PNAS
www.pnas.org

Supplementary Information for

Small protein folds at the root of an ancient metabolic network

Hagai Raanan, Saroj Poudel, Douglas H. Pike, Vikas Nanda, Paul G. Falkowski

This PDF file includes:

www.pnas.org/cgi/doi/10.1073/pnas.1914982117

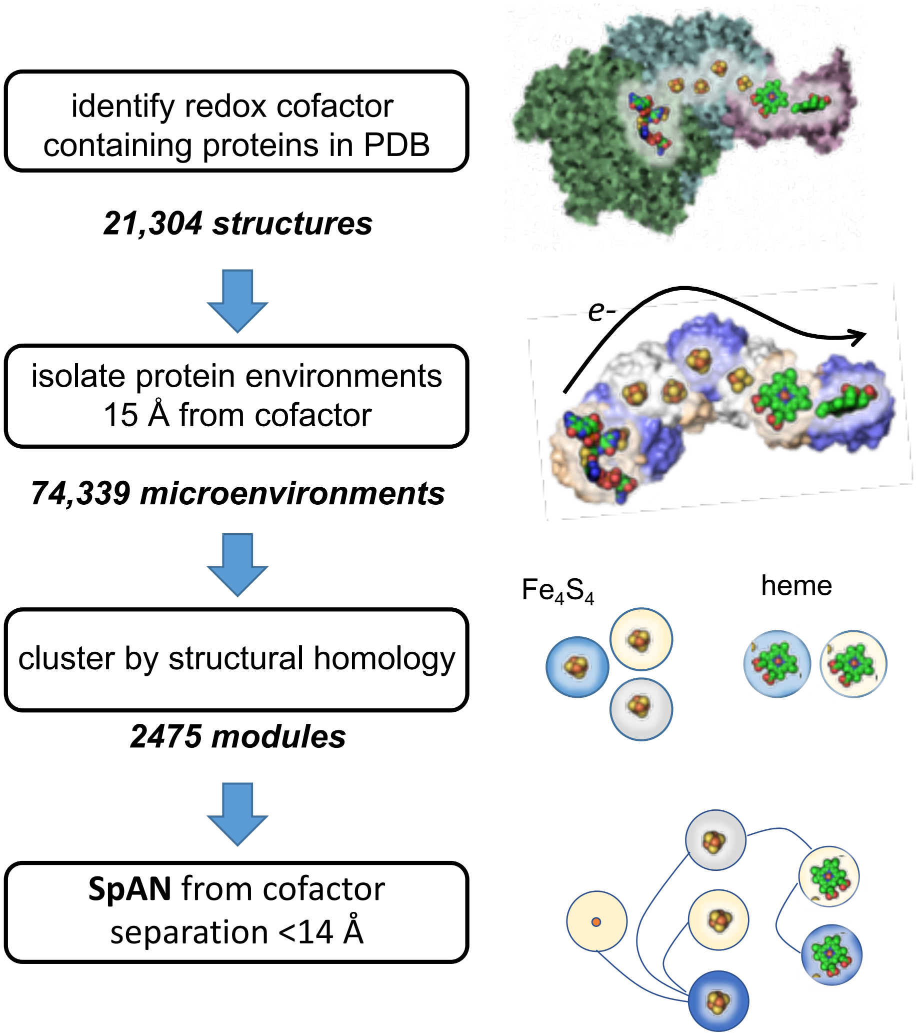

identify redox cofactor containing proteins in PDB
21,304 structures
isolate protein environments 15 Å from cofactor
74,339 microenvironments
cluster by structural homology
2475 modules
SpAN from cofactor separation <14 Å
e-
Fe_4S_4
heme

Fig. S1.

Generation of modules and the SpAN. Step 1 – cofactor containing proteins were identified in the PDB. Step 2 – microenvironments representing a 15Å radius sphere of residues from the center of the redox cofactor were isolated from each protein. Step 3 – microenvironments were clustered into modules based on structural similarity and sufficient superposition of cofactors in the alignment. Step 4 – a network of electron transfer paths (SpAN) was generated identifying pairs of microenvironments with cofactor edge-edge distances < 14Å apart. Example structure PDB 1q16 shown.

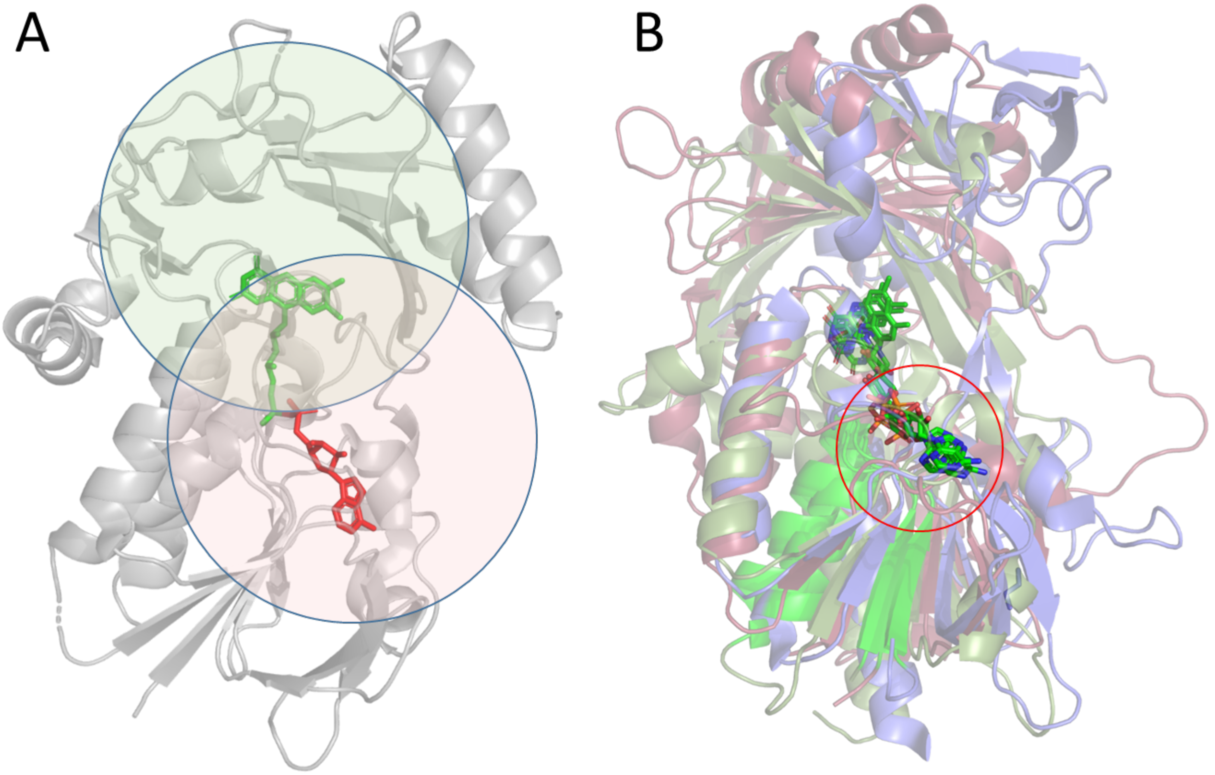


Fig. S2.

(A) The two microenvironments extracted from a nucleoside-based cofactor binding protein [3jq4]. The first microenvironment represents a 15Å distance around the geometric center of the expected electron transfer atoms (green) and the second around the nucleoside base (red). (B) Alignment example of three FAD containing proteins [5x68,2e48,3r9u]. The aligned protein around the nucleotide base of the FAD, which circled in red. All three nucleotide base microenvironments clustered into the Rossmann-like fold group (Rm) due to the similarity of the nucleoside binding site.

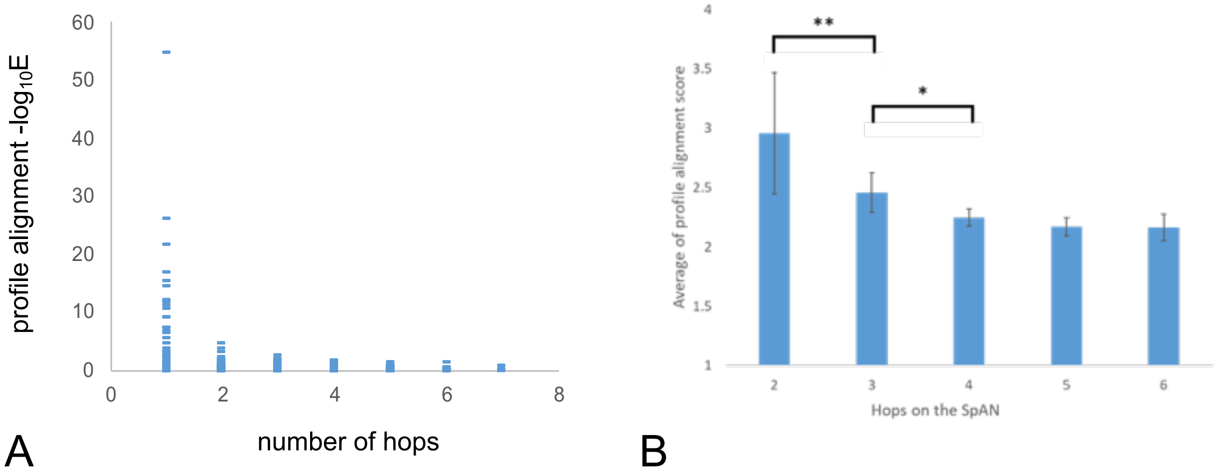


Fig. S3.

(A) Pairwise profile-alignment scores between modules ordered by number of hops in the SpAN, including single-hop scores. Single-hop scores were not included in our analysis as the overlap of protein microenvironments was expected to artifactually inflate alignment similarity. **(B)** Average of the profile-profile alignments score on SpAN. The 7th hop is excluded due to low sample size (n=3). Standard errors are presented. ** *p-value* <0.01. * *p-value* <0.05. Welch's t-test was used to calculate significance in distribution between the hops.

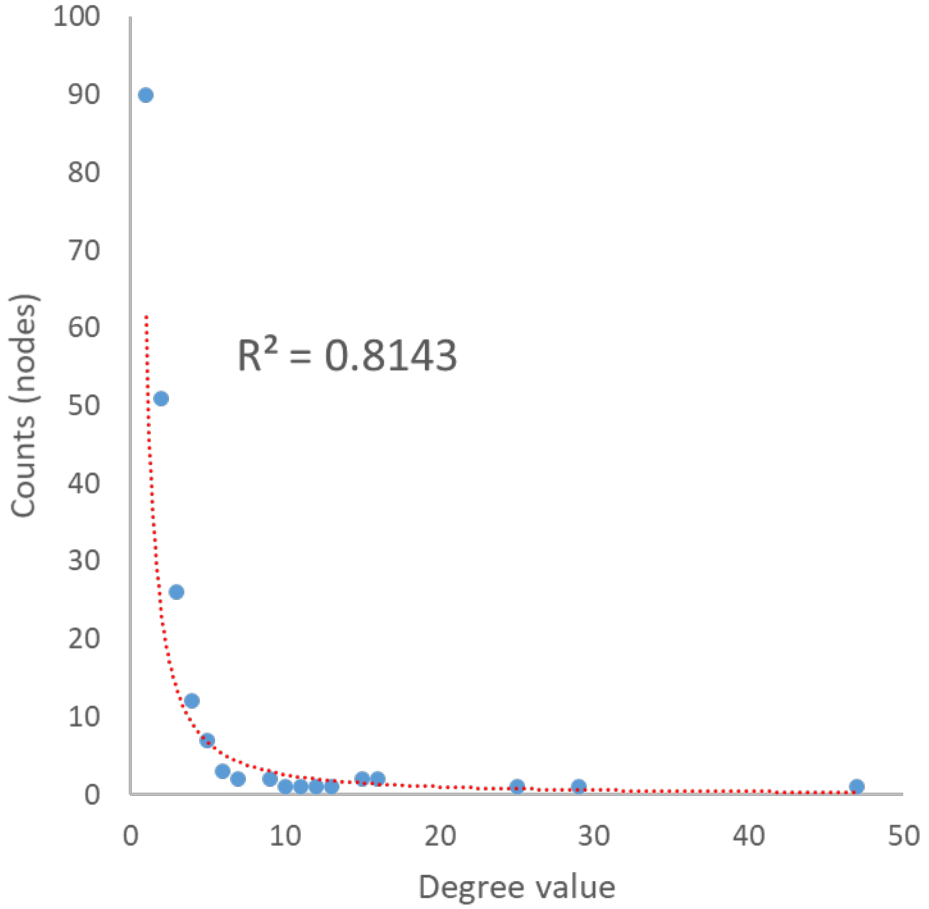


Fig. S4.

The degree distribution of SpAN. Data fit (red dashed line) shows that the SpAN follows the power law distribution, wherein a few nodes are highly connected, while the majority of nodes have only a few connections. A growing network and preferential attachment of new nodes can explain the power law distribution.

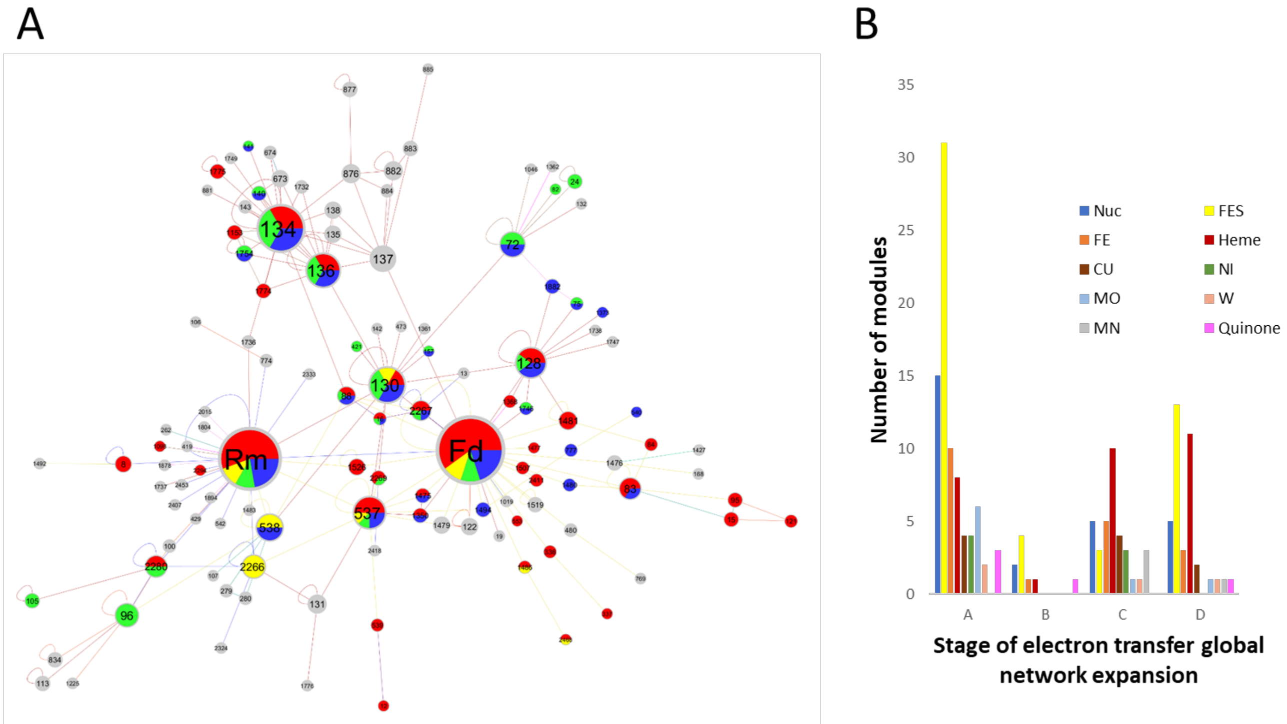


Fig. S5.

Estimated age of modules relative to the four stages of global electron transfer network expansion, inferred from the geological record (1). (A) SpAN is colored according to the module involvement in core microbial metabolic pathway stages. Stage A (red): methanogenesis, sulfur reduction, sulfate reduction, anoxygenic photosynthesis, and heterotrophy and autotrophy. Stage B (yellow): nitrogen fixation and oxygenic photosynthesis. Stage C (green): methane oxidation, nitrification, denitrification, sulfur oxidation, and sulfide oxidation. Stage D (blue): aerobic respiration, ammonification, and oxidation/reduction of other elements. Node size is relative to the node degree. (B) Cofactor distribution in each of the four stages of global electron transfer network expansion.

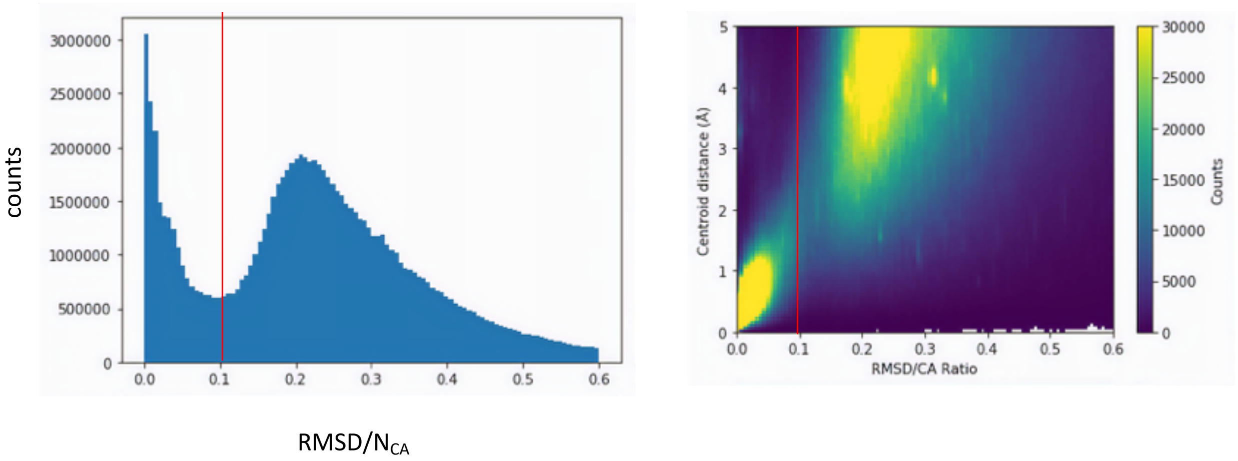


Fig. S6.

Pairwise alignments of microenvironments were clustered based on calculated RMSD per aligned backbone alpha-carbon (CA) determined by the pyMol 'align' program, and the centroid separation between the cofactors in each microenvironment. 0.1 Å RMSD per residue cleanly separates good alignments (low RMSD and centroid distance) from those large centroid separation.

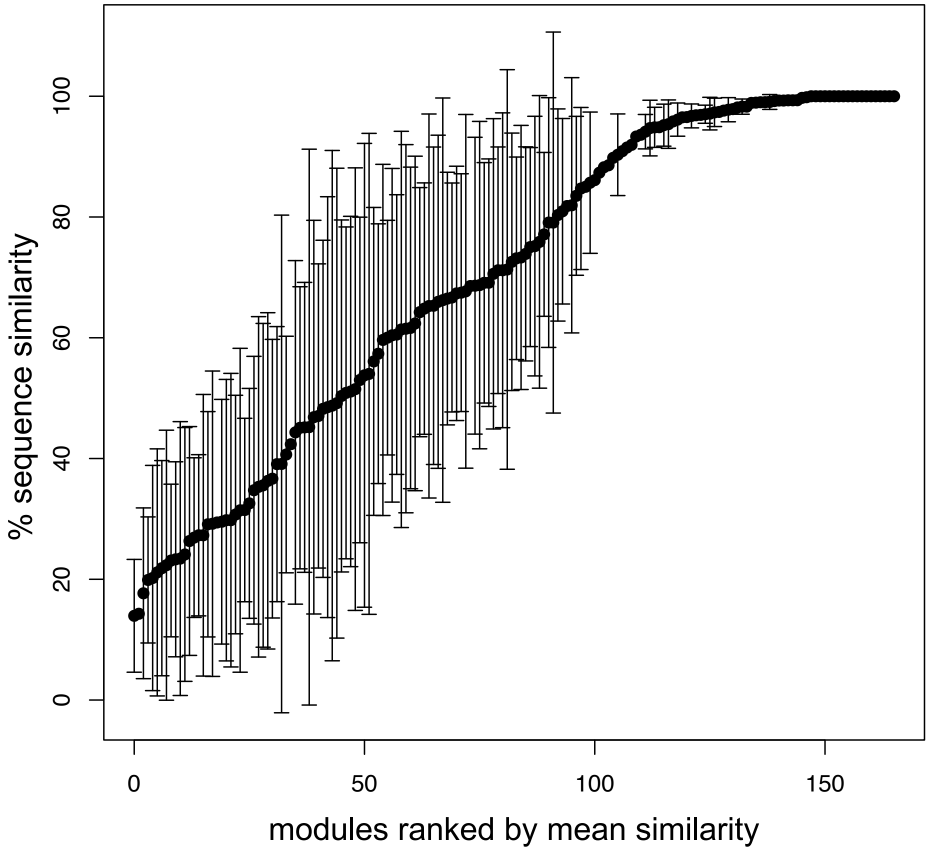
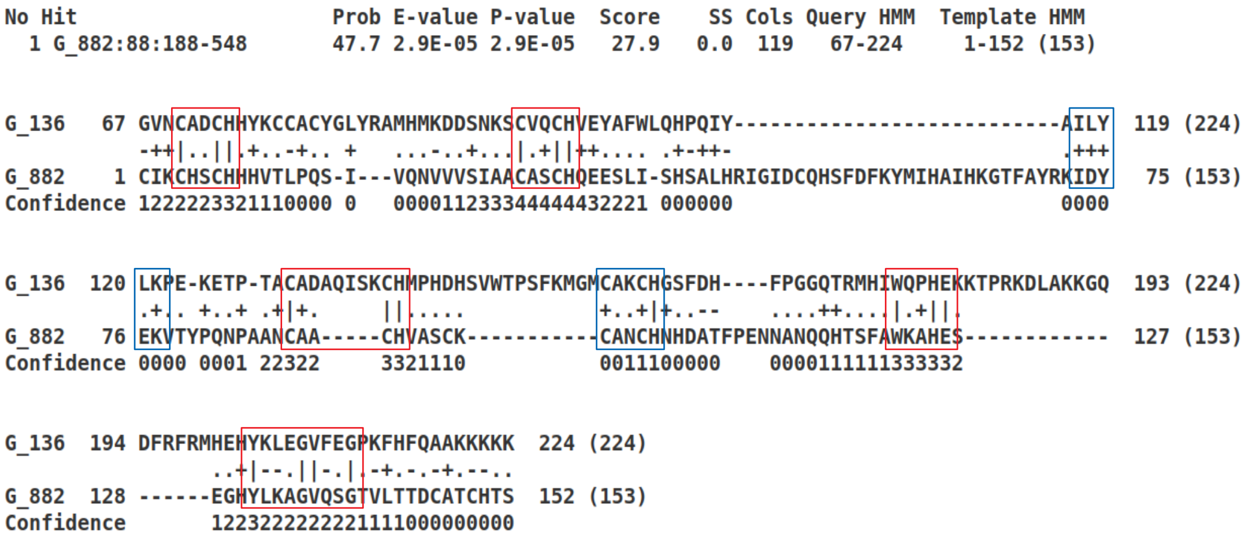


Fig. S7. Sequence diversity of a total of 160 connected modules.

The large connected component of the SpAN (**Fig. 2, main manuscript**) is comprised of over 160 modules. This plot shows the variation in module diversity from those that have high sequence diversity / low mean % sequence identity to low diversity / high identity.



```
No Hit                      Prob E-value P-value  Score    SS Cols Query HMM  Template HMM
  1 G_882:88:188-548        47.7 2.9E-05 2.9E-05   27.9   0.0  119   67-224     1-152 (153)


G_136   67 GVNCADCHHYKCCACYGLYRAMHMKDDSNKSCVQCHVEYAFWLQHPQIY---------------------------AILY  119 (224)
           -++|..||.+..-+.. +   ...-..+...|.+||++.... .+-++-                           .+++
G_882    1 CIKCHSCHHHVTLPQS-I---VQNVVVSIAACASCHQEESLI-SHSALHRIGIDCQHSFDFKYMIHAIHKGTFAYRKIDY   75 (153)
Confidence 1222223321110000 0   000011233344444432221 000000                           0000


G_136  120 LKPE-KETP-TACADAQISKCHMPHDHSVWTPSFKMGMCAKCHGSFDH----FPGGQTRMHIWQPHEKKTPRKDLAKKGQ  193 (224)
           .+|. +..+ .+|+.     ||.....           +..+|+..--    ....++....|.+||.
G_882   76 EKVTYPQNPAANCAA-----CHVASCK-----------CANCHNHDATFPENNANQQHTSFAWKAHES------------  127 (153)
Confidence 0000 0001 22322     3321110           0011100000    0000111111333332


G_136  194 DFRFRMHEHYKLEGVFEGPKFHFQAAKKKKK  224 (224)
                 ..+|--.||-.|.-+.-.-+.--..
G_882  128 ------EGHYLKAGVQSGTVLTTDCATCHTS  152 (153)
Confidence       1223222222221111000000000
```

Figure S8:

Profile-profile comparison of group 136 and group 882. The profile comparison was made using the program HHblits package (2). The highlighted red and blue boxes represent segments of profile with >= 2 and <2 confidence score respectively. E-value and Probability score are also provided.

Table S1

Proteins involved in the Wood-Ljungdhal pathway along with their 3D protein structures of *Moorella thermoacetica*

Reaction	Enzyme	Gene id	PDB	Template	Coverage	Seq identity
1	NADPH formate dehydrogenase	Moth_2312	NA	1fdo.A	0.75	43.28
1	NADPH formate dehydrogenase	Moth_2314	NA	4yry_B	0.58	45.48
2	10-Formyl-H4 folate synthetase (ADP forming)	moth_0109	4ioj	NA	NA	NA
3	5,10-Methenyl-H4 folate cyclohydrolase	Moth_1516	NA	4a5o.A	0.99	56.83
4	5,10-Methylene-H4 folate dehydrogenase [NAD(P)H]	Moth_1516	NA	4a5o.A	0.99	56.83
5	5,10-Methylene-H4 folate reductase (Fd2?red)	Moth_1191	NA	3apt	0.92	24.64
6	Methyl-H4 folate corrinoid iron-sulfur protein methyltransferase	Moth_1197	4djd A	NA	NA	NA
7	Corrinoid iron-sulfur protein (CFeSP)	Moth_1198/1201	4djd CD	NA	NA	NA
8	CO dehydrogenase/acetyl-CoA synthase (Fd2?red)	Moth_1198/1201/1202	3i01	NA	NA	NA
9	Pyruvate synthase	Moth_0064-	6cin	NA	NA	NA

SUPPLEMENTAL REFERENCES

1. E. K. Moore, B. I. Jelen, D. Giovannelli, H. Raanan, P. G. Falkowski, Metal availability and the expanding network of microbial metabolisms in the Archaean eon. *Nat Geosci* **10**, 629-636 (2017).
2. M. Remmert, A. Biegert, A. Hauser, J. Soding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173-175 (2011).