Published in final edited form as:

Tuberculosis (Edinb). 2013 January ; 93(1): 18–25. doi:10.1016/j.tube.2012.11.012.

Reannotation of Translational Start Sites in the Genome of *Mycobacterium Tuberculosis*

Michael DeJesus¹, James C. Sacchettini², and Thomas R. Ioerger^{1,*}

¹Department of Computer Science and Engineering, Texas A&M University, College Station, TX USA

²Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX USA

Abstract

Identification and correction of incorrect ORF start sites is important for a variety of experimental and analytical purposes, ranging from cloning to inference of operon structure. The genome of the H37Rv reference strain of *Mycobacterium tuberculosis* (*Mtb*) was originally annotated when it was first sequenced nearly 15 years ago. While this annotation has served the TB research community well as a standard of reference for over a decade, it has been demonstrated experimentally that the actual start sites for an estimated 5-10% of open reading frames differ from the annotation. In this paper, we present a comprehensive bioinformatic analysis of all 3989 ORFs (open reading frames) in the *M. tuberculosis* H37Rv genome. Our method combines information from comparative analysis (alignment to start sites of orthologs in other Actinobacteria), sequence conservation, “protein likeness,” putative ribosome binding sites, and other data to identify translational start sites. The features are combined in a linear model that is trained on dataset of known start sites verified by mass spectrometry, with a cross-validated accuracy of 94%. The method can be viewed as an augmentation of Hidden Markov Model-based tools such as Glimmer and GeneMark by incorporating more information than just the raw genomic sequence to decide which position is the legitimate translational start site for each ORF. Using this analysis, we identify 269 genes that most likely need to be re-annotated, and identify the best alternative translational start site for each. These revised ORF definitions could be used in the reannotation of the H37Rv genome, as well as to prioritize genes for experimental start-site validation.

Introduction

The genome of the H37Rv reference strain of *Mycobacterium tuberculosis* (*Mtb*) was originally annotated when it was first sequenced nearly 15 years ago¹, yielding 3924 open reading frames (ORFs). Subsequently, several additional open reading frames were added in a re-annotation that occurred in 2002, mostly involving the addition of several small ORFs (<100 amino acids), such as additional ribosomal proteins.² While this annotation has served the TB research community well for a decade as a standard of reference, it has been

© 2012 Elsevier Ltd. All rights reserved.

*Corresponding author.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Appendix A: Supplementary data

Supplementary material related to this article can be found online at xxxx.

observed that the actual start site for some open reading frames differs from the annotation. As an example, GlpX was recently shown to have a 34-amino acid N-terminal extension, relative to its annotated start site, which was validated by mass-spectrometry.³ Similarly, RibH was shown to start 6 amino acids further upstream, overlapping with the stop codon on RibA2.³ Epitope tagging was used to show that the translation product of Rv1955 actually begins 45 amino acids downstream from the annotated start site.⁴ In general, identification and correction of incorrect ORF start sites is important for a variety of experimental purposes. In addition, changes to the start sites of genes could affect operon structure, since many genes are predicted to be in operons on the basis of proximity (distance in nucleotides) from the start of the gene to the end of the gene upstream.⁵

The coding regions of genes in prokaryotic genomes is frequently determined, as part of the annotation process, by identifying the largest open reading frames (ORFs) spanning between a candidate start codon and stop codon for the organism, constrained to avoid (or minimize) overlaps. Tools such as Glimmer^{6, 7} and GeneMark^{8, 9} extend this by implementing a Hidden Markov Model, which can be trained on similar organisms to recognize local sequence patterns associated with translational start sites. Modern integrated annotation packages such as RAST¹⁰, can also take advantage of annotations of related organisms to identify the boundaries of ORFs based on a comparison to its orthologs. Bakke et al.¹¹ compared RAST with two other automated annotation servers (IMG and the JCVI pipeline), and found that they had comparable performance in terms of number and lengths of ORFs identified, though RAST predicted unique start sites for around 10% of them.

True start sites of purified proteins can be determined experimentally using techniques such as N-terminal sequencing by Edman degradation, or mass spectroscopy (MS) of proteolytic fragments. Recently, proteomic analysis using MS has been applied to whole-cell lysates and culture filtrates of H37Rv grown in 7H9+OADC media to identify start sites on a genome-wide basis.¹² While not all proteins are identifiable by this method (due to lack of expression, inefficient tryptic digestion, etc.), they were able to confirm the start sites of 727 ORFs in the H37Rv genome, while identifying N-terminal peptides that showed that 33 genes started upstream or downstream of the annotated start site.

Bioinformatics analysis can be used to identify other genes potentially misannotated in the rest of the genome. Characteristics of the currently annotated start site for each ORF can be compared with alternative start sites nearby in the genome. In *M. tuberculosis*, ATG and GTG are predominantly used as start codons, and TTG is used more rarely.¹ Among the 3989 genes in the original annotation of the H37Rv genome, ATG is used as a start codon 60.9% of the time, GTG is used in 33.5% of the genes, and TTG is used only 4.5% of the time. The use of non-ATG start codons, which are recognized less efficiently by the fMet-tRNA, likely plays a regulatory role for these select genes such as translation initiation factor IF5a.^{13, 14}

While ORFs in other organisms are often associated with purine-rich ribosome binding sites (RBS), many ORFs in *Mtb* do not seem to have clear RBSs. In other prokaryotes, RBSs consist of ~6 bp purine-rich sequences usually around 10 bp upstream of a translational start site.¹⁵ The consensus RBS sequence in *E. coli* and *B. subtilis* is AAGGAGG.¹⁶ Many variations around this consensus occur as well, and substitutions affect relative translation initiation rate.¹⁷ However, not all ORFs have an RBS (for example, some genes are transcribed directly from the 5' end of leaderless transcripts). In a survey of 30 sequenced genomes spanning eubacteria and archbacteria, RBS sequences were typically associated with 40-60% of ORFs (though extremes outside this range exist), and only around half (47.7%) of ORFs in the *Mtb* genome displayed a recognizable purine-rich RBS sequence

pattern in the upstream region.¹⁸ Thus RBS analysis provides only a partial sequence-based signal to indicate the true start site of a gene in the *Mtb* genome.

A valuable source of information can be comparison of the start sites of genes with the start sites of orthologs in other organisms. Inconsistency of start sites can easily be recognized in a multiple-sequence alignment, if one organism has a longer or shorter N-terminus compared to the others. Inconsistency does not guarantee a start site is incorrect (as some differences are genuine), and this approach can be subject to systematic biases in how the other genomes may have been annotated. Nonetheless, a comparative-genomics approach using alignments with other mycobacteria was used by Rison et al.³ to identify 11 candidate genes with potential alternative start sites, which yielded GlpX and RibH upon experimental validation.

In addition to start site consistency, there is additional information that can be brought to bear. Sequence conservation among the orthologs can be exploited to help identify a likely start site. The nucleotides downstream of a legitimate start site are typically more conserved, due to constraints of maintaining a stable fold, whereas the non-coding sequence upstream typically accumulates mutations at a higher rate (although there may be short conserved regulatory motifs in this region). Another relevant piece of evidence is “protein-likeness.” If the upstream region were translated into hypothetical amino acids, they are often found to be low-complexity and contain a high proportion of amino acids that would be unusual in an expressed protein product (such as a cluster of cysteines or glycines). For similar reasons, differences in predicted secondary structure might be considered to help discriminate translated regions downstream from a start site versus untranslated regions upstream of a start site, since the amino acid translation of the region upstream of a start site is not constrained to have a regular secondary structure. However, the presence of predicted secondary structure downstream of a true translational start site might not be a reliable indicator, since the N-terminus of many proteins is often disordered¹⁹.

In this paper, we present a comprehensive analysis of the ORFs in the *M. tuberculosis* H37Rv genome. Our bioinformatic method identifies genes with potentially misannotated start sites, based on combining evidence from multiple sequence alignments and analysis of putative peptide sequences. In the end, a summary score is assigned to each ORF that can be used to indicate ORFs with the highest probability of misannotation, and to predict the most likely alternative start site. This prediction could be used to prioritize genes for additional proteomic studies (e.g. validation by mass spectrometry). Furthermore, this data should prove helpful in the process of reannotating the H37Rv genome, and should be of value to researchers attempting to clone and express these genes.

Methods

Identification and Alignment of Orthologs

Orthologs for each gene in H37Rv were identified by running the program BlastP²⁰ against databases constructed for the amino acid sequences in the genomes of 10 other organisms. A significance threshold of $E < 10^{-10}$ was used to select only confident matches. For each ortholog, the nucleotide sequence in the region from -240 to +240 bp relative to the annotated start site was translated into an amino acid sequence (± 80 amino acids). These putative peptide were used to construct a multiple-alignment using ClustalW.²¹

Start site coordinates for predicted ORFs were obtained by running Glimmer v3.0.2⁷ and GeneMarkS v4.7a⁸ on the H37Rv genome. Glimmer was run on the webserver provided by NCBI (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi), using default options for bacterial genomes. GeneMarkS, was run on the server provided by the Georgia

Institute of Technology (<http://exon.gatech.edu/genemarks.cgi>), for prokaryotic genomes (in self-training mode).

Results

Comparative Analysis

For each gene in the H37Rv genome, orthologs were identified in N=10 related organisms (see Table 1). The genomes and public annotations for ten related organisms were retrieved from Genbank (Table 1). The ten species were chosen to represent a range of mycobacteria (including fast- and slow-growing species), but outside the *Mtb* Complex (for example, genes in *M. bovis* are nearly identical to H37Rv, and hence non-informative). Genomes of *Corynebacterium glutamicum* and *Nocardia farcina*, which are close relatives of mycobacteria within the Actinobacteria order, were chosen for additional diversity.

Orthologs for each gene in H37Rv were identified in the genomes of the other 10 organisms using BlastP. Only 1134 (24.43%) genes had orthologs in all 10 organisms; 465 (11.66%) genes had no orthologs, and were excluded from analysis. A histogram is shown in Figure 1 for the number of orthologs for each gene.

The genomic regions within ± 240 bp of the annotated start sites for each of the orthologs were extracted and translated into hypothetical peptide sequences (representing the N-terminus of the predicted protein, fused with a hypothetical prefix of 80 amino acids from the upstream non-coding region, possibly including stop codons). These putative peptide were used to construct a multiple-alignment. An example alignment is shown in Figure 2 for PimB (Rv0557), for which all the orthologs start 3 codons downstream from the annotated start site in H37Rv.

A total of 419 (10.50%) genes had start sites perfectly consistent (aligned) with the orthologs in all the other organisms. In some cases, there was one or a small number of orthologs displaying different start sites, but the start site in H37Rv was still consistent with the majority of the other orthologs. In other cases, there was high variability in the start sites. A histogram of the number of different start sites observed among the orthologs (and H37Rv) is shown in Figure 3.

Focusing on sites where H37Rv could be considered to be inconsistent in this comparative analysis, 1355 (33.98%) genes in H37Rv had start sites that were different than the majority start site based on the alignment to the other organisms. To quantify this difference, we define a score called δ , which represents the difference between the number of orthologs starting at the site identified by the majority in the alignment, and the number of orthologs starting at the same site as H37Rv. This score reflects the degree of consistency with start sites in other organisms (ranging between 0, for cases where they all agree, to n-1, where all the other orthologs start at a site different from H37Rv). The distribution of δ scores is shown in Figure 4. We define ω as the offset (distance measured in amino acids) of the majority start site to the start site in H37Rv. The histogram of offsets other than $\omega=0$ is shown in Figure 5. This shows that the potential difference in start positions is usually small (1-10 amino acids), though sometimes as large as 50 amino acids. If we let $s(i)$ be the number of organisms with a start codon at position i (relative to alignment with *Mtb*), and ω^* be the offset at which the largest number of organisms (other than *Mtb*) start, the δ can be expressed as $\delta = s(i+\omega^*) - s(i)$.

Sequence Conservation Score

Another aspect of the multiple sequence alignments that can be exploited to help identify true start sites is degree of sequence conservation. At each site i the set of amino acids from

all the organisms is collected (excluding gaps and stop codons) and mapped to groups based on similarity as follows: (FYWH)(STNQC)(AVILM)(DE)(KR)(G)(P). This step allows substitutions by chemically similar amino acid to be counted as effectively equivalent. After the most frequent group among the observed amino acids is identified, the conservation score is taken as the ratio of the size of this group to the total number of amino acids observed at the site in the alignment. For example, if there are 7 hydrophobic residues at site i among 10 aligned orthologs, then $\gamma(i) = 0.7$.

Note that if any of the sequences in the alignment contains a gap at this site, the score $\gamma(i)$ is automatically capped at maximum of 0.5, reflecting low conservation. Typically, the sequence conservation is higher downstream of a true start site and lower upstream, however the conservation score can be highly variable depending on the quality of the alignment. In order to reduce the noise in the signal, the conservation score is smoothed by averaging a

window of 10 amino acids downstream of a given start site (i.e. $\bar{\gamma}(i) = \frac{1}{10} \sum_{j=0}^9 \gamma(i+j)$). Figure 6 shows a plot of the smoothed sequence conservation relative to the annotated start site in H37Rv, averaged over all genes. While sequence conservation is higher downstream of the annotated start site (as expected), alternative start sites downstream of a true start site (in the translated region) will obtain a high sequence conservation score even though they are incorrect. In order to distinguish the transition in the signal that occurs at an actual start site, we calculate a finite difference approximation of the derivative of the smoothed conservation score as follows: $\gamma'(i) = \bar{\gamma}(i) - \bar{\gamma}(i-2)$. Figure 6 shows a plot of $\gamma'(i)$ superimposed on the plot of $\bar{\gamma}(i)$, clearly showing a peak at the position of the true start site.

Protein-likeness

In many cases, a transition can be observed in the distribution of hypothetical translated amino acids upstream of a start site to downstream. Whereas the amino acids downstream are expressed as the N-terminus, and thus subject to constraints for stable folding, the untranslated sequence upstream is not constrained, and often displays low complexity. To assess this in the H37Rv sequence, a protein-likeness score $\pi(i)$ is calculated for each site i based on the amino acids in a sliding window of size $L=10$ aa, $a[i..i+L-1]$ (ignoring gaps and stop codons). The similarity of this set of amino acids is compared to the overall distribution of amino acids observed in all 3989 ORFs in H37Rv, $\rho(a)$, using a χ^2 statistic,

$$\pi(i) = \chi^2 = \sum_a \frac{(\text{observed}(a) - \text{expected}(a))^2}{\text{expected}(a)}$$

Where $\text{expected}(a) = L * \rho(a)$. In many cases, the translation of the nucleotide sequence upstream of a start site exhibits an excess of rare amino acids like Cys (frequency=0.89%), which is a good indicator that it is unlikely to be part of the translated protein product. Figure 7 shows a plot of $\pi(i)$ (the χ^2 statistic) relative to the start site in H37Rv, averaged over all 3989 genes. Regions downstream of the annotated start site are more protein-like (i.e. smaller value of $\pi(i)$). As with the conservation score, to capture the transition that occurs at the correct start site and distinguish this position from those downstream which be protein-like as well, we estimate the derivative, $\pi'(i)$, as follows: $\pi'(i) = \pi(i-2) - \pi(i+2)$.

Ribosome Binding Sites

Putative RBSs can also be used as a feature to distinguish some translational start sites. Purine-rich sequences have been observed upstream of genes such as *csaR*²² in the *Mtb* genome. However, no single nucleotide pattern occurs consistently upstream of a large subset of ORFs. For example, by extracting all 6 bp subsequences with 20 bp upstream of

the start codon among all 3989 ORFs, the most frequent (GAGGAG) was only associated with 170 ORFs. Nonetheless, there is a weak signal indicating a bias for purines positioned around 8-11 bp upstream (see Figure 8a). By relaxing the criteria to include any 6 bp purine sequence (regardless of order of As and Gs), this pattern can be identified upstream of 1297 (32.5%) of ORFs. This is statistically significant, as random 20 bp windows sampled in this GC-rich genome contain such a consecutive sequence of purines only ~5.4% of the time. The positions of these subsequences relative to the start codon are shown in Figure 8b. A shorter sequence of 5 consecutive purines occurs upstream of 1841 ORFs (46.2%), while appearing ~13.7% of the time in random 20 bp windows throughout the genome. Because this more relaxed pattern (5 bp A+G) is associated with more ORFs, we use it as a binary feature in our model.

Glimmer and GeneMark

Additional information for reannotation of the H37Rv genome can be garnered by running ORF-prediction models based on HMMs that are trained to detect sequence patterns associated with translational start sites. To make use of these sources of information, Glimmer v3.0.2⁷ and GeneMarkS⁸ were used to make ORF predictions on the H37Rv genome. Both methods predict close to 4,200 ORFs for H37Rv (i.e. see Table 2).

Using start site predictions made by both methods, a score, $\mu(i)$, was developed which quantified whether both, one, or neither of the methods predict a given start site, i , to be the beginning of an ORF. Through this score, we can combine the information obtained from comparative analysis with the information of well-known gene prediction models to make improved predictions about start sites.

Combined Start Site Score

To determine whether a gene might have a misannotated start site based on this data, we developed a combined start site score, which is a linear combination of the features previously described. The features are assigned coefficients (weights) that determine how well they are able to discriminate correct start sites from incorrect ones. We refer to this as the BASS score (for 'Bioinformatic Analysis of Start Sites'):

$$BASS(i) = w_1 * s(i) + w_2 * \delta(i) + w_3 * \gamma'(i) + w_4 * \pi'(i) + w_5 * \mu(i) + w_6 * r(i) + w_7$$

where:

- $s(i)$ = number of orthologs starting at site i
- $\delta(i) = \max_j s(j) - s(i)$ ($0 \leq \delta(i) \leq n-1$)
- $\gamma'(i)$ = derivative of sequence conservation score ($0 \leq \gamma'(i) \leq 1$)
- $\pi'(i)$ = derivative of protein-likeness (χ^2) ($\pi'(i) \geq 0$)
- $\mu(i)$ = Number of other methods (Glimmer or GeneMarkS) that predict i to be a start codon ($\mu(i) \in [0, 1, 2]$)
- $r(i)$ = Binary value describing whether site i contains an RBS-like sequence within 20 bp upstream. ($r(i) \in [0, 1]$)

Stochastic gradient descent was used to determine the value of coefficients which minimized the mean-squared error of the combined score on training data (see Supplementary Material for a description of the algorithm). In order to train the algorithm, a set of 722 genes with translational start sites determined by mass-spectrometry¹² was obtained. Of this initial set of genes, 693 had correctly annotated start sites, and 29 had incorrectly annotated start sites.

Because there were significantly fewer examples of incorrect start sites, the training set was augmented by considering a random sample of alternative start sites among the correctly annotated genes as examples of incorrect start sites, thus increasing the final training set to a total of 1386 examples (i.e. 693 correct sites, and 693 incorrect sites). In order to estimate the overall performance of the linear function, 10-fold cross-validation was used, which ensures that the training and testing sets were non-overlapping. After 1,000 iterations of stochastic gradient descent, the linear combination of features converged to the following set of weights:

$$0.11 * s(i) + -0.28 * \delta(i) + 0.0001 * \gamma'(i) + 0.049 * \pi'(i) + 0.26 * \mu(i) + 0.037 * r(i) - 0.52$$

achieving an average cross-validated accuracy of 94.37%. The difference, $\delta(i)$, between the number of orthologs starting at the most frequent site versus the number of orthologs starting at a given site was assigned the highest magnitude weight (-0.28), followed by $\mu(i)$, the number of methods (Glimmer and GeneMarkS) that agree with given site (0.26), suggesting that these two features were the most relevant for discriminating start sites. The derivative of sequence conservation, $\gamma'(i)$, obtained the lowest weight (0.0001), suggesting that sequence conservation may not be as relevant for discriminating start sites, once the other features are taken into account.

The combined start site score was calculated for all potential start sites within a gene. To qualify as a potential start site, the start codon must be ATG, GTG, or TTG, and must not include a stop codon in-frame. To determine which genes were most likely to have incorrect start sites, the currently annotated start site i was compared to the best-scoring start site j within the surrounding 240 bp region: $\Delta BASS(i) = \max_j BASS(i) - BASS(j)$. Those genes with the largest difference in score are those most likely to have a misannotated start site according to our analysis. While a negative score $\Delta BASS < 0$ indicates that there is an alternative start site with a higher value than the annotated start site, this difference might not be significant. To distinguish the cases that are significant, we use a threshold on the $\Delta BASS$ score of -0.7 . This was determined by calculating the precision and recall on the ORFs from the Kelkar dataset for different thresholds τ , and then calculating the F_1 score,

which is the geometric mean of precision and recall $\left(F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}\right)$. The plot of the F_1 score is shown in Supplementary Figure S1 and displays a peak at $\tau = -0.7$, which is the optimized threshold for detecting misannotated ORFs.

A total of 3531 genes in H37Rv with at least one ortholog were analyzed in this manner (see Table S1 in the Supplementary Material). Of these, 269 genes (6.74% of all ORFs in the genome) exceeded our threshold ($\Delta BASS$ score < -0.7), suggesting that the translational start site for these genes might be misannotated. An example of one of these genes is PimB (Rv0557), which is predicted to start at +3 amino-acids downstream ($\Delta BASS$ score = -2.108), which is where all the orthologs start (Figure 2). RibH has the 12th highest rank by BASS score ($\Delta BASS = -1.888$), and the alternative start site with the highest $BASS(i)$ value is the codon 6 amino acids upstream of the annotated start site, consistent with what has been experimentally verified³. GlpX ($\Delta BASS = -0.576$) is also predicted by our method to start 34 amino-acids upstream of the annotated start site, consistent with the results from mass spectrometry.

To compare the accuracy of the BASS method, Glimmer, and GeneMarkS, their predictions were evaluated on the set of genes with experimentally determined start sites¹² (See Table 3). Among the 693 genes with confirmed start sites, the BASS score predicted 674 (97.26%) of genes had correctly annotated start sites, while Glimmer and GeneMarkS predicted start

sites that corresponded to 551 (79.51%) and 578 (83.41%) of these genes. Among the 29 genes with incorrectly annotated start sites, the BASS score predicted 13 (44.83%) were incorrectly annotated, while Glimmer and GeneMarkS predicted 20 (68.97%) and 19 (65.52%). The lower detection rate among the genes with incorrect start sites could be due to the fact that some of these genes (e.g. *cysG*, *leuA*, *devR*) had annotated start sites that coincided with the start sites in multiple orthologs, which could result from systematic biases in the annotation of these Actinobacterial genomes.

Changes in Operon Structure

The reannotation of translational start sites might affect the prediction of operons, as distance between genes is a criterion commonly considered when making inferences about operon structure²³. Genes within the same operon are expected to have fewer genetic requirements upstream of their start site, as only a single regulatory signal is necessary. Indeed, pairs of genes belonging to the same operon have shorter intergenic distances than expected.²⁴

To determine what changes in operon structure are suggested by the reannotation of start sites, we used the OP-method⁵ to make predictions. The OP-method relies solely on intergenic distances between pairs of genes on the same strand, considering a pair of genes to belong to the same operon as long as their intergenic distance is less than a specified threshold.

Utilizing a distance threshold of 100 nucleotides, pairs of genes were compared using the coordinates in the original annotation of H37Rv and the updated coordinates suggested by our reannotation. Of the 1955 pairs of adjacent genes on the same strand, 1925 (98.46%) were unaffected by the reannotation (i.e. OP-method prediction was the same). Of the remaining 30 pairs of adjacent genes, 10 were newly predicted to share the same operon, i.e. the genes were close enough to be joined in an operon, whereas 20 were now found to be too far apart to be part of the same operon, i.e. an existing operon was split.

One example that illustrates a correction to operon structure that has experimental support is in the gene cluster for mycobactin biosynthesis (*mbtA-mbtJ*). *mbtH-mbtB* are organized in a contiguous fashion on the same strand (Rv2377c-Rv2383c), followed by *mbtA* and *mbtJ* (Rv2384 and Rv2385) on the opposite strand, and finally *mbtI* (Rv2386c). The protein products of *mbtA-F* form a large non-ribosomal peptide synthase complex that perform a sequence of enzymatic reactions to synthesize the siderophore mycobactin, starting from salicylate produced by *mbtI* (salicylate synthase), with the other enzymes making modifications to the side-chain substituents.^{25, 26} The OP-method⁵ predicts that these genes *mbtH-mbtB* are expressed in two groups, *mbtH-mbtE* and *mbtD-mbtB*, due to a gap of 140 bp between *mbtE* and *mbtD*. However, our start site prediction method predicts that the start site of *mbtE* is actually 147 bp farther upstream (on the negative strand), eliminating the intergenic gap altogether, resulting in an overlap of 7 bp. That *mbtH-mbtB* would be expressed as a contiguous operon makes sense, given their functional relationship, and there is evidence from transcriptional profiling that these genes are co-transcribed.²⁷

Discussion

Correct annotation of genomes and accurate identification of ORF boundaries is critical for many reasons. While existing tools for ORF prediction such as GeneMark and Glimmer are fairly accurate, these computational procedures based solely on genomic sequence analysis still make mistakes, as revealed by mass-spectrometry studies. In an assessment on 30 annotated prokaryotic genomes in the NCBI database (<http://www.cbcb.umd.edu/software/glimmer/g3.table2.jun01.shtml>), Glimmer3 was found to predict ORF start sites that agree

with the existing annotation 77.5% of the time on average (range 56-89%). Interestingly, Glimmer3 had the lowest accuracy for the *M. tuberculosis* genome (56.2%). This is unlikely to be due solely to high-GC content, as Glimmer3 had higher accuracy on other GC-rich genomes, such as *Pseudomonas fluorescens* (74.9% accuracy; GC content is 63%, similar to 65% for *Mtb*), and the genome on which Glimmer3 had the second lowest accuracy (*Chorobium tepidum*, 58%) has a GC content of 56%.

Computational analysis based on other information can be used to augment and correct the annotation in some cases. The primary source of information comes from comparative analysis in an alignment with orthologs from related species. By aligning the N-terminus of a gene with orthologs from related species, it is often clear that there is a consensus start site among other organisms, and it can be determined whether the start site in *Mtb* agree (lines up) or disagrees (is offset) from this consensus site. This comparative method is not completely reliable, however, as the annotations of all the other genomes might suffer from the same bias. Also, in some cases the start sites utilized by orthologs are variable, as observed in the multiple alignment, i.e. there might be no clear consensus. In these cases, additional information can be brought to bear. In particular, the degree of sequence conservation provides a strong signal, as sequence downstream of the translational start site is more highly conserved. In addition, the amino acids downstream of the true start site are constrained to be more protein-like, whereas the amino acids in the untranslated region upstream are often found to be low-complexity and have a higher density of low-frequency amino acids. While many ORFs in the *Mtb* genome do not appear to have clear ribosome binding sites¹⁸, a subset do have purine-rich sequences which can also be used to help distinguish which potential start codon is the biologically relevant one. While none of these features by themselves can perfectly identify translational start sites for all ORFs, all of this information can be combined together in a linear model to predict which start sites are potentially misannotated, and which alternative start site is most likely the correct one.

The BASS method we have described can be thought of as an enhancement of Glimmer and GeneMarkS, since it uses their predictions as an input feature, but augments them with other information such as conservation of start sites among orthologs. Although Glimmer and GeneMarkS are capable of making inferences about potential start sites of ORFs given a nucleotide sequence, they generally predict a significantly larger number of incorrect start sites than expected. For example, of the 3531 genes used in our analysis, Glimmer and GeneMarkS predict start sites that correspond to 66.07% and 69.67% of the genes, suggesting that more than 30% of the genes have misannotated start sites. This is significantly larger than the 5-10% of genes expected to be misannotated, based on independent studies by Kelkar et al.¹² and deSouza et al.²⁸ In contrast, the BASS method predicts 6.74% of ORFs to be misannotated, and has an overall cross-validated accuracy of 94%. Thus, by complimenting these methods with a comparative analysis to other genomes, the annotation of ORFs in a genome can be improved.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by NIH grant P01AI095208, and the Robert A. Welch Foundation, grant number A-0015 (JCS).

References

1. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*. 1998; 393:537–544. doi: 10.1038/31159. [PubMed: 9634230]
2. Camus JC, Pryor MJ, Medigue C, Cole ST. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology*. 2002; 148:2967–2973. [PubMed: 12368430]
3. Rison SC, Mattow J, Jungblut PR, Stoker NG. Experimental determination of translational starts using peptide mass mapping and tandem mass spectrometry within the proteome of *Mycobacterium tuberculosis*. *Microbiology*. 2007; 153:521–528. [PubMed: 17259624]
4. Smollett KL, Fivian-Hughes AS, Smith JE, Chang A, Rao T, Davis EO. Experimental determination of translational start sites resolves uncertainties in genomic open reading frame predictions - application to *Mycobacterium tuberculosis*. *Microbiology*. 2009; 155:186–197. doi: 10.1099/mic.0.022889-0. [PubMed: 19118359]
5. Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D. Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. *Genome Biol*. 2003; 4:R59. doi: 10.1186/gb-2003-4-9-r59. [PubMed: 12952538]
6. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*. 1999; 27:4636–4641. doi: gkc675 [pii]. [PubMed: 10556321]
7. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*. 1998; 26:544–548. doi: gkb128 [pii]. [PubMed: 9421513]
8. Borodovsky M, Mills R, Besemer J, Lomsadze A. Prokaryotic gene prediction using GeneMark and GeneMark.hmm. *Curr Protoc Bioinformatics*. 2003 Chapter 4:Unit4 5. doi: 10.1002/0471250953.bi0405s01.
9. Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res*. 2005; 33:W451–454. doi: 33/suppl_2/W451 [pii] 10.1093/nar/gki487. [PubMed: 15980510]
10. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*. 2008; 9:75. doi: 1471-2164-9-75 [pii] 10.1186/1471-2164-9-75. [PubMed: 18261238]
11. Bakke P, Carney N, Deloache W, Gearing M, Ingvorsen K, Lotz M, McNair J, Penumetcha P, Simpson S, Voss L, Win M, Heyer LJ, Campbell AM. Evaluation of three automated genome annotations for *Halorhabdus utahensis*. *PLoS One*. 2009; 4:e6291. doi: 10.1371/journal.pone.0006291. [PubMed: 19617911]
12. Kelkar DS, Kumar D, Kumar P, Balakrishnan L, Muthusamy B, Yadav AK, Shrivastava P, Marimuthu A, Anand S, Sundaram H, Kingsbury R, Harsha HC, Nair B, Prasad TS, Chauhan DS, Katoch K, Katoch VM, Chaerkady R, Ramachandran S, Dash D, Pandey A. Proteogenomic analysis of *Mycobacterium tuberculosis* by high resolution mass spectrometry. *Mol Cell Proteomics*. 2011; 10:M111 011627. doi: M111.011445 [pii] 10.1074/mcp.M111.011445. [PubMed: 21969609]
13. Kozak M. Initiation of translation in prokaryotes and eukaryotes. *Gene*. 1999; 234:187–208. doi: S0378111999002103 [pii]. [PubMed: 10395892]
14. La Teana A, Pon CL, Gualerzi CO. Translation of mRNAs with degenerate initiation triplet AUU displays high initiation factor 2 dependence and is subject to initiation factor 3 repression. *Proc Natl Acad Sci U S A*. 1993; 90:4161–4165. [PubMed: 8483930]
15. Shine J, Dalgarno L. Determinant of cistron specificity in bacterial ribosomes. *Nature*. 1975; 254:34–38. [PubMed: 803646]

16. Ringquist S, Shinedling S, Barrick D, Green L, Binkley J, Stormo GD, Gold L. Translation initiation in *Escherichia coli*: sequences within the ribosome-binding site. *Mol Microbiol.* 1992; 6:1219–1229. [PubMed: 1375310]
17. de Smit MH, van Duin J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc Natl Acad Sci U S A.* 1990; 87:7668–7672. [PubMed: 2217199]
18. Ma J, Campbell A, Karlin S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol.* 2002; 184:5733–5745. [PubMed: 12270832]
19. Lobanov MY, Garbuzynskiy SO, Galzitskaya OV. Statistical analysis of unstructured amino acid residues in protein structures. *Biochemistry (Mosc).* 2010; 75:192–200. doi: BCM75020236 [pii]. [PubMed: 20367606]
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990; 215:403–410. doi: 10.1016/S0022-2836(05)80360-2 S0022-2836(05)80360-2 [pii]. [PubMed: 2231712]
21. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. *Bioinformatics.* 2007; 23:2947–2948. doi: btm404 [pii] 10.1093/bioinformatics/btm404. [PubMed: 17846036]
22. Liu T, Ramesh A, Ma Z, Ward SK, Zhang L, George GN, Talaat AM, Sacchettini JC, Giedroc DP. CsoR is a novel *Mycobacterium tuberculosis* copper-sensing transcriptional regulator. *Nat Chem Biol.* 2007; 3:60–68. doi: nchembio844 [pii] 10.1038/nchembio844. [PubMed: 17143269]
23. Ermolaeva MD, White O, Salzberg SL. Prediction of operons in microbial genomes. *Nucleic Acids Res.* 2001; 29:1216–1221. [PubMed: 11222772]
24. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci U S A.* 2000; 97:6652–6657. doi: 10.1073/pnas.110147297 110147297 [pii]. [PubMed: 10823905]
25. Quadri LE, Sello J, Keating TA, Weinreb PH, Walsh CT. Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. *Chem Biol.* 1998; 5:631–645. doi: S1074-5521(98)90291-5 [pii]. [PubMed: 9831524]
26. McMahon MD, Rush JS, Thomas MG. Analyses of MbtB, MbtE, and MbtF suggest revisions to the mycobactin biosynthesis pathway in *Mycobacterium tuberculosis*. *J Bacteriol.* 2012; 194:2809–2818. doi: JB.00088-12 [pii] 10.1128/JB.00088-12. [PubMed: 22447909]
27. Roback P, Beard J, Baumann D, Gille C, Henry K, Krohn S, Wiste H, Voskuil MI, Rainville C, Rutherford R. A predicted operon map for *Mycobacterium tuberculosis*. *Nucleic Acids Res.* 2007; 35:5085–5095. doi: gkm518 [pii] 10.1093/nar/gkm518. [PubMed: 17652327]
28. de Souza GA, Malen H, Softeland T, Saelensminde G, Prasad S, Jonassen I, Wiker HG. High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using *Mycobacterium tuberculosis* as an example. *BMC Genomics.* 2008; 9:316. doi: 1471-2164-9-316 [pii] 10.1186/1471-2164-9-316. [PubMed: 18597682]

Figure 1. Histogram of number of orthologs for genes in H37Rv among 10 related mycobacterial species.

```

M_vanbaalenii dsaargriqcmgppsslvw---rpqacfcrrtgnsdislaatsptggavtvtgVRVAIVAESFLPNVNGVTNSVLRVIEHLRRTGHEVLVIAPDTPRGQPAADK
M_marinum --vglalldrldkrpdqradtmstp*gvpadfaprm-pggnladsrr*hgggVRVAIVAESFLPNVNGVSNVVRVLEHLRRTGHEALVIAPDNPPGEPRADR
M_ulcerans --vglalldrldkrpdqradtmstp*gvpadfaprm-pggnladsrr*hgggVRVAIVAESFLPNVNGVSNVVRVLEHLRRTGHEALVIAPDNPPGEPRADR
M_avium gwlgaaavllvlla-ladkwldrrrgkgag-padfa*rm-pagnlsdkrr*hrdgVRVAIVAESFLPEVNGVSNVIRVLEHLRRTGHEALVIAPDTPPGQPRAE
M_smegmatis -tagrgedhsrqy-rtqhpghphggnglsghrvfecri-pwrnlaancdgdhrdVRVAIVAESFLPNVNGVTNSVLRVIDHLRRTGHEVLVIAPDTPRGQPPADR
M_gilvum -deryayhhprcg-vddghafrrrapvmshphfv*hi-parnladrcrcqcrgrVRVAIVAESFLPNVNGVTNSVLRVIEHLRRTGHEALVIAPDTPRGEPPADR
M_abscessus --sswwcgrspegccwsgpgcgaeaplltlkfa*rm-pggnlts*s*havdVRVAIVAESFLPNVNGVSNVLRVLEHLRRTGHEALVIAPDTPRGQAPAT
N_farcinia --crcggcwcgraganpwrstelgrrrgvrrnfpqps-rrcgavarraghpgvVRVAIVSEFLPNMNGVVNSVLRVLDHLRHRGHQALVIAPDNPRGLPPAPG
C_glutamicum ---ddanettv*ssmpgaftm*pkrsghlnkvldkamfg*lrcrhdvaiirpMRVAIVAESFLPNVNGVTNSVLRVLEHLKANGHDALVIAPGARDFEEIGH
M_tuberculosis g*sppprwc*rcwtsgwnvkvirrlqraeqqsshavcryatistagadsVCGVRVAIVAESFLPQVNGVSNVVRVLEHLRRTGHEALVIAPDTP---PGEDR

```

Figure 2.

Multiple alignment of orthologs of Rv0557 (PimB). The best ortholog of Rv0557 in 10 other organisms was identified, and the genomic region within ± 150 bp of the annotated start sites were translated into amino acids and aligned by ClustalW. Start codons are underlined. Gaps are indicated by '-'. Stop codons are indicated by '*'. The hypothetical translation of upstream regions is shown in lower case letters. The positions where a start codon was observed in any other organism are shown in yellow. The positions where legitimate alternative start codons occur in the *Mtb* sequence are shown in magenta. Annotated start in *Mtb* is colored in cyan.

Figure 3. Histogram of number of different start sites observed in multiple alignment among orthologs for each gene in H37Rv.

Figure 4.

Histogram of delta score, representing consistency of start sites. The number of orthologs with start sites matching the annotated start site in H37Rv is compared to the number of orthologs agreeing with the start site with most orthologs overall. Smaller delta values represent more consistency.

Figure 5. Histogram of codon distance between the annotated start site in H37Rv and the alternative start site agreed upon by the majority of orthologs.

Figure 6.

Average conservation score. The line with filled circles shows the plot of smoothed conservation score, $\bar{\gamma}(i)$, characterizing the average sequence conservation of 10 amino-acids downstream of a given position was calculated for locations around the annotated start site ('offset'), averaged across all genes. Larger values represent higher sequence conservation. The line with empty squares shows an approximation of the derivative, $\gamma'(i)$, used to distinguish the transition point that is observed at start sites.

Figure 7.

Average Protein-Likeness score. The line of filled circles shows a plot of protein-likeness score, $\pi(i)$, characterizing how protein-like a sequence of 10 amino-acids downstream of a given start site was calculated for positions around the annotated start site ('offset'), averaged across all genes. Smaller values represent sequences that are more protein-like. The line with empty squares represents an approximation of the derivative, $\pi'(i)$, which was used to distinguish the transition that is observed at start sites.

a)

b)

Figure 8.

a) Distribution of nucleotides in the untranslated region upstream of start codons for all ORFs in the *Mtb* genome. b) The distribution of positions at which 5 bp purine subsequences start (if multiple instances occur, the one closest to the start codon is chosen).

Table 1

Organisms used for comparative analysis.

organism	Genbank accession	# ORFs	% identity of 16S rRNA
<i>M. tuberculosis</i> H37Rv	NC_000962.2	3,989	100%
<i>M. marinum</i> M	NC_010612.1	5,423	99%
<i>M. ulcerans</i> Agy99	NC_008611.1	4,160	99%
<i>M. avium</i> 104	NC_008595.1	5,120	98%
<i>M. leprae</i> Br4923	NC_011896.1	1,604	97%
<i>M. gilvum</i> PYR GCK	NC_009338.1	5,241	96%
<i>M. smegmatis</i> mc ² 155	NC_008596.1	6,717	96%
<i>M. vanbaalenii</i> PYR 1	NC_008726.1	5,979	95%
<i>M. abscessus</i> ATCC 19977	NC_010397.1	4,920	94%
<i>N. farcinia</i> IFM 10152	NC_006361.1	5,681	92%
<i>C. glutamicum</i> ATCC 13032	NC_003450.3	2,993	90%

Table 2

ORF predictions by Glimmer and GeneMarkS.

method	Total ORFs predicted	ORFs with same start and stop codon as in H37Rv	ORFs with different start codon than in H37Rv	ORFs with different stop codon than in H37Rv
Glimmer	4343	2532 (58.30%)	1147 (26.41%)	1170 (26.94%)
GeneMark	4152	2697 (65.96%)	861 (20.74%)	893 (21.51%)

Table 3Performance on experimentally determined start sites¹²

Method	Correct start sites (out of 693 genes)	Incorrect start sites (out of 29 genes)
Δ BASS -0.7	674 (97.26%)	13 (44.83%)
Glimmer	551 (79.51%)	20 (68.97%)
GeneMark	578 (83.41%)	19 (65.52%)