A pipeline to explore alternative splicing events

Hendrik Schäfer*, Ina Koch*

*Molekulare Bioinformatik, Robert-Mayer-Str. 11-15, 60325 Frankfurt am Main

Alternative splicing (AS) is an import biological process, which enables a high variability in the proteom of eukaryotes. Up to 95 % of all multi exon genes in humans undergo AS [4]. Errors during the splicing can cause diseases, e.g. cancer and neuronal disorders [2][3].

In this work we present a method to collect information of alternative spliced proteins from several sources, which are combined into a single entry. A similar study was done by Stephanie Boué in 2002 for human proteins manually [1]. Seeing that many experimentally determined structure data are freely available online (∼ 58,000 new PDB entries over the last ten years), an automatic analysis of the effects of alternative splicing on the protein structure becomes possible.

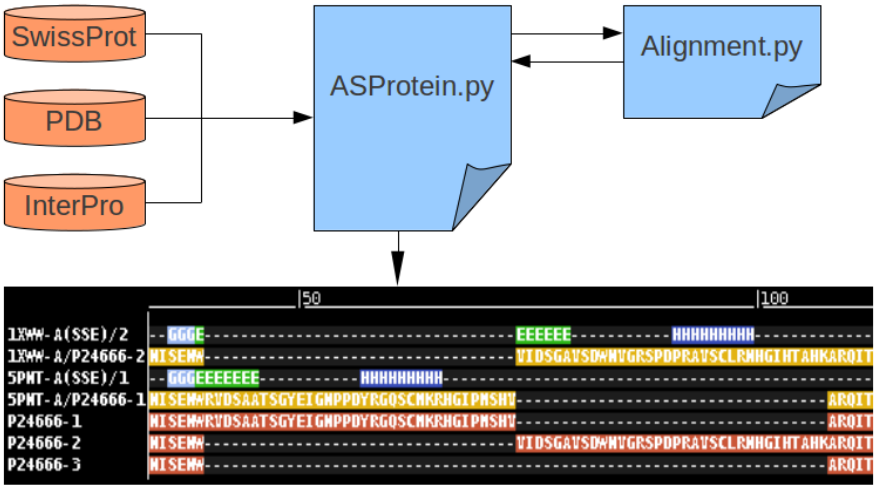

Figure 1: Overview of the program pipeline: As primary sources three data bases are used. The complete pipeline is implemented in Python, in this picture only the two main modules are shown. The different data sources are connected using an adapted alignment.

Therefore we implemented a pipeline which processes AS data based on SwissProt, PDB and InterPro. We identified possible changes of the protein fold on secondary structure level. Therefore we used DSSP to assign secondary structure elements (SSEs), which are then classified according the splicing effects. To connect all data sources, we developed an adapted alignment, which contains all isoform sequences, PDB sequences and the corresponding secondary structures. This approach makes it possible to

compare alternative splicing events and protein structure directly. Moreover, we implemented a visualization tool for the results and all alignments can be watched in Portable Network Graphics format.

Parsing 2,289 proteins in SwissProt we identified 758 proteins, which contain at least one alternatively spliced SSE. Ten proteins with PDB data of both variants of a single splicing event were found and a direct comparison of the tertiary structure is possible. We provide a statistical overview and explain a case study in detail. At the current state almost all entries of the resulting data set belong to mammals, thus an analysis of different organism groups is not possible. In most cases the reason for that are missing 3D-structure information.

The pipeline we implemented can be used for large-scale analysis of AS data. As we are using SSEs to identify effects on the protein structure most calculations can be done on primary structure level and thus are not computationally expensive. The complete run of all 2,289 proteins took 18 h (on a single 2000 MHz CPU). We store all processed information in flat file format and all data can be quickly accessed for further investigations.

References

[1] S. Boué, M. Vingron, E. Kriventseva, and I. Koch. Theoretical analysis of alternative splice forms using computational methods. *Bioinformatics Vol. 18 Suppl. 2*, 2002.

[2] B. K. Dredge, A. D. Polydorides, and R. B. Darnell. The splice of life: alternative splicing and neurological disease. *Nat Rev Neurosci. 2001 Jan;2(1):43-50*, 2001.

[3] M. M. Feldkamp, L. Angelov, and Guha A. Neurofibromatosis type 1 peripheral nerve tumors: aberrant activation of the Ras pathway. *Surg Neurol. 1999 Feb;51(2):211-8.*, 1999.

[4] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics 40, 1413 - 1415 (2008)*, 2008.