# SLIDE 1

Hello, I’m David Teixeira and my thesis project is to develop a gene regulatory network for the bacteria Pseudomonas Putida. My supervisor is Leighton Pritchard, hes been a great help so far when I ran into issues.

# SLIDE 2

So what is a gene regulatory network?

Genes are the regions of DNA which encode a product (Protein, RNA,…)

Intergenic regions are all the non-coding regions between genes, they contain the regulatory sequences that control gene expression by binding transcription factors.

Transcription factors and their regulators are gene products, therefore genes influence gene expression, the products of which affect gene expression in a constantly changing and connected system

The GRN visualises the knock-on effect changes to one genes expression may have on all others, either directly due to the gene encoding a TF that regulates other genes, or indirectly by the gene encoding another regulator.

Since bacteria group genes into operons, I have done the same with the networks to reduce clutter.

# SLIDE 3

Now, why make one for Pseudomonas putida?

Firstly, P.putida is a gram negative bacillus, or rod shaped, species of bacteria.

It’s very diverse with over 100 strains found in environments around the world.

This is because the strains contain extra “accessory” genes that enable survival in their environment, alongside the core genes crucial to cell function

Its commonly used in labs and industry for natural product and protein manufacture due to ease of culture and diverse metabolism.

And Its an opportunistic pathogen, albeit an uncommon one, causing sepsis and bacteremia

For those reasons, this project aims to develop incomplete gene regulatory networks produced over an internship this summer. The results of this could serve to optimise industrial/lab use, allow phylogeny study of strains, and act as a framework for computer models of the cell.

# SLIDE 4

This is just a summary of all the work behind the project so far. In outlined in red is the portion I did during the internship, it won’t be discussed.

The internship took 222 NCBI sourced genomes and TF binding data for 52 TFs from CollecTF and through a series of selfmade and sourced programs, ended with partially complete GRNs with IGRs linked to operon(s) they regulate, and operons to IGRs they encode a TF for.

Some foolish intern, definitely not me, made an error coding the final script that links operons to IGRs in the network, so this was corrected when the thesis began, hence it being outside the red border.

# SLIDE 5

So unfortunately, there isn’t much data on transcription factor binding sites on Pputida, so I went for a different approach.

PIGGY was previously used to identify all IGRs in all genomes, these IGRs aren’t *only* the regulatory sequences to which TFs bind but *the entire sequence* between two genes, so two IGRs may bind the same TFs but be classed differently due to differences in the non-regulatory parts.

TF to DNA binding is highly specific with TFs only binding specific sequences – these sequences must be conserved between IGRs that bind the same TF, otherwise the TF couldn’t bind, these sequences are a form of motif.

Logically speaking, functionally similar IGRs should bear the same features, and conversely IGRs bearing the same features should be functionally similar

This was the logic behind motif comparison – A tool called MEME searches the IGR sequences of a given genome and lists common recurring features(motifs), then a tool called MAST finds those motifs in the IGR sequences

If the motif composition of two IGRs were similar, they were considered the same functionally and merged on the graph, you’ll see what I mean by that in a second.

# SLIDE 6

This is just showing the overall process. A script collated the IGR sequences in multifasta format for each genome, used meme to find 50 motifs in that file, then used mast to find those motifs in each IGR sequences.

The resulting mast file contained the motif structure as shown underneath MOTIF DIAGRAM, the numbers in square brackets are the motifs, the other numbers are the length of the interspersing region.

A self-made python script would then compare each motif diagram to eachother (length wise, order, shared motifs…) and those deemed similar would be merged on the internship networks, forming what I called the MAST\_updated networks

# SLIDE 7

As a side-project, I’ve also made tools to study the typical IGRs associated with each gene per genome, and conversely to study the typical genes found in operons associated to each IGR.

Doing this, I can hopefully get an idea of how differing IGR to gene association can cause phenotypic differences in otherwise highly genetically similar putida strains, and why the same environmental factors in different strains cause different changes to gene expression.

At the time I submitted my presentation, I didn’t have results yet, but I have some now.

# SLIDE 8

Back to Motif analysis, this image shows what I meant by node merging. In this genome, these four IGRs (which are labelled as cluster followed by unique number) were found to have a similar motif composition which can be seen above them, meaning most of the motifs were shared, in the same order, and with similar interspersing regions.

In the previous graph, each cluster only linked onto a single operon, but as a result of my Motif\_comparator script, they were combined into a single node (with a horrendously long name) that links onto 4 operons.

# SLIDE 9

This figure shows how the motif comparison and merge affected the networks on a larger scale. Note that the networks themselves are so large I can only show a fraction of the whole thing for a single genome.

Above and below are before and after comparison respectively. In the before, I chose to only show the network sections where there were atleast 5 nodes linked (igr + operon)

In the after section, I showed the same regions as shown in the above in the same position, as well as any other feature that contained a merged IGR node.

On average this reduced the number of IGR nodes by 20-150

# SLIDE 10

Unfortunately I don’t have any results for the IGR association analysis yet, I only got the code working around the 8th so I had to submit the presentation.

So to review,

The motif comparison definitely reduced the amount of network clutter and introduced some interesting “hubs” of operon regulation, the accuracy of which is unverified but may be so by sequence comparison of merged IGRs. Unfortunately, the networks are still very discontinuous, with thousands of IGR-operon pairs not linking onto anything.

The tools for IGR-association analysis have been produced, but results will take time due to needing to manually review the files produced by said tools, or developing a complex script that can do so for me but I think it would be difficult for myself

The most obvious next step would be to obtain more TF binding site data for pseudomonas putida, during the internship I used everything I could get from a database called collecTF, and have not found another viable source.

Beyond that each genome contained over 2000 unidentified genes – ID’ing those, or atleast getting an idea on the function, would help graph development (Some might be TFs, others might indirectly affect gene expression)

Theres also the fact that regulation comes in 2 main forms – positive and negative – which currently my networks do not account for. This doesn’t currently limit network development, but it will need to be accounted for if these networks are ever to be used.

Finally as I stated theres the IGR-gene association analysis, which is likely to be my first next step given I managed to make the tools.

Thank you for your time, I hope I didn’t ramble or stutter too much, are there any questions.