RAPPORT DU PROJET DATA MINING

Réalisé par:

Bellakhal Taher

Jaouadi Oussama

3 éme Génie Mathématiques Appliquées et Modélisation

Introduction générale:

La bio-informatique est une discipline relativement récente, le terme ayant été créé dans les années 80. Cette notion englobe l'ensemble des applications de l'informatique aux sciences de la vie, domaine très vaste qui recouvre tous les axes de recherche, allant des applications en robotique aux techniques les plus avancées en intelligence artificielle. Pour la plupart des membres de la communauté scientifique, cette notion semble dans la pratique s'adapter plus particulièrement aux outils informatiques qui permettent de stocker, d'analyser et de visualiser les informations contenues dans les séquences des gènes et des protéines des êtres vivants. L'histoire de la bioinformatique est donc étroitement liée à celle de la biologie moléculaire, l'étude des molécules du vivant.

Les protéines sont des macromolécules biologiques présentes dans toutes les cellules vivantes. Elles sont formées d'une ou de plusieurs chaînes polypeptidiques. Chacune de ces chaînes est constituée de l'enchaînement de résidus d'aminés liés entre eux par des liaisons peptidiques. Les protéines adoptent une structure en trois dimensions qui leur permet d'assurer leur fonction biologique. Cette structure particulière est déterminée avant tout par leur séquence en acides aminés dont les propriétés physico-chimiques divers conduit la chaîne protéique à adopter un repliement stable. Chaque protéine est caractérisée par un ensemble de séquences de données présentées sous forme des caractères alphabétiques.

I. Outils et méthodes utilisées

1. Data mining

- Extraction d'information intéressante (non triviale, implicite, non connue précédemment et potentiellement utile) ou de patterns.
- Découverte de connaissance (mining) dans des Bdd, extraction de connaissance, analyse de données/pattern.
- Propose des résumés d'information (rapports multidimensionnels, résumés statistiques).

2. Les méthodes utilisées

Dans cette partie on va présenter les méthodes d'apprentissage supervisé que nous avons utilisé dans notre travail :

Bootstrap 1
Train-test
SVM(cross-validation)

* Propriétés :

Pour une matrice de taille $N \times M$, on définit les propriétés suivantes :

- **Booléenne :** indique si l'élément x_i^j est présent dans la ligne i ou non.

$$\bar{w}_i^j = 1 \text{ si } x_i^j > 0, \text{ et } 0 \text{ si non.}$$

- **Occurrence :** indique le nombre d'occurrence de l'élément x_i^j dans la ligne i.

$$\bar{w}_i^j = x_i^j.$$

- **Fréquence :** indique le fréquence relative de l'élément x_i^j par rapport aux autres éléments qui compose la ligne i.

$$\bar{w}_i^j = \frac{x_i^j}{x_i^*}, \text{ où } x_i^* = \sum_{j=1}^{M} x_i^j.$$

- **TF*IDF :** indique le fréquence relative de l'élément x_i^j par rapport aux autres éléments qui compose la colonne j.

$$\bar{w}_i^j = x_i^j \log \frac{N}{x_*^j}, \text{ où } x_*^j = \sum_{i=1}^{N} x_i^j.$$

II. Les démarches du travail

I. La base de données

La base de données d'entrée utilisée dans ce travail est un ensemble de familles (classes) de protéines extraites regroupées dans un seul fichier ".txt". Pour chaque famille, chaque ligne, composée d'un nombre de caractères alphabétiques, représentant une séquence de protéine. La figure suivante montre une partie de la base de données choisie :

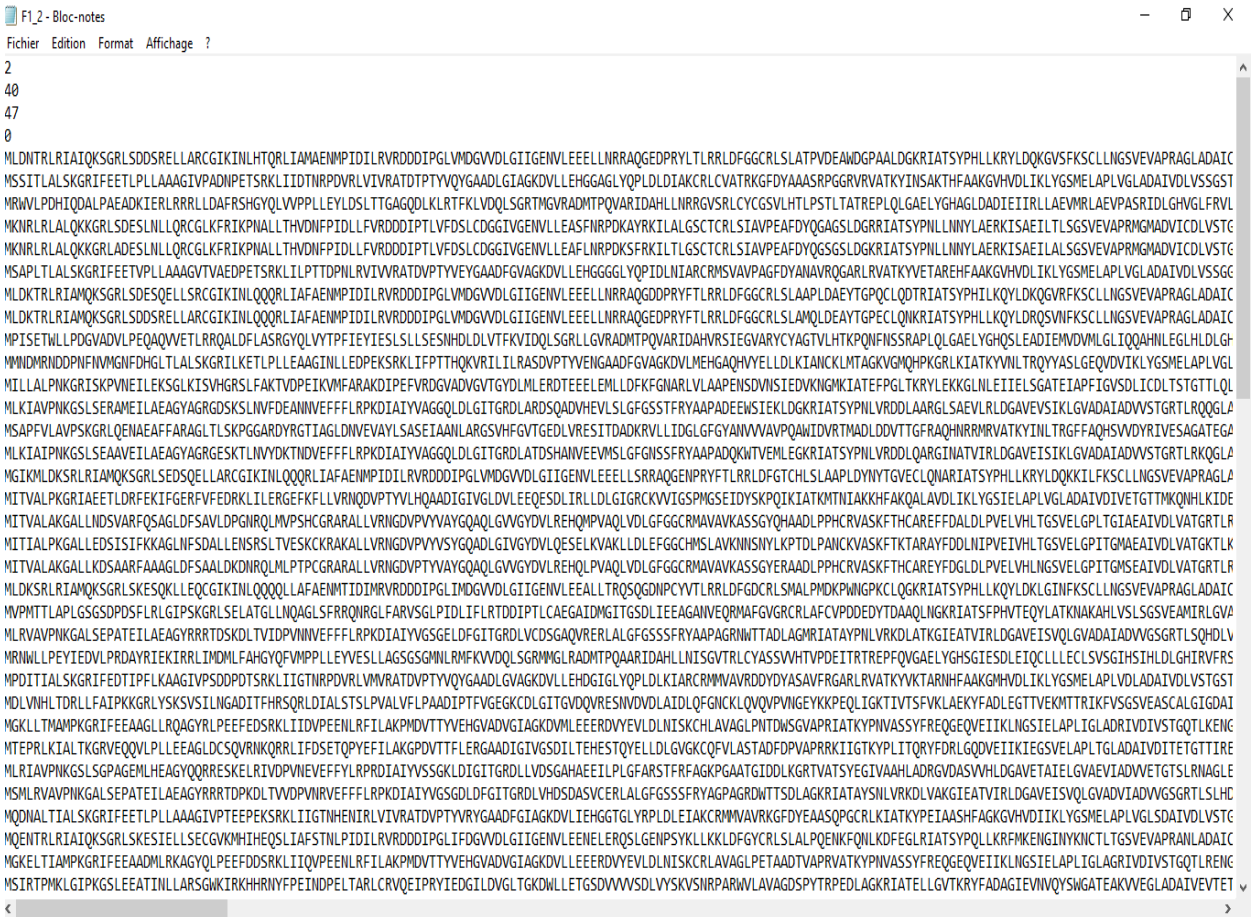

Figure1.base de données utilisée

2 éme étape : généreration des fichiers d’occurrence des n-grams pour n=2

DN	NT	TR	RL	LR	RI	IA	AI	IQ	QK	KS	SG	GR	LS	SD	DD	DS	SR	RE	EL	LL	LA	AR	RC	CG	GI	IK	KI	IN	NL	LH	HT	TQ	QR	LI	AM	MA
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	0	1	1	1	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	0	0	0	0	1	0	0
0	0	1	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	1	1	0	0	1	0	0
1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	1	1
1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	1	1
0	0	0	1	1	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	1	0	0	1	0	0	0	0	1	0	0
0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0
0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0
0	0	1	1	1	1	1	1	1	0	0	1	1	1	1	1	0	1	0	1	1	1	1	0	0	1	0	1	1	1	1	1	0	1	1	0	0
0	1	1	1	1	1	1	0	0	1	1	0	1	1	1	1	0	1	0	1	1	1	0	0	0	1	1	1	1	1	0	0	0	1	1	0	0
0	0	1	1	0	1	1	1	0	0	1	1	1	1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1	1	0	0	0	0	1	0	0
1	0	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0	1	1	0	0	0	0	1	0
1	0	1	1	1	1	1	0	0	0	0	1	1	1	1	1	0	1	1	0	1	1	1	0	0	0	0	0	1	1	1	0	0	0	1	0	1
1	0	0	1	1	1	1	1	0	1	0	1	1	1	0	1	1	0	1	1	0	1	1	0	0	1	1	1	1	1	1	0	0	0	0	1	0
0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0
0	0	0	1	0	1	1	1	0	0	0	0	1	0	1	0	0	0	1	1	1	1	0	1	0	1	1	1	0	0	1	0	0	1	1	0	0
1	0	0	1	1	1	1	1	1	0	0	1	1	0	0	0	1	0	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	1	1	1
0	0	0	1	1	0	1	1	0	0	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0	1	0	1	0	0	0	0	0	0	1	0	1
1	0	0	1	1	0	0	1	0	0	0	1	1	1	0	1	1	0	1	1	1	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1
1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	0	1	1	1	1
0	0	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	0	0	0	0	1	1	1	1
0	0	0	1	1	1	1	1	0	0	0	1	1	1	1	0	1	0	1	1	0	1	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0
0	0	1	1	1	1	1	0	1	0	1	1	1	1	1	0	1	0	1	1	1	1	1	0	0	1	1	1	0	1	0	1	0	0	1	1	0
0	1	0	1	1	1	1	1	1	0	0	0	1	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1	0	1	0	0	0	0	1	0	0
0	0	1	1	0	1	1	1	0	0	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	0	1	1	0	0
0	1	0	1	1	1	1	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1	0	0	0	0	1	1	1
0	0	1	1	0	0	1	1	0	0	0	0	1	0	1	0	1	0	1	1	1	1	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0
0	0	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0	0	0	1	0	0	1	0	1	0
0	0	0	1	1	1	1	1	0	0	0	1	1	1	1	0	1	0	0	1	0	1	1	0	0	1	0	0	1	1	1	0	0	0	1	0	0
1	1	0	1	0	1	1	1	1	0	1	0	1	1	1	0	0	1	0	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	1	0	0
0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0	1	1	0	1	1	1	1	0	0	0	1	0	0
0	0	0	1	1	1	1	0	1	0	0	0	1	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	0	1	0	0	0	0	1	1	1
1	1	1	1	1	1	1	1	0	0	0	1	0	0	1	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0	1	0	0
1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0
0	0	1	1	1	1	1	1	0	0	0	1	1	0	1	0	1	1	0	0	1	1	1	0	0	1	1	1	1	1	0	0	0	0	1	0	1
0	1	0	1	1	1	1	1	1	0	1	0	1	1	1	1	0	1	0	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	1	0	0

Figure2. Fichiers d’occurrence des 2-grams

3éme étape : Exemple d’application de la méthode ‘train-test’sur Tanagra

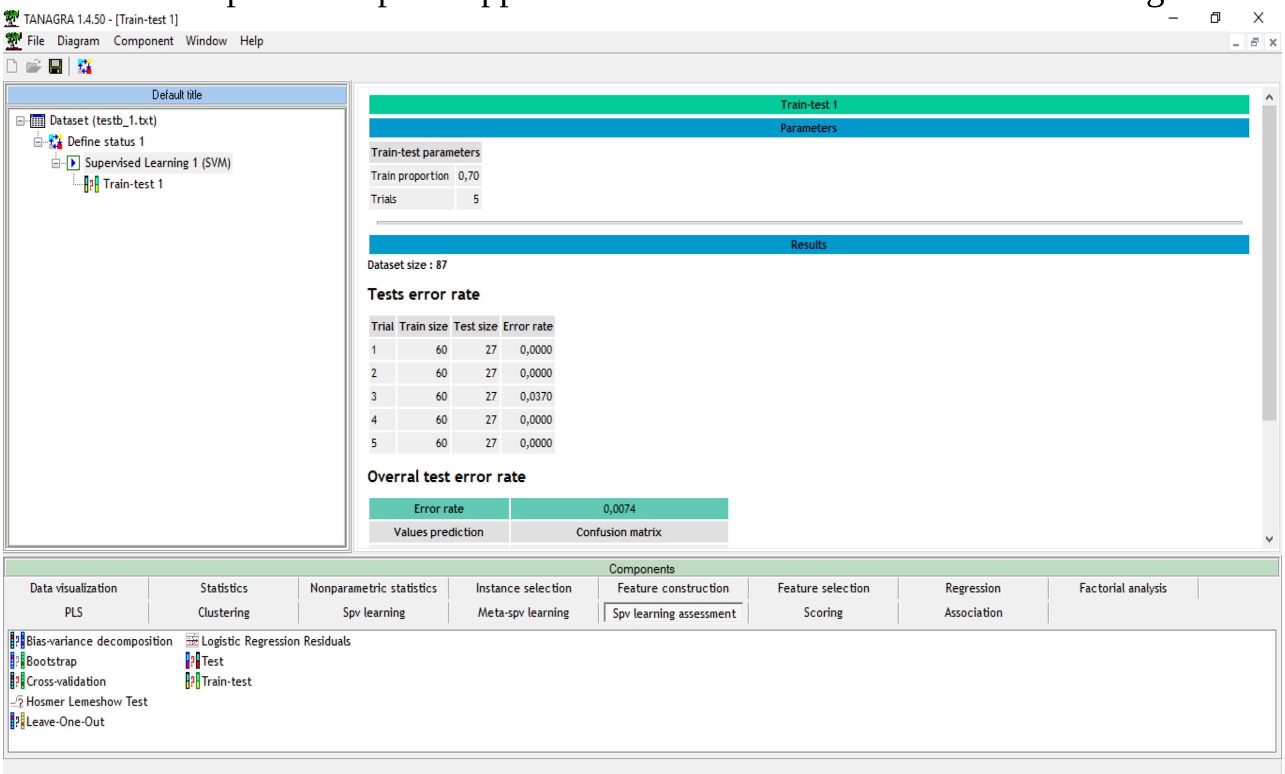

Figure3.application et exécution sur Tanagra

III. RESULTATS DES TESTS SUR TANAGRA

1. Pour les 2-grammes :

Pour la base des données booléenne :

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.0103	0.0160	0.0349	0.0038	0.0085	0.0147
Train-test	0.0370	0.0000	0.000	0.0370	0.0000	0.0148
SVM(cross-validation)	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118

Pour la base des données occurrence :

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.0103	0.0160	0.0038	0.0085	0.0123	0.01018
train-test	0.0000	0.0000	0.0370	0.0000	0.0000	0.0074
SVM(cross-validation)	0.0125	0.0125	0.0125	0.0125	0.0125	0.0125

Pour la base des données fréquence:

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0	0	0.025	0.025	0.025	0.015
Train-test	0	0	0.025	0.025	0.025	0.015
SVM(cross-validation)	0	0	0	0	0	0

Pour la base des données TF-IDF:

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.0698	0.0233	0.0349	0.0930	0.0465	0.0535
Train-test	0.5349	0.5349	0.5465	0.5465	0.5349	0.395
SVM(cross-validation)	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116

2. Pour les 3-grammes :

Pour la base des données booléenne :

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.0167	0.0234	0.0127	0.0134	0.0185	0.01694
Train-test	0.0349	0.233	0.0233	0.0233	0.0233	0.0256
SVM(cross-validation)	0.0125	0.0125	0.0125	0.0125	0.0125	0.0125

Pour la base des données occurrence :

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.0257	0.0358	0.0266	0.0220	0.0283	0.02768
Train-test	0.0000	0.0000	0.0370	0.0000	0.0000	0.0074
SVM(cross-validation)	0.0125	0.0125	0.0125	0.0125	0.0125	0.0125

Pour la base des données fréquence:

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0	0	0.025	0.025	0.025	0.015
Train-test	0	0	0.025	0.025	0.025	0.015
SVM(cross-validation)	0	0	0	0	0	0

Pour la base des données TF-IDF:

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.0698	0.0233	0.0349	0.0930	0.0465	0.0535
Train-test	0.5349	0.5349	0.5465	0.5465	0.5349	0.395
SVM(cross-validation)	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116

3. Pour les 4-grammes :

Pour la base des données booléenne :

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.1163	0.0233	0.0349	0.0465	0.0465	0.0535
Train-test	0.0349	0.233	0.0233	0.0233	0.0233	0.0256
SVM(cross-validation)	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116

Pour la base des données occurrence :

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.0698	0.0233	0.0349	0.0930	0.0465	0.0535
Train-test	0.4186	0.4186	0.407	0.3605	0.407	0.4023
SVM(cross-validation)	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116

Pour la base des données fréquence:

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0	0	0.025	0.025	0.025	0.015
Train-test	0	0	0.025	0.025	0.025	0.015
SVM(cross-validation)	0	0	0	0	0	0

Pour la base des données TF-IDF:

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.0698	0.0233	0.0349	0.0930	0.0465	0.0535
Train-test	0.5349	0.5349	0.5465	0.5465	0.5349	0.395
SVM(cross-validation)	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116

4. Pour les 5-grammes :

Pour la base des données booléenne :

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.1163	0.0233	0.0349	0.0465	0.0465	0.0535
Train-test	0.0349	0.233	0.0233	0.0233	0.0233	0.0256
SVM(cross-validation)	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116

Pour la base des données occurrence :

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.0698	0.0233	0.0349	0.0930	0.0465	0.0535
Train-test	0.4186	0.4186	0.407	0.3605	0.407	0.4023
SVM(cross-validation)	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116

Pour la base des données fréquence:

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0	0	0.025	0.025	0.025	0.015
Train-test	0	0	0.025	0.025	0.025	0.015
SVM(cross-validatio)	0	0	0	0	0	0

Pour la base des données TF-IDF:

méthodes	Erreur it1	Erreur it2	Erreur it3	Erreur it4	Erreur it5	Moy. erreur
Bootstrap 1	0.0698	0.0233	0.0349	0.0930	0.0465	0.0535
Train-test	0.5349	0.5349	0.5465	0.5465	0.5349	0.395
SVM(cross-validation)	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116

Conclusion

- On remarque bien que l'erreur pour chaque méthode d'apprentissage (bootstrap1, Train-test, SVM) varie selon le nombre de grammes et le type d'extraction (booléen,occurrence,...).
- Pour n = 2,3,4,5, on remarque la méthode SVM est la meilleure méthode d'apprentissage avec un taux d'erreur nul ou presque nul pour tous les types.
- On constate bien que la méthode SVM est la méthode dominante avec des erreurs nulles pour presque tous les types en particulier le type fréquence.