T-HOD (Text-mined Hypertension, Obesity, Diabetes Candidate Gene Database)

Hong-Jie Dai[1,2], Chi-Yang Wu[1], Jian-Ming Chen[1], Richard Tzong-Han Tsai[3*], Wen-Harn Pan[4*], Wen-Lian Hsu[1,2*]
[1]Institute of Information Science, Acdemia Sinica, Taipei, Taiwan, R.O.C.
[2]Dep. of Computer Science, National Tsing-Hua Univ., HsinChu, Taiwan, R.O.C.
[3]Dept. of Computer Science & Engineering, Yuan Ze Univ., Taoyuan, Taiwan, R.O.C.
[4]Institute of Biomedical Sciences, Acdemia Sinica, Taipei, Taiwan, R.O.C.

Background on the system:

Our database, T-HOD, employed the state-of-art text-mining technologies we recently developed, including a named entity recognition-gene normalization (GN) system [1,2] and a disease-gene relation extraction system [3]. T-HOD collected lists of genes that have proven to be relevant to three kinds of cardiovascular diseases – hypertension, obesity and diabetes, with the last disease specified into Type 1 and Type 2. It can be used for affirm the association of genes with these 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 which can be ranked based on their number of appearance, protein-protein interactions (PPI) and single nucleotide polymorphisms (SNPs).

T-HOD interface and implementation:

As shown in Figure 1, T-HOD interface is divided into four regions. We will elucidate the function of each region in the following section, respectively.

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.

* Corresponding authors

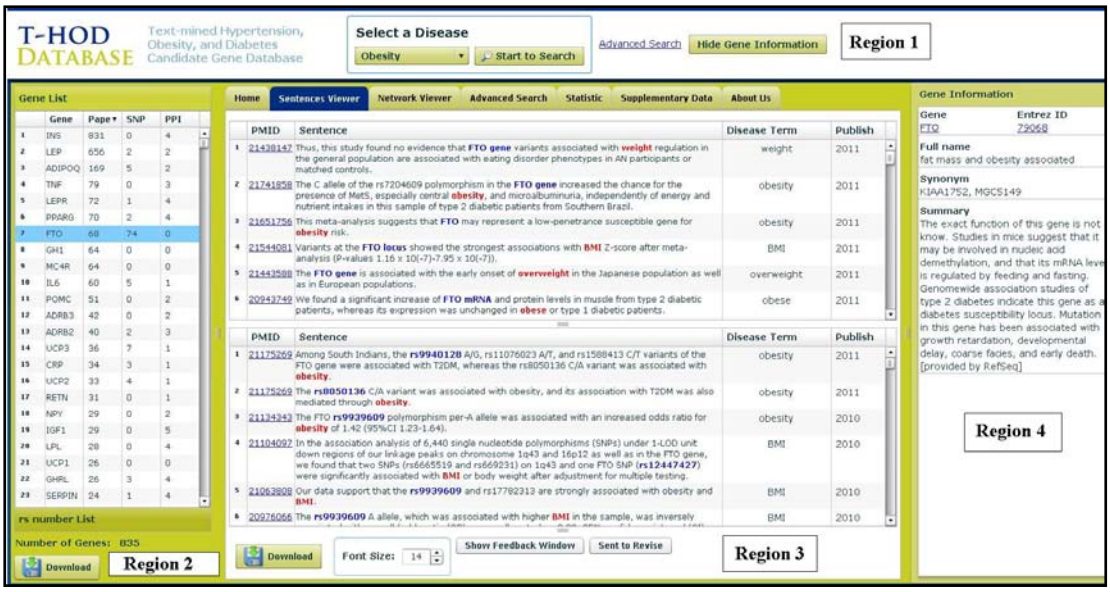

Figure 1. User interface of T-HOD database. The user interface is divided into four regions for precise introduction.

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, as well as the number of SNPs and number of PPIs in separate columns. The list 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 sentence viewer, network viewer, advanced search option tabs, and statistics viewer. Users can switch between different viewers by clicking on the upper tags in this region.

Sentence Viewer: The sentence 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 sentence 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, and in respect of valuable feedbacks, 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, respectively.

Network Viewer: Figure 2 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 HPRD [4] to illustrate the gene-gene network. It 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

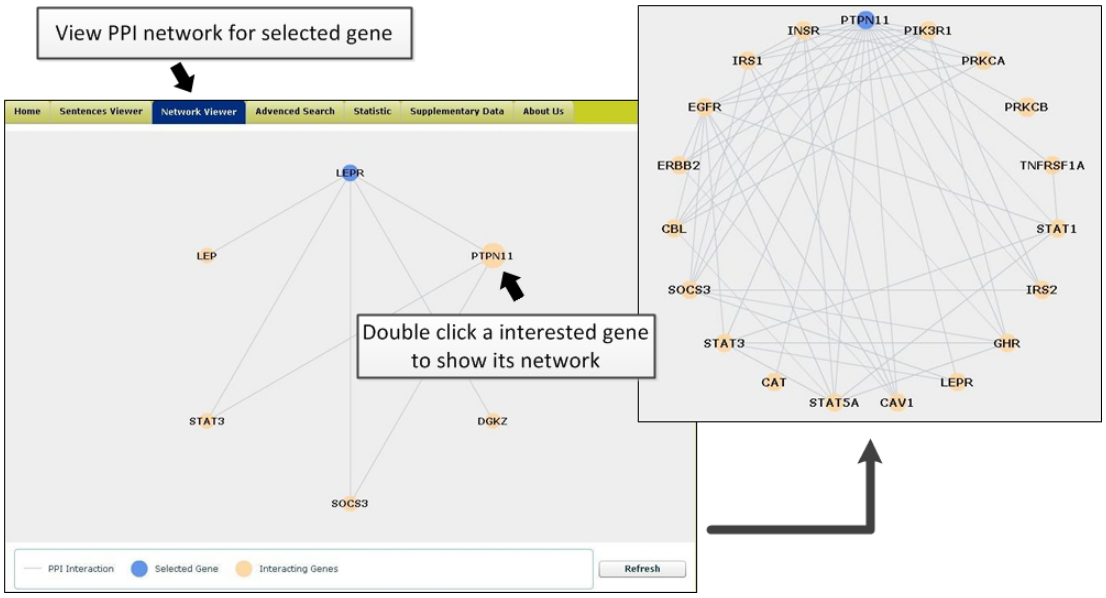

Figure 2. The Network viewer of the T-HOD database

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

Statistics Viewer: The number of candidate genes and candidate SNP sites contained in T-HOD are summarized in the viewer. The statistics viewer also plots the number

of candidate genes and the number of new candidate genes each year in bar charts as shown in Figure 3.

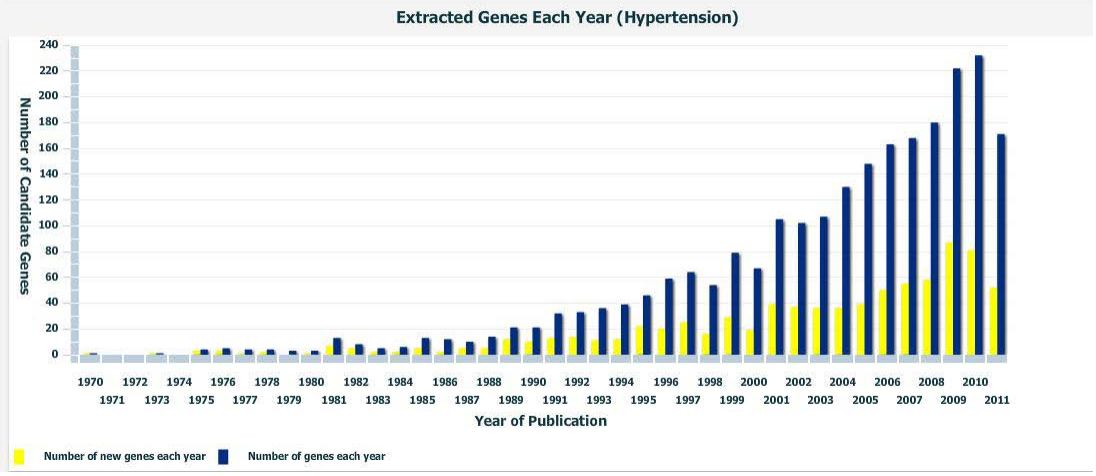

Figure 3. The statistics of the extracted hypertension candidate genes. The blue bars indicate the number of genes extracted each year, while the yellow bars specify the number of novel genes discovered each year.

Region 4: Gene and SNP information

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

Performance

To evaluate our extracted results for obesity candidate genes, we choose the obesity candidate gene list from Rankinen et al.'s "The Human Obesity Gene Map: The 2005 Updated" [5], because they offered a precise gene list. They reported 199 candidate genes related to obesity in human. In Table 1, the result of comparison of extract gene from the review paper and T-HOD is shown. Our system retrieved 492 genes till 2010, and the number of retrieved genes before 2006 is 251. Because the review paper only extracted gene before 2006, we also use genes extracted before 2006 in T-HOD to compare. The recall rate before 2006 of T-HOD is 60.08% (120/199). If we use all curated candidate genes in T-HOD to compare, the recall rate increases to 73.36%.

Table 1. Comparison of obesity candidate genes in T-HOD with Rankinen et al.'s review paper.

T-HOD content	**Extracted candidate genes**	**Overlap with the review paper of obesity**	**Recall**
Before 2006	251	120	60.30%(120/199)
All	492	146	73.36%(146/199)

Proposed task for TRACKIII for T-HOD:

1. When given a set of abstract related to a specific disease:
   a. Identify whether the abstracts contain disease-related gene information (curatable abstracts).
   b. As for curatable abstracts, extract the following information: PMID of the abstract, status of abstract (curatable or not), gene terms and its corresponding gene ID from Entrez Gene, disease terms, relation assertion (positive or negative), and the evidence sentence containing the gene-disease pair.

The task will be run manually and using the T-HOD system.

Manual Task: Users 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 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 | Curatable | Gene name | Gene ID (human) | Disease term | Relation assertion (positive or negative)| Evidence sentence

PMID	Curatable	Gene name	Gene ID	Disease term	Relation assertion	Evidence sentence
21068087	Yes	Renin	5972	hypertension	Positive	The data suggest that upregulation of
21700709	Yes	PPARG	5468	adipogenesis	Positive	Consistent with these results, a defic
19506323	Yes	CTLA4	1493	type 1 diabetes	Positive	Moreover, polymorphisms in CTLA4 have
16249443	Yes	Leptin	3952	insulin resistance	Positive	Thus central administration of leptin

Figure 4. An example of output file.

References

1. Dai H-J, Lai P-T, Tsai RT-H: **Multi-stage gene normalization and SVM-based ranking for protein interactor extraction in full-text articles**. *IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS* 2010, **7**(3):412-420.
2. Tsai RT-H, Sung C-L, Dai H-J, Hung H-C, Sung T-Y, Hsu W-L: **NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition**. *BMC Bioinformatics* 2006, **7**(Suppl 5):S11.
3. Tsai RT-H, Lai P-T, Dai H-J, Huang C-H, Bow Y-Y, Chang Y-C, Pan W-H, Hsu W-L: **HypertenGene: Extracting key hypertension genes from biomedical literature with position and automatically-generated template features**. *BMC Bioinformatics* 2009, **10**(Suppl 15):S9.
4. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al*: **Human Protein Reference Database--2009 update**. *Nucleic Acids Res* 2009, **37**(Database issue):D767-772.
5. Rankinen T, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, Walts B, Pérusse L, Bouchard C: **The Human Obesity Gene Map: The 2005 Update**. *Obseity* 2006, **14**:529-644.