

Conceptual Modeling of Genetic Studies and Pharmacogenetics

Xiaohua Zhou and Il-Yeol Song

College of Information Science and Technology, Drexel University

3141 Chestnut Street, Philadelphia, PA 19104, USA

{xiaohua.zhou, song}@drexel.edu

Abstract. Genetic Studies examine relationships between genetic variation and disease development. Pharmacogenetics studies the responses to drugs against genetic variation. These two lines of research evaluate relationships among genotype, phenotype, and environment regarding subjects. These studies demand a variety of other information; such as clinical observations, disease development history, demographics, life style, and living environment. Correct and informative modeling of these data is critical for bioinformaticians; the model affects the capacity of data manipulation and the types of queries they can ask as well as performance of the implemented system. In this paper, we present a conceptual model on genetic studies and Pharmacogenetics using Unified Modeling Language (UML). Our model provides a comprehensive view of integrated data for genetic studies and Pharmacogenetics by incorporating genomics, experimental data, domain knowledge, research approaches, and interface data for other publicly available resources into one cohesive model. Our model can support diverse biomedical research activities that use both clinical and biomedical data to improve patient care through incorporation of the roles of environment, life style and genetics. Our model consists of a set of class diagrams organized into a hierarchy of packages diagrams to clearly and intuitively show inter-object relationships at different levels of complexity.

1 Introduction

Completion of a high quality comprehensive sequence of the human genome marked the arrival of the Genomic Era. Genomics heavily impacts research in the fields of biology, health, and society. Clinical opportunities for gene-based pre-symptomatic prediction of illness and adverse drug responses are emerging at a rapid pace [2].

Genetic studies examine relationships between genetic variation and disease susceptibility. Pharmacogenetics is the study of response to drugs against genetic variation. These two lines of research address the triad of relationships among genotype, phenotype, and environment. Apart from genomic information, a variety of other information is required such as clinical observations, disease development history, demographics, life style, and living environment. Comprehensive and correct modeling of these data is necessary for bioinformaticians; the model affects the capacity of data manipulation and the types of queries they can ask as well as performance of the implemented system.

There have been a lot of cutting-edge works on conceptual modeling of biological data. However, most focus on a single type of data such as genome sequences, protein structures, protein interactions, and metabolic pathways (see Section 2 for details). Few works address the problem of conceptual modeling on comprehensive applications like genetic and Pharmacogenetics research. Though a large

number of works address implementation of complex biologic database systems, especially distributed ones through recent technology like CORBA [8] [9], they address the difficulty of integrating heterogeneous data sources rather than the complexity of one comprehensive data model.

In this paper, we present a conceptual model on genetic studies and Pharmacogenetics using Unified Modeling Language (UML); the standard for object-oriented analysis and design [14]. Our conceptual model supports biomedical research activities that use both clinical and biomedical data to improve patient care by incorporating the roles of environment, life style and genetics. We developed class and package diagrams for these domains to clearly and intuitively show inter-object relationships at different levels of complexity. We used UML to represent conceptual models. Conceptual models can be the basis for diverse research activities for bioinformaticians.

One distinction between our research and previous studies is that our conceptual model covers more comprehensive and integrated data to support various genetic studies and pharmacogenetic research. Thus, the model is more useful for the requirements of complex and comprehensive real world projects. The rest of the paper is organized as follows: Section 2 reviews the related work. Section 3 presents our conceptual models. Section 4 concludes our paper.

2 Related Work

In the literature, many publications address the topic of conceptual modeling of bioinformatic data and processes. Papers range from representations of single type of data like genome sequences [16] [17], protein structures [6] [18], protein interactions [3] [16], and metabolic pathways [4] [10] [21], to complex applications such as genetic study and Pharmacogenetics that often cover experimental data and knowledge management system.

Recently, conceptual modeling of genome sequence has received a lot of attention. Approaches include object-oriented models [16] [22], semi-structured models [5], relational [1], and extended ER models [17, 18]. Ram and Wei [18] create their own notations such as Sequences, Sequential Aggregate, Subsequences and Fragment to express semantics of biologic sequences. The work by Paton *et al.* [16] is most similar to ours in terms of approach. Both use UML notation to address representation of biologic data and process. Their conceptual model covers genomic data, transcriptome data, and protein interactions. Their work, however, does not include Pharmacogenetics. Thus, our work is complementary to the work by Paton *et al.*

Metabolic pathways are an important source to discover the gene functionality [10]. Usually it is helpful for provision of gene candidates for genetic studies. How to represent, store, compare and mine metabolic pathway data is also a hot topic for the research community. Schreiber [21] represents pathways as a directed bipartite graph and develops an algorithm to compute a better visual comparison of metabolic pathways. Like Schreiber, Heymans and Singh [7] use graphs to represent enzymes involved in metabolic pathways and, further, offer an iterative approach to score the similarity of metabolic pathways among multiple organisms.

With many fundamental biologic databases available, researchers in bioinformatics turn to apply established databases to address complex real world problems such as cancer prevention, diagnosis and treatment. Usually those research activities involve complex experimental data and multiple-source knowledge systems. One interesting work is the ontology development for a Pharmacogenetics knowledge base (PharmGKB) [24]. The research group of PharmGKB implemented and compared the ontological and relational approach to genomic sequences for Pharmacogenetics [20]. Further, they used

a frame-based system to represent ontology for experimental data and domain knowledge [15]. Their work on modeling of pharmacogenetic data is similar to ours. We borrow some ideas from them and use a different approach, UML, to model data for Pharmacogenetics as well as genetic studies.

3 Systems and Methods

This section presents the information models using the package diagram and class diagram notation of the Unified Modeling Language (UML). Due to the high complexity of Information models for genetic studies and Pharmacogenetics, all classes are not shown in one big diagram, but are split into subsystems each of which is represented by a package. A package in UML is a mechanism that groups semantically related items. A subsystem, an independent part of the system, is usually comprised of classes loosely coupled and very cohesive.

In UML class diagrams, classes are drawn in rectangles with the class name at the top and optionally with attributes and operations listed below. In this paper, we only list attributes important to understanding the model for the sake of space saving. Besides, inter-class relationships including generalization, aggregation, association and realization and its cardinality are shown in the diagram. The remainder of this section is organized as follows: Section 3.1 shows the overview of the whole system as a hierarchical package diagram; Sections 3.1-3.6 show the class diagram of each package.

3.1 A System Model

The system model shown in Figure 1 presents all subsystems in the form of a package diagram. A package is rendered by a tabbed folder and inter-relationships between two packages are represented by a dotted arrow. Because each subsystem depends on others, only heavy dependencies are shown in the diagram.

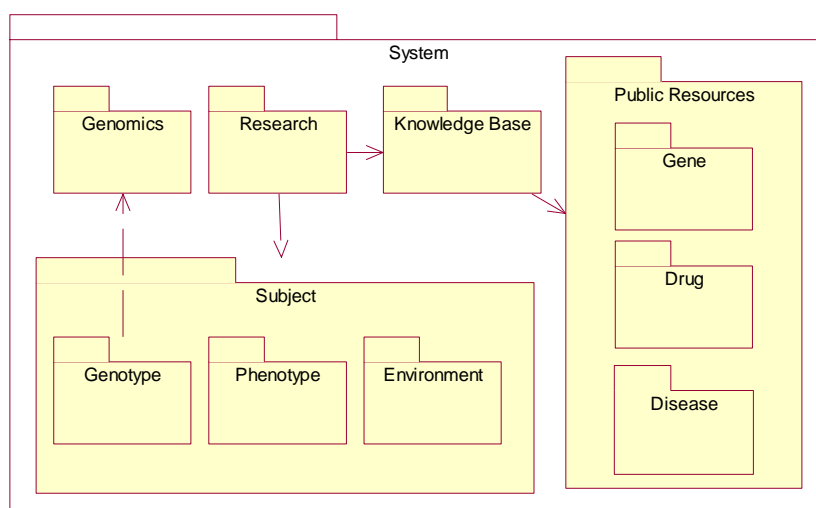


Fig. 1. A System Model of Data for Genetic Studies and Pharmacogenetics

Subject package contains classes related to experimental data of subjects including phenotype, genotype and environment. Models for phenotype, genotype and environment are still quite complex;

thus, they are separated into *Phenotype* subsystem, *Genotype* subsystem, and *Environment* subsystem, respectively.

Research package includes research dataset preparation, research approaches, and results. Dataset preparation heavily depends on the subject package. A dotted arrow from *Research* package to *Subject* package shows the dependency.

Knowledge base serves as an important supporting tool for genetic studies and pharmacogenetic research. First, it provides controlled vocabulary and alternative names of objects in the system, which allows us to query, retrieve, merge and analyze data more efficiently and effectively. Second, annotated knowledge about relationships among genes, diseases and drugs allows us to interpret evidence of intermediate or final research results.

Genomics and *Public Resources* are two other important packages. However, the details of these two packages are not shown in the paper due to the limit of space.

3.2 A Subject Model

The genotype is defined as all or part of the genetic constitution of an individual or a group [25]. Environment includes subjects' lifestyle, living environment, and demographic information. The phenotype represents visible properties of an organism produced by interaction of the genotype and environment [25].

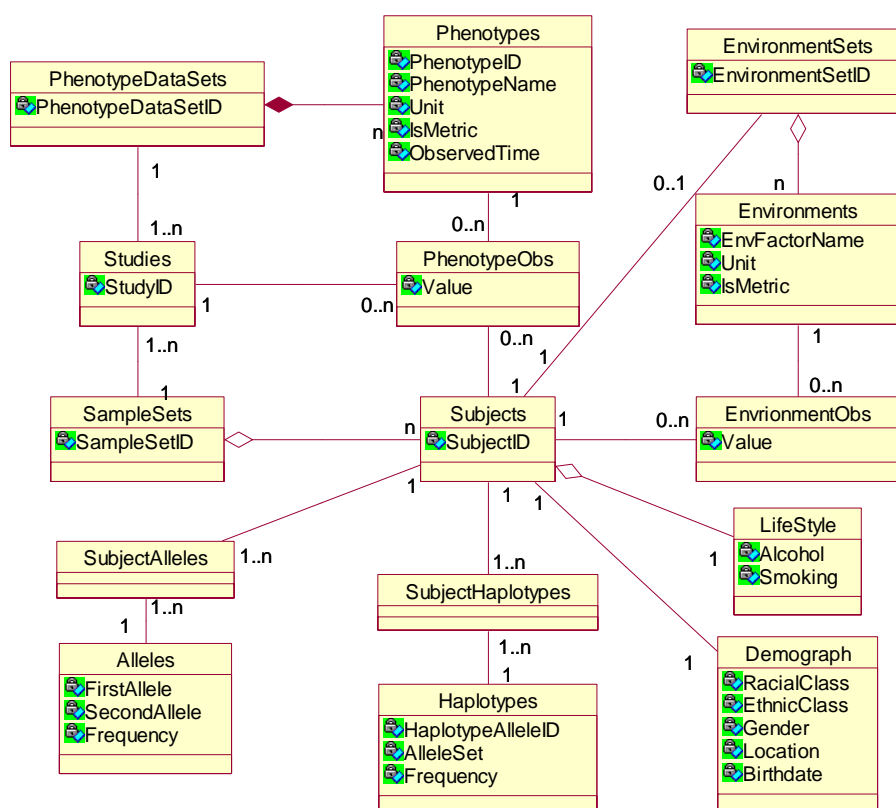


Fig. 2. A Subject Information Model shows the triad of relationship among genotype, phenotype and environment.

The genotype here is denoted by a subset of polymorphisms of interest. SNP is a site in the DNA where different chromosomes differ in their base. Haplotype [27] is a set of Single Nucleotide

Polymorphisms (SNPs) of interest in a DNA sequence block. The motivation of the Haplotype for research is explained in the section on the Genotype Model. A subject is associated with multiple alleles of polymorphisms and with multiple haplotype alleles.

Phenotypes of interest depend on the specific clinic trial or research experiment. For the sake of flexibility, each clinical trial or research is associated with an aggregation of phenotypes. Since observed phenotypes could vary across research experiments, an object called Trial Observation is created to store the observation of the subject phenotype in a specific clinical trial or research.

Environment not only describes the general sense of living environment factors, but also includes lifestyle and demographic information of the subject. Similar to phenotypes, living environment factors of interest are specific to a clinic trial or a research experiment. Distinct from phenotypes, although living environment factors of interest may be different for subjects belonging to different samples, they are fixed for a subject across trials or research. Therefore, an aggregation of environment factors is created to associate with a subject rather than a trial. Attributes of demographics of interest, such as ethnic or racial class, almost reach consensus. Lifestyle attributes, such as smoking or not, are in the same case. Each subject is simply associated with one demographic and one lifestyle object.

3.3 A Research Model

Availability of map, both physical and genetic, provided the infrastructure to boost studies linking phenotype to the interaction of the genotype and environment [11] [12]. Two particular study areas of interest are genetic studies and Pharmacogenetics.

Genetic studies use genetic markers including SNPs and Short Tandem Repeats (STRs) as tools to identify the genes responsible for disease susceptibility [28]. Researchers typically utilize one of two approaches, linkage and association studies. Linkage studies typically use families with multiple affected individuals, ideally including three or more generations to identify genetic regions more likely inherited with a disease or biologic response than expected by random chance [28].

Linkage studies are useful to indicate single-gene disorders, but not suitable to identify genes involved in polygenetic disorders. Polygenetic disorders are more challenging because the multiple disease genes tend to diminish statistical significance of a linkage to any one gene. Conversely, association studies that compare genetic differences in case and control samples provide more statistical power to identify disease susceptibility genes [28]. Modern association studies can be categorized into three groups: candidate-gene-based association studies, candidate-region-based association studies, and whole genome association studies.

Pharmacogenetics is a discipline that seeks to describe how inherited differences in genetic sequences among people influence their response to drugs [13, 19]. Pharmacogenetics helps determine why some medicines work better for some people than others and why some people are more likely to experience serious side effects. Knowledge that scientists gain from this research results in the delivery of safer, more effective medicines.

For any genetic study or pharmacogenetic research, a research data set associated with a sample and a data structure is required. For both association studies and pharmacogenetic research, samples consist of both case and control subjects. For linkage studies, family-based samples include only individuals with diseases.

A data structure is defined as a variable set necessary for the research. The variable set includes three types of variables; genotypes (polymorphisms or haplotypes), environment, and phenotypes. Usually

phenotypes are treated as dependent variables, while genotypes and environmental factors are as independent variables. Polymorphisms here usually refer to SNPs as they are the most powerful genetic markers thus far. A haplotype is a set of ordered SNP alleles in a region of a chromosome. The motivation for introduction of haplotypes will be explained in Section 3.4. Environment factor here has a broad sense; it refers to subjects' demographic information, lifestyle and living environment factors. A data structure is associated with either one genetic disease or one or multiple drugs. In some cases, researchers test the response of combined drugs on subjects.

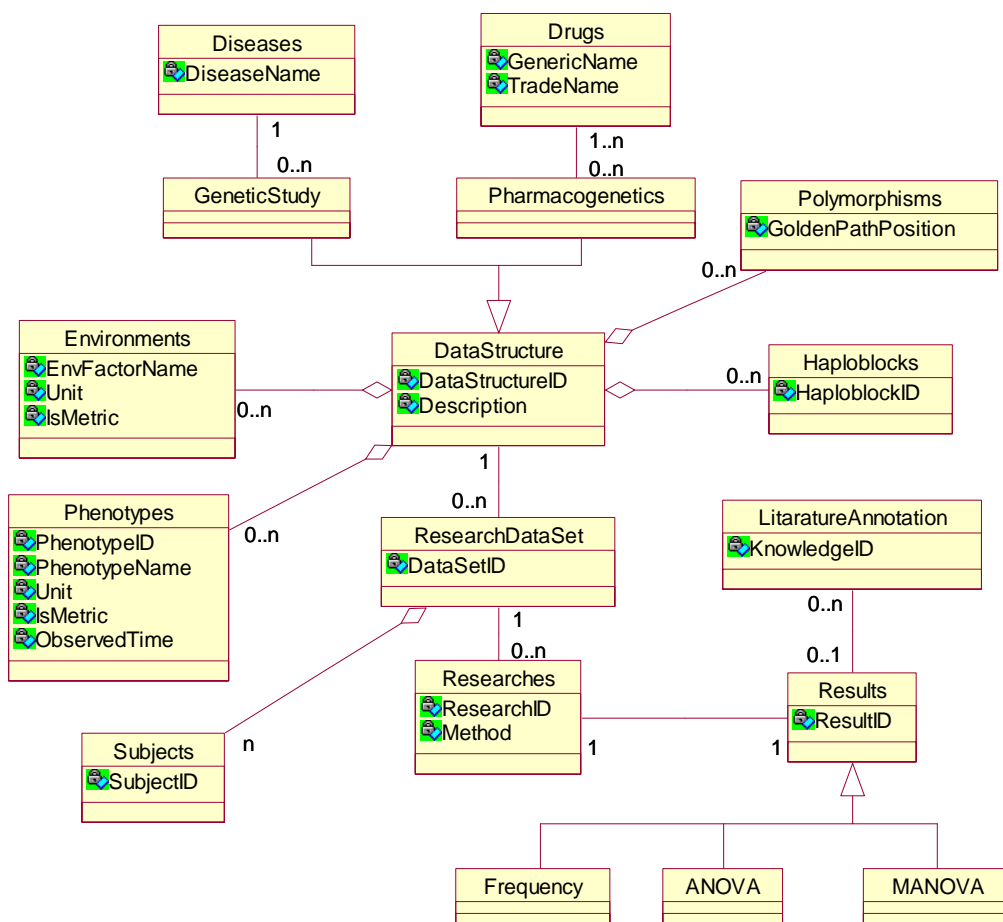


Fig. 3. A Research Model presents the data items for genetic studies and Pharmacogenetics and the intermediate or final research results.

A research data set can be used by many research experiments which apply a specific method such as frequency, ANOVA (Analysis of Variance), and MANOVA (Multivariate Analysis of Variance). A research will yield a result; format of research results depends on the statistical approaches applied. Some significant findings might be published and stored in the knowledge base. Significant findings can be roughly categorized into two groups. One is binary relationships between phenotypes and genotypes. For instance, a gene is associated with or without a disease; a set of SNPs is associated with or without a drug response. The other is to the degree to which the genotype affects the phenotype. For instance, different alleles might influence disease.

3.4 A Genotype Model

Herein, an individual's genotype is defined as a set of polymorphisms of interest. A polymorphism is a locus in a reference sequence with a variable length, where the sequence content might be different against individuals and each of possible sequence content is called an allele. Thus, a polymorphism has at least two alleles. The main types of polymorphisms include SNPs, STRs, Insertions and Deletions. SNPs and STRs have played significant roles in genetic research as powerful genetic markers due to their high densities in a genomic sequence.

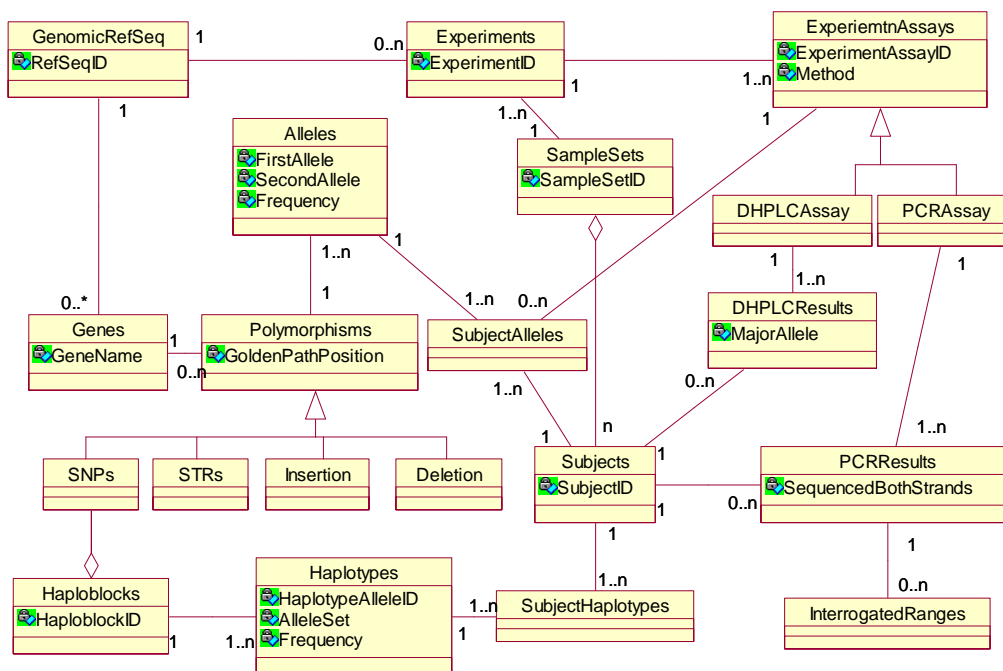


Fig. 4. A Model of Genotype and Experiment Essays shows the genotype of each subject that is observed from a variety of experiment assays.

The number of SNPs in humans is known to be between 10 and 30 million SNPs. To find the regions with genes that contribute to a disease, the frequencies of many SNP alleles are compared in individuals with and without a disease. When a particular region has an SNP allele that is more frequent in individuals with a disease than those without the disease, those SNPs and their alleles are hypothesized to be associated with the disease.

A haplotype is the set of SNP alleles along a region of a chromosome. Theoretically, there could be many haplotypes in a chromosome region. But some recent studies found that a haplotype occurs in a block pattern: the chromosome region of a block has just a few common haplotypes, followed by another block region also with just a few common haplotypes [27]. The recent studies also show that the common haplotypes are found in all populations studied, and that the population-specific haplotypes are generally rare [27].

The cost of genotyping is currently too high for whole-genome association studies. If a region has only a few haplotypes, then only a few SNPs need to be typed to determine which haplotype a chromosome has and whether the region is associated with a disease. That is the reason haplotypes are introduced for genetic studies and other genotype related studies.

To detect the alleles of a set of polymorphisms on a certain subject, we need experiments. Usually a sample set is assigned to a group of experiments, each of which is to query a set of polymorphisms within a reference sequence. Each experiment contains a group of assays. The purpose of an assay is to detect a particular polymorphism. Different types of assays are designed to detect different types of polymorphisms. For instance, a genotyping assay can detect the allele of an SNP, and a Polymerase Chain Reaction (PCR) sizing assay can detect the allele of an STR. Though there are at least eight types of assays available, most of them just determine the allele of a polymorphism for a subject and do not yield additional information. For this reason, only an attribute named method indicates the type of the assay. Denaturing High Performance Liquid Chromatography (DHPLC) Assay and PCR Assay produce additional information so that child classes are created for them. DHPLC Assay is associated with DPHLC result that additionally analyzes whether the allele is major or not.

3.5 A Phenotype Model

The phenotype represents visible properties of an organism produced by interaction of the genotype and environment. In this paper, phenotypes refer to disease symptoms and drug responses. Phenotypes of subjects are mainly collected either from clinical observations or laboratory tests with the permission of patients. In some cases, patients voluntarily submit their clinical profile for a research purpose.

Standardization of phenotypes across studies is important to the success of systems. The standardization brings at least two advantages. First, standardized measures will reduce learning cost and misunderstanding among researchers. As soon as a submission is accepted by the system, the data set will be used by various research groups in the world. Second, it makes the merge of separately small size sample sets from different research groups possible. In order to obtain stable statistical findings, most research experiments have limitations on sample sizes. For example, a candidate-gene-based association study might involve more than 1000 samples including cases and controls [28]. However, it is extremely costly to collect samples for various reasons and most separate sample sets contain much less than 1000 samples. If all the measures in different sample sets are standardized, the problem can be solved easily by ad hoc merging for a specific study.

In general, the standardization of phenotype measures involves the following three problems. The first problem is the unit inconsistency. Most phenotypes measures are metric. So an attribute called Unit is to identify the unit of a measure.

The second problem is the term inconsistency. It is very common that a term has multiple alternative names in the areas like medicines, biology and bioinformatics. In our models, three methods are used to relieve this problem. A normalized phenotype and a hierarchically organized attribute are employed to modify phenotypes. Besides, in the model of knowledge base systems, which we will explain in Section 3.6, all synonyms for a term are listed.

The third problem is the dimensional limitation. Now the system can only accept two-dimensional data. One dimension is a subject or a case. The other is a phenotype measure (or column). However, one phenotype is often measured several times on a subject at different time points for a real world clinical study. In our model, a phenotype name might include observing time information, "Blood Pressure Day 1", "Blood Pressure Day 10", for example. And a phenotype is associated with a normalized phenotype which does not include time dimensional information. We also set an attribute to store time dimensional information for each phenotype measure, which enables at least semi-automatic phenotype measure comparisons across sample sets. For the flexibility, we assign a new phenotype ID to a phenotype

submitted by any research groups because we are not sure it is identical to any extant phenotypes in the system the moment the data is submitted.

For any clinical studies or research experiments, more than one phenotype will be observed or tested. Thus each study is associated with one phenotype dataset. And the same phenotype dataset might be used for different studies conducted by the same research group. For the reason stated in the previous paragraph, a phenotype can belong to only one phenotype dataset. A phenotype dataset may involve multiple genes, diseases, and drugs (for Pharmacogenetics only).

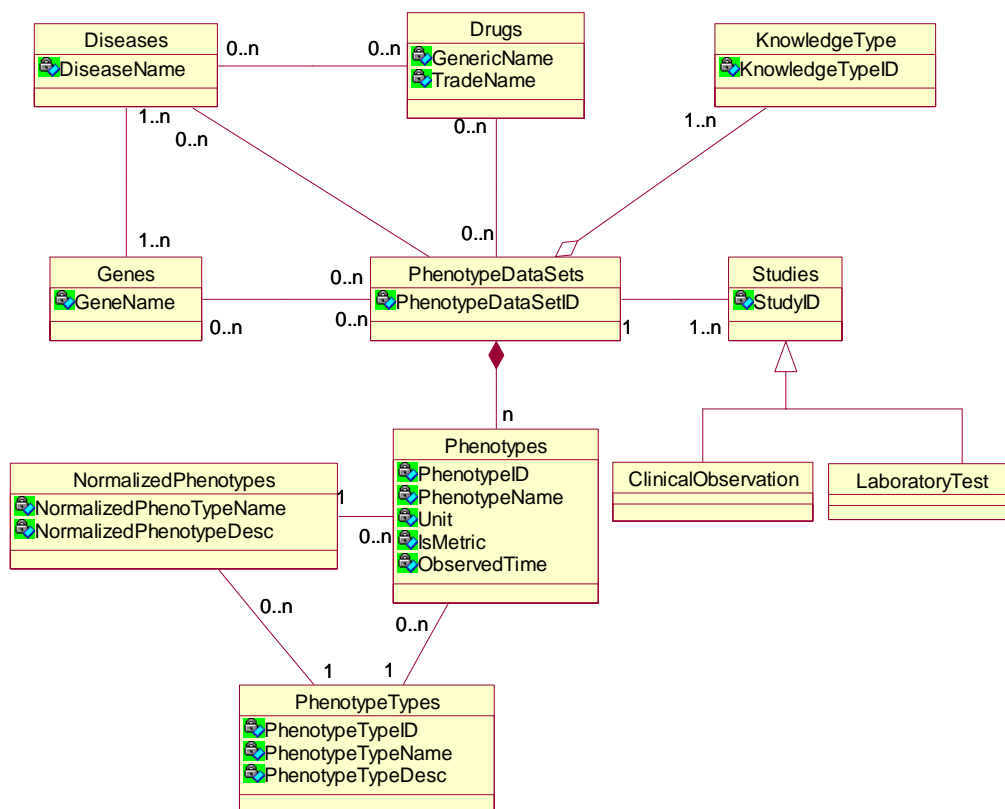


Fig. 5. A Phenotype Model addresses the representations of phenotypes in clinical trial or other experiments.

A phenotype dataset is always designed on purposes. Like PharmGKB [24], we represent the purpose by a set of knowledge types. In PharmGKB project, knowledge is categorized into five groups: clinical outcome, pharmacodynamics and drug responses, pharmacokinetics, molecular & cellular functional assays, and genotypes.

3.6 Knowledge Representation Model

The knowledge base contains only domain knowledge. The experimental data, for example genetic assays or tests, medicine response tests, and clinic studies, are not represented in the model. The knowledge system provides the following four types of knowledge [15]: controlled vocabulary terms, alternative names, accession numbers, and literature annotation.

Controlled Vocabulary Terms. The submitted experimental data must use the name of an object defined in the knowledge base. This convention ensures the consistency of concepts through the system.

The controlled vocabulary terms include drugs, diseases, genes, normalized phenotypes, environment, polymorphisms, and alleles. They are all inherited from the common superclass *Object*.

Alternative Names. Objects may have many synonyms. Maintenance of alternative names in the system assists a user in searching for a concept.

Accession Numbers. Accession numbers are unique identifiers for entities in external databases. Accession numbers are stored in object XRef (cross references) to facilitate communications with those databases. An object can have more than one cross reference which is associated with one standard resource such as dbSNP [33], dbSTS [34], GenBank [29], MedLine [35], or OMIM [32].

Literature Annotation. Genetic studies and Pharmacogenetics explore relationships among genes, diseases, and drugs. One relationship can involve multiple genes, diseases, and drugs. The instance of such a relationship is first of all normalized by an aggregation of knowledge types, and then annotated by a publication which contains the findings with respect to the instance of the relationship. From the predefined types the knowledge is associated with, a user may have ideas like what the knowledge is about. By further reading the abstract or full text of the publication, a user knows the details.

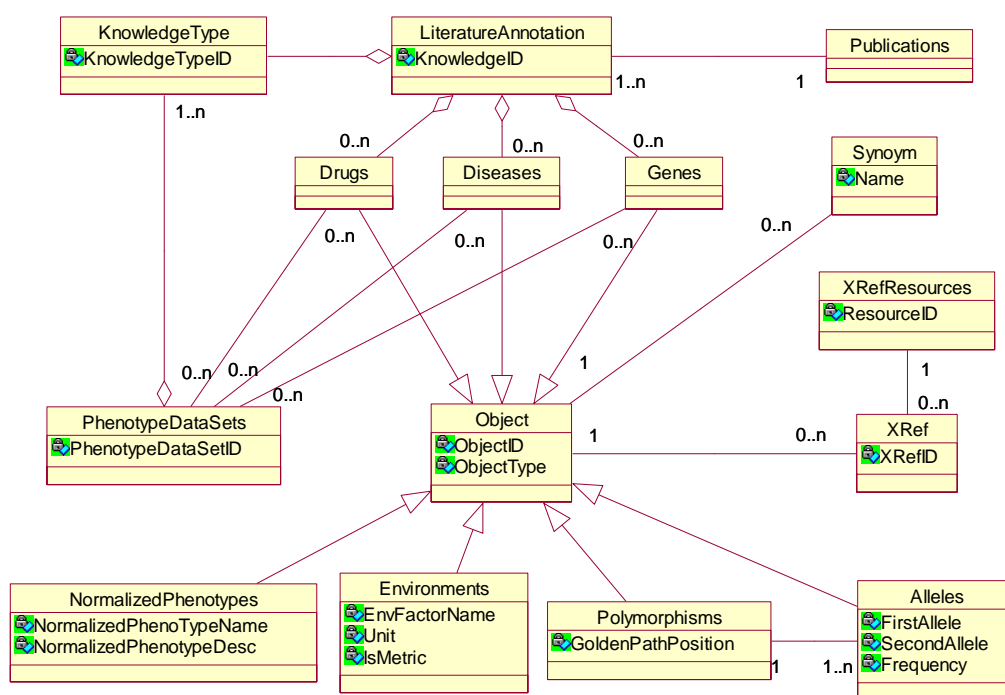


Fig. 6. A Model for Knowledge Representation provides the ontological information of drugs, diseases and genes that is evidenced by scientific literature.

Though the domain knowledge base system has no direct contributions to genetic studies and pharmacogenetic research, it is a very important supporting tool. First, the literature annotations serve as interpretations or evidences for intermediate or final findings produced by statistical approaches. Second, it is important to the reparation of research data.

4 Conclusions and Discussions

In this paper we presented a conceptual model for genetic studies and Pharmacogenetics. Comparing previous studies in conceptual modeling of bioinformatics data and process, our model provides a comprehensive view of integrated data for genetic studies and Pharmacogenetics by incorporating genomics, experimental data, domain knowledge, research approaches, and interface data for other publicly available resources into one model.

We used the class diagram notation of UML to represent the conceptual model. Observing the notation of inter-class relationships such as generalization, realization, aggregation, association, and their cardinalities, a user can intuitively understand the schema of the model. In order to show different levels of complexity of the model, we organized the whole system as a hierarchy of packages. By aggregating a set of low coupled and highly cohesive classes into a package, the whole system can be split to a set of subsystems. The hierarchical package diagram is helpful for users to easily understand the complex system. We believe that our conceptual models are comprehensive and easy to understand. They can be a basis in supporting various aspects of Pharmacogenetics and genetic studies.

Bioinformatics is a new field. Development of genome-based approaches to disease susceptibility and drug response is still in its infant stage. We can not come up with in advance all queries users will use and dataset researchers expect. Thus, it is inevitable to iteratively improve our model in practice after the system is implemented.

In the future, we plan to enhance our model as follows. First, we will enhance the model to capture recording of experimental approaches and procedures as well as recording experimental results. Second, we also plan to support longitudinal studies as well as cross-sectional studies. Experimental data like genotype, phenotype and living environment factors are organized around subjects. Though longitudinal data of a subject can be stored and accessed in the current model, a more extensive support may be desirable.

References

1. Berghozlz, A., Heymann, S., Schenk, J. and Freytag, J.: Sequence comparison using a relational database approach, *Proceedings of the International Database Engineering and Application Symposium (IDEAS)*, Montreal, August 1997.
2. Collins, F., Green, E., Guttacher, A., Guyer, S.: A vision for the future of genomics research, *Nature*, Vol. 422, 24 April 2003.
3. Eilbeck, K., Brass, A., Paton, N., and Hodgman, C.: INTERACT: an object oriented protein-protein interaction database, *Proceedings of Intelligent Systems in Molecular Biology (1999)* 87-94.
4. Ellis, L.B., Speedie, S.M., and McLeish, R.: Representing metabolic pathway information: an object-oriented approach, *Bioinformatics*, Vol. 14, (1998) 803-806.
5. Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., and Wildom, J.: The TSIMMIS approach to mediation: Data models and languages. *Proceedings of Second International Workshop on Next Generation Information Technologies and System*, June, (1995) 185-193.
6. Gray, P.M.D., Paton, N.W., Kemp, G.J.L., and Fothergill, J.E: An object-oriented database for protein structure analysis. *Protein Engineering*, 3, (1990) 235-243.
7. Heymans, M. and Singh, A.: Deriving phylogenetic trees from the similarity analysis of metabolic pathways, *Bioinformatics*, Vol. 19, Suppl. 1 (2003) i138-i146.
8. Hu, J., Mungall, C., Nicholson, D., and Archibald, A.: Design and implementation of a CORBA-based genome mapping system prototype, *Bioinformatics*, Vol.14 No.2 (1998) 112-120.

9. Jungfer, K., Rodriguez-Tome, and Mapplet, P.: a CORBA-based genome map viewer, *Bioinformatics*, Vol. 14, No. 8 (1998) 734-738.
10. Karp, P.: Representing, Analyzing, and Synthesizing Biochemical Pathways, *IEEE Expert*, April 1994.
11. Kell, D: Genotype—Phenotype mapping: genes as computer programs, *Trends in Genetics*, Vol. 18, No. 11 (2002).
12. Kim, J.: Computers are from Mars, Organisms are from Venus. *Computer*, July 2002, 25-32.
13. Krynetski, E.Y. and Evans, W.E.: Pharmacogenetics as a molecular basis for individualized drug therapy: the thiopurine S-methyltransferase paradigm, *Pharm. Res.*, 16(1999) 342-349.
14. Larman, C.: *Applying UML and Patterns*, 2nd edition, Prentice Hall (2001).
15. Oliver, D.E., Rubin, D.L., Stuart J.M., Hewett, M., Klein, T.E., and Altman, R.B.: Ontology Development for a Pharmacogenetics Knowledge Base, *Proceedings of Pacific Symposium on Biocomputing* (2002) 65-76.
16. Paton, N. et al: Conceptual Modeling of Genomic Information. *Bioinformatics*, Vol. 16 No. 6 (2000) 548-557.
17. Ram, S. and Wei, W.: Semantic Modeling of Biological Sequences, *Proceedings of the 13th Workshop on Information Technologies and Systems*, Seattle, (2003) 183-188.
18. Ram, S. and Wei, W.: Modeling the Semantics of Protein Structures, *Proceedings of the 23rd International Conference on Conceptual Modeling (ER 2004)*, Shanghai, China (2004).
19. Roses, A.D.: Pharmacogenetics and the practice of medicine, *Nature*, 2000 Jun 15, 405(6788):857-865.
20. Rubin, D. et al.: Representing genetic sequence data for Pharmacogenetics: an evolutionary approach using ontological and relational models, *Bioinformatics*, Vol. 18 Suppl. (2002) S207-S215.
21. Schreiber, F.: Comparison of Metabolic Pathways using Constraint Graph Drawing, *First Asia-Pacific Bioinformatics Conference (APBC2003)*, Adelaide, Australia (2003).
22. Wong, L.: Kleisli: a functional query system, *J. Functional Programming*, Vol.10, No.1 (2000) 19-56.
23. Wang, Z. and Moulton, J.: SNPs, Proteins Structures and Diseases, *Human Mutation*, Vol. 17 (2001) 263-270.
24. [PharmGKB] Pharmacogenetics Knowledge Base. <http://www.pharmgkb.org/>
25. Primer on Molecular Genetics, taken from the June 1992 DOE Human Genome 1991-92 Program Report.
26. Reference Sequences, <http://www.ncbi.nlm.nih.gov/RefSeq/>
27. Developing a Haplotype Map of the Human Genome for Finding Genes Related to Health and Disease. <http://www.genome.gov/10001665>.
28. White Paper: SNPs—Powerful Tools for Association Studies. August 2003, Applied Biosystems.
29. GenBank, <http://www.ncbi.nih.gov/Genbank/>.
30. EMBL Nucleotide Sequence Database, <http://www.ebi.ac.uk/embl/>
31. DNA Database of Japan (DDBJ), <http://www.ddbj.nig.ac.jp/>
32. Online Mendelian Inheritance in Man (OMIM), http://www.nslj-genetics.org/search_omim.html
33. dbSNP, <http://www.ncbi.nlm.nih.gov/projects/SNP/>
34. dbSTS, <http://www.ncbi.nlm.nih.gov/dbSTS/>
35. MedLine, <http://www.medline.com/>