Presentation of the Gene Renaming Task

Julien Jourde, Monday 1st June 2009

Transys – Training Session

Named-Entities Recognition (NER)

- What is a Named-Entity (NE) ?

  - A "**rigid designator**"
  - Designates an object (physical or conceptual)
  - Rigid (stable)

  - A name (a place, a person, ...)
  - Formula, date, price ...

- What is it used for ?

  - Recognition of NE is used to retrieve relevant informations written in natural language

  - NE represent the main objects of the domain (genes, proteins, species in Biology)

  - Their recognition represents the first step of semantic analysis

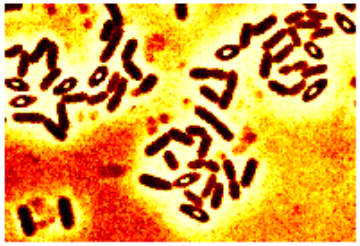
Synonymy

- A major problem for NER :
  - ➔ A given object is designated by different equivalent names
- How to interpret recognized NE ?
  - ➔ Once names are recognized in a text, they should be related to the relevant object

  ex : **spolll G** is synonym of **sigma G**

  - ➔ All so that what is known on the object can be gathered from different sources (different papers, databanks, ...)
- A specific case of synonymy is renaming.

Soft application



spoIIIG sporulation bacillus subtilis Search

Concept all [+]

- gene function
  - sporulation gene
    - forespore gene
      - forespore-spec
- gene concept

Species all [+]

Bacillus subtilis

Genes all [+]

sigG

Authors all [+]

Setlow P

Dates all [+]

2005

Query details: bacillus(lemma) subtilis(lemma) sporulation(Subtilist functional classification/cell envelope and

1-10 among 65 results in 611 categories

Expression of the Bacillus subtilis spoIVB gene is under dual sigma F/sigma G cont

However, during **sporulation**, only sigma G directs significant levels of spoIVB expression.

sigG sigF spoIVB Subtilist functional classification/information pathway/RNA synthesis/initiation of RNA synth organisms/Eukaryota/Fungi-Metazoa group/Metazoa/Eumetazoa/Bilateria/Coelomata/Protostomia/Panarthropoda/Arthropoda/Mandibulata/Pancru subtilis Subtilist functional classification/information pathway/RNA synthesis/initiation of RNA synthesis/sigF S function/enzyme/polymerase/RNA-polymerase Subtilist functional classification/cell envelope and cellular pro Factor Gene Expression Regulation, Bacterial Transcription Factors Base Sequence Molecular Sequence Data

Analysis of the interaction between the transcription factor sigmaG and the anti-s

The activation of sigma(G), a transcription factor, in **Bacillus subtilis** is coupled to the completion o

SpoIIAB SpoIIAA sigG sigF Subtilist functional classification/cell envelope and cellular process/sporulation/Sp Subtilis Molecular Biology Concept/regulator/transcription factor/anti-sigma-factor Subtilis Molecular Biology process/sporulation/SpoIIAA Subtilis Molecular Biology Concept/cell concept/cell-cycle/sporulation J Bacteriol Bacterial 2003 Evans Louise Errington Jeff Feucht Andrea Clarkson Joanna Yudkin Michael D Bacillus subtilis

Transcription of spoIVB is the only role of sigma G that is essential for pro-sigma K

Activation of pro-sigma K processing in the mother cell at late stages of **sporulation** in **Bacillus sub**

- Improvement by feeding the synonymy table

Gene and Protein Renaming

• It's a frequent phenomenon, as in *B. subtilis* :

➔ *E. coli* ortholog gene names were adopted during the 90's. Those gene names were considered as references.

Ex : ***div+ -> secA*** (in 1992, PMID : 1385592)

➔ Names are recycled.

Ex : ***guaA -> guaB*** and ***guaB -> guaA***

➔ There were already 350 genes in the genetic map. The first genomics sequence revealed many unknown genes. Formerly named **yxxX**, they are regularly renamed when the function is discovered.

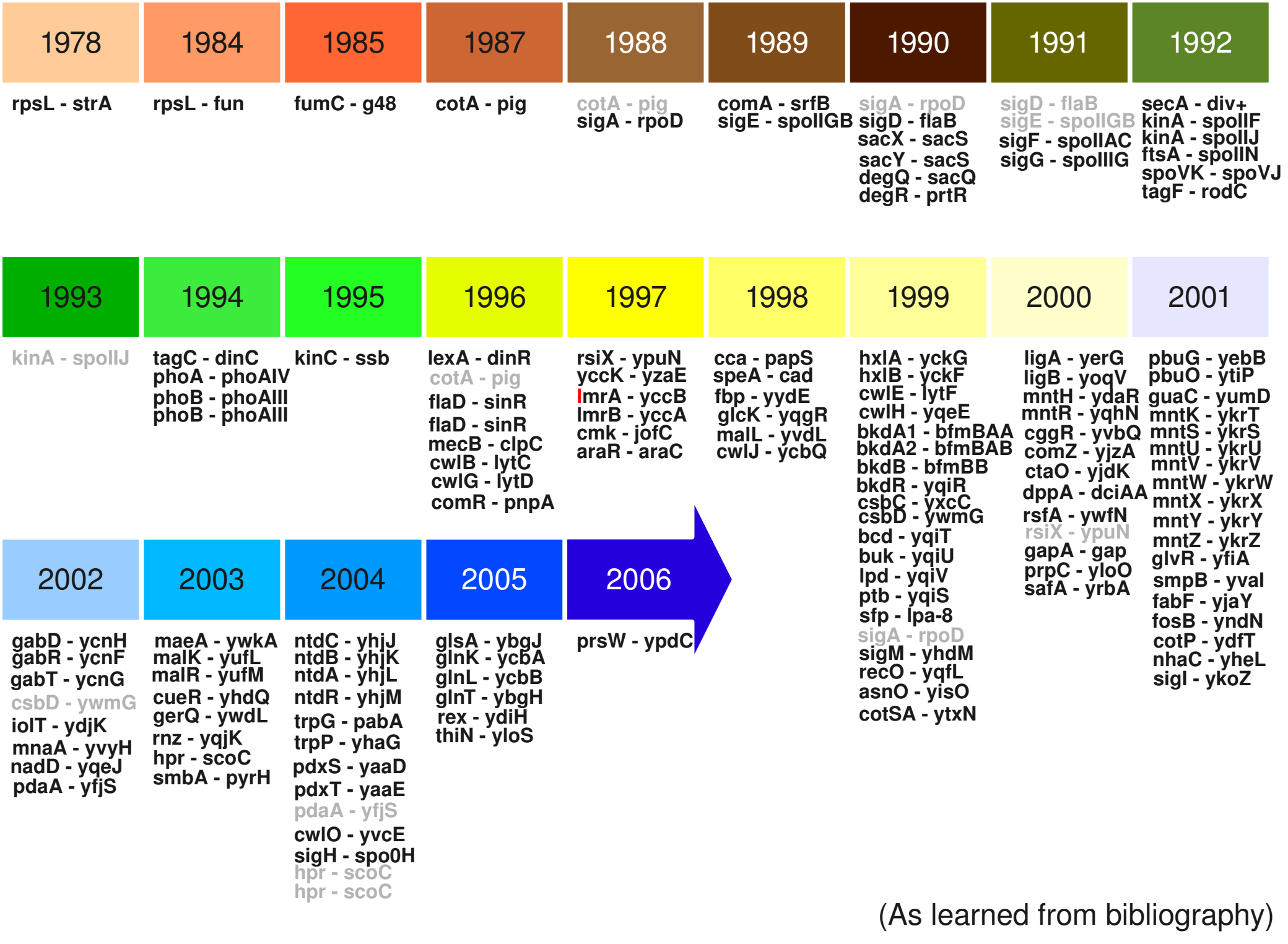
Ex : ***ydiH -> rex*** (in 2005, PMID : 16207915)

• New renaming each year (average of 8 for *B. subtilis*). Current renaming are all due to function discovery.

• The renaming information is stored in central databases.

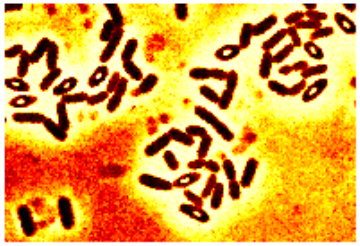
➔ Many disagreement among databases and many information in papers only.

➔ We started to look at it in 2006 (first construction of the EloAnn synonymy table [Goeltzer et al., 2007]). A new table named USST is under construction (integration of 7 different data sources : **BSORF**, **genetic map**, **EloAnn** table, **Genome Reviews**, **GenBank**, **SubtiList-GenoList** and **Swissprot**).



(As learned from bibliography)

Soft application



spolIIG sporulation bacillus subtilis Search

Concept all [+]

- gene function
  - sporulation gene
    - forespore gene
      - forespore-spec
- gene concept

Species all [+]

Bacillus subtilis

Genes all [+]

sigG

Authors all [+]

Setlow P

Dates all [+]

2005

Query details: bacillus(lemma) subtilis(lemma) sporulation(Subtilist functional classification/cell envelope and

1-10 among 65 results in 611 categories

Expression of the Bacillus subtilis spoIVB gene is under dual sigma F/sigma G cont

However, during **sporulation**, only sigma G directs significant levels of spoIVB expression.

sigG sigF spoIVB Subtilist functional classification/information pathway/RNA synthesis/initiation of RNA synth organisms/Eukaryota/Fungi-Metazoa group/Metazoa/Eumetazoa/Bilateria/Coelomata/Protostomia/Panarthropoda/Arthropoda/Mandibulata/Pancru subtilis Subtilist functional classification/information pathway/RNA synthesis/initiation of RNA synthesis/sigF S function/enzyme/polymerase/RNA-polymerase Subtilist functional classification/cell envelope and cellular pro Factor Gene Expression Regulation, Bacterial Transcription Factors Base Sequence Molecular Sequence Data

Analysis of the interaction between the transcription factor sigmaG and the anti-s

The activation of sigma(G), a transcription factor, in **Bacillus subtilis** is coupled to the completion o

SpoIIAB SpoIIAA sigG sigF Subtilist functional classification/cell envelope and cellular process/sporulation/Sp Subtilis Molecular Biology Concept/regulator/transcription factor/anti-sigma-factor Subtilis Molecular Biology process/sporulation/SpoIIAA Subtilis Molecular Biology Concept/cell concept/cell-cycle/sporulation J Bacteriol Bacterial 2003 Evans Louise Errington Jeff Feucht Andrea Clarkson Joanna Yudkin Michael D Bacillus subtilis

Transcription of spoIVB is the only role of sigma G that is essential for pro-sigma K

Activation of pro-sigma K processing in the mother cell at late stages of **sporulation** in **Bacillus sub**

- Improvement by adding dates to the synonymy table

Example of renaming in text

PMID : 14612444

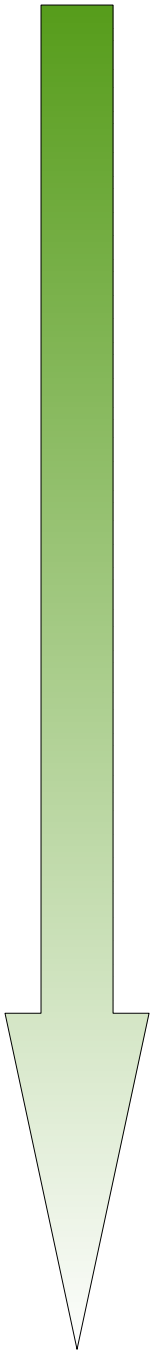
RNA polymerase mutation activates the production of a dormant antibiotic 3,3'-neotrehalosadiamine via an autoinduction mechanism in Bacillus subtilis.

Bacillus and Streptomyces species possess the ability to produce a variety of commercially important metabolites and extracellular enzymes. We previously demonstrated that antibiotic production in Streptomyces coeli-color A3(2) and Streptomyces lividans can be enhanced by RNA polymerase (RNAP) mutations selected for the rifampicin-resistant (Rif(r)) phenotype. Here, we have shown that the introduction of a certain Rif(r) rpoB mutation into a B. subtilis strain resulted in cells that overproduce an aminosugar antibiotic 3,3'-neotrehalosadiamine (NTD), the production of which is dormant in the wild-type strain. Mutational and recombinant gene expression analyses have revealed a polycistronic gene **ntdABC (formally yhjLKJ)** and a monocistronic gene **ntdR (formally yhjM)** as the NTD biosynthesis operon and a positive regulator for ntdABC, respectively. Analysis of transcriptional fusions to a lacZ reporter revealed that NTD acts as an autoinducer for its own biosynthesis genes via NtdR protein. Our results also showed that the Rif(r) rpoB mutation causes an increase in the activity of sigma(A)-dependent promoters including ntdABC promoter. Therefore, we propose that unlike the wild-type RNAP, the mutant RNAP efficiently recognized the sigma(A)-dependent promoters, resulting in the dramatic activation of the NTD biosynthesis pathway by an autoinduction mechanism.

Some examples of renaming in texts

- *The CARD15 (**also known as** NOD2) gene translation of trpP **(**yhaG**)**,*
- *XX, **previously known as** YYY*
- *We propose to **rename** yusC, yusB **and** yusA **as** metN, mete **and** metO, **respectively**.*
- *has been proposed to **rename** the nap **gene** of B. subtilis 168 **into** cesA **and** the ybfK **gene into** cesB.*
- *A new D,L-endopeptidase gene product, YojL **(renamed** CwlS**)**, plays a role*
- *we propose to **rename** YrzC CymR**, for** "cysteine metabolism repressor."*
- *we propose **renaming** ywdL **as** a spore germination gene, gerQ.*
- *Thus, the DinR protein is structurally and functionally analogous to the E. coli LexA protein, and accordingly, we propose **renaming** the protein B. subtilis LexA.*
- *It is proposed to **rename** the B. subtilis gene qdoI.*
- *The gene product of cotA **(formerly** pig**)** shows significant similarity to one of these 18 genes, cotA (**originally called** pig)*

Renaming Task history



Before 90's

• Biologists work on genes and name them by their phenotypes.

• Many labs study different phenotypes but same genes that are then merged.

During 90's

• Most of known genes are renamed as orthologs in *E. coli*. Most of names are in the form of **abc** (3 lower case letters).

1995

• Genomics sequence incoming. Conventions used to name *E. coli* genes are adopted by the *B. subtilis* community. Gene names should now be like **abcD**, and unknown genes should now be named like **yxxX**.

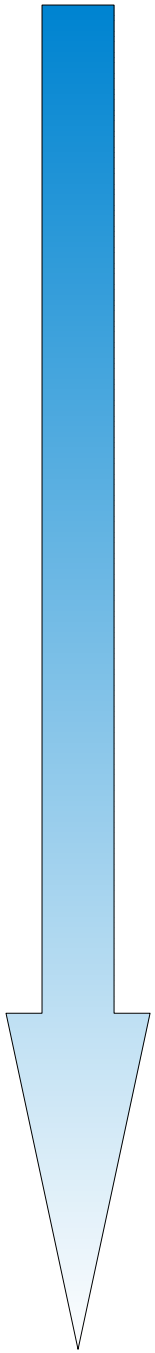
2006

• A. Goeltzer and E. Marchadier create the “EloAnn” synonymy table. They integrate data from SubtiList, Swissprot and litterature.

Since 09/2008

• J. Jourde is working on synonymy acquisition from texts to improve and finalize the “USST” table.

The new MIG Renaming Task



• Biologists working on bacteria have problems with synonymy, to retrieve and gather relevant info on given genes, especially for species they are not familiar with.

• Bibliome group works on automatic analysis of document content in Biology. What could be done about synonymy ?

09/2008 • Specific task on renaming recognition is started.

• *B. subtilis* is selected as a representative case for designing generic and reusable methods of renaming acquisition from scientific papers (corpus).

• First corpus is designed by automatic search of co-citations of synonyms (extracted from the EloAnn table) in PubMed and manual annotation of renaming relation. → **bacsuRename-192**

• Patterns for renaming recognition are designed by empiric methods.

01/2009 • First results are obtained.

→ Recall = **64%** and Precision = **100%**

• To be continued ...

Some figures

- First corpus : BacsuRename-192

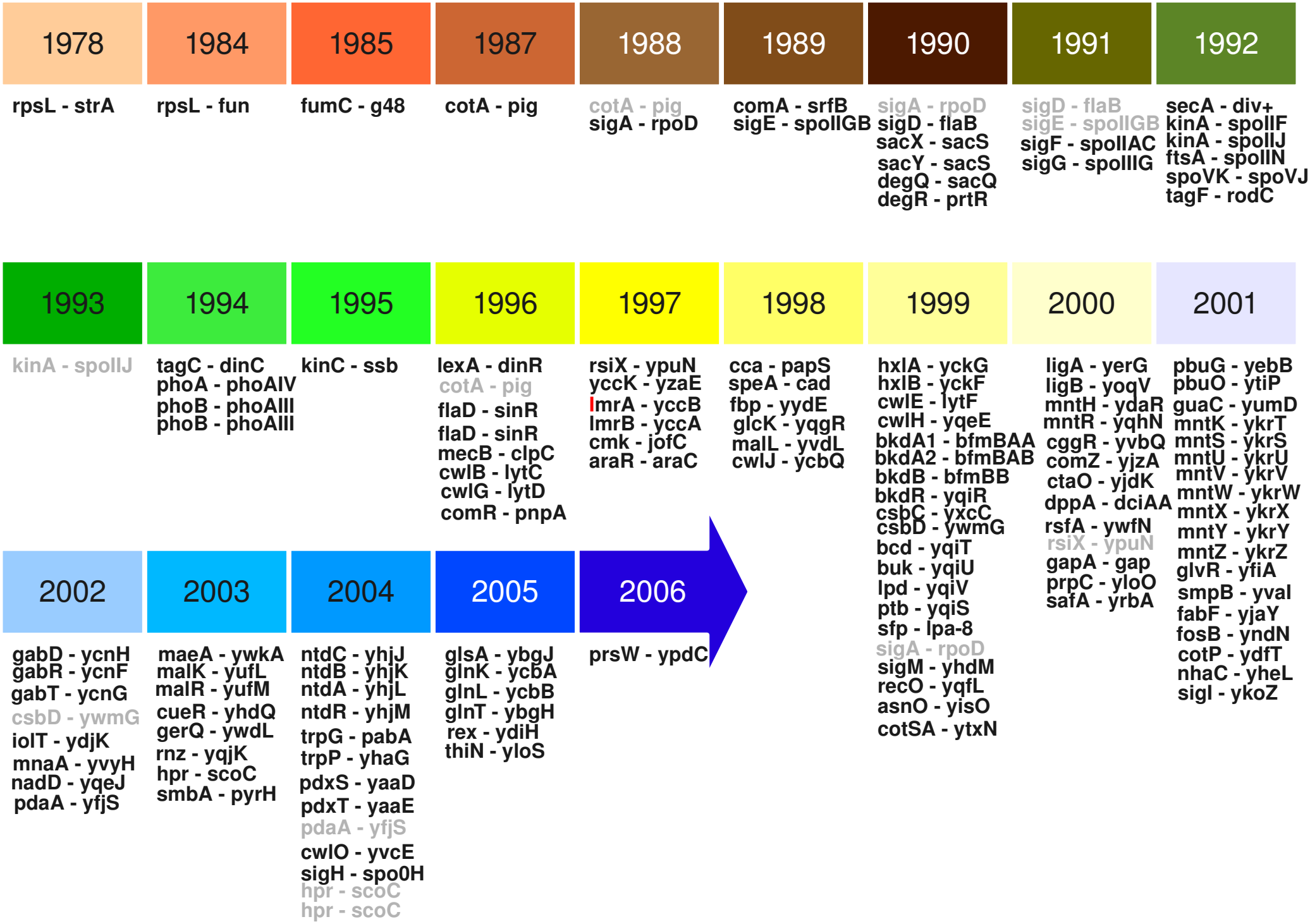
  - **192** texts BUT only **94** containing renaming
    - Automatic annotation by Alvis system (using TagEN with dictionary of hypothetical renaming couples)
    - Manual correction

Ex : "*the **ypdC** (**prsW**) gene displayed a strong effect on RsiW stability*." (PMID 17020587)

Ex : "*The DNA binding proteins **ArgR** and **AhrC** are essential for regulation of arginine metabolism in Escherichia coli and Bacillus subtilis, **respectively***." (PMID 14762010)

  - **159** citations BUT only **125** different couples

  - Recall of **64 %** on average and Precision of **100%** with Patterns


1978
rpsL - strA
1984
rpsL - fun
1985
fumC - g48
1987
cotA - pig
1988
cotA - pig
sigA - rpoD
1989
comA - srfB
sigE - spoIIGB
1990
sigA - rpoD
sigD - flaB
sacX - sacS
sacY - sacS
degQ - sacQ
degR - prtR
1991
sigD - flaB
sigE - spoIIGB
sigF - spoIIAC
sigG - spoIIIG
1992
secA - div+
kinA - spoIIF
kinA - spoIIJ
ftsA - spoIIN
spoVK - spoVJ
tagF - rodC
1993
kinA - spoIIJ
1994
tagC - dinC
phoA - phoAIV
phoB - phoAIII
phoB - phoAIII
1995
kinC - ssb
1996
lexA - dinR
cotA - pig
flaD - sinR
flaD - sinR
mecB - clpC
cwlB - lytC
cwlG - lytD
comR - pnpA
1997
rsiX - ypuN
yccK - yzaE
lmrA - yccB
lmrB - yccA
cmk - jofC
araR - araC
1998
cca - papS
speA - cad
fbp - yydE
glcK - yqgR
malL - yvdL
cwlJ - ycbQ
1999
hxlA - yckG
hxlB - yckF
cwlE - lytF
cwlH - yqeE
bkdA1 - bfmBAA
bkdA2 - bfmBAB
bkdB - bfmBB
bkdR - yqiR
csbC - yxcC
csbD - ywmG
bcd - yqiT
buk - yqiU
lpd - yqiV
ptb - yqiS
sfp - lpa-8
sigA - rpoD
sigM - yhdM
recO - yqfL
asnO - yisO
cotSA - ytxN
2000
ligA - yerG
ligB - yoqV
mntH - ydaR
mntR - yqhN
cggR - yvbQ
comZ - yjzA
ctaO - yjdK
dppA - dciAA
rsfA - ywfN
rsiX - ypuN
gapA - gap
prpC - yloO
safA - yrbA
2001
pbuG - yebB
pbuO - ytiP
guaC - yumD
mntK - ykrT
mntS - ykrS
mntU - ykrU
mntV - ykrV
mntW - ykrW
mntX - ykrX
mntY - ykrY
mntZ - ykrZ
glvR - yfiA
smpB - yval
fabF - yjaY
fosB - yndN
cotP - ydfT
nhaC - yheL
sigI - ykoZ
2002
gabD - ycnH
gabR - ycnF
gabT - ycnG
csbD - ywmG
iolT - ydjK
mnaA - yvyH
nadD - yqeJ
pdaA - yfjS
2003
maeA - ywkA
malK - yufL
malR - yufM
cueR - yhdQ
gerQ - ywdL
rnz - yqjK
hpr - scoC
smbA - pyrH
2004
ntdC - yhjJ
ntdB - yhjK
ntdA - yhjL
ntdR - yhjM
trpG - pabA
trpP - yhaG
pdxS - yaaD
pdxT - yaaE
pdaA - yfjS
cwlO - yvcE
sigH - spo0H
hpr - scoC
hpr - scoC
2005
glsA - ybgJ
glnK - ycbA
glnL - ycbB
glnT - ybgH
rex - ydiH
thiN - yloS
2006
prsW - ypdC

Complexity of renaming linguistic structure

- Many structures are too complex for the pattern based approach :

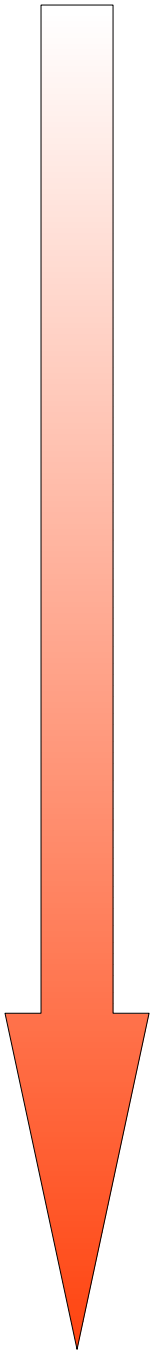
"Functional analysis showed that yebB encodes the previously characterized hypoxanthine-guanine permease PbuG and that ytiP encodes another guanine-hypoxanthine permease and is now named pbuO. yumD encodes a GMP reductase and is now named guaC."

"Decreased yqjK expression leads to an accumulation of a population of B.subtilis tRNAs in vivo, none of which have a CCA motif encoded in their genes, and YqjK cleaves tRNA precursors with the same specificity as plant RNase Z in vitro. We have thus renamed the gene rnz."

"A new gene, bkdR (formerly called yqiR), encoding a regulator with a central (catalytic) domain was found in Bacillus subtilis. This gene controls the utilization of isoleucine and valine as sole nitrogen sources. Seven genes, previously called yqiS, yqiT, yqiU, yqiV, bfmBAA, bfmBAB, and bfmBB and now referred to as ptb, bcd, buk, lpd, bkdA1, bkdA2, and bkdB, are located downstream from the bkdR gene in B. subtilis."

A parsing of syntactic dependencies and the use of machine learning are necessary

The new MIG Renaming Task



01/2009

• Complex structures can not be found with patterns.

• Machine learning could improve our soft.

04/2009

• Creation of a new corpus by automatic search of co-citations of hypothetical renaming couples (extracted from 7 different sources) in PubMed.

→ **bacsuRename-1843**

06/2009

• Annotation of the new corpus during Transys training session.

Later ...

• Soft improvement : corpus extension, richer document representation, machine learning (optimal and exhaustive example coverage).

09/2009

• update of the synonymy table and improvement of the BioAlvis service.

Corpus Extension

- Need for corpus extension :
  - To improve existing patterns
  - Machine learning will be applicable
  - To extract maximum number of renaming forms
- Extension :
  - Search of more than 4500 hypothetical renaming couples
  - Search of caracteristical forms of renaming (“*termed*”, “*designated as*”, ...)
  - Bootstrapping (planned for later, to extend the corpus again)

Corpus BacsuRename-1843

- We obtained **703** abstracts with co-citations. We added **1140** abstracts containing a renaming term for a total of **1843** texts.

- We detected **1014** citations. That's **588** different couples.

- Only names of genes and proteins of *B. subtilis* in renaming mentions have to be annotated.

- Expecting annotation time : **8** people / day

Results of the annotation session

annotator	abstracts	without annotation	with annotation	renaming pairs	uncertain way	uncertain species	uncertain synonymy	uncertain element	uncertain association	comment
aalbiniak	30	21	9	17	4	0	0	0	0	0
jbaglieri	26	17	9	21	10	0	1	0	1	0
pllorisgarc	33	21	12	30	6	0	9	0	1	0
vgoosens	52	37	15	35	22	0	2	0	3	0
cmonteferra	66	39	27	61	6	0	0	0	1	0
mskwark	23	12	11	25	7	0	2	0	0	0
slundstrom	12	4	8	22	6	0	1	0	1	0
cmackichan	57	18	39	77	4	0	0	0	3	0
mferens	16	15	1	1	0	0	0	0	0	0
Total	315	184	131	289	65	0	15	0	10	0

Results of the annotation session