

QUIZ I
L519: Bioinformatics: Theory and Applications
School of Informatics
Indiana University – Bloomington
Jan. 23, 2003

Command syntax Your program should follow the exact syntax (command name and arguments) as specified in each problem. The command name should be `your-network-id` followed by the letter `p` and followed by *the problem number*. For example, the solution for problem 1 should named as `your-network-id.p1`.

Data Below is the .ptt file for *Borrelia burgdorferi* (AE000783).

Borrelia burgdorferi complete genome - 0..910724

850 proteins

Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product
105..677	+	190	2687895	BB0001	-	-	-	-
768..1796	-	342	2687889	BB0002	-	G	COG1472	-
1784..3148	-	454	2687894	BB0003	-	-	-	-
3395..5188	+	597	2687888	BB0004	-	G	COG1109	-

....

The 8th column under COG denotes that the protein coding sequence belongs to the family. For example, the second gene with id 2687889 belongs to the family COG1472.

There is another genome, *Treponema pallidum* (AE000520).

Problem I: Family membership (60 pts) Write a program that reports the family name for a given sequence id.

Command syntax: `$ your-network-id.p1 sequence-id` should report the COG family name.

For example, `$ your-network-id.p1 2687889` will report COG1472.

Problem II: family membership (30 pt) Write a program that reports genes with respect to their family memberships.

Command syntax: `$ your-network-id.p2 AE000520.ptt AE000783.ptt` will reads in the two .ptt files and report genes with respect to their family memberships.

For example, sequence 2688629 and 3323222 belong to the same family COG0335. Then the result from the execution of your program need to include output as below. *The output format can be in any form.*

Family COG0335
2688629
3323222

Problem III: sequence extraction (10 pt) Once you know sequences that belong to the same family, you may want to look at or compare the sequences. Write a program that extract a sequence given a sequence id for .faa files.

Command syntax: \$ your-network-id.p3 sequence-id file.faa
will extract the amino acid sequence for sequence-id.

For example, \$ your-network-id.p3 2688629 AE000783.faa will extract the amino acid sequence as below.

```
>gi|2688629|gb|AAC67044.1| ribosomal protein L19 (rp1S) [Borrelia burgdorferi]  
MDLIRKIEAQNKKNEAFVFNVDTVRVVYKIIEGSNERLQSFEGIVISFQNKGGIGKTFLIRKISSGIGVE  
KIFPVYSPIIEKVEVLRGKVRRAKLYYMRNRIGKAAMKIKERLTIKKVKH
```