Sequence analysis

Analysis of HIV-1 pol sequences using Bayesian Networks: implications for drug resistance


Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium, 1Helsinki Institute for Information Technology, Helsinki, Finland, 2Virology Laboratory, Hospital Egas Moniz, Lisbon, Portugal, 3Chaim Sheba Medical Center, Ministry of Health, Tel-Aviv, Israel, 4Departamento de Genética, Universidade Federal do Rio de Janeiro, Brazil, 5Brown University, Providence, RI and 6ESAT, Katholieke Universiteit Leuven, Leuven, Belgium

Received on May 25, 2006; revised on September 26, 2006; accepted on September 28, 2006
Advance Access publication October 4, 2006
Associate Editor: John Quackenbush

ABSTRACT

Human Immunodeficiency Virus-1 (HIV-1) antiviral resistance is a major cause of antiviral therapy failure and compromises future treatment options. As a consequence, resistance testing is the standard of care. Because of the high degree of HIV-1 natural variation and complex interactions, the role of resistance mutations is in many cases insufficiently understood. We applied a probabilistic model, Bayesian networks, to analyze direct influences between protein residues and exposure to treatment in clinical HIV-1 protease sequences from diverse subtypes. We can determine the specific role of many resistance mutations against the protease inhibitor nelfinavir, and determine relationships between resistance mutations and polymorphisms. We can show for example that in addition to the well-known major mutations 90M and 30N for nelfinavir resistance, 88S should not be treated as 88D but instead considered as a major mutation and explain the subtype-dependent prevalence of the 30N resistance pathway.

Contact: koen.deforche@uz.kuleuven.ac.be
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

HIV escapes the inhibitory effect of antiretrovirals by selection of mutations that increase resistance against those antiretrovirals. Resistance testing, where the activity of available antiretrovirals against the possibly mutated virus from a patient is measured or predicted from the genotype, is therefore the standard of care. Longitudinal data, where virus from the same patient undergoing treatment is sequenced at multiple time points, are difficult to obtain, but would offer a direct insight in mutations that are selected under therapy. In this study, we used the prevalence of mutations in \textit{in vivo} cross-sectional sequence data to study resistance evolution.

Because of its high evolution rate, HIV shows a large amount of natural variation. This variation is currently classified in nine identified subtypes within the HIV-1 main group, and 16 circulating recombinant forms (CRFs), which are epidemic strains reflecting mosaic recombinations of the pure subtypes with each other or with other CRFs. In addition, unclassified strains and new recombinants are being increasingly reported. Subtype B is the most prevalent HIV subtype in the Western world, and antiretroviral drugs have been targeted mainly at this subtype. The applicability of current drugs has been studied for non-B subtypes (Kantor and Katzenstein, 2004) and it is currently believed that these drugs work also for non-B subtypes. However there is concern that some minor resistance-associated mutations in subtype B are in fact wild-type amino acids in other subtypes. In addition, different prevalences of known resistance-associated mutations and new mutations are seen in different subtypes.

To model the effect of mutations and of their interactions on drug resistance, we used Bayesian networks (see also ‘Introduction to Bayesian Networks’ in Supplementary material). A Bayesian network (BN) is a probabilistic model that describes statistical independencies between multiple variables (Pearl, 1998). Dependencies are represented in a directed acyclic graph and form the qualitative component of a BN. A BN may be learned from data by searching for the network that explains a maximum of the observed correlations in the data using a minimum number of arcs (Hackerman, 1999), and we learned a BN from amino acid sequence data annotated with treatment. Because BN learning has previously successfully been used to discover relationships between amino acids for predicting protein secondary structure (Klingler and Brutlag, 1994), they are well suited to learn interactions between different protein residues. We show how a semantic meaning can be postulated with respect to drug resistance for the presence of arcs in the network between amino acids and for the network structure around the drug node.

The objective of this work was to evaluate the usefulness of BN learning to gain insight in resistance pathways against the protease inhibitor nelfinavir, by identifying the role of mutations selected during treatment. In addition, we discovered interactions between background polymorphic positions and resistance mutations that

© 2006 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
may explain observed subtype differences. Our results may be relevant for genotypic resistance interpretation algorithms that intend to predict the therapy response for various drugs (Van Laethem et al., 2002), based on the presence of mutations at positions associated with drug resistance.

2 MATERIALS AND METHODS

Data

Sequences from patients naive for treatment with protease inhibitors (PI) and from patients with only experience to nelfinavir as PI treatment were collected from five clinical databases: PT (Hospital Egas Moniz, Lisbon, Portugal), LEU (University Hospital, Leuven, Belgium), ISR (Ministry of Health, Tel-Aviv, Israel), BR (Universidade Federal do Rio de Janeiro, Brazil) and NONB (an international database containing sequences from subtypes other than subtype B (Kantor et al., 2005)). At most one treated and one naive sequence per patient was included and identical sequences were removed. The subtype was determined using the REGA HIV-1 subtyping tool (de Oliveira et al., 2005) on the protease and partial reverse transcriptase. CRFs were not detected since this information did not contribute anything for CRFs without breakpoint in protease. Different mutations at a single position were treated independently.

Nelfinavir-associated mutations were detected by testing for conditional independence between a mutation and nelfinavir experience using the Cochran–Mantel–Haenszel \( \chi^2 \)-test, correcting for an unequal ratio of sequences from treated versus untreated patients within each combination of subtype and database and for multiple comparisons using Benjamini & Hochberg with a False Discovery Rate of 0.05. The dataset for BN learning was compiled by stratifying for an equal ratio of sequences from untreated versus treated patients, within each combination of subtype and database. Nelfinavir-associated mutations and wild-type amino acids (with prevalence over 10% in the PI naive population) at polymorphic positions and nelfinavir-associated positions were included as separate boolean variables. For a position with two amino acids, only the one with largest variation was included, and properties of the other amino acid, (such as antagonistic or protagonistic effects) were derived from the properties of the included amino acid.

BN

BN learning was done using the B-Course software (Myllymäki et al., 2002), scoring models by maximizing the posterior probability of the model. Robustness of network features were assessed with a non-parametric bootstrap using 100 replicates (Friedman et al., 1999). Only arcs, but ignoring arc direction, with a bootstrap support over 65% were considered robust.

The explanation of the network for effects, such as the reported subtype dependence of selection during nelfinavir treatment of the 30N resistance pathway (Gomes et al., 2002; Grossman et al., 2004) was investigated using the exact Bayesian inference in the network. The network arc that was most important for explaining the effect was searched by evaluating for each arc how well the network lacking the arc still predicted the effect.

In the network graph, nodes indicating the presence of different amino acids at a single position were grouped together, and obvious strong antagonistic effects between presence of these amino acids were not shown. Network arcs were coloured according to their estimated semantic meaning. Arcs that indicated a direct influence between nelfinavir-associated mutations were discriminated from arcs that indicated a direct influence from background polymorphisms on nelfinavir-associated mutations. While this distinction is ambiguous at polymorphic nelfinavir-associated positions, this facilitated the interpretation of the network.

BN

Network semantics were based on the correspondence between an unconditional dependency implied by an arc between two mutations and a possible epistatic fitness interaction between these two mutations (Klingler and Brutlag, 1994). A mutation that confers phenotypic resistance on its own and plays a key role in drug resistance is considered a major mutation (Shafer, 2002). A minor mutation further increases resistance mostly only in presence of a major mutation, or compensates for a possible fitness impact of other mutations, and is therefore selected only in the presence of these other mutations. Since a minor mutation interacts epistatically with a corresponding major mutation, the BN indicates this relationship by an arc between these mutations. The presence of a minor mutation is predicted mostly by the presence of the corresponding major mutation, and thus expected to be unconnected to treatment in the network. In contrast, the selection of a major mutation, as a first mutation, is unconditionally dependent on treatment, and thus expected to be connected to the treatment node. Since the dataset included both resistance mutations and wild-type amino acids at some positions, the positive protagonistic association with a resistance mutation may also be indicated through an inverse antagonist association with the wild-type at that position.

Although it has been shown that arc direction in a learned BN may indicate causality (Pearl, 1998), this assumes an acyclic causality graph, since the BN is by definition acyclic. An epistatic effect between two mutations implies a bi-directional causality: the presence of anyone of these mutations changes the fitness and thus the prevalence of the other mutation. Thus arc direction cannot be interpreted in a causal way. However, robust arc directions can indicate a complex non-additive multivariable effect of multiple parent nodes through conditional dependencies, and provide valuable information, when confirmed by analysis of the contingency table of the nodes involved. Around the drug node, this information was used to infer different resistance pathways, as the parent nodes use robust conditional dependencies to indicate mutual exclusivity.

3 RESULTS AND DISCUSSION

Dataset

The dataset for BN learning included 340 sequences from patients on nelfinavir therapy of various subtypes (see Supplementary Table 1), and 967 sequences from PI naive patients (same subtype distribution). Statistical analysis identified 18 polymorphic positions, and 30 nelfinavir-associated mutations with a prevalence over 0.5%. These were 10F, 13V, 20T/V, 23I, 30N, 33I/F, 35D, 62V, 63P and 82I. See Supplementary Figure 1 for the prevalence and subtype distribution of these mutations in naive and treated patients. Many of these mutations (20V, 33I, 35D, 62V, 64V, 66F, 71I/T/V, 74S, 82I, 88D/S, 89I/T/V, 90M and 93M, of which five were polymorphisms (13V, 35D, 62V, 63P and 82I). See Supplementary Figure 1 for the prevalence and subtype distribution of these mutations in naive and treated patients. Many of these mutations (20V, 33I, 35D, 62V, 64V, 66F, 71I/T/V and 93M) have been previously reported as selected by protease inhibitors in general (Wu et al., 2003), but most of them have not been associated with nelfinavir experience in particular. Selection of these mutations was more pronounced or sometimes exclusively in non-B subtypes.

The final dataset held 53 variables: 30 nelfinavir-associated mutations, 22 wild-type amino acids (including polymorphisms) and the nelfinavir treatment variable.

Nelfinavir experience BN

The most probable network is shown in Figure 1. Of 193 arcs included in the most probable network, 118 had a bootstrap support over 65%. Polymorphic positions 15, 37, 41 and 70 did not directly influence any nelfinavir-associated position and were omitted from the figure.

The semantics we postulated allowed us to consider only mutations directly connected to the cNFV node as major mutations.
Mutations 30N, 88S and 90M were connected with nelfinavir treatment as parent with bootstrap support 100% and mutations 20T, 74S, 89I/T and 35D/G/N (through an inverse association with 35E) as child, with lower robustness. Arc direction was found to be highly robust around eNFV (98%, data not shown), and a contingency table of mutations 30N, 88S 90M and 20T (Supplementary Table 2) revealed a multivariable effect indicating that nelfinavir treated sequences contained mostly only one of 30N, 88S or 90M, while mutation 20T was rarely selected alone, instead being mostly in the presence of one of 30N, 88S or 90M. Mutations 30N, 88S and 90M should thus be considered major resistance mutations of different resistance pathways and mutations 20T, 74S, 89I/T and 35D/G/N as minor mutations since rarely observed alone. It has been widely documented that
nelfinavir resistance follows two distinct pathways characterized by the primary mutations 30N and 90M (Grossman et al., 2004). Our analysis extends this with a new pathway characterized by the primary mutation 88S. This finding is in agreement with public phenotypic fold changes of 8.9 for NFV. Our analysis also confirms the current knowledge that 88D is a minor mutation selected in the presence of 30N (bootstrap support 100%), and thus clearly indicates the different role of 88S versus 88D.

The network indicated how background polymorphisms affect selection of resistance-associated mutations (blue arcs). As described in Materials and Methods, Bayesian inference was used on network variants each lacking a particular arc to determine how the network explains the lower prevalence of 30N versus 90M in subtypes G and A1, compared to subtype B. As the network variant without the arc between 30N and 89L lost most of this effect, this arc was used by the network to explain the effect. Thus 89L/30N exhibits a higher resistance or a higher replication capacity than 89M/30N. The fitness interaction between 89M and 30N has been confirmed using in vitro mutagenesis (Gonzalez et al., 2004; Abecasis et al., 2005).

Application of knowledge learned from the network

Despite the evidence for interactions between drug resistance mutations, linear models for predicting drug resistance are usually found to perform very well (Wang et al., 2004). This apparent contradiction is explained by considering that evolution does not select for unfavorable combinations of mutations, such as for example 88D without 30N. Therefore, these sequences are rarely seen, and thus a model may consider separate contributions of 30N and 88D to drug resistance. However, to predict therapy response, next to drug resistance phenotype, also evolution to drug resistance, quantified by the genetic barrier, is important (Vandamme and De Clercq, 2001). To calculate the genetic barrier, it is necessary to model constraints in evolution of resistance mutations (Beerenwinkel et al., 2005). In particular, the presence of polymorphisms or minor mutations at baseline may accelerate the selection of major resistance mutations and thus impact the genetic barrier.

A comparison of mutations before and after therapy is generally not considered to be conclusive evidence for a phenotypic or clinical role of a mutation. To evaluate the impact of compensatory secondary mutations using in vitro mutagenesis experiments it is necessary to consider the genetic context in which they contribute to resistance, which is indicated by the BN. This approach was used by Abecasis et al. (2005) to investigate the phenotypic role of mutations at position 89 in subtype G in a context of mutations at positions 71 and 90.

Our approach of BN learning follows a setup that bears resemblance to the work of Beerenwinkel et al. (2004), where a mixture of Bayesian tree models, which are a constrained version of BNs, are used to model HIV drug resistance evolution from similar cross-sectional data. Their model represents ordered accumulation of mutations following a number of possible trees. While such a model captures protagonistic epistatic fitness effects between resistance mutations, the current version available cannot represent antagonistic effects and epistasis of resistance mutations with non-mutating polymorphisms, and it is not clear how this could be done. We used the latter capability to investigate subtype differences, while Beerenwinkel et al. used only subtype B sequences.

The presented application of BN learning provides novel information on evolution of drug resistance missed by other descriptive data mining techniques, such as pairwise correlation and clustering techniques (Sing et al., 2005) or mixtures of mutagenic trees, by investigating at the same time interactions (antagonistic or synergistic) between all amino acid residues, including non-mutating polymorphisms. In future work, we plan to quantitate the fitness function experienced by the virus during treatment, using the BN structure as a blueprint for all epistatic interactions.

ACKNOWLEDGEMENTS

The authors wish to thank all collaborators of the non-B workgroup (http://hivdb.stanford.edu/pages/collaborators/Non_B_group.html), Ana Abecasis, and Esmeralda AJM Soares and all clinicians for their generous help in obtaining data, and Hendrik Blockeel for review of the manuscript. K.D. was funded by a PhD grant of the Institute for the Promotion of Innovation through Sciences and Technology in Flanders (IWTF). Y.M. is a post-doctoral researcher with the FWO-Vlaanderen; his research is supported by KULeuven EF/05/007 SymBioSys, GOA-Mefisto-666 and GOA-Ambiorics, Belspo IUAP V-22, and EU FP6 NoE Biopattern. This work was supported in part by the AIDS Reference Laboratory of Leuven that receives support from the Belgian Ministry of Social Affairs through a fund within the Health Insurance System, by FWO-Vlaanderen grant G.0266.04, and by the Katholieke Universiteit Leuven through Grant OT/04/43. KD was funded by a Ph.D grant of the Institute for the Promotion of Innovation through Sciences and Technology in Flanders (IWTF), which also provided funding to pay for Open Access publication charges.

Conflict of Interest: none declared.

REFERENCES


