

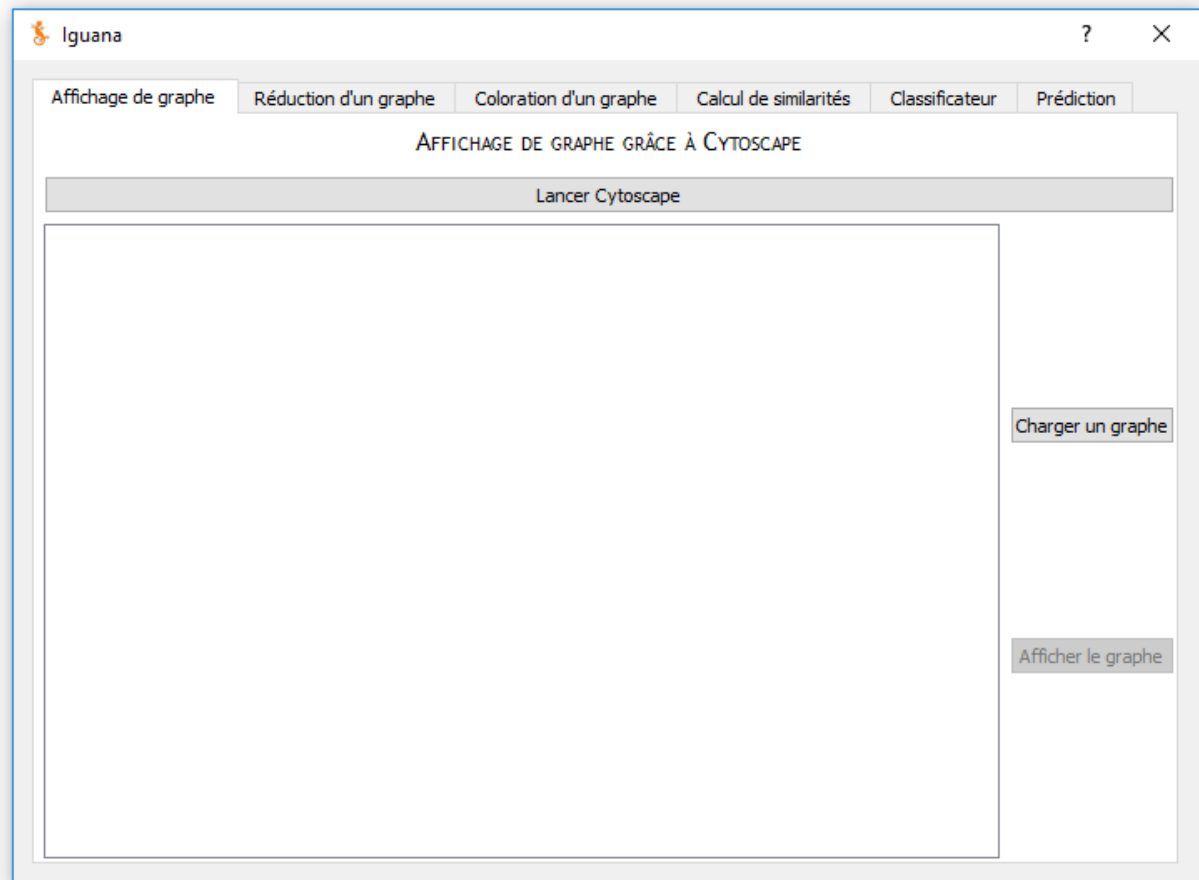
Guide d'utilisation Iguana



Auteurs :
Khalil Boulkenafet
Pierre Le Jeune
Jinhui Liu
Jules Paris
Justin Voïnéa

TABLE DES MATIERES

1.	Installation :	4
1.1	Installer Iguana	4
1.2	Installer Cytoscape	4
2.	Affichage d'un graphe	5
3.	Réduction d'un graphe	7
4.	Recherche de colorations	8
4.1	Identification des colorations	8
4.3	Afficher les différents composants dans Cytoscape	9
4.4	Export des n composantes	11
5.	Calcul de similarite	13
5.1	Fonctionnement	13
5.2	Réalisation dans Iguana	13
5.3	Options proposées par Iguana	15
5.3.1	Graphe de similarité	15
5.3.2	Création de données pour la prédiction	17
6.	Création d'un classificateur	18
6.1	Module de création	18
6.1.1	Fichier nécessaire à la création	18
6.1.2	Validation du modèle	18
6.2	Module de test	20
6.2.1	Fichier en entrée	20
6.2.2	Données résultantes	20
7.	Prédictions	22
8.	Aide interne à Iguana	23
9.	Conseils généraux	24



1. INSTALLATION :

1.1 Installer Iguana

Télécharger le fichier *Iguana_executable.zip* à partir du lien suivant :

https://mega.nz/#!4vhXWbrZ!J_RUe2IVf0x3x11wvAqWV9goKPGIFldKoWNbCKWLc1Y

Extraire le fichier dans un dossier et lancer le fichier *Iguana.exe*.

1.2 Installer Cytoscape

Télécharger Cytoscape sur ce site : <http://chianti.ucsd.edu/cytoscape-3.5.1/>.

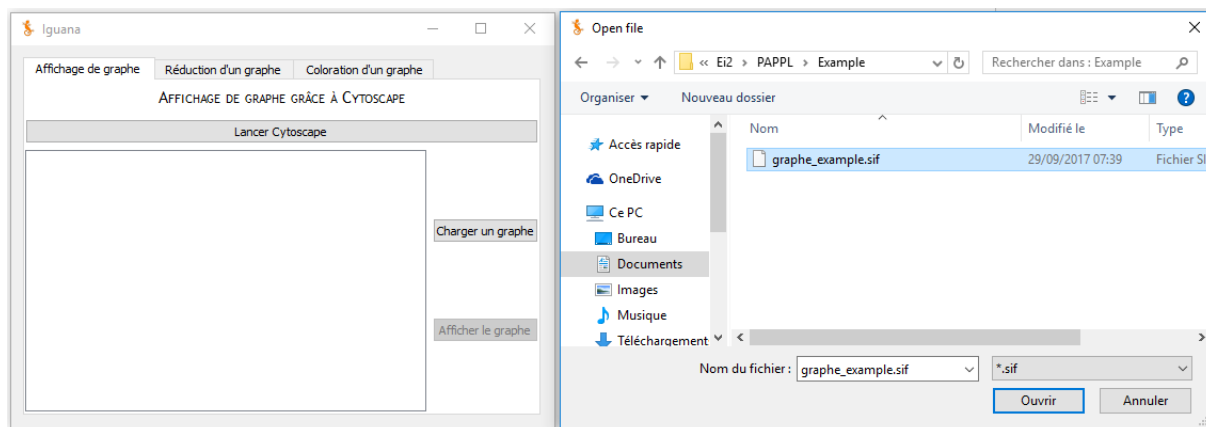
Prenez le fichier .exe de la version qui correspond à votre système d'exploitation (32 ou 64 bits).

Installer le logiciel en modifiant le répertoire d'installation par : **C:\Program Files\Cytoscape_v3.5.1**

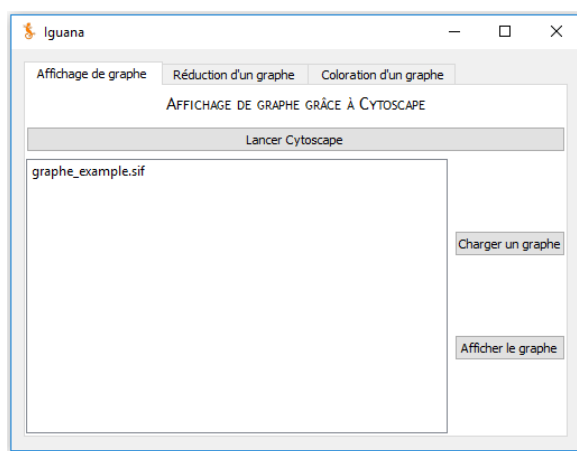
Une fois terminé, vous pouvez commencer à utiliser l'application.

2. AFFICHAGE D'UN GRAPHE

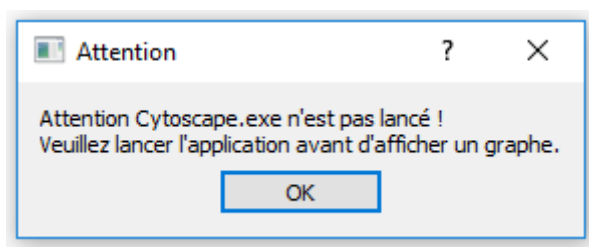
Pour d'afficher un graphe dans Cytoscape, il faut d'abord **charger un graphe** en allant le chercher dans vos dossiers.



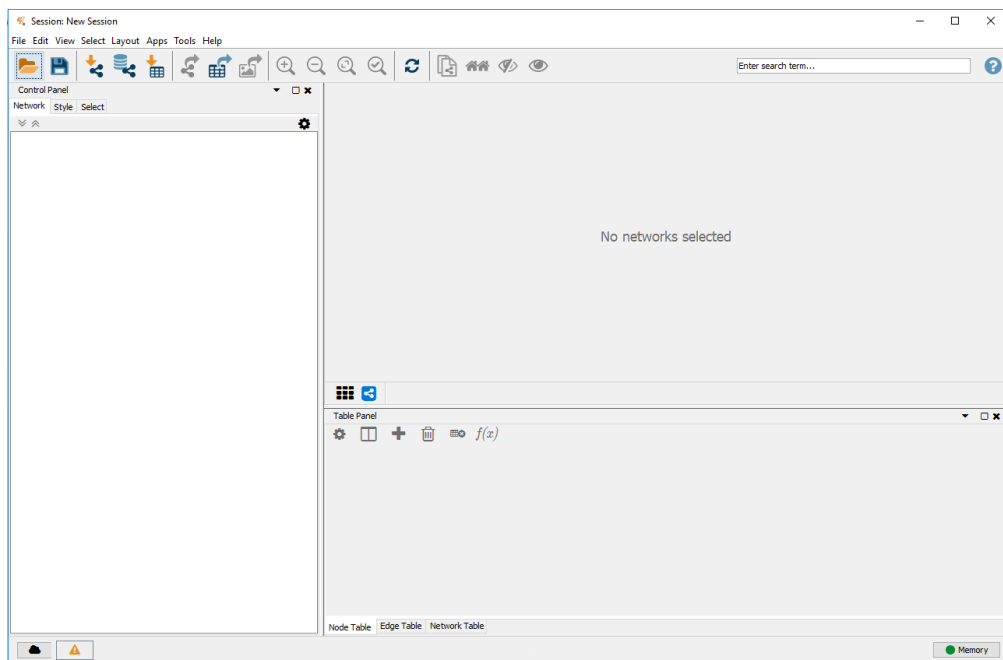
Une fois le graphe chargé, il est censé apparaitre dans la fenêtre principale.



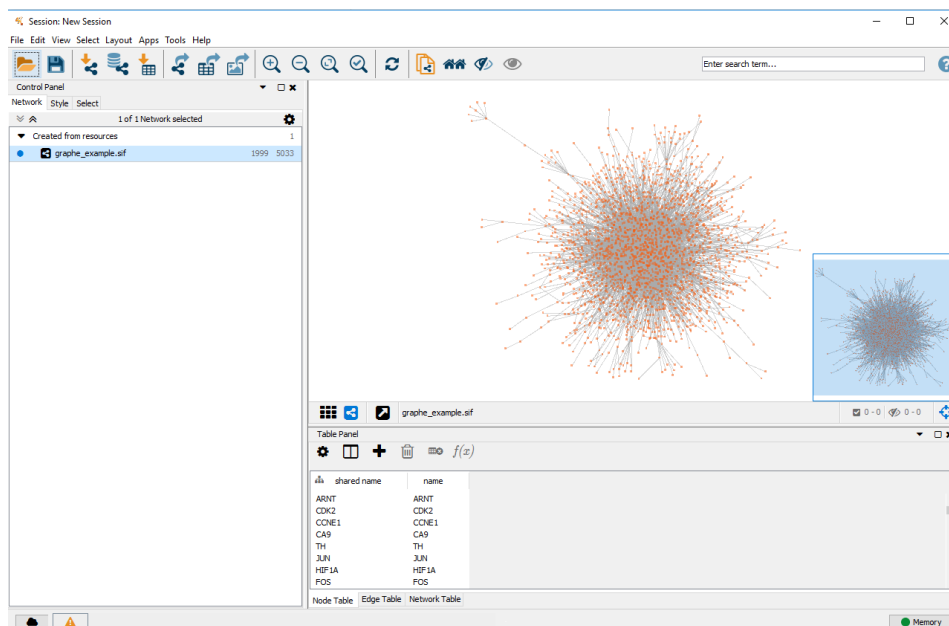
Il faut ensuite **cliquer sur le graphe à afficher** de sorte à le sélectionner puis cliquer sur le bouton **Afficher le graphe**. Cependant, il est nécessaire qu'une session Cytoscape soit ouverte sur votre ordinateur pour que le graphe soit affiché. Dans le cas contraire vous obtiendrez l'erreur suivante :



Vous pouvez cependant lancer Cytoscape depuis l'application en cliquant sur le bouton **Lancer Cytoscape**. Une fois lancée, la session de Cytoscape doit ressembler à celle-ci.

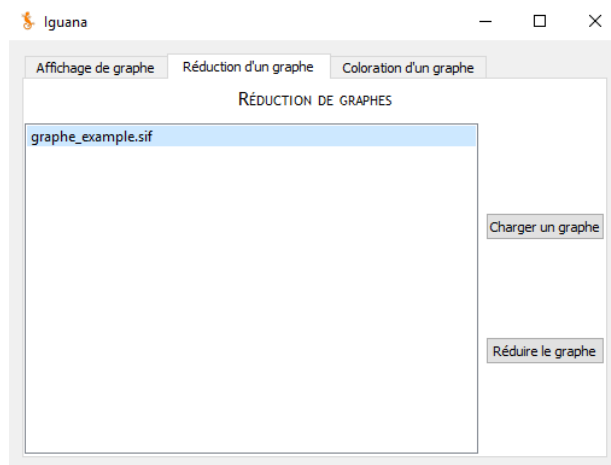


Le graphe s'affiche ensuite dans la session Cytoscape.



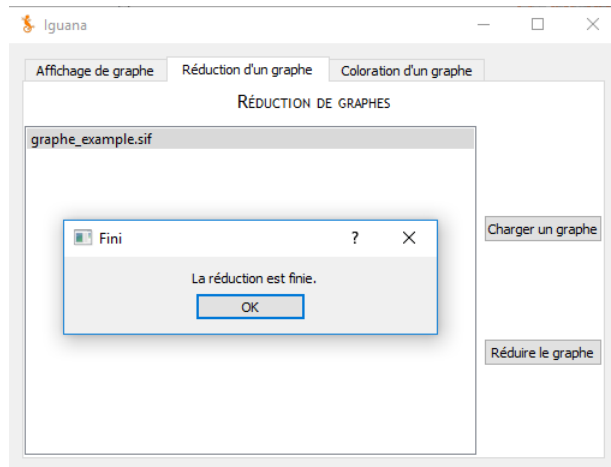
3. REDUCTION D'UN GRAPHE

Pour réduire un graphe, il faut passer dans le deuxième onglet, « Réduction d'un graphe ».







Si vous avez déjà chargé un graphe dans l'onglet d'affichage, celui-ci apparaîtra comme ci-dessus. Sinon reportez-vous au chargement d'un graphe dans le chapitre précédent.

Il ne vous reste plus qu'à **sélectionner un graphe** et à cliquer sur le bouton **Réduire le graphe** pour lancer la réduction. Un message vous signalera que l'opération est terminée.



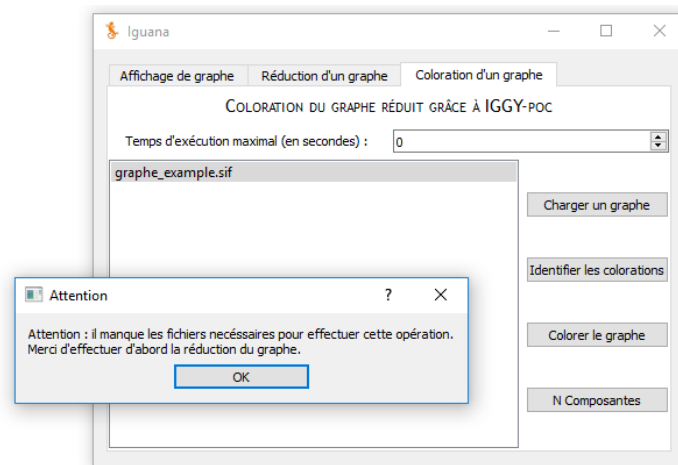
Vous pouvez vérifier que la réduction s'est bien passée en regardant les fichiers présents dans le dossier de votre graphe. Vous devez avoir trois nouveaux fichiers tel que suit.

Ce PC > Documents > Centrale > Ei2 > PAPPL > Exemple				
Nom	Modifié le	Type	Taille	
 graphe_exemple.sif	29/09/2017 07:39	Fichier SIF	66 Ko	
 graphe_exemple-reduced.sif	04/12/2017 15:21	Fichier SIF	405 Ko	
 graphe_exemple-reduced-hash.txt	04/12/2017 15:21	Document texte	29 Ko	
 graphe_exemple-reduced-logic.txt	04/12/2017 15:21	Document texte	42 Ko	

4. RECHERCHE DE COLORATIONS

4.1 Identification des colorations

Pour effectuer la recherche des colorations d'un graphe, il faut au préalable avoir effectué la réduction du graphe. En effet cette étape nécessite les fichiers qui sont générés par la réduction du graphe. Dans le cas contraire vous obtiendrez une erreur :



Si vous avez bien réalisé la réduction du graphe, il suffit de **sélectionner le graphe** dont vous voulez identifier les colorations et cliquer sur le bouton **Identifier les colorations**.

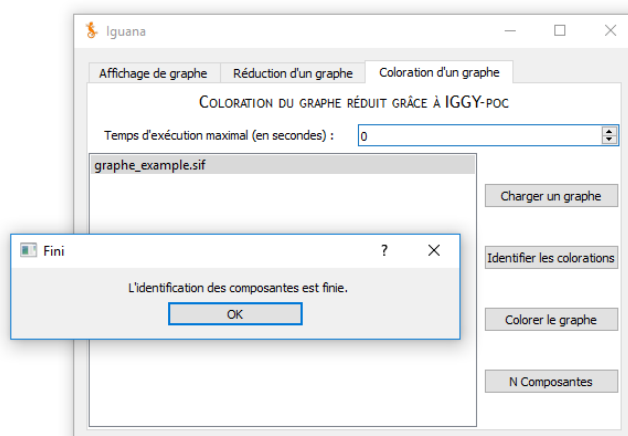
Option d'identification :

L'exécution de l'algorithme d'identification peut se révéler très long pour des graphes volumineux, c'est pourquoi vous pouvez régler le temps d'exécution de cette fonctionnalité. Pour cela, il suffit d'entrer le temps d'exécution souhaité (en seconde) dans le champ prévu à cet effet.








Temps d'exécution maximal (en secondes) :

En laissant 0, le programme s'exécutera jusqu'à l'obtention d'un optimum, ce qui n'est pas forcément le cas lors que vous entrez un temps fini.

Un message signale enfin la fin de l'identification.



De plus, à l'issue de l'identification de nouveaux fichiers sont ajoutés dans le dossier du graphe.

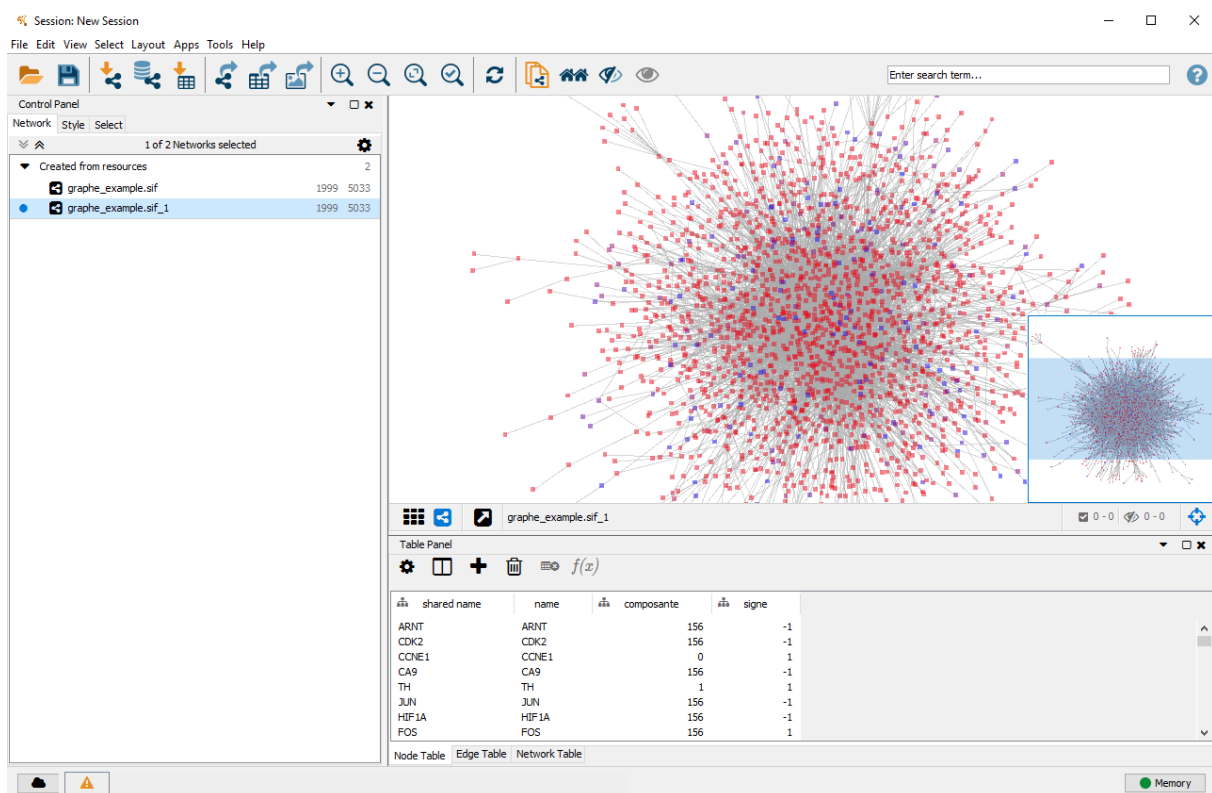
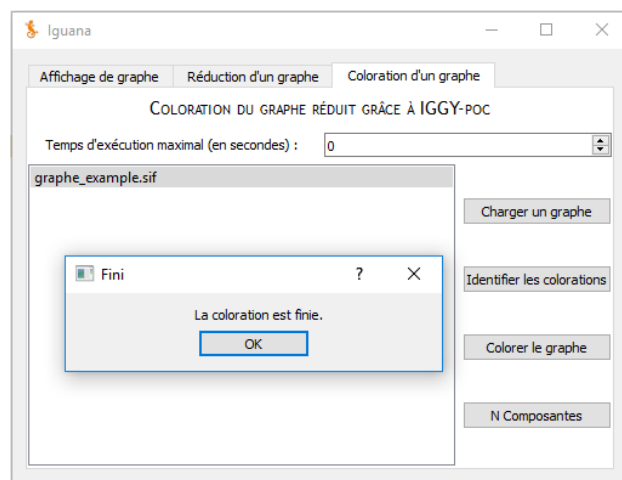
Ce PC > Documents > Centrale > Ei2 > PAPPL > Exemple				
Nom	Modifié le	Type	Taille	
 graphe_exemple.sif	29/09/2017 07:39	Fichier SIF	66 Ko	
 graphe_exemple-coloration-table.csv	04/12/2017 15:26	Fichier CSV Micro...	34 Ko	
 graphe_exemple-reduced.sif	04/12/2017 15:21	Fichier SIF	405 Ko	
 graphe_exemple-reduced-hash.txt	04/12/2017 15:21	Document texte	29 Ko	
 graphe_exemple-reduced-logic.txt	04/12/2017 15:21	Document texte	42 Ko	
 graphe_exemple-reduced-logic-colorations.txt	04/12/2017 15:26	Document texte	768 Ko	
 graphe_exemple-reduced-logic-colorations-processed.txt	04/12/2017 15:26	Document texte	40 Ko	

4.3 Afficher les différents composants dans Cytoscape

Pour lancer la coloration, il suffit de **sélectionner le graphe** puis de cliquer sur **Colorer le graphe**.

Cette fois, il faut avoir réalisé l'identification des colorations pour pouvoir colorer le graphe. Cependant aucun message d'erreur n'apparaîtra si ce n'est pas le cas. Le graphe sera simplement affiché dans Cytoscape (à condition qu'une session Cytoscape soit ouverte cf. *Affichage d'un graphe*) sans ses colorations.

Une fois terminée, un message apparaîtra et un nouveau graphe sera ajouté à la session Cytoscape.



Attention, chaque couleur dans le graphe représente un composant, il ne faut pas les confondre avec une coloration des nœuds du graphe, qui correspond à attribuer à chaque nœud la valeur +1 ou -1 pour activé ou désactivé. Une coloration est parfaite si elle est en accord avec les règles logiques qui lient les différents nœuds du graphe, comme cela est décrit dans la thèse de M. Miannay. Les composants eux, sont obtenus à partir d'une coloration et correspondent à un ensemble de nœuds qui « fonctionnent ensemble », c'est-à-dire que lorsqu'on passe l'activation de l'un d'eux, d'activée à inhibée, alors tous les autres nœuds de ce composant sont contraints de changer également leurs activations.

4.4 Export des n composantes

Pour l'export des composantes, il faut encore une fois **sélectionner le graphe** puis cliquer sur **N Composantes**. Une fois réalisé, un nouveau dossier « Composantes » sera créé, contenant n fichiers .sif représentant les graphes des composantes.

Ce PC > Documents > Centrale > Ei2 > PAPPL > Example

Nom	Modifié le	Type	Taille
Composantes	04/12/2017 15:31	Dossier de fichiers	
graphe_example.sif	29/09/2017 07:39	Fichier SIF	66 Ko
graphe_example-coloration-table.csv	04/12/2017 15:26	Fichier CSV Micro...	34 Ko
graphe_example-reduced.sif	04/12/2017 15:21	Fichier SIF	405 Ko
graphe_example-reduced-hash.txt	04/12/2017 15:21	Document texte	29 Ko
graphe_example-reduced-logic.txt	04/12/2017 15:21	Document texte	42 Ko
graphe_example-reduced-logic-colorations.txt	04/12/2017 15:26	Document texte	768 Ko
graphe_example-reduced-logic-colorations-processed.txt	04/12/2017 15:26	Document texte	40 Ko

Nom	Modifié le	Type	Taille
composante1.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante2.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante3.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante4.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante5.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante6.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante7.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante8.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante9.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante10.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante11.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante12.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante13.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante14.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante15.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante16.sif	04/12/2017 15:31	Fichier SIF	1 Ko
composante17.sif	04/12/2017 15:31	Fichier SIF	1 Ko

Ici encore, on a besoin des fichiers générés par l'identification des colorations, cependant, il n'y aura pas d'erreur en cas d'exécution sans ces fichiers, un fichier *composante1.sif* vide sera généré dans le dossier *Composantes* (créé si besoin).

Enfin, un message notifiera de la fin de l'exportation.

5. CALCUL DE SIMILARITE

5.1 Fonctionnement

Le calcul de similarité permet de relier la partie traitement des graphes à la partie classification des patients, qui est l'objectif final d'Iguana. En effet, le but de l'application est de pouvoir déterminer, à partir de l'expression de certains gènes, si un patient atteint par le myélome multiple est à risque ou non.

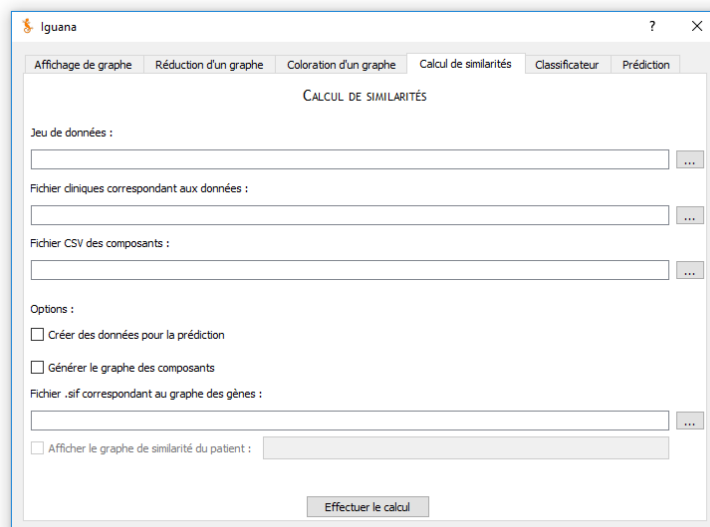
La première partie de l'application consistait à rechercher des structures (appelées composants) dans des graphes d'interactions entre les gènes. La seconde partie, elle consiste à calculer pour chacune de ces structures un « score » (entre 0 et 1) permettant de quantifier l'activation générale de chaque composant. On peut ensuite créer un classificateur à partir des données cliniques des patients et des scores de similarité. L'intérêt de travailler sur les composants est que l'on ne dispose pas nécessairement des données d'activation pour chaque gène, cependant, on a statistiquement des données au sein de chaque composants (le nombre de composants est très inférieur au nombre de gènes, de l'ordre de 1 pour 100).

Le calcul de similarité au sein de chaque composant revient à additionner les activations des gènes de chaque composant dont on a la valeur. Pour cela, on distingue deux cas :

- Gènes correspondant à un nœud positif : on ajoute simplement la valeur de l'activation du gène.
- Gènes correspondant à un nœud négatif : on ajoute $1 - a_i$, où a_i est l'activation du gène.

Par nœud négatif ou positif, on entend des nœuds qui sont activés ou inhibés dans la coloration, c'est-à-dire après l'étape 4.3 de ce tutoriel.

5.2 Réalisation dans Iguana

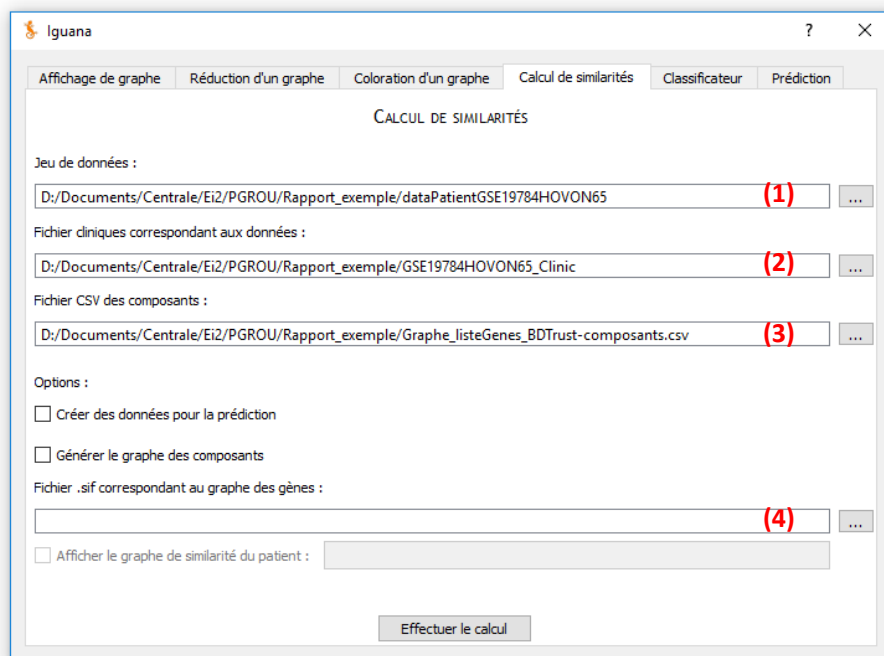


The screenshot shows the Iguana application window with the 'CALCUL DE SIMILARITÉS' tab selected. The interface includes the following elements:

- Navigation tabs:** Affichage de graphe, Réduction d'un graphe, Coloration d'un graphe, Calcul de similarités (selected), Classificateur, Prédiction.
- Section: CALCUL DE SIMILARITÉS**
 - Jeu de données :** A text input field with a dropdown arrow.
 - Fichier cliniques correspondant aux données :** A text input field with a dropdown arrow.
 - Fichier CSV des composants :** A text input field with a dropdown arrow.
 - Options :**
 - ☐ Créer des données pour la prédiction
 - ☐ Générer le graphe des composants
 - Fichier .sif correspondant au graphe des gènes :** A text input field with a dropdown arrow.
 - ☐ Afficher le graphe de similarité du patient : A text input field.
- Buttons:** 'Effectuer le calcul' at the bottom right.

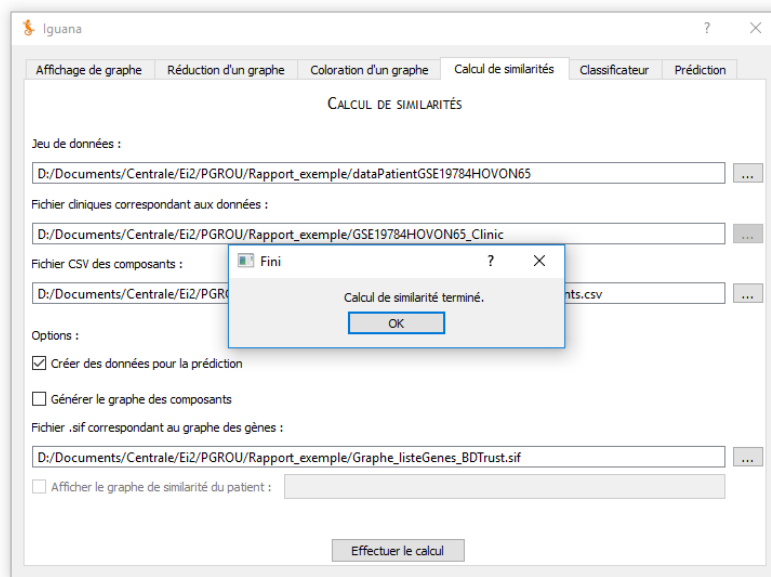
Tout d'abord, il faut aller chercher les différents fichiers nécessaires à ces opérations :

- Les fichiers contenant les données patients, ils contiennent les activations de chaque gène pour un patient. Il faut donc sélectionner le dossier contenant l'ensemble de fichiers patient. Il faut faire attention à ce qu'aucun autre fichier ne soit présent dans ce dossier. **(1)**
- Le fichier clinique : ce fichier contient le résultat clinique pour chaque patient (s'il s'agit d'un patient à risque ou non). Il est représenté par une liste associant le nom du patient avec le résultat. Ce fichier n'est pas nécessaire lorsque l'on souhaite réaliser le calcul de similarité dans le but de faire des prédictions sur l'état des patients, voir 5.3.2. **(2)**
- Le fichier des composants : il s'agit du fichier qui contient la liste des gènes/nœuds qui forment les différents composants. Ce fichier est un fichier .csv avec comme séparateur une tabulation. Il n'est nécessaire uniquement lorsque l'on veut afficher ou construire le graphe de similarité pour un patient voir 5.3.1. **(3)**
- Le graphe des gènes : il s'agit du graphe qui représente les interactions entre les gènes qui ont un lien avec la maladie. Ce fichier doit être un fichier .sif et correspondre avec le fichier des composants dont on a parlé juste au-dessus. Plus précisément, le fichier composant précédent doit avoir été créé à partir de ce graphe grâce au module « Coloration d'un graphe » d'Iguana. **(4)**



Une fois tous ces fichiers sélectionnés, il suffit de cliquer sur le bouton **Effectuer le calcul** ainsi que de sélectionner les options correspondantes avec ce que vous souhaitez faire.

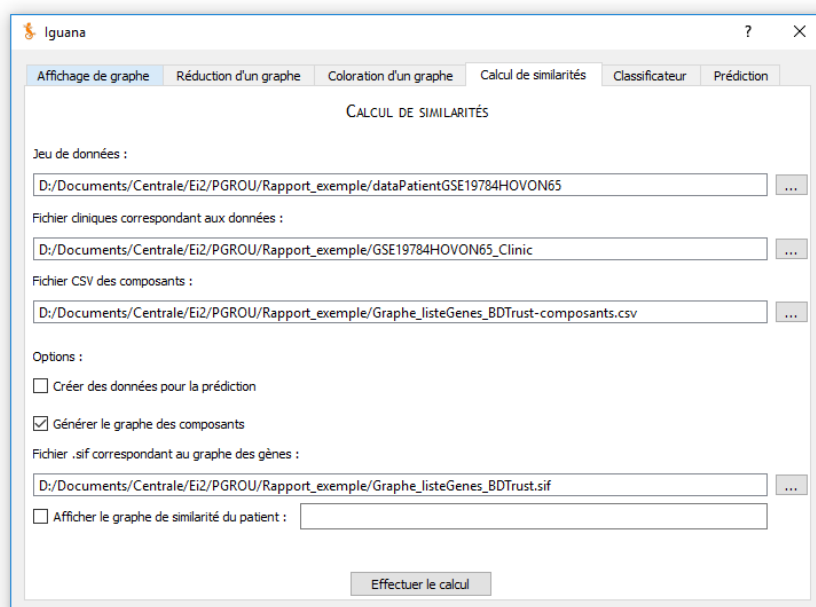
Un message vous signalera que le calcul est terminé ou s'il manque des fichiers.



5.3 Options proposées par Iguana

5.3.1 Graphe de similarité

La première option proposée pour le calcul de similarité est la création du graphe des composants ainsi que son affichage dans Cytoscape. Pour sélectionner ces options, il suffit de cocher les check box qui se trouvent en dessous des champs pour sélectionner les fichiers.



Le graphe des composants est un graphe qui représente l'interaction entre les différents composants. Chaque nœud représente un composant et sa couleur (entre le rouge et le vert) représente la valeur de la similarité du composant. Pour en faire l'affichage dans Cytoscape, il faut donner le nom d'un patient en le renseignant dans la boîte de dialogue à côté de l'option « Afficher le graphe de similarité du patient ».

☒ Générer le graphe des composants

Fichier .sif correspondant au graphe des gènes :

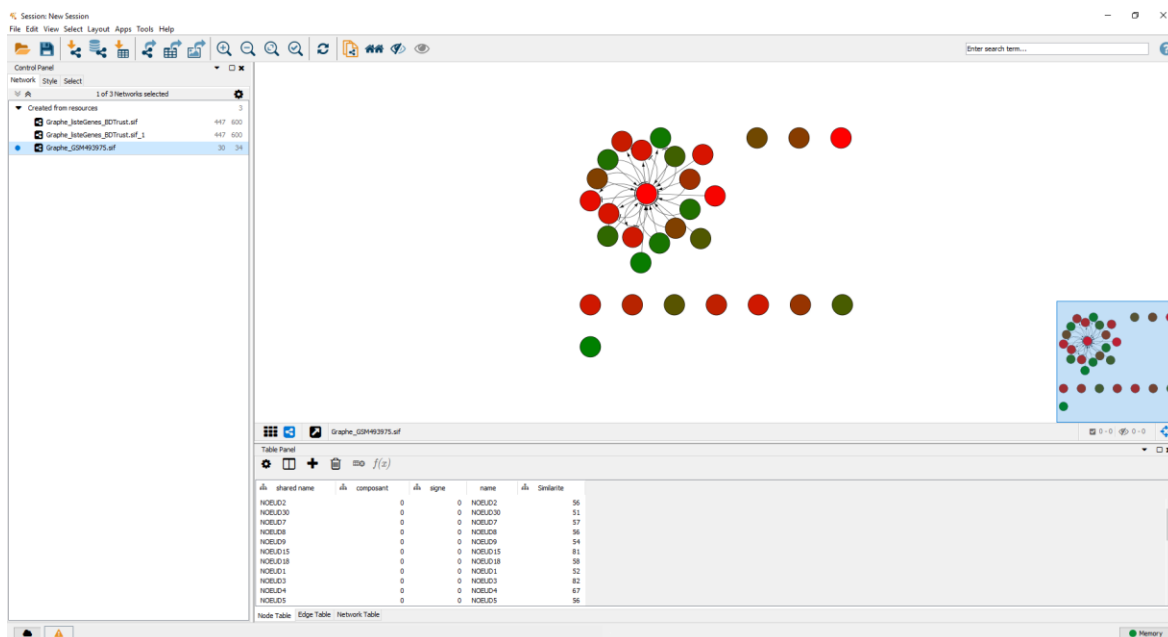
D:/Documents/Centrale/Ei2/PGROU/Rapport_exemple/Graphe_listeGenes_BDTrust.sif

☒ Afficher le graphe de similarité du patient : GSM493975

Le nom du patient est censé correspondre au nom du fichier d'un patient, comme suit.

Nom	Modifié le	Type	Taille
GSM493958	15/03/2018 17:28	Fichier	69 Ko
GSM493959	15/03/2018 17:28	Fichier	69 Ko
GSM493960	15/03/2018 17:28	Fichier	69 Ko
GSM493962	15/03/2018 17:28	Fichier	70 Ko
GSM493963	15/03/2018 17:28	Fichier	69 Ko
GSM493964	15/03/2018 17:28	Fichier	70 Ko
GSM493965	15/03/2018 17:28	Fichier	70 Ko
GSM493966	15/03/2018 17:28	Fichier	70 Ko
GSM493967	15/03/2018 17:28	Fichier	70 Ko
GSM493968	15/03/2018 17:28	Fichier	69 Ko
GSM493969	15/03/2018 17:28	Fichier	70 Ko
GSM493970	15/03/2018 17:28	Fichier	70 Ko
GSM493971	15/03/2018 17:28	Fichier	69 Ko
GSM493972	15/03/2018 17:28	Fichier	70 Ko
GSM493973	15/03/2018 17:28	Fichier	70 Ko
GSM493974	15/03/2018 17:28	Fichier	70 Ko
GSM493975	15/03/2018 17:28	Fichier	69 Ko

Une fois le calcul terminé, un fichier résultat est créé dans le dossier de départ et le graphe est affiché si cela était demandé.



5.3.2 Création de données pour la prédiction

Le fichier qui est créé par défaut par l'onglet de similarité est un fichier .csv qui contient sur chaque ligne :

- Le nom du patient.
- Les scores de chacun des composants pour le patient en question.
- Le résultat clinique du patient (TRUE ou FALSE)

Lorsqu'on veut réaliser des prédictions sur un (ou des) patient, on ne connaît évidemment pas le résultat clinique de celui-ci. Lorsque l'on coche l'option « Création de données pour la prédiction », le fichier créé ne contiendra pas la dernière colonne correspondant au résultat clinique. On pourra ainsi utiliser ce fichier pour faire des prédictions.

Options :

☒ Créer des données pour la prédiction

☐ Générer le graphe des composants

Fichier .sif correspondant au graphe des gènes :

D:/Documents/Centrale/Ei2/PGROU/Rapport_exemple/Grappe_listeGenes_BDTrust.sif



☐ Afficher le graphe de similarité du patient :

Lorsque l'option est décochée, le fichier est créé comme décrit plus haut et peut être utilisé soit pour créer un classificateur, soit pour tester un classificateur déjà existant, cf. 6. Dans ce cas, il n'est pas nécessaire de spécifier un fichier clinique, en revanche si un fichier est spécifié, il ne sera simplement pas pris en compte.

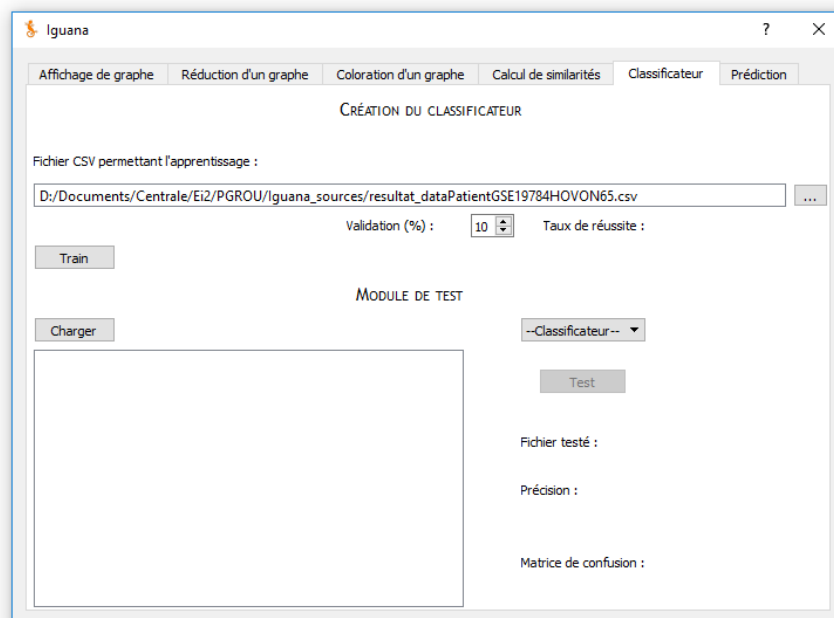
6. CREATION D'UN CLASSIFICATEUR

Le classificateur est basé sur un algorithme de Machine Learning basé sur des arbres de décision boostés (grâce à la librairie XGBOOST).

6.1 Module de création

6.1.1 Fichier nécessaire à la création

Afin de créer le classificateur, il suffit de sélectionner un fichier .csv qui a été créé par l'onglet « Calcul de similarité », avec l'option « création de données pour la prédiction » **désactivée**. Une fois le fichier sélectionné, il suffit de cliquer sur **Train** pour lancer la création du modèle.



6.1.2 Validation du modèle

Afin de valider le modèle, on va séparer le jeu de données en deux parties, une pour l'apprentissage et une pour la validation. On peut sélectionner le pourcentage de données utilisées pour la validation dans une plage entre 1% et 45%.

Pour la validation, nous effectuons un mélange aléatoire des données ainsi qu'un apprentissage et enfin un test du modèle. Nous répétons ces étapes 100 fois afin d'éviter le bruit statistique et nous faisons la moyenne des précisions obtenues.

Fichier CSV permettant l'apprentissage :

D:/Documents/Centrale/Ei2/PGROUP/Rapport_exemple/dataPatient.csv



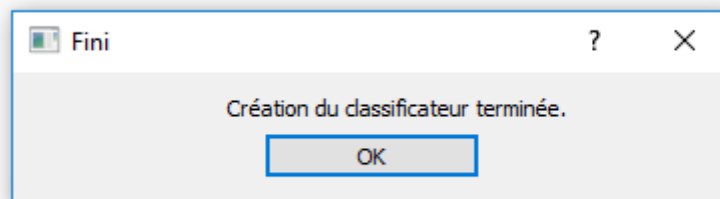
Validation (%) :

10

Taux de réussite : 78.2%

Train

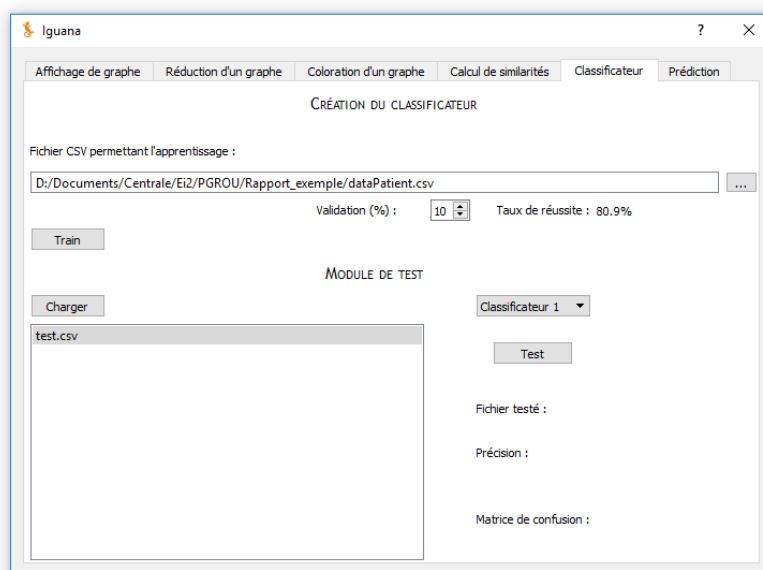
La création du modèle est également notifiée par une boîte de dialogue.



6.2 Module de test

6.2.1 Fichier en entrée

Dans ce module, on peut tester notre classificateur sur d'autres jeux de données. Pour cela, il suffit de **charger** un fichier de données (ce fichier a exactement la même structure que celui utilisé pour créer un classificateur. Ensuite, il faut sélectionner un fichier parmi les fichiers qui ont été chargés puis sélectionner le classificateur que l'on veut utiliser à l'aide de la liste déroulante (voir fig. x). Une fois cela effectué, il suffit de cliquer sur **Test** pour lancer le calcul.



6.2.2 Données résultantes

Fichier testé : test.csv

Précision : 0.88

	Positive	Negative
Matrice de confusion : Positive	469	67
Negative	1	21

Une fois que le test réalisé, Iguana va nous donner trois informations sur les résultats :

- Le nom du fichier testé.
- La précision du modèle sur ce jeu de données. Cela correspond au nombre de prédiction juste.
- La matrice de confusion. Elle présente les détails de la prédiction du modèle comme suit :

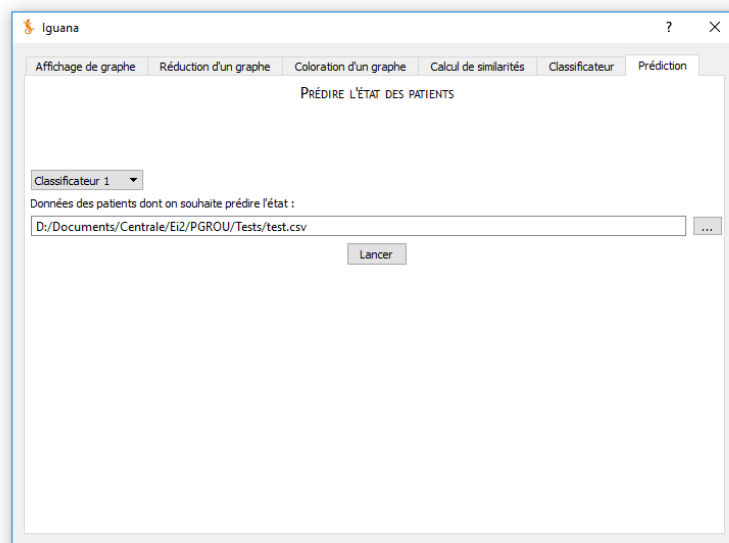
Nombre de prédiction "patient non à risque" dans le cas où le patient est réellement à risque	Nombre de prédiction "patient non à risque" dans le cas où le patient est à risque
Nombre de prédiction "patient à risque" dans le cas où le patient n'est pas à risque	Nombre de prédiction "patient à risque" dans le cas où le patient n'est réellement pas à risque

Figure 1 : Matrice de confusion

7. PREDICTIONS

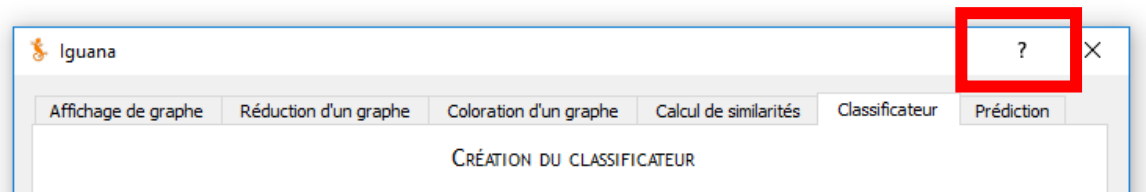
Pour effectuer les prédictions sur un ou plusieurs patients, il suffit de charger un fichier de données. Ce fichier doit être au format .csv et contenir autant de ligne que de patient et sur chaque ligne, la première colonne doit contenir le nom du patient et les colonnes suivantes les scores de similarité par composants. On peut construire des fichiers de cette forme via l'onglet « Calcul de similarité » en cochant l'option « Création de données pour la prédiction ».

Une fois le fichier chargé, il suffit de lancer la prédiction en cliquant sur **Lancer**.

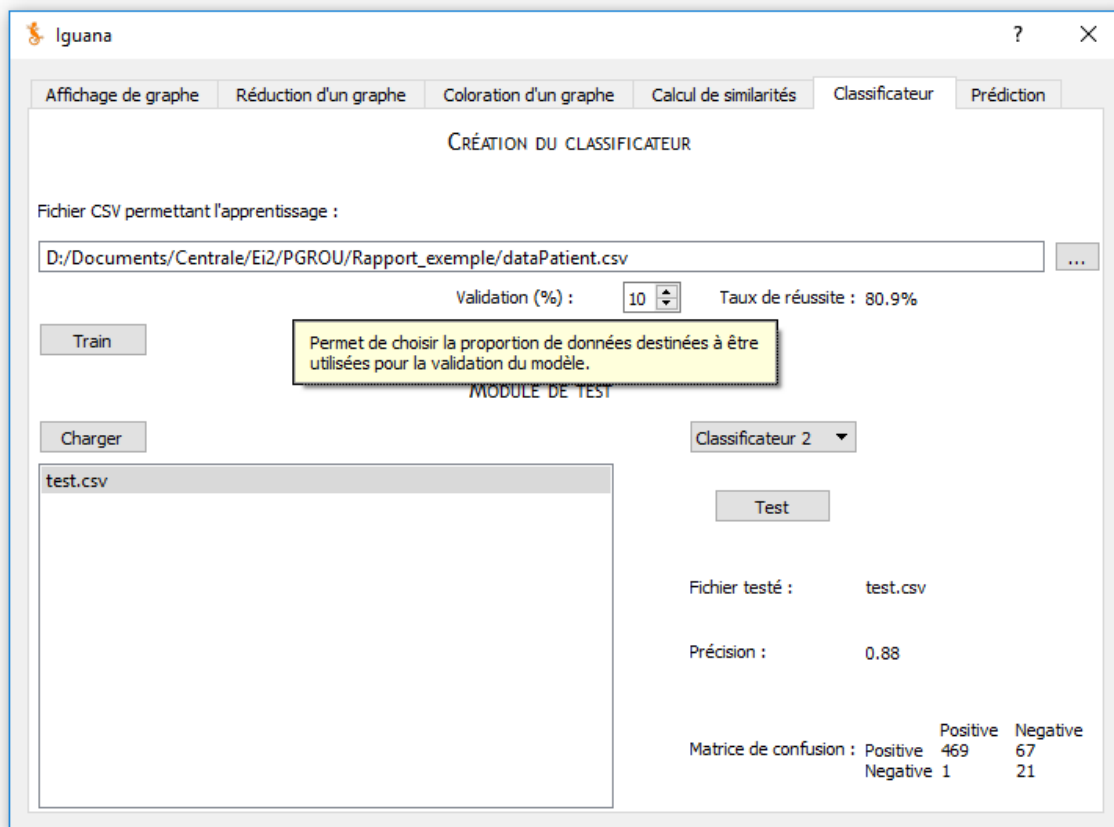


8. AIDE INTERNE A IGUANA

Un module d'aide est directement disponible à l'intérieur d'Iguana. Pour l'activer, il suffit de cliquer sur le **Point d'interrogation** en haut à droite de la fenêtre (voir ci-contre).



Une fois que l'on a cliqué dessus, on peut ensuite cliquer sur n'importe quel bouton ou champ pour faire apparaître une description succincte de la fonctionnalité.



9. CONSEILS GENERAUX

- Lorsque l'on ouvre un nombre important de graphe dans l'application, il faut faire attention à bien sélectionner le graphe sur lequel on veut effectuer une opération. En effet, certaines opérations modifient le graphe qui est sélectionné, il est donc nécessaire de le (re)sélectionner.
- Si l'exécution de l'identification des colorations prends trop de temps, vous pouvez l'arrêter manuellement en ouvrant le gestionnaire des tâches et en arrêtant le processus *clingo.exe*.
- Faites des dossiers séparés pour chaque graphe que vous voulez traiter de sorte à ne pas mélanger les fichiers générés par l'application.
- Ne modifiez pas les fichiers générés pendant le traitement ou entre les étapes du traitement.
- Vous pouvez modifier le style d'affichage des graphes directement dans Cytoscape, manuellement ou en important des styles personnalisés.
- Il se peut que les fichiers résultats du calcul de similarité s'écrivent les uns par-dessus les autres quand on lance l'algorithme à partir des mêmes données. Il est préférable d'en faire une sauvegarde au préalable.
- On ne peut pas ouvrir plusieurs graphes de similarité en même temps dans Cytoscape car la table utilisé est un attribut de la session Cytoscape et non du graphe en lui-même, ainsi les valeurs de similarité qui apparaitront dans la table du deuxième graphe seront celles du premier.
- Si les fichiers donnés à Iguana ne sont pas au bon format, l'application peut devenir instable. Si un tel comportement est observé, il convient de relancer Iguana.