



aneurIST

Integrated biomedical informatics for the management of cerebral aneurysms

## 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](mailto: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.



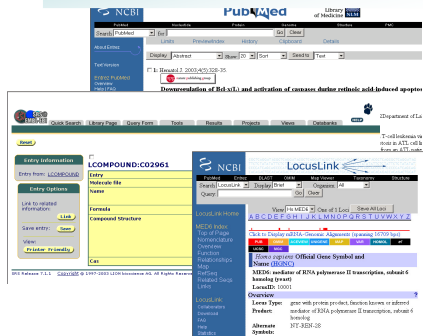
## Outline

- Introduction
- Microarray Workflow
- Text mining – Named Entity Recognition for Genes
- Live Demo - Combining Text mining and Microarray analysis

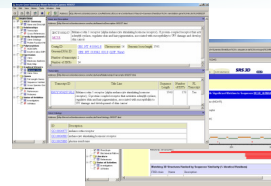


# @neuLink: Linking Genetics to Disease

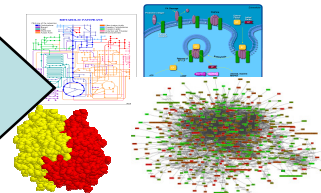
Textual information



Public Biomedical Databases



Disease Specific Genes

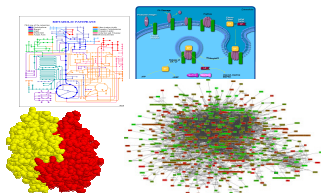


Text Mining

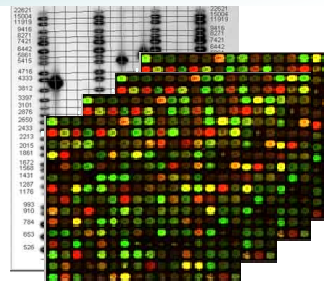


Candidate network of Genes with high Evidence

Disease Specific Genes



Experimental test



Data Mining

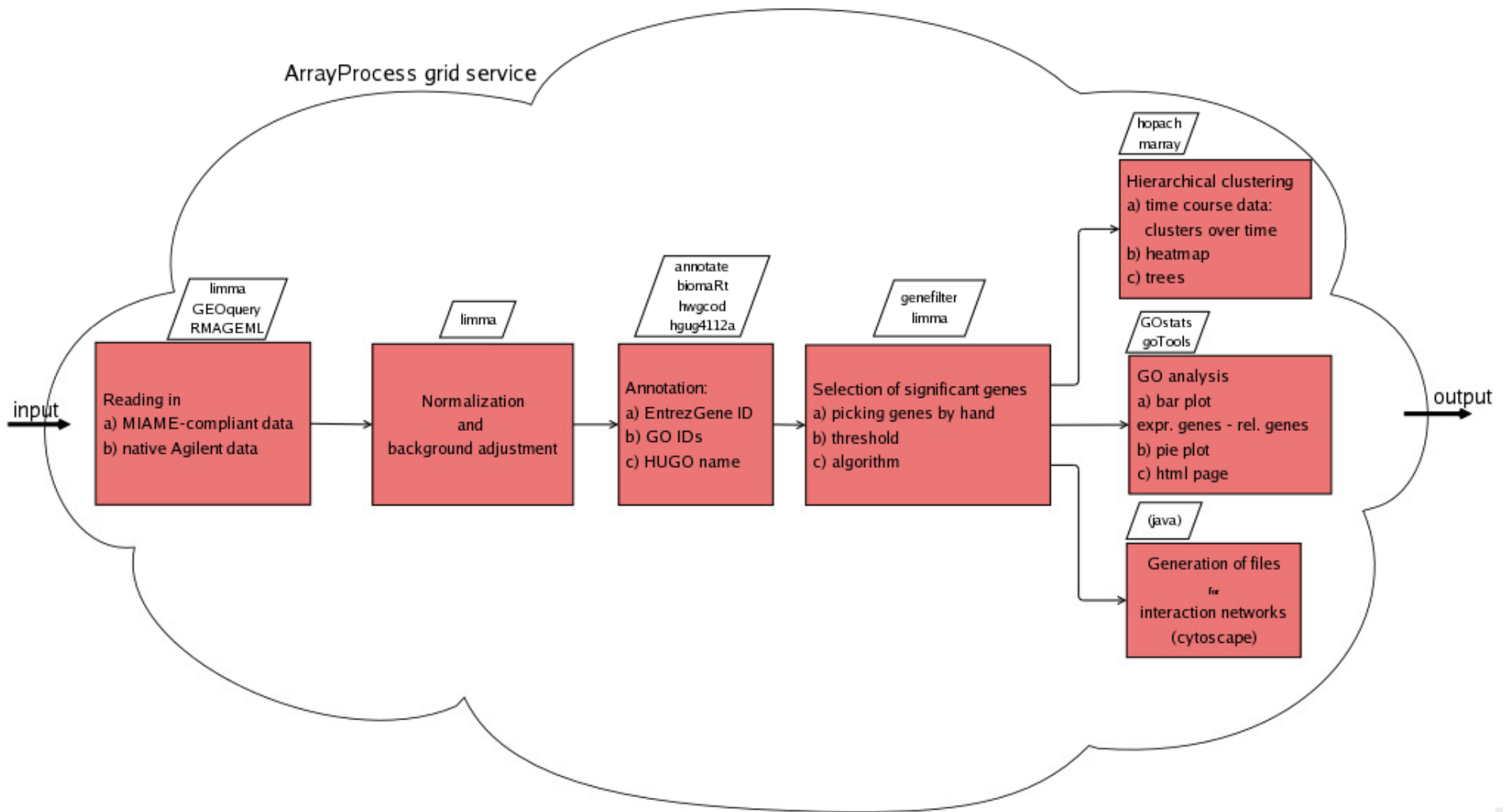
ATCGAATTA  
AT



Friedrich, C. M.; Dach, H.; Gattermayer, T.; Engelbrecht, G.; Benkner, S. & Hofmann-Apitius, M.  
**@neuLink: A Service-oriented Application for Biomedical Knowledge Discovery**  
*Proceedings of the HealthGrid 2008, IOS Press, 2008, 165-172*



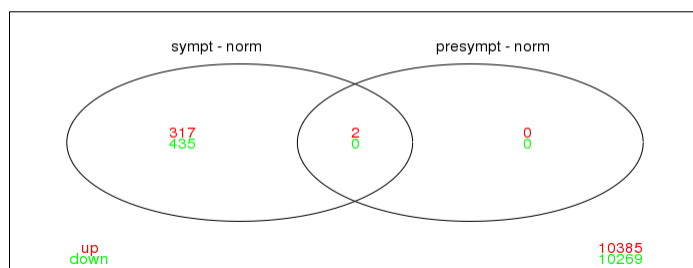
# Microarray Workflow





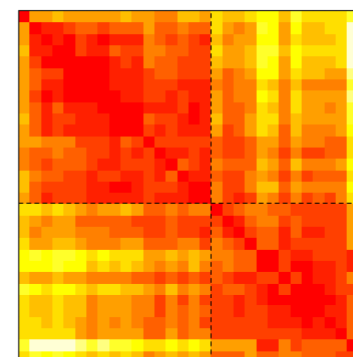
## Microarray Workflow results master thesis Michaela Gündel (B-IT)

Venn Diagram (differentially expressed genes)

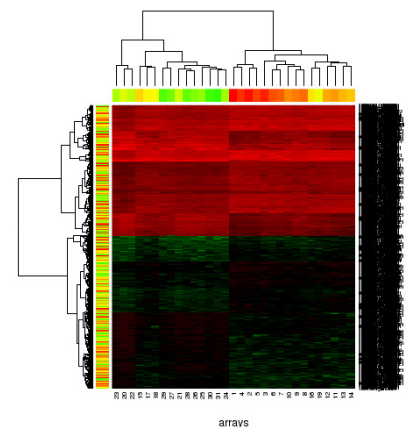
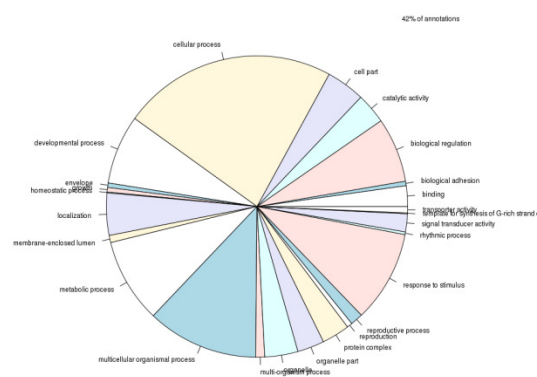


Clustering results

Array Distance Matrix



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.

J. Tamames and A. Valencia “**The success (or not) of HUGO nomenclature**”, Genome Biol. 2006; 7(5): 402.



We need Named Entity Recognition but:

Gene and protein name constraints:

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

<b>TNC</b>	<b>Neuronectin, GMEM, tenascin, HXB, cytotactin, hexabrachion</b>
	<b>Interleukin 1 alpha Tumor necrosis factor beta</b>
<b>COL1A1</b>	<b>Collagen, type I, alpha 1 Collagen alpha 1(I) chain Alpha 1 collagen Alpha-1 type I collagen</b>
	<b>TNF receptor 1 collagen, type I, alpha receptor</b>



# ProMiner: Entity Recognition and Normalization

Entrez Gene

GeneID: 3371  
 Official Symbol: TNC  
 Name: tenascin C (hexabrachion)

Accession number: P24821  
 Protein Name: tenascin

<b>TNC</b>	Neuronectin, GMEM, tenascin, HXB, cytotactin, hexabrachion
------------	--

Entrez Gene

GeneID: 1277  
 Official Symbol: COL1A1  
 Name: collagen, type I, alpha 1

Accession number: P02452  
 Protein Name: Collagen alpha-1(I) chain

<b>COL1A1</b>	Collagen, type I, alpha 1 Collagen alpha 1(I) chain Alpha 1 collagen Alpha-1 type I collagen
---------------	---

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

	Mouse BioCreAtIvE I		Fly BioCreAtIvE I		Yeast BioCreAtIvE I		HUMAN BioCreAtIvE II	
	best automatic system	<b>ProMiner system</b>	best automatic system	<b>ProMiner system</b>	best automatic system	<b>ProMiner system</b>	best automatic system	<b>ProMiner system</b>
F- measure	0,79	0,79	0,82	0,82	0,92	0,9	0,81	0,8

- 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
- Alexander A. Morgan & Lynette Hirschmann, “**Overview of BioCreative II Gene Normalization**” *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, 2007, 17-27
- Special Issue on BioCreative II , *Genome Biology* 2008, 9(Suppl 2):S2





## Combining Textmining and Microarray analysis SCAIVIEW – Knowledge Environment

The screenshot shows the SCAIVIEW web application interface. The browser window displays the URL 'http://aneulink.aneurist.org/'. The main content area shows search results for 'intracranial AND aneurysm\* AND'. A table lists 121 items found, displaying 1 to 100. A magnifying glass highlights a specific entry in the table:

Select	Entity	Relative Entropy	Aneurysm Linker Degree	Refer
<input checked="" type="checkbox"/>	ELN	0.3183	0.0	
<input type="checkbox"/>	PKD1	0.2950	3.0	
<input checked="" type="checkbox"/>	NOS3	0.1457	4.0	
<input checked="" type="checkbox"/>	COL3A1	0.1380	0.0	
<input checked="" type="checkbox"/>	SERPINA1	0.1203	0.0	
<input checked="" type="checkbox"/>	MMP9	0.1087	5.0	
<input type="checkbox"/>	APOE	0.0981	0.0	
<input checked="" type="checkbox"/>	COL1A2	0.0769	1.0	
<input checked="" type="checkbox"/>	TMP1	0.0715	2.0	
<input checked="" type="checkbox"/>	TIMP1: TIMP1: TIMP metalloproteinase inhibitor 1: H5027452	0.0552	4.0	
<input checked="" type="checkbox"/>	MMP2	0.0636	5.0	
<input type="checkbox"/>	MFAP4	0.0831	0.0	
<input checked="" type="checkbox"/>	ACE	0.0503	0.0	
<input checked="" type="checkbox"/>	TIMP3	0.0488	4.0	871
<input type="checkbox"/>	CCL4A1	0.0480	2.0	77
<input checked="" type="checkbox"/>	FBN2	0.0453	0.0	109

# SCAIVIEW

Best presented  
in a Live Demo

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 @neurIST consortium

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

