BIOINFORMATICS

Vol. 22 no. 14 2006, pages e281–e289
doi:10.1093/bioinformatics/btl247

Finding novel genes in bacterial communities isolated from the environment

Lutz Krause[1,*], Naryttza N. Diaz[1], Daniela Bartels[1], Robert A. Edwards[2,3,4], Alfred Pühler[6], Forest Rohwer[3,4], Folker Meyer[1] and Jens Stoye[5]

[1]Bielefeld University, Center for Biotechnology (CeBiTec) D-33594 Bielefeld, Germany, [2]Fellowship for Interpretation of Genomes, Burr Ridge IL, [3]Department of Biology, San Diego State University, San Diego, CA, [4]Center for Microbial Sciences, San Diego, CA, [5]Universität Bielefeld, Technische Fakultät D-33594 Bielefeld, Germany and [6]Universität Bielefeld, Lehrstuhl für Genetik, Fakultät für Biologie D-33594 Bielefeld, Germany

ABSTRACT

Motivation: Novel sequencing techniques can give access to organisms that are difficult to cultivate using conventional methods. When applied to environmental samples, the data generated has some drawbacks, e.g. short length of assembled contigs, in-frame stop codons and frame shifts. Unfortunately, current gene finders cannot circumvent these difficulties. At the same time, the automated prediction of genes is a prerequisite for the increasing amount of genomic sequences to ensure progress in metagenomics.

Results: We introduce a novel gene finding algorithm that incorporates features overcoming the short length of the assembled contigs from environmental data, in-frame stop codons as well as frame shifts contained in bacterial sequences. The results show that by searching for sequence similarities in an environmental sample our algorithm is capable of detecting a high fraction of its gene content, depending on the species composition and the overall size of the sample. The method is valuable for hunting novel unknown genes that may be specific for the habitat where the sample is taken. Finally, we show that our algorithm can even exploit the limited information contained in the short reads generated by 454 technology for the prediction of protein coding genes.

Availability: The program is freely available upon request.

Contact: Lutz.Krause@CeBiTec.Uni-Bielefeld.DE

1 INTRODUCTION

Novel sequencing methods have recently revolutionized the field of genome research. The sequencing of samples isolated directly from the environment allows access to organisms that can not be cultivated in the laboratory (Breitbart *et al*. (2002), Tyson *et al*. (2004), Venter *et al*. (2004)). Additionally, the massively parallel pyrosequencing system which was recently developed by 454 Life Science, Inc, has dramatically dropped the time and cost constraints of DNA sequencing (Margulies *et al*. (2005)). The application of 454 technology provides larger amounts of sequences at a lower cost compared to traditional DNA sequencing methods. These sequences are of great value for the identification of novel genes that can not be found in organisms cultured with traditional methods. The importance of such approaches is stressed by the fact that only a fraction of the living organism found in natural environments can be cultured by conventional methods (Tringe and Rubin (2005)).

The isolation and sequencing of DNA derived from diverse and mixed microbial communities is known as metagenomics, environmental genomics or ecogenomics. Although still in its infancy, this rapidly developing field has provided striking insights into the ecology and evolution of natural occurring microbial communities. Fields such as health and biotechnology have already benefited from metagenomics (Lombardot *et al*. (2006), Furrie (2006), Schloss and Handelsman (2003), Edwards and Rohwer *et al*. (2005), Edwards *et al*. (2006)).

Gene finding in environmental samples

Two different approaches are applied for predicting protein coding genes in bacterial genomes; intrinsic and extrinsic methods. Intrinsic methods (e.g. GLIMMER Delcher *et al*. (1999), GENEMARK Besemer and Borodovsky (1999)) analyze sequence properties of genomes to discriminate between coding sequences (CDS) and non-coding ORFs (NORFs). These methods exploit the different compositional properties of coding and non-coding sequences, which are mainly caused by a bias on codon usage in the CDS to optimize the translation efficiency (Gouy and Gautier (1982)).

In contrast, extrinsic methods (e.g. CRITICA Badger and Olsen (1999), ORPHEUS Frishman *et al*. (1998)) predict genes by searching for stretches of DNA that were conserved during evolution. The success of extrinsic methods can be explained by the fact that during evolution most of the new genes are formed by duplication, rearrangement and mutation events of existing genes (Chothia *et al*. (2003)).

The prediction of protein coding genes in environmental samples is problematic for several reasons. One is the low sequence quality of the assembled contigs which may lead to frame shifts and in-frame stop codons in the CDSs contained therein. Another problem is that assembled contigs may be too short to reveal the genome specific sequence properties, which are crucial in the application of intrinsic gene prediction methods. These reasons limit their application to environmental samples. Currently, the majority of the CDSs in environmental samples are identified based on a BLAST search against databases of known proteins.

*To whom correspondence should be addressed.

Species that are abundant in natural environments will also be over-represented in the samples. These species do not represent a problem while assembling them, and large stretches of their genomes can be obtained. But, the under-represented species constitute a challenge since for those only short contigs with low coverage are obtained. One problem related to the low coverage is that these contigs are even more prone to contain in-frame stop codons or frame shifts. Therefore, applying existing gene finders to environmental samples is fraught with difficulties because they were not designed to cope with this type of errors and short contigs.

Strategy

The main idea for the novel gene prediction method presented in this work is to search for stretches of DNA that are conserved within the environmental sample. Here, the algorithm does not rely on a pairwise sequence comparison, but instead it combines information from all BLAST hits at the same time. Conserved coding sequences are discriminated from conserved non-coding regions based on their synonymous substitution rate.

In functional proteins, the coding genes show a much higher number of synonymous substitutions than in non-coding sequences. The rate of synonymous to non-synonymous substitutions (k_S/k_A) reflects the interchange of positive selection and neutral evolution. Therefore, investigating the number of synonymous and non-synonymous substitutions can supply valuable information on whether or not a sequence stretch is under constraint for functional selection. This information can be used for the identification of genes in bacterial and eukaryotic genomes (Badger and Olsen (1999), Nekrutenko *et al*. (2003a), Nekrutenko *et al*. (2003b) and Moore and Lake (2003)).

For the prediction of protein coding sequences contained in a contig from an environmental sample, first a BLAST search against a nucleotide database is conducted. For this in principle any nucleotide database can be used, e.g databases containing complete genomes, metagenomes or known genes. To search for novel habitat specific genes a BLAST search against a database that exclusively contains all sequences from that sample can be employed. Subsequently, the algorithm needs to discriminate if the BLAST hits match conserved coding sequences, conserved non-coding regions, or shadows of CDSs in another reading frame. Additionally, a CDS may be embedded in long BLAST hits. For this case, the gene boundaries need to be identified. Given all BLAST hits for a contig, the algorithm will find the best path through all hits at the same time. In order to accomplish this task, several different features are taken into account: (a) the synonymous substitution rate at each position in the contig, (b) the positions of stop codons in the contig and (c) the position of stop codons in matching database sequences. Additionally, the end of BLAST hits are considered as possible indications for the boundaries of coding regions.

In the gene prediction process the algorithm will avoid in-frame stop codons, but otherwise will favor regions with a high synonymous substitution rate. The outcome of the BLAST hits are used to assign six scores to each nucleotide, one for each of the six possible reading frames, reflecting the nucleotides coding potential in this reading frame. Scores are assigned by counting the number of synonymous and non-synonymous substitutions at each position for each of the six reading frames. As a result, a scoring matrix with scores for each nucleotide in the contig is obtained. Based on these scores, a dynamic programming method is applied to find the optimal path through the matrix that maximizes the overall score (the sum of all scores on the path). The usage of a combined score for all BLAST hits should result in a superior performance compared to methods that rely on simple pairwise sequence alignments. The advantage should be particularly profound when a database of low quality with short contigs and many frame shifts is used for the BLAST based search for conserved sequences.

2 METHODS

The gene prediction algorithm

The algorithm can be divided into four phases: (1) a BLAST based search for conserved sequences (2) the calculation of combined scores (3) the prediction of coding sequences by dynamic programming and (4) the postprocessing.

Phase 1: Blast based search for conserved sequences During the first phase of the algorithm a BLAST search against a nucleotide database is conducted. Hereby, the contig as well as all sequences in the database are translated into all six reading frames (if the database contains known genes only the contig will be translated into all six reading frames). As the BLAST search is conducted on the amino acid level, each obtained hit is associated with a specific reading frame in the contig. The BLAST hits obtained are filtered, hits with $k_S/k_A < 1$ are excluded from the subsequent analysis as these do not indicate the presence of a coding sequence.

Phase 2: Calculation of combined scores In the second phase of the algorithm, the remaining hits are used to assess the coding potential of each nucleotide in the contig. Given a contig c of length n, $c[i]$ denotes the nucleotide at position i of that contig ($1 \leq i \leq n$). A nucleotide $c[i]$ could be coding in one of the six reading frames $k \in \{-3, -2, -1, +1, +2, +3\}$, or non-coding, denoted by $k = 0$. For each position i and for each reading frame k, the number of synonymous and non-synonymous substitutions at position i are counted (Figure 1). This is done by comparing the nucleotide sequence of the contig to the nucleotide sequence of all BLAST hits in this reading frame. The number of synonymous and non-synonymous substitutions are used to score that $c[i]$ is coding in reading frame k. Synonymous substitutions contribute with a positive score, non-synonymous substitutions with a negative score. Additionally, the correct ends of the coding sequences need to be determined. Therefore, stop codons in the contig are penalized with a negative score in the according frame. For a given BLAST hit both the contig and the matching database sequence of the BLAST hit may contain stop codons. To discriminate between real stop codons and stop codons introduced by sequencing errors, additionally negative scores are applied for: (a) all stop codons in the database sequences of the BLAST hits, (b) for ends of BLAST hits, as these also may indicate the boundaries of genes (Figure 1). Subsequently, each score obtained is normalized by the number of hits that contribute to that score. Using this strategy for all BLAST hits in reading frame k, a single combined score that reflects the coding potential of the contig at position i in this reading frame is derived. Additionally, for $k = 0$ a score of zero is assigned to each position i of the contig. As a result, a scoring matrix s_{ik} is derived which provides a position specific score that the contig is coding in one of the six reading frames or non-coding (Figure 1).

Phase 3: Prediction of coding sequences Coding sequences are predicted in the third phase. To assign one of the six reading frames k (or $k = 0$ for non-coding) to each position of the contig, the algorithm searches for the path in the scoring matrix s_{ik} that maximizes the sum of all scores on the path. According to the optimal path, each position i of the contig is subsequently labeled with the frame k it passes through at this position. Depending on their reading frame, genes may only start or

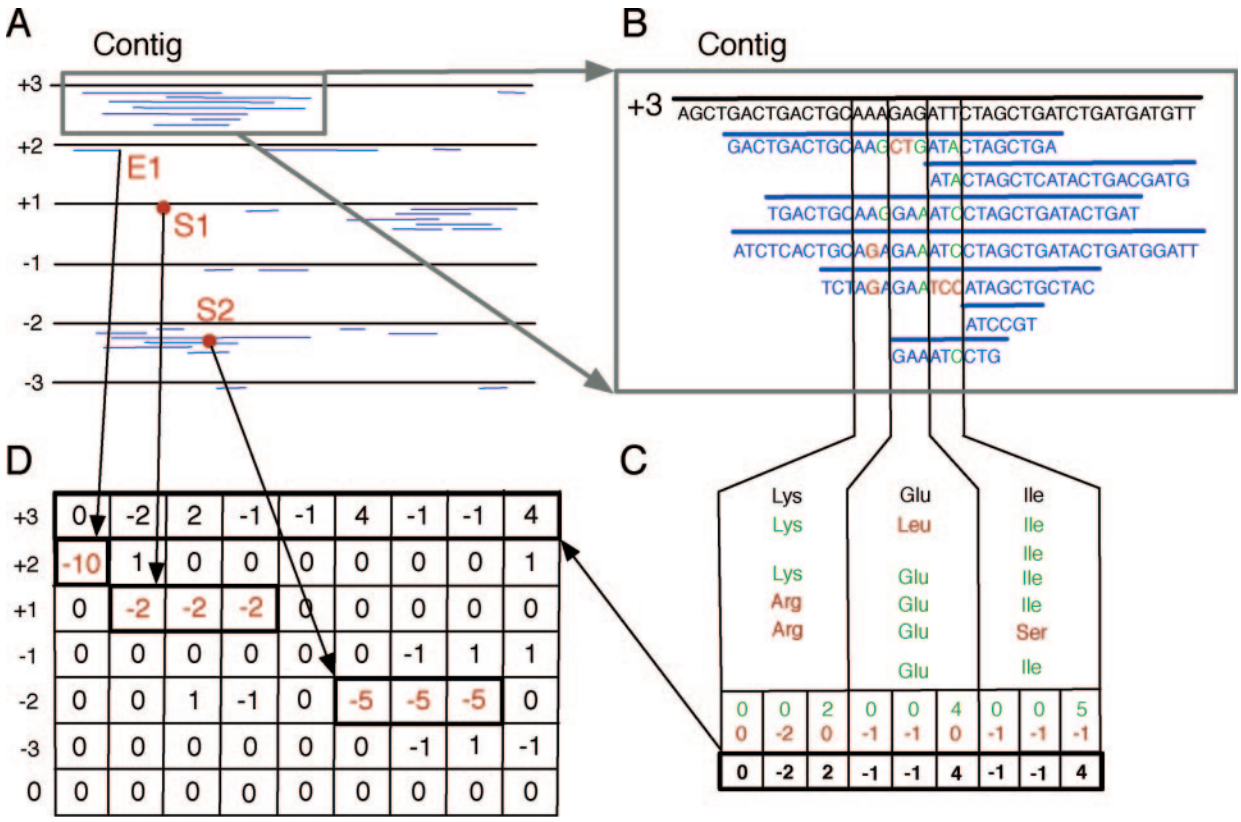

Fig. 1. Calculating combined scores. All scores are depicted without normalization. **A**) all six reading frames of a contig are shown (the continuous lines). BLAST hits matching the respective reading frames are displayed as blue bars below the reading frame. **B**) The nucleotide sequence of each reading frame of the contig is compared with all database sequences matching this reading frame. The number of synonymous and non-synonymous substitutions at each position is used as a score that the contig at this position is coding in the respective reading frame. **C**) The number of synonymous substitutions at each position are used as a positive score. The number of non-synonymous substitutions at each position contribute with a negative score. **D**) The calculated scores for each position and reading frame are stored in a matrix. For $k = 0$ a score of zero is assigned to each position i of the contig. Penalties are additionally added to the respective position and reading frame for stop codons in the contig (S1), in the matching database sequence (S2) as well as for the end of BLAST hits (E1).

stop at certain positions. Therefore, a valid path may not jump arbitrarily between frames, but instead underlies certain restrictions. To be precise, the set $V(i, k)$ of all valid precursors of a frame k at position i is defined as:

$$V(i,k) = \begin{cases} \{j, 0, -j\} & \text{if} \quad k = 0 \\ \{k, 0, -k\} & \text{if} \quad |k| = j \\ \{k\} & \text{otherwise.} \end{cases}$$

where $j = (i - 1) \bmod 3 + 1$. Figure 2 depicts the scoring matrix of combined scores and the calculation of the optimal path. This figure also introduces several terms used in the following. The optimal valid path for a scoring matrix s_{ik} can be calculated using dynamic programming by the following recursion:

$$f_i(k) = \max_{k' \in V(i,k)} \begin{cases} f_{i-1}(k') + s_{ik} + 2q & \text{if } k < 0 \text{ and } k' > 0 \\ f_{i-1}(k') + s_{ik} + q & \text{if } k \neq 0 \text{ and } k' \neq k \\ f_{i-1}(k') + s_{ik} & \text{otherwise} \end{cases}$$

where q is a negative score that is added to leave a gene on the forward strand or to enter a gene on the reverse strand ($2q$ are added if a gene on the forward strand is left and a gene on the reverse strand is entered at the same time). Thus, q is added for each 5′ end of a gene on a path. The penalty q was introduced to predict genes only in areas with sufficient coding evidence. The calculated value $f_i(k)$ is the maximal score of all paths that enter s at position 1 and pass through k at position i.

Phase 4: Postprocessing During the post-processing phase, the predictions are joined and frame shifts are identified. When a BLAST search against a database of short contigs is employed, genes may be covered only partially by hits which may result in the prediction of several fragments. This is particularly profound when a BLAST search against reads provided by the 454 technology is conducted. Therefore adjacent predictions within the same reading frame are joined if (a) their distance on the contig does not exceed 400 bp and (b) the sequence of the contig that separates the predictions does not contain an in-frame stop codon.

To identify frame shifts that were introduced by sequencing errors all adjacent predictions located on the same strand but within a different reading frame are predicted as frame shifts if (a) their distance on the contig is less than 200 bp and (b) they do do not have an in-frame stop codon close to the potential frame shift. As an optional postprocessing step, our algorithm can also extend predicted CDS to the longest possible ORF available for that prediction.

Implementation

The algorithm was implemented in PERL using an object oriented approach.

Measuring the performance

To evaluate the performance of the novel gene finder predictions were compared to known annotated genes. For this purpose, two measurements

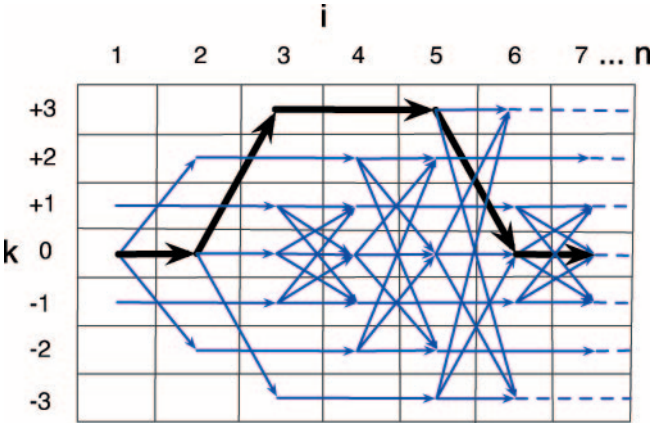

Fig. 2. Predicting coding sequences by calculating the optimal path in scoring matrix of combined scores. This figure shows the scoring matrix s_{ik} for the first seven positions of a contig. All valid paths in the scoring matrix are indicated with arrows. A gene is entered, if a path passes through frame $k \neq 0$ with the precursor frame $k' \neq k$. Accordingly, a gene is left, if a path that comes from a precursor frame $k' \neq 0$ enters a frame $k \neq k'$. The bold arrows depict an example path predicting a 3 bp long gene on the reading frame +3

are widely used: sensitivity and specificity. Sensitivity is a measure of the ability of the algorithm to predict known genes and is defined by $Sens = \frac{TP}{TP+FN}$. The specificity is a measure of the reliability of the predictions, given by the ratio $Spec = \frac{TP}{TP+FP}$. For the evaluation of the performance predictions were extended to the next 5′ stop codon. If an annotated CDS ends at that stop codon the prediction was counted as true positive (TP). Otherwise, the prediction was regarded as a false positive. All genes that are not completely embedded in the contigs are named truncated genes. These genes may appear at the end or beginning of the assembled contigs, therefore lacking the start or termination site of the gene. Truncated genes were excluded from the analysis.

Training GLIMMER on a synthetic metagenome

The prokaryotic gene finder GLIMMER version 3.01b was used to predict the genes of a synthetic metagenome (described in Materials). For the training step, all fragments of this metagenome were chained to one continuous contig. Adjacent fragments were concatenated with a linker sequence containing a stop codon in each of the six reading frames. Subsequently the GLIMMER ICM model was trained on the chained contig.

3 MATERIALS

Metagenome obtained with pyrosequencing

The performance of the algorithm was evaluated on a metagenome of a bacterial community isolated from the Solar Salterns in San Diego, CA (B. Rodriguez-Brito, R. Edwards, and F. Rohwer, Unpublished). Total community DNA was purified as described elsewhere (Edwards *et al*. (2006)) and sequenced using pyrosequencing by 454 Life Sciences, Inc, (Branford, CT). Using the 454 technology ≈60 Mb were obtained with an average read length of 100 bp. The reads were assembled using Phrap (Green (1994)). This resulted in 80,878 contigs with 16 Mb in total. In the following, this set is called *all contigs*. From this set a subset of contigs longer than 1,000 bp (2,244 contigs with 3.8 Mb in total) was selected, called hereafter *long contigs*.

Table 1. Annotated and published genomes used to create a synthetic metagenome

Organism	Accession number
Bacteria	
Alphaproteobacteria	
Candidatus pelagibacter ubique HTCC1062	NC_007205
Rhodobacter sphaeroides 2.4.1 chromosome 1	NC_007493
Gammaproteobacteria	
Shewanella oneidensis MR-1	NC_004347
Thiomicrospira crunogena XCL-2	NC_007520
Vibrio cholerae O1 biovar eltor str. N16961 chromosome 1	NC_002505
Cyanobacteria	
Prochlorococcus marinus subsp. *pastoris* str. CCMP1986	NC_005072
Synechococcus sp. WH 8120	NC_005070
Archaea	
Euryarchaeota	
Pyrococcus horikoshii OT3	NC_000961
Crenarchaeota	
Sulfolobus solfataricus P2	NC_002754

Species names and accessions numbers downloaded form the NCBI database.

The environmental sample from the Sargasso Sea

For the prediction of protein coding genes in metagenomes, the environmental sample from the Sargasso Sea (Venter *et al*. (2004)) was used as BLAST database during the search for conserved regions in the first phase of the algorithm. To save computational time, only half (≈390 Mb) of the entire Sargasso Sea sample was used.

Generating a synthetic metagenome

As a proof of concept, the algorithm was evaluated on a set of nine completely sequenced and annotated genomes (seven Bacteria and two Archaea, see Table 1). Members from the alpha, gammaproteobacteria and cyanobacteria groups were selected as they were reported to be abundant in the Sargasso Sea sample (Venter *et al*. (2004)). We also added Archaea to the evaluation set because they can be regarded as under-represented species in surface water marine environments. All genomic sequences and their respective annotations were downloaded from the NCBI Reference Sequence database (RefSeq) release 15 (Pruitt *et al*. (2005)). A synthetic metagenome with known CDSs was created by splitting the genome of each of the nine organism into fragments of length 4000 bp. A subset of *non-hypothetical* genes was created based on the annotated gene products from the public annotations. In this set all annotated genes with a gene product description of 'hypothetical protein' were excluded. Additionally, artificial sequencing errors (frame shifts and in-frame stop codons) were incorporated into all genes of the synthetic metagenome. In order to perform a systematic evaluation, all artificial sequencing errors were added to the synthetic metagenome in a controlled way. In one experiment, in-frame stop codons were added to the center of each gene of the original synthetic metagenome. In a second experiment frame shifts were incorporated to the center of all genes of the original synthetic metagenome.

4 RESULTS

Gene prediction in a synthetic metagenome using the environmental sample from the Sargasso Sea

Our algorithm can be used to identify genes contained in an environmental sample by directly searching for conserved regions within the sample. This approach may elucidate novel unknown genes present in the environmental sample which may be specific for the habitat the sample was taken from. The performance of the algorithm of predicting genes in an environmental sample by running a BLAST search against the sample itself was evaluated on the environmental sample data from the Sargasso Sea (Venter *et al*. (2004)). But, instead of drawing the contigs for which the genes are predicted directly from the Sargasso Sea sample, we used a more controlled and reliable data set. We chose several completely sequenced and annotated genomes from Bacteria groups that were also reported to be present in the species composition of the Sargasso Sea sample (Venter *et al*. (2004)). The genomes of these organisms were split into fragments of size 4000 bp, together forming a synthetic metagenome as a reliable standard of truth. Subsequently the genes of these contigs were predicted with our algorithm based on a BLAST search against the Sargasso Sea sample. To accurately evaluate the prediction performance that can be expected for a 'real' metagenome, sequences from alpha and gammaproteobacteria which are reported as over-represented in the Sargasso Sea sample, cyanobacteria which are modest abundant, as well as sequences from extremely scarce groups (two Archaea members) were included. The performance of the algorithm was measured by comparing the genes predicted for the synthetic metagenome with the known genes from the public genome annotations. To additionally evaluate the performance for sequencing errors that may frequently occur in metagenomes, three validation sets were used: (1) synthetic metagenome without artificial sequencing errors, (2) synthetic metagenome with in-frame stop codons and (3) synthetic metagenome with frame shifts.

Experiment 1: Gene prediction in a synthetic metagenome without artificial sequence errors The sensitivity and specificity reached by the algorithm for each organism contained in the synthetic metagenome is shown in Table 2. The results show that the sensitivity of the method strongly depends on the abundance of the different groups of Bacteria in the sample. While for the more abundant alpha, gammaproteobacteria and cyanobacteria an average sensitivity of 79% for all genes and 89% for the subset of non-hypothetical genes is achieved, for the Archaea the sensitivity is strongly reduced. The lower sensitivity for the Archaea was expected because this group is very rare in surface water marine environments and hence extremely scarce in the environmental sample from the Sargasso Sea. For the two cyanobacteria contained in the synthetic metagenome even a sensitivity of more than 94% is achieved for the non-hypothetical genes. In contrast, with a specificity between 88% and 99% the algorithm is highly specific for all groups. On average the specificity is 95%. The considerably lower overall sensitivity ($Sens_{all}$) when compared to the sensitivity for the subset of non-hypothetical genes ($Sens_{nh}$) can be explained by the fact that most of the genes labeled as 'hypothetical protein' in the public annotations were originally predicted with intrinsic methods. Many of these genes are either orphans (genes without sequence similarity to any known gene) or in fact non-coding and hence wrong annotations.

Table 2. Performance for a synthetic metagenome evaluated on the Sargasso Sea environmental sample

Organism	$Sens_{all}$	$Sens_{nh}$	Specificity
Bacteria			
Alphaproteobacteria			
C. pelagibacter	91.07	93.76	97.63
R. sphaeroides	62.01	77.62	97.02
Gammaproteobacteria			
S. oneidensis	85.36	95.12	90.44
T. crunogena	65.38	79.33	97.54
V. cholerae	69.66	87.66	93.88
Cyanobacteria			
P. marinus	93.29	94.42	89.75
Synechococcus sp.	82.99	94.13	87.83
Archaea			
Euryarchaeota			
P. horikoshii	29.99	66.40	97.77
Crenarchaeota			
S. solfataricus	26.69	43.44	98.89
Average	67.38	81.32	94.53

$Sens_{all}$ refers to the sensitivity calculated over all genes contained in the synthetic metagenome. $Sens_{nh}$ is the sensitivity calculated over all non-hypothetical genes. The entire Bacteria group represents the most common organisms in the Sargasso Sea sample. While the Archaea is the extremely scarce set for surface water marine environment.

Experiment 2: Gene prediction in a synthetic metagenome with artificial in-frame stop codons In the second experiment the performance of the algorithm was evaluated on genes containing in-frame stop codons. Therefore, an in-frame stop codon was added to the center of each annotated gene in the synthetic metagenome. In addition to the sensitivity and specificity, the percentage of true positives (TP) that span the artificially added stop codons was measured. In comparison to the synthetic metagenome without artificial sequence errors, for the genes with in-frame stop codons only a slight reduction in sensitivity and specificity was registered. The sensitivity is reduced by 1.7% for all genes and 1.3% for the subset of non-hypothetical genes. The reduction in specificity is 0.3%. On average, for 77% of all identified genes (TP) the prediction also spans the added in-frame stop codon (Table 3) and therefore correctly recognizes the stop codon as sequencing error. Strikingly, for the synthetic metagenome without artificial sequence errors only 4 predictions wrongly span a 'real' stop codon terminating the translation. These results demonstrate that the algorithm is quite robust for the task of identifying functional genes containing in-frame stop codons, generated by sequencing errors. These results also reveal the strength of our method to incorporate several features to determine the boundaries of coding sequence and to discriminate between 'real' stop codons and those introduced by sequencing errors.

Experiment 3: Gene prediction in a synthetic metagenome with artificial frame shifts In the third experiment the performance of the novel algorithm to predict frame shifts introduced by sequencing errors was evaluated. Therefore, an artificial frame shift was

Table 3. Performance for a synthetic metagenome with artificial in-frame stop codons

Organism	$Sens_{all}$	$Sens_{nh}$	Spec	SC predicted
Bacteria				
Alphaproteobacteria				
C. pelagibacter	88.87	92.20	97.58	71.71
R. sphaeroides	61.62	77.18	97.33	73.10
Gammaproteobacteria				
S. oneidensis	83.90	94.24	89.58	82.15
T. crunogena	64.95	79.01	97.88	80.23
V. cholerae	68.89	86.94	94.00	79.52
Cyanobacteria				
P. marinus	88.97	90.18	88.51	74.71
Synechococcus sp.	78.74	92.69	85.65	70.75
Archaea				
Euryarchaeota				
P. horikoshii	29.27	65.20	98.69	80.97
Crenarchaeota				
S. solfataricus	25.96	42.71	99.02	81.41
Average	65.69	80.04	94.25	77.17

$Sens_{all}$ is the sensitivity calculated over all genes contained in the synthetic metagenome. $Sens_{nh}$ is the sensitivity calculated over the subset of all non-hypothetical genes. SC predicted: percentage of true positives (TP) that correctly span in-frame stop codons.

Table 4. Performance for a synthetic metagenome with artificial frame shifts

Organism	$Sens_{all}$	$Sens_{nh}$	Spec	Percentage of TP predictions correctly identified as frame shift
Bacteria				
Alphaproteobacteria				
C. pelagibacter	86.39	89.32	97.73	57.71
R. sphaeroides	56.80	72.08	98.03	92.22
Gammaproteobacteria				
S. oneidensis	81.15	90.93	93.02	68.65
T. crunogena	59.69	73.33	98.33	66.67
V. cholerae	71.54	83.05	96.19	69.11
Cyanobacteria				
P. marinus	84.65	88.88	92.62	58.77
Synechococcus sp.	72.46	90.39	91.02	79.19
Archaea				
Euryarchaeota				
P. horikoshii	26.09	54.42	96.45	54.64
Crenarchaeota				
S. solfataricus	23.23	39.41	98.80	48.51
Average	62.44	75.76	95.80	66.16

$Sens_{all}$ is the sensitivity calculated over all genes contained in the contigs. $Sens_{n\ h}$ is the sensitivity calculated over the subset of non-hypothetical genes

added to each of the genes of the synthetic metagenome. For this data set, those predictions that do not match a fragment of an annotated gene where counted as false positives (FP). For those annotated genes of which at least one of its fragments is identified were counted as true positives (TP). Compared to the synthetic metagenome with no artificial mutations, the sensitivity and specificity is again only slightly reduced (Table 4). For this data set, 66% of the identified genes (TP) were also correctly predicted to have a frame shift. Noteworthy, for the synthetic metagenome without artificial errors only 357 frame shifts out of 11,686 true positive predictions were registered. This finding shows the high reliability of the method to predict frame shifts. As for the above experiments, the specificity values obtained by each genome are high, the average specificity value is 95%.

Gene identification in environmental samples obtained by 454 technology

At present, the main drawback of the recently developed high throughput parallel pyrosequencing is the short length of the reads obtained (≈100 bp on average). This is particularly undesirable when dealing with environmental data sets, since the sample is a large mixture of different species. To verify whether our algorithm is still able to identify genes in metagenomes obtained with the 454 technology, we assembled the 454 reads from the Solar Salterns sample into contigs and predicted the genes for the subset of all *long contigs*. For this verification we performed two experiments: First, a BLAST search against a database made from the set of *all contigs* from the Solar Salterns sample was conducted. Second, a direct BLAST search against a database of all 454 reads without prior assembly was employed. To validate the outcome from both experiments the respective predictions (extended to the longest possible ORF for that prediction) were compared with known proteins from the KEGG database (Ogata *et al*. (1999)) using BLAST.

Table 5. KEGG supported predictions. Number of predicted genes for a metagenome sequenced with 454 technology that have hit in the KEGG database.

Database	Number of predictions	Number of predictions with E-value up to 10^{-50}	10^{-20}	10^{-10}	10^{-5}
KEGG					
Contigs	3219	467	1544	2451	2858
Reads	3496	556	1699	2585	3044

Assembled contigs and 454 reads without prior assembly were used for BLAST search.

For both experiments, a high fraction of the predicted genes has significant BLAST hits against known proteins from the KEGG database. Remarkably, the number of predicted genes for the reads without assembly does not differ much when compared to the contigs (see Table 5). It should be pointed out that when looking at the BLAST hits against the KEGG database it seems that many of the predicted genes are fragmented due to internal frame shifts. Therefore during the BLAST search against the KEGG database, weaker E-values are obtained for these fragments. The predicted genes that do not match any known protein in the KEGG database constitute an interesting set for further studies as they could be either of false predictions, known genes with no or only a weak sequence similarity to the genes contained in KEGG, or more interestingly novel unknown genes. These results for the Solar Salterns sample demonstrate that the novel algorithm is well

Table 6. Performance for a synthetic metagenome evaluated on sequences obtained by pyrosequencing

Organism	$Sens_r$	$Sens_c$	$Sens_{nhr}$	$Sens_{nhc}$	$Spec_r$	$Spec_c$
Bacteria						
Alphaproteobacteria						
C. pelagibacter	38.96	28.02	43.92	31.66	85.29	91.54
R. sphaeroides	27.74	19.13	38.03	26.81	70.67	87.18
Gammaproteobacteria						
S. oneidensis	25.37	16.19	38.23	25.12	80.86	88.21
T. crunogena	36.21	23.49	46.31	30.29	86.42	93.43
V. cholerae	29.71	18.84	43.19	28.69	82.22	90.70
Cyanobacteria						
P. marinus	30.40	20.94	43.08	29.91	89.67	91.53
Synechococcus sp.	23.24	17.01	43.67	31.82	77.14	87.73
Archaea						
Euryarchaeota						
P. horikoshii	33.61	31.87	66.80	62.60	89.64	94.67
Crenarchaeota						
S. solfataricus	26.35	25.15	41.97	41.32	90.34	95.46
Average	30.18	22.29	45.02	34.25	83.58	91.16

$Sens_r$ and $Sens_c$ is the sensitivity for the synthetic metagenome when blasting against all 454 reads or against all assembled contigs. $Sens_{nhr}$ and $Sens_{nhc}$ is the sensitivity calculated for the subset of non-hypothetical genes of the synthetic metagenome when a BLAST search is done against the 454 reads and assembled contigs, respectively.

suited to predict genes in 'real' metagenomes, even if these samples are sequenced using the 454 technology.

Gene prediction in synthetic metagenomes using contigs and reads derived by pyrosequencing

We further evaluated the performance of the new gene finding algorithm for sequences obtained with the 454 technology (see Table 6), taking the synthetic metagenome dataset as a controlled standard of truth. The genes were predicted for the synthetic metagenomes dataset by employing a BLAST search against two different databases: one containing all assembled contigs from the Solar Salterns sample, and another containing all unassembled reads from the same sample.

In respect to the small size of the database used in the BLAST search (16 Mb for the assembled contigs and 60 Mb for the reads without prior assembly) the sensitivity obtained is very good. The highest sensitivity is reached for *Pyrococcus horikoshii*, 67% and 63% (for the subset of all non-hypothetical genes) calculated for the reads without assembly and the assembled contigs, respectively. Interestingly, these findings indicate that in contrast to the sample from the Sargasso Sea, the Archaea group is more abundant in the sample from the Solar Salterns. A second interesting observation is the good performance when running BLAST against the 454 reads without assembly, despite the fact that the average length of the reads is 100 bp. A specificity of 84% is achieved on average. Moreover, when compared to the assembled contigs the sensitivity is increased by ≈11%. In particular, these results for the short 454 reads reveal one of the strengths of our method: to consider all BLAST hits at the same time by calculating the optimal path through the matrix of combined scores instead of analyzing simple pairwise BLAST hits. This strategy allows us to identify genes that get only several short hits, even if all of the single hits are not significant.

Yet, determining the correct boundaries of the CDS when running BLAST against a small database of 454 reads is difficult, many genes are only partially covered by hits. As an optional postprocessing step our algorithm therefore can automatically extend predictions to the longest possible ORF.

Time efficiency of the novel algorithm

The running time of the novel algorithm highly depends on the size of the BLAST database since most of the running time is consumed during the BLAST based search for conserved regions, for the parsing of BLAST results as well as for the calculation of combined scores. For the evaluation presented in this survey all runs of the algorithm were executed on a compute cluster located at the Center of Biotechnology (CeBiTec), Bielefeld University. The cluster is composed of 128 Sun Fire V20z nodes. Each node has two 1.8 GHz AMD Opteron 244 CPUs and 2 Gb of RAM. The overall running time was 1 hour and 50 minutes for predicting the genes of the synthetic metagenome (≈24 Mb) when a BLAST search against half of the Sargasso Sea sample (≈390 Mb) was employed. The running time in average is 28s for the BLAST search, 17s for parsing the BLAST results and calculating the combined scores and 1s for predicting coding sequence by dynamic programming and postprocessing for a 4 Kb fragment when run on a single node using one CPU.

GLIMMER performance on synthetic metagenome

Most of the contemporary gene finding methods model frequencies of short oligonucleotides to discriminate between coding and non-coding sequences (e.g. by using a Markov chain or a Hidden Markov model). Before these methods can be used for gene prediction, usually as a first step the model needs to be trained to learn the organism specific sequence composition of the genome under study. As most of these methods model average sequence properties they may fail to adequately learn the oligonucleotide frequencies of diverse microbial assemblages. Pitfalls of existing gene finding technologies were examined by employing the state-of-the art microbial gene finder GLIMMER as an example. GLIMMER was trained on the synthetic metagenome itself as described in the Methods section. Subsequently, the trained GLIMMER was applied on each fragment. Although GLIMMER is very accurate for complete genomes (http://www.cbcb.umd.edu/software/glimmer/) the accuracy for the synthetic metagenome is strongly reduced (Table 7). Table 7 also points to one substantial problem that may affect intrinsic methods when applied to environmental data: the diverse compositional biases of different organisms contained in the sample. Another problem may be the unequal abundance of species, as overrepresented species have a stronger influence during training which may result in an unbalanced model. Also the synthetic metagenome on which GLIMMER was trained is unbalanced as it contains fragments from seven genomes with a low GC content and from two genomes with a high GC content (GC > 55%). The average GC content is ≈47%. The prediction accuracy of GLIMMER for the synthetic metagenome strongly depends on whether the fragments come from a genome with a high or low GC content. While GLIMMER has a good performance for the genomes with a low GC content, for the two genomes with a high GC content the performance is highly reduced. For the

Table 7. GLIMMER performance for a synthetic metagenome

Organism	Contig size (Mb)	GC (%)	$Sens_{all}$	$Sens_{nh}$	Spec
C. pelagibacter	1.3	30	94.34	95.54	78.39
P. marinus	1.7	31	91.80	94.87	74.04
S. solfataricus	3.0	36	91.46	93.89	72.29
P. horikoshii	1.7	42	88.28	95.60	76.92
T. crunogena	2.4	43	98.72	99.12	71.10
S. oneidensis	5.0	46	95.61	98.08	67.22
V. cholerae	3.0	48	90.68	98.48	69.52
Synechococcus sp.	2.4	59	43.80	51.60	60.41
R. sphaeroides	3.2	69	12.16	14.65	23.69
Average	2.6	47	78.53	82.42	65.95

Genomes ordered by GC content. GLIMMER was trained on the synthetic metagenome itself. $Sens_{all}$ and $Sens_{n\ h}$ is the GLIMMER sensitivity for set of all and for the subset of non-hypothetical genes. Spec: Specificity

genome with the highest GC content (*R. sphaeroides*) the accuracy is close to the one expected by a random decision drawn by a flipping a coin experiment. Owing to the diverse composition, high species richness and unequal species abundance, real metagenomes isolated from natural occurring organism assemblages possess a considerably higher complexity than the synthetic metagenome used in this study. Therefore, it is reasonable to expect that for real metagenomes the problems that affect intrinsic methods should be even more profound.

5 DISCUSSION

In this paper we presented a novel algorithm that was designed to predict genes in environmental samples. The algorithm is robust for the most common problems encountered when predicting genes in these data sets: short length of the assembled contigs and a low sequence quality.

Although, the focus of the algorithm is directed on the detection of novel genes, our algorithm can also be used to identify known genes in environmental samples: instead of searching against a database containing all fragments from the environmental sample a direct search against a database containing the sequences of known genes can be conducted.

Our results show that for large samples like the Sargasso Sea, a high fraction of the gene content can be identified based on the search for sequence conservation within the sample.

The results further demonstrate that even the short reads obtained by pyrosequencing can be used to identify protein coding genes. Therefore, environmental samples sequenced with the 454 technology may be a valuable resource to identify unknown (habitat-specific) genes. To search for novel genes our algorithm requires that at least a fraction of reads is assembled into contigs. Subsequently the complete database of reads can be used to predict the genes of these contigs. Our results therefore suggest the following strategy to identify novel (habitat-specific) genes in environmental samples: to sequence part of the sample with conventional methods to obtain longer fragments that can be assembled into contigs and additionally to sequence large amounts of data at low cost with the 454 technology to increase the size of the database that can be used to search for conserved sequences.

As our method relies on sequence similarities for the prediction of protein coding genes when running BLAST against the sample itself, the method strongly depends on the size and species composition of the sample. The sensitivity of the algorithm may be improved by incorporating general sequence properties of coding sequences or proteins.

ACKNOWLEDGEMENTS

LK was supported by the DFG Graduiertenkolleg 635 Bioinformatik. RAE and FR were supported by a grant NSF DEB-BE 04-21955 from the NSF Biocomplexity program. We thank Beltran Rodriguez-Brito for generating the environmental data. NND was supported by the Deutscher Akademischer Austausch Dienst. Thanks to the anonymous reviewers for valuable comments and helpful remarks.

REFERENCES

Badger,H. and Olsen,G.J. (1999) Article title. *Mol. Biol. Evol*, **16**, 512–524.

Besemer,J. and Borodovsky,M. (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.*, **27**, 3911–3920.

Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. *Genome Res.*, **8**, 175–185.

Breitbart,M., Salamon,P., Andresen,B., Mahaffy,J.M., Segall,A.M., Mead,D., Azam,F. and Rohwer,F. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A*, **99**, 14250–14255.

Chothia,C., Gough,J., Vogel,C. and Teichmann,S.A. (2003) Evolution of the protein repertoire. *Science*, **300**, 1701–1703.

Delcher,A.L., Harmon,D. and Kasif,S. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

Edwards,R.A. and Rohwer,F. (2005) Viral metagenomics. *Nat Rev Microbiol*, **3**, 504–510.

Edwards,R.A., Rodriguez-Brito,B., Wegley,L., Haynes,M., Breitbart,M., Peterson,D., Saar,M., Alexander,S., Alexander,E.C. and Rohwer,F. (2006) Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. *BMC Genomics*, **7**, 57.

Frishman,D., Mironov,A., Mewes,H. and Gelfand,M. (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.*, **26**, 2941–2947.

Furrie,E. (2006) A molecular revolution in the study of intestinal microflora. *Gut*, **55**, 141–143.

Gouy,M. and Gautier (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055–7074.

Green,P. (1994) Documentation for PHRAP. http://www.genome.washington.edu/UWGC/analysistools/phrap.htm.

Lombardot,T., Kottman,R., Pfeffer,H., Richter,M., Teeling,H, Quast,C. and Gloeckner,F.O. (2006) Mex.net-database resources for marine ecological genomics. *Nucleic Acids Res.*, **34**, D390–D393.

Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z., Dewell,S.B., Du,L., Fierro,J.M., Gomes,X.V., Godwin,B.C., He,W., Helgesen,S., Ho,C.H., Irzyk,G.P., Jando,S. C., Alenquer,M.L., Jarvie,T.P., Jirage,K.B., Kim,J.B., Knight,J.R., Lanza,J.R., Leamon,J.H., Lefkowitz,S.M., Lei,M., Li,J., Lohman,K.L., Lu,H., Makhijani,V.B., McDade,K.E., McKenna,M.P., Myers,E.W., Nickerson,E., Nobile,J.R., Plant,R., Puc,B.P., Ronan,M.T., Roth,G.T., Sarkis,G.J., Simons,J.F., Simpson,J.W., Srinivasan,M., Tartaro,K.R., Tomasz,A., Vogt,K.A., Volkmer,G.A., Wang,S.H., Wang,Y., Weiner,M.P., Yu,P., Begley,R.F. and Rothberg,J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Moore,J. E. and Lake,J. A. (2003) Gene structure prediction in syntenic DNA segments. *Nucleic Acids Res.*, **31**, 7271–7279.

Nekrutenko,A., Chung,W.Y. and Li,W.Y. (2003) An evolutionary approach reveals a high protein-coding capacity of the human genome. *Trends Genet*, **19**, 306–310.

Nekrutenko,A., Chung,W.Y. and Li,W.Y. (2003) ETOPE: evolutionary test of predicted exons. *Nucleic Acids Res.*, **31**, 3564–3567.

Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., **27**, 29–34.

Pruitt,K., Tatusova,T. and Maglott,R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, 501–504.

Schloss,P.D. and Handelsman,J. (2003) Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol*., **14**, 303–310.

Tringe,S. G. and Rubin,E. M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet*., **6**, 805–814.

Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V, Rubin,E.M, Rokhsar,D.S. and Banfield,J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.

Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K., Nelson,W., Fouts,D.E., Levy,S., Knap,A.H., Lomas,M.W., Nealson,K., White,O., Peterson,J., Hoffman,J., Parsons,R., Baden-Tillson,H., Pfannkoch,C., Rogers,Y-H and Hamilton,S.O. (2004) Environmental genome shotgun sequencing of the sargasso sea. *Science*, **304**, 66–74.