Combining Text Mining and Microarray Analysis

Christoph M. Friedrich and Michaela Gündel
Fraunhofer Institute Algorithms and Scientific Computing (SCAI)
Department of Bioinformatics

Contact: friedrich@scai.fraunhofer.de

The financial support of the European Commission is gratefully acknowledged. Material in this presentation reflects only the author’s views and the Commission is not liable for any use that may be made of the information contained herein.
Outline

- Introduction
- Microarray Workflow
- Text mining – Named Entity Recognition for Genes
- Live Demo - Combining Text mining and Microarray analysis
@neuLink: Linking Genetics to Disease

Textual information + Public Biomedical Databases + Disease Specific Genes

Text Mining

Candidate network of Genes with high Evidence

Experimental test + Data Mining

ATCGAATTA AT

Friedrich, C. M.; Dach, H.; Gattermayer, T.; Engelbrecht, G.; Benkner, S. & Hofmann-Apitius, M.
@neuLink: A Service-oriented Application for Biomedical Knowledge Discovery
Microarray Workflow results master thesis Michaela Gündel (B-IT)

Venn Diagram (differentially expressed genes)

Gene Ontology analysis

Gündel, M. ArrayProcess: Work Flow for Microarrays; Masters thesis, Life Science Informatics at Bonn-Aachen International Center for Information Technology (B-IT); Germany, 2007
ProMiner: Dictionary based Named Entity Recognition

A Nomenclature Human for Gene names exists (HUGO) but nobody uses it.


We need Named Entity Recognition but:

Gene and protein name constraints:

- Multiple synonyms
- Multi word terms
- Spelling variants
- Nested names
- Common names – AND, CAD

<table>
<thead>
<tr>
<th>Gene</th>
<th>Multi-synonym examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNC</td>
<td>Neuronectin, GMEM, tenascin, HXB, cytotactin, hexabrachion</td>
</tr>
<tr>
<td>COL1A1</td>
<td>Collagen, type I, alpha 1 Collagen alpha 1(I) chain Alpha 1 collagen Alpha-1 type I collagen</td>
</tr>
<tr>
<td></td>
<td>TNF receptor 1 collagen, type I, alpha receptor</td>
</tr>
</tbody>
</table>
In the second case, a missense mutation in COL1A1 (substitution of arginine by cysteine) results in a type I EDS phenotype with clinically normal-appearing dentition. Tooth samples are investigated by using light microscopy (LM), transmission electron microscopy (TEM) and immunostaining for types I and III collagen, and tenascin.
ProMiner: Performance in International Benchmarking

Participation of SCAI in „Critical Assessments of Text Mining in Biology“ (BioCreAtIvE) 2004 and 2006

<table>
<thead>
<tr>
<th></th>
<th>Mouse BioCreAtIvE I</th>
<th>Fly BioCreAtIvE I</th>
<th>Yeast BioCreAtIvE I</th>
<th>HUMAN BioCreAtIvE II</th>
</tr>
</thead>
<tbody>
<tr>
<td>best automatic system</td>
<td>ProMiner system</td>
<td>ProMiner system</td>
<td>ProMiner system</td>
<td>ProMiner system</td>
</tr>
<tr>
<td>F-measure</td>
<td>0,79</td>
<td>0,79</td>
<td>0,82</td>
<td>0,92</td>
</tr>
</tbody>
</table>

Lynette Hirschman; Alexander Yeh; Christian Blaschke & Alfonso Valencia, "Overview of BioCreAtIvE: critical assessment of information extraction for biology." BMC Bioinformatics, 2005, 6 Suppl 1, S1


Special Issue on BioCreative II, Genome Biology 2008, 9(Suppl 2):S2
Combining Textmining and Microarray analysis

SCAIView – Knowledge Environment

M. Hofmann-Apitius; J. Fluck; L. I. Furlong; O. Fornes; C. Kolarik; S. Hanser; M. Boeker; S. Schulz; F. Sanz; R. Klinger; H.-T. Mevissen; T. Gattermayer; B. Oliva & C. M. Friedrich, "Knowledge Environments Representing Molecular Entities for the Virtual Physiological Human", Philosophical Transactions of the Royal Society A, 2008, 366(1878), 3091-3110.
Acknowledgements

- Martin Hofmann-Apitius
- Juliane Fluck
- Theo Mevissen, Tobias Gattermayer, Bernd Müller, Patricia Laine, Christian Ebeling, Roman Klinger, Ye Cao
- Partners at Kings College London: Saliha Yilmaz and John McGregor
- Partners at IMIM (Barcelona) especially Laura I. Furlong, Oriol Fornes, Anna Bauer-Mehren and Baldo Oliva
- Partners of the @neurlIST consortium

This work has been partially funded in the framework of the European integrated project @neurlIST, which is co-financed by the European Commission through the contract no. IST-027703 (see http://www.aneurist.org)