# An Algorithm to Identify Abbreviations from MEDLINE

**Hiroko Ao**[1,2] aohiroko@ims.u-tokyo.ac.jp
**Toshihisa Takagi**[1] tt@k.u-tokyo.ac.jp

[1] Department of Computational Biology, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8561, Japan
[2] Basic Research Laboratory, *Kanebo*, LTD., 5-3-28 Kotobuki-cho, Odawara-shi, Kanagawa 250-0002, Japan

**Keywords:** abbreviation, expansion, MEDLINE

## 1 Introduction

With the rapid growth of machine-readable literature, such like MEDLINE database, new abbreviations are increasing. They make it difficult for researchers to follow the enormous size of vocabulary. To automatically identify abbreviation and its expansion (e.g., EaggEC; enteroaggregative *Escherichia coli*), we propose an effective mining system called ALICE (Abbreviation LIfter using Condition-based Extraction). This method also accepts acronym and its definition (e.g., HPLC; high performance liquid chromatography). Synonyms (e.g., cyclin-dependent kinase inhibitor 1A (p21)), hypernyms (e.g., interleukin receptors (IL2-R)), citations (e.g., John *et al.* 2000), and measures (e.g., 20 mg/d) are excluded as much as possible. It helps to construct ontology and lexicons across different fields and also facilitate computer analysis of corpora.

## 2 Method

ALICE is divided into three phases, information retrieval (IR), information extraction (IE), and conformation judgment (CJ). The outline of ALICE system is shown in Figure 1.

**IR phase**: We manually crafted rules and prepared stop words to search parentheses. A parenthesis here means a parenthesis "()", a bracket "[]", or a brace "{}". Candidate abbreviations (acronyms) or expansions (definitions) in parentheses are classified as four types according to existence of a space before them or to characters which compose them. When a parenthesis matches one of the types and it isn't a member of stop words, the chunk of the words in front of it (i.e., candidate expansions or abbreviations) is evaluated.
**IE phase**: We first formulated a hypothesis that parentheses were used for abbreviations. In this case, chunks (i.e., candidate expansions) are classified as eight types according to how they are composed of. In order to deal with parentheses used for expansions, then we added the ninth type, thereby, made it possible to extract candidate abbreviations.

Our rules don't restrict the first word of a expansion to begin with its initial letter of the abbreviation. For example, oestrogen receptor (ER), rt-PA-APSAC patency study (TAPS), isoelectric pH value (pI), or 50% of their VC (V50). Besides the case, this system enables to extract a expansion containing parentheses (e.g., 2'-(4-methylumbelliferyl)-alpha-D-N-acetylneuraminic acid (MUNANA)).
**CJ phase**: We established 13 prohibitive conditions to eliminate inappropriate candidate expansions and abbreviations. If a candidate is rejected (i.e., it is a synonym, a hypernym, a citations, or a measure), the program returns null.

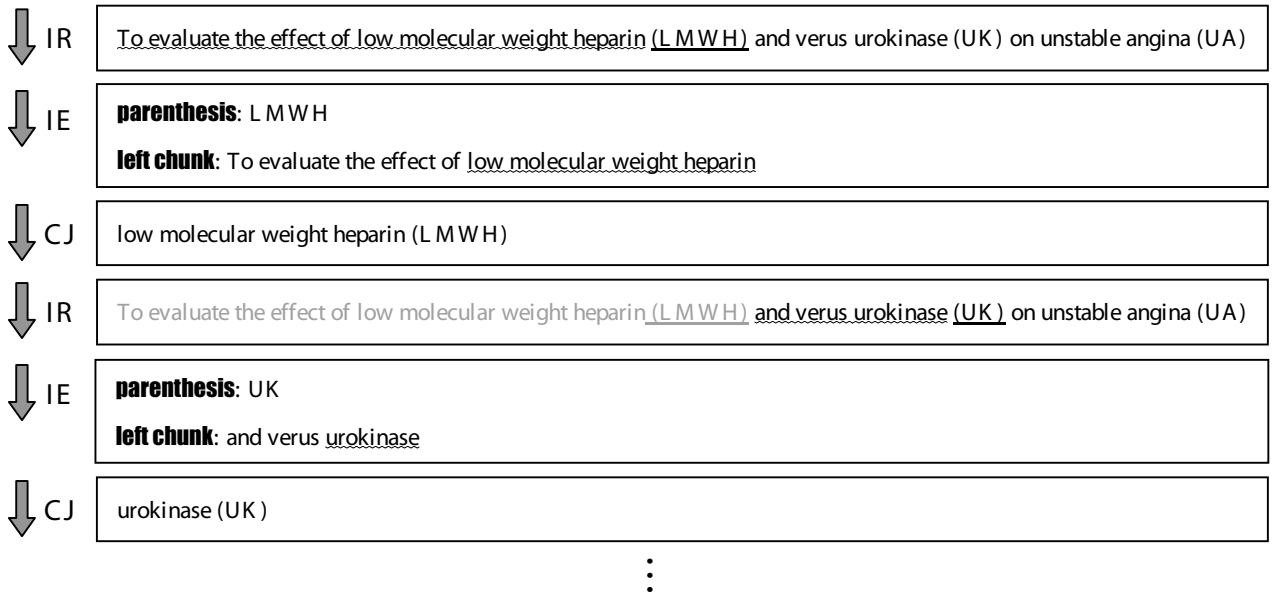


Figure 1: System of ALICE.

## 3 Results

ALICE was applied to 1,000 abstracts with their titles. They were selected randomly from MEDLINE (PMID 12500000∼12599999). This system extracted pairs of abbreviations and expansions, or acronyms and definitions with 96% precision and 98% recall. The total extracted number of the pairs by ALICE was 1,107. All of the number of the parentheses used in the 1,000 abstracts was 4,571.

## 4 Discussion

ALICE is effective in extracting abbreviations from MEDLINE abstracts as it covers any kinds of them without limiting research fields. Though there are many algorithms to extract abbreviations, some of them restrict parentheses to be used for abbreviations (acronyms) only [1]. Others limit the abbreviations to acronyms or protein names [2, 3]. To realize how the authors coinage or quote abbreviations, a comprehensive tool to accept all types of abbreviations is desired. In this regard, we believe our system made a contribution to realizing it.

As for the limitations of our system, the case of naming is difficult to be eliminated because they are similar to abbreviations (e.g., authentic PPAR/RXR binding element (Aco-PPRE), xylan degradation (xlnA), or class I(B) PI3K (PI3Kgamma); it was extracted "PI3K (PI3Kgamma)"). In addition, it is impossible to retrieve expansions split by enumeration (e.g., topoisomerase I (topo I) or II (topo II), or diffusion- (DWI) and perfusion-weighted (PWI) imaging). Furthermore, it is also incapable of retrieving when an order of words is different between a expansion and an abbreviation (e.g., 70 kDa heat shock cognate gene (HSC70) or protein kinase C epsilon (epsilonPKC)). These problems need to be solved in the future.

## References

[1] Chang, J.T., Schutze, H., and Altman, R.B., Creating an online dictionary of abbreviations from MEDLINE, *Journal of the American Medical Informatics Association*, 9(6):612–620, 2002.

[2] Yeates, S., Bainbridge, D., and Witten, I.H., Using compression to identify acronyms in text, *Data Compression Conference (DCC '00)*, 2000.

[3] Yoshida, M., Fukuda, K., and Takagi, T., PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary, *Bioinformatics*, 16(2):169–175, 2000.