An iterative strategy combining biophysical criteria and duration hidden Markov models for structural predictions of Chlamydia trachomatis σ66 promoters

Mallios, Ronna Reuben

UNIVERSITY OF CALIFORNIA, MERCED

An iterative strategy combining biophysical criteria and duration hidden Markov models for structural predictions of *Chlamydia trachomatis* σ^{66} promoters

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Quantitative Systems Biology

by

Ronna Reuben Mallios

Committee in charge:

Professor David Ojcius, Chair
Professor Michael Colvin
Professor David Ardell
Professor Miriam Barlow
Professor Jason Raymond

2010

Dedication

This dissertation is dedicated to Gabriella Eireen Mallios,

Annabelle Dahl Mallios and Samuel Zhenhua Dahl Mallios

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am grateful to California tax payers and the UC Regents for building a University of California in the Central Valley, allowing me to fulfill an adulthood dream. The pioneering faculty and administrators of UC Merced have expeditiously provided California's newest university with quality, state-of-the-art programs. I thank Professors David Ojcius and Michael Colvin, and Provost Keith Alley for accepting me, a non-traditional graduate student, into the Quantitative Systems Biology Program. As advisor and committee chair, Professor Ojcius has been steadfastly supportive, generous, and insightful. I also thank Professor David Ardell for sharing his profound expertise in bioinformatics and for his dedication to producing exemplary scientific research. Appendix K is a reprint of the article co-authored by Professors Ojcius, Ardell and myself which describes a major portion of the research presented here. Thanks also to Maria Avila and her student assistant Amanda Fox who put in countless hours perfecting the 5' RACE assay.

At the University of California, San Francisco – Fresno, Dr. Donna Hudson has been my Director, mentor and friend for 27 years. Were it not for her beneficent management, I would not have had the opportunity to pursue three advanced degrees during that time. I am also indebted to Hossein Moazamipour, CTO, for keeping many computers up and running through the years.

My family is my inspiration. My mother, Betty Reuben, earned a Bachelor's and Master's degree after raising four children; and my father, Jack Reuben, received his PhD degree after retiring from civil service. My husband, William, was the first PhD in the family, studying experimental statistics with legends in the field. He has

been the ultimate consultant, giving me confidence and pride in my data analysis. Our sons, Peter and Seth, continue to amaze me with their accomplishments and humanity – academically, professionally and personally. And together with their wives, Elizabeth Dahl and Gretchen Mallios, they have brought us three precious grandchildren – Annabelle, Sam and Gabriella, who were all born while this dissertation was being conceived. It is my fervent hope that opportunities for affordable high-quality public university education, especially at the University of California, will be as readily available to them as it was to their parents, grandparents and great grandparents.

VITA

Education

1960-61	University of California, Santa Barbara	Mathematics
1961-64	University of California, Los Angeles	Bachelor of Science
1981-82	University of New Mexico	Computer Science
1982-87	California State University, Fresno	Master of Science
1999-01	California State University, Fresno	Master of Public Health
2005-10	University of California, Merced	Doctor of Philosophy

Positions and Employment

1964-66	Booz, Hamilton Applied Research	Programmer/Analyst
1976-78	Carmel Unified School District	GATE Teacher
1979-80	Santa Catalina School, Monterey	Mathematics Teacher
1982-00	UCSF-Fresno Medical Information Resources	Programmer/Analyst III
2000- present	UCSF-Fresno Grants and Research Office	Research Analyst

Publications

[1-14]

1. Pantoja Zuzuarregui, J.R., R. Mallios, and J. Murphy, *The effect of obesity on kidney length in a healthy pediatric population.* Pediatr Nephrol, 2009. **24**(10): p. 2023-7.
2. Mallios, R.R., D.M. Ojcius, and D.H. Ardell, *An iterative strategy combining biophysical criteria and duration hidden Markov models for structural*

predictions of Chlamydia trachomatis sigma66 promoters. BMC Bioinformatics, 2009. **10**: p. 271.

3. Leigh, H., H. Cruz, and R. Mallios, *Telepsychiatry appointments in a continuing care setting: kept, cancelled and no-shows.* J Telemed Telecare, 2009. **15**(6): p. 286-9.
4. Leigh, H., R. Mallios, and D. Stewart, *Teaching psychiatry in primary care residencies: do training directors of primary care and psychiatry see eye to eye?* Acad Psychiatry, 2008. **32**(6): p. 504-9.
5. Mallios, R.R., *An iterative approach to class II predictions.* Methods Mol Biol, 2007. **409**: p. 341-53.
6. Leigh, H., D. Stewart, and R. Mallios, *Mental health and psychiatry training in primary care residency programs. Part II. What skills and diagnoses are taught, how adequate, and what affects training directors' satisfaction?* Gen Hosp Psychiatry, 2006. **28**(3): p. 195-204.
7. Leigh, H., D. Stewart, and R. Mallios, *Mental health and psychiatry training in primary care residency programs. Part I. Who teaches, where, when and how satisfied?* Gen Hosp Psychiatry, 2006. **28**(3): p. 189-94.
8. Mallios, R.R., *A consensus strategy for combining HLA-DR binding algorithms.* Hum Immunol, 2003. **64**(9): p. 852-6.
9. Mallios, R.R., *Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm.* Bioinformatics, 2001. **17**(10): p. 942-8.
10. Mallios, R.R., *Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm.* Bioinformatics, 1999. **15**(6): p. 432-9.
11. Mallios, R.R., *Iterative stepwise discriminant analysis: a meta-algorithm for detecting quantitative sequence motifs.* J Comput Biol, 1998. **5**(4): p. 703-11.
12. Mallios, R.R., *An iterative algorithm for converting a class II MHC binding motif into a quantitative predictive model.* Comput Appl Biosci, 1997. **13**(3): p. 211-5.
13. Mallios, R.R., *Multiple regression analysis suggests motifs for class II MHC binding.* J Theor Biol, 1994. **166**(2): p. 167-72.
14. Mallios, R.R., *Predicting the probability of helper T cell immunodominant sites through discriminant analysis.* Ann Clin Biochem, 1993. **30 (Pt 2)**: p. 152-6.

Abstract

Promoter identification is crucial for understanding gene regulation in bacteria. It has been demonstrated that the initiation of bacterial transcription depends upon the stability and topology of DNA in the promoter region as well as the binding affinity between the RNA polymerase σ-factor and promoter. However, promoter prediction algorithms to date have not explicitly used an ensemble of these factors as predictors. In addition, most promoter models have been trained on data from *Escherichia coli*. Although it has been shown that transcriptional mechanisms are similar among various bacteria, it is quite possible that the differences between *Escherichia coli* and *Chlamydia trachomatis* (*C. trachomatis*) are large enough to recommend an organism-specific modeling effort for *C. trachomatis*.

The intracellular life-cycle of Chlamydiae impedes the study of gene regulation. The bacteria are difficult to purify in large quantities and are resistant to standard genetic manipulation techniques. Consequently, less than 40 *C. trachomatis* σ^{66} promoters had been mapped at the inception of this study. Utilizing 29 of these experimentally identified promoters as a training set, this research develops an iterative model building procedure that combines such biophysical metrics as DNA stability, curvature, twist and stress-induced DNA duplex destabilization along with duration hidden Markov model parameters to model *C. trachomatis* σ^{66} promoters. The resulting model, MMCTPP1 (Multiple Metric *Chlamydia Trachomatis* Promoter Prediction), predicts the training set with a high degree of accuracy and provides insights into the structure of the promoter region.

MMCTPP1 *C. trachomatis* genome-wide predictions are provided, as well as co-predictions with three other algorithms. The substantial overlap between MMCTPP1 predictions and others bolsters the credibility of all four algorithms.

To validate the genome-wide predictions, 317 recently mapped transcription start sites of annotated *C. trachomatis* genes were combined with predictions from MMCTPP1 and TSS-PREDICT. The result maps 169 *C. trachomatis* σ^{66} promoters, yielding a four-fold increase in established promoters. These will assist researchers in studying gene regulation in *C. trachomatis* and enhance the training set for the development of MMCTPP2. This second generation multiple metric model will predict *C. trachomatis* σ^{66} promoters with increased accuracy and reveal a more refined characterization of structural features.

Chapter 1

Introduction

1.1 Background

Identifying mechanisms that regulate gene expression in bacteria is essential for understanding and eventually controlling their pathogenicity. Bacterial gene expression emanates from the transcription process whereby bacterial RNA polymerase (RNAP) enzymes synthesize messenger RNA via DNA templates. To initiate transcription, the RNAP slides along double stranded DNA until it binds to a promoter region. The promoter itself ranges in length from 27 to 33 nucleotides (nt). Hence, promoter identification is a first step in the quest to explain gene regulation in bacteria.

All known bacteria share a well conserved RNAP [1]. This transcriptional enzyme is comprised of a 3-subunit catalytic core plus a variable σ-factor subunit that provides DNA binding specificity. After a σ-factor joins the catalytic core, the resulting RNAP holoenzyme searches the DNA for a promoter that matches the specificity of the σ-factor. Once transcription is successfully initiated, the σ-factor leaves the complex, allowing the core enzyme mobility to proceed with transcript elongation.

The aim of this research is to answer the question, “Where’s the promoter?” Similar to looking for Waldo in a children’s book, this task involves searching for a pattern among configurations that appear similar. Figure 1 illustrates the challenge of identifying a promoter pattern within the upstream region of a bacterial gene.

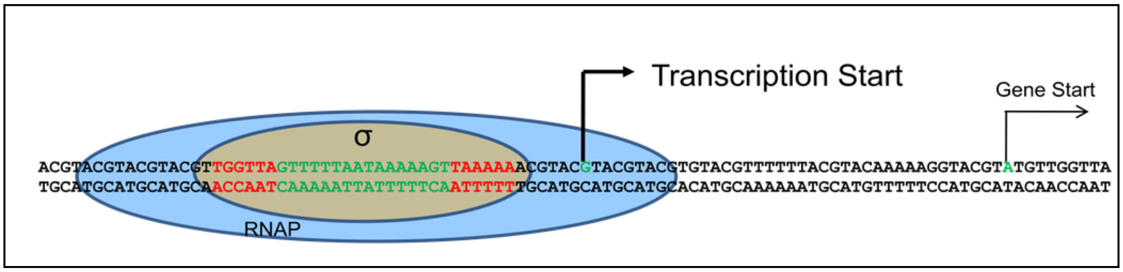

Transcription Start
Gene Start
σ
ACGTACGTACGTACGTTGGTTAGTTTTTAATAAAAAGTTAAAAAACGTACGTACGTACGTGTACGTTTTTTACGTACAAAAAGGTACGTATGTTGGTTA
TGCATGCATGCATGCAACCAATCAAAAATTATTTTTCAATTTTTTGCATGCATGCATGCACATGCAAAAAATGCATGTTTTTCCATGCATACAACCAAT
RNAP

Figure 1. Research aim: To find promoters (red) in the regions upstream from gene start sites.

E. coli promoter sequences have been studied since 1983 when Hawley and McClure [2] catalogued 112 promoters that were well-defined by genetic criteria (promoter mutations) or biochemical criteria (determination of the 5' terminal nucleotides of the mRNA transcript). Once the transcription start sites (TSS) were determined, the sequences were aligned to maximize the homology to a previously proposed consensus sequence for *E. coli* promoters consisting of two hexamers [3, 4]: TTGACa centered ~35 nt upstream from the transcription start site, and TAtAaT centered ~10 nt upstream from the TSS. The two hexamers mark the sites in the DNA where the σ-factor binds. The spacer region between hexamers in the 112 catalogued promoters varied from 16 to 18 nts.

The *E. coli* promoter list was updated to 300 by Lisser and Margalit in 1993 [5]. Because σ^{70} participates in the transcription of a majority of *E. coli* genes including those with housekeeping functions, σ^{70} promoters dominate the literature. Currently there are over 700 known *E. coli* σ^{70} promoters [6]. Over the years, many promoter prediction algorithms have been developed and disseminated, most of them based upon *E. coli* σ^{70} binding sites.

In 1996 Hertz and Stormo observed **"…** the polymerase needs to bind the DNA, open the DNA, initiate transcription, and release the promoter for elongation [7]**."** Since then, it has been demonstrated that the initiation of bacterial transcription depends upon the stability and topology of DNA in the extended promoter region, as well as the binding affinity between the RNAP σ-factor and promoter. Specifically, evidence from the profiling of DNA curvature, bendability, twist, stability and propensity for stress-induced destabilization in *E. coli*, *B. subtilis*, *C. trachomatis*, plants and vertebrates [8-10] suggests that there are peaks for these measures near the TSS.

Although biophysical metrics have been analyzed individually with respect to transcription initiation, promoter prediction algorithms to date have not explicitly used an ensemble of these factors as predictors. In fact, the predictors of an optimal promoter algorithm would most likely include RNAP σ-factor/DNA binding propensity and multiple biophysical metrics of the extended promoter region.

Bacteria of the genus *Chlamydia* are obligate intracellular parasites that were genetically isolated from other bacteria nearly a billion years ago when they moved into their intracellular environment [11]. In humans, *Chlamydia* infections are responsible for infertility, blindness, arthritis and cardiovascular disease [12]. *C. trachomatis* is the main cause of preventable blindness in the world.

Chlamydiae have a biphasic life-cycle. The infectious elementary body (EB) is metabolically inert outside of a host cell. Upon entrance into a host cell, the EB is enveloped by a vacuole, termed an inclusion. Within the inclusion the EB converts to a reticulate body (RB) which is metabolically active and replicates. Eventually, the RBs revert to EBs, the host cell lyses, and the EBs are freed to re-infect.

The chlamydial life-cycle impedes the study of gene regulation in two ways. First, unlike free-growing bacteria, chlamydiae are difficult to purify in large quantities; and second, there are no genetic tools to manipulate the chlamydial genome [13]. A heterologous in vivo transcription system in *E. coli* was developed where a hybrid holoenzyme was formed from a chlamydial σ-factor expressed from a plasmid and native *E. coli* core enzyme [14]. Although this system was somewhat successful, less than 40 *Chlamydia trachomatis* promoters had been experimentally identified at the inception of this study [1, 15-17].

1.2 Motivation

Why, then, should an organism-specific promoter model be developed for *Chlamydia trachomatis*?

1. *Chlamydia trachomatis* (*C. trachomatis*) is a major cause of

    a. bacterial sexually transmitted diseases leading to pelvic inflammatory disease, infertility, and infections in newborns; and

    b. ocular infections leading to trachoma/conjunctivitis and blindness.

2. Surveys of known bacterial promoters suggest that their structures are relatively diverse [18]. In particular, some established *C. trachomatis* promoters display obvious differences from the established consensus hexamers of *E. coli* [1, 15-17].

3. Although σ^{66}, the *C. trachomatis* analogue of *E. coli* σ^{70}, has DNA binding domains homologous to the DNA binding domains of σ^{70}, sequence based phylogenetic

analysis of bacterial RNAP subunits has shown discernable evolutionary distance between the *C. trachomatis* and *E. coli* RNAP in all four subunits [19].

4. Although standard genetic manipulation techniques are insufficient to study *C. trachomatis* gene regulation, *C. trachomatis* makes an excellent candidate for *in silico* analysis because of its small genome of ~1 Mbp and 895 genes.
5. Most existing promoter models over-predict, predicting many false positives.
6. This would be an opportunity to explore promoter models where the predictors include RNAP σ-factor/DNA binding propensity and multiple biophysical metrics of the extended promoter region.

Together, all of these reasons make it is plausible that an organism-specific model is appropriate for *C. trachomatis*.

1.3 Research aims

This study aims to address the following questions:

1. Can established *C. trachomatis* σ^{66} promoters be accurately predicted by a strategy that employs a small training set of *C. trachomatis* σ^{66} promoters?
2. If so, do higher order DNA structures within the extended promoter region, as well as the primary structure (sequence) of the promoter, contribute to the predictive model?
3. Do *C. trachomatis* genome-wide predictions based on the study model facilitate the identification of new *C. trachomatis* σ^{66} promoters?

This research also aims to set the groundwork to address the following question:

4. Do *C. trachomatis* σ^{66} promoters differ significantly from *E. coli* σ^{70} promoters?

1.4 Research design

1.4.1 Initial training set of 29 experimentally identified *C. trachomatis* σ^{66} promoters

A significant challenge for bioinformaticians is to model data that has been collected by multiple laboratories using different assays, protocols and equipment. This phenomenon is compounded in the study of *C. trachomatis* where the organism is metabolically active only inside an infected host-cell. One way to minimize the use of conflicting and/or controversial data is to rely upon reviews written by informed biologists. For this reason, the reviews of Mathews & Timms (2006) [17] and Tan (2006) [1] were consulted to compile a list of 16 experimentally identified σ^{66} promoters. Added to this list are 13 promoters that were experimentally identified by Grech *et al* (2007) [15] and Hefty *et al* (2007) [16] after the previously cited reviews were written. Table 1 describes the 29 experimentally identified σ^{66} promoters from 27 genes that form the basis of the training set for this study.

Table 1. Training set: 29 experimentally identified *C. trachomatis* σ^{66} promoters.

CT	Name	Ref[a]	-35 Hex	Spacer (16-20)	-10 Hex	h PI[b]
CT046	*hctB*	M	TGGTTA	GTTTTTAATAAAAAGT(16)	TAAAAA	16
CT062	*tyrS*	G	TTGCTA	TAAAAAGAACAGGATAGA(18)	TAAGAT	8
CT080	*ltuB*	M,T	TTATGA	AAAACAATTTTTTAATT(17)	TAAAAT	24
CT091	*yscU*	H	TTGAGA	AAAACATTTATATACGG(17)	TAACTT	8
CT098	*rs1*	M,T	TTGCCT	TTTTTAAGGTGAATATT(17)	TACACT	3
CT111	*groES*	M,T	TTGCAA	AAAAGCGAGGACTTTGC(17)	TATCGT	1
CT286	*clpC*	G	TTGCAT	CATTATCATAAATGTCG(17)	TATATG	8
CT322	*tuf*	M,T	TTGATA	ATAATCCGCGTCTGAAGT(18)	TACTAT	3
CT323	*infA*	M,T	TTGACA	TTTTCTGTTTAGTCGA(16)	TATAAT	3
CT377	*ltuA*	M,T	TGCAGA	GTTTTTATTTTAAATATGT(19)	TATAAT	16
CT394	*hrcA*	M,T	TTGACC	AGTGGAGACGGTTTTCT(17)	TATAAT	16
CT439m	*rpsL*	G	TTGCAA	ACAAAGATATTCTTATTC(18)	TATATT	3
CT442	*crpA*	M	GGGTTT	TTGAAAAAAACAAGTGTTT(19)	GTGTAG	16
CT444a	*omcA*	M,T	TTGATA	TAATTTTTATTTTATAA(17)	TGTAAT	16
CT444b	*omcA*	M,T	AATTGC	TTTTATCGATAAAAGAAAC(19)	TTCAAG	16
CT518	*rl14*	M	CTGTTG	TTGTTCGAGTCGAAAGGG(18)	TATACT	3
CT557	*lpdA*	H	TTGAGA	TTTTATCCACCCAGATG(17)	TACAAC	8
CT559	*yscJ*	G	TTGGCA	CTAATCTCCCCATTTGC(17)	TATGGT	16
CT576	*lcrH_1*	H	TTGTTA	AATCAGATCGTTAGAATT(18)	TAATAT	16
CT596	*exbB*	G	TTGGTT	CTATACAAGAAATTTGT(17)	TAGGAT	3
CT665	-	H	TTGTAT	CTTTTTAGAACGGGAAGGG(19)	TTGAAA	8
CT674	*yscC*	H	TTGCAA	GATAGAGGGCAAATAGA(17)	TATATT	16
CT681a	*ompA*	M,T	TATACA	AAAATGGCTCTCTGCTT(17)	TATTGC	8
CT681b	*ompA*	M,T	GTGCCG	CCAGAAAAAGATAGCGAG(18)	CACAAA	8
CT701	*secA_2*	M	TGTATA	GGCGCCTTTAAATAAGAGGG(20)	TAGGTT	8
CT708	-	G	TTGATT	TAGCGGAAGTAAAAAGG(17)	TACAAG	16
CT743	*hctA*	M,T	TTGCAT	GAATTTGAACAAACAAAC(18)	TAATTA	24
CT752	*efp_2*	G	TGGACA	AAGCTTAGAAGAGAACGA(18)	TAACAT	8
CT863	-	H	TTGCAT	GAAAAATACTTTTTAGA(17)	TAAGTT	16

[a]References: M: Mathews & Timms [17]; G: Grech *et al* [15]; H: Hefty *et al* [16]; T: Tan [1]

[b]hour Post Infection of transcriptional activation [20]

1.4.2 Research design: 3 degrees of iteration

1.4.2.1 Outermost iteration: project overview

The study design incorporates three levels of iteration. Figure 2 provides an overview of the outermost cycle. Each pass through the outermost cycle produces a Multiple Metric *Chlamydia Trachomatis* Promoter Prediction (MMCTPP) model, MMCTPP1 being the first. The research reported here chronicles steps 1-5 of the first level of iteration.

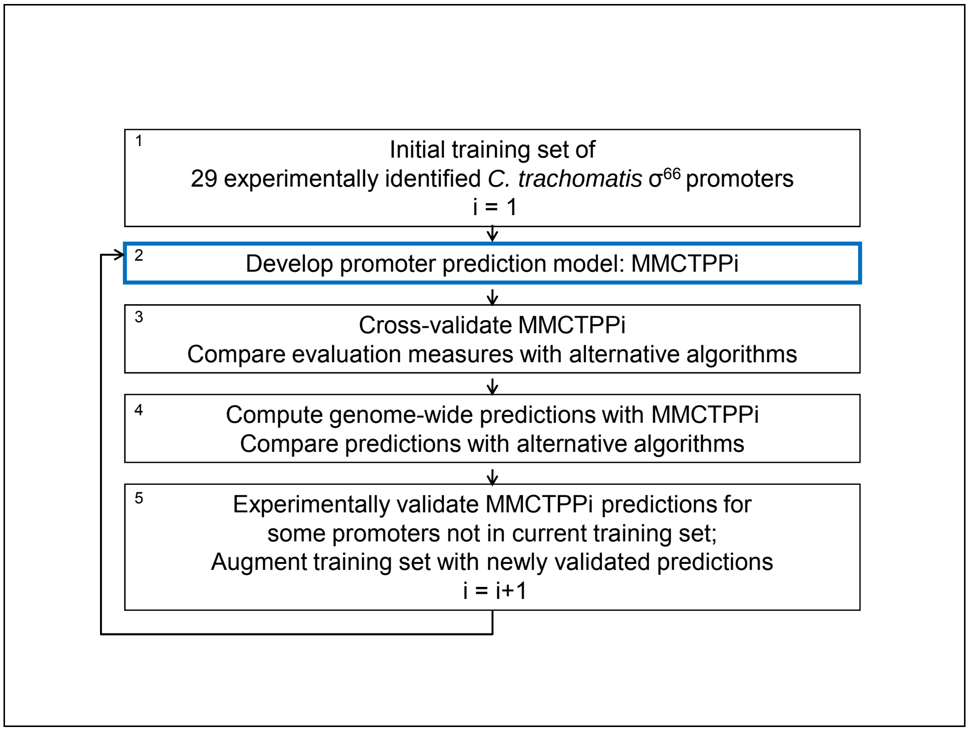

1
Initial training set of
29 experimentally identified C. trachomatis σ66 promoters
i = 1
2
Develop promoter prediction model: MMCTPPi
3
Cross-validate MMCTPPi
Compare evaluation measures with alternative algorithms
4
Compute genome-wide predictions with MMCTPPi
Compare predictions with alternative algorithms
5
Experimentally validate MMCTPPi predictions for
some promoters not in current training set;
Augment training set with newly validated predictions
i = i+1

Figure 2. Project overview.

Twenty-nine experimentally identified *C. trachomatis* σ^{66} promoters from the scientific literature form the initial training set. The second level of iteration produces MMCTPP1 (blue box) and is described next in section 1.4.2.2. Evaluation measures of MMCTPP1 are calculated via stratified K-fold cross-validation and compared with corresponding measures from three alternative algorithms. MMCTPP1 is then used to predict promoters for all 895 genes in the *C. trachomatis* genome. It was planned that these predictions, along with the predictions from two comparable algorithms, would be used to select candidates for experimental testing with 5'-rapid amplification of cDNA ends (5'-RACE). Confirmed promoters would then be added to the training set of experimentally identified promoters. However, a timely study using high-throughput RNA deep sequencing identified 371 *C. trachomatis* transcription start sites which could then be integrated with promoter predictions to identify new *C. trachomatis* σ^{66} promoters.

1.4.2.2 Second level of iteration: promoter prediction model

Sources of error that could lead to unreliable prediction models include: (i) imprecise laboratory procedures in defining and identifying promoters (including false positive promoters), (ii) presence of more than one promoter population, (iii) failure to include relevant predictor variables, and (iv) random variation. The purpose of the second level of iteration is to remove members of the training set that appear to be outliers. The reason for the "bad fit" might be experimental error or that the promoter belongs to a different family of promoters.

Figure 3 outlines the second level of iteration. In this initial study, the training set referred to in step 1 is the same initial training set of 29; this set will be expanded in future studies. The first task is to develop a duration hidden Markov model (HMM) to characterize and quantify RNAP-σ^{66}/DNA binding (red box). This step involves the third level of iteration and is discussed next in section 1.4.2.3. The duration HMM provides predicted binding scores which, along with biophysical metrics, comprise the set of potential predictor variables for Stepwise Binary Logistic Regression (SBLR). The resultant SBLR multiple variable model is then used to predict the training set. Those members of the training set which fit very poorly are eliminated from the training set, and the process is repeated until the training set stabilizes. It is also possible to replace an eliminated member if there is supporting evidence.

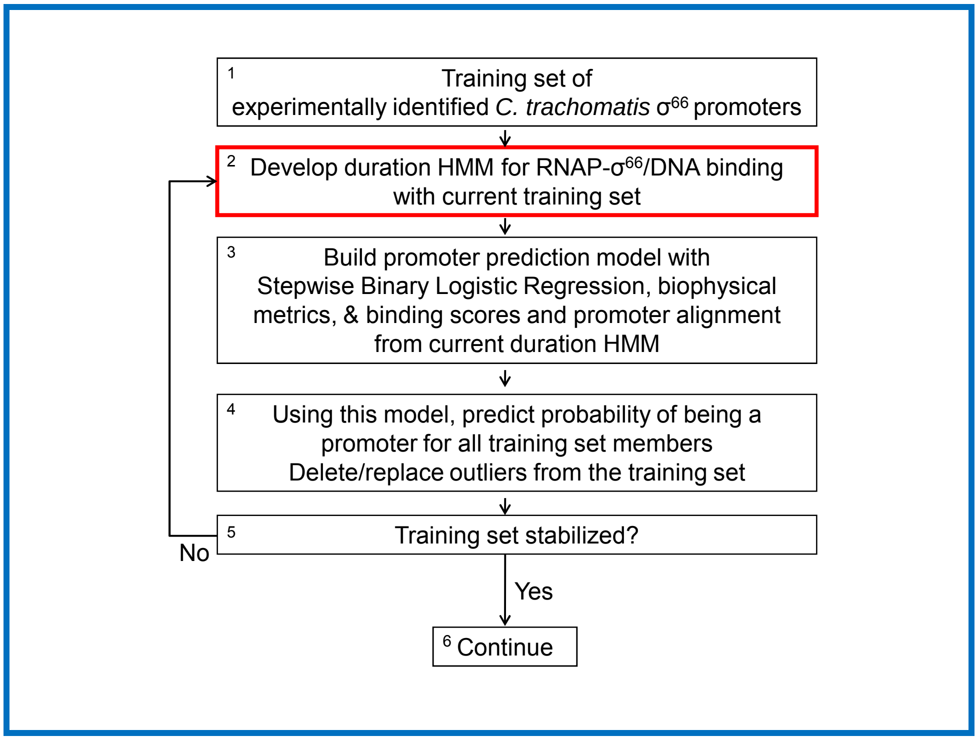

1 Training set of experimentally identified C. trachomatis σ66 promoters
2 Develop duration HMM for RNAP-σ66/DNA binding with current training set
3 Build promoter prediction model with Stepwise Binary Logistic Regression, biophysical metrics, & binding scores and promoter alignment from current duration HMM
4 Using this model, predict probability of being a promoter for all training set members
Delete/replace outliers from the training set
5 Training set stabilized?
No
Yes
6 Continue

Figure 3. Promoter prediction model iteration.

1.4.2.3 Third level of iteration: duration HMM to quantify RNAP-σ^{66}/DNA binding

Duration HMMs are used here to characterize and quantify RNAP-σ^{66}/DNA binding because they efficiently accommodate variable spacer regions between hexamers. However, the training set must define the hexamers. It is possible for a promoter region to be identified correctly but the hexamers imprecisely defined. The purpose of the third level of iteration is to refine the alignment of hexamers so as to optimize the duration HMM scores.

Figure 4 describes the third level of iteration. The input training set is that which is currently being analyzed by the second level of iteration. This level does not modify the census of the training set, only slightly realigns some members. The process always begins with the members of the training set in their original alignment as documented in Table 1. In step 2, a duration HMM is built with the current alignment and the resultant model is used to adjust the hexamers within the originally defined promoter or slightly towards either end (step 3). After some hexamers are realigned, the process continues with a new duration HMM until the configuration of the promoters stabilizes.

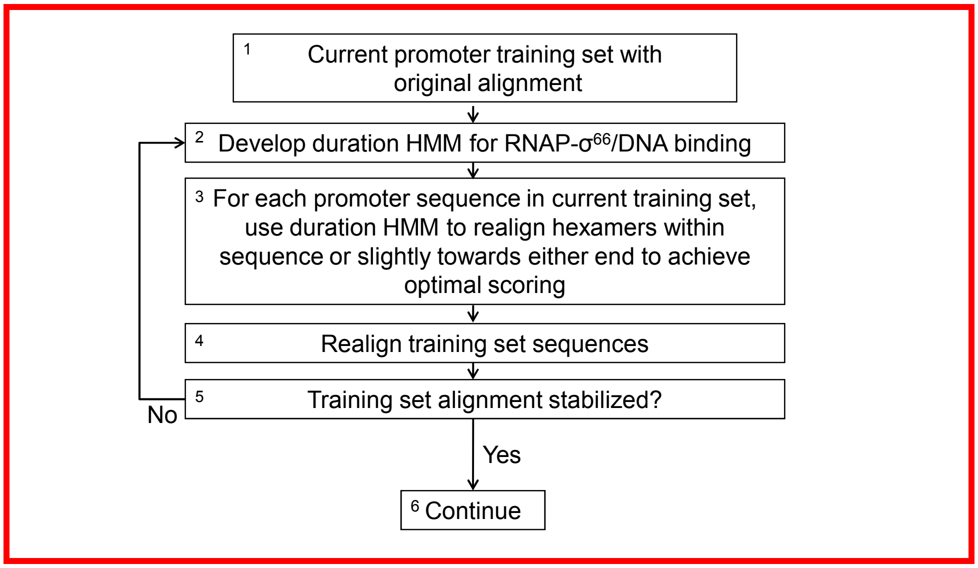

1 Current promoter training set with original alignment
2 Develop duration HMM for RNAP-σ66/DNA binding
3 For each promoter sequence in current training set, use duration HMM to realign hexamers within sequence or slightly towards either end to achieve optimal scoring
4 Realign training set sequences
5 Training set alignment stabilized?
No
Yes
6 Continue

Figure 4. Duration HMM for RNAP-σ^{66}/DNA binding iteration.

1.5 Manuscript organization

After presenting relevant algorithms for building promoter prediction models, implementation is described starting from the innermost iteration (duration HMM) toward the outside (promoter prediction model). The results of the model building strategy are presented, compared with alternative algorithms and discussed. Finally, genome wide predictions are presented, compared with those of alternative algorithms and used to collaborate in the identification of new *C. trachomatis* σ^{66} promoters.

Chapter 2

Algorithms for building promoter prediction models

Algorithms for separating DNA promoter sequences from non-promoter sequences can be categorized by how they define the null model. The first group, promoter vs background, compares position-specific nucleotide frequencies in a set of promoter sequences (promoter training set) with the background composition of the entire genome (null model). The second group, promoter vs non-promoter, utilizes a training set of non-promoter sequences (null model) as well as a training set of promoter sequences, and searches for features that separate the two populations.

The strategy presented here for building a promoter prediction model utilizes one algorithm from each group. From the first, a duration hidden Markov model (HMM) predicts RNAP-σ^{66}/DNA binding; and from the second, a Stepwise Binary Logistic Regression model predicts *C. trachomatis* σ^{66} promoters utilizing multiple predictors that include the duration HMM score and biophysical metrics.

Three algorithms are used as comparable algorithms to validate the results of this study and to identify co-predictions with the model developed here. The phylogenetic footprinting algorithm, Footy, utilizes a background composition null model, while the time delay artificial neural network NNPP2.2 and the support vector machine TSS-PREDICT use positive and negative training sets.

2.1 Promoter vs background algorithms

2.1.1 Position weight matrices

Position weight matrices (PWMs) were the first models to quantify the *E. coli* hexamer motifs [7, 21]. To illustrate the mechanics of PWMs, consider a hypothetical organism that has promoters with the pattern: trimer, spacer of length 1-3, trimer; and four hypothetical promoters TTGATAT, TTGCTAA, TGGCCTAA and TTAAACAAC. Table 2 displays the four promoters as sequences, aligned sequences, trimers and position-specific monomers.

Table 2. Hypothetical promoters.

Promoter	Aligned	Trimer 1	Trimer 2	M1	M2	M3	M4	M5	M6
TTGATAT	TTGA--TAT	TTG	TAT	T	T	G	T	A	T
TTGCTAA	TTGC--TAA	TTG	TAA	T	T	G	T	A	A
TGGCCTAA	TGGCC-TAA	TGG	TAA	T	G	G	T	A	A
TTATAGAAC	TTATAGAAC	TTA	AAC	T	T	A	A	A	C

A PWM is a matrix that quantifies a list of sequences of equal length such that each column refers to a position in the sequence and each row refers to a letter in the sequence alphabet. Each matrix element is necessarily a function of the frequency of the referent letter in the referent position. Table 3 shows the frequency matrix and the position-specific probability matrix (PSPM) for the four promoters. With regard to PWMs, probability is equivalent to relative frequency.

Table 3. PWMs: Frequency matrix and PSPM.

	Frequency matrix							PSPM					
	M1	M2	M3	M4	M5	M6		M1	M2	M3	M4	M5	M6
A	0	0	1	1	4	2		0	0	.25	.25	1	.5
C	0	0	0	0	0	1		0	0	0	0	0	.25
G	0	1	3	0	0	0		0	.25	.75	0	0	0
T	4	3	0	3	0	1		1	.75	0	.75	0	.25

In *Chlamydia*, the background probability of both A and T is .285, and of both C and G is .215. The ratio of the observed probability to the background probability is referred to as the likelihood ratio or odds. It is a measure of how strongly an event occurs in the sequence set relative to the background. Table 4 displays the position-specific odds matrix and the position-specific $\log_2$ odds matrix for the sequence set. The asterisks in the log-odds matrix mark $\log_2(0)$. In practice, a pseudo-count is added to each cell to prevent this problem. If the pseudo-count = .001, the asterisk becomes $\log_2(.001) = -10.0$.

Table 4. PWMs: Odds matrix and log-odds matrix.

	PS odds matrix						PS log-odds matrix (PSSM)					
	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6
A	0	0	.88	.88	3.51	1.75	*	*	-.19	-.19	1.8	.81
C	0	0	0	0	0	1.16	*	*	*	*	*	.21
G	0	1.16	3.49	0	0	0	*	.21	1.8	*	*	*
T	3.51	2.63	0	2.63	0	.88	1.8	1.4	*	1.4	*	-.19

Once a PWM is constructed, it can be used for predictive purposes. The log-odds matrix is often used to score new sequences. It is then called a position-specific scoring matrix (PSSM). The score of a new sequence, length n, is defined by

$$\text{Score} = \sum_{\text{position } i=1}^{n} \log_2\left({}^{p_{i,\propto i}}\!/_{b_{\propto i}} \right)$$

where αi is the character (A,C,G or T) at position i of the new sequence and $\log_2(p_{i,\alpha i}$ $/b_{\alpha i})$ is the corresponding element of the PS log-odds matrix.

The score of a sequence with trimers TTT and AAA would then be 1.8+1.4-10-.19+1.8+.81 = -4.38. It quantifies how much the new sequence resembles the set used to create the PSSM relative to the background.

Among the challenges encountered by PWM models is defining a threshold for the score of a new sequence that is sensitive enough to include known promoters without predicting numerous false positives. PWM models have been expanded to quantify the variable-length spacer region between hexamers [7, 22, 23], but duration hidden Markov models are more appropriate.

2.1.2 Duration hidden Markov models (HMM)

Duration HMMs are natural extensions of PWMs that explicitly model the empirical spacing distribution between motifs. "Duration" refers to this explicit representation of a spacer length distribution, as opposed to the geometrically distributed lengths that are expected from components of profile hidden Markov models [24].

An HMM consists of states, transition probabilities between states, and probabilities that a given state will emit various symbols. To illustrate the mechanics of duration HMMs, we continue with the example above. The probability of transmission from one motif or match (M) state to another is always 1, because a trimer always proceeds from the first letter to the second, and from the second to the third. The emission probabilities for the M states are the corresponding columns of the PSPM. For the M states, the duration HMM behaves exactly like the PSSM.

The spacer regions of the example promoter set are described in Table 5. The spacer positions (S1, S2 and S3) correspond to spacer states as long as they emit letters. At the end of a spacer, except if the spacer has length three, the model transitions to an End_Spacer state.

Table 5. The spacer region of hypothetical promoters.

Promoter	Aligned	Spacer	S1	S2	S3
TTGATAT	TTGA--TAT	A--	A	-	-
TTGCTAA	TTGC--TAA	C--	C	-	-
TGGCCTAA	TGGCC-TAA	CC-	C	C	-
TTATAGAAC	TTATAGAAC	TAG	T	A	G

In contrast to the M states, the emission probabilities for the S states are not calculated on a position-specific basis. The probabilities are based upon the entire spacer region. S1, S2, and S3 all have the same probabilities for emitting an A (.29), C (.43), G (.14) or T (.14). The End_Spacer state always emits “nothing” with a probability of 1.

Figure 5 illustrates the duration HMM for this example with the transition probabilities between states. All sequences have a letter in the first position of the spacer, S1. Hence, the transition probability from M3→S1 is 1. Half of the sequences have only 1 letter in the spacer region, so they transition down to the End_Spacer state and the transition probability from S1→End_Spacer is .5. The arrows that leave a state must add up to 1, so S1→S2 is also .5. One fourth of the sequences leave the spacer region with 2 letters in the spacer, so the transition probability from S2→End_Spacer is .25 and from S2→S3 is .75. All sequences in the End_Spacer state and all sequences in S3 must proceed to M4, so these transition probabilities are 1.

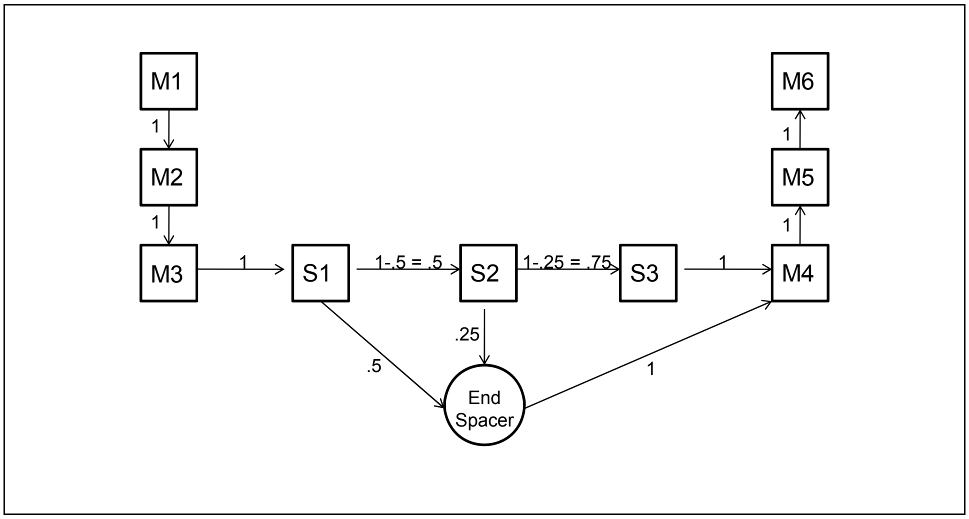


Figure 5. Example duration HMM.

Now that the duration HMM has been built according to the parameters of the four training sequences, it can be used to score a new sequence as was done with the PSSM. Scoring of the trimers is exactly the same as with the PSSM because all transition probabilities between sequential M states are 1. Thus, we can just calculate the spacer score according to the formula

$$\text{Spacer Score} = \sum_{\text{state i=S1}}^{\text{min(End\_Spacer,S3)}} \left[\log_2\left(\mathrm{p}_{\propto \mathrm{i}} / \mathrm{b}_{\propto \mathrm{i}}\right) + \log_2\left(\mathrm{t}_{\mathrm{i,i+1}}\right)\right]$$

where state i = S1, S2, S3 or End_Spacer, αi is the character (A,C,G,T or “nothing”) occupying state i of the new spacer, $p_{\alpha i}$ is the corresponding emission probability, $b_{\alpha i}$ is the background probability as before, and $t_{i,i+1}$ is the transition probability from state i to the next. The score for a spacer region AC would be calculated as: $[\log_2(.29/.285) + \log_2(.5)] + [\log_2(.43/.215) + \log_2(.25)] + [\log_2(1) + \log_2(1)] = -1.97$.

The program **durahmmer** (Ardell D.H., in preparation) builds a duration HMM from aligned training set sequences and writes parameters of the model in an HMM file compatible with the HMMER software suite [24]. From the HMMER suite, we use **hmmsearch** to read the specifications in the HMM file and search new sequences for the highest scoring subsequences.

2.1.3 Footy: Phylogenetic footprinting

Phylogenetic footprinting takes advantage of motif conservation among related species. Grech *et al* (2007) [15] developed an algorithm that combines *E. coli* trained PWMs and chlamydial phylogenetic footprinting. *C. trachomatis* upstream regions were screened with the PWMs and high-scoring hexamer pairs were filtered with an algorithm that accepts only conserved sequences in a consensus of *C. trachomatis*, *C. pneumoniae* and *C. caviae*. *C. trachomatis* σ^{66} promoter predictions are published along with the Footy algorithm methodology.

2.2 Promoter vs non-promoter algorithms

2.2.1 NNPP2.2: A time-delay artificial neural network

Given a training set of cases with identified features and known classification outcomes, an artificial neural network (ANN) seeks a mathematical separator that maximizes correct classifications [25]. The simplest application is a two-class problem where each training set case is quantified by a vector of features, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and an outcome state of class 1 or class 2. A decision surface

$$\mathrm{D}(\mathbf{x}) = \sum_{i=1}^{n} w_i x_i$$

is sought, such that for as many cases as possible the following conditions hold: D(**x**) > 0 → **x** is a member of class 1; D(**x**) < 0 → **x** is a member of class 2; and D(**x**) = 0 → no decision. The art lies in designing an algorithm that identifies the weight coefficients, w_i. Many algorithms start with an initial set of weights and adopt an iterative "learning" scheme that makes small changes in the weights such that classification is improved. In addition, many ANNs utilize hidden layers and non-linear decision surfaces.

NNPP2.2 [26] is a time-delay ANN [27] that combines two simple ANNs (each models an individual binding site) with a variable-length spacer region. The online algorithm is accessible at http://www.fruitfly.org/seq_tools/promoter.html for both eukaryotic and prokaryotic promoter predictions. NNPP2.2 was first derived for eukaryotes using a training set of *Drosophila melanogaster* promoters. Consequently, the published methodology describes the development of the eukaryotic algorithm. What follows is a description of a prokaryotic algorithm that might be inferred from the original documentation.

The training set of 272 *E. coli* promoter regions is available at http://www.fruitfly.org/seq_tools/datasets/Prokaryotic/. Each promoter region entry is 50-60 nts in length and identifies the two binding hexamers, the spacer region in between, flanking regions on each end, and TSS. Since the transcription start sites are known, an appropriate window can be selected for the development of each binding site ANN. For example, after assigning a location of +1 to the TSS, a window of -16 to -1 is appropriate for the -10 hexamer ANN. Within that window of a training sequence, each subsequence of length 6 is considered a case with feature vector **x** = (A1, C1, G1, T1, … , A6, C6, G6, T6) and outcome state = 1 if the hexamer is the identified binding site and 0 otherwise. The values for A1 … T6 are assigned a 1 if the referent letter occupies the referent position it the hexamer under consideration and 0 otherwise. The resulting D(**x**) is a scoring function similar to a PWM. In fact, previously published *E. coli* hexamer PWMs are used to initialize the weight matrices. The time delay ANN scans and scores overlapping sliding window sequences of length 50 and step size 1 with both the -35 hexamer ANN and the -10 hexamer ANN separated by a defined spacer interval. It reports a score in the range (0, 1) that indicates the likelihood of the sequence containing an RNAP binding region.

2.2.2 TSS-PREDICT: An ensemble of support vector machines

The decision surface described for ANNs can be viewed as a hyperplane. A support vector machine (SVM) differs from an ANN in that the algorithm does not seek just “any” hyperplane, it determines the maximum-margin hyperplane between two classes of a training set [28].

TSS-PREDICT [29] predicts bacterial promoters with an ensemble of SVMs. DNA sequences of length 200 are represented using the *tagged mismatch string kernel.* A subsequence of length 5 is tagged with its location relative to the TSS rounded to the nearest 10. In a sequence of 200, there will be 20,480 potential features (4^5 x 20 locations). When feature frequencies are tallied, one mismatch is allowed. Thus, ATAAT(-10) and ATACT(-10) will both be included in the frequencies of

ATAAT(-10) and ATACT(-10). After the feature frequencies are tallied, they are weighted by a measure of symmetric uncertainty [30] and the 200 highest-scoring features are selected for training the SVM.

For TSS-PREDICT, 40 SVMs were trained on 450 *E. coli* promoter sequences of length 200. The positive training set was the same for all 40 machines – for each sequence the TSS was placed at position 50 from the right end (150 from the left). The negative training sets were all different, but sequences continued to be of length 200. Each negative sequence contained a known TSS, but it was offset either upstream or downstream from position 50 in increments of 5 nt.

When presented with a new sequence of length 200, a trained SVM classifier returns the distance of that sequence from the optimal hyperplane separating positive and negative training sets. This distance, or score, reflects the likelihood of the nt in position 50 being the TSS. For the ensemble of 40 SVMs, the score for a given sequence/TSS was determined by averaging the scores provided by each SVM.

To predict the promoter for a given gene, all positions from -250 to -1 with respect to the gene start site were considered as possible TSSs. Top ranking scores were selected as TSS predictions. Once a TSS is predicted, two PWMs trained on 250 known *E. coli* promoters predict the two hexamers. The SVM approach has the advantage of quantifying the primary structures of the regions upstream and immediately downstream from the σ-factor binding region, as well as the binding region itself. *C. trachomatis* σ^{66} promoter predictions are published along with the TSS-PREDICT algorithm methodology.

2.2.3 Stepwise Binary Logistic Regression: The algorithm for building MMCTPPi

Stepwise Binary Logistic Regression (SBLR) [31, 32], as implemented in SPSS version 17.0 statistical software (SPSS Inc., Chicago, IL), selects an optimal set of independent variables (continuous and/or categorical) to classify observations into two populations. Logistic regression does not assume a linear relationship between the dependent and independent variables, normal distributions, or homoscedasticity (equal variances). It does, however, assume independence of observations. This requirement is addressed in section 4.3 which describes the selection of non-redundant observations.

The mathematical model (prediction equation) fitted by SBLR has the form

$$\mathbf{u} = b_0 + \sum_{i=1}^{n} b_i v_i$$

where n is the number of steps, v_1 through v_n are the predictor variables selected, and b_0 through b_n are coefficients determined by the analysis.

u is the logit for the dependent variable, which means that

$$\mathbf{u} = \ln(\text{odds(event)}) = \ln(\text{prob(event)}/\text{prob(non-event)})$$
$$= \ln(\text{prob(event)}/(1\text{-prob(event)})).$$

Here, the event is class membership. When P denotes the prob(class = promoter), the equation can be rewritten as

$$\mathbf{u} = \ln(P/(1-P));\ e^{\mathbf{u}} = P/(1-P);\ \text{and } P = e^{\mathbf{u}}/(1+e^{\mathbf{u}}) = 1/(1+e^{-\mathbf{u}}).$$

Selecting a threshold for P, most commonly 0.5, converts P into a classifier. When 0.5 is the probability threshold, $e^{\mathbf{u}} = 1$ and the classification threshold for **u** is 0.

The effectiveness of a model can be evaluated by its ability to correctly classify the training data.

The SPSS SLBR analysis procedure provides many user-defined options. The Forward Conditional stepwise procedure was selected for all analyses. At each step, a score statistic is calculated for each variable excluded from the model. The score statistic is based on Maximum-Likelihood Estimation criteria and is asymptotically distributed as a χ^2 variable [32]. The variable with the highest significant χ^2 value is entered into the model. If no significant variables remain, then the procedure stops with the current model. Similarly there is a mechanism for stepwise removal. After a new model has been generated, score statistics are calculated for all variables in the model. If the p-value for any variable in the model is greater than the probability for stepwise removal, then the variable is removed from the model. The default probabilities for stepwise entry (.05) and removal (.10) were retained, thus ensuring that the significance of all model variables is less than 0.10.

Chapter 3

Implementation I: Building a duration HMM for RNAP-σ^{66}/DNA binding

3.1 Data-file preparation (documentation in Appendix A-1)

3.1.1 Upstream regions of all 895 genes in the *C. trachomatis* genome are transformed for analysis in parse32.xls

The major data-file for this study, parse32.xls, is used for genome-wide predictions. A second file, parse32ts27.xls, contains the subset of 27 training set genes extracted from parse32.xls. After a duration HMM is constructed, a modified version of parse32ts27.xls provides input for **hmmsearch** which produces the duration HMM scores used in SPSS SBLR analysis. parse32.xls and parse32ts37.xls are created here according to the following two steps, and augmented in Chapter 4, Implementation II.

1. Files containing the *C. trachomatis* genome (NC_000117.fna) and genome table (NC_000117.ptt) were retrieved from the NCBI website, ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/, on July 2, 2007 and last modified by NCBI on January 23, 2007. A column indicating the hour of gene expression onset [20] was added to the genome table. An R script (script in Appendix B-1) was written to extract the 600 nt upstream regions of all 895 protein-coding genes as annotated in the genome table. The length of 600 was determined by first noting that the maximum distance from promoter to the gene start site (GSS) is 296 nt in the training set. The upstream region then was defined as 600 nt to allow for biophysical

structures 275 nt upstream from a predicted promoter. For predictions, the upper limit on promoters was set to 325 nt.

2. For each gene, the 600 nt upstream region was parsed into overlapping sliding window sequences of length 32 (6 nt for each hexamer and a maximum spacer of 20 nt) and step-size 1. Each subsequence (SEQ32) was labeled according to its parent gene and position occupied in the upstream region: e.g. the first SEQ32 was labeled CT001_600 because the initial nt is found 600 nts upstream from the CT001 GSS. Table 6 shows the portion of parse32.xls and parse32ts27.xls that pertains to positions 117 to 101 of CT046.

Table 6. Portion of parse32.xls and parse32ts27.xls.

SEQ32_ID	NAME	SEQ32	POSITION	HOUR
CT046_117	hctB	TAATTGTGTGTGGTTAGTTTTTAATAAAAAGT	117	16
CT046_116	hctB	AATTGTGTGTGGTTAGTTTTTAATAAAAAGTT	116	16
CT046_115	hctB	ATTGTGTGTGGTTAGTTTTTAATAAAAAGTTA	115	16
CT046_114	hctB	TTGTGTGTGGTTAGTTTTTAATAAAAAGTTAA	114	16
CT046_113	hctB	TGTGTGTGGTTAGTTTTTAATAAAAAGTTAAA	113	16
CT046_112	hctB	GTGTGTGGTTAGTTTTTAATAAAAAGTTAAAA	112	16
CT046_111	hctB	TGTGTGGTTAGTTTTTAATAAAAAGTTAAAAA	111	16
CT046_110	hctB	GTGTGGTTAGTTTTTAATAAAAAGTTAAAAAC	110	16
CT046_109	hctB	TGTGGTTAGTTTTTAATAAAAAGTTAAAAACT	109	16
CT046_108	hctB	GTGGTTAGTTTTTAATAAAAAGTTAAAAACTA	108	16
CT046_107	hctB	TGGTTAGTTTTTAATAAAAAGTTAAAAACTAA	107	16
CT046_106	hctB	GGTTAGTTTTTAATAAAAAGTTAAAAACTAAC	106	16
CT046_105	hctB	GTTAGTTTTTAATAAAAAGTTAAAAACTAACC	105	16
CT046_104	hctB	TTAGTTTTTAATAAAAAGTTAAAAACTAACCA	104	16
CT046_103	hctB	TAGTTTTTAATAAAAAGTTAAAAACTAACCAT	103	16
CT046_102	hctB	AGTTTTTAATAAAAAGTTAAAAACTAACCATT	102	16
CT046_101	hctB	GTTTTTAATAAAAAGTTAAAAACTAACCATTT	101	16

3.1.2 Files for duration HMM iteration

1. ts0.txt: training set sequences from Table 1 in original alignment (text file in Appendix B-2).
2. ts5e.txt: each sequence in ts0.txt extended by 5 nt on each end (text file in Appendix B-3).
3. ts32.txt: extracted from the first 3 columns of parse32ts27.xls defined above.

3.2 Duration HMM iteration

3.2.1 Overview of duration HMM iteration

Here, duration HMMs are used to generate a score that characterizes RNAP-σ^{66}/DNA binding, the primary promoter predictor. A set of known promoters is used to train the duration HMM which then scans new sequences to identify the highest scoring subsequence. The results for each scanned sequence are: HMM_SCORE = the score associated with the highest scoring subsequence, START = position of the lead nucleotide in the -35 hexamer and END = position of the last nucleotide of the -10 hexamer. Thus, if TTGTGTGTGGTTAGTTTTTAATAAAAA is the highest scoring subsequence in CT046_117, TAATTGTGTGTGGTTAGTTTTTAATAAAAAGT, START = 4 and END = 30.

Minor modifications in the alignment of the training set promoters can improve classification accuracy. To accomplish this, each promoter is allowed to vary within a neighborhood that extends the sequence by 5 nts on each side. A limit of 5 nts ensures that a modified hexamer will not locate completely outside of the original promoter sequence.

For example, when the training set promoter CT377 is extended, it becomes TTGTTTGCAGAGTTTTTATTTTAAATATGTTATAATCTGTC, with the underlined nts marking the extensions. As diagrammed in Figure 6, a duration HMM is initially generated by the training set in the original alignment which includes TGCAGAGTTTTTATTTTAAATATGTTATAAT for CT377. When the set of extended promoters is searched for the highest scoring instances of the duration HMM, it identifies TTGCAGAGTTTTTATTTTAAATATGTTATAAT as the highest scorer in TTGTTTGCAGAGTTTTTATTTTAAATATGTTATAATCTGTC. Consequently, TTGCAGAGTTTTTATTTTAAATATGTTATAAT replaces the original alignment of CT377 in the training set file.

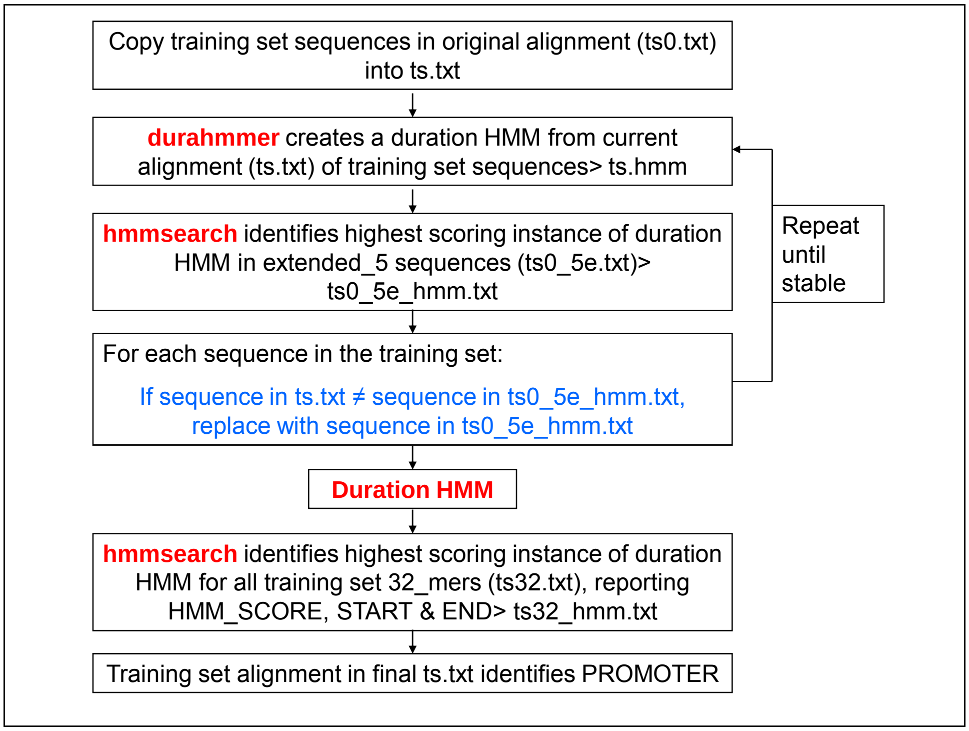

Copy training set sequences in original alignment (ts0.txt) into ts.txt
durahmmer creates a duration HMM from current alignment (ts.txt) of training set sequences> ts.hmm
hmmsearch identifies highest scoring instance of duration HMM in extended_5 sequences (ts0_5e.txt)> ts0_5e_hmm.txt
For each sequence in the training set:
If sequence in ts.txt ≠ sequence in ts0_5e_hmm.txt, replace with sequence in ts0_5e_hmm.txt
Repeat until stable
Duration HMM
hmmsearch identifies highest scoring instance of duration HMM for all training set 32_mers (ts32.txt), reporting HMM_SCORE, START & END> ts32_hmm.txt
Training set alignment in final ts.txt identifies PROMOTER

Figure 6. Duration HMM iteration.

At the end of each iteration, all training set sequences are realigned to match the high scorers identified in ts0_5e_hmm.txt. The process is repeated until the training set stabilizes.

The final duration HMM is then used to identify HMM_SCORE, START and END for all records in the ts32.txt file. These results are merged with the SPSS data-file for the analysis described in 4.1.2 and 4.2. Additionally, the final training set alignment in ts.txt defines the PROMOTER variable discussed in 4.1.1 and 4.2.

3.2.2 Specifics of duration HMM iteration (documentation in Appendix A-2)

A training set of promoter sequences, ts.txt, is supplied as input to **durahmmer** which creates a duration HMM with the command: **durahmmer** -5 6 -3 6 -s 16 -S 20 -p 1 -u 28.5:21.5:21.5:28.5 –C ts.txt >ts.hmm. The options to the command specify the following model parameters: 6 matched states (hexamers) at the 5’ and 3’ sequence ends; minimum and maximum spacer lengths of 16 and 20 respectively; a background compositional model of 28.5% A, 21.5% C, 21.5% G, and 28.5% T; and spacers should be modeled to have their empirical composition in the training set (which in this case was: 38% A, 12% C, 17% G, 33% T). Two hexamers plus a maximum spacer of 20 nts results in a model with 32 possible nodes.

The program **durahmmer** produces a valid HMMer 2.3.2 [24] model file representing a duration HMM. The model file supplies the parameters of the duration HMM to the program **hmmsearch** [24] which searches sequences for instances of the model. The model file for the final model of this study, ts.hmm, is provided as an example (hmm file in Appendix B-4).

Complete documentation for the contents of the file ts.hmm can be found in the HMMER User’s Guide at http://www.psc.edu/general/software/packages/hmmer/. Briefly, the first 17 lines are header information with the main model section following. There are 3 lines for each of the 32 possible nodes. The first and last 6 nodes refer to the -35 and -10 hexamers, while nodes 7 through 26 refer to possible spacer positions. The first line for each node displays the contribution to the final score (multiplied by 10^3) for the corresponding nucleotide matching A, C, G or T. The third line is particularly relevant to nodes 22 through 25, which correspond to spacer nucleotides 17 through 20. As nucleotides in these positions may or may not be present in the sequence being scored due to variable spacer length, the third line provides the odds of transitioning to another spacer nucleotide or to the -10 hexamer.

The command: **hmmsearch** -E 9000 ts.hmm ts32.txt>ts32_hmm.txt directs **hmmsearch** to identify the optimal promoters and HMM scores for all 16,200 SEQ32 observations from the 27 training set genes in ts32.txt. The high E-value was used because we are interested in the maximum score regardless of its magnitude. Finally, ts32_hmm.txt is merged with the SPSS data-file for analysis in Chapter 4.

Chapter 4

Implementation II: Building a promoter prediction model with Stepwise Binary Logistic Regression

4.1 Potential observations, dependent variable and independent variables for Stepwise Binary Logistic Regression (SBLR)

The potential observations, or experimental units, for SBLR are the overlapping 32-mers identified by SEQ32 in Table 6 (Chapter 3) excerpted from parse32ts27.xls and ts32.txt. The dependent variable, PROMOTER, and all of the independent variables are functions of the 32-mer in SEQ32.

Cases with upstream positions ≤ 325 and ≥ 40 were selected as potential observations to restrict the analysis to the range of the training set data. The upper bound is 30 nt upstream from the furthest upstream training set promoter and the lower bound is equal to the furthest downstream training set promoter. This restriction results in 286 potential observations per gene.

4.1.1 Dependent variable

The dependent variable, PROMOTER, is assigned a 1 if the promoter sequence defined by the final alignment in ts.txt (described in 3.2.1) is totally contained in SEQ32, and 0 otherwise. Thus, 1's identify potential promoter observations and 0's identify potential non-promoter observations.

4.1.2 Independent variables

1. HMM_SCORE: the duration HMM score for SEQ32 from ts32_hmm.txt described in Chapter 3.
2. POSITION: the location of SEQ32 in the upstream region relative to the GSS. e.g. For CT046_101, POSITION=101.
3. HOUR: hour of gene expression onset [20]. Possible times of expression onset include 1, 3, 8, 24 and 40 hours post infection (h PI). Mutually exclusive binary variables H1, H3, H8, H16, H24 and H40 are created to mark the time of expression onset.
4. Measures of curvature (CURVE) [33] and %GC content (GC) for each 600 nt upstream region, determined by the online bend.it Server (http://hydra.icgeb.trieste.it/dna/bend_it.html) with a window-size of 32.
5. Free energy change (ΔG) of DNA melting (parameter #33 [34], dinucleotide, window size 2), bendability (parameter #31 [35], trinucleotide, window size 3) and twist angle (parameter #44 [36], dinucleotide, window size 2), determined for each 600 nt upstream region by the online plot.it Server (http://hydra.icgeb.trieste.it/dna/plot_form.html). All measurements were then averaged over each SEQ32. ΔG always has a negative sign and is interpreted as greater values having lower stability. For statistical analysis this variable was transformed by STABLE = -ΔG so that the sign is always positive and the interpretation is that larger values have greater stability. Stability is also of interest in the immediate downstream region, so positions 27-37 (STABLE27_37) and 1-37 (STABLE1_37) were quantified. Since the

bendability measure increases with rigidity, it was renamed RIGID. The twist angle measurement, TWIST, was not transformed.

6. Stress-induced DNA duplex destabilization (SIDD) [10] measures the propensity for strand separation under negative superhelical stress based on structural and energetic properties of DNA. A low SIDD score indicates a high propensity for strand separation. SIDD measurements were determined by the WebSIDD Server [37] (http://www.genomecenter.ucdavis.edu/benham/sidd/websidd.php) with the default parameters except for Open Region Size = 63. Because Niehaus *et al* [38] have shown a time dependent response to chlamydial DNA supercoiling, interactions between the time of expression onset and SIDD were included [20]. The SIDD/hour of onset interaction is quantified by SIDD_H# = SIDD*H#.
7. For variables defined in 4-6, lagged variables were created for the four non-overlapping upstream subsequences of length 32: e.g. for CT046_100, CURVE_L32 was set equal to the CURVE value of CT046_132; CURVE_L64 was set equal to the CURVE value of CT046_164; CURVE_L96 was set equal to the CURVE value of CT046_196; and CURVE_L128 was set equal to the CURVE value of CT046_228.

4.2 Data-file preparation

1. To parse32.xls and parse32ts27.xls, append the following independent variables: CURVE, GC, STABLE, STABLE27_37, STABLE1_37, RIGID, TWIST and SIDD (documentation in Appendix A-1).
2. Read 32ts27.xls into an SPSS data-file. Create lags and interactions with SPSS syntax file, trans.sps (SPSS syntax file in Appendix B-5).
3. For each new duration HMM, copy START, END and HMM_SCORE from ts32_hmm.txt into columns 3-5 of SPSS data-file (documentation in Appendix A-2).
4. Set all PROMOTER = 0. Enter 1's manually.

4.3 Selection of non-redundant observations from potential observations

As stated in Chapter 2, SBLR assumes independent observations. To address this requirement, we select for analysis a subset of the overlapping potential observations that are non-redundant with respect to the hexamer pair that is most likely to bind the RNAP σ-factor.

Table 7 displays the first five columns of the SPSS data-file used for analysis. Each potential observation occupies a row. A row includes: the SEQ32 label (SEQ_ID); the SEQ32 literal sequence (SEQ32); the score of the optimal HMM instance in SEQ32 (HMM_SCORE); the position of the last nt in the -10 hexamer of the optimal HMM instance (END); and PROMOTER as previously defined.

Table 7. Selecting rows with END = 32 (*) ensures non-redundant observations with regard to hexamers and HMM_SCORE.

SEQ32_ID	PRO- MOTER	END	HMM_ SCORE	SEQ32:bold underline locates optimal HMM instance
CT046_117	0	* 32	-5.9	TAA**TTGTGT**GTGGTTAGTTTTTAATA**AAAAGT**
CT046_116	0	31	-5.9	AA**TTGTGT**GTGGTTAGTTTTTAATA**AAAAGT**T
CT046_115	0	30	-5.9	A**TTGTGT**GTGGTTAGTTTTTAATA**AAAAGT**TA
CT046_114	0	29	-5.9	**TTGTGT**GTGGTTAGTTTTTAATA**AAAAGT**TAA
CT046_113	0	29	-13.7	T**GTGTGT**GGTTAGTTTTTAATAA**AAAGTT**AAA
CT046_112	0	* 32	-11.4	**GTGTGT**GGTTAGTTTTTAATAAAAAG**TTAAAA**
CT046_111	1	* 32	-2.1	TGTG**TGGTTA**GTTTTTAATAAAAAGT**TAAAAA**
CT046_110	1	31	-2.1	GTG**TGGTTA**GTTTTTAATAAAAAGT**TAAAAA**C
CT046_109	1	30	-2.1	TG**TGGTTA**GTTTTTAATAAAAAGT**TAAAAA**CT
CT046_108	1	29	-2.1	G**TGGTTA**GTTTTTAATAAAAAGT**TAAAAA**CTA
CT046_107	1	28	-2.1	**TGGTTA**GTTTTTAATAAAAAGT**TAAAAA**CTAA
CT046_106	0	31	-11.9	GG**TTAGTT**TTTAATAAAAAGT**TAAAAA**CTAAC
CT046_105	0	30	-11.9	G**TTAGTT**TTTAATAAAAAGT**TAAAAA**CTAACC
CT046_104	0	* 32	-7.6	**TTAGTT**TTTAATAAAAAGTTAAAAAC**TAACCA**
CT046_103	0	* 32	-7.8	TAG**TTTTTA**ATAAAAAGTTAAAAACT**AACCAT**
CT046_102	0	31	-7.8	AG**TTTTTA**ATAAAAAGTTAAAAACT**AACCAT**T
CT046_101	0	30	-7.8	G**TTTTTA**ATAAAAAGTTAAAAACT**AACCAT**TT

If we select only those cases where END = 32, we eliminate all of the redundant optimal HMM hexamer pairs while retaining most optimal HMM instances (information). Table 7 demonstrates how this selection ensures that neighboring optimal HMM instances that match are included only once. Six potential observations, CT046_111 through CT046_106, all contain the experimentally identified promoter with hexamer pair TGGTTA and TAAAAA. Consequently, they all have PROMOTER=1 and HMM_SCORE = -2.1. But only CT046_111 has END = 32 and is selected to represent the experimentally identified CT046 promoter. Similarly, only CT046_117 represents the maximal non-promoter hexamer pair TTGTGT and AAAAGT with score = -5.9. This process incidentally aligns each selected SEQ32 such that the optimal downstream hexamer is at the far right end.

This selection process does not eliminate overlapping sequences, but it does eliminate overlapping likely binding sites. CT046_111 and CT046_112 overlap a great deal. However, the last hexamer of CT_046_111 (TAAAAA) is not present in CT046_112 and the first hexamer of CT_046_112 (GTGTGT) does not appear in CT046_111.

While selecting sequences with non-redundant HMM_SCORES does mitigate the problem of dependent observations, it may not entirely eliminate it. As this appears to be the first study to analyze biological sequence data with logistic regression, there are no available suggestions from others. Additionally, whereas there are numerous studies that affirm the robustness of Baysian Discriminant Analysis with regard to violating the assumptions of a linear relationship between the dependent and independent variables, normal distributions, and homoscedasticity [39], there are no similar studies regarding the robustness of logistic regression. An alternative to the current analysis would be to use Stepwise Discriminant Analysis, knowing that we are violating some assumptions.

There are versions of logistic regression, including generalized estimating equations (GEE) [32], that are specifically designed for correlated data such as longitudinal studies. In these procedures there are subject variables and within subject variables. It might be possible to force this study data into such a format, but as yet there are no readily available stepwise procedures to scan multiple possible predictors. A final alternative would be to select non-overlapping sequences with the penalty of losing information and perhaps introducing a selection bias.

SLBR is a procedure for model identification. It is only after a model has been identified that it can be evaluated for independence. Given that, we elect to analyze the non-redundant observations with SLBR and then examine the error terms for independence. In Time Series Analysis (which this analysis most resembles), this is done by checking that the error term is normally distributed with zero mean, and that autocorrelations and partial autocorrelations of the error term are not significant [40].

4.4 SBLR iteration

Deletion of members of the initial training set can eliminate promoters that are outliers or members of a different promoter population. This is accomplished via the iterative scheme diagrammed in Figure 7. Initially, the complete set of 29 verified promoters determines the duration HMM and the independent observations selected for SBLR analysis. SBLR delivers a mathematical model that produces a predicted probability of class membership (P) for each observation. A threshold on P of .5 is used to classify each observation as a predicted promoter or non-promoter.

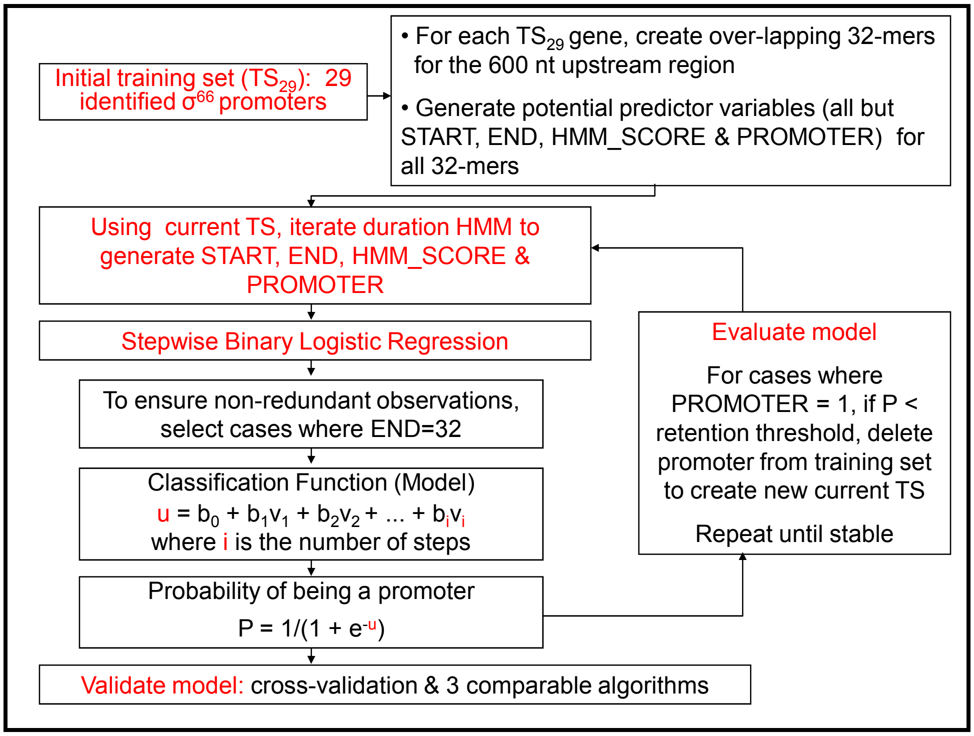

Initial training set (TS_{29}): 29 identified σ^{66} promoters
• For each TS_{29} gene, create over-lapping 32-mers for the 600 nt upstream region
• Generate potential predictor variables (all but START, END, HMM_SCORE & PROMOTER) for all 32-mers
Using current TS, iterate duration HMM to generate START, END, HMM_SCORE & PROMOTER
Stepwise Binary Logistic Regression
To ensure non-redundant observations, select cases where END=32
Classification Function (Model)
$u = b_0 + b_1v_1 + b_2v_2 + ... + b_iv_i$
where i is the number of steps
Probability of being a promoter
$P = 1/(1 + e^{-u})$
Evaluate model
For cases where PROMOTER = 1, if P < retention threshold, delete promoter from training set to create new current TS
Repeat until stable
Validate model: cross-validation & 3 comparable algorithms

Figure 7. SBLR iteration.

Each SBLR model is evaluated on the basis of the observation classifications. If TP = true positive, FP = false positive, TN = true negative and FN = false negative, then sensitivity or recall = TP/(TP+FN), specificity = TN/(FP+TN), positive predictive value (PPV) or precision = TP/(TP+FP), negative predictive value (NPV) = TN/(FN+TN), and accuracy = (TP+TN)/(TP+TN+FP+FN). ROC analysis Area Under the Curve is also reported. The total number of observations for each model differs according to the promoter training set being used.

For those 29 cases where PROMOTER = 1, we also use the value of P to determine when a promoter appears to be an outlier and should be eliminated from (or reinstated to) the training set. After observing the 29 probabilities, a retention threshold on P between 0 and .1 is established. If a training gene has only one identified promoter and that promoter has a P less than the retention threshold, then all observations for that gene are deleted from the analysis. Similarly, if a training set gene has two identified promoters and they are both selected for deletion, all observations for that gene are deleted. However, if a training set gene has two identified promoters and only one is selected for deletion, all upstream observations for that gene remain in the analysis dataset and only observations within the remaining promoter are assigned PROMOTER = 1.

Modifying the training set in any way necessitates the determination of a new duration HMM, which in turn determines which observations will be aligned such that END = 32 and subsequently included in the next SBLR analysis. The iteration process continues until the training set stabilizes, finalizing the promoter prediction model.

4.5 SBLR model validation

4.5.1 Stratified K-fold cross-validation

Once the final training set and model are selected, it is necessary to validate the model to protect against over-fitting and to allow for comparisons with algorithms trained on other datasets. In the case of dichotomous classification, stratified K-fold cross-validation [41] partitions the training set into K subsamples such that each subsample has approximately the same proportions of class membership. Here we designate each training gene as a subsample; hence K equals the number of genes in the training set. Then, one gene (1-2 promoters and approximately 90 non-promoters) is retained as a validation set while the remaining genes are used as training data. Evaluation measures are calculated by aggregating the results of each validation set.

4.5.2 Comparable algorithms

The following three algorithms were used to compare performance and to identify co-predictions with the model developed in this study: NNPP2.2 [26], TSS-PREDICT [29], and Footy[15]. NNPP2.2 is an online time-delay neural network that is accessible for promoter predictions at http://www.fruitfly.org/seq_tools/promoter.html. We used the following options: organism = prokaryote and minimum promoter score = 0.95 to define promoters in the 325 nt upstream region of all *C. trachomatis* genes. For the support vector machine algorithm TSS-PREDICT, the top two ranking predictions for each *C. trachomatis* gene are posted as supplementary material at doi:10.1016/j.combiolchem.2008.07.009. The 42 *C. trachomatis* promoters predicted by the phylogenetic footprinting algorithm, Footy, are reported directly in the publication that describes the algorithm.

Chapter 5

Model building results and discussion

5.1 Finding the best model

The initial model, M0, utilizes the initial training set of 29 promoters with non-redundant observations from their 27 parent genes. The duration HMM model converged after one iteration, modifying the alignment of 7 promoters. For all models, Table 8 reports the variables that were selected for the model and evaluation measures.

Table 8. Models produced by Stepwise Binary Logistic Regression iteration.

SBLR Model	M0	M1	M2	M3
Training Set Deletion	none	CT665 CT681a CT681b CT743	CT665 CT681a CT681b	CT681a CT681b
Variables in Model[a]	+HMM_SCORE +STABLE1_37 -POSITION +CURVE_L32 -GC_L128 +RIGID_L96 +CURVE	+HMM_SCORE +STABLE1_37 -GC_L32 -POSITION +CURVE_L32 -CURVE_L64 -GC_L128 +TWIST	+HMM_SCORE +STABLE1_37 -POSITION +CURVE_L32 -STABLE_L32 +SIDD_H24 -CURVE_L128 -SIDD_L128 +RIGID_L96	+HMM_SCORE +STABLE1_37 -POSITION +CURVE_L32 -STABLE_L32 -STABLE27_37 +CURVE
Sensitivity or Recall	19/29 (0.655)	25/25 (1.0)	26/26 (1.0)	25/27 (0.926)
Specificity	2426/2428 (0.999)	2083/2083 (1.0)	2226/2226 (1.0)	2322/2323 (1.0)
PPV or Precision	19/21 (0.905)	25/25 (1.0)	26/26 (1.0)	25/26 (0.962)
NPV	2426/2436 (.996)	2083/2083 (1.0)	2226/2226 (1.0)	2322/2324 (0.999)
Accuracy	2445/2457 (0.995)	2108/2108 (1.0)	2252/2252 (1.0)	2347/2350 (0.999)
AUC[b]	0.995	1.0	1.0	0.999

[a] The variables are listed in order of entrance into the model and the sign indicates the sign of the coefficient.

[b] ROC analysis Area Under the Curve

For model M0, 19 of the 29 promoters were classified correctly, with 2 false positives. There is always the possibility that false positives are yet to be recognized promoters, but at this point they are counted as misclassifications. For the 10 established promoters that were misclassified, the predicted probabilities ranged from 0.001 to 0.42. Since a natural separation appeared to between 0.07 and 0.10, P = 0.08 was selected as the retention threshold and promoters CT665, CT681a, CT681b and CT743 (along with all observations from their parent genes) were deleted from the training set for the next model, M1.

The duration HMM model for M1 converged after two iterations, modifying the alignment of 5 promoters. Table 8 shows that M1 classified the modified training set perfectly, indicating that perhaps too many promoters had been deleted from the original training set. The retention threshold was reset to 0.07 and CT743 was reinstated for model M2.

The duration HMM model for M2 converged after one iteration. Table 9 displays the alignments of the 6 promoters that were modified. M2 also classified the modified training set perfectly. Again the results indicated that the next model, M3, should reset the retention threshold to 0.06 and reinstate CT665. However, Table 8 reports that M3 is not as good as models M1 and M2 because of classification errors.

Table 9. M2 duration HMM sequence alignment modifications.

CT	Name	To GSS	-35 Hex	Spacer (16-20)	-10 Hex
CT323	*infA*	145	TTGACA	TTTTCTGTTTAGTCGA(16)	TATAAT
		149	TTGTTT	GACATTTTCTGTTTAGTCGA(20)	TATAAT
CT377	*ltuA*	74	TGCAGA	GTTTTTATTTTAAATATGT(19)	TATAAT
		75	TTGCAG	AGTTTTTATTTTAAATATGT(20)	TATAAT
CT442	*crpA*	66	GGGTTT	TTGAAAAAAACAAGTGTTT(19)	GTGTAG
		60	TTGAAA	AAAACAAGTGTTTGTG(16)	TAGACT
CT444b	*omcA*	61	AATTGC	TTTTATCGATAAAAGAAAC(19)	TTCAAG
		59	TTGCTT	TTATCGATAAAAGAAAC(17)	TTCAAG
CT518	*rl14*	198	CTGTTG	TTGTTCGAGTCGAAAGGG(18)	TATACT
		195	TTGTTG	TTCGAGTCGAAAGGGTA(17)	TACTCG
CT701	*secA_2*	57	TGTATA	GGCGCCTTTAAATAAGAGGG(20)	TAGGTT
		61	TTGTTG	TATAGGCGCCTTTAAA(16)	TAAGAG

Given two models, one training set a subset of the other, that both classify their respective training sets with 100% accuracy, we reasoned that the model trained on the largest set would provide the most sensitive genome-wide prediction. Thus, M2 was selected as the best and final model because of the perfect classification with the largest training set and hereafter will be referred to as MMCVPP1. The complete data file used to build MMCVPP1 is available online at http://www.biomedcentral.com/1471-2105/10/271 so that others may replicate or modify the model.

As discussed in section 4.3, the error terms of MMCVPP1 were checked for independence. Residuals, PROMOTER – P, were calculated for all selected observations and shown to be normally distributed with zero mean. Additionally, the autocorrelations and partial autocorrelations of the residuals were not significant. Thus, the independence assumption of SBLR was not violated by this model.

5.2 Validation of the MMCVPP1 model

Aggregated results of the stratified K-fold (25-fold) MMCVPP1 cross-validation are reported in the last column of Table 10. For the 25 genes and 26 promoters in the MMCVPP1 training set, 3 promoters (CT322_298, CT743_085, and CT752_064) were not identified (sensitivity = 0.885) and there were 11 false-positive predictions (precision = 0.676). The incorrect classifications are most likely due to incomplete representation of the sample space, but may indicate additional populations or absent predictors.

Table 10. MMCVPP1 cross-validation.

SBLR Model	MMCVPP1	MMCVPP1 Cross- Validation
Training Set Deletion	CT665 CT681a CT681b	CT665 CT681a CT681b
Sensitivity or Recall	26/26 (1.0)	23/26 (0.885)
Specificity	2226/2226 (1.0)	2215/2226 (0.995)
PPV or Precision	26/26 (1.0)	23/34 (0.676)
NPV	2226/2226 (1.0)	2215/2218 (0.999)
Accuracy	2252/2252 (1.0)	2238/2252 (0.994)
AUC	1.0	0.992

Table 11 compares the performance of the stratified K-fold cross-validation performance of MMCVPP1 with that of comparable algorithms when predicting promoters in the 25 cross-validation genes. The tally is in the form hits/predictions/gene. For NNPP2.2, a prediction was considered a hit if the hexamer pair in Table 1 was fully contained in the 50-mer NNPP2.2 prediction using a threshold of 0.95. The last two rows of the table show the cumulative sensitivity and precision of each prediction algorithm. MMCVPP1 cross-validation is the most sensitive (0.885), while Footy is the most precise (1.0).

Table 11. Comparing predictions of MMCVPP1 cross-validation and comparable algorithms for 25 training set genes.

CT	MMCVPP1 Cross-Validation	NNPP2.2	TSS-PREDICT	Footy
CT046	1/1	0/4	0/2	0/0
CT062	1/2	0/0	1/1	1/1
CT080	1/2	1/4	0/2	0/0
CT091	1/3	1/1	1/1	0/0
CT098	1/1	0/1	1/2	1/1
CT111	1/1	1/3	0/2	1/1
CT286	1/1	1/2	1/1	1/1
CT322	0/0	0/0	0/2	0/0
CT323	1/1	1/3	1/1	1/1
CT377	1/2	1/3	1/1	0/0
CT394	1/1	1/2	1/1	0/0
CT439m	1/1	0/3	0/0	1/1
CT442	1/2	1/1	1/1	0/0
CT444	2/5	2/5	1/2	0/0
CT518	1/1	0/0	1/1	0/0
CT557	1/1	0/1	1/1	0/0
CT559	1/1	1/1	0/2	1/1
CT576	1/2	1/3	1/2	0/0
CT596	1/1	0/1	0/2	1/1
CT674	1/2	1/2	0/0	0/0
CT701	1/1	1/2	0/2	0/0
CT708	1/1	1/2	1/1	1/1
CT743	0/0	0/2	1/5	0/0
CT752	0/0	1/1	0/2	1/1
CT863	1/1	1/1	1/1	0/0
Sensitivity	23/26 (0.89)	17/26 (0.65)	15/26 (0.58)	10/26 (0.39)
Precision	23/34 (0.68)	17/48 (0.35)	15/38 (0.40)	10/10 (1.0)

Table 12 reports the hits and misses for the 2 genes that were not used in the development of MMCVPP1. The only hit was scored by NNPP2.2, with 2 accompanying false positives.

Table 12. Comparing predictions of MMCVPP1 and comparable algorithms for 2 training set genes not in MMCVPP1 training set.

CT	MMCVPP1	NNPP2.2	TSS-PREDICT	Footy
CT665	0/1	1/3	0/2	0/0
CT681	0/1	0/1	0/2	0/1

5.3 MMCVPP1 model interpretation

The MMCVPP1 duration HMM describes and quantifies the RNAP-σ^{66}/DNA binding observed in the training set. For the MMCVPP1 duration HMM model, the input data file for **durahmmer** is ts.txt (text file in Appendix B-6). The output file, ts.hmm (hmm file in Appendix B-4), was discussed in Chapter 3 as an example of an HMMer2.3.2 model file. A visualization of the MMCVPP1 parameters is shown in Figure 8. The -35 hexamer is dominated by the initial TTG motif, while the initial T with frequent As and Ts describe the -10 hexamer. The C and G compositions (12% and 17%, respectively) of the spacer region are much smaller than those of the genome (21.5% each). Spacer lengths of 17 predominate, while spacers of length 19 are absent.

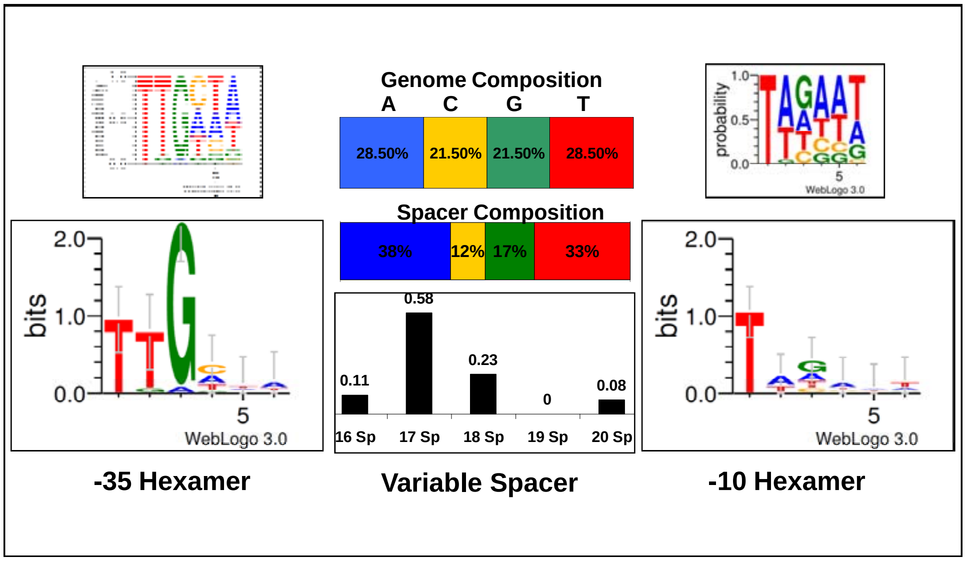


Figure 8. Visualization of the MMCVPP1 duration HMM.

The MMCVPP1 prediction equation generated by SBLR is**:**

$$\mathbf{u} = -1408.301 + 85.305*\text{HMM\_SCORE} + 1816.454*\text{STABLE1\_37} - 1.399*\text{POSITION} + 23.330*\text{CURVE\_L32} - 408.085*\text{STABLE\_L32} + 25.445*\text{SIDD\_H24} - 13.757*\text{CURVE\_L128} - 21.675*\text{SIDD\_L128} + 45.042*\text{RIGID\_L96}$$

Because it is the strongest predictor, HMM_SCORE is selected in the first step of the SBLR procedure. The prediction equation for step one is

$$\mathbf{u} = -0.237 + 0.700*\text{HMM\_SCORE}$$

Using a classification cutoff of P = 0.5 and setting u = 0 yields HMM_SCORE = 0.339 as the threshold for step 1 classification. At step 1, 14/26 promoters and 2220/2226 non-promoters were classified correctly. Thus, the remaining eight model variables moved 12 promoters with HMM_SCORE < .339 to promoter classification and 6 non-promoters with HMM_SCORE ≥ .339 to non-promoter classification (without altering the classification of 2,234 correctly classified observations).

The predictor variables and their coefficients describe the established promoters and their upstream regions. Promoters have high HMM_SCORE and low POSITION. The near upstream region is curved and unstable, whereas the further upstream region is uncurved and unstable under superhelical stress. For late-cycle genes where expression onset occurs at 24h PI, the effect of superhelical stress is less than at other times (a positive SIDD coefficient indicates there is little destabilization of DNA under superhelical stress). The upstream characteristics may reflect transcription factor binding and/or additional interaction with the RNAP holoenzyme.

The interpretation of the positive coefficient for STABLE1_37 is more subtle. In the second step of the SBLR, 4 observations change from FP to TN and 5 observations change from FN to TP. The means of STABLE, STABLE1_37 and STABLE33_37 are all larger in the second group than in the first. Although STABLE33_37 shows the greatest mean difference, the most statistically significant is STABLE1_37.

5.4 *C. trachomatis* genome-wide MMCVPP1 σ^{66} promoter predictions

σ^{66} promoters predicted for the entire *C. trachomatis* genome by MMCVPP1 are reported in *Appendix C: MMCTPP1 genome-wide promoter predictions* (documentation in Appendix A-4). The file lists 479 predicted promoters in 361 unique genes, along with their HMM scores and genome locations. Thus, for 534 of the total 895 *C. trachomatis* genes, this model does not find any 32-mers with a probability >0.5. This suggests a conservative prediction that emphasizes specificity over sensitivity. Other explanatory factors may include alternate binding patterns for σ^{66}, alternative σ-factors, and operon configurations.

Characteristics of the MMCVPP1 genome-wide prediction can be summarized by looking at all 479 predictions, or by looking at the 361 unique genes and selecting the predictions closest to the GSS. The two views produce similar results. Approximately 64% of predicted promoters are completely contained in non-coding upstream regions, 50% are on the positive strand, and time of activation distributes as follows: 5% hour 1, 23% hour 3, 51% hour 8, 20% hour 16 and 2% hour 24. The strand and hour distributions for all 895 genes in the genome are equivalent to the predicted promoter distributions, indicating that there is no strand or temporal preference for the predicted *C. trachomatis* σ^{66} promoters.

Figure 9 displays a histogram of predicted promoter positions. POSITION marks the 5' end of the data-file 32-mer, and is consequently ~ 40 nt upstream from the TSS. Thus, the POSITION distribution peaks with the 5' end around 68 nts upstream from the GSS which is equivalent to a distance of 28nts from TSS to GSS. The peak and shape of this distribution closely resemble the *E. coli* histogram from Burden *et al* (2005) [42].

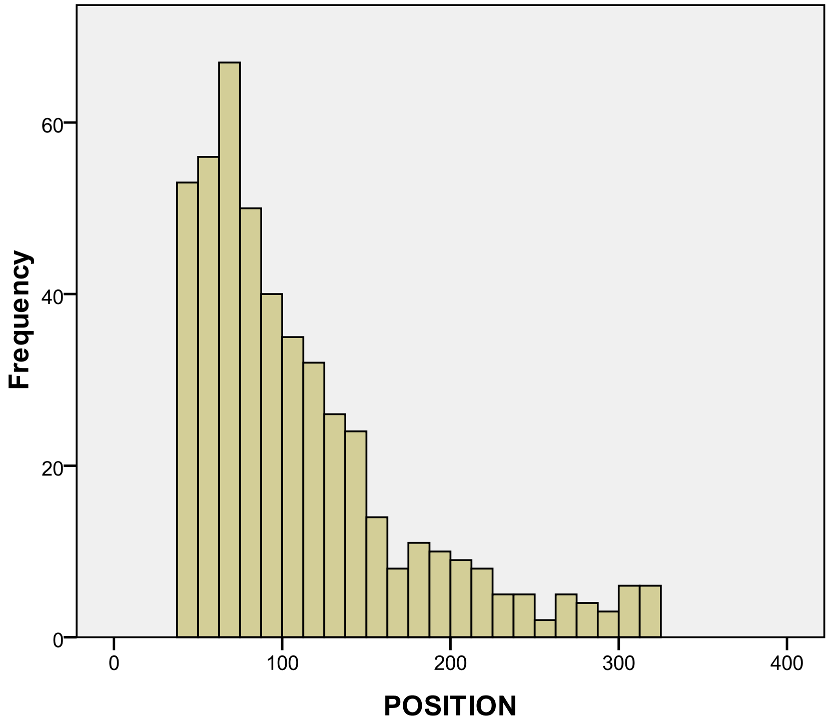


Figure 9. Histogram of predicted promoter POSITION, n=479. The peak at 68 nts upstream from the GSS is equivalent to a distance of 28 nts from the TSS to GSS.

5.5 Discussion of the MMCVPP1 model

The final model produced by the iterative strategy was generated by a training set with three of the original members, CT665, CT681a and CT681b, removed. An explanation of how these three sequences differ from the remainder would be informative. The last column of Table 1 (Chapter 1) reports that CT665 and CT681 are both expressed at 8 h PI, classifying them as mid-cycle genes. Niehus *et al* (2008) [38] recently demonstrated that chlamydial promoters show a differential response to changes in DNA supercoiling that correlates with the lifecycle expression pattern. Specifically, two mid-cycle genes (8 h PI) responded to supercoiling, while three late-cycle genes (≥ 16 h PI) did not. Their experimental set included *ompA*/CT681 in the mid-cycle group and *omcA*/CT444, *hctA*/CT743 & *ltuB*/CT080 in the late-cycle group. Thus, it is likely that there exists a set of mid-cycle promoters that differ topologically from other promoters to enhance their ability to respond to supercoiling, and this may explain the anomolous characteristics of these promoters that we observed.

A possible explanation for the large number of genes without promoter predictions by MMCVPP1 is heterogeneity requiring different models, for example for response to supercoiling. While investigating the initial model M0, we explored stepwise nominal regression, which allows for the discovery of more than two dependent variable categories. However, we did not find that a third category was substantiated. Nonetheless, we suspect that future promoter identifications may confirm the existence of more than two promoter populations for σ^{66} in *Chlamydiales*.

5.6 Comparison with *C. trachomatis* genome-wide NNPP2.2 and TSS-PREDICT predictions

As stated in the project overview (1.4.2.1), it was planned that MMCVPP1 co-predictions with NNPP2.2 and TSS-PREDICT would be used to select candidates for experimental testing with 5'-rapid amplification of cDNA ends (5'-RACE). R scripts (documentation in Appendix A-5; scripts in Appendices B-7 and B-8) scanned the promoters predicted by NNPP2.2 and TSS-PREDICT for matches with the promoters predicted by MMCVPP1. An NNPP2.2 match was declared when the MMCVPP1 prediction was contained within the 50 nt NNPP2.2 prediction. A TSS_PREDICT match was declared when the TSS_PREDICT predicted hexamer pair was contained within the MMCVPP1 prediction.

There was a substantial overlap among predictions by different methods. *Appendix D: NNPP2.2 co-predictions* lists the 209 promoters (176 unique genes) co-predicted by M2 and NNPP2.2, while *Appendix E: TSS-PREDICT co-predictions* lists the 175 promoters (162 unique genes) co-predicted by MMCVPP1 and TSS-PREDICT. *Appendix F: Co-predictions of all 3 algorithms* reports the 98 promoters (90 unique genes) co-predicted by MMCVPP1, NNPP2.2 and TSS-PREDICT. All predictions are for 325 ≥ POSITION ≥ 40, consistent with the range of the modeling procedure.

Of the 42 promoters predicted by Footy, 11 were members of the MMCVPP1 training set, 4 (CT265_111, CT342_102, CT547_065 and CT606_149) were co-predicted by MMCVPP1 and NNPP2.2, and 6 (CT267_097, CT269_82, CT446_245, CT546_050, CT646_071, and CT837_088) were predicted by all four algorithms.

Chapter 6

Strategy validation: Identification of 169 *C. trachomatis* σ^{66} promoters augments the list of mapped promoters and enhances the training set for MMCTPP2

Three months after the strategy for developing MVCTPP1 and the accompanying *C. trachomatis* σ^{66} promoter predictions were published [43] (Appendix K), Albrecht *et al* [44] reported 317 transcription start sites (TSSs) for putative coding genes of the *C. trachomatis* L2b/UCH-1/proctitis (L2b) genome (NC_010280) that they had determined via a deep sequencing RNA-Seq approach [45]. Their results presented an opportunity to partner the newly mapped TSSs with promoter predictions from MMCTPP1 and TSS-PREDICT to map new *C. trachomatis* σ^{66} promoters.

6.1 Establishing a homologous alignment between L2b/UCH-1/proctitis and D/UW-3/CX

Strains of *C. trachomatis* are classified serologically by their outer membrane protein (OmpA) [46]. Serovars A-C invade mucosal epithelia in the ocular tissue and are associated with trachoma. Serovars D-K infect the urogenital tract and are associated with sexually transmitted cervicitis in women and urogenital infections in men. Serovars L1, L2, and L3 are also associated with sexually transmitted infections, causing lymphogranuloma vernerum. Because forecasts by MMCTPP1 and TSS-PREDICT are specific for the *C. trachomatis* D/UW-3/CX (UW-3) genome

(NC_000117), synchronizing the algorithmic predictions with the TSS maps requires a homologous alignment between UW-3 and L2b (a variant of L2).

Thomson *et al* [46] demonstrated a high degree of homology among genomes representing serovars A, D, and L2. They found that 846 predicted and functional coding sequences were shared among all three strains, and 876 were shared between D and L2. Thus, it is likely that nearly all of the newly discovered L2b TSSs are also present in UW-3.

Genome annotations for L2b (NC_010280), L2 (NC_010287) and UW-3 (NC_000117) are available at ftp://ftp.ncbi.nih.gov/genomes/Bacteria/. A list of homologous gene-pairs between *C. trachomatis* L2/434/Bu (L2) and UW-3 was generously provided by Nicholas Thomson (Wellcome Trust Sanger Institute; personal communication). The L2 and L2b gene maps were easily aligned, which in turn provided a mapping from L2b to UW-3. The alignment of all three genomes is provided in *Appendix G: Alignment of strains L2b, L2 and UW-3.*

Before proceeding with the analysis, it was necessary to verify the homology between the upstream regions of L2b and UW-3 gene-pairs and to confirm synchronization with the MMCTPP1 algorithm. (MMCTPP1 and TSS-PREDICT have previously been shown to agree with regard to UW-3 genome annotation.) Table 1 of the deep sequencing study describes transcripts of 84 L2b genes along with their UW-3 homologues. Since 22 of these were correctly predicted by the MMCTPP1 algorithm, these 22 gene-pairs were selected for visual inspection of upstream regions. Genome locations and details of the 317 L2b TSSs are available in Supplementary Data at http://nar.oxfordjournals.org/cgi/content/full/gkp1032/DC. The upstream regions surrounding the homologous TSSs from both strains were visualized and compared with the GBpro Genome Browser V3.0. (http://prodoric.tu-bs.de/gbpro.php). BCM Search Launcher: Sequence Utilities (http://searchlauncher.bcm.tmc.edu/seq-util/seq-util.html) provided the reverse complement of negative strand sequences.

Appendix H: Homology verification Tables H-1 and H-2 demonstrate a high degree of homology between gene-pairs. The few existing nucleotide differences do not significantly impact the transcription start sites or predicted binding sites.

6.2 Matching MMCTPP1 UW-3 forecasts with L2b TSSs

MMCTPP1 UW-3 forecasts are detailed in *Appendix C: MMCTPP1 genome-wide promoter predictions*. The discriminator length is the number of nucleotides (nts) between the downstream end of the -10 hexamer and the TSS. To calculate the discriminator length of a MMCTPP1 prediction, first the distance from the downstream end of the predicted -10 hexamer to the annotated GSS in UW-3 (*tail to gss* {MMCTPP1}) is calculated. Then the distance from TSS to GSS in L2b (*tss to gss* {deep-seq}) is calculated. The discriminator length of the prediction is the difference between the two (*tail to gss* {MMCTPP1} - *tss to gss* {deep-seq}). Figure 10 (top) illustrates the calculation of the discriminator length. An MMCTPP1 forecast was considered correct if the discriminator length of the prediction was in the range {4-14}. The promoter region can then be defined by the location of the upstream end of the -35 hexamer, the spacer, and the sequence from the -35 hexamer to the TSS.

6.3 Matching TSS-PREDICT UW-3 forecasts with L2b TSSs

TSS-PREDICT UW-3 predictions appear as Supplementary Data with the article [29]. The distance from predicted TSS to the annotated GSS (*tss to gss* {TSS-PRED}) is reported for each predicted TSS. Subtracting *tss to gss* {TSS-PRED} from *tss to gss* {deep-seq} leaves an *off-set* which measures the distance between the TSS-PREDICT forecast and the experimental TSS. Figure 10 (bottom) diagrams the calculation of *off-set*. TSS-PREDICT was considered correct if |*off-set*| ≤ 6. After predicting the TSS, TSS-PREDICT uses a position weight matrix to identify the -35 and -10 hexamers which complete the definition of the promoter region.

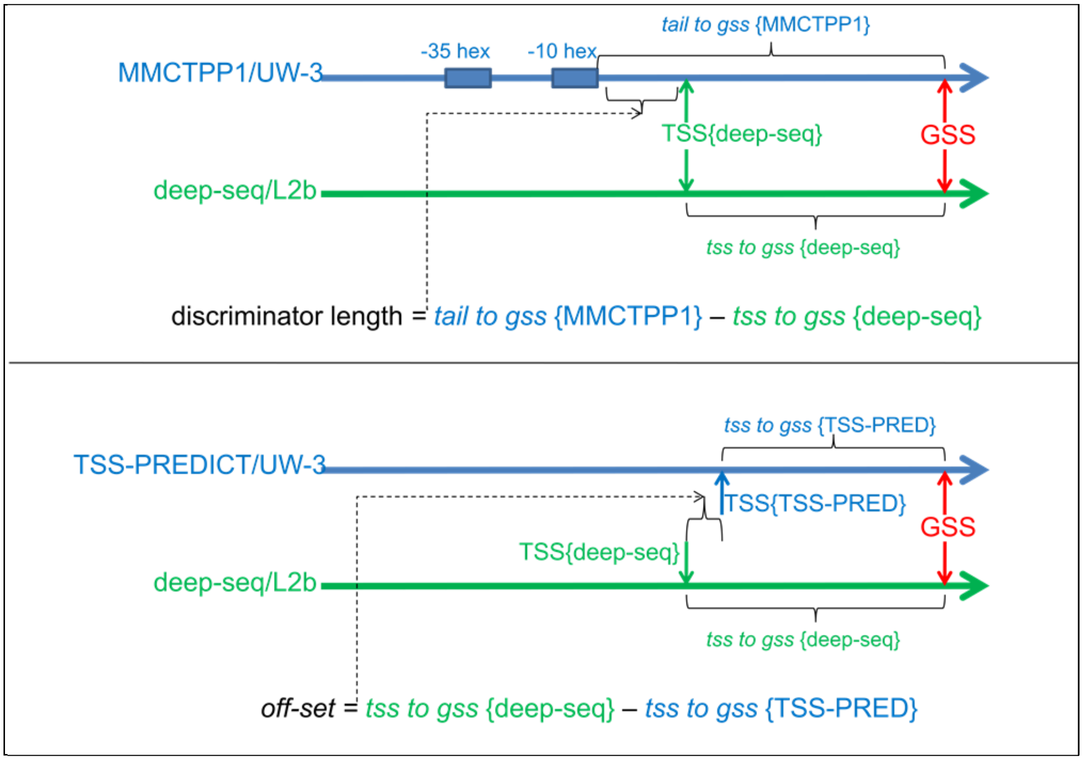


Figure 10. Diagrams illustrating the calculations of the discriminator length (top) and *off-set* (bottom).

6.4 169 mapped *C. trachomatis* promoters

The 317 identified L2b TSS locations were mapped onto UW-3, and the 317 UW-3 TSS locations were compared with the promoter predictions of the two algorithms. TSS-PREDICT was considered correct if the predicted TSS was within 6 nts (on either side) of the experimental TSS. MMCTPP1 was considered correct if the predicted discriminator length was in the range {4-14}.

Figure 11 illustrates the effectiveness of the prediction algorithms. Of the total 317 identified transcription start sites, 148 were not predicted by either MMCTPP1 or TSS-PREDICT. Eighty-nine TSSs were predicted by MMCTPP1, 138 by TSS-PREDICT, and 58 were jointly predicted by the two algorithms.

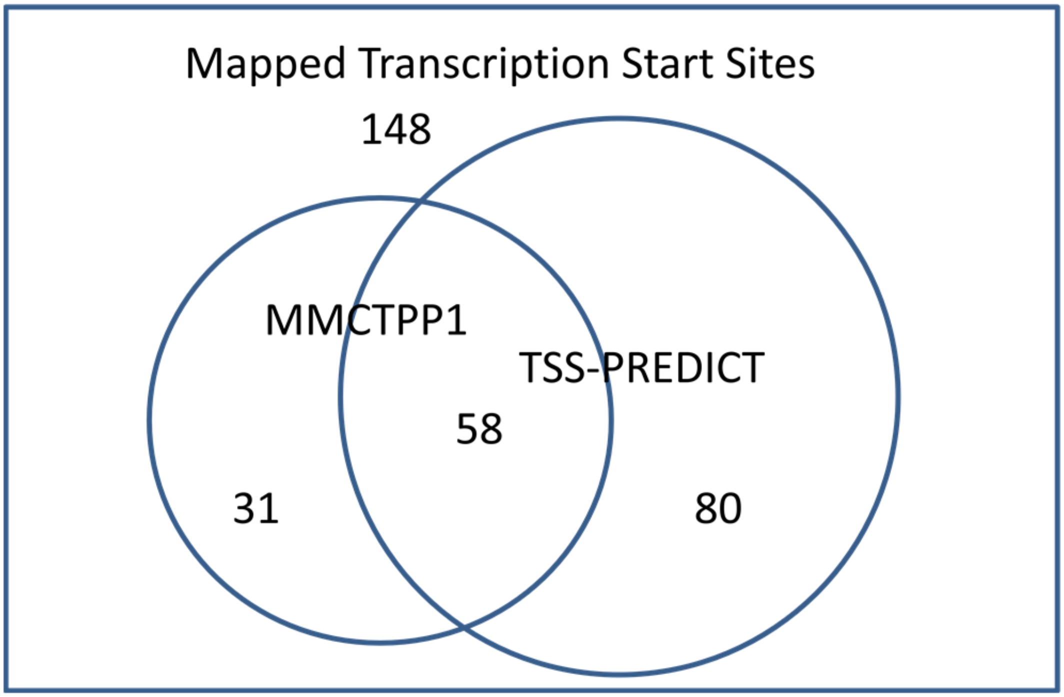


Figure 11. Effectiveness of MMCTPP1 and TSS-PREDICT in predicting 317 mapped transcription start sites.

Appendix I: MMCTPP1/L2b TSS matches describes the 89 MMCTPP1 hits and also notes the 58 TSSs that are jointly predicted by TSS-PREDICT. Table 13 displays the first 4 rows of the data-file. All discriminator lengths are restricted to be between 4 and 14. *seq32 ID* is the identifier used by MMCTPP1 for the 32-mer in a given position upstream from each GSS. The remainder of the columns define the promoter region and indicate the co-predictions with TSS-PREDICT.

Table 13. The first 4 of 89 rows in *Appendix I: MMCTPP1/L2b TSS matches*. Each row defines a matched promoter and indicates if the promoter is co-predicted by TSS-PREDICT.

disc len {MMCTPP1}	seq32 ID	-35 hex loc	-10 hex	spacer	-35 hex to tss	TSS-PREDICT
5	CT007_112	7142	TATGAT	17	TTGCTAAAAATTTTATTAAGCAGTATGATCTACCA	1
6	CT016_067	17572	TACAAT	17	TTGTCAAAAATGTACCCCTTAACTACAATGCCGAGG	1
6	CT022_102	27393	TAAAAT	17	GTGCATTTTTTCTTGCTTTTTCATAAAATGTTCGGG	0
6	CT025_060	29880	TATCCT	18	TTGAAAATCAAGCTAATGATGCTGTATCCTCTGGGGA	0

Appendix J: TSS-PREDICT only/L2b TSS matches enumerates the 80 TSSs that were predicted by TSS-PREDICT alone. Table 14 exhibits the first 4 rows of the data-file. The difference between experimental and predicted TSS locations relative to the GSS, *off-set*, is restricted to be between -6 and +6. The remaining columns define each promoter region.

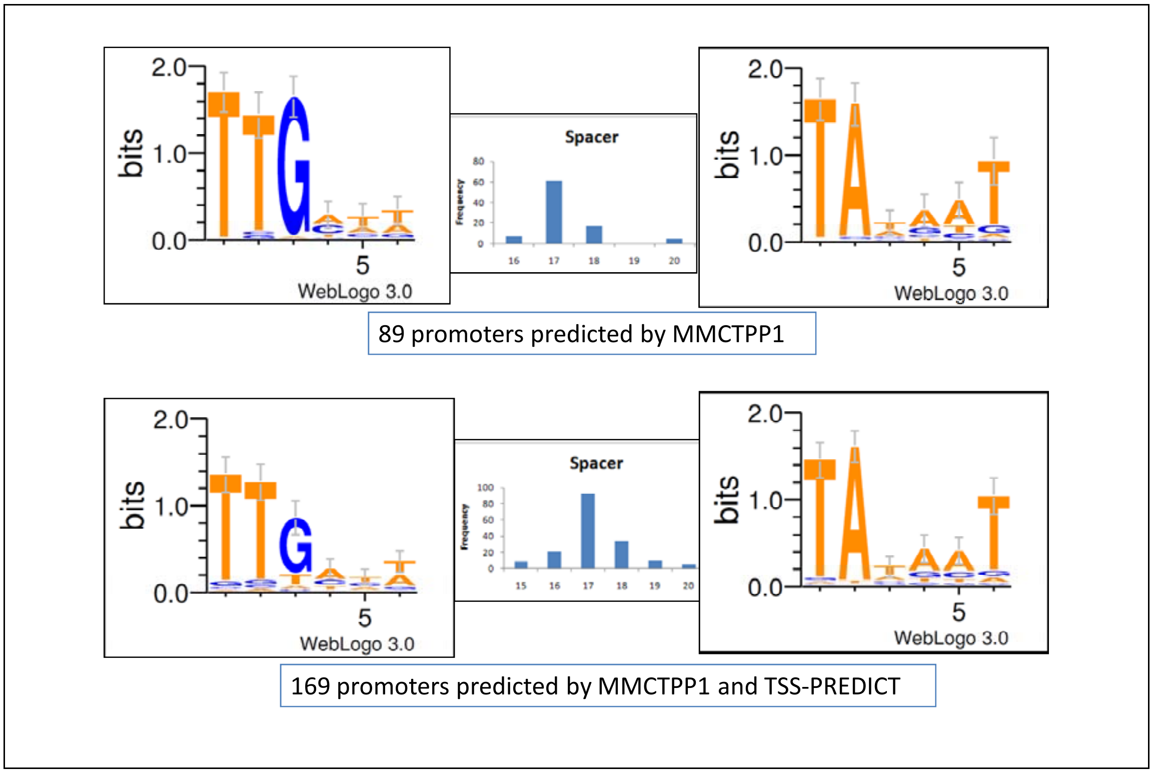
Table 14. The first 4 of 80 rows in *Appendix J: TSS-PREDICT only/L2b TSS matches*. Each row defines a matched promoter and indicates the difference between experimental and predicted TSS locations relative to the GSS (off-set).

UW-3 num	tss loc {TSS-PRED}	-35 hex	spacer	-10 hex	disc len {TSS-PRED}	off-set
CT005	6275	ctccaa	15	tatact	7	2
CT013	13578	gtgaca	19	tatact	5	-2
CT017	18453	ttgact	17	aataat	6	1
CT021	27455	ttgaca	18	tagtat	6	0

The MMCTPP1 algorithm was trained on 26 experimentally identified *C. trachomatis* σ^{66} promoters from 25 genes. For those 25 genes, 16 promoters were confirmed by the deep sequencing transcriptome. Five training set homologous gene transcripts (CT062, CT322, CT518, CT701 and CT752) were not present in the deep sequencing transcriptome and 4 of the reported TSSs (CT046, CT439m, CT442, and CT674) differed significantly from those in the training set.

The analysis reported here maps 169 *C. trachomatis* σ^{66} promoters, resulting in a four-fold increase in experimentally defined *C. trachomatis* σ^{66} promoters. The results demonstrate substantial agreement between the experimental *C. trachomatis* transcriptome and the two promoter prediction algorithms. Because it makes two predictions for each gene, TSS-PREDICT is expected to strike more hits than MMCTPP1 which was designed to avoid false-positives and predicts only 479 promoters for the entire *C. trachomatis* genome. Additionally, MMCTPP1 did quite well for a first-pass model built with a training set of only 26 promoters.

The WebLogos (http://weblogo.threeplusone.com/create.cgi) shown in Figure 12 illustrate the differences between the 89 promoters correctly predicted by MMCTPP1 (with 58 TSS-PREDICT co-predictions) and the entire set of 169 promoters that includes predictions by TSS-PREDICT alone. The TTGxxx motif of the -35 hexamer and the TAxaaT motif of the -10 hexamer continue to be dominant in the larger group, but are not as pronounced as in the smaller subset. A major difference can be observed in the distributions of the spacers. Spacers of length 15 and 19 are present in the larger group, while absent in those predicted by MMCTPP1. The spacers predicted by MMCTPP1 are 27% GC, compared to a composition of 43% GC for the entire *C. trachomatis* genome.



89 promoters predicted by MMCTPP1

169 promoters predicted by MMCTPP1 and TSS-PREDICT

Figure 12. Comparison of the promoter profiles generated by the 89 promoters predicted by MMCTPP1 and the 169 promoters predicted by MMCTPP1 and TSS-PREDICT together.

The 169 mapped promoters defined in Appendices I and J will inform the investigation *C. trachomatis* gene expression regulation. They will also enhance the second-pass multivariate *C. trachomatis* σ^{66} promoter prediction algorithm, MMCTPP2. The diversity of spacer regions and binding hexamers, as well as the substantially larger training set, will contribute to improved models of the σ^{66} binding site and biophysical features that characterize the extended promoter region.

Chapter 7

Conclusions and Outlook

7.1 Conclusions

The research presented here has demonstrated the following:

1. Established *C. trachomatis* σ^{66} promoters can be accurately predicted by a strategy that employs a small training set of *C. trachomatis* σ^{66} promoters.

2. Higher order DNA structures within the extended promoter region, as well as the primary structure of the promoter, contribute to the predictive model.

3. Genome-wide predictions based on MMCTPP1 facilitate the mapping of new *C. trachomatis* σ^{66} promoters.

Promoter predictions (Appendix C) and newly mapped promoters (Appendices I and J) will inform the investigation of *C. trachomatis* gene expression regulation.

7.2 Outlook

The stage has been set for an exciting second round of *C. trachomatis* σ^{66} promoter modeling. MMCTPP2, trained on the 169 newly mapped *C. trachomatis* σ^{66} promoters, is expected to refine the multiple metric model. The new model should more accurately characterize the promoter region in terms of multiple biophysical structures as well as predictablilty. Once an accurate organism-specific promoter model for *C. trachomatis* σ^{66} is developed, it will be possible to explore the similarities and/or differences between *C. trachomatis* σ^{66} and *E. coli* σ^{70} promoters.

References

1. Tan M: **Regulation of gene expression**. In: *Chlamydia genomics and pathogenesis.* Edited by Bavoil PM, Wyrick PB. Wymondham, U.K.: Horizon Bioscience; 2006: 103-132.
2. Hawley DK, McClure WR: **Compilation and analysis of Escherichia coli promoter DNA sequences**. *Nucleic Acids Res* 1983, **11**(8):2237-2255.
3. Rosenberg M, Court D: **Regulatory sequences involved in the promotion and termination of RNA transcription**. *Annu Rev Genet* 1979, **13**:319-353.
4. Siebenlist U, Simpson RB, Gilbert W: **E. coli RNA polymerase interacts homologously with two different promoters**. *Cell* 1980, **20**(2):269-281.
5. Lisser S, Margalit H: **Compilation of E. coli mRNA promoter sequences**. *Nucleic Acids Res* 1993, **21**(7):1507-1516.
6. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H *et al*: **RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation**. *Nucleic Acids Res* 2008, **36**(Database issue):D120-124.
7. Hertz GZ, Stormo GD: **Escherichia coli promoter sequences: analysis and prediction**. *Methods Enzymol* 1996, **273**:30-42.
8. Kanhere A, Bansal M: **Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes**. *Nucleic Acids Res* 2005, **33**(10):3165-3175.
9. Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW: **A DNA structural atlas for Escherichia coli**. *J Mol Biol* 2000, **299**(4):907-930.
10. Wang H, Benham CJ: **Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress**. *BMC Bioinformatics* 2006, **7**:248.
11. Brunelle BW, Nicholson TL, Stephens RS: **Microarray-based genomic surveying of gene polymorphisms in Chlamydia trachomatis**. *Genome Biol* 2004, **5**(6):R42.
12. Bavoil PM, Hsia R, Ojcius DM: **Closing in on Chlamydia and its intracellular bag of tricks**. *Microbiology* 2000, **146 (Pt 11)**:2723-2731.
13. Mathews SA, Volp KM, Timms P: **Development of a quantitative gene expression assay for Chlamydia trachomatis identified temporal expression of sigma factors**. *FEBS Lett* 1999, **458**(3):354-358.
14. Mathews SA, Stephens RS: **DNA structure and novel amino and carboxyl termini of the Chlamydia sigma 70 analogue modulate promoter recognition**. *Microbiology* 1999, **145 (Pt 7)**:1671-1681.
15. Grech B, Maetschke S, Mathews S, Timms P: **Genome-wide analysis of chlamydiae for promoters that phylogenetically footprint**. *Res Microbiol* 2007, **158**(8-9):685-693.

16. Hefty PS, Stephens RS: **Chlamydial type III secretion system is encoded on ten operons preceded by sigma 70-like promoter elements**. *J Bacteriol* 2007, **189**(1):198-206.
17. Mathews S, Timms P: **In silico identification of chamydial promoters and their role in the regulation of development**. In: *Chlamydia genomics and pathogenesis.* Edited by Bavoil PM, Wyrick PB. Wymondham, U.K.: Horizon Bioscience; 2006: 133-156.
18. Wagner R: **Transcription regulation in prokaryotes**. Oxford ; New York: Oxford University Press; 2000.
19. Iyer LM, Koonin EV, Aravind L: **Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer**. *Gene* 2004, **335**:73-88.
20. Belland RJ, Zhong G, Crane DD, Hogan D, Sturdevant D, Sharma J, Beatty WL, Caldwell HD: **Genomic transcriptional profiling of the developmental cycle of Chlamydia trachomatis**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8478-8483.
21. Staden R: **Computer methods to locate signals in nucleic acid sequences**. *Nucleic Acids Res* 1984, **12**(1 Pt 2):505-519.
22. Huerta AM, Collado-Vides J: **Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals**. *J Mol Biol* 2003, **333**(2):261-278.
23. Shultzaberger RK, Chen Z, Lewis KA, Schneider TD: **Anatomy of Escherichia coli sigma70 promoters**. *Nucleic Acids Res* 2007, **35**(3):771-788.
24. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**(9):755-763.
25. Hudson DL, Cohen ME: **Neural networks and artificial intelligence for biomedical engineering**. New York: IEEE Press; 2000.
26. Reese MG: **Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome**. *Comput Chem* 2001, **26**(1):51-56.
27. Hampshire JB, Waibel AH: **A novel objective function for improved phoneme recognition using time-delay neural networks**. *IEEE Trans Neural Netw* 1990, **1**(2):216-228.
28. Gordon JJ, Towsey MW, Hogan JM, Mathews SA, Timms P: **Improved prediction of bacterial transcription start sites**. *Bioinformatics* 2006, **22**(2):142-148.
29. Towsey M, Timms P, Hogan J, Mathews SA: **The cross-species prediction of bacterial promoters using a support vector machine**. *Comput Biol Chem* 2008, **32**(5):359-366.
30. Liu H, Wong L: **Data mining tools for biological sequences**. *J Bioinform Comput Biol* 2003, **1**(1):139-167.
31. Agresti A: **An introduction to categorical data analysis**. New York: Wiley; 1996.
32. Hosmer DW, Lemeshow S: **Applied logistic regression**, 2nd edn. New York: Wiley; 2000.

33. Munteanu MG, Vlahovicek K, Parthasarathy S, Simon I, Pongor S: **Rod models of DNA: sequence-dependent anisotropic elastic modelling of local bending phenomena**. *Trends Biochem Sci* 1998, **23**(9):341-347.
34. SantaLucia J, Jr., Allawi HT, Seneviratne PA: **Improved nearest-neighbor parameters for predicting DNA duplex stability**. *Biochemistry* 1996, **35**(11):3555-3562.
35. Satchwell SC, Drew HR, Travers AA: **Sequence periodicities in chicken nucleosome core DNA**. *J Mol Biol* 1986, **191**(4):659-675.
36. Uljanov N, James, T.: **Statistical analysis of DNA duplex structural features**. *Methods in Enzymology* 1995, **261**:90-115.
37. Bi C, Benham CJ: **WebSIDD: server for predicting stress-induced duplex destabilized (SIDD) sites in superhelical DNA**. *Bioinformatics* 2004, **20**(9):1477-1479.
38. Niehus E, Cheng E, Tan M: **DNA Supercoiling-Dependent Gene Regulation in Chlamydia**. *J Bacteriol* 2008.
39. Afifi AA, Clark V, May S: **Computer-aided multivariate analysis**, 4th edn. Boca Raton: Chapman & Hall/CRC; 2004.
40. Box GEP, Jenkins GM, Reinsel GC: **Time series analysis : forecasting and control**, 3rd edn. Englewood Cliffs, N.J.: Prentice Hall; 1994.
41. Picard RR, Cook RD: **Cross-Validation of Regression Models**. *Journal of the American Statistical Association* 1984, **79**(387):575-583.
42. Burden S, Lin YX, Zhang R: **Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences**. *Bioinformatics* 2005, **21**(5):601-607.
43. Mallios RR, Ojcius DM, Ardell DH: **An iterative strategy combining biophysical criteria and duration hidden Markov models for structural predictions of Chlamydia trachomatis sigma66 promoters**. *BMC Bioinformatics* 2009, **10**:271.
44. Albrecht M, Sharma CM, Reinhardt R, Vogel J, Rudel T: **Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome**. *Nucleic Acids Res* 2009.
45. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**(1):57-63.
46. Thomson NR, Holden MT, Carder C, Lennard N, Lockey SJ, Marsh P, Skipp P, O'Connor CD, Goodhead I, Norbertzcak H *et al*: **Chlamydia trachomatis: genome sequence analysis of lymphogranuloma venereum isolates**. *Genome Res* 2008, **18**(1):161-171.

Appendices

Appendix A: Strategy Details

bold: R, Unix, or SPSS scripts; HMM programs
italics:data-files

A-1. create basic datafile; biophysical variables for upstream regions of all 895 genes

1a. create biophysical variables for each 32mer of each upstream region; 100 genes/file*: *parse32_1_100.xls*

*xlsx files of Office 2007 can be longer

upstream600.txt (*gtable.txt, genome.txt*) -> *ups600out.txt* -> *ups600out.xls*

parse32.txt (*ups600out.txt*) -> *parse32_1_100.txt* -> *parse32_1_100.xls*
copy worksheet to temporary Excel worksheet; close *parse32_1_100.xls*.
delete all but fasta, position and seq_32;
add _ between fasta & position; open column before seq_32;
=CONCATENATE(A1,B1,TEXT(C1,"000"));
copy and paste (special) id to column K of *parse32_1_100.xls*

from ups600out.xls, copy rows 101-200 into doc, untable and submit to
bend.it: curve(window32) & gc(window 32) and
plot.it: stable(#33, window 2), twist(#44 window 2), rigid(#31 window 3)
paste in rows 33, 3, 4 (delete columns X, V and U)

from *ups600out.xls*, copy 15 rows (limit 10Kb) into doc, untable and submit to
webSIDD (window 63) insert in columns W-Z of *parse32_1_100.xls;* delete W at end

insert column after P, copy column S and change to 1s and 0s
insert row for titles; copy parse title row
from *parse32_1_100.xls*, copy and paste all computed values

1b. create training set *parse32ts.xls*

extract 27 genes from *parse32_x_y.xls* to *parse32ts27.xls*
copy columns K & L to Word, untable; save as *ts32.txt*

A-2. iterate dhmm & score training set

ts0.txt: verified hexamer-spacer-hexamer for each training set promoter
ts5e.txt: each sequence in *ts0.txt* extended by 5 nt on each end
ts32.txt: from #2

```
rmallios@engapps00:~/p0$ cat dh
durahmmer -5 6 -3 6 -s 16 -S 20 -p 1 -u 28.5:21.5:21.5:28.5 -C ts0.txt>ts0_1.hmm
```

```
rmallios@engapps00:~/p0$ cat hs
hmmsearch -E 90000 ts0_1.hmm  ts5e.txt>ts5e_ts0_1.txt
hmmsearch -E 90000 ts0_1.hmm  ts32.txt>ts32_ts0_1.txt
```

modify ts0.txt -> ts0_1.txt according to results of **hmmsearch** until stable

```
rmallios@engapps00:~/p0$ cat dog
cat ts32_ts0_1.txt |egrep from|tr -d ":,"|tr "_" " "|cut -d' ' -f1,2,8,10,12|sort -t" " -rn -k2|sort -t" " -k1> ts32_ts0_1score.txt
```

download *ts5p_ts0_1.txt; ts32_ts0_1.txt, ts32_ts0_1score.txt*
open *ts32_ts0_1score.txt* in Excel; sort; copy to *parse32ts27.xls* (delim)
add variable promoter0 to *parse32ts27.xls* (from information in *ts5p_ts0_1.txt*)

A-3. logistic regression model

open *parse32ts27.xls* in SPSS
run **lr29.sps** (lr29.txt) to transform
run Stepwise Binary Logistic Regression
add model coefficients to **lr29.sps** (lr29.txt)
SPSS output file for model 2: **m2.pdf**

A-4. predict genome

open *parse32_1_100.xls*
copy and untable genes 1-50 (rows 2 to 30,001) columns K & L to *g1_50.txt*

```
rmallios@engapps00:~/p2$ cat hs_dog
hmmsearch -E 90000 ts2_1.hmm g1_50.txt>g1_50_ts2_1.txt
cat g1_50_ts2_1.txt|egrep from|tr -d ":,"|tr "_" " "|cut -d' ' -f1,2,8,10,12|sort -t" " -rn -k2|sort -t" " -k1>g1_50_ts2_1score.txt
```

download *g1_50_ts2_1score.txt*; open with Excel; sort
copy and paste to end of *parse32_1_100.xls*; label start2 end2 hmm2

open *parse32_1_100.xls* with SPSS;
run syntax file **lr29.spv** (lr29.txt): hmm_score and p2 function

copy p2 to *parse32_1_100.xls*
select cases where p2>=.5; start2=32; position >=40 and position <=325
copy and paste values into *p2_predictions.xls*
filter final results into *m2_ct_genome.xls*

A-5. compare predictions

prepare tables and run **match_nnpp.txt** (*m2_table.txt, nnpp_table.txt, g_table.txt*) -> *match_nnpp_out.txt*
prepare tables and run **match_tssp.txt** (*m2_table.txt*, *t_table.txt*, *g_table.txt*) -> *match_tssp_out.txt*

with Excel, delete unnecessary columns from *match_nnpp_out.txt* and *match_tssp_out.txt*. Name columns name1 and POSITION.

open *p2_predictions.xls* with SPSS. Sort on name1 and POSITION. Merge Excel versions of *match_nnpp_out.txt* and *match_tssp_out.txt*.

filter desired tables.

Appendix B: Scripts and txt files

B-1. R script: Extracts 600 nt upstream region from all 895 *CT* genes

```
ngenes<- 895
n<- 600
gtable <-
read.table("gtable.txt",header=TRUE,colClasses=c("integer","character",
"character","integer","integer","integer","integer","integer"))
attach(gtable)
genome <- scan("genome.txt", what=character())
f= file("ups600out.txt","w")
for (i in 1:ngenes){

cat(name1[i], file=f)
cat(" ", file=f)
cat(name2[i], file=f)
cat(" ", file=f)
cat(start[i], file=f)
cat(" ", file=f)
cat(end[i], file=f)
cat(" ", file=f)
cat(strand[i], file=f)
cat(" ", file=f)
cat(up_space[i], file=f)
cat(" ", file=f)
cat(hour[i], file=f)
cat(" ", file=f)
cat(">", file=f)
cat(name1[i], file=f)
cat(" ", file=f)
if (strand[i]==1) {
s<- start[i] - n -1
for(j in 1:n) {
cat(genome[s+j],file=f)
}
}
else{
s<- end[i] + n + 1
for(j in 1:n) {
x<- genome[s-j]
if(x=="A")
y<- "T"
if(x=="C")
y<- "G"
if(x=="G")
y<- "C"
if(x=="T")
y<- "A"
cat(y,file=f)
}
}
cat(file=f, sep="\n")
}
close(f)
```

B-2. ts0.txt: initial training set sequences

```
>1CT046 hctB
TGGTTAGTTTTTAATAAAAAGTTAAAAA
>1CT080 ltuB
TTATGAAAAACAATTTTTTAATTTAAAAT
>1CT091 yscU
TTGAGAAAAACATTTATATACGGTAACTT
>1CT098 rs1
TTGCCTTTTTTAAGGTGAATATTTACACT
>1CT111 groES
TTGCAAAAAAAGCGAGGACTTTGCTATCGT
>1CT322 tuf
TTGATAATAATCCGCGTCTGAAGTTACTAT
>1CT323 infA
TTGACATTTTCTGTTTAGTCGATATAAT
>1CT377 ltuA
TGCAGAGTTTTTATTTTAAATATGTTATAAT
>1CT394 hrcA
TTGACCAGTGGAGACGGTTTTCTTATAAT
>1CT442 crpA
GGGTTTTTGAAAAAAACAAGTGTTTGTGTAG
>1CT444a omcA
TTGATATAATTTTTATTTTATAATGTAAT
>1CT444b omcA
AATTGCTTTTATCGATAAAAGAAACTTCAAG
>1CT518 rl14
CTGTTGTTGTTCGAGTCGAAAGGGTATACT
>1CT557 lpdA
TTGAGATTTTATCCACCCAGATGTACAAC
>1CT576 lrcH_1
TTGTTAAATCAGATCGTTAGAATTTAATAT
>1CT665 -
TTGTATCTTTTTAGAACGGGAAGGGTTGAAA
>1CT674 yscC
TTGCAAGATAGAGGGCAAATAGATATATT
>1CT681a ompA
TATACAAAAATGGCTCTCTGCTTTATTGC
>1CT681b ompA
GTGCCGCCAGAAAAAGATAGCGAGCACAAA
>1CT701 secA_2
TGTATAGGCGCCTTTAAATAAGAGGGTAGGTT
>1CT743 hctA
TTGCATGAATTTGAACAAACAAACTAATTA
>1CT863 -
TTGCATGAAAAATACTTTTTAGATAAGTT
>2ct062_064
TTGCTATAAAAAGAACAGGATAGATAAGAT
>2ct286_067
TTGCATCATTATCATAAATGTCGTATATG
>2ct439m_069
TTGCAAACAAAGATATTCTTATTCTATATT
>2ct559_055
TTGGCACTAATCTCCCCATTTGCTATGGT
>2ct596_066
```

```
TTGGTTCTATACAAGAAATTTGTTAGGAT
>2ct708_069
TTGATTTAGCGGAAGTAAAAAGGTACAAG
>2ct752_064
TGGACAAAGCTTAGAAGAGAACGATAACAT
```

B-3. ts5e.txt: ts0.txt extended by 5 nt on each end

```
>1CT046 hctB
GTGTGTGGTTAGTTTTTAATAAAAAGTTAAAAACTAAC
>1CT080 ltuB
ATGGTTTATGAAAAACAATTTTTTAATTTAAAATTAGAA
>1CT091 yscU
CTTTCTTGAGAAAAACATTTATATACGGTAACTTGCGAA
>1CT098 rs1
AAATCTTGCCTTTTTTAAGGTGAATATTTACACTACTCT
>1CT111 groES
ACCAGTTGCAAAAAAGCGAGGACTTTGCTATCGTTCTTC
>1CT322 tuf
AAAGCTTGATAATAATCCGCGTCTGAAGTTACTATGCTCG
>1CT323 infA
GTTGTTTGACATTTTCTGTTTAGTCGATATAATCGCTC
>1CT377 ltuA
TTGTTTGCAGAGTTTTTATTTTAAATATGTTATAATCTGTC
>1CT394 hrcA
AATTCTTGACCAGTGGAGACGGTTTTCTTATAATGACAC
>1CT442 crpA
TAGATGGGTTTTTGAAAAAAACAAGTGTTTGTGTAGACTCC
>1CT444a omcA
AACAATTGATATAATTTTTATTTTATAATGTAATATTGT
>1CT444b omcA
AAAAGAATTGCTTTTATCGATAAAAGAAACTTCAAGAGCCC
>1CT518 rl14
AAAAACTGTTGTTGTTCGAGTCGAAAGGGTATACTCGCAC
>1CT557 lpdA
CCTCATTGAGATTTTATCCACCCAGATGTACAACCCGGG
>1CT576 lrcH_1
TTAACTTGTTAAATCAGATCGTTAGAATTTAATATTGTTA
>1CT665 -
TCGCATTGTATCTTTTTAGAACGGGAAGGGTTGAAATATAA
>1CT674 yscC
TGAAGTTGCAAGATAGAGGGCAAATAGATATATTCTGCC
>1CT681a ompA
AAAGATATACAAAAATGGCTCTCTGCTTTATTGCTAAAT
>1CT681b ompA
ACGCAGTGCCGCCAGAAAAAGATAGCGAGCACAAAGAGAG
>1CT701 secA_2
CTTGTTGTATAGGCGCCTTTAAATAAGAGGGTAGGTTCGTTT
>1CT743 hctA
AATGGTTGCATGAATTTGAACAAACAAACTAATTAAAAAT
>1CT863 -
CCAACTTGCATGAAAAATACTTTTTAGATAAGTTCCCTC
>2ct062_064
TTGCCTTGCTATAAAAAGAACAGGATAGATAAGATGTTGC
>2ct286_067
AAAAGTTGCATCATTATCATAAATGTCGTATATGCTTGA
>2ct439m_069
```

```
ACCCCTTGCAAACAAAGATATTCTTATTCTATATTTCCCT
>2ct559_055
CCCGATTGGCACTAATCTCCCCATTTGCTATGGTGAGTG
>2ct596_066
GGATCTTGGTTCTATACAAGAAATTTGTTAGGATCGTCT
>2ct708_069
TTTCATTGATTTAGCGGAAGTAAAAAGGTACAAGTAACA
>2ct752_064
TCTTCTGGACAAAGCTTAGAAGAGAACGATAACATAGATG
```

B-4. ts.hmm: duration HMM model file for final model M2, output from durahmmer

```
HMMER2.0  [2.3.2]
NAME  ts
DESC  durahmmer model
LENG  32
ALPH  Nucleic
RF    no
CS    no
MAP   no
COM   durahmmer -5 6 -3 6 -s 16 -S 20 -p 1 -u 28.5:21.5:21.5:28.5 -C
ts.txt
COM   hmmcalibrate ts.hmm
NSEQ  26
DATE  Thu Sep  4 15:44:25 2008
CKSUM 0
XT      -8234     -5  -1000  -1000  -8234     -5  -8234     -5
NULT      -5  -8234
NULE     189   -218   -218    189
EVD    -5.953111   0.549023
HMM        A      C      G      T
         m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
            0      *      *
     1  -4755  -4755  -4755   1772
     -      0      0      0      0
     -      0      *      *      *      0      *      0      0      *
     2  -4755  -4755  -1390   1658
     -      0      0      0      0
     -      0      *      *      *      0      *      0      *      *
     3  -2582  -4755   2119  -4755
     -      0      0      0      0
     -      0      *      *      *      0      *      0      *      *
     4    111    839  -1483   -305
     -      0      0      0      0
     -      0      *      *      *      0      *      0      *      *
     5    111   -483   -898    570
     -      0      0      0      0
     -      0      *      *      *      0      *      0      *      *
     6    918  -2483   -898    111
     -      0      0      0      0
     -      0      *      *      *      0      *      0      *      *
     7    412   -845   -296    194
     -      0      0      0      0
     -      0      *      *      *      0      *      0      *      *
     8    412   -845   -296    194
     -      0      0      0      0
```

```
 -      0      *      *      *      0      *      0      *      *
 9    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
10    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
11    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
12    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
13    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
14    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
15    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
16    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
17    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
18    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
19    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
20    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
21    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      *      0      *      *
22    412   -845   -296    194
 -      0      0      0      0
 -   -179      *  -3096      *      0      *      0      *      *
23    412   -845   -296    194
 -      0      0      0      0
 -  -1217      *   -811      *      0      *      0      *      *
24    412   -845   -296    194
 -      0      0      0      0
 -   -377      *  -2119      *      0      *      0      *      *
25    412   -845   -296    194
 -      0      0      0      0
 -     -5      *  -8050      *      0      *      0      *      *
26    412   -845   -296    194
 -      0      0      0      0
 -      0      *      *      *      0      0      *      *      *
27  -4755  -4755  -4755   1772
 -      0      0      0      0
```

```
     -      0      *      *      *      0      *      0      *      *
    28   1658  -4755  -2256  -2582
     -      0      0      0      0
     -      0      *      *      *      0      *      0      *      *
    29    280    102  -1483    280
     -      0      0      0      0
     -      0      *      *      *      0      *      0      *      *
    30    918   -898   -161   -889
     -      0      0      0      0
     -      0      *      *      *      0      *      0      *      *
    31   1017   -898  -1483   -305
     -      0      0      0      0
     -      0      *      *      *      0      *      0      *      *
    32  -1889  -2483   -161   1280
     -      *      *      *      *
     -      *      *      *      *      *      *      *      *      0
//
```

B-5 trans.sps: SPSS syntax file for transformations

```
USE ALL.
COMPUTE filter_$=(end=32 and position>=40 and position<=325 ).
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
compute curveL32=lag(curve,32).
compute curveL64=lag(curve,64).
compute curveL96=lag(curve,96).
compute curveL128=lag(curve,128).
compute gcL32=lag(gc,32).
compute gcL64=lag(gc,64).
compute gcL96=lag(gc,96).
compute gcL128=lag(gc,128).
compute stableL32=lag(stable,32).
compute stableL64=lag(stable,64).
compute stableL96=lag(stable,96).
compute stableL128=lag(stable,128).
compute twistL32=lag( twist,32).
compute  twistL64=lag( twist,64).
compute  twistL96=lag( twist,96).
compute  twistL128=lag( twist,128).
compute rigidL32=lag(  rigid,32).
compute  rigidL64=lag(  rigid,64).
compute rigidL96=lag(  rigid,96).
compute rigidL128=lag(  rigid,128).
compute siddL32=lag(  sidd,32).
compute siddL64=lag(  sidd,64).
compute siddL96=lag(  sidd,96).
compute siddL128=lag(  sidd,128).
RECODE hour (1=1) (ELSE=0) INTO hour1.
RECODE hour (3=1) (ELSE=0) INTO hour3.
RECODE hour (8=1) (ELSE=0) INTO hour8.
RECODE hour (16=1) (ELSE=0) INTO hour16.
RECODE hour (24=1) (ELSE=0) INTO hour24.
compute sidd_h1 = sidd*hour1.
compute sidd_h3 = sidd*hour3.
compute sidd_h8 = sidd*hour8.
compute sidd_h16 = sidd*hour16.
compute sidd_h24 = sidd*hour24.
execute.
```

B-6 ts.txt: training set sequences for M2 model

```
>CT046
TGGTTAGTTTTTAATAAAAAGTTAAAAA
>CT062
TTGCTATAAAAAGAACAGGATAGATAAGAT
>CT080
TTATGAAAAACAATTTTTTAATTTAAAAT
>CT091
TTGAGAAAAACATTTATATACGGTAACTT
>CT098
TTGCCTTTTTTAAGGTGAATATTTACACT
>CT111
TTGCAAAAAAGCGAGGACTTTGCTATCGT
>CT286
TTGCATCATTATCATAAATGTCGTATATG
>CT322
TTGATAATAATCCGCGTCTGAAGTTACTAT
>CT323
ttgtTTGACATTTTCTGTTTAGTCGATATAAT
>CT377
tTGCAGAGTTTTTATTTTAAATATGTTATAAT
>CT394
TTGACCAGTGGAGACGGTTTTCTTATAAT
>CT439m
TTGCAAACAAAGATATTCTTATTCTATATT
>CT442
TTGAAAAAAACAAGTGTTTGTGTAGact
>CT444a
TTGATATAATTTTTATTTTATAATGTAAT
>CT444b
TTGCTTTTATCGATAAAAGAAACTTCAAG
>CT518
TTGTTGTTCGAGTCGAAAGGGTATACTcg
>CT557
TTGAGATTTTATCCACCCAGATGTACAAC
>CT559
TTGGCACTAATCTCCCCATTTGCTATGGT
>CT576
TTGTTAAATCAGATCGTTAGAATTTAATAT
>CT596
TTGGTTCTATACAAGAAATTTGTTAGGAT
>CT674
TTGCAAGATAGAGGGCAAATAGATATATT
>CT701
ttgtTGTATAGGCGCCTTTAAATAAGAG
>CT708
TTGATTTAGCGGAAGTAAAAAGGTACAAG
>CT743
TTGCATGAATTTGAACAAACAAACTAATTA
>ct752
TGGACAAAGCTTAGAAGAGAACGATAACAT
>CT863
TTGCATGAAAAATACTTTTTAGATAAGTT
```

B-7. R script: Scans promoters predicted by NNPP2.2 for matches with m3

```
n_m3<- 485
n_nnpp<- 614
n_g<- 895
m3_table <-
read.table("m3_table.txt",header=TRUE,colClasses=c("character","integer
","integer"))
attach(m3_table)
nnpp_table <-
read.table("nnpp_table.txt",header=TRUE,colClasses=c("character","integ
er","integer","integer","integer","integer","integer","integer","intege
r","integer","integer","integer","integer","integer","integer","integer
","integer","integer","integer","character","character","character","ch
aracter","character","character","character","character","character"))
attach(nnpp_table)
g_table <-
read.table("g_table.txt",header=TRUE,colClasses=c("character","characte
r","integer","integer"))
attach(g_table)
f= file("match_nnpp_out.txt","w")
for (i in 1:n_m3){
cat(file=f, m_name[i])
cat(file=f, " ")
cat(file=f, m_pos[i])
cat(file=f, " ")
cat(file=f, m_start[i])
for (k in 1:n_g){
if (m_name[i] == g_name1[k]){
cat(file=f, " ")
cat(file=f, g_name2[k])
cat(file=f, " ")
cat(file=f, g_hour[k])
cat(file=f, " ")
cat(file=f, g_op[k])
}
}
ms<-m_pos[i]-m_start[i]+1
me<-m_pos[i]-31
max_p<-0
for (j in 1:n_nnpp){
if (m_name[i] == n_name[j]){
ns<-326-s1[j]
ne<-ns-49
if(ns>=ms && ne<=me){
max_p<- p1[j]
max_s<- ns
max_e<- ne
max_seq<- seq1[j]
}
ns<-326-s2[j]
ne<-ns-49
if(ns>=ms && ne<=me){
max_p<- p2[j]
max_s<- ns
max_e<- ne
max_seq<- seq2[j]
}
ns<-326-s3[j]
ne<-ns-49
if(ns>=ms && ne<=me){
max_p<- p3[j]
max_s<- ns
max_e<- ne
max_seq<- seq3[j]
}
ns<-326-s4[j]
ne<-ns-49
if(ns>=ms && ne<=me){
max_p<- p4[j]
max_s<- ns
max_e<- ne
max_seq<- seq4[j]
}
ns<-326-s5[j]
ne<-ns-49
if(ns>=ms && ne<=me){
max_p<- p5[j]
max_s<- ns
max_e<- ne
max_seq<- seq5[j]
}
ns<-326-s6[j]
ne<-ns-49
if(ns>=ms && ne<=me){
max_p<- p6[j]
max_s<- ns
max_e<- ne
max_seq<- seq6[j]
}
ns<-326-s7[j]
ne<-ns-49
if(ns>=ms && ne<=me){
max_p<- p7[j]
max_s<- ns
max_e<- ne
max_seq<- seq7[j]
}
ns<-326-s8[j]
ne<-ns-49
if(ns>=ms && ne<=me){
max_p<- p8[j]
max_s<- ns
max_e<- ne
max_seq<- seq8[j]
}
ns<-326-s9[j]
ne<-ns-49
if(ns>=ms && ne<=me){
max_p<- p9[j]
max_s<- ns
max_e<- ne
max_seq<- seq9[j]
}
```

```
if(max_p>0){
cat(file=f, " ")
cat(file=f, max_p)
cat(file=f, " ")
cat(file=f, max_s)
cat(file=f, " ")
cat(file=f, max_e)
cat(file=f, " ")
cat(file=f, max_seq)
}
}
}
cat(file=f, sep="\n")
}
close(f)
```

B-8. R script: Scans promoters predicted by TSS-PREDICT for matches with m3

```
n_m3 <- 485
n_t <- 1508
n_g <- 895
m3_table <-
read.table("m3_table.txt",header=TRUE,colClasses=c("character","integer
","integer","integer","integer"))
attach(m3_table)
t_table <-
read.table("t_table.txt",header=TRUE,colClasses=c("character","integer"
,"character","integer","character","integer","integer"))
attach(t_table)
g_table <-
read.table("g_table.txt",header=TRUE,colClasses=c("character","characte
r","integer","integer","integer"))
attach(g_table)
f= file("match_tssp_out.txt","w")
for (i in 1:n_m3){
cat(file=f, m_ct[i])
cat(file=f, " ")
cat(file=f, m_pos[i])
cat(file=f, " ")
cat(file=f, m_start[i])
cat(file=f, " ")
cat(file=f, m_seqloc[i])
cat(file=f, " ")
cat(file=f, m_h35loc[i])
for (k in 1:n_g){
if (m_ct[i] == g_ct[k]){
cat(file=f, " ")
cat(file=f, g_name[k])
cat(file=f, " ")
cat(file=f, g_hour[k])
cat(file=f, " ")
cat(file=f, g_strand[k])
cat(file=f, " ")
cat(file=f, g_operon[k])
}
}
found <- 0
for (j in 1:n_t){
if (m_ct[i] == t_ct[j]){
if (t_strand[j] == 1 && ((m_seqloc[i]+31) <= t_tssloc[j]) &&
(t_tssloc[j] <= (m_seqloc[i]+31+14))){
found <- 1
f_35hex <- t_35hex[j]
f_10hex <- t_10hex[j]
f_spacer <- t_spacer[j]
f_disclen <- t_disclen[j]
f_tssloc <- t_tssloc[j]
}
if (t_strand[j] == -1 && ((m_seqloc[i]-31-14) <= t_tssloc[j]) &&
(t_tssloc[j] <= (m_seqloc[i]-31))){
found <- 1
f_35hex <- t_35hex[j]
f_10hex <- t_10hex[j]
f_spacer <- t_spacer[j]
f_disclen <- t_disclen[j]
f_tssloc <- t_tssloc[j]
}
}
}
if(found>0){
cat(file=f, " ")
cat(file=f, f_35hex)
cat(file=f, " ")
cat(file=f, f_10hex)
cat(file=f, " ")
cat(file=f, f_spacer)
cat(file=f, " ")
cat(file=f, f_disclen)
cat(file=f, " ")
cat(file=f, f_tssloc)
}
cat(file=f, sep="\n")
}
close(f)
```

Appendix C: MMCTPP1 genome-wide promoter predictions online at http://www.biomedcentral.com/1471-2105/10/271

seq32_id	hpi	s	DHMM	h35_loc	seq32 M2 Predicted Sequence	P M2
CT001_113	16	5	-3.9	1702	TCTTTTGCGACAGCAAAACGCATCTTTAATAG	1
CT001_055	16	1	1.1	1648	TTGCAAAAAAGATTAAAAGTCAGAGTTAAGTA	1
CT002_105	8	3	-4.5	1691	TTTTGCTGTCGCAAAAGACGCTTCAGTAAGAG	1
CT003_053	8	4	-3.5	2058	TCGTTGGGAGGATTAGTCAAAGTCCCTACAGT	1
CT006_080	3	5	-1.9	7014	TTTTTTGTCAGAATATCTCTCGCATATAAATT	1
CT007_112	16	4	6.5	7142	AGTTTGCTAAAAATTTTATTAAGCAGTATGAT	1
CT009_098	3	4	3.4	9278	ACTTTGATACTAAAAAAGAGGAAATCTAAGAG	1
CT015_067	8	5	-2.4	17490	TTTCTGGATATGATACACAATCAAAGTACTAT	1
CT016_067	8	4	1.5	17572	TGTTTGTCAAAAATGTACCCCTTAACTACAAT	1
CT022_102	3	4	-3.2	27393	AGGGTGCATTTTTTCTTGCTTTTTCATAAAAT	1
CT022_089	3	3	-1.1	27405	TCTTGCTTTTTCATAAAATGTTCGGGTATGCT	1
CT023_066	3	4	-5.2	27943	TGATTGCCAATGTACACTCTGGTCTTTGTTAG	1
CT025_098	3	4	-0.8	29843	GTGTTGAAAGATTATGCGCAATTGGATAGGTT	1
CT025_060	3	3	-3	29880	CCTTGAAAATCAAGCTAATGATGCTGTATCCT	1
CT029_132	8	1	-1.9	33022	TTGAGAGGAAAAACTGGTAAGGCTGCTAAAGT	1
CT031_108	8	4	-1.4	34303	AAATTGCTGCCGCTAGCGAATTTGATTATGTT	1
CT032_205	8	4	-2.8	34508	AAATTGCTAGAGGAGATGTTCGTTCTTCTAAT	1
CT035_040	1	3	-2.8	40316	TTTTGGTATAATGAGAAAAAGCTTTTTGTAAG	1
CT038_320	8	4	-2.1	43532	CCTTTGCTTTAGCTCGCAGCGTGGAATATTTT	1
CT039_124	8	1	-6.5	43927	AGGCATTAAAAAGTTGCGTCTTTTCATACAAT	0.79
CT040_045	16	1	2.2	44040	TTGATTTGGATTAGCGAATAAATAACTACTAT	1
CT043_324	8	4	2	48907	GTATTGCGGAAAATAACCAAAAAAAATATCCT	1
CT046_111	16	5	-1.6	51397	TGTGTGGTTAGTTTTTAATAAAAAGTTAAAAA	1
CT049_065	8	4	4.7	54051	TGTTTGTTAATTTAATTTTTTCTAATTAAAAG	1
CT053_255	8	4	2.4	60798	ACTTTGCTTTTTTTTTAAGTATCGAATACCCT	1
CT054_078	8	3	-0.9	60802	GTTTGCTTTTTTTGAAAAATAAAAATTTTGCT	1
CT054_076	8	1	2.4	60802	TTGCTTTTTTTGAAAAATAAAAATTTTGCTAT	1
CT054_054	8	4	0.7	60827	ATTTTGCTATGGGAATTTTCTAAAAGTATCAC	1
CT061_043	8	1	1.8	70947	TTGATTTAGCCTTATTTTTTAGTTTGTAAAAG	1
CT062_064	8	3	4	71848	CCTTGCTATAAAAAGAACAGGATAGATAAGAT	1
CT065_060	1	3	-2.8	78299	CGTTGATTTGATCAACAAAGAAAACTTAACAA	1
CT066_169	8	3	0.4	79235	GATTGCGAAAAAAGCAAAAACCACTATAGAAT	1
CT068_151	8	4	-0.6	80551	GCTTTGAGAAAGATTGTTTCTTGCTCTAAGAG	1
CT072_322	16	4	0.9	84823	TTCTTGGCATCAAATAGTTCCTAAATTACAAG	1
CT072_118	16	4	-1	85027	TCTTTGTTTTGTTGAGCGTCTTATAGTATCTT	1
CT072_068	16	1	-0.8	85074	TTGTATAGCAGCTGTTTTAAGTAGAGTATAGT	1
CT074_100	3	1	-0.2	89280	TTGTTCAATTAGGTATCTCAGATTCTTACAAT	1
CT076_305	16	4	-0.4	90258	TCATTGCGGGAGATAACGAATTTCATTATTTT	1
CT078_081	8	4	-1.5	92854	CTTTTGTAAACTGATCAAGAGAGGTCTAGACT	1
CT079_150	16	4	1.6	93484	ATATTGCTGTAAATCATTCTATCTTTTATAGT	1
CT080_071	24	4	0.5	93546	GGTTTATGAAAAACAATTTTTTAATTTAAAAT	1
CT082_123	16	4	-2.8	94150	GCCTTGCGTTTTTTTTGTCCAAATGTTTTTAT	1
CT084_154	8	4	2.7	97681	TTTTTGTTTAAAAACAGATTGGAAAATAGATT	1
CT084_120	8	1	1.1	97650	TTGTTTTGTTTTTAATTAAAAGAAAATAAATA	1

CT085_154	8	1	-1.2	99593	TTGCTAGAGCTAAAAGCAAGGGTTTTTAGTCT	1
CT090_040	16	4	-0.4	105494	ACCTTGAGAATAAAAACATTAACCAATTCGAT	1
CT091_071	8	4	1.3	106607	TTCTTGAGAAAAACATTTATATACGGTAACTT	1
CT091_043	8	3	-6.7	106580	ACTTGCGAAGTATTCCTTATAGCGCTTAAGCA	1
CT098_072	3	4	3.7	115743	ATCTTGCCTTTTTTAAGGTGAATATTTACACT	1
CT102_076	8	4	-0.6	117863	CTCTTGATTAATAAGCTTTGGCTTCGTAGGAT	1
CT102_062	8	4	-3.7	117877	GCTTTGGCTTCGTAGGATGAGGGACATATCTT	1
CT105_215	8	1	-0.4	120237	TTGAGAGAGGAGAGTAGCAGATTCTTTATTAT	1
CT110_144	1	1	0.4	128114	TTGGTAACATCGTTTTAATTGATAAATATTCT	1
CT111_132	1	4	-1	128445	CAGTTGCAAAAAAGCGAGGACTTTGCTATCGT	1
CT114_266	3	5	0	132984	ACGTTTGCTGGAAAGAAAAATTTCGTTAACTT	1
CT114_163	3	4	1.5	133086	CGTTTGATACAGCAAAAAGTAGTGACTATGTT	1
CT114_079	3	5	0.6	133171	ACTATGGCAAAAGATGTAGTTGTGTTTAAAAT	1
CT115_132	1	4	5.1	134807	AGTTTGTTTTAAATAGTTTTTTTAGTTAAAAT	1
CT115_072	1	4	-1.1	134867	TTTTTGCGTCCGAAACATTGTTTTATTAAGTG	1
CT123_041	3	5	-3	139363	GTACTTGTGAGTATATTCAACGCGTCTAAATT	1
CT125_060	3	4	5.3	141437	TGTTTGGAAAAAATAATCATCAAAATTATAAT	1
CT133_104	1	1	2.8	151320	TTGCCAAAGATTTTTGTTAAGAACTTTACATG	1
CT134_137	3	1	-1.2	151534	TTGGTTTTCATTCCAATTAAAGAGGATATGCT	1
CT139_208	3	3	0.6	157306	ATTTGCAATACAAATAATGTCTCGTTTAACAT	1
CT140_135	8	4	1.6	157145	TTGTTGCCGGTTCTATTCTAGAAACGTATAAT	1
CT141_061	8	4	2.4	158089	GGGTTGCAAAGAAGGTTCTTTGTAATTAATTT	1
CT144_043	8	5	-4	160655	TGATTTGGTTTCCTTTTGGTTCTTCTTATAAG	1
CT145_109	8	4	-1.3	161631	CTTTTGTTGTAGACGGATCGGAAAGCTACAAG	1
CT146_119	8	4	0.8	163476	ATGTTGGAAAAGTCTAAAAGATGATCTACGTT	1
CT149_121	8	4	0.6	172881	GAGTTGTTTTATCCAGTAATTTACCTTATTAT	1
CT150_071	3	4	2.8	173094	ATCTTGATTATTTTTGAAAATAGGTATAGCAT	1
CT151_265	8	4	2.8	173094	ATCTTGATTATTTTTGAAAATAGGTATAGCAT	1
CT156_068	16	5	-2.2	180666	TATATTGAAAAAGCGAACAACAAAACTAAAAC	1
CT159_219	16	1	-1.1	183886	TTGTGGAACAACAGTGGATCAGAGAATATGAT	1
CT161_090	16	5	-5.3	185139	CGTTTTGTTAAAACGCGAAACGCGGCTAAGGT	1
CT162_137	8	5	-0.5	185573	TATTTTGTGTTTGTAACTTAAAGAGTTACATT	1
CT164_084	8	1	-4.9	187288	TTGAAGGGCAGTATTGTCGCATTCGTTATTAG	1
CT164_047	8	4	0.4	187328	CACTTGAGAGGTTTTCTCTATTTCGATACGAT	1
CT165_073	16	3	-2.5	188035	ATTTGTTTTTAACAGGTAATAACAAATACCCG	1
CT170_106	8	1	1.9	192806	TTGATTAAGAGATGTTCTTATAGAAGTAAGAG	1
CT172.1_294	8	1	1.2	196074	TTGCGACTTTCGAAATAGAACGAACATATAAG	1
CT173_145	8	1	2.4	196265	TTGCAGATGTTAAGAAACAAAAAGAATACCAG	1
CT175_098	3	4	1.2	196762	TTTTGGCTTTTTGCTATGGTTTTTTGTACAAT	1
CT181_116	16	1	3.4	203437	TTGTAAATATCTAGAAGTGTATGATTTATGAT	1
CT182_138	16	4	-1.3	203554	ATCTTGATTGTTAACGATCTCCTTGTTAGGAT	1
CT182_118	16	3	-5.1	203573	CCTTGTTAGGATGGTCGGGTTCTGTGTACTAG	1
CT186_043	8	4	-6.1	208106	AGCTTGGCGGCCCTATTAATTTGTTTTGCAAG	0.91
CT188_133	8	5	-2.9	210539	CGTCTTGTCAATCTTCGAGAAGGAGATACGCT	1
CT189_285	3	5	0.3	213203	CACCTTGAAAGAGGTAATAGACTACTTAAAAG	1
CT191_091	8	3	-1.8	215793	CTTTGACTTTTTTAGTTCGCAACAAGTATAAG	1
CT191_089	8	1	-0.6	215793	TTGACTTTTTTAGTTCGCAACAAGTATAAGAG	1
CT195_186	8	4	3.4	221021	TTTTTGCAAAAGAAATGAAATAGCATTATAAC	1

CT195_047	8	1	-4.5	220885	TGGCTTCAACTTTGTAAAAGTAAAATTTTTAG	1
CT196_093	3	4	0.3	220972	CGCTGGTTTAATAAAACCCAACTAGTTATAAT	1
CT196_049	3	4	-0.7	221016	CTTTTGCAAAAACTCTCTACTTTAAATTAATT	1
CT196_048	3	3	-0.6	221016	TTTTGCAAAAACTCTCTACTTTAAATTAATTT	1
CT197_268	8	1	0	221190	TTGCATATTATTAGGAGCTCTAATCTTAACAG	1
CT197_088	8	4	-3.4	221373	CTCTTGCGATTAACGCCTTGCTTGATTAACAA	1
CT197_074	8	4	3.4	221387	GCCTTGCTTGATTAACAATCTCATGATACGAT	1
CT199_145	3	4	-3.6	224011	TCCGTGCGTAGAAAAGATTCTCTCTTTAAAAT	1
CT199_049	3	3	0.4	224106	ACTTGTTATCTGTGATTAGATCGCAATACAAT	1
CT200_104	3	4	-7.3	225025	TTGTTGCGACCTTCTGCAAGCGTGGATAGATC	1
CT203_221	8	4	0	227644	CTTTTGCGTGTCAACTTAATGTTGTTTAGATT	1
CT203_077	8	4	-2.4	227788	TCGTTGACAGAAGATTTTTGGCATGCTACGAC	1
CT204_043	8	4	2	228607	CTTTTGAGTCATAGTTTTTATCCAGATAAAAT	1
CT209_107	8	3	-3.3	238330	ACTTGATTTCGATCTCAGACATAGTGTATGTT	1
CT213_083	3	4	-0.7	241515	GCTTTGAAGGAATCTTGAAGAGGTTGTAGGAT	1
CT213_063	3	4	-1	241495	AGGTTGTAGGATTACTGTTTGAGGAATAGAAG	1
CT214_130	16	4	-0.5	243300	TCTTTGAGAAGGTGAGGTAAAGAAAATGAAAG	1
CT215_114	8	4	-0.7	243366	TGTTTGCATGCTAAGAAAGATTTATAGACAAT	1
CT217_099	8	4	-0.2	245995	ACCTGGATTTCTAATTCGCAAATCTTTAAAAT	1
CT218_090	8	4	0.5	246839	TACTTGGTTTATTTTTCTTATTATTTTAAAAA	1
CT221_242	8	4	1.7	250504	GCGTTGTGAACAAAAAACAGTGTATTTAAAGT	1
CT221.1_171	8	1	-1.3	250865	TTGTTTGAGTTTAGGGCTTTGTAAACTATTTT	0.96
CT221.1_140	8	1	-5.1	250834	TTGCGTATGAAGCGAGTGTCTTTGGATAAAGC	1
CT223_063	8	1	4.4	252127	TTGATTTCTTTAAAAAAATGTTTCTGTACAAT	1
CT226_075	3	1	0	254061	TTGGTTGGGTAGATTAGGTATTTAACTACCAT	1
CT226_047	3	4	2.3	254030	CCATTGAAATATATAAAAATTTTTACTTTAAT	1
CT228_063	1	4	-1.2	255424	GATTTGCGAATAAAGCGCCGTTCTGATACAGT	1
CT229_045	1	5	-0.1	256202	TACTTTGCCAAGTTTTGTTTAGGCATTAAGTT	1
CT230_076	3	4	-2.4	256314	TTCTTGCAAAAAAAATCTCTCCTTTCTTACAT	1
CT230_075	3	3	-0.2	256314	TCTTGCAAAAAAAATCTCTCCTTTCTTACATT	1
CT232_095	3	4	2.1	259303	GAGTTGCTTGTAAGTCTTTTGCATGATATACT	1
CT235_088	8	5	1.9	263144	AGCTTTGCTAAGAAACAAAAAACCTCTATATT	1
CT236_072	8	4	-3.7	264004	AAATGGCCATAAGTCGACCGTCTGCTTAAGAT	1
CT239_309	8	4	-0.7	267279	AAGTTGCATGCGAAGTAGAAAACTCTTGTAAT	1
CT243_056	16	4	3	271031	TTCTTGATGATTCTTTTCAAAATAATTAACAT	1
CT246_097	3	4	-5.9	274484	GATTTGCGTCAAACAAGCAAAACAGCTGTCCT	1
CT248_081	8	5	0.1	279406	TCTGTTGAGAATAGAAATCTTTCTTTTAATAT	1
CT249_128	3	4	5.1	279360	TGCTTGTGAATAAGAAATATTAATATTAAAAG	1
CT249_060	3	4	5.4	279428	GTCTTGATATTCGGTAAAAAATCAAGTAAAAT	1
CT252_082	8	1	-1.9	284909	TTGCAAGATTGCTAGCAAAAAATTTTTGGCAT	1
CT252_077	8	4	0.5	284901	AGATTGCTAGCAAAAAATTTTTGGCATAGTCT	1
CT253_129	8	4	0.1	284952	TAGTTGCTTTTTGAAAATACTCATGCTAGAGT	1
CT253_060	8	4	-2.5	285021	AGCTTACAAGAGTGTTGCTAGGGACGTAAAAT	1
CT253_046	8	1	-3.2	285032	TTGCTAGGGACGTAAAATCGAATCAATTTTTT	1
CT257_077	16	4	-3.1	288439	CGCTTGGGGACAATTTGTATTCCAAATACTAG	1
CT259_102	8	4	-0.6	291706	AGTTTGCTTTCTTTTTTAAAAAAATCTTTGCT	1
CT260_041	8	1	-5.9	291889	TTGCCTCCCTTTTAATAAGCCGTACTTAGAGG	1
CT261_131	3	4	-2.6	292302	AAGTTGATTTGTGATAGCTTCTTCCATGCAAT	1

CT261_126	3	4	-2.7	292307	GATTTGTGATAGCTTCTTCCATGCAATAGGAT	1
CT261_062	3	1	-2	292368	TTGAAAAAGCAAGCAAAACTAGAAAATAACCA	1
CT262_059	8	1	0.3	293066	TTGCTACAACCAGAAAATAAAGAAATTAAAGC	1
CT265_140	3	5	3.1	297515	AGGCTTGATTTCTTTTTAAGAGAAATTAAAAT	1
CT265_111	3	3	-0.3	297488	AATTGTTTGCGTGAAACAAAGGTCATTATAAT	1
CT265_064	3	4	4.4	297440	TGGTTGTTTTTGATTATTGTTTGTATTAAAAT	1
CT265_048	3	1	3	297427	TTGTTTGTATTAAAATAACTCTTTTTTATAAG	1
CT267_097	8	5	-1.7	299206	AGTCTTGAATCCAAAGGATGAATGCATATTAT	1
CT269_082	16	5	0.3	301556	AAGCTTGACAACGAATATGTGTATAGTAAACT	1
CT269_051	16	4	-6.7	301526	TATTTGAGAAAGCTTTTTGAAGCCCTTGTGTG	1
CT270_145	16	1	-3.2	303889	TTGCGTCAGATAGAACAAGAAAATGTTGCGCT	0.99
CT270_075	16	3	-4.1	303817	TGTTGCAGATAGCTTTTCTTCCAGAATACCAG	1
CT273_175	8	4	0.9	305026	AGGTTGCAGTGTATCGCATGTTTAGTTAAAAG	1
CT274_139	8	3	-2.5	305634	GCTTGAAGTAGAAGAGGAAGTCTTAGTACGAG	1
CT275_170	16	4	0.7	306266	TTTTTGTTTCTGATTTCAGGAAAAATTAATAG	1
CT275_093	16	1	-1.5	306340	TTGCTTATCAATAGCAATGCTACATTTGCGAT	1
CT275_070	16	3	-7.1	306365	ATTTGCGATTTTCTCCAGGAGGAGGGTGCATT	1
CT286_067	8	4	1.2	317906	AAGTTGCATCATTATCATAAATGTCGTATATG	1
CT287_136	8	5	1.1	321723	TGATTTGTCTTTTTGTAATTAAGATTTAAAAG	1
CT287_070	8	5	-4.1	321657	TTTTTTGCTAACATGATGGCCCTTTGTAATGT	1
CT288_062	1	4	5.9	321764	TTGTTGTAAAAAAACAATATTTATTCTAAAAT	1
CT293_094	8	1	1.1	327450	TTGTAGAAATGAGCAAGCAATAATTTTATTAT	1
CT293_065	8	3	4.1	327419	TATTGATTGGTTAAAAAAAATTACAATAAAAT	1
CT294_107	8	5	-0.3	328151	GCTCTTGACAGTGAATATTTTATTGATAATCT	1
CT302_219	8	4	0.7	340082	TCTTTGACAAAAGATGGTAAAGAAATTTTAAG	1
CT303_113	8	4	-4.4	340618	CAGTGGTATCGTAATGGGTATAGCTGTAGGAT	1
CT311_147	8	4	-1.5	348114	TATTTGCTTTAATTTATCTTGAGCGCTGAGAT	1
CT313_063	8	4	-1.3	350435	CCCTTGAATAGTATCGTTTTTTTTGGTAGGCT	1
CT317_052	3	3	-3	359813	CGTGGATACTAGGGAGTTGATTGCGTTATAAT	1
CT321_208	8	5	-1	362041	AGATTTGCGATTCGTGAAGGTGGTCGTACAAT	1
CT322_298	3	3	-2.1	363460	GCTTGATAATAATCCGCGTCTGAAGTTACTAT	1
CT323_149	3	1	1.6	363882	TTGTTTGACATTTTCTGTTTAGTCGATATAAT	1
CT324_264	3	4	2.3	363884	AACTTGTCAAAAAACAGAAGGAAAAGTATCTT	1
CT326.1_086	8	3	-6	367501	CATGGCAATGAGGGGATTGGGTCGTTTAGAAG	1
CT326.1_054	8	4	-0.5	367468	CATTTGTGGTTGTTAAATAACAATTTTAAAGG	1
CT326.2_063	8	4	3.6	367849	CTCTTGAGTTTATTTTCTAAGAAGGATAAAAT	1
CT327_169	8	4	-1.2	369301	ATTTTGTAAGATAGATCAAAGCGTAATACTCG	1
CT327_136	8	5	2	369267	TTTCTTGCAAGGAAGGCTTATTTTTATATGAT	1
CT327_096	8	3	-0.4	369229	TATTGCTTTGATATAAATCTCTTGGATATGCT	1
CT327_091	8	3	-2.3	369224	CTTTGATATAAATCTCTTGGATATGCTAATCT	1
CT328_073	8	4	-1.5	369262	TCCTTGCAAGAAATCGAGTATTACGCTTTGAT	1
CT337_096	3	5	-0.8	385104	AGCTTTGAAAAGAAGCTCTAAGGGTTTATTAT	1
CT338_054	16	4	-7.3	385661	CATTTGGACGTTCGGATAACGCGAAATAGTTT	1
CT340_048	8	4	-2.6	389577	ACTTTGCTGTATAGAAAGAAGGATCTTTTTAG	1
CT340_045	8	1	-1.6	389577	TTGCTGTATAGAAAGAAGGATCTTTTTAGCTG	1
CT341_307	3	3	0	391050	TCTTGAAGCCTAAATAAAAGTGGTGTTACAAT	1
CT342_162	3	4	3.8	391109	GTTTTGTTTACTTGTTTGGTAATTTTTATAAT	1
CT342_102	3	3	0	391050	TCTTGAAGCCTAAATAAAAGTGGTGTTACAAT	1

CT343_082	16	1	-2.5	391861	TGGCGACAGCGAATTAATTTTTGAATTAAAAC	1
CT343_064	16	3	0.9	391841	TTTTGAATTAAAACGGTTTTAACGGTTATAAT	1
CT344_104	16	5	-2.4	391989	ATTGTTGCGATTTTGTAGTATATCGGTAAGAA	1
CT344_094	16	3	-1.4	391997	TTTTGTAGTATATCGGTAAGAAGAAGTAATAT	1
CT344_045	16	4	-5.2	392047	CCCTTGCTCTTCCAAGGATTTTGAACTAGTTG	1
CT346_210	8	4	1.4	395158	TTCTTGTGTTTAGGCTTAGTGGAAGTTATAAT	0.99
CT347_114	16	4	-1	396230	TATTTGCTGCAGAGGAATTCTGTAGCTACGAG	1
CT355_224	8	4	0.3	406825	GGCTTGAGGATATAACGCTTTTTTGTTAAAAG	1
CT355_103	8	4	-1	406704	TGTTTGCCACATCTCTAGGGAGGCGGTAAGAT	1
CT356_116	24	4	-1.6	409136	ATTTTGGGTTGTTTATAACCATTTTTTATTAG	1
CT357R_047	3	4	-5.4	409672	CATAGGCAAGACTTCCAGGGGAAGCAAGCAAG	1
CT357_195	3	1	-2.9	409285	CCCTCAACGGGGACCGGGGGTTCAAATCCCTT	1
CT359_074	8	4	-2.5	411095	AAGTTGAAGGATGTCTGTAGTTGGTTTAAAGG	1
CT367_306	8	4	2	419032	GCTTTGCAAAAAATCCATCGCGCTTGTATAAT	1
CT368_140	8	5	-1.9	419746	TACTTTGTCTCAACGATTAGACAAACTACGAT	1
CT372_116	8	3	-4	424345	TCTGGCAAAAAAAATCTTTTTTCCACTACACG	1
CT373_048	8	4	-3.4	425811	CGCTTGCATCCTTGTGGTTAACCATTTACCAA	1
CT374_079	8	4	-0.1	426383	TTTTTGCTTAACAGCTCTCGGTTTCTTAAATT	1
CT376_280	1	1	-4.2	430442	TTGGCTGCTAAAAAGGGACAACAAGTTATGCG	1
CT376_107	1	4	-0.5	430266	TACTTGATTCTTTTATCATCCAAACGTATGTT	1
CT376_060	1	4	-3.7	430219	TAGTTGCAAACGTAGTGTTGAGAGTATGTGCG	1
CT376_043	1	1	-3	430205	TTGAGAGTATGTGCGTTTTGTAAAAGTAAAGA	1
CT377_075	16	1	5.3	430568	TTGCAGAGTTTTTATTTTAAATATGTTATAAT	1
CT378_080	3	4	-3.6	432262	GAATCGCGAAAGATCACGAAAGATAGTACAAG	1
CT382.1_088	8	4	1.6	436574	AAATTGTTTTCATTTGAATTTAATTTTATTTT	1
CT383_075	3	3	1.5	436599	CATTGAAGACAAAGAAAAACTTTTGTTAAAAT	1
CT385_045	16	1	-1.6	439428	TTGTGAAGAAGCACTGAATTTAGTGTTAAAAC	1
CT390_081	8	4	-2.3	444410	CGCTGGACAGATGAGAGTCTCATCTTTATATT	1
CT392_071	16	3	-0.1	447895	ACTTGATGTTTCTTTTGTTTGTTTCTTAAAAT	1
CT392_060	16	4	1.3	447883	CTTTTGTTTGTTTCTTAAAATTAATTTAAGCG	1
CT393_183	8	3	0.1	447846	ATTTGCAAAAACGCTTAAATTAATTTTAAGAA	1
CT393_071	8	4	-2.1	447959	AGATTGATCTAGAAACACTCCTATGCTAAGAT	1
CT394_043	16	4	-1.5	449801	TTCTTGACCAGTGGAGACGGTTTTCTTATAAT	1
CT396_145	3	4	0.6	451472	ATCTTGGAGGAGTTTACTAAAGGTTATAAGAT	1
CT399_101	8	4	-3.6	458500	GGATGGAATTTACGGTGGAGAGTTTTTAACAG	1
CT399_070	8	4	1.4	458469	GCTTTGAAAAATCGCTTAGAGTCTGTTACGAT	1
CT400_069	16	1	-0.4	459669	TTGATAATCTTCTTCTCATATTGAGTTAACAG	1
CT408_094	8	4	-3.2	465736	CGTTGGCAAGATTAATGGCAATCCCTTACGCT	1
CT410_115	16	5	-1.6	467897	GCCCTTGTAAACGTAATTTTTTCTATTATAGG	1
CT411_100	16	1	-3.5	469246	TGGAGAGTGGCTTAAACATTATGAATTAGCAT	1
CT412_077	16	4	-5.1	471202	GGATCGCTAGAAAAGCGCTTTCTGTTTATAGT	1
CT413_073	16	4	-1.9	474272	ACCTTGCCTAATTTACTTTTCTGATTTATCTA	1
CT416_137	8	4	-5.6	486332	ATTTTGCTCAAGCATGCGGGGAAACGTAGTAG	1
CT418_142	8	1	1.2	489113	TTGTAATGAAAAAAACAGATCGTACTTACGTT	1
CT421.1_121	8	4	-2	490984	CTATTGAAAAAAAAACGGCGCTCTACTATTCT	1
CT429_061	8	4	-4	499561	AAATTGAGGATAACACAAGAGAGAGTTGCTAT	1
CT435_061	8	1	-1.4	505227	TTGTCTTTGCCTGCTGGAGTAGATATTAAGAT	1
CT437_157	3	4	-0.1	507746	GTCTTGCTAATGAGTTGATCGATTGCTTCAAT	1

CT439m_069	3	3	1.8	508595	CCTTGCAAACAAAGATATTCTTATTCTATATT	1
CT440_198	8	1	0	508590	TTGCAAGGGGTTATTTCTAGGTCTAGTAAGAG	1
CT442_064	16	5	-0.9	511852	GTTTTTGAAAAAAACAAGTGTTTGTGTAGACT	1
CT444_177	16	4	-1.9	514219	GCTTTGATTTGCTAATTACCTGTTATTAGACG	1
CT444_130	16	4	3.2	514172	CAATTGATATAATTTTTATTTTATAATGTAAT	1
CT444_062	16	4	0.2	514104	GAATTGCTTTTATCGATAAAAGAAACTTCAAG	1
CT444.1_115	8	1	1.1	514267	TTGCCTACAAACAAAACAAACTTCGATAGAAT	1
CT449_188	3	3	-1.5	524363	GGTTGCGGTTTAGCAGCTTAGTTTGGTAAAAT	1
CT450_185	8	3	0.7	524457	ACTTGATAATAATCATTATCTATGGGTACCAT	1
CT456_104	8	4	5.6	530807	TGTTTGTTTTTAAAAACAAATAAAAATAAACT	1
CT458_115	16	4	-4.4	535493	CGCTGGTGAAAGATGTAAGGACTGGATATGAA	1
CT459_051	16	1	2.2	536536	TTGAAAAAGAAGAGTTTAAAGTTGGATATTTT	1
CT460_090	8	3	-2.2	536717	TTTTGACGATAAACCTAGTTAAGGCATAAAAG	1
CT461_052	8	4	-3.1	537074	AAATTGACTCACGTGTTCCTCGTCTTTAAGAT	1
CT465_248	8	1	1.8	540044	TTGCATCGCTAAAGAATAATATTGGCTAAAAG	1
CT471_217	8	4	-3.3	546072	TTTTTGCGGAAGTGGCGGAATTGGTATACGCG	1
CT471_092	8	3	-1	545948	TTTTGATAATCTTTTTATCTTTCTAGTATGCT	1
CT471_047	8	4	-6.9	545902	TGCTTGCATAGTTCGTTGTGCATGCGTAGAGC	1
CT475_058	8	4	0.6	548337	TACTTGCTACTATACACGCCACTTCGTAAAAT	1
CT480_091	1	1	-3.1	557633	TTACTAGAATTCAAAACCCAGATTGATAAGAG	1
CT480.1_149	3	4	1.7	558103	TCCTTGCCTATATTTTCTGATTTTCTTAAAGT	1
CT482_223	8	4	-1.1	559916	TCTTTGTAAACAGAGAAGATCCTTCTTAAAAA	1
CT485_083	16	1	-2.9	562690	TGGCAAAAATTCGTCGATCAGGAAGATACAAA	1
CT487_076	8	4	-1.1	564105	GATTTGGGAATATCCGAGGAATTGAGTACACT	1
CT487_055	8	1	-6.8	564087	TTGAGTACACTCTCGTGGCTGCCGATTACGTT	1
CT488_066	8	4	-2.4	564826	CTGTTGACAAAGGTAAAAGTGGTTGGCAAAAT	1
CT488_047	8	4	-3.4	564807	TGGTTGGCAAAATAAAGAGCTGATTCTTCTAT	1
CT489_056	3	3	2	566231	TCTGGCTTTTTTAATAATTTATTTTTTAAAAT	0.99
CT490_121	8	4	3.1	566252	TTATTGCATGAATAATAAATCGATTTTATTTT	1
CT494_048	8	4	-1.3	572281	GTTTTGTTGCTAATGCTAAAAGCGATTAATAG	1
CT494_045	8	4	3.1	572278	TTGTTGCTAATGCTAAAAGCGATTAATAGATT	1
CT496_153	3	1	0.9	574868	TTGTTTGTTTGAATGTTTTTTGTTGATAAGCT	1
CT496_088	3	4	-2.4	574800	TGCTTGCTTTTGGAGTGTCTATGTTTCATAAT	1
CT496.1_072	0	1	-2.6	575332	TTGACACTGGTCAACATACACCCCAGTACAAT	1
CT503_280	8	3	3.4	582180	CATTGCTTTTTTTTAGAACTTTAACATACAAT	1
CT503_160	8	4	-2	582059	TGGTGGAATGGTAGACACTAGGGACTTAAAAT	1
CT503_046	8	1	-3.1	581948	TTGATTTATAGATGAGGAGAGCTATTTAACTC	1
CT505_091	8	5	-2.3	584157	AGGGTTGGTGATAATGCTCAAAAGTGTATTAT	1
CT507_180	3	1	-0.3	585861	TGGATTAAAAGAAGTTGAAGTAGGCTTAAAAG	1
CT508_186	3	4	-7.8	586283	GATTTGCGCCGTCGTGTGCAATCTGATATCAA	1
CT509_062	3	4	3.6	586549	ACCTTGAAAAATAACAATTTTTGACCTAAGAT	1
CT512_289	3	4	0.7	589152	CCGTTGTAAAGACAAATAAGCATGTTTATGTG	1
CT512_100	3	1	-1.7	588966	TTGTTTTCGATCGAGGAGCTCATAAGTATCAT	1
CT514_048	3	1	2.4	589873	TTGCTTTGTTTGGTTTGGTAGCAAATTAAAAG	1
CT518_198	3	4	-4.2	591727	CTGTTGTTGTTCGAGTCGAAAGGGTATACTCG	1
CT519_241	3	4	-0.6	592038	GATTTGTTAAGCGTGTGGAAAGGGTATAGTAT	1
CT519_115	3	1	-2.7	591915	TTGCGTGCGGAAGCTGCTTTACAGAATAAAGT	1
CT524_065	3	1	-2.9	593831	TTGAAAACTCGTGATAAGCGTAAGAGTAATAA	1

CT525_227	3	5	-5.5	594849	ATTGTTGCTGGGGACGCCACGAAGCCTATGAT	1
CT525_199	3	4	-0.4	594822	TGATTGCTGAAGCCATAGAAGCAATTTATTCT	1
CT526_131	3	1	-3.7	595116	TTACTTACGGAGAGAATATCAGCGGATACGAT	1
CT526_057	3	4	-2.9	595039	GAATTGCTTGTCGAGAGTCTTGTCTCTACAAC	1
CT527_104	3	4	-3.8	595770	GACTTGGATAGAAAGGTAATGCTCGTTAAGGG	1
CT528_191	3	4	-2.6	596531	CTCTTGTGTGGATGTCGCAGGTCATATACCAT	1
CT533_060	8	5	-1.1	601008	TCGCTTGCTTGCTAAAAAAAAAAAAAGGATAAT	1
CT533_058	8	3	3.3	601008	GCTTGCTTGCTAAAAAAAAAAAAAGGATAATAT	1
CT533_055	8	4	6.7	601004	TGCTTGCTAAAAAAAAAAAAAGGATAATATACG	1
CT535_122	8	4	0.2	603346	TTCTTGATTTTCTTCTATATCGAGACTACTAT	1
CT544_160	8	1	-0.8	610998	TTGACGACATTACTACAATTGTCCAATATGAT	1
CT545_250	3	4	-3.8	612295	GGATTGGCAGCTGCAGAATTGTCTCATAAAAA	1
CT546_078	16	5	-1.7	617427	GCTTTTGTTTAGCAAAAACAAAAAAATTTATT	1
CT546_076	16	3	-0.8	617427	TTTTGTTTAGCAAAAACAAAAAAATTTATTTG	1
CT546_050	16	4	2.9	617400	TATTTGGCATTGCTGTTTTTATTTATTAAAAT	1
CT546_045	16	5	0.1	617394	GGCATTGCTGTTTTTATTTATTAAAATAAATA	1
CT546_041	16	1	3.7	617394	TTGCTGTTTTTATTTATTAAAATAAATAAAAA	1
CT547_097	8	4	1.2	617411	TTTTTGTTTTTGCTAAACAAAAGCTATAAGAG	1
CT547_065	8	3	0.1	617442	ATTTGACAAATTCTCTTTTTCTTTTTTATGAT	1
CT555_218	8	4	-0.7	624337	ATCTTGTTTTTGCACTTACGATACTCTATAAG	1
CT556_137	8	5	-1.4	628196	TACCTTGATTTTTTCCTCCTAGGAACTATAAT	1
CT557_165	8	4	-3.4	628855	TCATTGAGATTTTATCCACCCAGATGTACAAC	1
CT559_055	16	4	-6	631398	CGATTGGCACTAATCTCCCCATTTGCTATGGT	1
CT563_050	16	4	0.2	635112	ACCTTGCTACTAGAAGGGTTGATGATTAGCTT	1
CT564_079	24	1	-8.4	635371	TGGAGGCTGGCTGTGTAGCATGATTTTACGGT	1
CT564_047	24	1	-1.2	635403	TTGCTGCGCAGATTTTCCAAAATTTTTACAAA	1
CT565_061	8	4	2	636898	CTTTTGATAATTGTTTAGTACGAGATTAATTT	1
CT565_053	8	3	-0.5	636891	AATTGTTTAGTACGAGATTAATTTAATAAAAA	1
CT566_053	8	4	-2.8	637973	GCCTTGCCACCAAGAACCCTTTCCTATATTTT	1
CT569_136	8	1	1.6	639342	TTGCTAGGCGGTTTTATTGGATTGATTATGTT	1
CT573_061	8	4	0.4	645514	AAATTGATTTTTTTTCTCAAATCAGTTACTTT	1
CT575_124	16	4	0.9	648516	GCCTTGCAATTTTTTTTAACAGGAGGTACCGT	1
CT576_077	16	3	0.6	648615	ACTTGTTAAATCAGATCGTTAGAATTTAATAT	1
CT577_045	8	1	-2.7	649360	CTGCTTCTAACAAGAAAAAAGCAAAGTAAAAG	1
CT578_162	24	1	-4.4	649639	TTGACAAAAGGGGTGTTTGACAATCCTAAAGA	1
CT579_165	24	1	-4.9	651130	TGGCAACAAGCAAGTAAAATTGCAGCTAAACA	1
CT580_114	8	4	-3.1	653786	ACCTGGGTTTTTTTCAAGAGATGCCCTACAAT	1
CT581_179	8	1	-0.9	653801	TTGTTATGGATCTCTATACAATCTCCTAAAAG	1
CT581_044	8	4	-3	653939	TCTTTGCAGCGGTTTGCAGTAAATTTTTTATT	1
CT584_064	16	4	2.9	657806	AAATTGATTAAAAAGTTACAAAAAGTTACAAC	1
CT585_084	8	4	-4.9	658536	GTCTTGCATCGATAAATTAAGAATTAAAGAAA	1
CT585_081	8	1	-1.6	658536	TTGCATCGATAAATTAAGAATTAAAGAAAAAG	1
CT586_142	16	5	-6.3	659544	AGCCTTGCAACAAGGAACGGAAAAAATGCGTA	1
CT586_138	16	1	-1	659544	TTGCAACAAGGAACGGAAAAAATGCGTACCGT	1
CT590_198	8	1	-2.9	670213	TTGCCTTTAACAGAAAGCATTTCAGCTATGGG	1
CT595_197	8	4	-1.1	676921	TTCTTGAAAGGACTGAATGATGGGATTATTTG	1
CT596_066	3	4	0.5	676817	ATCTTGGTTCTATACAAGAAATTTGTTAGGAT	1
CT603_149	16	4	-3.1	682335	TTCTTGATCATTTAGCGAAAGCATGGTATCCT	1

CT603_069	16	1	-8.6	682258	TGGCTACGAACCAGGTGGTCAGAGGTTCAAAT	0.99
CT604_061	16	4	-2.7	684057	TGTTTGACAAGAATAAATCGCCTTTCTATATC	1
CT606.1_061	16	5	-0.5	686402	GGTGTTGCCACTTTTCAAAAATTAATTATCTG	1
CT608_127	3	3	-4.7	686975	TCTTGCCAAGACGGGGGGAGTTGCTCTATAGT	1
CT608_107	3	1	-2.3	686993	TTGCTCTATAGTAAAAGCCTCGTGCATACACT	1
CT611_117	3	4	-1.9	691963	GTTTTGTCACACATATCAAAAAGGAATATTGG	1
CT613_080	3	4	-5.4	693726	GACTTGCGAGTAAGTAAAGAGCGTCCTCCGAT	1
CT614_069	3	1	1.2	694089	TTGGAAGAAGAAAAAATTGGTTCGGGTAAGAT	1
CT618_126	8	4	-2.1	698047	GCATCGATTTAAAAGCGATTTCTTTTTACAAT	1
CT621_069	8	4	2.5	706918	AAGTTGTAAAAAAAATATTATTGGGATAGGTT	1
CT626_135	3	4	2.4	714179	CCGTTGTAGAAAATTGGAAATAGAACTAGAAT	1
CT628_191	8	4	0.7	716549	GCGTTGTAATACAAAACGGAAAAAGTTAGGAG	1
CT632_068	8	4	-0.8	720038	TTTTTGCTGAAAAACTTTGAGTGTTTTGTTAT	1
CT632_065	8	1	0.4	720038	TTGCTGAAAAACTTTGAGTGTTTTGTTATGTG	1
CT634_040	8	4	0.5	722747	AAATTGGATTCCCTTATAAAAACTTCTATAAT	1
CT636_236	3	1	-4.5	723044	TTGGTAGTCCAGATCGTTGTCGTCAGTAAGAG	0.78
CT636_056	3	4	-0.5	723227	ATATTGCTCTTTTTTGTTATTCGGCGTATATT	1
CT636_043	3	1	-3.5	723237	TTGTTATTCGGCGTATATTTCCGGACTAAAAA	1
CT641_043	16	4	-6.4	733886	ACGTTGCTTGGCTGTGAGATAGTTCATTCATC	1
CT646_071	3	4	5.7	741250	TTCTTGAAAAAGATGTTTTTATTTTTTAAAAT	1
CT647_083	3	3	-9.2	742616	GCTTGGATCGGTGCGCGTGTTCTTTCTAAAAA	1
CT650_091	3	4	-0.1	745252	ATATTGCAAACTGACGCCACGAAATATAAGAG	1
CT651_140	8	3	0.9	746546	CTTTGAAAGGTTAAAATTTTTTGGTGTAAACT	1
CT658_074	8	4	-0.9	753473	AGGTTGAATAAATCTTTTCCGAACCGTATCAT	1
CT664_048	16	4	-0.8	760785	GCGTGGCTAGAAGATTTAGGACTAAGTAAAAC	1
CT665_076	8	4	1.9	763290	GGGTTGAAATATAAAATTGAGTACAATAAATA	1
CT667_075	8	1	-6.8	763836	TTGACTGCGGTGAATACTGAAATGATCACCAT	1
CT667_044	8	1	2.1	763867	TGGCAAGAGCTGTTAAAGGAAGTTAATAAAAT	1
CT673_212	8	1	-2.1	768819	TTGAATGGAGCTAAGGTTGGACGTGGTAACAT	1
CT674_122	16	4	4	770381	AAGTTGCAAGATAGAGGGCAAATAGATATATT	1
CT680_234	3	3	0.9	778741	TTTTGCAAGCCAAAATAAAATTTCTCTAAAAG	1
CT680_200	3	3	-2.4	778707	GATTGCATAAAAATCCTTGCTTCCAGTACTAT	1
CT680_187	3	4	-3.3	778693	TCCTTGCTTCCAGTACTATATCGGTCTACTTG	1
CT681_099	8	1	-5.5	780159	TTGCTACAGGACATCTTGTCTGGCTTTAACTA	1
CT682_083	16	4	1.9	780588	GTTTTGATAGAAATCGTTTTTATTACTACAAA	1
CT682_066	16	3	-5.6	780604	TTTTATTACTACAAAGAAAGTGTTTTTAAACG	1
CT683_046	16	3	-3	783930	TTTTGCTCCTGTAAAAGGTAAGGTATTACTTT	1
CT684_270	8	4	2.5	785057	GTGTTGTAGAAAAATTCCATTTTTTTTACGAT	1
CT684_093	8	5	-5.1	785235	TGTTTTGCCTTTTTTGAGACAGAGGAGATAAT	1
CT684_091	8	3	-2.1	785235	TTTTGCCTTTTTTGAGACAGAGGAGATAATAG	1
CT687_060	16	3	-1.5	788671	TTTTGACATTAGATATAGAGAAACCGTATTTT	1
CT688_098	16	4	1.8	789975	CTCTTGATTTGCATTATAAGATTTTGTATCTT	1
CT688_060	16	1	-4.7	790010	TTGTATTCTGGGGACGGTCTGTAGAATACAAC	1
CT690_246	8	4	1.4	792938	ACATTGATAGTCGAAGAAACTAATTCTAAATG	1
CT691_072	8	3	3.1	792786	GATTGCAAATATATATATGAAGGAGGTATATT	1
CT691_070	8	1	3.5	792786	TTGCAAATATATATATGAAGGAGGTATATTTT	1
CT691_040	8	1	-2.8	792816	TTGGGAGCATTTTTCTCAAAGACAATTAAAAA	1
CT693_059	8	4	-2.9	794885	GGCTTGAGTTTTCCTTTGCTTAGGCCTATAAG	1

CT693_047	8	4	1.2	794897	CCTTTGCTTAGGCCTATAAGAAAATTTAGGTT	1
CT694_080	16	3	2.7	796334	AATTGTTTTTCTTAAAAAGAAGTTTTTAAAAT	1
CT698_075	8	3	2.5	801940	TCTTGCTTAGAAAGATCATCCAAAAGTATAAT	1
CT700_306	8	4	-3.7	802801	AGGTTGCCTACGTGGAAGTAGGGGCGTTAAAT	1
CT701_065	8	5	-3.3	804627	ATCCTTGTTGTATAGGCGCCTTTAAATAAGAG	1
CT706_149	8	4	-2.1	813156	CGCTTGACCCAAGAGACACTTAAACATAGAAT	1
CT706_112	8	4	-2.5	813119	ATTTTGATGCGTAGGTGTATTTTGAATTAGAT	1
CT708_069	16	4	2.6	814796	TCATTGATTTAGCGGAAGTAAAAAGGTACAAG	1
CT709_090	16	4	-3.3	818271	GTCTTGTAAAGAAAGTGATCAATTCTGATGAT	1
CT712_059	24	5	-2	823641	ATTTTTGCTAAGTTGATCCGTAGATTTAAATA	1
CT712_055	24	1	5.3	823641	TTGCTAAGTTGATCCGTAGATTTAAATAAAAT	1
CT712_047	24	1	-5.4	823649	TTGATCCGTAGATTTAAATAAAATCGTTTTAG	1
CT714_097	8	4	0.2	827147	TATTTGAATTAGAAGCGGATTTTTATTACCCT	1
CT719_087	8	5	-7.6	831781	TTCTTTGTAGGATACTTAGCGGGCTCTACCCT	1
CT723_144	8	4	0.1	834695	ATCTTGATATGATGGAGGGTCGTTTTTATTTT	1
CT725_148	3	1	-4.4	836083	TTGCTTTCGAAAAAGCCGCGGAAACCTATGCG	1
CT728_155	8	3	-2.8	840925	GATTGTTGGTCGCTATTTTAGAAAATTATCAG	1
CT729_068	8	4	-2.1	842160	CAGTTGTTGCTCAGACAAAACTTCCATATACT	1
CT731_110	8	4	-4.4	843267	GGGTCGCAGAAAGTAGAGGTTTGTGTTATACT	1
CT733_230	8	4	-1.7	846714	ACCTTGCCCCTAACAAAAAATCATGTTAGCAT	1
CT734_092	1	1	-1.3	846780	TTGATTGGGATGTTGAAGCGCCTTTATAAACT	1
CT735_042	1	3	-5	847602	ATTTGCCTCCGAAGCGGGTGAAAGCATAAGAG	1
CT736_324	8	3	0.6	848745	GCTTGCGCAGAAAAAAGTTTAGAATATATGAT	1
CT740_108	8	5	-0.6	861014	GGCTTTGCTTCCTTGTAATAAAACTTTAAAAA	1
CT740_073	8	4	1.5	860980	CCGTTGATTTTTTATAGAGTAACCTATAACTT	1
CT741_315	8	4	0.6	861709	TCGTTGCGGCATGCAAAATAGAGCCTTAAAAT	1
CT743_085	24	3	-3.8	863340	GGTTGCATGAATTTGAACAAACAAACTAATTA	0.99
CT744_045	8	4	-2.3	863611	TCTTTGCAGTCTTCTTTGAAGAGGAATAAACA	1
CT745_190	3	4	-1.2	867574	GCTTTGATCGATACTTTGCAGATAGCTATGAT	1
CT752_103	8	1	-4.9	884405	TTGCAAGAAAGAGGGAAAGCGTTTGTTTCGCG	1
CT752_064	8	3	-3.5	884446	TCTGGACAAAGCTTAGAAGAGAACGATAACAT	1
CT755_056	3	4	-4.8	887911	AACTGGCGAGTTGGTAGTGTTCTAGGTATGAA	1
CT756_160	16	4	2.3	888149	AGGTTGAATTGATAATGGATATAGAATAATTT	1
CT757_077	16	5	1.3	889731	TAACTTGCCTTTTGAAAGCTTAAGTTTAAGAT	1
CT758_051	16	1	-4.6	890775	TTGTGTAGTAGTTGGGATCATTGCAGTATTTG	1
CT761_044	16	3	-3.8	893850	AATTGCTAATATGTGTGGCATGGGATTGCTAT	1
CT762_114	16	4	-4.9	894844	TACTTGCTTTAGATCCTGCAACCAGTGAAAAT	1
CT763_097	8	5	-1	897915	AATATTGACGCTTTTTTAGAATTTCATATATT	1
CT766_230	8	1	0.2	899046	TTGCAAAAGGACGTGTTAATGTTTTGTTAAAT	1
CT766_209	8	3	1	899069	TTTTGTTAAATATGGAGAAAGTTTTTTATATG	0.65
CT766_207	8	1	3.6	899069	TTGTTAAATATGGAGAAAGTTTTTTATATGAG	1
CT766_105	8	3	-1.8	899173	TGTTGTCTTAGAGATGAAGTTGCTGATATTAT	1
CT767_066	16	4	-2	901415	GTCTTGACAAAAAAAACAATTGGTAGTCAAAG	1
CT769_244	8	1	0.2	903340	TTGAAATAAAAATATGTCCATAAGGATAGCAG	1
CT773_113	8	1	-2.3	906504	TTGTCATTGTGAAAAATGTTGAAAAGTTACTT	1
CT774_105	1	3	-1.7	908754	TCTTGCTAAGAAACTATCTCTCTGGTTAGGAT	1
CT775_115	8	4	0.4	908838	TTTTTGATAATAAATGAAAAGAATAGTTCTTT	1
CT776_178	8	1	-5.3	909541	TTGCGGTGCGGAATTTTCTTCATGCCTAAACG	1

CT777_071	3	4	-4.3	911307	CTTTGGCAAAACGATTATTTAGCGAATAAGGA	1
CT779_125	3	1	-2.5	915457	TTGCTTTATCTAGCCTTTAATATGATAACAAT	1
CT779_085	3	3	-3.9	915415	GATTGCAATGAATTCGTCTGGGCTGATAACCT	1
CT781_053	8	3	-0.7	916163	GATTGTCGTAAGAAGAAAAATATTGCTACTAT	1
CT783_125	8	5	1.1	919418	TTGTTTGCTTTTAATGAAAAAAAGAATATACA	1
CT783_123	8	3	2.9	919418	GTTTGCTTTTAATGAAAAAAAGAATATACACG	1
CT783_073	8	4	-7	919469	GCTTTGGGAGAGGGGTTCTCTGGGTTTTCGAT	1
CT784_086	8	1	-4.1	921114	TGGCTACAAAAAGCGGAAGAAATCTTTTGAAT	1
CT785_091	8	4	-1.4	921266	GTATTGATCTTTTAGGTGAGATCCTTTAAACT	1
CT786_312	3	4	3	921346	TTATTGTCAAAAAGTGAGGAAAAGGGTAAATT	1
CT788_112	8	5	-0.9	922763	GTTTTTGTTTCTCGAGAAAAAGGTACTATGAT	1
CT790_152	8	4	-0.9	923158	TTATTGACTTAGATTGCAGTAACTTTTACCCT	1
CT790_081	8	4	0.5	923229	GTGTTGCTGTTAAAAATTTTTTGGCATAGAAA	1
CT790_065	8	5	-2.1	923246	TTTTTTGGCATAGAAATAGAGCTGAATAGAAG	1
CT793_052	8	5	-2.2	928144	AGTATTGCGTGATCGAAAAAATTTTTTATTCG	1
CT794_152	8	4	-3.4	928314	ACATTCCCAAGAAAATTATAGTGCGTTACCAT	1
CT794.1_056	8	4	3.9	930282	ATATTGCTTATTAGTTTCTTTTGTTATAGAAT	1
CT795_044	1	3	-3.4	930707	CCTTGATCGCAAAATTCTTATTTTAGTAGAAG	1
CT798_070	16	4	0	936900	ATCTTGATTATTTTTGTTAAAAGAAATAATTA	0.85
CT799_146	16	5	-3.3	937268	ACGCTTGCTTGGAGAGGATTTCTTCGTATGTG	1
CT802_058	3	4	-4.2	938952	GCTGTGAAAGAAGTTTTAGAATTCGCTACATT	1
CT808_121	16	1	1	948134	TTGTAGAATTTTTTACCTAATCGACTTATAAT	1
CT811_109	8	3	-2.7	949243	AATTGCAAGCAAGCTTTTATTCCTCATACAGT	1
CT814_111	16	5	1.6	956760	TCTATTGATTGGGAAAAAATTTATTCTAAATT	1
CT816_145	8	1	-0.1	958422	TTGTTAAGATATTCTGGTACAGAAAATATTTG	1
CT816_118	8	3	-4.9	958451	ATTTGCAGAGTTATGGTCGAGGGGACTAAAAA	1
CT816_071	8	3	0.1	958498	CCTTGCTAAGACAATTGTTGATGTTGTAGAAG	1
CT817_131	3	4	2	960362	ACTTGGCTAAATCTGTTACTGTAGAGTAAAAT	1
CT817_085	3	4	-1.4	960408	CTCTTGATGAATAGCATAAGCGTCTGTATCTT	1
CT819_070	8	4	-1.9	963036	ATCTTGCAACCCTGTATTATACGTTTTAGAAA	1
CT821_060	16	5	3.4	964785	AAAGTTGATTGAAGTAAAAAGAATAATAAAAG	1
CT823_107	8	4	0.8	966926	CCCTTGATTTGCATCATTAGATTTTGTATGCT	1
CT826_314	8	1	-0.6	974053	TTGCAAAAAGAAGCAGCTTACCTCTCTATCAT	1
CT826_106	8	3	-1.9	973843	GTTTGCATAAATGCAAATTCAAGCCATAAAAA	1
CT827_212	8	5	1.4	974167	AACGTTGCTAGCTTCTATATATGGTATACAAG	1
CT836_066	3	5	-3.6	983239	GGGTTTGCTGTAGTAGCTACCCAAGCTAAACT	1
CT837_088	8	3	1.6	984553	ATTTGAAAGCTAATTCATTTATAAAATAAACT	1
CT837_048	8	4	0.8	984594	ATCTTGAAATAATAAAAAGACTATTTTTCACT	1
CT838_278	8	4	0.7	987987	CGTTTGCATTTAGCAGAAATACTCCGTAGAAT	1
CT838_055	8	3	-1.7	987765	CATGGCTTATTTTTTCTTGGAGAATATACACT	1
CT839_044	8	1	-3.7	988823	TTGTTAGCTAAAACGTTCAACTGATTTTCTAT	1
CT840_076	8	4	-3.6	988804	CAGTTGAACGTTTTAGCTAACAAAACTATGCG	1
CT849_066	8	4	-0.4	998732	TTTTTATTAAAGAGAGAAATTGCTGGTAAAAT	1
CT854_064	8	4	-4.3	1E+06	CTCTTGCCGCATATGCTCTCTTCCCCTATGAT	1
CT856_129	3	3	-2.4	1E+06	GCTTGCGTTTCTATTCGCGATAAAAATAGAAG	1
CT859_177	16	4	-1.4	1E+06	ATCTTGATAAGGTCAAAAAATTGGTTTGTCAG	1
CT860_147	8	4	-1.7	1E+06	AAATTGGGAACGGGATGAATATTTTTTACGTG	1
CT863_074	16	4	4	1E+06	AACTTGCATGAAAAATACTTTTTAGATAAGTT	1

CT865_113	8	4	3.3	1E+06	ATGTTGATAATAACGTTTCTAAAAGGTACCAT	1
CT867_082	16	4	-1.7	1E+06	CGTTTGTTTTTGATTAATTTTAACTGGAAAAT	1
CT870_172	8	3	0.3	1E+06	GTTTGCATCACACAAAAGCTGAGAGATAAAAT	1

Appendix D: NNPP2.2 co-predictions
online at http://www.biomedcentral.com/1471-2105/10/271

seq32_id	**P NN**	**NNPP2.2 Predicted Sequence**
>CT006_080	98	CGATTTTTTGTCAGAATATCTCTCGCATATAAATTTTTGTGCCCGCAATG
>CT007_112	98	AGAAGTTTGCTAAAAATTTTATTAAGCAGTATGATCTACCAGATCCTAAA
>CT015_067	99	ATTTTTCTGGATATGATACACAATCAAAGTACTATCTCCCCCTTACACTT
>CT016_067	99	ATCTGTTTGTCAAAAATGTACCCCTTAACTACAATGCCGAGGAAAGTGAG
>CT025_060	96	TTTCCTTGAAAATCAAGCTAATGATGCTGTATCCTCTGGGGAGGTTTCTG
>CT025_098	97	CGCGTGTTGAAAGATTATGCGCAATTGGATAGGTTTTTTTTCCTTGAAAA
>CT029_132	98	TATTTGAGAGGAAAAACTGGTAAGGCTGCTAAAGTTAAAGAGCTTATCGG
>CT040_045	99	TTCTTGATTTGGATTAGCGAATAAATAACTACTATTGCGAATACTTAT
>CT049_065	100	TTTTGTTTGTTAATTTAATTTTTTCTAATTAAAAGAAAATTAGAATTTAA
>CT054_054	99	AAAATTTTGCTATGGGAATTTTCTAAAAGTATCACCTTCAATCGATAGAT
>CT061_043	99	TATTTGATTTAGCCTTATTTTTTAGTTTGTAAAAGAAATTTTTTTT
>CT066_169	96	TGAGATTGCGAAAAAAGCAAAAACCACTATAGAATTCACCAAAAGATAAA
>CT074_100	98	AACTTGTTCAATTAGGTATCTCAGATTCTTACAATCCAGGAATCATTACA
>CT078_081	98	CAGCTTTTGTAAACTGATCAAGAGAGGTCTAGACTCTTTCCCTCTAAACA
>CT080_071	98	TATGGTTTATGAAAAACAATTTTTTAATTTAAAATTAGAATAGATTTTGA
>CT084_120	96	TTTTTGTTTTGTTTTTAATTAAAAGAAAATAAATAGCCGTAAAAAATCAT
>CT084_154	100	TTTTTTGTTTAAAAACAGATTGGAAAATAGATTTTTTGTTTTGTTTTTAA
>CT091_071	99	TCTTTCTTGAGAAAAACATTTATATACGGTAACTTGCGAAGTATTCCTTA
>CT102_062	98	TAAGCTTTGGCTTCGTAGGATGAGGGACATATCTTGATTAGGATCCGGCG
>CT102_076	98	TATTGATTAGAACAAGGGCTCTTGATTAATAAGCTTTGGCTTCGTAGGAT
>CT110_144	98	AGGTTGGTAACATCGTTTTAATTGATAAATATTCTGGCCAAGAACTTACT
>CT111_132	97	AACCAGTTGCAAAAAAGCGAGGACTTTGCTATCGTTCTTCCTCTGAACGT
>CT114_163	98	AAGCGTTTGATACAGCAAAAAGTAGTGACTATGTTGTCTAAAGCTCTTTT
>CT115_132	99	AGGTGTATTTGAAAAAAGTTTGTTTTAAATAGTTTTTTTAGTTAAAATGG
>CT125_060	100	TTTTGTTTGGAAAAAATAATCATCAAAATTATAATCATTCCCTCTGATAA
>CT134_137	99	TTATTGGTTTTCATTCCAATTAAAGAGGATATGCTGGCTGCCCTTTAGAG
>CT139_208	99	AGGATTTGCAATACAAATAATGTCTCGTTTAACATAAATGAAACTCACTT
>CT140_135	100	TGTTTGTTGCCGGTTCTATTCTAGAAACGTATAATACTTCCCCAGAGAGC
>CT141_061	95	CTGGGGTTGCAAAGAAGGTTCTTTGTAATTAATTTTACGAATGAGAAGAG
>CT144_043	98	ACGTGATTTGGTTTCCTTTTGGTTCTTCTTATAAGGAGGCAGATTA
>CT145_109	100	TTGTTGTAGACGGATCGGAAAGCTACAAGTATAGTACTGGGAAGAGCAAA
>CT146_119	99	AGGATGTTGGAAAAGTCTAAAAGATGATCTACGTTGTGCACATCGCCATC
>CT156_068	96	AACTATATTGAAAAAGCGAACAACAAAACTAAAACGGAATCGTAACCCAT
>CT159_219	100	CTCTTGTGGAACAACAGTGGATCAGAGAATATGATCTTTCAGTAAATAAC
>CT164_047	96	CATCACTTGAGAGGTTTTCTCTATTTCGATACGATATACTTTTTTGAGTT
>CT170_106	97	TATTTGATTAAGAGATGTTCTTATAGAAGTAAGAGCGTCTTTTTTGCGCA
>CT172.1_294	99	ATTTTGCGACTTTCGAAATAGAACGAACATATAAGAATGGGAGTATCGTC
>CT175_098	97	CAGTTTTGGCTTTTTGCTATGGTTTTTTGTACAATCCCCTGGCCAGGTAA
>CT186_043	96	CTAGAGATGGAAAAGCTTGGCGGCCCTATTAATTTGTTTTGCAAGAGGTT
>CT189_285	99	GTATGTTTGTTACACCTTGAAAGAGGTAATAGACTACTTAAAAGGCCTTG
>CT195_186	100	GAGTTTTTGCAAAAGAAATGAAATAGCATTATAACTAGTTGGGTTTTATT
>CT196_048	99	TTCTTTTGCAAAAACTCTCTACTTTAAATTAATTTGGACTTTAAAACCTT
>CT196_049	99	TTCTTTTGCAAAAACTCTCTACTTTAAATTAATTTGGACTTTAAAACCTT
>CT199_049	97	TCAACTTGTTATCTGTGATTAGATCGCAATACAATATACAAAGGAATCTA
>CT200_104	96	ACTTTGTTGCGACCTTCTGCAAGCGTGGATAGATCCACAAATTCGTTATT
>CT203_221	96	TATCTTTTGCGTGTCAACTTAATGTTGTTTAGATTCGCCTTGTCTGCTTA

>CT204_043	99	TTCCTTTTGAGTCATAGTTTTTATCCAGATAAAATTATAGGGAGCC
>CT213_083	99	AAAGCTTTGAAGGAATCTTGAAGAGGTTGTAGGATTACTGTTTGAGGAAT
>CT215_114	100	AAATGTTTGCATGCTAAGAAAGATTTATAGACAATCTGCCATGTCGCATC
>CT221_242	96	ATTGCGTTGTGAACAAAAAACAGTGTATTTAAAGTATTTATTCTGATAAA
>CT221.1_171	100	TGTTTGTTTGAGTTTAGGGCTTTGTAAACTATTTTTGCGTATGAAGCGAG
>CT223_063	99	TTTTTGATTTCTTTAAAAAAATGTTTCTGTACAATTTCTTTCCTGTTTTA
>CT226_047	98	CTACCATTGAAATATATAAAAATTTTTACTTTAATTAGCTTTCCTGTAGC
>CT226_075	99	CTTTTGGTTGGGTAGATTAGGTATTTAACTACCATTGAAATATATAAAAA
>CT228_063	99	AATGATTTGCGAATAAAGCGCCGTTCTGATACAGTTTTTCCGTTTTAAAA
>CT230_075	96	CTTTCTTGCAAAAAAAATCTCTCCTTTCTTACATTTAGTCTCAAAAGATA
>CT230_076	96	CTTTCTTGCAAAAAAAATCTCTCCTTTCTTACATTTAGTCTCAAAAGATA
>CT246_097	96	GATGATTTTAAAGATTTGCGTCAAACAAGCAAAACAGCTGTCCTAGAAGC
>CT248_081	99	TCGTCTGTTGAGAATAGAAATCTTTCTTTTAATATTAATATTTCTTATTC
>CT249_060	98	AGGGTCTTGATATTCGGTAAAAAATCAAGTAAAATGTTCGCCTTTTTTAG
>CT249_128	100	GTTTGCTTGTGAATAAGAAATATTAATATTAAAAGAAAGATTTCTATTCT
>CT253_129	95	TATTAGTTGCTTTTTGAAAATACTCATGCTAGAGTTCTCCTTAATACATA
>CT261_062	96	TGGATGAATTGAAAAAGCAAGCAAAACTAGAAAATAACCAATCCGATCAG
>CT261_126	99	GTTGATTTGTGATAGCTTCTTCCATGCAATAGGATTGATCTCGACAGGAA
>CT261_131	99	GTTGATTTGTGATAGCTTCTTCCATGCAATAGGATTGATCTCGACAGGAA
>CT265_048	100	TTATTGTTTGTATTAAAATAACTCTTTTTTATAAGAAGGGAGCGTGCTAC
>CT265_064	99	TTTTGGTTGTTTTTGATTATTGTTTGTATTAAAATAACTCTTTTTTATAA
>CT265_111	99	TTGTTTGCGTGAAACAAAGGTCATTATAATCAAATAGTTGGTTTTTGGTT
>CT267_097	95	AAAAGTCTTGAATCCAAAGGATGAATGCATATTATACGCATATATTGCGG
>CT269_082	96	ACAAAGCTTGACAACGAATATGTGTATAGTAAACTATTTGAGAAAGCTTT
>CT286_067	96	AAAGTTGCATCATTATCATAAATGTCGTATATGCTTGAAAAATATTCCAC
>CT287_070	99	TGTTTTTTTGCTAACATGATGGCCCTTTGTAATGTGCAAACTTCAATAAA
>CT287_136	99	AATTGATTTGTCTTTTTGTAATTAAGATTTAAAAGAGGCGGCTTGAGAAA
>CT288_062	100	TAATTGTTGTAAAAAAACAATATTTATTCTAAAATAATAACCACAGTTAC
>CT293_065	99	TATTGATTGGTTAAAAAAAATTACAATAAAATTATTGCTCAATTTTTTGT
>CT293_094	100	GCTTTGTAGAAATGAGCAAGCAATAATTTTATTATTGATTGGTTAAAAAA
>CT321_208	97	ATGAGATTTGCGATTCGTGAAGGTGGTCGTACAATCGGTGCTGGAACTAT
>CT323_149	99	AAGTTGTTTGACATTTTCTGTTTAGTCGATATAATCGCTCTCTCGAGTTT
>CT324_264	98	AACAACTTGTCAAAAAACAGAAGGAAAAGTATCTTTTGACATTCTCTCTT
>CT326.2_063	96	AAGCTCTTGAGTTTATTTTCTAAGAAGGATAAAATCTCGAGCCAATATTA
>CT327_091	95	TATTGCTTTGATATAAATCTCTTGGATATGCTAATCTTCTTGTCTTACTT
>CT327_096	95	TATTGCTTTGATATAAATCTCTTGGATATGCTAATCTTCTTGTCTTACTT
>CT327_136	98	CGATTTCTTGCAAGGAAGGCTTATTTTTATATGATTTATTTTCTATTGCT
>CT337_096	99	TTTAGCTTTGAAAAGAAGCTCTAAGGGTTTATTATCGTATTCTTTTGATT
>CT338_054	99	AGCCATTTGGACGTTCGGATAACGCGAAATAGTTTATCCGCTCCTTATGA
>CT341_307	97	ATCTCTTGAAGCCTAAATAAAAGTGGTGTTACAATCCCCGGTCTCTTGTG
>CT342_102	97	ATCTCTTGAAGCCTAAATAAAAGTGGTGTTACAATCCCCGGTCTCTTGTG
>CT343_064	99	AATTTTTGAATTAAAACGGTTTTAACGGTTATAATCCTTTGTCTAAATCA
>CT343_082	97	GTTTTGATTTGGCGACAGCGAATTAATTTTTGAATTAAAACGGTTTTAAC
>CT344_094	99	CGATTTTGTAGTATATCGGTAAGAAGAAGTAATATTAAACATGTGCACAC
>CT346_210	96	TTTTTCTTGTGTTTAGGCTTAGTGGAAGTTATAATTTTTCTCTGAAACTG
>CT355_103	99	GGGTGTTTGCCACATCTCTAGGGAGGCGGTAAGATCCAAAGAAAAAGGGG
>CT355_224	96	TTTGGCTTGAGGATATAACGCTTTTTTGTTAAAAGTGTTCTGACGGCTGG
>CT356_116	97	ATTTTGGGTTGTTTATAACCATTTTTTATTAGGTTTTGTCGAGTAAAATA
>CT367_306	99	GCTGCTTTGCAAAAAATCCATCGCGCTTGTATAATGCGTTGGGAGCAAAG
>CT374_079	98	TGGTTTTTGCTTAACAGCTCTCGGTTTCTTAAATTTCGAAAATGCAGCGC

>CT376_107	96	AGGTACTTGATTCTTTTATCATCCAAACGTATGTTGGGACCAAAATTAGT
>CT377_075	100	TTTTTTGTTTGCAGAGTTTTTATTTTAAATATGTTATAATCTGTCTATTA
>CT378_080	97	CAAAAATGGAATCGCGAAAGATCACGAAAGATAGTACAAGTAAAAAGAAA
>CT382.1_088	99	GGTTTTAAATTAAATTGTTTTCATTTGAATTTAATTTTATTTTTTAAGAT
>CT383_075	98	ATTCATTGAAGACAAAGAAAAACTTTTGTTAAAATTTTTTCGCTATACCG
>CT385_045	100	GAGTTGTGAAGAAGCACTGAATTTAGTGTTAAAACAGCAGAGATTAGT
>CT393_183	98	AGATTTAAAAACAAAATTTGCAAAAACGCTTAAATTAATTTTAAGAAACA
>CT394_043	97	AAATTCTTGACCAGTGGAGACGGTTTTCTTATAATGACACCGACTT
>CT396_145	97	ACTATCTTGGAGGAGTTTACTAAAGGTTATAAGATAGGAGATCGTCCTAT
>CT399_070	100	ACAGCTTTGAAAAATCGCTTAGAGTCTGTTACGATGAGCTAAAAGACATT
>CT421.1_121	99	GCTCTATTGAAAAAAAAACGGCGCTCTACTATTCTGTTCCTTTCATAAGC
>CT440_198	100	ATCTTTGTTTGCAAGGGGTTATTTCTAGGTCTAGTAAGAGTTTATCGATC
>CT442_064	100	TGGGTTTTTGAAAAAAACAAGTGTTTGTGTAGACTCCCTGCTCACAACCA
>CT444_062	100	TTTATTGTTAAAAGAATTGCTTTTATCGATAAAAGAAACTTCAAGAGCCC
>CT444_130	100	ATTTGTTTTAAAAAACAATTGATATAATTTTTATTTTATAATGTAATATT
>CT444.1_115	97	AGTTTGCCTACAAACAAAACAAACTTCGATAGAATAAATAAACTAAAACT
>CT449_188	96	TTTGTTTTGATGGTTGCGGTTTAGCAGCTTAGTTTGGTAAAATGGACGAA
>CT450_185	96	GAGACTTGATAATAATCATTATCTATGGGTACCATCCCTGCTCTGAGTTC
>CT456_104	100	ATTTGTTTGTTTTTAAAAACAAATAAAAATAAACTTTGTAGCAAATAATA
>CT459_051	99	TCTTTGAAAAAGAAGAGTTTAAAGTTGGATATTTTGCGAAAGGTTAGCAA
>CT460_090	97	TTATTTTGACGATAAACCTAGTTAAGGCATAAAAGAGTTGCGAAGGAAGA
>CT461_052	95	TAGAAATTGACTCACGTGTTCCTCGTCTTTAAGATGAGGAACTAGTTCAT
>CT471_092	100	TCATTTTGATAATCTTTTTATCTTTCTAGTATGCTGACGGTAGGTTTTTG
>CT487_076	99	CTGGATTTGGGAATATCCGAGGAATTGAGTACACTCTCGTGGCTGCCGAT
>CT488_066	98	GAGCTGTTGACAAAGGTAAAAGTGGTTGGCAAAATAAAGAGCTGATTCTT
>CT489_056	99	GTTTTGTTTGTCTGGCTTTTTTAATAATTTATTTTTTAAAATTATTTTTT
>CT490_121	99	TAATTATTGCATGAATAATAAATCGATTTTATTTTTAGATGTTTGAAAAA
>CT494_045	100	GTTTTGTTGCTAATGCTAAAAGCGATTAATAGATTTATTAGGGTTTGC
>CT494_048	100	GTTTTGTTGCTAATGCTAAAAGCGATTAATAGATTTATTAGGGTTTGC
>CT496_153	99	TTTTTGTTTGTTTGAATGTTTTTTGTTGATAAGCTGGGGGAAATGGCGGG
>CT503_160	99	AGATGGTGGAATGGTAGACACTAGGGACTTAAAATCCCTTGGGCTTTGGC
>CT505_091	99	GGTTGGTGATAATGCTCAAAAGTGTATTATAGAATTTTTAGCATAGTGAT
>CT507_180	97	TTCTGGATTAAAAGAAGTTGAAGTAGGCTTAAAAGGAACTGGTGCAGGGC
>CT509_062	99	AAGACCTTGAAAAATAACAATTTTTGACCTAAGATGCTTATATTACTTTA
>CT512_100	99	GAGTTGTTTTCGATCGAGGAGCTCATAAGTATCATGGTGTAGTAGCTATG
>CT519_241	97	CACGATTTGTTAAGCGTGTGGAAAGGGTATAGTATGGGAGCAAAAAAGAA
>CT525_227	98	TTTATTGTTGCTGGGGACGCCACGAAGCCTATGATTGCTGAAGCCATAGA
>CT533_055	97	GCTTGCTTGCTAAAAAAAAAAAAGGATAATATACGGGGTCTCTTTGTCAG
>CT533_058	97	GCTTGCTTGCTAAAAAAAAAAAAGGATAATATACGGGGTCTCTTTGTCAG
>CT533_060	97	GCTTGCTTGCTAAAAAAAAAAAAGGATAATATACGGGGTCTCTTTGTCAG
>CT535_122	98	AATTTCTTGATTTTCTTCTATATCGAGACTACTATCCATTACAGAGATGA
>CT544_160	97	GTATTGACGACATTACTACAATTGTCCAATATGATGCCCGGGCTGGTCTA
>CT545_250	99	TGATGATTGGATTGGCAGCTGCAGAATTGTCTCATAAAAAAGCAGCTGGA
>CT546_041	99	ATTTATTTGGCATTGCTGTTTTTATTTATTAAAATAAATAAAAAGGTTGG
>CT546_045	99	ATTTATTTGGCATTGCTGTTTTTATTTATTAAAATAAATAAAAAGGTTGG
>CT546_050	99	ATTTATTTGGCATTGCTGTTTTTATTTATTAAAATAAATAAAAAGGTTGG
>CT547_065	97	GAGATTTGACAAATTCTCTTTTTCTTTTTTATGATGACGCTTTGTTATTA
>CT547_097	97	TTTTTGTTTTTGCTAAACAAAAGCTATAAGAGATTTGACAAATTCTCTTT
>CT559_055	96	TCCCGATTGGCACTAATCTCCCCATTTGCTATGGTGAGTGAAAAGGTGTG
>CT565_053	97	GATAATTGTTTAGTACGAGATTAATTTAATAAAAAGAAAAAAATCAGGTA

```
>CT565_061       99  TATCTTTTGATAATTGTTTAGTACGAGATTAATTTAATAAAAAGAAAAAA
>CT573_061       96  TTTTCTTTTGTAAAAATTGATTTTTTTTCTCAAATCAGTTACTTTATACA
>CT576_077      100  ACTTGTTAAATCAGATCGTTAGAATTTAATATTGTTAGTAGTAATTTGTT
>CT603_149       96  AATTTCTTGATCATTTAGCGAAAGCATGGTATCCTCCTGCCCTATAGTTT
>CT604_061       99  GGTTGTTTGACAAGAATAAATCGCCTTTCTATATCCAAGACTGCAGCTGA
>CT611_117       99  AGCTTGTTTTGTCACACATATCAAAAAGGAATATTGGGGCGATACTTTCT
>CT614_069       99  CTATTGGAAGAAGAAAAAATTGGTTCGGGTAAGATTAAAAGTTATAAAAA
>CT621_069       99  AGAAAGTTGTAAAAAAAATATTATTGGGATAGGTTCGCGACAAGTACAAC
>CT646_071       98  TTTTTCTTGAAAAAGATGTTTTTATTTTTTAAAATGAGCGCTCTTCATTT
>CT651_140      100  TTTCTTTGAAAGGTTAAAATTTTTTGGTGTAAACTCCACGGATCTTTGGT
>CT658_074       97  GGAAGGTTGAATAAATCTTTTCCGAACCGTATCATGGAAGGGTTTCAAAA
>CT665_076      100  ATCTTTTTAGAACGGGAAGGGTTGAAATATAAAATTGAGTACAATAAATA
>CT673_212       99  ATTTTGAATGGAGCTAAGGTTGGACGTGGTAACATCATTGCTTTGCAAGA
>CT674_122       97  CTGAAGTTGCAAGATAGAGGGCAAATAGATATATTCTGCCAAACAGATAA
>CT680_234      100  TCTTTTTGCAAGCCAAAATAAAATTTCTCTAAAAGAAGATTGCATAAAAA
>CT683_046       97  ATTCGTGATTGGCACGGTTTTTGCTCCTGTAAAAGGTAAGGTATTACTTT
>CT684_091       99  ATTTGTTTTGCCTTTTTTGAGACAGAGGAGATAATAGGCTCTTTTCTCAC
>CT684_093       99  ATTTGTTTTGCCTTTTTTGAGACAGAGGAGATAATAGGCTCTTTTCTCAC
>CT684_270       96  TTTGTGTTGTAGAAAAATTCCATTTTTTTTACGATTCGTAGCCAACAAGT
>CT687_060       96  GCTTTTTGACATTAGATATAGAGAAACCGTATTTTCCAAAACTGCAGAAA
>CT688_098       95  CATCTCTTGATTTGCATTATAAGATTTTGTATCTTGAGTTTTTGTATTCT
>CT691_070       97  CTAGATTGCAAATATATATATGAAGGAGGTATATTTTGGGAGCATTTTTC
>CT691_072       97  CTAGATTGCAAATATATATATGAAGGAGGTATATTTTGGGAGCATTTTTC
>CT706_149       96  AAACGCTTGACCCAAGAGACACTTAAACATAGAATTCATCATTTTGATGC
>CT708_069       96  TTTTCATTGATTTAGCGGAAGTAAAAAGGTACAAGTAACAGGTCTGTCAA
>CT712_047       97  TTTTTGCTAAGTTGATCCGTAGATTTAAATAAAATCGTTTTAGAGGGCAA
>CT712_055       97  TTTTTGCTAAGTTGATCCGTAGATTTAAATAAAATCGTTTTAGAGGGCAA
>CT712_059       97  TTTTTGCTAAGTTGATCCGTAGATTTAAATAAAATCGTTTTAGAGGGCAA
>CT723_144       97  GAAATCTTGATATGATGGAGGGTCGTTTTTATTTTTACCCTTGAATCTAA
>CT728_155       96  GGATTGGCAACGCCAAGATTGTTGGTCGCTATTTTAGAAAATTATCAGCA
>CT733_230       99  AGATCTTTGCAAGAACCTTGCCCCTAACAAAAAATCATGTTAGCATGAAG
>CT734_092      100  AATTTGATTGGGATGTTGAAGCGCCTTTATAAACTATTTAGTTAATAAGA
>CT740_073       99  AAACTTTAAAAAACTCCGTTGATTTTTTATAGAGTAACCTATAACTTGAC
>CT752_064       96  TCTTCTGGACAAAGCTTAGAAGAGAACGATAACATAGATGGAGAAAAGAT
>CT756_160       97  TAGAGGTTGAATTGATAATGGATATAGAATAATTTTTAAGACTGCTAGTA
>CT757_077       99  TTTTAACTTGCCTTTTGAAAGCTTAAGTTTAAGATAGAGAATTTCTTATA
>CT763_097       95  AAATATTGACGCTTTTTTAGAATTTCATATATTCTTCCCACAATCTTGGG
>CT766_105       97  ATCTGTTGTCTTAGAGATGAAGTTGCTGATATTATTCGTATCGCAGGATT
>CT766_207       99  GTTTTGTTAAATATGGAGAAAGTTTTTTATATGAGTAGTGCAGGTTTACG
>CT766_209       99  GTTTTGTTAAATATGGAGAAAGTTTTTTATATGAGTAGTGCAGGTTTACG
>CT773_113       98  TTCTTTTTGCTTTTATTCTTGTCATTGTGAAAAAATGTTGAAAAGTTACTT
>CT779_125       97  AATTGTAAGGAATTGCTTTATCTAGCCTTTAATATGATAACAATGATGCG
>CT781_053       98  GATTGTCGTAAGAAGAAAAATATTGCTACTATTTTTGAGCCAAAGGACGG
>CT788_112       98  ATTGTTTTTGTTTCTCGAGAAAAAGGTACTATGATGATCTTTTTTTTAGC
>CT790_065       99  AAATTTTTTGGCATAGAAATAGAGCTGAATAGAAGAGACAAGATCACGAG
>CT795_044       98  CATTGTGATAGCCTTGATCGCAAAATTCTTATTTTAGTAGAAGAGGAGTT
>CT808_121      100  TTTTTGTAGAATTTTTTACCTAATCGACTTATAATCCGCCTTCTGCTTAA
>CT811_109       97  AGTAATTGCAAGCAAGCTTTTATTCCTCATACAGTTTGTGCTTCTTGTGG
>CT814_111       96  ACCTCTATTGATTGGGAAAAAATTTATTCTAAATTGGTTACGCAAGAAGT
>CT816_118       99  AATATTTGCAGAGTTATGGTCGAGGGGACTAAAAAGCATCAAGTGGACTC
```

```
>CT816_145       99  GTATTTTGTTAAGATATTCTGGTACAGAAAATATTTGCAGAGTTATGGTC
>CT817_085       96  TCTCTCTTGATGAATAGCATAAGCGTCTGTATCTTAGATGGAATCGAGAA
>CT817_131       96  GGAACTTGGCTAAATCTGTTACTGTAGAGTAAAATCTACAGTTTTTTCTC
>CT821_060      100  AGTTGATTGAAGTAAAAAGAATAATAAAAGATAAGGAGGAAAAATTAAAG
>CT826_106       97  TTTGTTTGCATAAATGCAAATTCAAGCCATAAAAAGAAGCTTCTCAACAA
>CT837_088       99  AATATTTGAAAGCTAATTCATTTATAAAATAAACTAGAAGACAATCTTGA
>CT838_278       98  GCTCGTTTGCATTTAGCAGAAATACTCCGTAGAATTGCTATCGGCCTATT
>CT849_066       98  TTATTTTTATTAAAGAGAGAAATTGCTGGTAAAATAAAAAATAAAAAAAC
>CT854_064       96  ATTCTCTTGCCGCATATGCTCTCTTCCCCTATGATTCTTCCTTCATGAAG
>CT863_074       97  TCCAACTTGCATGAAAAATACTTTTTAGATAAGTTCCCTCCTTTCTAAAA
>CT865_113       97  GCTATGTTGATAATAACGTTTCTAAAAGGTACCATGGAATAGCTCTCCAT
>CT867_082       99  AAAATTTTTATAAAACGTTTGTTTTTGATTAATTTTAACTGGAAAATCCC
>CT870_172       97  GACGTTTGCATCACACAAAAGCTGAGAGATAAAATTAATTACTCCACTTC
```

Appendix E: TSS-PREDICT co-predictions
online at http://www.biomedcentral.com/1471-2105/10/271

seq32_id	tss_h35	tss_h10	tss_spacer	disc len	tss_loc
>CT006_080	ttgtca	taaatt	16	7	6979
>CT007_112	ttgcta	tatgat	17	7	7178
>CT015_067	tggata	tactat	16	8	17454
>CT016_067	ttgtca	tacaat	17	7	17608
>CT022_089	ttgctt	tatgct	18	6	27441
>CT029_132	aggaaa	taaagt	15	5	33059
>CT049_065	ttgttt	tctaat	15	9	54083
>CT053_255	ttgctt	taccct	17	4	60765
>CT054_076	ttgaaa	gggaat	17	7	60847
>CT062_064	ttgcta	gataag	16	4	71880
>CT066_169	ttgcga	tagaat	18	6	79199
>CT072_068	ttgtat	tagagt	15	9	85110
>CT078_081	ttgtaa	tagact	17	7	92818
>CT079_150	ttgctg	tatagt	17	6	93449
>CT091_071	ttgaga	taactt	17	6	106572
>CT098_072	ttgcct	tacact	17	8	115706
>CT105_215	ttgaga	tttatt	18	6	120273
>CT110_144	ttggta	aatatt	18	5	128079
>CT114_079	atggca	taaaat	17	6	133205
>CT114_163	ttgata	tatgtt	17	6	133121
>CT125_060	ttggaa	tataat	17	8	141474
>CT140_135	ttgccg	tataat	17	8	157182
>CT141_061	ttgcaa	taattt	17	5	158123
>CT145_109	tagacg	tatagt	17	4	161670
>CT149_121	ttgttt	tattat	17	6	172846
>CT150_071	ttgatt	tagcat	17	7	173130
>CT162_137	ttgtgt	tacatt	16	6	185539
>CT164_047	ttgaga	tacgat	17	12	187369
>CT175_098	ttggct	tacaat	18	6	196797
>CT182_118	ttgtta	tgtact	16	7	203608
>CT189_285	ttgaaa	taaaag	16	6	213169
>CT191_089	ttgact	tataag	18	6	215757
>CT191_091	ttgact	tataag	18	6	215757
>CT196_093	gtttaa	tataat	15	6	221007
>CT197_074	ttgctt	tacgat	17	8	221424
>CT199_049	ttgtta	tacaat	18	4	224140
>CT200_104	ttgcga	tagatc	17	5	225059
>CT203_221	ttgcgt	tagatt	17	6	227679
>CT209_107	ttgatt	tatgtt	18	5	238295
>CT215_114	ttgcat	gacaat	17	5	243400
>CT217_099	tggatt	taaaat	17	5	246029
>CT218_090	atcata	tattat	19	10	246870
>CT221.1_140	ttgcgt	gataaa	18	6	250798
>CT223_063	tttctt	tacaat	16	5	252090
>CT226_075	ttggtt	ttaact	15	5	254029
>CT230_075	ttgcaa	tacatt	18	6	256350

>CT230_076	ttgcaa	tacatt	18	6	256350
>CT232_095	ttgctt	tatact	17	7	259339
>CT235_088	ttgcta	tattat	18	5	263179
>CT248_081	ttgaga	taatat	16	7	279371
>CT249_060	ttgata	taaaat	17	6	279463
>CT253_129	ttgctt	tagagt	17	7	284988
>CT259_102	tttctt	tatacc	18	6	291666
>CT261_062	ttgaaa	gaaaat	15	6	292401
>CT265_064	ttggtt	tattaa	18	6	297408
>CT265_140	ttgatt	taaaat	16	7	297480
>CT267_097	ttgaat	tattat	16	7	299171
>CT269_082	ttgaca	taaact	16	6	301522
>CT274_139	ttgaag	tagtac	15	7	305668
>CT275_093	ttgatt	tacatt	18	6	306372
>CT286_067	ttgcat	tatgct	19	5	317942
>CT287_070	ttgcta	taatgt	16	5	321624
>CT288_062	ttgtaa	taaaat	17	4	321797
>CT293_065	ttgatt	taaaat	18	6	327383
>CT294_107	ttgaca	taatct	16	5	328118
>CT313_063	ttgaat	taggct	17	6	350400
>CT317_052	tggata	tataat	18	8	359775
>CT323_149	ttgaca	tataat	16	6	363844
>CT326.2_063	ttgagt	taaaat	17	7	367813
>CT327_136	ttgcaa	tatgat	16	6	369233
>CT328_073	ttgcaa	tttgat	17	7	369298
>CT337_096	ttgaaa	tattat	16	6	385070
>CT343_064	ttgaat	tataat	18	6	391805
>CT346_210	ttgtgt	tataat	17	7	395194
>CT355_224	ttgagg	taaaag	17	7	406789
>CT374_079	ttgctt	taaatt	17	7	426419
>CT376_060	ttgcaa	tatgtg	15	5	430187
>CT377_075	tgcaga	tataat	19	6	430530
>CT383_075	ttgaag	taaaat	18	7	436636
>CT390_081	tggaca	tatatt	17	7	444446
>CT393_071	ttgatc	taagat	17	7	447995
>CT394_043	ttgacc	tataat	17	7	449837
>CT396_145	ttggag	taagat	17	6	451507
>CT399_070	ttgaaa	tacgat	17	5	458435
>CT400_069	ttctca	tactaa	16	5	459624
>CT410_115	ttgccc	tattat	19	5	467927
>CT412_077	tcgcta	tatagt	17	4	471235
>CT413_073	ttgcct	tctaac	19	4	474307
>CT421.1_121	ttgaaa	tattct	17	4	491017
>CT435_061	ttgtct	gatatt	15	6	505194
>CT437_157	ttgcta	ttcaat	17	4	507713
>CT442_064	ttgaaa	tagact	16	5	511819
>CT444_130	ttgata	taatat	19	4	514137
>CT444.1_115	tacaaa	tagaat	15	5	514304
>CT449_188	ttgcgg	taaaat	18	5	524328
>CT450_185	ttgata	taccat	18	4	524491

>CT459_051	ttgaaa	gatatt	18	6	536500
>CT471_092	ttgata	tagtat	15	7	545914
>CT471_217	ttgcgg	tatacg	15	7	546038
>CT496_088	ttgctt	cataat	17	6	574765
>CT512_100	ttgttt	aagtat	17	6	588931
>CT514_048	ttgctt	caaatt	15	6	589840
>CT518_198	ttgttg	tatact	15	7	591693
>CT524_065	ttgaaa	aagagt	15	7	593797
>CT527_104	ttgact	gctcgt	16	10	595737
>CT528_191	ttgtgt	taccat	17	5	596497
>CT533_055	ttgctt	gataat	16	9	600971
>CT533_058	ttgctt	gataat	16	9	600971
>CT533_060	ttgctt	gataat	16	9	600971
>CT535_122	ttgatt	tactat	17	4	603313
>CT544_160	ttgacg	tccaat	15	10	611035
>CT546_045	ttggca	taaaat	17	5	617366
>CT546_050	ttggca	taaaat	17	5	617366
>CT547_065	ttgaca	tatgat	18	5	617477
>CT547_097	ttgttt	tataag	15	9	617447
>CT556_137	ttgatt	tataat	16	6	628230
>CT557_165	ttgaga	tacaac	17	5	628889
>CT565_061	ttgata	taattt	17	5	636864
>CT566_053	ttgcca	tatatt	15	7	637939
>CT569_136	ttgcta	gattat	17	6	639307
>CT573_061	ttttct	tacaat	16	6	645472
>CT595_197	ttgaaa	tatttg	17	4	676888
>CT603_069	tacgaa	tcaaat	16	9	682217
>CT604_061	ttgaca	tatatc	17	7	684021
>CT608_127	ttgcca	tatagt	18	6	687011
>CT621_069	ttgtaa	taggtt	17	6	706883
>CT626_135	ttgtag	tagaat	17	6	714144
>CT632_065	ttgctg	tgttat	17	8	720001
>CT632_068	ttgctg	tgttat	17	8	720001
>CT634_040	ttggat	tataat	17	6	722712
>CT636_056	ttgctc	tatatt	17	7	723263
>CT646_071	ttgaaa	taaaat	17	6	741285
>CT651_140	ttgaaa	taaact	18	5	746581
>CT665_076	ttgaaa	aataaa	15	6	763323
>CT681_099	ttgcta	tttaac	18	6	780123
>CT682_083	ttgttt	tattac	16	7	780618
>CT683_046	ttgctc	tattac	15	6	783963
>CT684_091	ttgcct	gataat	16	7	785270
>CT684_093	ttgcct	gataat	16	7	785270
>CT688_060	ttgtat	tagaat	15	6	790043
>CT690_246	ttgata	ctaaat	16	6	792904
>CT691_070	ttgcaa	tatatt	18	6	792822
>CT691_072	ttgcaa	tatatt	18	6	792822
>CT693_059	ttgagt	tataag	17	8	794922
>CT694_080	ttgttt	taaaat	18	6	796370
>CT698_075	ttgctt	tataat	18	7	801903

>CT706_112	ttgatg	ttagat	17	7	813083
>CT708_069	ttgatt	tacaag	17	8	814833
>CT712_059	atcaga	tagatt	19	6	823669
>CT729_068	ttgttg	tatact	17	7	842124
>CT731_110	tcgcag	tatact	17	6	843302
>CT734_092	ttgatt	tataaa	18	7	846817
>CT740_073	ttgatt	tataac	15	8	860945
>CT752_103	ttgcaa	tttgtt	15	9	884441
>CT756_160	ttgaat	aataat	15	6	888182
>CT757_077	ttttaa	taagtt	17	6	889759
>CT763_097	ttgacg	tatatt	16	8	897879
>CT766_105	ttgtct	tattat	18	4	899207
>CT773_113	ttgtca	aaaagt	15	6	906537
>CT774_105	ttgcta	taggat	18	6	908718
>CT779_085	ttgcaa	taacct	18	6	915379
>CT781_053	ttgtcg	tactat	18	7	916200
>CT788_112	ttgttt	tactat	19	5	922733
>CT790_081	ttgctg	tagaaa	17	6	923264
>CT794.1_056	ttgctt	tagaat	17	6	930317
>CT799_146	ttggag	catttt	19	6	937309
>CT808_121	tagaat	tataat	17	7	948095
>CT817_085	ttgatg	tatctt	17	7	960444
>CT817_131	ttggct	taaaat	18	5	960396
>CT827_212	ttgcta	tacaag	16	5	974200
>CT837_048	ttgaaa	ttcact	17	5	984628
>CT837_088	ttgaaa	aataaa	16	5	984586
>CT849_066	tttatt	taaaat	18	5	998698
>CT854_064	ttgccg	tatgat	17	7	1004522
>CT863_074	ttgcat	taagtt	17	5	1017779

Appendix F: Co-predictions of all 3 algorithms
online at http://www.biomedcentral.com/1471-2105/10/271

seq32_id	NNPP2.2 Predicted Sequence	tss_h35	tss_loc
>CT006_080	CGATTTTTTGTCAGAATATCTCTCGCATATAAATTTTTGTGCCCGCAATG	ttgtca	6979
>CT007_112	AGAAGTTTGCTAAAAATTTTATTAAGCAGTATGATCTACCAGATCCTAAA	ttgcta	7178
>CT015_067	ATTTTTCTGGATATGATACACAATCAAAGTACTATCTCCCCCTTACACTT	tggata	17454
>CT016_067	ATCTGTTTGTCAAAAATGTACCCCTTAACTACAATGCCGAGGAAAGTGAG	ttgtca	17608
>CT029_132	TATTTGAGAGGAAAAACTGGTAAGGCTGCTAAAGTTAAAGAGCTTATCGG	aggaaa	33059
>CT049_065	TTTTGTTTGTTAATTTAATTTTTTCTAATTAAAAGAAAATTAGAATTTAA	ttgttt	54083
>CT066_169	TGAGATTGCGAAAAAAGCAAAAACCACTATAGAATTCACCAAAAGATAAA	ttgcga	79199
>CT078_081	CAGCTTTTGTAAACTGATCAAGAGAGGTCTAGACTCTTTCCCTCTAAACA	ttgtaa	92818
>CT091_071	TCTTTCTTGAGAAAAACATTTATATACGGTAACTTGCGAAGTATTCCTTA	ttgaga	106572
>CT110_144	AGGTTGGTAACATCGTTTTAATTGATAAATATTCTGGCCAAGAACTTACT	ttggta	128079
>CT114_163	AAGCGTTTGATACAGCAAAAAGTAGTGACTATGTTGTCTAAAGCTCTTTT	ttgata	133121
>CT125_060	TTTTGTTTGGAAAAAATAATCATCAAAATTATAATCATTCCCTCTGATAA	ttggaa	141474
>CT140_135	TGTTTGTTGCCGGTTCTATTCTAGAAACGTATAATACTTCCCCAGAGAGC	ttgccg	157182
>CT141_061	CTGGGGTTGCAAAGAAGGTTCTTTGTAATTAATTTTACGAATGAGAAGAG	ttgcaa	158123
>CT145_109	TTGTTGTAGACGGATCGGAAAGCTACAAGTATAGTACTGGGAAGAGCAAA	tagacg	161670
>CT164_047	CATCACTTGAGAGGTTTTCTCTATTTCGATACGATATACTTTTTTGAGTT	ttgaga	187369
>CT175_098	CAGTTTTGGCTTTTTGCTATGGTTTTTTGTACAATCCCCTGGCCAGGTAA	ttggct	196797
>CT189_285	GTATGTTTGTTACACCTTGAAAGAGGTAATAGACTACTTAAAAGGCCTTG	ttgaaa	213169
>CT199_049	TCAACTTGTTATCTGTGATTAGATCGCAATACAATATACAAAGGAATCTA	ttgtta	224140
>CT200_104	ACTTTGTTGCGACCTTCTGCAAGCGTGGATAGATCCACAAATTCGTTATT	ttgcga	225059
>CT203_221	TATCTTTTGCGTGTCAACTTAATGTTGTTTAGATTCGCCTTGTCTGCTTA	ttgcgt	227679
>CT215_114	AAATGTTTGCATGCTAAGAAAGATTTATAGACAATCTGCCATGTCGCATC	ttgcat	243400
>CT223_063	TTTTTGATTTCTTTAAAAAAATGTTTCTGTACAATTTCTTTCCTGTTTTA	tttctt	252090
>CT226_075	CTTTTGGTTGGGTAGATTAGGTATTTAACTACCATTGAAATATATAAAAA	ttggtt	254029
>CT230_075	CTTTCTTGCAAAAAAAATCTCTCCTTTCTTACATTTAGTCTCAAAAGATA	ttgcaa	256350
>CT248_081	TCGTCTGTTGAGAATAGAAATCTTTCTTTTAATATTAATATTTCTTATTC	ttgaga	279371
>CT249_060	AGGGTCTTGATATTCGGTAAAAAATCAAGTAAAATGTTCGCCTTTTTTAG	ttgata	279463
>CT253_129	TATTAGTTGCTTTTTGAAAATACTCATGCTAGAGTTCTCCTTAATACATA	ttgctt	284988
>CT261_062	TGGATGAATTGAAAAAGCAAGCAAAACTAGAAAATAACCAATCCGATCAG	ttgaaa	292401
>CT265_064	TTTTGGTTGTTTTTGATTATTGTTTGTATTAAAATAACTCTTTTTTATAA	ttggtt	297408
>CT267_097	AAAAGTCTTGAATCCAAAGGATGAATGCATATTATACGCATATATTGCGG	ttgaat	299171
>CT269_082	ACAAAGCTTGACAACGAATATGTGTATAGTAAACTATTTGAGAAAGCTTT	ttgaca	301522
>CT286_067	AAAGTTGCATCATTATCATAAATGTCGTATATGCTTGAAAAATATTCCAC	ttgcat	317942
>CT287_070	TGTTTTTTTGCTAACATGATGGCCCTTTGTAATGTGCAAACTTCAATAAA	ttgcta	321624
>CT288_062	TAATTGTTGTAAAAAAACAATATTTATTCTAAAATAATAACCACAGTTAC	ttgtaa	321797
>CT293_065	TATTGATTGGTTAAAAAAAATTACAATAAAATTATTGCTCAATTTTTTGT	ttgatt	327383
>CT323_149	AAGTTGTTTGACATTTTCTGTTTAGTCGATATAATCGCTCTCTCGAGTTT	ttgaca	363844
>CT326.2_063	AAGCTCTTGAGTTTATTTTCTAAGAAGGATAAAATCTCGAGCCAATATTA	ttgagt	367813
>CT327_136	CGATTTCTTGCAAGGAAGGCTTATTTTTATATGATTTATTTTCTATTGCT	ttgcaa	369233
>CT337_096	TTTAGCTTTGAAAAGAAGCTCTAAGGGTTTATTATCGTATTCTTTTGATT	ttgaaa	385070
>CT343_064	AATTTTTGAATTAAAACGGTTTTAACGGTTATAATCCTTTGTCTAAATCA	ttgaat	391805
>CT346_210	TTTTTCTTGTGTTTAGGCTTAGTGGAAGTTATAATTTTTCTCTGAAACTG	ttgtgt	395194
>CT355_224	TTTGGCTTGAGGATATAACGCTTTTTTGTTAAAAGTGTTCTGACGGCTGG	ttgagg	406789
>CT374_079	TGGTTTTTGCTTAACAGCTCTCGGTTTCTTAAATTTCGAAAATGCAGCGC	ttgctt	426419
>CT377_075	TTTTTTGTTTGCAGAGTTTTTATTTTAAATATGTTATAATCTGTCTATTA	tgcaga	430530
>CT383_075	ATTCATTGAAGACAAAGAAAAACTTTTGTTAAAATTTTTTCGCTATACCG	ttgaag	436636

>CT394_043	AAATTCTTGACCAGTGGAGACGGTTTTCTTATAATGACACCGACTT	ttgacc	449837
>CT396_145	ACTATCTTGGAGGAGTTTACTAAAGGTTATAAGATAGGAGATCGTCCTAT	ttggag	451507
>CT399_070	ACAGCTTTGAAAAATCGCTTAGAGTCTGTTACGATGAGCTAAAAGACATT	ttgaaa	458435
>CT421.1_121	GCTCTATTGAAAAAAAAACGGCGCTCTACTATTCTGTTCCTTTCATAAGC	ttgaaa	491017
>CT442_064	TGGGTTTTTGAAAAAAACAAGTGTTTGTGTAGACTCCCTGCTCACAACCA	ttgaaa	511819
>CT444_130	ATTTGTTTTAAAAAACAATTGATATAATTTTTATTTTATAATGTAATATT	ttgata	514137
>CT444.1_115	AGTTTGCCTACAAACAAAACAAACTTCGATAGAATAAATAAACTAAAACT	tacaaa	514304
>CT449_188	TTTGTTTTGATGGTTGCGGTTTAGCAGCTTAGTTTGGTAAAATGGACGAA	ttgcgg	524328
>CT450_185	GAGACTTGATAATAATCATTATCTATGGGTACCATCCCTGCTCTGAGTTC	ttgata	524491
>CT459_051	TCTTTGAAAAAGAAGAGTTTAAAGTTGGATATTTTGCGAAAGGTTAGCAA	ttgaaa	536500
>CT471_092	TCATTTTGATAATCTTTTTATCTTTCTAGTATGCTGACGGTAGGTTTTTG	ttgata	545914
>CT512_100	GAGTTGTTTTCGATCGAGGAGCTCATAAGTATCATGGTGTAGTAGCTATG	ttgttt	588931
>CT533_055	GCTTGCTTGCTAAAAAAAAAAAAGGATAATATACGGGGTCTCTTTGTCAG	ttgctt	600971
>CT535_122	AATTTCTTGATTTTCTTCTATATCGAGACTACTATCCATTACAGAGATGA	ttgatt	603313
>CT544_160	GTATTGACGACATTACTACAATTGTCCAATATGATGCCCGGGCTGGTCTA	ttgacg	611035
>CT546_045	ATTTATTTGGCATTGCTGTTTTTATTTATTAAAATAAATAAAAAGGTTGG	ttggca	617366
>CT547_065	GAGATTTGACAAATTCTCTTTTTCTTTTTTATGATGACGCTTTGTTATTA	ttgaca	617477
>CT565_061	TATCTTTTGATAATTGTTTAGTACGAGATTAATTTAATAAAAAGAAAAAA	ttgata	636864
>CT573_061	TTTTCTTTTGTAAAAATTGATTTTTTTTCTCAAATCAGTTACTTTATACA	ttttct	645472
>CT604_061	GGTTGTTTGACAAGAATAAATCGCCTTTCTATATCCAAGACTGCAGCTGA	ttgaca	684021
>CT621_069	AGAAAGTTGTAAAAAAAATATTATTGGGATAGGTTCGCGACAAGTACAAC	ttgtaa	706883
>CT646_071	TTTTTCTTGAAAAAGATGTTTTTATTTTTTAAAATGAGCGCTCTTCATTT	ttgaaa	741285
>CT651_140	TTTCTTTGAAAGGTTAAAATTTTTTGGTGTAAACTCCACGGATCTTTGGT	ttgaaa	746581
>CT665_076	ATCTTTTTAGAACGGGAAGGGTTGAAATATAAAATTGAGTACAATAAATA	ttgaaa	763323
>CT683_046	ATTCGTGATTGGCACGGTTTTTGCTCCTGTAAAAGGTAAGGTATTACTTT	ttgctc	783963
>CT684_091	ATTTGTTTTGCCTTTTTTGAGACAGAGGAGATAATAGGCTCTTTTCTCAC	ttgcct	785270
>CT691_070	CTAGATTGCAAATATATATATGAAGGAGGTATATTTTGGGAGCATTTTTC	ttgcaa	792822
>CT708_069	TTTTCATTGATTTAGCGGAAGTAAAAAGGTACAAGTAACAGGTCTGTCAA	ttgatt	814833
>CT712_059	TTTTTGCTAAGTTGATCCGTAGATTTAAATAAAATCGTTTTAGAGGGCAA	atcaga	823669
>CT734_092	AATTTGATTGGGATGTTGAAGCGCCTTTATAAACTATTTAGTTAATAAGA	ttgatt	846817
>CT740_073	AAACTTTAAAAAACTCCGTTGATTTTTTATAGAGTAACCTATAACTTGAC	ttgatt	860945
>CT756_160	TAGAGGTTGAATTGATAATGGATATAGAATAATTTTTAAGACTGCTAGTA	ttgaat	888182
>CT757_077	TTTTAACTTGCCTTTTGAAAGCTTAAGTTTAAGATAGAGAATTTCTTATA	ttttaa	889759
>CT763_097	AAATATTGACGCTTTTTTAGAATTTCATATATTCTTCCCACAATCTTGGG	ttgacg	897879
>CT766_105	ATCTGTTGTCTTAGAGATGAAGTTGCTGATATTATTCGTATCGCAGGATT	ttgtct	899207
>CT773_113	TTCTTTTTGCTTTTATTCTTGTCATTGTGAAAAATGTTGAAAAGTTACTT	ttgtca	906537
>CT781_053	GATTGTCGTAAGAAGAAAAATATTGCTACTATTTTTGAGCCAAAGGACGG	ttgtcg	916200
>CT788_112	ATTGTTTTTGTTTCTCGAGAAAAAGGTACTATGATGATCTTTTTTTTAGC	ttgttt	922733
>CT808_121	TTTTTGTAGAATTTTTTACCTAATCGACTTATAATCCGCCTTCTGCTTAA	tagaat	948095
>CT817_085	TCTCTCTTGATGAATAGCATAAGCGTCTGTATCTTAGATGGAATCGAGAA	ttgatg	960444
>CT837_088	AATATTTGAAAGCTAATTCATTTATAAAATAAACTAGAAGACAATCTTGA	ttgaaa	984586
>CT849_066	TTATTTTTATTAAAGAGAGAAATTGCTGGTAAAATAAAAAATAAAAAAAC	tttatt	998698
>CT854_064	ATTCTCTTGCCGCATATGCTCTCTTCCCCTATGATTCTTCCTTCATGAAG	ttgccg	1004522
>CT863_074	TCCAACTTGCATGAAAAATACTTTTTAGATAAGTTCCCTCCTTTCTAAAA	ttgcat	1017779

Appendix G: Alignment of strains L2b, L2 and UW-3

				UW-3		
L2b Num	**Location**	**L2 Num**	**Location**	**Num**	**start**	**end**
CTLon_0001	1..1017	CTL0001	1..1017	CT633	720271	721287
CTLon_0002	1043..2440	CTL0002	1043..2440	CT634	721313	722710
CTLon_0003	2462..2896	CTL0003	2462..2896	CT635	722732	723166
CTLon_0004	3010..5157	CTL0004	3010..5157	CT636	723280	725427
CTLon_0005	5345..6547	CTL0005	5345..6547	CT637	725615	726817
CTLon_0006	6522..7532	CTL0006	6522..7532	CT638	726792	727559
				CT638.1	727592	727801
CTLon_0007	7548..10628	CTL0007	7548..10628	CT639	727817	730897
CTLon_0008	10630..13650	CTL0008	10630..13650	CT640	730899	733913
CTLon_0009	13663..15342	CTL0009	13663..15342	CT641	733926	735605
CTLon_0010	15362..16177	CTL0010	15362..16177	CT642	735625	736440
CTLon_0011	17070..19643	CTL0011	17072..19645	CT643	737335	739908
CTLon_0012	19673..20677	CTL0012	19675..20679	CT644	739938	740942
CTLon_0013	20656..20952	CTL0013	20658..20954	CT645	740921	741217
CTLon_0014	21057..22436	CTL0014	21059..22438	CT646	741318	742697
CTLon_0015	22436..23014	CTL0015	22438..23016	CT647	742697	743275
CTLon_0016	23005..24279	CTL0016	23007..24281	CT648	743266	744540
CTLon_0017	24282..24818	CTL0017	24284..24820	CT649	744543	745079
CTLon_0018	25078..26136	CTL0018	25080..26138	CT650	745340	746398
CTLon_0019	26422..28248	CTL0019	26424..28250	CT651	746684	748510
CTLon_0020	28245..29735	CTL0020	28247..29737	CT652	748507	749997
CTLon_0021	29801..29980	CTL0021	29803..29982	CT652.1	750063	750242
CTLon_0022	30081..30800	CTL0022	30083..30802	CT653	750343	751062
CTLon_0023	30809..31297	CTL0023	30811..31299	CT654	751071	751559
CTLon_0024	31294..32103	CTL0024	31296..32105	CT655	751556	752365
CTLon_0025	32589..32882	CTL0025	32591..32884	CT656	752850	753143
CTLon_0026	32882..33199	CTL0026	32884..33201	CT657	753143	753460
CTLon_0027	33281..34288	CTL0027	33283..34290	CT658	753544	754548
CTLon_0028	34388..34624	CTL0028	34390..34626	CT659	754647	754883
CTLon_0029	34763..36235	CTL0029	34765..36237	CT660	755022	756494
CTLon_0030	36250..38067	CTL0030	36252..38069	CT661	756509	758326
CTLon_0031	38447..39454	CTL0031	38449..39456	CT662	758706	759713
CTLon_0032	40173..40574	CTL0032	40168..40569	CT663	760425	760826
CTLon_0033	40578..43067	CTL0033	40573..43062	CT664	760830	763319
CTLon_0034	43111..43362	CTL0034	43106..43357	CT665	763363	763614
CTLon_0035	43389..43640	CTL0035	43384..43635	CT666	763641	763892
CTLon_0036	43659..44108	CTL0036	43654..44103	CT667	763911	764360
CTLon_0037	44128..44799	CTL0037	44123..44794	CT668	764380	765051
CTLon_0038	44801..46129	CTL0038	44796..46124	CT669	765053	766381
CTLon_0039	46152..46658	CTL0039	46147..46653	CT670	766404	766910
CTLon_0040	46662..47513	CTL0040	46657..47508	CT671	766914	767765
CTLon_0041	47523..48644	CTL0041	47518..48639	CT672	767775	768896
CTLon_0042	48779..50251	CTL0042	48774..50246	CT673	769031	770503
CTLon_0043	50248..53013	CTL0043	50243..53008	CT674	770500	773265
CTLon_0044	53327..54397	CTL0044	53322..54392	CT675	773578	774648
CTLon_0045	54387..54908	CTL0045	54382..54903	CT676	774638	775159
CTLon_0046	55280..55819	CTL0046	55274..55813	CT677	775530	776069
CTLon_0047	55816..56553	CTL0047	55810..56547	CT678	776066	776803
CTLon_0048	56566..57414	CTL0048	56560..57408	CT679	776816	777664
CTLon_0049	57411..58259	CTL0049	57405..58253	CT680	777661	778509
CTLon_0050	58643..59827	CTL0050	58637..59821	CT681	778879	780060
CTLon_0051	60435..63677	CTL0051	60428..63670	CT682	780668	783910
CTLon_0052	63741..64748	CTL0052	63734..64741	CT683	783974	784981

CTLon_0053	65090..66541	CTL0053	65083..66534	CT684	785324	786775
CTLon_0054	66544..67311	CTL0054	66537..67304	CT685	786778	787545
CTLon_0055	67315..68502	CTL0055	67308..68495	CT686	787549	788736
CTLon_0056	68495..69700	CTL0056	68488..69693	CT687	788729	789934
CTLon_0057	69836..70681	CTL0057	69829..70674	CT688	790070	790915
CTLon_0058	70673..71503	CTL0058	70666..71496	CT689	790907	791737
CTLon_0059	71496..72461	CTL0059	71489..72454	CT690	791730	792695
CTLon_0060	72622..73296	CTL0060	72615..73289	CT691	792856	793530
CTLon_0061	73299..74579	CTL0061	73292..74572	CT692	793533	794813
CTLon_0062	74707..75918	CTL0062	74711..75922	CT693	794941	796152
CTLon_0063	76178..77146	CTL0063	76182..77150	CT694	796412	797383
CTLon_0064	77197..78393	CTL0064	77201..78397	CT695	797434	798630
CTLon_0065	78479..79657	CTL0065	78483..79661	CT696	798716	799894
CTLon_0066	79654..80289	CTL0066	79658..80293	CT697	799891	800526
CTLon_0067	80296..81630	CTL0067	80300..81634	CT698	800533	801867
CTLon_0068	81898..82803	CTL0068	81902..82807	CT699	802134	803039
CTLon_0069	82869..84194	CTL0069	82872..84197	CT700	803104	804429
CTLon_0070	84453..87362	CTL0070	84456..87365	CT701	804688	807597
CTLon_0071	87456..87983	CTL0071	87459..87986	CT702	807691	808218
CTLon_0072	88060..89532	CTL0072	88063..89535	CT703	808295	809767
CTLon_0073	89648..90880	CTL0073	89651..90883	CT704	809883	811115
CTLon_0074	90895..92154	CTL0074	90898..92157	CT705	811130	812389
CTLon_0075	92164..92775	CTL0075	92167..92778	CT706	812399	813010
CTLon_0076	92942..94270	CTL0076	92945..94273	CT707	813177	814505
CTLon_0077	94629..98120	CTL0077	94632..98123	CT708	814862	818353
CTLon_0078	98125..99225	CTL0078	98128..99228	CT709	818358	819458
CTLon_0079	99222..101021	CTL0079	99225..101024	CT710	819455	821254
CTLon_0080	101133..103436	CTL0080	101136..103439	CT711	821366	823669
CTLon_0081	103463..104635	CTL0081	103466..104638	CT712	823696	824868
CTLon_0082	104701..105723	CTL0082	104704..105726	CT713	824894	825916
CTLon_0083	105857..106861	CTL0083	105860..106864	CT714	826049	827053
CTLon_0084	106858..108225	CTL0084	106861..108228	CT715	827050	828417
CTLon_0085	108237..108602	CTL0085	108240..108605	CT716	828429	828794
CTLon_0086	108595..109899	CTL0086	108598..109902	CT717	828787	830091
CTLon_0087	109973..110497	CTL0087	109976..110500	CT718	830165	830689
CTLon_0088	110502..111506	CTL0088	110505..111509	CT719	830694	831698
CTLon_0089	111780..112562	CTL0089	111783..112565	CT720	831971	832753
CTLon_0090	112559..113713	CTL0090	112562..113716	CT721	832750	833904
CTLon_0091	113668..114348	CTL0091	113671..114351	CT722	833859	834539
CTLon_0092	114643..115368	CTL0092	114646..115371	CT723	834836	835561
CTLon_0093	115487..116011	CTL0093	115490..116014	CT724	835680	836204
CTLon_0094	116038..116592	CTL0094	116041..116595	CT725	836231	836785
CTLon_0095	116599..117738	CTL0095	116602..117741	CT726	836792	837931
CTLon_0096	117830..119809	CTL0096	117833..119812	CT727	838023	840002
CTLon_0097	119833..120579	CTL0097	119836..120582	CT728	840026	840772
CTLon_0098	120616..121902	CTL0098	120619..121905	CT729	840809	842095
CTLon_0099	121944..123071	CTL0099	121947..123074	CT730	842137	843264
CTLon_0100	123181..124455	CTL0100	123184..124458	CT731	843374	844648
CTLon_0101	124424..124897	CTL0101	124427..124900	CT732	844617	845090
CTLon_0102	124948..126294	CTL0102	124951..126297	CT733	845141	846487
CTLon_0103	126679..127344	CTL0103	126682..127347	CT734	846872	847537
CTLon_0104	127461..128828	CTL0104	127452..128819	CT735	847642	849009
CTLon_0105	128886..129338	CTL0105	128877..129329	CT736	849067	849519
CTLon_0106	129347..130006	CTL0106	129338..129997	CT737	849528	850187
CTLon_0107	130003..130791	CTL0107	129994..130782	CT738	850184	850972
CTLon_0108	130795..133194	CTL0108	130786..133185	CT739	850976	853375
CTLon_0109	139465..140760	CTL0109	139457..140752	CT740	859615	860910

CTLon_0110	140903..141247	CTL0110	140895..141239	CT741	861053	861397
CTLon_0111	141352..142542	CTL0111	141344..142534	CT742	861502	862692
CTLon_0112	142730..143107	CTL0112	142722..143099	CT743	862880	863257
CTLon_0113	143503..145971	CTL0113	143495..145963	CT744	863653	866118
CTLon_0114	145963..147237	CTL0114	145955..147229	CT745	866113	867387
CTLon_0115	147234..148607	CTL0115	147226..148599	CT746	867384	868757
CTLon_0116	148580..149590	CTL0116	148572..149582	CT747	868730	869740
CTLon_0117	149603..152842	CTL0117	149595..152834	CT748	869753	872992
CTLon_0118	152818..155445	CTL0118	152810..155437	CT749	872968	875595
CTLon_0119	161270..163270	CTL0119	161262..163262	CT750	881422	883422
CTLon_0120	163267..164136	CTL0120	163259..164128	CT751	883419	884288
CTLon_0121	164392..164928	CTL0121	164384..164920	CT752	884508	885080
CTLon_0122	164944..165168	CTL0122	164936..165160	CT753	885096	885320
CTLon_0123	165215..166087	CTL0123	165207..166079	CT754	885367	886239
CTLon_0124	166168..167706	CTL0124	166160..167698	CT755	886320	887858
CTLon_0125	168154..169506	CTL0125	168146..169498	CT756	888306	889658
CTLon_0126	169652..170662	CTL0126	169644..170654	CT757	889804	890814
CTLon_0127	170674..171924	CTL0127	170666..171916	CT758	890826	892076
CTLon_0128	171921..172658	CTL0128	171913..172650	CT759	892073	892810
CTLon_0129	172674..173831	CTL0129	172666..173823	CT760	892826	893983
CTLon_0130	173740..174798	CTL0130	173732..174790	CT761	893892	894950
CTLon_0131	174803..177214	CTL0131	174795..177206	CT762	894955	897366
CTLon_0132	177251..177670	CTL0132	177243..177662	CT763	897403	897822
CTLon_0133	177828..178634	CTL0133	177820..178626	CT764	897980	898786
CTLon_0134	178788..179120	CTL0134	178780..179112	CT765	898940	899272
CTLon_0135	179124..180143	CTL0135	179116..180135	CT766	899276	900295
CTLon_0136	180148..181200	CTL0136	180140..181192	CT767	900300	901352
CTLon_0137	181290..182972	CTL0137	181282..182964	CT768	901442	903130
CTLon_0138	183426..183785	CTL0138	183418..183777	CT769	903584	903943
CTLon_0139	183796..185052	CTL0139	183788..185044	CT770	903954	905210
CTLon_0140	185049..185501	CTL0140	185041..185493	CT771	905207	905659
CTLon_0141	185582..186211	CTL0141	185574..186203	CT772	905740	906369
CTLon_0142	186459..187499	CTL0142	186451..187491	CT773	906617	907657
CTLon_0143	187465..188493	CTL0143	187457..188485	CT774	907623	908651
CTLon_0144	188791..189552	CTL0144	188783..189544	CT775	908950	909711
CTLon_0145	189560..191173	CTL0145	189552..191165	CT776	909719	911332
CTLon_0146	191216..192349	CTL0146	191208..192341	CT777	911375	912508
CTLon_0147	192268..194529	CTL0147	192260..194521	CT778	912427	914688
CTLon_0148	194484..195173	CTL0148	194476..195165	CT779	914643	915332
CTLon_0149	195329..195823	CTL0149	195321..195815	CT780	915488	915982
CTLon_0150	196055..197635	CTL0150	196047..197627	CT781	916214	917794
CTLon_0151	197632..199125	CTL0151	197624..199117	CT782	917791	919284
CTLon_0152	199378..200427	CTL0152	199370..200419	CT783	919539	920588
CTLon_0153	200505..200867	CTL0153	200497..200859	CT784	920666	921028
CTLon_0154	200880..201017	CTL0153A	200872..201009	CT785	921041	921178
CTLon_0155	201494..201631	CTL0154	201486..201623	CT786	921655	921792
CTLon_0156	201655..201960	CTL0155	201647..201952	CT787	921816	922121
CTLon_0157	201994..202494	CTL0156	201986..202486	CT788	922155	922655
CTLon_0158	202616..202867	CTL0157	202608..202859	CT789	922777	923028
CTLon_0159	203147..203641	CTL0158	203139..203633	CT790	923307	923801
CTLon_0160	203688..205427	CTL0159	203679..205418	CT791	923846	925642
CTLon_0161	205509..207971	CTL0160	205500..207962	CT792	925667	928129
				CT793	928192	928461
CTLon_0163	208304..210091	CTL0162	208295..210082	CT794	928463	930250
CTLon_0164	210176..210463	CTL0163	210167..210454	CT794.1	930335	930622
CTLon_0165	210590..211081	CTL0164	210580..211071	CT795	930749	931240
CTLon_0166	211164..214175	CTL0165	211154..214165	CT796	931322	934333

CTLon_0167	214629..215237	CTL0166	214619..215227	CT797	934794	935402
CTLon_0168	215244..216668	CTL0167	215234..216658	CT798	935409	936833
CTLon_0169	217245..217802	CTL0168	217235..217792	CT799	937410	937967
CTLon_0170	217810..218349	CTL0169	217800..218339	CT800	937975	938514
CTLon_0171	218487..218825	CTL0170	218477..218815	CT801	938652	938990
CTLon_0172	218842..219087	CTL0171	218832..219077	CT802	939007	939252
CTLon_0173	219108..219611	CTL0172	219098..219601	CT803	939273	939776
CTLon_0174	219675..220541	CTL0173	219665..220531	CT804	939840	940706
CTLon_0175	220844..222196	CTL0174	220834..222186	CT805	941009	942361
CTLon_0176	222375..225245	CTL0175	222365..225235	CT806	942540	945410
CTLon_0177	225297..226298	CTL0176	225287..226288	CT807	945467	946462
CTLon_0178	226311..227849	CTL0177	226301..227839	CT808	946475	948013
CTLon_0179	228413..228733	CTL0178	228403..228723	CT809	948577	948888
CTLon_0180	228673..228999	CTL0179	228663..228989			
CTLon_0181	228983..229162	CTL0181	228973..229152	CT810	949148	949327
CTLon_0182	229184..230149	CTL0182	229174..230139	CT811	949350	950315
CTLon_0183	230369..234961	CTL0183	230359..234951	CT812	950536	955131
CTLon_0184	235162..235956	CTL0184	235152..235946	CT813	955374	956168
CTLon_0185	236040..236339	CTL0185	236030..236329	CT814	956252	956653
CTLon_0186	236360..236722	CTL0186	236350..236712	CT814.1	956572	956934
CTLon_0187	236967..238343	CTL0187	236957..238333	CT815	957180	958556
CTLon_0188	238354..240174	CTL0188	238344..240164	CT816	958567	960387
CTLon_0189	240277..241473	CTL0189	240267..241463	CT817	960490	961686
CTLon_0190	241656..242849	CTL0190	241646..242839	CT818	961868	963061
CTLon_0191	242891..243607	CTL0191	242881..243597	CT819	963103	963819
CTLon_0192	243704..244558	CTL0192	243695..244549	CT820	963917	964771
CTLon_0193	244628..245788	CTL0193	244619..245779	CT821	964841	966001
CTLon_0194	245803..246678	CTL0194	245794..246669	CT822	966016	966891
CTLon_0195	246818..248311	CTL0195	246808..248301	CT823	967030	968523
CTLon_0196	248513..251437	CTL0196	248502..251426	CT824	968723	971647
CTLon_0197	251451..252734	CTL0197	251440..252723	CT825	971661	972944
CTLon_0198	252738..253529	CTL0198	252727..253518	CT826	972948	973739
CTLon_0199	254164..257307	CTL0199	254153..257296	CT827	974375	977518
CTLon_0200	257345..258385	CTL0200	257334..258374	CT828	977556	978596
CTLon_0201	258645..259319	CTL0201	258634..259308	CT829	978967	979530
CTLon_0202	259435..260019	CTL0202	259424..260008	CT830	979646	980230
CTLon_0203	260016..260891	CTL0203	260005..260880	CT831	980227	981102
CTLon_0204	261024..261530	CTL0204	261013..261519	CT832	981235	981741
CTLon_0205	261962..262522	CTL0205	261951..262511	CT833	982172	982699
CTLon_0206	262509..262694	CTL0206	262498..262683	CT834	982710	982904
CTLon_0207	262713..263084	CTL0207	262702..263073	CT835	982923	983294
CTLon_0208	263091..264119	CTL0208	263080..264108	CT836	983301	984329
CTLon_0209	264430..266406	CTL0209	264419..266395	CT837	984639	986615
CTLon_0210	266403..267503	CTL0210	266392..267492	CT838	986612	987712
CTLon_0211	267506..268570	CTL0211	267495..268559	CT839	987715	988779
CTLon_0212	268668..269633	CTL0212	268657..269622	CT840	988877	989842
CTLon_0213	269794..272535	CTL0213	269783..272524	CT841	990003	992744
CTLon_0214	272750..274837	CTL0214	272739..274826	CT842	992959	995046
CTLon_0215	274866..275135	CTL0215	274855..275124	CT843	995075	995344
CTLon_0216	275361..275852	CTL0216	275350..275841	CT844	995570	996061
CTLon_0217	275884..276102	CTL0217	275873..276091	CT845	996033	996311
CTLon_0218	276074..276778	CTL0218	276063..276767	CT846	996283	996987
CTLon_0219	276913..277431	CTL0219	276902..277420	CT847	997122	997640
CTLon_0220	277444..277950	CTL0220	277433..277939	CT848	997656	998162
CTLon_0221	277978..278457	CTL0221	277967..278446	CT849	998190	998669
CTLon_0222	278774..278962	CTL0222	278763..278951	CT849.1	998988	999176
CTLon_0223	278981..280198	CTL0223	278970..280187	CT850	999195	1000412

CTLon_0224	280155..281030	CTL0224	280144..281019	CT851	1000369	1001244
CTLon_0225	281063..281704	CTL0225	281052..281693	CT852	1001302	1001916
CTLon_0226	281721..282320	CTL0226	281710..282309	CT853	1001934	1002533
CTLon_0227	282515..284284	CTL0227	282504..284273	CT854	1002728	1004497
CTLon_0229	285811..287514	CTL0231	285800..287503	CT855	1004550	1005941
CTLon_0230	287745..287906	CTL0231A	287734..287895	CT856	1006022	1007725
CTLon_0231	287946..288155	CTL0231B	287935..288144			
CTLon_0232	288269..289534	CTL0232	288258..289523	CT857	1008482	1009747
CTLon_0233	289644..291449	CTL0233	289633..291438	CT858	1009861	1011666
CTLon_0234	291544..292467	CTL0234	291533..292456	CT859	1011761	1012684
CTLon_0235	292519..294000	CTL0235	292508..293989	CT860	1012736	1014217
CTLon_0236	294012..295532	CTL0236	294001..295521	CT861	1014229	1015749
CTLon_0237	295560..296156	CTL0237	295549..296145	CT862	1015776	1016372
CTLon_0238	296078..297526	CTL0238	296067..297515	CT863	1016294	1017742
CTLon_0239	298387..299289	CTL0243	298375..299277	CT864	1018603	1019505
CTLon_0240	299537..300526	CTL0244	299525..300514	CT865	1019754	1020743
CTLon_0241	300541..302757	CTL0245	300529..302745	CT866	1020758	1022974
CTLon_0242	302768..303787	CTL0246	302756..303775	CT867	1022985	1024004
CTLon_0243	304014..305225	CTL0247	303998..305203	CT868	1024215	1025471
CTLon_0244	305402..308299	CTL0248	305380..308277	CT869	1025648	1028542
CTLon_0245	308302..311400	CTL0249	308280..311378	CT870	1028545	1031649
CTLon_0246	311565..314606	CTL0250	311543..314581	CT871	1031814	1034855
CTLon_0247	314637..317666	CTL0251	314612..317632	CT872	1034886	1037936
CTLon_0248	318081..318677	CTL0252	318047..318643	CT873	1038518	1038835
CTLon_0249	318772..321408	CTL0254	318738..321374	CT874	1039010	1041646
CTLon_0250	321683..323455	CTL0255	321649..323421	CT875	1041920	1176
CTLon_0251	323600..323872	CTL0256	323566..323838	CT001	1321	1593
CTLon_0252	324073..324375	CTL0257	324039..324341	CT002	1794	2096
CTLon_0253	324387..325862	CTL0258	324353..325828	CT003	2108	3583
CTLon_0254	325864..327330	CTL0259	325830..327296	CT004	3585	5051
CTLon_0255	327429..328520	CTL0260	327395..328486	CT005	5150	6241
CTLon_0256	328648..329217	CTL0261	328614..329183	CT006	6369	6938
CTLon_0257	329531..330481	CTL0262	329497..330447	CT007	7251	8201
CTLon_0258	330497..331399	CTL0263	330463..331365	CT008	8217	9119
CTLon_0259	331655..332086	CTL0264	331621..332052	CT009	9373	9804
CTLon_0260	332073..333440	CTL0265	332039..333406	CT010	9791	11158
CTLon_0261	333572..334828	CTL0266	333538..334794	CT011	11290	12546
CTLon_0262	334825..335619	CTL0267	334791..335585	CT012	12543	13337
CTLon_0263	335894..337234	CTL0268	335860..337200	CT013	13612	14952
CTLon_0264	337246..338307	CTL0269	337212..338273	CT014	14964	16025
CTLon_0265	338405..339709	CTL0270	338371..339675	CT015	16123	17427
CTLon_0266	339918..340646	CTL0271	339884..340612	CT016	17636	18364
CTLon_0267	340832..342133	CTL0272	340798..342099	CT017	18550	19851
CTLon_0268	342195..342668	CTL0273	342161..342634	CT018	19913	20386
CTLon_0269	343714..346824	CTL0274	343680..346790	CT019	21432	24542
CTLon_0270	346883..348769	CTL0275	346848..348734	CT020	24601	26487
CTLon_0271	348955..349698	CTL0276	348920..349663	CT021	26673	27416
CTLon_0272	349774..350100	CTL0277	349739..350065	CT022	27492	27818
CTLon_0273	350288..351367	CTL0278	350253..351332	CT023	28006	29085
CTLon_0274	351351..352223	CTL0279	351316..352188	CT024	29069	29941
CTLon_0275	352220..353566	CTL0280	352185..353531	CT025	29938	31284
CTLon_0276	353557..353907	CTL0281	353522..353872	CT026	31275	31625
CTLon_0277	353923..354981	CTL0282	353888..354946	CT027	31641	32699
CTLon_0278	355007..355372	CTL0283	354972..355337	CT028	32725	33090
CTLon_0279	355436..356089	CTL0284	355401..356054	CT029	33154	33807
CTLon_0280	356080..356697	CTL0285	356045..356662	CT030	33798	34415
CTLon_0281	356690..356992	CTL0286	356655..356957	CT031	34408	34710

CTLon_0282	356992..358644	CTL0287	356957..358609	CT032	34710	36362
CTLon_0283	358783..361023	CTL0288	358748..360988	CT033	36502	38742
CTLon_0284	361271..362251	CTL0289	361236..362216	CT034	38990	40015
CTLon_0285	362633..363406	CTL0290	362599..363372	CT035	40354	41127
CTLon_0286	363371..364528	CTL0291	363337..364494	CT036	41092	42303
	365074..365418			CT037	42310	42666
CTLon_0288	365440..366012	CTL0293	365040..365384	CT038	42865	43215
CTLon_0289	366099..366251	CTL0294	365406..365978	CT039	43231	43803
CTLon_0290	366293..367297	CTL0295	366065..366217	CT039.1	43891	44043
CTLon_0291	367299..368111	CTL0296	366259..367263	CT040	44085	45089
CTLon_0292	368173..370173	CTL0297	367265..368077	CT041	45091	45903
CTLon_0293	370290..370793	CTL0298	368138..370138	CT042	45965	47965
CTLon_0294	371259..371732	CTL0299	370259..370762	CT043	48083	48586
CTLon_0295	372075..373574	CTL0300	371228..371701	CT044	49052	49525
CTLon_0296	373712..374383	CTL0301	372044..373543	CT045	49866	51365
CTLon_0297	374538..375482	CTL0302	373681..374352	CT046	51504	52115
CTLon_0298	375577..376290	CTL0303	374507..375451	CT047	52270	53214
CTLon_0299	376381..377853	CTL0304	375546..376259	CT048	53309	54022
CTLon_0300	377912..379579	CTL0305	376350..377822	CT049	54113	55585
CTLon_0301	379615..381312	CTL0306	377881..379548	CT050	55644	57254
CTLon_0302	381374..382504	CTL0307	379584..381275	CT051	57358	58920
CTLon_0303	382494..382940	CTL0308	381336..382466	CT052	58974	60110
CTLon_0304	383272..385983	CTL0309	382456..382902	CT053	60100	60546
CTLon_0305	385987..387084	CTL0310	383234..385945	CT054	60878	63595
CTLon_0306	387113..387844	CTL0311	385949..387046	CT055	63599	64696
CTLon_0307	387853..389661	CTL0312	387075..387806	CT056	64725	65456
CTLon_0308	389813..390904	CTL0313	387815..389623	CT057	65465	67273
CTLon_0309	390984..391259	CTL0314	389775..390866	CT058	67425	68528
CTLon_0310	391441..393258	CTL0315	390946..391221	CT059	68608	68883
CTLon_0311	393366..394127	CTL0316	391403..393220	CT060	69065	70882
CTLon_0312	394286..395524	CTL0317	393328..394089	CT061	70990	71751
CTLon_0313	395534..396976	CTL0318	394248..395486	CT062	71910	73148
CTLon_0314	397037..398845	CTL0319	395496..396938	CT063	73158	74600
CTLon_0315	399031..400617	CTL0320	396999..398807	CT064	74661	76469
CTLon_0316	400970..401446	CTL0321	398993..400579	CT065	76655	78241
CTLon_0317	402107..403087	CTL0322	400930..401406	CT066	78592	79068
CTLon_0318	403080..403859	CTL0323	402067..403047	CT067	79726	80706
CTLon_0319	403860..405215	CTL0324	403040..403819	CT068	80699	81478
CTLon_0320	405205..406161	CTL0325	403820..405175	CT069	81479	82834
CTLon_0321	406188..407327	CTL0326	405165..406121	CT070	82824	83780
CTLon_0322	407523..409382	CTL0327	406148..407287	CT071	83807	84946
CTLon_0323	409329..410306	CTL0328	407483..409342	CT072	85142	87001
CTLon_0324	410464..411561	CTL0329	409289..410266	CT073	86948	87925
CTLon_0325	411561..412811	CTL0330	410424..411521	CT074	88083	89180
CTLon_0326	412941..413396	CTL0331	411521..412771	CT075	89180	90430
CTLon_0327	413374..414324	CTL0332	412901..413356	CT076	90560	91015
CTLon_0328	414294..415157	CTL0333	413334..414284	CT077	90993	91943
CTLon_0329	415275..415670	CTL0334	414254..415117	CT078	91913	92776
CTLon_0330	415995..416288	CTL0335	415235..415630	CT079	92894	93337
CTLon_0331	416321..416641	CTL0336	415955..416248	CT080	93614	93907
CTLon_0332	416651..418333	CTL0337	416281..416601	CT081	93964	94260
CTLon_0333	418330..418812	CTL0338	416611..418293	CT082	94270	95952
CTLon_0334	418826..419911	CTL0338A	418290..418772	CT083	95949	96431
CTLon_0335	420083..421822	CTL0339	418786..419871	CT084	96445	97530
CTLon_0336	421942..422211	CTL0340	420042..421781	CT085	97700	99439
CTLon_0337	422360..423943	CTL0341	421901..422170	CT086	99565	99834
CTLon_0338	423947..424387	CTL0342	422319..423902	CT087	99983	101566

CTLon_0339	424425..425690	CTL0343	423906..424346	CT088	101570	102010
CTLon_0340	425708..427834	CTL0344	424384..425649	CT089	102048	103313
CTLon_0341	427834..428916	CTL0345	425667..427793	CT090	103331	105457
CTLon_0342	429173..430273	CTL0346	427793..428875	CT091	105457	106539
CTLon_0343	430282..431187	CTL0347	429132..430232	CT092	106796	107896
CTLon_0344	431144..431869	CTL0348	430241..431146	CT093	107905	108810
CTLon_0345	431907..432278	CTL0349	431103..431828	CT094	108767	109492
CTLon_0346	432285..434975	CTL0350	431866..432237	CT095	109530	109901
CTLon_0347	434932..436236	CTL0351	432244..434934	CT096	109908	112586
CTLon_0348	436354..438063	CTL0352	434891..436195	CT097	112543	113847
CTLon_0349	438308..439363	CTL0353	436313..438022	CT098	113965	115674
CTLon_0350	439369..439728	CTL0354	438267..439322	CT099	115919	116974
CTLon_0351	439729..440190	CTL0355	439328..439687	CT100	116980	117339
CTLon_0352	440326..440787	CTL0356	439688..440149	CT101	117340	117801
CTLon_0353	440822..441718	CTL0357	440285..440746	CT102	117936	118397
CTLon_0354	441708..442604	CTL0358	440781..441677	CT103	118432	119328
CTLon_0355	442916..444886	CTL0359	441667..442563	CT104	119318	120214
CTLon_0356	445086..445910	CTL0360	442875..444845	CT105	120452	122422
CTLon_0357	445957..447063	CTL0361	445045..445869	CT106	122535	123446
CTLon_0358	447117..447872	CTL0362	445916..447022	CT107	123493	124602
CTLon_0359	447885..448673	CTL0363	447076..447831	CT108	124652	125407
CTLon_0360	448802..450436	CTL0364	447844..448632	CT109	125420	126208
CTLon_0361	450475..450783	CTL0365	448761..450395	CT110	126336	127970
CTLon_0362	450955..452781	CTL0366	450434..450742	CT111	128008	128316
CTLon_0363	453083..455686	CTL0367	450914..452740	CT112	128488	130314
CTLon_0364	455712..457172	CTL0368	453042..455645	CT113	130617	133220
CTLon_0365	457358..457840	CTL0369	455671..457131	CT114	133246	134706
CTLon_0366	457907..458305	CTL0370	457317..457799	CT115	134936	135361
CTLon_0367	458310..458624	CTL0371	457866..458264	CT116	135444	135842
CTLon_0368	458711..459214	CTL0372	458269..458583	CT117	135847	136161
CTLon_0369	459380..460120	CTL0373	458670..459173	CT118	136248	136751
CTLon_0370	460247..460522	CTL0374	459340..460080	CT119	136917	137738
CTLon_0371	460621..461307	CTL0375	460207..460482	CT120	137816	138058
CTLon_0372	461294..461851	CTL0376	460581..461267	CT121	138142	138843
CTLon_0373	461864..462358	CTL0377	461254..461811	CT122	138830	139387
CTLon_0374	462362..463735	CTL0378	461824..462318	CT123	139400	139894
CTLon_0375	463958..464410	CTL0379	462322..463695	CT124	139898	141271
CTLon_0376	464436..464825	CTL0380	463918..464370	CT125	141494	141946
CTLon_0377	464873..465724	CTL0381	464396..464785	CT126	141972	142361
CTLon_0378	465820..466557	CTL0382	464833..465684	CT127	142409	143260
CTLon_0379	466764..467408	CTL0383	465780..466517	CT128	143356	144093
CTLon_0380	467405..468106	CTL0384	466724..467368	CT129	144300	144944
CTLon_0381	468109..471525	CTL0385	467365..468066	CT130	144941	145642
CTLon_0382	471522..472799	CTL0386	468069..471485	CT131	145645	149061
CTLon_0383	472877..473680	CTL0387	471482..472759	CT132	149058	150335
CTLon_0384	474137..474550	CTL0388	472838..473641	CT133	150413	151216
CTLon_0385	474608..475690	CTL0389	474097..474510	CT134	151671	152084
CTLon_0386	475780..476499	CTL0390	474568..475650	CT135	152143	153225
CTLon_0387	476518..477363	CTL0391	475740..476459	CT136	153315	154034
CTLon_0388	477341..478288	CTL0392	476478..477323	CT137	154053	154898
CTLon_0389	478285..479565	CTL0393	477301..478248	CT138	154876	155823
CTLon_0390	479741..480427	CTL0394	478245..479525	CT139	155820	157100
CTLon_0391	480611..481057	CTL0395	479701..480387	CT140	157277	157963
CTLon_0392	481457..482314	CTL0396	480571..481017	CT141	158147	158593
CTLon_0393	482316..483158	CTL0397	481417..482274	CT142	158993	159850
CTLon_0394	483158..484024	CTL0398	482276..483118	CT143	159852	160694
CTLon_0395	484210..486054	CTL0399	483118..483984	CT144	160694	161551

CTLon_0396	486065..488056	CTL0400	484170..486014	CT145	161737	163581
CTLon_0397	488154..492503	CTL0401	486025..488016	CT146	163592	165583
CTLon_0398	492566..494089	CTL0402	488114..492463	CT147	165681	170030
CTLon_0399	494289..495236	CTL0403	492526..494049	CT148	170093	171616
CTLon_0400	495635..495793	CTL0404	494249..495196	CT149	171816	172763
CTLon_0401	495829..497340	CTL0405	495595..495753	CT150	173162	173320
CTLon_0402	497345..498022	CTL0406	495789..497300	CT151	173356	174867
CTLon_0403	498173..500605	CTL0407	497305..497982	CT152	174874	175551
CTLon_0404	502342..503280	CTL0408	498133..500565	CT153	175702	178134
				CT154	178320	179471
CTLon_0406	503794..504996	CTL0411	502302..503240	CT155	179440	180381
				CT156	180261	180602
CTLon_0407	507345..508091	CTL0413	503754..504959	CT157	180738	181952
				CT158	182076	182792
				CT159	182735	183667
				CT160	183813	184316
CTLon_0410	508703..510232	CTL0417	507306..508052	CT161	184313	185053
				CT162	185201	185440
CTLon_0412	510285..510545	CTL0419	508663..510192	CT163	185673	187319
CTLon_0413	511448..511732	CTL0419A	510245..510505	CT164	187372	187632
				CT165	188106	188552
				CT166	188549	190468
				CT167	190524	191855
				CT168	191921	192223
CTLon_0416	512081..513259	CTL0422	511407..511691	CT169	192279	192563
CTLon_0417	513252..514013	CTL0423	512040..513218	CT170	192912	194090
CTLon_0418	514204..514725	CTL0424	513211..513972	CT171	194083	194844
CTLon_0419	514765..514956	CTL0425	514163..514684	CT172	195091	195582
CTLon_0420	515333..516922	CTL0426	514761..514916	CT172.1	195610	195780
				CT173	195848	196120
				CT174	196207	196662
CTLon_0422	516949..517356	CTL0427	515293..516882	CT175	196857	198446
CTLon_0423	517353..518069	CTL0428	516909..517316	CT176	198473	198880
CTLon_0424	518215..519429	CTL0429	517313..518029	CT177	198877	199593
CTLon_0425	519431..519943	CTL0430	518175..519389	CT178	199739	200953
CTLon_0426	520087..520779	CTL0431	519391..519903	CT179	200955	201467
CTLon_0427	521087..521797	CTL0432	520047..520739	CT180	201611	202303
CTLon_0428	522165..522929	CTL0433	521047..521757	CT181	202611	203321
CTLon_0429	522905..524524	CTL0434	522125..522889	CT182	203689	204453
CTLon_0430	524511..524957	CTL0435	522865..524484	CT183	204429	206048
CTLon_0431	525074..526597	CTL0436	524471..524917	CT184	206035	206481
CTLon_0432	526622..527392	CTL0437	525034..526557	CT185	206802	208121
CTLon_0433	527389..528261	CTL0438	526582..527352	CT186	208146	208916
CTLon_0434	528275..528886	CTL0439	527349..528221	CT187	208913	209785
CTLon_0435	528888..531398	CTL0440	528235..528846	CT188	209799	210410
CTLon_0436	531413..533827	CTL0441	528848..531358	CT189	210412	212922
CTLon_0437	533830..534180	CTL0442	531373..533787	CT190	212937	215351
CTLon_0438	534327..535022	CTL0443	533790..534140	CT191	215354	215704
CTLon_0439	535126..536244	CTL0444	534287..534982	CT192	215851	216624
CTLon_0440	536371..537783	CTL0445	535086..536204	CT193	216651	217769
CTLon_0441	538223..539314	CTL0446	536331..537743	CT194	217896	219308
CTLon_0442	539540..539860	CTL0447	538183..539274	CT195	219747	220838
CTLon_0443	539936..540952	CTL0448	539500..539820	CT196	221062	221382
CTLon_0444	540916..542472	CTL0449	539896..540912	CT197	221458	222474
CTLon_0445	542631..543572	CTL0450	540876..542432	CT198	222438	223994
CTLon_0446	543604..544449	CTL0451	542591..543532	CT199	224153	225094
CTLon_0447	544442..545275	CTL0452	543564..544409	CT200	225126	225971

CTLon_0448	545290..546033	CTL0453	544402..545235	CT201	225964	226797
CTLon_0449	546345..547094	CTL0454	545250..545993	CT202	226812	227555
CTLon_0450	547124..548539	CTL0455	546305..547054	CT203	227862	228617
CTLon_0451	548559..550220	CTL0456	547084..548499	CT204	228647	230062
CTLon_0452	550229..551077	CTL0457	548519..550180	CT205	230082	231743
CTLon_0453	551059..552705	CTL0458	550189..551037	CT206	231752	232600
CTLon_0454	552951..554246	CTL0459	551019..552665	CT207	232582	234228
CTLon_0455	554243..556702	CTL0460	552911..554206	CT208	234474	235769
CTLon_0456	556900..558168	CTL0461	554203..556662	CT209	235766	238225
CTLon_0457	558160..558729	CTL0462	556859..558127	CT210	238422	239690
CTLon_0458	558749..559195	CTL0463	558119..558688	CT211	239682	240251
CTLon_0459	559185..559913	CTL0464	558708..559154	CT212	240271	240717
CTLon_0460	560008..561651	CTL0465	559144..559872	CT213	240707	241435
CTLon_0461	561955..563001	CTL0466	559967..561610	CT214	241530	243173
CTLon_0462	563015..564415	CTL0467	561914..562960	CT215	243477	244523
CTLon_0463	564509..565273	CTL0468	562974..564374	CT216	244537	245937
CTLon_0464	565404..566255	CTL0469	564468..565232	CT217	246091	246795
CTLon_0465	566367..567275	CTL0470	565363..566214	CT218	246926	247777
CTLon_0466	567272..567850	CTL0471	566326..567234	CT219	247889	248797
CTLon_0467	567847..568743	CTL0472	567231..567809	CT220	248794	249372
CTLon_0468	569033..569173	CTL0473	567806..568702	CT221	249369	250265
CTLon_0469	569200..569586	CTL0474	568992..569132	CT221.1	250554	250694
CTLon_0470	569728..570534	CTL0475	569159..569545	CT222	250721	251110
CTLon_0471	570926..571369	CTL0476	569687..570493	CT223	251252	252064
CTLon_0472	571477..571845	CTL0477	570885..571328	CT224	252455	252898
CTLon_0473	571944..572459	CTL0477A	571436..571804	CT225	252989	253357
CTLon_0474	572588..572989	CTL0478	571903..572418	CT226	253456	253986
CTLon_0475	573270..573860	CTL0479	572547..572948	CT227	254102	254503
CTLon_0476	574010..574657	CTL0480	573229..573819	CT228	254774	255364
CTLon_0477	574883..576130	CTL0481	573969..574616	CT229	255514	256161
CTLon_0478	576314..577786	CTL0482	574842..576089	CT230	256387	257634
CTLon_0479	577891..578238	CTL0483	576273..577745	CT231	257818	259290
CTLon_0480	578315..578851	CTL0484	577850..578197	CT232	259395	259742
CTLon_0481	578909..581695	CTL0485	578274..578810	CT233	259819	260355
CTLon_0482	581724..582137	CTL0486	578868..581654	CT234	260413	263199
CTLon_0483	582198..582431	CTL0487	581683..582096	CT235	263228	263641
CTLon_0484	582800..583546	CTL0488	582157..582390	CT236	263702	263935
CTLon_0485	583543..584469	CTL0489	582759..583505	CT237	264304	265050
CTLon_0486	584486..585469	CTL0490	583502..584428	CT238	265047	265973
CTLon_0487	585607..586209	CTL0491	584445..585428	CT239	265990	266973
CTLon_0488	586583..588961	CTL0492	585566..586168	CT240	267111	267713
CTLon_0489	589028..589549	CTL0493	586542..588920	CT241	268088	270466
CTLon_0490	589577..590641	CTL0494	588987..589508	CT242	270535	271056
CTLon_0491	590638..591834	CTL0495	589536..590600	CT243	271084	272148
CTLon_0492	592056..593078	CTL0496	590597..591793	CT244	272145	273341
CTLon_0493	593071..594057	CTL0497	592015..593037	CT245	273563	274585
CTLon_0494	594062..595351	CTL0498	593030..594016	CT246	274578	275564
CTLon_0495	595378..597822	CTL0499	594021..595310	CT247	275569	276858
CTLon_0496	597978..598328	CTL0500	595337..597781	CT248	276885	279329
CTLon_0497	598382..599752	CTL0500A	597937..598287	CT249	279485	279835
CTLon_0498	599829..602192	CTL0501	598341..599711	CT250	279889	281259
CTLon_0499	602502..603320	CTL0503	599788..602151	CT251	281336	283699
CTLon_0500	603571..604218	CTL0504	602461..603279	CT252	284009	284827
CTLon_0501	604222..604992	CTL0505	603530..604177	CT253	285078	285725
CTLon_0502	605200..605583	CTL0506	604181..604951	CT254	285729	286499
CTLon_0503	605772..607016	CTL0507	605159..605542	CT255	286707	287090
CTLon_0504	607006..608220	CTL0508	605731..606975	CT256	287279	288523

CTLon_0505	608224..609348	CTL0509	606965..608179	CT257	288513	289727
CTLon_0506	609354..610100	CTL0510	608183..609307	CT258	289731	290855
CTLon_0507	610419..610910	CTL0511	609313..610059	CT259	290861	291607
CTLon_0508	610919..611617	CTL0512	610378..610869	CT260	291930	292421
CTLon_0509	611614..612384	CTL0513	610878..611576	CT261	292430	293128
CTLon_0510	612377..612967	CTL0514	611573..612343	CT262	293125	293895
CTLon_0511	612988..614928	CTL0515	612336..612926	CT263	293888	294478
CTLon_0512	614894..615868	CTL0516	612947..614887	CT264	294499	296439
CTLon_0513	616001..617182	CTL0517	614853..615827	CT265	296405	297379
CTLon_0514	617300..617602	CTL0518	615960..617141	CT266	297512	298693
CTLon_0515	617822..618601	CTL0519	617259..617561	CT267	298811	299113
CTLon_0516	618516..619967	CTL0520	617781..618560	CT268	299333	300112
CTLon_0517	620290..622233	CTL0521	618475..619926	CT269	300027	301478
CTLon_0518	622220..622507	CTL0522	620249..622192	CT270	301801	303744
CTLon_0519	622507..623409	CTL0523	622179..622466	CT271	303731	304018
CTLon_0520	623687..624253	CTL0524	622466..623368	CT272	304018	304920
CTLon_0521	624260..624679	CTL0525	623646..624212	CT273	305198	305764
CTLon_0522	624921..626288	CTL0526	624219..624638	CT274	305771	306190
CTLon_0523	626340..626924	CTL0527	624880..626247	CT275	306433	307800
CTLon_0524	626896..627555	CTL0528	626299..626883	CT276	307839	308423
CTLon_0525	627557..629068	CTL0529	626855..627514	CT277	308395	309054
CTLon_0526	629072..630022	CTL0530	627516..629027	CT278	309056	310567
CTLon_0527	630012..630653	CTL0531	629031..629981	CT279	310571	311521
CTLon_0528	630659..631393	CTL0532	629971..630612	CT280	311511	312152
CTLon_0529	631417..631770	CTL0533	630618..631352	CT281	312158	312892
CTLon_0530	631790..633862	CTL0534	631376..631729	CT282	312916	313269
CTLon_0531	634057..635481	CTL0535	631749..633821	CT283	313289	315385
CTLon_0532	635487..636206	CTL0536	634016..635440	CT284	315556	316980
CTLon_0533	636471..639035	CTL0537	635446..636165	CT285	316986	317705
CTLon_0534	639016..640092	CTL0538	636430..638994	CT286	317970	320534
CTLon_0535	640323..642017	CTL0539	638975..640051	CT287	320515	321591
CTLon_0536	642104..643249	CTL0540	640283..641977	CT288	321823	323514
CTLon_0537	643306..643983	CTL0541	642064..643209	CT289	323601	324740
CTLon_0538	643986..644462	CTL0542	643267..643944	CT290	324798	325475
CTLon_0539	644464..644901	CTL0543	643947..644423	CT291	325478	325954
CTLon_0540	644938..645864	CTL0544	644425..644862	CT292	325956	326393
CTLon_0541	645935..646555	CTL0545	644899..645825	CT293	326430	327356
CTLon_0542	646687..648468	CTL0546	645896..646516	CT294	327428	328048
CTLon_0543	648620..649087	CTL0547	646648..648429	CT295	328180	329961
CTLon_0544	649196..649891	CTL0548	648581..649048	CT296	330113	330580
CTLon_0545	649875..651239	CTL0549	649157..649852	CT297	330689	331384
CTLon_0546	651283..651942	CTL0550	649836..651200	CT298	331368	332732
CTLon_0547	652738..655542	CTL0551	651244..651903	CT299	332710	333435
				CT300	333828	334175
CTLon_0549	655557..658376	CTL0553	652698..655502	CT301	334228	337032
CTLon_0550	658503..659018	CTL0554	655517..658336	CT302	337047	339866
CTLon_0551	658939..659364	CTL0555	658463..658978	CT303	339993	340508
CTLon_0552	659426..661375	CTL0556	658899..659324	CT304	340429	340854
CTLon_0553	661381..661992	CTL0557	659386..661335	CT305	340917	342866
CTLon_0554	661977..663293	CTL0558	661341..661952	CT306	342872	343483
CTLon_0555	663296..665071	CTL0559	661937..663253	CT307	343468	344784
CTLon_0556	665065..665865	CTL0560	663256..665031	CT308	344787	346562
CTLon_0557	666032..666658	CTL0561	665025..665825	CT309	346556	347356
CTLon_0558	666767..667477	CTL0562	665992..666618	CT310	347523	348149
CTLon_0559	667485..667856	CTL0563	666727..667437	CT311	348258	348968
CTLon_0560	667901..668884	CTL0564	667445..667816	CT312	348976	349347
CTLon_0561	668996..673186	CTL0565	667861..668844	CT313	349392	350375

CTLon_0562	673211..676969	CTL0566	668956..673146	CT314	350487	354677
CTLon_0563	677330..677722	CTL0567	673171..676929	CT315	354702	358460
CTLon_0564	677754..678272	CTL0568	677290..677682	CT316	358821	359213
CTLon_0565	678294..678992	CTL0569	677714..678232	CT317	359245	359763
CTLon_0566	679015..679440	CTL0570	678254..678952	CT318	359785	360483
CTLon_0567	679546..680094	CTL0571	678975..679400	CT319	360506	360931
CTLon_0568	680098..680346	CTL0572	679506..680054	CT320	361037	361585
CTLon_0569	680489..681673	CTL0573	680058..680306	CT321	361589	361837
CTLon_0570	682021..682242	CTL0574	680449..681633	CT322	361980	363164
CTLon_0571	682653..683564	CTL0575	681981..682202	CT323	363512	363733
CTLon_0572	683561..684007	CTL0576	682615..683526	CT324	364145	365056
CTLon_0573	686220..686402	CTL0577	683523..683969	CT325	365053	365499
CTLon_0576	687125..687751	CTL0580	686181..686363	CT326	365551	367242
				CT326.1	367229	367417
				CT326.2	367607	367789
CTLon_0577	687947..688771	CTL0581	687086..687712	CT327	368509	369135
CTLon_0578	688785..690335	CTL0582	687908..688732	CT328	369332	370156
CTLon_0579	690319..690537	CTL0583	688746..690296	CT329	370170	371720
CTLon_0580	690544..690816	CTL0583A	690280..690498	CT330	371929	372201
CTLon_0581	690813..692735	CTL0584	690505..690777	CT331	372198	374120
CTLon_0582	692832..694289	CTL0585	690774..692696	CT332	374217	375674
CTLon_0583	694312..699672	CTL0586	692793..694250	CT333	375697	381057
CTLon_0584	699890..701290	CTL0587	694273..699633	CT334	381275	382675
CTLon_0585	701283..701573	CTL0588	699851..701251	CT335	382668	382958
CTLon_0586	701583..703298	CTL0589	701244..701534	CT336	382968	384683
CTLon_0587	703298..703627	CTL0590	701544..703259	CT337	384683	385012
CTLon_0588	703765..704226	CTL0591	703259..703588	CT338	385149	385610
CTLon_0589	704283..704465	CTL0592	703726..704187	CT339	385894	387423
CTLon_0590	704510..706039	CTL0592A	704244..704426			
CTLon_0591	706111..708147	CTL0593	704471..706000	CT340	387496	389532
CTLon_0592	708182..709360	CTL0594	706072..708108	CT341	389567	390745
CTLon_0593	709389..709565	CTL0595	708143..709321	CT342	390774	390950
CTLon_0594	709763..710395	CTL0596	709350..709526	CT343	391147	391779
CTLon_0595	710705..713164	CTL0597	709724..710356	CT344	392089	394548
CTLon_0596	713266..713631	CTL0598	710666..713125	CT345	394650	395015
CTLon_0597	713974..714888	CTL0599	713227..713592	CT346	395365	396279
CTLon_0598	714950..715897	CTL0600	713935..714849	CT347	396341	397288
CTLon_0599	715951..717540	CTL0601	714911..715858	CT348	397342	398928
CTLon_0600	717576..718166	CTL0602	715912..717501	CT349	398964	399554
CTLon_0601	718148..719848	CTL0603	717537..718127	CT350	399536	401236
CTLon_0602	719855..721957	CTL0604	718109..719809	CT351	401243	403336
CTLon_0603	722031..722339	CTL0605	719816..721918	CT351a	403410	403718
CTLon_0604	722495..723040	CTL0606	721992..722300	CT352	403499	403804
				CT353	403874	404419
CTLon_0605	723315..724148	CTL0607	722456..723001	CT354	404694	405527
CTLon_0606	724164..725225	CTL0608	723276..724109	CT355	405543	406604
CTLon_0607	725530..727644	CTL0609	724125..725186	CT356	406909	409023
CTLon_0608	728101..728433	CTL0610	725491..727605	CT357R	409341	409628
CTLon_0609	729053..729643	CTL0611A	728062..728394	CT357	409480	409812
				CT358	409838	410374
				CT359	410434	411024
CTLon_0611	729895..730521	CTL0613	729014..729604	CT360	411276	411902
CTLon_0612	730683..731543	CTL0614	729856..730482	CT361	412064	412924
CTLon_0613	731553..732848	CTL0615	730644..731504	CT362	412934	414229
CTLon_0614	732841..733845	CTL0616	731514..732809	CT363	414222	415226
CTLon_0615	733855..734616	CTL0617	732802..733806	CT364	415236	415997
CTLon_0616	734793..736520	CTL0618	733816..734577	CT365	416174	417901

CTLon_0617	736690..738012	CTL0619	734754..736481	CT366	418071	419393
CTLon_0618	737954..738508	CTL0620	736651..737973	CT367	419335	419889
CTLon_0619	738501..739574	CTL0621	737915..738469	CT368	419882	420955
CTLon_0620	739571..740692	CTL0622	738462..739535	CT369	420952	422073
CTLon_0621	740673..742115	CTL0623	739532..740653	CT370	422054	423490
CTLon_0622	742157..742942	CTL0624	740634..742076	CT371	423532	424317
CTLon_0623	743083..744411	CTL0625	742118..742903	CT372	424459	425787
CTLon_0624	745083..746534	CTL0626	743044..744372	CT373	425856	426443
				CT374	426459	427910
CTLon_0626	746706..747764	CTL0628	745044..746495	CT375	428082	429140
CTLon_0627	747806..748786	CTL0629	746667..747725	CT376	429182	430162
CTLon_0628	748976..749116	CTL0630	747767..748747	CT377	430353	430493
CTLon_0629	749231..750808	CTL0631	748937..749077	CT378	430608	432185
CTLon_0630	750921..752264	CTL0633	749192..750769	CT379	432298	433641
CTLon_0631	752277..753077	CTL0634	750882..752225	CT380	433654	434454
CTLon_0632	753106..753879	CTL0635	752238..753038	CT381	434483	435256
CTLon_0633	753948..754784	CTL0636	753067..753840	CT382	435325	436161
CTLon_0634	754921..755112	CTL0637	753909..754745	CT382.1	436298	436489
CTLon_0635	755295..756026	CTL0638	754882..755073	CT383	436672	437403
CTLon_0636	756037..757656	CTL0639	755256..755987	CT384	437414	439033
CTLon_0637	757671..758006	CTL0640	755998..757617	CT385	439048	439383
CTLon_0638	758003..758872	CTL0641	757632..757967	CT386	439380	440249
CTLon_0639	759159..761234	CTL0642	757964..758833	CT387	440536	442611
CTLon_0640	761248..761595	CTL0643	759120..761195	CT388	442625	442972
CTLon_0641	761777..763003	CTL0644	761209..761556	CT389	443154	444380
CTLon_0642	763111..764295	CTL0645	761738..762964	CT390	444488	445672
CTLon_0643	764303..765310	CTL0646	763072..764256	CT391	445680	446687
CTLon_0644	765319..766449	CTL0647	764264..765271	CT392	446693	447826
CTLon_0645	766650..768395	CTL0648	765280..766410	CT393	448027	449772
CTLon_0646	768464..769642	CTL0649	766611..768356	CT394	449841	451019
CTLon_0647	769639..770211	CTL0650	768425..769603	CT395	451016	451588
CTLon_0648	770237..772219	CTL0651	769600..770172	CT396	451614	453596
CTLon_0649	772512..774596	CTL0652	770198..772180	CT397	453889	455973
CTLon_0650	774852..775616	CTL0654	772473..774557	CT398	456229	456993
CTLon_0651	776039..777025	CTL0655	774813..775577	CT399	457416	458402
CTLon_0652	777057..778223	CTL0656	776000..776986	CT400	458434	459600
CTLon_0653	778469..779707	CTL0657	777018..778184	CT401	459846	461084
CTLon_0654	779704..780813	CTL0658	778430..779668	CT402	461081	462340
CTLon_0655	780979..781788	CTL0659	779665..780774	CT403	462356	463165
CTLon_0656	781776..782603	CTL0660	780940..781749	CT404	463153	463980
CTLon_0657	782600..783199	CTL0661	781737..782564	CT405	463977	464576
CTLon_0658	783592..784056	CTL0662	782561..783160	CT406	464969	465433
CTLon_0659	784070..784444	CTL0663	783553..784017	CT407	465447	465821
CTLon_0660	784450..784953	CTL0664	784031..784405	CT408	465827	466330
CTLon_0661	785055..786416	CTL0665	784411..784914	CT409	466432	467793
CTLon_0662	786632..787909	CTL0666	785016..786377	CT410	468008	469285
CTLon_0663	787970..789793	CTL0667	786593..787870	CT411	469346	471169
CTLon_0664	789900..792827	CTL0668	787931..789754	CT412	471276	474203
CTLon_0665	792966..798215	CTL0669	789861..792788	CT413	474342	479597
CTLon_0666	798391..803715	CTL0670	792927..798176	CT414	479774	485086
CTLon_0667	804268..805098	CTL0671	798351..803675	CT415	485639	486469
CTLon_0668	805095..805805	CTL0672	804228..805058	CT416	486466	487176
CTLon_0669	805796..806677	CTL0673	805055..805765	CT417	487167	488048
CTLon_0670	806593..807600	CTL0674	805756..806637	CT418	487964	488971
CTLon_0671	807690..807941	CTL0675	806553..807560	CT419	489061	489312
CTLon_0672	807972..808295	CTL0676	807650..807901	CT420	489343	489666
CTLon_0673	808845..809546	CTL0677	807932..808255	CT421	490216	490917

CTLon_0674ae	809731..809892	CTL0678	808805..809506	CT421.1	491102	491263
CTLon_0675af	809909..810070	CTL0679	809691..809852	CT421.2	491280	491441
CTLon_0676	810084..810569	CTL0680	809869..810030	CT422	491454	491939
CTLon_0677	810702..811811	CTL0681	810043..810528	CT423	492072	493181
CTLon_0678	812000..812350	CTL0682	810661..811770	CT424	493370	493720
CTLon_0679	812528..814393	CTL0683	811959..812309	CT425	493898	495763
CTLon_0680	814590..815699	CTL0684	812487..814352	CT426	495960	497069
CTLon_0681	815672..816493	CTL0685	814549..815658	CT427	497042	497863
CTLon_0682	816429..817118	CTL0686	815631..816452	CT428	497799	498488
CTLon_0683	817144..818133	CTL0687	816388..817077	CT429	498514	499503
CTLon_0684	818194..819021	CTL0688	817103..818092	CT430	499564	500391
CTLon_0685	818990..819568	CTL0689	818153..818980	CT431	500360	500938
CTLon_0686	819580..821073	CTL0690	818949..819527	CT432	500950	502443
CTLon_0687	821437..822108	CTL0691	819539..821032	CT433	502807	503478
CTLon_0688	822211..822747	CTL0692	821396..822067	CT434	503581	504117
CTLon_0689	822744..823796	CTL0693	822170..822706	CT435	504114	505166
CTLon_0690	823813..824130	CTL0694	822703..823755	CT436	505183	505500
CTLon_0691	824138..826222	CTL0695	823772..824089	CT437	505508	507592
CTLon_0692	826264..826737	CTL0696	824097..826181	CT438	507634	508107
CTLon_0693	826787..827176	CTL0697	826223..826696	CT439m	508157	508528
CTLon_0694	827418..827756	CTL0698	826746..827135	CT440	508788	509126
CTLon_0695	827914..829863	CTL0699	827377..827715	CT441	509284	511218
CTLon_0696	829970..830422	CTL0700	827873..829822	CT442	511340	511792
CTLon_0697	830600..832261	CTL0701	829929..830381	CT443	511971	513632
CTLon_0698	832408..832674	CTL0702	830559..832220	CT444	513779	514045
CTLon_0699	833011..833235	CTL0703	832367..832633	CT444.1	514382	514606
CTLon_0700	833232..834752	CTL0704	832970..833194	CT445	514603	516123
CTLon_0701	835023..835574	CTL0705	833191..834711	CT446	516395	516946
CTLon_0702	835978..837732	CTL0706	834982..835533	CT447	517351	519105
CTLon_0703	837850..842052	CTL0707	835937..837691	CT448	519223	523425
CTLon_0704a	842472..842804	CTL0708a	837809..842011	CT449	523845	524177
CTLon_0705	843266..844027	CTL0709	842431..842763	CT450	524640	525401
CTLon_0706	844033..844950	CTL0710	843225..843986	CT451	525407	526324
CTLon_0707	844947..845597	CTL0711	843992..844909	CT452	526321	526971
CTLon_0708	845594..846244	CTL0712	844906..845556	CT453	526968	527618
CTLon_0709	846259..847950	CTL0713	845553..846203	CT454	527633	529324
CTLon_0710	847990..849324	CTL0714	846218..847909	CT455	529362	530696
CTLon_0711	849536..852541	CTL0715	847947..849281	CT456	530908	533925
CTLon_0712	852593..853309	CTL0716	849494..852511	CT457	533977	534693
CTLon_0713	853492..853995	CTL0717	852563..853279	CT458	534878	535381
CTLon_0714	853992..855099	CTL0718	853462..853965	CT459	535378	536485
CTLon_0715	855419..855679	CTL0719	853962..855069	CT460	536805	537065
CTLon_0716	855737..856726	CTL0720	855389..855649	CT461	537123	538112
CTLon_0717	856716..857375	CTL0721	855707..856696	CT462	538102	538761
CTLon_0718	857384..858187	CTL0722	856686..857345	CT463	538770	539573
CTLon_0719	858139..858813	CTL0723	857354..858157	CT464	539525	540199
CTLon_0720	858906..859547	CTL0724	858109..858783	CT465	540292	540933
CTLon_0721	859841..860170	CTL0725	858876..859517	CT466	541227	541556
CTLon_0722	860148..861206	CTL0726	859811..860140	CT467	541534	542592
CTLon_0723	861249..862409	CTL0727	860118..861176	CT468	542635	543795
CTLon_0724	862636..863172	CTL0728	861219..862379	CT469	544022	544558
CTLon_0725	863142..863873	CTL0729	862606..863142	CT470	544528	545259
CTLon_0726	863870..864472	CTL0730	863112..863843	CT471	545256	545858
CTLon_0727	864950..865744	CTL0731	863840..864442	CT472	546322	547116
CTLon_0728	865750..866064	CTL0733	864920..865714	CT473	547122	547436
CTLon_0729	866003..866932	CTL0734	865720..866034	CT474	547375	548304
CTLon_0730	867020..869392	CTL0735	865973..866902	CT475	548392	550764

CTLon_0731	869389..870354	CTL0736	866990..869362	CT476	550761	551726
CTLon_0732	870373..870885	CTL0737	869359..870324	CT477	551745	552257
CTLon_0733	870882..872618	CTL0738	870343..870855	CT478	552254	553990
CTLon_0734	872620..874035	CTL0739	870852..872588	CT479	553992	555470
CTLon_0735	874080..876170	CTL0740	872590..874005	CT480	555452	557542
CTLon_0736	876776..877507	CTL0741	874050..876140	CT480.1	557793	557957
CTLon_0737	877669..878322	CTL0742	876746..877477	CT481	558150	558881
CTLon_0738	878790..879155	CTL0743	877639..878292	CT482	559043	559696
CTLon_0739	879134..880132	CTL0744	878760..879125	CT483	560164	560529
CTLon_0740	880289..881233	CTL0745	879104..880102	CT484	560508	561506
CTLon_0741	881255..882040	CTL0746	880259..881203	CT485	561663	562607
CTLon_0742	882086..882658	CTL0747	881225..882010	CT486	562629	563414
CTLon_0743	882655..883389	CTL0748	882056..882628	CT487	563460	564032
CTLon_0744	883478..884803	CTL0749	882625..883359	CT488	564029	564763
CTLon_0745	884994..885245	CTL0750	883448..884773	CT489	564852	566177
CTLon_0746	885255..886649	CTL0751	884965..885216	CT490	566370	566621
CTLon_0747	886646..887254	CTL0752	885226..886620	CT491	566631	568025
CTLon_0748	887248..889848	CTL0753	886617..887225	CT492	568022	568630
CTLon_0749	889865..890860	CTL0754	887219..889819	CT493	568624	571224
CTLon_0750	891000..892622	CTL0755	889836..890831	CT494	571241	572236
CTLon_0751	892834..893340	CTL0756	890971..892593	CT495	572375	573997
CTLon_0752	893736..893885	CTL0757	892805..893311	CT496	574209	574715
CTLon_0753	893854..895272	CTL0758	893707..893856	CT496.1	575111	575260
CTLon_0754	895567..897399	CTL0759	893825..895243	CT497	575229	576647
CTLon_0755	897389..898090	CTL0760	895538..897370	CT498	576941	578773
CTLon_0756	898147..898572	CTL0761	897360..898061	CT499	578763	579464
CTLon_0757	898732..899334	CTL0762	898118..898543	CT500	579521	579946
CTLon_0758	899354..899866	CTL0763	898703..899305	CT501	580106	580708
CTLon_0759	899974..900423	CTL0764	899325..899837	CT502	580728	581240
CTLon_0760	900815..901681	CTL0765	899945..900394	CT503	581348	581902
CTLon_0761	901692..902696	CTL0766	900786..901652	CT504	582189	583055
CTLon_0762	902740..903165	CTL0767	901663..902667	CT505	583066	584070
CTLon_0763	903174..904307	CTL0768	902711..903136	CT506	584114	584539
CTLon_0764	904328..904726	CTL0769	903145..904278	CT507	584548	585681
CTLon_0765	904748..905116	CTL0770	904299..904697	CT508	585702	586100
CTLon_0766	905172..906545	CTL0771	904719..905087	CT509	586122	586490
CTLon_0767	906568..907002	CTL0772	905143..906516	CT510	586546	587919
CTLon_0768	906995..907492	CTL0773	906539..906973	CT511	587942	588376
CTLon_0769	907507..907878	CTL0774	906966..907463	CT512	588369	588866
CTLon_0770	907900..908451	CTL0775	907478..907849	CT513	588881	589252
CTLon_0771	908479..908880	CTL0776	907871..908422	CT514	589274	589825
CTLon_0772	908898..909440	CTL0777	908450..908851	CT515	589853	590254
CTLon_0773	909442..909777	CTL0778	908869..909411	CT516	590272	590814
CTLon_0774	909790..910158	CTL0779	909413..909748	CT517	590816	591151
CTLon_0775	910175..910426	CTL0780	909761..910129	CT518	591164	591532
CTLon_0776	910419..910637	CTL0781	910146..910397	CT519	591549	591800
CTLon_0777	910639..911055	CTL0782	910390..910608	CT520	591793	592011
CTLon_0778	911088..911762	CTL0783	910610..911026	CT521	592013	592429
CTLon_0779	911772..912107	CTL0784	911059..911733	CT522	592462	593136
CTLon_0780	912126..912392	CTL0785	911743..912078	CT523	593146	593481
CTLon_0781	912398..913252	CTL0786	912097..912363	CT524	593500	593766
CTLon_0782	913276..913611	CTL0787	912369..913223	CT525	593772	594626
CTLon_0783	913627..914295	CTL0788	913247..913582	CT526	594650	594985
CTLon_0784	914304..914969	CTL0789	913598..914266	CT527	595001	595669
CTLon_0785	915404..916300	CTL0790	914275..914940	CT528	595678	596343
CTLon_0786	916466..917416	CTL0791	915375..916271	CT529	596778	597674

CTLon_0787	917406..918248	CTL0792	916437..917387	CT530	597840	598790
CTLon_0788	918260..918721	CTL0793	917377..918219	CT531	598780	599622
CTLon_0789	918718..919578	CTL0794	918231..918692	CT532	599634	600095
CTLon_0790	919677..921305	CTL0795	918689..919549	CT533	600092	600952
CTLon_0791	921369..921851	CTL0796	919649..921277	CT534	601053	602681
CTLon_0792	921987..922739	CTL0797	921341..921823	CT535	602745	603227
CTLon_0793	922743..923216	CTL0798	921959..922711	CT536	603363	604115
CTLon_0794	923195..923911	CTL0799	922715..923188	CT537	604119	604592
CTLon_0795	924633..924941	CTL0800	923167..923883	CT538	604571	605287
CTLon_0796	924991..925446	CTL0801	924605..924913	CT539	606009	606317
CTLon_0797	925462..926193	CTL0802	924963..925418	CT540	606367	606822
CTLon_0798	926330..928078	CTL0803	925434..926165	CT541	606838	607569
CTLon_0799	928059..929345	CTL0804	926303..928051	CT542	607707	609455
CTLon_0800	929781..931151	CTL0805	928032..929318	CT543	609436	610722
CTLon_0801	931165..934878	CTL0806	929754..931124	CT544	611158	612528
CTLon_0802	935107..935976	CTL0807	931138..934851	CT545	612542	616255
CTLon_0803	936128..937084	CTL0808	935080..935949	CT546	616484	617353
CTLon_0804	937109..937693	CTL0809	936101..937057	CT547	617505	618461
CTLon_0805	937680..938120	CTL0810	937082..937666	CT548	618486	619070
CTLon_0806	938127..938552	CTL0811	937653..938093	CT549	619057	619497
CTLon_0807	938660..939973	CTL0812	938100..938525	CT550	619504	619929
CTLon_0808	940106..940513	CTL0813	938633..939946	CT551	620037	621068
CTLon_0809	940510..941616	CTL0814	940079..940486	CT552	621483	621890
CTLon_0810	941766..943001	CTL0815	940483..941589	CT553	622015	622992
CTLon_0811	943175..946774	CTL0817	941739..942974	CT554	623142	624377
CTLon_0812	946953..947432	CTL0818	943148..946747	CT555	624552	628151
CTLon_0813	947641..949038	CTL0819	946926..947405	CT556	628329	628808
CTLon_0814	949035..949970	CTL0820	947614..949011	CT557	629017	630414
CTLon_0815	950074..951051	CTL0821	949008..949943	CT558	630411	631346
CTLon_0816	951052..951888	CTL0822	950047..951024	CT559	631450	632430
CTLon_0817	952164..952835	CTL0823	951025..951861	CT560	632431	633267
CTLon_0818	952848..953768	CTL0824	952137..952808	CT561	633543	634214
CTLon_0819	953780..954064	CTL0825	952821..953741	CT562	634227	635147
CTLon_0820	954071..954940	CTL0826	953753..954037	CT563	635159	635443
CTLon_0821	955019..955462	CTL0827	954044..954913	CT564	635450	636319
CTLon_0822	955552..956544	CTL0828	954992..955435	CT565	636397	636840
CTLon_0823	956550..957074	CTL0829	955525..956517	CT566	636931	637923
CTLon_0824	957059..957514	CTL0830	956523..957047	CT567	637929	638453
CTLon_0825	957498..957827	CTL0831	957032..957487	CT568	638438	638893
CTLon_0826	957890..959065	CTL0832	957471..957800	CT569	638877	639206
CTLon_0827	959077..960582	CTL0833	957862..959037	CT570	639268	640443
CTLon_0828	960566..962848	CTL0834	959049..960554	CT571	640455	641960
CTLon_0829	962849..964078	CTL0835	960538..962820	CT572	641944	644226
CTLon_0830	964207..965277	CTL0836	962821..964050	CT573	644227	645456
CTLon_0831	965288..967018	CTL0837	964179..965249	CT574	645584	646654
CTLon_0832	967298..967996	CTL0838	965260..966990	CT575	646665	648395
CTLon_0833	968013..968372	CTL0839	967270..967968	CT576	648690	649388
CTLon_0834	968409..969872	CTL0840	967985..968344	CT577	649405	649764
CTLon_0835	969903..971222	CTL0841	968381..969844	CT578	649801	651264
CTLon_0836	971300..972283	CTL0842	969875..971194	CT579	651295	652614
CTLon_0837	972588..974495	CTL0843	971272..972255	CT580	652692	653675
CTLon_0838	974947..975714	CTL0844	972560..974467	CT581	653980	655887
CTLon_0839	975719..976510	CTL0845	974919..975686	CT582	656338	657105
CTLon_0840	976476..977027	CTL0846	975691..976482	CT583	657110	657901
CTLon_0841	977206..978246	CTL0847	976448..976999	CT584	657867	658418
CTLon_0842	978271..980277	CTL0848	977178..978218	CT585	658617	659657
CTLon_0843	980439..981713	CTL0849	978243..980249	CT586	659682	661688

CTLon_0844	981840..983792	CTL0850	980411..981685	CT587	661850	663124
CTLon_0845	983917..985725	CTL0851	981813..983765	CT588	663251	665203
CTLon_0846	985740..988604	CTL0852	983890..985698	CT589	665328	667136
CTLon_0847	988701..989399	CTL0853	985713..988577	CT590	667151	670015
CTLon_0848	989478..991358	CTL0854	988674..989372	CT591	670112	670810
CTLon_0849	992363..993154	CTL0855	989451..991331	CT592	671047	672768
				CT593	672765	673334
				CT593.1	673360	673551
CTLon_0851	993239..995317	CTL0858	992336..993127	CT594	673773	674564
CTLon_0852	995470..996168	CTL0859	993212..995290	CT595	674649	676727
CTLon_0853	996165..996572	CTL0860	995443..996141	CT596	676880	677578
CTLon_0854	996575..997282	CTL0861	996138..996545	CT597	677575	677982
CTLon_0855	997282..998577	CTL0861A	996548..997255	CT598	677985	678692
CTLon_0856	998574..999140	CTL0862	997255..998550	CT599	678692	679987
CTLon_0857	999130..999732	CTL0863	998547..999113	CT600	679984	680550
CTLon_0858	999803..1000195	CTL0864	999103..999705	CT601	680540	681142
CTLon_0859	1000192..1000779	CTL0865	999776..1000168	CT602	681213	681605
CTLon_0860	1000987..1002588	CTL0866	1000165..1000752	CT603	681602	682189
CTLon_0861	1002668..1003894	CTL0867	1000960..1002561	CT604	682398	683999
CTLon_0862	1004091..1004720	CTL0868	1002641..1003867	CT605	684076	685305
CTLon_0863	1004694..1004933	CTL0869	1004063..1004692	CT606	685503	686132
CTLon_0864	1004918..1005607	CTL0870	1004666..1004905	CT606.1	686106	686345
CTLon_0865	1005688..1007592	CTL0871	1004890..1005579	CT607	686330	687019
CTLon_0866	1007596..1008906	CTL0872	1005660..1007564	CT608	687100	689004
CTLon_0867	1009014..1009709	CTL0873	1007568..1008878	CT609	689008	690318
CTLon_0868	1009706..1010437	CTL0874	1008986..1009681	CT610	690426	691121
CTLon_0869	1010409..1010888	CTL0875	1009678..1010409	CT611	691118	691849
CTLon_0870	1010885..1012237	CTL0876	1010381..1010860	CT612	691821	692300
CTLon_0871	1012234..1012608	CTL0877	1010857..1012209	CT613	692297	693649
CTLon_0872	1012627..1014342	CTL0878	1012206..1012580	CT614	693646	694020
CTLon_0873	1014491..1015780	CTL0879	1012599..1014314	CT615	694039	695754
CTLon_0874	1016229..1016525	CTL0880	1014463..1015752	CT616	695903	697192
CTLon_0875	1016758..1017558	CTL0881	1016201..1016497	CT617	697641	697937
CTLon_0876	1017615..1020242	CTL0882	1016730..1017530	CT618	698170	698970
CTLon_0877	1020357..1022873	CTL0883	1017586..1020213	CT619	699026	701659
CTLon_0878	1022937..1025435	CTL0884	1020329..1022845	CT620	701774	704290
CTLon_0879	1025663..1027618	CTL0885	1022909..1025407	CT621	704354	706852
CTLon_0880	1027725..1029026	CTL0886	1025635..1027590	CT622	707080	709023
CTLon_0881	1029270..1030853	CTL0887	1027698..1028999	CT623	709131	710429
CTLon_0882	1031053..1031919	CTL0888	1029243..1030826	CT624	710672	712282
CTLon_0883	1032015..1032644	CTL0889	1031026..1031892	CT625	712455	713321
CTLon_0884	1033028..1034011	CTL0890	1031988..1032617	CT626	713418	714047
CTLon_0885	1034084..1034959	CTL0891	1033001..1033984	CT627	714431	715414
CTLon_0886	1035015..1035632	CTL0892	1034057..1034932	CT628	715486	716361
CTLon_0887	1035685..1036368	CTL0893	1034988..1035605	CT629	716417	717034
CTLon_0888	1036548..1036802	CTL0894	1035658..1036341	CT630	717087	717770
CTLon_0889	1036983..1038572	CTL0895	1036521..1036775	CT631	717949	718203
CTLon_0890		CTL0897	1036956..1038545	CT632	718384	719973

Appendix H: Homology verification tables

Table H-1. Homology check for 13 promoters predicted by MMCTPP1 for positive strand. Red sequence: predicted promoter; Green: experimental TSS; Red singleton: GSS; Underlined: noted non-homologies

	TSS-40	TSS	Annotated Gene Start	Promoter to GSS Sequence
CTLon_0027	33208	33248	33281	AAGGTTGAAT AAAATCTTTT CCGAACCGTA TCATAGAAGG GTTTCAAAAG ACGAAGTCCT GTTTTAAGGA GGCT
CT658	753471	753511	753544	GGTTGAATAA ATCTTTTCCG AACCGTATCA TGGAAGGGTT TCAAAAGACG AAGTCCTGTT TTAAGGAGGC TTGA
CTLon_0077	94558	94598	94629	TTTCATTGAT TTAGCGGAAG TAAAAAGGTA CAAGTAACAG GTCTGTCAAC CCCCTATGTT TTAGAGGAGA AA
CT708	814791	814831	814862	TTTCATTGAT TTAGCGGAAG TAAAAAGGTA CAAGTAACAG GTCTGTCAAC CCCCTATGTT TTAGAGGAGA AA
CTLon_0152	199253	199293	199378	TTGTTTGCTT TTAATGAAAA AAAGAATATA CACGAAAAGT GTTCGAAAAG CTGCTTTGGG AGAGGGTTTC TCTGGGTTTT CGATGGTGTC GTTATTTCTA ACGAACAAGT AAGGAGTAGG AATTCA
CT783	919414	919454	919539	TTGTTTGCTT TTAATGAAAA AAAGAATATA CACGAAAAGT GTTCGAAAAG CTGCTTTGGG AGAGGGGTTC TCTGGGTTTT CGATGGTGTC GTTATTTCTA ACGAACAAGT AAGGAGTAGG AATTCA
CTLon_0331	415922	415962	415995	ATGGTTTATG AAAAACAATT TTTTAATTTA AAATTAGAAT AGATTTTGAA ATAAATTATT CTGGTTTCTG CTCA
CT080	93541	93581	93614	ATGGTTTATG AAAAACAATT TTTTAATTTA AAATTAGAAT AGATTTTGAA ATAAATTATT CTGGTTTCTG CTCA
CTLon_0376	463896	463936	463958	TTTGTTTGGA AAAAATAATC ATCAAAATTA TAATCATTCC CTCTGATAAG GTGATTTAAG TTA
CT125	141432	141472	141494	TTTGTTTGGA AAAAATAATC ATCAAAATTA TAATCATTCC CTCTGATAAG GTGATTTAAG TTA
CTLon_0501	603439	603479	603571	TATTAGTTGC TTTTTGAAAA

				TACTCATGCT AGAGTTCTCC TTAATACATA AGTTCCTCAG GTCTTTTGCG CAAGCTTACA AGAGTGTTGC TAGGGACATA AAATCGAATC AATTTTTTCA CTGAGTTGCG TTA
CT253	284946	284986	285078	TATTAGTTGC TTTTTGAAAA TACTCATGCT AGAGTTCTCC TTAATACATA AGTTCCTCAG GTCTTTTGCG CAAGCTTACA AGAGTGTTGC TAGGGACGTA AAATCGAATC AATTTTTTCA CTGAGTTGCG TTA
CTLon_0534	636404	636444	636471	AAGTTGCATC ATTATCATAA ATGTCGTATA TGCTTGAAAA ATATTCCACC TTGCCATTCA GGTTTTTA
CT286	317903	317943	317970	AAGTTGCATC ATTATCATAA ATGTCGTATA TGCTTGAAAA ATATTCCACC TTGCCATTCA GGTTTTTA
CTLon_0536	640260	640300	640323	AATTGTTGTA AAAAACAATA TTTATTCTAA AATAATAACC ATAGTTACGG GGGAATCTCT TTCA
CT288	321760	321800	213823	ATTGTTGTAA AAAAACAATA TTTATTCTAA AATAATAACC ACAGTTACGG GGAAATCTCT TTCA
CTLon_0624	742964	743004	743083	AAACTCTGGC AAAAAAATCT TTTTTCCACT ACACGGGTGG AAAAGCTTTA TTAGAGGTTG TTGTGTCCTT CCGTTCGGTT TTACTGACTG CTCTGCTCTC CCTTTCTTTT ACGACCACCA
CT372	424340	424380	424459	AACTCTGGCA AAAAAAATCT TTTTTCCACT ACACGGGTGG AAAAGCTTTG TTAGAGGTTG TTGTGTCCTT CCGTTCGGTT TTACTGACTG CTCTGCTCTC CCTTTCTTTT ACGACTACCA
CTLon_0666	792891	792931	792966	ATACCTTGCC TAATTTACTT TTCTGATTTA TCTAACGCCT ATCGAGTTCG TACATATTCA ATAGGTTTGT CTTCTA
CT413	474267	474307	474342	ATACCTTGCC TAATTTACTT TTCTGATTTA TCTAACGCCT ATCGAGTTCG TACATATTCA ATAGGTTTGT CTTCTA
CTLon_0816	950016	950056	950074	TCCCGATTGG CACTAATCTC CCCATTTGCT ATGGTGAGTG AAAAGGTGTG CGTGAGTTA
CT559	631392	631432	631450	TCCCGATTGG CACTAATCTC CCCATTTGCT ATGGTGAGTG AAAAGGTGTG CGTGGGTTA
CTLon_0833	967218	967258	967298	ATCAACTTGT TAAATCAGAT

				CGTTAGAATT TAATATTGTT AGTAGTAATT TGTTATTTTA TTTTTTTAGG AATTATCGCG A
CT576	648610	648650	648690	TTAACTTGTT AAATCAGATC GTTAGAATTT AATATTGTTA GTAGTAATTT GTTATTTTTA TTTTTTTAGG AATTATCGCG A
CTLon_0876	1016631	1016671	1016758	TGCATCGATT TAAAAGCGAT TTCTTTTTAC AATGTCTTCC CGATATGCCT CCTTTTGAGT CATAAACCTT TGGTTTCACA AGATTTTTTA CGCAAAGGAC CCTTAATTTT TTTTGGAGGT TTCCACAA
CT618	698043	698083	698170	TGCATCGATT TAAAAGCGAT TTCTTTTTAC AATGCTTTCC CGATATGCCT CCTTTTGAGT CATAAACCTT TGGTTTCACA AGATTTTTGA CGCAAAAGGC CCTTAATTTT TTTTGGAGGT TTCCACAA

Table H-2. Homology check for 9 promoters predicted by MMCTPP1 for negative strand. Red sequence: predicted promoter; Green: experimental TSS; Red singleton: GSS; Underlined: noted non-homologies

	Annotated Gene End	TSS	TSS+40	Reverse Complement	Promoter to GSS Sequence
CTLon_0002	2440	2440	2480	TATCTCGGGA TTATAGAAGT TTTTATAAGG GAATCCAATT T	AAATTGGATT CCCTTATAAA AACTTCTATA ATCCCGAGAT A
CT634	722710	722710	722750	TATCTCGGGA TTATAGAAGT TTTTATAAGG GAATCCAATT T	AAATTGGATT CCCTTATAAA AACTTCTATA ATCCCGAGAT A
CTLon_0515	617602	617661	617701	TGGTAGCTTG TTGCCTCCTA GTTAAAGTGG TAACCCTTGC GTCCCTAAAT ACCGCAATAT ATGCGTATAA TATGCATTCA TCCTTTGGAT TCAAGACTTT	AAAGTCTTGA ATCCAAAGGA TGAATGCATA TTATACGCAT ATATTGCGGT ATTTAGGGAC GCAAGGGTTA CCACTTTAAC TAGGAGGCAA CAAGCTACCA
CT267	299113	299172	299212	TGGTAGCTTG TTGCCTCCTA GTTAAAGTGG TAACCCTTGC GTCCCTAAAT ACCGCAATAT ATGCGTATAA TATGCATTCA TCCTTTGGAT TCAAGACTTT	AAAGTCTTGA ATCCAAAGGA TGAATGCATA TTATACGCAT ATATTGCGGT ATTTAGGGAC GCAAGGGTTA CCACTTTAAC TAGGAGGCAA CAAGCTACCA
CTLon_0561	668884	668909	668949	TAAACACCTT GTCAATTTTT GACTCTAAGT GAAGCCTACC AAAAAAAACG ATACTATTCA AGGGGG	CCCCCTTGAA TAGTATCGTT TTTTTTGGTA GGCTTCACTT AGAGTCAAAA ATTGACAAGG TGTTTA
CT313	350375	350400	350440	TAAACACCTT GTCAATTTTT GACTTTAAGT GAAGCCTACC AAAAAAAACG ATACTATTCA AGGGGG	CCCCCTTGAA TAGTATCGTT TTTTTTGGTA GGCTTCACTT AAAGTCAAAA ATTGACAAGG TGTTTA
CTLon_0607	725225	725411	725451	TCCTTGATTT GTAGTTTTTA GGTAGAAAAA CCTTAAGAAT TTTGGGTTGT TCCTCCTCCC	TTGGCTTGAG GATATAACGC TTTTTTGTTA AAAGTGTTCT GACGGCTGGG TCCCTCCTCC

				CTTTTTCTTTGGATCTTACCGCCTCCCTAGAGATGTGGCAAACACCCAAACAAAGCAGCTTATATTCTAAAAAAGAACCTACCCTCGTCCTAGGTAAAAGCTATAGGGGAGGAGGGACCCAGCCGTCAGAACACTTTTAACAAAAAAGCGTTATATCCTCAAGCCAA	CCTATAGCTTTTACCTAGGACGAGGGTAGGTTCTTTTTTAGAATATAAGCTGCTTTGTTTGGGTGTTTGCCACATCTCTAGGGAGGCGGTAAGATCCAAAGAAAAAGGGGAGGAGGAACAACCCAAAATTCTTAAGGTTTTTCTACCTAAAAACTACAAATCAAGGA
CT355	406604	406790	406830	TCCTTGATTTGTAGTTTTTAGGTAGAAAAACCTTAAGAATTTTGGGTTGTTCCTCCTCCCCTTTTTCTTTGGATCTTACCGCCTCCCTAGAGATGTGGCAAACACCCAAACAAAGCAGCTTATATTCTAAAAAAGAACCTACCCTCGTCCTAGGTAAAAGCTATAGGGGAGGAGGGACCCAGCCGTCAGAACACTTTTAACAAAAAAGCGTTATATCCTCAAGCCAA	TTGGCTTGAGGATATAACGCTTTTTTGTTAAAAGTGTTCTGACGGCTGGGTCCCTCCTCCCCTATAGCTTTTACCTAGGACGAGGGTAGGTTCTTTTTTAGAATATAAGCTGCTTTGTTTGGGTGTTTGCCACATCTCTAGGGAGGCGGTAAGATCCAAAGAAAAAGGGGAGGAGGAACAACCCAAAATTCTTAAGGTTTTTCTACCTAAAAACTACAAATCAAGGA
CTLon_0628	748786	748855	748895	TAAGCCACCCTCTCTTTACTTTTACAAAACGCACATACTCTCAACACTACGTTTGCAACTAACTAATTTTGGTCCCAACATACGTTTGGATGATAAAAGAATCAAGTACC	GGTACTTGATTCTTTTATCATCCAAACGTATGTTGGGACCAAAATTAGTTAGTTGCAAACGTAGTGTTGAGAGTATGTGCGTTTTGTAAAAGTAAAGAGAGGGTGGCTTA
CT376	430162	430231	430271	TAAGCCACCCTCTCTTTACTTTTACAAAACGCACATACTCTCAACACTAC	GGTACTTGATTCTTTTATCATCCAAACGTATGTTGGGACCAAAATTAGTT

				GTTTGCAACTAACTAATTTTGGTCCCAACATACGTTTGGATGATAAAAGAATCAAGTACC	AGTTGCAAACGTAGTGTTGAGAGTATGTGCGTTTTGTAAAAGTAAAGAGAGGGTGGCTTA
CTLon_0629	749116	749153	749193	TACAGCCCCTAAAAAAACGATTTTAAGAGAGAAGTGATAGACAGATTATAACATATTTAAAATAAAAACTCTGCAAAC	GTTTGCAGAGTTTTTATTTTAAATATGTTATAATCTGTCTATCACTTCTCTCTTAAAATCGTTTTTTTAGGGGCTGTA
CT377	430493	430530	430570	TAAAGCCCCTAAAAAAACGATTTTAAGAGAGAAGTAATAGACAGATTATAACATATTTAAAATAAAAACTCTGCAAAC	GTTTGCAGAGTTTTTATTTTAAATATGTTATAATCTGTCTATTACTTCTCTCTTAAAATCGTTTTTTTAGGGGCTTTA
CTLon_0699	832674	832765	832805	TAACTTCCAGACTCCTTTCTAGAAAAGGGCTCTTGAAGTTTCTTTTATCGATAAAAGCAATTCTTTTAATAATAAAAGAAACTAGCCCTCATAGACAATATTACATTATAAAATAAAAATTATATCAATTGT	ACAATTGATATAATTTTTATTTTATAATGTAATATTGTCTATGAGGGCTAGTTTCTTTTATTATTAAAAGAATTGCTTTTATCGATAAAAGAAACTTCAAGAGCCCTTTTCTAGAAAGGAGTCTGGAAGTTA
CT444	514045	514136	514176	TAACTTCCAGACTCCTTTCTAGAAAAGGGCTCTTGAAGTTTCTTTTATCGATAAAAGCAATTCTTTTAACAATAAAAGAAACTAGCCCTCATAGACAATATTACATTATAAAATAAAAATTATATCAATTGT	ACAATTGATATAATTTTTATTTTATAATGTAATATTGTCTATGAGGGCTAGTTTCTTTTATTGTTAAAAGAATTGCTTTTATCGATAAAAGAAACTTCAAGAGCCCTTTTCTAGAAAGGAGTCTGGAAGTTA
CTLon_0752	893340	893455	893495	TAACGTTGATTCGGAAGCTTGTCTTGAAAATTCTTAGCCTCCAGCGCAATGACACATATT	TTTTGTTTGTTTGAATGTTTTTTTGTTGATAAGCTGGGGGAAATGGCGGGAACATAGCTA

				ATGAAACATAGACACTCCAAAAGCAAGCAGAGAGTTTAGCTATGTTCCCGCCATTTCCCCCAGCTTATCAACAAAAAAACATTCAAACAAACAAAA	AACTCTCTGCTTGCTTTTGGAGTGTCTATGTTTCATAATATGTGTCATTGCGCTGGAGGCTAAGAATTTTCAAGACAAGCTTCCGAATCAACGTTA
CT496	574715	574830	574870	TAACGTTGATTCGGAAGCTTGTCTTGAAAATTCTTAGCCTCCAGCGCAATGACACATATTATGAAACATAGACACTCCAAAAGCAAGCAGAGAGTTTAGCTATGTTTCCCGCCATTTCCCCCAGCTTATCAACAAAAAAACATTCAAACAAACAAAA	TTTTGTTTGTTTGAATGTTTTTTGTTGATAAGCTGGGGGAAATGGCGGGAAACATAGCTAAACTCTCTGCTTGCTTTTGGAGTGTCTATGTTTCATAATATGTGTCATTGCGCTGGAGGCTAAGAATTTTCAAGACAAGCTTCCGAATCAACGTTA
CTLon_0879	1025435	1025466	1025506	TTGCCAACTCTCTCAACTCTACAGTTGTACTTGTCGCGAACCTATCCCAATAATATTTTTTTTACAACTTTC	GAAAGTTGTAAAAAAAATATTATTGGGATAGGTTCGCGACAAGTACAACTGTAGAGTTGAGAGAGTTGGCAA
CT621	706852	706883	706923	TTGCCAACTCTCTCAACTCTACAGTTGTACTTGTCGCGAACCTATCCCAATAATATTTTTTTTACAACTTTC	GAAAGTTGTAAAAAAAATATTATTGGGATAGGTTCGCGACAAGTACAACTGTAGAGTTGAGAGAGTTGGCAA

Appendix I: MMCTPP1/L2b TSS matches

disc len	**seq32 ID**	**-10** **hex**	**spacer**	**-35 hex to tss**	**T-** **P**
5	CT007_112	TATGAT	17	TTGCTAAAAATTTTATTAAGCAGTATGATCTACCA	1
6	CT016_067	TACAAT	17	TTGTCAAAAATGTACCCCTTAACTACAATGCCGAGG	1
6	CT022_102	TAAAAT	17	GTGCATTTTTTCTTGCTTTTTCATAAAATGTTCGGG	0
6	CT025_060	TATCCT	18	TTGAAAATCAAGCTAATGATGCTGTATCCTCTGGGGA	0
9	CT054_076	TGCTAT	20	TTGCTTTTTTTGAAAAATAAAAATTTTGCTATGGGAATTTTC	1
6	CT080_071	TAAAAT	17	TTATGAAAAACAATTTTTTAATTTAAAATTAGAATA	0
7	CT091_071	TAACTT	17	TTGAGAAAAACATTTATATACGGTAACTTGCGAAGTA	1
4	CT098_072	TACACT	17	TTGCCTTTTTTAAGGTGAATATTTACACTACTCT	1
6	CT102_076	TAGGAT	17	TTGATTAATAAGCTTTGGCTTCGTAGGATGAGGGAC	0
5	CT111_132	TATCGT	17	TTGCAAAAAAGCGAGGACTTTGCTATCGTTCTTCC	0
6	CT125_060	TATAAT	17	TTGGAAAAAATAATCATCAAAATTATAATCATTCCC	1
8	CT140_135	TATAAT	17	TTGCCGGTTCTATTCTAGAAACGTATAATACTTCCCCA	1
7	CT141_061	TAATTT	17	TTGCAAAGAAGGTTCTTTGTAATTAATTTTACGAATG	1
8	CT149_121	TATTAT	17	TTGTTTTATCCAGTAATTTACCTTATTATGTTCTGCCA	1
6	CT150_071	TAGCAT	17	TTGATTATTTTTGAAAATAGGTATAGCATAGGGGCT	1
6	CT218_090	TAAAAA	17	TTGGTTTATTTTTCTTATTATTTTAAAAAGATCTAA	1
9	CT232_095	TATACT	17	TTGCTTGTAAGTCTTTTGCATGATATACTCCTTGGCCGA	1
5	CT235_088	TATATT	16	TTGCTAAGAAACAAAAAACCTCTATATTATCCCG	1
6	CT243_056	TAACAT	17	TTGATGATTCTTTTCAAAATAATTAACATGCGAAGC	0
6	CT249_060	TAAAAT	17	TTGATATTCGGTAAAAAATCAAGTAAAATGTTCGCC	1
5	CT253_129	TAGAGT	17	TTGCTTTTTGAAAATACTCATGCTAGAGTTCTCCT	1
12	CT259_102	TTTGCT	17	TTGCTTTCTTTTTTAAAAAAATCTTTGCTATACCTCCGAGAA	1
5	CT265_064	TAAAAT	17	TTGTTTTTGATTATTGTTTGTATTAAAATAACTCT	1
6	CT267_097	TATTAT	16	TTGAATCCAAAGGATGAATGCATATTATACGCATA	1
8	CT269_082	TAAACT	16	TTGACAACGAATATGTGTATAGTAAACTATTTGAGAA	1
8	CT286_067	TATATG	17	TTGCATCATTATCATAAATGTCGTATATGCTTGAAAAA	1
7	CT288_062	TAAAAT	17	TTGTAAAAAAACAATATTTATTCTAAAATAATAACCA	1
5	CT293_065	TAAAAT	18	TTGATTGGTTAAAAAAAATTACAATAAAATTATTGC	1
6	CT313_063	TAGGCT	17	TTGAATAGTATCGTTTTTTTTGGTAGGCTTCACTTA	1
7	CT323_149	TATAAT	20	TTGTTTGACATTTTCTGTTTAGTCGATATAATCGCTCTCT	1
6	CT327_096	TATGCT	18	TTGCTTTGATATAAATCTCTTGGATATGCTAATCTTC	0
6	CT342_102	TACAAT	18	TTGAAGCCTAAATAAAAGTGGTGTTACAATCCCCGGT	0
7	CT343_064	TATAAT	18	TTGAATTAAAACGGTTTTAACGGTTATAATCCTTTGTC	1
13	CT346_210	TATAAT	17	TTCAGAGAAAAATTATAACTTCCACTAAGCCTAAACACAAGAA	1
6	CT355_224	TAAAAG	17	TTGAGGATATAACGCTTTTTTGTTAAAAGTGTTCTG	1
6	CT367_306	TATAAT	17	TTGCAAAAAATCCATCGCGCTTGTATAATGCGTTGG	0
5	CT372_116	TACACG	18	TGGCAAAAAAAATCTTTTTTCCACTACACGGGTGGA	0
6	CT376_107	TATGTT	17	TTGATTCTTTTATCATCCAAACGTATGTTGGGACCA	0
6	CT377_075	TATAAT	20	TTGCAGAGTTTTTATTTTAAATATGTTATAATCTGTCTA	1
7	CT378_080	TACAAG	17	TCGCGAAAGATCACGAAAGATAGTACAAGTAAAAAGA	0
6	CT383_075	TAAAAT	18	TTGAAGACAAAGAAAAACTTTTGTTAAAATTTTTTCG	1
7	CT390_081	TATATT	17	TGGACAGATGAGAGTCTCATCTTTATATTACCGTCCA	1
6	CT392_071	TAAAAT	18	TTGATGTTTCTTTTGTTTGTTTCTTAAAATTAATTTA	0
6	CT393_071	TAAGAT	17	TTGATCTAGAAACACTCCTATGCTAAGATGCTCTTC	1
6	CT394_043	TATAAT	17	TTGACCAGTGGAGACGGTTTTCTTATAATGACACCG	1
6	CT413_073	TATCTA	17	TTGCCTAATTTACTTTTCTGATTTATCTAACGCCTA	1
7	CT444_130	TGTAAT	17	TTGATATAATTTTTATTTTATAATGTAATATTGTCTA	1
4	CT460_090	TAAAAG	18	TTGACGATAAACCTAGTTAAGGCATAAAAGAGTTG	0
8	CT489_056	TAAAAT	18	TGGCTTTTTTAATAATTTATTTTTTAAAATTATTTTTTA	0
6	CT496_153	TAAGCT	20	TTGTTTGTTTGAATGTTTTTTGTTGATAAGCTGGGGGAA	0
7	CT496_088	CATAAT	17	TTGCTTTTGGAGTGTCTATGTTTCATAATATGTGTCA	1
6	CT509_062	TAAGAT	17	TTGAAAAATAACAATTTTTGACCTAAGATGCTTATA	0
7	CT533_058	TAATAT	18	TTGCTTGCTAAAAAAAAAAAAGGATAATATACGGGGTC	1

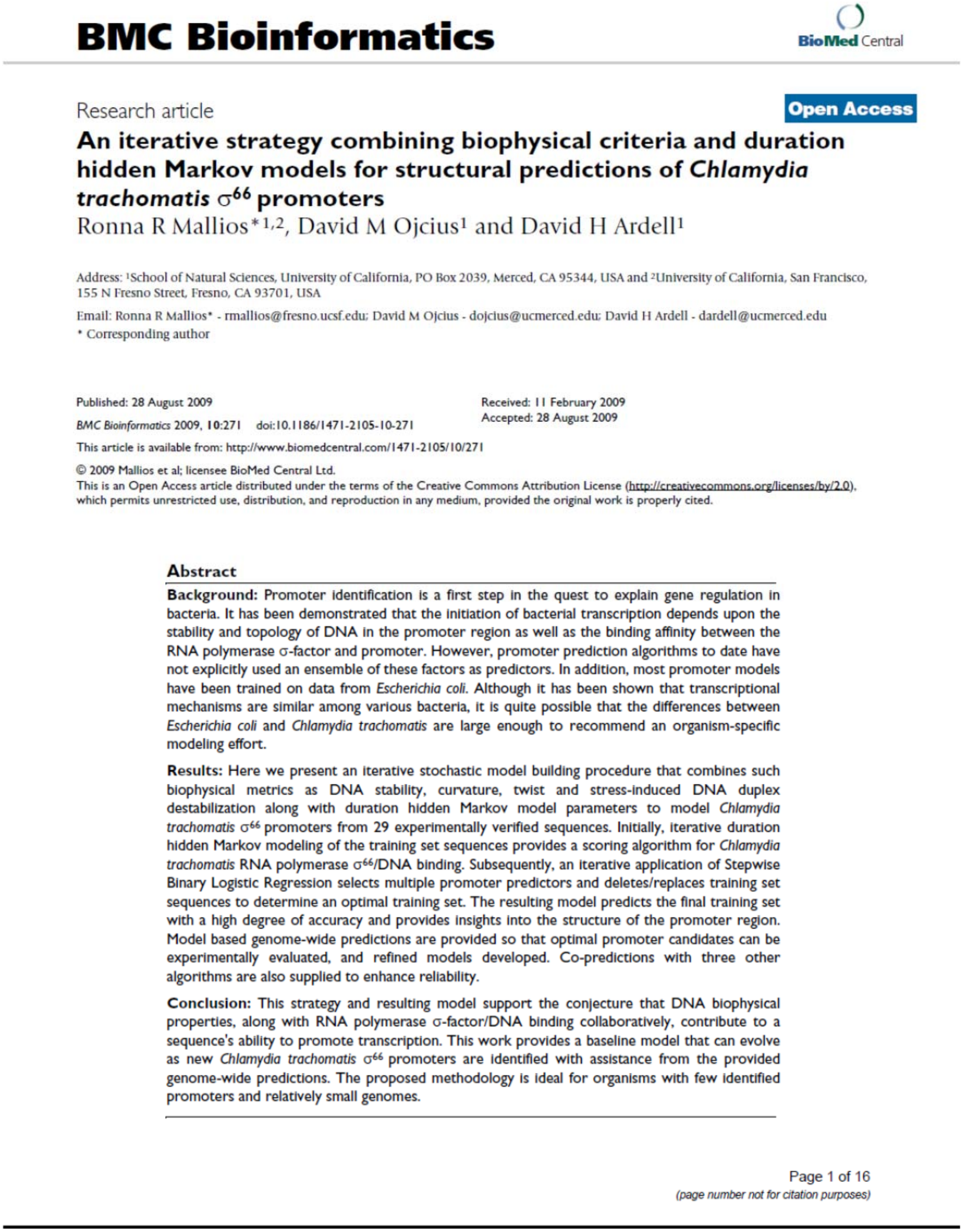
4	CT546_050	TAAAAT	17	TTGGCATTGCTGTTTTTATTTATTAAAATAAATA	1
5	CT547_065	TATGAT	18	TTGACAAATTCTCTTTTTCTTTTTTATGATGACGCT	1
5	CT556_137	TATAAT	16	TTGATTTTTTCCTCCTAGGAACTATAATTCGGGA	1
5	CT557_165	TACAAC	17	TTGAGATTTTATCCACCCAGATGTACAACCCGGGA	1
5	CT559_055	TATGGT	17	TTGGCACTAATCTCCCCATTTGCTATGGTGAGTGA	0
6	CT573_061	TACTTT	17	TTGATTTTTTTTCTCAAATCAGTTACTTTATACAAT	0
5	CT576_077	TAATAT	18	TTGTTAAATCAGATCGTTAGAATTTAATATTGTTAG	0
5	CT584_064	TACAAC	17	TTGATTAAAAAGTTACAAAAAGTTACAACTAACCT	0
6	CT596_066	TAGGAT	17	TTGGTTCTATACAAGAAATTTGTTAGGATCGTCTAG	0
6	CT608_127	TATAGT	18	TTGCCAAGACGGGGGGAGTTGCTCTATAGTAAAAGCC	1
7	CT618_126	TACAAT	17	TCGATTTAAAAGCGATTTCTTTTTACAATGCTTTCCC	0
6	CT621_069	TAGGTT	17	TTGTAAAAAAAATATTATTGGGATAGGTTCGCGACA	1
7	CT632_065	TATGTG	20	TTGCTGAAAAACTTTGAGTGTTTTGTTATGTGTGGTAGGC	1
8	CT634_040	TATAAT	17	TTGGATTCCCTTATAAAAACTTCTATAATCCCGAGATA	1
7	CT636_056	TATATT	17	TTGCTCTTTTTTGTTATTCGGCGTATATTTCCGGACT	1
9	CT658_074	TATCAT	17	TTGAATAAATCTTTTCCGAACCGTATCATGGAAGGGTTT	0
6	CT691_072	TATATT	18	TTGCAAATATATATATGAAGGAGGTATATTTTGGGAG	1
6	CT693_059	TATAAG	17	TTGAGTTTTCCTTTGCTTAGGCCTATAAGAAAATTT	1
6	CT708_069	TACAAG	17	TTGATTTAGCGGAAGTAAAAAGGTACAAGTAACAGG	1
6	CT709_090	GATGAT	17	TTGTAAAGAAAGTGATCAATTCTGATGATGAAGTCG	0
7	CT729_068	TATACT	17	TTGTTGCTCAGACAAAACTTCCATATACTCAACCTGA	1
6	CT731_110	TATACT	17	TCGCAGAAAGTAGAGGTTTGTGTTATACTCTGCGCA	1
6	CT733_230	TAGCAT	17	TTGCCCCTAACAAAAAATCATGTTAGCATGAAGCCG	0
6	CT740_073	TAACTT	17	TTGATTTTTTATAGAGTAACCTATAACTTGACGCTA	1
6	CT743_085	TAATTA	18	TTGCATGAATTTGAACAAACAAACTAATTAAAAATTA	0
6	CT757_077	TAAGAT	16	TTGCCTTTTGAAAGCTTAAGTTTAAGATAGAGAAT	1
8	CT763_097	TATATT	16	TTGACGCTTTTTTAGAATTTCATATATTCTTCCCACA	1
6	CT783_123	TACACG	18	TTGCTTTTAATGAAAAAAAGAATATACACGAAAAGTG	0
6	CT790_081	TAGAAA	17	TTGCTGTTAAAAATTTTTTGGCATAGAAATAGAGCT	1
6	CT794.1_056	TAGAAT	17	TTGCTTATTAGTTTCTTTTGTTATAGAATATTAGCT	1
5	CT798_070	TAATTA	17	TTGATTATTTTTGTTAAAAGAAATAATTAATGAGT	0
9	CT821_060	TAAAAG	16	TTGATTGAAGTAAAAAGAATAATAAAAGATAAGGAGGA	0
5	CT823_107	TATGCT	17	TTGATTTGCATCATTAGATTTTGTATGCTGCATAT	0
4	CT849_066	TAAAAT	17	TTATTAAAGAGAGAAATTGCTGGTAAAATAAAAA	1
8	CT854_064	TATGAT	17	TTGCCGCATATGCTCTCTTCCCCTATGATTCTTCCTTC	1
6	CT863_074	TAAGTT	17	TTGCATGAAAAATACTTTTTAGATAAGTTCCCTCCT	1

Appendix J: TSS-PREDICT only/L2b TSS matches

UW-3 num	-35 hex	spacer	-10 hex	disc len	off-set
CT005	ctccaa	15	tatact	7	2
CT013	gtgaca	19	tatact	5	-2
CT017	ttgact	17	aataat	6	1
CT021	ttgaca	18	tagtat	6	0
CT035	ttgata	15	tataat	7	4
CT043	tttact	17	taggtt	5	3
CT046	ttttca	15	gatcaa	4	2
CT047	ctgagt	18	taaaat	6	0
CT048	tgtata	19	tagaat	9	4
CT065	ttgcct	17	tatact	7	2
CT066	ttaata	19	tagaac	9	-1
CT067	ttgagg	17	taaaag	6	2
CT072	gtggtg	19	tagtat	6	0
CT082	ttcata	16	taaaat	13	6
CT099	tttaga	18	tataaa	8	2
CT113	ttgact	17	tatgga	8	-1
CT134	ttagga	18	tacaat	6	2
CT139	gtgtat	17	tatact	6	-1
CT147	tttata	16	aatact	5	2
CT148	ttcctt	17	tatact	7	1
CT153	tttaaa	17	tagggt	6	6
CT169	tatatt	15	tattat	5	2
CT177	tttttt	15	ctcaat	4	0
CT210	tttttt	17	tacact	6	1
CT214	ttttcc	17	tatatt	9	-2
CT229	ttattt	17	tataat	8	1
CT241	tagtca	17	tattgt	7	1
CT251	gtgcat	17	taaaat	6	0
CT254	ttgata	19	tatgat	6	-5
CT273	ttgact	16	tatgat	6	-1
CT275	ttgatt	18	tacatt	6	1
CT295	tttaca	17	tatagt	6	1
CT319	gtgtat	18	tataat	7	2
CT324	ttacct	18	tataaa	5	-2
CT332	ttgaag	17	tattct	5	-2
CT339	atgaca	17	tattct	7	1
CT340	tttttt	16	tagaaa	5	0
CT344	tgcact	16	tagtat	6	0
CT364	ttggag	18	tatact	4	-2
CT365	atgaaa	18	tataat	6	1
CT381	tttttt	17	tatagt	7	0
CT382.1	ttttaa	19	tataat	6	-1
CT398	tttact	17	taaact	7	2
CT415	ttgaac	15	tagaat	7	-1
CT446	ttgatt	17	gaaaat	6	-1
CT449	ttgcgg	18	taaaat	5	-1
CT482	cctaca	19	tacact	6	0
CT483	attcca	18	taaaat	7	0
CT490	ttgaca	17	ctttat	10	4
CT502	tagtaa	17	taaaat	4	-3
CT503	ttgctc	16	gacact	5	-3
CT529	ttttcg	19	tataat	7	2
CT543	tagtcc	17	taacat	6	1
CT588	ttgctt	16	gagagt	12	0

CT606	gtgaaa	18	tatcat	7	-2
CT617	tttcta	18	taaaat	6	0
CT622	tgggct	15	tataat	7	0
CT625	tttata	19	tctggc	11	0
CT630	ttgaaa	15	tattag	10	3
CT631	tcttca	16	tagact	5	-6
CT642	ttggca	17	tatggt	5	-1
CT645	tttaca	18	tatcct	4	3
CT652.1	ctggca	17	aaagat	4	4
CT655	gtacca	17	tatagt	5	0
CT656	gtgaaa	17	taaaat	6	2
CT711	tttaaa	16	taaaat	6	0
CT713	tttcct	18	tttgct	12	1
CT719	ttgttt	18	tacaga	8	5
CT723	tggact	17	tatgat	6	-1
CT734	tttagt	15	tataat	8	3
CT739	tcgatt	16	tataac	5	0
CT741	ttgatc	17	taacct	6	2
CT755	ttgtga	17	tagtgt	7	0
CT792	taaaaa	16	aatact	6	-4
CT818	ttgcta	16	tagtat	8	-4
CT832	ttcata	16	tatact	5	0
CT833	ttgata	18	taccat	4	-2
CT843	tgtaca	17	tataat	7	3
CT853	ttcact	16	tacact	5	-2
CT867	ttaact	19	tataat	6	3

Appendix K: BMC Bioinformatics publication



BMC Bioinformatics

BioMed Central

Research article

Open Access

An iterative strategy combining biophysical criteria and duration hidden Markov models for structural predictions of *Chlamydia trachomatis* σ^{66} promoters

Ronna R Mallios*[1,2], David M Ojcius[1] and David H Ardell[1]

Address: [1]School of Natural Sciences, University of California, PO Box 2039, Merced, CA 95344, USA and [2]University of California, San Francisco, 155 N Fresno Street, Fresno, CA 93701, USA

Email: Ronna R Mallios* - rmallios@fresno.ucsf.edu; David M Ojcius - dojcius@ucmerced.edu; David H Ardell - dardell@ucmerced.edu

* Corresponding author

Published: 28 August 2009

BMC Bioinformatics 2009, **10**:271 doi:10.1186/1471-2105-10-271

Received: 11 February 2009
Accepted: 28 August 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/271

Abstract

Background: Promoter identification is a first step in the quest to explain gene regulation in bacteria. It has been demonstrated that the initiation of bacterial transcription depends upon the stability and topology of DNA in the promoter region as well as the binding affinity between the RNA polymerase σ-factor and promoter. However, promoter prediction algorithms to date have not explicitly used an ensemble of these factors as predictors. In addition, most promoter models have been trained on data from *Escherichia coli*. Although it has been shown that transcriptional mechanisms are similar among various bacteria, it is quite possible that the differences between *Escherichia coli* and *Chlamydia trachomatis* are large enough to recommend an organism-specific modeling effort.

Results: Here we present an iterative stochastic model building procedure that combines such biophysical metrics as DNA stability, curvature, twist and stress-induced DNA duplex destabilization along with duration hidden Markov model parameters to model *Chlamydia trachomatis* σ^{66} promoters from 29 experimentally verified sequences. Initially, iterative duration hidden Markov modeling of the training set sequences provides a scoring algorithm for *Chlamydia trachomatis* RNA polymerase σ^{66}/DNA binding. Subsequently, an iterative application of Stepwise Binary Logistic Regression selects multiple promoter predictors and deletes/replaces training set sequences to determine an optimal training set. The resulting model predicts the final training set with a high degree of accuracy and provides insights into the structure of the promoter region. Model based genome-wide predictions are provided so that optimal promoter candidates can be experimentally evaluated, and refined models developed. Co-predictions with three other algorithms are also supplied to enhance reliability.

Conclusion: This strategy and resulting model support the conjecture that DNA biophysical properties, along with RNA polymerase σ-factor/DNA binding collaboratively, contribute to a sequence's ability to promote transcription. This work provides a baseline model that can evolve as new *Chlamydia trachomatis* σ^{66} promoters are identified with assistance from the provided genome-wide predictions. The proposed methodology is ideal for organisms with few identified promoters and relatively small genomes.

Background

Identifying mechanisms that regulate gene expression in bacteria is essential for understanding and eventually controlling their pathogenicity. All known bacteria share a well conserved transcriptional holoenzyme, RNA polymerase (RNAP). The RNAP is comprised of a 3-subunit catalytic core plus a variable σ-factor subunit that provides DNA binding specificity. One of these σ-factors, σ^{70} in *Escherichia coli*, participates in the transcription of a majority of genes including those with housekeeping functions.

E. coli is the best studied bacterial model with regard to promoter identification and prediction. As such, most promoter predictions are based upon the analysis of *E. coli* σ^{70} promoter data. The earliest collections of *E. coli* σ^{70} promoters revealed the -35 and -10 hexamer consensus motifs, TTGACA and TATAAT, that serve as recognition sites for the 2.4 and 4.2 domains of σ^{70} [1-3].

Position weight matrices (PWMs) were the first models to quantify the hexamer motifs [4]. PWM models were expanded to quantify the variable-length spacer region between hexamers [5-7], which is important in orienting the hexameric motifs for interaction with the sigma binding factors [8]. Challenges encountered by PWM models include defining thresholds that are sensitive enough to include known promoters without predicting numerous false positives.

Most of the quantitative modeling efforts that ensued require training sets comprised of both positive and negative sequences. Artificial neural networks (ANNs) [9] have been trained on sequences of identified *E. coli* promoters and non-promoters. A hidden layer in the ANN architecture quantifies interactions among pairs and triplets of nucleotides. The resulting ANN scans and scores overlapping sequences, and reports a score in the range (0, 1) that indicates the likelihood of the sequence being a promoter. A time-delay neural network (TDNN) can combine two simple ANNs (one for each hexamer) with a variable-length spacer region [10].

Burden *et al* (2005) [11] measured the distance from the transcription start site (TSS) to the translation start site (TLS) of 771 *E. coli* promoters. They showed that the distribution peaks sharply around 30 nt, and that combining the TSS-TLS distribution with the NNPP2.2 TDNN [10] significantly enhances the specificity of the prediction.

In another machine learning approach that has been applied to model promoters, support vector machines (SVMs) were trained on *E. coli* promoter sequences of length 200 [12]. Although the SVM approach has the advantage of comprehensively quantifying the primary structure of the upstream region, it does not examine structures of higher order that motivate our approach.

A natural extension to PWMs that explicitly models an empirical spacing distribution between motifs is given by duration hidden Markov models (HMMs). Here "duration" refers to this explicit representation of a spacer length distribution, as opposed to the geometrically distributed lengths that are expected from components of profile hidden Markov models [13]. Although the variable-length spacer region between hexamers has been incorporated into promoter modeling and predictors before [5-7], none of these earlier efforts have integrated an explicit probabilistic representation of the spacer distribution into a reusable predictor as a duration HMM, which is arguably its most natural representation. On the other hand, while duration HMMs have been introduced into genome analysis (for example, in intron-exon modeling, see Winters-Hilt 2006 http://www.biomedcentral.com/1471-2105/7/S2/S14), they have not to our knowledge been applied to modeling transcriptional or translational signals before.

Bacteria of the genus *Chlamydia* are obligate intracellular parasites that were genetically isolated from other bacteria nearly a billion years ago when they moved into their intracellular environment [14]. In humans, *Chlamydia* infections are responsible for infertility, blindness, arthritis and cardiovascular disease [15]. Because chlamydiae have an intracellular life-cycle, standard genetic techniques are often insufficient to study gene regulation [16]. Hence, only 30 to 40 promoters have been experimentally verified [16-19]. However, with a small genome of only about 1 Mbp and 895 genes, *Chlamydia trachomatis* (*CT*) makes a good candidate for *in silico* analysis.

Surveys of known bacterial promoters suggest that their structures are relatively diverse [8]. Additionally, established *CT* promoters display obvious differences from the established consensus hexamers of *E. coli* [16-19]. Although σ^{66}, the *CT* analog of *E. coli* σ^{70}, has DNA binding domains homologous to domains 2.4 and 4.2 in σ^{70}, sequence based phylogenetic analysis of bacterial RNAP subunits has shown discernable evolutionary distance between the *CT* and *E. coli* RNAP subunits [20]. Therefore, it is plausible that an organism-specific model is appropriate for *CT*.

Phylogenetic footprinting takes advantage of relative conservation of motifs among related species. Grech *et al* (2007) [17] developed an algorithm that combines *E. coli* trained PWMs and chlamydial phylogenetic footprinting. *CT* upstream regions are screened with the PWMs and the potential promoter hexamers are filtered with an algorithm that accepts only conserved sequences in a consensus of *C. trachomatis*, *C. pneumoniae* and *C. caviae*. Although this is a promising approach, because they used an *E. coli* trained PWM, their results may be strongly influenced by prior expectations that all bacterial promoters are structured as in *E. coli*. We believe that more development is needed in *ab initio* approaches for predicting promoters using sequence information directly from the organism under study (and perhaps from close phylogenetic relatives) in combination with biophysical metrics that derive from known models about the biology of transcription in general.

This study aims to develop *CT* promoter models using only known *CT* promoters. To do so, it considers DNA stability and topological features of the upstream region as well as RNAP σ-factor/DNA binding. As Hertz and Stormo (1996) [5] aptly wrote "... the polymerase needs to bind the DNA, open the DNA, initiate transcription, and release the promoter for elongation." The TDNNs and SVMs that consider extended promoter sequences are addressing this issue from a sequence perspective. This study utilizes measures that have been developed to quantify stability and other aspects of DNA structure. Evidence from the profiling of DNA curvature, bendability, twist, stability and propensity for stress-induced destabilization in *E. coli*, *B. subtilis*, *C. trachomatis*, plants and vertebrates [21-23] suggests that there are peaks for these measures near the TSS. Here we use a stochastic model building procedure that allows for the combination of relevant predictive measures selected from RNAP σ-factor/DNA binding propensity, as quantified by duration HMMs, and structural features of the upstream region, as quantified by biophysical metrics.

Methods

Stochastic Model Building

Stepwise Binary Logistic Regression (SBLR) [24,25], as implemented in SPSS version 17.0 statistical software (SPSS Inc., Chicago, IL), selects an optimal set of independent variables (continuous and/or categorical) to classify observations into two populations. Logistic regression does not assume a linear relationship between the dependent and independent variables, normal distributions, or homoscedasticity (equal variances). It does, however, assume independence of observations. We address this requirement in a separate section describing the selection of non-redundant observations.

The mathematical model (prediction equation) fitted by SBLR has the form

$$\mathbf{u} = b_0 + b_1 v_1 + b_2 v_2 + \ldots + b_i v_i,$$

where i is the number of steps, v_1 through v_i are the predictor variables selected, and b_0 through b_i are coefficients determined by the analysis.

u is the logit for the dependent variable, which means that

$$\begin{aligned}\mathbf{u} &= \ln(\text{odds(event)}) = \ln(\text{prob(event)}/\text{prob(nonevent)}) \\ &= \ln(\text{prob(event)}/[1 - \text{prob(event)}]).\end{aligned}$$

Here, the event is class membership. When P denotes the prob(class membership), the equation can be rewritten as

$$\mathbf{u} = \ln(P/(1-P); e^{\mathbf{u}} = P/(1-P); \text{ and } P = e^{\mathbf{u}}/(1+e^{\mathbf{u}}) = 1/(1+e^{-\mathbf{u}}).$$

Selecting a cutoff for P, most commonly 0.5, converts P into a classifier. When 0.5 is the probability threshold, $e^{\mathbf{u}}$ = 1 and the classification threshold for **u** is 0. The effectiveness of a model can be evaluated by its ability to correctly classify the training data.

The SPSS SBLR analysis procedure provides many user-defined options. We selected the Forward Conditional stepwise procedure for all analyses. At each step, a score statistic is calculated for each variable excluded from the model. The score statistic is based on Maximum-Likelihood Estimation criteria and is asymptotically distributed as a χ^2 variable [25]. The variable with the highest significant χ^2 value is entered into the model. If no significant variables remain, then the procedure stops with the current model. Similarly there is a mechanism for stepwise removal. After a new model has been generated, score statistics are calculated for all variables in the model. If the p-value for any variable in the model is greater than the probability for stepwise removal, then the variable is removed from the model. We retained the default probabilities for stepwise entry (.05) and removal (.10), thus ensuring that the significance of all model variables is less than 0.10.

Potential Observations and Dependent Variable

A significant challenge for bioinformaticians is to model data that has been collected by multiple laboratories using different assays, protocols and equipment. This phenomenon is compounded in the study of *CT* where the organism is metabolically active only inside an infected host-cell. One way to minimize the use of conflicting and/or controversial data is to rely upon reviews written by informed biologists. For this reason, we consulted the reviews of Mathews & Timms (2006) [19] and Tan (2006) [16] to compile a list of 16 experimentally verified σ^{66} promoters. To this list we added 13 promoters that were experimentally verified by Grech *et al* (2007) [17] and Hefty *et al* (2007) [18] after the previously cited reviews were written. For the purposes of this study, we consider these 29 sequences to be the known *CT* σ^{66} promoters.

Table 1 describes the 29 experimentally verified σ^{66} promoters from 27 genes that form the basis of the training set for this study. We derived potential observations for analysis according to the following procedure:

1. Files containing the *CT* genome (NC_000117.fna) and genome table (NC_000117.ptt) were retrieved from the NCBI website, ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/, on July 2, 2007 and last modified by NCBI on January 23, 2007. An R script was written to extract the 600 nt upstream regions of all 895 protein-coding genes as annotated in the genome table. The regions were verified at several stages throughout the research with other sources, including a *CT* genome database previously available from Los Alamos National Laboratories and comparable prediction algorithms. Table 1 displays the distance from promoter to TLS for all training set genes. Since the maximum distance is 296 nt, 325 nt was set as the upper limit for data analysis. Then, the upstream region was defined as 600 nt to allow for biophysical structures 275 nt upstream from a predicted promoter.

2. For each of the 27 genes listed in Table 1, the 600 nt upstream region was parsed into overlapping sliding window sequences of length 32 (6 nt for each hexamer and maximum spacer of 20 nt) and step-size 1. Each subsequence (SEQ32) was labeled according to its parent gene and position occupied in the upstream region: e.g. the first SEQ32 was labeled CT046_600 because the initial nt is found 600 nts upstream from the CT046 TLS.

3. The dependent variable, PROMOTER, was assigned a 1 if a promoter sequence listed in Table 1 was totally contained in SEQ32, and 0 otherwise. Thus, 1's identify potential promoter observations and 0's identify potential non-promoter observations.

4. Cases with upstream positions ≥ 40 and ≤ 325 were selected as potential observations to restrict the analysis to the range of the training set data. The upper bound is 30 nt upstream of the furthest upstream training set promoter and the lower bound is equal to the furthest downstream training set promoter.

Independent Variables

The primary variable for promoter prediction is the pattern that characterizes the binding between the RNAP σ-factor and DNA. Here we use duration HMMs to describe and quantify RNAP-σ^{66}/DNA binding. After a set of known promoters is used to train a duration HMM, the duration HMM scans a new sequence to identify the hexamer-spacer-hexamer subsequence that scores the highest with regard to potential binding. The variable HMM_SCORE is assigned the score associated with the highest scoring subsequence, while the variable START denotes the position of the lead nucleotide in the -35 hexamer and END denotes the position of the last nucleotide of the -10 hexamer.

Specifically, a training set of promoter sequences was placed in the file ts.txt. The initial ts.txt contained the contents of Table 1, columns "-35 Hex", Spacer and "-10 Hex". This file was supplied as input to **durahmmer** (Ardell D.H., in preparation) which was used to create a duration HMM with the command: **durahmmer** -5 6 -3 6 -s 16 -S 20 -p 1 -u 28.5:21.5:21.5:28.5 -C ts.txt > ts.hmm. The options to the command specify the following model parameters: 6 matched states (hexamers) at the 5' and 3' sequence ends; minimum and maximum spacer lengths of 16 and 20 respectively; a background compositional model of 28.5% A, 21.5% C, 21.5% G, and 28.5% T; and spacers should be modeled to have their empirical composition in the training set (which in this case was: 38% A, 12% C, 17% G, 33% T). The program **durahmmer** produces a valid HMMer 2.3.2 [13] model file representing a duration HMM. For the final model of this study, the model file and the input data file are provided as Additional Files 1 and 2. All 16,200 SEQ32 observations from the 27 genes were placed in the file all.txt so that optimal promoters and HMM scores could be calculated by **hmmsearch** [13] with the command: **hmmsearch** -E 9000 ts.hmm all.txt. We ran **hmmsearch** with a high E-value because we were interested in combining the score of the maximum scoring hit with other metrics in a composite procedure regardless of its magnitude.

In combination with the duration HMM model score described above, we also used the following biophysical metrics of promoter position and structure as possible independent variables for the SBLR model:

1. POSITION, which indicates the location of SEQ32 in the upstream region relative to the TLS. For CT046_101, POSITION = 101.

2. Measures of curvature (CURVE) [26] and %GC content (GC) for each 600 nt upstream region, which were determined by the online bend.it Server http://hydra.icgeb.trieste.it/dna/bend_it.html with a window-size of 32.

3. Free energy change (ΔG) of DNA melting (parameter #33 [27], dinucleotide, window size 2), bendability (parameter #31 [28], trinucleotide, window size 3) and twist angle (parameter #44 [29], dinucleotide, window size 2), which were determined for each 600 nt upstream region by the online plot.it Server http://hydra.icgeb.trieste.it/dna/plot_form.html. All measurements were then averaged over each SEQ32. ΔG always has a negative sign and is interpreted as greater values having lower stability. For statistical analysis this variable was transformed by STABLE = -ΔG so that the sign is always positive and the interpretation is that larger values have greater stability. Stability is also

Table 1: 29 experimentally verified σ^{66} promoters.

CT	Name	To TLS[a]	Ref[b]	-35 Hex	Spacer (16–20)	-10 Hex	h PI[c]
CT046	*hctB*	107	M	TGGTTA	GTTTTTAATAAAAAGT(16)	TAAAAA	16
CT062	*tyrS*	62	G	TTGCTA	TAAAAAGAACAGGATAGA(18)	TAAGAT	8
CT080	*ltuB*	68	M,T	TTATGA	AAAACAATTTTTTAATT(17)	TAAAAT	24
CT091	*yscU*	68	H	TTGAGA	AAAACATTTATATACGG(17)	TAACTT	8
CT098	*rs1*	69	M,T	TTGCCT	TTTTTAAGGTGAATATT(17)	TACACT	3
CT111	*groES*	129	M,T	TTGCAA	AAAAGCGAGGACTTTGC(17)	TATCGT	1
CT286	*clpC*	64	G	TTGCAT	CATTATCATAAATGTCG(17)	TATATG	8
CT322	*tuf*	296	M,T	TTGATA	ATAATCCGCGTCTGAAGT(18)	TACTAT	3
CT323	*infA*	145	M,T	TTGACA	TTTTCTGTTTAGTCGA(16)	TATAAT	3
CT377	*ltuA*	74	M,T	TGCAGA	GTTTTTATTTTAAATATGT(19)	TATAAT	16
CT394	*hrcA*	40	M,T	TTGACC	AGTGGAGACGGTTTTCT(17)	TATAAT	16
CT439m	*rpsL*	67	G	TTGCAA	ACAAAGATATTCTTATTC(18)	TATATT	3
CT442	*crpA*	66	M	GGGTTT	TTGAAAAAAACAAGTGTTT(19)	GTGTAG	16
CT444a	*omcA*	127	M,T	TTGATA	TAATTTTTATTTTATAA(17)	TGTAAT	16
CT444b	*omcA*	61	M,T	AATTGC	TTTTATCGATAAAAGAAAC(19)	TTCAAG	16
CT518	*rl14*	198	M	CTGTTG	TTGTTCGAGTCGAAAGGG(18)	TATACT	3
CT557	*lpdA*	162	H	TTGAGA	TTTTATCCACCCAGATG(17)	TACAAC	8
CT559	*yscJ*	52	G	TTGGCA	CTAATCTCCCCATTTGC(17)	TATGGT	16
CT576	*lcrH_1*	75	H	TTGTTA	AATCAGATCGTTAGAATT(18)	TAATAT	16
CT596	*exbB*	63	G	TTGGTT	CTATACAAGAAATTTGT(17)	TAGGAT	3
CT665	-	98	H	TTGTAT	CTTTTTAGAACGGGAAGGG(19)	TTGAAA	8
CT674	*yscC*	119	H	TTGCAA	GATAGAGGGCAAATAGA(17)	TATATT	16
CT681a	*ompA*	282	M,T	TATACA	AAAATGGCTCTCTGCTT(17)	TATTGC	8
CT681b	*ompA*	60	M,T	GTGCCG	CCAGAAAAAGATAGCGAG(18)	CACAAA	8
CT701	*secA_2*	57	M	TGTATA	GGCGCCTTTAAATAAGAGGG(20)	TAGGTT	8
CT708	-	66	G	TTGATT	TAGCGGAAGTAAAAAGG(17)	TACAAG	16
CT743	*hctA*	83	M,T	TTGCAT	GAATTTGAACAAACAAAC(18)	TAATTA	24
CT752	*efp_2*	62	G	TGGACA	AAGCTTAGAAGAGAACGA(18)	TAACAT	8
CT863	-	71	H	TTGCAT	GAAAAATACTTTTTAGA(17)	TAAGTT	16

[a] nt distance from the lead nt of the -35 hexamer to the TLS.
[b] References: M: Mathews & Timms [19]; G: Grech *et al* [17]; H: Hefty *et al* [18]; T: Tan [16]
[c] hour Post Infection of transcriptional activation [31]

of interest in the immediate downstream region, so positions 27-37 (STABLE27_37) and 1-37 (STABLE1_37) were quantified. Since the bendability measure increases with rigidity, it was renamed RIGID. The twist angle measurement, TWIST, was not transformed.

4. Possible times of expression onset include 1, 3, 8, 24 and 40 hours post infection (h PI). Mutually exclusive binary variables H1, H3, H8, H16, H24 and H40 were created to mark time of expression onset.

5. Stress-induced DNA duplex destabilization (SIDD) quantification utilizes structural and energetic properties of DNA to measure the propensity for strand separation under negative superhelical stress [22]. A low SIDD score indicates a high propensity for strand separation. SIDD measurements were determined by the WebSIDD server [30] http://www.genomecenter.ucdavis.edu/benham/sidd/websidd.php with default parameters except for Open Region Size = 63. Because Niehaus *et al* [31] have shown a time dependent response to chlamydial DNA supercoiling, interactions between the time of expression onset and SIDD were included [32]. The SIDD/hour of onset interaction is quantified by SIDD_H# = SIDD*H#.

6. For variables based on the entire SEQ32, lagged variables were created for the four non-overlapping upstream subsequences of length 32: e.g. for CT046_100, CURVE_L32 was set equal to the CURVE value of CT046_132, CURVE_L64 was set equal to the CURVE value of CT046_164; CURVE_L96 was set equal to the CURVE value of CT046_196; and CURVE_L128 was set equal to the CURVE value of CT046_228.

Selection of Non-redundant Observations from Potential Observations

As mentioned earlier, SBLR assumes independent observations. To address this requirement, we select for analysis a subset of the overlapping potential observations that are non-redundant with respect to the pair of hexamers that are most likely to bind the RNAP σ-factor.

Table 2 displays the first six columns of a portion of the data file used for analysis. Each potential observation occupies a row. A row includes: the SEQ32 label (SEQ_ID); the SEQ32 literal sequence (SEQ32); the score of the optimal HMM instance in SEQ32 (HMM_SCORE); the position of the lead nt in the -35 hexamer of the optimal HMM instance (START); the position of the last nt in the -10 hexamer of the optimal HMM instance (END); and PROMOTER as previously defined.

If we select only those cases where END = 32, we eliminate all of the redundant optimal HMM hexamer pairs while retaining most optimal HMM instances (information). Table 2 demonstrates how this selection ensures that neighboring optimal HMM instances that match are included only once. Six potential observations, CT046_111 through CT046_106, all contain the verified promoter with hexamer pair TGCTTA and TAAAAA. Consequently, they all have PROMOTER = 1 and HMM_SCORE = -2.1. But only CT046_111 has END = 32 and is selected to represent the verified CT046 promoter. Similarly, only CT046_117 represents the maximal non-promoter hexamer pair TTGTGT and AAAACT with score = -5.9. This process incidentally aligns each selected SEQ32 such that the optimal downstream hexamer is at the far right end.

This selection process does not eliminate overlapping sequences, but it does eliminate overlapping likely binding sites. CT046_111 and CT046_112 overlap a great deal. However, the last hexamer of CT_046_111 (TAAAAA) is not present in CT046_112 and the first hexamer of CT_046_112 (GTGTGT) does not appear in CT046_111.

It should be noted that although each training set gene begins with the same number of potential observations, this selection process causes the number of selected non-redundant observations to differ among genes. Each gene starts with around 5 potential observations with PROMOTER = 1 for each verified promoter, and around 325-40-5 = 280 potential observations with PROMOTER = 0. However, selection for non-redundant observations always results in the number of designated non-promoters being reduced to approximately 90.

While selecting sequences with non-redundant HMM_SCORES does mitigate the problem of dependent observations, it may not entirely eliminate it. While there are numerous studies that affirm the robustness of Baysian Discriminant Analysis with regard to violating the assumptions of a linear relationship between the dependent and independent variables, normal distributions, and homoscedasticity [33], we could not find similar studies regarding the robustness of logistic regression. An alternative to the current analysis would be to use Stepwise Discriminant Analysis, knowing that we are violating some assumptions.

There are versions of logistic regression, including generalized estimating equations (GEE) [25], that are specifically designed for correlated data such as longitudinal studies. In these procedures there are subject variables and within-subject variables. It might be possible to force this study data into such a format, but as yet there are no readily available stepwise procedures to scan multiple possible

Table 2: Selecting rows with END = 32 (*) ensures non-redundant observations with regard to hexamers and HMM_SCORE.

SEQ32_ID	PRO-MOTER	START	END	HMM_SCORE	SEQ32:bold italics locates optimal HMM instance
CT046_117	0	4	* 32	-5.9	TAA*TTGTGT*GTGGTTAGTTTTTAATA*AAAAGT*
CT046_116	0	3	31	-5.9	AA*TTGTGT*GTGGTTAGTTTTTAATA*AAAAGT*T
CT046_115	0	2	30	-5.9	A*TTGTGT*GTGGTTAGTTTTTAATA*AAAAGT*TA
CT046_114	0	1	29	-5.9	*TTGTGT*GTGGTTAGTTTTTAATA*AAAAGT*TAA
CT046_113	0	2	29	-13.7	T*GTGTGT*GGTTAGTTTTTAATAA*AAAGTT*AAA
CT046_112	0	1	* 32	-11.4	*GTGTGT*GGTTAGTTTTTAATAAAAAG*TTAAAA*
CT046_111	1	5	* 32	-2.1	TGTG*TGGTTA*GTTTTTAATAAAAAGT*TAAAAA*
CT046_110	1	4	31	-2.1	GTG*TGGTTA*GTTTTTAATAAAAAGT*TAAAAA*C
CT046_109	1	3	30	-2.1	TG*TGGTTA*GTTTTTAATAAAAAGT*TAAAAA*CT
CT046_108	1	2	29	-2.1	G*TGGTTA*GTTTTTAATAAAAAGT*TAAAAA*CTA
CT046_107	1	1	28	-2.1	*TGGTTA*GTTTTTAATAAAAAGT*TAAAAA*CTAA
CT046_106	0	3	31	-11.9	GG*TTAGTT*TTTAATAAAAAGT*TAAAAA*CTAAC
CT046_105	0	2	30	-11.9	G*TTAGTT*TTTAATAAAAAGT*TAAAAA*CTAACC
CT046_104	0	1	* 32	-7.6	*TTAGTT*TTTAATAAAAAGTTAAAAAC*TAACCA*
CT046_103	0	4	* 32	-7.8	TAG*TTTTTA*ATAAAAAGTTAAAAACT*AACCAT*
CT046_102	0	3	31	-7.8	AG*TTTTTA*ATAAAAAGTTAAAAACT*AACCAT*T
CT046_101	0	2	30	-7.8	G*TTTTTA*ATAAAAAGTTAAAAACT*AACCAT*TT

predictors. A final alternative would be to select non-overlapping sequences with the penalty of losing information and perhaps introducing a selection bias.

SBLR is a procedure for model identification. It is only after a model has been identified that it can be evaluated for independence. Given that, we elected to analyze the non-redundant observations with SBLR and then examine the error terms for independence. In Time Series Analysis (which this analysis most resembles), this is done by checking that the error term is normally distributed with zero mean, and that autocorrelations and partial autocorrelations of the error term are not significant [34].

Iterative Modeling Strategy

Sources of error that could lead to misclassification include (i) imprecise laboratory procedures in defining and identifying promoters (including false positive promoters), (ii) presence of more than one promoter population, (iii) failure to include relevant predictor variables, and (iv) random variation. To minimize the first two error sources, an iterative strategy was developed. Duration HMM iteration (Figure 1) addresses error source (i), while SBLR iteration (Figure 2) addresses source (ii).

Duration HMM Iteration (Figure 1)

Minor modifications in the configuration of the training set promoters can improve classification accuracy. To accomplish this, we allowed each promoter to vary within a neighborhood that extends the sequence by 5 nts on each side. A limit of 5 nts ensures that a modified hexamer will not locate completely outside of the original promoter sequence. For example, when the promoter CT377 is extended, it becomes **TTGTT**TGCAGAGTTTTTATTTTAAATATGTTATAAT**CTGTC**, with the bolded nts marking the extensions. Initially, a duration HMM is determined

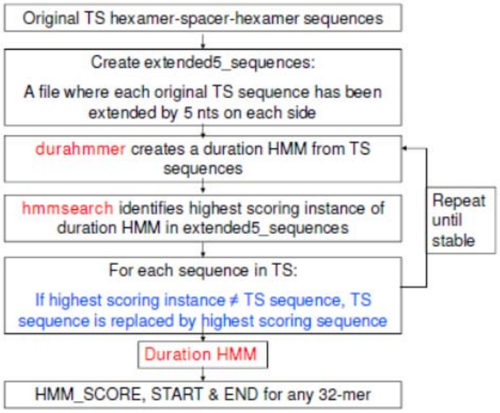

Original TS hexamer-spacer-hexamer sequences
Create extended5_sequences:
A file where each original TS sequence has been extended by 5 nts on each side
durahmmer creates a duration HMM from TS sequences
hmmsearch identifies highest scoring instance of duration HMM in extended5_sequences
For each sequence in TS:
If highest scoring instance ≠ TS sequence, TS sequence is replaced by highest scoring sequence
Repeat until stable
Duration HMM
HMM_SCORE, START & END for any 32-mer

Figure 1
Flowchart of duration HMM iteration.

by the original, non-extended, promoter training set. Then the set of extended promoters is searched for the highest scoring instance of the duration HMM in each extended sequence. If a high-scorer is not the same as the original promoter, it replaces the original in the training set. The iteration continues until stabilization. For the final model, CT377 was modified to TTGCAGAGTTTT-TATTTTAAATATGTTATAAT.

SBLR Iteration (Figure 2)

Deletion and subsequent replacement of members of the training set can eliminate promoters that are likely to be members of a different promoter population. This is accomplished via the iterative scheme diagrammed in Figure 2. Initially, the complete set of 29 verified promoters determines the duration HMM and the independent observations selected for SBLR analysis. SBLR delivers a mathematical model that produces a predicted probability of class membership (P) for each observation. A threshold on P of .5 is used to classify each observation as a predicted promoter or non-promoter.

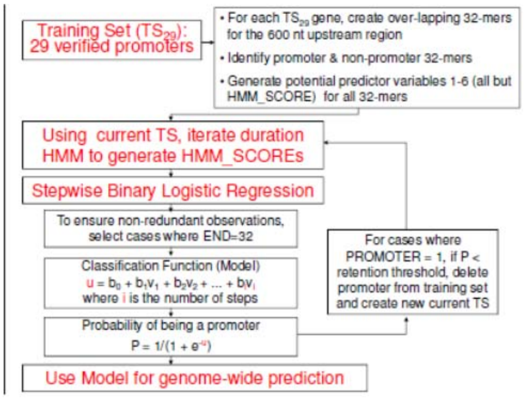

Training Set (TS_{29}): 29 verified promoters
• For each TS_{29} gene, create over-lapping 32-mers for the 600 nt upstream region
• Identify promoter & non-promoter 32-mers
• Generate potential predictor variables 1-6 (all but HMM_SCORE) for all 32-mers
Using current TS, iterate duration HMM to generate HMM_SCOREs
Stepwise Binary Logistic Regression
To ensure non-redundant observations, select cases where END=32
Classification Function (Model)
$u = b_0 + b_1v_1 + b_2v_2 + ... + b_iv_i$
where i is the number of steps
Probability of being a promoter
$P = 1/(1 + e^{-u})$
For cases where PROMOTER = 1, if P < retention threshold, delete promoter from training set and create new current TS
Use Model for genome-wide prediction

Figure 2
Flowchart of Stepwise Binary Logistic Regression iteration.

For those 29 cases where PROMOTER = 1, we also use the value of P to determine when a promoter appears to be an outlier and should be eliminated from the training set. After observing the 29 probabilities, a retention threshold on P between 0 and .1 is established. If a training gene has only one identified promoter and that promoter has a P less than the retention threshold, then all observations for that gene are deleted from the analysis. Similarly, if a training set gene has two identified promoters and they are both selected for deletion, all observations for that gene are deleted. However, if a training set gene has two identified promoters and only one is selected for deletion, all upstream observations for that gene remain in the analysis dataset and only observations within the remaining promoter are assigned PROMOTER = 1.

Modifying the training set in any way necessitates the determination of a new duration HMM, which in turn determines which observations will be aligned such that END = 32 and subsequently included in the next SBLR analysis. The iteration process continues until the training set stabilizes.

Stratified K-fold Cross-Validation

Once the final training set and model are selected, it is necessary to validate the model to ensure against over-fitting and to allow for comparisons with algorithms trained on other datasets. In the case of dichotomous classification, stratified K-fold cross-validation [35] partitions the training set into K subsamples such that each subsample has approximately the same proportions of class membership. Here we designate each training gene as a subsample; hence K equals the number of genes in the training set. Then, one gene (1-2 promoters and approximately 90 non-promoters) is retained as a validation set while the remaining genes are used as training data. Evaluation measures are calculated by aggregating the results of each validation set.

Comparable Algorithms

The following three algorithms were used to compare performance and to identify co-predictions with the model developed in this study: NNPP2.2, TSS-PREDICT, and Footy. NNPP2.2 [10] is an online time-delay neural network that is accessible for promoter predictions at http://www.fruitfly.org/seq_tools/promoter.html. We used the following options: organism = prokaryote and minimum promoter score = 0.95 to define promoters in the 325 nt upstream region of all *CT* genes. For the support vector machine algorithm TSS-PREDICT [12], the top two ranking predictions for each *CT* gene are posted as supplementary material at doi:10.1016/j.combiolchem.2008.07.009. The 42 *CT* promoters predicted by Footy [17], an algorithm that utilizes phylogenetic footprinting, are reported directly in the publication that describes the algorithm.

R scripts scanned the promoters predicted by NNPP2.2 and TSS-PREDICT for matches with the promoters predicted by the study model. An NNPP2.2 match was declared when the study prediction was contained within the 50 nt NNPP2.2 prediction. A TSS_PREDICT match was declared when the TSS_PREDICT predicted hexamer pair was contained within the study prediction.

Results

Finding the Best Model

The initial model, M0, utilizes the initial training set of 29 promoters with observations from their 27 parent genes. The duration HMM model converged after one iteration, modifying the alignment of 7 promoters. For all models, Table 3 reports the variables that were selected for the model and evaluation measures. If TP = true positive, FP = false positive, TN = true negative and FN = false negative, then sensitivity or recall = TP/(TP+FN), specificity = TN/(FP+TN), positive predictive value (PPV) or precision = TP/(TP+FP), negative predictive value (NPV) = TN/(FN+TN), and accuracy = (TP+TN)/(TP+TN+FP+FN). The total number of observations for each model differs according to the promoter training set being used.

For model M0, 19 of the 29 promoters were classified correctly, with 2 false positives. There is always the possibility that these are yet to be recognized promoters, but at this point they are counted as misclassifications. For the 10 verified promoters that were missed, the predicted probabilities ranged from 0.001 to 0.42. Since a natural separation appeared to between 0.07 and 0.10, P = 0.08 was selected as the retention threshold and promoters CT665, CT681a, CT681b and CT743 (along with all observations from their parent genes) were deleted from the training set for the next model, M1.

Table 3: Models produced by Stepwise Binary Logistic Regression Iteration and M2 Cross-Validation.

SBLR Model	M0	M1	M2	M3	M2 Cross-Validation
Training Set Deletion	none	CT665 CT681a CT681b CT743	CT665 CT681a CT681b	CT681a CT681b	CT665 CT681a CT681b
Variables in Model[a]	+HMM_SCORE +STABLE1_37 -POSITION +CURVE_L32 -GC_L128 +RIGID_L96 +CURVE	+HMM_SCORE +STABLE1_37 -GC_L32 -POSITION +CURVE_L32 -CURVE_L64 -GC_L128 +TWIST	+HMM_SCORE +STABLE1_37 -POSITION +CURVE_L32 -STABLE_L32 +SIDD_H24 -CURVE_L128 -SIDD_L128 +RIGID_L96	+HMM_SCORE +STABLE1_37 -POSITION +CURVE_L32 -STABLE_L32 -STABLE27_37 +CURVE	
Sensitivity or Recall	19/29 (0.655)	25/25 (1.0)	26/26 (1.0)	25/27 (0.926)	23/26 (0.885)
Specificity	2426/2428 (0.999)	2083/2083 (1.0)	2226/2226 (1.0)	2322/2323 (1.0)	2215/2226 (0.995)
PPV or Precision	19/21 (0.905)	25/25 (1.0)	26/26 (1.0)	25/26 (0.962)	23/34 (0.676)
NPV	2426/2436 (.996)	2083/2083 (1.0)	2226/2226 (1.0)	2322/2324 (0.999)	2215/2218 (0.999)
Accuracy	2445/2457 (0.995)	2108/2108 (1.0)	2252/2252 (1.0)	2347/2350 (0.999)	2238/2252 (0.994)
AUC[b]	0.995	1.0	1.0	0.999	0.992

[a]The variables are listed in order of entrance into the model and the sign indicates the sign of the coefficient.
[b]ROC analysis Area Under the Curve

The duration HMM model for M1 converged after two iterations, modifying the alignment of 5 promoters. Table 3 shows that M1 classified the modified training set perfectly, indicating that perhaps too many promoters had been deleted from the original training set. The retention threshold was reset to 0.07 and CT743 was reinstated for model M2.

The duration HMM model for M2 converged after one iteration. Table 4 displays the alignments of the 6 promoters that were modified. M2 also classified the modified training set perfectly. Again the results indicated that the next model, M3, should reset the retention threshold to 0.06 and reinstate CT665. However, Table 3 reports that M3 is not as good as models M1 and M2 because of classification errors.

Given two models, one training set a subset of the other, that both classify their respective training sets with 100% accuracy, we reasoned that the model trained on the largest set would provide the most sensitive genome-wide prediction. Thus, M2 was selected as the best and final model because of the perfect classification with the largest training set. The complete data file used to build M2, Additional File 3, is supplied so that others may replicate or modify the model.

Finally, the error terms of M2 were checked for independence. Residuals, PROMOTER - P, were calculated for all selected observations and shown to be normally distributed with zero mean. Additionally, the autocorrelations and partial autocorrelations of the residuals were not significant. Thus, the independence assumption of SBLR was not violated by this model.

Aggregated results of the stratified K-fold (25-fold) M2 cross-validation are reported in the last column of Table 3. For the 25 genes in the M2 training set, 3 promoters (CT322_298, CT743_085, and CT752_064) were not identified (sensitivity = 0.885) and there were 11 false-positive predictions (precision = 0.676). The incorrect classifications are most likely due to incomplete representation of the sample space, but may indicate additional populations or absent predictors.

Table 5 compares the performance of the stratified K-fold cross-validation performance of the M2 model with that of comparable algorithms when predicting promoters in the 25 cross-validation genes. The tally is in the form hits/predictions/gene. For NNPP2.2, a prediction was considered a hit if the hexamer pair in Table 1 was fully contained in the 50-mer NNPP2.2 prediction using a threshold of 0.95. The last two rows of the table show the cumulative sensitivity and precision of each prediction algorithm. M2 cross-validation is the most sensitive (0.885), while Footy is the most precise (1.0). Table 6 reports the hits and misses for the 2 genes that were not

Table 4: M2 duration HMM sequence alignment modifications.

CT	Name	To TLS	-35 Hex	Spacer (16-20)	-10 Hex
CT323	*infA*	145	TTGACA	TTTTCTGTTTAGTCGA(16)	TATAAT
		149	TTGTTT	GACATTTTCTGTTTAGTCGA(20)	TATAAT
CT377	*ltuA*	74	TGCAGA	GTTTTTATTTTAAATATGT(19)	TATAAT
		75	TTGCAG	AGTTTTTATTTTAAATATGT(20)	TATAAT
CT442	*crpA*	66	GGGTTT	TTGAAAAAAACAAGTGTTT(19)	GTGTAG
		60	TTGAAA	AAAACAAGTGTTTGTG(16)	TAGACT
CT444b	*omcA*	61	AATTGC	TTTTATCGATAAAAGAAAC(19)	TTCAAG
		59	TTGCTT	TTATCGATAAAAGAAAC(17)	TTCAAG
CT518	*rl14*	198	CTGTTG	TTGTTCGAGTCGAAAGGG(18)	TATACT
		195	TTGTTG	TTCGAGTCGAAAGGGTA(17)	TACTCG
CT701	*secA_2*	57	TGTATA	GGCGCCTTTAAATAAGAGGG(20)	TAGGTT
		61	TTGTTG	TATAGGCGCCTTTAAA(16)	TAAGAG

used in the M2 model. The only hit was scored by NNPP2.2, with 2 accompanying false positives.

Model Interpretation

The M2 duration HMM describes and quantifies the RNAP-σ^{66}/DNA binding observed in the training set. A visualization of the M2 parameters is shown in Figure 3. The -35 hexamer is dominated by the initial TTG motif, while the initial T with frequent As and Ts describe the -10 hexamer. The C and G compositions (12% and 17%, respectively) of the spacer region are much smaller than those of the genome (21.5% each). Spacer lengths of 17 predominate, while spacers of length 19 are absent.

The input data file for **durahmmer** (ts1.txt) and the resulting output data file (ts1_hmm.txt) are provided as Additional Files 1 and 2. The output data file is an HMMer 2.3.2 model file which supplies the parameters of the M2 duration HMM to **hmmsearch**. Complete documentation for the contents of the file can be found in the HMMER User's Guide at http://www.psc.edu/general/software/packages/hmmer/. Briefly, the first 17 lines are header information with the main model section following. There are 3 lines for each of the 32 possible nodes. The first and last 6 nodes refer to the -35 and -10 hexamers, while nodes 7 through 26 refer to possible spacer positions. The first line for each node displays the contribution to the final score (multiplied by 10^3) for the corresponding nucleotide matching A, C, G or T. The third line is particularly relevant to nodes 22 through 25, which correspond to spacer nucleotides 17 through 20. As nucleotides in these positions may or may not be present in the sequence being scored due to variable spacer length, the third line provides the odds of transitioning to another spacer nucleotide or to the -10 hexamer.

The M2 prediction equation generated by SBLR is:

$$\begin{aligned}\mathbf{u} = &-1408.301 + 85.305 * \mathrm{HMM\_SCORE} + 1816.454 * \mathrm{STABLE1\_37} - \\ &1.399 * \mathrm{POSITION} + 23.330 * \mathrm{CURVE\_L32} - 408.085 * \mathrm{STABLE\_L32} + \\ &25.445 * \mathrm{SIDD\_H24} - 13.757 * \mathrm{CURVE\_L128} - 21.675 * \mathrm{SIDD\_L128} + \\ &45.042 * \mathrm{RIGID\_L96}\end{aligned}$$

Being the strongest predictor, HMM_SCORE is selected in the first step of the SBLR procedure. The prediction equation for step one is

$$\mathbf{u} = -0.237 + 0.700 * \mathrm{HMM\_SCORE}$$

Using a classification cutoff of P = 0.5 and setting u = 0 yields HMM_SCORE = 0.339 as the threshold for step 1 classification. At step 1, 14/26 promoters and 2220/2226 non-promoters were classified correctly. Thus, the remaining eight model variables moved 12 promoters with HMM_SCORE < .339 to promoter classification and 6 non-promoters with HMM_SCORE ≥ .339 to non-promoter classification (without altering the classification of the previous 2234 observations).

The predictor variables and their coefficients describe the verified promoters and their upstream regions. Promoters have high HMM_SCORE and low POSITION. The near upstream region is curved and unstable, whereas the further upstream region is uncurved and unstable under superhelical stress. For late-cycle genes where expression onset occurs at 24 h PI, the effect of superhelical stress is less than at other times (a positive SIDD coefficient indicates there is little destabilization of DNA under superhelical stress). The upstream characteristics may reflect transcription factor binding and/or additional interaction with the RNAP holoenzyme.

The interpretation of the positive coefficient for STABLE1_37 is more subtle. In the second step of the SBLR, four observations change from FP to TN and 5 observations change from FN to TP. The means of STABLE, STABLE1_37 and STABLE33_37 are all larger in the second group than in the first. Although STABLE33_37 shows the greatest mean difference, the most statistically significant is STABLE1_37.

Model Exercise: Predicting Promoters for the **CT** *Genome*

Finally, the M2 model was used to predict promoters for the entire *CT* genome. Additional File 4 reports 479 predicted promoters in 361 unique genes, along with their HMM scores and genome locations. Thus, for 534 of the total 895 *CT* genes, this model does not find any 32-mers with a probability > 0.5. This suggests a conservative prediction that emphasizes specificity over sensitivity. Other explanatory factors may include alternate binding patterns for σ^{66}, alternative σ-factors, and operon configurations.

There was a substantial overlap among predictions by different methods. Additional File 5 lists the 209 promoters (176 unique genes) co-predicted by M2 and NNPP2.2, while Additional File 6 lists the 175 promoters (162 unique genes) co-predicted by M2 and TSS-PREDICT. Additional File 7 reports the 98 promoters (90 unique genes) co-predicted by M2, NNPP2.2 and TSS-PREDICT. All predictions are for 40 = POSITION = 325, consistent with the range of the modeling procedure.

Of the 42 promoters predicted by Footy, 11 were members of the M2 training set, 4 (CT265_111, CT342_102, CT547_065 and CT606_149) were co-predicted by M2 and NNPP2.2, and 6 (CT267_097, CT269_82, CT446_245, CT546_050, CT646_071, and CT837_088) were predicted by all four algorithms.

Characteristics of the M2 genome-wide prediction can be summarized by looking at all 479 predictions, or by look-

Table 5: Comparison of M2 Cross-Validation and predictions of comparable algorithms for 25 training set genes.

CT	HMM2 SCORE	M2 Cross-Validation	NNPP2.2	TSS-PREDICT	Footy
CT046	-1.6	1/1	0/4	0/2	0/0
CT062	4.0	1/2	0/0	1/1	1/1
CT080	0.5	1/2	1/4	0/2	0/0
CT091	1.3	1/3	1/1	1/1	0/0
CT098	3.7	1/1	0/1	1/2	1/1
CT111	-1.0	1/1	1/3	0/2	1/1
CT286	1.2	1/1	1/2	1/1	1/1
CT322	-2.1	0/0	0/0	0/2	0/0
CT323	1.6	1/1	1/3	1/1	1/1
CT377	5.3	1/2	1/3	1/1	0/0
CT394	-1.5	1/1	1/2	1/1	0/0
CT439m	1.8	1/1	0/3	0/0	1/1
CT442	-0.9	1/2	1/1	1/1	0/0
CT444	3.2	2/5	2/5	1/2	0/0
CT518	-4.2	1/1	0/0	1/1	0/0
CT557	-3.4	1/1	0/1	1/1	0/0
CT559	-2.7	1/1	1/1	0/2	1/1
CT576	0.6	1/2	1/3	1/2	0/0
CT596	0.5	1/1	0/1	0/2	1/1
CT674	4.0	1/2	1/2	0/0	0/0
CT701	-3.3	1/1	1/2	0/2	0/0
CT708	2.6	1/1	1/2	1/1	1/1
CT743	-3.8	0/0	0/2	1/5	0/0
CT752	-3.5	0/0	1/1	0/2	1/1
CT863	4.0	1/1	1/1	1/1	0/0
Sensitivity		23/26 (0.89)	17/26 (0.65)	15/26 (0.58)	10/26 (0.39)
Precision		23/34 (0.68)	17/48 (0.35)	15/38 (0.40)	10/10 (1.0)

Table 6: Comparing predictions of M2 and other algorithms for 2 training set genes not in M2 training set.

CT	M2	NNPP2.2	TSS-PREDICT	Footy
CT665	0/1	1/3	0/2	0/0
CT681	0/1	0/1	0/2	0/1

ing at the 361 unique genes, and selecting the predictions closest to the TLS. The two views produce similar results. Approximately 64% of predicted promoters are completely contained in non-coding upstream regions, 50% are on the positive strand, and time of activation distributes as follows: 5% hour 1, 23% hour 3, 51% hour 8, 20% hour 16 and 2% hour 24. The strand and hour distributions for all 895 genes in the genome are equivalent to the predicted promoter distributions, indicating that there is no strand or temporal preference for the predicted *CT* σ^{66} promoters.

Figure 4 displays a histogram of predicted promoter positions. POSITION marks the 5' end of the data file 32-mer, and is consequently ~40 nt upstream from the TSS. Thus, the POSITION distribution peaks with the 5' end around 68 nts upstream from the TLS and the TSS around 28 nts upstream from the TSS. The peak and shape of this distribution closely resemble the *E. coli* histogram from Burden *et al* (2005) [11].

Discussion

The final model produced by the iterative strategy was generated by a training set with three of the original members, CT665, CT681a and CT681b, removed. An explanation of how these three sequences differ from the remainder would be informative. The last column of Table 1 reports that CT665 and CT681 are both expressed at 8 h PI, classifying them as mid-cycle genes. Niehus *et al* (2008) [36] recently demonstrated that chlamydial promoters show a differential response to changes in DNA supercoiling that correlates with the lifecycle expression pattern. Specifically, two mid-cycle genes (8 h PI) responded to supercoiling, while three late-cycle genes (≥ 16 h PI) did not. Their experimental set included *ompA*/CT681 in the mid-cycle group and *omcA*/CT444, *hctA*/CT743 &*ltuB*/CT080 in the late-cycle group. Thus, it is likely that there exists a set of mid-cycle promoters that differ topologically from other promoters to enhance their ability to respond to supercoiling, and this may explain the anomolous characteristics of these promoters that we observed.

A possible explanation for the large number of genes without promoter predictions by the M2 model is heterogeneity requiring different models, for example for response to supercoiling. While investigating the initial model M0, we explored stepwise nominal regression, which allows for the discovery of more than two dependent variable categories. However, we did not find that a third category was

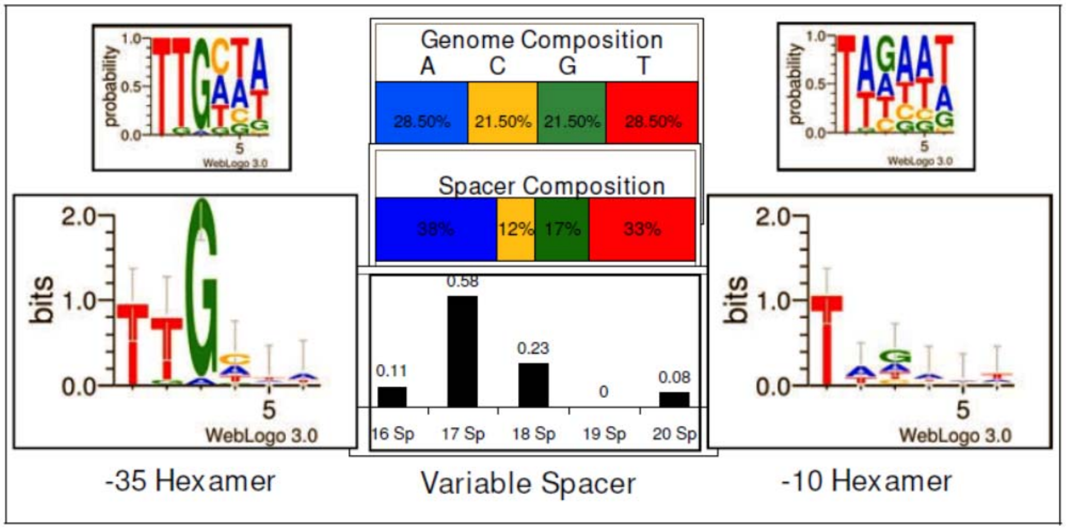


Figure 3

Visualization of the M2 duration HMM. The top WebLogos illustrate nucleotide frequencies in each of the hexamer positions. The bottom WebLogos convert the frequencies to bits of information.

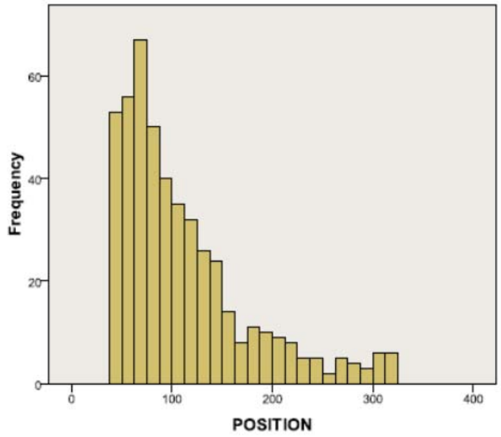


Figure 4
Histogram of predicted promoter position, n = 479. POSITION marks the 5' end of the data-file 32-mer, and is consequently ~40 nt upstream from the TSS. This distribution peaks with the 5' end around 68 nts upstream from the TLS and the TSS around 28 nts upstream from the TSS.

substantiated. Nonetheless, we suspect that future promoter identifications may confirm the existence of more than two promoter populations for σ^{66} in Chlamydiales.

A chief limitation of our study includes the challenge of collecting a reliable training set that was discussed earlier. We also feel that it would be advisable in future studies to relax the range of possible spacer lengths in the duration HMM for increased generalization, which might have allowed the discovery of more promoters in the whole genome analysis. Additionally, it is quite possible that there are structural features downstream from the TLS, as well as upstream, which would aid in promoter discovery. Future modeling efforts should extend the region of interest to 100 nt downstream from the TLS.

The high priority assigned to the duration HMM scores by the SBLR procedure reinforces that the duration hidden Markov model is an encouraging approach for modelling core promoters, that deserves further development. Also by implementing our model in HMMer our duration HMM is reusable, generalizable, easily adapted to other organisms and open-source. This approach explicitly incorporates spacing preferences of elements in a likelihood framework. Two natural further developments of this approach would include further iteration of the model development in *Chlamydia* using an expanded training set, exploiting computational criteria and measurements to define expanded training sets. Another possible extension would be to model extended promoter elements using further elaborations of the hidden Markov modeling framework.

The *CT* genome-wide promoter predictions and co-predictions with other algorithms provide the basis for future research in promoter identification. The fact that 20% of M2 predicted promoters were co-predicted by NNPP2.2 and TSS-PREDICT supports the validity of all three predictions. The expected confirmation of these promoters will augment the list of verified promoters. However, confirming or rejecting the predictions made by only M2 will provide more valuable information. Confirmation will strengthen the current model in a direction that diverges from *E. coli*, while rejection will add new non-promoter observations that differ from the current training set.

Conclusion

Models M1 and M2 support the conjecture that measures of DNA biophysical criteria along with measures of RNAP σ-factor/DNA binding collaboratively contribute to a sequence's ability to promote transcription. Whereas a measure of RNAP σ-factor/DNA binding ensures a sensitive prediction, adding measures of position relative to the TLS, stability, curvature, SIDD and twist provide specificity. The stratified K-fold cross-validation of M2 indicates that the model performs well by absolute criteria as well as compared to other predictive algorithms. Additionally, there is considerable overlap between the genome-wide predictions of M2 and NNPP2.2, TSS-PREDICT and Footy.

The modeling procedure we describe here seems especially appropriate for bacterial species where the set of known promoters is limited and the genome is relatively small.

Outlook

The model derived by the method described here is a first pass model that serves as proof of concept. The *CT* genome-wide promoter predictions, along with co-predictions by NNPP2.2, TSS-PREDICT and Footy, will allow researchers to select optimal candidates for validation mapping of transcript 5' ends by primer extension. As more chlamydial promoters are identified, the model will be updated, and a refined list of promoter predictions may be developed. More interactions among predictor variables may also be explored. A final model will provide insight into the process of chlamydial transcription initiation. Then, too, it will be possible to determine if chlamydial promoters differ significantly from those of other bacteria.

Authors' contributions

RM participated in conceiving the study, designed the strategies, retrieved data from various websites, conducted the data analysis, performed calculations and wrote the R scripts. DO participated in conceiving the study and provided bacteriological expertise. DA participated in conceiving the study and provided modeling expertise. Research was performed under the advice and supervision of DO and DA. All authors contributed to the draft of the paper, and all authors read and approved the final manuscript.

Additional material

Additional file 1

Input data file for durahmmer used to build M2 duration HMM.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-271-S1.txt]

Additional file 2

Parameters of the M2 duration HMM that were provided for hmmsearch by durahmmer, an HMMer 2.3.2 model file.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-271-S2.txt]

Additional file 3

Data file used to build M2.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-271-S3.xls]

Additional file 4

CT promoters predicted by M2.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-271-S4.xls]

Additional file 5

CT promoters predicted by M2 and NNPP2.2.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-271-S5.xls]

Additional file 6

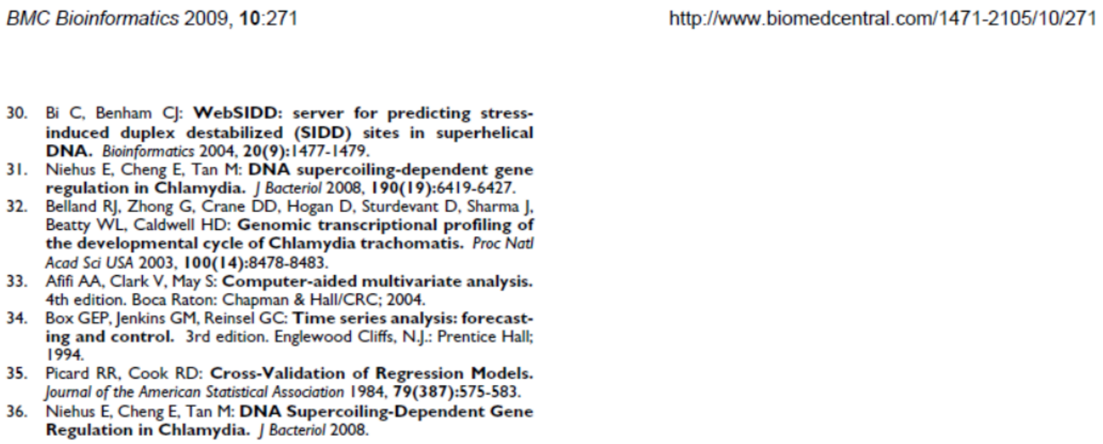
CT promoters predicted by M2 and TSS-PREDICT.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-271-S6.xls]

Additional file 7

CT promoters predicted by M2, NNPP2.2 and TSS-PREDICT.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-271-S7.xls]

References

1. Harley CB, Reynolds RP: **Analysis of E. coli promoter sequences.** *Nucleic Acids Res* 1987, **15(5)**:2343-2361.
2. Hawley DK, McClure WR: **Compilation and analysis of Escherichia coli promoter DNA sequences.** *Nucleic Acids Res* 1983, **11(8)**:2237-2255.
3. Lisser S, Margalit H: **Compilation of E. coli mRNA promoter sequences.** *Nucleic Acids Res* 1993, **21(7)**:1507-1516.
4. Staden R: **Computer methods to locate signals in nucleic acid sequences.** *Nucleic Acids Res* 1984, **12(1 Pt 2)**:505-519.
5. Hertz GZ, Stormo GD: **Escherichia coli promoter sequences: analysis and prediction.** *Methods Enzymol* 1996, **273**:30-42.
6. Huerta AM, Collado-Vides J: **Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals.** *J Mol Biol* 2003, **333(2)**:261-278.
7. Shultzaberger RK, Chen Z, Lewis KA, Schneider TD: **Anatomy of Escherichia coli sigma70 promoters.** *Nucleic Acids Res* 2007, **35(3)**:771-788.
8. Wagner R: **Transcription regulation in prokaryotes.** Oxford; New York: Oxford University Press; 2000.
9. O'Neill MC: **Escherichia coli promoters: neural networks develop distinct descriptions in learning to search for promoters of different spacing classes.** *Nucleic Acids Res* 1992, **20(13)**:3471-3477.
10. Reese MG: **Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome.** *Comput Chem* 2001, **26(1)**:51-56.
11. Burden S, Lin YX, Zhang R: **Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences.** *Bioinformatics* 2005, **21(5)**:601-607.
12. Towsey M, Timms P, Hogan J, Mathews SA: **The cross-species prediction of bacterial promoters using a support vector machine.** *Comput Biol Chem* 2008, **32(5)**:359-366.
13. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14(9)**:755-763.
14. Brunelle BW, Nicholson TL, Stephens RS: **Microarray-based genomic surveying of gene polymorphisms in Chlamydia trachomatis.** *Genome Biol* 2004, **5(6)**:R42.
15. Bavoil PM, Hsia R, Ojcius DM: **Closing in on Chlamydia and its intracellular bag of tricks.** *Microbiology* 2000, **146(Pt 11)**:2723-2731.
16. Tan M: **Regulation of gene expression.** In *Chlamydia genomics and pathogenesis* Edited by: Bavoil PM, Wyrick PB. Wymondham, U.K.: Horizon Bioscience; 2006:103-132.
17. Grech B, Maetschke S, Mathews S, Timms P: **Genome-wide analysis of chlamydiae for promoters that phylogenetically footprint.** *Res Microbiol* 2007, **158(8-9)**:685-693.
18. Hefty PS, Stephens RS: **Chlamydial type III secretion system is encoded on ten operons preceded by sigma 70-like promoter elements.** *J Bacteriol* 2007, **189(1)**:198-206.
19. Mathews S, Timms P: **In silico identification of chamydial promoters and their role in the regulation of development.** In *Chlamydia genomics and pathogenesis* Edited by: Bavoil PM, Wyrick PB. Wymondham, U.K.: Horizon Bioscience; 2006:133-156.
20. Iyer LM, Koonin EV, Aravind L: **Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer.** *Gene* 2004, **335**:73-88.
21. Kanhere A, Bansal M: **Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes.** *Nucleic Acids Res* 2005, **33(10)**:3165-3175.
22. Wang H, Benham CJ: **Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress.** *BMC Bioinformatics* 2006, **7**:248.
23. Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW: **A DNA structural atlas for Escherichia coli.** *J Mol Biol* 2000, **299(4)**:907-930.
24. Agresti A: **An introduction to categorical data analysis.** New York: Wiley; 1996.
25. Hosmer DW, Lemeshow S: **Applied logistic regression.** 2nd edition. New York: Wiley; 2000.
26. Munteanu MG, Vlahovicek K, Parthasarathy S, Simon I, Pongor S: **Rod models of DNA: sequence-dependent anisotropic elastic modelling of local bending phenomena.** *Trends Biochem Sci* 1998, **23(9)**:341-347.
27. SantaLucia J Jr, Allawi HT, Seneviratne PA: **Improved nearest-neighbor parameters for predicting DNA duplex stability.** *Biochemistry* 1996, **35(11)**:3555-3562.
28. Satchwell SC, Drew HR, Travers AA: **Sequence periodicities in chicken nucleosome core DNA.** *J Mol Biol* 1986, **191(4)**:659-675.
29. Uljanov N, James T: **Statistical analysis of DNA duplex structural features.** *Methods in Enzymology* 1995, **261**:90-115.



30. Bi C, Benham CJ: **WebSIDD: server for predicting stress-induced duplex destabilized (SIDD) sites in superhelical DNA.** *Bioinformatics* 2004, **20(9)**:1477-1479.
31. Niehus E, Cheng E, Tan M: **DNA supercoiling-dependent gene regulation in Chlamydia.** *J Bacteriol* 2008, **190(19)**:6419-6427.
32. Belland RJ, Zhong G, Crane DD, Hogan D, Sturdevant D, Sharma J, Beatty WL, Caldwell HD: **Genomic transcriptional profiling of the developmental cycle of Chlamydia trachomatis.** *Proc Natl Acad Sci USA* 2003, **100(14)**:8478-8483.
33. Afifi AA, Clark V, May S: **Computer-aided multivariate analysis.** 4th edition. Boca Raton: Chapman & Hall/CRC; 2004.
34. Box GEP, Jenkins GM, Reinsel GC: **Time series analysis: forecasting and control.** 3rd edition. Englewood Cliffs, N.J.: Prentice Hall; 1994.
35. Picard RR, Cook RD: **Cross-Validation of Regression Models.** *Journal of the American Statistical Association* 1984, **79(387)**:575-583.
36. Niehus E, Cheng E, Tan M: **DNA Supercoiling-Dependent Gene Regulation in Chlamydia.** *J Bacteriol* 2008.