SCAIView - Lucene for Life Science Knowledge Discovery

Dr. Christoph M. Friedrich
E-mail: friedrich@scai.fraunhofer.de

Fraunhofer Institute Algorithms and Scientific Computing
Department of Bioinformatics

Schloss Birlinghoven

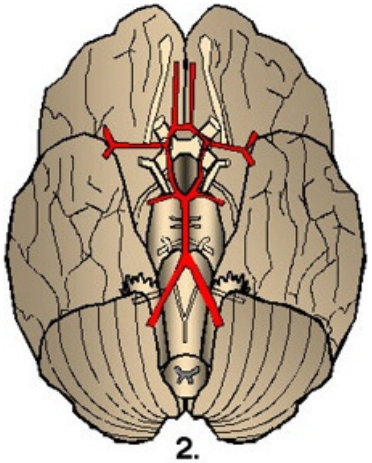
Outline

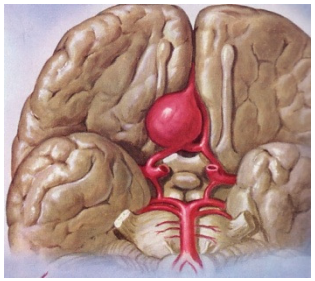
- Introduction to the European Project @neurIST and its vision
- Named Entity Recognition for the Life Sciences
- Semantic/Ontological Search concepts
- Lucene based SCAIView Knowledge Discovery Environment (Live Demo)
- Acknowledgements

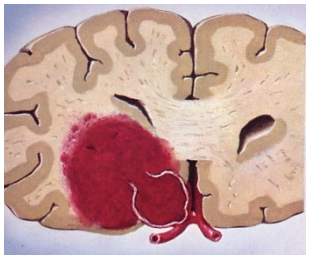
SCAI
Fraunhofer Institute Algorithms and Scientific Computing

Intracranial Aneurysms, a model disease

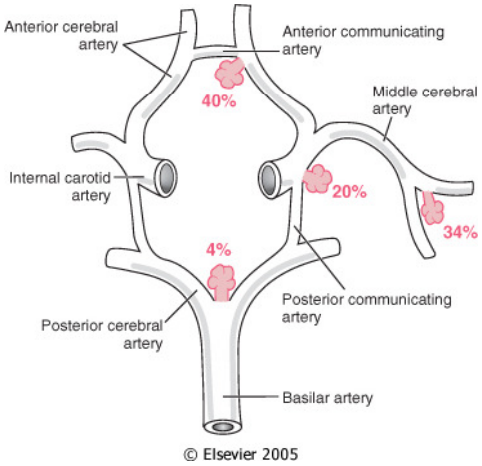
- **Intracranial Aneurysms (IA) prevalence of approx. 2-5% in the european population**
- **Risk of rupture low (subarachnoid hemorrhage) approx. 0.01% p.a. (36,000 p.a. in Europe) – mortality approx. 1/3**
- **Better imaging → more and more asymptomatic IA are detected (patients feel to have a time bomb in their head)**

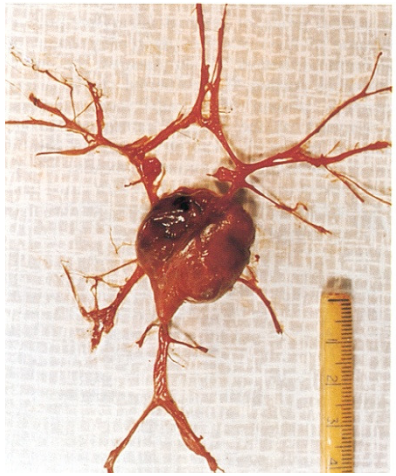






Circle of Willis





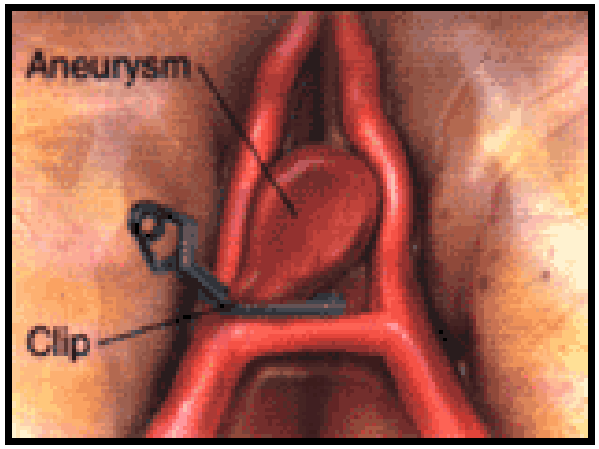
Giant Aneurysm

SCAI
Fraunhofer Institute Algorithms and Scientific Computing

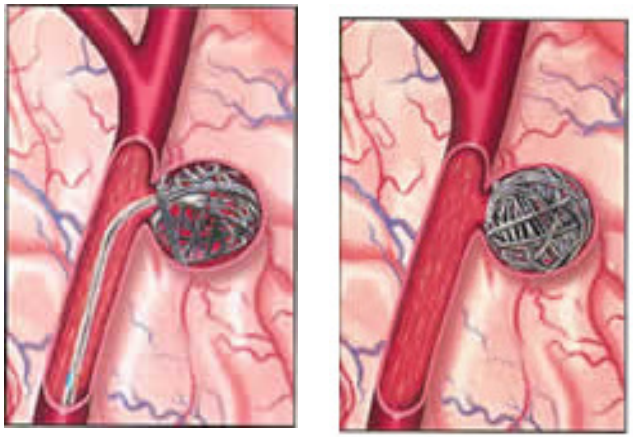
Intracranial Aneurysms, treatment options

- In general 4 treatment options, all are risky and experts discuss controversely

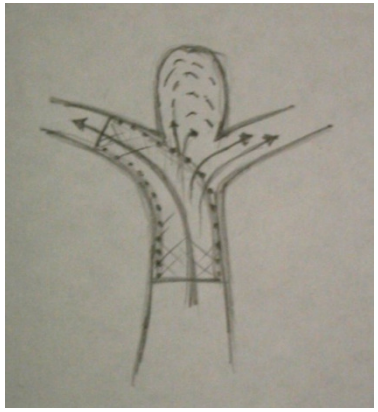
  1. Do nothing and wait
  2. Neurosurgical intervention with clipping
  3. Endovascular treatment with platinum coils
  4. Endovascular treatment with flow diverting stent (new in @neurIST)



Clipping



Coiling



Stenting

Fraunhofer Institute Algorithms and Scientific Computing

Known Risk factors

Risk factors assessed by Internal Cochrane Report
(Mike Clarke, University of Oxford)

- **Risk factors to develop an IA**
  - Genetic Factors: Ehlers Danlos Syndrome, Polycistic Kidney Disease, Moya Moya, ...
  - Family history, Hypothesis of Viral infections, ...
  - **Gender** - relative risk men to women **0.8** (95% CI 0.5 to 1.1)
- **Risk factors for rupture**
  - Size and Location (Posterior higher risk than Anterior)
  - Family history, Multiple Aneurysms
  - Hypertension, Stimulant Consumption
  - **Gender** (females have a higher relative risk **2.1** (95% CI 1.1 to 3.9))
  - Age ...

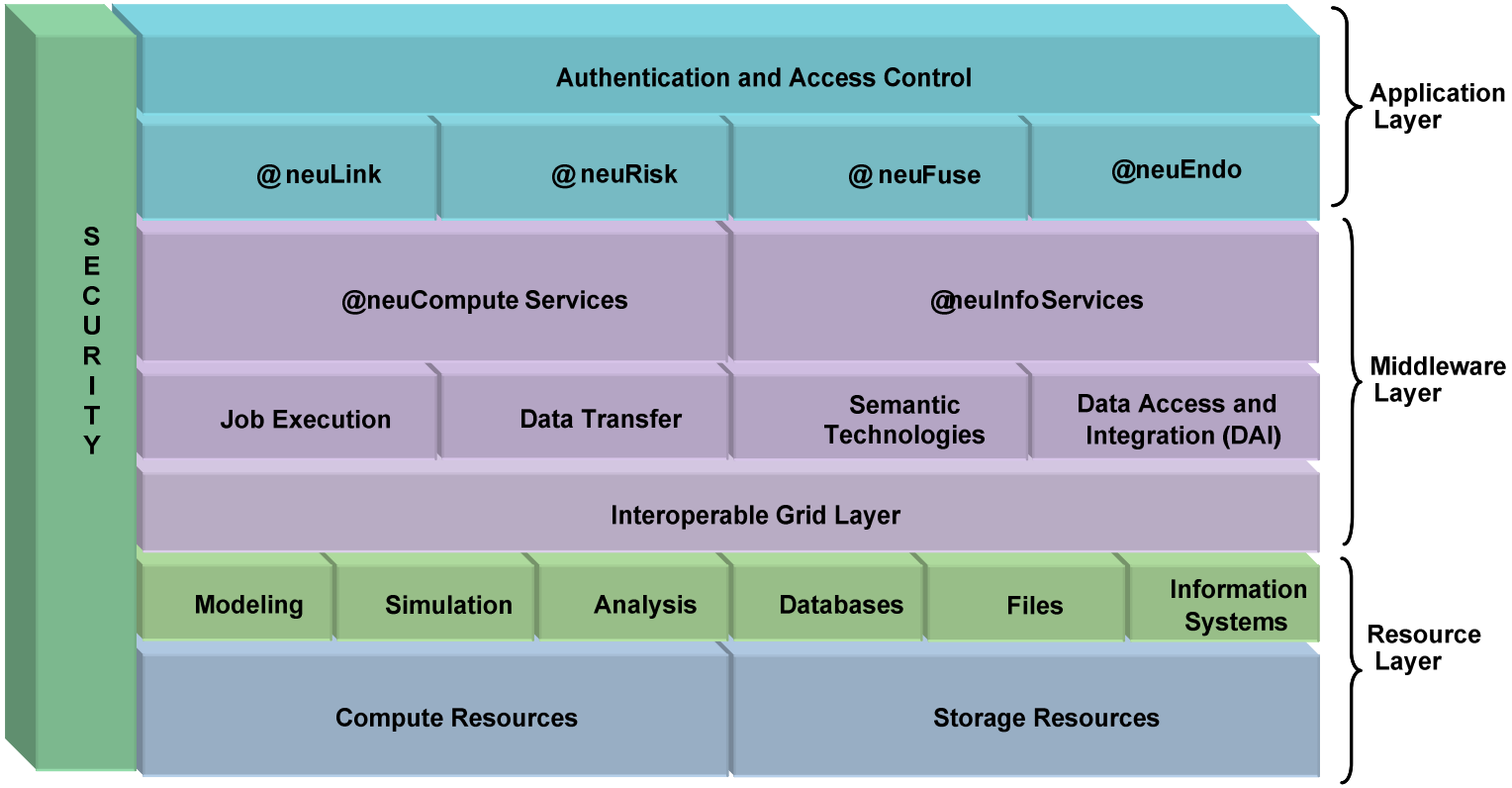
SCAI
Fraunhofer Institute Algorithms and Scientific Computing

European Integrated Project @neurIST

- Development of an integrated healthcare infrastructure to improve the decision support for IA
- Integrated European FP6 Project with 32 partners, 12 Mio EUR funding, 1/2006-4/2010 **http://www.aneurist.org**
- 7 clinical centers (+ external centers in a Virtual Hospital e.g. Uni Bonn), study size: 1200 patients
- Objective: **predict the risk of rupture for an individual patient**
- Multimodal data:
  - Imaging data, Haemodynamic models
  - Clinical data (phenotypes)
  - Genetic data (SNP Illumina 610Quad, Illumina HumanRef-8 V2 expression analysis data)
  - Epidemiological data (Erasmus MC, several databases, e.g. IPCI)
  - **Literature data** (Medline)

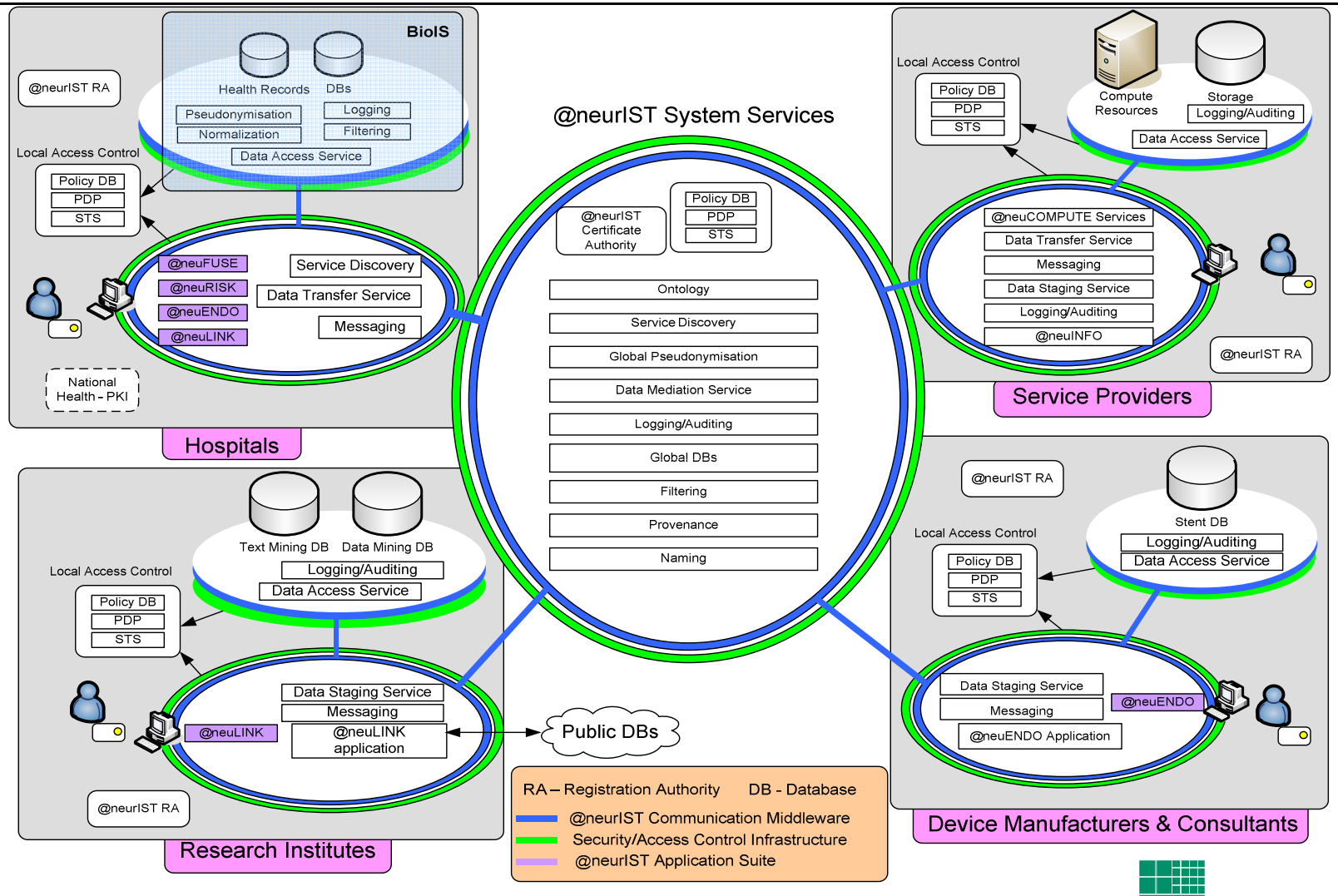
Fraunhofer Institute Algorithms and Scientific Computing

Layered Architecture View of the Service oriented architecture



H. Rajasekaran; L. L. , Iacono; P. Hasselmeyer; J. Fingberg; P. Summers; S. Benkner; G. Engelbrecht; A. Arbona; A. Chiarini; C. M. Friedrich; M. Hofmann-Apitius; K. Kumpf; B. Moore; P. Bijlenga; J. Iavindrasana; H. Mueller; R. D. Hose; R. Dunlop & A.F. Frangi „**@neurIST – Towards a System Architecture for Advanced Disease Management through Integration of Heterogeneous Data, Computing, and Complex Processing Services**“ *Proceedings of 21st IEEE International Symposium on computer-based medical systems,* **2008**, 361-366.

Fraunhofer Institute Algorithms and Scientific Computing

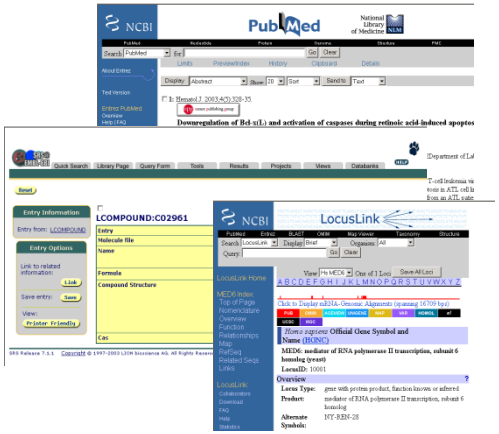


SCAI

Fraunhofer Institute Algorithms and Scientific Computing

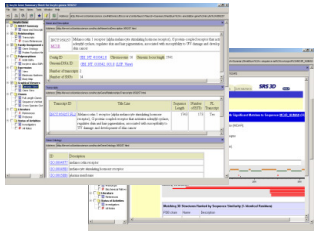
@neuLink: Linking Genetics to Disease

Textual information



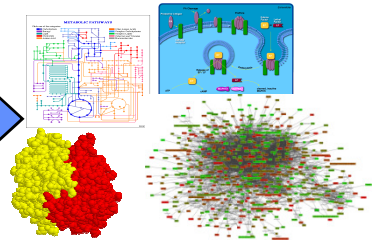
+

Public Biomedical Databases

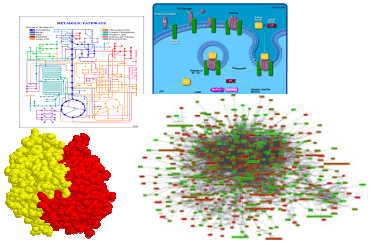


Disease Specific Interaction Networks

Textmining

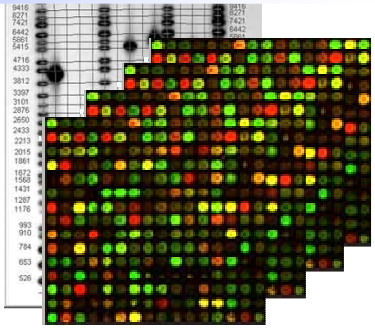


Disease Specific Interaction Networks



+

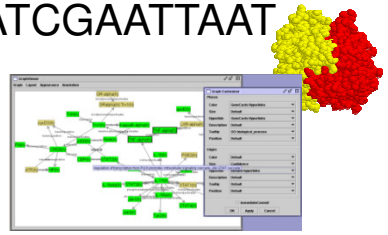
Experimental data/ Clinical data



Datamining

Candidate network of Genes with high Evidence

ATCGAATTAAT



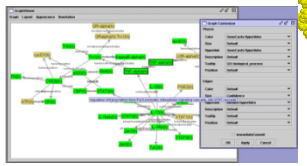
SCAI

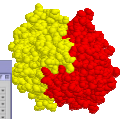
Fraunhofer Institute Algorithms and Scientific Computing

@neuLink: Linking Genetics to Disease (2)

Candidate network of Genes with high Evidence

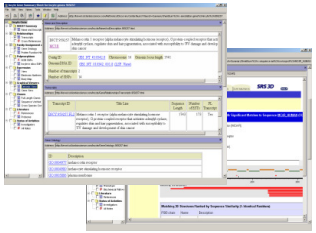
ATCGAATTAAT





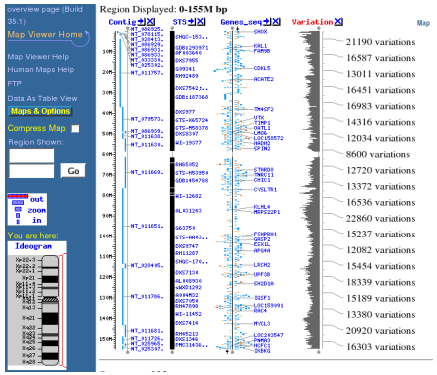
+

Public Biomedical Databases



Textmining
Datamining

Genetic Disease Marker (SNP)



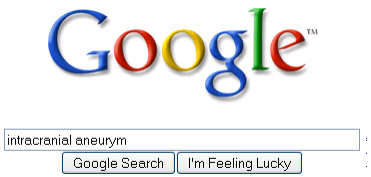
Friedrich, C. M.; Dach, H.; Gattermayer, T.; Engelbrecht, G.; Benkner, S. & Hofmann-Apitius, M.
@neuLink: A Service-oriented Application for Biomedical Knowledge Discovery
Proceedings of the HealthGrid 2008, IOS Press, **2008**, 165-172

SCAI
Fraunhofer Institute Algorithms and Scientific Computing

Some Search Concepts and definitions

What we are used to do:

- **Ad hoc fulltext Queries:**
  Non predefined queries for keywords in documents, Google type „Aspirin“



Large Set of “Relevancy?” Ranked Documents, now we have to skim through ☹

Is this *Knowledge Discovery?*

Let‘s go beyond Google, What technologies are available? What do we want?

Typically for **decision support**, „**Is a side effect for drug *x* in disease *y* or related diseases known?**“, „ stop project *x*, it‘s patented already“

SCAI
Fraunhofer Institute Algorithms and Scientific Computing

Information Extraction from Unstructured Text

- Most information in the Life Sciences is contained in Publications (at the moment 19Mio in Medline)
- Every day approx. 3000 new articles are indexed
- Human curated Databases for Disease specific Candidate Genes e.g. AlzGene DB
- Textmining is an automated way to extract this information
- Done with Dictionary, rule based and machine learning methods
- Finding and linking to a database (normalization/disambiguation)
- In this context genes, cytobands, Marker Identifiers, Variations and Risk Factors are of interest
- Knowledge Discovery expects novelty → Statistically aggregated or normalized information provides this novelty
- Knowing the published helps to reconfirm results or prevent duplication of work

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**

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

SCAI

Fraunhofer Institute Algorithms and Scientific Computing

ProMiner: Entity Recognition and Normalization

Entrez Gene

GeneID: 3371

Official Symbol: TNC

Name: tenascin C (hexabrachion)

swissprot

Acession number: P24821

Protein Name: tenascin

TNC	Neuronectin, GMEM, tenascin, HXB, cytotactin, hexabrachion

Entrez Gene

GeneID: 1277

Official Symbol: COL1A1

Name: collagen, type I, alpha 1

swissprot

Acession number: P02452

Protein Name:

Collagen alpha-1(I) chain

COL1A1	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.**

SCAI
Fraunhofer Institute Algorithms and Scientific Computing

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	ProMiner system	best automatic system	ProMiner system	best automatic system	ProMiner system	best automatic system	ProMiner system
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” to appear.

Gene Variations in Text

A Nomenclature exists, but it is not widely adopted

J. T. den Dunnen & S. E. Antonarakis "**Nomenclature for the description of human sequence variations.**" *Hum Genet,* **2001***, 109*, 121-124

Example: The FGFR2 exon 7 sequencing showed the classical Apert syndrome **c.758C > G** transversion (**p.Pro253Arg**).

- More often you find the old Nomenclature or individual adoptions:

Example: Nine polymorphisms were identified, 3 located in TIMP-1 **(-19C>T**, **261C>T**, **372T>C**), …

- Or the difficult natural language represented ones:

Example: This SNP induces **Ala** to **Pro substitution** at **amino acid 459** located on a triple-helical domain.

- Or the easy way:

Example: Only one variant, **rs767603**, at chromosome 14q23, …

Fraunhofer Institute Algorithms and Scientific Computing

Finding Gene Variation mentions in text

BACKGROUND AND PURPOSE: The collagen alpha2(I) gene (COL1A2) on chromosome 7q22.1, a positional and functional candidate for intracranial aneurysm (IA), was extensively screened for susceptibility in Japanese IA patients. METHODS: Twenty-one single nucleotide polymorphisms (SNPs) of COL1A2 were genotyped in genomic DNA from 260 IA patients (including 115 familial cases) (mean age, 59.9 years) and 293 controls (mean age, 61.6 years). Differences in allelic and genotypic frequencies between the patients and controls were evaluated with the chi(2) test. Circular dichroism spectrometry was monitored with collagen-related peptides that mimic triple-helical models of type I collagen with Ala-459 and Pro-459 to estimate the conformation and stability of alterations. RESULTS: Significant genotypic association in the dominant model was observed between an exonic SNP of COL1A2 and familial IA patients (chi(2)=11.08; df=1; P=0.00087; odds ratio=3.19; 95% CI, 2.22 to 6.50). This SNP induces Ala to Pro substitution at amino acid 459, located on a triple-helical domain. Circular dichroism spectra showed that the Pro-459 peptide had a higher thermal stability than the Ala-459 peptide. CONCLUSIONS: The variant of COL1A2 could be a genetic risk factor for IA patients with family history.

state, *location*, *gene*, *type*

rs42524

But its a typo: it is at position **549**

Can be seen in a followup article

Yoneyama et al. "**Collagen type I alpha2 (COL1A2) is the susceptible gene for intracranial aneurysms.**", Stroke, 2004.

Followup: Arnold et al. "**Collagen morphology is not associated with the Ala549Pro polymorphism of the COL1A2 gene.**", Stroke 2005.

SCAI
Fraunhofer Institute Algorithms and Scientific Computing

Conditional Random Fields for SNP mention detection

- Conditional Random Fields (CRF) are a family of probabilistic graphical models
- Machine Learning method specially suited for sequential data
- Not affected by unbalanced data
- Its an undirected model in contrast to Hidden Markov Models → dependencies allowed
- We created a training set of 207 abstracts with Variation mentions + trained a CRF
- Necessary Gene Names are detected by ProMiner
- Disambiguation: (Gene + Variation Mention) → dbSNP (rsNumbers)

Klinger, R.; Furlong, L. I.; Friedrich, C. M.; Mevissen, H. T.; Fluck, J.; Sanz, F. & Hofmann-Apitius, M. „**Identifying Gene Specific Variants in Biomedical Text**“ Journal of Bioinformatics and Computational Biology, 2007, 5(6), 1277-1296.

SCAI
Fraunhofer Institute Algorithms and Scientific Computing

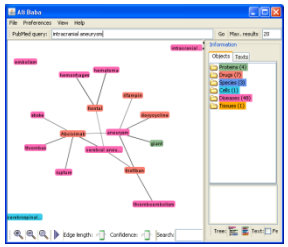
What can be done with text and extracted entities?

- **Semantic Search, sometimes called Entity Search (SS):**
  Search for documents containing Entities of selected
  Concept classes, e.g. Protein, Drug, Side Effect

- **Entity Result Aggregation and Analysis (AA):**
  Entities found in selected documents are analysed and aggregated,
  e.g. tag-cloud

- **Enrichment and Link-outs (LO):**
  Enrich the information of a text-source/snippet with additional information and refer to an external datasource.

- **Relational Networks (RN):**
  visualization of relations with network graphs. Examples are Co-occurrence networks.

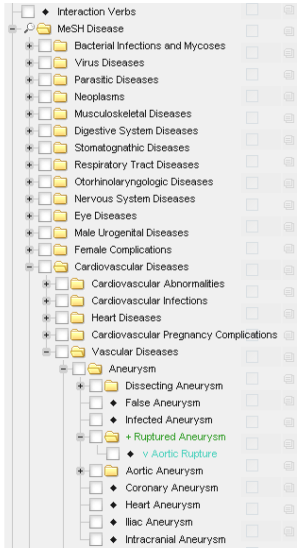


SCAI
Fraunhofer Institute Algorithms and Scientific Computing

More Technologies

- **Navigational Search:**
  Typically uses a tree-like or network based selection strategy to define the search query. Related to Semantic Search

- **Ontological Search (OS):**
  Uses Relational information defined in Ontologies/Databases for Search support.
  „**give me all documents mentioning oral contraceptives** “.
  Sometimes this is realized with tree-like interfaces.
  Real Ontological Search can reason over Ontologies.
  Is **Semantic Web** the „**Silver Bullet**“?



- **Facetted Search (FS):**
  Narrowing down the search results incrementally, with selection of known subcategories e.g. in e-Commerce

SCAI
Fraunhofer Institute Algorithms and Scientific Computing

More Technologies

- **Relevance Ranking (RR)**
  Ranked by relevancy, raw frequency is seldom working, more elaborated rankings like Relative Entropy(Kullback-Leibler Divergence), Z-Score are better.

- **Parametric Search (PS):**
  Specifying values and ranges of attributes during search, e.g. date ranges (similar to database queries)

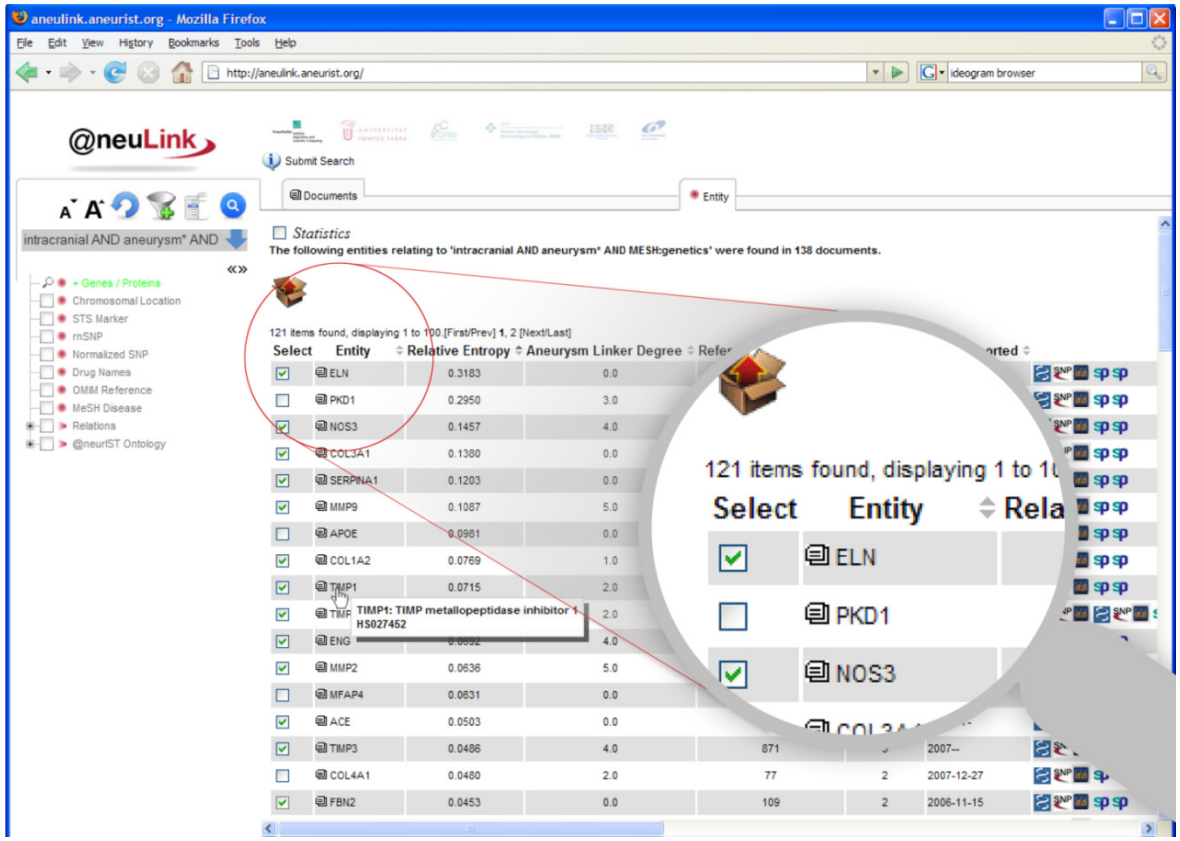
SCAI
Fraunhofer Institute Algorithms and Scientific Computing

SCAIVIEW

- **SCAIView is a part of @neuLink, a broader Knowledge Discovery suite partly developed in the @neurIST project**
- Data-Source: Medline including 19Mio documents (80GB text) + Billions of Taggings (ProMiner + Machine Learning based taggers – 40GB) + Life Science Ontologies
- **History:** for one year we tried building a Knowledge Discovery suite with a well known industrial relational database management system + text extension → **Too slow**
- Multi-threading and own MapReduce analysis
- Fulltext search with full-Medline statistics even with Millions of hits (not only restricted to newest 1000) – Query „cancer“ and mentioned Genes (210,000 docs + full analysis – in 2 seconds)
- Named entity recognition results are directly stored in the Index (waiting for the new TermAttributes in Lucene 2.9)
- Fulltext + Semantic + Ontological Search (+ simple Inference)
- Ranking via Relative Entropy (Kullback-Leibler Divergence), needs full analysis
- Performance Adjustment with Entity Confidences
- **API:** Webservice based API available for integration into other packages

Fraunhofer Institute Algorithms and Scientific Computing

SCAIView – Knowledge Environment



SCAIVIEW

Best presented

in a Live Demo

Demoserver: 4000 EUR PC, bought Jan 2008, 2*Dual Core, 8GB RAM + 24GB RAMDisk

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.

SCAI

Fraunhofer Institute Algorithms and Scientific Computing

Uptake via Webservice in the Health-e-Child project

aneulink.aneurist.org | 3DBrowser - Test

3D-BROWSER

Paediatric Brain Tumours (Astrocytomas)

maat

SCAIVIEW

Selected Levels: *Individual.Disease;1;Molecular.Chemical_ProteinGene;1*

Query (optional): astrocytoma

Tools

100 %

Restore

Level: Individual_Disease/1
Concept: (C0009326) - collagen disorder
Sem.Type: Disease or Syndrome

Molecular_Chemical_ProteinGene

gene

gene suppressor

Individual_Disease

[Documents] [@neurIST] [Tree] **Granularity**: depth 1 Build Map

Reset Selected Levels

Document: 1278 Go
Gene: COL1A2
Relevance: 498

Document: 2147 Go
Gene: F2
Relevance: 362

Document: 2006 Go
Gene: ELN
Relevance: 358

Document: 7040 Go
Gene: TGFB1
Relevance: 342

Document: 1277 Go
Gene: COL1A1
Relevance: 334

Document: 7450 Go

Acknowledgements

- Prof. Dr. Martin Hofmann-Apitius
- Dr. Juliane Fluck
- Theo Mevissen, Tobias Gattermayer, Bernd Müller, Patricia Laine, Christian Ebeling, Roman Klinger, Ye Cao
- 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)

Information Society Technologies

SCAI

Fraunhofer Institute Algorithms and Scientific Computing