Original article

T-HOD: a literature-based candidate gene database for hypertension, obesity and diabetes

Hong-Jie Dai[1], Johnny Chi-Yang Wu[1], Richard Tzong-Han Tsai[2], Wen-Harn Pan[3] and Wen-Lian Hsu[1,]*

[1]Intelligent Agent Systems Lab, Institute of Information Science, Academia Sinica, Taipei, [2]Department of Computer Science & Engineering, Yuan Ze University, Taoyuan and [3]Division of Preventive Medicine and Health Services Research, Institute of Population Health Sciences, National Health Research Institutes, Taipei, Taiwan, Republic of China

***Corresponding author**: Tel: +886-2-2788-3799 ext. 1804; Fax: 886-2-2782-4814; Email: hsu@iis.sinica.edu.tw

Submitted 18 June 2012; Revised and Accepted 19 December 2012

Citation details: Dai,H.-J., Wu,J.C.-Y., Tsai,R.T.-H., *et al*. T-HOD: a literature-based candidate gene database for hypertension, obesity and diabetes. *Database* (2012) Vol. 2012: article ID bas061; doi:10.1093/database/bas061.

Researchers are finding it more and more difficult to follow the changing status of disease candidate genes due to the exponential increase in gene mapping studies. The Text-mined Hypertension, Obesity and Diabetes candidate gene database (T-HOD) is developed to help trace existing research on three kinds of cardiovascular diseases: hypertension, obesity and diabetes, with the last disease categorized into Type 1 and Type 2, by regularly and semiautomatically extracting HOD-related genes from newly published literature. Currently, there are 837, 835 and 821 candidate genes recorded in T-HOD for hypertension, obesity and diabetes, respectively. T-HOD employed the state-of-art text-mining technologies, including a gene/disease identification system and a disease–gene relation extraction system, which can be used to affirm the association of genes with three diseases and provide more evidence for further studies. The primary inputs of T-HOD are the three kinds of diseases, and the output is a list of disease-related genes that can be ranked based on their number of appearance, protein–protein interactions and single-nucleotide polymorphisms. Unlike manually constructed disease gene databases, the content of T-HOD is regularly updated by our text-mining system and verified by domain experts. The interface of T-HOD facilitates easy browsing for users and allows T-HOD curators to verify data efficiently. We believe that T-HOD can help life scientists in search for more disease candidate genes in a less time- and effort-consuming manner.

Database URL: http://bws.iis.sinica.edu.tw/THOD

Introduction

Hypertension, obesity and diabetes (HOD) are three well-known components of metabolic syndromes, which are associated with numerous degenerative complex diseases. The study of HOD diseases has become increasingly difficult because of the diverse factors in disease progression, such as gene variation, chromosomal defects, genetic variations, environmental factors and family history. In most cases, development of these diseases is modulated by the variations of multiple genes and their interactions with environmental factors (1). Therefore, it is challenging to elucidate the pathogenic mechanisms of HOD.

In the past, many small-scale studies have been carried out to find HOD-related genetic variants; however, recent trend is to systematically analyze the collaborative action of multiple genetic variants to understand the pathogenic mechanisms of HOD. Researchers have been using various high-throughput experimental platforms such as microarrays (2) in transcriptomics and co-immunoprecipitation purification and mass spectrometry in proteomics to screen all possible candidate genes (3), generating large amounts of data. To study HOD genetics systematically, it is necessary to integrate the findings of both small-scale studies and high-throughput research. However, there are only a few databases and review papers that compile HOD-related genes from literature.

In the field of diabetes, T1Dbase (4) integrates valuable information on candidate genes from several databases for Type 1 diabetes, while T2D-Db (5) compiles from PubMed human, mouse and rat genes involved in the pathogenesis of Type 2 diabetes. For obesity genetics, the review paper 'The Human Obesity Gene Map: The 2005 Update' (6) lists candidate genes and/or potential loci up until the end of 2005. For hypertension genes, the genetic association database (GAD) (7) lists hundreds of hypertension candidate genes along with genes for several other diseases. All the above mentioned resources were compiled manually. However, due to limited human resources, manually curated databases cannot always be kept up-to-date. In recent years, various groups have proposed using automated text-mining approaches to reduce human effort in constructing and updating such databases (8–12). SNPs3D (11) and PubMeth (8) are two databases that are constructed using text-mining approaches coupled with manual review and annotation steps. SNPs3D compiles candidate genes and single-nucleotide polymorphism (SNP) sites related to cancers, neurodegenerative diseases and metabolic syndromes. PubMeth contains information on DNA methylation for several cancers. These two databases extract gene names that have a high co-occurrence with the target diseases. However, the co-occurrence-based approaches usually tend to yield a huge number of false-positive relations because of the lack of syntactic and semantic analysis.

Our database, Text-mined Hypertension, Obesity and Diabetes candidate gene database (T-HOD), employed the state-of-art text-mining technologies, including a gene identification (GI) system (13, 14), a disease term recognition system and the disease-gene relation extraction system—HypertenGene (15). Because gene names vary a great deal, different genes may contain the same name. Moreover, gene names may be ambiguous and easily confused with terms employed in other research fields. The employed GI system was designed to alleviate the above problems, which was used to recognize gene terms and link them to their corresponding Entrez Gene IDs using a collective entity linking approach (16). For extracting hypertension-related genes, we formulated the task as a binary classification problem in HypertenGene: for each recognized disease–gene pair from sentences in an abstract, determine whether it is a key relation. HypertenGene applies a maximum entropy model with a set of features, such as *n*-gram, chunk, parse tree and template features. We then rank all extracted genes according to their probability as calculated by the model. We extended and optimized the above systems to extract HOD genes in our T-HOD.

Database content and analyses

Figure 1 shows the number of newly discovered hypertension candidate genes by year, which was generated by our T-HOD statistics viewer. Charts for obesity and diabetes can also be viewed by the same viewer. The steeply climbing curves observed in Figure 1 are due to steadily increased number of hypertension-related genetic studies over the past 40 years.

We list the numbers of candidate genes contained in our database and GAD in Table 1. This version of T-HOD was constructed from abstracts recorded in PubMed from 1970 to 2011. Currently, there are 837, 835 and 821 candidate genes and 282, 317 and 258 rs numbers recorded in T-HOD for hypertension, obesity and diabetes, respectively. The rs number means it is officially registered and given a reference SNP identifier by dbSNP. This result reveals that T-HOD contains more candidate genes and SNP sites than other related databases or papers. One of the reasons is that T-HOD has included the most recently published candidate genes. According to the statistics shown in Figure 1, new HOD-related genes are constantly being discovered, and therefore, a continuously updated database is crucial. The other is that our relation extraction method does not rely on frequency of gene disease co-occurrences, which could improve the chance of finding promising but infrequent candidate genes supported by few papers.

T-HOD interface and implementation

Figure 2 shows the interface of T-HOD, which can be divided into four regions. We will elucidate the function of each region in the following section.

Region 1: control bar

Region 1 at the top of the frame contains a pull-down display menu. By clicking on the menu, users can select the disease of interest (Hypertension, Obesity or Type 1/2 diabetes). Users can also decide whether to show specific gene information or use our advanced search function in this region.

Region 2: candidate gene list

After disease selection, Region 2 shows a list of curated candidate genes. Along each candidate gene, the list also displays the number of papers containing evidence sentences and the number of SNPs and number of protein–protein interactions (PPIs) in separate columns. The list

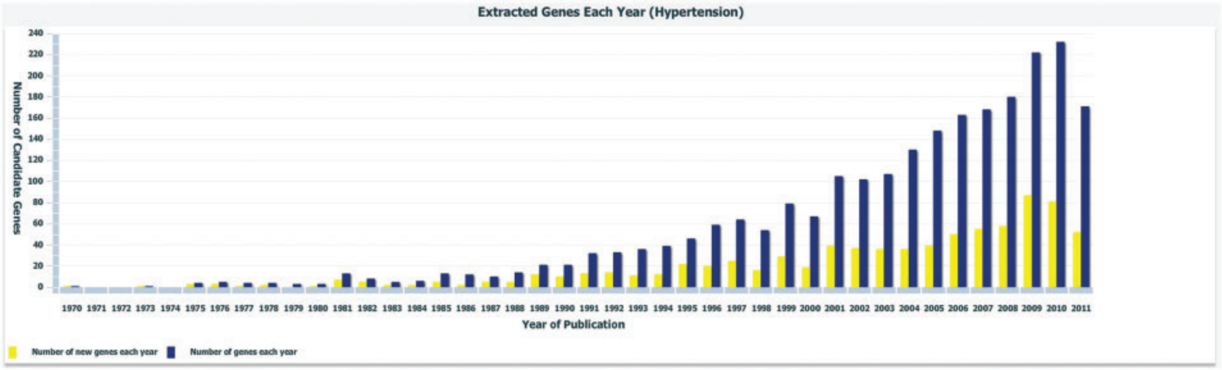
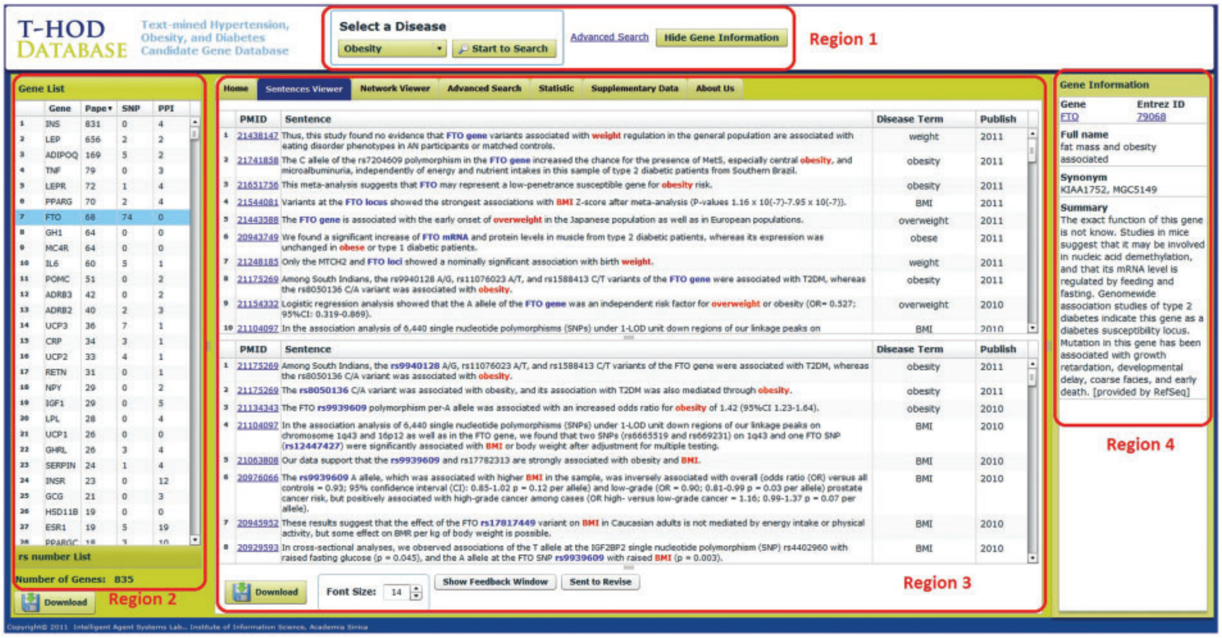


Figure 1. Statistics of the extracted hypertension candidate genes. The blue bars indicate the number of genes extracted each year, while the yellow bars specify the number of novel genes discovered each year.

Table 1. Comparison of candidate genes in T-HOD with GAD

	True positive	False positive	False negative	Precision	Recall	*F*-score	Number of documents
Hypertension	165	42	49	0.797	0.771	0.784	150
Obesity	105	35	29	0.75	0.784	0.766	115
Type 1 diabetes	60	24	27	0.714	0.69	0.702	73
Type 2 diabetes	127	35	37	0.784	0.774	0.779	140
Overall	457	136	142	0.771	0.763	0.767	608



	Gene	Pape▾	SNP	PPI
1	INS	831	0	4
2	LEP	656	2	2
3	ADIPOQ	169	5	2
4	TNF	79	0	3
5	LEPR	72	1	4
6	PPARG	70	2	4
7	FTO	68	74	0
8	GH1	64	0	0
9	MC4R	64	0	0
10	IL6	60	5	1
11	POMC	51	0	2
12	ADRB3	42	0	2
13	ADRB2	40	2	3
14	UCP3	36	7	1
15	CRP	34	3	1
16	UCP2	33	4	1
17	RETN	31	0	1
18	NPY	29	0	2
19	IGF1	29	0	5
20	LPL	28	0	4
21	UCP1	26	0	0
22	GHRL	26	3	4
23	SERPIN	24	1	4
24	INSR	23	0	12
25	GCG	21	0	3
26	HSD11B	19	0	0
27	ESR1	19	5	19
28	PPARGC	18	3	10

Number of Genes: 835

	PMID	Sentence	Disease Term	Publish
1	21438147	Thus, this study found no evidence that **FTO gene** variants associated with **weight** regulation in the general population are associated with eating disorder phenotypes in AN participants or matched controls.	weight	2011
2	21741858	The C allele of the rs7204609 polymorphism in the **FTO gene** increased the chance for the presence of MetS, especially central **obesity**, and microalbuminuria, independently of energy and nutrient intakes in this sample of type 2 diabetic patients from Southern Brazil.	obesity	2011
3	21651756	This meta-analysis suggests that **FTO** may represent a low-penetrance susceptible gene for **obesity** risk.	obesity	2011
4	21544081	Variants at the **FTO locus** showed the strongest associations with **BMI** Z-score after meta-analysis (P-values 1.16 x 10(-7)-7.95 x 10(-7)).	BMI	2011
5	21443588	The **FTO gene** is associated with the early onset of **overweight** in the Japanese population as well as in European populations.	overweight	2011
6	20943749	We found a significant increase of **FTO mRNA** and protein levels in muscle from type 2 diabetic patients, whereas its expression was unchanged in **obese** or type 1 diabetic patients.	obese	2011
7	21248185	Only the MTCH2 and **FTO loci** showed a nominally significant association with birth **weight**.	weight	2011
8	21175269	Among South Indians, the rs9940128 A/G, rs11076023 A/T, and rs1588413 C/T variants of the **FTO gene** were associated with T2DM, whereas the rs8050136 C/A variant was associated with **obesity**.	obesity	2011
9	21154332	Logistic regression analysis showed that the A allele of the **FTO gene** was an independent risk factor for **overweight** or obesity (OR= 0.527; 95%CI: 0.319-0.869).	overweight	2010
10	21104097	In the association analysis of 6,440 single nucleotide polymorphisms (SNPs) under 1-LOD unit down regions of our linkage peaks on	BMI	2010

	PMID	Sentence	Disease Term	Publish
1	21175269	Among South Indians, the **rs9940128** A/G, rs11076023 A/T, and rs1588413 C/T variants of the FTO gene were associated with T2DM, whereas the rs8050136 C/A variant was associated with **obesity**.	obesity	2011
2	21175269	The **rs8050136** C/A variant was associated with obesity, and its association with T2DM was also mediated through **obesity**.	obesity	2011
3	21134343	The FTO **rs9939609** polymorphism per-A allele was associated with an increased odds ratio for **obesity** of 1.42 (95%CI 1.23-1.64).	obesity	2010
4	21104097	In the association analysis of 6,440 single nucleotide polymorphisms (SNPs) under 1-LOD unit down regions of our linkage peaks on chromosome 1q43 and 16p12 as well as in the FTO gene, we found that two SNPs (rs6665519 and rs669231) on 1q43 and one FTO SNP (**rs12447427**) were significantly associated with **BMI** or body weight after adjustment for multiple testing.	BMI	2010
5	21063808	Our data support that the **rs9939609** and rs17782313 are strongly associated with obesity and **BMI**.	BMI	2010
6	20976066	The **rs9939609** A allele, which was associated with higher **BMI** in the sample, was inversely associated with overall (odds ratio (OR) versus all controls = 0.93; 95% confidence interval (CI): 0.85-1.02 p = 0.12 per allele) and low-grade (OR = 0.90; 0.81-0.99 p = 0.03 per allele) prostate cancer risk, but positively associated with high-grade cancer among cases (OR high- versus low-grade cancer = 1.16; 0.99-1.37 p = 0.07 per allele).	BMI	2010
7	20945952	These results suggest that the effect of the FTO **rs17817449** variant on **BMI** in Caucasian adults is not mediated by energy intake or physical activity, but some effect on BMR per kg of body weight is possible.	BMI	2010
8	20929593	In cross-sectional analyses, we observed associations of the T allele at the IGF2BP2 single nucleotide polymorphism (SNP) rs4402960 with raised fasting glucose (p = 0.045), and the A allele at the FTO SNP **rs9939609** with raised **BMI** (p = 0.003).	BMI	2010

Gene	Entrez ID
FTO	79068

Full name
fat mass and obesity associated

Synonym
KIAA1752, MGC5149

Summary
The exact function of this gene is not know. Studies in mice suggest that it may be involved in nucleic acid demethylation, and that its mRNA level is regulated by feeding and fasting. Genomewide association studies of type 2 diabetes indicate this gene as a diabetes susceptibility locus. Mutation in this gene has been associated with growth retardation, developmental delay, coarse facies, and early death. [provided by RefSeq]

Copyright© 2011 Intelligent Agent Systems Lab., Institute of Information Science, Academia Sinica

Figure 2. User interface of the T-HOD. The user interface is divided into four regions for precise introduction.

can be sorted by clicking on the column header, and it is accessible by hitting the 'Download' button at the bottom.

Region 3: viewers

Region 3 provides several viewers, including sentences viewer, network viewer, advanced search option tabs and statistic viewer. Users can switch between different viewers by clicking on the upper tags in this region.

Sentences viewer. The viewer provides curated evidence sentences for each selected candidate gene. If the candidate genes possess corresponding SNP information, the SNP evidence sentence would also be shown below the candidate gene evidence sentences. For each evidence sentence, the sentences viewer shows the source article's PMID and year of publication with highlighted gene and disease terms. Display of the system can be adjusted by changing the font size of the texts. In respect of valuable feedbacks after participating the BioCreative 2012 workshop (17), we constructed a user-friendly interface for users to express their thoughts. In addition, for those who are interested in our database and plan to adopt its use in other studies, the information of T-HOD is attainable by hitting the 'Download' button below the gene list and supporting sentences, allowing them to acquire the disease-related genes and their supporting proof.

Network viewer. Figure 3 shows the network viewer that presents a graphic-based gene–gene network for a selected candidate gene. For each selected candidate gene, the viewer integrates the corresponding PPI information recorded in the Human Protein Reference Database (http://www.hprd.org/) to illustrate the gene–gene network. The viewer allows users to discover the relations among extracted candidate genes. The blue node at the top of the window represents the gene that the user chose in Region 2. To cross examine the candidate genes, the user can double click on the nodes of other candidate genes shown in the same network. Accordingly, the network viewer will redraw the network graph based on the selected gene so that the user can navigate the database more smoothly.

Advanced search. The tab provides advanced search options that allow users to narrow down and specify the desired search results by the following items: publication date, Entrez Gene ID, gene name and PubMed ID.

Statistic viewer. The number of candidate genes and candidate SNP sites contained in T-HOD are summarized in the viewer. The statistic viewer also plots the number of candidate genes and the number of new candidate genes each year in bar charts as shown in Figure 1.

Region 4: gene and SNP information

For each selected candidate gene, the information integrated from different resources is shown in Region 4. In this region, we integrate the following information from Entrez Gene and SNP databases: the gene's official symbol, Entrez Gene ID, full name, synonyms and function summary. Users can also link to the corresponding database for further information.

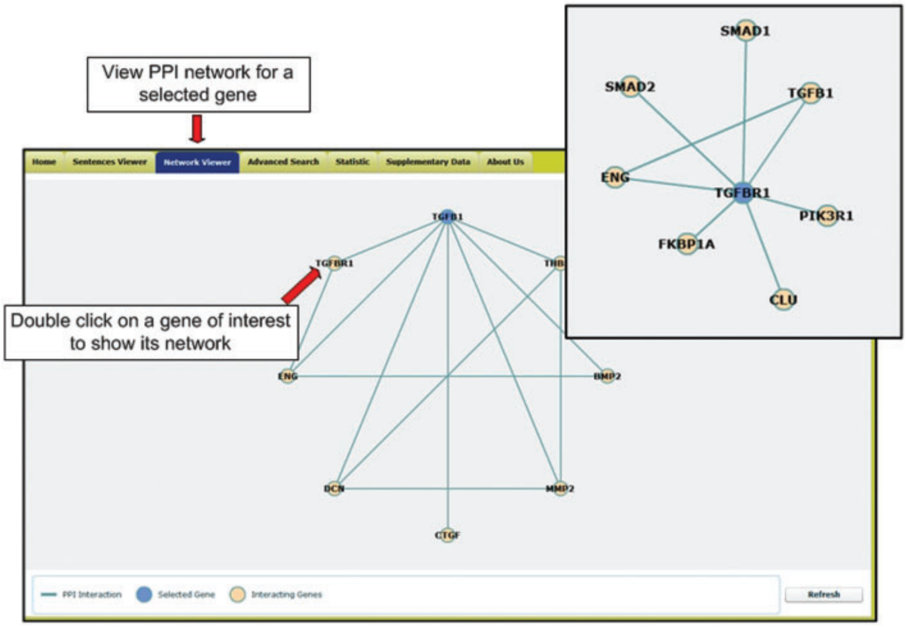


Figure 3. The network viewer of the T-HOD.

Text-mining-based database curation

Figure 4 shows the flowchart for constructing the T-HOD. It comprises three stages: (i) dataset collection and pre-processing, (ii) candidate genes extraction and (iii) manual verification. We collected abstracts on HOD from PubMed and used text-mining techniques to extract HOD candidate genes and SNPs from them. The T-HOD curators verify the extracted list and curate the knowledge into the T-HOD. In the following subsections, we will describe each stage in detail.

Stage 1: dataset collection and pre-processing

In this stage, we collect HOD-related abstracts from PubMed and filter out those that are non-genetic by using a genetic research filter. The filter uses a list of keywords that are frequently used in abstracts of genetic research, such as 'polymorphism', 'alleles', 'variant' and regular expressions as shown in Table 2 to determine whether the abstract is genetically related. The full list of keywords used by our genetic research filter is given in the supplementary materials available in the T-HOD website. The filtered dataset is then pre-processed by several natural language processing components. For example, a sentence splitter from LingPipe (18) is used to split sentences in a genetic-related abstract. Each sentence is then processed by the GENIA tagger (19) to generate part-of-speech information. The information will be used by the following stages for recognizing entities and extracting disease–gene pairs.

Stage 2: candidate genes extraction

In Stage 2, we extract HOD-related candidate genes from the pre-processed dataset through the following steps. First, a disease named entity recognition system is employed to recognize disease terms in a sentence. Second, the GI system is used to recognize and link mentioned genes to their corresponding Entrez Gene IDs. Based on the results of the previous steps, if a disease term and a gene are present in the same sentence, they are extracted as a disease–gene (D-G) candidate pair. Finally, the D-G extraction system determines whether a relation indeed exists within this D-G pair.

Stage 3: manual curation

Although the employed text-mining components have shown satisfactory scores (compare with Table 1), the text-mined candidate genes are examined by all T-HOD curators in Stage 3 to further ensure the quality of the curated content. In this stage, newly extracted candidate genes and their corresponding evidence sentences and abstracts are presented to the T-HOD curators. T-HOD curators review each extracted candidate gene and remove the incorrect results. Currently, the curation process has only been done on abstracts before 2011. Because all annotated error cases are recorded to our SQL database, we can also use such data to enhance our text-mining components in the future.

Results and discussion

Evaluation of a candidate gene database is difficult. Different standards and perspectives can produce different results. In our previous work (15), the employed D-G extraction system has shown satisfactory area under the curve scores of 81.4 and 83% for hypertension and diabetes, respectively. In this study, we compared the performance of T-HOD with the contents of GAD. The bench marking results are shown in Table 1.

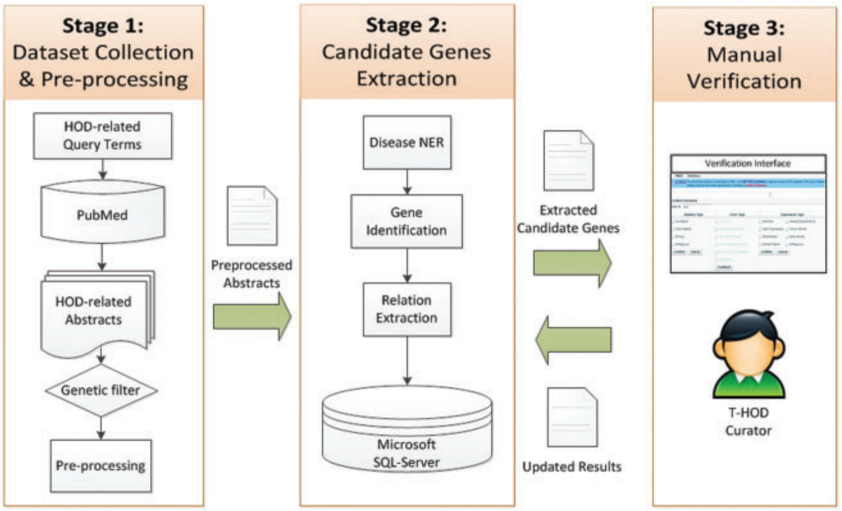


Figure 4. The flowchart of T-HOD construction.

Table 2. Regular expressions used for filtering genetic research articles[a]

Regular expression	Example
\d+\s?[A-Z]/[A-Z]	3123 C/A, 3123C/A
\d+[A-Z]	123C, 825T
[A-Z]\s?\d+	C 123, T 825
[A-Z]\d+[A-Z]	C825T
\d+\s?[A-Z](-{0,2}>|to)\s?[A-Z]	866C–>T, 825 C to T, 926G>C
Amino acid–>*Amino acid*\d+	Gly–>Arg483
[A-Z] to [A-Z]\s?\d+	C to T 825
\d+\s?([A-Z]>[A-Z])	969(G>C), 969 (G>C)
Amino acid(\d+)*Amino acid*	Ile(655)Val
SNP|rs\d{2,}	rs7805828, SNP
(-?\d*)del[A-Z]	−36delG, delT
[ATCG]\.\d+[ATCG]>[ATCG]	C.553G>T

[a] Amino acid refers to the abbreviation of amino acids, such as ‘Gly’ and ‘Arg’.

Disease-related literatures that exist in both databases were chosen for evaluation. Performance for the identification of gene–disease relations in hypertension, obesity and Type 2 diabetes documents all achieved a score around 75%. In contrast, relations of Type 1 diabetes only obtained a score around 70%.

There are several possible reasons that may result in the difference between T-HOD and GAD results. In order for curating a gene–disease relation into our T-HOD, the identity of both the candidate gene and disease terms must be normalized. In the current implementation, identities of gene terms are identified by using a collective entity disambiguation method, whereas disease terms are recognized through a list of vocabularies. Error in the identification of genes and the imperfect list of disease terms utilized may lead to the loss of relations that are present within documents. In addition, the difficulty of extract D-G relation will increase when one or both the disease and gene are expressed with an anaphoric expression. Furthermore, T-HOD only recognizes D-G relations within the same sentence. Currently, cross-sentence relations are not available, but it is a topic worth studying in the future. Finally, determining negations within a sentence is also an important issue. The current T-HOD system can only deal with negation descriptions in basic phrase structures, which may be insufficient in distinguishing more complex negation narratives.

T-HOD in the BioCreative 2012 interactive text-mining task

In addition to the aforementioned discussion, T-HOD was evaluated by the four voluntary curators from Pfizer, MGI Phenotype, GAD and Reactome in the BioCreative 2012 interactive text-mining task (IT) (17). For the user evaluation task of the IT task, we propose the following task for biocurators.

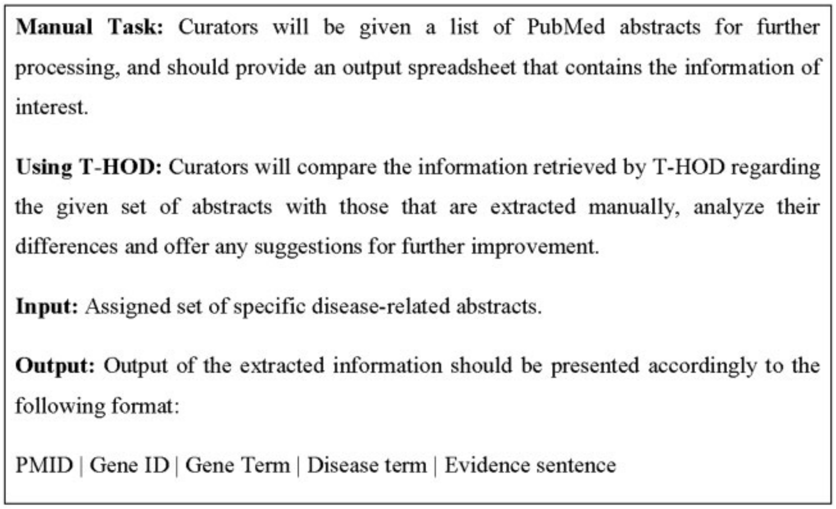
Proposed task for biocuration: HOD curation task

When given a set of abstracts (compiled from those published in 2011) related to a specific disease, a biocurator should:

(1) Identify whether the abstracts contain disease-related gene information (curatable abstracts).
(2) As for curatable abstracts, extract the following information: PMID of the abstract, gene terms and its corresponding gene ID from Entrez Gene, disease terms, relation assertion (positive or negative) and the evidence sentence containing the D-G pair.

Figure 5 shows the formal task descriptions (20). Based on the configuration of the IT task, the dataset for this evaluation consisted of 50 abstracts from 2011 that were randomly selected based on a PubMed search that included the disease name and the keywords ‘associated’ and ‘gene’. Curators were each assigned to curate 25 abstracts manually and 25 using the text-mining tools. In addition, curators validated the literature with the system using a gene-centric approach. The information captured included gene name, Entrez gene ID, disease, relation and evidence sentence. Note that because the abstract set used for the HOD curation task is publications from 2011, it is not yet verified by our T-HOD curators. The biocurator then compares their manually curated results with the text-mined results processed by T-HOD. For the convenience of biocurators in analyzing the results, we developed an interface that directly provides the information of T-HOD in the desired output format with additional PubMed and Entrez Gene links. These results are also available for download. Furthermore, this interface is able of notifying the curators when an abstract is not found in our database or it does not contain any relations of interest. An example of the interface output is shown in Figure 6, which is available at http://bws.iis.sinica.edu.tw:8080/THOD/request_sentence_list.

The results of the IT evaluation are shown in Table 3. Two evaluation results (gene or article centric) are reported. Both are based on the standard precision/recall/F-measure evaluation scheme. The ‘Gene Centric’ scheme compares the linked candidate gene identities with the annotations of curators to calculate the true/false positive/negative counts. In the ‘Article Centric’ scheme, only the case that the identities annotated by biocurators for an article were all identified by our T-HOD is considered as a true positive. The ‘Article Centric’ scheme is stricter than ‘Gene Centric’. Based on the official subjective measure (17), one curator satisfies with the usability of the current



Manual Task: Curators will be given a list of PubMed abstracts for further processing, and should provide an output spreadsheet that contains the information of interest.

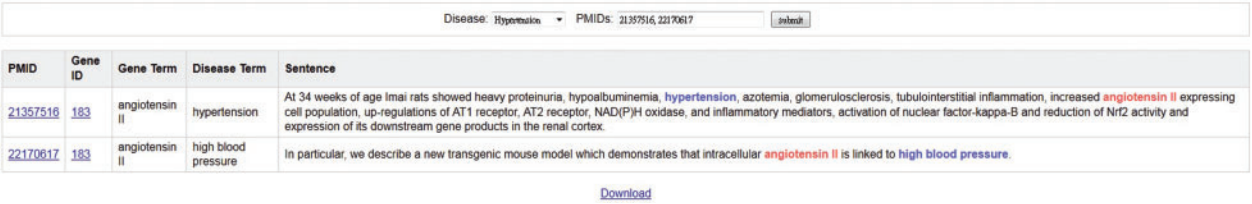
Using T-HOD: Curators will compare the information retrieved by T-HOD regarding the given set of abstracts with those that are extracted manually, analyze their differences and offer any suggestions for further improvement.

Input: Assigned set of specific disease-related abstracts.

Output: Output of the extracted information should be presented accordingly to the following format:

PMID | Gene ID | Gene Term | Disease term | Evidence sentence

Figure 5. Illustration of the proposed task for biocurators.



Disease: Hypertension ▾ PMIDs: 21357516, 22170617 submit

PMID	Gene ID	Gene Term	Disease Term	Sentence
21357516	183	angiotensin II	hypertension	At 34 weeks of age Imai rats showed heavy proteinuria, hypoalbuminemia, hypertension, azotemia, glomerulosclerosis, tubulointerstitial inflammation, increased angiotensin II expressing cell population, up-regulations of AT1 receptor, AT2 receptor, NAD(P)H oxidase, and inflammatory mediators, activation of nuclear factor-kappa-B and reduction of Nrf2 activity and expression of its downstream gene products in the renal cortex.
22170617	183	angiotensin II	high blood pressure	In particular, we describe a new transgenic mouse model which demonstrates that intracellular angiotensin II is linked to high blood pressure.

Figure 6. The interface for biocurators.

Table 3. BioCreative interactive text-mining task evaluation results

	Precision	Recall	*F*-score
Gene centric	0.795	0.7	0.745
Article centric	0.721	0.543	0.620

T-HOD implementation (score: 6.2/7) and feels that T-HOD could help him to complete the curation task (score: 6/7). (Subjective measure was conducted by the BioCreative 2012 IT organizers by using user survey. In the survey, the ranking corresponded to 1 indicating less positive/agreeable and the value is up to 7.) In general, the current T-HOD user interface for database curators is easy to use (the average score: 5.2/7) and could highlight and simplify the curation step (4.5/7). But the overall score (4.1/7) of our database could be improved by enhancing the following responses from these curators. One suggestion was that the database should be navigated through different approaches, instead of solely from the disease point of view. We have modified our advanced search options and look to integrate them into one function that is similar to search engine bars, which can accept all kinds of query terms. Other suggestions are mostly related to the layout of our interface (e.g. display window does not auto-adjust into the proper size of the browser; users are unable to return to their last step of action). These problems are due to the limit of the primary programs and structure of our database. To solve these problems, we are now working on an updated version of the database that contains new information, more user-friendly interface and reorganized programs. Currently, this version is under construction.

Conclusion

T-HOD is regularly updated by our text-mining systems and verified by domain experts. In addition, T-HOD not only extracts candidate gene names but also looks up their Entrez Gene IDs and rs numbers for integrating information of candidate gene properly. In summary, the literature-based candidate gene database consists of HOD-related candidate genes and includes the following features: (i) up-to-date candidate genes and SNPs information; (ii) a context sentence for each extracted gene; (iii) visualization of candidate genes' interaction; (iv) annual statistics for the number of HOD candidate genes and (v) a feedback interface that allows T-HOD curators to comment on our extracted sentences. We hope these features will help users to find and study candidate HOD-related genes systemically. Similar techniques will also be adopted to search for candidate genes for other diseases in the future.

Funding

National Core Facility Program for Biotechnology, Taiwan (Bioinformatics Consortium of Taiwan, NSC100-2319-B-010-002).

References

1. Doris,P.A. (2002) Hypertension genetics, single nucleotide polymorphisms, and the common disease:common variant hypothesis. *Hypertension*, **39**(2 Pt 2), 323–331.
2. Grant,S.F. and Hakonarson,H. (2008) Microarray technology and applications in the arena of genome-wide association. *Clin. Chem.*, **54**, 1116–1124.
3. Sundsten,T. and Ortsater,H. (2009) Proteomics in diabetes research. *Mol. Cell. Endocrinol.*, **297**, 93–103.
4. Hulbert,E.M., Smink,L.J., Adlem,E.C. *et al*. (2007) T1DBase: integration and presentation of complex data for type 1 diabetes research. *Nucleic Acids Res.*, **35**(Database issue), D742–D746.
5. Agrawal,S., Dimitrova,N., Nathan,P. *et al*. (2008) T2D-Db: an integrated platform to study the molecular basis of Type 2 diabetes. *BMC Genomics*, **9**, 320.
6. Rankinen,T., Zuberi,A., Chagnon,Y.C. *et al*. (2006) The human obesity gene map: the 2005 update. *Obseity*, **14**, 529–644.
7. Becker,K.G., Barnes,K.C., Bright,T.J. *et al*. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
8. Ongenaert,M., Van Neste,L., De Meyer,T. *et al*. (2008) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.*, **36**(Database issue), D842–D846.
9. Hahn,U., Wermter,J., Blasczyk,R. *et al*. (2007) Text mining: powering the database revolution. *Nature*, **448**, 130.
10. Fang,Y.C., Huang,H.C., Chen,H.H. *et al*. (2008) TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. *BMC Complement Altern. Med.*, **8**, 58.
11. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
12. Fang,Y.C., Lai,P.T., Dai,H.J. *et al*. (2011) MeInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature. *BMC Bioinformatics*, **12**, 471.
13. Dai,H.-J., Chang,Y.-C., Tsai,R.T.-H. *et al*. (2011) Integration of gene normalization stages and co-reference resolution using a Markov logic network. *Bioinformatics*, **27**, 2586–2594.
14. Dai,H.-J., Lai,P.-T. and Tsai,R.T.-H. (2010) Multistage gene normalization and svm-based ranking for protein interactor extraction in full-text articles. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **7**, 412–420.
15. Tsai,R.T.-H., Lai,P.-T., Dai,H.-J. *et al*. (2009) HypertenGene: extracting key hypertension genes from biomedical literature with position and automatically-generated template features. *BMC Bioinformatics*, **10(**Suppl 15), S9.
16. Dai,H.-J., Tsai,R.T.-H. and Hsu,W.-L. (2011) Entity disambiguation using a Markov-logic network. In: *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP).* Chiang Mai, Thailand, pp. 846–855.
17. Arighi,C.N., Carterette,B., Cohen,K.B. *et al*. (2012) An overview of the BioCreative Workshop 2012 Track III: interactive text mining task. In: *Proceedings of 2012 BioCreative Workshop,* Washington, DC.
18. LingPipe 4.1.0. http://alias-i.com/lingpipe (1 October 2008, date last accessed).
19. Teteisi,Y. and Ji,T. (2006) GENIA Annotation Guidelines for Tokenization and POS Tagging. In: *Technical Report(TR-NLP-UT-2006-4).* Tsujii Laboratory, University of Tokyo.
20. Wu,J.C.-Y., Dai,H.-J., Tsai,R.T.-H. *et al*. (2012) T-HOD: text-mined hypertension, obesity, diabetes candidate gene database. In: *Proceedings of the BioCreative 2012 Workshop,* Washington D.C.