

1 **ASM-Clust: classifying functionally diverse protein families using alignment score matrices**

2 Daan R. Speth¹ and Victoria J. Orphan¹

3

4 ¹ Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena,
5 CA, USA

6

7 **Correspondence:**

8 dspeth@caltech.edu or vorphan@gps.caltech.edu

9

10 **Abstract**

11 Rapid advances in sequencing technology have resulted in the availability of genomes from
12 organisms across the tree of life. Accurately interpreting the function of proteins in these
13 genomes is a major challenge, as annotation transfer based on homology frequently results in
14 misannotation and error propagation. This challenge is especially pressing for organisms whose
15 genomes are directly obtained from environmental samples, as interpretation of their physiology
16 and ecology is often based solely on the genome sequence. For complex protein (super)families
17 containing a large number of sequences, classification can be used to determine whether
18 annotation transfer is appropriate, or whether experimental evidence for function is lacking. Here
19 we present a novel computational approach for de novo classification of large protein
20 (super)families, based on clustering an alignment score matrix obtained by aligning all sequences
21 in the family to a small subset of the data. We evaluate our approach on the enolase family in the
22 Structure Function Linkage Database.

23

24 **Availability and implementation**

25 ASM-Clust is implemented in bash with helper scripts in perl. Scripts comprising ASM-Clust are
26 available for download from https://github.com/dspeth/bioinfo_scripts/tree/master/ASM_clust/

27

28 **Introduction**

29 The rapid advances in sequencing technology have led to a dramatic increase in available
30 genome sequences. This genomic data has provided new perspectives on big questions in
31 biology, such as the diversity of life, the distribution of metabolic traits across the tree of life,
32 and the origin of eukaryotes (Hug et al. 2016; Zaremba-Niedzwiedzka et al. 2017; Borrel et al.
33 2019). In addition, each newly available genome sequence contains novel protein sequences,
34 yielding novel protein families of unknown function and expanding families with previously
35 characterized representatives. Automatic functional annotation of novel protein sequences is
36 generally done by annotation transfer from known homologous proteins, either using sequence
37 alignment or hidden markov models (Altschul et al. 1990; Finn, Clements, and Eddy 2011). This
38 approach can, and often does, result in misinterpretation of the function of proteins in
39 mechanistically diverse superfamilies, and is prone to subsequent error propagation (Schnoes et
40 al. 2009). Accurately interpreting the function of novel proteins is one of the grand challenges in
41 biology, and relies heavily on availability of experimental data. Classifying mechanistically
42 diverse protein superfamilies provides insight in knowledge gaps, can indicate whether
43 annotation transfer is appropriate, and can help guide future experiments.

44 There are various automatic tools available for classification of proteins into isofunctional
45 families using sequence similarity, active site characteristics, and phylogenetic relationships
46 (Brown, Krishnamurthy, and Sjölander 2007; Lee, Rentzsch, and Orengo 2010; de Melo-
47 Minardi, Bastard, and Artiguenave 2010; Leuthaeuser et al. 2016; Knutson et al. 2017).
48 Alternatively, the structure of a protein family can be interactively explored using sequence
49 similarity networks (SSNs) (Atkinson et al. 2009; Copp et al. 2018). SSNs are constructed based
50 on pairwise all vs all alignment, with each node in the network representing a sequence, and each
51 edge between two nodes representing the alignment between sequences. Clusters of nodes can be
52 manually selected, or identified using a clustering algorithm such as MCL (Enright, Van
53 Dongen, and Ouzounis 2002). SSNs are a powerful method to investigate protein families, but
54 the network visualization limits the number of sequences that can be included, and alignments
55 between separate domains of multi-domain proteins may confuse the analysis.

56 Here we present ASM-Clust, an alternative method that uses alignment score matrix (ASM)
57 clustering. For each input sequence, ASM-Clust generates a profile consisting of a large number
58 of alignment scores, including both presence/absence and weight, and uses this profile to classify

59 each sequence. For a dataset containing N sequences, alignments are generated for all N
60 sequences against a randomly selected subset of n sequences, and taking each alignment score, or
61 a zero if the sequence did not align to the reference. This results in a matrix of $N \times n$ values
62 which is subsequently visualized using t-distributed stochastic neighbor embedding (t-SNE)
63 (Van der Maaten and Hinton 2008; Van der Maaten 2014), and can be clustered using DBscan
64 (Ester et al. 1996).

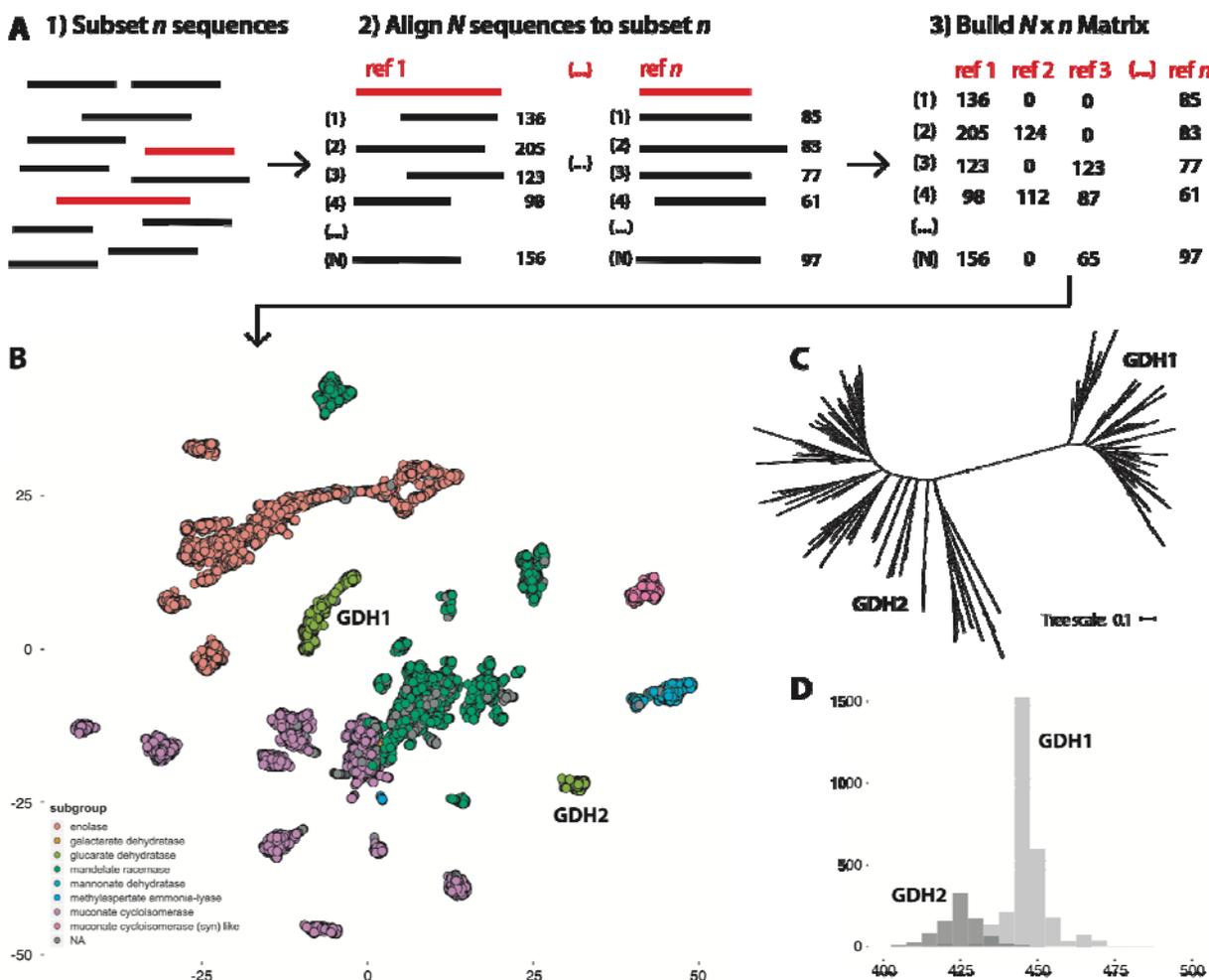
65

66 **Implementation**

67 ASM-Clust is implemented in bash with helper scripts in perl, and will take a protein fasta file as
68 the sole input. Fasta files are processed with `ASM_clust.sh`, which 1) randomly selects a subset
69 of n sequences (default 1000), 2) aligns the entire dataset to the subset of n sequences, 3)
70 combines all scores into a matrix (inserting 0 for query-database pairs that did not produce an
71 alignment), and 4) reduces the matrix to 2 dimensions using t-SNE (Figure 1a). For flexible
72 usage, ASM-Clust supports alignment using DIAMOND (Buchfink, Xie, and Huson 2015),
73 BLAST (Altschul et al. 1990), or MMSeqs2 (Steinegger and Söding 2017), and uses MMSeqs2
74 as default alignment software. Clustering results are comparable between different alignment
75 software (Supplemental Figure S1). Other user-defined options are the number of sequences in
76 the subset (default 1000), the main t-SNE parameter “perplexity” (default 1000) and maximum
77 iterations (default 5000) for dimensionality reduction, and the number of threads used by the
78 alignment software (default 1). Although the clustering is generally similar with multiple
79 randomly chosen subsets (Supplemental Figure S2), the subset can be defined for reproducibility.
80 The output of `ASM_clust.sh` can be visualized as a scatterplot where each dot represents a
81 sequence, and clusters are readily apparent (Figure 1b, Supplemental figure S1-S3). This format
82 allows additional annotation with sequence features, such as taxonomy, length, or composition.
83 The visualization in Figure 1b and Supplementary Figures S1-S3 was created using R, with the
84 `ggplot2` package, and clusters were called using the `dbscan` package. The t-SNE result and the
85 annotation data downloaded from SFLD were combined prior to visualization. Clusters obtained
86 with ASM-Clust can be further refined by iteratively applying the method to a subset of poorly
87 resolved data, such as the “hub”’s cluster (Supplementary Figure S3). The smaller total number of
88 sequences in the second iteration, combined with lowering the perplexity value of the t-SNE,

89 increases the resolving power of the analysis for clades with a small number of sequences, thus
 90 resolving rare classes with few members (Supplementary Figure S3).

91
 92



93

94

95 **Figure 1. ASM-clust workflow overview and enolase superfamily example.**

96 A) ASM-Clust workflow overview and B) example of the ASM-Clust output on the structure
 97 function linkage database (SFLD) enolase superfamily (48,850 sequences). The clusters are
 98 colored by SFLD subgroup: enolase (red), galactarate dehydratase (orange), glucarate
 99 dehydratase (light green), mandelate racemase (dark green), mannonate dehydratase (light blue),
 100 methylaspartate-ammonia lyase (dark blue), muconate cycloisomerase (purple), muconate
 101 cycloisomerase (syn) like (pink), and no assigned subgroup (gray). The isofunctional ‘glucarate

102 dehydratase' subgroup is split in two clusters, indicated with GDH1 and GDH2. C) Phylogenetic
103 analysis of the 'glucarate dehydratase' subgroup (clustered at 70% identity), and D) sequence
104 length comparison confirms the clear separation of the two clusters.

105

106 **Results**

107 ASM-Clust was tested on the enolase superfamily in the gold-standard Structure Function
108 Linkage Database (SFLD) (Akiva et al. 2014). Sequences and annotation data table were
109 downloaded from the SFLD website (<http://sfld.rbvi.ucsf.edu>) and all 48,850 sequences were
110 clustered using ASM-Clust with default settings, and visualized using R (Fig 1b). The
111 'mannonate dehydratase' and 'muconate cycloisomerase (syn) like' subgroups, each containing
112 only a single isofunctional family, are well resolved. As expected, the functionally diverse
113 'muconate cycloisomerase' and 'mandelate racemase' subgroups each partition into multiple
114 discrete clusters (Figure 1b). The isofunctional 'enolase' and 'glucarate dehydratase' subgroups
115 also result in multiple clusters (Figure 1b). Phylogenetic analysis of the 'glucarate dehydratase'
116 subgroup confirms that the observed clusters respond to distinct clades that can also be separated
117 by sequence length (Figure 1c & 1d, Supplementary methods). The smaller methylaspartate
118 ammonia-lyase and galactarate dehydratase subgroups (307 and 25 sequences respectively) are
119 more clearly resolved when ASM-clust is iteratively rerun on the "hub" cluster with a lower
120 perplexity value (Supplemental Figure S3). When prior high-quality annotation is not available,
121 clusters can be inspected for phylogeny, taxonomic distribution, and conserved residues to assess
122 whether they represent functionally divergent sequences.

123 ASM-Clust can retrieve clades from a complex superfamily with tens of thousands of sequences,
124 without prior reduction of the dataset. We expect this to become increasingly relevant as the
125 amount of sequence data from phylogenetically diverse organisms continues to grow rapidly, and
126 meaningful information can be overlooked while pre-clustering a sequence dataset.

127

128 **Funding**

129 This work was supported by the US Department of Energy, Office of Science, Office of
130 Biological and Environmental Research under award number DE-SC0016469 to Victoria J.
131 Orphan. Daan R. Speth was supported by the Netherlands Organisation for Scientific Research,
132 Rubicon award 019.153LW.039.

133 **References**

- 134
- 135 Akiva, Eyal, Shoshana Brown, Daniel E. Almonacid, Alan E. Barber 2nd, Ashley F. Custer,
136 Michael A. Hicks, Conrad C. Huang, et al. 2014. “The Structure-Function Linkage Database.”
137 *Nucleic Acids Research* 42 (Database issue): D521–30.
- 138
- 139 Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local
140 Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.
- 141
- 142 Atkinson, Holly J., John H. Morris, Thomas E. Ferrin, and Patricia C. Babbitt. 2009. “Using
143 Sequence Similarity Networks for Visualization of Relationships across Diverse Protein
144 Superfamilies.” *PloS One* 4 (2): e4345.
- 145
- 146 Borrel, Guillaume, Panagiotis S. Adam, Luke J. McKay, Lin-Xing Chen, Isabel Natalia Sierra-
147 García, Christian M. K. Sieber, Quentin Letourneur, et al. 2019. “Wide Diversity of Methane and
148 Short-Chain Alkane Metabolisms in Uncultured Archaea.” *Nature Microbiology* 4 (4): 603–13.
- 149
- 150 Brown, Duncan P., Nandini Krishnamurthy, and Kimmen Sjölander. 2007. “Automated Protein
151 Subfamily Identification and Classification.” *PLoS Computational Biology* 3 (8): e160.
- 152
- 153 Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein
154 Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60.
- 155
- 156 Copp, Janine N., Eyal Akiva, Patricia C. Babbitt, and Nobuhiko Tokuriki. 2018. “Revealing
157 Unexplored Sequence-Function Space Using Sequence Similarity Networks.” *Biochemistry* 57
158 (31): 4651–62.
- 159
- 160 Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. “An Efficient Algorithm for Large-
161 Scale Detection of Protein Families.” *Nucleic Acids Research* 30 (7): 1575–84.
- 162

163 Ester, M., H. P. Kriegel, J. Sander, and X. Xu. 1996. "A Density-Based Algorithm for
164 Discovering Clusters in Large Spatial Databases with Noise." KDD: Proceedings / International
165 Conference on Knowledge Discovery & Data Mining. International Conference on Knowledge
166 Discovery & Data Mining. <https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf>.
167

168 Finn, Robert D., Jody Clements, and Sean R. Eddy. 2011. "HMMER Web Server: Interactive
169 Sequence Similarity Searching." *Nucleic Acids Research* 39 (Web Server issue): W29–37.
170

171 Hug, Laura A., Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J.
172 Probst, Cindy J. Castelle, Cristina N. Butterfield, et al. 2016. "A New View of the Tree of Life."
173 *Nature Microbiology* 1 (April): 16048.
174

175 Knutson, Stacy T., Brian M. Westwood, Janelle B. Leuthaeuser, Brandon E. Turner, Don
176 Nguyendac, Gabrielle Shea, Kiran Kumar, et al. 2017. "An Approach to Functionally Relevant
177 Clustering of the Protein Universe: Active Site Profile-Based Clustering of Protein Structures
178 and Sequences: Functionally Relevant Clustering of Protein Superfamilies." *Protein Science: A
179 Publication of the Protein Society* 26 (4): 677–99.
180

181 Lee, David A., Robert Rentzsch, and Christine Orengo. 2010. "GeMMA: Functional Subfamily
182 Classification within Superfamilies of Predicted Protein Structural Domains." *Nucleic Acids
183 Research* 38 (3): 720–37.
184

185 Leuthaeuser, Janelle B., John H. Morris, Angela F. Harper, Thomas E. Ferrin, Patricia C.
186 Babbitt, and Jacquelyn S. Fetrow. 2016. "DASP3: Identification of Protein Sequences Belonging
187 to Functionally Relevant Groups." *BMC Bioinformatics* 17 (1): 458.
188

189 Melo-Minardi, Raquel C. de, Karine Bastard, and François Artiguenave. 2010. "Identification of
190 Subfamily-Specific Sites Based on Active Sites Modeling and Clustering." *Bioinformatics* 26
191 (24): 3075–82.
192

193 Schnoes, Alexandra M., Shoshana D. Brown, Igor Dodevski, and Patricia C. Babbitt. 2009.
194 “Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme
195 Superfamilies.” *PLoS Computational Biology* 5 (12): e1000605.
196
197 Steinegger, Martin, and Johannes Söding. 2017. “MMseqs2 Enables Sensitive Protein Sequence
198 Searching for the Analysis of Massive Data Sets.” *Nature Biotechnology* 35 (11): 1026–28.
199
200 Van der Maaten, Laurens, and Geoffrey Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal*
201 *of Machine Learning Research: JMLR* 9 (Nov): 2579–2605.
202
203 Van Der Maaten, L. 2014. “Accelerating T-SNE Using Tree-Based Algorithms.” *Journal of*
204 *Machine Learning Research: JMLR*.
205 <http://www.jmlr.org/papers/volume15/vandermaaten14a/vandermaaten14a.pdf>.
206
207 Zaremba-Niedzwiedzka, Katarzyna, Eva F. Caceres, Jimmy H. Saw, Disa Bäckström, Lina
208 Juzokaite, Emmelien Vancaester, Kiley W. Seitz, et al. 2017. “Asgard Archaea Illuminate the
209 Origin of Eukaryotic Cellular Complexity.” *Nature* 541 (7637): 353–58.
210
211

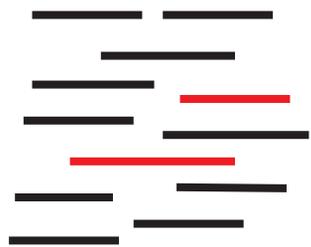
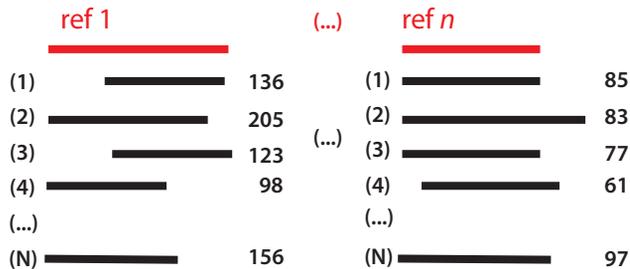
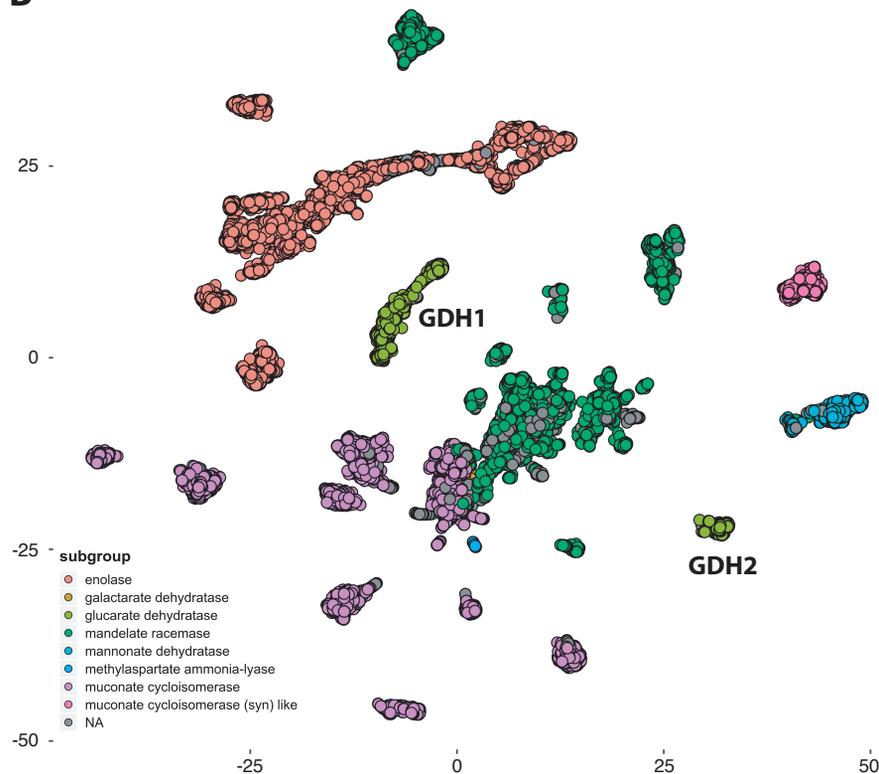
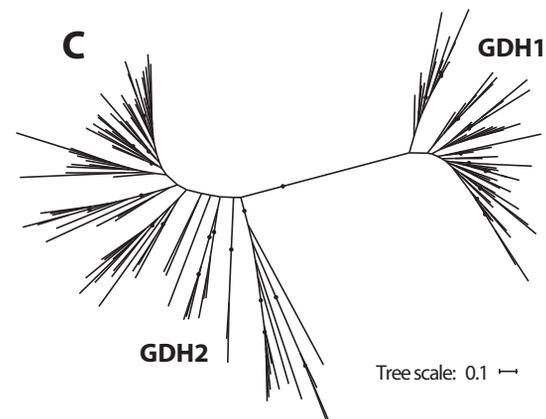
A 1) Subset n sequences**2) Align N sequences to subset n** **3) Build $N \times n$ Matrix**

Diagram illustrating the construction of an $N \times n$ matrix from the alignment results. The matrix is shown as a table of values representing the alignment scores for each sequence pair.

	ref 1	ref 2	ref 3	(...)	ref n
(1)	136	0	0		85
(2)	205	124	0		83
(3)	123	0	123		77
(4)	98	112	87		61
(...)					
(N)	156	0	65		97

B**C****D**