

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

Megon Walker, Vladimir Pavlovic, and Simon Kasif*
Bioinformatics Department
Boston University
Boston, MA 02115, USA

February 18, 2002

* Abstract The ever-growing number of completely sequenced prokaryotic genomes facilitates genomic annotation algorithms that make use of cross-species comparisons. 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 nonsynonymous (amino acid changing) and synonymous (silent) substitution rates that differentiate prokaryotic coding from intergenic regions. The new system is evaluated using a set of orthologous *Pyrococcus* gene pairs, for which it demonstrates an improved accuracy of detection. The new architecture has a number of attractive features that distinguish it from previous comparative models such as Paired-Hmms that are discussed in the paper.

Keywords: product hidden Markov model, comparative genomics, prokaryotic genome annotation, translation initiation site detection, *Pyrococcus*

Availability: The new product hidden Markov model-based translation initiation site finding program currently under development is available upon request from vladimir@bu.edu.

*To whom correspondence should be addressed. Tel: +1 617 358 1845; Email: kasif@bu.edu

Introduction

The genomic revolution that started in 1995 with the sequencing of *H. influenza* genome has produced almost a hundred 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 finished or ongoing bacterial sequenc-

ing projects have flooded microbiologists with sequence data and its initial interpretation (<http://www.tigr.org> and <http://www.ncbi.nlm.nih.gov>). This data creates a need to identify and catalog differences and similarities between all organisms and then to mine these comparative events in the attempt to 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.

Comparative gene finding methods train on information from similarity search procedures using as queries the putative proteins derived from lists of open reading frames. Homology information recommends itself to genome annotation because a reliable way to find genes is by detection of close similarity between putative encoded proteins and known proteins from the same or other organisms, by recognition of putative gene similarity to cDNAs from the same or a closely related organisms, or by comparison between closely related genomes. Homology information alone does not solve the annotation problem completely, because many genes (roughly 20-40%) have no significant similarity with other known sequences, or display only partial similarity to known proteins. BLAST and PSI-BLAST (1) On the other hand, *ab initio* finding methods train on DNA sequence only and extract information on 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 and represent protein-coding regions by fixed order or interpolated inhomogeneous three-periodic Markov models. Other techniques use a global approach to determine maximum likelihood sequence parse with regard to a probabilistic hidden Markov model (HMM) (2).

We propose a new probabilistic method for prokaryotic genomic annotation that combines *ab initio* and compara-

tive methods. In particular, the method employs a product hidden Markov model (PROD-HMM) with several hidden states modeling the trinucleotide frequency patterns specific for aligned orthologous sequences. The product hidden Markov model is a composite of two separate hidden Markov models with identical state structure and transitions following the logic of prokaryotic gene organization. It models statistics of pairs of orthologous DNA sequences through species-specific joint 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 sites is valuable for analysis of the putative protein product of a gene and elucidation of signaling information in the 5' region (ribosome binding signals, Shine-Delgarno motifs, and promoters). The difficulty is caused by the absence of 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 (3). Hence, the existing tools such as Glimmer (4) and GeneMark (5) that rely on simple ORF statistic 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 (6). Prediction of bacterial start sites was the focus of Hennenhali et al, 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/noncoding 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 (?). Borodovsky's group incorporated intergenic region, start codon, RBS score, downstream sequence, and pre-start signal in a three-periodic second order Markov model of protein-coding sequence and a second order homogeneous Markov model of non-coding sequence (2). The latest implementation of GenMarkS 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 (3). Yada's GeneHacker Plus HMM uses duration and dicodon statistics to model the coding region locally as well as the upstream translation control signals (7). Finally, the RB-Sfinder developed by Salzberg's group 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 better RBSs are found either upstream or downstream of the originally predicted start site, then the system moves the translation initiation site accordingly (8).

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 noncoding regions. Coding region is characterized by non-synonymous substitution and synonymous substitutions whose ratio is one feature of that region. On the other hand, the substitutions in the non-coding regions are almost random. The PROD-HMM effectively estimates the synonymous-nonsynonymous 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 combining homology and ab initio methods in 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 testing orthologs was accessed by comparison with the genomic sequence annotations given in Genbank.

Materials and Methods

Pyrococcus Strains Used

The two *Pyrococcus* genomes analyzed using the new tool are archeal thermophiles, *P. abyssi* (1.765 Mbp), and *P. horikoshii* (1.738 Mbp), whose genomes have been fully sequenced and annotated (?)¹. The anaerobic organisms have extremely thermostable proteins and enzymes,

¹Of the seven known species of *Pyrococcus*, three have been fully sequenced and annotated, including *P. furiosus* (1.908 Mbp).

some of which are already in use for commercial purposes (?).

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 translocations and insertion/deletion events (9).

P. abyssi and *P. horikoshii* share the same common ancestor, diverged from the ancestor of *P. furiosus*. This is evidenced by intergenomic disruptions to synteny involving rearrangement, translocation, transposition, indel, recombination, and inversion events. In addition, there are longer preserved chromosomal segments between *P. abyssi* and *P. horikoshii* than those observed 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. horikoshii*, and average amino acid identities between close homologs of *P. abyssi* and *P. horikoshii* exceed those between either and *P. furiosus*. Pairwise comparison of the two genomes also reveals high nucleotide conservation (1122 kb in common). However, the resultant number of predicted ORFs is quite different between the two remaining *Pyrococcus* species, despite the comparable genome sizes of *P. abyssi* (1765 ORFs) and *P. horikoshii* (2061 ORFs) (9). Hence, the two species are neither too closely related to offer additional comparative genomic information, nor too far diverged and are amenable to our approach.

Data Collection and Input Preparation

Data files were obtained from Genbank records for *Pyrococcus abyssi* (Genbank accession number (AL096836) and *Pyrococcus horikoshii* (Genbank accession number BA000001). We designed and implemented a semi-automated system that contains core modules written in PERL. It performs the following steps: 1) BLASTP ortholog determination, 2) extraction of genomic coordinates, coding sequences, and up to 200 nucleotides of upstream intergenic sequence, 3) alignment of orthologous nucleotide sequences using the global Smith-Waterman module of MUMmer (10), 4) determination of the nonsynonymous/synonymous substitution ratio for each aligned pair of orthologs, 5) determination of percent identity in the 60 bp surrounding the start sites of each aligned pair of orthologs.

Ortholog selection was performed by converting amino acid sequences of all protein products from *P. horikoshii* into separate databases against which every individual *P. abyssi* peptide sequence from the other genome was compared (BLASTP). A total of 1443 protein orthologs were

identified, similar to (9). During sequence data processing, nucleotide residues corresponding to each of the amino acid sequences and extending from at most 200 nucleotides upstream of the (Genbank annotated) start site to the stop codon were extracted from the genomic DNA sequence and Smith-Waterman aligned to the orthologous sequence from the related genome, including internal gaps in step 3. Inclusion of upstream intergenic sequence data enables the probabilistic model to learn the grammatical structure of genes and upstream regions from the data set during training in order to differentiate between coding and intergenic regions during start codon designation.

As a result, the final input to the model was composed of 183 pairs of aligned orthologous nucleotide sequences of protein coding regions and adjoining upstream sequences satisfying the above criteria. All ortholog pairs for model derivation are Watson strand ORFs in both genomes, and 136 of the pairs include sequence upstream of each start codon and coding region that does not overlap or lie directly adjacent to the preceding coding region. Although not included in this subset, any Crick strand sequences input to the model would be adjusted by formulating the complementary sequence in the opposite direction such that only 5' to 3' analysis is necessary overall.

Computational Model

We model the joint statistics of two related DNA sequences using a *Product Hidden Markov Model* or PROD-HMM. A PROD-HMM is a composite of two (or, in general, two or more) separate HMMs, each with a potentially different transition structure but with a single *joint* emission model. In this discussion we will assume, without loss of generality, that both models have the same N_{hmm} states $\{s_1, \dots, s_{N_{hmm}}\}$ and can emit the same M_{hmm} symbols $\{n_1, \dots, n_{M_{hmm}}\}$. The PROD-HMM will then be a model with $N = N_{hmm}^2$ product states $\{sp_1, \dots, sp_N\}$ describing every possible combination of states of the two separate models. 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 . In a PROD-HMM, however, 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. Similarly, the emission matrix describes the probability of jointly seeing a pair of emission symbols in one product state, as displayed in Equation 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 noncoding region (state NC), any of the three start codon position (S1, S2, S3), any position of an ORF triplet codon

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

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

(C1, C2, C3), and any position of the stop codon (E1, E2, E3). We constructed two identical HMMs of the gene structure of both species. Each constituent model itself does not cover the entire genome but only fits to one gene. Thus, there are $N = 100$ possible states in the this PROD-HMM (all possible paired combinations of the ten labels), each of them emitting $M = 16$ combinations of basepairs. 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 basepair 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 has to satisfy the condition that

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

Similarly, the transition table into a gapped ortholog 2 has to have zeros for all $j \neq q$. This ensures that the state of the last non-gap basepair 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:

$$E_{g1}(i \mid (p, q)) = \Pr (\text{emit symbol } i \text{ in ortholog 1} \mid \text{product state } (p, q))$$

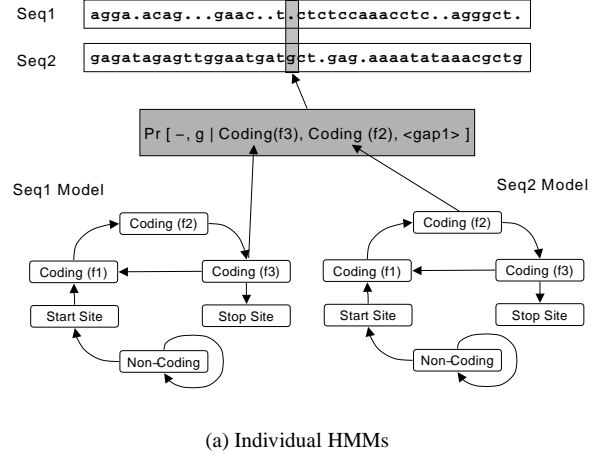
$$E_{g2}(j \mid (p, q)) = \Pr (\text{emit symbol } j \text{ in ortholog 2} \mid \text{product state } (p, q)).$$

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

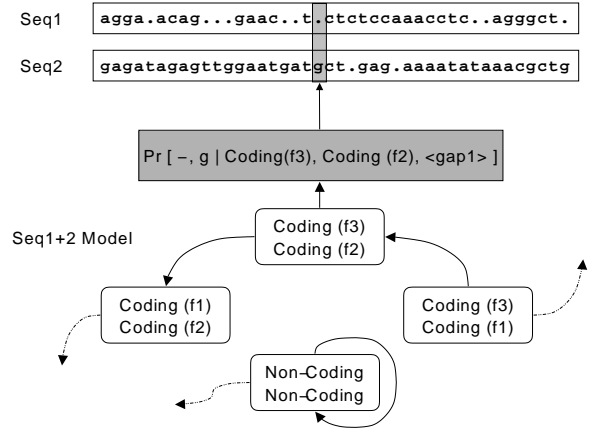
$$T((i, j) \text{ at } k \mid (p, q) \text{ at } k - 1) = \begin{cases} T_{aligned}((i, j) \mid (p, q)) & , \quad a(k) = \text{aligned} \\ T_{g1}(i \mid (p, q)) & , \quad a(k) = \text{gap in 1} \\ T_{g2}(j \mid (p, q)) & , \quad a(k) = \text{gap in 2} \end{cases},$$

and

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



(a) Individual HMMs



(b) Product HMM

Figure 1: Integration of comparative information for start site prediction in prokaryotic genomes using a product 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 (m) or not (n). Otherwise, they are matched with a gap (-) in the other sequences. Each pair of bases (or a base and a gap) can be labeled with one of 100 annotations: the pair (gap,T) is emitted by the (gap,Coding (f1)) state in frame 1 of the second 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.

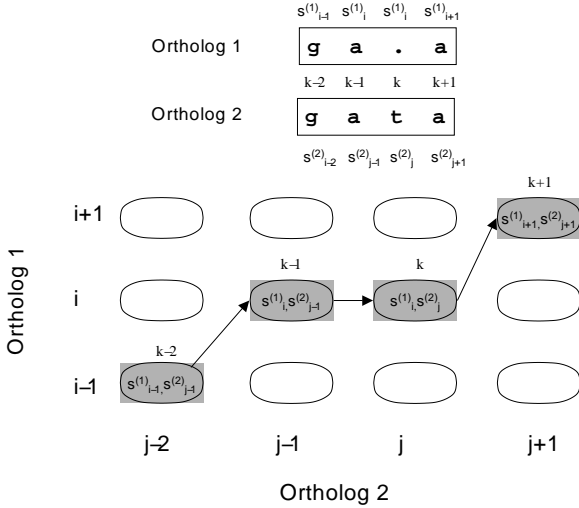


Figure 2: Product HMM needs to be modified when gaps are allowed in alignment of two orthologous sequences. If the two orthologs (1 & 2) are aligned as shown by the path $k - 2, k - 1, k, k + 1$ in the trellis of all possible alignments, then the state of the ortholog 1 at alignment position $k - 1$ ($s_i^{(1)}$) has to remain the same in the 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).

From the above description it is obvious that a PROD-HMM can be represented as an inhomogeneous HMM with a (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 in principle be directly applied to PROD-HMMs. In general, this straightforward application could be computationally expensive because PROD-HMMs state space grows exponentially with the number of compared sequences. However, in the 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 feasible².

Training

During training, we consider pairs of aligned (homologous) nucleotide sequences terminated at stop codons and their Genbank annotation (using one of the ten labels). The emission probability parameters of the PROD-HMM are estimated jointly for the related pair of genomes using maximum likelihood estimation. Counts of how many times an aligned base pair of nucleotides in the input train-

²Approximate inference algorithms based on, for instance, mean field variational factorization or collapsing inference could be used to address the computation complexity in practice.

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-nonsynonymous paradigm. Rare nucleotide substitutions (high conservation) of in-frame states ((f1,f1) and (f2,f2)) can differ significantly from higher substitutions in the third codon position (f3,f3) as well as almost random out-of-frame (e.g., (f1,f2)) and non-coding states. Statistics were obtained from 189 P. horikoshii and P. abyssi pairs.

ing sequences coincide with each of the 100 pairs of states are tabulated in 100 4-by-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-by-100 transition matrix. Emission from gapped pairs (E_{g1} and E_{g2}) and transition-into-gaps 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 nonsynonymous 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 nu-

cleotides from the train/test subset. Each input alignment terminates at the last nucleotides of the stop codons. 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 codons.

Results and Discussion

Algorithm accuracy

The product HMM-based algorithm was evaluated following analysis of a qualifying train/test subset of orthologous genes from two complete prokaryotic genomes, *Pyrococcus abyssi* and *Pyrococcus horikoshii*. The accuracy of the algorithm was estimated by its ability to find the start sites for all orthologs in the training data (self-evaluation) and was based on the amount of offset between translation initiation codons predicted by the algorithm and Genbank annotations for both strains of *Pyrococcus* orthologs. Although the start sites of the sequences in this database are not always verified experimentally, Genbank annotation 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.

Panels a and b of Figure 4 display the offset distributions of *P. abyssi* and *P. horikoshii* translation initiation site predictions made by the PROD-HMM-based algorithm. Prediction errors occurred predominantly downstream (97 predictions), with the remaining 42 of the inexact (offset) predictions placing the start site too far into the upstream intergenic region. As the maximum length of upstream noncoding sequence submitted with each coding ortholog was 200 nucleotides, upstream offsets never exceeded 200, while downstream predictions ran farther afield. The largest offsets resulted from global Smith-Waterman alignment inputs with the following attributes: 1) excessive, erratic gaps dispersed throughout the input alignment of the orthologous upstream intergenic and coding sequences, 2) extremely short upstream sequences in either or both orthologs, and 3) poor homology as indicated by BLASTP bitscores below 100.

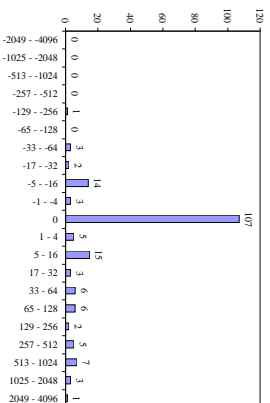
The 229 exact predictions constituted 149 pairs of orthologs: 84 pairs in which both ortholog start sites were predicted exactly, and 61 pairs in which one of the two predictions was offset. Genbank functional annotations revealed that 76 of these 149 pairs were composed of two proteins of undetermined function (labeled “hypo-

thetical proteins” by Genbank and “Function unknown” by COG), while the remaining 73 ortholog pairs contained at least one characterized member. COG functional annotation between the members of these 73 pairs generally corresponded (3 exceptions). 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. The predominant localization of correctly predicted orthologs to regions I and II is a result of high sequence conservation in the regions containing the replication origin, despite the substantial inversion of region 1 post-speciation. Metabolic enzymes predicted in other regions included five amino acid transport and metabolism proteins and four nucleotide transport and metabolism protein. The ribosomal operon was particularly well conserved following speciation, enabling our model to correctly predict five energy production and conversion proteins. The largest correctly annotated functional group, comprised of thirteen ortholog pairs vaguely classified as “General function prediction only” by COG, contained two NADH-dependent dehydrogenase related proteins, an NADH oxidase, methionyl and an alanyl tRNA synthetase fragments, among others. Eleven of the thirteen proteins exhibited coordinates within regions I and II of both genomes.

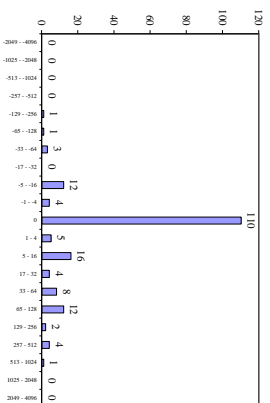
TBLASTN and TBLASTX

Benchmark TBLASTN and TBLASTX programs of the BLAST suite of local alignment tools were utilized for translation start site detection by modeling the task purely as a sequence alignment problem. To predict the gene structures in two given 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. Notwithstanding the similar sequence input to each program (for each ortholog, at most 200 upstream intergenic nucleotides and the entire coding region were submitted), TBLASTX comparisons involve two sets of six reading frames between which to form alignments, resulting in more inaccurate annotations than TBLASTN, whose single reading frame protein query divulges complete annotation for one genome. Reversing the query and database orthologs separately predicts the start site for each member of an ortholog pair.

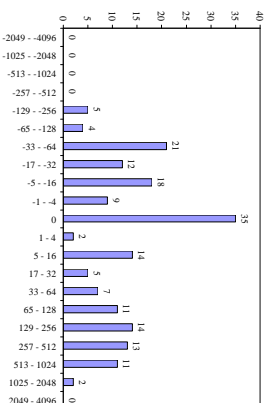
The model predictions are informed by cumulative species-specific trinucleotide patterns and substitution ratios learned from all train/test subset orthologs and characterized in the emission and transition matrices. An exact prediction by the algorithm occurred when the S1, S2, and S3 labels emitted by the model matched ex-



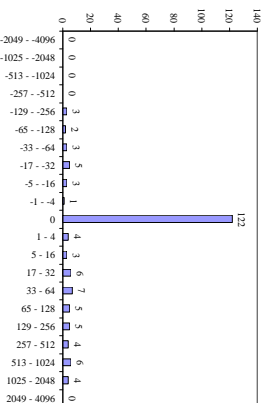
(a)



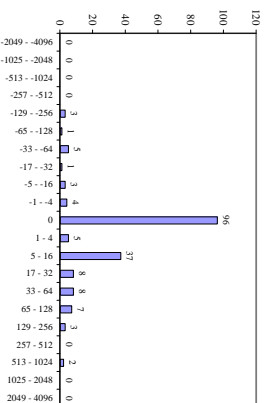
(b)



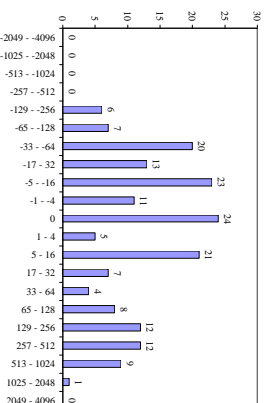
(c)



(d)



(e)



(f)

Figure 4: Prediction of start sites in 183 *P. abyssii* and *P. horikoshii* ortholog pairs using comparative methods. Panels (a)-(f) show offset statistics of translation initiation sites predicted by the product hidden Markov model-based algorithm, TBLASTN, and TBLASTX with respect to Genbank annotation. All bins are 100 base 2. Upstream predictions offset in the intergenic region are displayed in negative bins; prediction offsets distributed downstream are in positive bins. Significantly higher PROD-HMM prediction accuracy was observed for *P. horikoshii* (122 exactly predicted orthologs) than for *P. abyssii* (107 exactly predicted orthologs), however the opposite trend is observed for TBLASTN and TBLASTX, with higher accuracy for *P. abyssii* orthologs. The cumulative species-specific trinucleotide and synonymous/nonsynonymous substitution patterns that informed the PROD-HMM's predictions after training on all train/test subset orthologs proved to be a more accurate basis of prediction than the pairwise protein-sequence similarity information enabling the BLAST alignments.

actly 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 TBLASTN and TBLASTX output, the initial aligned codons of the hsp with the lowest expectation value and maximal bitscore were evaluated as the start site predictions for each 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 TBLASTN (panels c and d) and TBLASTX (panels e and f) for orthologs of both 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%), but TBLASTN captures the largest number of precise predictions (61%). For the *P. horikoshii* subset, the model performs better overall (67%), surpassing the number of exactly predicted (zero offset) start site locations found by TBLASTN (52%) and TBLASTX (13%). Significantly higher model prediction accuracy was observed for *P. horikoshii* (122 exactly predicted orthologs) than for *P. abyssi* (107 exactly predicted orthologs). Following speciation, large collinear regions of *P. horikoshii* were differentially lost, including segment b9 and the entire maltose and phosphate operons (ref). It is not clear how this relates to the transition and emission probabilities learned by the model from pairwise ortholog sequence alignments, as the extent of chromosomal rearrangements between closely related species appears to be independent of sequence conservation (ref to Poch). Furthermore, the opposite trend is observed for exact predictions by TBLASTN and TBLASTX, with higher accuracy for *P. abyssi* orthologs.

Does each annotation system perform best on (exactly predict the start sites of) the same set of orthologs? 78 of the *P. abyssi* orthologs exactly predicted by the model are also exactly annotated by TBLASTN, while the zero-offset predictions by the model and TBLASTX coincided in only 26 cases for *P. abyssi*. 76 of the *P. horikoshii* orthologs exactly predicted by the model are also exactly annotated by TBLASTN compared to 21 of the *P. horikoshii* orthologs exactly predicted by the model that were also exactly annotated by TBLASTX.

Model vs TBLASTX: specific cases

As neither the model nor TBLASTX involve start site specification by query or database sequences, further prediction analysis was conducted on members of the 60 paired orthologs for which the algorithm predicts both start sites exactly but for which both TBLASTX predictions are offset. The algorithm's accuracy of start site prediction exceeded that of conservation detection sys-

tem 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 nucleotides 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, thus only one output is reproduced for each pair discussed here.

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 codon. Figure 5(a) displays partial TBLASTX output for an 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). The PRODHMM-based algorithm predictions were exact for both orthologs. *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. TBLASTX detects the sequence similarity in the highest bitscoring hsp 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 nucleotides downstream of the start codons). The initial 68 amino acids skipped in 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 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 4 residues downstream of the true initial methionine in the *P. abyssi* sequence (3 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,

```

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  LASNIAIGHVRYSTSGSLSEVQPLEVRCCEYELAIHNGTLTNFIPLRRLYEGMGIKFHS 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  *VLPLWLCWCLSNVGDSTRFLVLGVHRG*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 frequent presence of high scoring HSPs upstream or downstream of the true start site. Panels (a) and (b) display partial TBLASTX output annotating a glutamine-dependent phosphoribosyl amidotransferase and a glucose-1-phosphate thymidyltransferase, 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. Panel (c) displays 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.

then aligned words (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 nucleotides upstream of the true start codons (52 codons upstream of the initial methionines). The alignment upstream of the hydrogenase protein displays about 50% amino acid identity in the highest scoring hsp when the *P. horikoshii* query is compared to the *P. abyssi* database, while an 8-residue-long low complexity region is masked in the upstream alignment of the highest scoring hsp when the *P. abyssi* query is compared to 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 codon (methionine encoded by a *P. abyssi* ATG and methionine encoded by a *P. horikoshii* ATG) 52 codons downstream.

Thresholds (Correlation of Bitscore & Percent Identity & dn/ds substitution ratio)

Three features descriptive of each ortholog pair were designated for further analysis: BLASTP bitscore, percent identity surrounding the start codon, and nonsynonymous/synonymous substitution rates in the coding regions. These criteria were correlated with the binned prediction offsets, and 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, which reflects the raw score independently of the scoring matrix used and indicates the extent of sequence similarity between the two orthologous proteins. Of the 183 ortholog pairs in the train/test subset, 8 pairs scored below 100. The model’s predictions

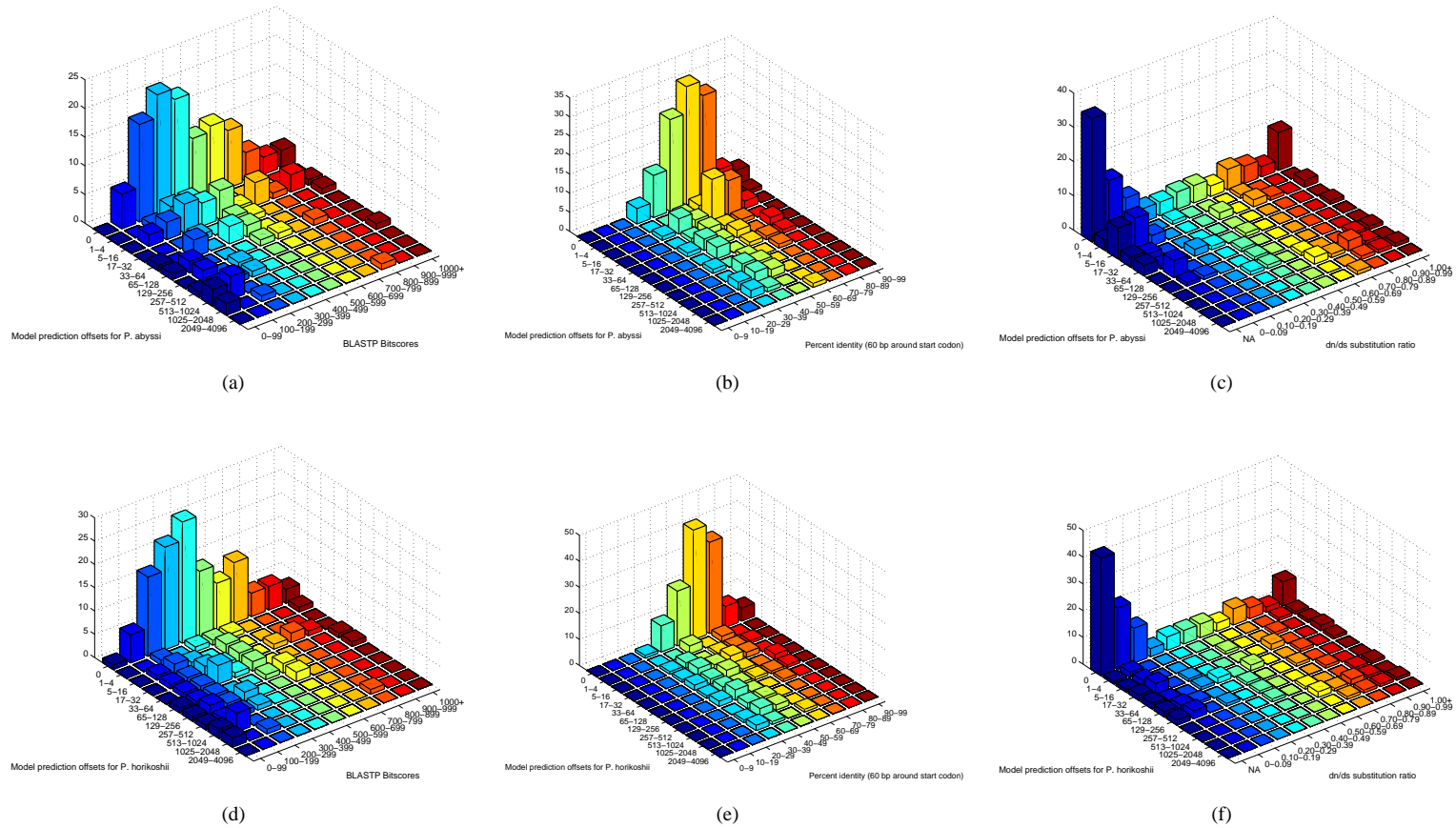


Figure 6: Three criteria were correlated with the binned offsets from each strain in order to inform future ortholog selection: BLASTP bitscore, percent identity surrounding the start codon, and nonsynonymous/synonymous substitution rates in the coding regions. The resultant thresholds will constitute the biological rationale by which orthologs are selected for model training (emission and transition matrix parameterization) in future analyses: bitscore $\zeta= 100$, percent identity $\zeta= 60\%$ & dn/ds substitution ratio $\zeta= 0.60$ or 0.40 .

for the 8 corresponding translation initiation sites were all significantly offset by more than 70 bp, with no exact predictions occurring until after the bitscore exceeded 100. As indicated by Figure 6(a 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.

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 panels b and e of Figure 6, 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 translation start site, contrasting nonsynonymous (amino acid changing) and synonymous (silent) substitution rates serve to differentiate prokaryotic coding from intergenic regions. Stronger selective constraints for synonymous changes than for nonsynonymous substitutions suppress function-compromising mutations in the protein coding regions, resulting in nonsynonymous/synonymous substitution ratios significantly smaller than one. The nonsynonymous/synonymous substitution rates in the aligned, in frame coding regions of all 183 orthologs were estimated using the SNAP program (11). 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 (?). The dn/ds ratio is calculated for all orthologs excepting mutational saturation cases when a 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 6, panels d 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 best 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 less than 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-based algorithm performs comparative modeling for successful homology-based genomic annotation by using a pair of orthologous DNA sequences from two related organisms to simultaneously annotate both. Depending on the intricacy of the features modeled by the hidden state architecture, intergenic, regulatory, promoter, and coding regions could be delimited by this method. This study restricts application to microbial start site delineation. The cumulative species-specific trinucleotide frequency patterns and synonymous/nonsynonymous substitution ratios that informed the model's predictions after training on all train/test subset *Pyrococcus* orthologs proved to be a more accurate basis of prediction than the pairwise protein-specific sequence similarity information enabling the BLAST alignments. 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 carefully tested with cross-validation across a number of bacterial and archeal taxa, including strains with starts experimentally verified by N-terminal protein sequencing, and published elsewhere. It will be properly expanded to work on multiple genomes and evaluated for its effectiveness for organisms of different evolutionary distances. 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.

Acknowledgments

M.W. thanks Yu Zheng for running the initial BLASTP and for retrieving and parsing the COG functional annotation for both strains. M.W. was supported in part by the National Science Foundation (IGERT).

11. Ota, T. and Nei, M. (1994) Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. *Molecular Biology and Evolution*, **11**(4), 613–619.

REFERENCES

1. 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 Research*, **25**(17), 3389–3402.
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**(11), 874–886.
3. 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 Research*, **29**(12), 2607–2618.
4. S.L., S., A.L., D., S., K., and O., W. Microbial gene identification using interpolated markov models..
5. Borodovsky, M. and McIninch, J. (1993) Genemark:parallel gene recognition for both dna strands. *Comput. Chem.*, **17**.
6. T., S., E., N., and H., M. Identification and characterization of e.coli ribosomal binding sites by free energy computation.
7. Yada, T., Yasushi, T., Toshihisa, T., and Nakai, K. (2001) A novel bacterial gene-finding system with improved accuracy in locating start codons. *DNA Research*, **8**(3), 97–106.
8. Suzek, B., Ermolaeva, M., Schreiber, M., and Salzberg, S. (2001) A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, **17**(12), 1123–1130.
9. 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 Research*, **11**.
10. Delcher, A., Kasif, S., Fleischmann, R., Peterson, J., White, O., and Salzberg, S. (1999) Alignment of whole genomes. *NAR Nucleic Acids Research*, **11**.

Figure Legends

* List of Figures

- 1 Integration of comparative information for start site prediction in prokaryotic genomes using a product 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 (m) or not (n). Otherwise, they are matched with a gap (-) in the other sequences. Each pair of bases (or a base and a gap) can be labeled with one of 100 annotations: the pair (gap,T) is emitted by the (gap,Coding (f1)) state in frame 1 of the second 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. 4
- 2 Product HMM needs to be modified when gaps are allowed in alignment of two orthologous sequences. If the two orthologs (1 & 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 the 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). 5
- 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-nonsynonymous paradigm. Rare nucleotide substitutions (high conservation) of in-frame states ((f1,f1) and (f2,f2)) can differ significantly from higher substitutions in the third codon position (f3,f3) as well as almost random out-of-frame (e.g., (f1,f2)) and non-coding states. Statistics were obtained from 189 *P. horikoshii* and *P. abyssi* pairs. 5
- 4 Prediction of start sites in 183 *P. abyssi* and *P. horikoshii* ortholog pairs using comparative methods. Panels (a)-(f) show offset statistics of translation initiation sites predicted by the product hidden Markov model-based algorithm, TBLASTN, and TBLASTX with respect to Genbank annotation. All bins are lob base 2. Upstream predictions offset in the intergenic region are displayed in negative bins; prediction offsets distributed downstream are in positive bins. Significantly higher PROD-HMM prediction accuracy was observed for *P. horikoshii* (122 exactly predicted orthologs) than for *P. abyssi* (107 exactly predicted orthologs), however the opposite trend is observed for TBLASTN and TBLASTX, with higher accuracy for *P. abyssi* orthologs. The cumulative species-specific trinucleotide and synonymous/nonsynonymous substitution patterns that informed the PROD-HMM's predictions after training on all train/test subset orthologs proved to be a more accurate basis of prediction than the pairwise protein-specific sequence similarity information enabling the BLAST alignments. 7
- 5 Relatively poor performance of TBLASTX for 60 ortholog pairs annotated correctly by the PROD-HMM can be explained by frequent presence of high scoring HSPs upstream or downstream of the true start site. Panels (a) and (b) display partial TBLASTX output annotating a glutamine-dependent phosphoribosyl amidotransferase and a glucose-1-phosphate thymidyltransferase, 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. Panel (c) displays 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. 9
- 6 Three criteria were correlated with the binned offsets from each strain in order to inform future ortholog selection: BLASTP bitscore, percent identity surrounding the start codon, and nonsynonymous/synonymous substitution rates in the coding regions. The resultant thresholds will constitute the biological rationale by which orthologs are selected for model training (emission and transition matrix parameterization) in future analyses: bitscore $\zeta = 100$, percent identity $\zeta = 60\%$ & dn/ds substitution ratio $\zeta = 0.60$ or 0.40 10