1. Purpose

This document describes the mechanism and process by which prokaryotic operons are predicted for gene sequences stored in the BioHealthBase database.

2. Method Description

Operon finding software (OFS) helps identify co-expressed bacterial genes, using only the genetic code as a starting point. This is an important activity because co-expression of prokaryotic genes controls intracellular enzyme concentrations, influences the operation of metabolic pathways, impacts drug discovery, and is essential in understanding microarray and other gene expression data associated with the target micro-organism and the host it infects. Operons are predicted for the prokaryotes: *Francisella* and *Mycobacterium.*

Two software methods are used by BioHealthBase to identify operons: the SRI "Pathway Tools 11.0" operon finding algorithm ("Ptools-OFS") and the Washington University-St Louis operon finding software package ("WUStL-OFS"). These computational methods first gather information about genetic open reading frames (ORFs), determine the genes directionality (plus or minus strand), calculates the base pair distance between flanking open reading frames, detects the presence of those genes in the same metabolic pathway (PTools only), and determine the presence of common Gene Ontology (GO) terms within the two protein Genbank records (WUStL only). The data is then compiled and used to make a prediction about which genes are present in operon units.

Actual operation of the two prediction tools is dramatically different (below). While the WUStL tool uses precompiled protein table data files (PTT), it also uses different algorithmic weighting parameters and a novel Genbank file mining step so that the operon prediction decisions made by WUStL are sometimes different than PTools-OFS.

The **WUStL-OFS tool** is operated as a shell script in which informant genomes are concatenated, run through formatdb; and then BLASTed against the target genomes. The algorithm calculates orthologs; base pair distances between ORFs, directionality of ORFs, and clusters the output into pairs and operon groups.

The **PTools-OFS tool** is run using a graphical user interface (GUI). Using genome data, the tool generates a pathway map for the organism, finds holes in the built pathway using BLAST, compares the pathway maps of target and informant organisms, and predicts operons using this data.

3. Input Data Preparation

PTools-OFS and WUStL-OFS both require input files to predict operons. These data files are collected for each organism under study, in the format provided at the NCBI FTP site[1]. The OFS algorithms use Genbank (*.gbk), fasta amino acid (*.faa), fasta nucleotide (*.fna), and pathway tools (*.ptt) files.

a) WUStL-OFS

The WUStL-OFS tool is run using only the PTT and FAA files as input. These files need to be generated for the target and informant organisms, and the shell script takes care of all downstream processing in an automated fashion.

b) PTools-OFS

The preparation for making PTools-OFS input files is more complicated.

After the pathologic tools application is started up in its GUI, a new organism database is created as specified. A bacterial taxon number is entered using the NC**** identifier and the "phyla control" is set to "bacteria". Genbank (gbk) and nucleotide files (fna) are then loaded into the PTools system, and the software application is set to run overnight, thus building a pathway map. After a trial parse, a pathway hole filler is run, name matching is verified, consistency is checked, and an automated pathway genome database (PGDB) is built. This PDGB and the four input files above form the collection of data files needed to obtain operon predictions from PTools.

The intermediate file produced in this process is as follows:

```
Directon:   (MT4041.2 MT0001 MT0002 MT0003 MT0004 MT0005
MT0006 MT0007 TRNA-ILE-1 TRNA-ALA-1)
Known TUs:  None
Predicted TU 1: (MT4041.2)  MT4041.2 hypothetical protein
Predicted TU 2: (MT0001)  MT0001  chromosomal DnaA
Predicted TU 3: (MT0002 MT0003 MT0004)
MT0002  DNA polymerase III, beta subunit
MT0003  recF protein
MT0004  conserved hypothetical protein
(etc)
```

This file is used as input to the Operon Finding (Transcription Unit) predictor, a software script. Running the OFS predictor generates a single file, the "operon-prediction-report.txt" which is post-parsed as outlined below.

1 ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/

4. Output Data Post-Processing and Display

Output of the operon predictions is a single tab-delimited file.

The **WUSTL output** is a single file specifying the likelihood that juxtaposed, co-directed genes are present on the same operon:

```
Operon findings at probability cutoff 0.73
1,MAV_0003,MAV_0004,0.818
2,MAV_0005,MAV_0006,0.859
3,MAV_0017,MAV_0019,0.988
4,MAV_0020,MAV_0021,0.818
4,MAV_0021,MAV_0022,0.850
5,MAV_0029,MAV_0030,0.902
```

There is one line for each pair of genes belonging to the same operon. The leftmost number is the number of the operon, where all ORFs led by the same first column number are physically present in the same operon. The 6 data rows in this example above represent 5 operons.

The **PTools output** operon prediction report contains no scores, and is post-parsed using a perl script, opr_parse_codeonly.pl . It looks like this:

```
1,MSMEG_0001a,MSMEG_0002
1,MSMEG_0002,MSMEG_0003
1,MSMEG_0003,MSMEG_0004
2,MSMEG_0005a,MSMEG_0006
3,MSMEG_0008a,MSMEG_0009
3,MSMEG_0009,MSMEG_0010
4,MSMEG_0011a,MSMEG_0012
```

This data shows 4 operons, in 7 data rows; one consisting of genes MSMEG0001a, 0002, 0003, 0004; one containing 0005a and 0006; one containing 0008a, 0009, and 0010; and one containing 0011a and 0012. This is the final file which is to be imported into the database.

5. References

Karp, P.D., Paley, S., and Romero, P. (2002) The Pathway Tools software. Bioinformatics *18* Suppl 1:S225-32

Westover, B.P., Buhler, J.D., Sonnenburg, J.L., Gordon, J.I. (2005) Operon prediction without a training set. *Bioinformatics* 21(7): 880.

NCBI ftp site: URL ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/