**Université Henri Poincaré Nancy 1**
**Université Nancy 2**
**Institut National Polytechnique de Lorraine**

**D.E.S.S. Information Scientifique et Technique**
**Intelligence Economique**
**Année universitaire 2000-2001**

# Développement des applications de DILIB "IMD" et "Transcriptome" (Annexes)

**par**

## Claude Nemurat

## Maîtres de stage :

**Jacques DUCLOY** **Dr Bertrand RIHN**

Stage effectué du 1er Mai au 31 Juillet 2001 à :

**Institut National de Recherche et de Sécurité**
**et**
**Institut National de l'Information Scientifique et Technique**

# ANNEXES

# *ANNEXE 1 :*

## *Exemple de structure de notices Medline.*

```
UI  - 21100868
PMID- 11161789
DA  - 20010222
DCOM- 20010607
IS  - 0888-7543
VI  - 70
IP  - 3
DP  - 2000 Dec 15
TI  - Structure, sequence, and promoter analysis of human disabled-2 gene
      (DAB2).
PG  - 381-6
AB  - Disabled-2 (DAB2 for human and Dab2 for other species) is one of two
      mammalian orthologues of Drosophila Disabled. DAB2 exhibits
      properties of a tumor suppressor gene: the expression of DAB2 is
      eliminated in 85-95% of breast and ovarian tumors; homozygous
      deletions of the gene have been found in some of these tumors; and
      reintroduction of DAB2 expression suppresses tumorigenicity of
      carcinoma cells. To study the mechanisms of loss of expression and to
      detect possible mutations in tumors, we have investigated the genomic
      structure of the DAB2 gene. The complete DAB2 gene was identified and
      sequenced from four overlapping BAC clones found to contain the gene.
      Complement factor 9 (C9) gene was localized next to the DAB2 gene at
      the 3'-end of the BAC DNA fragments. The human DAB2 gene is about 35
      kb in size and consists of 15 exons and 14 introns, producing an
      approximately 4-kb message. A spliced variant corresponding to mouse
      Dab2 p93 and a 3'-end spliced variant were also identified. The
      translation initiation site resides in the second exon, and the
      noncoding first exon is separated from the second exon by a 14-kb
      intron. The 420-bp sequence 5' of exon 1 contains a CpG island (39
      CpG sites). This 420-bp       putative promoter was found to contain
      the site for transcription initiation, identified by RNase protection
      assay, and is sufficient for active transcription in epithelial
      cells. The information about the gene structure of DAB2 will enable
      us to analyze possible mutations and the mechanisms of loss of DAB2
      expression in tumors.
      AD  - Department of Biochemistry and Winship Cancer Institute, Emory
      University School of Medicine, Atlanta, Georgia 30322, USA.
AU  - Sheng Z
AU  - He J
AU  - Tuppen JA
AU  - Sun W
AU  - Fazili Z
AU  - Smith ER
AU  - Dong FB
AU  - Xu XX
LA  - eng
SI  - GENBANK/AF188298
SI  - GENBANK/AF205890
SI  - GENBANK/AF218839
SI  - GENBANK/NM001343
SI  - GENBANK/U18869
SI  - GENBANK/U39050
ID  - R01 CA75389/CA/NCI
ID  - R01 CA79716/CA/NCI
PT  - Journal Article
```

CY  - United States
TA  - Genomics
JC  - GEN
JID - 8800135
RN  - 0 (Chromosomes, Bacterial Artificial)
RN  - 0 (DNA, Complementary)
RN  - 0 (Proteins)
RN  - 0 (disabled-2 protein)
SB  - IM
MH  - Base Sequence
MH  - Chromosomes, Bacterial Artificial
MH  - DNA, Complementary
MH  - Human
MH  - Molecular Sequence Data
MH  - *Promoter Regions (Genetics)
MH  - Proteins/*genetics
MH  - Support, Non-U.S. Gov't
MH  - Support, U.S. Gov't, P.H.S.
EDAT- 2001/02/13 11:00
MHDA- 2001/06/08 10:01
AID - 10.1006/geno.2000.6383 [doi]
AID - geno.2000.6383 [pii]
PST - ppublish
SO  - Genomics 2000 Dec 15;70(3):381-6.

# *ANNEXE 2 :*
## *Exemples de notices Medline au format SGML.*

```
<medline>
  <UI>21100868</UI>
  <PMID>11161789</PMID>
  <DA>20010222</DA>
  <DCOM>20010607</DCOM>
  <IS>0888-7543</IS>
  <VI>70</VI>
  <IP>3</IP>
  <DP>2000 Dec 15</DP>
  <TI>Structure, sequence, and promoter analysis of human disabled-2 gene
(DAB2).</TI>
  <PG>381-6</PG>
  <AB>
    <s>Disabled-2 (DAB2 for human and Dab2 for other species) is one of
two</s>
    <s>mammalian orthologues of Drosophila Disabled. DAB2 exhibits
    properties of
</s>
    <s>a tumor suppressor gene: the expression of DAB2 is eliminated in 85-
95&percnt; of</s>
    <s>breast and ovarian tumors; homozygous deletions of the gene have
been</s>
    <s>found in some of these tumors; and reintroduction of DAB2
expression</s>
    <s>suppresses tumorigenicity of carcinoma cells. To study the
mechanisms of</s>
    <s>loss of expression and to detect possible mutations in tumors, we
have</s>
    <s>investigated the genomic structure of the DAB2 gene. The complete
DAB2</s>
    <s>gene was identified and sequenced from four overlapping BAC clones
found</s>
    <s>to contain the gene. Complement factor 9 (C9) gene was localized
next to</s>
    <s>the DAB2 gene at the 3&apos;-end of the BAC DNA fragments. The human
DAB2 gene</s>
    <s>is about 35 kb in size and consists of 15 exons and 14 introns,
producing</s>
    <s>an approximately 4-kb message. A spliced variant corresponding to
mouse</s>
    <s>Dab2 p93 and a 3&apos;-end spliced variant were also identified.
The</s>
    <s>translation initiation site resides in the second exon, and the
noncoding</s>
    <s>first exon is separated from the second exon by a 14-kb intron. The
420-bp</s>
    <s>sequence 5&apos; of exon 1 contains a CpG island (39 CpG sites).
This 420-bp</s>
    <s>putative promoter was found to contain the site for
transcription</s>
    <s>initiation, identified by RNase protection assay, and is sufficient
for</s>
 <s>active transcription in epithelial cells. The information about the
gene</s>
```

```
    <s>structure of DAB2 will enable us to analyze possible mutations and
the</s>
    <s>mechanisms of loss of DAB2 expression in tumors.</s>
  </AB>
  <AD>Department of Biochemistry and Winship Cancer Institute, Emory
University
School of Medicine, Atlanta, Georgia 30322, USA.</AD>
  <AU>
    <e>Sheng Z</e>
    <e>He J</e>
    <e>Tuppen JA</e>
    <e>Sun W</e>
    <e>Fazili Z</e>
    <e>Smith ER</e>
    <e>Dong FB</e>
    <e>Xu XX</e>
  </AU>
  <LA>eng</LA>
  <SI>GENBANK&sol;AF188298</SI>
  <SI>GENBANK&sol;AF205890</SI>
  <SI>GENBANK&sol;AF218839</SI>
  <SI>GENBANK&sol;NM001343</SI>
  <SI>GENBANK&sol;U18869</SI>
  <SI>GENBANK&sol;U39050</SI>
  <ID>R01 CA75389&sol;CA&sol;NCI</ID>
  <ID>R01 CA79716&sol;CA&sol;NCI</ID>
  <PT>Journal Article</PT>
  <CY>United States</CY>
  <TA>Genomics</TA>
  <JC>GEN</JC>
  <JID>8800135</JID>
  <RN>
    <e>0 (Chromosomes, Bacterial Artificial)</e>
    <e>0 (DNA, Complementary)</e>
    <e>0 (Proteins)</e>
    <e>0 (disabled-2 protein)</e>
  </RN>
  <SB>IM</SB>
  <DE>
    <e>Base Sequence</e>
    <e>Chromosomes, Bacterial Artificial</e>
    <e>DNA, Complementary</e>
    <e>Human</e>
    <e>Molecular Sequence Data</e>
    <e>*Promoter Regions (Genetics)</e>
    <e>Proteins&sol;*genetics</e>
    <e>Support, Non-U.S. Gov&apos;t</e>
    <e>Support, U.S. Gov&apos;t, P.H.S.</e>
  </DE>
  <EDAT>2001&sol;02&sol;13 11:00</EDAT>
  <MHDA>2001&sol;06&sol;08 10:01</MHDA>
  <AID>10.1006&sol;geno.2000.6383 &lsqb;doi&rsqb;</AID>
  <AID>geno.2000.6383 &lsqb;pii&rsqb;</AID>
  <PST>ppublish</PST>
  <SO>Genomics 2000 Dec 15;70(3):381-6.</SO>
</medline>
```

# *ANNEXE 3 :Fichier IMD.desc.ed*

Ce fichier permet de paramétrer le serveur d'investigation.

```
<server code=IMD>
 <--declaration des dictionnaires contenant des informations necessaires à
 l'affichage des pages HTML-->
 <www lang=FR code=STD>
        <dict>IMD.dict
        <desc env=DILIB>tools/Server/minibib.desc.www.sg
  </www>
  <www lang=FR code=FR>
        <dict>IMD.dict
        <desc env=DILIB>tools/Server/minibib.desc.www.sg
  </www>

  <--ddefinition du type generique des bases-->
  <generic type=base code=airs>
      <--chemin d'acces aux elements titre et auteur utilises pour
      l'affichage des documents pertinents-->
     <title>doc/TITR#
     <author>doc/AUTE/e#
     <--declaration des index-->
     <index code=AUTE>
        <path>doc/AUTE/e#
        <minibibTag>AU
     </index>
     <index code=TITR text=EN>
        <path>doc/TITR
     </index>
     <index code=DESC lexicon=yes>
     <reject level=assoc>DE.reject.table
     <path>doc/DESC/e#
     </index>
  </generic>

  <--declaration de la structure de la base INOR-->
  <base code=INOR>
     <title>doc/PAYS_FR#
     <author>doc/ORG_ORIG/e#
     <input type=command>bidon
     <declaration des index-->
     <index code=AUTE>
         <--chemin d'acces dans le document xml contenant les references
      bibliographiques-->
         <path>doc/ORG_ORIG/e#
         <--declaration du separateur d'occurence-->
         <sep>;
     </index>
     <index code=DESC>
         <path>doc/THEME_FR/e#
         <sep>;
     </index>
  </base>
```

```
<base code=NGS from=airs>
 <title>doc/TITR#
  <author>doc/AUTE/e#
   <input type=command>bidon
   <index code=AUTE>
       <path>doc/AUTE/e#
       <sep>,
   </index>
   <index code=DESC>
      <path>doc/DESC/e#
      <sep>,
  </index>
</base>

<base code=NFOR>
  <title>doc/SIGLE#
  <author>doc/ORGA/e#
  <input type=command>bidon
  <index code=AUTE>
      <path>doc/ORGA/e#
      <sep>,
  </index>
  <index code=DESC>
       <path>doc/DISC/e#
       <sep>,
  </index>
</base>

<base code=NLICE from=airs>
  <title>doc/TITR#
  <author>doc/AUTE/e#
  <input type=command>bidon
  <index code=AUTE>
      <path>doc/AUTE/e#
      <sep>,
  </index>
  <index code=DESC>
      <path>doc/DESC/e#
      <sep>,
  </index>
</base>

<base code=NDOC from=airs>
   <title>doc/TITR#
   <author>doc/AUTE/e#
   <input type=command>cat Text/NDOC.BIB.0 |   MedlineFromWww
   <inputDep>Text/NDOC.BIB.0
   <input type=command>bidon
   <index code=AUTE>
       <path>doc/AUTE/e#
       <sep>,
   </index>
   <index code=DESC>
       <path>doc/DESC/e#
       <sep>,
   </index>
</base>
```

```
<base code=NTV from=airs>
  <title>doc/TITR#
  <author>doc/AUTE/e#
  <input type=command>bidon
  <index code=AUTE>
      <path>doc/AUTE/e#
      <sep>,
  </index>
  <index code=DESC>
      <path>doc/DESC/e#
      <sep>,
  </index>
</base>


<base code=NEE from=airs>
   <title>doc/TITR#
   <author>doc/AUTE/e#
   <input type=command>bidon
   <index code=AUTE>
       <path>doc/AUTE/e#
       <sep>,
   </index>
   <index code=DESC>
       <path>doc/DESC/e#
       <sep>,
   </index>
</base>

<base code=NGPS from=airs>
  <title>doc/TITR#
  <author>doc/AUTE/e#
  <input type=command>bidon
  <index code=AUTE>
      <path>doc/AUTE/e#
      <sep>,
  </index>
  <index code=DESC>
      <path>doc/DESC/e#
      <sep>,
  </index>
</base>

<base code=NVIB from=airs>
   <title>doc/TITR#
   <author>doc/AUTE/e#
   <input type=command>bidon
   <index code=AUTE>
        <path>doc/AUTE/e#
        <sep>,
   </index>
   <index code=DESC>
        <path>doc/DESC/e#
        <sep>,
   </index>
</base>
```

```
<base code=NEBPM from=airs>
   <title>doc/TITR#
   <author>doc/AUTE/e#
   <input type=command>bidon
   <index code=AUTE>
        <path>doc/AUTE/e#
        <sep>,
   </index>
   <index code=DESC>
        <path>doc/DESC/e#
        <sep>,
   </index>
</base>

<base code=NCIMPE from=airs>
   <title>doc/TITR#
   <author>doc/AUTE/e#
   <input type=command>bidon
   <index code=AUTE>
       <path>doc/AUTE/e#
       <sep>,
   </index>
   <index code=DESC>
       <path>doc/DESC/e#
       <sep>,
   </index>
</base>

<base code=NPSY from=airs>
  <title>doc/TITR#
  <author>doc/AUTE/e#
  <input type=command>bidon
  <index code=AUTE>
       <path>doc/AUTE/e#
       <sep>,
  </index>
  <index code=DESC>
       <path>doc/DESC/e#
       <sep>,
  </index>
</base>

<base code=NMEDIA from=airs>
    <title>doc/TITR#
    <author>doc/AUTE/e#
    <input type=command>bidon
    <index code=AUTE>
         <path>doc/AUTE/e#
         <sep>,
    </index>
    <index code=DESC>
         <path>doc/DESC/e#
         <sep>,
    </index>
</base>
```

```
  <base code=NESS from=airs>
      <title>doc/TITR#
      <author>doc/AUTE/e#
      <input type=command>bidon
      <index code=AUTE>
           <path>doc/AUTE/e#
           <sep>,
           </index>
      <index code=DESC>
           <path>doc/DESC/e#
           <sep>,
      </index>
  </base>

  <base code=PDOC>
      <title>doc/TITRE#
      <author>doc/AUTEUR/e#
      <input type=command>bidon
      <index code=AUTE>
           <path>doc/AUTEUR/e#
           <sep>;
           </index>
      <index code=DESC>
           <path>doc/DE1/e#
           <sep>;
      </index>
  </base>

   <base code=NSTRESS from=airs>
      <title>doc/TITR#
      <author>doc/AUTE/e#
      <input type=command>bidon
      <index code=AUTE>
           <path>doc/AUTE/e#
           <sep>;
           </index>
      <index code=DESC>
           <path>doc/DESC/e#
           <sep>;
      </index>
  </base>

</server>
```

# *ANNEXE 4 : Fichier IMD.base.liste*

```
<!-- Ce fichier Sgml contient la liste des bases ainsi qu'un certain nombre
d'informations sur celles ci:son nom, le chemin d'acces aux notices de la
base,son intitule, et le chemin d'acces a son fichier de description.>
<!-- le chemin de la balise "data" n'est pas definit par une variable
d'environnement, il convient d'indiquer correctement le chemin des donnees,
moyennant quoi il n'est plus necessaire d'importer les donnees ou
l'application, est generee.>
<!-- pour l'instant toute les bases sont decrites dans le fichier
IMD.desc.ed. A termes chaque base aura son propre document de description
dont le chemin sera contenu dans la balise "desc">

<imd>
      <base>
            <code>NDOC</code>
            <data>/users/nemurat/IMD1/Text/NDOC.BIB.0</data>
            <desc>$SERVER_DIR/NDOC.desc.ed</desc>
            <intitule>Base Bibliographique commune du Centre de Recherche
et de Formation</intitule>
      </base>
      <base>
            <code>NTV</code>
            <data>/users/nemurat/IMD1/Text/NTV.BIB.0</data>
            <desc>$SERVER_DIR/NTV.desc.ed</desc>
            <intitule>Base Bibliographique du D&eacute;partement
Ing&eacute;nierie des Proc&eacute;d&eacute;s</intitule>
      </base>
      <base>
            <code>NEE</code>
            <data>/users/nemurat/IMD1/Text/NEE.BIB.0</data>
            <desc>$SERVER_DIR/NEE.desc.ed</desc>
            <intitule>Base Bibliographique de l'unit&eacute;
Epid&eacute;miologie en Entreprise</intitule>
      </base>
      <base>
            <code>NGPS</code>
            <data>/users/nemurat/IMD1/Text/NGPS.BIB.0</data>
            <desc>$SERVER_DIR/NGPS.desc.ed</desc>
            <intitule>Base Bibliographique de l'unit&eacute; Valorisation
Information Communication</intitule>
      </base>
      <base>
            <code>NVIB</code>
            <data>/users/nemurat/IMD1/Text/NVIB.BIB.0</data>
            <desc>$SERVER_DIR/NVIB.desc.ed</desc>
            <intitule>Base Bibliographique de l'unit&eacute;
Pr&eacute;vention Technique des Machines</intitule>
      </base>
      <base>
            <code>NEBPM</code>
            <data>/users/nemurat/IMD1/Text/NEBPM.BIB.0</data>
            <desc>$SERVER_DIR/NEBPM.desc.ed</desc>
            <intitule>Base Bibliographique de l'unit&eacute; Surveillance
Biologique Exposition Substances Inorganiques</intitule>
      </base>
```

```
      <base>
            <code>NCIMPE</code>
            <data>/users/nemurat/IMD1/Text/NCIMPE.BIB.0</data>
            <desc>$SERVER_DIR/NCIMPE.desc.ed</desc>
            <intitule>Base Bibliographique du Centre Interr&eacute,gional
de Mesures Physiques de l'Est</intitule>
      </base>
      <base>
            <code>NGS</code>
            <data>/users/nemurat/IMD1/Text/NGS.BIB.0</data>
            <desc>$SERVER_DIR/NGS.desc.ed</desc>
            <intitule>Base Bibliographique de l'unit&eacute; Gestion de la
S&eacute;curit&eacute;</intitule>
      </base>
      <base>
            <code>NFOR</code>
            <data>/users/nemurat/IMD1/Text/NFOR.BIB.0</data>
            <desc>$SERVER_DIR/NFOR.desc.ed</desc>
            <intitule>R&eacute;pertoire des Adresses des Organismes de
Formation continue</intitule>
      </base>
      <base>
            <code>NLICE</code>
            <data>/users/nemurat/IMD1/Text/NLICE.BIB.0</data>
            <desc>$SERVER_DIR/NLICE.desc.ed</desc>
            <intitule>Base Bibliographique du Laboratoire
Interr&eacute;gional de Chimie de l'Est</intitule>
      </base>
      <base>
            <code>NPSY</code>
            <data>/users/nemurat/IMD1/Text/NPSY.BIB.0</data>
            <desc>$SERVER_DIR/NPSY.desc.ed</desc>
            <intitule>Base Bibliographique de l'unit&eacute; Ergonmie et
Psychologie appliqu&eacute;es &agrave; la Pr&eacute;vention</intitule>
      </base>
      <base>
            <code>NMEDIA</code>
            <data>/users/nemurat/IMD1/Text/NMEDIA.BIB.0</data>
            <desc>$SERVER_DIR/NMEDIA.desc.ed</desc>
            <intitule>Base Bibliographique de la M&eacute;diath&egrave;que
du Service Formation de Neuves Maisons</intitule>
      </base>
      <base>
            <code>NESS</code>
            <data>/users/nemurat/IMD1/Text/NESS.BIB.0</data>
            <desc>$SERVER_DIR/NESS.desc.ed</desc>
            <intitule>Base Bibliographique de l'unit&eacute; Suret&eacute;
des Syst&egrave;mes Electroniques</intitule>
      </base>
      <base>
            <code>INOR</code>
            <data>/users/nemurat/IMD1/Text/INOR.BIB.0</data>
            <desc>$SERVER_DIR/INOR.desc.ed</desc>
            <intitule>Base Bibliographique de l'unit&eacute; Suret&eacute;
des Syst&egrave;mes Electroniques</intitule>
      </base>
```

```
      <base>
            <code>PDOC</code>
            <data>/users/nemurat/IMD1/Text/PDOC.BIB.0</data>
            <desc>$SERVER_DIR/PDOC.desc.ed</desc>
            <intitule>Base Bibliographique de l'unit&eacute; Suret&eacute;
des Syst&egrave;mes Electroniques</intitule>
      </base>
      <base>
            <code>NSTRESS</code>
            <data>/users/nemurat/IMD1/Text/NSTRESS.BIB.0</data>
            <desc>$SERVER_DIR/NSTRESS.desc.ed</desc>
            <intitule>Base Bibliographique de l'unit&eacute; Suret&eacute;
des Syst&egrave;mes Electroniques</intitule>
      </base>


</imd>
```

# *ANNEXE 5 : GenereShell.sh*

```
#!/bin/sh

################## GenereShell.sh
########################################
#
# ce shell permet de genere les shells de pretraitement de chaque base, a
partir
# des informations contenues dans les fichier IMD.desc.ed et
IMD.bases.liste
# Les differents shells genere sont stockes dans le repetoire "Prog"
#
#############################################################################
##

# Gestion de la ligne de commande

# Cas 1 : pas de parametre
if [ $# -lt 1 ]
then
        echo "
        Usage :\n\t$0 'fichier'

        Entree : Fichier SGML de description

        ">&2
        exit 0
fi

# Cas 2 : le fichier n'existe pas
if [ ! -f $1 ]
then
        echo "Le fichier \"$1\" n'existe pas">&2
        exit 1
fi

# On realise ici une boucle permettant d'appliquer le meme traitement pour
# chaque base definies dans les fichier IMD.bases.liste et IMD.desc.ed

# BOUCLE

for name in `SgmlSelect -s imd/base/code# -p @s1 < $1`
    do

# Definition des variables $DESC et $AUTE . Ces variables contiennent le
nom des champs indexes sous DESC et AUTE.
# Pour certaines bases les champs DESC et AUTE sont inexistants, il a donc
fallut en utiliser d'autre
# qui restent cependant indexes sous DESC et AUTE afin de conserver un
traitement identique pour chaque bases
# par la suite
```

```
  DESC=`cat $SERVER_ROOT/IMD.desc.sg|SgmlSelect -s
server/base@code=$name/index@code=DESC/path# -p @s1|awk -F/ '{print $2}'`
  AUTE=`cat $SERVER_ROOT/IMD.desc.sg|SgmlSelect -s
server/base@code=$name/index@code=AUTE/path# -p @s1|awk -F/ '{print $2}'`

# Definition des variables $DESCSEP et $DESCAUTE. Ces variables contiennent
les separateurs d'occurence
# des champs DESC et AUTE, dont on doit necessairement tenir compte pour
creer les index
# Or ce separateur peut differer selon les bases, il est donc specifie dans
# le fichier de description des bases: IMD.desc.ed


  DESCSEP=`cat $SERVER_ROOT/IMD.desc.sg|SgmlSelect -s
server/base@code=$name/index@code=DESC/sep# -p @s1`
  AUTESEP=`cat $SERVER_ROOT/IMD.desc.sg|SgmlSelect -s
server/base@code=$name/index@code=AUTE/sep# -p @s1`

# Si le separateur est un ";" on le transforme en "," avant la suite des
traitements.

if  test "$DESCSEP" = ";"
      then
      DESCPROCESS="sed 's/ ; / , /g' | SgmlTextProc -P strtok -is \",\" -F
first doc/$DESC -R -G $DESC -E e "
        else
        DESCPROCESS="SgmlTextProc -P strtok -is \"$DESCSEP\" -F first
doc/$DESC -R -G $DESC -E e "
fi

if  test "$AUTESEP" = ";"
      then
      AUTEPROCESS="sed 's/ ; / , /g' | SgmlTextProc -P strtok -is \",\" -F
first doc/$AUTE -R -G $AUTE -E e "
        else
        AUTEPROCESS="SgmlTextProc -P strtok -is \"$AUTESEP\" -F first
doc/$AUTE -R -G $AUTE -E e "
fi

# Definition de la variable $DATA (chemin des notices) pour une base
($name)

 DATA=`cat IMD.bases.list | SgmlSelect -g imd/base/code#$name/../.. -p @g1
| SgmlSelect -g base/data# -p @g1`

# Generation des shell de creation de chaque bases

echo "#! /bin/sh\n\n" > $SERVER_DIR/Prog/$name.sh

    echo "
    # -------------------------------------
    # Base $name
    # -------------------------------------

    BASE=\$SERVER_ROOT/$name

    rm -rf \$SERVER_ROOT/$name.bib.*
```

```
        cat  $DATA  \\
          | removeCR \\
          | AirsToEd  \\
          | SgmlCharSetTr -f ed \\
          | MiniBibFromEd -T \\
          | $DESCPROCESS \\
          | SgmlTextProc -P tableReplace -F all doc/DESC/e# -R -t
$SERVER_DIR/DE.syn.table \\
          | $AUTEPROCESS \\
          | DamHfdBuild -h $SERVER_DIR/Server/$name.bib
    "     >> $SERVER_DIR/Prog/$name.sh

 chmod 775 $SERVER_DIR/Prog/$name.sh
 done

# FIN DE BOUCLE

#le chmod permet de s'affranchir des problemes de droits È l'ouverture des
fichier de donnÅes.

############ "deroulement des commandes pipÅes au "cat $DATA"
#################
#
# removeCR: permet de suprimer les /r introduit lors du passage de windows
È
# unix.
# AirsToED: passage des donnees au format ed
# SgmlCharSetTr: transcodage des caracteres (ici de ed È Sgml)
# MiniBibFromEd: rajout des balises de fin abscentes dans le format ed
# SgmlTextProc -p strtok: ($DESCPROCESS et $AUTEPROCESS): separation des
differentes occurence d'un champs que l'on veut indexer separement
# SgmlTextProc-P tableReplace: application d'une table de synonymes pour
eviter la generation d'associations non pertinentes.
# DamHfdBuild: Generation du hfd
#
################################################################################
####
```

# *ANNEXE 6 : GenereMakeFile.sh*

```
# Ce shell permet de generer a partir des shells de pretraitement des bases
un fichier de makefile complementaire integre au fichier de makefile
standard (Figure 10 page 23)

# Gestion de la ligne de commande

# Cas 1 : pas de parametre
if [ $# -lt 1 ]
then
        echo "
        Usage :\n\t$0 'fichier'

        Entree : Fichier SGML de description

        ">&2
        exit 0
fi

# Cas 2 : le fichier n'existe pas
if [ ! -f $1 ]
then
        echo "Le fichier \"$1\" n'existe pas">&2
        exit 1
fi

echo "#-------------------------------------------------------------------------
# Makefile de l'application genere a partir de
# $1
#-------------------------------------------------------------------------

"
#BOUCLE


for name in `SgmlSelect -s imd/base/code# -p @s1 < $1`
  do

AUTEPATH=`cat $SERVER_ROOT/IMD.desc.sg|SgmlSelect -s
server/base@code=$name/index@code=AUTE/path# -p @s1`
DESCPATH=`cat $SERVER_ROOT/IMD.desc.sg|SgmlSelect -s
server/base@code=$name/index@code=DESC/path# -p @s1`


  echo "

#-------------------------------------------------------------------------
# Base $name :
#-------------------------------------------------------------------------
\$(SERVER_ROOT)/$name.bib.hcs:    `cat IMD.bases.list  \
  | SgmlSelect -g imd/base/code#${name}/../.. -p@g1 \
  | SgmlSelect -g base/data# -p @g1 \
 # | sed 's/\//)\//' \
 # | sed 's/SERVER_/(SERVER_/'
`  \\
                                  \$(SERVER_PROG)/$name.sh
      \$(SERVER_PROG)/$name.sh
```

```
#---- AUTE ----

\$(SERVER_ROOT)/$name.AUTE.i.hcs:   \$(SERVER_ROOT)/$name.bib.hcs     \\
                                    \$(SERVER_PROG)/$name.sh       \\
                                \$(SERVER_PROG)/reject.tab
      IndexBuildUsual -k $AUTEPATH -i AUTE  -b \$(SERVER_ROOT)/$name -r
\$(SERVER_PROG)/reject.tab


#---- DESC ----

\$(SERVER_ROOT)/$name.DESC.i.hcs:   \$(SERVER_ROOT)/$name.bib.hcs      \\
                                    \$(SERVER_PROG)/$name.sh       \\
                                    \$(SERVER_PROG)/reject.tab
      IndexBuildUsual -k $DESCPATH -i DESC -b \$(SERVER_ROOT)/$name -r
\$(SERVER_PROG)/reject.tab


  "
  done

#FIN BOUCLE

#-------------------------- End of : $name.mk -----------------------------
```

# *ANNEXE 7 : Genere.def.path.input.sh*

```
#!/bin/sh

# Ce shell permet de genere le fichier def.path.input (qui devrait
d'ailleurs s'appeler def.data.intitule)
# È partir du fichier IMD.bases.liste
# Le fichier def.path.input definit des variables contenant les
informations des balises
# "data" et "intitule" du fichier IMD.bases.liste

rm -f def.path.input
# touch def.path.input

for name in `SgmlSelect -s imd/base/code# -p @s1 < $1`
    do

    echo "

    DATA_$name=`SgmlSelect -g imd/base/code#$name/../.. -p @g1|SgmlSelect -
g imd/base/data# -p @g1 < $1`
    export DATA_$name
    INTITULE_$name=\" `SgmlSelect -g imd/base/code#$name/../.. -p
@g1|SgmlSelect -g imd/base/intitule# -p @g1 < $1` \"
    export INTITULE_$name
    " >> def.path.input

    done

chmod 775 def.path.input
```

# *ANNEXE 8 :Shell IMD.create.sh.*

```
# Shell de generation du serveur

#!/bin/sh

# SERVER variables must be defined

. bin/def.sh
echo $SERVER_DIR

# Generation du fichier def.path.input contenant les chemin d'acces des
donnees
# ainsi que l'intitule des bases, ces informations Åtant affectÅes È des
# variables d'environnement
# Ce shell prend IMD.bases.liste en parametre

sh genere.def.path.input.sh IMD.bases.list


#-----------------------------------------------
#step 2 defining Server descriptor
#

# Passage au format Sgml

SgmlFromEd <$SERVER_DIR/IMD.desc.ed  >$SERVER_ROOT/IMD.desc.sg

echo "------ end step 2 ( Server/IMD.desc.sg generated)"


#-------------------------------------------------------
#step 2bis generating base shell
#

# generation des shell de préparation. cf GenereShell.sh pour details

sh GenereShell.sh IMD.bases.list

echo "------end step 2bis "


#
# step 3 generating makefile
#

sh GenereMakeFile.sh IMD.bases.list > IMD.bases.mk
(
ServerGener -s $SERVER_ROOT/IMD.desc.sg -m
cat $SERVER_DIR/IMD.bases.mk
) |  StrSearchMerge -m '$(SERVER_ROOT)/IMD.date' > $SERVER_BIN/IMD.mk


echo "------ end step 3 ( bin/IMD.mk generated)"


#
#step 4 generating the server
#
```

```
make -f $SERVER_BIN/IMD.mk

echo "------ end step 4 (make -f $SERVER_BIN/IMD.mk) "
echo


#
# Génération dynamique du programme de statistique:
$SERVER_PROG/all.base.stat.sh
#

sh genere.all.base.stat.sh IMD.bases.list

echo " "
echo "********** End of: $SERVER_PROG/genere.all.base.stat.sh"
echo "------> $SERVER_PROG/all.base.stat ------> Généré"
echo " "


chmod 775 $SERVER_PROG/all.base.stat

#
#launch of statistic program file
#

sh $SERVER_PROG/all.base.stat.sh
echo " "
echo "********** End of: $SERVER_PROG/all.base.stat.sh"
echo " "
#
# Génération dynamique de la page d'intro: FR.IMD.index.html
#

sh genere.page.intro.sh


echo " "
echo "********** End of: genere.page.intro.sh (custom dilib for INRS ->
FR.IMD.index.sh)"
echo "------> $SERVER_ROOT/FR.IMD.index.html ------> Généré"
echo " "
#
# Génération dynamique du cgi $SERVER_DIR/htbin/desc.cgi
#

sh genere.desc.cgi.sh IMD.bases.list

echo " "
echo "********** End of: genere.desc.cgi.sh "
echo "------> $SERVER_DIR/htbin/desc.cgi         ------> Généré"
echo " "


# Pour ne pas avoir de problèmes de droits:

chmod -R 775 *
rm -f 0
```

```
echo " "
echo "*************************************************************"
echo "*                                                           *"
echo "*   END OF: $SERVER_DIR/IMD.make.sh                         *"
echo "*   La Procédure d'installation est achevée,                *"
echo "*   j'espère que tout s'est bien passé, les statistiques    *"
echo "*   ci dessous sont généralement révélateurs de la réussite *"
echo "*                                                           *"
echo "*   Sinon recommencez                                       *"
echo "*   Bonne journée ou bonne soirée                           *"
echo "*************************************************************"


#---------------------------- End of IMD.create.sh
```

# *ANNEXE 9: Importation d'une nouvelle base*
## *Liste des fichiers et répertoire à modifier avant restructuration.*

- ***REPERTOIRE Text***

Import des données conçernant la base dans un fichier: NOMBASE.BIB.1

- ***REPERTOIRE Prog***

Rajout d'un fichier contenant le shell de génération de la base: NOMBASE.bib.sh

```
 #!/bin/sh

BASE=$SERVER_ROOT/NOMBASE

rm -rf $SERVER_ROOT/NOMBASE.bib.*

cat $SERVER_TEXT/NOMBASE.BIB.1  \
          | AirsToEd \
          | SgmlCharSetTr -f ed \
          | MiniBibFromEd -T  \
          | SgmlTextProc -P strtok -is "," -F first doc/DESC -R -G DESC -E
e \
          | SgmlTextProc -P strtok -is "," -F first doc/AUTE -R -G AUTE -E
e \
          | DamHfdBuild -h $SERVER_ROOT/NOMBASE.bib


#-------------------------end NOMBASE.bib.sh---------------
```

- ***FICHIER liste_bases***

Complèter le fichier par le nom de la nouvelle base:

```
NCIMPE
NDOC
NEBPM
NEE
NGPS
NTV
NVIB
NECO
NFOR
NLICE
NMEDIA
NPSY
NESS
NOMBASE
```

### ➢ *FICHIER BASES.DECLARATIONS*

Complèter le fichier en déclarant la nouvelle base

```
NDOC        $SERVER_DIR/Text/NDOC.BIB.1
      (~Base~du~Centre~de~Recherche~et~de~Formation)
NTV         $SERVER_DIR/Text/NTV.BIB.1          (~ventilation)
NEE         $SERVER_DIR/Text/NEE.BIB.1          (~epidémiologie)
NGPS        $SERVER_DIR/Text/NGPS.BIB.1   (~Veille
scientifique~et~technique)
NVIB        $SERVER_DIR/Text/NVIB.BIB.1         (~vibrations)
NEBPM       $SERVER_DIR/Text/NEBPM.BIB.1        (~toxicite~metaux)
NCIMPE      $SERVER_DIR/Text/NCIMPE.BIB.1       (~Acoustique)
NECO        $SERVER_DIR/Text/NECO.BIB.1   (~Economie)
NFOR        $SERVER_DIR/Text/NFOR.BIB.1   (~Formation)
NLICE       $SERVER_DIR/Text/NLICE.BIB.1  (~Licence)
NMEDIA            $SERVER_DIR/Text/NMEDIA.BIB.1 (~Media)
NPSY        $SERVER_DIR/Text/NPSY.BIB.1   (~Psychologie)
NESS        $SERVER_DIR/Text/NESS.BIB.1   (~saispas)
NOMBASE         $SERVER_DIR/Text/NOMBASE.BIB.1  (~essai)
```

REMARQUE: Le contenu de ce fichier ne sert qu'à remplir le fichier def.path.input (appelÅ dans le shell genere.page.intro.sh) par l'intermédiaire du shell genere.def.path.input .

### ➢ *FICHIER IMD.index.mk*

Compèter le fichier par le script permettant de créer les différents index

```
#-------------------------------------------------------------------------
----
#
#     Base NOMBASE
#
#-------------------------------------------------------------------------
----

#---- AUTE ----

$(SERVER_ROOT)/NOMBASE.AUTE.i.hcs: $(SERVER_ROOT)/NOMBASE.bib.hcs     \
                                   $(SERVER_PROG)/NOMBASE.bib.sh      \
                                   $(SERVER_PROG)/reject.tab
      IndexBuildUsual -k doc/AUTE/e# -i AUTE  -b $(SERVER_ROOT)/NOMBASE -r
$(SERVER_PROG)/reject.tab

#---- DESC ----

$(SERVER_ROOT)/NOMBASE.DESC.i.hcs: $(SERVER_ROOT)/NOMBASE.bib.hcs      \
                                   $(SERVER_PROG)/NOMBASE.bib.sh       \
                                   $(SERVER_PROG)/reject.tab
      IndexBuildUsual -k doc/DESC/e# -i DESC -b $(SERVER_ROOT)/NOMBASE -r
$(SERVER_PROG)/reject.tab

#---------- END OF: NOMBASE.exception ----------------------------------
-----
```

- ***FICHIER IMD.bib.mk***

Déclararion du makefile de génération de la base

```
#------------------------------------------------------------------------
----
#                          MAKEFILE de generation de la base sur NOMBASE
#------------------------------------------------------------------------
----

$(SERVER_ROOT)/NOMBASE.bib.hcs: $(SERVER_TEXT)/NOMBASE.BIB.1 \
                                $(SERVER_PROG)/NOMBASE.bib.sh
      $(SERVER_PROG)/NOMBASE.bib.sh


#------ End of NOMBASE.mk --------
```

- ***FICHIER IMD.desc.ed***

Declaration de la structure de la base

```
<base code=NOMBASE FROM=AIRS>
    <input type=command>bidon
    <title>doc/TITR#
     <author>doc/AUTE/e#
        <index code=AUTE>
            <path>doc/AUTE/e#
        </index>
      <index code=DESC>
              <path>doc/DESC/e#
        </index>
  </base>
```

# *ANNEXE 10 :Procédure utilisateur.*

## *INTRODUCTION :*

L'application Génome est destinée à exploiter les résultats des expériences de puces à ADN. Elle permet une analyse de corpus Medline liés à une liste de gênes .
Pour pouvoir inclure plus de détails, la procédure est basée sur la configuration du réseau informatique à l'INIST. Elle a été établie en fonction des besoins éventuels rencontré par un utilisateur quasi débutant. Il est cependant préférable que celui ci connaisse les commandes Unix de base.

## *1-Importer le répertoire Génome .*

Utiliser un raccourci bureau vers Osiris. Pour le créer :

Sous windows, cliquer sur démarrer/rechercher/ordinateur
Une fenêtre apparaît demandant le nom rechercher, taper osiris
Ouvrir osiris, puis le répertoire correspondant à votre nom.
Affecter le raccourci sur le lecteur désiré

## *2-Importer les données à traiter .*

- Importer les numéros des gènes intéressants, ainsi que leurs noms et expression à partir du tableau excell contenant les résultats de l'expérimentation :

  Sous Windows, ouvrir le tableau.xls des résultats
  Sélectionner dans le tableau les lignes correspondant aux gènes sur exprimés intéressants. , puis faire édition/copier
  Ouvrir le bloc note (démarrer/accessoires/bloc note), faire édition/coller
  Depuis le blocs note, enregistrer le fichier sous :
  Nom sur osiris/Genome/Import, sous le nom TAB1.txt
  Recommencer l'opération pour les gènes sous exprimés, enregistrer le fichier au même endroit sous le nom TAB2.txt

- Il faut ensuite, importer les corpus de données à traiter :
  Pour faciliter la tâche , un programme de l'application permet de générer automatiquement les requêtes à partir à partir des fichiers TAB1.txt et TAB2.txt.
  Pour utiliser ce programme, ouvrir Xwin32, se placer dans le répertoire users/nom/Genome, pour cela taper la commande : cd /users/nom/Genome
  Lancer le programme « TraitementsPrelim.sh » : sh TraitementPrelim.sh
  Ce shell génère les requêtes et les enregistre dans les fichiers :
  /users/nom/Genome/Import/requete_corpus1
  /users/nom/Genome/Import/requete_corpus2
  Retourner sous windows
  Ouvrir le fichier requete_corpus1 (nom sur osiris/Genome/Import)
  Copier la requête
  Lancer un navigateur, se placer sur le site :
  http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
  Coller la requête, cliquer sur « go »

Afficher tout les résultats au format Medline sur la même page et les enregistrer (save) sous :

Nom sur osiris/Genome/Import/corpus_medline1.txt

Recommencer l'opération avec la requête « requete_corpus2 », enregistrer au même endroit « corpus_medline2.txt.

## *3-Initialisation*

- Se placer dans le répertoire Genome.

Ce répertoire contient initialement le répertoire « Import » et le fichier « local.sh »
Ouvrir le fichier « local.sh » et définir le chemin d'accès au répertoire « Genome :
SERVER_DIR=/……/Genome
Enregistrer le fichier modifié
Sous « Genome », taper la commande : . local.sh pour initialiser la variable SERVER_DIR.
Verifier l'initialisation en tapant la commande : echo $SERVER_DIR

## *4-Utilisation d'anti-dictionnaires à priori de l'indexation*

Pour utiliser des anti-dictionnaires à priori de l'indexation, il suffit de modifier les tables de rejet prévues à cet effet. Pour cela , ouvrir les fichiers correspondants, entrer les mots (un par ligne), puis enregistrer avant de lancer l'application.
Pour cela utiliser l'éditeur de texte sous unix pour ouvrir la table que l'on veut enrichir.
Pour les tables de rejet entrer un mots par ligne, avec un retour chariot à la fin.
Pour les tables de synonymes , taper sur une même ligne le mots suivi d'une tabulation et du mots retenu pour l'application.
Les tables disponibles,situées dans le répertoire « Import » sont les suivantes :
DE.reject.table
DE.syn.table , pour le champs Mots Clés DE

Des tables similaires sont disponibles pour les champs ABS(mots du résumé), KW(mots clés), TI(mots du titre),AU(auteurs), et RN(registry number).

NB : Il existe deux champs mots clés(DE et KW). Pour DE la table de rejet s'applique directement sur l'index, pour KW la table s'applique sur la génération des associations et clusters.

## *5-Lancement de l'application.*

Sous « Genome » lancer la commande : sh Import/Genome.sh
L'application est générée, lorsque la génération est terminée, un répertoire « Server » à été crée sous « Genome ». Le vérifier éventuellement en tapant sous « genome » la commande : ls

## *6-Visualisation*

Sous « Genome » lancer la commande : netscape Server/FR.Genome.index.html &

Les différents champs indexés sont les suivant :

aut, pour les auteurs
mots clés, pour les mots clés filtrés au niveau des associations et clusters
mots du titre pour les mots du titre
résumé pour les mots du résumé
Numéros Genbank pour les numéros Genbank
Titre pour la liste des titres des documents traités
Registry Number pour les Registry Number
Numéros Medline pour les numéros des notices medline

## *7-Mise à jour des tables de rejet à posteriori de l'indexation.*

Se placer dans le répertoire « Import »,(cd Import) puis lancer la commande :

sh Extract.sh

Il est alors demander de choisir un champs, taper le numéro du champs désiré puis valider par « entrée ».

L'index est alors extrait et apparaît à l'écran sous Xemacs.Tout les mots de l'index et de la table de rejet sont listés sont listés :

Ceux précédés d'un x sont ceux qui sont déjà dans la table de rejet, supprimer le x pour les réhabiliter.

Ceux précédés d'une simple tabulation sont les mots de l'index non contenus dans la table de rejet, ajouter un x en début de ligne pour les ajouter à la table de rejet

Ceux précédé d'un o sont les mots présent dans la table de rejet mais non prèsent dans l'index de l'appli en cours.

Une fois les modifications apportées, sauver le fichier et le fermer.
La mise à jour de la table de rejet s'effectue automatiquement et on récupéré la main.

NB : Pour que ces modifications soient prises en compte sur le server il faut relancer l'application. Pour cela revenir sous « Genome », et relancer le shell « Genome.sh » après avoir effacer la première application.

```
cd ..
rm –Rf Server
sh Import/Genome.sh
```

# *ANNEXE 11 : GenereRequête.sh*

```
#generation des listes AN.txt a partir du tableau


cat Import/TAB1.txt | awk -F\t '{print $23}' > Import/AN11.txt
cat Import/TAB2.txt | awk -F\t '{print $23}' > Import/AN22.txt


#Generation des requetes

cat Import/AN11.txt | grep '[0-9]' > Import/AN1.txt
cat Import/AN1.txt | sed 's/$/[SI]/g'| perl -e 'while(<>){chomp; print"$_
or "} print"<\n"' | sed 's/ or <//g'  > Import/requete_corpus1

cat Import/AN22.txt | grep '[0-9]' > Import/AN2.txt
cat Import/AN2.txt | sed 's/$/[SI]/g'| perl -e 'while(<>){chomp; print"$_
or "} print"<\n"' | sed 's/ or <//g'  > Import/requete_corpus2

rm AN11.txt
rm AN22.txt
rm AN1.txt
rm AN2.txt
```

# *ANNEXES 12 : ExtractIndex1.sh.*

```
# shell permettant l'extraction d'un index et sa presention dans un fichier
texte presente à l'utilisateur. Ce shell et une partie du shell
ExtractIndex.sh.

clear
status="ok"

# elaboration d'un menu permettant le choix d'un index

textemenu="Choix de l'index a extraire\n
    1 quitter
    2 Abstract Gene1
    3 Descriptor Gene1
    4 Registry Number Gene1
    5 Title Gene1
    6 Autor Gene1
    7 Keyword Gene1
    8 Abstract Gene2
    9 Descriptor Gene2
    10 Registry Number Gene2
    11 Title Gene2
    12 Autor Gene2
    13 Keyword Gene2

Entrez la rÅponse :\n"

echo "$textemenu"

# definition des variables en fonction de la reponse utilisateur

while [ $status = "ok" ]
  do
  read reponse
  clear
  case $reponse in
     1) TABLE=""
        status="fin";;
     2) INDEX=Gene1.ABS.i.hfd
        TABLE=ABS1.reject.table
        status="fin";;
     3) INDEX=Gene1.DE.i.hfd
        TABLE=DE1.reject.table
        status="fin";;
     4) INDEX=Gene1.RN.i.hfd
        TABLE=RN1.reject.table
        status="fin";;
     5) INDEX=Gene1.TI.i.hfd
        TABLE=TI1.reject.table
        status="fin";;
     6) INDEX=Gene2.ABS.i.hfd
        TABLE=ABS2.reject.table
        status="fin";;
     7) INDEX=Gene2.DE.i.hfd
        TABLE=DE2.reject.table
        status="fin";;
     8) INDEX=Gene2.RN.i.hfd
         TABLE=RN2.reject.table
```

```
        status="fin";;
    9) INDEX=Gene2.TI.i.hfd
        TABLE=TI2.reject.table
        status="fin";;
    esac
  export TABLE;
  done

# extraction de l'index choisi par l'utilisateur et presentation de l'index
apres comparaison avec la table de rejet

DamCat ../Server/$INDEX|SgmlSelect -s idx/kw# -p @s1|sed 's/^*//g' | sort -
u > index1

echo "index1 contient `wc -l < index1` lignes\n"


comm -23 index1 $TABLE | sed 's/^/ /g'  > index.tmp
comm -12 index1 $TABLE | sed 's/^/o /g'  >> index.tmp
comm -13 index1 $TABLE | sed 's/^/x /g'  >> index.tmp

cat index.tmp | sort -u > index
echo "index contient `wc -l < index` lignes\n"
```

# *ANNEXE 13 : ExtractIndex3.*

```
# derniere partie du shell ExtractIndex.sh permettant la mise a jour de la
table de rejet avec les termes rejettes par l'utilisateur.

 cat index | grep '^x' | SgmlSelect -p @2 | sort > index1.reject
 cat index | grep '^0' | SgmlSelect -p @2 | sort > index2.reject

 cat index1.reject index2.reject | sort -n > $TABLE


 rm  index.tmp
 rm index
 rm index1
 rm index1.reject
 rm index2.reject
```

# *ANNEXE 14 : Genome.desc.ed.*

```
<--description du serveur d'investigation, même principe que pour IMD-->

<server code=Genome>
  <www lang=FR code=FR>
       <dict>Import/Genome.FR.dict
       <dict>Import/Medmono.dict
       <desc env=DILIB>tools/Server/minibib.desc.www.sg
  </www>
  <www lang=EN code=EN>
       <dict>Import/Genome.EN.dict
       <dict>Import/MedMono.en.dict
       <desc env=DILIB>tools/Server/minibib.desc.www.sg
  </www>

  <generic type=base code=medline>
     <title>medline/TI#
     <author>medline/AU/e#
     <index code=aut>
        <reject level=index>Import/AU.reject.table
        <replace>Import/AU.syn.table
        <path>medline/AU/e#
        <minibibTag>AU
        <cross>aut
        <cross>TI
        <cross>tit
        <cross>DE
        <cross>RN
        <cross>ABS
        <cross>SI
        <cross>UI
     </index>
     <index code=TI text=EN>
        <reject level=index>Import/TI.reject.table
        <replace>Import/TI.syn.table
        <path>medline/TI
        <cross>TI
        <crosss>aut
        <cross>DE
        <cross>RN
        <cross>ABS
        <cross>SI
        <cross>UI
     </index>
     <index code=tit>
        <path>medline/TI#
        <cross>aut
        <cross>DE
        <cross>RN
        <cross>ABS
        <cross>SI
        <cross>UI
     </index>
```

```
<index code=DE lexicon=yes>
   <replace>Import/DE.syn.table
   <reject level=index>Import/DE.reject.table
   <path>medline/DE/e#
   <cross>DE
   <cross>aut
   <cross>TI
   <cross>tit
   <cross>RN
   <cross>ABS
   <cross>SI
   <cross>UI
   <cross>AD
   <cross>TA
</index>
<index code=RN lexicon=yes>
   <replace>Import/RN.syn.table
   <reject level=index>Import/RN.reject.table
   <path>medline/RN/e#
   <cross>RN
   <cross>aut
   <cross>TI
   <cross>tit
   <cross>DE
   <cross>ABS
   <cross>SI
   <cross>UI
</index>
<index code=ABS text=EN>
   <reject level=index>Import/ABS.reject.table
   <replace>Import/ABS.syn.table
   <path>medline/AB
   <cross>ABS
   <cross>aut
   <cross>TI
   <cross>tit
   <cross>DE
   <cross>RN
   <cross>SI
   <cross>UI
</index>
<index code=SI lexicon=yes>
   <reject level=index>Import/SI.reject.table
   <path>medline/SI#
   <cross>SI
   <cross>aut
   <cross>TI
   <cross>tit
   <cross>DE
   <cross>RN
   <cross>ABS
   <cross>UI
   <cross>AD
   <cross>TA
</index>
```

```
      <index code=UI lexicon=yes>
         <path>medline/UI#
         <cross>UI
         <cross>aut
         <cross>TI
         <cross>tit
         <cross>DE
         <cross>RN
         <cross>ABS
         <cross>SI
      </index>
      <index code=AD>
         <path>medline/AD#
      </index>
      <index code=TA>
         <path>medline/TA#
      </index>
   </generic>


   <base code=Gene1 from=medline>
      <input type=command>cat Import/corpus_medline1_dedoublonne
      <inputDep>Import/corpus_medline1_dedoublonne
      <index code=aut>
      <index code=DE lexicon=yes>
         <reject level=index>Import/DE1.reject.table
      </index>
      <index code=TI text=EN>
         <reject level=index>Import/TI1.reject.table
      </index>
      <index code=ABS text=EN>
         <reject level=index>Import/ABS1.reject.table
      </index>
      <index code=RN lexicon=yes>
         <reject level=index>Import/RN1.reject.table
      </index>
   </base>


   <base code=Gene2 from=medline>
      <input type=command>cat Import/corpus_medline2_dedoublonne
      <inputDep>Import/corpus_medline2_dedoublonne
      <index code=DE lexicon=yes>
         <reject level=index>Import/DE2.reject.table
      </index>
      <index code=TI text=EN>
         <reject level=index>Import/TI2.reject.table
      </index>
      <index code=ABS text=EN>
         <reject level=index>Import/ABS2.reject.table
      </index>
      <index code=RN lexicon=yes>
         <reject level=index>Import/RN2.reject.table
      </index>
   </base>

</server>
```

# ANNEXE 15 : GenereTableMotsCommuns.

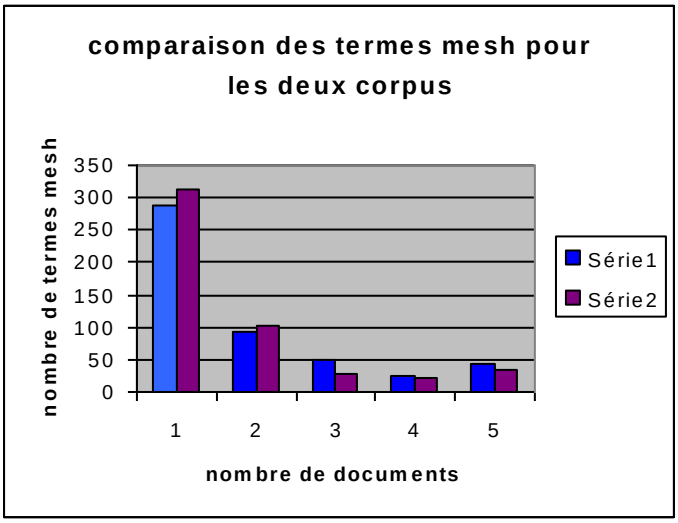
```
# Ce shell genere des tables de marquage des termes communs au deux corpus
bibliographiques. Ces tables seront prises en compte lors de la generation
par le shell TraitementPrelim.sh
#marquage des termes communs pour les mots du MESH

DamCat /applis/dps/INRS/Genome/Server/Gene1.DE.i.hfd | SgmlSelect -s
idx/kw# -p @s1|sed 's/^*//g' | sort -u > Import/DE1
DamCat /applis/dps/INRS/Genome/Server/Gene2.DE.i.hfd | SgmlSelect -s
idx/kw# -p @s1|sed 's/^*//g' | sort -u > Import/DE2
comm -12 Import/DE1 Import/DE2 > Import/DEcommuns
cat Import/DEcommuns | awk -F\t '{printf "%s\t%s(c)\n", $1,$1}' >
Import/DEcommuns.table

#marquage registry number communs

DamCat /applis/dps/INRS/Genome/Server/Gene1.RN.i.hfd | SgmlSelect -s
idx/kw# -p @s1|sed 's/^*//g' | sort -u > Import/RN1
DamCat /applis/dps/INRS/Genome/Server/Gene2.RN.i.hfd | SgmlSelect -s
idx/kw# -p @s1|sed 's/^*//g' | sort -u > Import/RN2
comm -12 Import/RN1 Import/RN2 > Import/RNcommuns
cat Import/RNcommuns | awk -F\t '{printf "%s\t%s(c)\n", $1,$1}' >
Import/RNcommuns.table

#marquage des numeros GB communs

DamCat /applis/dps/INRS/Genome/Server/Gene1.SI.i.hfd | SgmlSelect -s
idx/kw# -p @s1|sed 's/^*//g' | sort -u > Import/SI1
DamCat /applis/dps/INRS/Genome/Server/Gene2.SI.i.hfd | SgmlSelect -s
idx/kw# -p @s1|sed 's/^*//g' | sort -u > Import/SI2
comm -12 Import/SI1 Import/SI2 > Import/SIcommuns
cat Import/SIcommuns | awk -F\t '{printf "%s\t%s(c)\n", $1,$1}' >
Import/SIcommuns.table

rm Import/DE1
rm Import/DE2
rm Import/DEcommuns
rm Import/RN1
rm Import/RN2
rm Import/RNcommuns
rm Import/SI1
rm Import/SI2
rm Import/SIcommuns
```

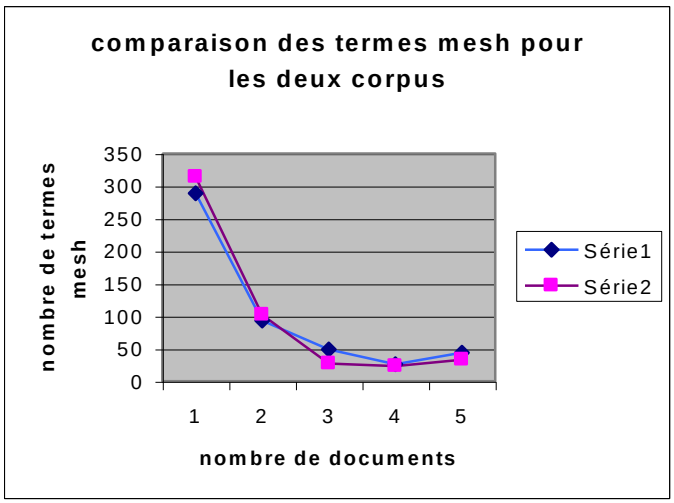
# *ANNEXE 16 : TraitementPrelim.sh.*

```
# Ce shell permet de modifier le contenu de certain champs dans le fichier
XML contenant les references bibliographiques.Il permet de prendre en
compte toutes les modifications enumerees dans la partie 6.5.

#!/usr/bin/ksh

#Mise sous format XML des fichiers medline

Import/MedlineFromWwwSplit < Import/corpus_medline1.txt
>Import/corpus_medline1.sg
Import/MedlineFromWwwSplit < Import/corpus_medline2.txt
>Import/corpus_medline2.sg

#supression des * precedent certains termes du MESH

cat Import/corpus_medline1.sg | sort -u | SgmlSelect -s medline/DE/e# -p
@s1 | grep '^*' | sed 's/^*//g' > Import/tableau1
cat Import/tableau1 | awk -F\t '{printf "*%s\t%s\n", $1,$1}' >
Import/tableau11

cat Import/corpus_medline2.sg | sort -u | SgmlSelect -s medline/DE/e# -p
@s1 | grep '^*' | sed 's/^*//g' > Import/tableau2
cat Import/tableau2 | awk -F\t '{printf "*%s\t%s\n", $1,$1}' >
Import/tableau22



#Generation de la table de marquage des genes

cat Import/TAB1.txt | awk -F\t '{printf "GENBANK/%s\t(%s) %s %s \n",
$23,$4,$23,$20}' | sed 's/ {.*}//g' | sort -u > Import/table1
cat Import/TAB2.txt | awk -F\t '{printf "GENBANK/%s\t(%s) %s %s \n",
$23,$4,$23,$20}' | sed 's/ {.*}//g' | sort -u > Import/table2

#creation de la table de marquage des titres

cat Import/corpus_medline1.sg | SgmlSelect -s medline/EDAT# -s medline/TI#
-p @s1 -p @s2 > Import/tab1
cat Import/corpus_medline2.sg | SgmlSelect -s medline/EDAT# -s medline/TI#
-p @s1 -p @s2 > Import/tab2

cat Import/tab1 | awk -F\t '{printf "%s\t(%s) %s\n", $2,$1,$2}'>
Import/tab11
cat Import/tab2 | awk -F\t '{printf "%s\t(%s) %s\n", $2,$1,$2}'>
Import/tab22
```

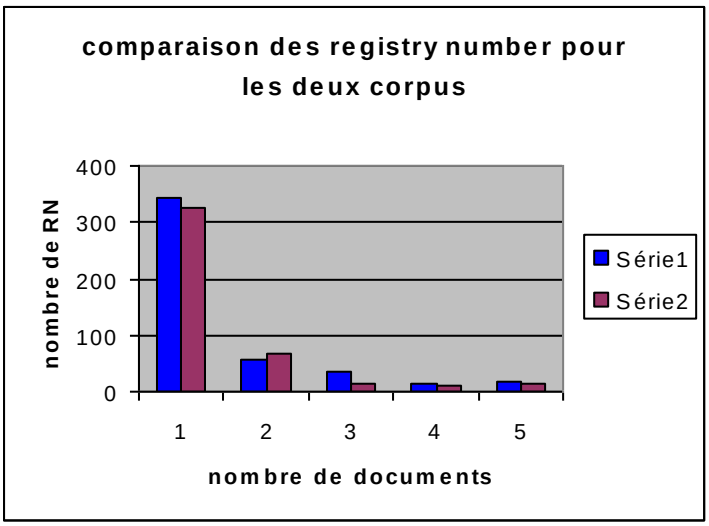
```
#Dedoublonage des notices et marquage des differents index

cat Import/corpus_medline1.sg | sort -u
      | SgmlTextProc -P tableReplace -F all medline/SI# -R -t Import/table1
      | SgmlTextProc -P tableReplace -F all medline/TI# -R -t Import/tab11
      | SgmlTextProc -P tableReplace -F all medline/DE/e# -R -t
      Import/tableau11
      | SgmlTextProc -P tableReplace -F all medline/DE/e# -R –t
      Import/DEcommuns.table
      | SgmlTextProc -P tableReplace -F all medline/RN/e# -R -t
      Import/RNcommuns.table
      | SgmlTextProc -P tableReplace -F all medline/SI# -R -t
      Import/SIcommuns.table
> Import/corpus_medline1_dedoublonne

cat Import/corpus_medline2.sg | sort -u
      | SgmlTextProc -P tableReplace -F all medline/SI# -R -t Import/table2
      | SgmlTextProc -P tableReplace -F all medline/TI# -R -t Import/tab22
      | SgmlTextProc -P tableReplace -F all medline/DE/e# -R -t
      Import/tableau22
      | SgmlTextProc -P tableReplace -F all medline/DE/e# -R -t
      Import/DEcommuns.table
      | SgmlTextProc -P tableReplace -F all medline/RN/e# -R -t
      Import/RNcommuns.table
      | SgmlTextProc -P tableReplace -F all medline/SI# -R -t
      Import/SIcommuns.table
      > Import/corpus_medline2_dedoublonne

rm Import/table1
rm Import/table2
rm Import/tab1
rm Import/tab11
rm Import/tab2
rm Import/tab22
rm Import/tableau1
rm Import/tableau11
rm Import/tableau2
rm Import/tableau22
```

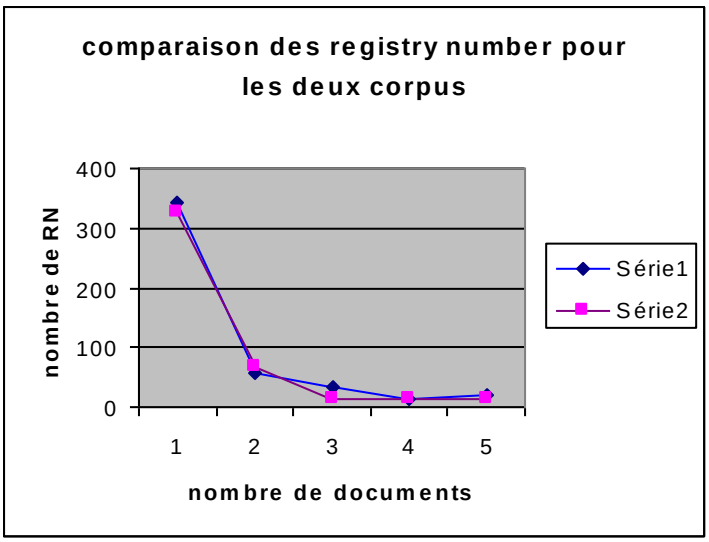
# ANNEXE 17 : Statistique sur les termes du MeSH.



# **ANNEXE 18 :** Statistiques sur les registry number.



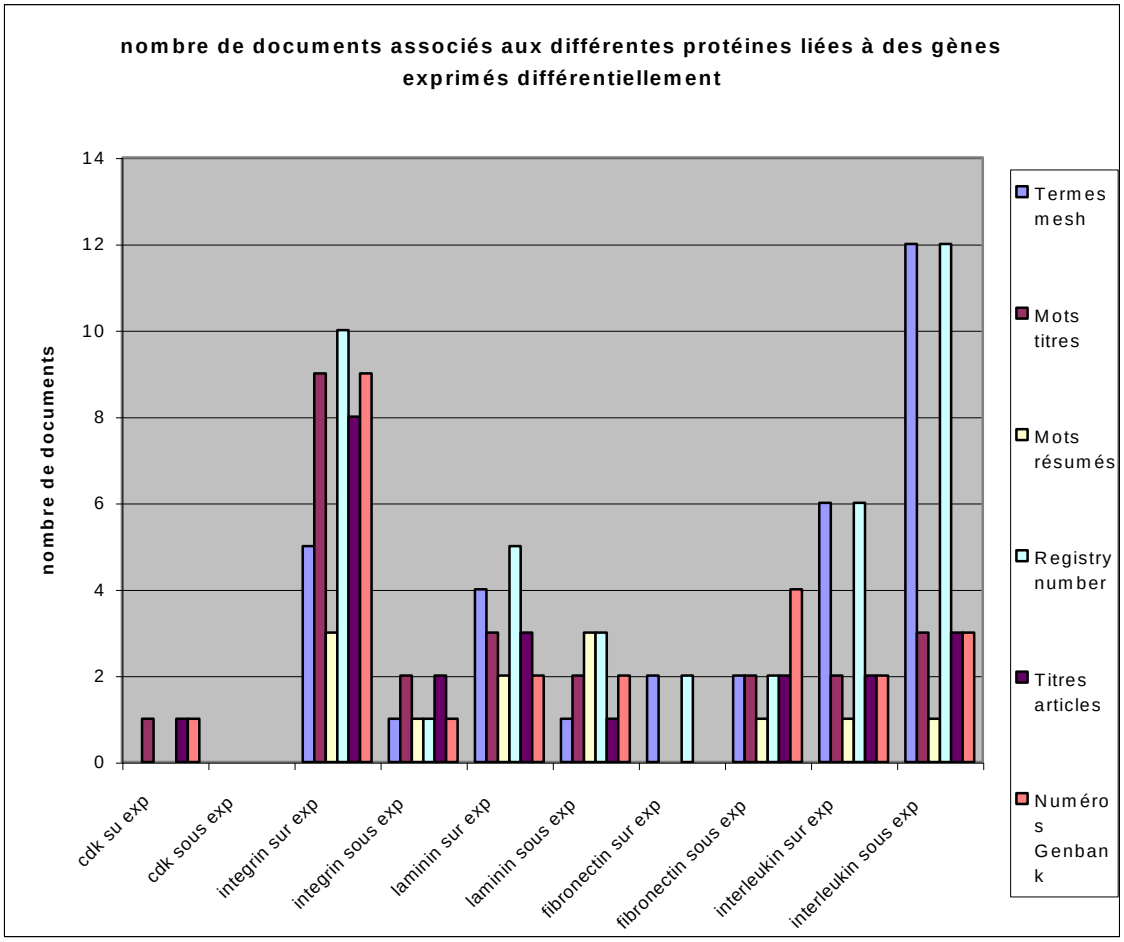
# ANNEXE 19 :

## Recherche sur certaines protéines liées à des gènes exprimés différentiellement.

| nombre de documents associés aux protéines liées a des gènes exprimés différentiellement | | | | | | |
|---|---|---|---|---|---|---|
| mots recherché | Termes mesh | Mots titres | Mots résumés | Registry number | Titres articles | Numéros Genbank |
| cdk su exp | 0 | 1 | 0 | 0 | 1 | 1 |
| cdk sous exp | 0 | 0 | 0 | 0 | 0 | 0 |
| integrin sur exp | 5 | 9 | 3 | 10 | 8 | 9 |
| integrin sous exp | 1 | 2 | 1 | 1 | 2 | 1 |
| laminin sur exp | 4 | 3 | 2 | 5 | 3 | 2 |
| laminin sous exp | 1 | 2 | 3 | 3 | 1 | 2 |
| fibronectin sur exp | 2 | 0 | 0 | 2 | 0 | 0 |
| fibronectin sous exp | 2 | 2 | 1 | 2 | 2 | 4 |
| interleukin sur exp | 6 | 2 | 1 | 6 | 2 | 2 |
| interleukin sous exp | 12 | 3 | 1 | 12 | 3 | 3 |

Corpus lié aux gènes sur-exprimés

Corpus lié aux gènes sous-exprimés


nombre de documents associés aux différentes protéines liées à des gènes exprimés différentiellement
nombre de documents
14
12
10
8
6
4
2
0
cdk su exp
cdk sous exp
integrin sur exp
integrin sous exp
laminin sur exp
laminin sous exp
fibronectin sur exp
fibronectin sous exp
interleukin sur exp
interleukin sous exp
Termes mesh
Mots titres
Mots résumés
Registry number
Titres articles
Numéros Genbank

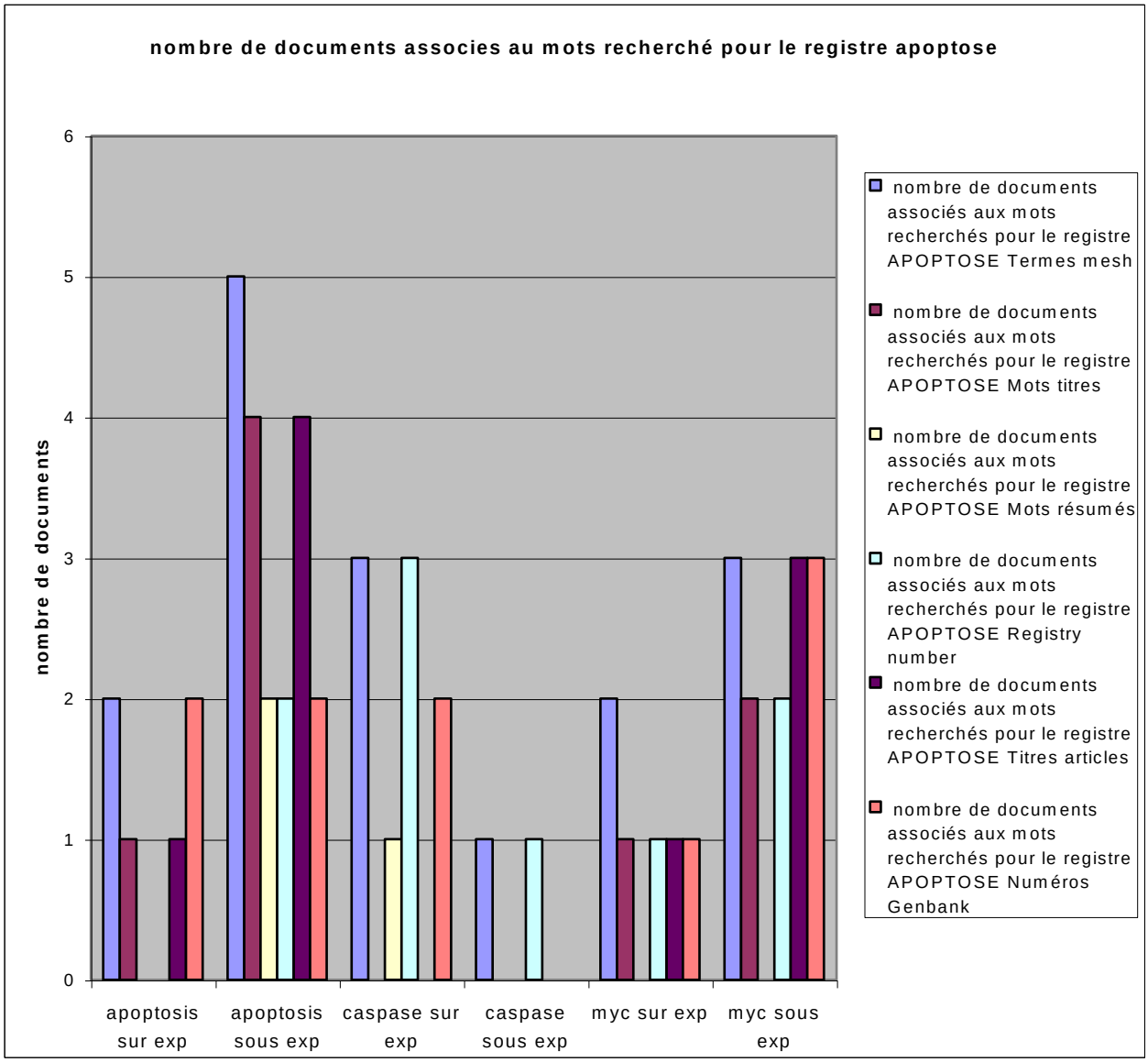
# ANNEXE 20 :

## Recherche sur le registre « Apoptose ».

| nombre de documents associés aux mots recherchés pour le registre APOPTOSE | | | | | | |
|---|---|---|---|---|---|---|
| mots recherchés | Termes mesh | Mots titres | Mots résumés | Registry number | Titres articles | Numéros Genbank |
| apoptosis | 2 | 1 | 0 | 0 | 1 | 2 |
| apoptosis | 5 | 4 | 2 | 2 | 4 | 2 |
| caspase | 3 | 0 | 1 | 3 | 0 | 2 |
| caspase | 1 | 0 | 0 | 1 | 0 | 0 |
| myc | 2 | 1 | 0 | 1 | 1 | 1 |
| myc | 3 | 2 | 0 | 2 | 3 | 3 |

Sur corpus lié aux gènes sur-exprimés

Sur corpus lié aux gènes sous-exprimés


nombre de documents associes au mots recherché pour le registre apoptose
nombre de documents
0
1
2
3
4
5
6
apoptosis sur exp
apoptosis sous exp
caspase sur exp
caspase sous exp
myc sur exp
myc sous exp
nombre de documents associés aux mots recherchés pour le registre APOPTOSE Termes mesh
nombre de documents associés aux mots recherchés pour le registre APOPTOSE Mots titres
nombre de documents associés aux mots recherchés pour le registre APOPTOSE Mots résumés
nombre de documents associés aux mots recherchés pour le registre APOPTOSE Registry number
nombre de documents associés aux mots recherchés pour le registre APOPTOSE Titres articles
nombre de documents associés aux mots recherchés pour le registre APOPTOSE Numéros Genbank