

POSTER PRESENTATION

Open Access

Functional variation of alternative splice forms in their protein interaction networks: a literature mining approach

Şenay Kafkas^{1,2*}, Ekrem Varoğlu¹, Dietrich Rebholz-Schuhmann², Bahar Taneri³

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

Analyzing protein interactions and protein functions is crucial for the analysis of complex biological processes as well as the consequences from aberrant gene products [1]. Protein-Protein Interaction Networks (PPIN) are invaluable means enabling scientists to get a global understanding of interactomes, while analyzing individual protein functions [2]. High-throughput experiments and complete literature mining analyses have been used to deliver well-structured data into scientific databases reporting on proteins, their interactions and functions. These repositories form a precious resource to scientists, but only cover a portion of the proteome and often underrepresented alternative splice forms [3].

Alternative splicing (AS) is a cellular process that produces from a single gene different physical variants of a given protein which may differ in its structure or its function. This process produces molecular variability and contributes to the complexity of the proteomes and their interactomes. The analysis of AS shows that this process is most relevant to molecular regulation processes. In this research, we attempt to identify functional variability linked to alternative splice forms within their PPINs from the scientific literature. For this purpose, we gather AS events and analyze the transcript data for 16,826 different genes from the HumanSDB3 database [4,5]. We have collected around 4 million abstracts from NCBI's PubMed by utilizing a rich search term set for each individual isoform by using Gene DB, Swissprot DB and synonym generation. We then utilize an SVM classifier which uses in-domain features together with standard term weights and have trained it on the

BioCreative-II IAS corpus (81.31% F_1 -measure on test set) for selecting those abstracts which are likely to contain interaction data. Finally, we employ another SVM classifier based on syntactic features and have trained it on the AIMed corpus (F_1 -measure of 54.20% cross validation performance) for extracting PPI information from the selected abstracts. The obtained PPIN comprises a total of 31,819 distinct interactions between 7,161 distinct proteins out of which 5,615 are considered to represent an isoform from HumanSDB3.

To the best of our knowledge, neither a genome-wide PPIN for the human protein isoforms has been built nor has their variability concerning interactions and functions been analyzed. Currently, we analyze the distribution of functional annotations based on the GO terms from the literature for all the isoforms. Both PPIN and functional annotations of the isoforms will be suitable for identifying potential interactions or functional variations of AS. Our findings are linked to the HumanSDB3 database and will be available through a publicly accessible web interface for further use.

Author details

¹Department of Computer Engineering, Eastern Mediterranean University, Famagusta, North Cyprus. ²European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. ³Faculty of Arts and Sciences, Eastern Mediterranean University, Famagusta, North Cyprus.

Published: 6 October 2010

References

1. Jaeger S, Gaudan S, Leser U, Rebholz-Schuhmann D: Integrating Protein-Protein Interactions and Text Mining for Protein Function Prediction. *BMC Bioinformatics* 2008, **9**(Suppl.8):S2.
2. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, Mering C: STRING 8-a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 2008, **37**(Database Issue):D412-D416.

* Correspondence: kafkas@ebi.ac.uk

¹Department of Computer Engineering, Eastern Mediterranean University, Famagusta, North Cyprus

Full list of author information is available at the end of the article

3. Mathivanan S, Periaswamy B, Gandhi TKB, Kandasamy K, Suresh S, Mohmood R, Ramachandra YL, Pandey A: **An evaluation of human protein-protein interaction data in the public domain.** *BMC Bioinformatics* 2006, **7**(Suppl. 5):S19.
4. Taneri B, Novoradovsky A, Snyder B, Gaasterland T: **Databases for comparative analysis of human-mouse orthologous alternative splicing.** *Lecture Notes in Bioinformatics* 2005, **3388**:123-131.
5. Taneri B, Snyder B, Novoradovsky A, Gaasterland T: **Alternative splicing of mouse transcription factors affect their DNA-binding domain architecture and is tissue specific.** *Genome Biol* 2004, **5**:R75.

doi:10.1186/1471-2105-11-S5-P1

Cite this article as: Kafkas et al.: Functional variation of alternative splice forms in their protein interaction networks: a literature mining approach. *BMC Bioinformatics* 2010 **11**(Suppl 5):P1.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

