

Algorithms in Bioinformatics I, WS2006/7

Assignment sheet # 7

Christian Rausch

November 29, 2006

1 Clusters of orthologous genes (3 points)

COGs – Clusters of Orthologous Groups – natural system of gene families from complete genomes. Clusters of Orthologous Groups (COGs) were delineated by comparing protein sequences encoded in complete unicellular genomes representing 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain. The Initial Version of COGs includes 44 organisms. The Updated Version of COGs includes 66 organisms in the Unicellular Clusters, plus Eukaryotic Clusters (called KOGs). More organisms will be added in the future. NCBI's COGs website is found at <http://www.ncbi.nlm.nih.gov/COG>.

Please make yourself familiar with the COGs website and answer the following questions:

- What are COGs?
- How are COGs created ?
- Which are the organisms used to create the NCBI COGs database?
- What is the COG function code for “Cell cycle control, cell division, chromosome partitioning”?
- How many COGs are currently in the NCBI COGs database that have to do with “Lipid metabolism”?
- How many COGs are currently in COG for “Mevalonate kinase”?
- For a specific COG you view the COGnitor phylogenetic tree for the proteins used to create the COG. Which color would the diamond for “*Saccharomyces cerevisiae*” be?
- By which letter is “*Mycoplasma pneumoniae*” represented in a phylogenetic patterns search?
- How many COGs include proteins from both “*Pyrococcus horikoshii*” (k) and “*Borrelia burgdorferi*” (o)?
- Use the phylogenetic patterns search to find differences in closely related genomes: e.g. “*Bacillus subtilis*” vs. “*Bacillus halodurans*” in the context of “information storage and processing” processes. How many COGs relevant to transcription are shared, only in “*Bacillus subtilis*”, and only in “*Bacillus halodurans*”, respectively?
- How many COGs related to cellular processes do they both share?
- How many COGs are shared by all organisms used to create the COGs database?
- How many COGs are involved in FATTY ACID BIOSYNTHESIS ?

2 Sorting signed permutations (7 points)

2.1 reduced permutations and framed intervals (1 point)

- What are *reduced permutations*? – Please write down the definition in your words.
- Give an example for a *reduced permutation* (one that is not in the script/slides; use caps, that means 0 at the beginning and $n + 1$ at the end).
- Please define the expression *framed interval* and write down all such framed intervals in your example of a reduced permutation.

2.2 Breaking hurdles (4.5 points)

2.2.1 (0.5 points)

Take a look at Section *Breaking Hurdles* in the script. The example permutation (0 2 4 3 1 5) contains one *hurdle*. Please write it down. What is meant by position $i + 1$ and $i + k$?

2.2.2 (4 points)

After cutting the hurdle the permutation is: (0 -3 -4 -2 1 5). Please show that this permutation can be sorted in 4 reversals by applying *Algorithm 1* (Give all intermediate and final induced reversals as well as their scores).

2.3 Super hurdles (1.5 points)

2.3.1 (1 point)

Please give all hurdles in the example permutation (0 2 5 4 3 6 1 7). Write down all steps to cut and sort the hurdle beginning at 2.

2.3.2 (0.5 points)

What are the two hurdles contained in the example permutation (0 2 4 3 5 1 6 8 7 9)? Please *merge* them.

Due by 10am, Monday, 4 December 2006