# AUTOMATIC ANNOTATION OF BACTERIAL COMMUNITY SEQUENCES AND APPLICATION TO INFECTIONS DIAGNOSTIC

Victor Solovyev[1], Asaf Salamov[2], Igor Seledtsov[2] Denis Vorobyev[2] and Alexander Bachinsky[2]
[1]*Department of Computer Science, Royal Holloway, University of London, London, U.K.*
[2]*Softberry Inc., 116 Radio Circle, Suite 400, Mount Kisco, NY 10549, U.S.A.*

Keywords: Bacterial community, Genome annotation, Sequence assembling, RNA-seq data, Computational pipelines, classification and diagnostic pathogenic bacteria, Tanscriptome analysis.

Abstract: To annotate bacterial sequences from an environmental sample, we have developed an automatic annotation pipeline Fgenesb_annotator that includes self-training of gene-finding parameters, prediction of CDS, RNA genes, operons, promoters and terminators. New version of pipeline includes frame shift corrections and special module with improved prediction accuracy of ribosomal proteins. To analyze next-generation sequencing data we have developed OligiZip assembler and Transomics pipeline that provide solutions to the following tasks: 1) de novo reconstruction of genomic sequence; 2) reconstruction of sequence with a reference genome; 3) SNP discovery; 4) mapping RNA-Seq data to a reference genome, assemble them into alternative transcripts and quantify the abundance of these transcripts. Using the OligoZip assembler and gene Fgenesb pipeline we have developed a novel computational approach of identification toxic and non-toxic bacterial serotypes using next-generation sequencing data. It can be used for detection of bacterial infections in wounds, water or food contamination.

## 1 INTRODUCTION

Appearance of metagenomic sequencing projects that sample the genomes of multiple species and strains with shorter average sequence fragment length, higher frequency of sequencing errors, and the phylogenetic heterogeneity of the organisms in the environmental sample, presents significant additional challenges in computational gene finding ((Venter et al., 2004), (Mavromatis et al., 2007), (Krause et al., 2007)). It is not surprising that some popular gene–finding algorithms demonstrate very poor performance on these currently abundant data (Krause et al., 2007). To overcome this problem a few new gene-finders have been developed such as GISMO (Noguchi et al., 2006), which is using Support Vector Machine for ORF classification and MetaGene (Lowe and Eddy, 1997), which is designed to predict genes from fragmented genomic sequences.

The bacterial genome annotation pipeline Fgenesb_annotator presented here was initially developed to provide a completely automatic and comprehensive annotation of huge amount of sequences generated by one of the first published metagenomic project (Venter et al., 2004). The pipeline includes a suit of original algorithms for protein coding gene identification, operon assignment and promoters and terminators identification. The parameters of gene prediction are automatically trained using uncharacterized genomic sequence. The pipeline also includes a module for rRNA genes identification

and incorporates well-established public programs such as tRNAscan-SE (Lowe and Eddy, 1997), and Blast (Altschul et al., 1997). Fgenesb_annotator is widely used to annotate sequences such as scaffolds of bacterial genomes or short reads of DNA extracted from a bacterial community ((Oliynyk et al., 2007), (Martin et al., 2006), (Frigaard et al., 2006), (Perez-Brocal et al., 2006), (Badger and Olsen, 1999)).

Recent introduction of next-generation sequencing instruments made possible sequencing

Solovyev V., Salamov A., Seledtsov I., Vorobyev D. and Bachinsky A..
AUTOMATIC ANNOTATION OF BACTERIAL COMMUNITY SEQUENCES AND APPLICATION TO INFECTIONS DIAGNOSTIC.
DOI: 10.5220/0003333703460353
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (Meta-2011), pages 346-353
ISBN: 978-989-8425-36-2

hundreds DNA of environmental sample at dramatically reduced cost. To analyze these data we have developed OligiZip assembler and visualization tools that provides effective solutions to the following three tasks: 1) ab initio reconstruction of genomic sequence; 2) reconstruction of sequence using a reference genome from the same or close organism; 3) mutation profiling and SNP discovery in a given set of genes. Using the OligoZip assembler and our metagenomics gene prediction pipeline FgenesB we introduce a novel computational approach for differentiation between toxic and non-toxic bacterial serotypes using next-generation sequences data that can be applied for rapid diagnostic of infections and environmental and food contamination.

# 2 LEARNING PARAMETERS AND PREDICTION OF PROTEIN-CODING GENES

Prediction of protein-coding genes depends on a set of parameters describing a bacterial gene model. Many of them such as, for example, codon usage are specific for each new genome. The Fgenesb algorithm includes software module, which iteratively learns these parameters using a given genomic sequence. It starts with a compilation of all relatively long ORFs (> 200 bp) that serve as an initial gene set for calculating organism-specific gene finding parameters. A few algorithms that use long ORFs (or ORFs with a significant match to a protein from a different organism) for deriving models of coding sequences have been described earlier ((Salzberg et al., 1998), (Larsen and Krogh, 2003),(Borodovsky et al., 1986). In the subsequent iterations we assign each potential ORF a score which reflects a joint probability to observe various features associated with protein-coding sequences such as coding content of a reading frame, oligonucleotide composition of sequence regions surrounding start and stop codons and the closeness of ORF's length to the length distribution of 'real' genes.

Protein coding sequences are modeled by 3-periodic fifth-order Markov-chains using 5th-order Markov transition probabilities computed for hexamers ending at each of the three codon positions. Non-coding sequences are also modeled using 5th-order Markov transition matrix computed for hexamers ending at any non-coding nucleotide. Markov models or Hidden Markov models using Markov transition probabilities of second or fifth-order have earlier been implemented in many bacterial gene-finding algorithms ((Larsen and Krogh, 2003), (Borodovsky et al., 1986), (Krogh et al., 1994)). Start of coding region including ribosome-binding site (RBS) is modeled by second order Markov model using second-order Markov transition probabilities computed for triplets ending at nucleotides in the upstream of the start codon sequence [-20 - +1] (here +1 is the position of the 1st nucleotide of start codon). To derive the log-likelihood ratios (start/nonstart probabilities) we also computed second-order Markov non-start transition matrix using sequences upstream of non-start triplets ATG, GTG, TTG and TGT. These probabilities have been averaged over windows of 5 bp long to account for variable location of RBS in 5'-gene region. The RBS weight matrix was applied in one of earlier works on bacterial gene finding to improve gene identification (Markowitz, 2007). To model the stop codon we used second-order Markov transition matrix computed for triplets ending at the positions +1 to +3 after the stop codon. To derive the log-likelihood ratios (stop/nonstop probabilities) we also computed second-order Markov non-start transition matrix computed for triplets ending at the positions +1 to +3 of TAA, TAG and TGA non-stop codon triplets. We also exploit a distribution of the ORF lengths that is accounted in combined scoring of potential protein-coding sequences.

If an ORF's start and end positions are L1 and L2 (excluding the stop codon), then the coding score of the ORF (S) is defined as

$$S = \sum_{i=L1}^{L2-5} llh(i) + \sum_{i=L1-20}^{L1} llww(i) + \log(P(L2-L1+1) + \sum_{i=L2+1}^{L2+3} llwm(i),$$

where llh(i) is the log-likelihood ratio for the probability of generating nucleotide X*i* in sequence position *i* by Markov model of coding region to the probability of generating nucleotide X*i* by Markov model of non-coding region computed using corresponding Markov transition matrices of the fifth-order described above; llww(i) is the log-likelihood ratio for the probability of generating nucleotide X*i* in sequence position *i* using Markov model of start region to the probability of generating nucleotide X*i* by Markov model of non-start region computed using corresponding Markov transition matrices of the second-order described above; P(L2-L1+1) is the probability of a coding region to have L2-L1+1 length; and llwm(i) is the log-likelihood ratio for the probability of generating nucleotide X*i* in sequence position *i* using Markov model of the stop codon region to the probability of generating nucleotide X*i* by Markov model of non-stop region computed using corresponding Markov transition matrices of the second-order described above.

At each iteration step the algorithm uses gene finding parameters produced on the previous stage and selects ORFs with the scores higher than some predefined threshold, sorts them in the descending order by score and accounts any selected ORF as a protein-coding gene if it overlaps with no more than 2 previous higher scoring genes and the fraction of the overlap sequence does not exceed 50% of the ORF length. Iteration stops when at least 99% of predicted genes between two successive iterations are the same. Optionally, we provide generic parameters to annotate unknown bacteria, archaea or their mixture sequences. These parameters have been used to annotate short sequences extracted from the environmental samples (Venter et al., 2004).

## 3 ACCURACY OF PROTEIN CODING GENE PREDICTION BY FGENESB

Fgenesb gene prediction engine is one of the most accurate prokaryotic gene finders. Its performance has been estimated on various datasets. Recently the pipeline performance has been tested on prediction of coding sequences for the representative set of assembled contigs (which are the typical results of bacterial sequencing) and as well as for the unassembled reads (Krause et al., 2007). The other pipeline in that study was a combination of Critica (Delcher et al., 1999) and Glimmer (Hayes and Borodovsky, 1998) (called here CG pipeline). The Fgenesb correctly identified 10-30% more reference genes on the contig sequences than the CG pipeline in every analyzed data set. The accuracy of the gene calls on unassembled reads (where many species are usually represented) was also evaluated. Fgenesb correctly identified ~70% and missed ~20% of reference genes on unassembled reads in all data sets. The CG pipeline exhibited significantly more poor results (~7% accurately predicted and 85% missed genes) (Krause et al., 2007). Both pipelines predicted 7-10% of genes inaccurately.

To make another comprehensive test we downloaded sequences and annotations of complete genomes from the IMG database (Besemer et al., 2001). 216 genomes have been selected that have percent of “unusual” annotated start codons (i.e. not atg, gtg, tgt, ctg, att, atc, or ata) and stop codons (not tag, taa, tga) less than 0.5%. The accuracy data for Fgenesb, Metagene (Lowe and Eddy, 1997) and latest versions of GeneMarkS (Delcher et al., 2007) and Glimmer (Tyson et al., 2005) are presented in Table 1.

Table 1: Gene prediction accuracies for 216 bacterial complete genomes.

| | Sn (%) exact | Sp (%) | Sn (%) exact+overlapping |
|---|---|---|---|
| Fgenesb | 76.8 | 91.2 | 96.0 |
| GeneMarkS | 73.5 | 90.7 | 96.5 |
| Metagene* | 76.8 | 93.7 | 95.9 |
| Glimmer | 73.3 | 88.5 | 95.2 |

*) Metagene uses a set of regression functions for estimation the coding scores, which have been computed using annotations of 170 bacterial genomes (many of them are among our test set sequences).

One more test of gene predictors we made for artificial shotgun sequences (700 bp fragments from a set of 216 bacterial genomes). We extracted these sequences with real genes covering only part of each sequence (by its 5’- or 3’-fragment) to make gene prediction more difficult. Correct gene assignment for these sequences was extremely high for all three gene finders (Table 2), while Metagene shows slightly lower specificity than Fgenesb and GeneMark.

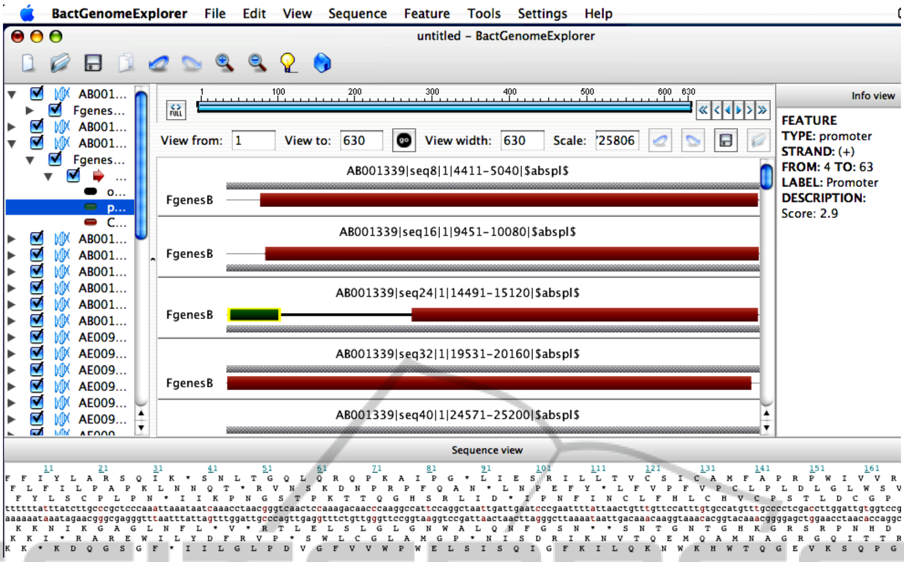
Table 2: Gene prediction accuracies for short bacterial sequences.

| | (Sn+Sp)/2 (%) | Sn (%) | Sp (%) |
|---|---|---|---|
| Fgenesb | 95.55 | 98.5 | 93.5 |
| GeneMark | 94.05 | 98.8 | 91.3 |
| Metagene | 91.65 | 97.6 | 82.9 |

Our results presented in Table 1 and 2 are in close agreement with the performance analysis of Metagene and GenMark tested on different data sets earlier (Lowe and Eddy., 1997).

The final Fgenesb_annotator results can be presented in the GenBank format to be readable by visualization software such as Softberry Bacterial Genome Explorer by server or local versions (Figure 1) or converted into the format that is required by SEQIN software for submission of new sequences with their annotations to the GenBank.

The Fgenesb_annotator has been applied in dozens of published studies for annotation of whole bacterial genomes, as well as short sequences


BactGenomeExplorer File Edit View Sequence Feature Tools Settings Help
untitled – BactGenomeExplorer
View from: 1 View to: 630 View width: 630 Scale: 25806
Info view
AB001339|seq8|1|4411-5040|$abspl$
FgenesB
AB001339|seq16|1|9451-10080|$abspl$
AB001339|seq24|1|14491-15120|$abspl$
AB001339|seq32|1|19531-20160|$abspl$
AB001339|seq40|1|24571-25200|$abspl$
Sequence view

FEATURE
TYPE: promoter
STRAND: (+)
FROM: 4 TO: 63
LABEL: Promoter
DESCRIPTION:
Score: 2.9

Figure 1: Bacterial Genome Explorer. Annotation of a set of matagenomics sequences is visualized.

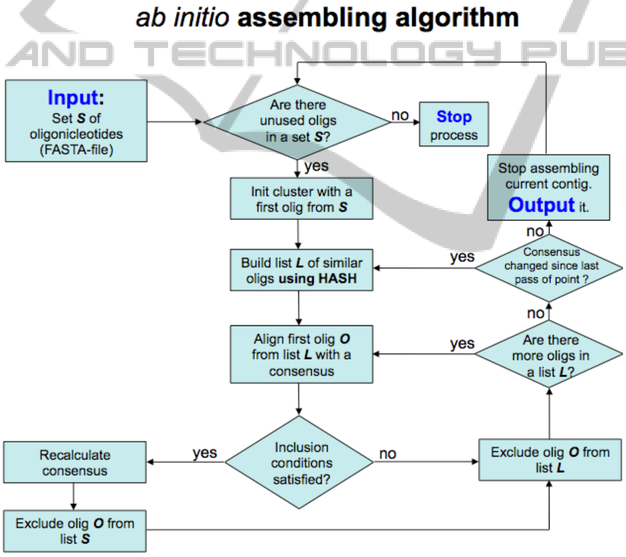

ab initio assembling algorithm
Input: Set S of oligonicleotides (FASTA-file)
Are there unused oligs in a set S?
no
Stop process
yes
Init cluster with a first olig from S
Build list L of similar oligs using HASH
Align first olig O from list L with a consensus
Inclusion conditions satisfied?
yes
Recalculate consensus
Exclude olig O from list S
no
Exclude olig O from list L
Are there more oligs in a list L?
yes
no
Consensus changed since last pass of point ?
yes
no
Stop assembling current contig. Output it.

Figure 2: OligoZip. de novo assembling algorithm.

extracted form environmental samples (for examples, ((Martinez et al., 2007), (McClain et al., 2009), (Yan et al., 2004)). The operons identified by Fgenesb_annotator have been used as initial operon models that were further experimentally investigated and improved ((Yan et al., 2006), (Pothier et al., 2007)). Other pipeline components such as Bprom (promoter predictor) and Bterm (terminator predictor) also have been actively used in many studies of functional characterization of bacterial sequences ((Budde et al., 2007), (Pilhofer et al., 2007), (Kosaka et al., 2006), (Grieshaber et al., 2006), (Warren et al., 2007)).

## 4 OLIGOZIP TOOLS FOR RECONSTRUCTING SEQUENCES FROM NEXT-GENERATION SEQUENCING DATA

Recent introduction of next-generation sequencing instruments: 454 (Roche), Genome Analyzer

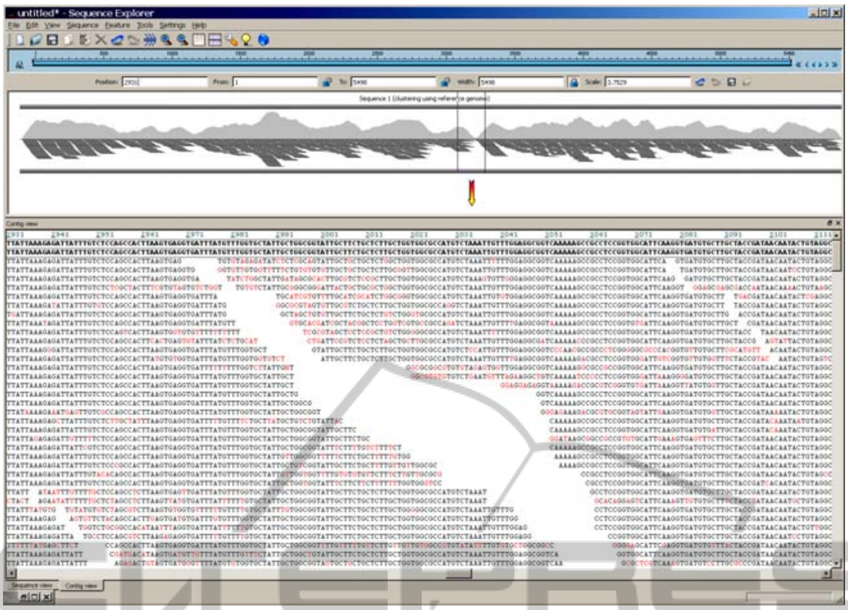

untitled* - Sequence Explorer
File Edit View Sequence Feature Tools Settings Help
Contig view
Sequence view
Contig view

Figure 3: Sequence Explorer to visualizing contigs and assembled reads.

(Illumina), ABI-SOLiD (Life Technologies), and several others, all capable of producing millions of DNA sequence reads in a single run, made possible sequencing hundreds of genes in parallel at dramatically reduced cost. We have developed a new algorithm (Fig.2) for processing short reads from next-generation sequencing machines. OligoZip tool uses L-plets hashing technique to achieve fast data processing, and it takes into account reads quality information.

OligoZip provides effective solutions the following tasks: 1) *De novo* reconstruction of genomic sequence; 2) Reconstruction of sequences based on reference genome from same or close species; 3)Mutation profiling and SNP discovery in a given set of genes.

De novo sequence reconstruction was tested on assembling several phage and bacterial genomes demonstrated its superior clustering power compared to earlier published approach [35]. Simulated error-free 25mers of bacteriophage φX174 and coronavirus SARS TOR2 were assembled perfectly and in reconstruction of Haemophilus influenzae genome contigs assembled by OligoZip were almost twice as long as those assembled by the published SSAKE software.

To test reconstruction of bacterial sequences using reference genome, we assembled genomic sequence of Methanopyrus kandleri TAG11 on known genome of Methanopyrus kandleri AV19. Solexa reads, about 6 million each for AV19 and TAG 11, were produced by sequencing lab of Harvard Partners HealthCare Center for Genetics and Genomics. AV19 genome itself has been assembled perfectly, with one extra contig that happened to be the genome of phage φX174. TAG 11 reads were assembled into several hundred contigs. Similar results were achieved on five other Methanopyrus stains. We also develop sequence assembling viewer to work with the reads data and assembling results interactively (Fig 3).

## 5 DIFFERENTIATION BETWEEN TOXIC AND NON-TOXIC BACTERIAL SEROTYPES

Using the OligoZip assembler and our metagenomics gene prediction pipeline FgenesB we have developed a novel computational approach of differentiation between toxic and non-toxic bacterial serotypes using next-generation sequences data.

Initially we found a way to construct a plot that separates well non-pathogenic and pathogenic genomes by using comparison of the bacterial proteomes predicted by Fgenesb (fig. 4).

Further, we have modeled next generation sequencing data extracted from the environmental sample. We generated a set of oligs 200 bp long for a mixture of 10 genomes. After that we tested how accurate a pathogenic genome can be identified in the mixture by aligning all oligs to a set of known

Table 3: Alignment of a genome oligs to the same or other genomes.

| Genome \ Oligs from | Gallinarum (nonpathogenic) | | Typhi_CT18 (pathogenic) | | Bacillus subtilis 168 | |
|---|---|---|---|---|---|---|
| | Hits | Total Score | Hits | Total Score | Hits | Total Score |
| Gallinarum | **215622** | **6407000** | 208461 | 5946000 | 67930 | 591066 |
| (nonpathogenic) | **206264** | **6340000** | 192903 | 5837000 | 15278 | 234807 |
| Typhi_CT18 | 210048 | 6171000 | **226918** | **7406000** | 70377 | 615816 |
| (pathogenic) | 198050 | 6087000 | **219868** | **7355000** | 16126 | 248026 |

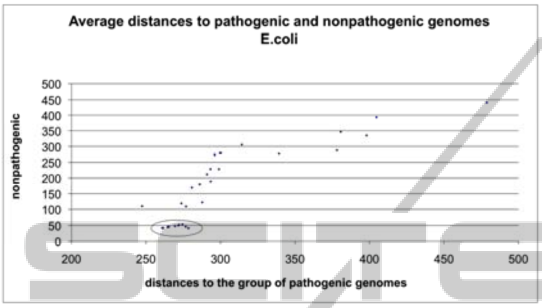


Figure 4: The plot of comparison of 32 known complete E.coli. genomes. All non-pathogenic genomes are cauterized in the circled region.

genomes. A typical example of alignment scores is presented in table 3 where significant alignments (hits) is counted to the same and different genomes. We computed the total number of hits for 2 score thresholds as well as total score of these hits. We can observe from the table that if a pathogenic genome is present in the mixture it will be recognized, while some bacterial strains have very similar genomes.

We are observing many common oligs between close pathogenic and non-pathogenic genomes. To increase sensitivity of classification of genomes from environmental sample we applied OligoZip for assembling a set of generated oligs from the genome mixture. OligoZip has produced ~ 100 contigs for each genome (with using 10 times coverage) or ~ 50 contigs for each genome (using 20 times coverage). These contigs cover ~ 99% of genome sequences. The contigs sizes range from 755000 bases to 2000 bp. Using sequences of these contigs we can identify the genomes from the mixture with no ambiguity. Annotating these sequences by Fgenesb gene-finder we can compare proteome content of known genomes with proteins in the mixture contigs and localize the mixture genomes on the plot of pathogenic and non-pathogenic genomes.

Currently we are collaborating with Veterinary Laboratories Agency to receive next generation sequencing data of environmental samples. The describe approach can be applied for analyzing such DNA for detection of bacterial toxic and non-toxic serotypes in wounds, water or food contaminations.

In addition we implemented Transomics computational pipeline to work with next generation transcriptome data. The pipeline maps RNA-Seq data to a reference genome, assemble them into transcripts and quantify the abundance of these transcripts in particular datasets. The pipeline will include the following key components: alignment of RNA-Seq reads to genome, identification of alternative transcripts, (for eukaryotic genes) and measuring expression levels of transcript isoforms. This pipeline also can be applied to analysis of transcriptomics data to detect pathogenic viruses or bacteria.

Fgenesb, OligoZip or Transomics pipeline components and other software programs are available to run independently at www.softwberry.com or as a part of integrated environment of the Molquest software package that can be downloaded at www.molquest.com for Windows, MAC and Linux OS.

# REFERENCES

Venter, J. C., Remington, K., Heidelberg, J., Halpern, A., Rusch, D., Eisen, J., Wu, D., Paulsen, I., Nelson, K., Nelson, W., *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science,* 304, 66–74.

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman,E., McHardy, A. C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., *et al*. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nature Methods, 4, 495-500.

Krause, L., McHardy, A., Nattkemper, T., Puhler, A., Stoye, J., Meyer, F. (2007) GISMO - gene identification using a support vector machine for ORF classification. *Nucleic Acids Res.*, 35, 2, 540-549.

Noguchi, H., Park, J., Takagi, T. (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.,* 34, 19, 5623–5630.

Lowe, T. M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*., 25, 955-964.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., 25, 3389-3402.

Oliynyk, M., Samborskyy, M., Lester, J., Mironenko, T., Scott, N., Dickens, S., Haydock, S., Leadlay, P. (2007) Complete genome sequence of the erythromycin-producing bacterium Saccharopolyspora erythraea NRRL23338. *Nature Biotechnology,* 25, 447–453.

Martin, H., Ivanova, N., Kunin,V., Warnecke, F., Barry, K., McHardy, A., Yeates, C., He, S., Salamov, A., Szeto, S., *et al*. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature Biotechnology*, 24, 1263–1269.

Frigaard, N., Martinez, A., Mincer, T., DeLong, E. (2006) Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature*, 439, 847-850.

Perez-Brocal,V., Gil, R., Ramos, S., Lamelas, A., Postigo, M., Michelena, J., Silva, F., Moya, A., Latorre, A. (2006) A Small Microbial Genome: The End of a Long Symbiotic Relationship? *Science*, 314, 312-313.

Badger, J. H., Olsen, G. J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol*., 16, 512–524.

Delcher, A., Harmon, D., Kasif, S., White, O., Salzberg, S. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*., 27, 4636–4641.

Hayes,W., Borodovsky, M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res*., 8, 1154-1171.

Salzberg, S., Delcher, A., Kasif, S., White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*., 26, 544-548.

Larsen, T., Krogh, A. (2003) EasyGene – a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, 4, 21, 1-15.

Borodovsky, M. Y., Sprizhitskii, Y. A., Golovanov, E. I., Aleksandrov, A. A. (1986) Statistical patterns in primary structures of functional regions in E. coli genome: III. Computer recognition of coding regions. *Mol. Biol.*, 20, 1390-1398.

Krogh, A., Mian, I., Haussler, D. (1994) A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res*., 22, 4768–4778.

Frishman, D., Mironov, A., Mewes, H. W., Gelfand, M. (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res*., 26, 2941-2947.

Markowitz, V. M., Microbial genome data sources. *Curr. Opin. Biotechnol.* 18, 267–272 (2007).

Besemer, J., Lomsadze, A., Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*., 29, 2607-2618.

Delcher, A, Bratke, K., Powers, E., Salzberg, S. (2007)

Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23, 6, 673–679.

Tyson, G., Lo, I., Baker,B., Allen, E., Hugenholtz, P., Banfield, J. (2005) Genome-directed isolation of the key nitrogen fixer Leptospirillum ferrodiazotrophum sp. nov. from an acidophilic microbial community. *Appl. Envir. Microbiol*., 71, 6319-6324.

Martinez, A., Bradley, A. S., Waldbauer, J. R., Summons, R. E., and DeLong, E. F. (2007) Proteorhodopsin photosystem gene expression enables photophosphorylation in a heterologous host. *PNAS,* 104, 5590-5595.

McClain, M., Shaffer, C., Israel, D., Peek, R., Cover, T. (2009) Genome sequence analysis of Helicobacter pylori strains associated with gastric ulceration and gastric cancer. *BMC Genomics,* 10, 3.

Yan, B., Methé, B. A., Lovley, D. R., Krushkal, J. (2004) Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family Geobacteraceae. *J. Theor. Biol.*, 230, 133-144.

Yan, B., Núñez, C., Ueki, T., Esteve-Núñez, A., Puljic.M., Adkins, R. M., Methé, B. A., Lovley, D. R., Krushkal, J. (2006). Computational prediction of RpoS and RpoD regulatory sites in Geobacter sulfurreducens using sequence and gene expression information. *Gene,* 384, 73-95.

Pothier, J. F., Wisniewski-Dye, F., Weiss-Gayet, M., Loccoz, Y., Prigent-Combaret, C. (2007) Promoter-trap identification of wheat seed extract-induced genes in the plant-growth-promoting rhizobacterium Azospirillum brasilense Sp245. *Microbiology*, 153, 3608–3622.

Gil, H., Platz, G. J., Forestal, C. A., Monfett, M., Bakshi, C., Sellati, T. J., Furie. M. B., Benach, J. L., Thanassi, D. G. (2006) Deletion of TolC orthologs in Francisella tularensis identifies roles in multidrug resistance and virulence. *PNAS,* 103, 12897-12902.

Michel, G. P., Durand, E., Filloux, A. (2007) XphA/XqhA, a Novel GspCD Subunit for Type II Secretion in Pseudomonas aeruginosa. *J. Bacteriol*., 189, 3776-3783.

Budde, P., Davis, B., Yuan, J., Waldor, M. (2007) Characterization of a higBA Toxin-Antitoxin Locus in Vibrio cholerae. *J. Bacteriol*., 189, 491-500.

Pilhofer, M., Bauer, A., Schrallhammer, M., Richter, L., Ludwig, W., Schleifer, K., Petroni, G. (2007) Characterization of bacterial operons consisting of two tubulins and a kinesin-like gene by the novel Two-Step Gene Walking method. *Nucleic Acids Res*., 35, e135.

Kosaka, T., Uchiyama, T., Ishii, S., Enoki, M., Imachi, H.,

Kamagata, Y., Ohashi, A., Harada, H., Ikenaga, H., and Watanabe, K. (2006) Reconstruction and Regulation of the Central Catabolic Pathway in the Thermophilic Propionate-Oxidizing Syntroph Pelotomaculum thermopropionicum. *J. Bacteriol*., 188, 202-210.

Grieshaber, N. A., Grieshaber, S. S., Fischer, E. R., Hackstadt, T. (2006) A small RNA inhibits translation of the histone-like protein Hc1 in Chlamydia trachomatis. *Mol. Microbiol*., 59, 541-550.

Warren R., Sutton G, Jones S and Holt R. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 2007, 23(4), 500-501.