W416–W422 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* *Published online 15 May 2008*
doi:10.1093/nar/gkn286

E3Miner: a text mining tool for ubiquitin-protein ligases

Hodong Lee[1], Gwan-Su Yi[2] and Jong C. Park[1,*]

[1]Department of Computer Science, KAIST, 335 Gwahangno, Yuseong-gu, Daejeon 305-701 and [2]School of Engineering, Information and Communications University, 119 Munjiro, Yuseong-gu, Daejeon 305-732, South Korea

Received February 20, 2008; Revised April 17, 2008; Accepted April 26, 2008

ABSTRACT

Ubiquitination is a regulatory process critically involved in the degradation of >80% of cellular proteins, where such proteins are specifically recognized by a key enzyme, or a ubiquitin-protein ligase (E3). Because of this important role of E3s, a rapidly growing body of the published literature in biology and biomedical fields reports novel findings about various E3s and their molecular mechanisms. However, such findings are neither adequately retrieved by general text-mining tools nor systematically made available by such protein databases as UniProt alone. E3Miner is a web-based text mining tool that extracts and organizes comprehensive knowledge about E3s from the abstracts of journal articles and the relevant databases, supporting users to have a good grasp of E3s and their related information easily from the available text. The tool analyzes text sentences to identify protein names for E3s, to narrow down target substrates and other ubiquitin-transferring proteins in E3-specific ubiquitination pathways and to extract molecular features of E3s during ubiquitination. E3Miner also retrieves E3 data about protein functions, other E3-interacting partners and E3-related human diseases from the protein databases, in order to help facilitate further investigation. E3Miner is freely available through http://e3miner.biopathway.org.

INTRODUCTION

The ubiquitin is a small protein that works as a tag covalently attached to proteins to influence nearly all cellular processes in eukaryotes (1). In particular, the tagging process, or ubiquitination, is critically involved in the degradation of 80–85% proteins, where they are specifically recognized by a key enzyme or a ubiquitin-protein ligase (E3) (2,3). Since nondegradation of proteins gives rise to cellular toxicity, the malfunction of E3s is often implicated in serious human diseases, such as cancers and neurodegenerative disorders (4,5).

In order to discover novel therapeutic solutions for such diseases, biological and biomedical researchers have paid much attention to molecular mechanisms of E3s. However, due to the distributed nature of such discoveries, the findings are scattered over a number of protein databases [e.g. UniProt (6), IntAct (7), GO (8) and OMIM (9)], making it difficult to coordinate efforts to access their data efficiently and conveniently. Moreover, due to the unfocused nature of such databases, they do not give stream-lined and specific information about E3s and their mechanisms. Furthermore, the UbiProt database, constructed specifically to provide data about target substrates of E3s, neither contains a comprehensive range of data about E3s nor covers up-to-date substrates reported in the literature, mostly due to the manual curation method that it employs (10). In this work, we address this insufficiency and out-of-datedness of E3 data by automatically extracting and managing data from MEDLINE abstracts and relevant protein databases.

Given the biological importance of E3s, it is not surprising that a rapidly growing body of the published literature describes novel findings about E3s (e.g. around 20 000 MEDLINE abstracts can be retrieved with the 'ubiquitin' MeSH term and around 10 000 abstracts with the MeSH term 'ubiquitin-protein ligase or E3 ligase' as of February 2008). However, text-mining tools have not yet targeted specifically at the extraction of E3 or ubiquitination-related data from the literature (11–14). In this article, we describe a text-mining tool, E3Miner, that extracts information about molecular mechanisms of E3s from published literature and biological databases. The main contribution of E3Miner is to make available information about protein interactions in E3-specific ubiquitination pathways, where ubiquitins are transferred to target substrates through interactions among enzymes. E3Miner extracts E3s and their interacting proteins from text to present an integrated view of such pathways using a graphic browser, and incorporates the protein

*To whom correspondence should be addressed. Tel: +82 42 869 3541; Fax: +82 42 869 5581; Email: park@nlp.kaist.ac.kr

interactions as reported in research articles to help advance the understanding of the underlying mechanisms. It also extracts E3-related molecular features, such as domains, ubiquitination types and ubiquitination sites. The molecular features of this kind suggest important biological implications of E3s, since they are closely associated with cellular functions and processes of ubiquitination (15,16).

Protein databases provide information as important as, if not more so than, literature databases. E3Miner integrates E3-dependent functions/processes with GO terms, E3-related human diseases from OMIM and other E3-interacting proteins from IntAct. The users could consult (i) ontological information of GO terms for knowledge inference such as protein function prediction (17), (ii) human disease names to look into pathological implications of E3s or their substrates and (iii) additional E3-interacting proteins to investigate unknown proteins involved in the ubiquitination process or the regulation of E3s. Taken together, we believe that the E3 data extracted by our tool enable a systematic understanding of E3s and their roles.

MATERIALS AND METHODS

E3Miner is an information extraction system that extracts E3s and their ubiquitination-related data from the published literature. Currently, the focus is on the extraction of E3-related information from the sentences, which mention E3s for definition, apposition, example, role and activity. After selecting E3-mentioning sentences, the system identifies E3 names from the text, based on a three-step approach: (i) tagging parts-of-speech (POSs) to the words in each sentence; (ii) identifying candidate protein names of E3s by using phrasal rules for definition, apposition, example, role and activity and then (iii) grounding (or linking) protein names into the corresponding entries in the UniProt database.

Figure 1 shows a sample procedure of the E3Miner system. Given a MEDLINE abstract, the system first selects sentences with enzyme markers for E3s (e.g. *E3 ligase*, *ubiquitin ligase* and *ubiquitin-protein ligase*). It then identifies candidate expressions for E3 protein names from the sentences, by using the clausal and phrasal rules for E3s. If candidate expressions include protein names, the system grounds (or links) the protein names into the corresponding entries of UniProt. It then identifies the protein names for E3-interacting proteins, such as target substrates, ubiquitin-activating enzymes (E1s), ubiquitin-conjugating enzymes (E2s) and deubiquitinating enzymes (DUBs), by using clausal rules for the E3-substrate interactions and other protein interactions. E3Miner then extracts E3-related data, such as GO terms, E3 domains, ubiquitination types and ubiquitination sites by matching them with the sentences, and then integrates the disease information from OMIM and other interacting proteins from IntAct, using the identified UniProt ID (UPID). In the output shown in Figure 1, IDs beginning with 'UPID', 'GO' and 'MIM' indicate UniProt ID, GO term ID and OMIM ID, respectively. In addition, E3Miner provides statistical information from the precompiled E3 data. We describe each step of the procedure further in the following subsections.

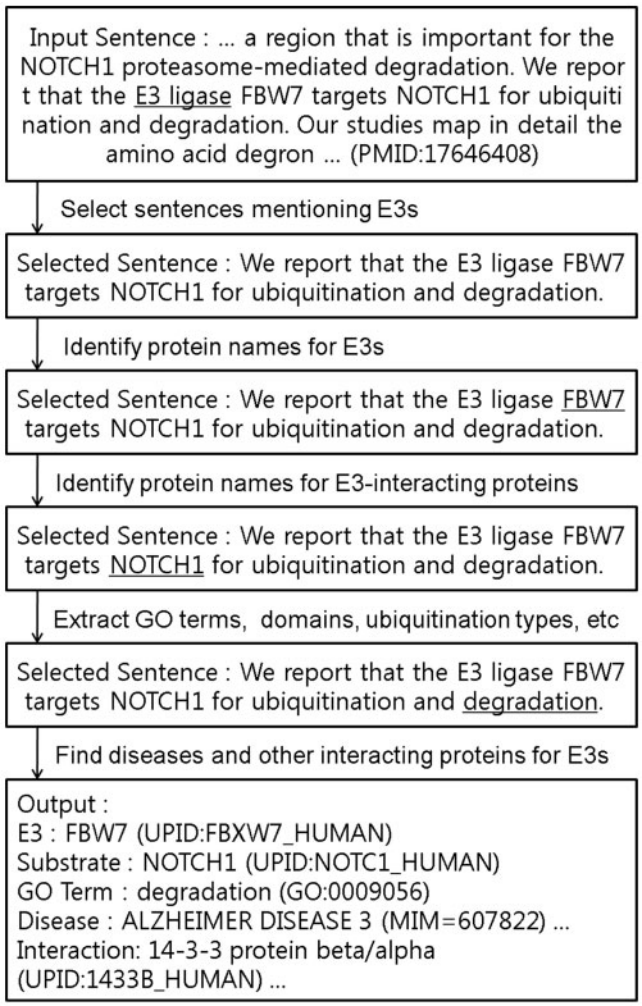

Figure 1. An example procedure for E3 data extraction.

Selecting relevant sentences

E3Miner splits the input text into individual sentences and then selects sentences with mentions of E1, E2, E3 and DUB, using regular expressions of enzyme class names. For example, the E3 class names include markers, such as *E3, ubiquitin-protein ligase, ubiquitin ligase* and *ub-ligase*. Table 1 shows example patterns for enzyme class marker. E3Miner utilizes the regular expressions to search for them from each sentence, with consideration for such morphological variations as plural endings. The details of regular expressions are shown in 'Supplementary Material' Section 1.

Identifying E3s

If there are sentences with mentions of E3 markers, E3Miner attaches POS tags, such as NP (noun phrase) and VP (verb phrase), to the words of each such sentence. It then extracts sequential NPs with E3 protein/gene names, using clausal and phrasal rules for definition, apposition, example, role, activity and noun complex.

Table 1. Examples of enzyme class markers

Class	Examples
E1	E1, ubiquitin-activating enzyme, ub-activating enzyme
E2	E2, ubiquitin-conjugating enzyme, ubiquitin carrier enzyme
E3	E3, ubiquitin-protein ligase, ubiquitin ligase, ub-ligase
DUB	DUB, de-ubiquitinating enzyme, deubiquitinase

Table 2. Example clausal and phrasal patterns for E3 protein names

Rule type	Example patterns
Definition	'P is E3', 'P are E3'
Apposition	'E3, P,', 'P, E3,'
Example	'E3, such as, P', 'E3, for example, P'
Role	'P function as E3', 'P act as E3'
Activity	'P have E3 activity', 'P exhibit E3 activity'
Noun Complex	'E3 P', 'P E3'

Table 2 shows example patterns for such rules. In this table, P and E3 indicate the noun phrases for E3 protein/gene names and the E3 class markers, respectively.

Taking into account parentheses ('('and')'), hyphenation ('-') and coordination items ('and' and 'or'), E3Miner recognizes E3 protein/gene names from extracted noun phrases, and discards E3 class names, if present. It then identifies UPID and synonyms from UniProt for the recognized E3 names.

In this work, we developed (i) a POS tagger that assigns to each word its most frequent POS tag by looking up a manually curated POS dictionary together with domain-specific correction rules, (ii) a noun phrase recognizer that looks for noun phrases that begin or end at words involved in determiners (e.g. 'a' and 'the'), prepositions (e.g. 'for' and 'with') or verbs, since the noun phrases for 'P' and 'E3' in the patterns are adjacent to such words and (iii) a protein name linker that finds out UPIDs by using an organism name and a protein/gene name identified from the same sentences. If the procedure fails to identify the organism name, it attempts to search for UPIDs with an organism name from the title of abstract or from the prefixes (e.g. 'h' for 'human' and 'y' for 'yeast') of the identified protein/gene names. In this process, the procedure performs the exact match of the identified organism and protein/gene names, along with their variations by whitespace (' '), hyphenation ('-') and symbols (e.g. 'I' and 'II'). If this match results in a single UPID, then the procedure assigns it to the protein mention; but if multiple UPIDs are found, the procedure shows all of them without further disambiguation. In this case, the procedure assigns the first UPID for 'human' to the mention, as a default, for further uses in our precompiled E3 data.

Identifying E3-interacting proteins

If protein names of E3s are identified, the tool finds other E3-mentioning sentences that do not contain enzyme class markers. E3Miner then extracts target substrates from such sentences, using clausal rules encoding the ubiquitinating relation, such as 'E3 ubiquitinate P', 'E3 target P' and 'ubiquitination of P by E3', where P indicates the protein name of a target substrate. The interacting relations used in our system are further elaborated in Section 2 of 'Supplementary Material'. E3Miner identifies enzymes, such as E1, E2 and DUB, using a method similar to the E3 identification, and checks for co-occurring E3s in the same sentence, in order to ensure that their protein interactions are positively involved in ubiquitination pathways.

Extracting GO terms and ubiquitination-related features

E3Miner identifies GO terms that occur in sentences with mentions of E3s. It locates a part of noun phrases in sentences with a complete list of GO terms, by using the longest-first matching method. If a UPID is identified for an E3, this procedure imports GO terms from the corresponding UniProt entry. It then identifies ubiquitin-related molecular features, such as E3 domain, auto-ubiquitination, ubiquitination types of E3 and ubiquitin-binding sites of substrates. E3Miner extracts such features by matching the following patterns: (i) E3 domain names by using a marker 'domain' or words ending with '-dependent' and '-containing'; (ii) auto-ubiquitination by the word-level patterns, such as auto- or self-ubiquitination; (iii) ubiquitination types by the patterns, such as poly-/multi-/mono-/oligo-ubiquitination and (iv) ubiquitination sites by patterns, such as 'lysine #-linked', 'lys(#)-linked' or 'K#-linked', where '#' indicates the location number of a lysine residue.

Integrating data for human diseases and other E3-interacting proteins

The system searches for human disease names for E3s and their substrates from OMIM. It utilizes protein/gene names and their synonyms in the search process. It then integrates other E3-interacting protein names from IntAct, using UPIDs for E3s. Such interacting proteins may include (i) unknown substrate proteins binding to an E3, (ii) unknown enzymes involved in a ubiquitination pathway, (iii) unknown proteins having an important role in the regulation of an E3 and (iv) unknown proteins to be regulated by an E3. We believe that such protein information is useful for in-depth investigation of E3-related proteins.

Providing statistical information

E3Miner generates statistical information from the precompiled E3 data. The system looks up the precompiled data, in order to provide the statistics of E3-interacting proteins, together with their references to source articles for further investigation. Using this statistics, users may infer the importance, relevance, and research interest of a particular E3.

IMPLEMENTATION

E3Miner enables users to exploit a comprehensive collection of available E3 data through an interactive web-based interface. It runs on a Linux machine with dual Xeon 2.8 GHz CPUs and Apache 2.0.55 as a web-server

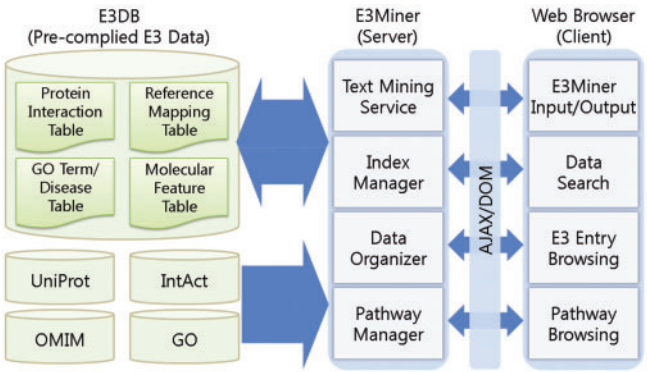

Figure 2. E3Miner provides E3 data through an interactive interface of pathway browsing and search functions. The graphic module of the pathway browser is based on GraphViz, and the search modules are based on the indexing function of MySQL. Given an input, the web interface invokes such modules using web development techniques, i.e., AJAX and DOM.

platform. The web interface is implemented using Python, AJAX/DOM (Asynchronous Javascript and XML/Document Object Model), MySQL 5.0.22 and GraphViz 2.2.1. Figure 2 shows the system architecture of E3Miner and its web interface.

The system has been running for >6 months as of February 2008. It is also tested on a corpus with 9300 MEDLINE abstracts published before 23 June 2006, obtained by a PubMed search with the MeSH term 'E3 ligase OR ubiquitin protein ligase'. We compiled the corpus to extract E3 relevant data reported in the literature. E3Miner extracted 2757 overall mentions of E3s in 3.09 s per abstract. Among them, 1696 protein names are successfully cross-linked to the UniProt database, resulting in 796 distinct entries. The smaller number of the corresponding entries is due to duplicate mentions for the same proteins from different abstracts. The resulting data from the corpus is used for the construction of a precompiled collection of E3 data, called E3DB.

In order to obtain indicative measures for the system performance, we evaluated the precision and recall of our mining method on a testset of 100 abstracts randomly selected from the ones published after 1 January 2005. We found that 47 abstracts of the testset contain expressions of E3 protein names with enzyme markers, whereas the other abstracts contain expressions either only for E3 markers or for protein names. Given the testset, E3Miner correctly extracted E3 data for 34 abstracts over a total of 35 answered cases, i.e. 97% (34/35) precision and 74% (34/47) recall. The system tends to miss mentions of E3 protein complexes expressed in lengthy noun phrases, since the protein names are syntactically far from their E3 markers. Nevertheless, we believe that this is acceptable since our current priority is to achieve precision over recall for the practical usefulness of our database. In addition, the system's I/O process is automatically monitored by an error checking module and is also manually tested by about 15 persons working on computer science and bioinformatics. E3Miner is freely accessible from http://e3miner.biopathway.org.

WEB INTERFACE AND EXAMPLE

Web interface

Through a web interface, E3Miner not only provides a function of mining E3 data from text, but also supports three types of E3 search from the precompiled E3DB in an interactive manner: quick search, advanced search and pathway browsing. The quick search provides a keyword search over all fields, such as gene name, protein name, UniProt ID, target protein, E1, E2, DUB, GO term, organism, domain, ubiquitin structure and disease. For more detailed data retrieval, the advanced search allows users to search for E3 entries by using a combination of keywords in a particular field. Users can also search for E3 entries by browsing the ubiquitination pathways in the pathway browser. In order to help users with knowledge inference, the browser presents an integrated view of pathways, and supports interactive protein navigation. The browsing of E3 entry and ubiquitination pathway is also possible from the output of E3Miner, if the corresponding E3 entry is registered in E3DB.

Program input

E3Miner receives input as PMID(s) or text in journal abstract(s). PMIDs should be separated by comma (','), and the input text should contain at least a single sentence. If PMIDs are given, E3Miner retrieves the corresponding titles and texts from PubMed through Internet access, and extracts E3 data from them.

Program output

Given input text or PMIDs, E3Miner produces structured data entries about E3s identified from the input text. A data entry consists of E3 relevant data items from the input text and relevant databases, including interacting proteins and their specificity, Gene Ontology (GO) terms, human disease names, additional E3-interacting protein names and statistical information, as described subsequently. (i) The interacting proteins are members of target substrates and ubiquitin-transferring proteins, such as E1s, E2s and DUBs. The ubiquitination types and sites for a target substrate are also identified from the input text. (ii) The GO terms both extracted from the input text and imported from UniProt are used to describe functions or processes of an E3. The extracted GO terms are the sentence fragments and their morphological variations matched with those in GO. (iii) The disease names related to E3s and their substrates are imported from OMIM. The output includes the disease names with their cross-links to the OMIM database. (iv) The additional E3-interacting protein names from IntAct are enumerated with their cross-links to UniProt. (v) The output also includes statistical information both by the number of journal abstracts that refer to the identified E3s and by the number of interacting proteins for the E3s and target substrate. The statistical information of this kind is expected to convey the importance and relevance of E3s. Further details of the output are available through http://e3miner.biopathway.org/help_manual.html.

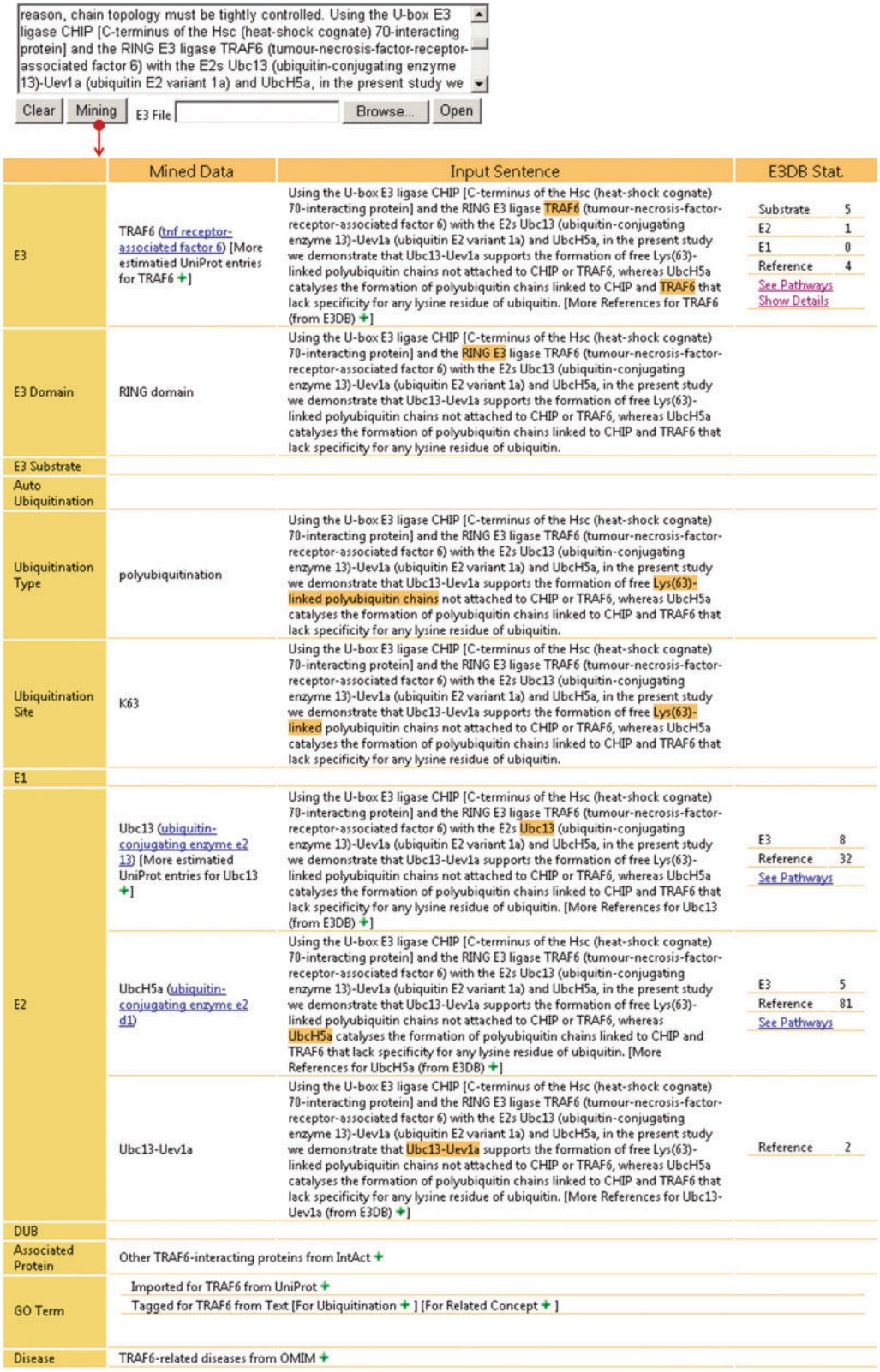

Figure 3. An example output of E3Miner for the MEDLINE abstract (PMID:18042044).

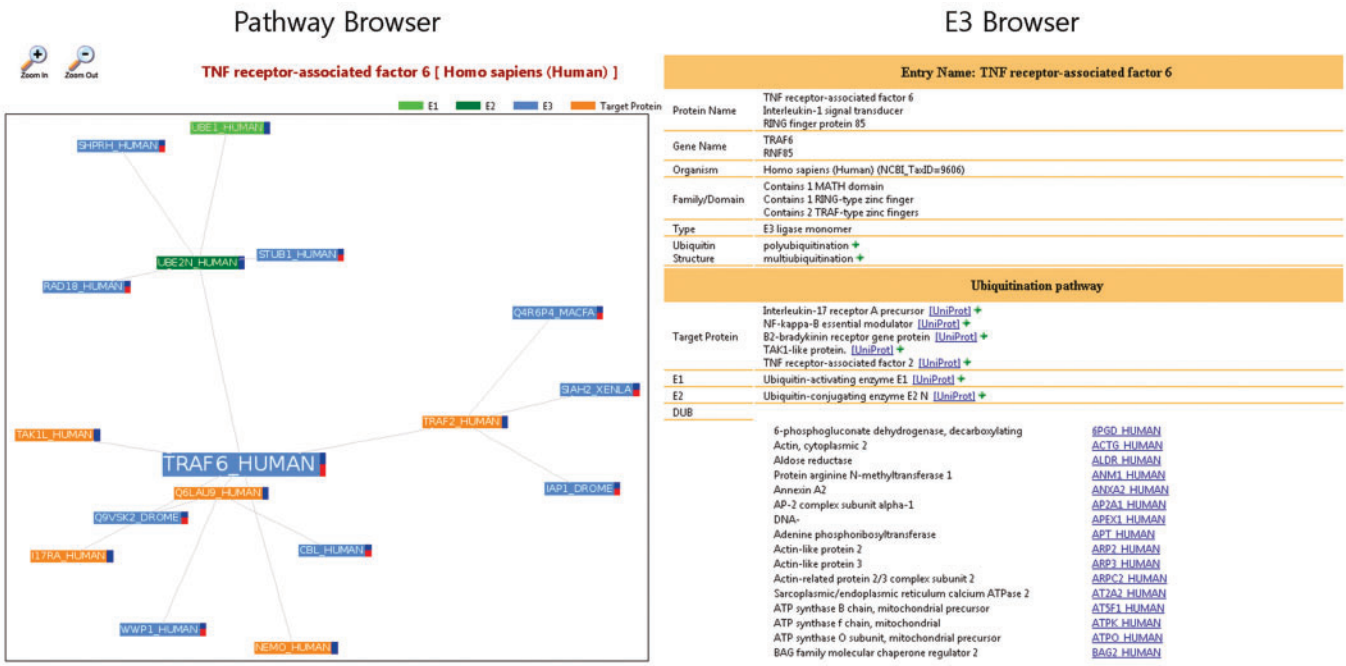

Figure 4. Snapshots of Pathway Browser and E3 Brower linked from the output for TRAF6 in Figure 3.

Example

Given input text for the MEDLINE abstract (PMID:18042044), E3Miner generates data entries about E3 ligase CHIP and TRAF6 from the text. Figure 3 shows an example entry for TRAF6.

Triggered by the 'mining' button, the system first identifies TRAF6 from the input text, and then attempts to find its standard protein name with a UPID from UniProt. It then identifies the E3-interacting proteins, such as substrates and E2s. In this case, the system extracts Ubc13, UbcH5a and Ubc13-Uev1a as E2s, but fails to find the corresponding UniProt entry for Ubc13-Uev1a complex, which is not registered in UniProt. The system then finds GO terms and ubiquitination-related features, such as 'RING domain' for the E3 domain, 'polyubiquitination' for the ubiquitination type and 'K63' for the ubiquitin-binding site. The GO terms are separately categorized for those imported from UniProt and those automatically tagged from text, and the GO terms tagged from text are also categorized as the terms directly related to the ubiquitination and the other ubiquitination-related terms. E3Miner also integrates the disease names from OMIM and other E3-interacting proteins from IntAct. In the output, the 'Mined Data' column shows the E3 data identified from text, and the 'Input Sentence' column shows the source sentence for the mined data. The system then calculates the statistical information and shows them in the 'E3DB Stat.' column. Using the links of this column, users can browse the ubiquitination pathways and further details of a particular E3, as depicted in Figure 4.

CONCLUDING REMARKS

E3Miner is still evolving in its format and capability for text mining. Nevertheless, we find that it is now mature and useful enough in its present form as a specialized biological tool. Since its mining process is fully automated, E3Miner can be flexibly fine-tuned to extract novel E3 discoveries and important findings related to specific E3s from the literature. This aspect would be particularly important for the usability of E3Miner, since there are apparently quite a large body of knowledge about E3s and their biological mechanisms that remain undiscovered. As a practical tool for biological research, E3Miner makes available ubiquitination pathways, molecular features, protein functions and processes and human disease information associated with E3s that could all work together in the search for potent drug targets for hereditary diseases, such as cancers and neurodegenerative disorders.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Tak-eun Kim deeply for his expert and comprehensive help with the web interface. This work was supported in part by grant No. R01-2005-000-10824-0 from the Basic Research Program of the Korea Science & Engineering Foundation, and in part by the Korea Research Foundation Grant funded by the Korean

Government (MOEHRD, Basic Research Promotion Fund) (KRF-2007-313-D00738).

Conflict of interest statement. None declared.

REFERENCES

1. Conaway,R.C., Brower,C.S. and Conaway,J.W. (2002) Emerging roles of ubiquitin in transcription regulation. *Science*, **296**, 1254–1258.
2. Hershko,A. and Ciechanover,A. (1998) The ubiquitin system. *Annu. Rev. Biochem.*, **67**, 425–479.
3. von Mikecz,A. (2006) The nuclear ubiquitin-proteasome system. *J. Cell Sci.*, **119**, 1977–1984.
4. Ardley,H.C. and Robinson,P.A. (2004) The role of ubiquitin-protein ligases in neurodegenerative disease. *Neurodegener. Dis.*, **1**, 71–87.
5. Burger,A.M. and Seth,A.K. (2004) The ubiquitin-mediated protein degradation pathway in cancer: therapeutic implications. *Eur. J. Cancer*, **40**, 2217–2229.
6. Bairoch,A., Apweiler,R., Wu,C., Barker,W., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
7. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct – open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
8. The GO Consortium (2007) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.
9. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
10. Chernorudskiy,A., Garcia,A., Eremin,E., Shorina,A., Kondratieva,E. and Gainullin,M. (2007) UbiProt: a database of ubiquitylated proteins. *BMC Bioinform.*, **8**, 126.
11. Cohen,A. and Hersh,W. (2005) A survey of current work in biomedical text mining. *Brief Bioinform.*, **6**, 57–71.
12. Krallinger,M., Erhardt,R. and Valencia,A. (2005) Text-mining approaches in molecular biology and biomedicine. *Drug Discov. Today*, **10**, 439–445.
13. Lussier,Y., Borlawsky,T., Rappaport,D., Liu,Y. and Friedman,C. (2006) PhenoGO: assigning phenotypic context to Gene Ontology annotations with natural language processing. In *Pac. Symp. Biocomput.*, **11**, 64–75.
14. Karamanis,N., Lewin,I., Seal,R., Drysdale,R. and Briscoe,E. (2007) Integrating natural language processing with FlyBase curation. In *Pac. Symp. Biocomput.*, **12**, 245–256.
15. Pickart,C.M. and Eddins,M.J. (2004) Ubiquitin: structures, functions, mechanisms. *Biochim. Biophys. Acta*, **1695**, 55–72.
16. Welchman,R.L., Gordon,C. and Mayer,R. (2005) Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nat. Rev. Mol. Cell Biol.*, **6**, 599–609.
17. Tao,Y., Sam,L., Li,J., Friedman,C. and Lussier,Y. (2007) Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*, **23**, i529–i539.