
Characterization of structural and free energy properties of promoters associated with Primary and Operon TSS in *Helicobacter pylori* genome and their orthologs

ADITYA KUMAR and MANJU BANSAL*

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

*Corresponding author (Fax, +91-80-2360 0535; Email, mb@mbu.iisc.ernet.in)

Promoter regions in the genomes of all domains of life show similar trends in several structural properties such as stability, bendability, curvature, etc. In current study we analysed the stability and bendability of various classes of promoter regions (based on the recent identification of different classes of transcription start sites) of *Helicobacter pylori* 26695 strain. It is found that primary TSS and operon-associated TSS promoters show significantly strong features in their promoter regions. DNA free-energy-based promoter prediction tool PromPredict was used to annotate promoters of different classes, and very high recall values (~80%) are obtained for primary TSS. Orthologous genes from other strains of *H. pylori* show conservation of structural properties in promoter regions as well as coding regions. PromPredict annotates promoters of orthologous genes with very high recall and precision.

[Kumar A and Bansal M 2012 Characterization of structural and free energy properties of promoters associated with Primary and Operon TSS in *Helicobacter pylori* genome and their orthologs. *J. Biosci.* 37 423–431] DOI 10.1007/s12038-012-9214-6

1. Introduction

Helicobacter pylori is a gram-negative, microaerophilic, spiral, flagellated ϵ -proteobacterium, which colonizes in the highly acidic environment of the human gut (Blaser 1990). Most of the human population is infected by this dominant species of human gastric microbiome. Its adaptations, such as its spiral shape (Hazell *et al.* 1986), flagellin structure (Andersen-Nissen *et al.* 2005), urease (Marshall *et al.* 1990), LPS Lewis antigens (Aspinall and Monteiro 1996), etc., facilitate its colonization in hazardous environments. Its colonization in gastric epithelium of host causes gastritis that lasts for decades. Infected host is associated with an increased risk of noncardia, peptic ulcer and gastric lymphoma (Cover and Blaser 2009). Gastritis induced by *Helicobacter pylori* increases risk for stomach cancers, while only a small proportion of infected hosts develop malignancy (Polk and Peek 2010). Several strains of this bacterium have been identified and 10 genomes have been fully sequenced, and the data is available in the NCBI database.

The circular genome of the most well-studied strain, *Helicobacter pylori* 26695, is 1.67 million base pairs in

length and carries 1576 open reading frames (Tomb *et al.* 1997). Single base resolution transcriptome study of *Helicobacter pylori* 26695 has been reported recently and provides newer insights into the genome, such as occurrence of different categories of transcripts such as primary, secondary, internal and antisense (Sharma *et al.* 2010). It also shows the presence of an extensive operon structure, and transcripts have been identified associated with primary operons, suboperons and monocistrons. Genomes of *Helicobacter pylori* strains are extremely diverse in nature as panmictic population (population characterized by extensive recombination and genetic diversity) (Polk and Peek 2010). Genetic variability is achieved by both intra-genomic diversification and inter-genomic recombination (Dorer *et al.* 2009). Intra-genomic diversification comes from point mutations and recombination. Genome comparison of two different strains j99 and 26695 of *Helicobacter pylori* from unrelated isolates exhibit differences in DNA sequence of orthologs. This DNA sequence variation does not get translated into highly divergent proteomes (Alm *et al.* 1999). Hence, genome annotation and comparative analysis of such a clinically important species is very important.

Keywords. *Helicobacter pylori*; operons; orthologs; promoters

Promoter annotation is an important component of genome annotation of completely sequenced genomes. In light of recent experimental evidences it has been suggested that promoter regions, which are the sites for transcription initiation, have characteristic structural properties, such as lower stability, low bendability and higher curvature, relative to the flanking genomic sequences. It has been shown that unlike organism specific regulatory sequence motifs, these properties are observed across organisms (Florquin *et al.* 2005; Kanhere and Bansal 2005). The relevance of these structural properties to transcription regulation is implicated in their universal presence in all domains of life, such as prokaryotes, vertebrates and plants (Kanhere and Bansal 2005). In the present work we studied DNA sequence-based structural properties of promoter regions in *Helicobacter pylori* 26695, for which the transcriptome data has recently become available (Sharma *et al.* 2010). The promoter annotation tool PromPredict, developed based on relative free energy calculations of genomic sequences, has been shown to perform uniformly well across a wide range of genomes with GC content varying from 13% to 74% in prokaryotes (Rangannan and Bansal 2010). Whole genome promoter prediction of *Helicobacter pylori* 26695 gives ~60% recall value with ~61% precision (Rangannan and Bansal 2011) when a prediction is scored as True Positive if the low stability signal is located within the 500 nucleotides region upstream of a translation start sites (TLS) of an annotated gene. However, the recent transcriptome analysis clearly indicates the presence of internal transcription start sites (located within the coding region of a neighbouring gene) as well as operons that are well established constructs of prokaryotic genomes and are responsible for a single transcription start site (TSS) regulating transcription of more than one gene. Hence, the above recall and precision values may not provide an accurate idea about the efficiency of the prediction program. In the current study we addressed this issue by analysing the results for experimentally verified TSS and found that the promoter prediction performance was far better than indicated by the above-mentioned recall and precision values. We have also extended the promoter analysis to orthologous genes from nine other strains of *H. pylori* for which whole genome sequence data is available.

2. Materials and methods

2.1 Promoter sequence dataset

Whole genome sequence of *Helicobacter pylori* 26695 was downloaded from NCBI database. TSS information of *Helicobacter pylori* 26695 was acquired from published data (Sharma *et al.* 2010). All protein-coding genes associated TSSs of different categories were extracted. Out of 816 primary TSSs, 714 were found to correspond to protein

coding genes. Similarly, 106 and 426 out of a total of 119 and 439, respectively, of the secondary and internal TSSs correspond to protein coding genes. Operon TSS dataset, which is a subset of primary TSS, is also classified into 3 categories. A total of 309 primary operons, 117 suboperons and 62 monocistrons were obtained for protein coding genes. Promoter sequences of 1001 nucleotides length (-500 upstream to 500 downstream with respect to TSS positioned at 0) were extracted from whole genome sequence in NCBI database. TLS information was also taken from NCBI database.

2.2 Ortholog selection

Orthologous genes were identified by using GeneOrder 4.0 (Mahadevan and Seto 2010) for all 10 available strains of *H. pylori* in NCBI (<ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/>). It uses BLAT ('BLAST-like alignment tool'), which performs 'all against all' protein comparisons and the protein similarity score between two genomes indicating the synteny (gene order). Gene pairs with the highest BLAT scores (200 – very likely to be a homolog) were chosen for the study.

2.3 Free energy calculation and promoter annotation

The stability of double-stranded DNA molecule can be expressed in terms of free energy of its constituent dinucleotides. The free energy of 10 unique dinucleotides is available from melting studies (SantaLucia 1998) of oligo- and polynucleotides. The average free energy of a 1001-nucleotide-long promoter sequences was calculated by dividing each sequence into overlapping 15 base pair windows (or 14 dinucleotide steps). Free energy was calculated for each window by using a set of 10 unique dinucleotide values (Kanhere and Bansal 2005). The average free energy profile of promoter sequences shows low stability peak in the vicinity of the TSS and is used to annotate the promoter (Rangannan and Bansal 2010; Rangannan and Bansal 2011).

2.4 Bendability prediction of sequences

Bendability of promoter sequences was calculated by using two different trinucleotide models – DNase I sensitivity (Brukner *et al.* 1995) and nucleosomal positioning preferences (Satchwell *et al.* 1986). Bendability profile was calculated by using a 30 nucleotide window size and assigning the published bendability values to each consecutive overlapping trinucleotide in the sequence (Kanhere and Bansal 2005). Both models, DNase I sensitivity and nucleosomal positioning preferences, generally show similar bendability profiles with a low bendability peak being observed near the TSS. To

compare stability and bendability properties, all profiles were normalized. The Z-score was plotted for each nucleotide position of the DNA sequence. The Z-score of the i -th observation of a random variable X is given by

$$z_i = \frac{X_i - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of all observations $X_1, X_2 \dots X_n$.

2.5 Evaluation of promoter prediction using PromPredict

Whole genome promoter prediction was performed on *Helicobacter pylori* 26695 genome sequence by using our in-house program PromPredict. Predictions obtained by whole genome predictions were analysed with respect to experimentally determined transcription start sites as well as annotated translation start sites. Predictions that were found in -200 nucleotides upstream to $+100$ downstream region, with respect to TSS, were considered as true positive (TP) (table 1), while -500 upstream to 0 region was considered to define TP prediction, in case of TLS (table 2). Predictions that fell within the coding region were considered as false-positive (FP). Recall, precision and F-score (harmonic mean of recall and precision) evaluation parameters were used to analyse the predictions are defined as follows.

$$\text{Recall} = \frac{\text{No of genes with an identified TP}}{\text{Total no of genes}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$F\text{-score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

The average DE value (viz. the difference in average free energy between neighbouring 100-nucleotide- long regions) was calculated. $\text{PP_DE}_{\text{ave}}$ (DE corresponding to each predicted promoter region) was compared with average

DE value of all predictions, represented as $\text{WPP_DE}_{\text{ave}} (\mu)$, in the particular genome, to assign the reliability level to each prediction (Rangannan and Bansal 2010). The criteria used to classify each prediction according to its reliability level are as follows:

Level 1 (Low) $\rightarrow \text{PP_DE}_{\text{ave}} \leq \mu - 1\sigma$

Level 2 (Medium) $\rightarrow \mu - 1\sigma > \text{PP_DE}_{\text{ave}} \leq \mu$

Level 3 (High) $\rightarrow \mu > \text{PP_DE}_{\text{ave}} \leq \mu + 1\sigma$

Level 4 (Very high) $\rightarrow \mu + 1\sigma > \text{PP_DE}_{\text{ave}} \leq \mu + 1\sigma$

Level 5 (Highest) $\rightarrow > \text{PP_DE}_{\text{ave}} > \mu + 2\sigma$

A prediction score is defined by D_{max} and corresponds to the highest DE value (in kcal/mol) within a predicted region.

3. Results and discussion

3.1 Sequence composition analysis of promoter regions

Base composition analysis of core promoter regions of *Helicobacter pylori* 26695 shows DNA sequences are AT rich in the vicinity of the TSS (data not shown here). To verify if this is due to presence of a TATA box, we analysed the relative frequencies of all tetrameric nucleotide sequences in the core promoter region (-50 to 0 nucleotides with respect to TSS) as compared to the background flanking region (-500 to 500 nucleotides with respect to TSS) in the six different classes of TSSs mentioned above (figure 1). Several AT-rich tetramer nucleotides, including the TATA sequence, show a higher preference for core promoter region. However, the promoter regions associated with internal category of TSSs do not show such preference in tetramer occurrence when compared with the 1001-nucleotide-long background region, and so it will be very difficult for any motif-based promoter annotation tool to predict these promoters.

Table 1. The prediction statistics for promoters of protein coding genes of *Helicobacter pylori* 26695 strain, with experimentally verified TSSs of different categories

Category	TSS	Protein coding genes			Orthologous genes		
		No. of genes	TP genes	Recall	No. of genes	TP genes	Recall
Primary	816	714	559	78.3	578	467	80.8
Secondary	119	106	59	55.7	88	54	61.3
Internal	439	426	152	35.7	323	111	34.4
Operon	320	309	271	87.7	250	226	90.4
Sub-operons	120	117	57	48.7	103	51	49.5
Monocistron	65	62	34	54.8	49	27	55.1

The prediction results for genes of *Helicobacter pylori* 26695 which have orthologs in all other 9 strains are also shown. Predictions in the range of -200 to $+100$ nucleotides with respect to TSS were considered as TP.

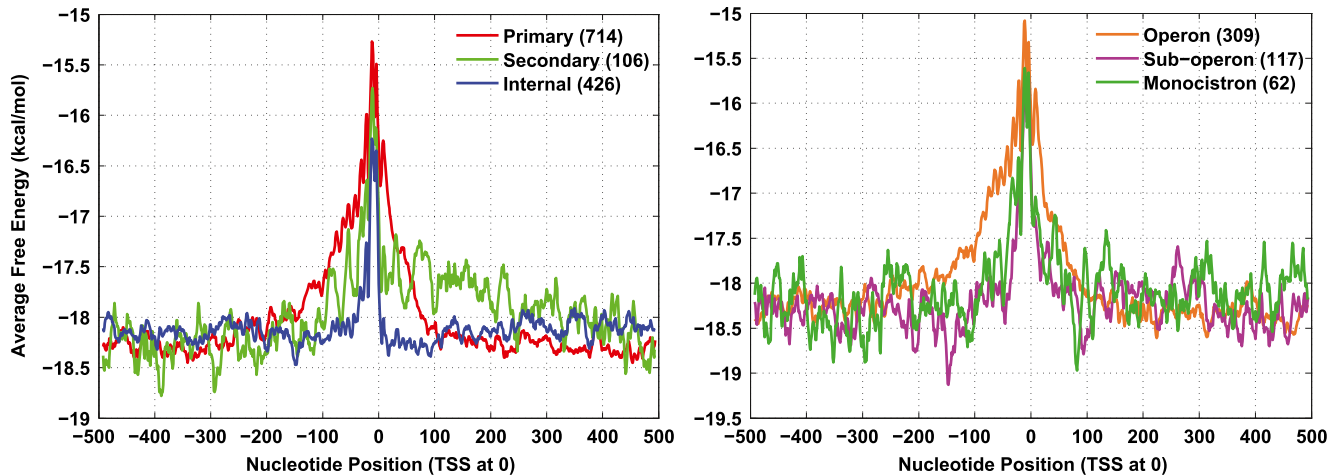


Figure 2. Average free energy profiles of 1001-nucleotide-long genomic sequences (extending from -500 nucleotides upstream to $+500$ nucleotides downstream with respect to various categories of transcription start sites at 0 position, for protein coding genes).

Among the TSSs assigned to operons, the free energy profile for the TSS corresponding to primary operons shows higher and broader low stability peak compared to others. Similarly, bendability profiles show marked variation for different category of TSSs. Both bendability parameters indicate a less bendable or ‘nucleoid free’ region, extending up to ~ 100 nucleotides upstream of the primary TSS (figure 3). Interestingly, DNase I sensitivity, which indicates relative preference for bending towards the major groove (Brukner *et al.* 1995), shows two peaks corresponding to regions of lower bendability, flanking the primary TSS position, while the internal TSS locations seem to correspond to regions with enhanced bending towards the major groove. A smaller preference for AT-rich tetramer sequences in internal category TSS perhaps also accounts for lower height of the bendability peaks.

The profiles for operon, suboperon and monocistron TSS show similar peak heights, because they are subsets of the primary TSS category while their spread varies. This indicates that different categories of promoters do show differences in structural properties even within the same organism.

3.3 Promoter prediction analysis

Whole genome prediction analysis for protein coding genes gives 78% recall value for the primary TSS category of promoter sequences (table 1). The recall value is considerably lower for promoter sequences corresponding to secondary and internal category of TSSs, being 56% and 36% respectively, because of weaker signals in the average free energy plots (as seen in figure 2, where the low stability peak for these TSSs is smaller as compared to the primary category of TSS). In the case of internal TSSs, which lie

within the coding regions, the identified promoter signals would normally have been scored as FPs in our whole genome analysis. The TSSs for operons have a recall value of 88%, which is 10% higher than that for the primary category of TSSs. This indicates that operon promoter signals are more ubiquitous than the other primary category promoters, even though the operons TSSs are a subset of primary TSS dataset. The suboperons and monocistrons achieve recall values of $\sim 49\%$ and $\sim 55\%$ respectively.

The whole genome annotation of promoters, in regions upstream of annotated TLS, in nine other strains of *H. pylori* achieved recall and precision values similar to those for the 26695 strain (Rangannan and Bansal 2010). As described in the section Methods, a common set of orthologous genes present in *Helicobacter pylori* 26695 (corresponding to genes that have orthologs in all nine available strains) was created, using gene pairs obtained from GeneOrder 4.0 server. The orthologous genes present in *Helicobacter pylori* 26695 show marginally higher recall value for promoter annotation as compared to all protein-coding genes corresponding to all categories of TSS, except internal category (table 1). This indicates that along with the coding regions, the promoter regions are also highly conserved for orthologous genes.

Whole genome promoter prediction studies carried out earlier in our laboratory, with respect to TLS, showed a recall value of $\sim 60\%$; however, in light of the reported operon structure, when we calculated recall value only for the protein-coding genes associated with the primary TSS of *Helicobacter pylori* 26695 strain, we achieved a recall value of $\sim 79\%$ (table 2). This indicates that most of the genes that do not have any associated promoter predictions probably lack independent TSSs and form part of an operon. The precision value improves from 61% to 66%, since some of

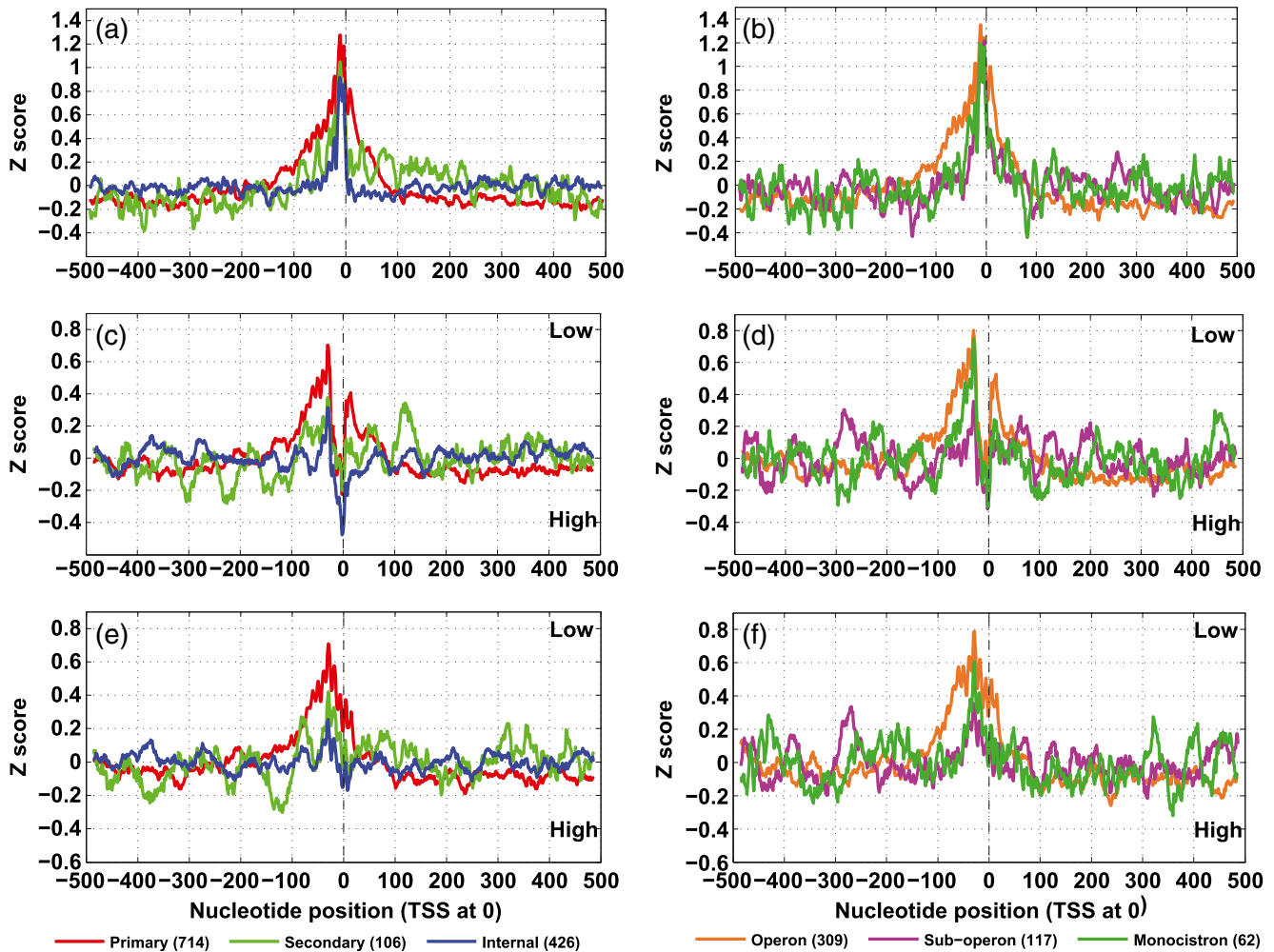


Figure 3. Normalized Z-scores for the DNA sequence dependent structural properties showing characteristic features in the vicinity of TSSs. Average stability profiles (a and b) and average bendability profiles (c, d) using DNase I sensitivity and (e, f) using Nucleosomal positioning preference of 1001 nucleotides long DNA sequences.

the predictions earlier identified as FPs are found to be TPs for internal TSS and hence not counted as FPs in the present analysis. The ~12% increase in F-score value arises because recall and precision are both improved.

The orthologous genes that have primary TSSs identified in the 26695 strain have their promoters annotated with a recall value of 82% and 68% precision (table 3). Predictions are consistent, with average ~81% recall and ~65% precision values, for orthologous genes present in each of the nine strains. Combined performance of recall and precision for the common orthologous gene dataset gives an F-score ranging between 71% and 74%. This indicates that PromPredict is a valid algorithm to predict the promoters for orthologous genes in related bacterial strains.

Reliability levels of promoter predictions for the 26695 strain show that level 3 (high) and higher predictions are comparatively more frequent for primary and secondary

TSSs than for the internal category TSSs (figure 4). Similarly, for operon category TSSs, level 3 (high) and higher predictions are more as compared to suboperon TSS. This suggests that the reliability levels of the predictions are indicative of the TSS category.

3.4 Correlation of D_{max} scores between promoters for orthologous genes

Previous studies on orthologous genes of rice and Arabidopsis (Morey *et al.* 2011) showed that their prediction scores (D_{max}) are correlated. In the current study we analysed prediction scores of genome wide orthologous genes of 10 strains of *Helicobacter pylori* and found that that the D_{max} values (which vary between 1 to 7 kcal/mol) for promoter predictions for orthologous gene pairs (corresponding to

Table 3. The prediction statistics for orthologous genes in 10 different strains of *Helicobacter pylori* genome sequences available in NCBI database

Organism	Genes*	Genes with TP	No. of TP	No. of FP	Recall	Precision	F-score
<i>H. pylori</i> 26695	578	471	564	264	81.5	68.1 [†]	74.2
<i>H. pylori</i> J99	590	481	577	313	81.5	64.8	72.2
<i>H. pylori</i> HPAG1	595	479	553	321	80.5	63.3	70.9
<i>H. pylori</i> Shi470	582	475	563	292	81.6	65.8	72.9
<i>H. pylori</i> G27	594	485	577	324	81.6	64.0	71.8
<i>H. pylori</i> P12	590	487	580	295	82.5	66.3	73.5
<i>H. pylori</i> B38	579	470	561	302	81.2	65.0	72.2
<i>H. pylori</i> B8	598	475	567	324	79.4	63.6	70.7
<i>H. pylori</i> PeCan4	589	480	567	317	81.5	64.1	71.8
<i>H. pylori</i> SJM180	589	483	566	313	82.0	64.4	72.1

*Genes with experimentally determined primary TSS were considered for the 26695 strain, while for all other strains their orthologs were considered. Predictions in the range of -500 to 0 nucleotides with respect to TLS were considered as TP and predictions falling in the coding region were considered as FP.

[†] FPs that were found to be TPs for internal TSSs were removed while calculating the precision value for the 26695 strain.

experimentally verified primary TSS in reference strain 26695) were well correlated (correlation coefficient >0.8) (figure 5). This implies that promoter regions of the orthologous genes exhibit very similar properties.

4. Conclusion

Promoter sequences corresponding to various categories of TSSs such as primary, secondary and internal show some

differences in their structural properties, indicating that classification of their promoter regions is also possible. The differences in the tetramer nucleotide frequencies between the internal category promoter regions and the flanking regions are quite small. Hence, it would be very difficult to annotate them using motif-based promoter annotation tools. However, they do show characteristic structural properties, and so promoter annotation tools, like PromPredict, which are based on these properties, will be more efficient in annotating these regions.

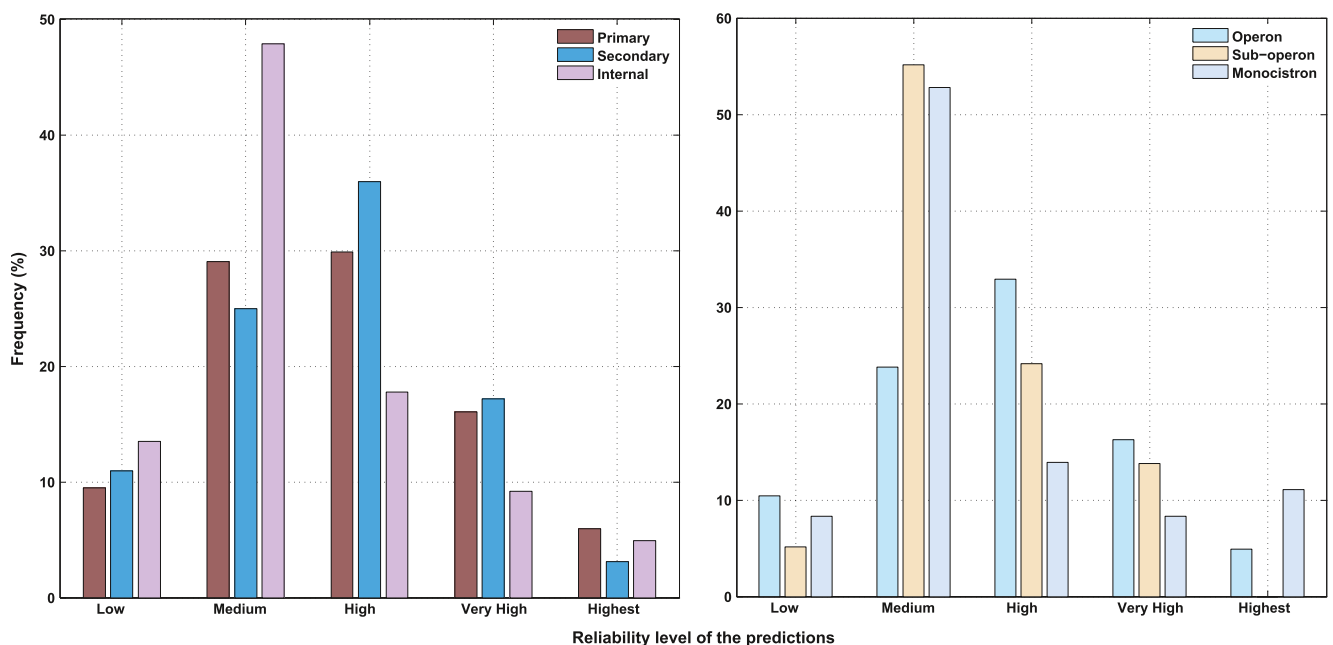


Figure 4. Reliability levels of the predictions for different categories of TSSs for the *H. pylori* 26695 strain.

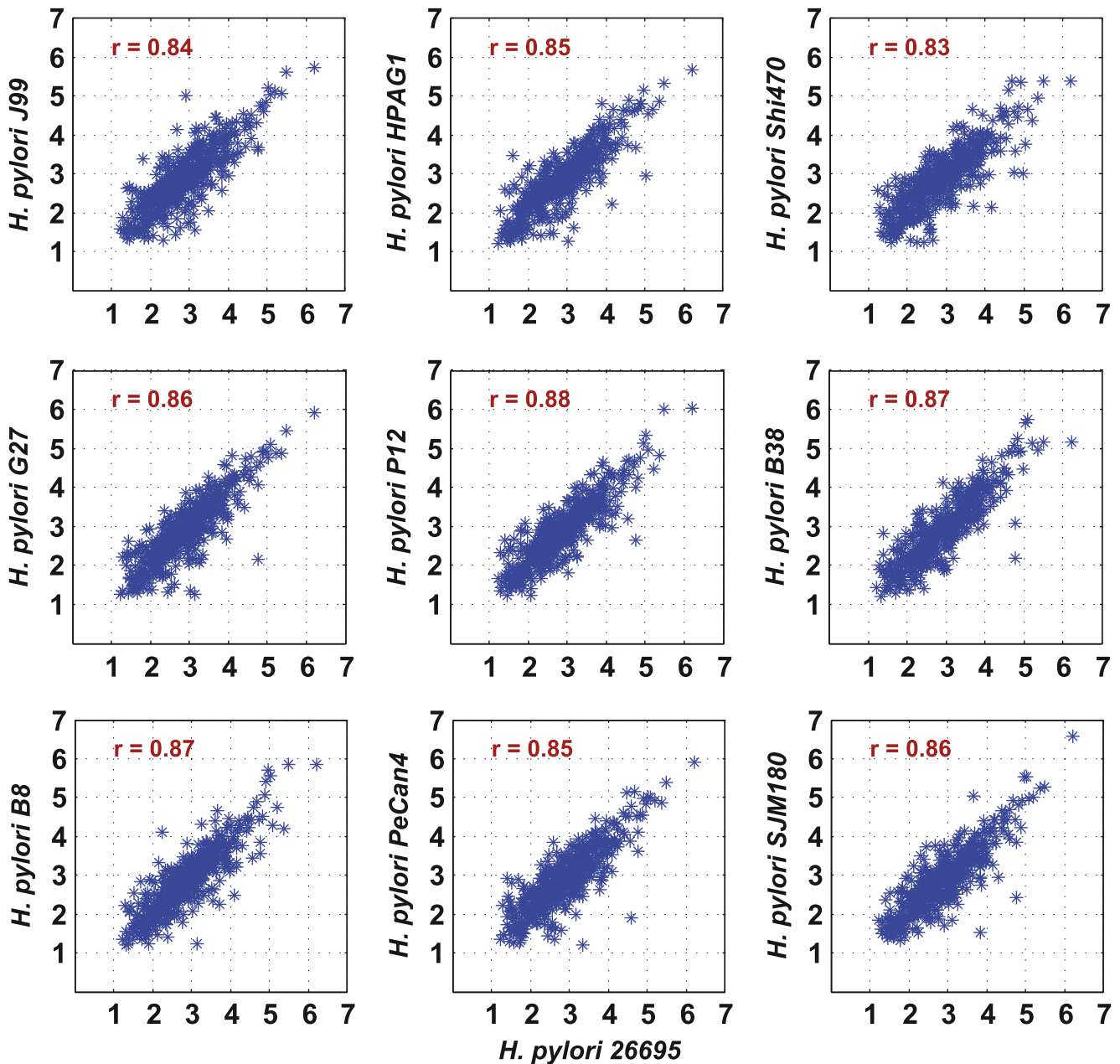


Figure 5. Prediction score (Dmax) correlation between orthologous genes in various *Helicobacter pylori* genomes. Strain 26695 (x-axis) was taken as a reference for orthologous gene predictions. Each point represents the prediction score of an orthologous gene pair, corresponding to a primary TSS in reference strain. Multiple ortholog predictions with respect to reference gene were removed by selecting the ortholog with nearest prediction score. The correlation coefficient value for each pair of genomes is indicated on the top right of the plot.

Whole genome promoter annotation of *H. pylori* genomes using PromPredict, with respect to TLS dataset, achieves a recall value of ~60% only, which can be attributed to the presence of operon architecture in the prokaryotic genome. Most of the genes that do not have predictions probably lack their personal TSS. TSSs associated with operons have their promoter regions very well annotated

by our methodology and a recall value of ~88% is achieved. Orthologous genes in all *H. pylori* strains also show conserved structural properties in their promoter region, and PromPredict also annotates the promoter regions of the orthologous genes with very high recall and hence can facilitate identification of operons in other strains of *H. pylori*.

Acknowledgements

The authors thank Vasumathi M for her contribution in the analysis of orthologous genes.

References

- Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, *et al.* 1999 Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397** 176–180
- Andersen-Nissen E, Smith KD, Strobe KL, Barrett SL, Cookson BT, Logan SM and Aderem A 2005 Evasion of Toll-like receptor 5 by flagellated bacteria. *Proc. Natl. Acad. Sci. USA* **102** 9247–9252
- Aspinall GO and Monteiro MA 1996 Lipopolysaccharides of *Helicobacter pylori* strains P466 and MO19: structures of the O antigen and core oligosaccharide regions. *Biochemistry* **35** 2498–2504
- Blaser MJ 1990 *Helicobacter pylori* and the pathogenesis of gastro-duodenal inflammation. *J. Infect. Dis.* **161** 626–633
- Brukner I, Sanchez R, Suck D and Pongor S 1995 Trinucleotide models for DNA bending propensity: comparison of models based on DNaseI digestion and nucleosome packaging data. *J. Biomol. Struct. Dyn.* **13** 309–317
- Cover TL and Blaser MJ 2009 *Helicobacter pylori* in health and disease. *Gastroenterology* **136** 1863–1873
- Dorer MS, Talarico S and Salama NR 2009 *Helicobacter pylori*'s unconventional role in health and disease. *PLoS Pathog.* **5** e1000544
- Florquin K, Saeys Y, Degroeve S, Rouze P and Van de Peer Y 2005 Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Res.* **33** 4255–4264
- Hazell SL, Lee A, Brady L and Hennessy W 1986 Campylobacter pyloridis and gastritis: association with intercellular spaces and adaptation to an environment of mucus as important factors in colonization of the gastric epithelium. *J. Infect. Dis.* **153** 658–663
- Kanhere A and Bansal M 2005 Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Res.* **33** 3165–3175
- Mahadevan P and Seto D 2010 Rapid pair-wise synteny analysis of large bacterial genomes using web-based GeneOrder4.0. *BMC Res. Notes* **3** 41
- Marshall BJ, Barrett LJ, Prakash C, McCallum RW and Guerrant RL 1990 Urea protects *Helicobacter* (Campylobacter) *pylori* from the bactericidal effect of acid. *Gastroenterology* **99** 697–702
- Morey C, Mookherjee S, Rajasekaran G and Bansal M 2011 DNA free energy-based promoter prediction and comparative analysis of Arabidopsis and rice genomes. *Plant Physiol* **156** 1300–1315
- Polk DB and Peek RM Jr 2010 *Helicobacter pylori*: gastric cancer and beyond. *Nat. Rev. Cancer* **10** 403–14
- Rangannan V and Bansal M 2010 High-quality annotation of promoter regions for 913 bacterial genomes. *Bioinformatics* **26** 3043–3050
- Rangannan V and Bansal M 2011 PromBase: a web resource for various genomic features and predicted promoters in prokaryotic genomes. *BMC Res. Notes* **4** 257
- SantaLucia J Jr 1998 A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* **95** 1460–1465
- Satchwell SC, Drew HR and Travers AA 1986 Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191** 659–675
- Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, *et al.* 2010 The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464** 250–255
- Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, *et al.* 1997 The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388** 539–547