

Hôpitaux Universitaires de Genève
Division d'Informatique Médicale

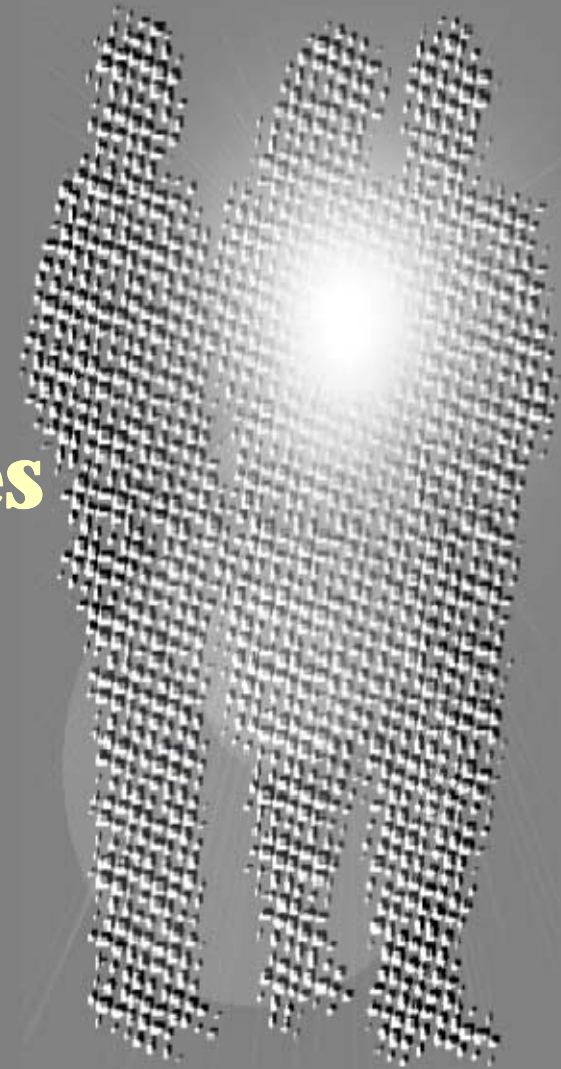
Abstracting natural language knowledge representation schema for biomedical sciences

Robert H Baud, PhD

**Synergy between Research in Medical Informatics, Bio-informatics
and Neuro-Informatics:**

Knowledge Empowering Individualised Healthcare and Well-Being

Brussels, 14 December 2001



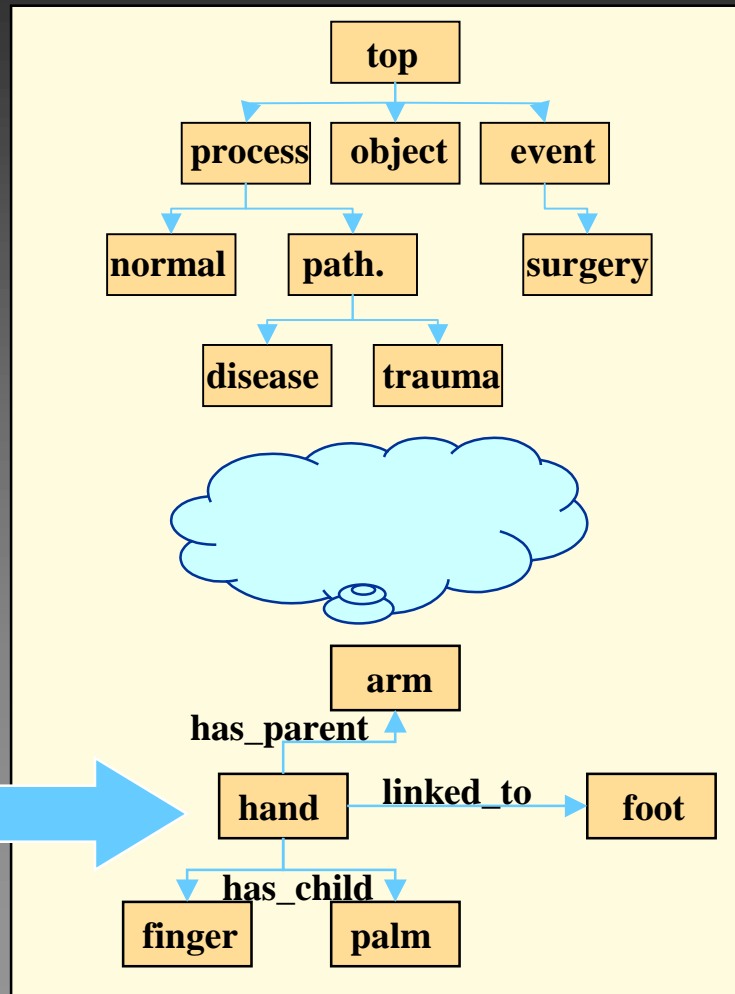
Summary

- ♣ In order to bridge the gap between genomic information and clinical relevance and application, we propose the recasting of NLP tools in new domains
- ♣ Building the necessary models requires the synergy of domain experts and NLP experts
- ♣ Awaited benefits are: data mining, automated classification and terminology developments.

Target

- ♣ To search for gene-gene interactions, protein-protein interactions, etc
- ♣ To transfer knowledge from unstructured data – the literature – to a structured form to be included in a knowledge base
- ♣ To look for co-occurring concepts with a certain proximity in the text
- ♣ To qualify the links between co-occurring items

Modeling a domain



Hyperonyms

Local proximity

Tagging

- ♣ A hyperonym is a high level label grouping concepts into broad classes. Also called a tag.
- ♣ The UMLS semantic net can be considered as a set of tags for the medical domain (132 classes).
- ♣ Tagging techniques are asking for smaller tagset, with an extension of 40 tags. Grouping classes of the semantic net achieves this goal (see Medtag).
- ♣ Categorisation by tags is the basic instrument for disambiguation of texts.

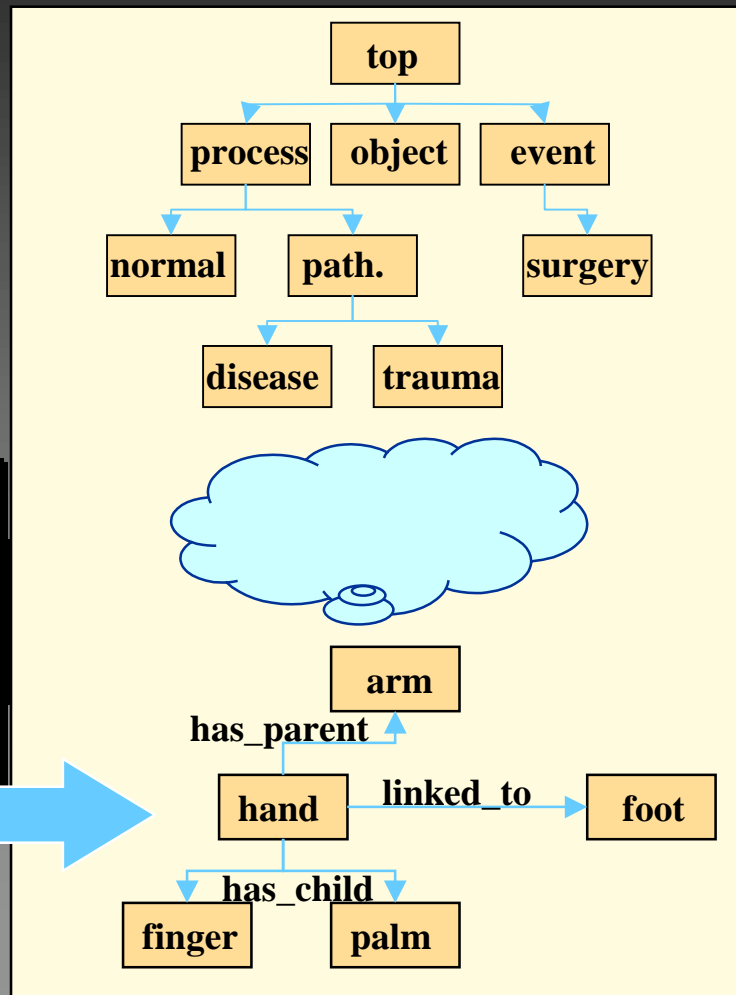
MedTag: a tagset for clinical texts

- ♣ abn: abnormality
- ♣ loc: body location
- ♣ bosu: body substance
- ♣ find: sign or symptom
- ♣ mental: mind or cognitive process
- ♣ live: live organism
- ♣ ther: therapeutic procedure
- ♣ tiss: cell or body tissues
- ♣ . . .

Modeling local proximities

- ♣ A local model is just modeling a chunk of knowledge without direct relations with other local models
- ♣ A local model is built with the purpose of resolving simple local inferences
- ♣ A local model is simple to build and easy to manage. It may be immediately available
- ♣ However, a local model is certainly not able to bring general inference capabilities. It should be replaced at some point in time by a more formal ontological solution.

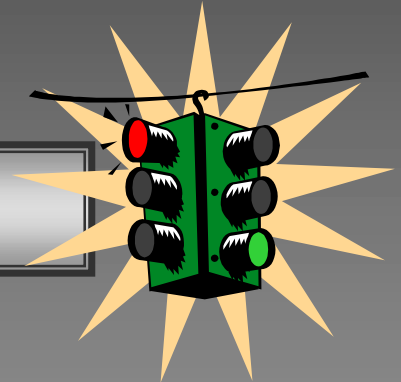
Modeling a domain



Hyperonyms

Ontology

Local proximity



Multiple steps for modelling

- ♣ Tokenisation of the input text
- ♣ Disambiguation
- ♣ Shallow parsing
- ♣ Semantic tagging
- ♣ Syntax-driven rule:
IF `papr-(hasNCompl)-loc`
THEN `[]-(hasLoc)-[]`
- ♣ `Haemorrhage/noun;`
`of/prep|art;encephale/noun`
- ♣ Remove art for « of »
- ♣ `[Haemorrhage]-`
`(hasNCompl)-[Encephale]`
- ♣ `Haemorrhage/papr`
`Encephale/loc`
- ♣ `[Haemorrhage]-(hasLoc)-`
`[Encephale]`

Coping with similar terms

♣ Examples:

- ♣ Spondylodiny, chronic
- ♣ Permanent vertebral pain
- ♣ Continuous vertebralgies

♣ Numerous wordings are equivalent

♣ No hope to « educate » the physicians; voice recognition systems will not improve the situation!

♣ Working with syntax and morphology at one hand, with semantic proximities at the other hand.

Example: spondylodiny, chronic

♣ Tokenisation:

♣ spondyl:cl_Vertebral; odiny:cl_Pain; chronic:cl_Chronicity

♣ Shallow parsing:

♣ [odiny:cl_Pain]
 (hasPrefix) [spondyl:cl_Vertebral]
 (hasAdjective) [chronic:cl_Chronicity]

♣ Tagging:

♣ [find] (hasPrefix) [loc] and [find] (hasAdjective) [temp]

♣ Validation rules:

♣ If [find] (hasPrefix) [loc] then [find] (hasLocation) [loc]

♣ Result:

♣ cl_Pain]
 (hasLocation) [cl_Vertebral]
 (hasTempAttr) [cl_Chronicity]

Example: permanent vertebral pain

♣ Tokenisation:

♣ permanent:cl_Chronicity; vertebral:cl_Vertebral; pain:cl_Pain

♣ Shallow parsing:

♣ [pain:cl_Pain]

(hasAdjective) [vertebral:cl_Vertebral]

(hasAdjective) [permanent:cl_Chronicity]

♣ Tagging:

♣ [find] (hasAdjective) [loc] and [find] (hasAdjective) [temp]

♣ Validation rules:

♣ If [find] (hasAdjective) [loc] then [find] (hasLocation) [loc]

♣ Result:

♣ cl_Pain]

(hasLocation) [cl_Vertebral]

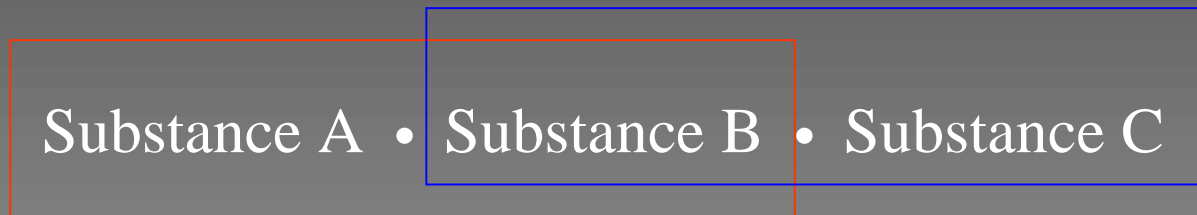
(hasTempAttr) [cl_Chronicity]

GeneWays System

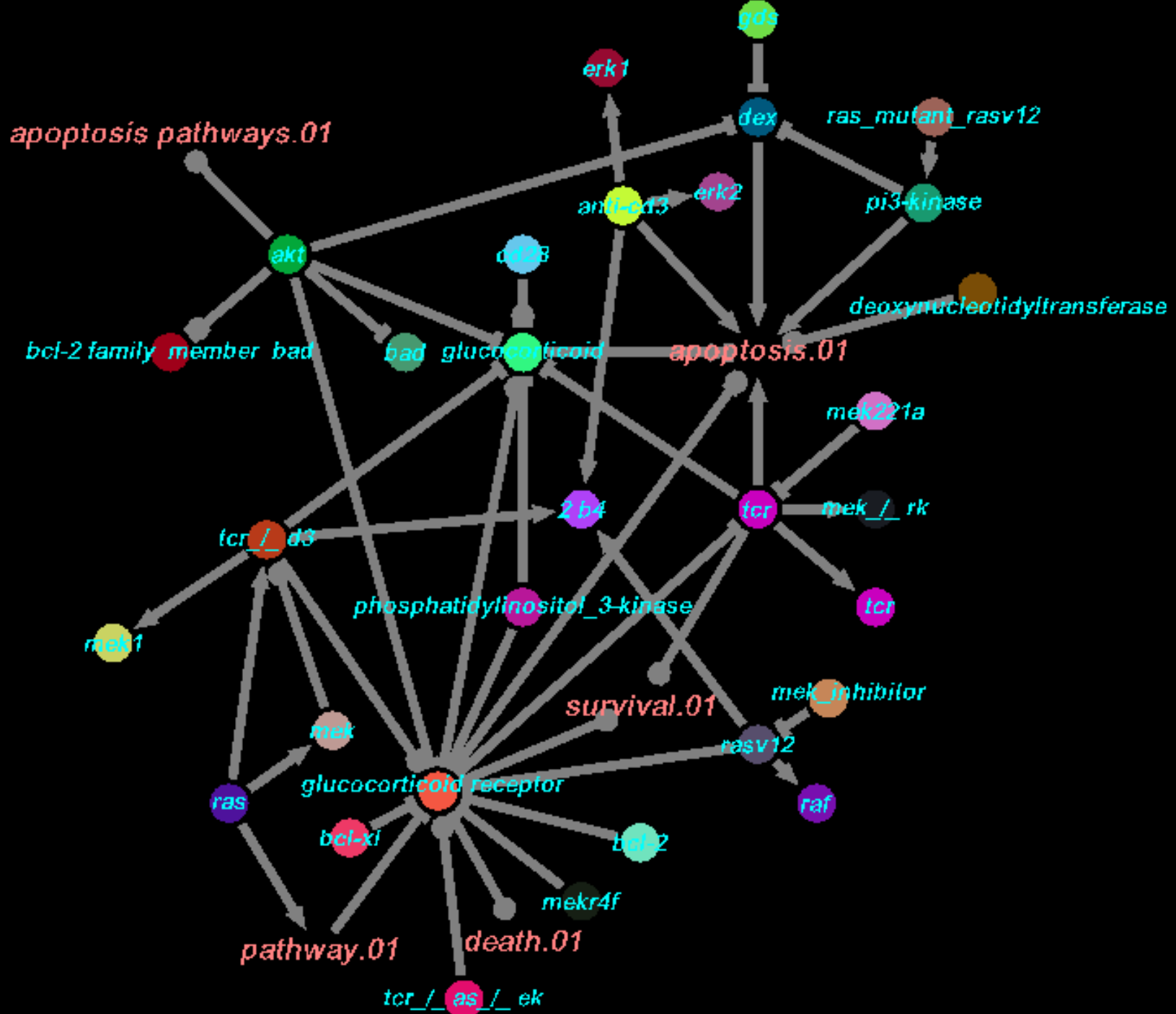
- ❖ Automatic extraction, analysis, and visualization of molecular pathway data from the research literature
- ❖ Extraction: literature gathered from the WWW
- ❖ Analysis: NLP and statistical analysis for determining the molecular interactions defined in the literature
- ❖ Visualization: Graphical interface to allow researchers to better understand interactions and explore literature resources

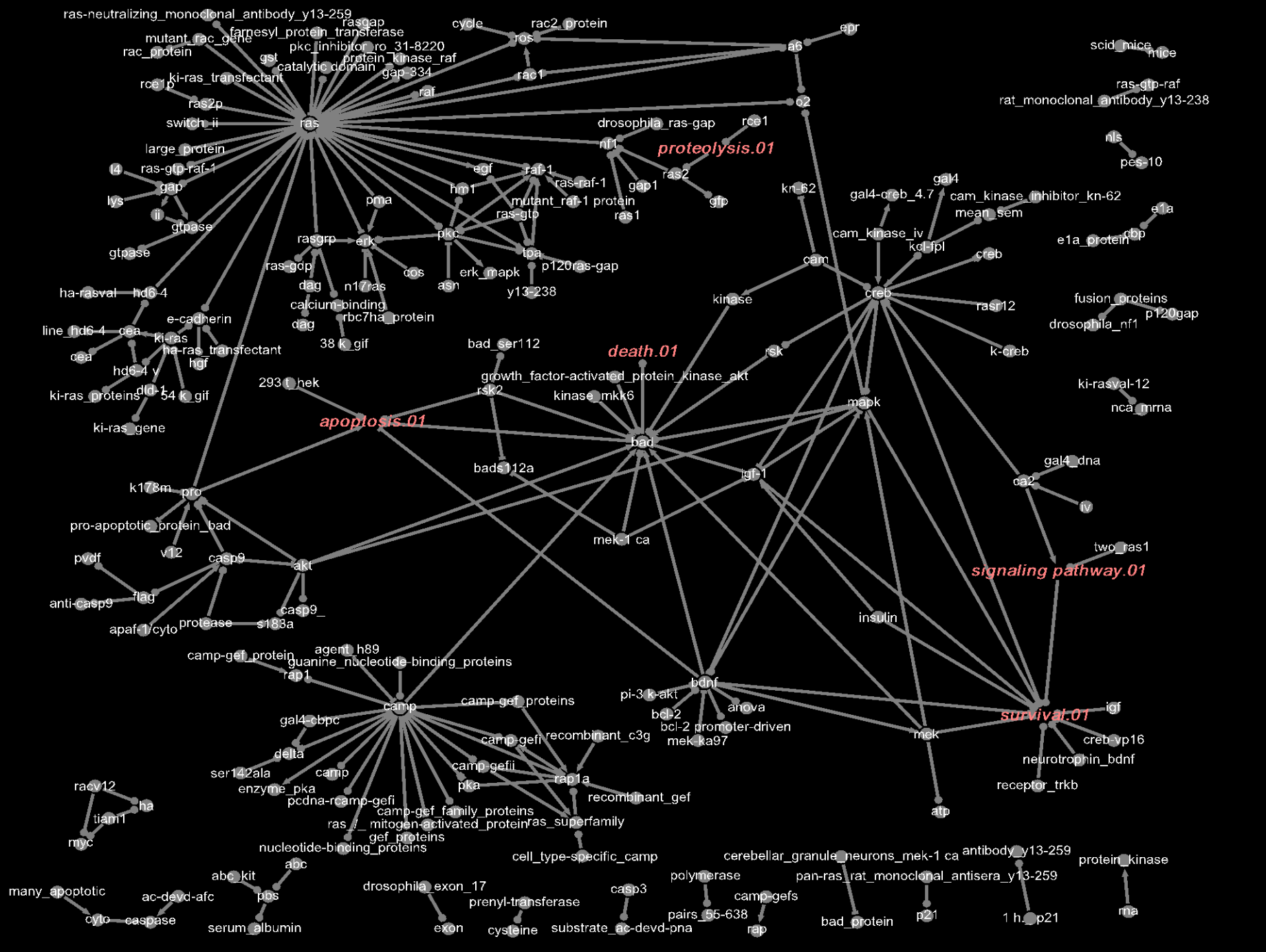
Basic work: find concepts and links

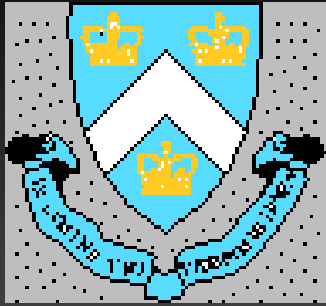
- “**rap1** functions as a negative regulator of **Tcr**-mediated **il-2 gene** transcription”



- **Rap1** inhibits **Tcr**
- **Tcr** mediates **il-2**







The GeneWays Project

Department of Medical Informatics, Columbia University

Carol Friedman, Pauline Kra, Michael Krauthammer, Hong Yu, Andrey Rzhetsky

Department of Computer Science, Columbia University

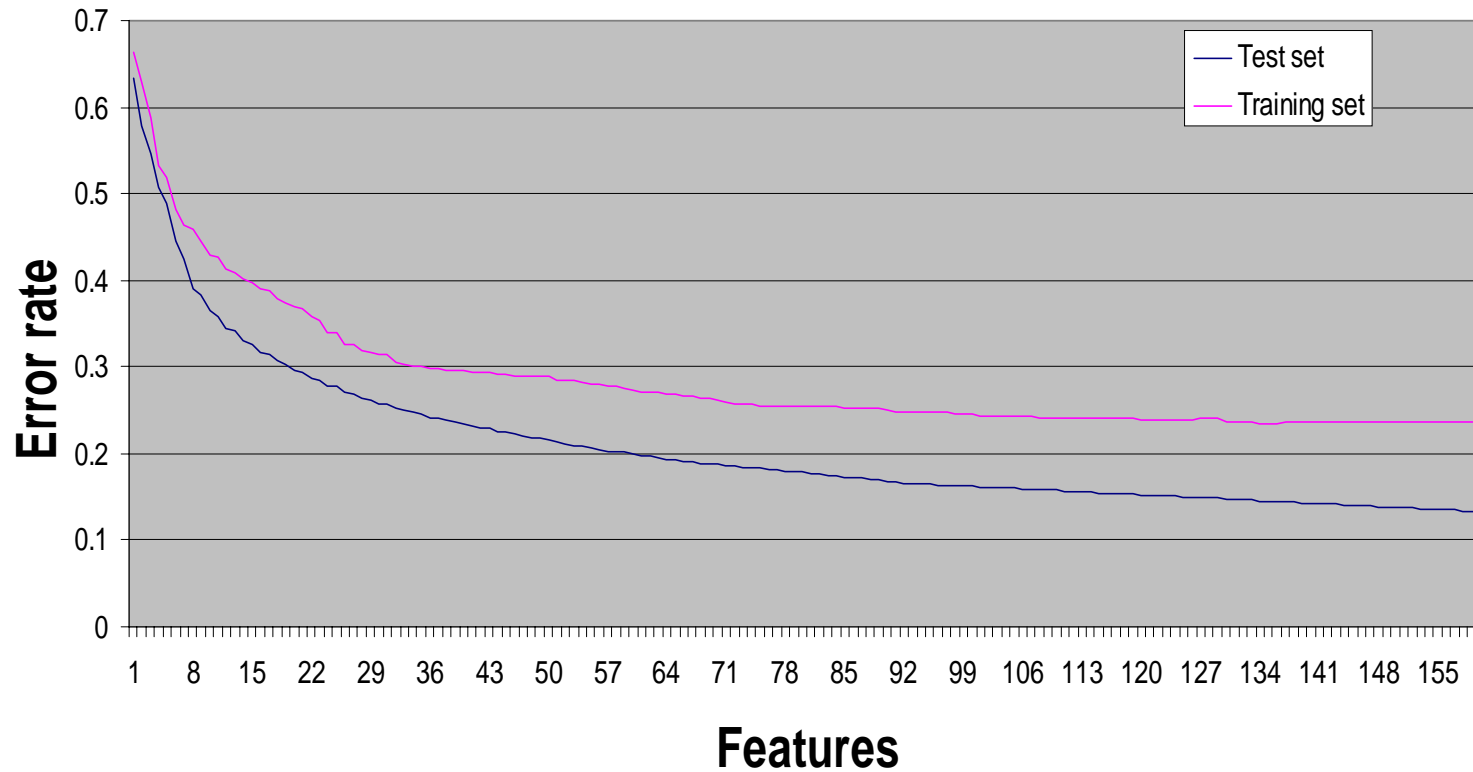
Vasileios Hatzivassiloglou, Pablo Ariel Duboue, Wubin Weng

Columbia Genome Center, Columbia University

Pavel Morozov, Tomohiro Koike, Shawn Gomez, Sabina Kaplan, Sergey Kalachikov, Jim Russo, Andrey Rzhetsky

Topic filtering: typical results

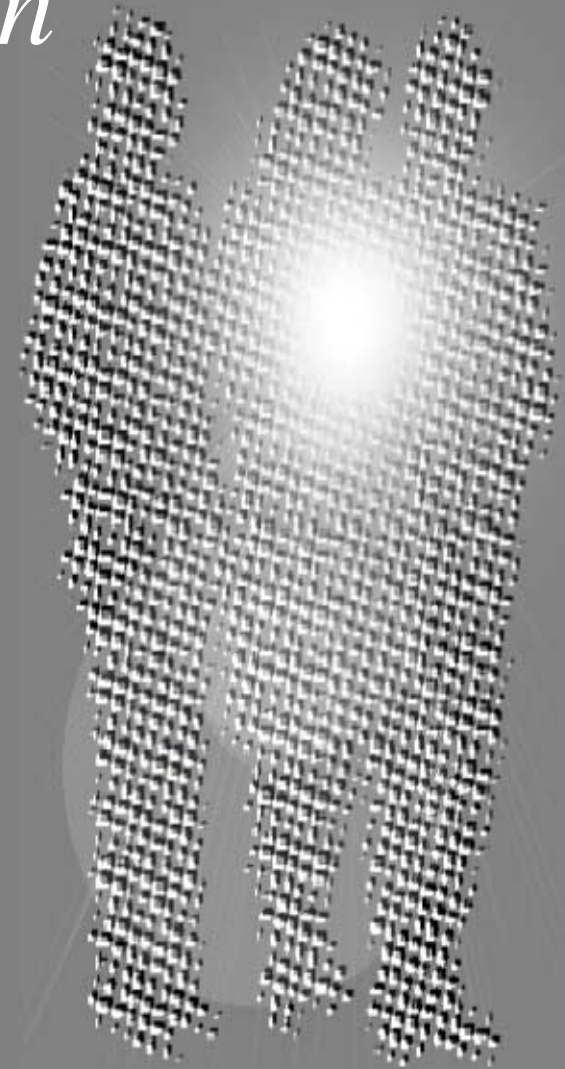
Extraction of conclusion sentences in genomic abstracts



Conclusion

- ♣ Knowledge extraction schema from the clinical domain may be recasted for genomics
- ♣ A priori success rate is 75 %
- ♣ Patterns of knowledge may be represented in networks of concepts and relations
- ♣ Proximity between documents may be computed and clusters may be automatically set up from the literature
- ♣ It is time for experts from the different domains to meet.

Thank you for your attention



Author email:
Robert.Baud@dim.hcuge.ch