

Cataloguing sequences homologous to the *Rhodobacter flagellar motor*

Supervisors: Judy Armitage & Jotun Hein

Motivation. Firstly, the coming years will see the number of known bacterial genomes go into the thousands, which gives a unique chance to apply comparative annotation on a massive scale, which will allow much more ambitious goals and also create computational challenges. Secondly, genomes will increasingly be supplemented with higher level data like structures, inferred networks, interaction data and kinetic parameters, that must also be compared across species, making the activity multilevel. Current genome annotation often classifies genes into pathways on the basis of the best understood homologue but many species have multiple homologues of well characterised pathways, which may have different functions within a bacterial cell. A major motivation behind this project is to develop searching algorithms to identify not only the structural but also the regulatory pathways and the fingerprints relating to control within these homologous pathways. Finally, we have an initial specific object of interest (the flagellar motor of *Rhodobacter sphaeroides*) and all annotation will be directed towards this, thus making it focussed.

The questions that can be addressed by pure sequence/genome data and of the focus of this project are:

- The strength of selection of different regions/nucleotides
- Annotation of regulatory signals by comparisons of closely related sequences
- Which protein structures (known through their gene) are conserved through which bacteria?

More detailed analysis also incorporating other data types, could be the topic for a DPhil.

Goal: To catalogue the existing genome and sequence information available for the study of the *R. sphaeroides* flagellar motor as function of sequence distance, sequencing time and function. Getting an overview of the available relevant data is a prerequisite to using the information optimally. Tabulating as a function of distance is valuable as different distances are optimal for different purposes, as function of sequencing time is informative as it will give a feeling for how strong the growth in the data is and finally function is important to trace overall changes in functionality. Finally in detail investigation of a single gene will yield insights into how much will be gained when analyzing so much data simultaneously. The main techniques needed for this project is sequence comparison, phylogeny reconstruction and annotation algorithms.

Background: A central object of study for the Oxford Centre for Integrative Systems Biology (OCISB) is the flagellar motor in *R. sphaeroides*. Flagella are made up of multiple copies of over 20 proteins, with over 50 genes involved in the assembly of the flagellum. The system is closely related to, and may share an evolutionary origin with, the Type III secretion system used by pathogens to deliver toxins directly into host cells. Many bacteria have both systems. The genes in general, but not always, are arranged in high order operons (regulons). Sequencing suggests many species have more than one set of flagellar genes, although whether both are expressed and when is unclear. Given the universality of the bacterial flagellum, this is an ideal complex structure to develop searching algorithms to identify not only the structural but the regulatory pathway and the fingerprints relating to that control.

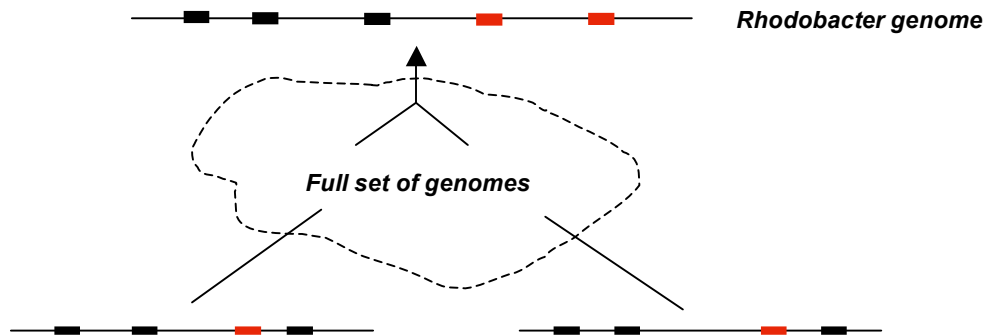
A small example is the sigma 54 regulation in *R. sphaeroides*. *R. sphaeroides* has 4 sigma 54 homologues. Experimental work shows that while one regulates nitrogen metabolism, of the other 3 one regulates flagellar expression and another chemotaxis gene expression. What however regulates the activity of these 2 sigma 54s is unknown. Simple annotation would not identify this subtle but biologically highly significant action, but kinetic analysis of very large data sets, with experimental confirmation, will identify the key components behind such a network. Bacteria move using flagella or pili (gliding). The 2 completely different systems are both controlled by the best understood sensory system in biology, the bacterial chemosensory system. All the chemotaxis proteins known, and the copy number, crystal structures and kinetics have been measured in *E.coli*.

Again genome analysis suggests that most species have multiple sets of chemosensory genes encoding more than one pathway. Experimental work however suggests that while some are involved in chemotaxis, others are involved in e.g. development or pathogenicity. However, most are historically designated "chemotaxis" operons. Given we cannot culture most species, we need a reliable mechanism to identify which pathway regulates the activity of the flagella motor and which regulates something else.

The Project

The present project will yield points in the genome that can be traced phylogenetically through speciations, duplications and deletions. Tracing each homologous set of points will define a set of regions, where combined phylogeny, alignment and annotation should be performed.

Ideally alignment and phylogenies are treated together and this has been done in many methods going back to Sankoff (1973). However, this can only be done correctly for up to 5-7 sequences, after that one will have to resort to heuristics, where there is not guarantee of an optimal solution, but that will hopefully perform well. It is crucial to design the applied heuristic well. An extreme reduction to the phylogenetic problem is illustrated to the left to in the in HMM profile alignment (Krogh et al., 1994), where all sequences are viewed symmetrically as generated by the same HMM.



Genomic alignments are all based on heuristic algorithms that typically takes sets anchors – that each align small regions in two genomes – and extend these to complete pairwise and multiple alignments. Embedded in such multiple genomic alignments are regions of interest that can be traced from the Rhodobacter genome into the remaining genomes through events like insertions, deletions, substitutions and full region loss/duplications.

Project Plan:

- Read key articles on the flagellar motor (Wadhams and Armitage,...), bacterial bioinformatics (Binneweis, Stothard and Wishart,...) and sequence analysis (Krogh, Sankoff,...).
- Make global phylogeny of bacteria. This can be done with programs such as PHYLIP, PAUP and a series of others. Given the very large size of this phylogeny, methods of to reduce the presentation must be used.
- For the set of sequences of interest investigate the number of homologous as function of evolutionary distance of genomes.
- Chose a single sequence from the flagellar motor set, find its gene and annotate this gene maximally for gene structure, signals and rate of evolution.

Comments: i. This project is very suited as preparation for a complete DPhil with focus on the flagellar motor in Rhodobacter.
 ii. It could be extended to a collaboration project if additional information involving structures and networks were also catalogued. At the protein structure level: To find sets of determined structures from bacterial genomes at suitable evolutionary distance to allow their use in homology modelling with the Rhodobacter motor as target. At the regulatory network, interaction and kinetic parameter level: Use information from other bacterial species to annotated the Rhodobacter motor and its regulation.

References:

- Binneweis, TT et al. (2006) "Ten years of bacterial genome sequencing: comparative-genomics-based discoveries" *Funct. Integr Genomics* 6:165-186
- Eddy SR (2002) Computational genomics of non-coding RNA genes. *Cell*. 2002 Apr 19;109(2):137-40. Review.
- Hobolth A, Jensen JL (2005) Applications of hidden Markov models for characterization of homologous DNA sequences with a common gene. *J. Comput. Biol.* 12(2):186-203.
- A. Krogh, M. Brown, I.S. Mian, K. Sjolander and D. Haussler (1994) *Hidden Markov models in computational biology*, *J. Mol. Biol.*, 235, 1501-1531.
- Overbeek, R., et al. (2007) "Annotation of Bacterial and Archeal Genomes: Improving Accuracy and Consistency" *Chem. Rev.* 1007.3431-3447
- Pallen and Matzke (2006) Origin of Species to the origin of bacterial flagella. *Nat Rev Microbiol.*4(10):784-90.
- Price, Dehal and Arkin (2007) "Orthologous Transcription Factors in Bacteria Have Different Functions and Regulate Different Genes" *PLOS Compu. Biol.* 3.9:1739-50
- D Sankoff (1995) Minimal Mutation Trees of Sequences. *SIAM Journal on Applied Mathematics*, Vol. 28, No. 1, 35-42.
- Stothard and Wishart (2006) "Automated bacterial genome analysis and annotation?" *Curr.Opin.Microbiol.*9:505-510
- Ulrich, L. E., E. V. Koonin and I. B. Zhulin. 2005. One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol.* 13: 52-56.
- Wadhams, G.H. and Armitage, J.P. (2004) Making sense of it all: Bacterial chemotaxis. *Nature Revs Mol Cell Biol.* 5: 1024-1037
- Wuichet, K. and I. B. Zhulin. 2003. Molecular evolution of sensory domains in cyanobacterialchemoreceptors. *Trends Microbiol.* 11: 200-203.
- Yu, GX et al. (2007) "A Versatile computational pipeline for bacterial genome annotation improvement and comparative analysis wit Brucellas as use case. *Nuc. Ac. Res.* 35.12.3952-62.