Journal of
Health & Medical Informatics

GHEMTIO et al. J Health Med Informat 2011, 2:1
DOI: 10.4172/2157-7420.1000104

Research Article Open Access

HIV-PDI: A Protein-Drug Interaction Resource for Structural Analyses of HIV Drug Resistance: 1. Concepts and Associated Database

GHEMTIO Leo[1]*, SMAÏL-TABBONE Malika[1], DJIKENG Appolinaire[2], DEVIGNES Marie-Dominique[1], KEMINSE Lionel[3,4], KELBERT Patricia[1], FOKAM Joseph[3], MAIGRET Bernard*[1,4] and OUWE-MISSI-OUKEM-BOYER Odile[3]*

[1]*Nancy Université, LORIA, Orpailleur Team, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France*
[2]*Biosciences eastern and central Africa (BecA) Hub at the International Livestock Research Institute (ILRI). P.O. Box Nairobi, Kenya*
[3]*Centre International de Référence Chantal Biya (CIRCB) pour la Recherche sur la Prévention et la Prise en charge du VIH/SIDA, BP 3077, Yaoundé, Cameroun*
[4]*Harmonic Pharma, Espace Transfert, 615 rue du Jardin Botanique 54600 Villers-lès-Nancy, France*

Abstract

Overcoming the problem of resistance to antiretroviral drugs (ARVs) in HIV-infected patients is a major issue in AIDS research today. Advances in genome sequencing have facilitated the identification of a growing number of individual genotypes. Hence, it is now possible to understand HIV drug resistance at the molecular level by considering the three-dimensional (3D) structural interactions between ARVs and the mutated viral proteins of patients. Therefore, identification of the critical interactions lost further to one or several HIV mutations, and consequently the modifications of other molecular factors, could be indicators to propose appropriate ARVs escaping the resistance. This paper introduces the HIV-PDI (Protein-Drug Interactions) resource designed to be a decision making tool to propose alternative ARVs against a particular mutated viral protein, and thus to provide a personalized antiretroviral treatment. The HIV-PDI was conceived to serve as an integrated resource for studying HIV drug resistance at the structural level of the protein-drug interaction, with a special emphasis on the active site of the HIV drug target. As a first step, we focus on the well documented protease and related drugs. The HIV-PDI includes clinical information on patients, resistance to given ARVs treatments, HIV proteins structures and mutations, HIV protein/ARV drugs and their 3D interactions. The HIV-PDI may be queried using multiple combinations of fields including protein, drug and treatment conditions and coupled to visualization/analysis tools of 3D Protein-Drug interactions. The HIV-PDI resource can be used in order to help understand the appearance of resistance and to promote further novel drug and treatment developments based on analyses of 3D pattern of protein-drug interactions. A web-based version of HIV-PDI is available at http://hiv-pdi.loria.fr.

Introduction

Despite the availability of several antiretroviral drug molecules (ARVs), the rapid emergence of resistance to ARVs by the HIV virus has become a major obstacle to effective control of HIV infection and the treatment of infected patients [1,2]. Several retrospective and prospective studies have shown that the presence of a resistant strain of HIV before starting ARV treatment can often affect the outcome [3,4]. Thus, resistance to ARVs is now considered a top priority for HIV/AIDS research and actions [5,6].

Many documented examples of resistance are directly related to mutations affecting the genetic make-up of the virus, and especially to those which relate to proteins targeted by ARVs. It is now well established that mutation-induced resistance involves structural modifications of viral proteins which reduce the ability of drug molecules to bind near the active site of the viral target but which do not otherwise impede the viral function. Indeed, X-ray crystallography studies have clearly shown that certain mutations can change the geometry of the viral protein active site, thus reducing the binding affinities of ARVs [7-10]. Consequently, the structural basis of ARV resistance can be understood by careful inspection of the interactions between the mutated target (i.e. the HIV variant) and its ligand (i.e. the drug) at the molecular level [11-13]. Therefore, to carry out molecular analyses of HIV disease mechanisms, it is necessary to augment current HIV-related databases with detailed three-dimensional (3D) protein and ligand structural information. Integrating and exploiting this knowledge could help to optimise the use of existing ARVs in first and second line treatment regimes [14,15].

Several HIV-related databases have been described for the collection and storage of information on known mutations, drug resistance, clinical data, and additional data on HIV proteins including correlated information for AIDS treatment [16-23] (see Table, Supplementary information 1, which shows the examples of commonly used HIV-related databases). For example, the Stanford HIV drug resistance database (HIVdb) [21] is a very well documented public database designed to represent, store, and analyze the variations in protein sequences which are responsible for HIV drug resistance. The HIVdb contains several hundred HIV-1 mutations in several flat files along with resistance data, results on drugs susceptibility, and clinical patient data. It also provides several tools for analyzing mutated viral sequences in order to help understand drug resistance [21,24]. Moreover, it has been used recently to support some very interesting studies on the relationship between HIV protein mutations and disease treatments in a given population [7,25,26].

Despite the availability of these resources linked to existing HIV

***Corresponding authors:** GHEMTIO Leo, Nancy Université, LORIA, Orpailleur Team, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France, Tel: +(33) 3 54 95 85 92; Fax: +(33) 3 83 27 83 19; E-mail: leo.ghemtio@loria.fr

MAIGRET Bernard, Nancy Université, LORIA, Orpailleur Team, Campus scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France, E-mail: bernard.maigret@loria.fr

OUWE-MISSI-OUKEM-BOYER Odile, Centre International de Référence Chantal Biya (CIRCB) pour la Recherche sur la Prévention et la Prise en charge du VIH/SIDA, BP 3077, Yaoundé, Cameroun, E-mail: oukem@gmail.com

Received April 11, 2011; **Accepted** July 10, 2011; **Published** July 12, 2011

Citation: GHEMTIO L, SMAÏL-TABBONE M, DJIKENG A, DEVIGNES MD, KEMINSE L, et al. (2011) HIV-PDI: A Protein-Drug Interaction Resource for Structural Analyses of HIV Drug Resistance: 1. Concepts and Associated Database. J Health Med Informat 2:104. doi:10.4172/2157-7420.1000104

databases [27], there is a need for bioinformatics tools which can relate information concerning viral mutations with 3D structural features in the corresponding proteins. Those mutations may affect the pattern of 3D interactions which characterize how a particular drug might bind to its cognate HIV protein. Therefore, identifying changes in such patterns could provide new insights into the molecular basis of HIV drug resistance. Our basic hypotheses are that (i) for a given patient, observed resistance to a treatment may rely on a loss of affinity of the delivered ARVs targeting HIV proteins, (ii) this loss of affinity is a consequence of structural modifications at the drug binding site due to mutations, and (iii) recovering lost affinity can be achieved, for example, by restoring a similar drug interaction pattern by identifying new interaction points in another class of ARVs to compensate for lost ones. Therefore, by using molecular modeling and protein-ligand docking, we presume that it will be possible to identify the main structural and interaction features leading to a loss of affinity and hence propose a possible route to overcome drug resistance. Therefore, in addition to being able to store and manipulate clinical and biological data, we explicitly designed our HIV-PDI database to include protein and ligand structural information, physico-chemical descriptors, and 3D protein-ligand interaction patterns. Nonetheless, for the initial version of our database, we decided to focus on the well-characterized and documented HIV protease and its associated inhibitors [28-31] with the expectation that the lessons learned would be directly applicable to other HIV protein targets.

In this paper, we describe the design and implementation of our HIV-Protein Drug Interaction (HIV-PDI) database which is the centerpiece of a bioinformatics resource designed to integrate epidemiological, clinical, pharmacological, biological, chemical, and structural data collected from HIV infected patients and which allows close interaction with several protein structure analysis and visualizing tools for decision support.

Methods

Database requirements analysis

After extensive discussions with people on the ground, it became clear that in order to integrate current biological knowledge of viral proteins with clinical data, it would be desirable to create a repository which would provide unified access to a very diverse range of information drawn from equally diverse sources of data. For example, the repository would need to be able to store details on HIV protein sequences and their 3D structures, the chemical structures of known drug molecules and their physico-chemical properties, information about 3D protein-drug interactions, patient information including data on CD4 counts, viremia, drug treatment regimes, and treatment outcomes, for example. Furthermore, this data would need to be extracted from various textual sources such as scientific articles and patents, existing HIV databases (Supplementary information 1), as well as other protein sequence and structural databases such as the Protein Data Bank (PDB) [32], Swiss-Prot [33], GenBank [34], and various other data sources at the European Molecular Biology Laboratory (EMBL) [35], for example.

Our HIV-PDI database aims to span experimental, clinical, theoretical and computational domains related to HIV resistance and to provide access to all of these diverse resources via a user friendly graphical user interface (GUI) in an integrated manner which avoids information redundancy and facilitates data integrity. Furthermore, all relevant scientific publications and descriptions of the methods used to produce the data are also stored in our database. Hence, in case of inconsistent or contradictory data, the user has all of the necessary information to make his own interpretation. As illustrated (Figure 1), the HIV-PDI database can therefore be considered as a decision support system based on extensive collections of heterogeneous HIV-related data.

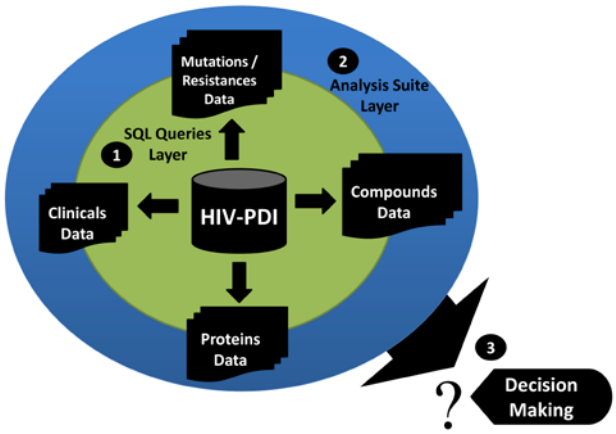

Figure 1: Overview of the HIV-PDI decision support resource.

The HIV-PDI database can be queried easily on multiple fields to access data about ligands (e.g. ligand conformations or chemical groups), proteins (e.g. different protein conformation of protein structures), 3D protein-ligand interactions, as well as patients' clinical data. For example, queries have been constructed which can search for examples of ligands which have a special chemical group, data on treatments of one or several patients, information about sequence mutations found in one or several patients, or data on ARV drug resistance observed in one or several patients. More complicated queries can also be formulated which can, for example, select ligands and proteins that are in complex and which have a given type of chemical interaction, provide the details of patient treatments which provoke mutations relating to drug resistance, identify treatment changes which have overcome previous cases of resistance, and select and visualize in 3D protein-drug interactions associated with mutations.

We expect that the clinical data, mutation and resistance information stored in the HIV-PDI database will be a very useful resource for finding relationships between patient genotype and drug resistance profiles using data mining techniques. Furthermore, the protein structure information and molecular descriptors [36-39] of the drugs stored in the HIV-PDI database will be useful for developing statistical models of drug potency and resistance.

The HIV-PDI conceptual data model

The HIV-PDI database was designed using the entity-relational (ER) database model to take into account the complexity of the data to be entered and subsequently extracted [40-42]. The ER approach was used to facilitate the processes of data requirement specification and conceptual database modeling. The resulting ER data model was then translated into a set of normalized relational tables. This phase was performed automatically using the publicly available DB Designer [43] computer-aided software engineering (CASE) tools. Details of the various data types and their relationships necessary to achieve our goal were integrated in a single ER conceptual model, as illustrated (Figure 2). The main entities in this model are described in further detail below.

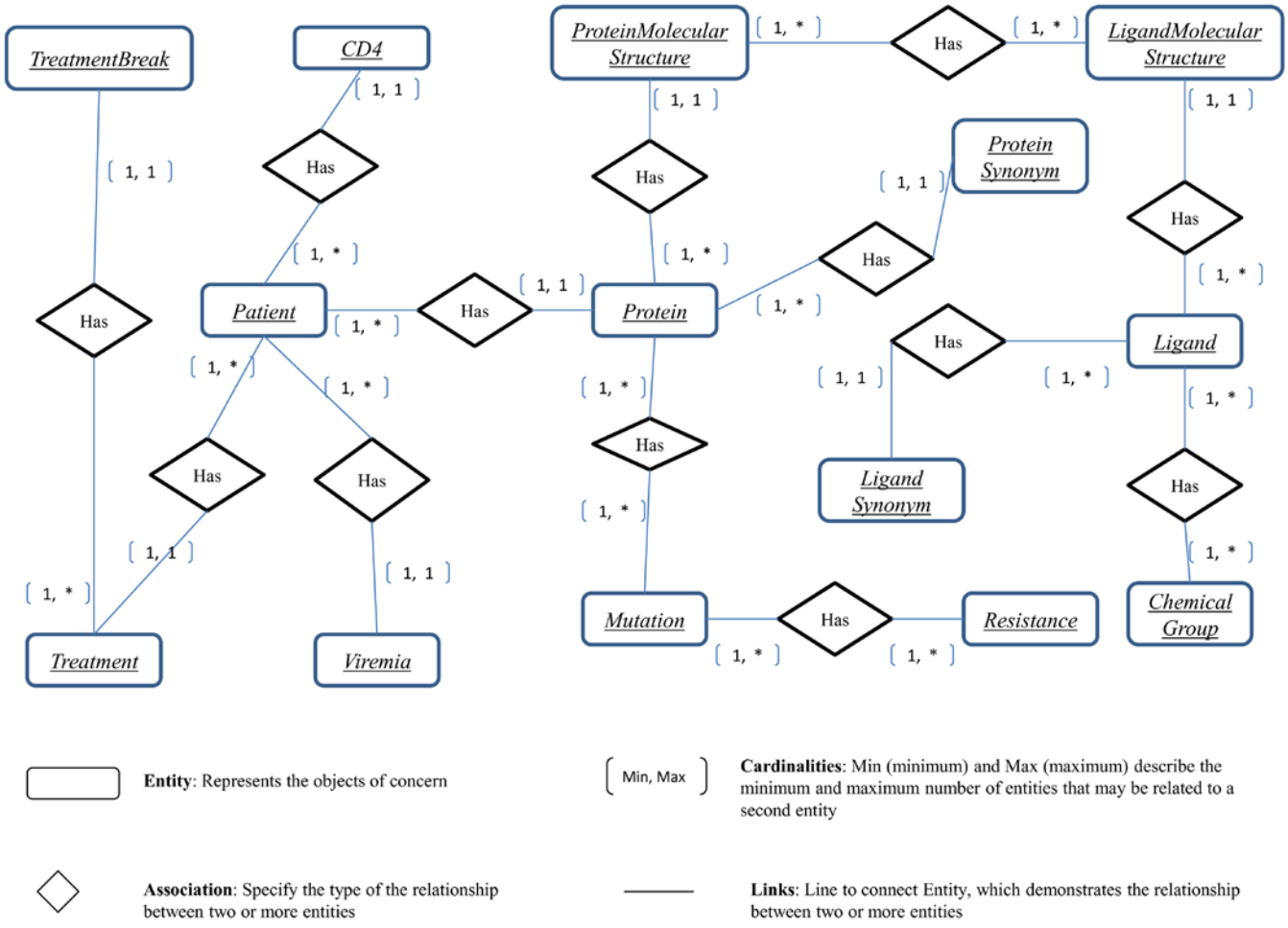

Figure 2: The ER diagram of the HIV-PDI database (entity attributes are not shown for clarity).

The entity *Protein* represents a wild type (WT) or a mutant protein sequence found in a patient. Each *Protein* instance is identified by a unique key (specific to this database) and is associated with at least one instance of the entity *Patient*. This association has a temporal descriptor to track the appearance or disappearance of a specific mutation in a patient according to the time-course of the patient's treatment. The entity *Protein* is also associated with one or more instances of the entity *Mutation* because the sequence of a *Protein* instance may contain one or several mutations, and conversely because a single mutation can be related to several *Protein* instances. The current version of the database only stores information on the HIV protease, although future versions will also store information on other HIV protein targets. The entity *Ligand* represents instances of the ligands (i.e. experimental compounds or clinical drug candidates). This entity contains information characterizing the chemical and pharmaceutical properties of each compound stored.

To represent structural information about proteins and drugs, the entity *Protein-MolecularStructure* is related to the entity *LigandMolecularStructure* in order to define the 3D structures of protein-ligand complexes as well as the molecular interactions occurring within the complex. Wherever possible, crystallographic data from the PDB is used, but if this is not available the protein structure is modeled by homology and protein-ligand interactions are modeled using the Glide docking program [44]. A given *Protein* may be associated with more than one *Protein Molecular Structure* because it can have more than one structural conformation. Similarly, a given *Ligand* may be associated with multiple *Ligand Molecular Structure* instances, each of which represents a single conformation of the corresponding *Ligand* instance. Other descriptors that characterize protein-ligand complexes are also stored. These descriptors include affinity data (e.g. EC-50, KI, and IC-50) and the root mean squared deviation (RMSD) measure of their shape complementarities calculated from spherical harmonic (SH) analyses of their surface areas and volumes.

The entity *Resistance* represents the fact that a ligand no longer interacts optimally with a mutated form of the viral protease. Each instance of *Resistance* is thus associated with an instance of the *Mutation* entity which is itself documented by validated information from other databases or provided by patient data collected from various institutions. The relevant mutations can be obtained through the association between the entity *Protein* and the entity *Mutation*. The entity *Patient* represents instances of individual patients, and is associated with other entities which represent temporal clinical information such as the *Treatment* (drugs and regimen), *Viremia* and *CD4* entities.

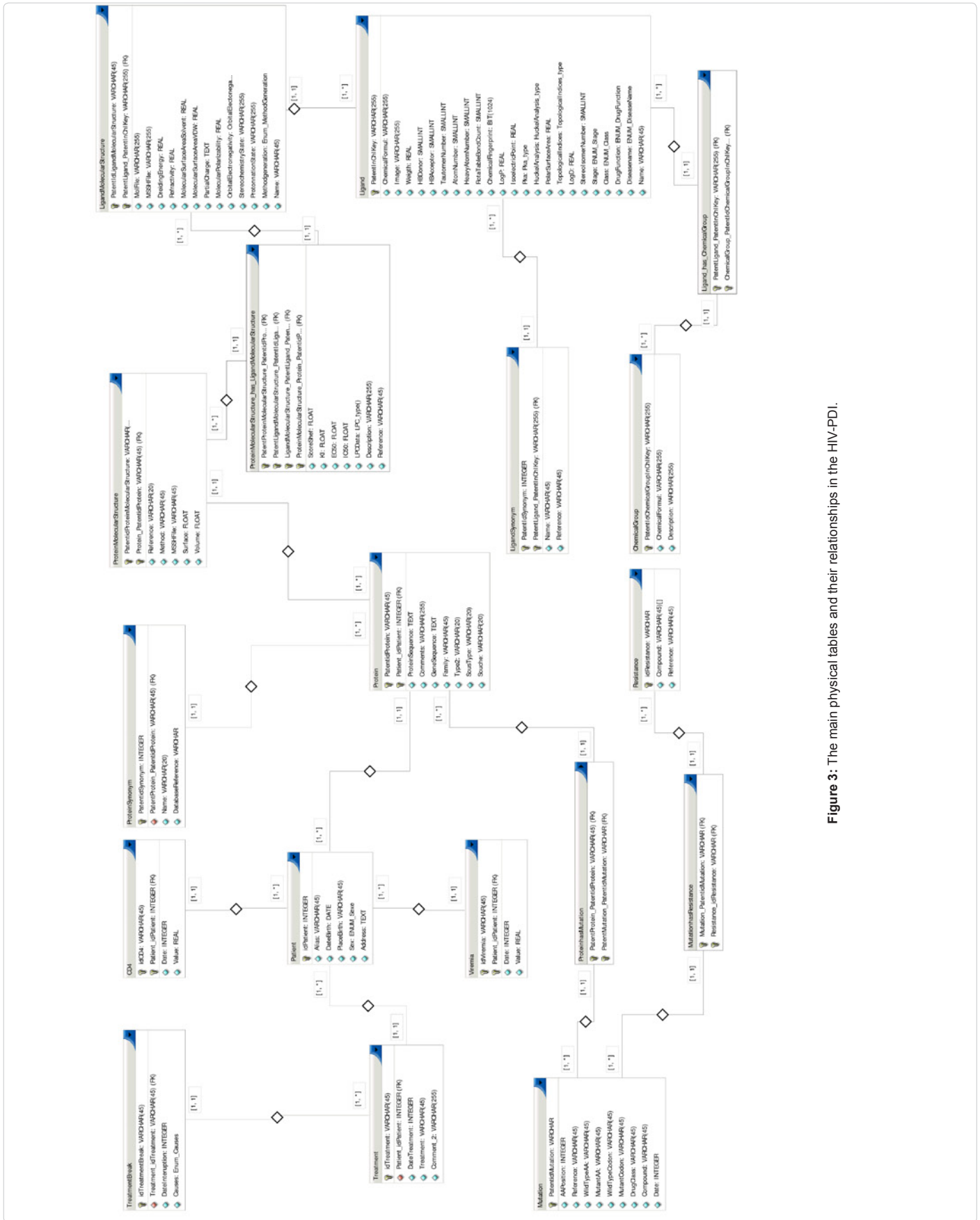

Figure 3: The main physical tables and their relationships in the HIV-PDI.

	Ligand	***Protein***	***Patient***	***Mutation***	***Viremia***	***Treatment***
Number of entries	5,962	2,493	2,029	7,791	28,447	13,629

Table 1: A summary of the total number of entries in the HIV-PDI database, version 1.0.

Inchi key	Hb donor	Tautomer number	Atom number	Logp	Polar surface area	Stage	Name	Logd
YMARZQAQMVYCKC-RKPZHZFKBN	3	2	70	2.38	139.57	Commercial	FDA-Agenerase	2.5
NZPXPXAGXYTROM-YKPJHVMUBZ	0	10	75	7.14	114	Commercial	FDA-Aptivus	6.2
QDRMCFDXPIEYGX-XSBXPMGLBM	1	162	133	0.72	195.25	Commercial	FDA-Atripla	3
LHCOVOKZWQYODM-AKUSELONBT	2	9	58	-2.47	108.3	Commercial	FDA-Combivir	-3.4
ZZHIRSCAVGFBKE-UHFFFAOYAA	0	3	23	1.02	45.76	R&D	LC-PR-1	1.1
CBVCZFGXHXORBI-IXYYIQBABV	4	9	92	2.58	118.03	Commercial	FDA-Crixivan	2.1
XQSPYNMVSIKCOC-KIJMNMGRBZ	2	3	26	-1.31	113.45	Commercial	FDA-Emtriva	0.1
JTEGQNOMFQHVDC-ZCIVMJCZBB	2	3	26	-1.27	113.45	Commercial	FDA-Epivir	-0.3
MBFKCGGQTYQTLR-ZUCZGPQLBE	3	6	46	0.62	101.88	Commercial	FDA-Epzicom	0.4
WREGKURFCTUGRC-KSLURUABBM	2	3	28	-1.23	88.15	Commercial	FDA-Hivid	-0.4
PYGWGZALEOIKDF-UHFFFAOYAM	2	6	43	5.24	120.64	Commercial	FDA-Intelence	5.2
ZEJGORPKXHJSER-FYWRMAATBT	2	44	53	0.14	152.85	Commercial	FDA-Isentress	-2.9

Table 2: Sample selection from the Ligand table.

Id protein	Id patient	Protein sequence	Strain	Family	Type
St-20245-0-I13V-L63A	20245	PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFIKVRQY-DQIAIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF	Unknown	Protease	HIV-1
St-19227-0-P39S-L63P-A71T	19227	PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLSGRWKPKMIGGIGGFIKVRQYDQI-PIEICGHKTIGTVLVGPTPVNIIGRNLLTQIGCTLNF	Unknown	Protease	HIV-1
Sw-O41798		PQITLWQRPLVTVKIGGQLIEALLDTGADDTVLEGINLPGKWKPKMIGGIGGFIKVRQYDQI-LIEIGGKKAIGTVLVGPTPINIIGRNMLTQIGCTLNF	M	Protease	HIV-1
Sw-O89290		PQITLWQRPLVTIRVGGQLKEALLDTGADDTVLEDVNLPGKWKPKMIGGIGGFIKVKQYDSI-LIEICGHRAIGTVLVGPTPVNIIGRNMLTQIGCTLHF	M	Protease	HIV-1

Table 3: Sample selection extracted from the protein table.

Id treatment	Id patient	Date treatment	Treatment	Comment
29440_AZT_-359	29440	-359	AZT	NOTHING
29440_DDI_-162	29440	-162	DDI	NOTHING
29440_AZT_0	29440	0	AZT	NOTHING
29440_DDI_0	29440	0	DDI	NOTHING
29440_3TC_0	29440	0	3TC	NOTHING
20245_AZT_0	20245	0	AZT	NOTHING
20245_3TC_0	20245	0	3TC	NOTHING
20246_D4T_0	20246	0	D4T	NOTHING
20246_D4T_23	20246	23	D4T	NOTHING
20246_3TC_23	20246	23	3TC	NOTHING
20247_D4T_0	20247	0	D4T	NOTHING
20247_3TC_0	20247	0	3TC	NOTHING
20248_D4T_0	20248	0	D4T	NOTHING
20249_AZT_0	20249	0	AZT	NOTHING
20249_3TC_0	20249	0	3TC	NOTHING
20250_AZT_0	20250	0	AZT	NOTHING
20250_3TC_0	20250	0	3TC	NOTHING
20251_AZT_0	20251	0	AZT	NOTHING
20251_3TC_0	20251	0	3TC	NOTHING
20252_D4T_0	20252	0	D4T	NOTHING

Table 4: Sample selection extracted from the treatment table.

Data tables and data sources

The HIV-PDI database is implemented as a set of relational tables, as shown (Figure 3). Each main entity or relationship in the conceptual model typically maps to a relational table in the physical model. For example, the *Ligand* table contains several one-dimensional (1D) and two-dimensional (2D) attributes describing molecular properties such as compound name, chemical formula, pKa, hydrogen bond donor and acceptor information, along with other structural features. Each entry in this table is labeled according to its International Union of Pure and Applied Chemistry (IUPAC) International Chemical Identifier (InChiKey) [45]. An InChiKey identifier is a textual identifier for chemical substances, and is designed to provide a standard human-readable way to encode molecular information and to facilitate chemical property searching in databases and on the internet. Nonetheless, a given ligand may be commonly known using several alternative names, and these are stored in the *LigandSynonym* table. The *LigandMolecularStructure* table stores the 3D molecular descriptors of drug structures such as atomic coordinates, SH molecular surface shape coefficients [46-48], molecular polarizability, and stereochemical state, for example. The *ChemicalGroup* table stores details of the chemical subgroups contained within each ligand compound.

The data on compounds defined as HIV protease inhibitors was collected from four main sources. These are compounds approved as commercial drugs by the US Food and Drug Administration (FDA), patents, compounds undergoing clinical trials, compounds from

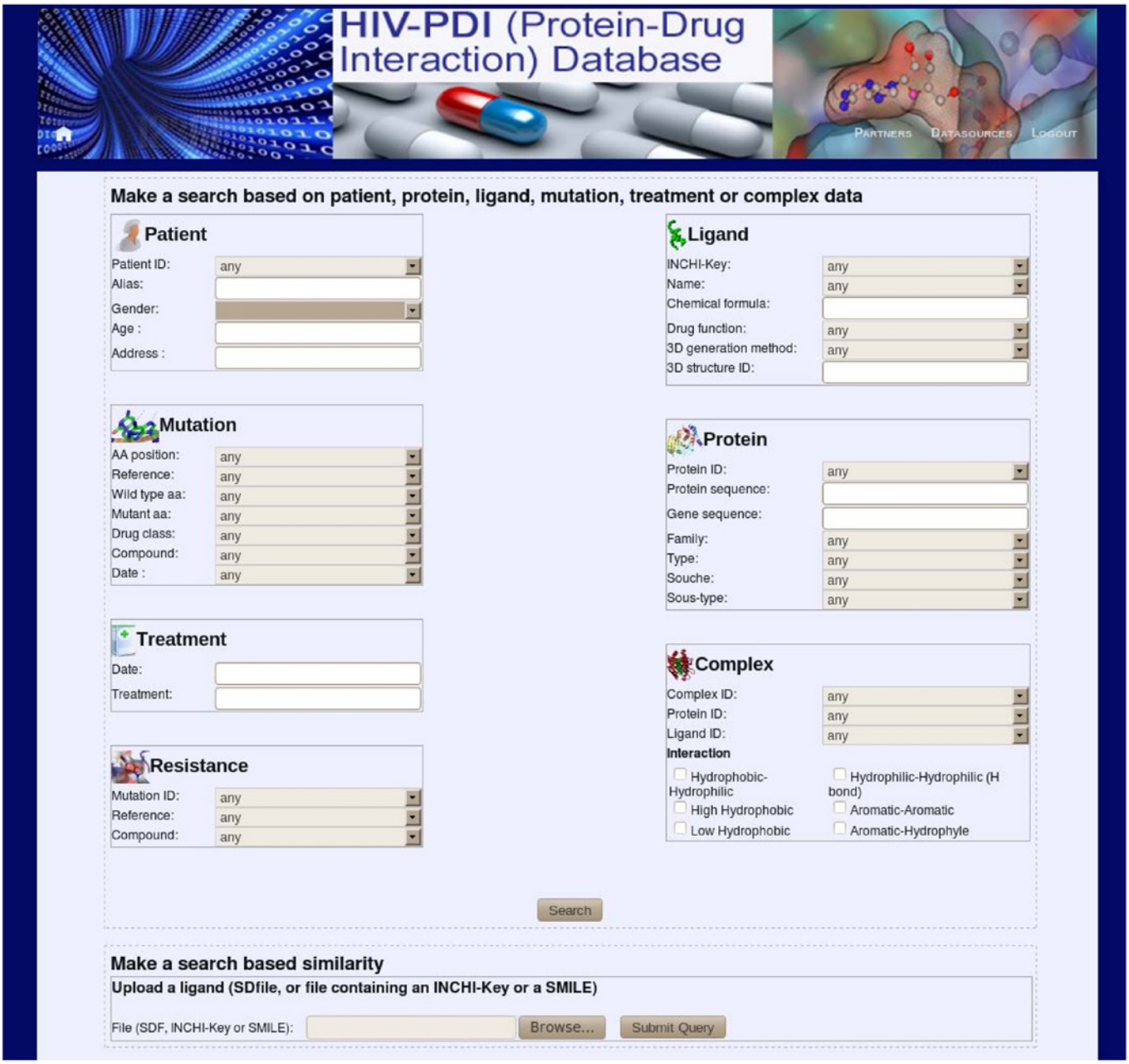

Figure 4: The screenshot of the main GUI page for accessing all fields in the HIV-PDI.

the Life Chemical [49] database which are reported as HIV-related compounds, and HIV-related compounds described in articles published in various international peer-reviewed journals. The data on each compound identified from these sources is stored in the HIV-PDI database. Additional 3D conformations and molecular descriptors were calculated and stored, using Omega [50] to generate multiple conformations for each compound, and by using the Chemaxon and InChI tools [45,51] to calculate the descriptors for each conformer. In a similar manner, the *Protein* table stores 1D and 2D information about HIV-related proteins, such as their family and sub-family types, and genetic and amino acid sequences. The *ProteinMolecularStructure* table stores 3D protein structure information, such as the atomic coordinates of all of the amino acids in a protein, the SH coefficients of a protein's surface shape and the shape of its ligand binding site. If a protein is known using several synonyms, these may be stored in the *ProteinSynonym* table.

The main sources of protein information are Swiss-Prot [33] and the PDB. The WT sequence of HIV-1 was collected from the Stanford HIVdb. This sequence has a Swiss-Prot identifier of Q9WFL7 and the PDB code is 2QMP. As is usual practice, this WT sequence is taken as the reference sequence against which all mutations are defined. When

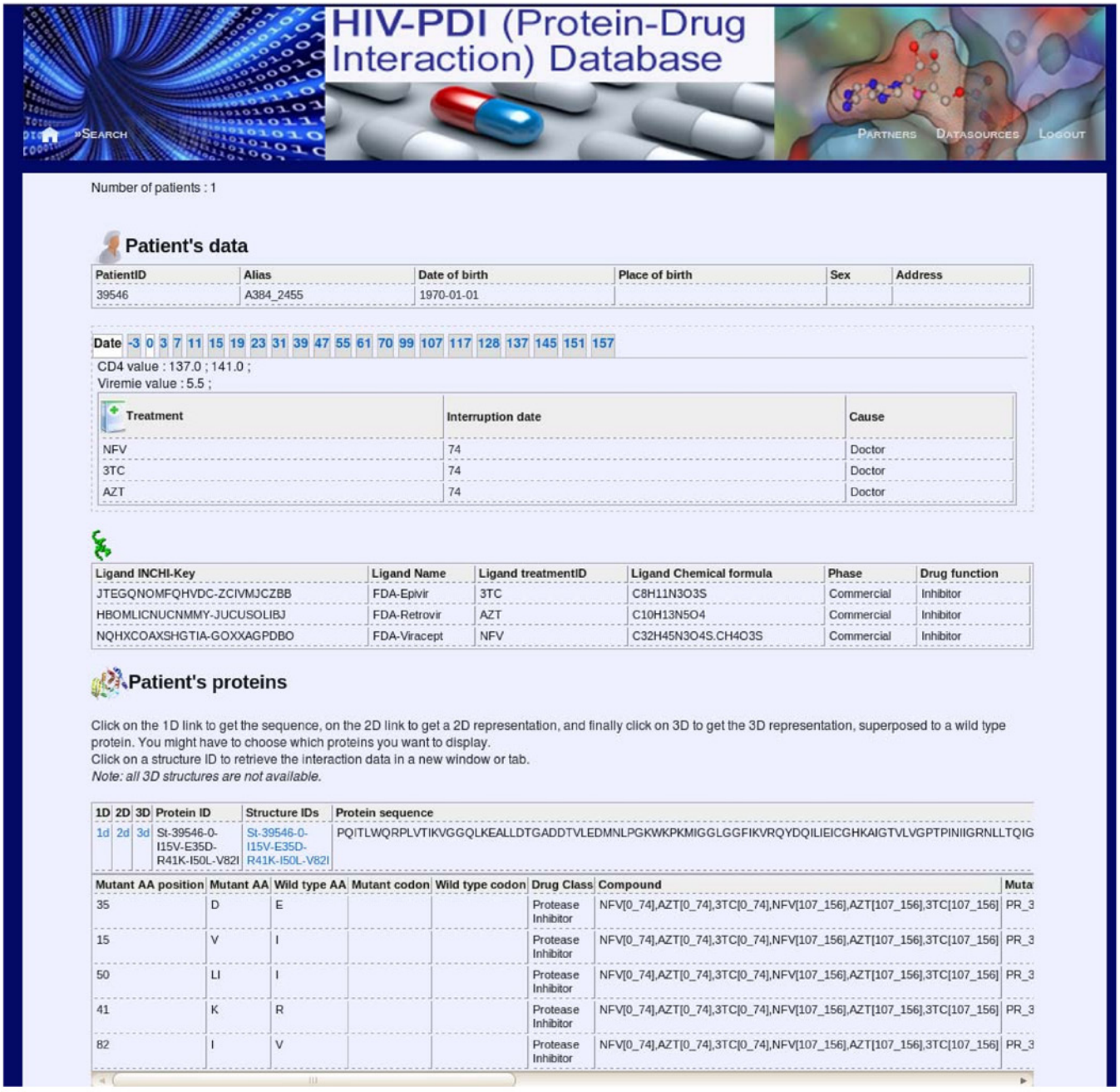

Figure 5: The screenshot of the GUI for patient's data in the HIV-PDI.

no X-ray structure for a given mutation is available in the PDB, a 3D model of the corresponding sequence was built by homology using the 2QPM structure as a starting template in the Modeler program [52,53]. The experimental and model-built protein structures were then used as starting structures for molecular dynamics (MD) simulations which were used to generate multiple conformational samples of each protein. Each MD run consisted of a 1 ns simulation in explicit solvent, as calculated by the NAMD software [54]. This produced a set of conformational states for each protein for subsequent 3D shape analyses and protein-ligand docking simulations. The MSSH program was used to compute SH-based descriptors of the volumes and surfaces of each protein conformation [46,55].

The *ProteinMolecularStructure_has_LigandMolecularStructure* table stores inform-ation concerning protein-drug complexes, such as affinity data (EC-50, KI, and IC-50), lists of 3D interactions, and the RMS difference between the SH shapes of the protein and ligand surfaces [46]. Data on protein-drug interactions were extracted from the 3D structures of known HIV protease-drug complexes. The affinity of protein-drug complexes was extracted from the corresponding scientific literature. For compounds for which no protein-ligand structure is available in the PDB, flexible protein-ligand docking was performed using the Glide program [44,45,56-58] in order to generate 3D structures of complexes between each ligand and the WT and mutated HIV proteases. Lists of residues interacting with each ligand were extracted from the modeled complex using the LPC software [59]. The SH coefficient RMSD (quality of fit) between each protein pocket and ligand molecule was calculated using the SHEF program [55].

The *Patient* table stores information relating to individual patients such as their name, sex, and date of birth. The *CD4* table stores data on the CD4 T-lymphocyte counts recorded for individual patients (cells/μL, measurement date). The *Viremia* table stores data on the plasma viral load of patients (copies of viral genome/mL, measurement date). Similarly, the *Treatment* and *Treatment Break* tables store data on the patients' treatments, such as the date of the treatment, the treatment specificity, details on interruption schedules and any clinical remarks. The *Mutation* table stores data on each patient's HIV mutations, such as the WT position and type of the mutated amino acid and the corresponding mutated codons. The *Resistance* table stores details of the compounds for a patient found to have a resistant strain of the virus.

The clinical data on HIV-infected patients was extracted from the Stanford HIVdb. These data are related by a unique identification number for each patient. The list of known mutations was extracted from the available literature and from the Stanford HIVdb, and from the Los Alamos, Swiss-Prot, PDB, and International AIDS Society-USA (IAS-USA) [60] databases [61,62] (Supplementary information 1). The full sequences of mutated proteins were deduced by comparison with the reference sequence of the WT HIV protease.

Results

Database implementation

The relational database is implemented in PostgreSQL [63] (version 8.3.7) on a 64-bit Linux operating system running the Apache web server. Python wrapper scripts were used to import data from external data sources into the database. The Structured Query Language (SQL) [40,64] is used to retrieve and manage the data stored in the HIV-PDI database. For users who are not familiar with using SQL, a web-based graphical user interface (GUI) has been implemented using Python and the object-oriented web application framework CherryPy [65].

Database statistics

Table 1 lists some overall statistics on the number of records currently stored in the HIV-PDI database. Currently, the *Ligand* table contains a total of 5,962 non-redundant entries concerning HIV-related compounds. This data can be extracted using a SQL query such as "*SELECT PatentInChiKey, HBDonor, TautomerNumber, AtomNumber, LogP, PolarSurfaceArea, Stage, Class, DrugFunction, Diseasename, Name FROM Ligand;*". Table 2 shows a partial listing of the results of this query.

The *Protein* table contains 2,493 entries. This table may be listed with the SQL query "*SELECT PatentidProtein, Patient_idPatient, ProteinSequence, GeneSequence, Family, Type2, Souche FROM Protein;*" (see Table 3). The *Patient* table holds 2,029 entries. It may be queried using "*SELECT idPatient, Alias, DateBirth, PlaceBirth, Sex, Address FROM Patient;*". The *Mutation* table contains 7,791 entries. It may be queried using: "*SELECT PatentidMutation, AAPosition, Reference, WildTypeAA, MutantAA, DrugClass, Compound FROM Mutation;*". The *Treatment* table holds 13,629 records. It may be queried using "*SELECT idTreatment, Patient_idPatient, DateTreatment, Treatment, Comment_2 FROM Treatment;*" (see Table 4).

Graphical user interface

The main page of the graphical user interface (GUI) (see Figure 4, which shows the screenshot of the main GUI page for accessing all fields in the HIV-PDI) allows access to all of the data present in the database using several main request categories. All of the database entities may be explored using this page. For example, the result of a request to visualize the data stored for patients (see Figure 5, which shows the screenshot of the GUI for patient's data in the HIV-PDI) presents additional links to view the data in more detail (1D, 2D, and 3D visualizations are available).

The database may also be queried directly by requests on entities such as *Mutation* (see Figure 5), *Ligand, Protein* or *Patient*, for example. More complex queries are also possible by using a combination of requests on different entities. Because all of the entity levels are linked to each other, the GUI pages allow the user to traverse easily from patient data to molecular structural information or from protein sequences and therapies to drug resistance and treatments, for example.

Discussion

An important aspect of the present work is that a conceptual data model was employed to link data and information from diverse domains throughout the database design process. By transferring clinical and biological user requirements to a graphical data model, the conceptual design approach allowed all relevant data and relationships to be considered at a glance. The resulting tables that make up the relational database contain data on clinical and experimental measurements, 3D structural data relating to proteases and anti-protease drugs, along with and physico-chemical properties and molecular descriptors of known drug molecules and drug candidates.

Taken together, this integrated resource now makes it possible to perform a wide range of database queries which will be useful for helping to understand the structural causes of ARV resistance phenomenon observed in patients. To our knowledge, the HIV-PDI is the first HIV-related database and resource which fully integrates clinical and biological data on HIV proteins with 3D structural and physico-chemical information on ARV drugs and drug candidates. Overall, the HIV-PDI database supports patient-oriented queries, efficient follow

up of patient treatments, convenient reviews of treatment outcomes, and systematic documentation of ARV drug resistance. It also brings the promise of being able to identify alternative treatments for observed cases of resistance. Hence one ultimate goal of our HIV-PDI resource is to support an integrated bioinformatics platform to provide a decision-support tool to help clinicians making informed choices for second line treatments against HIV in patients presenting resistance. Furthermore, this platform could also be used for identifying and designing new ARVs which can address the on-going problem of increasing viral resistance to existing therapies.

Conclusion

We have designed and implemented the HIV-PDI database to relate the problems of ARV drug resistance in patients to protein-ligand interactions at the molecular level. The HIV-PDI database therefore provides a useful repository for the collection and interpretation of diverse information which could be crucially important in understanding HIV disease processes at the molecular level and in proposing alternative ARV therapies on a patient-specific basis. The introduction of 3D structural analysis about the protein/ligand complexes in a database related to HIV and the consequences of mutations on the stability of the protein/ligand complexes make the main difference with the others known databases. A subsequent paper will present examples of using the HIV-PDI resource in the context of clinical decision support for dealing with patients presenting resistance to current ARV therapies.

Future developments of HIV-PDI resource will include adding data for additional HIV protein targets and drugs, linking out to more chemoinformatics modules to provide richer and more sensitive analyses of protein-ligand interactions and to perform more sensitive comparisons of ligand physico-chemical properties. We will also provide links to knowledge extraction modules to predict virological responses to therapy or resistant mutations [10,11,27,66-68]. Therefore, our database platform could open new avenues for identifying and evaluating new drugs which can bind to specific mutated HIV proteins with high affinity and thus provide new and more tailored treatments for ARV-resistant HIV mutations.

Authors' contributions

All authors have jointly developed the research concept and collaborated on the writing of the manuscript. As the main author Leo GHEMTIO has initiated the study, carried out the computational analyses, has interpreted the results, and drafted the manuscript. All authors revised the manuscript and approved its final version.

Acknowledgments

The authors thank the *Bill & Melinda Gates Foundation* for financial support through the Grand Challenge Exploration grant N° 52034 (Round I). They are grateful to Michel Souchet, Dave Ritchie, Birama Ndiaye, and Florent Petronin for their respective contribution to the present work.

Leo Ghemtio was supported by grants from INRIA (Institut National de Recherche en Informatique et en Automatique), CNRS (Centre National pour la Recherche Scientifique) and the *Bill & Melinda Gates Foundation*;

Lionel Keminse and Joseph Fokam were supported by grant from the *Bill & Melinda Gates Foundation*.

We thank Openeye and Chemaxon for providing academic licences for their software.

This work was also supported in part by Region Lorraine within the framework of the PRST MISN (MBI operation).

References

1. Dau B, Holodniy M (2009) Novel targets for antiretroviral therapy: clinical progress to date. Drugs 69: 31-50.
2. Temesgen Z, Warnke D, Kasten MJ (2006) Current status of antiretroviral therapy. Expert Opin Pharmacother 7: 1541-1554.
3. Paar C, Palmetshofer C, Flieger K, Geit M, Kaiser R, et al. (2008) Genotypic antiretroviral resistance testing for human immunodeficiency virus type 1 integrase inhibitors by use of the TruGene sequencing system. J Clin Microbiol 46: 4087-4090.
4. Shafer RW (2002) Genotypic testing for human immunodeficiency virus type 1 drug resistance. Clin Microbiol Rev 15: 247-277.
5. Mascolini M, Larder BA, Boucher CA, Richman DD, Mellors JW (2008) Broad advances in understanding HIV resistance to antiretrovirals: report on the XVII International HIV Drug Resistance Workshop. Antivir Ther 13: 1097-1113.
6. Pillay D (2007) The priorities for antiviral drug resistance surveillance and research. J Antimicrob Chemother 60: i57-58.
7. Garriga C, Perez-Elias MJ, Delgado R, Ruiz L, Najera R, et al. (2007) Mutational patterns and correlated amino acid substitutions in the HIV-1 protease after virological failure to nelfinavir- and lopinavir/ritonavir-based treatments. J Med Virol 79: 1617-1628.
8. Kovalevsky AY, Chumanevich AA, Liu F, Louis JM, Weber IT, et al. (2007) Caught in the Act: the 1.5 A resolution crystal structures of the HIV-1 protease and the I54V mutant reveal a tetrahedral reaction intermediate. Biochemistry 46: 14854-14864.
9. Tie Y, Kovalevsky AY, Boross P, Wang YF, Ghosh AK, et al. (2007) Atomic resolution crystal structures of HIV-1 protease and mutants V82A and I84V with saquinavir. Proteins 67: 232-242.
10. Wang YF, Tie Y, Boross PI, Tozser J, Ghosh AK, et al. (2007) Potent new antiviral compound shows similar inhibition and structural interactions with drug resistant mutants and wild type HIV-1 protease. J Med Chem 50: 4509-4515.
11. Hou T, Zhang W, Wang J, Wang W (2009) Predicting drug resistance of the HIV-1 protease using molecular interaction energy components. Proteins 74: 837-846.
12. Kontijevskis A, Prusis P, Petrovska R, Yahorava S, Mutulis F, et al. (2007) A look inside HIV resistance through retroviral protease interaction maps. PLoS Comput Biol 3: e48.
13. Shuman CF, Markgren PO, Hamalainen M, Danielson UH (2003) Elucidation of HIV-1 protease resistance by characterization of interaction kinetics between inhibitors and enzyme variants. Antiviral Res 58: 235-242.
14. Beerenwinkel N, Sing T, Lengauer T, Rahnenfuhrer J, Roomp K, et al. (2005) Computational methods for the design of effective therapies against drug resistant HIV strains. Bioinformatics 21: 3943-3950.
15. Ghosh AK, Chapsal BD, Weber IT, Mitsuya H (2008) Design of HIV protease inhibitors targeting protein backbone: an effective strategy for combating drug resistance. Acc Chem Res 41: 78-86.
16. Altuglu I, Cavusoglu C, Cicek C, Tunger O (2007) Development of a database for tracking HIV positive/AIDS patients. Mikrobiyol Bul 41: 101-108.
17. Cohen J (2009) HIV/AIDS Tangled patent dispute over 'free' drug-resistance database. Science 323: 1156-1157.
18. Dunn D, Pillay D (2007) UK HIV drug resistance database: background and recent outputs. J HIV Ther 12: 97-98.
19. Kantor R, Machekano R, Gonzales MJ, Dupnik K, Schapiro JM, et al. (2001) Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database: an expanded data model integrating natural language text and sequence analysis programs. Nucleic Acids Res 29: 296-299.
20. Pan C, Kim J, Chen L, Wang Q, Lee C (2007) The HIV positive selection mutation database. Nucleic Acids Res 35: D371-375.
21. Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 31: 298-303.
22. [http://www.hiv.lanl.gov/content/index]
23. [http://hivdb.stanford.edu/]
24. Rhee SY, Fessel WJ, Liu TF, Marlowe NM, Rowland CM, et al. (2009)

Predictive value of HIV-1 genotypic resistance test interpretation algorithms. J Infect Dis 200: 453-463.

25. http://www.ncbi.nlm.nih.gov/pubmed/17369658
26. [http://www.hivrdi.org/]
27. Wang D, Larder B, Revell A, Montaner J, Harrigan R, et al. (2009) A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy. Artif Intell Med 47: 63-74.
28. Anderson J, Schiffer C, Lee SK, Swanstrom R (2009) Viral protease inhibitors. Handb Exp Pharmacol: 85-110.
29. Hong L, Zhang XC, Hartsuck JA, Tang J (2000) Crystal structure of an in vivo HIV-1 protease mutant in complex with saquinavir: insights into the mechanisms of drug resistance. Protein Sci 9: 1898-1904.
30. Lexa KW, Damm KL, Quintero JJ, Gestwicki JE, Carlson HA (2009) Clarifying allosteric control of flap conformations in the 1TW7 crystal structure of HIV-1 protease. Proteins 74: 872-880.
31. Perryman AL, Lin JH, Andrew McCammon J (2006) Optimization and computational evaluation of a series of potential active site inhibitors of the V82F/I84V drug-resistant mutant of HIV-1 protease: an application of the relaxed complex method of structure-based drug design. Chem Biol Drug Des 67: 336-345.
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN (2000) The Protein Data Bank. Nucleic Acids Res. 28: 235-242.
33. [http://www.expasy.ch/sprot/]
34. [http://www.ncbi.nlm.nih.gov/Genbank/]
35. [http://www.ebi.ac.uk/embl/]
36. Bajorath J (2001) Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. J Chem Inf Comput Sci 41: 233-245.
37. Bajorath J (2002) Chemoinformatics methods for systematic comparison of molecules from natural and synthetic sources and design of hybrid libraries. Mol Divers 5: 305-313.
38. Godden JW, Bajorath J (2002) Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis. J Chem Inf Comput Sci 42: 87-93.
39. Oprea TI, Matter H (2004) Integrating virtual screening in lead discovery. Curr Opin Chem Biol 8: 349-358.
40. Kabachinski J (2008) Databases, Tuples, and SQL. Biomed Instrum Technol 42: 385-387.
41. Teorey TJ (2006) Database Modeling And Design. Morgan Kaufmann 4: 275.
42. Thompson CB, Sward K (2005) Modeling and teaching techniques for conceptual and logical relational database design. J Med Syst 29: 513-525.
43. [http://fabforce.net/dbdesigner4/]
44. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, et al. (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47: 1739-1749.
45. Coles SJ, Day NE, Murray-Rust P, Rzepa HS, Zhang Y (2005) Enhancement of the chemical semantic web through the use of InChI identifiers. Org Biomol Chem 3: 1832-1834.
46. Cai W, Shao X, Maigret B (2002) Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. J Mol Graph Model 20: 313-328.
47. Mavridis L, Hudson BD, Ritchie DW (2007) Toward high throughput 3D virtual screening using spherical harmonic surface representations. J Chem Inf Model 47: 1787-1796.
48. Yamagishi ME, Martins NF, Neshich G, Cai W, Shao X, et al. (2006) A fast surface-matching procedure for protein-ligand docking. J Mol Model 12: 965-972.
49. http://www.lifechemicals.com/
50. http://www.eyesopen.com/
51. http://www.chemaxon.com/
52. Eswar N, Eramian D, Webb B, Shen MY, Sali A (2008) Protein structure modeling with MODELLER. Methods Mol Biol 426: 145-159.
53. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2007) Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci 2:2-9.
54. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E et.al (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26:1781-1802.
55. Cai W, Xu J, Shao X, Leroux V, Beautrait A, et al. (2008) SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces. J Mol Model 14: 393-401.
56. Miteva MA, Lee WH, Montes MO, Villoutreix BO (2005) Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. J Med Chem 48: 6012-6022.
57. Bajorath J (2002) Integration of virtual and high-throughput screening. Nat Rev Drug Discov 1: 882-894.
58. Good AC, Krystek SR, Mason JS (2000) High-throughput and virtual screening: core lead discovery technologies move towards integration. Drug Discov Today 5: 61-69.
59. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M (1999) Automated analysis of interatomic contacts in proteins. Bioinformatics 15: 327-332.
60. http://www.iasusa.org/
61. Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, et al. (2009) Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. PLoS One 4 :e4724.
62. Johnson VA, Brun-Vezinet F, Clotet B, Gunthard HF, Kuritzkes DR (2008) Update of the Drug Resistance Mutations in HIV-1. Top HIV Med 16: 138-145.
63. http://www.postgresql.org/
64. Jamison DC (2003) Structured Query Language (SQL) fundamentals. Curr Protoc Bioinformatics 9: 2.1-2.29.
65. http://cherrypy.org/
66. Glasgow J, Jurisica II, Ng R (2000) Data mining and knowledge discovery in molecular databases. Pac Symp Biocomput: 365-366.
67. Ghose AK, Viswanadhan VN, Wendoloski JJ (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. J Comb Chem 1: 55-68.
68. Tsai YS, King PH, Higgins MS, Pierce D, Patel NP (1997) An expert-guided decision tree construction strategy: an application in knowledge discovery with medical databases. Proc AMIA Annu Fall Symp: 208-212.