Modelling HIV resistance development under the selective pressure of rare drug combinations

Author(s): Perraudin, Caroline

Publication Date: 2015

Permanent Link: https://doi.org/10.3929/ethz-a-010569979

Rights / License: In Copyright - Non-Commercial Use Permitted
Modelling HIV resistance development under the selective pressure of rare drug combinations

Submitted to the Computational Biology Group for the Master’s program in Biotechnology at the Department of Biosystems Science and Engineering ETH Zürich

March - November 2015

Author: Caroline Perraudin (MSc D-BSSE ETHZ)
Supervisor: Hesam Montazeri (PhD D-BSSE ETHZ)
Examiner: Niko Beerenwinkel (Prof D-BSSE ETHZ)
# Table of Contents

1 Motivation ................................................................. 3

2 Introduction .............................................................. 4
   2.1 Susceptibility to HIV ............................................. 4
   2.2 Human immunodeficiency virus - Life cycle ................. 4
   2.3 Antiretroviral therapies ......................................... 6
   2.4 Resistance mutations ............................................. 6
   2.5 Aims of the project ............................................... 7

3 Methods ................................................................. 9
   3.1 Continuous Time Conjunctive Bayesian Networks (CT-CBNs) ... 9
   3.2 Mixture model implementation using existing CBN algorithms ... 10
      3.2.1 New likelihood function ................................... 10
      3.2.2 Parameter inference using a Monte Carlo Expectation Maximisation (MCEM) ... 10
         3.2.2.1 E-step .................................................. 10
         3.2.2.2 M-step .................................................. 11
      3.2.3 Accounting for noise in clinical data ..................... 12
      3.2.4 Package and software used .................................. 12

4 Results .............................................................. 13
   4.1 Research question ............................................... 13
   4.2 Validation of the model extension ............................. 14
      4.2.1 Reconstruction of the true underlying networks ........ 14
      4.2.2 Accuracy of rate optimisation ............................. 15
   4.3 Simulation study .................................................. 17
      4.3.1 Experimental setup ......................................... 17
      4.3.2 Weight computation: first approach ....................... 18
         4.3.2.1 Assessment of the performance of the different algorithms ... 19
      4.3.3 Weight computation: second approach ..................... 19
         4.3.3.1 Assessment of the performance of the different algorithms ... 21
         4.3.3.2 Analysis of the effect of different noise levels ... 22
   4.4 Analysis of clinical data from the Swiss HIV Cohort Study (SHCS) database ... 23
      4.4.1 Experimental setup ......................................... 24
      4.4.2 Weight computation: first approach ....................... 25
         4.4.2.1 Assessment of the performance of the different algorithms ... 25
      4.4.3 Weight computation: second approach ..................... 25
         4.4.3.1 Assessment of the performance of the different algorithms ... 26
      4.4.4 Weight computation: third approach ....................... 26
         4.4.4.1 Assessment of the performance of the different algorithms ... 27
      4.4.5 Weight computation: fourth approach ..................... 27
      4.4.6 Analysis of RT dataset ..................................... 28
         4.4.6.1 Assessment of the performance of the different algorithms on the RT dataset ... 30
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4.6.2</td>
<td>Comparison of the number of edges detected in the three approaches using the RT dataset</td>
<td>31</td>
</tr>
<tr>
<td>4.4.6.3</td>
<td>Examples of mutational pathways as a consequence of RT drugs</td>
<td>32</td>
</tr>
<tr>
<td>4.4.7</td>
<td>Analysis of PR dataset</td>
<td>35</td>
</tr>
<tr>
<td>4.4.7.1</td>
<td>Assessment of the performance of the different algorithms on the PR dataset</td>
<td>36</td>
</tr>
<tr>
<td>4.4.7.2</td>
<td>Comparison of the number of edges detected in the three approaches using the PR dataset</td>
<td>37</td>
</tr>
<tr>
<td>4.4.7.3</td>
<td>Examples of mutational pathways as a consequence of PR drugs</td>
<td>38</td>
</tr>
<tr>
<td>5</td>
<td>Discussion</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>Outlook</td>
<td>42</td>
</tr>
<tr>
<td>7</td>
<td>Acknowledgements</td>
<td>42</td>
</tr>
<tr>
<td>8</td>
<td>References</td>
<td>43</td>
</tr>
<tr>
<td>9</td>
<td>Additional Material</td>
<td>45</td>
</tr>
<tr>
<td>9.1</td>
<td>Summary plots from the different unsuccessful approaches</td>
<td>45</td>
</tr>
<tr>
<td>9.2</td>
<td>Comparison of the number of edges detected in the three approaches using the RT dataset using the separate jaccard index</td>
<td>46</td>
</tr>
<tr>
<td>9.3</td>
<td>Comparison of the number of edges detected in the three approaches using the PR dataset using the separate jaccard index</td>
<td>46</td>
</tr>
<tr>
<td>9.4</td>
<td>Names of all the treatments used for network learning</td>
<td>47</td>
</tr>
</tbody>
</table>
1 Motivation

The human immunodeficiency virus (HIV) causes the acquired immunodeficiency syndrome (AIDS), a progressive and life-threatening failure of the immune system, if left untreated. Since the first discovery of the virus in 1981 almost 78 million infections have been reported and about 39 million individuals died due to this disease. At the end of 2013, the World Health Organisation (WHO) reported a total of 35 million people that were living with HIV. Fig. 1 shows the global prevalence of HIV in adults in 2013 [7].

![HIV Prevalence Map]

**Figure 1: HIV Prevalence.** Percentage of HIV infected adults (15 - 49 years) in 2013 by WHO region; Western pacific: 0.1%, Eastern Mediterranean: 0.1%, South East Asia: 0.3%, Europe: 0.4%, Americas: 0.5%, Africa: 4.5%, Globally: 0.8% [7].

Current medical treatments show very good successes in lowering the viral load in infected patients. For this purpose antiretroviral therapies (ARTs) consisting of different combinations of a large number of available drugs are used. Mostly, a combination of three different drugs is used to treat the patients efficiently. In general, the different drugs are chosen from at least two different drug classes, such that there is an impact at different stages in the viral lifecycle [17]. Finding the optimal combination of drugs for HIV treatment remains a challenging and highly personalised task. Factors that make the choice of the drug combination complicated include drug interactions and toxicities, adherence to therapy and development of drug resistance. The combinatorial complexity of the different drugs and viral mutational patterns indicate the utility of statistical models [14]. For this purpose, conjunctive Bayesian networks (CBNs) have been used to help find the optimal drug combinations based on the actual genotype of individual patients. The individualised genetic barrier (IGB) has recently been suggested as a useful predictor of treatment outcome [12]. The IGB has been defined as the probability that a virus will not become resistant to a certain drug; a high IGB means that the virus is unlikely to become resistant to the
drug. This approach has been applied to data from patients that were under two of the frequent treatments, namely zidovudine plus lamivudine and nelfinavir. However, it has not yet been extended to all the possible drug combinations due to a lack of suitable data.

My project specifically focuses on the further development of the existing CBN models to allow treatment outcome predictions for any desired drug combination. In an attempt to model the temporal order of resistance mutation accumulation for any drug combination, a mixture model will be implemented using the already existing model. Then, the whole data can be used to reconstruct the different mutational pathways. This should yield a better network prediction and a more reliable estimation of the parameters involved. Eventually, it should allow predictions for any drug combination [15]. The extended model will be tested with simulated data and its performance will be compared with previous approaches. Finally, the extended model will be applied to real data from the Swiss HIV Cohort Study (SHCS) database.

2 Introduction

2.1 Susceptibility to HIV

Among infectious killers, HIV is the most deadly. According to WHO an estimated 39 million people have died so far since the first cases in 1981. In 2013, 1.5 million deaths were due to AIDS-related causes [7]. Another big issue is HIV infections in children. In the sub-Saharan Africa, most infections with HIV in children happen during pregnancy, childbirth or breastfeeding. Treating the mothers with appropriate ARTs would decrease the viral load sufficiently such that children could be born without infections.

The first signs of an HIV infection are, in most cases, flu-like symptoms 2-4 weeks after the exposure to the virus, the so-called acute retro viral syndrome (ARS) or primary infection. The symptoms may include fever, swollen glands, sore throat, headache, muscle and joint aches and pains. If left untreated the disease will progress to the next stage called the clinical latency stage. During this period, the virus populates the host and this produces little to no symptoms. Without ART, this period lasts an average of 10 years. Eventually, this latency period is over and the virus has a symptomatic impact on the immune system. When an HIV infection gets symptomatic, the transition to the last stage has happened, the stage of AIDS. Some symptoms associated with AIDS are rapid weight loss, recurring fever, diarrhea for more than a week, sores of the mouth, anus or genitals. However, the only way for a reliable diagnosis is to get tested in an official test center as the symptoms mentioned above can also be related to many other infectious illnesses [7].

2.2 Human immunodeficiency virus - Life cycle

HIV is a lentivirus, a subgroup of the retroviruses, and is depicted in Fig. 2.
Introduction

Figure 2: Human immunodeficiency virus. Schematic representation of HIV with all the involved proteins labeled accordingly [5].

HIV enters the body through sexual contact, transfusion with infected blood or when using needles for injections that are contaminated. The virus seems to attach to immune cells, namely to dendritic cells which are found in mucosal membranes in the mouth, vagina, rectum, penis and in the upper gastrointestinal tract. Dendritic cells transport the virus from the site of infection to the lymph nodes where infections of other immune cells can occur. The main targets are CD4 lymphocytes, also called T-cells, where viral replication takes place [31, 18].

An infection starts with the binding of HIV to a specific receptor and co-receptor on the surface of the CD4 cell followed by a fusion process where the viral genetic material is released into the target cell. A schematic representation of the viral genome can be seen in Fig. 7. After infection, the reverse transcriptase (RT) converts the single-stranded viral RNA to double-stranded DNA. The newly synthesised DNA enters the nucleus and is integrated into the host DNA by the integrase. It may stay there, inactive, for several years (clinical latency). Once the host cell becomes activated, the virus uses the enzymes of its host to replicate the viral genetic material and to create longer proteins. The protease then cuts the longer HIV proteins into the individual, functional proteins. The processed proteins and the replicated viral genetic material together can then form a new virus. The newly replicated virus can be released by budding and can then infect other host cells [3].

Figure 3: HIV Genome. Schematic representation of the full genome of HIV-1 including the precursor proteins as well as the processed, functional proteins [2].
2.3 Antiretroviral therapies

The treatments in use, ARTs, consist of a combination of drugs that interfere with the viral life cycle at different stages. The various drugs are classified into six major drug classes based on their function and are shortly described in the following [1, 7].

1. **Entry inhibitors**
   Prevent the virus from binding to the receptors on the human target cell, hence HIV cannot infect the cell.

2. **Fusion inhibitors**
   Interfere with the fusion process of the virus with the cellular membrane of target cells, hence HIV cannot infect the cell.

3. **Reverse transcriptase inhibitors**
   The enzyme for the reverse transcription, a process that converts single-stranded HIV RNA into double-stranded HIV DNA, is inhibited.
   - (a) **nucleoside/nucleotide RT inhibitors (NRTIs)**
     These inhibitors act as faulty DNA building blocks. Integration of one of these leads to a stop in HIV DNA synthesis.
   - (b) **non-nucleoside RT inhibitors (NNRTIs)**
     Bind to the enzyme RT directly and inhibit the HIV DNA synthesis.

4. **Integrase inhibitors**
   Inhibit the enzyme integrase, which integrates viral genetic material into the DNA of the infected cell. Therefore, the viral genetic material will not be transcribed and is just floating around in the cell.

5. **Protease inhibitors (PIs)**
   Inhibit the enzyme protease, which cuts long chains of HIV proteins into smaller functional proteins. Inhibiting the protease’s function leads to unfunctional proteins, hence the virus cannot assemble new particles.

6. **Multi-class combination products**
   Combine the effects of drugs from at least two classes into one single product.

2.4 Resistance mutations

The intensive use of ARTs to fight HIV has increased the immune pressure, which has led to a considerable number of mutations that make the virus resistant to the respective drugs [27, 30]. Fig. 4 shows the known resistance mutations for the two main drug classes. A selection of these are also the mutations that will be used to generate the mutational pathways later on.

<table>
<thead>
<tr>
<th>Major Non-Nucleoside RT Inhibitor (NNRTI) Resistance Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>Cons</td>
</tr>
<tr>
<td>NVP</td>
</tr>
<tr>
<td>EFV</td>
</tr>
<tr>
<td>ETR</td>
</tr>
<tr>
<td>RPV</td>
</tr>
</tbody>
</table>

Master Thesis 6 Caroline Perraudin
Figure 4: Resistance mutations. Extract from the Stanford resistance database showing the mutations in two different proteins that lead to resistance against the respective drugs (leftmost column), last update: March 2, 2014 [6]

2.5 Aims of the project

Despite huge advances in combination therapies, drug resistance is still one of the major factors contributing to HIV-1 treatment failure [23]. Drug-specific escape mutations are accumulated in the viral genome during the treatment. This eventually leads to a viral genotype that is completely resistant to the drug in use. Statistical models based on CBNs have been developed in an attempt to model the mutational pathways that a virus can take from the wild type to a fully resistant virus. These models allow the computation of time to complete viral resistance for a particular treatment combination, given the individual genotype of a patient [10, 13]. Due to data sparseness the models are not yet capable of predicting networks for any desired drug combination. Fig. 5 illustrates the data sparseness that we are confronted with for the majority of the drug combinations. The histogram shows the number of observations of the 75 most abundant treatments. For the remaining 478 treatments, for which the SHCS database contains data, even less observations are available. This makes it impossible for current algorithms to learn networks for all treatments.
The aim of this master thesis is to further develop existing CBN models in order to fight the data sparseness. Until now, data from one treatment only was used to learn a CBN model for the respective treatment. The idea here is to use a mixture model to be able to use the whole data set to learn each of the CBN models. The contribution of the different drug combinations will be expressed by weights that reflect their treatment similarity. This should eventually allow statistically relevant statements for any drug combination. In the course of this thesis, the current CBN models will be extended using a mixture model. On the theoretical side, I will learn how to extend an existing model and show that this extension is valid. Simulated data will initially be used to assess the extended model. On the application side, the new model will be used on clinical data from the SHCS database and the performance will be characterised. I will use various statistical tools to characterise the performance of the extended models on simulated and real data and compare them with current approaches.

Specific aims:

- Extension of the current CBN model using the theory for mixture models
- Validation of the model extension using simulated data
- Analysis of the accuracy of predictions of the algorithm using simulated data
- Assessment of the performance of the new model in comparison with previous approaches
- Real data analysis with clinical data from the SHCS database
3 Methods

3.1 Continuous Time Conjunctive Bayesian Networks (CT-CBNs)

CBNs [19] are graphical models which constrain events in their order of occurrence. They can be used to model the order of events that accumulate over time, such as resistance mutations in an organism. CBNs are a generalisation of oncogenetic tree models with the big advantage that they allow that one mutational event depends on more than one predecessor event, whereas in oncogenetic trees one event can only depend on one predecessor event [10]. Specifically, the aim is to reconstruct the order in which drug-induced resistance mutations in the HIV genome occur. For this purpose, we use a binary representation of genotypes, indicating which mutations are present, and the corresponding sampling times. Fig. 6 shows one example of a CBN:

![Diagram of Continuous Time Conjunctive Bayesian Networks (CT-CBNs)](image)

**Figure 6: Example.** A CT-CBN [13] consisting of four events and some temporal ordering restrictions

To give an example, event 3 in the poset $P$, can only happen once events 1 and 2 have happened. The waiting times in the CT-CBNs can be modelled with exponential random variables [13]. More formally, for each event $i$, we define an exponentially distributed random variable $Z_i \sim \text{Exp}(\lambda_i)$ and a random variable $T_i$ as:

$$T_i = \max_{j \in \text{pa}(i)} T_j + Z_i$$

where $\text{pa}(i)$ represents all the predecessor events of event $i$ in the poset $P$. In the case of mutational events, the parameter of the exponentials, $\lambda_i$, is the respective mutation rate. Accordingly, the waiting time for event 3 consists of the maximum time required for events 1 and 2 to happen plus the time until event 3 itself happens. The distributive lattice, $J(P)$, shows the set of all genotypes that are compatible with the poset $P$.

For a fixed poset, the likelihood function (Eq. 2) for the described model and the corresponding log-likelihood (Eq. 3) can be written down as follows:

$$f_\lambda(t) = \prod_{i=1}^{P} \lambda_i \exp(-\lambda_i (t_i - \max_{j \in \text{pa}(i)} t_j)), \quad \text{if } t_i > \max_{j \in \text{pa}(i)} t_j \text{ for all } i \in [P]$$

$$l(\lambda) = \sum_{k=1}^{N} \sum_{i=1}^{P} \left( \log \lambda_i - \lambda_i (t_i^{(k)} - \max_{j \in \text{pa}(i)} t_j^{(k)}) \right)$$

Master Thesis 9 Caroline Perraudin
where $P$ denotes the number of mutations included in the CBN and $N$ is the number of observed genotypes. The rates are estimated using a Monte Carlo expectation maximisation (MCEM) algorithm [25].

### 3.2 Mixture model implementation using existing CBN algorithms

To date, all the observations had to stem from patients that followed the same regimen. This is the major drawback of existing CBN models, because, as shown in Fig. 5, for most treatments a very small number of observations is available - for many there are as few as one or two observations only. Using the CT-CBN model with so little data will not lead to a satisfactory result. This problem was tackled in the course of this project with the attempt to use a mixture model approach such that all the data could be used as input for the network learning and rate estimation.

#### 3.2.1 New likelihood function

Inference of the CT-CBN model using an MCEM algorithm has previously been implemented in the MC-CBN tool [4]. This tool was used as starting point for the new approach. The aim of this thesis was to allow the model to consider data from patients that followed different treatments to deal with the prevalent data sparseness in clinical datasets. To reach this goal, the contributions from patients that followed different treatments had to be weighted with a number that reflects the similarity between the applied ARTs. For this purpose, we used a mixture model approach. We obtained the following expected hidden log-likelihood function:

$$
I^*(\lambda) = \sum_{k=1}^{N} \gamma^{(k)} \cdot \log(\beta) + \sum_{i=1}^{P} \sum_{k=1}^{N} \gamma^{(k)} \cdot \left( \log \lambda_i - \log \left( \frac{t_i}{\max_{j \in \text{pa}(i)} t^{(j)}_j} \right) \right)
$$

where $\gamma^{(k)}$ is the responsibility of the respective treatment for the observed genotype $k$. The responsibilities are fixed, unlike in the mixture model theory, where these are a function of the optimised parameters. The parameter $\beta$ is the fraction of cases in the data that followed the treatment for which the network learning is carried out. In our case, the first term does not depend on the parameters or the structure of the network. The second term computes the sum of the weighted probability of each observation given the network structure and the mutation rates.

#### 3.2.2 Parameter inference using a Monte Carlo Expectation Maximisation (MCEM)

The sufficient statistics for the estimation of $\lambda_i$ is $t_i - \max_{j \in \text{pa}(i)} t^{(j)}_j$. However, we only know the sampling time ($t_s$), not the occurrence time ($t$) of the mutations. Therefore, the expectation of the occurrence time of the mutations has to be computed, such that the optimisation can be carried out. The estimation of the mutation rates is carried out using an MCEM algorithm [22]. The algorithm was adapted from [25] for the new model and is described below.

#### 3.2.2.1 E-step

As outlined in [25], in the E-step of the MCEM algorithm, the expected time differences are computed for all mutations, given the observation (genotype $(g)$, sampling time $(t_s)$)
and the ML estimate of the parameters from the previous iteration. This expectation is defined as \( e_i(g,t_s) = E[T_i - \max_{j \in \text{pa}(i)} T_j | g, t_s, \lambda, P] \) and is approximated using Monte Carlo integration by importance sampling. From a proposal distribution \( q(t) \), \( L \) samples are drawn and \( e_i(g,t_s) \) is computed as:

\[
e_i(g,t_s) \approx \frac{1}{L} \sum_{k=1}^{L} \frac{f_\lambda(t^{(k)})}{q(t^{(k)})} (t_i^{(k)} - \max_{j \in \text{pa}(i)} t_j^{(k)})
\]

\[= \frac{\sum_k w_k (t_i^{(k)} - \max_{j \in \text{pa}(i)} t_j^{(k)})}{\sum_k w_k}
\]

(5)

The fraction \( \frac{f_\lambda(t^{(k)})}{q(t^{(k)})} \) is called importance weights, where \( f_\lambda(t^{(k)}) \) denotes the CBN density (Eq. 2) and \( q(t^{(k)}) \) the proposal distribution (Eq. 6). In the case of \( L=1 \) the MCEM is called the stochastic EM. The proposal distribution has a huge impact on the efficiency and accuracy of the estimation. We use the following proposal distribution and the equation \( T_i = \max_{j \in \text{pa}(i)} T_j + Z_i \) to generate the occurrence times \( t_i \) of the mutations, given the genotype \( g \), and its sampling time \( t_s \):

\[
Z_i \sim \text{TExp}(\lambda_i, 0, t_s - \max_{j \in \text{pa}(i)} t_j), \quad \text{for } i \in g^{(k)}
\]

\[
Z_i \sim \text{Exp}(\lambda_i), \quad \text{otherwise}
\]

(6)

where \( \text{TExp}(\lambda, a, b) \) is an exponential, truncated to the interval \( a \) and \( b \).

### 3.2.2.2 M-step

In the M-step, the new estimate of the rate is computed. Differentiating Eq. 4 with respect to \( \lambda_i \) yields the following equation:

\[
\sum_{k=1}^{N} \frac{\gamma^{(k)}}{\lambda_i} - \sum_{k=1}^{N} \gamma^{(k)} \left( t_i^{(k)} - \max_{j \in \text{pa}(i)} t_j^{(k)} \right) = 0
\]

(7)

Solving above equation for \( \lambda_i \), we obtain the new estimate of the mutation rates:

\[
\hat{\lambda}_{i,\text{new}} = \frac{\sum_{k=1}^{N} \gamma^{(k)} (t_i^{(k)} - \max_{j \in \text{pa}(i)} t_j^{(k)})}{\sum_{k=1}^{N} \gamma^{(k)}}
\]

(8)

It has been shown that an improvement in the Markov chains can be achieved by averaging over the last iterations \([26]\). The maximum number of iterations, \( \text{maxIter} \), can be defined and the averaging happens over the last \( \zeta \cdot \text{maxIter} \) iterations, where \( \zeta \in (0, 1] \).
3.2.3 Accounting for noise in clinical data

In clinical datasets, however, observations are subject to noise. Since the ML poset is the largest poset compatible with all the observations, when applying it to clinical data the ML posets will be very sparse, even if most of the data would suggest a denser graph. This problem was previously addressed and is solved as outlined in [13].

Let $P_\delta$ be the ML poset obtained, allowing that at most a fraction of $\delta$ of all genotypes violate the relations in $P_\delta$. It is assumed that the incompatible genotypes are generated with a uniform probability $q_\delta = 1/(2^P - |J(P_\delta)|)$, where $P$ is the number of mutations. The probability of an observation $(g^{(k)}, t_s^{(k)})$ for a given regimen is then defined as:

$$
\text{Prob}(g^{(k)}, t_s^{(k)}|\lambda, \alpha) = \begin{cases} 
\alpha \cdot \text{Prob}(g^{(k)}, t_s^{(k)}|\lambda) & \text{if } g^{(k)} \in J(P_\delta), \\
(1 - \alpha) \cdot q_\delta & \text{otherwise}
\end{cases}
$$

The mixing parameter $\alpha$ is computed as the weighted fraction of genotypes that are compatible with the poset $P_\delta$ and is defined as follows:

$$
\hat{\alpha} = \frac{\sum_{g^{(k)} \in J(P_\delta)} \gamma^{(k)}}{\sum_{k=1}^{N} \gamma^{(k)}}
$$

The numerator is the sum of the observed weights of the compatible genotypes, while the denominator is the sum of the weights of all genotypes irrespective of their compatibility. This can be interpreted as a weighted frequency of compatible genotypes in the data. This additional mixture model allows to account for a certain level of noise in the data. The final output is the ML poset and the corresponding ML rates. It has been shown that the ML poset is the densest graph that is compatible with all the observations [13].

3.2.4 Package and software used

The existing package MC-CBN [4] was used. The further development of the model to deal with data sparseness was implemented in R version 3.2.1.
4 Results

4.1 Research question

The existing CT-CBN model was extended by the implementation of a mixture model. This allowed using data from patients that followed different regimens as input to the model. The required weights were computed such that the contributions from different regimens were weighted according to the similarities of the prescribed treatments. So, the whole dataset could be used to learn networks for all the regimens, even for the very rare ones. The following analyses investigate the performance of the new implementation. The specific question we would like to answer, is:

How does the new model perform in comparison with previous implementations that do not use the mixture model approach?

In order to answer this question, the following three approaches were used for the network reconstruction:

- **One-for-all**
  The whole dataset containing data from various regimens was used as input. All the weights used were identical, leading to no differentiation of the contribution from the different observations. One "consensus" poset is the result of this approach.

- **Separate**
  This approach uses only the observations from patients that followed the same regimen. The number of observations that can be used ranges from 255, for the most frequent treatment, to 1 for various rare treatments. Technically, this can be achieved by using the identity matrix as similarity matrix as input to the new implementation. This then assigns a weight of one to observations from the regimen itself, while it assigns a weight of zero to the remaining observations and restores the previous implementation.

- **Weighted**
  The whole dataset containing data from various regimens was used as input. Similarity scores between the different treatments were used as weights. The idea is that in comparison with the one-for-all approach, here, the noise can be decreased and in comparison with the separate approach, the sample size can be increased.

The following two testing values were computed, in order to compare the performance of the three above mentioned approaches for network reconstruction:

- **Hamming distance**
  The Hamming distance between the observed genotype and the maximum a posteriori (MAP) estimate of the genotype for the observed sampling time was computed. This served as an indicator for the genotype prediction accuracy using the resulting ML network and rates.

- **Likelihood of testdata**
  Another strong indicator that was computed, is the likelihood of the testdata given the ML network and the corresponding rates.
4.2 Validation of the model extension

4.2.1 Reconstruction of the true underlying networks

In order to find out whether the CT-CBN model in general can reconstruct the true network, the following experiment was carried out: three posets consisting of five mutations with corresponding mutation rates were randomly generated (Fig. 7). From each of these, 200 observed genotypes and sampling times were generated [29].

**Figure 7:** Networks used for data generation. From each of the networks 200 observed genotypes and sampling times were generated without addition of noise.

Fig. 8 gives an impression of what the simulated database looked like. Fig. 8a shows the number of patients that had zero to five mutations in their simulated genotype and Fig. 8b shows the frequency of the different mutations in the simulated database. The whole database contained 600 observations.

**Figure 8:** Simulated database. Investigation of the data before using it for the network reconstruction.

The barplots show that all the mutations appeared and that there were patients with zero to five mutations, corresponding to wildtype to full resistance. This simulated database was then used as input for the algorithm. The separate approach was used. Given a correct implementation of the new algorithm, the networks should be reconstructed correctly, because, the separate approach restores the previous algorithm using the new implementation. With a sample size of 200 observations, the previously implemented algorithm was able to reconstruct the networks correctly. The aim was to rule out technical issues in the implementation. Fig. 9 shows the real and the reconstructed ML...
networks. The first row shows the real networks that were used for the data simulation and the second row displays the networks obtained when applying the new CT-CBN model using the separate approach. This was repeated ten times with different networks. The results shown here are one of the ten runs. The results from all the other repetitions were in line with the results presented here.

![Diagram of real and reconstructed posets](image)

**Figure 9: Successful network reconstruction. One example**

In conclusion, with these settings the new implementation of the algorithm succeeded in reconstructing the true networks. This was a quick validation check for the implementation of the new model. It showed that the implementation was fine. Unfortunately, when it comes to data from a clinical database, it is impossible to get as many observations for each treatment. Furthermore, clinical samples are subject to noise, while here, the genotypes were generated without noise. This explains the need of the weighted approach.

### 4.2.2 Accuracy of rate optimisation

It could be shown that the reconstruction of the true posets was possible with the implemented model under optimal conditions, where a decent number of noise-free observations was available. In a second step, the accuracy of the rate prediction was assessed. This was carried out using the three different approaches. The weights for the weighted approach were computed as the Jaccard indexes among the transitively-closed adjacency matrices, representing the real networks (for further details see 4.3.3). In order to assess the rate prediction accuracy, ten networks were generated (Fig. 10). The simulated dataset consisted of 200 noise-free observations from each of the true networks. This dataset was then used for the rate optimisation process. It was assumed that the true network had been reconstructed. The aim was to analyse the accuracy of the rate prediction in an ideal case where the true network had been reconstructed and a decent amount of observations was available. In order to get a more robust estimate, the simulation of the dataset and the rate optimisation was repeated thirty times with the same ten networks. The mean squared errors (MSE) between the real and the ML rates for the three approaches are shown in Fig. 10. Only three of the ten examples are showed here. The results for the seven remaining are in line with the results presented here.
The rate optimisation, given the true network, was carried out for the three different approaches: weighted, one-for-all, separate. It can be seen that the MSE for the weighted approach is very close to zero in most cases. Using the other two approaches, some rate predictions show big deviations from the true rates. The poor performance of the rate prediction can in most cases be explained by the fact that the corresponding mutations happen at the very end of the network. These mutations occur much less often, so in the separate approach, less data is available for the rate prediction in comparison with the weighted approach, where the sample size is much larger. Moreover, the errors from the rate prediction from earlier mutations accumulate and lead to much more variation in the rate prediction for later mutations.

In conclusion, using a bigger sample size, as in the weighted approach, definitely increases the rate prediction accuracy in comparison with the separate approach. However, further increasing the sample size by using all the observations as input, as in the
one-for-all approach, increases the noise in the data and leads to a much less accurate rate prediction. The trade-off we are confronted with is: a big sample size or low noise levels. Finding the optimal balance between these two, by defining appropriate weights, is the main issue in the new model.

4.3 Simulation study

For the simulation study, data was simulated to mimic the clinical data such that the performance of the new tool could systematically be analysed. Samples were simulated from a varying number of randomly generated networks containing 10 mutations. A different treatment was assigned to each network and so, to the observations generated from these networks. In order to test the resulting ML networks, 50 observations were generated from each network. The three different approaches were used for the network learning and rate optimisation (weighted, one-for-all, separate) and their performance was characterised using the log-likelihood of the test data and the Hamming distance between observed genotype and its MAP estimate for the observed sampling time.

4.3.1 Experimental setup

A different number of observations (N = 2, 5, 10, 25, 50, 100, 200) was generated from each of the networks to simulate the real situation. The number of networks varied, such that the total number of observations generated remained constant, namely 400 observations. E.g. for N = 2 observations per network, 200 random networks (M) were generated, in order to get a total of 400 observations (N_{\text{tot}}).

<table>
<thead>
<tr>
<th></th>
<th>N = 2</th>
<th>N = 5</th>
<th>N = 10</th>
<th>N = 25</th>
<th>N = 50</th>
<th>N = 100</th>
<th>N = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>200</td>
<td>80</td>
<td>40</td>
<td>16</td>
<td>8</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>N_{\text{tot}}</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
</tbody>
</table>

The networks all consisted of 10 mutations. They were randomly generated and the rates were uniformly drawn from the interval 0.5 - 2. Using these, the 400 observations, consisting of a genotype and a sampling time, were simulated. The frequency of the different mutations for one example are plotted in Fig. 11.

![Figure 11: Data investigation.](image)

One example of the observed mutation frequencies for the case where 25 observations were generated per treatment, from 16 different networks.
The simulated dataset was used as input for the network learning using the three different settings (weighted, one-for-all, separate) and the performance was compared using above mentioned testing values.

### 4.3.2 Weight computation: first approach

In the simulation setting the first problem was that no real treatment was assigned to the different, randomly generated networks. Still, the similarity among them had to be computed, in order to be able to use the new model. The first attempt was to compute the weights from data generated from the different networks using ideas from the responsibility computation used for mixture models (Eq. 11, Fig. 12) [9].

![Diagram showing the first attempt for the weight computation](image-url)

**Figure 12: First attempt for the weight computation.** From each of the networks 200 observations (represented as coloured bars) were generated, the respective average probabilities are computed as depicted and processed to reach the final similarities.

In order to compute the similarities, 200 observations from each of the networks were generated, represented as coloured bars in Fig. 12. Then, the average probabilities of these observations, given any of the networks, were computed. The idea being that two very different networks will only have a few common compatible genotypes, where a probability can be computed. For all the other observations, the probability will be zero and therefore, reduce the average probability. Further, the similarity from the network with itself should be one. Therefore, a row-wise normalisation by the diagonal elements was implemented, such that the diagonal contained ones. This resulted in an almost symmetric matrix. In order to make it perfectly symmetric a heuristic step was introduced, namely, averaging over the symmetric entries (Eq. 11).
\[ \gamma_u^{(v)} = \frac{1}{2} \left( \frac{P(X^{(k \in u)}|\text{poset}_u, \lambda_u)}{P(X^{(k \in u)}|\text{poset}_u, \lambda_u)} + \frac{P(X^{(k \in v)}|\text{poset}_v, \lambda_v)}{P(X^{(k \in v)}|\text{poset}_v, \lambda_v)} \right) \]  

(11)

**Figure 13: Weight distribution.** One example of the distribution of the observed weights for the case with \( N=25 \) simulated observations per treatment.

The range resulting of the weights (Fig. 13) was between zero and one, with all the similarities between 0.45 - 1. The computation of the weights was computationally expensive and the resulting weights only showed a slight differentiation between observations from the regimen itself and all the remaining ones.

### 4.3.2.1 Assessment of the performance of the different algorithms

The network learning and rate optimisation was carried out for the experimental setup as outlined in Table 1. The testing values were computed and the performance of the new model was analysed. With these weights, it was impossible to show that introducing weights to account for similarities between different treatments could increase the performance of the predictions. This can be explained with the range of the weights. All of them were between 0.45 - 1. This did not lead to a clear distinction of observations from the regimen itself and the remaining ones. The contradictions introduced by the observations from all the other treatments were not weighted down enough. Due to all these contradictions, the ML networks in almost all cases were the empty network. It led to the conclusion that the weights need to have a different distribution. The weights from the observations from the regimen itself should be significantly higher than the rest, such that the major impact comes from these observations. The other observations should have a much lower impact on the outcome and therefore, the corresponding weights should be much lower.

### 4.3.3 Weight computation: second approach

In a second approach, the weights were computed as the Jaccard index [21] between the two true networks \( u \) and \( v \) (Eq. 12). The Jaccard index is defined as the fraction of the intersection and the union of two samples. Specifically, the samples are two transitively-closed adjacency matrices. For the intersection the common entries of the two matrices are counted and for the union the unique number of entries is used.
Results

\[ \gamma_u^{(v)} = \frac{|\text{poset}_u \cap \text{poset}_v|}{|\text{poset}_u \cup \text{poset}_v|} \]  

Figure 14: Weight distribution. One example of the distribution of the observed weights for the case with \(N=25\) observations.

Fig. 21 shows the resulting distribution of the weights. They show a clear distinction between observations from the treatment itself and all the remaining ones.

The network learning was repeated with the new weights as outlined in Table 1. One of the major aims of implementing the mixture model, was to be able to increase the sample size, especially, for the very rare treatments in order to get more accurate predictions. Therefore, first the sample sizes for the different settings were compared and the results are plotted in Fig. 15.

Figure 15: Effective sample sizes in comparison. For the one-for-all case the sample size was constant at \(N=400\) observations, for the weighted and separate case it was computed as the sum of the observed weights.

Using the weighted approach led to an increase in the effective sample size compared to the case using the separate approach. The increase is biggest for the very rare networks. The one-for-all approach, of course, always had the biggest sample size with 400
observations. However, it also had the big disadvantage that the observations were not weighted, which led to an increased number noise level in the data.

4.3.3.1 Assessment of the performance of the different algorithms

The analysis on the simulated database was again carried out using the new weights. The network learning was repeated 20 times in order to get a robust result. The following graphs show the resulting testing values for the different settings. According to these, the performance was assessed. In order to find significant performance differences, a two-sided wilcoxon test was carried out. The significant differences at a significance level of 0.05 are indicated in the figures below.

(a) log-likelihood of the testdata. For \( N < 25 \) the log-likelihood for the separate approach was orders of magnitude lower than the range plotted

(b) Hamming distances

Figure 16: Average test values for 20 repetitions. Significant differences based on a two-sided wilcoxon test are indicated in the plots.

The log-likelihood and the Hamming distances were used as performance indicators for the different settings. A good performance means that the log-likelihood of the testdata and the average Hamming distance are closest to zero.

The summary plots (Fig. 16) show two things. First, the likelihood of the testdata using the weighted approach is much higher as soon as we have less than 50 samples available. Then, using the separate approach, the algorithm overfits the resulting network to the small number of available observations. These results are further emphasised by Fig. 43b, showing the Hamming distances for the different cases. For low numbers of observations the Hamming distance gets significantly bigger when using the separate approach compared to the other two approaches. Lastly, as expected, the one-for-all approach performs worse because of the increased noise level in comparison with the other approaches.

In summary, these plots show that, if appropriate weights are used, the weightet approach outperforms the others, as soon as the number of available observations decreases below 50. This is a satisfactory results, as in the clinical data, for most treatments less than 50 samples are available. This means, that in these cases the new approach using weights, should perform better.
4.3.3.2 Analysis of the effect of different noise levels

Up to now, noise-free observations were simulated. However, we wanted to get a notion of the effects of noise in the data on the performance of the models. So, the noise level was gradually increased. This means, that the genotypes simulated contained different percentages of incorrect mutations. Epsilon is the parameter indicating the noise level. A value of \( \varepsilon = 0.1 \) means that, on average, in each genotype 10\% of the mutations were incorrect. The analysis was carried out for noise levels ranging from 0 - 0.3. The upper bound representing a huge noise level. As before, the total number of observations was set to 400 and the number of networks was defined accordingly for the different cases: 400/N. The data generation and subsequent network learning was repeated 20 times for each value of \( \varepsilon \) and the results are summarised in Fig. 17 and 18.

![Graphs showing the effect of different noise levels](image)

(a) \( \varepsilon = 0 \)

(b) \( \varepsilon = 0.1 \)

(c) \( \varepsilon = 0.2 \)

(d) \( \varepsilon = 0.3 \)

Figure 17: Average test values for 20 repetitions. Comparison of performance using different noise levels.
Results

(a) $\epsilon = 0$

(b) $\epsilon = 0.1$

(c) $\epsilon = 0.2$

(d) $\epsilon = 0.3$

Figure 18: Average test values for 20 repetitions. Comparison of performance using different noise levels.

The overall conclusion from these plots is that the general statement that the \textit{weighted} method outperforms the other two approaches as soon as the number of available samples drops below a certain threshold, remains. As expected, the performance decreases with increasing noise levels. We also can see that the performance of the \textit{separate} approach decreases faster than for the \textit{weighted} approach. With increased noise levels, the first time, the \textit{weighted} model outperforms all the others, according to the log-likelihood, increases from \(N<50\) to \(N<100\). Further, the fact that the increased noise levels do not lead to substantial changes, shows that the model seems to be able to cope with a substantial amount of noise in the data.

4.4 Analysis of clinical data from the Swiss HIV Cohort Study (SHCS) database

In the SHCS database there are two datasets that are relevant, namely, the ones containing the mutations in the reverse transcriptase and in the protease. The reverse transcriptase is needed to reverse transcribe the viral genome such that it can enter the human nucleus and be integrated into the host genome. Reverse transcriptase inhibitors block this process. Similarly, the protease inhibitors block the protease. The protease is needed to cleave the proteins in order to render them functional for their task, the virion processing and assembly (Fig. 19). Due to the additional selective pressure of the drugs, the virus accumulates so-called resistance mutations that eventually make the virus resistant to prescribed the drugs.
The data from the SHCS database was filtered for observations where the regimen, the full genotype and the sampling time were known. Fifteen known resistance mutations from the Stanford Resistance Database (Fig. 4) were selected and used for learning the CBN models. The mutations were selected using results from previous publications [10, 11]. Further, the treatments were clustered according to the number of observations that were available per treatment in order to get a better picture of the performance of the different approaches in various situations. In the different groups the respective number of test samples from each regimen were excluded from the training data and used for subsequent testing (Table 2). For the *separate* approach the separation of sufficient training and testing data could only be guaranteed when at least six observations were available.

**Table 2: Experimental setup:** splitting up the treatments according to the number of available observations

<table>
<thead>
<tr>
<th>N\textsubscript{obs} per treatment</th>
<th>1-2</th>
<th>3-5</th>
<th>6-9</th>
<th>10-29</th>
<th>30-100</th>
<th>&gt;100</th>
</tr>
</thead>
<tbody>
<tr>
<td>N\textsubscript{test separate}</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>4</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>N\textsubscript{test weighted}</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>Number of treatments</td>
<td>103</td>
<td>58</td>
<td>32</td>
<td>33</td>
<td>13</td>
<td>7</td>
</tr>
</tbody>
</table>

In the *one-for-all* case the union of the testing data of all treatments in the given group was used as testing data. For the respective testing values it was accounted for the different size of the testing data. A list of all treatments that were included in the analysis can be found in 9.4.

### 4.4.1 Experimental setup

The filtered data was then used for the network learning and rate estimation. The whole analysis was repeated twenty times with different data in order to get a more...
accurate estimate of the performance. With a small number of observations available the assumption was that the mixture model would outperform the other approaches, whereas when a lot of data was available, the approach using the data from one treatment only might perform equally well or better.

4.4.2 Weight computation: first approach

For the clinical data analysis, initial weights were provided by Jasmina Bogojeska from IBM Zürich. They were computed as the fraction of common resistance mutations between two treatments $u$ and $v$ (Eq. 13).

$$\gamma^{(v)}_{u} = \frac{|\text{mut}_u \cap \text{mut}_v|}{|\text{mut}_u \cup \text{mut}_v|}$$  \hspace{1cm} (13)

where $\text{mut}_u$ respectively $\text{mut}_v$ is the set of known resistance mutations for the treatments $u$ and $v$.

![Figure 20: Weight distribution. One example of the RT dataset analysis for the class with N=6-9 observations.](image)

The resulting weight distribution (Fig. 13) from this approach was very similar to the first approach in the simulation study. The weights spanned all the way from zero to one, showing no great differentiation between observations from the regimen itself compared to all the others.

4.4.2.1 Assessment of the performance of the different algorithms

Using this initial approach for the weight computation, it was impossible to differentiate the performances of the weighted and the one-for-all approach, which means that the extension of the model using these weights did not improve the performance at all. The results described here are summarised in Fig. 41.

4.4.3 Weight computation: second approach

In a second approach, the idea was to mimic the weights from the results from the simulation study. For that purpose, the weights were artificially rescaled (Fig. 21). The weights of one were not changed, whereas all the lower weights were linearly rescaled such that they were between 0 - 0.15. It was an arbitrary step, which was supposed to give further insight into the importance of the distribution of the weights used.
4.4.3.1 Assessment of the performance of the different algorithms

However, the results after rescaling the weights showed no significant differences with the previous approach. The results for the whole data analysis are summarised in Fig. 42. This led to the conclusion, that the current weight computation did not reflect the similarity of the treatments accurately and therefore, a new method had to be found for the weight computation.

4.4.4 Weight computation: third approach

A completely new method for the weight computation was used in the third approach. The weights were computed using a logistic regression model using a five-fold cross-validation and various regularisation parameters ranging from $10^{-5}$ to $10^5$. This model assigns a probability to each observation that it belongs to any of the treatments. The best regularisation parameter was $10^{-4}$. The average overall accuracy of the logistic regression model using this regularisation parameter was 16%.

The resulting weights (Fig. 22) were all very low and again, there was no separation between the observations from the treatment itself and all the other observations.

Figure 21: Weight distribution. One example of the RT dataset analysis for the class with N=6-9 observations.

Figure 22: Weight distribution. One example of the RT dataset analysis for the class with N=6-9 observations.
4.4.4.1 Assessment of the performance of the different algorithms

The results with these weights were not significantly different from the previous approaches. The results from the clinical data analysis are summarised in Fig. 43. In conclusion, the introduction of weights did not manage to increase the performance of the model in comparison with the one-for-all case. Hence, this was not yet an appropriate method for the weight computation for the clinical data.

4.4.5 Weight computation: fourth approach

Finally, the similarities between two treatments were computed as the Jaccard index among the drugs from one drug class for two treatments $u$ and $v$.

$$J_{uv} = \frac{|\text{drugs}_u \cap \text{drugs}_v|}{|\text{drugs}_u \cup \text{drugs}_v|}$$

(14)

where $\text{drugs}_u$ and $\text{drugs}_v$ are the drugs from one drug class only, here RT respectively PR drugs. The reason for that separation is that the PR and RT drugs trigger different resistance mutations that are treated independently. Therefore, two treatments can be very similar when modelling the PR mutations networks, while they are very different when looking at the RT mutations. For example, one patient follows the regimen "ABC+DDI+ETC+SQV" and another one "AZT+EFV+SQV". Considering the RT drugs (ABC, DDI, ETC, AZT, EFV), they have none in common, which results in a Jaccard index of zero, while they have 100% of the PR drugs (SQV) in common, which yields a Jaccard coefficient of one. As the PR and the RT mutational pathways are learned independently, the similarities were computed as the drug specific Jaccard indexes. The idea was that the more identical drugs were prescribed to patients, the more similar the development of resistance mutations should be, as they are a direct consequence of the drugs applied.

In a second step, the additional knowledge about the other drug class was included in the weight computation. The assumption was that if e.g. the PR drug is very efficient, this will have an impact on the replication of the virus and therefore, also on the mutation rates. Hence, including the knowledge from the drug from the respectively other drug class might increase the accuracy of the predictions. For this purpose the weights were computed as follows:

$$\gamma_u^{(v)} = \gamma^{(RT)}_u \cdot \max(J^{(PR)}_{uv}, 0.2)$$

(15)

respectively,

$$\gamma_u^{(v)} = \gamma^{(PR)}_u \cdot \max(J^{(RT)}_{uv}, 0.2)$$

(16)

where $J^{(RT)}_{kl}$ is the Jaccard index for treatment $u$ and $v$ among the RT drugs. Accordingly, $J^{(PR)}_{uv}$ is the Jaccard index for treatment $u$ and $v$ among the PR drugs. In the second term the maximum of the Jaccard index and 0.2 is taken, as in many cases, otherwise this index becomes zero, because the two treatments share none of the drugs of the second drug class. However, we cannot afford to ignore all the data where the second Jaccard index is zero. Therefore, 0.2 was chosen as a very small value that should penalise the treatment without overlap in the drugs from the second drug class.

The resulting weight distributions for the two different weight computation approaches
are shown in Fig. 23. These are one example of the weight distributions, namely, for the treatments for which we have six to nine observations available.

\[(a)\] RT weight distribution: separate Jaccard index (Eq. 14)  
\[(b)\] RT weight distribution: combined Jaccard index (Eq. 15)  
\[(c)\] PR weight distribution: separate Jaccard index (Eq. 14)  
\[(d)\] PR weight distribution: combined Jaccard index (Eq. 16)

**Figure 23: Weight distributions.** One example from the two datasets for the class with N=6-9 observations.

In Fig. 23a, we can see that for the RT dataset there is a bigger variation in the weights than for the PR dataset (Fig. 23c). The reason for that is that mostly, two or three RT drugs and only one PR drug is combined in the drug cocktail. This most often leads to a Jaccard index of zero or one in the PR case. The combined Jaccard indexes then look very similar for the PR and the RT dataset (Fig. 23d, 23b), with the only difference that a lot more observations are ignored in the case of the PR mutations. The final analysis of the two datasets was carried out with these similarities.

**4.4.6 Analysis of RT dataset**

First, the dataset containing the mutations in the reverse transcriptase was analysed. The following figure gives a first impression of the frequency of occurrence of the selected resistance mutations:
Results

Figure 24: RT resistance mutation frequencies. The observed frequency of the selected RT resistance mutations after filtering out observations with missing data.

For the two classes N=1-2 and N=3-5, it was impossible to apply the separate approach, as otherwise the separation of sufficient training and testing data could not be ensured. The major aim of the use of weights was to increase the input sample size. Therefore, the effective sample sizes using the three different approaches were compared in a first step. The results are plotted in Fig. 25. For the weighted approach, the resulting sample size using the weights resulting from Eq. 14 is plotted in yellow and the one resulting from Eq. 15 with a dotted yellow line.

Figure 25: Effective sample sizes in comparison. For the one-for-all case the sample size was constant at N = 2000 observations, for the weighted and separate approach the effective sample size was computed as the sum of the observed weights.

This plot shows that the effective sample size could be increased many fold using the weighted in comparison with the separate approach. Even when using the combined RT-PR-Jaccard indexes (Eq. 16, 15) as weights, the sample size could be largely increased. In the one-for-all case the total number of observations was constantly 2000, however, the noise in these 2000 observations was much higher, because they were just randomly chosen among all the available observations, no matter which treatment the patient followed.
4.4.6.1 Assessment of the performance of the different algorithms on the RT dataset

The whole analysis was first carried out with the weights as defined in Eq. 14 and the performance of the different methods was compared. For the two classes N=1-2 and N=3-5, the separate approach could not be applied. The separation of sufficient training and testing data could not be guaranteed. The results of the analysis are plotted in Fig. 26.

![Comparison of performance](image1)

**Figure 26: Summary plot.** Significant differences based on a two-sided wilcoxon test are indicated in the plots. The analysis could not be carried out for the separate approach for less than 6 observations per treatment.

The log-likelihood and the Hamming distances were used as performance indicators for the different settings. A good performance means that the log-likelihood of the test data and the average Hamming distance are closest to zero.

The log-likelihood plot (Fig. 26a) shows that the weighted approach consistently results in a better performance than the other two approaches. A two-sided wilcoxon test was carried out and the significant differences at a significance level of 0.05 are indicated in the plots. Only in some cases, could we report a significantly better performance of the weighted approach. It can also be seen that the separate approach performs very badly as soon as the number of observations decreases to less than 100. Further, it could be expected that the separate approach performs best for N>100, as the noise in the data is smallest while still a decent number of observations is available. However, the weighted approach does not perform significantly worse than the separate approach in that case.

The results from the Hamming distances are completely in line with the results from the log-likelihood. Due to the fact that the Hamming distance is an integer number, this indicator is less precise. Still, the Hamming distances are consistently the smallest for the weighted approach. However, the differences are small and only in a few cases significant.

In summary, in most cases, it could not be shown, in terms of likelihood or Hamming distance, that the one-for-all approach was significantly outperformed by the weighted approach. Therefore, in a second step, the Jaccard indexes from the two drug classes were combined (Eq. 16, 15). The assumption was that the further reduction of noise in the data might increase the performance of the weighted approach. The results with the new definition of the weights are summarised in Fig. 27.
4.4.6.2 Comparison of the number of edges detected in the three approaches using the RT dataset

In order to further analyse the differences between the three approaches, the number of reported edges in the ML networks were counted and compared. One of the assumptions was that the initial, unexpectedly good performance of the one-for-all method was due to the fact that it reported a lot less edges, while our aim is to report the densest graphs possible. Fig. 28 shows boxplots indicating the number of edges in the transitively closed adjacency matrices describing the resulting networks. The boxplots here show the results when using the combined Jaccard indexes as weights. The boxplots from the first analysis with the Jaccard indexes per drug class (Fig. 44) are in line with the results presented here.
Whenever there are many observations available, all three approaches seem to report similar numbers of edges in the ML networks (N>100). However, with a decreasing number of available samples, the weighted method manages to report more edges. In comparison with the separate method, in the weighted approach, more observations can be used as input and therefore, more different genotypes will be present, which allows to recover more edges. Then, in the one-for-all case, the sample size is the biggest. This, however, leads to an increased number of contradictions in the data. Contradictions mean that the order, in which the events happen, are often not identical. The consequence of these contradictions is that less edges can be recovered in the one-for-all case, in comparison with the weighted case, where contradictions are weighted down.

In conclusion, we are confronted with a trade-off between a big sample size and the number of contradictions in the data. The weighted approach, using the combined Jaccard indexes as weights, combines these in the better way than the other two.

4.4.6.3 Examples of mutational pathways as a consequence of RT drugs

The performance of the different approaches has largely been discussed, however, the resulting networks have not appeared yet. In the following, the output of the different approaches will be compared with each other and with previously published resistance mutation accumulation paths. Fig. 29 summarises the findings in literature [24, 13]. The pathway including the mutations 41, 210 and 215 is also called thymidine analogue mutation one (TAM1) pathway; the one including the mutations 67, 70, 219, accordingly, TAM2 pathway [20, 8]. The TAM1 and TAM2 pathways are known mutational pathways as a response to the increased immune pressure. Mutation 184 has previously been reported in [24].

Figure 28: Number of edges in ML network. The density of the graphs resulting from the network learning using the three approaches was compared.
Results

Figure 29: Resistance mutation accumulation as a response to zidovudine containing treatments. Results from previous publications [24, 13, 20, 8]. Wherever two mutations appear at the same place, the sequence of events could not be defined unambiguously.

These mutational pathways were published as a response to treatment combinations that contained zidovudine (AZT), one of the most frequently prescribed RT drugs in ARTs. The treatment 3TC+AZT was chosen for the comparison with the known resistance mutation accumulation paths, because it was the most frequent treatment in our data. The networks for 3TC+AZT that resulted from our analysis are expected to recover a considerable number of the relations in Fig. 29.

In the one-for-all case, the output consists of one "consensus" network. In the two other cases, one network per regimen is obtained. The networks reported here only contain the relations that appeared in at least 50% of the ML networks from the 20 network learning repetitions that were carried out in total.

Figure 30: Results from the one-for-all approach. The relations reported here are only the ones that were reported in more than 50% of the 20 repetitions; the orange boxes indicate the recovered relations from Fig. 29.

The one-for-all approach managed to recover 215 → 41 → 210 and 67/70 → 219. These two seem to be suggested by the majority of observations in the SHCS database. The TAM1 and TAM2 pathways seem to be suggested by most observations.
Results

Figure 31: Results from the separate approach. The relations reported here are only the ones that were reported in more than 50% of the 20 repetitions; the orange boxes indicate the recovered relations from Fig. 29.

Then, using the separate approach, sparser parts of the pathway in Fig. 29 were recovered, namely, 215 $\rightarrow$ 210 and 70 $\rightarrow$ 219. Neither the TAM1 nor the TAM2 pathway could be reconstructed completely.

Figure 32: Results from the weighted approach. The relations reported here are only the ones that were reported in more than 50% of the 20 repetitions; the orange boxes indicate the recovered relations from Fig. 29.

Finally, the weighted approach managed to reconstruct all the relations from Fig. 29, namely, 184 $\rightarrow$ 215 $\rightarrow$ 41 $\rightarrow$ 210 and 184 $\rightarrow$ 67/70 $\rightarrow$ 219. The TAM1 and TAM2 pathways could be detected perfectly.

In conclusion, in this example, the weighted approach managed to produce the densest graph. It also recovered most of the known relations. This is further evidence for the advantages of the model using the weights suggested in this thesis.
4.4.7 Analysis of PR dataset

Then, the dataset containing mutations in the protease was analysed in the same way as the RT dataset. The following figure gives a first impression of the frequency of occurrence of the selected resistance mutations:

![Frequency of PR resistance mutations](image)

**Figure 33: RT resistance mutation frequencies.** The observed frequency of the selected RT resistance mutations after filtering out observations with missing data.

It can be seen, that the PR mutations are in general a lot less frequent than the RT mutations. Again, the analysis was carried out with the resulting weights from the Jaccard indexes among the drugs from the same drug class (Eq. 14) and then, with the combined Jaccard indexes (Eq. 16) as weights. The resulting effective sample sizes are reported in Fig. 34.

![Comparison of sample sizes](image)

**Figure 34: Effective sample sizes in comparison.** For the one-for-all case the sample size was constant at \( N = 2000 \) observations, for the weighted and the separate approach the effective sample size was computed as the sum of the observed weights.

For the cases with one or two observations per treatment, the sample sizes for the weighted approach in some repetitions reached zero. Therefore, this group of treatments was not included in the following analysis. This can be explained as follows: Usually more RT drugs than PR drugs are prescribed. Therefore, in many cases, the Jaccard indexes among the PR drugs were zero, because there were no shared PR drugs in the regimens. As a consequence the required input sample size of 2000 observations was not reached.
For all other cases, the sample size could clearly be increased in the *weighted* approach compared with the *separate* approach. Again, the sample size is biggest for the *one-for-all* approach, however, this is linked with poorer quality of the input data.

4.4.7.1 Assessment of the performance of the different algorithms on the PR dataset

First, the complete analysis was carried out with the Jaccard indexes per drug class. The network learning was repeated twenty times in order to get a robust estimate of the performances. As the sample sizes for the treatments with 1-2 observations sometimes reached zero, this class was not included in the summary plots below (Fig. 35). The *separate* approach was only applied to the groups with at least six observations to ensure sufficient training and testing data and a separation of the two.

![Figure 35: Summary plot.](image)

Significant differences based on a two-sided wilcoxon test are indicated in the plots. The analysis could not be carried out for the *separate* approach for less than 6 observations per treatment.

The results here are a bit more contradictory. The log likelihood of the *weighted* approach is still mostly closer to zero than for the other two approaches. However, in many cases the differences are not significant and very small.

Surprisingly, the average Hamming distance using the *one-for-all* approach is slightly smaller than for the *weighted* approach, while the log-likelihood for the *one-for-all* approach is lower than for the *weighted* approach. These contradictions, however, are only significant in two cases.

After this first analysis, the Jaccard indexes of the two drug classes were combined in order to find out if, similarly to the RT analysis, the performance of the model could be increased with further noise reduction in the input data.
The general picture is similar to the previously discussed. However, for the log-likelihood, now in some cases, differences between the one-for-all and the weighted case are significant. Showing that the weighted model outperforms the other two. The Hamming distances in the one-for-all case are still slightly lower on average, but only in one case, this difference is significant. The poorer performance of the defined weights on the PR dataset can be due to the fact that usually a maximum of one PR drug is prescribed per treatment. Therefore, the Jaccard indexes are either one or zero, which is not a perfect measure to reflect similarities. Also, the mutation frequencies for the PR dataset are much lower than for the RT dataset. In total, the mutations are observed less often, therefore, the reconstruction of the networks gets more difficult.

In conclusion, the results here are worse than for the RT data analysis, but still, it can be shown that the weighted model clearly performs better than the separate approach. Compared to the one-for-all approach, in terms of likelihood, a better performance of the weighted model can be shown in some cases. However, in terms of Hamming distances this can not be confirmed. Further research on the optimisation of the suggested weights for the PR drugs is required, in order to improve the performance of the new model.

4.4.7.2 Comparison of the number of edges detected in the three approaches using the PR dataset

Lastly, the number of edges in the reconstructed networks were compared for the different settings of the PR dataset analysis.
Figure 37: **Number of edges in ML network.** The density of the graphs resulting from the network learning using the three approaches was compared.

For the very frequent treatments (N>100), the *separate* approach reported the most edges. This can be explained with the fact that the noise is the smallest in this approach, as only observations from the treatment itself are considered for the network learning. However, such that this approach performs well, more than 100 observations have to be available, which is not the case for many treatments. As soon as the number of observations per treatment dropped below 100, the *weighted* approach consistently reported the highest number of edges. The fact that the *weighted* approach manages to report more edges is another advantage of the new model.

### 4.4.7.3 Examples of mutational pathways as a consequence of PR drugs

Not so much has been published for the resistance mutation accumulation in the protease. For the drug "IDV", the previously reported relations are the following: 54 → 82 and 90 → 84 [11]. The most frequent treatment from our database that included the drug "IDV" was selected for comparison, namely "3TC+AZT+IDV". The networks reported here only contain the relations that appeared in at least 50% of the ML networks from the 20 network learning repetitions that were carried out.
Results

Figure 38: Results from the separate approach. The relations reported here are only the ones that were reported in more than 50% of the 20 repetitions.

Figure 39: Results from the weighted approach. The relations reported here are only the ones that were reported in more than 50% of the 20 repetitions.

Figure 40: Results from the one-for-all approach. The relations reported here are only the ones that were reported in more than 50% of the 20 repetitions.

In none of the cases, the above mentioned relations have been detected. Also in another regimen containing "IDV", none of these relations could be recovered. The data does not seem to support these, on the contrary, in most cases, they seem to be completely independent of each other. Not much research has been carried out on the PR mutations and therefore, validation of the pathways shown here is difficult. One possible explanation could be that the Jaccard indexes for the PR drugs are mostly one or zero. Consequently, a big amount of observations was ignored completely and we might just consider a too small input sample size to detect above mentioned temporal ordering of the respective mutations.
5 Discussion

Through the course of this thesis, we further developed an existing algorithm in order to deal with the prevalent data sparseness in the clinical data from the SHCS database. A mixture model approach was used to increase the sample size. The contributions of the different observations were weighted according to the similarity between two treatments. The performance of the new model was characterised using the log-likelihood of the test data and the Hamming distance between the observed genotype and the MAP estimate of the genotype for the observed sampling time.

First, the new implementation of the model had to be validated. For this purpose three networks consisting of five mutations with rates were randomly generated. Using 200 simulated observations, the algorithm reconstructed the true underlying networks in all cases. Hence, given enough noise-free samples, it is possible to recover the true network. Secondly, the rate-prediction accuracy was assessed. For this purpose, ten networks consisting of ten mutations with mutation rates were randomly generated. From each of these, 200 observations were simulated and used for the rate optimisation for the three different approaches (weighted, one-for-all, separate). It was assumed that the true network was recovered and only the rate optimisation was carried out, in order to see how the different approaches influence the rate optimisation. The mean squared error (MSE) between the real and the ML rates was computed and compared. It could be shown that the MSE was smallest for the weighted approach. Introducing weights, led to an increase in the sample size. More data was available to assess the mutation rates of events, especially those which occurred late in the pathway. This increase in sample size explains the increased performance in the weighted approach. The inaccuracies using the separate approach could be explained by the fact that the mutations with inaccurate ML rates were the ones from mutations that occurred later in the pathway. Therefore, less data was available and the rate optimisation was less accurate. As expected, the one-for-all approach was generally less accurate than the other two because the noise level, compared to the other two approaches, was much higher. The further increased sample size could not compensate that.

Then, the simulation study was designed in such a way that we could characterise the performance of the different approaches in various settings. The data was simulated from a varying number of networks that were randomly generated. The total number of observations was kept constant at 400. These were used as input data for various different cases. The number of observations generated per network varied between 2 - 200, mimicking the clinical data. The corresponding number of networks generated varied between 200 - 2. The test data consisted of 50 samples generated from each of the networks. For the one-for-all approach the union of the test data was used. For the weighted approach, two weight-computation methods were tested. The successful method computed the weights as the Jaccard index of two transitively-closed adjacency matrices. The transitively-closed adjacency matrices are the binary representations of the mutational pathways. The results of this analysis showed that for a big sample size the separate method showed the best performance in terms of log-likelihood and Hamming distance. However, as soon as the number of observations dropped below 50 observations per treatment, the weighted method outperformed the other two. Then, noise was added to the simulated database (10 - 30 %). In these cases, the weighted approach performed equally well as the separate approach for the frequent treatments and significantly better for the rare treatments. As in the clinical data noise is present and for most treatments only a few observations are available—the method of choices
should be the *weighted* method. This assumption was investigated next using clinical data from the SHCS database.

For the clinical dataset, the weights could not be computed as in the simulation study, because the true mutational pathways were not known. Therefore, different weight computation approaches were tested. The method suggested here was to compute the weights as the combination of the Jaccard indexes among PR and RT drugs. In a first step, the available data was split in groups according to the number of observations that were available per treatment. The resistance mutation accumulation was analysed separately for the RT and the PR mutations. For the RT mutations, we could show a significantly better performance of the *weighted* approach as soon as the number of observations dropped below 100 per treatment. Moreover, for the treatments with more than 100 observations available, using the *separate* approach did not perform significantly better than the *weighted* approach. Further, the *weighted* approach reported most edges in the ML networks, which is another advantage of the new model. For the PR mutations, the results related to the performance of the *separate* approach were identical.

The main difference is that the Hamming distance for the *weighted* approach was, on average, slightly bigger than for the *one-for-all* approach, while the log-likelihood was consistently better for the *weighted* approach. In summary, we could confirm the results from the simulation study with the analysis of the clinical datasets. The mixture model, using fixed similarities between treatments as weights, significantly outperforms previous approaches.

The robust ML networks for the treatment 3TC+AZT, modelling the RT mutation accumulation, were reported and compared with previous publications. The *weighted* approach recovered most of the known relations (TAM1, TAM2). The *separate* approach recovered fewer relations from the TAM1 and TAM2 pathways. Overall, the *separate* approach, reported less edges than the *weighted* approach. Finally, the *one-for-all* approach also recovered the TAM1 and TAM2 pathways as they seem to be supported by the majority of the observations. The increased number of contradictions in the input data led to a sparser graph in comparison with the other two approaches.

The same was repeated for the PR mutations. Here, the treatment 3TC+AZT+IDV was chosen for further discussion. The only PR drug in that treatment is IDV. In general, much less has been published for the PR mutations. For this drug only two relations have been published in one paper. No other publication confirmed these relations. In our analysis, none of these were recovered through our approaches. Another database could be used for the mutational pathway reconstruction in an attempt to confirm the results presented here.

In conclusion, besides developing a new model, a performance analysis was carried out. First, the performance of the new model was studied using simulated data. In addition, the clinical dataset from the SHCS was analysed. The crucial point for the mixture model approach to work well was the weight computation. In this thesis, different weight computation approaches were presented and tested. The combination of the Jaccard indexes among the separate drug classes was the most promising method for the weight computation using clinical data. Using this approach, we managed to confirm the advantage of the use of the extended model that had already been shown in the simulation study. The *weighted* approach, using the weights as suggested in this thesis, typically performs better in terms of likelihood and Hamming distance and it manages to report more edges in the ML networks in comparison to all the other approaches, especially in the presence of limited data availability and noise in the data.
6 Outlook

This thesis was a first step in the fight against the data sparseness that we are confronted with when it comes to clinical data from HIV patients. A possible extension for the simulation study would be to fix some parts of the pathways in all the simulated mutational pathways. In this thesis, the simulated networks were completely randomly generated. Fixing some parts of the pathways might further increase the performance of the weighted approach and especially, it would be more similar to the real situation.

Regarding the clinical data analysis, in further approaches the weights might be further optimised. For example, by using a complete mixture model, where the weights are not fixed from the start. The weights as suggested in this thesis could be used as prior knowledge. Also, for the mutations in the protease, another weight computation that is more differentiated might yield better results. The problem is that the Jaccard index among PR drugs is typically zero or one because of prescription trends of PR drugs in combination with RT drugs.

7 Acknowledgements

This thesis has benefited greatly from the support of many people, some of whom I would like to mention here. To begin with, I am very grateful to Niko Beerenwinkel who provided me with a fascinating project in his research group. Special thanks go to Hesam Montazeri who was my supervisor throughout my research. Not only did he introduce me to advanced R computation, but he also patiently answered all my questions on the theory of bayesian networks and others. Further, thanks to Jasmina Bogojeska who helped with her experience from previous research on a similar problem. Finally, I would like to thank the whole Computational Biology Group for making my time so enjoyable.
8 References


9 Additional Material

9.1 Summary plots from the different unsuccessful approaches

**Figure 41:** First approach. blue: weighted, green: one-for-all, red: separate.

**Figure 42:** Second approach. blue: weighted, green: one-for-all, red: separate.

**Figure 43:** Third approach. blue: weighted, green: one-for-all, red: separate.
9.2 Comparison of the number of edges detected in the three approaches using the RT dataset using the separate jaccard index

![Box plots showing the number of edges in ML network before rescaling.](image1)

**Figure 44:** Number of edges in ML network before rescaling.

9.3 Comparison of the number of edges detected in the three approaches using the PR dataset using the separate jaccard index

![Box plots showing the number of edges in ML network before rescaling.](image2)

**Figure 45:** Number of edges in ML network before rescaling.
9.4 Names of all the treatments used for network learning

![Figure 46: The 246 treatments used for the analysis.](image-url)