Journal of Clinical Monitoring and Computing (2005) 19: 263–278
DOI: 10.1007/s10877-005-0673-2


# A GRID-BASED HIV EXPERT SYSTEM

*Peter M.A. Sloot,*[1] *Alexander V. Boukhanovsky,*[2] *Wilco Keulen,*[3] *Alfredo Tirado-Ramos,*[1] *and Charles A. Boucher*[4]

Sloot P MA, Boukhanovsky AV, Keulen W, Tirado-Ramos A, Boucher CA. A grid-based HIV expert system.
J Clin Monit Comput 2005; 19: 263–278

**ABSTRACT. Objectives**. This paper addresses Grid-based integration and access of distributed data from infectious disease patient databases, literature on *in-vitro* and *in-vivo* pharmaceutical data, mutation databases, clinical trials, simulations and medical expert knowledge. **Methods**. Multivariate analyses combined with rule-based fuzzy logic are applied to the integrated data to provide ranking of patient-specific drugs. In addition, cellular automata-based simulations are used to predict the drug behaviour over time. Access to and integration of data is done through existing Internet servers and emerging Grid-based frameworks like Globus. Data presentation is done by standalone PC based software, Web-access and PDA roaming WAP access. The experiments were carried out on the DAS2, a Dutch Grid testbed. **Results**. The output of the problem-solving environment (PSE) consists of a prediction of the drug sensitivity of the virus, generated by comparing the viral genotype to a relational database which contains a large number of phenotype-genotype pairs. **Conclusions**. Artificial Intelligence and Grid technology are effectively used to abstract knowledge from the data and provide the physicians with adaptive interactive advice on treatment applied to drug resistant HIV. An important aspect of our research is to use a variety of statistical and numerical methods to identify relationships between HIV genetic sequences and antiviral resistance to investigate consistency of results.

**KEY WORDS.** computational Grids, HIV, PSE, expert system, artificial intelligence, bio-statistics.

## 1. INTRODUCTION

### *1.1. Motivation*

Forty two million people worldwide have been infected with HIV (Human Immunodeficiency Virus) and 12 million have died, over the last 20 years. Figure 1 shows the pan-epidemic extent of HIV infections.

Effective antiretroviral therapy has lead to sustained HIV viral suppression and immunological recovery in patients who have been infected with the virus. The incidence of AIDS has declined in the Western world with the introduction of effective antiretroviral therapy, though questions on "When to start treatment? What to start with? How to monitor patients?" remain heavily debated. Adherence to antiretroviral treatment remains the cornerstone of effective treatment, and failure to adhere is the strongest predictor of virological failure. Long-term therapy can lead to metabolic complications. Other treatment options are now available, with the recent introduction to clinical practice of fusion inhibitors, second-generation non-nucleoside reverse transcriptase inhibitors, and nucleotide

From the [1]Section Computational Science, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands, [2]Institute for High Performance Computing and Information Systems, Bering St, 38, St. Petersburg, Russia, [3]Virology Education, 69042 Utrecht, The Netherlands, [4]University Medical Center, University of Utrecht, 3508 GA Utrecht, The Netherlands.

Received and accepted for publication June 30, 2005.

Based on "A Grid-based HIV Expert System", by P.M.A. Sloot, A.V. Boukhanovsky, W. Keulen, and C.A. Boucher, which appeared in the IEEE/ACM International Symposium on Cluster Computing and the Grid, Cardiff, UK, May 9-12, 2005. ©2005 IEEE.

Address correspondence to Peter M.A. Sloot, Section Computational Science, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
E-mail: sloot@science.uva.nl

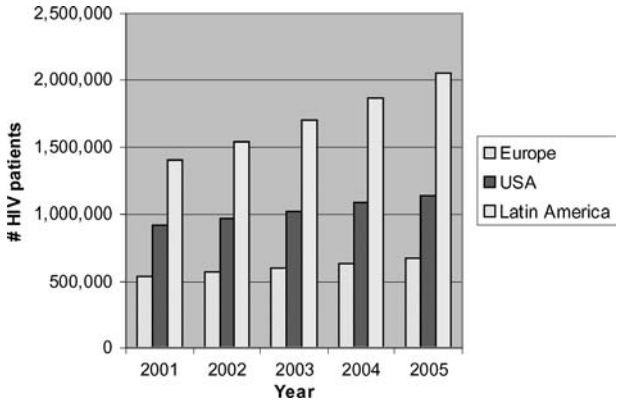


*Fig. 1. Worldwide spread of HIV infections, history and near future perspective.*

reverse transcriptase inhibitors. The sheer complexity of the disease, the distribution of the data, the required automatic updates to the knowledgebase and the efficient use and integration of advanced statistical and numerical techniques necessary to assist the physician motivated us to explore the novel possibilities supported by Grid technology.

In this position paper we describe ongoing research in our 3 laboratories (Utrecht, St. Petersburg and Amsterdam) addressing the development of a Grid based medical decision support system. The goal of the research is to investigate novel computational methods and techniques that support the development of a user friendly integrated support system for physicians. We use emerging Grid-technology to combine data discovery, data mining, statistical analyses, numerical simulation and data presentation [1].

The paper is organized as follows. The rest of Chapter 1 describes the background of HIV research and a prototypical rule-based approach to data analysis. In chapter 2 we give an overview of the two computational techniques we study to understand the temporal variability of HIV populations through stochastical modeling, the evolution of HIV infection and the onset of AIDS through Cellular Automata (CA) modeling. Chapter 3 describes a first approach to advanced data presentation through roaming devices such as Personal Digital Assistants (PDAs), as well as our Virtual Organization-based (VO) Grid approach. Finally, Chapter 4 offers a brief discussion on our conclusions and future work.

## *1.2. Background*

### *1.2.1. Clinical aspects of HIV*

The clinical management of patients infected with Human Immunodeficiency Virus (HIV) is based on studies on the pathogenesis of the disease and the results of trials evaluating the effects of anti-HIVdrugs. Retrospective analysis of large cohorts has identified laboratory markers for disease progression, such as the amount of virus (HIV-RNA) and the number of T helper cells (CD4 + cells) in blood. In addition the results of prospective drug trials have generated data on effectiveness of individual drugs and drug combinations and the effect of drug resistant viruses on therapy outcome. Currently clinicians are limited in the practical use of this information because in most cases they are only provided with statistical relationships between individual parameters and disease or therapy outcome. Large data sets have not been analyzed and made available in such a way that it allows a clinician to use the available data in more clinical settings. The availability of large databases and the development of innovative data mining approaches create the opportunity to develop systems which allow the practicing clinician to determine the risk profile for disease development, or the change or success for a given regimen for his individual patients. Such a system will determine the rate of success for different drug regimens by taking into account the effect and interaction of all relevant laboratory and clinical parameters and by comparing the results for similar patients available in the database.

Currently there are fifteen drugs licensed for treatment of individuals infected with HIV. These drugs belong to two classes, one inhibiting the viral enzyme reverse transcriptase and another inhibiting the viral protease. These drugs are used in combination with therapy to maximally inhibit viral replication and decrease HIV-RNA to below levels of detection levels (currently defined as below 50 copies per ml) in blood. Treatment with drug combinations is successful in inhibiting viral replication to undetectable levels in only 50% of the cases. In the remaining 50% of cases viruses can be detected with a reduced sensitivity to one or more drugs from the patients' regimen. The molecular base for resistance has been, and still is, focus of extensive research. Over 80 amino acid positions in the viral enzyme reverse transcriptase (RT) and 40 positions in the protease enzyme can undergo changes when exposed to selective drug pressure in vitro or in vivo. For some drugs, at certain positions, a change towards a specific new amino acid is seen. At other positions several alternative amino acids may appear and cause (variable) levels of resistance to one or more drugs. In theory, therefore, an infinite number of combinations of amino acid changes could appear and cause resistance in vivo. Preliminary clinical observations however show that specific amino acid changes at a limited number of positions and a limited number of combinations prevail. In addition to changing drug sensitivity some amino acid changes may also influence the replication potential of HIV. Amino acids selected initially during a failing regimen cause resistance to the drugs the patient is taking,

but at the same time may decrease the capacity of the virus to replicate. Changes appearing later do not function to further increase resistance but merely function to restore the capacity of the virus to replicate ("viral fitness"). Several clinical studies have been performed recently to evaluate the clinical benefit of resistance-guided therapy. These studies show that a better virological response is obtained in patients who are failing their therapy, when their new regimen is chosen on the basis of their resistant profile. In three out of the four studies from last year the results showed that if new regimens were selected on the basis of the mutations (viral resistance genotype) the results were better as compared to standard care approaches. Currently, the basis for clinical interpretation of the viral genotype is based on data sets relating mutations to changes in drug sensitivity, and/or data sets directly relating mutations present in the virus to clinical responses to specific regimens. Initially, experts compared the observed mutations to lists of published sequences taken from the literature, and based on this comparison would select a regimen.

### *1.2.2. Prototype support system*

Recently, first generation bioinformatics software programs have been developed to support clinicians. Examples of such systems are the Virtual Phenotype developed by Virco NV, and a first generation decision support system (Retrogram TM) developed by Virology Networks BV in collaboration with parts of our research team. The output of these programs consists of a prediction of the drug sensitivity of the virus generated by comparing the viral genotype to a relational database containing a large number of phenotype-genotype pairs. The Retrogram decision software interprets the genotype of a patient by using rules developed by experts on the basis of the literature, taking into account the relationship of the genotype and phenotype. In addition, it is based on (limited) available data from clinical studies and on the relationship between the presence of genotype directly to clinical outcome. It is important to realise however that these systems focus on biological relationships and are limited to the role of resistance. The next step will be to use clinical databases and investigate the relationship between the viral resistance profile (mutational profile and/or phenotypic data) and therapy outcome measures such as amount of virus (HIV-RNA) and CD4+ cells. A summary of the flow of data is shown in Figure 2.

### *1.2.3. Data collection*

Large high quality clinical and patient databases are used to explore the relationships described above and to develop a first prototype matching system.

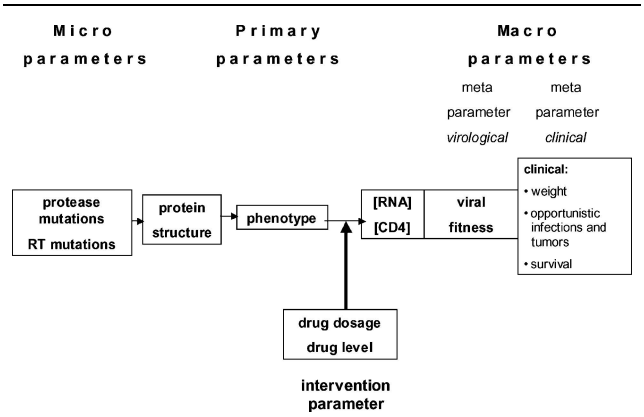


*Fig. 2. From molecule to man: Hierarchical data flow model for infectious diseases.*

The Viradapt study showed that the virological response was better in the patient group in which genotype and rule-based interpretation was used as compared to the standard of care arm [2]. On the basis of these results, a more elaborate decision support software system (Retrogram version 1.0) was built in collaboration with Virology Networks B.V. This system ranks the efficacy of the antiretroviral drugs within each class. The ranking is based on expert interpretation of two types of data. The software system estimates the drug sensitivity for the fifteen drugs by interpreting the genotype of a patient by using mutational algorithms. These mutational algorithms are developed by a group of experts on the basis of the scientific literature, taking into account the published data relating genotype to phenotype. In addition, the ranking is based on data from clinical studies on the relationship between the presence of particular mutations and clinical or virological outcome.

The Athena cohort is a large Dutch observational clinical cohort study aiming at the surveillance of antiretroviral treatment supported by the Dutch government. The cohort consists of 3000 patients from whom clinical, virological, immunological and data on drug side effects are centrally collected through a decentralised data entry system. Within this cohort 600 patients are studied intensively, phenotypic and genotypic data, drug levels and CD4+ and HIV-RNA patterns are collected. From two large international trials (sponsored by Roche Pharmaceuticals) evaluating the effect of a new fusion inhibitor drug (T20), representing 1000 patients from whom also phenotype, genotype, viral fitness, drug levels as CD4+ and HIV-RNA patterns will be collected. The third database will be from the international multi-center Great study sponsored by Virology Networks BV, within this study the value of the Retrogram decision support program is evaluated and similar parameters a described above will be collected, within

this study 360 patients will be enrolled. Another dataset will come from the Italian Musa study, in this trial data will be collected from 450 patients followed over a year. Entry point to the trial is failing a fist or second regimen, subsequently patients will be genotyped and a new regimen will be selected on the basis of Retrogram 1.4 or the Virtual Phenotype from Virco (Belgium).

Throughout the duration of the project we will collect additional datasets. These datasets may serve to further refine our models and first version software and may also be use to perform validation studies.

### *1.2.4. Data analysis*

The primary goal of the data analysis is to identify patterns of mutations (or naturally occurring polymorphisms) associated with resistance to antiviral drugs and to predict the degree of *in-vitro* or *in-vivo* sensitivity to available drugs from an HIV genetic sequence. The statistical challenges in doing such analyses arise from the high dimensionality of these data. A variety of approaches have been developed to handle this type of data, including clustering, recursive partitioning, and neural informatics. Neural informatics is used for synthesis of heuristic models received by methods of knowledge engineering, and results of the formal multivariate statistical analysis in uniform systems. Clustering methods have been used to group sequences that are "near" each other according to some measure of genetic distance [3]. Once clusters have been identified, recursive partitioning can be used to determine the important predictors of drug resistance, as measured by *in-vitro* assays or by patient response to antiviral drugs. Principle component analyses can help to identify what are the most important sources of variability in the HIV genome. An important aspect of our research is to use a variety of methods to identify relationships between HIV genetic sequences and antiviral resistance to validate the consistency of results.

The molecular sequences of the viral enzymes reverse transcriptase and protease are the micro parameters in the model. In theory an infinite number of combinations of mutations could appear and cause (variable) changes in viral drug sensitivity and viral replication capacity (See also Table 1). Clinical datasets however show that specific amino acid changes at a limited numbers of positions in a limited number of combinations prevail. HIV-RNA and CD4 are the primary parameters determining disease outcome. HIV-RNA, the amount of HIV-RNA genomic copies per ml plasma, has been validated as being highly predictive of clinical outcome. HIV-RNA and CD4+ cell numbers are now the standard endpoint in clinical trials for approval of new antiretroviral drugs. A patient's HIV-RNA may range between a few hundred to millions of RNA copies per ml plasma. The CD4+ cell numbers in peripheral blood range typically between zero and thousand. Whereas the predictive clinical value of both parameters has been determined initially in untreated individuals, they have also been shown to be of predictive value also for patients under antiretroviral therapy. Recently observations have been published indicating that in some patients under highly active antiretroviral therapy (HAART) a disconnect may occur between the response in HIV-RNA and in CD4 counts. Typically, in these patients a rise in HIV-RNA as consequence of incomplete inhibition of viral replication under therapy is not paralleled by a continuous decrease in CD4 counts. This disconnect has been explained by a decrease in the viral replicative capacity ('viral fitness') which leads to a decrease in capacity to lower CD4 counts.

The patient's weight and secondary opportunistic infections and/or malignancies are parameters that determine disease outcome and survival time. Currently there are fifteen drugs licensed for treatment of individuals infected with HIV: More than ten inhibitors have been developed which inhibit the reverse transcriptase process. These inhibitors can be classified in two sub-categories that differ in the way they inhibit the RT-enzyme, nucleoside (analogue) RT-inhibitors (NRTI) and the non-nucleoside RT-inhibitors (NNRTI). These compounds inhibit the protease enzyme, which acts much later on in the HIV replication cycle than reverse transcriptase.

The protease is responsible for cleaving a long polyprotein into smaller functional proteins. The overall exposure to antiretroviral drugs has been shown to be an

*Table 1. Parameters for the data analyses. Here the hierarchical approach shown in Figure 2 is extended to detail the content of the parameters*

| | | |
|---|---|---|
| Micro Parameter | | Protease Mutations Reverse Transcriptas Mutations |
| Primary Parameter | | HIV-RNA CD4 Drug Resistance |
| Macro Parameter | Meta Parameter: Virological | Viral Fitness |
| | Meta Parameter: Clinical | Weight Opportunistic Infections and Tumors Survival |
| Intervention Parameter | | Drug Dosage Bio-availability of Drug/Drug Level |

important factor for the degree of success for a given therapy. The overall exposure can be captured by parameters as dosage and bio-availability which will codetermine the drug level within an individual patient. Given the relationships between exposure and antiviral efficacy, variability in drug levels (which may be due to differences in patient adherence to their regimens) will contribute to virological and immunological outcome. Individuals with relatively low exposure are more likely to experience virological failure than those with a high exposure.

## 2. METHODS AND MATERIALS

### *2.1. Modeling the dynamics and temporal variability of HIV-1 populations*

In addition to rule based and parameter based decision support we developed statistical models and cellular automata based models to study the dynamics of the HIV populations. These 2 numerical models run on Grid-resources. The output is integrated with the medical support system and accessible to the end-user. In this paragraph we briefly outline the two computational methods. Details are beyond the scope of this paper; we refer to the references provided.

#### *2.1.1. A cellular automata model to study the evolution of HIV infection and the onset of AIDS*

A cellular automata model to study the evolution of HIV infection and the onset of AIDS is developed. The model takes into account the global features of the immune response to any pathogen, the fast mutation rate of the HIV, and a fair amount of spatial localization, which may occur in the lymph nodes. The dynamics of the cellular automata requires high throughput computing, which is provided by the resource management of the Grid. In this section, we employ non-uniform Cellular Automata (CA's) to simulate drug treatment of HIV infection, in which each computational domain may contain different CA rules, in contrast to normal uniform CA models. Ordinary (or partial) differential equation models are insufficient to describe the two extreme time scales involved in HIV infection (days and decades), as well as the implicit spatial heterogeneity. Zorzenon dos Santos et al. [7] reported a cellular automata approach to simulate three-phase patterns of human immunodeficiency virus (HIV) infection consisting of primary response, clinical latency and onset of acquired immunodeficiency syndrome. We developed a non-uniform CA model to study the dynamics of drug therapy of HIV infection, which simulates four-phases (acute, chronic, drug treatment responds and onset of AIDS). Our results indicate that both simulations (with and without treatments) evolve to the same steady state. Three different drug therapies (mono-therapy, combined drug therapy and HAART) can also be simulated in our model. Our model for prediction of the temporal behaviour of the immune system to drug therapy qualitatively corresponds to clinical data.

Pseudo Code 1a: HI Model (Adapted from Zorzenon dos Santos R. M., Phys. Rev. Let. 2001). H = healthy cell, A1 and A2 are infected cells at different time steps.

| | |
|---|---|
| Assume: | {H, A1(t), A2(t+ $\tau$), D}; 1 time-step = 1 week; Simulation of lymph-node; Moore neighbourhood and square lattices used |
| Rule 1: | (a) If it has at least one infected-A1 neighbor, it becomes infected-A1 |
| | (b) If it has no infected-A1 neighbor but does have at least R ($2 < R < 8$) infected-A2 neighbors, it becomes infected-A1 |
| | (c) Otherwise it stays healthy |
| Rule 2: | An infected-A1 cell becomes infected-A2 after $\tau$ time steps |
| Rule 3: | Infected-A2 cells become dead cells |
| Rule 4: | (a) Dead cells can be replaced by healthy cells with probability *p*repl *in the next step.* |
| | (b) Each new healthy cell introduced may be replaced by an infected-A1 with probability *p*infec |

This CA (Pseudo-code 1a) mimics in a simple way the dynamical properties of a HIV infection; next we introduce drug therapy into the model by modelling a response function Presp and changing only rule 1.

Pseudo Code 1b: Advanced HI Model, taking into account drug therapy effects.

Rule 1:

(a) If there is one A1 neighbor after the starting of drug therapy, $N(0 \leq N \leq 7)$ neighbor healthy cells become infected-A1 in *the next time steps* with probability *p*resp. Otherwise, all of eight neighbors become infected-A1.

$N$ represents effectiveness of drugs.

$N = 0$: no replication;

$N = 7$: less effective for the drug.

*P*resp $(t - t_s)$ represents certain response function of drug effects over the time steps ($t$). The $t_s$ is the starting of treatment.

The main success of the presented CA model is the adequate modeling of the four-phases of HIV infection with different time scales into one model. Moreover, we could also integrate all of the three different therapy procedures. The simulations show a qualitative correspondence to clinical data. During the phase of drug therapy response, temporal fluctuations for $N > 3$ were observed, this is due to the relative simple form of the response distribution function ($P_{\text{dis}}$)applied to the drug effectiveness parameter $N$ at each time-step. The simulation results indicate that, in contrast to ODE/PDE, our model supports a more flexible approach to mimic different therapies through the use of mapping the parameter space of $P_{\text{dis}}$ to clinical data. Therefore there is ample room to incorporate biologically more relevant response functions into the model. The data integration required for the CA, the parametric computation and the data presentation are supported by the Grid.

### *2.1.2. Multivariate stochastic modeling*

The modeling of Human Immunodeficiency Virus (HIV-1) genotype datasets has a goal to identify patterns of mutations (or naturally occurring polymorphisms) associated with resistance to antiviral drugs and to predict the degree of *in-vitro* or *in-vivo* sensitivity to available drugs from an HIV-1 genetic sequence. The statistical challenges in doing such analyses arise from the high dimensionality of these data. Direct application of the well-known genetic approaches [5] to analysis of HIV-1 genotype results in a lot of problems. Principal difference is in the fact that, in HIV DNA analysis, the main scope of interests is the so-called relevant mutations – a set of mutations, associated with the drug resistance. These mutations might exist in different positions over the amino-acid chains. Moreover, the sheer complexity of the disease and data require the development of the reliable statistical technique for its analysis and modeling. A multivariate stochastic model for describing the dynamics of complex non-numerical ensembles, such as observed in the (HIV) genome, has been developed in [6]. This model was based on principle component analyses for numerated variables. Generally speaking, the interpretation of numerated variables in terms of relevant mutations is not clear. Below we develop this model directly for the ensemble of relevant mutations in the RT and protease parts of the HIV-1 genome. Each element of the ensemble is presented as the cortege $\Xi_k = \{\xi_j\}_{j=1}^{n_k}$, $k = \overline{1, M}$ with the variable dimension $n_k$-the total number of the mutations in the gene. Each value $\xi_k$ is a literal index and corresponds the position and new value of the amino acid (e.g., 184 V, 77I, etc.). It allows to associate each mutation with the categorical random variable $i \in 1 \ldots K$, where $K$ is the total number of possible mutations. Each sub sample of genomes with a fixed number of mutations $n = \text{const}$ may be considered as the realizations of a categorical random vector.

The representation above is based on the proximity to the "wild-type" virus and takes into account only the relevant mutations in a genome. It allows for significant compression of the DNA representation and simplifies the interpretation of the results.

*Principle of the modeling approach*. The joint variability of different mutations in the HIV-1 genomes is a complicated phenomenon. The dimension of the probabilistic characteristics is high, and its analytical investigations and interpretation are hard. Hence, for the studying of HIV-1 populations we use a computational statistical approach that allows to numerically generate an ensemble with the same probabilistic properties by means of a Monte-Carlo procedure. This is a well-known powerful method to study complex system variability.

The idea of the stochastic modeling is shown in the Figure 5. It is based on the evolutionary hypothesis, considering the group with $n+1$ mutations as subgroup of group with $n$ mutations in a previous step. For each gene the transit from $n$ to $n+1$ mutation groups is driven by a stochastic operator $D_{(n+1)}$, which defines the mutations on the $n+1$ step, when the mutations on the previous $n$ steps are known. The initial step of the stochastic procedure begins from the whole ensemble of wild-type viruses. The number of the genomes that has been mutated at each step of the stochastic procedure is in accordance with $M_n = \rho_n M$, where $\rho_n$ are the probabilities of the occurrence of genotypes with $n$ mutations in a total population of $M$ genes.

The stochastic operator $D$ may be considered as a "black box". It is formalized in terms of the conditional probabilities of the occurrence of mutation $\xi_i$, if the mutation $\xi_j$ arise in the previous step of the generation. For genotypes with 2 mutations only the values $D_{ij}$ are the conditional

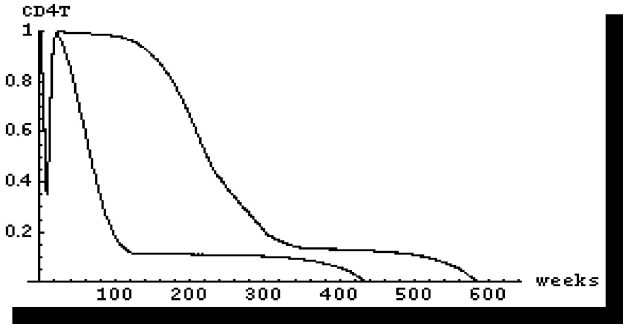


*Fig. 3. Temporal behaviour of the CD4 count, with modeled Brownian movement for lymphocytes [8].*

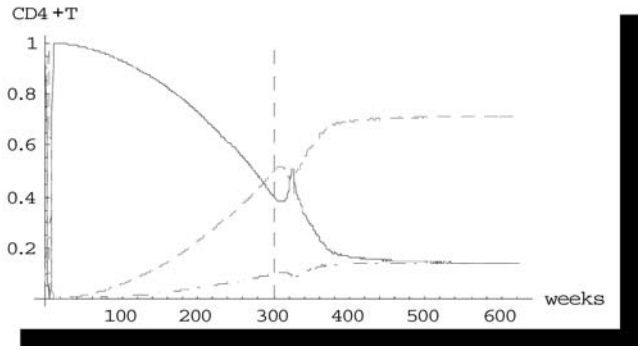


*Fig. 4. As in Figure 3, with additionally modeled mono therapy in week 300 [8].*

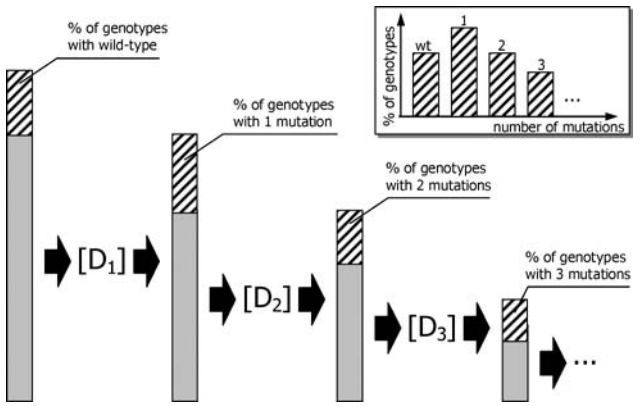


*Fig. 5. Principle of the modeling.*

probabilities of the pairs. In this case the matrix $\{D_{ij}\}$ is the transition Markov probability matrix, containing the conditional probabilities for simple Markov chains with the number of these states corresponding to quantity of the relevant mutations. In more complicate cases, where $n > 2$, the probability matrix $\{D_{ij}\}$ consists of the conditional probabilities to meet mutation $\xi_j$ in certain gene, when the mutation $\xi_i$ is present.

This approach allows us to reduce the complicated statistical description of the dataset to a rather simple model, using only three probabilistic distributions as the initial parameters of the model: distribution of number $n$ of the mutations $\rho_n$;

- distribution $P_{\xi}^{(1)}$ for the relevant mutations in the group $n = 1$;
- transient probability matrix $D$.

All these parameters might be identified on the sample datasets of the HIV-1 population.

*Identification of the model.* For the identification of parameters of the model, a large database of HIV-infected patients, collected over several years in USA, is used [4]. These databases contain genotypes of 43620 patients examined from August 9, 1998 to May 5, 2001. We observed 59 different mutations in the RT genome, including 17 mixed mutations, and 77 different mutations in the protease genome, including 34 mixed mutations.

*Distribution $\rho_n$ of number of mutations.* The practice of HIV treatment however, has shown that the variability of the number of mutations $n$ is high, due to the complexity of the drug combinations that has been applied. The sample estimate of distribution $\rho_n$ of the number of mutations in protease is shown in the Figure 6. It is seen, that the distributions have a clear first peak ($n = 1$), and a shelf (or second peak), corresponding to $n = 3 \div 5$. Therefore we expect that there are two groups of genomes in the database, corresponding to the low and high number of mutations. The possible interpretation of the discovered bi-modal distribution is that we have two groups of patients. One group is the "new" patients who had one or two treatments, thus their genotype contains relative small numbers of mutations. The second group is the "old" patients, which have a long treatment history, or new patients, infected through treated HIV-1 patients [15].

*Distributions of the relevant mutations $P_{\xi}$.* Distribution $\rho_n$ allows describe the variability of the groups of the "new"

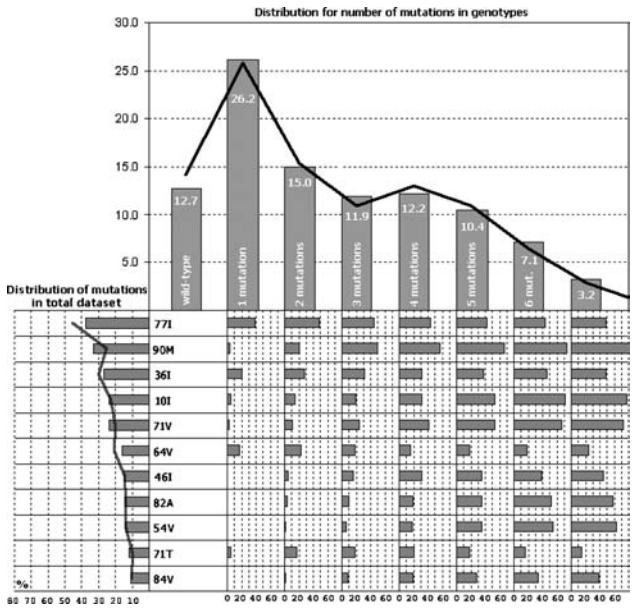


*Fig. 6. Statistical description for distribution of mutations in Protease.*

and "old" patients, only. For a more detailed study of the virus mutations driving by the certain drugs combinations, the probabilities of occurrence of the relevant mutations $\xi$ should be considered. They are estimated by the sample frequencies:

$$P_\xi = \frac{\{\text{Number of genes with mutation } \xi\}}{M}. \quad (1)$$

Here $M$ is the total number of genomes in the dataset. Equation (1) describes the marginal impact of each mutation in the total population, without any information about number and occurrences of other mutations. The probabilities of the most significant relevant mutations $\xi_k$ (in decreasing order of its probability) are shown in Figure 6. The marginal estimates of $P_\xi$ over the total dataset show only general impacts of the mutations. For a detailed analysis of its behavior we also consider the occurrences $P_\xi^{(n)}$ of mutations in the groups of genotypes with exactly $n$ mutations. These values were computed also by means of Equation (1), where $M \stackrel{\text{def}}{=} M_n = \rho_n M$ – the number of genes with $n$ mutations in a database. The sample estimates of these occurrences are also shown in the Figure 1. It is clearly seen that the inputs of some mutations are rather different for different $n$, both for the protease and RT parts of the genome. E.g., for RT, for $n = 1$, the mutations 184 V and 103 N have the main input. The distribution $P_\xi^{(1)}$ is the limit distribution from the procedure shown in Figure 5.

From Figure 1 we also observe that the total sum $\sum_k P_{\xi_k} > 100\%$, excluding case $n = 1$. This demonstrates that the analysis of the marginal mutations is not enough for general statistical description of all DNA ensemble variability, because some positions of DNA may be statistically dependent [15], especially in relation to viral fitness. Hence, the joint characteristics of its variability must be taking into account.

*Transient probability matrix D.* The conditional probability of the occurrence of mutation $\xi_i$, if the mutation $\xi_j$ arises from the previous steps of the generation, is estimated by:

$$D_{ij} = \frac{\{\text{Number of genotypes with mutations } \xi_i \text{ and } \xi_j \text{ simult\&aneously}\}}{\{\text{Number of genotypes with mutation } \xi_i\}}. \quad (2)$$

The dimensionality of the related matrix, obtained from Equation (2), may be rather high. In order to decrease the dimensionality we consider the algebraic technique of orthogonal expansion, applied to transient probability matrices [16].

$$D = \Phi \Lambda^{1/2} \Psi. \quad (3)$$

*Table 2. Normalized (%) values of the expansion coefficients $\lambda_k$ in Equation (4)*

| | # of PC | | | | | | |
|---|---|---|---|---|---|---|---|
| Part of the genome | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| RT | 61.3 | 8.2 | 5.4 | 2.8 | 2.1 | 1.7 | 1.6 |
| Protease | 55.0 | 6.3 | 4.5 | 4.2 | 3.4 | 2.7 | 2.4 |

where $\Phi$ are the eigenvectors of matrix $DD^T$, and $\Psi$-of matrix $D^T D$. It allows considering the coefficients $a_k = \sqrt{\lambda_k}$ as the principal components (PC) [13], and represents the probability (2) as a series:

$$D_{ij} = \sum_k \sqrt{\lambda_k} \phi_{ik} \psi_{jk}. \quad (4)$$

The values $\lambda_k$ shows the part of the probability, explained by $k$-th PC. The sum of the first $k$-th coefficients $\lambda_k$ may be interpreted as a measure of convergence of the series (4). In Table 2 the values of the first 7 $\lambda_k$ for the RT and protease parts of the HIV-1 genome are shown. These data were obtained for the total database. It can be seen that the series (4) converges rather fast in both cases: e.g. for the RT part only the first term of the series explain more 60% of conditional probability (the first five terms explain 80%).

Let us consider the normalized bases $\tilde{\phi}_{ik} = \lambda_k^{0.25} \phi_{ik}$, $\tilde{\psi}_{jk} = \lambda_k^{0.25} \psi_{jk}$. It allows to present the terms in Equation (4) as the $p_k^{ij} = \tilde{\phi}_{ik} \tilde{\psi}_{jk}$ and interpreted these values as the independent factor loadings, driving the changes of the conditional probability $D_{ij}$ over all the mutations $\xi_i, \xi_j$ in the database. For example, in the Figure 7 the estimates of the first basic functions are shown for RT and protease parts of the genotype (the input of multiplication of functions are in the Table 2). It is clearly seen, that the first term $p_1^{ij} = \tilde{\phi}_{i1} \tilde{\psi}_{j1}$ reflects the total occurrence of the mutations in a genotype (see Figure 6): for the mutations with the maximal occurrences the input to conditional probabilities of its pairs is also high.

*Model validation.* The simulation model is based on the $\rho_n$, $P_\xi^{(1)}$, $D$ distributions of the mutations only. No information of more complicate mechanisms (distributions of pairs, triples, etc.) has been used for this identification.

The main goal of the verification is the possibility to reproduce these features of the ensemble through the dependencies formalizing the matrix $D$. We compared the total occurrences of all mutations in genotypes, estimated on the initial and simulated samples, see also Figure 6 (solid line). It is seen, that the results of the simulation and sample are rather close.

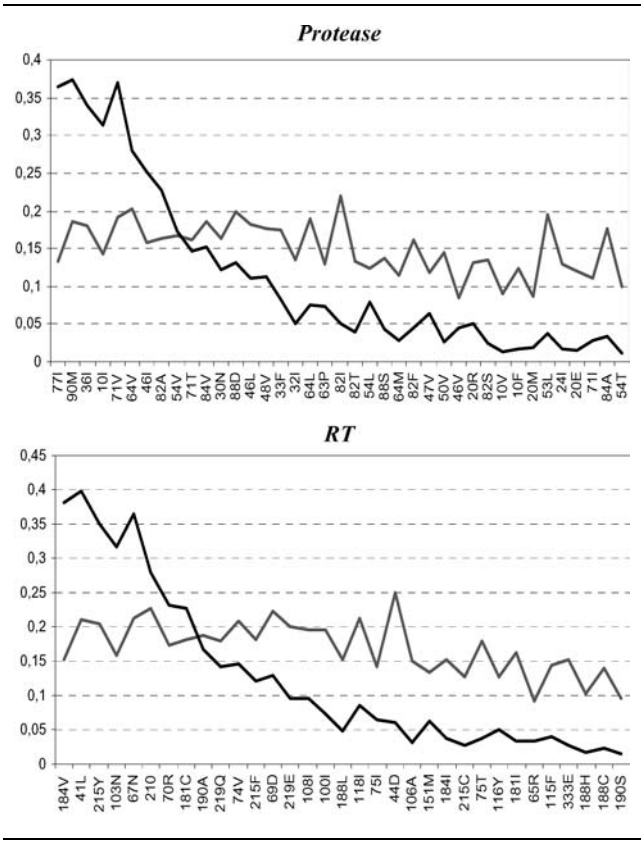


*Fig. 7. Orthogonal basic functions of expansion (4) for transient probability matrix.*

The error of the simulation increases proportionally to absolute value of the occurrences. Nevertheless, for some cases the error of the simulation is larger then the boundary of the confidence interval. This systematic error may be explained by possible variations in matrix $D$ for groups of the "old" and "new" patients.

*Application to forecast of HIV-1 evolution in time*. The evolution of total world populations of HIV-1 and the associated changing of the related drug resistance levels should be taken into account. The stochastic models, used to describe the HIV-1 genotype ensemble in terms of parameters and shown in the Figure 5, can be used for the analysis of its temporal variability during the observation period (VIII.1998–V.2001). The temporal variability of the data may be considered in terms of the samples of the seasons (3-months periods). The volumes of seasonal samples are from 1500 till 4500 genotypes; that is enough for obtaining the stable estimations. Only the hypothesis of linear trends is considered: $\xi(t) = at + b + \delta(t)$, where $a$ is the most interesting parameter—value of the trend, $b$ is the shift parameter, and $\delta$ is the white noise. In the Table 3 the integral parameters of trends of the various parameters of the HIV-1 population (mean value of the parameter, value of the trend, determination coefficient $R^2$ and the sample value of $F$-criterion) are shown.

*Trends of single mutations occurrence* $P_\xi$. The database allowed us to investigate trends in codon frequency in the period of 1998 till 2001. Results for Protease and RT are shown in Table 3. The majority of the mutations in the genotype have a negative trend, only 77I in Protease has significant positive trend.

*Trends of bi-modal distribution for number of mutations in genotypes* $\rho_n$. For the decreasing of the data dimensionality and the statistical discrimination of two groups in the dataset

*Table 3. Trend analysis of the parameters of the HIV-1 genotype population (F is compared with Fisher's test F(1,31,95%) = 4.14)*

| | Occurrence of mutations, % | | | | $p_g$, %, | Coefficients $\sqrt{\lambda_k}$, Equation (4) | | |
|---|---|---|---|---|---|---|---|---|
| Parameter | 77I | 90M | 10I | 71V | Equation (5) | $k=1$ | $k=2$ | $k=3$ |
| Protease part | | | | | | | | |
| Mean | 37.78 | 32.69 | 27.97 | 23.64 | 48 | 5.78 | 1.67 | 0.83 |
| $a$ (1/month) | 0.20 | −0.43 | −0.72 | 0.32 | 0.74 | 0.13 | 0.06 | 0.06 |
| $R_2$ | 0.68 | 0.91 | 0.61 | 0.82 | 0.67 | 0.80 | 0.73 | 0.54 |
| $F$ | 16.7 | 77.6 | 9.6 | 47.1 | 64.0 | 23.6 | 26.8 | 11.8 |
| RT part | | | | | | | | |
| | 41L | 215Y | 103N | 67N | | $k=1$ | $k=2$ | $k=3$ |
| Mean | 32.86 | 31.37 | 30.66 | 27.21 | 47 | 6.65 | 2.20 | 2.08 |
| $a$ (1/month) | −0.51 | −0.50 | −0.32 | −0.39 | 0.49 | 0.11 | 0.17 | 0.07 |
| $R^2$ | 0.88 | 0.93 | 0.88 | 0.84 | 0.75 | 0.68 | 0.78 | 0.71 |
| $F$ | 57.4 | 98.7 | 59.8 | 41.8 | 94.3 | 21.4 | 36.1 | 25.3 |

we consider the model of the mixture of two Bernoulli distributions:

$$\rho_n = p_g C_{m_1}^k q_1^k (1-q_1)^{m_1-k} + (1-p_g) C_{m_2}^k q_2^k (1-q_2)^{m_2-k} \quad (5)$$

where $p_g$ is an input of the first group of mutations (and $p_g$ is an input of the second group, $m_1, m_2$-are maximal numbers of mutations in groups and $q_1, q_2$-are probabilities to find each one (arbitrary) mutation in the groups. The use of Bernoulli distribution logic (based on the repetition of the independent events) is more close to the description of the mutation process, then the Poisson distribution, generally applying to description of rare events. Temporal variability of the parameters $(p, q_1, q_2, m_1, m_2)_t$ of the $\rho_n$ approximation by Equation (5) are shown in Table 3. In both cases only the parameter $p_g$ (weight of the left part for group of $m_1$ mutations) has a clear significant positive trend. For protease value $p_g$ increased from 39% in Summer, 1998 to 62% in Summer 2001 (with average increment $a = 0.74\%$ per month). Taking into account trends for separate mutations we observed a "degradation" of genotypes: the number of patients with simple genotypes (small number of mutations) is growing but a number of patients with big count of mutations is decreased.

*Trends of transient probabilities D.* The analysis of the trends of parameters for distribution (1) shows that the input of the first group of mutations with low number $n$ is increased. Hence, it may be a consequence of the temporal variations of the interdependencies between different mutations, governed by the developing of the drug therapy. For the analysis of these hypothesis, let us consider the trends for the matrix $D$, Equation (2). Taking into account the expansions (3, 4), we may reduce the complicate problem for joint trend analysis for components $D_{ij}$ to the procedure of trend analysis for independent time series – components of expansions (4). From the Table 3 it can be seen, that all the components have a clear positive trends. Taking into account the shape of first bases functions, see Figure 7, it is clear, that generally the joint probabilities $D_{ij}$ of the mutations is increased also; moreover, the power of increasing corresponds to the total occurrences of the mutation in the ensemble.

The discrimination of the groups of "old" and "new" patients in terms of bi-modal distribution (5) allow to forecast the growth of the total number of HIV-infected people in time:

$$N(t) = N_{\underset{\text{patients}}{\text{new}}}(t) + N_{\underset{\text{patients}}{\text{old}}}(\varepsilon t), \varepsilon \ll 1. \quad (6)$$

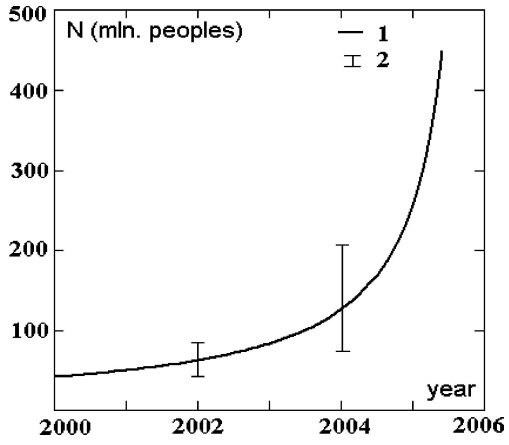


*Fig. 8. Qualitative forecast of HIV-1 population grows. 1 – mean value (7), 2 – 90% confidence interval.*

Here $\varepsilon$ – is the slow time parameter, which shows the rapid increasing of the new patients group in comparison with the old patients. The part of "new" patients of the sample is $p_g$ (old patients$-(1-p_g)$) from (5). Hence, the growth curve is:

$$N(t) = N_{\underset{\text{patients}}{\text{old}}}(0)\left[1 + \frac{p_g(t)}{1-p_g(t)}\right], \quad (7)$$

where $p_g(t) = p_0 + a_g t$-is the linear trend with the parameters from Table 3, and $N_{\underset{\text{patients}}{\text{old}}}(0)$ is the initial value of "old" (treated) patients on the beginning of the forecast.

In Figure 8 the "crucial" forecast of the HIV-1 population growth are shown. It is based on the fact that altogether 42 million people worldwide have been infected with HIV at the beginning of XXI century, and 12 million have died over the last 20 years. Moreover, not taken into account is the arising of new drugs and different prophylactic and social preventive activities for restriction of HIV-1 infection. Really, this result is qualitative only; for quantitative conclusions the more sophisticated research should be done.

## 3. RESULTS

### *3.1. Data presentation: Roaming PDA access*

#### *3.1.1. User Scenario*

RetroGram™ (www.retrogram.com) is a unique HIV-genotype expert based interpretation software program,

which weighs the effect of specified genotype changes on clinical drug activity. It accepts a list of substitutions to the protease and reverse transcriptase genes with respect to the NL4-3 reference strain. This is accomplished by running a "simulation", which applies some hundred rules relating substitutions on the HIV genome to knowledge of effects on drug response. The latter comes from over hundreds of references from the clinical literature. The rules are checked against the reported substitutions, and each drug is evaluated for its suitability. In a later stage we added Web-access where a Web interface is used to submit the input and take out the output. We want to make the simulations wireless-accessible. Developing a wireless Internet version from scratch will not be cost-efficient and causes maintainability problems. For example, the rules mentioned above are often changed and these changes have to be reflected in both versions. Furthermore, for privacy and security reasons the developer is not granted access to the source code of the "simulation". Thus, it is much more convenient to have wireless access to the Web-based interface. In this case the "simulation" take places in a unique server and privacy and security are guaranteed. A typical user scenario is described below and the associated graphical representation of the Retrogram Web access is given in Figure 9.

After the user has successfully logged in, the *Patient Detail* page is displayed (Figure 10). The form, taking place in this page is used to enter the personal data of the patient. Two fields are required in the form, *Patient ID* and *Data of Sample.*

According to the information taken from the laboratory the user enters the laboratory test results (i.e. Protease or RT substitutions) for the patient in the *Laboratory Information* page. Next a script invoked on the server does the following:

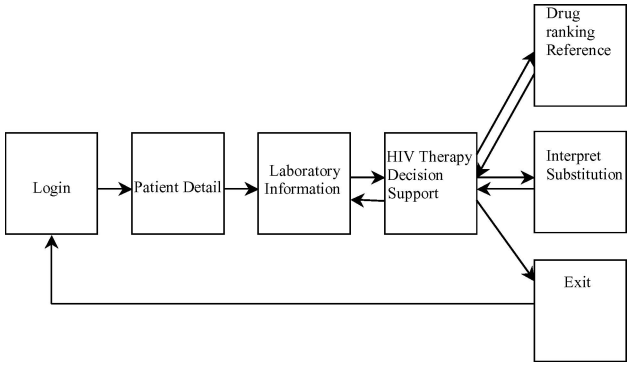


*Fig. 9. Web-based Retrogram use case sequence.*

Script 1: Server validation script

Validate inputs:
Validate Protease or RT substitutions if they conform to certain rules.
A single substitution should be represented by an integer (for position in the gene) and a letter (for the amino acid). The position in the gene is in the rage from 1 to 99 for Protease position and from 1 to 599 for RT position. The amino acid code is one of the following codes: A C D E F G H I K L M N P Q R S T U V W Y.
Submit the inputs to the "simulation" program and take back the drugs ranking result.
Show the Drugs ranking result in the *'HIV Therapy decision support'* screen:
After applying certain rules on the laboratory test result return to the final drugs ranking or drug's level of suitability indication as follows:
A (green): This drug can be used
B (yellow): Consider use if no class A drug available
C (amber): Consider use if no class A or B drug available
D (red): Consider use if no class A, B or C drug available
U (grey): Unranked, insufficient data available

In the '*HIV Therapy decision support*' screen, clicking on any drug name in the ranking lists will display a list of available references from the scientific literature supporting the particular ranking for that drug. In the 'HIV Therapy decision support' screen, clicking on the 'Interpret substitution' button will show classification of the patient's substitutions into *relevant, natural or additional.*

### *3.1.2. Roaming, wireless access*

In the designing phase of wireless versions of the application the constraints of the mobile devices should be considered. At the same time we have tried to maintain the same level of usability and readability as in the original Web version. This is accomplished by maintaining the same structure as that in the Web but with some modifications. For example, the Patient detail form has many fields and putting them in one screen would cause problems in the usability of the program (it's supposed that the mobile device has a resolution comparable to a normal PDA, i.e., something around 160 × 160 pixels). Thus we use three screens for

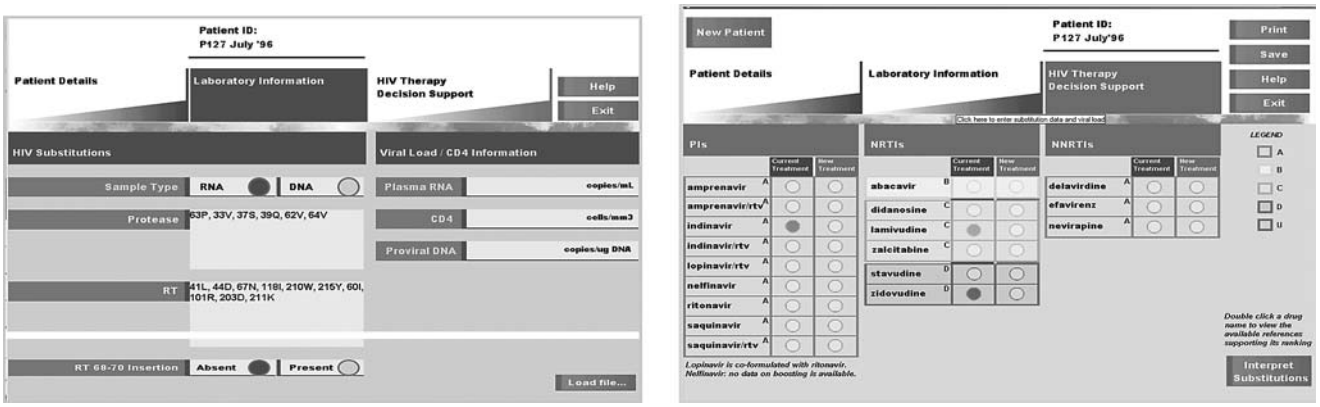


*Fig. 10. Web Retrogram: user enters patient substitutions (left), drug ranking results (right).*

Patient Detail data. The Patient Detail Web page has 2 required fields. We put them in the first screen after the 'login' screen. In this way, if the user is not interested in entering optional data, she can directly go to the Laboratory Information.

*Proxy method Implementation*. A Proxy method is implemented for accessing the web-based software from mobile devices. The Proxy server takes places between the remote server (the Retrogram server) and the mobile device. A *mininavigator* script developed in the Proxy is responsible for the following:

- Take the patient data from the mobile user (i.e. patient detail, laboratory information)
- Create an HTTP communication with the remote server,
- Submit data to the remote server. These data are basically the input for the Retrogram 'simulation'.
- Take the result from the remote server (HTML code generated from retrogram.asp script),
- Parse HTML code and retrieve only relevant information (i.e. drug ranking, error messages, drug references etc.). It uses this relevant information to build wireless pages (i.e. WML page in case of WAP or Web-clipping page).
- Send the wireless pages to the mobile device.

The Proxy is implemented using PHP: Hypertext Preprocessor as a server-site scripting language [9–11] running on the Apache Web server [12].

Two versions are developed using the Proxy method: WAP version and web clipping. If a user wants to enter the 'patient details' fields, he has to move from one screen to the other and come back again. The fields already filled in the previous screens should not be lost. Thus maintaining the client's state is necessary. In the WAP case we simply use cookies but in web clipping cookies are supported only in PALM OS 4.0 version or higher. For this reason the "hidden field" method is used this is another method used for maintaining state in the Internet. The following figures are the user interfaces that have been captured. They track the user's path through the running of the application, as shown in Figures 11(a) and 11(b), where the user enters the patient's details and accesses ranking results.

*J2ME Implementation*. The same user interface is applied in the J2ME implementation. There are two main differences between the J2ME implementation and the Proxy one:

1. J2ME enables the device to communicate directly to the Retrogram server without an intermediate Proxy
2. In J2ME the client's interface is contained within the device. In the Proxy method, every time the interface should be changed, the Proxy is responsible for generating a new page.

The following illustrates the necessary steps one should take in order to fetch an HTML page generated from a script in the remote host. Specifically this is an example illustrating how the user can login to a script in the Retrogram server and extract the cookie from the header response:

1. Open an HTTP connection
2. Open an input stream
3. Make an HTTP POST request
4. Extract the cookie from the header response
5. Close the connection

In the J2ME implementation of Retrogram the entire client's interface takes places in the device. The connection

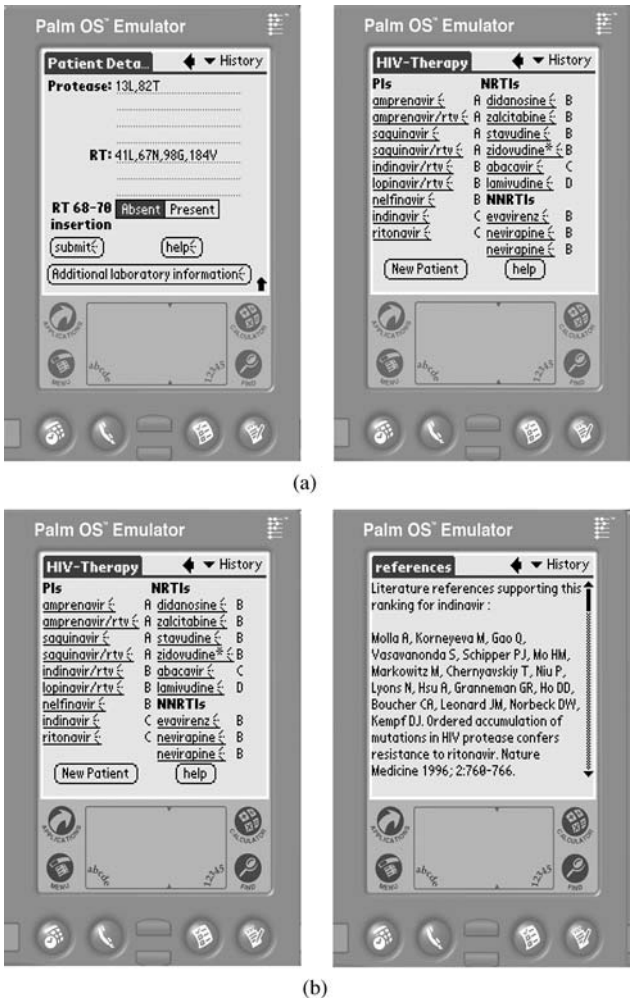


*Fig. 11. (a) User corrects the input and submit again (left), drug ranking results (right). (b) Users clicks to the drug 'indinavir' (left), references supporting this ranking (right).*

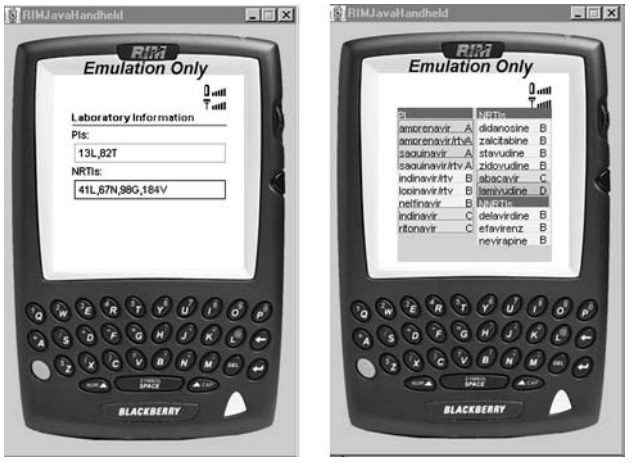


*Fig. 12. J2ME method; user enters patient's substitutions (left), drug ranking results (right).*

to the server is established in the following cases: user login, with connection with the server is necessary in order to validate the user and/or password. The user submits the username and password, and the application judges them for their correctness by scanning the HTML response from the Retrogram server. The user submits the patient's laboratory information data. The application should connect to the server in order to submit the data, take the result (HTML format) and extract the drugs ranking. Next the user looks for the references that suggest a certain drug ranking. The database with all the references exists in the Retrogram server, therefore the connection is necessary. The application submits to a Retrogram script the cookie and the name of the drug. The drug references are given back from the server in HTML format. The application should clean up the HTML tags and show the references as plain text. Finally the user looks for classification of the patient's substitutions. This classification is part of the Retrogram 'simulation' and thus the connection to the server is still necessary. In Figure 12 we illustrate the process of taking the drugs ranking using the J2ME method.

Currently we have the J2ME version in use for different users to study the usability and extendibility. More details on the implementation can be found in reference [13].

### *3.2. Virtual laboratory infrastructure*

#### *3.2.1. A virtual organization for retrogram-centered workflow*

Grid technology is a major cornerstone of today's computational science and engineering, with its basic unit of Grid organization called the Virtual Organization (VO). A VO is a set of Grid entities, such as individuals, applications, services or resources, which are related to each other by some level of trust. In the most basic example, service providers would only allow access to the members of the same VO. We are currently building a distributed Grid-based overall decision support infrastructure to support the Retrogram-centered workflow shown in Figure 13.

This VO will offer a Grid virtual laboratory that will assist users in the interpretation the genotype of a patient by using rules developed by experts on the basis of the literature, taking into account the relationship between the genotype and phenotype. The workflow is based on highly distributed available data from clinical studies and on the

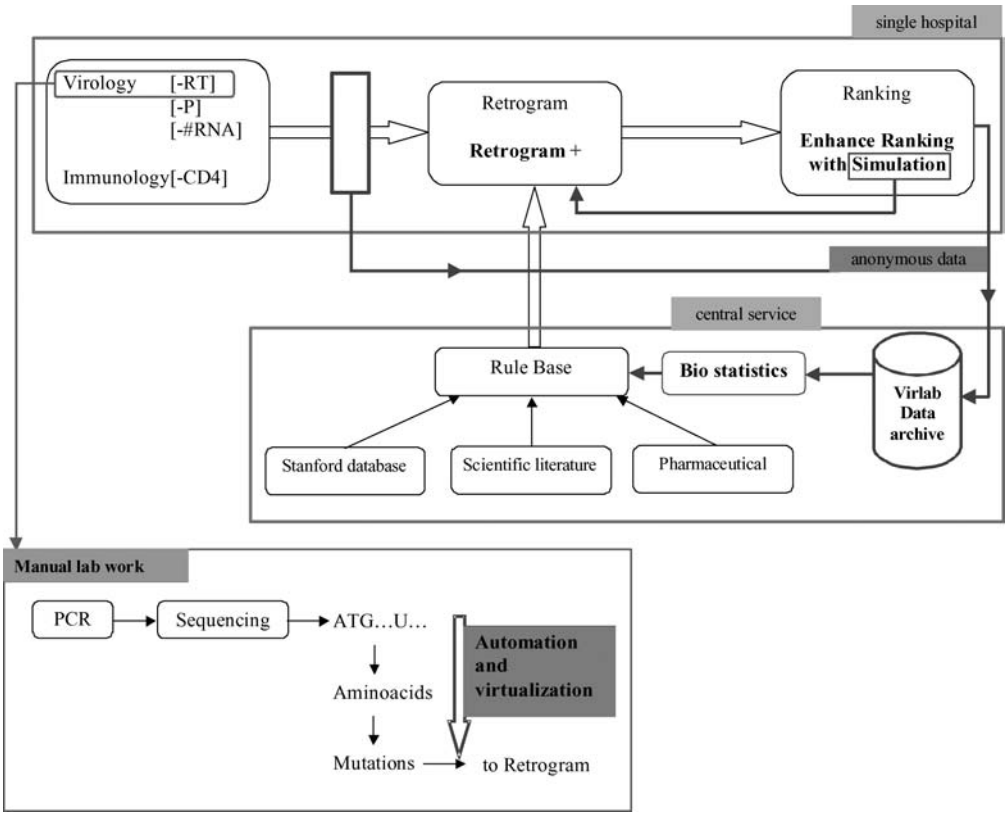


*Fig. 13. A Retrogram-centered workflow.*

relationship between the presence of genotype and the clinical outcome. In order to cover the fast temporal and spatial scales required to infer information from a molecular (genomic) level up to patient medical data multi-scale methods are applied, where simulation, statistical analysis and data mining are combined and used to enhance the rule-based decision. In this scenario, information sources are widely distributed, and the data processing requirements are highly variable, both in the type of resources required and the processing demands. Experiment design, integration of information from various sources, as well as transparent scheduling and execution of experiments will be supported by this support system based on distributed Grid middleware. The DAS2 testbed (Netherlands) will initially provide the additional computational power for our compute intensive jobs. We will reuse Grid middleware from successful European projects such as CrossGrid (www.crossGrid.org) and VL-e (www.vl-e.nl) to provide basic Grid services of data management, resource management, and information services on top of Globus. For transparent use of this infrastructure we will build a presentation layer that will provide a user-friendly interface to both medical doctors and scientists.

## 4. DISCUSSION

### *4.1. Conclusions and future work*

In this paper we discussed an integrative approach to biomedicine at large and to infectious diseases in particular. We showed how in the understanding of processes 'from molecule to man' Grid technology can play a crucial role. In order to cover the fast time and spatial scales required to infer information from a molecular (genomic) level up to patient medical data, we need to apply multi-scale methods where simulation, statistical analysis, data-mining is combined in an efficient way. Moreover the required integrative approach asks for distributed data collection (e.g. HIV mutation databases, patient data, literature reports etc.) and a virtual organization (physicians, hospital administration, computational resources etc.). Also the access to and use of large-scale computation (both high performance as well as distributed) is essential since many of the computations involved require near real-time response and are to complex to run on a personal computer or PDA. Finally data presentation is crucial in order to lower the barrier of actual

usage by the physicians, here the Grid technology (server-client approach) can play an important role.

Although many of the aspects discussed in this paper have proven to work in concept, the complete integration of the systems and the evaluation of day-to-day use is still under development [17]. In addition each of the underlying methods (Rule-based, statistical and CA based models) remain topics of further studies. We will set up a use-base with the system described running under various European Grid testbeds. The first testbed we will use is the so-called DAS2, and eventually the CrossGrid testbed, which supports specific features for interactive computation, an essential ingredient for a medical decision support system.

The authors gratefully acknowledge Fan Chen and Ferdinand Alimadhi for assistance in implementing the CA models and the roaming PDA access. The Dutch Virtual Laboratory on e-science project supported parts of the research presented here: http://www.VL-e.nl.

## GLOSSARY

Grid: Distributed architecture for solving computational problems by making use of the resources from the members of a virtual organization, treating them as a virtual cluster.

CA: Cellular Automata, a discrete model studied in computational theory and mathematics, which consists of regular grid of cells, each in one of a finite number of states.

Decision Support System: Computer-based system that helps in the process of decision-making.

Web Interface: User interfaces for information available via the web.

Proxy: Computer service which allows clients to make indirect network connections to other services.

HTTP: Hyper Text Transfer Protocol, a request/response protocol for transferring information on the Web.

HTML: Hyper Text Markup Language, a markup language designed for the creation of web pages.

WML: Wireless Markup Language, a markup language used in mobile phones.

J2ME: Java 2 Platform Micro Edition, a collection of Java interfaces for embedded consumer appliances such as cellular phones.

DAS2: Distributed ASCI Super Computer 2, a wide-area distributed computer connecting 5 Dutch Universities.

## REFERENCES

1. Zhao Z, Belleman RG, van Albada GD, Sloot PMA. AG-IVE: An Agent-Based Solution to Constructing Interactive Simulation Systems, in Series Lecture Notes in Computer Science, April 2002; 2329: 693–703.
2. Durant J, Clevenbergh P, Halfon P, Delguidice P, Porsin S, Simonet P, Montagne N, Dohin E, Schapiro JM, Boucher C, Dellamonica P. Improving HIV therapy with drug resistance genotyping: The Viradapt Study. Lancet 1999; 353: 2195–2199.
3. Sevin AD, DeGruttola, Nijhuis M, Schapiro JM, Foulkes AS, Para MF, Boucher CAB. Methods for Investigation of the Relationship between Drug-Susceptibility Phenotype and Human Immunodeficiency Virus Type 1 Genotype with Applications to AIDS Clinical Trials Groupw 333. The Journal of Infectious Diseases 2000; 182: 59–67.
4. The Genotype database is obtained from a large service testing laboratory from the US. It contains the resistance profiles of the Protease and Reverse Transcriptase genes of the HIV-1 virus obtained from plasma samples of HIV-1 infected patients. No clinical background information on medication or drug history is available.
5. Mathematical Methods for DNA Sequences. In. Waterman MS, eds. CRC Press Inc., Boca Raton, Florida, 1999.
6. Kiryukhin I, Saskov K, Boukhanovsky AV, Keulen W, Boucher, CA, Sloot PMA. Stochastic modeling of temporal variability of HIV-1 population. In: Sloot PMA, Abrahamson D, Bogdanov AV, Dongarra JJ, Zomaya AY, Gorbachev YE, eds. Computational Science – ICCS 2003, Melbourne, Australia and St. Petersburg, Russia, Proceedings Part I, in series Lecture Notes in Computer Science, vol. 2657, pp. 125–135. Springer Verlag, June 2003. ISBN 3-540-40194-6.
7. Zorzenon dos Santos RM, Coutinho S. Dynamics of HIV infection: A cellular automata approach. Phys Rev Lett 2001; 87(16): 168102–1–4.
8. Sloot PMA, Chen F, Boucher CA. Cellular automata model of drug therapy for HIV infection. In: Bandini S, Chopard B, Tomassini M, eds. 5th International Conference on Cellular Automata for Research and Industry, ACRI 2002, Geneva, Switzerland, October 9–11, 2002. Proceedings, in series Lecture Notes in Computer Science, vol. 2493, pp. 282–293. October 2002.
9. PHP: Hypertext Preprocessor: http://www.php.net.
10. The resource for PHP developers: http://www.phpbuilder.com
11. Zend Technologies – PHP tools for the development, protection and scalability of PHP applications – PHP for Linux, Unix and Apache, Encoder, Accelerator Studio, Debugger: http://www.zend.com.
12. The Apache Software Foundation: http://www.apache.org.
13. Alimadhi F. Mobile Internet: Wireless access to Web-based interfaces of legacy simulations, MSc thesis, University of Amsterdam, The Netherlands, September 2002: http://www.science.uva.nl/research/pscs/papers/master.html.
14. Cross-Grid: Grid technology of Interactive Distributed Computation: http://www.eu-crossGrid.org/.

15. Little SJ, Holte S, Routy JP, Daar ES, Markowitz M, Collier AC, Koup RA, Mellors JW, Connick E, Conway B, Kilby M, Wang L, Whitcomb JM, Hellmann NS, Richman DD. Antiretroviral-drug resistance among patients recently infected with HIV. N Engl J Med 2002; 8;347(6): 385–394.
16. Karlin S. A First Course in Stochastic Processes. Academic Press. NY-London, 1968.
17. Sloot PMA, Boucher CA, Kiryukhin I, Saskov K, Boukhanovsky AV. A grid-based problem-solving environment for biomedicine. In: Nørager S, ed. Proceedings of the First European HealthGrid Conference, January, 16th-17th, 2003, pp. 300–323. Commission of the European Communities, Information Society Directorate-General, Brussels, Belgium, 2003.