Deliverable 2.2

Tools for data enhancement

Project acronym: *DYNANETS*

Project full title: Computing Real-World Phenomena with Dynamically Changing Complex Networks.

Grant agreement no.: 233847

Due-Date:	M 12
Delivery:	M14 M17 resubmission following Review 1
Lead Partner:	EMC
Dissemination Level:	Public
Status:	Re-submission
Approved:	Q Board, Project Steering Group
Version:	V8.0 Resubmission following review 1 October 2010

DOCUMENT INFO

0.1 Authors

Date and version number	Author	Details
September and October, 2010	David van de Vijver	Revision of D2.2 that was submitted to EC but was rejected.
28/10/2010	George Kampis	Editing, Revisions QAP

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

Objectives

The objective of this deliverable is to develop tools for data enhancement.

Context of this deliverable within workpackage 2

In deliverable 2.1 we described a large number of datasets that is available to DynaNets. Deliverable 2.2 will develop tools which can be used to uncover relevant and non-trivial correlations between the structure and dynamics of the network on the one hand (e.g. the sexual network) and the nodes features on the other (e.g. demographic data). The latter will be reported in deliverable 2.3.

Rationale for use of phylogenetics in DynaNets

One of the main topics that DynaNets will study is transmission of drug resistant HIV-1. To address this important issue, DynaNets can strongly benefit from phylogenetics which can identify clusters of genetically related HIV strains which will arise in transmission events.

1. Onward transmission of drug resistant variants among patients that have not been treated (yet) with antiretroviral drugs is an important factor driving transmission of drug resistant HIV. Phylogenetics can determine the extent by which onward transmission plays a role in the spread of HIV.
2. Phylogenetics can be used to study the spatial spread of micro-organisms.
3. Phylogenetics could be used as approximation of transmission networks.
4. DynaNets has access to >15,000 HIV-1 infected patients for whom data about drug resistance is available (see deliverable 2.1).

Need for data enhancement tools

Several analytical tools exist for phylogenetic analysis. Unfortunately, there is no consensus on definition of clusters. We previously found that well-established phylogenetic techniques as maximum-likelihood cannot identify automatically all transmission events. We therefore first developed a quick novel tool which may be of help in identifying these events, without the need of estimating a phylogenetic tree. In addition, when considering phylogenetic trees, identification of clusters has to be done

manually. Automatic identification, which is needed as we have access to substantial numbers of patients, cannot be done. Our second tool analyses a tree phylogeny and can automatically identify phylogenetically related clusters.

Tool 1 – Phylogeny-free identification of clusters

We developed a new clustering method which is named TBC (Threshold Bootstrap Clustering). The advantage of TBC is that it is computationally less intensive as other currently used methods, which saves DynaNets a lot of time in analyzing the data. In addition, TBC can automatically identify clusters. TBC takes advantage of resampling techniques and models of sequence evolution, but does not require phylogenetic tree estimation. The TBC uses as input a multiple alignment of molecular sequences and its output is a crisp partition of the taxa into an automatically determined number of clusters.

We evaluated TBC in the problem of transmission event identification using a dataset from patients followed up at the Catholic University of the Sacred Heart in Rome. TBC was unfortunately less good in correctly identifying transmission events than manual inspection of a phylogenetic tree that was calculated using the well-established technique of maximum-likelihood. We therefore also developed a second tool that exploited the phylogenetic information (since the TBC is not based on phylogenetic tree estimation).

Tool 2 – automatic identification of clusters of phylogenetically related sequences

We developed a novel method for automatic identification of phylogenetically related clusters. The method was based on a statistical comparison of intra- and inter-cluster branch length distance distributions. The method was used to identify clusters using the ARCA repository which includes >10,000 sequences. The method was validated on an independent test set of known transmission events and yielded >80% of concordance with the visual inspection of the phylogenetic tree. In addition, when applied to the ARCA data base, it was able to identify transmission events that were previously confirmed, reported and published.

Why are these tools of benefit to DynaNets

The first tool will not be used by DynaNets for the transmission event identification, as we found that it is not reliable enough for this task. The second tool will be used in deliverable 2.3. In this deliverable we will analyse the sequences from the SPREAD network. SPREAD includes patients newly diagnosed with HIV-1 that are representative

for the risk group and geographical distribution of HIV-1 across Europe. Importantly, SPREAD also contains a large number of sequences (>3,000) and the method proposed in this deliverable is therefore helpful. We will also repeat the analysis using sequences collected among patients newly diagnosed with HIV-1 through MSM-contacts in the ErasmusMC in Rotterdam, the Netherlands. An advantage of the data collected in Rotterdam is that the ErasmusMC is the only hospital treating HIV-1 infected individuals in the South Western region of the Netherlands. The catchment area includes more than one million individuals. Data obtained at ErasmusMC are therefore a dense sample.

Work on (partially) missing data

In this deliverable we also report on methods for handling missing data in a experiment on human contacts that took place in Dublin. We do however recognize that we have to do more work on tools for handling (partially) missing data. We will therefore report on this during the second review.

2 Contributors

Several institutes have contributed to this deliverable. The focus on phylogenetics was agreed upon by all partners in December of 2009 in Amsterdam. The contents were presented in May 2010 in Budapest. In particular, the contribution was as follows:

Partner	Contribution
EMC	Writing of executive summary, introduction, conclusions. Organization of several teleconferences to discuss progress
UCSC	Development of phylogenetic tools. Writing of research papers
ISI	Work on missing data in human contacts experiment

2.1 Introduction

DynaNets decided during the kick-off meeting in Amsterdam in June 2009 that the project will use data available from phylogenetic analysis. In section 2.1.1, we will first outline why we have a focus on phylogenetic analysis. Section 2.1.2 will explain that novel methods for data enhancement in phylogenetics are required.

2.1.1 Rationale for use of phylogenetic analysis in DynaNets

DynaNets will develop and use complex network analysis to study, amongst others, infectious diseases. One of the main topics that will be investigated is the spread of drug resistant HIV-1. This topic is important for public health as currently about 10% of all patients newly diagnosed with HIV-1 became infected with a virus that is resistant to at least one antiretroviral drug [1, 2]. Transmission of drug resistance can have important clinical repercussions as patients may have to start more toxic second line treatment.
We have identified four rationales why DynaNets can benefit from phylogenetics:

1. Onward transmission of drug resistant HIV-1
2. Geographical spread
3. Phylogenetic clusters follow a scale free network
4. Availability of data

Rationale 1 Onward transmission of drug resistant HIV-1

A drug resistant variant can be transmitted from a patient receiving antiretroviral drug treatment that shows virological failure and in whom a drug resistant variant has emerged. The mutational patterns in these patients that failed treatment are frequently complex and involve resistance to a large range of different antiretroviral drug classes [3]. On the other hand, a patient can become infected through onward transmission of drug resistant variants between patients that have not been treated for HIV. Mutational patterns in transmitted resistance usually involve single mutations and resistance to single drugs [4].

Phylogenetic analysis have shown that transmission of drug resistant HIV-1 is frequently the consequence of onward transmission [5-8]. Moreover, transmission of HIV is for a large part ascribed to patients that are in the early acute stages of infection when

patients have frequently not been identified as infected and are not receiving treatment yet [7, 8]. Finally, transmission clusters with sustained drug-resistant variants have been identified across Europe [9, 10].

Phylogenetic analysis can therefore provide valuable insights into the spread of (transmitted drug resistant) HIV-1. Previous modelling efforts have not taken onward transmission among untreated patients into account [11].
The extent to which onward transmission plays a role in the spread of HIV should be taken into account in DynaNets.

Rationale 2 Geographical spread
Phylogenetic analysis can be used to monitor the spatial spread of micro-organisms [12-14]. Using phylogenetic information may therefore allow DynaNets to include spatial information.

Rationale 3 Phylogenetic clusters follow scale free network
Obtaining information about sexual networks is difficult as available information may not always be reliable. DynaNets wants to investigate if phylogenetics can be used as a proxy for sexual networks. Preliminary work that will be reported in D2.3 showed that phylogenetically identified clusters in data available to DynaNets followed a scale free network. The number of sexual partners that an individual has also follows a scale free network [15-17]. Other have also reported that phylogenetically identified clusters follow a scale-free network [18].

Rationale 4 Availability of sequences that can be used for phylogenetics
Treatment guidelines recommend to perform a genotypic resistance test as soon as patients are diagnosed with HIV-1 [19-21]. As a consequence, substantial numbers of HIV-1 sequences are available that can be used in phylogenetic analysis. DynaNets has access to large numbers of sequences through SPREAD (>3,000 sequences), ARCA (>10,000), ViroLab (>3,000) and the hospitals participating in this project. Availability of these large numbers of data is a clear strength of our project and can for instance not be obtained in other parts of the World.

2.1.2 Need for data enhancement tools

A large number of analytical tools exist for phylogenetic analysis [22]. Nonetheless, there are two important reasons why new tools for data enhancement should be developed:

1. No consensus on definition of a cluster
2. No good method for automatic identification of clusters

Need 1 No consensus on definition of a cluster

There is no clear consensus on the definition of clusters of phylogenetically related sequences. Archer et al. developed a method which classifies clusters using rigorous statistical standards. The method is implemented in a software packaged named CTree [23]. The method is however computationally intensive and therefore requires a lot of time. Also, the maximal number of taxa allowed in CTree is 125 and DynaNets has access to >10,000 sequences (see deliverable 2.1). Lewis et al. found that sequences a cluster could be defined within HIV-1 sequences that have a crude genetic distance to each other of at most 4.8% [18]. The cut-off was identified using available data. Known transmission pairs were not used. It is unknown if this arbitrary cut-off is also valid in other populations. It is also not known if the method can be applied to other viruses.
In this deliverable, we will present work we did on a novel method for classifying clusters.

Need 2 No good method for automatic identification of clusters

DynaNets has access to a substantial number of sequences. Identification of phylogenetically related sequences (so-called clusters) can only be done manually. This means that the result of the analysis (a phylogenetic tree) should be printed and then manually checked for relevant clusters. This is problematic when thousands of sequences have to be analysed. (Such large numbers are needed as transmission of resistance only occurs in 10% of patients).
In this deliverable we describe a novel method for automatic clustering using the ARCA database which included >10,000 sequences.

2.2 Definition of phylogenetically related clusters

In this section we will describe a novel method for defining phylogenetically related sequences. The method is described as a scientific paper that has been accepted for publication by Plos One.

One of the main objections of the reviewers to the previous version of this deliverable was that we had to summarize the methods and results instead of giving entire publications. We subscribe to this point of view. Nonetheless, we also believe that we should include the efforts that have gone into development of the tools. We have therefore included the paper, but also included a summary and discussion at the end of the section. Parts of the scientific paper are in grey. The summary and discussion is given as a boxed text.

2.2.1 Introduction

During the past forty years a plethora of methods that infer phylogenetic trees have been introduced, based on genetic distances, evolutionary parsimony, maximum-likelihood and Bayesian theory [22].

By cutting a phylogenetic tree at some level(s), it is possible to induce a partition of the taxa and define clusters, identifying thus non-overlapping groups of taxa or transmission events. However, the procedures for selecting optimal phylogenetic tree cut points have not been widely explored. The state-of-the-art method is a heuristic procedure that examines inter-cluster and intra-cluster distance distributions and gives a partition of the set of taxa in a phylogenetic tree, by considering the patristic distance matrix, implemented in a software named CTree [23]. This algorithm has a drawback in its complexity, which is cubic in the number of taxa. CTree has been successfully validated with the classification of type-1 human immunodeficiency virus (HIV-1) group M subtypes, but would be hardly applied for the identification of transmission clusters within large phylogenetic trees (up to several thousands of taxa, whilst the maximal number of taxa allowed in CTree for automatic cluster determination is 125). In fact, recent literature that addressed the HIV-1 transmission event identification, defined a partially-overlapping set of clusters based on a thresholding of the genetic distance matrix of the viral sequences [2, 18]. The identified clusters were then confirmed by looking at the phylogenetic tree and verifying that they were together in a subtree highly supported by the resampling statistics.

This manuscript introduces a new partition technique, the threshold bootstrap clustering (TBC), to address the taxa clustering, the transmission group identification, and the intra-patient quasispecies characterisation. This new algorithm is remarkably linked with the Chinese restaurant process, previously employed both for the clustering of microarray gene expression data [24] and for haplotype identification in ultra-deep sequencing [25].

The TBC uses models of sequence evolution and performs resampling of sequence alignments, and does not require phylogenetic tree estimation. TBC automatically determines the number of clusters without additional steps. The computational complexity of the TBC has a quadratic upper bound, lesser in one order of magnitude than the complexity of the CTree algorithm.
Finally, coupled with the TBC, we define a methodology for assessing its robustness, calculating partition likelihood and cluster reliability, which indeed is independent on the clustering techniques and can be used with any other partition method.

2.2.2 Methods

The core of TBC method is inspired by a Chinese restaurant process (also known as Dirichlet process), a discrete-time stochastic process [26, 27]. The process can be described with the metaphor of a (Chinese) restaurant with infinite tables, where customers walk in and sit down at a table. The tables are chosen according to the following random process: (a) the first customer always chooses the first table; (b) the n^{th} customer chooses the first unoccupied table with probability $a/(n-1+a)$, and an occupied table with probability $c/(n-1+a)$, where c is the number of people sitting at that table and a is a scalar parameter of the process. Intuitively, each customer entering the restaurant sits at a table with probability proportional to the number of customers already sitting at it, and sits at a new table with probability proportional to a. Thus, customers tend to sit at most “popular” tables, that become even more crowded. By this, the process has a “power law” behaviour, where a few tables attract the majority of the customers, and the parameter a determines how likely a customer is to sit at a new table. Usually, in real-world problems, the Chinese restaurant process is used a prior and a Gibbs’ sampler is employed [24, 25].

In the TBC the probability assigned to any particular cluster slightly depends on the cluster size itself (this is accounted indeed in the refinement step), whilst the chance for a given object to join a cluster or to form a new one depends on how much the object is “similar” to other objects in a cluster, with respect to a known distribution that describes the overall object (dis)similarity. Since we intend to cluster molecular sequences, the measure of dissimilarity can be a genetic distance calculated via a specific evolutionary model. In addition, the TBC is run on a column-wise bootstrap sample of the original alignment, shuffling the sequence alignment order: this allows to obtain potentially different partitions when executing the TBC, using random seeds for shuffling and

bootstrap (see the next section for the likelihood assessment of partitions and the cluster reliability calculation).

The TBC algorithm starts with a multiple sequence alignment A, $|A|=n$. The algorithm initially shuffles the sequence order and draws a column-wise bootstrap sample of the alignment B. A preliminary phase creates an a-priori distribution D_B of random pair-wise distances from B and calculates a threshold value t, corresponding to a x^{th} (usually 5^{th} or 10^{th}) percentile of D_B. Then an empty list of clusters C is initialised. The sequences are scanned sequentially and the first sequence s_1 induces a first cluster $s_1=c_1 \epsilon C$. The second sequence is compared with the first cluster and if the median value of the distance distribution obtained by comparing s_2 with all the elements in c_1 (now there is only one element in c_1) is below the threshold t, then s_2 is assigned to c_1, otherwise forms a new cluster. The same holds with sequence s_3, which is compared with c_1, and eventually with c_2, if s_2 had formed a new cluster. Either the cluster list or the size of a cluster grow by continuing the sequence scan and distance threshold comparison. At iteration i, sequence s_i id compared with the cluster list $C = c_1, ..., c_k,..., c_j$, where $j<=i$. The comparison starts from $k=1$ and proceeds until j, stopping in between if s_i joins a certain cluster c_k. If s_i is assigned to cluster c_k, $c_k=c_k \cup (s_i)$, otherwise the cluster list is incremented by a new cluster $c_{j+1}=(s_i)$, i.e. $C=C \cup (c_{j+1})$. After each sequence has been examined ($i=n$), a post-processing phase starts. By following the Chinese restaurant dogma, for which popular clusters tend to attract single elements, we calculate a distribution of cluster sizes and we delete the clusters whose size is below the 5^{th} (or 10^{th}) percentile. Then the cluster assignment phase is re-run for those sequences belonging to the deleted clusters. Finally, the number of clusters corresponds to the size of $|C|$. The TBC algorithm is explained in detail in Table 1. Of note, the TBC shares a few common points with the cluster affinity search technique proposed for gene-expression data clustering, although the cluster construction procedure is quite different.

The computational complexity of the TBC is $O(n^2)$, where n is the number of objects to be clustered: in fact, the threshold assessment phase requires n^2 random object comparisons for the estimation of D_B, and the sequence scan phase in the worst cases would create either a unique cluster or n distinct clusters, corresponding to $n(n+1)/2$ comparisons.

In order to speed up the algorithm in our implementation, we limited the number of random distances to be calculated in the interval [1000, 500000], with an additional

control on the threshold t, calculated at every 100th iteration, stopping the procedure if the difference between two consecutive estimated thresholds was below 0.0000001. In the sequence scan phase, each distribution D_{ij} was approximated by considering a limited number of comparisons with the objects in a cluster equal to the square root of the cluster size, with a minimum of 10 comparisons (unless the cluster size was smaller) and a maximum of 100, i.e. $min(max(sqrt(|c_i|),10),100)$.

Assessment of partition likelihood and cluster reliability

The TBC clustering induces a full partition of the taxa objects into p clusters, where p is automatically determined. By varying the initial conditions (i.e. random seeds for taxa shuffling and sequence bootstrap), TBC can produce different partitions, both in the number of clusters and in the elements belonging to each cluster. As maximum-likelihood and Bayesian estimations are used to select both for best phylogenetic trees under a set of model parameters, and to infer node reliability, we might be interested to assess the most plausible partition(s) obtained from multiple runs of the TBC and to determine reliability of each cluster. Of note, such a methodology would apply to any clustering technique that can produce different partitions by varying its initial conditions.

By reviewing the literature, this problem has gained growing attention in the recent years, acquiring the name of “consensus” or “ensemble” clustering [28]. Consensus clustering tries to find a single partition which is a better fit under some goodness-of-fit functions with respect to other existing partitions. The consensus partition does not necessarily coincide with any of the original partitions. The cluster-based similarity partitioning algorithm, the hyper-graph partitioning algorithm, or k-means based algorithms [29] are a few of the many variations on a theme.

We propose here, differently from most of consensus clustering algorithms, a methodology that selects *one* particular partition in a set of obtained partitions.

Partitions can be compared statistically to determine their agreement, using the adjusted Rand index (ARI) [30], an indicator of cluster agreement which corrects for chance and takes values in [0,1]. By using the ARI, given a set of partitions P, $|P|=m$, we can compute the likelihood of a partition with respect to the others $p_i \in P$ as $L(P|p_i)=Pr(p_i|P)=\prod_{j\neq i}a_{ji}$, where a_{ji} is the ARI between partition p_j and p_i, and then select the best partition p^b with the maximum likelihood. In this case, we are assuming that the ARI

is directly proportional to a probability, i.e. $a_{ji} \propto L(p_j|\, p_i)=Pr(p_j|p_i)$.

Once the best partition is determined, we estimate the reliability of each cluster with a procedure similar to the posterior probability estimation for nodes of a phylogenetic tree under Bayesian monte-carlo analysis [22]. In detail, the partitions $p_i \in P$ are ordered decreasingly by their associated likelihood and the last x^{th} percentile (usually from the 75^{th} or above) of partitions is deleted. Each retained partition p_i is compared with the best partition p^b and for each cluster $c^b_i \in p^b$ a support value is defined as follows: (i) for each partition $p_j \in P$, $j \neq b$, identify the cluster $c^*_{jk} \in p_j$ such as $c^*_{jk} \cap c^b_i$ is the maximum among all possible intersections $c_{jk} \cap c^b_i$; (ii) calculate the support s^b_{ij} as $s^b_{ij} = c^*_{jk} \cap c^b_i / c^*_{jk} \cup c^b_i$. Then the overall support s^b_i for a cluster c^b_i is the average value of all s^b_{ij}.

Data sets, software and settings of comparison methods

The TBC has been entirely implemented in java. Procedures for likelihood and cluster reliability assessment have been written using the R mathematical software suite. The whole source code is available as a supplementary material. The CTree [23] algorithm was used as a comparison method, along with the PAM (using the LogDet estimator as a distance measure) where the optimal number of clusters was assessed via the average silhouette value maximisation [31].

As a final evaluation (vi), the TBC was also applied to a set of HIV-1 group M subtype B polymerase sequences obtained from the private CUSH clinical data base, identifying viral isolates coming from patients with known transmission history (collecting any sequence at any time point). A set of control sequences was added to this data set: specifically, samples coming from other HIV positive patients followed up at CUSH, with unknown transmission history, two outgroups (HIV-1 subtypes C and J), and the reference HIV-1 subtype B HXB2 strain. Sequences were aligned using ClustalW [32]. Resistance of each viral sequence with respect to an antiretroviral class (nucleoside-tide/non-nucleoside/protease inhibitors) was defined as the presence of at least one amino-acidic mutation panelled by the International AIDS Society (any major mutation for protease) [33], by aligning pair wisely each sequence against the HIV-1 consensus B, with an in-house modified version of the local Smith-Waterman-Gotoh alignment algorithm implemented in java [34]. Columns of the multiple alignments corresponding to codon positions previously associated to drug resistance were deleted, in order to avoid the possible bias coming from convergent evolution due to treatment experience. (Note: in the original manuscript TBC was also evaluated for identification of HIV and hepatitis C sequences. These analysis are available upon request).

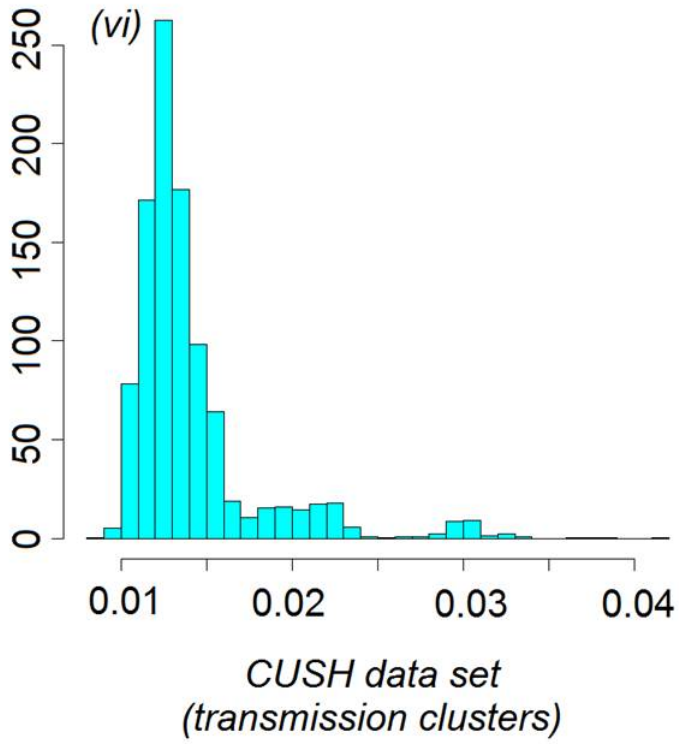
For each TBC analysis, the distance threshold percentile was evaluated in the interval *t*=[1.0, 45] with step sizes of 0.5/0.25, and 200 bootstrap runs were executed. For the analyses of HIV/HCV subtyping (reference data sets), CTree has been executed on phylogenetic trees constructed by neighbour-joining and LogDet estimator. A patient from the data set analysed by Shankarappa et al. [35] was analysed in depth by estimating a Bayesian phylogeny using BEAST [36], with the default parameter settings, checking after 10 million generations both the auto-optimisation benchmarks and the effective sample size. For the analysis of transmission clusters on CUSH data, a maximum-likelihood tree was estimated, setting up a general-time-reversible model, with a 20-parameter gamma optimisation, and a mix of nearest-neighbor interchanges and subtree-prune-regraft moves for tree topology search, using the FastTree software [37]. Reliability of each tree split was calculated by a Shimodaira-Hasegawa test. The internal sensitivity parameter of CTree was always set to 1 (slowest as concerns computational time, but correspondent to maximal accuracy).

2.2.3 Results

Transmission cluster identification

The last analysis (vi) was executed on the CUSH dataset, comprising HIV-1 subtype B isolates from patients with known transmission history: there were 12 known transmission events from patient-to-patient (n=66 sequences, with 5.5 sequences per transmission event, and exactly two patients in each transmission event), 6 control patients from CUSH (12 sequences), two outgroups (subtype C and J), and the subtype B HXB2 reference isolate. The best ARI between transmission groups and clusters generated by TBC was 0.682 at *t*=7, with a median (IQR) cluster support of 75% (56%-88%). All control sequences except one were correctly placed, whereas, when looking at the transmission events, only 3/12 (25%) transmission events were uniquely determined by placing both patients (and only those) in the same cluster (supplementary Figure S1). The TBC was able in general to identify and cluster viral isolates belonging to the same patient, but not extremely sensitive to identify transmission events, although the sample size of this experiment was small and sparse as concerns times of sampling. The CTree algorithm yielded a poorer performance, with an ARI=0.35, identifying correctly only 2/12 (16.7%) transmission events. The PAM did not identify any transmission group, and selected only 2 clusters, with an ARI of 0.001.

As an additional comparison, we used also the method proposed in [2, 18], which can be considered the state-of-the-art with respect to HIV-1 transmission cluster identification. The procedure identified 3/12 (25%) clusters that were confirmed by node reliability values >90% from the maximum-likelihood phylogenetic tree.



2.2.4 Summary and Discussion

In this manuscript we introduced the threshold bootstrap clustering, a new methodology for partitioning molecular sequences. The TBC is inspired by a stochastic Chinese restaurant process and takes advantage of resampling techniques and models of sequence evolution. The TBC uses as input a multiple alignment of molecular sequences and its output is a crisp partition of the taxa into an automatically determined number of clusters. By varying initial conditions, TBC can produce different partitions.

TBC was evaluated in the problem of transmission event identification. Using a data set of patients followed up at the Catholic University of Sacred Heart in Rome, Italy, with known transmission history, TBC was able to identify transmission events in 25% of cases. The transmission event dataset was also evaluated using a previously published method [2, 18], specifically tuned for HIV transmission cluster identification, and that

method identified 25% of transmission events. With a human-visual evaluation of subtrees and node reliability of a maximum-likelihood phylogenetic tree, we were able to infer correctly 50% of transmission events. Thus, even a detailed phylogenetic analysis was not able to resolve all transmission events. In fact, for HIV it has been shown previously that many factors (such as long period of infectivity, sparse time and space sampling) can limit the concordance of phylogenetic reconstruction and the reported epidemiological evidence [38, 39]. The transmission event data set of CUSH was composed by sequence samples of patients taken at different times and disease stage: some patients were sequenced multiple times either before treatment initiation or at treatment failures, whilst others had only one sequence sample taken. We recognise that a larger and less sparse data set would be desirable in order to assess better the TBC performance on this particular problem.

TBC has the advantage of a quadratic complexity, and there is the possibility to identify a consensus partition and a measure of cluster reliability. Although conceptually different and presumably with less expressional power than a full phylogenetic analysis, the TBC might be useful for the processing of large-scale sequence data sets, where both the phylogenetic software and the standard clustering algorithm might be hardly applicable.

Why is this tool of benefit to DynaNets?

There is a discussion ongoing about methods for defining phylogenetically related clusters. As such, definition of clusters is arbitrary. We attempted to develop a new method for identification of clusters. The method was as better or as good as established methods. Unfortunately, the method is not good enough to be used in DynaNets. (We also used this method successfully for classifying HIV-1 subtypes and HCV genotypes).

What does this tool mean for DynaNets

Based on this analysis we decided to focus on a different method. A method that can automatically identify clusters of phylogenetically related sequences. This is described in section 2.3

2.3 Automatic identification of clusters of phylogenetically related sequences

In this section we will describe the tools we developed for automatic identification of phylogenetically related sequences. Similarly to section 2.3, we have included the scientific paper and made that grey as it could be skipped. The summary and discussion is boxed.

2.3.1 Introduction

Recently, medium/large-scale phylogeny, coupled with new methodologies of cluster/network analysis, was applied to European HIV genomic data [40], specifically in the United Kingdom [2, 9, 18] and Switzerland [10, 41]. These works mark out foundations of our study: in this paper we will introduce a new methodological approach for cluster analysis of large phylogenetic trees and apply it for analysing the subtype B HIV-1 epidemic in Italy, considering viral genomic samples from a large national cohort and linked demographic and clinical information. The data will be extracted from the Italian Antiretroviral Resistance Cohort Analysis (ARCA) observational database [42]. ARCA started in 1995 and is an almost nationwide initiative, that stores data from HIV-infected patients followed at 105 centres, collecting demographic information, hepatitis B and C status, AIDS defining events, HIV-RNA load, CD4+, therapy change episodes, and HIV genotypes (the oldest sequence is dated in 1991). As of Apr 9, 2010, data from 20,200 patients and 23,051 HIV-1 polymerase/envelope sequences were available.
The newly developed method will also be used to determine which demographical, clinical and epidemiological factors are associated with clustering.

2.3.2 Methods

Data

HIV-1 polymerase sequences of ART-naive and ART-experienced patients were extracted from ARCA, with sampling dates, plus corresponding information (where available) on patients' age, gender, country of origin, zone of residence, mode of HIV transmission, date of first HIV positive test, date of first ART, HIV-RNA load and CD4+ cell count contemporary to the sampling date (+/- 30 days), and seroconversion period.

Sequences

Viral subtype was assigned by employing the Rega subtyping tool [43]. All subtype B sequences were aligned using the parallel implementation of ClustalW, and a multiple alignment was generated for subsequent phylogenetic analysis. Resistance of each viral sequence with respect to an antiretroviral class (nucleoside-tide/non-nucleoside/protease inhibitors) was defined as the presence of at least one amino-acidic mutation panelled by the International AIDS Society (any major mutation for protease) [44], by aligning pairwisely each sequence against the HIV-1 consensus B, with an in-house modified version of the local Smith-Waterman-Gotoh alignment algorithm implemented in java. Columns of the multiple alignment corresponding to codon positions previously associated to drug resistance were deleted.

New method for clustering

Maximum-likelihood phylogenetic analysis was performed on the aligned sequences, adding HIV-1 subtypes J and C as outgroups. The parallel implementation of FastTree software [37] was used, setting up a general-time-reversible model, with a 20-parameter gamma optimisation, and a mix of nearest-neighbor interchanges and subtree-prune-regraft moves for tree topology search. Reliability of each tree split was calculated by a Shimodaira-Hasegawa test.

After obtaining a phylogenetic tree, tree topology was analysed with a breadth-first visit by considering the number of subtrees/clusters (and associated number of leaves) with a number of distinct patients >=3, and a node reliability >=0.9 on each tree level. Additionally, by performing a depth-first search on the tree, we analysed the variations in the number of clusters comparing each intra-cluster median branch length difference with the overall inter-cluster branch length difference distribution, selecting clusters with a number of distinct patients >=3, a node reliability >=0.9 and a median intra-cluster distance below a variable threshold (ranging from the 5^{th} to the 100^{th} percentile of the inter-cluster distance distribution with a step of 0.05). Of note, this approach is similar (but faster) to the algorithm for automatic cluster detection introduced by Archer [23]. Cluster partitions were compared by using the adjusted rand index [30].

Statistical analysis

Multivariable logistic analysis was performed in order to identify prognostic factors of transmission clusters. Missing values of numerical variables were replaced by the average. Because the same patient could contribute more than one sequence at different times, a generalised-estimating-equations model was performed [45].

2.3.3 Results

Data

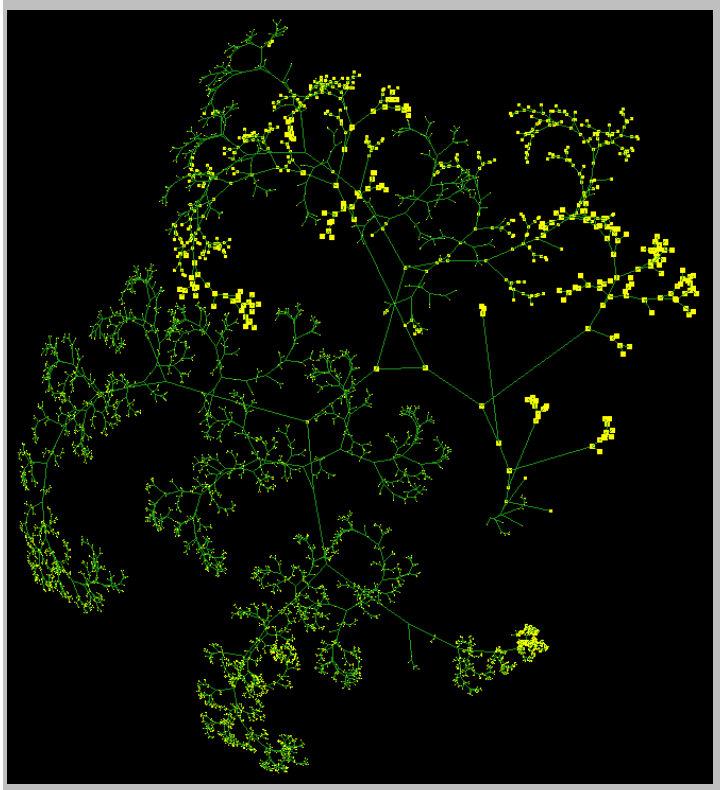
A total of 11,543 HIV-1 subtype B sequences were included. Multiple alignment required approximately 10 hours on a two quad-core 64bit Intel Xeon X5550@2.66GHz, with Hyper-Threading technology, and 24GB DDR3 RAM. Pairwise alignments and parallel FastTree phylogeny software run for ≈15 and ≈30 minutes.

Phylogenetic analysis

Figure 1 shows the phylogenetic tree that was generated using maximum-likelihood phylogeny. The complexity of the tree shows how difficulty of manual identification of clusters.

Figure 1.

Maximum-likelihood phylogenetic tree of n=11,543 HIV-1 subtype B sequences (plus two outgroups HIV-1 subtype C and J) from the Italian ARCA data base



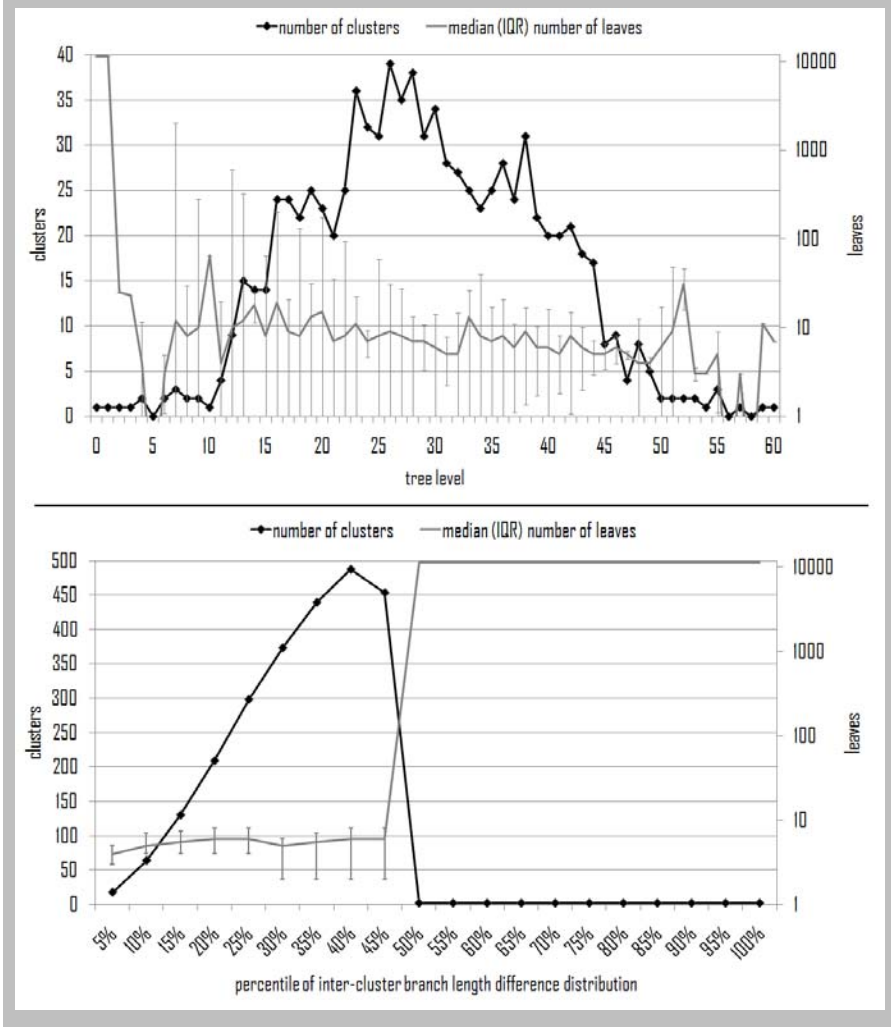
By performing a breadth-first visit on the tree, the number of clusters *n* (nodes with reliability>0.9 and at least 3 sequences from 3 distinct patients) followed a bell-shaped trend by increasing the tree level *t*. The maximum number of clusters *n*=39 was found at

t=26, where the maximum tree level was 60. The median number of leaves *l* in each cluster showed a slightly decreasing trend by increasing the tree level (Figure 2 upper panel).

By applying a depth-first search and selecting clusters with the additional constraint of a median intra-cluster branch length difference below a variable inter-cluster percentile threshold difference *t ϵ [5, 100]*, the number of clusters retrieved at the 10th and 25th percentile were 64 and 298, respectively. The peak was at the 40th percentile, consisting of 487 clusters. After the 50th percentile (where the median inter-cluster is equal to median inter-cluster branch length difference) the number of clusters decreased at the constant value of 2, which was the number of subtrees with a reliability >0.9 starting from the outgroup J root.

Figure 2

The number of clusters and the median (IQR) number of leaves in each cluster by executing either the breadth-first (upper panel) or the depth-first (lower panel) tree



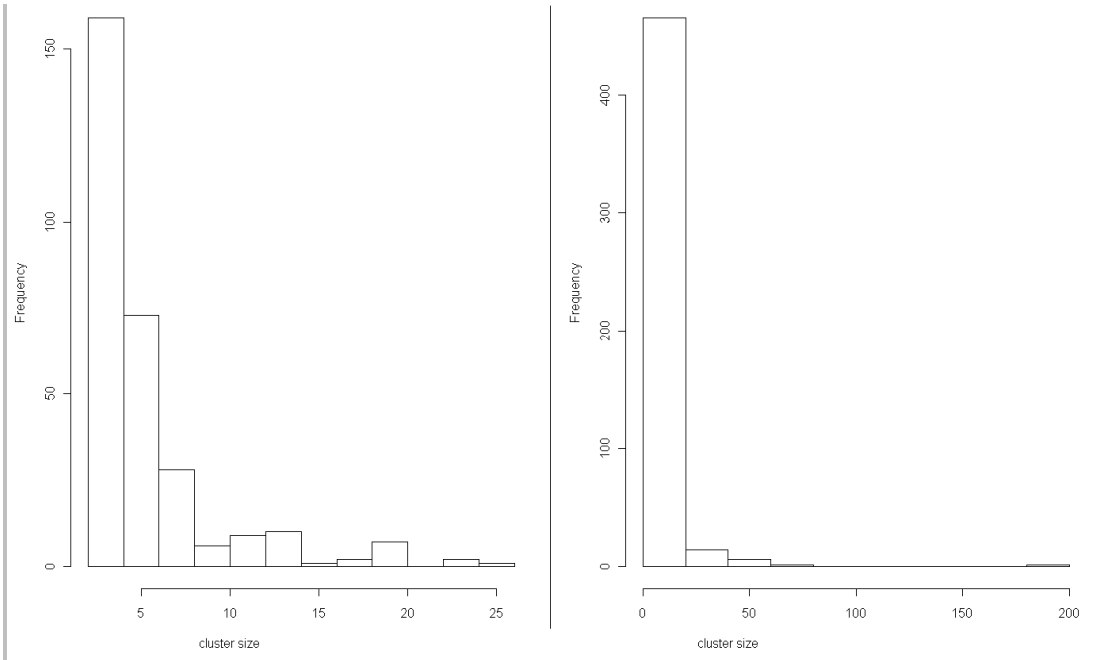
If we compare the two procedures for cluster determination, on average in the breadth-first the number of clusters is lower and the cluster size is higher than in the depth-first

Of note, for the depth-first search, the calculation of the branch length difference distribution for a (sub)tree with *k* leaves (either inter-patient, which is calculated among all leaves, or inter-patient, for each subtree), requires as upper bound *k*(k-1)/2* pairwise comparisons, corresponding in our settings to 66.6 millions. For computational issues, we limited the number of comparisons at a maximum of 5 millions, selecting at random without replacement the leaves from the tree (a complete run required ≈20 minutes on the same hardware platform described above).

By considering the cluster partition obtained with the depth-first search at the 25th percentile threshold, each of the 298 clusters contained a median (IQR) of 4 (3-6) distinct patients (Figure 3, left panel). At the 40th percentile threshold, each of the 487 clusters contained a median (IQR) of 5 (3-7) distinct patients (Figure 3, right panel). While the maximum cluster size obtained with the 25th percentile threshold was 26, with the 40th percentile threshold there were eight clusters composed by >40 patients (n=42, 42, 46, 47, 47, 50, 64, and 197, respectively). However, the two partitions (the clustered sequences), were highly similar, with an adjusted rand index of 0.84. For this reason, subsequent analyses were carried out considering the partition with the 25th percentile threshold.

Figure 3.

Distribution of cluster sizes (one sequence per patient) by applying a depth-first tree search and selecting nodes with reliability>0.9, at least 3 sequences from 3 distinct patients, and median intra-cluster branch length difference below either the 25th (left panel) or the 40th (right panel) inter-cluster percentile threshold difference.



Statistical analysis

Descriptive statistics of the study population is shown in table 1.

Table 1.

Descriptive statistics of the study population.

Variables		n	%
Sequences		11541	100%
Patients		9855	85.39%
Mode of HIV transmission	Injecting drug user	1956	19.85%
	Heterosexual	1984	20.13%
	Homo/bisexual	1874	19.02%
	Other/unknown	4041	41.00%
Gender	Female	2351	23.86%
	Male	6909	70.11%
Country of origin	Italy	5560	56.42%
	Out-of-Italy	300	3.04%
	Unknown	3995	40.54%
ART-naive		1014	8.79%
ART-experienced		6959	60.30%
Unknown ART status		3568	30.92%
Any class resistance	ART-naive	119	11.74%
	ART-experienced	5224	75.07%
Resistance to NRTI	ART-naive	67	6.61%
	ART-experienced	4952	71.16%

Variables		n	%
Resistance to NNRTI	ART-naive	59	5.82%
	ART-experienced	2371	34.07%
Resistance to PI	ART-naive	41	4.04%
	ART-experienced	2665	38.30%
Numerical markers		**median**	**IQR**
HIV-RNA Log10 copies/ml	ART-naive	4.33	4.09-5.06
	ART-experienced	4.09	3.73-4.27
CD4+ count cells/mm3	ART-naive	373	263-471
	ART-experienced	373	273-388
Age (years)	ART-naive	38	32-43
	ART-experienced	41	38-43
Time from first HIV+ test (years)	ART-naive	0	0-2
	ART-experienced	11	7-15

IQR: interquartile range

ART: antiretroviral therapy

NRTI: nucleoside/tide reverse transcriptase inhibitors

NNRTI: non-nucleoside reverse transcriptase inhibitors

PI: protease inhibitors

Twenty-three-percent of the ART-naive patients clustered. There were 6/298 (2%) clusters composed exclusively by ART-naive patients, all resistance-free. Table 2 gives a description of the univariable odds of a unique-factor versus a mixed-factor clustering (excluding clusters with missing information) and compares them with the overall factor proportions observed in unclustered sequences.

Table 2.

Description of odds of a unique-factor versus a mixed-factor clustering (eliminating clusters with unknown information) and comparison with the overall proportion observed in unclustered sequences.

		Clustered patients/sequences							Unclustered patients/sequence (n=8170)	
all patients		tot no. of clusters	unique	% unique	not unique	% unique over total	odds	p-value	no. of patients	%
mode of HIV transmission	IDU	126	35	76.1%	11	27.8%	3.18	0.0004	1726	21.1%
	heterosexual		38	71.7%	15	30.2%	2.53	0.0016	1628	19.9%
	homo/bisexual		36	81.8%	8	28.6%	4.50	<0.0001	1493	18.3%
ART status	naive	158	6	7.0%	80	3.8%	0.08	0.0039	769	9.4%
	experienced		137	56.1%	107	86.7%	1.28	0.0548	4558	55.8%
country of origin	Italy	109	100	94.3%	6	91.7%	16.67	<0.0001	4679	57.3%
	out-of-Italy		3	33.3%	6	2.8%	0.50	0.3173	246	3.0%
gender	female	263	55	49.5%	56	20.9%	0.98	0.9244	1921	23.5%
	male		152	73.1%	56	57.8%	2.71	<0.0001	5716	70.0%

presence of drug resistance mutations	any class	298	113	48.3%	121	37.9%	0.93	0.6010	4987	61.0%
	NRTI		104	46.2%	121	34.9%	0.86	0.2571	4608	56.4%
	NNRTI		28	18.8%	121	9.4%	0.23	<0.0001	2262	27.7%
	PI		54	35.1%	100	18.1%	0.54	0.0002	2181	26.7%
time from first HIV+ test	<4 years	74	10	45.5%	12	13.5%	0.83	0.6698	769	9.4%
	>= 4 and <9 years		5	15.2%	28	6.8%	0.18	0.0001	674	8.2%
	>=9 and <14 years		3	6.7%	42	4.1%	0.07	<0.0001	768	9.4%
	>=14 years		6	17.6%	28	8.1%	0.21	0.0002	895	11.0%

		Clustered patients/sequences							Unclustered patients/sequence (n=796)	
ART-naive patients		tot no. of clusters	unique	% unique	not unique	% unique over total	odds	p-value	no. of patients	%
presence of drug resistance mutations	any class	86	9	69.2%	4	10.5%	2.25	0.1655	96	12.5%
	NRTI		7	77.8%	2	8.1%	3.50	0.0956	54	7.0%
	NNRTI		5	62.5%	3	5.8%	1.67	0.4795	46	6.0%
	PI		5	83.3%	1	5.8%	5.00	0.1025	28	3.6%
time from first HIV+ test	<4 years	49	40	60.6%	5	46.5%	8.00	0.0116	417	54.2%
	>= 4 and <9 years		3	23.1%	5	3.5%	0.60	0.0522	55	7.2%
	>=9 and <14 years		0	0.0%	3	0.0%	NA	NA	20	2.6%
	>=14 years		0	NA	0	0.0%	NA	NA	13	1.7%

Heterosexuals and homo/bisexuals were less likely to appear in clusters composed by mixed risk groups rather than in unique clusters (the odds being the strongest in homo/bisexuals), along with patients from central Italy and males. On the contrary, patients from southern Italy, ART-naive, carrying non-nucleoside or protease resistance mutations, at any time from first HIV+ test, were more likely to cluster with patients of different factor strata. In the subset of ART-naive patients, those carrying any resistance mutation (overall and in all classes) were less likely to cluster with patients not harboring any resistance (although not significantly due to the limited sample size), along with those recently infected.

For the patients with known seroconversion date, we found evidence of transmission from ART-experienced patients in 10/312 (3.2%, 95% CI 1.6%-6.7%) cases, when there was a patient in a cluster whose seroconversion date was precedent than the sequencing date of another ART-experienced patient in the same cluster, using as a denominator the total number of patients with known seroconversion date. Of these 10 cases, 4 (40%, 95% CI 13%-95%) where carrying at least one resistance mutation. There was one additional case of resistance transmission (from an ART-naive patient), that hardly allowed us to estimate the percentage of onward transmission as 0.321% (95% CI 0.008%-1.77%). Finally, the overall estimated proportion of resistance transmission in patients with known seroconversion date was 1.6% (95% CI 0.6%-3.5%).

When performing multivariable analysis on the full data set, factors significantly associated with transmission clusters were: a more recent calendar year, living in the central Italy, having acquired HIV infection via sexual contacts as compared to injecting drug user route, a younger age, a more recent HIV+ test, and carrying at least one mutation in the protease gene associated to major drug resistance. Table 2 summarises odds-ratios, confidence intervals and robust p-values for each variable in the model.

By restricting the analysis on the subset of patients with complete demographic information (age, gender, country of origin, zone of residence, mode of HIV transmission, ART status, and time from first HIV+ test), all odds-ratios were confirmed. In addition, the presence of at least one resistance mutation to non-nucleoside reverse transcriptase inhibitors showed a lower risk of transmission (OR= 0.65, 95% CI 0.56-0.76, p=0.005), along with a lower HIV-RNA load (OR=0.88 per one log10 copies/ml lower, 95% CI 0.83-0.94, p=0.044). The same analysis on the ART-naive patients (first sequence) showed consistent results, although most of p-values were above 0.05. Of note, the same analyses carried out using the clustering obtained with the 40th percentile threshold were in agreement with presented results (not shown).

Table 3.

Adjusted odds-ratios of transmission clustering evidence from fitting a multivariable generalised-estimating-equations model (with binomial logistic link).

Factor		*OR*	*95% CI*	*p-value*
calendar year (per 10 more recent)		1.92	(1.51-2.43)	<0.0001
mode of HIV transmission	heterosexual versus IDU	1.42	(1.14-1.77)	0.00178
	homo/bisexual vs IDU	1.67	(1.33-2.09)	<0.0001
	other/unknown vs IDU	1.84	(1.48-2.29)	<0.0001
gender	male vs female	0.92	(0.79-1.07)	0.2903
	unknwn vs female	0.72	(0.52-1)	0.05073
country of origin	out-of-Italy vs Italy	0.87	(0.61-1.23)	0.42676
	unknown vs Italy	1.46	(1.23-1.74)	<0.0001
age (per 10 years older)		0.84	(0.77-0.9)	<0.0001
ART status	ART-experienced vs ART-naive	1.18	(0.95-1.47)	0.14241
	unknown vs ART-naive	0.59	(0.47-0.74)	<0.0001
time from first HIV+ test	<4 years vs >14 years	2.13	(1.59-2.85)	<0.0001
	between 4 and 9 years vs >14 years	1.41	(1.06-1.86)	0.01727
	between 9 and 14 years vs >14 years	1.25	(0.96-1.64)	0.09787

	unknown vs >14 years	1.36	(1.06-1.73)	0.01457
presence of at least one drug resistance mutation	resistance to NRTI yes vs no	0.88	(0.76-1.01)	0.063
	resistance to NNRTI yes vs no	0.88	(0.77-1.002)	0.05409
	resistance to PI yes vs no	1.56	(1.36-1.78)	<0.0001
HIV-RNA per one log10 copies/ml higher		1.06	(0.98-1.14)	0.14644
CD4+ count per 50 cells/mm3 higher		1.004	(0.989-1.02)	0.57323

2.3.4 Discussion

Main finding

We have presented a novel method for automatic identification of phylogenetically related clusters. For this purpose, we used a collection of more than 10,000 sequences. Such a large sample is required as DynaNets will, amongst others, study transmission of drug resistant HIV which occurs in ~10% of patients newly diagnosed with HIV-1 [46].

Two partition methods (a breadth-first and depth-first visit with specific node constraints) were compared on the sequence hierarchy. The depth-first search was found more appropriate for our purposes, being able to select a higher number of clusters with a lower number of patients and an intra-cluster branch length difference sensibly lower than the overall inter-cluster difference distribution. The new partition algorithms proposed in this work may be a methodological advancement for the cluster analysis of large (and very large) phylogenetic trees. A previous study by Archer at al. [23] described a rigorous statistical methodology, but is practically unfeasible due to its high computational complexity. The studies by Lewis, Hughes et al. [2, 18] used a simpler partitioning method based on the linkage of sequences with a genetic distance below a certain threshold: the authors verified that the partition did not change significantly by varying the threshold and then confirmed the clusters by looking at the phylogenetic tree. Although this approach might be faster and easier to implement, even for the pre-selection of sequences to be included in the analysis, needs a fine tuning of the distance threshold (by considering the whole distance distribution) and a further analysis on the phylogenetic tree.

We also used multivariable logistic regression analysis to identify factors that are associated with clustering.

This work relied on the ARCA repository which lacks complete demographic and clinical information on patients, especially when considering seroconversion date and ART

status (known for 3% and 31% of the population, respectively). In fact, the estimation of infection transmission from ART-experienced patients was subject to a large uncertainty, and even higher was the confidence interval in the proportion of transmitted and onward-transmitted drug resistance. Another problem is the non-uniform sequence sampling in the population: although current guidelines recommend an HIV genotype resistance testing before starting any HAART and at each virological failure, in our study population the proportion of sequences from ART-naive patients was 8.8%.

Why is this of benefit to data enhancement in DynaNets

The novel method that we outlined in this section addresses the first need we defined in the introduction section. In summary this was the need for a good method that assesses clusters in large samples of phylogenetically related sequences.

Use of this method in DynaNets

ARCA was used in this study for development of the phylogenetic method. Analysis of the clusters showed that the ARCA database has limitations for identifying epidemiologically relevant clusters. We will therefore repeat the analysis using data from the SPREAD network. SPREAD includes patients newly diagnosed with HIV-1 that are representative for the risk group and geographical distribution of HIV-1 across Europe. Importantly, SPREAD also contains a large number of sequences (>3,000) and the method proposed in this deliverable is therefore helpful. We will also repeat the analysis using sequences collected among patients newly diagnosed with HIV-1 through MSM-contacts in the ErasmusMC in Rotterdam, the Netherlands. An advantage of the data collected in Rotterdam is that the ErasmusMC is the only hospital treating HIV-1 infected individuals in the South Western region of the Netherlands. The catchment area includes more than one million individuals. Data obtained at ErasmusMC are therefore a dense sample. We will present these results in deliverable 2.3 which is due at month 24.

In summary, using the tool described in this section will allow us to identify epidemiologically meaningful clusters in other large datasets. Information that can be retrieved and that is meaningful for DynaNets includes:

1. Level of onward transmission of drug resistant HIV
2. Impact of treated population on transmission of drug resistant HIV
3. Cluster size
4. Geographical spread of different HIV variants

2.4 Missing data in human contact experiment

2.4.1 Introduction

One of the objectives of deliverable 2.2 was to develop tools for handling (partially) missing data. In this section we will describe how missing data were handled in an experiment on human contacts. This experiment has been described more extensively in wp 3. We decided that describing this part of the experiment here and not in wp3 as it deals with handling of missing data which is one of the objectives of deliverable 2.2

2.4.2 Missing data

The data campaign of the SG in Dublin lasted about three months and recorded the interactions of roughly 10000 visitors. The sheer magnitude, both timewise and numberwise, of the campaign inevitably led to some amount of noise in the collected data. Among the possible reasons for measurement errors, one can list hardware or battery failure or tag mishandling by the science gallery personnel devoted to collecting tags worn by visitors leaving the science gallery before redistributing them to incoming visitors. The first step in the data analysis consists in post-processing the raw data without applying any filter. All the relevant quantities of interest are therefore calculated on the original dataset. As a result, one is then able to look for outliers and/or look for RFID devices showing malfunctioning patterns possibly along several days. The most frequently encountered experimental errors are tags reporting contacts outside museum opening hours, tags reporting an unusually high number of simultaneous contacts, tags recording extremely long visit durations and finally tags recording extremely short contact durations. Science gallery personnel forgetting to switch off tags collected from visitors leaving the science gallery is the most plausible cause of the first three experimental errors listed above, whereas a high number of very short visits is typically an example of erratic behaviour of a tag undergoing multiple reboots due to a low battery level. In particular, we look for associations between recurring suspicious patterns and specific tags. This leads to flagging all the data produced by tags falling into the categories described above. The second step produces a dataset, hereafter referred to as cleaned dataset, in which all the data flagged as suspicious by the previous analysis are weeded out and the quantities of interest are again recalculated. As a third step, we try whenever possible to reconstruct corrupted data by e.g. joining together contiguous short visits registered by the same tag and/or getting rid only of the data collected outside museum

working hours and we finally recalculate the quantities of interest on this dataset hereafter referred to as partially recovered dataset. The results obtained by the post processing of the original dataset, the cleaned dataset and the partially recovered dataset are then compared in order to test the impact of noisy data on the quality of the statistical analysis. Further to that, we perform stress tests by varying the choice of the filtering parameters and in some cases by removing artificially some data. In the case of the data campaign at the Science Gallery, the collected data proved rather robust and even a removal of about 30% of the recorded interactions did not qualitatively alter the conclusions about the contact statistics nor the aggregated network analysis.

3 Conclusions

In this deliverable we presented several tools that were developed for data enhancement. We focused on phylogenetic analysis as such analysis can help us to identify transmission networks. One of the tools is currently used to analyze datasets available to DynaNets. We will report on these results in deliverable 2.3.

References

1. Vercauteren J, Wensing AM, van de Vijver DA, et al. *Transmission of drug-resistant HIV-1 is stabilizing in Europe*. J Infect Dis 2009;200:1503-8

2. Hughes GJ, Fearnhill E, Dunn D, et al. *Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom*. PLoS Pathog 2009;5:e1000590

3. Frentz D, Boucher CA, Assel M, et al. *Comparison of HIV-1 genotypic resistance test interpretation systems in predicting virological outcomes over time.* PLoS One 2010;5:e11505

4. Spread-programme. *Transmission of drug-resistant HIV-1 in Europe remains limited to single classes.* Aids 2008;22:625-35

5. Bezemer D, van Sighem A, Lukashov VV, et al. *Transmission networks of HIV-1 among men having sex with men in the Netherlands.* Aids 2010;24:271-82

6. Brenner BG, Roger M, Moisi DD, et al. *Transmission networks of drug resistance acquired in primary/early stage HIV infection.* Aids 2008;22:2509-15

7. Brenner BG, Roger M, Routy JP, et al. *High rates of forward transmission events after acute/early HIV-1 infection.* J Infect Dis 2007;195:951-9
8. Yerly S, Vora S, Rizzardi P, et al. *Acute HIV infection: impact on the spread of HIV and transmission of drug resistance.* AIDS 2001;15:2287-2292
9. Hue S, Gifford RJ, Dunn D, Fernhill E and Pillay D. *Demonstration of sustained drug-resistant human immunodeficiency virus type 1 lineages circulating among treatment-naive individuals.* J Virol 2009;83:2645-54
10. Yerly S, Junier T, Gayet-Ageron A, et al. *The impact of transmission clusters on primary drug resistance in newly diagnosed HIV-1 infection.* Aids 2009;23:1415-23
11. Smith RJ, Okano JT, Kahn JS, Bodine EN and Blower S. *Evolutionary dynamics of complex networks of HIV drug-resistant strains: the case of San Francisco.* Science 2010;327:697-701
12. Paraskevis D, Pybus O, Magiorkinis G, et al. *Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach.* Retrovirology 2009;6:49
13. Worobey M, Gemmel M, Teuwen DE, et al. *Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960.* Nature 2008;455:661-4
14. Gilbert MT, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE and Worobey M. *The emergence of HIV/AIDS in the Americas and beyond.* Proc Natl Acad Sci U S A 2007;104:18566-70
15. Liljeros F, Edling CR, Amaral LA, Stanley HE and Aberg Y. *The web of human sexual contacts.* Nature 2001;411:907-8
16. Liljeros F, Edling CR and Nunes Amaral LA. *Sexual networks: implications for the transmission of sexually transmitted infections.* Microbes Infect 2003;5:189-96
17. Schneeberger A, Mercer CH, Gregson SA, et al. *Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in Britain and Zimbabwe.* Sex Transm.Dis. 2004;31:380-387
18. Lewis F, Hughes GJ, Rambaut A, Pozniak A and Leigh Brown AJ. *Episodic sexual transmission of HIV revealed by molecular phylodynamics.* PLoS Med 2008;5:e50
19. Hammer SM, Saag MS, Schechter M, et al. *Treatment for adult HIV infection: 2006 recommendations of the International AIDS Society-USA panel.* JAMA 2006;296:827-843
20. Volberding PA, Deeks SG. *Antiretroviral therapy and management of HIV infection.* Lancet 2010;376:49-62
21. Thompson MA, Aberg JA, Cahn P, et al. *Antiretroviral treatment of adult HIV infection: 2010 recommendations of the International AIDS Society-USA panel.* Jama 2010;304:321-33

22. Salemi M, Vandamme AM. *The phylogenetic handbook.* Cambridge: Cambridge university press, 2003

23. Archer J, Robertson DL. *CTree: comparison of clusters between phylogenetic trees made easy.* Bioinformatics 2007;23:2952-3

24. Qin ZS. *Clustering microarray gene expression data using weighted Chinese restaurant process.* Bioinformatics 2006;22:1988-97

25. Zagordi O, Geyrhofer L, Roth V and Beerenwinkel N. *Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction.* J Comput Biol 2010;17:417-28

26. Ferguson T. *Bayesian analysis of some nonparametric problems.* Ann Stat 1973;1:209-30

27. Rusmussen C. *The infinite Gaussian mixture model.* Advances in neural information processing systems 2000;12:554-560

28. Strehl A, Ghosh A. *Cluster ensembles - a knowledge reuse framework for combining multiple partitions.* J Mach Learn Res 2003;3:583-617

29. Kuncheva L, Vetrov D. *Evaluation of stability of k-means cluster ensembles with respect to random initialization.* IEEE transactions on pattern analysis and machine intelligence 2006;28:1798-1808

30. Hubert L, Arabie P. *Comparing partions.* Journal of Classification 1985;2:193-218

31. Kaufman L, Rousseeuw PJ. *Finding groups in data: an introduction to cluster analysis.* New York: Wiley, 1990

32. Thompson JD, Higgins DG and Gibson TJ. *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res 1994;22:4673-80

33. Johnson VA, Brun-Vezinet F, Clotet B, et al. *Update of the Drug Resistance Mutations in HIV-1: December 2009.* Top HIV med 2009;17:138-145

34. Gotoh O. *An improved algorithm for matching biological sequences.* J Mol Biol 1982;162:705-8

35. Shankarappa R, Margolick JB, Gange SJ, et al. *Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection.* J Virol 1999;73:10489-502

36. Drummond AJ, Rambaut A. BEAST: *Bayesian evolutionary analysis by sampling trees.* BMC Evol Biol 2007;7:214

37. Price MN, Dehal PS and Arkin AP. *FastTree 2--approximately maximum-likelihood trees for large alignments.* PLoS One 2010;5:e9490

38. Brown AE, Gifford RJ, Clewley JP, et al. *Phylogenetic reconstruction of transmission events from individuals with acute HIV infection: toward more-rigorous epidemiological definitions.* J Infect Dis 2009;199:427-31

39. Hue S, Clewley JP, Cane PA and Pillay D. *HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy.* Aids 2004;18:719-28

40. Drumright LN, Frost SD. *Sexual networks and the transmission of drug-resistant HIV.* Curr Opin Infect Dis 2008;21:644-52

41. Kouyos RD, von Wyl V, Yerly S, et al. *Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland.* J Infect Dis 2010;201:1488-97

42. Maserati R, De Silvestri A, Uglietti A, et al. *Emerging mutations at virological failure of HAART combinations containing tenofovir and lamivudine or emtricitabine.* Aids 2010

43. de Oliveira T, Deforche K, Cassol S, et al. *An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics.* 2005;21:3797-3800

44. Johnson VA, Brun-Vezinet F, Clotet B, et al. *Update of the Drug Resistance Mutations in HIV-1: Spring 2008.* Top HIV Med 2008;16:62-8

45. Armitage P, Berry G and Matthews JNS. *Statistical methods in medical research. Oxford: Blackwell Science, 2002*

46. van de Vijver DAMC, Wensing AMJ, Boucher CAB, et al. *The epidemiology of transmission of drug resistant HIV-1. HIV Sequence Compendium 2006/2007: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.* LA-UR 07-4826, 2007:17-36