HydDB: A web tool for hydrogenase classification and analysis

Dan Søndergaard[a], Christian N. S. Pedersen[a], Chris Greening[c, d] *

[a] Aarhus University, Bioinformatics Research Centre, C.F. Møllers Allé 8, Aarhus DK-8000, Denmark

[b] Australian National University, Research School of Chemistry, Sullivans Creek Road, Acton, ACT 2601, Australia

[c] The Commonwealth Scientific and Industrial Research Organisation, Land and Water Flagship, Clunies Ross Street, Acton, ACT 2060, Australia

[d] Monash University, School of Biological Sciences, Clayton, VIC 2800, Australia

Correspondence:

Dr Chris Greening (chris.greening@monash.edu), Monash University, School of Biological Sciences, Clayton, VIC 2800, Australia

Dan Søndergaard (das@birc.au.dk), Aarhus University, Bioinformatics Research Centre, C.F. Møllers Allé 8, Aarhus DK-8000, Denmark

Short Title: HydDB, a hydrogenase database

Figures: Two colour figures, two tables

Abstract

H_2 metabolism is the most ancient and diverse mechanism of energy-generation. The metalloenzymes mediating this metabolism, hydrogenases, are encoded by over 60 microbial phyla and are present in all major ecosystems. We developed a classification system and web tool, HydDB, for the structural and functional analysis of these enzymes. We show that hydrogenase function can be predicted by primary sequence alone using an expanded classification scheme (comprising 29 [NiFe]-hydrogenase subgroups, 8 [FeFe]-hydrogenase subtypes, [Fe]-hydrogenases). Using this scheme, we built a web tool that rapidly and reliably classifies hydrogenase primary sequences using a combination of *k*-nearest neighbors' algorithms and CDD referencing. Demonstrating its capacity, the tool reliably predicted hydrogenase content and function in 12 newly-sequenced bacteria, archaea, and eukaryotes. HydDB also provides the capacity to browse 3248 annotated sequences and contains a detailed repository of physiological, biochemical, and structural information about the 38 hydrogenase classes defined here. The database and classifier are freely and publicly available at http://services.birc.au.dk/hyddb/

Introduction

Microorganisms conserve energy by metabolizing H_2. Oxidation of this high-energy fuel yields electrons that can be used for respiration and carbon-fixation. This diffusible gas is also produced in diverse fermentation and anaerobic respiratory processes [1]. H_2 metabolism contributes to the growth and survival of microorganisms across the three domains of life: chemotrophs and phototrophs, lithotrophs and

heterotrophs, aerobes and anaerobes, mesophiles and extremophiles alike [1,2]. On the ecosystem scale, H_2 supports microbial communities in most terrestrial, aquatic, and host-associated ecosystems [1,3]. It is also generally accepted that H_2 was the primordial electron donor [4]. In biological systems, metalloenzymes known as hydrogenases are responsible for oxidizing and evolving H_2 [1,5]. Our recent survey showed there is a far greater number and diversity of hydrogenases than previously thought [2]. It is predicted over 55 microbial phyla and up to half of all microorganisms harbor hydrogenases [2,6]. Better understanding H_2 metabolism and the enzymes that mediate it also has wider implications, particularly in relation to human health and disease [3,7], biogeochemical cycling [8], and renewable energy [9,10].

There are three classes of hydrogenase, the [NiFe], [FeFe], and [Fe] hydrogenases, that are distinguished by their metal composition. Whereas the [Fe]-hydrogenases are a small methanogenic-specific family [11], the [NiFe] and [FeFe] classes are widely distributed and functionally diverse. They comprise numerous different groups and subgroups/subtypes with distinct biochemical features (e.g. directionality, affinity, redox partners, and localization) and physiological roles (i.e. respiration, fermentation, bifurcation, sensing) [1,5]. For example, while Group 2a and 2b [NiFe]-hydrogenases share > 35% sequence identity, they have distinct roles as respiratory uptake hydrogenases and H_2 sensors respectively [12,13]. Building on previous work [14,15], we recently created a comprehensive hydrogenase classification scheme predictive of biological function [2]. This scheme was primarily based on amino acid sequence phylogeny, but also factored in genetic organization, metal-binding motifs, and functional information. This analysis identified 22 subgroups (within four groups)

of [NiFe]-hydrogenases and six subtypes (within three groups) of [FeFe]-hydrogenases, each with unique physiological roles [2].

In this work, we build on these findings to develop the first web database for the classification and analysis of hydrogenases. We developed an expanded classification scheme that captures the full sequence diversity of hydrogenase enzymes and predicts their biological function. Using this information, we developed a classification tool based on the *k*-nearest neighbors' (*k*-NN) method. This tool is a more reliable, efficient, and user-friendly method for hydrogenase classification than standard approaches involved phylogenetic tree construction, with a precision of more than 99.8%.

Results and Discussion

A sequence-based classification scheme for hydrogenases

We initially developed a classification scheme to enable prediction of hydrogenase function by primary sequence alone. To do this, we visualized the relationships between all hydrogenases in sequence similarity networks [16], in which nodes represent individual proteins and the distances between them reflect BLAST *E*-values. As reflected by our analysis of other protein superfamilies [17,18], SSNs allow robust inference of sequence-structure-function relationships for large datasets without the problems associated with phylogenetic trees (e.g. long-branch attraction). Consistent with previous phylogenetic analyses [2,14,15], this analysis showed the hydrogenase sequences clustered into eight major groups (Groups 1 to 4 [NiFe]-hydrogenases, Groups A to C [FeFe]-hydrogenases, [Fe]-hydrogenases), six of which separate into multiple functionally-distinct subgroups or subtypes at narrower

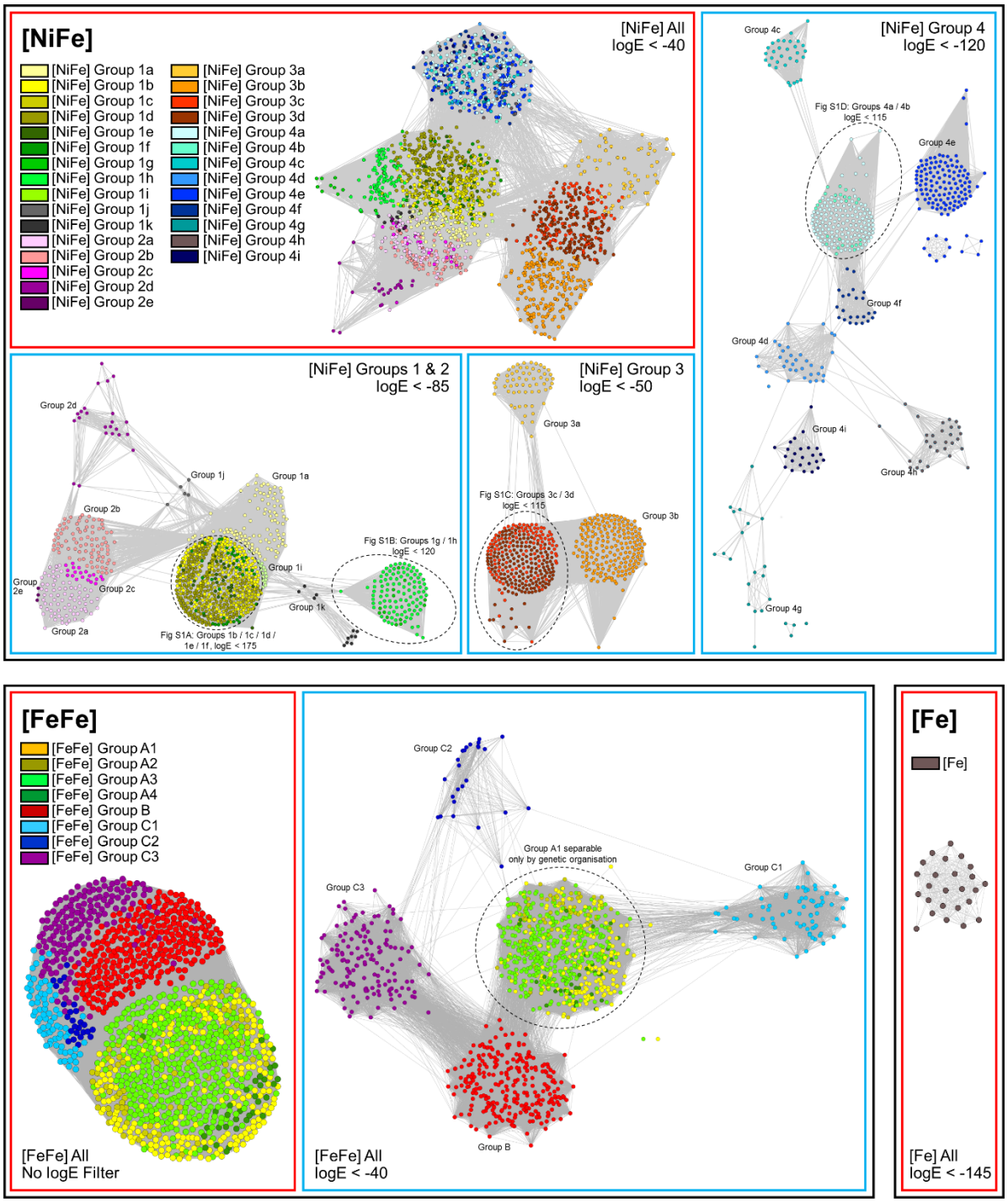


Figure 1. Sequence similarity network of hydrogenase sequences. Nodes represent individual proteins and the edges show the BLAST E-values between them at the logE filter defined at the bottom-left of each panel. The sequences are colored by class as defined in the legends. Figure S1 shows the further delineation of the encircled [NiFe] hydrogenase classes.

log*E* filters **(Figure 1; Figure S1)**. The SSNs demonstrated that all [NiFe]-hydrogenase subgroups defined through phylogenetic trees in our previous work [2] separated into distinct clusters, which is consistent with our evolutionary model that such hydrogenases diverged from a common ancestor to adopt multiple distinct functions [2]. The only exception were the Group A [FeFe]-hydrogenases, which as previously-reported [2,15], cannot be classified by sequence alone as they have principally diversified through changes in domain architecture and quaternary structure. It remains strictly necessary to analyze the organization of the genes encoding these enzymes to determine their specific function, e.g. whether they serve fermentative or electron-bifurcating roles.

The SSN analysis revealed that several groups and subgroups that clustered together in the phylogenetic tree analysis [2] separate into several subclades of probable distinct function **(Figure 1)**. On this basis, we refined and expanded the hydrogenase classification scheme to reflect the sequence diversification observed **(Table 1)**. Three lineages originally classified as Group 1a [NiFe]-hydrogenases were reclassified as new subgroups, the Coriobacteria (Group 1i), Archaeoglobi (Group 1j), and Methanosarcinales (Group 1i). The previously-defined 4b and 4d subgroups [2] were dissolved, as the SSN analysis confirmed they were highly polyphyletic. These sequences are reclassified here into five new subgroups: the formate- and carbon monoxide-respiring Mrp-linked complexes (Group 4b) [19], the ferredoxin-coupled Mrp-linked complexes (Group 4d) [20], the well-described methanogenic Eha (Group 4h) and Ehb (Group 4i) supercomplexes [21], and a more loosely clustered class of unknown function (Group 4g). Three crenarchaeotal hydrogenases were also classified as their own family (Group 2e); these enzymes

[NiFe] Group 1: Respiratory H_2-uptake [NiFe]-hydrogenases			
1a	Periplasmic	Electron input for sulfate, metal, and organohalide respiration. [NiFeSe] variants.	[2]
1b	Prototypical	Electron input for sulfate, fumarate, metal, and nitrate respiration.	[2]
1c	Hyb-type	Electron input for fumarate, nitrate, and sulfate respiration. Physiologically reversible.	[2]
1d	Oxygen-tolerant	Electron input for aerobic respiration and oxygen-tolerant anaerobic respiration.	[2]
1e	Isp-type	Electron input primarily for sulfur respiration. Physiologically reversible.	[2]
1f	Oxygen-protecting	Unresolved role. May liberate electrons to reduce reactive oxygen species.	[2]
1g	Crenarchaeota-type	Electron input primarily for sulfur respiration.	[2]
1h	Actinobacteria-type	Electron input for aerobic respiration. Scavenges electrons from atmospheric H_2.	[2,22]
1i	Coriobacteria-type (putative)	Undetermined role. May liberate electrons for anaerobic respiration.	This work
1j	Archaeoglobi-type	Electron input for sulfate respiration [23].	This work
1k	Methanophenazine-reducing	Electron input for methanogenic heterodisulfide respiration [24].	This work
[NiFe] Group 2: Alternative and sensory uptake [NiFe]-hydrogenases			
2a	Cyanobacteria-type	Electron input for aerobic respiration. Recycles H_2 produced by other cellular processes.	[14]
2b	Histidine kinase-linked	H_2 sensing. Activates two-component system controlling hydrogenase expression.	[14]
2c	Diguanylate cyclase-linked (putative)	Undetermined role. May sense H_2 and regulate processes through cyclic di-GMP production.	[2]
2d	Aquificae-type	Unresolved role. May generate reductant for carbon fixation or have a regulatory role.	[2]
2e	Metallosphaera-type (putative)	Undetermined role. May liberate electrons primarily for aerobic respiration [25].	This work
[NiFe] Group 3: Cofactor-coupled bidirectional [NiFe]-hydrogenases			
3a	F_{420}-coupled	Couples oxidation of H_2 to reduction of F_{420} during methanogenesis. Physiologically reversible. [NiFeSe] variants.	[14]
3b	NADP-coupled	Couples oxidation of NADPH to evolution of H_2. Physiologically reversible. May have sulfhydrogenase activity.	[14]
3c	Heterodisulfide reductase-linked	Bifurcates electrons from H_2 to heterodisulfide and Fd_{ox} in methanogens. [NiFeSe] variants.	[14]
3d	NAD-coupled	Interconverts electrons between H_2 and NAD depending on cellular redox state.	[14]
[NiFe] Group 4: Respiratory H_2-evolving [NiFe]-hydrogenases			
4a	Formate hydrogenlyase	Couples formate oxidation to fermentative H_2 evolution. May be H^+-translocating.	[2]
4b	Formate-respiring	Respires formate or carbon monoxide using H^+ as electron acceptor. Na^+-translocating via Mrp [19].	This work
4c	Carbon monoxide-respiring	Respires carbon monoxide using H^+ as electron acceptor. H^+-translocating.	[2]

4d	Ferredoxin-coupled, Mrp-linked	Couples Fd_{red} oxidation to H^+ reduction. Na^+-translocating via Mrp complex [20].	This work
4e	Ferredoxin-coupled, Ech-type	Couples Fd_{red} oxidation to H^+ reduction. Physiologically reversible via H^+/Na^+ translocation.	[2]
4f	Formate-coupled (putative)	Undetermined role. May couple formate oxidation to H_2 evolution and H^+ translocation.	[2]
4g	Ferredoxin-coupled (putative)	Undetermined role. May couple Fd_{red} oxidation to proton reduction and H^+/Na^+ translocation.	This work
4h	Ferredoxin-coupled, Eha-type	Couples Fd_{red} oxidation to H^+ reduction in anaplerotic processes. H^+/Na^+-translocating [21].	This work
4i	Ferredoxin-coupled, Ehb-type	Couples Fd_{red} oxidation to H^+ reduction in anabolic processes. H^+/Na^+-translocating [21].	This work
[FeFe] Hydrogenases			
A1	Prototypical	Couples ferredoxin oxidation to fermentative or photobiological H_2 evolution.	[2,15]
A2	Glutamate synthase-linked (putative)	Undetermined role. May couple H_2 oxidation to NAD reduction, generating reductant for glutamate synthase.	[2,15]
A3	Bifurcating	Reversibly bifurcates electrons from H_2 to NAD and Fd_{ox} in anaerobic bacteria.	[2,15]
A4	Formate dehydrogenase-linked	Couples formate oxidation to H_2 evolution. Some bifurcate electrons from H_2 to ferredoxin and NADP.	[2,15]
B	Colonic-type (putative)	Undetermined role. May couple Fd_{red} oxidation to fermentative H_2 evolution.	[15]
C1	Histidine kinase-linked (putative)	Undetermined role. May sense H_2 and regulate processes via histidine kinases [2].	This work
C2	Chemotactic (putative)	Undetermined role. May sense H_2 and regulate processes via methyl-accepting chemotaxis proteins [2].	This work
C3	Phosphatase-linked (putative)	Undetermined role. May sense H_2 and regulate processes via serine/threonine phosphatases [2].	This work
[Fe] Hydrogenases			
All	Methenyl-H_4MPT dehydrogenase	Reversibly couples H_2 oxidation to 5,10-methenyltetrahydromethanopterin reduction.	[14]

Table 1. Expanded classification scheme for hydrogenase enzymes. The majority of the classes were defined in previous work [2,14,15,22]. The [NiFe] Group 1i, 1j, 1j, 2e, 4d, 4g, 4h, and 4i enzymes and [FeFe] Groups C1, C2, and C3 enzymes were defined in this work based on their separation into distinct clusters in the SSN analysis **(Figure 1)**. HydDB contains detailed information on each of these classes, including their taxonomic distribution, genetic organization, biochemistry, and structures, as well a list of primary references.

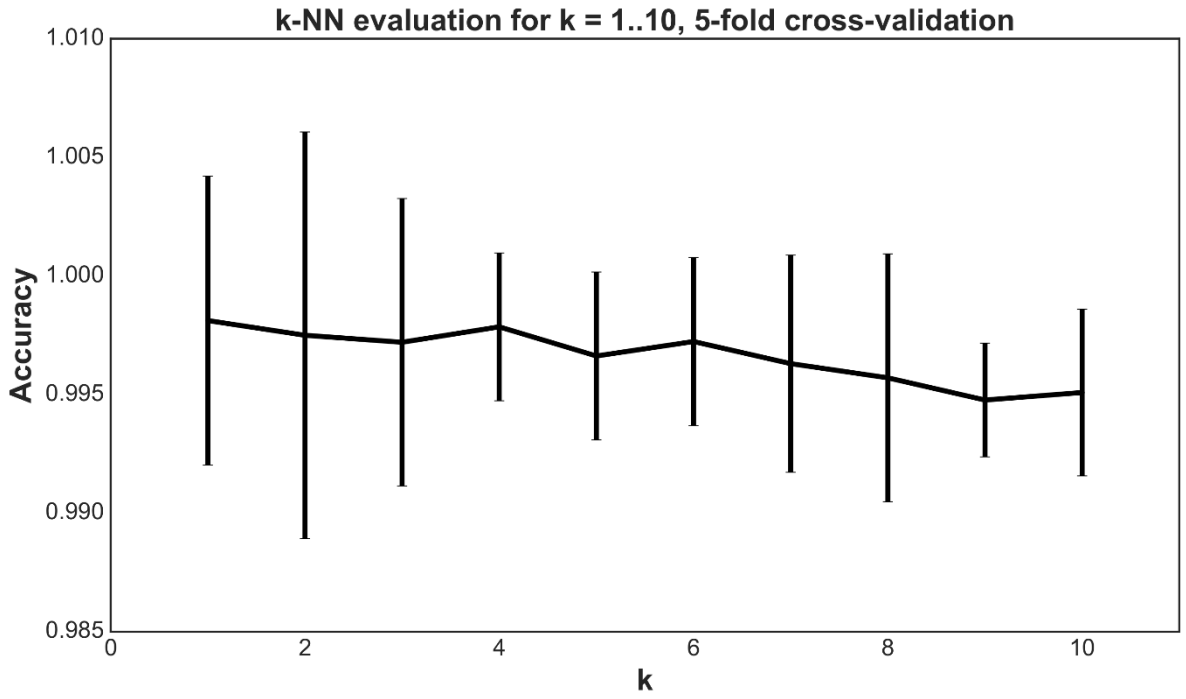
enable certain crenarchaeotes to grow aerobically on O_2 [25,26] and hence may represent a unique lineage of aerobic uptake hydrogenases currently underrepresented in genome databases. The Group C [FeFe]-hydrogenases were also separated into three main subtypes given they separate into distinct clusters even at relatively broad logE values **(Figure 1)**; these enzymes likely have a sensory role [2,15] and are each co-transcribed with different regulatory elements **(Table 1)**.

HydDB reliably predicts hydrogenase class using the *k*-NN method and CDD referencing

Using this information, we built a web tool to classify hydrogenases. Hydrogenase classification is determined through a two-step process following input of the catalytic subunit sequence. In the first, the Conserved Domain Database (CDD) [27] is referenced to confirm that the inputted sequence has a hydrogenase catalytic domain, i.e. “Complex1_49kDa superfamily” (cl21493) (for NiFe-hydrogenases), “Fe_hyd_lg_C superfamily” (cl14953) (for FeFe-hydrogenases), and “HMD” (pfam03201) (for Fe-hydrogenases). The sequence is subsequently classified through the *k*-NN method that determines the most similar sequences listed in the HydDB reference database. To determine the optimal *k* for the dataset, we performed a 5-fold cross-validation for $k = 1 \ldots 10$ and computed the accuracy for each *k*. The results are shown in **Figure 2**. The classifier predicted the classes of the 3248 hydrogenase sequences with 99.8% accuracy and high robustness when performing a 5-fold cross-validation (as described in the Methods section) for $k = 4$. The six sequences where there were discrepancies between the SSN and *k*-NN predictions are shown in **Table S1**. The classifier has also been trained to detect and

exclude protein families that are homologous to hydrogenases but do not metabolize H_2 (Nuo, Ehr, NARF, HmdII [1,2]) using reference sequences of these proteins.

Figure 2. Evaluation of the *k*-NN classifier for *k* = 1…10. For each *k*, a 5-fold cross-validation was performed. The mean accuracy ± two standard deviations of the folds is shown in the figure (note the y-axis). *k* = 1 provides the most accurate classifier. However, *k* = 4 provides almost the same accuracy and is more robust to errors in the training set (reflected by the lower standard deviation). In general, the standard deviation is very small, indicating that the predictions are robust to changes in the training data.



Sequences of the [FeFe] Group A can be classified into functionally-distinct subtypes (A1, A2, A3, A4) based on genetic organization [2]. The classifier can classify such hydrogenases if the protein sequence immediately downstream from the catalytic subunit sequence is provided. The classifier references the CDD to search for conserved domains in the downstream protein sequence. A sequence is classified

as [FeFe] Group A2 if one of the domains “GltA”, “GltD”, “glutamate synthase small subunit” or “putative oxidoreductase”, but not “NuoF”, is found in the sequence. Sequences are classified as [FeFe] Group A3 if the domain “NuoF” is found and [FeFe] Group A4 if the domain “HycB” is present. If none of the domains are found, the sequence is classified as A1. These classification rules were determined by collecting 69 downstream protein sequences. The sequences were then submitted to the CDD and the domains which most often occurred in each subtype were extracted.

In addition to its accuracy, the classifier is superior to other approaches due to its usability **(Figure S2)**. It is accessible as a free web service at http://services.birc.au.dk/hyddb/ HydDB allows the users to paste or upload sequences of hydrogenase catalytic subunit sequences in FASTA format and run the classification. When analysis has completed, results are presented in a table that can be downloaded as a CSV file. This provides an efficient and user-friendly way to classify hydrogenases, in contrast to the previous standard which requires visualization of multiple sequence alignments in phylogenetic trees [28].

HydDB infers the physiological roles of H_2 metabolism

As summarized in **Table 1**, hydrogenase class is strongly correlated with physiological role. As a result, the classifier is capable of predicting both the class and function of a sequenced hydrogenase. To demonstrate this capacity, we used HydDB to analyze the hydrogenases present in 12 newly-sequenced bacteria, archaea, and eukaryotes of major ecological significance. The classifier correctly classified all 24 hydrogenases identified in the sequenced genomes, as validated

198 with SSNs **(Table 2)**. On the basis of these classifications, the physiological roles of
199 H_2 metabolism were predicted **(Table 2)**. For five of the organisms, these predictions
200 are confirmed or supported by previously published data [26,29–32]. Other predictions
201 are in line with metabolic models derived from metagenome surveying [33–35]. In some
202 cases, the capacity for organisms to metabolize H_2 was not tested or inferred in
203 previous studies despite the presence of hydrogenases in the sequenced genomes
204 [30,36–38].
205
206 While HydDB serves as a reliable initial predictor of hydrogenase class and function,
207 further analysis is recommended to verify predictions. Hydrogenase sequences only
208 provide organism with the genetic capacity to metabolise H_2; their function is
209 ultimately modulated by their expression and integration within the cell [1,39]. In
210 addition, some classifications are likely to be overgeneralized due to lack of
211 functional and biochemical characterization of certain lineages and sublineages. For
212 example, it is not clear if two distant members of the Group 1h [NiFe]-hydrogenases
213 (*Robiginitalea biformata*, *Sulfolobus islandicus*) perform the same H_2-scavenging
214 functions as the core group [8]. Likewise, it seems probable that the Group 3a [NiFe]-
215 hydrogenases of *Thermococci* and *Aquificae* use a distinct electron donor to the
216 main class [40]. Prominent cautions are included in the enzyme pages in cases such
217 as these. HydDB will be updated when literature is published that influences
218 functional assignments.
219
220 **HydDB contains interfaces for hydrogenase browsing and analyzing**
221 In addition to its classification function, HydDB is designed to be a definitive
222 repository for hydrogenase retrieval and analysis. The database presently contains

Organism	Phylum	Hydrogenase accession no.	HydDB classification	SSN classification	Predicted H_2 metabolism	Confirmed H_2 metabolism
Pyrinomonas methylaliphatogenes	Acidobacteria	WP_041979300.1	[NiFe] Group 1h	[NiFe] Group 1h	Persistence by aerobic respiration of atmospheric H_2	Confirmed experimentally [29]
Phaeodactylibacter xiamenensis	Bacteroidetes	WP_044227713.1 WP_044216927.1 WP_044227053.1	[NiFe] Group 1d [NiFe] Group 2a [NiFe] Group 3d	[NiFe] Group 1d [NiFe] Group 2a [NiFe] Group 3d	Chemolithoautotrophic growth by aerobic H_2 oxidation	Bacterium grows aerobically, but H_2 oxidation untested [30]
Bathyarchaeota archaeon BA1	Bathyarchaeota	KPV62434.1 KPV62673.1 KPV62298.1	[NiFe] Group 3c [NiFe] Group 3c [NiFe] Group 4g	[NiFe] Group 3c [NiFe] Group 3c [NiFe] Group 4g	Couples Fd_{red} oxidation to H_2 evolution in energy-conserving and bifurcating processes	Unconfirmed but consistent with metagenome-based models [34]
Lenisia limosa	Proteobacteria (Epsilon class)	LenisMan28	[FeFe] Group A1	[FeFe] Group A	Fermentative evolution of H_2	Confirmed experimentally [32]
Acidianus copahuensis	Crenarchaeota	WP_048100721.1 WP_048100713.1 WP_048100378.1 WP_048100359.1	[NiFe] Group 1g [NiFe] Group 1g [NiFe] Group 1h [NiFe] Group 2e	[NiFe] Group 1g [NiFe] Group 1g [NiFe] Group 1h [NiFe] Group 2e	Chemolithoautotrophic growth by H_2 oxidation using O_2 or S_0 as electron acceptors	Partially confirmed experimentally [26]
Arcobacter sp. E1/2/3	Proteobacteria (Epsilon class)	Arc.peg.2312	[NiFe] Group 1b	[NiFe] Group 1b	Chemolithoautotrophic growth by anaerobic H_2 oxidation	Confirmed experimentally [32]
Methanoperedens nitroreducens	Euryarchaeota (ANME)	WP_048088262.1 WP_048090768.1	[NiFe] Group 3b [NiFe] Group 3b	[NiFe] Group 3b [NiFe] Group 3b	Secondary role for H_2 metabolism limited to fermentative evolution of H_2	Unconfirmed but consistent with metagenome-based models [33]
Kryptonium thompsoni	Kryptonia	CUU03002.1 CUU06124.1	[NiFe] Group 1d [NiFe] Group 3b	[NiFe] Group 1d [NiFe] Group 3b	Chemolithoautotrophic growth by aerobic H_2 oxidation, fermentative evolution of H_2.	Untested, candidate phylum identified by metagenomics [37]
Lokiarchaeum sp. GC14_75	Lokiarchaeota	KKK40681.1	[NiFe] Group 3c	[NiFe] Group 3c	Bifurcates electrons between H_2, heterodisulfide, and ferredoxin	Unconfirmed but consistent with metagenome-based models [41]

Nitrospira moscoviensis	Nitrospirae	WP_053379275.1	[NiFe] Group 2a	[NiFe] Group 2a	Chemolithoautotrophic growth by aerobic H_2 oxidation	Confirmed experimentally [31]
Bacterium GW2011_GWE1_35_17	Moranbacteria	KKQ46070.1 KKQ45273.1	[NiFe] Group 1a [NiFe] Group 3b	[NiFe] Group 1a [NiFe] Group 3b	Chemolithoautotrophic growth by anaerobic H_2 oxidation, fermentative evolution of H_2.	Unconfirmed but consistent with metagenome-based models [35]
Bacterium GW2011_GWA2_33_10	Peregrinibacteria	KKP36897.1	[FeFe] Group A3	[FeFe] Group A	Bifurcates electrons between H_2, NADH, and ferredoxin	Unconfirmed but consistent with metagenome-based models [35]
Entotheonella sp. TSY1	Tectomicrobia	ETW97737.1 ETW94065.1	[NiFe] Group 1h [NiFe] Group 3b	[NiFe] Group 1h [NiFe] Group 3b	Persistence by aerobic respiration of atmospheric H_2, fermentative evolution of H_2	Untested, candidate phylum identified by metagenomics [38]

Table 2. Predictive capacity of the HydDB. HydDB accurately determined hydrogenase content and predicted the physiological roles of H_2 metabolism in 12 newly-sequenced archaeal and bacterial species.

entries for 3248 hydrogenases, including their NCBI accession numbers, amino acid sequence, hydrogenase class, taxonomic affiliation, and predicted behavior **(Figure S2)**. To enable easy exploration of the data set, the database also provides access to an interface for searching, filtering, and sorting the data, as well as the capacity to download the results in CSV or FASTA format. There are individual pages for the 38 hydrogenase classes defined here **(Table 1)**, including descriptions of their physiological role, genetic organization, taxonomic distribution, and biochemical features. This is supplemented with a compendium of structural information about the hydrogenases, which is integrated with the Protein Databank (PDB), as well as a library of over 1000 literature references **(Figure S5)**.

Conclusions

To summarize, HydDB is a definitive resource for hydrogenase classification and analysis. The classifier described here provides a reliable, efficient, and convenient tool for hydrogenase classification and functional prediction. HydDB also provides browsing tools for the rapid analysis and retrieval of hydrogenase sequences. Finally, the manually-curated repository of class descriptions, hydrogenase structures, and literature references provide a deep but accessible resource for understanding hydrogenases.

Materials and Methods

Sequence datasets

The database was constructed using the amino acid sequences of all curated non-redundant 3248 hydrogenase catalytic subunits represented in the NCBI RefSeq

database in August 2014 [2] **(Dataset S1)**. In order to test the classification tool, additional sequences from newly-sequenced archaeal and bacteria phyla were retrieved from the Joint Genome Institute's Integrated Microbial Genomes database [42].

Sequence similarity networks

Sequence similarity networks (SSNs) [16] were used to visualize the distribution and diversity of the 3248 retrieved hydrogenase sequences. In this analysis, nodes represent individual proteins and edges represent the all-versus-all BLAST *E*-values. Three networks were constructed using Cytoscape, namely for the [NiFe]-hydrogenase large subunit sequences, [FeFe]-hydrogenase catalytic domain sequences, and [Fe]-hydrogenase sequences. The relationships between them were viewed at different log*E* cutoffs using different subsets of sequences.

Classification method

The k-NN method is a well-known machine learning method for classification [43]. Given a set of data points $x_1, x_2, \ldots x_N$ (e.g. sequences) with known labels $y_1, y_2, \ldots, y_N$ (e.g. type annotations), the label of a point, x, is predicted by computing the distance from x to $x_1, x_2, \ldots x_N$ and extracting the k labeled points closest to x, i.e. the neighbors. The predicted label is then determined by majority vote of the labels of the neighbors. The distance measure applied here is that of a BLAST search. Thus, the classifier corresponds to a homology search where the types of the top k results are considered. However, formulating the classification method as a machine learning problem allows the use of common evaluation methods to estimate the accuracy of the method and perform model selection. The classifier was evaluated

using k-fold cross-validation. The dataset is first split in to k parts of equal size. $k-1$ parts (the *training set*) are then used for training the classifier and the labels of the data points in the remaining part (the *test set*) are then predicted. This process, called a *fold*, is repeated k times. The predicted labels of each fold are then compared to the known labels and an accuracy can be computed.

Acknowledgements

We thank A/Prof Colin J. Jackson, Dr Hafna Ahmed, Dr Andrew Warden, and Dr Stephen Pearce for their helpful advice and comments regarding this manuscript. This work was supported by a PUMPkin Centre of Excellence PhD Scholarship awarded to DS, an Australian National University PhD Scholarship awarded to FHA, and a CSIRO Office of the Chief Executive Postdoctoral Fellowship awarded to CG.

Author Contributions

CG and DS designed experiments. DS and CG performed experiments. CG, DS, and CNSP analysed data. CNSP supervised students. CG and DS wrote the paper.

The authors declare no conflict of interest.

References

1. Schwartz, E., Fritsch, J. & Friedrich, B. *H2-metabolizing prokaryotes*. (Springer Berlin Heidelberg, 2013).

2. Greening, C. *et al.* Genome and metagenome surveys of hydrogenase diversity indicate H2 is a widely-utilised energy source for microbial growth and survival. *ISME J.* **10,** 761–777 (2016).

3. Cook, G. M., Greening, C., Hards, K. & Berney, M. in *Advances in Bacterial Pathogen Biology* (ed. Poole, R. K.) **65,** 1–62 (Academic Press, 2014).

4. Lane, N., Allen, J. F. & Martin, W. How did LUCA make a living? Chemiosmosis in the origin of life. *BioEssays* **32,** 271–280 (2010).

5. Lubitz, W., Ogata, H., Rüdiger, O. & Reijerse, E. Hydrogenases. *Chem. Rev.* **114,** 4081–148 (2014).

6. Peters, J. W. *et al.* [FeFe]- and [NiFe]-hydrogenase diversity, mechanism, and maturation. *Biochim. Biophys. Acta - Mol. Cell Res.* (2014). doi:10.1016/j.bbamcr.2014.11.021

7. Carbonero, F., Benefiel, A. C. & Gaskins, H. R. Contributions of the microbial hydrogen economy to colonic homeostasis. *Nat Rev Gastroenterol Hepatol* **9,** 504–518 (2012).

8. Greening, C. *et al.* Atmospheric hydrogen scavenging: from enzymes to ecosystems. *Appl. Environ. Microbiol.* **81,** 1190–1199 (2015).

9. Levin, D. B., Pitt, L. & Love, M. Biohydrogen production: prospects and limitations to practical application. *Int. J. Hydrogen Energy* **29,** 173–185 (2004).

10. Cracknell, J. A., Vincent, K. A. & Armstrong, F. A. Enzymes as working or inspirational catalysts for fuel cells and electrolysis. *Chem. Rev.* **108,** 2439–2461 (2008).

11. Shima, S. *et al.* The crystal structure of [Fe]-Hydrogenase reveals the geometry of the active site. *Science* **321,** 572–575 (2008).

12. Lenz, O. & Friedrich, B. A novel multicomponent regulatory system mediates H_2 sensing in *Alcaligenes eutrophus*. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 12474–12479 (1998).

13. Greening, C., Berney, M., Hards, K., Cook, G. M. & Conrad, R. A soil actinobacterium scavenges atmospheric H2 using two membrane-associated, oxygen-dependent [NiFe] hydrogenases. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 4257–4261 (2014).

14. Vignais, P. M., Billoud, B. & Meyer, J. Classification and phylogeny of hydrogenases. *FEMS Microbiol. Rev.* **25,** 455–501 (2001).

15. Calusinska, M., Happe, T., Joris, B. & Wilmotte, A. The surprising diversity of clostridial hydrogenases: a comparative genomic perspective. *Microbiology* **156,** 1575–1588 (2010).

16. Atkinson, H. J., Morris, J. H., Ferrin, T. E. & Babbitt, P. C. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* **4,** e4345 (2009).

17. Ahmed, F. H. *et al.* Sequence-structure-function classification of a catalytically diverse oxidoreductase superfamily in mycobacteria. *J. Mol. Biol.* **427,** 3554–3571 (2015).

18. Ney, B. *et al.* The methanogenic redox cofactor F_{420} is widely synthesized by aerobic soil bacteria. *ISME J.* In press (2016).

19. Kim, Y. J. *et al.* Formate-driven growth coupled with H_2 production. *Nature* **467,** 352–5 (2010).

20. McTernan, P. M. *et al.* Intact functional fourteen-subunit respiratory membrane-bound [NiFe]-hydrogenase complex of the hyperthermophilic archaeon *Pyrococcus furiosus*. *J. Biol. Chem.* **289,** 19364–19372 (2014).

21. Lie, T. J. *et al.* Essential anaplerotic role for the energy-converting hydrogenase Eha in hydrogenotrophic methanogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 15473–8 (2012).

22. Constant, P., Chowdhury, S. P., Pratscher, J. & Conrad, R. Streptomycetes contributing to atmospheric molecular hydrogen soil uptake are widespread and encode a putative high-affinity [NiFe]-hydrogenase. *Environ. Microbiol.* **12,** 821–829 (2010).

23. Stetter, K. O. Archaeoglobus fulgidus gen. nov., sp. nov.: a new taxon of extremely thermophilic archaebacteria. *Syst. Appl. Microbiol.* **10,** 172–173 (1988).

24. Deppenmeier, U. & Blaut, M. Analysis of the vhoGAC and vhtGAC operons from Methanosarcina mazei strain Gö1, both encoding a membrane-bound hydrogenase and a cytochrome b. *Eur. J. Biochem.* **269,** 261–269 (1995).

25. Auernik, K. S. & Kelly, R. M. Physiological versatility of the extremely thermoacidophilic archaeon *Metallosphaera sedula* supported by transcriptomic analysis of heterotrophic, autotrophic, and mixotrophic growth. *Appl. Environ. Microbiol.* **76,** 931–935 (2010).

26. Giaveno, M. A., Urbieta, M. S., Ulloa, J. R., González Toril, E. & Donati, E. R. Physiologic versatility and growth flexibility as the Main characteristics of a novel thermoacidophilic *Acidianus* strain isolated from Copahue geothermal area in Argentina. *Microb. Ecol.* **65,** 336–

346 (2012).

27. Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* **32,** W327–31 (2004).

28. Berney, M., Greening, C., Hards, K., Collins, D. & Cook, G. M. Three different [NiFe] hydrogenases confer metabolic flexibility in the obligate aerobe *Mycobacterium smegmatis*. *Environ. Microbiol.* **16,** 318–330 (2014).

29. Greening, C. *et al.* Persistence of the dominant soil phylum *Acidobacteria* by trace gas scavenging. *Proc. Natl. Acad. Sci.* **112,** 10497–10502 (2015).

30. Chen, Z. *et al. Phaeodactylibacter xiamenensis* gen. nov., sp. nov., a member of the family *Saprospiraceae* isolated from the marine alga Phaeodactylum tricornutum. *Int. J. Syst. Evol. Microbiol.* **64,** 3496–3502 (2014).

31. Koch, H. *et al.* Growth of nitrite-oxidizing bacteria by aerobic hydrogen oxidation. *Science* **345,** 1052–1054 (2014).

32. Hamann, E. *et al.* Environmental *Breviatea* harbour mutualistic *Arcobacter* epibionts. *Nature* **534,** 254–258 (2016).

33. Haroon, M. F. *et al.* Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature* **500,** 567–70 (2013).

34. Evans, P. N. *et al.* Methane metabolism in the archaeal phylum *Bathyarchaeota* revealed by genome-centric metagenomics. *Science* **350,** 434–438 (2015).

35. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain *Bacteria*. *Nature* **523,** 208–211 (2015).

36. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521,** 173–179 (2015).

37. Eloe-Fadrosh, E. A. *et al.* Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat Commun* **7,** (2016).

38. Wilson, M. C. *et al.* An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506,** 58–62 (2014).

39. Greening, C. & Cook, G. M. Integration of hydrogenase expression and hydrogen sensing in bacterial cell physiology. *Curr. Opin. Microbiol.* **18,** 30–8 (2014).

40. Greening, C. *et al.* Physiology, biochemistry, and applications of F_{420}- and F_o-dependent redox reactions. *Microbiol. Mol. Biol. Rev.* **80,** 451–493 (2016).

41. Sousa, F. L., Neukirchen, S., Allen, J. F., Lane, N. & Martin, W. F. Lokiarchaeon is hydrogen dependent. *Nat. Microbiol.* **1,** 16034 (2016).

42. Markowitz, V. M. *et al.* IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research* **40,** D115–22 (2012).

43. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13,** (1967).

Supporting Information

Figure S1. Sequence similarity networks showing the relationships between closely related subgroups of [NiFe]-hydrogenases as narrow logE filters.

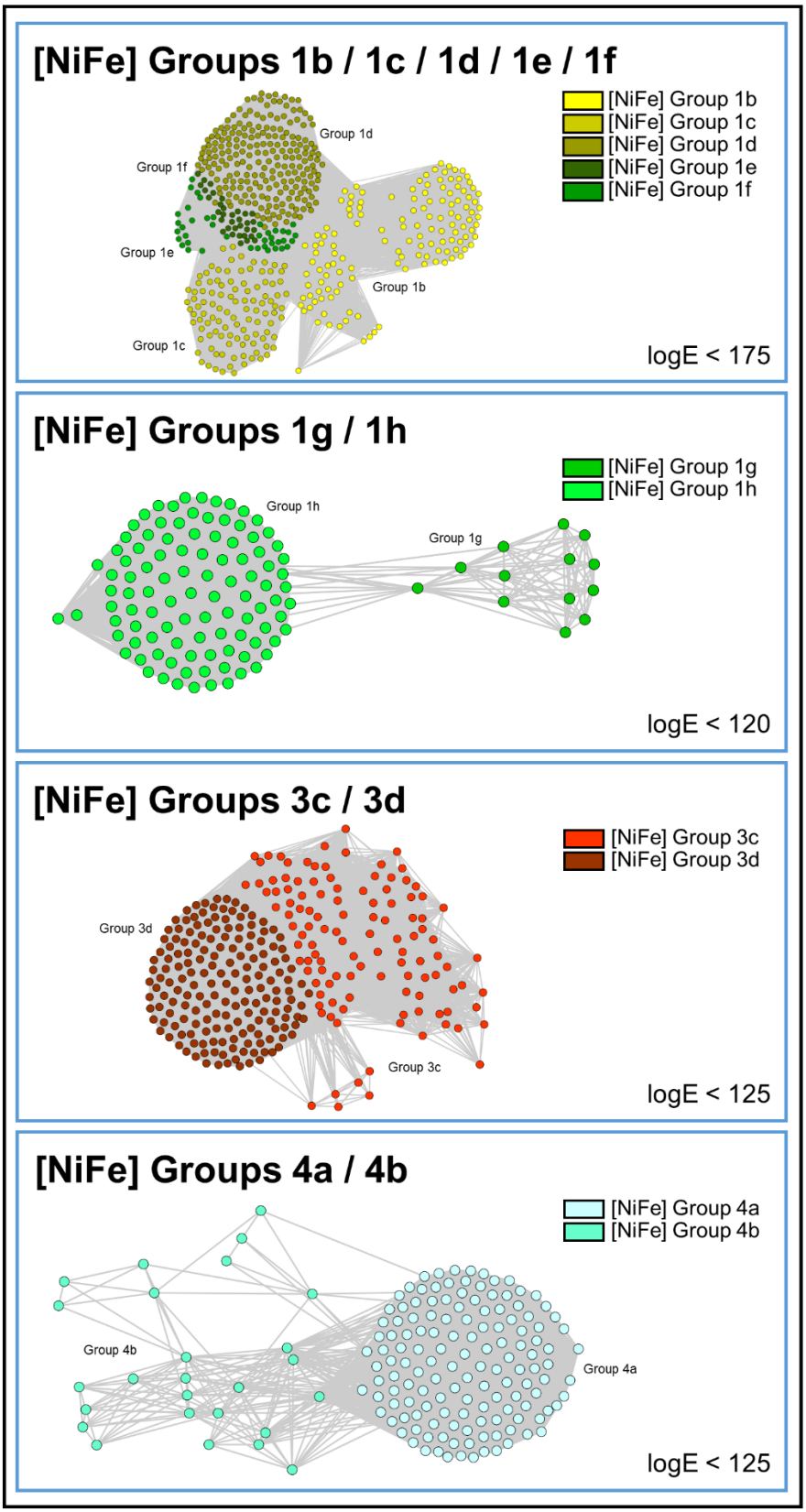


Figure S2. Screenshot showing interface of HydDB classification page.

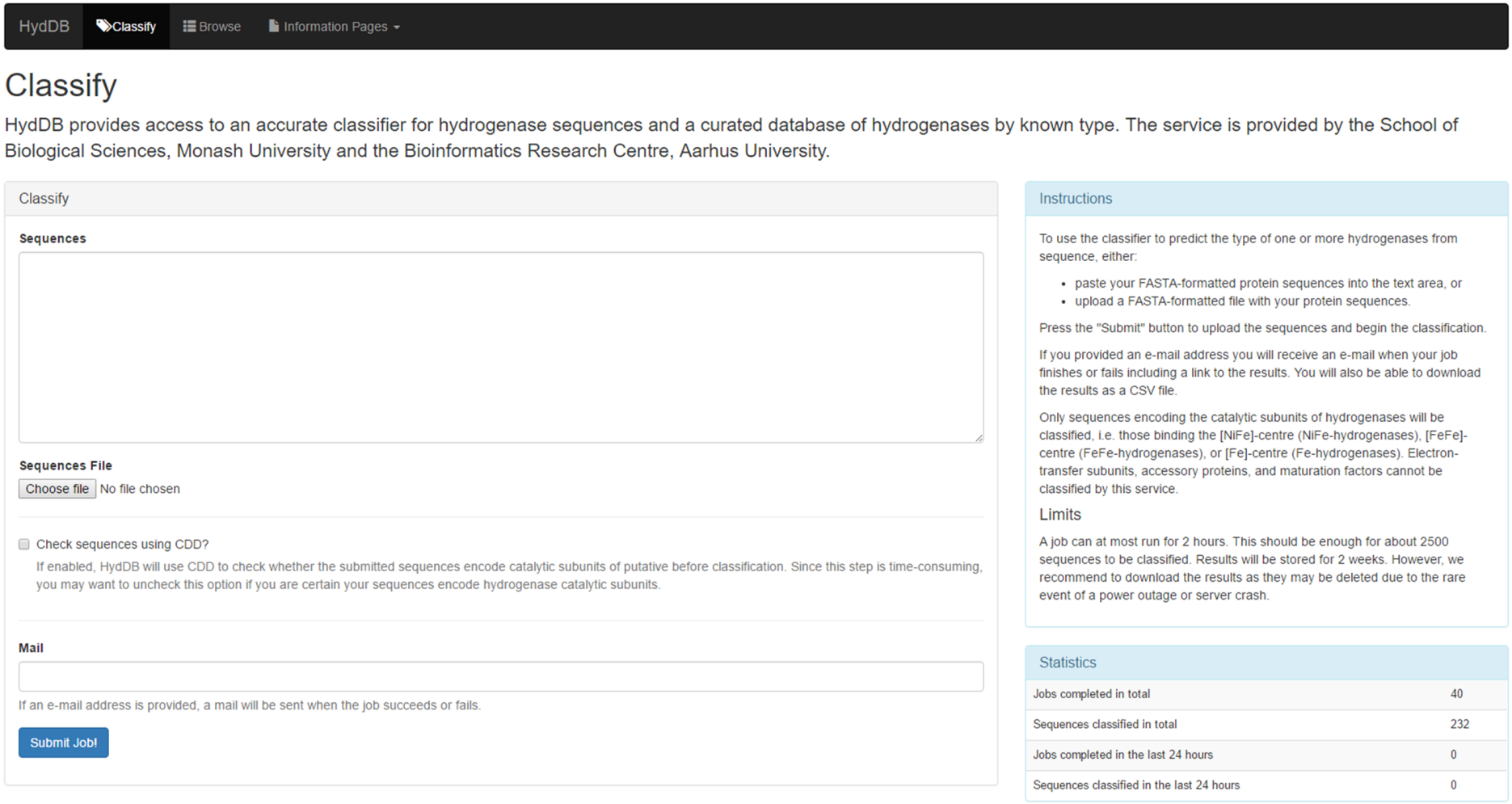


Figure S3. Screenshot showing the information provided in the data entry pages for 3248 individual hydrogenases in HydDB.

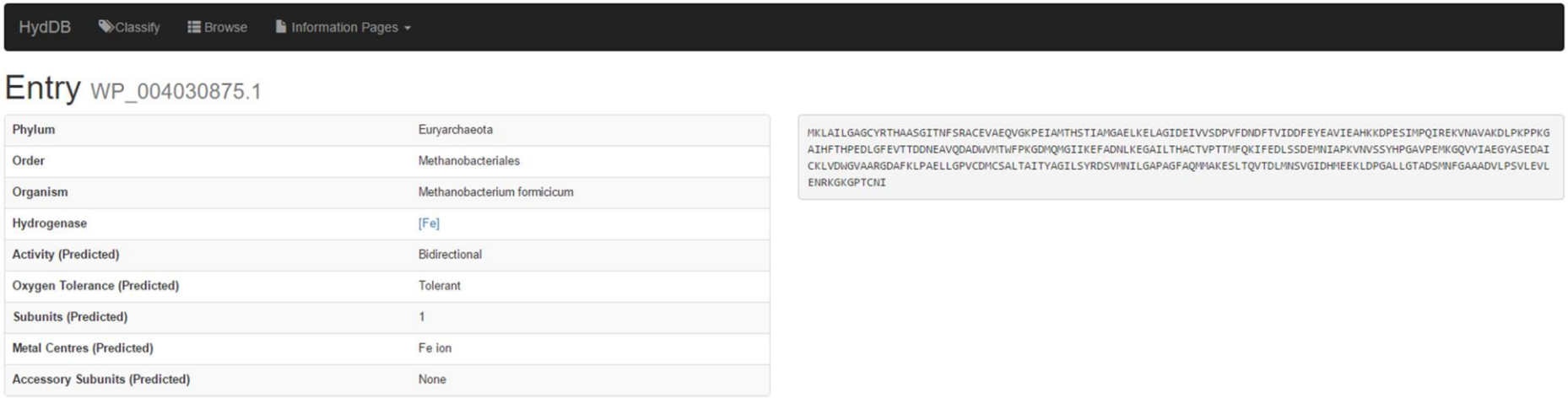
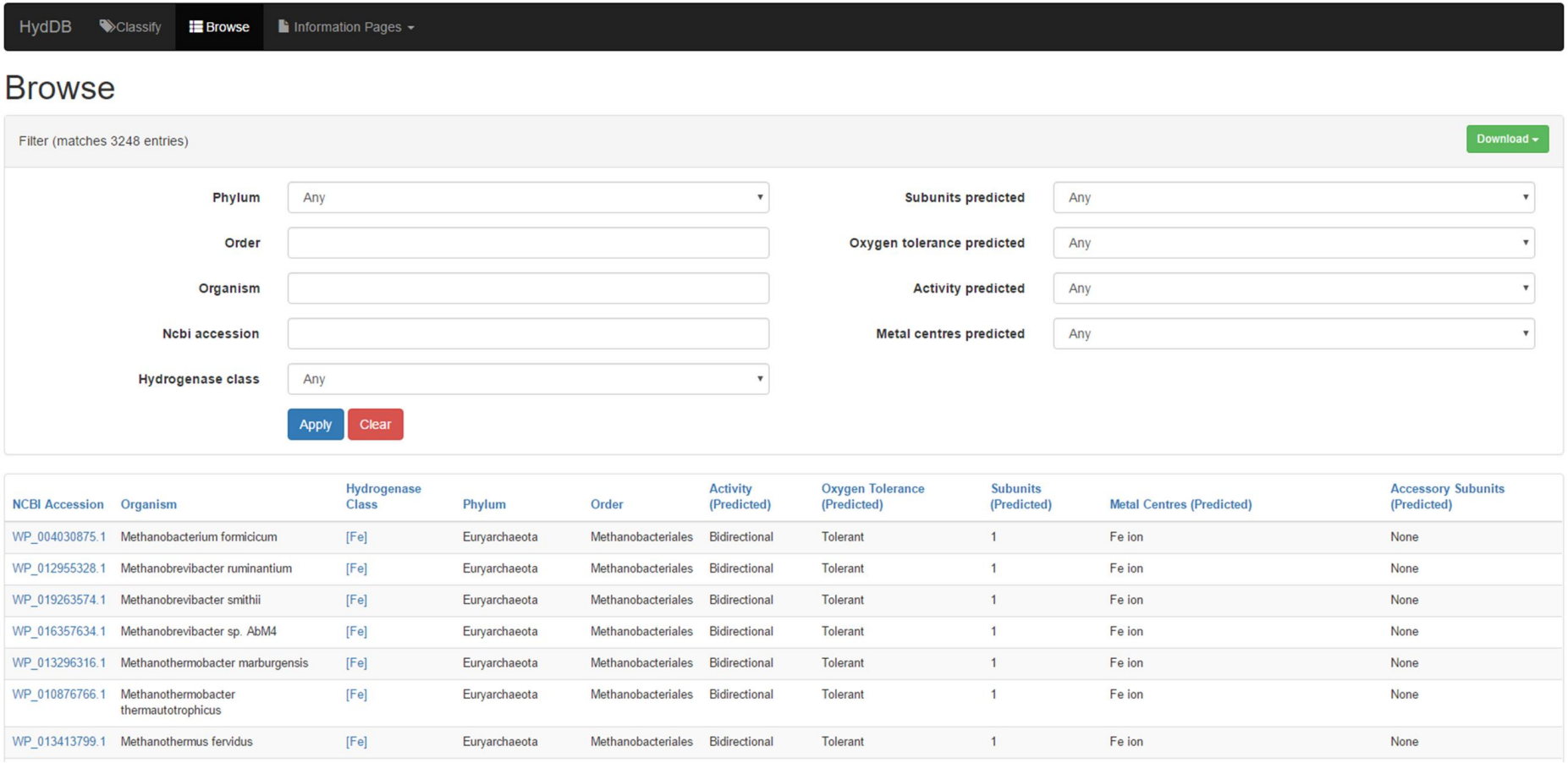


Figure S4. Screenshot showing the capacity for browsing hydrogenase data entries in HydDB.



NCBI Accession	Organism	Hydrogenase Class	Phylum	Order	Activity (Predicted)	Oxygen Tolerance (Predicted)	Subunits (Predicted)	Metal Centres (Predicted)	Accessory Subunits (Predicted)
WP_004030875.1	Methanobacterium formicicum	[Fe]	Euryarchaeota	Methanobacteriales	Bidirectional	Tolerant	1	Fe ion	None
WP_012955328.1	Methanobrevibacter ruminantium	[Fe]	Euryarchaeota	Methanobacteriales	Bidirectional	Tolerant	1	Fe ion	None
WP_019263574.1	Methanobrevibacter smithii	[Fe]	Euryarchaeota	Methanobacteriales	Bidirectional	Tolerant	1	Fe ion	None
WP_016357634.1	Methanobrevibacter sp. AbM4	[Fe]	Euryarchaeota	Methanobacteriales	Bidirectional	Tolerant	1	Fe ion	None
WP_013296316.1	Methanothermobacter marburgensis	[Fe]	Euryarchaeota	Methanobacteriales	Bidirectional	Tolerant	1	Fe ion	None
WP_010876766.1	Methanothermobacter thermautotrophicus	[Fe]	Euryarchaeota	Methanobacteriales	Bidirectional	Tolerant	1	Fe ion	None
WP_013413799.1	Methanothermus fervidus	[Fe]	Euryarchaeota	Methanobacteriales	Bidirectional	Tolerant	1	Fe ion	None

Figure S5. Screenshot showing the detailed content of the information pages about each hydrogenase class on HydDB. Equivalent information pages are available for all 38 hydrogenase classes defined in this work **(Table 1)**.

[NiFe] Group 1h-hydrogenase

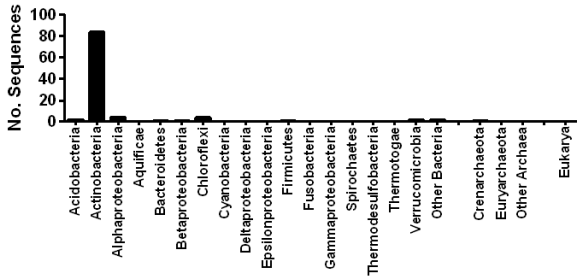
This entry was last updated at: June 13, 2016, 11:11 a.m.

Properties

Group	[NiFe] Group 1: Respiratory H_2-uptake [NiFe] hydrogenases
Subgroup	[NiFe] Group 1h: Actinobacteria-type
Function	Hydrogenotrophic respiration using O_2 as terminal electron acceptor. Enzyme scavenges electrons from atmospheric H_2 to fuel respiratory chain during carbon-starvation. Route of electron transfer unresolved.
Activity	H_2-uptake (unidirectional, high-affinity)
Oxygen tolerance	O_2-tolerant or O_2-insensitive
Localisation	Membrane-associated?

Distribution

Ecosystem distribution	Upland soils, plant tissues, possibly surface waters
Taxonomic distribution	Widespread among obligately aerobic soil bacteria, especially Actinobacteria, Acidobacteria, and Chloroflexi



Genetic Organisation

e.g. *Mycobacterium smegmatis* (WP_003893634.1):

Architecture

Structures	5AA5 (*Ralstonia eutropha*, 2.5 Å resolution, active)
Subunits	3?
Subunit description	HyhL (hydrogenase large subunit) HyhS (hydrogenase small subunit) HyhE (putative iron-sulfur protein and proposed physiological electron acceptor)
Catalytic site	[NiFe]-centre
FeS clusters	Proximal: 3Cys1Asp[4Fe4S] Medial: 4Cys[4Fe4S] Distal: 3Cys1His[4Fe4S]

Important Notes

The *Robiginitalea biformata* and *Sulfolobus islandicus* enzymes are relatively to distantly related to the main group. No studies have yet tested whether these enzymes have a H_2-scavenging role like other Group 1h [NiFe]-hydrogenases. They may instead represent founding members of a functionally-distinct lineage.

Sequences in this class

NCBI Accession	Organism	Hydrogenase Class	Phylum	Order	Activity (Predicted)	Oxygen Tolerance (Predicted)	Subunits (Predicted)	Metal Centres (Predicted)	Accessory Subunits (Predicted)
WP_014267363.1	Granulicella mallensis	[NiFe] Group 1h	Acidobacteria	Acidobacteriales	Aerobic Uptake	Tolerant	3	[NiFe]-centre, 3 x [4Fe4S] clusters	[FeS] protein
WP_011688202.1	Solibacter usitatus	[NiFe] Group 1h	Acidobacteria	Solibacterales	Aerobic Uptake	Tolerant	3	[NiFe]-centre, 3 x [4Fe4S] clusters	[FeS] protein
WP_021597135.1	Actinomadura madurae	[NiFe] Group 1h	Actinobacteria	Actinomycetales	Aerobic Uptake	Tolerant	3	[NiFe]-centre, 3 x [4Fe4S] clusters	[FeS] protein
WP_026402909.1	Actinomadura rifamycini	[NiFe] Group 1h	Actinobacteria	Actinomycetales	Aerobic Uptake	Tolerant	3	[NiFe]-centre, 3 x [4Fe4S] clusters	[FeS] protein
WP_018330638.1	Actinomycetospora chiangmaiensis	[NiFe] Group 1h	Actinobacteria	Actinomycetales	Aerobic Uptake	Tolerant	3	[NiFe]-centre, 3 x [4Fe4S] clusters	[FeS] protein
WP_007735075.1	Rhodococcus qingshengii	[NiFe] Group 1h	Actinobacteria	Actinomycetales	Aerobic Uptake	Tolerant	3	[NiFe]-centre, 3 x [4Fe4S] clusters	[FeS] protein
WP_003935326.1	Rhodococcus ruber	[NiFe] Group 1h	Actinobacteria	Actinomycetales	Aerobic Uptake	Tolerant	3	[NiFe]-centre, 3 x [4Fe4S] clusters	[FeS] protein
WP_005443931.1	Saccharomonospora azurea	[NiFe] Group 1h	Actinobacteria	Actinomycetales	Aerobic Uptake	Tolerant	3	[NiFe]-centre, 3 x [4Fe4S] clusters	[FeS] protein

« 1 2 3 »

Literature

Genetics:

- Berney, M., Greening, C., Hards, K., Collins, D., and Cook, G.M. (2014) Three different [NiFe] hydrogenases confer metabolic flexibility in the obligate aerobe *Mycobacterium smegmatis*. *Environ. Microbiol.* **16**: 318-330.
- Constant, P., Chowdhury, S.P., Hesse, L., and Conrad, R. (2011) Co-localization of atmospheric H_2 oxidation activity and high affinity H_2-oxidizing bacteria in non-axenic soil and sterile soil amended with *Streptomyces* sp. PCB7. *Soil Biol. Biochem.* **43**: 1888-1893.
- Constant, P., Chowdhury, S.P., Hesse, L., Pratscher, J., and Conrad, R. (2011) Genome data mining and soil survey for the novel group 5 [NiFe]-hydrogenase to explore the diversity and ecological importance of presumptive high-affinity H_2-oxidizing bacteria. *Appl. Environ. Microbiol.* **77**: 6027-6035.
- Greening, C., Biswas, A., Carere, C.R., Jackson, C.J., Taylor, M.C., Stott, M.B., Cook, G.M., and Morales, S.E. (2016) Genomic and metagenomic surveys of hydrogenase distribution indicate H_2 is a widely utilised energy source for microbial growth and survival. *ISME J.* **10**: 761-777.
- Khdhiri, M., Hesse, L., Popa, M.E., Quiza, L., Lalonde, I., Meredith, L.K., Röckmann, T., and Constant, P. (2015) Soil carbon content and relative abundance of high affinity H_2-oxidizing bacteria predict atmospheric H_2 soil uptake activity better than soil microbial community composition. *Soil Biol. Biochem.* **85**: 1-9.

Physiology:

- Berney, M., Greening, C., Conrad, R., Jacobs, W.R., and Cook, G.M. (2014) An obligately aerobic soil bacterium activates fermentative hydrogen production to survive reductive stress during hypoxia. *Proc. Natl. Acad. Sci. U. S. A.* **111**: 11479-11484.
- Constant, P., Chowdhury, S.P., Pratscher, J., and Conrad, R. (2010) Streptomycetes contributing to atmospheric molecular hydrogen soil uptake are widespread and encode a putative high-affinity [NiFe]-hydrogenase. *Environ. Microbiol.* **12**: 821-829.

Table S1. Hydrogenase sequences where there is disagreement between classification by SSN and *k*-NN methods. These sequences represent six out of the total 3248 sequences analyzed, i.e. 0.0018%.

NCBI Accession	Organism	*k*-NN Classification	SSN Classification
WP_027414715.1	Aneurinibacillus terranovensis	[NiFe] Group 1e	[NiFe] Group 1d
WP_027358538.1	Desulforegula conservatrix	[NiFe] Group 3d	[NiFe] Group 3c
WP_012532312.1	Geobacter bemidjiensis	[NiFe] Group 3d	[NiFe] Group 3c
WP_012469611.1	Geobacter lovleyi	[NiFe] Group 3d	[NiFe] Group 3c
WP_004512544.1	Geobacter metallireducens	[NiFe] Group 3d	[NiFe] Group 3c
WP_015839165.1	Geobacter sp. M21	[NiFe] Group 3d	[NiFe] Group 3c

Dataset S1. Excel spreadsheet listing the sequence, taxonomy, and hydrogenase class of all 3248 hydrogenase catalytic subunit sequences listed in HydDB