03-727 Project Description

Did horizontal transfer contribute to colonization of the human eye by Streptococcus pneumoniae?

Streptococcus pneumoniae is the primary cause of earache (otitis media) in small children. It can also causes serious diseases, including meningitis and pneumonia. *S. pneumoniae* is most commonly found in the ears, nose and throat, but some strains of *S. pneumoniae* can cause an eye infection called conjunctivitis ("pink eye"). Recently, the genomes of *S. pneumoniae* strains sampled from conjunctivitis patients were sequenced, making it possible to ask whether these genomes harbor specific genes that facilitate eye infection. Dr. Luisa Hiller has identified 93 candidate genes in 21 loci (spatial groups) that are unique to the conjunctivitis isolates within *S. pneumoniae.* Since these genes are not found in *S. pneumoniae* genomes sampled from other parts of the human body, they are promising candidates for functions related to colonization of the human eye. This locus has already been investigated computationally and experimentally and the preliminary evidence supports the hypothesis that genes in this cluster promote adhesion to human corneal cells.

This year's Phylogenetics project will be a phylogenetic analysis of one of the candidate conjunctivitis genes. We will provide each student in 03-727 with an amino acid sequence of one of the previously untested conjunctivitis candidate genes. Your goal is to investigate how the pneumococcal conjunctivitis strains acquire these candidate genes.

Milestones:

Week One:

- Determine the phylogenetic distribution of your gene within the Streptococcus genus and in more distantly related bacteria; select an appropriate set of sequences for phylogenetic analysis
- Pieter and I will review your proposed sequence set during next Thursday's class (**Nov. 19**).

Week Two:

- Construct a multiple alignment of those sequences and refine and trim it manually
- Select a model of sequence evolution using TOPALI

- Infer the phylogenetic history of your gene family, using bootstrapping to assess the reliability of your tree.

Week Three:

- Interpret your results, using phylogenetic reconciliation to infer the history of gene duplications and horizontal gene transfer that gave rise to the present day family.

Week Four:

- Present your results to the class and Dr. Hiller in the final meeting (**Dec 10th**)
- Final paper: 3 – 5 pages (not including appendices), **due December 11th**.

Project content:

1. *Collect appropriate sequences:* Decide which sequences and species to include in your data set. Consider genomes from other Streptococcus species and from more distantly related bacteria. You will have to find the right compromise between thorough sampling and limiting computational and conceptual complexity. Avoid bias as much as possible by sampling from a broad set of taxa and keep in mind the importance of outgroups. Beware of paralogs and domain-based matches.

2. *Multiple sequence alignment:* Try several different programs to align your data and experiment with different parameters, substitution matrices, etc. Consider refining your alignment manually.

3. *Phylogeny estimation:* Consider which algorithmic strategy is appropriate for your data set, given the age and conservation of the sequences. You will want to use of TOPALI or a similar program to select the best sequence evolution model. As above, try several different programs and check to see if the results agree.

4. *Assess the reliability of your tree:* Bootstrapping can be used to assess the reliability of individual branches in the tree. In a probabilistic framework, various approaches for testing the reliability of the entire tree are also available. You can also assess the reliability of your tree by comparing its "predictions" with known biological relationships.

5. *Reconcile your tree(s)* with a bacterial species tree to infer gene duplications and horizontal transfers.
6. *Interpret results:* Discuss your results in light of the original question. What conclusions can be drawn from your analysis? You will also want to discuss the limitations of your

results. Given your assessment of reliability, how robust are your conclusions? If you had more time, what additional tests would you perform?

Deliverables:

Prepare proposed sequence set for evaluation in class: **November 19th**

- Bring your laptop to class, with the blast output associated with each of your searches with your assigned sequence open in a browser. To prepare for this, you should conduct at least one search of the nucleic acid data base and at least one amino acid data base, using your assigned sequence as a query.
- A proposed list of sequences for a DNA-based phylogenetic analysis.
- A proposed list of amino acid sequences for a protein-based phylogenetic analysis.

Final presentation: in class **December 10th**

Your final presentation should include 5 - 9 slides and cover:

- Phylogenetic distribution
- MSAs, if tricky
- Gene trees
- Reconciliation
- Pitfalls

Final paper: 3 – 5 pages (not including appendices), **due December 11th**

Your paper should include the following sections: Introduction, Methods, Results/Discussion, an appendix, and references. Given the nature of this project, the methods section will be a substantial part of your paper. The key goal is to demonstrate what you have learned about using phylogenetics to investigate a biological question, the relative merits of various phylogenetic methods, and which methods to use in a given context. Your final paper should include an appendix with supplementary figures, raw data, alignments and trees. You must also turn in fasta files with your unligned sequences, fasta files with your multiple sequence alignments (before and after trimming), as well as files containing the trees you inferred in newick format and reconciled trees in notung format.