Construction of Viral Hepatitis Bilingual Bibliographic Database with Protein Text Mining and Information Integration Functions

Heng Chen*
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Yueyang Road 320, Shanghai 200031, China
chenheng@sibs.ac.cn

Yongjuan Zhang
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Yueyang Road 320,Shanghai 200031, China
zhangyj@sbs.ac.cn

Chunhong Lin
ShangTex Workers' College, Changshou Road 652, Shanghai 200060, China
linch@fzzd.sh.cn

Liwen Zhang
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Yueyang Road 320, Shanghai 200031, China
zhangliwen@sibs.ac.cn

Tao Chen
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Yueyang Road 320, Shanghai 200031, China
Chentao01@sibs.ac.cn

Abstract

With fast development of viral hepatitis research, a large number of the research achievements have been generated and scattered in various literatures. Information service providers are meeting the challenge of satisfying readers' needs for more efficient and intelligent retrieval. Data mining and information integration are basically the promising and effective ways which become more and more important. Our study describes how to build the viral hepatitis bibliographic database, how the viral hepatitis related protein information is mined from the viral hepatitis bibliographic database, and integrated with corresponding information in the Universal protein resource - the Uniprot database from EBI. With the help of Chinese and English bilingual protein control vocabulary built by ourselves, mining of the viral hepatitis related protein text in the bilingual bibliographic database is realized and integration with corresponding protein information in the Uniprot database is achieved. In a word, our paper describes the integration and mapping between Chinese-English bilingual bibliographic databases and the authoritative factual databases (the Uniprot database) through relevant text mining works. It would be useful for extension, utilization and mining of Chinese-English bilingual bibliographic resources, as well as cross lingual information retrieval, integration, and mining.

1 Introduction

At present, global mass information floods and affects all aspects of human life. As one of the most active research fields, life science generates countless achievements and datasets that scatter in various literatures every year. In life science field, viral hepatitis is a seriously infecting disease resulted from various hepatitis viruses. So, viral hepatitis is, arguably, one of the most intensely studied viruses in the history of biomedical research over the world. With fast development of viral hepatitis research, a large number of the research achievements have been generated and scattered in various literatures. Although most of them are accessible through databases and web sites, it is still a problem for readers to identify what they really need from enormous search results. So mining and information integration are essential to meet readers' needs for more efficient and intelligent retrieval. Different useful information resources can be further integrated after the information is filtered , digitized and mined, The integration of information resources could be chosen, organized and processed according to the needs of different readers or users so as to yield the new information resources and new knowledge formation. The integration of digital information resources includes: data integration, information integration, knowledge integration, in which knowledge integration is at the highest level of resource integration system, which is based on the inevitable requirement and result of data and information integration to a certain stage.

In: Proceedings of IJCAI Workshop on Semantic Machine Learning (SML 2017), Aug 19-25 2017, Melbourne, Australia.

Knowledge mining is a complex process of identifying effective, novel, potentially useful information and knowledge from the information database (Feng and Wang, 2008). Information integration allows users to get the most extensive information, while knowledge mining allows users to quickly find the knowledge they want from the infinite information ocean. The application of information integration and knowledge mining technology and the establishment of linked and integrated database knowledge service system will allow users to quickly and efficiently find the necessary information and knowledge (Zhang *et al*., 2010).

Nowadays, many professional databases have been developed to the era of data mining and integration, knowledge mining and discovery, and greatly focus on information integration and knowledge mining so as to realize link and integration between different type of database through the one-way or two-way mode, which makes the relevant different types of database connected into a interactive organic whole, and enriches the extension and expansion capabilities of the relevant database. Some successful works have been carried out, such as GOPubMed, which can automatically recognize concepts from user's search query to PubMed and display papers containing relevant terms (Doms and Schroeder, 2005), and Entrez, an integrated search system that enables access to multiple National Center for Biotechnology Information (NCBI) databases (Maglott *et al*., 2011). Similar works are also reported by Alexopoulou *et al* (2008), Chen *et al*. (2013), McGarry *et al*. (2006), Pasquier (2008), and Sahoo *et al*. (2007). Different useful information resources can be further integrated after this information is filtered, digitized and mined. The innovation of database design and construction makes users deeply experience the charm and potential of information integration and knowledge mining.

In summary, with the development of international scientific database, information integration and knowledge mining has become the mainstream and the trend of digital information resources processing and utilization. the semantic network is the environment of information integration, ontology is the core of semantic web construction and foundation. Construction of the professional domain ontology, based on the integration and mining of digital information resources will become the focus of information integration and knowledge mining research (Yan, 2008). Based on the analysis of domestic and foreign database information integration and knowledge mining theory and application, authors learning from advanced foreign information integration and knowledge mining technology explore the association and integration of the Chinese and English bilingual literature databases of viral hepatitis and the related scientific data databases at home and abroad in the innovation construction of the viral hepatitis special literature knowledge database, moreover, the authors further study the deep processing of the subject classification index of the literature in the knowledge database from the user's needs so as to facilitate the readers' use and retrieval.

As you know, literature database and protein science database are the ones of the most important support source for hepatitis virus researchers. So in this paper, we build the viral hepatitis bilingual bibliographic database and perform viral hepatitis related protein text mining and integrating with the Uniprot protein database so as to give our vigorous support for the sino-foreign hepatitis virus researchers' information retrieval and knowledge discovery.

2 Materials, Methods, Design and Results

2.1 Materials

Data resources: Medline database which is from NCBI for English dataset, CNKI database which is from China National Knowledge Infrastructure for Chinese dataset, and Uniprot protein database which is from EBI (European Bioinformatics Institute) for protein dataset.

Methods and procedure:

① Collect, select and process the viral hepatitis and hepatitis virus A, B, and C related dataset (literature data) from the above Chinese and English database;

②Build the bilingual text mining control vocabulary (dictionary);

③Perform text mining of viral hepatitis related proteins in the viral hepatitis bilingual literature database;

④ Perform preliminary research on eliminating the false positive ones from mining results;

⑤Integrate the viral hepatitis bilingual literature database with the Uniprot protein database on the basis of the mined hepatitis virus A, B and C related protein.

2.2 Design

System design

1. System architecture: 3-tier structure based on B/S model (separateness of web server and database server). See fig.1 as follows:

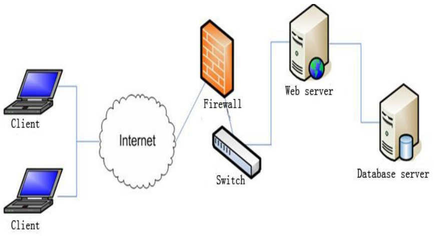


Figure 1 System architecture

2. System hardware platform: IBM 4 core servers
3. System software platform:
Operating system: Linux, Ubuntu 9.04
WEB server: Nginx 0.87
Database software: MySql 5.6.22
Development language: C++ for information index module and data mining module, and PHP for web application module.
4. Integration design architecture of database system platform. See fig.2 as follows:

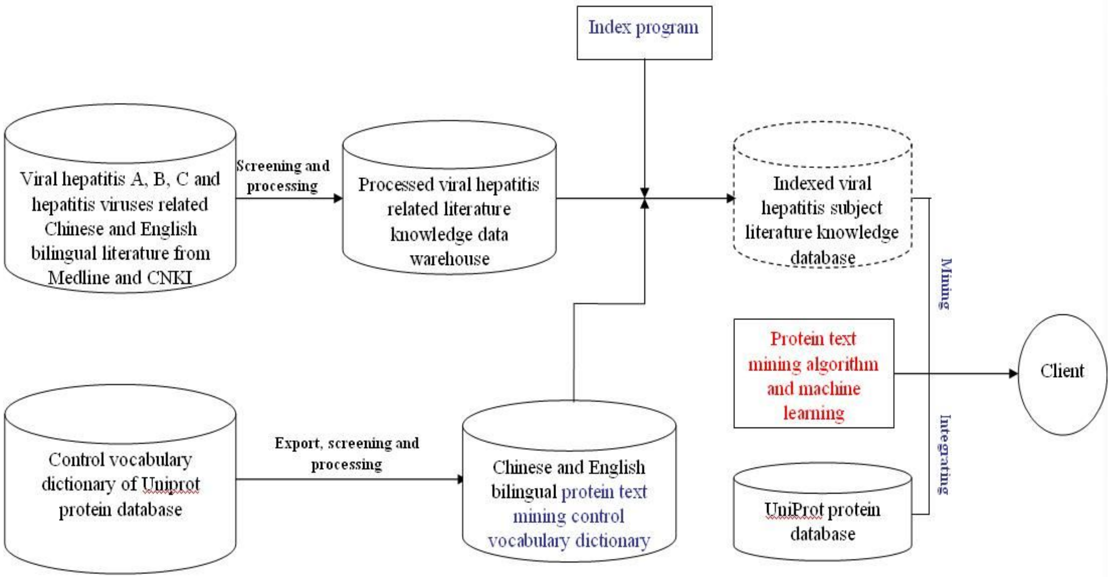

Index program
Viral hepatitis A, B, C and hepatitis viruses related Chinese and English bilingual literature from Medline and CNKI
Screening and processing
Processed viral hepatitis related literature knowledge data warehouse
Indexed viral hepatitis subject literature knowledge database
Mining
Protein text mining algorithm and machine learning
Client
Integrating
Control vocabulary dictionary of Uniprot protein database
Export, screening and processing
Chinese and English bilingual protein text mining control vocabulary dictionary
UniProt protein database

Figure 2 Database system platform structure

Figure 2 demonstration: On the one hand, literature records about viral hepatitis A, B and C from Medline database of Web of Science platform in English and from CNKI database of China in Chinese were screened, collected and processed into the viral hepatitis related literature knowledge data warehouse. On the other hand. The control vocabulary of Uniprot protein database from EBI was also screened, collected, processed and translated into the Chinese & English bilingual viral hepatitis related protein text mining control vocabulary. Then the indexed viral hepatitis subject literature knowledge database was built by index program including improved index procedure control and optimizing index algorithm through application of the protein text mining control vocabulary in the processed viral hepatitis related literature data warehouse. Finally, integration of the indexed viral hepatitis subject literature knowledge database and Uniprot protein database was realized by mapping ruler through protein text or knowledge mining algorithm and machine learning.

5. Viral hepatitis related literature indexing and processing. See fig.3 as follows:

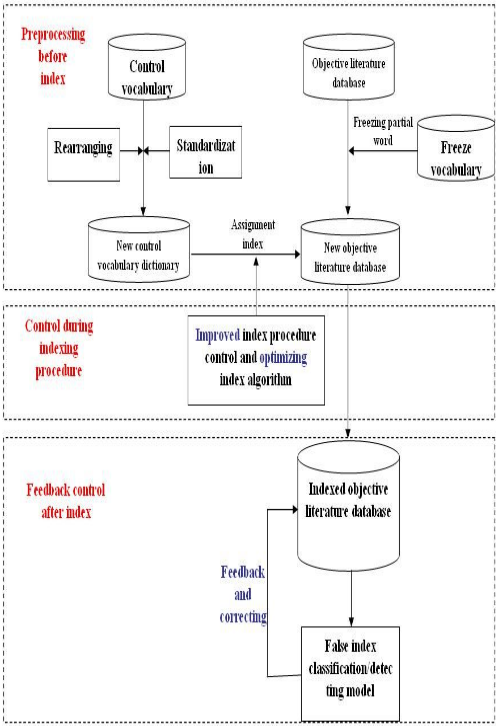

Preprocessing before index
Control vocabulary
Objective literature database
Freezing partial word
Freeze vocabulary
Rearranging
Standardization
New control vocabulary dictionary
Assignment index
New objective literature database
Control during indexing procedure
Improved index procedure control and optimizing index algorithm
Feedback control after index
Indexed objective literature database
Feedback and correcting
False index classification/detecting model

Figure 3 literature indexing and processing flow chart

Figure 3 demonstration: The literatures in the viral hepatitis knowledge data warehouse were indexed and processed according to three stages in the flow chart. Stage 1 is preprocessing before index. Stage 2 is control during indexing procedure. Stage 3 is feedback control after index. Aim of all three stages above is to protect protein text mining from false positive indexing and mining results.

6. Database system function module components:

① Information issue/management system

② Literature knowledge database processing/maintaining system

③ Administration system for user right and IP address

④ Information index system

⑤ Knowledge mining system

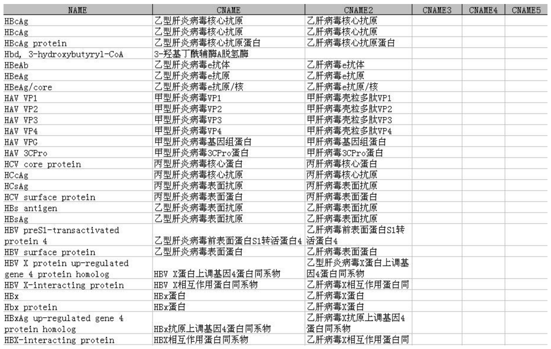
⑥ Knowledge inquiry system

⑦ Data maintaining system

⑧ Web site visiting and statistical system

Construction of Chinese English bilingual control vocabulary dictionary

Part exemplary diagram for the bilingual control vocabulary. See fig.4 as follows:



NAME	CNAME	CNAME2	CNAME3	CNAME4	CNAME5
HBcAg	乙型肝炎病毒核心抗原	乙肝病毒核心抗原			
HBcAg	乙型肝炎病毒核心抗原	乙肝病毒核心抗原			
HBcAg protein	乙型肝炎病毒核心抗原蛋白	乙肝病毒核心抗原蛋白			
Hbd, 3-hydroxybutyryl-CoA	3-羟基丁酰辅酶A脱氢酶				
HBeAb	乙型肝炎病毒e抗体	乙肝病毒e抗体			
HBeAg	乙型肝炎病毒e抗原	乙肝病毒e抗原			
HBeAg/core	乙型肝炎病毒e抗原/核	乙肝病毒e抗原/核			
HAV VP1	甲型肝炎病毒VP1	甲肝病毒壳粒多肽VP1			
HAV VP2	甲型肝炎病毒VP2	甲肝病毒壳粒多肽VP2			
HAV VP3	甲型肝炎病毒VP3	甲肝病毒壳粒多肽VP3			
HAV VP4	甲型肝炎病毒VP4	甲肝病毒壳粒多肽VP4			
HAV VPG	甲型肝炎病毒基因组蛋白	甲肝病毒基因组蛋白			
HAV 3CPro	甲型肝炎病毒3CPro蛋白	甲肝病毒3CPro蛋白			
HCV core protein	丙型肝炎病毒核心蛋白	丙肝病毒核心蛋白			
HCcAg	丙型肝炎病毒核心抗原	丙肝病毒核心抗原			
HCsAg	丙型肝炎病毒表面抗原	丙肝病毒表面抗原			
HCV surface protein	丙型肝炎病毒表面蛋白	丙肝病毒表面蛋白			
HBs antigen	乙型肝炎病毒表面抗原	乙肝病毒表面抗原			
HBsAg	乙型肝炎病毒表面抗原	乙肝病毒表面抗原			
HBV preS1-transactivated protein 4	乙型肝炎病毒前表面蛋白S1转活蛋白4	乙肝病毒前表面蛋白S1转活蛋白4			
HBV surface protein	乙型肝炎病毒表面蛋白	乙肝病毒表面蛋白			
HBV X protein up-regulated gene 4 protein homolog	HBV X蛋白上调基因4蛋白同系物	乙型肝炎病毒X蛋白上调基因4蛋白同系物			
HBV X-interacting protein	HBV X相互作用蛋白同系物	乙肝病毒X相互作用蛋白同			
HBx	HBx蛋白	乙肝病毒X蛋白			
Hbx protein	HBx蛋白	乙肝病毒X蛋白			
HBxAg up-regulated gene 4 protein homolog	HBx抗原上调基因4蛋白同系物	乙肝病毒X抗原上调基因4蛋白同系物			
HBX-interacting protein	HBX相互作用蛋白同系物	乙肝病毒X相互作用蛋白同			

Figure 4 Demonstration diagram of part exemplary for the bilingual control vocabulary of viral hepatitis (A, B, C) protein

Information integrating and hyperlinking regulation and examples for the mined protein text in literature using Chinese English bilingual control vocabulary

Using the HBV related protein text as example to demonstrate information integrating and hyperlinking regulation for the mined English protein text in literature. See as follows:

① HBeAg,
http://lifecenter.sgst.cn/protein/cn/quickSearch.do?entrezWord=HBeAg

② Capsid protein,

http://lifecenter.sgst.cn/protein/cn/quickSearch.do?entrezWord=Capsid%20protein

③ Large envelope protein,
http://lifecenter.sgst.cn/protein/cn/quickSearch.do?entrezWord=Large%20envelope%20protein

④ RNA-directed DNA polymerase
http://lifecenter.sgst.cn/protein/cn/quickSearch.do?entrezWord=RNA-irected%20DNA%20polymerase

While for the mined Chinese protein text in literature:

Translate the Chinese protein into English protein text in advance, such as "乙型肝炎 e 抗原" is translated into " HBeAg", "衣壳蛋白质" is translated into "Capsid protein " , then performing information integrating and hyperlinking according to regulations above and examples.

Main performance index of the database system:

1. The biggest record number for the literature information: 0.2 billion.

2. Index and data mining time:

at current condition of the database system containing one million four hundred and seventy thousand (1,470,000) control vocabularies and about twenty thousand (20,000) literature records, the index and data mining time is about eighteen minutes.

The index and data mining time is about five minutes after the single literature record is added.

3. The average retrieval time: < 0.03s (second)

4. The amount of concurrency (the number of users simultaneous access): >50 people

Viral hepatitis subject literature knowledge database extends three functions through data mining, information integration and hyperlinking

1. Obtain the protein sequence and annotation information

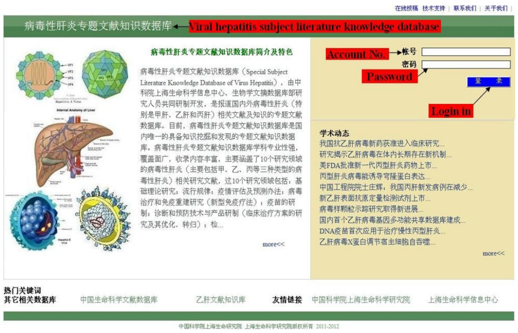
2. Perform homological analysis of the protein sequences (BLAST)

3. Perform different alignment of the protein sequences and evolutionary tree mapping

2.3 Results

Function realization and result display of the viral hepatitis subject literature knowledge database

Homepage of the viral hepatitis subject literature knowledge database. See fig.5 as follows:


在线投稿 技术支持 | 联系我们 | 关于我们 |
病毒性肝炎专题文献知识数据库
Viral hepatitis subject literature knowledge database
Account No.
帐号
密码
Password
登　录
Login in
more<<
more<<
热门关键词
其它相关数据库
中国生命科学文献数据库
乙肝文献知识库
友情链接
中国科学院上海生命科学研究院
上海生命科学信息中心
中国科学院上海生命研究院 上海生命科学研究院版权所有 2011-2012

病毒性肝炎专题文献知识数据库简介及特色

病毒性肝炎专题文献知识数据库（Special Subject Literature Knowledge Database of Virus Hepatitis），由中科院上海生命科学信息中心、生物学文摘数据库部研究人员共同研制开发，是报道国内外病毒性肝炎（特别是甲肝、乙肝和丙肝）相关文献及知识的专题文献数据库。目前，病毒性肝炎专题文献知识数据库是国内唯一的具备知识挖掘和发现的专题文献知识数据库。病毒性肝炎专题文献知识数据库学科专业性强，覆盖面广，收录内容丰富，主要涵盖了10个研究领域的病毒性肝炎（主要包括甲、乙、丙等三种类型的病毒性肝炎）相关研究文献，这10个研究领域包括：基础理论研究；流行规律；疫情评估及预测办法；病毒治疗和免疫重建研究（新型免疫疗法）；疫苗的研制；诊断和预防技术与产品研制（临床治疗方案的研究及其优化、转归）；检...

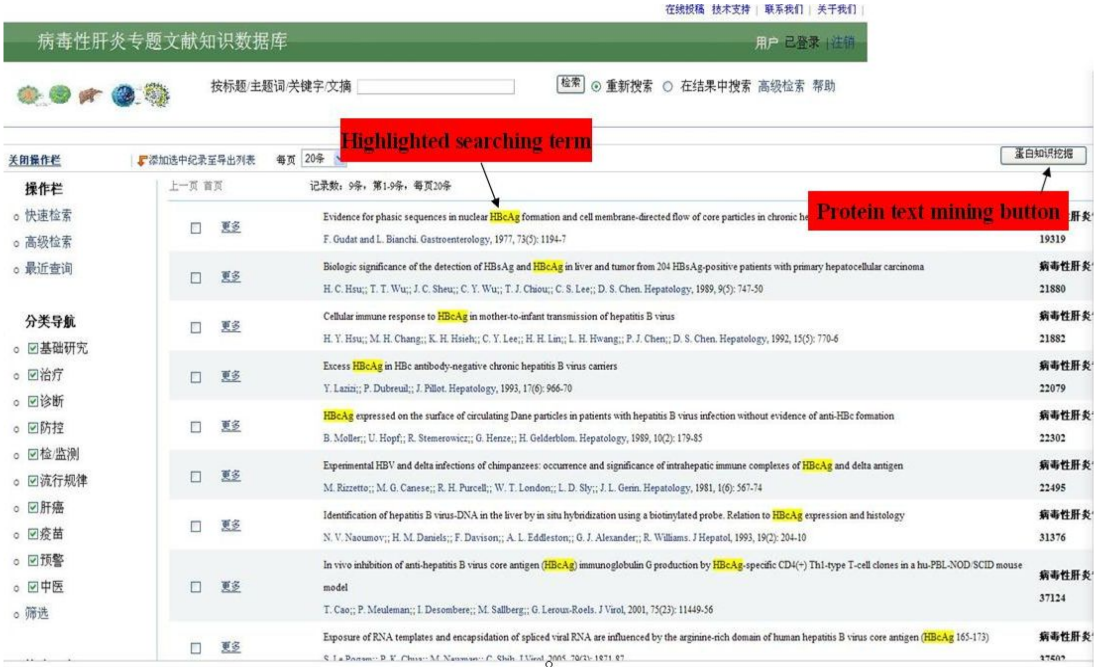
学术动态

我国抗乙肝病毒新药获准进入临床研究...
研究揭示乙肝病毒在体内长期存在新机制...
美FDA批准新一代丙型肝炎药物上市...
丙型肝炎病毒能诱导穹隆蛋白表达...
中国工程院院士庄辉：我国丙肝新发病例在减少...
新乙肝表面抗原定量检测试剂上市...
病毒样颗粒示踪研究取得新进展...
国内首个乙肝病毒基因多功能共享数据库建成...
DNA疫苗首次应用于治疗慢性丙型肝炎...
乙肝病毒X蛋白调节宿主细胞自吞噬...

Figure 5 Homepage of the viral hepatitis subject literature knowledge database

Realization of protein mining for the viral hepatitis literature knowledge database.

The viral hepatitis related proteins are successfully mined by using the bilingual control vocabulary, algorithm and computer program in the viral hepatitis bilingual bibliographic database. Moreover, the viral hepatitis bilingual bibliographic database is protein database through the protein mining and information integration. See the fig.6, 7, 8 as follows:



在线投稿 技术支持 | 联系我们 | 关于我们 |

病毒性肝炎专题文献知识数据库

用户 已登录 | 注销

按标题/主题词/关键字/文摘 [检索] ● 重新搜索 ○ 在结果中搜索 高级检索 帮助

Highlighted searching term

关闭操作栏 | 添加选中纪录至导出列表 每页 20条 [蛋白知识挖掘]

Protein text mining button

操作栏
- 快速检索
- 高级检索
- 最近查询

分类导航
- ☑基础研究
- ☑治疗
- ☑诊断
- ☑防控
- ☑检/监测
- ☑流行规律
- ☑肝癌
- ☑疫苗
- ☑预警
- ☑中医
- 筛选

上一页 首页 记录数：9条，第1-9条，每页20条

□ 更多 Evidence for phasic sequences in nuclear HBcAg formation and cell membrane-directed flow of core particles in chronic he
F. Gudat and L. Bianchi. Gastroenterology, 1977, 73(5): 1194-7 — 19319

□ 更多 Biologic significance of the detection of HBsAg and HBcAg in liver and tumor from 204 HBsAg-positive patients with primary hepatocellular carcinoma — 病毒性肝炎
H. C. Hsu;; T. T. Wu;; J. C. Sheu;; C. Y. Wu;; T. J. Chiou;; C. S. Lee;; D. S. Chen. Hepatology, 1989, 9(5): 747-50 — 21880

□ 更多 Cellular immune response to HBcAg in mother-to-infant transmission of hepatitis B virus — 病毒性肝炎
H. Y. Hsu;; M. H. Chang;; K. H. Hsieh;; C. Y. Lee;; H. H. Lin;; L. H. Hwang;; P. J. Chen;; D. S. Chen. Hepatology, 1992, 15(5): 770-6 — 21882

□ 更多 Excess HBcAg in HBc antibody-negative chronic hepatitis B virus carriers — 病毒性肝炎
Y. Lazizi;; P. Dubreuil;; J. Pillot. Hepatology, 1993, 17(6): 966-70 — 22079

□ 更多 HBcAg expressed on the surface of circulating Dane particles in patients with hepatitis B virus infection without evidence of anti-HBc formation — 病毒性肝炎
B. Moller;; U. Hopf;; R. Stemerowicz;; G. Henze;; H. Gelderblom. Hepatology, 1989, 10(2): 179-85 — 22302

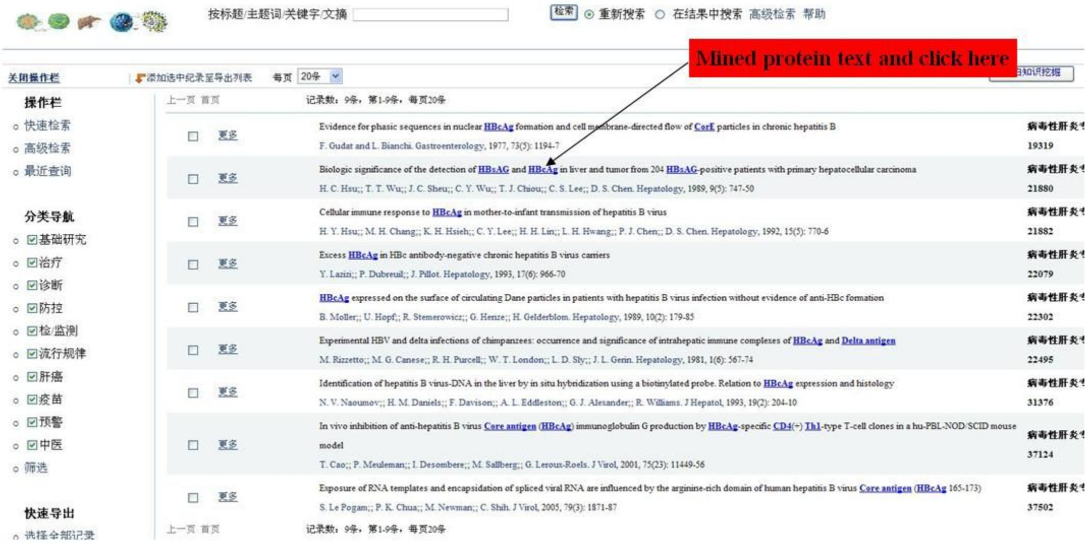
□ 更多 Experimental HBV and delta infections of chimpanzees: occurrence and significance of intrahepatic immune complexes of HBcAg and delta antigen — 病毒性肝炎
M. Rizzetto;; M. G. Canese;; R. H. Purcell;; W. T. London;; L. D. Sly;; J. L. Gerin. Hepatology, 1981, 1(6): 567-74 — 22495

□ 更多 Identification of hepatitis B virus-DNA in the liver by in situ hybridization using a biotinylated probe. Relation to HBcAg expression and histology — 病毒性肝炎
N. V. Naoumov;; H. M. Daniels;; F. Davison;; A. L. Eddleston;; G. J. Alexander;; R. Williams. J Hepatol, 1993, 19(2): 204-10 — 31376

□ 更多 In vivo inhibition of anti-hepatitis B virus core antigen (HBcAg) immunoglobulin G production by HBcAg-specific CD4(+) Th1-type T-cell clones in a hu-PBL-NOD/SCID mouse model — 病毒性肝炎
T. Cao;; P. Meuleman;; I. Desombere;; M. Sallberg;; G. Leroux-Roels. J Virol, 2001, 75(23): 11449-56 — 37124

□ 更多 Exposure of RNA templates and encapsidation of spliced viral RNA are influenced by the arginine-rich domain of human hepatitis B virus core antigen (HBcAg 165-173) — 病毒性肝炎
S. Le Pogam;; P. K. Chua;; M. Newman;; C. Shih. J Virol, 2005, 79(3): 1871-87 — 37502

Figure 6 Page of the hepatitis viral protein mining (1)



按标题/主题词/关键字/文摘 [检索] ● 重新搜索 ○ 在结果中搜索 高级检索 帮助

Mined protein text and click here

关闭操作栏 | 添加选中纪录至导出列表 每页 20条 [白知识挖掘]

操作栏
- 快速检索
- 高级检索
- 最近查询

分类导航
- ☑基础研究
- ☑治疗
- ☑诊断
- ☑防控
- ☑检/监测
- ☑流行规律
- ☑肝癌
- ☑疫苗
- ☑预警
- ☑中医
- 筛选

快速导出
- 选择全部记录

上一页 首页 记录数：9条，第1-9条，每页20条

□ 更多 Evidence for phasic sequences in nuclear HBcAg formation and cell membrane-directed flow of CorE particles in chronic hepatitis B — 病毒性肝炎
F. Gudat and L. Bianchi. Gastroenterology, 1977, 73(5): 1194-7 — 19319

□ 更多 Biologic significance of the detection of HBsAG and HBcAg in liver and tumor from 204 HBsAG-positive patients with primary hepatocellular carcinoma — 病毒性肝炎
H. C. Hsu;; T. T. Wu;; J. C. Sheu;; C. Y. Wu;; T. J. Chiou;; C. S. Lee;; D. S. Chen. Hepatology, 1989, 9(5): 747-50 — 21880

□ 更多 Cellular immune response to HBcAg in mother-to-infant transmission of hepatitis B virus — 病毒性肝炎
H. Y. Hsu;; M. H. Chang;; K. H. Hsieh;; C. Y. Lee;; H. H. Lin;; L. H. Hwang;; P. J. Chen;; D. S. Chen. Hepatology, 1992, 15(5): 770-6 — 21882

□ 更多 Excess HBcAg in HBc antibody-negative chronic hepatitis B virus carriers — 病毒性肝炎
Y. Lazizi;; P. Dubreuil;; J. Pillot. Hepatology, 1993, 17(6): 966-70 — 22079

□ 更多 HBcAg expressed on the surface of circulating Dane particles in patients with hepatitis B virus infection without evidence of anti-HBc formation — 病毒性肝炎
B. Moller;; U. Hopf;; R. Stemerowicz;; G. Henze;; H. Gelderblom. Hepatology, 1989, 10(2): 179-85 — 22302

□ 更多 Experimental HBV and delta infections of chimpanzees: occurrence and significance of intrahepatic immune complexes of HBcAg and Delta antigen — 病毒性肝炎
M. Rizzetto;; M. G. Canese;; R. H. Purcell;; W. T. London;; L. D. Sly;; J. L. Gerin. Hepatology, 1981, 1(6): 567-74 — 22495

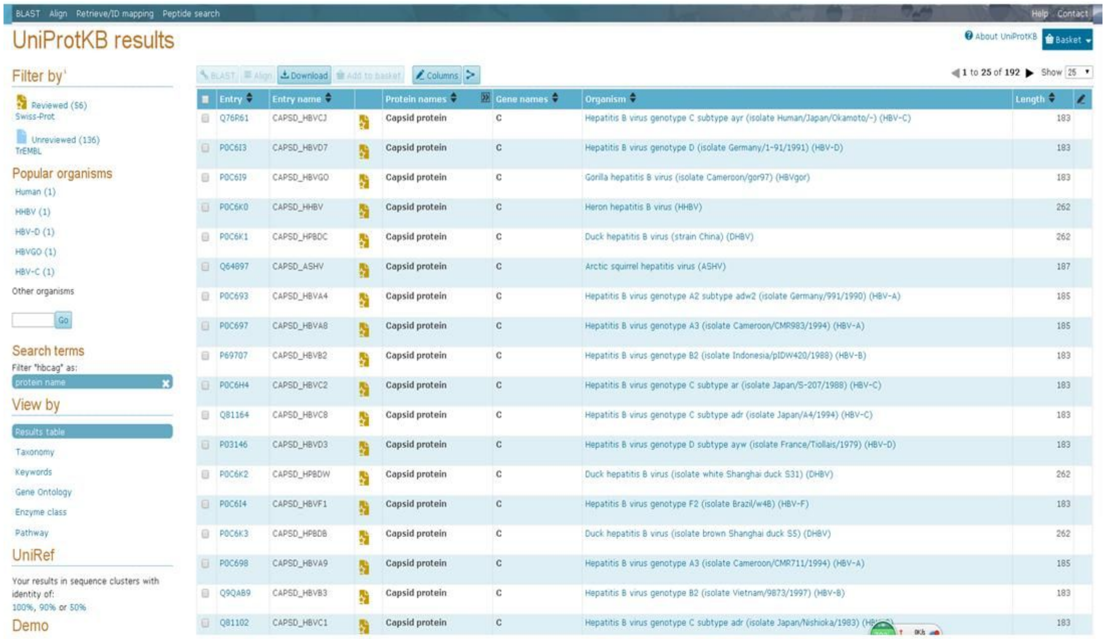
□ 更多 Identification of hepatitis B virus-DNA in the liver by in situ hybridization using a biotinylated probe. Relation to HBcAg expression and histology — 病毒性肝炎
N. V. Naoumov;; H. M. Daniels;; F. Davison;; A. L. Eddleston;; G. J. Alexander;; R. Williams. J Hepatol, 1993, 19(2): 204-10 — 31376

□ 更多 In vivo inhibition of anti-hepatitis B virus Core antigen (HBcAg) immunoglobulin G production by HBcAg-specific CD4(+) Th1-type T-cell clones in a hu-PBL-NOD/SCID mouse model — 病毒性肝炎
T. Cao;; P. Meuleman;; I. Desombere;; M. Sallberg;; G. Leroux-Roels. J Virol, 2001, 75(23): 11449-56 — 37124

□ 更多 Exposure of RNA templates and encapsidation of spliced viral RNA are influenced by the arginine-rich domain of human hepatitis B virus Core antigen (HBcAg 165-173) — 病毒性肝炎
S. Le Pogam;; P. K. Chua;; M. Newman;; C. Shih. J Virol, 2005, 79(3): 1871-87 — 37502

上一页 首页 记录数：9条，第1-9条，每页20条

Figure 7 Page of the hepatitis viral protein mining (2)



UniProtKB results

Filter by

Reviewed (56)
Swiss-Prot

Unreviewed (136)
TrEMBL

Popular organisms

Human (1)
HHBV (1)
HBV-D (1)
HBVGO (1)
HBV-C (1)
Other organisms

Search terms

Filter "hbcag" as:
protein name

View by

Results table
Taxonomy
Keywords
Gene Ontology
Enzyme class
Pathway

UniRef

Your results in sequence clusters with identity of:
100%, 90% or 50%

Demo

1 to 25 of 192 Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
Q76R61	CAPSD_HBVCJ	Capsid protein	C	Hepatitis B virus genotype C subtype ayr (isolate Human/Japan/Okamoto/-) (HBV-C)	183
P0C6I3	CAPSD_HBVD7	Capsid protein	C	Hepatitis B virus genotype D (isolate Germany/1-91/1991) (HBV-D)	183
P0C6I9	CAPSD_HBVGO	Capsid protein	C	Gorilla hepatitis B virus (isolate Cameroon/gor97) (HBVgor)	183
P0C6K0	CAPSD_HHBV	Capsid protein	C	Heron hepatitis B virus (HHBV)	262
P0C6K1	CAPSD_HPBDC	Capsid protein	C	Duck hepatitis B virus (strain China) (DHBV)	262
Q64897	CAPSD_ASHV	Capsid protein	C	Arctic squirrel hepatitis virus (ASHV)	187
P0C693	CAPSD_HBVA4	Capsid protein	C	Hepatitis B virus genotype A2 subtype adw2 (isolate Germany/991/1990) (HBV-A)	185
P0C697	CAPSD_HBVA8	Capsid protein	C	Hepatitis B virus genotype A3 (isolate Cameroon/CMR983/1994) (HBV-A)	185
P69707	CAPSD_HBVB2	Capsid protein	C	Hepatitis B virus genotype B2 (isolate Indonesia/pIDW420/1988) (HBV-B)	183
P0C6H4	CAPSD_HBVC2	Capsid protein	C	Hepatitis B virus genotype C subtype ar (isolate Japan/S-207/1988) (HBV-C)	183
Q81164	CAPSD_HBVC8	Capsid protein	C	Hepatitis B virus genotype C subtype adr (isolate Japan/A4/1994) (HBV-C)	183
P03146	CAPSD_HBVD3	Capsid protein	C	Hepatitis B virus genotype D subtype ayw (isolate France/Tiollais/1979) (HBV-D)	183
P0C6K2	CAPSD_HPBDW	Capsid protein	C	Duck hepatitis B virus (isolate white Shanghai duck S31) (DHBV)	262
P0C6I4	CAPSD_HBVF1	Capsid protein	C	Hepatitis B virus genotype F2 (isolate Brazil/w4B) (HBV-F)	183
P0C6K3	CAPSD_HPBDB	Capsid protein	C	Duck hepatitis B virus (isolate brown Shanghai duck S5) (DHBV)	262
P0C698	CAPSD_HBVA9	Capsid protein	C	Hepatitis B virus genotype A3 (isolate Cameroon/CMR711/1994) (HBV-A)	185
Q9QAB9	CAPSD_HBVB3	Capsid protein	C	Hepatitis B virus genotype B2 (isolate Vietnam/9873/1997) (HBV-B)	183
Q81102	CAPSD_HBVC1	Capsid protein	C	Hepatitis B virus genotype C subtype adr (isolate Japan/Nishioka/1983) (HB	183

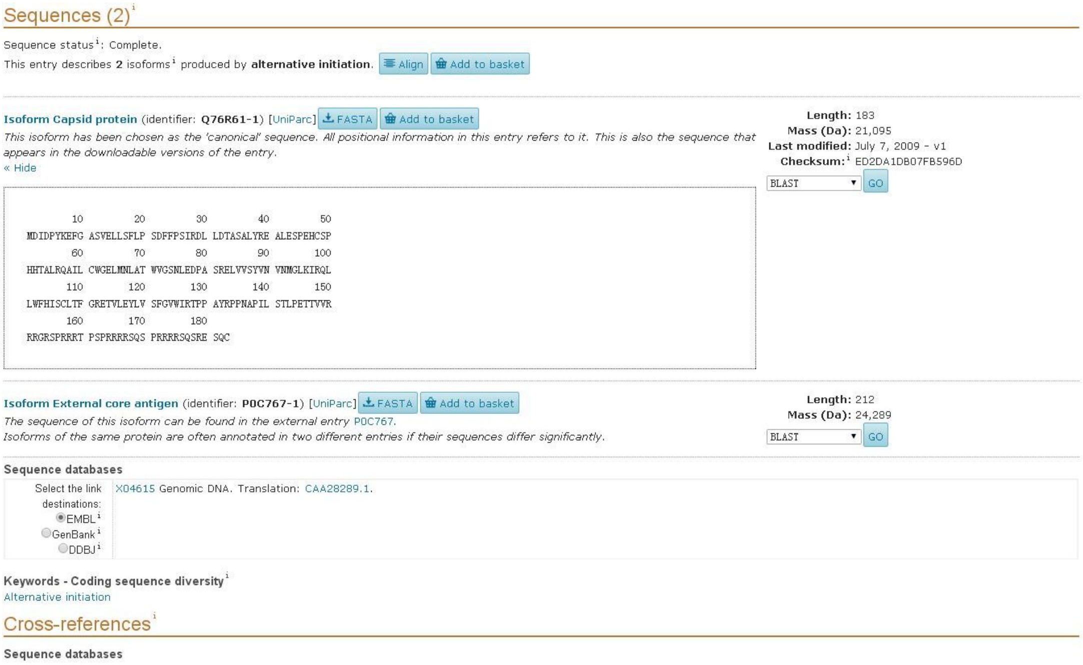
Figure 8 Page of the hepatitis viral protein of literature database integrating and hyperlinking to the Uniprot protein scientific database

Viral hepatitis subject literature knowledge database extends three functions through data mining, information integration and hyperlinking

Obtain the hepatitis viral protein sequence and annotation information. See fig.9 as follows:

Result of homological analysis of the protein sequences (BLAST). See fig.10 as follows:

Obtain the evolutionary tree mapping. See fig.11 as follows:



Sequences (2)

Sequence status: Complete.

This entry describes **2** isoforms produced by **alternative initiation**.

Isoform Capsid protein (identifier: **Q76R61-1**) [UniParc]

This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

« Hide

Length: 183
Mass (Da): 21,095
Last modified: July 7, 2009 - v1
Checksum: ED2DA1DB07FB596D

```
        10         20         30         40         50
MDIDPYKEFG ASVELLSFLP SDFFPSIRDL LDTASALYRE ALESPEHCSP
        60         70         80         90        100
HHTALRQAIL CWGELMNLAT WVGSNLEDPA SRELVVSYVN VNMGLKIRQL
       110        120        130        140        150
LWFHISCLTF GRETVLEYLV SFGVWIRTPP AYRPPNAPIL STLPETTVVR
       160        170        180
RRGRSPRRRT PSPRRRRSQS PRRRRSQSRE SQC
```

Isoform External core antigen (identifier: **P0C767-1**) [UniParc]

The sequence of this isoform can be found in the external entry P0C767.
Isoforms of the same protein are often annotated in two different entries if their sequences differ significantly.

Length: 212
Mass (Da): 24,289

Sequence databases

Select the link destinations:
● EMBL
○ GenBank
○ DDBJ

X04615 Genomic DNA. Translation: CAA28289.1.

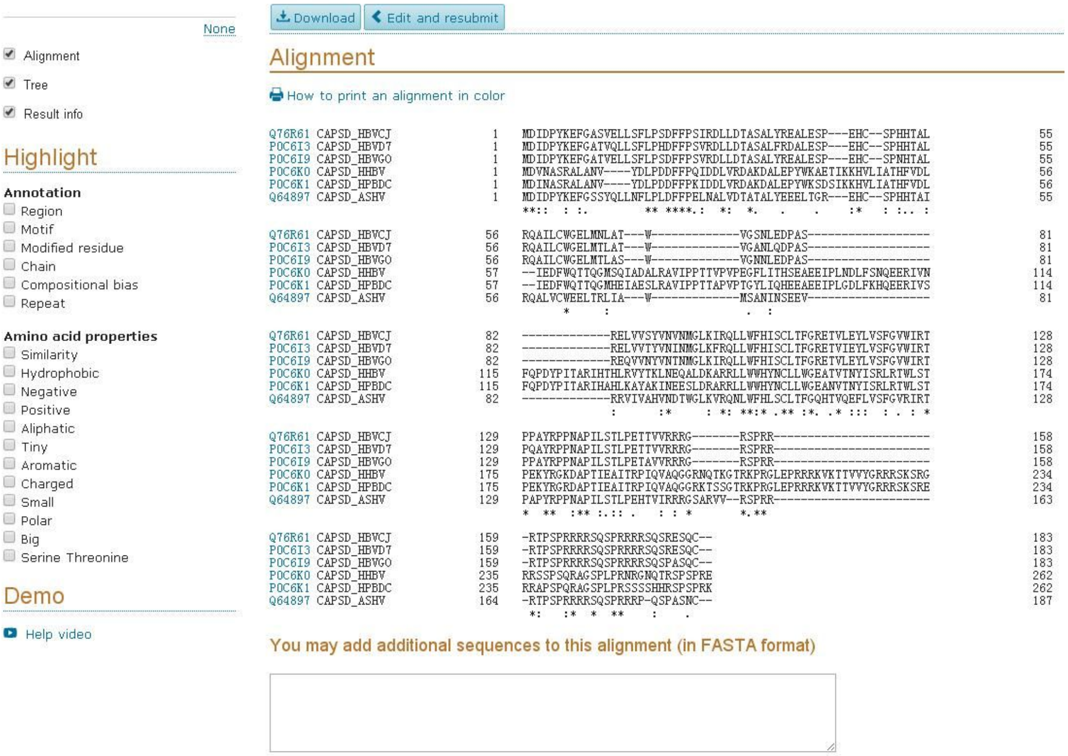
Keywords - Coding sequence diversity

Alternative initiation

Cross-references

Sequence databases

Figure 9 Page of the protein sequence and annotation information of HBcAg



☑ Alignment

☑ Tree

☑ Result info

Highlight

Annotation

□ Region
□ Motif
□ Modified residue
□ Chain
□ Compositional bias
□ Repeat

Amino acid properties

□ Similarity
□ Hydrophobic
□ Negative
□ Positive
□ Aliphatic
□ Tiny
□ Aromatic
□ Charged
□ Small
□ Polar
□ Big
□ Serine Threonine

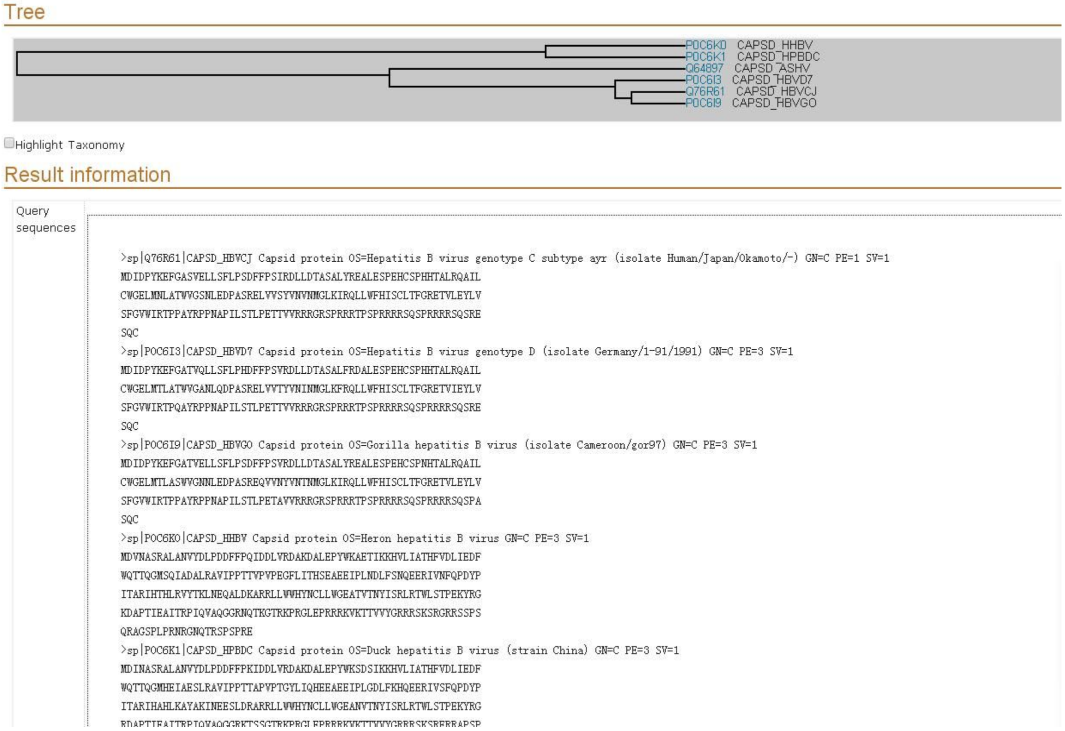
Demo

Alignment

```
Q76R61 CAPSD_HBVCJ        1    MDIDPYKEFGASVELLSFLPSDFFPSIRDLLDTASALYREALESP---EHC--SPHHTAL     55
P0C6I3 CAPSD_HBVD7        1    MDIDPYKEFGATVQLLSFLPHDFFPSVRDLLDTASALFRDALESP---EHC--SPHHTAL     55
P0C6I9 CAPSD_HBVGO        1    MDIDPYKEFGATVELLSFLPSDFFPSVRDLLDTASALYREALESP---EHC--SPNHTAL     55
P0C6K0 CAPSD_HHBV         1    MDVNASRALANV----YDLPDDFFPQIDDLVRDAKDALEPYWKAETIKKHVLIATHFVDL     56
P0C6K1 CAPSD_HPBDC        1    MDINASRALANV----YDLPDDFFPKIDDLVRDAKDALEPYWKSDSIKKHVLIATHFVDL     56
Q64897 CAPSD_ASHV         1    MDIDPYKEFGSSYQLLNFLPLDFFPELNALVDTATALYEEELTGR---EHC--SPHHTAI     55
                               **::  : :.        ** ****.:  *:  *.   .    .    :*   : :.. :

Q76R61 CAPSD_HBVCJ       56    RQAILCWGELMNLAT---W-------------VGSNLEDPAS------------------     81
P0C6I3 CAPSD_HBVD7       56    RQAILCWGELMTLAT---W-------------VGANLQDPAS------------------     81
P0C6I9 CAPSD_HBVGO       56    RQAILCWGELMTLAS---W-------------VGNNLEDPAS------------------     81
P0C6K0 CAPSD_HHBV        57    --IEDFWQTTQGMSQIADALRAVIPPTTVPVPEGFLITHSEAEEIPLNDLFSNQEERIVN    114
P0C6K1 CAPSD_HPBDC       57    --IEDFWQTTQGMHEIAESLRAVIPPTTAPVPTGYLIQHEEAEEIPLGDLFKHQEERIVS    114
Q64897 CAPSD_ASHV        56    RQALVCWEELTRLIA---W-------------MSANINSEEV------------------     81
                                     *     :                    .  :

Q76R61 CAPSD_HBVCJ       82    -------------RELVVSYVNVNMGLKIRQLLWFHISCLTFGRETVLEYLVSFGVWIRT    128
P0C6I3 CAPSD_HBVD7       82    -------------RELVVTYVNINMGLKFRQLLWFHISCLTFGRETVIEYLVSFGVWIRT    128
P0C6I9 CAPSD_HBVGO       82    -------------REQVVNYVNTNMGLKIRQLLWFHISCLTFGRETVLEYLVSFGVWIRT    128
P0C6K0 CAPSD_HHBV       115    FQPDYPITARIHTHLRVYTKLNEQALDKARRLLWWHYNCLLWGEATVTNYISRLRTWLST    174
P0C6K1 CAPSD_HPBDC      115    FQPDYPITARIHAHLKAYAKINEESLDRARRLLWWHYNCLLWGEANVTNYISRLRTWLST    174
Q64897 CAPSD_ASHV        82    -------------RRVIVAHVNDTWGLKVRQNLWFHLSCLTFGQHTVQEFLVSFGVRIRT    128
                                            :      :*     : *: **:* .** :*. .* :::  : . : *

Q76R61 CAPSD_HBVCJ      129    PPAYRPPNAPILSTLPETTVVRRRG-------RSPRR-----------------------    158
P0C6I3 CAPSD_HBVD7      129    PQAYRPPNAPILSTLPETTVVRRRG-------RSPRR-----------------------    158
P0C6I9 CAPSD_HBVGO      129    PPAYRPPNAPILSTLPETAVVRRRG-------RSPRR-----------------------    158
P0C6K0 CAPSD_HHBV       175    PEKYRGKDAPTIEAITRPIQVAQGGRNQTKGTRKPRGLEPRRRKVKTTVVYGRRRSKSRG    234
P0C6K1 CAPSD_HPBDC      175    PEKYRGRDAPTIEAITRPIQVAQGGRKTSSGTRKPRGLEPRRRKVKTTVVYGRRRSKSRE    234
Q64897 CAPSD_ASHV       129    PAPYRPPNAPILSTLPEHTVIRRRGSARVV--RSPRR-----------------------    163
                               *  **  :** :.:: .   : : *       *.**

Q76R61 CAPSD_HBVCJ      159    -RTPSPRRRRSQSPRRRRSQSRESQC--                                    183
P0C6I3 CAPSD_HBVD7      159    -RTPSPRRRRSQSPRRRRSQSRESQC--                                    183
P0C6I9 CAPSD_HBVGO      159    -RTPSPRRRRSQSPRRRRSQSPASQC--                                    183
P0C6K0 CAPSD_HHBV       235    RRSSPSQRAGSPLPRNRGNQTRSPSPRE                                    262
P0C6K1 CAPSD_HPBDC      235    RRAPSPQRAGSPLPRSSSSHHRSPSPRK                                    262
Q64897 CAPSD_ASHV       164    -RTPSPRRRRSQSPRRRP-QSPASNC--                                    187
                                *:   :*  *  **    :    .
```

You may add additional sequences to this alignment (in FASTA format)

Figure 10 Page of homological analysis result of the HBcAg protein sequences (BLAST)



Tree

□Highlight Taxonomy

Result information

Query sequences

```
>sp|Q76R61|CAPSD_HBVCJ Capsid protein OS=Hepatitis B virus genotype C subtype ayr (isolate Human/Japan/Okamoto/-) GN=C PE=1 SV=1
MDIDPYKEFGASVELLSFLPSDFFPSIRDLLDTASALYREALESPEHCSPHHTALRQAIL
CWGELMNLATWVGSNLEDPASRELVVSYVNVNMGLKIRQLLWFHISCLTFGRETVLEYLV
SFGVWIRTPPAYRPPNAPILSTLPETTVVRRRGRSPRRRTPSPRRRRSQSPRRRRSQSRE
SQC
>sp|P0C6I3|CAPSD_HBVD7 Capsid protein OS=Hepatitis B virus genotype D (isolate Germany/1-91/1991) GN=C PE=3 SV=1
MDIDPYKEFGATVQLLSFLPHDFFPSVRDLLDTASALFRDALESPEHCSPHHTALRQAIL
CWGELMTLATWVGANLQDPASRELVVTYVNINMGLKFRQLLWFHISCLTFGRETVIEYLV
SFGVWIRTPQAYRPPNAPILSTLPETTVVRRRGRSPRRRTPSPRRRRSQSPRRRRSQSRE
SQC
>sp|P0C6I9|CAPSD_HBVGO Capsid protein OS=Gorilla hepatitis B virus (isolate Cameroon/gor97) GN=C PE=3 SV=1
MDIDPYKEFGATVELLSFLPSDFFPSVRDLLDTASALYREALESPEHCSPNHTALRQAIL
CWGELMTLASWVGNNLEDPASREQVVNYVNTNMGLKIRQLLWFHISCLTFGRETVLEYLV
SFGVWIRTPPAYRPPNAPILSTLPETAVVRRRGRSPRRRTPSPRRRRSQSPRRRRSQSPA
SQC
>sp|P0C6K0|CAPSD_HHBV Capsid protein OS=Heron hepatitis B virus GN=C PE=3 SV=1
MDVNASRALANVYDLPDDFFPQIDDLVRDAKDALEPYWKAETIKKHVLIATHFVDLIEDF
WQTTQGMSQIADALRAVIPPTTVPVPEGFLITHSEAEEIPLNDLFSNQEERIVNFQPDYP
ITARIHTHLRVYTKLNEQALDKARRLLWWHYNCLLWGEATVTNYISRLRTWLSTPEKYRG
KDAPTIEAITRPIQVAQGGRNQTKGTRKPRGLEPRRRKVKTTVVYGRRRSKSRGRRSSPS
QRAGSPLPRNRGNQTRSPSPRE
>sp|P0C6K1|CAPSD_HPBDC Capsid protein OS=Duck hepatitis B virus (strain China) GN=C PE=3 SV=1
MDINASRALANVYDLPDDFFPKIDDLVRDAKDALEPYWKSDSIKKHVLIATHFVDLIEDF
WQTTQGMHEIAESLRAVIPPTTAPVPTGYLIQHEEAEEIPLGDLFKHQEERIVSFQPDYP
ITARIHAHLKAYAKINEESLDRARRLLWWHYNCLLWGEANVTNYISRLRTWLSTPEKYRG
RDAPTIEAITRPIQVAQGGRKTSSGTRKPRGLEPRRRKVKTTVVYGRRRSKSRERRAPSP
```

Figure 11 Page of the evolutionary tree mapping of the HBcAg protein

3 Discussion, Conclusion and Future Work

3.1 Discussion

The viral hepatitis bilingual bibliographic database was successfully built, and protein text was also successfully mined, and two different classes of databases were also triumphantly integrated, but we encountered some problems, especially such as false positive mining results in bilingual protein text mining. Having investigated the false positive questions, we think there are probably three causes resulting in the false positive mining results:

1) Low quality of the original datasets collected;

2) The accuracy and unity of a specialized word usage is not enough in building of bilingual control vocabulary;

3) In data mining and integration, computer algorithms, mining mode and route selection, and algorithm itself are unreasonable or the system has defects.

As for the problems above, we use artificial quality control to handle the collected original datasets; refer to specialized dictionary and consult the experts to solve the accuracy and unity question of a specialized word usage; try to explore different algorithms, mining mode and route to solve accuracy and efficiency question of data mining and integration.

After the viral hepatitis bilingual bibliographic database was used and demonstrated, we have got many feedbacks from users. Most of them love the convenience of easily searching hepatitis viral protein names, locating highlighted viral protein names in search results, and accessing UniProt database for the detailed protein information through information integration and links. But they also raised some questions and proposed many advices. Overall, however, the feedback is very positive so far. According to users' suggestions and problems, we have discovered, following issues are currently being considered and actually some of them are being undertaken in order to further enhance the system and make it more efficient and convenient:

1) add more hepatitis viral protein names and their features into the English-Chinese Controlled-vocabulary dictionary. This work is continuously being conducted and actually we also plan to add relationships of hepatitis viral proteins and other relevant information so as to finally construct a Chinese hepatitis viral protein ontology. Then it would be possible to realize semantic-based text mining and provide users with knowledge-based information service.

2) integrate more factual scientific databases, especially factual gene databases. Some users are also interested in other special fields, such as evidence-based medicine, AIDS, etc. If search results of a special topic from a bibliographic database can be integrated with relevant factual scientific databases, it is certainly very helpful and convenient for users. This is an interesting direction for information integration and knowledge mining.

3.2 Conclusion

With the fast development of the viral hepatitis research, to satisfy user's information needs is becoming an inevitable challenge. So, construction of the viral hepatitis bilingual literature database is important, significant and useful. Integration of two different classes of databases via data mining and linking is innovative and trend for database development. Moreover, information integration and data mining are playing a more and more important role in big data era.

3.3 Future work

In order to solve the problems above, future work must be done as follows:

1) Constantly extend and update datasets in viral hepatitis bilingual literature database;

2) Constantly improve mining and integrating quality so as to decrease the false positive results as low as possible through algorithm improvement and machine learning;

3) Further improve accuracy and unity of the bilingual control vocabulary;

4) The viral hepatitis bilingual literature database will be linked more factual scientific atabase via data mining and information integration.

Acknowledgements

This work is supported by The Chosen Excellent Program for Introduced Outstanding Talent of Chinese Academy of Sciences in the Fields of Bibliographical Information and Periodical Publication 2010 (Subject field 100 talent program) and Chinese National Science and Technology Support Project (No.2013BAH21B06)

Reference

Alexopoulou, D., Wächter, T., Pickersgill, L., Eyre, C. and Schroeder, M.: Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. ***BMC Bioinformatics***. 9(Suppl 4), S2, 2008

Chen Heng, Jin Yi, Zhao Yan, Zhang Yongjuan, Chen Chengcai, Sun Jilin, Zhang Shen. Mining and Information Integration Practice for Chinese Bibliographic Database of Life Sciences. Book title : ***Advances in Data Mining: Applications and Theoretical Aspects***; Vol.7987, pp.1-10, 2013. Publisher: Springer Berlin Heidelberg. Book subtitle**:** ***13th Industrial Conference, ICDM 2013,*** NewYork, NY, USA, July 16-21, 2013, Proceedings. (DOI: 10.1007/978-3-642-39736-3)

Doms, A. and Schroeder, M.: GoPubMed: exploring PubMed with the gene ontology.

Nucleic Acids Research. Vol.33: 783-786, 2005

Feng Xinmin and Wang Jiandong. The concept dilemma of knowledge mining and the broad-sense knowledge mining. ***Journal of Information,*** Vol.27 (7): 63-65, 2008

Maglott, D., Ostell, J., Pruitt, K. and Tatusova, T.: Entrez Gene: gene-centered information at NCBI. ***Nucleic Acids Research***. Vol.39: 52-57, 2011

McGarry, K., Garfield, S. and Morris, N.: Recent trends in knowledge and data integration for the life sciences. ***Expert Systems***. Vol.23(5): 330-341, 2006

Pasquier, C.: Biological data integration using Semantic Web technologies. ***Biochimie***. Vol.90: 584-594, 2008

Sahoo, S., Bodenreider, O., Zeng, K. and Sheth, A.: An experiment in integrating large biomedical knowledge resources with RDF: Application to associating genotype and phenotype information. In: ***16th International World Wide Web Conference (WWW2007) on Health Care and Life Sciences Data Integration for the Semantic Web***, pp. 8-12. Banff, Canada(2007)

Yan Zhihong. Research on the integration mode of digital information resources in Chinese University libraries. ***Thesis for Master degree***, Chong Qing University, 2008

Zhang Xiaojuan, Zhang Yutao, Zhang Jieli and Wang Juncheng. The central research issues of information resources integration in china. ***Journal of the China Society for Scientific andTechnical Information,*** Vol.28 (5): 791-800, 2010