We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

18

Functional Analysis of Intergenic Regions for Gene Discovery

Li M. Fu
Pacific Tuberculosis and Cancer Research Organization, Anaheim, CA USA

1. Introduction

Gene finding can be defined as a problem of identifying a stretch of the genomic DNA sequence that is biologically functional. Such a genomic DNA sequence is known as a gene. A gene performs a function like protein coding or regulation at the molecular level and plays a biological role, such as growth, metabolism, and intelligence. Traditionally, gene finding relies on numerous biological experiments and statistical analysis to pinpoint the location of a new gene in a genetic map. With the advent of bioinformatics, gene finding has largely become a computational problem. Genes are predictable based on the genomic sequence alone. However, the determination of the specific function and biological role of a gene would still demand *in vivo* experimentation, which is hoped to be reduced or even replaced by new bioinformatics algorithms in the future.

A newly sequenced genome is annotated thoroughly so that the information it carries can be utilized. In essence, genome annotation is to identify the locations of genes and all of the coding regions in a genome, and determine their protein products as well as functions. Hundreds of bacterial genome sequences are publicly available and the number will soon reach a new milestone. Gene annotation by hand is almost impossible to handle the deluge of new genome sequences appearing at this pace. The need for automated, large-scale, high-throughput genome annotation is imminent (Overbeek, Begley et al. 2005; Van Domselaar, Stothard et al. 2005; Stothard and Wishart 2006). The basic level of genome annotation is the use of BLAST (Altschul, Gish et al. 1990) for finding similarities between related genomic sequences. Integration with other sources of information and experimental data is a trend in genome annotation.

A recent study indicates that many genomes could be either over-annotated (too many genes) or under-annotated (too few genes), and a large percentage of genes may have been assigned a wrong start codon (Nielsen and Krogh 2005). The fact that the original genome annotation is accurate and complete upon submission does not guarantee that it will not be changed, as new experimental evidence and knowledge would continue to arrive and constant updates would be inevitable. However, re-annotation of the whole genome is not very fruitful, as most of the genes have been identified in the first annotation. For example, the re-annotation of the H37Rv genome resulted in about 2% of new protein-coding sequences (CDS) added to the genome. The result reflects the limitation with current genome annotation technology. To address the issue, we developed a new method for gene finding in an annotated genome. We select the genome of *Mycobacterium tuberculosis*, the

causative pathogen of tuberculosis, as the experimental genome for this study. The availability of the complete genome sequence of *M. tuberculosis* H37Rv (Cole, Brosch et al. 1998) has led to a better understanding of the biology and pathogenicity of the organism, and new molecular targets for diagnostics and therapeutics can be invented at a fast pace by focusing on genes with important functions.

In our previous studies, we found that some intergenic sequences in *M. tuberculosis* genome exhibited expression signals, as detected by the Affymetrix GeneChip (Fu 2006; Fu and Fu-Liu 2007; Fu and Shinnick 2007). The same observation has been made for other bacteria, such as *Bacillus subtilis* (Lee, Zhang et al. 2001), and also holds true in the eukaryotic system (Zheng, Zhang et al. 2005). At present, it is not clear whether or how intergenic expression represents gene activity. Here, we presented our research work concerning gene discovery in the intergenic sequences based on transcription activity. In this work, new protein-coding genes were identified by the bioinformatics criteria based on the gene structure, protein coding potential, and ortholog evidence, in conjunction with microarray-based transcriptional evidence.

2. Research methods and design

The developed method of gene finding in the intergenic sequences proceeds as follows:

1. Transcription analysis to identify intergenic regions exhibiting significant gene expression activity.
2. Coding potential and gene structure analysis on active intergenic elements identified based on transcription evidence.
3. Protein domain search to identify functional domains in each active intergenic element with significant transcription activity and coding potential.
4. Homology search based on BLAST to seek homologue evidence.

The flowchart of the method is displayed in Figure 1.

The method was applied to the originally annotated *M. tuberculosis* H37Rv genome (Cole, Brosch et al. 1998). The genes discovered in the intergenic sequences were validated against recent findings in the literature. The research protocols (Fu and Shinnick 2007) were described below.

2.1 RNA isolation

M. tuberculosis strain H37Rv was obtained from the culture collection of the Mycobacteriology Laboratory Branch, Centers for Disease Control and Prevention at Atlanta. Bacterial lysis and RNA isolation were performed following the procedure at the CDC lab, Atlanta (Fisher, Plikaytis et al. 2002). Briefly, cultures were mixed with an equal volume of RNALater™ (Ambion, Austin, TX) and the bacteria harvested by centrifugation (1 min, 25000g, 8°C) and transferred to Fast Prep tubes (Bio 101, Vista, CA) containing Trizol (Life Technologies, Gaithersburg, MD). Mycobacteria were mechanically disrupted in a Fast Prep apparatus. The aqueous phase was recovered, treated with Cleanascite (CPG, Lincoln Park, NJ), and extracted with chloroform-isoamyl alcohol (24:1 v/v). Nucleic acids were ethanol precipitated. DNAase I (Ambion) treatment to digest contaminating DNA was performed in the presence of Prime RNase inhibitor (5′-3′, Boulder, CO). The RNA sample was precipitated and washed in ethanol, and redissolved to make a final concentration of 1 mg/ml. The purity of RNA was estimated by the ratio of the readings at 260 nm and 280 nm (A260/A280) in the UV. 20 ul

RNA samples were sent to the UCI DNA core and further checked through a quality and quantity test based on electrophoresis before microarray hybridization.

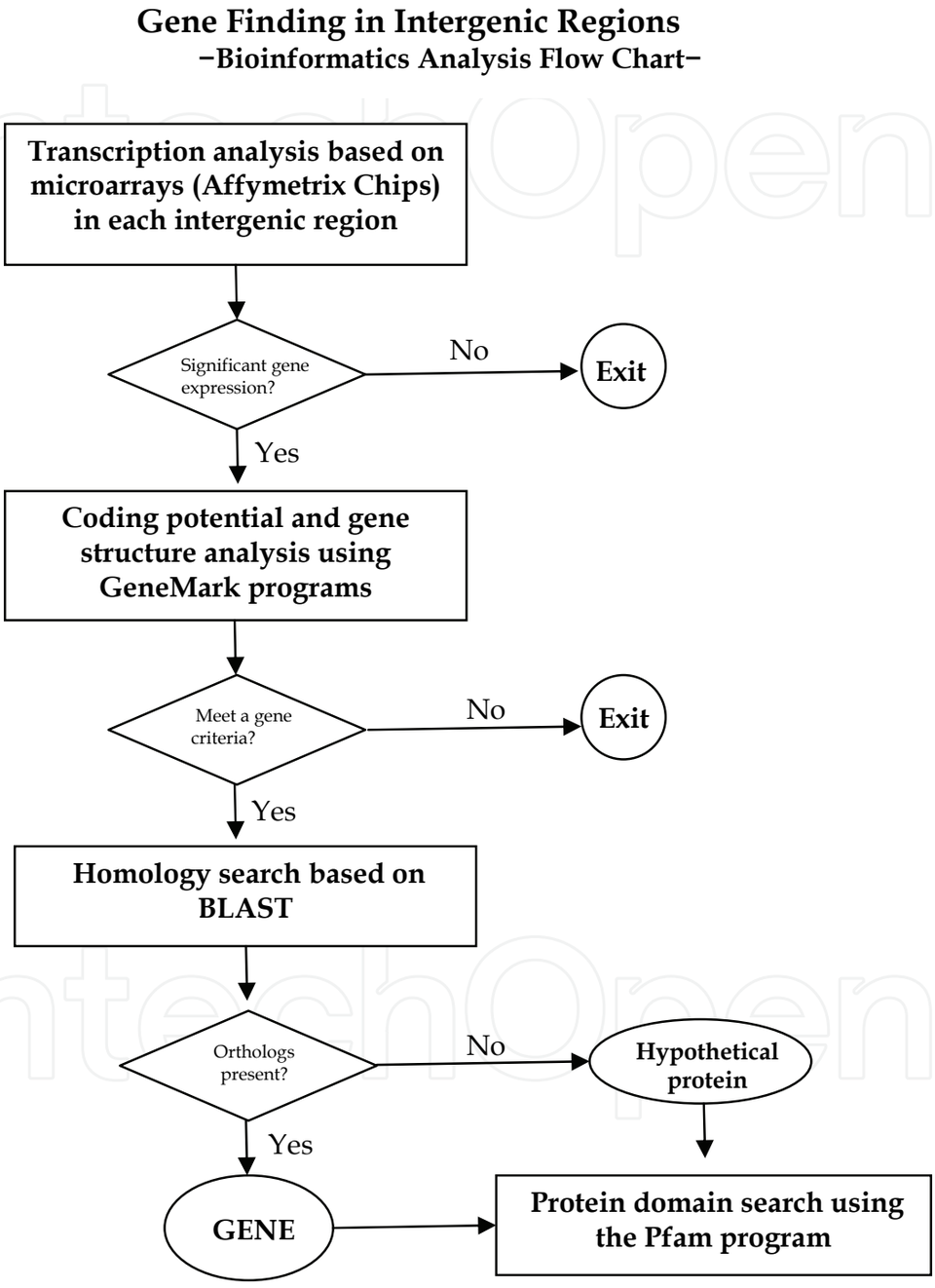


Fig. 1. The bioinformatics method with a flowchart developed for finding genes in intergenic regions.

2.2 Microarray hybridization

In this study, we used the anti-sense Affymetrix *M. tuberculosis* genome array (GeneChip). The probe selection was based on the genome sequence of *M. tuberculosis* H37Rv (Cole, Brosch et al. 1998). Each annotated ORF (Open Reading Frame) or IG (Intergenic Region) was interrogated with oligonucleotide probe pairs. An IG referred to the region between two consecutive ORFs. The gene chip represented all 3924 ORFs and 740 intergenic regions of H37Rv. The selection of these IGs in the original design was based on the sequence length. Twenty 25-mer probes were selected within each ORF or IG. These probes were called PM (Perfect-Match) probes. The sequence of each PM probe was perturbed with a single substitution at the middle base. They were called MM (Mismatch) probes. A PM probe and its respective MM probe constituted a probe pair. The MM probe served as a negative control for the PM probe in hybridization.

Microarray hybridization followed the Affymetrix protocol. In brief, the assay utilized reverse transcriptase and random hexamer primers to produce DNA complementary to the RNA. The cDNA products were then fragmented by DNAase and labeled with terminal transferase and biotinylated GeneChip DNA Labeling Reagent at the 3' terminal.

Each RNA sample underwent hybridization with one gene array to produce the expression data of all genes on the array. We performed eleven independent bacterial cultures and RNA extractions at different times, and collected eleven sets of microarray data for this study. A global normalization scheme was applied so that each array's median value was adjusted to a predefine value (500). The scale factor for achieving this transformed median value for an array was uniformly applied to all the probe set values on a specific array to result in the determined signal value for all the probe sets on the array. In this manner, corresponding probe sets can be directly compared across arrays.

2.3 Gene expression analysis

The gene expression data were analyzed by the program GCOS (GeneChip Operating Software) version 1.4. In the program, the Detection algorithm determined whether a measured transcript was detected (P Call) or not detected (A Call) on a single array according to the Detection *p*-value that was computed by applying the one-sided Wilcoxon's signed rank test to test the Discrimination scores (R) against a predefined adjustable threshold τ. The parameter τ controlled the sensitivity and specificity of the analysis, and was set to a typical value of 0.015, and the Detection *p*-value cutoffs, α_1 and α_2, set to their typical values, 0.04 and 0.06, respectively, according to the Affymetrix system.

2.4 Gene prediction

Protein-coding region identification and gene prediction were performed by the programs, GeneMark and GeneMark.hmm (Lukashin and Borodovsky 1998; Besemer and Borodovsky 2005) (http://exon.gatech.edu/GeneMark/), respectively. The prokaryotic version and the *M. tuberculosis* H37Rv genome were selected. Both programs use inhomogeneous Markov chain models for coding DNA and homogeneous Markov chain models for non-coding DNA. GeneMark adopts Bayesian formalism, while GeneMark.hmm uses a Hidden Markov Model (HMM).

2.5 Protein domain search

The Pfam program version 20.0 (Finn, Mistry et al. 2006) (http://pfam.wustl.edu/) was employed to conduct protein domain search after the input DNA sequence was translated

into a protein sequence in six possible frames. The search mode was set to "global and local alignments merged", and the cut-off E-value set to 0.001, which was more stringent than the default value of 1.0. Pfam maintained a comprehensive collection of multiple sequence alignments and hidden Markov models for 8296 common protein families based on the Swissprot 48.9 and SP-TrEMBL 31.9 protein sequence databases.

2.6 Homology search

The BLASTx program (Altschul, Gish et al. 1990) (http://www.ncbi.nlm.nih.gov/BLAST/) was used to identify high-scoring homologous sequences. The program first translated the input DNA sequence into a protein sequence in six possible frames, and then matched it against the non-redundant protein sequence database (nr) in the GenBank and calculated the statistical significance of the matches. The default cut-off E-value was 10.0 but we set it to 1.0×10^{-10}. Orthologs refer to homologs in different strains of the same species. Orthologs provide critical evidence for gene finding and characterization in a new genome sequence. A typical prokaryotic gene has the following structure: the promoter, transcription initiation, the 5' untranslated region, translation initiation, the coding region, translation stop, the 3' untranslated region, and transcription stop.

3. Results

In our previous research, we conducted a genome-wide expression analysis on intergenic regions using the Affymetrix GeneChip (Fu and Shinnick 2007). The transcriptional activity of intergenic regions was measured based on a set of eleven independent RNA samples extracted from *M. tuberculosis* culture. Each RNA sample contained the information of genome-wide expression of genes and intergenic elements. The Affymetrix GeneChip was uniquely suited to this study since it had the advantage of encoding both genes and intergenic sequences whereas other types of microarray like the cDNA array could not profile intergenic sequences. As an additional strength, the Affymetrix array was designed to minimize cross-hybridization by using unique oligonucleotide probes and the pair of PM (Perfect-Match) and MM (Mismatch) probes. The cross-hybridization of related or overlapping gene sequences often contributed to false positive signals, especially in the case when long cDNA sequences were used as probes. A study demonstrated that the Affymetrix GeneChip produced more reliable results in detecting changes in gene expression than cDNA microarrays (Li, Pankratz et al. 2002).

In this work, genes in the intergenic sequences were recognized based on transcriptional activity, structural patterns, and coding potential, and subsequently validated through sequence comparison with orthologs from other *M. tuberculosis* strains.

3.1 Transcriptional analysis

An intergenic region was assumed to transcribe if there existed transcripts in the RNA sample that were bound to the probes encoding that intergenic sequence. The presence or absence of a given transcript was determined in accordance with the Detection algorithm of the Affymetrix system. In this study, a gene or intergenic region was determined to express (transcriptionally active) only if the derived mRNA was present (P-call) in more than 90% of the collected RNA samples with a Detection p-value < 0.001. The status of active

transcription assigned to an intergenic sequence signified the possible presence of a gene within that sequence. We focused on finding protein-coding genes and neglected regulatory genes that transcribed into a regulatory RNA instead of mRNA. Furthermore, it was not clear how much cross-hybridization would occur between genes and intergenic sequences. As a result, the functional criterion based on expression activity was strengthened by structural analysis for minimizing false positives in gene identification.

3.2 Sequence-based gene prediction

In the sequence-based approach to gene prediction, gene structure and coding potential are the two mutually supportive elements. The GeneMark algorithm was applied to an intergenic sequence for checking whether it contained a probable coding region, and the GeneMark.hmm algorithm was used for predicting a gene within the sequence. The criteria based on the predefined transcriptional evidence, coding potential, and gene structure yielded 65 candidate genes in the intergenic regions of *M. tuberculosis* H37Rv.

3.3 Protein domain search

The biological function of a gene is determined using *in vivo* experimentation in a traditional approach. Recently, the wealth of bioinformatics knowledge in the functional domains of proteins has enabled the function of a molecular sequence to be characterized directly, subject to *in vivo* validation. Thus, the "candidate" genes within the intergenic sequences that satisfied the criteria based on transcription activity, gene structure, and coding potential were further examined for embedded functional domains. To this end, Pfam was applied to search on the protein sequences of the candidate genes. Twelve of them were found to have a known domain (Tables 1); a found domain was generally located within the predicted gene sequence, but there were a few exceptions (i.e., IG398 and IG1140) in which a domain was found within the intergenic sequence but outside the predicted gene sequence. The biological function and role of a gene is deducible from its associated functional domains; yet sufficient evidence from homology or biochemistry would serve to corroborate it.

3.4 Homologue evidence

In evolutionary biology, a reliable means for predicting the function of an unknown gene sequence is based on homologs or orthologs. BLAST is a bioinformatics program for database search, allowing functional and evolutionary inference between sequences. In this study, BLAST was employed to retrieve from sequence databases all proteins that produced statistically significant alignment with a given intergenic sequence under consideration. The sequences retrieved by BLAST were homologous to the query sequence. It turned out that the highest-scoring homologous sequences with ≥98% identity were consistently those belonging to the same strain (H37Rv) or different strains of *Mycobacterium tuberculosis* (e.g., CDC1551, F11, and C). These sequences are coding sequences described in the currently annotated genome of *M. tuberculosis*.

A homologous sequence found in different strains of the same species often represents an ortholog that shares similar function, whereas a homologous sequence found in the same organism is a paralog (which is produced via gene duplication within a genome) that tends to have a different function. No evidence suggested paralogs in our analysis, as argued based on the following observation. We noted that, given an intergenic sequence, when BLAST returned a homologous sequence pertaining to the H37Rv strain, it was apparently

the same protein-coding sequence contained in the intergenic sequence. This is because the intergenic sequence used as a query and its homologous sequence returned by BLAST occupied the same physical location within the genome, as inferred from information given by the Affymetrix Genechip. This coincidence was further explained by the fact that the intergenic sequence was named according to the early version of the H37Rv genome annotation while the homologous sequence was retrieved from the GenBank which contained all up-to-date genes. The results are significant. First, we demonstrated that our method was able to identify protein-coding genes in intergenic regions previously considered as non-coding sequences. Secondly, our method based on bioinformatics and transcriptional evidence correctly predicted these changes on a high-throughput, genomic scale. The changes refer to

- IG1061 → (containing) Rv1322A
- IG499 → Rv0634B
- IG617 → Rv0787A
- IG1741 → Rv2219A
- IG2500 → Rv3198A
- IG2053 → Rv2631
- IG1179 → Rv1489A
- IG2522 → Rv3224B
- IG1291 → Rv1638A
- IG398 → Rv0500A
- IG2870 → Rv3678A
- IG188 → Rv0236A
- IG2498 → Rv3196A,
- IG2591 → Rv3312A
- IG595 → Rv0755A
- IG1814 → Rv2309A
- IG1030 → Rv1290A
- IG2141 → Rv2737A

In the above findings, each intergenic region contained an independent gene/CDS with the only exception that part of IG2053 was incorporated in its left-flanking CDS. The presence of a gene structure in an IG and its lack of functional correlation with its adjacent genes suggested that it was not a run-away segment from adjacent genes.

In our analysis, predicted genes located within intergenic sequences that met the criteria defined based on protein-coding potential, structural patterns, and transcription evidence, were called "candidate" genes. If a candidate gene of unknown function was homologous to another gene of know function, the candidate gene was assigned the function associated with its homologous gene. Nonetheless, the strategy of inferring the function of an uncharacterized sequence from its orthologs had limited value in analyzing intergenic data mainly because most of the orthologs found in this study were hypothetical proteins with unknown function. We did not assign a specific function to a candidate gene until it had an ortholog of known function, whether or not the candidate gene carried a known functional domain. In the absence of a specific function assigned, a CDS was termed a hypothetical protein rather than a gene in our system.

In this work, six intergenic sequences were identified that met the criteria we defined, including protein coding, structural patterns, transcription, and ortholog evidence: IG499,

IG617, IG1741, IG2500, IG1567, and IG2229, among which four genes had been reported in the *M. tuberculosis* H37Rv genome (Table 1). A hypothetical protein was found in 52 intergenic sequences and 14 among them had been reported in the H37Rv genome. Overall, this research discovered two genes with a specific function and 38 hypothetical proteins that had not been reported in the H37Rv genome (Fu and Shinnick 2007). The two new genes discovered were a DNA-binding protein in the CopG family and a nickel binding GTPase, located in IG1567 and IG2229, respectively (Figure 2). It was worth noting that 4.3% of intergenic regions exhibiting transcriptional activity contained a gene described in the re-annotated H37Rv genome, compared with 1.0% of intergenic regions in the absence transcriptional activity. The four-fold increase in likelihood suggested that microarray-based transcriptional analysis would facilitate genome-wide gene finding.

IG	Lt Flank	Rt Flank	Domain ID	Re-annotated H37Rv Gene
IG1061	Rv1322	Rv1323	Glyoxalase	Rv1322A*
IG499	Rv0634c	Rv0635	Ribosomal_L33	Rv0634B
IG617	Rv0787	Rv0788	PurS	Rv0787A
IG398	Rv0500	Rv0501	DUF1713	Rv0500A*
IG1741	Rv2219	Rv2220	RDD	Rv2219A
IG2500	Rv3198c	Rv3199c	Glutaredoxin	Rv3198A
IG2053	Rv2631	Rv2632c	UPF0027	Rv2631*
IG1179	Rv1489c	Rv1490	MM_CoA_mutase	Rv1489A*
IG1140	Rv1438	Rv1439c	TetR_N	None
IG2522	Rv3224	Rv3225c	YbaK	Rv3224B*
IG1567	Rv1991c	Rv1992c	RHH_1	None
IG2229	Rv2856	Rv2857c	cobW	None

*: Hypothetical protein

Table 1. The functional domains of the predicted genes located within the intergenic sequences of *M. tuberculosis* H37Rv genome. Each intergenic sequence shown is characterized by its flanking genes or ORFs and the functional domain identified in the translated protein sequence. Most of IGs with a functional domain contain a gene in the re-annotated H37Rv genome (Fu and Shinnick 2007).

4. Discussion

Computational algorithms for gene prediction are divided in two classes: One is based on sequence similarity and the other based on gene structure and signal. The latter is known as *ab initio* prediction. The first class of algorithm, represented by BLAST (Altschul, Gish et al. 1990), finds sequences (DNA, protein, or ESTs) in the database that match the given sequence, whereas the second class of algorithm, such as Hidden Markov Model (Burge and Karlin 1997; Lukashin and Borodovsky 1998; Besemer and Borodovsky 2005), builds a model

of gene structure from empirical data. They both have limitations. For instance, the sequence-based approach is not applicable if no homology is found, whereas the model-based approach is not workable if no adequate training data is available for model parameter estimation. To explore an alternative in a different perspective, the method developed in our research combined sequence alignment, transcriptional evidence, and homology. In particular, the transcriptional activity of a piece of DNA is direct evidence that it is functioning. This is important because a gene means a piece of genomic DNA that is functional. In the absence of functional evidence, any gene computed by whatever algorithms will remain hypothetical.

New gene 1:
[Location]: Between Rv1991c and Rv1992c
[Product]: DNA-binding protein, CopG family
[Nucleotide Sequence]:
atcgtccatggtttctagcacgcggtatgcgttggccacggcgagggcctccgcttcgtcggtgccatggatgctctctagagccctgtcgatctggcccgtgagcaattggg
cgtccagctcgtgcaggtagcgctgcgcagccttcgtgaagaactcggaccgactcatgccgagctcactcgcacgccgcgatacccgatcgaacgtctcatccggcagag
aaatagctgtcttcat
[Protein Sequence]:
mktaislpdetfdrvsrraselgmsrsefftkaaqrylheldaqlltgqidralesihgtdeaealavanayrvletmdd

New gene 2:
[Location]: Between Rv2856 and Rv2857c
[Product]: Nickle binding GTPase involved in regulation of expression urease and hydrogenase
[Nucleotide Sequence]:
atggtctcctcggtcaccgagggcaaggacaagccgctgatgtacccggcgacgttccgctcgagggatgtagtgctgctcgacaagatcgacttggtgccctttctggac
gccgacgtggacgcgtatatcgcgcatgtccgcgaggtcaacgcagccgcgacgatcctgccgaccagcacgcgcaccggagccggcatggggtcctggtcatga
[Protein Sequence]:
mvssvtegkdkplmypatfrsrdvvlldkidlvpfldadvdayiahvrevnaaatilptstrtgagmgsws

Fig. 2. Examples of new genes with a predicted function found in the genome of *M. tuberculosis* H37Rv (Fu and Shinnick 2007).

The whole *M. tuberculosis* H37Rv genome has been sequenced and annotated comprehensively (Cole, Brosch et al. 1998). Transcriptional analysis of intergenic regions is a means of exploring unknown genes. Our idea capitalized on transcription analysis in gene finding, which was useful especially when applied to an annotated genome. Current genome annotation technology allowed all genes to be identified by a computational algorithm, and it was unlikely to add new genes through re-annotation at the same time unless using a different algorithm. Thus, it was within expectation that the number of new protein-coding sequences due to re-annotation was merely 2% of that in the original submission of *M. tuberculosis* genome (Camus, Pryor et al. 2002). Through homology and pattern-based search, most protein-coding sequences with a predicted function have been reported. Yet, transcriptional evidence could quickly hint at potential protein-coding genes in the intergenic regions. It is encouraging that we are still able to find new genes in this study given the fact that the current knowledge concerning *M. tuberculosis* genes is derived from intensive research in the field involving *in vivo* biological experiments, such as gene deletion and complementation. Thus the integration of the sequence- and function-based analyses would be a useful approach to not just gene characterization but also gene prediction. As the experiment was based on the standard in vitro growth condition of *M.*

tuberculosis, silent genes under this condition were not under examination in this study, but the same idea should be applicable to other genomes under other conditions, and contribute to the improvements of current gene databases.

The methods presented here did not address the issue of genes that did not code proteins. There are a number of regulatory, non-coding RNAs assuming a distinct role from mRNA, rRNA and tRNA. Many such RNAs have been identified and characterized both in prokaryotes and eukaryotes and their main functions are posttranscriptional regulation of gene expression and RNA-directed DNA methylation (Erdmann, Barciszewska et al. 2001; Pickford and Cogoni 2003). A non-coding RNA has neither a long open reading frame nor a gene structure. The DNA sequence that encodes a non-coding RNA is called a gene if its regulatory function can be defined. Thus it is possible that an isolated expression element lacking a gene structure is a non-coding, regulatory RNA. However, it was confirmed that the potential protein-coding genes found in this study did not match any RNA family published in the RNA-families database (www.sanger.ac.uk/ Software/Rfam/).

5. Conclusion

High-throughput gene finding on a newly sequenced genome is enabled through advanced computational genome annotation software. However, genome annotation does not guarantee all genes to be identified since knowledge and concepts about what constitutes a gene are evolving and yet to be perfected. Genome re-annotation using the same kind of computational heuristics offers limited help, unless supported with new *in vivo* experimental evidence, but such evidence often slowly arrives. We developed a method that integrated sequence-based and transcriptional information for gene finding in the intergenic regions of an annotated genome. In the experiment with the *M. tuberculosis* H37Rv genome, the method discovered genes with a specific function, such as a DNA-binding protein in the CopG family and a nickel binding GTPase, as well as hypothetical proteins that have not been reported in the *M. tuberculosis* H37Rv genome. This work has demonstrated that microarray-based transcriptional analysis could play an important role in gene finding on the genomic scale.

6. Acknowledgments

We would like to thank Dr. Thomas Shinnick at CDC for collaboration and the use of the facilities, and thank UCI for providing service for microarray hybridization. We thank Thomas R. Gingeras at Affymetrix, Inc. for designing *Mycobacterium tuberculosis* GeneChip. Bacterial culture and RNA isolation were performed by Pramod Aryal.

7. References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol 215(3): 403-10.

Besemer, J. and M. Borodovsky (2005). "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses." Nucleic Acids Res 33(Web Server issue): W451-4.

Burge, C. and S. Karlin (1997). "Prediction of complete gene structures in human genomic DNA." J Mol Biol 268(1): 78-94.

Camus, J. C., M. J. Pryor, C. Medigue and S. T. Cole (2002). "Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv." Microbiology 148(Pt 10): 2967-73.

Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, et al. (1998). "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence." Nature 393(6685): 537-44.

Erdmann, V. A., M. Z. Barciszewska, A. Hochberg, N. de Groot and J. Barciszewski (2001). "Regulatory RNAs." Cell Mol Life Sci 58(7): 960-77.

Finn, R. D., J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, et al. (2006). "Pfam: clans, web tools and services." Nucleic Acids Res 34(Database issue): D247-51.

Fisher, M. A., B. B. Plikaytis and T. M. Shinnick (2002). "Microarray analysis of the Mycobacterium tuberculosis transcriptional response to the acidic conditions found in phagosomes." J Bacteriol 184(14): 4025-32.

Fu, L. M. (2006). "Exploring drug action on Mycobacterium tuberculosis using affymetrix oligonucleotide genechips." Tuberculosis (Edinb) 86(2): 134-43.

Fu, L. M. and C. S. Fu-Liu (2007). "The gene expression data of Mycobacterium tuberculosis based on Affymetrix gene chips provide insight into regulatory and hypothetical genes." BMC Microbiol 7: 37.

Fu, L. M. and T. M. Shinnick (2007). "Genome-Wide Analysis of Intergenic Regions of Mycobacterium tuberculosis H37Rv Using Affymetrix GeneChips." EURASIP J Bioinform Syst Biol: 23054.

Fu, L. M. and T. M. Shinnick (2007). "Understanding the action of INH on a highly INH-resistant Mycobacterium tuberculosis strain using Genechips." Tuberculosis (Edinb) 87(1): 63-70.

Lee, J. M., S. Zhang, S. Saha, S. Santa Anna, C. Jiang and J. Perkins (2001). "RNA expression analysis using an antisense Bacillus subtilis genome array." J Bacteriol 183(24): 7371-80.

Li, J., M. Pankratz and J. A. Johnson (2002). "Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays." Toxicol Sci 69(2): 383-90.

Lukashin, A. V. and M. Borodovsky (1998). "GeneMark.hmm: new solutions for gene finding." Nucleic Acids Res 26(4): 1107-15.

Nielsen, P. and A. Krogh (2005). "Large-scale prokaryotic gene prediction and comparison to genome annotation." Bioinformatics 21(24): 4322-9.

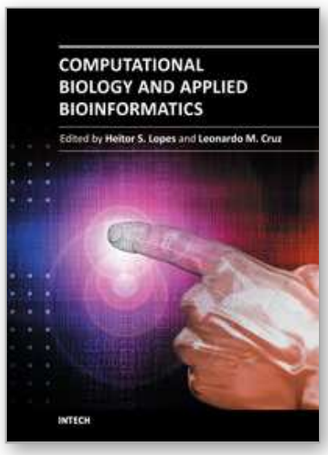
Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, et al. (2005). "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes." Nucleic Acids Res 33(17): 5691-702.

Pickford, A. S. and C. Cogoni (2003). "RNA-mediated gene silencing." Cell Mol Life Sci 60(5): 871-82.

Stothard, P. and D. S. Wishart (2006). "Automated bacterial genome analysis and annotation." Curr Opin Microbiol 9(5): 505-10.

Van Domselaar, G. H., P. Stothard, S. Shrivastava, J. A. Cruz, A. Guo, X. Dong, et al. (2005). "BASys: a web server for automated bacterial genome annotation." Nucleic Acids Res 33(Web Server issue): W455-9.

Zheng, D., Z. Zhang, P. M. Harrison, J. Karro, N. Carriero and M. Gerstein (2005). "Integrated pseudogene annotation for human chromosome 22: evidence for transcription." J Mol Biol 349(1): 27-45.



Computational Biology and Applied Bioinformatics

Edited by Prof. Heitor Lopes

ISBN 978-953-307-629-4

Hard cover, 442 pages

Publisher InTech

Published online 02, September, 2011

Published in print edition September, 2011

Nowadays it is difficult to imagine an area of knowledge that can continue developing without the use of computers and informatics. It is not different with biology, that has seen an unpredictable growth in recent decades, with the rise of a new discipline, bioinformatics, bringing together molecular biology, biotechnology and information technology. More recently, the development of high throughput techniques, such as microarray, mass spectrometry and DNA sequencing, has increased the need of computational support to collect, store, retrieve, analyze, and correlate huge data sets of complex information. On the other hand, the growth of the computational power for processing and storage has also increased the necessity for deeper knowledge in the field. The development of bioinformatics has allowed now the emergence of systems biology, the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of a living being. This book presents some theoretical issues, reviews, and a variety of bioinformatics applications. For better understanding, the chapters were grouped in two parts. In Part I, the chapters are more oriented towards literature review and theoretical issues. Part II consists of application-oriented chapters that report case studies in which a specific biological problem is treated with bioinformatics tools.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen