Capturing Whole-Genome Characteristics in Short Sequences Using a Naïve Bayesian Classifier

Rickard Sandberg,1,2,3 Gösta Winberg,1,2 Carl-Ivar Bränden,1 Alexander Kaske,2 Ingemar Ernberg,1 and Joakim Cöster2

1Microbiology and Tumor Biology Center, Karolinska Institute, S-171 77 Stockholm, Sweden; 2Virtual Genetics Laboratory AB, S-171 77 Stockholm, Sweden

Bacterial genomes have diverged during evolution, resulting in clearcut differences in their nucleotide composition, such as their GC content. The analysis of complete sequences of bacterial genomes also reveals the presence of nonrandom sequence variation, manifest in the frequency profile of specific short oligonucleotides. These frequency profiles constitute highly specific genomic signatures. Based on these differences in oligonucleotide frequency between bacterial genomes, we investigated the possibility of predicting the genome of origin for a specific genomic sequence. To this end, we developed a naïve Bayesian classifier and systematically analyzed 28 eubacterial and archaeal genomes. We found that sequences as short as 400 bases could be correctly classified with an accuracy of 85%. We then applied the classifier to the identification of horizontal gene transfer events in whole-genome sequences and demonstrated the validity of our approach by correctly predicting the transfer of both the superoxide dismutase (sodC) and the bioC gene from Haemophilus influenzae to Neisseria meningitis, correctly identifying both the donor and recipient species. We believe that this classification methodology could be a valuable tool in biodiversity studies.

The complete genome sequences of many organisms are now available. This permits comprehensive comparative analysis of genome structures. Recent investigations have reported differences in both the subsets of proteins the genomes encode (Rubin et al. 2000) and the frequency of occurrence of many short oligonucleotides (Karlin and Burge 1995), hereafter called “motifs”. The comparative studies have mostly focused on short motifs, such as dinucleotides (Karlin et al. 1992; Goldman 1993; Karlin and Ladunga 1994; Karlin and Burge 1995; Karlin et al. 1997; Nakashima et al. 1997, 1998), tri-nucleotides (Karlin et al. 1992; Goldman 1993; Karlin and Ladunga 1994; Karlin and Ladunga 1994; Karlin et al. 1997) and tetranucleotides (Karlin and Ladunga 1994; Karlin et al. 1997). Motifs up to eight nucleotides long were recently analyzed and compared using the chaos game representation (CGR) (Deschavanne et al. 1999). The existence of specific genomic signatures (motif frequency profiles) has been reported for all motif lengths. It has also been shown that intergenic differences are generally higher than intragenomic differences (Karlin and Ladunga 1994; Karlin and Burge 1995; Karlin et al. 1997; Nakashima et al. 1998; Deschavanne et al. 1999). Genes from different prokaryotic and eukaryotic organisms have been classified using dinucleotide composition (Nakashima et al. 1997, 1998). In the present study, we examined the conditions for identifying the genome of origin for a specific genomic sequence, using the genomic signature concept. The naïve Bayesian classifier used here is a probabilistic technique commonly used in text classification (Robertson and Sparck-Jones 1976; Langley 1992; Lewis and Gale 1994). The methodology is illustrated in Figure 1. First, all genomes are scanned for the occurrences of all possible overlapping motifs with a length of \( n \) nucleotides (4\(^n\) possible motifs). Then, a genomic sequence is chosen at random from anywhere inside a genome (coding or noncod-
From this genomic sequence, all overlapping motifs are extracted. The naïve Bayesian classifier uses the extracted motifs to predict their most probable genomic origin by comparing the frequencies of the extracted motifs with the motif frequencies of the different genomes. In the present study, we determined how the performance of the Bayesian classifier of genomic sequences depends on motif length and sequence length. We demonstrate its generalizing ability and its capacity to discriminate between closely related microorganisms using a sequence sample of only a few hundred nucleotides. We also demonstrate how these properties of the classifier can be applied to the problem of pinpointing horizontal gene transfer events (Doolittle 1999), identifying both donor and recipient (Kroll et al. 1998). As discussed by Eisen (Eisen 2000), there are very few well-documented cases of horizontal gene transfer events where both donor and recipient strains are known. The horizontal gene transfer events from H. influenzae to N. meningitis are one of the best-documented cases available and were therefore chosen as a reference system.

RESULTS

Visualizing the Genomic Signature Concept

To visualize the difference in motif frequencies between and within the genomes of prokaryotic species, we performed a principal components analysis on the motif frequency profiles (Fig. 2). Different eubacterial and archaeal genomes form clusters in the three-dimensional space drawn by principal components 4, 5, and 6 ("PCA space"). Closely related microorganisms cluster together in PCA-space as shown for Helicobacter and Pyrococcus.

Dependence of Classification Accuracy on Motif and Sequence Length

The performance of the classifier depends on the motif length used for establishing the genomic signature. Training was performed on 90% of the genome sequences, and the remaining 10% was used to evaluate classification accuracy. Longer motifs result in a more specific representation of the genome (Fig. 3). Classification accuracy increases with motif length (Fig. 3), and highest accuracy was achieved with eight and nine-nucleotide motifs. We classified sequences of six different lengths: 35, 60, 100, 200, 400, and 1000 nucleotides (nt), and monitored the classification accuracy. The accuracy increases with the sequence length, and sequences of 400-nt length were correctly classified in 85 of 100 cases (Fig. 3). For 100 nucleotide sequences, the mean classification accuracy is 60%, and for short sequences (35 nt) the classifier correctly predicts 36 of 100 test sequences on average. In further studies, we used nine-nucleotide motifs. Inspection of the conditional probabilities within the classifiers indicates that the classification depends not on a few species-specific motifs, but rather on the whole set of motifs (data not shown).

Figure 2 Visualizing the genomic signature concept. Principal components analysis (PCA) was performed on the motif frequencies of 25 genomic sequences from each eubacterial and archaeal genome. The sequences are mapped into a three-dimensional PCA-space, drawn by three principal components (here components 4, 5, and 6). Each sequence was randomly chosen and had a length of 1000 bp. Closely related microorganisms cluster together in PCA-space here shown for Pyrococcus strains pabyssi and horikoshii and for Helicobacter pylori strains 26695 and J99. For clarity, arrows indicate each distinct genome cluster when similar colors were used to plot the sequences from different eubacterial and archaeal genomes. The figure was plotted using Spotfire (Spotfire Inc.).
Pylori results are presented in Figure 5. For both species with highly similar motif frequency profiles. The strain MC58 and serotype A strain Z2491), by discriminating different strains of the same species, exemplified by H. pylori (serotypes A and B strain M53B and serotype A strain Z2491), Pyrococcus abyssi and horkoshi OT3), and Chlamydia trachomatis (strain Nigg and Serovar D [D/UW-3/Cx]). Each new classifier trained had to correctly discriminate between two different strains of the same species with highly similar motif frequency profiles. The results are presented in Figure 5. For both N. meningitidis and H. Pylori, roughly 200 nucleotides were needed for accuracy around 90% (Fig. 5), but for Pyrococcus and Chlamydia, 60 nucleotides sufficed for discrimination with 90% accuracy. However, prediction accuracy was also enhanced because only two classes were to be discriminated, compared to the previous experiments with 25 classes. These results suggest a “hierarchical classification” procedure that first classifies a sequence to correct species and then, using a species-specific classifier, correctly identifies the strain.

Identifying Donor and Recipient Strains in Horizontal Gene Transfer Events

The possibility of using the classifier for identifying regions of horizontally transferred genes was examined. Many different methods have been proposed for finding putative cases of horizontal gene transfer (Mrazek and Karlin 1999; Eisen 2000; García-Vallve et al. 2000). However, most methods are designed to search genomes for putative horizontally transferred genes without identifying the donor (Eisen 2000). A general problem in analyzing horizontal gene transfer events is the validation of the results (Eisen 2000). However, one case where strong evidence for horizontal gene transfer exists is from H. influenzae to N. meningitidis (Kroll et al. 1998; Eisen 2000). The SodC gene and the Bio gene cluster show strong homology to H. influenzae genes, and the 29-nt long Haemophilus Uptake Sequence (HmUS) was found downstream of both of the genes (Kroll et al. 1998). The horizontal gene transfer events between H. influenza and N. meningitidis were used to evaluate our classifier for identifying both the donor and the recipient in a horizontal gene transfer event.

The availability of complete sequences for both N. meningitidis serotypes A and B further gave us the possibility to constrain the method by conducting the in silico experiment on two highly similar genomes. Two percent of the N. meningitidis genomes (serotype A 1.8% and serotype B 2.3%) was classified as being of H. influenzae origin. Scanning the whole genomes of N. meningitidis serotypes A and B for the 29-bp HmUS gave us three perfect hits in each genome and a few HmUs containing only a few mismatches. All HmUs hits (perfect matches and with one mismatch) were located in regions of the N. meningitidis genome that our tool classified as being of H. influenzae origin. Both the gene regions described by Kroll et al. (1998) were correctly classified as being of H. influenzae origin in the genomes of both serotypes A and B, demonstrating that the classifier can correctly identify both the recipient and the donor in a horizontal gene transfer event (Fig 6).

Interestingly, three additional regions were classified as being of H. influenzae origin in both N. meningitidis genomes (Fig. 6, Table 1). In all three cases, one or more HmUS were also found within the genomic region. The identified regions contained a putative virulence-associated protein (NMA1725), putative

Lack-of-Knowledge Experiments Support Global Motif Patterning in Bacterial Genomes

Because the classifier does not depend on alignments, but instead uses motif frequencies, it can be trained using only a subset of the sequence. This subset is subsequently excluded when performing the classification task. This approach assumes that the intergenomic differences in motif frequency between genomes are greater than the intragenomic differences, as previous studies indicate (Karlin and Ludunga 1994; Karlin and Burge 1995; Nakashima et al. 1998; Deschavanne et al. 1999). We thus excluded regions from the different genomes when training the classifier (i.e., when recording motif frequencies) and then randomly picked genomic sequences from the excluded regions for assessing the classification accuracy (Fig. 4). We systematically increased the percentage of the genome excluded when training the classifier to find its limits (Fig. 4). Even when 90% of a genome is excluded during the training, the classifier still produces reliable results. The decrease in classification accuracy could, to some extent, be compensated for by increasing the sequence sample length (Fig. 4).

Classification of Closely Related Microorganisms

We investigated whether the classifier was able to correctly discriminate different strains of the same species, exemplified by H. pylori strains 26695 and J99, N. meningitidis (serotype B strain MC58 and serotype A strain Z2491), Pyrococcus abyssi and horkoshi OT3), and Chlamydia trachomatis (strain Nigg and Serovar D [D/UW-3/Cx]). Each new classifier trained had to correctly discriminate between two different strains of the same species with highly similar motif frequency profiles. The results are presented in Figure 5. For both N. meningitidis and H. Pylori, roughly 200 nucleotides were needed for accuracy.
restriction enzyme (NMA1591), putative methyltransferase (NMA1590), and a conserved hypothetical protein (NB1979). For the latter three genes, BLASTX searches identified striking homologs in Haemophilus proteins (Table 1). Those genes are likely to represent previously undetected instances of horizontal gene transfer from H. influenzae to N. meningitis.

DISCUSSION
We investigated the possibility of classifying genomic sequences based on motif frequency distributions. The classifier presented needs a sample of only 400 nucleotides to correctly classify its origin from 25 totally sequenced bacteria with >85% accuracy. This demonstrates that genome characteristics are captured in the frequencies of overlapping motifs in very short sequences. The lack-of-knowledge experiments demonstrate the feasibility of using the classifier on partial genome sequences. The classifier produced the best results when representing the genomes with eight- or nine-nucleotide motifs, although the optimum motif length is likely to depend on the amount of genomic data available. Longer genomic sequences permit a more specific motif representation, particularly if the motif length is increased. The classifier is able to generalize the genomic motif distribution from a sampled region of the genome to other regions, a functional consequence of the observation that overall variation in motif frequency within a genome is lower than the variation between genomes of different species.

Motif frequency classification does not depend on alignment methods (BLAST, Smith-Waterman) because it is position-independent (“scrambled”). It is therefore computationally inexpensive. Because the bacterial genome sequences are stored in the form of a motif frequency table, the original sequence entry is not required for comparison to the target sequence, in contrast to optimal alignment methods. The genome representation does not grow with more genomic sequences, only with new species identified (i.e., new classes). The genomes are represented as motif frequency vectors with a set dimensionality, which enables further preprocessing (vector transformations) to find better genome representation and possible improvements of classification accuracy.

In the present configuration, the classifier was used for identifying horizontal gene transfer events in whole genome sequences. The classifier was able to correctly identify both the donor and recipient strains in known horizontal gene transfer events from H. influenzae to N. meningitis, in contrast to most methods that only detect genes with abnormal sequence composition without predicting a likely donor. Using the classifier, we found three new potential examples of horizontal gene transfer from H. influenzae into N. meningitis. Finding both HmUS in the proposed regions as well as highly homologous genes in the H. influenzae genome supported the classifier results.

Surprisingly, short sequences with only 60 base pairs were correctly classified in 46 of 100 cases. Because of the remarkable resolution of the classifier, it is intriguing to speculate whether it could be applied to diagnostics of microbial diseases. New techniques for high-throughput sequencing of short genomic sequences have been developed (Ronaghi et al. 1996). The classifier could possibly be used to complement existing diagnostic tools in conjunction with these new sequencing techniques.
for each motif, Mj, in each genome were calculated for each lengths, m, of one to nine nucleotides, and frequency tables classifiers that only discriminate between different strains of the (resulting in 25 different classes). We then designed new classifiers that only discriminate between different strains of the same species.

**METHODS**

**Data**

The complete genomic sequence of 28 archaea and eu- bacterium organisms, with genome sizes ranging from 580 kb for *Mycoplasma genitalium* to 4.639 kb for *Escherichia coli*, were obtained from GenBank and TIGR at 05/00. The genomes were “scanned” for overlapping motif occurrence using motif lengths, m, of one to nine nucleotides, and frequency tables for each motif, Mj, in each genome were calculated for each motif length. For example, when using a motif length of nine nucleotides, 4^9 (262,144) possible unique motifs (“words”) exist. Species with multiple strains sequenced (see also Fig. 1). The naive Bayesian classifier assumes each motif to be independent of the other motifs, which is clearly false. In fact, in most real-world tasks the independence assumptions is violated, but the method has still proven successful (Domingos and Pazzani 1997).

$$P(S | G_i) = \prod_{j=1}^{k} P(M_j | G_i)$$

(2)

The classifier assigns genomic origin to sequences by taking the maximum \(P(G_i | S)\) value, calculated for all available genomes. The probability of finding sequence \(S\) in a genome \(G_i\) can be expressed as the product of the \(N\) (eq. 1) probabilities to find each motif \(M_j\) in genome \(G_i\) (eq. 2). The naive Bayesian classifier uses principal component analysis as a useful strategy to further analyze and visualize the differences in motif frequency distributions between bacterial genomes.

Finally, it should be stressed that although further improvements may be necessary for different applications, the methodology is general and could easily be applied to other classification tasks on biological sequences.

**ACKNOWLEDGMENTS**

We thank Per Liden and Mikael Huss at Virtual Genetics Laboratory for helpful comments on the manuscript. This work was supported in part by the Swedish Cancer Society and the Swedish Technical Research Council (1FR).
The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES


Nakashima, H., Ota, M., Nishikawa, K., and Ooi, T. 1998. Genes from nine genomes are separated into their organisms in the dinucleotide composition space. DNA Res. 5: 251–259.


Received February 22, 2001; accepted in revised form May 25, 2001.

Table 1. List of Putative Horizontally Transferred Genes

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>ORF</th>
<th>Classification (kb)</th>
<th>BLAST E-value</th>
<th>BLASTX E-value</th>
<th>HmUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SodC</td>
<td></td>
<td>1/1</td>
<td>0.00; 87%</td>
<td>0.00; 88%</td>
<td></td>
</tr>
<tr>
<td>bio gene cluster - bioC</td>
<td></td>
<td>0.75/1.5</td>
<td>0.00; 87%</td>
<td>0.00; 88%</td>
<td></td>
</tr>
<tr>
<td>Conserved hypothetical protein</td>
<td></td>
<td>1.25/1.25</td>
<td>0.00; 92%</td>
<td>0.00; 92%</td>
<td></td>
</tr>
<tr>
<td>Type III restriction enzyme</td>
<td></td>
<td>2.75/3.5</td>
<td>0.00; 92%</td>
<td>0.00; 92%</td>
<td>28/29</td>
</tr>
<tr>
<td>Type III methyltransferase</td>
<td></td>
<td>0.5/0.5</td>
<td>0.00; 92%</td>
<td>0.00; 92%</td>
<td>28/29</td>
</tr>
<tr>
<td>Virulence associated protein</td>
<td></td>
<td>1.75/1.25</td>
<td>2e-45; 95%</td>
<td></td>
<td>2 * 29/29</td>
</tr>
</tbody>
</table>

Genes identified as horizontally transferred from Haemophilus influenzae to Neisseria meningitidis. ORF and Bayesian classification data are provided for both N. meningitidis serotype A and, in parentheses for serotype B. Classification column indicates block length identified as being of H. influenzae origin. The genomic regions identified in N. meningitidis were compared for homology using BLAST toward H. influenzae and the homologous sequences reported with accession number, e-value and percent identical nucleotides. The best hit after searching genomic regions using BLASTX with accession number and e-value. HmUS indicates the number of perfect matching nucleotides between the N. meningitidis region and the consensus. The two N. meningitidis genomes gave almost identical results (data not shown).

1Putative horizontal gene transfer events identified by another study (Kroll et al. 1998).