MarkerRIF: An Interactive Curation System for Biomarker

Hong-Jie Dai¹*, Chi-Yang Wu², Wei-San Lin¹, Richard Tzong-Han Tsai³, Wen-Lian Hsu²

¹Graduate Institute of BioMedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan, R.O.C., ²Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C., ³Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan, R.O.C.

*Corresponding author: E-mail: hjdai@tmu.edu.tw

Abstract
Disease-related biomedical researches nowadays focus on uncovering biomarkers, which are genes and protein that can act as indicators of the biological state of the ongoing disorder. Observing the expression and behavior of biomarkers can greatly benefit clinical researches and decisions. To efficiently and precisely extract biomarker-related knowledge buried within biomedical texts, we developed a text mining-based curation system named MarkerRIF, which allows curators to retrieve biomarker-related narrations and store their annotations directly while browsing through PubMed.

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**
1. Download the file “mrif.xpi” from http://bws.iis.sinica.edu.tw/MarkerRIF/ or use the direct link http://bws.iis.sinica.edu.tw/MarkerRIF/mrif.xpi
2. Open your Firefox browser. Go to the directory “Tools”, and the option “Extensions” under it.
3. Drag and drop “mrif.xpi” onto the Extensions page, and a confirmation of adding this tool will appear in a few seconds.
4. After installing MarkerRIF, an X-shaped symbol will appear at the lower right corner of the browser.
5. Please make sure MarkerRIF is enabled under the Extensions page.

**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.
4. Following step 3, grant the access of your Google account to MarkerRIF.
- Google Chrome: On the same page where you loaded the gene list, click on “Grant Google Access” and you will be directed to a new page. Accept the request of MarkerRIF, and you will be redirected to the MarkerRIF extension page with your account name at the top.
- Firefox: On the same pop-up window where you loaded the gene list, click on “Grant Google Access” and you will be directed to a new page. Accept the request of MarkerRIF, and you will be redirected to the MarkerRIF pop-up window with your account name at the top.
- An alternative way of granting account access is also provided on http://bws.iis.sinica.edu.tw/MarkerRIF/Account/Register
- Note that after the grant, MarkerRIF can only access your full name and email information associated with your Google account, no further information will be accessed by MarkerRIF. Unauthorized users will not be able to see the function-related sentences provided by MarkerRIF. Figure 1 shows the results after step 3 and 4.

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

Welcome Johnny Wu

<table>
<thead>
<tr>
<th>Choose File</th>
<th>default gist</th>
<th>Load</th>
</tr>
</thead>
</table>

- default gist (m8) - 716 bytes, last modified: 5/30/2013
  - 57016 (aldo-keto reductase family 1, member B10 (aldo reductase))
  - 51280 (golgi membrane protein 1)
  - 8842 (prominin 1)
  - 1116 (chitinase 3-like 1 (cartilage matrix chitinase-3))
  - 14734 (glypican 3)
  - 4072 (epithelial cell adhesion molecule)
  - 2719 (glypican 3)
  - 1499 (catenin (cadherin-associated protein), beta 1, B8E10)
  - 3569 (interleukin 6 (interferon, beta 2))
  - 7015 (telomerase reverse transcriptase)
  - 30158 (hepatoma-derived growth factor (high-mobility group protein 1-like))
  - 1737 (dihydrodihydrolase S-acetyltransferase)
  - 174 (alpha-fetoprotein)
  - 11576 (alpha fetoprotein)
  - 7422 (vascular endothelial growth factor A)
  - 213 (serum albumin)
  - 7157 (tumor protein p53)
  - 1261665 (telomerase)
  - 3481 (nasal-like growth factor 2 (somatomedin A))
  - 4634 (TCAKd)

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.

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.
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**

Users will be given a list of genes related to liver cancer, along with a total of three different sets of abstracts. For these abstracts, please 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 assigned a set of PubMed abstracts for further processing, and should submit their annotations manually at http://bws.iis.sinica.edu.tw/MarkerRIF/Annotation/Create as shown in Figure 5.
• Using MarkerRIF: In contrast to the manual task, curators will extract the information of interest from the two other set of abstracts with the assistance of MarkerRIF, The curated results will be stored and accessible through the MarkerRIF website upon logging in (http://bws.iis.sinica.edu.tw/MarkerRIF/Account/Login). Curators can then compare and analyze the differences between the two approaches, and offer suggestions for further improvement.

Figure 5. The manual curation interface.

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

Curation dataset selection
To perform the proposed curation task, a curation dataset along with a list containing genes of interest is provided. The gene list contains probable liver cancer biomarkers collected from several review papers (12, 13, 14). The curation dataset consists of 190 abstracts retrieved from PubMed using two different query terms. One is the predefined query listed on our website
(((blood[Title/Abstract] OR serum[Title/Abstract] OR urine[Title/Abstract]) AND clinical[Title/Abstract]) OR diagnosis[Title/Abstract]) AND liver cancer[Title/Abstract],

and the other being

(carcinoma, hepatocellular[MeSH Terms]) AND biomarker.

The former query term is defined by domain experts in search of their information of interest within abstracts, and the latter is a more straightforward query used to look for liver cancer biomarkers. The 190 abstracts are divided into three sets in correspondence with the three curators participating in the full biocuration, with each containing 63, 63, 64 abstracts, respectively. Each curator will perform complete manual curation on one set, and MarkerRIF-assisted curation on the other two sets. Due to the limit of time, we have asked the curators to only curate the first 30 abstracts of each set, with a total of 90 curated abstracts for each curator.

**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 instance-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.

**Preliminary Curation Results**

A total of three curators were involved in the full participation of the biocuration track. Three sets of liver cancer-related abstracts, each containing 30 abstracts, were assigned to the curators for manual or MarkerRIF-assisted curation, respectively. Judging by current annotations, the notion and function of MarkerRIF were well understood by the participating curators. For instance, the sentence “Additionally, the expression characteristics of annex A2 during hepatocarcinogenesis were detected in p21-HBx gene knockin transgenic mice model.” were deemed as “Non-RIF-related” regarding annexin A2 by two of the curators. The sentence “Between January 2003 and December 2005, we enrolled 115 treatment-naive patients who received TACE as an initial treatment modality.” was marked as “Entity recognition error”, since TACE is a treatment rather than a gene in this case. As for the sentence “Expression of BRM mRNA, but not BRG1 mRNA, was significantly reduced in primary HCC tumours, compared to non-tumour tissue counterparts.”, it was considered as “Negation” regarding BRG1, since its expression remains unaffected in primary HCC tumours. After curators complete the task, we will analyze the inter-annotator agreement to validate the consistency of these annotations and then report the overall evaluation results of our biomarker-related sentence extraction.

Acknowledgements
This work was supported by the National Science Council of Taiwan under the grant number NSC102-2319-B-010-002, NSC-102-2218-E-038-001, and the Taipei Medical University under the grant number TMU101-AE1-B55.

References