Arch Microbiol (2008) 190:281–299
DOI 10.1007/s00203-008-0389-z

ORIGINAL PAPER

Genome information management and integrated data analysis with HaloLex

Friedhelm Pfeiffer · Alexander Broicher · Thomas Gillich · Kathrin Klee · José Mejía · Markus Rampp · Dieter Oesterhelt

Received: 7 January 2008 / Revised: 1 April 2008 / Accepted: 8 May 2008 / Published online: 1 July 2008

Abstract HaloLex is a software system for the central management, integration, curation, and web-based visualization of genomic and other *-omics* data for any given microorganism. The system has been employed for the manual curation of three haloarchaeal genomes, namely *Halobacterium salinarum* (strain R1), *Natronomonas pharaonis*, and *Haloquadratum walsbyi*. HaloLex, in particular, enables the integrated analysis of genome-wide proteomic results with the underlying genomic data. This has proven indispensable to generate reliable gene predictions for GC-rich genomes, which, due to their characteristically low abundance of stop codons, are known to be hard targets for standard gene finders, especially concerning start codon assignment. The proteomic identification of more than 600 N-terminal peptides has greatly increased the reliability of the start codon assignment for *Halobacterium salinarum*. Application of homology-based methods to the published genome of *Haloarcula marismortui* allowed to detect 47 previously unidentified genes (a problem that is particularly serious for short protein sequences) and to correct more than 300 start codon misassignments.

Communicated by Erko Stackebrandt.

F. Pfeiffer · A. Broicher · T. Gillich · K. Klee · D. Oesterhelt (✉)
Department of Membrane Biochemistry,
Max-Planck-Institute of Biochemistry,
Am Klopferspitz 18, 82152 Martinsried, Germany
e-mail: oesterhe@biochem.mpg.de

J. Mejía · M. Rampp
Computing Center (RZG) of the Max-Planck-Society,
Max Planck Institute of Plasma Physics,
Boltzmannstrasse 2, 85748 Garching, Germany

Keywords Halophilic archaea · Genome information system · Genome browser · Proteomics · Biological data curation · Start codon assignment · Dinucleotide bias

Introduction

In the era of high-throughput biochemical experiments and large-scale systems-modeling approaches, the availability of high-quality input data like the complete gene or protein inventory of an organism is of paramount importance (cf. Kitano 2002). In practice, however, progress is often hampered by the lack of access to the relevant data, their insufficient integration with related information, or simply inadequate reliability of the data. Although the current era is commonly referred to as "postgenomic," many problems related to the quality of (microbial) genome annotations are still not satisfactorily solved. The GC-rich genomes of halophilic archaea, for example, are known to pose particular challenges for the bioinformatic prediction of their gene and protein inventory. Unsupervised, automatic gene prediction likely fails and blindly relying on such data can apparently compromise any further analysis or experiment. Thus, there is not only a need for making genomic and other related data available to the end-user in a most convenient and comprehensive way, but also tools are required that support generating, managing, and manually curating the data and allow experts to assess and improve their quality.

To this end, we have developed HaloLex, which serves both the aforementioned purposes. Halolex is a software system for the central management, integration, and web-based visualization of genomic and other *-omics* data for any given microorganism. Centered on the genomic information, HaloLex provides a comprehensive and user-

friendly web interface (http://www.halolex.mpg.de) with many different interlinked views and various search functionalities to the underlying database. Advanced data mining tasks can be performed by employing high-level programming interfaces to access and automatically process bulk-data with computer scripts and programs.

The main scientific purpose of HaloLex is to support in-depth analysis of selected prokaryotic genomes and to assist knowledge-based manual revision and refinement of their annotation, in particular by taking into account also nongenomic, experimental data (e.g., proteomics). Typically, genomic data enter the system after automatic gene identification, classification, and basic annotations have been accomplished using a general-purpose genome annotation system like, e.g., "GenDB" (Meyer et al. 2003).

We are not trying to parallel seemingly similar efforts like the "integrated microbial genomes browser IMG" (Markowitz et al. 2006), the "UCSC Archaeal Genome Browser" (Schneider et al. 2006) or "PEDANT" (Riley et al. 2007), "AGMIAL" (Bryson et al. 2006), or alike (see Bryson et al. 2006 for a recent overview), which are full-blown (automatic) genome annotation and/or information systems and provide exhaustive data repositories to the community. The focus of the HaloLex system is rather to assist experts in achieving an extraordinarily high data quality for a selection of (model) organisms and to make that data available for further analysis like systems modeling, experiment design, etc. To this end, we place particular emphasis on integrating standard genomic data with proteomic (see also Pleissner et al. 2004), transcriptomic, and metabolomic data. This has, for example, enabled a number of genome-scale proteomics (Tebbe et al. 2005; Klein et al. 2005; Bisle et al. 2006; Falb et al. 2006; Aivaliotis et al. 2007; Konstantinidis et al. 2007) and transcriptomic (Twellmeyer et al. 2007) analyses as well as a whole genome metabolic flux simulation (Gonzalez et al. 2008).

So far, scientific applications with HaloLex have mainly been focussing on a number of halophilic archaea, in particular on *Halobacterium salinarum* strain *R1* (*DSM 671*, Pfeiffer et al. 2008), *Natronomonas pharaonis* strain *Gabara* (*DSM 2160*, Falb et al. 2005), and *Haloquadratum walsbyi* strain *HBSQ001* (*DSM 16790*, Bolhuis et al. 2006). These genomes, together with *Halobacterium salinarum* strain NRC-1 (Ng et al. 2000), *Haloarcula marismortui* (Baliga et al. 2004) and *Haloferax volcanii* (J. Eisen, unpublished), are of primary interest to our own group and our collaborators. For specific examples, the reader is referred to the articles of Teufel et al. (2008), Scheuch et al. (2008), Dambeck and Soppa (2008) (all three references in this issue of *Archives of Microbiology*), and to the general review on the genomics and functional genomics of halophilic archaea by Soppa et al. (2008) (this issue of *Archives of Microbiology*).

However, HaloLex is not limited to halophilic archaea but currently covers all publicly available archaeal and a few selected bacterial genomes as obtained from the NCBI (ftp.ncbi.nih.gov/genomes/Bacteria). Besides demonstrating that the system is in principle capable of handling data for a large number of organisms, it allows users to browse the underlying GenBank data (augmented with additional information like bioinformatic predictions) within the same, coherent web interface. The integrated access within the same data model and software environment has shown to be a key prerequisite for conducting comprehensive statistical analyses like the ones presented in the second part of this paper.

The paper is organized as follows: in "Section 1: overview of the HaloLex system", we describe the main functionalities of the HaloLex system and give some notes on its implementation. "Section 2: integrated data analysis with HaloLex" highlights a number of biological problems that have been addressed with HaloLex, and points out bioinformatic solutions for the specific challenges posed by the GC-rich genomes of halophilic archaea. In particular, we shall present new, significantly improved annotation data for *Haloarcula marismortui*.

Section 1: overview of the HaloLex system

In short, HaloLex is based on a relational database serving as the central repository for all kinds of data, which are available for a given microorganism. A dynamic web application provides integrated access to the data and supports the daily work with the genomic and proteomic information in an economical way. The web interface is complemented by a programming interface, which enables (computationally experienced) local users to perform complex data mining tasks, based on a coherent data model and query methods.

Main functionalities of the web application

The primary and most accessible interface to the data stored in HaloLex is a web application, which allows to conveniently browse and query data over the Internet with a minimum technical effort (http://www.halolex.mpg.de). Depending on their individual role, anonymous or appropriately authorized users get read-only access to various browsing and search functionalities or are equipped with additional privileges for data curation and management, respectively. Access rights can be granted separately for each individual strain allowing us to handle all data within the same data store and code base.

Wherever applicable, graphics are rendered in the SVG (Scalable Vector Graphics) format. This greatly facilitates

postprocessing of results and improves the quality of their presentation as compared to working with pixel-based formats (GIF, JPEG, PNG, etc.), which are conventionally employed by the majority of existing web applications.

Genome viewer

The available information about an individual coding sequence is summarized by a central "details page" listing sequences (coding region and protein translation), functional information (e.g., protein name, gene name, EC number, functional classification), general gene and protein characteristics (e.g., sequence length, start and stop codons, GC content, theoretical pI value), and results from several bioinformatic tools, e.g., transmembrane and signal peptide prediction with "Phobius" (Kall et al. 2004), protein export signals with "Tatfind" (Rose et al. 2002), codon adaptation index (Sharp and Li 1987), etc. In addition, the details page shows homologous sequences as well as cross-references to entries of the same protein in major public sequence databases like GenBank, UniProt, Kegg, and also links to relevant PubMed abstracts.

Usually, the details page is reached by selecting an organism and directly specifying an identifier or name for the gene of interest. In addition, also less specific searches and browsing functionalities are supported, including the option to obtain complete lists of genes or proteins, which can optionally be filtered by various characteristics like pI value range, type of proteomic identification, etc. (cf. Fig. 1).

If the organism or gene of interest is not specified a priori, the user can alternatively start out with a blast-based search (Altschul et al. 1997) for all sequences in the HaloLex database, which are similar to a given query.

To reach the details page, one may also start from a graphical display of a particular region on the genome. The corresponding "region viewer" page provides standard genome browsing functionalities and allows to color-code genes according to a variety of characteristics like the annotation status, assigned function class, GC content, proteomic identification (see Fig. 2), and many more.

Genome curation

For the manual curation of genome-based data, the web interface provides basic forms for updating the protein function annotation of individual genes (i.e., changing protein name, gene name, EC number, etc.).

In addition, the gene assignment itself can be revised. HaloLex supports the introduction of newly identified genes, which, e.g., may have been missed by some automatic gene prediction tool. Such tools may also have produced false positives, i.e., open reading frames (ORFs) that are eventually found not to code for proteins. Such "spurious ORFs", which are especially frequent in GC-rich genomes (cf. "Section 2: integrated data analysis with HaloLex"), are not eliminated from the database but get appropriately tagged. This allows to optionally retain such ORFs in viewing and data mining tools (cf. Fig. 2).

Furthermore, start codons may have been misassigned, which is also a common problem for GC-rich genomes (cf. "Section 2: integrated data analysis with HaloLex"). HaloLex assists the curator in assessing and revising the setting of the start codon by showing a number of characteristic quantities like the resulting amino-acid distribution or pI values corresponding to all relevant alternative choices of the start codon.

Viewers for proteomic data

Figure 3 illustrates a navigation path from a spot on a two-dimensional gel image via the spectrum taken in a mass-spectrometric experiment to the identified protein. Individual spots on the gel image, for which spectra have been taken, are classified and color-coded according to the type and quality of the protein identification. The corresponding mass-intensity spectra are annotated and rendered such that the interpretation of the "raw" spectrum immediately gets transparent for the user.

Data mining capabilities

Naturally, not all conceivable types of data analysis can be anticipated and implemented in a web application with limited effort. For example, we opted not to provide sophisticated web-based cross-genome comparison functionalities. To still support complex and highly customizable data mining applications, HaloLex offers full programmatic access to all data and tools within a well-structured data model. Being able to work in such a coherent environment has proven to be a fundamental prerequisite for a large variety of research projects, which have been conducted with HaloLex in the course of several years, a few current examples of which shall be highlighted in the subsequent section. The corresponding application programming interface (API) requires analysis programs to be written in the Java language and to run in the same local-area network where the HaloLex server is located. Both restrictions can, however, be relaxed by means of a SOAP-based web service interface, which we are internally already employing successfully (for a nontechnical introduction to web services and their role in biosciences, see Stein 2002).

Integration of other *-omics* data

The HaloLex database allows storing and accessing other *-omics* data in an integrated way and links them with the

Search Proteins By Proteomics Results

Organism tag: Halobacterium_salinarum.R1.public_MPIB ...

Proteomic Status: ☑ Reliable ☐ Normal ☐ Manual ☐ Questionable ☐ Questionablestill ☐ Unlikely

SELECT ALL | DESELECT ALL | SUBMIT

Reliable	Normal	Manual	Questionable	Questionablestill	Unlikely
1739	252	1	1106	3	0
1991					
1992					

Number of hits found: 1739

ID↑	Gene	FC	Protein name	Contig	Position	Rstatus	MS
OE1001F	-	CHY	conserved hypothetical protein	CHR	248-1453F	ok	Reliable
OE1004F	-	TP	ABC-type transport system ATP-binding protein	CHR	1450-2115F	ok	Reliable
OE1005F	-	TP	ABC-type transport system permease protein	CHR	2112-3254F	ok	Reliable
OE1008F	-	GEN	homolog to oligosaccharyl transferase	CHR	3322-5643F	ok	Reliable
OE1013R	glmS	CHM	glutamine--fructose-6-phosphate aminotransferase (isomerizing) (EC 2.6.1.16)	CHR	5646-7451R	ok	Reliable
OE1014R	graD5	CHM	sugar nucleotidyltransferase (EC 2.7.7.-)	CHR	7454-8641R	ok	Reliable
OE1016R	graD2	CHM	sugar nucleotidyltransferase (EC 2.7.7.-)	CHR	8655-9860R	ok	Reliable
OE1019R	-	ISH	IS1341-type transposase (TCE32)	CHR	11478-12734R	ok	Reliable
OE1022R	-	CHY	conserved hypothetical protein	CHR	13462-13812R	ok	Reliable
OE1023R	-	CHY	conserved hypothetical protein	CHR	13809-14201R	ok	Reliable
OE1062F	-	CHY	conserved hypothetical protein	CHR	32104-32358F	ok	Reliable
OE1067R	-	CHY	conserved hypothetical protein	CHR	33613-33918R	ok	Reliable
OE1077R	ugd2	CHM	UDP-glucose 6-dehydrogenase (EC 1.1.1.22)	CHR	40626-41918R	ok	Reliable
OE1078F	graD6	CHM	sugar nucleotidyltransferase (EC 2.7.7.-)	CHR	42006-42734F	ok	Reliable
OE1079F	-	CHY	conserved hypothetical protein	CHR	42895-43674F	ok	Reliable
OE1081R	gth6	GEN	probable glycosyltransferase, type 1	CHR	45260-46471R	ok	Reliable

Fig. 1 Screenshot of the search functionality of HaloLex. Example output of a query for all genes of *Halobacterium salinarum* (R1), which were "reliably" identified by proteomics (indicated in the *rightmost column*). The complete list of 1,992 identifications was truncated for brevity

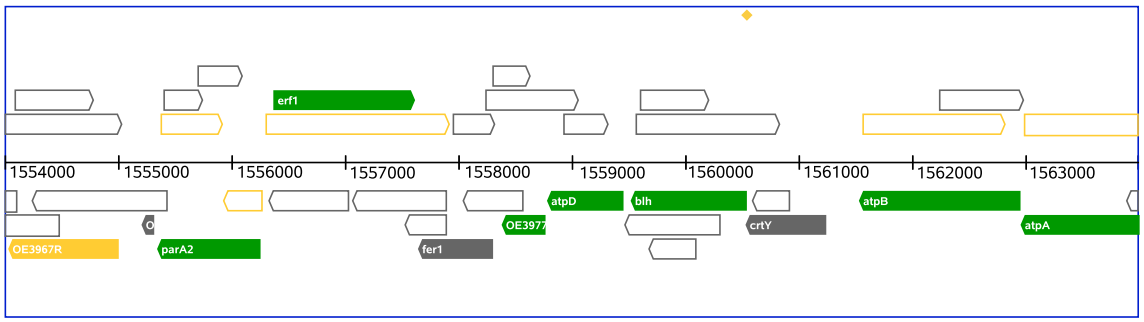

Fig. 2 Screenshot of the region viewer of HaloLex. Genomic region on the *Halobacterium* chromosome with ORFs color-coded according to different trust levels of proteomic identification. "Spurious" ORFs (which are hidden by default) are rendered as *open symbols*

corresponding genomic data. As shown above, this is well established for proteomic data (currently limited to database searches using MASCOT) and also applies to transcriptomic (Twellmeyer et al. 2007), as well as to curated metabolic data based on KEGG information (Falb et al. 2008). Access to the latter, however, is currently restricted to internal data mining applications, i.e., transcriptomic and metabolic data have not yet been made publicly available via the HaloLex web interface.

Notes on the implementation

HaloLex was originally implemented as a classic *"LAMP"* system, i.e., it has been operated on a *L*inux platform, using an *A*pache webserver, the *M*ysql relational database management system, and employing the *P*erl programming language. The system has been mainly used for department-internal purposes and covered only a few genomes.

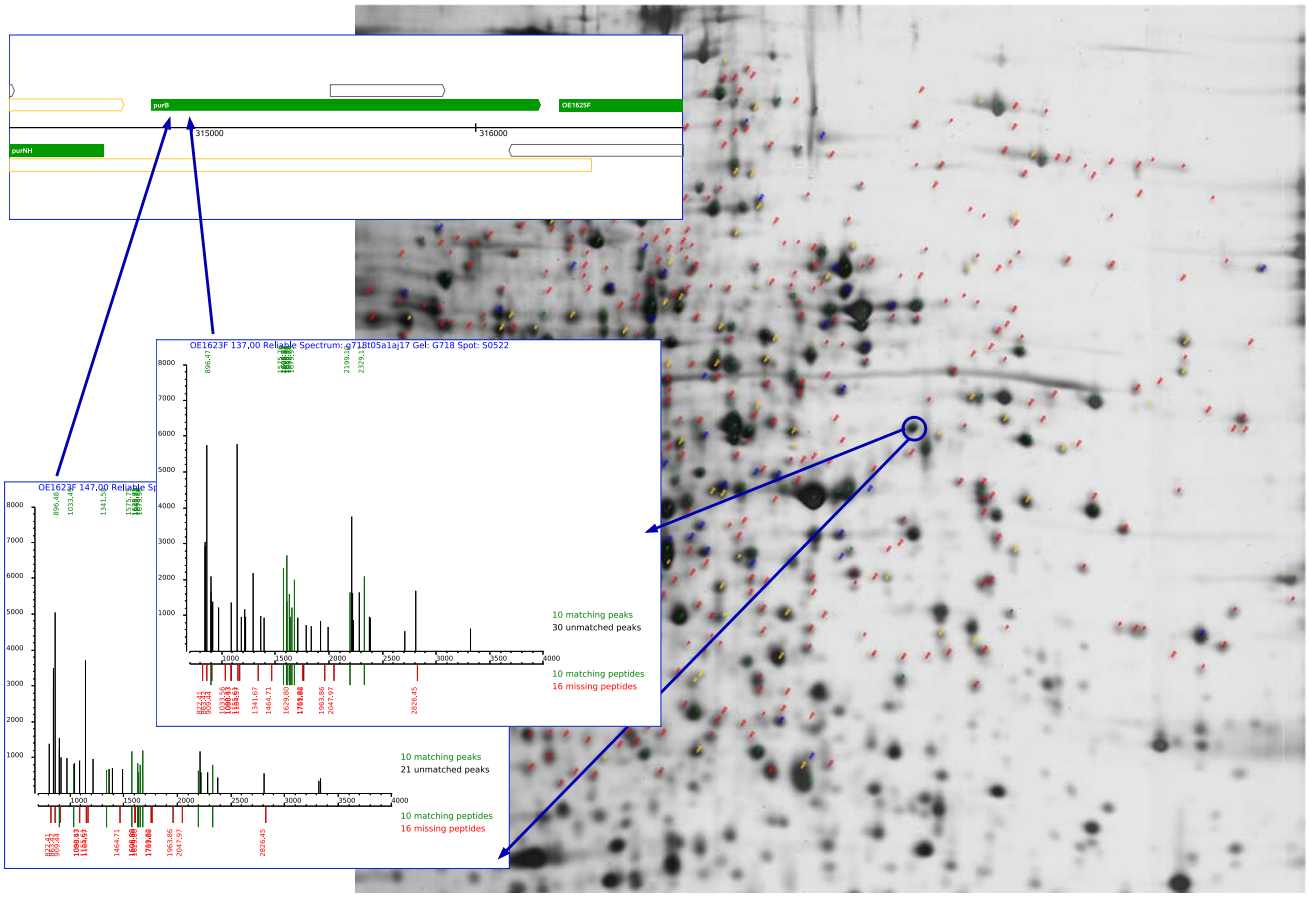

Fig. 3 Integrated access to genomic and proteomic data. Montage of different views of the HaloLex web interface on proteomic data. *Blue arrows* indicate example navigation tracks from a particular spot on a 2D gel image via two different mass-spectra to the identified protein, and its location on the genome, respectively

To substantially extend the system with respect to the amount and complexity of data and to provide user-friendly public web access to the wealth of internal HaloLex functionalities, the system was recently reimplemented based on the Java Enterprise Edition 5 platform (see, e.g., Stearns et al. 2006). Besides many other well-established benefits delivered by this technology, we take advantage of the so-called "distributed components" approach, which promotes (loose) coupling of different stand-alone services through standardized interfaces. Specifically, HaloLex uses remote services offered by the MIGenAS sequence analysis platform (Rampp et al. 2006), e.g., for computing bioinformatic predictions like the transmembrane topology and for cross-referencing database identifiers (cf. Wu et al. 2004). For genome sequences imported from GenBank, we employ the SIMAP web service (Rattei et al. 2006) to retrieve precalculated and regularly updated similarities of proteins with public sequence databases like UniProt, PDB, etc. Data mining applications are enabled by an Application Programming Interface (API), which is built upon the "Remote Interface" component of Java's Enterprise Edition. The same technology is easily exploited to export a web service interface.

Section 2: integrated data analysis with HaloLex

A typical gene prediction problem, ORF overprediction, was chosen as a principal topic to illustrate several applications of HaloLex. We describe the statistical basis for this problem and how an integrated analysis of proteomic and genomic data allows to overcome it. In addition, we describe homology-based methods to detect and resolve gene prediction problems. Using the manual curation tools of HaloLex, we were able to substantially improve the gene prediction for the published genome of *Haloarcula marismortui* (Baliga et al. 2004).

GC-rich genomes like those of halophilic archaea are known to challenge standard gene prediction tools (Nielsen and Krogh 2005; McHardy et al. 2004). Two types of problems are encountered: (1) the existence of alternative long open reading frames (Veloso et al. 2005) makes it

difficult to discriminate protein-coding genes from spurious ORFs; (2) start codon selection is highly error-prone due to long N-terminal ORF extensions in front of the start codon used in vivo (Aivaliotis et al. 2007). In both cases, which we summarize as the "ORF overprediction problem", noncoding DNA may be erroneously "translated" into protein sequences upon unwary application of gene predictors. This markedly deteriorates the quality of the resulting protein-coding gene set. A high-quality gene set is, however, essential for genetic experiments, analysis of transcription and translation signals, or the analysis of protein export signals, which are commonly located in the N-terminal region, not to speak of systems biology applications such as metabolic modeling.

ORF overprediction is illustrated in Fig. 2, which shows a 10 kb region of the *Halobacterium salinarum* strain R1 genome. Protein-coding genes are outnumbered by spurious ORFs, which are all longer than 100 codons. In many cases, a spurious ORF is even longer than the protein-coding gene with which it overlaps. Spurious ORFs with a length of up to 1,300 codons have been found in the *Halobacterium* genome (Pfeiffer et al. 2008).

The ORF overprediction problem is also strikingly illustrated by the fact that 20% of the predicted protein sequences of strain NRC-1 of *Halobacterium salinarum* are inconsistent with those of strain R1, although the DNA sequences of both strains are virtually identical (four single-base differences, five one-base frameshifts, three indels; see Pfeiffer et al. 2008). Among the genes with a start codon assignment discrepancy is the TATA-binding protein tbpA (Scheuch et al. 2008).

Genome statistical data

ORF overprediction is caused by the low number of stop codons in GC-rich genomes (Veloso et al. 2005). Because of the reduced frequencies of T and A, there is a low expectation value for each of the three stop codons (TAA, TAG, and TGA). The problem is further aggravated, because the number of stop codons actually found in prokaryotic genomes is even lower than that predicted by basic statistics of single-nucleotide frequencies. In case of *Halobacterium*, only 66% of the statistically expected stop codons are found. It is interesting to note that nearly all prokaryotic genomes have less stop codons than expected (Fig. 4). While the reduction is moderate for AT-rich genomes, it is significant for genomes with a GC content larger than 60%, where 28% of the expected stop codons are missing on average.

This observation can be explained by an additional bias at the dinucleotide level, which exists on top of the aforementioned bias due to an altered GC content. For the *Halobacterium* chromosome, as an example, this is

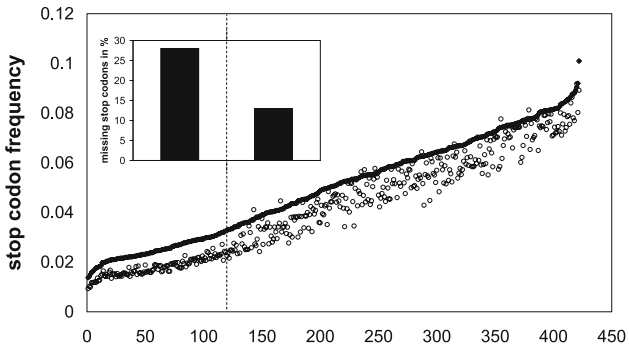

Fig. 4 Expected and actual frequency of stop codons for 425 microbial genomes. For the chromosomes of 425 microbial strains, the expected and the actual number of stop codons was counted and normalized by the total number of codons. Species are sorted along the abscissa by decreasing GC content. For nearly all genomes, the number of actually present stop codons (*open circles*) is significantly lower than that expected (*filled symbols*). The *small inset* shows that for the group of genomes with a GC content >60% (to the *left of the dashed vertical line*), only 72% of the expected stop codons are found, whereas more than 85% of the expected stop codons are actually present in the group of genomes with a GC content <60% (to the *right of the dashed vertical line*). The GenBank data for all microbial strains were downloaded from ftp.ncbi.nih.gov/genomes/Bacteria. Only the chromosome (more precisely: the longest replicon) was chosen for each strain and only one representative strain was used for each species

illustrated by Fig. 5a, which shows that dinucleotides with the same number of A or T residues do not occur with equal frequencies. In particular, the "TA" dinucleotide, which appears in two of the three stop codons, is especially rare. Reduced "TA" dinucleotide frequencies have been found in most prokaryotic genomes (Karlin et al. 2002).

In *Halobacterium*, the "CG" dinucleotide is much more frequent than the other dinucleotides consisting only of G and C residues. As already noted by Karlin et al. (2002), an excess of "CG" is rather exceptional for prokaryotic genomes, which commonly are enriched for "GC." Indirectly, this "CG excess" facilitates gene selection and start codon assignment in *Halobacterium* to some extent, as it results in an excess of four trinucleotides, which correspond to arginine codons. Thus, translations of random stretches of DNA (spurious ORFs) are preferentially arginine-rich and thus highly alkaline, while halophilic proteins are known to be rich in aspartic acid and highly acidic: the pI value of 82% of the *halobacterial* proteins is between 3.5 and 5.5 (Tebbe et al. 2005).

Like spurious ORFs, N-terminal gene extensions in front of the correct start codon tend to be highly alkaline, whereas the rest of the N-terminal region of the protein tends to be acidic. In combination, this results in a large pI upshift in front of the correct start codon (see Fig. 6), which can help to assign it properly (Tebbe et al. 2005). In

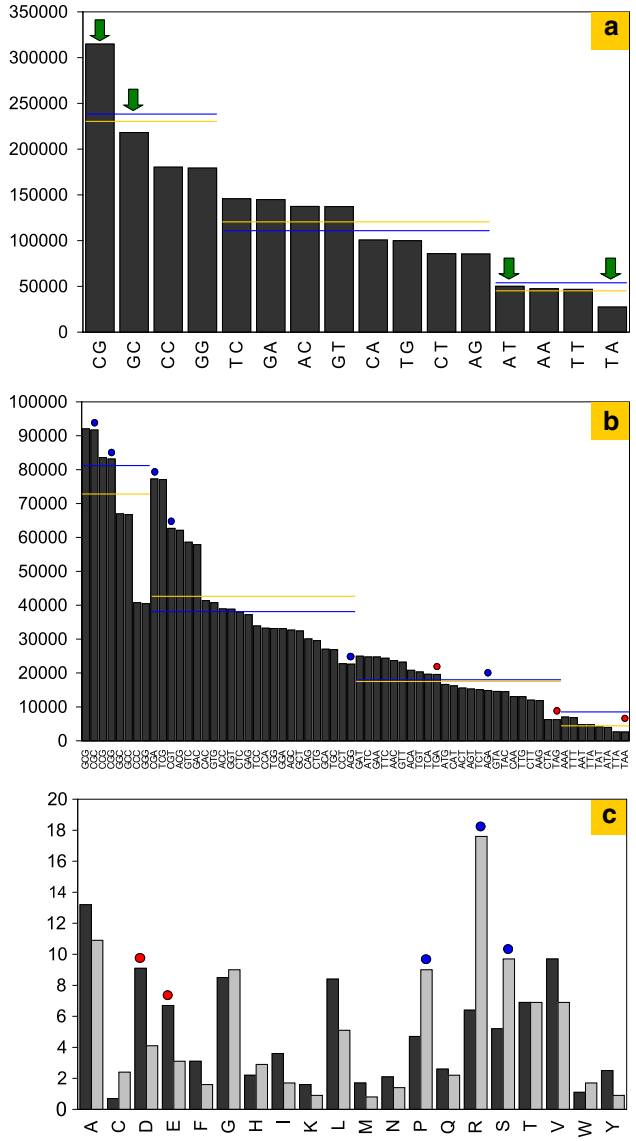

Fig. 5 Dinucleotide bias for *Halobacterium salinarum*. **a** Counts of dinucleotides in the *Halobacterium salinarum* chromosome. Dinucleotides are grouped according to the number of G or C residues. Within each group, each dinucleotide is adjacent to its reverse complement (e.g., TC and GA). The four palindromic dinucleotides are indicated by *green arrows*. For each group, the theoretically expected average (*blue line*) is compared with the average, which is actually observed (*yellow line*). **b** Same as (**a**) but showing the counts of trinucleotides. *Red circles* highlight stop codons and *blue circles* highlight trinucleotides that correspond to arginine codons. **c** The amino acid composition as computed from the protein-coding gene set (*black*) and from trinucleotide counts (*gray*). The over-representation of the acidic amino acids aspartate and to a lesser extent glutamate (*red circles*) in protein-coding genes contrasts with the over-representation of the basic amino acid arginine, prolines and serines (*blue circles*) in translations of random stretches of DNA. This is the basis for a strong pI difference between these two sets of ORFs

the HaloLex web interface, the indicative pI values are shown to assist the annotator in assigning the correct start codon (see "Section 1: overview of the HaloLex system").

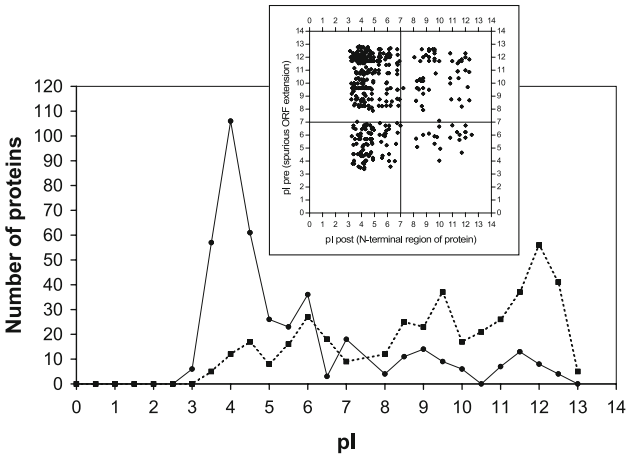

Fig. 6 pI shift around start codons. The distribution of pI values of the 20 N-terminal residues excluding the initial Met (*solid line*) and the 20 residues of the spurious ORF extension (*broken line*) that precedes the start codon is plotted for *Halobacterium*. Transmembrane proteins and proteins with a signal sequence or twin-arginine export motif have been excluded from the analysis. The *small inset* shows the correlation of the pI value of the N-terminal region of the protein (pI-post, plotted on the *x-axis*) and the pI value of the spurious ORF extension (pI_pre, plotted on the *y*-axis). The majority of the N-terminal regions are acidic, while a large fraction of the spurious extensions is highly alkaline

Integrated analysis of proteomic and genomic data

Gene selection and start codon assignment are greatly facilitated by experimental evidences, especially by proteome analysis. We have collected genome-scale proteomic data for *Halobacterium salinarum* (68% of all proteins identified, Tebbe et al. 2005; Klein et al. 2005; Bisle et al. 2006; Falb et al. 2006; Aivaliotis et al. 2007) and for *Natronomonas pharaonis* (43% of all proteins identified, Konstantinidis et al. 2007). This allowed to address and solve the two problems associated with gene prediction in GC-rich genomes, as an ORF is unambiguously confirmed as gene if the protein product is identified by a proteomic experiment. More than 100 orphans (ORFs that potentially code for proteins but do not have any homologs in the databases) could therewith be confirmed as genes. In many cases, initial gene predictions had to be corrected on the basis of proteomic data (see Tebbe et al. 2005).

No evidence for "ORF overprinting" (i.e., more than one gene is located on the same genomic sequence stretch, Keese and Gibbs 1992) was found in *Halobacterium* and *Natronomonas*, although throughout their chromosomes, more than one reading frame is open at a given genome location (cf. Fig. 2). Searching for protein identifications resulting from alternative overlapping reading frames, we did not find a single pair of identified overlapping proteins (see Aivaliotis et al. 2007). Therefore, we conclude that, if

ORF overprinting occurs at all, it is a very rare event (Konstantinidis et al. 2007; Pfeiffer et al. 2008).

To address the problem of start codon assignment, we selected N-terminal peptides from the aforementioned set of proteomic data. In addition, we designed experiments in an attempt to specifically identify N-terminal peptides (Aivaliotis et al. 2007). In total, N-termini from 606 proteins in *H. salinarum* and from 328 in *N. pharaonis* were identified (Falb et al. 2006; Aivaliotis et al. 2007). On the basis of these experimental data, the subsequent integrated analysis of proteomic and genomic data in the HaloLex system confirmed that commonly applied gene finders have a high error rate with respect to start codon selection (Falb et al. 2005; Aivaliotis et al. 2007). Major difficulties to assign correct start codons are also evident from the fact that several hundred start codon assignment discrepancies exist between *Halobacterium salinarum* strains R1 (Pfeiffer et al. 2008) and NRC-1 (Ng et al. 2000), although the DNA sequences are virtually identical. Whenever N-terminal peptides could be identified by proteomics, they confirmed the start codon assignment for strain R1 (Pfeiffer et al. 2008).

A selection of additional results from our proteomic analysis illustrates the power of integrated analysis with the HaloLex system.

Our set of experimentally validated N-terminal peptides is among the largest in the prokaryotic world and allowed to unravel N-terminal protein maturation in halophilic archaea, which consists of methionine cleavage and N-terminal protein acetylation (Falb et al. 2006). N-terminal protein maturation critically depends on the penultimate residues (the one following the initiator-methionine). The set of proteins with N-terminally identified peptide contains 90 integral membrane proteins (again being one of the largest sets currently available). The data show that a major fraction of the integral membrane proteome is synthesized without a cleavable signal sequence and processed analogous to cytosolic proteins (Falb et al. 2006).

One focus of our group is on membrane proteins, which we have extensively analyzed by proteomics (Klein et al. 2005, Bisle et al. 2006). While identification of integral membrane proteins has become highly efficient, our data show that the identification of peptides that form the transmembrane domain is still in its infancy. Most of the integral membrane proteins are identified exclusively through loop peptides. Statistical analysis shows that this hampers protein modification-based quantitative proteomics of integral membrane proteins (Bisle et al. 2006).

Yet another issue concerning gene selection could be solved by experimental means. We were uncertain if our protein-coding gene set would show a major overprediction of small genes. Indicative of such an overprediction were two statistical results: (1) proteins smaller than 20 kDa are severely underrepresented in the set of proteomically identified proteins (Tebbe et al. 2005; Klein et al. 2007); (2) although we had used gel systems that are able to separate proteins below 20 kDa, the number of 2D gel spots in this size range seems much smaller than expected from a theoretical 2D gel (Tebbe et al. 2005). Experimental analysis showed that the small proteins indeed exist, but have so far been missed due to technical problems in standard biochemical experiments. There is a severe washout of small proteins upon standard SDS gel handling procedures (Klein et al. 2007). Also, the low number of peptides upon tryptic digestion severely hampers proteomic identification. With improved experimental techniques, 380 proteins smaller than 20 kDa could be identified (which increased the fraction of identified small proteins by a factor of six).

Homology-based checking of ORF prediction

Small protein-coding genes easily escape upon gene prediction. Therefore, we implemented a semiautomatic homology-based procedure to detect yet unannotated small genes. To this end, short protein sequences from closely related organisms are used for independent homology searches using blastP (protein vs. protein) and tblastN (protein vs. six-frame translation of the genome). Proteins with a higher score in tblastN as compared to blastP are selected for subsequent manual curation. Annotations of new genes, which are detected by this procedure, can be generated using a six-frame translator implemented in HaloLex. We applied this procedure to the published genome of *Haloarcula marismortui*, using proteins with up to 150 residues from *H. salinarum* strain R1, *N. pharaonis*, and *H. walsbyi* as a seed. This enabled us to detect 47 previously missed genes in *Haloarcula* (Table 1); among them, four were ribosomal proteins and 10 were small CPxCG-related zinc finger proteins, which are a prominent class of potential gene regulators found in all archaeal genomes (Tarasov et al. 2008).

In a similar way, sequence homology analysis allows to identify such genes, whose start codons were very likely incorrectly assigned. For this purpose, we analyze the results of a blastP search in closely related organisms. For each organism, the best homolog is used (provided the *e*-value is better than 1E-20). The alignment start position for query and hit is used to categorize the alignment. Alignments are considered to indicate a start codon misassignment if (1) the alignment starts very close to the N-terminus for one sequence but far away for the other and (2) when the alignment starts at the initiator-methionine for one sequence and this methionine aligns with a potential start codon translation (Met or Val) in the other sequence. Candidates are further analyzed by manual inspection. Table 2 lists 337 genes from the published genome of *Haloarcula marismortui*, where we have reassigned the

Table 1 Newly assigned genes in *Haloarcula marismortui*

ORF	Length (aa)	Best homolog	Function	Seq id. (%)	Other homologs
rrnAC0103_A	75	NP3662A	rib_prot S28.eR	86	OE2664F, HQ2884A
rrnAC0208_A	60	NP0350A	CHY	48	–
rrnAC0216_A	126	OE2874F	CHY	42	HQ1719A
rrnAC0301_A	80	HQ2541A	Small ZnF	45	–
rrnAC0669_A	150	NP0856A	CHY	66	HQ1219A, OE1540R
rrnAC0678_A	53	NP0788A	Small ZnF	73	HQ1109A, OE1789R, HQ2748A, OE7210R
rrnAC0696_A	99	NP0816A	Small ZnF	47	OE1556F
rrnAC0797_A	57	HQ2892A	rib_prot L37.eR	92	OE3141R, NP4310A
rrnAC0991_A	48	NP2998A	CHY	79	OE3047F
rrnAC1044_A	86	HQ1848A	moaD family protein	33	NP2500A, NP5020A, NP3946A, OE3595R
rrnAC1515_A	66	NP1736A	Small ZnF	58	HQ3220A, OE3365R
rrnAC1588_A	146	OE5063R	IS200-type transposase	72	NP4630A, OE1439F, rrAC0815, OE4728F
rrnAC1597_A	61	NP4882A	rib_prot S14	72	OE3408F, HQ2828A, NP1768A
rrnAC1603_A	94	NP4870A	RNAseP comp. 1	52	OE3398F, HQ2834A
rrnAC1676_A	44	NP4282A	CHY	77	NP2940A
rrnAC1676_B	122	HQ1297A	CHY	60	–
rrnAC1678_A	100	HQ1827A	CHY	59	–
rrnAC1706_A	141	NP1764A	CHY	52	–
rrnAC1831_A	63	NP1510A	CHY	63	HQ1704A, OE1775R
rrnAC1867_A	52	HQ1176A	Small ZnF	69	OE1435R, NP5316A
rrnAC1929_A	212	OE3249F	Cob cluster protein	54	NP5310A, HQ1412A, NP1896A
rrnAC1936_A	89	NP3612A	CHY	48	–
rrnAC1983_A	231	NP1896A	CHY	43	–
rrnAC2105_A	134	NP0772A	CHY	54	HQ1375A, OE4661R
rrnAC2167_A	96	NP4084A	CHY	35	HQ2323A
rrnAC2212_A	142	HQ1071A	CHY	38	OE2090R
rrnAC2268_A	116	NP2558A	Transcription regulator	74	OE2591R, NP3596A, rrnAC3399, HQ1949A
rrnAC2270_A	59	HQ1365A	Small ZnF	68	NP0928A, OE4676F
rrnAC2286_A	84	NP5336A	CHY	43	–
rrnAC2448_A	54	NP5086A	CHY	51	–
rrnAC2530_A	130	HQ2261A	CHY	52	–
rrnAC2569_A	49	HQ3677A	Small ZnF	74	NP0778A, OE4167C1R, HQ2748A, OE7210R
rrnAC2574_A	52	NP0788A	Small ZnF	92	OE1789R, HQ1109A, HQ2748A
rrnAC2592_A	98	HQ1034A	CHY	79	NP1820A
rrnAC2764_A	111	NP5102A	CHY	59	HQ3659A, OE4054F
rrnAC2791_A	115	NP0196A	CHY	70	HQ3647A, OE3914R
rrnAC2834_A	129	OE4148F	CHY	31	HQ3411A
rrnAC2897_A	73	OE4475R	Small ZnF	58	HQ3437A, NP0708A
rrnAC2982_A	137	HQ2813A	CHY	42	HQ1789A, HQ2547A, pNG7092, NP1808A
rrnAC3115_A	57	NP0186A	rib_prot HL32	75	HQ3421A
rrnB0024_A	139	OE6004F	Small ZnF	64	NP6252A, HQ1149A
rrnB0146_A	118	OE1549F	CHY	54	NP1698A, HQ1429A, OE3894R
rrnB0177_A	89	HQ2065A	CHY	50	–
pNG3034_A	47	NP4282A	CHY	80	NP2940A
pNG6117_A	85	OE6052R	CHY	59	–
pNG6164_A	115	OE6242R	CHY	79	–
pNG6170_A	53	NP0788A	CHY	75	OE1789R, HQ1109A, HQ2748A, OE7210R

Using tblastN, previously unannotated genes were detected and realized by the manual curation options within HaloLex. For each newly assigned gene, its code, length, the best homolog (with a brief function indication and percentage of sequence identity) and other homologous genes are given. Codes are systematically assigned using the number of the upstream ORF and a letter attached with an intervening underscore (commonly _A). Function assignment abbreviations: *CHY* conserved hypothetical protein, *rib_prot* ribosomal protein, *small ZnF* small CPxCG-related zinc finger protein (Tarasov et al. 2008)

```
VNG2591C   mrlvqvtvptgkrdavlaalddegvdyvvtpetasteytavvhfplptaavsdvldalqdvglsqdaytvvvdaetvvsr
OE4634F    MRLVQVTVPTGKRDAVLAALDDEGVDYVVTPETASTEYTAVVHFPLPTAAVSDVLDALQDVGLSQDAYTVVVDAETVVSR
HQ3141A    MRFVQVLVPAGTRDAVVEVITDENIEYAITDETGTDEYEAIITFPLPTAAVEPVLDQLRTVGIDTDATTVVLEAETVVSD
NP0578A    MRLVQVTIPAGKREAVLRVLDEEGIDYVVTDETSGREYTAVAYFPLPTSAVEPILEQLRDVGLEREAYTVVVSAETVVSK
rrnAC2377  mrlvqlliptgkrdavlgvlteegidyvltdetsgreftavvtfpvptnalepvlealrdvginddgytvvvdantviss
           ** **   * * * **      *   *  * **   *      ** ** *    *  *  **      ***  * ** *

VNG2591C   rfdeldarysadsdaadd--hiareelvaraadlapnrvtyavltlvstiiatagllldspatvvgsMVIAPLLGPAMSA
OE4634F    RFDELDARYSADSDAADD--HIAREELVARAADLAPNRVTYAVLTLVSTIIATAGLLLDSPATVVGSMVIAPLLGPAMSA
HQ3141A    KFDELDEAYSNDSDDTTDGDRIARDELLARANELAPGIGPFILMTIVSAIVATAGLLLDSPAVVVGSMVIAPLIGPAMST
NP0578A    RFDDLKDSYAEKEESEE---RIARQEIEARAEELAASIPTYVVMTIVSAVIATAGLLLDSPATVVGSMVIAPLIGPAMTT
rrnAC2377  qfeeveetyaeeeded----riareeltskakdlapslsnyalMTIISAIIATAGLLLDSPAVVVGSMVIAPLIGPAMTA
            *      *            *** *    *  **         *  *   ********** ********** ****

VNG2591C   SVGTVVDDDDLFARGIRLQVVGVALAVVGAAAFAFLVKTTHLVP-PGLDVLSLSEVRERLRPDFLSLVVALGSGVAGVYS
OE4634F    SVGTVVDDDDLFARGIRLQVVGVALAVVGAAAFAFLVKTTHLVP-PGLDVLSLSEVRERLRPDFLSLVVALGSGVAGVYS
HQ3141A    SVGTVVDDTSLVARGVKLQILGGVLAIVSAAGFAFMLRTVQIVPLSAVEVFEIGEVSQRLAPDVLSLVIALGAGAAGAVS
NP0578A    AVGSVIDDAELFQRGVSLQVVGIVLAVAAATVFAVFVQVMNLVP-PGLDPLSLAEVEERLSPNFLSLAVAIGAGIAGAVS
rrnAC2377  NVGTVVDDNEMFARGVKLQAVGLGLAVASATAFALLVRYANVIP-PLADVTAVGQIRERVAPDFLSLIVALGAGAAGVVS
            ** * **     **  **  *  **   *  **         *              *  *  ***  * * *  *  *
```

Fig. 7 Homology-based start codon checking for the detection of ORFs, which are too short. A sequence alignment of four homologous proteins of *H. salinarum* (strains R1 and NRC-1), *N. pharaonis*, *H. walsbyi* and *H. marismortui* is shown. Codes starting with OE are from *H. salinarum* strain R1, those with VNG from strain NRC-1, NP from *N. pharaonis*, HQ from *H. walsbyi* and those starting with rrnAC from *H. marismortui*. *Uppercase letters* indicate the protein sequence as obtained from the current database, the first methionine being *bold*. *Lowercase letters* indicate additional residues obtained by our correction of the start codon assignment. Residues conserved in all sequences are indicated by *asterisks*

```
VNG2422C                        MAYSGPPKHHAQPSGPHHRPQRSHAPTHYATP
OE4429F                         maysgppkhhaqpsgphhrpqrshapthyatp
HQ3640A                                                         -
NP0462A    mtpalsipylcrshqkgldsaglgpncqpldvlnrrfaadavlvgdglrgr
rrnAC2722                                      MTQNASGFIPGSTTGS


VNG2422C  [33]MIAVLADTHSDTDHALTGHARQAVADADAVVHAGDFTTESSLDAFHDAATRLHAVHGNADSPAVRDRLPP
OE4429F    [1]MIAVLADTHSDTDHALTGHARQAVADADAVVHAGDFTTESSLDAFHDAATRLHAVHGNADSPAVRDRLPP
HQ3640A    [1]MLTVISDTHSTDNHQLSGQTYEAVQNAEMVAHAGDFMCESVLDALQREATQLVGVAGNNDDTGIRERLPT
NP0462A    [1]MLAVLSDTHGRDSPRLSGRTADAVAEASRVVHAGDFMTEAVLDAFEERGP-LAAVYGNNATAAVRERLPA
rrnAC2722 [17]MLTAISDTHGTDNHRLTGRTLDAVREADHVLHAGDFMTEQVLDAIDAESDELTGVVGNNDRPAVRARLSD
           *     ***      * *    **     * *****  *  ***       *  * **      * **
```

Fig. 8 Homology-based start codon checking for the detection of ORFs, which are too long. A sequence alignment of four homologous proteins of *H. salinarum* (strains R1 and NRC-1), *N. pharaonis*, *H. walsbyi* and *H. marismortui* is shown. Codes starting with OE are from *H. salinarum* strain R1, those with VNG from strain NRC-1, NP from *N. pharaonis*, HQ from *H. walsbyi* and those starting with rrnAC from *H. marismortui*. The protein sequences are highly homologous. Residues conserved in all sequences are indicated by *asterisks* (*lower alignment block*). Spurious N-terminal sequence extensions are possible in three of the four species, but are considered to be incorrect as they are not homologous to each other (*upper alignment block*). *Uppercase letters* indicate the protein sequence as obtained from the current database, the first methionine being *bold*. The position of the probable initiator methionine in the current database sequence is indicated. *Lowercase letters* indicate gene extensions, which are possible but are considered spurious

start codon (196 genes are shortened and 141 extended). We briefly discuss two example cases by showing the corresponding multiple sequence alignments (Figs. 7, 8).

Figure 7 shows a gene, which needs to be extended in *Haloarcula marismortui* (and also in *H. salinarum* strain NRC-1). Met-1 of rrnAC2377 aligns with Met-121 of NP0578A. Using the longer sequence (here NP0578) for tblastN shows that the homologous region extends beyond the assigned start codon (lowercase sequence letters for rrnAC2377). VNG2591C can also be extended to match OE4634F, as the genome sequences of strains R1 and NRC-1 are identical in this region (also indicated by lowercase sequence letters).

Figure 8 shows an example of a gene, which needs to be shortened in *Haloarcula marismortui* (and also in *H. salinarum* strain NRC-1). The methionine at position 17 in the rrnAC2722 sequence aligns with the methionine at position 1 of NP0462A. Using the longer sequence (here rrnAC2722) for tblastN does not result in an extension of the homologous region as compared to the shorter sequence (NP0462A), which indicates that the extension may be spurious. Spurious ORF extensions are possible in three of the four halophiles, but they are not homologous to each other.

It should be stressed that the homology-based procedures described above are not suitable for performing automatic, unsupervised gene predictions. They rather serve to preselect candidates with probable gene prediction errors, which then need to be manually inspected. The HaloLex system is well suited to support such manual

Table 2 Genes with corrected start codon assignments in *Haloarcula marismortui*

ORF	Corrected length (aa)	Original length (aa)	Direction of change	Homologous ORFs
rrnAC0004	1,551	1,356	Extended	NP4364A, OE3175F, HQ3018A
rrnAC0005	624	360	Extended	OE2052F, NP3952A
rrnAC0012	957	1,083	Shortened	HQ1987A, NP4816A
rrnAC0041	279	339	Shortened	NP3812A, HQ1503A, OE2950R
rrnAC0053	864	717	Extended	NP2042A, HQ3014A
rrnAC0080	1,347	1,383	Shortened	NP3698A, OE2648F, HQ2890A
rrnAC0083	936	654	Extended	NP1382A, HQ2430A
rrnAC0101	2,238	2,328	Shortened	OE2656R, NP3690A
rrnAC0115	774	909	Shortened	NP4142A, OE2472F
rrnAC0137	882	993	Shortened	OE2860R, NP3116A, rrnAC1236
rrnAC0145	1,125	843	Extended	OE4359F, HQ1341A
rrnAC0171	984	1,080	Shortened	OE1599F
rrnAC0178	1,812	1,725	Extended	OE1613R
rrnAC0181	1,866	1,413	Extended	NP4200A, OE3010F, HQ2528A
rrnAC0198	1,236	1,305	Shortened	NP1646A
rrnAC0199	480	537	Shortened	NP1296A
rrnAC0213	951	726	Extended	HQ1261A
rrnAC0215	723	522	Extended	NP0356A, HQ2398A, OE1636F
rrnAC0239	999	1,110	Shortened	NP1902A, OE2918F
rrnAC0240	333	372	Shortened	OE3652F, HQ2230A
rrnAC0249	879	1,017	Shortened	OE3606R, NP3184A
rrnAC0261	267	300	Shortened	HQ2898A, NP3686A, OE3683R
rrnAC0280	333	381	Shortened	NP2580A, HQ2722A
rrnAC0284	447	225	Extended	NP2596A, HQ2724A
rrnAC0304	1,104	1,155	Shortened	OE2360R, NP5226A
rrnAC0305	957	1,011	Shortened	OE2551F, NP3722A
rrnAC0322	1,176	1,203	Shortened	NP3076A, OE2763F
rrnAC0324	411	153	Extended	NP3362A, OE4451F
rrnAC0329	1,245	1,035	Extended	NP2206A, HQ2700A
rrnAC0374	414	324	Extended	NP2642A, OE2237F, HQ2615A
rrnAC0394	1,008	1,056	Shortened	NP1082A, OE4339R, HQ3696A
rrnAC0426	1,386	1,464	Shortened	pNG7203, NP0964A, HQ3464A
rrnAC0430	936	975	Shortened	HQ2692A, NP1916A, OE2547R
rrnAC0436	1,530	1,605	Shortened	OE2288F, NP2702A, HQ2668A
rrnAC0481	1,002	894	Extended	NP2798A
rrnAC0494	249	306	Shortened	NP4192A, OE1860F, HQ1556A
rrnAC0497	708	471	Extended	HQ1562A, OE1794R
rrnAC0505	1,710	1,485	Extended	OE3490R, NP1742A, HQ3347A
rrnAC0506	1,527	1,575	Shortened	NP3956A, OE2049R
rrnAC0536	573	606	Shortened	NP2906A, HQ2751A
rrnAC0546	1,791	1,833	Shortened	HQ1573A, OE1495R, NP1746A
rrnAC0568	699	435	Extended	HQ1615A, rrnAC3127, NP1996A
rrnAC0572	804	855	Shortened	HQ3669A, rrnAC2557, NP0792A, OE3115F
rrnAC0589	657	846	Shortened	rrnAC2321, OE8048F
rrnAC0617	1,335	1,413	Shortened	NP0212A, HQ2634A, OE8010R
rrnAC0619	1,272	963	Extended	HQ3141A, OE4634F, NP0578A
rrnAC0620	1,566	1,431	Extended	NP4066A, HQ2635A, OE7174R
rrnAC0628	912	957	Shortened	NP4072A, HQ2637A, OE1748R
rrnAC0629	738	774	Shortened	NP4074A, HQ2638A, OE1752F

Table 2 continued

ORF	Corrected length (aa)	Original length (aa)	Direction of change	Homologous ORFs
rrnAC0631	1,122	642	Extended	NP0380A, OE1582R, HQ1531A
rrnAC0633	969	1,008	Shortened	NP0384A, OE1578F, HQ1670A
rrnAC0638	474	606	Shortened	HQ3692A, NP2228A
rrnAC0651	1,191	1,059	Extended	OE4393R, NP0888A
rrnAC0655	603	657	Shortened	NP1390A, HQ1666A
rrnAC0660	333	417	Shortened	NP4090A, HQ2743A, OE1651F
rrnAC0663	1,029	240	Extended	NP0372A, OE1646R, HQ2392A
rrnAC0666	708	582	Extended	NP0100A, HQ3478A, OE1004F
rrnAC0674	810	945	Shortened	rrnAC0848, OE7042R, rrnAC2044, HQ2141A, NP6028A
rrnAC0687	1,269	1,347	Shortened	OE4207F
rrnAC0696	762	900	Shortened	NP0818A, OE1554R
rrnAC0717	636	699	Shortened	NP1230A, OE1713F, HQ1537A
rrnAC0721	1,371	1,395	Shortened	OE3467R, HQ1298A
rrnAC0753	879	936	Shortened	OE2785R, HQ2762A
rrnAC0777	903	492	Extended	HQ2440A
rrnAC0779	861	915	Shortened	OE2138F, NP1596A
rrnAC0801	951	1,071	Shortened	NP4302A, OE3145F, HQ2933A
rrnAC0825	1,284	1,383	Shortened	NP4134A, HQ2196A
rrnAC0833	795	858	Shortened	NP1462A, OE1641R, HQ2394A
rrnAC0838	1,779	1,827	Shortened	NP2726A, HQ1873A, OE2653R
rrnAC0841	954	1,062	Shortened	NP2730A, OE2561R, HQ1874A
rrnAC0843	1,524	1,587	Shortened	NP2738A, OE2555R, HQ2402A
rrnAC0852	1,017	966	Extended	OE3343R
rrnAC0875	543	405	Extended	OE3121R, HQ2788A
rrnAC0878	810	783	Extended	OE3119R, NP4334A
rrnAC0883	1,116	504	Extended	NP4236A
rrnAC0896	1,152	1,293	Shortened	HQ1590A, OE2358F, NP3650A
rrnAC0917	1,065	1,140	Shortened	HQ1663A, OE1669F
rrnAC0925	990	1,035	Shortened	NP2796A, OE2451R
rrnAC0934	423	471	Shortened	NP2710A, OE2005F, HQ2301A
rrnAC0942	1,611	1,416	Extended	OE3436R
rrnAC0944	1,302	1,350	Shortened	HQ1663A, OE1669F
rrnAC0956	462	537	Shortened	HQ1497A, OE2934R
rrnAC1042	1,806	1,851	Shortened	rrnAC1570, HQ3533A
rrnAC1083	1,965	2,010	Shortened	NP4322A, OE2871F
rrnAC1106	519	420	Extended	NP4198A, OE2985F, HQ2561A
rrnAC1107	1,476	1,176	Extended	NP4904A, HQ1686A
rrnAC1115	270	324	Shortened	NP4036A, OE2903R, HQ2458A
rrnAC1138	849	507	Extended	OE2020F, NP1592A
rrnAC1169	1,311	867	Extended	NP3742A, OE2827R, HQ2339A
rrnAC1218	1,794	1,821	Shortened	HQ1754A
rrnAC1220	663	606	Extended	HQ1752A
rrnAC1261	1,155	1,182	Shortened	NP4050A, HQ2389A
rrnAC1263	798	894	Shortened	OE2913R, NP3970A
rrnAC1281	2,529	2,592	Shortened	OE2573F, NP1526A
rrnAC1299	1,881	1,704	Extended	OE1143R, HQ3344A, NP1442A
rrnAC1308	339	399	Shortened	NP2066A, HQ1665A, OE1673F
rrnAC1336	399	426	Shortened	NP4972A
rrnAC1341	1,683	1,800	Shortened	NP0164A

Table 2 continued

ORF	Corrected length (aa)	Original length (aa)	Direction of change	Homologous ORFs
rrnAC1350	1,050	1,191	Shortened	NP3216A, OE1906R, HQ2500A
rrnAC1361	687	498	Extended	OE2276F, NP2980A, HQ1692A
rrnAC1365	1,026	1,161	Shortened	rrnAC0576
rrnAC1377	582	777	Shortened	rrnAC0508, NP3954A
rrnAC1383	210	342	Shortened	NP1548A
rrnAC1395	849	1,029	Shortened	NP4160A
rrnAC1438	429	351	Extended	NP3220A, OE2139R, HQ1579A
rrnAC1443	1,275	1,398	Shortened	NP3228A, HQ1584A, OE2149R
rrnAC1444	1,623	1,746	Shortened	HQ1934A, HQ2096A, pNG7256
rrnAC1447	414	534	Shortened	NP2292A, HQ1637A, OE1953F
rrnAC1454	303	207	Extended	OE1963F, HQ1645A, NP2308A
rrnAC1477	1,047	1,251	Shortened	OE2014F, HQ2353A
rrnAC1497	1,431	1,707	Shortened	NP4594A, OE3274R
rrnAC1500	1,092	1,155	Shortened	OE3278R, NP4774A
rrnAC1504	846	528	Extended	NP4780A, HQ2866A, OE3286F
rrnAC1516	189	489	Shortened	OE3330F
rrnAC1530	765	459	Extended	NP1786A, HQ3174A
rrnAC1532	711	579	Extended	NP1788A, OE3352R, HQ3173A
rrnAC1536	1,332	1,395	Shortened	HQ1685A, NP4902A, OE5298F
rrnAC1542	1,416	1,644	Shortened	OE1133F
rrnAC1567	291	474	Shortened	HQ2131A
rrnAC1588	1,308	882	Extended	OE5062R
rrnAC1621	570	459	Extended	HQ2801A, OE3367F
rrnAC1626	2,532	2,808	Shortened	rrnAC2044, rrnAC0848, OE7042R, HQ2141A
rrnAC1628	366	216	Extended	OE3324R, HQ2783A, NP3352A
rrnAC1630	591	411	Extended	OE2334R, NP0028A
rrnAC1638	975	1,107	Shortened	NP1214A
rrnAC1647	402	222	Extended	NP1834A
rrnAC1655	750	675	Extended	NP1082A, OE4339R, HQ3696A
rrnAC1665	642	735	Shortened	NP2884A
rrnAC1669	279	351	Shortened	OE1371R, HQ1283A, NP5232A
rrnAC1680	390	429	Shortened	NP0612A, OE1371R, HQ1286A
rrnAC1690	1,545	1,509	Extended	NP0624A, HQ1292A
rrnAC1702	1,719	1,218	Extended	NP1742A, OE3490R, HQ3347A
rrnAC1708	1,347	1,089	Extended	NP4502A, HQ3336A, OE3496R
rrnAC1718	1,359	1,065	Extended	NP4542A, OE3506F, HQ3330A
rrnAC1726	1,527	1,485	Extended	HQ3326A, OE3511F, NP4534A
rrnAC1743	549	486	Extended	HQ1673A, NP5358A, rrnAC3526
rrnAC1764	924	978	Shortened	rrnAC1777, NP5168A, OE1385F, HQ1277A
rrnAC1774	339	411	Shortened	HQ1279A, OE1379R
rrnAC1776	588	633	Shortened	NP5166A, OE1384F, HQ1278A
rrnAC1779	651	516	Extended	NP5170A
rrnAC1782	933	963	Shortened	HQ1276A, NP5174A, OE4651F
rrnAC1797	858	903	Shortened	NP4932A, OE3445F, rrnAC0317
rrnAC1809	735	444	Extended	OE1445R, NP1134A
rrnAC1812	705	525	Extended	OE1451F, HQ1168A, NP1178A
rrnAC1822	717	666	Extended	NP1636A, OE1793F, HQ1712A
rrnAC1826	522	237	Extended	NP1498A, HQ1709A, OE1785F
rrnAC1840	2,676	2,727	Shortened	NP1516A, OE1770F, HQ1701A

Table 2 continued

ORF	Corrected length (aa)	Original length (aa)	Direction of change	Homologous ORFs
rrnAC1849	1,746	1,083	Extended	HQ1189A, NP5206A
rrnAC1853	3,003	2,616	Extended	NP5214A, HQ1185A
rrnAC1855	891	330	Extended	NP5218A, OE1417F, HQ1183A
rrnAC1867	522	756	Shortened	HQ1177A, OE1434R, NP5318A
rrnAC1870	1,455	1,209	Extended	NP4702A, OE3960F
rrnAC1880	1,239	1,284	Shortened	NP0438A, OE3971R, rrnAC3166
rrnAC1905	429	279	Extended	NP4526A, pNG6069
rrnAC1930	924	366	Extended	OE3253F, NP5308A, HQ1411A
rrnAC1931	804	504	Extended	HQ1410A, NP5306A, OE3255F
rrnAC1950	1,893	1,680	Extended	NP0158A, HQ1329A
rrnAC1957	1,491	1,671	Shortened	HQ2578A
rrnAC1979	795	591	Extended	NP1462A, OE1641R
rrnAC1983	1,218	1,755	Shortened	NP1754A, OE2013R
rrnAC1992	738	591	Extended	NP1470A
rrnAC2014	792	747	Extended	NP5122A, OE1306F, HQ1416A
rrnAC2085	360	522	Shortened	NP0342A, OE4713R, HQ3071A
rrnAC2098	1,878	1,905	Shortened	NP0198A, OE4671R, HQ1369A
rrnAC2105	522	1,014	Shortened	NP0774A, OE4663F, HQ1347A
rrnAC2127	1,005	711	Extended	NP0962A
rrnAC2129	864	894	Shortened	OE4355R, NP3186A
rrnAC2158	1,974	2,034	Shortened	NP0404A, OE4613F, HQ3117A
rrnAC2159	462	609	Shortened	NP0954A, HQ3116A, OE4610R
rrnAC2181	1,098	1,227	Shortened	NP1140A, OE4571R, HQ1074A
rrnAC2221	435	579	Shortened	OE4541F, NP1718A, HQ1065A, rrnAC2455
rrnAC2223	372	387	Shortened	NP1710A, OE4544R, HQ1063A
rrnAC2245	1,137	870	Extended	OE4034R, HQ3066A, NP0030A
rrnAC2247	1,296	1,443	Shortened	NP1050A
rrnAC2258	624	435	Extended	NP0018A
rrnAC2261	954	1,026	Shortened	OE1151R, NP0014A, HQ1359A
rrnAC2278	528	600	Shortened	NP3368A, HQ2565A, rrnAC0868, OE2992R
rrnAC2284	1,038	993	Extended	NP5368A, OE2438R
rrnAC2352	1,167	1,188	Shortened	pNG7026, OE5170F, HQ1989A
rrnAC2356	999	1,251	Shortened	NP5048A, HQ1275A, OE4196R
rrnAC2359	432	180	Extended	NP4806A, rrnAC0738, OE3162F, HQ2346A
rrnAC2377	1,281	924	Extended	NP0578A, OE4634F, HQ3141A
rrnAC2440	684	780	Shortened	NP0956A, OE4360R, HQ3733A
rrnAC2460	1,977	2,127	Shortened	NP1264A, HQ3402A, OE4140R
rrnAC2469	1,299	1,422	Shortened	HQ2809A, HQ2192A
rrnAC2473	1,524	1,569	Shortened	OE2133R, NP3020A
rrnAC2474	288	351	Shortened	NP1258A, OE4136R, HQ3399A
rrnAC2476	936	582	Extended	NP1312A, OE4133R
rrnAC2518	1,266	1,221	Extended	NP1318A, OE3943R, HQ3056A
rrnAC2525	1,026	735	Extended	HQ1021A, OE4201R
rrnAC2526	390	477	Shortened	NP1272A, HQ2001A, OE4217R
rrnAC2529	741	906	Shortened	NP1268A, OE4218F, HQ1025A
rrnAC2532	2,766	3,069	Shortened	NP0538A, OE1272R, HQ1460A
rrnAC2550	2,154	2,241	Shortened	OE1267R, NP0536A, HQ1456A
rrnAC2558	1,863	1,443	Extended	OE3889R, HQ3102A, NP1576A
rrnAC2565	978	1,038	Shortened	HQ3671A, NP0900A, OE4195F

Table 2 continued

ORF	Corrected length (aa)	Original length (aa)	Direction of change	Homologous ORFs
rrnAC2582	699	870	Shortened	NP1406A, HQ1040A, OE4235F
rrnAC2586	1,200	1,065	Extended	HQ3704A, NP1412A, OE4236F
rrnAC2592	975	1,281	Shortened	HQ1035A, NP1818A
rrnAC2627	2,061	2,106	Shortened	NP1344A, HQ2213A
rrnAC2629	804	864	Shortened	NP5160A
rrnAC2630	858	993	Shortened	OE4085R, NP0606A, HQ3650A
rrnAC2633	1,356	888	Extended	rrnAC0404, OE3070R
rrnAC2636	624	720	Shortened	NP5088A, OE3906F
rrnAC2642	738	246	Extended	NP5114A, HQ2624A, OE2740F
rrnAC2656	1,125	1,182	Shortened	HQ2450A, OE2317R
rrnAC2657	1,194	1,062	Extended	OE5132F, rrnB0290, NP1412A
rrnAC2714	1,569	1,179	Extended	NP0482A, HQ1003A, OE4390F
rrnAC2722	510	558	Shortened	NP0462A, HQ3640A, OE4429F
rrnAC2748	435	582	Shortened	NP5152A, HQ3265A, OE4027F
rrnAC2749	453	186	Extended	NP5150A, OE4028R, HQ3266A
rrnAC2753	648	687	Shortened	HQ1473A
rrnAC2754	366	417	Shortened	NP5146A, OE4039F
rrnAC2755	1,206	1,257	Shortened	OE4034R, HQ3066A, NP0030A
rrnAC2756	2,046	1,770	Extended	NP5144A, HQ3065A, OE4041F
rrnAC2761	1,143	960	Extended	OE2170R, HQ2450A
rrnAC2772	1,521	1,617	Shortened	NP1074A, HQ2660A
rrnAC2776	1,488	1,260	Extended	pNG7305, OE2076F, HQ2506A
rrnAC2780	1,356	1,443	Shortened	NP4376A, HQ3643A, OE3922R
rrnAC2781	894	711	Extended	NP0190A, OE3921F, HQ3644A
rrnAC2782	1,236	1,140	Extended	NP0192A
rrnAC2783	1,437	1,296	Extended	NP5292A, OE2063R
rrnAC2798	552	393	Extended	OE3905F, HQ1379A, NP0086A
rrnAC2800	612	699	Shortened	NP0092A, OE3902R, HQ1377A
rrnAC2804	516	447	Extended	NP1700A, OE3895F
rrnAC2806	381	228	Extended	NP1698A, HQ1429A, OE3894R
rrnAC2810	909	990	Shortened	OE3892R, NP1688A, HQ3137A
rrnAC2811	1,905	1,128	Extended	OE3889R, HQ3102A, NP1576A
rrnAC2818	1,344	1,479	Shortened	NP2252A, HQ3104A, OE3882R
rrnAC2822	951	309	Extended	NP2248A, OE3879F, HQ3106A
rrnAC2831	540	411	Extended	NP5076A, HQ1339A, OE3871R
rrnAC2834	1,032	1,677	Shortened	NP1066A, OE4144R, HQ1407A
rrnAC2836	1,206	1,359	Shortened	HQ2439A, NP3538A
rrnAC2851	744	786	Shortened	NP0554A, OE4165R, HQ3686A
rrnAC2857	2,196	2,238	Shortened	NP1350A, OE4181R, HQ3684A
rrnAC2859	630	678	Shortened	NP1332A, HQ3517A
rrnAC2867	1,467	1,353	Extended	OE4370R, NP5292A
rrnAC2870	1,092	951	Extended	HQ3382A, OE4359F
rrnAC2891	1,698	1,773	Shortened	NP1008A, OE4122R, HQ3049A
rrnAC2893	975	873	Extended	NP0698A, HQ3439A, OE2975F
rrnAC2901	1,728	1,278	Extended	NP0898A, OE4471R
rrnAC2933	720	558	Extended	NP0238A, HQ1390A
rrnAC2937	2,844	2,895	Shortened	OE1286R, NP0232A
rrnAC3005	927	999	Shortened	NP1116A, HQ3034A, OE3214F
rrnAC3008	1,158	723	Extended	NP1114A, HQ3033A, OE3216F

Table 2 continued

ORF	Corrected length (aa)	Original length (aa)	Direction of change	Homologous ORFs
rrnAC3046	1,104	1,146	Shortened	NP0884A, HQ2919A
rrnAC3050	2,199	2,514	Shortened	rrnAC0848, rrnAC2044, HQ2141A, NP6028A
rrnAC3062	954	1,032	Shortened	rrnAC1698, OE1358R, HQ2259A
rrnAC3071	915	1,149	Shortened	OE3959R, HQ3234A, NP5036A
rrnAC3074	699	963	Shortened	NP0072A, HQ3236A, OE3964R
rrnAC3079	573	624	Shortened	NP3402A
rrnAC3083	375	423	Shortened	NP0948A, OE4292F, HQ3465A
rrnAC3100	2,010	2,049	Shortened	NP2262A, HQ1094A, OE3832F
rrnAC3121	651	1,194	Shortened	HQ1495A
rrnAC3130	939	645	Extended	HQ1618A, rrnB0227
rrnAC3132	1,392	1,419	Shortened	HQ1619A, rrnAC2624
rrnAC3137	1,047	798	Extended	NP0860A, HQ3045A, OE4446R
rrnAC3167	687	750	Shortened	OE1188F, rrnAC1953
rrnAC3182	627	666	Shortened	NP3516A, rrnAC1228
rrnAC3198	1,215	1,044	Extended	NP1072A
rrnAC3210	1,311	1,356	Shortened	NP4992A, OE3792F, HQ3101A
rrnAC3214	864	936	Shortened	HQ3098A, OE3787R, NP2524A
rrnAC3226	1,671	1,599	Extended	NP5136A
rrnAC3236	1,383	420	Extended	OE1018F, rrnAC1586
rrnAC3256	546	663	Shortened	NP3054A, HQ3084A, OE3752R
rrnAC3268	678	498	Extended	NP5010A, OE3731R, HQ3131A
rrnAC3272	981	834	Extended	NP5006A, HQ3129A, OE3735F
rrnAC3279	1,107	1,032	Extended	NP2398A, OE3722F, HQ3125A
rrnAC3302	786	735	Extended	OE3439F, HQ2249A
rrnAC3328	531	621	Shortened	NP1380A, OE1858F, HQ1684A
rrnAC3342	891	945	Shortened	NP4916A, OE3430F, HQ2764A
rrnAC3345	453	531	Shortened	HQ1964A, rrnAC2948, OE2717R
rrnAC3348	876	963	Shortened	NP4524A, HQ3322A, OE3531R
rrnAC3352	609	765	Shortened	NP4518A, OE3537R, HQ2230A
rrnAC3371	2,361	1,971	Extended	pNG2034, NP3562A, HQ1851A, OE5286R
rrnAC3385	573	396	Extended	OE3854R, NP2906A
rrnAC3394	444	525	Shortened	NP1284A, rrnB0323
rrnAC3420	843	885	Shortened	OE3633R, NP2432A, HQ3036A
rrnAC3450	366	399	Shortened	OE3588C1R, NP2486A, HQ3026A
rrnAC3452	894	951	Shortened	NP2484A, OE3586R, HQ3027A
rrnAC3462	1,929	2,019	Shortened	NP2410A, OE3580R, HQ2579A
rrnAC3475	1,113	594	Extended	NP6258A
rrnAC3486	636	687	Shortened	NP4286A
rrnAC3509	624	471	Extended	OE1814R, HQ1569A
rrnAC3528	459	477	Shortened	rrnAC3526, pNG2015
rrnAC3536	276	153	Extended	NP0840A, OE1853R
rrnAC3537	1,089	681	Extended	HQ1674A, OE1854R, NP0842A
rrnAC3551	1,026	750	Extended	NP2032A, HQ3010A, rrnAC1926, OE5141R
rrnB0092	1,533	1,599	Shortened	HQ1972A
rrnB0172	591	507	Extended	NP1842A
rrnB0198	1,530	1,623	Shortened	NP0556A, OE4115F
rrnB0242	1,173	1,287	Shortened	NP2606A, HQ2307A
rrnB0257	834	888	Shortened	NP4244A, OE1942F
rrnB0265	1,683	1,851	Shortened	NP4242A

Table 2 continued

ORF	Corrected length (aa)	Original length (aa)	Direction of change	Homologous ORFs
rrnB0266	1,404	1,011	Extended	NP3416A
rrnB0275	1,101	1,065	Extended	rrnAC0899, OE4576F
rrnB0325	1,107	1,278	Shortened	OE5142F, NP2128A, rrnAC3284, HQ3147A
pNG2007	939	1,119	Shortened	OE4023F, NP1282A, rrnAC2744, HQ3263A
pNG2015	516	615	Shortened	rrnAC3526, NP1956A
pNG4017	1,254	1,125	Extended	NP2168A, rrnAC2207, OE2401F
pNG4035	561	516	Extended	OE3768F, NP5358A
pNG5001	1,134	1,104	Extended	rrnAC0252, NP0102A, HQ1815A, OE1005F
pNG5004	1,488	1,068	Extended	rrnAC0250
pNG5010	1,632	879	Extended	OE5248F, HQ1543A, NP2464A
pNG5131	633	579	Extended	rrnAC3384, OE4753R
pNG5139	1,251	312	Extended	HQ2051A, NP6268A, OE1070R
pNG6047	591	618	Shortened	HQ1118A, OE2691R, rrnAC0503, NP2664A
pNG6054	423	1,047	Shortened	NP5022A
pNG6069	417	444	Shortened	NP4526A, rrnAC1905
pNG6075	378	477	Shortened	OE7144R, NP3058A
pNG6092	294	324	Shortened	pNG6058, OE7057F, NP3002A, HQ2407A
pNG6120	861	921	Shortened	OE5424R
pNG6141	615	585	Extended	OE5415R
pNG7012	1,281	1,026	Extended	OE1077R, rrnAC3239, HQ2680A, NP2322A
pNG7037	1,488	1,725	Shortened	NP5056A
pNG7040	1,182	999	Extended	pNG7041, OE4576F
pNG7050	750	705	Extended	HQ3696A, rrnAC0479, NP1198A, OE3661F
pNG7058	528	501	Extended	OE1252R, HQ2374A, NP1606A
pNG7060	1,272	1,302	Shortened	NP6204A
pNG7066	1,017	1,107	Shortened	OE2128F, HQ2746A, NP1386A
pNG7078	1,071	1,242	Shortened	NP1388A, HQ1592A
pNG7081	399	363	Extended	HQ4010A
pNG7101	1,971	2,004	Shortened	HQ1729A
pNG7106	897	819	Extended	OE2497F, HQ2422A, NP1346A
pNG7178	984	834	Extended	HQ2189A
pNG7227	747	612	Extended	NP0054A, HQ1091A, OE3843F
pNG7244	2,481	2,166	Extended	pNG7246, HQ1944A
pNG7252	1,659	1,788	Shortened	OE2316R, rrnAC2655, HQ2451A
pNG7278	1,041	1,155	Shortened	OE4674F, HQ1124A
pNG7280	540	489	Extended	pNG6134, NP5298A
pNG7297	603	630	Shortened	NP0672A
pNG7321	381	408	Shortened	HQ1769A, pNG7235, OE3930R, NP0566A
pNG7327	1,512	1,572	Shortened	HQ1784A, NP0802A, OE1568F
pNG7342	1,098	1,128	Shortened	pNG7026, OE5170F
pNG7351	1,065	1,089	Shortened	rrnAC0191, NP1260A, OE4674F, HQ3648A
pNG7377	432	324	Extended	HQ3372A
pNG7380	1,746	1,932	Shortened	HQ1768A

Using our semiautomatic checking procedure, candidate genes with probable errors in start codon assignment were identified and subjected to manual curation. When sufficiently strong evidences were found, the start codon was reassigned using the manual curation options of HaloLex. For each gene in the list (*first column*), we provide the corrected (*second column*) and original length (*third column*) of the amino-acid sequence, and the set of homologous genes that support our decision for the new start codon assignment (*fifth column*). The redundant *fourth column* facilitates a quick overview of whether sequences were extended or shortened with respect to their original length

curation, as it does not only support detailed analysis but, once a decision is taken, allows it to be conveniently made persistent with a few clicks.

Conclusions and outlook

We have described HaloLex, a software system for the central management, integration, and web-based visualization of genomic and other *-omics* data. A number of HaloLex functionalities are specifically tailored to halophilic archaea, but the system can handle any given microorganism.

HaloLex has proven an indispensable tool for the data management, curation, and in-depth bioinformatic analysis of three halophilic archaea sequenced in-house, namely *Halobacterium salinarum* (strain R1), *Natronomonas pharaonis*, and *Haloquadratum walsbyi.* HaloLex summarizes all available data for a given organism including experimental data, like, e.g., proteomics, in an easy-to-use web interface. This proved to be of enormous importance for both, the daily user of genome information as well as for the manual curator of the gene annotation in these organisms.

In this article, we further reviewed a number of selected, biologically relevant results we obtained for these species, thus highlighting the capabilities of HaloLex for prediction and curation of gene assignment, in particular by the integrated analysis of genomic with proteomic data.

Lately, we have applied HaloLex functionalities to the published genome of another halophilic archaeon, *Haloarcula marismortui,* which resulted in a significantly improved version of the original gene prediction.

Other halophiles (also from the bacterial kingdom) like *Halobacillus halophilus* are currently being annotated by different collaborations, which shows that HaloLex could be a useful tool also for a broader user-community. Based on our promising experiences, we thus encourage potential collaborators to consider employing our HaloLex server as a data repository and a tool for curation and analysis of their genomes (and proteomes, etc.) of interest. At the same time, HaloLex would allow such groups to make their data available to the public (or restricted user groups) without having to take up the burden of developing and hosting their own software and hardware infrastructure.

Our ongoing and future activities are focussed on making those data and methods fully available, which so far can be used only internally (e.g., data from transcriptomics experiments). Moreover, HaloLex functionalities are continuously being improved and extended. Currently, we are about to couple software modules for text-mining and metabolic modeling, which we are developing in our group to the HaloLex web application. We also plan to release our web-service interface to support mining of HaloLex data over the Internet.

Acknowledgments We thank Volker Hickmann and Jan Wolfertz who were involved in early phases of the HaloLex project, and Karin Gross who has recently joined the team. Development of the Halolex system benefited greatly from the MIGenAS infrastructure for bioinformatics software and data, which was developed with funds from the Max-Planck-Society. Part of this work was funded by the Deutsche Forschungsgemeinschaft within the priority program SP1112.

References

Aivaliotis M, Gevaert K, Falb M, Tebbe A, Konstantinidis K, Bisle B, Klein C, Martens L, Staes A, Timmerman E, Van Damme J, Siedler F, Pfeiffer F, Vandekerckhove J, Oesterhelt D (2007) Large-scale identification of N-terminal peptides in the halophilic archaea *Halobacterium salinarum* and *Natronomonas pharaonis*. J Proteome Res 6:2195–2204

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Baliga NS, Bonneau R, Facciotti MT, Pan M, Glusman G, Deutsch EW, Shannon P, Chiu Y, Weng RS, Gan RR, Hung P, Date SV, Marcotte E, Hood L, Ng WV (2004) Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea. Genome Res 14:2221–2234

Bisle B, Schmidt A, Scheibe B, Klein C, Tebbe A, Kellermann J, Siedler F, Pfeiffer F, Lottspeich F, Oesterhelt D (2006) Quantitative profiling of the membrane proteome in a halophilic archaeon. Mol Cell Proteomics 5:1543–1558

Bolhuis H, Palm P, Wende A, Falb M, Rampp M, Rodriguez-Valera F, Pfeiffer F, Oesterhelt D (2006) The genome of the square archaeon *Haloquadratum walsbyi*: life at the limits of water activity. BMC Genomics 7:169

Bryson K, Loux V, Bossy R, Nicolas P, Chaillou S, van de Guchte M, Penaud S, Maguin E, Hoebeke M, Bessieres P, Gibrat JF (2006) AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res 34:3533–3545

Dambeck M, Soppa J (2008) Characterization of a *Haloferax volcanii* member of the enolase superfamily: deletion mutant construction, expression analysis, and transcriptome comparison. Arch Microbiol. doi:10.1007/s00203-008-0379-1

Falb M, Pfeiffer F, Palm P, Rodewald K, Hickmann V, Tittor J, Oesterhelt D (2005) Living with two extremes: conclusions from the genome sequence of *Natronomonas pharaonis*. Genome Res 15:1336–1343

Falb M, Aivaliotis M, Garcia-Rizo C, Bisle B, Tebbe A, Klein C, Konstantinidis K, Siedler F, Pfeiffer F, Oesterhelt D (2006) Archaeal N-terminal protein maturation commonly involves N-terminal acetylation: a large-scale proteomics survey. J Mol Biol 362:915–924

Falb M, Müller K, Königsmaier L, Oberwinkler T, Horn P, von Gronau S, Gonzalez O, Pfeiffer F, Bornberg-Bauer E, Oesterhelt

D (2008) Metabolism of halophilic archaea. Extremophiles 12:177–196
Gonzalez O, Gronau S, Falb M, Pfeiffer F, Mendoza E, Zimmer R, Oesterhelt D (2008) Reconstruction, modeling & analysis of *Halobacterium salinarum* R-1 metabolism. Mol Biosyst 4:148–159
Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036
Karlin S, Brocchieri L, Trent J, Blaisdell BE, Mrazek J (2002) Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes. Theor Popul Biol 61:367–390
Keese PK, Gibbs A (1992) Origins of genes: "big bang" or continuous creation? Proc Natl Acad Sci USA 89:9489–9493
Kitano H (2002) Systems biology: a brief overview. Science 295:1662–1664
Klein C, Garcia-Rizo C, Bisle B, Scheffer B, Zischka H, Pfeiffer F, Siedler F, Oesterhelt D (2005) The membrane proteome of *Halobacterium salinarum*. Proteomics 5:180–197
Klein C, Aivaliotis M, Olsen JV, Falb M, Besir H, Scheffer B, Bisle B, Tebbe A, Konstantinidis K, Siedler F, Pfeiffer F, Mann M, Oesterhelt D (2007) The low molecular weight proteome of *Halobacterium salinarum*. J Proteome Res 6(4):1510–1518
Konstantinidis K, Tebbe A, Klein C, Scheffer B, Aivaliotis M, Bisle B, Falb M, Pfeiffer F, Siedler F, Oesterhelt D (2007) Genome-wide proteomics of *Natronomonas pharaonis*. J Proteome Res 6:185–193
Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I, Lykidis A, Mavromatis K, Ivanova N, Kyrpides NC (2006) The integrated microbial genomes (IMG) system. Nucleic Acids Res 34:D344–348
McHardy AC, Goesmann A, Pühler A, Meyer F (2004) Development of joint application strategies for two microbial gene finders. Bioinformatics 20:1622–1631
Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, Pühler A (2003) GenDB—an open source genome annotation system for prokaryote genomes. Nucleic Acids Res 31:2187–2195
Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, Swartzell S, Weir D, Hall J, Dahl TA, Welti R, Goo YA, Leithauser B, Keller K, Cruz R, Danson MJ, Hough DW, Maddocks DG, Jablonski PE, Krebs MP, Angevine CM, Dale H, Isenbarger TA, Peck RF, Pohlschroder M, Spudich JL, Jung KW, Alam M, Freitas T, Hou S, Daniels CJ, Dennis PP, Omer AD, Ebhardt H, Lowe TM, Liang P, Riley M, Hood L, DasSarma S (2000) Genome sequence of *Halobacterium* species NRC-1. Proc Natl Acad Sci USA 97:12176–12181
Nielsen P, Krogh A (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics 21:4322–4329
Pfeiffer F, Schuster SC, Broicher A, Falb M, Palm P, Rodewald K, Ruepp A, Soppa J, Tittor J, Oesterhelt D (2008) Evolution in the laboratory: the genome of *Halobacterium salinarum* strain R1 compared to that of strain NRC-1. Genomics 91:335–346
Pleissner KP, Eifert T, Buettner S, Schmidt F, Boehme M, Meyer TF, Kaufmann SH, Jungblut PR (2004) Web-accessible proteome databases for microbial research. Proteomics 4:1305–1313
Rampp M, Soddemann T, Lederer H (2006) The MIGenAS integrated bioinformatics toolkit for web-based sequence analysis. Nucleic Acids Res 34:W15–W19
Rattei T, Arnold R, Tischler P, Lindner D, Stumpflen V, Mewes HW (2006) SIMAP: the similarity matrix of proteins. Nucleic Acids Res 34:D252–D256
Riley ML, Schmidt T, Artamonova II, Wagner C, Volz A, Heumann K, Mewes HW, Frishman D (2007) PEDANT genome database: 10 years online. Nucleic Acids Res 35:D354–D357
Rose RW, Bruser T, Kissinger JC, Pohlschröder M (2002) Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway. Mol Microbiol 45:943–950
Scheuch S, Marschaus L, Sartorius-Neef S, Pfeifer F (2008) Regulation of gvp genes encoding gas vesicle proteins in halophilic Archaea. Arch Microbiol. doi:10.1007/s00203-008-0362-x
Schneider KL, Pollard KS, Baertsch R, Pohl A, Lowe TM (2006) The UCSC Archaeal Genome Browser. Nucleic Acids Res 34:D407–D410
Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295
Soppa J, Baumann A, Brenneis M, Dambeck M, Hering O, Lange C (2008) Genomics and functional genomics with haloarchaea. Arch Microbiol. doi:10.1007/s00203-008-0376-4
Stearns J, Chinnici R, Sahoo (2006) An Introduction to the Java EE 5 Platform. http://java.sun.com/developer/technicalArticles/J2EE/intro_ee5/
Stein L (2002) Creating a bioinformatics nation. Nature 417:119–120
Tarasov VY, Besir H, Schwaiger R, Klee K, Furtwängler K, Pfeiffer F, Oesterhelt D (2008) A small protein from the bop-brp intergenic region of *Halobacterium salinarum* contains a zinc finger motif and regulates bop and crtB1 transcription. Mol Microbiol 67:772–780
Tebbe A, Klein C, Bisle B, Siedler F, Scheffer B, Garcia-Rizo C, Wolfertz J, Hickmann V, Pfeiffer F, Oesterhelt D (2005) Analysis of the cytosolic proteome of *Halobacterium salinarum* and its implication for genome annotation. Proteomics 5:168–179
Teufel K, Bleiholder A, Griesbach T, Pfeifer F (2008) Variations in the multiple tbp genes in different Halobacterium salinarum strains and their expression during growth. Arch Microbiol. doi:10.1007/s00203-008-0383-5
Twellmeyer J, Wende A, Wolfertz J, Pfeiffer F, Panhuysen M, Zaigler A, Soppa J, Welzl G, Oesterhelt D (2007) Microarray analysis in the Archaeon *Halobacterium salinarum* strain R1. PLoS ONE 2:e1064
Veloso F, Riadi G, Aliaga D, Lieph R, Holmes DS (2005) Large-scale, multi-genome analysis of alternate open reading frames in bacteria and archaea. Omics 9:91–105
Wu CH, Huang H, Nikolskaya A, Hu Z, Barker WC (2004) The iProClass integrated database for protein functional analysis. Comput Biol Chem 28:87–96