

# Semi-automatic Construction of a Taxonomy for a Biological Field



Oriane MATTE-TAILLIEZ<sup>1</sup>, Mathieu ROCHE<sup>2</sup>, Yves KODRATOFF<sup>2</sup>, Michel TERMIER<sup>1</sup>

<sup>1</sup> Équipe Bioinformatique des Génomes, Institut de Génétique et Microbiologie - <sup>2</sup> Équipe Inférence et Apprentissage du Laboratoire de Recherche en Informatique - Université Paris-Sud XI - Orsay - FRANCE

E-mail: {Oriane.Matte, Mathieu.Roche, Yves.Kodratoff}@lri.fr; Michel.Termier@igmors.u-psud.fr

The amount of published data shows such a fast increase that it cannot be anymore managed manually. This is true for many topics and in particular for the data resulting from genomic, transcriptomic and proteomic methods in the field of yeast molecular biology. We apply to this field a semi-automatic method for collecting information from texts.

## First step : Cleaning

- Homogenization of 6000 abstracts.

## Second step : Terminology

- Grammatical tag : Brill's tagger (Brill, 1994) for the field.
- Extraction of the more relevant terms in the field : Jacquemin's association measure with addition of specific heuristics.

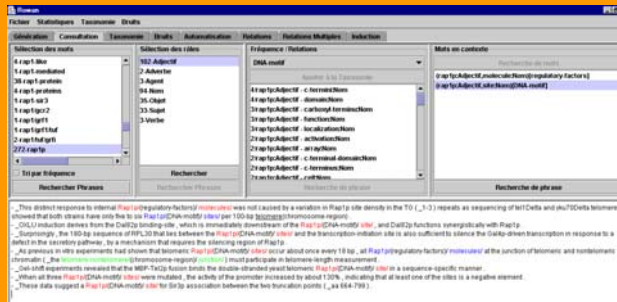
## Third step : Taxonomy

Input : terminology cleaned corpus  
Taxonomy construction with an expert :

- **First approach :**  
Examining the list of terms and determining the belonging or not to the field concept (% of terms belonging to a concept : 70).

### Second approach semi-automatic :

- Xerox's Shallow parsing.
- Using Rowan, a software developed by our team for taxonomy construction.



The next stages of our work will be :

- (i) Apply this taxonomy to the corpus of the whole papers corresponding to the abstracts.
- (ii) Build the necessary patterns for extracting the information available on *Saccharomyces cerevisiae* DNA-binding proteins in order to sharpen their characterization, and to elucidate the binding mechanisms.

Query : "DNA-binding protein yeast" in database Medline

6000 abstracts = CORPUS

1: Biochim Biophys Acta 2001  
The modulation of the biological activities of mitochondrial histone Abf2p by yeast PKA and its possible role in the regulation of mitochondrial DNA content during glucose repression.  
Cho JH, Lee YK, Chae CB.  
Department of Life Science and Division of Molecular and Life Science, Pohang University of Science and Technology.  
The mitochondrial histone, Abf2p, of *Saccharomyces cerevisiae* is essential for the maintenance of mitochondrial DNA and appears to play an important role in the recombination and copy number determination of mitochondrial-DNA.  
PMID: 11779632 [PubMed]

Cleaner

Cleaned corpus

The modulation of the biological activities of mitochondrial histone Abf2p by yeast PKA and its possible role in the regulation of mitochondrial DNA content during glucose repression.  
The mitochondrial histone, Abf2p, of *Saccharomyces cerevisiae* is essential for the maintenance of mitochondrial DNA and appears to play an important role in the recombination and copy number determination of mitochondrial-DNA.

Shallow Parser

ADJ(16@possible 17@role)  
ADJ(9@histone 10@Abf2-protein)  
ADJ(8@mitochondrial 10@Abf2-protein)  
ADJ(5@biological 6@activity)  
NNPREP(17@role 18@in 20@regulation)  
NNPREP(10@Abf2-protein 11@by 12@yeast)  
NNPREP(6@activity 7@of 10@Abf2-protein)  
NNPREP(2@modulation 3@of 6@activity)  
SUBJ(5@Abf2-protein 22@play)  
SUBJ(5@Abf2-protein 20@appear)

Answering a biological question :  
Characterization of the *S. cerevisiae* DNA-binding proteins

- "Best" terms for the field
- 1 C terminal-> 1.85e+303
  - 2 DNA binding-> 1.61e+298
  - 3 N terminal-> 1.59e+286
  - 4 TATA binding-> 2.91e+122
  - 5 temperature sensitive-> 4.46e+91
  - 6 carboxy terminal-> 5.73e+67
  - 7 mating type-> 4.78e+61
  - 8 mitochondrial DNA-> 3.22e+49
  - 9 cell cycle -> 7.08e+40
  - 10 zinc finger-> 7.07e+40

