max planck institut
informatik

Introduction to HIV Databases

Alejandro Pironti

Dar es Salaam

November 14th, 2011

euresist

What is a Database?

- An organized collection of data
- Usually organized to model relevant aspects of reality
- Examples:
  - Rooms in a hotel
  - Books in a library
  - Items in a store
  - Clinical HIV data
  - Protein data

Image: http://www.jofwidata.com/images/database-design-development.jpg

Database Management Systems

- Digital databases are stored in computers
- The software collection used to create and manage a database is called a database management system (DBMS)
- Today, we will focus on a DBMS with a relational model

Image:http://www.life123.com/technology/computer-software/database-software/a-guide-to-database-management-software.shtml

The Relational Database Model

- Data in a relational database is kept in tables
- Tables have a fixed structure of rows and columns
- Tables relate to each other by the means of identifiers called *keys*
- A computer language is used to alter and query the database

Patients			
patientID: int	Name: varchar	Gender: varchar	Born: date
1	Robert Williams	Male	14/01/1955
2	Jaime Gonzalez	Male	25/09/1983
3	Lisa Schmidt	Female	30/01/1990

PatientDiagnoses				
pDiagnoseID : int	diagnoseID : int	patientID: int	statusID: int	Diagnosis Date: date
1	1	1	1	20/02/2011
2	2	1	1	25/08/2011
3	2	2	2	18/10/2011

The Relational Database Model

Patients			
patientID: int	Name: varchar	Gender: varchar	Born: date
1	Robert Williams	Male	14/01/1955
2	Jaime Gonzalez	Male	25/09/1983
3	Lisa Schmidt	Female	30/01/1990

PatientDiagnoses				
pDiagnoseID : int	diagnoseID : int	patientID: int	statusID: int	Diagnosis Date: date
1	1	1	1	20/02/2011
2	2	1	1	25/08/2011
3	2	2	2	18/10/2011

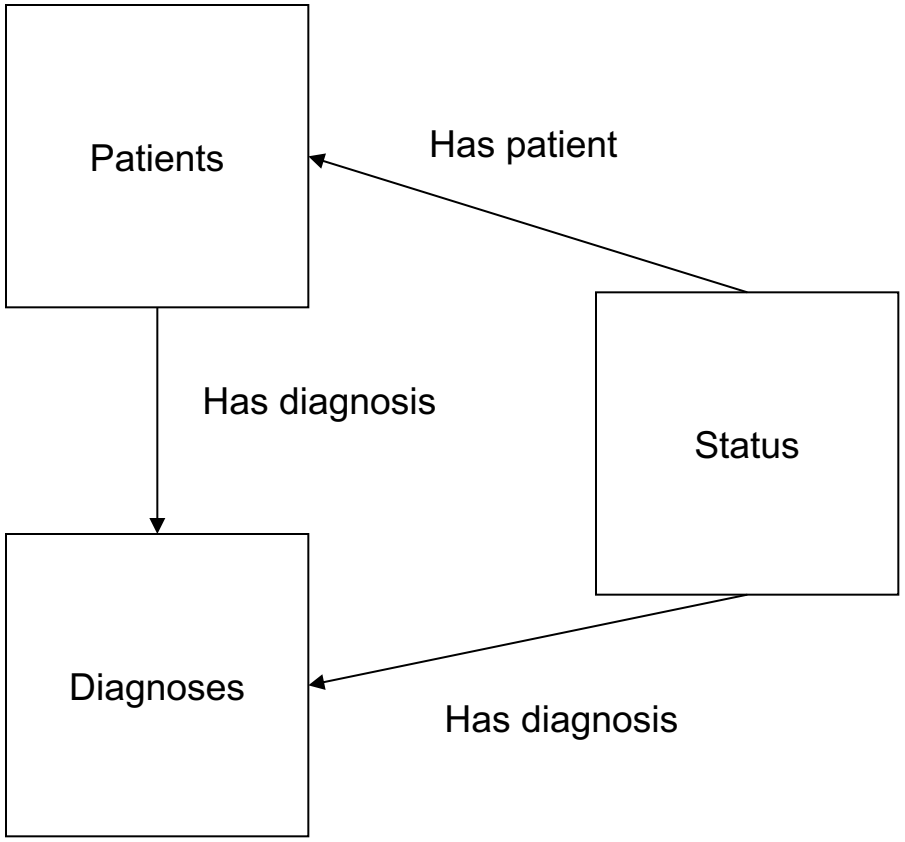
Diagnoses	
diagnoseID: int	Diagnose: varchar
1	Viral pneumonia
2	Nasopharyngitis
3	Dengue fever

Status	
statusID: int	status: int
1	Resolved
2	On treatment
3	Resulted in death

The Relational Database Model

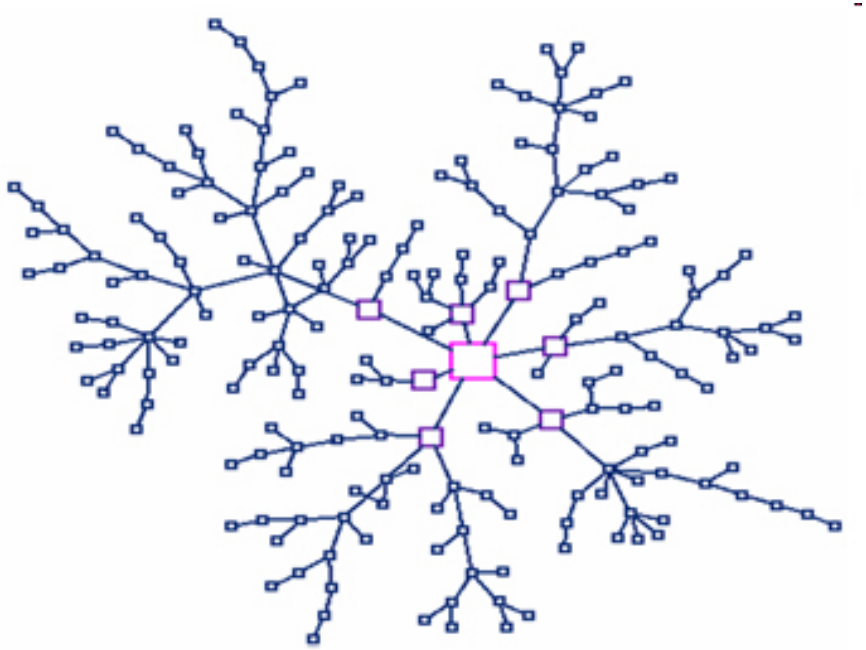
- A data model is needed to store data efficiently
- Tables represent either entities, e.g.
  - Patients
  - Diagnoses
  - Status

  or the relationships between the entities, e.g.
  - Which diagnosis was given to which patient and when
  - What was the status of a patients diagnosis and when



Database Models

- Databases should be capable of holding large amounts of complexly interrelated information
- Entities and relationships may change
- Flexibility is required
- When designing a database, you have to look well ahead



Database Models

- A good compromise between flexibility and ease of use has to be found
- Separating information into many tables increases flexibility
- However, this complicates use, since tables have to be joined

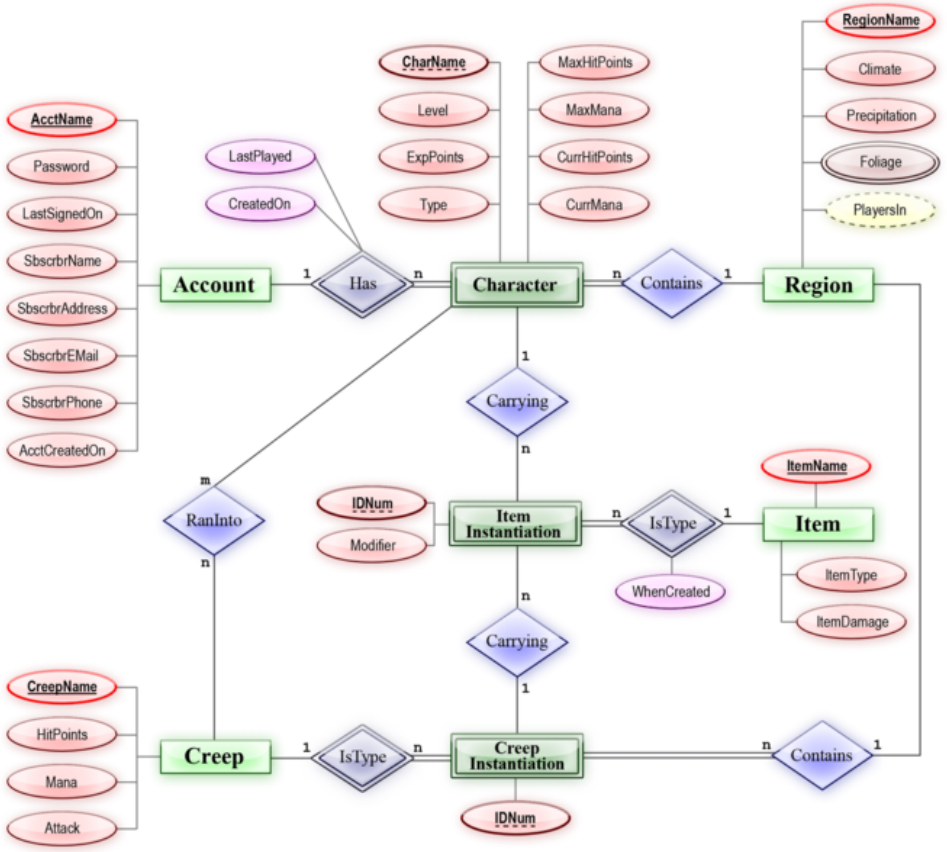


Image: http://en.wikipedia.org/wiki/Database_schema

MySQL

- Relational database management system
- SQL: Structured query language
- Developed in 1994 by Michael Widenius and David Axmark
- Freely available

Important Table Column Characteristics in MySQL

- Data types:
  - INT: integer numbers
  - DECIMAL: floating point numbers
  - VARCHAR: text
  - DATETIME: dates and times
  - DATE: dates
- Column characteristics:
  - Primary Key: ID for a row
  - Not null: Empty entries not allowed
  - Autoincrement: Automatically increment value (ID assignment)

Patients			
patientID: int	Name: varchar	Gender: varchar	Born: date
1	Robert Williams	Male	14/01/1955
2	Jaime Gonzalez	Male	25/09/1983
3	Lisa Schmidt	Female	30/01/1990

Diagnoses	
diagnoseID: int	Diagnose: varchar
1	Viral pneumonia
2	Nasopharyngitis
3	Dengue fever

Steps in Creating a Database

1. Think
    1. What do you want to store?
    2. What are the requirements?
    3. How could the requirements change with time?
2. Design a database model
3. Implement the model in the computer
4. Fill with data

Demonstration

Creating a Database Scheme

- A database scheme is a collection of tables

create schema *schema_name*

Creating Tables

- Tables organize information in the database

use *schema_name*

create table table_name (*column1 type1, column2, type2*,…,primary_key(*column*))

Inserting Information

- Information can be stored in the tables

insert into table_name values (*value1, value2*, …)

Retrieving Information: WHERE

- If we want to retrieve information in tables that fulfills certain characteristics we use WHERE clauses

select *column1, column2*, … from *table1, table2*, … where *column1=property1* and *column2 =property2* and …

Retreiving Information: JOIN

- If we want to join two tables we use a JOIN query

```
select column1, column2, … from table1 join
    table2 on (table1.column1 = table2.column2)
    where column1=property1 and column2
    =property2 and …
```

Summary

- Data is stored in databases
- Databases are collections of tables
- Tables represent real-world objects or notions and the relationships among them
- You need to put thought into database design
- MySQL is a popular database management system
- Basic database commands:
  - Create Tables
  - Fill information into tables
  - Query for specific information in tables and reconstruct relationships among tables

EuResist

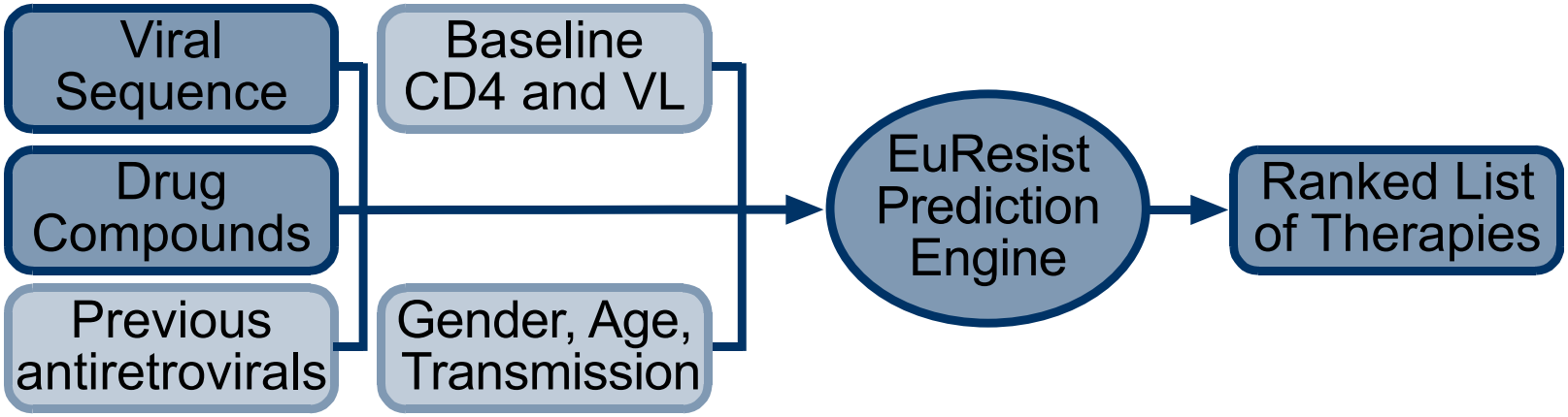
- EuResist is a non-profit network composed of members:
  - The Karolinska Institute, University of Siena, University of Cologne, Max Planck Institute for Informatics, Informa s.r.l.
- And partners:
  - IBM Research, Rega Institute, irsiCaixa, CPR-Santé, Centro Hospitalar de Lisboa Ocidental
- The EuResist Integrated Database (EIDB) a database comprising clinical, demographic, and genomic data of HIV-infected patients:
  - > 49 000 patients
  - >127 000 treatments
  - > 50 000 Pol sequences
  - > 548 000 viral load measurements
  - >1000 V3 Loops

max planck institut informatik

euresist

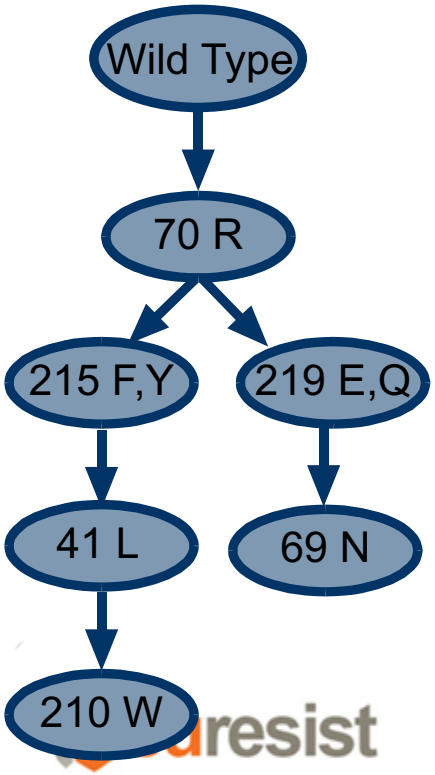
The EuResist Prediction Engine

- A data-driven system to assist in the selection of combination antiretroviral therapies.
- Given a viral sequence and a set of drugs, EuResist ranks a list of combination therapies according to their success probability
- The engine can be provided with additional information in order to enhance prediction accuracy



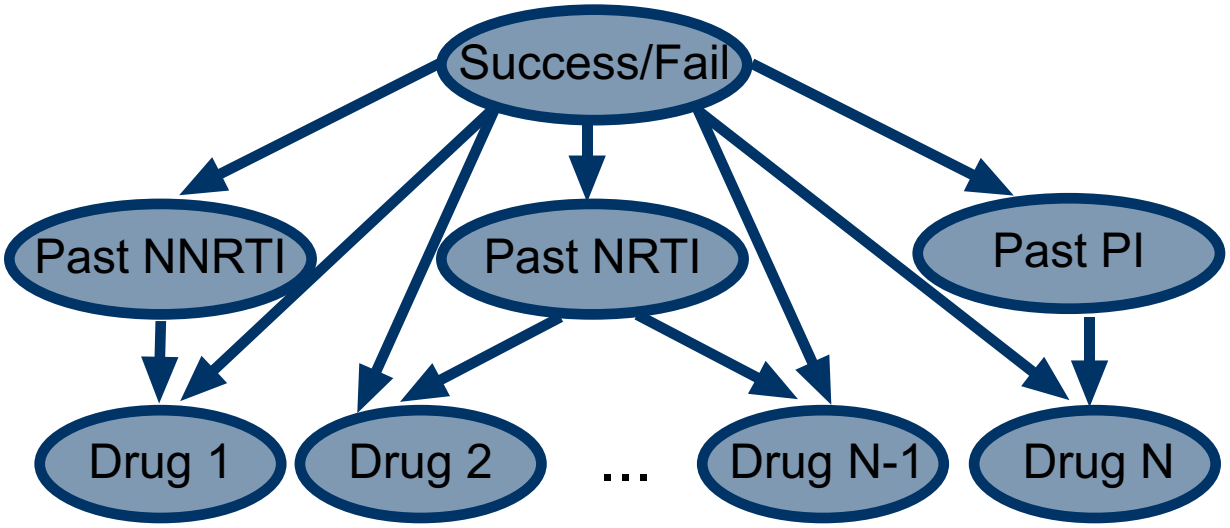
The Statistical Learning Engines

- The EuResist Prediction Engine is composed of three sub-engines. Each of the engines concentrates on different aspects:
  - The **Evolutionary Engine** uses mutagenetic trees to compute the genetic barrier to drug resistance. This feature is used in a logistic regression.



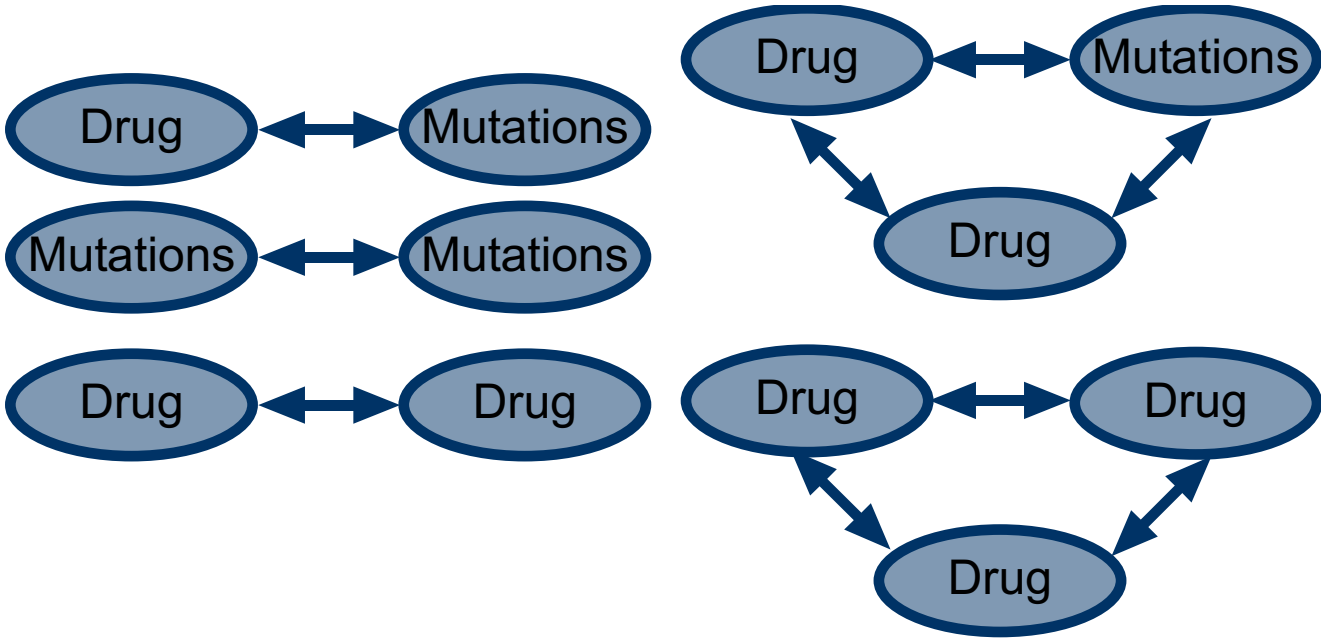
The Statistical Learning Engines

- The EuResist Prediction Engine is composed of three sub-engines. Each of the engines concentrates on different aspects:
  - The **Generative Discriminative Engine** employs logistic regression and a Bayesian network modeling interactions between current and past antiretroviral drugs



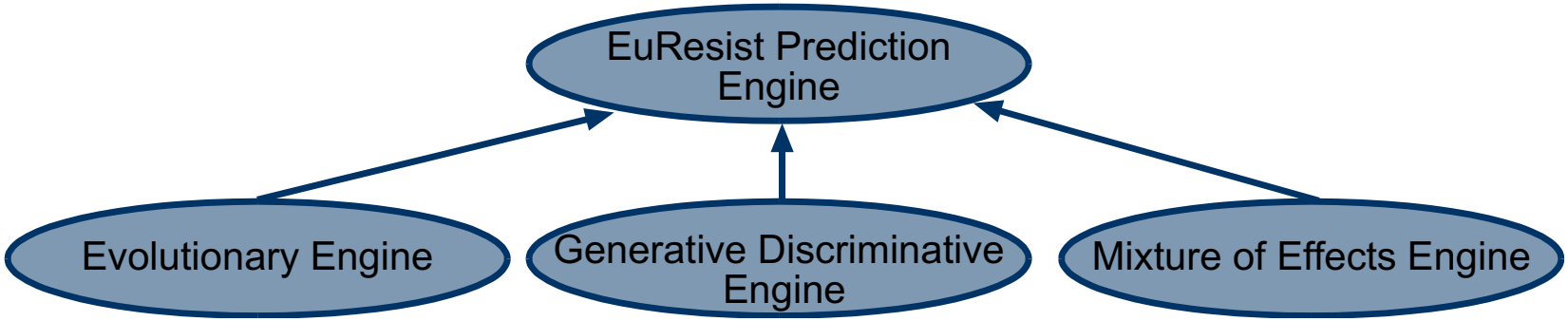
The Statistical Learning Engines

- The EuResist Prediction Engine is composed of three sub-engines. Each of the engines concentrates on different aspects:
  - The **Mixture of Effects Engine** includes second and third-order variable interactions between drugs and mutations and uses them in a random forest



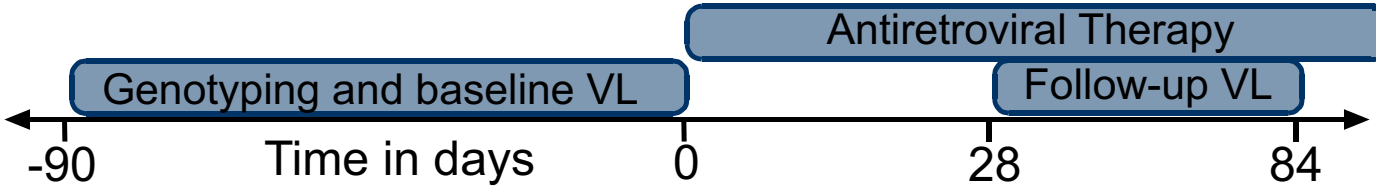
The Statistical Learning Engines

- The EuResist Prediction Engine is composed of three sub-engines. All engines predict therapy success by using logistic regression but concentrate on different aspects:
  - The predictions of each of the engines are combined into one by consensus, which amounts to the average



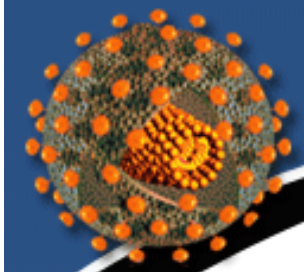
The Standard Datum Definition

- In order to be trained, the prediction engines need examples of successful and failing therapies
- An adequate definition of success and failure is essential for good performance
- Success: ΔVL at least 100-fold or VL ≤ 500 cp/ml at week 8 (within weeks 4-12)
- Failure: Otherwise



EuResist Satellite DB

- Local database tool for management of HIV patient data
- Permits uploading of stored data to the EuResist Database, thus facilitating cooperations
- Freely available at http://satellite.euresist.org



Los Alamos National Laboratory HIV Databases

- Contains many databases
  - HIV Sequence Database
    - Sequence Analyses
    - Clinical information
  - Vaccine Trials Database
  - CTL/CD8+ T-Cell Epitope Database
  - T Helper/CD4+ T-Cell Epitope Database
  - Antibody Database
- Publicly available on the internet
- http://www.hiv.lanl.gov/content/index

Stanford HIV Database

- Mutations that arise during antiretroviral therapy
- Drug susceptibility information for isolates with certain mutations
- Summaries of clinical studies
- Collection of treatment change episodes
- Drug resistance tables
- Listing of literature evidence for specific mutations
- Mutations by treatment and subtype
- Drug-resistance-mutation interpretation tools