A method to retrieve papers from MEDLINE: PETER system

Hiroko Ao, Yasunori Yamamoto, Toshihisa Takagi
aohiroko@ims.u-tokyo.ac.jp, yayamamo@ims.u-tokyo.ac.jp, tt@cb.k.u-tokyo.ac.jp
Dept. of Computational Biology, University of Tokyo

We attempted to eliminate non-relevant papers from results of PubMed searches for each topic. The system is called PETER (PubMed Enhancer Toward Efficient Research) and it works as follows.

1. get LocusLink IDs manually.
2. collect information of gene names (AKA synonyms) from public databases.
3. make synonym variations automatically.
4. search papers by PubMed with each synonym.
5. extract titles and abstracts.
6. take another information about synonyms from the extracted titles and abstracts.
7. extract information about abbreviations from the titles and the abstracts.
8. retrieve appropriate papers by using the synonyms and the abbreviations.

Keywords to be used for the PubMed searches were synonyms which were collected from public databases (e.g., SWISS-PROT, LocusLink, etc.). The retrieval method PETER employs is rule-based and rules were constructed from the observations that a potential abstract usually includes a synonym of a query and at least one word from the other synonyms. We call these words "selected words", each of which must have no less than four letters and should not be stopwords we prepared.

The scoring system was designed to evaluate a paper in terms of whether or not it contains a query's abbreviation, another synonym (full spelling), or a selected word. The one-sentence splitter called JASMINE (Just A Sentence-splitter Maximizing Intelligence of kNowledge Extraction) and the abbreviation extractor called ALICE (Abbreviation LIfter using Condition-based Extraction) were also developed for PETER system.

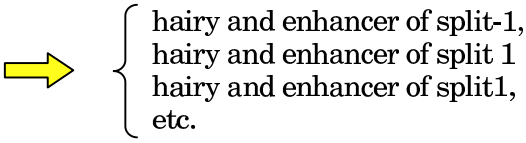
Making out a list of synonyms was the hardest work due to the insufficiency of the databases for gathering appropriate ones. Some entries related to gene names stored in the databases are inappropriate as gene names (e.g., hypothetical protein FLJ20006, Hirschsprung disease and EST), some are not gene specific (e.g., A1, DNA binding protein and tumor necrosis factor), and some are not appeared in real papers. In order to overcame these difficulties and achieve high

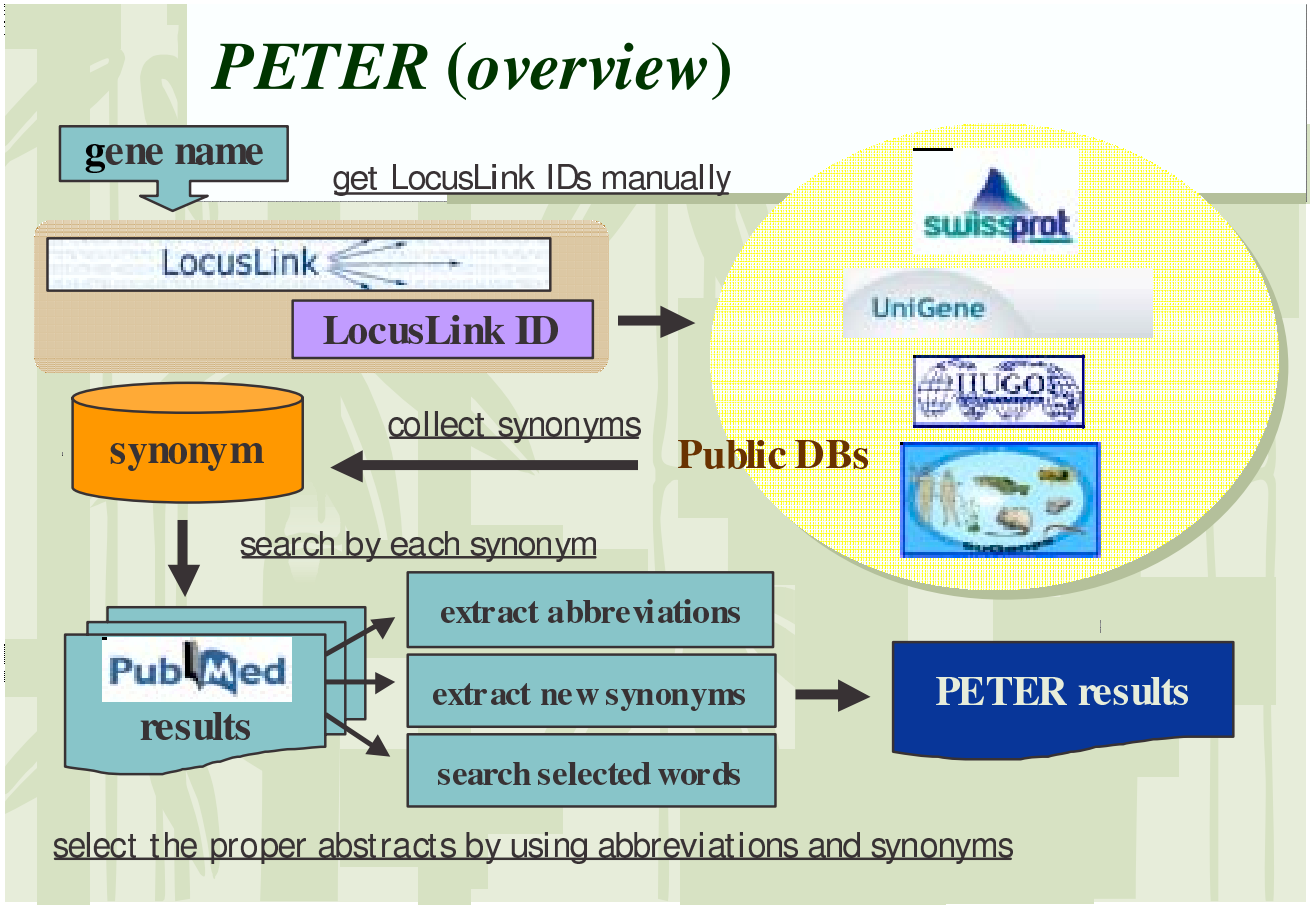
recall, we made a variant generator to add synonym variations for PubMed searches[1], and collected the other synonyms from the retrieved papers' titles and abstracts after the searches. To get high precision, at the same time, we established empirical selection rules toilsomely.

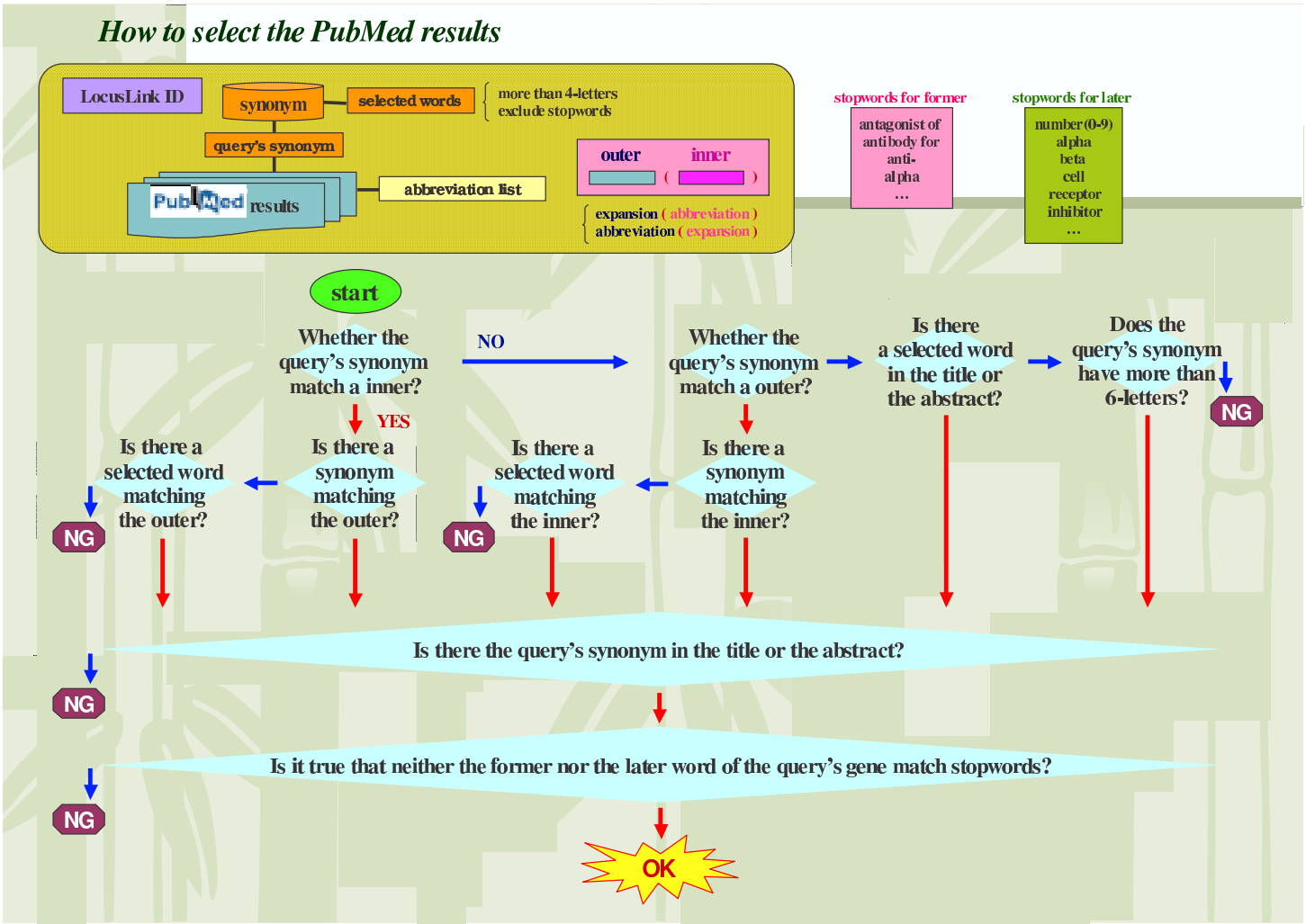
As a conclusion, we got high precision and recall concerning human UV-regulated genes. The reason is that we have developed PETER for dermatologists in the first place as a joint research with a cosmetic company. It was built to retrieve papers about UV-regulated genes from MEDLINE. Our approach was to make a system which worked as much the same as biologists' do to get papers. While we tried to improve PETER to work well for all genes, it was quite difficult to adjust our method even to general human genes. Through this TREC project, we recognized that, to get better results, it was important for a retrieval system to be able to tune the rules upon biologist's requests. There is no perfect rule, and there is no specialist to establish an all-round method and to predict the results which the system provides. Although it is impossible to create a flawless method for all genes, we want to make an effort to improve PETER as much as we can. We take pleasure in discussing scholars engaged in Bioinformatics, Biology, and Information Retrieval.

---

[1] For example, hairy and enhancer of split-1, (Drosophila)




PETER (overview)
gene name
get LocusLink IDs manually
LocusLink
LocusLink ID
swissprot
UniGene
HUGO
Public DBs
synonym
collect synonyms
search by each synonym
PubMed
results
extract abbreviations
extract new synonyms
search selected words
PETER results
select the proper abstracts by using abbreviations and synonyms


How to select the PubMed results
LocusLink ID
synonym
selected words
more than 4-letters
exclude stopwords
query's synonym
PubMed results
abbreviation list
outer
inner
( )
expansion (abbreviation)
abbreviation (expansion)
stopwords for former
antagonist of
antibody for
anti-
alpha
...
stopwords for later
number (0-9)
alpha
beta
cell
receptor
inhibitor
...
start
Whether the query's synonym match a inner?
NO
YES
Whether the query's synonym match a outer?
Is there a selected word in the title or the abstract?
Does the query's synonym have more than 6-letters?
NG
Is there a selected word matching the outer?
Is there a synonym matching the outer?
Is there a selected word matching the inner?
Is there a synonym matching the inner?
NG
NG
Is there the query's synonym in the title or the abstract?
NG
Is it true that neither the former nor the later word of the query's gene match stopwords?
NG
OK