DNA-energetics-based analyses suggest additional genes in prokaryotes

GARIMA KHANDELWAL[1,2], JALAJ GUPTA[1,2] and B JAYARAM[1,2,3,*]

[1]*Department of Chemistry*

[2]*Supercomputing Facility for Bioinformatics and Computational Biology*

[3]*Kusuma School of Biological Sciences*
Indian Institute of Technology, New Delhi 110 016, India

**Corresponding author (Fax, +91-11-2658 2037; Email, bjayaram@chemistry.iitd.ac.in)*

We present here a novel methodology for predicting new genes in prokaryotic genomes on the basis of inherent energetics of DNA. Regions of higher thermodynamic stability were identified, which were filtered based on already known annotations to yield a set of potentially new genes. These were then processed for their compatibility with the stereo-chemical properties of proteins and tripeptide frequencies of proteins in Swissprot data, which results in a reliable set of new genes in a genome. Quite surprisingly, the methodology identifies new genes even in well-annotated genomes. Also, the methodology can handle genomes of any GC-content, size and number of annotated genes.

[Khandelwal G, Gupta J and Jayaram B 2012 DNA-energetics-based analyses suggest additional genes in prokaryotes. *J. Biosci.* **37** 433–444] **DOI** 10.1007/s12038-012-9221-7

1. Introduction

Since the sequencing of the first microbial genome (Fleischmann *et al.* 1995), the wave of genome sequence data has risen exponentially, with more than 1761 microbial genomes already sequenced and several getting re-sequenced (Pagani *et al.* 2012). In the high-throughput sequencing technology scenario, computational methods are rapidly becoming an integral part of automated data annotation (Pati *et al.* 2010). The most important part of sequence annotation is gene prediction, which refers to the process of detecting functional stretches of DNA sequences, most commonly, the protein coding regions on DNA (Fickett 1982). Gene prediction is the main focus of many annotation methods.

Gene prediction methods use either an *ab initio* or homology approach. Most of the *ab initio* approaches are based on training the model parameters on known annotations by either looking at the disparity in statistics of nucleotides in different regions of DNA such as the probability of occurrence of a stop codon in a coding region as opposed to that in a random sequence/non-coding region, or by looking at the content of various regions such as the codon bias, oligomer frequencies, etc. The homology-based approaches match the sequence with its homologs in the annotated databases using alignment methods (Korf *et al.* 2001). Apart from these, some hybrid methods have also been developed that utilize both the above approaches for gene prediction. As compared to manual curation, these methods are relatively faster but less accurate (Guigó *et al.* 2006; Harrow *et al.* 2006; DeCaprio *et al.* 2007), leading to a need for better automated methods.

Most of the gene prediction softwares claim an accuracy of 90% or more, but their predictions vary as they utilize different models for their predictions on different systems. The gene prediction methods are limited to the known sequences either for training them or for finding a homolog, making them severely database dependent (Claverie *et al.* 1997). This is evident from the fact that the new genes are being discovered in already annotated genomes (Glusman *et al.* 2006; Jensen *et al.* 2006) leading to regular updates of the initial genomic annotations assigned using different annotation softwares, mostly based on statistical models. A recent study has shown that training-based models are not

Keywords. DNA; energetics; gene prediction; prokaryotes

Published online: 25 June 2012

able to determine all the protein coding regions present in a genome (Pati *et al.* 2010), which could be due to the lack of generalization in these models that should detect genes irrespective of the training data applied to capture them in the newly sequenced organisms, especially if the training data is unavailable for the organism in question. Also, predictions are fraught with many false-positives and their specificities are worse than their sensitivities (Yok and Rosen 2011).

Studies in the recent past have focused on improvements in predictions and on the development of next-generation prediction servers for better annotations. The newer methods are characterized by improved annotation/prediction in both prokaryotes and eukaryotes by either reducing the predicted number of false-positives (Baren and Brent 2006) or predicting new genes or improving some part of the prediction such as translational start sites (TSS). A few methodologies also utilize an extra step over and above their predictions like frameshift detection (Tech and Meinicke 2006), removal of overlapping genes, adjustment of translational initiation site (Stormo *et al.* 1982; Zhu *et al.* 2004) and comparison with closely related genomic sequences (Yu *et al.* 2007). Some of the methodologies for improving predictions that have been incorporated in one or more softwares are the following: combining comparative genomic approaches with *ab initio* results (Frishman *et al.* 1998), utilizing phylogenetic fingerprints (Wu *et al.* 2006), employing protein multiple sequence alignments (Keller *et al.* 2011), modelling conserved features (Meyer and Durbin 2004), using cDNA (Stanke *et al.* 2008) or Expressed Sequence Tag (EST) data or by consensus prediction by two or more methods (Yok and Rosen 2011). Several gene predictors like ORPHEUS (Frishman *et al.* 1998), AUGUSTUS (Stanke *et al.* 2004; Keller *et al.* 2011), Procustes (Gelfand *et al.* 1996), N-SCAN (Gross and Brent 2006), GenomeScan (Yeh, *et al.* 2001), SLAM (Alexandersson *et al.* 2003), GeneMapper (Chatterji and Pachter 2006), GeneWise (Birney and Durbin 2000), MED (Zhu *et al.* 2007), GeneComber (Shah *et al.* 2003), GenePrimp (Pati *et al.* 2010), etc., have incorporated these or similar methodologies to improve their predictions. A program called 'Combiner' (Allen *et al.* 2004) utilizes data about the gene boundary locations from *ab initio* methods, splice-site prediction, protein sequence alignments, EST and cDNA alignments along with other evidences to predict complete gene models in eukaryotes.

Overall, the acute database dependence, and the paucity of experimental information, strongly underscores the need for a methodology that is truly *ab initio* and requires no training. Methods to tackle this problem have been proposed for prokaryotes (Audic and Claverie 1998; Besemer and Borodovsky 1999), but the results were not completely convincing (Mathé *et al.* 2002). This is feasible only if the intrinsic properties of DNA and the proteins that they code for are understood and built-in. Difficulties arising due to compositional variance of DNA, available gene models, etc., can be overcome if properties inherent to the DNA molecule itself are brought to fore in the gene prediction methodologies.

An earlier attempt to predict genes using the physicochemical properties of DNA and the conjugate rule combined into a 'J-vector' has proven to be quite successful (Jayaram 1997; Singhal *et al.* 2008), with sensitivity and specificity values of 0.87 (86.53%) and 0.64 (63.91%) respectively for 372 prokaryotic genomes. This method is universally applicable to all prokaryotic genomes and does not need alterations of input parameters specific to any organism. Also, a separate study based on the melting temperatures of DNA (Khandelwal and Jayaram 2010) showed different stabilities for various functional regions of genomic sequences, notably higher stability for genic regions as compared to the non-genic regions, which was similar to the previous work done in this area utilizing similar approaches (Maeda and Ohtsubo 1987; Wada and Suyama 1983, 1984a, 1984b, 1985a, 1986; Huang and Kowalski 2003).

Building on these findings, we started looking for regions on a genome that possess gene-like characteristics such as higher stability, a gene-like 'J-vector', the right combination of stereochemical properties as well as the observed tripeptide frequencies in the protein products. We indeed found several regions on completely sequenced and annotated genome sequences that had gene-like characteristics and also showed homology with already annotated gene or protein sequences, but were not annotated in the organism studied. Encouraged by these initial results, we developed a complete methodology for predicting new genes in a sequenced genome, which were missed by the initial annotation softwares. The methodology is provided in the materials and methods section and the findings on a few genomes are discussed. It was already proposed that unidentified genes are yet to be mined from the intergenic regions (Dhar *et al.* 2009) and was interesting to see that even in highly studied and annotated genomes; numerous potential genes were found, even after considering the possibility of error in the prediction, especially with overprediction. Such large numbers hint that there must still be some regions on a genome that are not annotated but do function as coding sequences. It is also possible that these potential genes may represent evolutionarily transient stages of gene decay or gene gain (Knowles and McLysaght 2009; Siepel 2009). This suggests that our knowledge of even the location of genes and other functional units on genomic DNA is still incomplete, which again emphasizes the need for the development of alternative and better methods for genome annotation.

2. Materials and methods

The methodology works on the principle that basic energetic interactions such as hydrogen bonding between the Watson–Crick base pairs and stacking between the adjacent base

pairs contributes to the stability of the DNA molecule. In this study, the stability of the molecule was established on the basis of the melting temperature of the system, derived from the use of the energetic contributions mentioned above. The hydrogen bonding and stacking energy parameters were derived from all-atom molecular dynamics simulations on all tetranucleotide sequences (Dixit *et al.* 2005; Lavery *et al.* 2009). The energies for all the 10 unique dinucleotides were determined as a special case of averaging over trinucleotide data (table 1). Both hydrogen bonding and stacking energies contributions were considered together for each dinucleotide in the form of 'strength parameter' and denoted as 'E' (table 1), as in a previous study (Khandelwal and Jayaram 2010). The strength parameter was then used along with the Na^+ ion concentration and DNA concentration of the sequences to generate a regression equation (equation 1) for predicting the melting temperature of DNA sequences, on a training set of 123 oligonucleotides for which the experimental melting temperatures were known.

$$\mathrm{Tm}(^\circ\mathrm{C}) = \{(-8.69 \times \mathrm{E}) + [6.07 \times \ln(\mathrm{Len})] + [4.97 \times \ln(\mathrm{Conc})] + [1.11 \times \ln(\mathrm{dna})]\} - 233.45 \quad (1)$$

Here, E is the strength parameter, Len is the length of DNA sequence, Conc is the the Na^+ ion concentration and dna is the oligonucleotide concentration.

The training dataset gives a Pearson Product moment correlation coefficient of 0.98 between the experimental and predicted melting temperatures and an average error of 1.38°C. Equation 1 is then used to predict the melting temperatures of a dataset of 225 oligonucleotides for which the experimental melting temperatures were known, in order to check the reliability of the prediction (figure 1). The predicted melting temperatures of the test dataset yield a Pearson Product moment correlation coefficient of 0.99 against experiment and the average error of prediction is 1.27°C. It may be noted that the MD-derived strength parameters correlate quite well with the experimental melting temperatures even without any training. The training here is only to make the predicted melting temperatures quantitative and absolute.

Table 1. Energy parameters (in kcal) for dinucleotides derived from molecular dynamics simulations

Dinucleotide	Hydrogen bond	Stacking energy	Strength parameter
AA	−6.92	−26.92	−33.84
AC	−9.64	−27.87	−37.51
AG	−8.78	−26.91	−35.69
AT	−7.05	−27.34	−34.38
CA	−9.34	−27.23	−36.57
CC	−11.84	−26.33	−38.17
CG	−11.37	−27.83	−39.20
CT	−8.78	−26.91	−35.69
GA	−10.12	−26.98	−37.10
GC	−12.03	−28.27	−40.30
GG	−11.84	−26.33	−38.17
GT	−9.64	−27.87	−37.51
TA	−7.16	−27.15	−34.31
TC	−10.12	−26.98	−37.10
TG	−9.34	−27.23	−36.57
TT	−6.92	−26.92	−33.84

There are other methods for melting temperature prediction, the most popular one being that of SantaLucia, which is based on stacking parameters (SantaLucia 1998). The predicted melting temperatures using SantaLucia's parameters on the above test dataset as reported in the literature (Panjkovich and Melo 2005) show a Pearson Product moment correlation coefficient of 0.98 and an average error of 2.96°C, which is double that obtained by the current method.

Other factors such as solvation and ion atmosphere also contribute to the stability of DNA (Jayaram and Beveridge 1990). It is interesting that this simple method based on hydrogen bonding and stacking works so well in predicting melting temperatures of oligonucleotides.

To deal with longer sequences extending to the level of genomes, the sequence is first broken into overlapping windows of 70 base pairs and the melting temperature of each window is calculated, generating a continuous array of

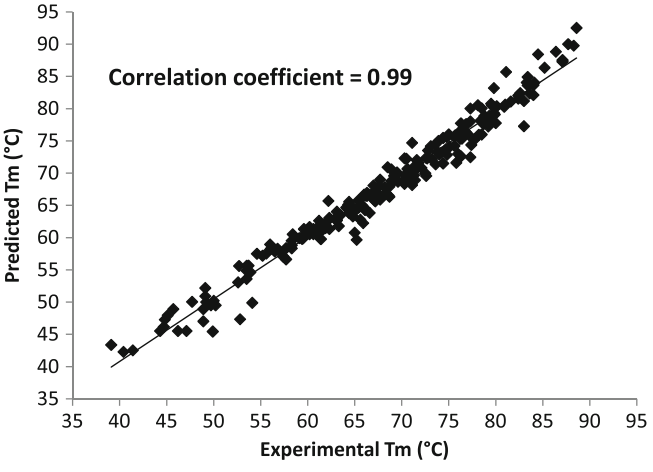

Figure 1. Correlation between experimental and predicted melting temperatures of a test dataset consisting of 225 oligonucleotides.

melting temperatures over the sequence termed as the 'melting profile' of the sequence. For a sequence of 'N' bases, a total of (N−70) data points are obtained forming a profile. The average melting temperature for a sequence is calculated as:

$$Avg\ Tm(^{\circ}C) = \frac{\sum_{1}^{[(N-70)+1]} Tm}{[(N-70)+1]}$$

The computed melting temperatures of DNA sequence of any given length were found to correlate very well with the experimental melting temperatures of complete genomic melting data of different organisms (data not shown). The methodology for the generation of melting profiles is presented in detail elsewhere (Khandelwal and Jayaram 2010).

2.1 *Extraction of thermodynamically more stable regions*

The melting temperatures correspond to the stability of the DNA sequence – higher temperatures indicating higher thermodynamic stability of those sequences. This property is utilized to determine relatively more stable regions in the genome. Genome sequences along with their annotations are downloaded from the National Centre for Biotechnology Information (NCBI, *http://www.ncbi.nlm.nih.gov/*). Only the genomes where the sequencing and assembly was complete and had varied GC-content were considered for the study. The melting profile and the average melting temperature of the sequences were calculated as mentioned above. The sequences of all the regions depicting higher stability with respect to the average melting temperature for that genome were extracted from the genomic sequence. The minimum length of a sequence to be considered as a potential gene for extraction was set at 100, and minimum segment of low stability to separate two potential genic regions of high stability was fixed at 35 bases. The extracted sequences were then extended on both the 5′ and 3′ ends in all the six frames to form a complete Open Reading Frame (ORF), as this methodology is not based on a codon system and does not look for an ORF in its initial step. The ones that form an ORF are retained for further analysis and the rest of them are discarded. If there is more than one ORF with the same stop position but with different start positions, then only the longest ORF is considered for further analysis.

2.2 *Comparison with protein sequence databases*

The extended sequences are then converted into their corresponding protein sequences and searched against the non-redundant database of protein sequences using BLASTP (Version: 2.2.25+) modules of the stand-alone version of BLAST (Atshul 1990), downloaded from the ftp site of NCBI (*ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/*). The BLAST database was also downloaded from ftp site of NCBI (*ftp://ftp.ncbi.nlm.nih.gov/blast/db/*), updated as of December 2011. The genomes used are also updated as of December 2011, so as to provide consistency between the annotation of genomes and the data present in the databases used for BLAST searches. Sequences having an E-value of less than 0.01 (Group I) are separated from the ones which show low or no similarity (Group II) with the already known protein sequences. Sequences in Group I that match with ones already present in the query genome are rejected as they have already been annotated. Sequences that match with some other organisms are not scrutinized through any other filter and are marked as 'new annotations' for the query genome sequence.

2.3 *Determining protein coding potential of new sequences*

The ORFs of Group II sequences that have extremely low or no similarity with known sequences after BLASTP analysis are then evaluated on the basis of their 'J-vector' orientation, which separates gene-like sequences from non-gene-like sequences using physico-chemical properties of DNA (Dutta *et al.* 2006). Sequences that are predicted to be having gene-like features are then converted to amino acid sequences and screened on the basis of stereo-chemical properties (linear or branched, hydrogen bond donors, conformationally flexible, and short or long) of amino acid side chains in naturally occurring proteins (Jayaram 2008). An initial version of this filter was incorporated in an earlier gene prediction software (ChemGenome 2.0, *http://www.scfbio-iitd.res.in/chemgenome/chemgenomenew.jsp*), which was modified to improve its accuracy as checked against the latest Swissprot/Uniprot data (O'Donovan *et al.* 2002; The Uniprot Consortium 2011). A third filter developed on the basis of standard deviations in the frequency of occurrence of tripeptides from Swissprot data (with evidence at the protein and the transcript levels) is also used to reduce the number of false-positives. A threshold value of ≤2.5 was set to discriminate gene-like sequences from non-gene-like sequences. The sequences that are obtained after the above methodology are termed as 'potential new genes'. The flowchart of the complete process is presented in figure 2.

The methodology described above has been utilized as an illustration for detecting potential new genes in 12 different genomes with varied GC percentages ranging from 22% to 74%. The scientific names of the organisms along with their NCBI IDs and the percentage GC-content is presented in table 2.

Further testing was done on Synthetic *Mycoplasma genitalium*, *Bacillus subtilis* and *Ecsherichia coli* genomes, downloaded from the NCBI Website along with their annotations.

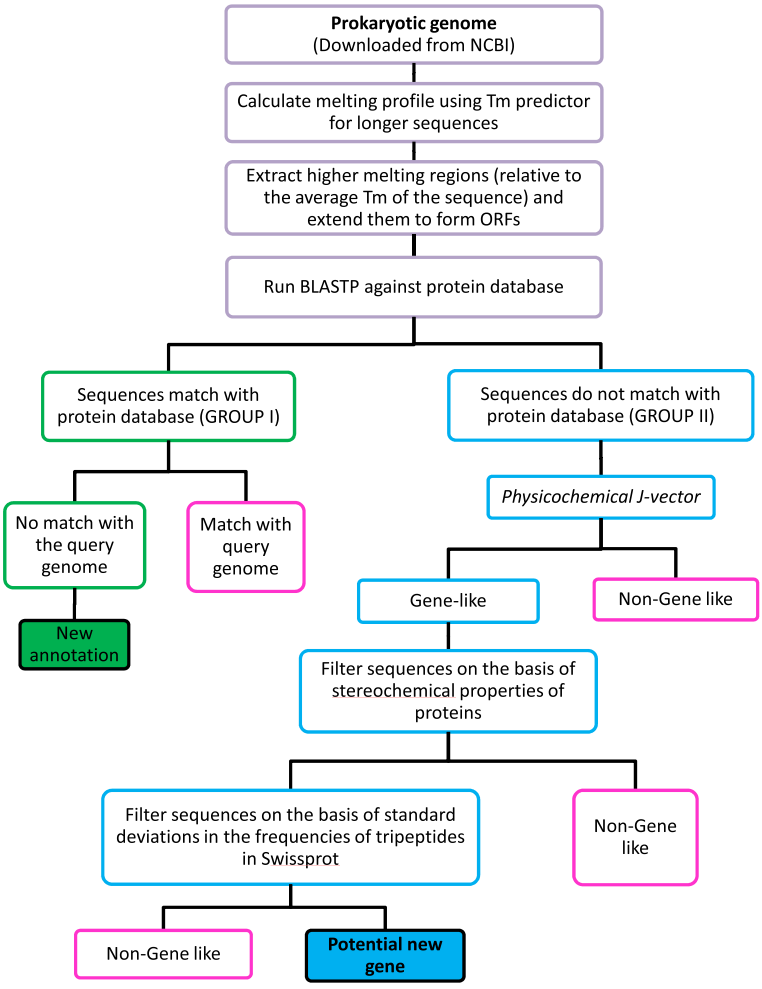

Figure 2. Flowchart for predicting new genes using physico-chemical approach.

3. Results and discussion

Thermodynamics of DNA plays a vital role in its functionality, as studies have shown that different regions having different functionalities show different energetic signatures (Maeda and Ohtsubo 1987; Wada and Suyama 1983, 1984a, b, 1985a, 1986; Umek and Kowalski 1988; Natale *et al.* 1993; Lin and Kowalski 1994; Kanhere and Bansal 2005a). A few investigators have utilized this fact in the past using one or more of the energy components together with or without structural properties of the DNA molecule to predict functional regions over a sequence (Kanhere and Bansal 2005b; Dutta *et al.* 2006; Rangannan and Bansal 2007; Singhal *et al.* 2008; Abeel *et al.* 2008; Dineen *et al.* 2009; Morey *et al.* 2011). Previous studies have also documented the effect of various energy components on the stability of DNA and mutational stability of DNA sequences (Delcourt and Blake 1991; Hunter 1993; Sugimoto *et al.* 1996; Owczarzy *et al.* 1999; Sponer *et al.* 2001, 2004; Protozanova *et al.* 2004; Yakovchuk *et al.* 2006).

We utilized the stacking energy between the base pairs and the hydrogen bonding energy between the Watson–Crick base pairs as the strength parameter 'E' to compute the energetic contributions towards the stability of DNA. This was then used to derive a regression equation to predict the melting temperatures of oligonucleotides. The same equation was used to predict the melting temperatures of longer DNA sequences at the level of genomes, and to develop their melting profiles as shown in figure 3.

Sequences having higher stability with respect to the average value for that genome are then screened using BLAST, J-vector, stereo-chemical properties of proteins and deviations from the Swissprot frequencies of tripeptides. It may be noticed from figure 3 that there are two regions that have a higher stability (relative to genomic average Tm) without any annotation provided to them. When we analysed

Table 2. Genomes analyzed in this study with the number of gene-like sequences predicted after each step of the methodology

Genome	GC-content (%)	NCBI ID	Number of bases	Number of annotated CDS	Higher stability ORF	Potential gene sequences obtained after each step						
						Group I	Match with query genome	New Annotations	Group II	J-vector	Stereo-chemical properties	Tripeptide frequency
Candidatus Sulcia muelleri DMIN	22.5	NC_014004	243933	226	158	135	132	**3**	23	7	5	**1**
Brachyspira pilosicoli 95/1000	27.9	NC_014330	2586443	2299	2045	1470	1442	**28**	575	246	166	**56**
Weissella koreensis KACC *15510*	35.5	NC_015759	1422408	1335	1323	823	803	**20**	500	228	142	**27**
Taylorella equigenitalis MCE9	37.4	NC_014914	1695860	1556	1536	959	918	**41**	577	277	155	**34**
Glaciecola nitratireducens FR1064	42.3	NC_016041	4134229	3654	4301	2391	2284	**107**	1910	780	465	**145**
Acaryochloris marina MBIC11017	47.3	NC_009925	6503724	6254	7307	3769	3567	**202**	3538	2505	1596	**586**
Prevotella denticola F0289	50.4	NC_015311	2937589	2386	4435	1868	1511	**357**	2567	2086	1727	**1218**
Acetobacter Pasteur-ianus IFO 3283-01	53.0	NC_013209	2907495	2628	5097	2191	1629	**562**	2906	2408	1788	**1123**
Candidatus Tremblaya princeps	58.8	NC_015736	138927	121	225	85	65	**20**	140	114	102	**57**
Rhodopseudomonas palustris TIE-1	64.9	NC_011004	5744041	5246	13160	4810	3476	**1334**	8350	7639	7258	**5743**
Pseudoxanthomon as spadix BD-a59	67.7	NC_016147	3452554	3149	8162	2921	1974	**947**	5241	4740	4454	**3669**
Isoptericola variabilis 225 chromosome	73.9	NC_015588	3307740	2881	7053	2564	1951	**613**	4489	4080	4018	**3345**

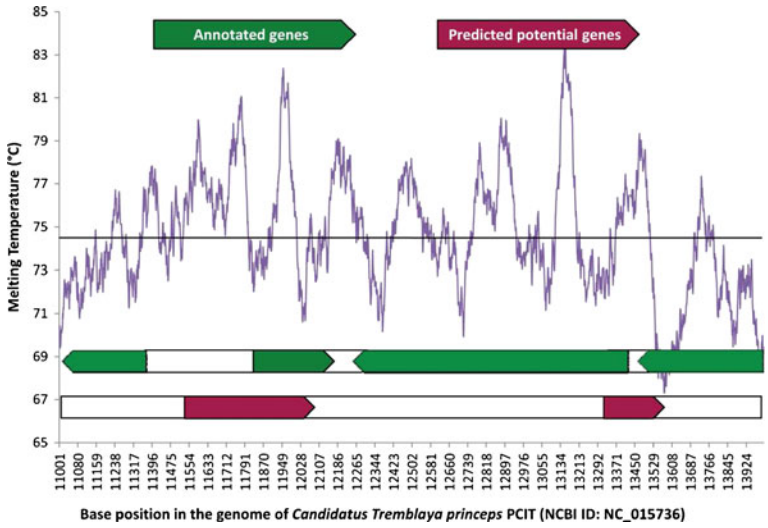

Figure 3. Melting profile and the annotated and predicted genes in the genome of *Candidatus Tremblaya princeps* PCIT. The black line represents the average melting temperature of the genome and arrows indicate the direction of transcription.

these sequences, we found that both these regions exhibit properties of a gene and were bracketed as potential new genes with the above-mentioned methodology. The number of gene-like sequences predicted after each step for all the 12 genomes considered in this study are presented in table 2.

Results in table 2 show that even after incorporating filters to remove DNA sequences that do not have genic character, a considerable number of new genes is obtained in each case. The number of sequences that match with the annotations of the genome in study are presented in column 8 (match with query genome) of table 2. Each of the predicted genes forms an ORF, so the loss of start codons, frame changes and hence error due to inclusion of pseudogenes is not an issue in the current study. Also, the predicted genes have passed the filters incorporating the stereo-chemical properties and tripeptides frequencies, both of which handle the sequences at the level of proteins. The total number of gene-like sequences predicted for each genome varies from a minimum of 1 in case of *Candidatus Sulcia muelleri* DMIN to a maximum of 5743 in case of *Rhodopseudomonas palustris* TIE-1. It may also be seen from table 2 that the number of genes predicted in each case is not affected either by the GC-content of the genomes or by the number of annotated genes in each case. For example, the genome of *Acaryochloris marina* MBIC11017 (6503724 base pairs) has a GC-content of 47.3% and the number of annotated genes is 2137 and the predicted number of new genes in this case is 586, while for two other genomes, *Prevotella denticola* F0289 (2937589 base pairs, 50.4% GC 2386 annotated genes) and *Acetobacter pasteurianus* IFO 3283-01 (2907495 base pairs, 53.0% GC 2682 annotated genes), the number of new genes predicted is 1218 and 1123 respectively, which is approximately double that predicted in the case of *Acaryochloris marina* MBIC11017 although the GC-content of these organisms does not vary much.

The size of the genome also has no effect on the number of new predictions as seen from table 2, where the smallest genome is of *Candidatus Tremblaya princeps,* which has a GC-content of 58.8% and a size of 138927 base pairs. The total number of new genes predicted in this case is 57, which is 50 times that predicted for another genome, *Candidatus Sulcia muelleri* DMIN (GC-content=22.5%), which is approximately double (243933 base pairs) of the smallest one but has only one predicted new gene. The number of sequences obtained after all the filters for each case is not even related to the number of sequences obtained in the initial step of the methodology involving melting temperature discrimination. It may be noticed from table 2 that for *Glaciecola nitratireducens* FR1064, the number of sequences obtained at the first step is 3654 and the final number attained is just 145, while in the case of *Acaryochloris marina* MBIC11017, the number of sequences extracted at the initial step is 2137 and those remaining after the complete methodology is 586, which is approximately four times that obtained for *Glaciecola nitratireducens* FR1064 . It is also evident from table 2 that there is a considerable reduction of potential new genes after each filter, demonstrating strongly the need for physico-chemical filters as considered in this study so as to keep the number of false-positives in each case to a minimum.

The genome of *Glaciecola nitratireducens* FR1064 (4134229 bp) has 145 new predicted genes, while that of *Bacillus subtilis* subsp. subtilis str. 168 (4215606) has 519 prediction. Although the size as well as the GC-content (42.3

and 43.5 respectively) of both of them are almost similar, the number of predictions in the latter is thrice that of the former, highlighting the lack of correlation of the GC-content and the size of genome to the number of new predictions.

It can be noticed from table 2 that *Rhodopseudomonas palustris* TIE-1 has the maximum number of predicted genes (5743), although it is not the largest genome in terms of base pairs (*Acaryochloris marina* MBIC11017 6503724 base pairs) or nor does it have the highest GC-content (*Isoptericola variabilis* 225 chromosome, 73.9% GC) among all the genomes considered in the current study, which reiterates the independence of prediction on the size or the GC-content of the genome.

The column illustrating new annotations in table 2 clearly reveals that there are a number of sequences in each genome that have not yet been annotated in the query genome although they show significant matches with proteins present in other species. These genes could have been horizontally transferred, and so they do not show characteristics similar to most of the protein coding sequences present in the query genome and hence are missed by *ab initio* gene prediction softwares, which are used to provide annotation of genomes as present in NCBI.

3.1 *A test case of Synthetic Mycoplasma genitalium JCVI-1.0 (Synthia)*

The first synthetic genome constructed by Venter and coworkers (Gibson *et al.* 2010), where the genome is completely annotated, was also adapted to determine the accuracy of the predictions made by this methodology. It was observed that the number of new genes predicted in this case was just 5, while the number of annotated genes for this organism is 483. The details of the prediction for Synthia are given in table 3. The number of annotated genes that lie in the higher stability regions of Synthia is 265.

When we compare this with the number of predicted potential genes, which in case of Synthia would be false-positives as the genome has been synthetically constructed and the functionality of each region is known, the error in prediction turns out to be 0.19, which is extremely low and the reliability of prediction is as high as 0.98. This shows that only 2% of the total genes predicted are false-positives, implying that the accuracy of the new gene prediction methodology presented is high.

The new annotations lie in both the overlapping (21 sequences) and the intergenic regions (9 sequences), while the new predictions are all present in the overlapping regions. It is further observed that the melting temperatures and hence the stability of new genes occurring in intergenic regions are marginally higher than those occurring in the overlapping regions in case of new annotations and are similar to those noted for the new genes.

Table 3. Results of the methodology on Synthetic Mycoplasma genitalium JCVI-1.0 showing the number of predicted sequences after each step

Genome	GC-content (%)	GenBank ID	Number of bases	Number of annotated CDS	Higher stability ORF	Group I	Match with query genome	New Annotations	Group II	J-vector	Stereo-chemical properties	Tripeptide frequency
Synthetic *Mycoplasma genitalium* JCVI-1.0	31.7	CP000925.1	582970	483	551	295	265	30	256	47	28	5

There are a number of sequences that match with proteins from other genomes but are not annotated in Synthia and add up to the error in new annotations as these might not be getting expressed. This error in new annotation in case of Synthia turns out to be 10%, which is calculated on the basis of highly stable ORF sequences matching with the annotations provided in NCBI.

3.2 *Analysis on Bacillus subtilis and Escherichia coli genomes*

When a similar analysis was done on two model genomes, *Bacillus subtilis* and *Escherichia coli*, the results were astounding as these genomes are highly studied and the annotations are supposed to be near complete. Several potential new genes were identified in these two cases as shown in table 4. The number of potential new genes predicted in *Escherichia coli* and *Bacillus subtilis* genomes are 856 and 519 respectively. Even if we consider the error in prediction (2% of the highly stable annotated genes as estimated from the completely annotated genome of Synthia), the number of potential new genes remaining would be 803 in case of *Escherichia coli* and 460 for *Bacillus subtilis*. Additionally, there are 1102 and 272 new annotations in *Escherichia coli* and *Bacillus subtilis* respectively, which are enormous in number even if one were to provide allowance for possible errors in new annotations.

It is extremely likely that all the potential genes and new annotations detected by this methodology are not true genes due to the lack of promoter or promoter-like sequences, absence of ribosomal binding sites and other features that prevent their expression. But, even if all the odds are taken into consideration, there would still be substantial number of sequences that could code for proteins. Additional filters will help resolve these issues between potentially protein coding regions vis-à-vis protein coding regions.

4. Conclusion

In a nutshell, this study clearly suggests that the complete potential of the physico-chemical properties of DNA has not yet been tapped, which is particularly relevant as a number of DNA sequences especially of the eukaryotic genomes are not fully annotated till date. More sophisticated methods need to be developed for gene prediction even for prokaryotes. We are learning to look at DNA from a different perspective other than just reading and comparing either the sequence characters or statistics based on them. There is considerable optimism that all the issues of gene prediction would be answered and methodologies that are universally applicable for all the genomes will be evolved.

Table 4. Analysis results on *Escherichia coli* and *Bacillus subtilis* genomes

Genome	GC-content (%)	NCBI ID	Number of bases	Number of annotated CDS	Higher stability ORF	Potential gene sequences obtained after each step						
						Group I	Match with query genome	New Annotations	Group II	J-vector	Stereo-chemical properties	Tripeptide frequency
Escherichia coli str. K-12 substr. MG1655	50.8	NC_000913	4639675	4320	6863	3752	2650	1102	3111	2347	1562	856
Bacillus subtilis subsp. subtilis str. 168	43.5	NC_000964	4215606	4176	5222	2750	2478	272	2472	1692	1053	519

Acknowledgements

The authors thank the Ascona B-DNA Consortium for providing the molecular dynamics simulation data and to NCBI for free access to genomic data and BLAST. Programme support to Supercomputing Facility for Bioinformatics & Computational Biology from Department of Biotechnology, Government of India, is greatly acknowledged. GK is a recipient of DBT-SRF.

References

Abeel T, Saeys Y, Rouzé P and de Peer YV 2008 ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. *Bioinformatics* **24** i24–i31

Alexandersson M, Cawley S and Pachter L 2003 SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.* **13** 496–502

Allen JE, Pertea M and Salzberg SL 2004 Computational Gene Prediction Using Multiple Sources of Evidence. *Genome Res.* **14** 142–148

Audic S and Claverie J-M 1998 Self-identification of protein-coding regions in microbial genomes. *Proc. Natl. Acad. Sci. USA* **95** 10026–10031

Besemer J and Borodovsky M 1999 Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.* **27** 3911–3920

Birney E and Durbin R 2000 Using GeneWise in the Drosophila annotation experiment. *Genome Res.* **10** 547–548

Baren MJ van and Brent MR 2006 Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Res.* **16** 678–685

Chatterji S and Pachter L 2006 Reference based annotation with GeneMapper. *Genome Biol.* **7** R29

Claverie JM, Poirot O and Lopez F 1997 The difficulty of identifying genes in anonymous vertebrate sequences. *Comput. Chem.* **21** 203–214

DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M and Galagan JE 2007 Conrad: Gene prediction using conditional random fields. *Genome Res.* **17** 1389–1398

Delcourt SG and Blake RD 1991 Stacking energies in DNA. *J. Biol. Chem.* **266** 15160–15169

Dhar PK, Thwin, ST, Tun K, Tsumoto Y, Maurer-Stroh, Eisenhaber F and Surana U 2009 Synthesizing non-natural parts from natural genomic template. *J. Biol. Engg.* **3** 2

Dineen DG, Wilm A, Cunningham P and Higgins DG 2009 High DNA melting temperature predicts transcription start site location in human and mouse. *Nucleic Acids Res.* **37** 7360–7367

Dixit SB, Beveridge DL, Case DA, Cheatham 3rd TE, Giudice E, Lankas F, Lavery R, Maddocks JH, *et al.* 2005 Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.* **89** 3721–3740

Dutta S, Singhal P, Agrawal P, Tomer R, Kritee, Khurana E, *et al.* 2006 A physico-chemical model for analyzing DNA sequences. *J. Chem. Inf. Model* **46** 78–85

Fickett JW 1982 Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* **10** 5303–5318

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness, EF, Kerlavage AR, *et al.* 1995 Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. *Science* **269** 496–512

Frishman D, Mironov A, Mewes HW and Gelfand M 1998 Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.* **26** 2941–2947

Gelfand, MS, Mironov AA and Pevzner PA 1996 Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci.* **93** 9061–9066

Gibson DG, Glass JI, Lartigue C, Noskov VN and Chuang R-Y 2010 Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329** 52–56

Glusman G, Qin S, El-Gewely MR, Siegel AF, Roach JC, Hood L, *et al.* 2006 Third approach to gene prediction suggests thousands of additional human transcribed regions. *PLoS Comput. Biol.* **2** e18

Gross SS and Brent MR 2006 Using multiple alignments to improve gene prediction. *J. Comput. Biol.***13** 379–393

Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, *et al.* 2006 EGASP: The human ENCODE genome annotation assessment project. *Genome Biol.* **7**S2

Harrow J, Denoeud F, Frankish A, Reymond A, Chen C-K, Chrast J, *et al.* 2006 GENCODE: Producing a reference annotation for ENCODE. *Genome Biol*. **7**S4

Huang Y and Kowalski D 2003 WEB-THERMODYN: sequence analysis software for profiling DNA helical stability. *Nucleic Acids Res.* **31** 3819–3821

Hunter CA 1993 Sequence-dependent dna-structure - the role of base stacking interactions. *J. Mol. Biol.* **230** 1025–1054

Jayaram B 1997 Beyond the wobble: the rule of conjugates. *J. Mol. Evol.* ***45*** 704–705.

Jayaram B 2008 Decoding the design principles of amino acids and the chemical logic of protein sequences. *Nat. Precedings* (*http://hdl.handle.net/10101/npre.2008.2135.1*)

Jayaram B and Beveridge DL 1990 Free Energy of an arbitrary charge distribution imbedded in coaxial cylindrical dielectric continua: Application to conformational preferences of DNA in aqueous solutions. *J. Phys. Chem*. **94** 4666–4671

Jensen KT, Petersen L, Falk S, Iversen P, Andersen P, Theisen M, *et al.* 2006 Novel overlapping coding sequences in *Chlamydia trachomatis*. *FEMS Microbiol Lett.* **265** 106–117

Kanhere A and Bansal M 2005a Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Res*. **33** 3165–3175

Kanhere A and Bansal M 2005b A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics* **6** 1–10

Keller O, Kollmar M, Stanke M and Waack S 2011 A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27** 757–763

Khandelwal G and Jayaram B 2010 A phenomenological model for predicting melting temperatures of DNA sequences. *PLoS ONE* **5** e12433

Knowles DG and McLysaght A 2009 Recent de novo origin of human protein-coding genes. *Genome Res.* **19** 1752–1759

Korf I, Flicek P, Duan D and Brent MR 2001 Integrating genomic homology into gene structure prediction. *Bioinformatics* **17** S140-S148

Lavery R, Zakrzewska K, Beveridge DL, Bishop TC, Case TA, Cheatham IIIT, Dixit S, Jayaram B, *et al.* 2009 A systematic molecular dynamics study of nearest neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.* **38** 299–313.

Lin S and Kowalski D 1994 DNA helical instability facilitates initiation at the SV40 replication origin. *J. Mol. Biol.* **235** 496–507

Maeda Y and Ohtsubo E 1987 Relationship between helix-coil transition and gene organization of ColEl plasmid DNA differential scanning calorimetric and theoretical studies. *J. Mol. Biol*. **194** 691–698

Mathé C, Sagot M-F, Schiex T and Rouzé P 2002 Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **30** 4103–4117

Meyer IM and Durbin R 2004 Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res.* **32** 776–783

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

Natale DA, Umek RM and Kowalski D 1993 Ease of DNA unwinding is a conserved property of yeast replication origins. *Nucleic Acids Res.* **21** 555–560

O'Donovan C, Martin MJ, Gattiker A, Gasteiger, E, Bairoch A and Apweiler R 2002 High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **3** 275–284

Owczarzy R, Vallone PM, Goldstein RF and Benight AS 1999 Studies of DNA dumbbells VII: Evaluation of the next nearest-neighbor sequence-dependent interactions in duplex DNA. *Biopolymers* **52** 29–56

Pagani I, Konstantinos L, Jansson J, Chen I-Min A, Smirnova T, Bahador N, *et al.* 2012 The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **40** D571-D579

Panjkovich A and Melo F 2005 Comparison of different melting temperature calculation methods for short DNA sequences. *Bioinformatics* **21** 711–722

Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, *et al.* 2010 GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat. Methods* **7** 455–457

Protozanova E, Yakovchuk P and Frank-Kamenetskii MD 2004 Stacked-unstacked equilibrium at the nick site of DNA DOI: dx.doi.org . *J. Mol. Biol.* **342** 775–785

Rangannan V and Bansal M 2007 Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability. *J. Biosci.* **32** 851–862

SantaLucia J Jr 1998 A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* **95** 1460–1465

Shah SP, McVicker GP, Mackworth AK, Rogic S and Ouellette BFF 2003 GeneComber: combining outputs of gene prediction programs for improved results. *Bioinformatics* **19** 1296–1297

Siepel A 2009 Darwinian alchemy: Human genes from noncoding DNA. *Genome Res.* **19** 1693–1695

Singhal P, Jayaram B, Dixit SB and Beveridge DL 2008 Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. *Biophys. J.* **94** 4173–4183

Sponer J, Leszczynski J and Hobza P 2001 Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers* **61** 3–31

Sponer J, Jurecka P and Hobza P 2004 Accurate interaction energies of hydrogen-bonded nucleic acid base pairs. *J. Am. Chem. Soc.* **126** 10142–10151

Stanke M, Steinkamp R, Waack S and Morgenstern B 2004 AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**W309-W312

Stanke M, Diekhans M, Baertsch R and Haussler D 2008 Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24** 637–644

Stormo GD, Schneider TD, Gold L and Ehrenfeucht A 1982 Use of the 'Perceptron' algorithm to distinguish translation initiation site in *E. coli. Nucleic Acids Res.* **10** 2997–3011

Sugimoto N, Nakano S, Yoneyama M and Honda K 1996 Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.* **24** 4501–4505

Tech M and Meinicke P 2006 An unsupervised classification scheme for improving predictions of prokaryotic TIS. *BMC Bioinformatics* **7** 121

The UniProt Consortium 2011 Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **39** D214-D219

Umek RM and Kowalski D 1988 The ease of DNA unwinding as a determinant of initiation at yeast replication origins. *Cell* **52** 559–567

Wada A and Suyama A 1983 Correlation between physical stability maps and genetic map of DNA double strand. *J. Phys. Soc. Jpn.* **52** 4417–4422

Wada A and Suyama A 1984a Stability distribution in the phage g-DNA double helix: A correlation between physical and genetic structure. *J. Biomol. Struct. Dyn.* **2** 573–591

Wada A and Suyama A 1984b Variation of double-helix stability along DNA molecular thread and its biological implications: Homostabilizing propensity of gene double-helix; in *Molecular basis of cancer* (ed) R Rein (New York: Alan R. Liss Inc.) pp 37–46

Wada A and Suyama A 1985a Homogeneous double-helix-stability in individual genes; in *4th Conversation in Biomolecular Stereodynamics* (ed) RH Sarma (State University of New York at Albany) p 65

Wada A and Suyama A 1986 Local stability of DNA and RNA secondary structure and its relation to biological functions. *Prog. Biophys. Mol. Biol.* **47** 113–157

Wu J, Hu Z and DeLisi C 2006 Gene annotation and network inference by phylogenetic profiling. *Bioinformatics* **7** 80

Yakovchuk P, Protozanova E and Frank-Kamenetskii MD 2006 Base-stacking and base-pairing contributions into thermal stability of the DNA double helix DOI:dx.doi.org . *Nucleic Acids Res.* **34** 564–574

Yeh R-F, Lim LP and Burge CB 2001 Computational inference of homologous gene structures in the human genome. *Genome Res.* **11** 803–816

Yok NG and Rosen GL 2011 Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinformatics* **12** 20

Yu GX, Snyder EE, Boyle SM, Crasta OR, Czar M, Mane SP, *et al.* 2007 A versatile computational pipeline for bacterial genome annotation improvement and comparative analysis, with Brucella as a use case. *Nucleic Acids Res.* **35** 3953–3962

Zhu HQ, Hu GQ, Ouyang ZQ, Wang J and She ZS 2004 Accuracy improvement for identifying translation initiation sites in microbial genomes. *Bioinformatics* **20** 3308–3317

Zhu HQ, Hu GQ, Yang YF, Wang J, and She ZS 2007 MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. *BMC Bioinformatics* **8** 97