**Universität Bielefeld**

Technische Fakultät
Graduiertenkolleg Bioinformatik
Institut für Genomforschung

# Gene Finding and the Evaluation of Synonymous Codon Usage Features in Microbial Genomes

Zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften der Universität Bielefeld
vorgelegte Dissertation

von

Alice Carolyn McHardy

| | |
|---|---|
| Thesis advisors: | Prof. Dr. Robert Giegerich |
| | Prof. Dr. Alfred Pühler |

UNIVERSITÄT BIELEFELD

Alice McHardy
Siegfriedstraße 51
D-33615 Bielefeld
Alice.McHardy@Genetik.Uni-Bielefeld.DE

# Preface

Genome research is a rather young area of scientific interest which has evolved within the last few years since the first complete genome sequences of both pro- and eukaryotic organisms became available [1, 2]. Today, the production of genomic sequence data – involving DNA sequencing, assembly and gap closure [3] – has become a routine operation which is performed at many institutions worldwide. The current challenge is the development of methods for the analysis of the data. Besides methods directly related to the prediction [4] and functional characterization of genes and other genetic elements [5, 6], related topics such as the analysis of genome structure [7], phylogeny [8], molecular evolution [9, 10] and comparative genomics [11, 12, 13] are hot spots of scientific research.

An important application of sequence analysis methods is in genome annotation. At this stage of whole genome projects, the aim is the production of a first, preliminary description of the genomic sequence, which specifies the function and location of biological features such as genes and regulatory genetic elements. The corresponding information can either be derived by sequence analysis methods or is available from laboratory experiments described in literature. The latter is usually only true for a small fraction of the genes. Thus, sequence-derived evidence plays a major role in annotation.

Genome annotation may be seen as only the first step and a necessary means for the generation of further biological knowledge for an organism. Based on such data, transcriptome and proteome experiments can be designed on a genome-wide scale. The recent establishment of high-throughput analysis methods in these research areas allows the large-scale, systematic research of biological properties. A novel challenge that results from this development is the inference of knowledge by linking, combining and integrating information derived from such different sources.

In this work, topics related both to genome annotation and to linking sequence-derived information with experimental data are dealt with. Parts of it have already been published or accepted for publication in different journals. The content of chapter 3 will appear in the first issue of *Proteomics* in 2004. The implementation of the gene prediction component of the GenDB genome annotation system (chapter 5) is described in the publication on GenDB which appeared in *Nucleic Acids Research* in April, 2003. In the following, the aims and topics of this work are explained in more detail.

Chapter 1 describes the development of joint gene finding strategies for microbial genomes which combine the strengths of two commonly used programs. A large number of genome projects have either recently been finished or are currently underway, and it is becoming increasingly important to have performant methods for this task. To begin with, the gene finding performance of the two programs was determined for a data set of 114 prokaryotic genome sequences belonging to a wide variety of microbial organisms. Based on the information obtained hereby, joint application strategies were optimized, using different parameters with relevance to the gene finding problem. The resulting combined methods are significantly improved in performance, especially for GC-rich genomes. Using the implementation within the GenDB genome annotation system (chapter 5), they are already being applied in several microbial genome projects.

The second part focuses on the evaluation of synonymous codon usage features of prokaryotic coding sequences. For classification based on such features, a novel method of log-odds ratio scoring is introduced, which has several favorable properties. Chapter 2 contains a description of the method and its implementation with the CoBias program.

In chapter 3, the novel method is applied to the prediction of highly expressed genes and for estimation of gene expression levels. The implicit assumption is that expression level-dependent features in codon usage can be taken as estimates of protein expression rates. This is supported by a comparison with data on protein abundance from the *Escherichia coli* and *Bacillus subtilis* exponential growth phase. A comparison with another frequently used method to estimate gene expression levels shows the favorable properties of the new approach. It is finally demonstrated how the results can be used for an application we named ’predictive proteomics’ – to improve the *in silico* simulation of 2-dimensional gel electrophoretic experiments.

In chapter 4, the method is used for the detection of horizontally transferred genes contained in contemporary microbial genomes and the proposal of a putative donor species. As foreign genetic material is likely to be transferred in complete functional units, instead of searching for single genes, the search was extended to clusters of atypical genes (CAGs) in the genomic sequence. The method allows the inference of a potential donor genome based on

genome-specific sequence properties, which is an innovation compared to existing methods. For a simulation experiment with artificial gene transfer events between current microbial genomes, the method is shown to have high discriminatory power and sensitivity in donor detection. An in detail evaluation is performed for the bacterial hyperthermophile *Thermotoga maritima*, for which ample evidence of horizontal gene transfer events from archaeal species exists. For *T. maritima*, the predicted CAGs and their putative donor genomes agree with previous studies of the genome.

## Acknowledgments

I would like to thank my advisors Prof. Dr. Robert Giegerich and Prof. Dr. Alfred Pühler for their scientific advice and interest in my work. This was carried out with my being a PhD student in the 'Graduiertenkolleg Bioinformatik' of the Faculty of Technology at Bielefeld University. I am also very grateful to Dr. Folker Meyer and Dr. Jörn Kalinowski of the Bielefeld Institute for Genome Research, who gave me the opportunity to be associated with this institution and participate in large-scale research projects such as the whole genome sequencing and annotation of a number of microbial organisms. I am especially indebted to all members of the Bioinformatics Resource Facility of the Institute for Genome Research who supported me in this work and patiently answered many questions, in particular Burkhard Linke, Michael Dondrup, Alexander Goesmann, Oliver Rupp and Lutz Krause. I also thank Dr. Ellen Baake, Dr. Sebastian Böcker, Dr. Marc Rehmsmeier, Dion Whitehead and the members and associates of the Graduiertenkolleg Bioinformatik for their help, comments and ideas on the subject. Dion Whitehead also supplied me with the habitat data table used in chapter 4.

# Contents

**1 Development of joint application strategies for two microbial gene finders** 1

1.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

1.2.1 Data sets . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

1.2.2 Measuring performance and classification accuracy . . . . . . . . . 4

1.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

1.3.1 Composition of the data set . . . . . . . . . . . . . . . . . . . . . 6

1.3.2 Gene finding performance of Glimmer and Critica . . . . . . . . . 7

1.3.3 Glimmer(ct): Improving gene finding performance for the GC-rich genomes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

1.3.4 Gene finding peformance for different gene lengths . . . . . . . . . 10

1.3.5 Diagnostic accuracy of the Glimmer scores . . . . . . . . . . . . . 11

1.3.6 Development of combined strategies . . . . . . . . . . . . . . . . . 12

1.3.7 Performance evaluation of OTS and VTS . . . . . . . . . . . . . . 14

1.4 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

**2 CoBias - Using log-odds ratio scores for classification based on trends in codon usage** 19

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20

2.1.1 The Codon Adaptation Index (CAI) . . . . . . . . . . . . . . . . . 20

2.1.2 The PHX and PA classifications . . . . . . . . . . . . . . . . . . . 20

2.1.3 The $\chi^2$ statistic . . . . . 21
2.1.4 The Mahalanobis distance . . . . . 22
2.1.5 Motivation . . . . . 22
2.2 A probabilistic method for the evaluation of synonymous codon usage features 23
2.2.1 Assessing a trend in codon usage . . . . . 23
2.2.2 Significance estimation . . . . . 24
2.2.3 The ’strength’ of a feature . . . . . 25
2.2.4 Information content of the scoring matrix . . . . . 25
2.3 Implementation . . . . . 26
2.3.1 Parameters of CoBias . . . . . 26
2.4 Discussion . . . . . 28

**3 An analysis of 'predictive proteomics'** 31
3.1 Introduction . . . . . 32
3.2 Materials and Methods . . . . . 34
3.2.1 Data sets . . . . . 34
3.2.2 Assessing expression level-dependent features in codon usage . . . 35
3.3 Results . . . . . 36
3.3.1 Probably highly expressed (PHE) genes in major metabolic pathways 38
3.3.2 PHE genes in protein biosynthesis . . . . . 39
3.3.3 PHE genes in stress response . . . . . 39
3.3.4 Comparison with the Codon Adaptation Index (CAI) . . . . . 40
3.3.5 Comparison with *B. subtilis* proteome data . . . . . 40
3.3.6 Comparison with *E. coli* proteome data . . . . . 42
3.3.7 Simulation of a ’virtual’ 2-D gel . . . . . 46
3.4 Discussion . . . . . 48
3.4.1 Modeling expression level-dependent features . . . . . 48
3.4.2 Relation to skewedness of genomic GC-content . . . . . 48
3.4.3 Relation to protein abundance at exponential growth . . . . . 49
3.4.4 Favorable properties compared to the CAI . . . . . 50
3.4.5 *In silico* 2-D gel simulation . . . . . 50
3.4.6 Summary . . . . . 51

**4 Predicting the origin for horizontal gene transfer events** 53
4.1 Introduction . . . . . 53
4.2 Materials and Methods . . . . . 55
4.2.1 Sequence data . . . . . 55
4.2.2 Detecting atypical genes and a potential donor genome . . . . . 56
4.2.3 Detection of clusters of atypical genes (CAGs) . . . . . 57
4.2.4 Phylogenetic methods . . . . . 57
4.2.5 Performance evaluation . . . . . 58

4.3 Results . . . 59
4.3.1 The discriminatory power of codon usage differences between contemporary genomes . . . 59
4.3.2 Predicting the origin of artificially transferred genes . . . 61
4.3.3 Taking a closer look: CAGs in the *Thermotoga maritima* genome . 63
4.4 Discussion . . . 67

**5 Integration into the genome annotation system GenDB** 69
5.1 Introduction . . . 69
5.1.1 Data model . . . 70
5.1.2 Tool concept . . . 71
5.2 Integration of CoBias . . . 73
5.3 The GenDB gene prediction component . . . 75
5.3.1 Gene finding . . . 75
5.3.2 Automatic annotation . . . 78
5.3.3 Application in microbial genome projects . . . 80

**Bibliography** 81

# List of Figures

1.1 Performance of Glimmer for genomes annotated using Glimmer versus other gene finders in the annotation process. . . . . . 6

1.2 Comparison of tool performance for Glimmer, Critica and the Critica-trained Glimmer on 113 prokaryotic genome sequences. . . . . . 8

1.3 Relation of gene finding performance to genomic GC-content and gene length. 9

1.4 Diagnostic accuracy of the different Glimmer scores. . . . . . 11

1.5 Performance of the combined strategies. . . . . . 12

3.1 Comparison of $S_{AV}(g)$ values of codon usage with the Codon Adaptation Index. . . . . . 41

3.2 Comparison of relative protein abundance with $S(g)$ values of codon usage for 173 *E. coli* proteins and corresponding genes. . . . . . 42

3.3 Comparison of relative protein abundance with $S(g)$ values of codon usage for 173 *E. coli* proteins and corresponding genes, colored by functional categories. . . . . . 45

3.4 Comparison of a virtual 2-D gel with a 2-D gel electrophoresis experiment for *E. coli*. . . . . . 46

4.1 The discriminatory power of codon usage differences between microbial genomes. . . . . . 59

4.2 Distribution of scores for pairs of genomes with different degrees of evolutionary relatedness. . . . . . 60

4.3 Clusters of atypical genes in the *Thermotoga maritima* genome. . . . . . 63

4.4 Phylogenetic analysis of the combined translated sequences of oligopeptide ABC transporter genes of *Thermotoga maritima*, *Pyrococcus abyssi* and *Pyrococcus horikoshi* . . . . . 65
4.5 Clusters of oligopeptide ABC transporter genes in the *Thermotoga maritima* genome. . . . . . 66

5.1 The core data model of GenDB in UML . . . . . 71
5.2 The CoBias tool configuration page. . . . . . 73
5.3 Using CoBias observations to simulate different amounts of protein in the virtual 2D-gel, . . . . . 74
5.4 Default and extended application of Glimmer in GenDB. . . . . . 76
5.5 The Glimmer tool configuration page. . . . . . 77
5.6 The GenDB contig view displaying a genomic region subsequent to gene prediction and automatic annotation according to the OTS strategy. . . . . . 79

# List of Tables

1.1 Mean sensitivity, selectivity and overall performance of different gene finding methods. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

1.2 Mean sensitivity, selectivity and overall performance of different gene finding methods for genes of known function or other confirmation. . . . . . . 10

1.3 Using the Glimmer vote score to divide predictions into probably correct ones and less certain candidates in need of manual validation. . . . . . . . . 13

1.4 Comparison to the Yacop and Glimmer $\cap$ ZCURVE combined strategies . . 14

1.5 Sensitivity and false positive proportion of predictions (1 - Specificity) for Glimmer, Glimmer(ct), OTS and VTS for 27 genome sequences with a GC-content $>$ 0.56. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

2.1 CoBias results for the S-layer gene *cspB* from 28 different *Corynebacterium glutamicum* strains. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

3.1 Information content of the scoring matrix for the three genomes. . . . . . . 34

3.2 Top scoring genes with the highest $S_{AV}(g)$ values . . . . . . . . . . . . . . 36

3.3 Probably highly expressed genes $\geq$ 100 codons in the three genomes . . . . 37

3.4 Functional classification of *E. coli* genes expressed at exponential growth. . 43

4.1 List of species with completely sequenced genomes in the data set. . . . . . 55

4.2 Mean sensitivity ($z_1$) of detecting the donor genome for artificial HGT events for different phylogenetic relationships and GC-content of donor and acceptor genomes. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62

4.3 Habitat or location of isolation of the predicted donor species for *T. maritima* CAGs. Unless specified otherwise, the data is from [14]. . . . . . . . . . . 64

CHAPTER 1

# Development of joint application strategies for two microbial gene finders

As a starting point in annotation of bacterial genomes, gene finding programs are used for the prediction of functional elements in the DNA sequence. Due to the faster pace and increasing number of genome projects currently underway, it is becoming especially important to have performant methods for this task.

This study describes the development of joint application strategies which combine the strengths of two microbial gene finders to improve the overall gene finding performance. Critica is very specific in the detection of similarity-supported genes, as it uses a comparative sequence analysis-based approach. Glimmer employs a very sophisticated model of genomic sequence properties and is sensitive also in the detection of organism-specific genes. Based on a data set of 114 microbial genome sequences, we optimized a combined application approach using different parameters with relevance to the gene finding problem. This results in a significant improvement in specificity while being similar in sensitivity to Glimmer. The improvement is especially pronounced for GC-rich genomes. The method is currently being applied for the annotation of several microbial genomes.

## 1.1 Introduction

Microbial whole genome projects have become quite frequent today. Following sequencing and assembly, in the annotation phase a functional description of the sequence is produced. For the storage, retrieval and processing of hereby necessary information, annotation systems such as Artemis [15], ERGO [16], GenDB [17] or MAGPIE [18] have been developed. As the first step in annotation, gene finders are usually applied for the prediction of functional elements such as coding sequences (CDSs) in the DNA sequence.

Compared to the more complex genetic organization in higher organisms, protein coding sequences in prokaryotic genomes possess a relatively simple structure. The task in microbial CDS prediction is to separate Open Reading Frames (ORFs) that correspond to *in vivo* transcribed and translated regions of protein coding sequence from the purely hypothetical ORFs, that do not constitute functional elements of the organisms chromosome. A further issue is the determination of the correct start position, which contrary to the stop position of a coding sequence is not uniquely defined.

Different classes of microbial gene finders exist. *Ab initio* methods rely on the evaluation of intrinsic sequence properties, such as the biased distribution of DNA oligomers in coding sequences. Examples for programs which implement this approach include Glimmer [19, 20], GeneMark.hmm/S [21, 22], ZCURVE [23] and EasyGene [24]. Extrinsic gene finders additionally use pairwise sequence similarity as ’external evidence’ for their predictions; examples for these are the Critica [25] and Orpheus [26] programs. A still different approach uses a ’Bio-dictionary’ of prokaryotic protein sequence patterns for gene identification [27]. For some genomes, a performance improvement has been obtained by combining the results from two or more programs [23, 28]. These methods have been named the Glimmer $\cap$ ZCURVE [23] and Yacop $\big(\text{Critica} \cup (\text{Glimmer} \cap \text{ZCURVE})\big)$ [28] strategies. For start site prediction, characteristic features of gene starts and the surrounding sequence are utilized, such as preferred start codons and ribosome binding site (RBS) patterns [22, 23, 25, 29].

There is a large number of microbial genome projects either recently finished or currently underway. It is becoming increasingly important to have performant gene prediction methods. These should allow the creation of high quality genome annotation data while reducing superfluous human validation effort. In this study, this is tackled by the development of joint gene finding strategies based on the gene finders Glimmer and Critica. Both have different strengths, are freely available and can be utilized in automated high-throughput analysis on a Unix system. Information regarding their performance is currently scarce and available only for smaller sets of 7 [28] or 18 [23] genomes. As this may not give a representative picture for all genomes available today, initially their performance was evaluated on 114 genome sequences belonging to a wide variety of microbial organisms. Hereby, Glimmer was found to be the more sensitive program but its performance decreases strongly for GC-rich genomes. For example, for the genomes of *Sinorhizobium meliloti* and *Streptomyces coelicolor* there are 1507 and 5817 false positive CDS predictions, respectively. If relying on the program results without further modifications, this means an enormous manual validation effort for human annotators in genome projects. We tackled this problem with the development of joint application strategies for the two programs. Using different parameters with relevance to the gene finding problem, combined strategies with optimized performance were devised.

## 1.2 Methods

### 1.2.1 Data sets

The EMBL annotations of 114 genomic sequences of eubacterial and archaeal microorganisms were used in this study. A complete list can be found at `http://www.Genetik.Uni-Bielefeld.DE/~alice/Geneprediction/Sequences`. To exclude annotation ambiguities, CDSs annotated with a non-integer number of codons or ending without a stop codon were excluded. Getorf from the EMBOSS package [30] was used for ORF determination. Critica and Glimmer-2.1 were run with the option to use ribosome binding site (RBS) information to locate the correct start position. For comparison of the Glimmer performance for genomes annotated using Glimmer versus those using other gene finders, the available Glimmer version at the time of obtaining the annotation data was used (Glimmer-2.1.0). In the further analyses, performance was compared to the latest version of Glimmer (Glimmer-2.1.3). Data sets of genes with known function or other supporting evidence were prepared for every genome based on the information given in the CDS gene product description. For this, all CDSs described without an indication of either function, experimental confirmation, sequence conservation or the occcurence of functional domains were classified as uncertain. Of the total set of 305613 CDSs annotated for the 114 genomes, this was the case for 58889 entries. The genomic sequence data with the corresponding annotated CDSs, gene finding results and ORFs can be browsed using the GenDB web frontend (`http://www.Genetik.Uni-Bielefeld.DE/~alice/geneprediction/gendb_cds.html`). The genes considered as uncertain in this study can be identified by their 'Status function', which was set to 'putative'.

### 1.2.2 Measuring performance and classification accuracy

In a two-class classification problem such as discriminating between hypothetical ORFs and CDSs, the classification performance of a method can be evaluated by determining the numbers of true positive (*TP*), true negative (*TN*), false positive (*FP*) and false negative (*FN*) classified items, where $TP+FP+TN+FN=N$. Positives correspond to ORFs described as CDSs in the annotation, negatives are the remaining 'hypothetical' ORFs. Based on the sensitivity $x=TP/(TP+FN)$ and specificity $y=TP/(TP+FP)$, the correlation coefficient

$$CC(P,A)=\frac{N\cdot x\cdot y-TP}{\big((N\cdot x-TP)(N\cdot y-TP)\big)^{\frac{1}{2}}} \tag{1.1}$$

can be determined, that represents the accuracy of the predictive classification *P* with respect to the annotation *A*. This provides a summary of gene finding performance based on all four parameters [31]. The significance of differences in performance, sensitivity and specificity of the different methods was determined using two-sample t-tests with pooled variance for similar variance samples and the Welch approximation to the degrees of freedom otherwise.

In gene finding, a statistical model of CDS properties is used to evaluate the ’coding potential’ of the analyzed ORF, which is usually represented by a continuous numerical value. Besides such a score, additional parameters such as overlap with neighboring predictions are typically employed for the prediction. The predictive result of a gene finder thus is not identical with a classification based on a single numerical measure. To determine the discriminatory power of the internally used scoring methodologies, ROC analysis [32] was carried out for the different numerical measures used by Glimmer, for which the scores assigned to the ORFs during the analysis are available from the output. The receiver (or relative) operating characteristic (ROC) is a plot of the sensitivity versus the false positive proportion ($FP/(FP+TN)$) of the hypothetical ORFs for various settings of the decision threshold. The area under the ROC curve measures the probability of correct classification and can be used as a single-valued, general measure of classification accuracy [32]. Since in bacterial genomes the number of hypothetical ORFs largely exceeds the number of CDSs, a truncated ROC is calculated, similar to the use in performance evaluation of protein database search methods [33, 34]. The $\mathrm{ROC}_{0.1}$ corresponds to the area under the ROC curve up to a false positive proportion of 10 percent.

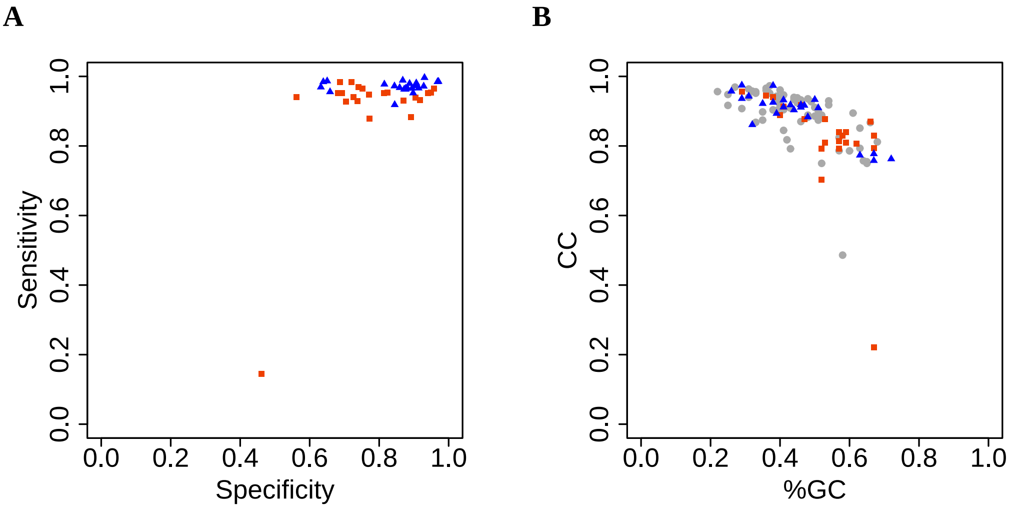


Figure 1.1: Performance of Glimmer for genomes annotated using Glimmer or other gene finders in the annotation process. **A:** Specificity versus sensitivity of Glimmer for genomes annotated using Glimmer ($G$, orange squares) and genomes where other gene finders were employed ($\overline{G}$, blue triangles). **B:** Decreasing Glimmer performance with increasing GC-content. Correlation of Glimmer predictions with annotation data versus genomic GC-content for the $G$, $\overline{G}$ and remaining genomes (grey circles).

## 1.3 Results

### 1.3.1 Composition of the data set

Current practice in microbial genome projects is to use one or more gene finders in combination with sequence database search methods such as BLASTX [35] to locate potential coding sequences, followed by additional manual effort of validation. Therefore, it seemed necessary to first evaluate whether any of the utilized annotations mostly reflects the predictions of the employed gene finder as CDS content, which would render it unsuitable as standard of truth in performance evaluation. To the best of our knowledge, we do not know of any annotation in the data set where Critica has been applied in the gene prediction step. Glimmer has been frequently used and its performance was thus compared for 22 genomes annotated using Glimmer ($G$) to that for 23 genomes where other gene finders were applied ($\overline{G}$, Figure 1.1a). Surprisingly, the mean Glimmer performance is better for $\overline{G}$ ($CC(P,A) = 0.89$) than for the $G$ set ($CC(P,A) = 0.82$). Of the 114 genome sequences, 14 have $CC(P,A)$-values between 0.95 and 0.97. Three of these belong to the $\overline{G}$ and only one to the $G$ set. The two sequences for which Glimmer performs best are the *Clostridium perfringens* and *Listeria monocytogenes* genomes, which both belong to $\overline{G}$.

Rather than the gene finder used, genomic GC-content has the major influence on prediction quality. Figure 1.1b shows the decreasing Glimmer performance for genomes with higher GC-content. These are more frequent in $G$ than in $\overline{G}$. Thus, no genome was excluded because of the gene finder used in the annotation process.

For the genome of the archaebacterium *Aeropyrum pernix*, the sensitivity of both gene finders in reproducing the annotation data was found to be rather low (Glimmer 0.59%, Critica 0.56%). The *Aeropyrum pernix* annotation contains all ORFs longer than 200 codons annotated as CDSs, which has been estimated to result in approximately 100% overannotation [36]. As this annotation thus is no good representation of CDS content, the genome was excluded from further analyses. The remaining 113 genome sequences comprise the data set used in this study.

### 1.3.2 Gene finding performance of Glimmer and Critica

For the complete data set of 113 bacterial and archaeal genomes, the overall gene finding performance of both Glimmer and Critica is quite high. The mean correlation between predicted and annotated CDS is 0.88 for Glimmer and 0.93 for Critica (Table 1.1). Glimmer has a statistically significant higher sensitivity than Critica (+5%, $p = 2.2 \cdot 10^{-12}$, determined with a two-sample t-test, see Methods), but lacks in specificity (-13%, $p = 6.4 \cdot 10^{-22}$).

Some exceptions exist. For the *Mycobacterium leprae* genome, the specificity of Glimmer is only 22%, compared to 81% for Critica. This may be due to the unusually high content of pseudogenes among the annotated CDSs (40%). The resulting coverage of functional CDSs for this intracellular pathogen is 500 per megabase of genome sequence. This is about half the usual coverage for bacterial genomes and has been explained as an extreme case of reductive evolution [37].

Also, for a number of GC-rich genomes the performance of Critica is noticeably better than that of Glimmer (Figure 1.2a). Examples for these are the genomes of *Pseudomonas aeruginosa* (GC-content 67%), *Ralstonia solanacearum* (67%), and most pronounced, the genome of *Streptomyces coelicolor* (72%).

### 1.3.3 Glimmer(ct): Improving gene finding performance for the GC-rich genomes

A problem which occurs in high GC-content genomes when using Glimmer is how to obtain an adequate training set of coding sequences. This is needed for parameter estimation of the Glimmer interpolated context model of CDSs. Per default, Glimmer applies a script called *long-orfs* for this. Up to Glimmer, version 2.1.0, *long-orfs* detects all non-overlapping ORFs longer than 500bp in a given genomic sequence. But the number of such non-overlapping, long ORFs decreases strongly with increasing GC-content of a genome. At some point it is too small to be used [23]. Recently, a novel version of *long-orfs* was released (Glimmer-

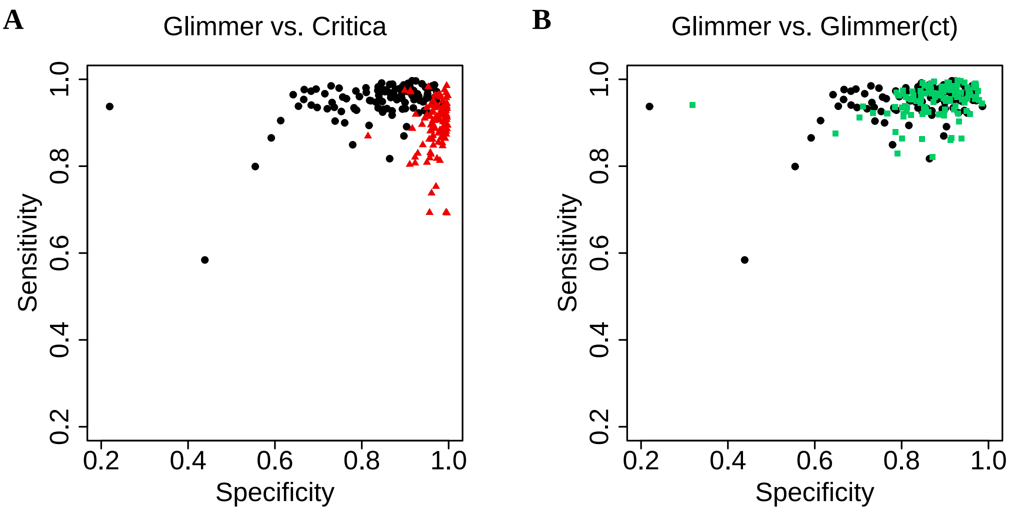


Figure 1.2: Comparison of tool performance for Glimmer, Critica and the Critica-trained Glimmer(ct) on 113 prokaryotic genome sequences. **A, B**: Sensitivity versus specificity for Glimmer (black circles) versus Critica (red triangles) and versus Glimmer(ct) (green squares). With Glimmer(ct), Critica was used to generate the training set of CDSs for parameter estimation of the Glimmer model.

2.1.3), which computes an optimal minimum length of 'long orfs' to enlarge the training set. Still, a dramatic performance difference is evident for Glimmer on GC-rich ($> 56\%$) genomes compared to sequences of lower GC-content (Table 1.1). Both sensitivity (-3%) and specificity (-18%) are reduced for the GC-rich genomes. Figure 1.3a shows the decreasing performance of Glimmer with increasing GC-content of the individual genome sequences.

We thus tried how further changing the composition of the training set can be used to improve the gene prediction performance of Glimmer. An iterative usage, that is using an initial set of predictions as a training set for another Glimmer run, did not lead to any improvement (data not shown). With Glimmer(ct), the more specific Critica CDS predictions were used as the training set. This results in a statistically significant 2% performance improvement compared to the standard application ($p = 0.04$, Figure 1.2c). The Glimmer(ct) prediction is more specific (+3%, $p = 0.02$) without loosing in sensitivity. For GC-rich genomes, the improvement is even more pronounced (+9% in specificity, +1% in sensitivity, Table 1.1).

For Critica, there is a slight loss in both sensitivity and specificity, which results in a 2% ($p = 0.027$) difference in overall gene finding performance between GC-rich and the remaining genomes (Figure 1.3b).

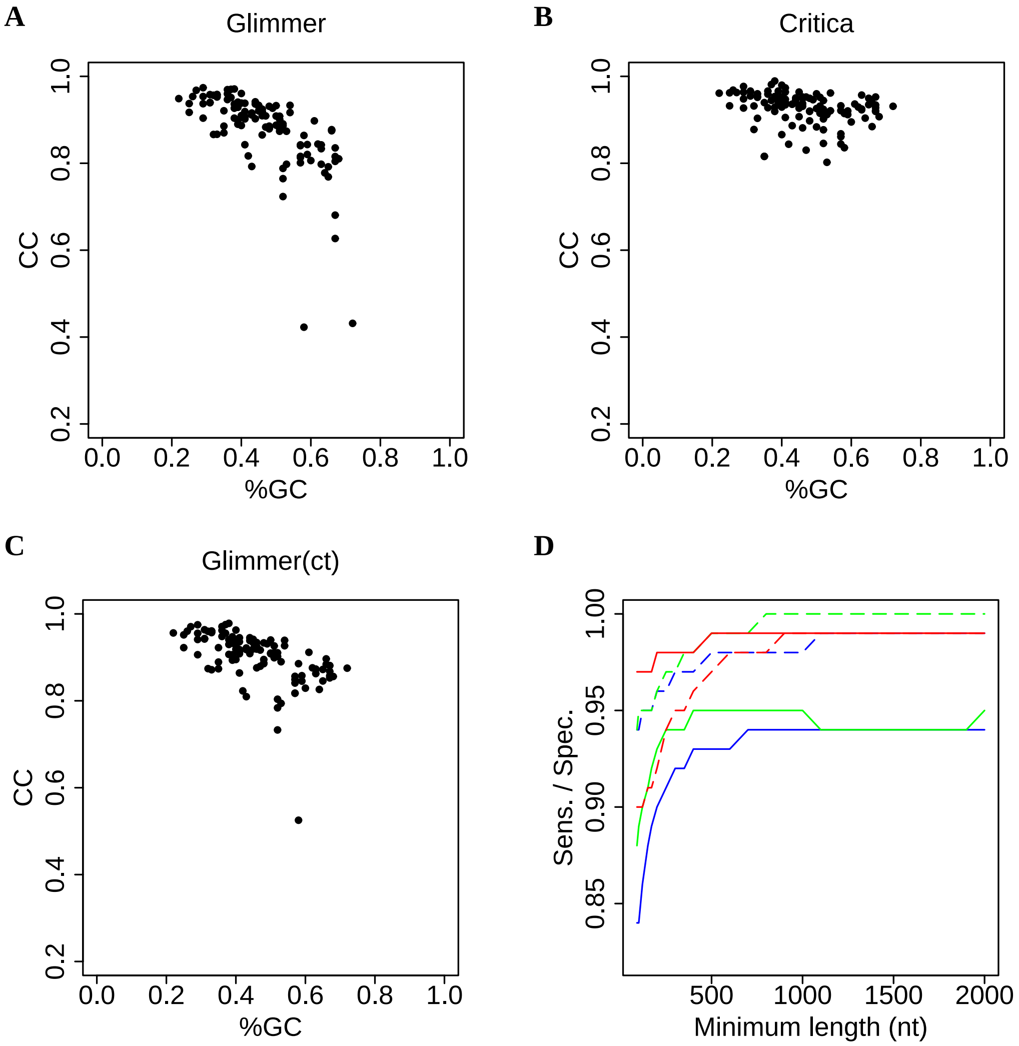


Figure 1.3: Relation of gene finding performance to genomic GC-content and gene length. **A, B, C:** Performance of Glimmer, Critica and Glimmer(ct) versus genomic GC-content for 113 microbial genomes. **D:** Sensitivity (dashed line) and specificity (solid line) of Glimmer (blue), Glimmer(ct) (green) and Critica (red) for different minimum gene length settings.

Table 1.1: Mean sensitivity, selectivity and overall performance of different gene finding methods on 113 bacterial and archaeal genomes.

| Gene finder | $CC(P,A)$ | Sensitivity | Specificity |
|---|---|---|---|
| Glimmer[a] | 0.88 ± 0.10 (0.77 ± 0.13)[b] | 0.95 ± 0.08 (0.93 ± 0.16) | 0.84 ± 0.12 (0.68 ± 0.11) |
| Glimmer[c] | 0.88 ± 0.09 (0.78 ± 0.12) | 0.95 ± 0.05 (0.93 ± 0.08) | 0.84 ± 0.12 (0.70 ± 0.13) |
| Glimmer(ct)[d] | 0.90 ± 0.06 (0.85 ± 0.07) | 0.95 ± 0.04 (0.93 ± 0.03) | 0.87 ± 0.08 (0.80 ± 0.10) |
| Critica | 0.93 ± 0.04 (0.91 ± 0.03) | 0.90 ± 0.06 (0.88 ± 0.04) | 0.97 ± 0.03 (0.96 ± 0.04) |
| Union | 0.90 ± 0.06 (0.85 ± 0.07) | 0.95 ± 0.04 (0.94 ± 0.03) | 0.87 ± 0.08 (0.80 ± 0.10) |
| OTS | 0.92 ± 0.05 (0.91 ± 0.08) | 0.94 ± 0.04 (0.92 ± 0.03) | 0.92 ± 0.07 (0.91 ± 0.12) |
| VTS | 0.93 ± 0.04 (0.92 ± 0.06) | 0.93 ± 0.05 (0.91 ± 0.03) | 0.95 ± 0.05 (0.94 ± 0.09) |

[a] version 2.1.0
[b] The values in parenthesis are for the 27 genomes with a genomic GC-content $> 0.56$.
[c] version 2.1.3, using a new version of *long-orfs* for training set creation.
[d] version 2.1.3, using Critica for training set creation.

Table 1.2: Mean sensitivity, selectivity and overall performance of different gene finding methods for genes of known function or with other confirmation.

| Gene finder | $CC(P,A)$ | Sensitivity | Specificity |
|---|---|---|---|
| Glimmer | 0.79 ± 0.12 (0.72 ± 0.13)[a] | 0.98 ± 0.04 (0.96 ± 0.08) | 0.68 ± 0.16 (0.59 ± 0.15) |
| Glimmer(ct)[b] | 0.81 ± 0.10 (0.79 ± 0.10) | 0.98 ± 0.02 (0.98 ± 0.02) | 0.71 ± 0.15 (0.67 ± 0.14) |
| Critica | 0.86 ± 0.10 (0.87 ± 0.09) | 0.95 ± 0.03 (0.94 ± 0.03) | 0.81 ± 0.15 (0.83 ± 0.15) |
| Union | 0.81 ± 0.10 (0.79 ± 0.10) | 0.98 ± 0.02 (0.98 ± 0.01) | 0.71 ± 0.15 (0.67 ± 0.14) |
| OTS | 0.84 ± 0.10 (0.85 ± 0.11) | 0.98 ± 0.02 (0.97 ± 0.02) | 0.74 ± 0.15 (0.78 ± 0.17) |
| VTS | 0.86 ± 0.10 (0.87 ± 0.10) | 0.97 ± 0.02 (0.96 ± 0.02) | 0.79 ± 0.15 (0.80 ± 0.16) |

[a] The values in parenthesis are for the 27 genomes with a genomic GC-content $> 0.56$.
[b] Using Critica for training set creation.

### 1.3.4 Gene finding peformance for different gene lengths

To examine the relation of gene length and prediction performance for Glimmer, Critica and Glimmer(ct), the sensitivity and selectivity for different settings of the minimum CDS length were compared (Figure 1.3d). The values at a minimum length of 90bp correspond to those given in Table 1.1. The specificity of all three gene finders decreases for shorter CDS lengths. This is more pronounced for Glimmer and Glimmer(ct) than for Critica, which is the most specific tool for all lengths. Glimmer(ct) has the highest sensitivity in detecting longer CDSs. Only if considering the complete set of CDSs longer than 90bp, it becomes identical to that of the standard application.

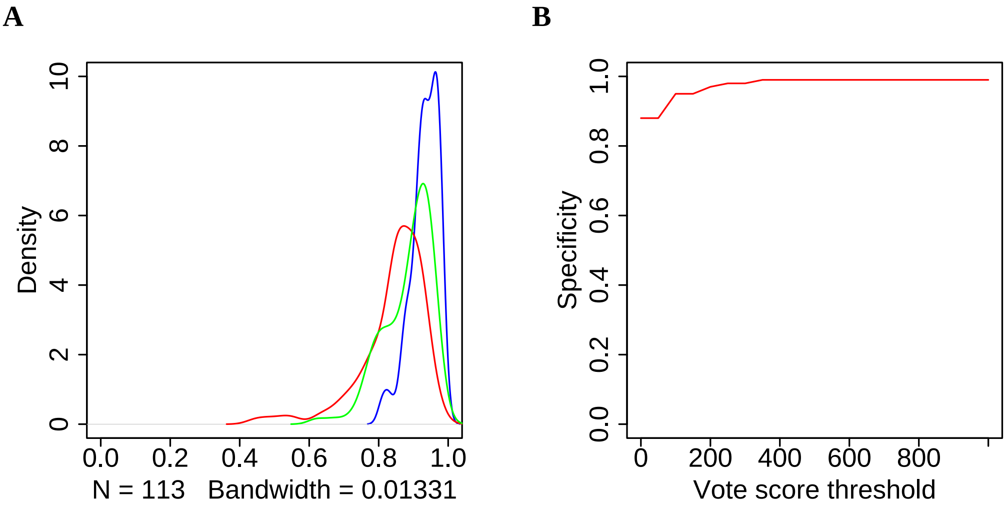


Figure 1.4: Diagnostic accuracy of the different Glimmer scores. **A:** Density estimate of the $\text{ROC}_{0.1}$ distribution for the vote (blue), raw (red) and probability score (green) for the 113 genomes. **B:** Specificity of the remaining Glimmer(ct) predictions for different settings of the vote score threshold.

## 1.3.5 Diagnostic accuracy of the Glimmer scores

Three numerical scores are available from the Glimmer output for the analyzed ORFs. These are a length-normalized raw log-score from the utilized interpolated context model, a probability value and a vote score, which is the sum of the probability scores for subregions contained within the analyzed ORF sequence in other frames. The primary decision criterion Glimmer uses is the probability score, optionally also ORFs with vote scores above a certain threshold are predicted. We were interested to determine which of these scores allows the most reliable prediction of CDSs. As a measure of predictive accuracy, the $\text{ROC}_{0.1}$ was determined for the different measures. Figure 1.4a shows a density estimate for the $\text{ROC}_{0.1}$ distributions of the raw, probability and vote scores for the 113 genomes. With a mean $\overline{\text{ROC}}_{0.1}$ of 0.93, the vote score allows the most accurate discrimination between CDSs and hypothetical ORFs. The raw and probability scores are less informative ($\overline{\text{ROC}}_{0.1}$ of 0.81 and 0.88).

The vote score may be used to divide Glimmer(ct) predictions into probably correct and less certain ones. Determination of the lowest vote score setting with which the maximum specificity (within 2 digits) could be obtained resulted in a threshold setting of 400 (Figure 1.4b). 99% of the predicted CDSs with vote scores $\geq$ 400 are correct predictions, which covers 56% of all annotated CDSs (Table 1.3). The remaining, lower scoring Glimmer(ct) predictions contain a high percentage of false positives, which makes their manual validation seem especially important.

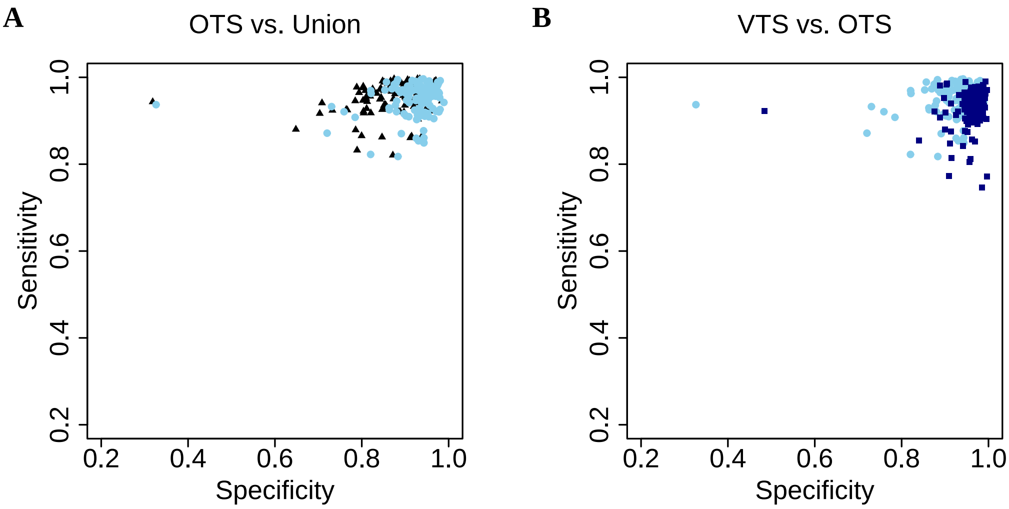


Figure 1.5: Performance of the combined strategies. **A, B:** Sensitivity versus specificity for OTS (light blue circles) versus the union set (black triangles) and versus VTS (dark blue squares).

### 1.3.6 Development of combined strategies

Typical for bacterial genome sequences, the number of hypothetical ORFs largely exceeds the number of annotated CDSs. For the analyzed genomes, the ratio of CDSs to hypothetical ORFs lies between 0.03 (*Mycobacterium leprae*) and 0.15 (*Sulfolobus tokodaii*) for ORFs longer than 90bp. Therefore, it is considerably less effort in annotation for a human annotator to manually discard false positive predictions rather than check for false negatives among the rejected ORFs. A gene prediction strategy based on a combination of different tool results should thus improve the specificity without significantly loosing in sensitivity compared to the individual tools. To achieve this, we pursued the following idea for two parameters with relevance to the gene finding problem: Given the set of highly specific Critica predictions and the set of additional Glimmer(ct) predictions, which contain more false positives but some true positive predictions nevertheless – can parameter settings be determined which allow the removal of mostly the false positive additional Glimmer(ct) predictions? The parameters we focussed on sequentually were the allowed overlap length of additional Glimmer(ct) prediction with Critica ones and the Glimmer(ct) vote score, which was determined to be the most accurate measure for CDS prediction.

The simple union of Critica and Glimmer(ct) predictions did not result in any significant change of performance compared to Glimmer(ct), as the set of Critica predictions is almost completely contained in the Glimmer(ct) ones (Table 1.1). With the Overlap Threshold

Table 1.3: Using the Glimmer vote score to divide predictions into probably correct ones and less certain candidates in need of manual validation. Given is the lowest vote score setting with which the maximal specificity could be obtained.

| Gene finder | $CC(P,A)$ | Sensitivity | Specificity |
|---|---|---|---|
| Glimmer(ct) | 0.90 | 0.95 | 0.87 |
| Vote score $> 400$ | | 0.56 | 0.99 |
| OTS | 0.92 | 0.94 | 0.92 |
| Vote score $> 200$ | | 0.91 | 0.97 |

Strategy (**OTS**), additional Glimmer(ct) predictions are discarded if their overlap length with Critica predictions exceeds a given threshold.

For parameter estimation, 15 different settings of maximum allowed overlap length were tried and Glimmer(ct) predictions with more overlap than this removed. The maximal correlation coefficient $CC(P,A)$ was achieved with an allowed overlap length of 10bp. For the individual genomes, the optimal setting was $\leq$ 50bp for 99 genomes and between 100 and 600bp for another 11. Only for three genomes (*Fusobacterium nucleatum*, *Escherichia coli* CFT073 and *Leptospira interrogans*) performance could thus not be increased. To account for genomes where the 10bp setting is too strict, 50 bp was used as the final parameter setting with OTS. This increases specificity (+4%, $p = 6.5 \cdot 10^{-5}$) without significantly loosing in sensitivity (Table 1.1).

The Vote Score Threshold Strategy (**VTS**) uses the Glimmer(ct) vote scores to further improve specificity. Additional Glimmer(ct) predictions are discarded if their vote score is lower than a given threshold setting. For determination of the optimal threshold setting, 15 different settings of the vote score threshold between 0 and 1000 were tried. For 90 of the individual genomes, threshold settings were found which led to a performance improvement. The maximum overall performance was obtained when disregarding all predictions with vote scores $< 100$ (Table 1.1). Using this parameter setting further significantly increases specificity by 4% ($p = 2.67 \cdot 10^{-6}$), but is also associated with some loss in sensitivity (-2%, $p = 0.004$). As disregarding Glimmer(ct) predictions with low vote scores results in some sensitivity loss, these may instead be used to single out ’uncertain’ candidate genes in need of human attention. Determination of the lowest vote score setting for which the set of higher scoring OTS predictions retains the maximum specificity (within 2 digits) led to a threshold of 200. In combination with the Critica predictions, the higher scoring Glimmer(ct) predictions of the OTS strategy cover 91% of the annotated CDSs with an associated probability of 0.97 that these are correct (Table 1.3). The more uncertain additional Glimmer(ct) predictions with lower vote scores remaining with OTS should be given special attention in the manual validation process.

Table 1.4: Comparison to the Yacop and Glimmer ∩ ZCURVE combined strategies.

| Gene finder | $CC(P,A)$ | Sensitivity | Specificity |
|---|---|---|---|
| **I** [a] | | | |
| Glimmer | 0.91 ± 0.03 | 0.97 ± 0.01 | 0.87 ± 0.04 |
| Yacop | 0.96 ± 0.01 | 0.98 ± 0.01 | 0.95 ± 0.02 |
| OTS | 0.94 ± 0.02 | 0.97 ± 0.02 | 0.91 ± 0.03 |
| VTS | 0.96 ± 0.01 | 0.95 ± 0.01 | 0.97 ± 0.01 |
| **II** [b] | | | |
| Glimmer | 0.82 ± 0.11 | 0.94 ± 0.05 | 0.75 ± 0.14 |
| ZCURVE ∩ Glimmer | 0.94 ± 0.02 | 0.97 ± 0.01 | 0.92 ± 0.03 |
| OTS | 0.95 ± 0.02 | 0.96 ± 0.01 | 0.94 ± 0.04 |
| VTS | 0.96 ± 0.01 | 0.95 ± 0.00 | 0.97 ± 0.01 |

[a] For the seven genome data set used in [28].

[b] For the four genome data set used in [23].

### 1.3.7 Performance evaluation of OTS and VTS

Both OTS and VTS exhibit a significant performance improvement compared to the Glimmer standard application. For OTS, the specificity is improved (+8%, $p = 2.8 \cdot 10{-}08$) without loosing significantly in sensitivity (0%, $p = 0.45$), this is also true for the function-known or otherwise confirmed subsets of genes (+4% in performance, +7% in specificity, no loss in sensitivity). VTS is even more specific (+11%, $p = 2.4 \cdot 10{-}17$), but has some loss in sensitivity (-2%, $p = 3.4 \cdot 10{-}4$). For the function-known subsets of genes, there is no significant sensitivity loss with VTS. For both strategies, the performance improvement is most pronounced for the GC-rich genomes (Table 1.5). As an example, the number of false positive predictions for the *Sinorhizobium meliloti* chromosome is reduced from 1507 for Glimmer to 100 / 47 with OTS and VTS.

For the seven genomes used in that study, the performance of VTS is similar to that described for Yacop, which uses a Critica $\cup$ (Glimmer $\cap$ ZCURVE) combination of gene finding results [28]. OTS performs slightly worse (-2%). But of the seven analyzed genomes, only one has a GC-content $> 50\%$. Compared to a Glimmer ∩ ZCURVE strategy evaluated on a four genome data set with two GC-rich genomes [23], both OTS and and VTS perform better (Table 1.4).

Table 1.5: Sensitivity and false positive proportion of predictions (1 - Specificity) for Glimmer, Glimmer(ct), OTS and VTS for 27 genome sequences with a GC-content $> 0.56$.

| Organism | GenBank | Glimmer | | Glimmer(ct) | | OTS | | VTS | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc. No. | Sens. | 1- Spec. | Sens. | 1- Spec. | Sens. | 1- Spec. | Sens. | 1- Spec. |
| *D. radiodurans* | AE000513 | 2521 (0.98) | 1107 (0.31) | 2483 (0.96) | 517 (0.17) | 2423 (0.94) | 156 (0.06) | 2415 (0.94) | 135 (0.05) |
| *M. tuberculosis* | AE000516 | 3910 (0.93) | 758 (0.16) | 3873 (0.93) | 621 (0.14) | 3780 (0.9) | 300 (0.07) | 3671 (0.88) | 209 (0.05) |
| *D. radiodurans* | AE001825 | 338 (0.95) | 124 (0.27) | 334 (0.94) | 83 (0.2) | 330 (0.92) | 28 (0.08) | 329 (0.92) | 25 (0.07) |
| *P. aeruginosa* | AE004091 | 4814 (0.87) | 3323 (0.41) | 5375 (0.97) | 1315 (0.2) | 5323 (0.96) | 165 (0.03) | 5303 (0.95) | 135 (0.02) |
| *C. crescentus* | AE005673 | 3584 (0.96) | 1156 (0.24) | 3476 (0.93) | 584 (0.14) | 3427 (0.92) | 175 (0.05) | 3404 (0.91) | 155 (0.04) |
| *C. tepidum* | AE006470 | 2013 (0.89) | 452 (0.18) | 1942 (0.86) | 352 (0.15) | 1912 (0.85) | 116 (0.06) | 1829 (0.81) | 80 (0.04) |
| *A. tumefaciens* | AE008688 | 2579 (0.93) | 846 (0.25) | 2548 (0.91) | 620 (0.2) | 2520 (0.9) | 193 (0.07) | 2506 (0.9) | 134 (0.05) |
| *A. tumefaciens* | AE008689 | 1753 (0.93) | 489 (0.22) | 1721 (0.92) | 373 (0.18) | 1708 (0.91) | 100 (0.06) | 1698 (0.91) | 61 (0.03) |
| *B. melitensis* | AE008917 | 1926 (0.94) | 688 (0.26) | 1895 (0.92) | 461 (0.2) | 1878 (0.91) | 134 (0.07) | 1858 (0.9) | 58 (0.03) |
| *B. melitensis* | AE008918 | 1061 (0.93) | 293 (0.22) | 1055 (0.93) | 245 (0.19) | 1039 (0.91) | 72 (0.06) | 1037 (0.91) | 28 (0.03) |
| *X. campestris* | AE008922 | 4083 (0.98) | 2033 (0.33) | 4010 (0.96) | 943 (0.19) | 3946 (0.94) | 231 (0.06) | 3933 (0.94) | 122 (0.03) |
| *X. axonopodis* | AE008923 | 4160 (0.96) | 2320 (0.36) | 4036 (0.94) | 1105 (0.21) | 3942 (0.91) | 254 (0.06) | 3931 (0.91) | 113 (0.03) |
| *M. kandleri* | AE009439 | 1660 (0.98) | 322 (0.16) | 1661 (0.98) | 269 (0.14) | 1639 (0.97) | 180 (0.1) | 1620 (0.96) | 117 (0.07) |
| *B. suis* | AE014291 | 1913 (0.9) | 677 (0.26) | 1828 (0.86) | 456 (0.2) | 1819 (0.86) | 147 (0.07) | 1781 (0.84) | 112 (0.06) |
| *B. suis* | AE014292 | 1033 (0.9) | 325 (0.24) | 1009 (0.88) | 275 (0.21) | 999 (0.87) | 122 (0.11) | 973 (0.85) | 95 (0.09) |
| *B. longum* | AE014295 | 1612 (0.93) | 625 (0.28) | 1592 (0.92) | 482 (0.23) | 1592 (0.92) | 217 (0.12) | 1589 (0.92) | 174 (0.1) |
| *P. putida* | AE015451 | 5240 (0.98) | 1768 (0.25) | 5099 (0.95) | 1107 (0.18) | 5063 (0.95) | 263 (0.05) | 5006 (0.94) | 213 (0.04) |
| *P. syringae pv. tomato* | AE016853 | 5253 (0.96) | 1359 (0.21) | 5174 (0.95) | 959 (0.16) | 5152 (0.94) | 346 (0.06) | 5059 (0.92) | 254 (0.05) |
| *R. solanacearum* | AL646052 | 2747 (0.8) | 2204 (0.45) | 3227 (0.94) | 766 (0.19) | 3182 (0.93) | 64 (0.02) | 3160 (0.92) | 40 (0.01) |
| *M. loti* | BA000012 | 6649 (0.98) | 2468 (0.27) | 6457 (0.96) | 1460 (0.18) | 6348 (0.94) | 320 (0.05) | 6257 (0.93) | 180 (0.03) |
| *C. efficiens* | BA000035 | 2740 (0.93) | 738 (0.21) | 2713 (0.92) | 488 (0.15) | 2670 (0.91) | 94 (0.03) | 2656 (0.9) | 54 (0.02) |
| *B. japonicum* | BA000040 | 7930 (0.95) | 3971 (0.33) | 7665 (0.92) | 2337 (0.23) | 7563 (0.91) | 765 (0.09) | 7528 (0.91) | 431 (0.05) |
| *Halobacterium* | HSPNRC1XX | 1990 (0.97) | 793 (0.28) | 1925 (0.94) | 446 (0.19) | 1871 (0.91) | 89 (0.05) | 1844 (0.9) | 74 (0.04) |
| *M. leprae* | MLEPRAE | 1527 (0.94) | 5438 (0.78) | 1533 (0.94) | 3285 (0.68) | 1526 (0.94) | 3150 (0.67) | 1503 (0.92) | 1603 (0.52) |
| *M. tuberculosis* | MTBH37RV | 3786 (0.97) | 886 (0.19) | 3776 (0.97) | 692 (0.15) | 3710 (0.95) | 385 (0.09) | 3687 (0.94) | 215 (0.06) |
| *S. coelicolor* | SCO645882 | 4546 (0.58) | 5817 (0.56) | 7393 (0.95) | 1473 (0.17) | 7165 (0.92) | 165 (0.02) | 7114 (0.91) | 105 (0.01) |
| *S. meliloti* | SME591688 | 3249 (0.97) | 1507 (0.32) | 3237 (0.97) | 872 (0.21) | 3227 (0.97) | 100 (0.03) | 3201 (0.96) | 47 (0.01) |

## 1.4 Conclusion

This work describes the development of joint application strategies for two microbial gene finders, which combine the strengths of both tools to improve the overall gene finding performance. The comparative sequence analysis approach Critica employs ensures its high specificity in the detection of similarity-supported genes. In the interpretation of the results of pairwise DNA sequence comparisons, Critica makes use of the degeneracy of the genetic code to discriminate conserved coding from conserved non-coding regions [25]. Similar approaches are also increasingly becoming popular in the field of eukaryotic gene prediction [38]. Compared to approaches which use similarity on amino acid level, an advantage is that it does not depend on existing accurate annotation, which is used to generate the contents of the protein sequence databases. If using comparisons on amino acid level, genes may be missed whose homologs have not been annotated or annotated to short. In our analyses, we found Critica to be very robust. It performs well on sequences with a high GC-content and also on the *Mycobacterium leprae* genome, which contains a large number of pseudogenes. Its strength is its high specificity, which is also evident in the detection of function-known or otherwise confirmed genes. It also is the most specific in predicting short genes.

The gene finder Glimmer completely relies on an *ab initio* approach in gene identification. It uses a very sophisticated model of sequence properties of prokaryotic CDSs [20]. It is highly sensitive, also in the detection of genes supported by additional evidence. For GC-rich genomes, it strongly looses in prediction performance, which is mainly due to a specificity loss. We found that by using the very specific Critica predictions as a training set for the Glimmer CDS model, performance in terms of both sensitivity and specificity can be significantly improved.

A troublesome issue is the unknown quality of many CDS entries in the current annotation data. The annotation describes the CDS content of a genomic sequence and thus is per definition the standard of truth against which gene finding performance is evaluated. In its creation, considerable human effort is also often involved to achieve a high quality. Still, for no genome all annotated CDSs are supported by experimental or otherwise convincing evidence. A comparison of the length distribution of annotated genes with genes matching a known protein led to the conclusion that many genomes might currently be over-annotated, especially concerning the short genes [36]. Because of the size of the analyzed data set, the results deduced in this study are unlikely to be much influenced by erroneous strategies of individual annotation projects. They were also given further confirmation and found to achieve an even higher sensitivity in validation on the subsets of function-known genes.

In the development of combined gene prediction strategies, the very specific Critica predictions were initially set as fixed and combined with different subsets of additional Glimmer(ct) predictions to improve the overall performance. For specification of this additional subset, two different parameters with relevance to the gene finding problem were evaluated. The first is the allowed overlap length of neighboring genes, as genes of longer overlap length are generally considered unlikely for prokaryotic organisms, although there is no systematic research to this issue yet. From a biological perspective this may be explained by the extreme constraints which are placed on a sequence which is coding in two different frames. We found that by removing additional predictions with long overlaps, the specificity in gene identification can be considerably improved without a significant loss of sensitivity. The second parameter is the Glimmer(ct) vote score, which was determined to be the Glimmer scoring method that allows the most accurate discrimination between hypothetical ORFs and CDSs. Discarding low vote score predictions results in a further gain in specificity, but is accompanied by a slight sensitivity loss. Interestingly, there is no significant sensitivity loss of VTS for the subsets of function-known or otherwise classified as more reliable genes. The additional genes missed by VTS thus are both low-scoring according to sequence composition and without indication of function or biological activity, according to the annotation data. They are either falsely annotated or real genes which are difficult to determine, such as the genes contained in prophage DNA. Using OTS allows considerable reduction of the necessary manual validation effort of the gene finding results for the human annotators, especially for GC-rich genomes. As an example, with OTS the false positive prediction rate for the *Sinorhizobium meliloti* chromosome is reduced from 32% for Glimmer to 2%, without a loss of sensitivity.

The described methods have been implemented within the GenDB genome annotation system and are currently being applied in several bacterial genome projects. We hope that the software and additional information presented in this work will be helpful to annotators in producing a high quality genome annotation.

CHAPTER 2

# CoBias - Using log-odds ratio scores for classification based on trends in synonymous codon usage

Recent studies have shown synonymous codon usage in microbial organisms to be related to a number of different factors. Among these are genomic GC-content, position on leading / lagging strand, gene expression level or growth at high temperatures. Furthermore, interspecies comparisons confirmed it to be quite specific for the genes of a genome. Thus, such features may be used to predict properties such as an alien origin or a high expression level of a gene. Different methods have been defined to specifically solve one of these tasks.

Here, a probabilistic modeling approach to classification based on trends in synonymous codon usage is described, which has been implemented with the CoBias program. The method has several favorable properties. It is generally applicable to any two-class discrimination problem which may be solved using codon usage properties, such as discrimination between highly versus not highly expressed, leading versus lagging strand, genome $x_a$ versus genome $x_b$ or real versus hypothetical genes. It also places the analysis of codon usage features on a firm statistical basis, which allows estimation of the strength and significance of an observed feature using standard procedures.

## 2.1 Introduction

The synonymous codon usage in prokaryotic genomes has been related to a wide variety of factors [39], such as expression level of a gene [40, 41, 42], overall GC-content [43], growth at high temperatures [44] or the location on leading or lagging strand [45]. Several methods have been defined for classification of genes into different categories based on such features. Examples for specifically defined measures for the prediction of highly expressed genes are the Codon Adaptation Index (CAI) [40] and the PHX classification [46]. For the detection of putative alien (PA) genes which originate from horizontal gene transfer (HGT) events, the PA classification [47] has been introduced. In combination with additional criteria based on evaluating GC-content, amino acid composition or the genomic gene position, the $\chi^2$ [48] and Mahalanobis distance [49] are also used for the detection of potential horizontally transferred genes by codon usage properties. In the following, the different formalisms are briefly described.

### 2.1.1 The Codon Adaptation Index (CAI)

Sharp and Li [40] define the Codon Adaptation Index, which is the geometric average of 'relative adaptiveness' values for the codons of a gene. Let $I_j$ be the set of all synonymous codons $c_i$ encoding amino acid $a_j$, with $i \in I_j$ and $a_j$ being one of the 20 natural amino acids for $j = 1, \dots, 20$. The relative adaptiveness values for the 61 amino acid encoding codons (stop codons excluded) are calculated using a reference set of highly expressed genes $H$. Let $f_{i,j}$ be the frequency of codon $c_i$ and $f_{max,j}$ the frequency of the most often used codon for the amino acid $a_j$ in $H$. The relative adaptiveness value $w_{i,j}$ for $c_i$ is defined as:

$$w_{i,j} := \frac{f_{i,j}(H)}{f_{max,j}(H)} \tag{2.1}$$

Given a gene sequence $g$ of $l$ codons, let $w_{i_k,j_k}$ be the relative adaptiveness value for codon $c_i$ encoding $a_j$ at position $k$ in the sequence. The CAI for $g$ is defined as:

$$\text{CAI} := \Big(\prod_{k=1}^{l} w_{i_k,j_k}\Big)^{1/l} \tag{2.2}$$

CAI values range from 0 to 1, with higher values indicating a higher similarity in codon usage to the reference of highly expressed genes.

### 2.1.2 The PHX and PA classifications

Karlin *et al.* predict highly expressed (PHX) [46] and putative alien (PA) genes [47] using a formalism based on the weighted total variation distance [50]. Let $R$ be a reference set of genes corresponding to a subset of all annotated genes of a genome and $g$ be an arbitrary gene. The 'bias' of $g$ relative to $R$ is

$$B(g|R) := \sum_{j=1}^{20} p_j(g) \sum_{i\in I_j} |f_{i,j}(g) - f_{i,j}(R)| \tag{2.3}$$

where $p_j(g)$ is the relative frequency of amino acid $a_j$, and $f_{i,j}(g)$ and $f_{i,j}(R)$ are relative frequencies of the synonymous codons $c_i \in I_j$ in $g$ and a reference set $R$, respectively. The utilized reference sets $R$ are the gene classes $C$ (all protein coding genes), $RP$ (ribosomal protein encoding genes), $CH$ (chaperone-degradation encoding genes) and $TF$ (translation and transcription processing factor encoding genes). $B(g|R)$ is used in the general expression measure

$$E(g) := \frac{B(g|C)}{0.5\cdot B(g|RP) + 0.25B\cdot(g|CH) + 0.25\cdot B(g|TF)} \tag{2.4}$$

and the class-specific expression measures

$$E_{RP}(g) := \frac{B(g|C)}{B(g|RP)}, \quad E_{CH}(g) := \frac{B(g|C)}{B(g|CH)}, \quad E_{TF}(g) := \frac{B(g|C)}{B(g|TF)} \tag{2.5}$$

Large expression values result if the codon usage of a gene is more similar to one of the classes of highly expressed genes than to the average usage of a genome.

**Definition I.** A gene is predicted highly expressed (PHX) if two or more of the class-specific expression values exceed 1.05 and if $E(g) \geq 1.00$ holds.

**Definition II.** A gene is putative alien (PA) if $B(g|RP) > M + 0.15$, $B(g|CH) > M + 0.15$, $B(g|TF) > M + 0.15$ and $B(g|C) > M + 0.12$ with $M$ being the median of $B(g|C)$ for all annotated genes $g \in C$ of a genome.

### 2.1.3 The $\chi^2$ statistic

Lawrence and Ochman predict horizontally transferred genes based on the GC-content of the first and third codon position [48]. In addition, synonymous codon usage is evaluated by calculating CAI and $\chi^2$ values of codon usage. Using the 'goodness-of-fit' test

$$\chi^2(g) := \sum_{j=1}^{20} \sum_{i\in I_j} \frac{(Y_{i,j}(g) - n_j(g)p_j)^2}{n_j p_j} \tag{2.6}$$

the deviation of a gene $g$ from the uniform usage of all synonymous codons is determined. Here, $Y_{i,j}(g)$ is the number of occurrences of codon $c_i$ encoding amino acid $a_j$ and $n_j$ is the number of occurrences of amino acid $a_j$ in $g$. According to the null hypothesis that synonymous codon usage has a uniform distribution, the probabilities for synonymous codons of amino acid $a_j$ are $p_j = \frac{1}{|I_j|}$. Genes with both high $\chi^2$ and CAI values are subsequently discarded, as these are likely to display atypical sequence properties due to expression-level dependent features.

### 2.1.4 The Mahalanobis distance

Garcia-Vallvé *et al.* use the Mahalanobis distance to detect atypical genes which deviate from the genomic average in terms of codon usage [49]. In combination with criteria based on evaluation of GC-content, amino acid composition, and gene position, atypical genes satisfying all conditions are classified as originating from horizontal gene transfer. The genes predicted as putative horizontally transferred genes with this method make up the content of the HGT-DB [51]. The Mahalanobis distance is

$$d^M\big(X(g),X(C)\big)^2 = \big(X(g)-X(C)\big)^T S^{-1}\big(X(g)-X(C)\big) \tag{2.7}$$

with $X(g)$ and $X(C)$ being codon usage vectors of 61 dimensions (stop codons excluded) of the relative frequencies of the codons for a gene $g$ and all annotated genes $C$ for the organism. $T$ is the transposition operator and $S^{-1}$ the inverse matrix of the 61 $\times$ 61 covariance matrix

$$S_{m,n} := \sum_{k=1}^{|C|} [X_m(g_k)-X_m(C)][X_n(g_k)-X_n(C)] \quad (m,n=1,2,\ldots,61) \tag{2.8}$$

where $|C|$ is the number of genes of the organism.

### 2.1.5 Motivation

For the evaluation of features in synonymous codon usage, a simple scoring model is introduced and implemented with the CoBias program, which is also used for the probabilistic interpretation of the scores obtained in pairwise sequence alignments [52]. The method is supported by a firm statistical basis and places the analysis of codon usage features within such a framework. This also allows estimation of the strength and significance of an observed feature using standard procedures. To assess the significance, a Bayesian approach of model comparison is applied [53]. In [42], we show the method to be well suited for the analysis of expression level-dependent features in codon usage. It is furthermore generally applicable to any other two-class discrimination problem which may be solved using codon usage properties, such as discrimination between leading versus lagging strand, genome $x_a$ versus genome $x_b$, or real versus hypothetical genes.

## 2.2 A probabilistic method for the evaluation of synonymous codon usage features

### 2.2.1 Assessing a trend in codon usage

Assume the analysis of a biological property which is related to synonymous codon usage is of interest – for example genes originating from genome $x_a$ are to be discriminated from the genes of another organism's genome $x_b$. This can be done by creating a probabilistic model of the synonymous codon usage differences between these two genomes. Subsequently, the codon usage of a gene sequence $g$ is scored using this model. The observed score can be seen as a likelihood ratio, which reflects the relative odds that the gene originates from $x_b$ as opposed to originating from $x_a$.

Initially, a log-odds ratio scoring matrix is created, which reflects the differences in synonymous codon usage between two references sets of genes. Let $a_j$ be one of the 20 natural amino acids or the stop symbol $*$. We define a codon usage model $M$ for a single reference set of genes to consist of the probabilities $p_{i,j}$ for the synonymous codons $c_i$ encoding an amino acid $a_j$, such that $\sum_{i\in I_j} p_{i,j} = 1$ holds for all $j$. Given a target model $T = \{p_{i,j}(t)\}$ derived from a target reference set $t$ and a background model $B = \{p_{i,j}(t)\}$ derived from a background reference set $b$, log-odds ratio scores

$$s_{i,j} := \ln \frac{p_{i,j}(t)}{p_{i,j}(b)} \quad \text{for } j = 1,\ldots,21 \text{ and } i \in I_j \tag{2.9}$$

can be calculated for the synonymous codons $c_i$ of amino acid $a_j$. The values of $s_{i,j}$ are stored in a $21 \times 64$ scoring matrix $M_{T,B}$, setting $s_{i,j} = 0$ for all $i \notin I_j$, $j = 1,\ldots,21$. The matrix $M_{T,B}$ may be seen as a probabilistic model which represents codon usage differences between the $T$ and $B$. The parameters of $T$ and $B$ can be estimated from the relative frequencies of the synonymous codons occurring in target and background reference sets of genes, respectively.

For the evaluation of synonymous codon usage properties of a gene sequence $g$, let $a_j$ be one of the evaluated amino acids or the stop symbol, with $j = 1,\ldots,N, N \leq 21$. Unless specified otherwise, the stop symbol (*) and the amino acids methionine (M) and tryptophane (W), which are encoded by only one codon are excluded and thus $N = 18$. The number of occurrences of codon $c_i$ encoding amino acid $a_j$ in $g$ is denoted $n_{i,j}(g)$. Given a codon usage model $M$, the likelihood $P(g|M)$, which is the probability that $M$ assigns to the occurrence of $g$, can be calculated:

$$P(g|M) := \prod_{j=1}^{N} \prod_{i\in I_j} (p_{i,j})^{n_{i,j}(g)} \tag{2.10}$$

The log-likelihood ratio

$$S(g) := \ln \frac{P(g|T)}{P(g|B)} \tag{2.11}$$

reflects the relative odds that the target $T$ as opposed to the background model $B$ correctly represents the codon usage of $g$. According to the Neyman-Pearson lemma [54], the likelihood ratio is the best test statistic for testing a hypothesis $H_0$ against an alternative hypothesis $H_1$, which maximizes the power of the test. Substituting $P(g|T)$ and $P(g|B)$ using 2.10, we find:

$$S(g) = \sum_{j=1}^{N} \sum_{i \in I_j} n_{i,j}(g) \cdot s_{i,j} \tag{2.12}$$

$S(g)$ can thus be calculated by summation of the log-odds ratio scores for all codons of the evaluated amino acids (or the stop symbol) of $g$.

### 2.2.2 Significance estimation

The significance of an obtained score $S(g)$ can be estimated with the posterior probability $P(B|g)$ via a Bayesian approach of model comparison [53]. This is the probability of the model $B$, given the observed gene sequence of synonymous codons $g$. According to Bayes' rule,

$$P(B|g) = \frac{P(g|B) \cdot P(B)}{P(g)} \tag{2.13}$$

Assuming that either $T$ or $B$ is the correct model for $g$, $P(g)$ can be written as the sum of the joint probabilities:

$$P(g) = \sum_{X \in \{B,T\}} P(g,X) = \sum_{X=B,T} P(g|X) \cdot P(X) \tag{2.14}$$

Substitution of $P(g)$ using 2.14 yields:

$$P(B|g) = \frac{P(g|B) \cdot P(B)}{P(g|B) \cdot P(B) + P(g|T) \cdot P(T)} \tag{2.15}$$

By division with $P(g|T) \cdot P(T)$ and using 2.11, we obtain:

$$P(B|g) = \frac{e^{-S(g)} \cdot \frac{P(B)}{P(T)}}{e^{-S(g)} \cdot \frac{P(B)}{P(T)} + 1}. \tag{2.16}$$

The prior odds ratio $\frac{P(T)}{P(B)}$ reflects the *a priori* expectation $P(B)$ and $P(T) = 1 - P(B)$ for the different models to correctly represent $g$. For significance estimation with $P(B|g)$ values, this parameter needs to be specified in advance.

### 2.2.3 The 'strength' of a feature

Both $S(g)$ and $P(B|g)$ are length-dependent. For estimation of the ’strength’ of an observed feature, the average degree of evidence per codon in favor of model $T$ is assessed by normalizing the log-likelihood ratio

$$S_{AV}(g) := \frac{S(g)}{l} \tag{2.17}$$

with the length $l$. Genes more similar in codon usage to the target than the background model have normalized log-likelihood scores $> 0$.

### 2.2.4 Information content of the scoring matrix

Finding genes with a higher similarity in codon usage to one set than the other requires a significant difference in codon usage between these two sets. The codon usage difference of the target relative to the background model is assessed with the relative entropy or Kullback-Leibler distance [55]

$$H(T||B) := \frac{1}{N} \cdot \sum_{j=1}^{N} \sum_{i \in I_j} p_{i,j}(t) \cdot \ln \frac{p_{i,j}(t)}{p_{i,j}(b)} \tag{2.18}$$

normalized per amino acid. We call this the information content of the scoring matrix (measured in *nats* - natural digits). $H(T||B)$ is always $\geq 0$ and only zero in case the codon usage of the two references is identical.

## 2.3 Implementation

CoBias is a user-level program that allows application of log-odds ratio scoring, calculation of CAI values [40] and a simplified general expression measure [41] for the analysis of synonymous codon usage features of DNA sequences. The described methodology along with the additional functionality necessary for the parsing of sequence files and feature extraction from the DNA sequences has been implemented in a number of modules in the programming language PERL [56].

### 2.3.1 Parameters of CoBias

For the computation of values for a set of DNA sequences, CoBias requires the specification of a matrix file of log-odds ratio scores of codon usage. Optionally, target and background reference sets of sequences can directly be specified for the calculation of log-odds ratio scores. The accepted sequence format for all input files is the multiple fasta format. Additional parameters which can be defined for the program are the following:

-l Minimum sequence length in codons. The default is a length of 50.

-g Set genetic code. The default is the standard genetic code 1.

-t <Seq-file> Specify a set of sequences for computation of synonymous codon usage of the target reference.

-b <Seq-file> Specify a set of sequences for computation of synonymous codon usage of the background reference.

-m <file> Specify a matrix file of log-odds ratio scores of codon usage. Such files can be created and saved during a CoBias run where target and background reference sets of sequences have been specified. This allows their reuse in subsequent program runs.

-c Set a minimum number of synonymous codons for an amino acid which must be present in a sequence to proceed with further evaluation. The default setting is 0.

-p Specify the prior odds ratio $\frac{P(T)}{P(B)}$to be used for calculation of the posterior probability value $P(B|g)$. The default setting is 0.05.

-a <string> The amino acids or stop symbol (*) which are to be excluded from the calculation of log-likelihood ratio scores. These are specified concatenated in a string, the default is “MW*”.

-s Save the log-odds ratio scoring matrix.

-f <val> Use a different method for codon bias calculations.
1: Calculate CAI [40] values relative to a specified target reference set of sequences.
2: Calculate the simplified expression measure $E(g)$ [41], similar to the measures used in the PHX classification. This requires the specification of target and background reference sets of sequences.

## 2.4 Discussion

There is one important difference of the presented method to the Codon Adaptation Index, the PA classification, the Mahalanobis distance and the $\chi^2$-test. The latter measure the absolute deviation of a gene in terms of codon usage from one or more reference sets. For deviant genes, they do not indicate to what other kind of reference the similarity is higher. Thus, for instance, atypical genes with respect to a genome-specific reference can be determined. For the prediction of 'alien' genes from this set, an additional filtering step is applied to remove highly expressed genes which deviate from the genomic average due to expression level-dependent features in codon usage [48, 47]. To estimate gene expression levels, deviation from a reference set of highly expressed genes is commonly measured using the CAI for *Escherichia coli* genes [57, 58]. For genomes with a very biased base composition, this procedure can lead to wrong conclusions [39] as other genome-specific properties may contribute significantly to the synonymous codon usage of the highly expressed genes.

Instead of measuring the deviation from one reference, log-odds ratio scores of codon usage model the differences in synonymous codon usage which can be observed between two sets of genes. The method is generally applicable to any two-class discrimination problem which can be solved using features in synonymous codon usage. Applied to the detection of highly expressed genes and estimation of gene expression rates, it has the advantage that common features in sets of highly and not highly expressed genes, such as the influence of a skewed genomic base composition are discarded in the modeling procedure. Using a classification by relative similarity to different reference sets of genes, the PHX classification applies a similar approach, but does not allow significance estimation for an observed feature.

For the evaluation of expression level-dependent features in synonymous codon usage, the method thus has several favorable properties. This is demonstrated by a comparison with experimental data in chapter 3. The analysis of such features in codon usage can also be of great practical interest to the experimental researcher. To obtain satisfactory expression levels of recombinant protein, it sometimes is necessary to improve the 'translational fitness' of a recombinant gene sequence with respect to the utilized expression system. For *Escherichia coli*, many examples can be found where optimizing synonymous codon usage resulted in increased expression rates of recombinant protein [59, 60, 61, 62].

In [63], the CoBias program has been used to determine the strength of expression level-dependent features in the codon usage of the surface (S)-layer gene *cpsB* from 28 *Corynebacterium glutamicum* strains. S-layer genes encode the building blocks of the outermost cell wall layer and are known to be very highly expressed prior to cell division. The gene product makes up approximately 20% of the total protein content of the cell [64, 65]. For this, a matrix of expression level-dependent features in synonymous codon usage was created from the annotation data of the *C. glutamicum* ATCC 13032 genome, a strain which itself does not possess any S-layer genes. In agreement with their high expression level, all 28 genes were found to exhibit very significant evidence of expression level-dependent features in codon usage ($P(B|g) \leq 1.4 \cdot 10^{-38}$, Table 2.1). The strength of these features as estimated using $S_{AV}(g)$ ranges from $0.18 - 0.27$. Only 19 genes with higher values of $S_{AV}(g)$ are annotated for the *C. glutamicum* ATCC 13032 genome. Genes with a similar strength of these features are two very highly expressed *C. glutamicum* genes from glycolysis and the translation machinery (Table 2.1).

Applied to the detection of horizontally transferred genes, the method offers another important innovation. By modeling differences in synonymous codon usage between pairs of genomes, log-odds ratio scores can be used to detect genes with a higher similarity to another organisms' chromosome. This allows the prediction of a donor genome for putative alien genes, which is an innovation for sequence composition-based approaches to the characterization of horizontal gene transfer events. In chapter 4, the performance of this approach is evaluated with a simulation experiment and by analyzing the *Thermotoga maritima* genome, which has been reported to possess genetic material of archaeal origin [66, 67, 68].

Table 2.1: CoBias results for S-layer gene *cspB* from 28 different *Corynebacterium glutamicum* strains. For the analysis, a matrix of expression level-dependent features of synonymous codon usage for *C. glutamicum* ATCC 13032 was used. The lower part of the table contains selected examples of highly expressed genes from the *C. glutamicum* ATCC 13032 genome with similar $S_{AV}(g)$ values.

| $P(B\|g)$ | $S_{AV}(g)$ | $S(g)$ | *C. glutamicum* Strain |
|---|---|---|---|
| 2.09e-59 | 0.27 | 138.11 | DSM 20137 |
| 2.95e-58 | 0.27 | 135.46 | ATCC 31832 |
| 1.51e-56 | 0.26 | 131.53 | ATCC 17965 |
| 5.19e-56 | 0.26 | 130.29 | ATCC 17966 |
| 2.71e-55 | 0.26 | 128.64 | ATCC 14752 |
| 4.12e-55 | 0.26 | 128.22 | ATCC 14068 |
| 1.03e-53 | 0.25 | 125.00 | DSM 447 |
| 4.93e-53 | 0.25 | 123.44 | ATCC 19223 |
| 4.95e-53 | 0.25 | 123.43 | ATCC 14020 |
| 1.09e-52 | 0.24 | 122.64 | 22243 |
| 1.76e-52 | 0.25 | 122.16 | ATCC 14915 |
| 5.55e-52 | 0.24 | 121.02 | 22220 |
| 2.30e-51 | 0.24 | 119.59 | ATCC 14017 |
| 2.22e-50 | 0.24 | 117.33 | ATCC 14067 |
| 1.02e-49 | 0.24 | 115.80 | ATCC 15354 |
| 3.29e-49 | 0.23 | 114.63 | ATCC 15243 |
| 4.36e-48 | 0.23 | 112.05 | ATCC 13745 |
| 4.36e-48 | 0.23 | 112.05 | ATCC 14751 |
| 1.08e-47 | 0.23 | 111.14 | ATCC 21341 |
| 1.44e-47 | 0.22 | 110.86 | DSM 46307 |
| 1.64e-47 | 0.23 | 110.73 | ATCC 14747 |
| 2.43e-47 | 0.23 | 110.33 | DSM 20598 |
| 5.17e-47 | 0.22 | 109.57 | ATCC 19240 |
| 8.48e-47 | 0.22 | 109.08 | ATCC 13058 |
| 1.14e-46 | 0.22 | 108.78 | ATCC 21645 |
| 9.34e-46 | 0.22 | 106.68 | ATCC 31380 |
| 1.28e-45 | 0.22 | 106.37 | ATCC 13744 |
| 1.40e-38 | 0.18 | 90.16 | ATCC 31808 |
| **Reference genes** | | | **Gene name** |
| 3.22e-48 | 0.27 | 112.35 | *eno*, enolase (EC 4.2.1.11) |
| 1.49e-15 | 0.18 | 37.13 | *rpsE* 30S ribosomal protein S5 |

CHAPTER 3

# Comparing expression-level dependent features in codon usage with protein abundance: An analysis of 'predictive proteomics'

Synonymous codon usage is a commonly used means for estimating gene expression levels of *Escherichia coli* genes and has also been used for predicting highly expressed genes for a number of prokaryotic genomes. By comparison of expression level-dependent features in codon usage with protein abundance data from two proteome studies of exponentially growing *E. coli* and *B. subtilis* cells, we try to evaluate whether the implicit assumption of this approach can be confirmed with experimental data. Log-odds ratio scores are used to model differences in codon usage between highly expressed genes and genomic average. Using these, the strength and significance of expression level-dependent features in codon usage were determined for the genes of the *Escherichia coli, Bacillus subtilis* and *Haemophilus influenzae* genomes. The comparison of codon usage features with protein abundance data confirmed a relation between these to be present, although exceptions to this, possibly related to functional context, were found. For species with expression level-dependent features in their codon usage, the applied methodology could be used to improve *in silico* simulations of 2-D gel electrophoretic experiments.

## 3.1 Introduction

The choice of synonymous codons within the coding sequences of a genome is known to be non-random and thought to reflect a balance among the forces of selection, mutation and random genetic drift [69, 70]. Since early studies of *Escherichia coli* and *Saccharomyces cerevisiae* genes, codon usage in these organisms has been found to exhibit a bias towards some ’preferred’ or ’major’ codons, with the extent of the bias being related to the expression level of a gene [71, 72]. Although the set of preferred codons can differ [73], preferred codons in organisms where codon usage strongly represents expression level-dependent features have been found to be those recognized by the most abundant tRNA for each amino acid or have perfect Watson-Crick pairing [71, 74, 72, 75, 76]. The relation of gene expression levels, relative tRNA abundance and the strength of codon bias has been explained with the presence of a selective force for translational ’efficiency’ [77], meaning that codon choice by speed and nature of the interaction with the cognate tRNA is thought to influence the speed and accuracy of the translation process. Besides expression level-dependent features, recent studies on multiple microbial genomes have determined additional factors which affect codon usage in microbial genomes [43, 78, 44]. These include forces related to generating an organisms GC-content [43, 39, 44], growth at high temperatures [44], strand-specific forces [78, 39], gene length [79], the context of bases surrounding each codon [80, 81, 57, 82] and the position in a gene [83, 84, 85].

Classification of genes by codon usage similarity relative to a reference set of highly expressed genes has for some species been utilized as predictive indicator of gene expression levels [86, 87, 82, 57, 58]. Commonly used measures for evaluating codon usage similarity or bias relative to a reference are e.g. the Codon Adaptation Index (CAI) [40] or the ’frequency of optimal codons’ [72]. In an approach to use codon usage differences between different reference sets of genes to predict highly expressed genes for a number of microbial genomes, Karlin *et al.* [50] used a weighted version of the total variation distance as measure of codon bias. Predicted highly expressed (PHX), by their definition, are genes which are more similar in codon usage relative to a number of reference sets of highly expressed genes than to the average codon usage of the genome [46].

Expression level-dependent features in codon usage have been determined to influence the codon usage of many of the complete microbial genomes available today. The above mentioned predictive approaches rely on the implicit assumption that these features can be taken as representative for gene expression levels. Here, we use a log-odds ratio scoring approach to create a model of expression level-dependent features in codon usage. The model only represents codon usage differences between highly expressed genes and genomic average for a given genome and thus excludes expression level-independent features of codon usage, which are present in both references. Using this, the strength and significance of these features in codon usage is determined for the genes of three bacterial genomes. By comparison of these results with protein abundance data from *Escherichia coli* and *Bacillus subtilis* proteome studies, the relation of expression level-dependent features in codon usage and protein abundance is evaluated. As a possible application of the methodology, we explore its use for the creation of a more realistic virtual 2-D gel [88], an *in silico* simulation of a 2-D gel electrophoretic experiment.

## 3.2 Materials and Methods

### 3.2.1 Data sets

The sets of coding sequences were derived from the current EMBL annotations of the *Escherichia coli* K-12 [89], *Bacillus subtilis* [90] and *Haemophilus influenzae* [1] genomes. To exclude annotation ambiguities, coding sequences annotated with a non-integer number of codons were discarded. Functional classification is based on the information given in the annotations. The category ’Protein biosynthesis’ (Table 3.4 and Figure 3.3) combines genes of the categories ’Transcription’, ’Translation factors’, ’tRNA synthetases’, and ’ribosomal proteins’ listed in Table 3.3. For comparison with protein abundances, relative estimates of absolute protein abundance were obtained from [91] and [92],which correspond to percent of total protein visualized on the 2-D gel using fluorescence staining [91] or silver-staining [92] methods. The *E. coli* data set published in [92] comprises estimates of protein abundance for 173 *E. coli* proteins which were determined in the cell extract of *E. coli* cells in late exponential growth. For proteins occurring in more than one spot, abundance estimates of all spots containing the same protein were added. The image displaying a silver-stained 2-D gel of *E. coli* cell extract (Figure 3.4) was published in the same work and could be obtained as electronic version from the Swiss-2DPAGE database (`http://www.expasy.ch/ch2d/`). The *B. subtilis* data set contains estimates of protein abundance for the 50 most abundant proteins determined in a cell extract of exponentially growing *B. subtilis* cells on minimal medium [91]. For comparison with $S_{AV}(g)$ values, CAI values relative to $H$ reference set were calculated for all the genes of the three genomes. The result files and $H$ references used for the three organisms are available at `http://www.Genetik.Uni-Bielefeld.DE/~/alice/PHE/`.

Table 3.1: Information content of the scoring matrix for the three genomes.

| Organism | $H$ reference set[a] | $H(H\|\|N)$ | PHE (Total) |
|---|---|---|---|
| *E. coli* | | | |
| | RPs, TFs, RPol, tRSs | 0.130 | 334 |
| *B. subtilis* | | | |
| | RPs, TFs, RPol, tRSs | 0.072 | 249 |
| *H. influenzae* | | | |
| | RPs, TFs, RPol, tRSs | 0.063 | 186 |

[a] Consists of genes involved in protein biosynthesis which encode ribosomal proteins (RPs), translation factors (TFs), RNA polymerase subunits (RPol) and tRNA synthetases (tRSs).

### 3.2.2 Assessing expression level-dependent features in codon usage

Using the CoBias program (chapter 2), expression level-dependent features in codon usage are modeled with a log-odds ratio scoring matrix which represents differences between codon usage models of highly and not highly expressed genes. The target model $H$ representing the codon usage of highly expressed genes was created from genes involved in protein biosynthesis (ribosomal protein, translation factor, RNA polymerase subunit and tRNA synthetase encoding genes, Table 3.1), similar to the gene sets described in [46]. The background model $N$ representing the codon usage of not highly expressed genes was created from all the coding sequences of the respective genome. The prior odds ratio $\frac{P(H)}{P(N)}$ (0.05) was chosen under consideration of a previous report, which predicted between 4 and 17 % of the genes $\geq 100$ codons of diverse bacterial genomes to be highly expressed [46]. Subsequently $S(g)$, $S_{AV}(g)$ and $P(N|G)$ values were calculated for all coding sequences of the three genomes. Assuming that expression level-dependent features in codon usage can be taken as representative for the expression level of a gene, $P(N|G)$ can be called the probability that $G$ is not highly expressed, according to codon usage evidence. For genes with $P(N|G) > 0.5$ the $N$ codon usage model is more probable to be the right one. Upon inspection of results $P(N|G) = 0.2$ was set as significance threshold to discriminate real features present from random fluctuations in codon usage. Genes with $P(N|G)$ values below this were classified as 'probably highly expressed' (PHE) genes.

## 3.3 Results

The number of ’probably highly expressed’ (PHE) genes identified in the genomes of *E. coli*, *B. subtilis* and *H. influenzae* range from 334 for *E. coli* to 186 for *H. influenzae*. Generally, the most significant evidence of expression level-dependent features in codon usage is attributed to the *E. coli* genes, which is due to the high information content of the scoring matrix (Table 3.1, 3.2).

For all three organisms PHE genes are common in functional categories which contain many genes with ’house-keeping’ functions (Table 3.3), such as energy generation (glycolysis, tricarboxylic acid cycle and respiratory chain), DNA replication and protein biosynthesis, folding and degradation. In addition to this, genes highly expressed during exposure to stress are among the PHE genes. Well known examples from *E. coli* include the strongly induced type I response cold-shock proteins RbfA, NusA, Pnp and CspA [93], proteins participating in response to oxidative stress, such as KatG, SodA, SodB, AhpCF and Dps [94] as well as the major chaperones and chaperonins of heat-shock response (e.g. DnaK, DnaJ, GrpE and GroEL [95]). Table 3.2, in which the genes with the highest $S_{AV}(g)$ values of the three genomes are displayed, demonstrates the wide range of functional categories in which genes with strong expression level-dependent influences in their codon usage can be found. Besides genes encoding ribosomal proteins and translation factors from protein biosynthesis, genes which code for proteins involved in glycolysis (*gapA, eno* from *E. coli*; *gapdh, eno* from *H. influenzae*), detoxification (*ahpC* from *B. subtilis*), cold shock response (*cspA* from *E. coli*) as well as major constituents of the outer membrane of the gram-negative *E. coli* (*lpp, ompC* and *ompA*) are present. Although the set of functional categories covered by the PHE genes is largely identical for the analyzed organisms, the coverage of the proteins in these categories varies.

Table 3.2: Top scoring genes with the highest $S_{AV}(g)$ values

| *Escherichia coli* | | *Bacillus subtilis* | | *Haemophilus influenzae* | |
|---|---|---|---|---|---|
| *lpp* | 0.36 | *rplT* | 0.27 | *rpS15*(HI1468) | 0.21 |
| *gapA* | 0.31 | *tufA* | 0.27 | *gapdh* | 0.19 |
| *eno* | 0.31 | *rplS* | 0.26 | *rpS15*(HI1328) | 0.18 |
| *ompC* | 0.31 | *rpsM* | 0.25 | *eno* | 0.18 |
| *rplL* | 0.31 | *rpsI* | 0.25 | *mopI* | 0.18 |
| *rpmH* | 0.31 | *rpsD* | 0.25 | *tufB* | 0.18 |
| *rpsI* | 0.30 | *sspA* | 0.24 | *rpL32* | 0.18 |
| *tufA* | 0.30 | *rplB* | 0.23 | *tufA* | 0.18 |
| *ompA* | 0.29 | *rpsG* | 0.23 | *rpS9* | 0.17 |
| *cspA* | 0.29 | *ahpC* | 0.23 | *rpL1* | 0.16 |

Table 3.3: Probably highly expressed genes $\geq$ 100 codons in the three genomes

| Functional category | PHE genes |
|---|---|
| ***Escherichia coli*** | |
| Transcription | *nusABG, rpoBCDH, rho, deaD, hepA* |
| Translation factors | *infB, efp, fusA, frr, tsf, tufA, tufB, prfC, rbfA* |
| Ribosomal proteins | 35 genes |
| tRNA synthetases | *alaS, argS, asnS, aspS, glnS, gltX, glyQ, glyS, ileS, leuS, lysS, metG, pheS, pheT, proS, serS, tyrS, valS, trpS* |
| Protein folding | *fkpA, mopA (groEL), dnaJ, dnaK, htpG, ppiB, slyD, tig, dsbA, grpE* |
| RNA degradation | *pnp, rne* |
| DNA repair | *gyrA, gyrB, ssb, recA* |
| Nucleotide biosynthesis | *adk, carB, guaAB, ndk, nrdA, nrdD, purABC, prsA, nrdB* |
| Nucleotide salvage | *deoBCD, upp, gpt* |
| Detoxification | *katG, sodA, sodB, tpx, ahpCF* |
| Glycolysis | *eno, fba(A), gpmA, gapA, pfkA, pgi, pgk, pykF, tpiA* |
| TCA cycle | *acnB, gltA, icdA, fumB, mdh, sdhA, sucABCD* |
| Respiratory chain | *atpADF, cydAB, cyoBC, fldA, frdABD, nuoCGLN, atpC* |
| Pentose-phosphate pathway | *talB, tktA, gnd* |
| Fatty acid biosynthesis | *accABC, acpD, fabABI,fabF* |
| | |
| ***Bacillus subtilis*** | |
| Transcription | *rpoA, rpoB, rpoC, rho* |
| Translation factors | *infC, efp, frr, fus, tsf, tufA,infB* |
| Ribosomal proteins | 31 genes |
| tRNA synthetases | *lysS, thrS, tyrS, alaS, argS, asnS, aspS, ileS, leuS, metS, valS, serS, gltX* |
| Protein folding | *dnaK, groES, groEL, ppiB, tig* |
| RNA degradation | *pnpA* |
| DNA repair | *recA, gyrA* |
| Nucleotide biosynthesis | *ctrA,guaAB, nrdE, purABQ* |
| Detoxification | *ahpCF, katA, sodA* |
| Glycolysis | *fbaA, pgk, pgm, pykA, tpi, eno, gap, pgi* |
| TCA cycle | *citBC, citH* |
| Respiratory chain | *atpAD, qoxBD, qoxA, atpF* |
| Fatty acid biosynthesis | *accB* |

| | |
|---|---|
| ***Haemophilus influenzae*** | |
| Transcription | *rpoB, rpoC, deaD* |
| Translation factors | *infB, infC, efp, fusA, rrf, tsf, tufB, tufA* |
| Ribosomal proteins | 35 genes |
| tRNA synthetases | *asnS, aspS, glyQ, glyS, lysU, thrS, valS, glnS, pheS, proS, serS, tyrS* |
| Protein folding | *dnaK, groEL, slyD, tig, prsA* |
| RNA degradation | *pnp* |
| DNA repair | *gyrA, ssb* |
| Nucleotide biosynthesis | *guaAB, nrdAB, purA, prsA, purM* |
| Nucleotide salvage | *deoD* |
| Detoxification | *sodA, hktE* |
| Glycolysis | *eno, fba, gpmA, pfkA, pgK, pykA, tpiA, gapdH* |
| TCA cycle | *mdh* |
| Respiratory chain | *atpAD, cydB, frdAB, atpF, cydA* |
| Pentose-phosphate pathway | *talB, tktA, zwf* |
| Fatty acid biosynthesis | *fabB, fabI, accBC* |

### 3.3.1 Probably highly expressed (PHE) genes in major metabolic pathways

Of the genes which encode the major glycolytic enzymes, nearly all are among the PHE classified enzymes in the analyzed organism (except phosphofructokinase Pfk from *B. subtilis* and glucose-6-phosphate isomerase Pgi from *H. influenzae*). In addition to some of these enzymes, isoenzymes with minor activity in glycolysis exist, which were not classified PHE. Examples of these from *E. coli* are PfkB (Table 3.4), FbaB, PgmI and PykA. PfkB encodes a phosphofructokinase with only a tenth of the activity of the glycolytic enzyme PfkA [96], FbaB has minor fructose-1,6-bisphosphate aldolase activity and the cofactor-independent phosphoglycerate mutase PgmI and pyruvate kinase PykA also exhibit only minor activity in glycolysis [97]. PykF and PykA are also sequence homologs [98], indicative of a paralogous relationship resulting from an act of gene duplication. PHE genes are also common in other major metabolic pathways such as the tricarboxylic acid cycle (TCA) and the respiratory chain, the central intermediate metabolism, biosynthetic pathways like the fatty acid biosynthesis and pentose-phosphate pathway, which provides reduction equivalents for biosynthetic purposes as well as the pentose-sugars needed for DNA and RNA biosynthesis. Also, genes from nucleotide biosynthesis and turnover are commonly classified PHE in the analyzed organisms.

### 3.3.2 PHE genes in protein biosynthesis

As the major participant in the transcription process, all subunits of the RNA polymerase core enzyme ($\alpha_2\beta\beta'$) are classified PHE for *B. subtilis*. The subunit $\alpha$ encoding *rpoA* gene from *E. coli* and *H. influenzae* lies slightly below the cut-off. Other PHE classified participants in DNA transcription are the transcription termination factor $\rho$ (*rho*) of *B. subtilis* and *E. coli* and two *E. coli* transcription initiation factors; the major $\sigma$ factor (RpoD), which participates in transcription initiation of nearly all genes expressed during normal growth conditions and heat-shock transcription factor $\sigma_{32}$ (RpoH), which controls the expression of heat-shock response genes. Of the genes involved in the translation process, nearly all genes $\geq$ 100 codons which encode ribosomal proteins or the major factors of translation initiation (*infB, infC*), elongation (*tufA, tufB, tsf, fusA*) and termination (*frr*; called *rrf* in *H. influenzae*) are classified PHE (Table 3.3). Proteins involved in the the translation process are among the most abundant ones and constitute a significant proportion of the proteins detected by 2-D PAGE analysis of experimentally growing *E. coli* (Figures 3.4,3.2, 3.3) and *B. subtilis* cells. Especially pronounced are expression level-dependent features in codon usage in ribosomal protein encoding genes, which for all three organisms are among the top-scoring genes (Table 3.2). For *B. subtilis*, this effect is strongest with seven of the ten top-scoring genes of Table 3.2 encoding ribosomal proteins and the information content of the corresponding scoring matrix (Table 3.1) being the highest of all listed matrices. The number of tRNA synthetase genes classified as PHE varies strongly from 19 in *E. coli* to only eight in *B. subtilis*.

### 3.3.3 PHE genes in stress response

A number of genes which encode proteins involved in the response to different kinds of stress conditions, such as sudden exposure to heat, cold or oxidative conditions are classified PHE. Among these are molecular chaperones, which play major roles in protein folding and turnover under both stress and non-stress conditions, such as $\sigma_{32}$-induced DnaK-DnaJ-GrpE chaperone complex, which blocks aggregation of newly synthesized and denatured protein and the heat-shock GroEL/GroES chaperonin involved in protein folding. Similarly, a number of proteins strongly induced during response to cold shock are among the PHE classified genes, such as CspA (length 70 amino acids) and cold-shock ribosome-binding factor A (RbfA). CspA is the major cold shock protein of *E. coli*, which accounts for more than 10% of protein biosynthesis during the acclimation phase and has been proposed to function as an RNA chaperone [99] and transcription antiterminator [100] at low temperatures. Another strongly induced cold-shock protein is Pnp of *E. coli*, which is a ribonuclease also involved in regulation of the cold-shock response [101]. The *pnp* gene is PHE for all analyzed organisms and among the top scoring PHE genes (Table 3.2). Also among the PHE classified genes are a number of proteins induced upon oxidative stress to protect the cell from oxidative damage. Examples of these are the *E. coli* isoenzymes of superoxide dismutase, SodA and SodB, which catalyze the decomposition of reactive superoxide radical anions, as well as hydroperoxidase I (KatG) and alkyl hydroperoxide reductase (AhpCF), needed for the detoxification of peroxides.

### 3.3.4 Comparison with the Codon Adaptation Index (CAI)

$S_{AV}(g)$ values were compared to CAI [40] values calculated relative to the same reference of highly expressed genes for the three genomes. Common variance (measured using the coefficient of determination $r^2$) of the genes relative to the model of expression level-dependent features and the set of highly expressed genes can be attributed to the presence of these features in the reference set of highly expressed genes. Figure 3.1 shows that these features are dominant in the codon usage of the highly expressed genes for all three genomes. For the genomes of *H. influenzae* and *B. subtilis*, the contribution to the codon usage of the reference set of highly expressed genes is less than for *E. coli* (95%), only 76% (81%) of the variance can be explained with them. The remaining 5 to 24% variance cannot be accounted for by expression level-dependent influences on codon usage and must be attributed to other influences which act on these genes.

A performance comparison of the model of expression level-dependent features and the Codon Adaptation Index in predicting protein abundance was undertaken by comparison with the protein abundance values of the *B. subtilis* dataset. As the common variance of CAI and $S_{AV}(g)$ values is only 81% and not as high as for *E. coli*, differences in performance should be observable. For the comparison protein abundance values were divided by the molecular weight of the corresponding proteins to obtain an estimate of the molar abundance of the corresponding proteins. The relation of the $S_{AV}(g)$ scores and CAI values with these estimates of molar abundance in both cases is weak, but the correlation was higher for $S_{AV}(g)$ ($r = 0.37$) than for CAI values ($r = 0.27$).

### 3.3.5 Comparison with *B. subtilis* proteome data

Büttner *et al.* recently gave a quantitative assessment of protein abundance for the 50 most abundant proteins in a *B. subtilis* cell extract of exponentially growing cells [91]. The majority of these proteins perform 'house-keeping' functions as components of translation apparatus (7), glycolysis (8), tricarboxylic acid cycle (3), amino acid metabolism (11) or protein folding (4). With the exception of amino acid metabolism, these categories correspond well with the functional categories listed in Table 3.3. 40 of the 50 corresponding genes of these proteins are classified PHE by codon usage evidence. Of the 10 that were not, 6 (*metC, serA, rocD, argF, leuB* and *ilvD*) are involved in amino acid metabolism.

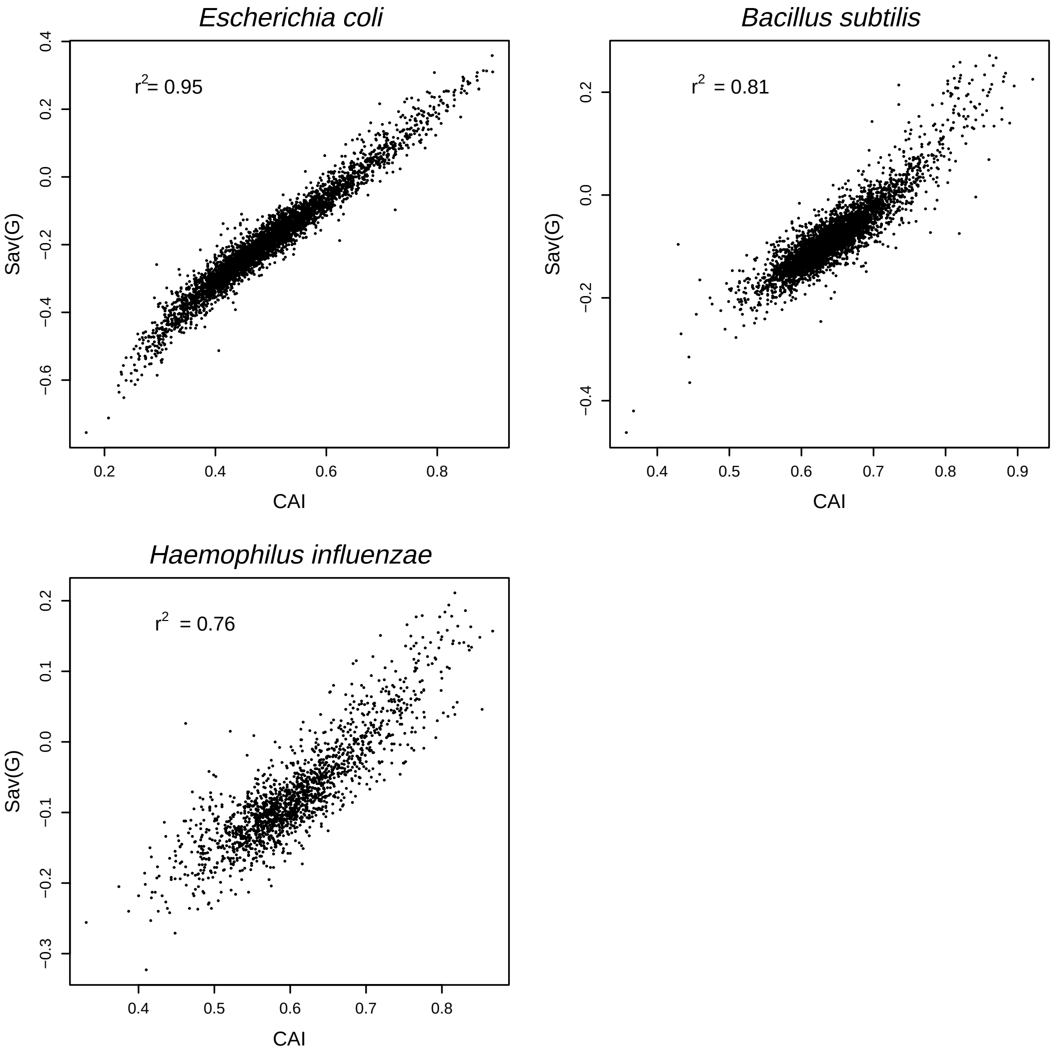


Figure 3.1: Comparison of $S_{AV}(g)$ with the Codon Adaptation Index (CAI) [40]. The coefficient of determination ($r^2$) measures the common variance of the two measures on the analyzed data set.

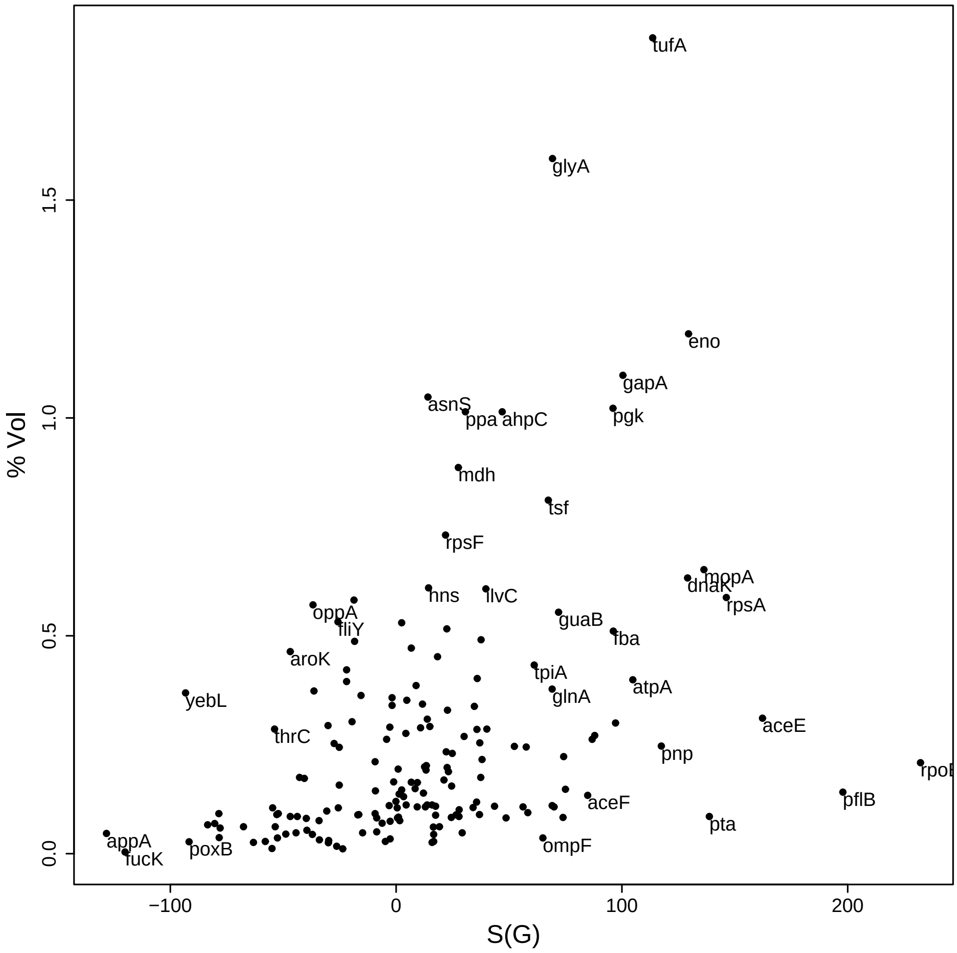


Figure 3.2: Comparison of relative protein abundance with $S(g)$ values of codon usage for 173 *E. coli* proteins and corresponding genes. %Vol is an estimate of protein abundance obtained by image analysis from the relative spot size in a silver-stained 2-D gel [92].

### 3.3.6 Comparison with *E. coli* proteome data

Expression level-dependent features in codon usage were compared with experimental data on the abundance of 173 proteins identified in the late exponential growth phase of *E. coli* cells. Of these 173 proteins, 94 are encoded by PHE classified genes (Table 3.4). A plot of relative protein abundance versus the $S(g)$ values of the corresponding genes is displayed in

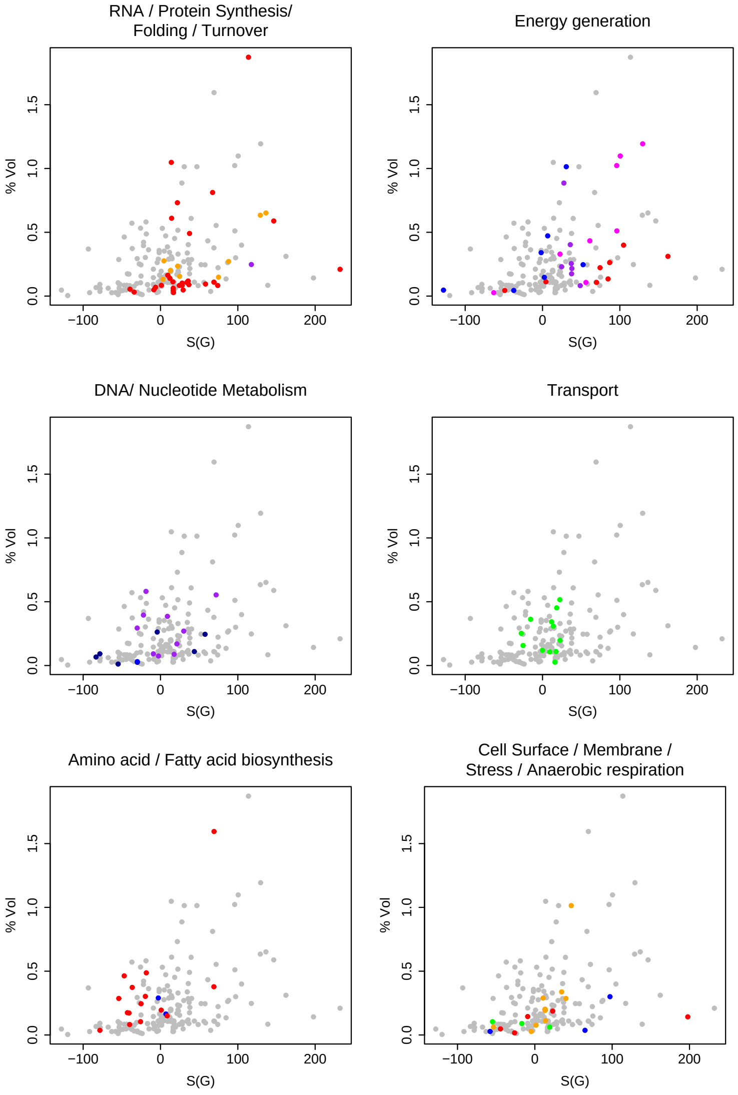
Figure 3.2. Points in this plot can be divided into three different categories: Points where protein abundance correlates with the strength of codon usage influences, which supports the analyzed hypothesis, points with low protein abundance but strong codon usage evidence, which might be genes stronger expressed under other than the measured experimental conditions and, thirdly, points corresponding to abundant proteins encoded by genes with weak evidence for expression level-dependent influences in their codon usage, which serve as counter-evidence for the hypothesis that abundant proteins are encoded by genes with strong evidence of expression level-dependent influences in their codon usage. As the analysis presented in Figure 3.2 shows, an overall relation of protein abundance and $S(g)$ scores is evident and the majority of the points belong to either the first or second category. Some genes, such as *aroK, thrC, oppA, fliY* and *yebL* in Figure 3.2 belong to the more abundant proteins but are not encoded by genes with expression level-dependent features evident in their codon usage.

Table 3.4: Functional classification of *E. coli* genes expressed at exponential growth.

| Functional category | PHE | not PHE |
|---|---|---|
| Amino acid biosynthesis | *dapD, glnA, glyA, ilvC* | *argI, aroG, aroK, aspC, cysK, dapA, gdhA, hisA, ilvE, leuC, lysA, metH, serC, thrC, trpA, trpD* |
| Anaerobic respiration | *pflB, yfiD* | *glpR, hypB, hypD* |
| Cell division | *ftsZ* | |
| Cell surface | *rfaD* | *kdsB, murE* |
| Central intermediary metabolism | *gltD, metK, ppa* | *appA, glpK, hdhA, sgaH* |
| Cofactor biosynthesis | | *folA, gor, ispA, trxA* |
| DNA degradation | | *xthA* |
| DNA replication / repair | *gyrA, recA* | *dksA, dnaB, dnaQ, mutS, polA* |
| Degradation | *pta* | *fucK, poxB* |
| Energy metabolism | *aceE, aceF, ackA, atpA, atpC, atpD, lpdA* | *glp*D |
| Fatty acid biosynthesis | *accB* | *fabD* |
| Global regulators | *sspA* | *fur* |
| Gluconeogenesis | | *fbp, ppsA* |
| Glycolysis | *eno, fba, gapA, gpmA, pfkA, pgk, tpiA* | *pfkB* |

| | | |
|---|---|---|
| Membrane | *ompA, ompF* | *phoE* |
| Nucleotide metabolism | *adk, guaB, ndk, udp, upp* | *carA, dut, pyrB, pyrI, trxB* |
| Pentose-phosphate pathway | *talB* | *rpiA, zwf* |
| Protein biosynthesis | *alaS, argS, asnS, aspS, frr, glnS, hns, leuS, lysS, metG, nusA, nusG, pheS, pheT, rbfA, rplI, rplU, rpoB, rpsA, rpsF, serS, tsf, tufA, tyrS* | *fmt, greA, hisS, map, rimL* |
| Protein folding | *dnaK, dsbA, fkpA, htpG, mopA, ppiB, tig* | *grpE, mopB* |
| RNA degradation | *pnp* | |
| Stress response | *ahpC, cspC, hslU, sodA, sodB, tpx, uspA* | *clpB, hslV, ibpA* |
| TCA cycle | *icdA, mdh, sdhA, sucA, sucB, sucC, sucD* | |
| Transport | *artI, crr, dppA, glnH, malE, mglB, ptsH, ptsI* | *hisJ, livJ,livK, oppA, potD* |
| Various | *tolB* | *cheZ* |
| not classified | *yjgF* | *bcp, fliY, hdeA, pfs, yadK, yceB, ydaA, yebL, yfiA, ygiN, yhhF, ylaD* |

Figure 3.3 shows the same data with the spots colored according to the functional category of the corresponding gene/protein, for 18 of the 23 functional categories given in Table 3.4. Functional categories with a common context such as energy generation for proteins participating glycolysis, TCA cycle, the respiratory chain or central intermediary metabolism were placed in one plot. Highly expressed genes during exponential growth include genes involved in RNA / protein biosynthesis, energy generation and DNA synthesis. For the genes of this functional context the relation of protein abundance with $S(g)$ values seems the most pronounced. Plot f in Figure 3.3 displays genes which encode proteins from other cellular compartments, such as the cell surface or outer membrane and proteins with stronger expression under other conditions than the one measured, such as exposure to stress or lack of oxygen. The proteins which belong to this category (with the exception of the highly expressed *ahpC* gene) display varying degrees of expression level-dependent features in their codon usage, but are not very abundant in the exponential growth phase. Plot e in Figure 3.3 displays genes involved in amino (red) or fatty acid (blue) biosynthesis, for which, similar to the genes involved in amino acid metabolism of the *B. subtilis* dataset, a relation of expression level-dependent features in codon usage and protein abundance is not apparent. A number of these moderately abundant proteins do not exhibit any significant evidence of expression level-dependent features in their codon usage (Table 3.4).



**A:** RNA / Protein synthesis / Folding / Turnover: Protein biosynthesis (red), protein folding (orange), RNA degradation (purple).

**B:** Energy generation: Central intermediary metabolism (blue), energy metabolism (red), glycolysis (magenta), TCA cycle (purple).

**C:** DNA / Nucleotide Metabolism: DNA degradation (blue), DNA replication/repair (darkblue), nucleotide metabolism (purple).

**D:** Transport (green).

**E:** Amino acid / Fatty acid biosynthesis: Amino acid biosynthesis (red), fatty acid biosynthesis (blue).

**F:** Cell surface / Membrane/Stress / Anaerobic respiration: Anaerobic respiration (red), cell surface (green), outer membrane (blue), stress (orange).

Figure 3.3: Comparison of relative protein abundance with $S(g)$ values of codon usage for 173 *E. coli* proteins and corresponding genes. Proteins are displayed colored according to functional categories. %Vol is an estimate of protein abundance obtained by image analysis from the relative spot size in a silver-stained 2-D gel [92].

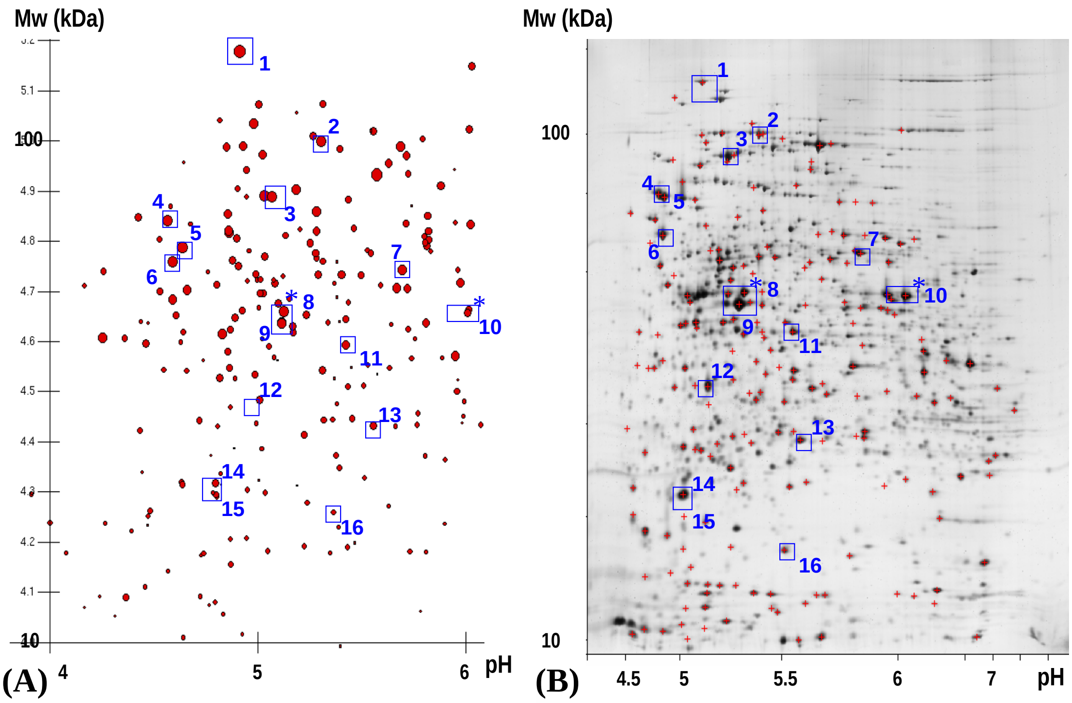


Figure 3.4: Comparison of a virtual 2-D gel with a 2-D gel electrophoresis experiment [92] for *E. coli*. Assuming a spherical spot shape, spot volume in the virtual gel (A) was set proportional to the $S(g)$ values of the corresponding genes. Abundant proteins determined in the real gel (B) and their counterparts in the simulation are denoted with boxes numbered from 1 to 16. Boxes marked with an asterisk surround two spots corresponding both to *GlyA* (Box 10) and *Eno* (Box 8) in (B). Proteins assigned to marked spots on the gels: 1 RpoB; 2 AceE; 3 Pta; 4 DnaK; 5 MopA; 6 RpsA; 7 AtpA; 8 Eno; 9 TufA; 10 GlyA; 11 Fba; 12 Tsf; 13 TpiA; 14 Ppa; 15 AhpC; 16 PpiB.

### 3.3.7 Simulation of a 'virtual' 2-D gel

Figure 3.4 displays a comparison of a ’virtual’ 2-D gel created from the theoretical molecular weight ($M_r$) and isoelectric point (p*I*) of the gene products described in the *E. coli* annotation with a silver-stained 2-D gel displaying a separation of cell extract from exponentially growing *E. coli* cells. Different protein abundance levels were simulated by setting the spot volume proportional to the corresponding $S(g)$ scores for every gene, which were taken as estimates of absolute protein abundance. Due to the non-linear pH gradient in the real gel, which is much steeper between pH 4.5-5.0 than between pH 5.0-6.0, the proteins spots in the virtual gel appear to be more spread out than those in the real gel. Spots in the real 2-D gel which are marked by a red cross have been linked to annotated *E. coli* genes by experimental evidence. To demonstrate the relation of codon usage properties and protein abundance some of the most abundant proteins in the real gel as well a s their counterparts of the simulation were surrounded with a blue box. Boxes marked with an asterisk contain proteins which probably due to post-translational modifications have been identified in two spots in the real gel as opposed to one spot in the virtual gel (serine hydroxymethyltransferase *GlyA* in the left box and enolase *Eno* in the upper part of the box on the right). Spots 14 (AhpC) and 15 (Ppa) in the virtual gel are both contained in the same spot in the real gel.

## 3.4 Discussion

### 3.4.1 Modeling expression level-dependent features

Codon usage is a commonly used means for classification of *E. coli* genes by expression levels and has recently also been applied to predicting highly expressed genes for diverse bacterial genomes [50]. In this work, the relation of expression level-dependent features in codon usage with gene expression rates was investigated by comparing the strength of expression level-dependent features in codon usage with protein abundance data from two proteome studies of *E. coli* [92] and *B. subtilis* [91]. To specifically determine the strength and significance of these features in the genes of three bacterial genomes, log-odds ratio scoring [53] was utilized to model expression level-dependent differences in codon usage.

To get an accurate estimate of expression level-dependent features in codon usage, the choice of the reference sets, especially the set of highly expressed genes, is critical. Because of this, different combinations of sets of genes were tried, which are similar in nature to the sets described in literature [50]. These contain genes involved in protein biosynthesis, which are known to be very highly expressed under exponential growth conditions. As these genes are highly expressed under a certain condition, there is the danger that this might lead to a misrepresentation of expression level-dependent features in codon usage under a different condition, if there is a condition-dependency of these influences. The wide range of functional categories represented by the PHE classified genes (Table 3.3) and the very high $S_{AV}(g)$ values attributed to some major genes of shock response to some extent confutes this possibility. Examples of such genes are the genes which encode the major *E. coli* cold shock protein CspA (Table 3.2), major chaperones and chaperonins of heat-shock response and proteins involved in oxidative stress response. Also, clustering of *E. coli* and *B. subtilis* genes by codon usage did not lead to such an observation [102, 103]. Thus, the differences in codon usage observed between the reference of highly expressed genes and genomic average for every genome are assumed to be condition-independent and representative for expression level-dependent features in codon usage of this genome.

### 3.4.2 Relation to skewedness of genomic GC-content

As can be inferred from the differing information content of the codon usage matrices (Table 3.1), the strength of expression level-dependent features in codon usage varies for the three genomes. The most pronounced are codon usage differences between highly expressed genes and genomic average for *E. coli* (0.13 nats of information per codon), they are weaker for *B. subtilis* (0.072 nats) and the least evident in the *H. influenzae* genome (0.063 nats). This order is inversely related to the the skewedness of the genomic GC-content for the analyzed genomes (*E. coli*: 51%, *B. subtilis*: 44%, *H. influenzae*: 38%). Overall GC-content of a genome has been found in a recent, large-scale study of 40 bacterial genomes to constitute the major influence on codon usage [44]. Possibly, expression-level dependent influences are only present if there is ’room’ in codon usage and their effect is not overridden by stronger forces such as maintaining the genomic GC-content in genomes with a biased base composition. In organisms where the information content of the codon usage matrix is weak, expression level-dependent features in codon usage must not only be present but even be stronger than in the genes of other organisms in order to be judged significant. Thus, in genomes with a very biased GC-composition prossibly only few, very highly expressed genes, such as ribosomal protein encoding genes, are subject to sufficient selectional pressure to bear significant evidence of these forces in their codon usage.

### 3.4.3 Relation to protein abundance at exponential growth

The existence of expression level-dependent features in codon usage has been explained with the presence of an external selective force, which optimizes a property of the translation process, such as ’translational efficiency’ or mRNA stability. These also are the basis for predicting highly expressed genes by codon usage properties [50], which implicitly assumes that these features can be taken as representative for gene expression rates. To evaluate the predictive value of this kind of information in codon usage with respect to gene expression rates, the strength of these features was compared with estimates of relative protein abundance for two sets of genes experimentally identified in two large-scale studies of the *E. coli* and the *B. subtilis* proteome under exponential growth conditions (Figures 3.4, 3.2). Ideally, such features should be compared with protein synthesis rates; here we assumed that in fast growing organisms such as *E. coli* and *B. subtilis* during exponential growth where a steady loss of protein due to cell division is to be compensated, abundant proteins also correspond to proteins synthesized at high rates. For both datasets, the majority of the proteins detected in the 2-D gel are also encoded by genes with significant evidence of expression level-dependent features in their codon usage. 94 of the 173 identified *E. coli* proteins, 40 of the 50 most abundant *B. subtilis* proteins in the 2-D gel are encoded by PHE classified genes. A plot displaying a comparison of protein abundance versus $S(g)$ values for the *E. coli* data set (Figures 3.2, 3.3) shows that most of the abundant proteins identified during exponential growth are also encoded by genes with expression level-dependent features evident in their codon usage. For some of the proteins determined to be present with medium abundance (up to 0.5 %vol) this relation does not seem to hold. Similar to the results found for the *B. subtilis* dataset, where 6 of the 10 not PHE classified genes are involved in amino acid metabolism, many of the corresponding *E. coli* genes are involved in amino acid biosynthesis (Figure 3.3, Table 3.4). Amino acid metabolism also is no frequent functional category found among the PHE classified genes in the three genomes (Table 3.3).

### 3.4.4 Favorable properties compared to the CAI

To determine how classification of genes by codon usage similarity relative to a reference set of highly expressed genes agrees with our method of estimating the strength of expression level-dependent features in codon usage, $S_{AV}(g)$ values were compared to the Codon Adaption Index (CAI) [40] (Figure 3.1). Classification by both procedures was found to agree very well for the genes of the *E. coli* genome, were 95% of the codon usage variance relative to the reference set of highly expressed genes can be explained with the presence of expression level-dependent features. For the genomes of *Bacillus subtilis* and *Haemophilus influenzae* genomes, these features are less evident in the codon usage of the highly expressed genes and only 81% (76%) of the variance can be explained with them. Thus, with increasing skewedness of the genomic GC-content, classification of genes by the strength of expression level-dependent features in codon usage differs increasingly from classification by codon usage similarity relative to one reference set of highly expressed genes. $S_{AV}(g)$ and CAI values were also compared with the estimates of molar protein abundance derived for the *B. subtilis* dataset, for which differences in CAI and $S_{AV}(g)$ values should be observable. Although the quantitative correlation of the $S_{AV}(g)$ scores and CAI values with these estimates of molar abundance in both cases were weak, correlation was higher for $S_{AV}(g)$ ($r = 0.37$) than for CAI values ($r = 0.27$).

### 3.4.5 *In silico* 2-D gel simulation

We used a simple heuristic to explore the use of expression level-dependent features in codon usage for the creation of an *in silico* 2-D gel simulation. Assuming a spherical spot shape, spot volume was set proportional to $S(g)$ values in a virtual 2-D gel calculated from the annotation data of the *E. coli* genome (Figure 3.4). Instead of a huge cloud of equally-sized spots the number of spots with spot volumes visible to the eye is greatly reduced, which confers a more realistic appearance to the simulation. In the comparison with a real 2-D gel experiment conducted with cell extract from exponentially growing *E. coli* cells, many abundant proteins identified in the real gel also correspond to abundant proteins in the simulation. The spots represented in the simulation of Figure 3.4 probably are a summation of abundant proteins expressed under different kinds of conditions. An additional refinement we plan to explore is to take into account the condition-dependency of gene expression, e.g. by displaying subsets of genes known to be expressed under a certain condition. But overall, codon usage features seem to be useful to confer a more realistic appearance to a virtual 2-D gel simulation. Using the implementation integrated into the GenDB system [17], experimental researchers who analyze the *C. glutamicum* proteome already use this kind of data to improve the 2-D gel simulation for their organism.

### 3.4.6 Summary

Using log-odds ratio scores to model codon usage differences between highly expressed genes and genomic average allows estimation of expression level-dependent features in codon usage. Forces which influence both the target and background reference used for creation of the codon usage models, such as forces related to generating an organisms GC-content, are excluded. The comparison of these features with experimental data on protein abundances has shown a relation between these to be present. This gives some justification to the approach of predicting highly expressed genes by their codon usage. By using such data for the creation of a more realistic virtual 2-D gel simulation, we propose an application which is closely related to experimental research.

CHAPTER 4

# Predicting the origin for horizontal gene transfer events by codon usage properties

Horizontal gene transfer (HGT) between the genomes of microbial species is thought to be an important force in shaping the gene-content of prokaryotic genomes. Here, differences in synonymous codon usage between microbial genomes are used for the detection and characterization of horizontal transfer events. As foreign genetic material is likely to be transferred in complete functional units, we searched for clusters of atypical genes (CAGs) in the genomic sequence data. An innovation compared to other sequence composition-based methods is the inference of a potential donor genome. On a data set of all currently available microbial genomes, the sensitivity of detecting the correct donor was found to be quite high for the *in silico* simulation of HGT events between contemporary microbial genomes. The prediction of CAGs and their putative donor genomes was evaluated for the genome of the bacterial hyperthermophile *Thermotoga maritima*. The result found hereby agree with previous phylogenetic, similarity-based and structural analyzes of the genome.

## 4.1 Introduction

Horizontal or lateral gene transfer (HGT) between different species is thought to play an important role in prokaryotic genome evolution. Generally, the prediction of genes which are foreign to a genome is based on the determination of ’atypical’ features, such as unusual sequence similarity or phylogenetic tree topology, distribution of homologs among species (phyletic pattern), or deviation from genome-specific sequence properties (see [104, 105, 9] for reviews on HGT). Examples for the latter are GC-content [48, 106], synonymous codon usage [51] or dinucleotide bias [107]. The sets of genes detected with these methods can be quite different [108, 109]. Sequence similarity-based methods allow the detection of HGT events which can be very ancient, but their application is limited to genes with sufficient known homologs present. Composition-based methods do not suffer from this restriction but are assumed to allow only the detection of more recent gene transfer events, due to the process of amelioration [48]. Among the genes with ’atypical’ sequence composition many highly mobile genetic elements, such as transposases or phage genes can be found.

The most significant contribution of horizontal gene transfer to the genome composition of modern organisms is thought to occur in situations which require adaptation by sequence evolution and acquisition, such as the colonization of a novel biological niche [104, 9]. Unless a selective advantage results through the contribution of a useful function to the cell, the duration of the newly acquired sequences is thought to be fleeting. Under this assumption, a successful transfer event requires the mobilization of all necessary genes in a single step. Selection will then be for the transfer of complete gene clusters or operons, which can be expressed in the recipient cell by a host promoter at the site of insertion [110].

For the genomes of different microbial species, intra-genomic variation in synonymous codon usage has been shown to be small compared to the inter-genomic differences [44]. In this study, the synonymous codon usage differences between microbial genomes are applied for the detection of genes with ’atypical’ codon usage in a genome. A novelty compared to methods which detect putative horizontally transferred genes based on atypical sequence composition [48, 49] is the inference of a potential donor genome. The sensitivity of donor genome detection is evaluated by the *in silico* simulation of HGT events for all currently available genomes of microbial species (state of January 2003). For the detection of ’alien’ genes in the real genomic data, the search is extended to clusters of genes with atypical codon usage, as foreign genetic material is likely to be transferred in complete functional units. For this, the concept of a cluster of atypical genes (CAG) is introduced. A CAG consists of neighboring atypical genes in the genomic sequence, which all show a higher similarity in codon usage to the same potential donor genome than to the acceptor genome. The score for this donor must lie above a previously defined significance threshold and must be the highest of all potential donors for which evidence exists in the cluster.

A detailed analysis of the predicted CAGs is performed for the bacterial hyperthermophile *Thermotoga maritima*, for which similarity-based, phylogenetic and structural evidence of gene transfer events from archaeal species has been reported [66, 67, 68].

## 4.2 Materials and Methods

### 4.2.1 Sequence data

The data set consists of all EMBL annotations available for complete microbial genomes (state of January 2003), totaling in 106 annotations with 90 eubacterial and 16 archaebacterial organisms represented. This corresponds to 88 different species. For 10 species genomes from multiple strains were available (Table 4.1). Using modules of the BioPerl library [111], sets of CDSs were extracted from the DNA sequence according to the annotation information. To exclude annotation ambiguities, coding sequences described with 'fuzzy' start or stop positions, containing a non-integer number of codons or internal stop codons (annotation errors or pseudogenes) were discarded.

Table 4.1: List of species with completely sequenced genomes in the data set. The number of strains is given in parentheses.

| Species (No. of Strains) | |
|---|---|
| *Aeropyrum pernix* (1) | *Agrobacterium tumefaciens* (1) |
| *Aquifex aeolicus* (1) | *Archaeoglobus fulgidus* (1) |
| *Bacillus halodurans* (1) | *Bacillus subtilis* (1) |
| *Bifidobacterium longum* (1) | *Borrelia burgdorferi* (1) |
| *Bradyrhizobium japonicum* (1) | *Brucella melitensis* (1) |
| *Brucella suis* (1) | *Buchnera aphidicola* (1) |
| *Buchnera* sp. (1) | *Campylobacter jejuni* (1) |
| *Caulobacter crescentus* (1) | *Chlamydia muridarum* (1) |
| *Chlamydia trachomatis* (1) | *Chlamydophila pneumoniae* (3) |
| *Chlorobium tepidum* (1) | *Clostridium tetani* (1) |
| *Clostridium acetobutylicum* (1) | *Clostridium perfringens* (1) |
| *Corynebacterium efficiens* (1) | *Corynebacterium glutamicum* (1) |
| *Deinococcus radiodurans* (1) | *Escherichia coli* (4) |
| *Fusobacterium nucleatum* (1) | *Haemophilus influenzae* (1) |
| *Halobacterium* sp. (1) | *Helicobacter pylori* (2) |
| *Lactococcus lactis* (1) | *Leptospira interrogans* (1) |
| *Listeria innocua* (1) | *Listeria monocytogenes* (1) |
| *Lactobacillus planetarum* (1) | *Mesorhizobium loti* (1) |
| *Methanobacterium thermoautotrophicum* (1) | *Methanococcus jannaschii* (1) |
| *Methanopyrus kandleri* (1) | *Methanosarcina acetivorans* (1) |
| *Methanosarcina mazei* (1) | *Mycobacterium leprae* (1) |

| | |
|---|---|
| *Mycobacterium tuberculosis* (2) | *Mycoplasma genitalium* (1) |
| *Mycoplasma penetrans* (1) | *Mycoplasma pneumoniae* (1) |
| *Mycoplasma pulmonis* (1) | *Neisseria meningitidis* (2) |
| *Nostoc* sp. (1) | *Oceanobacillus iheyensis* (1) |
| *Pasteurella multocida* (1) | *Pseudomonas aeruginosa* (1) |
| *Pseudomonas putida* (1) | *Pyrobaculum aerophilum* (1) |
| *Pyrococcus abyssi* (1) | *Pyrococcus furiosus* (1) |
| *Pyrococcus horikoshii* (1) | *Ralstonia solanacearum* (1) |
| *Rickettsia conorii* (1) | *Rickettsia prowazekii* (1) |
| *Salmonella typhimurium* (2) | *Shewanella oneidensis* (1) |
| *Shigella flexneri* (1) | *Sinorhizobium meliloti* (1) |
| *Staphylococcus aureus* (3) | *Staphylococcus epidermidis* (1) |
| *Streptococcus agalactiae* (1) | *Streptococcus mutans* (1) |
| *Streptococcus pneumoniae* (2) | *Streptococcus pyogenes* (1) |
| *Streptomyces coelicolor* (1) | *Sulfolobus solfataricus* (1) |
| *Sulfolobus tokodaii* (1) | *Synechocystis* (1) |
| *Thermoanaerobacter tengcongensis* (1) | *Thermoplasma acidophilum* (1) |
| *Thermoplasma volcanium* (1) | *Thermosynechococcus elongatus* (1) |
| *Thermotoga maritima* (1) | *Treponema pallidum* (1) |
| *Ureaplasma urealyticum* (1) | *Vibrio cholerae* (1) |
| *Vibrio vulnificus* (1) | *Wigglesworthia brevipalpis* (1) |
| *Xanthomonas campestris* (1) | *Xanthomonas citri* (1) |
| *Xylella fastidiosa* (2) | *Yersinia pestis* (2) |

### 4.2.2 Detecting atypical genes and a potential donor genome

Synonymous codon usage is assumed to be a genome-specific characteristic for the genomes of different species, with intraspecies differences being small compared to those between different species [44]. Thus, a possible explanation for the existence of genes $g$ in genome $x_a$ with a codon usage more similar to another genome $x_b$, is that they were transferred from $x_b$ or a closely related species to $x_a$. To detect atypical genes and a potential donor genome, the following procedure is applied:

Let $x_a$ be the 'acceptor' and $x_b$ a possible 'donor' genome for a horizontal gene transfer event, with $x_a, x_b \in X$, where $X$ is a set of genomes of different organisms. For a transfer event from another organism's genome to genome $x_a$, let $X_{\hat{a}} = X \setminus \{x_a\}, a \neq b$ be the set of possible donor genomes. As described in chapter 2, log-odds ratio scoring of synonymous codon usage can be used for detecting genes in $x_a$ with a higher similarity in codon usage to another genome $x_b$. For this, codon usage differences between the genomes $x_a \in X$ and $x_b \in X_{\hat{a}}$ were modeled using log-odds ratio scoring matrices $M_{b,a}$ of synonymous codon usage. For every genome, the annotated set of CDSs was utilized as the reference set in matrix creation. For every gene $g$ of $x_a$, log-likelihood ratio scores of synonymous codon usage $S(g)$ were calculated using the $M_{b,a}$ matrices of the potential donor genomes $x_b \in X_{\hat{a}}$. Calculations were carried out using the CoBias program [42].

### 4.2.3 Detection of clusters of atypical genes (CAGs)

As foreign genetic material is likely to be transferred in complete functional units, instead of searching for single genes, we searched for clusters of genes with atypical codon usage properties. To detect such clusters, genes of $x_a$ satisfying $S(g) > 0$ in comparison against one or more other genomes $x_b \in X_{\hat{a}}$ were taken as initial candidates. Then, the following procedure was applied: Let $g \in G_{u_j}$ be the set of genes of a cluster $u_j$ in genome $x_a$. Based on the absolute start position in the genome, every gene of the initial set is assigned to a cluster $u_j$, with the start of neighboring genes in a cluster not being allowed to be more than $k$ bp apart. To propose a donor genome for a cluster $u_j$, we proceed as follows: For every genome $x_b \in X_{\hat{a}}$ with a score $S(g) > 0$ for one or more genes $g \in G_{u_j}$, a cluster score $T_{b,j}$ is calculated:

$$T_{b,j} = \sum_{g \in G_{u_j}} S(g) \tag{4.1}$$

Of the obtained cluster scores, let $T_{b,j}^{max}$ be the maximum cluster score for cluster $u_j$ and $M_{b,a}^{max}$ the corresponding applied matrix. If using $M_{b,a}^{max}$ resulted in a score $S(g) > 0$ for all $g \in G_{u_j}$, the corresponding genome $x_b$ is proposed as donor for $u_j$. Otherwise, the cluster is split into two or more clusters $u_{\tilde{k}}$ of neighboring genes in the sequence which either satisfy

$$S(g) > 0, \quad \text{for all } g \in G_{u_{\tilde{k}}} \tag{4.2}$$

or

$$S(g) \leq 0, \quad \text{for all } g \in G_{u_{\tilde{k}}}. \tag{4.3}$$

using $M_{b,a}^{max}$. For these new clusters, the procedure is repeated recursively until all have been assigned a donor. Neighboring clusters with the same proposed donor are subsequently merged. The clusters with values of $T_{b,j}^{max} \geq c$ are predicted as CAGs. The applied parameter settings were $k = 5000$ and $c = 10$. Using such a large value of $k$ allows the detection of cluster containing individual not atypical genes between the atypical ones.

### 4.2.4 Phylogenetic methods

For comparison with the alien gene predictions by codon usage properties, phylogenetic analyzes were performed for trees automatically generated for the genes of the analyzed genomes with the Pyphy program [112]. The Pyphy system allows automatic retrieval of homologs from a non-redundant database consisting of SWISS-PROT, TrEMBL and TrEMBLnew entries [113], multiple alignment creation with CLUSTALW [114] and phylogenetic tree construction with PAUP [115] using the neighbor-joining method [116] with 100 bootstrap steps.

### 4.2.5 Performance evaluation

To measure discriminatory power of codon usage differences between genomes, the relative (or receiver) operating characteristic (ROC) [32] was used. For a dataset consisting of known events of two types, the ROC is evaluated in a graph of the true positive proportion of positive items (sensitivity) versus the false positive proportion of negative items, for various settings of the decision threshold. The ROC corresponds to the area under the curve and measures the probability of correct classification. It can be used as a single-valued, general measure of classification accuracy. The ROC is seen to vary between 0.5 and 1.0, with a value of 0.5 meaning no discrimination exists. A ROC of 1.0 means perfect discrimination, the true positive proportion is one for all values of the false positive proportion. For analyzes the Bioconductor ROC library was used with the R statistical package [117].

The significance of differences in the donor genome detection sensitivity for different subsets of the possible donor-acceptor combinations of genomes in the data set was evaluated by using two-sample t-tests with pooled variance for similar variance samples and the Welch approximation to the degrees of freedom otherwise.

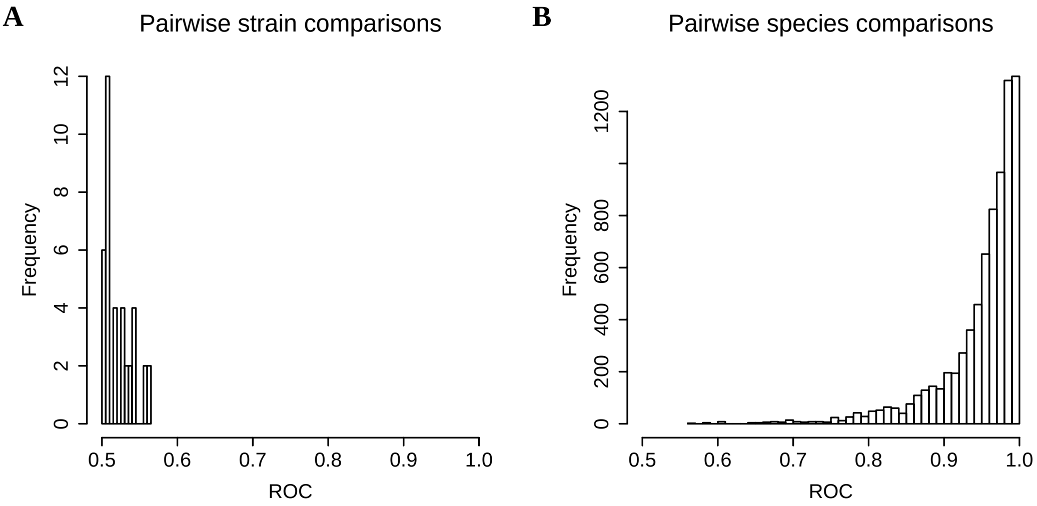


Figure 4.1: The discriminatory power of codon usage differences between microbial genomes. Displayed is the frequency of the obtained ROC scores obtained, which measure the discriminatory power of log-odds ratio scores of codon usage for discriminating between the genes of the different genomes. **A:** All combinations of strains of the same species. **B:** All combinations of different microbial species.

## 4.3 Results

### 4.3.1 The discriminatory power of codon usage differences between contemporary genomes

Prerequisite for the identification of horizontally transferred genes by codon usage properties is a sufficient degree of dissimilarity in codon usage between a donor $x_b$ and acceptor genome $x_a$ to allow discrimination of newly acquired genes from the other genes present in $x_a$. The receiver operating characteristic (ROC) [32] was used to determine the discriminatory power of codon usage differences for the genes of the complete set of microbial genomes used in this study. For the genomes of different microbial species, differences in codon usage allow a very accurate classification of genes (Figure 4.1). A mean $\overline{\text{ROC}}$ for interspecies comparisons of $0.95 \pm 0.06$ indicates that on average a gene belonging to genome $x_b$ will achieve a higher score than a gene of genome $x_a$ in 95 out of 100 cases. Contrasting this, differences in codon usage for genomes belonging to strains of the same species are nearly not detectable. For discrimination between the genes from two strains of the same species, the $\overline{\text{ROC}}$is $0.52 \pm 0.02$, which is very close to the performance of a system deciding on chance alone. To illustrate the distribution of the $S(g)$ values, some examples for pairs of genomes with different degrees of evolutionary relatedness are given in Figure 4.2.

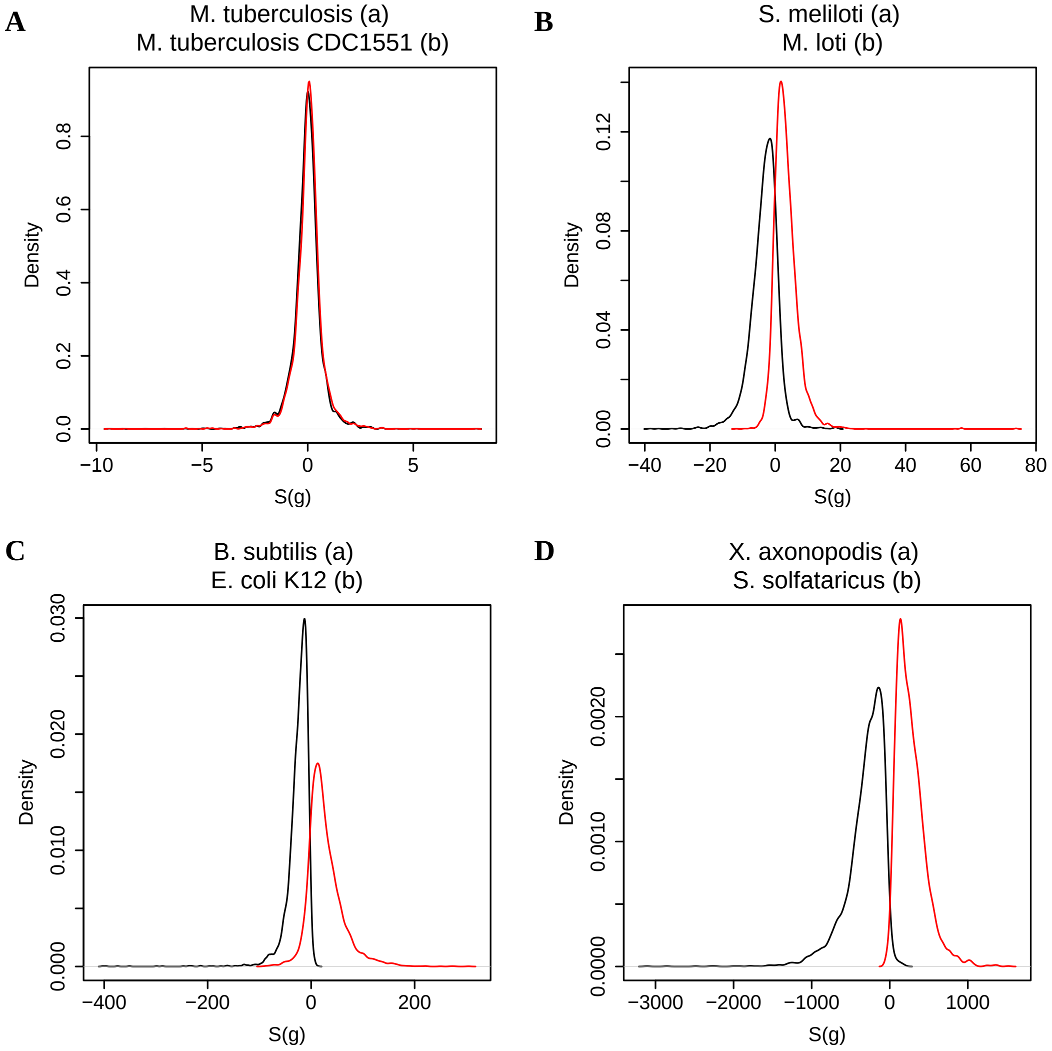


Figure 4.2: Distribution of scores for pairs of genomes with different degrees of evolutionary relatedness. The distribution of $S(g)$ values for the genes of $x_a$ (black) versus the distribution for the genes of $x_b$ (red) using the $M_{b,a}$ matrix of synonymous codon usage differences between $x_a$ and $x_b$. **A:** Two strains of *Mycobacterium*. **B:** Two α-proteobacteria. **C:** A member of the Bacillus / Clostridium group and a proteobacterium. **D:** An eu- and an archaebacterium.

### 4.3.2 Predicting the origin of artificially transferred genes

Using synonymous codon usage differences between microbial genomes for the detection of the donor of horizontally transferred genes was evaluated in a simulation experiment. For every genome of the 88 analyzed species, artificial gene transfer was performed by randomly sampling without replacement 100 genes from the 87 genomes of other species.

For every gene transferred to genome $x_a$, the $M_{b,a}$ matrices were used to calculate $S(g)$ values, with $x_b \in X_{\hat{a}}$. The $S(g)$ scores obtained for a gene with the different matrices were sorted in decreasing order. Based on this, the sensitivity in donor genome detection $z_r$ was determined. $z_r$ is the proportion of genes for which the score $S(g)$ obtained with the $M_{b,a}$ matrix of the correct donor genome $x_b$ has a rank $\leq r$ and a value $> 0$. The mean sensitivity in donor genome detection $z_1$ for inter-species transfers is 62%. It increases to 75% of correctly detected donors for $r = 2$ and 81% for $r = 3$. This is far from the expected sensitivity when randomly sampling (without replacement) a donor from the set of $k$ genomes. For $k = 88$, this is 1% ($r = 1$), 2% ($r = 2$) and 3% ($r = 3$) for different values of $r$.

The relation of the sensitivity in donor genome prediction to the degree of evolutionary relatedness of the donor and acceptor genomes as well as similarity in overall genomic GC-content was investigated (Table 4.2). For all evaluated phylogenetic categories, the sensitivity of donor genome detection was higher for gene transfers between species of different categories than for those belonging to the same one. The same effect can be seen if moving downwards from phylogenetic categories which contain a large number of distantly related species, such as the domain, to categories which contain smaller numbers of more closely related species, such as the group. As GC-content is known to have a major influence on the synonymous codon usage of microbial organisms, it was also investigated whether sensitivity is lower for transfer events between genomes with similar base composition or of high GC-content. For this, the set of genomes was divided into three nearly equal sized groups of low ($<$ 39%), middle and high ($\geq$ 51%) GC-content. The sensitivity of donor genome detection is slightly higher for transfer events in which either one or both of the involved genomes have a lower GC-content. A significant difference in donor detection for genomes with either similar or different GC-content was not observed (Table 4.2).

Table 4.2: Mean sensitivity ($z_1$) of detecting the donor genome for artificial HGT events for different phylogenetic relationships and GC-content of donor and acceptor genomes.

| Phylogenetic category | Identical | Different | P-value[a] |
|---|---|---|---|
| Domain | $0.60 \pm 0.17$ (5496[b]) | $0.65 \pm 0.17$ (2336) | $8.96 \cdot 10^{-23}$ |
| Class | $0.59 \pm 0.18$ (1604) | $0.63 \pm 0.17$ (6228) | $4.19 \cdot 10^{-20}$ |
| Subclass | $0.55 \pm 0.17$ (634) | $0.63 \pm 0.17$ (7198) | $3.00 \cdot 10^{-23}$ |
| Group | $0.57 \pm 0.18$ (232) | $0.62 \pm 0.17$ (7600) | $1.39 \cdot 10^{-05}$ |
| Low Donor GC | $0.64 \pm 0.16$ (930) | $0.64 \pm 0.16$ (1798) | 0.85 |
| Middle Donor GC | $0.62 \pm 0.18$ (812) | $0.62 \pm 0.18$ (1740) | 0.99 |
| High Donor GC | $0.60 \pm 0.17$ (812) | $0.60 \pm 0.17$ (1740) | 0.87 |

[a] The P-value estimates the significance of the observed sensitivity values for inter- versus intragroup comparisons of the different categories. The lower the P-value, the more significant is the observed difference in sensitivity.

[b] Number of evaluated donor and acceptor combinations.

### 4.3.3 Taking a closer look: CAGs in the *Thermotoga maritima* genome

As a suitable genome for a more detailed evaluation on real genomic data, the *T. maritima* genome was chosen, because ample evidence for the existence of archaeal genes in the organism has been reported [66, 67, 68]. Application of the described procedure resulted in the detection of 37 clusters of atypical genes consisting a total of 80 genes (Figure 4.3). This corresponds to 4.4% of the 1.9 Mb genomic sequence, a slightly lower estimate than the 6.4% detected by Ochman *et al.* based on atypical sequence properties [104]. The gene product descriptions of the functionally characterized atypical genes are mostly related to small molecule uptake and transport, sugar and cell wall biosynthesis and regulators of transcription. For the majority of these genes (51), an archaeal species is predicted as donor. In total, there are nine different species of the archaeal domain predicted as donor organisms (Table 4.3).

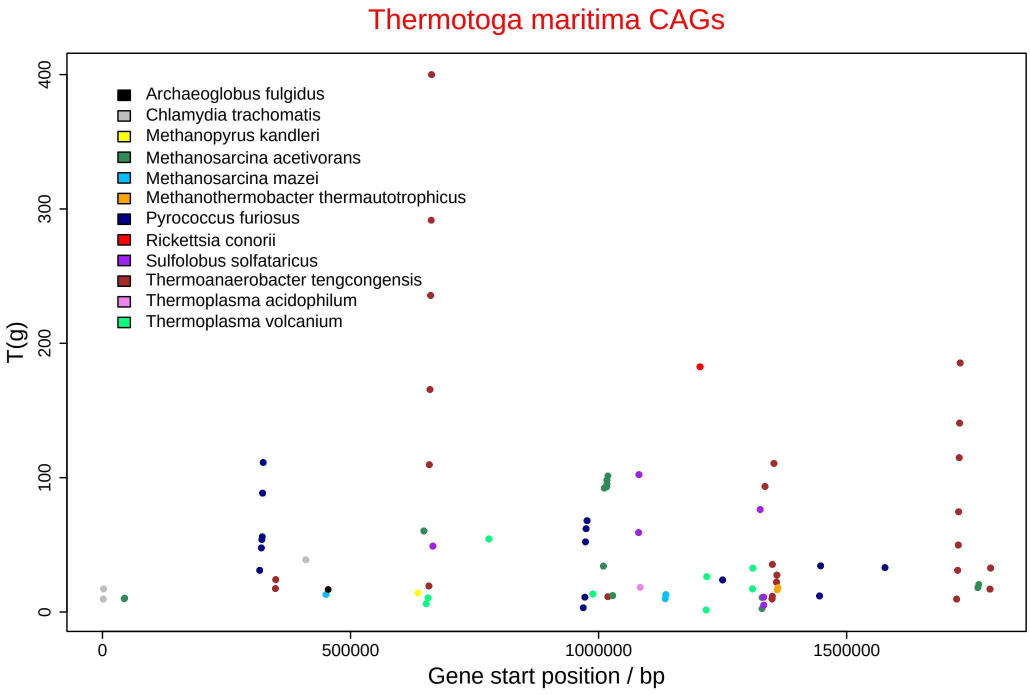


Figure 4.3: Clusters of atypical genes in the *Thermotoga maritima* genome. Moving from start to the end of the genomic sequence, for every gene of a CAG the cumulative score $T(g)$ against the predicted donor genome is displayed at the start position in the genome sequence.

Table 4.3: Habitat or location of isolation of the predicted donor species for *T. maritima* CAGs. Unless specified otherwise, the data is from [14].

| Species | Domain | Habitat or location of isolation |
|---|---|---|
| *Archaeoglobus fulgidus* | Archaea | Marine hydrothermal systems of Italy |
| *Chlamydia trachomatis* | Bacteria | Obligate intracellular parasite |
| *Methanopyrus kandleri* | Archaea | Isolated from the sea floor at the base of a 2,000-m-deep black smoker chimney in the Gulf of California [118] |
| *Methanosarcina acetivorans* | Archaea | Marine Sediment [119] |
| *Methanosarcina mazei* | Archaea | Sewage[a] and isolated from an aquaculture fishpond [120] |
| *Methanothermobacter thermautotrophicus* | Archaea | Sewage sludge and isolated from Cuyahoga River, Cleveland, Ohio[a] |
| *Pyrococcus furiosus* | Archaea | Volcanic marine sediment, Italy[a] |
| *Rickettsia conorii* | Bacteria | Strict intracellular pathogenic bacterium |
| *Sulfolobus solfataricus* | Archaea | Volcanic area near Naples |
| *Thermoanaerobacter tengcongensis* | Bacteria | Isolated from a hot spring in China [121] |
| *Thermoplasma acidophilum* | Archaea | Its usual habitat is in heat-generating coal slag-heaps, but it can also can be found in vulcanic hot springs |
| *Thermoplasma volcanium* | Archaea | Solfatoras around the world |

[a] Information from `http://www.atcc.org/SearchCatalogs`

For a direct gene transfer event between different bacterial species, the donor and acceptor species need to inhabit, at least transiently, a common habitat. With the exception of two species of pathogenic intracellular bacteria, the habitats of the donor species all share common properties with that of the thermophilic *T. maritima*, which was isolated from a geothermal heated marine sediment at Vulcano, Italy (Table 4.3). To the pathogenic bacteria *Chlamydia trachomatis* and *Rickettsia conorii* a total of four genes was assigned, which have neither homologs nor any evident functional context in the genome and thus possibly are erroneously annotated. All other species originate from either thermophilic or aquatic environments. The species to which most of the atypical genes (25) have been assigned as potential donor is *Thermoanaerobacter tengcongensis*, a thermophilic bacterium which has been isolated from a hot spring in China. Among the CAGs assigned to this donor is a cluster of seven putative lipopolysaccharide biosynthetic genes and a cluster of seven genes containing an ABC transporter operon.

The most frequently predicted archaeal donor is *Pyrococcus furiosus*, to which 5 CAGs containing 15 genes have been assigned. These correspond to complete or parts of operons of oligopeptide ABC transporter genes (Figure 4.5), ribose ABC transporter genes, a glycerol uptake operon and a clostripain-related protein. Although *P. furiosus* was not included in that particular analysis, a relation of the 'Archaea-like sequences' [66] in the *T. maritima* genome to another *Pyroccocus* species has already been reported, based on the periodicity of structural DNA sequence properties [68]. Figure 4.5 shows a family of related ABC transporter operons, which has previously been proposed to originate from horizontal gene transfer, based on atypical phylogenetic connections to different *Pyrococcus* species [112]. The phylogenetic analyzes performed on the complete operons I - IV in this study provided evidence for a complex evolutionary scenario, involving both gene duplication and horizontal gene transfer events (Figure 4.4). Based on synonymous codon usage properties, the following was inferred. Three of the depicted six operons or operon remnants are predicted as CAGs. A *Pyrococcus* species (*P. furiosus*) is also proposed as donor. The evidence also supports a complex evolutionary scenario with more than one transfer event. In case a single operon would have been transferred and subsequently undergone gene duplications, under the assumption of similar amelioration rates, the operons would be expected to exhibit a similar degree of atypicality in codon usage properties, reflective of the residence time in the genome of the organism. As this is not the case, a more plausible scenario includes at least two transfer events of the operons I - IV, with the transfer of operon IV being the most recent.

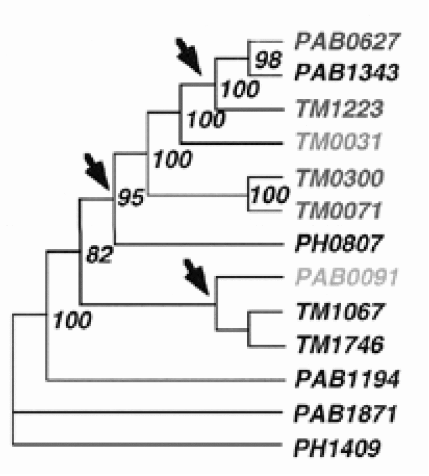


Figure 4.4: Phylogenetic analysis of the combined translated sequences of oligopeptide ABC transporter genes in *T. maritima* (TM), *Pyrococcus abyssi* (PAB) and *Pyrococcus horikoshi* (PH) performed in [112]. Values at the nodes indicate the percentage of 1000 neighbour joining bootstraps. Arrows indicate sites of putative horizontal gene transfer events.

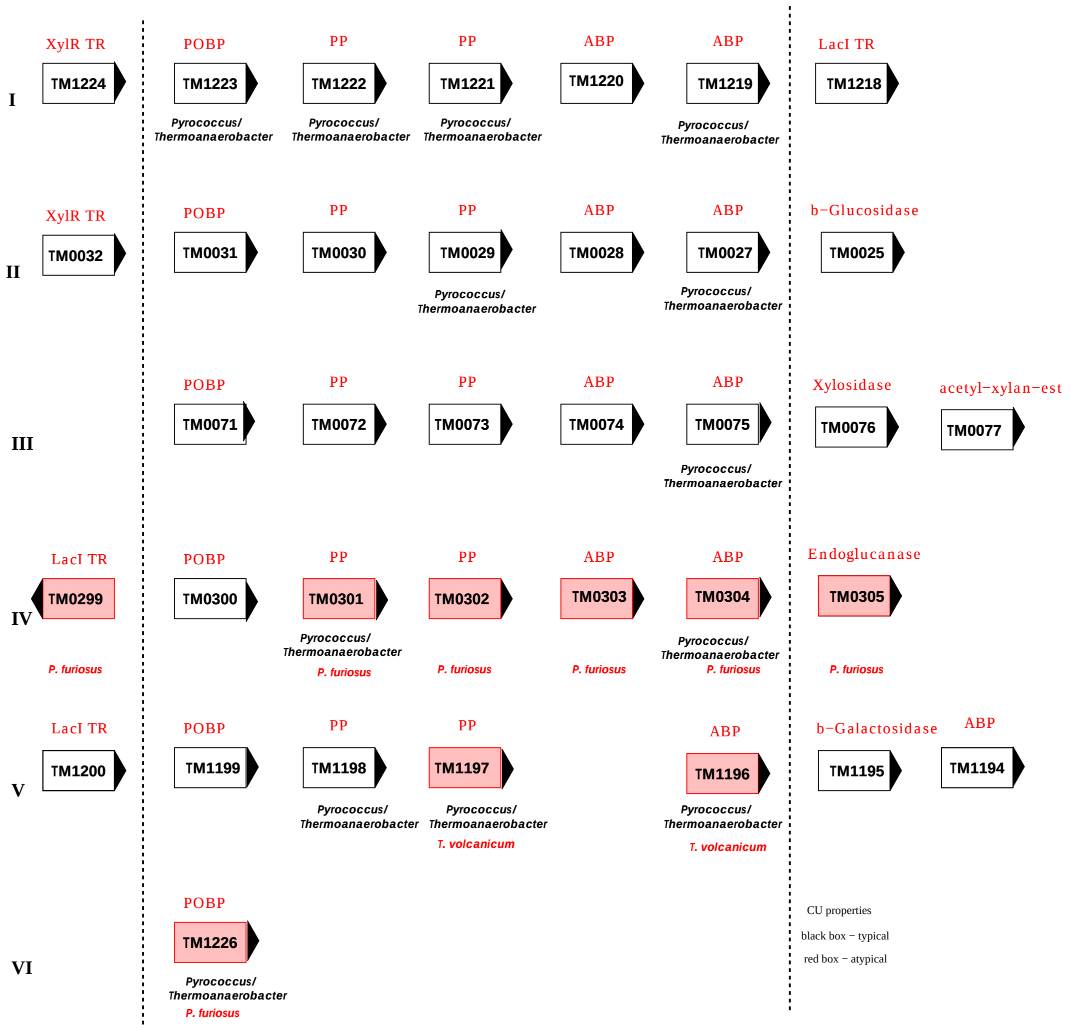


Figure 4.5: Clusters of oligopeptide ABC transporter genes with atypical phylogenetic connections [112] in the *Thermotoga maritima* genome[a]. The schematic representation shows the co-location, genomic context, phylogenetic and synonymous codon usage relations of the gene family. Genes colored red are members of CAGs with the deduced donor organism given in red below. For many of the genes, besides the homologous *T. maritima* genes, only genes of *Pyrococcus* species and *Thermoanaerobacter* were found as nearest neighbors connected by the same parental node in the phylogenetic trees (specified in black below). Abbreviations for the gene products: *POBP - periplasmatic oligopeptide binding protein, PP - permease protein, ABP - ATP binding protein, TR - transcriptional regulator*.

[a] V and VI were not included in the phylogenetic analysis of [112].

## 4.4 Discussion

The prediction of putative alien genes based on a higher similarity in codon usage to another genome is an innovation which allows the proposal of a potential donor genome based on codon usage properties for horizontally transferred genes. As foreign genetic material is likely to be transferred in complete functional units, instead of searching for single genes, we searched for cluster of atypical genes (CAGs), consisting of neighboring atypical genes in the genomic sequence. The method requires a training set of organism-specific genes of sufficient size, thus the donor genome or that of a closely related species must be available for the analysis. With the rapidly increasing number of complete genome sequences which are currently being published, this will likely be the case for many horizontally transferred genes in the near future.

To ascertain the suitability of codon usage differences between microbial genomes to place a gene with its 'host' genome, preliminary studies on the discriminatory power of these differences and a simulation experiment with artificially transferred contemporary genes were conducted. Here, it was found in agreement with a previous study on the genes of 40 microbial genomes [44], for the 88 different microbial species analyzed in this study, codon usage differences between genomes are much larger than intra-genomic variation. The discriminatory power of codon usage differences is very high. We found that it does not only suffice for accurate discrimination between the genes of two genomes, but even allows to reach a high sensitivity in predicting the correct out of 88 possible donors for artificial gene transfer events.

For the detection of horizontally transferred genes in real genomic data, the situation is complicated by the process of amelioration [48], the adaptation of a gene's sequence properties to the host genome. As no functional constraints related to the encoded gene product are associated with the use of the different synonymous codons, this occurs much faster than evolution on amino acid level. Thus, only more recent transfer events are detectable by synonymous codon usage properties compared to phylogenetic methods. To increase the accuracy in horizontal gene transfer detection, we searched for clusters of atypical genes with a higher similarity to the same donor genome in synonymous codon usage. This is sensible from the biological perspective and – as the scores obtained for the individual genes against a common donor may be seen as independent and thus additive evidence – enables a more accurate discrimination between horizontally transferred regions and individual genes which display atypical properties due to random fluctuations or other sources of intragenomic variation.

For evaluation of the methodology, the genome of *Thermotoga maritima* was chosen, as ample evidence for the transfer of archaeal genes into the genome has been found [66, 67, 68]. Of the predicted CAGs for the *T. maritima* genome, most have been assigned to archaeal donors. This finding is given confirmation by previous reports based on phylogenetic, similarity-based and structural analyzes of the genome. Interestingly, the habitat conditions of *T. maritima* and the nine different deduced archaeal donor species are similar, with all species inhabiting high-temperature or aquatic environments. A recent study discovered common features in the synonymous codon usage of thermophiles, which were assigned to a selective force related to the thermophilic habitat conditions [44]. As common features found in both references are discarded with the applied methodology, these do not contribute to the analysis. Rather, the predictions reflect the higher similarity in codon usage to the inferred donor genome as opposed to the current residence genome, and are based on the differences in synonymous codon usage between the two genomes. Even more convincingly, the *T. maritima* CAGs corresponding to a family of homologous oligopeptide ABC-transporters, have previously been identified as likely gene transfer candidates based on their atypical phylogenetic connections [112]. In agreement with the phylogenetic analysis performed in this study, the evidence also supports a gene transfer scenario involving more than one transfer event and a *Pyrococcus* species is proposed as donor. Thus, based on different sources of information – phylogenetic relationships deduced from the succession of amino acid letters in the sequence and features determined from sequence composition in terms of synonymous codon usage – the same result is obtained.

Sequence similarity and sequence composition-based approaches to HGT characterization have been noted to detect different sets of genes [108, 109]. Compared to methods which use sequence similarity on amino acid level, sequence composition-based approaches only allow the detection of more recent transfer events. As they are using global, genome-specific properties, sequence composition-based approaches are not limited to the detection of atypical genes with known homologs from other species. As we have shown, with our method independent and novel evidence is generated, which can be consistently combined with the results of phylogenetic analyses. It thus may serve as a useful aid to phylogenetic approaches in the characterization of HGT events.

CHAPTER 5

# Integration into the genome annotation system GenDB

GenDB is an open-source genome annotation system, which is currently in worldwide use for the annotation of more than a dozen microbial genomes. This chapter describes the integration of the developed methods and programs of this work into the system. This allows their use in the microbial genome annotation projects and other whole-genome related applications. In combination with the virtual 2D-gel utility of GenDB, the CoBias program can be used for the *in silico* simulation of 2D-gel electrophoretic experiments. A number of programs for gene prediction and the combined gene finding strategies developed in chapter 1 were incorporated into the gene prediction component of GenDB. The combined strategies are currently applied in several microbial genome projects.

## 5.1 Introduction

The process of genome annotation can be defined as assigning meaning to sequence data that would otherwise be almost devoid of information. By identifying regions of interest and defining putative functions for those areas, the genome can be understood and further research initiated. Annotation is generally thought to possess the best quality when performed by a human expert. Due to the vast amount of data which has to be evaluated, software assistance for computation, storage, retrieval and analysis of relevant data has become essential for the success of any genome project.

A number of genome annotation systems intended for the analysis of prokaryotic and eukaryotic organisms have been designed and presented in the last few years. The first generation focussed primarily on generating human readable HTML documents based on tables and sometimes in-line graphics. These include the MAGPIE [18], GeneQuiz [122] and Pedant [123] systems. The intuitive visualizations provided by MAGPIE and the splitting of results by significance levels to enable comparison of different tools (also MAGPIE) can be found in most of the later on developed systems. A second generation of mostly commercial genome annotation systems was published, including ERGO (Integrated Genomics, Inc.), Pedant-Pro (successor to Pedant, Biomax Informatics AG), Phylosopher (Gene Data, Inc.), BioScout (Genequiz, Lion AG), WIT [124] and the open source systems Artemis [125]. Some systems (MAGPIE, Artemis, and Phylosopher) contain extensive visualizations or include multiple genome comparison based annotation strategies (most notably by ERGO [126]). With the exception of Artemis, all systems provide an automatic annotation feature, which except for ERGO are variants of a "best blast hit" strategy. Only MAGPIE, Artemis and the newer versions of Pedant allow the integration of expert knowledge through manual annotation.

The substantial commercial interest in the area of genome annotation led to a situation where, with the noted exception of Artemis, no genome annotation system was in the public domain. The resulting need for a well designed and documented open source genome annotation system led to the development of GenDB. GenDB is a flexible and easily extensible system, which is currently in worldwide use for the annotation of more than a dozen novel microbial genomes. As with the very successful Linux computer operating system, the open source license of GenDB enables the cooperative development of high quality software for genome annotation. The system is intended to provide a flexible, transparent infrastructure for genome projects, which can easily be adopted and modified to meet the requirements of different groups. For a more detailed and technical description of the system design, the reader is referred to [17]. In the following section, the GenDB data model and tool concept, which are important for understanding parts of this work are described in more detail.

### 5.1.1 Data model

GenDB uses a very simple data model, that is based on only three core types of classes. Regions describe arbitrary (sub-) sequences. A region can be related to a parent region, for example a coding sequence (CDS) is part of a contig. Observations correspond to information computed by bioinformatics tools such as BLAST [35] or InterPro [127] for a region. Annotations store the interpretation of a (human) annotator. They describe regions based on the evidence stored in the observations. Figure 5.1 shows the relationships between the

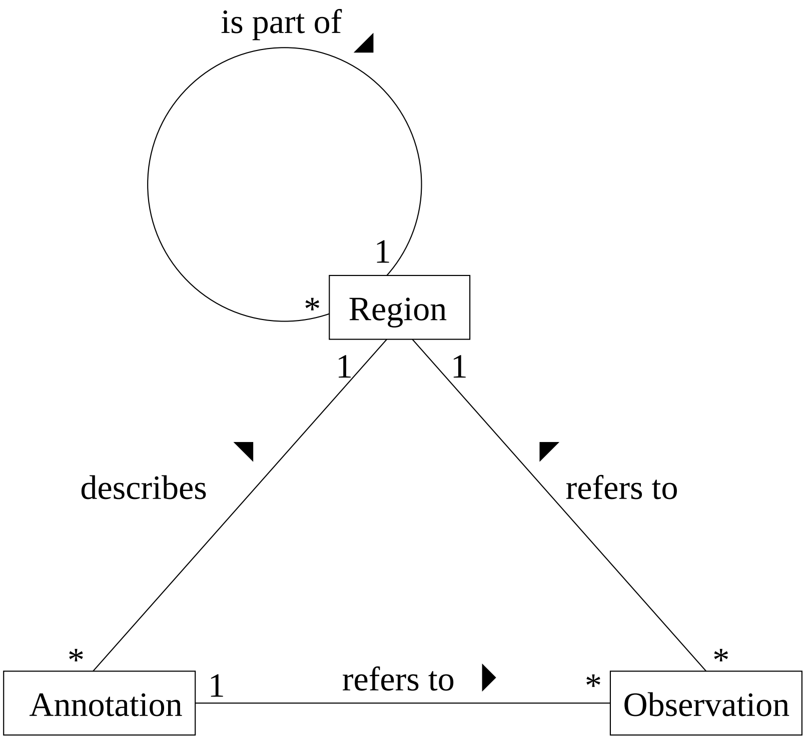


Figure 5.1: The core data model of GenDB in UML [129]. Only the three central classes are shown, these actually represent a hierarchy of specialized classes such as BLAST and InterPro observation classes.

different core classes. As can be seen, there is a clear distinction between the results from the bioinformatics tools (observations) and their interpretation (annotations), implemented in the data model. The data model represents a hierarchy of classes, including the prokaryotic components of the EMBL feature set [128] with several extensions. The three core classes are complemented by additional classes such as tools and annotators.

Since data access is via instances of the classes described above, the classes in GenDB themselves form the API (application programmers interface). This object-oriented approach makes code maintenance easy and also the data and methods of the system accessible to other programs. At the same time, a means to extend the GenDB system is provided.

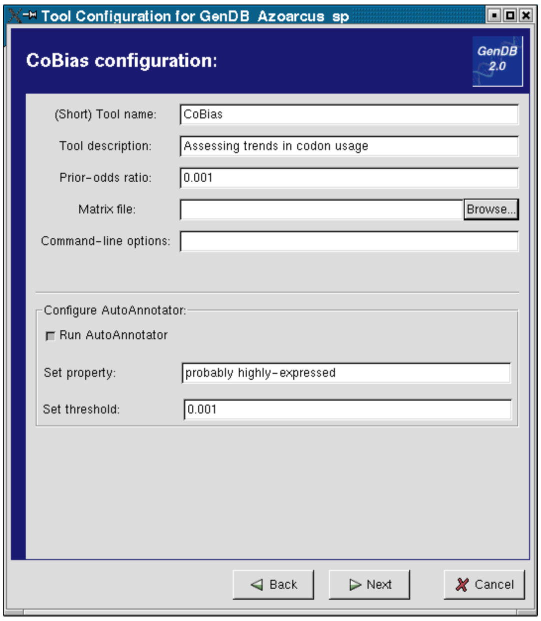
### 5.1.2 Tool concept

GenDB allows the incorporation of arbitrary programs for different kinds of computational analysis. According to the system design, these programs are integrated as tools that create observations for a specific kind of region. Within the GenDB tool concept two major subclasses of tools exist, the *region* and the *function* tools. The former are programs intended for the prediction of different types of regions realized in the data model, such as coding sequence (CDS), tRNA, rRNA, ribosome binding site (RBS) or operon. The latter comprise programs which deliver function-related information such as the result of BLAST [35] or hmmpfam [130] database searches. For the prediction of regions, Glimmer [20], Critica [25] Getorf [30] and tRNAscan-SE [131] have been integrated into the system. Similarity searches on DNA or amino acid level against arbitrary sequence databases can be done using the BLAST program suite. In addition to using HMMER [130] for motif searches, the BLOCKS [132] and INTERPRO databases can also be searched to classify sequence data based on a combination of different kinds of motif search tools. A number of additional tools have been integrated for the characterization of certain features of coding sequences, such as TMHMM [133] for the prediction of $\alpha$-helical transmembrane regions, SignalP [134] for signal peptide prediction or CoBias [42] for analyzing trends in codon usage. The gene prediction component of GenDB, which was developed as part of this work and the integration of the CoBias program are described in more detail in the following sections.

## 5.2 Integration of CoBias

As described in chapter 2, the CoBias program can be applied to two-class discrimination problems which may be solved using features in synonymous codon usage. Examples are the discrimination between highly and not highly expressed genes (chapter 3) or detecting genes introduced by horizontal gene transfer (chapter 4). To faciliate its application, CoBias was integrated as a *function* tool in the GenDB system. Figure 5.2 shows the configuration page for CoBias within GenDB.

CoBias can also be applied to confer a more realistic appearance to a 2D-gel simulation (chapter 3). For this, $S(g)$-scores of expression-level dependent features in synonymous codon usage are used for the simulation of different absolute amounts of protein. Assuming a spherical spot size, the spot radius is calculated based on the $S(g)$-scores. Figure 5.3 shows such a simulation for the *Corynebacterium glutamicum* genome with the GenDB virtual 2D-gel utility.



(Short) Tool name: CoBias

Tool description: Assessing trends in codon usage

Prior-odds ratio: 0.001

Matrix file:

Command-line options:

Configure AutoAnnotator:

☑ Run AutoAnnotator

Set property: probably highly-expressed

Set threshold: 0.001

Figure 5.2: The CoBias configuration page of the GenDB tool configuration wizard. For the creation of a novel CoBias tool, a matrix file with log-odds scores of codon usage has to be specified. The automatic annotation method for the CoBias results can be configured in the lower section.

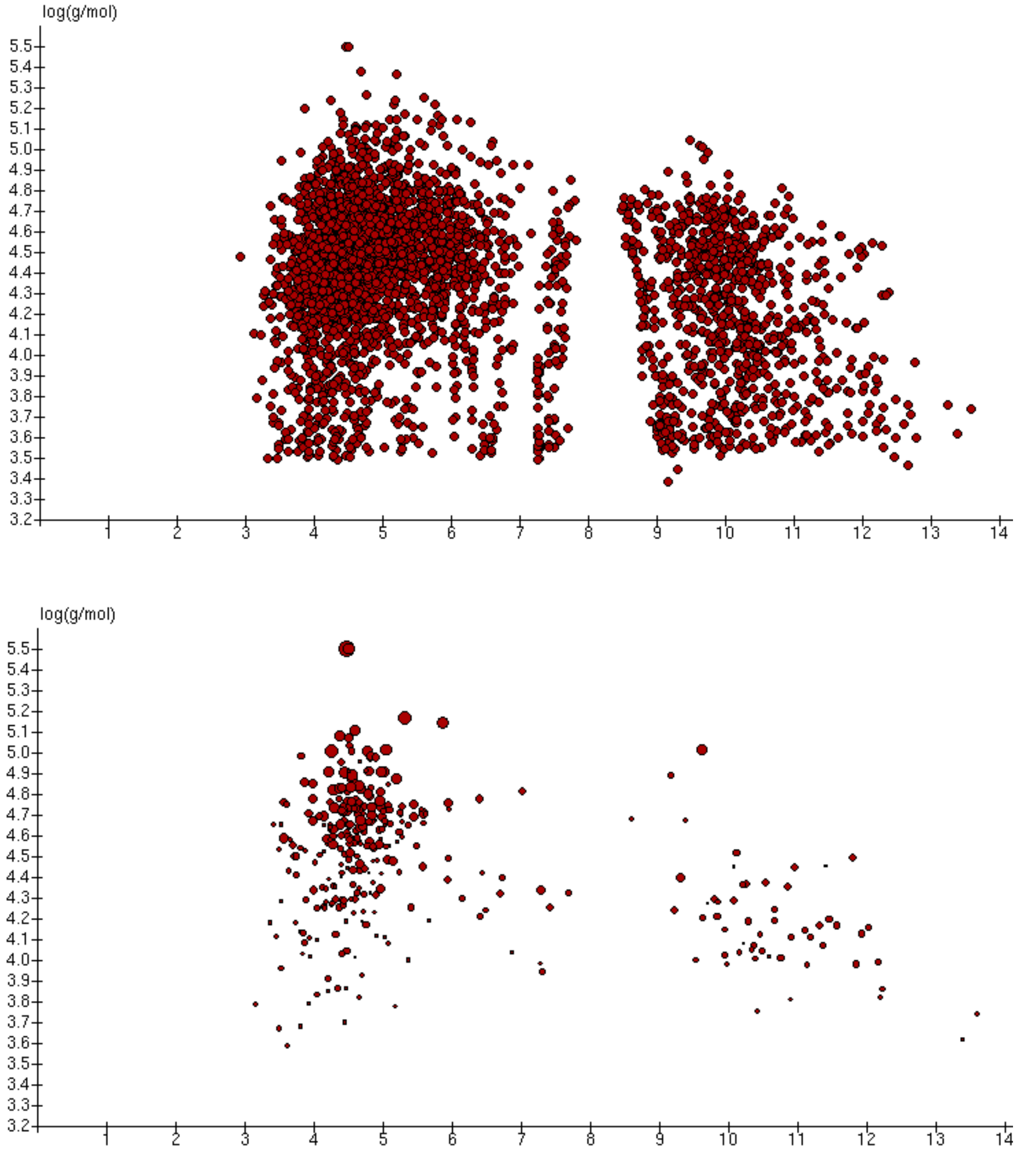


Figure 5.3: Using CoBias observations to simulate different amounts of protein in the virtual 2D-gel. Displayed is a 2D-gel simulation for the *Corynebacterium glutamicum* genome without (top) and with the use of CoBias observations (bottom). For the calculations, a codon usage matrix of expression-level dependent features in the codon usage of *C. glutamicum* was used. Spot volume was calculated by setting $S(g)$-scores proportional to the spot volume under the assumption of a spherical spot size.

## 5.3 The GenDB gene prediction component

The region annotation step of prokaryotic genome projects usually combines running gene finding tools with further automated or manual steps of result interpretation and validation. For the complete procedure, different groups have developed different strategies. A key requirement for the gene prediction component of a genome annotation system is thus flexibility. Novel tools for region prediction should be easy to integrate. Different strategies for the automation of the result interpretation step (autoannotation strategies) should be realizable. The aim is the generation of high-quality annotation data by combining both automated and human annotation efforts and reducing superfluous manual annotation efforts. Due to its flexible design, this can easily be realized within GenDB. As described in the previous section, novel tools can simply be integrated and used to create observations. Because of the separation of the tool run and the result interpretation (annotation) step, there is further room for flexibility in the choice of strategies for the (partial) automation of the region annotation step. These can be implemented as automatic annotation methods.

### 5.3.1 Gene finding

For the prediction of regions such as CDSs, tRNA-genes or RBSs, several programs have been integrated into GenDB. The program Getorf from the EMBOSS package [30] allows the determination of all open reading frames (ORFs) in a DNA sequence. For the prediction of CDSs, Glimmer [19, 20] and Critica [25] have been integrated into the system. The tRNAscan-SE program [131] can be used for the prediction of tRNA genes. In the following, the integation of Glimmer and Critica is described.

#### Glimmer

Glimmer is an *ab initio* gene finder which uses a statistical model of CDS properties to discriminate between CDSs and hypothetical ORFs. Its application can be divided into a training and a prediction phase (Figure 5.4). In the training phase, a statistical model of coding sequences is built from a set of CDSs. In the prediction phase, the model is applied to detect CDSs among the ORFs contained in a piece of genomic sequence. The Glimmer package contains a number of different programs, among these are *long-orfs*, *extract*, *build-icm* and *glimmer2*. The application of these programs in the intended order on a piece of genomic sequence results in the (default) prediction of CDSs (Figure 5.4). In that case, the *training set* of CDSs are the *long-orfs* results on the sequence.

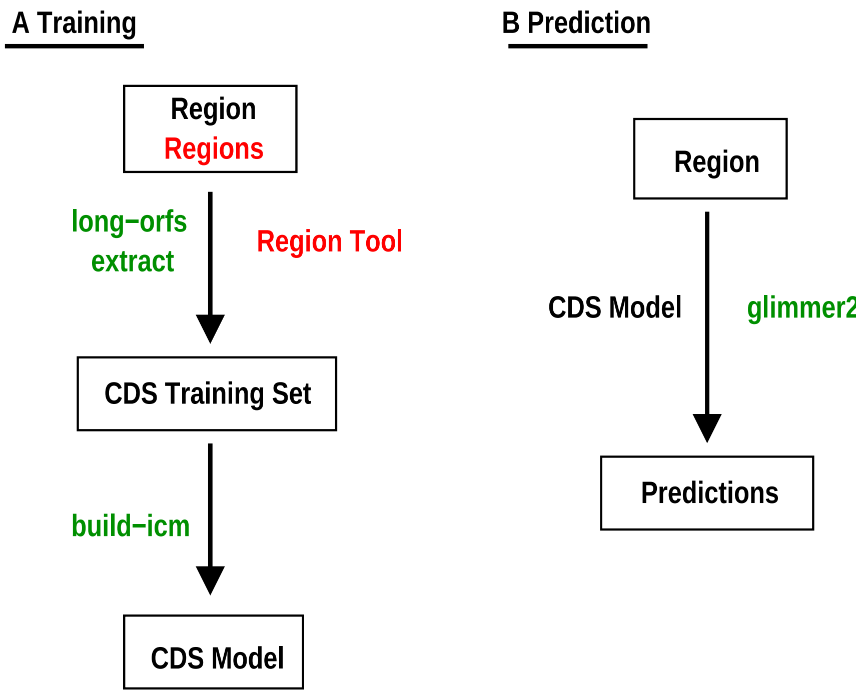
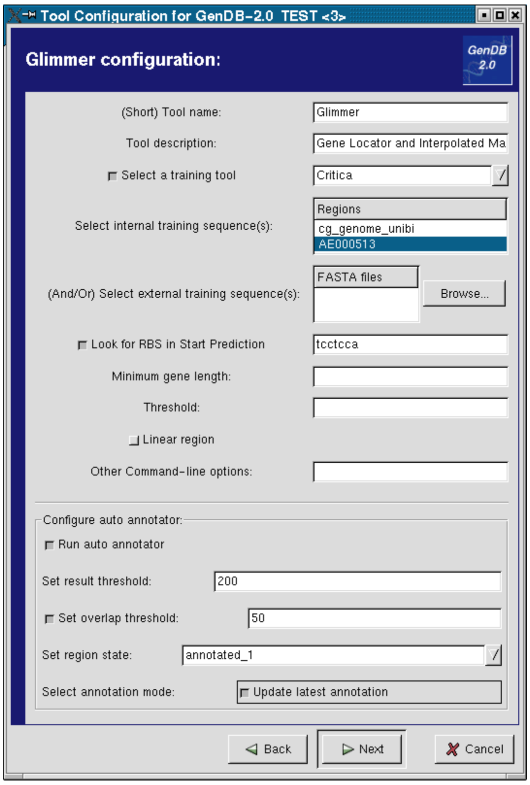


Figure 5.4: Default (green) and extended (red) application of Glimmer [19, 20] in GenDB. (A) Training of the Glimmer model. The training set of CDSs is extracted from a set of genomic *training sequences* using either *long-orfs* or a GenDB *training tool*. (B) Predictive step in which a previously created model is used.

The composition of the *training set* of CDSs can strongly influence the characteristics of the generated model and thus the quality of the gene finding results (chapter 1). Ideally, the set should consist of a representative and sufficiently large sample of the CDSs for the respective genome. To confer flexibility for different applications, the *training set* can be created in multiple ways in GenDB (Figure 5.4), which has been modelled as part of the Glimmer tool configuration. The creation of novel Glimmer tool requires specification of a set of *training sequences* and, optionally, a *training tool*. A *training sequence* can be any region object of a suitable type (Source, Contig, Partial Region) in GenDB or an external sequence file in FASTA format. As an option to *long-orfs*, other gene finding programs which have been run on the specified *training sequences* may be used as *training tool*.

Examples where a deviation from the default application may result in improved gene finding is the prediction of genes for GC-rich genomes (chapter 1), local recomputation of results after frame-shift corrections of the sequence or gene prediction during the assembly phase of a genome project, when the sequence is still split into more than one contig.



Tool Configuration for GenDB-2.0 TEST <3>

Glimmer configuration:

(Short) Tool name: Glimmer

Tool description: Gene Locator and Interpolated Ma

☑ Select a training tool: Critica

Select internal training sequence(s): Regions: cg_genome_unibi, AE000513 (selected)

(And/Or) Select external training sequence(s): FASTA files

☑ Look for RBS in Start Prediction: tcctcca

Minimum gene length:

Threshold:

☐ Linear region

Other Command-line options:

Configure auto annotator:

☑ Run auto annotator

Set result threshold: 200

☑ Set overlap threshold: 50

Set region state: annotated_1

Select annotation mode: ☑ Update latest annotation

Figure 5.5: The Glimmer configuration page of the GenDB tool configuration wizard. As part of the tool specification, *training sequence(s)* and, optionally, a *training tool* can be defined. In the lower section, the automatic Glimmer annotator can be configured.

Given a specified *training set* of CDSs, this is subsequently used by *build-icm* to train the Glimmer ICM model. In the final predictive step, the model is utilized by *glimmer2* for CDSs prediction on a genomic sequence. The results are stored as observation objects of the type CDS in GenDB. Figure 5.5 shows the Glimmer tool configuration page of GenDB.

### Critica

The gene finder Critica does not have the modular design to allow the separation of the training from the predictive step. Within GenDB, the default application of the program is possible, which is running the program on a specified region object. In addition to CDS finding results, from the program output information about the location of ribosome binding sites (RBSs) and frame shifts in the sequence is available. This data is stored by generating observations of the corresponding types in GenDB.

### 5.3.2 Automatic annotation

#### Simple automatic annotators

For automation of the result interpretation, simple automatic annotation methods have been implemented for all region tools. Using these, observations can be used to automatically create or update regions of the corresponding type. Several concepts apply to these autoannotators:

- **Create or update:**
  An observation is used to create a region, if there is no corresponding region yet. Otherwise, it is used to update the region. For CDS regions, this means altering the start position, if the current start of the region disagrees with the start of the observation.

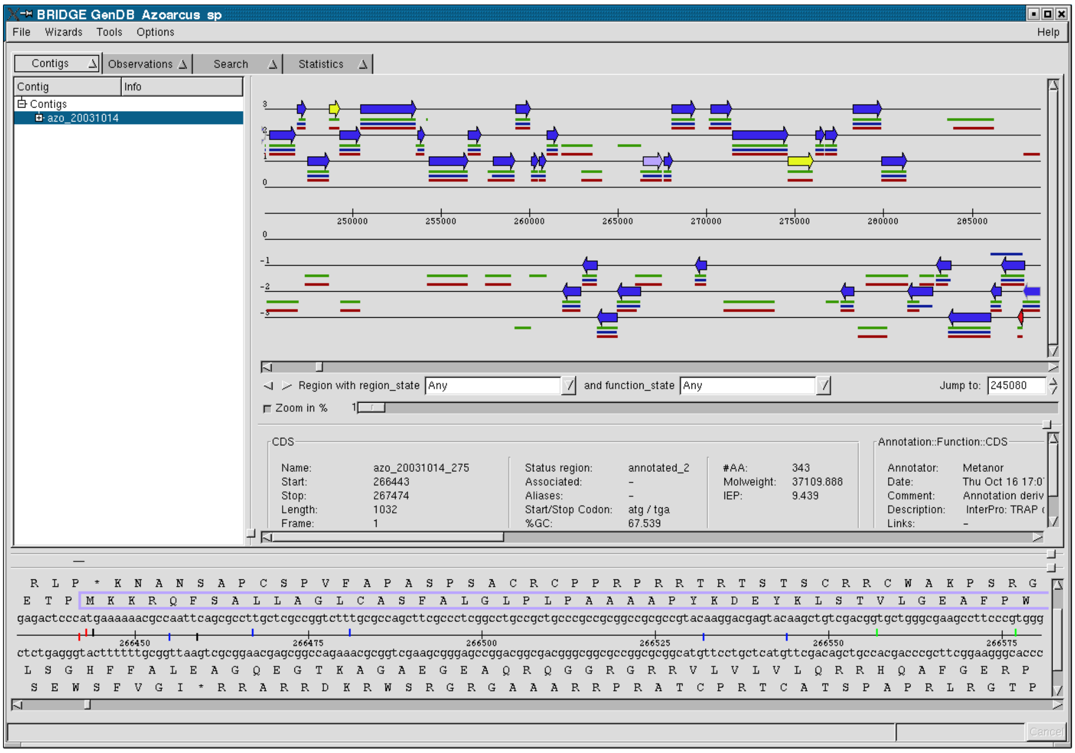
- **Regions with human annotations are left unchanged:**
  In case an annotation of a human annotator exists for a region, it is left unchanged. This ensures that no manual annotation effort is lost.

- **Update depending on the region status:**
  Every region can be assigned one of six region stati, which represent a confidence estimate that a biological counterpart to the modeled region exists. During configuration of the autoannotation methods, the region status that is to be assigned to the newly created or updated methods must be specified. The region states possess an internal order, which is ’attention needed’< ’ignored’< ’putative’ < ’annotated_1’ < ’annotated_2’ < ’annotated_3’ < ’finished’. To avoid a re-annotation of regions which have been annotated by more reliable methods, an update is only performed if the newly assigned status is higher than the current status of a region.

- **Threshold setting for tool results:**
  Some tools return a numerical score as a confidence estimate in a prediction. During configuration of the region autoannotators a threshold can be specified. Observations with scores equal to or higher than this setting are assigned the configured region status. Observations with lower scores than threshold are set to ’attention needed’.

#### Combined automatic annotation methods

By combining the automatic annotation methods described above, more complex annotation strategies can be realized. As an example, realization of the **Overlap Threshold Strategy** (**OTS**) and **Vote Score Threshold Strategy** (**VTS**) (chapter 1) is described:



Region with region_state Any and function_state Any Jump to: 245080

Zoom in % 1

CDS

```
Name:         azo_20031014_275     Status region:      annotated_2     #AA:        343
Start:        266443               Associated:         -               Molweight:  37109.888
Stop:         267474               Aliases:            -               IEP:        9.439
Length:       1032                 Start/Stop Codon:   atg / tga
Frame:        1                    %GC:                67.539
```

Annotation::Function::CDS

```
Annotator:    Metanor
Date:         Thu Oct 16 17:0
Comment:      Annotation deriv
Description:  InterPro: TRAP (
Links:        -
```

```
 R  L  P  *  K  N  A  N  S  A  P  C  S  P  V  F  A  P  A  S  P  S  A  C  R  C  P  P  R  P  R  R  T  R  T  S  T  S  C  R  R  C  W  A  K  P  S  R  G
E  T  P  M  K  K  R  Q  F  S  A  L  L  A  G  L  C  A  S  F  A  L  G  L  P  L  P  A  A  A  A  P  Y  K  D  E  Y  K  L  S  T  V  L  G  E  A  F  P  W
gagactcccatgaaaaaacgccaattcagcgccttgctcgccggtctttgcgccagcttcgccctcggcctgccgctgcccgccgcggccgcgccgtacaaggacgagtacaagctgtcgacggtgctgggcgaagccttcccgtggg
        266450                   266475                   266500                   266525                   266550                   266575
ctctgagggtactttttttgcggttaagtcgcggaacgagcggccagaaacgcggtcgaagcgggagccggacggcgacgggcggcgccggcgcggcatgttcctgctcatgttcgacagctgccacgacccgcttcggaagggcaccc
L  S  G  H  F  F  A  L  E  A  G  Q  E  G  T  K  A  G  A  E  G  E  A  Q  R  Q  G  G  R  G  R  R  V  L  V  L  V  L  Q  R  R  H  Q  A  F  G  E  R  P
 S  E  W  S  F  V  G  I  *  R  R  A  R  R  D  K  R  W  S  R  G  R  G  A  A  A  R  R  P  R  A  T  C  P  R  T  C  A  T  S  P  A  P  R  L  R  G  T  P
```

Figure 5.6: The GenDB contig view displaying a section of the *Azoarcus* sp. BH72 genome after gene prediction and automatic annotation according to the OTS strategy. Region objects such as CDSs or tRNA genes are represented by arrows on the corresponding reading frames. Regions are colored according to their region status (*blue* - *annotated_2*, *yellow* - *annotated_1*, *red* - *attention neeeded*). Glimmer (*green*), Glimmer(ct) (*red*) and Critica (*blue*) observations are displayed below the different frames.

- **OTS**
  Create a Critica tool and configure the autoannotator to annotate all Critica observations as CDSs with region status ’annotated_2’. Run Critica. Create a Glimmer tool which uses the Critica observations as a training set. Configure the autoannotator to annotate only additional Glimmer observations which do not overlap more than 50bp with existing regions (set overlap threshold to 50). Configure autoannotator to annotate Glimmer observations with region status ’annotated_1’. Run Glimmer.

- **VTS**
  Proceed as described with the **OTS** strategy. In configuration of the Glimmer autoannotator, set the ’tool result threshold’ to 100. This results in the assignment of the region

status ’attention needed’ to additional Glimmer observations with a tool result (Vote score) smaller than 100.

### 5.3.3 Application in microbial genome projects

At the time of this writing, the described gene prediction and autoannotation components have already been used in several microbial genome projects. Among these is the recently finished *Corynebacterium glutamicum* genome [135]. In all cases, tRNAscan-SE was run for the detection and autoannotation of tRNA genes. Glimmer, Critica and Glimmer(ct) (chapter 1) were used for the prediction of CDSs. An autoannotation step based on the OTS strategy was applied for annotation of *Alcanivorax borkumensis*, *Azoarcus* sp., *Xanthomonas campestris* pv. vesicatoria, *Listeria welshimeri*, *Clavibacter michiganensis* subsp. michiganensis and *Bdellovibrio bacteriovorus* (Figure 5.6).

# Bibliography

[1] R. D. Fleischmann, M.D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269:496–512, 1995.

[2] H. W. Mewes, K. Albermann, M. Bahr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S. G. Oliver, F. Pfeiffer, and A. Zollner. Overview of the yeast genome. *Nature*, 387:7–65, 1997.

[3] C. M. Fraser and Fleischmann. R. D. Strategies for whole microbial genome sequencing and analysis. *Electrophoresis*, 18:1207–1216, 1997.

[4] C. Mathé, M. Sagot, T. Schiex, and P. Rouze. Current methods of gene prediction, their strength and weaknesses. *Nucleic Acids Res.*, 30:4103–4117, 2002.

[5] H. Ge, A. J. Walhout, and M. Vidal. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.*, 19:551–560, 2003.

[6] A. Osterman and R. Overbeek. Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.*, 7:238–251, 2003.

[7] Eichler E. E. and D. Sankoff. Structural dynamics of eukaryotic chromosome evolution. *Science*, 301:793–797, 2003.

[8] M. J. Sanderson and A. C. Driskell. The challenge of constructing large phylogenetic trees. *Trends Plant Sci.*, 8:374–379, 2003.

[9] C. G. Kurland, B. Canback, and O. G. Berg. Horizontal gene transfer: A critical view. *Proc. Natl. Acad. Sci. USA*, 100:9658–9662, 2003.

[10] B. Snel, P. Bork, and M. A. Huynen. Genomes in flux: The evolution of archaeal and proteobacterial gene content. *Genome Research*, 12:17–25, 2002.

[11] K. A. Frazer, L. Elnitski, D. M. Church, I. Dubchak, and R. C. Hardison. Cross-species sequence comparisons: A review of methods and available resources. *Genome Research*, 13:1–12, 2003.

[12] P. Chain, S. Kurtz, E. Ohlebusch, and T. Slezak. An application-focused review of comparative genomics tools: Capabilities, limitations and future challenges. *Brief. Bioinf.*, 4:105–123, 2003.

[13] L. Wei, Y. Liu, I. Dubchak, J. Shon, and J. Park. Comparative genomics approaches to study organism similarities and differences. *J. Biomed. Inform.*, 35:142–150, 2002.

[14] D. R. Boone, R. W. Castenholz, and G. M. Garrity. *Bergey's manual of systematic bacteriology*. Springer, New York, 2nd edition, 2001.

[15] K. M. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.-A. Rajandream, and B. Barrell. Artemis: sequence visualization and annotation. *Bioinformatics*, 16:944–945, 2000.

[16] R. Overbeek, N. Larsen, T. Walunas, M. D'Souza, G. Pusch, E. Selkov, K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, et al. The ERGO genome analysis and discovery system. *Nucleic Acids Res.*, 31:164–171, 2003.

[17] F. Meyer, A. Goesmann, A. C. McHardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinowski, B. Linke, O. Rupp, R. Giegerich, and A. Puhler. GenDB - an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, 31:2187–2195, 2003.

[18] T. Gaasterland and C. W. Sensen. MAGPIE: automated genome interpretation. *Trends Genet.*, 12:76–78, 1996.

[19] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, 26:544–548, 1998.

[20] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, 27:4636–4641, 1999.

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

[22] J. Besemer, A. Lomsadze, and M. Borodovsky. 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, 2001.

[23] F.-B. Guo, H.-Y. Ou, and C.-T. Zhang. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.*, 31:1780–1789, 2003.

[24] T. S. Larsen and A. Krogh. EasyGene - a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, 4:21, 2003.

[25] J. H. Badger and G. J. Olsen. CRITICA: Coding Region Identification Tool Invoking Comparative Analysis. *Mol. Biol. Evol.*, 16:512–524, 1999.

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

[27] T. Shibuya and I. Rigoutsos. Dictionary-driven prokaryotic gene finding. *Nucleic Acids Res.*, 30:2710–2725, 2002.

[28] M. Tech and R. Merkl. YACOP: Enhanced gene prediction obtained by a combination of existing methods. *Bioinformatics*, in press.

[29] B. E. Suzek, M. D. Ermolaeva, M. Schreiber, and Salzberg S. L. A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, 17:1123–1130, 2001.

[30] S. A. Olson. EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief. Bioinf.*, 3:87–91, 2002.

[31] P. Baldi and S. Brunak. *Bioinformatics - The machine learning approach*, pages 155–163. MIT Press, Cambridge Massachusetts, London, England, 2001.

[32] J.A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293, 1988.

[33] M. Gribskov and N. L. Robinson. The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, 20:25–33, 1996.

[34] A. A. Schaeffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, 29:2994–3005, 2001.

[35] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.

[36] M. S. Skovgard, L. J. Jensen, S. Brunak, D. Ussery, and A. Krogh. On the total number of genes and their length distribution in complete microbial genomes. *Trends. Genet.*, 17:425–427, 2001.

[37] S. T. Cole, K. Eiglmeier, K. D. James, N. R. Thomson, P. R. Wheeler, N. Honore, T. Garnier, C. Churcher, D. Harris, K. Mungall, et al. Massive gene decay in the leprosy bacillus. *Nature*, 409:1007–1011, 2001.

[38] I. B. Rogozin, D. D'Angelo, and L. Milanesi. Protein-coding regions prediction combining similarity searches and conservative evolutionary properties of protein-coding sequences. *Gene*, 226:129–137, 1999.

[39] R. J. Grocock and P. M. Sharp. Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene*, 289:131–139, 2002.

[40] P. M. Sharp and W.-H. Li. The Codon Adaption Index - a measure of directional synonymous codon usage bias, and its partial applications. *Nucleic Acids Res.*, 15:1281–1295, 1987.

[41] R. Jansen, H. J. Bussemaker, and M. Gerstein. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using variety of models. *Nucleic Acids Res.*, 31:2242–2251, 2003.

[42] A. C. McHardy, A. Pühler, J. Kalinowski, and F. Meyer. Comparing expression-level dependent features in codon usage with protein abundance: An analysis of 'predictive proteomics'. *Proteomics*, in press.

[43] R. D. Knight, S. J. Freeland, and L. F. Landweber. A simple model based on mutation and selection explains the trends in codon and amino acid usage and GC composition within and across genomes. *Genome Biol.*, 2:RESEARCH0010, 2001.

[44] D. J. Lynn, G. D. Singer, and D. A. Hickey. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.*, 30:4272–4277, 2002.

[45] B. Lafay, A. T. Lloyd, M. J. McLean, K. M. Devine, P. M. Sharp, and K. H. Wolfe. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res.*, 27:1642–1649, 1999.

[46] S. Karlin and J. Mrázek. Predicting highly expressed genes of diverse procaryotic genomes. *J. Bacteriol.*, 182:5238–5250, 2000.

[47] J. Mrazék and S. Karlin. Detecting alien genes in bacterial genomes. *Ann. N. Y. Acad. Sci.*, 870:314–329, 1999.

[48] J. G. Lawrence and H. Ochman. Amelioration of bacterial genomes: Rates of change and exchange. *J. Mol. Evol.*, 44:383–397, 1997.

[49] S. Garcia-Vallvé, A. Romeu, and J. Palau. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Research*, 10:1719–1725, 2000.

[50] S. Karlin, J. Mrázek, and A. M. Cambell. Codon usage in different gene classes of the *Escherichia coli* genome. *Mol. Microbiol.*, 29:1341–1355, 1998.

[51] S. Garcia-Vallvé, E. Guzman, M. A. Montero, and A. Romeu. HGT-DB: a database of putative horizontally transferrred genes in prokaryotic complete genomes. *Nucleic Acids Res.*, 31:187–189, 2003.

[52] S. F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, 219:555–565, 1991.

[53] R. Durbin, S. Eddy, A. Krogh, and G. Mitchinson. *Biological sequence analysis*. Cambridge University Press, Cambridge, 1998. pp.36-41.

[54] W. J. Ewens and G. R. Grant. *Statistical methods in bioinformatics: An introduction*, chapter Statistics(ii): Classical estimation and hypothesis testing. Springer-Verlag, New York, 2001.

[55] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[56] L. Wall, T. Christiansen, and J. Orwant. *Programming Perl*. O' Reilly, $3^{rd}$ edition, 2000.

[57] O. G. Berg and P. J. Silva. Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection. *Nucleic Acids Res.*, 25:1397–1404, 1997.

[58] S. D. Hooper and O. G. Berg. Gradients in nucleotide and codon usage along *Escherichia coli* genes. *Nucleic Acids Res.*, 28:3517–3523, 2000.

[59] A. Yadava and C. F. Ockenhouse. Effect of codon optimization on expression levels of a functionally folded malaria vaccine candidate in prokaryotic and eukaryotic expression systems. *Infect. Immun.*, 71:4961–4969, 2003.

[60] S. J. Park, S. K. Lee, and B. J. Lee. Effect of tandem rare codon substitution and vector-host combinations on the expression of the EBV gp110 C-terminal domain in *Escherichia coli*. *Protein. Expr. Purif.*, 24:470–480, 2002.

[61] Y. Li, C. X. Chen, B. U. von Specht, and H. P. Hahn. Cloning and hemolysin-mediated secretory expression of a codon-optimized synthetic human interleukin-6 gene in *Escherichia coli*. *Protein Expr. Purif.*, 25:437–447, 2002.

[62] D. L. Lakey, R. K. Voladri, K. M. Edwards, C. Hager, B. Samten, R. S. Wallis, P. F. Barnes, and D. S. Kernodle. Enhanced production of recombinant *Mycobacterium tuberculosis* antigens in *Escherichia coli* by replacement of low-usage codons. *Infect. Immun.*, 68:233–238, 2000.

[63] N. Hansmeier, A. Tauch, A. Pühler, and J. Kalinowski. Classification of *Corynebacterium glutamicum* surface-layer proteins by sequence analyses and atomic force microscopy. *manuscript in preparation*.

[64] U. B. Sleytr. Basic and applied S-layer research: an overview. *FEMS Microbiol. Reviews*, 20:5–12, 1997.

[65] H. J. Boot and P. H. Pouwels. Expression, secretion and antigenic variation of bacterial S-layer proteins. *Mol. Microbiol.*, 21:1117– 1123, 1996.

[66] K. E. Nelson, R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, et al. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, 399:323–329, 1999.

[67] C. L. Nesbø, S. L'Haridon, K. O. Stetter, and W. F. Dolittle. Phylogenetic analysis of two 'archaeal' genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria. *Mol. Biol. Evol.*, 18:362–375, 2001.

[68] P. Wornign, L. J. Jensen, K. E. Nelson, S. Brunak, and D. W. Ussery. Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. *Nucleic Acids Res.*, 28:706–709, 2000.

[69] P. M. Sharp and W.-H. Li. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.*, 24:23–28, 1986.

[70] M. Bulmer. The selection-mutation drift theory of synonymous codon usage. *Genetics*, 129:897–907, 1991.

[71] T. Ikemura. Correlation between the abundance of *Escherichia coli* transfer tRNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.*, 151:389–409, 1981.

[72] T. Ikemura. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, 2:13–34, 1985.

[73] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.*, 9:R43–R74, 1981.

[74] T. Ikemura. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in its protein genes: differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNA. *J. Mol. Biol.*, 158:573–597, 1982.

[75] H. Dong, L. Nilsson, and C.G. Kurland. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.*, 260:649–663, 1996.

[76] K. Kanayama, Y. Yamada, Y. Kudo, and T. Ikemura. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238:143–155, 1999.

[77] H. Grosjean and W. Fiers. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene*, 18:199–209, 1982.

[78] M. J. McLean, K. H. Wolfe, and K. M. Devine. Base composition skews, replication origin, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.*, 47:691–696, 1998.

[79] A. Eyre-Walker. Synonymous codon bias is related to gene length in *Escherichia coli*: Selection for translational accuracy. *Mol. Biol. Evol.*, 13:7055–7074, 1996.

[80] M. Yarus and L.S. Folley. Sense codons are found in specific contexts. *J. Mol. Biol.*, 182:529–540, 1985.

[81] M. Gouy. Codon contexts in enterobacterial and coliphage genes. *Mol. Biol. Evol.*, 4:426–444, 1987.

[82] G. McVean and G. Hurst. Evolutionary lability of context-dependent codon bias in bacteria. *J. Mol. Evol.*, 50:264–275, 2000.

[83] M. Bulmer. Codon usage and intragenic position. *J. Theor. Biol.*, 133:67–71, 1988.

[84] A. Eyre-Walker and M. Bulmer. Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res.*, 21:4599–4603, 1993.

[85] A. Eyre-Walker. The close proximity of *Escherichia coli* genes: consequences for stop codon and synonymous codon use. *J. Mol. Evol.*, 42:73–78, 1996.

[86] H. Sakai, C. Imamura, Y. Osada, R. Saito, T. Washio, and M. Tomita. Correlation between Shine-Dalgarno sequence conservation and codon usage of bacterial genes. *J. Mol. Evol.*, 52:164–170, 2001.

[87] A. Pan, C. Dutta, and J. Das. Codon usage in highly expressed genes of *Haemophilus influenca* and *Mycobacterium tuberculosis*: translational selection versus mutational bias. *Gene*, 215:405–413, 1998.

[88] D. Medjahed, G. W. Smythers, D. A. Powell, R. M. Stephens, P. F. Lemkin, and D. J. Munroe. VIRTUAL2D: a web-accessible predictive databasse for proteomics analysis. *Proteomics*, 3:129–138, 2003.

[89] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, J. Gregor, N.W. Davis, H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau, and Y. Shao. The complete genome sequence of *Escherichia coli* K-12. *Science*, 277:1453–1474, 1997.

[90] F. Kunst, N. Ogasawara, I. Moszer, A. M. Albertini, G. A. Alloni, V. Azevedo, M.G. Bertero, P. Bessieres, A. Bolotin, and S. Borchert. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, 390:249–56, 1997.

[91] K. Büttner, J. Bernhardt, C. Scharf, R. Schmid, U. Mäder, C. Eymann, H. Antelmann, A. Völker, and M. Hecker. A comprehensive two-dimensional map of cytosolic proteins of *Bacillus subtilis*. *Electrophoresis*, 22:2908–2935, 2001.

[92] L. Tonella, B. J. Walsh, J. Sanchez, K. Ou, et al. '98 *Escherichia coli* SWISS-2DPAGE database update. *Electrophoresis*, 19:1960–1971, 1998.

[93] S. Phadtare, J. Alsina, and M. Inouye. Cold-shock response and cold-shock proteins. *Curr. Opin. Microbiol.*, 2:175–180, 1999.

[94] G. Storz and J.A. Imlay. Oxidative stress. *Curr. Opin. Microbiol.*, 2:188–194, 1999.

[95] M. E. Gottesman and W. A. Hendrickson. Protein folding and unfolding by *Escherichia coli* chaperones and chaperonins. *Curr. Opin. Microbiol.*, 3:197–202, 2000.

[96] D. G. Fraenkel. In F.C. Neidhardt, editor, *Escherichia coli and Salmonella cellular and molecular biology*, pages 190–191. ASM Press, Washington, D. C., 2. edition, 1996.

[97] E. Ponce, N. Flores, A. Martinez, F. Valle, and F. Bolivar. Cloning of the two pyruvate kinase isoenzyme structural genes from *Escherichia coli*: the relative roles of these enzymes in the pyruvate biosynthesis. *J. Bacteriol.*, 177:5719–5722, 1995.

[98] P. Karp, M. Riley, M. Saier, I. T. Paulsen, J. Collado-Vides, S.M. Paley, A. Pelligrini-Toole, C. Bonavides, and S. Gama-Castro. The EcoCyc database. *Nucleic Acids Res.*, 30:56–58, 2002.

[99] W. Jiang, Y. Hou, and M. Inouye. CspA, the major cold shock protein of *Escherichia coli*, is an RNA chaperone. *J. Biol. Chem.*, 272:196–202, 1997.

[100] B. Weonhye, X. Bing, M. Inouye, and K. Severinov. *Escherichia coli* CspA-family RNA chaperones are transcription antiterminators. *Proc. Natl. Acad. Sci. USA*, 97:7784–7789, 2000.

[101] K. Yamanaka and M. Inouye. Selective mRNA degradation by polynucleotide phosphorylase in cold shock adaptation in *Escherichia coli*. *J. Bacteriol.*, 183:2808–2816, 2001.

[102] C. Medique, T. Rouxel, P. Vigier, A. Henaut, and A. Danchin. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.*, 222:851–856, 1991.

[103] I. Moszer, E. Rocha, and A. Danchin. Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr. Opin. Microbiol.*, 2:524–528, 1999.

[104] H. Ochman, J. G. Lawrence, and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405:299–304, 2000.

[105] E. V. Koonin, K. S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.*, 55:709–742, 2001.

[106] J. G. Lawrence and H. Ochman. Molecular archaeology of the *Escherichia coli* genome: rate of change and exchange. *Proc. Natl. Acad. Sci. USA*, 95:9513–9417, 1998.

[107] S. D. Hooper and O. G. Berg. Detection of genes with atypical nucleotide sequence in microbial genomes. *J. Mol. Evol.*, 54:365–375, 2002.

[108] M. A. Ragan. Detection of lateral gene transfer among microbial genomes. *Curr. Opin. Genet. Dev.*, 11:620–626, 2001.

[109] J. G. Lawrence and H. Ochman. Reconciling the many faces of lateral gene transfer. *Trends Microbiol.*, 10:1–3, 2002.

[110] J. G. Lawrence and J. R. Roth. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143:1843–1860, 1996.

[111] J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. R. Gilbert, I. Korf, H. Lapp, H. Lehväslaiho, et al. The Bioperl Toolkit: Perl modules for the life sciences. *Genome Research*, 12:1611–1618, 2002.

[112] T. Sicheritz-Ponten and S. G. E. Andersson. A phylogenomic approach to microbial evolution. *Nucleic Acids Res.*, 29:545–552, 2001.

[113] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, 31:365–370, 2003.

[114] J. D. Thompson, D. G. Higgins, and T. J. Gibson. ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

[115] D. L. Swofford. *PAUP: Progressive Analysis Using Parsimony (and other methods*. Sinauer Associates, Sunderland, MA, 1998. version 4.

[116] N. Saitou and M. Nei. The neighbor-joining method; a new method for reconstruction of phylogenetic trees. *Mol. Biol. Evol.*, 4:672–674, 1987.

[117] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *J. Comp. Graph. Stat.*, 5:299–314, 1996.

[118] A. I. Slesarev, K. V. Mezhevaya, K. S. Makarova, N. N. Polushin, O. V. Shcherbinina, V. V. Shakhova, G. I. Belova, L. Aravind, D. A. Natale, I. B. Rogozin, R. L. Tatusov, Y. I Wolf, K. O. Stetter, A. G. Malykh, E. V. Koonin, and S. A. Kozyavkin. The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. *Proc. Natl. Acad. Sci. USA*, 2002.

[119] K. Sowers and S. Baron. *Methanosarcina acetivorans* sp. nov., an acetotrophic methane-producing bacterium isolated from marine sediments (from the sumner branch of Scripps Canyon located near La Jolla, California). *Appl. Env. Microbiol.*, 47:971–978, 1984.

[120] M. Lai, C. Shu, M. Chiou, T. Hong, M. Chuang, and Hua J. J. Characterization of *Methanosarcina mazei* N2M9705 isolated from an aquaculture fishpond. *Curr. Microbiol.*, 39:79–84, 1999.

[121] Y. Xue, Y. Xu, Y. Liu, Y. Ma, and P. Zhou. *Thermoanaerobacter tengcongensis* sp. nov., a novel anaerobic, saccharolytic, thermophilic bacterium isolated from a hot spring in Tengcong, China. *Int. J. Syst. Evol. Microbiol.*, 51:1335–1341, 2001.

[122] G. Casari, C. Ouzounis, A. Valencia, and A. Sander. Genequiz II: automatic function assignment for genome sequence analysis. In *Proceedings of the First Annual Pacific Symposium on Biocomputing*, pages 707–709, Hawaii, 1996. World Scientific.

[123] D. Frishman, K. Albermann, J. Hani, K. Heumann, A. Metanomski, A. Zollner, and H.W. Mewes. Functional and structural genomics using PEDANT. *Bioinformatics*, 17:44–57, 2001.

[124] R. Overbeek, N. Larsen, G. D. Pusch, M. D'Souza, E. Selkov, N. Kyrpides, M. Fonstein, N. Maltsev, and E. Selkov. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, 28:123–125, 2002.

[125] K. M. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M-A. Rajandream, and B. Barrell. Artemis: sequence visualisation and annotation. *Bioinformatics*, 16:944–945, 2000.

[126] R. Overbeek, M. Fontstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA*, 96:2896–2901, 1999.

[127] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. R. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, T. M. Mulder, N. J.and Oinn, M. Pagni, F. Servant, C. J. A. Sigrist, and E. M. Zdobnov. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, 29:37–40, 2001.

[128] `http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html%`.

[129] B. Oestereich. *Objektorientierte Softwareentwicklung: Analyse und Design mit der UML*. R. Oldenbourg, $5^{th}$ edition, 2001.

[130] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[131] T. M. Lowe and S. R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, 25:955–964, 1997.

[132] S. Henikoff and J. G. Henikoff. Automated assembly of protein blocks for database searching. *Nucleic Acids Res.*, 19:6565–6572, 1991.

[133] E. L.L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden Markov model for predicting transmembrane helices in protein sequences. In J. Glasgow, T. Littlejohn, R. Major, F. Lathrop, D. Sankoff, and C. Sensen, editors, *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, pages 175–182, Menlo Park, CA, 1998. AAAI Press.

[134] H. Nielsen, J. Engelbrecht, S. Brunak, and G. Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 10:1–6, 1997.

[135] J. Kalinowski, B. Bathe, D. Bartels, N. Bischoff, M. Bott, A. Burkovski, N. Dusch, L. Eggeling, B. J. Eikmanns, L. Gaigalat, A. Goesmann, M. Hartmann, K. Huthmacher, R. Krämer, B. Linke, A. C. McHardy, F. Meyer, B. Möckel, W. Pfefferle, A. Pühler, D. A. Rey, C. Rückert, O. Rupp, H. Sahm, V. F. Wendisch, I. Wiegräbe, and A. Tauch. The complete genome sequence of the amino acid producing bacterium *Corynebacterium glutamicum* and its impact on amino acid production. *J. Biotech.*, 104:5–25, 2003.

# Erklärung

Hiermit versichere ich, die vorliegende Dissertation selbständig angefertigt und keine weiteren als die angegebenen Hilfsmittel und Quellen verwendet zu haben.

Bielefeld, im 2003 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Alice McHardy