A multi-scale approach for deciphering HIV infection
Ertaylan, G.

Citation for published version (APA):
INTRODUCTION

“The capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problem whose solution is required for objectively rational behavior in the real world or even for a reasonable approximation to such objective rationality.”

Herbert Simon, 1957

Today, there are 6.9 billion people living on earth [1], interacting continuously with one another forming an immense network of interactions. A person waking up with an infectious disease in Amsterdam is able to infect someone in New York by the end of the day. In an environment like this dealing with infectious diseases poses a great problem, which needs to be addressed thoroughly. The good news is that we are also equipped with tools, which allow us to understand the problem in considerable detail.

One particular example is the Human Immunodeficiency Virus (HIV). Since it was recognized in 1981, acquired immunodeficiency syndrome (AIDS) has claimed more than 25 million lives.

Genetic analyses of archived blood samples from early AIDS patients revealed that the epidemic started in Africa but arrived in the U.S. in 1969 from a single common ancestor from Haiti [2]. Once the virus got to the U.S., then it spread explosively around the world. Since then the HIV epidemic has become a pandemic and today at least 33 million people are living with HIV [20], among which substantial proportion remains unaware of their serostatus [21]. In a report by Mathers and Loncar from the World Health Organization (WHO), projections of global mortality up to 2030 identified HIV/AIDS as one of the three leading causes of death [22]. Although almost every step of the virus life cycle has been unraveled in the past 25 years, and more than twenty distinct drugs targeted against HIV have been approved, all efforts to achieve an overall eradication of the virus have turned out to be ineffective up to now.

There are several reasons why treating HIV infection or developing a preventive vaccine is exceptionally difficult. On the cellular level, the high virus production rate of $10^3 - 10^4$ viral particles per infected cell [4] forces the system wide activation of the immune system causing more CD4+ T cells to be actively infected. This in turn leads to a high turnover rate for the virus where both viral production and the clearance rate increase. The continuous system wide
Introduction

activation of the immune system causes the structure of the lymph nodes to be compromised [4] leading to somewhat diminished functioning of the innate immune system. In time, the resistance to the infection weakens if the patient is not using any anti-retrovirals and the patient develops AIDS.

For patients who are continuously under anti retroviral therapy, HIV’s ability to rapid reproduction is greatly suppressed. However, the virus has the ability to remain hidden within the host cells genome where it can stay latent for several weeks to months. Latently infected cells can start producing virus particles after activation. Hence the infection is re-ignited by those latently infected cells if the medication is interrupted.

On the genotypic level there is another major problem; the extra ordinarily high nucleotide substitution rate of HIV due to the lack of proof reading mechanism of the viral enzyme Reverse Transcriptase (RT). It is estimated to be several million times faster than an average eukaryotic genome. Such a high substitution rate enables the presence of slightly different versions of the viral proteins in different viruses simultaneously. This coupled with the current estimate of total HIV-1 particles (5.10^{10} units) and the total number of productively infected CD4+ T cells (10^8 cells), translates to a virus population that exists in a cloud of genotypes called *quasispecies* [5], [6]. This rapid turnover rate and vast genotypic diversity in the quasispecies allows adaptation to environmental stresses (such as developing drug resistance to anti-retroviral medication) and changing immune response over time.

Consequently, the infection with HIV has inherent complexity emerging from the underlying processes in all different scales. We know many aspects of the disease, we have dedicated large databases for storing those information, and we have developed over twenty drugs targeting various proteins of the virus, yet we have still many unanswered questions.

On the other hand we have an idea where to go from here. Today we have the ability to measure clinical markers of infection (such as viral count and CD4+ T cell count) even in resource poor settings. We sequence viral RNA from chronically infected patients as daily practice in our medical centers. We have databases filled with protein interaction information as well as HIV drug resistance and still growing. Today not only that we have unprecedented processing power on the tip of our finger and are able store large quantities of data in our repositories, but also we continue to generate new data with a pace that has not even imagined before. We conduct on average seventy-five medical randomized control trials vs eleven systematic reviews a day, we generate data much faster than we are analyzing [9]. We are at a crucial point in the history of biology that, for the first time, the improvements in our techniques exceeded our ability to grasp the complexity lying underneath our findings.

In this manuscript we will argue that we can and should approach the problem of HIV infection more systematically and from the computational science point of view. If we want to understand the disease we *have to register, quantify and analyze how the virus percolates through its life cycle, infection after infection and how do the small changes in this process translates back to the clinical observations*.

This goal can be achieved by developing models that allow us to bridge the gap between different spatio-temporal scales while applying novel numerical algorithms that will make such multiscale information processing computationally tractable and reproducible. Furthermore, the feedback from these modeling efforts will be used to direct the experiments in the future.

The development of the computational models brings inevitable questions; what is the level of abstraction one can choose while keeping the functioning of the system intact; what is the fundamental information that needs to be transferred from one model or scale to another; what physical principles must be satisfied during the transfer of information?

*This work primarily focuses on unraveling the interplay between HIV and different cell...*
types of the human immune system. Since there is ample data on molecular scale up to population level, the real challenge is identifying a strategy for systematically tackling this data that will eventually help to answer some of the fundamental questions we are looking for: How does HIV hijacks the cellular machinery; are there any crucial pathways or proteins for the virus; can we predict the type of cells HIV is interacting by analyzing reported HIV-human protein interactions; what can we learn about the evolutionary dynamics of HIV by modeling the interactions between HIV and types of different immune system cells and finally, do those interactions of HIV with various cell types have an effect on developing AIDS? Throughout this thesis each and every one of these questions is answered. We explain how and which pathways and processes primarily accessed by HIV proteins using protein interaction networks. We infer types of cellular proteins that are potential targets for interacting with HIV and report cell types expressing those proteins. We also explain the effect of different cell types on HIV infection by modeling the co-receptor tropism switch of HIV.

The rest of this chapter will briefly cover necessary background. How the genetic information is stored, transcribed and translated to produce proteins. How these proteins function together to form a cell, the building block of multicellular organisms. Later, human immune system is introduced. Finally there is a short introduction about HIV and its life cycle in human host.

Background

DNA
All living cells, without any exception, store their hereditary information in the form of double-helix of stranded molecules of unbranched paired polymer chains, formed always of the same types of monomers- adenine (A), thymine (T), cytosine (C), guanine (G). These monomers—that is, each nucleotide- are strung together in a long linear sequence that encodes the genetic information, just as the sequence of 1s and 0s encodes the information in a computer file [23].

RNA
To carry out its information-storage function, DNA must do more than copy itself before cell division. It must also express this information, putting it to use so as to guide the synthesis of other molecules in the cell. This process begins with a templated polymerization called transcription, in which segments of the DNA sequence are used as templates to guide the synthesis of shorter molecules of the closely related polymer ribonucleic acid, or RNA. RNA is composed of the same nucleotides other than T becoming urasil (U). Later in the more complex process of translation, many of these RNA molecules serve to direct the synthesis of polymers of a radically different chemical class-the proteins [23].

PROTEIN
Protein molecules, like DNA and RNA, are long unbranched polymer chains, formed by the stringing together of monomeric building blocks drawn from a standard repertoire that is the same for all living cells. Like DNA and RNA, they carry information in the form of a linear sequence of symbols, in the same way as a human message written in an alphabetic script. There are many different protein molecules in each cell, and – apart from water- they form most of the cell’s mass [23].
Figure 1.1: From DNA to Protein: Genetic information is read out and put to use through a two-step process. First, in transcription, segments of the DNA sequence are used to guide the synthesis of molecules of RNA. Then, in translation, the RNA molecules are used to guide the synthesis of molecules of protein [23].

CELL
It is estimated that there are more than 10 million living species on Earth today. Each species is different, and reproduces itself faithfully, yielding a progeny that belong to the same species. Most of those organisms are single cells; others such as ourselves, are vast multicellular cities in which groups of cells perform specialized functions and are linked by intricate systems of communication. But in all cases the whole organism has been generated by cell divisions from a single cell. The single cell, therefore, is the vehicle for the hereditary information that defines the species. Specified by this information, the cell includes the machinery to gather raw materials from the environment, and to construct out of them a new cell in its own image, complete with a new copy of the hereditary information [23].

IMMUNE SYSTEM
Like all other multicellular organisms, we have evolved several mechanisms to resist infection by pathogens. To combat especially powerful pathogens that breach these barricades, vertebrates use two types of immune defense, which are carried out by specialized proteins and cells: innate immune responses spring into action immediately after an infection begins and do not depend on the host’s prior exposure to the pathogen, while more powerful adaptive immune responses operate later in an infection and are highly specific for the pathogen that induced them [23].

Human immune system is a system of cells, chemicals and processes that is responsible for protecting the organism against diseases by identifying and killing pathogens and tumor cells. Detection is complicated as pathogens can evolve rapidly, producing adaptations that avoid the immune system and allow the pathogens to successfully infect their hosts.
The immune system protects organisms from infection with layered defenses with increasing specificity. In simple terms, physical barriers prevent pathogens such as bacteria and viruses from entering the organism. If a pathogen breaches the physical barriers of the organism, the innate immune system provides an immediate, but non-specific response. If the innate response is successfully evaded, vertebrates possess a third layer of protection, the adaptive immune system, which is activated by the innate response. Here, the immune system adapts its response during an infection to improve its recognition of the pathogen. This improved response is then retained after the pathogen has been eliminated, in the form of immunological memory, and allows the adaptive immune system to mount faster and stronger attacks each time the same (or similar) pathogen is encountered.

The members of the adaptive immune system are special types of cells, called lymphocytes. The major types of lymphocytes are B cells and T cells. They are derived from hematopoietic stem cells in the bone marrow. B cells are involved in the humoral immune response, whereas T cells are involved in cell-mediated immune response.

Helper T cells (Th) are type of lymphocytes that are important in the infection with HIV. They regulate both the innate and adaptive immune responses and help determine which types of immune responses the body will produce against a particular intruder [7][8]. These cells have no cytotoxic activity and do not kill infected cells or clear pathogens directly. They instead control the immune response by directing other cells to perform these tasks. They are also the types of immune system cells primarily infected with HIV.

**HUMAN IMMUNODEFICIENCY VIRUS (HIV)**

Human immunodeficiency virus (HIV) is a lentivirus (a member of the retrovirus family) that causes acquired immunodeficiency syndrome (AIDS), a condition in humans in which the immune system begins to fail, leading to life-threatening opportunistic infections.

HIV-1 initially infects T cells and macrophages directly or is carried by dendritic cells to the lymph nodes where T cells and macrophages reside. Viral replication in the regional lymph nodes leads to viremia and the spreading of the infection in the lymphoid tissue. If the viremia is controlled by the host’s immune response, then the patient enters a phase of clinical latency. During this phase, viral replication in T cells, macrophages and dendritic cells continues. There continues a gradual erosion of CD4+ cells by productive infection. When the CD4+ cells that are destroyed cannot be replenished, CD4+ cell numbers decline and the patient develops clinical symptoms of full-blown AIDS. Macrophages, which are infected but not lysed by the immune system, can transport the virus to various tissues, particularly the brain (See Figure-1.2).

If untreated most people infected with HIV-1 eventually develop AIDS [10]. These individuals mostly suffer from opportunistic infections associated with the progressive failure of the immune system [11].
Individuals progress to AIDS at a variable rate affected by viral, host, and environmental factors; most will progress to AIDS within 10 years of HIV infection: some will progress much sooner, and some will take much longer [12][13]. Treatment with anti-retroviral drugs significantly increases the life expectancy of people infected with HIV.
HIV LIFE CYCLE
HIV infects primarily vital cells in the human immune system such as helper T cells (CD4+ T cells), macrophages, and dendritic cells [14]. HIV infection leads to low levels of CD4+ T cells through three main mechanisms: First, direct viral killing of infected cells; second, increased rates of apoptosis in infected cells; and third, killing of infected CD4+ T cells by CD8 cytotoxic lymphocytes that recognize infected cells. When CD4+ T cell numbers decline below a critical level, cell-mediated immunity is lost, and the body becomes progressively more susceptible to opportunistic infections.

HIV enters macrophages and CD4+ T cells by fusion of the viral envelope with the cell membrane and the release of the viral capsid enzyme into the cell [15][16].

Entry to the cell begins through interaction of the envelope complex (gp160 spike) and both CD4 and a chemokine receptor (generally either CCR5 or CXCR4) on the cell surface.
The first step in fusion involves the high-affinity attachment of the CD4 binding domains of gp120 to CD4. Once gp120 is bound with the CD4 protein, the envelope complex undergoes a structural change, exposing the chemokine binding domains of gp120 and allowing them to interact with the target chemokine receptor \[15\][16]. This allows for a more stable two-pronged attachment, which allows the N-terminal fusion peptide gp41 to penetrate the cell membrane \[15\][16]. Repeat sequences in gp41, HR1 and HR2 then interact, causing the collapse of the extracellular portion of gp41 into a hairpin. This loop structure brings the virus and cell membranes close together, allowing fusion of the membranes and subsequent entry of the viral capsid \[15\][16].

After HIV has bound to the target cell, the HIV RNA and various enzymes, including reverse transcriptase, integrase, ribonuclease, and protease, are injected into the cell \[15\] (See Figure 1.3).

Shortly after the viral capsid enters the cell, an enzyme called reverse transcriptase liberates the single-stranded (+) RNA genome from the attached viral proteins and copies it into a complementary DNA (cDNA) molecule \[17\]. The process of reverse transcription is extremely error-prone, and the resulting mutations cause drug resistance or allow the virus to evade the body's immune system. The reverse transcriptase also has ribonuclease activity that degrades the viral RNA during the synthesis of cDNA. Together, the cDNA and its complement form a double-stranded viral DNA and is then transported into the cell nucleus. The integration of the viral DNA into the host cell's genome is carried out by another viral enzyme called integrase \[17\].

This integrated viral DNA then may lie dormant, in the latent stage of HIV infection (latency) \[17\]. To actively produce the virus, certain cellular transcription factors need to be present. This means that those cells most likely to be killed by HIV are those currently fighting infection.

During viral replication, the integrated DNA provirus is transcribed into mRNA, which is then spliced into smaller pieces. These small pieces are exported from the nucleus into the cytoplasm, where they are translated into the regulatory proteins Tat (which encourages new virus production) and Rev. As the newly produced Rev protein accumulates in the nucleus, it binds to viral mRNAs and allows unspliced RNAs to leave the nucleus, where they are otherwise retained until spliced \[18\]. At this stage, the structural proteins Gag and Env are produced from the full-length mRNA. The full-length RNA is actually the virus genome; it binds to the Gag protein and is packaged into new virus particles.

The final step of the viral cycle, assembly of new HIV-1 virons, begins at the plasma membrane of the host cell. The Env polyprotein (gp160) goes through the endoplasmic reticulum and is transported to the Golgi complex where it is cleaved by protease and processed into the two HIV envelope glycoproteins gp41 and gp120. These are transported to the plasma membrane of the host cell where gp41 anchors the gp120 to the membrane of the infected cell. The Gag (p55) and Gag-Pol (p160) polyproteins also associate with the inner surface of the plasma membrane along with the HIV genomic RNA as the forming virion begins to bud from the host cell (Figure 1.3). Maturation either occurs in the forming bud or in the immature virion after it buds from the host cell. During maturation, HIV proteases cleave the polyproteins into individual functional HIV proteins and enzymes. The virus structural components then assemble to produce a mature HIV virion. The mature virus is then able to infect another cell.

HIV differs from other viruses with its high genetic variability. The diversity within the HIV population is a result of its fast replication cycle, with the generation of about $10^{10}$ virions every day, coupled with a high mutation rate of approximately $3 \times 10^{-5}$ per nucleotide base per cycle of replication and recombinogenic properties of reverse transcriptase \[19\]. This scenario
leads to the production of many variants of HIV in a single infected patient in the course of a day [3].

**Aim and outline of this thesis**

The research described in this manuscript is aimed at bridging the gap between the protein interaction data (without spatio-temporal scale) of HIV and human proteins, all the way up to the clinical observation of the co-receptor switch of HIV infection (where both time and space are of foremost importance) with the focus on extra-cellular proteins.

We choose a *bottom-up* approach with the protein-level being the starting point due to the availability of data and high degree of freedom for analysis, thanks to the wide range of interactions and the lack of temporal scale. First we construct the complete HIV-human protein interaction map and analyze it for centrality and relative importance of its components. Then we introduce spatial information by concentrating only on the interactions that take place outside the cell membrane. Later we use these interactions to predict novel proteins, which are potentially involved in HIV infection. Finally, for studying the evolutionary dynamics of HIV infection, time is introduced in the computational model of HIV tropism (virus affinity for different receptors).

![Figure 1.4](image)

*Figure 1.4 The schema showing the outline of the manuscript. Figure a) represents the total HIV – human protein interaction network where the blue nodes represent human proteins and red nodes represent HIV proteins. Figure b) focuses on a specific part of the total network, extracellular proteins namely. In figure c) novel potential HIV-interacting extracellular proteins are predicted (green nodes) with our network centrality inspired algorithm (See chapter 4 Methods). Figure d) focuses on the temporal aspects of the HIV infection with extracellular proteins from figure c).*

In chapter 2, we introduce the problems related to HIV infection and the current status of the field today. The aid of computational science in today’s treatment is briefly described on ViroLab experimental platform, which contains computational models that cover various spatial and temporal scales from atomic level interactions in nanoseconds up to sociological interactions on the epidemiological level, spanning years of disease progression.

Recently, the National Institute of Allergy and Infectious Diseases has launched the HIV-1 Human Protein Interaction Database in an effort to catalogue all published interactions between HIV-1 and human proteins. In order to systematically investigate these interactions functionally and dynamically, we have constructed an HIV-1 human protein interaction network. **Chapter 3** focuses on the protein interactions between HIV and human proteins and the underlying complex network. This network was analyzed for the most relevant proteins and associated
processes that are specific for the HIV life cycle. In order to expose viral strategies, network motif analysis was carried out showing reoccurring patterns in virus-host dynamics.

Our analyses in the third chapter show that human proteins interacting with HIV form a densely connected sub-network within the total human protein interaction network. The evaluation of this sub-network for connectivity and centrality identified a set of proteins essential for the HIV life cycle. Remarkably, we were able to associate proteins involved in RNA polymerase II transcription with hubs and proteasome formation with bottlenecks, which were overlooked otherwise. Inferred network motifs show significant over-representation of positive and negative feedback patterns between virus and host. Strikingly, such patterns have never been reported in combined virus-host systems.

The results of the third chapter indicate that HIV infection results in a reprioritization of cellular processes reflected by an increase in the relative importance of the transcriptional machinery and proteasome formation. We conclude that during the evolution of HIV, some patterns of interaction have been selected which result in a system where virus proteins preferably interact with central human proteins for direct control and with proteasomal proteins for indirect control over the cellular processes. Finally, the patterns described by the network motifs illustrate how virus and host interact with one another. The results of this chapter are of paramount importance since it underlines important features of “the complete HIV – human protein interaction network”.

HIV infection affects the populations of T helper cells, dendritic cells and macrophages. Besides, it has a serious impact on the central nervous system. It is yet not clear whether this list is complete and why specifically those cell types are affected. To address this question, in chapter 4 we introduce spatial information to the HIV-Human protein interaction network from chapter-3 by focusing only on the surface membrane proteins interacting with HIV proteins. In this chapter we describe our method to identify cellular surface proteins that permit, mediate or enhance HIV infection in different cell/tissue types in HIV-infected individuals. Receptors associated with HIV infection share common functions and domains, and are involved in similar cellular processes. These properties are exploited by graph theory and a novel gene-ranking algorithm to predict unprecedented surface membrane proteins (SMP) potentially interacting with HIV.

We compiled a set of 13 SMPs that are known to interact with HIV from the HIV-1 protein interaction network from chapter 3. This set is extended by proteins that have direct interaction and share functional similarity. This resulted in a comprehensive network around the initial SMP set. Using network centrality analysis, graph theory, GeneOntology and a gene-ranking algorithm we generate a set of 21 surface membrane factors that constitutes a well-founded starting point for experimental testing of cell/tissue susceptibility of different HIV strains as well as for cohort studies evaluating patient specific disease progression.

Among these 21 surface membrane factors three have confirmed functions in HIV infection, seven have been identified by at least two other studies, and 11 are novel predictions and thus excellent targets for experimental investigation. In conclusion, our findings constitute the necessary background for future research investigating the role of SMPs during infection with HIV.

In chapter 5 we use a computational model with a clear temporal scale for studying the longitudinal dynamics of the well-established receptors such as CD4, CCR5 and CXCR4 from chapter 4 and HIV virions. In this chapter we will focus on the two main co-receptors of HIV-1, namely CCR5 and CXCR4, and their role in the infection dynamics thus introducing the temporal component of the system.
CCR5 and CXCR4 are expressed on the cells of the immune system. They are primarily responsible for successful entry of the viruses into the CD4+ T cells. Viruses using the CCR5 receptor (R5-tropic) are predominant in the early infection and viruses using the CXCR4 receptor appear in the later stages. This so called “co-receptor tropism switch” in HIV-1 infected individuals, is observed in 50% of the cases, accompanied with a rapid decline of CD4+ count and a spike in the number of viruses in the bloodstream (viremia).

To investigate the co-receptor switch of HIV-1 we have developed a model, which takes into account only target cells and infecting virions. Computational model implementation has been carried out as an agent based model with special emphasis on the spatial characteristics of virus/cell interactions in the lymph nodes. Robustness of the computational model is verified through perturbation analysis of the key model parameters.

Given the transmission with an R5 tropic virus, we evaluate the fitness of the viral quasispecies by tuning the co-receptor tropism. The emergence rate of the X4 tropic variants, the factors leading to it and the necessary conditions for the co-receptor switch, are investigated to elucidate whether accumulation of mutations induce the co-receptor switch.

Our results in chapter 5 are two fold; first, we have developed and verified an intuitive simulation platform of HIV-1 tropism for testing potential scenarios. Second, the natural evolution of the viral population over time resulted in the co-receptor switch without inherent assumption of X4-tropic variant superiority. Moreover, our studies indicate that there is an optimum mutation rate for the co-receptor switch and interestingly, it coincides with the experimentally reported mutation rate of HIV-1.

The time of the co-receptor switch depends on the fluctuations in the HIV population size. The bottlenecks in the viral population size increases the chance of occurrence of the co-receptor switch, therefore reducing the time required for the co-receptor switch. R5-tropic variants benefit from the emergence of X4 tropic variants and outcompete dual-tropic variants together leading to sympatric speciation of HIV-1 in-vivo. This implies a strong link to in-vivo compartmentalization of the virus while adapting to its host.

Finally, in chapter 6 we discuss the significance of our findings, put them in perspective and conclude with final remarks.

Chapter References


Bastian et al. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?. PLoS Med (2010) vol. 7 (9) pp. e1000326


