RESEARCH ARTICLE

Community detection in sequence similarity networks based on attribute clustering

Janamejaya Chowdhary[1,2¤]*, **Frank E. Löffler**[3,4,5]*, **Jeremy C. Smith**[1,2,5]*

1 Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, Tennessee, United States of America, **2** University of Tennessee-Oak Ridge National Laboratory, Joint Institute for Biological Sciences and Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, United States of America, **3** Department of Microbiology, Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, Tennessee, United States of America, **4** Center for Environmental Biotechnology, University of Tennessee, Knoxville, Tennessee, United States of America, **5** Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, Tennessee, United States of America

¤ Current address: Private Residence, Oak Ridge, Tennessee, United States of America
* janamejaya.chowdhary@gmail.com (JC); frank.loeffler@utk.edu (FL); smithjc@ornl.gov (JCS)

OPEN ACCESS

Citation: Chowdhary J, Löffler FE, Smith JC (2017) Community detection in sequence similarity networks based on attribute clustering. PLoS ONE 12(7): e0178650. https://doi.org/10.1371/journal.pone.0178650

Editor: Irene Sendiña-Nadal, Universidad Rey Juan Carlos, SPAIN

Received: January 27, 2017

Accepted: May 16, 2017

Published: July 24, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was funded by the University of Tennessee.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Networks are powerful tools for the presentation and analysis of interactions in multi-component systems. A commonly studied mesoscopic feature of networks is their community structure, which arises from grouping together similar nodes into one community and dissimilar nodes into separate communities. Here, the community structure of protein sequence similarity networks is determined with a new method: Attribute Clustering Dependent Communities (ACDC). Sequence similarity has hitherto typically been quantified by the alignment score or its expectation value. However, pair alignments with the same score or expectation value cannot thus be differentiated. To overcome this deficiency, the method constructs, for pair alignments, an extended alignment metric, the link attribute vector, which includes the score and other alignment characteristics. Rescaling components of the attribute vectors qualitatively identifies a systematic variation of sequence similarity within protein superfamilies. The problem of community detection is then mapped to clustering the link attribute vectors, selection of an optimal subset of links and community structure refinement based on the partition density of the network. ACDC-predicted communities are found to be in good agreement with gold standard sequence databases for which the "ground truth" community structures (or families) are known. ACDC is therefore a community detection method for sequence similarity networks based entirely on pair similarity information. A serial implementation of ACDC is available from https://cmb.ornl.gov/resources/developments

1. Introduction

A core objective of bioinformatics is classification of sequences. Given a set of sequences, each sequence is assigned to a sequence family which corresponds to evolutionarily related sequences sharing significant sequence, structure, and/or mechanistic similarity. Sequence families sharing overall structural and/or mechanistic similarity suggestive of a common evolutionary origin are grouped together into a superfamily. As only a tiny fraction of known sequences have been experimentally characterized, such a classification of sequences is important for developing, for example, functional hypotheses [1], identification of erroneous functional annotation [2], and understanding evolutionary mechanisms underlying enzyme sequence variation within superfamilies [3]. Due to availability of millions of protein and gene sequences due to affordable sequencing, it is imperative that accurate classification (e.g., [4]) of sequences be performed for making sense of the sequence universe.

Membership of a sequence in a sequence family is based on similarity between pairs of sequences. Sequence similarity is typically quantified using pair sequence alignments. For a given amino-acid or nucleotide substitution matrix and associated gap open and gap extension penalties, an objective function is defined which assigns a cost for identical and mismatched residues based on the substitution matrix, and a penalty for introducing and extending gaps. During pair sequence alignment, this objective function is maximized and an optimal gapped alignment is identified [5]. Such an alignment is fully quantified by the alignment length (L_a), the number of identical residues (N_id), the number of mismatched residues (N_m) and the number of gaps (N_g). The problem of finding an optimal alignment is equivalent to finding the ground state of a directed path in a 2-dimensional random medium [6] to which the optimal alignment corresponds to the ground state of the directed path and the alignment score (S) is then analogous to the negative free energy of the ground state [7]. The statistical significance of a score is quantified in terms of an expectation value (E-val) and measures the excess similarity for the pair alignment relative to alignment with random sequences [8, 9]. These alignment attributes then fully quantify similarity between sequence pairs.

A convenient framework for simultaneous incorporation of all pair similarity information is that of networks. A network is a visually efficient and mathematically convenient representation of information contained in a relational system [10, 11]. In general, a network can be represented as an undirected attributed graph $G = \{V, E, A_\mathrm{V}, A_\mathrm{E}\}$ which, at a minimum, contains as its elements the set V of N_V nodes or vertices and a set E of N_E links or edges. Each node i in set V corresponds to an indivisible interacting component of the system and each link in set E connects a pair of nodes depending of the presence of some relationship. The topology of a network is completely specified by the set of nodes and links. Additionally, each node may be associated with a set of attributes or metadata, e.g., the name and age of individuals in a social network. Similarly, each link may be associated with attributes, e.g., shared hobbies between two individuals in a social network. With the set of node and link attributes, A_v and A_E, respectively, all information in a network is completely specified. In general, a network's topology and attributes are correlated [12]. When combined with the availability of network data from different sources and the rapid development of tools for their analysis, network analysis has become a cornerstone of machine learning and data mining with applications in diverse research areas, involving such as social [13], biological [14], and spatial [15] networks.

One of the objectives of network analysis is to reveal the local, mesoscopic, and global relationships in the data. At the local level, network structure can be characterized, say, in terms of the degree of a node, which corresponds to the number of links involving the node. At the mesoscale, a commonly studied feature is that of communities [16–18], which can be interpreted as sets of nodes or links with more similarity between them than with other nodes or links in the network. For example, a social network with links arising solely due to shared hobbies between individuals may contain communities of individuals sharing a common hobby. The global structure of the network contains information on distributions of network characteristics such as the degree. Based on suitable measures of network structure at different scales, data mining can lead to valuable information on the important elements and interactions within the network.

In the context of sequence classification, a sequence similarity networks (SSNs) [19, 20] represents each sequence as a node in a network and each pair of nodes is connected with an undirected link representative of the similarity between them. The similarity between any two sequences in the SSN can be incorporated as the link attribute vector based on a suitable choice of similarity attributes. Due to their ability to simultaneously incorporate and present sequence similarity information for multiple sequences, SSNs have become an important tool for sequence classification [21–26]. The hypotheses behind the use of SSNs for sequence classification is that communities in the network correspond to sequence "families" with more significant intra-community than inter-community sequence similarity and that these communities are separable.

As each node in an SSN participates in N_v-1 links, the local network structure is identical for each node if link and node attributes are ignored. So, community detection methods that depend on the local structure around each node will fail [16]. An alternative approach is link community detection [27]. In this approach, node communities are defined as groups of similar or related links based on the line graph of the network, i.e, a graph where each node corresponds to a link in the original network [28]. However, all links in an SSN are topologically equivalent. Hence, communities cannot be identified based solely on network topology when link or node attributes are ignored.

Most methods for detecting communities in SSNs make the reasonable assumption that the score or E-val are good measures of sequence similarity and use them as edge attributes for the network. The inclusion of weights breaks the local symmetry of the SSN and makes it possible to analyze it. However, this progress comes with a cost. In order to identify communities, SSN analysis methods typically require a user specified cutoff value for these edge metrics or cutoff values for some method-dependent parameters. With these assumptions and parameters in place, communities can be detected based on approaches such as cut-based identification of connected components [19, 20, 29] where only edges with weights exceeding a single cutoff value of the score or expectation value, applicable to all links, are retained in the network. Alternative methods that depend on flow/random walks on the network (MCL [30]), or network geometry (spectral clustering [31] and weighted graph cluster editing TransClust [32]) are available. Such methods can lead to communities for which the smallest edge weight within the community differs, unlike a fixed value in a single-cut based approach, and this represents a more general strategy for community detection. Not surprisingly, different methods emphasize different aspects of the network and often lead to differences in community membership of sequences [33].

Here we propose a novel method, Attribute Clustering Dependent Communities (ACDC), for community structure detection in protein SSN which utilizes attribute vector clustering and excludes user-tunable parameters. First, new link attribute vectors with pair alignment information as their components are proposed. Rescaling components of **a** link attribute vector reveals a novel systematic variation of sequence similarity within protein superfamilies. Second, the distribution of attribute vectors in space is analyzed in order to identify the number of clusters. Links in the identified clusters are filtered in a systematic manner such that an objective function that quantifies network structure, the link partition density [27], is maximized in order to identify and refine the community structure of the SSN. The use in ACDC of partition density maximization eliminates the need for any user defined cutoff, unlike most methods for network clustering which do need have tunable user selected parameter(s). Third, in order to provide a reference best performance community structure, multidimensional grid searches are performed to identify optimal cutoff attribute vectors that best reproduce the "ground truth" community structure for selected superfamilies. A key outcome of the analysis is the identification of the topology of attribute space and key features relevant to community detection. The method is applied successfully to gold standard datasets for which "ground truth" community structure is known.

2. Methods

2.1. Link attribute selection for community detection

The SSN for a set of sequences is based entirely on pair alignments between them. The E-val for a pair alignment is calculated based on comparison of its score S with scores for sequence alignments between randomly generated sequences. Such random sequences are external to the sequences in the SSN and consequently, the E-val is considered an extrinsic measure of similarity. Here, we discard such extrinsic measures. By comparison, the alignment score, S, is calculated for each pair of sequences in the network and is an intrinsic network attribute. It should be noted that S and E-val are closely related; small E-vals tend to be associated with large values of S. Typically most alignment programs convert scores to bit-scores. In the following bit-scores will be referred to as the score.

Two pair alignments with very similar or identical values of S need not have the same set of values for L_a, N_{id}, N_m, and N_g (S1 Table). Thus, although score is a useful quantity, additional information is necessary in order to differentiate between these two pair alignments. Here, it is proposed that the score be supplemented with three other alignment attributes, i.e., L_a, N_{id}, and N_m. Since $L_a = N_{id} + N_m + N_g$, any two attributes on the right side of the equation can be selected as independent variables. Without any loss of information, N_g can be ignored. Now, pair alignments with very similar scores should be differentiateable based on their attribute vectors (L_a, N_{id}, N_m, S).

Each attribute vector embodies the function $S = \mathrm{S}(L_a, N_{id}, N_m)$ and the magnitude of S depends parametrically on the substitution matrix. In order to identify universal features of a multivariate function, it is often useful to scale the variables by some natural scale factor implicit in the system and consider the functional variation after scaling. This is particularly useful for free energy functions, such as the alignment score[7]. Furthermore, the magnitude of individual variables is typically different and normalizing the data is a good practice. To incorporate these potential benefits, scale factors will be presented in the following and the resultant scaled attribute vector is presented.

For the pair alignment of two sequences, A and B, each with identical sequence lengths, the maximum value of S ($= S_{max}$) is obtained when A and B are identical. As sequence B diverges from A, S is expected to decrease. Thus, S takes on values from 0 to S_{max}, and S_{max} is a reasonable scale factor. In the case of distinct sequence lengths, the largest score is obtained when one of the sequences is a subsequence of the other. The score, S_{min}, then corresponds to the self-alignment score for the shorter sequence or the smaller self-alignment score, if the shorter sequence does not have the smaller self-alignment score. With S_{min} as an appropriate scale factor, the scaled variable, $s_a = S/S_{min}$, then varies between 0 and ≈ 1, with the largest value depending on the amino acid composition and the substitution matrix.

Similarly, for two sequences A and B with different sequence lengths, L_a is typically equal to or less than the length of the smaller sequence, L_{min}, although some values of L_a may exceed L_{min} when gaps are present. Thus, L_{min}, is an appropriate scale factor and the scaled alignment length, $l_a = L_a/L_{min}$, varies between 0 and ≈ 1. Scaled values for the remaining two attributes, N_{id} and N_m, can be obtained by scaling them by L_a. Now they represent the fraction of identical residues, f_{id}, and the fraction of mismatched residues, f_m, of the alignment length. The rescaled set of link attributes is now defined as $\mathrm{A}_{E,S} = \{(l_a, f_{id}, f_m, s_a)_{(i)}\}$ where (i) represents the i^{th} links attribute vector. A useful consequence of this scaling is that any distance between points in attribute space is not dominated by the variation of a numerically large attribute[34]. It should be noted that the scaling proposed here is by no means the only possibility but is nevertheless clearly based on naturally existing scales.

For a perfect alignment, the scaled attribute vector is (1, 1, 0, 1). The deviation of any pair alignment from this reference perfect alignment can be quantified in terms of the Euclidean distance, *d*, between the two attribute vectors which can take values between 0 (for a perfect alignment) and ≈ 2 (for the worst possible alignment). Thus, *d* is a measure of the evolutionary distance between pairs of sequences.

2.2. Cut based grid search for community detection (GridS)

Cut-based methods are among the simplest methods for community detection in weighted networks. The main idea is to partition the network by selecting all link weights that exceed a single preselected cutoff value. For a network where each link is associated with a link-attribute vectors, instead of a scalar weight, a cutoff value for each link attribute vector component can be combined to construct a cutoff link attribute vector, $A_{E,c} = (l_{a,c}, f_{id,c}, f_{m,c}, s_{a,c})$, where the subscript c stands for cutoff value. The set of attribute vectors is then partitioned based on $A_{E,c}$ such that only links for which $l_a \geq l_{a,c}$, $f_{id} \geq f_{id,c}$, $f_m \leq f_{m,c}$, and $s_a \geq s_{a,c}$ are retained. Nodes can be connected based on the retained links and naturally group together into communities. An optimal choice of $A_{E,c}$ is such that the resulting communities are in best agreement with the "ground truth" community structure of the network.

A simple brute force search method for identifying an optimal $\{A_{E,c}\}$ is by (a) varying cutoff attribute values along a four multidimensional grid, (b) identifying links corresponding to each cutoff attribute vector, (c) connecting the set of identified links in order to assemble communities in the network, (d) quantifying the agreement of the derived community structure with "ground truth" community structure based on some measure, *Q*, which will be defined later, and (e) selecting the maximum value of *Q*, $Q_{max,n}$ where *n* is the dimensionality of the space being searched. This method will be referred to as GridS in the following

2.3. Attribute clustering dependent communities (ACDC)

2.3.1. Community detection strategy.

Inspired by a cut-based connected component approach to community detection, we propose the following hypotheses:

1. Alignments between sequence pairs with significant sequence similarity (the related sequences) and those between sequence pairs with insignificant sequence similarity (the unrelated sequences) occupy topologically distinct regions in attribute space.
2. The two distinct regions are connected with a transition region in attribute space containing continuously varying attribute vectors.
3. There exists a separatrix that partitions attribute space into two distinct regions, one containing sequence pairs with significant sequence similarity and the other containing the remaining attribute vectors.
4. There exists a function whose maximum value is achieved for the "correct" community structure for the network.

In hypothesis 1, sequence pairs with high sequence similarity implies that they either share common ancestry, i.e., they are homologs, or they have different ancestral sequences but are still similar, i.e. they are analogs. Highly similar sequence pairs tend to have large *S*, L_a and N_{id}, and small N_m values. Unrelated sequences, on the other hand, have alignments with small percentage identities arising from pairs of random sequences or very distant related sequences. Such unrelated sequence pairs tend to have small S, L_a, and N_{id}, and large N_m values. Clearly, these two types of attribute vectors (or their scaled versions) are likely to occupy distinct regions of attribute space and the hypothesis is reasonable.

It is well known that each alignment attribute contributing to the link attribute vector is a continuous variable[35], and hence hypothesis 2 is also reasonable. For a given protein superfamily, the distribution of attribute values may be discrete but that is only a result of the finiteness of the sequence dataset.

Hypothesis 3 is the equivalent of the cut-based connected approach to community detection formulated in terms of attribute space. In the case of a weighted network using S as the link weight, a cutoff value of the link weight is selected such that all links with weights lower than the cutoff value are excluded. By analogy, it is proposed that there exists a separatrix passing through the transition region which separates attribute space into regions containing pair alignments between related or unrelated sequences.

Hypothesis 4 states a general approach towards community detection according to which there exists an objective function that implicitly or explicitly incorporates some property of community structure such that its extrema correspond to a good community structure of the network [36, 37]. A number of quality functions exist, each optimal for specific contexts. An ideal feature of these methods is that a uniform cutoff scheme, as explicit in the Separatrix of Hypothesis 3, is now replaced with an implicit variable cutoff scheme where each community can have a different cutoff attribute vector. Thus, more variability of attribute vectors within communities, independent of other communities, can be accommodated.

2.3.2. Clustering and objective function maximization for community detection.

In the context of an SSN, a community of nodes represents a set of sequences that are more similar to each other (i.e., homologous) than to sequences from a different community. Consider two sequences A and B. If A and B belong to different communities, they are expected to have a much smaller measure of sequence similarity than when they belong to the same community. In this case, all pairs of sequences within a community are more similar than all pairs of sequences from distinct communities. Thus, intra-community pair similarities and inter-community pair similarities define two limiting regimes of sequence similarity.

For communities with more than two nodes, any three, or more, nodes, connected together with links, can be related through transitive homology. As per the transitive property of homology, if sequence A is homologous to sequence B, and sequence B is homologous to sequence C, then sequences A and C are also homologous, even if they do not share significant sequence similarity. If within a community, sequences A and C have diverged sufficiently, their sequence similarity may be indistinguishable from inter-community sequence similarity. On the other hand, they may be sufficiently similar to be indistinguishable from intra-community pair sequence similarity. Thus, all sequences related through transitive homology have pair alignment similarity measures that span a broad range of values intermediate to the two limiting sequence similarity measures for inter- and intra-community pair similarity.

Consider the three sequences A, B and C within a community. Provided the similarity cutoff is large enough to include the most distant homologous sequence pairs in the community, as long as the links between A and B, and B and C are included, the transitive homology relation between them will always place A, B and C within the same community. Thus, a minimal set of links associated with large pair sequence similarity attributes should be sufficient for community detection. Here, it is proposed that in scaled pair alignment attribute space, all such pair attributes represent points that should occupy the same high sequence similarity part of attribute space, and that this is approximately separable from all other similarity attributes. Based on this hypothesis, such a region could be identifiable based on the unsupervised learning method of clustering. The problem of community detection in SSNs is now converted in the ACDC method into the detection of clusters of points in link-attribute space and selecting the cluster with attribute values corresponding to high sequence similarity.

Since a clearly defined hypersurface separating intra- from inter-community links or between clusters, may not exist, it is likely that several links that contribute to the high sequence similarity cluster will need to be excluded in the community detection. The identification of an optimal set of links based on their attribute vectors could be implemented either as a simple multidimensional-cut based method that would select only the links for which attribute values satisfy some constraint. Alternatively, this objective could be implemented as a functional optimization in which the subset of links is selected such that a function that quantifies the community structure of the network based on attribute values is optimized. Unfortunately, for most real networks with known ground truth community structures, no single optimal quality measure of the networks community structure exists[36]. The search for quality measures whose optimal value represents a good community structure is an area of ongoing research [37].

The overall structure of a network community is defined by the number of nodes in it and the distribution of links between them. The most commonly used objective function for community structure optimization is Modularity[16]. Modularity is a measure of how compact the arrangement of nodes in a network community is, as compared to a network with randomly assigned links between the same set of nodes, while keeping the order for each node unchanged. Alternatively, the density of nodes in a community can be quantified by Partition Density[27] which compares the distribution of links within a community with reference structures that correspond to limiting cases of link connectivity, i.e., a chain of links versus a maximally connected clique. Due to its clearer structural interpretation and better performance in link based community detection[27], partition density was selected as the objective function for community structure optimization. For a network with N links and a partition of these links $\{P_1, P_2, \ldots, P_C\}$ into C subsets, let subset P_C have m_C links and n_C nodes. Then, the partition density D is defined as

$$D = \frac{2}{N}\Sigma_C m_C \frac{m_c - (n_C - 1)}{(n_C - 2)(n_C - 1)} \tag{1}$$

The summation is limited to communities with more than two nodes in them. Note that for a single community network, in the limit of a linearly connected chain of nodes, $m_c = n_c$-1 and D is zero. In the limit of a fully connected single community network (as for the unfiltered SNN), $m_c = n_c(n_c$-1)/2 and D is unity. The partition density prefers maximal inter-node connectivity in communities, although this tendency is limited by averaging over all communities.

2.3.3. Community structure refinement. Due to the absence of a universal objective function whose optimization leads to an optimal community structure prediction for the network, any resulting solution is likely to contain communities with erroneously assigned links. In this scenario, there is a need for a procedure for refining the community structure based on the scaled attribute vectors.

While the potential number of errors in community structure prediction is *a priori* unknown, at least two types of error can be anticipated. By the consensus definition of a community, similarity between nodes within a community should exceed similarity between nodes between distinct communities. One possible error arises if this simple criterion is not satisfied. In the following, this error will be referred to as a Type 1 error. Let Z correspond to the similarity measure between two nodes connected by a link and let small values of Z imply highly similar nodes. Given the simple definition of Type 1 errors, the following procedure can be used to account for this error. Let $Z_{\max}(i,i)$ and $Z_{\max}(j,j)$ be the largest values of some pre-selected intra-community similarity measure between pairs of nodes in communities *i* and *j*, respectively. Let $\mathrm{Z}_{\min}(\mathrm{i,j})$ be the smallest value of the inter-community similarity measure between communities *i* and *j*. Then, the presence of a Type 1 error implies that the condition C_1: $\mathrm{Z}_{\min}(\mathrm{i,j})\leq\max[\mathrm{Z}_{\max}(\mathrm{i,i}),\mathrm{Z}_{\max}(\mathrm{j,j})]$ is not satisfied. In this case, the solution is to simply merge communities *i* and *j*. The set of inter-community links to be added to the merged community should ensure that condition C_1 is strictly enforced.

A second type of error may arise if a single "ground truth" community is split into two communities during optimization of the objective function such that condition C_1 is satisfied. Such an error will be referred to as a Type 2 error in the following. A Type 2 error indicates that the optimization converged on a solution for which the final $Z_{\max}(i,i)$ values are smaller than would be required for the two communities to be merged automatically. Alternatively, this may indicate that the link required to merge the split communities does not exist. A solution to Type 2 error is to override condition C_1 and include links with *Z* values up to some preset value, Z_0. In this case, the value of Z_0 should be small enough that this correction does not make the merged community eligible for merger with some other community for which links satisfying C_1 exist. If such a situation arises, the split communities are not merged.

For a network with *M* communities resulting from function optimization, the following implementation of a community refinement procedure to account for Type1 and Type 2 errors is proposed: Select a scale factor, $s_{\mathrm{f}}>1$, and a maximum value of the similarity measure, Z_0.

1. Let *M* be the number of communities. Construct the *M*x*M* matrices $Z_{\max}$ and $Z_{\min}$
2. Identify the largest value of *Z*, Z_{m}, from the set of intra-community links.
3. communities i≠j, if C_1 is false, tag the community pairs and go to Step 4, else Step 5.
4. Merge communities *i≠j*. Add links between communities *i* and *j* for which the similarity measure $Z\leq\max[\mathrm{Z}_{\max}(\mathrm{i,i}),\mathrm{Z}_{\max}(\mathrm{j,j})]$.
5. Let *N* be the new number of communities. Construct *N*x*N* matrices $Z_{\max}$ and $Z_{\min.}$
6. $Z_{\mathrm{m}} \leftarrow s_{\mathrm{f}}\, Z_{\mathrm{m}}$
7. Find left outliers, Z_{L}, from all values of $Z_{\min}$.
8. communities *i≠j*, if $Z\leq\min\{Z_{\mathrm{L}}, Z_{\mathrm{m}}, Z_0\}$, merge communities *i* and *j* with those links. If no new links found, go to Step 11.
9. Evaluate Z_{m}.
10. If $Z_{\mathrm{m}}\geq Z_0$, go to Step 11, else, go to Step 1
11. End

The proposed method starts by testing for Type 1 errors and merging communities, as required. Next, it considers Type 2 errors. Based on the hypothesis that split communities are likely to have inter-community $Z_{\min}$ values that corresponds to a process distinct from the process of formation of two distinct communities, the former is expected to be an outlier in the distribution of the inter-community similarity measure values. Starting with a given community structure, the proposed method searches for pairs of communities for which $Z_{\min}(\mathrm{i,j})\leq Z_{\mathrm{L},}$ where Z_{L} is the lower outlier bound, and merges pairs of communities that are so identified. In the case $\mathrm{Z}_{\mathrm{L}}\leq\mathrm{Z}_{\mathrm{m}}$, the analysis reduces to that for Type 1 errors. If $Z_{\mathrm{L}}\geq Z_{\mathrm{m}}$, the score of the search and merger is extended by scaling Z_{m} by the factor $f_{\mathrm{s}}\geq1$.

As a systematic procedure, a two-dimensional grid in the (s_f, Z_0) plane is considered. For each grid point, the procedure outlined in the previous paragraph is followed and the partition density evaluated for the resultant community structure. The grid point that leads to the largest value of partition density is selected and the corresponding community structure is selected as the final community structure for the network. While this refinement procedure is expected to improve the agreement with the "ground truth" community structure, it may still not lead to elimination of all split communities.

A third type of error may arise when two "ground truth" communities are merged during optimization of the objective function. While it is tempting to consider such merged communities as equivalent to overlapping communities, the presence of additional links between them may complicate the splitting of such merged communities. Such splitting of merged communities will not be considered here.

2.4. Data sets, sequence similarity network construction, and visualization

The gold standard (GS) protein sequence dataset of Brown *et al.* [38] with known "ground truth" community structure was selected for sequence similarity network construction. This dataset consists of 866 protein sequences from 91 families (or communities) that belong to five mechanistically diverse enzyme superfamilies. The classification into communities is based on experimental information on the reactions catalyzed by these enzymes. The GS dataset is considered in its entirety as well as split into five subsets, each consisting of a single constituent superfamily from the GS dataset. The number of sequences and communities in each superfamily are summarized in Table 1. Validating ACDC against these six GS datasets then tests its ability to distinguish between families within superfamilies as well as in a mixture of superfamilies. Recently, NCBI has replaced GI numbers with the corresponding Accession number to refer to individual sequences. As the Gold Standard sequence dataset identifies sequences by the GI number, we have included the mapping of each GI number to the corresponding NCBI accession number as S2 Table.

In order to identify the region of attribute space occupied by potential evolutionarily unrelated sequences with insignificant sequence similarity, 10000 randomly generated sequences, each with 100 residues, were constructed. Using HMMscan in the HMMER3 suite of programs [39], all sequences were searched against the library of Pfam [22] profile hidden Markov models for potential domains using the default gathering threshold for each profile hidden Markov model. Sequences with any domain matches were removed. For the remaining sequences, an all-by-all pair sequence comparison was performed. Pairs of sequences that share at most 15–20% identity over less than 75 residues were identified. Only 19 sequences were found and these constitute the R15 dataset. Given the small percentage identity in R15, they are expected to be unrelated or at least very distantly related. As a result, it is reasonable to assume that each of these sequences constitutes a separate community/protein family.

The numbers of sequences and families in each superfamily from the selected dataset are listed in Table 1. The Amidohydrolase (AH) and Vicinyl Oxygen Chelatase (VOC) superfamily were selected at random from the GS dataset for learning the structure of attribute space and benchmarking the community detection method. The remaining datasets were used to perform the actual evaluation of the community detection method.

An SSN was constructed for each dataset by using each sequence in the dataset as a query sequence and all sequences in the dataset as the target database. The sequence similarity attributes were calculated with the Smith-Waterman [40] method for optimal alignment detection as implemented in the SSEARCH program that is part of the FASTA suite of programs [41].

Table 1. Sequence datasets: Network properties, number of communities, and their use in method development.

Dataset	#Sequences	#Links	#Communities	Usage
Amidohydrolase	*232*	*26796*	*29*	*Benchmark*
Crotonase	*91*	*4095*	*16*	*Validation*
Enolase	*285*	*40470*	*9*	*Validation*
Haloacid Dehalogenase	*125*	*7750*	*20*	*Validation*
Vicinyl Oxygen Chelatase	*113*	*6328*	*17*	*Benchmark*
Gold Standard	*866*	*374545*	*91*	*Validation*

https://doi.org/10.1371/journal.pone.0178650.t001

The default amino-acid substitution matrix used by SSEARCH, BLOSUM50, was selected along with the default gap open and extension penalties. Since protein sequences often contain repeats and composition biases, the program segmasker in the NCBI BLAST suite [9] was used to mask these regions in all protein sequences prior to their pair alignment. For each pair of sequences in the dataset, the pair alignment information was stored in BLAST tabular format. In general, for a pair of sequences A and B, selection of A as the query and B as the library sequence during pair alignment typically leads to a different alignment from that of B as the query and A as the library sequence. To account for this asymmetry, the highest scoring pair alignment was selected to represent the pair. In this manner, all pair alignments were collected for each protein superfamily.

Alignment metrics for all pair alignments in each selected groups of clusters were collated into a tab separated table. The network table was imported into Cytoscape 3.2.1 [42], a versatile network visualization and analysis application. The network was visualized with the Yfiles-Organic Layout, which has been demonstrated to group nodes connected by edges representing large percentage identities close together in space [19].

2.5. Data processing for community detection

2.5.1. Cut based grid search. An important factor in grid searches is the increment in each component value between neighboring grid points. For a fine-grained grid, performing a full grid search calculation is computationally prohibitive. On the other hand, a coarse-grained grid may miss important set of cutoffs. A compromise used here is to use a fine grid with a spacing of 0.01 along three dimensional subspaces, selecting a subset of best performing grid points, and then extending the grid search with a spacing of 0.01 along the fourth dimension.

Since the attribute space is four dimensional, all combinations of three dimensional subspaces $\{(l_a, f_{id}, f_m), (l_a, f_{id}, s_a), (l_a, f_m, s_a), (s_a, f_{id}, f_m)\}$ are selected. Grid searches are performed and Q is calculated for all cutoff attribute vectors $\{(l_{a,c}, f_{id,c}, f_{m,c}), (l_{a,c}, f_{id,c}, s_{a,c}), (l_{a,c}, f_{m,c}, s_{a,c}), (s_{a,c}, f_{id,c}, f_{m,c})\}$. The largest value of Q from the subspace search, $Q_{max,3}$, is identified and the set of cutoff attribute vectors for which $Q > = 0.9\ Q_{max,3}$ is shortlisted. The scale factor of 0.90 is selected in order to limit the fourth dimensional search to the region most likely to contain large or larger values of Q. For each shortlisted cutoff attribute vector, a one-dimensional scan of the missing fourth dimension is performed using a grid spacing of 0.01. Only cutoff attribute vectors for which $Q \geq 0.90 Q_{max,3}$ are saved. These cutoff attribute vectors approximately represent an optimal hypersurface which brackets the part of attribute space required for optimal community detection. Knowledge of this hypersurface as well as $Q_{max,4}$ values provides a useful reference for comparing the performance of community detection methods.

2.5.2. Attribute clustering dependent communities. To identify clusters in attribute space, the implementation of k-means clustering in MATLAB (R2015a, The MathWorks Inc., Natick, MA, 2000) was selected for cluster detection. The optimal number of clusters in attribute space for each superfamily is *a priori* unknown. Since higher-dimensional attribute vectors cannot be visualized easily, Principal Component Analysis (PCA) [43] is performed using MATLAB to identify linear combinations of attribute components along which the variation of attribute values is largest. For the benchmark datasets, the two principal components along which the attribute space shows the largest variation were identified and points in attribute space projected onto them. A first estimate of the number of clusters in attribute space can be obtained based on visual inspection of this projection. Clustering methods may also perform better on the transformed data than raw data even if the number of dimensions is not reduced. The Euclidean distances between points in attribute space were clustered using *k*-means after projecting each point along the principal directions.

The optimal number of clusters is evaluated with the implementation of the Calinski-Harabasz (CH) method [44] in MATLAB and was called through the evalcluster function. The number of clusters is varied between one and 10. For each selected number of clusters, the CH coefficient is calculated. The optimal number of clusters (q) is selected as the number of clusters for which the variation of the CH clusters showed an elbow. If the optimal number is 8 or 9, the maximum number of clusters was increased to 15 in order to verify if an alternative solution existed beyond 10 clusters.

Since *k*-means clustering can result in a partition that is a local minimum, a total of 100 replicas of clustering were performed on each data set in order to improve the sampling for finding an optimal clustering. It was found that q varied between 2 and 9 for all datasets as a result of diversity of pair attribute variation with superfamilies, even if visual inspection broadly identified 2 or 3 clusters. In order to achieve a consensus between k-means and visual inspection based number of clusters, a coarse-graining procedure is applied. For each of the q clusters, the mean value of each component of the attribute vector was calculated. The resulting mean attribute vectors for all clusters were clustered into two groups. The group with the largest mean attribute vectors was selected as representative of Region 1 for community detection. Since k-means clustering is unaware of any Separatrix that may demarcate Region 1 from the remaining data, some clusters that may belong to Region 1 are excluded if the they are closer to clusters from the intermediate region. Such errors are corrected by using the benchmark datasets and identifying the smallest mean attribute cutoff vector, $A_{\mathrm{E,m}}$, above which most attribute vectors in the cluster contribute to Region 1.

Another potential shortcoming of the *k*-means clustering method is that some data points which are well separated from a cluster may be incorrectly assigned to the cluster if the distance of these points from the cluster is smaller than the distance from any other cluster. In such cases, the incorrectly assigned points are outliers that should be excluded. In order to identify outliers, the value of a_{l} for all links contributing to Region 1 are collected and sorted in order of decreasing magnitude. Based on an implementation of a modified boxplot, the range of a_{l} values that correspond to outliers is identified. All links with outlier values of a_{l} are excluded and the remaining links are sorted based on the value of d. The resultant set of links, free from outliers, is then used for community detection. Partition density minimization based on the resulting set of links representative of Region 1 should then lead to a reasonable first approximation for the number of communities. As there is no need for any user input during this stage of data analysis, the outlined procedure for community detection is entirely unsupervised. Inspection of data at each stage is nonetheless prudent.

Since the scaled attribute vectors are four-dimensional objects for any sequence dataset, it would be beneficial to combine them into a composite variable whose variation reflects the collective variation of the attribute vectors. The distance d from a perfect alignments attribute vector is selected as such a composite variable. The community structure refinement procedure is applied to the community structure obtained by Partition density maximization. For this purpose, a simple grid in the (s_f, Z_0) plane in the range $1.01 \leq s_f \leq 1.10$ and $1.100 \leq Z_0 \leq 1.110$ and grid point separation of $\Delta s_f = 0.01$ and $\Delta Z_0 = 0.001$ is constructed. Partition density maximization over the selected set of grid points is then used to identify the final refined community structure. To ensure that sufficiently large inter-community d_{min} values are considered, a final value of $Z_m > 1$ is enforced and this excludes some grid points from consideration. The community structure refinement procedure is entirely unsupervised, making the entire implementation of ACDC an unsupervised method.

2.5.3. Comparison of predicted and ground-truth community structures. In order to quantitatively evaluate the performance of the ACDC method, the predicted community structure should be compared with the "ground truth" communities in the datasets. One measure of performance is simply comparing the number of communities. However, this does not provide any information on the similarity of predicted and "ground truth" communities. A number of measures of cluster similarity are available which compare the internal or external properties of clusters. Here, we utilize the *F*-measure [45, 46], which is a composite index based on the Precision and Recall of the communities.

Let *N* be the number of sequences, n_i the number of sequences in community *i* within the "ground truth" network, m_j the number of sequences in community *j* based on the selected method, and let O_{ij} be the number of sequences common to ground truth community *i* and identified community *j*. The precision (*P*) of cluster *j* is a measure of the fraction of sequences in the "ground truth" community that is correctly detected by the method and is evaluated as $P_{ij} = O_{ij}/n_i$. Similarly, recall (*R*) is defined as the fraction of sequences common to the detected and "ground truth" communities that is present in community *j*, and is defined as $R_{ij} = O_{ij}/m_j$. The *F*-measure then combines Precision and Recall into a single measure defined as

$$F = \frac{1}{N}\Sigma_i n_i \left[max_j \left\{ \frac{2}{\frac{1}{R_{ij}} + \frac{1}{P_{ij}}} \right\} \right] \tag{2}$$

For a perfect community detection method, $P = R = 1$ and the corresponding *F*-measure is also 1. In the other extreme, in the case of complete failure when every community has exactly one sequence, both Precision and Recall are small, but not zero, and the corresponding F-measure is also small, but not zero. Note that a low F-measure does not necessarily imply bad performance, rather it indicates a harder task [47]. Here, the F-measure was calculated for the detected communities structure and compared to the "ground truth" communities for each sequence dataset.

2.5.4 Outlier detection. A computationally simple method for outlier detection from a univariate sampling of data is the boxplot. In its simplest form this requires the calculation of the sample median(m), the first (q_1) and third (q_3) quartiles. As proposed by Tukey [48], for the set of observations $X = \{x_1, x_2, \ldots, x_n\}$, where all data is sorted in increasing order of value, data points outside the range [q_1-1.5(q_3-q_1), q_1+1.5(q_3-q_1)] are selected as outliers. For data sampled from a normal distribution, this range corresponds to a probability of 0.007 that a randomly selected data point will be outside the selected range. The resulting outliers roughly correspond to 0.7% of the data. However, this method performs well only for distributions that are mostly symmetrical within the inter-quartile region. There is no *a priori* reason to make such an assumption for sequence similarity data. Instead, the medcouple (mc)-based outlier detection method proposed by Dovoedo [49] is selected. The medcouple is a measure of skewness of the data[50] and is defined as

$$mc = \underset{x_i \leq q_2 \leq x_j}{\overset{med}{}} h(x_i, x_j) \tag{3}$$

where h(x_i,x_j) is a kernel function and med corresponds to the median value of $h(x_i,x_j)$. For $x_i \neq x_j$,

$$h(x_i, x_j) = \frac{(x_j - q_2) - (q_2 - x_i)}{(x_j - x_i)} \tag{4}$$

For $x_i = x_j = q_2$, let there be k observations m_1 to m_k that are equal to the sample median, then

$$h(x_j - x_i) = \begin{cases} -1 & i + j - 1 < k \\ 0 & i + j - 1 = k \\ 1 & i + j - 1 > k \end{cases} \tag{5}$$

For right-skewed sample distributions, mc >0, for left-skewed distributions, mc<0, and for symmetric distributions, mc = 0. Building on the medcouple based outlier detection method of Hubert *et. al*[51], Dovoedo [49] proposed that points outside the range [q_2-4e$^{(-1.93\ mc)}$(q_2-q_1): q_2+4e$^{(2.18\ mc)}$(q_3-q_2)] should be considered to be outliers. The probability that a randomly selected data point lies outside the fences is maintained at 0.007. The stability of this proposed outlier detection method for skewed distributions and its simple implementation makes this method suitable for automated outlier detection.

3. Results

3.1. Attribute vectors reveal systematic variation of alignment attributes

A key difference between other SSN analysis methods and ACDC is the use in the latter of a new scaled attribute vector. The question thus arises as to whether the introduction of this new attribute improves the classification of pair alignments? A good selection of alignment attributes should (a) have enough variables to distinguish between different pair alignments, (b) show features in their distributions that allow for clear separation of the attribute vectors into regions that could be interpreted as pair alignments between homologous or unrelated sequence pairs, and (c) preferably reveal systematic variation in the data. To address these issues, the frequency distribution P(S) of alignment scores, scatter plots of the unscaled pair alignment attributes in the (L_a, S) plane and in (L_a, P_{id}, S) space (here P_{id} is the percentage of identical residues), as well as the scaled pair alignment attributes in (l_a, f_{id}, s_a) space were calculated. For all these choices of attributes, data based on intra-community links was also included in order to understand the ability of the selected attributes to distinguish between intra- and inter-community links. Results for the AH superfamily are presented in Fig 1 while data for all other superfamilies is presented in S1 Fig.

The frequency distribution of alignment scores is one of the simplest summaries of alignment information and is shown in Fig 1a for the AH superfamily. The distribution is peaked at small values of the alignment score, S, and shows at least three sub-populations. Since two alignments with the same S may be indistinguishable, it is not clear how insightful P(S) might be. Further, the small S region (at ~ 0<S<100) and the intermediate peak (at 100<S<700) overlap and separating the distributions into distinct domains is not straightforward. Another consequence of this overlap is that the intra-community alignment scores cannot be separated from the inter-community alignment scores.

To remedy this situation, increasing the dimensionality of the attribute data is one option. As an example, a projection of attribute vectors onto the (L_a, S) plane is presented in 1(b). Now, there is improvement in separation of data into distinct regions with respect to the distribution of scores. Thus, the selection of these two attributes helps distinguish intra- from inter-

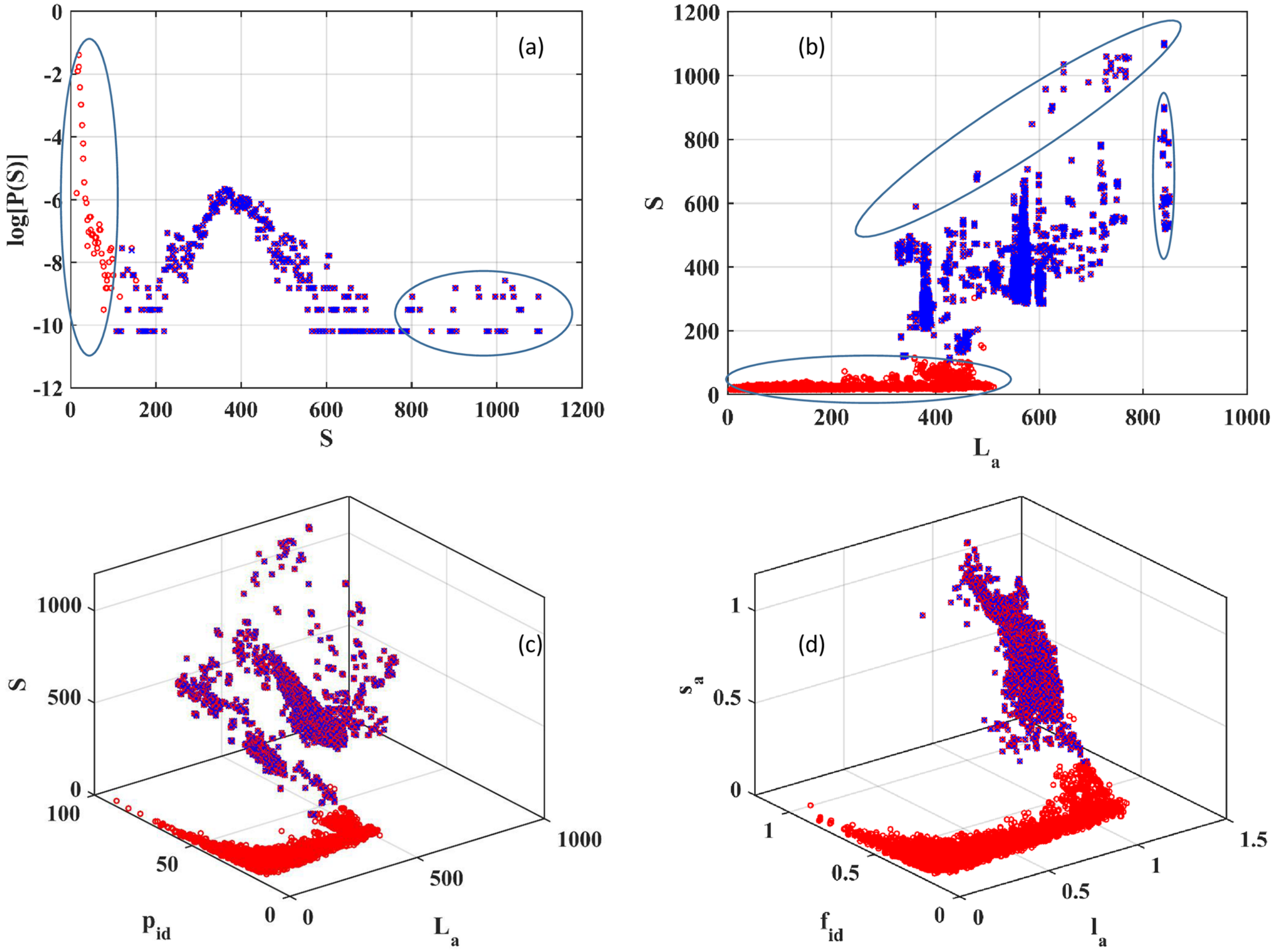

Fig 1. Alignment attributes from the Amidohydrolase superfamily can be represented in several ways. (a) Frequency distribution of scores (S), (b) Scatter plot of attribute vectors in the (L_a, S) plane, (c) variation of attribute vectors in (L_a, p_{id}, S) space, and (d) the scaled attribute vectors in (l_a, f_{id}, s_a) space.

https://doi.org/10.1371/journal.pone.0178650.g001

community pair sequence alignments. At least four clusters of points can be discerned, shown in Fig 1b. There is no systematic variation between the clusters. Note that only the cluster with the smallest S values contains most of the inter-community attributes.

By going to three dimensions, a scatter plot of attributes in the (L_a, P_{id}, S) space leads to a more complicated distribution of points, shown in Fig 1c. Now, a large number of clusters is visible although no systematic overall trends are apparent, except for the increase in L_a at almost constant S for small values of L_a, a feature present in Fig 1b but not Fig 1a. A second feature that is apparent at small S and L_a is the increase in P_{id} which is weakly dependent on S and L_a. This feature cannot be discerned in Fig 1b. The introduction of scaled attributes transforms Fig 1c into Fig 1d where, instead of distinct clusters, there is a clear continuous variation of alignment attributes in attribute space. The separation of intra- from inter-community attributes is no worse than the unscaled attributes in three dimensions. The same overall improvement when using scaled attributes is apparent for all other gold standard datasets (see S1 Fig).

Thus, attribute scaling brings out the hidden systematic variation of alignment attributes within a superfamily which would be missed with any of the other selections of attribute vectors considered here.

3.2. Overall distribution of points in attribute space is superfamily independent

In order for the variation of points in attribute space, as in Fig 1d, to be useful for partitioning the SSN, it is important to establish the generality of the variation of pair alignments in attribute space. Since the scatter plots reveal a large variation, it would be useful to coarse-grain attribute space in order to identify a less noisy trend. A local coarse-grained representation of the populated attribute space can be obtained by averaging the alignment attributes for neighboring points. To do so, a set of neighboring points has to be found. As it turns out, k-means clustering works by collecting points into clusters such that points within the cluster are closer to each other than to points in different clusters. With an appropriate selection of the number of clusters, a small number of points that are representative of the local features of the attribute space can be identified. The number of points in each cluster was selected to be approximately 100, except for the R15 dataset (see Methods-Data Sets) where all points are included. This represents a choice that is expected to be small enough that the averages represent local features and large enough that the averages are well behaved. The average attribute vector for each cluster was calculated and cluster indices ordered such that the scaled scores are in ascending order for the clusters.

In Fig 2, the variation of the coarse-grained attribute vectors is presented for the R15, VOC (Vicinyl Oxygen Chelatase), and AH datasets. The variation of points in (l_a, f_{id}, s_a) and (l_a, f_m, s_a) space clearly shows the presence of two distinct limiting regions, labeled Region 1 and Region 2 in the figure. Region 1 contains attribute vectors that could be thought of as starting from the highest attribute values ($\approx$1) in Fig 2a, except for f_m which starts from its lowest values ($\approx$0) in Fig 2b. As s_a decreases (rightmost points in Fig 2a and 2b), l_a values tend to stay mostly unchanged until s_a reaches a value smaller than $\approx$0.2 in Region 1. Given that $l_a \approx 1$ in Region 1, such attribute vectors correspond to pair alignments that roughly cover the entire length of the shorter sequence in the pair alignment, akin to a global alignment. As the number of mutations increases over the entire length of the sequence, the alignment score, and therefore s_a, decreases, while f_{id} decreases to about 0.3 and f_m increases to about 0.6. Thus, alignments in Region 1 correspond to sequence pairs that have not diverged significantly and are likely to be homologous.

For s_a smaller than 0.2, l_a starts to decrease from a value of about 1 in Region 1 to much smaller values in the intermediate region. The decrease in l_a implies that sequence pairs can be aligned only over small fractions of their sequence lengths. Along with this variation in l_a, f_{id} gradually decreases from 0.3 to 0.2 (see Fig 2c) and f_m gradually increases from $\approx$0.6 to $\approx$0.8 (see Fig 2d) between Regions 1 and 2. Thus, the number of mismatched residues increases over the transition region. Together, these observations suggest that alignments in the intermediate region, going from Region 1 to Region 2, correspond to increasingly divergent pairs of sequences and may represent inter-community pair alignments or perhaps alignments between highly divergent sequence pairs within the same community.

In Region 2, s_a and l_a stay small, f_{id} starts increasing and f_m starts decreasing from its typical values in the intermediate region. The large values of percentage identity and small alignment lengths seen for Region 2 have previously been demonstrated to exist for non-homologous sequence pairs, for example, by Rost[35]. Thus, $f_{id} = 0.3$ or a percentage identity of 30% between two sequences need not correspond to homologous sequence pairs if the attribute

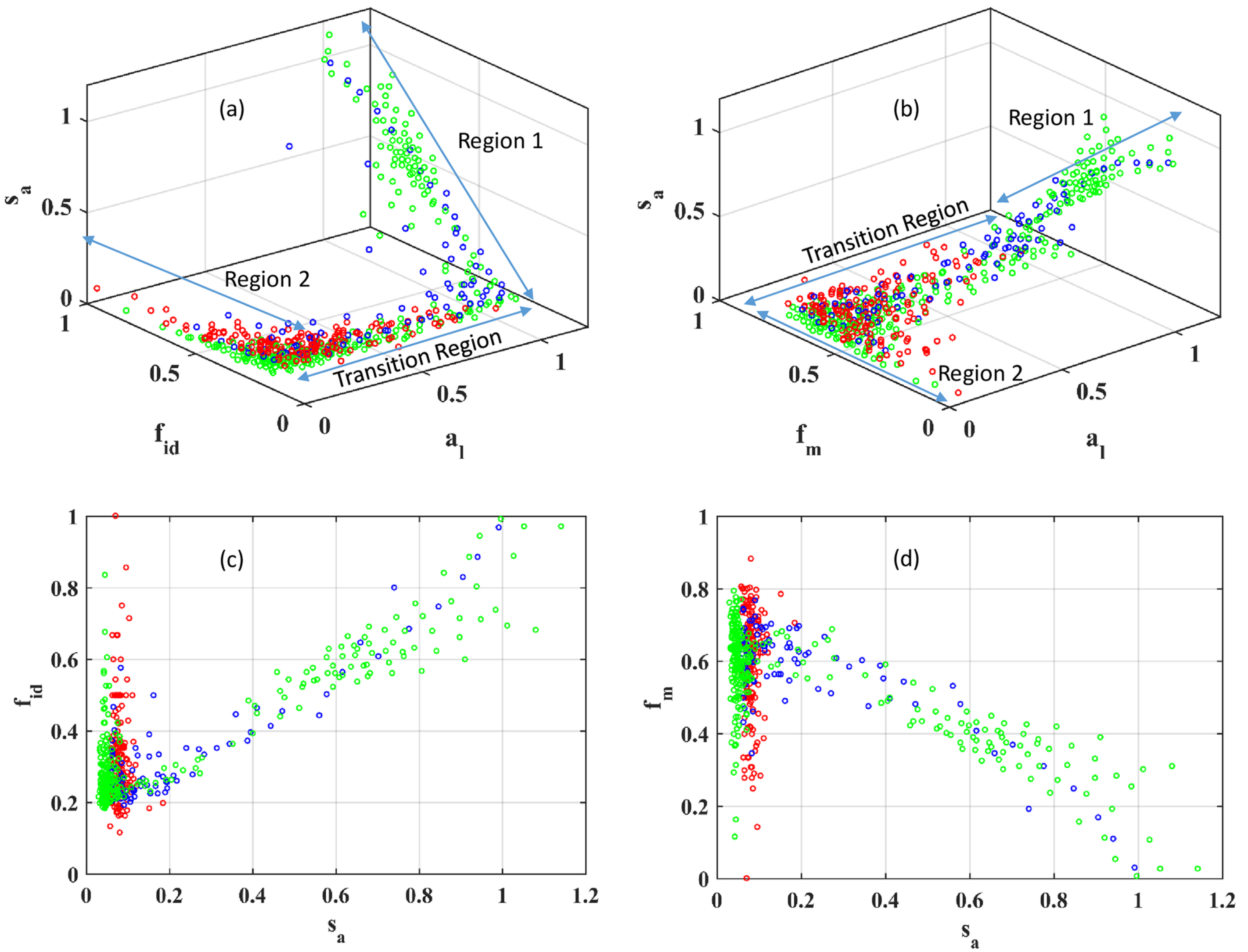

Fig 2. Distribution of attribute vectors in attribute space partition is superfamily independent. (a) in (l_a, f_{id}, s_a) space, (b) in (l_a, f_m, s_a) space, (c) in (f_{id}, s_a) plane, and (d) in (f_m, s_a) plane for the R15 dataset (in red), Amidohydrolase superfamily (in green), and Vicinyl Oxygen Chelatase superfamily (in blue).

https://doi.org/10.1371/journal.pone.0178650.g002

vector lies in Region 2. For the R15 dataset, pair alignments primarily occupy Region 2 and part of the intermediate region. Since R15 contains sequences with very low sequence similarity, it is reasonable to propose that Region 2 contains pair alignments between unrelated sequences or very distantly related sequence pairs.

Attribute vectors for the AH and VOC datasets mostly follow the overall variation outlined in the preceding paragraphs although there is some superfamily dependent variation. It is anticipated that for other protein superfamilies, additional parts of attribute space may be available and the distribution of points in attribute space may differ from the pathway presented here. Such deviations may be indicative of a different evolutionary history for sequences in the superfamily.

The overall presence of two regions of different variation of attributes and an intermediate region between them in Fig 2a and 2b suggests a gradual transition in topology of attribute space between regions corresponding to significant and insignificant sequence similarity.

Thus, coarse-graining of attribute vectors clearly demonstrates a transition path in attribute space and the changes in attributes along the pathway. To the best of the author's knowledge, the overall variation in attribute space has never been presented before in any community detection method for any sequence dataset.

3.3 Detecting communities with ACDC

Having identified a scaled attribute vector representing pair alignments, the overall distribution of points in attribute space and the potential implications of different regions in attribute space, now the structure of attribute space is used to detect communities. In ACDC, the first step is to cluster all attribute vectors in the 4-dimensional space. The optimal number of clusters was identified using the variation of the Calinski-Harabasz coefficient (S2 Fig) and the distribution of points in each cluster is presented for the AH and VOC superfamilies in Fig 3a and 3b respectively. The differences in the distribution of points in attribute space leads to different numbers of clusters. The number of clusters contributing to Region 1 for the AH and VOC superfamiles are 1, Cluster 3 in Fig 3a, and 3, Clusters 1 and 3 in Fig 3b, respectively. There is no spatial demarcation of these clusters from the adjoining cluster, cluster 1 for AH and cluster 2 for VOC. Such a variation is expected given the continuity of attribute values.

The second step of ACDC involves identifying the mean attribute value for each cluster followed by clustering these mean attribute vectors. The mean cluster values are presented in Fig 3c and 3d for the AH and VOC superfamilies respectively. For the AH superfamily, there is clear separation between the large mean s_a valued clusters and the other two clusters. In the case of VOC, clustering of the mean attribute vectors leads to two clusters as shown in Fig 1d. One cluster, Cluster 1, in Fig 3d, is a natural constituent of Region 1 and contains the largest s_a attribute vectors. Of the remaining attribute clusters, one (indicated with an arrow in Fig 3d), has a large value of the mean s_a, but is included in the second cluster of mean attribute vectors. As discussed in the Methods section, imposing a lower cutoff on the mean attribute vector can lead to better assignment. For VOC, a cutoff of $s_a = 0.41$ correctly assigns this erroneously assigned cluster to Region 1. However, to make it more generally applicable, the minimum mean attribute vector $A_{E,m}$, was selected such that it satisfies the condition $s_a \geq 0.30$. The value of 0.30 is selected since (a) most attribute vectors in the intermediate region have $s_a > 0.24$ and (b) the change in variation from Region 1 to the intermediate region starts at $s_a \approx 0.30$.

Following the selection of attribute space clusters for Region 1, the l_a values for all selected attribute vectors are collected and outlier analysis performed in order to finalize the set of attribute vectors to be analyzed in the next stage. As clear in both Fig 3a and 3b, the distribution of points in Region 1 is broad and has disconnected regions that clearly contain outliers. Links corresponding to these outliers are removed (see Methods) and the shortlisted attribute vectors are considered for the optimization of D. Analysis of these shortlisted attribute vectors in 4-dimensional space is simplified by introducing the collective variable, d, which measures the distance of each attribute vector from a perfect alignments attribute vector. The list of attribute vectors is sorted in increasing order of d. For each cut-off value of d, all links with smaller d values are collected, the community structure identified and the magnitude of D evaluated.

The variation of D as a function of d is presented in Fig 4a and 4b for the AH and VOC superfamilies. For both superfamilies, D(d) shows a non-monotonic variation with d with large fluctuations occurring associated with small changes in d. Links with similar values of d when added to different communities may change the shape of these communities, possibly by merger of communities, thereby leading to these large fluctuations. Since the objective is to optimize community structure, the variation with d is not as important as the community structure obtained by maximization of D.

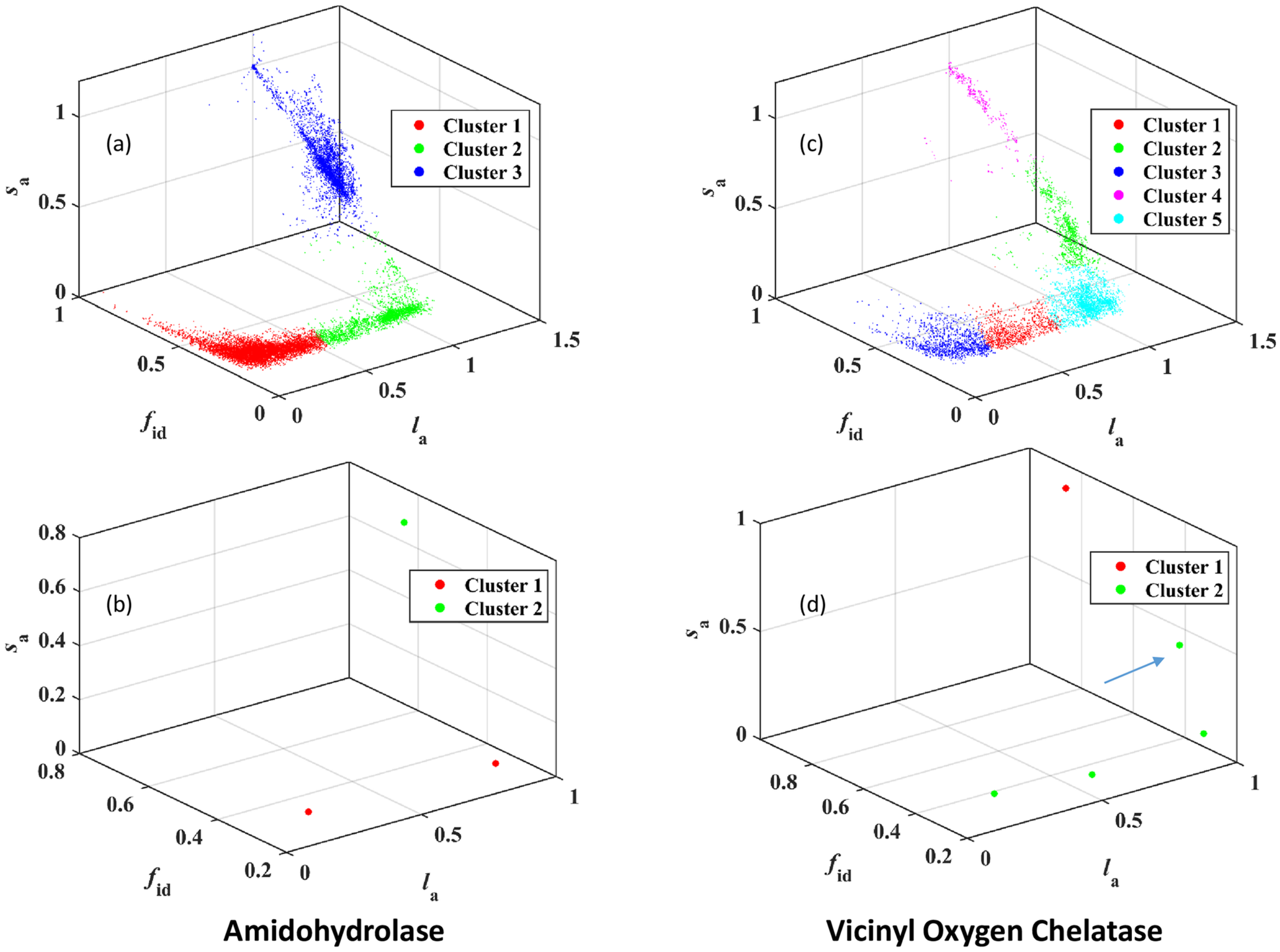

Fig 3. Partition of attribute vectors and mean attribute vectors into clusters by k-means clustering. (a) Cluster assignment for each attribute vector from the Amidohydrolase superfamily, (b) Assignment of each clusters mean attribute vector for the Amidohydrolase superfamily to Region 1 or otherwise, (c) Cluster assignment for each attribute vectors from the Vicinyl Oxygen Chelatase superfamily, and (d) Assignment of cluster mean attribute vector for the Vicinyl Oxygen Chelatase superfamily. Clusters with highest s_a values exceeding 0.3 are included in Region 1 for both superfamiles. An error in cluster assignment to Region 1 is indicated in (d) with an arrow and is corrected in the analysis.

https://doi.org/10.1371/journal.pone.0178650.g003

Defining d_{min} to be the minimum inter-community value of d for a pair of communities, the last step of community refinement in ACDC is based on the distribution of d and includes contributions from all pairs of communities present in the network. After the optimization of D, the distribution of inter-community d_{min} values is calculated and the results are presented in Fig 4c and 4e for the AH and VOC superfamilies, respectively. Both figures contain a broad distribution centered at $d_{min}{\approx}1.5$ and contains a small number of points at small values. If the broad distribution were to be interpreted as arising due to some unspecified evolutionary process, the small d_{min} components would be thought of as outliers with respect to the distribution. These outliers are selected (see Methods), and the corresponding communities are merged iteratively. After the merger of communities, the resulting distribution of d_{min} values is presented in Fig 4d and 4f. As expected, the low d_{min} outliers are eliminated. The broad distribution on the other hand becomes less noisy and can be approximated, to first order, as a

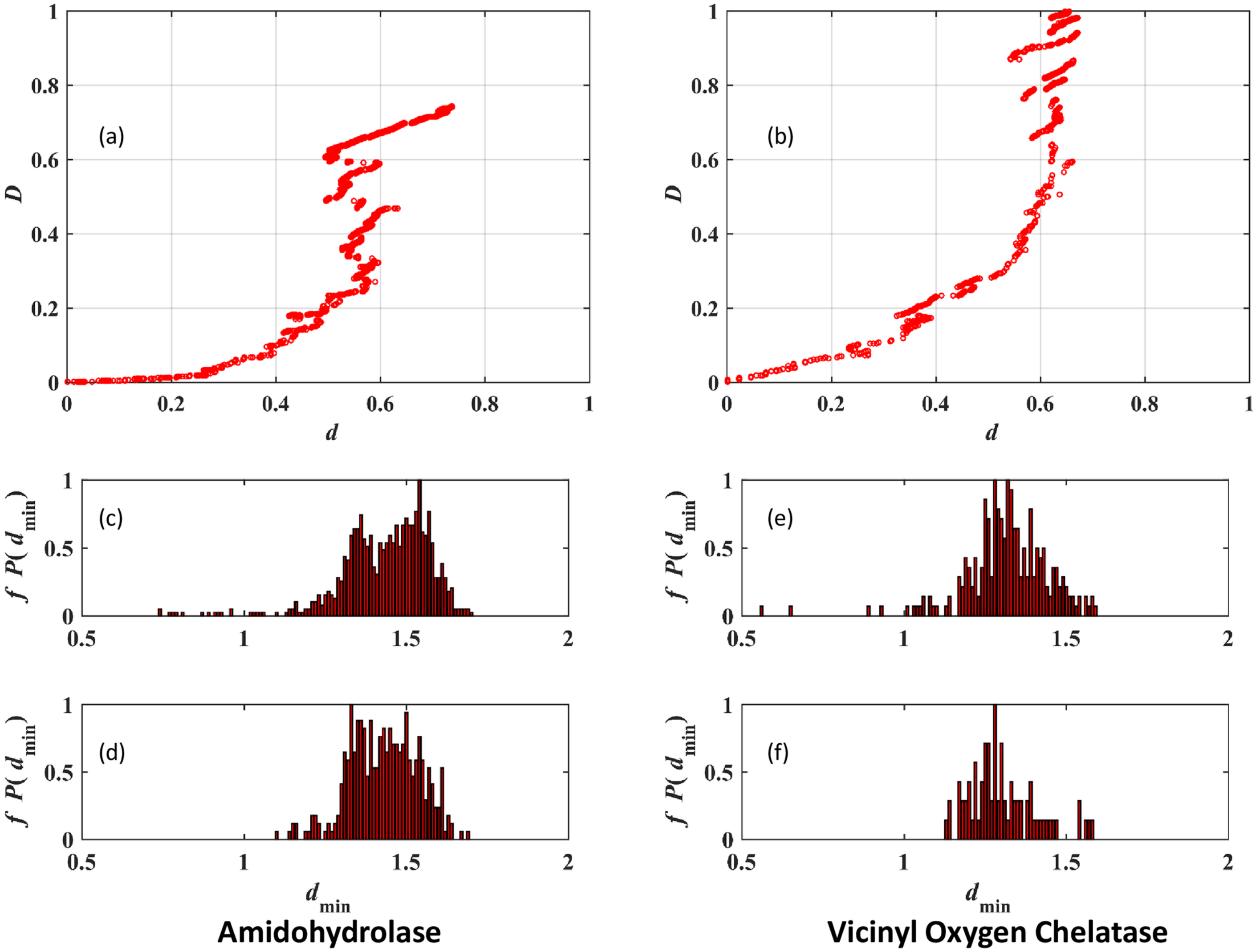

Fig 4. Community structure optimization and refinement with ACDC. (a) Optimization of partition density (*D*) with distance *d* from the perfect alignments attribute vector for the Amidohydrolase superfamily, (b) Partition density (D) versus distance *d* for the Vicinyl Oxygen Chelatase superfamily, (c) the distribution of inter-community d_{min} before community refinement and (d) after community refinement for the Amidohydrolase, (e) the distribution of inter-community d_{min} before community refinement, and (f) after community refinement for the Vicinyl Oxygen Chelatase superfamily. The distributions have been scaled such that the bin with the largest frequency has a value of 1.

https://doi.org/10.1371/journal.pone.0178650.g004

Gaussian distribution. A few large d_{min} value outliers are apparent, particularly for the VOC superfamily, and they perhaps represent the most divergent pairs of communities within the superfamilies.

The community structures for the AH and VOC SSNs predicted by ACDC are presented in Fig 5. In the figure, each node is connected by a link whose color is an indicator of the value of *d* for that link. Maximization of D invariably leads to highly connected communities, and this is indeed the case. Some communities have irregular shapes but these are likely a consequence of the pair similarity. An example of such a community is shown in Fig 5b (top row rightmost) where the community appears to be formed by merger of two communities; however, these correspond to isofunctional sequences from the same community.

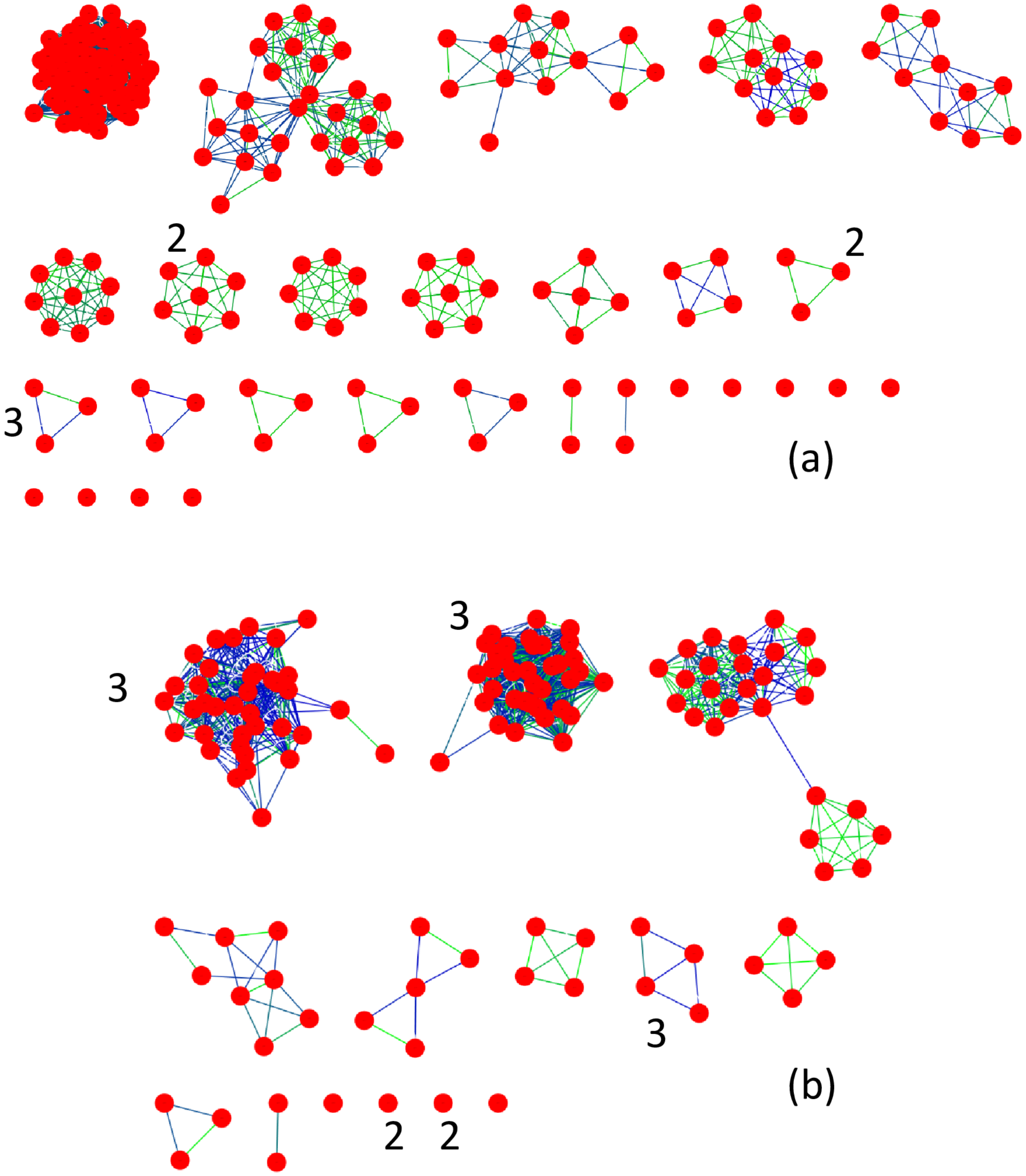

Fig 5. Community structure based on ACDC and the resulting error types. (a) Amidohydrolase superfamily community structure and (b) Vicinyl Oxygen Chelatase superfamily community structure. Each node (red) corresponds to a sequence. The set of links obtained with the ACDC method are also shown in increasing order of d from 0 (green) to the maximum d value (blue). Also, shown are the type of errors associated with communities (if any). Communities with Type 2 errors are merged in the "ground truth" community structure. Type 3 communities are formed by merger of independent "ground truth" communities.

https://doi.org/10.1371/journal.pone.0178650.g005

Next, the community structures predicted by ACDC are compared with the ground truth community structures based on the number of communities and the F-measure. The results for all datasets are summarized in Table 2 and the intermediate data are presented in S2 Fig. The F-measure values for most datasets are around 0.9 or higher, which is indicative of good

Table 2. Performance evaluation of the ACDC and GridS methods.

	Ground Truth	ACDC		GridS	
Dataset	**#Communities**	**#Communities**	**F-measure**	**#Communities**	**F-measure**
Amidohydrolase	29	28	0.9859	29	0.9877
Crotonase	16	16	0.8474	14	0.9741
Enolase	9	10	0.9684	9	0.9797
Haloacid Dehalogenase	20	23	0.9825	20	1
Vicinyl Oxygen Chelatase	17	14	0.7821	17	0.8315
Gold Standard	91	86	0.9368	86	0.9484

https://doi.org/10.1371/journal.pone.0178650.t002

community detection. ACDC predicts a smaller number of communities for the Amidohydrolase, Vicinyl Oxygen Chelatase, and Gold Standard datasets, which is likely a consequence of merger of communities that ought to be distinct. The SSN for the Crotonase, Enolase and Haloacid Dehalogenase superfamilies are over-fragmented, perhaps due to the presence of fragmented communities. The worst performance of ACDC is for the VOC superfamily for which the F-measure≈0.78 and this is sub-optimal.

To explore the reasons for the imperfect agreement of the ACDC method, the community membership of sequences from the AH superfamily is analyzed. One community based on the ACDC method contains three sequences, gi|4033703 (NCBI Acc # Q52725.2), gi|6226558 (NCBI Acc #: P72156.2), and gi|11890745 (NCBI Acc #: AAG41202), which are assigned different functions in the gold standard database, namely, s-triazine hydrolase, atrazine chlorohydrolase, and melamine deaminase, respectively. For pair alignments between these three sequences, S = (145.8, 152.5,301.1) for the pairs (gi|4033703, gi|6226558), (gi|4033703,gi|11890745) and (gi|6226558, gi|11890745). These alignment scores are high and imply that they have significant sequence similarity. As a consequence of this similarity, these sequences constitute one community instead of the three expected based on the "ground truth" community structure. In Fig 5a, this Type 3 error is shown for one community. For the VOC dataset, additional occurrences of Type 3 errors are shown in Fig 5b. Note that within the limitation of a two-dimensional projection of the network and the visualization scheme employed here, the presence of overlapping community structure is not apparent. Manual inspection of the merged community nodes does not provide support for interpreting overlapping communities as the cause of Type 3 errors.

It would be tempting to infer that the sequences involved in Type 3 errors either correspond to multifunctional enzymes with as yet unexplored functional similarity or that mutations in the active site have significantly altered their substrate preference. Alternatively, the choice of the substitution matrix may somehow result in pair alignments which are incorrectly assigned a high score for these sequences. In any case, the potential presence of such sequences with high sequence similarity scores but with different functions makes it difficult to assign them to separate communities. Hence, caution must be exercised in judging community detection methods.

A second type of error that arises is where a "ground truth" community splits into two or more communities. By definition, a community contains nodes with more intra-community than inter-community similarity. It was verified that this condition is satisfied for the split communities found in both AH and VOC datasets. Such Type 2 errors are shown in Fig 5 for both superfamilies for which only two fragments result from the error. The fragmentation of a "ground truth" community is a consequence of the absence of a high-enough score for the pair alignment between sequences in each fragment. Such errors are likely to be a product of the alignment and do not necessarily reflect shortcomings of the community detection method.

A good test of any community detection method for SSNs is its ability to distinguish between communities that belong to distinct superfamilies when sequences from multiple superfamilies are present. The gold standard dataset is composed of five superfamilies and the number of "ground truth" communities in the Gold Standard dataset is 91. Assuming no overlap, the union of the set of communities identified by ACDC for each superfamily within the Gold Standard dataset leads to a total of 91 communities (from Table 2), in agreement with the "ground truth" number of communities. However, when all sequences are included together, the number of communities detected by ACDC is 86, instead of the "ground truth" number of 91. Either ACDC leads to Type 3 errors involving inter-superfamily community merger or from intra-superfamily community merger. Examination of superfamily membership of all nodes in each community did not reveal any instances of the mergering of communities from different superfamilies. Thus, ACDC erroneously merges communities from the same superfamily in the presence of additional superfamilies. Nevertheless, the high F-measure of 0.9368 indicates good overall performance of ACDC.

3.4 Performance comparison of ACDC with other community detection methods

In a recent evaluation of the performance of different clustering/community detection methods on SSNs [33], TransClust was found to be the most successful method with an F-measure≈0.914 for the Gold Standard dataset. At the outset, an F-measure value of 0.9368 for ACDC suggests that it outperforms TransClust, at least for the GS dataset. However, the SSN used with ACDC is based on the BLOSUM50 substitution matrix and differs from the SSN generated with the BLOSUM62 substitution matrix employed by Bernardes *et al* [33], complicates a direct comparison. Furthermore, it is important to note that ACDC utilizes a multidimensional pair similarity attribute vector, unlike TransClust which is typically applied to SSNs based on the choice of negative logarithm of pair similarity E-values as the alignment attribute. Unfortunately, apart from ACDC, no community detection methods based on multidimensional attribute vectors have been applied to SSNs. Therefore, a straightforward four-dimensional grid search (GridS) is employed for comparison with ACDC. GridS assumes the existence of a Separatrix cutoff attribute vector that partitions attribute space into a region that contains links relevant for community detection and the rest of attribute space. The detailed implementation of the GridS method is presented in Methods.

For comparison of GridS with ACDC, the maximum F-measure and the corresponding number of communities are summarized in Table 2. All F-measure values based on GridS are superior to ACDC for all datasets. For the Gold Standard dataset, the best F-measure value of 0.9484 exceeds the corresponding values of 0.9368 based on ACDC and 0.914 obtained with TransClust [33]. The best performance of GridS is for the Haloacid Dehalogenase superfamily for which the exact "ground truth" community structure is reproduced. The worst performance of GridS is for the VOC dataset, for which the F-measure of 0.8315 is only slightly better than the value of 0.7821 for ACDC. Thus, sequence similarities for the VOC dataset may be symptomatic of problems in using sequence similarity to distinguish between mechanistically distinct sequence communities or perhaps due to the choice of substitution matrices. Turning to the number of communities, GridS leads to the "ground truth" number of communities except for the Crotonase and Gold Standard datasets, for which the number is underestimated. Overall, GridS leads to the best community structure for the "ground truth" SSNs.

Given the optimal set of cutoff attribute vectors identified by GridS, the validity of the assumption that Region 1 contains the set of links, with large l_a values, that are most relevant to community structure detection can be tested. The set of cutoff attribute vectors, $\{(l_{a,c}, f_{id,c},$

$f_{m,c}$, $s_{a,c}$)}, that leads to the largest F-measure value for each dataset are summarized in Table 3 as a range of attribute values. The best solution does include the large l_a region of attribute space, as expected. However, attribute value combinations that extend to small values of l_a are also possible. These small l_a attribute vectors represent distant relationships which perhaps provide alternative routes for connecting nodes while keeping the F-measure unchanged. Thus, the GridS method supports the assumption that links with large l_a are most relevant to community structure detection.

The large variation in f_m values for all datasets suggests that community structure is least sensitive to the fraction of mismatched residues. In contrast, $f_{id,c}$ and $s_{a,c}$ values for best community structure are typically in the range $0.31 \leq f_{id,c} \leq 0.38$ and $0.14 \leq s_{a,c} \leq 0.33$. The variation of the former attribute is consistent with the choice of 30–40% identity of residues for homologous sequences [8, 35]. The narrow range of values for $s_{a,c}$ suggests that homology detection may benefit from including the scaled score as well.

Given the larger F-measure values obtained with GridS, it would be useful to identify potential differences between the optimal GridS and ACDC solutions that may contribute to the better performance of former. During a grid search, each value of F-measure can be obtained for a number of different attribute cutoff vectors. For the AH and VOC datasets, the cutoff attribute vectors (0.95, 0.32, 0.83, 0.28) and (0.10, 0.38, 0.80, 0.22) are selected as they deviate most from Region 1. The attribute vectors that satisfy these cutoff values in GridS are shown in Fig 6 along with those predicted by ACDC and all available links in the SSN.

For both the AH and VOC datasets, ACDC and GridS excluded most of the available links and selected only the large l_a region of attribute space. For the AH dataset, ACDC selects more links at smaller l_a values than GridS even though the F-measure values are quite similar. Given that GridS leads to 29 communities instead of the 28 found by ACDC, it would appear that the additional links included in ACDC may lead to a Type 3 error that is avoided in GridS. For the VOC dataset, GridS selects a number of links at small l_a values that are not included by ACDC. It was verified that including these additional links to the set of links identified by ACDC does not change the number of communities. GridS excludes a few links at small s_a values that are included by ACDC. Given that GridS finds 17 communities instead of the 14 found by ACDC, it is likely that the links excluded by GridS, as compared to ACDC, lead to a larger number of communities. The missing links in GridS occupy the region near the inter-cluster boundary, as shown in Fig 3c, and excluding them in ACDC may not be straightforward. Thus, small differences in the set of selected links, probably in the inter-cluster region of ACDC, may be responsible for differences in ACDC and GridS.

4. Discussion

An SSN is a network where each node corresponds to a protein (or nucleotide) sequence and each link indicates a similarity relationship between pairs of sequences. As with any real or

Table 3. Cutoff attribute vector range based on the GridS method.

Dataset	Cutoff attribute vector range
Amidohydrolase	(0.95–0.96, 0.32–0.33, 0.52–0.83, 0.28)
Crotonase	(0.14–0.71, 0.38, 0.48–0.80, 0.14–0.33)
Enolase	(0.99, 0.31, 0.49–0.69, 0.25/0.28)
Haloacid Dehalogenase	(0.39–0.71, 0.31, 0.46–0.62, 0.29)
Vicinyl Oxygen Chelatase	(0.10–0.81, 0.38, 0.0–0.83, 0.14–0.28)
Gold Standard	(0.19–0.25, 0.37, 0.0–0.80, 0.24)

https://doi.org/10.1371/journal.pone.0178650.t003

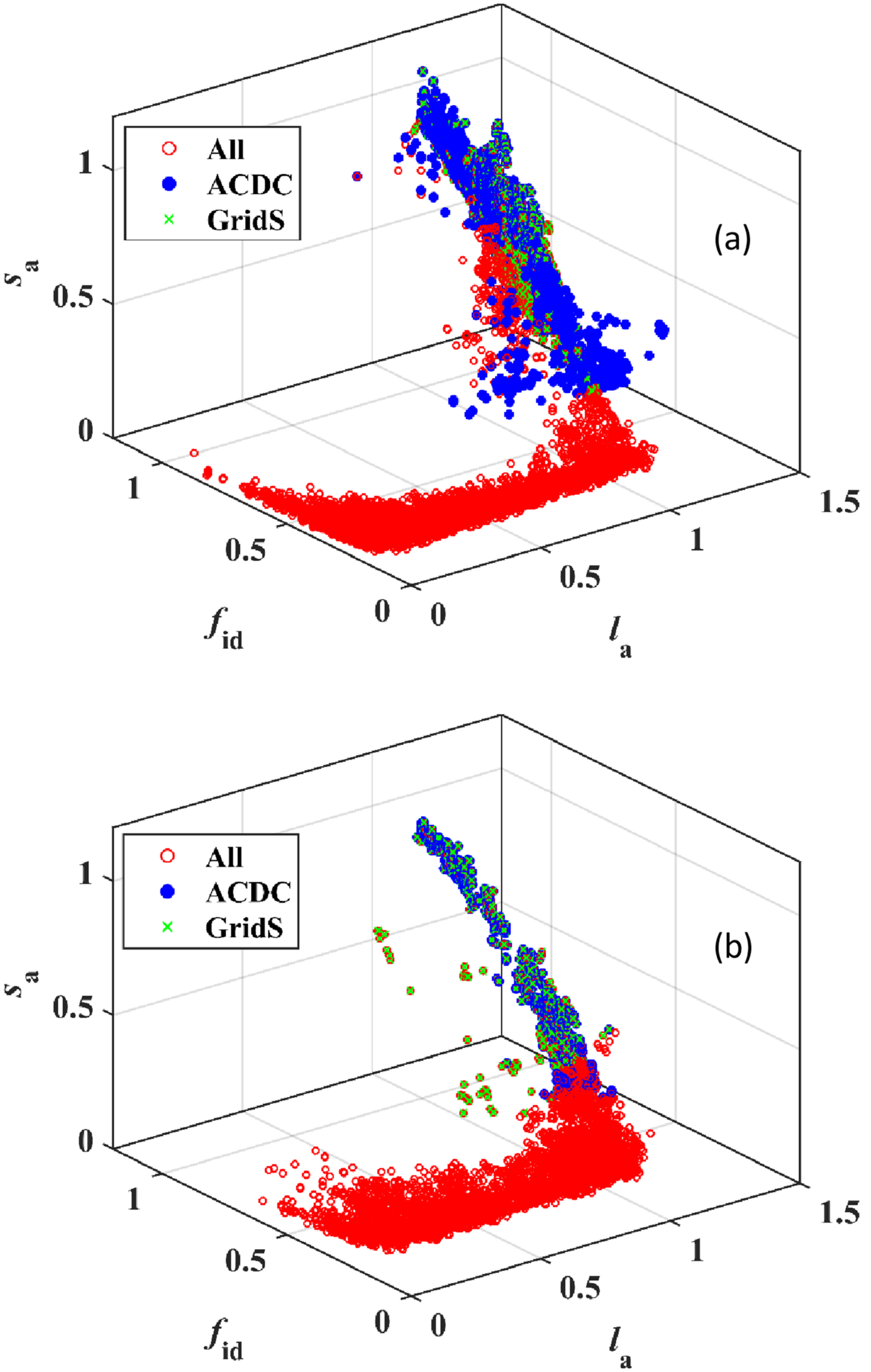

Fig 6. Distribution of links in attribute space relevant for community identification. (a) For the Amidohydrolase and (b) Vicnyl Oxygen Chelatase superfamily. All links (red), links selected by ACDC (blue), and links that give the maximum F-measure based on GridS (green) are presented for comparison. The cutoff attribute vectors selected by GridS for the Amidohydrolase and Vicnyl Oxygen Chelatase superfamiles are (0.95, 0.32, 0.83, 0.28) and (0.10, 0.38, 0.80, 0.22), respectively.

https://doi.org/10.1371/journal.pone.0178650.g006

artificial network, an important mesoscopic structural motif in a network is its community structure. For protein SSNs, each community corresponds to a functionally (and evolutionarily) related set of sequences and groups of related communities constitute a superfamily. Thus, community structure of SSNs is an important tool for sequence classification. In this manuscript, a new method for detecting the community structure in protein SSNs is presented.

A crucial feature of an SSN is the similarity between pairs of sequences which can be quantified with a number of attributes such as the alignment score, fraction of identical residues, fraction of mismatched residues and the alignment length. These standard alignment attributes are automatically output by most sequence pair alignment programs. In order to differentiate between two pair alignments with the same alignment score, a novel sequence similarity attribute vector is constructed by supplementing the alignment score with standard alignment metrics, i.e., the alignment length, percentage identity and percentage of mismatched positions. Invoking a natural length and score scale for each pair alignment, a scaled attribute vector is proposed as a link attribute for the SSN. Now, each scaled attribute vector corresponds to a point in four-dimensional attribute space.

The distribution of points in attribute space varies continuously between two limiting regions via an intermediate region. To the best of the author's knowledge, this attribute scaling and the resultant distribution of points in attribute space is absent from published literature. The two limiting regions were identified as a high sequence similarity Region 1 at large l_a values, and low sequence similarity Region 2 at small l_a values. For both these regions, there is substantial variation in the all alignment attributes, except l_a. In Region 1, $f_{id} \geq 0.3$ for all attribute vectors and they primarily correspond to intra-community sequence pairs which are expected to be evolutionarily related, *i.e.*, they are homologous, and likely to be structurally similar [8, 35, 52].

In the intermediate region, $0.2 \leq f_{id} \leq 0.3$, sequence pairs belong to the twilight zone of sequence homology. About 10% of all such sequence pairs in the twilight zone have been estimated to share the same tertiary structure[35]. Since the intermediate region appears to contain a sizable fraction of inter-community attribute vectors (see Fig 1d and Figure d in S1 Fig), the percentage of such homologous sequence pairs in the intermediate region may differ from 10% and is superfamily dependent. It is likely that pair attribute vectors in this region correspond to divergent domains shared between distantly related sequence pairs.

Approaching Region 2, l_a decreases with respect to the intermediate region and $f_{id}<0.2$ at first. Such f_{id} values correspond to the midnight region of sequence homology. The fraction of intra-community attribute vectors in this region is expected to be very small highlighting the problem of identifying homologous sequences in the midnight region. Such pair alignments may correspond to subdomain fragments with large similarity and perhaps represent "ancestral" peptides [53] shared by sequences in a superfamily. Alternatively, they may correspond to sequence fragments that share significant sequence similarity independent of evolutionary origin. Further analysis will be required to clearly identify the significance of this region.

The new community detection method, ACDC, makes use of the structure of the distribution of attribute vectors in attribute space. It is based on the hypothesis that links in Region 1 correspond to homologous sequence pairs that form the minimal set of links required for community detection. Such links in Region 1 are then identified as components of a subset of clusters identified with *k*-means clustering of points in attribute space. Based on a suitable choice of outliers, the minimal set of attribute vectors is shortlisted. Network structure is then utilized during partition density maximization for a first solution to the community structure. Community structure is then refined in order to eliminate Type 1 errors and minimize Type 2 errors which arise when the minimum inter-community similarity, *d*, is small than or slightly larger than the maximum intra-community similarity. The entire process is set up as an unsupervised method with no user specified parameters.

The performance of ACDC was evaluated by comparing the number of communities and the F-measure values with respect to "ground truth" community structure of the gold standard datasets of Brown et al[38] as well as the five functionally distinct protein superfamiles that constitute it. For the entire dataset, the F-measure value of 0.9368 based on ACDC exceeds a recently published best F-measure value of 0.914 obtained with TransClust [33]. At least for the Brown dataset[38], ACDC is the best performing community detection method for SSNs. For most constituent superfamilies, ACDC leads to F-measure values of about 0.9 or higher, except for the Vicinyl Oxygen Chelatase superfamily for which the F-measure of 0.78 suggests poor reproduction of the "ground truth" community structure. Analysis of the errors in community detection reveals (a) the partition of communities due to smaller similarity between sequences belonging to the two fragment communities or (b) merger of communities due to higher similarity between sequences that belong to different communities. Such errors may arise due to shortcomings of the substitution matrix used to calculate sequence similarity attributes. A study of the effect of the substitution matrix on community detection in SSNs using ACDC and its application to sequence databases with more distantly related homologous sequence will be presented in a following manuscript.

In order to compare the performance of ACDC to a method that is based on multidimensional attribute vectors, a four-dimensional grid search GridS was performed. The set of cutoff attribute vectors that lead to the best F-measure for all datasets was identified. GridS leads to F-measure values that exceed ACDC for all datasets. For the Gold Standard dataset, the GridS F-measure of 0.948 is much better than ACDC or TransClust. Note that this F-measure value is not significantly smaller than the F-measure of 0.959 that Bernardes *et al.* [33] obtained for the same dataset using a Hidden Markov Model profile-profile similarity network. However, note that a grid search by itself cannot be used to predict community structure, unless coupled with some criterion for selecting a best prediction in some efficient manner. It is hoped that future developments in community detection will benefit from comparison of their performance with benchmark GridS results. Comparison of the set of attribute vectors selected by ACDC and GridS indicates that ACDC includes links that contribute to the inter-cluster region that are excluded by GridS. For ACDC to perform better, this boundary region will have to be addressed in subsequent work.

ACDC incorporates partition density optimization, which was originally formulated in the context of link community detection; link communities in turn lead to the identification of overlapping communities of nodes. Since overlapping communities may arise in SSNs of multi-domain protein sequences the question arises as to whether ACDC can detect them. Among the superfamilies considered here for validating ACDC, the Enolase superfamily contains multi-domain sequences, each carrying the following PFAM domains pairs or one domain from each pair: (PF02746 and PF01188) or (PF03592 and PF00113) or (PF05034 and PF07476). The Enolase superfamily contains 9 "ground truth" communities and ACDC detects 10 communities. At least, then, for this example of a multi-domain sequence superfamily, ACDC performs well and the F-measure for the resulting community structure is 0.9684. Due to alignment length scaling and the preferential selection of attribute vectors with large scaled aligned lengths, ACDC preferentially selects alignments over the full length of the shorter sequence. As a result, multi-domain sequences aligned over the entire length of the shorter sequence are likely to be selected by the method and these will minimize the presence of overlapping communities. It is certainly possible that ACDC may not be as successful for other more complicated multi-domain sequence superfamilies. Then, the use of other overlapping community detection methods may be necessary. Further investigation is required to address this issue.

A general shortcoming of link based community detection methods is their scalability. For a superfamily with N sequences, there are n = N(N-1)/2 links in the SSN. For large superfamilies, such as the alpha/beta hydrolase fold super-family, with N≈O(10^5) sequences, the number of links n≈O(10^{10}) is clearly a huge number. A first step towards improving the performance of any SSN analysis method on large sequence datasets is to filter the dataset such that no two sequences have a percentage identity exceeding 99% or 95% identity is included. There are two computationally demanding steps in ACDC; k-means clustering and partition density optimization. Efficient implementations of k-means-like or grid based clustering methods have a time-complexity of O(nk) or O(n), where k is the number of clusters, or grid size [54]. In the partition density optimization step, all links are sorted by the distance d, a process that has a time complexity of O(n log n). From the sorted list of attributes, links are added to the network in increasing magnitude of d until a maximum value of the partition density is reached. Graph traversal and checking if two nodes are connected have time-complexities of O(n+N) and O(1)[55]. So, partition density optimization should scale reasonably well. As these two computationally demanding steps have reasonable scaling behavior, ACDC can be used for community detection in large sequence datasets, particularly once parallel implementations have been incorporated.

Most methods for community detection utilize one of, for example, the negative logarithm of the E-value, score, percentage identity, as link attributes. For each choice of attribute, a different SSN is obtained although the community structure is hopefully similar. The set of SSNs then comprises a multiplex network[56], one network for each attribute, and the community structure represents a balance between link attributes at each layer of the network. By integrating all attributes, ACDC in effect serves as a community detection method for multiplex networks. Although formulated for SSNs, it should be possible to extend ACDC to the analysis of network structure in multiplex networks. In any case, given the simplicity, intuitive interpretability, and unsupervised nature, it is hoped that the ACDC will prove useful for the scientific community.

Supporting information

S1 Table. Pair alignments with same scores can have different sets of alignment characteristics.
(PDF)

S2 Table. Mapping of Gold Standard dataset sequence gi numbers to NCBI accession numbers.
(CSV)

S1 Fig. Various representations for alignment attributes from the validation sequence datasets.
(PDF)

S2 Fig. Summary of results from all stages of ACDC for all sequence datasets.
(PDF)

S1 Dataset. All pair alignment data in BLAST tabular format for the Amidohydrolase superfamily.
(TXT)

S2 Dataset. All pair alignment data in BLAST tabular format for the Crotonase superfamily.
(TXT)

S3 Dataset. All pair alignment data in BLAST tabular format for the Enolase superfamily.
(TXT)

S4 Dataset. All pair alignment data in BLAST tabular format for the Haloacid Dehalogenase superfamily.
(TXT)

S5 Dataset. All pair alignment data in BLAST tabular format for the Vicinyl Oxygen Chelatase superfamily.
(TXT)

S6 Dataset. All pair alignment data in BLAST tabular format for the Gold Standard dataset.
(TXT)

S7 Dataset. All scaled attribute vectors for the Amidohydrolase superfamily.
(TXT)

S8 Dataset. All scaled attribute vectors for the Crotonase superfamily.
(TXT)

S9 Dataset. All scaled attribute vectors for the Enolase superfamily.
(TXT)

S10 Dataset. All scaled attribute vectors for the Haloacid Dehalogenase superfamily.
(TXT)

S11 Dataset. All scaled attribute vectors for the Vicinyl Oxygen Chelatase superfamily.
(TXT)

S12 Dataset. All scaled attribute vectors for the Gold Standard dataset.
(TXT)

Acknowledgments

The authors acknowledge Steve Moulton and Michael Galloway for invaluable assistance with the Center for Molecular Biophysics' computer clusters. The authors also acknowledge helpful input on several features of SSEARCH from Dr. William Pearson.

Author Contributions

Conceptualization: JC.

Data curation: JC.

Formal analysis: JC.

Funding acquisition: FL JCS.

Investigation: JC.

Methodology: JC.

Project administration: JCS FL.

Resources: JCS FL.

Software: JC.

Supervision: JCS FL.

Validation: JC.

Visualization: JC.

Writing – original draft: JC.

Writing – review & editing: JCS FL JC.

References

1. Schlessinger A, Matsson P, Shima JE, Pieper U, Yee SW, Kelly L, et al. Comparison of human solute carriers. Protein Science. 2010; 19(3):412–28. https://doi.org/10.1002/pro.320 PMID: 20052679
2. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. PLoS Comput Biol. 2009; 5(12):e1000605. https://doi.org/10.1371/journal.pcbi.1000605 PMID: 20011109
3. Brown SD, Babbitt PC. New Insights about Enzyme Evolution from Large Scale Studies of Sequence and Structure Relationships. Journal of Biological Chemistry. 2014; 289(44):30221–8. https://doi.org/10.1074/jbc.R114.569350 PMID: 25210038
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000; 25(1):25–9. https://doi.org/10.1038/75556 PMID: 10802651
5. Vingron M, Waterman MS. Sequence alignment and penalty choice. Journal of Molecular Biology. 1994; 235(1):1–12. http://dx.doi.org/10.1016/S0022-2836(05)80006-3. PMID: 8289235
6. David F, Ginsparg P, Zinn-Justin J. Fluctuating geometries in statistical mechanics and field theory: Elsevier; 1996.
7. Wolfsheimer S, Melchert O, Hartmann AK. Finite-temperature local protein sequence alignment: Percolation and free-energy distribution. Physical Review E. 2009; 80(6):061913.
8. Pearson WR. An Introduction to Sequence Similarity ("Homology") Searching. Current Protocols in Bioinformatics: John Wiley & Sons, Inc.; 2002.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990; 215(3):403–10. http://dx.doi.org/10.1016/S0022-2836(05)80360-2. PMID: 2231712
10. Newman M. Networks: An Introduction: Oxford University Press, Inc.; 2010. 720 p.
11. Newman M, Barabási A-L, Watts DJ. The structure and dynamics of networks: Princeton University Press; 2006.
12. Cerina F, De Leo V, Barthelemy M, Chessa A. Spatial Correlations in Attribute Communities. PLoS ONE. 2012; 7(5):e37507. https://doi.org/10.1371/journal.pone.0037507 PMID: 22666361
13. Borgatti SP, Everett MG, Johnson JC. Analyzing Social Networks: SAGE; 2013.
14. Junker BHaS F. Analysis of Biological Networks. John Wiley & Sons, Inc.; 2007.
15. Barthélemy M. Spatial networks. Physics Reports. 2011; 499(1–3):1–101. http://dx.doi.org/10.1016/j.physrep.2010.11.002.
16. Fortunato S. Community detection in graphs. Physics Reports. 2010; 486(3–5):75–174. http://dx.doi.org/10.1016/j.physrep.2009.11.002.
17. Bothorel C, Cruz JD, Magnani M, Micenkova B. Clustering attributed graphs: Models, measures and methods. Network Science. 2015; 3(03):408–44. https://doi.org/10.1017/nws.2015.9
18. Yang J, Leskovec J. Overlapping Communities Explain Core–Periphery Organization of Networks. Proceedings of the IEEE. 2014; 102(12):1892–902.
19. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. PLoS ONE. 2009; 4(2):e4345. https://doi.org/10.1371/journal.pone.0004345 PMID: 19190775
20. Halary S, McInerney J, Lopez P, Bapteste E. EGN: a wizard for construction of gene and genome similarity networks. BMC Evol Biol. 2013; 13(1):1–9. https://doi.org/10.1186/1471-2148-13-146 PMID: 23841456
21. Akiva E, Brown S, Almonacid DE, Barber AE, Custer AF, Hicks MA, et al. The Structure–Function Linkage Database. Nucleic Acids Research. 2014; 42(D1):D521–D30. https://doi.org/10.1093/nar/gkt1130 PMID: 24271399

22. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Research. 2014; 42(D1):D222–D30. https://doi.org/10.1093/nar/gkt1223 PMID: 24288371

23. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, et al. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Research. 2015; 43(D1):D376–D81. https://doi.org/10.1093/nar/gku947 PMID: 25348408

24. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology. 1995; 247(4):536–40. http://dx.doi.org/10.1016/S0022-2836(05)80134-2. PMID: 7723011

25. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Research. 2002; 30(1):281–3. https://doi.org/10.1093/nar/30.1.281 PMID: 11752315

26. Weston J, Elisseeff A, Zhou D, Leslie CS, Noble WS. Protein ranking: From local to global structure in the protein similarity network. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(17):6559–63. https://doi.org/10.1073/pnas.0308067101 PMID: 15087500

27. Ahn Y-Y, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. Nature. 2010; 466(7307):761–4. http://www.nature.com/nature/journal/v466/n7307/abs/nature09182.html#supplementary-information. PMID: 20562860

28. Evans TS, Lambiotte R. Line graphs of weighted networks for overlapping communities. Eur Phys J B. 2010; 77(2):265–72. https://doi.org/10.1140/epjb/e2010-00261-8

29. Gerlt JA, Bouvier JT, Davidson DB, Imker HJ, Sadkhin B, Slater DR, et al. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. Biochimica et Biophysica Acta (BBA)—Proteins and Proteomics. 2015; 1854(8):1019–37. http://dx.doi.org/10.1016/j.bbapap.2015.04.015.

30. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research. 2002; 30(7):1575–84. https://doi.org/10.1093/nar/30.7.1575 PMID: 11917018

31. Nepusz T, Sasidharan R, Paccanaro A. SCPS: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. BMC Bioinformatics. 2010; 11(1):1–13. https://doi.org/10.1186/1471-2105-11-120 PMID: 20214776

32. Wittkop T, Emig D, Lange S, Rahmann S, Albrecht M, Morris JH, et al. Partitioning biological data with transitivity clustering. Nat Meth. 2010; 7(6):419–20. http://www.nature.com/nmeth/journal/v7/n6/suppinfo/nmeth0610-419_S1.html.

33. Bernardes J, Vieira F, Costa L, Zaverucha G. Evaluation and improvements of clustering algorithms for detecting remote homologous protein families. BMC Bioinformatics. 2015; 16(1):1–14. https://doi.org/10.1186/s12859-014-0445-4 PMID: 25651949

34. van den Berg R, Hoefsloot H, Westerhuis J, Smilde A, van der Werf M. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006; 7 (1):142. https://doi.org/10.1186/1471-2164-7-142 PMID: 16762068

35. Rost B. Twilight zone of protein sequence alignments. Protein Engineering. 1999; 12(2):85–94. https://doi.org/10.1093/protein/12.2.85 PMID: 10195279

36. Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth. Knowledge and Information Systems. 2013; 42(1):181–213. https://doi.org/10.1007/s10115-013-0693-z

37. Creusefond J, Largillier T, Peyronnet S. On the Evaluation Potential of Quality Functions in Community Detection for Different Contexts. In: Wierzbicki A, Brandes U, Schweitzer F, Pedreschi D, editors. Advances in Network Science: 12th International Conference and School, NetSci-X 2016, Wroclaw, Poland, January 11–13, 2016, Proceedings. Cham: Springer International Publishing; 2016. p. 111–25.

38. Brown S, Gerlt J, Seffernick J, Babbitt P. A gold standard set of mechanistically diverse enzyme superfamilies. Genome Biology. 2006; 7(1):R8. https://doi.org/10.1186/gb-2006-7-1-r8 PMID: 16507141

39. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998; 14(9):755–63. https://doi.org/10.1093/bioinformatics/14.9.755 PMID: 9918945

40. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology. 1981; 147(1):195–7. http://dx.doi.org/10.1016/0022-2836(81)90087-5. PMID: 7265238

41. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences. 1988; 85(8):2444–8.

42. Hauser M, Mayer C, Soding J. kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics. 2013; 14(1):248. https://doi.org/10.1186/1471-2105-14-248 PMID: 23945046

43. Wold S, Esbensen K, Geladi P. Proceedings of the Multivariate Statistical Workshop for Geologists and GeochemistsPrincipal component analysis. Chemometrics and Intelligent Laboratory Systems. 1987; 2 (1):37–52. http://dx.doi.org/10.1016/0169-7439(87)80084-9.

44. Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics. 1974; 3 (1):1–27. https://doi.org/10.1080/03610927408827101

45. Tan P-N, Steinbach M, Kumar V. Introduction to Data Mining, (First Edition): Addison-Wesley Longman Publishing Co., Inc.; 2005.

46. Powers DM. What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes. arXiv preprint arXiv:150306410. 2015.

47. Rodriguez-Esteban R. Biomedical Text Mining and Its Applications. PLoS Comput Biol. 2009; 5(12): e1000597. https://doi.org/10.1371/journal.pcbi.1000597 PMID: 20041219

48. Tukey JW. Exploratory Data Analysis: Addison-Wesley Publishing Company; 1977.

49. Dovoedo YH. Contributions to outlier detection methods: Some theory and applications: The University of Alabama Tuscaloosa; 2011.

50. Brys G, Hubert M, Struyf A. A Robust Measure of Skewness. Journal of Computational and Graphical Statistics. 2004; 13(4):996–1017. https://doi.org/10.1198/106186004X12632

51. Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. Computational Statistics & Data Analysis. 2008; 52(12):5186–201. http://dx.doi.org/10.1016/j.csda.2007.11.008.

52. Haggerty LS, Jachiet P-A, Hanage WP, Fitzpatrick DA, Lopez P, O'Connell MJ, et al. A Pluralistic Account of Homology: Adapting the Models to the Data. Molecular Biology and Evolution. 2014; 31 (3):501–16. https://doi.org/10.1093/molbev/mst228 PMID: 24273322

53. Alva V, Söding J, Lupas AN. A vocabulary of ancient peptides at the origin of folded proteins. eLife. 2015; 4:e09410. https://doi.org/10.7554/eLife.09410 PMID: 26653858

54. Aggarwal CC, Reddy CK. Data Clustering: Algorithms and Applications: Chapman \& Hall/CRC; 2013. 652 p.

55. Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms, Third Edition: The MIT Press; 2009. 1312 p.

56. Lee K-M, Min B, Goh K-I. Towards real-world complexity: an introduction to multiplex networks. Eur Phys J B. 2015; 88(2):1–20. https://doi.org/10.1140/epjb/e2015-50742-1