Biofilm-forming capacity in *S. aureus*: Relationship to promoter sequence
Problem Set

This problem set makes use of StaphyloBIKE, accessible through the BioBIKE portal.

1. Investigation of ica cluster

The *ica* locus of *Staphylococcus aureus* has been implicated in adhesion of the bacteria to surfaces. Kim et al (2008, J Microbiol Biotechnol 18:28-34) found that all 112 strains of *S. aureus* that were tested possess the *ica* cluster.

a. Can we verify this claim? Is the cluster specific to *S. aureus* and not found in other *Staphylococcus*? First, how many *Staphylococcus* strains are available in StaphyloBIKE? How many of them are *S. aureus*? To answer this question, you'll probably make use of COUNT-OF (Lists-Tables button, List-Analysis submenu) and the organismal subsets found through the DATA button.

b. Look for *ica* clusters in *Staphylococcus* strains. Do this by displaying the regions surrounding *icaA* genes. For the first go at this, execute GENES-DESCRIBED-BY (Genes-Proteins button, Description-Analysis submenu), provide "icaA" as the argument, and select the DISPLAY option. How many instances are found? Notice from the display that the genes found are not called *icaA*. You might reasonably be concerned about this, so click on one of the links in the display to learn more about the gene. Do you think these genes really are examples of *icaA*?

c. Check in another way. Define a variable (maybe *icaA-genes*) as the list found in the previous step. Be sure to simplify the list (using SIMPLIFY-LIST, found from the Lists-Tables button, List-Production submenu) to get rid of the structure of the list[*]. Then find the genetic context surrounding all the genes. They should be part of the *icaADBC* operon. To find the context, execute CONTEXT-OF (with the DRAW option) acting on the variable you just defined. You should get two displays. The top display gives you list of the genes surrounding each instance of the putative *icaA* gene. To the right of each gene is the number of nucleotides separating it from the next gene. The second display (perhaps behind the first) shows these relationships graphically. You can click on the graphical representation of any gene to learn more about it.

In most cases *icaC* won't show up, because it is three genes away from *icaA*, and the default is to show only two genes away. To rectify this, re-execute CONTEXT-OF with GENE-WIDTH set to 3.

Are all instances of putative *icaA* part of *icaADBC* operons?

d. There are a lot of *Staphylococcus* strains missing. Perhaps Kim et al are wrong about the ubiquity of *icaADBC* or perhaps our method of finding the cluster was faulty. Try another way. Choose one of the *icaA* genes (by name, e.g. *nwmn_2565*) and use it as the argument for ORTHOLOG-OF (Genes-Proteins button), which will give you all evolutionarily synonymous genes amongst sequenced Staphylococci. Define a new variable containing these genes and obtain their context as before. Now how many

[*] The genes are clustered by organism, allowing for the possibility that there may be multiple genes per organism with "icaA" in the description.

S. aureus strains do you believe carry *ica* clusters? How many don't? What about other *Staphylococcus* strains?

e. You found *ica* clusters in two ways. How can you account for the differences in the results from the two methods?

2. Investigation of ica cluster

The *ica* locus appears to be an operon: the genes follow one another without a significant intergenic region. The region upstream of the first gene, *icaA*, should contain the sequences governing the regulation of the operon.

a. Obtain the upstream region of the *icaA* genes (using the larger set of orthologs you found in 1d). You'll want to use the SEQUENCE-UPSTREAM-OF function (Genes-Proteins button, Gene-Neighborhood submenu). Use the LABELED option to maintain the connection between each upstream sequence and the gene from which it is derived.

b. Align the upstream sequences, using ALIGNMENT-OF (Strings-Sequences button, Bioinformatic-Tools submenu). How similar are they?

c. Two sequences stand out as different. How can you account for their difference in sequence? Test any hypothesis you might come up with.

3. Investigation of AgrD protein

[This is a good time to clear your screens] *agrD* encodes a protein that is the precursor for an extracellular peptide signal that allows *S. aureus* to sense cell density. How similar are these proteins amongst Staphylococci?

a. Define a variable containing as large a list as possible of all *agrD* genes in Staphylococcus, using the two methods described in Problem 1. Why do the lists differ from one another?

b. Define a variable containing the list of corresponding AgrD proteins. You'll want to make use of the PROTEIN-OF function (Genes-Proteins button).

c. Align the protein, using ALIGNMENT-OF. How big are the proteins? Proteins are typically several hundred amino acids. (The size of the protein is related to the perhaps surprising result of Problem 3a).

d. Construct a phylogenetic tree (i.e. a family tree) based on the protein sequences, using TREE-OF (Strings-Sequences button, Phylogenetic-Tree submenu). Paste the alignment (or the completed ALIGNMENT-OF function) into the argument calling for an alignment. For the argument labeled "tree-project", make up any name you like, enclosed in quotation marks. Absorb the tree you obtain by executing the function (which could take a while).

e. The tree may not be in as helpful a form as you might like. It would be nice to know which ***organisms*** cluster together. Find a way to display a list showing the organism of each of the AgrD protein. You might want to recall your old experience with APPLY-FUNCTION or with loops.

4. Investigation of RsbU protein

Repeat the operations of Problem 3 with the regulatory protein RsbU. Notice in the alignment that two of the protein stand out, shorter than the others by many amino acids. Explain this result.