The importance of short protein-coding genes (usually defined as no longer than 100 codons) and corresponding small proteins and peptides, in prokaryotic and eukaryotic organisms is becoming increasingly obvious as the pervasive role of small proteins as signaling molecules, and as regulators of protein expression and functionality is being uncovered (e.g., [1,2]).

In E. coli, the smallest known functional gene-product is a 29-amino acid peptide involved in K⁺ transport (KdpF) [3]. ORFs with as few as 14 amino acids have been predicted to encode functional genes in E. coli, and with as few as 28 amino acids in Saccharomyces cerevisiae [4], while many short peptides have been identified, including a 13 aa peptide encoded within the E. coli Shiga-like toxin operon. Moreover, artificial constructs encoding just six amino acids were able to transcribe and result in functional gene-products involved in intracellular signaling in B. subtilis [5]. In eukaryotes, many important signaling molecules are short peptides, including various peptide hormones, cytokines and co-repressors or co-activators [6,7]. In eukaryotes, evidence is accumulating on the existence of widespread very short ORFs, called uORFs, located 5’ of a reference gene, which post-transcriptionally regulate translation of the gene [8-10].

Computational Identification of Small Genes in Bacterial Genomes

In spite of the high sensitivity of computational prokaryotic-gene predictor methods, it is recognized that short genes are still often overlooked by published annotations. Computational prediction of short genes is risky. Short sequences contain less information that can signal their coding capacity, and small genes may not follow the same codon-composition properties of the average gene, and may encode for peptides with unique amino acid composition. Among the most common predictors, issues of specificity limit the ability of gene predictors to identify small genes with high sensitivity. For example, among popular predictors, Glimmer tends to predict many more short genes than others, but among the many predicted short genes the rate of false positives appears to be so high that a cut-off on the minimal length of predicted ORFs has been included in more recent versions of Glimmer [11], preventing prediction of very short genes.

In an attempt to identify coding regions missed by published annotations, we developed procedures to identify genomic regions with significant 3-base sequence periodicity, which when associated with ORF structures could signal the presence of a coding sequence [12]. These procedures were implemented in the N-Profile Analysis Computational Tool (NPACT), a web-based bioinformatics tool available at http://genome.ufl.edu/npact. We collected with NPACT all genes predicted by the annotations of 1000 prokaryotic genomes, by four other popular prediction methods, and ORFs identified by sequence 3-base periodicity, recording conservation of all genes across different phyla. We identified a total of 4,421,545 predicted genes, among which 889,837 ORFs were not included in the published genome annotations. Most of these excluded genes (83%) corresponded to ORFs no longer than 100 codons (Table 1). This collection of short ORFs would represent almost three-times as many short genes than others, but among the many predicted short genes the rate of false positives appears to be so high that a cut-off on the minimal length of predicted ORFs has been included in more recent versions of Glimmer [11], preventing prediction of very short genes.

In an attempt to identify coding regions missed by published annotations, we developed procedures to identify genomic regions with significant 3-base sequence periodicity, which when associated with ORF structures could signal the presence of a coding sequence [12]. These procedures were implemented in the N-Profile Analysis Computational Tool (NPACT), a web-based bioinformatics tool available at http://genome.ufl.edu/npact. We collected with NPACT all genes predicted by the annotations of 1000 prokaryotic genomes, by four other popular prediction methods, and ORFs identified by sequence 3-base periodicity, recording conservation of all genes across different phyla. We identified a total of 4,421,545 predicted genes, among which 889,837 ORFs were not included in the published genome annotations. Most of these excluded genes (83%) corresponded to ORFs no longer than 100 codons (Table 1). This collection of short ORFs would represent almost three-times as many short genes than others, but among the many predicted short genes the rate of false positives appears to be so high that a cut-off on the minimal length of predicted ORFs has been included in more recent versions of Glimmer [11], preventing prediction of very short genes.

In spite of the high sensitivity of computational prokaryotic-gene predictor methods, it is recognized that short genes are still often overlooked by published annotations. Computational prediction of short genes is risky. Short sequences contain less information that can signal their coding capacity, and small genes may not follow the same codon-composition properties of the average gene, and may encode for peptides with unique amino acid composition. Among the most common predictors, issues of specificity limit the ability of gene predictors to identify small genes with high sensitivity. For example, among popular predictors, Glimmer tends to predict many more short genes than others, but among the many predicted short genes the rate of false positives appears to be so high that a cut-off on the minimal length of predicted ORFs has been included in more recent versions of Glimmer [11], preventing prediction of very short genes.

In an attempt to identify coding regions missed by published annotations, we developed procedures to identify genomic regions with significant 3-base sequence periodicity, which when associated with ORF structures could signal the presence of a coding sequence [12]. These procedures were implemented in the N-Profile Analysis Computational Tool (NPACT), a web-based bioinformatics tool available at http://genome.ufl.edu/npact. We collected with NPACT all genes predicted by the annotations of 1000 prokaryotic genomes, by four other popular prediction methods, and ORFs identified by sequence 3-base periodicity, recording conservation of all genes across different phyla. We identified a total of 4,421,545 predicted genes, among which 889,837 ORFs were not included in the published genome annotations. Most of these excluded genes (83%) corresponded to ORFs no longer than 100 codons (Table 1). This collection of short ORFs would represent almost three-times as many short genes than others, but among the many predicted short genes the rate of false positives appears to be so high that a cut-off on the minimal length of predicted ORFs has been included in more recent versions of Glimmer [11], preventing prediction of very short genes.

In an attempt to identify coding regions missed by published annotations, we developed procedures to identify genomic regions with significant 3-base sequence periodicity, which when associated with ORF structures could signal the presence of a coding sequence [12]. These procedures were implemented in the N-Profile Analysis Computational Tool (NPACT), a web-based bioinformatics tool available at http://genome.ufl.edu/npact. We collected with NPACT all genes predicted by the annotations of 1000 prokaryotic genomes, by four other popular prediction methods, and ORFs identified by sequence 3-base periodicity, recording conservation of all genes across different phyla. We identified a total of 4,421,545 predicted genes, among which 889,837 ORFs were not included in the published genome annotations. Most of these excluded genes (83%) corresponded to ORFs no longer than 100 codons (Table 1). This collection of short ORFs would represent almost three-times as many short genes than others, but among the many predicted short genes the rate of false positives appears to be so high that a cut-off on the minimal length of predicted ORFs has been included in more recent versions of Glimmer [11], preventing prediction of very short genes.

In an attempt to identify coding regions missed by published annotations, we developed procedures to identify genomic regions with significant 3-base sequence periodicity, which when associated with ORF structures could signal the presence of a coding sequence [12]. These procedures were implemented in the N-Profile Analysis Computational Tool (NPACT), a web-based bioinformatics tool available at http://genome.ufl.edu/npact. We collected with NPACT all genes predicted by the annotations of 1000 prokaryotic genomes, by four other popular prediction methods, and ORFs identified by sequence 3-base periodicity, recording conservation of all genes across different phyla. We identified a total of 4,421,545 predicted genes, among which 889,837 ORFs were not included in the published genome annotations. Most of these excluded genes (83%) corresponded to ORFs no longer than 100 codons (Table 1). This collection of short ORFs would represent almost three-times as many short genes than others, but among the many predicted short genes the rate of false positives appears to be so high that a cut-off on the minimal length of predicted ORFs has been included in more recent versions of Glimmer [11], preventing prediction of very short genes.

In an attempt to identify coding regions missed by published annotations, we developed procedures to identify genomic regions with significant 3-base sequence periodicity, which when associated with ORF structures could signal the presence of a coding sequence [12]. These procedures were implemented in the N-Profile Analysis Computational Tool (NPACT), a web-based bioinformatics tool available at http://genome.ufl.edu/npact. We collected with NPACT all genes predicted by the annotations of 1000 prokaryotic genomes, by four other popular prediction methods, and ORFs identified by sequence 3-base periodicity, recording conservation of all genes across different phyla. We identified a total of 4,421,545 predicted genes, among which 889,837 ORFs were not included in the published genome annotations. Most of these excluded genes (83%) corresponded to ORFs no longer than 100 codons (Table 1). This collection of short ORFs would represent almost three-times as many short genes than others, but among the many predicted short genes the rate of false positives appears to be so high that a cut-off on the minimal length of predicted ORFs has been included in more recent versions of Glimmer [11], preventing prediction of very short genes.

In an attempt to identify coding regions missed by published annotations, we developed procedures to identify genomic regions with significant 3-base sequence periodicity, which when associated with ORF structures could signal the presence of a coding sequence [12]. These procedures were implemented in the N-Profile Analysis Computational Tool (NPACT), a web-based bioinformatics tool available at http://genome.ufl.edu/npact. We collected with NPACT all genes predicted by the annotations of 1000 prokaryotic genomes, by four other popular prediction methods, and ORFs identified by sequence 3-base periodicity, recording conservation of all genes across different phyla. We identified a total of 4,421,545 predicted genes, among which 889,837 ORFs were not included in the published genome annotations. Most of these excluded genes (83%) corresponded to ORFs no longer than 100 codons (Table 1). This collection of short ORFs would represent almost three-times as many short genes than others, but among the many predicted short genes the rate of false positives appears to be so high that a cut-off on the minimal length of predicted ORFs has been included in more recent versions of Glimmer [11], preventing prediction of very short genes.
of each expressed protein-coding sequence. The ribosome-profiling sequencing technology can provide deeper measurements and more accurate quantification than mass spectrometry proteomics, and can provide information not only on the amount of protein produced in given conditions but also on the dynamics of protein expression [15].

In contrast to computational gene prediction and to other experimental approaches, the identification power of ribosome profiling is independent on gene length, allowing detection of very short expressed genes and regulatory peptides (Figure 1). Furthermore, translation initiation sites (TIS) can be identified exploiting the activity of inhibitors stalling ribosomes at or proximal to TISs [16-18], a strategy also referred to as global translation initiation sequencing or GTI-seq [18]. Ribosome profiling provides the opportunity not only to identify precisely coding region, but also to uncover events and mechanisms of post-transcriptional control of protein expression in response to environmental stimuli. By genome-wide profiling, previously unrecognized widespread post-transcriptional regulation of gene translation and translational response to stress have been newly identified providing ample evidence of post-transcriptional regulation in eukaryotes (e.g., [10,14]). New fundamental biological processes have been discovered [19,20] and opportunities for biotechnological innovation have been identified [21]. It was demonstrated that in vertebrates the majority of expressed genes are associated with the translation of peptides encoded by uORFs in the 5’UTR or by internal out-of-frame ORFs (AltORFs) [10]. By bridging the gap between global measurements of mRNA and protein levels, ribosome profiling provides the most advanced tool for accurately and directly measuring levels of protein expression, and can provide the necessary information for building an optimal protein sequence search database for MS-based proteomics [22].

Although identification of small proteins by experimental data on expression and functionality is facilitated by the ever-growing availability of these high throughput genomic methods, it is unlikely that computational gene predictions will be soon superseded by experimental methods, whose sensitivity is limited by the necessity to identify the conditions required for expression of many genes. Integration of information from sequence features, conservatism, and

<table>
<thead>
<tr>
<th>Gene set</th>
<th>Total</th>
<th>Conserved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotated</td>
<td>3,531,708</td>
<td>3,239,662</td>
</tr>
<tr>
<td>Annotated ≤300 nt</td>
<td>390,233</td>
<td>235,413</td>
</tr>
<tr>
<td>New</td>
<td>889,837</td>
<td>182,294</td>
</tr>
<tr>
<td>New ≤300 nt</td>
<td>740,695</td>
<td>97,836</td>
</tr>
</tbody>
</table>

Table 1: Total number and conserved genes identified among annotated or newly-predicted genes in 1000 bacterial genomes [12].

transcriptomic, translatomic, and proteomic analyses, will most likely provide the best strategy for obtaining the most complete picture of the coding potential of prokaryotic and eukaryotic organisms [23].

Acknowledgment

This work is supported by NIH Grant 5R01GM87485-2.

References


Figure 1: RIBO-seq evidence of expression of the newly identified gene R49, encoding a 17aa-long peptide 5’ of annotated gene PA4163. Three rows of symbols represent potential start codons (circles) and stop codons (vertical red bar) in the three reading frames of the strand encoding the represented gene, with the top row representing the frame of the gene. Bigger circlets indicate the position of canonical start codons ATG or GTG, and smaller circlets indicate alternative start codons TTG, CTG, or ATT.


