

# ESIL 2<sup>ème</sup> Année

## Examen de Génomique et Bioinformatique

Mars 2004

Durée 2 heures  
Documents non autorisés

### Question 1. (6 pt)

Dans le cadre d'une analyse génétique, des chercheurs ont pu identifier un gène de prédisposition à une classe de lymphome. Ce gène présente des similitudes de séquence avec plusieurs autres gènes impliqués dans des voies de signalisation cellulaire, ainsi qu'avec plusieurs facteurs de transcription, mais son produit n'a pas encore été caractérisé expérimentalement. Votre tâche est de compenser ces lacunes en exploitant les données de génomique fonctionnelle suivantes :

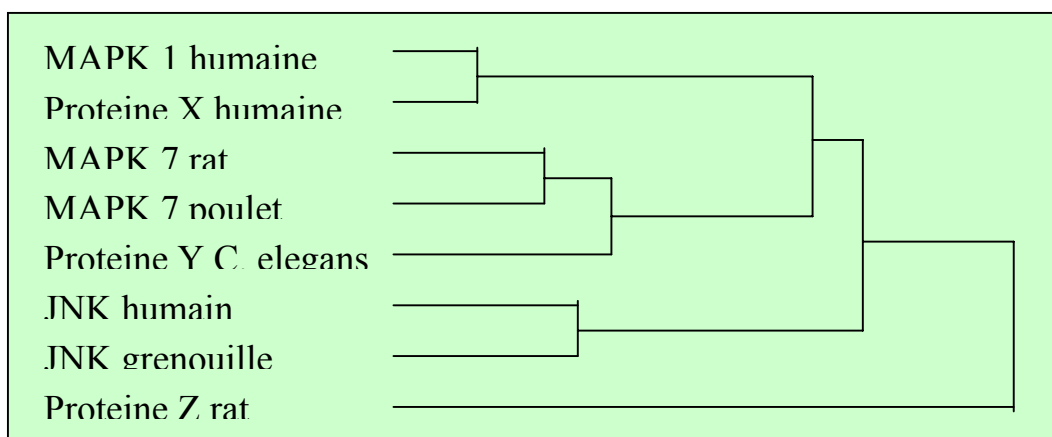
- des données de puces à ADN couvrant des milliers de gènes humains, y compris le gène identifié sur base génétique;
- des données de spectrométrie de masse sur une série de complexes cellulaires présents à l'intérieur des lymphocytes;
- des données ChIP-chip couvrant un grand nombre de régions promotrices dont l'activité est contrôlée spécifiquement dans les lymphocytes.

Décrivez brièvement comment vous envisageriez d'exploiter ces trois types de données pour préciser la fonction du produit du gène identifié sur base génétique, en insistant sur les complémentarités (avantages versus limitations respectives).

Le tout doit tenir dans une feuille d'examen (feuille séparée SVP).

### Question 2. (4 pt)

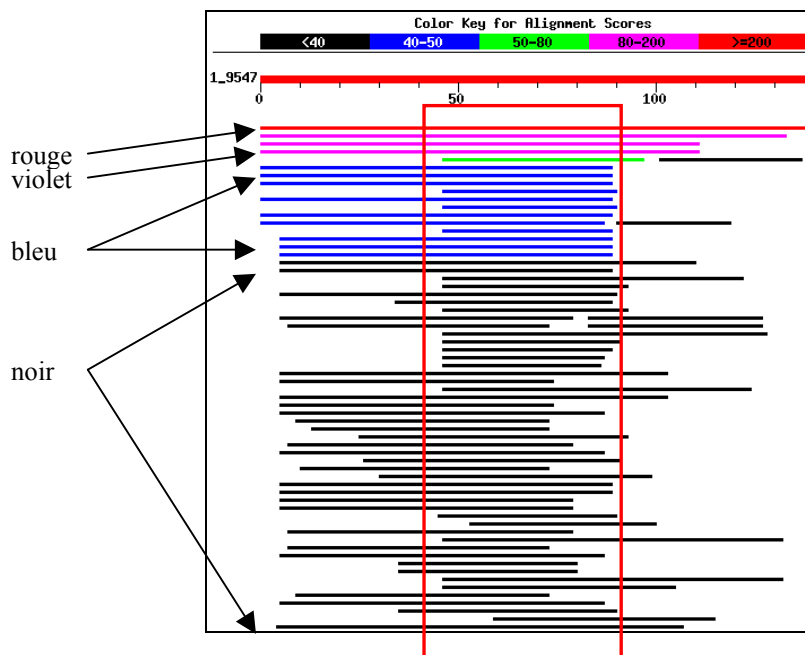
Les *stress-activated protein kinases* (JNK) et *mitogen-activated protein kinases* (MAPK) forment une famille de protéines caractérisées par un domaine MAPK commun. Il existe plusieurs MAPK (MAPK1, MAPK3, MAPK7, etc.) dans les génomes d'eucaryotes supérieurs. A partir d'une séquence de MAPK1 humaine, on récupère par Blast sept autres séquences qui, après alignement multiple, produisent l'arbre phylogénétique suivant :



A partir de leur position dans l'arbre et de l'espèce d'origine, faites une proposition de classification fonctionnelle pour chacune des séquences X, Y et Z.

### Question 3. (10 pt)

RNPA (ou C5) est le composant protéique de l'endoribonucléase RNase P. On réalise un Blast de la RNPA de la bactérie *Chlamydomophila caviae* (environ 130 aa) contre Swissprot. On obtient la sortie graphique suivante :



- Que signifient les couleurs des différentes lignes ?
- Les lignes du bas correspondent-elles forcément à des résultats non significatifs ? pourquoi ?
- Donnez une explication possible pour la fréquence plus élevée des lignes dans la région encadrée.

L'annexe A présente un sous-ensemble des alignements produits par Blast.

- Lesquels de ces alignements sont statistiquement significatifs ?
- Comment obtenir un arbre phylogénétique à partir de ces résultats ?
- Sur la seule base de l'annexe A, classez les séquences trouvées en orthologues, paralogues, xénologues, ou non-homologues.
- Dans quelle zone placeriez-vous la signature de l'activité RNPA ?
- A quoi correspondent les « + » entre les séquences et les « - » dans les séquences ?
- Imaginez que vous trouviez parmi les dernières solutions une protéine notée « protein of unknown function » dont la E-value est 2.3. Proposez deux moyens de vérifier s'il s'agit ou non d'une RNPA :
  - En regardant seulement l'alignement de Blast
  - Au moyen d'un serveur Web autre que Blast

## ANNEXE A

>[gi|21263939|sp|Q8XEW2|RNPA\\_SALTY](#) Ribonuclease P protein component (RNaseP protein) (RNase P protein) (Protein C5)  
Length = 119

Score = 45.4 bits (106), Expect = 3e-05  
Identities = 29/91 (31%), Positives = 48/91 (52%), Gaps = 2/91 (2%)

Query: 1 MHRSTLPKYARVLKRRKQFLYISRAGSHCQGSQVIFHVAPSRYSGCCCKLGITVSKK-FGKA 59  
M + P+ R+L F ++ + Q+ + G ++G+TV+KK +A  
Sbjct: 1 MVKLAFPRELRLRLTPAHFTFVFVQPPQRAGTPQITI-LGRLNSLGHPRIGLTVAKKNVRRRA 59

Query: 60 HKRNYFKRIVREAFRKKRHSPLACQIVVMPK 90  
H+RN KR+ RE+FR ++H LPA VV+ K  
Sbjct: 60 HERNRIKRLTRESFRLRQHELPAMDFVAVAK 90

../..

>[gi|13878718|sp|P57915|RNPA\\_PASMU](#) Ribonuclease P protein component (RNaseP protein) (RNase P protein) (Protein C5)  
Length = 119

Score = 43.5 bits (101), Expect = 1e-04  
Identities = 22/45 (48%), Positives = 31/45 (68%), Gaps = 1/45 (2%)

Query: 47 KLGITVSKK-FGKAHKRNYFKRIVREAFRKKRHSPLACQIVVMPK 90  
+LG+TV+KK KAH RN KR+ RE+FR ++ LP C V++ K  
Sbjct: 46 RLGLTVAKKHLKKAHDRNRIRKRLCRESFRLAQYKLPNCDFVIVAK 90

../..

>[gi|13878724|sp|Q9HT05|RNPA\\_PSEAE](#) Ribonuclease P protein component (RNaseP protein) (RNase P protein) (Protein C5)  
Length = 135

Score = 37.7 bits (86), Expect = 0.007  
Identities = 27/79 (34%), Positives = 44/79 (55%), Gaps = 7/79 (8%)

Query: 47 KLGITVSKKFGK-AHKRNYFKRIVREAFRKKRHSPLACQIVVMP-KNKQPKFEDLLQDF 104  
+LG+ + KK K A +RN KR++RE+FR + +L IVV+ K + + +L Q F  
Sbjct: 47 RLGLVIGKKNVKLAVQRNRLKRLIRESFRHNQETLAGWDIVVIARKGLGELENPELHQQF 106

Query: 105 AQQIPEALSSKLAKNKPTT 123  
+ L +L +N+P T  
Sbjct: 107 GK-----LWKRLLRNRPRT 120

../..

>[gi|3915853|sp|P50069|RNPA\\_BORBU](#) Ribonuclease P protein component (RNaseP protein) (RNase P protein) (Protein C5)  
Length = 119

Score = 29.6 bits (65), Expect = 1.8  
Identities = 13/27 (48%), Positives = 19/27 (70%)

Query: 50 ITVSKKFGKAHKRNYFKRIVREAFRKK 76  
+T SK F + KRN +R+ +EAFRK+  
Sbjct: 47 VTFSGGFRGSKRNRIRRLFKEAFRKR 73

>[gi|3913168|sp|P77989|BGAL\\_THEET](#) Beta-galactosidase (Lactase)  
Length = 743

Score = 28.9 bits (63), Expect = 3.1  
Identities = 14/37 (37%), Positives = 23/37 (62%), Gaps = 1/37 (2%)

Query: 102 QDFAQQIPEALSSKLAKNKPTTGVEYSPKNEKCESVL 138  
DF +++ E ++ KL +++PT GV Y +EK E V  
Sbjct: 393 HDFYKEMNE-IAHKLDRSRPTGGVRYLRDSEKLEDVF 428

>[gi|13878722|sp|Q9CJ73|RNPA\\_LACLA](#) Ribonuclease P protein component (RNaseP protein) (RNase P protein) (Protein C5)  
Length = 117

Score = 28.9 bits (63), Expect = 3.6  
Identities = 15/45 (33%), Positives = 27/45 (60%)

Query: 47 KLGITVSKKFGKAHKRNYFKRIVREAFRKKRHSPLACQIVVMPKN 91  
++G++VSKK G A RN KR++R A + + L V++ ++  
Sbjct: 43 RVGLSVSKKLGHAVLRNRIRKRLLRHAIAEFKPYLTDEDFVVIARS 87