Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

MOCASSIN-prot: A Multiple Objective Clustering Approach for Protein Similarity Networks

Brittney N. Keel [1], **Bo Deng** [1] **and Etsuko N. Moriyama** [2,*]

[1]USDA, ARS, U.S. Meat Animal Research Center, Clay Center, NE 68933, USA
[2]Department of Mathematics, University of Nebraska-Lincoln, Lincoln, NE 68588, USA
[3]School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Proteins often include multiple conserved domains. Various evolutionary events including duplication and loss of domains, domain reshuffling, as well as sequence divergence generate complex proteins and affect their functions. As a consequence, a large variation exists in the numbers, combinations, and orders of domains among protein families and subfamilies, and their evolutionary history is best modeled through networks that incorporate information from the entire domain content of the proteins. We have previously proposed a game-theoretic approach to constructing protein networks. In this study we adapt our method into the framework of multi-objective optimization and examine its application to cluster multi-domain proteins.
Results: We applied our method to cluster a multi-domain protein family, the Regulator of G-Protein Signaling family, as well as protein sets from ten genomes. We compared our classification results with the results from two other methods, Markov clustering and phylogenetic clustering. We showed that compared to other techniques, our approach, which uses both domain composition and quantitative sequence similarity information, can generate more functionally coherent protein clusters and better differentiate protein subfamilies.
Availability: MOCASSIN-prot source code, implemented in Perl and Matlab, is freely available on the web at NEED TO FILL THIS IN!!.
Contact: emoriyama2@unl.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Proteins often include conserved sequence regions. They are called domains, and in eukaryotes, 70% or more proteins are multi-domain proteins (Chothia and Gough, 2009; Levitt, 2009). Domains are often associated to discrete functions, and shuffling and accretion of domains (Graur and Li, 2000; Koonin *et al.*, 2000) are important mechanisms for evolutionary innovation in protein functions. Among multi-domain proteins, families are often recognized based on their domain composition and sometimes their specific arrangements of domains (Chothia and Gough, 2009; Vogel *et al.*, 2004).

One way to infer protein function is through phylogenomic analysis, where protein functions are assigned in the context of protein families based on evolutionary relationships. Classifying proteins into families and subfamilies has been shown to improve the accuracy of functional classification (Sjölander, 2004). Protein clustering is usually done based on similarity among their sequences. One of the simplest and most common ways to identify sequence similarity is to perform a pairwise alignment-based sequence similarity search, such as BLAST (Altschul *et al.*, 1997) or FASTA (Pearson and Lipman, 1988). More sensitive similarity searches can be done using profile hidden Markov models (pHMMs) (Finn *et al.*,

[1] Mention of a trade name, proprietary product, or specified equipment does not constitute a guarantee or warranty by the USDA and does not imply approval to the exclusion of other products that may be suitable.
[2] The USDA is an equal opportunity provider and employer.

2014). Clustering algorithms often use the information from such similarity searches to generate clusters of protein sequences. For proteins that share one or more domains, phylogenetic analysis can be performed based on a multiple sequence alignment generated from these domain(s).

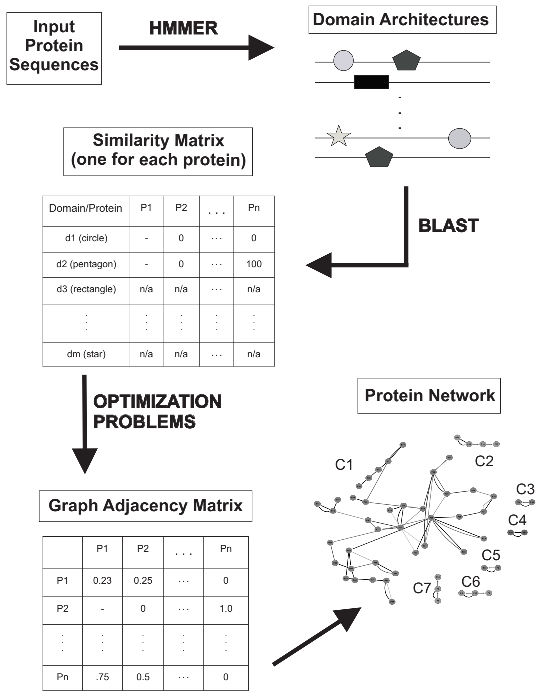
There are, however, two fundamental issues in using phylogenetic approaches when trying to classify groups of divergent proteins in protein families. The first is how to construct clusters given that the sequences are not alignable in their entirety when the proteins have multiple domains in varied composition and arrangement, limiting their application to proteins that share at least one alignable domain. If common domains are not found throughout the proteins, each subgroup of proteins needs to be independently analyzed based on different sets of domains. Another drawback to phylogenetic clustering is that it usually assumes evolution to be a bifurcating process. However, reticulate evolutionary events, such as domain shuffling, lead to evolutionary histories that are more accurately represented by networks.

In order to get a more complete picture of the evolutionary process of multi-domain proteins the domain architectures of the proteins must be considered. Phylogenetic profile methods (Bhardwaj *et al.*, 2012; Chang *et al.*, 2008) have tried to address this issue by constructing a phylogenetic tree that takes into consideration the entire domain content by viewing each protein sequence as a vector of domain scores. A tree is built using the Euclidean distance between the vectors of domain scores as the pairwise distance between the proteins. Just as for classic phylogenetic methods, network relationships among the proteins cannot be detected using this approach.

Protein similarity networks have been introduced to address the multi-domain protein clustering problem (Atkinson *et al.*, 2009; Pipenbacher *et al.*, 2002). Many protein similarity networks are constructed using local sequence similarities such as BLAST E-values (Altschul *et al.*, 1997). The Markov clustering algorithm (TRIBE-MCL), a graph clustering algorithm that simulates random walks within a graph, has been used to cluster proteins in a similarity network into families (Enright *et al.*, 2002; Van Dongen, 2000). Sequence similarity networks based only on local similarities, such as TRIBE-MCL, are used to cluster proteins on a larger, *e.g.* proteome, scale, where more variation in domain composition exists among proteins and commonly shared domains across all proteins are not required. In some sense these methods incorporate domain conservation information. However, they use information from only one region of similarity between two proteins. More detailed domain architecture information (such as the entire domain content and domain order) needs to be utilized in order to get a clearer picture of the evolutionary histories.

Domain co-occurrence networks (Wang *et al.*, 2011; Wuchty and Almaas, 2005) and related graph-theoretic approaches (Kummerfeld and Teichmann, 2009; Przytycka *et al.*, 2006; Xie *et al.*, 2011) incorporate domain composition (and sometimes order) information. However, these methods are employed to depict relationships among the domains, and the relationships between the proteins are usually not considered. Bipartite graphs have also been used to identify co-occurring domain sets in proteins (Cohen-Gihon *et al.*, 2007; Nacher *et al.*, 2009).

Existing methods construct domain networks and protein networks individually, with practically no connection between them. With the realization that protein functions cannot be understood fully without integrating their constitutive domain information, our aim is to build protein networks in terms of domain architectures and to improve and enhance protein function prediction. Protein sequence evolution is primarily governed by selective constraints on their sequences to maintain functions and also by modularity of domains that allows functional innovation. With this assumption, we have previously introduced a game-theoretic method for constructing protein networks (Deng *et al.*, 2013). In this work we adapt our approach into the framework of multi-objective optimization. Our method, MOCASSIN-prot, not only provides protein classifications using network clustering, but also gives us a better interpretation of the relationships between proteins and domains. Comparing our approach with other clustering methods, including phylogenetic clustering and Markov clustering, through analysis of both small and large-scale protein data sets, we illustrate the advantages of our method over others.



Domain/Protein	P1	P2	. . .	Pn
d1 (circle)	-	0	···	0
d2 (pentagon)	-	0	···	100
d3 (rectangle)	n/a	n/a	···	n/a
⋮	⋮	⋮	⋮	⋮
dm (star)	n/a	n/a	···	n/a

	P1	P2	. . .	Pn
P1	0.23	0.25	···	0
P2	-	0	···	1.0
⋮	⋮	⋮	⋮	⋮
Pn	.75	0.5	···	0

Fig. 1. Workflow for constructing a directed protein similarity network using MOCASSIN-prot. First the domain architecture of each protein in the set is identified by a profile hidden Markov model search (using the HMMER software). Then for each of the proteins a similarity matrix is constructed, using a log-transformed BLAST E-value as the similarity score. This matrix serves as the input to a multi-objective optimization problem. Each of these optimization problems is solved, and the solutions are used to construct a graph adjacency matrix for the protein network.

2 Methods

2.1 Workflow for MOCASSIN-prot protein clustering

A protein space is defined to be a set of proteins, each of which is in turn defined by a set of domains. Given the protein space and the corresponding domain space we can construct a similarity network that gives us a set of protein clusters, where a protein cluster is defined to be a weakly connected component in the similarity network. A *directed protein similarity graph*, $G = (V, E)$, for a given set of proteins, is a directed graph such that each vertex in the set $V = \{P_1, P_2, ..., P_n\}$ uniquely corresponds to one protein and all edges have nonzero weights with the incoming edges to any given vertex summing to 1. This definition is similar to that of Holloway and Beiko (2010).

The complete workflow for obtaining the clusters for a protein set via multi-objective optimization is shown in Figure 1. The first step is to identify the domain architecture for each of the proteins. This can be done using any domain sequence search algorithm. In our analysis we use a profile hidden Markov model search. Next, for each of the proteins we construct a similarity matrix using each domain as the unit, which serves as the input to a multi-objective optimization problem. We solve one optimization problem for each protein and use the solutions to construct a graph adjacency matrix for the protein network. Each step in this process is described in greater detail in the following sections.

2.1.1 Similarity matrix construction

As mentioned before, the network graph is constructed in a protein-by-protein approach. We begin by first determining the domain architecture for each protein. This is done by using the `hmmscan` program of HMMER3 (version v3.1) (Eddy, 2011) to search against the Pfam protein families database (release 27.0) (Finn *et al.*, 2014). These results are then filtered to include only those domains whose E-value lies below a given user-specified threshold. For our analysis we used non-overlapping domains and an E-value threshold of 1.0.

Once the domains have been identified, we construct a set of similarity matrices using each protein as the reference. In these matrices, we compare the amino acid sequences of all domain regions found in the reference protein to all other protein sequences using `blastp` with the default E-value threshold of 10.0. For a given reference protein, P_i, we extract the amino acid sequence from the top hit region for each domain on the protein. Suppose there are a total of m domains, $d_1, d_2, ..., d_m$, found on the proteins in the protein space, V. Then, as exhibited in Figure S1, the similarity matrix, $A_i = a_i(s, j)$, for the reference protein P_i is an $m \times n$ matrix, where $a_i(s, j)$ is the similarity score of domain s from protein P_i to protein P_j, and n is the number of proteins in the protein space.

For each entry in the similarity matrix we used a log-transform of the BLAST E-value, $a_i(s, j) = -\log(\alpha_{sj})$, where α_{sj} is the E-value obtained for the query domain d_s of protein P_i against the subject protein P_j. This score could be considered a proxy for the mutual information between proteins P_i and P_j with respect to domain d_s in that the higher the value the more similar the pair are in domain d_s. If a domain, d_k, was not found on the subject protein, P_j, then the similarity score was assigned to be $a_i(k, j) = -\log(2870)$, an arbitrarily chosen low similarity score. The values in the reference column (the i^{th} column) are not used in calculation and are therefore marked as '-' as shown in Figure S1. It should be noted that in our current method, each domain is represented only once in the similarity matrix. If a protein includes repetitive domains, such information is not included.

2.1.2 Protein similarity network construction

The construction of the edge weights is based on the assumption that protein sequence evolution is primarily governed by selective constraints on their sequences to maintain function along with modularity of domains that allows functional innovation. This assumption leads to maximizing the sequence similarity between proteins along shared domains. The outcome is a set of protein clusters with similar domain architectures in the protein space.

For each protein, P_i, we search for the other proteins in the protein space to which it is most uniquely similar (in terms of domain architecture). This is done using our multi-objective optimization method, which starts by defining S_i to be the index of integers $\{1, 2, ..., m\}$ such that $s \in S_i$ if and only if d_s is a domain of protein P_i. For ease of notation, let $y_j = w_{ij}$ for all $j = 1, 2, ..., n$, denote the weight of the edge from protein P_j to P_i. Thus we have $y_i = 0$, $\sum_j y_j = 1$, and $y_j \geq 0$. The incoming edge weight vector $\mathbf{y} = (y_1, y_2, ..., y_n)$ to protein P_i is then chosen to simultaneously "maximize" the mean similarity score

$$f_s(y_1, y_2, ..., y_n) = \sum_j a_i(s, j) y_j, \quad s \in S_i,$$

for each domain d_s of protein P_i (namely for each s from S_i). The theory of multi-objective optimization does not imply that each of the similarity scores is maximized in the strict mathematical sense, but rather in the sense of Pareto optimality, where any other choice of the probability vector **y** will make at least one of the objective functions f_s assume a smaller similarity score.

In our method, for each reference protein, P_i, we solve for the vector **y** in the following linear programming (LP) problem

$$\begin{aligned} &\max_{(\mathbf{y}, v)} && E_i = v \\ &\text{subject to} && f_s = \sum_j a_i(s, j) y_j \geq v, \text{for } s \in S_i \\ & && \sum_j y_j = 1,\ y_j \geq 0,\ j = 1, 2, ..., n,\ y_i = 0. \end{aligned} \tag{1}$$

Holloway and Beiko (2010) used this approach to construct a genome network. Because this is a LP problem a solution must exist and in general is unique. It is straight forward to verify that the solution is Pareto efficient, namely, for any probability vector (i.e. any non-solution) **y** there is at least one domain index $s \in S_i$ whose corresponding mean similarity score is smaller than the optimal value, i.e. $f_s < v$ for at least one $s \in S_i$.

By the theory of LP we know that the primary LP (1) has a dual linear programming problem (Deng *et al.*, 2013; Nash, 1951):

$$\begin{aligned} &\max_{(\mathbf{x}, v)} && E_i = v \\ &\text{subject to} && g_j = \sum_{s \in S_i} a_i(s, j) x_s \leq v, \text{for } j = 1, 2, ..., n, j \neq i \\ & && \sum_{s \in S_i} x_s = 1,\ x_s \geq 0. \end{aligned} \tag{2}$$

The solution probability vector **x** is the so-called Lagrange multiplier (or shadow price) of the primary LP problem (1). Conversely, the solution **y** of the primary LP (1) is the Lagrange multiplier of the dual LP (2). The optimal objective values, v, of (1) and (2) are identical.

This dual LP problem solves for the "maximal" divergence distribution of protein P_i with respect to the domain space. That is, for each protein P_j its least similar (i.e., most divergent) structure to the reference protein P_i will be a domain, d_s, shared with protein P_i having the smallest similarity score $a_i(s, j)$. Since this may happen on different domains for different proteins, the task is to find a domain distribution vector **x** to simultaneously "minimize" all of the mean similarity scores

$$g_j = \sum_{s \in S_i} a_i(s, j) x_s \quad j = 1, 2, ..., n, j \neq i.$$

The dual LP problem (2) is a way to solve this multi-objective optimization problem. The solution vector **x** is Pareto efficient (or optimal), that is, any other choice of the probability vector **x** will violate at least one of the objectives so that $g_j > v$ for at least one $1 \leq j \leq n, j \neq i$.

In the final step of our method we use the solutions of (1) and (2) for each of the proteins in the protein space to construct both the directed protein similarity network and the diversity profile. The edge weight w_{ij} from node P_j to P_i is assigned to be y_j from the solution to the primary LP (1) for node P_i. That is, the **y** solution, which obviously depends on i but the dependence is suppressed for simplicity, for node P_i gives the i^{th} row of the network matrix **W** in Figure S1. Arranging **W** in a block diagonal form results in the cluster formation for the network with each irreducible block defining a cluster. A high edge weight in the graph indicates a strong similarity between proteins. This network can be interpreted to be optimally conserved in that any other choice of network topology or edge weights will result in a network with at least one weaker conserved protein (namely a node in the network with weaker connections to the others in its cluster).

The solution vector **x** (which obviously depends on the node number i) is referred to as the domain diversity vector for protein P_i. A high diversity weight, x_s indicates that the reference protein, P_i, is more dissimilar or unique to the other proteins with respect to domain d_s. Arranging the **x** solution vectors for the proteins in the protein space into rows for their corresponding proteins yields the so-called diversity profile matrix.

Thus, by our framework the edge weights of the protein similarity network and the corresponding diversity profile are the result of the search for minimally shared regions (in **x**) of maximal similarity (in **y**) for each protein in relation to all other proteins in the protein space. The optimal objective values also yield important information about the network clusters. In fact, the optimal value $E_i = v$ is a measure for how tight the connection of protein P_i is to the other proteins. The higher the value, the higher the mean similarity scores f_s for all $s \in S_i$, which can be interpreted to mean that protein P_i is more conserved with respect to the proteins in the protein space. In addition, for two topologically identical clusters, it is their average optimal objective values that set them apart, by which the cluster with higher average optimal objective value is a 'tighter' or more similar subnetwork than the other.

2.1.3 Network refinement

The MOCASSIN-prot method described above gives rise to clustered networks on minimally shared regions of maximal similarity. We can extend the method to obtain secondary clusters from within a primary network cluster. This is done by removing the least conserved domain for each protein in the protein space. That is, for the domain d_k having the largest diversity weight for protein P_i, we purposely remove its objective function, f_k in (1). We then find the solution to the new corresponding LP problem. This secondary clustering structure will capture the next minimally shared region of maximal similarity for the proteins inside the primary cluster, increasing each protein's optimal objective value and identifying further subgroupings of similar proteins within the primary cluster. This zoom-in procedure can continue to reveal the tertiary, the quaternary, and so on, similarity relationship of the proteins. Thus unlike other methods, we are able to define a more refined network clustering structure from the original network without the use of arbitrary thresholds for pruning.

2.2 Data sets used in this study

Two types of protein data sets were used in this study. The first, the Regulator of G-Protein Signaling (RGS) protein family data set includes 55 proteins from the mouse (*Mus musculus*) genome (NCBI Annotation Release 103). They were found by performing HMMER3 (Eddy, 2011) search using Pfam (**?**) pHMMs PF00615 (RGS) and PF09128 (RGS-like) as queries, with an E-value threshold of 1.0. This RGS family sequence set was subsequently used to HMMER3 search against the entire Pfam database to find other domains that coexist in the sequences. Twenty six Pfam domains (including RGS and RGS-like) were identified on the proteins. The 55 RGS proteins and the Pfam domains found in each protein are shown in Table S1.

Large-scale protein sets from ten genomes (including seven bacteria and three eukaryotes) were also examined in this study. For each proteome set obtained from the UniProt Knowledgebase (UniProtKB, www.uniprot.org, release 2014_08), proteins from the SwissProt sector that had at least one identifiable domain and UniProt protein family annotation were collected. Table S2 lists the numbers of proteins and domains for each of these protein sets.

2.3 Evaluation of the method

The protein clustering results of MOCASSIN-prot were compared against two other methods, phylogenetic clustering and Markov clustering.

2.3.1 Phylogenetic clustering

For a regular phylogenetic analysis of multi-domain proteins, a multiple sequence alignment of a domain shared across all proteins is required. Therefore phylogenetic clustering could only be applied to the RGS data set in this study. A multiple alignment of the commonly shared domain (RGS or RGS-like domain) of the 55 RGS sequences was done using MAFFT (version v7.182, Katoh and Standley, 2013) using the L-INS-i algorithm with the default parameters. The maximum likelihood phylogeny was reconstructed using PHYML (version v3.1, Guindon *et al*., 2010) using the LG amino-acid substitution model, the gamma distribution shape parameter with the maximum-likelihood estimate, and bootstrap analysis with 1,000 pseudoreplicates. Bootstrap values of 70% were used to define the clusters of RGS sequences.

2.3.2 Markov clustering

The TRIBE-MCL algorithm (Enright *et al*., 2002) clusters proteins using the following steps: (a) for a given set of proteins, an all-vs.-all BLAST hit table is generated using the `blastp` program, (b) the `mcxdeblast` application is used to parse the BLAST table and generate an all-vs.-all similarity matrix using an E-value threshold of 1.0, (c) the `mcxassemble` program creates a probability matrix from the similarity matrix, and (d) the probability matrix is used as the input to the `mcl` program, which generates the protein clusters. Clusters were obtained using varied values of the inflation parameter, including the default I = 2.0 as well as I = 3.0, 4.0, 5.0. The MCL package (version v12-068) was downloaded from `www.micans.org/mcl`.

2.3.3 Cluster comparison metric

To compare the sets of clusters generated by two different clustering methods, we use a symmetric similarity measure similar to the average maximum Jaccard index (Prelic *et al*., 2006). Given two sets of protein clusters, A and B, the per-cluster similarity is given by

$$S(A,B) = \frac{\sigma(A,B) + \sigma(B,A)}{|A| + |B|} \tag{3}$$

where $\sigma(A,B) = \sum_{A_1 \in A} \max_{B_1 \in B} \frac{|A_1 \bigcap B_1|}{|A_1 \bigcup B_1|}$. Thus, this comparison metric has values ranging from 0 to 1, with 0 indicating no proteins were clustered correctly and 1 indicating all proteins were clustered correctly.

3 Results

3.1 MOCASSIN-prot clustering of RGS family proteins

The set of 55 mouse RGS proteins was clustered using MOCASSIN-prot. A total of 26 Pfam domains were identified on these proteins (Table S1). Nine clusters were found using our method (Figure 2). The clusters are labeled according to their average optimal objective values, in descending order. As mentioned before, these objective values yield important information about the inter-cluster similarities in the network. For example, Cluster 2 and Cluster 5 each include three proteins and are topologically identical. However, their average optimal objective values are 146.7074 and 71.9517, respectively, indicating that Cluster 2 is a 'tighter' or more conserved subnetwork than Cluster 5.

Even though all 55 proteins in this data set belong to the RGS family (they all contain either the RGS or RGS-like domain), the proteins in each of the network clusters differ from those in the other clusters with respect to their domain composition and varying levels of similarity between their domain sequences. The domain profile across the proteins in the RGS network is exhibited in Figure 3, where the proteins are grouped according to the clusters (1-9) in the network graph.

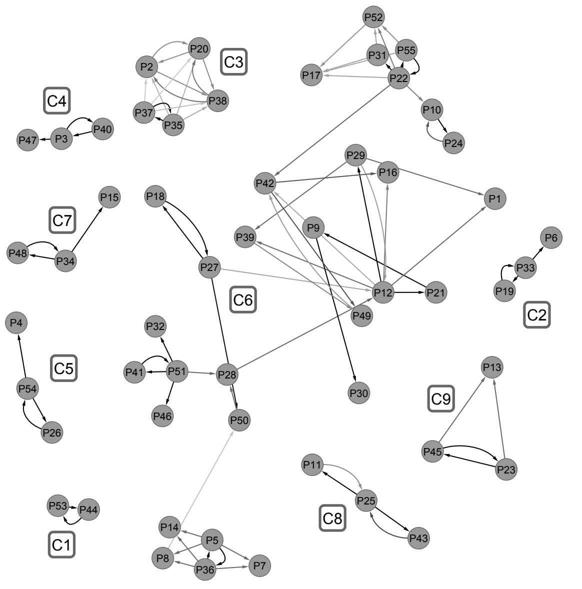


Fig. 2. RGS family primary MOCASSIN-prot network. From 55 mouse RGS family proteins, 9 clusters were identified using MOCASSIN-prot. The nodes represent distinct proteins, and the edges are directed so that the incoming edge weights of each node sum to 1. The edge color (in varying shades of gray) indicates the edge weight, with darker edges indicating high edge weights and lighter edges indicating low weights. The optimal objective value for each protein is represented in the network by the length of its incoming edges, with longer edges corresponding to small objectives values. Clusters in the network are labeled in descending order according to their average optimal objective value.

There are some clear differences in profile patterns between the clusters. For many of the clusters, specifically Clusters 1, 2, 7, 8, and 9, the proteins within the cluster all contain the same set of domains, and the weights placed on these domains are the same. The profile also highlights domains that are unique to specific clusters. For example, Pfam domain PF00018.23 is hallmark to Cluster 8 because it is present in all members of Cluster 8 but none of the other proteins in the network.

As an initial validation of the network clusters, we examined the protein type for the proteins within each cluster, taken from the NCBI website (Table S3). We saw that, generally speaking, proteins of the same type tended to fall in the same cluster. Figure S2 clearly shows that different domain architectures are represented in different clusters. Sequence divergence within the same domain type (e.g., RGS domain for RGS 17 vs. RGS 19/20 proteins) is recognized in separating Clusters 1 and 2. Note also that two isoforms of the same beta-adrenergic receptor kinase 2 gene (P3 and P47) fall into the same cluster (Cluster 4in Figure 2) even though one isoform (P47) lacks two domains (Figure S2). Hence with our multi-objective optimization framework, we can incorporate not only domain architecture information, but also sequence similarity, which produced an RGS protein network with valid clusters.

3.1.1 Comparison of RGS clustering to other clustering methods

To evaluate the protein clusters generated using MOCASSIN-prot against other methods, the 55 RGS protein sequences were clustered using the maximum-likelihood phylogenetic method and TRIBE-MCL. As mentioned before, for the phylogenetic method, only sequence information from the domain shared across all proteins (here, RGS or RGS-like domain) can be used. The MCL algorithm uses only the one most significantly similar region between pairs of proteins, which is identified by a `blastp` search prior to the clustering. In contrast, our multi-objective optimization technique incorporates both domain architecture and sequence similarity information. The clustering patterns were compared based on the symmetric Jaccard index (Equation (3)) calculated for each pairing of the three methods as shown in Table 1 (default TRIBE-MCL inflation paraemeter). Results for the other TRIBE-MCL inflation parameter values are shown in Table S4.

Table 1. Comparison of methods for RGS proteins.

	MOC (9)	PHY (25)	MCL^a (5)
MOC	-	-	-
PHY	0.4473	-	-
MCL	0.2018	0.1605	-

^{a} Default inflation parameter, I=2.0, used.

The maximum-likelihood phylogenetic analysis produced 25 clusters when 70% bootstrap values were used as the clustering threshold (Figure S3). Four of those clusters, Clusters 1, 11, 18, and 22 indicated by circled cluster numbers, were identical to clusters found by MOCASSIN-prot (Table S5). Although the phylogenetic clusters exhibited some similarity to the MOCASSIN-prot clusters ($S(PRD, PHY) = 0.4473$), the phylogenetic analysis cannot represent information from domains that are included in the multiple alignment. Therefore, it misses some relationships that our network approach reveals.

For example, seven of the RGS proteins (P2, P18, P20, P27, P35, P37, and P38) contain the RGS-like domain (PF09128.6). Proteins P2, P20, P35, P37, and P38 are grouped into the same cluster by both MOCASSIN-prot and the phylogenetic method. However, for the other two proteins

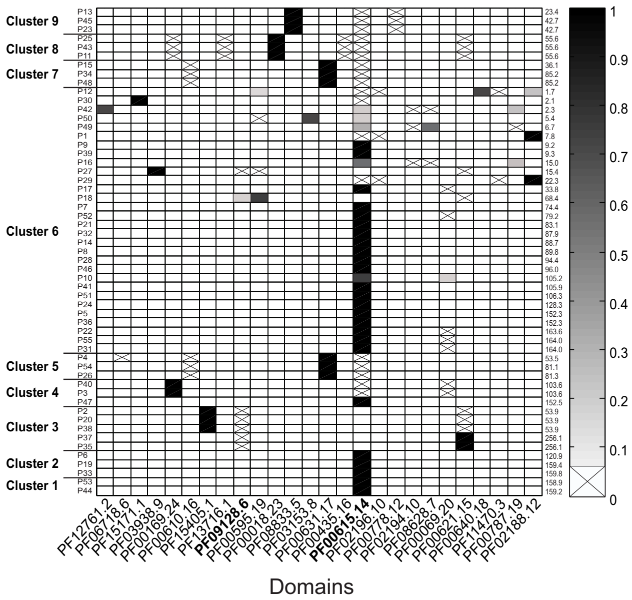


Fig. 3. RGS family diversity profile. Each row represents one of the 55 RGS proteins, while each column represents one of the 26 Pfam domains (RGS and RGS-like domains are shown in bold). Each cell is color-coded based on the diversity weights x from the LP problem (1): black for $x_i = 1$, a black X for $x_i = 0$ (but domain exists), and from light gray to dark gray for $0 > x_i > 1$. Blank cells indicate domain absence. Proteins in each cluster (identifiers shown on the left) are arranged according to their optimal objective value, with the largest appearing as the lowest row in the cluster. The objective value for each protein is shown on the right.

containing the RGS-like domain, proteins P18 and P27, there is a discrepancy between the two methods in cluster membership. Although the phylogenetic method clusters P18 with the five proteins containing RGS-like domain mentioned above (Cluster 23 in Figure S3), the phylogeny also shows that the RGS-like domain sequence of P18 is significantly different from those of the five proteins (the bootstrap value for the node clustering P18 with the others is 99.4%). P27 exists as a single protein cluster in the phylogeny (Cluster 15 in Figure S3).

In contrast, MOCASSIN-prot places both P18 and P27 in another cluster (Cluster 6 in Figure 2). Moreover, the edge weights indicate that relative to the other proteins in the RGS data set, P18 is P27's closest neighbor and vice versa. This relationship is a direct result of our method's use of information from all domains. As shown in Figure 3 (also see Figure S2), P18 contains two other domains, PF00595.19 (PDZ) and PF00621.15 (RhoGEF), in addition to the RGS-like domain (PF09128,6). All three of these domains are shared with P27, and one additional domain (PF03938.9, OmpH-like) is unique to P27. In the MOCASSIN-prot network P27 also exhibits similarity to both P50 and P12. Figure 3 shows that similar regions to domain PF00595.19 (PDZ) are uniquely found in proteins P12, P18, P27, and P50. While uniquely divergent domain compositions of P18 and P27 compared to other proteins containing the RGS-like domain combined with their less-conserved RGS-like domain sequences were enough to cluster them separately from other proteins with the RGS-like domain, sequence similarities of domains shared with proteins belonging to Cluster 6 were strong enough to make them cluster with these proteins. The same clustering pattern for P18 and P27 was seen in the secondary MOCASSIN-prot network (Figure S??). Since the phylogenetic clustering is based only on the RGS and RGS-like domains, these similarity relationships are not captured, illustrating the limitation of alignment-based phylogenetic clustering.

Clusters for the RGS sequences were also constructed using TRIBE-MCL. With the default inflation parameter, I = 2.0, the method identified 5 clusters (Figure S4(a)). Results for the other values of the inflation parameter are shown in Figure S4(b)-(d). The number of clusters generated by TRIBE-MCL was much smaller compared to MOCASSIN-prot and the phylogenetic method. None of the clusters generated by TRIBE-MCL clusters were 100% consistent with a cluster from MOCASSIN-prot (Table S5). The seven proteins containing the RGS-like domain were spread between two clusters in all runs of TRIBE-MCL (C1 and C5 in Figure S4(a); C1 and C7 in Figure S4(b); C1 and C4 in Figure S4(c)) except for the case when I = 5.0, where they were found in three of the clusters (C1, C5, and C7 in Figure S4(d)). As mentioned before, the MCL method clusters the sequences based on the single most significant region between them. The existence of highly conserved domains common to all of the proteins, such as the RGS and RGS-like domains, obscures further subgroup relationships among proteins. By incorporating information from all of the domains on the proteins, our method is able to cluster the RGS proteins on a finer scale than the TRIBE-MCL method.

As described before, our method can correctly cluster isoforms of the same gene, even when the domain compositions are different. This is not the case with the other two methods. For the two beta-adrenergic receptor kinase 2 isoforms (P3 and P47) mentioned before, P47 lacks two domains and contains only the RGS domain (Figure S2). In the phylogenetic clustering (Figure S3) P3 clusters with P40, beta-adrenergic receptor kinase 1, comprising Cluster 16, while P47 makes up Cluster 24, a single protein cluster located distantly from Cluster 16. Regardless of the inflation parameter, TRIBE-MCL places each of the three proteins in a separate cluster. For example, in Figure S4(a), P3 is in C2, P40 is in C3, and P47 is in C5. In the MOCASSIN-prot network all three beta-adrenergic receptor kinase proteins, including the two isoforms, cluster together (Cluster 4 in Figure 2). Our approach was able to pick out the relationships between these proteins correctly because they were highly similar in their shared RGS domain sequence.

Table 2. Clustering accuracy for MOCASSIN-prot and TRIBE-MCL for ten reference protein sets.

Genome	# REF^a	S(MCL2,REF)^b	S(MOC1,REF)^c	S(MOC2,REF)^d
B. subtilis	1322	0.1334(226)	0.1881(198)	0.5968(1617)
E. coli	1766	0.1215(276)	0.1429(188)	0.6285(1921)
T. pallidum	298	0.0608(16)	0.0444(9)	0.6919(188)
S. pyogenes	358	0.2738(276)	0.0612(15)	0.7020(232)
S. epidermidis	604	0.0827(45)	0.0841(31)	0.6907(414)
S. aureus	662	0.1001(62)	0.1036(45)	0.6918(496)
Y. pestis	829	0.0653(45)	0.0786(44)	0.7207(604)
D. melanogaster	1528	0.0996(161)	0.1315(161)	0.5552(1502)
M. musculus	3702	0.1330(1217)	0.2565(1207)	0.4263(6936)
S. cerevisiae	2243	0.1423(428)	0.1711(291)	0.5833(2523)

[a] Total number of reference clusters.
[b] MCL2 denotes TRIBE-MCL with default inflation parameter, I=2.0.
[c] MOC1 denotes primary MOCASSIN-prot network.
[d] MOC2 denotes secondary MOCASSIN-prot network.

3.2 MOCASSIN-prot clustering of large-scale protein sets

To test the performance of MOCASSIN-prot on larger scale data, we used protein sets from seven prokaryote and three eukaryote genomes. Figures S5-S14 show the primary and secondary protein similarity networks obtained from MOCASSIN-prot for these protein sets. For comparison, the clusters generated using TRIBE-MCL for each protein set are shown in Figures S15-S24.

We used the UniProt family assignments as the set of reference clusters and tested the performance of MOCASSIN-prot compared to TRIBE-MCL for each of the protein sets. The clustering accuracy results, assessed using the similarity measure in Equation (3), are summarized in Tables 2 and S6.

The TRIBE-MCL method, in general, exhibited lower clustering performance compared to MOCASSIN-prot. In only a few cases, TRIBE-MCL outperformed the primary MOCASSIN-prot clustering. The secondary MOCASSIN-prot networks especially were highly consistent with the reference clusters, surpassing the performance of TRIBE-MCL in all data sets. The reason that TRIBE-MCL was consistently outperformed is that TRIBE-MCL uses information from only one region of local similarity, i.e., it does not use information from all conserved domains on the proteins. MOCASSIN-prot incorporates information from all domains, leading to clusters that are more consistent with the UniProt family assignments.

The secondary MOCASSIN-prot networks were much more accurate than the primary networks because removing the least conserved domain for each protein, i.e., the domain with the maximum diversity weight, in a primary cluster and reclustering increases each protein's optimal objective value, identifying subgroupings of highly similar proteins within the primary cluster. We also tested the tertiary and quaternary MOCASSIN-prot networks and found little increase in method performance (Table S6), suggesting that the secondary network is sufficient for classifying the proteins.

4 Conclusion

Large-scale clustering of protein sequences incorporating their domain composition information is a challenging problem. Traditional approaches to the protein clustering problem, including TRIBE-MCL and phylogenetic clustering, use information from only one local region of similarity between proteins or information from only domain sequences shared among the majority of proteins. To obtain an accurate classification of protein families one needs to incorporate information from the entire domain composition.

In this study, we presented MOCASSIN-prot, a multi-objective optimization approach for protein classification. This method utilizes quantitative sequence similarity information from all domains on the proteins and builds a network that houses clusters of similar protein sequences. The method is scalable to the complete proteome level. Evaluation of protein clusters from MOCASSIN-prot, especially those from secondary networks, and TRIBE-MCL showed that MOCASSIN-prot exhibited consistently higher performance.

We should note that with our method the network structures and domain profiles depend critically on the similarity matrices that serve as input to the model. Therefore, in future work we must examine different scoring schemes (e.g., E-values based on profile HMMs, composition-based similarity scores, etc.) for their robustness and sensitivity. Improving the protein-domain similarity scores will give us better resolution in protein-domain classification, reflecting more accurate evolutionary and functional relationships between proteins.

Acknowledgements

The authors would like to thank Steven Dunbar for his valuable comments, ideas, and assistance in the undertaking of the research described here.

Funding

NEEDS TO BE FILLED IN

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.

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

Bhardwaj, G., Ko, K.D., Hong, Y., Zhang, Z., *et al.* (2012) PHYRN: A robust method for phylogenetic analysis of highly divergent sequences. *PLoS ONE*, **7**, e34261.

Chang, G.S., Hong, Y., Ko, K.D., Bhardwaj, G., *et al.* (2008) Phylogenetic profiles reveal evolutionary relationships within the 'twilight zone' of sequence similarity. *Proceedings of The National Academy of Sciences, USA*,, **105**, 13474-13479.

Chothia, C., and Gough, J. (2009) Genomic and structural aspects of protein evolution. *Biochemical Journal*, **419**, 15-28.

Cohen-Gihon, I., Nussinov, R., and Sharan, R. (2007) Comprehensive analysis of co-occurring domain sets in yeast proteins. *BMC Genomics*, **8**, 161.

Deng, B., Hinds, B., Zheng, X., and Moriyama, E.N. (2013) Bioinformatic game theory and its application to biological affinity networks. *Applied Mathematics*, **4**, 92.

Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Computational Biology*, **7**, e1002195.

Enright, A.J., Van Dongen, S., and Ouzounis C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, **30**, 1575-1584.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., *et al.* (2014) Pfam: The protein families database. *Nucleic Acids Research*, **42**, 2014, D222-230.

Graur, D., and Li, W.H. (2000) *Fundamentals of Molecular Evolution* 2nd edn. Sinauer Associates, Inc., Sunderland, MA.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systems Biology*, textbf59, 307-321.

Holloway, C., and Beiko, R. (2010) Assembling networks of microbial genomes using linear programming. *BMC Evolutionary Biology*, **10**, 360.

Katoh, K., and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772-780.

Koonin, E.V., Aravind, L., and Kondrashov, A.S. (2000) The impact of comparative genomics on our understanding of evolution. *Cell*, **101**:6, 573-576.

Kummerfeld, S.K., and Teichmann, S.A. (2009) Protein domain organisation: adding order. *BMC Bioinformatics*, **10**, 39.

Levitt, M. (2009) Nature of the protein universe. *Proceedings of the National Academy of Sciences, USA*, **106**, 11079-11084.

Nacher, J.C., Ochiai, T., Hayashida, M., and Akutsu, T. (2009) A bipartite graph based model of protein domain networks. In: *Complex Sciences*, Vol. 4 (J. Zhou, ed.). pp. 525-535 Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer Berlin Heidelberg.

Nash, J. (1951) Non-cooperative games. *The Annals of Mathematics*, **54**:21, 286-295.

Pipenbacher, P., Schliep, A., Schneckener, S., Schonhuth, A., *et al.* (2002) ProClust: Improved clustering of protein sequences with an extended graph-based approach. *Bioinformatics*, **18** (Suppl. 2), S182-S191.

Pearson, W.R., and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proceedings of The National Academy of Sciences, USA*,, **85**, 2444-2448.

Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., *et al.* (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122-1129.

Przytycka, T., Davis, G., Song, N., and Durand, D. (2006) Graph theoretical insights into evolution of multidomain proteins. *Journal of Computational Biology*, **13**, 351-363.

Sjölander, K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**, 170-179.

Van Dongen, S. (2000) Graph clustering by flow simulation. PhD thesis. The Netherlands: University of Utrecht.

Vogel, C., Berzuini, C., Bashton, M., Gough, J., and Teichmann, S.A. (2004) Supra-domains: evolutionary units larger than single protein domains. *Journal of Molecular Biology*, **336**, 809-823.

Wang, Z., Zhang, X.C., Le, M.H., Xu, D., *et al.* (2011) A protein domain co-occurrence network approach for predicting protein function and inferring species phylogeny. *PLoS ONE*, **6**, e17906.

Wuchty, S., and Almaas, E. (2005) Evolutionary cores of domain co-occurrence networks. *BMC Evolutionary Biology*, **5**, 24.

Xie, X., Jin, J., and Mao, Y. (2011) Evolutionary versatility of eukaryotic protein domains revealed by their bigram networks. *BMC Evoluntionary Biology*, **11**, 242.