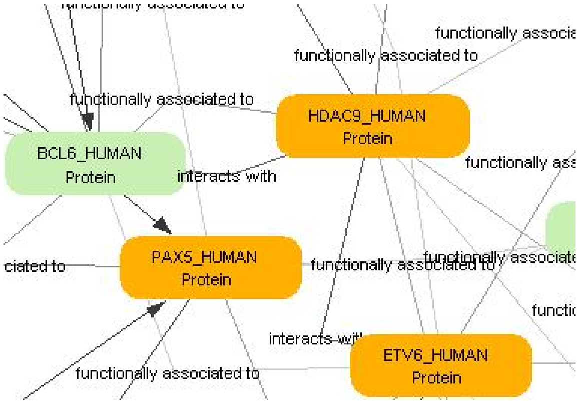
**Biomedical data mining, data-driven hypothesis generation in cancer research**



Research in life sciences is only possible today with access to online databases. Extracting information useful for medical researchers and practitioners is possible now with the methods of parallel data mining, simultaneously applied to high throughput data, molecular databases and medical publications.

## Mining mass spectrometric/proteomics data

High throughput mass spectrometry analysis produces large amounts of noisy data that have to be filtered and preprocessed with computational tools before subjected to detailed analysis and interpretation. Our strategy uses principles borrowed from cognitive psychology for identifying patterns in mass spectra. Namely, the human mind is able to capture holistic features in complex sensory inputs, and we trust that similar principles can be applied to abstract data structures. The bioinformatics support of proteomics research is a central theme in our projects. We develop new tools capable of filtering and processing large data streams characteristic of high throughput analysis workflows.

## Medical hypothesis generation

Hypothesis generation refers to generating surprising, non-trivial suppositions and explanations based on information extracted from textual resources. From a data-mining perspective, text-based hypothesis generation is a case of link discovery, i.e. a hypothesis can be considered as an undiscovered relation between pre-existing knowledge items. Early success stories include the discovery of therapies for Raynaud's disease and migraine. In the genomics era, hypotheses are often formulated as relations involving molecular entities, such as genes, proteins, drugs, metabolites, etc., so the use of textual resources needs to be combined with molecular databases, and often, with new experimental data generated by the user. A typical example of application is finding undiscovered links and synergisms between approved pharmaceuticals, as drug combinations can reach the applications phase much faster then novel drugs. A promising area is the study of synergisms that may exist between generic and targeted therapeutic agents or the design of cocktail therapies for complex diseases.

The emphasis of current cancer therapy is shifting from traditional chemotherapy to targeted drugs. Such therapies rest on two fundamental motives: i) the use of targeted pharmacons that act on one or a few molecular targets specific to tumor cells, and ii) identification of biomarkers suitable for the prediction of drug response. High throughput technologies provide massive amounts of data that can be processed from many viewpoints, the average research groups however lack the necessary and sometimes very extensive, bionformatics repertoire. Our aim is to develop on-line facilities that are able to integrate high throughput data with complex algorithmic procedures that allow identification of biomarkers or statistical targets. An additional goal is to create prediction systems that can help point of care diagnostics applications.

**Project Participants:**
Balázs Ligeti, PhD student
Prof. Sándor Pongor, PI

**Collaborators:**

Dr. Balázs Győrffy, Research Laboratory of Pediatrics and Nephrology, Hungarian Academy of Sciences, Semmelweiss University, Budapest, Hugary
Beáta Reiz, Szeged Biological Centre, Szeged, Hungary
Dr. Ingrid Petrič, Centre for Systems and Information Technologies, University of Nova Gorica, Slovenia
Dr. Mike Myers, International Centre for Genetic Engineering and Biotechnology, Trieste, Italy
Dr. Attila Kertész-Farkas, International Centre for Genetic Engineering and Biotechnology, Trieste, Italy

**References**

267. Reiz, B.; Kertész-Farkas, A.; Dhir, S.; Pongor, S.; Myers, M.P. (2013). Chemical rule-based filtering of MS/MS spectra. *Bioinformatics*. 29(7), 925-932

266. Petric, I.; Ligeti, B.; Gyorffy, B.; Pongor, S. (2013). Biomedical Hypothesis Generation by Text Mining and Gene Prioritization. *Protein Pept Lett.*.

263. Vera, R.; Perez-Riverol, Y.; Perez, S.; Ligeti, B.; Kertész-Farkas, A.; Pongor, S. (2013). JBioWH: an open-source Java framework for bioinformatics data integration. *Database*. 2013

261. Reiz, B.; Busa-Fekete, R.; Pongor, S.; Kovács, I. (2013). Closure Enhancement in a Model Network with Orientation Tuned Long-Range Connectivity. *Bistable Perception Special Issue of Learning & Perception*. 5, 119–148

257. Kertész-Farkas, A.; Reiz, B.; Myers, M.P.; Pongor, S. (2012). Database Searching In Mass Spectrometry Based Proteomics. *Current Bioinformatics*. 7(2), 221-230

256. Reiz, B.; Myers, M.P.; Pongor, S.; Kertész-Farkas, A. (2012). Precursor Mass Dependent filtering of Mass Spectra for Proteomics Analysis. *Protein and Peptide Letters(In Press)*.

255. Reiz, B.; Kertész-Farkas, A.; Pongor, S.; Myers, M.P. (2012). Data preprocessing and filtering in Mass Spectrometry based proteomics. *Current Bioinformatics*. 7(2), 212-220

251. Cserhati, M.; Turoczy, Z.; Zombori, Z.; Cserzo, M.; Dudits, D.; Pongor, S.; Gyorgyey, J. (2011). Prediction of new abiotic stress genes in Arabidopsis thaliana and Oryza sativa according to enumeration-based statistical analysis. *Mol Genet Genomics*. 285(5), 375-391

248. Kertész-Farkas, A.; Reiz, B.; Myers, M.P.; Pongor, S. (2011). PTMSearch: A Greedy Tree Traversal Algorithm for Finding Protein Post-Translational Modifications in Tandem Mass Spectra. *Lecture Notes in Computer Science 6912*. 2, 162-176

247. Reiz, B.; Pongor, S. (2011). Psychologically inspired, rule-based outlier detection in noisy data. *13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing: 2011*. 1, 131-136.

241. Kuzniar, A.; Dhir, S.; Nijveen, H.; Pongor, S.; Leunissen, J.A. (2010). Multi-netclust: an efficient tool for finding connected clusters in multi-parametric networks. *Bioinformatics*. 26(19), 2482-2483

235. Kuzniar, A.; Lin, K.; He, Y.; Nijveen, H.; Pongor, S.; Leunissen, J.A. (2009). ProGMap: an integrated annotation resource for protein orthology. *Nucleic Acids Res*. 37(Web Server issue), W428-34.

204. Csermely, P.; Agoston, V.; Pongor, S. (2005). The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol Sci*. 26(4), 178-182.