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

Daan R. Speth[1] and Victoria J. Orphan[1]

[1] Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA

Correspondence:

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

Abstract

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

Availability and implementation

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

Introduction

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

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

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

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

Implementation

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

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

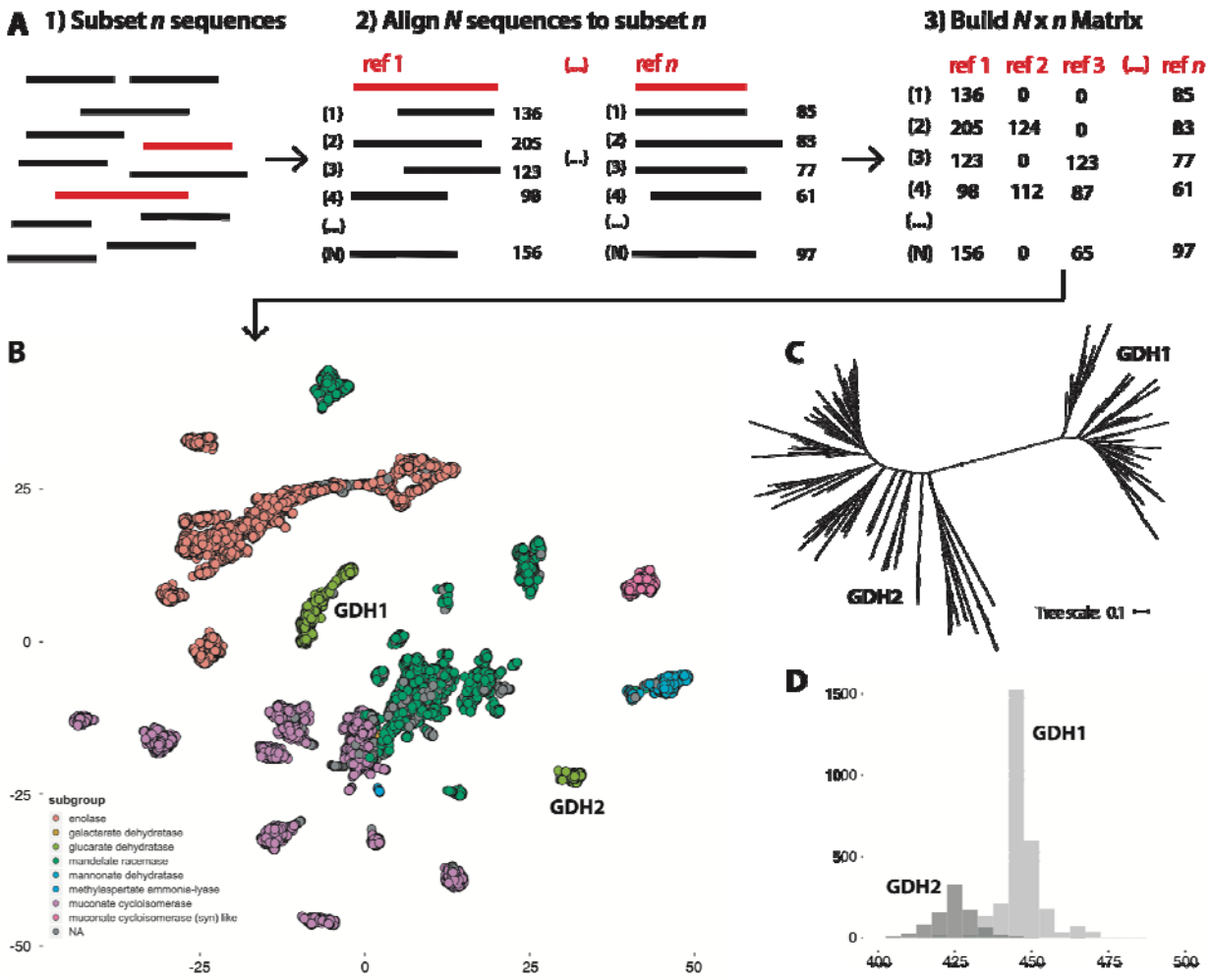


Figure 1. ASM-clust workflow overview and enolase superfamily example.

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

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

Results

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

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

Funding

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

References

Akiva, Eyal, Shoshana Brown, Daniel E. Almonacid, Alan E. Barber 2nd, Ashley F. Custer, Michael A. Hicks, Conrad C. Huang, et al. 2014. "The Structure-Function Linkage Database." Nucleic Acids Research 42 (Database issue): D521–30.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." Journal of Molecular Biology 215 (3): 403–10.

Atkinson, Holly J., John H. Morris, Thomas E. Ferrin, and Patricia C. Babbitt. 2009. "Using Sequence Similarity Networks for Visualization of Relationships across Diverse Protein Superfamilies." PloS One 4 (2): e4345.

Borrel, Guillaume, Panagiotis S. Adam, Luke J. McKay, Lin-Xing Chen, Isabel Natalia Sierra-García, Christian M. K. Sieber, Quentin Letourneur, et al. 2019. "Wide Diversity of Methane and Short-Chain Alkane Metabolisms in Uncultured Archaea." Nature Microbiology 4 (4): 603–13.

Brown, Duncan P., Nandini Krishnamurthy, and Kimmen Sjölander. 2007. "Automated Protein Subfamily Identification and Classification." PLoS Computational Biology 3 (8): e160.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." Nature Methods 12 (1): 59–60.

Copp, Janine N., Eyal Akiva, Patricia C. Babbitt, and Nobuhiko Tokuriki. 2018. "Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks." Biochemistry 57 (31): 4651–62.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. "An Efficient Algorithm for Large-Scale Detection of Protein Families." Nucleic Acids Research 30 (7): 1575–84.

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

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

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

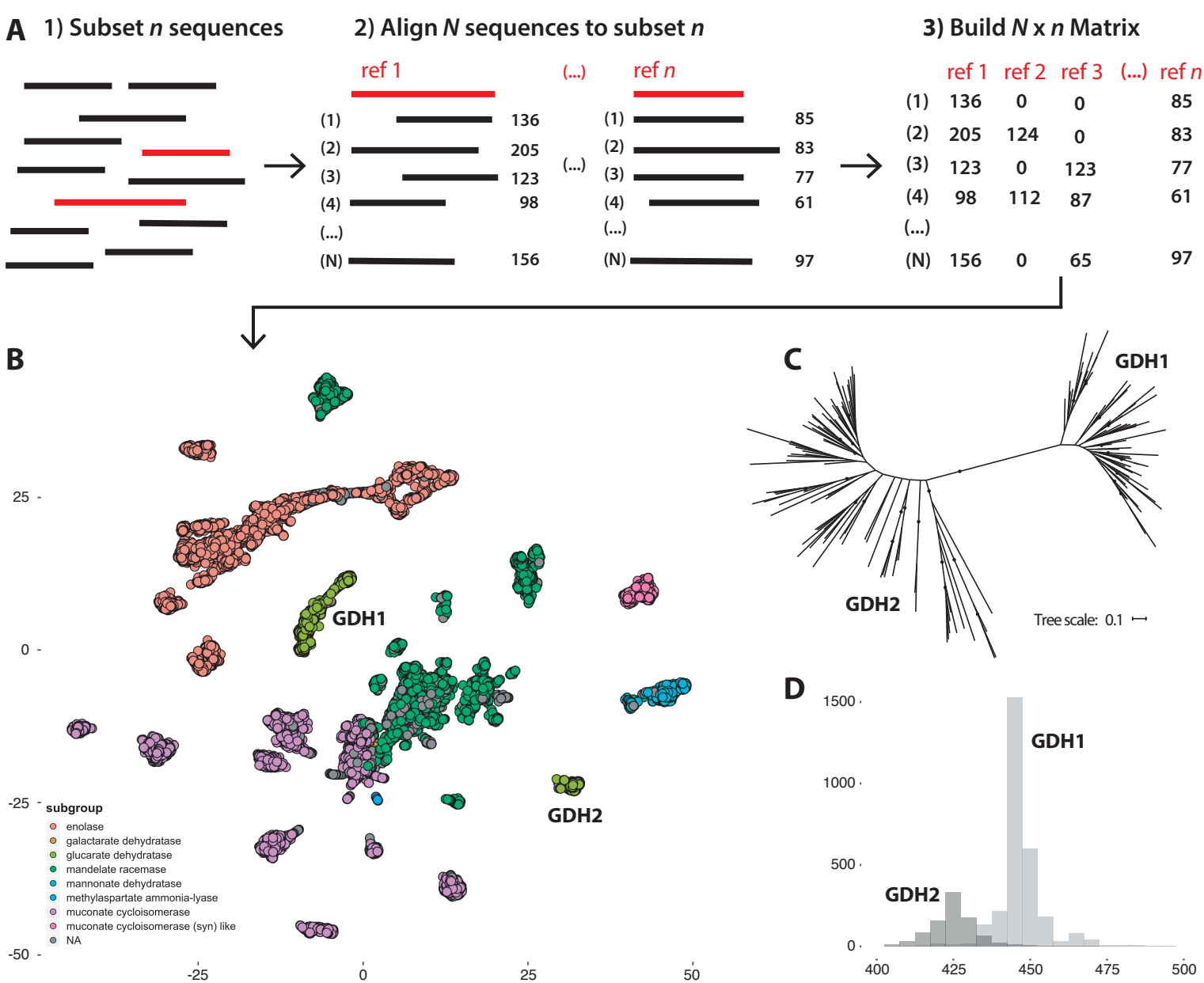
Knutson, Stacy T., Brian M. Westwood, Janelle B. Leuthaeuser, Brandon E. Turner, Don Nguyendac, Gabrielle Shea, Kiran Kumar, et al. 2017. “An Approach to Functionally Relevant Clustering of the Protein Universe: Active Site Profile-Based Clustering of Protein Structures and Sequences: Functionally Relevant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

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

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

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

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
ref 1
(...)
ref n
(1) 136
(2) 205
(3) 123
(4) 98
(...)
(N) 156
(1) 85
(2) 83
(3) 77
(4) 61
(...)
(N) 97
3) Build N x n Matrix
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
GDH1
GDH2
subgroup
enolase
galactarate dehydratase
glucarate dehydratase
mandelate racemase
mannonate dehydratase
methylaspartate ammonia-lyase
muconate cycloisomerase
muconate cycloisomerase (syn) like
NA
25
0
-25
-50
-25
0
25
50
C
GDH1
GDH2
Tree scale: 0.1
D
GDH1
GDH2
1500
1000
500
0
400
425
450
475
500