BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 24 2005, pages 4322–4329
doi:10.1093/bioinformatics/bti701

Genome analysis

Large-scale prokaryotic gene prediction and comparison to genome annotation

Pernille Nielsen* and Anders Krogh

Bioinformatics Centre, Institute of Molecular Biology and Physiology, University of Copenhagen,
Universitetsparken 15, 2100 Copenhagen, Denmark

Received on July 4, 2005; revised on September 13, 2005; accepted on September 30, 2005
Advance Access publication October 25, 2005

ABSTRACT

Motivation: Prokaryotic genomes are sequenced and annotated at an increasing rate. The methods of annotation vary between sequencing groups. It makes genome comparison difficult and may lead to propagation of errors when questionable assignments are adapted from one genome to another. Genome comparison either on a large or small scale would be facilitated by using a single standard for annotation, which incorporates a transparency of why an open reading frame (ORF) is considered to be a gene.

Results: A total of 143 prokaryotic genomes were scored with an updated version of the prokaryotic genefinder EasyGene. Comparison of the GenBank and RefSeq annotations with the EasyGene predictions reveals that in some genomes up to ~60% of the genes may have been annotated with a wrong start codon, especially in the GC-rich genomes. The fractional difference between annotated and predicted confirms that too many short genes are annotated in numerous organisms. Furthermore, genes might be missing in the annotation of some of the genomes. We predict 41 of 143 genomes to be over-annotated by >5%, meaning that too many ORFs are annotated as genes. We also predict that 12 of 143 genomes are under-annotated. These results are based on the difference between the number of annotated genes not found by EasyGene and the number of predicted genes that are not annotated in GenBank.

We argue that the average performance of our standardized and fully automated method is slightly better than the annotation.

Availability: The EasyGene 1.2 predictions and statistics can be accessed at http://www.binf.ku.dk/cgi-bin/easygene/search

Contact: pern@binf.ku.dk

INTRODUCTION

Hundreds of prokaryotic genomes have been sequenced. This has completely changed microbial research and facilitated large-scale comparative genomics. The annotation process often includes a lot of meticulous inspection done by experts, whose detailed biological knowledge is very valuable for this work. Unfortunately, the annotation quality of the genomes varies a lot and it makes comparisons difficult. Various sequencing groups use different tools for annotation, and the varying criteria for when to annotate an open reading frame (ORF) as a gene makes the annotation very uneven. It may in some cases lead to erroneous assignments of name and function, or to over-annotation when an assignment is adapted to other genomes. It is well known that many public databases contain such erroneous data and that the errors have a tendency to propagate (Doerks *et al*., 1998; Galperin and Koonin, 1998). This will become an even bigger problem than it is today as the rate of genome sequencing and annotation will keep increasing. A general standard for annotation would help to prevent this from happening and facilitate genome comparison.

Automated methods for prokaryotic genefinding like GLIMMER (Salzberg *et al*., 1998; Delcher *et al*., 1999; http://www.tigr.org/software/glimmer/), ORPHEUS (Frishman *et al*., 1998) and different versions of GeneMark (Lukashin and Borodovsky, 1998; Besemer and Borodovsky, 1999; Besemer *et al*., 2001) have been widely used in genome sequencing projects [see for instance (Fitz-Gibbon *et al*., 2002; Cerdeno-Tarraga *et al*., 2003; Wei *et al*., 2003; McLeod *et al*., 2004)]. GLIMMER uses interpolated Markov models whose predictions are based on oligomers of varying length and on local sequence composition (Salzberg *et al*., 1998; Delcher *et al*., 1999). ORPHEUS utilizes similarity derived genes to create codon usage and ribosome binding site statistics, which are used to predict additional genes (Frishman *et al*., 1998). GeneMark.hmm uses a hidden Markov model (HMM) framework whose initial parameters are predetermined from the statistics of annotated genes. In post-processing evaluation of ribosome binding sites is applied (Lukashin and Borodovsky, 1998). Later a heuristic approach for model derivation was implemented (Besemer and Borodovsky, 1999) and in GeneMarkS an iterative training procedure is added (Besemer *et al*., 2001).

Here we present a resource of fully automated and homogeneous annotation of protein coding genes in the publicly available genomes. It is based on an updated version of the prokaryotic genefinder EasyGene (Larsen and Krogh, 2003). The most important difference between EasyGene and other genefinders is the use of a statistical significance measure. EasyGene takes a genome sequence as input and gives a list of statistically significant genes as output. In comparison with the abovementioned genefinders, EasyGene is generally better at predicting the exact gene starts and has a very good combined sensitivity/specificity performance for both AT-rich and GC-rich genomes. Furthermore, the results of 10-fold cross-validation sensitivity indicate that EasyGene avoids over-fitting (Larsen and Krogh, 2003).

The aim is to create a homogeneous resource of gene predictions and other available information from Easygene 1.2 for all the prokaryotic genomes made public. Our objective is to present the

*To whom correspondence should be addressed.

information in a clear and easily comprehensible way and to facilitate cross-referencing to other data. We believe that this will be a valuable resource for genome comparison and hope that it is a step towards a more standardized annotation.

Statistics is provided on GenBank and RefSeq annotations as well as on the EasyGene predictions for each genome. The results indicate that wrong start codons and over-annotation of short genes pose the major problems in prokaryotic genome annotations. Furthermore, it seems that the fully automated predictions on average are more reliable than the annotation, although such a claim is difficult to prove.

METHODS

EasyGene is based on an HMM whose parameters are genome specific. The training sets for parameter estimation are made by extracting all maximal ORFs with a length above 120 bases and translating them into proteins. The Swiss-Prot database (Boeckmann *et al*., 2003) is searched for significant matches to the proteins using ungapped BLASTP (Altschul *et al*., 1990) and an E value of 10^{-5}. Hits to the query genus are discarded. Proteins with certain keywords (see below) are also removed. The ORFs with significant hits are processed into two sequence sets. One set contains all of these ORFs, which are considered to be certain genes. The other set contains certain genes with only one possible start codon (Frishman *et al*., 1998). Both sets are similarity reduced (Hobohm *et al*., 1992) and used for training of an HMM. The putative genes in the genome are then scored with the HMM and a measure of the statistical significance is calculated (R). R is the expected number of ORFs that may be predicted with the same length-adjusted score or better in one megabase of random DNA with the same third order Markov statistics as the genome. For each ORF length, the score is adjusted such that the expected number of predictions is constant. For more details see Larsen and Krogh (2003).

In the following section this paper defines 'certain genes with certain start' to be the maximal ORFs that have significant matches to Swiss-Prot in a BLASTP search, and where the hits are distributed in such a manner that only the most upstream start codon can be the start of the gene. The set of 'certain genes with uncertain start' denotes the ORFs whose matches leave more than one possible start of the gene. We consider all the 'certain genes' to be real genes as they have significant matches to Swiss-Prot (Frishman *et al*., 1998; Larsen and Krogh, 2003).

We have analysed 143 publicly available prokaryotic genomes with EasyGene 1.2 and present the predictions on the EasyGene 1.2 webpage.

For each genome the results are presented in four formats.

(1) GFF. It contains a header with genome name and identifier, model version, creation date, R cutoff, version of Swiss-Prot used in the BLASTP search and the genus from which proteins were excluded from the training set. The official GFF format contains eight mandatory and two optional fields (http://www.sanger.ac.uk/software/formats/gff/gff-spec.shtml). To meet our purpose the optional commentary field has been subdivided into six columns (Fig. 1).

(2) HTML. The HTML version of the GFF page is augmented with links to other information. For each prediction there are links to the BLASTP reports, predicted CDS sequences (in FASTA format), Swiss-Prot and GenBank. If the annotations differ, links to both are given. RefSeq annotation is written in parentheses. Furthermore, all predicted CDSs can be downloaded from this page (Fig. 1).

(3) NARROW HTML. It is similar to the HTML version but without columns 1, 2 and 8 (for columns see Fig. 1).

(4) CDS. It contains the same header as the GFF version and all the predicted CDSs for the organism in FASTA format.

The GenBank files for 25 of the genomes contain CDSs with the label '/pseudo', which means that the CDS is considered to be a pseudogene by the annotators. These are the pseudogenes we include in column 9.6 and in the subsequent analysis of our results.

A statistical evaluation of the output from EasyGene 1.2 and information for each of the 143 genomes are collected in three tables. Table 1 contains general information about the genomes as well as the annotation and the predictions. Table 2 focuses on the differences between the annotation and the predictions. Table 3 compares the annotation and the predictions with the certain genes (see Tables 1, 3 and 4 for examples). The full tables are used for comparative analysis of the genomes and are available on the EasyGene 1.2 webpage.

In our comparisons we say that an annotated or a certain gene is predicted if the stop codons are identical. Separate statistics is done on the correspondence of start codons.

Modifications to EasyGene

For the present work several adjustments have been made to the original EasyGene implementation. The most important alteration from the user's point of view is the prediction of alternative starts.

Training on multiple genomes or on separate parts from one genome is now possible. This is an advantage if proteins from plasmids are to be included in the training set or in cases where no whole-genome file exists.

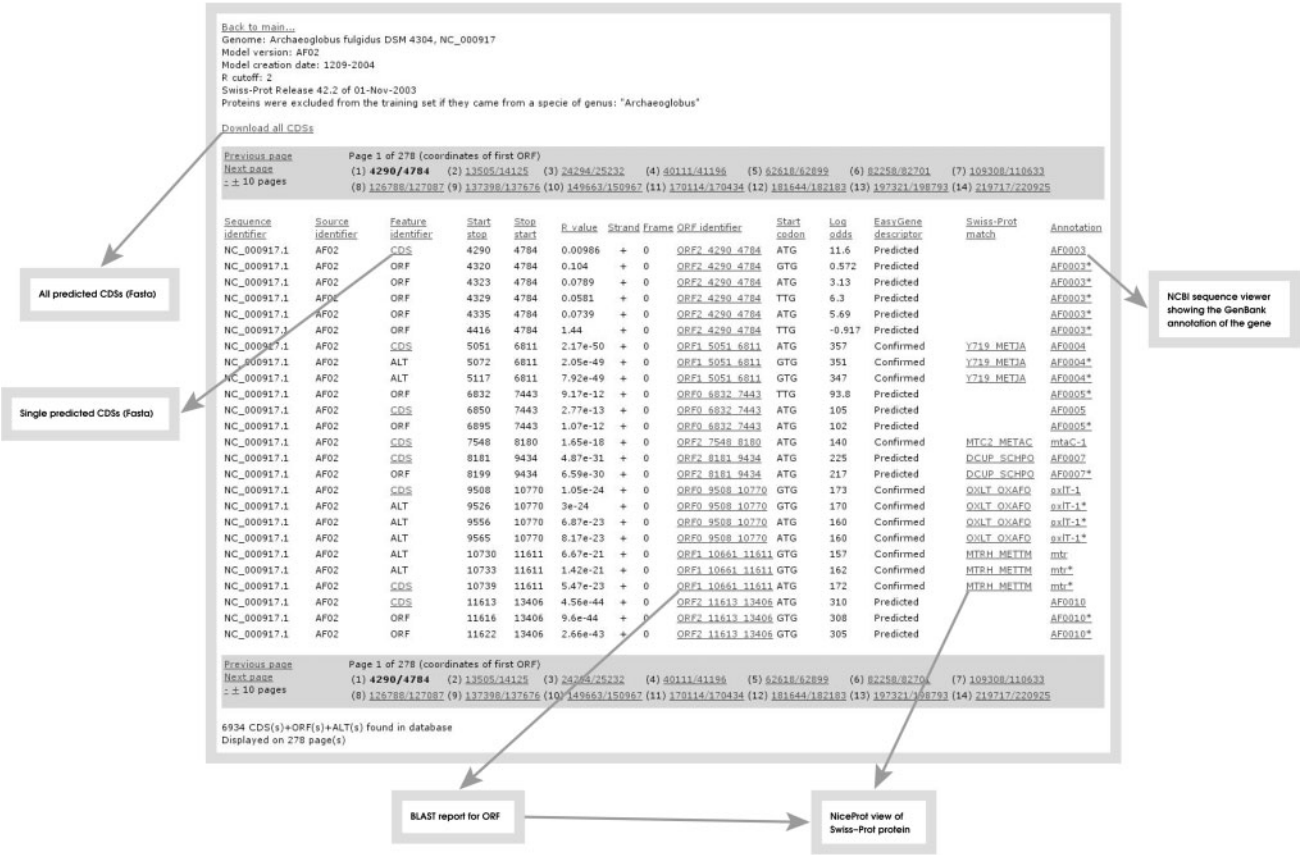
A low-complexity filtered (Wootton and Federhen, 1993) version of the Swiss-Prot database is used in the BLAST search to avoid that local amino acid over-representation in the query results in high scoring matches in low-complexity regions.

The content of the ID line in the FASTA version of Swiss-Prot is changed in order to control the information available for further processing. Once a BLAST report is created [see Larsen and Krogh (2003) for details], proteins are eliminated from further processing if they come from organisms with the same genus as that of the query-genome. Proteins are also removed if they belong to the superkingdom 'Viruses' or if they have one of the keywords in Table 2. In the original version of EasyGene a shorter list of keywords was used to exclude proteins based on the ID line written in the BLAST report. The ID line was copied from the Swiss-Prot description line. This was not an optimal approach as the ID line becomes truncated in the BLAST report. Sometimes proteins were processed because the keyword appeared after the truncation. In Version 1.2 the list has been expanded to include words from the Swiss-Prot keyword line and the elimination is based on both the entire description line and the keyword line. We found it necessary to search both lines because they do not always correspond.

Additional alterations have been made such that proteins from other organisms can be rejected. For instance one might wish to exclude proteins from Salmonellas in a model for *Escherichia coli* K12 because of their close relationship.

The HMM implementation uses scaling of probabilities (Durbin *et al*., 1998). The drawback of this method is that for very long ORFs with a high probability of being a CDS, the relative probability of a non-coding state in the model drops below machine precision. This happens when the ratio of the coding to non-coding probability is of the order of 10^{600}, which in practice means for some ORFs longer than ∼3000 bp. Therefore the probability of finding a subsequent coding region becomes zero. This underflow problem is dealt with by testing whether predictions are made both before and after all ORFs with a length of 3000 bases or more (lORF). The lORF has to be located more than 6000 bases from the sequence start or end. If the test fails, the entire sequence is split into segments, each matching an lORF or an intermediate region. Both types of segments are extended by 51 bases on either side as an attempt not to miss overlapping predictions at the edges. New predictions are made for each segment and subsequently combined into one output. If the test is passed, the original output is used for further processing.

The model gives a log-odds score for all possible starts of a gene. All start codons with a score that differs no more than 15 from the best is reported as alternative start positions. This number was chosen because it results in ∼1


All predicted CDSs (Fasta)
Single predicted CDSs (Fasta)
NCBI sequence viewer showing the GenBank annotation of the gene
BLAST report for ORF
NiceProt view of Swiss-Prot protein

Back to main...
Genome: Archaeoglobus fulgidus DSM 4304, NC_000917
Model version: AF02
Model creation date: 1209-2004
R cutoff: 2
Swiss-Prot Release 42.2 of 01-Nov-2003
Proteins were excluded from the training set if they came from a specie of genus: "Archaeoglobus"

Download all CDSs

Previous page
Next page
- ± 10 pages

Page 1 of 278 (coordinates of first ORF)
(1) **4290/4784** (2) 13505/14125 (3) 24294/25232 (4) 40111/41196 (5) 62618/62899 (6) 82258/82701 (7) 109308/110633
(8) 126788/127087 (9) 137398/137676 (10) 149663/150967 (11) 170114/170434 (12) 181644/182183 (13) 197321/198793 (14) 219717/220925

Sequence identifier	Source identifier	Feature identifier	Start stop	Stop start	R value	Strand	Frame	ORF identifier	Start codon	Log odds	EasyGene descriptor	Swiss-Prot match	Annotation
NC_000917.1	AF02	CDS	4290	4784	0.00986	+	0	ORF2_4290_4784	ATG	11.6	Predicted		AF0003
NC_000917.1	AF02	ORF	4320	4784	0.104	+	0	ORF2_4290_4784	GTG	0.572	Predicted		AF0003*
NC_000917.1	AF02	ORF	4323	4784	0.0789	+	0	ORF2_4290_4784	ATG	3.13	Predicted		AF0003*
NC_000917.1	AF02	ORF	4329	4784	0.0581	+	0	ORF2_4290_4784	TTG	6.3	Predicted		AF0003*
NC_000917.1	AF02	ORF	4335	4784	0.0739	+	0	ORF2_4290_4784	ATG	5.69	Predicted		AF0003*
NC_000917.1	AF02	ORF	4416	4784	1.44	+	0	ORF2_4290_4784	TTG	-0.917	Predicted		AF0003*
NC_000917.1	AF02	CDS	5051	6811	2.17e-50	+	0	ORF1_5051_6811	ATG	357	Confirmed	Y719_METJA	AF0004
NC_000917.1	AF02	ALT	5072	6811	2.05e-49	+	0	ORF1_5051_6811	GTG	351	Confirmed	Y719_METJA	AF0004*
NC_000917.1	AF02	ALT	5117	6811	7.92e-49	+	0	ORF1_5051_6811	GTG	347	Confirmed	Y719_METJA	AF0004*
NC_000917.1	AF02	ORF	6832	7443	9.17e-12	+	0	ORF0_6832_7443	TTG	93.8	Predicted		AF0005*
NC_000917.1	AF02	CDS	6850	7443	2.77e-13	+	0	ORF0_6832_7443	ATG	105	Predicted		AF0005
NC_000917.1	AF02	ORF	6895	7443	1.07e-12	+	0	ORF0_6832_7443	ATG	102	Predicted		AF0005*
NC_000917.1	AF02	CDS	7548	8180	1.65e-18	+	0	ORF2_7548_8180	ATG	140	Confirmed	MTC2_METAC	mtaC-1
NC_000917.1	AF02	CDS	8181	9434	4.87e-31	+	0	ORF2_8181_9434	ATG	225	Predicted	DCUP_SCHPO	AF0007
NC_000917.1	AF02	ORF	8199	9434	6.59e-30	+	0	ORF2_8181_9434	ATG	217	Predicted	DCUP_SCHPO	AF0007*
NC_000917.1	AF02	CDS	9508	10770	1.05e-24	+	0	ORF0_9508_10770	GTG	173	Confirmed	OXLT_OXAFO	oxlT-1
NC_000917.1	AF02	ALT	9526	10770	3e-24	+	0	ORF0_9508_10770	GTG	170	Confirmed	OXLT_OXAFO	oxlT-1*
NC_000917.1	AF02	ALT	9556	10770	6.87e-23	+	0	ORF0_9508_10770	ATG	160	Confirmed	OXLT_OXAFO	oxlT-1*
NC_000917.1	AF02	ALT	9565	10770	8.17e-23	+	0	ORF0_9508_10770	ATG	160	Confirmed	OXLT_OXAFO	oxlT-1*
NC_000917.1	AF02	ALT	10730	11611	6.67e-21	+	0	ORF1_10661_11611	GTG	157	Confirmed	MTRH_METTM	mtr
NC_000917.1	AF02	ALT	10733	11611	1.42e-21	+	0	ORF1_10661_11611	GTG	162	Confirmed	MTRH_METTM	mtr*
NC_000917.1	AF02	CDS	10739	11611	5.47e-23	+	0	ORF1_10661_11611	ATG	172	Confirmed	MTRH_METTM	mtr*
NC_000917.1	AF02	CDS	11613	13406	4.56e-44	+	0	ORF2_11613_13406	ATG	310	Predicted		AF0010
NC_000917.1	AF02	ORF	11616	13406	9.6e-44	+	0	ORF2_11613_13406	GTG	308	Predicted		AF0010*
NC_000917.1	AF02	ORF	11622	13406	2.66e-43	+	0	ORF2_11613_13406	GTG	305	Predicted		AF0010*

Previous page
Next page
- ± 10 pages

Page 1 of 278 (coordinates of first ORF)
(1) **4290/4784** (2) 13505/14125 (3) 24294/25232 (4) 40111/41196 (5) 62618/62899 (6) 82258/82701 (7) 109308/110633
(8) 126788/127087 (9) 137398/137676 (10) 149663/150967 (11) 170114/170434 (12) 181644/182183 (13) 197321/198793 (14) 219717/220925

6934 CDS(s)+ORF(s)+ALT(s) found in database
Displayed on 278 page(s)

Fig. 1. The predictions for each sequence are presented on a separate page. The pages provide for easy access to the sequence of predicted CDSs and facilitate navigation to the relevant BLASTP report, Swiss-Prot data and NCBI resources. Each column is described in links on the page. The first eight columns comply with the GFF format. Columns 9–14 are an extension of the GFF commentary field. These columns contain information about the maximal ORF, the predicted start codon, the sequence log-odds score, the type of sequence, the highest scoring match in Swiss-Prot and the match to GenBank and RefSeq annotation. For further description, see the webpage.

Table 1. Example of statistics in Table 1 of the EasyGene 1.2 output obtained for nine organisms using R-value cutoff = 2

Genome	A	P	G size	AT	GBc	GFc
Aeropyrum pernix	2694 (1841)	1695	1 669 695	43.7	77.3 (85.4)	88.1
Rickettsia conorii Malish 7	1374 (1374)	1183	1 268 755	67.6	80.5 (80.5)	75.8
Xylella fastidiosa 9a5c	2766 (2766)	2366	2 679 306	47.3	83.2 (83.2)	76.0
Vibrio cholerae chr. II	1091 (1092)	919	1 072 315	53.1	84.4 (84.5)	82.2
Leptospira interrogans chr. I	4358 (4360)	3206	4 332 241	65	77.9 (77.9)	73.4
L.interrogans chr. II	367 (367)	284	358 943	64.9	79.6 (79.6)	75.6
Pirellula sp.	7322 (7322)	4788	7 145 576	44.6	94.6 (94.6)	84.7
Clostridium tetani E88	2372 (2372)	2733	2 799 251	71.3	85.4 (85.4)	86.9
Bordetella pertussis	3805 (3806)	4172	4 086 189	32.3	82.7 (82.7)	88.5

A, number of annotated genes including CDSs labelled as pseudogenes; P, number of predicted genes; G size, genome size; AT, AT content in percent; GBc, coding percentage of genome as annotated in GenBank not including pseudogenes; GFc, coding percentage of genome as predicted by EasyGene. Values for the annotation are given for GenBank and (RefSeq).

Table 2. Proteins are removed if they have one of these keywords

Alternative initiation	Possible
Bacteriophage	Probable
Hypothetical	Putative
Hypothetical protein	Transposable element
Insertion	Transposon
Phage	Unknown
Phagocytosis	Viral occlusion body
Plasmid	Virulence

alternative start codon being reported for each certain start gene, for the genomes *Bordetella parapertussis* (AT = 31.9%) and *E.coli* K12 (AT = 49.2%).

Programs have been made to add information to the EasyGene output about the difference between predictions and annotations. Information is also available about correspondence between prediction and certain genes and about Swiss-Prot matches.

From the final model output, webpages are created automatically and put on the EasyGene 1.2 homepage. The webpages contain the predictions and links to NCBI and Swiss-Prot.

Table 3. Examples from Table 2 of the EasyGene 1.2 output

Genome	AeP	AnP	PnA	DS	Ap	ApE
A.pernix	1567 (1563)	1127 (278)	128 (132)	1001 (977)	0 (0)	0 (0)
R.conorii Malish 7	1149 (1149)	225 (225)	34 (34)	160 (160)	0 (0)	0 (0)
X.fastidiosa 9a5c	2043 (2043)	723 (723)	323 (323)	783 (783)	0 (0)	0 (0)
V.cholerae chr. II	880 (881)	211 (211)	39 (38)	298 (298)	0 (0)	0 (0)
L.interrogans chr. I	3168 (3168)	1190 (1192)	38 (38)	820 (820)	0 (0)	0 (0)
L.interrogans chr. II	282 (282)	85 (85)	3 (3)	83 (83)	0 (0)	0 (0)
Pirellula sp.	4728 (4728)	2594 (2594)	60 (60)	2314 (2314)	0 (0)	0 (0)
C.tetani E88	2347 (2347)	26 (26)	386 (386)	734 (734)	0 (0)	0 (0)
B.pertussis	3695 (3695)	110 (111)	477 (477)	1329 (1329)	359 (359)	281 (281)

AeP, number of annotated genes that are predicted (identical stop codons) including CDSs labelled as pseudogenes; AnP, number of genes annotated but not predicted including pseudogenes; PnA, number of genes predicted but not annotated; DS, number of genes from AeP annotated with a different start than predicted; Ap, CDSs annotated as pseudogenes; ApE, CDSs annotated as pseudogenes and predicted to be real genes. Values for the annotation are given for GenBank and (RefSeq).

Table 4. Examples from Table 3 of the EasyGene 1.2 output

Genome	C	CnA	CnP	CS	CSADS	CSPDS
A.pernix	736	5 (9)	6	203	72 (68)	7
R.conorii Malish 7	656	9 (9)	20	426	1 (1)	3
X.fastidiosa 9a5c	2475	27 (27)	29	699	16 (16)	24
V.cholerae chr. II	2317	25 (25)	5	857	1 (7)	7
L.interrogans chr. I	1578	3 (3)	15	706	20 (20)	20
L.interrogans chr. II	1578	1 (1)	1	706	5 (5)	4
Pirellula sp.	2041	9 (9)	37	413	8 (8)	36
C.tetani E88	1485	83 (83)	3	661	16 (16)	8
B.pertussis	2345	100 (100)	12	654	82 (82)	96

C, number of certain genes (possibly with unknown start); CnA, number of certain genes not annotated; CnP, number of certain genes not predicted; CS, number of certain genes with certain start in the training set; CSADS, CS annotated with a different start; CSPDS, CS predicted with a different start. The values for C and CS are model specific and will be the same for each sequence run by the same model. Values for the annotation are given for GenBank and (RefSeq).

We create three tables containing information and statistical evaluation of output data for each genome (see Tables 1, 3 and 4 for examples). The tables are also used to check for suspicious behaviour of EasyGene. If for instance the set of certain genes is poorly predicted for a particular genome, the predictions are inspected manually and a new model might be created.

RESULTS AND DISCUSSION

From the data of 143 prokaryotic genomes we observe that the fraction of certain genes with certain start (ORFs with significant Swiss-Prot matches and only one possible start codon) in the certain genes set increases with higher AT content (Fig. 2). These genes comprise more than half of the set when the AT content is >67.9%. This points to the difficulty of identifying the correct start codons in organisms with low AT content. In these organisms the G-rich start codon GTG is used much more frequently than in AT-rich organisms. Furthermore, in GC-rich organisms the frequency of GTG is higher than the frequency of stop codons (TAA, TGA, TAG). In the AT-rich organisms the frequency of start codons is the same or lower as that of the stop codons. Therefore, there are generally more possible start codons for each stop codon in GC-rich organisms.

EasyGene only considers the three most common start codons ATG, GTG and TTG. Prokaryotes can use five additional start codons and genes with these start codons will be predicted to have a wrong start. Furthermore, the set of certain genes with certain starts can potentially contain genes with the wrong start. This might bias the model to predict a wrong start codon even in the case where ATG, TTG or GTG is being used. However, ATG, GTG and TTG are considered to be the most abundant by far, and based on EasyGene's ability to predict experimentally verified start codons in *E.coli* (Larsen and Krogh, 2003), we believe the effects of excluding the alternative start codons to be minor.

In the following discussion we consider each sequence separately as it is annotated in GenBank, unless otherwise stated. The majority of plasmids and extra chromosomal elements are short sequences and do not have many genes compared to an entire genome. If the annotation and the predictions differ in a few genes, the fractional differences for each sequence become large and skew the general results. It is difficult to interpret these differences between annotation and prediction for plasmids and extra chromosomal elements and they are therefore omitted from the discussion. As expected, Figure 3 shows that there is an increasing tendency to both annotate (slope $= 0.62$) and predict (slope $= 0.80$) the most upstream start codon with higher AT content in a genome. However, in the annotations the most upstream start codon is generally preferred much more frequently than predicted by EasyGene, especially in genomes with AT content <50%. We also observe that the fraction of predicted genes with a different start codon than the annotated one is higher in GC-rich organisms (Fig. 4). The outliers *Aeropyrum Pernix* and *Pirellula* sp. are predicted to have 59.1 and 48.3% genes with a different start, respectively. We obviously do not know the true number of most upstream start codons being used. From the above discussion and the picture emerging from the EasyGene predictions, we believe there is a tendency to use the most upstream start codon too often in the annotations. There are of course well-annotated genomes with many experimentally validated starts where this discussion does not apply.

There is an increase (slope $= 0.07$) in the fraction of annotated genes that are predicted, with growing AT content (Fig. 4). Again, this indicates the difficulty in finding genes in GC-rich organisms. There are generally more annotated genes that are not predicted, than vice versa (data not shown).

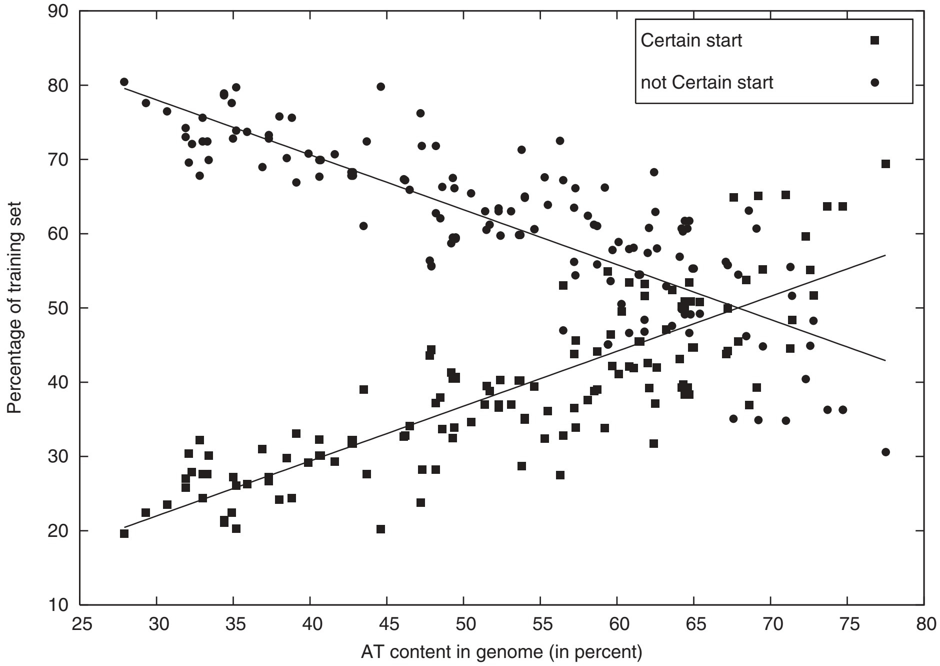


Fig. 2. The fraction of certain start genes in the training set increases in genomes with higher AT content and is generally >50% when AT exceeds 67.6%.

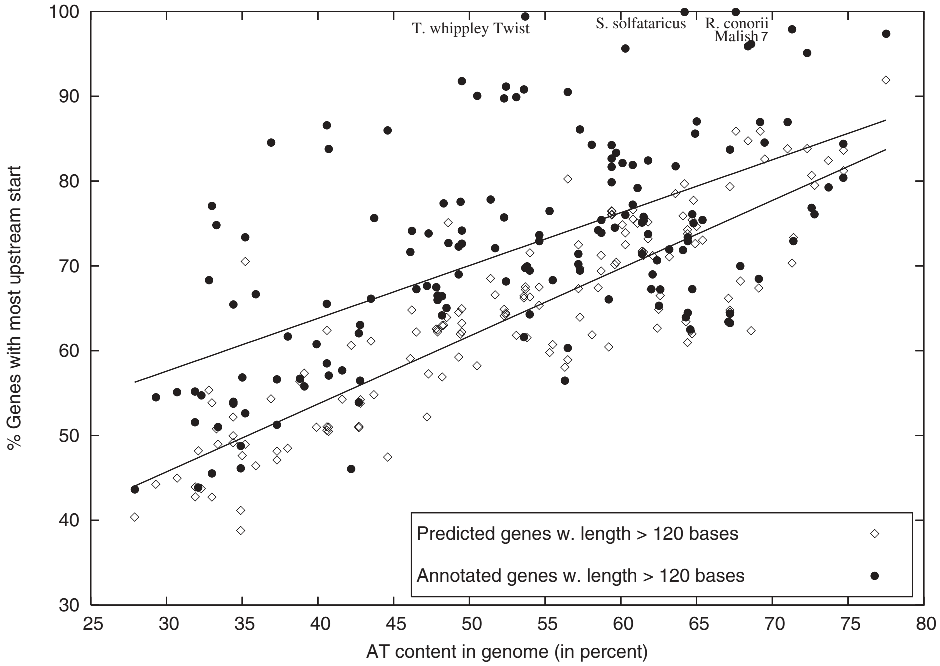


Fig. 3. There is a tendency towards annotating genes with the most upstream start codon. Particularly in GC-rich genomes this is done much more often than predicted by EasyGene.

Skovgaard *et al*. (2001) published a study of 33 microbial genomes, which indicated that the genomes were all over-annotated to a higher or lesser degree. Based on the predictions from EasyGene we wanted to investigate the over-annotation issue again, to see if the same trend applies to the larger number of available genomes. When comparing the genomic coding percentage predicted by EasyGene to the annotated value with regard to either genome size or AT content no clear trend can be seen. However, this may be due to EasyGene generally predicting longer genes, while there are many short genes annotated (Fig. 5). Skovgaard *et al.* (2001) showed that it is generally more difficult to discriminate between short genes and random ORFs in GC-rich organisms.

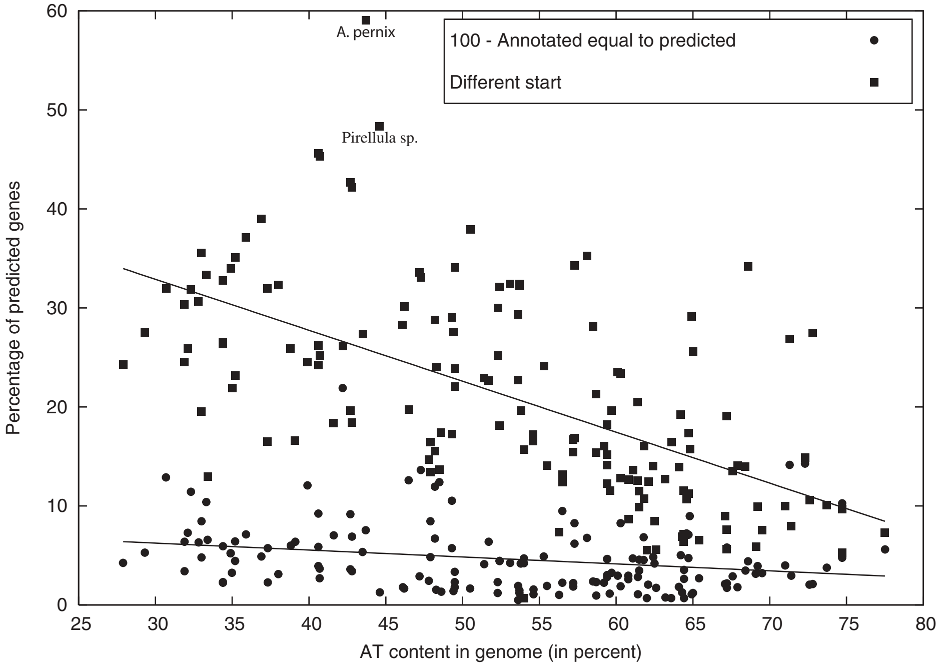


Fig. 4. The fraction of predicted genes with a start codon different from the annotated one is higher in genomes with low AT content. The largest differences are found in *A.pernix* and *Pirellula* sp. The fraction of annotated genes equal to predicted based on stop codon comparison increases with higher AT content.

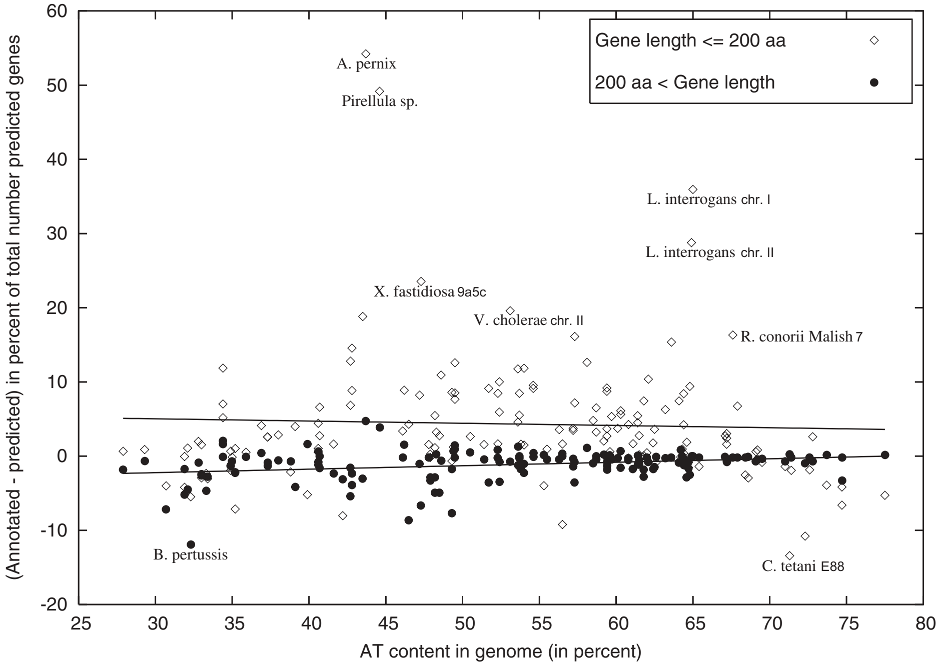


Fig. 5. In many genomes the fraction of small annotated genes is larger than the fraction of small predicted genes (most extreme in *Pirellula* sp., *L.interrogans* chr. I and II, and *X.fastidiosa* 9a5c). Fractions are with respect to all predicted genes. For a few genomes short genes could be missing from the annotation (e.g. *C.tetani* E88). EasyGene 1.2 generally predicts longer genes than what is annotated (*B.pertussis*).

We observe a tendency towards annotating a larger fraction of short genes in GC-rich organisms, which supports the previous study. To assess the over-annotation in terms of the difference between the number of annotated and predicted genes we need a single standard for comparison. We do not know the true number of genes in a genome. The genomes have been annotated by different methods which makes the total number of annotated genes unsuitable as a standard. Instead we use the total number of predicted genes. The over-annotation is defined as follows:

$$\text{over-annotation\%} = \frac{(\#\text{Annotated} - \#\text{Predicted}) * 100}{\#\text{Predicted}}. \qquad (1)$$

We do not include the annotated pseudogenes in this calculation. Using the difference between the number of annotated and predicted genes may lead to estimations that are too high or too low. However, based on a previous sensitivity study (Larsen and Krogh, 2003) we believe that the estimates are reasonable. We predict 41 of 143 organisms to be over-annotated by >5%. Especially *Rickettsia conorii Malish* 7 (16%), *Xylella fastidiosa* 9a5c (17%), *Vibrio cholerae* chromosome II (19 %), *Leptospira interrogans* chromosome I (36%) and II (29%), *Pirellula* sp. (53%) and *A.pernix* (59%) seem to have many annotated CDSs that are not likely to be real genes. Common to them all is that a large fraction of the annotated proteins is short (Fig. 5), which is an indication of over-annotation (Skovgaard *et al*., 2001). When considering the RefSeq annotation for *A.pernix* the predicted over-annotation drops to 9% because many of the short genes annotated in GenBank have been removed.

Of 143 organisms 12 are predicted to be under-annotated by >5%. Of these *Clostridium tetani* E88 (13%) and *Bordetella pertussis* (17%) show the largest deviation. For *C.tetani* E88 EasyGene predicts a larger percentage of short genes (Fig. 5) than what is annotated, indicating that perhaps there are short genes missing in the annotation. In *B.pertussis* 78% of the annotated pseudogenes are predicted to be real, which would explain the predicted under-annotation and point to a possible problem for EasyGene in handling pseudogenes.

Of the genomes 26 contain CDSs that are labelled as pseudogenes in the annotation. The number of pseudogenes range from 1 (*Tropheryma whippleii* TW08) to 359 (*B.pertussis*). EasyGene predicts on average 53% of the pseudogenes in a genome to be real genes. Among the 12 genomes predicted to be under-annotated there are 3 organisms (*B.pertussis*, *B.parapertussis* and *Nitrosomonas europaea*) with more than 100 pseudogenes annotated and a large percentage of these are predicted to be real. Hence, EasyGene cannot distinguish some of the pseudogenes from real genes. This is not a problem for all genomes with many pseudogenes and we have no explanation why the abovementioned genomes stand out.

To assess the performance of EasyGene we consider the certain genes not predicted with an *R*-cutoff of 2 (default) for models based on one sequence only. The percentage of certain genes that are not predicted range from 0 to 8% (mean 1.3%). In comparison, the annotation misses 0–7% (mean = 1.6%) of the genes. This indicates that EasyGene is able to find real genes that are not annotated and that the average performance of EasyGene is a little better than that of the annotation. For the certain genes with certain start EasyGene predicts the wrong start for 0–15% (mean = 4.6%). We expect the annotation to be doing better than EasyGene in this measurement because the most upstream start, which is always correct in our set with certain starts, is often preferred in the annotations. Indeed we observe that 0–35% (mean = 4.4%) are considered to have a different start by the annotation. This does not indicate that the annotation of all the start codons is better on average. Experimentally validated annotations will of course outperform EasyGene.

Some prokaryotic genomes have regions that are known to be horizontally acquired and therefore may have different sequence statistics. To investigate EasyGenes ability to predict genes in such regions we have looked at predictions in some of the pathogenicity islands in *E.coli* O157:H7 and *Salmonella typhimurium* (Table 5). In *E.coli* O157:H7 the locus of enterocyte effacement (LEE) contains 54 ORFs (Perna *et al*., 1998). EasyGene finds 94% of the confirmed genes and all the confirmed start genes in this region. Of the 54 ORFs 91% are predicted to be genes. In *S.typhimurium* pathogenicity islands 1 (SPI1) and 4 (SPI4) there are 48 and 18 ORFs, respectively (Marcus *et al*., 2000). There are no confirmed start genes found in SPI4. EasyGene predicts all confirmed genes in these regions. Ninety-four percent and 55% of the ORFs are predicted to be genes in SPI1 and SPI4, respectively. It seems that EasyGene is able to find genes in horizontally transferred segments, at least in the examples mentioned here.

Table 5. Predictions in pathogenicity islands (PAI) in *E.coli* O157:H7 (LEE) and *S.typhimurium* (SPI1 and SPI4)

PAI	C	CF	CS	CSF	ORFs	CDSs
LEE	16	15	5	5	54	49
SPI1	29	29	10	10	48	45
SPI4	5	5	0	0	18	10

C, number of certain genes (possibly with unknown start) in the region; CF, number of certain genes found; CS, number of certain genes with certain start in the region; CSF, CS found; ORFs, number of ORFs in the region; CDSs, number of CDSs predicted in the region ($R = 2$).

It is possible that the 'certain gene' sets contain errors because of matches to wrong genes in Swiss-Prot, matches to low complexity regions that have escaped detection, or because of random matches.

By removing hits from closely related organisms, we limit the first problem at the cost of losing lineage-specific genes in the training set. This will only be a problem if the statistics of these genes differ from the rest of the genes. Perhaps the prediction of lineage-specific short genes are affected too.

Low complexity regions are masked both in the database and in the BLAST search, so we believe that this problem is minimal.

Finally, we have a reasonably stringent *E*-value cut-off (10^{-5}), which should avoid random matches. We therefore believe that the number of wrong 'certain genes' is very small, and probably as small as one can hope for in a fully automated procedure.

CONCLUSION

We have shown that accurate annotation of prokaryotic genomes still is not an easy task to undertake and that problems with wrongly annotated protein coding genes seem to persist in many organisms. A major issue that needs attention is the discrimination between short random ORFs and real genes. This problem is the largest contributor to over-annotation, especially in GC-rich genomes. There is a tendency for annotators to choose the most upstream start codon for a gene and there are genomes where this has been done for >99% of the genes. This seems extreme, even though AT-rich organisms have fewer start codons per stop codon than GC-rich organisms. More transparency in the annotation process and in the presentation is needed for others to understand these choices.

On average our fully automated method performs slightly better than the manual annotation. Therefore, we believe that EasyGene can be used as a valuable time-saving tool in the annotation work flow. Combining a very good initial starting point and the biological knowledge of the annotators, which is difficult to incorporate in a fully automated method, could increase the time spent on the more challenging problems in prokaryotic genome annotation and increase the overall performance.

We have presented the gene predictions for 143 genomes made by EasyGene 1.2. These predictions are made using a single standard, which facilitates comparison of related genomes and large-scale studies like the current one. The predictions are available from our website along with the statistical tables used in this study. The website has easy access to related resources at NCBI and Swiss-Prot and will be updated with new genomes.

ACKNOWLEDGEMENTS

We thank Lars G. T. Joergensen, who has made a large contribution to the database and website, and Thomas Schou Larsen, Morten Lindow, Lise Petersen and Kasper Munch Terkelsen for valuable discussions. This work was supported by the Danish National Research Foundation and by EU grant no. QLRI-CT-2001-00015.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

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

Besemer,J. *et al*. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*., **29**, 2607–2618.

Boeckmann,B. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.

Cerdeno-Tarraga,A.M. *et al*. (2003) The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. *Nucleic Acids Res*., **31**, 6516–6523.

Delcher,A.L. *et al*. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*., **27**, 4636–4641.

Doerks,T. *et al*. (1998) Protein annotation: detective work for function prediction. *Trends Genet*., **14**, 248–250.

Durbin,R.M., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

Fitz-Gibbon,S.T. *et al*. (2002) Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. *Proc. Natl Acad. Sci. USA*, **99**, 984–989.

Frishman,D. *et al*. (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res*., **26**, 2941–2947.

Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol*., **1**, 55–67.

Hobohm,U. *et al*. (1992) Selection of representative protein data sets. *Protein Sci*., **1**, 409–417.

Larsen,T.S. and Krogh,A. (2003) EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, **4**, 21.

Lukashin,A.V. and Borodovsky,M. (1998) GeneMark. hmm: new solutions for gene finding. *Nucleic Acids Res*., **26**, 1107–1115.

Marcus,S.L. *et al*. (2000) *Salmonella* pathogenicity islands: big virulence in small packages. *Microbes Infect*., **2**, 145–156.

McLeod,M.P. *et al*. (2004) Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other *Rickettsiae*. *J. Bacteriol*., **186**, 5842–5855.

Perna,N.T. *et al*. (1998) Molecular evolution of a pathogenicity island from enterohemorrhagic *Escherichia coli* O157:H7. *Infect. Immun*., **66**, 3810–3817.

Salzberg,S.L. *et al*. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*., **26**, 544–548.

Skovgaard,M. *et al*. (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet*., **17**, 425–428.

Wei,J. *et al*. (2003) Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457t. *Infect Immun*., **71**, 2775–2786.

Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem*., **17**, 149–163.