Infection, Genetics and Evolution 11 (2011) 734–739

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Dragon exploratory system on Hepatitis C Virus (DESHCV)

Samuel K. Kwofie [a], Aleksandar Radovanovic [b], Vijayaraghava S. Sundararajan [a], Monique Maqungo [a], Alan Christoffels [a], Vladimir B. Bajic [b,*]

[a] *South African National Bioinformatics Institute, University of the Western Cape, Private Bag- X17, Modderdam Road, Bellville, Cape Town, South Africa*

[b] *Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia*

ARTICLE INFO

Article history:
Received 30 April 2010
Received in revised form 30 November 2010
Accepted 8 December 2010
Available online 29 December 2010

Keywords:
Hepatitis C Virus
Text-mining
Dictionaries
Biomedical concepts
Database
Hypotheses generation

ABSTRACT

Even though Hepatitis C Virus (HCV) cDNA was characterized about 20 years ago, there is insufficient understanding of the molecular etiology underlying HCV infections. Current global rates of infection and its increasingly chronic character are causes of concern for health policy experts. Vast amount of data accumulated from biochemical, genomic, proteomic, and other biological analyses allows for novel insights into the HCV viral structure, life cycle and functions of its proteins. Biomedical text-mining is a useful approach for analyzing the increasing corpus of published scientific literature on HCV. We report here the first comprehensive HCV customized biomedical text-mining based online web resource, dragon exploratory system on Hepatitis C Virus (DESHCV), a biomedical text-mining and relationship exploring knowledgebase was developed by exploring literature on HCV. The pre-compiled dictionaries existing in the dragon exploratory system (DES) were enriched with biomedical concepts pertaining to HCV proteins, their name variants and symbols to make it suitable for targeted information exploration and knowledge extraction as focused on HCV. A list of 32,895 abstracts retrieved via PubMed database using specific keywords searches related to HCV were processed based on concept recognition of terms from several dictionaries. The web query interface enables retrieval of information using specified concepts, keywords and phrases, generating text-derived association networks and hypotheses, which could be tested to identify potentially novel relationship between different concepts. Such an approach could also augment efforts in the search for diagnostic or even therapeutic targets. DESHCV thus represents online literature-based discovery resource freely accessible for academic and non-profit users via http://apps.sanbi.ac.za/DESHCV/ and its mirror site http://cbrc.kaust.edu.sa/deshcv/.

1. Introduction

The skyrocketing chronicity and global infection rate of Hepatitis C Virus (HCV) necessitate the need to unlock the molecular etiology underlying the pathophysiology of HCV related diseases such as liver cancer. The plethora of essential molecular data in the corpus of published biomedical literature could be leveraged to augment efforts towards discovery of novel anti-viral drugs, cellular receptors and appropriate predictive biomarkers. Most of the data derived from high throughput and "omics" experiments exist in variety of formats, thereby making cross data integration difficult. The development of HCV specific database as repositories of information utilizable in cross discipline biology research is therefore vital. The Los Alamos Hepatitis C Virus sequence database (http://hcv.lanl.gov) offers annotated sequences and analysis tools (Kuiken et al., 2005). The Los Alamos hepatitis C immunology database (http://hcv.lanl.gov/content/immuno/immuno-main.html) is a repository of biocurated immunological epitopes integrated with retrieval and analysis tools (Yusim et al., 2005). The Japanese HCV database integrated in the HVDB (http://s2as02.genes.nig.ac.jp) comprises data on phylogenetic and provides java embedded viewers for visualizing phylogenetic trees and the HCV genome. The European Hepatitis C Virus database (euHCVdb, http://euhcvdb.ibcp.fr) provides annotated sequences and tools for analysis, and information on protein structure and function (Combet et al., 2007). Hepatitis C Virus sequence and immunology database and analytical applications (HCVdb, http://www.hcvdb.org/index.asp?bhcp=1) offers data on analyzed protein sequence and features, epitopes, and curated knowledge on protein interactions and function. Binding site finder (BSFINDER, http://wilab.inha.ac.kr/bsfinder) enable prediction of HCV binding site residues and potential interacting protein partners using support vector machine (Chen and Han, 2009). A comprehensive review of selected HCV related database has highlighted the useful capabilities, utilities and applications of these resources (Kuiken et al., 2006). Hepatitis C Virus-specific database contain much useful information on molecular biology, sequences, immunology, protein structure and function, viral

* Corresponding author.
E-mail address: vladimir.bajic@kaust.edu.sa (V.B. Bajic).

1567-1348/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2010.12.006

evolution and genetics. Nevertheless, there is no resource that allows for the exploration of potential links (associations) between different biomedical concepts of relevance to HCV. We have developed one such resource, dragon exploratory system on Hepatitis C Virus (DESHCV, http://apps.sanbi.ac.za/DESHCV/ and its mirror site http://cbrc.kaust.edu.sa/deshcv/), based on text-mining approach to complement the existing HCV resources and to enable different insights into the molecular context of HCV functioning.

As reported (Cohen and Hersh, 2005), the biomedical knowledgebase is growing at an increasing rate. PubMed database is currently a repository of about 20 million citations for biomedical articles from MEDLINE and life science related journals. MeSH indexers index about 500,000 journal articles annually for PubMed/MEDLINE (Mitchell et al., 2003). A search by publication dates in PubMed shows over 20,000 HCV related records published over the last decade. At the time when this study is conducted a total of 32,895 HCV related documents were available that makes it virtually impossible for a single researcher or a research group to process in any reasonable time. However, biomedical text-mining approach could be utilized to analyze this large volume of scientific data and reports published on HCV. Biomedical text-mining employs different techniques to extract and summarize information from text (Cohen and Hunter, 2008). Its algorithms may derive putative relationships between disjunct sets of concepts to unravel potentially new associations and hypotheses for possible novel discovery. The co-occurrence of the concepts of interests, either in a portion of text (say abstract) or in a sentence, is identified computationally, and provides useful clues on the potential associations between these concepts, some of which may be completely new. For example, using text-mining techniques fish oil was proposed to have a potential therapeutic effect on Raynaud's disease (Swanson, 1986). This hypothesis was made by showing a relationship between fish oil and Raynaud's disease via physiological concepts such as high blood pressure and platelet aggregation. This has been proved to be the correct inference on the link of fish oil and Raynaud's disease. As another example, literature based discovery was previously used to predict thalidomide as a possible therapeutic drug for HCV infection after detecting implicit associations in biomedical text (Weeber et al., 2003). Text-mining approach was also employed in identifying some of the hepatocellular proteins used in generating the human HCV interactome (de Chassey et al., 2008).

The essential features and characteristics of some of the available text-mining tools have been discussed elsewhere (Bajic et al., 2005; Weeber et al., 2005). Shi and Campagne (2005) have described in detail the various concepts, principles, challenges and algorithms behind development of biomedical text-mining tools during the building of protein catalogue and implementation of the Textractor Framework (http://icb.med.cornell.edu/crt/textractor/index.xml). Anni 2.0 (http://biosemantics.org/anni/), a web-based biomedical text-mining tool offers an ontology-based interface to MEDLINE and enables retrieval of documents and possible association amongst biomedical concepts (Jelier et al., 2008). PolySearch (http://wishart.biology.ualberta.ca/polysearch), a web-based text-mining system allows the retrieval of relationships between human diseases, genes, mutations, drugs and metabolites (Cheng et al., 2008). Nowadays, customized knowledgebases or topic-specific text-mining resources are increasingly becoming popular. Typical examples are the dragon exploratory system (DES) based resources: dragon TF association miner (DTFAM, http://research.i2r.a-star.edu.sg/DRAGON/TFAM), useful for exploring functional association amongst transcription factors (Pan et al., 2004); dragon database for exploration of sodium channels in human (DDESC, http://apps.sanbi.ac.za/ddesc), which provides comprehensive text-mining information related to sodium channels (Sagar et al., 2008). DES and its previous versions were successfully used in compilation of several other resources such as in database on ovarian cancer (Kaur et al., 2009) and esophageal cancer (Essack et al., 2009), as well as in studies on prioritizing disease genes (Lombard et al., 2007; Tiffin et al., 2005).

To the best knowledge of the authors, there is no single database solely focused on HCV research that allows for the comprehensive exploration of the association between the biomedical concepts related to HCV. In this report, we present dragon exploratory system on Hepatitis C Virus (DESHCV), the previous version of it being reported earlier (Kwofie et al., 2009). DESHCV is developed using DES. The HCV proteins and their name variants have been integrated into the pre-compiled dictionaries of biological concepts present in DES. These concepts are cross-referenced to database such as gene ontologies (GO), UNIPROT, KEGG Pathway, REACTOME and Entrez Gene. A list of abstracts was retrieved via PubMed database using keywords related to HCV. These abstracts were analyzed using concepts in the following dictionaries: "human genes and proteins", "metabolites and enzymes", "pathways", "chemicals with pharmacological effects", "Hepatitis C Virus concepts", and "disease concepts".

The user-friendly online interface allows concepts, keywords and phrases searches. A concept query could generate networks and hypothesis. The computationally suggested associations between genes and other concepts such as diseases may assist experimental biologist to explore which genes amongst a pool of genes need to be characterized for further molecular analysis. Such an approach could in principle also lead to possible discovery of new vaccines; and enhance the development of appropriate diagnostic method. The user has the possibility to inspect the post-processed PubMed abstracts with colour-coded tagged concepts from the used dictionaries as found in the text. The downloadable concept lists sheet could be a primary source of data for biocuration. The paired concept list spreadsheet can serve as the essential preliminary data for exploring associations between concepts and can be converted easily into simple interaction file format (SIF) compatible with some of the interaction visualization and analysis tools such as Cytoscape (Killcoyne et al., 2009). Researchers with minimal or no knowledge on text-mining can explore DESHCV with ease via system's simplified user query interface. The integrated downloadable tutorial manual further aids easy use of the system. DESHCV is an online text-mining developed knowledgebase freely available for non-commercial use via http://apps.sanbi.ac.za/DESHCV and http://cbrc.kaust.edu.sa/deshcv.

2. Construction and content

2.1. Implementation

A list of 32,895 MEDLINE abstracts was collected via PubMed interface using the following keywords query: HCV OR "Hepatitis C Virus". The PubMed textual data downloaded in the extensible markup language (XML) format allow for easy data integration into DES for semantic processing and analysis. The DESHCV data files were generated by DES, a proprietary biomedical text-mining tool of OrionCell (http://www.orioncell.org). The HCV proteins symbols and name variants were added to the "human protein and genes" dictionaries, and were mapped onto external annotation database. The HCV name variants have been disambiguated and integrated in the database. For example, a concept query with the word "core" retrieves "core protein" and not the words "score" or "core". The word core apart from being part of the morphological features of HCV core protein has different meanings in various English dictionaries.

Post processed PubMed abstracts containing found concepts from used dictionaries constitute part of the database files. The DESHCV precompiled dictionary data files systems are composed of categories of terms such as names of "genes", "proteins", "metabolites", enzymes", "pathways", "pharmacological chemicals", and "diseases". The categories of terms consist of biological entities that we refer to as concepts. Therefore for the purpose of this work, each dictionary consists of a catalogue of related concepts. The DESHCV system is organized into 6 distinct dictionaries: "human genes and proteins", "metabolites and enzymes", "pathways", "chemicals with pharmacological effects", "Hepatitis C Virus concepts", and "disease concepts". Hierarchically, each dictionary is considered as a parent with no subcategories. For example, both human gene "interferon alpha 2" and human protein "alanine aminotransferase" are concepts belonging to the same parent "human genes and proteins" dictionaries and not separate subcategories. In some instances certain concepts are assigned to one dictionary, even though they could belong to another. For example the concept name ribavirin, a therapeutic drug for hepatitis C infection has been assigned to the "metabolites and enzymes" dictionary, even though it could also belong to "chemicals with pharmacological effects" dictionary. Our decision is influenced by the belief that most literature report ribavirin as a metabolite.

For the purpose of hypothesis generation, the DES association module generates association maps in the form of graphs. The nodes of these graphs represent various concepts and the edges linking the nodes are weighted with frequencies of occurrence in the PubMed abstracts. Hypotheses are generated if for example, concepts X and Y are correlated, as well as concepts Y and Z, but no correlation is found between X and Z. Then the hypotheses generator module will suggest a potential link between concepts X and Z.

2.2. Database architecture

The database comprises of an Apache HTTP web server integrated with a back-end MySQL server whilst HTML/CSS and JavaScript scripts constitute the front-end (Fig. 1). The data files consist of precompiled concept dictionaries, post processed PubMed abstracts, and cross-referenced annotations. The logic systems consist of Perl and PHP modules.

2.3. The database interfaces

DESHCV is based on a client-server model and can be accessed by any user with a standard web browser. The user interfaces in DESHCV allow easy navigation, query, inspection, and retrieval of data. It comprises of concept and abstract search menus (Fig. 2). The concept search menu allows users to search the database using specified concepts. The hypothesis generator is an integral functional component of the concept search menu.

3. Results and discussions

3.1. Concepts queries

DESHCV is the first text-mining based web accessible resource developed using published abstracts of scientific reports on HCV as referenced in PubMed. DESHCV provides comprehensive information on HCV and enables users to gain easy insights through exploration of potential associations between the concepts of interest. Users can query the database using concepts such as genes, proteins, metabolites, enzymes, pathway, disease concepts, and pharmacological chemicals to retrieve useful associations that may suggest new insights into the problem. We thus provide users the chance to query the compiled abstracts within the framework of concept-based retrieval and extraction system.

For example, a concept query "thalidomide" retrieves all associated concepts from their respective dictionaries found in the analyzed text. These could be displayed either in a graphical or tabular format. These identified concepts co-occur with thalidomide in the PubMed abstracts. The frequency of occurrence of the concepts is shown and the link can be clicked to view the abstracts.

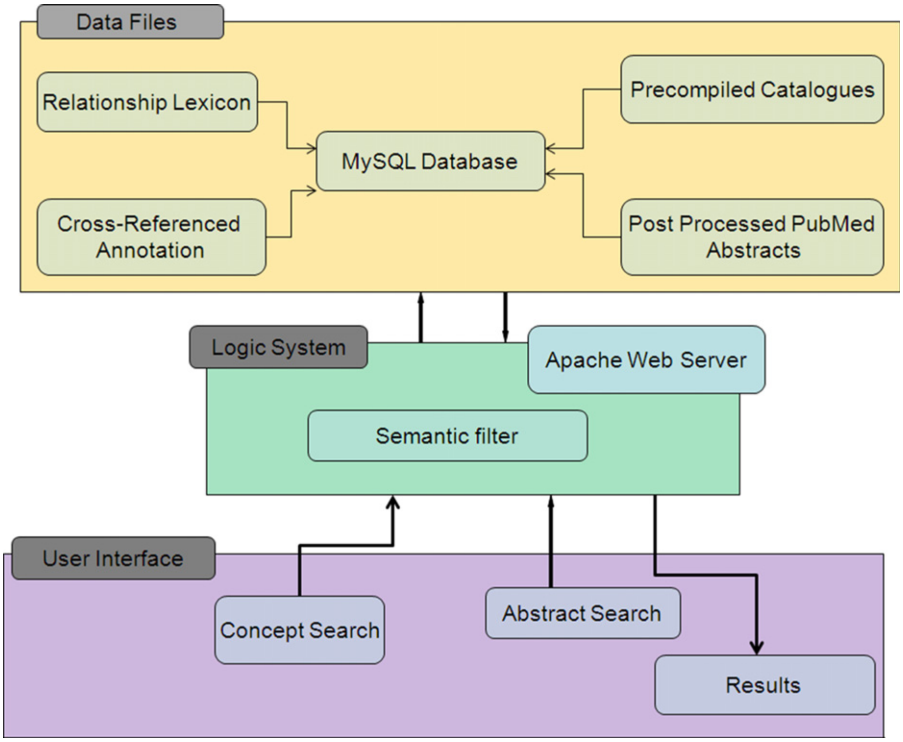

Fig. 1. Schematic diagram of the integrated database of DESHCV. A layout of the DESHCV database architecture showing the relationship between the incorporated data files, logic systems and user query interface.

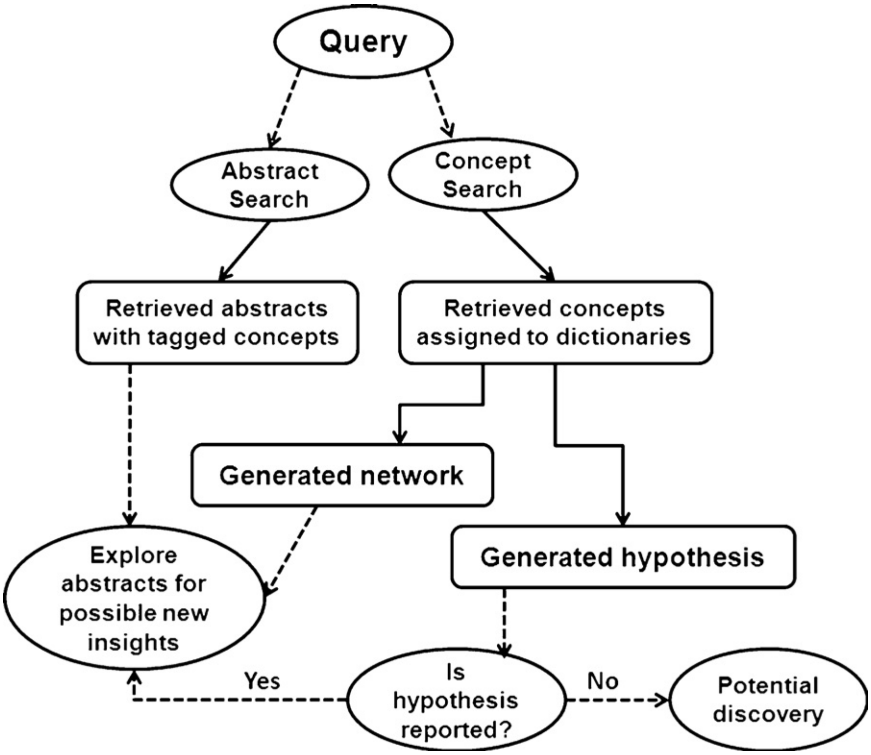
Query
Abstract Search
Concept Search
Retrieved abstracts with tagged concepts
Retrieved concepts assigned to dictionaries
Generated network
Generated hypothesis
Explore abstracts for possible new insights
Is hypothesis reported?
Yes
No
Potential discovery

Fig. 2. DESHCV data flow schema diagram. A structured workflow outlining the various steps and decision-making processes involved in retrieving enriched biological data from DESHCV.

The user has the option of either viewing the abstract with or without tagged concepts. All disease concepts associated with thalidomide are retrieved including chronic hepatitis C, which co-occurs with thalidomide within three abstracts. This result can be displayed in the form of an association map by using the "draw network" generator. The association map is a graph consisting of interacting network of nodes representing thalidomide and its associated concepts. The edges linking the nodes shows the relationship between the concepts and are weighted with the frequency of co-occurrence. To ensure effective interpretation and evaluation of results, users have the option of limiting the number of interactions for display by ignoring links with fewer frequencies. The association map can be resized from A0 to A5 and the detail slider can be used to alter viewing capabilities. For the purpose of this discussion, concepts with less than three links were ignored to obtain high degree of details without obscuring vital information. The association established between thalidomide and chronic hepatitis C is a loose relationship and no inference may be deduced until the literature is manually verified to either accept or reject the relationship (Supplementary Fig. S1). The linking abstracts were manually inspected to ascertain the proposed relationship. Accordingly, thalidomide is a promising novel compound for chronic hepatitis C therapy since thalidomide decreased liver enzymes in six out of eight patients suffering from chronic hepatitis C (Milazzo et al., 2006).

3.2. Abstract queries

The abstract search menu allows users to do keywords searches, use quotes for "phrase search", and Boolean logical operators such as "OR", "AND" or "NOT". An abstract search with the keyword "hypervariability" therefore returns a list of abstracts containing the queried keyword "hypervariability". The retrieved abstracts also contain tagged colour-coded concepts assigned to their respective dictionaries. The abstract search results can serve as rich source of data useful for curation since it contains automatically identified biomedical concepts. Even though manual curation using human expertise could yield high quality results (Muller et al., 2004), it is sometimes fraught with errors since the human eyes may unintentionally overlook useful data. In addition, the rapidly increasing volume of published literature sometimes could render it herculean for researchers to manually inspect and efficiently locate or identify biomedical concepts of interest embedded in literature. This feature of DESHCV is aimed at complimenting the already existing information search, retrieval and extraction resources.

3.3. Evaluation of DES by reproducing a known hypothesis

The performance of DES has been evaluated recently by Sagar et al. (2008) in the context of sodium channels in human. Since it was not possible to evaluate all the concepts embedded in the 5243 documents used in Sagar's analysis, SCN1A was chosen as a reference gene for the analysis. DES accurately identified most of the concepts present in the 131 abstracts associated with SCN1A. The analysis showed both precision and recall for identified concepts from different dictionaries that ranged between 81% and 100% whilst the *F*-measure was between 87% and 100%.

In this report, a different approach was used to evaluate the performance of DES by simulating an already confirmed scientific discovery. Such approach has been used during the implementation of the DAD-system, Swanson's Raynaud's disease-fish oil discovery was simulated to test the performance of the system (Weeber et al., 2000). Swanson's discovery of a hidden connection between disjunct literatures on magnesium and migraine was successfully re-implemented in the LitLinker systems (Meredith et al., 2005). Anni biomedical text-mining tool was used to reproduce a previously published thalidomide-chronic hepatitis C discovery (Jelier et al., 2008). Here, DESHCV is used successfully to simulate the thalidomide-chronic hepatitis C association (Weeber et al., 2003). By clicking on the hypothesis generator, the user can retrieve hypothesis to reveal potential relationships existing between concepts. The hypothesis generator query menu allows users to automatically or manually select categories with which to generate hypothesis. The open discovery approach previously described by Swanson and adapted by Meredith has been employed here. Thalidomide (from "metabolites and enzymes" dictionaries) was used as the starting term to retrieve all linking concepts within the human proteins and genes category. The tumor necrosis factor (TNF) was manually selected as the linking term whilst concepts within the diseases catalogues were defined as the target (Supplementary Fig. S2). The system successfully generated hypothesis to infer potential relationships between thalidomide and the following disease concepts: chronic liver disease, chronic viral hepatitis, chronic persistent hepatitis and chronic renal failure. The system used disjuncted literature between the different concepts to predict implicit relationship amongst them. Chronic liver disease, chronic viral hepatitis and chronic persistent hepatitis are possible name variants and these conditions are implicated in liver failure. By clicking on the test button linking the chronic viral hepatitis and thalidomide retrieved an abstract on a case report concerning thalidomide-associated hepatitis where the patient had medical history including chronic hepatitis C (Fowler and Imrie, 2001). The hypothesis generated by DESHCV was reasonable since the verification of the literature supported the "discovery".

3.4. Generation of thalidomide-amantadine association

DESHCV has also been used to generate potentially new hypotheses, which propose relationships between thalidomide and amantadine as possible combination therapy for chronic viral hepatitis C. All disease concepts associated with thalidomide were retrieved and chronic hepatitis C was selected as linking term with concepts in metabolites and enzymes dictionaries considered as target terms. This hypothesis was tested automatically by checking the auto test radio button. This inferred an implicit relationship between thalidomide and amantadine but no PubMed abstract was retrieved implying a potentially new discovery (Fig. 3). This means that these two concepts do not co-occur in any of the PubMed abstracts. This potential discovery is based on textual analysis of abstracts and not full text papers. The standard treatment for chronic hepatitis C is a combination therapy of pegylated IFN-alpha to elicit immune response and antiviral effect of ribavirin. Patients infected with certain genotypes of HCV do not respond to this treatment, necessitating the need for enhanced combination therapy. Reports on available data concerning triple therapy comprising of pegylated IFN-alpha, ribavirin and protease inhibitors targeting NS3-4SA protease looks promising, and this could become standard treatment feature in the near future (Flisiak and Parfieniuk, 2010; Zeuzem, 2008). The usage of thalidomide in the treatment of chronic hepatitis C unresponsive to alpha-interferon and ribavirin has been investigated (Caseiro, 2006; Milazzo et al., 2006), whilst amantadine has been combined with interferon-alpha plus ribavirin (Chrissafidou and Musch, 2009; von Wagner et al., 2008).

The possibility of a triple therapy for effective management of chronic hepatitis C should prompt researchers on the need to investigate combining any one of the available treatment drugs (IFN-alpha and ribavirin) or possibly protease inhibitors to thalidomide and amantadine as implied by the generated hypotheses. Although the beneficial effect of amantadine and other antiviral drug for the treatment of chronic hepatitis C is controversial and sometimes contentious, potent analogues with less toxicity, augmented pharmacophore and enhanced pharmacokinetics could be explored further. The possibility of augmenting the above therapy with hepatoprotective drug such as silibinin could be explored. Any discovery proposed using biomedical text-mining approach should undergo the rigour of laboratory evaluations and ethical consideration and possibly must conform to existing legislation before usage.

3.5. Systems reports

Comprehensive summary reports generated from DESHCV are made available on the left menu of the user interface. The concept lists report groups concepts by dictionaries and frequency of appearance inside documents. The frequency of document report lists abstract with their total number of embedded tagged concepts. The frequencies of pairs report group pairs of concepts that co-occur in documents. Another useful feature is the document clustering report which displays clusters of concepts sorted by frequency of appearance and is compiled using artificial neural networking algorithm. Its usefulness lies in the fact that similar documents tend to cluster together, thereby allowing

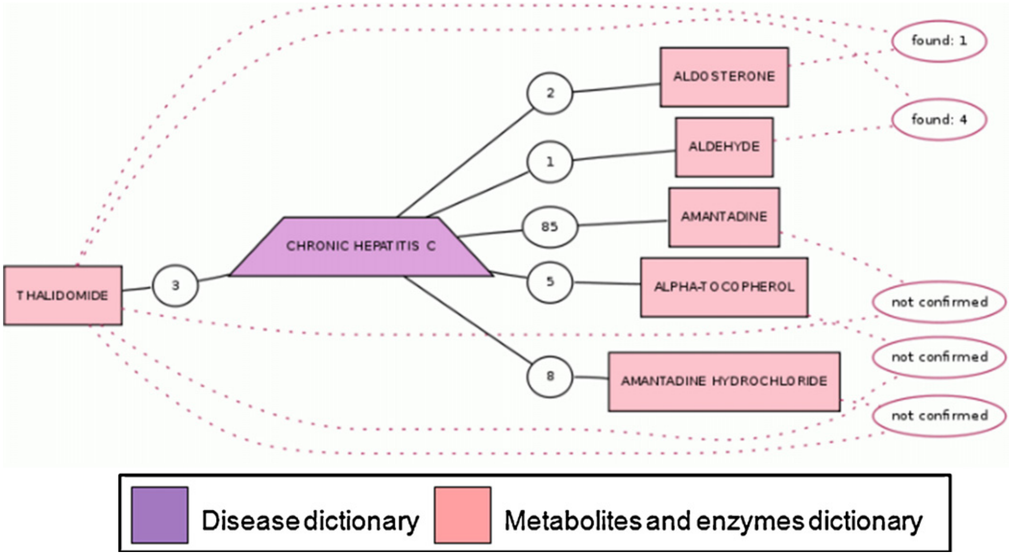

Fig. 3. A diagram displaying thalidomide-amantadine hypothesis. This shows an implicit relationship between thalidomide and amantadine inferring potentially new hypotheses. The biomedical concepts "thalidomide" and "amantadine" belong to the "metabolites and enzymes" dictionaries whilst "chronic hepatitis C" concept belongs to the "disease" dictionary.

biologist to harness potential information amongst clustered abstracts. Recommended readings displays the link to top 10 documents with most concepts, though not manually generated they could give an overview of HCV research to a new researcher in virology. For example, the most recommended reading is an abstract on a review describing in detail the current antiviral drugs in clinical usage (De Clercq, 2004).

4. Limitation

Associations generated between paired concepts are inferred from co-occurrences and may not necessarily relate to any molecular functionality. Textual data is obtained from abstract, which are easy to index. Some details of research are present in full body text and as such vital information may not be reported in abstract. Full text documents were not analyzed in this study.

5. Future directions

Integrating blast and identifier queries to enhance querying capabilities of DESHCV is currently being investigated. The possibility of integrating full text document is currently being explored and could be added to the database as a separate feature. The database would be updated every six months to meet the demands of ever increasing PubMed records related to HCV.

6. Conclusions

We have developed a Hepatitis C Virus customized web-based text-mining resource, which allows researchers to intuitively use the system to get insight into possible novel associations between concepts. We have also used the system successfully to reproduce already published thalidomide-chronic hepatitis C biomedical text-mining discovery (Weeber et al., 2003). DESHCV database is free to use for academic and non-profit purposes.

Acknowledgements

This work was partly supported by grants from the National Research Foundation (South Africa), National Bioinformatics Network (to S.K.K.), and DST/NRF Research Chair (to A.C.). We are also grateful to A.R. and V.B.B. for making available the DES resource.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2010.12.006.

References

Bajic, V.B., Veronika, M., Veladandi, P.S., Meka, A., Heng, M.W., Rajaraman, K., Pan, H., Swarup, S., 2005. Dragon plant biology explorer. A text-mining tool for integrating associations between genetic and biochemical entities with genome annotation and biochemical terms lists. Plant Physiol. 138, 1914–1925.

Caseiro, M.M., 2006. Treatment of chronic hepatitis C in non-responsive patients with pegylated interferon associated with ribavirin and thalidomide: report of six cases of total remission. Rev. Inst. Med. Trop. Sao Paulo 48, 109–112.

Chen, Y., Han, K., 2009. BSFINDER: finding binding sites of HCV proteins using a support vector machine. Protein Pept. Lett. 16, 373–382.

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., Wishart, D.S., 2008. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res. 36, W399–405.

Chrissafidou, A., Musch, E., 2009. Peripheral polyneuropathy and bilateral optic neuropathy during treatment of chronic hepatitis C. Dtsch. Med. Wochenschr 134, 927–930.

Cohen, A.M., Hersh, W.R., 2005. A survey of current work in biomedical text mining. Brief. Bioinform. 6, 57–71.

Cohen, K.B., Hunter, L., 2008. Getting started in text mining. PLoS Comput. Biol. 4, e20.

Combet, C., Garnier, N., Charavay, C., Grando, D., Crisan, D., Lopez, J., Dehne-Garcia, A., Geourjon, C., Bettler, E., Hulo, C., Le Mercier, P., Bartenschlager, R., Diepolder, H., Moradpour, D., Pawlotsky, J.M., Rice, C.M., Trepo, C., Penin, F., Deleage, G., 2007. euHCVdb: the European Hepatitis C Virus database. Nucleic Acids Res. 35, D363–366.

de Chassey, B., Navratil, V., Tafforeau, L., Hiet, M.S., Aublin-Gex, A., Agaugue, S., Meiffren, G., Pradezynski, F., Faria, B.F., Chantier, T., Le Breton, M., Pellet, J., Davoust, N., Mangeot, P.E., Chaboud, A., Penin, F., Jacob, Y., Vidalain, P.O., Vidal, M., Andre, P., Rabourdin-Combe, C., Lotteau, V., 2008. Hepatitis C Virus infection protein network. Mol. Syst. Biol. 4, 230.

De Clercq, E., 2004. Antiviral drugs in current clinical use. J. Clin. Virol. 30, 115–133.

Essack, M., Radovanovic, A., Schaefer, U., Schmeier, S., Seshadri, S.V., Christoffels, A., Kaur, M., Bajic, V.B., 2009. DDEC: dragon database of genes implicated in esophageal cancer. BMC Cancer 9, 219.

Flisiak, R., Parfieniuk, A., 2010. Investigational drugs for hepatitis C. Expert Opin. Invest. Drugs 19, 63–75.

Fowler, R., Imrie, K., 2001. Thalidomide-associated hepatitis: a case report. Am. J. Hematol. 66, 300–302.

Jelier, R., Schuemie, M.J., Veldhoven, A., Dorssers, L.C., Jenster, G., Kors, J.A., 2008. Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol. 9, R96.

Kaur, M., Radovanovic, A., Essack, M., Schaefer, U., Maqungo, M., Kibler, T., Schmeier, S., Christoffels, A., Narasimhan, K., Choolani, M., Bajic, V.B., 2009. Database for exploration of functional context of genes implicated in ovarian cancer. Nucleic Acids Res. 37, D820–823.

Killcoyne, S., Carter, G.W., Smith, J., Boyle, J., 2009. Cytoscape: a community-based framework for network modeling. Methods Mol. Biol. 563, 219–239.

Kuiken, C., Mizokami, M., Deleage, G., Yusim, K., Penin, F., Shin, I.T., Charavay, C., Tao, N., Crisan, D., Grando, D., Dalwani, A., Geourjon, C., Agrawal, A., Combet, C., 2006. Hepatitis C database, principles and utility to researchers. Hepatology 43, 1157–1165.

Kuiken, C., Yusim, K., Boykin, L., Richardson, R., 2005. The Los Alamos hepatitis C sequence database. Bioinformatics 21, 379–384.

Kwofie, S.K., Radovanovic, A., Maqungo, M., Bajic, V.B., Christoffels, A., 2009. Hepatitis C Virus discovery database (HCVdd): a biomedical text mining and relationship exploring knowledgebase. In: ISCB Africa ASBCB Joint Conference on Bioinformatics of Infectious Diseases, December 01–03, 2009, Bamako, Mali.

Lombard, Z., Tiffin, N., Hofmann, O., Bajic, V.B., Hide, W., Ramsay, M., 2007. Computational selection and prioritization of candidate genes for fetal alcohol syndrome. BMC Genomics 8, 389.

Meredith, M., Skeels, Kiera, H., Meliha, Y.Y., Wanda, P., 2005. Interaction design for literature-based discovery. In: CHI '05 Extended Abstracts on Human factors in Computing Systems, April 02–07, 2005, Portland, USA.

Milazzo, L., Biasin, M., Gatti, N., Piacentini, L., Niero, F., Zanone Poma, B., Galli, M., Moroni, M., Clerici, M., Riva, A., 2006. Thalidomide in the treatment of chronic hepatitis C unresponsive to alfa-interferon and ribavirin. Am. J. Gastroenterol. 101, 399–402.

Mitchell, J.A., Aronson, A.R., Mork, J.G., Folk, L.C., Humphrey, S.M., Ward, J.M., 2003. Gene indexing: characterization and analysis of NLM's GeneRIFs. AMIA Annu. Symp. Proc. 460–464.

Muller, H.M., Kenny, E.E., Sternberg, P.W., 2004. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol. 2, e309.

Pan, H., Zuo, L., Choudhary, V., Zhang, Z., Leow, S.H., Chong, F.T., Huang, Y., Ong, V.W., Mohanty, B., Tan, S.L., Krishnan, S.P., Bajic, V.B., 2004. Dragon TF association miner: a system for exploring transcription factor associations through text-mining. Nucleic Acids Res. 32, W230–234.

Sagar, S., Kaur, M., Dawe, A., Seshadri, S.V., Christoffels, A., Schaefer, U., Radovanovic, A., Bajic, V.B., 2008. DDESC: dragon database for exploration of sodium channels in human. BMC Genomics 9, 622.

Shi, L., Campagne, F., 2005. Building a protein name dictionary from full text: a machine learning term extraction approach. BMC Bioinform. 6, 88.

Swanson, D.R., 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect. Biol. Med. 30, 7–18.

Tiffin, N., Kelso, J.F., Powell, A.R., Pan, H., Bajic, V.B., Hide, W.A., 2005. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res. 33, 1544–1552.

von Wagner, M., Hofmann, W.P., Teuber, G., Berg, T., Goeser, T., Spengler, U., Hinrichsen, H., Weidenbach, H., Gerken, G., Manns, M., Buggisch, P., Herrmann, E., Zeuzem, S., 2008. Placebo-controlled trial of 400 mg amantadine combined with peginterferon alfa-2a and ribavirin for 48 weeks in chronic hepatitis C Virus-1 infection. Hepatology 48, 1404–1411.

Weeber, M., Klein, H., Aronson, A.R., Mork, J.G., de Jong-van den Berg, L.T., Vos, R., 2000. Text-based discovery in biomedicine: the architecture of the DAD-system. Proc. AMIA Symp. 903–907.

Weeber, M., Kors, J.A., Mons, B., 2005. Online tools to support literature-based discovery in the life sciences. Brief. Bioinform. 6, 277–286.

Weeber, M., Vos, R., Klein, H., De Jong-Van Den Berg, L.T., Aronson, A.R., Molema, G., 2003. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. J. Am. Med. Inform. Assoc. 10, 252–259.

Yusim, K., Richardson, R., Tao, N., Dalwani, A., Agrawal, A., Szinger, J., Funkhouser, R., Korber, B., Kuiken, C., 2005. Los alamos hepatitis C immunology database. Appl. Bioinform. 4, 217–225.

Zeuzem, S., 2008. Interferon-based therapy for chronic hepatitis C: current and future perspectives. Nat. Clin. Pract. Gastroenterol. Hepatol. 5, 610–622.