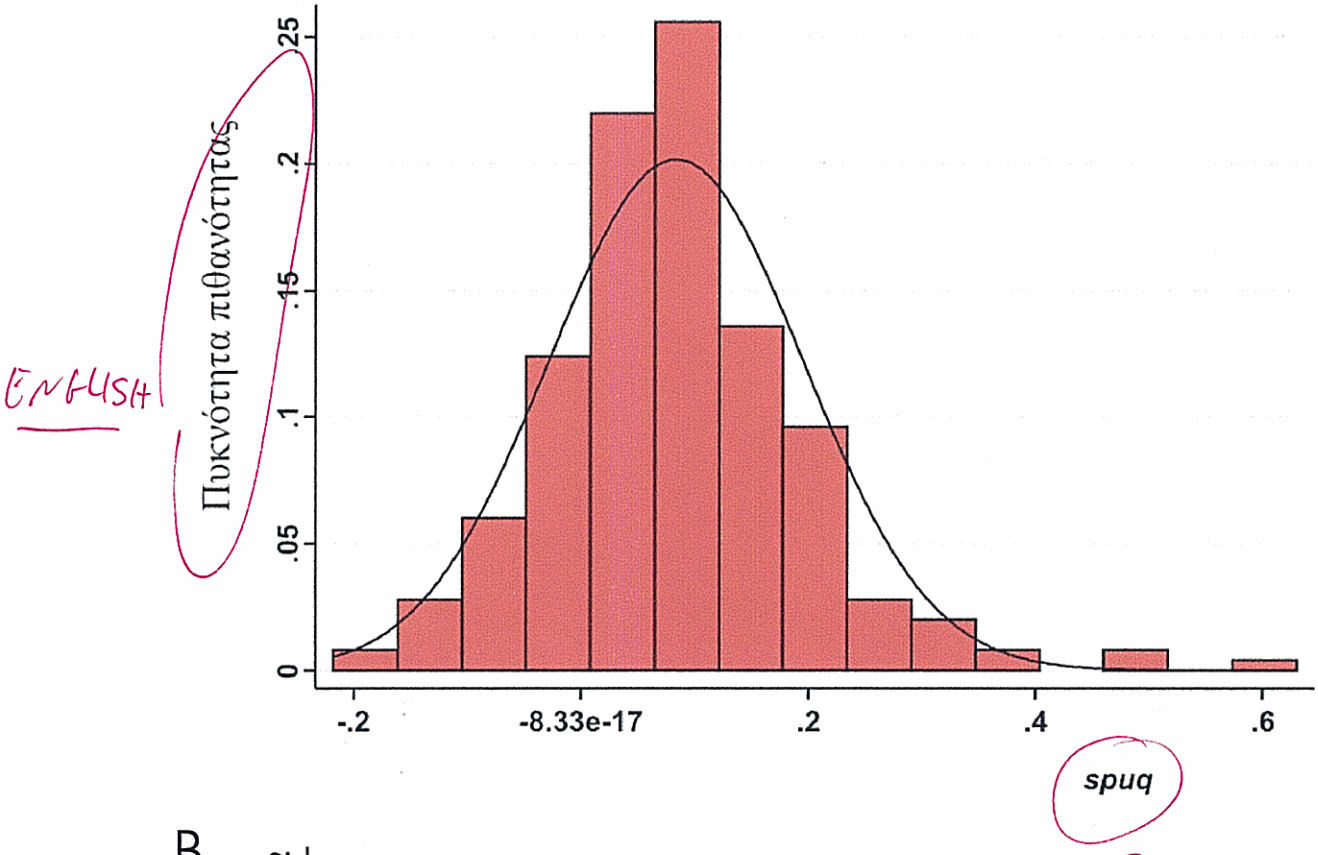
FOR

OR CONTAINED IN

NOT NECESSARILY. CLOSELY OR DISTANTLY RELA SEQUENCES MIGHT BE MORE SEVERE AS MIGHT BE CONTRADICTIONS OF DIFFERENT PARTS OF THE ALIGNMENT.

DO WE HAVE A BETTER (MORE CLEAR) EXAMPLE? DO WE ONLY USE A SINGLE RECOMBINANT HERE? OR ARE THERE SEVERAL.

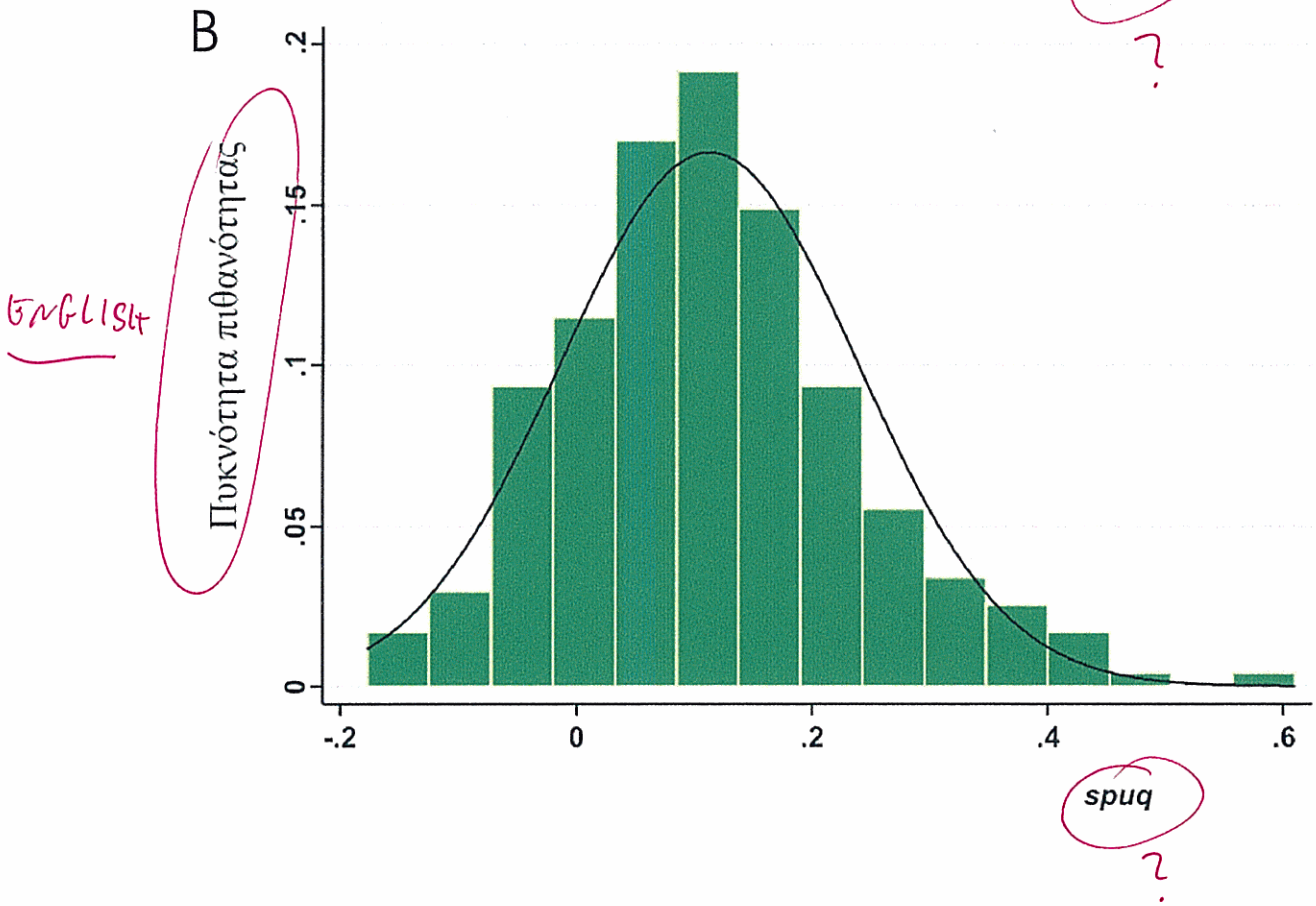
A



ENGLISH

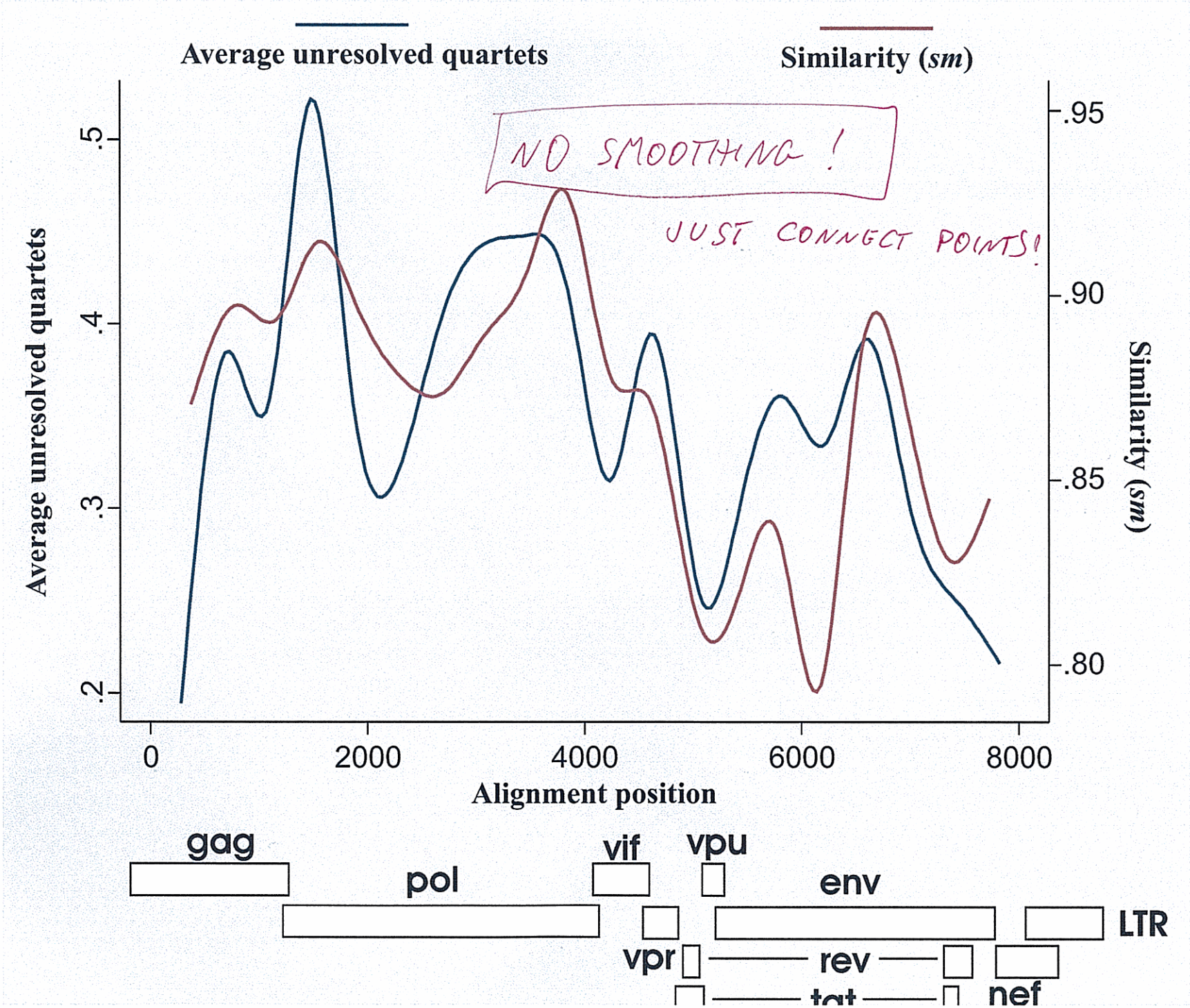
?

B



ENGLISH

?


Average unresolved quartets
Similarity (sm)
NO SMOOTHING !
JUST CONNECT POINTS!
Average unresolved quartets
.2
.3
.4
.5
.80
.85
.90
.95
Similarity (sm)
0
2000
4000
6000
8000
Alignment position
gag
pol
vif
vpu
env
LTR
vpr
rev
tat
nef