projects

Table of contents

1. Projects and Research Groups

NOTE: This list is under development and will be expanded in due course. If you (or your group) would like to be added to this list, please send us a message

2. People

- Jun-ichi Tsujii
- Nigel Collier
- Patrick Ruch
- Mark Craven
- Bob Futrelle
- Kristofer Franzen
- Rolf Apweiler head of SwissProt

3. Groups

- BioMinT is a EU-funded Research Project (2003-2005)
    - Aim: developing tools for content-based and knowledge-intensive information retrieval and extraction.
    - Applications: annotation of Swiss-Prot and PRINTS proteomics databases
    - Methods:

1. IR: Query expansion + Ranking
    1. query is protein or gene name
    2. expand it using synonym database (using 14 different databases)
    3. generate and execute PubMed query
    4. retrieve documents, filter and rank by relevance
2. Named Entity Recognition (recognition of Biological Entities), and IE
    1. evaluation of external tools: Yapex, KeX, GAPSCORE
    2. learning approaches for species classification
    3. plan to train a generic shallow parser over GENIA
3. providing results as database slot fillers
    - publications
        - good 'marketing' style presentation Focus in particular on pages 27-41, which regard requirements for Text Mining
        - Ranking for BioMinT , document retrieval and ranking, not particularly relevant to us
        - Bio Entity Recognition learn extraction patterns for classification of organisms

    - Evaluating Protein Name Recognition a comparison of two Protein Name Taggers: KeX and Yapex
    - Classifying Protein Fingerprints This appears to be a purely Data Mining approach to classify Bio Data.
    - BioMint short presentation by K. SeeWald
- BioNLP.org (Futrelle's page)
- Alex Morgan's HomePage with BioNLP resources
- Protein Annotation Tools (AnnBlast)
- Mining the Bibliome: Information Extraction from the Biomedical Literature (UPenn)
- Natural Language Processing and Computational Linguistics (Brandeis University)
- Semantic representation of biomedical text (Lister Hill Center, National Library of Medicine)
- Biomedical informatics (University of Sheffield)
- Helix (Inria - France)
- BioPath Project (University of Salford)
- Genia Project (Tsujii laboratory - University of Tokyo)
- MedSyndikate Project (University of Freiburg)
- Bioinformatics (University of Arizona)
- Text Mining Group (Protein Design Group, NCB - Spain)
- Language Technology Group (University of Edinburg) - Disp Project
- BioText Project (University of California)
- Georgetown University
- BioNLP People (Kevin Cohen's page)
- Textomy The PreBind/Textomy system was presented in (donaldson:bioinformatics03). Similar to BioMinT. Meant as a database curation aid for the **BIND** database of protein-protein interactions. It contains IR, IE, Domain Knowledge. It uses SVM for filtering relevant documents (for protein-protein interactions). Same SVM is used to find relevant sentences. No deep linguistics. Lists of protein names and synonyms are derived from public databases, and are used as domain knowledge. Morphological and contextual rules are used to find candidate interacting proteins. Follows a step of human validation. They test their results against **MIPS** (an indipendent interaction database)
- Yapex
- A*STAR
- TextPresso
    - IR, IE and QA
    - interface base on simple IR querys, **or** category based interface
        - works on text that has been pre-annotated (how?)

  - IE planned, not yet available
  - not using learning (markup done manually?)
  - one simple domain (C. elegans)
  - Corpus of 2700 papers and 16000 abstracts
  - open-source, freely available
- PASTA Result of an EPSRC project (1998-2001) Described recently in Bioinformatics
  - IE system (MUC style)
  - focusing on the role of amino acids residues in protein active sites
  - tokenizaton, POS tagging, NE recognition, parsing, discourse interpretation, template extraction, templates are then used to fill a Relat DB