

## Contrôle continu : Bioanalyse (3L5BC5M) – 23 octobre 2008

### Question 1

Définir en une phrase les concepts suivants :

- 1) homologie, 2) orthologie et 3) paralogie

### Question 2

- a) Expliquer la différence entre les banques de données SwissProt et TrEMBL
- b) Définir en quelques mots la banque de données OMIM
- c) Expliquer en quelques mots la technique de comparaison de deux séquences appelée matrice de points.
- d) Pour effectuer quel(s) type(s) d'analyse(s) utiliseriez-vous les programmes suivants :
  - 1) BlastP, 2) Dotmatcher de la suite EMBOSS, et 3) Water de la suite EMBOSS

### Question 3

- a) Utiliser la méthode de programmation dynamique pour déterminer l'alignement global optimal entre les deux séquences suivantes :

Séquence 1 : ATGGCTA

Séquence 2 : GGCG

Système de scores : identité = 0, substitution = +1, indel = +2 (Utilisation pour le calcul d'un score de distance)

Remplir la matrice de programmation dynamique et produire l'alignement final. Quel est le score de cet alignement ?

- b) Expliquer pourquoi la pénalité des indels doit être plus importante que la pénalité des substitutions.

### Question 4

- a) Vous voulez obtenir les protéines de *Emericella nidulans* codées par le gène *KapC*. Donner la requête que vous effectueriez sous le système d'interrogation Entrez. Pour vous aider, vous trouverez ci-dessous une partie de l'interface d'Entrez.
- b) Cette requête vous a permis d'obtenir la fiche donnée en Annexe 1 dont le champ séquence a été supprimé. Répondez aux questions suivantes :
  - à quel type d'organisme correspond *Emericella nidulans*?
  - dans quel journal scientifique les travaux concernant cette séquence ont-ils été publiés?
  - Donner la position et le nom des domaines protéiques rencontrés dans cette protéine ?
  - Quel est le motif présent sur cette séquence ?
  - Quelle est la fonction de cette protéine ?

### Question 5

La partie Features a été extraite d'une entrée provenant de la banque EMBL.

- de quel organisme est-elle issue ?
- quelles sont les positions des introns ?
- quelle est la fonction de la protéine codée par ce gène ?

```

FEATURES             Location/Qualifiers
     source            1..1093
                        /organism="Saccharomyces cerevisiae"
                        /mol_type="genomic DNA"
                        /strain="BMA41"
     gene              1..1093
                        /gene="TAD3"
     CDS                join(1..109,178..229,286..1093)
                        /gene="TAD3"
                        /note="Tad2p and Tad3p form a heterodimer
                        essential for cell viability"
                        /codon_start=1
                        /product="tRNA-specific adenosine-34 deaminase subunit
                        Tad3p/ADAT3"
                        /protein_id="CAB60630.1"
                        /db_xref="UniProtKB/Swiss-Prot:Q9URQ3"
                        /translation="MVKKVNNPLKIDYQNGI IENRLLQIRNFKDVNTPKLINVWSIRI
                        DPRDSKKVIELIRNDFQKNDPVSLRHLKRIKDIETSTLEVVLCSEYICDEGEINNK
                        LKSIWVGTKKYELSDIEVPEFAPSTKELNNAWSVKYWPLIWNPNQILNDYKIDM
                        QEVRNELSRASTLSVKMATAGKQFPMVSVFVDP SRKKDKVVAEDGRNCENSLPIDHSV
                        MVGIRAVGERLREGVDEDANSYLCLDYDVYLTHEPCSMCSMALIHSRVRVFLTEMQ
                        RTGSLKLTSGDGYCMNDNKQLNSTYEAFQWIGEEYPVGQVDRDVC"

```

## Question 6

Vous avez réalisé l'alignement suivant avec le programme stretcher de la suite EMBOSS.

- Quelle matrice de substitution a été utilisée ? Quels sont les pondérations utilisées pour les gaps (expliquer leur différence).
- Expliquer à quoi correspondent les différents pourcentages obtenus.
- Expliquer ce qui est représenté sur la ligne intermédiaire.

Aligned_sequences: 2	Length: 247
1: Q9CES4_LACLA	Identity: 137/247 (55.5%)
2: O34677_BACSU	Similarity: 182/247 (73.7%)
Matrix: EBLOSUM62	Gaps: 5/247 ( 2.0%)
Gap_penalty: 12	Score: 685
Extend_penalty: 2	

```

           10      20      30      40      50
Q9CES4  MGINTQIEVTDLHKSFGKNEVLKGITTKFEKGDVVCIIGPSGSGKSTFLR
      :   :   . . . . :   :   :   . . . . . :   :   :   :   :   :   :
O34677  M-----ITFQNVNKHYGDFHVLKQINLQIEKGEVVVIGPSGSGKSTLLR
           10      20      30      40
           60      70      80      90     100
Q9CES4  ALNGLETATSGDIIIDGFNLTDKNTNLNLVLRQNVGMVFQHFNLFPNMTVM
      . : . . : . . . : . . . . :   :   :   :   :   :   :   :   :
O34677  CINRLESINEGLTVNGTAINDRKTDINQVRQNIQGMVFQHFHLYPHKTVL
           50      60      70      80      90
           110     120     130     140     150
Q9CES4  QNITYAPVELKKMSKDDADKKAIQLLETVGLLDKKDAMPEMLSGGQQRV
      : : : : . . . . : . . . : : : : : : : : : : : : : : : : :
O34677  QNIMLAPVKVLRQSPEQAKETARYYLEKVGIPDKADAYPSQLSGGQQQRV
           100     110     120     130     140
           160     170     180     190     200
Q9CES4  AIARALAMNPVMLFDEPTSALDPEMVGDLAVMQKLAEEGTMMLIVTHE
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
O34677  AIARGLAMKPEVMLFDEPTSALDPEMIGEVLDMKTLAKEGMTMVVTHE
           150     160     170     180     190
           210     220     230     240
Q9CES4  MGFARKVANRVIFTDGGVILEDTPEELFDSPKHPRLQDFLSKVLNA
      : : : : : : : : : : : : : : : : : : : : : : : : : : :
O34677  MGFAKEVADRIVFIDEGKILEEAVPAEFYANPKEERARLFLSRILNH
           200     210     220     230     240
```

## Annexe 1

LOCUS P0C5H8 278 aa linear PLN 04-DEC-2007  
DEFINITION Putative transcription factor kapC.  
ACCESSION P0C5H8  
VERSION P0C5H8.1 GI:158705856  
DBSOURCE swissprot: locus KAPC\_EMENI, accession P0C5H8.1 release reviewed;  
KEYWORDS DNA-binding; Nucleus; Transcription; Transcription regulation.  
SOURCE Emericella nidulans (anamorph: Aspergillus nidulans)  
ORGANISM Emericella nidulans  
Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina;  
Eurotiomycetes; Eurotiomycetidae; Eurotiales; Trichocomaceae;  
Emericella.  
REFERENCE 1 (residues 1 to 278)  
AUTHORS Galagan, J.E et al.  
TITLE Sequencing of Aspergillus nidulans and comparative analysis with A.  
fumigatus and A. oryzae  
JOURNAL Nature 438 (7071), 1105-1115 (2005)  
PUBMED 16372000  
REMARK NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].  
STRAIN=FGSC 4  
COMMENT On Oct 25, 2007 this sequence version replaced gi:74596856.  
[FUNCTION] Putative transcription factor (By similarity).  
[SUBCELLULAR LOCATION] Nucleus (By similarity).  
[SIMILARITY] Belongs to the bZIP family.  
[SEQUENCE CAUTION] Sequence=EAA62953.1; Type=Erroneous gene model  
prediction; Note=The predicted gene AN3189 has been split into 2  
genes: AN10378 and AN10380.  
FEATURES  
Location/Qualifiers  
source 1..278  
/organism="Emericella nidulans"  
/db\_xref="taxon:162425"  
gene 1..278  
/gene="kapC"  
/locus\_tag="AN10378"  
Protein 1..278  
/gene="kapC"  
/locus\_tag="AN10378"  
/product="Putative transcription factor kapC"  
Region 1..278  
/gene="kapC"  
/locus\_tag="AN10378"  
/region\_name="Mature chain"  
/experiment="experimental evidence, no additional details  
recorded"  
/note="Putative transcription factor kapC."  
/FTId=PRO\_0000306819."  
Site 104..127  
/gene="kapC"  
/locus\_tag="AN10378"  
/site\_type="DNA binding"  
/experiment="experimental evidence, no additional details  
recorded"  
/note="Basic motif."  
Region 131..162  
/gene="kapC"  
/locus\_tag="AN10378"  
/region\_name="Domain"  
/experiment="experimental evidence, no additional details  
recorded"  
/note="Leucine-zipper."