Krallinger et al. BMC Bioinformatics 2010, 11(Suppl 5):O1
http://www.biomedcentral.com/1471-2105/11/S5/O1

BMC Bioinformatics

ORAL PRESENTATION Open Access

The PPI affix dictionary (PPIAD) and BioMethod Lexicon: importance of affixes and tags for recognition of entity mentions and experimental protein interactions

Martin Krallinger[1*], Ashish V Tendulkar[2], Florian Leitner[1], Andrew Chatr-aryamontri[3], Alfonso Valencia[1]

From Workshop on Advances in Bio Text Mining
Ghent, Belgium. 10-11 May 2010

Substantial text mining efforts are being devoted to detect protein mentions and protein-protein interaction (PPI) relations from scientific articles [1,2]. In this context, the BioCreative challenge showed that the correct identification of the individual interactor proteins is still a challenging task, especially when using full text articles [2]. A systematic analysis of particularities of protein mentions in the context of interaction descriptions was nonetheless missing. Experimental biologists often use specific fusion proteins or protein-tags such as -GST, -His, -Myc, FLAG-, antibodies or fluorescent protein (GFP, YFP, CFP and RFP) tags to detect and visualize interactions. These tags are often mentioned as affixes of the target proteins in the literature. The importance of affixes in biomedical text mining had been addressed in case of affixal negation expressions [3], to consider general posttranslational modifications of proteins [4] and can be observed in trigger verbs used for interaction extraction [5].

We carried out a detailed study on the presence of common affixes belonging to interactor protein mentions in full text sentences considered by database curators as evidential support for experimentally characterized physical protein interactions. Furthermore, we tried to determine whether specific affixes might be useful to detect PPI relevant articles and to correlate affix mentions with particular interaction detection methods.

Based on examination of over 3,000 of the previously referred interaction evidence passages we have compiled a collection of 277 interaction relevant affixes (89 suffixes, 176 prefixes and 12 that could be both), which were structured into 36 affix tag classes (26 super-affix and 10 combined or sub-affix classes). Figure 1A shows the frequency of mention of each of the affix tag classes. In the resulting PPI affix dictionary (PPIAD), each affix tag class has been manually linked to experimental qualifiers represented by associated PSI-MI ontology [5] concepts by considering their concept definitions. Additionally, statistical associations of affix tag classes to PSI-MI interaction detection method concepts have been derived through curator-based annotations of the evidence passages. To overcome the limited scope and lexical coverage of terms contained in the PSI-MI ontology we build the BioMethod Lexicon, a collection of experimental method terms important for protein interaction and gene regulation relations, and characterized method term co-mentions with affix tag classes.

Within a total set of 6,300 interaction evidence sentences, 1,946 (31 %) mentioned at least one interaction relevant affix, which shows that it is a relatively common feature of interaction descriptions. Using statistical analysis of associations between affix classes and interaction detection method annotations (Chi-square test) we discovered that some of the affix classes showed strong associations to interaction methods, such as between: MI:0096 - AF_21 (MI: pull down and PPIAD: gst_pull_down_tag), MI:0676 - AF_6 (tandem affinity purification and Tandem_Affinity_Purification_tag), MI:0018 - AF_10 (two hybrid and Gal4_tag), MI:0006 - AF_4 (anti bait coimmunoprecipitation and Antibody_tag), MI:0055 - AF_15 (fluorescent

* Correspondence: mkrallinger@cnio.es
[1]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre, Madrid, Spain
Full list of author information is available at the end of the article

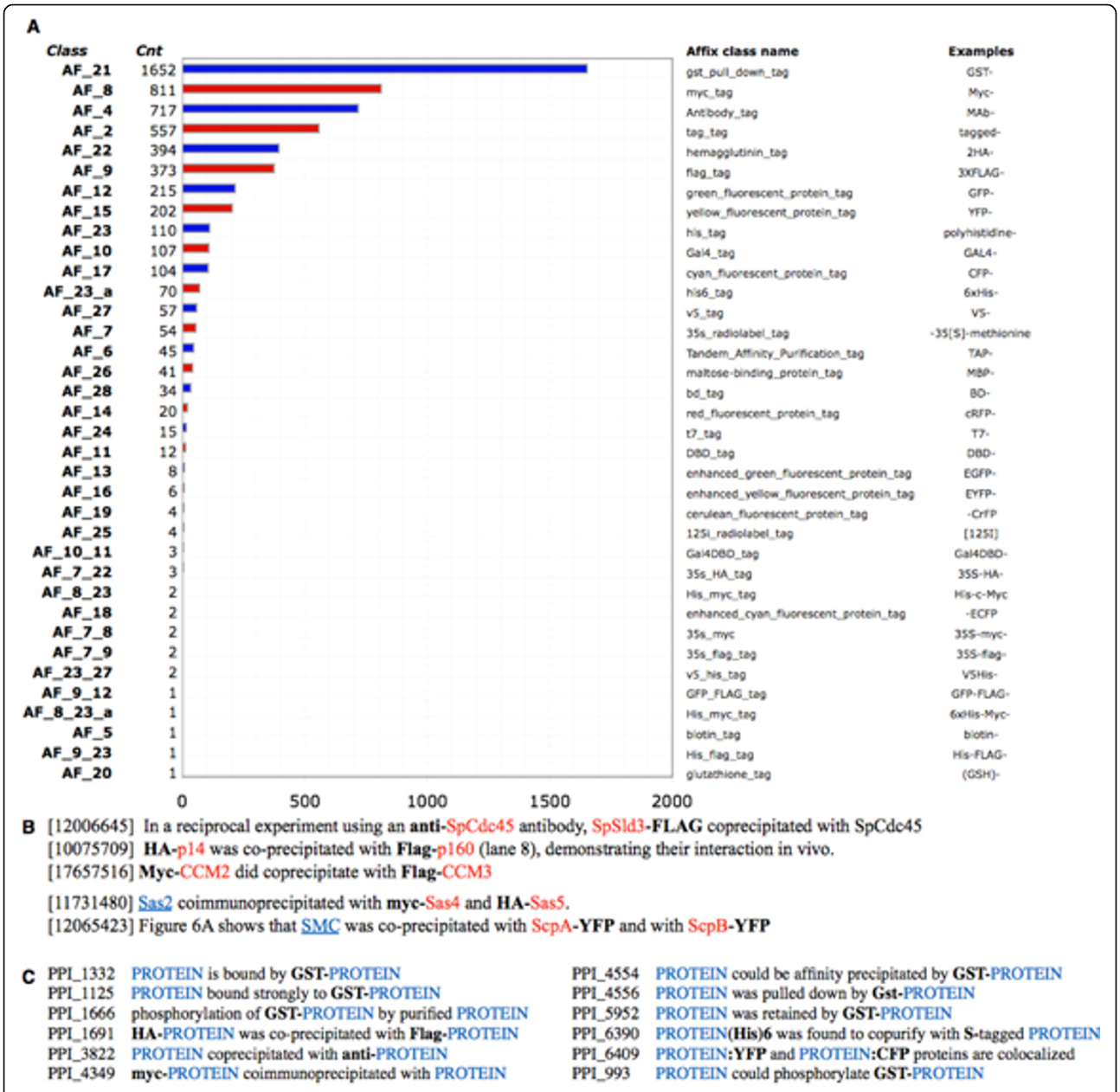

Figure 1 Main affix classes and their frequencies in the interaction evidence sentences (A). Example sentences with affix mentions (bold), affix target names (red), and interactor mentions without affixes (blue) are highlighted (B). Example interaction expression patterns containing affix related mentions (C).

resonance energy transfer and yellow_fluorescent_protein_tag), MI:0809 - AF_15 (bimolecular fluorescence complementation and yellow_fluorescent_protein_tag) or MI:0007 - AF_22 (anti tag coimmunoprecipitation and hemagglutinin_tag). This could be important to detect experimentally validated interactions and even to help associating some of them to potential interaction detection methods.

To determine if interaction affix mentions might be exploited for finding PPI relevant papers, the distribution of affix mentions across relevant and non-relevant full text articles from BioCreative II.5 training and test set was examined, showing that some of the affix classes were more frequently linked to PPI relevant articles. This indicates that they could be exploited as additional features for an article selection task.

At the level of identification of interactor proteins and interaction pairs through these affixes additional analysis is required. However, it is clear that dictionary look-up based strategies for detecting mentions of proteins need

to take into account affix handling for correct interactor identification from mention strings. For detecting interaction pairs, affix mentions can be a criterion for cases where other strategies are not able to retrieve interactions for co-mentioned entities or fail to determine whether the interaction has been experimentally proven. Difficulties encountered by affix-based PPI extraction relate to recall when one of the interactors does not display a valid affix and to precision when only some of the affix-mentioning proteins do show interactions (Fig. 1B). To address these issues, a manual collection of 799 affix relevant interaction expression patterns has been constructed (Fig. 1C). Additional materials and the PPIAD are available at: http://www.cse.iitm.ac.in/~ashishvt/research/PPIAD/.

Acknowledgements
This work was supported by the European Commission FP6 NoEs ENFIN LSHG-CT-2005-518254 and by grants BIO2007-66855 from the Spanish Ministerio de Ciencia e Innovación and by the Spanish National Bioinformatics Institute (http://www.inab.org/).

Author details
[1]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre, Madrid, Spain. [2]Department of Computer Science and Engineering, Indian Institute of Technology, Madras, India. [3]Wellcome Trust Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Mayfield Road, Edinburgh, EH9 3JR, UK.

Published: 6 October 2010

References

1. Smith L, *et al*: **Overview of BioCreative II gene mention recognition.** *Genome Biol* 2008, **9(Suppl 2)**:S2.
2. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A: **Overview of the protein-protein interaction annotation extraction task of BioCreative II.** *Genome Biol* 2008, **9(Suppl 2)**:S4.
3. Sanchez-Graillet O, Poesio M: **Negation of protein–protein interactions: analysis and extraction.** *Bioinformatics* 2007, **23(13)**:i424-i432.
4. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K: **Beyond the clause: extraction of phosphorylation information from medline abstracts.** *Bioinformatics* 2005, **Suppl 1:** i319-27.
5. Hakenberg J, Plake C, Royer L, Strobelt H, Leser U, Schroeder M: **Gene mention normalization and interaction extraction with context models and sentence motifs.** *Genome Biol* 2008, **9(Suppl2)**:S14.

doi:10.1186/1471-2105-11-S5-O1
Cite this article as: Krallinger *et al.*: The PPI affix dictionary (PPIAD) and BioMethod Lexicon: importance of affixes and tags for recognition of entity mentions and experimental protein interactions. *BMC Bioinformatics* 2010 **11**(Suppl 5):O1.