

TD : Annotation d'une séquence de la famille Sigma-70

Rappel de Biologie: les facteurs sigma sont des protéines du complexe d'initiation de la transcription. Ce sont les facteurs sigma qui reconnaissent le promoteur et recrutent le "core enzyme" de la RNA polymérase composé des facteurs alpha, beta et gamma. Il y a plusieurs sortes de facteurs sigma, qui reconnaissent des promoteurs différents grâce notamment à la région de fixation à l'ADN en position -35 qui est différente pour chaque facteur sigma (Figure 1).

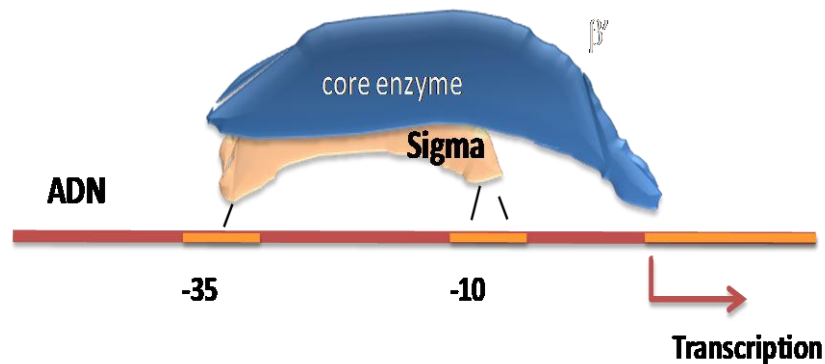


Figure 1: schéma de fonctionnement du Facteur sigma

On cherche à annoter un gène d'E. coli codant pour une protéine de 613 aa que nous appellerons protéine "unknown".

On analyse la constitution en domaines de "unknown" au moyen du site Interpro. On obtient le résultat suivant.

Banque	Domaine	Nom	E-value	position
PFAM	PF03979	Sigma70_r1_1	3.6e-30	[2-80]
PFAM	PF04542	Sigma70_r2	4.6e-23	[379-449]
PFAM	PF04539	Sigma70_r3	6.1e-29	[458-534]

Les informations trouvées dans PFAM sur chacun des 3 domaines sont les suivantes:

Family: Sigma70_r1_1 (PF03979)
[...] Region 1.1 is also involved in promoter binding

Family: Sigma70_r2 (PF04542)
Region 2 of sigma-70 is the most conserved region of the entire protein. All members of this class of sigma-factor contain region 2.
The high conservation is due to region 2 containing both the -10 promoter recognition helix and the primary core RNA polymerase binding determinant. The core binding helix, interacts with [...] the largest polymerase subunit, beta prime

Family: Sigma70_r3 (PF04539)
Region 3 forms a discrete compact three helical domain within the sigma-factor. [...] Region 3 primarily is involved in binding the core RNA polymerase in the holoenzyme.

Q1: Interprétez les résultats obtenus avec Interpro

On réalise un Blast de la séquence protéique contre Swissprot. On obtient 131 solutions de E-value <10, dont principalement des gènes des familles RpoS, RpoD, Rpo32 et RpoF.

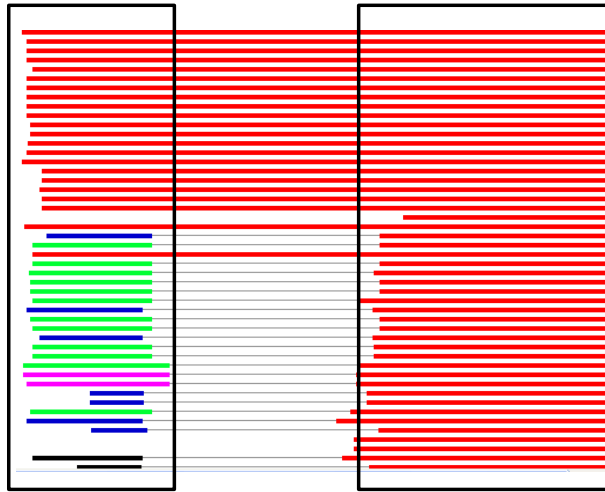


Figure 2. schéma des alignements Blast (on ne voit ici que des alignements de E-value < 10e-10):

Q2: A partir de la figure 2, sachant que le bloc conservé de droite est environ situé entre les positions 375 et 600 et le bloc moins conservé de gauche entre les positions 10 et 130, et en vous basant sur l'analyse des domaines, faites une hypothèse sur les fonctions des protéines trouvées par Blast.

On récupère les séquences d'une quinzaine de protéines ayant toutes une E-value Blast inférieure à 10e-10. Leurs noms sont les suivants :

RPOF_STRCO	RPOF_STRAU	RP32_PSEAE	RP32_PROMI	RP32_SERMA
RP32_ECOLI	RP32_CITFR	RPOD_LEPIN	RPOD_BUCBP	RPOD_BUCAI
RPOD_SALTY	RPOS_COXBU	RPOS_VIBCH	RPOS_SALTY	RPOS_ECOLI

Q3: Sans autre information que les E-values, les noms des protéines et la Figure 2, faites une hypothèse raisonnable quant à l'homologie/paralogie/orthologie des séquences sélectionnées.

On réalise un alignement multiple et un arbre phylogénétique des protéines sélectionnées à l'aide du serveur phylogeny.fr (options par défaut). On obtient l'alignement et l'arbre ci-dessous.

Q4: En analysant l'alignement multiple et en connaissant les différents domaines fonctionnels de la protéine, peut-on confirmer les hypothèses ci-dessus?

Q5: L'arbre phylogénétique confirme-t-il vos hypothèses? Comment?

CLUSTAL FORMAT: MUSCLE (3.7) multiple sequence alignment

```
RPOF_STRCO -----MPASTAFQAPPAPPAQAQA--
RPOF_STRAU -----MTVPASTAFQVFPQDPQVPHPQ
RP32_PSEAE -----MTTSLQPVHALVP-----
RP32_PROMI -----MTQEMQSLALVPQ-----
RP32_SERMA -----MTKEMQTLALVPQ-----
RP32_ECOLI -----MTDKMQSLALAPV-----
RP32_CITFR -----MTKEMQNLALAPV-----
RPOD_LEPIN MENLQSMPEVQKIISLGKANGEVSYDDINEILPDKILNSEKIDDFFTLLH--EMGIEIVE
RPOD_BUCBP -MEQNPFQSQLKLLVTVGKEQGYLTYSEINDHLSdniINSDQIEDIIQMIN--DMGIQVVE
RPOD_BUCAI -MDQNPQSQLKLLVTHGKEQGYLTYSEVNDHLPEDIIDSEQIDDIQMIN--DMGIPVVE
RPOD_SALTY -MEQNPFQSQLKLLVTRGKEQGYLTYAEVNDHLPEDIVDSQIEDIIQMIN--DMGIQVME
unknown_EC -MEQNPFQSQLKLLVTRGKEQGYLTYAEVNDHLPEDIVDSQIEDIIQMIN--DMGIQVME
RPOS_COXBU MKTKTTTKTIKKAARK-----IKKPSKRKIKKTAKRSPKPKKIKASDHGLIFAK
RPOS_VIBCH -----MSVSNVTVKVEEFDFEDEALEVLE
RPOS_SALTY -----MSQNTLKVHDLNEDAEPDENGVEAFD
RPOS_ECOLI -----MSQNTLKVHDLNEDAEPDENGVEVFD
```

```
RPOF_STRCO -----QAPAQAQEAPAPQRSRGADT---RALTQV--LFGELKGLAPGTP
RPOF_STRAU EPREEP-----HEEPPSPAAPRPQRSRGADT---RALTQV--LFGQLKGLQPGRTR
RP32_PSEAE -----GANLEA---YVHSVNSIPLLS
RP32_PROMI -----GSIEA---YIRAANSYFMLTA
RP32_SERMA -----GSLEA---YIRAANAYFMLTA
RP32_ECOLI -----GNLDS---YIRAANAWPMLSA
RP32_CITFR -----GNLES---YIRAANAWPMLSA
RPOD_LEPIN EYTRNTLEPASTL--VPKDDSKPARKKKESSASTS---GSEDPKLYLREIGKVSLSIG
RPOD_BUCBP KAPDSDDLILHEIKRNNETDIVEATAQVLSTVSELGRITDPVRMYMREMGTVELLTR
RPOD_BUCAI EAPDADDLILNEI--NTDDEDAVEAATQVLSSVESELGRITDPVRMYMREMGTVELLTR
RPOD_SALTY EAPDADDLLAEN--TSTDEDAEAAAQVLSSVESEIGRTDPVRMYMREMGTVELLTR
unknown_EC EAPDADDLMLA----ENTADEDAEAAAQVLSSVESEIGRTDPVRMYMREMGTVELLTR
RPOS_COXBU TKKET-----TEKEDAELANAKAKTKKRRET---RSSDPTQIYLRELGFQPLLNA
RPOS_VIBCH TDAEL-----TSDEELVAVEGASEDVREEFDASAKSLDATQMYLSEIGFSPLLTA
RPOS_SALTY EKALSE-----EEPSNDLAEELLSQGATQ---RVLDATQLYLGEIGYSPLLTA
RPOS_ECOLI EKALVE-----QEPSNDLAEELLSQGATQ---RVLDATQLYLGEIGYSPLLTA
```

```
RPOF_STRCO EHD-----
RPOF_STRAU EHE-----
RP32_PSEAE EQERELAERLFYQQ-----
RP32_PROMI EEEKELAERLHYEG-----
RP32_SERMA EEERELAERLHYQG-----
RP32_ECOLI DEERALAEKLYHG-----
RP32_CITFR DEERALAEKLYHG-----
RPOD_LEPIN ETEVFLAKRIE-KGEKIIIEETILSSSILRANYIKLLPKIRSKKIKVYDLIRVDKMYALNA
RPOD_BUCBP EGEIDIAKRIE-DGINQVQCSVAEYPEAITHLLEQYNRVELGELRLSELIH--GFVDPNA
RPOD_BUCAI EGEIDIAKRIE-EGINQVQCSVSEYPEAITYLLEQYDRVKTGQIRLSDIIT--GFVDPNA
RPOD_SALTY EGEIDIAKRIE-DGINQVQCSVAEYPEAITYLLEQYDRVEAEEARLSDLIT--GFVDPNA
unknown_EC EGEIDIAKRIE-DGINQVQCSVAEYPEAITYLLEQYDRVEAEEARLSDLIT--GFVDPNA
RPOS_COXBU KEELKIARRVH-KG-----
RPOS_VIBCH EEEVLYARRAL-RG-----
RPOS_SALTY EEEVYFARRAL-RG-----
RPOS_ECOLI EEEVYFARRAL-RG-----
.
:
```

```
RPOF_STRCO -----
RPOF_STRAU -----
RP32_PSEAE -----DL-----
RP32_PROMI -----DL-----
RP32_SERMA -----DL-----
RP32_ECOLI -----DL-----
RP32_CITFR -----DL-----
RPOD_LEPIN EEAHKLEELFFKNILVIQE-----QEKVLQEAVSKIRKYSSET
RPOD_BUCBP EAVHTSITNHINTNSEHHHNDNTEENN-----DDHLVDPELAREKFIALKNQYHI
RPOD_BUCAI EEIIFPTAIHIGSELLDEQNNNEDEENNQEDH---EDDHSIDPELANEKFSSELRIQYNN
RPOD_SALTY EEEMAPTATHVGSLSQEDLDDEDEDEEDGDGDDAADDNSIDPELAREKFAELRAQYVV
unknown_EC EEDLAPTATHVGSLSQEDLDDEDEDEEDGDGDDAADDNSIDPELAREKFAELRAQYVV
RPOS_COXBU -----DP-----
RPOS_VIBCH -----DE-----
RPOS_SALTY -----DV-----
RPOS_ECOLI -----DV-----
```

```
RPOF_STRCO -----
RPOF_STRAU -----
RP32_PSEAE -----
RP32_PROMI -----
RP32_SERMA -----
RP32_ECOLI -----
RP32_CITFR -----
```



```

RPOF_STRCO      DPA----LDGVEHRDL-----VRHLLVQLPEREQRILLRLRY----YSNLTQSQISAE
RPOF_STRAU      DPE----LAGVEHRDL-----VRHLLVQLPEREQRILLRLRY----YNNLTQSQISAE
RP32_PSEAE      EDHRYDPARQLEDADWSDSSANLHEALEGLDERSRDILQQRW--LSEKATLHDLAEKY
RP32_PROMI      EDKSSDFADGIEEDNWDNHAADRLTLAKTLDEERSQDIIRARW-LDEDNKSTLQELADKY
RP32_SERMA      QDKSSDFAEGIEEDNWNESNAADKLAYALEGLDERSQHIIRARW-LDDDNKSTLQELADQY
RP32_ECOLI      QDKSSNFADGIEEDNWEQAANRLTDAMQGLDERSQDIIRARW-LDEDNKSTLQELADRY
RP32_CITFR      QDKSSNFADGIEEDNWDQAANKLTHAMEGLDERSQDIIRARW-LDEDNKSTLQELADRY
RPOD_LEPIN      DTEVEVTPVNAASSILAEQ----IRQVLHTLPAREQKVI RMRFGLDGYPQTLEEVGYQF
RPOD_BUCBP      DTNIELPLDSATSASLRSA----TKNVLSGLTTREAKVLRMRFGIDMNTDHTLEEVGKQF
RPOD_BUCAI      DTTLELPLDSATSESLRSA----THDVL SGLTAREAKVLRMRFGIDMNTDHTLEEVGKQF
RPOD_SALTY      DTTLELPLDSATTESLRAA----THDVL SGLTAREAKVLRMRFGIDMNTDHTLEEVGKQF
unknown_EC      DTTLELPLDSATTESLRAA----THDVL SGLTAREAKVLRMRFGIDMNTDHTLEEVGKQF
RPOS_COXBU      DDNNIDPARLIQNVDLQDH----IERWLAQLDERHREVILRFGLHENEKGTLEAVGKAV
RPOS_VIBCH      DSHNADPEFSTQDDDIRES----LLNWLDELNPKQKEVLARRFGLGYPESTLEEVGREI
RPOS_SALTY      DEKENGPEDTTQDDMKQS----IVKWL FELNAKQREVLARRFGLGYEAAATLEDVGREI
RPOS_ECOLI      DEKENGPEDTTQDDMKQS----IVKWL FELNAKQREVLARRFGLGYEAAATLEDVGREI

```

```

: * . : : * : * :

```

```

RPOF_STRCO      GVSQMHVSRLRLARSFQRLRSANRIDA-----
RPOF_STRAU      GVSQMHVSRLRLARSFARLRSANRIEA-----
RP32_PSEAE      NVSAERIRQLEKNAMSKLKGRI LA-----
RP32_PROMI      GVSAERVRQLEKNAMKCLR LAIED-----
RP32_SERMA      GVSAERVRQLEKNAMKCLRMAIEA-----
RP32_ECOLI      GVSAERVRQLEKNAMKCLRRAAIEA-----
RP32_CITFR      GVSAERVRQLEKNAMKCLRRAAIEA-----
RPOD_LEPIN      KVTREIRIRQIEAKALRRLRHPSPRSKLLK--DYIDG---
RPOD_BUCBP      DVTREIRIRQIEAKALRRLRHPSPRSSEILR--SFLDD---
RPOD_BUCAI      DVTREIRIRQIEAKALRRLRHPSPRSSEVLR--SFLDD---
RPOD_SALTY      DVTREIRIRQIEAKALRRLRHPSPRSSEVLR--SFLDD---
unknown_EC      DVTREIRIRQIEAKALRRLRHPSPRSSEVLR--SFLDD---
RPOS_COXBU      GLTREVRVRIQIDALQQLRHILEMEGVGTGEEVED----
RPOS_VIBCH      NLTREVRVRIQVEGLRRLREILVKQGLNMEALFNVEYDN
RPOS_SALTY      GLTREVRVRIQVEGLRRLREILQ TQGLNIEALFRE----
RPOS_ECOLI      GLTREVRVRIQVEGLRRLREILQ TQGLNIEALFRE----

```

```

: : : : : *

```

