

## Recueil de stratégies de ré-annotations séquences protéiques « predicted » et « hypothetical »

Anne Poupon, IBBMC

LOCUS	ZP_00387963	142 aa	linear	BCT 12-APR-2005
DEFINITION	COG2246: Predicted membrane protein [Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365].			
ACCESSION	ZP_00387963			
VERSION	ZP_00387963.1	GI:62516616		
1 mdkekinelw gkykdiipyv fwgvmttlvn ifsywlcnsv fkwdimpatl maqflsivfa				
61 yltnrkvwfh sqastfkeya seiasffaar igtalldmai mfifaekmhl nsmvikilan				
121 vvvvivnyva skfwifkskd se				

**Blast:** beaucoup d'homologues avec des e-values très faibles (de 0 à 10<sup>-10</sup>), avec la même fonction: wall teichoic acid glycosylation protein GtcA

**Conclusion:** c'est aussi une wall teichoic acid glycosylation protein GtcA

LOCUS	ZP_00387964	136 aa	linear	BCT 12-APR-2005
DEFINITION	COG5652: Predicted integral membrane protein [Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365].			
ACCESSION	ZP_00387964			
VERSION	ZP_00387964.1	GI:62516617		
1 mgsvllisgl tfqnsqgsaq ltlkfvnlfv kvfhltgtm maeypiyhmm rklahtaeyf				
61 llgtstcyff srwqkhyaws alalsaafsl fdqtskllvp grefdptdfp fdiagyllgi				
121 gli111qrs w akraavn				

**Blast:** Pas de hit avec un score significatif, mais dans les non-significatifs, au moins 3 fois la même fonction (voir 62516617\_BlastIt1.html), avec des taux d'identité de séquence pas trop mauvais (20 à 25%), sur toute la longueur, et ce sont aussi des protéines bactériennes.

Si on fait une deuxième itération dans PsiBlast sans sélectionner ces protéines, leur e-value diminue, mais pas suffisamment pour être significative. Si on fait une deuxième itération en sélectionnant les deux premières protéines non significatives, on sort beaucoup de ces protéines (acetobutylicum phosphotransbutyrylase) avec des e-values très petites.

Si on fait un blast avec la première de ces protéines, on constate que les résidus conservés dans la famille sont bien les mêmes que ceux qui s'alignent dans la protéine de Bulgaricus. Balst détecte le domaine VanZ, si on regarde dans Pfam les résidus conservés dans ce domaine, il y a une assez bonne correspondance, mais Pfam ne détecte pas le domaine dans notre protéine (ni blast d'ailleurs).

Donc, c'est une hypothèse tout à fait plausible.

Pour confirmer il faut passer en manuel..., mais c'est un cas où ça peut être très utile de signaler que cette hypothèse est vraisemblable.

LOCUS ZP\_00387965 60 aa linear BCT 12-APR-2005  
DEFINITION hypothetical protein Ldelb01000003 [Lactobacillus delbrueckii  
subsp. bulgaricus ATCC BAA-365].  
ACCESSION ZP\_00387965  
VERSION ZP\_00387965.1 GI:62516618

1 mfglsiigim vfhyfehirs ahlllkseqf wdyygstgv dfflflsgmg lfysltanfp

**Blast:** rien, même non significatif.

**ScanProsite:** rien

**TMPred:** 2 hélices transmembranaires.

**Conclusion:** rien, même pas sûr que ce soit une protéine.

LOCUS ZP\_00387966 154 aa linear BCT 12-APR-2005  
DEFINITION COG0615: Cytidylyltransferase [Lactobacillus delbrueckii subsp.  
bulgaricus ATCC BAA-365].  
ACCESSION ZP\_00387966  
VERSION ZP\_00387966.1 GI:62516619

1 mkrvitygtf dllhyghinl lrrakaqgdy livalstdef nwnskhkkty fsyeqrkqll  
61 eairyvdlvi pendwdqkrs dmheyhidtf vmgndwkgkf dflkeegvnn vylprtpeis  
121 sskikhdyd anevteeskl shddldtdpd hdkk

**Blast:** Détecte un domaine CTP transférase 2. Homologue à de nombreuses glycerol-3-phosphate cytidylyltransferase avec des e-values très basses.

**Conclusion:** La fonction est glycerol-3-phosphate cytidylyltransferase. A signaler quand même: l'homologie avec les transférases va du début jusqu'au résidu 120 environ, il reste donc à peu près 30 résidus qui ne sont pas dans les autres transférases.

LOCUS ZP\_00387967 303 aa linear BCT 12-APR-2005  
DEFINITION COG3475: LPS biosynthesis protein [Lactobacillus delbrueckii  
subsp. bulgaricus ATCC BAA-365].  
ACCESSION ZP\_00387967  
VERSION ZP\_00387967.1 GI:62516620

1 mdniklklse dffkeevrnd ytvspemkkv waveldlldq ldrvcqkydi pwylsggsll  
61 gavrhqgyip wdddldlmyy rkdferlcev asqeftepyf fqteetdtgs irghaqlrns  
121 attailksee yfhyqfnqgi fidifpldnv pddpeerqaf vksvnqlkhr arqfynycng  
181 ypnknlsglk qfllytkfff ekkksggrni yydrfakeit kyddqatqev mmvmletekc  
241 cwkreypvanp vrvpfellsl pipkdydpl1 tkqygkwnif vrggslhgsv ifdpdksyke

**Blast:** beaucoup d'homologues avec la même fonction et des e-values basses (de l'ordre de  $10e-20$ ):  
LPS biosynthesis protein

**Conclusion:** on propage la fonction.

```

LOCUS       ZP_00387970                198 aa         linear       BCT 12-APR-2005
DEFINITION  COG2244: Membrane protein involved in the export of O-antigen and
teichoic acid [Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365].
ACCESSION   ZP_00387970
VERSION     ZP_00387970.1  GI:62516623

      1 mllanigigy ygnreiayvr dnkqkmaatf weiqivktvm tvvaylsfvv fmafysgnkt
      61 ymwvqsinli avafdiswly egiedfrtv lrntfvkivs mvaifvfiks ssdvalyiai
     121 laistflgnl tlphtfkml pgvnlaclkp lrhfkptiam fipqiatqly vqlnrtmlgl
     181 mvdqkasgft niqttwls

```

**Blast :** CD search détecte un domaine Polysaccharide biosynthesis protein. Blast donne beaucoup d'homologues impliqués dans le transport membranaire de sucres chez la bactérie. Mais ces prots font dans les 500 résidus, et la notre seulement 200.

**TMPred :** 3 ou 4 hélices transmembranaires.

**Conclusion :** Il s'agit d'une protéine impliquée dans le transport membranaire et/ou la biosynthèse de polysaccharides. La fonction qui a été retenue dans le fichier GenPep est celle du premier homologue dont la fonction est annotée. Cependant, il y a d'autres homologues, avec des e-values du même ordre, et dont la fonction est légèrement différente (en fait il ne s'agit pas du même sucre), il me semble donc hasardeux de conclure sur la nature du sucre qui est le substrat de cette protéine.

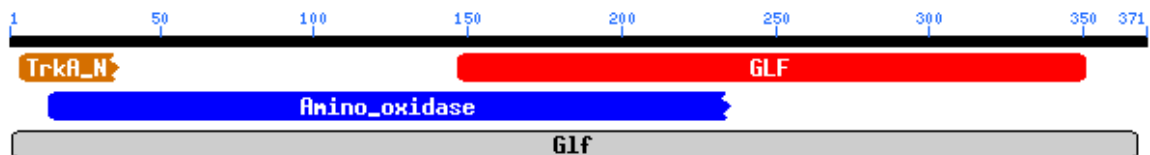
```

LOCUS       ZP_00387971                371 aa         linear       BCT 12-APR-2005
DEFINITION  COG0562: UDP-galactopyranose mutase [Lactobacillus delbrueckii
subsp. bulgaricus ATCC BAA-365].
ACCESSION   ZP_00387971
VERSION     ZP_00387971.1  GI:62516624

      1 mskylvvgag lfgavfarea akrghetvi ekrdhiagni ytkeidgiqv hqygahifht
      61 snkevweyvq qfaefnrytn spvanykgkm ynlpfnmntf tqmwgvrtpp eamdkiqr
     121 aemagktpqn leeqaislig rdiyeklikg ytekqwrka telpafiikr vpvrlidnn
     181 yfnddyqgip kgytklven mlkddkitve ldtddfkakd eylqkfdrvv ytgpidffid
     241 yklgeleyrs lrfeteeknv gnyqgnavin ytdaetpytr viehkhfefg kgdpdktiit
     301 reypadwhkg depyypvnd rnsdlyaqyk emadkedakv ifggrlgqyk yymndqvaaa
     361 aleavneefg k

```

**Blast :** CD search détecte plusieurs domaines qui se chevauchent :



Beaucoup d'homologues avec la même fonction (UDP-galactopyranose mutase), et des e-value très basses.

**Conclusion :** c'est une UDP-galactopyranose mutase

LOCUS	ZP_00387972	244 aa	linear	BCT 12-APR-2005
DEFINITION	hypothetical protein Ldelb01000012 [Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365].			
ACCESSION	ZP_00387972			
VERSION	ZP_00387972.1	GI:62516625		
	<pre> 1 mpddsvylpv lvgavknyka giayqrddeg dnisarnppy seltgvywaw knlkdvdaig 61 lvhyrryfyv skphdpdhva kgadyehfla dhdvivpkkrr nyyiesnydh yihahpaep1 121 dktrdiiada ypdylpafdm vmkrrsahmf nmfvmkrapaf esycefvgfv lsklegqidi 181 sgysvqdqrv ygyiserlld vwlyttkqdf vempwgqigq eavikkgylnl ikrklgigkk 241 qthf </pre>			

**Blast :** Pas de domaine détecté. Beaucoup d'homologues avec des e-values très basses. L'homologue le plus proche dont la fonction est annotée est une galactosyl transférase, mais après il n'y a presque que des glycosyl transférases.

**Enzyme :** Dans la banque ENZYME, beaucoup de glycosyl transférases, dont certaines sont également des galactosyl transférases. Il n'y a donc pas de contradiction. Par contre, cela ne précise pas entièrement la fonction.

**Conclusion :** C'est une glycosyl transférase.

LOCUS	ZP_00387973	188 aa	linear	BCT 12-APR-2005
DEFINITION	COG2148: Sugar transferases involved in lipopolysaccharide synthesis [Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365].			
ACCESSION	ZP_00387973			
VERSION	ZP_00387973.1	GI:62516626		
	<pre> 1 mlalivfspv flvlslliks rdggsaffaq erigkggkpf kmykfrsmkm daeeilksdp 61 elyqkyvand yklladedpr itpigrwmrr asvdelpqfv nilkgdmsii gprpvvekel 121 aeygnrkd kf lsvrpgamgl wqatgrsnis ypercdvele yidnisftyd vkiffqtifs 181 ilkkegay </pre>			

**Blast :** CD search détecte un domaine Bacterial sugar transferase. Dans le blast beaucoup d'homologues ayant la fonction sugar transferase.

**Conclusion :** C'est une tranférase de sucre. A nouveau, il n'est pas possible de conclure sur la nature du sucre, et donc pas sur la fonction précise, étant donné la diversité des homologues.

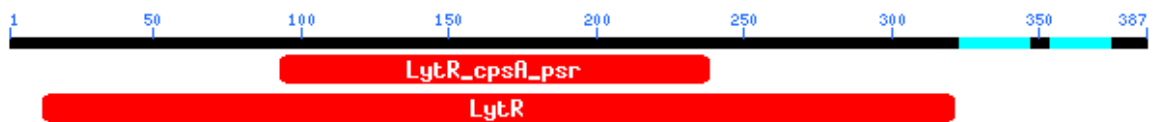
LOCUS	ZP_00387974	257 aa	linear	BCT 12-APR-2005
DEFINITION	hypothetical protein Ldelb01000014 [Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365].			
ACCESSION	ZP_00387974			
VERSION	ZP_00387974.1	GI:62516627		
	1 mgyprdfqln yatlfsvspl vidwgsysqi eaemrlfqaapgkyayyhl lsgldlplan			
	61	qdeihaffaa hpgkefitys sqkngaqla rvqkyhfthn frqpnkamrl frkiekaeqr		
	121	vfprkkfag tlafgsnwvs lendlvqvl reddrirtmf argflvdell vptmlniype		
	181	fkdriddydrp vhdrpeefqg slryinwdg spyvwrekdy etllaarrqg hlfsrkfdae		
	241	vdkaiidkia gqllleik		

**Blast :** homologue avec glycosyl transferase.

**Conclusion :** C'est une glycosyl transferase. Pas de fonction plus précise.

LOCUS	ZP_00387975	387 aa	linear	BCT 12-APR-2005
DEFINITION	COG1316: Transcriptional regulator [Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365].			
ACCESSION	ZP_00387975			
VERSION	ZP_00387975.1	GI:62516628		
	1 mdkntrvgkr kkrnkamrif wsvmavilva iggfalyeyn tvknaadtay rsgglgnaen			
	61	gsknsvisns kpiaillmgt dtgalgrtyk grtdsimvav lnpkttkttl vsferdqqvn		
	121	lpdypensps klnaayaygn akelakvlkk yynipinayv linmgglkti vnkvvvgvdia		
	181	pilsfsyegy tftkgkthm dgakalaysr mryddpegdy grqkrqrqvl sallnkaesa		
	241	tllnssfis slskqvqtdf tfsdmtsmak rylaatkdlk tdythgtsym qdgvsyqkvs		
	301	vserqrisnl irktlglktk tvstsdids tsssssssst ttdsdtgga gntagpsdng		
	361	aaagngadsg ttggysagnd qnnstgy		

**Blast :** CD search détecte 2 domaines au même endroit :



Le plus grand est un domaine retrouvé dans de nombreux régulateurs de la transcription bactériens, et se lie à une séquence ADN spécifique.

Beaucoup d'homologues avec des e-values très faibles, tous annotés comme transcription regulator

**Conclusion :** C'est un transcription regulator.

LOCUS	ZP_00387976	52 aa	linear	BCT 12-APR-2005
DEFINITION	hypothetical protein Ldelb01000017 [Lactobacillus delbrueckii			

```

      subsp. bulgaricus ATCC BAA-365].
ACCESSION   ZP_00387976
VERSION     ZP_00387976.1  GI:62516629
1 mqkltalllla sasflvaspa kassladrny qgqkvitvng drptffqkqp vd

```

**Blast :** Rien

**TMPred :** Une hélice transmembranaire

**Conclusion :** rien. De plus, cette séquence est très courte, ce n'est peut-être même pas une protéine.

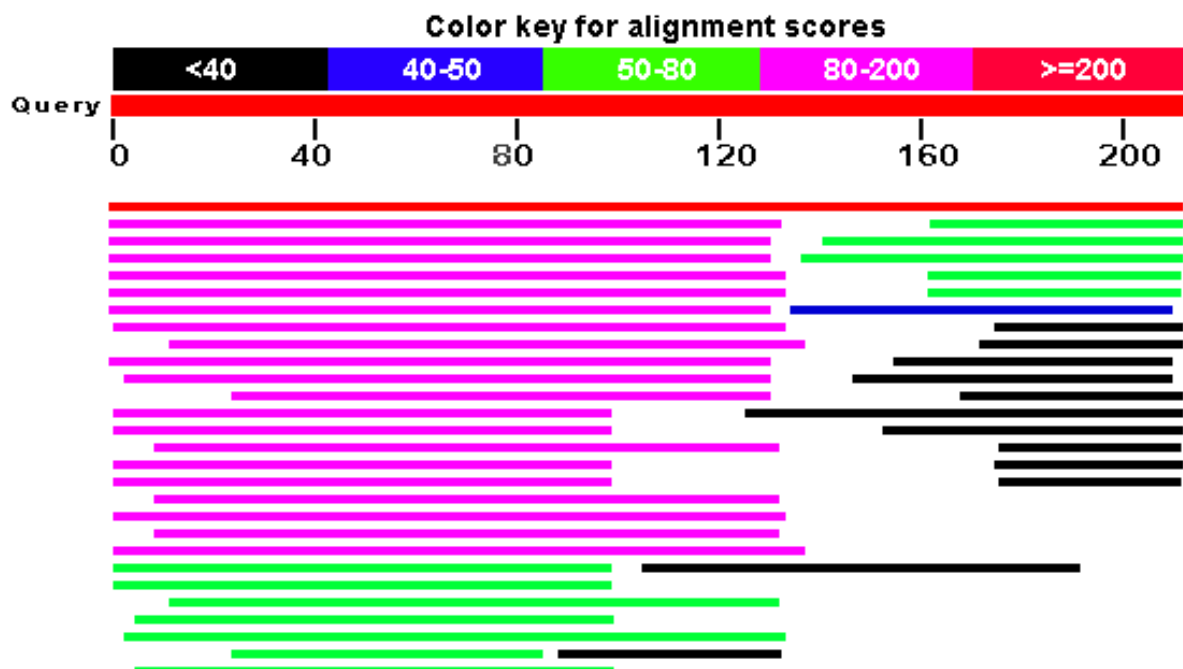
```

LOCUS       ZP_00387977                211 aa           linear   BCT 12-APR-
2005
DEFINITION  hypothetical protein Ldelb01000018 [Lactobacillus delbrueckii
            subsp. bulgaricus ATCC BAA-365].
ACCESSION   ZP_00387977
VERSION     ZP_00387977.1  GI:62516630

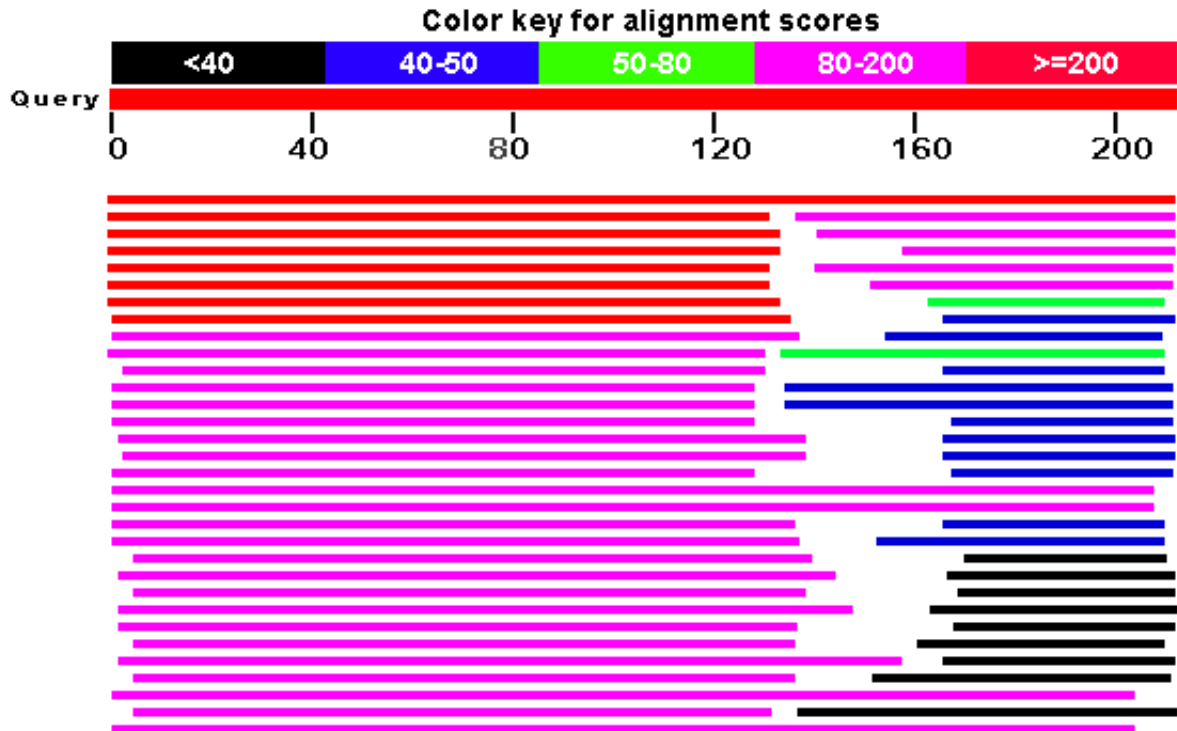
      1 mpakarkply wnptawhnkk iaggwlynrs hligyqltgq nnpknlitg arqlndpgml
      61 kyenqvasyi ksssrhyiry rvkpifrgre llargvemea qstgsnavrf hvyifnvqdg
     121 vklngsngts vvtgaakksa kktvvkkaa kktsskkki ktsttgrivg yrrykiyhvp
     181 ggagyhmnsa navyfretete akragyrral r

```

**Blast :** Pas de domaine détecté. Au premier tour de Psi-Balst, ça ressemble a une protéine à deux domaines longueur (voir fichier 62516630\_Balst1.html):



Mais si on fait une deuxième itération, certaines protéines s'alignent sur toute la longueur (voir fichier 62516630\_Balst2.html):



Si on regarde les protéines qui s'alignent sur toute la longueur, elles ont des e-values très basses ( $10^{-40}$ ), et les mêmes fonctions (desoxyribonucléase). C'est également proche de la fonction de la majorité des protéines qui ont des e-values basses mais s'alignent seulement sur les 120 premiers résidus (nucléases).

**Conclusion :** C'est une nucléase, probablement une desoxyribonucléase.

LOCUS	ZP_00387978	82 aa	linear	BCT 12-APR-2005
DEFINITION	hypothetical protein Ldelb01000020 [Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365].			
ACCESSION	ZP_00387978			
VERSION	ZP_00387978.1	GI:62516631		
<p>1 msdllvlava angritlnsl psyfpvggk pryfyrkkva iggtegsltg ayagsdmvcl 61 fadasgvywf kgnrgieki gs</p>				

**Blast :** rien, même non significatif.

**ScanProsite :** rien

**Pfam :** rien

LOCUS	ZP_00387979	86 aa	linear	BCT 12-APR-2005
DEFINITION	hypothetical protein Ldelb01000021 [Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365].			
ACCESSION	ZP_00387979			
VERSION	ZP_00387979.1	GI:62516632		
<p>1 mmkeknlwtw ltlvlllvgf aamgasvwl sqaakqtqrn sqlagglyqa srtsqeass</p>				

61 ikkrlldklas dnsslkkkns rlaken

**Blast** : rien.

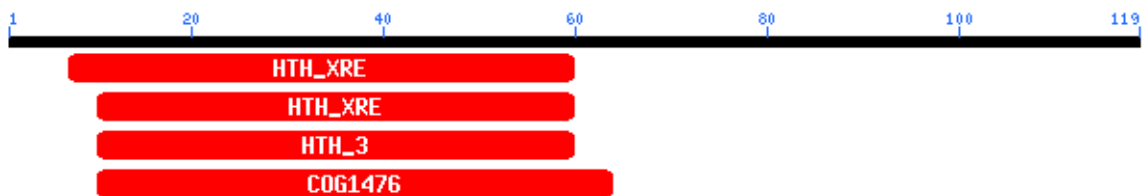
**ScanProsite** : rien

**Pfam** : rien

LOCUS ZP\_00387980 119 aa linear BCT 12-APR-2005  
DEFINITION COG1396: Predicted transcriptional regulators [Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365].  
ACCESSION ZP\_00387980  
VERSION ZP\_00387980.1 GI:62516633

1 meknmigkyl rdrrrrrgms qqelalalgv skqtisnwev grkvprmkav dkianifgvs  
61 rnsilaglpv emleqadrrv vldtldrdir1 tylgqqvpre yidiiek1mr wdiaerdaq

**Blast** : CD search détecte divers domaines HTH :



Ce sont des domaines Hélice-tour-hélice, donc de liaison à l'ADN.

Beaucoup d'homologues avec la même fonction et des e-values basses : transcription régulateur.

**Conclusion** : transcription regulator. Attention quand même : la très grande majorité de ces protéines sont annotées « Predicted transcriptional regulators », donc la fonction a été propagée...

LOCUS ZP\_00387981 324 aa linear BCT 12-APR-2005  
DEFINITION COG0240: Glycerol-3-phosphate dehydrogenase [Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365].  
ACCESSION ZP\_00387981  
VERSION ZP\_00387981.1 GI:62516634

1 malarmlsns ghevtvwsal pgevdelstr rraqnlpqmv ipdeikftke iaeacqdkdi  
61 ilfavpsvfv rstaktaaaf ipdgqiivdv akgiepdt11 tlteviadel nkdgkghgnvh  
121 yvamsgptha eevakdlptt ivsacedqav aekvqdvfmn knmrvytnsd rlgvelcgal  
181 knvialasgi cs1glgygdnm raaliirgma eikrlglkmg gkedsfdgla gmgdlivtat  
241 skesrnnnag yligkgsae eakkevgmvv eginaipaal eladkydvem pivfavdavv  
301 nrgadaretv dalmlrekks emtk

**Blast** : Glycerol-3-phosphate deshydrogenase.

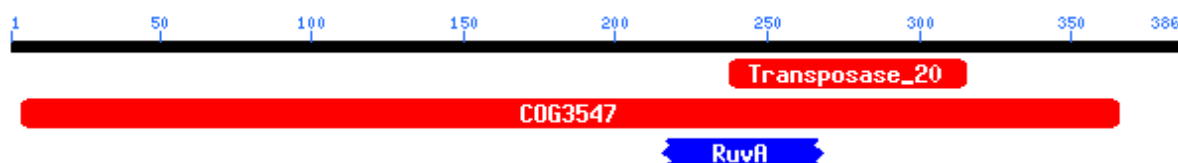
LOCUS ZP\_00387982 386 aa linear BCT 12-APR-2005  
DEFINITION COG3547: Transposase and inactivated derivatives [Lactobacillus

```

delbrueckii subsp. bulgaricus ATCC BAA-365].
ACCESSION   ZP_00387982
VERSION     ZP_00387982.1  GI:62516635
1  mvngqkvndy aisndmvgfn rllgdlkqvt kpqiifeatg vysrrlqaf1 dmhelryvmm
61 npleakrktk ddlhqnktdk ldalylaklq sehqqrlayv qnkeyqelma nnriyeqash
121 dlitnknrlh kavqltfpei ehllanprgk nywsialrfp hpdivletke vdiidflkgl
181 sgigkkrand itqslirlak vacpavkks ahirglkmai nnilsaeec qtalqemakl
241 apkrdleilt slpgigenta lriiselgdi rrfnpsqln afvgvdpqvy esgnltahls
301 iskrptaigr kvlylainqi qsakkagnpc hiadyyekrk rssetashkk aaiasihkll
361 rtifaliknd qlsydvakh nqrlls

```

**Blast** : Domaines détectés par CD search :



COG3547 : Transposase and inactivated derivatives [DNA replication, recombination, and repair]

Beaucoup d'homologues appartenant aussi à cette fonction.

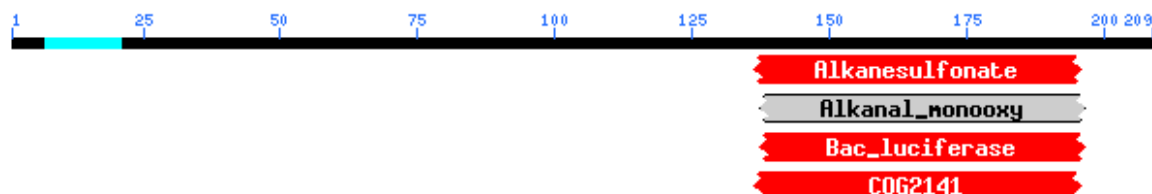
**Conclusion** : Transposase and inactivated derivatives [DNA replication, recombination, and repair]

```

LOCUS       ZP_00387983                209 aa           linear   BCT 12-APR-
2005
DEFINITION  COG2141: Coenzyme F420-dependent N5,N10-methylene
tetrahydromethanopterin reductase and related flavin-dependent
oxidoreductases [Lactobacillus delbrueckii subsp. bulgaricus ATCC
BAA-365] .
ACCESSION   ZP_00387983
VERSION     ZP_00387983.1  GI:62516636
1  mlitatlill vvfslslly vqrqqkaer dqaqalatgf srtrtlkgei ksaltlsttl
61 kelvlasdgd vdnfstvakd llaentaasn lqlapkgkvt eiypkgnv gkinlltdpi
121 rgplcrygis hqvtvsidr1 snhrflglv sdrknefka fgrdwnerse ifkegwldtk
181 kslthpnrhl fsgnyynklt gslscrasa

```

**Blast** : domaines partiellement détectés par CD search :



Annotée comme COG2141, Coenzyme F420-dependent N5,N10-methylene tetrahydromethanopterin reductase and related flavin-dependent oxidoreductases [Energy production and conversion]. Mais

seulement 18% du domaine s'aligne avec notre protéine, ce qui n'est pas assez pour conclure.

Dans le Blast, pas de protéine s'alignant sur toute la longueur.

A la troisième itération de Psi-Blast, on a des protéines qui s'alignent sur toute la longueur. Les fonctions sont relativement diverses, mais elles ont des points communs, en particulier elles semblent toutes fixer le FMN.

**Conclusion :** pour aller plus loin, il faudrait faire des alignements manuels avec les diverses familles de protéines concernées, et voir dans lesquelles on a le plus de résidus actifs conservés. Ici on pourrait conclure à quelque chose du genre « putative FMN dependent monooxygenase or oxydoreductase ».