Nucleic Acids Research, 2002, Vol. 30 No. 14 3181–3191

A comparative genomic method for computational identification of prokaryotic translation initiation sites

Megon Walker[1], Vladimir Pavlovic[1] and Simon Kasif[1,2,*]

[1]Bioinformatics Program and [2]Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

Received March 5, 2002; Revised and Accepted May 20, 2002

DDBJ/EMBL/GenBank accession nos AL096836, BA000001

ABSTRACT

The ever growing number of completely sequenced prokaryotic genomes facilitates cross-species comparisons by genomic annotation algorithms. This paper introduces a new probabilistic framework for comparative genomic analysis and demonstrates its utility in the context of improving the accuracy of prokaryotic gene start site detection. Our framework employs a product hidden Markov model (PROD-HMM) with state architecture to model the species-specific trinucleotide frequency patterns in sequences immediately upstream and downstream of a translation start site and to detect the contrasting non-synonymous (amino acid changing) and synonymous (silent) substitution rates that differentiate prokaryotic coding from intergenic regions. Depending on the intricacy of the features modeled by the hidden state architecture, intergenic, regulatory, promoter and coding regions can be delimited by this method. The new system is evaluated using a preliminary set of orthologous *Pyrococcus* gene pairs, for which it demonstrates an improved accuracy of detection. Its robustness is confirmed by analysis with cross-validation of an experimentally verified set of *Escherichia coli* K-12 and *Salmonella thyphimurium* LT2 orthologs. The novel architecture has a number of attractive features that distinguish it from previous comparative models such as pair-HMMs.

INTRODUCTION

The genomic revolution that started in 1995 with the sequencing of the *Haemophilus influenza* genome has produced almost 100 genomes and thousands of genes. Since this initial sequencing, we have witnessed an almost exponential increase in the amount of genomic sequence data. In particular, numerous ongoing or finished bacterial sequencing projects have flooded microbiologists with sequence data and its initial interpretation (http://www.tigr.org and http://www.ncbi.nlm.nih.gov). These data create an opportunity to identify, catalog and mine differences and similarities between all organisms in an attempt to comparatively discover causal events or surprising modifications. Comparative genomics research aims to develop models and techniques that identify all the genes, decipher how the genes are regulated, and to distinguish the interactions that produce higher levels of function and behavior (1).

Comparative gene finding methods train on information from similarity search procedures using as queries the putative proteins derived from lists of open reading frames (ORFs) (2). Homology information recommends itself to genome annotation because a reliable way to find genes is by detection of close similarity between putative proteins and known proteins from the same or other organisms, recognition of putative gene similarity to cDNAs from the same or a closely related organism, or comparison between closely related genomes. Homology information alone does not solve the annotation problem completely because many genes (~20–40%) have no significant similarity with other known sequences, or display only partial similarity to known proteins (3). On the other hand, intrinsic methods train on DNA sequence only and determine gene locations using statistical patterns of nucleotides inside and outside coding regions along with the patterns at gene boundaries. Some intrinsic computer methods for gene finding employ a local Bayesian approach, such as Glimmer (4) and GeneMark (5).

We propose a new probabilistic method for prokaryotic genome annotation. The method employs a novel product hidden Markov model (PROD-HMM). The PROD-HMM is a composite of two hidden Markov models that simultaneously model the statistics of pairs of orthologous DNA sequences through species-specific transition and emission probabilities. Depending on the intricacy of the features modeled by the hidden state architecture, intergenic, regulatory, promoter and coding regions can be delimited by this method.

In this paper, we apply our computational method to the problem of identifying bacterial translation initiation sites. Accurate knowledge of the translation initiation site is valuable for analysis of the putative protein product of a gene and for elucidation of such signaling information in the 5′ region as ribosome binding signals, Shine–Delgarno motifs and promoters. The difficulty is caused by the absence of

*To whom correspondence should be addressed at: Bioinformatics Program, Boston University, 44 Cummington Street, Boston, MA 02215, USA. Tel: +1 617 358 1845; Fax: +1 617 353 6766; Email: kasif@bu.edu
Correspondence may also be addressed to Vladimir Pavlovic. Tel: +1 617 358 2302; Fax: +1 617 353 6766; Email: vladimir@bu.edu

relatively strong sequence patterns identifying true translation initiation sites. Unlike eukaryotic ORFs, tightly packed prokaryotic genes frequently overlap each other, obscuring translation initiation sites and confounding exact predictions of prokaryotic genes (6). Hence, the existing tools such as Glimmer (4) and GeneMark (5) that rely on simple ORF statistics exhibit a relatively low accuracy of locating the precise position of translation start sites. For instance, the quality of Glimmer's predictions varies from 50 to 90%, depending on the benchmarks (4).

Several methods have been developed for improving start site prediction in prokaryotes. Pioneering work suggested the plausibility of computational translation start site characterization in prokaryotes based on calculations of the optimal binding energy between the 16S rRNA and the region upstream of start codons (7). Prediction of bacterial start sites was the focus of Hannenhalli *et al.* (8), who developed a specialized algorithm that detects various sequence features of start sites: ribosome binding site (RBS) binding energy, distance of the RBS from the start codon, distance from the beginning of the maximal ORF to the start codon, the start codon composition, and the coding/non-coding potential around the start site. Start sites in the training and testing sequences were designated either experimentally or by similarity searches, and the discriminatory system was optimized using mixed integer programming (8). Borodovsky and co-workers incorporated intergenic region, start codon, RBS score, downstream sequence and pre-start signal into a three-periodic second order Markov model for protein-coding sequence and a second order homogeneous Markov model for non-coding sequence (2). The latest implementation of GeneMarkS utilizes a non-supervised training procedure incorporating GeneMarkHMM to find bacterial start sites. This method also employs a Gibbs sampling multiple alignment program to create a two-component statistical model of a conserved site situated in upstream sequence (6). The Yada *et al.* (9) GeneHacker Plus HMM uses duration and dicodon statistics to model the coding region and upstream translation control signals locally. Finally, the RBSfinder developed by Salzberg and co-workers to post-process Glimmer and GeneMark annotation inputs an entire genomic sequence and first-pass annotation to train a probabilistic model that scores candidate ribosome sites surrounding previously annotated start codons. If a better RBS is found either upstream or downstream of the originally predicted start site, then the system moves the translation initiation site accordingly (10).

In this paper, we attempt to quantify the power of purely comparative statistics to detect important genomic features. Our method for identification of prokaryotic translation initiation sites employs a PROD-HMM. Unlike other methods, it relies on modeling the difference in nucleotide substitution rates in coding and non-coding regions. Coding region is characterized by non-synonymous substitutions and synonymous substitutions whose ratio is one feature of that region. Stronger selective constraints for synonymous changes than for non-synonymous substitutions suppress function-compromising mutations in the protein coding regions, resulting in non-synonymous/synonymous substitution ratios significantly <1 in the orthologous coding regions (11). On the other hand, the substitutions in the non-coding regions are almost random. The PROD-HMM effectively estimates the synonymous/non-synonymous characteristics of a pair of DNA sequences and uses them to discriminate the coding from intergenic regions. The algorithm in its current implementation relies on accurate knowledge of transcription termination sites to enhance its performance. Hence, it can most efficiently facilitate post-processing of putative ORFs designated by intrinsic bacterial genome annotation software packages such as Glimmer.

We show that the accuracy of gene start prediction can be improved by computational analysis using the PROD-HMM tool. Two related prokaryotic genomes were analyzed using the new tool. Parameters of the HMM (transition and emission probabilities) were estimated by using annotated DNA training sequences. The accuracy of detecting the translation initiation codons of test set orthologs was assessed by comparison with GenBank and EcoGene genomic sequence annotations.

MATERIALS AND METHODS

The data sets analyzed in this study are available at http://genomics10/megonw/PROD-HMM/.

Species studied

The success of our approach depends on selection of genomes that are not too similar (intergenic sequence similarity as high as that of coding regions) or too far diverged (coding sequence similarity as low as that of intergenic regions) to inform the comparative analysis. Each pair of strains considered is amenable to our approach because they are neither too closely related to offer additional comparative genomic information nor too distant.

Of the three fully sequenced *Pyrococcus* genomes (including *Pyrococcus furiosus* at 1.908 Mb), *Pyrococcus abyssi* (1.765 Mb) (GenBank accession no. AL096836) and *Pyrococcus horikoshii* (1.738 Mb) (12) were selected for this analysis. Divergence of the *P.furiosus* ancestor from the *P.abyssi* and *P.horikoshii* common ancestor prior to speciation of *P.abyssi* and *P.horikoshii* is evidenced by intergenomic disruptions to synteny involving rearrangement, translocation, transposition, insertion/deletion, recombination and inversion events. In addition, there are longer preserved chromosomal segments between *P.abyssi* and *P.horikoshii* than between *P.furiosus* and either of the other two species. Several major chromosomal features common to *P.abyssi* and *P.horikoshii* are different in *P.furiosus*, and average amino acid identities between close homologs of *P.abyssi* and *P.horikoshii* exceed those between either and *P.furiosus* (13).

The *Pyrococcus* genomes are delineated into four regions by conservation pattern such that no DNA fragment exchange between the regions has been observed. Best conserved are region I (replication origin, inverted between *P.abyssi* and *P.horikoshii*) and region IV (ribosomal operon). Region II displays better gene order and content conservation between *P.abyssi* and *P.horikoshii* than region III with its many translocation and insertion/deletion events (13). Pairwise comparison of the two genomes under study reveals high nucleotide conservation (1122 kb in common). However, the number of predicted ORFs is quite different, despite the

comparable genome sizes of *P.abyssi* (1765 ORFs) and *P.horikoshii* (2061 ORFs) (13).

Most prokaryotic translation initiation sites have been predicted computationally, not experimentally. Although the start sites of the sequences in this database are not always verified experimentally, GenBank annotations of complete prokaryotic genomes are frequently used to evaluate the algorithm performance of gene finders because the public database annotation represents expert opinion summarizing various types of evidence. The PROD-HMM's ability to give biologically correct annotation rather than to merely recognize GenBank annotation was evaluated using 801 ortholog pairs from enterobacteria *Escherichia coli* K-12 and *Salmonella thyphimurium* LT2 (4.857 Mb). *Escherichia coli* K-12 contains the largest number of genes with validated starts among all prokaryotes, with 811 proteins delimited by N-terminal sequencing in the EcoGene dataset (10,14). Three *E.coli* strains have been fully sequenced and annotated, including *E.coli* K-12 (5.498 Mb), *E.coli* 0157:H7 (4.639 Mb) and *E.coli* 0157:H7 EDL933 (5.528 Mb). However, these three strains are too closely related for simultaneous pairwise annotation by our method, as the nucleotide substitutions are not yet extensive enough to inform the PROD-HMM.

Data collection and input preparation

Data files were obtained from GenBank records for *P.abyssi* (GenBank accession no. AL096836), *P.horikoshii* (GenBank accession no. BA000001), *E.coli* K-12 (GenBank accession no. U00096) and *S.thyphimurium* LT2 (GenBank accession no. 003197). We designed and implemented a semi-automated system that contains core modules written in PERL. It performs the following steps: (i) BLASTP ortholog determination; (ii) extraction of genomic coordinates, coding sequences and up to 200 nt of upstream intergenic sequence; (iii) alignment of orthologous nucleotide sequences using the global alignment module of MUMmer (15); (iv) determination of the non-synonymous/synonymous substitution ratio for each aligned pair of orthologs; and (v) determination of percent identity in the 60 bp surrounding the start sites of each aligned pair of orthologs.

During sequence data processing in step (iii), nucleotide residues corresponding to each of the amino acid sequences and extending from, at most, 200 nt upstream of the GenBank annotated start site to the stop codon were extracted from the genomic DNA sequence and globally aligned to the orthologous sequence from the related genome, including internal gaps. Inclusion of upstream intergenic sequence in the data set trains the probabilistic model to differentiate the grammatical structure of genes from upstream intergenic regions, thereby enabling it to delimit coding and intergenic regions during start site designation.

A total of 1443 *Pyrococcus* protein orthologs were identified (BLASTP two-way best matches), similar to the number of orthologs obtained by Lecompte *et al.* (13). The final input to the model was composed of 183 pairs of aligned, orthologous, Watson strand nucleotide sequences consisting of protein coding regions and adjoining upstream intergenic sequences. Of the pairs, 136 included intergenic sequence upstream of each start codon that did not overlap or lie directly adjacent to the preceding coding region. The above criteria for filtering the initial ortholog set effectively reduced the number of operons lacking intergenic upstream sequence and resulted in a conservative input set for self-evaluation by the new algorithm.

801 *E.coli* K-12 genes that had been derived from the 811 verified sequences in the EcoGene data set and for which *S.thyphimurium* LT2 orthologs were delineated by BLAST searches qualified for analysis by our model. The *S.thyphimurium* LT2 protein ortholog for each experimentally validated *E.coli* K-12 gene was designated by one-way BLASTP similarity searches of each *E.coli* K-12 protein sequence against a database of all known *S.thyphimurium* LT2 protein sequences. All 801 pairs of aligned, orthologous nucleotide sequences consisting of protein coding regions and adjoining upstream intergenic sequences were annotated by the PROD-HMM. Neither overlapping nor Crick strand sequences were excluded, thereby providing a comprehensive input set for simultaneous annotation with 10-fold cross-validation by the PROD-HMM. Crick strand sequences input to the model were adjusted by formulating the complementary sequence in the opposite direction such that only 5′ to 3′ analysis is necessary overall. By applying our method with cross-validation to this experimentally validated set, its robustness is confirmed and the *Pyrococcus* results are corroborated.

Computational model

We simultaneously model the statistics of two related DNA sequences using a PROD-HMM. A PROD-HMM is a composite of two (in general two or more) HMMs. In this discussion we will assume, without loss of generality, that both HMMs have the same N_{hmm} states $\{s_1,\ldots,s_{N_{hmm}}\}$ and can emit the same M_{hmm} symbols $\{n_1,\ldots,n_{M_{hmm}}\}$. The PROD-HMM will then be a model with $N = N_{hmm}^2$ product states $\{sp_1,\ldots,sp_N\}$ describing every possible combination of states of the two HMMs. For instance, product state sp_1 is a pair $sp_1 = (s_1,s_1)$. Each product state will emit a pair of $M = M_{hmm}^2$ symbols, e.g., $np_1 = (n_1,n_1)$.

Like an ordinary HMM, a PROD-HMM is completely specified by a transition probability matrix T and an emission matrix E. However, in a PROD-HMM the transition matrix describes the probability of arriving in one product state (a composite of two simple states) from another, as shown in equation **1**:

$$\begin{aligned} T(sp_i \mid sp_j) &= \Pr(\text{product state } i \text{ at position } k \mid \text{product state } j \\ &\quad \text{at position } k-1) \\ &= \Pr[(s_{i1},s_{i2}) \text{ at position } k \mid (s_{j1},s_{j2}) \\ &\quad \text{at position } k-1] \end{aligned} \quad \mathbf{1}$$

Similarly, the emission matrix describes the probability of simultaneously seeing a pair of emission symbols in one product state, as displayed in equation **2**:

$$\begin{aligned} E(np_i \mid sp_j) &= \Pr(\text{emit pair of symbols } np_i \\ &\quad \text{at position } k \mid \text{product state } j \text{ at position } k) \\ &= \Pr[(n_{i1},n_{i2}) \text{ at position } k \mid (s_{j1},s_{j2}) \\ &\quad \text{at position } k] \end{aligned} \quad \mathbf{2}$$

In our case, we consider the PROD-HMM to be a composite of two separate HMMs with transitions following the logic of prokaryotic gene organization: intergenic non-coding region

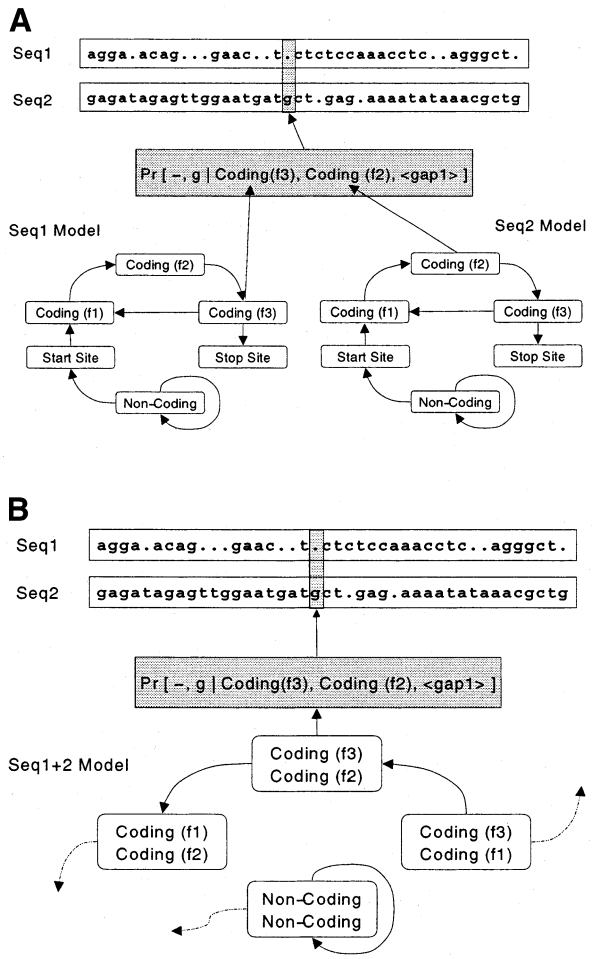

Figure 1. Integration of comparative information for start site prediction in prokaryotic genomes using a PROD-HMM. Composite of two 10-state HMMs (**A**), a 10×10 product HMM (**B**) models joint distribution of nucleotides in two aligned genomic sequences. When aligned, the bases of two nucleotides can match or not. Each aligned base pair can be labeled with one of 100 annotations. Otherwise, a nucleotide is matched with a gap in the other sequence. An additional 10 annotation states are required in the case of a sequence 1 nucleotide aligned with a gap in sequence 2 (pairwise combinations of the 10 state labels with the gap state). Ten final annotations denote any of the 10 nucleotide features in sequence 2 aligned with a gap in sequence 1 (see Fig. 2). The pair (gap, g) is emitted by the [gap, Coding (f3)] state in frame 3 of the first sequence. Given a pair of aligned genomic sequences, the PROD-HMM will detect start sites in both sequences as transitions from non-coding to coding states.

(state NC), any of the three start site positions (S1, S2, S3), any position of an ORF triplet codon (C1, C2, C3), and any position of the stop site (E1, E2, E3). In general, our model does not cover the entire genome but only fits to one gene. There are $N = 100$ states in this PROD-HMM (all possible paired combinations of the 10 labels), each of them emitting $M = 16$ combinations of base pairs. This is depicted in Figure 1.

The model assumes that the two orthologous sequences have been previously aligned. If the alignment allows gaps, the PROD-HMM needs to be modified in the following way. When a base pair in one sequence is aligned with a gap in the other, we need to maintain the state of the gapped sequence, as illustrated in Figure 2. To do so, we define two new transition tables, T_{g1} and T_{g2}. T_{g1} models the transitions when a gap is encountered in ortholog 1. It satisfies the condition that

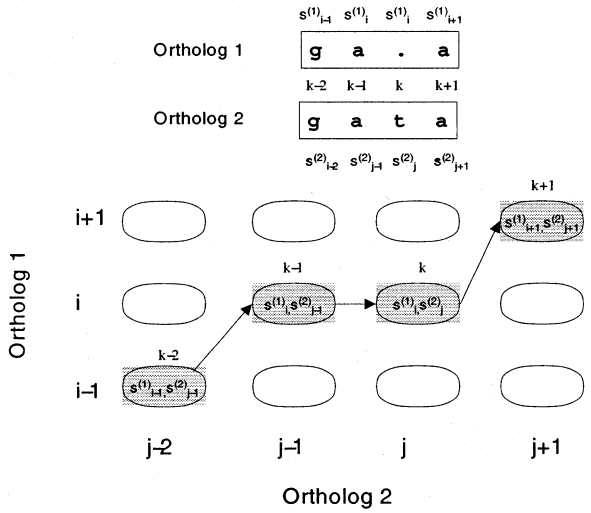

Figure 2. PROD-HMM needs to be modified when gaps are allowed in alignment of two orthologous sequences. If the two orthologs (1 and 2) are aligned as shown by the path $k-2$, $k-1$, k, $k+1$ in the trellis of all possible alignments, then the state of the ortholog 1 at alignment position $k-1 (s_i^{(1)})$ has to remain the same in position k. In the PROD-HMM, this can be imposed by an appropriate state transition matrix T followed by a modified emission matrix E (see text for details).

$T_{g1}[(i,j) \mid (p,q)] = 0$, when $i \neq p$.

Similarly, the transition table into a gapped ortholog 2 must have zeros for all $j \neq q$. This ensures that the state of the last non-gap base pair is conserved when gaps occur. Finally, for each of the two cases, we need to define two new emission tables, E_{g1} and E_{g2}, as follows:

$E_{g1}[i \mid (p,q)] =$
Pr [emit symbol i in ortholog 1 | product state (p,q)]
$E_{g2}[j \mid (p,q)] =$
Pr [emit symbol j in ortholog 2 | product state (p,q)].

Note that in the gapped alignment case the PROD-HMM essentially becomes inhomogeneous because its parameters T and E vary with the type of alignment at position k between the two sequences. If the alignment information at position k is denoted by $a(k)$, then the inhomogeneous PROD-HMM parameters are

$$T[(i,j) \text{ at } k \mid (p,q) \text{ at } k-1] = \begin{cases} T_{aligned}[(i,j) \mid (p,q)] & , \quad a(k) = \text{aligned} \\ T_{g1}[j \mid (p,q)] & , \quad a(k) = \text{gap in 1} \\ T_{g2}[i \mid (p,q)] & , \quad a(k) = \text{gap in 2} \end{cases}$$

and

$$E[(i,j) \text{ at } k \mid (p,q) \text{ at } k] = \begin{cases} E_{aligned}[(i,j) \mid (p,q)] & , \quad a(k) = \text{aligned} \\ E_{g1}[j \mid (p,q)] & , \quad a(k) = \text{gap in 1} \\ E_{g2}[i \mid (p,q)] & , \quad a(k) = \text{gap in 2} \end{cases}$$

A PROD-HMM can be represented as an inhomogeneous HMM with (possibly large) product state and emission spaces.

Hence, parameter and state estimation algorithms (Baum–Welch, forward–backward estimation and Viterbi decoding) of ordinary HMMs can be directly applied to PROD-HMMs. In our case of two orthologous DNA sequences with $N = 100$ product states and $M = 16$ emissions, direct application of ordinary HMM algorithms to PROD-HMM is computationally feasible.

Training

During training, we consider pairs of aligned homologous nucleotide sequences terminated at stop codons and their GenBank annotation (using 100 pairwise combinations of 10 labels). The emission probability parameters of the PROD-HMM are estimated simultaneously for the related pair of genomes using maximum likelihood estimation. Counts of how many times an aligned base pair of nucleotides in the input training sequences coincide with each of the 100 pairs of states are tabulated in 100 4×4 emission matrices. Transition probability parameters are estimated in a similar fashion: counts of how many times each possible aligned pair of states in the input training sequences (position $k - 1$) are directly followed downstream by any of the 100 pairs of aligned states (at position k) are tabulated in a 100×100 transition matrix. Emissions from gapped pairs (E_{g1} and E_{g2}) and transition-into-gap parameters (T_{g1} and T_{g2}) are separately estimated whenever there is a gap in one of the orthologs.

Thus, the grammatical structure and trinucleotide frequency patterns of orthologs identified during the homology search were directly embedded in the product HMM in order to improve its annotation. This captures species-specific oligonucleotide frequency patterns as well as synonymous and non-synonymous substitution patterns. The latter is illustrated in Figure 3.

Testing

During testing, input to the model consists of the aligned pairs of orthologous upstream and coding nucleotides. Each input alignment terminates at the last nucleotides of the stop codon. The initial nucleotides of the input are assumed to comprise intergenic upstream region, and the model determines the translation initiation start site using the PROD-HMM with architecture and probability parameters described above according to the dynamic programming Viterbi algorithm. A single occurrence of (S1, S2, S3) labels in each of the two decoded sequences designates the corresponding start codon nucleotides.

Comparison with other models

Our PROD-HMM approach is related to pioneering research based on the generalized pair-HMM, or GPHMM (16,17). Both models deal with the problem of joint genomic annotation of two related DNA sequences.

The PROD-HMM effectively models the possibility of simultaneously different annotations of two aligned base pairs (e.g., at position k one sequence is in state $C1$ while the other is in $C3$). For example, because of mutations, frame shifts or recombinations, a coding multi-domain gene might have lost or acquired a domain. In this case the model must allow for the possibility of aligning part of a coding region in one genome to non-coding in another. This feature is particularly important for the application of PROD-HMMs for eukaryotic annotation

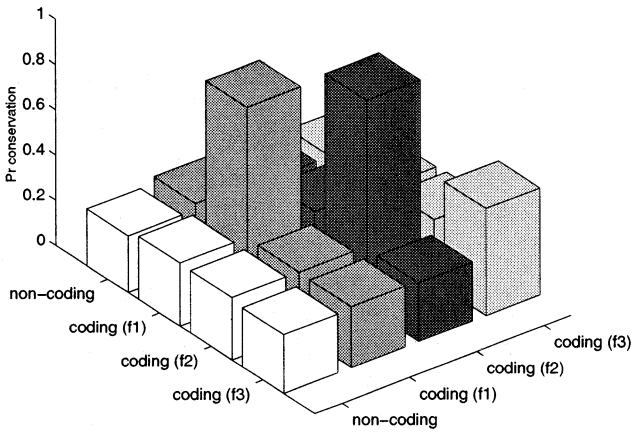

Figure 3. Nucleotide conservation captured by PROD-HMM emission matrix E. The emission statistics of a PROD-HMM exploit a more subtle conservation property of coding regions, described previously using the synonymous/non-synonymous paradigm. Rare nucleotide substitutions (high conservation) of in-frame states [(f1,f1) and (f2,f2)] can differ significantly from more frequent substitutions in the third codon position (f3,f3) as well as almost random out-of-frame [e.g. (f1,f2)] and non-coding state substitutions. Statistics were obtained from 183 *P.horikoshii* and *P.abyssi* pairs.

where exons or splice sites might become mutated and therefore a coding region in one sequence could align to an intron or intergenic region in another (L.Zhang, V.Pavlovic and S.Kasif, manuscript in preparation). Although the original generalized pair-HMM formulation did not explicitly model such possibilities, extension of GPHMMs to this setting is possible.

A GPHMM might be of relatively high computational complexity even for moderately long pairs of sequences. In effect, the running time of a GPHMM is quadratic in the length of the sequence or more precisely $O(NL^2)$ where N is the number of states and L is the length of the sequence. In this sense it is similar in running time to global alignment procedures. The PROD-HMM runs in time N^2L assuming the alignment has been carried out during pre-processing. The PROD-HMM architecture allows us to separate the alignment phase from the gene identification phase. Since alignment of genomes has received substantial attention, a number of feasible algorithms have been developed that run in almost linear time (e.g. MUMmer, GLASS and others). Although this paper does not focus on this aspect of the procedure, efficiency is an important feature of eukaryotic genome annotation.

Our approach is also related to the seminal work on eukaryotic annotation using extrinsic homology information as found in GenomeScan (18). There, the system also treats alignment information as additional evidence that is 'fed' to the decoding architecture of the HMM. See Pavlovic *et al.* (19) for a general methodology of incorporating multiple evidence into HMM architectures.

RESULTS AND DISCUSSION

Algorithm accuracy

The product HMM-based algorithm was evaluated following analysis of 183 *P.abyssi–P.horikoshii* orthologs (self-evaluation) and annotation of 801 *E.coli* K-12–

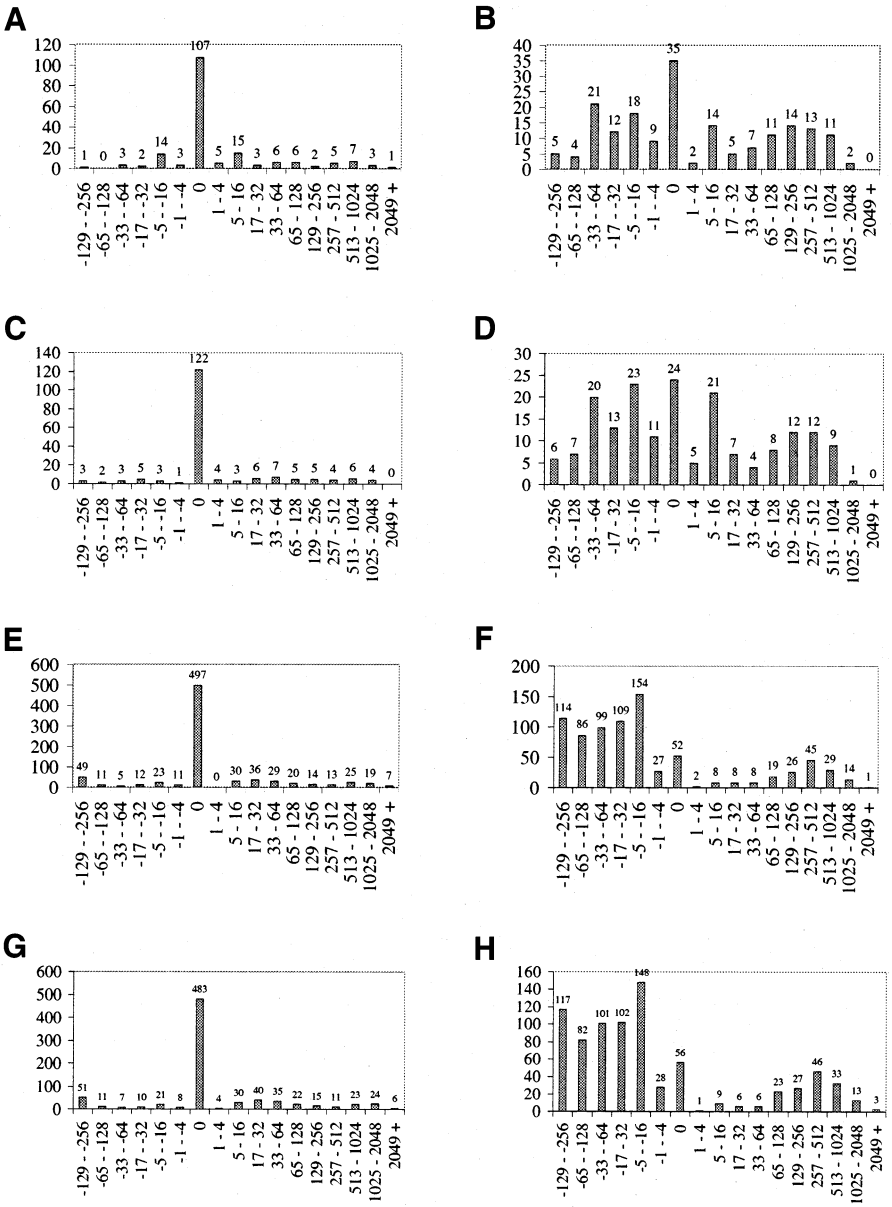

Figure 4. Prediction of start sites in 183 *P.abyssi* (**A** and **B**), *P.horikoshii* (**C** and **D**), *E.coli* K-12 (**E** and **F**) and *S.thyphimurium* LT2 (**G** and **H**) ortholog pairs using comparative methods. Panels show offset statistics of translation initiation sites predicted by the product hidden Markov model-based algorithm (A, C, E and G) and TBLASTX (B, D, F and G) with respect to GenBank and EcoGene annotation. All bins are log base 2. Upstream predictions offset in the intergenic region are displayed in negative bins; prediction offsets distributed downstream are in positive bins. Slightly higher PROD-HMM prediction accuracy was observed for *P.horikoshii* (122 exactly predicted orthologs) than for *P.abyssi* (107 precise predictions). Similarly, 497 *E.coli* K-12 orthologs and 483 *S.thyphimurium* LT2 orthologs were exactly predicted by the PROD-HMM. The robust performance of our algorithm on the experimentally validated set despite 10-fold cross-validation confirms the significance of the *Pyrococcus* results. The cumulative species-specific trinucleotide and synonymous/non-synonymous substitution patterns that informed the PROD-HMM's predictions after training proved to be a more accurate basis of prediction than the pairwise protein-specific sequence similarity information enabling the BLAST alignments.

S.thyphimurium LT2 orthologs (10-fold cross-validation). The accuracy of the algorithm was estimated by the amount of offset between translation initiation codons predicted by the PROD-HMM and the GenBank annotation (*Pyrococcus* and *Salmonella* orthologs), or by the amount of offset between the PROD-HMM start site genomic coordinates and the experimentally verified EcoGene annotation (*E.coli* K-12 orthologs).

As the maximum length of upstream non-coding sequence submitted with each coding ortholog was 200 nt, upstream offsets never exceeded 200, while downstream predictions ran farther afield. The largest offsets resulted from global alignment inputs with the following attributes: (i) excessive, erratic gaps dispersed throughout the input alignment of the orthologous upstream intergenic and coding sequences; (ii) extremely short upstream sequences in either or both orthologs; and (iii) poor homology as indicated by BLASTP bitscores <100.

Figure 4A and C display the offset distributions of *P.abyssi* and *P.horikoshii* translation initiation site predictions made by

the PROD-HMM-based algorithm. Of the 183 *P.abyssi* and 183 *P.horikoshii* start sites, 58 and 67%, respectively, are exactly predicted with zero offset, comparable with *E.coli* K-12 (62%, Fig. 4E) and *S.thyphimurium* LT2 (60%, Fig. 4G). The robust performance of our algorithm on the experimentally validated set despite 10-fold cross-validation confirms the significance of the *Pyrococcus* results.

Prediction errors in the *Pyrococcus* annotation occurred predominantly downstream of the true start (97 predictions), with the remaining 42 of the inexact predictions placing the start site in the upstream intergenic region. The 229 exact predictions constituted 145 pairs of orthologs: 84 pairs in which both ortholog start sites were predicted exactly and 61 pairs in which one of the two predictions was offset.

Below are details regarding the functional annotation of the best-predicted *P.abyssi* orthologs. GenBank functional annotations revealed that 72 of these 145 pairs encoded two proteins of undetermined function (labeled hypothetical proteins by GenBank with function unknown by COG), while the remaining 73 ortholog pairs contained at least one characterized member. When indicated for both orthologs, COG functional annotation between the members of these 73 pairs generally corresponded (three exceptions).

The predominant occurrence of correctly predicted orthologs in regions I and II is a result of high sequence conservation in the regions containing the replication origin, despite the substantial post-speciation inversion of region I. The largest correctly annotated functional group, composed of 13 ortholog pairs COG-classified by general function, contained two NADH-dependent dehydrogenase related proteins, an NADH oxidase, methionyl and alanyl tRNA synthetase fragments, among others. Of the 13 proteins, 11 exhibited coordinates within regions I and II of both genomes. Seven proteins involved in DNA replication, recombination and repair were correctly predicted at corresponding genomic coordinates of regions I and II, as were five inorganic ion transport and metabolism proteins, six proteins involved in coenzyme metabolism and six proteins for carbohydrate transport and metabolism.

Occurring broadly throughout all four genomic regions and exactly predicted for at least one ortholog by the new algorithm were five amino acid transport and metabolism proteins, four nucleotide transport and metabolism proteins and five energy production and conversion proteins.

PROD-HMM versus TBLASTX

The benchmark TBLASTX program of the BLAST suite of local alignment tools was utilized for translation start site detection by modeling the task purely as a sequence alignment problem. To predict the gene structures in two genomic sequences with orthologous genes, it is simplest to look for coding and regulatory regions by comparing the corresponding orthologous nucleotide and protein sequences for conserved regions. For each ortholog, at most 200 upstream intergenic nt and the entire coding region were submitted. TBLASTX comparisons involved two sets of translations in six reading frames. Reversing the query and database orthologs separately predicted the start site for each member of an ortholog pair.

The PROD-HMM predictions are informed by cumulative species-specific trinucleotide patterns and substitution ratios learned from input orthologs and characterized in the emission and transition matrices. An exact prediction by the algorithm occurred when the S1, S2 and S3 labels emitted by the model exactly matched the sequence positions of the initial start codon. BLAST alignments are based on the pairwise protein-specific sequence similarity observed between only the two sequences under analysis during any given prediction. For each TBLASTX output, the initial aligned codon of the hsp with the lowest expectation value and maximal bitscore was evaluated as the start site prediction for the corresponding ortholog. A zero offset prediction was made when hsp extension stopped along the diagonal in the upstream direction at exactly the initial methionine in the protein alignment of two translated sequences.

Figure 4 displays the offset distributions of translation initiation sites predicted by TBLASTX (Fig. 4B and D) for orthologs of both *Pyrococcus* strains. For the *P.abyssi* subset, the accuracy of the new algorithm (58% start sites exactly predicted with zero offset) exceeds that of TBLASTX (19%). For the *P.horikoshii* subset, the PROD-HMM performs better overall (67%), surpassing the number of start sites exactly predicted by TBLASTX (13%). Higher model prediction accuracy was observed for *P.horikoshii* (122 exactly predicted orthologs) than for *P.abyssi* (107 precise predictions). The opposite trend is observed for exact predictions by TBLASTX, with higher accuracy for *P.abyssi* orthologs. Does each annotation system perform best on the same set of orthologs? The zero offset predictions by the model and TBLASTX coincided in only 26 cases for *P.abyssi*. Of the *P.horikoshii* orthologs exactly predicted by the model, 21 are also exactly annotated by TBLASTX.

Figure 4F and H display the offset distributions of translation initiation sites predicted by TBLASTX for *E.coli* K-12 and *S.thyphimurium* LT2 orthologs. For *E.coli* K-12, the PROD-HMM annotation accuracy (62% start sites predicted with zero offset) exceeds that of TBLASTX (6%). Performance was similar for the *S.thyphimurium*: 60% zero offset predictions by the model, but only 7% by TBLASTX.

Further prediction analysis was conducted on members of 60 paired *Pyrococcus* orthologs for which the algorithm predicts both start sites exactly but for which both TBLASTX predictions are offset. The PROD-HMM's accuracy of start site prediction exceeded that of conservation detection system TBLASTX after analysis of the same subset of ortholog pairs. Notwithstanding the similar sequence input to each program (for each ortholog, at most 200 upstream intergenic nt and the entire coding region were submitted), TBLASTX comparisons involved many more reading frames between which to form potential alignments, resulting in more inaccurate annotations. Sufficiently similar lists of hsps (corresponding percent identity, query coverage, expectation values and bitscores) resulted from each pair of TBLASTX comparisons conducted that only one output is reproduced for each case discussed in Figure 5.

When the nucleotides of each coding region input to TBLASTX to form the query and the database search space have insufficient sequence similarity in the 5′-coding region, the aligned words (hits) in the upstream and initial coding regions are far spaced or on different diagonals such that TBLASTX hsp extension is restricted to the coding region. The resultant prediction is offset downstream of the true start

```
Query: pabyssi.ffn 227068 228611 (1544 letters)
Database: pyro.247.ffn 212758 214307 (1550 letters)
Score =  763 bits (1659), Expect(3) = 0.0
Identities = 311/379 (82%), Positives = 342/379 (90%)
Frame = +3 / +3

Query: 405  LASNIAIGHVRYSTSGSLSEVQPLEVRCCGYELAIAHNGTLTNFIPLRRLYEGMGIKFHS 584
            L  N  IGHVRYSTSGSLSEVQPLEV CCGY+++IAHNGTLTNF+PLRR YE  G KF S      . . .
Sbjct: 405  LNGNPVIGHVRYSTSGSLSEVQPLEVECCGYKVSIAHNGTLTNFLPLRRFYESRGFKFRS 584

       (a) TBLASTX annotation of phosphoribosyl amidotransferase offset downstream


Query: pabyssi.ffn 165156 166537 (1382 letters)
Database: pyro.ffn 158949 160289 (1341 letters)
Score =  693 bits (1508), Expect(2) = 0.0
Identities = 291/341 (85%), Positives = 312/341 (91%)
Frame = +3 / +1

Query: 45   *MLPLRWLRWSVSYPCDRGFIFLEVHRGEVHILYDLH*SLSRGGTKLRGGVPMKYDVVVV 224
            *+LPL WL W +S   D  F+ L VHRG VHI++DLH  +SR G KL GGV M+YDVVVV      . . .
Sbjct: 4    *VLPLWWLCWCLSNVGDSRFLVLGVHRG*VHIVHDLHKGMSRWGLKL*GGVSMRYDVVVV 183

         (b) TBLASTX annotation of geranylgeranyl hydrogenase offset upstream
```

Figure 5. Relatively poor performance of TBLASTX for 60 ortholog pairs annotated correctly by the PROD-HMM can be explained by the frequent presence of high scoring HSPs upstream or downstream of the true start site. (**a**) Partial TBLASTX output annotating a glutamine-dependent phosphoribosyl amidotransferase for which TBLASTX predictions were offset far downstream of the true start codon, but the PROD-HMM-based algorithm predictions were exact for both orthologs. (**b**) The partial TBLASTX output annotating a geranylgeranyl hydrogenase with predicted start site offset significantly upstream of the true start codon, while the PROD-HMM-based algorithm predictions were exact for both orthologs. The true start site that was correctly identified by the PROD-HMM (M-M encoded by ATG-ATG) is 52 codons downstream of the putative M-V (ATG-ATG) start site indicated by TBLASTX.

codon. Figure 5(a) displays the partial TBLASTX output for such an instance involving the glutamine-dependent phosphoribosyl amidotransferase (purF) transcribed in *P.abyssi* region II (gi5457654, 227268–228611) and in the corresponding region I of *P.horikoshii* (gi3256629, 212958–214307). PurF is involved in nucleotide transport and metabolism in the purine biosynthesis pathway, as it is the first step in *de novo* purine synthesis and part of the classically defined route for thiamine synthesis. The PROD-HMM-based algorithm predictions were exact for both orthologs. TBLASTX detects the sequence similarity in the hsp of highest bitscore and aligns the forward third reading frame of the protein translations from each ortholog (+3/+3), but it initiates the alignment 68 codons downstream of the initial methionine (204 nt downstream of the start codons). The initial 68 amino acids excluded from the best scoring hsp display high sequence identity but are offset by one reading frame until the 69th residue due to a missing amino acid in the *P.horikoshii* sequence. The alignment covers all remaining amino acids of the protein to the terminal codon, as the sequence identity between the coding regions of the orthologs is strong. The hsp with the sixth highest bitscore (120) aligns the forward third reading frame of the protein translations from each ortholog (+3/+3) beginning at an aligned methionine pair only four residues downstream of the true initial methionine in the *P.abyssi* sequence (three residues downstream in *P.horikoshii*). Its expectation value is 0.0, ruling out such sequence similarity by chance. However, the short coverage (60 residues) is contained entirely within the coding region upstream of the 69th residue.

On the other hand, when the intergenic nucleotides of each ortholog input to TBLASTX to form the query and the database search space have high sequence similarity, hits in the upstream intergenic region within the window length and on the same diagonal enable TBLASTX hsp extension upstream into the intergenic region, and the prediction is offset upstream of the true start codon. Figure 5(b) exemplifies those cases in which TBLASTX predictions were offset significantly upstream of the true start codon, but the PROD-HMM-based algorithm predictions were exact for both orthologs. Putative geranylgeranyl hydrogenase (energy production and conversion) orthologs are transcribed in *P.abyssi* region I (gi5457597, 165356–166537) and at the border between the corresponding regions I and II of *P.horikoshii* (gi3256567, 159108–160289). The hsp of highest bitscore following both TBLASTX comparisons aligns the third forward reading frame of the *P.abyssi* ortholog translation to the first forward reading frame of the *P.horikoshii* ortholog translation (+3/+1), with predicted start site 156 nt upstream of the true start codons (52 codons upstream of the initial methionines). The alignment upstream of the hydrogenase protein displays ~50% amino acid identity in the highest scoring hsp when the *P.horikoshii* query is compared with the *P.abyssi* database, while an eight residue long low complexity region is masked in the upstream alignment of the highest scoring hsp when the *P.abyssi* query is compared with the *P.horikoshii* database. Both alignments begin at an intergenic 'codon' (methionine encoded by a *P.abyssi* ATG and valine encoded by a *P.horikoshii* GTG) that was mistaken for the true start 52 codons downstream (methionine encoded by a *P.abyssi* ATG and methionine encoded by a *P.horikoshii* ATG).

Ortholog evaluation criteria

Three features descriptive of each *Pyrococcus* ortholog pair were designated for further analysis: BLASTP bitscore, percent identity surrounding the start codon and non-synonymous/synonymous substitution rates in the coding regions. These criteria were correlated with the binned prediction

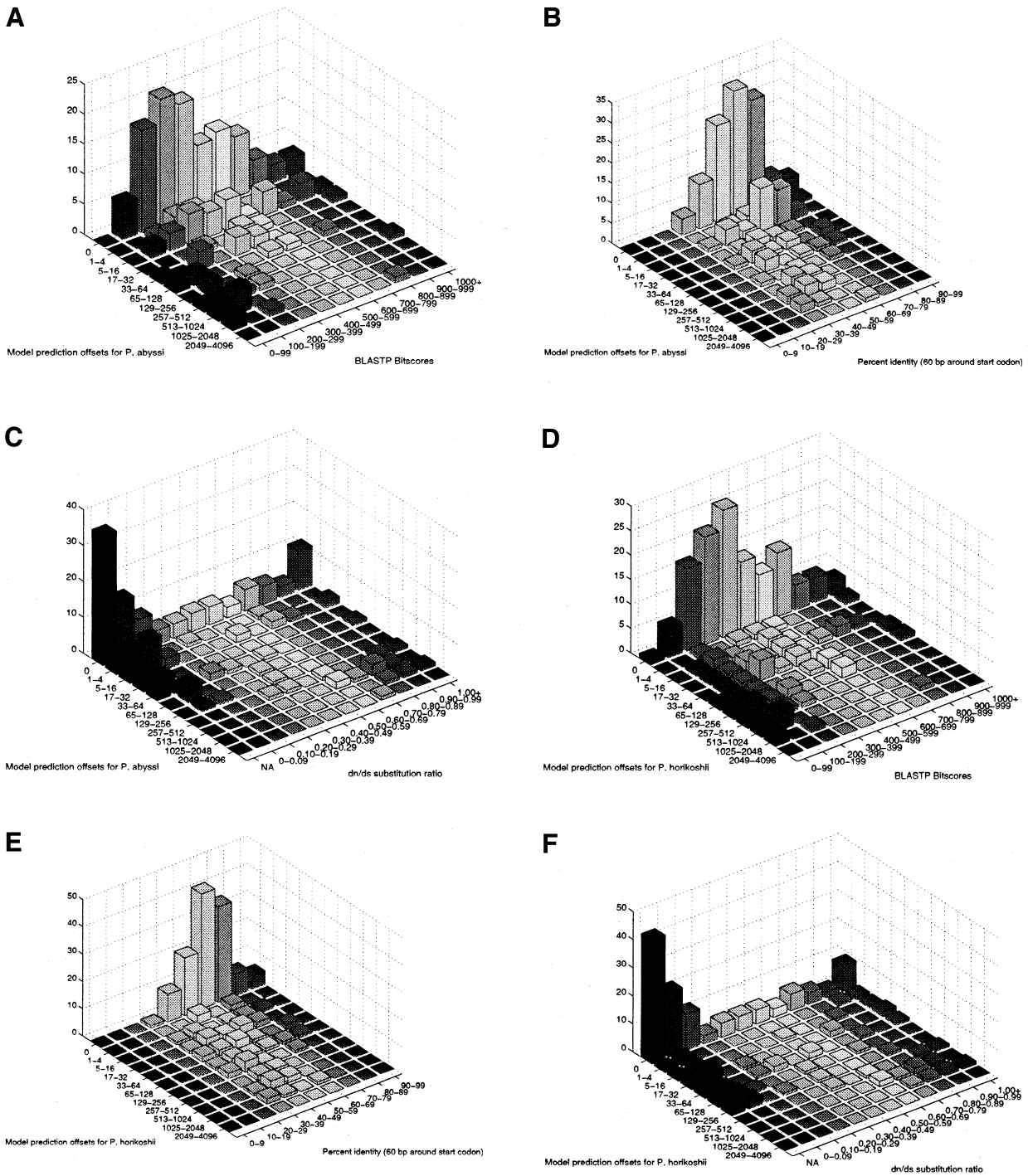

Figure 6. Three criteria were correlated with the binned offsets from *P.abyssi* (**A**–**C**) and *P.horikoshii* (**D**–**F**) in order to inform future ortholog selection and PROD-HMM prediction qualification: BLASTP bitscore (A and D), percent identity surrounding the start codon (B and E), and non-synonymous/synonymous substitution rates in the coding regions (C and F). The resultant optimal ranges for PROD-HMM annotation will constitute the biological rationale by which orthologs are selected for model training (emission and transition matrix parameterization) in future analyses: bitscore 100–800, percent identity 40–80% and dn/ds substitution ratio NA (mutational saturation) or <0.30.

offsets. The resultant thresholds will constitute the biological rationale by which orthologs are selected for model training (emission and transition matrix parameterization) and by which PROD-HMM predictions are qualified after testing in future analyses.

The results of the initial BLASTP search were categorized by bitscore. As indicated in Figure 6A and D, the counts of exactly predicted orthologs for each species peak in the 300–500 bitscore range, with declining counts of exactly predicted start codons at both extreme ends of the bitscore

range. This and the slight negative correlation coefficients for *P.abyssi* (–0.1242) and *P.horikoshii* (–0.2325) indicate degradation of model prediction performance not only when bitscore decreases and sequence similarity weakens, but also as bitscore increases and sequence similarity between the orthologous intergenic and coding regions is optimized. For example, of the 183 ortholog pairs in the *Pyrococcus* set, eight pairs scored <100. Only one of the 16 translation initiation site predictions generated by the PROD-HMM for this subset was precise (Fig. 6D).

The extent of alignment in the 60 bp surrounding the start codon was quantified by the percentage of identical, matching nucleotides in non-gapped base pairs. In Figure 6B and E, the largest counts of exactly predicted orthologs for each species occur in the 60–69% identity bin. The declining counts of exactly predicted start codons at both extreme ends of the percent identity range attest to insufficient information provided when orthologs are very far diverged (too little coding sequence similarity) or very closely related (too much intergenic sequence similarity). Nonetheless, there is a slight negative correlation between prediction offset and percent identity for both *P.abyssi* (–0.2031) and *P.horikoshii* (–0.2832).

In addition to the species-specific trinucleotide frequency patterns trained from sequences immediately upstream and downstream of the translation start sites, contrasting non-synonymous (amino acid changing) and synonymous (silent) substitution rates serve to differentiate prokaryotic coding from intergenic regions. Stronger selective constraints for synonymous changes than for non-synonymous substitutions suppress function-compromising mutations in the protein coding regions, resulting in non-synonymous/synonymous substitution ratios significantly <1 in the orthologous coding regions (11). The non-synonymous/synonymous substitution ratios in the aligned, in-frame coding regions of all 183 orthologs were estimated using the Nei–Gojobori method implemented by the SNAP program (20,21). As this model does not account for transition/transversion biases, disregards codon usage biases and ignores gapped codons, there is a potential to underestimate S, overestimate ds and underestimate the dn/ds ratio. Another option would be to estimate dn/ds using the codon/residue substitution model included in the PAML software package (22). The dn/ds ratio was calculated for all orthologs excepting mutational saturation cases when an NA label is assigned because ps or pn exceed 0.75. The latter case occurred in 30% of the 183 ortholog pairs. The results of this analysis are summarized in Figure 6C and F. The counts of exactly predicted orthologs for each species peak in the NA category, followed by the 0–0.09 ratio bin and, thirdly, either the 0.10–0.19 bin (*P.abyssi*) or the 1.00+ bin (*P.horikoshii*). Notwithstanding the aberrant prediction accuracy at the 1.00+ extreme end of the dn/ds ratio range, there exist positive correlation coefficients for *P.abyssi* (0.1357) and *P.horikoshii* (0.1648). Model prediction accuracy increases with the prevalence of synonymous substitutions in the coding regions.

If the most representative orthologs are to be qualified for model parameterization and the most accurate model predictions are to be earmarked as such, then at least one of the three following threshold criteria should be observed: bitscore 100–800, percent identity 40–80% and dn/ds substitution ratio NA (mutational saturation) or <0.30.

SUMMARY

With so many prokaryotic genomes completely sequenced, start site detection benefits from extensive use of cross-species comparisons. We have presented a novel probabilistic approach to comparative prokaryotic gene annotation. Our product hidden Markov model performs comparative, homology-based genomic annotation by using a pair of orthologous DNA sequences from two related organisms to simultaneously annotate both. This study restricts application to microbial start site delineation. The cumulative species-specific trinucleotide frequency patterns and synonymous/non-synonymous substitution ratios that informed the model's predictions proved to be a more accurate basis of prediction than the pairwise protein-specific sequence similarity information enabling the BLAST alignments. Application of our method with 10-fold cross-validation on experimentally validated *E.coli* K-12 genes and the corresponding *S.thyphimurium* LT2 orthologs confirmed its robustness and proved the significance of the *Pyrococcus* results.

While related to previously proposed GPHMMs (16), our PROD-HMM offers several practical advantages. The PROD-HMM's capability to simultaneously assign different annotations to two aligned base pairs (lacking in GPHMMs) is required when a multi-domain gene loses or acquires a domain or when a eukaryotic exon in one species becomes mutated. Furthermore, the computational complexity of a GPHMM is quadratic in the length of sequences, while in the PROD-HMM it remains linear because the alignment is done efficiently during pre-processing. Finally, the PROD-HMM approach differs slightly from other proposed methods that treat homology as extrinsic information (18).

The model architecture includes a sufficiently comprehensive set of biologically motivated start site features in its architecture of hidden states that it should be applicable to a broad range of species. Thus, this objective algorithm for start site prediction will be properly expanded to work on multiple genomes, evaluated for its effectiveness for organisms of different evolutionary distances, and the results published in a separate study. The software implementation resulting from this endeavor will facilitate post-processing of putative genes designated by automated bacterial genome annotation software packages such as Glimmer.

Depending on the intricacy of the features modeled by the hidden state architecture, intergenic, regulatory, promoter and coding regions could be delimited by PROD-HMM. The method can be further extended to applications beyond microbial genomics, such as eukaryotic annotation, or can even be coupled with probabilistic functional genomic annotation methods.

ACKNOWLEDGEMENTS

M.W. thanks Yu Zheng for running the initial BLASTP and for retrieving and parsing the COG functional annotation for both *Pyrococcus* strains. M.W. was supported in part by the National Science Foundation (IGERT) and NSF (KDI). V.P. and S.K. were supported by NSF (KDI).

REFERENCES

1. Batzoglou,S., Pachter,L., Mesirov,J., Berger,B. and Lander,E. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, **10**, 950–958.
2. Shmatkov,A., Melikyan,A., Chernousko,F. and Borodovsky,M. (1999) Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes. *Bioinformatics*, **15**, 874–886.
3. Altschul,S., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search algorithms. *Nucleic Acids Res.*, **25**, 3389–3402.
4. Salzberg,S., Delcher,A., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.
5. Borodovsky,M. and McIninch,J. (1993) GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem.*, **17**, 123–153.
6. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
7. Schurr,T., Nadir,E. and Margalit,H. (1993) Identification and characterization of *Escherichia coli* ribosomal binding sites by free energy computation. *Nucleic Acids Res.*, **21**, 4019–4023.
8. Hannenhalli,S., Hayes,W., Hatzigeorgiou,A. and Fickett,J. (1999) Bacterial start site prediction. *Nucleic Acids Res.*, **27**, 3577–3582.
9. Yada,T., Yasushi,T., Toshihisa,T. and Nakai,K. (2001) A novel bacterial gene-finding system with improved accuracy in locating start codons. *DNA Res.*, **8**, 97–106.
10. Suzek,B., Ermolaeva,M., Schreiber,M. and Salzberg,S. (2001) A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, **17**, 1123–1130.
11. Makalowski,W. and Boguski,M. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA*, **95**, 9407–9412.
12. Kawarabayasi,Y., Sawada,M., Horikawa,H., Haikawa,Y., Hino,Y., Yamamoto,S., Sekine,M., Baba,S., Kosugi,H., Hosoyama,A., Nagai,Y., Sakai,M., Ogura,K., Otuka,R., Nakazawa,H., Takamiya,M., Ohfuku,Y., Funahashi,T., Tanaka,T., Kudoh,Y., Yamazaki,J., Kushida,N., Oguchi,A., Aoki,K., Nakamura,Y., Robb,T., Horikoshi,K., Masuchi,Y., Shizuya,H. and Kikuchi,H. (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.*, **5**, 55–76.
13. Lecompte,O., Ripp,R., Puzos-Barbe,V., Duprat,S., Heilig,R., Dietrich,J., Thierry,J. and Poch,O. (2001) Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. *Genome Res.*, **11**, 981–993.
14. Rudd,K. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.
15. Delcher,A., Kasif,S., Fleischmann,R., Peterson,J., White,O. and Salzberg,S. (1999) Alignment of whole genomes. *Nucleic Acids Res.*, **11**, 2369–2376.
16. Pachter,L., Alexandersson,M. and Cawley,S. (2002) Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J. Comput. Biol.*, **9**, 389–399.
17. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.
18. Yeh,R.-F., Lim,L.P. and Burge,C.B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.*, **11**, 803–816.
19. Pavlovic,V., Garg,A. and Kasif,S. (2002) A Bayesian framework for combining gene predictions. *Bioinformatics*, **18**, 19–27.
20. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
21. Ota,T. and Nei,M. (1994) Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. *Mol. Biol. Evol.*, **11**, 613–619.
22. Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **45**, 725–736.