# Bioinformatics and Biological Databases

1) Sigma-54 Promoter Database – A database of sigma-54 promoters covering a wide range of bacterial genomes

2) ClusterMine360 – A database of PKS/NRPS biosynthesis

by

Kyle Conway

A thesis presented to the
University of Ottawa
in fulfillment of the
thesis requirement for the degree of
Masters in Chemistry

# Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

# Abstract

The Sigma-54 Promoter Database contains computationally predicted sigma-54 promoters from over 60 prokaryotic species. Organisms from all major phyla were analysed and results were made available online at http://www.sigma54.ca. This database is particularly unique due to its inclusion of intragenic regions, grouping of data by COG and COG category, and the ability to summarize results either by phylum or database-wide.

ClusterMine360 (http://www.clustermine360.ca/) is a database of microbial polyketide and non-ribosomal peptide gene clusters. It takes advantage of crowd-sourcing by allowing members of the community to make contributions while automation is used to help achieve high data consistency and quality. The database currently has over 200 gene clusters from over 185 compound families. It also features a unique sequence repository containing over 10,000 PKS/NRPS domains. The sequences are filterable and downloadable as individual or multiple sequence FASTA files. This database will be a useful resource for members of the PKS/NRPS research community enabling them to keep up with the growing number of sequenced gene clusters and rapidly mine these clusters for functional information.

# Acknowledgements

I would like to thank Dr Chris Boddy for his enthusiasm and ever-positive mindset. I would also like to thank Luis for enduring my constant grumbling.

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| A | adenylation domain |
| ACP | acyl-carrier protein |
| AT | acyl-transferase |
| ATP | adenosine triphosphate |
| C | condensation domain |
| CLF | chain length factor |
| COG | cluster of orthologous groups |
| DH | dehydratase domain |
| DNA | deoxyribonucleic acid |
| E | epimerization domain |
| ER | enoyl reductase domain |
| HMM | hidden Markov model |
| KR | keto-reductase |
| KS | keto-synthase |
| MeSH | medical subject headings |
| NCBI | National Center for Biotechnology Information |
| NP-hard | non-deterministic polynomial-time hard |
| NRPS | non-ribosomal peptide synthetase |
| ORF | open reading frame |
| PCP | peptidyl-carrier protein |
| PKS | polyketide synthase |
| RNA | ribonucleic acid |
| RNAP | RNA polymerase |
| TE | thioesterase |

# Chapter 1: Introduction

## 1.1 The Rapid Rise of DNA Sequencing

Over the past thirty years, the amount of data generated through DNA sequencing has grown exponentially. This is easily illustrated by the staggering increases in the amount of data stored in nucleotide databases (see Figure 1). Part of the reason the growth in data has been so enormous is that improvements in technology have enabled the sequencing of increasingly large DNA assemblies. For example, while only small RNA genes could be sequenced in the 1960s (1, 2), the 1970s saw technology that first allowed the sequencing of small lengths of DNA such as operator sequences (2), followed by gel based sequencing methods (2) that resulted in the sequencing of the first complete genome in 1977, that of bacteriophage ΦX174 (3). This trend continued into the 1980s and 1990s with the sequencing of ever larger viral, bacterial (4), and finally eukaryotic genomes (5–7) leading up to the first draft publication of the human genome in 2001 (8, 9).

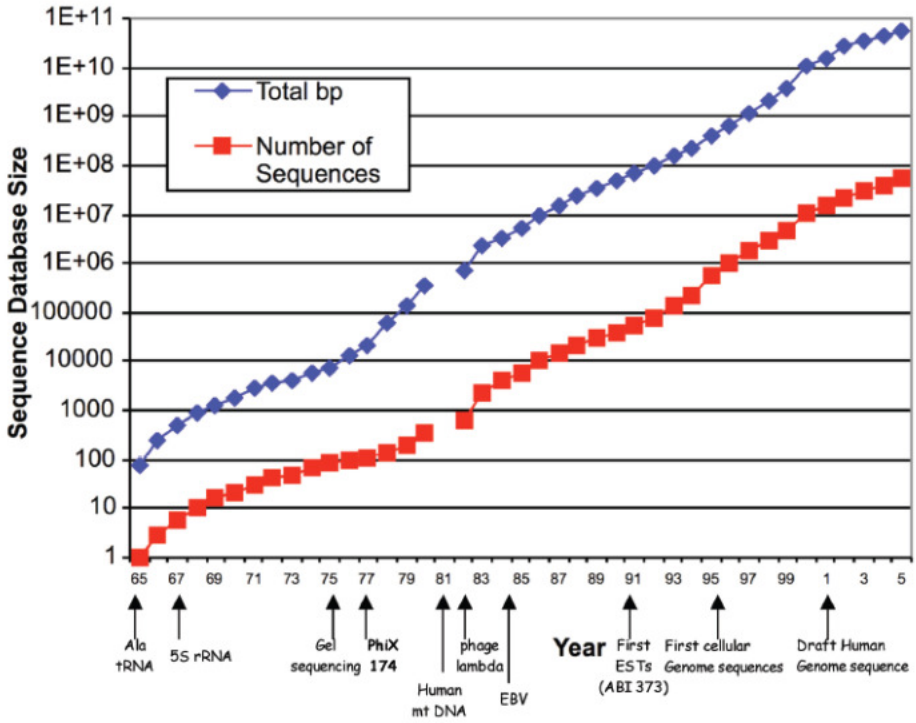



**Figure 1. Growth of nucleotide sequence databases.**
Data prior to 1981 are from the Dayhoff, Nucleotide Sequence Database, Vol 1, while those after 1981 are based on the sequences in GenBank (2). © Clyde A. Hutchison III and licensed for reuse under the Creative Commons Attribution Non-Commercial license (http://creativecommons.org/licenses/by-nc/2.0/uk/)

Since the turn of the millennium, improvements in technology have had dramatic impacts on sequencing costs. For example, around the year 2000, it cost several thousands of dollars to sequence a million base pairs of DNA. Today, the same amount of DNA now costs only approximately 10 cents to sequence representing a decrease of approximately 5 orders of magnitude (see Figure 2) (10).

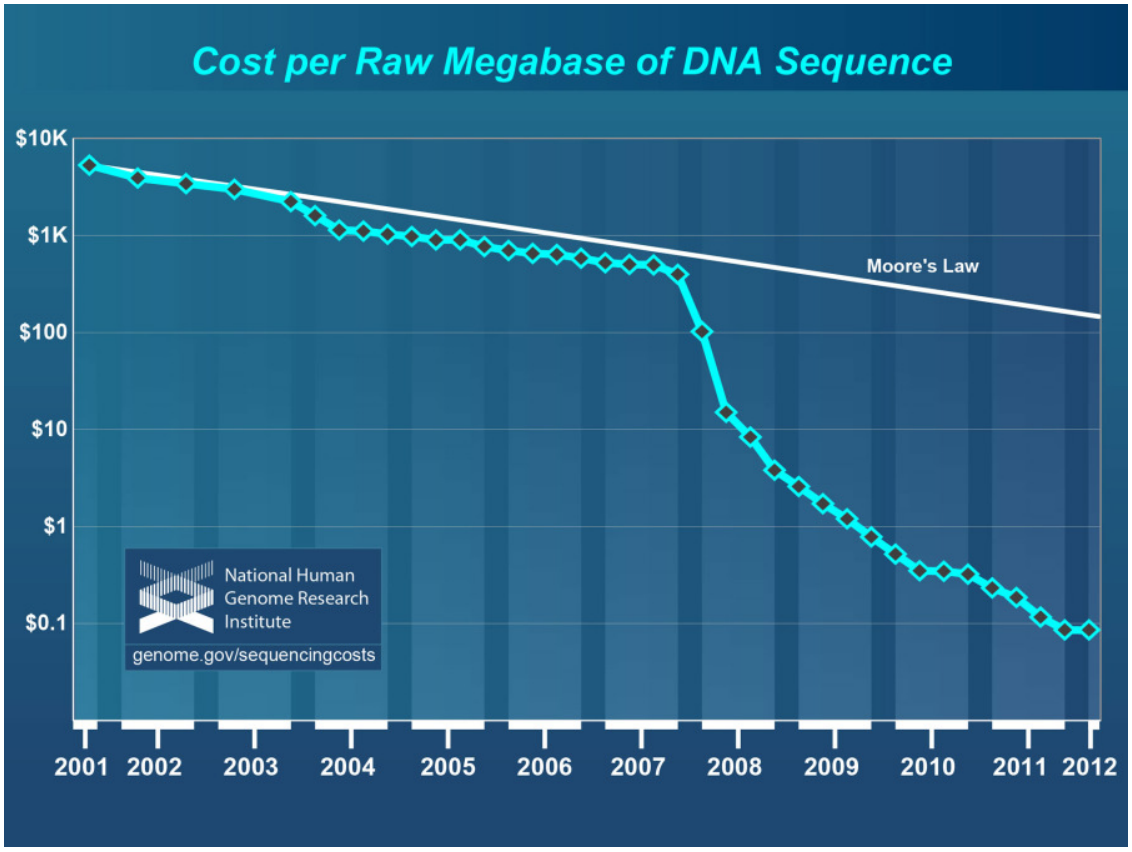



**Figure 2. Sequencing cost, over time, of 1 million base pairs of DNA sequence (10).**
Courtesy: National Human Genome Research Institute

Given the enormous decrease in sequencing costs, it has become affordable to do DNA sequencing on an immense scale. Indeed, with the dramatic decrease in costs, sequence data is now being used as a vital tool by researchers in a wide variety of disciplines such as molecular biology, medicine, chemistry, anthropology, physics, art, computer science, engineering and many more (11).

## 1.2 Importance of Bioinformatics

The rapidity with which large amounts of biological data can be generated is astounding. It, however, also creates a problem. The DNA sequence, by itself, is rarely useful. Some analysis needs to be undertaken to extract the information of interest from the raw sequence. Given the large amounts of raw sequence data, manual analysis is simply unmanageable (2). The analysis of large data sets is ideally suited for processing by computer algorithms (12). As such, the field of bioinformatics has grown in importance with the increasing amounts of biological data sets being generated (13).

### 1.2.1 Pattern Matching

One of the more important sub-disciplines of bioinformatics involves pattern identification algorithms (14). There are many approaches to pattern matching in biological systems but a full discussion is beyond the scope of this document. The focus, instead, will be on three frequently used methods: positionally weighted matrices, multiple sequence alignment, and hidden Markov models.

#### 1.2.1.1 Positionally Weighted Matrices

Positionally weighted matrices are frequently used to describe both conserved and variable elements involved in DNA-protein binding interactions and, in particular, the specificity of transcription factors (12). It is an extension of the concept of a consensus sequence. A simple consensus sequence, however, is often unable to capture the sequence variability that is inherent at the sites of DNA-protein interactions. This is particularly true in the promoter region since variations in sequence can allow for fine-tuned control of transcription. Therefore, due to the inherent variability, it is frequently difficult to come up with a simple consensus sequence that is not, to some extent, arbitrarily defined (12).

In the context of a DNA sequence, a positional weighted matrix is a matrix that has a value for each base at a possible site. To determine the overall score for a particular site, the sequence to be evaluated is aligned with the matrix and the score for the base at each position is summed to give a total score. For example, if the sequence GTCAT was evaluated using the matrix in Figure 3, the total score would be 175. The highest possible scoring sequence would be TCAAT with a score of 220 since this sequence matches the highest value at each possible site.

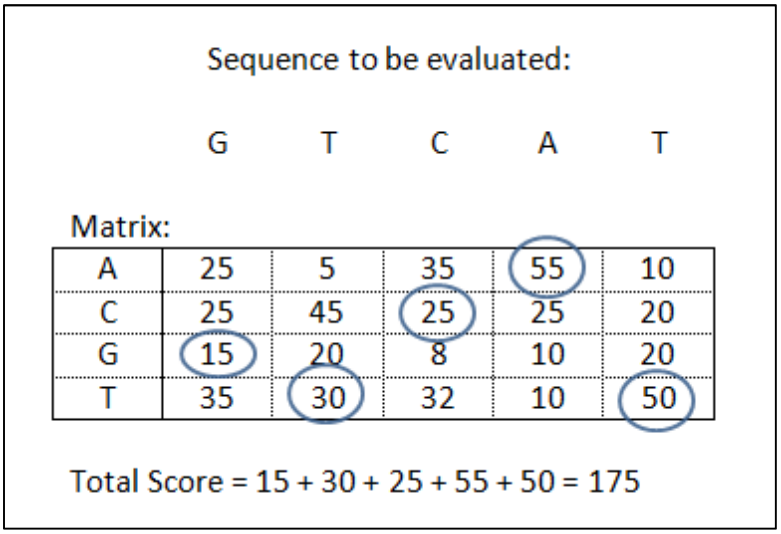



**Figure 3. Example of evaluating a sequence against a positionally weighted matrix using a simple scoring system.**
The sequence GTCAT is evaluated against the matrix. Each letter of the sequence is compared to its corresponding column in the matrix. The score for each position is assigned based the value of the matching base in the matrix. Therefore, when the first position is a G, a score of 15 is assigned since this is the value provided by the matrix. The values determined at each position are summed to give a total score.

While the simplest scoring system is used in Figure 3, positionally weighted matrices can be scored in more than one way. Using statistical mechanics, Berg and von Hippel showed that the log of the frequency of a base at a given position is proportional to the contribution of that base to the total binding energy (12, 15).

$$Contribution\ to\ Total\ Binding\ Energy \propto log_2 f_{b,i}$$

**Figure 4. The log frequency of the base at each position is proportional to its contribution to binding energy.**
$f_{b,i}$ is the frequency of the base, *b*, at a given position, *i*, in the matrix.

It follows, therefore, that the total binding energy is proportional to the sum of the log frequencies for each of the bases at each position (15).

$$Total\ Binding\ Energy \propto \sum log_2\ f_{b,i}$$

**Figure 5. The sum of the contributions from the individual bases is proportional to the total binding energy.**
$f_{b,i}$ is the frequency of the base, *b*, at a given position, *i*, in the matrix.

This interpretation matches the predictions arising from information theory in which the information content for a given sequence is also proportional to sum of the logs of the base

frequencies. The formula in Figure 6 can be used to determine the information content of a sequence when evaluated against a matrix. The total information content, $I_{sec}$, also called the Kullback Liebler distance, does not correspond to any physically relevant value. However, it is related to the binding energy in that higher values of information content correspond to a greater probability of binding compared to random sites in a genome.

$$I_{\mathrm{seq}}(i) = \sum_{b} f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

**Figure 6. Formula for the calculation of the information content, also known as the Kullback-Liebler distance** (12)**.**
*i* is the site position, *b* refers to each possible base, $f_{b,i}$ is the observed frequency of each base at that position *i*, $p_b$ is the frequency of base *b* in the whole genome.

It is important to note the $p_b$ factor in the denominator adjusts for variations in genome base content. If the normalized frequency of the base, *b*, at the given position, *i*, is greater than the probability of selecting that base randomly based on the genome GC content, then it will add positively to the score. If it is lower than the random probability based on GC content, it will subtract from the score.

One of the problems associated with positional weighted matrix methods is that there is no standard, accepted way of determining a threshold value (12). There have been some algorithms that have been developed that attempt to relate a score with *p*-values but they are non-trivial to implement (16–18). Therefore, choosing a threshold score becomes somewhat arbitrary in an attempt to balance the number of false positives and false negatives.

These problems, however, are countered by several advantages. One of the biggest advantages is that it is relatively simple to implement while providing more accurate results than simply searching for a simple pairwise alignment (ie. a BLAST search) since the matrix based method better takes into account the variability of the consensus sequence (12). Additionally, the use of positionally weighted matrices for predicting protein DNA interactions has been shown to be supported by statistical mechanical theory (15).

#### 1.2.1.2 Multiple Sequence Alignment and Phylogenetic Trees

Multiple sequence alignment is another method that is frequently used in bioinformatics. This technique involves taking sequences that are thought to be related and arranging them in a matrix that maximizes their similarity. In doing so, it can help highlight functionally or structurally important residues (19). Performing a multiple sequence alignment, however, is not simple. In mathematical terms, it is NP-hard (20). There are algorithms that use dynamic

programming to find an optimal solution but these are generally not feasible due to time and memory constraints (19).

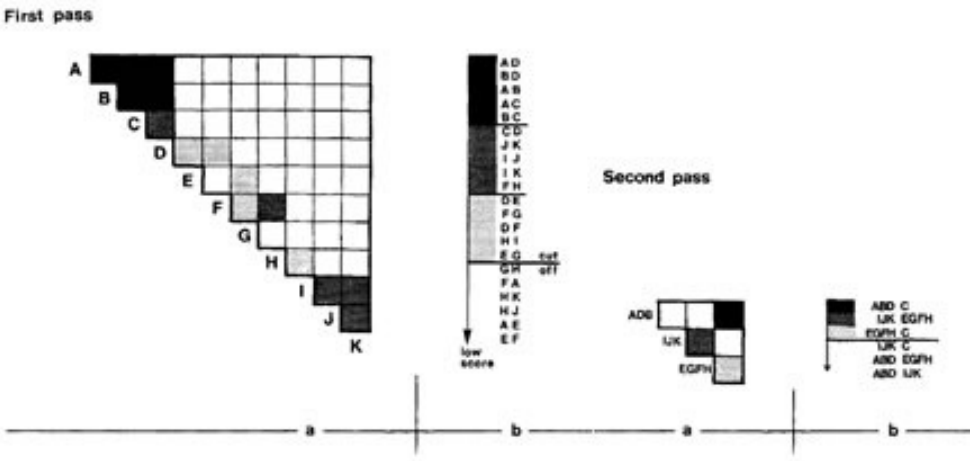



**Figure 7. Multiple sequence alignment by progressive alignment.**
In the first pass, each sequence is compared to every other sequence to determine a comparative similarity score (a). Sequences that are highly similar are indicated by dark squares in the matrix. The results are then ordered (b) and the most similar sequences are used to form a consensus sequence after which the process is repeated (Second pass). Modified from (21). Used with permission of the author.

An alternative heuristic approach that is frequently used is called progressive alignment. Some well-known implementations of this algorithm are Clustal (22), ClustalW (23), T-Coffee (24), and MUSCLE (25). This algorithm involves comparing each sequence against every other sequence in the query to find the two most related sequences. The consensus sequence of those two sequences is then used to find the next most related sequence. This repeats until all of the sequences are ordered (21) (see Figure 7).

The pairwise alignment method is not guaranteed to find the most optimal result. However, it is efficient and can be tuned to produce results with good accuracy (19). Another benefit of the pairwise alignment method is that its results can be used to create a phylogenetic tree (26).

A phylogenetic tree is a representation of the evolutionary relationships between a set of sequences (27). The multiple sequence alignment determines the similarity between each sequence and each group of sequences and a software program, such as MEGA5 (28), can use this information to infer and draw a phylogenetic tree (see Figure 8).

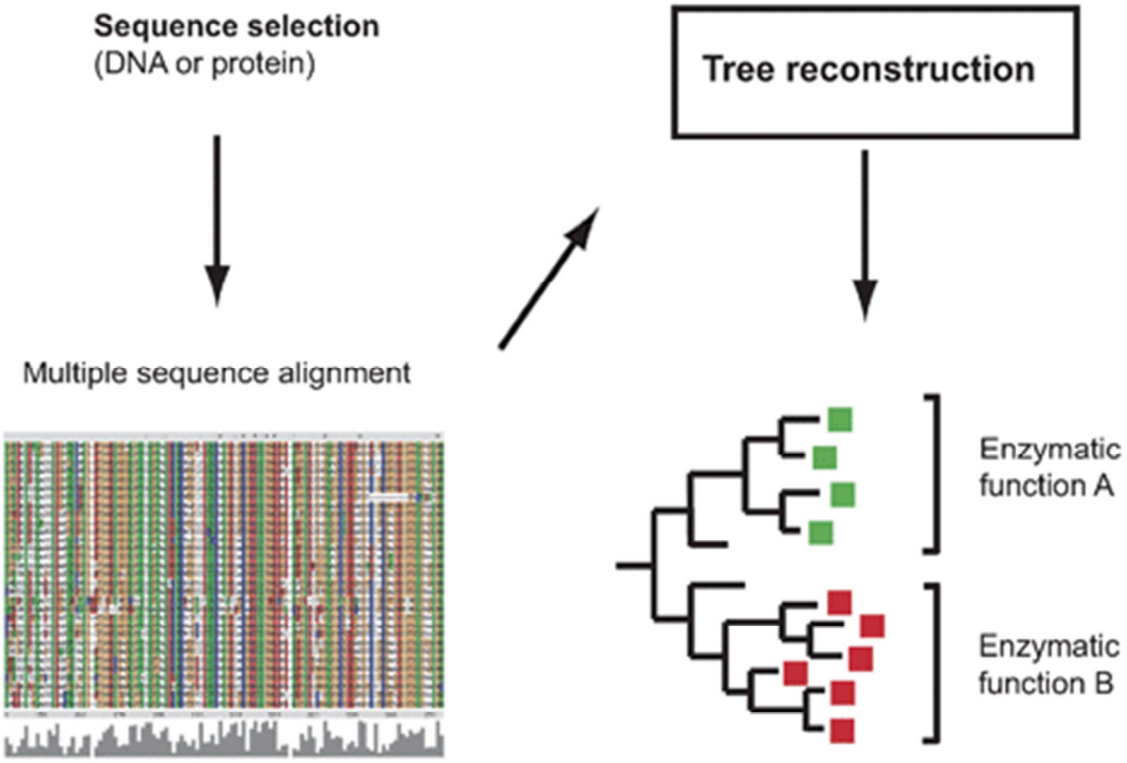



**Figure 8. Phylogenetic tree construction.**
Describes the steps involved in building a phylogenetic tree. First, the sequences are aligned using a Multiple Sequence Alignment algorithm. The aligned sequences are then supplied to a tree building algorithm with outputs the phylogenetic tree. Adapted from (29). Used with the permission of the author.

Phylogenetic trees can be useful in assigning protein function. By treeing a sequence of unknown function with a set of sequences with known functions, it is possible to assign a putative function to the sequence by determining the clade it groups with (29).

### 1.2.1.3 Hidden Markov Models

The mathematics of Markov models was first explored by Leonard Baum in the 1960s (30). While speech recognition was one of its first applications (31), hidden Markov models quickly emerged to be one of the most useful methods for finding patterns in biological systems (32).

A hidden Markov model (HMM) consists of a series of states and transitions. HMMs used to model biological sequences are $1^{st}$ order models which means that the states are linearly related. A special type of HMM, called a profile HMM, is generally used when modelling biological data. Profile HMMs have insert and delete states associated with each main state

and these can be used to model insertion or deletion gaps in the sequence (see Figure 9). The insert state will emit a letter of the alphabet while the delete state will not (it is 'mute') (32).

Each main state in the model has both a set of emission probabilities and a transition probabilities. The emission probabilities represents the likelihood of emitting a given alphabet letter at that position. The transition probabilities are the chances of remaining in the same state or of moving to one of the adjacent states. For example, there may be a 50% chance of staying in the same state, a 30% chance of moving to an insertion state, and a 20% chance of moving to the deletion state.

The end result of proceeding through the Markov model is a series of letters (or symbols) from a pre-defined alphabet. In the case of DNA, the HMM will have a 4 letter alphabet representing each of the bases: A,C,T,G. (33)

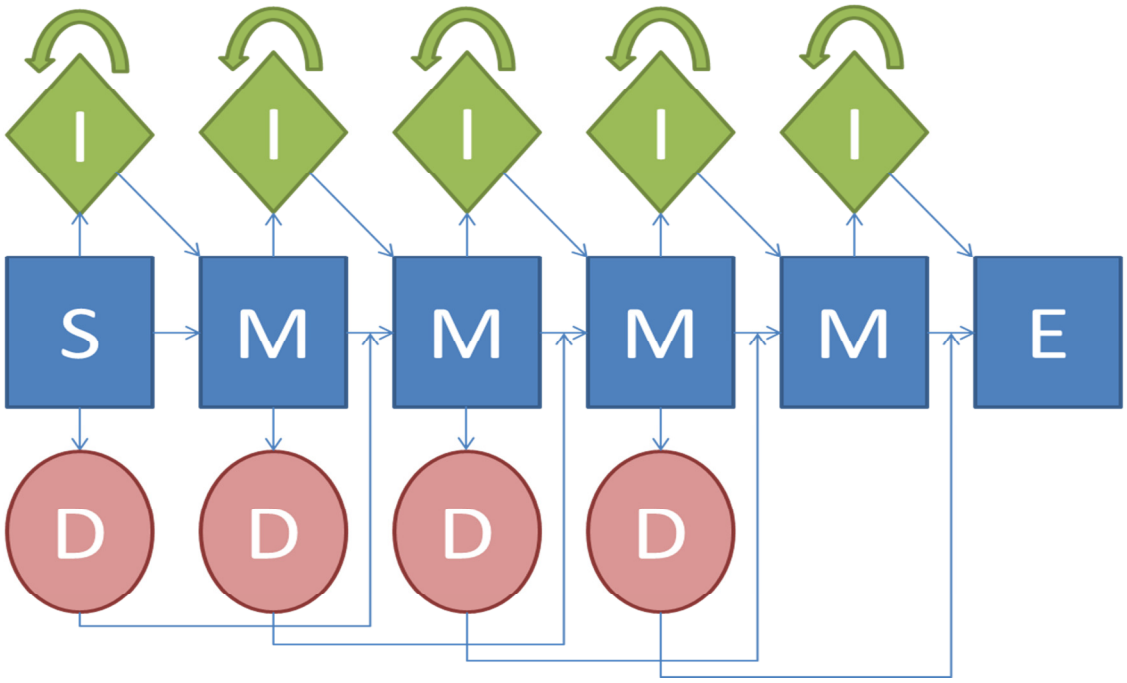



**Figure 9. Architecture of a 1st order profile hidden Markov model.**
The model starts at S, which is the starting state. From the start state, there is a probability for moving to an insert state, delete state, or the next main state which are represented by I, D and M, respectively. Each main state is associated with a probability that a symbol will be emitted. Each state also has a probability of moving on to one of the downstream states. The output is complete once the model has reached E, the end state. Diagram based on (32).

One of the benefits of hidden Markov models is that they can be trained from non-aligned sequences and can therefore discover patterns that simpler models might not pick up. There are a number of algorithms that can be used to train HMMs such as Baum-Welch, expectation maximization, gradient descent, genetic algorithms, Gibbs sampling, and simulated annealing (32).

Once a model has been obtained, sequences can be evaluated against the model to determine how likely they are to have been emitted from that model. The total probability of emitting a given sequence can be obtained by multiplying all of the emission and transition probabilities along the path. In the case of hidden Markov models, the path taken to emit those letters is not known. All that is known are the parameters and the end result. Therefore, since the path taken is not known, the total probability is the sum of all of the paths that could have emitted that sequence. This calculation is non-trivial and must be obtained using dynamic programming algorithms (32).

## 1.3 Bacterial Secondary Metabolite Biosynthesis

One area in which DNA sequencing and bioinformatics have had a major impact is the field of secondary metabolite biosynthesis. There is great interest in bacterial secondary metabolites as they have long been known to be an excellent source of molecules with significant biological and pharmacological activity. Up to 60% of known bioactive microbial metabolites demonstrate antibiotic activity, 25% display anti-tumor activity, and some 50% show other types of biological activity (34). It is no wonder then that most new drugs are either natural products or natural product derived (35). While these numbers are impressive, there are an almost innumerable number of microbial metabolites yet to be discovered (34). This large number of undiscovered metabolites is an excellent source of potential new drugs.

While it is believed that there are many metabolites that have not yet been identified, efforts to discover these new natural products have been hampered as most of the known bacterial species are not cultivatable (34). It is believed that the number of unknown bacterial species may be up to 250 times greater than the number of species that have been currently described. Many of these species are likely to produce novel secondary metabolites with diverse chemical structures. Uncultivatable bacteria thus represent one of the largest potential sources of new chemical diversity.

In order to access this chemical diversity, a better understanding of the pathways that produce these secondary metabolites is needed. For example, it is crucial to better know how these pathways are regulated so that they can be expressed under laboratory conditions. In addition, much of the sequence data on these pathways is scattered. Aggregation of this data would be highly useful in terms of looking for patterns that could lead to the discovery of previously undiscovered metabolic pathways and new chemical diversity.

### 1.3.1 Relationship between Sequence Data and Bacterial Natural Product Biosynthesis

With the rise of inexpensive sequencing techniques, sequence data from bacterial metabolic pathways has been used in a variety of ways. For example, sequence data has been used to discover new bacterial secondary metabolite clusters (36), to examine the regulation of these clusters (37), to predict the products of a cluster (38), to look at enzymatic structural relationships such as interdomain and intermodular interactions (39), to study horizontal gene transfer as well as the evolution of secondary metabolite pathways (40, 41) and also to engineer clusters that produce novel products (42–44). It is evident from the preceding list that sequence data from bacterial secondary metabolite pathways can be highly useful. Most of these topics mentioned, however, are far beyond the scope of this monograph. The focus of

the proceeding sections will be narrowed to the role of sigma-54 in bacterial metabolism and the analysis of gene clusters from sequence data.

### 1.3.1.1 Sigma-54 as a Regulator of Bacterial Metabolism

Sigma-54, also known as RpoN or sigma-N, is one of many alternative transcription factors that modulate bacterial metabolism. Under certain conditions, sigma-54 binds to the RNA polymerase (RNAP) complex and directs the complex to sigma-54 promoter sites. It is also able to bind DNA without first binding to the core RNA polymerase complex (45, 46). In order to create an open promoter complex and initiate transcription, sigma-54 must form a complex with core RNA polymerase, bind to DNA, interact with Enhancer Binding Proteins (EBPs) and hydrolyze ATP (5). Sigma-54 is unique among the bacterial sigma transcription factors in that it requires contact with an EBP in order to melt the DNA and proceed with transcription (47).

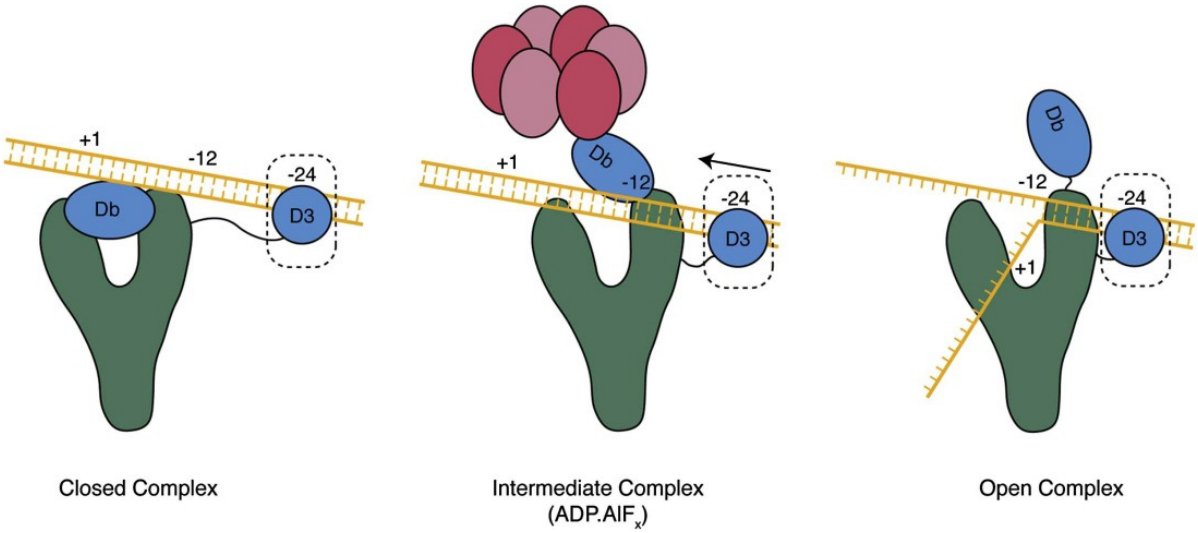



**Figure 10. A schematic representation of the closed, intermediate, and open complexes of the RNA polymerase holoenzyme complexed with DNA** (47)**.**
In the closed complex, the RNAP (green) binds to sigma-54 domains (blue) which is bound to DNA (yellow). Upon binding with an enhancer binding protein (pink) and ATP, the sigma-54-RNAP complex is able to melt the DNA. Once the DNA is melted, the enhancer binding protein can dissociate and transcription can proceed**.**


Initial studies linked sigma-54 to transcription of genes involved in nitrogen regulation and flagellar biosynthesis (48, 49). However, more recent research has uncovered that it also plays a role in regulating cell wall biosynthesis, transport systems, signal transduction, and central intermediary metabolism (49, 50). At this time, it is not significantly associated with secondary metabolite biosynthesis genes. However, this might be because the complete role of sigma-54 is not yet fully understood.

### 1.3.1.2 Sigma-54's Phylogenetic Distribution

While sigma-54 homologs are widely distributed in the bacterial world, they are not found in all bacterial phyla (see Figure 11). For example, two major classes of bacteria, Cyanobacteria and Actinobacteria, do not encode a sigma-54 homolog (49).

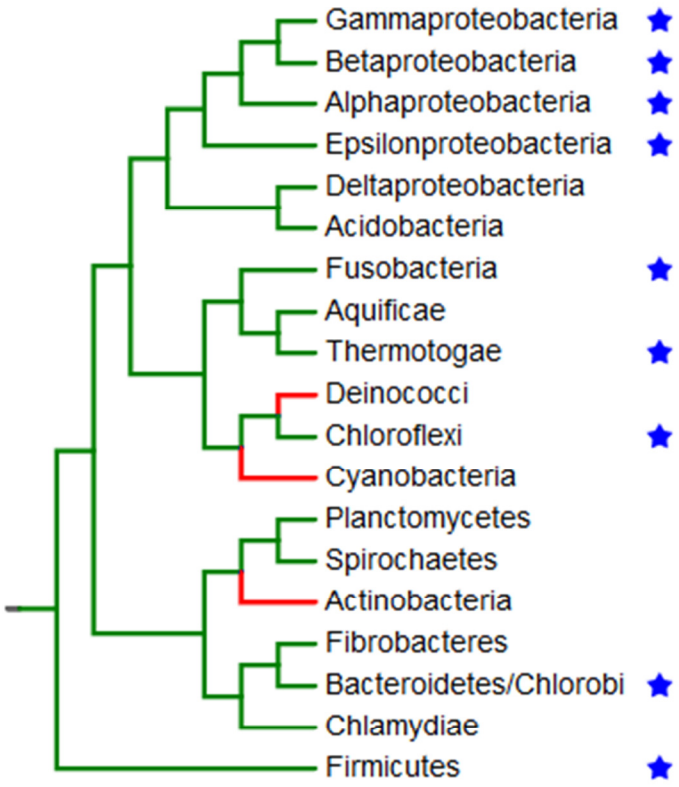



**Figure 11. Phylogenetic distribution of sigma-54 in bacteria.**
Green lines: at least one species contains a sigma-54 homolog
Red lines: no species with sigma-54
Blue stars: phylum contains some members that do not have sigma-54
Image was generated using EvolView (51) using bacterial phylogenetic data from the Interactive Tree of Life (52) and sigma-54 from (49)

Given its wide distribution, it is believed that the sigma-54 regulatory mechanism is relatively ancient. However, over time, some bacteria have lost the need for it. It has been noted that bacterial species with small genomes, such as endosymbionts, and those with larger than average genomes, such as the Actinobacteria and Cyanobacteria, are more likely to have lost sigma-54 then bacteria with an average sized genome (49).

### 1.3.1.3 Overexpression of Sigma-54 Allows Production of Oxytetracycline in Heterologous Hosts

While it has been known for some time that sigma-54 plays some role in bacterial metabolism, the true extent of its actions are still being uncovered. For example, it has been shown that overexpression of sigma-54 enables heterologous production of oxytetracycline in *E. coli*, a gamma-Proteobacteria acting as a heterologous host (53). The sigma-54 factor appears to recruit RNA polymerase to promoter sequences directly upstream of biosynthetic genes, enabling gene expression in the heterologous host (53).

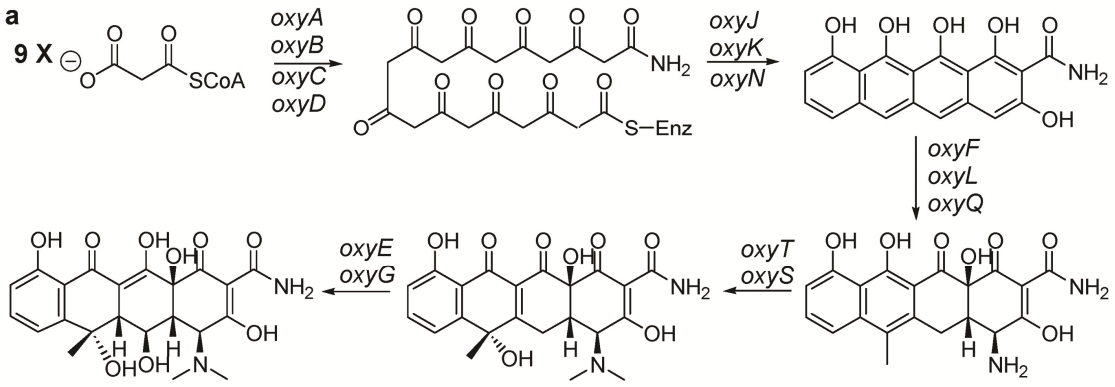



**Figure 12. The enzymatic pathway responsible for the biosynthesis of oxytetracycline** (54)**.**
Used with permission of the author.

Oxytetracycline is the product of a type II polyketide synthase secondary metabolite cluster isolated from the actinomycete *Streptomyces rimosus* (55) which does not encode a sigma-54 homolog. Curiously, however, it has been shown that it contains sigma-54 binding site upstream of the *oxyB* gene which encodes a β-ketosynthase (53).

### 1.3.1.4 Sigma-54 Consensus Sequence

Sigma-54's DNA binding consensus sequence has been determined through a variety of methods (50, 56). The consensus sequence can be represented as YTGGCACGrNNNTTGCW where Y represents C or T, R represents A or G, W represents A or T, and N represents any base (56). The most important contributions to binding are found at the -12 and -24 positions upstream of the transcriptional start site (see Figure 13).

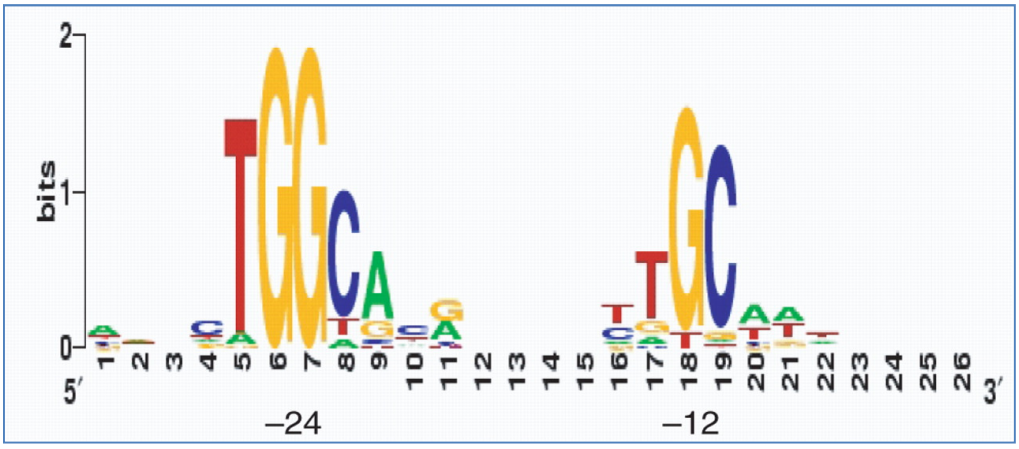



**Figure 13. Frequency based representation of the sigma-54 consensus sequence.**
The height of each letter indicates the relative frequency of that base at that position. Figure reprinted from (50). 

A positional weighted matrix can also be constructed to represent the consensus sequence. This is a more comprehensive way of representing the promoter binding site by taking into account the relative frequencies of each base at each position.

### 1.3.2 Analysis of Bacterial Secondary Metabolite Gene Clusters

Bacteria produce a wide range of secondary metabolites. Pathways containing polyketide synthases (PKS) or non-ribosomal peptide synthetases (NRPS) are among those that produce the most chemical diversity (57, 58). Therefore, there is great interest in better understanding how these pathways work to be able to harness their full potential (57, 59). Sequencing and bioinformatics analysis of these pathways have revealed patterns in the protein sequences that can be used to predict the products of new clusters as well as better understand the factors involved in the enzymatic synthesis of these natural products (60) .

#### 1.3.2.1 Pathway Types

There are several types of PKS/NRPS pathways. Type I PKS clusters consist of large modular, multifunctional enzymes that act in a stepwise fashion, although some Type I pathways have iterative components. Type II PKSs consist of many individual enzymes, each separately encoded by their own gene, that act together through an iterative process. Type III PKSs use Coenzyme A to hold the substrate instead of acyl carrier proteins (ACP). NRPS clusters are similar to Type I PKSs except that they use amino acids as the products backbone. Hybrid pathways have both a PKS and an NRPS component.

### 1.3.2.2 Cluster Prediction Tools

A variety of tools have been developed with the goal of predicting the products of PKS/NRPS clusters from sequence data. Examples of these include SearchPKS (61), NRPS-PKS (62), SEARCHGTr (63), NRPSpredictor (64), ClustScan (65), NPSearcher (66), ASMPKS/MAPSI (67), CLUSEAN (68), SBSPKS (69), SMURF (70), NRPSpredictor2 (71), NRPSsp (72) and antiSMASH (38).

Many of these tools use computational methods such as multiple sequence alignments, phylogenetic tree reconstruction or hidden Markov models to determine the type and function of the domains of a given cluster.

### 1.3.2.3 Defining the Cluster

The first step in analyzing a cluster is to define its boundaries. The simplest way is to search through a sequence for genes with homology to genes known to be involved in secondary metabolite biosynthesis. By identifying groupings of these genes, a cluster can be defined. Some sequences, such as genomes, can have many clusters. Without biochemical experimentation, however, it is not possible to define these borders without making some suppositions. For example, antiSMASH assumes the presence of a new cluster if there is an interval of more than 15 kb between genes related to secondary metabolite biosynthesis (38).

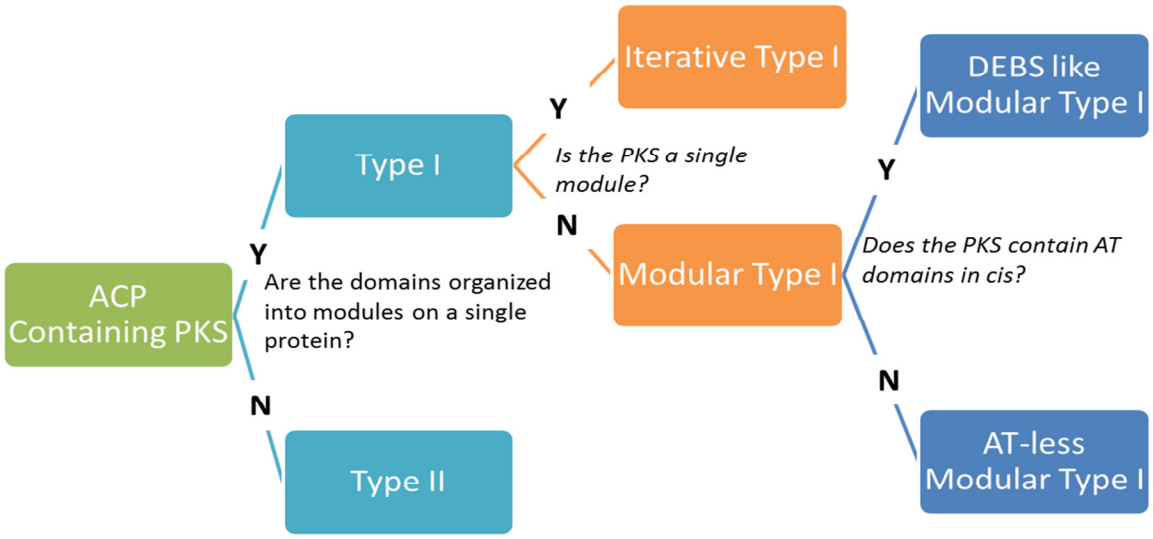



**Figure 14. Decision tree to determine PKS type.**
Based on diagram in (73). Abbreviations: ACP, acyl carrier protein; DEBS, 6-Deoxyerythronolide B synthase; AT, acyl transferase.

Once the borders of a putative cluster have been defined, the next step is to try to determine the type of pathway. This is done by identifying the presence of certain key enzymes. For example, an NRPS pathway will have adenylation (A) domains while PKS pathways would either have acyl transferase (AT) domains or potentially no acyl selecting domains at all such as in the case of in-trans systems. By determining whether each gene encodes multiple domains (Type I vs Type II), the number of modules (modular vs iterative), and the presence or absence of AT domains (*in cis* vs *in trans*), a determination of the pathway type can be made (see Figure 14) (73).

### 1.3.2.4 Predicting Domain Organization

The order of chain elongation can also be predicted by using the colinearity rule. The colinearity rule suggests that the ordering of PKS/NRPS genes in the DNA is related to the arrangement of modules and domains in the assembly line (74). Computational tools can take advantage of this relationship to make an estimation of the structure of the assumed product (29).

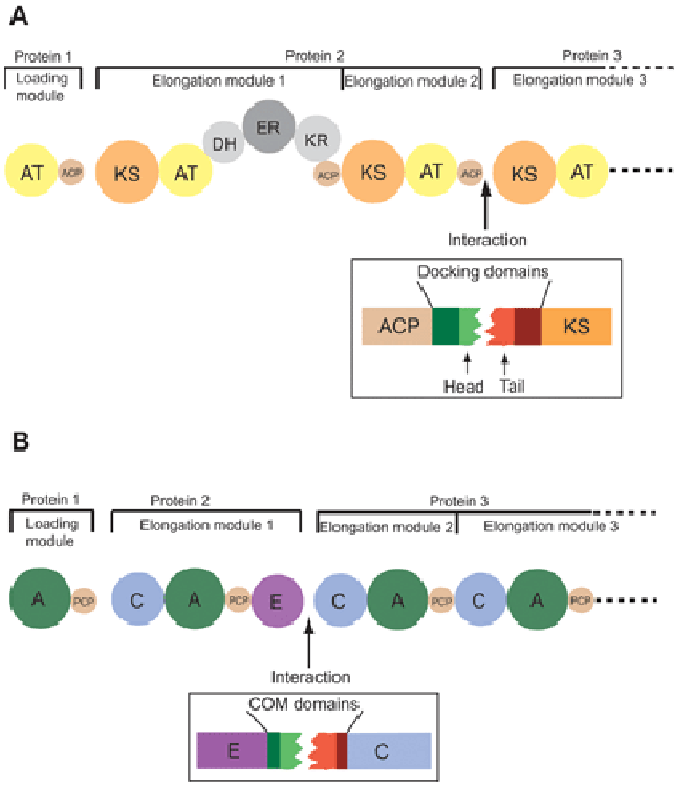



**Figure 15. Diagram depicting the role of protein-protein interactions between docking domains** (29)**.** The top figure (A) depicts a modular PKS assembly line while the bottom (B) shows a set of typical NRPS elongation modules. Abbreviations: KS, ketosynthase; AT: acyltransferase; ACP: acyl carrier protein; DH: dehydratase; ER: enoyl reductase; KR: ketoreductase; C: Condensation domain; A, adenylation domain; PCP: peptidyl carrier protein; E: epimerisation domain. Reprinted with permission of the authors.

While this is also used to determine the expected product(s) of a cluster, there are a number of other factors that can prevent an accurate prediction. For example, many newly discovered clusters display atypical properties such as domain skipping, stuttering (repeating a module more than one time), or multiple acyl carrier proteins (ACPs) per module (75–77). Currently, accurate modeling *in silico* of these clusters with unusual features is not possible. Additional biochemical investigations into these phenomena will be required in order to improve the predictive ability of these algorithms.

#### 1.3.2.5 Predicting Domain Activity and Specificity

The functional characteristics of many PKS/NRPS domains can be inferred from its polypeptide sequence (73, 78). Methods such as multiple sequence alignment, phylogenetic tree based grouping, and hidden Markov models are used to perform meta-analyses of biochemical studies to uncover these relationships.

Adenylation (A) domains from NRPS clusters, which are responsible for selecting the next amino acid to be added to the growing polypeptide, are among the most well studied and can be predicted based on a signature sequence of 10 binding site residues (79), homology modeling (80), transductive support vector machines (TVSMs) (64), or hidden Markov models (81).

Their equivalent in PKS clusters, acyl transferases (AT), have received less attention but consensus models have been created that can predict the selection of malonate and methyl-malonate (73) and there have been two separate attempts at creating predictive hidden Markov models for these domains (65, 81).

Keto-synthase (KS) domains catalyze the elongation of the polyketide chain and, as such, are not directly involved in substrate selection and tend to be have broad substrate specificity (82). However, they have been used to assist in predicting the product of a cluster (83, 84) and, especially in Type II systems where the inactive KS, also known as the chain length factor (CLF), is believed to be responsible for controlling chain elongation (85–87).

Keto-reductase (KR) domains have also been analyzed to determine their activity and the type of stereochemistry they impart to their substrate. KR domains reduce the β-ketone to a β-hydroxyl group and, during the process, sets a stereocenter. Based on the presence or absence of certain active site residues, it can be determined if the domain is active or inactive (73). In addition, the presence of key residues in the active site can be used to predict the stereochemical outcome of the reduction (88–90).

# Chapter 2: The Sigma-54 Promoter Database

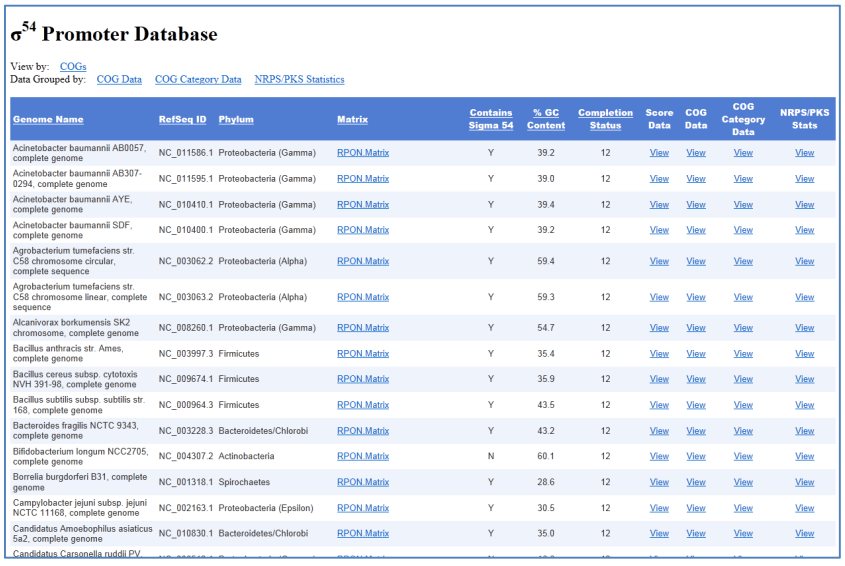
## 2.1 The Sigma-54 Promoter Database

In the past 25 years, approximately half of all drugs, and over 60% of new antibacterial and new anticancer drugs are natural products or natural product derived (35). Among these, many are from bacterial sources. Indeed, bacterial secondary metabolites are an excellent source of potential new drugs particularly since it is believed that they harbour a large number of active secondary metabolites that have yet to be discovered (34).

In order to access these potential new compounds, it is important to better understand the factors controlling secondary metabolite biosynthesis in bacteria. Sigma-54, a bacterial transcriptional regulator, has been shown to have to be intimately connected to important metabolic processes in bacteria but its full role is still not yet certain (49). By examining its transcriptional binding sites, some of the secrets of this unique transcriptional regulator might be uncovered.

## 2.2 Purpose of Database

The Sigma-54 Promoter Database is a unique resource that gives researchers access to computationally predicted sigma-54 promoter sites in over 60 bacterial genomes. The database home page (see Figure 16) can be accessed at www.sigma54.ca.



$\sigma^{54}$ Promoter Database

View by: COGs
Data Grouped by: COG Data COG Category Data NRPS/PKS Statistics

| Genome Name | RefSeq ID | Phylum | Matrix | Contains Sigma 54 | % GC Content | Completion Status | Score Data | COG Data | COG Category Data | NRPS/PKS Stats |
|---|---|---|---|---|---|---|---|---|---|---|
| Acinetobacter baumannii AB0057, complete genome | NC_011586.1 | Proteobacteria (Gamma) | RPON.Matrix | Y | 39.2 | 12 | View | View | View | View |
| Acinetobacter baumannii AB307-0294, complete genome | NC_011595.1 | Proteobacteria (Gamma) | RPON.Matrix | Y | 39.0 | 12 | View | View | View | View |
| Acinetobacter baumannii AYE, complete genome | NC_010410.1 | Proteobacteria (Gamma) | RPON.Matrix | Y | 39.4 | 12 | View | View | View | View |
| Acinetobacter baumannii SDF, complete genome | NC_010400.1 | Proteobacteria (Gamma) | RPON.Matrix | Y | 39.2 | 12 | View | View | View | View |
| Agrobacterium tumefaciens str. C58 chromosome circular, complete sequence | NC_003062.2 | Proteobacteria (Alpha) | RPON.Matrix | Y | 59.4 | 12 | View | View | View | View |
| Agrobacterium tumefaciens str. C58 chromosome linear, complete sequence | NC_003063.2 | Proteobacteria (Alpha) | RPON.Matrix | Y | 59.3 | 12 | View | View | View | View |
| Alcanivorax borkumensis SK2 chromosome, complete genome | NC_008260.1 | Proteobacteria (Gamma) | RPON.Matrix | Y | 54.7 | 12 | View | View | View | View |
| Bacillus anthracis str. Ames, complete genome | NC_003997.3 | Firmicutes | RPON.Matrix | Y | 35.4 | 12 | View | View | View | View |
| Bacillus cereus subsp. cytotoxis NVH 391-98, complete genome | NC_009674.1 | Firmicutes | RPON.Matrix | Y | 35.9 | 12 | View | View | View | View |
| Bacillus subtilis subsp. subtilis str. 168, complete genome | NC_000964.3 | Firmicutes | RPON.Matrix | Y | 43.5 | 12 | View | View | View | View |
| Bacteroides fragilis NCTC 9343, complete genome | NC_003228.3 | Bacteroidetes/Chlorobi | RPON.Matrix | Y | 43.2 | 12 | View | View | View | View |
| Bifidobacterium longum NCC2705, complete genome | NC_004307.2 | Actinobacteria | RPON.Matrix | N | 60.1 | 12 | View | View | View | View |
| Borrelia burgdorferi B31, complete genome | NC_001318.1 | Spirochaetes | RPON.Matrix | Y | 28.6 | 12 | View | View | View | View |
| Campylobacter jejuni subsp. jejuni NCTC 11168, complete genome | NC_002163.1 | Proteobacteria (Epsilon) | RPON.Matrix | Y | 30.5 | 12 | View | View | View | View |
| Candidatus Amoebophilus asiaticus 5a2, complete genome | NC_010830.1 | Bacteroidetes/Chlorobi | RPON.Matrix | Y | 35.0 | 12 | View | View | View | View |
| Candidatus Carsonella ruddii PV, | | | | | | | | | | |

**Figure 16. Sigma-54 Promoter Database home page.**
A larger view of the home screen can be seen in Appendix 7: Sigma-54 Promoter Database Screenshots, p. 101

The home page lists the organisms in the database along with the matrix they were analyzed with. The matrix can be downloaded by clicking on its name. The home page also includes links to data pages containing score data, COG data, and COG Category data for each organism. Database-wide or phylum based aggregated data can be accessed by following links at the top of the page.

By pressing the 'View' link underneath the Score Data column, detailed Score Data information for an organism can be viewed. A filter selection page will appear (see Figure 17) that allows for altering the score and location parameters of the results. This allows researchers the ability to adjust the result set according to their own preferences.

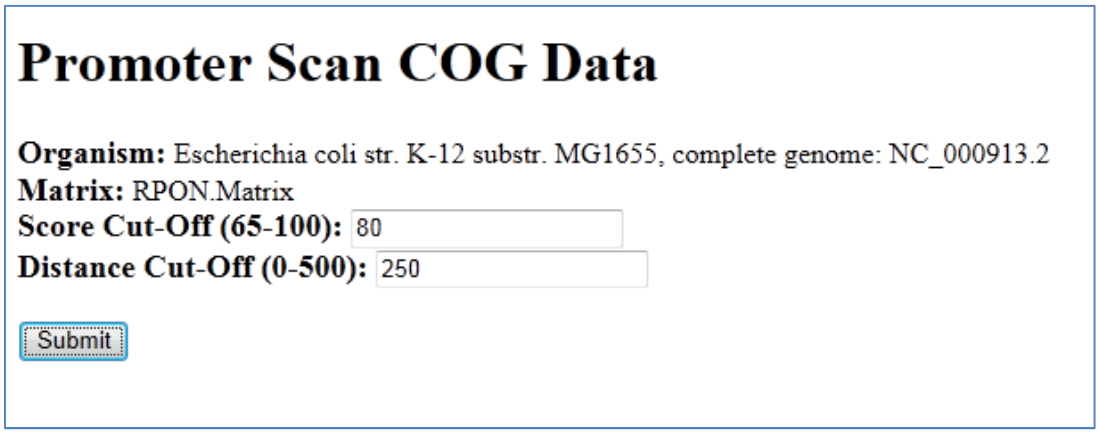
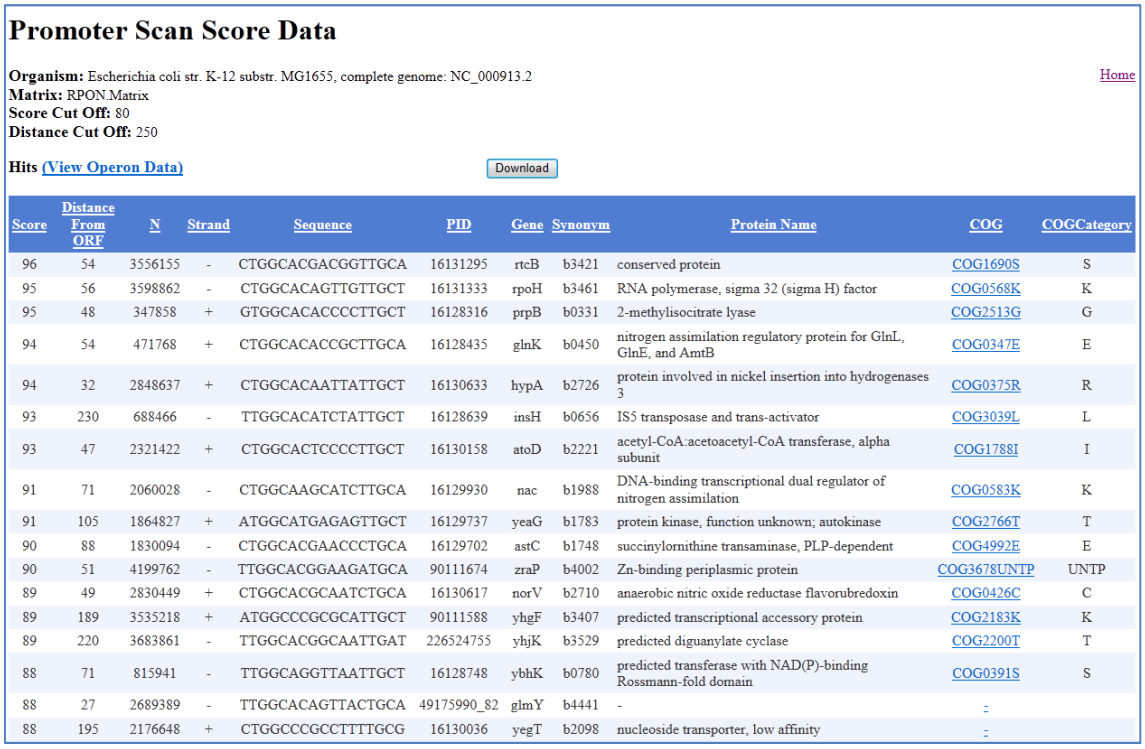



**Figure 17. Screen demonstrating the ability to filter results by score or location upstream of the gene.**



## Promoter Scan Score Data

**Organism:** Escherichia coli str. K-12 substr. MG1655, complete genome: NC_000913.2
**Matrix:** RPON.Matrix
**Score Cut Off:** 80
**Distance Cut Off:** 250

Home

**Hits** (View Operon Data) Download

| Score | Distance From ORF | N | Strand | Sequence | PID | Gene | Synonym | Protein Name | COG | COGCategory |
|---|---|---|---|---|---|---|---|---|---|---|
| 96 | 54 | 3556155 | - | CTGGCACGACGGTTGCA | 16131295 | rtcB | b3421 | conserved protein | COG1690S | S |
| 95 | 56 | 3598862 | - | CTGGCACAGTTGTTGCT | 16131333 | rpoH | b3461 | RNA polymerase, sigma 32 (sigma H) factor | COG0568K | K |
| 95 | 48 | 347858 | + | GTGGCACACCCCTTGCT | 16128316 | prpB | b0331 | 2-methylisocitrate lyase | COG2513G | G |
| 94 | 54 | 471768 | + | CTGGCACACCGCTTGCA | 16128435 | glnK | b0450 | nitrogen assimilation regulatory protein for GlnL, GlnE, and AmtB | COG0347E | E |
| 94 | 32 | 2848637 | + | CTGGCACAATTATTGCT | 16130633 | hypA | b2726 | protein involved in nickel insertion into hydrogenases 3 | COG0375R | R |
| 93 | 230 | 688466 | - | TTGGCACATCTATTGCT | 16128639 | insH | b0656 | IS5 transposase and trans-activator | COG3039L | L |
| 93 | 47 | 2321422 | + | CTGGCACTCCCCTTGCT | 16130158 | atoD | b2221 | acetyl-CoA:acetoacetyl-CoA transferase, alpha subunit | COG1788I | I |
| 91 | 71 | 2060028 | - | CTGGCAAGCATCTTGCA | 16129930 | nac | b1988 | DNA-binding transcriptional dual regulator of nitrogen assimilation | COG0583K | K |
| 91 | 105 | 1864827 | + | ATGGCATGAGAGTTGCT | 16129737 | yeaG | b1783 | protein kinase, function unknown; autokinase | COG2766T | T |
| 90 | 88 | 1830094 | - | CTGGCACGAACCCTGCA | 16129702 | astC | b1748 | succinylornithine transaminase, PLP-dependent | COG4992E | E |
| 90 | 51 | 4199762 | - | TTGGCACGGAAGATGCA | 90111674 | zraP | b4002 | Zn-binding periplasmic protein | COG3678UNTP | UNTP |
| 89 | 49 | 2830449 | + | CTGGCACGCAATCTGCA | 16130617 | norV | b2710 | anaerobic nitric oxide reductase flavorubredoxin | COG0426C | C |
| 89 | 189 | 3535218 | + | ATGGCCCGCGCATTGCT | 90111588 | yhgF | b3407 | predicted transcriptional accessory protein | COG2183K | K |
| 89 | 220 | 3683861 | - | TTGGCACGGCAATTGAT | 226524755 | yhjK | b3529 | predicted diguanylate cyclase | COG2200T | T |
| 88 | 71 | 815941 | - | TTGGCAGGTTAATTGCT | 16128748 | ybhK | b0780 | predicted transferase with NAD(P)-binding Rossmann-fold domain | COG0391S | S |
| 88 | 27 | 2689389 | - | TTGGCACAGTTACTGCA | 49175990_82 | glmY | b4441 | - | - | |
| 88 | 195 | 2176648 | + | CTGGCCCGCCTTTTGCG | 16130036 | yegT | b2098 | nucleoside transporter, low affinity | - | |

**Figure 18. Score data page displaying putative promoter details by decreasing score.**

The Score Data page has detailed information regarding putative promoters that were computationally predicted. An example Score Data page, for *E. coli*, is shown above in Figure 18. Each row represents a putative promoter and includes information such as the score, how far upstream it is from the ORF, where it is in the genome (N), the direction or strand, the sequence, and gene identification information such as its PID, gene name, gene synonym, name of the protein it encodes, and finally what COG and COG category it is classified under.

From the home page, it is also possible to view COG data for a given organism by selecting the 'View' link under the COG Data column. The COG Data page (see Figure 19) contains a list of COGs present in the organism summed by the number of times it occurs upstream of an ORF at a score level of 65, 75, and 80. The COG Data page provides information with regards to what type of genes are predicted to be under sigma-54 transcriptional regulation in the given organism.

## Promoter Scan COG Data

**Organism:** Escherichia coli str. K-12 substr. MG1655, complete genome: NC_000913.2
**Matrix:** RPON.Matrix

**Number of Hits by Gene Scoring Above 65 Grouped by COG** (View Operon Data) Download Table

| # Hits With Score > 65 | # Hits With Score > 75 | # Hits With Score > 80 | Total COG Count | COG | Description |
|---|---|---|---|---|---|
| 739 | 193 | 49 | 789 | - | Unassigned |
| 16 | 7 | 4 | 17 | COG0642T | Signal transduction histidine kinase |
| 44 | 12 | 4 | 46 | COG0583K | Transcriptional regulator |
| 12 | 7 | 3 | 13 | COG2200T | FOG: EAL domain |
| 34 | 11 | 3 | 36 | COG2814G | Arabinose efflux permease |
| 11 | 6 | 2 | 14 | COG0745TK | Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain |
| 3 | 2 | 2 | 3 | COG1177E | ABC-type spermidine/putrescine transport system, permease component II |
| 11 | 4 | 2 | 11 | COG1012C | NAD-dependent aldehyde dehydrogenases |
| 6 | 3 | 2 | 6 | COG2204T | Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains |
| 2 | 2 | 2 | 2 | COG3691S | Uncharacterized protein conserved in bacteria |
| 3 | 2 | 2 | 3 | COG0568K | DNA-directed RNA polymerase, sigma subunit (sigma70/sigma32) |
| 8 | 3 | 2 | 8 | COG1879G | ABC-type sugar transport system, periplasmic component |
| 7 | 3 | 2 | 7 | COG0834ET | ABC-type amino acid transport/signal transduction systems, periplasmic component/domain |
| 4 | 2 | 2 | 4 | COG0456R | Acetyltransferases |
| 3 | 2 | 2 | 3 | COG1363G | Cellulase M and related proteins |
| 3 | 2 | 2 | 3 | COG0754E | Glutathionylspermidine synthase |
| 6 | 4 | 2 | 6 | COG1762GT | Phosphotransferase system mannitol/fructose-specific IIA domain (Ntr-type) |
| 3 | 2 | 2 | 3 | COG4992E | Ornithine/acetylornithine aminotransferase |
| 3 | 2 | 2 | 3 | COG0078E | Ornithine carbamoyltransferase |
| 10 | 3 | 2 | 11 | COG1063ER | Threonine dehydrogenase and related Zn-dependent dehydrogenases |

**Figure 19. COG page for a single organism showing the total number of hits with scores greater than or equal to 65, 75, and 80.**

Similar to the COG Data page, the frequency of COG categories at various score levels is shown on the COG Category page (see Figure 20). This gives a more general overview of the types of genes in the given organism that are predicted to be under sigma-54 control.

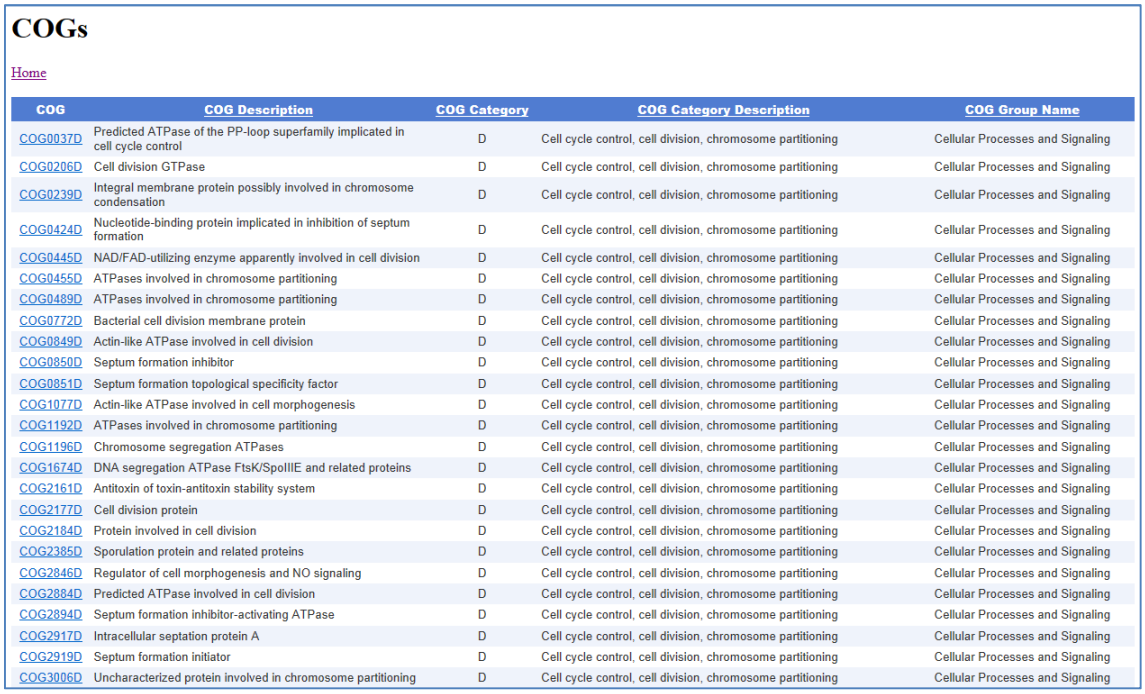
## Promoter Scan COG Data by Category

**Organism:** Escherichia coli str. K-12 substr. MG1655, complete genome: NC_000913.2
**Matrix:** RPON.Matrix

**Number of Hits in Front of Genes Scoring Above 65 Grouped by COG Category** (View Operon Data) Download Table

| Hits with Score > 65 | Hits with Score > 75 | Hits with Score > 80 | Total Count | COG Category | Description |
|---|---|---|---|---|---|
| 382 | 89 | 37 | 403 | R | General function prediction only |
| 358 | 109 | 24 | 371 | G | Carbohydrate transport and metabolism |
| 347 | 119 | 36 | 363 | E | Amino acid transport and metabolism |
| 304 | 85 | 21 | 318 | S | Function unknown |
| 286 | 83 | 25 | 306 | K | Transcription |
| 275 | 80 | 15 | 287 | C | Energy production and conversion |
| 219 | 54 | 12 | 229 | M | Cell wall/membrane/envelope biogenesis |
| 198 | 58 | 21 | 216 | P | Inorganic ion transport and metabolism |
| 197 | 54 | 19 | 206 | L | Replication, recombination and repair |
| 177 | 46 | 8 | 186 | J | Translation, ribosomal structure and biogenesis |
| 165 | 60 | 28 | 178 | T | Signal transduction mechanisms |
| 149 | 53 | 16 | 156 | H | Coenzyme transport and metabolism |
| 136 | 34 | 9 | 142 | O | Posttranslational modification, protein turnover, chaperones |
| 124 | 32 | 10 | 129 | U | Intracellular trafficking, secretion, and vesicular transport |
| 107 | 25 | 7 | 111 | N | Cell motility |
| 98 | 26 | 6 | 100 | I | Lipid transport and metabolism |
| 96 | 28 | 4 | 98 | F | Nucleotide transport and metabolism |
| 60 | 20 | 4 | 64 | Q | Secondary metabolites biosynthesis, transport and catabolism |
| 45 | 14 | 6 | 49 | V | Defense mechanisms |
| 31 | 7 | 1 | 33 | D | Cell cycle control, cell division, chromosome partitioning |
| 2 | 1 | 0 | 2 | A | RNA processing and modification |
| 1 | 0 | 0 | 1 | W | Extracellular structures |
| 0 | 0 | 0 | 0 | B | Chromatin structure and dynamics |
| 0 | 0 | 0 | 0 | Y | Nuclear structure |
| 0 | 0 | 0 | 0 | Z | Cytoskeleton |

**Figure 20. COG Category totals for scores greater than or equal to 65, 75, and 80.**



COGs

Home

| COG | COG Description | COG Category | COG Category Description | COG Group Name |
|---|---|---|---|---|
| COG0037D | Predicted ATPase of the PP-loop superfamily implicated in cell cycle control | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG0206D | Cell division GTPase | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG0239D | Integral membrane protein possibly involved in chromosome condensation | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG0424D | Nucleotide-binding protein implicated in inhibition of septum formation | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG0445D | NAD/FAD-utilizing enzyme apparently involved in cell division | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG0455D | ATPases involved in chromosome partitioning | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG0489D | ATPases involved in chromosome partitioning | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG0772D | Bacterial cell division membrane protein | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG0849D | Actin-like ATPase involved in cell division | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG0850D | Septum formation inhibitor | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG0851D | Septum formation topological specificity factor | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG1077D | Actin-like ATPase involved in cell morphogenesis | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG1192D | ATPases involved in chromosome partitioning | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG1196D | Chromosome segregation ATPases | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG1674D | DNA segregation ATPase FtsK/SpoIIIE and related proteins | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG2161D | Antitoxin of toxin-antitoxin stability system | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG2177D | Cell division protein | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG2184D | Protein involved in cell division | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG2385D | Sporulation protein and related proteins | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG2846D | Regulator of cell morphogenesis and NO signaling | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG2884D | Predicted ATPase involved in cell division | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG2894D | Septum formation inhibitor-activating ATPase | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG2917D | Intracellular septation protein A | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG2919D | Septum formation initiator | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |
| COG3006D | Uncharacterized protein involved in chromosome partitioning | D | Cell cycle control, cell division, chromosome partitioning | Cellular Processes and Signaling |

**Figure 21. List of COGs in the database. Displays a description for each COG along with its category and group information.**

Database-wide data filtered by COG is also available. A page listing all of the COGs can be accessed by pressing the COGs link at the top left hand corner of the home page. This page lists all of the COGs and shows their COG code, a description of the COG, its category name, and its group name (see Figure 20). By selecting the COG code, a new page appears showing all predicted sigma-54 sites, for all organisms in the database, that are upstream of a gene with that COG (see Figure 21).

**Promoter Scan Score Data**

**COG:** COG0568K
**COG Categories:** Transcription

Home

**Hits Scoring Above 65** Download

| Score | Distance From ORF | Phylum | Genome Name | Matrix | N | Strand | Sequence | PID | Gene | Synonym | Protein Name |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 97 | 41 | Proteobacteria (Gamma) | Salmonella enterica subsp. enterica serovar Typhi str. CT18, complete genome | RPON.Matrix | 4109175 | + | TTGGCACGGTTGTTGCT | 16762736 | rpoH | STY4243 | RNA polymerase factor sigma-32 |
| 95 | 56 | Proteobacteria (Gamma) | Escherichia coli str. K-12 substr. MG1655, complete genome | RPON.Matrix | 3598862 | - | CTGGCACAGTTGTTGCT | 16131333 | rpoH | b3461 | RNA polymerase, sigma 32 (sigma H) factor |
| 93 | 39 | Spirochaetes | Borrelia burgdorferi B31, complete genome | RPON.Matrix | 813278 | - | TTGGCACAGTTTTTGCA | 15595116 | - | BB0771 | RNA polymerase sigma factor (rpoS) |
| 92 | 57 | Proteobacteria (Gamma) | Escherichia coli str. K-12 substr. MG1655, complete genome | EColiExpConfirmed.Matrix | 3598863 | - | TCTGGCACAGTTGTTGC | 16131333 | rpoH | b3461 | RNA polymerase, sigma 32 (sigma H) factor |
| 88 | 242 | Proteobacteria (Delta) | Myxococcus xanthus DK 1622, complete genome | RPON.Matrix | 3904298 | - | TTGGCATTTCTGTTGCT | 108760621 | sigB | MXAN_3357 | RNA polymerase sigma-B factor |
| 88 | 223 | Proteobacteria (Gamma) | Salmonella enterica subsp. enterica serovar Typhi str. CT18, complete genome | RPON.Matrix | 3233543 | + | GTGGCACAAATGATGCT | 16761985 | rpoD | STY3390 | RNA polymerase sigma factor RpoD |
| 85 | 67 | Proteobacteria (Delta) | Myxococcus xanthus DK 1622, complete genome | RPON.Matrix | 7676489 | + | TTGGCACGTGGGTCGCA | 108762657 | sigC | MXAN_6209 | RNA polymerase sigma-C factor |
| 84 | 243 | Proteobacteria (Delta) | Myxococcus xanthus DK 1622, complete genome | MXanth.S54.Selected20.Matrix | 3904299 | - | GTTGGCATTTCTGTTGCTCAAG | 108760621 | sigB | MXAN_3357 | RNA polymerase sigma-B factor |
| 84 | 244 | Proteobacteria (Delta) | Myxococcus xanthus DK 1622, complete genome | MXanth.S54.Top40.Matrix | 3904300 | - | GGTTGGCATTTCTGTTGCTCA | 108760621 | sigB | MXAN_3357 | RNA polymerase sigma-B factor |
| 83 | 63 | Proteobacteria (Delta) | Myxococcus xanthus DK 1622, complete genome | MXanth.S54.Selected20.Matrix | 7676493 | + | CTTGGCACGTGGGTCGCAAAGG | 108762657 | sigC | MXAN_6209 | RNA polymerase sigma-C factor |
| 83 | 404 | Proteobacteria (Gamma) | Escherichia coli str. K-12 substr. MG1655, complete genome | EColiExpConfirmed.Matrix | 3210665 | + | CGTGGTACAAATAATGC | 16130963 | rpoD | b3067 | RNA polymerase, sigma 70 (sigma D) factor |
| 83 | 61 | Proteobacteria (Gamma) | Escherichia coli str. K-12 substr. MG1655, complete genome | EColiS54.Matrix | 3598867 | - | AAGCTCTGGCACAGTTGTTGCTA | 16131333 | rpoH | b3461 | RNA polymerase, sigma 32 (sigma H) factor |
| 83 | 55 | Spirochaetes | Treponema denticola ATCC 35405, complete genome | RPON.Matrix | 960036 | + | TTAGCATGGTTTTTGCT | 42526449 | - | TDE0937 | RNA polymerase sigma-70 factor family protein |
| 83 | 26 | Spirochaetes | Treponema denticola ATCC 35405, complete genome | RPON.Matrix | 77827 | + | TTAGCACACTTTTTGCT | 42525589 | - | TDE0070 | sigma-70 family RNA polymerase sigma factor |
| 81 | 244 | Proteobacteria (Delta) | Myxococcus xanthus DK 1622, complete genome | MXanth.S54.Top20.Matrix | 3904300 | - | GGTTGGCATTTCTGTTGCTCAA | 108760621 | sigB | MXAN_3357 | RNA polymerase sigma-B factor |
| 81 | 403 | Proteobacteria (Gamma) | Escherichia coli str. K-12 substr. MG1655, complete genome | RPON.Matrix | 3210666 | + | GTGGTACAAATAATGCT | 16130963 | rpoD | b3067 | RNA polymerase, sigma 70 (sigma D) factor |
| 81 | 318 | Proteobacteria (Gamma) | Legionella pneumophila subsp. pneumophila str. Philadelphia 1, complete genome | RPON.Matrix | 1411644 | + | AAGGCATAAATATTGCA | 52841515 | rpoS | lpg1284 | stationary phase specific sigma factor RpoS |

**Figure 22. Database wide putative promoters for a given COG ordered by decreasing score.**

Aggregated COG and COG Category statistics are available by selecting the COG or COG Category links located on the top left side of the home page. A filter screen appears allowing users to choose what types of organisms they wish to include in the aggregated data. Users can choose to have organisms that have a sigma-54 homolog, those that do not, or both. They can also filter by Phylum (see Figure 23).


**Figure 23. Page to select filter parameters for aggregated COG data.**

## Promoter Scan Aggregate COG Data

Home

**Contains $\sigma^{54}$:** All
**Phylum:** Proteobacteria
**Matrix:** RPON.Matrix

**Hits by COG by Gene** (View Operon Data) Download Table

| # Hits With Score ≥ 65 | # Hits With Score ≥ 75 | # Hits With Score ≥ 80 | Total COG Count | COG | Description | # of Organisms in Database with this COG (Score ≥ 80) | % of Organisms in Database with this COG (Score ≥ 80) | % of σ54 Containing Organisms in Database with this COG (Score ≥ 80) | % of Organisms not Containing σ54 in Database with this COG (Score ≥ 80) | % Difference between σ54 and non σ54 in Database with this COG (Score ≥ 80) | View Phylum Distribution |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 27075 | 6792 | 1819 | 29409 | - | Unassigned | 72 | 97.3 | 100.0 | 92.3 | 7.7 | View |
| 90 | 70 | 68 | 97 | COG1484L | DNA replication protein | 7 | 9.5 | 14.6 | 0.0 | 14.6 | View |
| 1112 | 281 | 52 | 1168 | COG0583K | Transcriptional regulator | 27 | 36.5 | 45.8 | 19.2 | 26.6 | View |
| 555 | 135 | 48 | 585 | COG0642T | Signal transduction histidine kinase | 26 | 35.1 | 45.8 | 15.4 | 30.4 | View |
| 182 | 80 | 44 | 190 | COG0834ET | ABC-type amino acid transport/signal transduction systems, periplasmic component/domain | 18 | 24.3 | 31.3 | 11.5 | 19.7 | View |
| 431 | 124 | 35 | 466 | COG1309K | Transcriptional regulator | 27 | 36.5 | 41.7 | 26.9 | 14.7 | View |
| 351 | 102 | 35 | 365 | COG2204T | Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains | 21 | 28.4 | 41.7 | 3.8 | 37.8 | View |
| 349 | 76 | 35 | 374 | COG2207K | AraC-type DNA-binding domain-containing proteins | 20 | 27.0 | 31.3 | 19.2 | 12.0 | View |
| 504 | 123 | 33 | 540 | COG2814G | Arabinose efflux permease | 32 | 43.2 | 45.8 | 38.5 | 7.4 | View |
| 460 | 134 | 30 | 489 | COG1028IQR | Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) | 32 | 43.2 | 47.9 | 34.6 | 13.3 | View |
| 115 | 52 | 28 | 119 | COG0683E | ABC-type branched-chain amino acid transport systems, periplasmic component | 14 | 18.9 | 20.8 | 15.4 | 5.4 | View |
| 125 | 41 | 26 | 137 | COG0715P | ABC-type nitrate/sulfonate/bicarbonate transport systems, periplasmic components | 16 | 21.6 | 27.1 | 11.5 | 15.5 | View |

**Figure 24. Database-wide aggregate COG data.**

The Aggregated COG Data page provides statistics on the frequency of genes with the given COG located downstream of a predicted sigma-54 promoter. By selecting the link on the right hand side of the page under the View Phylum Distribution column, additional data can be viewed regarding the distribution by phylum of genes that have that COG as well as a high scoring sigma-54 promoter (see Figure 25).

## Distribution by Phylum (Score ≥ 80) for COG1484L: DNA replication protein

**Total # of organisms in database:** 74
**# of organisms in database with $\sigma^{54}$:** 48
**# of organisms in database without $\sigma^{54}$:** 26

| Phylum | # of organisms that have this COG (regardless of score) | # of organisms with $\sigma^{54}$ | # of organisms without $\sigma^{54}$ | # of organisms (Total) |
|---|---|---|---|---|
| Actinobacteria | 7 | 0 | 0 | 0 |
| Aquificae | 1 | 0 | 0 | 0 |
| Bacteroidetes/Chlorobi | 2 | 0 | 0 | 0 |
| Chlamydiae/Verrucomicrobia | 0 | 0 | 0 | 0 |
| Chloroflexi | 1 | 0 | 0 | 0 |
| Chrysiogenetes | 0 | 0 | 0 | 0 |
| Crenarchaeota | 0 | 0 | 0 | 0 |
| Cyanobacteria | 2 | 0 | 0 | 0 |
| Deferribacteres | 0 | 0 | 0 | 0 |
| Deinococcus-Thermus | 0 | 0 | 0 | 0 |
| Dictyoglomi | 0 | 0 | 0 | 0 |
| Euryarchaeota | 0 | 0 | 0 | 0 |
| Fibrobacteres/Acidobacteria | 1 | 1 | 0 | 1 |
| Firmicutes | 10 | 2 | 0 | 2 |
| Fusobacteria | 0 | 0 | 0 | 0 |
| Gemmatimonadetes | 0 | 0 | 0 | 0 |

**Figure 25. Presence of COG with a high score by phylum.**

## 2.3 Unique Features

While there have been other databases with sigma-54 promoter data (91), this database is unique. Not only does it look at the largest and most broadly distributed range of bacteria, it also analyzes sequences between and within genes for promoter sites. All other published databases do not examine intragenic sequences. This database also groups results by COG (Clusters of Orthologous Groups) providing key gene function information (92). Lastly the database takes into account operon data from the DOOR Operon Database (93) to enhance its predictive capabilities. While other databases only look at the gene immediately following a predicted sigma-54 promoter site, this database takes into account all downstream genes from the same operon potentially giving a more accurate view of genes under direct control of sigma-54.

## 2.4 Results and Discussion

Approximately 60 prokaryotic genomes have been analyzed and their results posted online at http://www.sigma54.ca. The genomes were selected to represent a wide selection of sequenced prokaryotic genomes, including those that do not have sigma-54. Most bacterial phyla as well as a few archeal genomes are included.

All hits within 500 bp of a downstream transcription initiation site with a score larger than or equal to 65 were compiled and made available online at http://www.sigma54.ca. Intragenic hits were not excluded from the data set as long as they met the same criteria as other sites. That is, they had a score larger than or equal to 65 and were within 500 bp of an annotated transcription site. This is justified by the fact that up to 18% of sigma-54 promoters are located in intragenic regions (50).

A score of 65 was initially chosen to ensure that all potential sigma-54 sites would be picked up by the algorithm and that further filtering on the score could be used to determine the optimal score cut off value. The 500 bp cut-off upstream of the gene was chosen since most transcriptional start sites are believed to fall within this range.

### 2.4.1 Matrices

The genomes of the organisms in the database were analysed against the RponMatrix which was constructed using the sequence from 186 sigma-54 promoter sites (56). From the result set generated using the RponMatrix on the model organism *Myxococcus xanthus*, two additional matrices were generated using the top 20 and top 40 sigma-54 scoring sites as inputs.

The matrices can be downloaded from the URLs shown below:

RponMatrix:
http://www.sigma54.ca/promoterdata/Web/downloadmatrix.ashx?matrix=RPON.Matrix

Top20Matrix:
http://www.sigma54.ca/promoterdata/Web/downloadmatrix.ashx?matrix=MXanth.S54.Top20.Matrix

Top40Matrix:
http://www.sigma54.ca/promoterdata/Web/downloadmatrix.ashx?matrix=MXanth.S54.Top40.Matrix

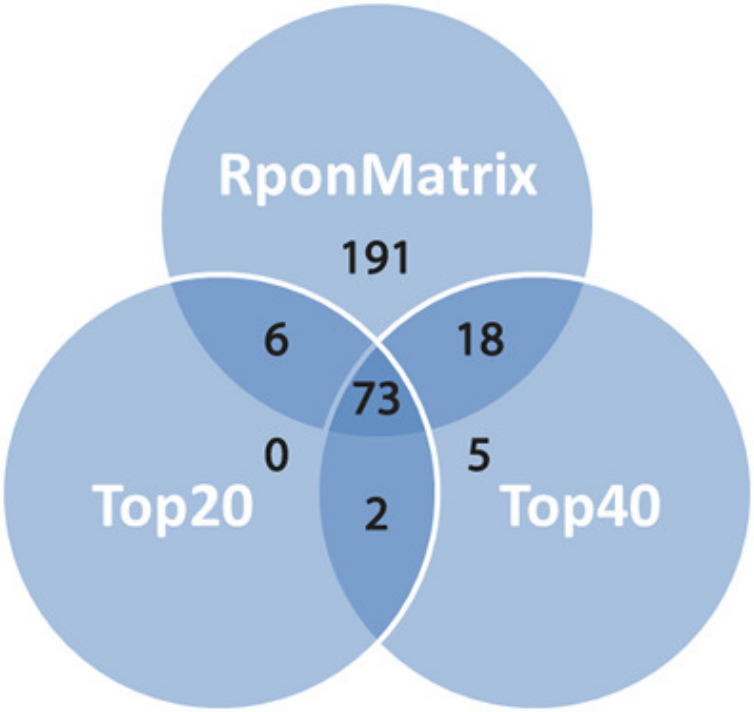



**Figure 26. Venn diagram showing the distribution of hits with scores ≥ 80 in *Myxococcus xanthus.***
Hits are categorized by the matrix used to perform the analysis. The RponMatrix predicted all but 7 of the sites. It, however, also predicted many more sites overall potentially indicating a high rate of false positives.

These matrices were scanned against the genome of *Myxococcus xanthus* in order to compare their results to determine the effect of using a different matrix on the result set. While the Top20Matrix and Top40Matrix produced less hits than the RpoNMatrix, only 7 sites where predicted by these other 2 matrices that were not also predicted by the RpoNMatrix. Therefore, the RponMatrix was selected as the primary matrix for the database since its result

set was the broadest and included almost all of the sites predicted by the two other matrices. In addition, this matrix has been previously used in similar analyses (91).

### 2.4.2 COG and COG Categories

COG and COG category data were collected at various score cut-offs. By looking at COG and COG category information, it is possible to get better overall of view of the function of sigma-54.

Database wide summaries were created that show which COGs are most often preceded by computationally predicted sigma-54 sites at a score larger than or equal to 75. Based on the COG data, it was discovered that there was a clear phylum based difference in the presence of sigma-54 promoter sites upstream of genes encoding COG1484L DNA replication protein (Table 1). In the Proteobacteria, most of the genes encoding this type of protein have a sigma-54 promoter binding site. However, only 1 gene in the Firmicutes and none in the Actinobacteria of the same COG code have sigma-54 promoters. This indicates a phylum divergence in the use of sigma-54.

**Table 1. Cross-phylum differences for COG1484L DNA replication protein.**
The first number indicates the total number of gene products that have a COG code of COG1484L and have a computationally predicted upstream $\sigma^{54}$ promoter site with a score larger than or equal to 80. The second number indicates the total number of gene products in the organism with that COG code.

| Cross-phylum differences for COG1484L DNA replication protein (Score ≥ 80) | |
|---|---|
| **All genomes** | 71/145 |
| **Proteobacteria** | 68/93 |
| **Firmicutes** | 1/15 |
| **Actinobacteria** | 0/20 |

While it is expected that there would not be sigma-54 promoter sites preceding these genes in Actinobacteria, as none of the Actinobacteria in the database have sigma-54, it is surprising to note the large difference between the Proteobacteria and the Firmicutes. Both of these phyla have organisms in the database that both contain and do not contain genes encoding sigma-54. In fact, 6 out of the 9 Firmicutes in the database encode a sigma-54 homolog. Despite this, gene products of type COG1484L in the Firmicutes are rarely preceded by a sigma-54 promoter, while similar genes in Proteobacteria are quite likely to have one.

This divergence suggests that sigma-54 may play different roles in each of these bacterial phyla. A better understanding of these divergences could lead to a better understanding of the current role, as well as the evolutionary role, of sigma-54 in bacterial organisms.

### 2.4.3 Relationship between GC Content and Predicted Binding Sites

It is important to note that scores generated by the PromScan algorithm are relative only to the sequence it was analyzed against. Scores from different sequences with different percentage GC content cannot be compared since the frequency of each base in the sequence is included in the scoring equation. Furthermore, sequences that share a similar GC content to the matrix are predicted to have more high scoring sites than sequences with more or less GC content than the matrix (Figure 27).

Given the need to closely examine the score cut-off, the adjusted hits (number of hits at a given score, divided by the genome length, times 10,000) was calculated for each organism and was plotted as a function of %GC content. As shown in Figure 27, there is a peak in the trend near the average GC content of the matrix. There is a positive linear relationship between GC content and number of adjusted hits up to the average matrix GC content of 56.6%, while afterwards there is a negative linear relationship.

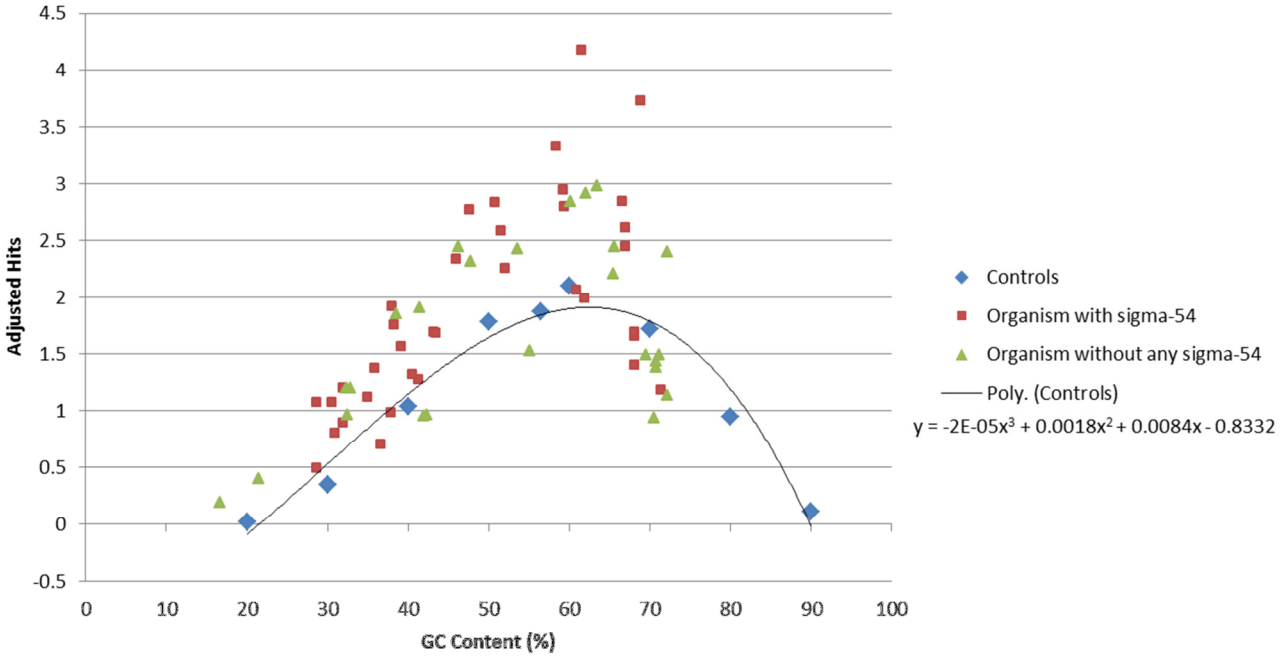



**Figure 27. Relationship between GC content and number of predicted binding sites.**
The adjusted hit score was determined by counting the number of sites in the genome with a score greater than or equal to 80. This number was then divided by the length of the sequence (genome size) and multiplied by 10,000 to bring the values to a more accessible range. Organisms that have a sigma-54 homologue are represented by red squares while those without a sigma-54 encoding gene are represented by green triangles. Randomly generated control sequences of varying %GC content are represented by blue diamonds. The curved line represents the line of best fit for the randomly generated controls and is described by the equation $y = -2E\text{-}05x^3 + 0.0018x^2 + 0.0084x - 0.8332$.

The fact that the number of hits reaches a maximum near the average GC content of the matrix even at high cut off scores indicates that there is some inherent matrix bias in the scoring algorithm.

While this relationship exists for all score cut-off values, it is more prominent at low cut-off score. While there is still a peak in the trend with higher cut-off scores, there is a much larger degree of spread. This indicates that the number of hits is no longer simply a function of the GC content of the organism. Indeed, only data from scores larger than or equal to at least 75, if not 80, should be used for any analysis as these values represent an ideal compromise between avoid both false positives and false negatives.

An interesting observation related to this involves the data points of *Streptomyces coelicolor*, an organism that does not contain sigma-54 (see Appendix 3: Charts of GC Content vs Adjusted Hits, p. 61). At low scores, *S. coelicolor* is an obvious anomaly to the general trend with a very high adjusted hits value relative to all the other organisms. As the score cut-off is increased, however, the data point for *S. coelicolor* comes into line with the remainder of the organisms. This highlights the importance of understanding the effect of score cut-off on hits. When using a low score cut-off, large numbers of putative sites are identified. However, many or most of these are likely false positives. At high score cut-off values, there is a risk of missing sigma-54 promoter sites resulting in false negatives. In the case of *S. coelicolor*, at low score cut off, it would seem to be an organism that has numerous sigma-54 sites despite the fact that it does not encode a sigma-54 homolog. This would, indeed, be an interesting finding. However, when a more stringent score is applied, the number of adjusted hits for *S. coelicolor* rejoins the general trend and indicates that the high number of adjusted hits seen at a low score cut off were due to false positives.

Intriguingly, whether the organisms contain or do not contain sigma-54 did not have an effect on their overall trends in terms of adjusted hits. A large part of this is likely due to matrix bias where the number of adjusted hits ends up being related more to the overall GC content of the organism than anything else.

Although more speculative in nature, another possible reason for the similarities between organisms that have a sigma-54 and those that don't is that sigma-54 regulation may have been a conserved mechanism among most or all bacteria, and that sigma-54 promoter sites still remain in organisms that have lost their gene that encodes for sigma-54. An additional hypothesis would be that another transcriptional regulator, present in the organisms that have lost sigma-54, has taken over the use of these promoter recognition sequences after the sigma-54 gene was lost.

### 2.4.4 Examination of False Positives and False Negatives

All experimentally validated sigma-54 promoters in the *E. coli* genome have scores ranging from the 80s to high 90s (50). In GC rich organisms such, as *M. xanthus,* a number of identified sigma-54 promoters have scores in the low to mid 70s (94, 95). Based on these data, a score of 75 is expected to provide the required low level of false negatives across a broad range of genomes.

To evaluate the level of false positives, a sequence randomized *E. coli* genome was analyzed. 25.6% of genes in the native *E. coli* genome and 18.4% of the genes in the sequence randomized genome were predicted to have sigma-54 promoters. 18.4% thus provides a baseline for the level false positive predictions. This high false positive rate is typical of PWM based promoter prediction tools (96, 97).

As false positives are not expected to be conserved across species, a comparative genomic approach that examines the predicted promoters across a wide range of bacterial genomes should decrease the false positive rate and increases the specificity (98). The data shows that more than 50 % of the bacterial genomes possessing a *glnA* ortholog (COG0174), the archetypical gene under direct sigma-54 transcriptional control, had one or more predicted sigma-54 promoters upstream of this gene. In comparison, only 10% of the genomes examined had predicted sigma-54 promoters upstream of the well-characterized non-sigma-54 regulated ribosomal proteins encoding genes (COG0093, COG0049) (99), transcription elongation factor encoding genes (COG0264, COG0231)(100, 101), and the ATP synthase encoding genes (COG005, COG0224) (102). Thus while individual promoter predictions have a high false positive rate, correlating promoter predictions across multiple species for individual orthologs substantially decreases the false positive predictions, and increases the ability to accurately predict orthologs with sigma-54 promoters.

### 2.4.5 Sigma-54 and PKS/NRPS Genes

One goal of this research was to see if there was any association between sigma-54 promoter sites and NRPS/PKS genes. The database is currently being used to look for sigma-54 promoters adjoining NRPS/PKS genes and operons and cursory analysis suggests that there are putative sigma-54 promoter sites located in over half of polyketide and non-ribosomal peptide biosynthetic pathways. One of these, the promoter upstream of the *oxyA* gene, has been shown to be regulated by sigma-54 over-expression, validating initial predictions from this database (53).

If sigma-54 promoters are commonly found in polyketide and non-ribosomal peptide biosynthetic pathways, sigma-54 over-expression may be a general route to heterologous

expression of polyketides and non-ribosomal peptides. To evaluate this possibility, the genomes of the 58 sequenced bacteria in the sigma-54 database where examined for the presence of sigma-54 promoters in polyketide and non-ribosomal peptide biosynthetic gene clusters. The species included major secondary metabolite producers such as 10 Actinobacteria, 20 Proteobacteria, 9 Firmicutes, as well as 19 species from diverse phyla. This selection of species contains representative examples of most of the sequenced bacterial phylogenic diversity.

Manual examination of all genes annotated as polyketide synthase, 3-oxoacyl acyl carrier protein synthase, and non-ribosomal peptide synthetase (COG3321, COG0304, and COG1020 respectively) (92) and their adjacent genes, identified 180 polyketide and non-ribosomal peptide biosynthetic gene clusters.

**Table 2. Sigma-54 distribution in bacterial gene clusters.**
Results of a bioinformatics analysis of 58 bacterial genomes containing 180 polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) biosynthetic gene clusters show that the majority of gene clusters contain putative sigma-54 promoters (score ≥ 75). It is particularly intriguing that Actinobacteria possess gene clusters with sigma-54 promoters since they lack the sigma-54 encoding gene.

| | strains | PKS/NRPS gene clusters | gene clusters with $\sigma^{54}$ promoters | percent with $\sigma^{54}$ promoters |
|---|---|---|---|---|
| ***Actinobacteria*** | 10 | 94 | 61 | 65% |
| ***Firmicutes*** | 9 | 15 | 11 | 73% |
| ***Proteobacteria*** | 20 | 49 | 36 | 74% |
| **Other phyla** | 19 | 22 | 16 | 73% |
| **Total** | **58** | **180** | **124** | **69%** |

Of the 180 polyketide and non-ribosomal peptide biosynthetic gene clusters identified, 124 (69%) contained one or more sigma-54 promoters appropriately positioned to regulate transcription (Table 2). 24 gene clusters possessed a single sigma-54 promoter, however the vast majority had two or more, with some pathways possessing up to two dozen sigma-54 promoters. These results demonstrate that a majority (69%) of polyketide and non-ribosomal peptide biosynthetic gene clusters possess putative sigma-54 promoters and that these promoters are appropriately positioned to regulate transcription of at least one operon in these gene clusters. These results are consistent with the hypothesis that sigma-54 over-expression may be a general method for ensuring transcription of polyketide and non-ribosomal peptide biosynthetic gene clusters in heterologous hosts, such as *E. coli*.

The identification of a functional sigma-54 promoter in the *S. rimosus* oxytetracycline biosynthetic gene cluster and 236 putative sigma-54 promoters in 61 gene clusters from seven different Actinobacterial genomes is highly unexpected. Actinobacteria do not contain a gene encoding sigma-54. Three hypotheses could account for this observation. These promoters are non-functional in Actinobacteria and are remnants of horizontal biosynthetic pathway transfer from non-Actinobacteria where these sigma-54 promoters are functional. Alternatively, a

functional equivalent of sigma-54 could be present. Extensive research efforts into regulation of polyketide biosynthesis in *Streptomyces* has not uncovered a functional equivalent to sigma-54, however the vast majority of these studies have been carried out in *S. coelicolor*, the only Streptomyces from the database to contain no sigma-54 promoters in any of its 10 polyketide and non-ribosomal peptide biosynthetic gene clusters. A third hypothesis is that these promoters represent false positives from the bioinformatics-based sigma-54 promoter prediction. While the prediction of promoters for orthologs across a wide range of bacterial species, as is done in this study, can decrease the false positive rate associate with positionally weighted matrix based analyses and improve the selectivity of genome wide predictions, experimentation is ultimately required to evaluate the functionality of individual promoters. Understanding the origins and roles of these putative sigma-54 promoter sequences thus remains a wide-open question.

### 2.4.6 Sigma-54 Directly Regulates Heterologous Production in *E. coli*

Sigma-54 can positively regulate transcription directly or indirectly. For direct regulation, the sigma-54-RNAP complex binds to a sigma-54 consensus promoter upstream of the operon under sigma-54 control. The sigma-54 consensus promoter is unique among bacterial promoters with highly conserved residues –12 and –24 from the transcriptional start site (T*GG*CACG-N4-TT*GC*(T/A)) (48) and can reliably be identified from sequence data. If sigma-54 is directly regulating transcription of *oxyB*, a sigma-54 promoter consensus sequence must be present upstream of the operon containing *oxyB*. To determine if sigma-54 promoters were present in the oxytetracycline biosynthetic pathway the entire gene cluster was examined for sigma-54 promoter consensus sequences using the PromScan algorithm (103).

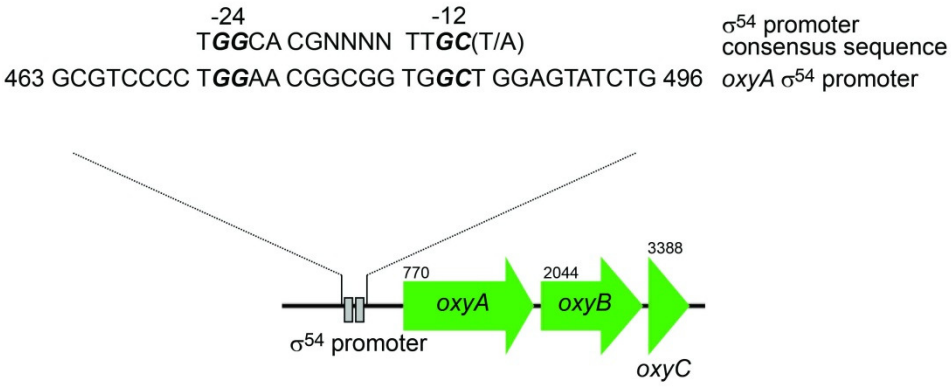



**Figure 28. A putative sigma-54 promoter exists in front of the *oxyABCDEF* operon in the oxytetracycline gene cluster from *S. rimosus*.**
Used with permission of the author.

The promoter shows high homology to the $\sigma^{54}$ promoter consensus sequence especially at the key -12 and -24 positions. This promoter could be responsible for direct, positive transcriptional regulation of the *oxyB* gene by $\sigma^{54}$ over-expression.

A high scoring consensus promoter was located upstream of the operon containing *oxyB* in a location that would allow direct transcription of the putative operon (Figure 28). The location and consensus sequence of the identified promoter provides strong evidence for direct regulation by sigma-54 of the heterologous production of oxytetracycline in *E. coli*.

## 2.5 Summary and Conclusion

A database of computationally determined sigma-54 promoter sites was created and was made available online at http://www.sigma54.ca. Distinctive to this database is the inclusion of intragenic regions as well as COG and COG category data that can be queried database wide, or by phylum. This unique online resource should provide fertile grounds for additional discovery related to the role of sigma-54 in bacteria.

## 2.6 Methods

### 2.6.1 General

Console programs were written in C# using Microsoft Visual C# 2008 Express Edition (free from http://www.microsoft.com/express/). The website and webservice were created in C# using Microsoft Visual Web Developer 2008 Edition (free from http://www.microsoft.com/express/) and the SQL database to store the data was created using SQL Server 2008 Express (free from http://www.microsoft.com/express/). Perl scripts were run using Strawberry Perl (free from http://strawberryperl.com/).

A C# console wrapper program was created calls a C# webservice, retrieves the organism's DNA sequence file (FASTA file from NCBI's Genome Database (10), http://www.ncbi.nlm.nih.gov/) and the sigma-54 matrix file which is based on 186 known sigma-54 promoter sites (6,8). The PromScan Perl script (see 2.6.2 Calculation of Predicted Sigma-54 Binding Sites) was run and its output file containing hits with scores larger than or equal to 65 was captured and sent back to the database through the webservice.

PTT files (from NCBI's Genome Database, http://www.ncbi.nlm.nih.gov/) containing annotation data were uploaded to the database. Using LINQ to SQL, hits generated from the

PromScan algorithm were linked to genes located within 500 bp upstream and in the same strand direction.

The complete list of COGs and COG categories were imported from NCBI's Clusters of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/COG/) (11).

Organisms were tagged as containing sigma-54 if they contained a gene named sigma-54 or if they had a gene product that matched COG1508, DNA-directed RNA polymerase specialized sigma subunit, sigma-54 homolog.

For each organism, the number of distinct gene products that matched a given COG was counted. The same process was repeated for COG categories.

Predictions from the DOOR Prokaryotic Operon Database (93) were used to define operons.

Nonribosomal peptide-synthetase (NRPS) and polyketide synthase (PKS) statistics were generated by comparing the number of hits of a given score that corresponded to NRPS/PKS gene products to the total number of gene products in the organism that corresponded to NRPS/PKS gene products. The COG codes used to identify NRPS/PKS gene products were:

- COG1020 Non-ribosomal peptide synthetase modules and related proteins
- COG2508 Regulator of polyketide synthase expression
- COG3208 Predicted thioesterase involved in non-ribosomal peptide biosynthesis
- COG3315 O-Methyltransferase involved in polyketide biosynthesis
- COG3319 Thioesterase domains of type I polyketide synthases or non-ribosomal peptide synthetases
- COG3320 Putative dehydrogenase domain of multifunctional non-ribosomal peptide synthetases and related enzymes
- COG3321 Polyketide synthase modules and related proteins

PKS/NRPS gene clusters were manually counted by searching the database for all genes with COG codes matching those listed above sorted by organism. A cluster was defined to be a region with PKS/NRPS genes, as annotated by COG code. Cluster boundaries were estimated by looking for 15 kbp or larger gaps in the presence of PKS/NRPS genes on either side of the cluster vc.

The data was aggregated by phylum. Database wide summaries were also prepared.
A website was setup at http://www.sigma54.ca to display results.

### 2.6.2 Calculation of Predicted Sigma-54 Binding Sites

Sigma-54 binding sites were predicted using PromScan (see Appendix 9: PromScan Perl Script, p. 105). PromScan is a Perl implementation of the calculation for the Kullback Liebler distance (see Figure 6, p. 5) that was developed by Studholme (103, 91). Before analyzing the DNA sequence, it determines the highest possible score and normalizes it to 100. It then scans the DNA sequence one nucleotide position at a time and assigns a normalized score for each frame of 16 nucleotides. It repeats the same procedure on the reverse strand. High scoring frames represent sequences that are highly similar to the matrix and are potential sigma-54 binding sites.

The PromScan Perl script was modified to output all hits within a genome, including intragenic hits, with a score of 65 or higher. It was run on Windows using Strawberry Perl (free from http://strawberryperl.com/). As inputs, the script used genomic DNA sequence files (FNA FASTA file from NCBI's FTP Genome Database, see details below) and the rpoN matrix file which is based on 186 known sigma-54 promoter sites (56). The script output consisting of a score, strand direction, and bp location for each hit was captured in a text file and imported into a SQL server database (Microsoft SQL Server 2008 R2 Express free from http://www.microsoft.com/express/). The results were cross referenced using annotation data from PTT files and RNT files containing gene and RNA gene data respectively (also from NCBI's FTP Genome Database).

# Chapter 3: ClusterMine360 Database

## 3.1 Introduction

The amount of information on microbial secondary metabolite biosynthesis has been growing explosively. Gene clusters responsible for the biosynthesis of polyketides and non-ribosomal peptides, identified by the presence of polyketide synthases (PKS) or non-ribosomal peptide synthetases (NRPS) encoding genes, have received significant attention, resulting in the sequencing of hundreds of gene clusters. With the power, speed and low cost of next generation sequencing methods, this number is expected to rapidly increase by at least an order of magnitude in the next few years.

To take advantage of this wealth of data, it needs to be easily accessible and discoverable. While the sequences themselves are available in NCBI databases (104, 105), they are frequently difficult to locate partially due to the large amounts of information that these databases host. There is no standardized annotation for these biosynthetic gene clusters. For example some are tagged with PKS and/or NRPS such as the cycloheximide (accession number JX014302) (106) and streptothricin (accession number AB684619) (107) gene clusters, while others are tagged with the term polyketide synthase or non-ribosomal peptide synthetase such as laidlomycin (accession number JQ793783) (108) and collismycin A (accession number HE575208) (109). With the rapid growth in bacterial genome sequencing, many new clusters are located within much larger genome sequence files and are occasionally unannotated, such as the antibiotic TA/myxovirescin biosynthetic gene cluster in the *Myxococcus xanthus* genome (accession number CP000113.1) (110). These problems are compounded by the fact that gene cluster discovery is being undertaken by researchers from diverse fields of expertise, including chemistry, biochemistry, microbiology, biotechnology, and drug discovery all with differing standards for gene cluster annotation. It is thus no surprise that given these issues, it can be extremely challenging, time consuming, and often frustrating to find appropriate genes cluster in the NCBI database.

In order to accelerate research and leverage existing data in PKS/NRPS biosynthesis, a focused and comprehensive database that gathers this gene cluster information together is required (29). While there are some existing databases that provide important resources on PKS/NRPS gene clusters and/or their products (62, 67, 111, 112), none have the features necessary to enable the community to maximize the benefit from sequence data. In particular, two key features have been identified that are required for the community. The first is to have a comprehensive up-to-date database. Because of the rapid emergence of new gene clusters across a broad range of disciplines, a resource that can be easily updated by any and all

community members is required to ensure that the database is both comprehensive and current. The second is that the difficulty in accessing multiple, diverse gene clusters has limited the ability of researchers to carry out comprehensive phylogenetic and functional analysis. Therefore the database must have the ability to generate multiple sequence FASTA files for individual catalytic domains found in PKS and NRPS biosynthesis.

In evaluating the existing PKS/NRPS databases, it was found that some of them, such as NRPS-PKS (62) and MAPSI (67), have not been updated in recent years. Others, such as NORINE (111), focus on the products of the cluster and do not contain information on the gene cluster itself. DoBISCUIT (http://www.bio.nite.go.jp/pks/) is a new and promising database but currently has limited amounts of data while PKMiner (112) is limited to Type II PKS clusters. Curated databases, such as those mentioned previously, can offer high levels of data quality but they are not always actively updated since few institutions or research groups have the resources to maintain ongoing manual curation. Additionally, there can be long lag times between the discovery of a new gene cluster and its inclusion in a traditionally curated database. Newly discovered clusters are often excluded from these databases as they do not meet curation criteria. For example, they may lack a characterized product as is seen for a large number of cryptic or silent gene clusters from whole genome sequencing efforts (36). The result of this is a bias towards a limited number of well known, archetypical clusters such as the erythromycin (accession number AY623658) (113) and tyrocidine (accession number AF004835) (114) gene clusters. This is a particularly important concern for researchers attempting to assign function to new gene clusters as well as those involved in bioprospecting as they need access to the breath and diversity of sequenced clusters and not simply the well-known prototypical textbook clusters. The best way to address these issues, which are limiting the research ability of the community, is to build a dynamic resource that allows users to make contributions, minimizes the amount of time-consuming manual curation by database administrators, but maintains the high standard of curated data quality.

New data, especially from bacterial genome sequencing, is being generated at an extraordinarily rapid rate (29). In order to keep up with this influx of data, while at the same time minimizing the amount of inefficient data entry, a server based workflow engine was developed to assist in curation of gene cluster data. Additionally, a community-based approach to the collection of data was adopted for this database. Researchers can sign up for a free account allowing them to add to or update the database. This crowd-sourcing allows participation by those who are most interested in using the data, ensuring broad coverage of the data across diverse fields, while decreasing the need for a dedicated full-time curator.

Community-based curation has some unique challenges. In particular, it can be difficult to ensure high levels of data quality (115, 116). To address this issue, the input from users is

limited such that only a few key details need be provided with the bulk of the data collection and analysis being performed in an automated fashion using known databases such as the NCBI databases and analysis tools including antiSMASH (38). The use of automation means the database can 'auto-curate' itself, reducing the amount of administrative burden and enabling the database to grow dynamically through community contributions.

## 3.2 Database Organization

The microbial PKS/NRPS database, ClusterMine360 (http://www.clustermine360.ca/), is organized around two key elements, the compound family and the gene cluster (see Figure 29).

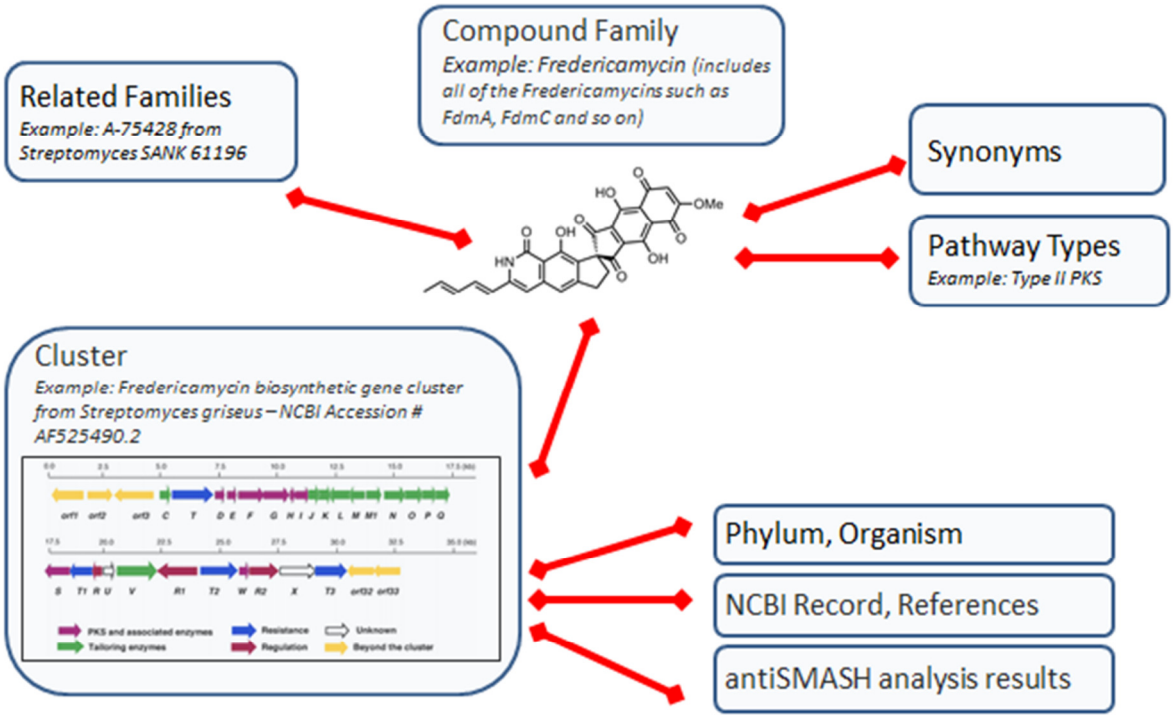



**Figure 29. Organization of ClusterMine360.**
The compound family and cluster represent the two major organization units of the database. Additional data fields connect to either the compound family or cluster. The organization of the fredericamycin gene cluster is shown in the cluster pane (117).

A compound family is a grouping of compounds that have the same core structure. This term is used since most gene clusters produce more than one compound although they tend to be highly related. As an example, the epothilone biosynthetic pathway produces four highly related polyketides, epothilones A-D, which differ by the presence or absence of a methyl

group and an epoxide moiety (118, 119). Thus by organizing by compound family, it is possible to capture the chemical diversity generated by a single biosynthetic gene cluster without duplicating data in the database. The 'Compound Families' page of the website has a listing of all of the families along with an image of the structure of a representative member of family (if available).

Since many natural products are known by more than one name, synonyms for each compound family can be added. This is essential to limit duplicate entries. For example the polyketide pimaricin is also widely known as natamycin. Prior to adding a new compound family, the list of existing names is checked to make sure it has not already been added. If the compound family has already been added under another name, the user is notified and is given the primary name for that family in the database. Additionally, the database queries ChemSpider to identify synonyms for each compound family and adds these to the compound family's details page, ensuring a comprehensive set of synonyms for compound family.

Because many compounds can be highly related, yet clearly not from the same compound family, each compound family can be linked to related families. For example erythromycin, megalomycin, and oleandomycin all share the same polyketide core but differ in the sugar residues attached to the core. These are clearly highly related compounds, thus they are linked together as related families. Identification of related families is highly subjective. While it is possible to evaluate similarity between structures using mathematical coefficients such as the Tanimoto similarity or Euclidian distance (120), no weighting scheme that captured the subjective relatedness of for example erythromycin, megalomycin, and oleandomycin without including for example methylmycin, narbomycin, pikromycin, or lankamycin was available. Compound families can also be related by similarities in the clusters that produce them. As part of the analysis undertaken by antiSMASH, it searches for similar clusters and the results of this are then used to automatically link the compound families. Links to related compound families are shown on the compound family's details page enabling users to easily access data for related compounds. To capture some of the broader relatedness between compound families, each family is associated with one or more overall biosynthetic pathway type such as PKS Type I, Type II, Type III, or NRPS. Clusters with both PKS and NRPS domains are identified as hybrid pathway. This enables the compound families to be rapidly sorted by a broad structural relatedness.

The second major organization unit of the database is the gene cluster. Multiple clusters can be associated with a given compound family. For example, epothilone biosynthetic gene clusters have been sequenced from two strains of *Sorangium cellulosum* (121, 122) and erythromycin gene clusters have been sequenced from both *Saccharopolyspora erythraea* (123) and *Aeromicrobium erythreum* (113). Each cluster is associated with an NCBI nucleotide

record. The NCBI record is used as the source for the lineage of the producing organism including the phylum, genus, and species. Links to primary literature references for the sequencing data are also retrieved from the NCBI record and displayed on the cluster's details page. Linked to each gene cluster is the annotation data for each gene in the cluster as well as each domain found in the PKS and NRPS encoding genes. This data is generated via antiSMASH analysis of each gene clusters (38). The domain sequences, extracted from the antiSMASH results, are also available from the gene cluster's details page.

## 3.3 Automation

Ensuring high data quality is very time consuming and makes database upkeep difficult. One of the most important requirements for the database was to integrate automation to make curation as easy as possible. Since most of the data is populated automatically, external users are able to contribute without much risk to data quality. This semi-automatic curation also means that large amounts of data can be added to the database in a relatively short amount of time.

The following steps occur once a cluster is added (see Figure 30). First, the NCBI nucleotide database is queried to retrieve important information about the sequence such as its description, the name and lineage of the organism it was isolated from, and any sequencing references that are associated with the record. Once this information has been retrieved, the cluster is submitted to antiSMASH for analysis. The database automatically tracks the progress of the antiSMASH submission and proceeds to download the results when completed. The results are then parsed to retrieve information such as the pathway types for that cluster, which is used to ensure that the pathway types of the linked compound family are correct. Finally, if antiSMASH has identified any PKS/NRPS domains, the amino acid sequence of those domains will be stored in the database's sequence repository along with key information such as domain substrate specificity, stereochemistry, and activity of the domain, as applicable. In addition, when a compound family is added, it is searched against the PubChem (124, 125) database to retrieve MeSH pharmacological identifiers that classify the compound's bioactivity. SMILES strings are also retrieved enabling users to search the databases by sub-structure. The typical time to complete these processes ranges from a few minutes to a few hours depending on server load.

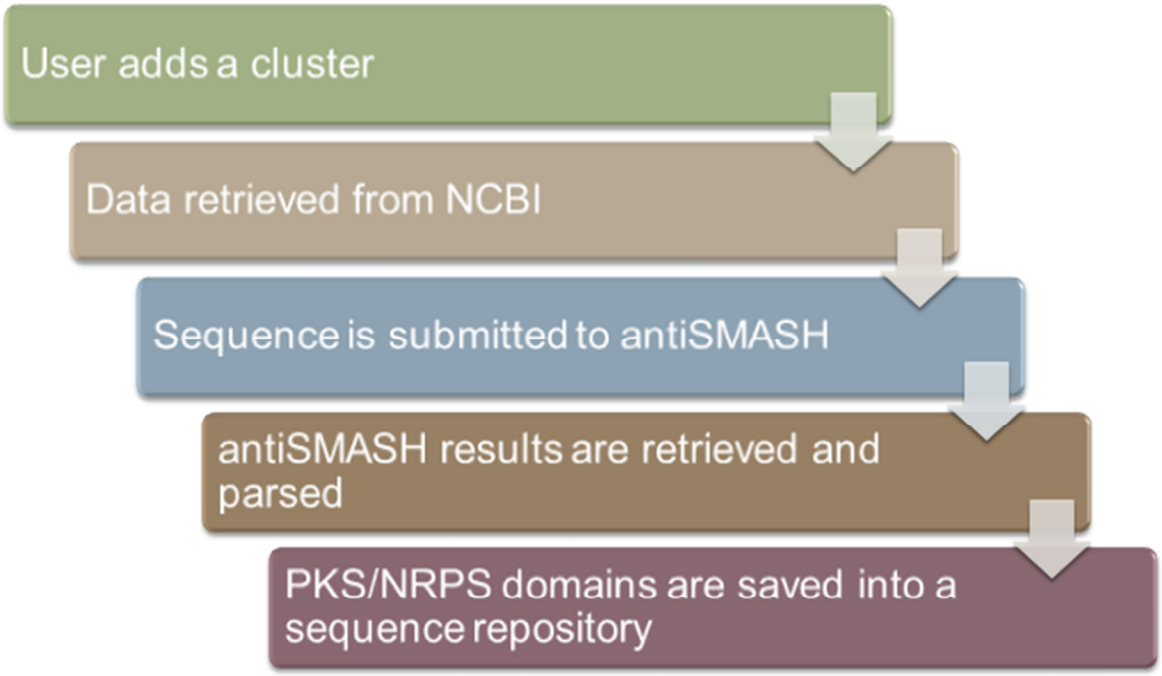



**Figure 30. ClusterMine360 Processing Steps**
ClusterMine360 has automated many of the steps required for curating the database. Automated curation is essential to enable crowd sourcing without sacrificing data quality.

In addition to the automated processes above, other features that make it particularly easy for users to add data have been incorporated. When a compound family is added to the database, a wizard guides the user through the process of entering information on pathway types, synonyms, and related families as well as helping the user in generating an image for the structure of the compound. To make it easy to associate an image, the ChemSpider database (http://www.chemspider.com) is queried to retrieve images that match the compound family name. Alternatively, an image can be generated from a user supplied SMILES string. Similarly, when adding synonyms, potential synonyms are returned from ChemSpider and the user can easily select those that are applicable.

## 3.4 antiSMASH

antiSMASH (*anti*biotics & *S*econdary *M*etabolite *A*nalysis *Sh*ell) is the bioinformatics tool used to provide analysis on clusters. antiSMASH can scan a cluster's sequence and determine the most likely pathway type for that cluster. For Type I PKS clusters, it also attempts to predict whether it is modular, iterative, or has trans-ATs. It is also able to make predictions for individual domains. It endeavours to determine the substrate specificity for AT and adenylation domains. For KR domains, it assesses whether it is active or inactive, and the probable stereochemistry of the product. More details can be found in (38). To ensure the

standardization of the large amounts of data in the database and to minimize manual curation, the results retrieved from antiSMASH by ClusterMine360 cannot be edited by individual users to include new biochemistry. However, as newly characterized PKS/NRPS domains are added to antiSMASH, the clusters in the database can be easily re-analyzed to take advantage of the improved analytics.

## 3.5 User Contributions

User contributions to the database are encouraged and acknowledged. In order to contribute to the database, users must register for a free account using a simple registration form. The name of the contributor, the name of their research group, and a link to their webpage is displayed on records that they have added to the database.

## 3.6 Present Content

Currently, the database has over 185 unique compound families, more than 200 clusters with known products, and over 300 clusters with no known products (silent or cryptic gene clusters). The sequence repository has 10,000+ PKS/NRPS domains from over 500 clusters available for download including 1300+ ACPs, 1000+ ATs, 1000+ KRs, 1300+ KSs, 250+ TEs, along with sequences from less common domains such as heterocyclization and epimerization domains.

## 3.7 Sequence Repository

One of the most unique aspects of this database is its sequence repository. The repository contains a large number of diverse PKS/NRPS domains extracted from the antiSMASH analysis of the clusters contained in the database. The ability to scan any NCBI nucleotide record and have the detected PKS/NRPS included in this repository is also included. This repository will likely become an invaluable tool to those involved in identifying sequence homologies and bioprospecting. The sequences can be downloaded individually in FASTA format. Alternatively, all of the domains in a given cluster can be downloaded at once in a zip file. Also included is the ability to filter the domains based on a variety of criteria following which they can be downloaded in a multi-sequence FASTA file. Importantly, the depth of information included in each sequence's header is exceptional. They are full of rich information such as accession number, producing organism, gene identifier, pathway type, domain type, as well as any predicted properties of that domain. There is also an option to output shortened headers for use with bioinformatics tools that have restrictions on the number of characters in the header.

## 3.8 ClusterMine360: A Powerful Tool For Phylogenetic Analysis

To demonstrate the utility of the ClusterMine360 database, NRPS heterocyclization domains were selected and used for cluster analysis. Heterocyclization domains play a key role in NRPS biosynthesis, coupling acyl and peptidyl groups onto Cys, Ser, and Thr residues followed by cyclization of the associated side chain to generate thiazol and oxazole rings (126–128) . This occurs during the biosynthesis of non-ribosomal peptides such as the antibiotic bacitracin and mixed non-ribosomal peptide/polyketides such as the antimitotic agents epothilone and rhizoxin.

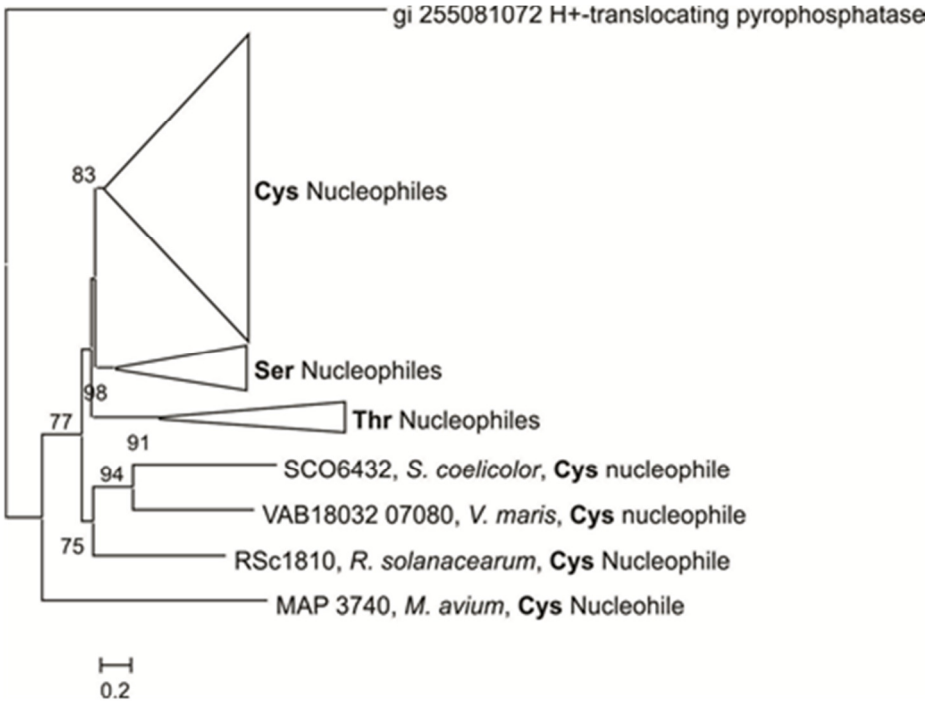



**Figure 31. Phylogenetic tree of heterocyclization domains from NRPS gene clusters.**
A rooted phylogenetic tree of heterocyclization domains from NRPS gene clusters shows that heterocyclization domains tree based on function. ClusterMine360 provides a rapid and powerful tool for generating and analyzing phylogenetic trees of PKS and NRPS domains.

A FASTA file of 106 heterocyclization domains was downloaded and aligned using MUSCLE (25). A phylogenetic tree was generated from the resulting alignment using the PhyML maximum likelihood method with the WAG model of amino acid substitution and Nearest Neighbor Interchange for the tree topology search (129). The tree shows that heterocyclization domains clustered by function, based on whether the domain used enzyme bound Cys, Ser or Thr as its substrate (Figure 31).

To evaluate which residues each heterocyclization domain used, the “detail of cluster” function in the sequence repository was examined to identify the specificity of adenylation

domain associated with the heterocyclization domain. Based on this analysis, the tree shows that Cys, Ser, and Thr specific heterocyclization domains all tree apart from each other. This analysis shows that with ClusterMine360 it is possible to rapidly develop phylogenetic tools to predict the function of an individual domain.

## 3.9 Conclusion

ClusterMine360 is a unique database of microbial PKS/NRPS clusters. It contains over 200 clusters from than 185 compound families and features a unique sequence repository containing over 10,000 PKS/NRPS domains. By leveraging automation and crowd-sourcing, it is hoped that this database will grow dynamically through contributions from interested parties as new clusters are discovered and sequenced. This database will be a useful resource for members of the PKS/NRPS research community enabling them to keep up with the growing number of sequenced gene clusters and rapidly mine these clusters for functional information.

## 3.10 Acknowledgements

The antiSMASH development team is thanked for providing an excellent tool to the natural products community. Kai Blin, in particular, deserves particular praise for his assistance in integrating the database with antiSMASH.

The team at GGA Software Services created and maintains the Indigo open-source chemistry toolkit (http://ggasoftware.com/opensource/indigo) which was used for sub-structure searching and for generating images from SMILES strings.

In addition, Dr Paul Thiessen from NIH/NLM/NCBI provided invaluable assistance with regards to interacting with the PubChem REST interface.

# Chapter 4: Contributions to Original Research

- Developed a database of sigma-54 promoters with results from over 60 prokaryotic organisms
  - Unique Features
    - The only sigma-54 promoter database currently available online
    - Includes intragenic results and low scoring sites for completeness
    - Results available by gene and by operon
    - Results can be grouped by COG (Clusters of Orthologous Groups) giving gene function information
    - Results are downloadable in spreadsheet format
  - Used to estimate the presence of putative sigma-54 promoter sites in PKS/NRPS gene clusters
  - Predicted the presence of a key sigma-54 promoter site in front of *oxyA* in the oxytetracycline gene cluster
    - Helped to explain the observation that oxytetracycline can be produced heterologously in *E. coli* when sigma-54 is overexpressed

- Created a database of microbial PKS/NRPS gene clusters
  - Currently the largest online database of PKS/NRPS biosynthesis
  - Has a unique 'crowd-sourcing' approach to data entry
  - Highly automated
  - Integrated with antiSMASH to provide exceptional analysis
  - Contains a unique sequence repository of over 10,000 PKS/NRPS domains
  - An important resource for the PKS/NRPS research community

# References


1. Holley,R.W., Apgar,J., Everett,G. a, Madison,J.T., Marquisee,M., Merrill,S.H., Penswick,J.R. and Zamir,A. (1965) Structure of a Ribonucleic Acid. *Science*, **147**, 1462–5.

2. Hutchison,C. a (2007) DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.*, **35**, 6227–37.

3. Sanger,F., Air,G.M., Barrell,B.G., Brown,N.L., Coulson,A.R., Fiddes,J.C., Hutchison,C.A., Slocombe,P.M. and Smith,M. (1977) Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature*, **265**, 687–695.

4. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R. a, Kirkness,E.F., Kerlavage, a R., Bult,C.J., Tomb,J.F., Dougherty,B. a and Merrick,J.M. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.

5. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M., et al. (1996) Life with 6000 genes. *Science*, **274**, 546–567.

6. The C. elegans Sequencing Consortium (1998) Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science*, **282**, 2012–2018.

7. Adams,M.D. (2000) The Genome Sequence of Drosophila melanogaster. *Science*, **287**, 2185–2195.

8. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A., et al. (2001) The sequence of the human genome. *Science*, **291**, 1304–51.

9. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

10. Wetterstrand,K. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. *www.genome.gov/sequencingcosts*.

11. Smaglik,P. (2002) Converging on DNA. *Nature*, **420**, 3–3.

12. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

13. Akalın,P.K. (2006) Introduction to bioinformatics. *Mol. Nutr. Food Res.*, **50**, 610–619.

14. Brazma, a, Jonassen,I., Eidhammer,I. and Gilbert,D. (1998) Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.*, **5**, 279–305.

15. Berg,O.G., Hippel,P.H. Von and Von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–50.

16. Touzet,H. and Varré,J.-S. (2007) Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol. Biol.*, **2**, 15.

17. Claverie,J.M. and Audic,S. (1996) The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.*, **12**, 431–9.

18. Staden,R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Bioinformatics*, **5**, 89–96.

19. Edgar,R.C. and Batzoglou,S. (2006) Multiple sequence alignment. *Curr. Opin. Struct. Biol.*, **16**, 368–73.

20. Just,W. (2001) Computational complexity of multiple sequence alignment with SP-score. *J. Comput. Biol.*, **8**, 615–23.

21. Taylor,W.R. (1988) A flexible method to align large numbers of biological sequences. *J. Mol. Evol.*, **28**, 161–9.

22. Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–44.

23. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–80.

24. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–17.

25. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–7.

26. Feng,D.F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–60.

27. Yang,Z. and Rannala,B. (2012) Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.*, **13**, 303–14.

28. Tamura,K., Peterson,D., Peterson,N., Stecher,G., Nei,M. and Kumar,S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, **28**, 2731–9.

29. Jenke-Kodama,H. and Dittmann,E. (2009) Bioinformatic perspectives on NRPS/PKS megasynthases: advances and challenges. *Nat. Prod. Rep.*, **26**, 874–83.

30. Baum,L.E. and Petrie,T. (1966) Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Ann. Math Stat*, **37**, 1554–1563.

31. Rabiner,L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *P. IEEE*, **77**, 257–286.

32. Eddy,S. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

33. Baldi,P., Chauvin,Y., Hunkapiller,T. and Mcclure,M.A. (1994) Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 1059–1063.

34. Berdy,J. (2005) Bioactive microbial metabolites. *J. Antibiot.*, **58**, 1–26.

35. Newman,D.J. and Cragg,G.M. (2007) Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod.*, **70**, 461–77.

36. Challis,G.L. (2008) Mining microbial genomes for new natural products and biosynthetic pathways. *Microbiology*, **154**, 1555–69.

37. Cundliffe,E. (2006) Antibiotic production by actinomycetes: the Janus faces of regulation. *J. Ind. Microbiol. Biotechnol.*, **33**, 500–6.

38. Medema,M.H., Blin,K., Cimermancic,P., De Jager,V., Zakrzewski,P., Fischbach,M.A., Weber,T., Takano,E. and Breitling,R. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.*, **39**, W339–W346.

39. Nair,D.R., Anand,S., Verma,P., Mohanty,D. and Gokhale,R.S. (2012) Genetic, biosynthetic and functional versatility of polyketide synthases. *Curr. Sci. India*, **102**, 277–287.

40. Ridley,C.P., Lee,H.Y. and Khosla,C. (2008) Evolution of polyketide synthases in bacteria. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 4595–600.

41. Jenke-Kodama,H., Sandmann,A., Müller,R. and Dittmann,E. (2005) Evolutionary implications of bacterial polyketide synthases. *Mol. Biol. Evol.*, **22**, 2027–39.

42. McDaniel,R., Ebert-Khosla,S., Hopwood,D. a and Khosla,C. (1993) Engineered biosynthesis of novel polyketides. *Science*, **262**, 1546–50.

43. Medema,M.H., Van Raaphorst,R., Takano,E. and Breitling,R. (2012) Computational tools for the synthetic design of biochemical pathways. *Nat. Rev. Microbiol.*, **10**, 191–202.

44. Wong,F.T. and Khosla,C. (2012) Combinatorial biosynthesis of polyketides--a perspective. *Curr. Opin. Chem. Biol.*, **16**, 117–23.

45. Buck,M. and Cannon,W. (1992) Specific binding of the transcription factor sigma-54 to promoter DNA. *Nature*, **358**, 422–4.

46. Doucleff,M., Pelton,J.G., Lee,P.S., Nixon,B.T. and Wemmer,D.E. (2007) Structural basis of DNA recognition by the alternative sigma-factor, sigma54. *J. Mol. Biol.*, **369**, 1070–8.

47. Bose,D., Pape,T., Burrows,P.C., Rappas,M., Wigneshweraraj,S.R., Buck,M. and Zhang,X. (2008) Organization of an activator-bound RNA polymerase holoenzyme. *Mol. Cell*, **32**, 337–46.

48. Buck,M., Gallegos,M. and Studholme,D. (2000) The Bacterial Enhancer-Dependent sigma-54 (sigma-N) Transcription Factor. *J. Bacteriol.*, **182**, 4129–4136.

49. Francke,C., Groot Kormelink,T., Hagemeijer,Y., Overmars,L., Sluijter,V., Moezelaar,R. and Siezen,R.J. (2011) Comparative analyses imply that the enigmatic Sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics*, **12**, 385.

50. Zhao,K., Liu,M. and Burgess,R.R. (2010) Promoter and regulon analysis of nitrogen assimilation factor, sigma54, reveal alternative strategy for E. coli MG1655 flagellar biosynthesis. *Nucleic Acids Res.*, **38**, 1273–83.

51. Zhang,H., Gao,S., Lercher,M.J., Hu,S. and Chen,W.-H. (2012) EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. *Nucleic Acids Res.*, **40**, W569–72.

52. Letunic,I. and Bork,P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, **39**, W475–8.

53. Stevens,D.C., Conway,K., Pearce,N., Garza,A. and Boddy,C.N. (2012) Alternative Sigma Factor Over-Expression Enables Heterologous Expression of a Type II Polyketide Biosynthetic Pathway in Escherichia coli. *PloS ONE*, **Submitted**.

54. Stevens,D.C., Henry,M.R., Murphy,K. a and Boddy,C.N. (2010) Heterologous expression of the oxytetracycline biosynthetic pathway in Myxococcus xanthus. *Appl. Environ. Microbiol.*, **76**, 2681–3.

55. Zhang,W., Ames,B.D., Tsai,S.-C. and Tang,Y. (2006) Engineered biosynthesis of a novel amidated polyketide, using the malonamyl-specific initiation module from the oxytetracycline polyketide synthase. *Appl. Environ. Microbiol.*, **72**, 2573–80.

56. Barrios,H., Valderrama,B. and Morett,E. (1999) Compilation and analysis of sigma(54)-dependent promoter sequences. *Nucleic Acids Res.*, **27**, 4305–13.

57. Salomon,C.E., Magarvey,N. a, Sherman,D.H. and Magarvey,A. (2004) Merging the potential of microbial genetics with biological and chemical diversity: an even brighter future for marine natural product drug discovery. *Nat. Prod. Rep.*, **21**, 105–21.

58. Walsh,C.T. (2008) The chemical versatility of natural-product assembly lines. *Acc. Chem. Res.*, **41**, 4–10.

59. Cane,D., Walsh,C. and Khosla,C. (1998) Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science*, **282**, 63–68.

60. Zucko,J., Starcevic,A., Diminic,J., Elbekali,M., Lisfi,M., Long,P.F., Cullum,J. and Hranueli,D. (2010) From DNA Sequences to Chemical Structures – Methods for Mining Microbial Genomic and Metagenomic Data Sets for New Natural Products. *Food Technol. Biotech.*, **9862**, 234–242.

61. Yadav,G., Gokhale,R.S. and Mohanty,D. (2003) SEARCHPKS: a program for detection and analysis of polyketide synthase domains. *Nucleic Acids Res.*, **31**, 3654–3658.

62. Ansari,M.Z., Yadav,G., Gokhale,R.S. and Mohanty,D. (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res.*, **32**, W405–13.

63. Kamra,P., Gokhale,R.S. and Mohanty,D. (2005) SEARCHGTr: a program for analysis of glycosyltransferases involved in glycosylation of secondary metabolites. *Nucleic Acids Res.*, **33**, W220–5.

64. Rausch,C., Weber,T., Kohlbacher,O., Wohlleben,W. and Huson,D.H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.*, **33**, 5799–808.

65. Starcevic,A., Zucko,J., Simunkovic,J., Long,P.F., Cullum,J. and Hranueli,D. (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic Acids Res.*, **36**, 6882–92.

66. Li,M.H.T., Ung,P.M.U., Zajkowski,J., Garneau-Tsodikova,S. and Sherman,D.H. (2009) Automated genome mining for natural products. *BMC Bioinf.*, **10**, 185.

67. Tae,H., Sohng,J.K. and Park,K. (2009) MapsiDB: an integrated web database for type I polyketide synthases. *Bioprocess Biosyst. Eng.*, **32**, 723–7.

68. Weber,T., Rausch,C., Lopez,P., Hoof,I., Gaykova,V., Huson,D.H. and Wohlleben,W. (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J. Biotechnol.*, **140**, 13–7.

69. Anand,S., Prasad,M.V.R., Yadav,G., Kumar,N., Shehara,J., Ansari,M.Z. and Mohanty,D. (2010) SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Res.*, **38**, W487–96.

70. Khaldi,N., Seifuddin,F.T., Turner,G., Haft,D., Nierman,W.C., Wolfe,K.H. and Fedorova,N.D. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.*, **47**, 736–41.

71. Röttig,M., Medema,M.H., Blin,K., Weber,T., Rausch,C. and Kohlbacher,O. (2011) NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.*, **39**, W362–7.

72. Prieto,C., García-Estrada,C., Lorenzana,D. and Martín,J.F. (2012) NRPSsp: non-ribosomal peptide synthase substrate predictor. *Bioinformatics*, **28**, 426–7.

73. Bachmann,B.O. and Ravel,J. (2009) Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. In *Methods in enzymology*. Elsevier Inc., Vol. 458, pp. 181–217.

74. Callahan,B., Thattai,M. and Shraiman,B.I. (2009) Emergent gene order in a model of modular polyketide synthases. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 19410–5.

75. Fischbach,M. a and Walsh,C.T. (2006) Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.*, **106**, 3468–96.

76. Wenzel,S.C. and Müller,R. (2005) Formation of novel secondary metabolites by bacterial multimodular assembly lines: deviations from textbook biosynthetic logic. *Curr. Opin. Chem. Biol.*, **9**, 447–58.

77. Haynes,S.W. and Challis,G.L. (2007) Non-linear enzymatic logic in natural product modular mega-synthases and -synthetases. *Curr. Opin. Drug Discovery Dev.*, **10**, 203–18.

78. Yadav,G., Gokhale,R.S. and Mohanty,D. (2009) Towards prediction of metabolic products of polyketide synthases: an in silico analysis. *PLoS Comput. Biol.*, **5**, e1000351.

79. Marahiel,M., Mootz,D. and Stachelhause,T. (1999) The specificity-conferring code of adenylation nonribosomal peptide synthetases. *Chem. Biol.*, **6**, 493–505.

80. Challis,G.L., Ravel,J. and Townsend,C. a (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.*, **7**, 211–24.

81. Minowa,Y., Araki,M. and Kanehisa,M. (2007) Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.*, **368**, 1500–17.

82. Jenke-Kodama,H., Börner,T. and Dittmann,E. (2006) Natural biocombinatorics in the polyketide synthase genes of the actinobacterium Streptomyces avermitilis. *PLoS Comput. Biol.*, **2**, e132.

83. Rausch,C., Hoof,I., Weber,T., Wohlleben,W. and Huson,D.H. (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.*, **7**, 78.

84. Ziemert,N., Podell,S., Penn,K., Badger,J.H., Allen,E. and Jensen,P.R. (2012) The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PloS One*, **7**, e34064.

85. Dreier,J. and Khosla,C. (2000) Mechanistic analysis of a type II polyketide synthase. Role of conserved residues in the beta-ketoacyl synthase-chain length factor heterodimer. *Biochemistry*, **39**, 2088–95.

86. Szu,P.-H., Govindarajan,S., Meehan,M.J., Das,A., Nguyen,D.D., Dorrestein,P.C., Minshull,J. and Khosla,C. (2011) Analysis of the ketosynthase-chain length factor heterodimer from the fredericamycin polyketide synthase. *Chem. Biol.*, **18**, 1021–31.

87. Tang,Y., Lee,T.S., Kobayashi,S. and Khosla,C. (2003) Ketosynthases in the initiation and elongation modules of aromatic polyketide synthases have orthogonal acyl carrier protein specificity. *Biochemistry*, **42**, 6588–95.

88. Caffrey,P. (2003) Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. *ChemBioChem*, **4**, 654–7.

89. Caffrey,P. (2005) The stereochemistry of ketoreduction. *Chem. Biol.*, **12**, 1060–2.

90. Keatinge-Clay,A.T. (2007) A tylosin ketoreductase reveals how chirality is determined in polyketides. *Chem. Biol.*, **14**, 898–908.

91. Studholme,D.J. and Dixon,R. (2003) Domain Architectures of s54 -Dependent Transcriptional Activators. *J. Bacteriol.*, **185**, 1757–1767.

92. Tatusov,R.L. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.

93. Mao,F., Dam,P., Chou,J., Olman,V. and Xu,Y. (2009) DOOR: a database for prokaryotic operons. *Nucleic Acids Res.*, **37**, D459–63.

94. Ossa,F., Diodati,M.E., Caberoy,N.B., Giglio,K.M., Edmonds,M., Singer,M. and Garza,A.G. (2007) The Myxococcus xanthus Nla4 protein is important for expression of stringent response-associated genes, ppGpp accumulation, and fruiting body development. *J. Bacteriol.*, **189**, 8474–83.

95. Giglio,K.M., Eisenstatt,J. and Garza,A.G. (2010) Identification of enhancer binding proteins important for Myxococcus xanthus development. *J. Bacteriol.*, **192**, 360–4.

96. Gordon,J.J., Towsey,M.W., Hogan,J.M., Mathews,S.A. and Timms,P. (2006) Improved prediction of bacterial transcription start sites. *Bioinformatics*, **22**, 142–8.

97. McLeay,R.C., Leat,C.J. and Bailey,T.L. (2011) Tissue-specific prediction of directly regulated genes. *Bioinformatics*, **27**, 2354–60.

98. Qiu,P. (2003) Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem. Biophys. Res. Commun.*, **309**, 495–501.

99. Nomura,M., Gourse,R. and Baughman,G. (1984) Regulation of the synthesis of ribosomes and ribosomal components. *Annu. Rev. Biochem.*, **53**, 75–117.

100. Aseev,L. V, Levandovskaya,A.A., Tchufistova,L.S., Scaptsova,N. V and Boni,I. V (2008) A new regulatory circuit in ribosomal protein operons: S2-mediated control of the rpsB-tsf expression in vivo. *RNA*, **14**, 1882–94.

101. Aoki,H., Adams,S.L., Chung,D.G., Yaguchi,M., Chuang,S.E. and Ganoza,M.C. (1991) Cloning, sequencing and overexpression of the gene for prokaryotic factor EF-P involved in peptide bond synthesis. *Nucleic Acids Res.*, **19**, 6215–20.

102. Kumar,R. and Shimizu,K. (2011) Transcriptional regulation of main metabolic pathways of cyoA, cydB, fnr, and fur gene knockout Escherichia coli in C-limited and N-limited aerobic continuous cultures. *Microb. Cell Fact.*, **10**, 3.

103. Studholme,D., Buck,M. and Nixon,T. (2000) Identification of potential sigma N-dependent promoters in bacterial genomes. *Microbiology*, **146**, 3021–3023.

104. Benson,D. a, Karsch-Mizrachi,I., Clark,K., Lipman,D.J., Ostell,J. and Sayers,E.W. (2012) GenBank. *Nucleic Acids Res.*, **40**, D48–53.

105. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–5.

106. Shen,B. and Yin,M. Unpublished results.

107. Maruyama,C., Toyoda,J., Kato,Y., Izumikawa,M., Takagi,M., Shinya,K., Katano,H., Utagawa,T. and Hamano,Y. Unpublished results.

108. Hwang,J.Y., Kim,H.S., Sedai,B. and Nam,D.H. Unpublished results.

109. Garcia,I., Vior,N.M., Braña,A.F., González-Sabin,J., Rohr,J., Moris,F., Méndez,C. and Salas,J. a (2012) Elucidating the biosynthetic pathway for the polyketide-nonribosomal peptide collismycin A: mechanism for formation of the 2,2'-bipyridyl ring. *Chem. Biol.*, **19**, 399–413.

110. Goldman,B.S., Nierman,W.C., Kaiser,D., Slater,S.C., Durkin, a S., Eisen,J. a, Eisen,J., Ronning,C.M., Barbazuk,W.B., Blanchard,M., et al. (2006) Evolution of sensory complexity recorded in a myxobacterial genome. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 15200–5.

111. Caboche,S., Pupin,M., Leclère,V., Fontaine,A., Jacques,P. and Kucherov,G. (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.*, **36**, D326–31.

112. Yi,G.-S. and Kim,J. (2012) PKMiner: a database for exploring type II polyketide synthases. *BMC Microbiol.*, **12**, 169.

113. Brikun,I. a, Reeves,A.R., Cernota,W.H., Luu,M.B. and Weber,J.M. (2004) The erythromycin biosynthetic gene cluster of Aeromicrobium erythreum. *J. Ind. Microbiol. Biotechnol.*, **31**, 335–44.

114. Mootz,H.D. and Marahiel,M. a (1997) The tyrocidine biosynthesis operon of Bacillus brevis: complete nucleotide sequence and biochemical characterization of functional internal adenylation domains. *J. Bacteriol.*, **179**, 6843–50.

115. Bücheler,T. and Sieg,J.H. (2011) Understanding Science 2.0: Crowdsourcing and Open Innovation in the Scientific Method. *Procedia. Comput. Sci.*, **7**, 327–329.

116. Meyer,P., Hoeng,J., Rice,J.J., Norel,R., Sprengel,J., Stolle,K., Bonk,T., Corthesy,S., Royyuru,A., Peitsch,M.C., et al. (2012) Industrial methodology for process verification in research (IMPROVER): toward systems biology verification. *Bioinformatics*, **28**, 1193–201.

117. Shen,B., Wendt-Pienkowski,E., Huang,Y., Zhang,J., Li,B., Jiang,H., Kwon,H. and Hutchinson,C.R. (2005) Cloning, sequencing, analysis, and heterologous expression of the fredericamycin biosynthetic gene cluster from Streptomyces griseus. *J. Am. Chem. Soc.*, **127**, 16442–52.

118. Gerth,K., Bedorf,N., Höfle,G., Irschik,H. and Reichenbach,H. (1996) Epothilons A and B: antifungal and cytotoxic compounds from Sorangium cellulosum (Myxobacteria). Production, physico-chemical and biological properties. *J. Antibiot.*, **49**, 560–3.

119. Hardt,I., Steinmetz,H. and Gerth,K. (2001) New Natural Epothilones from Sorangium c ellulosum, Strains So ce90/B2 and So ce90/D13: Isolation, Structure Elucidation, and SAR Studies. *J. Nat. Prod.*, **64**.

120. Willett,P. (2000) Chemoinformatics - similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol.*, **11**, 85–8.

121. Molnár,I., Schupp,T., Ono,M., Zirkle,R., Milnamow,M., Nowak-Thompson,B., Engel,N., Toupet,C., Stratmann,A., Cyr,D.D., et al. (2000) The biosynthetic gene cluster for the microtubule-stabilizing agents epothilones A and B from Sorangium cellulosum So ce90. *Chem. Biol.*, **7**, 97–109.

122. Tang,L., Shah,S., Chung,L., Carney,J. and Katz,L. (2000) Cloning and heterologous expression of the epothilone gene cluster. *Science*, **287**, 640–642.

123. Oliynyk,M., Samborskyy,M., Lester,J.B., Mironenko,T., Scott,N., Dickens,S., Haydock,S.F. and Leadlay,P.F. (2007) Complete genome sequence of the erythromycin-producing bacterium Saccharopolyspora erythraea NRRL23338. *Nat. Biotechnol.*, **25**, 447–53.

124. Bolton,E., Wang,Y., Thiessen,P. and Bryant,S. (2008) PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.*, **4**, 217–241.

125. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–33.

126. Chen,H., O'Connor,S., Cane,D.E. and Walsh,C.T. (2001) Epothilone biosynthesis: assembly of the methylthiazolylcarboxy starter unit on the EpoB subunit. *Chem. Biol.*, **8**, 899–912.

127. Kelly,W.L., Hillson,N.J. and Walsh,C.T. (2005) Excision of the epothilone synthetase B cyclization domain and demonstration of in trans condensation/cyclodehydration activity. *Biochemistry*, **44**, 13385–93.

128. Duerfahrt,T., Eppelmann,K., Müller,R. and Marahiel,M.A. (2004) Rational design of a bimodular model system for the investigation of heterocyclization in nonribosomal peptide biosynthesis. *Chem. Biol.*, **11**, 261–71.

129. Guindon,S., Dufayard,J.-F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–21.

# Supplementary Data

## Index to Supplementary Data

# Appendix 1: List of Organisms in the Sigma-54 Promoter Database

**Table S - 1. List of organisms in the sigma-54 promoter database.**

| Phylum | Genome Name | Contains sigma-54 | RefSeq ID |
|---|---|---|---|
| Actinobacteria | *Bifidobacterium longum NCC2705*, complete genome | N | NC_004307.2 |
| Actinobacteria | *Mycobacterium tuberculosis CDC1551*, complete genome | N | NC_002755.2 |
| Actinobacteria | *Mycobacterium ulcerans Agy99*, complete genome | N | NC_008611.1 |
| Actinobacteria | *Nocardia farcinica IFM 10152*, complete genome | N | NC_006361.1 |
| Actinobacteria | *Rubrobacter xylanophilus DSM 9941*, complete genome | N | NC_008148.1 |
| Actinobacteria | *Saccharopolyspora erythraea NRRL 2338*, complete genome | N | NC_009142.1 |
| Actinobacteria | *Salinispora tropica CNB-440*, complete genome | N | NC_009380.1 |
| Actinobacteria | *Streptomyces avermitilis MA-4680*, complete genome | N | NC_003155.4 |
| Actinobacteria | *Streptomyces coelicolor A3(2)*, complete genome | N | NC_003888.3 |
| Actinobacteria | *Streptomyces griseus subsp. griseus NBRC 13350*, complete genome | N | NC_010572.1 |
| Aquificae | *Sulfurihydrogenibium sp. YO3AOP1*, complete genome | Y | NC_010730.1 |
| Bacteroidetes/Chlorobi | *Bacteroides fragilis NCTC 9343*, complete genome | Y | NC_003228.3 |
| Bacteroidetes/Chlorobi | *Candidatus Amoebophilus asiaticus 5a2*, complete genome | Y | NC_010830.1 |
| Bacteroidetes/Chlorobi | *Gramella forsetii KT0803*, complete genome | Y | NC_008571.1 |
| Chlamydiae/ Verrucomicrobia | *Chlamydia trachomatis D/UW-3/CX*, complete genome | Y | NC_000117.1 |
| Chlamydiae/ Verrucomicrobia | *Chlamydophila pneumoniae AR39*, complete genome | Y | NC_002179.2 |
| Chlamydiae/ Verrucomicrobia | *Coraliomargarita akajimensis DSM 45221* chromosome, complete genome | N | NC_014008.1 |
| Chloroflexi | *Sphaerobacter thermophilus DSM 20745* chromosome 1, complete genome | Y | NC_013523.1 |
| | *Sphaerobacter thermophilus DSM 20745* chromosome 2, complete genome | | NC_013524.1 |
| Crenarchaeota | *Pyrobaculum arsenaticum DSM 13514*, complete genome | N | NC_009376.1 |
| Cyanobacteria | *Gloeobacter violaceus PCC 7421*, complete genome | N | NC_005125.1 |
| Cyanobacteria | *Microcystis aeruginosa NIES-843*, complete genome | N | NC_010296.1 |
| Cyanobacteria | *Nostoc punctiforme PCC 73102*, complete genome | N | NC_010628.1 |

| | | | |
|---|---|---|---|
| Cyanobacteria | *Synechocystis sp. PCC 6803*, complete genome | N | NC_000911.1 |
| Deinococcus-Thermus | *Deinococcus deserti VCD115*, complete genome | N | NC_012526.1 |
| Euryarchaeota | *Pyrococcus horikoshii OT3*, complete genome | N | NC_000961.1 |
| Fibrobacteres/ Acidobacteria | *Solibacter usitatus Ellin6076*, complete genome | Y | NC_008536.1 |
| Firmicutes | *Bacillus cereus subsp. cytotoxis NVH 391-98*, complete genome | Y | NC_009674.1 |
| Firmicutes | *Bacillus subtilis subsp. subtilis str. 168*, complete genome | Y | NC_000964.3 |
| Firmicutes | *Clostridium acetobutylicum ATCC 824*, complete genome | Y | NC_003030.1 |
| Firmicutes | *Clostridium kluyveri DSM 555*, complete genome | Y | NC_009706.1 |
| Firmicutes | *Clostridium perfringens str. 13*, complete genome | Y | NC_003366.1 |
| Firmicutes | *Lactobacillus brevis ATCC 367*, complete genome | N | NC_008497.1 |
| Firmicutes | *Listeria monocytogenes str. 4b F2365*, complete genome | Y | NC_002973.6 |
| Firmicutes | *Staphylococcus aureus subsp. aureus MRSA252*, complete genome | N | NC_002952.2 |
| Firmicutes | *Streptococcus pyogenes M1 GAS*, complete genome | N | NC_002737.1 |
| Proteobacteria (Alpha) | *Agrobacterium tumefaciens str. C58* chromosome circular, complete sequence | Y | NC_003062.2 |
| | *Agrobacterium tumefaciens str. C58* chromosome linear, complete sequence | | NC_003063.2 |
| Proteobacteria (Alpha) | *Methylobacterium extorquens DM4*, complete genome | Y | NC_012988.1 |
| Proteobacteria (Alpha) | *Rickettsia rickettsii str. Iowa*, complete genome | N | NC_010263.1 |
| Proteobacteria (Beta) | *Neisseria meningitidis MC58*, complete genome | Y | NC_003112.2 |
| Proteobacteria (Beta) | *Ralstonia solanacearum GMI1000* plasmid pGMI1000MP, complete sequence | Y | NC_003296.1 |
| | *Ralstonia solanacearum GMI1000*, complete genome | | NC_003295.1 |
| Proteobacteria (Delta) | *Geobacter sulfurreducens PCA*, complete genome | Y | NC_002939.4 |
| Proteobacteria (Delta) | *Myxococcus xanthus DK 1622*, complete genome | Y | NC_008095.1 |
| Proteobacteria (Delta) | *Sorangium cellulosum 'So ce 56'*, complete genome | Y | NC_010162.1 |
| Proteobacteria (Epsilon) | *Campylobacter jejuni subsp. jejuni NCTC 11168*, complete genome | Y | NC_002163.1 |
| Proteobacteria (Epsilon) | *Helicobacter pylori J99*, complete genome | Y | NC_000921.1 |
| Proteobacteria (Gamma) | *Candidatus Carsonella ruddii PV*, complete genome | N | NC_008512.1 |
| Proteobacteria (Gamma) | *Escherichia coli str. K-12 substr. MG1655*, complete genome | Y | NC_000913.2 |
| Proteobacteria (Gamma) | *Francisella tularensis subsp. tularensis SCHU S4*, complete genome | N | NC_006570.2 |

| | | | |
|---|---|---|---|
| Proteobacteria (Gamma) | *Legionella pneumophila subsp. pneumophila str. Philadelphia 1*, complete genome | Y | NC_002942.5 |
| Proteobacteria (Gamma) | *Pseudomonas aeruginosa PAO1*, complete genome | Y | NC_002516.2 |
| Proteobacteria (Gamma) | *Pseudomonas putida KT2440*, complete genome | Y | NC_002947.3 |
| Proteobacteria (Gamma) | *Pseudomonas syringae pv. tomato str. DC3000*, complete genome | Y | NC_004578.1 |
| Proteobacteria (Gamma) | *Salmonella enterica subsp. enterica serovar Typhi str. CT18*, complete genome | Y | NC_003198.1 |
| Proteobacteria (Gamma) | *Shewanella oneidensis MR-1*, complete genome | Y | NC_004347.1 |
| Proteobacteria (Gamma) | *Yersinia pestis CO92*, complete genome | Y | NC_003143.1 |
| Spirochaetes | *Borrelia burgdorferi B31*, complete genome | Y | NC_001318.1 |
| Spirochaetes | *Treponema denticola ATCC 35405*, complete genome | Y | NC_002967.9 |
| Tenericutes | *Candidatus Phytoplasma mali*, complete genome | N | NC_011047.1 |

## Appendix 2: RpoN Matrix

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A: | 12 | 2 | 0 | 12 | 139 | 11 | 55 | 51 | 46 | 44 | 38 | 13 | 4 | 1 | 9 | 76 |
| C: | 14 | 0 | 0 | 147 | 23 | 122 | 17 | 48 | 64 | 42 | 62 | 22 | 18 | 2 | 173 | 5 |
| G: | 10 | 184 | 186 | 6 | 18 | 10 | 103 | 69 | 36 | 35 | 43 | 15 | 10 | 181 | 1 | 17 |
| T: | 150 | 0 | 0 | 21 | 6 | 43 | 11 | 18 | 40 | 65 | 43 | 136 | 154 | 2 | 3 | 88 |

**Figure S - 1. RpoN matrix derived from 186 sigma-54 promoter sites** (56, 103, 91)**.**

# Appendix 3: Charts of GC Content vs Adjusted Hits

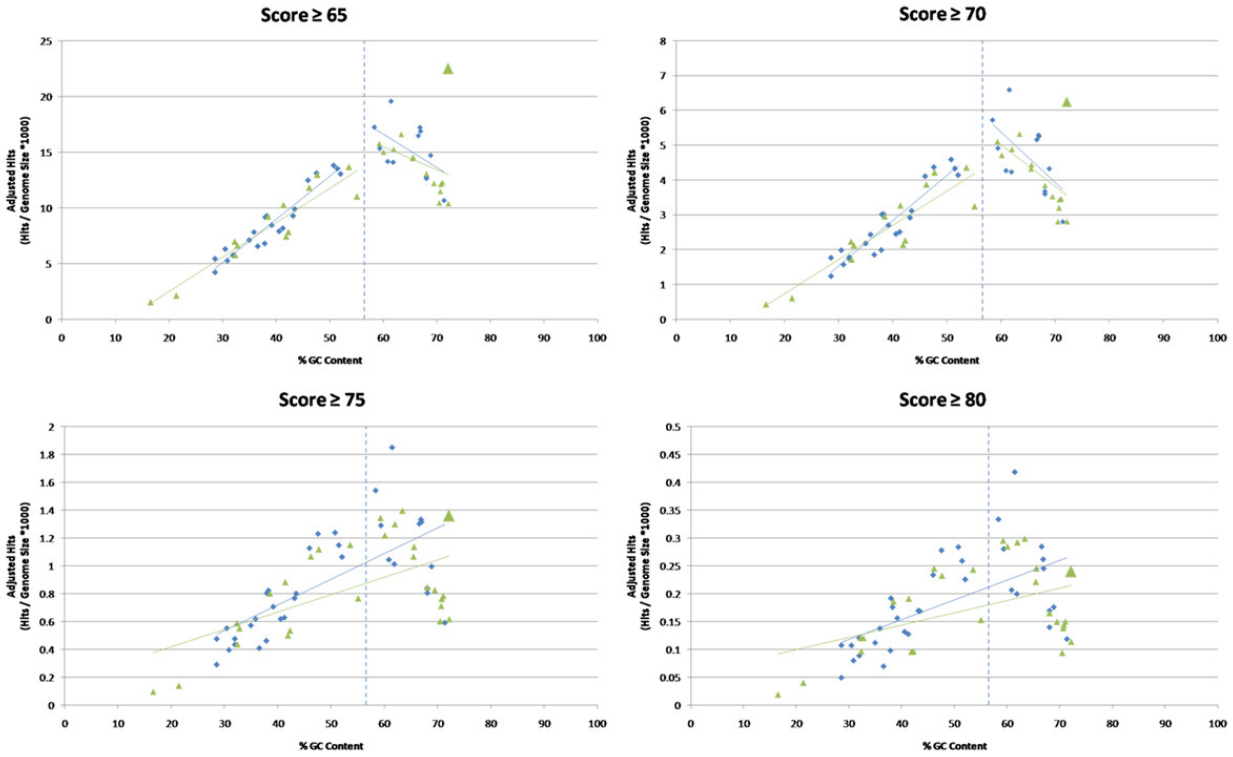



**Figure S - 2. Relationship between GC content and adjusted hits.**
Adjusted hits were calculated by taking the total number of hits at a given score for a given organism, dividing by the organism's genome size, then multiplying by 10 000. Green triangles represent organisms that do not have sigma-54, while blue diamonds indicate organisms that do. The dotted blue line at 56.6% indicates the average GC content of the RpoN matrix. The large triangular data point at %GC of 72.1 is that of *Streptomyces coelicolor*.

# Appendix 4: Putative Sigma-54 Promoters in the Oxytetracycline Gene Cluster

**Table S - 2. List of putative sigma-54 promoters identified from a bioinformatics analysis of the oxytetracycline gene cluster.**

| Oxytetracycline gene cluster from *Streptomyces rimosus DQ143963* | | | | | | |
|---|---|---|---|---|---|---|
| Score | Distance | N | Strand | Sequence | Gene | Protein Name |
| 82 | 285 | 485 | + | CTGGAACGGCGGTGGCT | oxyA | Ketosynthase |
| 75 | 50 | 9883 | - | GCGGCCCGAGTAGTGCA | oxyH | Acyl-CoA ligase |
| 80 | 104 | 13973 | + | CAGGCACTGCACCTGCT | oxyM | Ketoreductase |
| 82 | 68 | 19404 | - | GTGGCTCGAAGTGTGCT | oxyR | PNP-oxidase |

# Appendix 5: PKS/NRPS Gene Clusters with Predicted Sigma-54 Promoters by Phylum

**Table S - 3. Tabulation of PKS/NRPS gene clusters with predicted sigma-54 promoters at varying cut-off stringencies.**

| | Number of strains | PKS/NRPS gene clusters | Gene clusters with $\sigma^{54}$ promoters at cutoff = 75 | Gene clusters with $\sigma^{54}$ promoters at cutoff = 78 | Gene clusters with $\sigma^{54}$ promoters at cutoff = 79 | Gene clusters with $\sigma^{54}$ promoters at cutoff = 80 |
|---|---|---|---|---|---|---|
| Actinobacteria | 10 | 94 | 61 | 47 | 37 | 32 |
| Firmicutes | 9 | 15 | 11 | 8 | 7 | 7 |
| Proteobacteria | 20 | 49 | 36 | 26 | 26 | 21 |
| Other | 19 | 22 | 16 | 14 | 13 | 10 |
| Total | 58 | 180 | 124 | 95 | 83 | 70 |

# Appendix 6: PKS/NRPS Gene Clusters with Associated Sigma-54 Promoters

**Table S3**. List of PKS, NRPS and NRPS/PKS gene clusters with the associated $\sigma^{54}$ promoters identified from our bioinformatics analysis.

**NC_002755.2 Mycobacterium tuberculosis CDC1551, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS | | 105314-123821 | | | | | |
| | Score | Distance From ORF | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 430 | 106294 | + | TTGGCAGCGCTTCGGCT | - | MT0106 | dioxygenase, putative |
| | 75 | 157 | 106567 | + | CTGGCAGTCAAGGTGCC | - | MT0106 | dioxygenase, putative |
| | 81 | 323 | 109450 | + | GTGGCGTGAACATTGCG | - | MT0109 | hypothetical protein |
| | 81 | 325 | 122488 | - | GTGGAATAGCACTTGCC | - | MT0112 | cation transporter E1-E2 family ATPase |
| | 81 | 60 | 122247 | + | ATGGCCCGCTGGATGCA | - | MT0113 | hypothetical protein |
| **2** | PKS | | 485426-491388 | | | | | |
| **3** | PKS | | 1313164-1328792 | | | | | |
| | Score | Distance From ORF | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 341 | 1327340 | - | CTGGCTCAGACCCTGCG | - | MT1222 | acyl-CoA synthetase |
| **4** | PKS | | 1865066-1886258 | | | | | |
| | Score | Distance From ORF | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 121 | 1866089 | + | CTGGCACTGATGCTGGA | - | MT1701 | polyketide synthase |
| | 76 | 370 | 1878562 | + | TTGCCACCGACCTTGAT | - | MT1704 | polyketide synthase |
| **5** | PKS | | 2294295-2309319 | | | | | |
| **6** | NRPS/ PKS | | 2647576-2673168 | | | | | |
| | Score | Distance From ORF | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 344 | 2649233 | - | CCGGCCCCGCCGTTGCC | - | MT2439 | hypothetical protein |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 83 | 331 | 2651649 | - | GAGGAACGCCTTTTGCT | dnaJ-2 | MT2442 | chaperone protein DnaJ |
| | 78 | 334 | 2673502 | - | CTGGCAGCCGTTTGGCA | pchE | MT2451 | dihydroaeruginoic acid synthetase |
| **7** | PKS | | 3236513-3319822 | | | | | |
| | Score | Distance From ORF | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 385 | 3236128 | + | TCGGCACGCCTCGGGCT | - | MT2998 | thioesterase |
| | 75 | 1 | 3237290 | + | TTCGCTGGACATTTGCT | - | MT2998.1 | hypothetical protein |
| | 77 | 349 | 3256220 | + | CTGGGACTCGACTCGCT | - | MT3004 | polyketide synthase |
| | 78 | 23 | 3292184 | - | GTGGCCGACCGTTTGCG | - | MT3021.1 | polyketide synthase |
| | 77 | 323 | 3296988 | - | ATCGCACGGCGCTGGCG | - | MT3023 | acyl-CoA synthetase |
| | 79 | 490 | 3298272 | + | TTGGCATCGATCTTGGG | - | MT3026 | methyltransferase, putative |
| | 78 | 477 | 3298285 | + | TGGGCACGAATTTCGCC | - | MT3026 | methyltransferase, putative |
| | 77 | 104 | 3298658 | + | CTGCCATACTCCTTGCC | - | MT3026 | methyltransferase, putative |
| | 82 | 353 | 3303494 | + | GTGGCATGCTCATTCCT | - | MT3031 | glycosyl transferase |
| | 76 | 44 | 3303803 | + | GTGGCTGCGGTATTGCC | - | MT3031 | glycosyl transferase |
| | 79 | 276 | 3306596 | - | CTGGCAAGATCTTCGCC | - | MT3034 | UDP-glucoronosyl and UDP-glucosyltransferase |
| | 75 | 379 | 3308002 | - | TTGGCCAGCTAGTTACT | - | MT3036 | hypothetical protein |
| | 76 | 106 | 3308155 | - | GTGGCAGGGGCCGTGAT | - | MT3037 | hypothetical protein |
| | 89 | 439 | 3311348 | + | CTGGCACGCTACATGCA | - | MT3041.1 | hypothetical protein |

**NC_008611.1 Mycobacterium ulcerans Agy99, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | PKS | | 1803923-1835953 | | | | | |
| | Score | Distance From ORF | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 77 | 280 | 1820828 | + | CTGGACAAGTTATTGCT | pks11 | MUL_1656 | chalcone synthase, Pks11 |
| | 75 | 91 | 1821017 | + | ATGTCACATGTGTTGAA | pks11 | MUL_1656 | chalcone synthase, Pks11 |
| | 80 | 234 | 1834666 | - | ATGGTGCCGCAGTTGCT | - | MUL_1664 | hypothetical protein |
| | 75 | 443 | 1834875 | - | ACGGCACAGCATTTTCG | - | MUL_1664 | hypothetical protein |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 75 | 322 | 1835436 | - | CTGGATCCCGTTTTGCG | - | MUL_1665 | antibiotic resistance ABC transporter, efflux protein |
| **2** | PKS | | | 2211634-2271101 | | | | |
| | Score | Distance From ORF | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 133 | 2213368 | + | CGGGCACACCGAGTGCG | - | MUL_2003 | chorismate pyruvate-lyase |
| | 78 | 133 | 2216250 | + | GTGGCCGAACGGTTGCG | pks15/1 | MUL_2005 | polyketide synthase Pks15/1 |
| | 79 | 305 | 2228469 | + | GTGGCTCAACGGTGGCT | - | MUL_2009 | methyltransferase |
| | 75 | 67 | 2230104 | + | GTGTCACACCGGTTGAA | mas | MUL_2010 | multifunctional mycocerosic acid synthase |
| | 78 | 354 | 2240002 | - | ACGGCAAGATCATTGCC | drrB | MUL_2013 | daunorubicin-DIM-transport integral |
| | 76 | 153 | 2240793 | - | CTGGCGCGGTCGTTCCC | drrA | MUL_2014 | daunorubicin-DIM-transport ATP-binding |
| | 77 | 288 | 2240928 | - | TTGGCAATGCCGGTGCG | drrA | MUL_2014 | daunorubicin-DIM-transport ATP-binding |
| | 84 | 178 | 2245274 | - | GTGACACTGATTTTGCA | ppsE | MUL_2015 | phenolpthiocerol synthesis type-I polyketide |
| | 80 | 247 | 2250769 | - | ATGGCCCTCGAGTTGCG | ppsD | MUL_2016 | phenolpthiocerol synthesis type-I polyketide |
| | 82 | 34 | 2257161 | - | TTGGAACAGCTTTCGCA | ppsC | MUL_2017 | phenolpthiocerol synthesis type-I polyketide |
| | 76 | 121 | 2266630 | - | GTGGCCGACGTCGTGCT | ppsA | MUL_2019 | phenolpthiocerol synthesis type-I polyketide |
| **3** | NRPS | | 2921595-2959529 | | | | | |
| | Score | Distance From ORF | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 79 | 76 | 2921519 | + | TCGGCACGATAATTCCT | - | MUL_2620 | hypothetical protein |
| | 80 | 216 | 2929670 | - | ATGTCGCGGGCATTGCT | - | MUL_2625 | hypothetical protein |
| | 81 | 92 | 2930909 | - | CTGGCGCAGACCGTGCA | fadE25 | MUL_2626 | acyl-CoA dehydrogenase FadE25 |
| | 75 | 87 | 2934800 | - | GTCGCGCGGGGATTGCC | birA | MUL_2631 | bifunctional protein BirA |
| | 76 | 286 | 2952542 | - | GTGGCAGACCTGCGGCT | sigF | MUL_2640 | RNA polymerase sigma factor SigF |
| | 83 | 205 | 2953689 | - | GTGGCACGGCGCCGGCT | usfY | MUL_2642 | hypothetical protein |
| **4** | NRPS/ PKS | | 4029686-4059363 | | | | | |
| | Score | Distance From ORF | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 385 | 4033562 | - | TTGGACTACGGCTTGCT | mbtB | MUL_3632 | phenyloxazoline synthase MbtB |
| | 75 | 397 | 4038148 | - | CTGGCCCCACCGCTGCC | mbtF | MUL_3633 | non-ribosomal peptide synthetase MbtF |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 75 | 61 | 4042974 | + | GTTGAATGGCTGTTGCG | - | MUL_3635 | hypothetical protein |
| | 76 | 245 | 4044260 | + | CTGGCACTTCGCGTGAT | mbtC | MUL_3637 | polyketide synthase MbtC |
| | 76 | 56 | 4050456 | - | ATGGCAACGCTATTACC | mbtG | MUL_3640 | lysine-N-oxygenase MbtG |
| | 80 | 31 | 4053325 | + | ATCGCAGGCTCATTGCA | - | MUL_3643 | short-chain membrane-associated dehydrogenase |
| | 75 | 317 | 4057245 | - | CTGGCTCGGCAGCGGCA | desA1_1 | MUL_3646 | acyl-[acyl-carrier protein] desaturase DesA1_1 |
| **5** | NRPS | | 4833556-4838804 | | | | | |
| **6** | NRPS | | 5377263-5405727 | | | | | |
| | Score | Distance From ORF | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 77 | 255 | 5378620 | + | ACGGCACCAGCACTGCT | - | MUL_4854 | oxidoreductase |
| | 75 | 467 | 5379274 | + | CTGGCCAAGGTTTGGCA | - | MUL_4855 | hypothetical protein |
| | 84 | 398 | 5391216 | - | TCGGCCCGGATATTGCT | - | MUL_4861 | hypothetical protein |
| | 83 | 199 | 5396076 | - | ATGGCAATGACGTTGCG | - | MUL_4867 | hypothetical protein |
| **7** | PKS | | 5523988-5532849 | | | | | |

**NC_006361.1 Nocardia farcinica IFM 10152, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | PKS | | 184410-208907 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 160 | 207612 | + | GTGGCGCCGTCGGTGCT | - | nfa1980 | hypothetical protein |
| **2** | NRPS | | 742402-783087 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 5 | 782060 | + | GTGGCACGCGTGCCGCG | - | nfa7210 | putative esterase |
| **3** | NRPS/ PKS | | 809229-843740 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 426 | 815841 | - | GTGGGCCAACAACTGCT | - | nfa7550 | putative ATP-dependent protease |
| | 78 | 324 | 816789 | - | CTGGCCTGGGCGGTGCT | amiE | nfa7560 | acylamide amidohydrolase |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 78 | 239 | 820253 | - | TTGGCGCAACACGTGCG | - | nfa7600 | hypothetical protein |
| | 80 | 307 | 820321 | - | CTGGCGAGCCAGTTGCC | - | nfa7600 | hypothetical protein |
| | 76 | 231 | 827277 | + | CTGGCCGAGCGGATGCT | nbtD | nfa7660 | putative non-ribosomal peptide synthetase |
| | 76 | 64 | 837122 | + | CTGGCCGAGCTGTGGCA | nbtF | nfa7680 | putative non-ribosomal peptide synthetase |
| **4** | NRPS | | 1219367-1235955 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 77 | 475 | 1218892 | + | CTGGCCCGTTGCCTGCG | - | nfa11040 | hypothetical protein |
| **5** | NRPS | | 2965856-2985159 | | | | | |
| **6** | PKS | | 3189119-3223839 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 247 | 3191679 | - | CTGGCGCTGGTGCTGCG | - | nfa30090 | putative iron-sulfur oxidoreductase |
| | 75 | 103 | 3193577 | - | CTAGCGCGCGGATTGCG | - | nfa30110 | putative monooxygenase |
| | 75 | 380 | 3193854 | - | GTGGCATCCTGGTGGCC | - | nfa30110 | putative monooxygenase |
| | 76 | 406 | 3195498 | - | CTGGGGCAGTGATTGCG | - | nfa30120 | acyl-CoA synthetase |
| | 75 | 62 | 3198518 | + | CCGGCAAGCGCGATGCT | - | nfa30170 | hypothetical protein |
| | 76 | 56 | 3202150 | + | ATTGCTGGAATATTGCT | - | nfa30220 | putative sigma factor |
| | 75 | 459 | 3212616 | - | CTGGCGGACCAACTGCT | - | nfa30250 | putative polyketide synthase |
| | 76 | 221 | 3220860 | - | CTGGACTTCCTGTTGCT | - | nfa30320 | hypothetical protein |
| **7** | NRPS | | 3297940-3312573 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 179 | 3299108 | - | GTGGCCGGGACCGTGCA | - | nfa31160 | putative ABC transporter ATP-binding protein |
| | 77 | 296 | 3299225 | - | CTGGCGCGGGACTTCCG | - | nfa31160 | putative ABC transporter ATP-binding protein |
| **8** | PKS | | 4478704-4491458 | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 338 | 4489377 | - | ATGGCACAGGTGGCGCC | - | nfa43240 | putative polyketide synthase |
| | 77 | 493 | 4491951 | - | CCGGCAACGCACTTGCC | - | nfa43260 | putative transporter |
| **9** | NRPS | | 5244714-5291138 | | | | | |
| **10** | NRPS | | 5318160-5361768 | | | | | |
| **11** | PKS | | 5919882-5934805 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 80 | 97 | 5923205 | + | GCGGCACGTGAGCTGCT | - | nfa55870 | hypothetical protein |
| | 79 | 153 | 5926963 | - | GTGGCACGACCGTGCCA | - | nfa55900 | long-chain-fatty-acid--CoA ligase |
| | 78 | 189 | 5926962 | + | ATGGCACGGTCGTGCCA | - | nfa55910 | putative dioxygenase |

**NC_009142.1 Saccharopolyspora erythraea NRRL 2338, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | PKS | | 14813-51921 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 81 | 175 | 15677 | - | TTGGCCGGGCCGTTGCG | - | SACE_0012 | TetR family transcriptional regulator |
| | 76 | 129 | 49054 | + | GCGGCCCGGCTCTTTCT | - | SACE_0029 | RNA polymerase sigma factor |
| **2** | PKS | | 778214-832825 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 328 | 783452 | - | CTGGCGCCTACGGTGCA | eryCIV | SACE_0716 | eryCIV NDP-6-deoxyhexose 3,4- |
| | 76 | 128 | 784712 | - | CTGGCACAGGTGATCCG | eryBVI | SACE_0717 | NDP-4-keto-6-deoxy-glucose 2,3-dehydratase |
| | 77 | 362 | 787113 | - | GTGGCACGACGGCGGCG | eryBV | SACE_0719 | 6-DEB TDP-mycarosyl glycosyltransferase |
| | 81 | 224 | 831762 | - | GAGGTACGCGGCTTGCA | ermE | SACE_0733 | N-6-aminoadenine-N-methyltransferase,erythro |
| **3** | NRPS | | 1436967-1445722 | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **4** | PKS | | 2528343-2549947 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 242 | 2528101 | + | GTGGACCGGCAGTTGGT | cmtC | SACE_2339 | trehalose corynomycolyl transferase C |
| **5** | PKS | | 2795362-2818679 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 132 | 2798967 | - | TCGGCACGCGACTTGGG | - | SACE_2592 | glucose-methanol-choline oxidoreductase |
| | 80 | 116 | 2800190 | - | CTGGCGCACCGGGTGCT | - | SACE_2593 | glycopeptide antibiotics resistance protein |
| **6** | NRPS/ PKS | | 2838066-2886848 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 81 | 433 | 2844648 | - | GTGGCGCAGGTGCTGCA | - | SACE_2619 | non-ribosomal peptide synthetase |
| | 79 | 322 | 2855647 | - | GAGGCACTGCGCCTGCT | - | SACE_2627 | isopentenyl-diphosphate delta-isomerase II 2 |
| | 77 | 316 | 2883545 | - | GTGGCGCTGGCGCTGCA | - | SACE_2635 | major facilitator superfamily transporter |
| **7** | NRPS | | 2944491-2974637 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 112 | 2962692 | + | GTGGAACGGCCGCGGCA | - | SACE_2698 | putative regulatory protein |
| | 76 | 63 | 2974297 | + | GTGGCACGCGCGGCGCC | - | SACE_2709 | hypothetical protein |
| **8** | PKS | | 3127652-3157229 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 284 | 3131486 | - | TCGGCACGGTGTTCGCC | rifI | SACE_2867 | aminoquinate/shkimate dehydrogenase |
| | 79 | 195 | 3138945 | + | GTGCCATTCCGGTTGCT | - | SACE_2875 | modular polyketide synthase |
| | 77 | 158 | 3138982 | + | CAGGCATCGAGGTTGCG | - | SACE_2875 | modular polyketide synthase |
| | 78 | 428 | 3151884 | + | CTGGCGCACGTGTTCCT | mhpA | SACE_2877 | 3-(3-hydroxyphenyl)propionate |
| | 77 | 240 | 3152072 | + | CTGGCACGGCCTCTACC | mhpA | SACE_2877 | 3-(3-hydroxyphenyl)propionate |
| **9** | NRPS | | 3301269-3308199 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 77 | 55 | 3307140 | + | ACGGCGCGGATCGTGCT | - | SACE_3016 | SyrP-like protein |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **10** | NRPS | | 3326161-3344964 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 85 | 242 | 3325919 | + | GTGGCCAGGCGGTTGCT | - | SACE_3033 | putative peptide monooxygenase |
| **11** | PKS | | 4545937-4616845 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 59 | 4554457 | + | TTGACTAGAGTGTTGCT | - | SACE_4130 | hypothetical protein |
| | 78 | 290 | 4559173 | + | ATGGCCTCGCGGTTGCC | - | SACE_4132 | acyl carrier protein |
| | 86 | 49 | 4563179 | + | TGGGCACCGCCCTTGCT | - | SACE_4135 | MaoC-like dehydratase |
| | 80 | 259 | 4572145 | - | CTGGCCGGGCACCTGCT | - | SACE_4138 | type I PKS modular polyketide synthas |
| | 78 | 417 | 4588385 | - | GTGGAACAGGCCTTCCA | - | SACE_4139 | type I modular polyketide synthase |
| | 75 | 146 | 4612861 | - | GTGGCGCTGCGCGTGCC | - | SACE_4142 | cytochrome P450 monooxygenase |
| | 75 | 251 | 4612966 | - | CTGGCTCGGATGGAGCT | - | SACE_4142 | cytochrome P450 monooxygenase |
| | 83 | 27 | 4614155 | - | ATGGCAGGGAAGTTCCT | - | SACE_4143 | cytochrome P450 monooxygenase |
| **12** | NRPS | | 4759004-4809694 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 475 | 4772132 | - | CTGGCCCGGGCGCTGCG | - | SACE_4285 | AMP-dependent synthetase and ligase |
| | 76 | 498 | 4772155 | - | GTGGCGCGCTGCTGGCC | - | SACE_4285 | AMP-dependent synthetase and ligase |
| **13** | PKS | | 4821866-4828194 | | | | | |
| **14** | PKS | | 4992823-5006652 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 183 | 4996606 | - | GTGGACCGCCCGGTGCA | - | SACE_4472 | NAD-dependent epimerase/dehydratase |
| | 79 | 82 | 4997888 | - | GTGGCGCTGAAGCTGCT | qor | SACE_4474 | quinone oxidoreductase |
| **15** | PKS | | 5102367-5117363 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 79 | 327 | 5106524 | + | CTCGCACGCGGCCTGCA | - | SACE_4566 | hypothetical protein |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 78 | 359 | 5112790 | + | GTGCCGCAACGTTTGCT | pteA1 | SACE_4574 | modular polyketide synthase |
| **16** | PKS | | 5932907-5941187 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 213 | 5933803 | + | CGGCCACGCCGGTTGCA | chlB1 | SACE_5308 | iterative type I polyketide synthase |
| **17** | PKS | | | 6225358-6248235 | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 287 | 6239374 | - | TCGGCACCGCTCTTCCA | - | SACE_5540 | 3-oxoacyl-(acyl carrier protein) synthase III |

**NC_009380.1 Salinispora tropica CNB-440, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | PKS | | 671414-679577 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 214 | 671873 | + | TTGGAACGAAACGAGCT | - | Strop_0597 | hypothetical protein |
| | 83 | 374 | 672939 | + | ATGGCTGGCACGTTGCT | - | Strop_0598 | beta-ketoacyl synthase |
| **2** | NRPS/ PKS | | 144505-1172682 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 79 | 296 | 1150310 | - | CTGGCATGCCTCCGGCA | - | Strop_1020 | MbtH domain-containing protein |
| | 82 | 296 | 1164400 | + | GTGGCATGGCCTATGCC | - | Strop_1028 | hypothetical protein |
| | 76 | 157 | 1164539 | + | CTGGCCGGCCGATCGCT | - | Strop_1028 | hypothetical protein |
| | 78 | 220 | 1165260 | + | CTGCCACGGCTGCTGCC | - | Strop_1029 | cyclase family protein |
| | 81 | 438 | 1167460 | - | CTGGCACAGATAGCGCT | - | Strop_1030 | regulatory protein, LuxR |
| **3** | PKS | | 2502319-2520690 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 33 | 2505275 | - | TGGGCACAAAGCCTGCG | - | Strop_2212 | TAP domain-containing protein |
| | 75 | 447 | 2507023 | - | GCGGCACCCTGGTTGAT | - | Strop_2213 | hypothetical protein |
| | 78 | 480 | 2513894 | - | ATGGGACTGACCTCGCT | - | Strop_2220 | hypothetical protein |
| | 83 | 411 | 2514406 | + | CTGGTAGAGCAGTTGCA | - | Strop_2222 | dTDP-glucose 4,6-dehydratase |
| | 75 | 1 | 2517088 | + | TGGGCACGGGTCCGGCA | - | Strop_2224 | beta-ketoacyl synthase |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 76 | 463 | 2519253 | + | GCGGCCGGCCCGTTGCT | - | Strop_2227 | hypothetical protein |
| **4** | PKS | | 2795012-2816666 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 433 | 2798754 | - | CTGGTCGAACTCTTGCG | - | Strop_2487 | dehydrogenase, E1 component |
| | 76 | 124 | 2804213 | - | GAGGCCGACCTGTTGCT | - | Strop_2492 | AMP-dependent synthetase and ligase |
| | 81 | 84 | 2805809 | - | CTGGCCTGGCTTGTGCA | - | Strop_2494 | beta-ketoacyl synthase |
| | 80 | 205 | 2809136 | - | GTGGCGACCGTGTTGCT | - | Strop_2498 | antibiotic biosynthesis monooxygenase |
| | 77 | 357 | 2812320 | - | GTGGCCGAGATCCTGCA | - | Strop_2501 | cupin 2 domain-containing protein |
| | 77 | 112 | 2814752 | - | CTGGCGCACCTGGTGCC | - | Strop_2504 | acetyl-CoA carboxylase, biotin carboxyl carrier |
| **5** | NRPS/ PKS | | 2956815-2996840 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 393 | 2974666 | - | TTGGCGCAGCTCTGGCC | - | Strop_2648 | methyltransferase type 12 |
| | 80 | 85 | 2977211 | - | CTGGCACCCGCCGTGCC | - | Strop_2650 | ABC transporter related |
| | 75 | 466 | 2979339 | - | CTGGCGTTGACCCTGCA | - | Strop_2652 | binding-protein-dependent transport |
| | 77 | 150 | 2996990 | - | TTGGCCAAGCAAGTGCT | - | Strop_2659 | hypothetical protein |
| **6** | PKS | | 3020222-3042851 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 204 | 3023263 | - | GTGGCCTGGTTCTGGCC | - | Strop_2686 | alpha amylase, catalytic region |
| | 75 | 322 | 3025145 | - | TTGGCATGTGGTGCGCT | - | Strop_2688 | helix-turn-helix domain-containing protein |
| | 75 | 330 | 3026030 | - | GCGGCACCGGCGTTGTT | - | Strop_2689 | hypothetical protein |
| | 76 | 207 | 3027623 | - | ACGGCCCGTCTCGTGCT | - | Strop_2691 | hypothetical protein |
| | 81 | 377 | 3031869 | - | TTGGCAGGGCCGGTGCC | - | Strop_2695 | flavin reductase domain-containing protein |
| | 77 | 192 | 3037925 | - | CTGGCGCTGGCCTCGCA | - | Strop_2697 | beta-ketoacyl synthase |
| | 76 | 235 | 3043086 | - | ATGGCACGGGTGTTCTC | - | Strop_2701 | hypothetical protein |
| **7** | NRPS/ PKS | | 3112454-3249126 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 75 | 422 | 3116359 | + | ACGGCATGAGCGTCGCA | - | Strop_2766 | regulatory protein, TetR |
| | 75 | 117 | 3141159 | - | CTGGTCCGGCTGGTGCC | - | Strop_2770 | cytochrome P450 |
| | 80 | 282 | 3142490 | + | CTGGCCCGGACCGTGCG | - | Strop_2772 | aminotransferase, class I and II |
| | 76 | 440 | 3143585 | + | CTGGCCGGCCCGATGCA | - | Strop_2773 | hypothetical protein |
| | 76 | 344 | 3143681 | + | GTGGCCGGCGCCGTGCA | - | Strop_2773 | hypothetical protein |
| | 77 | 410 | 3144589 | + | TTGGCGCAGCGGCTGCG | - | Strop_2774 | AMP-dependent synthetase and ligase |
| | 79 | 25 | 3160851 | - | CTGGCAACGAGTTTGGT | - | Strop_2778 | beta-ketoacyl synthase |
| | 85 | 327 | 3165966 | - | CTGGCCCGGCACCTGCT | - | Strop_2779 | acyl transferase domain-containing protein |
| | 76 | 471 | 3188122 | - | GCGGCACGCCGATCGCC | - | Strop_2785 | hypothetical protein |
| | 77 | 312 | 3190952 | - | AGGGCAAGCGGATTGCC | - | Strop_2787 | AMP-dependent synthetase and ligase |
| | 75 | 112 | 3193950 | - | CTGGCCGCGACCCTGCT | - | Strop_2791 | alpha/beta hydrolase fold |
| | 75 | 303 | 3194141 | - | CTGGCCGACCCGCTGCT | - | Strop_2791 | alpha/beta hydrolase fold |
| | 78 | 427 | 3198676 | - | TTGGCACTCTCATGGCC | - | Strop_2795 | beta-ketoacyl synthase |
| | 78 | 168 | 3198676 | - | TTGGCACTCTCATGGCC | - | Strop_2796 | phosphopantetheine-binding |
| | 79 | 373 | 3198881 | - | GCGGCACTGCTCTGGCT | - | Strop_2796 | phosphopantetheine-binding |
| | 75 | 470 | 3198978 | - | CGGGAACGAGCCCTGCT | - | Strop_2796 | phosphopantetheine-binding |
| | 75 | 370 | 3201972 | - | CAGGCACGGCTGCGGCA | - | Strop_2800 | L-carnitine dehydratase/bile acid- |
| | 76 | 7 | 3203132 | + | CTGGCACTCTGGAGGCT | - | Strop_2803 | AMP-dependent synthetase and ligase |
| | 78 | 466 | 3212075 | + | CTGGCGCTGGAGGTGCT | - | Strop_2809 | hypothetical protein |
| | 78 | 205 | 3222961 | - | CGGGCACTGCGGGTGCT | - | Strop_2817 | thioesterase |
| | 75 | 410 | 3224293 | - | GTGGCTCGACTGGTACT | - | Strop_2818 | thiazolinyl imide reductase |
| | 78 | 126 | 3241571 | - | GCGCCACAGGACTTGCT | - | Strop_2825 | cytochrome P450 |
| **8** | PKS | | 3475927-3510099 | | | | | |
| **9** | NRPS | | 4996301-5013814 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 80 | 426 | 4997797 | - | GTGGCCCCGCTCTTCCT | - | Strop_4414 | methyltransferase type 12 |
| | 84 | 166 | 5003350 | - | ATGGCCCGCGCGTTGCG | - | Strop_4416 | amino acid adenylation domain-containing protein |
| | 79 | 496 | 5009133 | - | CTGGCCCGGCGGCTGCC | - | Strop_4419 | kynurenine 3-monooxygenase |

**NC_003155.4 Streptomyces avermitilis MA-4680, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | PKS | | 113361-118594 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| **2** | PKS | | 486648-567017 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 157 | 490476 | - | TTGGCAAGAAAGCGGCA | pteF | SAV_409 | LuxR family transcriptional regulator |
| | 77 | 370 | 567387 | - | CTGGCGAGATTTTCGCA | pteA1 | SAV_419 | modular polyketide synthase |
| **3** | NRPS | | 751402-762285 | | | | | |
| **4** | NRPS/ PKS | | 991484-1042269 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 136 | 992855 | - | CGGGTGCGTCTGTTGCT | cyp4 | SAV_838 | cytochrome P450 |
| | 77 | 39 | 1002360 | - | CTGGCCCGGCGGTTGGC | - | SAV_845 | modular polyketide synthase |
| | 75 | 51 | 1021253 | - | CTGGCACACAAAGAGCG | nrps7-8 | SAV_855 | non-ribosomal peptide synthetase |
| | 76 | 223 | 1022254 | - | TTGGCCGAGGACGTGCT | - | SAV_856 | thioesterase |
| | 87 | 293 | 1024835 | + | CTGGCACGGGCCTTCCA | nrps7-10 | SAV_859 | non-ribosomal peptide synthetase |
| | 81 | 433 | 1026118 | + | TTGGCGCGGGCGGTGCA | nrps7-11 | SAV_860 | non-ribosomal peptide synthetase |
| | 79 | 488 | 1036207 | + | GTGGCAGGCGAACTGCC | - | SAV_867 | ABC transporter ATP-binding protein |
| **5** | PKS | | 1132045-1212960 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 76 | 1168427 | + | ATGGCACCGCACCCGCC | aveC | SAV_940 | AveC |
| | 75 | 111 | 1208710 | - | CAGGCCCGCACCCTGCT | aveBV | SAV_949 | dTDP-4-keto-6-deoxyhexose 3,5- |
| | 76 | 410 | 1212380 | - | ACGGCCCGGATACTGCA | aveBVIII | SAV_952 | dTDP-4-keto-6-deoxy-L-hexose 2,3-reductase |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 77 | 281 | 1211927 | + | GGAGCACTGCTGTTGCT | aveG | SAV_953 | thioesterase |
| | 75 | 114 | 1212094 | + | GTGGGACACGGGCTGCC | aveG | SAV_953 | thioesterase |
| **6** | PKS | | 1549424-1554224 | | | | | |
| **7** | PKS | | 1893266-1913284 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 263 | 1910520 | - | GCGGTACGCCAGTTGCC | pks2-2 | SAV_1551 | modular polyketide synthase |
| | 76 | 443 | 1912725 | - | CTGGCCCTCCTTGTGCG | pks2-3 | SAV_1552 | non-ribosomal peptide synthetase |
| | 76 | 354 | 1912250 | + | CTGGTGCGAGGCGTGCT | - | SAV_1553 | methyltransferase |
| **8** | PKS | | 2773878-2784841 | | | | | |
| **9** | PKS | | 2877941-2894413 | | | | | |
| **10** | PKS | | 2896543-2914291 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 81 | 339 | 2904983 | - | CTGGCACCGCTCACGCT | - | SAV_2378 | polyketide oxidase/hydroxylase |
| | 75 | 296 | 2906293 | + | ACGGCCCGCCCGCTGCT | - | SAV_2381 | TetR family transcriptional regulator |
| **11** | PKS | | 3477658-3493243 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 77 | 112 | 3491806 | - | CTGGCCCTGTTCCTGCC | - | SAV_2846 | hypothetical protein |
| | 75 | 159 | 3492602 | - | GAGGCACCCCTTCTGCG | - | SAV_2847 | hypothetical protein |
| | 77 | 210 | 3493453 | - | ACGGCCAGCACCTTGCT | - | SAV_2849 | hypothetical protein |
| **12** | PKS | | 3534525-3634592 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 77 | 17 | 3630164 | - | ATGGCGCGGCTCGCGCA | olmRII | SAV_2901 | LuxR family transcriptional regulator |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **13** | NRPS | | 3930306-3937493 | | | | | |
| **14** | NRPS | | 3980213-3994940 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 11 | 3985530 | - | CTGGAAGGTACCGTGCA | nrps1-1 | SAV_3197 | non-ribosomal peptide synthetase |
| | 77 | 263 | 3989602 | - | CGGGCACGGGACTGGCC | nrps1-2 | SAV_3198 | non-ribosomal peptide synthetase |
| **15** | NRPS | | 4494250-4540450 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 79 | 266 | 4496242 | - | CTGGCGCAGATCCTGCC | pptA4 | SAV_3637 | phosphopantetheinyl transferase |
| | 83 | 421 | 4500418 | - | CTGGCCCGGGAACTGCT | - | SAV_3641 | ornithine carbamoyltransferase |
| | 80 | 163 | 4516448 | - | CTGGCACGCCGAGGGCA | nrps2-2 | SAV_3643 | non-ribosomal peptide synthetase |
| | 76 | 214 | 4521073 | - | GCGGCACAGCTGTTCCC | cysK2 | SAV_3648 | cysteine synthase |
| | 77 | 416 | 4521275 | - | CTGGCGCGCAGCTCGCC | cysK2 | SAV_3648 | cysteine synthase |
| | 76 | 211 | 4523419 | - | CTGGCCAGGCTGCTGCG | fadE25 | SAV_3650 | acyl-CoA dehydrogenase |
| | 75 | 154 | 4528046 | - | GTGGCCAAGGAGATGCT | - | SAV_3652 | isomerase |
| | 75 | 388 | 4536085 | - | ATGGCGCTGCGGGTGCG | - | SAV_3661 | hypothetical protein |
| **16** | PKS | | 8553602-8561604 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 252 | 8560553 | - | CTGGCCCGGACCGAGCT | - | SAV_7185 | UDP-glucose:sterol glucosyltransferase |
| **17** | PKS | | 8776963-8789766 | | | | | |

**NC_003888.3 Streptomyces coelicolor A3(2), complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | PKS | | 104989-119654 | | | | | |
| **2** | NRPS | | 513989-526783 | | | | | |

| | | |
|---|---|---|
| **3** | PKS | 1335793-1343779 |
| **4** | NRPS | 3543335-3585724 |
| **5** | PKS | 5508078-5508078 |
| **6** | PKS | 6432566-6465258 |
| **7** | PKS | 6895193-6961810 |
| **8** | NRPS | 7106284-7116513 |
| **9** | PKS | 7590412-7602007 |
| **10** | NRPS | 8506283-8523749 |

**NC_010572.1 Streptomyces griseus subsp. griseus NBRC 13350, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | PKS | | 294490-307047 | | | | | |
| **2** | NRPS | | 480685-538065 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 80 | 142 | 535949 | - | CTGGCCCCGGTGCTGCT | - | SGR_453 | iron ABC transporter |
| | 75 | 367 | 536174 | - | CTGGCCCGGCCGCTGAA | - | SGR_453 | iron ABC transporter |
| **3** | NRPS | | 659688-691515 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 80 | 48 | 664691 | + | GTGGCGTGCCTTCTGCA | - | SGR_578 | putative SyrP-like protein |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 76 | 167 | 676251 | + | CTGGCCCGGCTGGAGCT | - | SGR_583 | putative NRPS |
| | 82 | 298 | 685576 | - | CTGGCCCGCGCTCTGCT | - | SGR_588 | hypothetical protein |
| **4** | PKS | | 704021-717508 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 480 | 703541 | + | GTGCGACGGCGGTTGCC | tebC | SGR_604 | putative enediyne biosynthesis protein |
| | 76 | 474 | 704168 | + | AGGGCACCCGAGGTGCT | unbL1 | SGR_605 | putative enediyne biosynthesis protein |
| | 79 | 329 | 708297 | + | GTGGCACGCCAGTGGGA | unbU | SGR_608 | putative enediyne biosynthesis protein |
| **5** | NRPS | | 762595-774309 | | | | | |
| **6** | PKS | | 937762-968455 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 299 | 940053 | + | CTGCCACAGCCGATGCA | - | SGR_803 | putative large multi-functional protein |
| | 81 | 130 | 956182 | - | CTGGCACGGCAGCTCCA | - | SGR_813 | putative FAD-dependent oxidoreductase |
| **7** | NRPS | | 1044652-1071469 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 294 | 1046802 | - | GGGGCACGCACATCGCC | - | SGR_890 | hypothetical protein |
| | 75 | 437 | 1060513 | - | CTGGCTCACCGGATGCG | - | SGR_896 | putative O-methyltransferase |
| **8** | NRPS/ PKS | | 2924650-2941895 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 374 | 2930023 | + | ATGGTGCGGCAGTTGGA | - | SGR_2485 | putative NRPS |
| | 75 | 470 | 2935806 | + | ATGGCCCATGTGCTGCC | - | SGR_2486 | putative aminopeptidase 2 |
| | 78 | 400 | 2937233 | + | ATGGCGCACGCGGTGCA | - | SGR_2487 | putative thioesterase |
| **9** | NRPS | | 3048501-3074603 | | | | | |
| **10** | NRPS | | 3762124-3830096 | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **11** | PKS | | 7102079-7171428 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 362 | 7161653 | + | TCGGCACGCTCGGTGCC | pks1-7 | SGR_6083 | putative type-I PKS |
| **12** | PKS | | 7286041-7354154 | | | | | |
| **13** | PKS | | 7551740-7619551 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 83 | 426 | 7551314 | + | CTGGCAGGGCAGTTGGA | - | SGR_6359 | hypothetical protein |
| | 78 | 256 | 7570469 | - | CTGGGACTGTCCTCGCT | - | SGR_6360 | hypothetical protein |
| | 75 | 422 | 7574873 | - | GTGGCCGACGTAGTGCA | - | SGR_6365 | hypothetical protein |
| | 76 | 266 | 7596949 | - | CTGGCCGGGCACCTGCG | pks3-3 | SGR_6371 | putative type-I PKS |
| | 81 | 244 | 7619795 | - | GTGGCCCAGCGGCTGCA | pks3-1 | SGR_6373 | putative type-I PKS |
| **14** | NRPS | | 8011976-8029693 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 81 | 164 | 8012818 | + | GCGGCACGATCCGTGCA | - | SGR_6714 | putative ABC-type Fe3+-siderophore transporter |
| | 85 | 326 | 8015602 | + | CTGGCCCGGCTGGTGCT | - | SGR_6716 | putative NRPS |
| | 77 | 424 | 8030117 | - | CTGGCACATCGGCTGGT | - | SGR_6718 | putative esterase |
| **15** | NRPS | | 8049140-8065069 | | | | | |
| **16** | PKS | | 8100042-8154770 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 30 | 8100012 | + | CGGGCAGAGTCCTTGCG | pks4-1 | SGR_6776 | putative NRPS-type-I PKS fusion protein |
| | 77 | 383 | 8110593 | + | TTCGCACGGACGGTGCG | pks4-2 | SGR_6777 | putative type-I PKS |
| | 75 | 269 | 8124572 | + | TTCGGACGCATCGTGCT | - | SGR_6780 | putative malonyl-CoA:ACP transacylase |

**NC_003228.3 Bacteroides fragilis NCTC 9343, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS | | 3294535-3306572 | | | | | |

| 2 | NRPS | | 3635663-3639879 | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 357 | 3638482 | + | GCGGCCGACAGCTTGCT | - | BF3118 | hypothetical protein |

**NC_005125.1 Gloeobacter violaceus PCC 7421, complete genome**

| 1 | PKS | | 2057662-2100787 | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 304 | 2064191 | - | ATGGTCCGCATCTTCCT | - | gll1939 | cyclopropane-fatty-acyl-phospholipid synthase |
| | 83 | 457 | 2064344 | - | CTGGCCCGAGATCTGCT | - | gll1939 | cyclopropane-fatty-acyl-phospholipid synthase |
| | 77 | 280 | 2065156 | - | ATGGCCCAGTTGTGGCC | - | gll1940 | fatty acid desaturase |
| | 77 | 423 | 2072927 | - | GCGGCCCTGGGGTTGCT | - | gll1946 | fatty acid desaturase |
| | 81 | 329 | 2073836 | - | GTGGGGCGGCTTTTGCC | - | gll1947 | fatty acid desaturase |
| | 75 | 166 | 2078085 | - | CCGGCAGCACTCCTGCA | - | gll1950 | long-chain fatty-acid-CoA ligase |
| | 76 | 219 | 2078138 | - | ATGGCTTCGCACTCGCT | - | gll1950 | long-chain fatty-acid-CoA ligase |
| | 77 | 466 | 2078385 | - | GGGGCGCACCCCTTGCG | - | gll1950 | long-chain fatty-acid-CoA ligase |
| | 81 | 411 | 2079094 | - | CTGGAGCGCTGTTTGCA | - | gll1951 | phosphopantetheinyltransferase family protein |
| | 76 | 461 | 2088597 | - | GTGGCTGGACAGTTGAT | - | gll1954 | polyketide synthase |
| | 76 | 146 | 2088597 | - | GTGGCTGGACAGTTGAT | - | gll1955 | polyketide synthase |
| | 75 | 317 | 2088768 | - | ATGGCCCTGATCTTTCG | - | gll1955 | polyketide synthase |
| | 75 | 277 | 2094691 | - | CTGGACCAGTTGCTGCA | - | gll1957 | glycolipid synthase |
| 2 | PKS | | 2999221-3029646 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 77 | 396 | 2998825 | + | CTGGCACCAGAAACGCT | - | glr2821 | WD repeat-containing protein |
| | 81 | 288 | 3009343 | - | CTGGCGGGGCCTTTGCG | - | gll2826 | hypothetical protein |
| | 77 | 307 | 3010641 | - | ACGGGACATGATTTGCT | - | gll2827 | hypothetical protein |
| | 80 | 157 | 3014822 | - | CAGGCGCTGATCTTGCT | - | gll2829 | polyketide synthase |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 79 | 450 | 3023483 | - | GTGGCACAGACCGTCCT | - | gll2836 | alcohol dehydrogenase |
| | 79 | 174 | 3023483 | - | GTGGCACAGACCGTCCT | - | gsl2837 | hypothetical protein |
| **3** | PKS | | 3033832-3050434 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 221 | 3040466 | + | CTGGCGCAGTGTTTGAG | - | glr2850 | thioesterase |
| | 79 | 41 | 3040646 | + | CTGGCAACTCTGGTGCA | - | glr2850 | thioesterase |
| | 78 | 393 | 3048049 | + | GTGGCCCGTCCGATGCC | - | glr2859 | oxidoreductase |
| **4** | PKS | | 4421532-4440489 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 215 | 4422391 | - | CTGGAAAGTCCATCGCT | - | gll4219 | hypothetical protein |
| | 83 | 331 | 4422933 | - | CTGGCCGGAGGATTGCT | - | gll4220 | hypothetical protein |
| | 75 | 154 | 4431563 | - | CTGGTAGGCCCCATGCG | - | gll4225 | glycolipid synthase |

**NC_010296.1 Microcystis aeruginosa NIES-843, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS/ PKS | | 2507861-2539146 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 77 | 38 | 2518358 | + | TGGGCGCAAGCCTTGCG | - | MAE_27820 | amino acid adenylation |
| | 78 | 118 | 2526583 | + | TGGGCGCGAGCATTGCG | - | MAE_27860 | short-chain dehydrogenase/reductase |
| | 79 | 235 | 2537743 | + | GTGGCATCGGTTGTGCG | - | MAE_27910 | hypothetical protein |
| **2** | NRPS/ PKS | | 3486436-3541027 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 79 | 340 | 3528001 | + | GTGGCAATTGTCCTGCA | mcyF | MAE_38620 | McyF protein |
| | 79 | 114 | 3540068 | + | ATGGCAATATTTCTGCA | mcyJ | MAE_38660 | McyJ protein |
| **3** | NRPS | | 5194435-5220124 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 130 | 5208380 | - | TTGGCATGGTATTCTCA | - | MAE_56550 | hypothetical protein |
| | 80 | 148 | 5209406 | - | TTGGCTCGACCGTTTCT | - | MAE_56560 | bacilysin biosynthesis protein BacA-like protein |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 80 | 198 | 5209456 | - | ATGGCTTAGACCTTGCC | - | MAE_56560 | bacilysin biosynthesis protein BacA-like protein |
| **4** | NRPS | | 5519975-5552181 | | | | | |

**NC_010628.1 Nostoc punctiforme PCC 73102, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | PKS | | 56243-74142 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 81 | 379 | 58235 | - | GTGGCAATATAATTGCC | - | Npun_R0038 | short-chain dehydrogenase/reductase SDR |
| | 79 | 405 | 60343 | - | TTCGCAAAACTATTGCT | - | Npun_R0039 | polyketide synthase thioester reductase subunit HglB |
| | 76 | 291 | 68964 | - | ATGGCCCAAAGTTTTCA | - | Npun_R0042 | KR |
| **2** | PKS | | 1549256-1571247 | | | | | |
| **3** | PKS | | 2521301-2548634 | | | | | |
| **4** | NRPS/ PKS | | 2648282-2708401 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 183 | 2660116 | - | TTGTCATAGACATTGCC | - | Npun_R2175 | KR |
| | 79 | 484 | 2664987 | - | ATGGCAGGTATAATGCC | - | Npun_R2178 | WD-40 repeat-containing protein |
| | 76 | 206 | 2704134 | + | GTGATACAGTAATTGCA | - | Npun_F2186 | alcohol dehydrogenase |
| | 81 | 362 | 2705124 | + | CAGGCATTCCTATTGCT | - | Npun_F2187 | pyrroline-5-carboxylate reductase |
| **5** | NRPS | | 3054973-3081156 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 50 | 3055694 | + | TGGGCAAGGATATAGCA | - | Npun_F2460 | amino acid adenylation domain-containing protein |
| | 76 | 93 | 3065615 | + | GTGACTCAAGAGTTGCT | - | Npun_F2462 | amino acid adenylation domain-containing protein |
| **6** | NRPS/ PKS | | 3722118-3780010 | | | | | |
| **7** | NRPS/ PKS | | 3899401-3958883 | | | | | |

| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
|---|---|---|---|---|---|---|---|---|
| | 75 | 191 | 3899210 | + | TTCACACGTATGTTGCT | - | Npun_F3155 | hypothetical protein |
| | 86 | 415 | 3925855 | + | ATGGCGTGAATTTTGCT | - | Npun_F3168 | hypothetical protein |
| | 76 | 296 | 3950844 | + | GTGGCTAGGGTGCTGCG | - | Npun_F3173 | amino acid adenylation domain-containing protein |
| | 75 | 148 | 3950992 | + | AAGGCCCTAGTATTGCC | - | Npun_F3173 | amino acid adenylation domain-containing protein |
| | 75 | 54 | 3955358 | + | TAGGGAAGATAATTGCT | - | Npun_F3174 | isoprenylcysteine carboxyl methyltransferase |
| **8** | PKS | | 4180429-4210477 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 482 | 4181756 | - | CTCGAACATTTATTGCT | - | Npun_R3355 | thioesterase |
| | 75 | 474 | 4185955 | + | TTGTCAGGAGGGTTGCC | - | Npun_F3359 | beta-ketoacyl synthase |
| | 76 | 364 | 4186065 | + | CTGGCTAGAGAAATGCT | - | Npun_F3359 | beta-ketoacyl synthase |
| | 83 | 221 | 4201954 | + | ATGGCAGTGATATTGCC | - | Npun_F3363 | short-chain dehydrogenase/reductase SDR, |
| **9** | NRPS/ PKS | | 4256267-4330313 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 81 | 468 | 4259675 | - | ATGGCAAGAGTTTTGTA | - | Npun_R3419 | thioesterase |
| | 81 | 165 | 4259675 | - | ATGGCAAGAGTTTTGTA | - | Npun_R3420 | MbtH domain-containing protein |
| | 76 | 491 | 4260403 | - | ATTGCACGAGGCTTTCT | - | Npun_R3421 | hypothetical protein |
| | 77 | 499 | 4280094 | - | ACGGCCCGACTGTAGCT | - | Npun_R3429 | condensation domain-containing protein |
| | 76 | 482 | 4286311 | - | TTGGATCGGCGATCGCT | - | Npun_R3430 | beta-ketoacyl synthase |
| | 75 | 239 | 4307280 | - | ATGGCTCTCTTATTGAA | - | Npun_R3436 | amino acid adenylation domain-containing protein |
| | 78 | 331 | 4307372 | - | TTGGCATCCAGCTTGTT | - | Npun_R3436 | amino acid adenylation domain-containing protein |
| | 81 | 358 | 4321256 | - | ATGGCAGCACTTGTGCA | - | Npun_R3448 | taurine catabolism dioxygenase TauD/TfdA |
| **10** | NRPS/ PKS | | 8150738-8167006 | | | | | |

**NC_008536.1 Solibacter usitatus Ellin6076, complete genome**

| | | | |
|---|---|---|---|
| 1 | PKS | 2480784-2494608 | |

| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
|---|---|---|---|---|---|---|---|---|
| | 80 | 22 | 2489341 | + | CTGGCCGGGAGATTGCG | - | Acid_1973 | beta-ketoacyl synthase |
| | 77 | 463 | 2493837 | + | ACGGCAGACCCATTGCC | - | Acid_1974 | putative acyl carrier protein |
| 2 | PKS | | 3884932-3895290 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 213 | 3885870 | - | GTGGATCGAGTTTCGCT | - | Acid_3076 | 4'-phosphopantetheinyl transferase |
| | 78 | 300 | 3885957 | - | ATGGCACCGCGGCTGAT | - | Acid_3076 | 4'-phosphopantetheinyl transferase |
| | 77 | 50 | 3893616 | - | CGGGCACGCCGCTTGAG | - | Acid_3077 | beta-ketoacyl synthase |

**NC_009674.1 Bacillus cereus subsp. cytotoxis NVH 391-98, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS | | 417602-429629 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 80 | 491 | 422474 | + | CTGGTAGAGGATTTGCA | - | Bcer98_0368 | amino acid adenylation domain-containing protein |
| **2** | NRPS | | 1197549-1205485 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 77 | 484 | 1198103 | + | ACGTCACGAGGTTTGCA | - | Bcer98_1088 | 2-nitropropane dioxygenase NPD |
| **3** | NRPS | | 1853248-1878954 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 310 | 1855588 | + | TTGGAGCGTCTTTGGCT | - | Bcer98_1744 | AMP-dependent synthetase and ligase |
| | 80 | 293 | 1856961 | + | TAGGCGCTGGTGTTGCT | - | Bcer98_1745 | hypothetical protein |
| **4** | PKS | | 3245067-3256221 | | | | | |
| **5** | PKS | | 1781906-1859783 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 77 | 130 | 1785974 | + | CCGGCACCTTAGCTGCT | pksE | BSU17120 | enzyme involved in polyketide synthesis |
| | 81 | 446 | 1791566 | + | CTGCCACAGATATTGCC | pksI | BSU17170 | polyketide biosynthesis enoyl-CoA hydratase |
| **6** | NRPS | | 1949682-2002351 | | | | | |

| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
|---|---|---|---|---|---|---|---|---|
| | 81 | 138 | 1956312 | - | GTGGAACGGATGCTGCG | yngI | BSU18250 | AMP-binding domain protein |
| | 82 | 303 | 1956477 | - | CTGGCACGGCAAATGGT | yngI | BSU18250 | AMP-binding domain protein |
| | 79 | 340 | 1957700 | - | CGGGCACTGCTCTTGTT | yngJ | BSU18260 | acyl-CoA dehydrogenase, short-chain specific |
| | 76 | 263 | 1960350 | - | ATGGGAAGGCGCCTGCG | yngL | BSU18290 | putative integral inner membrane protein |
| | 77 | 161 | 1975017 | - | ATTGAACACTTTTTGCT | ppsD | BSU18310 | plipastatin synthetase |
| **7** | NRPS | | 3280519-3297919 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 391 | 3290651 | - | GTGGCACCTGTCCAGCG | dhbE | BSU31980 | 2,3-dihydroxybenzoate-AMP ligase |
| | 82 | 299 | 3291784 | - | TTGGCCTTGAGCTTGCA | dhbC | BSU31990 | isochorismate synthase DhbC |
| | 78 | 267 | 3296537 | - | CAGGCACAGCGACTGCC | yuiF | BSU32040 | amino acid transporter |
| | 80 | 328 | 3296598 | - | ATGGCAGGAACGTTTCT | yuiF | BSU32040 | amino acid transporter |
| **8** | NRPS | | 3952275-3958482 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 347 | 3957044 | + | ATGGCACCGCCTATGGT | ywaA | BSU38550 | branched-chain amino acid aminotransferase |

**NC_003030.1 Clostridium acetobutylicum ATCC 824, complete genome**

| | | | |
|---|---|---|---|
| **1** | PKS | 3529432-3534819 | |

**NC_009706.1 Clostridium kluyveri DSM 555, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS | | 1554646-1602560 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 321 | 1564208 | + | CTCGAATGGATATTGCT | - | CKL_1507 | NRPS cyclization domain containing protein |
| | 77 | 233 | 1578681 | + | GAGGCATAGAACATGCA | - | CKL_1515 | hypothetical protein |
| **2** | NRPS/ PKS | | 1818097-1833491 | | | | | |
| **3** | NRPS/ PKS | | 2414655-2426709 | | | | | |

| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
|---|---|---|---|---|---|---|---|---|
| | 76 | 361 | 2419626 | - | TTGATACAGTTTTTGCA | - | CKL_2354 | nonribosomal peptide synthetase |

**NC_008497.1 Lactobacillus brevis ATCC 367, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS | | 1293616-1300116 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 122 | 1297040 | - | GTGCAACTGTTGTTGCA | - | LVIS_1323 | D-alanyl transfer protein |
| | 81 | 491 | 1300607 | - | GGGGCACCACGGTTGCG | - | LVIS_1327 | D-Ala-teichoic acid biosynthesis protein (putative) |

**NC_002973.6 Listeria monocytogenes str. 4b F2365, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS | | 1002948-1007380 | | | | | |

**NC_002952.2 Staphylococcus aureus subsp. aureus MRSA252, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS | | 198393-206225 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 81 | 487 | 197906 | + | CTAGCATGCTTGTTGCT | - | SAR0180 | putative non-ribosomal peptide synthetase |

**NC_003063.2 Agrobacterium tumefaciens str. C58 chromosome linear, complete sequence**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS | | 74687-80557 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 121 | 79699 | - | ATGGCTGCTCGGTTGCT | - | Atu3073 | aspartate racemase |
| | 83 | 322 | 80879 | - | CCGGCACGGCGCTCGCT | - | Atu3074 | short chain dehydrogenase |
| **2** | NRPS/ PKS | | 716058-769035 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 463 | 729910 | - | CTGGTCCGGATTTCGCT | - | Atu3673 | siderophore biosynthesis protein |
| | 78 | 187 | 729910 | - | CTGGTCCGGATTTCGCT | - | Atu3674 | hypothetical protein |
| | 85 | 425 | 734796 | + | ATGGCACGCTACTGGCG | - | Atu3677 | polyketide synthase, siderophore biosynthesis |
| | 75 | 293 | 734928 | + | GCGGCGCGTTTCTCGCT | - | Atu3677 | polyketide synthase, siderophore biosynthesis |
| | 75 | 28 | 737448 | + | ATGCAAGGCTTCTTGCA | - | Atu3680 | putative siderophore biosynthesis protein |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 80 | 62 | 742445 | + | ATGGCAGCGCTGTTGAT | - | Atu3682 | non-ribosomal peptide synthetase, siderophore |
| | 86 | 328 | 765024 | + | TTGGCACGGATGTCGCC | fecD | Atu3690 | ABC transporter, membrane spanning protein (iron (III) |
| | 79 | 247 | 765105 | + | CCGGCAGCGCCGTTGCA | fecD | Atu3690 | ABC transporter, membrane spanning protein (iron (III) |
| | 75 | 470 | 767004 | + | CTGGCGCCCCAGGTGCC | - | Atu3692 | sigma factor |

**NC_012988.1 Methylobacterium extorquens DM4, complete genome**

| | | | |
|---|---|---|---|
| **1** | PKS | 2613383-2622159 | |

**NC_012988.1 Methylobacterium extorquens DM4, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS | 4859328-4866323 | | | | | | |
| 2 | NRPS | 5649466-5678659 | | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 86 | 134 | 5649332 | + | CAGGCACAGGCTTTGCT | - | METDI5670 | putative UDP-N-acetyl-D-mannosaminuronic acid |
| | 77 | 53 | 5653547 | + | ATGGCGCCCGTCTTCCT | - | METDI5674 | putative glycosyl transferase |
| | 75 | 105 | 5662917 | - | CCGGCAACCACCTTGCC | - | METDI5678 | hypothetical protein |
| | 75 | 46 | 5667529 | - | ATGGCAGGTCCGCCGCT | - | METDI5683 | hypothetical protein |
| | 75 | 148 | 5675379 | - | CTGGCACCGCGCTTCGA | murE | METDI5688 | UDP-N-acetylmuramoylalanyl-D-glutamate--2, 6- |
| | 75 | 355 | 5677962 | - | TTGGTGCGCGGCCTGCA | - | METDI5690 | hypothetical protein |

**NC_003296.1 Ralstonia solanacearum GMI1000 plasmid pGMI1000MP, complete sequence**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS/ PKS | 782247-820777 | | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 194 | 802722 | + | CTGGCCGAACTCTGGCA | RSp0642 | RS05859 | peptide synthetase protein |
| **2** | NRPS | 1782597-1797233 | | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 13 | 1787443 | + | CTCCCACCGCTTTTGCA | RSp1421 | RS03121 | hypothetical protein |
| **3** | NRPS/ PKS | 1945814-1979076 | | | | | | |

| Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
|---|---|---|---|---|---|---|---|
| 79 | 452 | 1949544 | - | GTCGCACGCGATCTGCA | tISRso15 | RSp1547 | ISRSO15-transposase protein |
| 83 | 50 | 1950406 | - | GTGGCACGTTCAATGCG | RSp1549 | RS02105 | hypothetical protein |
| 82 | 103 | 1951581 | - | CCGGCATTGCTTTTGCA | RSp1551 | RS02107 | hypothetical protein |
| 80 | 248 | 1951726 | - | GTGGCGCTGCTGGTGCA | RSp1551 | RS02107 | hypothetical protein |
| 75 | 279 | 1957715 | + | GTGGGATGGGATATGCC | hexR | RSp1556 | putative transcription regulation repressor HEXR transcription |
| 75 | 60 | 1957934 | + | GCGGCCCCGCGCTTGCC | hexR | RSp1556 | putative transcription regulation repressor HEXR transcription |
| 77 | 289 | 1960873 | - | GCGGCACTGGGCCTGCA | pgl | RSp1558 | 6-phosphogluconolactonase oxidoreductase protein |
| 77 | 106 | 1962311 | + | GAGGCACGCAATTCGCG | edd | RSp1560 | phosphogluconate dehydratase |
| 78 | 135 | 1964859 | - | CGGGCATCGTTGTTGCG | RSp1561 | RS02117 | hypothetical protein |
| 80 | 247 | 1964633 | + | ATGGCGCGGATGGTGCC | RSp1562 | RS02118 | hypothetical protein |
| 77 | 139 | 1965305 | + | TGCGCATGGATTTTGCA | RSp1563 | RS02119 | hypothetical protein |
| 75 | 312 | 1974525 | - | CTGGAATGGCGTTAGCG | RSp1571 | RS02127 | transcriptional regulator transcription regulator protein |

**NC_008095.1 Myxococcus xanthus DK 1622, complete genome**

| **1** | NRPS | 1488859-1520119 | | | | | |
|---|---|---|---|---|---|---|---|
| Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| 82 | 476 | 1490324 | + | GTGGCGCGGCCCTTGAA | - | MXAN_1276 | glutamate-cysteine ligase family 2 protein |
| 77 | 461 | 1497379 | - | TCTGGAACGCGTGCAGCAGGGG | - | MXAN_1281 | hypothetical protein |
| 75 | 1 | 1500305 | + | AGGGCAAGGCATTTCCA | - | MXAN_1284 | 2-isopropylmalate synthase/homocitrate synthase |
| 75 | 159 | 1503794 | + | TTGGACCGGCTCGTGCG | abcA | MXAN_1286 | ABC transporter permease/ATP-binding protein |
| 75 | 63 | 1503890 | + | TTGGCGCGGAAGGCGCA | abcA | MXAN_1286 | ABC transporter permease/ATP-binding protein |
| 76 | 304 | 1505775 | + | CATTGCCCGCGCCCTGCTGCTG | - | MXAN_1287 | hypothetical protein |
| 78 | 233 | 1505846 | + | GAGGCCCAACTGTTGCG | - | MXAN_1287 | hypothetical protein |
| 77 | 252 | 1517769 | - | CCCTGGCACGTGACGTGTTCAA | - | MXAN_1291 | non-ribosomal peptide synthetase |
| 79 | 252 | 1517769 | - | CCCTGGCACGTGACGTGTTCA | - | MXAN_1291 | non-ribosomal peptide synthetase |
| 80 | 250 | 1517767 | - | CTGGCACGTGACGTGTT | - | MXAN_1291 | non-ribosomal peptide synthetase |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 76 | 460 | 1517977 | - | GTGGTGCGGCGGATGCA | - | MXAN_1291 | non-ribosomal peptide synthetase |
| | 76 | 455 | 1518967 | - | GCTGGCGCGCGAGGAACTGAAG | - | MXAN_1292 | hypothetical protein |
| | 75 | 300 | 1520419 | - | ACCGGTACACACGTTGGTGGCG | hemG | MXAN_1293 | protoporphyrinogen oxidase |
| | 76 | 336 | 1520455 | - | GGAGACACCGGCGTTGTTGATG | hemG | MXAN_1293 | protoporphyrinogen oxidase |
| **2** | NRPS | | | 1832662-1910318 | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 298 | 1834249 | - | GTGGCACGTCGCTGGTG | - | MXAN_1560 | class I aminotransferase |
| | 79 | 84 | 1836766 | - | TCTGGCACGGTGACTGCTAAGC | ahpC | MXAN_1564 | alkyl hydroperoxide reductase C |
| | 89 | 85 | 1836767 | - | TTCTGGCACGGTGACTGCTAAG | ahpC | MXAN_1564 | alkyl hydroperoxide reductase C |
| | 89 | 83 | 1836765 | - | CTGGCACGGTGACTGCT | ahpC | MXAN_1564 | alkyl hydroperoxide reductase C |
| | 88 | 85 | 1836767 | - | TTCTGGCACGGTGACTGCTAA | ahpC | MXAN_1564 | alkyl hydroperoxide reductase C |
| | 77 | 420 | 1841516 | - | CGCTGGCGCGGTTGCTGCGCAT | - | MXAN_1567 | urea amidolyase-like protein |
| | 79 | 420 | 1841516 | - | CGCTGGCGCGGTTGCTGCGCA | - | MXAN_1567 | urea amidolyase-like protein |
| | 80 | 418 | 1841514 | - | CTGGCGCGGTTGCTGCG | - | MXAN_1567 | urea amidolyase-like protein |
| | 76 | 61 | 1842452 | - | GGCTGGCACGCCAGCGTCTCCG | - | MXAN_1568 | LamB/YcsF family/allophanate hydrolase family protein |
| | 78 | 61 | 1842452 | - | GGCTGGCACGCCAGCGTCTCC | - | MXAN_1568 | LamB/YcsF family/allophanate hydrolase family protein |
| | 75 | 461 | 1842852 | - | AGCTGCCACGCGCCTGGCGCG | - | MXAN_1568 | LamB/YcsF family/allophanate hydrolase family protein |
| | 76 | 293 | 1844083 | + | ATGACTCGCAGCTTGCG | - | MXAN_1570 | class V aminotransferase |
| | 75 | 376 | 1847137 | + | GCTGGCCTCCACGGTGGTGACG | - | MXAN_1573 | AMP-binding domain protein |
| | 77 | 227 | 1849786 | - | CCTGGCGCGCGCCGGGGTGAAG | - | MXAN_1574 | TfoX domain-containing protein |
| | 75 | 60 | 1854462 | - | CACGGCGCAGCGCTTGCTAAGG | - | MXAN_1578 | metallo-beta-lactamase family protein |
| | 75 | 61 | 1854463 | - | CCACGGCGCAGCGCTTGCTAAG | - | MXAN_1578 | metallo-beta-lactamase family protein |
| | 81 | 59 | 1854461 | - | ACGGCGCAGCGCTTGCT | - | MXAN_1578 | metallo-beta-lactamase family protein |
| | 76 | 104 | 1854518 | + | GGGGCGCAATCCTTGCG | - | MXAN_1579 | hypothetical protein |
| | 77 | 75 | 1854547 | + | GGGGCGCAGGTCTTGCG | - | MXAN_1579 | hypothetical protein |
| | 80 | 180 | 1860605 | - | AAGTGGCGCGGCCGCTGCTTGG | - | MXAN_1584 | hypothetical protein |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 80 | 180 | 1860605 | - | AAGTGGCGCGGCCGCTGCTTG | - | MXAN_1584 | hypothetical protein |
| | 84 | 178 | 1860603 | - | GTGGCGCGGCCGCTGCT | - | MXAN_1584 | hypothetical protein |
| | 75 | 146 | 1861734 | + | GTGGGACATCCGGTGCG | - | MXAN_1587 | hypothetical protein |
| | 76 | 130 | 1866438 | - | GGAGGCACGCCGCACGCTGGAA | - | MXAN_1590 | putative para-aminobenzoate synthase, component I |
| | 75 | 190 | 1871142 | - | CCTGCTCTGCACGGAGCTGAAG | - | MXAN_1592 | hypothetical protein |
| | 75 | 203 | 1871155 | - | GGCTGGAAGTGGACCTGCTCTG | - | MXAN_1592 | hypothetical protein |
| | 77 | 203 | 1871155 | - | GGCTGGAAGTGGACCTGCTCT | - | MXAN_1592 | hypothetical protein |
| | 78 | 130 | 1875006 | - | CATGGCACGGGGGACGAAGAGG | - | MXAN_1595 | hypothetical protein |
| | 75 | 428 | 1875304 | - | GGAGGATGACCGGCTGCTGATG | - | MXAN_1595 | hypothetical protein |
| | 76 | 277 | 1877522 | + | GGTGGCCAGCCGGGCGCGGACG | - | MXAN_1599 | putative methyltransferase |
| | 75 | 44 | 1877755 | + | GTGGTGCCGCTGCTGCT | - | MXAN_1599 | putative methyltransferase |
| | 75 | 398 | 1878919 | + | GGAGGCCCGCTCCGAGCTGTGG | - | MXAN_1600 | class I aminotransferase |
| | 79 | 267 | 1879050 | + | GGCTGGCTCGGGCGTTGCGTCC | - | MXAN_1600 | class I aminotransferase |
| | 85 | 270 | 1879047 | + | CTGGCTCGGGCGTTGCG | - | MXAN_1600 | class I aminotransferase |
| | 80 | 268 | 1879049 | + | GGCTGGCTCGGGCGTTGCGTC | - | MXAN_1600 | class I aminotransferase |
| | 79 | 398 | 1881868 | - | GTGGCCCGGGCACTGCG | - | MXAN_1601 | fatty acid desaturase family protein |
| | 76 | 297 | 1883277 | + | GTGGCCTGGGTGGTGCC | - | MXAN_1603 | putative non-ribosomal peptide synthetase |
| | 78 | 262 | 1889174 | + | TGCTGGAAGGGTCGCTGCTGGG | - | MXAN_1605 | putative permease |
| | 79 | 263 | 1889173 | + | TGCTGGAAGGGTCGCTGCTGG | - | MXAN_1605 | putative permease |
| | 78 | 265 | 1889171 | + | CTGGAAGGGTCGCTGCT | - | MXAN_1605 | putative permease |
| | 75 | 256 | 1890469 | + | GGTGGCCGCCTTCATGCTGGAG | - | MXAN_1606 | hypothetical protein |
| | 75 | 260 | 1890465 | + | GTGGCCGCCTTCATGCT | - | MXAN_1606 | hypothetical protein |
| | 75 | 294 | 1910612 | - | GTAGGAACGCACGACCCTCAAG | - | MXAN_1608 | hypothetical protein |
| **3** | NRPS/ PKS | | 3251570-3266656 | | | | | |
| **4** | PKS | | 4029025-4040172 | | | | | |

| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
|---|---|---|---|---|---|---|---|---|
| | 76 | 178 | 4029404 | + | GCTGGTCCACATCATGC TCGAG | - | MXAN_3460 | Gfo/Idh/MocA family oxidoreductase |
| | 76 | 182 | 4029400 | + | CTGGTCCACATCATGCT | - | MXAN_3460 | Gfo/Idh/MocA family oxidoreductase |
| **5** | NRPS/ PKS | | 4220394-4350875 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 319 | 4254532 | - | ATGGCGGAAGAGCTGC T | - | MXAN_3627 | hypothetical protein |
| | 75 | 398 | 4255390 | - | ACTGCCACGCCAGCAG CTCGGG | - | MXAN_3628 | non-ribosomal peptide biosynthesis thioesterase |
| | 75 | 399 | 4255391 | - | AACTGCCACGCCAGCA GCTCG | - | MXAN_3628 | non-ribosomal peptide biosynthesis thioesterase |
| | 75 | 308 | 4259689 | - | CTGGCGGACATCTGGC A | - | MXAN_3630 | polyketide synthase type I |
| | 75 | 287 | 4264309 | - | GAAGGAACAGCTGTTG CGCATG | - | MXAN_3631 | polyketide synthase type I |
| | 78 | 286 | 4264308 | - | AAGGAACAGCTGTTGC G | - | MXAN_3631 | polyketide synthase type I |
| | 76 | 211 | 4284118 | + | AAGGCGTGCGCATTGC T | - | MXAN_3635 | non-ribosomal peptide synthase/polyketide synthase |
| | 77 | 32 | 4334952 | + | CTGGTACTTCGAGTGCA | - | MXAN_3638 | M19 family peptidase |
| | 76 | 311 | 4347401 | - | GTGGAAAACCATCTGCT | - | MXAN_3644 | isochorismatase |
| | 75 | 268 | 4351143 | - | GTGGCACGCGAAATCC C | - | MXAN_3647 | 2,3-dihydroxybenzoate-2,3-dehydrogenase |
| **6** | PKS | | 4498820-4544105 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 284 | 4498536 | + | TTGAGACGGAGATTGC G | - | MXAN_3778 | DnaK family protein |
| **7** | PKS | | 4729983-4809194 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 337 | 4731192 | - | CGTGGGCCGGTTCCGG CTCAAG | - | MXAN_3931 | hypothetical protein |
| | 76 | 73 | 4740097 | - | TCAGACACGCCCATCGC TGATC | - | MXAN_3932 | polyketide synthase |
| | 75 | 90 | 4740114 | - | CTTGAAAAACCGGTTGC TCAGA | - | MXAN_3932 | polyketide synthase |
| | 75 | 103 | 4755572 | - | AGGTGGTACTTCGCTTA CTCG | - | MXAN_3933 | mixed type I polyketide synthase - peptide synthetase |
| | 76 | 101 | 4755570 | - | GTGGTACTTCGCTTACT | - | MXAN_3933 | mixed type I polyketide synthase - peptide synthetase |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 75 | 428 | 4784222 | - | ATTGGCTCGCGTCTGGCTGCGC | - | MXAN_3935 | non-ribosomal peptide synthase/polyketide synthase |
| | 77 | 221 | 4784015 | - | TGCTGGCACGAACCCTGGGCG | - | MXAN_3935 | non-ribosomal peptide synthase/polyketide synthase |
| | 82 | 427 | 4784221 | - | TTGGCTCGCGTCTGGCT | - | MXAN_3935 | non-ribosomal peptide synthase/polyketide synthase |
| | 78 | 219 | 4784013 | - | CTGGCACGAACCCTGGG | - | MXAN_3935 | non-ribosomal peptide synthase/polyketide synthase |
| | 75 | 315 | 4784109 | - | TTGGCGCTGAGCGTGCC | - | MXAN_3935 | non-ribosomal peptide synthase/polyketide synthase |
| | 78 | 333 | 4790662 | - | GTTGGTCCGCCGGGCGGTGGAG | - | MXAN_3936 | polyketide synthase |
| | 75 | 155 | 4797857 | - | GCTGTCGCTCGTGACGCTCAAG | - | MXAN_3938 | polyketide synthase |
| | 76 | 245 | 4800766 | - | GGCTGGCTCTGCAACTGCGCAG | - | MXAN_3941 | polyketide beta-ketoacyl:acyl carrier protein synthase |
| | 76 | 245 | 4800766 | - | GGCTGGCTCTGCAACTGCGCA | - | MXAN_3941 | polyketide beta-ketoacyl:acyl carrier protein synthase |
| | 75 | 243 | 4800764 | - | CTGGCTCTGCAACTGCG | - | MXAN_3941 | polyketide beta-ketoacyl:acyl carrier protein synthase |
| | 77 | 318 | 4804366 | - | GTGGCGCATGCCTTTCT | - | MXAN_3943 | cytochrome P450 family protein |
| | 77 | 97 | 4805951 | - | CCTGGACAGCCTGCGGCTGCAG | taF | MXAN_3945 | polyketide TA biosynthesis protein TaF |
| | 76 | 464 | 4806318 | - | CCTGTCGCGCGTGGTGCTGGAC | taF | MXAN_3945 | polyketide TA biosynthesis protein TaF |
| | 76 | 220 | 4806318 | - | CCTGTCGCGCGTGGTGCTGGAC | - | MXAN_3946 | putative acyl carier protein |
| **8** | NRPS/PKS | | 4875470-4906976 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 79 | 135 | 4900530 | - | CTCGCCCGCAGGTTGCA | - | MXAN_4001 | non-ribosomal peptide synthase/polyketide synthase |
| | 75 | 470 | 4906209 | - | AGGGCACGGCCCGTGAT | - | MXAN_4002 | nonribosomal peptide synthetase |
| **9** | NRPS/PKS | | 4996865-5026357 | | | | | |
| **10** | NRPS/PKS | | 5252320-5298262 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 350 | 5253698 | + | GCTGGAGCGCGTGCTGCTGGCC | - | MXAN_4290 | putative thioesterase |
| | 77 | 352 | 5253696 | + | TGCTGGAGCGCGTGCTGCTGG | - | MXAN_4290 | putative thioesterase |
| | 76 | 354 | 5253694 | + | CTGGAGCGCGTGCTGCT | - | MXAN_4290 | putative thioesterase |
| | 76 | 132 | 5263568 | - | ATGTCGCCGTGCTTGCA | - | MXAN_4295 | patatin-like phospholipase family protein |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 77 | 150 | 5267893 | - | AGCTGGGAGTGGACCTGCTCA | - | MXAN_4296 | non-ribosomal peptide synthetase |
| | 77 | 379 | 5279177 | - | GATGCGTCGGCAGTTGCTGATG | - | MXAN_4298 | polyketide synthase type I |
| | 80 | 178 | 5287908 | - | ATGGCGCTCGAGTTGCG | - | MXAN_4299 | non-ribosomal peptide synthase/polyketide synthase |
| | 77 | 183 | 5292178 | - | CGCTGGCGAAGCGGCTGCTGT | - | MXAN_4300 | polyketide synthase type I |
| | 76 | 181 | 5292176 | - | CTGGCGAAGCGGCTGCT | - | MXAN_4300 | polyketide synthase type I |
| **11** | PKS | | 5407416-5440677 | | | | | |
| **12** | NRPS/PKS | | 5444807-5471033 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 182 | 5448069 | + | CCTGGACGACCGGTTGGTGCTG | - | MXAN_4413 | hypothetical protein |
| | 78 | 328 | 5468596 | + | CCTGGGCCACGGATTGCTGCGC | - | MXAN_4416 | cephalosporin hydroxylase family protein |
| | 79 | 332 | 5468592 | + | CTGGGCCACGGATTGCT | - | MXAN_4416 | cephalosporin hydroxylase family protein |
| **13** | NRPS/PKS | | 5600883-5653548 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 259 | 5612126 | - | ATGGCGCTGGAGCTGCG | - | MXAN_4526 | polyketide synthase type I |
| | 78 | 373 | 5612240 | - | ATGGCGCTGCTGTTGGA | - | MXAN_4526 | polyketide synthase type I |
| | 81 | 466 | 5612333 | - | GTGGCAGGGCAGGTGCG | - | MXAN_4526 | polyketide synthase type I |
| | 75 | 170 | 5627582 | - | CTGGCTGAACAACTGCA | - | MXAN_4527 | polyketide synthase |
| | 79 | 187 | 5653735 | - | GTGGCACAAGCTGCGCT | - | MXAN_4532 | non-ribosomal peptide synthase |
| **14** | NRPS | | 5735538-5780049 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 39 | 5755794 | + | CAATGGAACGGCGCATCCTCC | - | MXAN_4598 | non-ribosomal peptide synthase |
| | 75 | 41 | 5755792 | + | ATGGAACGGCGCATCCT | - | MXAN_4598 | non-ribosomal peptide synthase |
| | 75 | 485 | 5759989 | + | CCTGGGCTACGGGGTGCTGCGC | - | MXAN_4599 | M28 family peptidase |
| | 75 | 493 | 5765386 | - | GGTGGCCTGGGCGTCGGTGGCG | - | MXAN_4600 | radical SAM domain-containing protein |
| | 75 | 40 | 5764933 | - | GCGGTAAAGTCTTTGCT | - | MXAN_4600 | radical SAM domain-containing protein |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 78 | 237 | 5774788 | + | CCTGGCCCGTCTGGAGCTCCAG | - | MXAN_4602 | hypothetical protein |
| | 75 | 241 | 5774784 | + | CTGGCCCGTCTGGAGCT | - | MXAN_4602 | hypothetical protein |
| | 76 | 62 | 5777754 | + | ATGTCGCTCGACTTGCT | - | MXAN_4604 | hypothetical protein |

**NC_010162.1 Sorangium cellulosum 'So ce 56', complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | PKS | | 488146-499710 | | | | | |
| **2** | PKS | | 1150403-1167578 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 300 | 1150103 | + | CTGGCCTGCGCGTCGCT | - | sce0818 | putative dioxygenase |
| | 75 | 299 | 1151780 | + | ATGGCGCTGATCTTCCG | - | sce0819 | polyketide synthase |
| **3** | NRPS | | 3204860-3216562 | | | | | |
| **4** | PKS | | 4376667-4461851 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 84 | 488 | 4376179 | + | CTGGCAGGATATCTGCT | - | sce3188 | polyketide synthase |
| | 77 | 52 | 4440389 | + | CTGGGGCGGATCTCGCT | - | sce3193 | polyketide synthase |
| **5** | PKS | | 5761328-5853706 | | | | | |
| **6** | PKS | | 6867494-6856887 | | | | | |
| **7** | PKS | | 402480-9513516 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 81 | 172 | 9407529 | - | TCGGCACGCTGCCTGCA | - | sce6747 | hypothetical protein |
| | 90 | 68 | 9411437 | - | GTGGCGCGCCGCTTGCT | - | sce6751 | hypothetical protein |
| | 75 | 485 | 9425010 | - | GTGGCGCCGCGGCTGCC | - | sce6759 | hypothetical protein |
| | 75 | 4 | 9436354 | + | ATGGCAGGAGGGGTGAT | - | sce6766 | putative secreted protein |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 81 | 80 | 9440680 | - | CGGGCACGGGCGCTGCT | - | sce6770 | hypothetical protein |
| | 80 | 438 | 9441038 | - | AGGGCATGATCGTTGCC | - | sce6770 | hypothetical protein |
| | 77 | 38 | 9452036 | - | GCGGTACACCCCTTGCG | - | sce6780 | hypothetical protein |
| | 75 | 93 | 9452091 | - | GAGGCGCGGCGGATGCT | - | sce6780 | hypothetical protein |
| | 81 | 344 | 9452342 | - | CTGGCTCGAGGCCTGCA | - | sce6780 | hypothetical protein |
| | 75 | 212 | 9452057 | + | GCGGCCCGCCCGCTGCT | - | sce6781 | TetR family transcriptional regulator |
| | 77 | 479 | 9456589 | + | ACGGCACGCGACGTGCG | - | sce6785 | hypothetical protein |
| | 77 | 260 | 9460836 | + | CTGGCACCGGGCGAGCT | - | sce6787 | hypothetical protein |
| | 78 | 361 | 9466227 | - | CTGGCGCTCGATCTGCT | - | sce6792 | hypothetical protein |
| | 77 | 107 | 9470202 | - | TTCGCATGGCTGTTGGA | - | sce6797 | hypothetical protein |
| | 75 | 293 | 9479287 | - | ATGGTTCGTTTCTCGCT | - | sce6805 | hypothetical protein |
| | 77 | 345 | 9497084 | - | GTGGATCGAGTCATGCT | - | sce6819 | hypothetical protein |
| **8** | NRPS/ PKS | | 11424644-11529102 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 81 | 427 | 1.1E+07 | - | GGGGCACGATCGCTGCT | - | sce8207 | sigma-54 dependent transcriptional regulator |
| | 84 | 42 | 1.1E+07 | - | CTGGCATCCCTCGTGCT | - | sce8210 | putative DNA helicase |
| | 75 | 419 | 1.1E+07 | + | AGGGCGCGGCGTCTGCA | - | sce8233 | hypothetical protein |
| | 82 | 43 | 1.1E+07 | + | AGGGCACGGCTCTTGGA | - | sce8234 | hypothetical protein |
| | 76 | 173 | 1.1E+07 | + | GTGGCCGACGTCCTGCT | yfbL2 | sce8235 | putative aminopeptidases |
| | 75 | 388 | 1.1E+07 | + | GTGGCGCGCTCCCCGCT | - | sce8244 | putative protein phosphatase |
| | 75 | 371 | 1.1E+07 | - | GAGGCCCAGCTCGTGCA | - | sce8246 | CobW/P47K family protein |

**NC_002163.1 Campylobacter jejuni subsp. jejuni NCTC 11168, complete genome**

| | | |
|---|---|---|
| **1** | NRPS | 1236809-1238584 |

**NC_000913.2 Escherichia coli str. K-12 substr. MG1655, complete genome**

| | | |
|---|---|---|
| **1** | NRPS | 608682-631222 |

| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
|---|---|---|---|---|---|---|---|---|
| | 78 | 373 | 612792 | + | ATCGCACCGTGGTTGCC | ybdZ | b4511 | conserved protein |
| | 80 | 361 | 619783 | - | CTGGTACGCCGGATGCG | fepC | b0588 | iron-enterobactin transporter subunit |
| | 79 | 153 | 620564 | - | CCGTCACGCTACTTGCT | fepG | b0589 | iron-enterobactin transporter subunit |
| | 79 | 138 | 621385 | + | CGGGCACGGCAATGGCG | entS | b0591 | enterobactin exporter, iron-regulated |
| | 76 | 428 | 624161 | - | TGGGCGCAAGTGTTGCC | fepB | b0592 | iron-enterobactin transporter subunit |
| | 79 | 291 | 625002 | + | CTGGCCTGTCTGCTGCA | entE | b0594 | 2,3-dihydroxybenzoate-AMP ligase component of |

**NC_002942.5 Legionella pneumophila subsp. pneumophila str. Philadelphia 1, complete genome**

| | | | |
|---|---|---|---|
| **1** | NRPS/ PKS | 2166115-2171151 | |
| **2** | NRPS/ PKS | 2442757-2470485 | |

**NC_002516.2 Pseudomonas aeruginosa PAO1, complete genome**

| | | | |
|---|---|---|---|
| **1** | NRPS | 2532669-2549458 | |

| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
|---|---|---|---|---|---|---|---|---|
| | 90 | 276 | 2539319 | - | CTGGCCCGGGCCTTGCT | - | PA2302 | non-ribosomal peptide synthetase |
| | 79 | 154 | 2545100 | - | ATGGCGCTGCTGTTGGT | - | PA2305 | non-ribosomal peptide synthetase |
| | 83 | 8 | 2545667 | - | ATGGAACGAATGTCGCT | - | PA2306 | hypothetical protein |
| | 84 | 55 | 2547555 | - | CTGGCAGGACTATTTCT | - | PA2308 | ABC transporter ATP-binding protein |
| **2** | NRPS | | 2636517-2687178 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 125 | 2640910 | + | ACGGCACTGTCCTTCCA | fpvR | PA2388 | FpvR |
| | 80 | 88 | 2651219 | + | CTGGCGCCGAGCCTGCT | pvdO | PA2395 | PvdO |
| | 75 | 140 | 2653197 | - | CTTCCACGTAAATTGCA | pvdF | PA2396 | pyoverdine synthetase F |
| | 77 | 421 | 2665565 | - | CTGGCGCGGATGCCGCT | pvdD | PA2399 | pyoverdine synthetase D |
| | 77 | 479 | 2672108 | - | CTGGCGCGGATGCCGCT | pvdJ | PA2400 | PvdJ |

| 3 | NRPS | | | 4724639-4746551 | | | | |
|---|---|---|---|---|---|---|---|---|
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 81 | 393 | 4727194 | - | CTGGCTCGCGGGCTGCT | fptA | PA4221 | Fe(III)-pyochelin outer membrane receptor precursor |
| | 78 | 265 | 4728882 | - | CCGGCGAGCCTGTTGCT | - | PA4222 | ABC transporter ATP-binding protein |
| | 80 | 286 | 4728903 | - | CTGGCCCGTGCGCTGCT | - | PA4222 | ABC transporter ATP-binding protein |
| | 75 | 43 | 4731415 | - | CTGGCGCAATCCTTGTC | pchG | PA4224 | pyochelin biosynthetic protein PchG |
| | 75 | 16 | 4736814 | - | CTGGAAGAGGGCGTGCT | pchF | PA4225 | pyochelin synthetase |
| | 81 | 64 | 4736862 | - | CTTGCCCGCCACTTGCA | pchF | PA4225 | pyochelin synthetase |
| | 77 | 465 | 4741576 | - | CTGGTCGGCGCCTTGCA | pchE | PA4226 | dihydroaeruginoic acid synthetase |

**NC_002947.3 Pseudomonas putida KT2440, complete genome**

| 1 | NRPS | | | 4767855-4798662 | | | | |
|---|---|---|---|---|---|---|---|---|
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 79 | 278 | 4795256 | - | TTGGCGCCGAAGGTGCA | - | PP_4222 | syrP protein, putative |
| | 77 | 431 | 4795409 | - | ATGGGGCAGCGGTTGCG | - | PP_4222 | syrP protein, putative |
| | 80 | 413 | 4796855 | - | ATGGGCCAGGCATTGCT | - | PP_4223 | diaminobutyrate--2-oxoglutarate aminotransferase |
| **2** | NRPS | | | 4818235-4832793 | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 122 | 4832915 | - | TTGGCCAGTGGCCTGCT | - | PP_4245 | siderophore biosynthesis protein, putative |

**NC_004578.1 Pseudomonas syringae pv. tomato str. DC3000, complete genome**

| 1 | NRPS | | | 2307009-2356936 | | | | |
|---|---|---|---|---|---|---|---|---|
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 81 | 315 | 2306694 | + | ATGGAATTCACCTTGCT | - | PSPTO_2134 | pyoverdine synthetase, thioesterase component |
| | 76 | 381 | 2322147 | + | CTGGTCGGTCATTTGCA | - | PSPTO_2137 | MbtH-like protein |
| | 78 | 274 | 2324978 | - | CCGGCGCCGAAATTGCT | - | PSPTO_2139 | cation ABC transporter, permease protein |
| | 80 | 491 | 2327813 | - | CTGGCACAAGTGGCGCT | - | PSPTO_2143 | hypothetical protein |
| | 75 | 384 | 2330265 | + | CTGGAAAACGACGTGCT | - | PSPTO_2147 | pyoverdine sidechain peptide synthetase I, epsilon-Lys module |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 77 | 389 | 2333571 | + | CCGACACACTTGTTGCT | - | PSPTO_2148 | pyoverdine sidechain peptide synthetase II, D-Asp-L-Thr |
| **2** | NRPS/ PKS | | 2863432-2890634 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 322 | 2868091 | - | TTGGCGCGGGTGCCGCA | - | PSPTO_2593 | multidrug resistance protein, AcrA/AcrE family |
| | 79 | 78 | 2872757 | - | GTGGCCAGGCTCTGGCA | irp4 | PSPTO_2598 | yersiniabactin synthetase, thioesterase component |
| | 81 | 235 | 2873968 | - | CTGGCGCAACACTCGCA | irp3 | PSPTO_2599 | yersiniabactin synthetase, thiazolinyl reductase component |
| | 80 | 112 | 2884576 | - | CTGGCGCGGCACTTGAG | - | PSPTO_2601 | hypothetical protein |
| | 78 | 384 | 2884848 | - | CTGGCGCTGACGTGGCT | - | PSPTO_2601 | hypothetical protein |
| | 75 | 89 | 2890723 | - | CAGGCTCGGAACTGGCA | - | PSPTO_2602 | yersiniabactin non-ribosomal peptide synthetase |
| **3** | NRPS | | 3151815-3185293 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 472 | 3152407 | + | TTGATACGGCTGATGCT | syfA | PSPTO_2829 | non-ribosomal peptide synthetase SyfA |
| **4** | NRPS | | 5087432-5106501 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 76 | 440 | 5088005 | + | CTGGCAGCCTGGTTCCT | - | PSPTO_4518 | non-ribosomal peptide synthetase, initiating |
| **5** | PKS | | 5282597-5304804 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 75 | 23 | 5284944 | + | CTGGTCCGCAAGTCGCT | cfa3 | PSPTO_4683 | coronafacic acid beta-ketoacyl synthetase component |
| | 75 | 194 | 5286445 | + | CTGCCAGGCCAACTGCT | cfa5 | PSPTO_4685 | coronafacic acid synthetase, ligase component |
| **6** | NRPS | | 5313602-5327197 | | | | | |

**NC_003198.1 Salmonella enterica subsp. enterica serovar Typhi str. CT18, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | NRPS | | 624926-644794 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 82 | 483 | 640548 | - | CTGGCGCAATTTCTGCT | fepB | STY0638 | iron-enterobactin transporter periplasmic binding protein |
| | 79 | 288 | 641267 | + | CTGGCCTGTCTGCTGCA | entE | STY0640 | enterobactin synthase subunit E |

**NC_004347.1 Shewanella oneidensis MR-1, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | PKS | | 1670313-1686159 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 78 | 441 | 1676645 | - | ATAGCATGAGTATTGCC | - | SO_1599 | beta keto-acyl synthase |

**NC_003143.1 Yersinia pestis CO92, complete genome**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | NRPS | | 839849-856936 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 79 | 186 | 841919 | - | CGGGCACTATTGCTGCT | - | YPO0774 | hypothetical protein |
| | 81 | 353 | 857289 | - | ATGGCCCCCCGCCTGCT | - | YPO0778 | putative siderophore biosysnthesis protein |
| **2** | NRPS/ PKS | | 2140840-2169669 | | | | | |
| | Score | Distance From | N | Strand | Sequence | Gene | Synonym | Protein Name |
| | 82 | 239 | 2145615 | - | CTGGCGCGACTGCTGCG | irp4 | YPO1908 | yersiniabactin biosynthetic protein YbtT |
| | 75 | 486 | 2166571 | + | ATGACACGCTGCTGGCG | irp8 | YPO1915 | putative signal transducer |
| | 83 | 441 | 2167924 | + | CTGGCCGGCGGATTGCT | ybtS | YPO1916 | salicylate synthase Irp9 |
| | 79 | 96 | 2168269 | + | GGGGCATTCACGTTGCT | ybtS | YPO1916 | salicylate synthase Irp9 |

# Appendix 7: Sigma-54 Promoter Database Screenshots

## $\sigma^{54}$ Promoter Database

View by: COGs
Data Grouped by: COG Data  COG Category Data  NRPS/PKS Statistics

| Genome Name | RefSeq ID | Phylum | Matrix | Contains Sigma 54 | % GC Content | Completion Status | Score Data | COG Data | COG Category Data | NRPS/PKS Stats |
|---|---|---|---|---|---|---|---|---|---|---|
| Acinetobacter baumannii AB0057, complete genome | NC_011586.1 | Proteobacteria (Gamma) | RPON.Matrix | Y | 39.2 | 12 | View | View | View | View |
| Acinetobacter baumannii AB307-0294, complete genome | NC_011595.1 | Proteobacteria (Gamma) | RPON.Matrix | Y | 39.0 | 12 | View | View | View | View |
| Acinetobacter baumannii AYE, complete genome | NC_010410.1 | Proteobacteria (Gamma) | RPON.Matrix | Y | 39.4 | 12 | View | View | View | View |
| Acinetobacter baumannii SDF, complete genome | NC_010400.1 | Proteobacteria (Gamma) | RPON.Matrix | Y | 39.2 | 12 | View | View | View | View |
| Agrobacterium tumefaciens str. C58 chromosome circular, complete sequence | NC_003062.2 | Proteobacteria (Alpha) | RPON.Matrix | Y | 59.4 | 12 | View | View | View | View |
| Agrobacterium tumefaciens str. C58 chromosome linear, complete sequence | NC_003063.2 | Proteobacteria (Alpha) | RPON.Matrix | Y | 59.3 | 12 | View | View | View | View |
| Alcanivorax borkumensis SK2 chromosome, complete genome | NC_008260.1 | Proteobacteria (Gamma) | RPON.Matrix | Y | 54.7 | 12 | View | View | View | View |
| Bacillus anthracis str. Ames, complete genome | NC_003997.3 | Firmicutes | RPON.Matrix | Y | 35.4 | 12 | View | View | View | View |
| Bacillus cereus subsp. cytotoxis NVH 391-98, complete genome | NC_009674.1 | Firmicutes | RPON.Matrix | Y | 35.9 | 12 | View | View | View | View |
| Bacillus subtilis subsp. subtilis str. 168, complete genome | NC_000964.3 | Firmicutes | RPON.Matrix | Y | 43.5 | 12 | View | View | View | View |
| Bacteroides fragilis NCTC 9343, complete genome | NC_003228.3 | Bacteroidetes/Chlorobi | RPON.Matrix | Y | 43.2 | 12 | View | View | View | View |
| Bifidobacterium longum NCC2705, complete genome | NC_004307.2 | Actinobacteria | RPON.Matrix | N | 60.1 | 12 | View | View | View | View |
| Borrelia burgdorferi B31, complete genome | NC_001318.1 | Spirochaetes | RPON.Matrix | Y | 28.6 | 12 | View | View | View | View |
| Campylobacter jejuni subsp. jejuni NCTC 11168, complete genome | NC_002163.1 | Proteobacteria (Epsilon) | RPON.Matrix | Y | 30.5 | 12 | View | View | View | View |
| Candidatus Amoebophilus asiaticus 5a2, complete genome | NC_010830.1 | Bacteroidetes/Chlorobi | RPON.Matrix | Y | 35.0 | 12 | View | View | View | View |
| Candidatus Carsonella ruddii PV | | | | | | | | | | |

**Figure S - 3. Screenshot of the main page of the Sigma-54 Promoter Database.**

# Appendix 8: ClusterMine360 Screenshots

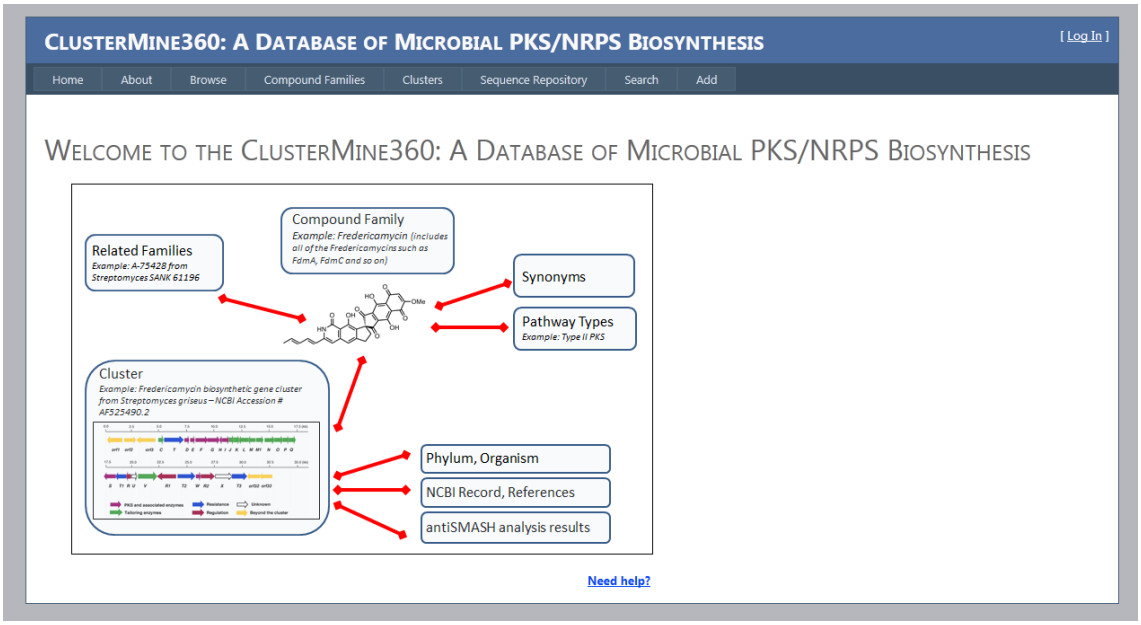



**Figure S - 4. Screenshot of ClusterMine360's home page.**

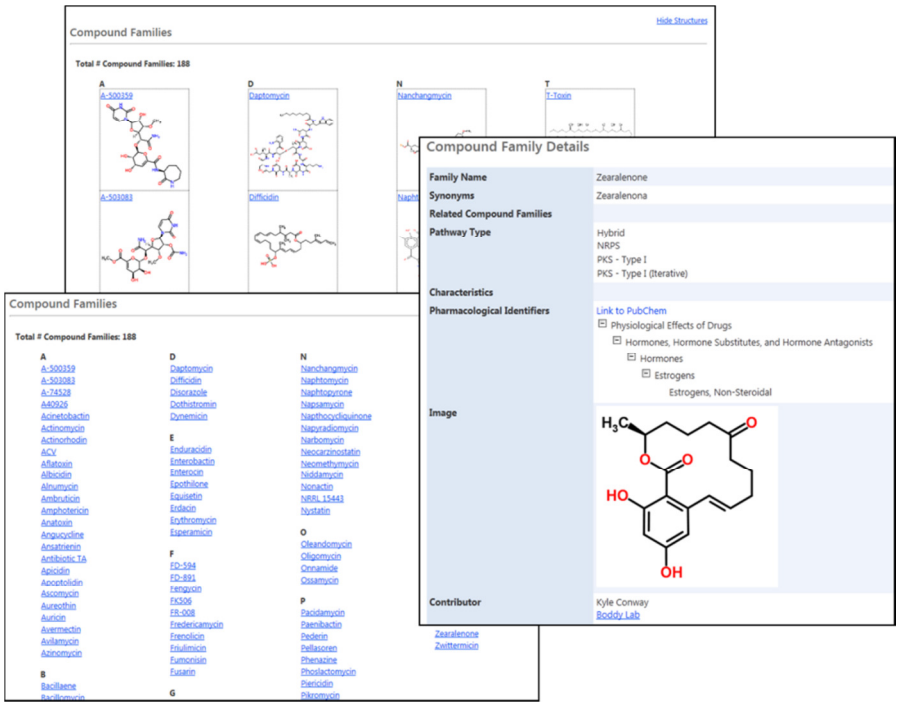



**Figure S - 5. Composite screenshot of the page listing compound families (top and bottom left) and the details page for a compound family (right).**

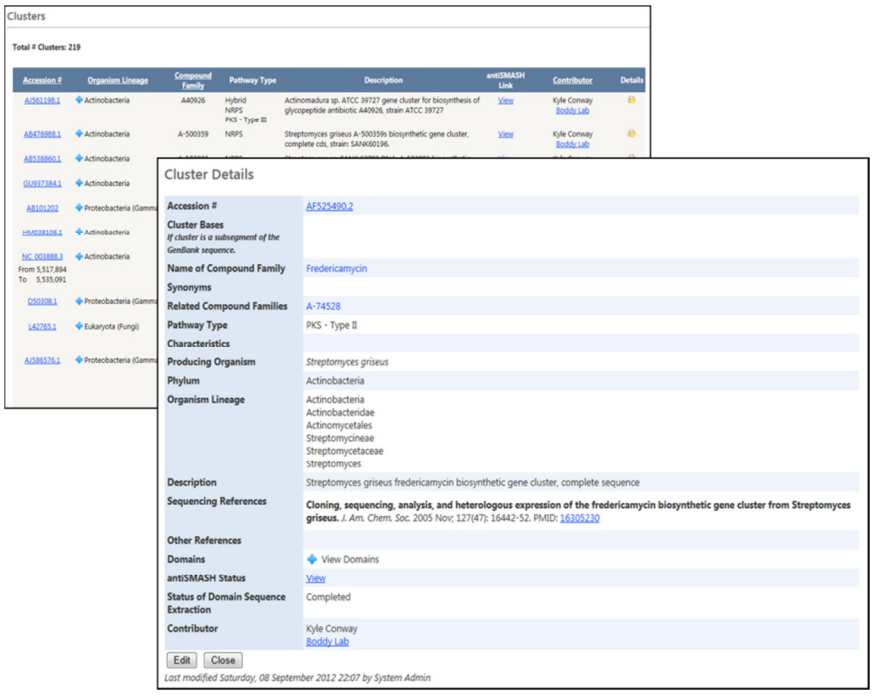



**Figure S - 6. Screenshots of the page listing clusters (top left) and a cluster details page (bottom right).**

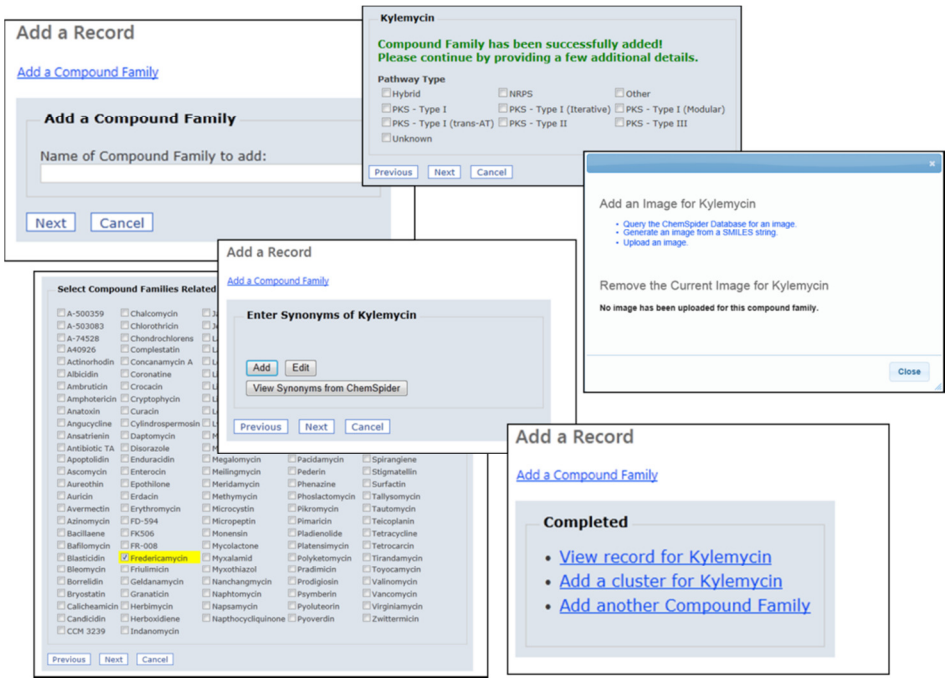



**Figure S - 7. Composite screenshot of steps involved in adding a compound family.**

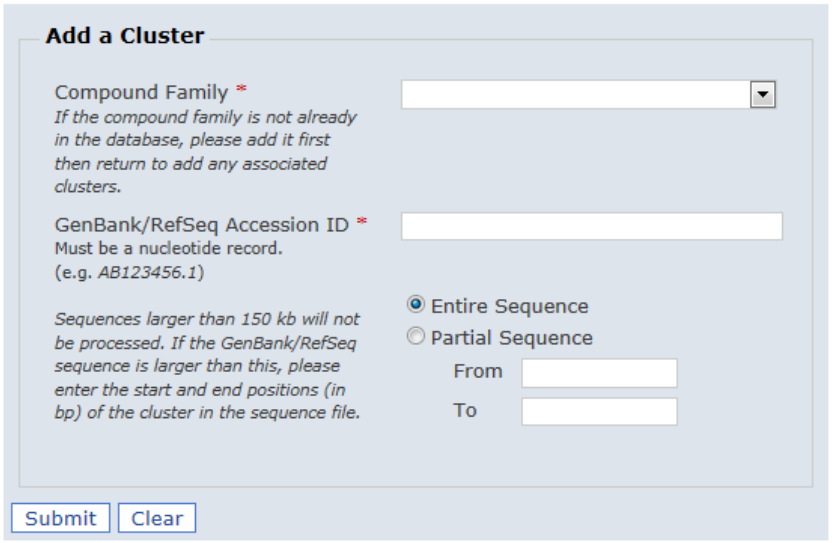



**Figure S - 8. Screenshot of form used for submitting clusters to ClusterMine360.**

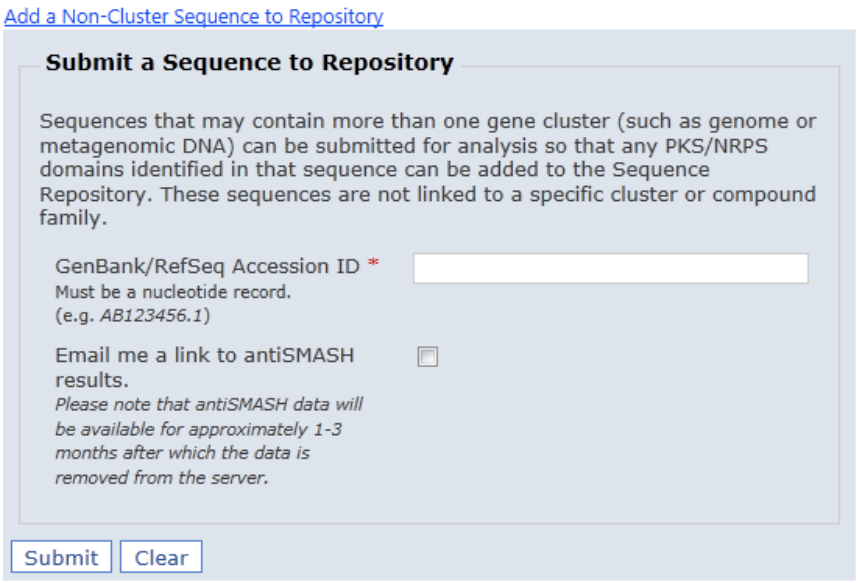



**Figure S - 9. Screenshot of form used to submit a sequence to the repository.**

# Appendix 9: PromScan Perl Script

```
#!/usr/bin/perl -w

################################################################################
#                                                              #
#   copyright       : (C) 2000-2003 by David J. Studholme            #
#   email           : ds2@sanger.ac.uk                          #
#                                                              #
################################################################################

################################################################################
#*                                                              #
#*   This program is free software; you can redistribute it and/or modify   #
#*   it under the terms of the GNU General Public License as published by   #
#*   the Free Software Foundation; either version 2 of the License, or      #
#*   (at your option) any later version.                                    #
#*                                                              #
################################################################################

### VERSION / UPDATE LOG ###

# 7/11/02 Corrected bugs caught by Richard Pau
# 21/11/02 Changed to log base 2 rather than natural log
# 07/03 Enabled reporting of distance upstream
# enabled normalising of scores.
# 27/08/03 Switched format of matrix files.
# 27/08/03 Integrated Hit object into same file.

use strict;
use Getopt::Std;
use IO::File;
use IO::Handle;


$| = 1;

### Parse command line arguments
### -m specify matrix file, -g specify genome name, -c specify cut-off
### -h: help, -:statistics, -n: ptt, -i unknown , -m: matrix, -g: genome -c:cutoff
my %option;
getopts( "hsnim:g:c:", \%option );

if ( $option{h} ) {
        show_help();
        exit;
}

my ($seqread);
my ( $seq_length, @seq_window );
my ( $mean_score, $total_dev, $sd, $variance );
my $GC_content;

### Set threshold score above which counts as a hit
my $cut_off = 80;
if ( $option{c} ) {
        $cut_off = $option{c};
}

### Set input/output filenames
my $genome = "ecoli";
```

```
if ( $option{g} ) {
	$genome = $option{g};
}

### Remove any filename extensions
if ( $genome =~ m/([\w\d\.]+)\.fna/ ) {
	$genome = $1;
}

### Name of file containing sequence to be scanned
my $sequence_file = "$genome.fna";

### Name of file containing details of ORFs
my @ptt_file;
unless ( $option{n} ) {
	my $ptt_file = "$genome.ptt";
	my $ptt_fh   = new IO::File;
	$ptt_fh->open("<$ptt_file")
	  or die "Failed to open ptt file $genome.ptt: $!\n";
	while ( my $readline = <$ptt_fh> ) {
		push @ptt_file, $readline;
	}
	$ptt_fh->close;
}

### Name of matrix file
my $matrix_file = $option{m} or die "You must specify a matrix file\n";

### Open error file

open OUTPUT, '>', $option{g}.".txt" or die $!;
STDOUT->fdopen( \*OUTPUT, 'w' ) or die $!;


### Begin the scans!!
promscan_kl( $matrix_file, $sequence_file, \@ptt_file );
exit;

sub promscan_kl {

	my ( $matrix_file, $sequence_file, $ptt_file ) = @_;

	my %all_scores_forward;
	my %all_scores_reverse;
	my @hits;

	my ( $sequence, $header );
	print STDERR "Reading the sequence file into memory ... ";
	my $seq_fh = new IO::File;
	$seq_fh->open("<$sequence_file")
	  or die "Failed to open $sequence_file: $!\n";

	while ( my $readline = <$seq_fh> ) {
		chomp $readline;
		$sequence .= $readline if $readline =~ m/^[A-Za-z]+$/;
		if ( $readline =~ m/>(.*)/ ) {
			$header = $1;
		}
	}
	$seq_fh->close;

	print STDERR "done.\n";
```

```
        print STDERR "Sequence: $header\n\n";
        print "Sequence|$header\n";

        print STDERR "Splitting the sequence into a list ... ";
        my @sequence = split //, $sequence;
        print STDERR "done.\n";

        ### Calculate the GC content of the sequence
        my $GC_content = calculate_GC(@sequence);
        print "G+C|$GC_content\n";
        print STDERR "G+C = $GC_content\n";
        my %content;
        $content{G} = $GC_content / 2;
        $content{C} = $GC_content / 2;
        $content{A} = ( 1 - $GC_content ) / 2;
        $content{T} = ( 1 - $GC_content ) / 2;

        ### Read matrix file
        my $pmatrix = read_matrix_file($matrix_file);
        print "Matrix|".$option{m}."\n";
        my %matrix  = %$pmatrix;


        ### Do sanity check on matrix, and calculate width of the matrix
        die "Matrix corrupted\n"
          unless ( @{ $matrix{A} } == @{ $matrix{C} }
                and @{ $matrix{C} } == @{ $matrix{G} }
                and @{ $matrix{A} } == @{ $matrix{T} } );
        my $seq_window_length = @{ $matrix{A} };

        ### Calculate maximum possible score
        my $overall_highest = 0;

        for ( my $i = 1 ; $i < $seq_window_length ; $i++ ) {
                my $highest_score = 0;
                my $highest_base;
                foreach my $base ( "A", "C", "G", "T" ) {
                        if ( $matrix{$base}[$i] > $highest_score ) {
                                $highest_score = $matrix{$base}[$i];
                                $highest_base  = $base;
                        }
                }
                $overall_highest +=
                  ( $highest_score * log( $highest_score / $content{$highest_base} ) /
                          log(2) )
                  if $highest_score;
        }

#print STDERR "(Highest possible score before normalisation = $overall_highest)\n" ;
#print "(Highest possible score before normalisation = $overall_highest)\n" ;

        print STDERR "Sites scoring > $cut_off\n";
###     print "Sites scoring > $cut_off\n";

        ### Initialise the first window of sequence to be scanned
        for ( my $i = 0 ; $i < $seq_window_length ; $i++ ) {
                $seq_window[$i] = $sequence[$i];
        }

        ### read sequence one base at a time
        my $n = 0;
```

```
while ( my $seqread = uc( shift @sequence ) ) {
        my $seq_window;
        if ( $seqread =~ m/[ACGTN]/ ) {    # sanity-check for legal bases

                ### move the sequence window along one residue
                shift(@seq_window);          # removes first residue
                push( @seq_window, $seqread ); # adds next residue

                ### Assign a score to current window
                ### This score, known as the Kullback-Liebler distance,
                ### reflects the theoretical binding energy of the
                ### DNA-protein interaction.
                ### For more information see: Stormo, G.D. 2000.
                ### DNA binding sites: representation and discovery.
                ### Bioinformatics. 16:16-23.

                ### First check the forward strand
                my $strand = "+";

                ### Calculate score. Remember, divide by log2 to get log to the base of 2
                my $current_score = 0;
                for ( my $i = 0 ; $i < $seq_window_length ; $i++ ) {
                        my $base = $seq_window[$i];
                        $current_score = $current_score + ( ( $matrix{$base}[$i] ) *
                                log( $matrix{$base}[$i] / $content{$base} ) / log(2) )
                         if $matrix{$base}[$i];
                }

                ### Normalise the score
                $current_score /= $overall_highest;
                $current_score *= 100;

                if ( $option{s} ) {
                        ### Record each score in a matrix %all_scores_forward
                        $all_scores_forward{$n} = $current_score;
                }

                ### If this window scores greater than the cut off, then record details ....
                if ( $current_score > $cut_off ) {

                        ### Convert @seq_window into a string, $seq_window
                        $seq_window = "";
                        my $q = 0;
                        while ( $seq_window[$q] ) {
                                $seq_window .= $seq_window[$q];
                                $q++;
                        }

                        ### Is hit in coding or non-coding region?
                        my ( $locate_match, $intergenic_size ) =
                          locate_match( $ptt_file, $n, $strand, $sequence,
                                $seq_window );

                        ### Round score to nearest whole number
                        $current_score = int($current_score);

                        ### Print details of this window
                        if ( $locate_match or $option{n} ) {

                                ### Create a new Hit object and set its values
                                my $hit = Hit->new();
                                $hit->position($n);
```

```
                        $hit->strand($strand);
                        $hit->score($current_score);
                        $hit->strand($strand);
                        $hit->sequence($seq_window);
                        $hit->locus($locate_match);

                        ### Record the existence of this Hit object in an array
                        push @hits, $hit;

                        printf STDERR "%8s|%1s|%s|%5d|%s\n", $n, $strand,
                          $seq_window, $current_score, $locate_match;
                }
}

### now check the Reverse strand
$strand = "-";

### Calculate score. Remember divide by log2 to get log to the base of 2
$current_score = 0;
for ( my $i = 0 ; $i < $seq_window_length ; $i++ ) {
                my $j    = $seq_window_length - $i - 1;
                my $base = revcomp( $seq_window[$j] );
                $current_score = $current_score + ( ( $matrix{$base}[$i] ) *
                                log( $matrix{$base}[$i] / $content{$base} ) / log(2) )
                 if $matrix{$base}[$i];
}

### Normalise the score
$current_score /= $overall_highest;
$current_score *= 100;

if ( $option{s} ) {
                ### Record each score in a matrix %all_scores_forward
                $all_scores_reverse{$n} = $current_score;
}

#### print current score if greater than threshold value
if ( $current_score > $cut_off ) {

                ### Convert @seq_window into a string, $seq_window
                $seq_window = "";
                my $q = 0;
                while ( $seq_window[$q] ) {
                        $seq_window = $seq_window . $seq_window[$q];
                        $q++;
                }

                ### Convert $seq_window to Reverse Complement
                my $revcomp = reverse($seq_window);
                $revcomp =~ tr/ACGTacgt/TGCAtgca/;
                $seq_window = $revcomp;

                ### Is hit in coding or non-coding region?
                my ( $locate_match, $intergenic_size ) =
                  locate_match( $ptt_file, $n, $strand, $sequence,
                        $seq_window );

                ### Round the score to nearest whole number
                $current_score = int($current_score);

                ### Print details of this window
                ### ... but only if it is intergenic
```

```
                                        if ( $locate_match or $option{n} ) {

                                                ### Create a new Hit object and set its values
                                                my $hit = Hit->new();
                                                $hit->position($n);
                                                $hit->strand($strand);
                                                $hit->score($current_score);
                                                $hit->strand($strand);
                                                $hit->sequence($seq_window);
                                                $hit->locus($locate_match);

                                                ### Record the existence of this Hit object in an array
                                                push @hits, $hit;

                                                printf STDERR "%8s|%1s|%s|%5d|%s\n", $n, $strand,
                                                  $seq_window, $current_score, $locate_match;
                                        }
                                }

                                ### The sequence pointer proceeds by one base
                                $n++;

                        }
                        ;    # end of if($seqread =~ /[ACGTN]/)

};    # end of while (read $sequence_file, $seqread, 1) loop

### The final value of $n, the sequence-pointer,
### gives the number of readable bases
$seq_length = $n;
### print "N = $seq_length\n";

###  Create a hash of arrays, to index Hit objects by score
###  Cannot use a simple hash because multiple Hits may have
###   the same score.
my %hitscores;
foreach my $hit (@hits) {
        my $score = $hit->score();
        unless ( $hitscores{$score} ) {
                $hitscores{$score} = [];
        }
        my $ref  = $hitscores{$score};
        my @list = @$ref;
        push @list, $hit;
        @$ref = @list;
}

### Sort the Hit objects into descending order of scores
my @unordered_scores = keys(%hitscores);
my @ordered_scores = sort { $b <=> $a } @unordered_scores;
my @ordered_hits;
foreach my $score (@ordered_scores) {
        my $ref  = $hitscores{$score};
        my @hits = @$ref;
        foreach my $hit (@hits) {
                push @ordered_hits, $hit;
        }
}
### Print out number of hits
print "Num of Hits|".scalar(@ordered_hits)."\n";
```

```
	### Now print out the hits ...
	foreach my $hit (@ordered_hits) {
		my ( $pos, $str, $seq, $score, $locus ) = (
			$hit->position(), $hit->strand(),
			$hit->sequence(), $hit->score(),
			$hit->locus()
		);
		printf "%8s|%1s|%s|%5d|%s\n", $pos, $str, $seq, $score, $locus;
	}

	### If we have opted to calculate statistics ...
	if ( $option{s} ) {

		### Calculate mean score for all possible windows in the
		### query sequence
		my $total_score = 0;
		for ( my $i = 0 ; $i < $seq_length ; $i++ ) {
			$total_score =
			  $total_score + $all_scores_forward{$i} + $all_scores_reverse{$i};
		}

		$mean_score = $total_score / ( $seq_length * 2 );
		print "Mean score = $mean_score\n";

		### Calculate variance and sd of scores for all
		### possible windows in the query sequence
		$total_dev = 0;
		for ( my $i = 0 ; $i < $seq_length ; $i++ ) {
			$total_dev =
			  $total_dev + ( ( $all_scores_forward{$i} - $mean_score )**2 ) +
			  ( ( $all_scores_reverse{$i} - $mean_score )**2 );
		}

		$variance = $total_dev / ( $seq_length * 2 );
		print "Variance = ", $variance, "\n\n\n";
		$sd = $variance**(0.5);
	}
	### print time - $^T;

	$seq_fh->close;

}   # end of promscan_kl()

sub locate_match {
	my $ptt_file        = shift or die "Failed to pass a PTT file";
	my $promoter_position = shift or die "Failed to pass a position\n";
	my $promoter_strand   = shift or die "Failed to pass a strand\n";
	my $sequence        = shift or die "Failed to pass a sequence\n";
	my $seq_window      = shift
	  or die
"Filed to pass a seq_window ($ptt_file,$promoter_position,$promoter_strand\n";

	unless ( defined $option{n} ) {

		my ( $start_pos1, $end_pos1, $strand1, $product1 ) =
		  ( 0, 0, "?", "ORIGIN" );
		my ( $start_pos2, $end_pos2, $strand2, $product2 ) =
		  ( 0, 0, "?", "ORIGIN" );

		my $promoter_description = "";
		### $promoter_description = "not found";
		foreach my $read_line (@$ptt_file) {
```

```
# Read each window of the protein table file
my @fields = read_ptt_line($read_line);

unless ( $fields[0] eq "failed" ) {

        ( $start_pos1, $end_pos1, $strand1, $product1 ) =
          ( $start_pos2, $end_pos2, $strand2, $product2 );
        ( $start_pos2, $end_pos2, $strand2, $product2 ) = @fields;

        my $size_of_intergenic_region = 0;
        my $upstream_sequence       = "";

        ### Check wether promoter falls within this window
        if (   $promoter_position > $end_pos1
                && $promoter_position < $start_pos2 )
        {

                ###  promoter is in intergenic region...
                my $size_of_intergenic_region_temp =
                  $start_pos2 - $end_pos1;

                if ( $promoter_strand eq "+" && $strand2 eq "+" ) {
                        $size_of_intergenic_region =
                          $size_of_intergenic_region_temp;
                        $upstream_sequence = substr(
                                $sequence,
                                ( $end_pos1 - 50 ),
                                ( $size_of_intergenic_region + 49 )
                        );

                        my $spacer = $start_pos2 - $promoter_position -
                          length($seq_window);
                        $promoter_description =
                          "-N($spacer)- $product2|$size_of_intergenic_region";

                }

                elsif ( $promoter_strand eq "-" && $strand1 eq "-" ) {
                        $size_of_intergenic_region =
                          $size_of_intergenic_region_temp;
                        my $seq_tmp = substr(
                                $sequence,
                                ( $end_pos1 - 2 ),
                                ( $size_of_intergenic_region + 52 )
                          )
                          or die
"Failed to get seq_tmp! end_pos1=$end_pos1, size_of_intergenic_region=$size_of_intergenic_region\n";

                        my $spacer = $promoter_position - $end_pos1;
                        $promoter_description =
                          "-N($spacer)- $product1|$size_of_intergenic_region";

                }

                elsif ( $promoter_strand eq "-" && $strand1 eq "+" ) {
                        $size_of_intergenic_region =
                          $size_of_intergenic_region_temp;
                        my $seq_tmp = substr(
                                $sequence,
                                ( $end_pos1 - 2 ),
                                ( $size_of_intergenic_region + 52 )
```

```
                                                                )
                                                                or die
"Failed to get seq_tmp! end_pos1=$end_pos1, size_of_intergenic_region=$size_of_intergenic_region\n";

                                                        my $spacer = $promoter_position - $end_pos1;
                                                        $promoter_description =
"-N($spacer)- $product1 (WRONG ORIENTATION)|$size_of_intergenic_region";

                                                }

                                                elsif ( $promoter_strand eq "+" && $strand1 eq "-" ) {
                                                        $size_of_intergenic_region =
                                                         $size_of_intergenic_region_temp;
                                                        my $seq_tmp = substr(
                                                                $sequence,
                                                                ( $end_pos1 - 2 ),
                                                                ( $size_of_intergenic_region + 52 )
                                                         )
                                                         or die
"Failed to get seq_tmp! end_pos1=$end_pos1, size_of_intergenic_region=$size_of_intergenic_region\n";

                                                        my $spacer = $promoter_position - $end_pos1;
                                                        $promoter_description =
"-N($spacer)- $product1 (WRONG ORIENTATION)|$size_of_intergenic_region";

                                                }
                                        }

                                        elsif ( ( $promoter_position >= $start_pos1 )
                                                and ( $promoter_position <= $end_pos1 )
                                                and ( $option{i} ) )
                                        {
                                                $promoter_description = "internal to $product1";
                                        }
                                }

                        }
                        my $size_of_intergenic_region = $start_pos2 - $end_pos1;
                        return ( $promoter_description, $size_of_intergenic_region );

                }
}   # End of sub locate_match

sub read_ptt_line {

        # Expects a scalar value containing one line of a .ptt file
        # Extracts the 'location', strand, and 'product'
        #   from that line and returns them, in that order

        my $read_line = $_[0];
        my $i;
        my $product;
        my @fields;

        if ( $read_line =~ /(\d+)\.\.(\d+)\s+([+-])/ ) {

                # Extract 'location' and 'strand'
                @fields  = split( /\s+/, $read_line );
                $i       = 5;
                $product = "";
                while ( $fields[$i] ) {
```

```
                                        # Extract 'product'
                                        $product = $product . " " . $fields[$i];
                                        $i++;
                                }
                                return ( $1, $2, $3, $product );
                        }
                        ;   # End of 'if( $read_line =~ ....' block

                        ### print ("variable read_line:$read_line\n");
                        return ("failed");

};   # end of sub read_ptt_line;

sub calculate_GC {
                        my (@sequence) = @_;
                        my ( $n, $GC, $AT ) = ( 0, 0, 0 );    # initialise counters
                        my $seqread;

                        foreach my $seqread (@sequence) {
                                        $GC++ if $seqread =~ /[GCgc]/;
                                        $AT++ if $seqread =~ /[ATat]/;
                        }
                        $n = $GC + $AT;
                        my $GC_content = $GC / $n;
                        return $GC_content;
}

sub revcomp {
                        my $string = shift @_ or print STDERR "Called revcomp on null string!";
                        my $seq    = "";
                        my @array  = split( //, $string );
                        while ( my $char = pop @array ) {
                                        $char =~ tr/ACGT/TGCA/;
                                        $seq = $seq . $char;
                        }

                        #print STDERR "Revcomp of $string -> $seq\n" ;
                        return $seq;
}

sub read_matrix_file {

                        my $matrix_file = shift;
                        my %matrix;

                        my $fh = new IO::File;
                        $fh->open("<$matrix_file")
                          and print STDERR "Opened matrix file: $matrix_file\n"
                          or die "Could not open Configuration file $matrix_file: $!\n";

                        ### print "\nScoring matrix:\n\n";

                        while ( my $readline = <$fh> ) {
                                        chomp($readline);

                                        if ( $readline =~ m/([ACGT]):([\s\d]+)/ ) {
                                        ###             print "$readline\n";
                                                        print STDERR "$readline\n";

                                                        my $base   = $1;
                                                        my @values = split /\s+/, $2;
```

```
                                $matrix{$base} = [];
                                foreach my $value (@values) {
                                        push @{ $matrix{$base} }, $value;
                                }
                        }
                }

###             print "\n\n";
                print STDERR "\n\n";

                $fh->close;
                print STDERR "Closed matrix file\n";

                return \%matrix;
}

sub show_help {
                print STDERR <<EOF;


=============================================================================================================
$0: scan a DNA sequence against a motif specified as a scoring matrix
=============================================================================================================

Usage: $0 <options>

   Available options

      -h          : Show this help.
                -l          : Use linear scoring formula ignoring G+C content,
                   rather than the Kullback-Leibler distance.
      -s          : Calculate statistics (this requires a lot of memory!)
      -i          : Show matches that are internal to ORFs as well as intergenic ones.
      -n          : Do not use a .ptt file (ie no ORF information)
                -g <file>     : Specify the input file (genome)
                   You can specify the .ppt file or the .fna file
                   or just the base name:
                   eg. $0 -g NC_000962.ptt
                      $0 -g NC_000962.fna
                      $0 -g NC_000962
                -m <file>     : Specify the matrix file.
                -c <number>   : Specify the threshold (cut-off) score (default is 1000).
                   You may wish to experiment with various threshold values
                    to get the optimum balance between sensitivity and accuracy.

=============================================================================================================

Output is printed to the output file as well as to STDOUT (ie the screen).
For each hit, ie a close match to the consensus sequence that is being sought,
the output is arranged into six columns:

Column 1          : The position of the hit in the input sequence.
Column 2          : The strand on which the hit is found (+/-).
Column 3          : Sequence of the hit.
Column 4          : A score for the hit.  The higher the score, the better the match!
Column 5          : The position of the hit relative to adjacent/downstream Open Reading Frames
Column 6          : The size of the intergenic region in which the match is found.

For more information see the PromScan website http://www.promscan.uklinux.net

=============================================================================================================
EOF
```

```
}

### Hit.pm class of objects representing a hit,
###  i.e. a good match to the sought sequence
###  defined by the frequency matrix
package Hit;

sub new {
	my $self = {};
	bless $self, "Hit";
	return $self;
}

sub position {
	my $self = shift @_;
	if ( my $specified = shift @_ ) {
		$$self{POSITION} = $specified;
	}
	return $$self{POSITION};
}

sub strand {
	my $self = shift @_;
	if ( my $specified = shift @_ ) {
		$$self{STRAND} = $specified;
	}
	return $$self{STRAND};
}

sub score {
	my $self = shift @_;
	if ( my $specified = shift @_ ) {
		$$self{SCORE} = $specified;
	}
	return $$self{SCORE};
}

sub sequence {
	my $self = shift @_;
	if ( my $specified = shift @_ ) {
		$$self{SEQ} = $specified;
	}
	return $$self{SEQ};
}

sub locus {
	my $self = shift @_;
	if ( my $specified = shift @_ ) {
		$$self{LOCUS} = $specified;
	}
	return $$self{LOCUS};
}

sub intergenic_size {
	my $self = shift @_;
	if ( my $specified = shift @_ ) {
		$$self{SIZE} = $specified;
	}
	return $$self{SIZE};
}

return 1;    # end of package
```