Metabolic modeling of Streptomyces and its relatives

Alam, Mohammad Tauqeer

Submitted as: **Alam MT**, Takano E, Breitling R (2011) : *Prioritizing orphan proteins for further study using phylogenomics and gene expression profiles in* Streptomyces coelicolor.

Chapter 6

Prioritizing orphan proteins for further study using phylogenomics and gene expression profiles in *Streptomyces coelicolor*

ABSTRACT

Background: Streptomyces coelicolor*, a model organism of antibiotic-producing bacteria, has one of the largest genomes of the bacterial kingdom, including 7825 predicted protein coding genes. Of these genes, a large number, more than 30%, are functionally orphan, i.e. they are encoding hypothetical proteins with unknown function. However, many of these functional orphan genes show interesting gene expression dynamics in large-scale transcriptome analyses.*

Results: *Here we present a new algorithm combining time-course gene expression datasets and comprehensive phylogenomic information to identify a list of high-priority orphan genes, which show the highest level of aggregated evidence of being biologically important. These genes are the ones most generally conserved and showing the most informative expression dynamics along the time course. They often feature conserved neighboring genes as well.*

Conclusions: *The identified high-priority orphan genes are promising candidates to be examined experimentally for further elucidation of their function.*

6.1 Introduction

Here we present an analysis of orphan genes (hypothetical genes with unknown function) in the *Streptomyces coelicolor* genome, combining gene expression analysis and comparative genomics. The aim is to prioritize orphan

genes for further study. In our gene expression studies (Nieselt et al., 2010; Alam et al., 2010b), we frequently encountered genes that showed interesting expression patterns, but had no known function. To identify which of these genes merit in-depth experimental analysis, we developed a strategy for prioritizing protein encoding genes for additional characterization, combining phylogenomic information (Alam et al., 2010a) (i.e. the level of evolutionary conservation of each protein), and gene expression data from a large gene expression time series (Nieselt et al., 2010). We postulate that widely conserved proteins that show a physiologically relevant dynamic expression pattern are the most promising candidates for further experimental study, e.g. using gene overexpression and knock-out or knock-down approaches.

The functional annotation of orphan genes is not only relevant for its basic biological interest, but is also an important help for the improvement of genome-scale metabolic models based on genome annotation. These models in their initial form almost always contain gaps that need to be filled by manual curation or automated gap-filling algorithms that add missing essential metabolic activities to the models (Thiele and Palsson, 2010; Henry et al., 2010; Kumar et al., 2007; Alam et al., 2010a; Medema et al., 2010).

During our previous studies of genome-scale metabolic models of *Streptomyces coelicolor* and its relatives, we regularly had to postulate enzymatic functions that had not been assigned to specific proteins in the organisms (Alam et al., 2010a; Medema et al., 2010, 2011b). Assigning specific enzyme-coding genes to these orphan metabolic activities is very important for the subsequent analysis and interpretation of the models, and several approaches have been developed to assign sequences to the orphan metabolic activities: they employ, for example, mRNA co-expression analysis (Kharchenko et al., 2004), phylogenetic profile information (Jothi et al., 2007; Chen and Vitkup, 2006; Snitkin et al., 2006), pattern recognition techniques (Cuff et al., 2009) or comparative genomics (Osterman and Overbeek, 2003). These approaches are organism specific and have mostly been employed for well-studied model organisms such as *Escherichia coli* and *Saccharomyces cerevisiae*.

6.2 Results and Discussion

Of the 7825 predicted protein coding genes in the *Streptomyces coelicolor* genome (Bentley et al., 2002), 2688 (34%) are coding for functionally orphan proteins, i.e. proteins that are annotated as "hypothetical protein", "conserved protein", "putative membrane protein" or "putative secreted protein". Of these orphan proteins, 27 are conserved in all and 384 are present in at least half (22/44) of the 44 analyzed complete actinomycete genomes (see Methods section for a complete species list). 686 orphan proteins are present in at least 11 (25%) and 179 are conserved in at least 33 (75%) actinomycete genomes.

Of the 384 generally conserved actinomycete orphan proteins (i.e., those that are present in at least half of the analyzed genomes), 27 are also encoded in all species in a representative set of five non-actinomycete bacterial genomes (*Bacillus subtilis, Escherichia coli* K12, *Lactobacillus plantarum* WCFS1, *Staphylococcus aureus*, and *Streptococcus pneumonia* AP200), and 73 are present in at least half of the representative bacterial genomes.

Of these 76 ultra-highly conserved bacterial orphan genes, 24 also have putative homologues (reciprocal best BLAST hits) in at least half of the species in a representative set of eight non-bacterial genomes (including the eukaryotes *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Plasmodium falciparum*, *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Homo sapiens*, and the archaea *Haloterrigena turkmenica* and *Methanosarcina acetivorans*). These proteins are therefore almost universally conserved; however, although there seems to be significant conservation of some orphan proteins, none of them is truly universal, i.e. none has a putative homologue in all of the 58 studied genomes. This is most likely due to the fact that some of the included genomes are highly reduced, as a result of the parasitic lifestyle of the organism.

To prioritize the orphan proteins for further characterization, we therefore summarized the phylogenomic information (i.e. the level of evolutionary conservation of each protein) in a single "conservation" score, which expresses the degree of conservation across the three domains examined (acti-

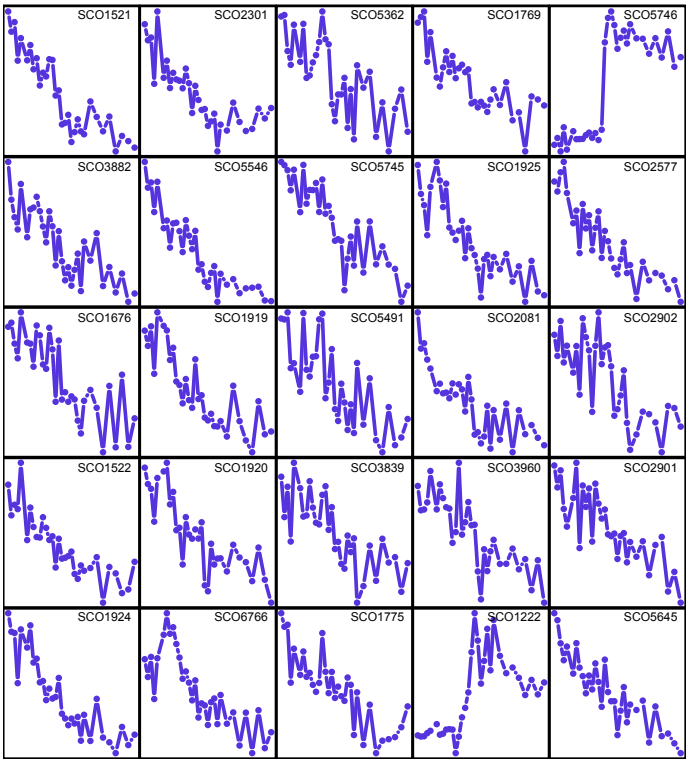

Figure 6.1: *Average expression profile of the top 25 candidate orphan genes.*

nomycetes, bacteria, non-bacteria). This score was combined with a second measure of expression dynamics across a large gene expression time series studying the metabolic switch caused by phosphate starvation. The "expression dynamics" score described in the Methods section identifies genes that show a smooth expression trend across (part of) the time series and favors those genes that show a particularly strong (step-like) expression change at one time point. This is intended to allow to focus on genes that are not only passively following the expression change during nutrient depletion but that show evidence for active regulation, which is indicative of a central function in cellular physiology. Based on the p-value of the "expression dynamics" score, we assigned a rank to each gene, and averaged this value with the rank of the "conservation" score.

Using the averaged conservation and expression dynamics rank, we arrived at a list of 30 top orphan proteins. These were examined in more detail to determine if their function was really unknown: we checked the most recent versions of the Uniprot (The UniProt–Consortium, 2009) and StrepDB database for annotations, performed a PSI-BLAST against the Uniprot database, compared the annotation of the homologs in *E. coli*, yeast and human where these were available, and analyzed the domain architecture using SMART tool (Simple Modular Architecture Research Tool) (Schultz et al., 1998). Using this information, we asked three microbiologist and bioinformaticians to independently score the genes according to their "orphanicity", i.e. their confidence in the absence of a known potential function. The average score of the three raters was combined with the average score of the conservation and expression dynamics to arrive at a final ranking for the most interesting orphan genes for further study: the top genes are those for which we have absolutely no information about their function, that are ultra-highly conserved across species, and show a highly significant dynamics in their gene expression (Table 6.1).

Based on the gene expression profiles (Figure 6.1), the candidate genes SCO5746 and SCO1222 are particularly interesting: they show a very strong switch upon phosphate starvation, and their expression increases upon entry into the stationary phase, similar to the expression pattern of the antibiotic biosynthesis gene clusters *act* and *red*. All other high-priority genes show a decrease of expression along the time course. SCO5746 has a putative uncharacterized homolog in *E. coli* and contains a domain of the DegT/DnrJ-/EryC1/StrS aminotransferase family. The aminotransferase activity was demonstrated for purified StsC protein, which acts as an L-glutamine:scyllo-inosose aminotransferase and catalyses the first amino acid transfer in the biosynthesis of the streptidine subunit of antibiotic streptomycin. It is therefore tempting to speculate that the SCO5746 gene has some role in the biosynthesis of a new antibiotic in *S. coelicolor* as well, and the same might be the case for the completely uncharacterized SCO1222. The closest putative antibiotic biosynthesis clusters are SCO5799-SCO5801 (siderophore synthetase type) and SCO1206-SCO1208 (chalcone synthetase type), both of which seem

Figure 6.2: *This figure shows annotation conservation of the neighbors of orphan genes in four sequenced* Streptomyces *genomes. The conserved orphan gene is shown in the centre, and the two neighbors on each side are shown in the form of arrows. Each arrow has four sections, corresponding to the four* Streptomyces *species:* S. coelicolor, S. avermitilis, S. griseus *and* S. scabies. *They are colored in blue where the annotation matches that of* S. coelicolor. *The annotation of the* S. coelicolor *homolog is listed above each gene if it is conserved in at least one of the other species; if at least two of the other species share another annotation, this is listed in brackets.*

Table 6.1: **Top 30 orphan proteins for further study.***The proteins are prioritized according to their conservation across actinomycetes, bacteria and non-bacteria; their expression dynamics (summarized in the p-value); and their orphanicity, i.e. the absence of any functional information.*

Gene Name	Annotation	Final rank	Orphanicity rank	Exp. quantile	p-value	act	bac	non-bac
SCO1521	hypothetical protein	1	1	0.21	3.71E-10	44	5	5
SCO2301	hypothetical protein	6	4	0.34	3.27E-07	43	5	5
SCO5362	hypothetical protein	6.5	9	0.13	2.02E-07	44	4	7
SCO1769	hypothetical protein	8	5	0.12	3.08E-08	40	3	1
SCO5746	hypothetical protein	8	7	0.18	4.38E-18	20	3	1
SCO3882	hypothetical protein	8.5	2	0.18	6.71E-08	38	5	1
SCO5546	hypothetical protein	8.5	14	0.35	7.62E-09	42	3	6
SCO5745	hypothetical protein	9.5	17	0.02	9.49E-10	43	4	6
SCO1925	hypothetical protein	11.5	18	0.09	1.24E-07	44	5	3
SCO2577	hypothetical protein	12	3	0.64	2.66E-07	41	5	1
SCO1676	hypothetical protein	12.5	15	0.32	7.05E-09	31	1	4
SCO1919	hypothetical protein	12.5	11	0.16	5.74E-07	44	4	2
SCO5491	hypothetical protein	12.5	6	0.35	3.07E-07	32	3	3
SCO2081	hypothetical protein	13	8	0.60	2.88E-08	38	2	1
SCO2902	hypothetical protein	14.5	22	0.37	3.05E-07	43	5	4
SCO1522	hypothetical protein	15.5	19	0.19	6.47E-07	43	3	5
SCO1920	hypothetical protein	16	12	0.27	1.71E-06	42	5	5
SCO3839	hypothetical protein	16.5	27	0.35	1.60E-08	35	3	2
SCO3960	hypothetical protein	17.5	13	0.30	5.66E-08	29	5	1
SCO2901	hypothetical protein	18	23	0.36	5.37E-07	41	3	5
SCO1924	hypothetical protein	18.5	20	0.08	6.81E-08	44	1	2
SCO6766	hypothetical protein	18.5	10	0.19	4.55E-08	20	1	2
SCO1775	hypothetical protein	21	16	0.32	3.00E-06	42	4	2
SCO1222	hypothetical protein	22	21	0.43	3.33E-09	27	1	1
SCO5645	hypothetical protein	22	28	0.07	3.11E-07	36	4	2
SCO1530	hypothetical protein	24.5	24	0.03	8.99E-07	43	1	5
SCO2497	hypothetical protein	26.5	29	0.52	2.38E-06	37	5	7
SCO5787	hypothetical protein	27	26	0.12	5.88E-06	44	3	7
SCO2599	hypothetical protein	27.5	25	0.13	4.17E-07	44	1	1
SCO5711	hypothetical protein	29.5	30	0.12	8.65E-06	44	5	5

unlikely candidates for interacting with SCO5746 or SCO1222. However, it is possible that these genes contribute to a dispersed biosynthetic pathway,

not involving a dense genomic clustering.

Interestingly, we see a strong neighborhood conservation of most of the candidate orphan genes in other *Streptomyces* species (Figure 6.2). In some cases, the annotation of the neighbors does suggest at least a broad functional category: for example, SCO1521/1522 might be involved in DNA remodeling during recombination, as their conserved neighbors are a Holliday junction resolvase and DNA helicase (RuvABC complex); and SCO2081 might play a role in cell division, matching its conserved neighbor, the cell division protein ftsZ (Jakimowicz et al., 2005). However, most of the conserved neighbors are hypothetical proteins themselves and do not seem to immediately identify a putative function for most of the orphan genes; nonetheless, the neighborhood information will be valuable for the design and interpretation of the most efficient experimental perturbations.

6.3 Materials and methods

6.3.1 Genome sequence analysis

For the phylogenomic profiling, we studied the complete genome sequences of the 44 actinomycete species, that were also used in our earlier phylogenetic study (Alam et al., 2010a): *Arthrobacter aurescens* TC1, *Acidothermus cellulolyticus* 11B, *Bifidobacterium adolescentis* ATCC 15703, *Bifidobacterium longum* NCC2705, *Corynebacterium diphtheriae* NCTC 13129, *Corynebacterium efficiens* YS-314, *Corynebacterium glutamicum* ATCC 13032, *Corynebacterium jeikeium* K411, *Clavibacter michiganensis* subsp michiganensis NCPPB 382, *Frankia alni* ACN14a, *Frankia sp* CcI3, *Frankia sp* EAN1pec, *Kineococcus radiotolerans*, *Leifsonia xyli* subsp xyli str CTCB07, *Mycobacterium avium* subsp, *Paratuberculosis str* k10, *Mycobacterium avium* 104, *Mycobacterium bovis* BCG Pasteur 1173P2, *Mycobacterium bovis* subsp bovis AF2122 97, M*ycobacterium gilvum* PYR-GCK, *Mycobacterium sp* JLS, *Mycobacterium sp* KMS, *Mycobacterium leprae* TN, *Mycobacterium sp* MCS, *Mycobacterium tuberculosis* H37Ra, *Mycobacterium smegmatis* str MC2155, *Mycobacterium tuberculosis* CDC1551, *Mycobacterium tuberculosis* F11, *Mycobacterium tuberculosis* H37Rv, *Mycobacterium ulcerans* Agy99,

Mycobacterium vanbaalenii PYR-1, *Nocardioides sp* JS614, *Nocardia farcinica* IFM 10152, *Propionibacterium acnes* KPA171202, *Rhodococcus sp* RHA1, *Renibacterium salmoninarum* ATCC 33209, *Salinispora arenicola* CNS 205, *Streptomyces avermitilis* MA 4680, *Saccharopolyspora erythraea* NRRL 2338, *Streptomyces griseus* strain IFO13350, *Streptomyces scabies* strain 8722, *Salinispora tropica* CNB 440, *Thermobifida fusca* YX, *Tropheryma whipplei* str Twist, *Tropheryma whipplei* TW08 27. This was complemented by the genomes of 6 eukaryotes (*Caenorhabditis elegans*, *Arabidopsis thaliana*, *Homo sapiens*, *Plasmodium falciparum* 3D7, *Drosophila melanogaster*, *Saccharomyces cerevisiae*), 2 archaea (*Haloterrigena turkmenica*, *Methanosarcina acetivorans*), and 5 other model bacteria from different taxonomical classes (*Bacillus subtilis*, *Escherichia coli* K12, *Lactobacillus plantarum* WCFS1, *Staphylococcus aureus*, *Streptococcus pneumonia* AP200). Putative homologs were identified as reciprocal best BLAST hits. The conservation score was calculated in three steps: (1) the genes were independently ranked according to the number of species of actinomycetes, other bacteria, and non-bacteria in which they have a putative homolog; (2) their ranks in the bacteria and non-bacteria lists were averaged; and (3) the resulting rank and the rank in the actinomycete list were averaged again to produce the final rank.

6.3.2 Gene expression data

Details about the gene expression dataset and experimental conditions can be found in (Nieselt et al., 2010; Alam et al., 2010b).

6.3.3 Dynamic expression detection

To identify genes that show a dynamic expression along the time course, and in particular genes that have a clear expression switch at one time point, we used the following iterative algorithm (in pseudocode):

Data: a vector v of gene expression data
Result: minPvalue, the p-value of the switch-like dynamic expression
minPvalue ← 1;
foreach *i in the set (*2 **to** *(length(v)* - 2*))* **do**
  j ← i + 1;
  MaxWindowSize ← min(i, length (v) - i);
  foreach *position p in the set ((i - MaxWindowSize + 1)* **to** *i - 1)* **do**
    q ← j + (i - p);
    Pvalue ← p-value of the t-test comparing v[p:i] and v[j:q];
    If (Pvalue < minPvalue) minPvalue ← Pvalue;
  end
end
return minPvalue;

Algorithm 1: Algorithm for dynamic expression switch detection

6.4 Conclusions

The aim of this paper was to prioritize protein coding orphan genes (i.e., genes encoding hypothetical proteins with unknown function) for further experimental characterization of their function. We combined two lines of evidence for this purpose: First, we developed an algorithm to score the most interesting dynamic switches in gene expression data. Second, we introduce a conservation score summarzing the level of evolutionary conservation across diverse domains (actinomycetes, other bacteria and non-bacteria). We combined the expression score and the conservation score, and identified a list of 30 high-priority orphan genes, which are promising candidates for future experimental study. In some of the cases, the neighboring genes of the candidate orphan genes show strong conservation and suggest at least a broad functional category for the candidate orphan genes.

6.5 Acknowledgments

M.T.A was funded by GBB scholarship, University of Groningen. E.T. was funded by a Rosalind Franklin Fellowship from the University of Groningen. R.B. is supported by an NWO-Vidi fellowship.