MET Pathway in PubMed: A Pathway Visualization and Curation System

Hong-Jie Dai[1], Chu-Hsien Su[2], Po-Ting Lai[2,3], Ming-Siang Huang[2], Nai-Wen Chang[2,4], Wen-Lian Hsu[2]

[1]Department of Computer Science and Information Engineering, National Taitung University, Taiwan, R.O.C.
[2]Institute of Information Science, Academia Sinica, Taiwan, R.O.C.
[3]Department of Computer Science, National Tsing-Hua University, Taiwan, R.O.C.
[4]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taiwan, R.O.C.
hjdai@nttu.edu.tw
jason@iis.sinica.edu.tw
s102062802@m102.nthu.edu.tw
elephant52381@iis.sinica.edu.tw
d00945020@ntu.edu.tw
hsu@iis.sinica.edu.tw

Abstract. Metastasis is the spread of a cancer/tumor from one organ to another organ, and it is the most dangerous stage in cancer progression, which causes more than 90% of cancer deaths. To improve the understanding of the complicated cellular mechanisms underlying, studies on the signaling pathways are required. To this end a metastasis pathway database, MetastasisWay, has been developed with the goal to store relations among genes, cancers, tissues and organ of metastasis mentioned in the large volume of literature and integrate them to construct a metastasis pathway through text-mining techniques. In light of the requirement to facilitate the curation process for pathway information, a METastasisway curation tool (MET) was developed as a Chrome browser extension which allows curators to review and edit concepts and relations related to metastasis directly in PubMed. PubMed users can view the metastatic pathways integrated from the large collection of research papers.

Keywords: Metastatic Pathway Extraction; Text Mining; Database Curation

1 Introduction

Metastasis refers to the spread of a cancer from its primary site to other parts of the body (secondary sites), while maintaining its malignant growth. Metastasis is often the major concern of patients and clinicians, as it results in the death of over 90% of cancer patients [1]. However, predication of metastasis is a highly challenging task due to the dynamic nature of cancers. Two tumors with the exact same diagnosis may differ in their progression, as one moves to a secondary site but the other does not. Recently, the increasing awareness of biological signaling pathways and their role in

metastasis has enabled life scientists to acquire a more comprehensive overview of the metastatic process. Studies have supported the potential use of gene-specific target therapies in treating metastasis. Additional clinical trials will then be conducted to validate this finding by examining drug-treated patient samples.

Increased understanding of the roles of genes in the metastatic mechanism can lead to improved treatment of cancer patients through the control of metastasis. However, the complexity of gene-cancer interactions stands as the major obstacle that prevents insight into these relations. In light of this, we developed MET[1], the interactive curation tool for our metastasis pathway database MetastasisWay. Through the using of MET, users of PubMed can easily access the pathway information of our database related to the accessed abstracts. Users can further contribute their reading about metastatic pathway by registering as a voluntary curator on our website[2]. Following our previous success in developing a text mining-assisted curation system, a browser extension to assist biomarker curation [2], MET is implemented as a Chrome browser extension to make it easy to install and update while keeping the entire curation process in PubMed.

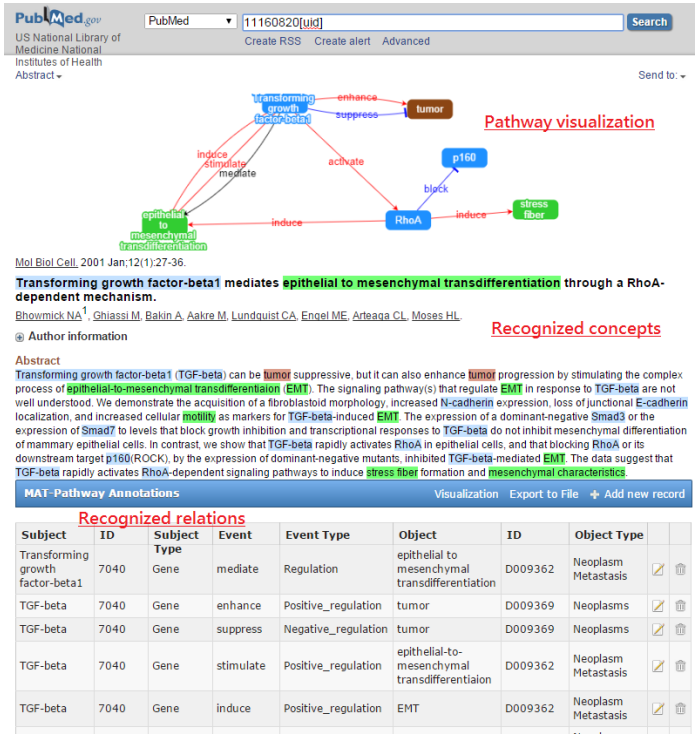

Fig. 1. The article (PMID: 11160820) processed by MET.

[1] MET stands for METastasisway curation tool.
[2] http://btm.tmu.edu.tw/metastasisway

2 Main Features of MET

The main features of MET include (1) display of the text-mined and user-curated recognition and normalization results of the metastasis-related concept terms, such as genes and cancers, described in an abstract, (2) illustration of the extracted metastasis relation information as a metastasis pathway diagram, and (3) the curation interface for curators to update the recognition and relation extraction results. Fig. 1 illustrates the concept recognition and pathway visualization results for the abstract (PMID: 11160820) on the PubMed web site.

When the mouse cursor is moved over the recognized concepts, a brief pop-up summary of each concept will be displayed as shown in Fig. 2. The pathway visualized by MET above the abstract is constructed based on the information of the curation table below the abstract. Users can zoom in to/out of the pathway or rearrange the nodes of the pathway by using their mouse cursor. The curated pathway information recorded in the curation table can be downloaded by clicking the "Export to File" button.

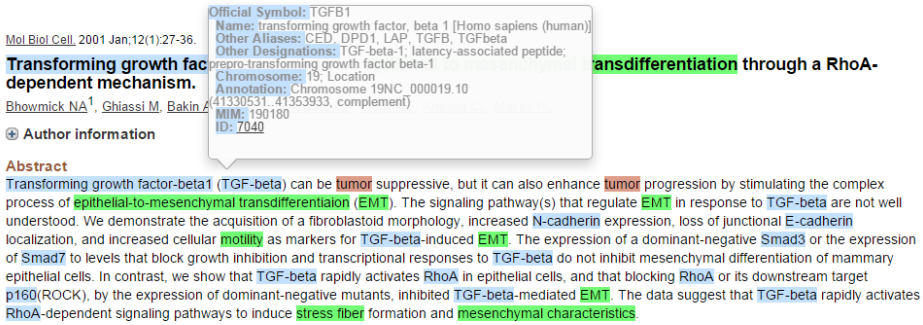

Fig. 2. Detail information about the recognized concept.

In order to generate the result shown in Fig. 1, MET sends the article information to our text mining services. The text mining services support a wide range of biomedical concepts including gene, microRNA, neoplasm metastasis, cytoskeleton, cell movement, cell adhesion, neoplasms, tissues and organ. Based on the recognized concepts, the relations among them are determined and sent back to MET for visualizing the pathway in the client side browser. If the user is registered as a curator, the curation function is available for him to modify the extracted concepts and relations.

In MET curators can modify the following properties of the recognized concepts:

1. The boundaries of the concept
2. The concept type
3. The normalized database ID for the concept

Fig. 3 shows the curation interface for the recognized concepts in MET. A curator can directly update the concept annotation through the "Edit" button in the pop-up summary window (Fig. 3.A). In the editing mode, the curator can change the concept

type by selecting the correct one from the dropdown list (Fig. 3.B). He can also assign an Entrez Gene ID for the concept if it is a gene/protein/microRNA concept. For cancer name the corresponding MeSH term ID can be assigned. The "Delete" button is used to remove the recognized concept (Fig. 3.A). The curator can also add a new concept by first highlighting the words and then selecting its concept type through the curation interface (Fig. 3.C, 3.D).

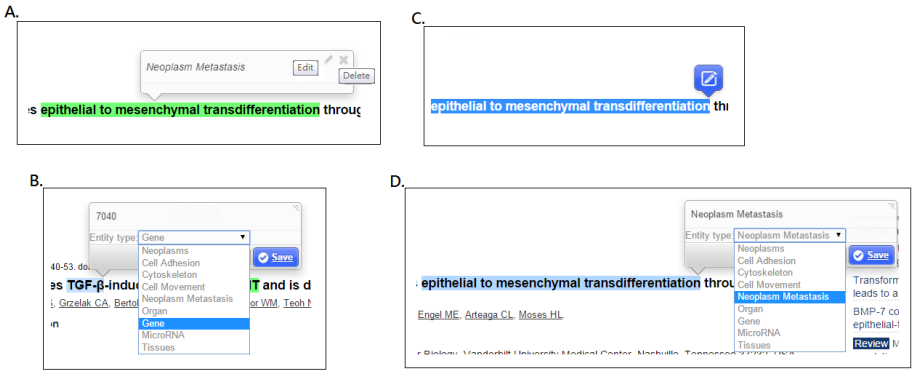

Fig. 3. The curation interface for the recognized concept

As shown in Fig. 1, the relations extracted by our text mining services are listed in the curation table below the abstract. The table shows the detail components of each relation. Curators can edit the participants of the relation and the type of event of the selected relation by using the editing function of the curation table shown on the left side of Fig. 4. They can also add relations that were not recognized by our text mining services by using the "Add new record" interface of the curation table as shown in the right side of Fig. 4. Once curators confirm and save the curation results, the results will be submitted and stored in the MetastasisWay database.

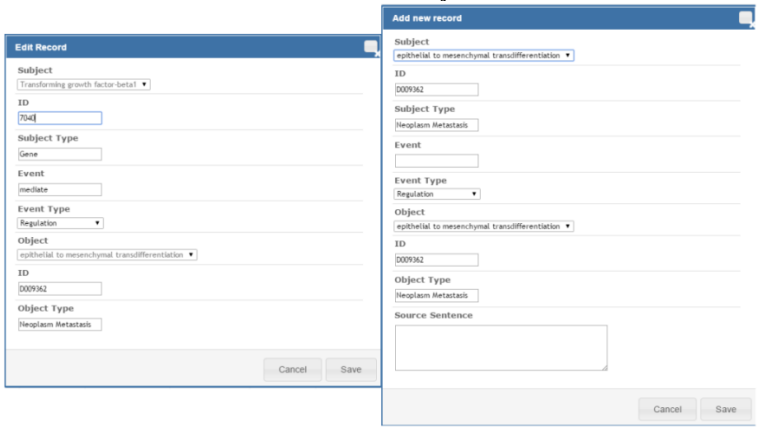

Fig. 4. The curation interface for recognized relations

3 Curation with MET

3.1 Target Curation Concept Types

Table 1. Concept description and instance

Concept Type	Description	Example
Gene MicroRNA	The gene, gene product, and microRNA names mentioned in an abstract	TGF-β miR-181a
Neoplasm Metastasis Cytoskeleton Cell Movement Cell Adhesion	Metastasis is a complex disease-contained series of biological processes. Therefore, descriptions related to cytoskeleton, cell movement, and cell adhesion are also considered as an instance of metastasis.	metastasis stress fiber cell aggregation cell adhesion
Neoplasms	The cancer names mentioned in an abstract	liver cancer
Organ	The organ names mentioned in an abstract	liver
Tissues	The tissue names mentioned in an abstract	adipose tissue

Table 1 summarizes the nine types of biomedical concepts a curator should curate. MET uses different colors, illustrated in Fig. 5, to depict the concepts recognized by our services or curated by curators.

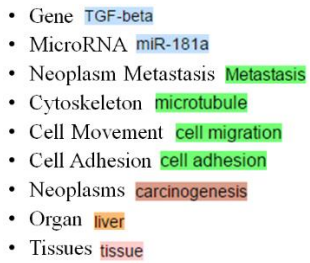

Fig. 5. The colors used in MET for different concept types

3.2 Target Curation Relation Types

The following four relation types involved in a pathway are our curation targets: metastasis relations, positive regulation, negative regulation, or neutral regulation if the mode of the control cannot be determined from the context. The participants of a pathway may include the entire concept types defined in Table 1. The order of the concepts involved in a pathway may elaborate how the biological process happened.

The following summarizes possible relation templates for constructing a metastasis pathway.

- Gene → Gene → Neoplasm Metastasis: The template means that the positive or negative regulation between genes (including microRNAs) resulted in the regulation of neoplasm metastasis (including concepts of cytoskeleton, cell movement, and cell adhesion).
- Gene → Gene → Neoplasms ⇒ (Tissues | Organ): The template means that the positive/negative regulation or regulation between genes (including microRNAs) caused cancer metastasis to certain organ or tissue. The sub-template (Neoplasms ⇒ Tissues | Organ) indicates that the relation involved a cancer concept as its subject, a tissue or organ concept as its object, and a metastasis concept as its trigger.

The above templates can be deconstructed into the following relation patterns:

- (Gene | MicroRNA) → (Gene | MicroRNA)
- (Gene | MicroRNA)→(Neoplasm Metastasis | Cytoskeleton | Cell Movement | Cell Adhesion)
- (Gene | MicroRNA) → (Neoplasms)
- (Neoplasms) ⇒ (Tissues | Organ)

3.3 Curation Tasks for the BioCreative Interactive Text Mining Task

In order to collect articles related to the metastasis biological process and the gene associated of metastasis, the query term "EMT[title/abstract] AND TGF-β[title/abstract]" was used to search PubMed, which results in 949 abstracts. 300 abstracts were randomly selected as the curation dataset for the interactive text mining task. Because there were six curators participating in our curation task, the dataset was split into six equal groups. Table 2 shows the dataset group assignment for each curator.

Table 2. The task assignments of the curators

Curator #	MET Week1	Manual Week1	MET Week2	Manual Week2
C1	G1	G3	G2	G4
C2	G2	G4	G1	G3
C3	G3	G5	G4	G6
C4	G4	G6	G3	G5
C5	G5	G1	G6	G2
C6	G6	G2	G5	G1

As shown in Table 2, there are two phases for the curators in two weeks. In each phase, the curators were asked to complete the following two curation tasks.

MET-assisted curation task. The curators used the Chrome browser extended by MET and conducted their curation directly on the PubMed website with the assistance of MET. MET highlights all recognized concepts and visualizes the constructed pathways in PubMed as illustrated in Fig. 1.

Manual curation task. The curators used BRAT [3] to complete the assignment manually. For each curator, the abstracts assigned to him were preloaded on our BRAT website. The abstracts were not preprocessed by our text mining services for recognizing concepts and relations. For each abstract the curator was asked to annotate all concepts and relations mentioned. In the first week curators were asked to annotate the original abstracts without any aid. In the second week the concept annotations curated by another curator, who worked on the same article in the first week, were provided, and the curator conducted his curation based on the concept annotations provided by the previous curator.

4 Results and Discussion

4.1 Curation Results and Observations

Table 3 shows the number of curated abstracts for the six groups after the two week task. According to the curated results, five curators successfully completed the curation task. One of the curators could not finish the assignment for personal reasons. The abstract curation rate ranged from 4 to 16 per hour.

Table 3. The number of completed abstracts of each task

Group #	MET Week1	MET Week 2	Manual Week 1	Manual Week 2
G1	6	7	10	15
G2	4	6	4	7
G3	10	n/a	8	8
G4	n/a	13	8	6
G5	10	13	16	n/a
G6	10	5	n/a	13
Total	40	44	46	49

Surprisingly, the numbers of completed abstracts by manual curation are more than the ones by MET-assisted in both sessions. The following summarizes our observations after discussing with curators and examining the curation results. Firstly, the information provided by MET is comprehensive, including all of the metastasis-related concepts recognized by our text mining techniques, their normalized IDs and the relations among them. Curators had to check all of these data and correct them if necessary. In contrast in the manual curation task, curators could annotate the sentence-contained relations along with the concepts involved, but ignore all other concepts appearing in the same abstract. Furthermore, most of the curators only annotated

the locations of the concepts but did not normalize the annotated concepts to proper database IDs.

Secondly, the BRAT tool used for manual curation is a mature tool with an easy-to-use annotation interface. Some of our curators already had experience using the tool for annotation. In contrast MET is incubating and many new features were implemented and integrated during the interactive task. Within the two weeks, some faults of MET and our text mining services were reported by curators, such as the concept annotation boundary errors, and the inconsistency between the curated relations and the visualized pathway. Some of the faults came from the text mining services. For example, for recognizing gene mentions in the accessed abstract, the gene annotations provided by Pubtator [4] were integrated. However, the boundaries of the annotations may be inconsistent with the abstract, which lead to the incorrect boundaries visualized by MET. The errors can be easily corrected by deleting the existing annotations and changing them into the right boundaries. Those bugs caused some delay in the curation process, but the interaction with curators also gave us the opportunities to improve MET. Before the start of the second week, we made a major update in MET, which fixed several bugs reported in the first week. It was reflected by the increased number of curated abstracts in the second week.

From the participation survey, most of the curators thought that the annotated concepts and visualized pathways were extremely helpful in giving an idea of which concept was involved in the pathway of metastasis. Through the use of MET can help them extract the knowledge of metastasis in the abstracts quickly. The concept of MET is also considered to be unique because there are no other tools that can help to curate literature for pathways contained as many as concept types like MET. Therefore, we believe that with a further improvement, MET would be very useful especially in building a database for metastasis pathway.

Finally, the design of our curation task was to give different curators the chance to curate the same article groups in the MET-assisted and manual curation tasks. Unfortunately, when we measured the inter-annotator agreement by using the overlapped article groups in MET-assisted and manually curated results, we observed that there was no overlap of articles. It was due to the difference of the article order between PubMed (ordered by publication dates) and the BRAT system (ordered by file names). Therefore, the results from the two curation methods cannot be directly compared.

Another interesting observation is that when comparing the manual curation results in the first week with the second week, there is no significant increase in the number of completed articles. It seems that providing the concept annotations do not reduce the effort of curators because the curators tend to check all annotated by themselves.

4.2 Relation Extraction Performance

Metastasis pathway consists of difference biological relationships between concepts such as gene-gene regulation, gene-cancer regulation and target organs of the metastasis. Our text mining service uses a pattern-based relation extraction approach to extract above relations. Each pattern consists of the concept type tags, such as

"<Gene>" and "<Metastasis>", the relation predicates, such as "inhibit" and "regulate" and the relation type, such as "Positive_regulation_SVO" and "Regulation". To evaluate the relation extraction performance, we assumed that the boundaries of concepts are given. 70 abstracts curated by our curators were then used as the testing set. The relation extraction component achieved precision/recall/F-measure of 0.68/0.77/0.72.

5 Conclusion

We have developed a pathway visualization and curation system that can facilitate the curation of metastatic pathways. Through the participation of the BioCreative interactive text mining task, several feedbacks and suggestions responded by the MET users were received. Overall, MET is considered to be a unique and easy-to-use tool for metastatic pathway curation. With further improvement, MET would be very useful in building a metastatic pathway database. MET would also have great potential and is possibly applicable in other curation concepts. In future work, we would like to fix errors in regards to the inconsistent annotation boundaries and implement new features as requested by curators, such as listing the source sentences of the extracted relations to help curators confirm the relations directly from the curation table, which should make reviewing the pathway annotations faster.

6 Acknowledgment

This work was supported by the Ministry of Science and Technology of Taiwan (MOST-104-2221-E-143-005).

REFERENCES

1. Mehlen P, Puisieux A: **Metastasis: a question of life or death**. *Nat Rev Cancer* 2006, **6**(6):449-458.
2. Dai HJ, Wu JC, Lin WS, Reyes AJ, Dela Rosa MA, Syed-Abdul S, Tsai RT, Hsu WL: **LiverCancerMarkerRIF: a liver cancer biomarker interactive curation system combining text mining and expert annotations**. *Database (Oxford)* 2014, **2014**.
3. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii Ji: **BRAT: a web-based tool for NLP-assisted text annotation**. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics: 2012*: Association for Computational Linguistics; 2012: 102-107.
4. Wei C-H, Harris BR, Li D, Berardini TZ, Huala E, Kao H-Y, Lu Z: **Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts**. *Database* 2012, **2012**.