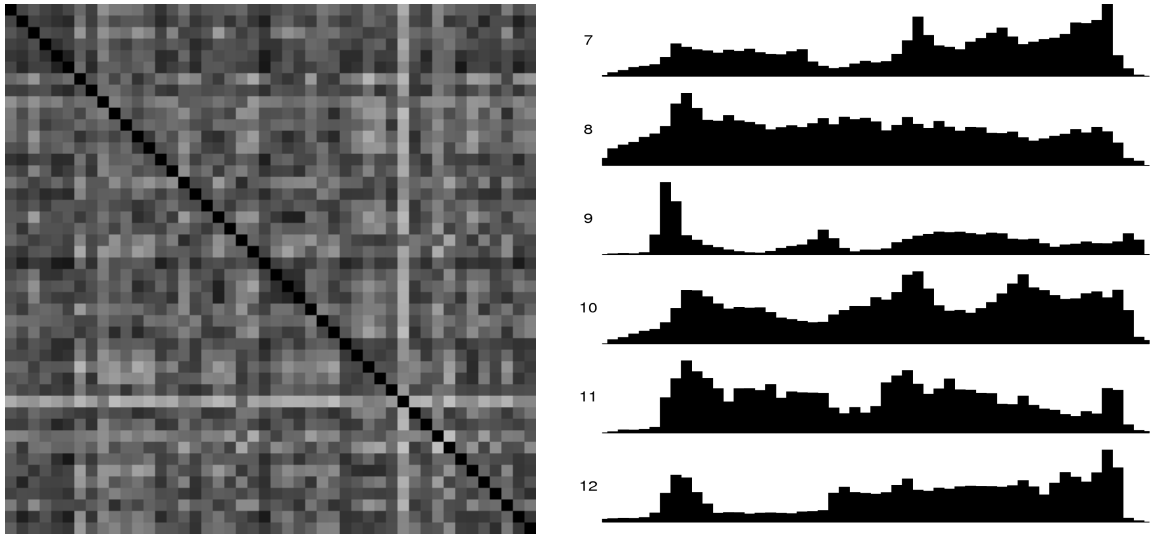
Success Stories in Intelligent Biomedical Data Analysis



Dr.techn. Alexander K. Seewald

Seewald Solutions

alex@seewald.at / alex.seewald.at

Intelligent Biomedical Data Analysis

Involves **Biological & Medical Data**

- Images (2/3D); Text (Research papers, lab notes, DNA); Multi-dimensional „classical“ data (EEG, clinical trials, patient records…)

Involves **Data Analysis**

- Applying statistical tests; parameter fitting to full or partial known models; exploratory data analysis (hypothesis-generating) etc..

Involves **Intelligence**

- Extensive computer use; Analytics which would otherwise be impossible or very expensive.

alex@seewald.at / alex.seewald.at

Intelligent Biomedical Data Analysis

Why is this a challenging field? Because the results of biomedical research are usually not available in computer-readable form!

- Biochemical pathway reaction constants and relations are not present in a formal language.
- Main communication between researchers is in the form of research papers (text), which can only be rudimentarily understood by computers
- Most online databases are designed to be human-readable rather than computer-readable.
- Image features easily discernable to a human observer are quite hard to teach computers.

...

alex@seewald.at / alex.seewald.at

BioMinT: Biological Text Mining

Three-year research project funded by the EU

- Develop a generic text mining tool for content-based and knowledge-intensive information retrieval and extraction
- Mining Information from scientific papers.
- Adapted to needs of biological researchers in general and specifically for annotating the Swiss-Prot and PRINTS proteomics databases.

Metaphor: In-silico research / curator assistant

biomint.pharmadm.com

alex@seewald.at / alex.seewald.at

The BioMinT Tool

General workflow

1. User enters protein / gene name
2. Name is looked up in comprehensive Gene and Protein Synonym Database (GPSDB). Selection criteria: species, taxonomic range, source database and source field.

   This expands Name with (almost) all known synonyms.
3. Generate & execute PubMed query with all synonyms.
4. Retrieve references, filter and rank by relevance.
5. Extract information for annotation purposes (PRINTS,SP)

alex@seewald.at / alex.seewald.at

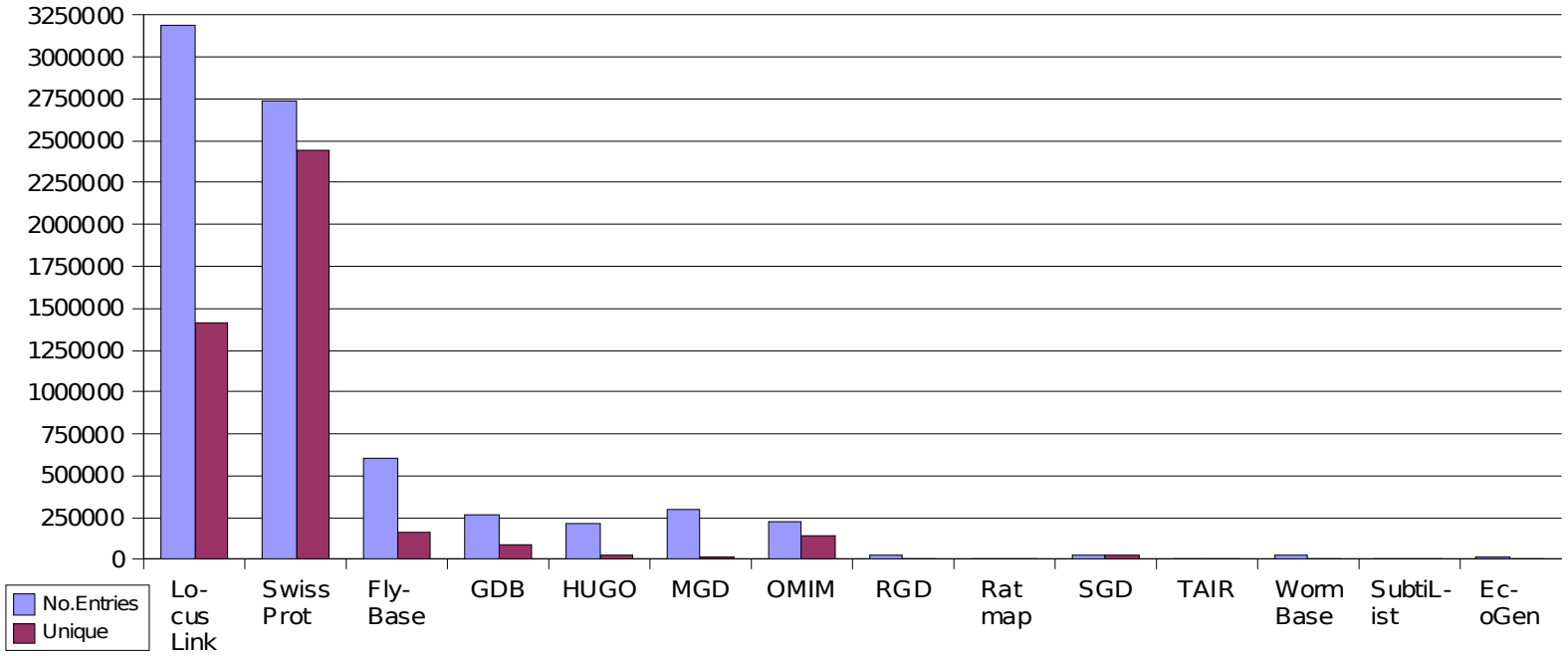
GPSDB

Download all 14 databases according to SIB

Extract all relevant fields & links from each DB

Create all synonym pairs (noting Source DB, field, ID)

Used for synonym expansion, homonym removal, named entity recognition, link network analysis..



alex@seewald.at / alex.seewald.at

Ranking

- Based on *medical annotation dataset* by Swiss Institute for BioInformatics. Initial experiments indicated that word *missense* significantly correlates with relevance.

Ranking via **Lucene** (query term extended via Gene/Protein Synonym DB plus *missense,* filtered by species)

$$score_d = coord_{qd} \sum_t tf_q \frac{idf_t}{norm_q} tf_d \frac{idf_t}{norm_{dt}} boost_t \quad (1)$$

where

$$score_d = \text{score for document d} \quad (2)$$

$$coord_{qd} = \text{number of terms in both query and document divided by number of terms in query} \quad (3)$$

$$tf_q = \text{the square root of the frequency of t in the query} \quad (4)$$

$$idf_t = \log \frac{numDocs}{docFreq_t + 1} + 1.0 \quad (5)$$

$$numDocs = \text{number of documents in index} \quad (6)$$

$$docFreq_t = \text{number of documents containing t} \quad (7)$$

$$norm_q = \sqrt{\sum_t (tf_q idf_t)^2} \quad (8)$$

$$tf_d = \text{the square root of the frequency of t in d} \quad (9)$$

$$norm_{dt} = \text{sqrt number of tokens in d and same field as t} \quad (10)$$

$$boost_t = \text{the user-specified boost for term t} \quad (11)$$

alex@seewald.at / alex.seewald.at

Homonymy Recognition

Synonym Group = A group of database entries connected by inter-database links, all dealing with same gene/protein entity.

Homonym = Name appears in several *Syn.Grp*

Each of ten queries was expanded with all synonyms, and then checked for homonyms. All found homonyms were verified by domain experts: *Accuracy*=100%.

However, homonyms have little impact on ranking performance.

Query	Homonyms
vhl	HRCA1,RCA1
xpc	p125
wrn	RECQL2,RECQL3
tulp1	RP14
wt1	WAGR

alex@seewald.at / alex.seewald.at

Species Domain Classification

Predict the domain of an organism from MEDLINE

- Simple: (*Bacteria*) ⇒domain=B, ⇒domain=E (85%)

Better Rules (97%)

⇒

(archaeon) => domain=A (163.0/0.0)
(Archaeal) and (!Bacterial) => domain=A (92.0/0.0)
(Halobacterium) => domain=A (22.0/3.0)
(archaebacterium) and (Bacterial) => domain=A (7.0/0.0)
(Methanobacterium) => domain=A (6.0/2.0)
(Archaea) and (!Proteins) => domain=A (2.0/0.0)
(Viral) => domain=V (351.0/18.0)
(Bacterial) and (!Animal) => domain=B (1665.0/14.0)
(Bacterial) and (!RNA) and (!cerevisiae) => domain=B (211.0/10.0)
(!Animal) and (Escherichia) and (!Proteins) and (!Fungal) and (!cDNA) => domain=B (26.0/2.0)
(!Animal) and (bacteria) and (!cDNA) => domain=B (19.0/3.0)
(strain) and (!Fungal) and (!Proteins) and (!2) => domain=B (17.0/1.0)
(!Animal) and (cyanobacterium) => domain=B (9.0/1.0)
(Bacteria) and (!Animal) => domain=B (6.0/1.0)
(Frames) and (operon) => domain=B (4.0/1.0)
(Salmonella) => domain=B (2.0/0.0)
(Streptomyces) and (!at) => domain=B (5.0/0.0)
(Anabaena) => domain=B (3.0/0.0)
(bacterium) => domain=B (5.0/1.0)
(Bacillus) and (!Animal) => domain=B (5.0/1.0)
(pneumoniae) => domain=B (2.0/0.0)
=> domain=E (2534.0/20.0)

alex@seewald.at / alex.seewald.at

Species Classification (20 most frequent)

Predict exact species of an organism from MEDLINE

- 19.0% Baseline (most common class *Human*)
- **75.5% Human domain expert's rules**
- 76.4% NaiveBayes
- 79.6% Mapping MeSH Terms to species manually
- 88.9% JRip Rule Learner, 172 rules
- **89.3% support vector machine (SMO, Weka)**

Comparing JRip rules to domain expert rules

- Expert: + precision, – recall; — F-Measure
- JRip: – precision, + recall; ++ F-measure

alex@seewald.at / alex.seewald.at

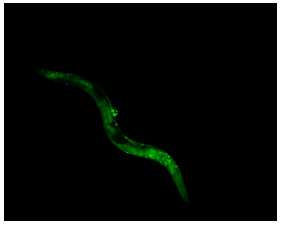
Watching C. elegans Think

Basic research project in Systems Neuroscience

Four Objectives

• Engineering	*Real-time tracking nerve cells*
• Methodological	*Validate nervous cell models*
• Holistic	*Understand complete N.S.*
• Insight	*Better learning algorithms*

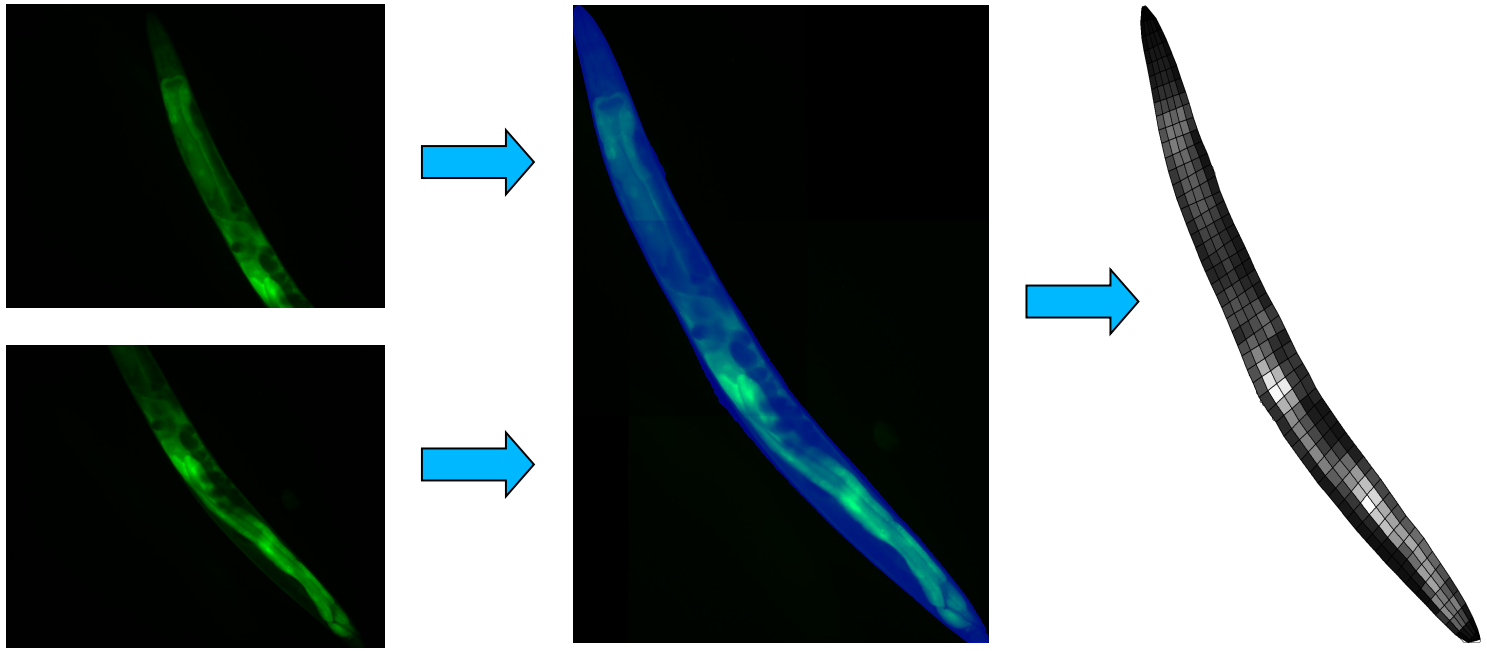
Model organism: C. elegans
~ 1000 cells, ~ 300 nerve cells
Might be feasible to simulate



alex@seewald.at / alex.seewald.at

Biological Image Mining

Results of an automated analysis of C.elegans images (data by Prof. T. Johnson's group)



Reduces workload by 80% for tagging worm images

Development time about 1PM, ongoing collaboration

alex@seewald.at / alex.seewald.at

Biological Image Mining (2)

Techniques used...

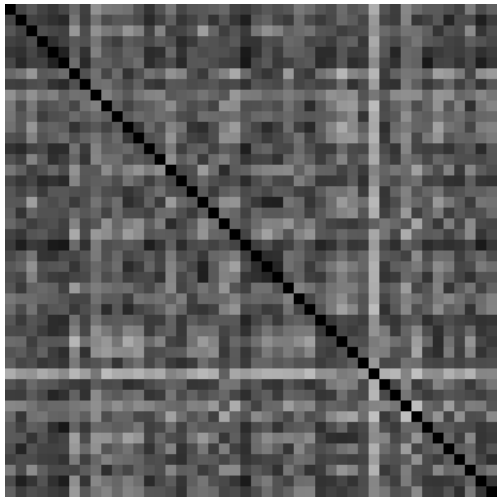
- Image correlation coefficient for combining head (a) and tail (b) images; two-step hierarchical search.
- Pixel classification via mean and standard deviation of green channel in 5x5 window around each pixel, after histogram equalization.
- Threshold optimization by testing minimum circularity and area of largest blob – significantly improves results!
- Closure (erode, dilate)
- Fill internal holes with circularity below threshold
- Heuristic search for breaks in contour, which are repaired with straight lines and filled on the inside.
- Final worm: Segmentation into 50 slices and 5 dices

alex@seewald.at / alex.seewald.at

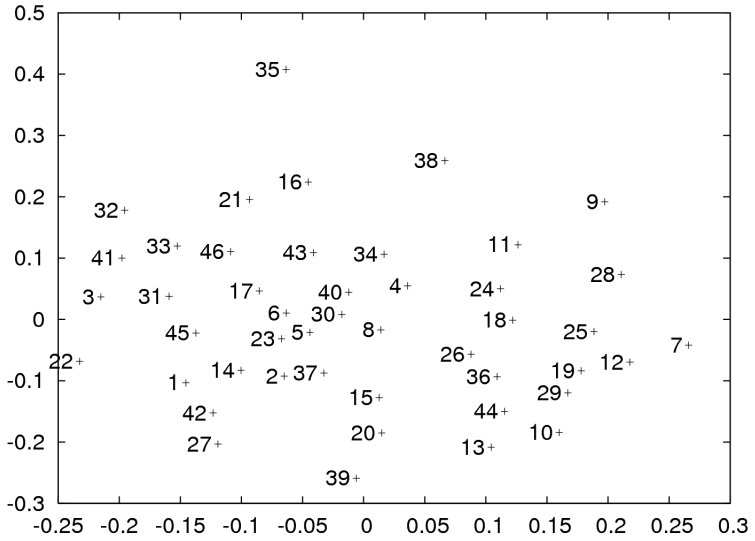
Preliminary Results

2D GFP activity patterns

- Distance matrix



- 2D visualization via Sammon mapping

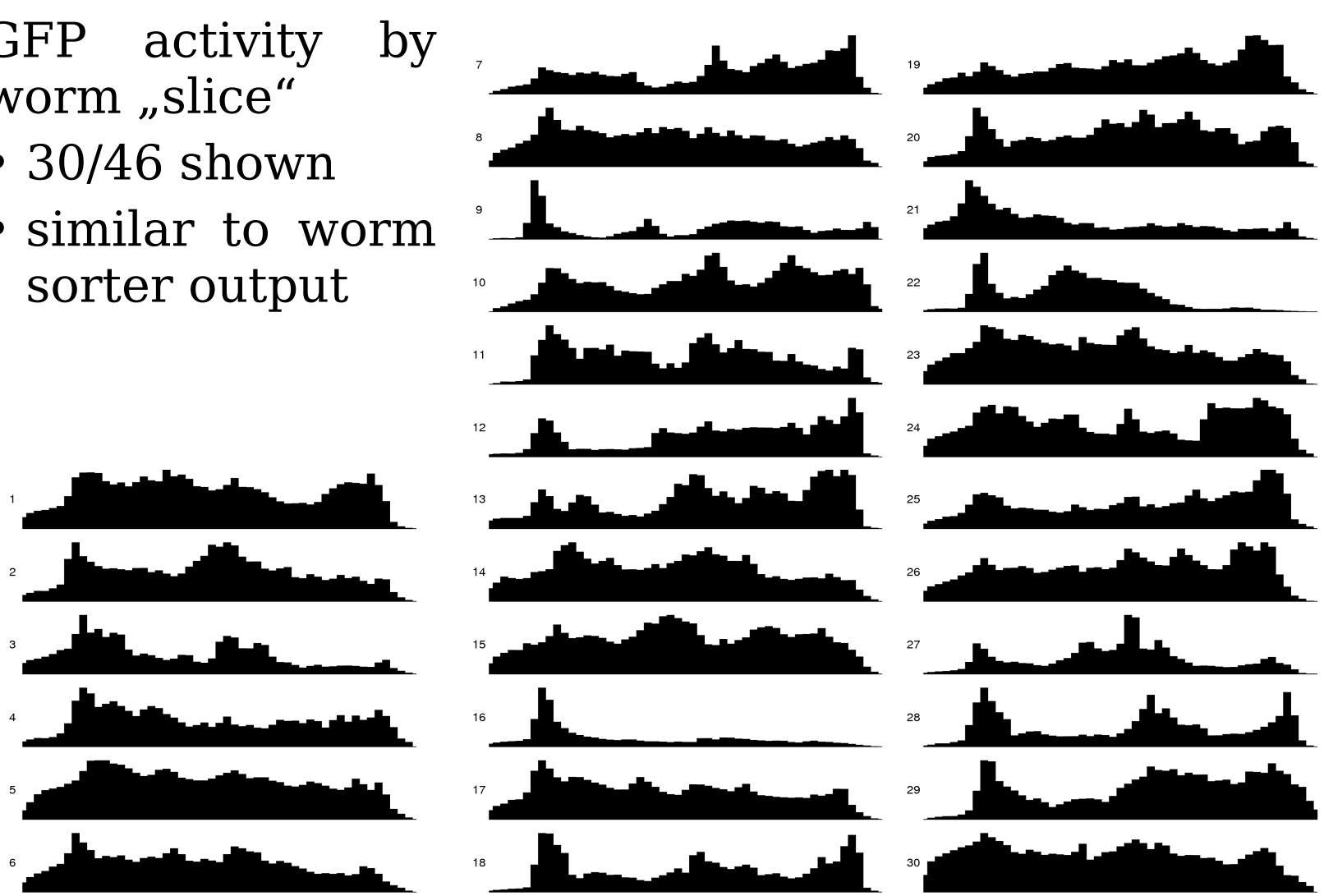


alex@seewald.at / alex.seewald.at

Preliminary Results (2)

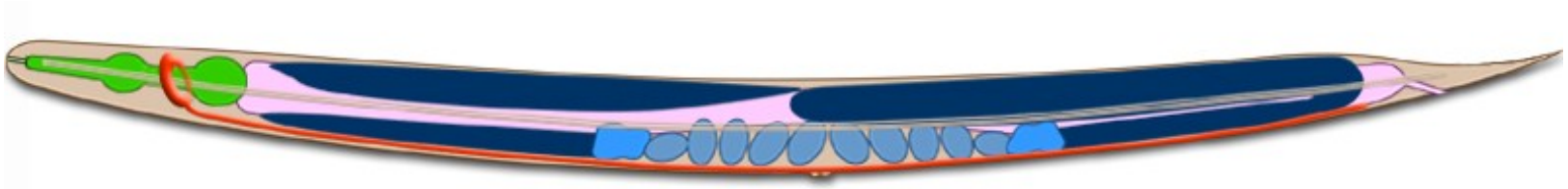
GFP activity by worm „slice“

- 30/46 shown
- similar to worm sorter output



alex@seewald.at / alex.seewald.at

Future Work



Integrate anatomical data via Slideable Worm (WormAtlas.org, Prof. Z. Altun) to trace GFP activity to single cell or small cell groups.

Pros

- 800 worm slices of C. elegans, partially tagged
- Unique anatomical data available nowhere else

Cons

- Only ~30 slices actually fully tagged
- Tagging is not on level of single cell, but cell types
- Need to determine 3D worm for best results

alex@seewald.at / alex.seewald.at

Conclusion

Successful projects...

- Include computer scientists as equal partners to biologists and medical doctors.
- Make an effort to create high-quality computer-readable data initially, and later for validation.
- Use resources that are already available whenever possible.
- Use Open Source software whenever possible.

alex@seewald.at / alex.seewald.at

Past Projects

2000-2005	Employed at OFAI as junior researcher
2001	EEG data analysis (contributed by Brain Research institute, Vienna)
2000-2002	*A New Modular Architecture for Data Mining* (FWF)
2002	*3DSearch* (multi-document summarization, EU & uma AG)
2002-2003	*Intelligent Go Board* (embedded device to capture moves of Japanese Go during play, presented at Innovation Workship in '05)
2003-2005	*BioMinT* (integrated system for biological text mining, EU FP5)
2004-2006	*SA Train* (Spam training methodology for SpamAssassin, Evaluation of commercial and OSS spam filter systems)
2005	*Digits* (handwritten digit recognition: open source corpus and preliminary experiments)
2006	Employed at GE Money Bank as CRM Analyst
2006-2007	*IGO-2* (image mining on images of Go final board states)
2007	*THOUCENS* (image mining on GFP/DIC images contributed by Univ. of Colorado at Boulder & Univ. f. Bodenkultur Wien)
Since 2007	Independent researcher (50% research, 50% commercial)

alex@seewald.at / alex.seewald.at