EC
Journals.ASM.org

A Machine Learning Approach To Identify Hydrogenosomal Proteins in *Trichomonas vaginalis*

David Burstein,[a] Sven B. Gould,[b] Verena Zimorski,[b] Thorsten Kloesges,[b] Fuat Kiosse,[b] Peter Major,[b] William F. Martin,[b] Tal Pupko,[a,c] and Tal Dagan[b]

Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel[a]; Institute of Molecular Evolution, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany[b]; and National Evolutionary Synthesis Center, Durham, North Carolina, USA[c]

The protozoan parasite *Trichomonas vaginalis* is the causative agent of trichomoniasis, the most widespread nonviral sexually transmitted disease in humans. It possesses hydrogenosomes—anaerobic mitochondria that generate H_2, CO_2, and acetate from pyruvate while converting ADP to ATP via substrate-level phosphorylation. *T. vaginalis* hydrogenosomes lack a genome and translation machinery; hence, they import all their proteins from the cytosol. To date, however, only 30 imported proteins have been shown to localize to the organelle. A total of 226 nuclear-encoded proteins inferred from the genome sequence harbor a characteristic short N-terminal presequence, reminiscent of mitochondrial targeting peptides, which is thought to mediate hydrogenosomal targeting. Recent studies suggest, however, that the presequences might be less important than previously thought. We sought to identify new hydrogenosomal proteins within the 59,672 annotated open reading frames (ORFs) of *T. vaginalis*, independent of the N-terminal targeting signal, using a machine learning approach. Our training set included 57 gene and protein features determined for all 30 known hydrogenosomal proteins and 576 nonhydrogenosomal proteins. Several classifiers were trained on this set to yield an import score for all proteins encoded by *T. vaginalis* ORFs, predicting the likelihood of hydrogenosomal localization. The machine learning results were tested through immunofluorescence assay and immunodetection in isolated cell fractions of 14 protein predictions using hemagglutinin constructs expressed under the homologous SCSα promoter in transiently transformed *T. vaginalis* cells. Localization of 6 of the 10 top predicted hydrogenosome-localized proteins was confirmed, and two of these were found to lack an obvious N-terminal targeting signal.

The anaerobic parabasalian flagellate *Trichomonas vaginalis* infects the urogenital tract of hundreds of millions of people annually (55). In this organism, ATP is produced in hydrogenosomes by substrate-level phosphorylation rather than by a proton-driven and membrane-bound ATP-synthase complex (49). Hydrogenosomes share an ancestor with the mitochondrion, but their scattered distribution over the eukaryotic supergroups (some fungi, parabasalids, amoeboflagellates, ciliates, and at least one animal) indicates that the specialization of these mitochondria to the anaerobic lifestyle occurred several times in independent lineages during evolution (20, 32, 59). With the exception of the ciliate *Nyctotherus ovalis* (1) and the human parasite *Blastocystis* sp. (61, 82), hydrogenosomes typically lack their own genome and translation machinery, reflecting reductive evolution. This necessitates the import of hundreds of nuclear-encoded proteins from the cytosol (17, 31, 32, 59).

Understanding the biochemistry and molecular evolution of hydrogenosomes is of medical importance as the most common drug treatments—nitroimidazole derivates such as metronidazole—target hydrogenosomal proteins (6, 46). The common point of view is that pyruvate:ferredoxin oxidoreductase oxidizes pyruvate within the hydrogenosomes, upon which ferredoxin reduces the nitro moiety of the drug by transferring the electrons, ultimately leading to the release of short-lived cytotoxic radicals (34, 58, 78). An alternative malate-dependent pathway has furthermore been suggested, which nevertheless is also part of the hydrogenosomal biochemistry (34). Resistance to nitroimidazole derivates has been observed in anaerobic parasites such as *Giardia*, *Entamoeba*, and *Trichomonas* and in the last of these is known to be increasing (78, 83). However, we do not possess an exhaustive list of hydrogenosomal proteins, and proteomic approaches contained many apparent cytosolic contaminations (31, 71). A better understanding of hydrogenosomal proteins and their import into the *Trichomonas* organelle is important to the development of treatment strategies.

Targeting and translocation of proteins into yeast mitochondria have been studied in detail (reviewed in references 12, 50, 56, and 77). In contrast, little is known about the targeting mechanisms or the import machinery in hydrogenosomes. Only a few homologs of mitochondrial import machinery components have been identified in *T. vaginalis*. Two of these were shown to localize to the outer hydrogenosomal membrane (Hmp35 and Sam50) (18, 73). Import of precursors was shown to be ATP dependent, and early *in vitro* analyses suggested that correct targeting requires an N-terminal leader (9, 11), referred to in this article as a hydrogenosomal targeting signal sequence (HTS).

The genome of *T. vaginalis* contains 59,672 open reading frames (ORFs) (TrichDB, version 1.1 [5]), 226 of which encode the canonical HTS defined by Carlton and colleagues (11) as follows: ML(S/T/A)$X_{(1..15)}$R(N/F/E/XF) or MSL$X_{(1..15)}$R(N/F/XF) or MLR (S/N)F (11). The hydrogenosomal localization of only 30 proteins has been verified experimentally (11, 53, 63, 64, 79). The current

Received 29 August 2011 **Accepted** 17 November 2011
Published ahead of print 2 December 2011
Address correspondence to Tal Dagan, tal.dagan@hhu.de.
D. Burstein and S. B. Gould contributed equally to this article.
Supplemental material for this article may be found at http://ec.asm.org/.

doi:10.1128/EC.05225-11

estimate is significantly lower than the ~500 proteins expected to be found in the hydrogenosome (73). This is compounded by the finding that some HTS-lacking proteins are imported into hydrogenosomes, the alpha subunit of succinyl-coenzyme A (CoA) synthetase (TVAG_165340), and a thioredoxin reductase isoform (TVAG_125360) (53). Thus, protein properties in addition to an HTS are likely to serve as potential targeting precursors to the hydrogenosomes. Consequently, the *T. vaginalis* genome should encode hydrogenosomal proteins that have so far not been identified due to their lack of a canonical N-terminal HTS.

Our study aimed to predict proteins that are targeted to the hydrogenosome but with criteria that are independent of the canonical HTS. For that purpose, we have implemented a classification tool based on a machine learning approach to screen the entire *T. vaginalis* genome for proteins potentially targeted to the hydrogenosome. This approach allows us to extract information from various feature combinations in order to identify patterns within a known learning set (bait) and perform subsequent predictions on an unknown data set (prey). Machine learning algorithms have been used for biological data mining, including applications for prediction of protein targeting signals (see reference 70 for a review) or protein-protein interactions (37), and finding protein-encoding genes (72) and noncoding RNAs (51) within completely sequenced genomes. Using this approach we predicted and subsequently validated experimentally new hydrogenosomal proteins, some of which do not carry N-terminal targeting motifs.

MATERIALS AND METHODS

Machine learning classification. The machine learning analysis was implemented using the open source package WEKA, version 3.7.0 (29) with default parameters unless otherwise stated. Three learning phases were conducted. Predictions from the first two phases were experimentally validated. Information gained from these validations augmented the input for the subsequent learning phase. The learning procedures were performed on two data sets. The first data set includes the direct measures listed in Table 1. In the second data set all continuous variables were preprocessed into a discrete variable by binning their distribution into 10 equal-frequency bins.

Seven classifiers were used for the machine learning inference. The set of algorithms includes the naïve Bayes, which is a simple probabilistic classifier that assumes complete independence among the different features (48, 57). The Bayesian network classifier is based on a probabilistic representation of the relations between the features using graph theory (30). This classifier was used in combination with two different structure search algorithms: the K2 search algorithm (13, 14) with a maximum of 2, 3, or 4 parenting nodes and the tree-augmented network (TAN) Bayes search algorithm (26). The support vector machine (SVM) approach is based on a general linear model used to seek for possible patterns in the supplied features (10, 80). Two alternative kernels were used for the SVM learning process: the polynomial kernel and radial basis function (RBF) kernel. The performance of all classifiers was compared at the end of the learning process, and the best classification scheme was then selected for further analysis.

Feature selection was carried out to identify the subset of the 57 features that perform best with each combination of classifier and data set. The feature selection was performed by applying a "wrapper" (39, 44) using a best-first search algorithm including a greedy hill-climbing procedure augmented with a backtracking facility (15).

The performance of each learning scheme was evaluated by the area under the curve (AUC) score, which is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (24, 28). For the estimation of the classification performance, a 10-fold cross-validation was performed. The

TABLE 1 Features used for the learning[a]

Category and description of protein feature (reference)	*P* value[f]	
	First phase	Final phase
Sequence and function		
Protein sequence length	0.0014**	0.0013**
Gene sequence length	0.0014**	0.0013**
Import signal presence/absence[b]	2.20E−16**	2.20E−16**
Gene GC content	8.00E−06**	0.4109
Fit of gene GC content to the genomic total[c]	7.53E−10**	0.0008**
5′ UTR length	2.39E−05**	0.2062
3′ UTR length	0.0036**	0.6928
GO annotation (4)	NA	NA
Mean hydropathy index (47)	6.17E−08**	1.26E−10**
Positively charged amino acids content	0.0513*	2.67E−05**
Negatively charged amino acids content	0.0339**	0.6928
Neutral amino acids content	1.23E−07**	2.81E−09**
Polar amino acids content	1.29E−09**	1.68E−09**
Nonpolar amino acids content	7.06E−10**	2.35E−07**
Hydrophilic amino acids content	2.52E−13**	3.03E−08**
Hydrophobic amino acids content	1.61E−13**	1.94E−08**
Amino acid content for all 20 amino acids		
Alanine	1.48E−06**	5.59E−09**
Arginine	0.1942	0.4440
Asparagine	0.2024	0.6520
Aspartic acid	0.0002**	0.0004**
Cysteine	0.0014**	0.1338
Glutamic acid	0.0620	0.0052**
Glutamine	1.99E−11**	2.31E−06**
Glycine	0.0955	0.8803
Histidine	0.4409	0.7056
Isoleucine	0.5529	0.7056
Leucine	0.7944	0.5803
Lysine	0.4008	0.5757
Methionine	0.3706	0.0010**
Phenylalanine	0.0109**	0.0838
Proline	0.0017**	0.4345
Serine	0.0199**	0.0057**
Threonine	0.1561	0.6040
Tryptophan	0.3140	0.6520
Tyrosine	0.5148	0.4128
Valine	0.0552*	0.0894
Evolution		
Phylum of the nearest neighbor	NA	NA
No. of BBHs in the total data set[d]	8.25E−10**	4.03E−08**
No. and percentage of genomes with BBHs[e]	2.25E−13**	1.97E−10**
No. and percentage of BBHs in:		
Eukaryotes	0.0078**	2.31E−06**
Archaebacteria	3.42E−05**	0.2397
Alphaproteobacteria	3.42E−05**	1.14E−28**
Betaproteobacteria	1.36E−38**	3.40E−29**
Gammaproteobacteria	9.47E−47**	3.40E−29**
Epsilonproteobacteria	1.87E−44**	7.71E−15**
Deltaproteobacteria	7.10E−27**	1.05E−24**
Other bacteria	7.29E−42**	8.03E−25**

[a] Numerical features were compared between positives and negatives in first and final phases using Wilcoxon test and an FDR correction for multiple comparisons.
[b] Motif ML(S/T/A)$X_{(1..15)}$R(N/F/E/XF), MSL$X_{(1..15)}$R(N/F/XF), or MLR(S/N)F (25).
[c] Calculated by two features: (i) the *P* value of a χ^2 test with the total proteome and (ii) significance of the *P* value ($P < 0.01$, $0.01 < P < 0.05$, and $P > 0.05$).
[d] The best BLAST hit (BBH) is defined as the BLAST hit having the minimum E-value from among 687 prokaryotic genomes using a *T. vaginalis* ORF as a query.
[e] The *P* values of the number and percentages of all features concerning the distribution of BBHs in the different taxonomic groups are identical.
[f] **, Significant after FDR correction; *, significant but only before the FDR correction. NA, not applicable.

training set was shuffled and divided into 10 equally sized sets. The classifier was trained on 90% of the data, and the remaining 10% were used as an unseen test set to assess the classifier's performance. This procedure was repeated 10 times (10 folds), with a different 10% of the data randomly selected as the test set in each repeat. For each of the 10 folds, the AUC was calculated, and the mean AUC is reported. It should be noted that the data serving as a test set were excluded from the feature selection stage; i.e., the feature selection was performed separately for each fold of the cross-validation. This contributes to the independence between the data used for the learning process and evaluation process using unseen data. For the best-performing classifier an additional step of feature selection and training was performed on the entire training set. The resulting trained classifier was used to produce the import scores for all *T. vaginalis* ORFs. The unbalanced frequencies of imported and nonimported proteins included in the learning set (about 1:20) might render an overestimated AUC (38). In order to provide comparable performance estimates despite the bias of the training set, values for the area under the precision recall curve (AUPR) were calculated as well, using AUCCalculator, version 0.2 (38). The proteins selected for validation in the laboratory represent a mix of high- and low-import probabilities based on the presence and absence of the HTS motif (MOT+ and MOT− schemes, respectively). In two proteins (TVAG_129210 and TVAG_171100) the import scores of the two schemes were opposite.

Data. The draft genome sequence of *T. vaginalis* was downloaded from TrichDB, version 1.1 (5). A total of 15 eukaryotic and 687 prokaryotic (629 eubacterial and 60 archaebacterial) genomes were downloaded from the November 2009 version of the RefSeq database (62) for the evolutionary reconstruction (see Table S1 in the supplemental material). For each ORF of *T. vaginalis*, 57 features were included regarding the gene and protein sequence, protein function, evolutionary relationships, the existence of an import signal, and gene ontology (GO) annotation (Table 1; see Table S2 for a detailed description). For the inference of the evolutionary features, each of the *T. vaginalis* ORFs was subjected to a BLAST search (3) against the 702 query genomes. The BLAST hits were sorted using as thresholds an E-value of ≤1E−10 and ≥25% for the percentage of identical amino acids. Each ORF was aligned with its homologs using Muscle (19). Phylogenetic trees were reconstructed by the neighbor-joining (NJ) method (68) with the default Jones-Taylor-Thorton (JTT) substitution matrix (41) using the Phylip package (25).

Paralogous protein families were reconstructed by conducting a BLAST search using all ORFs against the complete *T. vaginalis* proteome. Query hit pairs with an E-value of ≤1E−10 and percentage identical amino acids of ≥25% were aligned with Needleman-Wunsch global alignment (60) using the needle software included in the EMBOSS package (66). Pairwise protein similarity was calculated as the percentage of identical amino acids between the two proteins in the global alignment. Clusters of paralogous protein families were reconstructed from the protein similarities with the Markov cluster (MCL) algorithm (22) using the default parameters. The clustering was repeated using increasing protein similarity thresholds for the inclusion in the data ranging between 30% and 95% (T_{30} and T_{95}, where T is threshold).

Secondary structure predictions of the proteins were performed using PSIPRED (40) with the Swiss-Prot (7) database as input. Only amino acids with a confidence score higher than 0.7 were included in the analysis. Proteins having a secondary structure prediction for less than 70% of their sequence were marked as secondary structure unknown.

The training set of the first learning phase included the experimentally validated imported proteins and 576 nonimported proteins that were chosen based on their GO annotation (4) indicating a strict cytosolic localization. GO terms that were used include ribosomal and flagellar proteins, proteins from various amino acid metabolism pathways, transcription factors, and RNA polymerase subunits. The training set of the third phase included 37 imported proteins and 736 nonimported ones (see Table S1 in the supplemental material). The imported proteins included the 30 known imported proteins, 6 proteins validated in this study, and one additional protein validated in another study in our lab. The proteins in the negative set were selected based on their annotation in the TrichDB database (5). The following keywords were used for the selection: nuclear, ribosomal, histone, polymerase, actin, tubulin, dynein, flagellar, helicase, and DNA. All proteins in the training set were chosen so that there is an indication that they are expressed (number of expressed sequence tags [ESTs] > 0). Six proteins that were localized to the cytosol as part of an additional study in our lab were added to the negative set as well.

Culture conditions and transfection. Strain T1 of *T. vaginalis* was cultured in TYM medium at 37°C as previously described (54). Full-length coding sequences (ORFs) were retrieved from http://trichdb.org/trichdb/ and amplified without the stop codon from genomic DNA isolated from 50 ml of culture using DNAzol, according to the manufacture's protocol (Invitrogen, Germany). Genes were cloned into pTagVag2 (35) providing the gene of interest with a 3′ encoded, double hemagglutinin (HA) tag. For transfection, an electroporation protocol developed by Delgadillo and colleagues (16) was used. Briefly, 50 ml of cells (exponential growth phase) was collected at 1,500 × *g* at 4°C for 10 min, and the cells were then passed four times through a 23-gauge needle. A total of 300 μl of cells (2.5×10^8 cells) and 50 μg of pTagVag2 plasmid (35) harboring the gene of interest plus a C-terminal HA tag were mixed and pipetted into a 0.4-cm electroporation cuvette. Electroporation was carried out at 350 V and 950 μF. After the transfection, cells were cooled on ice for 10 min and then inoculated in 12 ml of TYM medium (containing 1% [vol/vol] penicillin-streptomycin solution [MP Biomedicals]). For selection the medium was then supplemented with 100 μM G418.

Protein localization. Isolation of hydrogenosomes was based on the method described by Bradley et al. (9) with slight modifications. After the cells were ground, unlysed cells, glass beads, crude membranes, and nuclei were removed by centrifugation at 755 × *g* for 10 min at 4°C and the whole-cell lysate from the supernatant was collected. The cytosolic fraction (supernatant) was obtained by subsequent centrifugation of the whole-cell lysate at 7,500 × *g* for 10 min at 4°C. The pellet was resuspended in 45% Percoll; the hydrogenosomes were separated by isopycnic centrifugation as described by Bradley and colleagues (9). Protein concentrations were determined with a Bradford assay kit (Bio-Rad) according to the manufacturer's instructions. Protein samples (20 μg each) were run on 12% resolving gels (sodium dodecyl sulfate-polyacrylamide gel electrophoresis) and blotted onto nitrocellulose membranes (Hybond-C Extra; Amersham Biosciences) for Western blot analysis. Blots were washed (three times for 10 min each) in TBS (20 mM Tris-HCl, pH 7.5, 150 mM NaCl) and blocked for 1 h in TBS containing 3% (wt/vol) bovine serum albumin (BSA). Blots were incubated for 1 h at room temperature, with a subsequent 1 h of incubation with mouse anti-HA antibodies (dilution, 1:5,000; Sigma). Blots were washed as before and incubated with anti-mouse horseradish peroxidase conjugate (ImmunoPure goat at a dilution of 1:10,000; Pierce) in TBS containing 3% (wt/vol) dry milk powder for 1 h at room temperature. After three subsequent washes in TBS, signals were visualized using 4 ml of solution A (1.25 mM Luminol [Sigma] in 0.1 M Tris-HCl, pH 6.8), 400 μl of solution B (6 mM *para*-hydroxycoumaric acid [Sigma] in dimethyl sulfoxide [DMSO]), and 1.2 μl of 30% (vol/vol) H_2O_2 and Lumi-Film chemiluminescent detection film (Roche).

Expressed HA-tagged proteins and acetate:succinate CoA-transferase ([ASCT] a hydrogenosomal marker) were visualized in *T. vaginalis* cells with mouse anti-hemagglutinin monoclonal antibody (Sigma-Aldrich, Germany) and rabbit anti-ASCT polyclonal antibody (79) as primary antibodies and with secondary Alexa Fluor-488 donkey anti-mouse and Alexa-Fluor-594 donkey anti-rabbit antibodies (Invitrogen, Karlsruhe, Germany). Images were processed with an LSM 510 Meta confocal laser scanning microscope (Zeiss, Germany) using the software Image Browser (Zeiss). Cells from a logarithmic phase *T. vaginalis* culture were placed on glass silane-coated microscopic slides (Electron Microscopy Sciences, Hatfield, PA) for 15 min at 37°C in an anaerobic chamber and dried almost completely at room temperature. The cells were then fixed in two subsequent steps by methanol (5 min) and acetone (5 min) at −20°C and

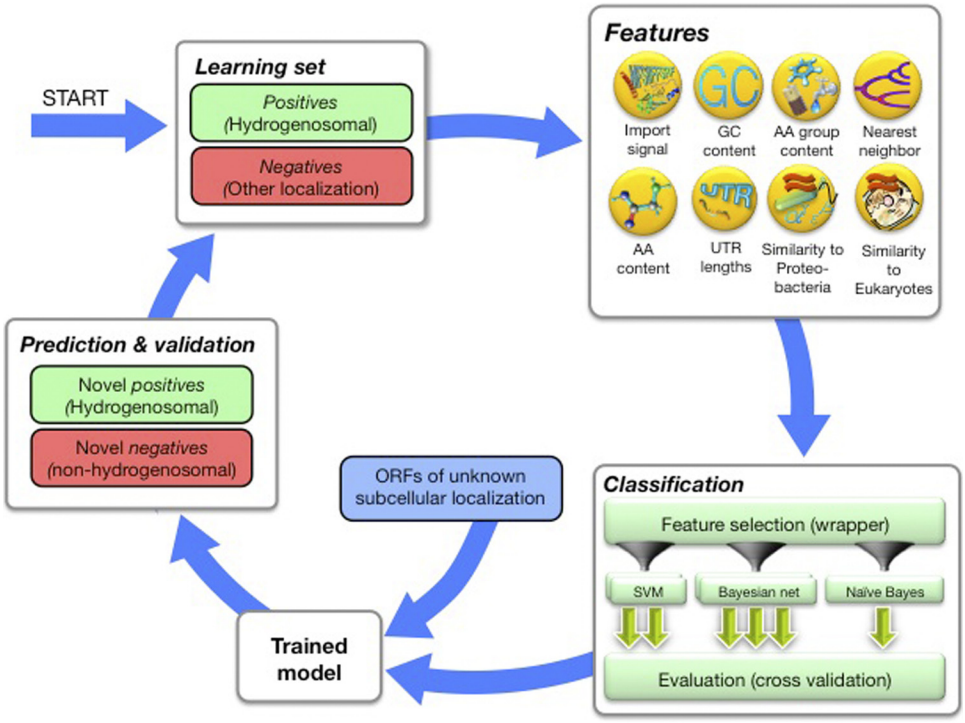

FIG 1 The machine learning procedure. For a learning set comprising all proteins known to be targeted to the hydrogenosome (positive set) and a set of nontargeted proteins (negative set), 57 different features were calculated. These values are passed to several classifiers, which aim to identify feature combinations that best differentiate between the positive and negative sets. In order to choose the best-performing classifier, 10-fold cross validation is performed. Within each fold, an inner cross validation is done to choose the best-performing features (feature selection). After the best classifier has been chosen, it is trained again over all of the learning set and is used to perform the prediction for each ORF in the *T. vaginalis* genome. The localization of the top-scoring predictions is experimentally tested. Newly identified hydrogenosomal proteins are added to the positive set, and another phase of learning can be performed.

treated with 0.25% gelatin and 0.25% BSA in phosphate-buffered saline ([PBS] 8% [wt/vol] NaCl, 0.2% [wt/vol] KCl, 1.44% [wt/vol] Na_2HPO_4, 0.24% [wt/vol] KH_2PO_4, pH 7.4) for 1 h at room temperature. The slides were then flooded with both primary antibodies (diluted 1:500) and incubated for 1 h at room temperature. After three 10-min washes in PBS, the slides were incubated with secondary antibodies (diluted 1:1,000) for 1 h at room temperature in the dark. After the slides were washed as described above, they were mounted in Vectashield with 4′,6′-diamidino-2-phenylindole (DAPI; Vector Laboratories, Burlingame, CA).

RESULTS

Hydrogenosomal localization prediction. The input for the machine learning classifiers includes 57 features, measured for each of the 59,672 protein annotations based on the *T. vaginalis* genome. These features comprise information about the gene sequence, HTS presence, physiochemical properties, function, and phylogeny of the protein (Table 1). All proteins are divided into three groups. The first includes proteins whose hydrogenosomal localization was known prior to the machine learning analysis, and these are designated positives. The second group includes proteins that localize to other parts of the cell and are, hence, designated negatives. Together, these two groups comprise the learning set. The third group includes all remaining *T. vaginalis* proteins whose subcellular localization is unknown. The learning set is used for both training and testing the classification algorithms. Three machine learning classifier algorithms were tested: naïve Bayes, Bayesian networks, and a support vector machine (SVM). For each classifier, a phase of feature selection was performed in which the best-separating subset of features is selected. The accuracy of each classifier is the average performance over 10-fold cross-validations (see Materials and Methods), and the best-performing algorithm was subsequently used. The classification process results in a prediction score, S_{import}, that quantifies the likelihood for a given protein to be localized to the hydrogenosome. A protein having a high S_{import} score (close to 1) has features similar to the imported proteins in the learning set and is predicted to be imported into the hydrogenosome. To test the essentiality of the import motif for hydrogenosome targeting, we executed the machine learning twice, with and without the HTS presence/absence feature. We designate these two schemes MOT+ and MOT−, for with and without the HTS motif, respectively. During the study we conducted three phases of machine learning prediction and validation in the lab. The initial learning set included proteins whose hydrogenosomal localization was reported in the literature (positives) and proteins whose function is unique to other subcellular localizations (negatives). In each phase we added the results of the localization experiments from the previous round into the learning set. In what follows we present the results of the final classification phase (Fig. 1).

Forty-one of the 55 numeric features were found to differ significantly between the positive and negative learning sets (Table 1). The remaining 14 numeric features, all measuring amino acid properties, were included in the inference procedure as well since it is possible that synergistic effects exist among different features that can be identified only during the learning process. The feature

selection process that was applied to the data prior to the classification step aims to select combinations of features, which maximizes the classifier performance in distinguishing between positive and negative proteins. To estimate the prediction robustness for each feature, a feature stability score was used. This score is calculated as the fraction of 10-fold cross-validation repeats in which the feature was selected by the feature selection process. For example, the score of a feature that was selected in 2 out of the 10 (10-fold) cross validations is 0.2. Features that are found informative by the feature selection process in all 10 folds are highly robust, and their stability score is set to 1. The two learning schemes resulted in overall similar feature stability scores (Fig. 2). In both MOT+ and MOT− schemes, the most robust feature was sequence similarity to *Betaproteobacteria* that was consistently selected in all cross validations. Other features that received high stability scores (>0.7) in both schemes include the length of 5′ untranslated regions (UTRs), hydrophobic and hydrophilic amino acid content, arginine count, and the number of homologous sequences in eukaryotes. Notably, the lengths of the 5′ UTRs that received very high stability scores (0.9 and 0.8 in MOT+ and MOT−, respectively) do not differ significantly between the positives and negatives in the learning set. It is possible that this feature alone is not informative for a distinction between imported and nonimported proteins but in combination improves the classification performance. Interestingly, the phylogenetic features received high stability scores, including the number of hits (homologs) in the various *Proteobacteria* classes and the identity of the nearest neighbor in the phylogenetic tree (Fig. 2).

The accuracy of the machine learning inference was measured by the area under the curve (AUC) of the receiver operating characteristics (ROC) curve. This measure quantifies the rate of true positive versus false positive in the classification procedure. Additionally, we calculated the area under the precision recall curve (AUPR), which is a more accurate performance estimator used for strongly biased data sets (38). The classification performance with both tested schemes was very high, with AUC values above 0.978 and AUPR values above 0.816 (Table 2). The mean AUCs of the various classifiers were 0.96 ± 0.003 and 0.95 ± 0.003 for the MOT+ and MOT− schemes, respectively. The best classifiers in both schemes were the Bayesian network classifiers; however, the tiny performance coefficient of variation among the different classifiers (0.3%) indicates that they performed similarly. Most of the proteins in both schemes received very low S_{import} values (see Table S1 in the supplemental material), in accordance with the observation that most *Trichomonas* proteins are not targeted to the hydrogenosome. A small fraction of proteins, however, obtained S_{import} values higher than 0.9: 720 (1.2%) proteins in the MOT+ learning scheme and 345 (0.57%) proteins in the MOT− learning scheme (see Fig. S1A). In both schemes, 53,654 (90%) proteins had S_{import} scores lower than 0.05, and 201 (0.33%) proteins had S_{import} scores higher than 0.9. However, the overall correlation between S_{import} values from the MOT+ and MOT− schemes is not high ($r_s = 0.43$; $P \ll 0.01$). Several proteins received high S_{import} scores using one scheme and low scores using the other. For example, 12 proteins had S_{import} scores higher than 0.95 in MOT− and lower than 0.05 in MOT+ (see Fig. S1B). Hence, an exclusion of the import motif feature from the machine learning analysis results in a different set of proteins that are predicted as targeted to the hydrogenosome. Importantly, the number of proteins with S_{import} scores higher than 0.95 in both the MOT+ and MOT− learning schemes is 673, which is close to the estimated number of about 500 hydrogenosomal proteins (73).

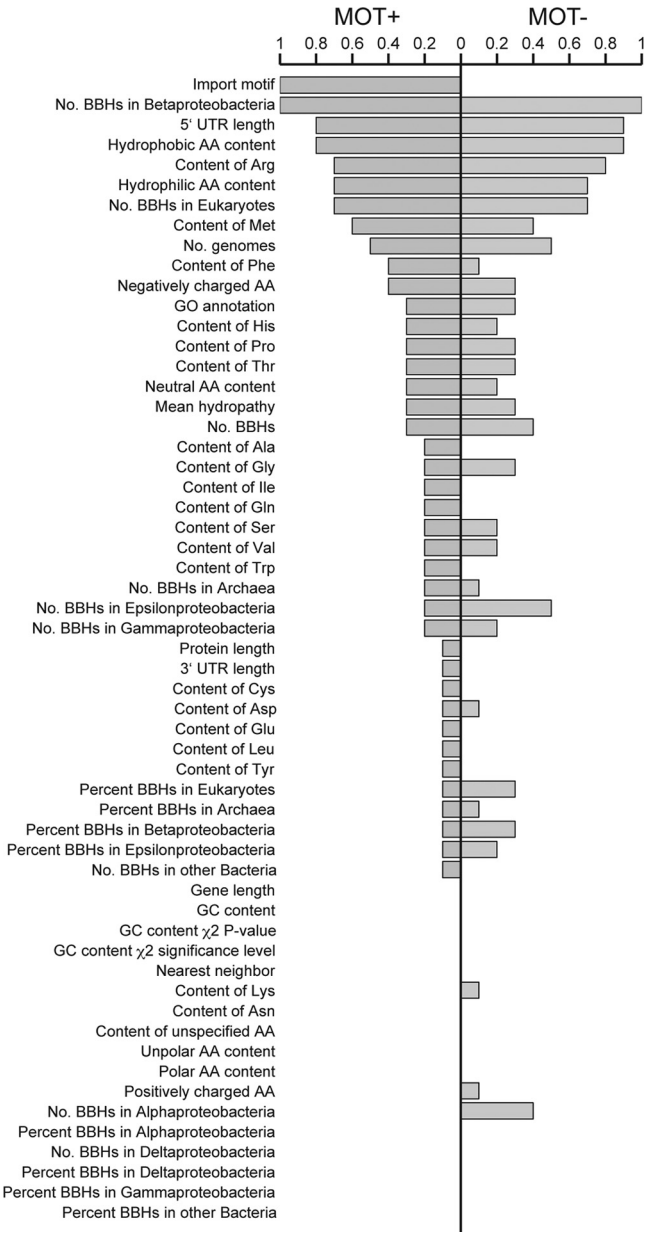

FIG 2 A comparison of feature stability score using the MOT+ and MOT− schemes. Using the 10-fold cross-validation approach, the estimation of the classifier performance is repeated 10 times (10 folds; see Materials and Methods for details). In each repeat, a different set of best features may be selected. Feature stability measures the fraction of the cross-validation repeats in which the feature was selected. A feature that was selected repeatedly in all of the 10 folds will receive a score of 1, indicating that the feature was found to be consistently informative for the distinction between positive and negative sets. BBH, best BLAST hits; AA, amino acid.

Hydrogenosomal localization validation. We selected 14 proteins for experimental validation (Table 3) based on their S_{import} scores. Ten out of these 14 have high scores at least in one of the learning schemes (MOT+ or MOT−) and are predicted to be localized to the hydrogenosome. Four had very low scores and are not predicted to be localized to the hydrogenosome. Out of the 10

TABLE 2 Machine learning predicted accuracy

Scheme	Accuracy measure	Predicted accuracy by learning phase	
		Initial	Final
MOT+	AUC	0.98	0.99
	AUPR	0.84	0.96
MOT−	AUC	0.98	0.99
	AUPR	0.82	0.90

high-scoring predictions, four include a canonical N-terminal import motif, as described previously (11). Proteins were hemagglutinin (HA) tagged at their C termini, and their subcellular localizations were determined by cell subfractionation and subsequent Western blot analysis without distinguishing between subhydrogenosomal localization. Potential contamination by cytosolic proteins within the hydrogenosomal fraction was monitored by control Western blots detecting actin, and the localization was furthermore checked by *in situ* immunolocalization (Fig. 3; see also Fig. S2 in the supplemental material). Altogether, our predictions were correct in 10 of these 14 proteins (71%). All four low-scoring predictions were found not to localize to the hydrogenosome (true negative). Out of the 10 high-scoring predictions, we localized six novel proteins to the hydrogenosomes of *T. vaginalis*, two of which lack the canonical HTS (Table 3).

Out of the four proteins harboring an HTS and for which hydrogenosomal localization was verified, TVAG_456770 and TVAG_361540 are paralogs of the iron sulfur biosynthesis protein IscA (Table 3; Fig. 3). The proteins contain an HTS slightly different from each other and overall share 69% identical amino acids. Together with another iron sulfur assembly protein (TVAG_055320), they form a three-member protein family at the threshold of 60% identical amino acids (T_{60}). The third member lacks the canonical HTS defined above but harbors a similar HTS prefix (Table 3). This protein received a low import score in both schemes (Table 3). Proteins such as IscS, IscU, and IscA involved in FeS cluster assembly are typically found present in mitochondrial, mitosomal, and hydrogenosomal organelles (20, 21, 75, 76). In *T. vaginalis* the IscS have been shown to localize in the hydrogenosome (74).

An additional HTS-harboring protein that we localized in the hydrogenosome is the chaperonin (HSP60) protein (TVAG_088050). This protein has two paralogs at T_{70}; one of them (TVAG_203620) has an HTS and was previously localized to the hydrogenosome (8). The import score of the other member (TVAG_167250), which has an HTS, too, is high in the MOT+ scheme (Table 3).

The final validated HTS-harboring protein (TVAG_129210) is of unknown function and is annotated as a conserved hypothetical protein (Table 3). No homologs for this protein were found within the genomes included in our study or by a global online BLAST query at NCBI. A sequence search against the *T. vaginalis* genome yielded 239 paralogous sequences at T_{95}. All of the paralogs have an identical 5′ sequence of the first 6 amino acids, but only TVAG_129210 has the known import motif "MSLSKSEREF." The import score of the paralogs is low, ranging between 0.0001 and 0.44, and none of them is expressed (EST frequency in TrichDB, 0). Hence TVAG_129210 is a *T. vaginalis*-specific protein that belongs to a huge protein family with a single member that is imported into the hydrogenosome.

Evidence for the HTS not solely being responsible for correct targeting comes from the 4-amino-acid short HTS of the pyruvate:ferredoxin oxidoreductase subunit A (PFOA), which is processed after the enzyme is imported (36). There are four copies of this gene in the nucleus, encoding four isoenzymes with at least 80% sequence identity, and two different HTSs: "MLRS" in TVAG_198110 and "MLRN" in TVAG_242960, TVAG_230580, and TVAG_254890. In a screening of almost 60,000 potential proteins, MLRS is found on 17, and MLRN is found on 13 proteins in total. These include among others an axonemal dynein light chain and a ubiquitin-dependent peptidase (TVAG_499270 and TVAG_050730, respectively) and a potential mannosyl-transferase of the endoplasmic reticulum (ER) membrane (TVAG_365830). We analyzed the latter and could localize the protein to the ER, which in *T. vaginalis* is tightly wrapped around the nucleus (Fig. 4). Intriguingly, the only HTS to our knowledge essential for import is that of a hydrogenosomal thioredoxin reductase (TrxRh1, TVAG_281360) (53), and that is found only once in the genome—on the TrxRh1 protein itself.

Two of the novel hydrogenosomal proteins harbor no N-terminal HTS as defined above (Table 3; Fig. 3). The first, TVAG_479680, carries an nitropropane dioxygenase (NPD)-like domain (52) and is annotated as a 2-nitropropane dioxygenase (EC 1.13.12.16) and might be involved in oxidative denitrification of nitroalkanes to carbonyl and nitrite compounds (52). This result exemplifies the utility of the machine learning approach to identify imported proteins that carry a noncanonical HTS. The TVAG_479680 protein has homologs in various bacteroidetes and several *Leishmania* species. A phylogenetic network analysis of this protein groups it with its eubacterial homologs rather than the *Leishmania* lineage (Fig. 5). The second HTS-lacking protein is TVAG_221830, which contains a Glo-EDI-BRP-like domain (52) and is annotated as a lactoylglutathione lyase (EC 4.4.1.5). The protein domain encoded by this gene groups it with a protein superfamily that includes metalloproteins and antibiotic resistance proteins (52). A BLAST search at NCBI using the protein sequence yielded several proteins having a similar domain in *Fusobacteria* (Fig. 6). Neither of the above two proteins has paralogs in *T. vaginalis*.

Four of the tested proteins for which high import scores were initially calculated by one or both of the learning schemes were found to be localized only in the cytosol (TVAG_064650, TVAG_062520, TVAG_204360, and TVAG_171100 in Table 3). The import scores calculated for these proteins in the final learning phase that included the newly identified hydrogenosomal proteins decreased considerably (Table 3). This result indicates that the addition of newly identified hydrogenosomal proteins to the learning set (positives) improved the accuracy of the algorithm.

Posthoc analysis. After the final learning phase, which included our localization results, we reexamined how the various features differ between hydrogenosome-imported and nonimported proteins. To that end, the Wilcoxon signed-rank test was used and corrected for multiple testing, using a false-discovery rate test (FDR) (33) (Table 1). Sequence similarities to *Gammaproteobacteria* and *Betaproteobacteria* homologs were the features with the most significant difference between imported and nonimported proteins (P value, 3.40×10^{-29}). Other features regarding similarity to proteobacteria and other eubacteria received very significant values as well (P values between 1.14×10^{-28} and 7.71×10^{-15}). Numerous features regarding the amino acid con-

TABLE 3 Results of experimental validation

Protein group and identification[a]	Annotation	Initial S_{import} by scheme		Final S_{import} by scheme		Import motif	Imported	No. of ESTs	Protein length	Import motif sequence	Import motif offset	Nearest neighbor
		MOT+	MOT−	MOT+	MOT−							
Hydrogenosomal import												
TVAG_361540 (1)	Iron sulfur assembly protein	1.00	1.00			+	+	10	151	MLSQAFRAF	1–9	*Opitutus terrae* PB90 1 (*Chlamydiae*)
TVAG_456770 (1)	Iron sulfur assembly protein	1.00	0.88			+	+	5	81	MLSSIIRSF	1–9	*Chlamydomonas reinhardtii*
TVAG_129210	Conserved hypothetical protein	1.00	0.00			+	+	0	487	MSLSKSEREF	1–10	No BLAST homologs
TVAG_088050 (2)	Chaperonin (HSP 60)	1.00	0.93			+	+	28	323	MLSKASSAFVRSF	1–13	*Entamoeba histolytica* HM-1:IMSS
TVAG_479680	2-Nitropropane dioxygenase	0.99	0.98			−	+	14	81			*Bacteroides fragilis* YCH46 (*Bacteroidetes*)
TVAG_221830	Conserved hypothetical protein	0.99	1.00			−	+	116	492			*Treponema denticola* ATCC 35405 (*Spirochaetes*)
TVAG_064650	Conserved hypothetical protein	1.00	1.00	0.00	0.37	−	−	108	323			*Listeria monocytogenes* HCC23 (*Firmicutes*)
TVAG_062520	Cyclophilin	1.00	0.99	0.25	0.27	−	−	187	323			*Pirellula* sp. (*Planctomycetes*)
TVAG_204360	Malate dehydrogenase	0.88	1.00	0.87	0.00	−	−	1186	397			*Entamoeba histolytica* HM-1:IMSS
TVAG_171100	Malate dehydrogenase	0.00	0.97	0.01	0.02	−	−	263	96			*Danio rerio*
TVAG_464170	Enolase	0.07	0.00	0.93	0.01	−	−	673	81			*Eukaryota*
TVAG_139300	PEP Carboxykinase	0.00	0.14	0.94	0.04	−	−	66	96			*Canditatus Koribacter versatilis* Ellin34s (*Acidobacteria*)
TVAG_069460	AMP-dependent ligase	0.00	0.64	0.01	0.75	−	−	6	323			*Dictyostelium discoideum* AX4
TVAG_023840	Conserved hypothetical protein	1.00E-05	1.20E-04	2.43E-04	3.32E-05	−	−	5	378			No BLAST homologs
Paralogs of validated proteins												
TVAG_055320 (1)	Iron sulfur assembly protein, putative			0.32	0.048	−	?	51	134			
TVAG_203620 (2)	Rubisco subunit binding-protein alpha subunit			0.99	0.21	+	+	18	558	MLSKQASSAFIRSF	1–14	
TVAG_167250 (2)	Chaperonin, putative			0.98	0.0027	+	?	2	560	MSLIEAAKHFTRAF	1–14	

[a] Protein families are identified by matching numbers in parentheses.

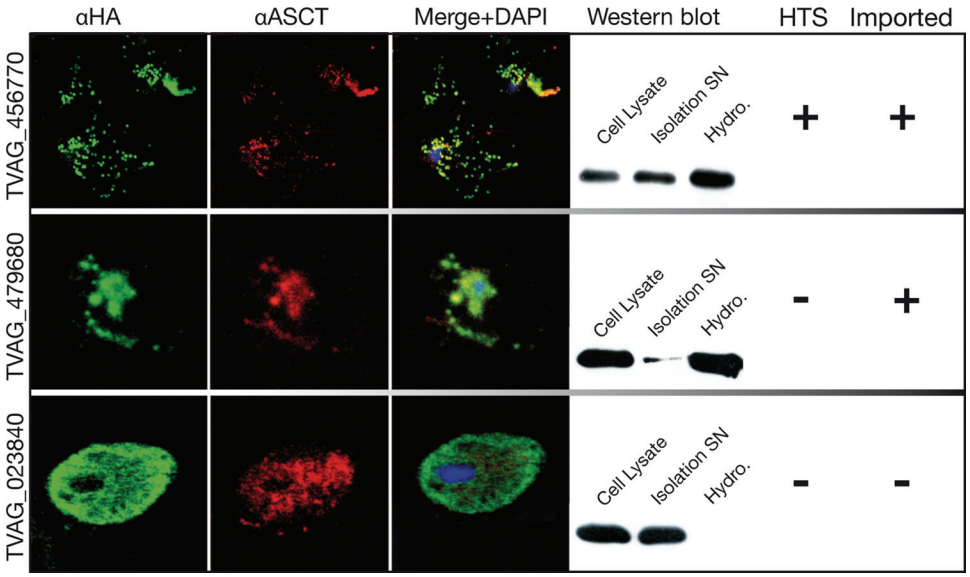

FIG 3 Results of the *in vivo* localization of two novel hydrogenosomal proteins: TVAG_456770 (a paralog of the iron sulfur biosynthesis protein IscA), TVAG_479680 (2-nitropropane dioxygenase), and, as a negative control, TVAG_023840 (glucokinase), together with the hydrogenosomal marker ASCT (TVAG_ 395550). α, anti.

tent of the proteins also showed highly significant differences. The most significant of these were mean hydropathy, polar and nonpolar amino acid content, and the content of arginines, all four with a P value lower than 10^{-8}. Another feature that received a very significant P values is the total number of BLAST hits. This is probably due to strong correlation between the number of BLAST hits in eubacteria and the total number of BLAST hits ($r_s = 0.782$; P values, $< 2.2 \times 10^{-16}$).

To test whether protein secondary structure correlates with the localization to the hydrogenosome, we compared the structural composition between top-scoring proteins in the last learning phase ($S_{\text{import}} > 0.9$) and the remaining proteins. We found that the top-scoring proteins are strongly enriched with beta sheet (P values of 2.57×10^{-9}, Wilcoxon test) and depleted of coiled segments (P values of $= 0.002$, Wilcoxon test). However, adding the secondary structure as a feature in the machine learning improved only slightly the AUC

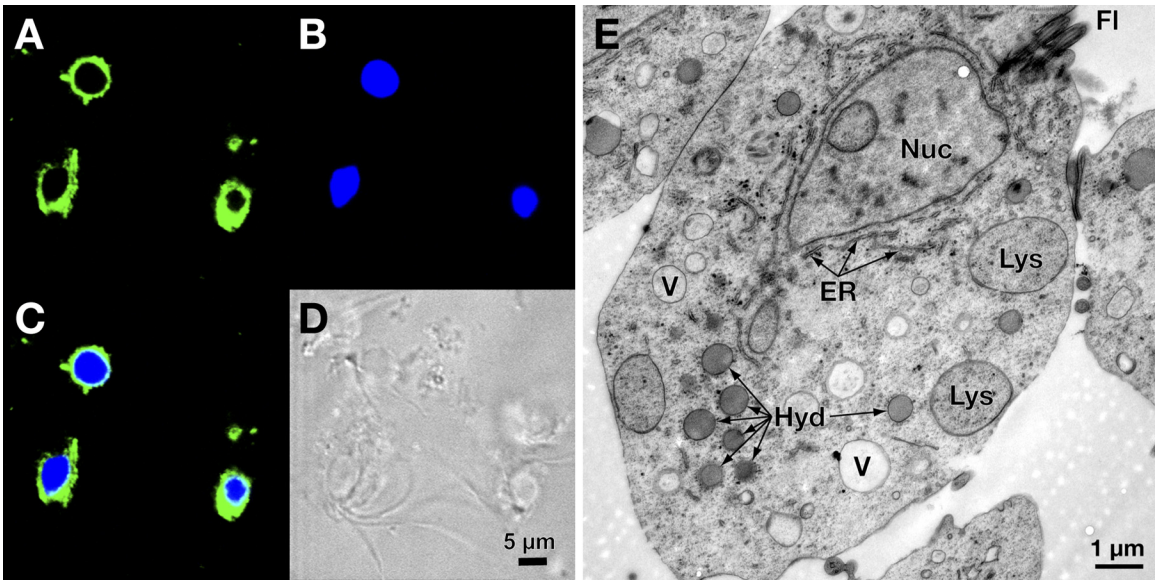

FIG 4 Localization of the mannosyl-transferase encoded by the TVAG_365830 gene. This mannosyl-transferase homologue possesses the same N-terminal sequence (MLRN) as found in PFO, but while PFO is imported into hydrogenosomes (Hyd) and the presequence is cleaved (36), TVAG_365830 is localized to the ER, despite possessing the same N terminus as pyruvate:ferredoxin oxidoreductase. (A) HA-tagged TVAG_365830. (B) DAPI staining. (C) Merge of the images in panels A and B. (D) Bright-field image. (E) An illustration of the typical arrangement of the ER (arrows) around the nucleus (Nuc) in a transmission electron microscopic image of *T. vaginalis*. When not attached to host tissue, flagellated *T. vaginalis* cells are pyriform and about 20 μm in length. A single cell can house several dozen hydrogenosomes, which are often found clustered in proximity to the axostyle (not visible in this section). Other membrane-bound structures include lysosomes (Lys) and vacuolar compartments (V).

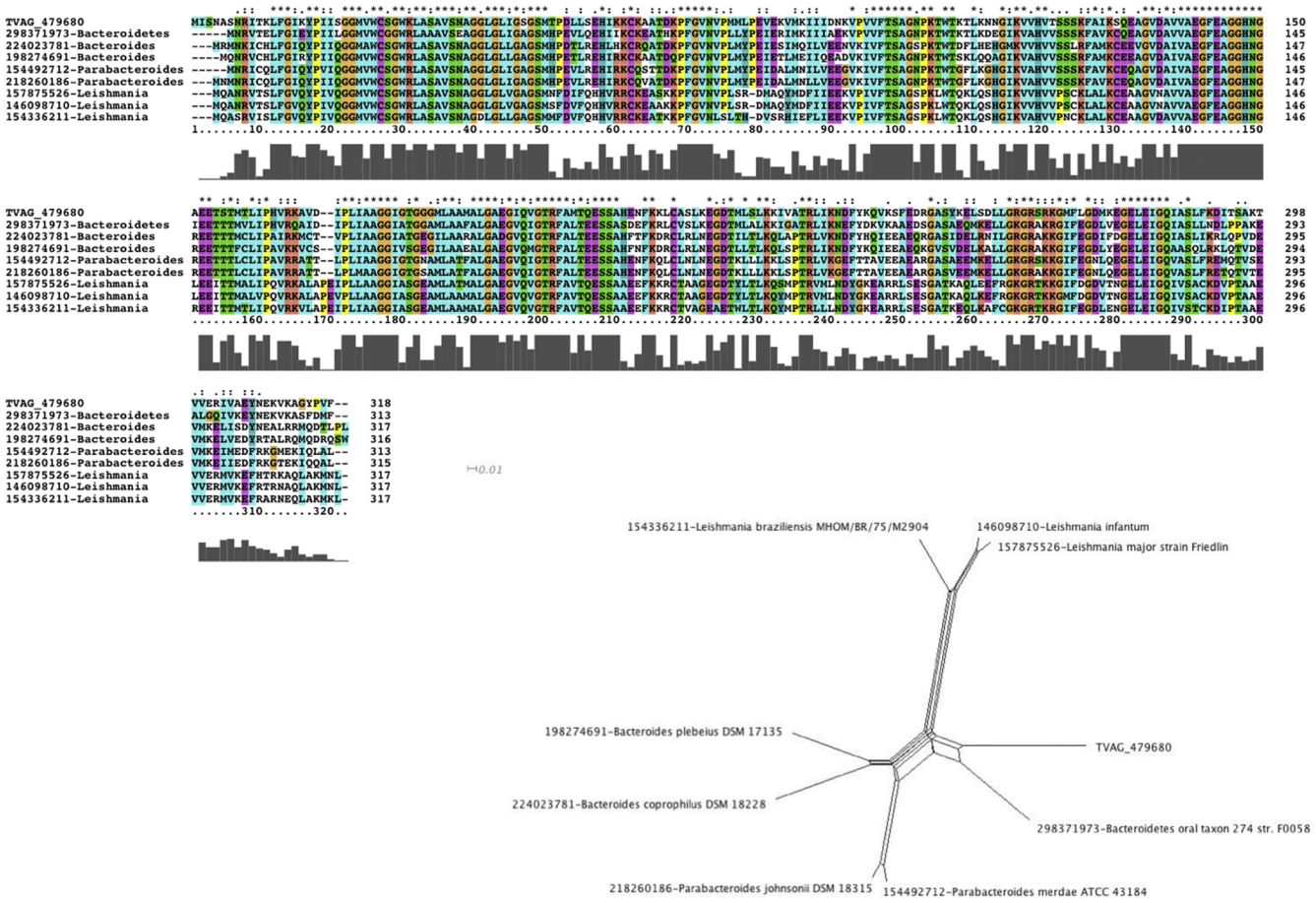

FIG 5 A multiple sequence alignment and phylogenetic network of TVAG_479680, a novel hydrogenosomal protein (annotated as 2-nitropropane dioxygenase), with its homologs.

and failed to explain the high import scores of several nonhydrogenosomal proteins (Table 3, TVAG_464170).

DISCUSSION

Trichomonas vaginalis encodes more than twice as many proteins as its human host, and instead of classical mitochondria it possesses hydrogenosomes. Like typical mitochondria, trichomonad hydrogenosomes synthesize ATP, but in contrast to mitochondria they must import all proteins from the cytosol as they lack a genome and translation machinery (11, 59). Many hydrogenosomal proteins are equipped with a short N-terminal hydrogenosomal targeting signal (HTS), which directs the preprotein to the organelle (11). Recently, though, the first proteins were identified that are apparently imported based on internal targeting signals (53). This could help to explain the discrepancy between the number of proteins estimated to be present in the hydrogenosome (about 500) (73) and those that harbor an HTS (about 220) (11).

In order to predict subcellular localization, we conducted a genome-wide screen for hydrogenosomal proteins using a set of machine learning classification algorithms. The algorithms do not depend solely on the presence of an HTS but include 57 features that measure various genomic, biochemical, and evolutionary traits of the proteins. Experimental validation revealed that 6 out of 10 proteins receiving a high import prediction score localized to the hydrogenosomes. As more hydrogenosomal proteins are discovered, the performance of the machine learning prediction will improve. When we included the six proteins that we localized *in vivo* (Table 3), the prediction score for those that failed to be imported dropped in four out of six cases. Furthermore, the prediction accuracy as calculated by the AUC and AUPR measures is higher in this final classification phase (Table 2).

Our total success rate for experimentally tested predictions was 71% (10/14). Assuming that 500 proteins are targeted to trichomonad hydrogenosomes, the probability of identifying one of the imported proteins by chance is 0.8371% (500/59,672). Although the success rate is 70-fold better than chance, still it is *de facto* much less accurate than the expected inference accuracy estimated by the AUC and AUPR measures (Table 2). According to these curves, for protein values with S_{import} values equal to or higher than the minimal S_{import} values for the 10 tested proteins, it is expected that 301 proteins having S_{import} values of >0.99 in the MOT+ scheme should be localized to the hydrogenosome. There could be several reasons for the discrepancy between the expected and observed prediction performance. For example, a learning set that includes a set of imported proteins whose properties differ significantly from those of nonimported proteins but also are much different from the properties of yet undiscovered hydrogenosomal proteins would lead to high accuracy and a low success rate. This is because the prediction accuracy is estimated as the

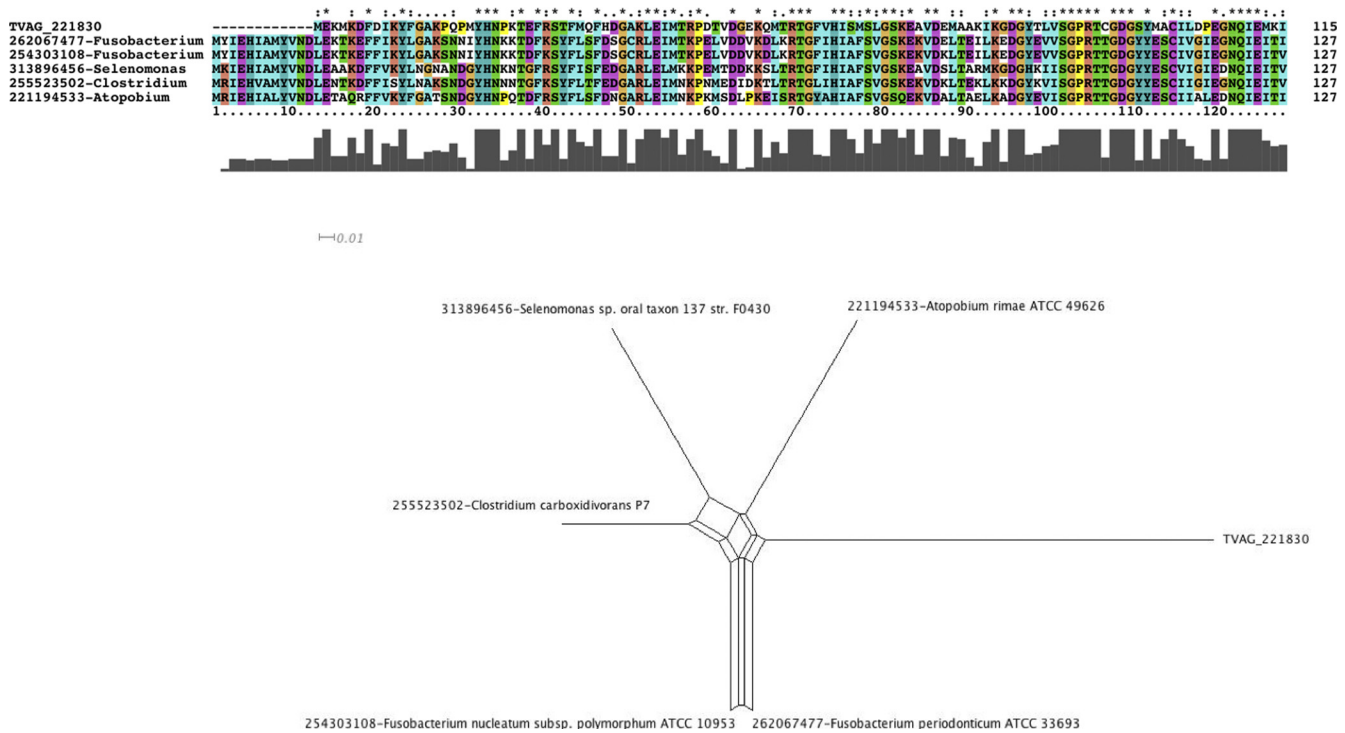

FIG 6 A multiple sequence alignment and phylogenetic network of TVAG_221830, a novel hydrogenosomal protein (containing a Glo-EDI-BRP-like domain), with its homologs.

ability of the classification algorithm to distinguish between imported and nonimported proteins, while for a high success rate we require a good distinction between the yet unknown imported proteins and nonimported proteins. Other possible reasons for the low success rate in our approach could be related to the vast amount of genes present in the *T. vaginalis* genome. Some proteins have dozens of highly similar duplicates for which similar protein characteristics are calculated, making the distinction between the rare imported proteins and the abundant cytosolic proteins very difficult. Moreover, it is possible that the 57 features we used are not those that best discriminate between imported and nonimported proteins; possibly other features such as structural information or yet to be discovered sequence signals would improve the prediction. The learning set is still very limited and biased toward proteins harboring the canonical HTS. As additional imported proteins are discovered, the performance of the machine learning approach is expected to further improve, as shown by our study.

The machine learning approach identified two HTS-lacking proteins that we also localize to the hydrogenosome (TVAG_479680 and TVAG_221830), and these are a putative 2-nitropropane dioxygenase and a protein of unknown function, respectively. These could not have been predicted as "hydrogenosomal" based on the presence of an HTS alone, and they are a representative example of the ability of our approach to identify hydrogenosomal proteins lacking an HTS. The localization experiments additionally confirmed the prediction of four new hydrogenosomal proteins. Taken together, our results suggest that the targeting information is not restricted to a motif such as the suggested HTS alone but might rather be a combination of factors including amino acid composition and protein conformation. This view is furthermore supported by the finding of the short PFOA targeting signal on proteins not targeted to the hydrogenosome. PFOA must contain internal information next to its short HTS assisting in hydrogenosomal targeting.

Streamlined import apparatuses exist in the mitosome-bearing protists *Giardia* and *Encephalitozoon* (50), and the same could be true for *T. vaginalis*. So far, only a few potential import components have been identified, which include TIM17/TIM23, TIM44, and PAM16/PAM18 of the inner membrane (73) and Hmp35 of the outer hydrogenosomal membrane (18). But other components might have been missed by a search based on sequence comparisons, due to the AT-rich genome of *T. vaginalis* altering the codon and amino acid usage and, additionally, the phylogenetic distance of *Trichomonas* from other characterized organisms.

A recent study by Rada and colleagues (65) analyzed the core components of the hydrogenosomal membranes using a proteome-based approach. To test their proteomic results, the authors verified the hydrogenosomal localization of 23 proteins using transfected cell lines. Within our prediction none of the Rada et al. protein set received high scores (see Table S1 in the supplemental material). The low scores are due to the very different characteristics of membrane proteins compared to those of soluble matrix proteins. Our initial training set included only two membrane proteins (Hmp31 and Hmp35) while the large majority represented soluble matrix proteins. If the situation in *Trichomonas* mirrors that in yeast, hydrogenosomal membrane proteins are most likely targeted and integrated into the membrane by a process different from that of the matrix proteins (42, 45, 81). The latter are either recognized by their N-terminal motifs or by an alternative internal signal that replaced the N-terminal motif (53), whereas the membrane proteins in yeast insert auton-

omously via a mechanism involving the Sam50 complex (42, 45, 81); in the hydrogenosomal membrane the mechanism could be similar. In either case, this will affect the prediction algorithm through the quality of the feature selection. From this observation we conclude that for future analyses one might need to train the algorithm on either matrix or membrane proteins separately and to balance the set of positives for the learning phase according to alternative import pathways.

Patterns of protein sequence similarity and phylogenetic reconstruction play an important role in hydrogenosomal targeting prediction using the machine learning approach. One of the strongest evolutionary features is the number of homologs in *Betaproteobacteria*. Furthermore, both of the HTS-lacking proteins that we have localized in the hydrogenosome here are eubacterial proteins (Fig. 5 and 6). One possibility for the origin of these proteins would be lateral gene acquisition from prokaryotic endosymbionts of human that share their habitat with *T. vaginalis* (2). However, because lateral gene transfer is a rare event among eukaryotes (69), a more tenable possibility would be that these proteins are vestiges of the common endosymbiotic origin of mitochondria and hydrogenosomes. Many proteins that are targeted into double membrane-bound organelles in eukaryotes (hydrogenosomes, mitochondria, mitosomes, and chloroplasts) are the products of genes that were transferred to the host nuclear genome during the course of endosymbiosis (43). Differential loss of genes from endosymbiotic origin and insufficient sampling density of sequenced eukaryotic genomes in the taxonomic neighborhood of *T. vaginalis* may lead to a phylogenetic signal that is similar to lateral gene acquisition. Indeed, the common ancestor of mitochondria and hydrogenosomes is assumed to have been an alphaproteobacterium (27); thus, the common expectation is that nuclear genes of mitochondrial origin would be more similar than their alphaproteobacterial homologs. However, owing to the substantial frequency of lateral gene transfer during prokaryote evolution, the alphaproteobacterial phylogenetic signal is scrambled over time (67), leading to a wider taxonomic distribution of eubacterial homologs with a tendency toward proteobacterial genes (23). Evidence for the role of the evolutionary component in hydrogenosomal targeting prediction is in line with the endosymbiotic origin of the organelle.

ACKNOWLEDGMENTS

The study was funded in part by the German Science Foundation (SFB-TR1). D.B. is a fellow of the Converging Technologies Program of the Israeli Council for Higher Education. T.P. is supported by an Israel Science Foundation (878/09) grant, Israel Ministry of Science and Technology Infrastructure grant, and by the National Evolutionary Synthesis Center, National Science Foundation EF-0905606.

We thank Katrin Henze and Gideon Dror for useful suggestions during the study.

REFERENCES

1. **Akhmanova A, et al.** 1998. A hydrogenosome with a genome. Nature **396**:527–528.
2. **Alsmark UC, Sicheritz-Ponten T, Foster PG, Hirt RP, Embley TM.** 2009. Horizontal gene transfer in eukaryotic parasites: a case study of *Entamoeba histolytica* and *Trichomonas vaginalis*. Methods Mol. Biol. **532**: 489–500.
3. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.
4. **Ashburner M, et al.** 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**:25–29.
5. **Aurrecoechea C, et al.** 2009. GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. Nucleic Acids Res. **37**:D526–530.
6. **Benchimol M.** 2008. The Hydrogenosome as a drug target. Curr. Pharm. Des. **14**:872–881.
7. **Boeckmann B, et al.** 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31**:365–370.
8. **Bozner P.** 1997. Immunological detection and subcellular localization of Hsp70 and Hsp60 homologs in *Trichomonas vaginalis*. J. Parasitol. **83**: 224–229.
9. **Bradley PJ, Lahti CJ, Plumper E, Johnson PJ.** 1997. Targeting and translocation of proteins into the hydrogenosome of the protist *Trichomonas*: similarities with mitochondrial protein import. EMBO J. **16**:3484–3493.
10. **Burges CJC.** 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. **2**:121–167.
11. **Carlton JM, et al.** 2007. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. Science **315**:207–212.
12. **Chacinska A, Koehler CM, Milenkovic D, Lithgow T, Pfanner N.** 2009. Importing mitochondrial proteins: machineries and mechanisms. Cell **138**:628–644.
13. **Cooper GF, Herskovits E.** 1991. A Bayesian method for constructing Bayesian belief networks from databases, p 86–94. *In* D'Ambrosio D, Smets P, and Bonissone P (ed), The seventh conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., Los Angeles, CA.
14. **Cooper GF, Herskovits E.** 1992. A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. **9**:309–347.
15. **Dechter R, Pearl J.** 1985. Generalized best-first search strategies and the optimality of A*. J. Assoc. Comput. Machinery **32**:505–536.
16. **Delgadillo MG, Liston DR, Niazi K, Johnson PJ.** 1997. Transient and selectable transformation of the parasitic protist *Trichomonas vaginalis*. Proc. Natl. Acad. Sci. U. S. A. **94**:4716–4720.
17. **Dyall SD, Johnson PJ.** 2000. Origins of hydrogenosomes and mitochondria: evolution and organelle biogenesis. Curr. Opin. Microbiol. **3**:404–411.
18. **Dyall SD, et al.** 2003. *Trichomonas vaginalis* Hmp35, a putative pore-forming hydrogenosomal membrane protein, can form a complex in yeast mitochondria. J. Biol. Chem. **278**:30548–30561.
19. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**:1792–1797.
20. **Embley TM, Martin W.** 2006. Eukaryotic evolution, changes and challenges. Nature **440**:623–630.
21. **Embley, TM, van der Giezen M, Horner DS, Dyal PL, and Foster P.** 2003. Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. Philos. Trans. R. Soc. Lond. B Biol. Sci. **358**:191–202.
22. **Enright AJ, Van Dongen S, Ouzounis CA.** 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. **30**:1575–1584.
23. **Esser C, Martin W, Dagan T.** 2007. The origin of mitochondria in light of a fluid prokaryotic chromosome model. Biol. Lett. **3**:180–184.
24. **Fawcett T.** 2006. An introduction to ROC analysis. Pattern Recognit. Lett. **27**:861–874.
25. **Felsenstein J.** 2005. PHYLIP, version 3.6: phylogeny inference package. University of Washington, Seattle, WA.
26. **Friedman N, Geiger D, Goldszmidt M.** 1997. Bayesian network classifiers. Mach. Learn. **29**:131–163.
27. **Gray MW, Burger G, Lang BF.** 1999. Mitochondrial evolution. Science **283**:1476–1481.
28. **Green D, Swets J.** 1966. Signal detection theory and psychophysics. Wiley, New York, NY.
29. **Hall M, et al.** 2009. The WEKA data mining software: an update. SIGKDD Explor. Newls. **11**:10–18.
30. **Heckerman D, Geiger D, Chickering DM.** 1995. Learning Bayesian networks: the combination of knowledge and statistical data. Mach. Learn. **20**:197–243.
31. **Henze K.** 2008. The Proteome of *T. vaginalis* Hydrogenosomes, p 163–178. *In* Tachezy J (ed), Hydrogenosomes and mitosomes: mitochondria of anaerobic eukaryotes. Springer, Berlin, Germany.
32. **Hjort K, Goldberg AV, Tsaousis AD, Hirt RP, Embley TM.** 2010. Diversity and reductive evolution of mitochondria among microbial eukaryotes. Philos. Trans. R. Soc. Lond. B Biol. Sci. **365**:713–727.
33. **Hochberg Y, Benjamini Y.** 1990. More powerful procedures for multiple significance testing. Stat. Med. **9**:811–818.
34. **Hrdý I, Cammack R, Stopka P, Kulda J, Tachezy J.** 2005. Alternative

pathway of metronidazole activation in *Trichomonas vaginalis* hydrogenosomes. Antimicrob. Agents Chemother. **49**:5033–5036.
35. **Hrdý I, et al.** 2004. *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature **432**:618–622.
36. **Hrdý I, Müller M.** 1995. Primary structure and eubacterial relationships of the pyruvate:ferredoxin oxidoreductase of the amitochondriate eukaryote *Trichomonas vaginalis.* J. Mol. Evol. **41**:388–396.
37. **Jansen R, et al.** 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science **302**:449–453.
38. **Jesse D, Mark G.** 2006. The relationship between precision-recall and ROC curves, p 233–240.In Proceedings of the 23rd International Conference on Machine Learning. Association for Computing Machinery, New York, NY.
39. **John GH, Kohavi R, Pfleger K.** 1994. Irrelevant features and the subset selection problem, p 121–129. *In* Proceedings of the 11th International Conference on Machine Learning. Morgan Kaufmann, San Mateo, CA.
40. **Jones DT.** 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. **292**:195–202.
41. **Jones DT, Taylor WR, Thornton JM.** 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–282.
42. **Kemper C, et al.** 2008. Integration of tail-anchored proteins into the mitochondrial outer membrane does not require any known import components. J. Cell Sci. **121**:1990–1998.
43. **Kleine T, Maier UG, Leister D.** 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annu. Rev. Plant Biol. **60**:115–138.
44. **Kohavi R, John GH.** 1997. Wrappers for feature subset selection. Artif. Intell. **97**:273–324.
45. **Kozjak V, et al.** 2003. An essential role of Sam50 in the protein sorting and assembly machinery of the mitochondrial outer membrane. J. Biol. Chem. **278**:48520–48523.
46. **Kulda J.** 1999. Trichomonads, hydrogenosomes and drug resistance. Int. J. Parasitol. **29**:199–212.
47. **Kyte J, Doolittle RF.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.
48. **Langley P, Iba W, Thompson K.** 1992. An analysis of Bayesian classifiers, p 223–228. *In* Swartout ED, Proceedings of the 10th National Conference on Artificial Intelligence. AAAI Press/MIT Press, San Jose, CA.
49. **Lindmark DG, Müller M.** 1973. Hydrogenosome, a cytoplasmic organelle of the anaerobic flagellate *Tritrichomonas foetus*, and its role in pyruvate metabolism. J. Biol. Chem. **248**:7724–7728.
50. **Lithgow T, Schneider A.** 2010. Evolution of macromolecular import pathways in mitochondria, hydrogenosomes and mitosomes. Philos. Trans. R. Soc. Lond. B Biol. Sci. **365**:799–817.
51. **Lu ZJ, et al.** 2011. Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome Res. **21**:276–285.
52. **Marchler-Bauer A, et al.** 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. **39**:D225–D229.
53. **Mentel M, Zimorski V, Haferkamp P, Martin W, Henze K.** 2008. Protein import into hydrogenosomes of *Trichomonas vaginalis* involves both N-terminal and internal targeting signals: a case study of thioredoxin reductases. Eukaryot. Cell **7**:1750–1757.
54. **Mertens E, Müller M.** 1990. Glucokinase and fructokinase of *Trichomonas vaginalis* and *Tritrichomonas foetus.* J. Protozool. **37**:384–388.
55. **Miller M, Liao Y, Gomez AM, Gaydos CA, D'Mellow D.** 2008. Factors associated with the prevalence and incidence of *Trichomonas vaginalis* infection among African American women in New York City who use drugs. J. Infect. Dis. **197**:503–509.
56. **Mokranjac D, et al.** 2009. Role of Tim50 in the transfer of precursor proteins from the outer to the inner membrane of mitochondria. Mol. Biol. Cell **20**:1400–1407.
57. **Morrison DF.** 1990. Multivariate statistical methods. McGraw-Hill, New York, NY.
58. **Müller M.** 1986. Reductive activation of nitroimidazoles in anaerobic microorganisms. Biochem. Pharmacol. **35**:37–41.
59. **Müller M.** 1993. The hydrogenosome. J. Gen. Microbiol. **139**:2879–2889.
60. **Needleman SB, Wunsch CD.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**:443–453.
61. **Perez-Brocal V, Clark CG.** 2008. Analysis of two genomes from the mitochondrion-like organelle of the intestinal parasite Blastocystis: complete sequences, gene content, and genome organization. Mol. Biol. Evol. **25**:2475–2482.
62. **Pruitt KD, Tatusova T, Maglott DR.** 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. **35**:D61–D65.
63. **Pütz S, et al.** 2006. Fe-hydrogenase maturases in the hydrogenosomes of *Trichomonas vaginalis.* Eukaryot. Cell **5**:579–586.
64. **Pütz S, Gelius-Dietrich G, Piotrowski M, Henze K.** 2005. Rubrerythrin and peroxiredoxin: two novel putative peroxidases in the hydrogenosomes of the microaerophilic protozoon *Trichomonas vaginalis.* Mol. Biochem. Parasitol. **142**:212–223.
65. **Rada P, et al.** 2011. The core components of organelle biogenesis and membrane transport in the hydrogenosomes of *Trichomonas vaginalis.* PLoS One **6**:e24428.
66. **Rice P, Longden I, Bleasby A.** 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. **16**:276–277.
67. **Richards TA, Archibald JM.** 2011. Cell evolution: gene transfer agents and the origin of mitochondria. Curr. Biol. **21**:R112–114.
68. **Saitou N, Nei M.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.
69. **Salzberg SL, White O, Peterson J, Eisen JA.** 2001. Microbial genes in the human genome: lateral transfer or gene loss? Science **292**:1903–1906.
70. **Schneider G, Fechner U.** 2004. Advances in the prediction of protein targeting signals. Proteomics **4**:1571–1580.
71. **Schneider RE.** 2009. Proteome analysis of the *Trichomonas vaginalis* hydrogenosome and putative import machinery. Doctor of Philosophy in microbiology, immunology, and molecular genetics. University of California. Los Angeles, CA.
72. **Schweikert G, et al.** 2009. mGene: accurate SVM-based gene finding with an application to nematode genomes. Genome Res. **19**:2133–2143.
73. **Shiflett AM, Johnson PJ.** 2010. Mitochondrion-related organelles in eukaryotic protists. Annu. Rev. Microbiol. **64**:409–429.
74. **Sutak R, et al.** 2004. Mitochondrial-type assembly of FeS centers in the hydrogenosomes of the amitochondriate eukaryote *Trichomonas vaginalis.* Proc. Natl. Acad. Sci. U. S. A. **101**:10368–10373.
75. **Tachezy J, Sanchez LB, Müller M.** 2001. Mitochondrial type iron-sulfur cluster assembly in the amitochondriate eukaryotes *Trichomonas vaginalis* and *Giardia intestinalis*, as indicated by the phylogeny of IscS. Mol. Biol. Evol. **18**:1919–1928.
76. **Tovar J, et al.** 2003. Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation. Nature **426**:172–176.
77. **Truscott KN, Brandner K, Pfanner N.** 2003. Mechanisms of protein import into mitochondria. Curr. Biol. **13**:R326–337.
78. **Upcroft P, Upcroft JA.** 2001. Drug targets and mechanisms of resistance in the anaerobic protozoa. Clin. Microbiol. Rev. **14**:150–164.
79. **van Grinsven KW, et al.** 2008. Acetate:succinate CoA-transferase in the hydrogenosomes of *Trichomonas vaginalis*: identification and characterization. J. Biol. Chem. **283**:1411–1418.
80. **Vapnik V.** 1999. The nature of statistical learning theory. Springer, New York, NY.
81. **Walther DM, Rapaport D.** 2009. Biogenesis of mitochondrial outer membrane proteins. Biochim. Biophys. Acta **1793**:42–51.
82. **Wawrzyniak I, et al.** 2008. Complete circular DNA in the mitochondria-like organelles of *Blastocystis hominis.* Int. J. Parasitol. **38**:1377–1382.
83. **Wright JM, Webb RI, O'Donoghue P, Upcroft P, Upcroft JA.** 2010. Hydrogenosomes of laboratory-induced metronidazole-resistant *Trichomonas vaginalis* lines are downsized while those from clinically metronidazole-resistant isolates are not. J. Eukaryot. Microbiol. **57**:171–176.