

Deliverable 4.3 Enhancement Plan

Project start:	1 March 2006
Project Duration:	36 months
Priority area:	2.4.11
Contract No.:	INFSO-IST-027446
Website:	http://www.virolab.org

Due-Date:	Month 18
Delivery:	Month 18
Lead Partner:	University of Amsterdam
Project Leader	Prof. P. M. A. Sloot
Dissemination Level:	public
Status:	final pre-approval
Approved:	
Version:	0.6

Log of Document

Version	Date	Changes Summary	Authors
1.1	27-09-2007	Addendum from Joc Cing	Breannán Ó Nualláin
1.0	27-09-2007	Final pre-approval version	Breannán Ó Nualláin
0.6	27-09-2007	Added list of abbreviations	Breannán Ó Nualláin
0.5	27-09-2007	Augmented chapter 11	Breannán Ó Nualláin
0.4	20-09-2007	Edits to contributions	Peter M. A. Sloot
0.3	18-09-2007	Reorganised by research initiative	Breannán Ó Nualláin
0.2	07-09-2007	Restructured and added overview	Breannán Ó Nualláin
0.1	07-08-2007	Collected inputs from partners	Breannán Ó Nualláin

Table of Contents

1	Executive summary	7
2	Overview & Objectives	7
3	Molecular dynamics of interactions between inhibitors and viral proteins.	10
3.1	Basis for the approach	12
3.2	Background & previous work	12
3.3	Work & publications completed in the context of Virolab	13
3.4	Plans for the remainder of the Virolab project	15
3.5	Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.	15
4	Building a kinetic model for HIV protease cleavage.	16
4.1	Basis for the approach	16
4.2	Background & previous work	17
4.3	Work & publications completed in the context of Virolab	17
4.4	Plans for the remainder of the Virolab project	18
4.5	Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.	18
5	Elucidating selection forces that act on the pathogenic potential of HIV.	19
5.1	Basis for the approach	19
5.2	Background & previous work	19
5.3	Work & publications completed in the context of Virolab	20
5.4	Plans for the remainder of the Virolab project	20
5.5	Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.	20
6	Modelling entry of HIV into target cells.	21
6.1	Basis for the approach	21
6.2	Background & previous work	22

6.3	Work & publications completed in the context of Virolab	22
6.4	Plans for the remainder of the Virolab project	23
6.5	Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.	23
7	Hybrid multi-agent modelling of lymph node cell movement.	24
7.1	Basis for the approach	24
7.2	Background & previous work	25
7.3	Work & publications completed in the context of Virolab	26
7.4	Plans for the remainder of the Virolab project	28
7.5	Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.	29
8	Information-theoretic measures of genetic distance.	30
8.1	Basis for the approach	31
8.2	Background & previous work	31
8.3	Work & publications completed in the context of Virolab	32
8.4	Plans for the remainder of the Virolab project	32
8.5	Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.	32
9	Complex-network models for HIV spreading.	33
9.1	Basis for the approach	33
9.1.1	Modeling the dynamics of the infected nodes	34
9.1.2	Direct Simulation Algorithm	35
9.2	Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.	35
9.2.1	Current research	36
9.2.2	Conclusion	38
10	Semi-automated Literature Mining.	38
10.1	Basis for the approach	39
10.2	Background & previous work	40
10.3	Work & publications completed in the context of Virolab	40
10.4	Plans for the remainder of the Virolab project	41

10.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.	42
11 Enhancements of rule-based interpretation.	42
11.1 Basis for the approach	42
11.2 Background & previous work	43
11.3 Work & publications completed in the context of Virolab	44
11.4 Plans for the remainder of the Virolab project	47
11.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.	48
12 Bayesian combination of evidence.	48
12.1 Basis for the approach	48
12.2 Background & previous work	49
12.3 Work & publications completed in the context of Virolab	51
12.4 Plans for the remainder of the Virolab project	51
12.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.	51
13 Abbreviations	51
Bibliography	52

List of Figures

2.1	Research initiatives at all levels can support decision making. . . .	11
7.1	Simulation of LPS effect on TNF and IL10	29
7.2	Chemotaxis Model based on Receptor Kinetics	30
9.1	Simulation results and reported data for the AIDS epidemic and reconstruction of HIV cases in the USA.	36
9.2	Direct simulation results for infection spreading through homogeneous and inhomogeneous networks. Number of vertices is 10^5 , probability of infection is 5%, number of experiments is 1000. a) homogeneous network with one edge at each node; b) scale free network with exponent $\gamma = 2.5$	37
9.3	The structure of the contact epidemiological network based on the k-kernel decomposition	38
10.1	Semi-automated Literature Mining	40

1 Executive summary

This deliverable ongoing work and plans for the remainder of the duration of the Virolab project on the subject of enhancements to the Virolab HIV drug-susceptibility interpretation system.

At the first annual Virolab project review, the reviewers expressed the opinion that, given the work that had already been completed on technology and building of computational infrastructure, the project should shift its emphasis somewhat towards simulation, modelling and enhancement of the Virolab HIV drug-susceptibility interpretation system. We have taken that recommendation to heart and have fostered a number of research initiatives, besides those already planned, together with our associate partners.

These research initiatives are not intended to provide complete coverage of what is, after all, a highly multiscale problem. We merely attempt to peer through some windows into the complex dynamics of HIV. However each of these windows into the problem can provide us with useful information which we can avail of in delivering decision support (see Figure 2.1).

In this deliverable we outline these research initiatives, which are of two kinds: those that supply data and knowledge to the Virolab HIV drug-susceptibility interpretation system as well as initiatives to improve the system itself and its capacity to process and combine evidence of various sorts.

The research initiatives are each described in terms of the basis for the approach, the background and previous work, the work completed already in the context of Virolab, the plans for the remaining duration of the project and the nature of the results expected and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

2 Overview & Objectives

The stated objective of the Virolab project is to develop a virtual laboratory for infectious diseases that facilitates medical knowledge discovery and decision support for HIV drug resistance in particular. This virtual laboratory for European researchers and medical doctors will function as a user-friendly, rule-

based decision-support system for HIV drug-resistance testing and treatment and reliably predict drug susceptibility and virological response while providing researchers with a support environment to study trends in HIV resistance on individual and epidemiological levels.

As described in the Virolab Description of Work [39], Work Package 4 is concerned with the virtualisation and enhancement of the state-of-the-art genotypic resistance interpretation tools and their integration into the virtual laboratory.

The virtualisation and automation of Retrogram were reported in Deliverable 4.1 [38]. In the current deliverable we report on ongoing work on enhancing the Virolab HIV drug-susceptibility interpretation system and plans for continuing it for the remainder of the project.

Whereas the virtualisation and automation phase was characterised by the development of tools and structures within which existing data and knowledge could be integrated, the present phase is more ambitious. We have embarked on a number of modest research initiatives to explore novel ways of providing data, evidence and knowledge for enhancing the Virolab HIV drug-susceptibility interpretation system and to evaluate their feasibility.

The mechanisms of HIV constitute a multiscale problem, spanning temporal and spatial scales from the level of the dynamics of DNA to the epidemiological spreading of a disease through a human population. The state of current knowledge is such that we have merely windows into this big picture. However, the information which we can glean from looking through each of these windows can be potentially useful for generating knowledge which can be expressed as rules in our decision support system. One small piece of data gained from a low-level molecular dynamics simulation, for example, can complement a statistic calculated from a large-scale clinical study. The fact that there are still large gaps in our knowledge of the working of HIV should not deter us from applying all of the knowledge and evidence we already have or are in a position to generate.

To this end we have undertaken a number of research initiatives under the aegis of the Virolab project based on the experience and skills of the project partners and international associate partners. International collaborations of the University of Amsterdam in Singapore and Saint Petersburg have been intensified with the context of Virolab to carry out this research. This has allowed us to broaden the scope of the work carried out in Virolab within the existing budget and shift the emphasis of the project from technology to simulation and mod-

elling as suggested the reviewers by the first annual project review. We feel that this emphasis on modelling and simulation distinguishes Virolab from other such projects.

These initiatives have often an exploratory, curiosity-driven character. The intention is not to attempt any kind of coverage of the potential research areas but to provide some insights which can be used to enhance the decision-making process.

As a case in point, entry inhibitors, such as enfuvirtide, have only relatively recently been licensed for use. To date, there is very little statistical clinical information available about their effectiveness against acute infection, let alone chronic infection. For this reason we anticipate that any information which we can gain from our work on simulation of cell entry and its inhibition will be extremely valuable.

Our research initiatives contribute to enhancing the decision support system in two ways: those enhancements which are based on knowledge, insights, data or evidence which has been gained from research initiatives in modelling and simulation of biological processes, and; enhancements internal to the Virolab HIV drug-susceptibility interpretation system itself such as providing an improved rule language and mechanisms for the combination of evidence from diverse sources.

In this deliverable we describe:

- the ways in which these research initiatives will provide information in the form of data, evidence and knowledge for the Virolab HIV drug-susceptibility interpretation system,
- the ways in which this information will be combined to provide coherent judgements on drug-susceptibility, and
- the envisaged modes of interaction between the Virolab HIV drug-susceptibility interpretation system and the computational experiments running at the various research initiatives.

The research initiatives aimed at enhancing the Virolab HIV drug-susceptibility interpretation system itself are:

- Enhancements of rule-based interpretation.
- Bayesian combination of evidence.

Beginning with the lowest level and moving upwards, the research initiatives which will provide input for the Virolab HIV drug-susceptibility interpretation system are as follows:

- Molecular dynamics of interactions between inhibitors and viral proteins.
- Building a kinetic model for HIV protease cleavage.
- Elucidating selection forces that act on the pathogenic potential of HIV.
- Modelling entry of HIV into target cells.
- Hybrid multi-agent modelling of lymph node cell movement.
- Information-theoretic measures of genetic distance.
- Complex-network models for HIV spreading.
- Semi-automated Literature Mining.

As is shown in Figure 2.1, results from the research initiatives at each level can serve to support the decision making process by hierarchical statistical inferencing which spans these levels.

In the sequel we describe each of these initiatives in detail under the headings:

1. Basis for the approach
2. Background & previous work
3. Work & publications completed in the context of Virolab
4. Plans for the remainder of the Virolab project
5. Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

In a couple of cases, the nature of the research initiative is exploratory and, as such, the results are too preliminary for it to be clear how they will be used to enhance the Virolab HIV drug-susceptibility interpretation system.

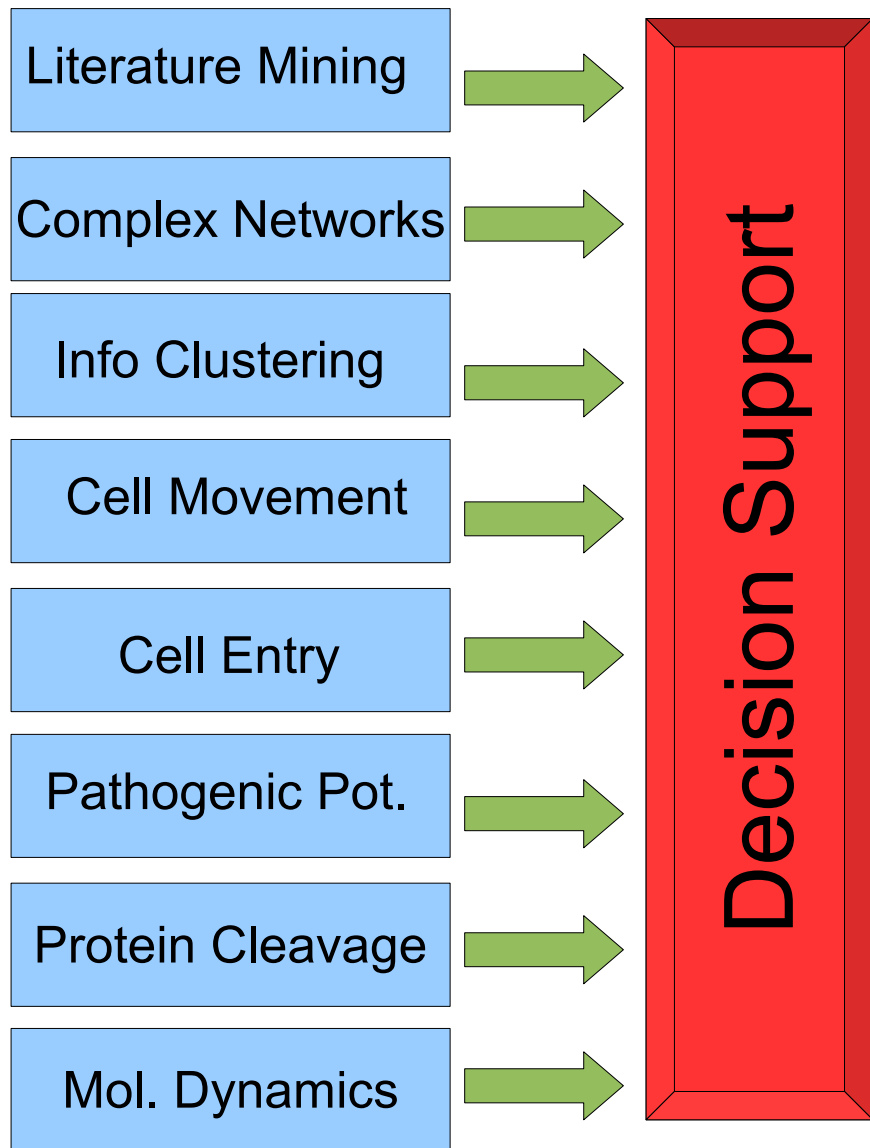


Figure 2.1: Research initiatives at all levels can support decision making.

3 Molecular dynamics of interactions between inhibitors and viral proteins.

This work is carried out by S. Kashif Sadiq, Ileana Stoica, Stefan Zasada, David Wright and Peter V. Coveney at University College London.

It is aimed at incorporating computational data at the molecular level into the decision-support system. The task relies ultimately on the capability of molecular simulations to achieve an accurate ranking of drug binding affinities on clinically relevant time scales.

3.1 Basis for the approach

The development of drug resistance by the human immuno-deficiency virus (HIV) continues to be a major problem in the treatment of AIDS. While several treatment regimens such as the highly active antiretroviral treatment (HAART) have been devised, involving inhibitors that target multiple viral proteins[11], emergence of mutations in these proteins is a contributing factor to the eventual failure of treatment.

While genotypic assaying of individuals infected with HIV is a standard procedure implemented to obtain portions of the viral sequence, the interpretation of such information, given the complexity of emergent mutational patterns, often means that clinicians have to resort to decision support software for assistance.

Incorporating computational data at the molecular level into such a decision support system would prove invaluable for patient-specific medical treatment. This is a challenging task, relying ultimately on the capability of molecular simulations to achieve an accurate ranking of drug binding affinities on clinically relevant time scales.

3.2 Background & previous work

The aim of Virolab is to develop a virtual laboratory which functions as a user-friendly, rule-based decision support system for the treatment of HIV drug resistance, and which reliably predicts patient-specific drug susceptibility and viro-

logical response[15]. UCL's contribution to Virolab is targeted at enhancing its Drug Ranking component. We are employing large scale molecular dynamics (MD) techniques using grid resources in order to study the interactions between inhibitors and viral proteins in atomic detail, and are using free energy methods to predict the effect of mutations on drug binding affinities.

3.3 Work & publications completed in the context of Virolab

Assessment of MM/PBSA capabilities for an accurate ranking of ligand binding energies in HIV-1 proteases

In a recently completed study we investigate the potential of the MM/PBSA methodology applied to binding affinity ranking for the inhibitor saquinavir and the wild-type (WT) and resistant variants of HIV-1 protease: L90M, G48V and G48V/L90M. L90M is a particularly important mutation as it is clinically associated with resistance to nearly all approved inhibitors, yet it does not lie in the active site[48]. G48V lies in the flaps and is found predominantly in association with L90M in response to saquinavir therapy[48].

We use MD and MM/PBSA to determine the structural and energetic determinants of binding, as a way to understand the mechanisms of emergence of the L90M and G48V mutations under chemotherapeutic pressure. We perform 10ns of fully unrestrained MD simulations for each protease-drug system, followed by ligand binding free energy calculations. By explicitly accounting for changes in solute entropy upon binding in addition to enthalpy, we are able to obtain a remarkable level of correlation with absolute experimental binding affinities, as well as to successfully rank the binding strengths of the inhibitor to the mutants versus the WT. Furthermore, we are able to provide a decomposed energetic description for the basis of observed drug resistance and to identify the effect of mutations on the mechanisms of enthalpy/entropy compensation.

Free energy ranking of viral fitness in saquinavir-bound HIV-1 proteases

Ultimately, it is the overall viral fitness of a particular sequence that directs its persistence in vivo. The emergence of mutations in HIV proteins in response to chemotherapeutic pressure is indicative of an interplay between drug binding and the binding of the natural substrates[56]. A complete description of

drug resistance must therefore incorporate the effects of a mutation on catalytic efficiency. Using the recently reported (11) crystal structure of inactive HIV-1 protease to the rate limiting gag polyprotein substrate Nc-p1, we study the differential effect of the L90M, G48V and G48V/L90M mutations on substrate and inhibitor binding. We develop a metric, the free energy potential of viral fitness V_f , solely in terms of the free energy differences of substrate and inhibitor binding and thus computable by molecular simulations. Such a metric correctly ranks the viral fitness of the selected HIV-1 proteases under chemotherapeutic pressure, as well as explains the emergence of mutational patterns such as L90M followed by G48V in response to treatment with saquinavir. A manuscript based on these results is in preparation for publication.

The Binding Affinity Calculator (BAC)

To perform atomically detailed computations under conditions of optimal efficiency, we have designed a distributed, grid-based, automated, binding affinity calculator (BAC). The BAC engine integrates and automates all the steps of the drug ranking process: the parameterisation and modelling of the protein-inhibitor or protein-substrate complexes, the preparation for MD production, the actual MD simulations using the highly scalable NAMD package, and the binding affinity calculations via the MM/PBSA module implemented in AMBER. The BAC makes use of the Application Hosting Environment — AHE[57] to run multiple molecular dynamics simulations using grid resources. To successfully integrate BAC with ViroLab, we expect to have access to EU grid infrastructure (specifically here DEISA), but there is no route available to access it. Therefore, we are making use of UK national resources (NGS) and even TeraGrid (USA) to be able to deliver. The EU must make its own infrastructure much more integrated with its funded R&D projects, and urgently.

Using the AHE and grid resources; integration with the ViroLab GSEngine

In collaboration with the ViroLab GSEngine developers from Cyfronet (partner 1), we have integrated our Application Hosting Environment with the GSEngine, so that applications can be launched on remote Grid resources using the GSEngine workflow development and execution environment. The AHE is a lightweight hosting environment for running unmodified applications — NAMD,

LB3D, LAMMPS, DL_POLY among others, on grid resources such as the UK NGS and EU DEISA grids. The AHE wraps legacy applications as web services, allowing users to interact with these applications via a web services interface. In the case of the ViroLab integration, the Ruby based GSEngine uses the AHE Java client API via JRuby to make calls to an AHE server. The AHE hides the details of running the application from the end user, interfacing with a number of different middleware solutions such as Globus and UNICORE. Within the ViroLab project we envisage grid based parallel applications such as NAMD will be represented as GSObjects in the GSEngine system, with the AHE will be used as a launching mechanism to manage the execution of the application on a wide variety of remote grid resources. This will provide a high level of abstraction for users wishing to incorporate such applications in their ViroLab experiments, for example using molecular level HIV/inhibitor modelling in support of the ViroLab drug ranking expert system. This work is undergoing further development by our partners at Cyfronet.

3.4 Plans for the remainder of the Virolab project

The strength of our approach is its ability to offer insight into resistance and the structure-affinity relationship at the molecular level, with a moderate computational cost and over clinically relevant timescales (~1 week). We are currently working to improve the reliability and robustness of the modelling and drug ranking protocols, thereby enlarging their applicability to a wider range of HIV-1 protease inhibitors and associated mutations. Furthermore, work is in progress in our group towards extending the use of the drug ranking protocol to the HIV reverse transcriptase and its NNRTI-s.

3.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

Once tested and optimised, the BAC application will be integrated with Virolab's Work Package 4 (Enhancement) and used to complement the existing components of WP4. The binding affinity component should prove an invaluable aid in understanding the molecular basis of drug resistance and in predicting patient-

specific responses to treatment.

4 Building a kinetic model for HIV protease cleavage.

This research initiative is being carried out by Viktor Müller and his collaborators at ELTE.

The plans involve a large-scale sequence analysis, of which the pilot has already been done, to show that the basal rate of HIV evolution may be much slower than generally thought, and a simulation model to demonstrate that this can be explained if continuous chains of transmission occur primarily in long-lived infected cells.

In collaboration with Carlo Torti (Partner 6, Università Degli Studi di Brescia) we will analyze their long-term clinical data (from the Italian MASTER cohort) to see whether there has been a time trend in the virulence (severity) of untreated infections since the start of the cohort.

ELTE is considering, together with Gökhan Ertalyan of the University of Amsterdam, to seek an explanation for the switch from R5 to X4 virus during HIV disease progression on the basis of the local spatial structure of infections. This would be primarily Ertalyan's project but Müller has contributed to its design and may to some extent supervise it.

4.1 Basis for the approach

We are building and analysing a reaction kinetics model of the cleavage of viral polyproteins by the HIV protease enzyme, which is a key step in the maturation of virus particles. We will prepare two implementations of the reaction system: a standard deterministic ODE model, and a stochastic simulation model. The main goal is to simulate the complex cleavage process to understand its role in the natural life cycle of the virus and to expose key reactions that may offer vulnerabilities for drug development or contribute to our understanding of existing drug resistance mutations.

4.2 Background & previous work

The proteins that make up a HIV virion are first produced in the form of polyproteins that have to be cleaved by the viral protease. This cleavage is essential for the maturation of virus particles, and blocking it by protease inhibitors has been a major route of therapy. The nature and complexity of the process call for a systems biology approach. There are 11 cleavage sites in the Gag and Gag-Pol polyproteins, and a full description of all reactions (including formation of complexes) involves more than 200 equations. A fascinating complication is that the protease itself is part of the Gag-Pol polyprotein, and therefore it catalyzes its own liberation. The process is started by the weak autocatalytic cleavage of the polyprotein. This generates interesting kinetics with a slow start then acceleration until all protease molecules are freed. A kinetic understanding of cleavage is essential for understanding the complex evolution of drug resistance against protease inhibitors. For instance, primary resistance mutations in the protease reduce the efficiency of processing, but this can be restored partially by compensatory mutations in the cleavage sites.

A qualitative description of the cleavage reactions has been compiled by previous work: key reactions, substrates and cleavage products are all known. However, a quantitative description is still lacking, and building a kinetic model will attempt just that. For the analysis we will comb the literature to find experimental data on the kinetic constants (rate parameters) of the reactions. Our work may pinpoint which measurements are still needed to parameterise the model.

4.3 Work & publications completed in the context of Virolab

We have formulated a deterministic model of the cleavage process involving the initial substrates, the major end products (functional virus proteins) and the intermediate substrates leading to the latter. The equation system has been implemented in SBML (Systems Biology Mark-up Language) format, which can be used as input to several analysis tools and will later facilitate dissemination in a widely used standard of biological modelling. We plan to analyse the system with the open-access software COPASI. We have started to mine the literature for estimations of kinetic constants, and will start the numerical analysis when all data have been collected.

4.4 Plans for the remainder of the Virolab project

The analysis of the system is still ahead of us. A literature study is assisted by a student and will take a few months to complete. A stochastic implementation of the cleavage reactions will also be implemented. This is justified because the relatively small number of molecules within a virion may render deterministic descriptions inappropriate. We will compare the behaviour of both implementations.

An important synergy within the project is that once we have compiled estimations on the cleavage rates, we are going to compare these with the binding affinity constants estimated by UCL (Partner 11) within Virolab. If a straightforward relationship can be established between cleavage rates and binding affinity, then molecular dynamics computational prediction of the latter would allow us to estimate missing kinetic rates, which could vastly improve the precision of the models.

4.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

A guaranteed result of this work is to test whether currently available kinetic data are sufficient for a quantitative analysis of the cleavage process. Lack of data may limit the applicability of the model, but even in this case the analysis would pinpoint the most important cleavage rates that should be measured to improve our understanding and complement the system. There is hope that missing key data can be obtained still within the time frame of Virolab through collaboration with experimental groups outside the consortium.

Given sufficient data, the model will allow us to identify rate-limiting steps in the cleavage process, which can guide drug development toward key vulnerabilities of the virus. Furthermore, if the action of drugs can be associated with individual cleavage reactions, then the model will be able to predict non-trivial drug interactions. Similarly, if the effect of drug resistance mutations can be interpreted in the parameterisation of the model, then gene interactions (epistasis) between different mutations will also be predicted by the model, which enhances drug susceptibility interpretation.

5 Elucidating selection forces that act on the pathogenic potential of HIV.

This research initiative is being carried out by Viktor Müller and his collaborators at ELTE. The plans involve ‘enhancing’ our general understanding of HIV, and do not contribute to drug susceptibility interpretation directly.

5.1 Basis for the approach

We are building ODE and stochastic simulation models to explore the ability of HIV to induce a chronic hyperactivation of the immune system. The main goal is to elucidate whether this viral characteristic is under selection at the level of within-host competition between virus strains.

5.2 Background & previous work

Accumulating evidence indicates that the pathogenesis of HIV infection is linked to the ability of the virus to induce a chronic generalized hyperactivation of the immune system. Remarkably, immune hyperactivation is absent in African primates naturally infected with related simian immunodeficiency viruses (SIV), and these infections are indeed largely nonpathogenic. This raises the question whether immune activation is an evolved (adaptive) trait of HIV or an “unwanted” side-effect of jumping to a new host species.

Activating immune cells may benefit the virus by increasing its supply of susceptible target cells. HIV replicates primarily in activated CD4+ T-lymphocytes, while quiescent cells that form the majority of this cell population are largely resistant to infection. Activating CD4+ T cells thus seems to be clearly beneficial for the virus, and indeed these cells are subject to chronic hyperactivation, which becomes gradually stronger as the disease progresses. However, the paradox still remains why this generalized activation is absent from natural SIV infections.

We propose that the key to resolve this paradox lies in the spatial structure of HIV/SIV infections. Using the standard modelling framework of HIV dynamics, we demonstrate that in a single well-mixed compartment of infections, mutations

affecting the ability to activate target cells are selectively neutral for the virus. However, HIV infects cells and is produced primarily in the lymphoid tissues, which have a highly organized spatial structure. Individual lymph follicles can be infected with different virus strains and may serve as relatively isolated local sites for several rounds of virus replication. This gives rise to a multi-compartment system connected by systemic lymph and blood circulation. Using simulation models we show that given such spatial structure, the ability to activate target cells can spread in the virus population. We thus demonstrate that local, but not global immune activation is selectively advantageous for the viruses, and the latter may therefore be lost during co-evolution with the host species.

5.3 Work & publications completed in the context of Virolab

This work is nearly finished and we expect to submit the manuscript to a high-impact, peer-reviewed journal in the coming months. We have developed and explored several variants for both the systemic ODE and the local simulation model. We have completed numerical analysis of the final ODE models, and the simulations are also nearly finished.

5.4 Plans for the remainder of the Virolab project

In the last stage of the work, we need to finalize the parameters of the models and complete the final simulation runs to be presented in the scientific paper. The manuscript is 50% completed; it needs to be finished and finalized.

5.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

Our results suggest that the ability of the virus to induce chronic systemic immune activation, which may largely be responsible for disease, is probably not directly under selection. Rather, it may be a side-effect of selection acting on immune activation that occurs locally at the microscopic sites of infection and there improves target-cell supply of the virus. This result improves our understanding of the selection forces that act on the pathogenic potential of HIV. Nonpathogenic

SIV infections suggest that systemic and local immune activation can be decoupled, and this may be a potential route also for the future evolution of HIV.

6 Modelling entry of HIV into target cells.

This work, carried out by Gökhan Ertalyan and Peter M. A. Sloot at the University of Amsterdam, is aimed at understanding how the dynamics of coreceptor usage is governed.

6.1 Basis for the approach

HIV entry into its target cell is mediated by a multi-step process. HIV-1 uses CD4 receptor as its primary receptor to gain entry. Binding to CD4 induces conformational changes in the viral envelope protein that leads to the engagement of the one of the chemokine co-receptors CCR5 or CXCR4. Early infection with human immunodeficiency virus is characterized by the predominance of CCR5-tropic (R5) virus. However, over the course of infection CXCR4-tropic (X4) virus appears in the later stage of the infection in approximately 50% of the infected individuals and usually precedes an accelerated CD4⁺ T cell depletion with rapid disease progression. The reason for this phenotypic switch and effects on the disease progression is still not clear.

Several CCR5 coreceptor antagonists are currently being developed for their potential use for patients infected with human immunodeficiency virus type 1 (HIV-1). They are also under consideration as topical microbicides to prevent HIV-1 sexual transmission[31]. There are potential problems associated with the treatment with CCR5 antagonists. For example, use of this therapy may lead to emergence of X4 strains, which could accelerate disease progression[20]. Furthermore, CCR5 inhibition could interfere with the normal immune and inflammatory responses. Despite the apparent normal phenotype of CCR5 Δ 32 homozygotic individuals, it is not clear how interference with CCR5 will affect the already impaired immune system in HIV-infected patients. Thus, many questions needed to be answered including “How the dynamics of the coreceptor usage is governed?” before CCR5 inhibitors are safely used in humans[33].

6.2 Background & previous work

There are three hypothesis aiming to explain the phenomenon mentioned above:

The first hypothesis tries to explain the predominance of R5 virus via the selection of R5 virus during transmission and the coreceptor switch via the results of evolution in the viral population. This hypothesis is called “Transmission & Mutation” hypothesis[42].

The second one is called the “Immune-Control” hypothesis which is based on the assumption that X4 viruses are better recognized by the immune system and therefore have a higher clearance rate than R5 viruses. The later emergence of the X4 viruses are the cause of immune system erosion during the infection[54, 36].

The third hypothesis is “Target-Cell-based” hypothesis and focuses on the differential target cell preferences of R5 and X4 viruses and the effects of this preference on the cell population level. This hypothesis emphasizes the activation of the immune system on cell population level and effects of different coreceptor designation and lifetime of various cell types.

To test the hypothesis mentioned above several models have been proposed [42, 43, 58]. Although the models shown to be useful for simulating the qualitative behaviour, their applicability to experimental data remained limited due to the innate properties of the modeling scheme used as well as the undetermined dynamical parameters.

To investigate the dynamics regarding coreceptor usage in HIV-1 infection, we are designing an agent based model with the emphasis on spatial interactions between various immune system cell types (naïve and memory T cells, helper T cells and B cells) and R5, X4 and R5X4 tropic HIV-1 strains.

In the current version immune cell types, virus strains and proliferation rates are defined from a rather limited parameter set. Therefore, currently it is qualitative and serves as the test-bed to verify our modeling framework.

6.3 Work & publications completed in the context of Virolab

This work is based on earlier work reported in [19] and [16] and also on work performed at the University of Amsterdam reported in [49].

6.4 Plans for the remainder of the Virolab project

After verifying the modeling framework with qualitative behavior we are going to address temporal aspects.

Setting the time

During the course of HIV infection, acute and chronic phases, each stands for different time scales from hours to years. On the cell level, cell entry of the virions, the average time between homeostatic divisions of T cells and the lifespan of memory T cells all correspond to different time scales such as hours, days and months respectively. So, setting the time scale is a crucial step if one wants to have a quantitative model of the immune system dynamics.

Parameter gathering & estimation

Literature overview of appropriate parameters and parameter estimation of the unavailable parameters are planned before testing different hypotheses.

Testing for the hypotheses

The hypotheses mentioned above are going to be created in the modeling framework and simulations within appropriate parameter ranges are going to be conducted.

Analysis of the results

Results of the simulations are going to be analyzed for future implications.

6.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

HIV entry inhibitors are a new class of drugs that will be widely available in the near future. Although we know that they are aiming for restricting virus target-cell coreceptor availability, we do not know the outcomes of this approach on the immune-system level.

HIV target cell entry model aims to be the theoretical framework which can be used to test various patient-inhibitor drug scenarios and/or refine the suggested therapy. Therefore complementing overall decision support system of Virolab.

7 Hybrid multi-agent modelling of lymph node cell movement.

This work is carried out by Tay Joc Cing and Narges Zarrabi at Nanyang Technological University in Singapore in collaboration with Gökhan Ertaylan and Peter M. A. Sloot at the University of Amsterdam.

7.1 Basis for the approach

The study of cellular properties and molecular dynamics in order to understand the emergent dynamics of cell interactions is an important fundamental aspect of modeling the human immune system. There are mathematical and hybrid multi-agent models available that try to represent the migration of cells through local forces involving diffusion, blood flow and chemotaxis, as well as lymphoid recirculations. We adopt a hybrid multi-agent modeling approach to extend the current model of cell motility to model the operation of an entire lymph node. Such a model will enable us to estimate the amount of viral load and infected cells throughout lymphatic compartments. In this research initiative, we describe the motivations and research challenges for developing an efficient, multi-scale model of HIV-1 drug treatment with a focus on the microbiology of defenses within a lymph node. We propose a high-level architecture for modeling a lymph node that is based on a cell movement model at the cell population level and a cell interactions model at the protein level.

The rapid spread of HIV and AIDS throughout the world has resulted in a global need for a vaccine that can stem HIV progression and prevent AIDS. At present there have been some successes in developing therapeutic drugs which can keep the amount of virions under a certain threshold for about 10 to 12 years from the start of infection. However, none of these drugs are able to completely or permanently suppress the viral pathogenicity. It is known that the amount of

HIV viral RNA (viral load) and CD4+ T cell count in the blood stream or lymphatic sites such as the lymph nodes are two important biomarkers in measuring the progress of HIV infection and of quantifying the onset of AIDS. In HIV infected patients, AIDS is considered to occur once the T cells count falls below a certain threshold value (200/ml).

The purpose of our research is to develop an agent-based computer simulation for monitoring and measuring the efficacy of drug treatments. This simulation is valuable since its results can be used in the decision support system to help medical practitioners and healthcare policymakers to prescribe a suitable therapeutic drug regime for HIV infected patients that would minimize the chance of developing anti-drug resistant strains while maximizing the lifespan of the patient. Such a simulation will also be used to predict the effect of existing HIV drugs and inhibitors in vitro and for designing new drugs that can inhibit HIV progression. These purposes must primarily be supported by an estimation of viral load and the levels of activated lymphocytes in the human body (during the therapy). Since 98% of all lymphocytes are aggregated within lymph nodes, it is therefore essential that we focus our efforts on modeling the structure, cell motilities and cell-antigen reactions within the lymph node.

7.2 Background & previous work

We have established in earlier work (CAFISS)[26], the feasibility of developing models that can elucidate the nature of causative agents in HIV-1 infection. The development of CAFISS in this respect was a multi-agent simulation of four commonly held HIV-1 infection hypothesis; namely, direct CD4+ infection, Rapid Mutation, Syncytium Formation and CD4 Receptor Filling. With only a handful of lymphocytes in the model, we were able to simulate and retrodict through in silico experiments that Syncytium Formation and Rapid Mutation were the possible etiological agents for enabling HIV to persist and develop into AIDS. Results were verified through qualitative correlations with the three-stage clinical cycle of HIV. However, CAFISS was limited in its use towards describing the virulence of the incumbent strain, and of how the results would appear under HAART treatment.

In order to simulate the virulence of the HIV-1 strain(s), we would require a microbiological model of how HIV-1 interacts with antigen presenting cells and how lymphocytes such as CD4 cells become infected by HIV-1 virions. This must

be supplemented by an abstraction of the HIV-1 strain that allows each strain to be classified according to its binding affinity and co-receptor requirements. In addition, we would also require a precise model of how cells and virions interact within a lymph node; be it due to a combination of diffusion, blood flow or chemotaxis. With a model of cell motility, we expect to be able to more accurately predict the infection rate per lymphocyte. What follows is then to simulate HIV cell entry (See Section 6p19), which is rather deterministic unlike the replication process.

7.3 Work & publications completed in the context of Virolab

We consider the human immune response to infections, subsequent eradication and return to homeostasis as a complex system. In simpler terms, it is the collective interaction of trillions of lymphocytes and regulating chemicals that serve to produce an emergent defense system against known and unknown pathogens. Such a system is inherently nonlinear and uncertainties abound in how the numerous causative feedforward and feedback links serve to regulate and produce the macroscopic behaviours that are observed. There is a need to consider individual-based models that are sufficiently granular to capture all the necessary characteristics before the simulation can produce useful results. Preliminary work we have done is to try to distinguish the use of equation-based models and agent-based models for immune system simulation[27].

The biological scenario highlights the multi-scale nature of the problem. From the interactions at the molecular level, to the cell and lymphoid compartment levels, there are at least 3–5 orders of difference in magnitude for temporal and spatial scales, leading to significant computational challenges. For example, during the first 60 seconds of T cell – APC contact, the T cell makes its initial activation ‘decision’ to either sustain the interaction or disengage. On the other hand, Cell division takes a few hours. Activated T cells are estimated to divide two to four times every 24 hours, and this process lasts for 3 to 5 days. The lifespan of cells also vary greatly. Lymphocytes (T cells and B cells) in a naïve or memory state are usually long-lived (in the order of years). Our preliminary research in this direction has been the application of event-driven agent-based simulation for B cell clonal selection[30] and distributed event-driven simulation of chemotaxis[41]. The differences in scale also mandates that both population

based models and individual based models are jointly necessary, and that simple models based on cellular automata (ie those of Selada and Seiden) are likely to produce highly inaccurate results since communication and cell and chemical motilities (at a lower scale than lymphocyte activation) are not explicitly modeled. We have developed a hybrid agent-based model that combines the use of equation based models to handle molecular quantities and agents to represent lymphocytes[29, 28]. A preliminary survey of HIV-1 drug treatment regimes has also been completed[55].

Finally, at a higher scale, there is recirculation through the lymphatic system, hence structure and organization of how agents and quantities convene or flow is important so as to estimate the 'input' rate for the lymph node simulation rather than employ stochastic input distribution models. The role of chemoattractant gradients in the movement of lymphocytes within the lymph node is hypothesized but not yet demonstrated. Recently, a new experimental system based on slices of a lymph node have been used to show that T cell motility within lymph nodes strongly depends on CCR7 chemokine receptor and its ligands CCL19 and CCL21. Both ligands CCL19 and CCL21 are present in the high endothelia venules (HEVs) and T cell zones of the lymph node. These ligands are essential for effective migration of T and B cells across HEVs into secondary lymphoid organs. It has been reported that the lack of CCL19 and CCL27 expression in lymphoid organs result in defective T cell trafficking into lymph nodes and impede lymphocyte homing. We are currently also investigating architectures for recirculation as well as theoretical models of the dynamics observed in the lymphatic system. Often times, a singular modeling approach does not suffice and some combination of network[53], equation and agent-based models may be needed. The proposed architecture comprises two layers:

Cell population level defines a population of cells motility and encountering mainly within the lymph node. Therefore we have completely studied the movement of cells inside the lymph node. Common properties of cells such as motility will be inherited from previous simulations on modeling movement of cells under environmental effects.

Protein Level describes interaction among different cells or cells and virions. These interactions will lead to cell activation, infection, and replication and finally producing new viruses. Only compartments which have significant

role in the binding and entry process will be modeled in this level. Protein level increases the model granularity. To increase the model accuracy we can also integrate binding affinity calculations within our hybrid agent-based models.

7.4 Plans for the remainder of the Virolab project

We have developed a Hybrid Multi-Agent model of locomotion of a population of cells toward a source of chemoattractants that employs a distributed event simulator as well as a simulation model of HIV replication process[37] within a cell. The immediate future work is to complete each of these models and integrate them together using a recirculation model so as to achieve a more accurate model of cells and molecules dynamics interacting in the lymph node. Specifically, these works include:

- Extending the chemotaxis model to simulate lymphocyte recirculation, and applying the hybrid scheme to build agent rules for simulating an immune response to HIV pathogenesis.
- Building the agent-based model at the protein level to include the function of all compartments involved in HIV entry and replication. It is then possible to test the effect of existing drugs on the model particularly entry inhibitors. The simulation results may also give insights to design more therapeutic regimes for HIV infection.
- Modeling the operation of a single lymph node based on multi-agent modeling methodology. This model can be considered as an independent class or a 'black box'. By object-oriented encapsulation of detailed functionality within composite classes, we can effectively mirror the biological structure and develop a system capable of estimating viral load and lymphocyte counts in lymph nodes and effectively measure the efficacy of drug treatments.

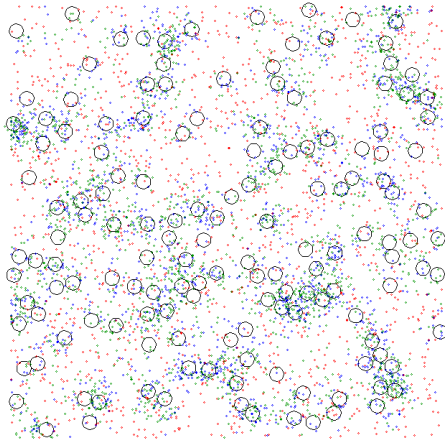


Figure 7.1: Simulation of LPS effect on TNF and IL10

7.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

The main contribution of the work are the frameworks and methodological considerations for designing the type of agent-based model for immune system simulation. The latter domain being disparate in functionality and complexity, our framework has therefore focused on the primary mechanisms of immune response which are broadly based on cell motility and inter-cell signaling.

As a fundamental requirement, all models of microbiology must address the multi-scales of time and space. We have developed a discrete-event scheduler that facilitates this and demonstrates its efficacy and efficiency based on a B-cell clonal expansion model. We use this mechanism to model the effect of chemotaxis on cell motility in a model of how lymphoblastic cells move within the lymph node. Using just discrete event simulation alone is insufficient for handling large numbers of molecules and cells modeled as individual agents. Hence, we have developed a hybrid approach to simulate the effect of diffusing and forced flow quantities over a mesh, on agents moving in a continuous space superimposed on this mesh. This modeling approach potentially allows one to add logical rules to each agent (cell) that can encode intra-cellular mechanisms (like protein production leading to receptor recycling).

Another type of immune system model we have created is one based on multi-threaded agents persisting on a mesh of cells. These agents are asynchronously updated using a uniform time-step. Although this model is less effec-

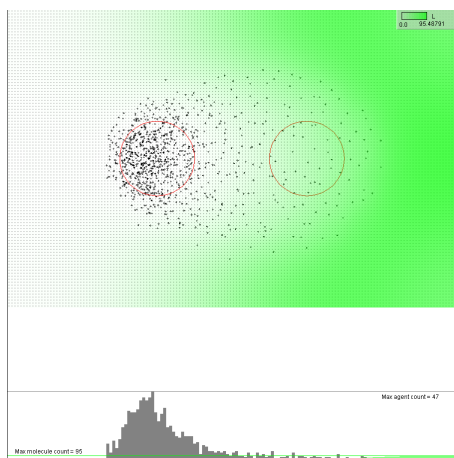


Figure 7.2: Chemotaxis Model based on Receptor Kinetics

tive in modeling large scale biological systems, it is nonetheless useful when the quantities of cells and molecules are small or can be rescaled only to observe a small subset (like when the cells and molecules are distributed randomly and uniformly). More recent progress in using these frameworks are for the simulation of airborne viral transmissions in a hospital environment, and for simulating the effect of lipopolysaccharides (LPS) on cytokines TNF and IL10.

8 Information-theoretic measures of genetic distance.

This work is carried out by Max Filatov, Breannán Ó Nualláin and Peter M. A. Sloot at the University of Amsterdam.

The aim of this research initiative is to apply information-theoretic measures of similarity to HIV sequences in order to classify them according to their resistance to antiretroviral drugs.

We develop a novel hierarchical clustering method founded on the Kolmogorov complexity-based Universal Similarity Metric[34].

8.1 Basis for the approach

In all fields, in an attempt at understanding, our initial and pervading attempt at organising the objects under study is to classify them, separating different objects and collecting similar ones into groups. This methodology has a long history in the biological sciences since the eighteenth century when Linnaeus classified organisms hierarchically. The field of phylogenetics applies classification methods in order to elucidate evolutionary relations among various groups of organisms.

At the genetic level, the standard approach to classification is to generate the pairwise similarities of a collection of organisms applying a (possibly weighted) distance measure to parts of the genomes of the organisms (after possible initial alignment). These distances are then used to generate a phylogenetic tree using parsimony, distance or maximum likelihood methods.

A whole plethora of such methods is available, together with software packages for performing the calculations and generating the trees. [3]

Genetic subtyping done by e.g. bootscanning, or direct phylogenetic analysis as in the Rega HIV Subtyping Tool.[17]

8.2 Background & previous work

Two goals can be distinguished when attempting to classify organisms. The cladistic goal is to arrange the HIV sequences only by their order of branching in an evolutionary tree and not by their morphological similarity. This approach can allow us to make inferences concerning the spread of the disease and the way in which it entered the human population.

In contrast, our goal is to classify HIV sequences, not based on their ancestry but based on their phenotype of resistance to various antiretroviral drugs. The underlying assumption is that sequences which are similar in nature will lead to similar resistance to drugs.

There are two components to our clustering approach. The first is the use of the Universal Similarity Metric, and information-theoretic measure based on the Kolmogorov complexity or compressibility of the sequences.[34]

The second is the use of a novel hierarchical agglomerative clustering method based on these similarities. All existing clustering methods are founded on bi-

nary similarity metrics and these have some limitations[18]. Trees are built from the pairwise similarity of the underlying sequences. The Universal Similarity Metric however lends itself to making not only two-way comparisons among sequences, but also three-way, four-way and many-way comparisons. This similarity hypermetric leads to a complete set of similarity data not only over the binary cartesian product of the set of sequences but over the entire power set. We will investigate whether exploring this entire power set can lead to a novel clustering technique.

Previously Krasnogor et al. [32] applied USM method with binary similarities and standard binary clustering techniques to small data sets of simple contact maps and achieved good results for protein structure comparison. However Rocha et al. [45] found these techniques to be less effective than other protein structure comparison methods on a larger, representative data set.

Our hope is that our novel clustering technique will make this approach viable for predicting HIV drug resistance.

8.3 Work & publications completed in the context of Virolab

This is a new research initiative in the Virolab project.

8.4 Plans for the remainder of the Virolab project

As his Masters thesis project, Max Filatov will perform a proof of concept study to implement and investigate the proposed, novel clustering method, experimentally tune the similarity hypermetric and test its effectiveness on datasets of HIV sequences with known resistances to antiretroviral drugs.

8.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

This research initiative is very exploratory. Should it bear fruits, the computational tools developed can be used to perform further, more thorough testing before the results can be incorporated into the Virolab HIV drug-susceptibility interpretation system.

9 Complex-network models for HIV spreading.

This work is carried out by Alexander Boukhanovsky and collaborators at the Saint Petersburg State University and Peter M. A. Sloot at the University of Amsterdam.

Despite the availability of a large number of mathematical models describing the spreading of HIV, a good understanding of the spreading dynamics through numerical analyses is still a major challenge. It is essential to combine epidemiological processes with sociometric models and network sciences. Many mathematical models have been suggested to simulate HIV population dynamics. For instance statistical techniques, like the back-calculation method and its related modifications, are widely used to estimate the incidence and short-term projection of HIV. Generally these methods are based on information from annual AIDS cases and incubation periods of the disease[52]. Popular epidemiological models like SIR models are often used to simulate HIV spreading, and new ways to account for homogeneous mixing and the impact for instance demographic effects or drug resistance have been discussed[12].

The goal of the approach presented below is to show that complex network based modeling techniques provide a universal and natural way to describe any kind of infection spreading and specifically HIV.

9.1 Basis for the approach

Formal Description of the Network Model

Let us consider a Complex-Network model (CN-model) as a set of the pairs $\langle G, \mathfrak{T} \rangle$, where G is a graph, that is, an ordered pair of disjoint sets (V, E) (vertices and edges), and \mathfrak{T} is an evolutionary operator, governing network changes in discrete time t :

$$\begin{aligned} \langle V, E \rangle_{t+1} &\stackrel{\text{def}}{=} \mathfrak{T} \langle V, E \rangle_t \\ \langle V, E \rangle_{t=0} &\stackrel{\text{def}}{=} \mathfrak{T} \langle V_0, E_0 \rangle \end{aligned} \tag{9.1}$$

The evolutionary operator in Equation 9.1 can be represented as a composition of distinct operators $\mathfrak{T} = \bigotimes_k \mathfrak{T}_k$ corresponding to different dynamical aspects. At each time step Equation 9.1 defines a graph with a casual ratio of the fraction of susceptible, infected and removed individuals. Generally, the interplay between these values is described in the form of a standard epidemic SIR model in terms of a system of differential equations.

Modeling of a Sexual Contact Network

We construct a model as a dynamical scale-free network in respect to Equation 9.1, wherein each individual is represented by a node and the edges are the connections between individuals. The scale-free property implies that network has a power-law degree distribution:

$$P(k) \sim k^{-\gamma}, \quad k \leq k_{\max} \quad (9.2)$$

where k - the number of sexual partners per year and γ - a parameter of the distribution. We shall be using for the homosexual contact network $\gamma_1 = 1.6$ and $k_{\max} = 250 \div 300$. For the heterosexual network we can take $\gamma_2 = 2.7$ and $k_{\max} = 60 \div 70$ [35]. The cut-off of the distribution is very important as it defines the number of superspreaders in the network. With respect to Equation 9.2, at each time step we use a configuration model for contact network generation. This flexible approach is based on the generation of a degree sequence which allows generating links between any two nodes according to its degrees taken from a specific distribution.

9.1.1 Modeling the dynamics of the infected nodes

Taking into account the duration of the incubation period, availability of treatment and the effect of diagnosis the Markov model with fixed number of states can be used. In particular, Aalen et al.[9] proposed the parameterisation of a multi-state Markov model to represent stages of HIV infection and the diagnosis and treatment.

9.1.2 Direct Simulation Algorithm

For simulation of heterosexual spreading the network is represented as a bipartite graph and the transmission probability from men to women is taken twice as efficient. The basic simulation procedure can be written down as a number of consequent steps:

1. Generate a network using a given node degree probability distribution given by Equation 9.2 with an initial number of randomly infected nodes.
2. Infect nodes surrounded by infected nodes for every link (per partner).
3. For each infected node apply a rule of progression from HIV to AIDS. Nodes with AIDS are removing from the network.
4. Apply the demographic rule.
5. Store the current nodes state and generate a new random network.
6. Repeat steps 2. to 5. till statistical significance has been obtained.

9.2 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

AIDS data

The United States data [21] are used for model identification and tuning mainly because they provide relative good statistics of AIDS cases, several kinds of infection spreading and a possibility to explore the effects of treatment. These data include at least three epochs of the epidemic evolution: Pre-ARV, ARV and HAART-treatment; and three kinds of HIV spreading. For each of these epochs and the three kinds of infection distinct HIV spreading behaviour was observed.

Validation: comparison with annual AIDS data

As we can observe from Figure 9.1 the simulation results are very close to the estimation of the officially registered annual AIDS cases. The obvious effect on all the figures shown is a substantial decline of AIDS cases after introduction of HAART (1996). The stable number of AIDS cases in recent years may be explained by the stationary number of HIV infection over a long period of time.

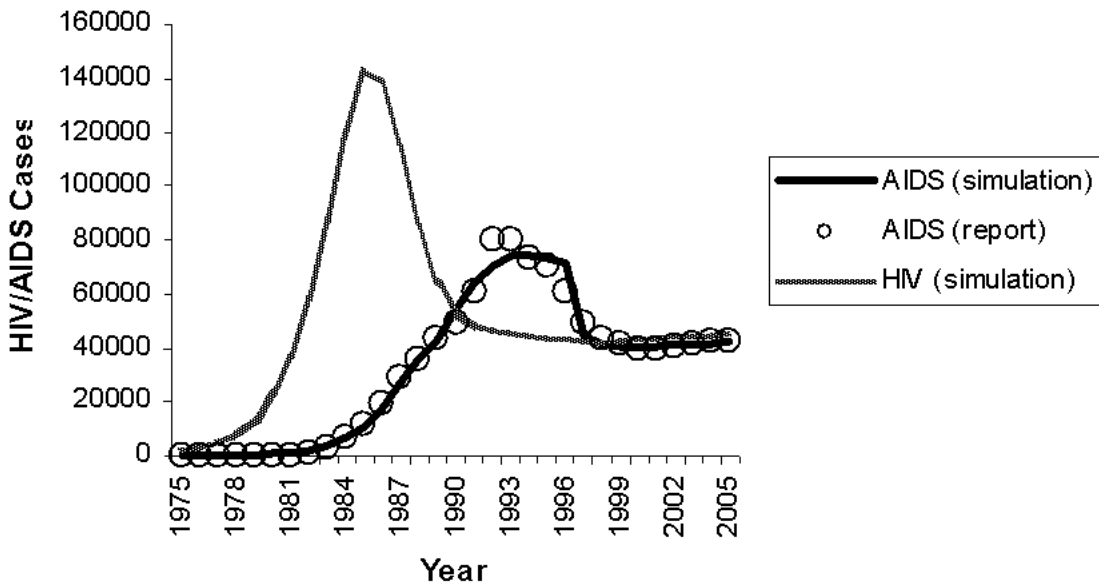


Figure 9.1: Simulation results and reported data for the AIDS epidemic and reconstruction of HIV cases in the USA.

Reconstruction: annual HIV cases

The simulation result of annual HIV cases reconstruction is presented in Figure 9.1. It is remarkable that those estimations are close to official estimation of CDC [21]. Some of the effects that we observe from Figure 9.1 for AIDS cases are present in the curve for HIV as well. For instance, the stationary flow of HIV epidemics over more than a decade reflects in the stationary flow of the number of annual AIDS cases.

9.2.1 Current research

The current research is focused on the experiments with different features of the direct simulation procedure. The key problem is to define the sufficient size and number of experiments of network for desired statistical significance. Direct simulation algorithm described in subsection 9.1.2 has a stochastic nature whereas the real data are not stochastic. It means that the model-based result should be taken as a mean value or median of many experiments with different initial states. On the one hand the number of experiments should be enough for the correct averaging. But on the other hand it should be accomplished at the reasonable time. There are two primary ways to solve this problem. The first one is

to apply parallel computing for those procedures. It allows making the total time several times smaller. Another way is to parameterize the whole process in the form of system of differential equations. In the case of known simulation parameters the direct simulation is more preferable because this result may not contain faults of parameterisation. But commonly the true simulation parameters should be defined as a result of optimization where the correct model based curve of infection cases has to be found for every set of parameters. In this case the number of computations may exceed reasonable value and estimation of parameters for direct simulation procedure is better to do with parameterized model.

The other important feature is the influence of heterogeneity on the shape of the curve of annual infection cases.

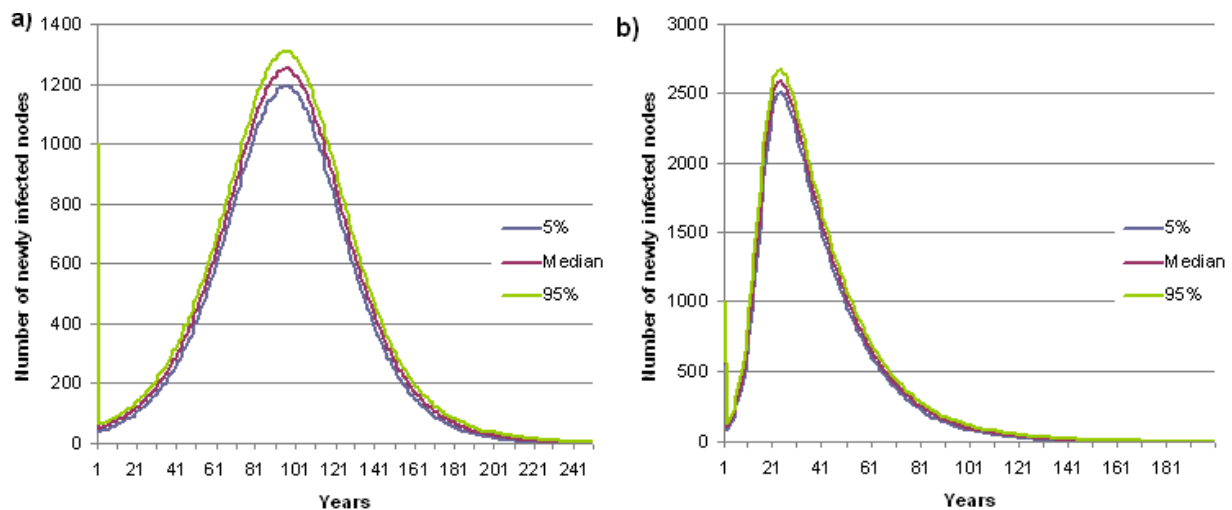


Figure 9.2: Direct simulation results for infection spreading through homogeneous and inhomogeneous networks. Number of vertices is 10^5 , probability of infection is 5%, number of experiments is 1000. a) homogeneous network with one edge at each node; b) scale free network with exponent $\gamma = 2.5$

As we see in Figure 9.2 the network structure has a great influence on the rate of infection spreading and the shape of the curve of annual infection cases. It is interesting that for homogeneous network the rate of infection growth and drop is approximately the same whereas for the scale-free network the rate of growth is noticeably higher.

Additional research is focused on the visualization of the epidemiological networks.

The k -kernel decomposition shown in Figure 9.3 is a useful approach to represent different groups ranged by risk to be infected (or to infect other nodes).

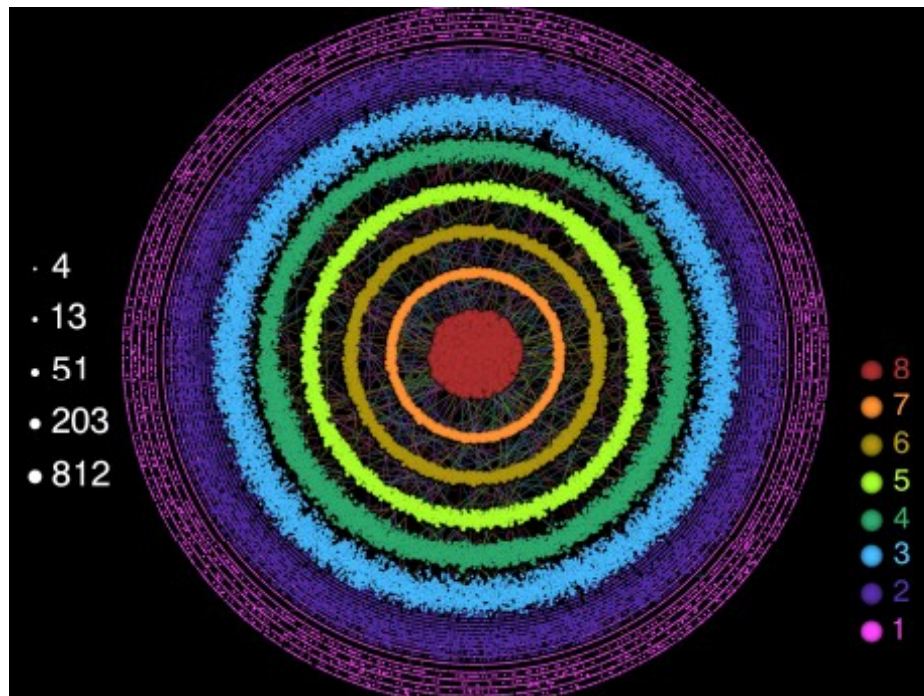


Figure 9.3: The structure of the contact epidemiological network based on the k-kernel decomposition

k-kernel of graph is the most connected subgraph which nodes degree is equal or higher than k . Different risk groups are marked with different colours and belong to the corresponding envelope. The envelopes with higher number of nodes degree are placed closer the center of the figure.

9.2.2 Conclusion

A parameterized CN model describing the dynamics of HIV spreading is presented. A publication is in press[50]. Homosexual and heterosexual spreading is described by scale-free network, drug users spreading is described with homogeneous mixing inside the exposure group. The experiments show a good correspondence between the model results and real demographic historical epidemiological data. We will include this epidemiological prediction model into the drug decision support system for HIV infections [7, 51].

10 Semi-automated Literature Mining.

This work is carried out by Breannán Ó Nualláin and Peter M. A. Sloot at the University of Amsterdam in collaboration with members of the Adaptive Information Disclosure group of the Virtual Laboratory for e-Science project[8].

10.1 Basis for the approach

The biomedical literature is vast and is augmented at an ever increasing rate. The number of articles published on HIV drug resistance alone makes it difficult for even the most dedicated researcher to keep pace with the literature. At time of writing, a search on PubMed[40] for articles during the past year on HIV drug resistance returns over 600 publications. Resources such as PubMed, which make abstracts of scientific articles available in an electronic form are invaluable sources of information but the sheer volume of that information makes it unmanageable without computational aid.

First generation rule systems for HIV drug resistance interpretation, such as Retrogram were compiled by experts gleaning knowledge from publications by hand. However state of the art techniques for information retrieval and data mining are available to assist in the task of identifying relevant publications and in formulating rules based on their content.

The biomedical literature is written with human readers in mind and, as such, is poorly structured to be understood by computer. Natural Language Processing is the field of study which concerns itself with the computational understanding of text written in natural or human languages. Applied to bioinformatics, the field continues to make progress and has its own journal (International Journal of Data Mining and Bioinformatics) but as yet has had successes only for well-defined problems in narrow areas.

We propose such a well-defined problem for which we will combine a number of tried and tested tools from the areas of Information Retrieval, Data Mining and Natural Language Processing.

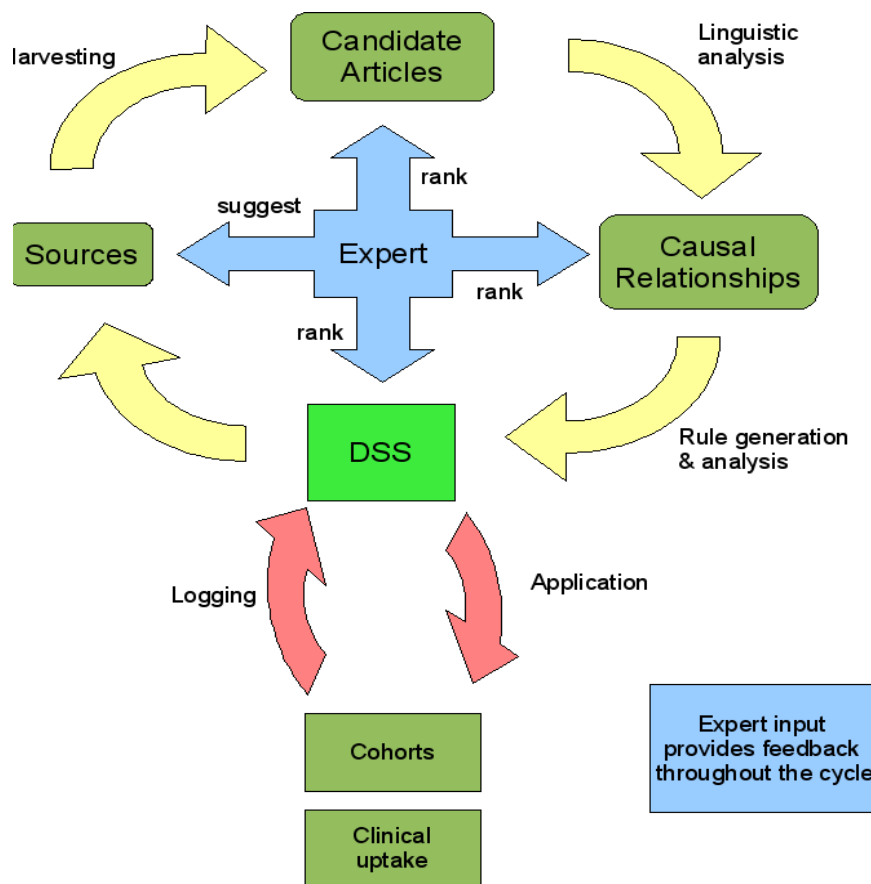


Figure 10.1: Semi-automated Literature Mining

10.2 Background & previous work

The focused task which we propose is the retrieval of causal relations between HIV sequence mutations and resistance to antiretroviral drugs by a surface linguistic analysis of the text of electronically published articles using PubMed. A similar task has been reported recently in extracting gene-disease relations from Medline[14].

10.3 Work & publications completed in the context of Virolab

This is a new research initiative in the Virolab project.

10.4 Plans for the remainder of the Virolab project

As training sets, we use the existing rule-based drug resistance interpretation systems, such as Retrogram, the Stanford HIVdb, Rega and ANRS. These rule sets provide a set of relations between HIV mutations and the resistance they impart on antiretroviral drugs. Several of these systems additionally provide links to the articles from the literature from which the rules in question have been derived. Using these articles, we will extract the phrases used to express the causal relations between mutations and drugs.

The first step in the process is the identification of salient entities and their synonyms. Examples of such entities are:

enzymes e.g. reverse transcriptase, protease.

drug classes e.g. nonnucleoside reverse transcriptase inhibitor (nNRTI).

drugs e.g. lamivudine, abacavir, amprenavir

mutations written with respect to a reference genome of HIV-1, commonly the NL4-3 wild type, and consisting of a position followed by an amino acid.

The identification of synonyms is not to be underestimated. Indeed quite a body of research has been dedicated to this task alone. Approaches such as ontologies, name dictionaries, contextual information and abbreviations have been tried. A valuable resource is MeSH, the U.S. National Library of Medicine's controlled vocabulary used for indexing articles for MEDLINE/PubMed.

A test for coverage will be performed on the set of articles obtained from the rule sets above.

A keyword search is performed on PubMed to harvest candidate articles. See the top left arrow in Figure 10.1. This amounts to constructing suitably formed boolean queries and sending them on PubMed.

This will provide us with a set of candidate abstracts which we can subject to a more structured linguistic analysis (top-right arrow of Figure 10.1). We identify a number of causal relations[24, 23]:

simple causatives synonyms for "cause" such as "give rise to", "effects".

resultative causatives linking verbs which refer to causal link plus resulting situation, such as "inhibit" or "suppress"

instrumental causatives express part of the causing event as well as the result, such as “activate” or “effectuate”

The text of the abstracts is parsed for the presence of such causal phrases linking mutations to drugs. This leads to a set of causal relationships.

10.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

The final stage is the generation of rules from these causal relationships (bottom-right arrow of Figure 10.1). These rules will be used to directly augment the Virolab HIV drug-susceptibility interpretation system.

The process is not entirely automated and each stage will be tuned by an expert who ranks sources, candidate articles, causal relationships and generated rules and provides feedback throughout the entire cycle. Additional feedback will be provided from provenance tracking.

11 Enhancements of rule-based interpretation.

This work is carried out by Breannán Ó Nualláin at the University of Amsterdam.

In Deliverable 4.1[38] we reported on the virtualisation of the Virolab HIV drug-susceptibility interpretation system and identified a number of shortcomings in both the existing Retrogram system and the state-of-the-art language (ASI) used to represent drug interpretation knowledge. Here we describe an improved language which is more expressive and enjoys a fully-specified, formal semantics, allowing for automated reasoning over rule sets.

11.1 Basis for the approach

Since the Virolab HIV drug-susceptibility interpretation system incorporates rule sets from many interpretation systems (Retrogram, Stanford HIVdb, Rega, ANRS)

we would like to make judgements of the following kinds about rule sets:

ambiguity Is the rule set internally ambiguous? Does it allow more than one interpretation?

completeness Does the rule set have complete coverage?

consistency Are there rules in the set which make contradictory predictions?

redundancy Do some rules of a rule set subsume others?

dissonance How do rule sets differ in their predictions?

predictive power Can one rule set make more specific predictions than another or can it make predictions in cases where the other is silent?

Further, recent findings have revealed the need to express multiplicative effects of certain mutations on drugs. ASI, in its present form, is limited to linear combinations of effects.

A formal language with a well-defined semantics will allow for making judgements of the above kinds using reasoning that is either completely automated or at least semi-automated. This potential to perform inference on data expressed in formally-specified languages has been recognised and is being developed by the Semantic Web project[5].

11.2 Background & previous work

The *de facto* state of the art for the expression of rules concerning mutation-influenced HIV drug resistance is the Algorithm Specification Interface[2] developed by the Stanford HIV Drug Resistance Database project[6].

ASI was an important step forward in the field since it provided a common language for the expression of such rules since it allowed for comparison of rule sets and rule systems and interoperability among them. ASI provides a formal language for describing the rules which are encoded in a drug interpretation system, independent of the language and computing environment used to program the system. Several institutes that develop such systems use ASI to publish their rule sets, among them Rega Institute[4], Agence Nationale de Recherche sur la SIDA[1] and Stanford HIVdb itself.

However, as was outlined in Deliverable 4.1[38], ASI in its current form has a number of shortcomings, both from a virological viewpoint as well as from an informatics viewpoint. For this reason we have defined a new formal language which subsumes ASI and has the additional advantage of having a fully-defined formal semantics. This allows inferences to be made about rule sets expressed in the language.

11.3 Work & publications completed in the context of Virolab

As already reported in Deliverable 4.1[38] we implemented the Virolab HIV drug-susceptibility interpretation system, in doing so designed a framework for representing several rule-based systems for HIV drug resistance interpretation, with staged representations of the knowledge contained in them. In this section we describe the formal syntax and semantics of the “core” representation language of the Virolab HIV drug-susceptibility interpretation system. This is the language needed to subsume the existing ASI-based systems but without any extensions needed to express, e.g. multiplicative virological effects.

Syntax of Virolab rule language

The syntax of the Virolab HIV drug-susceptibility interpretation system rule language is presented in Extended Backus-Naur Form (EBNF) notation. Literals are written in **bold face**. Nonterminals which are not further defined are written in *italic*. These are either well-defined mathematical notions e.g. $\langle nat \rangle$ is the representation of a natural number, or rule actions, $\langle action \rangle$, which are defined elsewhere. The Kleene star (*) and plus (+) indicate zero-or-more and one-or-more occurrences of the preceding, respectively.

$$\langle rule \rangle \longrightarrow \langle condition_rule \rangle \mid \langle interval_rule \rangle \mid \langle score_rule \rangle$$

$$\langle condition_rule \rangle \longrightarrow \langle mutation \rangle \langle action \rangle$$

$$\langle mutation \rangle \longrightarrow \langle atomic_mutation \rangle \mid \langle mutation \rangle \longrightarrow \langle indel \rangle$$

$$\langle mutation \rangle \longrightarrow \langle position \rangle \langle amino_acid_list \rangle$$

$$\langle mutation \rangle \longrightarrow \mathbf{not} \langle mutation \rangle$$

⟨mutation⟩ → ⟨mutation⟩ **and** ⟨mutation⟩
⟨mutation⟩ → ⟨mutation⟩ **or** ⟨mutation⟩

⟨indel⟩ → ⟨position⟩ (**I|D**)
⟨atomic_mutation⟩ → ⟨position⟩ ⟨amino_acid⟩
⟨position⟩ → ⟨*positive_nat*⟩
⟨amino_acid⟩ → **A|C|D|E|F|G|H|I|K|L|M|N|P|Q|R|S|T|V|W|Y**
⟨amino_acid_list⟩ → [**not**] ⟨amino_acid⟩+

⟨interval_rule⟩ → **range** [⟨*nat*⟩ , ⟨*nat*⟩] ⟨mutation⟩ ⟨*action*⟩
⟨score_rule⟩ → ⟨score_clause⟩ +
⟨score_clause⟩ → ⟨mutation⟩ ⇒ ⟨*float*⟩

Semantics of Virolab rule language

The semantics of expressions in the Virolab rule language are given relative to a wild type, namely the HIV reference genome NL4-3. Its protease segment is the 99 amino-acid string:

```
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGI  
GGFIKVGQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
```

and its reverse transcriptase segment is the 556 amino-acid string:

```
PISPIETVPVKLKP GMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKI  
GPENPYNTPVFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGL  
KQKKSVTVLDVGDAYFSVPLDKDFRKYTAFTIPSINNETPGIRYQYNVLP  
QGWKGSPAIFQCSMTKILEPFRKQNPDVVIYQYMDDL YVGS DLEIGQHRT  
KIEELRQHLLRWGF TTPDKKHQKEPPFLWMGYELHPDKWTVQP IVLPEKD  
SWTVNDIQKLVGKLNWASQIYAGIKVRQLCKLLRGTKALTEVVPLTEEEAE  
LELAENREILKEPVHGVYYDPSKDLIAEIQKQGQGWTYQIYQEPFRNLK  
TGKYARMKGAHTNDVKQLTEAVQKIATESIVIWGKTPKF KLP IQKETWEA  
WWTEYWQATWIPEWEFVNT PPLV KLWYQLEKEP IIGAETFYVDGAANRET  
KLGKAGYVTRGRQKVVPLTDTTN QKTELQAIHLALQDSGLEVNIVTDSQ  
YALGIIQAQPKSESELVSQIIEQLIKKEKVYLAWVPAHKGIGGNEQVDK  
LVSAGI
```

The semantics of each rule is given as a function of an input sequence, s , given as a set of point mutations on amino acids, for example 81F, 122K.

The semantics of a rule is an action, which will depend on the rule set in question. For example, in the case of Rega or Retrogram, it is a level of susceptibility. In recent versions of Stanford HIVdb, it is a score which is mapped to a set of susceptibility levels. Further details are omitted here.

The semantics of a condition is a predicate (i.e. a Boolean-valued function) on sequences.

$$\begin{aligned}
 \llbracket \langle \text{condition_rule} \rangle \rrbracket (s) &= \llbracket \langle \text{mutation} \rangle \langle \text{action} \rangle \rrbracket (s) \\
 &= \text{if } \llbracket \langle \text{mutation} \rangle \rrbracket (s) \text{ then } \llbracket \langle \text{action} \rangle \rrbracket \\
 \llbracket \langle \text{interval_rule} \rangle \rrbracket (s) &= \llbracket \text{range } [m, n] \langle \text{mutation} \rangle \langle \text{action} \rangle \rrbracket (s) \\
 &= \text{if } m \leq \#\{p \in s : \llbracket \langle \text{mutation} \rangle \rrbracket (p)\} \leq n \text{ then } \llbracket \langle \text{action} \rangle \rrbracket \\
 \llbracket \langle \text{score_rule} \rangle \rrbracket (s) &= \llbracket \langle \text{score_clause} \rangle + \rrbracket (s) \\
 &= \llbracket (\langle \text{mutation} \rangle \Rightarrow \langle \text{float} \rangle) + \rrbracket (s). \\
 &= \sum \{ \langle \text{float} \rangle : \llbracket \langle \text{mutation} \rangle \rrbracket (s) \}
 \end{aligned}$$

There are three kinds of rule: a direct condition rule; an interval rule in which the number of occurrences of the given mutation in the sequence is bounded above or below or both; and a score rule which assigns a score to the given sequence depending on which of its component mutations occur. Note that the score rule is limited to an additive effect.

The semantics of mutations are written in terms of positions, indicated by p and amino acids, indicated by a .

$$\begin{aligned} \llbracket \langle \text{atomic_mutation} \rangle \rrbracket (s) &= \llbracket \langle \text{position} \rangle \langle \text{amino_acid} \rangle \rrbracket (s) \\ &= \langle \langle \text{position} \rangle, \langle \text{amino_acid} \rangle \rangle \in s \\ \llbracket \langle \text{indel} \rangle \rrbracket (\langle p, a \rangle) &= \llbracket \langle \text{position} \rangle \mathbf{I} \rrbracket (\langle p, a \rangle) \\ &= p = \langle \text{position} \rangle \wedge a = \mathbf{I} \\ \llbracket \langle \text{indel} \rangle \rrbracket (\langle p, a \rangle) &= \llbracket \langle \text{position} \rangle \mathbf{D} \rrbracket (\langle p, a \rangle) \\ &= p = \langle \text{position} \rangle \wedge a = \mathbf{D} \\ \llbracket \langle \text{position} \rangle \langle \text{amino_acid_list} \rangle \rrbracket (s) &= \exists \langle p, a \rangle \in s \\ &\quad \text{such that } p = \langle \text{position} \rangle \wedge a \in \langle \text{amino_acid_list} \rangle \\ \llbracket \langle \text{mutation} \rangle_1 \mathbf{and} \langle \text{mutation} \rangle_2 \rrbracket (s) &= \llbracket \langle \text{mutation} \rangle_1 \rrbracket (s) \wedge \llbracket \langle \text{mutation} \rangle_2 \rrbracket (s) \\ \llbracket \langle \text{mutation} \rangle_1 \mathbf{or} \langle \text{mutation} \rangle_2 \rrbracket (s) &= \llbracket \langle \text{mutation} \rangle_1 \rrbracket (s) \vee \llbracket \langle \text{mutation} \rangle_2 \rrbracket (s) \\ \llbracket \mathbf{not} \langle \text{mutation} \rangle \rrbracket (s) &= \neg \llbracket \langle \text{mutation} \rangle \rrbracket (s) \end{aligned}$$

The notation for mutations is necessarily quite complex and is the source of some confusion in ASI. For example what is the intended semantics of the following?

$\text{not}(184, \text{not}(F, G))$

Does this mean that there is not a mutation at position 184 which is not either F or G?

11.4 Plans for the remainder of the Virolab project

- Validation of the Virolab HIV drug-susceptibility interpretation system encodings of the ASI-based systems against the existing systems. This will require making explicit the assumptions made for overcoming ambiguities in rule sets defined in ASI, consultation with their authors and testing of the resulting rule sets.
- Definition of an XML syntax for the Virolab HIV drug-susceptibility interpretation system rule language. ASI is expressed as a Document Type Definition (DTD). This has become a dated technology. A schema language such as Relax NG is a more suitable candidate.

- Implementation of software to perform the automated reasoning tasks specified in Section 11.1 and the accompanying comparison and analysis of the rule sets.
- Study of dissonance among rule sets based on patient data.

11.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

The work described in this chapter is at the very core of the Virolab HIV drug-susceptibility interpretation system, providing a flexible and adaptable system to which all of the other enhancements described in this deliverable can be added.

12 Bayesian combination of evidence.

This work is carried out by Breannán Ó Nualláin, Kaustubh Patil and Peter M. A. Sloot at the University of Amsterdam.

Information, data and evidence from many of the sources reported in this document as well as data from the provenance component of the virtual laboratory need to be combined within the Virolab HIV drug-susceptibility interpretation system to provide coherent judgements on drug-susceptibility. We apply Bayesian statistical data analysis to support this decision making.

12.1 Basis for the approach

Decision making can be supported by statistical data analysis. How can statistical inferencing serve to support the decision making process? The available evidence and data are, by their nature, incomplete so methods are required which can deal with missing data.

Classical regression has some well-known limitations for decision analysis. Accounting for widely ranging sample sizes is not easy and it cannot reflect hierarchical structure in a model. In addition sparse data make fitting such a model

relating response to covariates difficult to estimate.

As Goodman[25] points out, biological understanding and previous research play little formal role in the interpretation of quantitative results when classical, frequentist statistical methods are applied. P values and hypothesis tests have limited utility in this context.

Our approach will be to use Bayesian hierarchical modelling to make predictive distributions in the presence of uncertainty. The full chain of analysis will combine Bayesian hierarchical modelling with probabilistic decision analysis based on utility attribution and/or multi-objective optimisation of such quantities as cost, chance and duration of survival or quality-adjusted life years.

Provenance data indicating confidence of medical clinicians in rule sets and various sources of experimental data will also play a role.

Bayesian analysis will be used in two ways: as posterior inferences summarising uncertainties about predictive quantities, as well as within the decision analysis in multistage decision trees.

Attention will be given to the value of information; Balancing risks, delays and confidence in obtaining additional data needed to bolster inferences.

12.2 Background & previous work

As Brophy and Joseph [13] point out, decision making with incomplete evidence is a difficult but frequently occurring medical dilemma. A number of methods are available which when judiciously applied lead to a reasonable structure for decision making in clinical situations with incomplete evidence and even with sources evidence which are inconsistent[10]

The prevailing paradigm for clinical decision making revolves around evidence-based medicine, with randomised clinical trials representing the zenith of experimental comparative design. However in the absence of such idealised evidence, decisions must still be made. They examine three methods:

1. Objective Bayesian analysis
2. Bayesian analysis incorporating prior information
3. Indirect comparisons involving hierarchical Bayesian meta-analysis

They point out that “innovative means of using other data to complement and enhance the results of randomized trials are being increasingly examined” and further that “contemporary statistical theory suggests more informed decisions might be reached by attempting to incorporate prior information with the results of [. . .] trials.”

In order to provide coherent judgements taking into account evidence from literature, rule sets, modelling and simulation we will need to combine a number of kinds of knowledge, categorised by Shortliffe et al.[47] as:

- knowledge derived from data analysis (numerical or statistical)
- judgemental or empirical subjective knowledge
- common sense or scientific knowledge
- strategic knowledge or procedural knowledge

Potentially suitable is Dempster and Shafer’s framework for modelling the combination of evidence, which been extended to incorporate the uncertain nature of information retrieval and relevance feedback. [46]

Rocchio [44] identified the notion of “relevance feedback”: users being unable to state what information they need, but being able to mark information objects that are relevant to their needs. This will be of particular interest to us in developing a system for clinicians who may be neither knowledgeable nor indeed interested in the technical details of how a decision has been reached. This and many other issues are germane to making a user-friendly drug-resistance interpretation system which will gain broad acceptance among clinicians. Indeed in early research Shortliffe & Pagan[47] identify key questions asked by physicians about medical decision support systems which were crucial to the design of their MYCIN system:

1. Do I need this system?
2. Will it help without being dogmatic?
3. Does it justify its recommendations so that I can judge them for myself?
4. Is it fast and easy to use?
5. Is it designed to make me feel comfortable when I use it?

12.3 Work & publications completed in the context of Virolab

This is a new research initiative in the Virolab project.

12.4 Plans for the remainder of the Virolab project

We follow the methodology prescribed by Gelman et al. [22]:

1. Enumerate the space of possible decisions and outcomes, where outcomes include observables and parameters.
2. Determine conditional posterior probability distributions for outcomes over all decisions.
3. Define a utility function over outcomes *or* define multiple objectives to be Pareto optimised.
4. Compute the expected utility (or utilities) as a function of the decisions. Where a decision-making sequence is represented by a decision tree, the expected utility must be calculated at each node, conditional on all information available up to that point.

We favour this Bayesian approach above that of classical statistical decision theory, which relies on optimal point estimates of utility to the neglect of complete posterior distributions.

12.5 Nature of results and their possible use for enhancing the Virolab HIV drug-susceptibility interpretation system.

This research initiative is concerned with the very combination of results from the other research initiatives and their synthesis into a unified prediction of drug resistance.

13 Abbreviations

Abbreviation/Term	Explanation
AHE	Application Hosting Environment
AIDS	Acquired Immune Deficiency Syndrome
ANRS	Agence Nationale de Recherche sur la SIDA
ARV	Antiretroviral
ASI	Algorithm Specification Interface
BAC	Binding Affinity Calculator
CN	Complex Network
DNA	Deoxyribonucleic Acid
DTD	Document Type Definition
EBNF	Extended Backus-Naur Form
HAART	Highly Active Anti-Retroviral Therapy
HIV	Human Immunodeficiency Virus
MD	Molecular Dynamics
MM/PBSA	Molecular Mechanics/Poisson-Boltzmann Surface Area
RNA	Ribonucleic Acid
SBML	Systems Biology Mark-up Language
USM	Universal Similarity Measure
WT	Wild Type

Bibliography

- [1] Agence Nationale de Recherche sur la SIDA. <http://www.anrs.fr>.
- [2] Algorithm Specification Interface. <http://hivdb.stanford.edu/pages/asi/>.
- [3] Phylogeny programs. <http://evolution.genetics.washington.edu/phylip/software.html>.
- [4] Rega Institute for Medical Research. <http://www.kuleuven.ac.be/reg/>.
- [5] Semantic web. <http://www.w3.org/2001/sw/>.

- [6] Stanford HIV Drug Resistance Database project. <http://hivdb.stanford.edu/>.
- [7] Virolab: A decision support system for hiv drug ranking. <http://www.virolab.org/>.
- [8] The Virtual Laboratory for e-Science. www.vl-e.nl.
- [9] O. O. Aalen, V. T. Farewell, D. De Angelis, N. E. Day, and O. N. Gill. New therapy explains the fall in AIDS incidence with a substantial rise in number of persons on treatment expected. *AIDS*, 13(1):103–108, 1999.
- [10] A. E. Ades and A. J. Sutton. Multiparameter evidence synthesis in epidemiology and medical decision making: current approaches. *Journal of the Royal Statistical Society A*, 169(1):5–35, 2006.
- [11] M. Alfano and G. Poli. *Drug Design Rev.*, 2004.
- [12] F. Baryarama, J. Y. T. Mugisha, and L. S. Luboobi. Mathematical model for HIV/AIDS with complacency in a population with declining prevalence. *Computational and Mathematical Methods in Medicine*, 77(1):27–35, 2006. Available as <http://www.informaworld.com/smpp/title~content=t713653639~db=all~tab=issueslist~branches=7>.
- [13] James M. Brophy and Lawrence Joseph. Medical decision making with incomplete evidence. *Medical Decision Making*, 25(2):222–228, 2005.
- [14] Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun ichi Tsujii. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. In *Pacific Symposium on Biocomputing*, pages 4–15, 2006.
- [15] P.V. Coveney, R.S. Saksena, S.J. Zasada, M. McKeown, and S. Pickles. *Comp. Phys. Commun.*, 2007.
- [16] Rob de Boer. CD8+ T-cell response latency in human immunodeficiency virus infection. Technical report, University of Utrecht, May 2005. Excellent Tracé second phase.

- [17] T. de Oliveira, K. Deforche, S. Cassol, M. Salminen, D. Paraskevis, C. Seebregts, J. Snoeck, E. J. van Rensburg, A. M. J. Wensing, D.A. van de Vijver, C. A. Boucher, R. Camacho, and A-M. Vandamme. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, 21(19):3797–3800, 2005. <http://www.bioafrica.net/virus-genotype/html/subtypinghiv.html>.
- [18] Richard C. Deonier, Simon Tavaré, and Michael S. Waterman. *Computational Genome Analysis: An Introduction*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.
- [19] Gökhan Ertalyan. DNA vaccination for viral diseases: Towards HIV and hepatitis vaccines. Master's thesis, University of Utrecht, September 2005.
- [20] G. Fatkenheuer, A.L. Pozniak, M.A. Johnson, A. Plettenberg, S. Staszewski, A.L. Hoepelman, M.S. Saag, F.D. Goebel, J.K. Rockstroh, and B.J. Dezube. Efficacy of short-term monotherapy with maraviroc, a new CCR5 antagonist, in patients infected with HIV-1. *Nat. Med.*, 11:1170–72, 2005.
- [21] Center for Disease Control. <http://www.cdc.gov/hiv/>.
- [22] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edition, 2004.
- [23] R. Girju. Automatic detection of causal relations for question answering. In *Proc. of the 41st ACL, Workshop on Multilingual Summarization and Question Answering.*, 2003.
- [24] R. Girju and D. Moldovan. Text mining for causal relations. In *The 15th international FLAIRS conference in cooperation with AAAI*, 2002.
- [25] Stephen N. Goodman. Towards evidence-based medical statistics. 1. the P-value fallacy. *Ann. Intern. Med.*, 130:995–1004, 1999.
- [26] Z. Guo, H. K. Han, and J. C. Tay. Sufficiency verification of HIV-1 pathogenesis based on multi-agent simulation. In *Proceedings of the ACM Genetic and Evolutionary Conference 2005 (GECCO 2005)*, volume I, pages 305–312, 2005. Best Paper Nominee.

- [27] Z. Guo and J. C. Tay. A comparative study of modeling strategies of immune system dynamics under HIV-1 infection. In *In Proceedings of the International Conference for Artificial Immune Systems (ICARIS 2005)*, volume 3627, pages 220–233, Banff, Alberta, Canada, August 2005. Springer LNCS.
- [28] Zaiyi Guo and Joc Cing Tay. Assessing causal intuitions in agent-based microbiological models. In *Proceedings of the European Conference on Complex Systems (ECCS 2007)*, October 2007. abstract.
- [29] Zaiyi Guo and Joc Cing Tay. A hybrid agent-based model of chemotaxis. In *Proceedings of the 7th International Conference on Computational Science (ICCS 2007)*, volume 4487, pages 119—127. Springer LNCS, 2007.
- [30] Zaiyi Guo and Joc Cing Tay. Multi-timescale event scheduling in multi-agent immune simulation models. *Journal of Biosystems*, 2007. to appear.
- [31] P.J. Klasse, R.J. Shattock, and J.P. Moore. Which topical microbicides for blocking HIV-1 transmission will work in the real world? *PloS Med.*, 3(9):1501–07, 2006.
- [32] N. Krasnogor and D. A. Pelta. Measuring the similarity of protein structures by means of the universal similarity metric. *Bioinformatics*, 20(7):1015–1021, 2004.
- [33] J. Lama and V. Planelles. Host factors influencing susceptibility to hiv infection and aids progression. *retrovirology*. 2007, 4:52, in press.
- [34] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul Vitányi. The similarity metric. In *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 863–872, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.
- [35] F. Liljeros, C. R. Edling, L. A. Amaral, H. E. Stanley, and Y. Aberg. The web of human sexual contacts. *Nature*, 411(6840):907–908, 2001.
- [36] F. Miedema. AIDS pathogenesis: a dynamic interaction between HIV and the immune system. *Immunol. Today*, 11:293–97, 1990.

- [37] Amri Muchammad. Designing a model of HIV cell entry using UML. Final year report, EvoCom, 2007.
- [38] Breannán Ó Nualláin. Deliverable 4.1: Virtualisation and automation. Technical report, The Virolab Project, June 2007.
- [39] The Virolab Project. Description of work. Technical Report Contract nr 027446, Sixth Framework Programme Priority 2, Information Society Technologies, October 2005.
- [40] PubMed. www.pubmed.gov.
- [41] Karthik Raveendran. A distributed event-driven simulation of chemotaxis. Final year report, EvoCom, 2007.
- [42] R. Regoes and S. Bonhoeffer. The HIV coreceptor switch: a population dynamical perspective. *TRENDS in Microbiology*, 2006.
- [43] R.M. Ribeiro, M.D. Hazenberg, A.S. Perelson, and M.P. Davenport. Naïve and memory cell turnover as drivers of CCR5-to-CXCR4 tropism switch in human immunodeficiency virus type 1: implications for therapy. *Journal of Virology*, 80(2):802–9, 2006.
- [44] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART retrieval system: experiments in automatic document processing*, chapter 14, pages 313–323. Prentice-Hall, 1971.
- [45] Jairo Rocha, Francesc Rosselló, and Joan Segura. Compression ratios based on the universal similarity metric still yield protein distances far from CATH distances, 2006. <http://arxiv.org/abs/q-bio/0603007v2>.
- [46] I. Ruthven and M. Lalmas. Using Dempster-Shafer’s theory of evidence to combine aspects of information use. *Journal of Intelligent Information Systems*, 19:267–302, 2002.
- [47] Edward H. Shortliffe and Lawrence M. Pagan. Expert systems research: modeling the medical decision making process. Technical report, Stanford University, 1982.
- [48] P. M. A. Sloot, A. V. Boukhanovsky, W. Keulen, A. Tirado-Ramos, and C. A. Boucher. *J. Clinic. Monit. Comput.*, 2005.

- [49] P. M. A. Sloom, F. Chen, and C. A. Boucher. Cellular automata model of drug therapy for HIV infection. In S. Bandini, B. Chopard, and M. Tomassini, editors, *5th International Conference on Cellular Automata for Research and Industry (ACRI)*, volume 2493 of *Lecture Notes in Computer Science*, pages 282–293, Geneva, Switzerland, 2002.
- [50] P. M. A. Sloom, S. V. Ivanov, A. V. Boukhanovsky, D. van de Vijver, and C. Boucher. "stochastic simulation of HIV population dynamics through complex network modeling". *International Journal of Computer Mathematics*, 2007. Accepted for publication.
- [51] P. M. A. Sloom, A. Tirado Ramos, I. Altintas, M. T. Bubak, and C. A. Boucher. From molecule to man: Decision support in individualized e-health. *IEEE Computer*, 39(11):40–46, 2006. (cover feature).
- [52] M.J. Sweeting, D. De Angelis, and O. O. Aalen. Bayesian back-calculation using a multi-state model with application to HIV. *Statist. Med.*, 24:3991–4007, 2005.
- [53] Joc Cing Tay and Philip Tan. Evolving boolean networks to find intervention points in dengue pathogenesis. In *Proceedings of the IEEE Engineering in Medicine and Biology Conference (EMBC 2006)*, pages 5315–5321, 2006.
- [54] M. Tersmette and F. Miedema. Interaction between HIV and the host immune system in the pathogenesis of AIDS. *AIDS*, 4:57–66, 1990.
- [55] Yit Hong Toh and Joc Cing Tay. An overview of HAART. Technical report, EvoCom, 2005.
- [56] S. Wan, P. V. Coveney, and D. R. Flower. *Phil. Trans. Royal Soc. A.*, 2005.
- [57] W. Wang and P. A. Kollman. *Proc. Natl. Acad. Sci.*, 2001.
- [58] A. Yates, J. Stark, N. Klein, R. Antia, and R. Callard. Understanding the slow depletion of memory CD4+ T-cells in HIV infection. *PloS Med.*, 4(5):e177, 2007.