A Physicochemical Model for Analyzing DNA Sequences

Samrat Dutta, Poonam Singhal, Praveen Agrawal, Raju Tomer, Kritee, Ekta Khurana, and B. Jayaram*

Department of Chemistry and Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology, Hauz Khas, New Delhi-110016, India

Received April 11, 2005

In search of an ab initio model to characterize DNA sequences as genes and nongenes, we examined some physicochemical properties of each trinucleotide (codon), which could accomplish this task. We constructed three-dimensional vectors for each double-helical trinucleotide sequence considering hydrogen-bonding energy, stacking energy, and a third parameter, which we provisionally identified with DNA−protein interactions. As this three-dimensional vector moves along any genome, the net orientation of the resultant vector should differ significantly for gene and nongene regions to make a distinction feasible, if the underlying model has some merits. An analysis of 331 prokaryotic genomes comprising a total of 294 786 experimentally verified genes (nonoverlapping) and an equal number of nongenes presents a proof of concept of the model without the need for further parametrization. Also, initial analyses on *Saccharomyces cerevisiae* and *Arabidopsis thaliana* suggest that the methodology is extendable to eukaryotes. The physicochemical model (*ChemGenome1.0*) introduced has the potential to be developed into a gene-finding algorithm and, more pressingly, could be employed for an independent assessment of the annotation of DNA sequences.

I. INTRODUCTION

The regulation of gene expression is a matter of chemistry between DNA and proteins at the molecular level. While remarkable advances have been made over the past two decades in the analysis of DNA sequences and in gene prediction in particular, via statistical and mathematical models and artificial intelligence techniques based on genome, gene, cDNA, and protein sequence databases and the clever design of computational protocols,[1−28] an expeditious in silico gene-finding model which directly captures the physicochemical properties intrinsic to DNA sequences and the chemistry of protein−DNA interactions remains a goal yet to be realized. Proceeding along these lines, we sought to look for some simplifying universal principles working behind deciding "what can be a gene" in any species. Working with the hypothesis that both the structure of the DNA and its interactions with regulatory proteins and polymerases decide the function of a DNA sequence, we developed a simple three-parameter model based on Watson−Crick hydrogen-bonding energy, base-pair stacking energy, and a third parameter which we provisionally identified with DNA−protein interactions. Each of these parameters acts as a dimension for a three-dimensional unit vector, whose orientation differs for each trinucleotide. The premise that the cumulative vectors for gene and nongene regions should differ in orientation (Figure 1) stands verified on 331 prokaryotic genomes and 21 eukaryotic genomes. We introduce, here, the physicochemical model for analyzing DNA sequences, present a series of validation tests on a large number of genomes, and examine its merits and limitations and its potential utility in genome analyses.

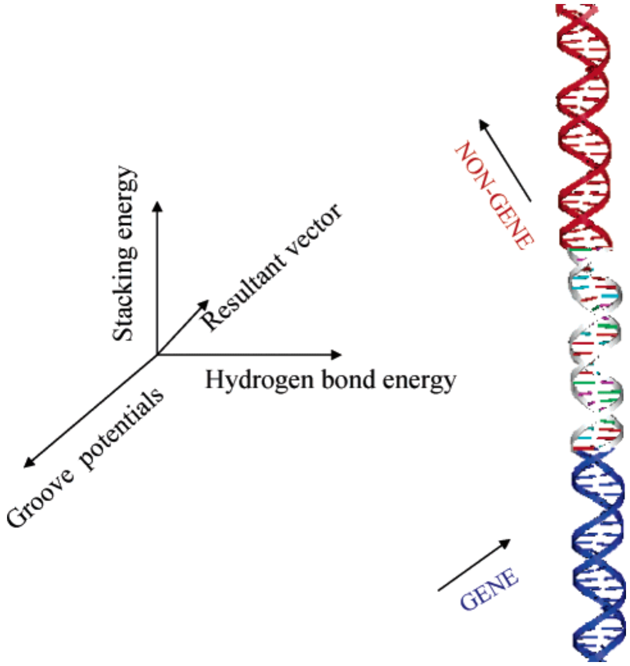

Figure 1. Physicochemical model for analyzing DNA sequences and the hypothesis for genome characterization as genes and nongenes.

II. METHODS

The physicochemical model proposed involves developing a three-dimensional (3-D) vector for double-helical deoxyribonucleic acid (DNA) base sequences, with each dimension representing one facet of DNA recognition[29] by proteins. Each of the 64 trinucleotides is assigned three coordinates,

* Author to whom correspondence should be addressed. Tel.: +91-11-2659 1505, +91-11-2659 6786. Fax: +91-11-2658 2037. E-mail: bjayaram@chemistry.iitd.ac.in.

10.1021/ci050119x CCC: $33.50 © 2006 American Chemical Society
Published on Web 10/12/2005

x, y, and z, in the interval of -1 to $+1$, ($x, y, z \in [-1, +1]$), corresponding to the three proposed chemical properties of DNA. For a given DNA sequence (genome segment), the resultant vector is found by accumulating the x, y, and z components of the individual codons ($X = \sum x$, $Y = \sum y$, $Z = \sum z$). The orientation of this resultant vector from the origin is given by the direction cosines. The x, y, and z parameters for each codon are developed as follows.

Step 1. The Watson−Crick (WC) hydrogen-bond energies between bases in a base pair embedded in B-DNA in an aqueous environment calculated using the finite difference Poisson−Boltzmann[30] (FDPB) method gave a value of ~ -2 kcal/mol/H bond (Jayaram and Honig, unpublished data, 1989). The number of WC hydrogen bonds formed by each codon in the double helix are counted, converted to energies, and mapped onto the $[-1, 1]$ interval giving the x coordinates.

$$E[i] = (\text{no. of H bonds} \times 2\ \text{kcal/mol}), \qquad i = 1-64$$

$$x[i] = [\{E[i] - \text{median}(E)\}/\{\text{maximum}(E) - \text{median}(E)\}]$$

E is the set of magnitudes of H-bond energies for all 64 codons, and x is the corresponding WC H-bond parameter. We observed earlier that WC hydrogen bonds and stacking could largely account for the dynamics and flexibility of codons,[31] offering an explanation for the effects of wobble during translation.

Step 2. The electrostatic, van der Waals, and hydrophobic contributions to stacking energies were calculated for the 32 unique double-helical trinucleotide sequences after building the structures in canonical B-DNA form and geometrically optimizing them.[32] AMBER[33] force-field parameters were used for all the calculations. The electrostatic contribution was computed using the FDPB method. The resultant energies were then linearly mapped onto the interval $[-1, 1]$ as shown in Table 1, giving the y coordinates for each trinucleotide.

$$E[i] = \text{stacking energy for } i\text{th codon}, \qquad i = 1-64$$

$$y[i] = [\{E[i] - \text{median}(E)\}/\{\text{maximum}(E) - \text{median}(E)\}]$$

Step 3. To obtain the z coordinates for each codon, a training set of 1500 gene/nongene (shifted-gene) pairs (where a gene is at least 100 nucleotides long) of the *E. coli* K12 genome[34] was used. The z parameters were optimized to give the best separation in orientation between the gene and nongene vectors. This can be achieved if we send the z components of the gene and nongene vectors to the extreme opposite ends on the unit sphere.

Let $z = \sum f_i z_i$, where f_i is the frequency fraction of the ith codon, and z_i is ith codon's z value. Since $-1 \le z_i \le 1$ and $\sum f_i = 1$, then $-1 \le z \le 1$.

Therefore, the problem of maximizing separation can be formulated as

$$\min \sum_{\text{all pairs}} [(z - 1)^2 + (z' + 1)^2]$$

where $-1 \le z_i \le 1$, $z = \sum f_i z_i$, and $z' = \sum f_i' z_i$. In the above function, the z component of the genes is being sent to the $+1$ extreme and that of the nongenes (z') is being sent to -1. This is a convex positive valued function and, hence, has its global minimum. We have used the "steepest descent method for constrained objective (optimization) space"[35] to locate the minimum and, thus, optimized the z parameters. We found that all the z values, without exception, were either $+1$ or -1 (though the optimization was carried out in the continuous space $[-1, 1]$), which gives the impression that some codons are more favorable to gene character and others to nongene character. To test this hypothesis, we then retrained the z parameters on 62 microbial genomes individually and recovered essentially the same z values (of 64 codons) across the 62 microbial genomes. This indicates that z values are capturing some inherent DNA structural/functional properties and are not merely database-trained parameters. We also noticed that the z values obtained were, by and large, consistent with the rule of conjugates[31] proposed earlier. The conjugate rule captures the observed quartet degeneracy without exception and is a macro-level manifestation of the molecular-level interactions at the decoding site in the translation step of gene expression. According to the rule of conjugates, adenine (A) is the conjugate of cytosine (C) and guanine (G) is the conjugate of thymine (T). A codon and its corresponding conjugate codon have equal and opposite values for the z parameter (namely, $+1$ or -1). A -1 value for a codon does not mean that it occurs only in the nongenes. In the training (optimization) function, the z values are weighted using gene/nongene fractions and the minimum of that function is favored by the corresponding $+1/-1$ values. Use of the fraction just

Table 1. Assigned Values for the x, y, and z Coordinates for Each of the 64 Codons

CODON	x	y	z	CODON	x	y	z
CCC	1.0	1.0	−1	TCC	0.33	0.16	−1
CCG	1.0	0.95	−1	TCG	0.33	0.13	−1
CCT	0.33	0.39	1	TCT	−0.33	−0.12	−1
CCA	0.33	0.27	−1	TCA	−0.33	−0.54	−1
CGC	1.0	0.72	1	TGC	0.33	−0.01	−1
CGG	1.0	0.95	−1	TGG	0.33	0.27	−1
CGT	0.33	0.17	1	TGT	−0.33	−0.41	−1
CGA	0.33	0.13	−1	TGA	−0.33	−0.54	−1
CTC	0.33	−0.11	−1	TTC	−0.33	−0.30	−1
CTG	0.33	−0.06	−1	TTG	−0.33	−0.28	−1
CTT	−0.33	−0.12	1	TTT	−1.0	−0.24	1
CTA	−0.33	−0.26	−1	TTA	−1.0	−0.37	−1
CAC	0.33	−0.19	1	TAC	−0.33	−0.38	−1
CAG	0.33	−0.06	−1	TAG	−0.33	−0.26	−1
CAT	−0.33	−0.18	1	TAT	−1.0	−0.30	1
CAA	−0.33	−0.28	1	TAA	−1.0	−0.37	−1
GCC	1.0	0.72	1	ACC	0.33	0.26	−1
GCG	1.0	0.72	−1	ACG	0.33	0.17	−1
GCT	0.33	0.16	1	ACT	−0.33	−0.34	1
GCA	0.33	−0.01	1	ACA	−0.33	−0.41	−1
GGC	1.0	0.72	1	AGC	0.33	0.16	1
GGG	1.0	1.0	−1	AGG	0.33	0.39	−1
GGT	0.33	0.26	1	AGT	−0.33	−0.34	1
GGA	0.33	0.16	1	AGA	−0.33	−0.12	1
GTC	0.33	−0.22	1	ATC	−0.33	−0.21	1
GTG	0.33	−0.19	1	ATG	−0.33	−0.18	−1
GTT	−0.33	−0.25	1	ATT	−1.0	−0.14	1
GTA	−0.33	−0.38	1	ATA	−1.0	−0.30	−1
GAC	0.33	−0.22	1	AAC	−0.33	−0.25	1
GAG	0.33	−0.11	1	AAG	−0.33	−0.12	−1
GAT	−0.33	−0.21	1	AAT	−1.0	−0.14	1
GAA	−0.33	−0.30	1	AAA	−1.0	−0.24	1

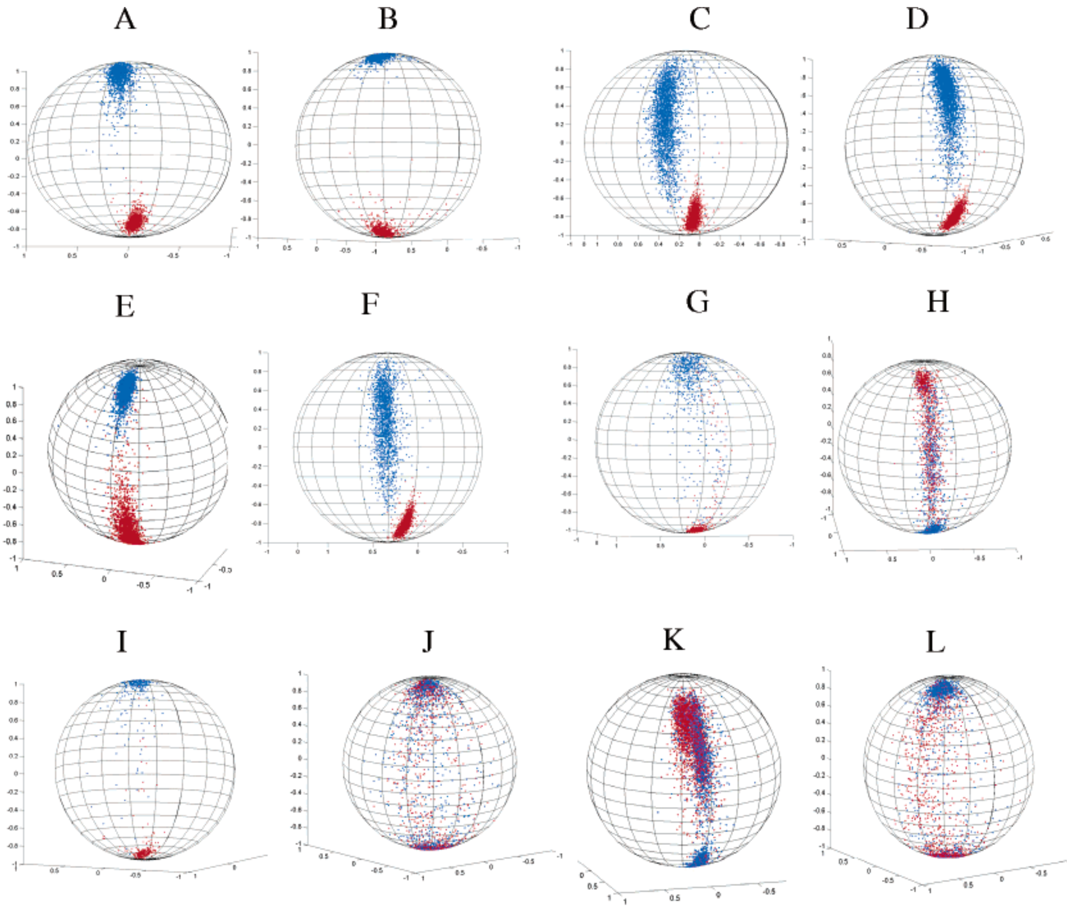

Figure 2. Three-dimensional plots of the distributions of gene and nongene direction vectors for the six best (A−F) and six worst (G−L) cases calculated from the genomes of (A) *Agrobacterium tumefaciens* (NC_003304), (B) *Wolinella succinogenes* (NC_005090), (C) *Rhodopseudomonas palustris* (NC_005296), (D) *Bordetella bronchiseptica* (NC_002927), (E) *Clostridium acetobutylicium* (NC_003030), (F) *Bordetella pertusis* (NC_002929), (G) *Thermococcus kodakaraensis* (NC_006624), (H) *Brucella suis* (NC_004310), (I) *Pyrococcus horikoshii* (NC_000961), (J) *Bacillus subtilis* (NC_000964), (K) *Mesorhizobium loti* (NC_002678), and (L) *Bacillus halodurans* (NC_002570).. Points colored blue correspond to genes, and those colored red correspond to nongenes. The plot is generated using MATLAB (release 13, version 6.5).

implies favorability of a codon toward gene/nongene character.

The adherence of z values to the rule of conjugates prompted us to search for some consistent patterns in the major groove of each base pair. Currently, work is in progress in this direction. We also note that the x, y, and z parameters fall into 31 unique sets. Out of them, 18 are found to follow symmetric considerations, suggesting that there is room for further improvement of the parameters. Table 1 gives the values for the x, y, and z parameters for each codon of the physicochemical model.

Orthogonalization of the Model Parameters. We noticed that the set of x, y, and z parameters developed above were not mutually orthogonal. We, therefore, orthogonalized these parameters by performing the method of “successive partialling”[36,37] on the obtained $x-y-z$ values. We took z as the first predictor, then residualized y with z, and finally residualized x with y and z combined. However, the results presented in Figure 2 and Tables 2 and 3 and discussed below differ only marginally whether orthogonalization is carried out. Hence, for the sake of simplicity and clarity of interpretation, we carried out our analysis with unorthogonalized x, y, and z values, as listed in Table 1.

Finding the Best Plane Dividing Genes from Nongenes. The best plane is generated for every genome using a pocket algorithm[38] (modified perceptron algorithm), which is a modification of perceptron learning[39] that makes perceptron learning well-behaved with nonseparable training data.

III. RESULTS

We tested the physicochemical model on 331 prokaryotic genomes, following the experimentally verified Genbank data.[40] Gene regions and each of their nongene (frame-shifted gene) regions in these genomes are separated, and the orientation of the resultant vectors for each DNA segment are calculated. The clusterings of gene and nongene points obtained in the best six and worst six cases are presented in Figure 2 (A−L). A clear separation of genes (blue) from nongenes (red) is discernible in almost all of the cases. By focusing on experimentally verified genes, the problem of overannotation in some prokaryotes is circumvented for the purposes of evaluating the model. Despite the simplicity of the model, the segregation of genes and nongenes noticed across different prokaryotic genera and species is significant. It is also clear from Figure 2 (G−L) that, even in the worst cases, clustering of genes and nongenes does occur.

The gene/nongene (frame-shifted gene) separation accuracies for various species are quantified and presented in Table 2 after finding the best plane for each genome that divides the unit sphere into gene and nongene hemispheres. It is

Table 2. Gene Evaluation Data[a] for Prokaryotic Genomes for Experimentally Verified Genes (Nonoverlapping) and Shifted Genes

species number	NCBI_ID	species name	genes	TP[b]	FP[b]	SS[b]	SP[b]	CC[b]
1	NC_000117	*Chlamydia trachomatis*	463	458	4	0.98	0.99	0.98
2	NC_000853	*Thermotoga maritima* MSB8	641	619	3	0.96	0.99	0.96
3	NC_000854	*Aeropyrum pernix* K1	561	532	7	0.94	0.98	0.93
4	NC_000868	*Pyrococcus abyssi* GE5	632	630	241	0.99	0.63	0.49
5	NC_000907	*Haemophilus influenzae*	955	953	7	0.99	0.99	0.99
6	NC_000908	*Mycoplasma genitalium* G-37	189	186	2	0.98	0.98	0.97
7	NC_000909	*Methanocaldococcus janaschii*	720	708	9	0.98	0.98	0.97
8	NC_000912	*Mycoplasma pneumoniae* M129	243	241	2	0.99	0.99	0.98
9	NC_000913	*Escherichia coli* K12	2759	175	659	0.63	0.72	0.39
10	NC_000915	*Helicobacter pylori*	731	727	4	0.99	0.99	0.98
11	NC_000916	*Methanobacterium thermoautotrophicum*	719	711	4	0.98	0.99	0.98
12	NC_000917	*Archaeoglobus fulgidus*	782	774	8	0.98	0.98	0.97
13	NC_000917	*Archaeoglobus fulgidus* DSM4304	782	774	8	0.98	0.98	0.98
14	NC_000918	*Aquifex aeolicus* VF5	584	575	3	0.98	0.99	0.97
15	NC_000921	*Helicobacter pylori* strain J99	658	648	9	0.98	0.98	0.97
16	NC_000922	*Chlamydophila pneumoniae* CWL029	597	590	9	0.98	0.98	0.97
17	NC_000948	*Borrelia burgdorferi* B31 plsmids cp32−1	11	11	0	1.0	1.0	1.0
18	NC_000949	*Borrelia burgdorferi* B31 plsmids cp32−3	11	11	0	1.0	1.0	1.0
19	NC_000950	*Borrelia burgdorferi* B31 plsmids cp32−4	11	11	0	1.0	1.0	1.0
20	NC_000951	*Borrelia burgdorferi* B31 plsmids cp32−6	10	10	0	1.0	1.0	1.0

[a] Data for the first 20 genomes in the order of NCBI IDs are shown in this table. Data for all 331 genomes are provided in Supporting Information Table 1 (Table S1). [b] True positives (TP): genes evaluated as genes. False positives (FP): nongenes evaluated as genes. True negatives (TN): nongenes evaluated as nongenes. False negatives (FN): genes evaluated as nongenes. Number of actual positives (AP) = TP + FN. Number of actual negatives (AN) = FP + TN. Predicted number of positives (PP) = TP + FP. Predicted number of negatives (PN) = TN + FN. Sensitivity (SS) = TP/(TP + FN). Specificity (SP) = TP/(TP + FP). Correlation − coefficient = $(\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN})/\sqrt{\mathrm{AN}\times\mathrm{PP}\times\mathrm{AP}\times\mathrm{PN}}$.

Table 3. Gene Evaluation Data[a] for Prokaryotic Genomes for Experimentally Verified Genes (Nonoverlapping) and Pregenes

species number	NCBI_ID	species name	genes	TP	FP	SS	SP	CC
1	NC_000117	*Chlamydia trachomatis*	463	426	93	0.92	0.82	0.73
2	NC_000853	*Thermotoga maritima* MSB8	641	579	149	0.90	0.79	0.67
3	NC_000854	*Aeropyrum pernix* K1	561	441	113	0.79	0.80	0.67
4	NC_000868	*Pyrococcus abyssi* GE5	632	329	8	0.52	0.98	0.45
5	NC_000907	*Haemophilus influenzae*	955	898	120	0.94	0.88	0.82
6	NC_000908	*Mycoplasma genitalium* G-37	189	175	22	0.92	0.89	0.80
7	NC_000909	*Methanocaldococcus janaschii*	720	646	117	0.90	0.85	0.80
8	NC_000912	*Mycoplasma pneumoniae* M129	243	218	53	0.90	0.80	0.69
9	NC_000913	*Escherichia coli* K12	2759	1946	645	0.70	0.75	0.47
10	NC_000915	*Helicobacter pylori*	731	679	155	0.93	0.81	0.71
11	NC_000916	*Methanobacterium thermoautotrophicum*	719	675	190	0.94	0.78	0.68
12	NC_000917	*Archaeoglobus fulgidus*	782	724	251	0.92	0.74	0.61
13	NC_000917	*Archaeoglobus fulgidus* DSM4304	782	724	251	0.92	0.74	0.61
14	NC_000918	*Aquifex aeolicus* VF5	584	534	186	0.91	0.74	0.62
15	NC_000921	*Helicobacter pylori* strain J99	658	615	109	0.93	0.85	0.76
16	NC_000922	*Chlamydophila pneumoniae* CWL029	597	573	170	0.96	0.77	0.72
17	NC_000948	*Borrelia burgdorferi* B31 plsmids cp32−1	11	11	1	1.0	0.92	0.90
18	NC_000949	*Borrelia burgdorferi* B31 plsmids cp32−3	11	11	0	1.0	1.0	1.0
19	NC_000950	*Borrelia burgdorferi* B31 plsmids cp32−4	11	10	1	0.91	0.91	0.80
20	NC_000951	*Borrelia burgdorferi* B31 plsmids cp32−6	10	10	0	1.0	1.0	1.0

[a] Data for the first 20 genomes in the order of NCBI IDs are shown in this table. Data for all 331 genomes are provided in Supporting Information Table 2 (Table S2).

noteworthy that the computed correlation coefficients are ≥ 0.90 or better for more than 300 genomes and 1.0 for 92 genomes, even when fixed values for x, y, and z were used for all 331 genomes. We also tested our model for experimentally verified genes versus the nongenic regions preceding them (pregenes) for 331 genomes (Table 3). For more than 80% of the genomes, separation sensitivity is ≥ 0.9. Further, on analyzing the worst case genomes, we checked for conformity of the gene sequences with some basic gene characteristics (a gene sequence is expected to have an initiation codon, a termination codon, and no internal stop codons). We found a large number of such nonconforming sequences in the worst case genomes and a negligibly low number in the best case genomes. We reperformed the analysis for the worst case genomes after removing such nonconforming sequences and found the separation correlations to be ≥ 0.9 (e.g., NC_000961 previous accuracy = 0.53, revised accuracy = 0.96; NC_002678 previous accuracy = 0.38, revised accuracy = 1.0). Thus, the results obtained using a fixed set of (x, y, z) parameters are promising for each genome without exception, which reconfirms our belief that the physicochemical model introduced is ab initio and independent of any species-dependent codon biasing.

The specificities and sensitivities[41] achieved in gene versus nongene separation with this simple physicochemical model

Table 4. Gene Evaluation Data for 21 Eukaryotic Genomes for Experimentally Verified tRNA Genes (Nonoverlapping) and Pregenes

species number	NCBI_ID	species name	genes	TP	FP	SS	SP	CC
1	NC_001133	*S. cerevisiae* chromosome I	6	5	0	0.83	1.0	0.91
2	NC_001134	*S. cerevisiae* chromosome II	14	14	0	1.0	1.0	1.0
3	NC_001135	*S. cerevisiae* chromosome III	12	11	0	0.92	1.0	0.95
4	NC_001136	*S. cerevisiae* chromosome IV	31	31	0	1.0	1.0	1.0
5	NC_001137	*S. cerevisiae* chromosome V	20	19	1	0.95	0.95	0.95
6	NC_001138	*S. cerevisiae* chromosome VI	12	12	0	1.0	1.0	1.0
7	NC_001139	*S. cerevisiae* chromosome VII	38	38	0	1.0	1.0	1.0
8	NC_001140	*S. cerevisiae* chromosome VIII	11	11	0	1.0	1.0	1.0
9	NC_001141	*S. cerevisiae* chromosome IX	10	10	1	1.0	0.91	0.95
10	NC_001142	*S. cerevisiae* chromosome X	26	26	0	1.0	1.0	1.0
11	NC_001143	*S. cerevisiae* chromosome XI	19	18	0	0.95	1.0	0.97
12	NC_001144	*S. cerevisiae* chromosome XII	24	22	4	0.92	0.85	0.87
13	NC_001145	*S. cerevisiae* chromosome XIII	25	24	1	0.96	0.96	0.96
14	NC_001146	*S. cerevisiae* chromosome XIV	18	18	0	1.0	1.0	1.0
15	NC_001147	*S. cerevisiae* chromosome XV	26	26	1	1.0	0.96	0.98
16	NC_001148	*S. cerevisiae* chromosome XVI	17	17	0	1.0	1.0	1.0
17	NC_003070	*A. thaliana* chromosome I	239	239	5	1.0	0.98	0.99
18	NC_003071	*A. thaliana* chromosome II	96	90	2	0.94	0.98	0.96
19	NC_003074	*A. thaliana* chromosome III	93	92	1	0.99	0.99	0.99
20	NC_003075	*A. thaliana* chromosome IV	79	77	1	0.97	0.99	0.98
21	NC_003076	*A. thaliana* chromosome V	108	108	1	1.0	0.99	0.99

Table 5. Gene Evaluation Data for 21 Eukaryotic Genomes for Experimentally Verified Genes[a] (Nonoverlapping Coding Sequences) and Pregenes

species number	NCBI_ID	species name	genes	TP	FP	SS	SP	CC
1	NC_001133	*S. cerevisiae* chromosome I	87	73	10	0.84	0.88	0.74
2	NC_001134	*S. cerevisiae* chromosome II	355	273	26	0.77	0.91	0.72
3	NC_001135	*S. cerevisiae* chromosome III	134	102	9	0.76	0.92	0.72
4	NC_001136	*S. cerevisiae* chromosome IV	667	606	62	0.91	0.91	0.82
5	NC_001137	*S. cerevisiae* chromosome V	229	202	19	0.88	0.91	0.81
6	NC_001138	*S. cerevisiae* chromosome VI	100	86	10	0.86	0.89	0.78
7	NC_001139	*S. cerevisiae* chromosome VII	449	398	43	0.89	0.90	0.80
8	NC_001140	*S. cerevisiae* chromosome VIII	252	219	28	0.87	0.89	0.77
9	NC_001141	*S. cerevisiae* chromosome IX	170	143	17	0.84	0.89	0.76
10	NC_001142	*S. cerevisiae* chromosome X	298	279	81	0.94	0.77	0.71
11	NC_001143	*S. cerevisiae* chromosome XI	270	223	10	0.82	0.96	0.81
12	NC_001144	*S. cerevisiae* chromosome XII	411	334	33	0.81	0.91	0.76
13	NC_001145	*S. cerevisiae* chromosome XIII	408	290	23	0.71	0.93	0.68
14	NC_001146	*S. cerevisiae* chromosome XIV	346	298	40	0.86	0.88	0.76
15	NC_001147	*S. cerevisiae* chromosome XV	457	367	38	0.80	0.91	0.74
16	NC_001148	*S. cerevisiae* chromosome XVI	397	355	50	0.89	0.88	0.78
17	NC_003070	*A. thaliana* chromosome I	38 568	28 882	1810	0.75	0.94	0.62
18	NC_003071	*A. thaliana* chromosome II	21797	17 873	1759	0.82	0.91	0.65
19	NC_003074	*A. thaliana* chromosome III	27 611	22 496	2166	0.81	0.91	0.65
20	NC_003075	*A. thaliana* chromosome IV	22 006	17 535	1491	0.80	0.92	0.64
21	NC_003076	*A. thaliana* chromosome V	30 924	24 070	2015	0.78	0.92	0.65

[a] Exons are treated as genes.

are comparable to the more sophisticated mathematical models built into the extant gene-finding algorithms. Tables 2 and 3 together convey the utility of the physicochemical model for a database-independent evaluation of genome annotation.

Shorter fragments of DNA, as in genes for tRNAs, are particularly problematic with conventional models. The physicochemical model when tested on experimentally verified tRNA genes in the cases of *Saccharomyces cerevisiae* and *Arabidopsis thaliana* yielded 98.26% true positives (Table 4).

Application of the physicochemical model to exonic regions of *A. thaliana* resulted in sensitivity, specificity, and correlation coefficients of 0.75, 0.94, and 0.62 for chromosome I and 0.82, 0.91, and 0.65 for chromosome II, respectively. Similar results were obtained for chromosomes III, IV, and V. Also, initial analyses of all gene/nongene regions in the 16 chromosomes of *S. cerevisiae*, a eukaryote (Table 5), show promise for the general applicability of the methodology regardless of the source of the genome. An analysis of chromosome I of *A. thaliana* at the gene level resulted in sensitivity, specificity, and correlation coefficients of 0.82, 0.81, and 0.62, respectively. We also tested our model on viruses where different open reading frames result in different gene products. Application of the model to more than 100 such genes (overlapping genes) and their pregenes results mostly in sensitivity, specificity, and correlation coefficients of 1.0, 1.0, and 1.0, respectively (Tables 6 and 7), thus, further validating the model.

A question arises whether the percentage of GC content, which provides an even simpler model for gene/nongene separation in prokaryotes, can produce results comparable

Table 6. *A. thaliana* (Exonic Region vs Pregene and Introns)

software	sensitivity	specificity
ChemGenome1.0	0.75	0.94
GeneMark.hmm	0.82	0.77
GenScan	0.63	0.70
MZEF	0.48	0.49
FGENF	0.55	0.54
Grail	0.44	0.38
FEX	0.55	0.32
FGENESP	0.42	0.59

to those of the current model, and the answer is in the negative. Gene sequences and the corresponding frame-shifted (nongene) sequences have nearly the same percentage GC. So, for the percentage GC algorithm, the two sequences are functionally similar.

The accuracies of the calculated correlation coefficients of most of the existing models depend on annotation accuracies in the database. For the case of *Aeropyrum pernix*, the calculated sensitivity, specificity, and correlation coefficients were 0.77, 0.97, and 0.91, when calculations were done relative to the existing annotated data with the physicochemical model. The results improved when the physicochemical model was employed for experimentally verified genes, and sensitivity, specificity, and correlation coefficients of 0.94, 0.98, and 0.93, respectively, were recovered. This can possibly be due to overannotation, as widely believed.[42] The quality of annotation in sequenced microbial genomes is presented in a recent commentary.[43]

IV. DISCUSSION

A novel physicochemical model is developed and tested on 331 prokaryotic genomes wherein genes and nongenes (shifted-genes and pregenes) could be distinguished with separation accuracies comparable to those of the existing methods. In the case of prokaryotes, the pregene regions are typically very small and consist of regulatory or other signal sites; hence, we considered shifted genes as nongenes for the purpose of performance appraisal of the model (Table 2). Moreover, when we checked the separation accuracies between the gene and pregene regions, the specificity tended to drop slightly (Table 3), that is, an increase in the false positives, which indicates that the physicochemical model captures the regulatory sites as well. As a matter of fact, shifted genes contain stop codons and are never expressed, implying that a shifted gene is a stronger nongene sequence than a pregene sequence and provides a good case study for the validation of the model.

Table 7. Gene Evaluation Data for Overlapping Genes and Nogenes (Pregenes) in Virus Genomes

species number	NCBI_ID	number of overlapping genes tested	start and stop position	TP	FP	SS	SP	CC
1	NC_001498	2	1807..3330; 1829..2389	2	0	1.0	1.0	1.0
2	NC_001507	3	1465..1067; 1601..1212; 2588..1543	3	0	1.0	1.0	1.0
3	NC_001515	6	175..411; 175..748; 797..2917; 797..810; 4076..2925; 5004..4045	6	1	1.0	0.86	0.84
4	NC_001661	5	12049..13833; 13757..15364; 15339..16061; 16155..16826; 16792..17295	5	0	1.0	1.0	1.0
5	NC_001664	16	80277..81035; 80812..82479; 99260..100588; 100545..101552; 105562..107028; 106965..107198; 108325..110667; 110636..112624; 27116..26259; 27349..26948; 53135..51723; 53916..53086; 62080..59588; 64214..62034; 128136..125989; 130043..127551	13	2	0.81	0.87	0.68
6	NC_001798	4	25100..24810; 26878..25016; 147533..146625; 147699..147244	3	0	0.75	1.0	0.77
7	NC_001819	3	619..2235; 2221..5835; 5775..7805	3	0	1.0	1.0	1.0
8	NC_002469	4	46..2535; 946..2580; 2535..4100; 4093..6381;	4	0	1.0	1.0	1.0
9	NC_003310	12	79938..80324; 80281..80739; 81358..82359; 82274..82831; 104072..104713; 104710..105456; 128269..129549; 129479..130042; 130062..131210; 131207..134701; 151299..151676; 151666..152388	12	1	1.0	0.92	0.91
10	NC_003518	2	383..877; 383..3178	2	0	1.0	1.0	1.0
11	NC_003519	3	405..1688; 1675..2025; 1889..2461	3	0	1.0	1.0	1.0
12	NC_003520	2	130..3942; 130..5466	2	0	1.0	1.0	1.0
13	NC_003680	3	2822..3421; 2822..4795; 2859..3323	3	0	1.0	1.0	1.0
14	NC_003977	4	1..1623; 155..835; 1901..2452; 2307..3215	4	0	1.0	1.0	1.0
15	NC_004002	14	57592..58593; 58508..59065; 73198..73935; 73932..74588; 74631..76991; 76988..78895; 103414..104706; 104675..105181; 9918..9649; 10390..9905; 10698..10396; 11219..10689; 111787..110876; 111980..111756	14	1	1.0	0.93	0.92
16	NC_004102	2	342..9377; 342..828	2	0	1.0	1.0	1.0
17	NC_004323	4	43277..43513; 43371..44468; 28214..27387; 28546..28208	4	0	1.0	1.0	1.0
18	NC_006883	16	123336..123869; 123866..124294; 126904..127383; 127380..128420; 144040..144372; 144369..144812; 154028..155035; 155022..156404; 167145..168116; 168113..170440; 196427..197362; 197331..198209; 198244..198975; 198972..199874; 205819..206724; 206717..270508	16	0	1.0	1.0	1.0

A comparison of the sensitivities obtained with the physicochemical model (*ChemGenome1.0*) vis-à-vis existing software such as GLIMMER (sensitivity > 97%), SOFTBERRY (sensitivity > 95%), and GENEMARK (sensitivity > 85%) clearly indicates a satisfactory performance. In addition, when the model was tested on eukaryotes (*A. thaliana* and *S. cerevisiae*), the results were equally encouraging (Table 6). The model behaves equally well in capturing tRNA genes and in separating the overlapping viral genes. The above data is indicative of the potential of the model to be developed into a stand-alone algorithm for gene finding. More immediately, the model could be adapted in conjunction with the current mathematical and statistical methods for evaluating gene predictions.

Some improvements envisioned in the model are a more obvious interpretation of the *z* parameter in terms of groove potentials and protein−DNA interactions and definition of a universal plane for all or at least a class of genomes. All sequence-dependent properties associated with the grooves of the double helix are captured in the *z* parameter—a scenario which can be improved by a systematic exploration of the diverse energy contributors to protein−DNA recognition. On the feasibility of identifying a universal plane to differentiate gene from nongene vectors, we considered fixing the plane and recovered a correlation of more than 0.90 for 260 out of 331 prokaryotic genomes. Work along these lines is in progress. Also, work is in progress to characterize the model on exons, introns, and intergenic regions in eukaryotic genomes. Further analyses are likely to lead to some new insights into the noncoding DNA. As of now, the segregation of genes and nongenes in over 350 genomes presents strong proof of the concept of the physicochemical model and points to the possibility of developing novel hypothesis-driven models for DNA sequence analyses utilizing the enormous and continually expanding genomic information.

V. CONCLUSIONS

A physicochemical model for gene evaluation is introduced and its performance appraised on 331 prokaryotic genomes, 21 eukaryotic genomes, and 18 viral genomes with highly encouraging results. The physicochemical model introduced is amenable to further systematic improvements in terms of incorporating more protein−DNA chemistry. Alternatively, information from atomic models for genome analysis[44−47] can be incorporated. The minimal database dependence, the existence of fixed parameters for separating gene and nongene regions, the goodness of the observed fit in relation to known annotation, the extendibility of the model to eukaryotes and tRNAs, and the scope for systematic improvements are some of the attractive features of the model. Furthermore, a reasonable expectation is that an investigation of the location of the gene regions on the unit sphere (Figure 2) and its relation to the underlying physicochemical principles built into the three dimensions could contribute to a new view of what can be a gene and its functional implications. Other potential applications include comparative genomics studies, a genomic signature for phylogenetic analysis, and evolving criteria for designing stable DNA sequences based on the location of the gene vectors for designed sequences.

ACKNOWLEDGMENT

Funding from the Department of Biotechnology is gratefully acknowledged. The suite of programs (*ChemGenome1.0*), written in C++, has been hosted at www.scfbio-iitd.res.in/ChemGenome for free access. S.D., P.S., P.A., R.T., and K. are equal contributors to this study.

Supporting Information Available: Table of gene evaluation data for all 331 genomes evaluated in this paper. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bordovsky, M. Y.; McIninch, J. D. GENMARK: Parallel gene recognition for both DNA strands. *Comput. Chem.* **1993**, *17*, 123−133.

(2) Lukashin, A. V.; Bordovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **1998**, *26*, 1107−1115.

(3) Bordovsky, M.; McIninch, J. D.; Koonin, E. V.; Rudd, K. E.; Medigue, C.; Danchin, A. Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res.* **1995**, *23*, 3554−3562.

(4) Burge, C.; Karlin, S. Prediction of Complete Gene Structures in Human Genomic DNA. *J. Mol. Biol.* **1997**, *268* (1), 78−94.

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

(6) Kulp, D.; Haussler, D.; Reese, M. G.; Eeckman, F. H. A generalized Hidden Markov Model for the Recognition of Human Genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1996**, *4*, 134−142.

(7) Meyer, I. M.; Durbin, R. Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics* **2002**, *18*, 1309−1318.

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

(9) Henderson, J.; Salzberg, S.; Fasman, K. H. Finding genes in DNA with a Hidden Markov Model. *J. Comput. Biol.* **1997**, *4*, 127−141.

(10) Tiwari, S.; Ramachandran, S.; Bhattacharya, A.; Bhattacharya, S.; Ramaswamy, R. Prediction of probable genes by Fourier analysis of Genomic sequences. *Bioinformatics* **1997**, *13*, 263−270.

(11) Issac, B.; Singh, H.; Kaur, H.; Raghava, G. P. Locating probable genes using Fourier transform approach. *Bioinformatics* **2002**, *18*, 196−197.

(12) Snyder, E. E.; Stormo, G. D. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res.* **1993**, *21*, 607−613.

(13) Xu, Y.; Uberbacher, E. C. Gene prediction by pattern recognition and homology search. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1996**, *4*, 241−251.

(14) Rogozin, I. B.; Milanesi, L.; Kolchanov, N. A. Gene structure prediction using information on homologous protein sequence. *Comput. Appl. Biosci.* **1996**, *12*, 161−170.

(15) Gotoh, O. Homology-based gene structure prediction: simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps. *Bioinformatics* **2000**, *16*, 190−202.

(16) Kan, Z.; Rouchka, E. C.; Gish, W. R.; States, D. J. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.* **2001**, *11*, 889−900.

(17) Rinner, O.; Morgenstern, B. AGenDA: gene prediction by comparative sequence analysis. *In Silico Biol.* **2002**, *2*, 195−205.

(18) Foissac, S.; Bardou, P.; Moisan, A.; Cros, M. J.; Schiex, T. EUGENE'HOM: A generic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Res.* **2003**, *31*, 3742−3745.

(19) Korf, I.; Flicek, P.; Duan, D.; Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **2001**, *17*, S140−S148.

(20) Solovyev, V. V.; Salamov, A. A.; Lawrence, C. B. Identification of human gene structure using linear discriminant functions and dynamic programming. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1995**, *3*, 367−375.

(21) Chen, T.; Zhang, M. Q. Pombe: a gene-finding and exon−intron structure prediction system for fission yeast. *Yeast* **1998**, *14*, 701−710.

(22) Shmatkov, A. M.; Melikyan, A. A.; Chernousko, F. L.; Bordovsky, M. Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes. *Bioinformatics* **1999**, *15*, 874−886.

(23) Yeramian, E.; Bonnefoy, S.; Langsley, G. Physics-based gene identification: proof of concept for Plasmodium falciparum. *Bioinformatics* **2002**, *18*, 190−193.

(24) Rogic, S.; Ouellette, B. F.; Mackworth, A. K. Improving gene recognition accuracy by combining predictions from two gene-finding programs. *Bioinformatics* **2002**, *18*, 1034−1045.

(25) Guo, F. B.; Ou, H. Y.; Zhang, C. T. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.* **2003**, *31*, 1780−1789.

(26) Zhang, C. T.; Zhang, R. Evaluation of gene-finding algorithms by a content-balancing accuracy index. *J. Biomol. Struct. Dyn.* **2002**, *19*, 1045−1052.

(27) Mathe, C.; Sagot, M. F.; Schiex, T.; Rouze, P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **2002**, *30*, 4103−4117.

(28) Rogic, S.; Mackworth, A. K.; Ouellette, F. B. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **2001**, *11*, 817−832.

(29) Kritee. *Towards a chemical model to predict genes & analyze prokaryotic genomes*; a dissertation submitted to Indian Institute of Technology, Delhi, India, in partial fulfilment of the requirement for the degree of five-year integrated Master of Technology in Biochemical Engineering & Biotechnology, 2001.

(30) Honig, B. H.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144−1149 and references therein.

(31) Jayaram, B. Beyond the Wobble: The rule of conjugates. *J. Mol. Evol.* **1997**, *45*, 704−705.

(32) Khurana, E. *Chemical Model for Genome Analysis*. A dissertation submitted to Indian Institute of Technology, Delhi, in partial fulfilment of the requirement for the degree of Master of Science in Chemistry, 2002.

(33) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L. *AMBER 6*; University of California: San Francisco, CA, 1999.

(34) Blattner, F. R.; Plunkett, G., III; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F. The complete genome sequence of *Escherichia coli* K-12. *Science* **1997**, *277*, 1453−1474.

(35) Reference deleted in press.

(36) Press, W. H.; Teukolsky, S. A.; Vellerling, W. T.; Flannery, B. P. *Numerical Recipes in C. The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: Cambridge, MA, 1992.

(37) Patricia, C. J. *Applied Multiple Regression*−*Correlation Analysis for the Behavioral Sciences*; Lawrence Erlbaum Associates: Mahwah, NJ, 1984.

(38) Gallant, S. I. Perceptron-Based Learning Algorithm. *IEEE Trans. Neural Networks* **1990**, *2* (1), 179−191.

(39) Rosenblatt, F. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*; Spartan Press: Washington, DC, 1961.

(40) ftp://ftp.ncbi.nih.gov/genomes/Bacteria/.

(41) David, W. *Mount Sequence and Genome Analysis*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2001; p 357.

(42) Bocs, S.; Danchin, A.; Medigue, C. Re-annotation of genome microbial CoDing-Sequences: finding new genes and inaccurately annotated genes. *BMC BioInformatics* **2002**, *3*, 5.

(43) Ussery, D. W.; Hallin, P. F. Genome Update: annotation quality in sequenced microbial genomes. *Microbiology* **2004**, *150*, 2015−2017.

(44) Kanhere, A.; Bansal, M. An assessment of three dinucleotide parameters to predict DNA curvature by quantitative comparison with experimental data. *Nucleic Acids Res.* **2003**, *31*, 2647−2658.

(45) Lafontaine, I.; Lavery, R. Optimization of Nucleic Acid Sequences. *Biophys. J.* **2000**, *79*, 680−685.

(46) Beveridge, D. L.; Dixit, S. B.; Barreiro, G.; Thayer, K. M. Molecular Dynamics Simulations of DNA Curvature, Flexibility, Dynamical Aspects of Helix Phasing and Premelting Phenomena. *Biopolymers NAS* **2005**, in press.

(47) Liu, Z.; Mao, F.; Guo bo Yan, J.-t.; Wang, P.; Qu, Y.; Xu, Y. Quantitative evaluation of protein−DNA interactions using an optimized knowledge potential. *Nucleic Acids Res.* **2005**, *33* (2), 546−548.

CI050119X