MarkerRIF: An interactive curation system for BioMarker

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

[1]Graduate Institute of BioMedical Informatics, Taipei Medical University, Taipei, Taiwan, R.O.C.

[2]Institute of Information Science, Acdemia Sinica, Taipei, Taiwan, R.O.C.

[3]Department of Computer Science & Engineering, Yuan Ze University, Taoyuan, Taiwan, R.O.C.

[*]Contact: hjdai@tmu.edu.tw

Motivation and Background

Entrez Gene is a repository for gene-specific knowledge of the National Center for Biotechnology Information (NCBI). In addition to general and genomic information, narrative evidences regarding the gene functions within publications can be found in the GeneRIF (Gene Reference Into Function) section. This section provides a platform that enables scientists to share and enrich gene-related functional annotations.

In view of GeneRIF, we developed a browser extension named BioMarker Reference Into Function (MarkerRIF), which allows users to view and edit gene-related functions described in the abstract instantly online. Replacement of the word “Gene” with “Marker” delivers the main purpose of our tool, which is to look for supporting evidence of disease biomarker candidates that were uncovered through previous text-mining processes.

MarkerRIF contains functions including gene name and disease term annotation, linking of the aforementioned terms to their corresponding database, and the extraction of MarkerRIF sentences. Furthermore, a user curation interface is available for curators to curate or modify the extracted RIF sentences. Once confirmed, users can also directly submit the function-describing sentence to our MarkerRIF database or the GeneRIF section of the Entrez Gene database to further elucidate the behavior of these genes. A collection of this knowledge from the literature should provide additional help in the study of biomarkers and may supplement clinical decision making.

MarkerRIF Installation

Google Chrome

1. Download the file “mrif.crx” from http://bws.iis.sinica.edu.tw/MarkerRIF/ or use the direct link http://bws.iis.sinica.edu.tw/MarkerRIF/mrif.crx.
2. Open your Google Chrome browser, and modify its settings from the upper right

panel.

3. Go to the directory “Tools”, and the option “Extensions” under it.
4. Drag and drop “mrif.crx” onto the Extensions page, and a confirmation of adding this tool will appear in a few seconds.
5. When installation is complete, a new page delineating the changes of MarkerRIF will be shown for your reference.
6. Please make sure MarkerRIF is enabled under the Extensions page.

Mozilla Firefox

MarkerRIF for Mozilla Firefox can be downloaded from http://bws.iis.sinica.edu.tw/MarkerRIF/mrif.xpi

Usage Scenario

Browsing with MarkerRIF

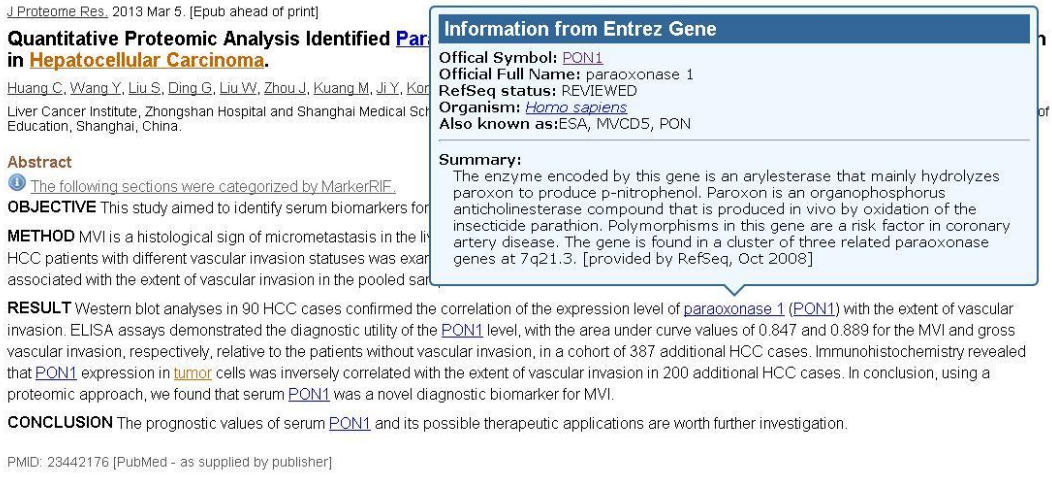
1. Go to the Extensions of the Google Chrome/Firefox Browser.
2. Google Chrome: Click on “Options” under MarkerRIF, and you will be directed to a new page.

   Firefox: After installing MarkerRIF, a X-shaped symbol will appear at the lower right corner of the browser. Single click on the symbol, and a pop-up window will appear.
3. On this page/pop-up window, two steps are required to enable MarkerRIF. First, choose and load the gene list of interest of which you would like to observe its function in abstracts. An example biomarker gene list file can be downloaded from http://bws.iis.sinica.edu.tw/MarkerRIF/default.glist.

Figure 1. Gene list loading and granting Google account access to MarkerRIF

4. Following step 3, grant the access of your Google account to MarkerRIF. Note that unauthorized users will not be able see the function-related sentences provided by MarkerRIF. Figure 1 shows the results after step 3 and 4.
5. Visit PubMed, and search the website with predefined query terms (e.g. liver cancer). We have discussed and created a set of query terms with collaborated curators, and it can be found at http://bws.iis.sinica.edu.tw/MarkerRIF.
6. When viewing the search results, please assign the format of the "Display Settings" as abstracts if you want to see the curation interface.
7. For abstracts that were primarily uncategorized, MarkerRIF can automatically arrange them into four different sections: Objectives, Methods, Results and Conclusions. Automatically sectioned abstracts are preceded by a note to inform researchers that they were categorized by MarkerRIF. Biomedical named entities including gene names, and disease terms are highlighted in the displayed results with different colors, respectively. When the mouse cursor is moved over the recognized entities, a brief pop-up summary of each entity will be displayed as shown in Figure 2. In addition, these entities can be hyperlinked to their corresponding Entrez gene or MeSH pages.

Figure 2. PubMed search results marked up by MarkerRIF.



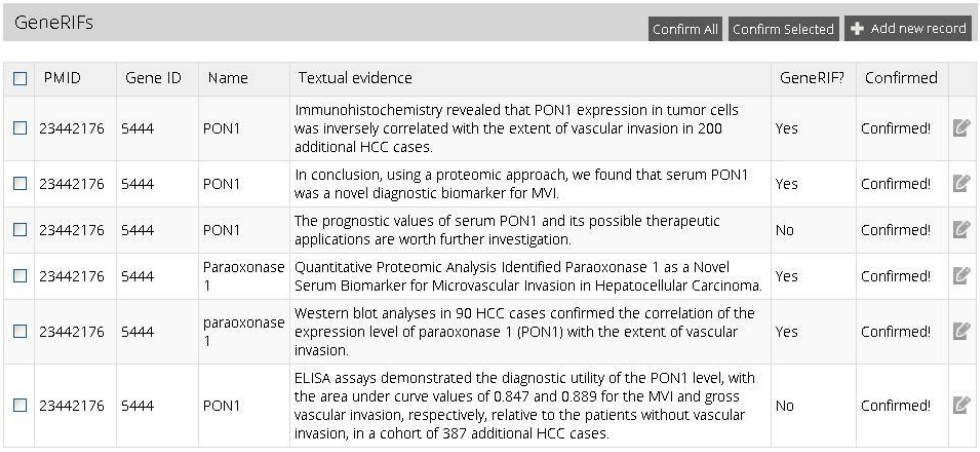
8. As shown in Figure 3, all probable RIF candidate sentences extracted by MarkerRIF will be listed on the bottom of each abstract, with ones that match the gene list arranged ahead of others.

Curation with MarkerRIF

Figure 4 shows the curation interface of MarkerRIF. All sentences containing the gene name in an abstract are analyzed by our sentence determiner and scored by MarkerRIF, and suggested candidate sentences are listed at the bottom. The record of each sentence can be edited using the curation interface, which allows users to add a new

RIF sentence, or to modify the content of an existing textual sentence and validate whether the sentence truly conveys RIF knowledge.

Figure 3. Candidate RIF sentences extracted by MarkerRIF.

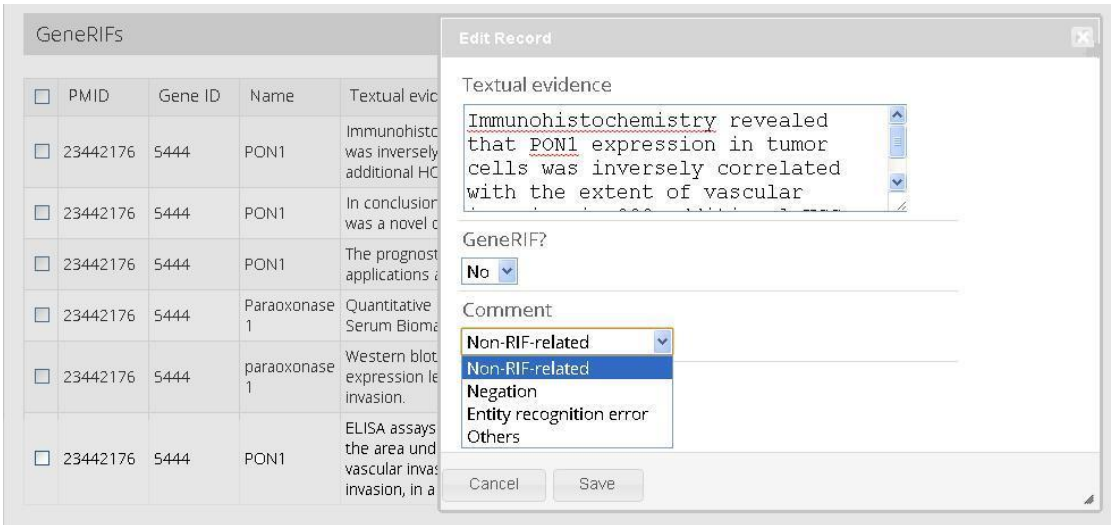


GeneRIFs

Confirm All | Confirm Selected | Add new record

	PMID	Gene ID	Name	Textual evidence	GeneRIF?	Confirmed	
☐	23442176	5444	PON1	Immunohistochemistry revealed that PON1 expression in tumor cells was inversely correlated with the extent of vascular invasion in 200 additional HCC cases.	Yes	Confirmed!	✎
☐	23442176	5444	PON1	In conclusion, using a proteomic approach, we found that serum PON1 was a novel diagnostic biomarker for MVI.	Yes	Confirmed!	✎
☐	23442176	5444	PON1	The prognostic values of serum PON1 and its possible therapeutic applications are worth further investigation.	No	Confirmed!	✎
☐	23442176	5444	Paraoxonase 1	Quantitative Proteomic Analysis Identified Paraoxonase 1 as a Novel Serum Biomarker for Microvascular Invasion in Hepatocellular Carcinoma.	Yes	Confirmed!	✎
☐	23442176	5444	paraoxonase 1	Western blot analyses in 90 HCC cases confirmed the correlation of the expression level of paraoxonase 1 (PON1) with the extent of vascular invasion.	Yes	Confirmed!	✎
☐	23442176	5444	PON1	ELISA assays demonstrated the diagnostic utility of the PON1 level, with the area under curve values of 0.847 and 0.889 for the MVI and gross vascular invasion, respectively, relative to the patients without vascular invasion, in a cohort of 387 additional HCC cases.	No	Confirmed!	✎

Figure 4. Curation interface of MarkerRIF



Furthermore, causes of false positive sentences are generally divided into four categories: non-RIF related, negation, entity recognition error, and others. Once the user confirms and saves the curation results, it will be submitted and stored on our server. Data provided from different users accounts are stored individually and can be used for additional comparison and analysis.

Proposed tasks and curators for the BioCreative user interactive task

When given a list of genes related to a specific disease:

a. Search PubMed using the predefined query terms, and identify whether the abstracts contain disease biomarker-related information (curatable abstracts).
b. As for curatable abstracts, extract the following information: PMID of the

abstract, gene terms and its corresponding gene ID from Entrez Gene, evidence sentence containing RIF information, and relation assertion (descriptive of RIF or not).

The task will be run both manually and using MarkerRIF.

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

Details of the protocol

Input: Assigned set of specific disease-related abstracts.

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

PMID | Gene ID |Gene name | Evidence sentence| Relation assertion

Figure 5. An example of output file.

PMID	Gene ID	Name	Textual evidence	GeneRIF?
23442176	5444	PON1	Immunohistochemistry revealed that PON1 expression in tumor cells was inversely correlated with the extent of vascular invasion in 200 additional HCC cases.	Yes
23442176	5444	PON1	In conclusion, using a proteomic approach, we found that serum PON1 was a novel diagnostic biomarker for MVI.	Yes
23442176	5444	PON1	The prognostic values of serum PON1 and its possible therapeutic applications are worth further investigation.	No
23442176	5444	Paraoxonase 1	Quantitative Proteomic Analysis Identified Paraoxonase 1 as a Novel Serum Biomarker for Microvascular Invasion in Hepatocellular Carcinoma.	Yes

Technical Details

Text-mining web server and system performance

The text-mining server comprises three REpresentational State Transfer (REST) architectural web services.

***Section Categorizer*:** The section categorizer demarcates abstracts into different paragraphs regarding their content. For a given abstract, if PubMed or the pre-sectioned check uncovers that the abstract does not contain obvious section tags, such as ‘Objective’ and ‘Conclusion’, a machine learning-based categorizer [1] is employed to dissect the given abstract.

***Named Entity Tagger*:** The service include two named entity taggers. The first is a machine learning-based gene mention tagger [2], which labels gene names in abstracts. Following entity recognition, an entity normalization module normalizes the

found gene names to their corresponding Entrez Gene database identifiers using a multi-stage approach [3]. Our gene mention tagger achieved an F-score of 86.24% on the BioCreAtIvE II corpus [4, 5]. The performance of our normalization system was evaluated on our instace-based gene mention linking corpus [6], and achieved F-scores of 0.856 and 0.71 for human genes on the article-wide and the instance-based levels, respectively. For cross species evaluation, it achieved the highest area under the precision/recall curve score (0.58) on the BioCreative II.5 interactor normalization dataset [7] and the threshold average precision score of 0.413, which used the median of the confidence scores among all 20th instances as the threshold; The system ranked second in the BioCreative III gene normalization dataset[8].

The second tagger is a dictionary-based disease name tagger based on the maximum matching algorithm. We compiled a dictionary of about 40,000 disease terms with corresponding unique identifiers from the MeSH database. It achieved a satisfactory F-score of 83.4% on the Jimeno *et al.*'s corpus [9].

***Sentence Determiner*:** The sentence determiner provides evidence sentences for genes of interest at the bottom of the abstract. A list of RIF related terms, such as "downregulate" and "induce", is organized, and sentences containing both the gene name and RIF related terms are extracted and ranked by a machine learning model. Several works have proposed effective features in GeneRIF indexing, such as [10, 11]. This work focuses on biomarker-related narratives. We hope to evaluate the effective of the employed features in the specific task by participating the interactive track. In respect of valuable feedbacks, we constructed a user friendly interface for users to curate these sentences and express their thoughts.

MarkerRIF client interface and database

The client interface of MarkerRIF is mainly written in JavaScript with Google Chrome application programming interface and the add-on software development kit of Mozilla Firefox. The OAuth 2.0 authorization framework[1] is employed to obtain curators' profile information to reduce the effort of user registration.

The MarkerRIF database is set up on a Windows server with ASP.NET, MongoDB and SQL server.

[1] http://tools.ietf.org/html/rfc6749

Checklist for the system requirements of BioCreative IV IAT task

Requirement	Support?
a. System should be compatible with the most commonly used web browsers.	Google Chrome and Mozilla Firefox
b. System should highlight entities and relationships (if applicable) relevant to the annotation task.	v
c. User should be able to edit the text mining results by correcting errors or adding missing information, and should be able to export the corrected data.	Once logged in, user can add/edit the results.
d. Use standard input and output formats.	Input: PubMed IDs/Any query terms
e. Full-text processing.	n/a
f. Interactive disambiguation of domain entities.	n/a
g. Ability to filter/sort the results according to different criteria; rank results based on what is more relevant to the user.	Rank results based on the user-provided gene list.
h. On/off for text mining tool, allowing manual annotation in off mode.	User can disable the tool through the browser setting.
i. Record time, be able to record time of curation session for each user (need log in as well).	Log in support; Will support the recording of time.
j. Load curation suggestions or warnings for display during curation.	n/a
k. Upload gene list or ontology term list for focused curation.	v

References

1. Lin RTK, Dai H-J, Bow Y-Y, Chiu JL-T, Tsai RT-H: **Using conditional random fields for result identification in biomedical abstracts** *Integrated Computer-Aided Engineering* 2009, **16**(4):339-352.
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. Dai H-J, Lai P-T, Tsai RT-H: **Multistage Gene Normalization and SVM-Based Ranking for Protein Interactor Extraction in Full-Text Articles**. *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY*

AND BIOINFORMATICS 2010, **7**(3):412-420.

4. Smith L, Tanabe LK, Ando RJn, Kuo C-J, Chung I-F, Hsu C-N, Lin Y-S, Klinger R, Friedrich CM, Ganchev K *et al*: **Overview of BioCreative II gene mention recognition**. *Genome Biology* 2008, **9**(Suppl 2):S2.
5. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J *et al*: **Overview of BioCreative II gene normalization**. *Genome Biology* 2008, **9**(Suppl 2):S3.
6. Dai H-J, Chang Y-C, Tsai RT-H, Hsu W-L: **Integration of gene normalization stages and co-reference resolution using a Markov logic network**. *Bioinformatics* 2011, **27**(18):2586-2594.
7. Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, Valencia A: **An Overview of BioCreative II.5**. *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS* 2010, **7**(3):385-399.
8. Lu Z, Kao H-Y, Wei C-H, Huang M, Liu J, Hsu C-JKC-N, Tsai RT-H, Dai H-J, Okazaki N, Cho H-C *et al*: **The gene normalization task in BioCreative III**. *BMC Bioinformatics* 2011, **12**(Suppl 9):S2.
9. Jimeno A, Jimenez-Ruiz E, Lee V, Gaudan S, Berlanga R, Rebholz-Schuhmann D: **Assessment of disease named entity recognition on a corpus of annotated sentences**. *BMC Bioinformatics* 2008, **9**(Suppl 3):S3.
10. Jimeno-Yepes A, Sticco J, Mork J, Aronson A: **GeneRIF indexing: sentence selection based on machine learning**. *BMC Bioinformatics* 2013, **14**(1):171.
11. Lu Z, Cohen KB, Hunter L: **Finding GeneRIFs via gene ontology annotations**. *Pac Symp Biocomput* 2006:52-63.