Journal of Infection and Public Health (2016) 9, 331—338

http://www.elsevier.com/locate/jiph

Human Disease Insight: An integrated knowledge-based platform for disease-gene-drug information

Munazzah Tasleem, Romana Ishrat, Asimul Islam, Faizan Ahmad, Md. Imtaiyaz Hassan*

Centre for Interdisciplinary Research In Basic Sciences, Jamia Millia Islamia, Jamia Nagar, New Delhi 110025, India

Received 5 April 2015; received in revised form 21 May 2015; accepted 23 October 2015

KEYWORDS
Human disease database;
Human genome project;
Data integration;
Knowledge management system;
Relational database management system

Summary The scope of the Human Disease Insight (HDI) database is not limited to researchers or physicians as it also provides basic information to non-professionals and creates disease awareness, thereby reducing the chances of patient suffering due to ignorance. HDI is a knowledge-based resource providing information on human diseases to both scientists and the general public. Here, our mission is to provide a comprehensive human disease database containing most of the available useful information, with extensive cross-referencing. HDI is a knowledge management system that acts as a central hub to access information about human diseases and associated drugs and genes. In addition, HDI contains well-classified bioinformatics tools with helpful descriptions. These integrated bioinformatics tools enable researchers to annotate disease-specific genes and perform protein analysis, search for biomarkers and identify potential vaccine candidates. Eventually, these tools will facilitate the analysis of disease-associated data. The HDI provides two types of search capabilities and includes provisions for downloading, uploading and searching disease/gene/drug-related information. The logistical design of the HDI allows for regular updating. The database is designed to work best with Mozilla Firefox and Google Chrome and is freely accessible at http://humandiseaseinsight.com.

Introduction

Scientists have documented diseases within specific categories in various online databases. Due to advancements in science and technology, especially

* Corresponding author. Tel.: +91 11 2698 3409; fax: +91 11 2698 3409; mobile: +91 9990323217.
E-mail address: imtiyaz.hassan@gmail.com (Md.I. Hassan).

http://dx.doi.org/10.1016/j.jiph.2015.10.018

in genomics and information technology, we have entered in an exciting era of modern biology. The major challenge that the medical science community is presently facing is the integration of vast and rapidly growing amounts of information on various diseases into a holistic understanding. Recently, there has been considerable progress in disease genetics and genome-related medicine, leading to the generation of extensive data. The remarkable approach adopted by the Human Genome Project [1,2] making human genome, transcriptome and proteome data publicly available through online databases has facilitated in-depth investigations of disease genetics.

Currently, databases containing information about human diseases are focused predominantly on a particular disease category, such as all known Mendelian disorders described in the Online Mendelian Inheritance in Man (OMIM) [3]; infectious disease information, such as that found in the Infectious Disease Biomarker Database [4]; rare childhood diseases (http://www.madisonsfoundation.org/index.php); hereditary ocular diseases (https://disorders.eyes.arizona.edu); dermatological diseases (http://www.aocd.org/); and gastrointestinal diseases (http://www.gastro.net.au/). Other databases, such as the Malaysian National Cardiovascular Disease Database (NCVD), give an overview of cardiovascular disease and maintain records of patients suffering from these diseases [5]. In addition, the Indian Genetic Disease Database (IGDD) is a mutation data repository for genetic diseases in India; however, the information stored in the IGDD is helpful only to researchers [6]. Another database that provides information about autoimmune disorders is the Autoimmune Disease Database, which gives descriptions of autoimmune disorders and links these diseases to candidate genes, which is, again, a database that useful only for researchers [7]. The Comparative Toxicogenomic Database (CTD) is a rich resource for researchers to access information about the etiology of environmental diseases and explore chemical-gene and protein interactions [8]. Such attempts have contributed enormously to efforts related to the prevention, diagnosis and treatment of diseases and have resulted in the development of new approaches to alleviate the consequences of life-threatening illnesses. However, no disease database providing guidance related to bioinformatics tools and information available to members of the non-scientific public currently exists. Hence, the integration of information on all human diseases from different categories within a common place has become an important issue in the field of bioinformatics.

In recent years, the scientific community has been able to gain information through a number of useful internet-accessible resources, in addition to text books of biological and medical information. In the database development field, internet-accessible information retrieval systems have recently become popular due to the reduced costs of data storage and transfer. Vast amounts of biomedical information can be accessed through the World-Wide Web (WWW), although this information is scattered. Additionally, the heterogeneity and complexity of the available resources means that some information cannot be retrieved in a timely fashion. Furthermore, the rapidly growing fields related to disease information, genomics and proteomics databases and drug discovery and the expeditious development of computational tools to solve biological queries necessitate the integration of all of these information sources in a well-organized and concise database. Over a period of time, advancements in diagnostic evaluation and treatment have emerged. To provide the community with the most recent knowledge on human diseases and the discovery of genes involved in diseases, we have created a Knowledge Management System (KMS) that includes information on various categories of human diseases, the drugs used to cure these diseases, and the genes involved in causing these diseases, as well as bioinformatics tools for analyzing the genes in question. The HDI is therefore a comprehensive database of human diseases belonging to various categories that is cross-linked to other databases to retrieve detailed information on genes, drugs and tools. The HDI exhibits broad utility as it renders clinical information for physicians, genetic information and tool classification for researchers and disease descriptions for the general public. Thus, the HDI aims to provide a better understanding of human diseases, genes and drugs, as well as their relationships with one another, and allows easy retrieval of information through its user-friendly web-based applications.

Materials and methods

Overview of the database, HDI

The Human Disease Insight (HDI) database introduces an integrated knowledge base of diseases, genes and drugs and a list of bioinformatics tools with a user-friendly interface. The database was developed to allow simple retrieval of disease/gene/drug information and exploratory analysis of disease-specific genes within a single

Table 1 Summary of the raw data included in the HDI.

Sections	Entries	Numbers
Diseases	Diseases of the Circulatory System	24
	Diseases of the Digestive System	14
	Diseases of the Eye and Adenexa	41
	Diseases of the Musculoskeletal System	15
	Disease of the Genitourinary System	15
	Diseases of the Respiratory System	16
	Diseases of the Skin and Subcutaneous Tissues	08
	Disorders of the Perinatal Period	05
	Endocrine, Nutritional and Metabolic Diseases	216
	Hematological Disorders	59
	Infectious Diseases	145
	Malignant Neoplasms	67
Drugs	Anesthetics	16
	Anti-infective Medications	56
	Anti-allergics and Medicines used in Anaphylaxis	8
	Antidotes	14
	Anti-epileptic Drugs	7
	Anti-migraine Medicines	4
	Antineoplastic and Immunosuppressive	41
	Anti-parkinsonism Medicines	02
	Blood products and Plasma substitutes	10
	Cardiovascular Medicines	28
	Dermatological Medicines	17
	Diagnostic Agents	10
	Disinfectants and Antiseptics	13
	Diuretics	4
	Gastrointestinal Medicines	18
	Hormones and Endocrine Medicines	22
	Immunologics	13
	Medicines Acting on the Respiratory Tract	06
	Medicines Affecting Blood	10
	Muscle Relaxants and Cholinesterase Inhibitors	05
	NSAIDs and Antipyretics—Analgesics	13
	Ophthalmological Preparations	17
	Oxytocics and Antioxytocics	07
	Psychotherapeutic Medicines	11
	Solutions Correcting Water, Electrolyte and Acid-base	09
	Vitamins and Minerals	09
Genes		1440
Tools	DNA Sequence Analysis	33
	Genomics	05
	Modeling	09
	Protein Structure	35

location. It is designed to assemble, store, organize and display information about human diseases, genes associated with human diseases and drugs used to cure diseases in conjunction with a classified list of bioinformatics tools for sequence analysis and structural protein modeling. The HDI currently contains information on 625 human diseases, 320 drugs, 1440 genes and a classified list of bioinformatics tools (see Table 1). The diseases have been classified into 12 categories, and each category has been populated with disease information that includes synonym(s), pathogens, general disease descriptions, genes, clinical features, pathways, investigations, prevention, treatments, risk factors, prevalences and references [9—11] (http://www.nlm.nih.gov/medlineplus/, http://www.medscape.com/). Drugs have been classified into 26 broad categories [12]. The assignment of genes to human diseases includes links to the National Center for Biotechnology Information

(NCBI) [13] and the Universal Protein Resource (UniProt) [14] databases for detailed information. Bioinformatics tools are broadly classified into 4 main categories, and each category is then categorized into further sub-categories [15]. Each entry in the list of tools provides a brief description and an external link. Information collected for diseases, drugs and genes is interconnected in such a way that (i) through the disease option, multiple genes and/or multiple drugs involved in a particular disease can be retrieved; (ii) through the drugs option, the number of diseases for which a particular drug can be used is retrieved; and (iii) through the gene option, the number of disease(s) with which a particular gene is associated can be displayed. These pieces of information can be accessed freely. The information available in the HDI is derived from various resources, including electronic media, articles published in PubMed and text books and is curated and updated regularly.

Database structure

The HDI is a knowledge-based data warehouse that provides an integrated and curated repository of human diseases and related information. The classification of bioinformatics tools, together with descriptions of and links to their respective web pages, assists in performing research analyses of gene/protein sequences. The HDI provides a user-friendly web interface to allow the user to retrieve, download and upload information through interactive web forms. A schematic representation of the logistics used in the HDI is shown in Fig. 1

Software design and implementation

The HDI data warehouse was developed and is executed based on a three-tier architecture-user/client web-interface and a relational database management system (RDBMS) at the backend. The user/client can be a physician, researcher, student and/or member of the general public. The web interface is composed of web pages and web forms, designed in HTML5, CSS, PHP, JavaScript, ajax, jquery and MySql queries, to provide a common gateway interface. At the backend, we have created data marts to persistently and securely store information pertaining to human diseases. This database has been dynamically constructed, and web pages and web forms are interlinked with the data warehouse created at the backend for querying the database as instructed by the end user through button clicks and drop-down menus. The data warehouse created at the backend is a relational database managed by MySql and developed on the Windows operating system. For web services, the Apache HTTP web server is used. Data mining was performed to retrieve information on human diseases, genes, drugs and tools through various web resources and text books. The data thus obtained were then subjected to curation and uploaded to the database.

The framework for the HDI primarily consists of tables containing information on diseases, drugs and genes together with bioinformatics tools. Each disease category is populated with a number of diseases. Each entry in HDI provides comprehensive information about a human disease, characterized by its synonym(s), general description, pathogen, involved genes, clinical features, pathways, investigations, prevention, treatment, related drugs, prevalence, risk factors and associated references. Each drug category is populated with a number of drugs together with descriptions and links to drug databanks for detailed information. Genes involved in human diseases are collected, and related links to NCBI and UniProt are provided in a drop-down menu to retrieve detailed knowledge. For user convenience, major bioinformatics tools, together with their descriptions and links, are classified in an effort to provide guidance for performing specified analyses of genes/proteins. The HDI can be freely accessed from any web browser.

Data curation

Information is stored in the HDI after extensive literature, web and text-mining, followed by data curation. The HDI is being enhanced through continued efforts to improve disease knowledge and the interlinking of disease, drug and gene tables to obtain optimum information. The information is made available to the user after an extensive data-mining process. The knowledge thus obtained is managed in a relational database through cross-linking to fetch the data stored in the HDI data warehouse and via cross-linking of web resources (NCBI, UniProt and DrugBank). Genes related to human diseases are included in the database and are interlinked with disease tables to obtain the name(s) of disease(s) governed by a specific gene.

Knowledgebase access

HDI data can be retrieved efficiently through the drop-down menus and search functions provided on each page of the web site. The user can access alphabetically ordered diseases, drugs, genes and tools through a drop-down menu. Diseases from different categories can be selected through a drop-down menu. Clicking on ''Disease'' displays

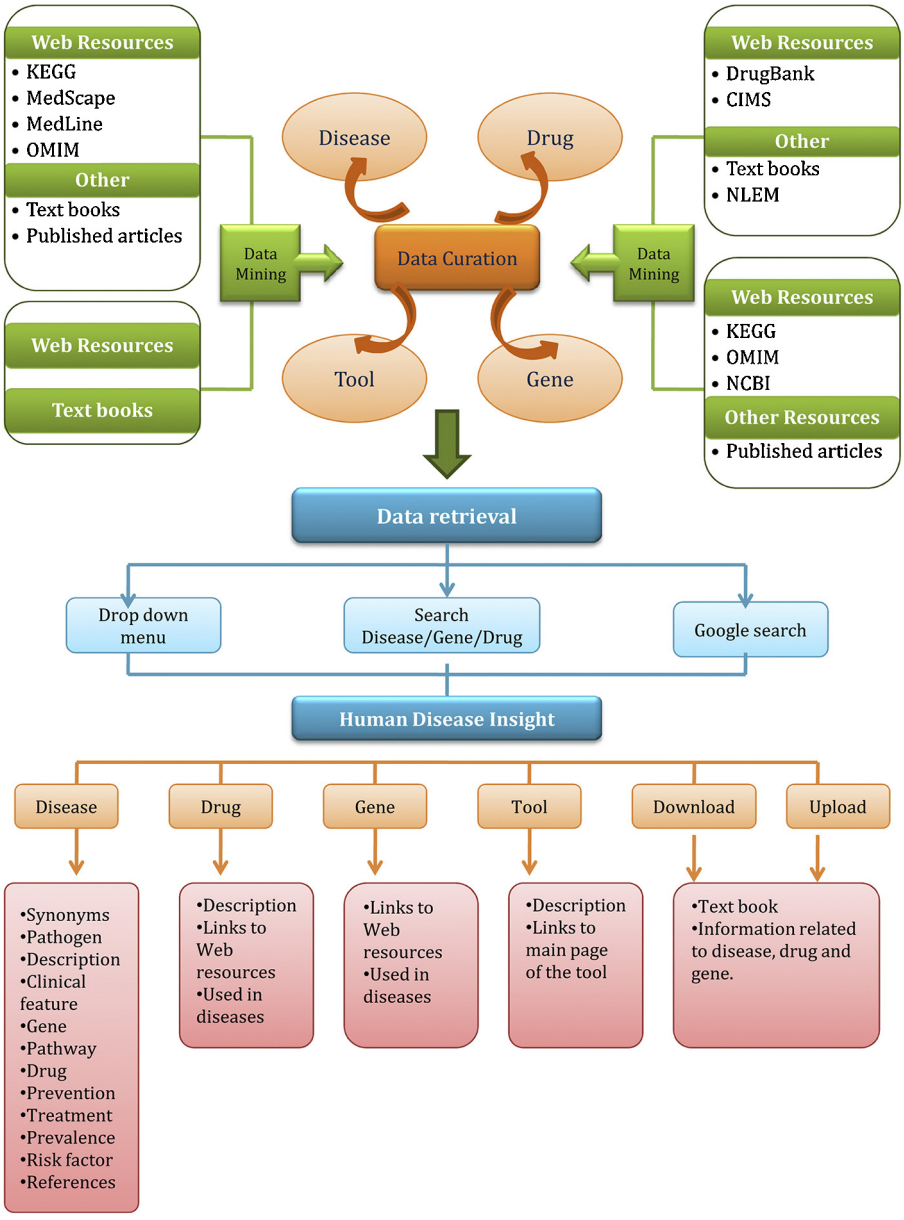

Figure 1 Schematic representation of data entry and retrieval in HDI.

some of the stored information about the disease, together with a ''More'' button. Clicking on the ''More'' button leads to detailed information about the selected disease. Similarly, drugs can be selected from different categories in the drop-down menu; clicking on any drug will provide its description, disease(s) that can be cured by the drug and links to a drug bank for details. In the same manner, a gene of interest can be selected from the drop-down menu, which will display a table with links to NCBI and UniProt to provide detailed information and the names of diseases caused by mutations in the gene. Two different search boxes are provided for user convenience. We have incorporated three search boxes on the home page that can be used to search the complete data mart for diseases, drugs and genes individually in the HDI data warehouse. To enhance the usability of these search boxes, codes were written to provide auto-complete search suggestions to the user, which will reduce search time and correct spelling errors. An example showing a working model of the search box is shown in Fig. 2 (Supplementary Figs. 1–5, describing the main web pages of the HDI). Another search included on each page is a Google search box that searches for the entered term in the database, as well as on the web. Our web site includes a provision for downloading and uploading published

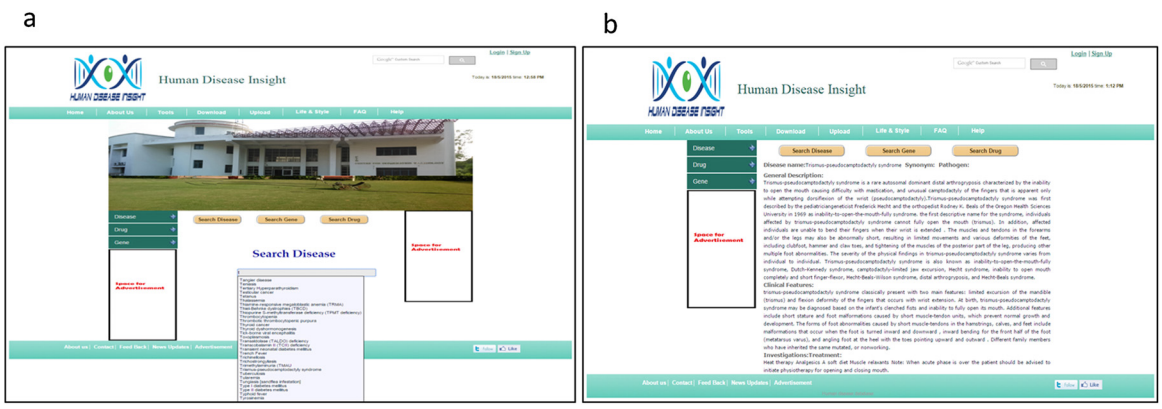

Figure 2 An example of using the HDI search box. (a) Disease search box showing suggestions for selecting the desired disease for retrieving information by typing only one letter. (b) Information retrieval page accessed after selecting the disease of interest.

articles, e-books and articles related to diseases, drugs and genes for registered users. All uploads from users will be updated in the database in a timely manner. For registration, a signup option is provided, and a registered user can login to download and upload related information. Medicinal and research-oriented news will be e-mailed to the provided by the user. A user feedback option is provided to improve the database. An advertisement option is provided for advertising companies to display their advertisements in the space of the web site after filling in the form. Furthermore, the database is connected to social networking sites to increase its popularity.

Results and discussion

Database availability

The database can be accessed free of charge to retrieve data on diseases, drugs and genes, as well as tool-related information. Free registration is required to download and upload related content.

Salient features of the HDI

The HDI is a robust knowledge management system that manages data-mined knowledge through cross-linking of data marts and web resources. This user-friendly, data-intensive repository provides the user with a platform to retrieve comprehensive disease-related information and perform gene/protein sequence-based analyses using direct links to classified bioinformatics tools. The HDI allows users to upload content to improve the database.

Comparison with other available disease databases

The HDI is a unique human disease database because although there are a number of existing databases that provide information for a particular category of disease; however, these databases are useful only for clinicians or researchers. Some of the existing disease databases include PedBase, the Online Mendelian Inheritance in Man (OMIM) database, the Indian Genetic Disease Database, the Global Infectious Diseases and Epidemiology Network (GIDEON), the Office of Rare Diseases Research (ORDR) database, the Dermatologic Disease Database, NeuroDnet and CADgene. PedBase (http://www.pedbase.org/) is a pediatric disease database that provides information related to childhood diseases that is only for clinicians. The OMIM database is a comprehensive database that provides knowledge related to human genetic disorders [3]. The Indian Genetic Disease Database is a database that provides information on mutations in causal genes for genetic diseases that are common in India [6]. The Global Infectious Diseases and Epidemiology Network (GIDEON) presents relevant information about infectious diseases and their epidemiology [16]. The Office of Rare Diseases Research (ORDR) database is a database registry for rare diseases and disorders linked to biorepositories [17]. The Dermatologic Disease Database (http://www.aocd.org/?page=DiseaseDatabaseHome) provides information regarding skin diseases. NeuroDnet is a database that provides relevant information about signaling molecules, genes and proteins and their interactions for constructing neurodegenerative disease networks [18]. CADgene provides detailed information on genes related to coronary artery diseases and tools to construct gene

networks but does not provide any information about the disease caused by the gene of interest [19].

Multiple databases have been created to address problems related to specific disease categories, which highlights the importance of compiling information on all human diseases from various categories under a common platform for analysis. This issue has been resolved by the creation of the HDI, which contains relevant information about human diseases from various categories, along with descriptions and cross-references of the genes involved and the drugs used to treat them. This database contains a well-classified list of computational tools, along with descriptions of these tools and links to their sources. The distinct characteristic of the HDI is the inclusion of information on genes, drugs and tools to aid in the exploration of human diseases. Additionally, a word-suggesting search engine for searching diseases/genes and drugs is provided. The HDI allows users to download and upload relevant content. The data in the HDI are updated regularly. The purpose of developing the HDI was to provide a simple solution to allow physicians, researchers and non-professionals to extract information on human diseases, genes, drugs and bioinformatics. The extensively cross-referenced, unified information provided, together with the facility of downloading and uploading content in a user-friendly manner, make the HDI unique. Thus, the concept of HDI is different from that of other available disease databases.

Future directions

The HDI provides optimum information required for the diagnosis and treatment of various human diseases. There are currently 625 diseases, 1440 genes, 320 drugs and 82 tools in the HDI. The information available in specific fields is rapidly expanding. Our aim is to collect a complete dataset for human diseases, genes, drugs and tools and to generate an integrated platform that can be employed to identify genes causing human disease. We also aim to integrate various bioinformatics tools to annotate human disease-specific genes. In the future, the main challenge will be to maintain an up-to-date dataset with the growing number diseases, genes, drugs and bioinformatics tools.

Conclusion

The HDI offers a premier platform that addresses all aspects of diseases, including their history, symptoms, causes, epidemiology, treatment, precautions, etc. Moreover, all diseases have been linked with data regarding their pharmacology and genomics and proteomics, as well as many other relevant databases. The HDI will not only contribute to a greater understanding of disease and provide primary data for research but will also enable the identification of interactions among various diseases through comparisons conducted with various tools provided in our database. The information provided by the HDI will lay the foundation for further advances in disease diagnosis and aid in the design of novel approaches for diagnosing and treating diseases. We believe that with the enrichment of the database, users will be able to access information about all human diseases.

Availability: This database is available at: http://www.humandiseaseinsight.com.

Funding

The HDI is supported by grants from the University Grants Commission (UGC) [F. No. 40-201/2011(SR)].

Competing interests

None declared.

Ethical approval

Not required.

Acknowledgement

The authors thank the FKT-Center for Information Technology (Jamia Millia Islamia, New Delhi, India) for providing technical support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jiph.2015.10.018.

References

[1] Christodoulou J. The human genome project: opportunities, challenges and consequences for population screening. Southeast Asian J Trop Med Public Health 2003;34(Suppl. 3):234—8.

[2] Hood L, Rowen L. The Human Genome Project: big science transforms biology and medicine. Genome Med 2013;5(9):79.
[3] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledge-base of human genes and genetic disorders. Nucleic Acids Res 2005;33(Database issue):D514—7.
[4] Yang IS, Ryu C, Cho KJ, Kim JK, Ong SH, Mitchell WP, et al. IDBD: infectious disease biomarker database. Nucleic Acids Res 2008;36(Database issue):D455—60.
[5] Wan Azman Wan Ahmad RZ, Ismail O, Sinnadurai J, Rosman A, Piawe CS, Abidin IZ, Kui-Hian S. Malaysian National Cardiovascular Disease Database (NCVD) — Acute Coronary Syndrome (ACS) registry: How are we different? CVD Prevention and Control 2011;6:8.
[6] Pradhan S, Sengupta M, Dutta A, Bhattacharyya K, Bag SK, Dutta C, et al. Indian genetic disease database. Nucleic Acids Res 2011;39(Database issue):D933—8.
[7] Karopka T, Fluck J, Mevissen HT, Glass A. The Autoimmune Disease Database: a dynamically compiled literature-derived database. BMC Bioinformatics 2006;7:325.
[8] Mattingly CJ, Colby GT, Forrest JN, Boyer JL. The Comparative Toxicogenomics Database (CTD). Environ Health Perspect 2003;111(6):793—5.
[9] Kasper DL, Harrison TR. Harrison's principles of internal medicine. 16th ed. New York: McGraw-Hill, Medical Pub. Division; 2005.
[10] Davidson S, Innes JA. Davidson's essentials of medicine. Edinburgh/New York: Elsevier/Churchill Livingstone; 2009.
[11] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27—30.
[12] Tripathi KD. Essentials of Medical Pharmacology. Jaypee Brothers Medical Publishers; 2008.
[13] Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2007;35(Database issue):D26—31.
[14] Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 2006;34(Database issue):D187—91.
[15] Jenny Gu PEB. Structural Bioinformatics. 2nd ed. Wiley-Blackwell; 2009, February.
[16] Edberg SC. Global Infectious Diseases and Epidemiology Network (GIDEON): a world wide Web-based program for diagnosis and informatics in infectious diseases. Clin Infect Dis 2005;40(1):123—6.
[17] Rubinstein YR, Groft SC, Bartek R, Brown K, Christensen RA, Collier E, et al. Creating a global rare disease patient registry linked to a rare diseases biorepository database: Rare Disease-HUB (RD-HUB). Contemp Clin Trials 2010;31(5):394—404.
[18] Vasaikar SV, Padhi AK, Jayaram B, Gomes J. NeuroDNet — an open source platform for constructing and analyzing neurodegenerative disease networks. BMC Neurosci 2013;14:3.
[19] Liu H, Liu W, Liao Y, Cheng L, Liu Q, Ren X, et al. CADgene: a comprehensive database for coronary artery disease genes. Nucleic Acids Res 2011;39(Database issue):D991—6.