ANALYSE STATISTIQUE DE DONNÉES TEXTUELLES

Isabelle STÉVANT – Master2 MSB - 2010-2011

Table des matières

I. Introduction....1

A .Contexte biologique....1

B .Intérêt de l'analyse statistique de données textuelles pour ce sujet....1

II. Résultats de l'analyse fréquentielle....2

A .Fréquence des mots....2

B .Analyse des métaclés....3

C .Représentations bidimensionnelles des métaclés....5

III. Conclusion....7

I. INTRODUCTION

Le corpus choisi provient d'une requête sur PubMed à partir du terme « hnRNP C ». Il s'agit d'une protéine présente dans le noyau des cellules eucaryotes et qui interviendrait dans le phénomène d'épissage alternatif (voir la partie contexte biologique). Ce choix a été motivé par le sujet de stage du second semestre, qui va être de prouver l'intervention de cette protéine dans l'épissage de l'ARN, et ce par une approche bioinformatique.

A . Contexte biologique

Afin de produire les protéines nécessaire à son fonctionnement, la cellule utilise une machinerie complexe qui va transcrire son ADN en ARN, puis cet ARN va être traduit en protéines. Plusieurs étapes intermédiaires sont nécessaires à la réalisation d'une protéine fonctionnelle. L'ARN dit pré-messager va subir une étape appelée épissage avant de sortir du noyau et être traduit. Cet ARN est constitué d'une succession d'Introns et d'Exons, nécessaires ou non à la production de la protéine. Un complexe de protéine, appelé splicéosome, est capable de reconnaître ces régions et va exciser les Introns de l'ARN, les exons étant les parties contenant l'information. Cette étape est délicate, car si un exons est enlevé à la place d'un intron, la protéine issue de cet ARN sera défectueuse et peut amener à des pathologies graves.

L'assemblage et le fonctionnement du splicéosome requière plus de 150 polypeptides et 5 ribonucléoprotéines, dont les hnRNP. Ces protéines ont un domaine de fixation à l'ARN et reconnaisse une séquence particulière. L'EBI (European Bioinformatique Institute) a mis en place une méthode de biologie moléculaire couplée à une analyse informatique des données qui a permis de déterminer la séquence sur laquelle se fixe la protéine hnRNP C. Cette protéine a de plus été identifiée comme responsable de mauvais épissage d'ARN conduisant à des pathologies graves.

B . Intérêt de l'analyse statistique de données textuelles pour ce sujet

La première étape du stage consistera à faire une étude bibliographique afin de faire le bilan de ce que l'on sait déjà sur cette protéine, et dans quelle pathologie elle interviendrait. L'analyse des données textuelles va permettre non pas de réduire le nombre d'articles à étudier mais d'en avoir une synthèse.

II. Résultats de l'analyse fréquentielle

A . Fréquence des mots

827 mots sont répétés plus de 11 fois et présents dans plus de 9 textes différents du corpus, composé de 457 textes. Les mots les plus représentés sont ceux en rapport directe avec la protéine hnRNP. On y retrouve sa nature (PROTEINE, RIBONUCLEOPROTEINE), un de ses rôle (BINDING, SPLICING), ou encore sa localisation (NUCLEAR). On a à peu près 50% des 100 mots les plus représentés qui ont un rapport plus ou moins directe avec la protéine (surlignés en jaunes).

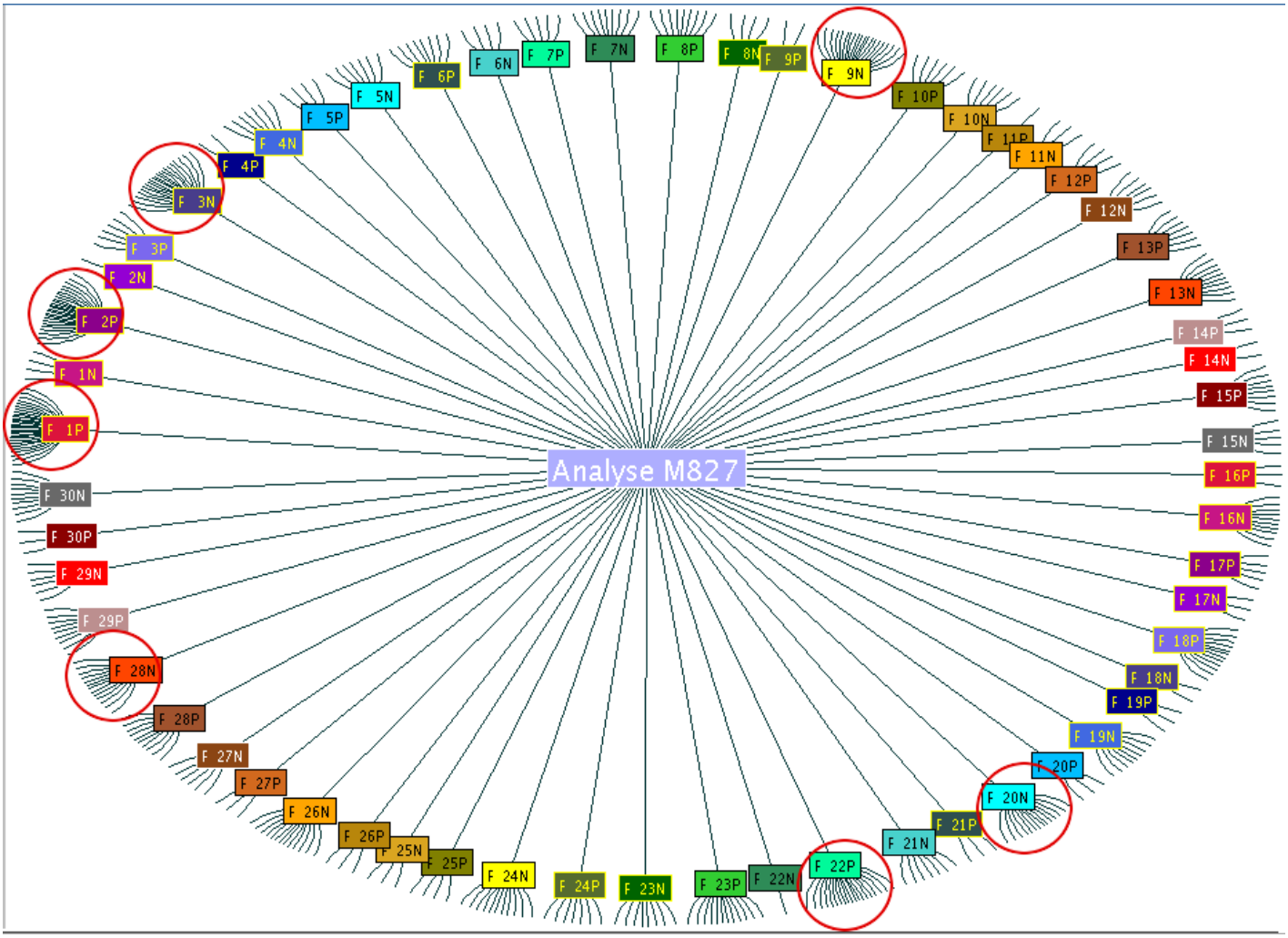
1dico	PROTEIN	1312	377
2dico	HNRNP	1217	295
3dico	PROTEINS	998	294
4dico	BINDING	948	312
5dico	RNA	937	292
6dico	NUCLEAR	740	277
7dico	MRNA	674	225
8dico	CELLS	501	223
9dico	SPLICING	419	107
10dico	EXPRESSION	353	158
11dico	HETEROGENEOUS	352	213
12dico	CELL	330	165
13dico	GENE	310	143
14dico	RIBONUCLEOPROTEIN	298	181
15dico	HUMAN	288	162
16dico	DNA	288	101
17dico	EWS	283	42
18dico	SEQUENCE	276	149
19dico	DOMAIN	275	129
20dico	REGION	245	138
21dico	COMPLEX	232	103
22dico	SPECIFIC	231	147
23dico	ACTIVITY	231	126
24dico	RICH	215	114
25dico	SITE	210	105
26dico	TRANSCRIPTION	200	109
27dico	VITRO	197	133
28dico	TERMINAL	195	113
29dico	EXON	190	52
30dico	RESULTS	184	158
31dico	ALPHA	181	44
32dico	POLY	181	79
33dico	PRE	178	93
34dico	DEPENDENT	177	115
35dico	FOUND	172	127
36dico	FACTOR	172	103
37dico	ELEMENT	167	81
38dico	VIVO	159	112
39dico	TRANSLATION	158	65
40dico	AUF	154	26
41dico	ROLE	149	125
42dico	SHOW	148	116
43dico	INTERACTION	145	97
44dico	REGULATION	144	104
45dico	KINASE	140	64
46dico	FUNCTION	140	102
47dico	FACTORS	138	100
48dico	CELLULAR	138	87
49dico	COMPLEXES	134	76
50dico	AMINO	133	76

51dico	MRNAS	131	77
52dico	DOMAINS	130	75
53dico	SUGGEST	130	125
54dico	HERE	129	128
55dico	FLI	129	23
56dico	ACID	128	86
57dico	SEQUENCES	127	83
58dico	ANALYSIS	127	93
59dico	CONTAINING	126	85
60dico	BIND	125	94
61dico	BETA	123	45
62dico	FAMILY	122	78
63dico	PHOSPHORYLATION	122	45
64dico	TRANSCRIPTIONAL	121	64
65dico	GENES	121	70
66dico	SINGLE	117	66
67dico	BINDS	117	84
68dico	ACTIVATION	117	64
69dico	NUCLEUS	116	64
70dico	FUSION	115	63
71dico	TLS	111	19
72dico	NOVEL	110	79
73dico	TYPE	109	76
74dico	STRUCTURE	109	63
75dico	SEVERAL	107	94
76dico	SPLICE	105	45
77dico	DIFFERENT	104	77
78dico	LEVELS	103	70
79dico	SITES	102	63
80dico	KDA	101	55
81dico	ALTERNATIVE	99	49
82dico	PREVIOUSLY	98	93
83dico	STUDY	97	89
84dico	ALPHACP	96	13
85dico	SMN	95	8
86dico	ADDITION	94	88
87dico	ELEMENTS	94	54
88dico	INTERACTIONS	92	72
89dico	CROSS	92	50
90dico	EWING	89	34
91dico	MOLECULAR	88	68
92dico	SHOWN	87	75
93dico	IMPORTANT	87	80
94dico	DATA	87	80
95dico	PROMOTER	87	39
96dico	MOTIF	86	48
97dico	ETS	86	22
98dico	PRESENT	86	73
99dico	BOUND	85	60
100dico	MYC	85	30

B . Analyse des métaclés

Une métaclé représente un ensemble de mots-clé décrivant un thème et peut être utilisé pour sélectionner un sous-ensemble de texte à l'intérieur de notre corpus.

Dans cet étude, nous avons 30 métaclés positives, et autant de négatives. Je pars du principe que plus il y a de mots-clés dans dans une métaclé, plus elle sera informative.



Représentation des métaclés en arbre.

On remarque de certaines métaclés contiennent beaucoup de mots et d'autres moins. J'ai entouré en rouge les métaclés contenant le plus de mots.

Il s'agits des métaclés suivantes :

- F 1P
- F 2P
- F 3N
- F 9N

- F 20N
- F 22P
- F 28P

Regardons leur contenu :

F 1P Wiki

abl, bcr, BONE, cell, cells, CYCLIN, DIFFERENTIATION, erg, ETS,
EWING, EWS, EXPRESSION, FLI, fus, FUSION, GENE, GENES, GROWTH,
liposarcoma, MARROW, myeloid, NEURAL, ONCOGENIC, ORIGIN, primary, PROGENITOR, SARCOMA,
SARCOMAS, tls, transcription, TRANSFORMATION, TRANSLOCATION, TRANSLOCATIONS, TUMOR, TUMORS,

F 2P Wiki

ALTERNATIVE, ANTISENSE, ASF, ATROPHY, ENHANCER, EXON, EXONIC, INCLUSION, intron,
INTRONIC, LOSS, MUSCULAR, MUTATION, MUTATIONS, NEURON, SILENCER, SITE, SKIPPING,
SMN, snrnp, SPINAL, SPLICE, SPLICING, SURVIVAL,

F 3N Wiki

abl, ACTIVITY, alpha, auf, bcr, CAP, CELLULAR, degradation, DEPENDENT,
differentiation, ENTRY, hcv, HEPATITIS, hur, INITIATION, INTERNAL, IRES, KNOCKDOWN,
MRNA, MYC, myeloid, pcbp, phase, REGION, RIBOSOMAL, RIBOSOME, SUPPRESSION,
TRANSLATION, TRANSLATIONAL, UTR, virus,

F 9N Wiki

auf, BEARING, CYTOPLASM, CYTOPLASMIC, domain, EXPORT, FLUORESCENT, hcv, heat,
hur, IMPORT, isoforms, liposarcoma, LOCALIZATION, nuclear, NUCLEOCYTOPLASMIC, NUCLEUS, PORE,
primary, shock, SHUTTLES, SHUTTLING, signal, TERMINAL, translocation, TRANSPORT, TRANSPORTIN,
YEAST,

F20N Wiki

aberrant, ANTI, apoptosis, APOPTOTIC, AUF, beta, bone, cell, chicken,
damage, DEFICIENT, DETECTABLE, dna, EPITOPE, export, gamma, hcv, hsp,
import, INVESTIGATE, isoforms, MAINTENANCE, marrow, MICE, nuclease, NUMBER, peptide,
peptides, regions, REPAIR, resistant, resolution, response, SEEN, subunit, viral,
virus,

F22P Wiki

abl, ACIDS, alternative, AMINO, bcr, CDNA, CLONE, CLONING, CONTAINS,
EMBRYONIC, enzyme, FRAME, GAMMA, gene, GLYCINE, HOMOLOGY, HYBRIDIZATION, MOUSE,
mutation, NORTHERN, NOVEL, phase, probe, SEQUENCE, SEQUENCING, subunit, TERMINUS,

F28P Wiki

activation, antibodies, antibody, enhancer, exon, hnrna, inclusion, intronic, matrix,
MONOCLONAL, mutation, neural, nucleus, POLYPYRIMIDINE, progression, PTB, PYRIMIDINE, REPEATS,
RRM, RRMS, SRC, state, TRACT, tumor, utr,

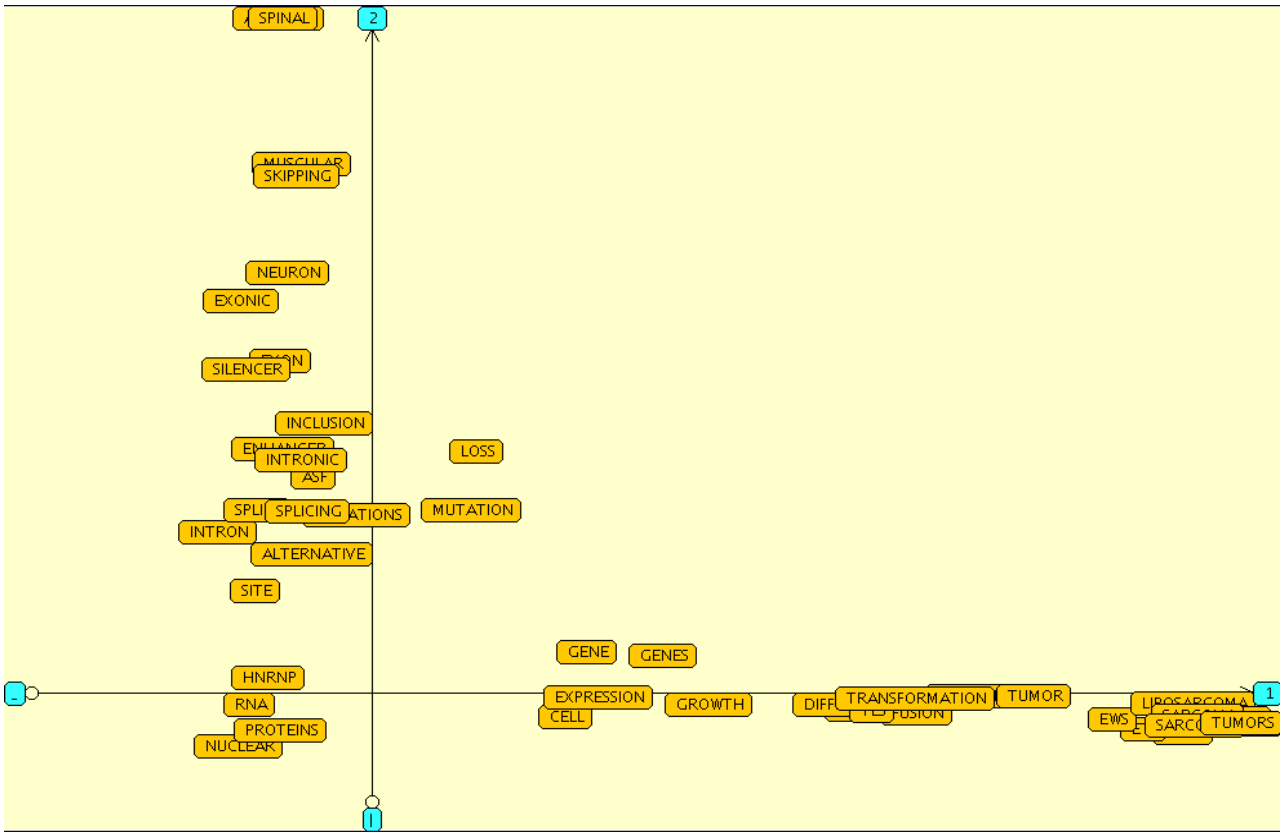
On peut dégager pour chaque métaclé un thème dominant :

- F 1P → Problèmes liés au dérèglement du cycle cellulaire (qui conduit au cancer)
- F 2P → Épissage alternatif
- F 3N → Traduction de l'ARN en protéine
- F 9N → Localisation cellulaire, adressage des protéines
- F 20N → Apoptose, mort cellulaire
- F 22P → Expériences
- F 28P → ARN et sa composition

C . Représentations bidimensionnelles des métaclés

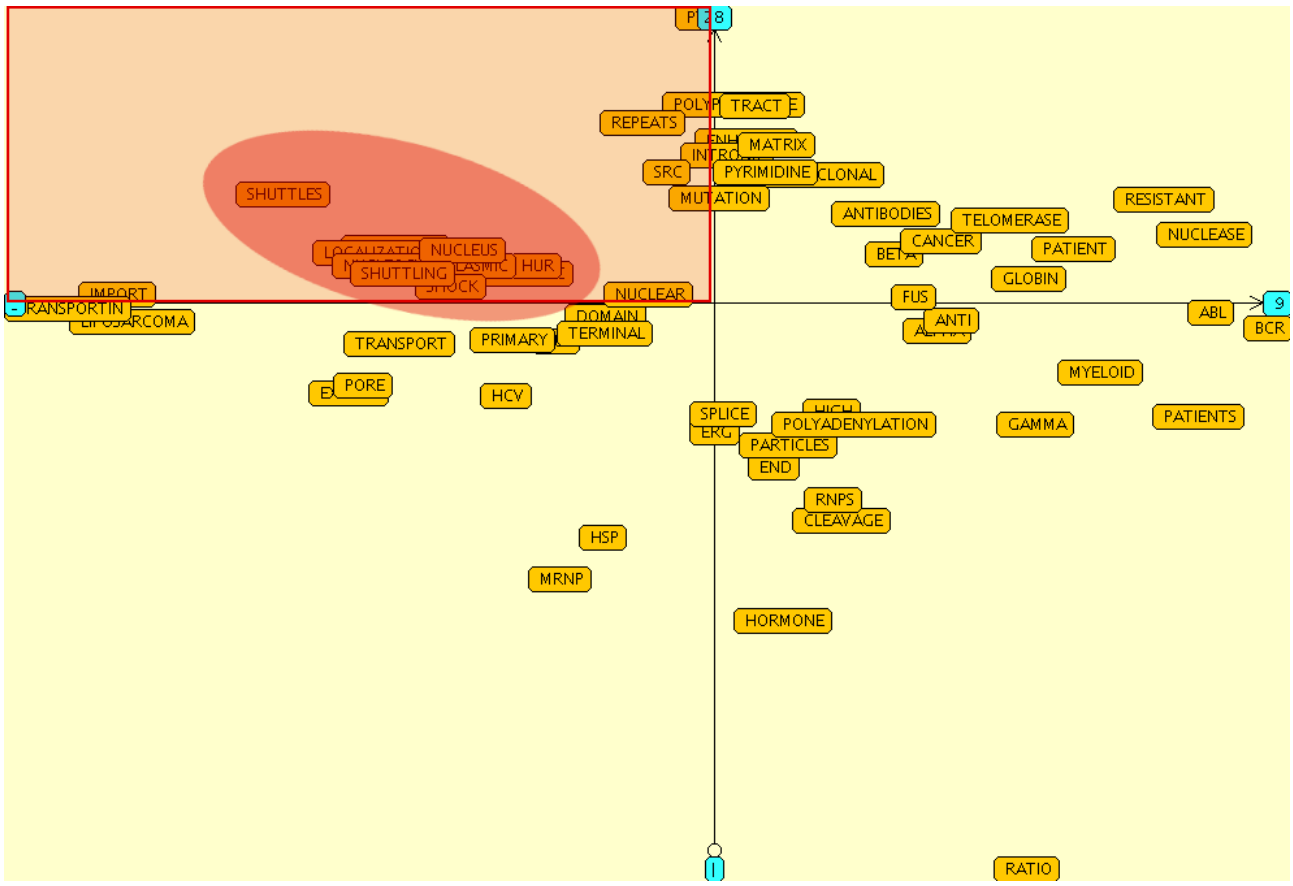
La représentation bidimentionnelle permet une superposition des métaclés et ainsi de mettre en évidence des associations de termes. Elle constitue un guide de lecture des sous-ensembles de textes déjà mis en évidence par les métaclés.

Nous avons représenté deux à deux chaque métaclé retenue précédemment et nous avons remarqué plusieurs profils différents :

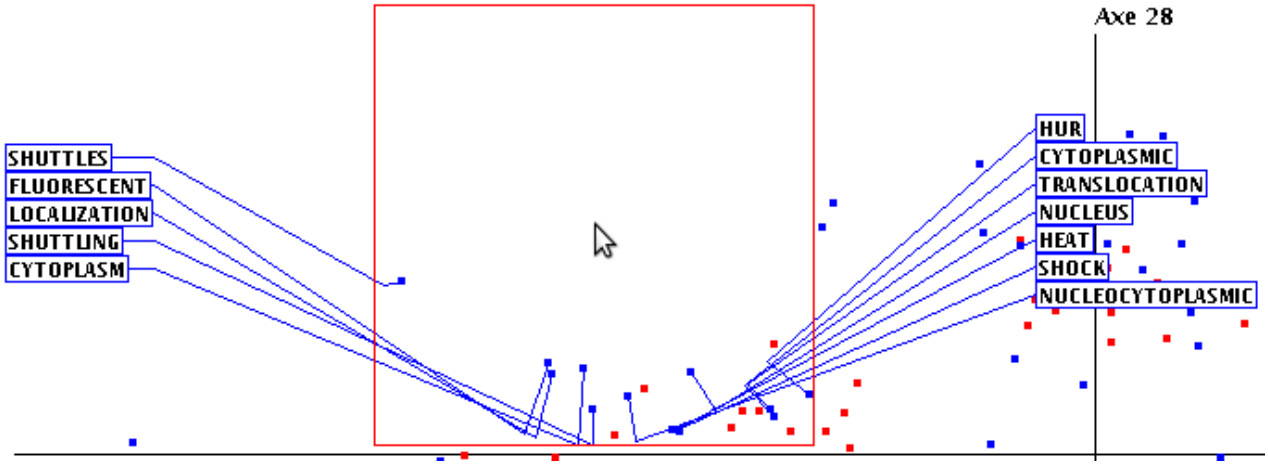


Métaclé F 2 en fonction de la Métaclé F 1

Aucun mots-clé en commun (à part mutation et loss). Les 2 métaclés dont indépendantes.



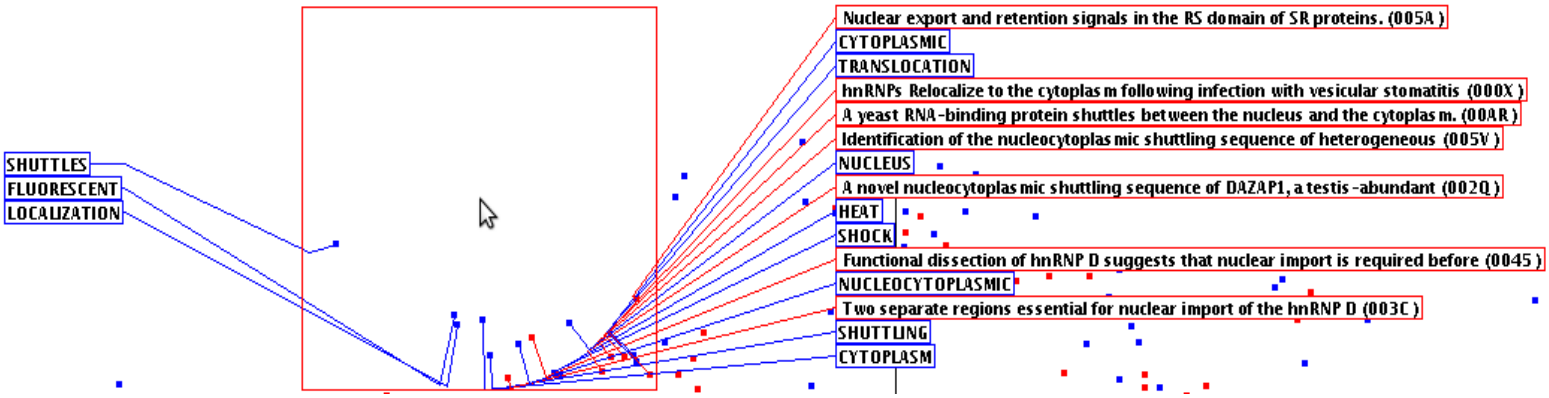
En rouge sur ce graphique, on retrouve la représentation des métaclés F 9N (Localisation cellulaire/adressage) et F 28P(ARN).



Représentation de ce même groupe de mots-clé avec Qnomis-3

On peut voir ici que les mots en communs son en rapport avec la localisation dans la cellule. Ce sous-ensemble doit dont traiter de la localisation de l'ARN dans la cellule, autrement dit, du passage de l'ARN du noyau au cytoplasme.

Si l'on souhaite en lire plus sur ce sujet, les publications partageant ces mots clés sont donné par le logiciel Qnomis-3 :



III. Conclusion

La combinaison des métaclés associées à une analyse par correspondance nous permet de mettre en évidence les différents thèmes abordés dans un corpus constitué d'une simple requête pubMed avec un mon de protéine peu parlant au premier abord. Si je n'avais pas de connaissance sur cette protéine au début de cette étude, j'aurais appris les choses suivantes :

La protéine hnRNP intervient dans :

- L'épissage, alternatif ou non
- Le passage de l'ARN du noyau au cytoplasme
- un certains nombre de pathologies telles que des cancers, notamment celui de la peau (sarcome), mais aussi dans la distrophie spinale

Si l'on regarde de plus près chaque métaclé, on y apprendra encore plus.

Cette étude a permis de mettre en évidence et ce de manière automatique des informations pertinentes sur un sujet bien précis, le tout à partir d'un corpus de 457 textes. Comme on l'a dit plus haut, l'analyse des données textuelles va permettre non pas de réduire le nombre d'articles à étudier mais d'en avoir une synthèse.