Advanced Gene Mention Tagging System for CALBC Challenge

Cheng-Ju Kuo[1] clarkkuo@iis.sinica.edu.tw

Chun-Nan Hsu[1] chunnan@iis.sinica.edu.tw

Maurice HT Ling[2,3] mauriceling@acm.org

[1] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2] School of Chemical and Life Sciences, Singapore Polytechnic, Republic of Singapore
[3] Department of Zoology, The University of Melbourne, Australia

In gene mention tagging task (GM) of BioCreative II (2007) challenge, we had built a system based on bi-directional parsing models of Conditional Random Fields (CRFs), achieved a F-score of 86.83 [2] on GM test corpus and ranked second among twenty-one participants. After which, we improved its performance to a F-score of 88.3 by integrating high dimensional bi-directional parsing models (up to 6 models) [1]. For inter-operability within the BioCreative MetaServer, the system input was changed from a single sentence to a PubMed abstract which can be segmented to multi-sentences. This means that we can take advantage of the contextual information among sentences to optimize the tagging performance. For example, "NAA" is a candidate gene mention tagged from a text of "...metabolites in 5 patients: N-acetyl-aspartate (NAA), creatine...". If we could link the definition of "NAA" to "N-acetyl-asparate" and know that "N-acetyl-asparate" is not been tagged as a gene mention, we can ignore all "NAA"s tagged in the following sentences to output and achieved performance gain by reducing false-positives. Thus, we developed a system, BIOADI [3], to identify abbreviation and its corresponding defintions for the purpose. In addition, mention collision among distinct models or systems may overwhelm the system performance and is not allowed by CALBC challenge as it will cause nested tags, such as <e><e>PKA</e> gene</e>. Hence, we built a gene mention evaluation system (GME) to assign a confidence score to each mention which is tagged by distinct models or systems. This allows us to select the one with the highest confidence for output to solve mention collision. To handle large quantity of CALBC test data, we implemented the system with MapReduce programming model and run it on a 9-nodes Hadoop cluster to process one million test abstracts. It took a time about 27 minutes to finish the entire process. We submitted 3 runs, each of them with different confidence thresholds (0.3, 0.45 and 0.6) for a balanced performance in precision and recall. See Figure 1 for an example.

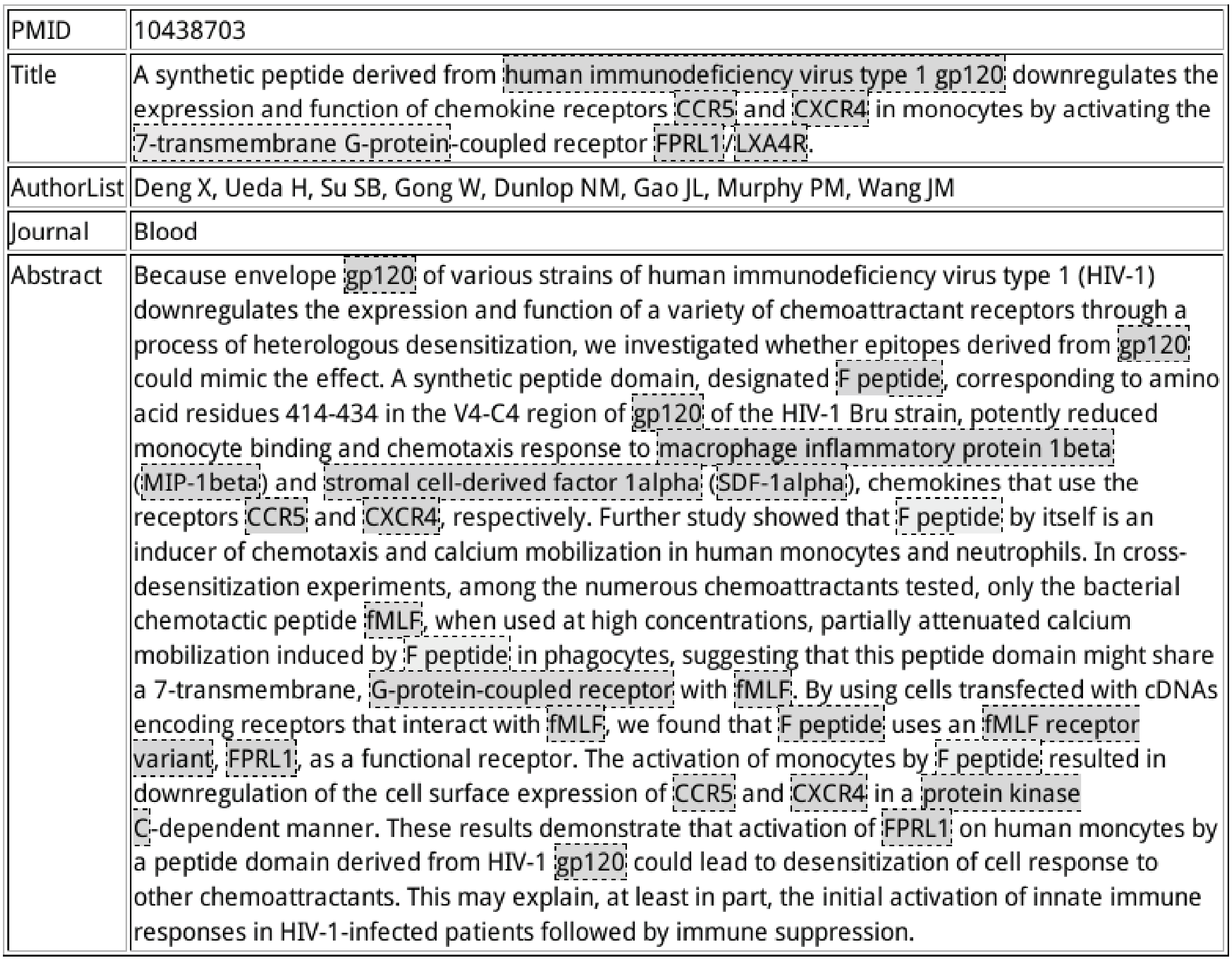
References

[1] C.-N. Hsu, Y.-M. Chang, C.-J. Kuo, Y.-S. Lin, H.-S. Huang, and I.-F. Chung. Integrating high dimensional bi-directional parsing models for gene mention tagging. *Bioinformatics*, 24(13):i286–i294, 2008.

[2] C.-J. Kuo, Y.-M. Chang, H.-S. Huang, K.-T. Lin, B.-H. Yang, Y.-S. Lin, C.-N. Hsu, and I.-F. Chung. Rich feature set, unification of bidirectional parsing and dictionary filtering for high f-score gene mention tagging. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 103–105, Centro Nacional de Investigaciones Oncologicas (CNIO), Madrid, Spain, 2007.

[3] C.-J. Kuo, M. Ling, K.-T. Lin, and C.-N. Hsu. BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. *BMC Bioinformatics*, 10(Suppl 15):S7, 2009.

Figure 1: An example of the run with the confidence threshold of 0.3



PMID	10438703
Title	A synthetic peptide derived from human immunodeficiency virus type 1 gp120 downregulates the expression and function of chemokine receptors CCR5 and CXCR4 in monocytes by activating the 7-transmembrane G-protein-coupled receptor FPRL1/LXA4R.
AuthorList	Deng X, Ueda H, Su SB, Gong W, Dunlop NM, Gao JL, Murphy PM, Wang JM
Journal	Blood
Abstract	Because envelope gp120 of various strains of human immunodeficiency virus type 1 (HIV-1) downregulates the expression and function of a variety of chemoattractant receptors through a process of heterologous desensitization, we investigated whether epitopes derived from gp120 could mimic the effect. A synthetic peptide domain, designated F peptide, corresponding to amino acid residues 414-434 in the V4-C4 region of gp120 of the HIV-1 Bru strain, potently reduced monocyte binding and chemotaxis response to macrophage inflammatory protein 1beta (MIP-1beta) and stromal cell-derived factor 1alpha (SDF-1alpha), chemokines that use the receptors CCR5 and CXCR4, respectively. Further study showed that F peptide by itself is an inducer of chemotaxis and calcium mobilization in human monocytes and neutrophils. In cross-desensitization experiments, among the numerous chemoattractants tested, only the bacterial chemotactic peptide fMLF, when used at high concentrations, partially attenuated calcium mobilization induced by F peptide in phagocytes, suggesting that this peptide domain might share a 7-transmembrane, G-protein-coupled receptor with fMLF. By using cells transfected with cDNAs encoding receptors that interact with fMLF, we found that F peptide uses an fMLF receptor variant, FPRL1, as a functional receptor. The activation of monocytes by F peptide resulted in downregulation of the cell surface expression of CCR5 and CXCR4 in a protein kinase C-dependent manner. These results demonstrate that activation of FPRL1 on human moncytes by a peptide domain derived from HIV-1 gp120 could lead to desensitization of cell response to other chemoattractants. This may explain, at least in part, the initial activation of innate immune responses in HIV-1-infected patients followed by immune suppression.