

# **BioMinT: Biological Text Mining**

## **EU FP5 Quality of Life Project**

**Dr. Dipl.-Ing. Alexander K. Seewald**  
Österreichisches Forschungsinstitut  
für Artificial Intelligence



“Economic and business pressures are forcing drug companies to deploy computing, but there are still **gaps between what users want and what can be achieved.**”

*(Peter Rees - Scientific computing world - Jul/Aug 2003)*

“To be honest I don’t really understand **why you can’t buy more** [off-the shelf bioinformatics software].”

*(Jim Fickett, global director bioinformatics, AstraZeneca - Scientific Computing World, Jul/Aug 2003)*

“What might help is if the [bioinformatics] manufacturers **have the scientists’ needs in mind.**”

*(Michael Man, Pfizer - Genome Technology, Jan 2003)*

Current frontier is **biological text mining** = finding research papers, extracting topics, ranking by relevance, extracting metabolic pathways...

- Still in its infancy
- Biology is **hard** domain for general text mining
- Chronic lack of large training corpora
- "Access is a bigger problem than algorithms"

**So, we concentrate on a small user group with clear requirements and address these issues.**

## Research project funded by the EU (2003 – 2005)

- develop a tool for content-based and knowledge-intensive information retrieval and extraction
- to be applied to the annotation of the Swiss-Prot and PRINTS proteomics databases with information mined from scientific papers
- adapted to the needs of biological researchers in general

= **In-silico research assistant for curators**

[www.biomint.org](http://www.biomint.org)

- **University of Manchester (U.K)**, School of biological sciences
  - Prints and Precis providers
- **Swiss Institute of Bioinformatics**
  - SwissProt providers and users
- **University of Antwerp (Belgium)**
  - Language technology providers
- **Österreichisches Forschungsinstitut für AI (ÖFAI, Austria)**
  - Information extraction/retrieval providers
- **University of Geneva (Swiss)**
  - Information extraction/retrieval providers
- **PharmaDM (Belgium)**
  - Relational data mining technology, architecture

## General workflow:

- User enters protein or gene name
- Look up name in integrated synonym database
- Generate and execute PubMed query
- Retrieve documents, filter and rank by relevance.

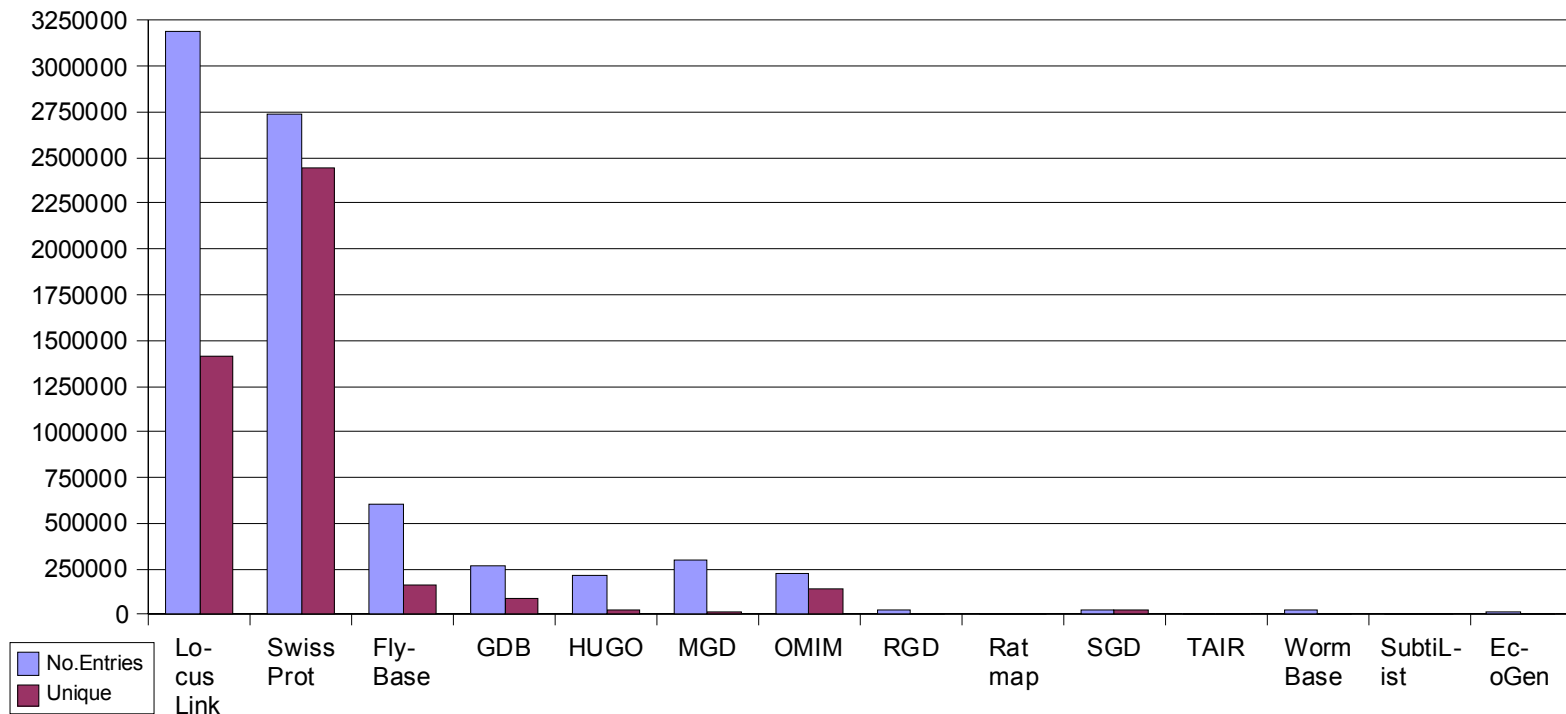
Implementation as Web form, available to our partners.

Download all 14 databases according to SIB (+ SwissProt)

Extract all relevant fields from each DB separately

Create all pairs of synonyms (noting Source DB, field, ID)

**7,652,510 pairs of synonyms; 737,040 unique names**



## Positive-only comparison allows to recognize...

- Competitive perf. of KeX & Yapex w/ sloppy comparison
- Overlong matches of KeX

<u>All DEs</u>	Yapex	KeX	GAPSCORE
Strict	<b>0.202±0.401</b>	0.097±0.296	0.192±0.394
PNP	0.606±0.423	0.529±0.374	<b>0.629±0.414</b>
Sloppy	0.732±0.443	<b>0.775±0.420</b>	0.761±0.427

## Recent work

- Competitive perf. of GAPSCORE vs. Yapex
- Ensemble of all approaches improves on best single system

## Learning approaches on top 20 species

- 75.5% Human domain expert
- 79.6% Mapping MeSH Terms to species
- 88.9% JRip Rule Learner, 172 rules
- 89.3% support vector machine (SMO)

## Conclusion

- Domain experts are good at creating precise rules, but bad at managing trade-off
- JRip is good at managing trade-off, but yields worse precision offset by better recall.

## TextPresso: Question answering

- Small domain with simple nomenclature (C. elegans)
- Corpus of 2,700 full-text papers and 16,000 abstracts
- Open-Source, freely available search: [www.textpresso.org](http://www.textpresso.org)

## QUOSA: Query, Organize, Share, Analyze

- Commercial product, launched late 2002
- Establishes local paper collection by downloading
- Prioritizes full-text papers during search
- Available to hundreds of researchers in two US hospitals

- Generating better PubMed queries
- Filtering and Ranking documents
- User-interface improvements
- Bootstrap human-generated corpora